Optimal Local Model Averaging for Divergent-Dimensional Functional-Coefficient Regressions*†

Yuying Sun[1,2], Shaoxin Hong[3‡], Zongwu Cai[4]

[1]*Academy of Mathematics and Systems Science & Center for Forecasting Science, Chinese Academy of Sciences, Beijing, Beijing 100190, China*

[2]*School of Economics and Management, University of Chinese Academy of Sciences, Beijing, Beijing 100190, China*

[3]*Center for Economic Research, Shandong University, Jinan, Shandong 250100, China*

[4]*Department of Economics, University of Kansas, Lawrence, KS 66045, USA*

ABSTRACT

This paper proposes a novel local model averaging estimator for divergent-dimensional functional-coefficient regressions, which selects optimal functional combination weights by minimizing a local leave-h-out forward-validation criterion. It is shown that the proposed leave-h-out forward-validation model averaging (FVMA) estimator is asymptotically optimal in the sense of achieving the lowest possible local squared error loss in a class of functional model averaging estimators, which is also extended to the ultra-high dimensional framework. The rate of the FVMA-based varying-weights converging to the optimal weights minimizing the expected local quadratic errors is derived. Besides, when correctly specified models are included in the candidate model set, the proposed FVMA asymptotically assigns all varying-weights to the correctly specified models. Furthermore, a simulation study and an empirical application highlight the merits of the proposed FVMA estimator relative to a variety of popular estimators with constant model averaging weights and model selection.

KEY WORDS: Asymptotic optimality; Functional-coefficient models; Forward-validation; Model averaging; Varying-weights

JEL Classification: C2, C13.

*Yuying Sun gratefully acknowledges the research support from the National Natural Science Foundation of China with the grant numbers 72322016. Shaoxin Hong gratefully acknowledges the research support from the National Natural Science Foundation of China with the grant numbers 72303131. Zongwu Cai gratefully acknowledges the research support from the National Natural Science Foundation of China with the grant numbers 72033008 (Key Project).

†E-mails: sunyuying@amss.ac.cn (Y. Sun); henryhong@sdu.edu.cn (S. Hong), and caiz@ku.edu (Z. Cai).

‡Corresponding author: Shaoxin Hong, E-mail: henryhong@sdu.edu.cn.

1 Introduction

Rational decisions and forecasts are often influenced by various economic and financial factors, such as monetary policies, interest rates, inflation, and business sentiment. For instance, in finance, Jansen et al. (2008) investigated the role of fiscal policy in explaining the behavior of the U.S. stock and bond markets, which was further studied by Tu & Wang (2020). In asset pricing models, factor loadings are considered as functions of certain state variables, which represent the unobserved information set of investors (Roussanov, 2014; Cai et al., 2015a,b, 2022). In labor economics, the marginal returns to education are dependent on work experience, assuming that work experience is a valued attribute by employers (Card, 2001; Cai et al., 2006). In exchange rate forecasting, Hong & Lee (2003) utilized a proxy variable to reveal useful information about the direction of changes, capturing nonlinearity in the mean for five exchange rates.

The functional coefficient model is a widely used nonparametric approach in applied science fields such as statistics and econometrics as well as finance to capture nonlinear features. This model allows coefficients to be represented as functions of observable state variables (Cai et al., 2000b, 2009; Jansen et al., 2008; Xiao, 2009; Phillips & Wang, 2022; Tu & Wang, 2020, 2022). For instance, the coefficients in the functional coefficient auto-regressive (FAR) model, which was initially proposed by Chen & Tsay (1993) and later extended by Cai et al. (2000b), are in unknown form depending on lagged terms. Notably, many well-known nonlinear models can be regarded as special cases of functional coefficient models. For example, the threshold model in Tong (1978) and Chan (1993) assumes the coefficients to be step functions of some observed state variables, including the lagged dependent variable. On the other hand, the smooth transition model proposed by Teräsvirta (1994) considers logistic functions of state variables.

Given the available data, we encounter a substantial number of functional coefficient candidate models. One popular approach is model selection which aims to choose an optimal model for making a prediction. Popular model selection criteria includes the nonparametric-

version of the bias-corrected AIC (Cai & Xu, 2008; Cai et al., 2015b), regularization method (Zou, 2006), and dimension reduction. However, many studies have shown that procedures that select the best model from a set, particularly in regression analysis, are intrinsically unstable (Stock & Watson, 2012).

Unlike model selection, model averaging incorporates all available information and constructs a weighted average of all potential candidate models. It is expected that model averaging serves as a form of insurance against selecting a poor candidate model (Hansen, 2014; Zhang & Zhang, 2022) and enhances robustness against model misspecification biases (Hsiao & Wan, 2014). Model averaging can be roughly categorized into Bayesian model averaging (BMA) and frequentist model averaging (FMA). For a literature review, refer to Claeskens & Hjort (2008) and Steel (2020).In contrast to BMA, where models are weighted based on posterior probabilities, FMA has gained increasing attention over the past decades. Various strategies include Mallows model averaging (Hansen, 2007; Zhu et al., 2019), jackknife model averaging (Hansen & Racine, 2012b), leave-subject-out cross-validation (Gao et al., 2016), forward-validation (Zhang & Zhang, 2022), k-fold cross-validation (Zhang & Liu, 2022), and AdaBoost semiparametric model averaging (Li et al., 2022). For instance, Zhu et al. (2019) proposed a Mallows-type model averaging for semiparametric varying-coefficient partially linear model and demonstrated the asymptotic optimality of the selected constant weights.

The aforementioned model averaging approaches are designed to select optimal constant combination weights for candidate models. To the best of our knowledge, there are only two papers to select the optimal non-constant weights in the model averaging literature. Specifically, Sun et al. (2021) proposed time-varying model averaging estimators based on local jackknife criterion. It is shown that the selected weight achieves the asymptotic optimality and the proposed model averaging estimator is consistent under certain mild conditions. Subsequently, Sun et al. (2022) derived the asymptotic normality of the penalized time-varying model averaging estimators, when the true model is included in candidate models. However, these two works mainly focus on time-varying coefficient regressions with low-dimensional covariates, potentially overlooking valuable information from the thousands

of predictors available in sophisticated information systems. When dealing with a large set of potential covariates, there arises substantial model uncertainty. Therefore, it becomes highly desirable to reduce model uncertainty and enhance forecast accuracy in functional-coefficient regressions with high-dimensional covariates.

Our attempt in this article is at developing an optimal model averaging method with varying weights for functional-coefficient regressions. However, compared to the case with constant weights, we face three distinct challenges. First, we need to devise a suitable local weight choice criterion, which varies over state variables. Existing literature typically considers the unbiased estimator of the quadratic loss risk over the entire sample, and thus, the selected weights are constant. Instead, our approach allows weights to change over state variables, rendering the traditional weight choice criterion unsuitable. Second, we seek to establish the asymptotic optimality and consistency of the combination weight estimator. However, proving these properties becomes significantly more intricate than in the constant weight setting. For example, some desirable properties of the projection matrix, such as symmetry and idempodence, cannot be applied anymore. Moreover, we allow the number of covariates to increase as the sample size grows, leading to a divergence in both dimension and the number of candidate models. This considerably complicates the mathematical proof. For example, the rate of parameter estimators converging to the well-defined limits in high-dimensional misspecified models is different from that in low-dimensional framework; see Lemma 3 in Appendix. Additionally, specific conditions must be assumed to illustrate the relationships among the sample size, the number of candidate models, and the dimension.

To address these challenges, we propose a local forward-validation model averaging method to select varying weights for functional-coefficient candidate models. This weight choice criterion is designed for selecting optimal weights in out-of-sample forecasts and suitable for highly persistent time-series data. The asymptotic optimality and consistency will be established for both the diverging dimension of covariates and the diverging number of candidate models. Besides, when the correctly specified models are included in candidate models, we demonstrate that the proposed method assigns all weights to the correctly spec-

ified models at any fixed point. We further extend our work to the ultra-high dimensional framework and the asymptotic optimality is investigated accordingly. A simulation study and an empirical application highlights the merits of the proposed model averaging estimator, relative to various popular estimators with constant model averaging weights and model selection.

The remainder of this paper is organized as follows. Section 2 introduces the model averaging estimation across different functional-coefficient models, together with proposing the local weight choice criterion and extending to the ultra-high dimensional model framework. Section 3 derives the asymptotic properties of the proposed method. Sections 4 and 5 present the numerical results in simulation and real data example. Finally, Section 6 concludes the article. Mathematical proofs are relegated to Appendix.

2 Model and Its Implementation

2.1 Model Setup

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

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

where Y_{t+h} is a dependent variable, $\mathbf{X}_t = (X_{t1}, X_{t2}, \cdots, X_{tq})'$ is a vector of covariate, $\mu_t = m(\mathbf{u}, \mathbf{x}) = \mathbb{E}(Y_t|\mathbf{U}_t = \mathbf{u}, \mathbf{X}_t = \mathbf{x})$ is the multivariate regression function, and ϵ_{t+h} is unobservable disturbance with $\mathbb{E}(\epsilon_{t+h}|\mathbf{X}_t) = 0$ almost surely (a.s.). Here, both $\mathbf{X}_t$ and $\mathbf{U}_t$ with the joint distribution $f(\mathbf{x}, \mathbf{u})$ might be allowed to consist of some lagged values of Y_{t+h}. For notational simplicity, let $\mathbf{Y} = (Y_{1+h}, \cdots, Y_{T+h})'$ be a $T \times 1$ vector of the observed values of the dependent variable, $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_T)'$, $\mathbf{X} = (\mathbf{X}_1, \cdots, \mathbf{X}_T)'$ be a $T \times q$ covariate matrix, and $\boldsymbol{\epsilon} = (\epsilon_{1+h}, \cdots, \epsilon_{T+h})'$. 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 (2010), the functional-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 are flexible enough to cover many applications. Here are some examples, including but not limited to, the functional-coefficient autoregressive model, generalized exponential autoregressive model, and threshold autoregressive model; see, for example, Cai et al. (2000b) for details, and with U_t being time, poisson regression model with time-varying coefficients as in Cai et al. (2000a) and trending time series models studied in Cai (2007) and Chen & Hong (2012), etc. This setting is particularly appealing in modeling economic and financial data; see, for example, Cai (2010) for details. Here, to ease notation, it is assumed that $\mathbf{U}_t$ is an observable scalar smoothing variable. Of course, one can consider the multivariate case for the smoothing variable. But, the estimation procedure and asymptotic results still hold for the multivariate case with much complicated notation. Therefore, in what it follows, it is assumed that $p = 1$ and $\mathbf{U}_t$ is changed to U_t.

Indeed, the functional 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

$$m(u, x) = \sum_{j=1}^{\infty} \partial^j m(u, x)/\partial x^j|_{x=0} x^j = \sum_{j=0}^{\infty} a_j(u)\, z_j,$$

where $z_j = x^j$ for all j. Suppose the data generating process is a model including q regressors with nonzero coefficients, i.e., $Y_{t+h} = \mathbf{X}_t' \boldsymbol{\alpha}(U_t) + \epsilon_{t+h}$, where $\mathbf{X}_t$ is a $q \times 1$ vector of explanatory variables, and $\boldsymbol{\alpha}(U_t)$ is a $q \times 1$ functional coefficient vector. Here, each element in $\boldsymbol{\alpha}(U_t)$ is nonzero, and ϵ_{t+h} is unobservable disturbance with $\mathbb{E}(\epsilon_{t+h}|\mathbf{X}_t) = 0$ almost surely.

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

$$Y_{t+h} = \sum_{j=1}^{q_m} \alpha_j^{(m)}(U_t) X_{tj} + \epsilon_{t+h}^{(m)} \equiv \mathbf{X}_t^{(m)'} \boldsymbol{\alpha}^{(m)}(U_t) + \epsilon_{t+h}^{(m)} = \mu_t^{(m)} + \epsilon_{t+h}^{(m)},$$

where the functions $\{\alpha_j^{(m)}(\cdot)\}$ are measurable functions from $\mathbb{R}^p$ to $\mathbb{R}$ and $\mathbf{X}_t^{(m)} = (X_{t1}, \cdots, X_{tq_m})'$ is a $q_m \times 1$ vector of regressors and $\boldsymbol{\alpha}^{(m)}(U_t) = (\alpha_1^{(m)}(U_t), \cdots, \alpha_{q_m}^{(m)}(U_t))'$. Note that we allow each candidate model has a divergent dimension of regressors as the sample size T increases; that is, q_m grows to infinity at some slower rates than the sample size.

2.2 Estimation Procedure

The unknown coefficient functions can be estimated by using a local constant estimation technique. For any given u_0 and U_t in a neighborhood of u_0, it follows from a 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:

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

where $k_t = k((U_t - u_0)/l)$, $k(\cdot)$ is a kernel function on $\mathbb{R}^1$, and $l > 0$ is a bandwidth which satisfies $l \to 0$ as $T \to \infty$. Let $\mathbf{X}^{(m)}$ denote a $T \times q_m$ matrix with $\mathbf{X}_t^{(m)'}$ as its t-th row, and $\mathbf{K}(u_0) = \text{diag}\{k_1, \cdots, k_T\}$, 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 $\boldsymbol{\alpha}^{(m)}(u_0) \equiv (\alpha_1^{(m)}(u_0), \cdots, \alpha_{q_m}^{(m)}(u_0))'$. 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 jth position. Define $\mathbf{P}_t^{(m)} \equiv \mathbf{P}^{(m)}(U_t) = \left[\mathbf{X}^{(m)'}\mathbf{K}(U_t)\mathbf{X}^{(m)}\right]^{-1}\mathbf{X}^{(m)'}\mathbf{K}(U_t)$ as a $q_m \times T$ 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 as follows:

$$\widehat{\boldsymbol{\mu}}^{(m)} = \begin{pmatrix} \widehat{\mu}^{(m)}(U_1) \\ \vdots \\ \widehat{\mu}^{(m)}(U_T) \end{pmatrix} \equiv \begin{pmatrix} \mathbf{X}_1^{(m)\prime}\mathbf{P}_1^{(m)} \\ \vdots \\ \mathbf{X}_T^{(m)\prime}\mathbf{P}_T^{(m)} \end{pmatrix} \mathbf{Y} = \mathbf{P}^{(m)}(\mathbf{X})\mathbf{Y}, \tag{3}$$

where the definition of $\mathbf{P}^{(m)}(\mathbf{X})$ is obvious in (3).

2.3 Selection of Local Weight

Let $\mathbf{w} = (w^1, \cdots, w^{M_T})'$ be a vector of weights in the unit simplex of $\mathbb{R}^{M_T}$, i.e., $\mathcal{H}_T = \left\{\mathbf{w} \in [0,1]^{M_T} : \sum_{m=1}^{M_T} w^m = 1\right\}$. Actually, these weights can be allowed to be dynamic such that they can be functional weights of some information. For simplicity, we write $\mathbf{w} = \mathbf{w}(u_0)$ with $w^m = w^m(u_0)$ given u_0. Define $\mathbf{P}(\mathbf{w}, \mathbf{X}) = \sum_{m=1}^{M_T} w^m \mathbf{P}^{(m)}(\mathbf{X})$. For given $\mathbf{w}$, in view of (3), an averaging estimator for the conditional mean is given by

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

and the model averaging estimator for $\boldsymbol{\alpha}(u_0)$ is given by

$$\widehat{\boldsymbol{\alpha}}(u_0, \mathbf{w}) = \sum_{m=1}^{M_T} w^m \boldsymbol{\Pi}^{(m)\prime} \widehat{\boldsymbol{\alpha}}^{(m)}(u_0),$$

where $\boldsymbol{\Pi}^{(m)}$ is a projection matrix of size $q_m \times q$ mapping $\boldsymbol{\alpha}(u_0)$ to $\boldsymbol{\alpha}^{(m)}(u_0)$.

Within heteroskedastic or autocorrelated errors, we propose the leave-h-out forward-validation estimator in the functional-coefficient linear regression model. Denote two selected matrixes as $\boldsymbol{\phi}_t = (\mathbf{I}_t, \mathbf{0}_{t\times(T-t)})$ for $1 < t \leq h$ and $\boldsymbol{\phi}_t = (\mathbf{0}_{h\times(t-h)}, \mathbf{I}_h, \mathbf{0}_{h\times(T-t)})$ for $h+1 \leq t \leq T$, and $\boldsymbol{\pi}_t = (\mathbf{0}_{1\times(t-1)}, 1)$ for $1 < t \leq h$ and $\boldsymbol{\pi}_t = (\mathbf{0}_{1\times(h-1)}, 1)$ for $h+1 \leq t \leq T$. Then, we obtain $\mathbf{Y}_{[t+h]} = \boldsymbol{\phi}_t \mathbf{Y}$ and $\mathbf{X}_{[t]}^{(m)} = \boldsymbol{\phi}_t \mathbf{X}^{(m)}$, which are the sets to be removed. Denote $\mathbf{Y}_{[-(t+h)]}$ and $\mathbf{X}_{[-t]}^{(m)}$ as the remaining sets of $\mathbf{Y}$ and $\mathbf{X}^{(m)}$ after removing $\mathbf{Y}_{[t+h]}$ and $\mathbf{X}_{[t]}^{(m)}$, respectively. For any fixed u_0, the following local constant estimator $\widehat{\boldsymbol{\alpha}}_{[-t]}^{(m)}$ is obtained from $\mathbf{Y}_{[-(t+h)]}$ and $\mathbf{X}_{[-t]}^{(m)}$:

$$\widehat{\boldsymbol{\alpha}}_{[-t]}^{(m)}(u_0) = \left(\mathbf{X}_{[-t]}^{(m)\prime}\mathbf{K}_{[-t]}(u_0)\mathbf{X}_{[-t]}^{(m)}\right)^{-1} \mathbf{X}_{[-t]}^{(m)\prime}\mathbf{K}_{[-t]}(u_0)\mathbf{Y}_{[-(t+h)]},$$

where $\mathbf{K}_{[-t]}(u_0) = \text{diag}\{k_1, \cdots, k_{t-h}, k_{t+1}, \cdots, k_T\}$ and the leave-h-out forward-validation estimator $\widetilde{\mu}_t^{(m)}(u_0)$ of $\mu_t^{(m)}$ is

$$\widetilde{\mu}_t^{(m)}(u_0) = \boldsymbol{\pi}_t \mathbf{X}_{[t]}^{(m)} \widehat{\boldsymbol{\alpha}}_{[-t]}^{(m)}(u_0). \tag{4}$$

Thus, the leave-h-out 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 $\widetilde{\boldsymbol{\mu}}(\mathbf{w}) = (\widetilde{\mu}_1(\mathbf{w}), \cdots, \widetilde{\mu}_T(\mathbf{w}))'$.

Now, define the local squared loss of $\boldsymbol{\mu}(\mathbf{w})$ as follows:

$$L_T(u_0, \mathbf{w}) = (\widehat{\boldsymbol{\mu}}(\mathbf{w}) - \boldsymbol{\mu})' \mathbf{K}(u_0) (\widehat{\boldsymbol{\mu}}(\mathbf{w}) - \boldsymbol{\mu}),$$

which is infeasible because of the unknown conditional mean $\boldsymbol{\mu}$. Thus, we propose the feasible leave-h-out forward validation criteria to develop the corresponding local constant averaging estimators

$$\text{FV}_T(u_0, \mathbf{w}) = (\mathbf{Y} - \widetilde{\boldsymbol{\mu}}(\mathbf{w}))' \mathbf{K}(u_0) (\mathbf{Y} - \widetilde{\boldsymbol{\mu}}(\mathbf{w})). \tag{5}$$

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

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

Then, the FVMA estimator of conditional mean at time t is $\widehat{\mu}_t(\widehat{\mathbf{w}}_{U_t})$, which is in a dynamic way.

Actually, the proposed model averaging procedure's implementation can be formulated as a quadratic programming problem. This formulation involves minimizing a quadratic objective function subject to linear constraints, and the following algorithm illustrates the computational procedure.

Algorithm 1: An Algorithm for Computing $\mathbf{w}$.

For any given u_0,

Step 1: Calculate the leave-h-out forward-validation estimator for μ_t under every candidate model.

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

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

    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.(4).

  end

  Calculate $\widehat{\boldsymbol{\mu}}^{(m)}$ by Eq.(3).

end

Step 2: Calculate the model averaging weight based on the local weight choice criterion.

(2.1) Calculate $\widetilde{\mathbf{e}} = (\mathbf{Y}, \cdots, \mathbf{Y}) - \begin{pmatrix} \widetilde{\mu}_1^{(1)}(u_0) & \cdots & \widetilde{\mu}_1^{(M_T)}(u_0) \\ \vdots & & \vdots \\ \widetilde{\mu}_T^{(1)}(u_0) & \cdots & \widetilde{\mu}_T^{(M_T)}(u_0) \end{pmatrix}$;

(2.2) Solve the constrained quadratic programming problem to obtain the model averaging weight

$$\widehat{\mathbf{w}}_{u_0} = \arg\min_{\mathbf{w}\in\mathcal{H}_T} \mathrm{FV}_T(u_0, \mathbf{w}),$$

where $\mathrm{FV}_T(u_0, \mathbf{w}) = (\mathbf{Y} - \widetilde{\boldsymbol{\mu}}(\mathbf{w}))'\mathbf{K}(u_0)(\mathbf{Y} - \widetilde{\boldsymbol{\mu}}(\mathbf{w})) = \mathbf{w}'\widetilde{\mathbf{e}}'\mathbf{K}(u_0)\widetilde{\mathbf{e}}\mathbf{w}$.

Output: $\widehat{\mathbf{w}}_{u_0}$

Note: Numerical solutions can be obtained using various optimization software packages. For instance, the *quadprog* package in the R language and the *quadprog* command in MATLAB are commonly used to solve such problems.

2.4 Extension to Ultra-High Dimensional Framework

So far, 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. This section is mainly motivated by the attempt to address the dimensionality issue encountered in regression problems with $q > T$ and reduce the computational burden of model averaging procedure. There are two steps involved.

In Step 1, we use model screening to prepare candidate models, which essentially selects a valid subset of all candidate models. Let M^* be a subset of $\{1, \cdots, M_T\}$ and thus, $\mathcal{H}_T^* =$

$\{\mathbf{w} \in [0,1]^{M_T} : \sum_{m\in M^*} w^m = 1 \text{ and } \sum_{m\notin M^*} w^m = 0\}$. Various model screening strategies are proposed in the existing literature, including threshold model screening (Zhang et al., 2016), top s model screening (Yuan & Yang, 2005), ordering model screening (Claeskens et al., 2006), and others. For example, we could follow Ando & Li (2014, 2017) to do 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. Then, we divide the q marginal correlation into $M_T^* + 1$ groups based on the ordering. The first group has the highest values, while the last group has the correlations closest to zero. And in each group, q_m is smaller than the sample size T. After that, we discard the $M_T^* + 1$ group and thus, the number of candidate mode is M_T^*. In this case, $M^* = \{1, \cdots, M_T^*\}$.

In Step 2, we construct model averaging based on the subset M^*, and the weight vector is derived from

$$\widehat{\mathbf{w}}_{u_0}^* = \arg\min_{\mathbf{w}\in\mathcal{H}_T^*} L_T(u_0, \mathbf{w}),$$

which will be shown to be asymptotically optimal under some regularity conditions, see, Theorem 5 later.

3 Asymptotic Properties

Let $\boldsymbol{\alpha}^{(m)*}(u_0)$ be parameter vector which is essentially derived from minimizing the MSE between Y_{t+h} and the mth candidate model at the point u_0, i.e.,

$$\boldsymbol{\alpha}^{(m)*}(u_0) = \arg\min_{\boldsymbol{\alpha}^{(m)}(u_0)} \mathbb{E}[Y_{t+h} - \mathbf{X}_t^{(m)'}\boldsymbol{\alpha}^{(m)}(u_0)]^2 = [\mathbb{E}(\mathbf{X}_t^{(m)}\mathbf{X}_t^{(m)'})]^{-1}\mathbb{E}(\mathbf{X}_t^{(m)}Y_{t+h}). \quad (6)$$

From Lemma 3 in Appendix, we know that under certain regularity conditions,

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

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 of 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* $O_p(qT^{-1}l^{-1})$*. Thus, the bias term is ignored in (6).*

Denote $L_T^*(u_0, \mathbf{w}) \equiv [\boldsymbol{\mu}^*(\mathbf{w}) - \boldsymbol{\mu}]'\mathbf{K}(u_0)[\boldsymbol{\mu}^*(\mathbf{w}) - \boldsymbol{\mu}]$, $\boldsymbol{\mu}^*(\mathbf{w}) \equiv \sum_{m=1}^{M_T} w^m \boldsymbol{\mu}^{(m)*}$, $\boldsymbol{\mu}^{(m)*} \equiv \widehat{\boldsymbol{\mu}}^{(m)}|_{\widehat{\boldsymbol{\alpha}}^{(m)}(u_0)=\boldsymbol{\alpha}^{(m)*}(u_0)}$ and $\xi_T(u_0) = \inf_{\mathbf{w}\in\mathcal{H}_T} \mathbb{E}L_T^*(u_0, \mathbf{w})$. Let $f(u, \mathbf{x})$ denote the joint density of $(U, \mathbf{X})$, $f_U(u)$ be the marginal density of U, $\zeta_{\max}(\mathbf{A})$ and $\zeta_{\min}(\mathbf{A})$ denote the maximum and minimum singular value of a matrix $\mathbf{A}$, respectively. Unless stated otherwise, all limiting processes refer to $T \to \infty$. Our derivation of the asymptotic optimality requires the following conditions.

Condition (C.1). *For all* $s \geq 1$ *and some positive constant* C, $|f(u, v|\mathbf{x}_0, \mathbf{x}_1; s)| \leq C < \infty$*, where* $f(u, v|\mathbf{x}_0, \mathbf{x}_1; s)$ *is the conditional density of* (U_0, U_s) *given* $(\mathbf{X}_0, \mathbf{X}_s)$*, and* $f(u|\mathbf{x}) \leq C < \infty$*, where* $f(u|\mathbf{x})$ *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* $c > 1-2/\iota$, $\sup_{1\leq t\leq T} T^{-1}l^{-1}||\mathbf{K}(u_0)\boldsymbol{\mu}||^2 = O_p(1)$, $\sup_{1\leq t\leq T} ||\mathbf{X}_t^{(m)}||/\sqrt{q} = O_p(1)$ *uniformly for all* m*, and* $\zeta_{\min}(T^{-1}l^{-1}\mathbf{X}^{(m)'}\mathbf{K}(u_0)\mathbf{X}^{(m)}) \geq C_0$ *for some positive constant* C_0*.*

Condition (C.3). *The error term* $\{\epsilon_{t+h}\}$ *is weakly stationary and satisfies* $\mathbb{E}(\epsilon_{t+h}|\mathbf{X}_t, U_t) = 0$ *almost surely,* $\mathbb{E}(\epsilon_{t+h}^2) = \sigma^2$ *and* $\mathbb{E}(\epsilon_{t+h}^2|I_t) = \sigma^2(I_t)$*.*

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* $\int_{-1}^{1} k(u)u^2du < \infty$*.*

Condition (C.5). *The bandwidth* $l = cT^{-1/5+\nu}$ *for some* $-4/5 < \nu \leq 0$ *and* $0 < c < \infty$*.*

Condition (C.6). *For any fixed* u_0, $qM_TT^{-1}l^{-1} = o(1)$ *and* $qM_T^{1/2}T^{1/2}l^{1/2}\xi_T^{-1}(u_0) = o(1)$*,*

Remark 2. *Condition (C.1) is the same as Condition 1 (ii) in Cai et al. (2000b), which is a standard condition for functional-coefficient regression models. Condition (C.2) imposes a standard requirement for the mixing coefficient and moments, which is commonly used in the*

existing literature (Fan & Yao, 2003). Condition (C.3) imposes that the forecast error is a martingale difference sequence when $h = 1$, and allows a non-diagonal covariance structure for regression errors with bounded eigenvalues.

Remark 3. *Condition (C.4) requires the two-sided kernel to be symmetric and bounded with a compact support $[-1, 1]$. Note that in out-of-sample forecasting, the functional-coefficient regression models need the two-sided kernel instead of one-sided kernel with a compact support $[-1, 0]$ in time-varying coefficient regression models. The commonly used kernels including the Epanechnikov and uniform kernels satisfy Condition (C.4). If $\nu = 0$, the optimal bandwidth $h_{opt} = O(T^{-1/5})$ satisfies Condition (C.5).*

Remark 4. *Condition (C.6) requires that $\xi_T(u_0)$ grows at a faster rate than $qM_T^{1/2}T^{1/2}l^{1/2}$, which is similar to Condition 7 of Ando & Li (2014) and Condition (C.6) of Zhang et al. (2016). Note that this condition implies $\xi_T(u_0) \to \infty$, which requires that all candidate models are misspecified. Specifically, suppose the m_0-th candidate model is correctly specified, then, we have $\boldsymbol{\alpha}^{(m_0)*}(u_0) = \boldsymbol{\alpha}(u_0)$, where $\boldsymbol{\alpha}(u_0)$ is the true value defined in date generating process. Thus, we have*

$$\xi_T(u_0) = \inf_{\mathbf{w}\in\mathcal{H}_T} \mathbb{E}[\boldsymbol{\mu}^*(\mathbf{w}) - \boldsymbol{\mu}]'\mathbf{K}(u_0)[\boldsymbol{\mu}^*(\mathbf{w}) - \boldsymbol{\mu}] \leq \mathbb{E}[\boldsymbol{\mu}^{(m_0)*} - \boldsymbol{\mu}]'\mathbf{K}(u_0)[\boldsymbol{\mu}^{(m_0)*} - \boldsymbol{\mu}] = 0,$$

and then, Condition (C.6) is violated. We first discuss asymptotic optimality when all candidate models to be misspeficified, and then, discuss the alternative cases where some models are correctly specified.

The following theorem states that the proposed criterion has the asymptotic optimality for diverging q_m, when all candidate models are misspecified.

Theorem 1 (Asymptotic Optimality)**.** *Suppose that Conditions (C.1)-(C.6) hold. Then, for any given point u_0, the FVMA estimator satisfies the asymptotic optimality (OPT) property, i.e.,*

$$\frac{L_T(u_0, \widehat{\mathbf{w}}_{u_0})}{\inf_{\mathbf{w}\in\mathcal{H}_T} L_T(u_0, \mathbf{w})} \xrightarrow{p} 1,$$

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

Theorem 1 shows that for any given u_0, the model averaging procedure is asymptotically optimal in the sense that its local squared loss is asymptotically identical to that of the infeasible but best possible model averaging estimator. This provides theoretical support for the advantages of the proposed method over other averaging or selection methods, including SAIC or SBIC, because the infeasible local loss of the best possible model averaging estimator is smaller or equal to that of other model averaging estimators and model selection estimators.

For any given u_0, denote the optimal weight $\mathbf{w}^0(u_0) = \arg\min_{\mathbf{w}\in\mathcal{H}_T} \mathbb{E}[L_T(u_0, \mathbf{w})]$, and $\widetilde{\xi}_T(u_0) = \min_{\mathbf{w}\in\mathcal{H}_T} \mathbb{E}[L_T(u_0, \mathbf{w})]$.

Condition (C.7). *For any given u_0, $\kappa_1 < \zeta_{\min}(T^{-1}l^{-1}\mathbf{\Lambda}'\mathbf{K}(u_0)\mathbf{\Lambda}) \leq \zeta_{\max}(T^{-1}l^{-1}\mathbf{\Lambda}'\mathbf{K}(u_0)\mathbf{\Lambda}) < \kappa_2 < \infty$ for some positive constants κ_1 and κ_2, where $\mathbf{\Lambda}$ is a $T \times M_T$ matrix with $\widehat{\mu}_s^{(m)}$ in its (s, m)th element.*

Condition (C.8). $\max_{1\leq m\leq M_T} \max_{1\leq t\leq T} P_t^{(m)} = O_p(M_T T^{-1} l^{-1})$, *where $P_t^{(m)}$ is the tth diagonal element of $\mathbf{P}^{(m)}(\mathbf{X})$.*

Condition (C.9). *For $1 \leq m \leq M_T$ and given u_0, $\zeta_{\max}(T^{-1}l^{-1}\mathbf{P}^{(m)'}(\mathbf{X})\mathbf{K}(u_0)\mathbf{P}^{(m)}(\mathbf{X})) = O_p(q)$ a.s., and $\Pr(\zeta_{\min}(T^{-1}l^{-1}(\mathbf{P}^{(m)'}(\mathbf{X})\mathbf{K}(u_0)\mathbf{P}^{(m)}(\mathbf{X})) > C > 0))$ tends to 1 for some positive constant C.*

Condition (C.10). *(i) $\widetilde{\xi}_T^{-1}(u_0)T^{-2\delta}l^{-2\delta}M_T^2 q = o(1)$ and $M_T^{1/2}q^{-1/2}T^{-1/2-\delta}l^{-1/2-\delta} = o(1)$, and (ii) $M_T^{3/2}q^{3/2}T^{-1/2+\delta}l^{-1/2+\delta} = o(1)$, where δ is a positive constant.*

Remark 5. *Condition (C.7) is similar to Condition (C.9) in Liao et al. (2019), which requires that the minimum and maximum singular values of $\mathbf{\Lambda}'\mathbf{K}(u_0)\mathbf{\Lambda}/(Tl)$ are asymptotically bounded. Condition (C.8) is similar to Condition (C.3) of Liao et al. (2019), which is related to cross-validation methods (Li, 1987; Gao et al., 2016). Condition (C.9) is commonly used in Fan & Peng (2004); Li et al. (2022). This condition is rather mild, because typical estimators satisfy the regularity condition that the maximum singular value of the corresponding matrix is bounded.*

Remark 6. *Condition (C.10) illustrates the relationships among $\widetilde{\xi}_T(u_0)$, Tl, M_T and q. Similar conditions can be found in the model averaging literature; see Liao et al. (2019); Li et al. (2022). Note that condition (C.10) allows all candidate models to be misspecified, as well as correctly specified models included. For example, suppose the m_0-th candidate model is correctly specified, then, we have $||\widehat{\boldsymbol{\alpha}}^{(m_0)}(u_0) - \boldsymbol{\alpha}(u_0)|| = O_p(q^{1/2}T^{-1/2}l^{-1/2})$ based on Lemma 3. Thus, we have*

$$\widetilde{\xi}_T(u_0) = \inf_{\mathbf{w}\in\mathcal{H}_T} \mathbb{E}[\widehat{\boldsymbol{\mu}}(\mathbf{w}) - \boldsymbol{\mu}]'\mathbf{K}(u_0)[\widehat{\boldsymbol{\mu}}(\mathbf{w}) - \boldsymbol{\mu}] = O_p(1),$$

and then, Condition (C.10) still holds.

Theorem 2 (Consistency of Weights Estimation)**.** *Suppose that Conditions (C.1)-(C.5) and (C.7)-(C.10) hold. Then, for any given u_0, there exists a local minimizer $\widehat{\mathbf{w}}_{u_0}$ of $FV_T(u_0, \mathbf{w})$ such that*

$$||\widehat{\mathbf{w}}_{u_0} - \mathbf{w}^0(u_0)|| = O_p(M_T q T^{-1/2+\delta} l^{-1/2+\delta}),$$

where δ is a positive constant given in Condition (C.10).

Remark 7. *From Theorem 2, it is observed that for any given u_0, $\widehat{\mathbf{w}}_{u_0}$ converges to the optimal weight $\mathbf{w}^0(u_0)$ at the rate $M_T q T^{-1/2+\delta} l^{-1/2+\delta}$. Given u_0 and the rate of $Tl \to \infty$, the slower the rates of $M_T \to \infty$ and $q \to \infty$, the faster the rate of $\widehat{\mathbf{w}}_{u_0}$ approaching to $\mathbf{w}^0(u_0)$ in probability. Theorem 2 holds in the case where all candidate models are misspeficied, as well as the alternative case where some models are correctly specified.*

Remark 8. *A linear regression model is correctly specified for $\mathbb{E}(Y_{t+h}|\mathbf{X}_t, \mathbf{U}_t)$ if $\mathbb{E}(Y_{t+h}|\mathbf{X}_t, \mathbf{U}_t) = \mathbf{X}_l'\boldsymbol{\alpha}(\mathbf{U}_t)$ for some $\boldsymbol{\alpha}(\mathbf{U}_t)$, which is equivalent to the condition that*

$$\mathbb{E}(\epsilon_{t+h}|\mathbf{X}_t, \mathbf{U}_t) = 0.$$

That is, correct model specification occurs if and only if the conditional mean of the linear regression error is zero. See more discussions in Hong (2005).

Theorem 3 (Consistency of MA Parameter Estimation)**.** *Suppose Conditions (C.1)-(C.5) and (C.7)-(C.10) hold. Then, for any given* u_0*,*

$$||\widehat{\boldsymbol{\alpha}}(u_0, \widehat{\mathbf{w}}_{u_0}) - \boldsymbol{\alpha}^*(u_0, \mathbf{w}^0(u_0))|| = O_p(M_T^{3/2} q^{3/2} T^{-1/2+\delta} l^{-1/2+\delta}),$$

where $\boldsymbol{\alpha}^*(u_0, \mathbf{w}^0(u_0)) = \sum_{m=1}^{M_T} w_m^0(u_0) \boldsymbol{\alpha}^{(m)*}(u_0)$*,* $w_m^0(u_0)$ *is the* m*-th element of* $\mathbf{w}^0(u_0)$*, and* $\boldsymbol{\alpha}^{(m)*}(u_0)$ *is defined in Remark 1.*

Theorem 3 shows that for any give u_0, the model averaging estimator $\widehat{\boldsymbol{\alpha}}(u_0, \widehat{\mathbf{w}}^0(u_0))$ converges to a well-defined limit $\boldsymbol{\alpha}^*(u_0, \mathbf{w}^0(u_0))$, even all candidate models are misspecified.

Next, we discuss whether the proposed local averaging estimator asymptotically assigns all weights to the correctly specified models, if they are included in candidate models.

Condition (C.11). *For any fixed* u_0*,* $qM_T T^{-1} l^{-1} = o(1)$ *and*

$$qM_T^{1/2} T^{1/2} l^{1/2} \{ \inf_{\mathbf{w} \in \widetilde{\mathcal{H}}_T} \mathbb{E} L_T^*(u_0, \mathbf{w}) \}^{-1} = o(1),$$

where $\widetilde{\mathcal{H}}_T = \{\mathbf{w} \in [0,1]^{M_T} : \sum_{m \notin \mathcal{D}} w^m = 1\}$ *and* $\mathcal{D}$ *is the subset of* $\{1, \cdots, M_T\}$ *which is composed of the correctly specified models.*

Remark 9. *Condition (C.11) is essentially equivalent to Condition (C.6), if* $\mathcal{D}$ *is empty, that is, all candidate models are misspecified.*

Theorem 4. *If there is one or more correctly specified models, and Conditions (C.1)-(C.5) and (C.11) are satisfied, then,*

$$\sum_{m \in \mathcal{D}} \widehat{w}_{u_0}^m \xrightarrow{p} 1,$$

where $\widehat{w}_{u_0}^m$ *is the* m*-th element of* $\widehat{\mathbf{w}}_{u_0}$*.*

Theorem 4 shows that the proposed criterion asymptotically assigns all weights to the the correctly specified models when the model set includes correctly specified models. If there is only one correctly specified model among the candidate models, Theorem 4 implies that the proposed criterion would select this correctly specified model asymptotically.

Finally, to establish the asymptotic theory for the model averaging estimator under the ultra-high dimensional framework as described in Section 2.4, the following condition is needed.

Condition (C.12). *For any fixed u_0, there exist a nonnegative series of $v_T(u_0)$ and a weight series of $\mathbf{w}_T \in \mathcal{H}_T$ such that $\xi_T^{-1}(u_0)v_T(u_0) \to 0$, $\inf_{\mathbf{w}\in\mathcal{H}_T} L_T(u_0, \mathbf{w}) = L_T(u_0, \mathbf{w}_T) - v_T(u_0)$, and $\Pr(\mathbf{w}_T \in \mathcal{H}_T^*) \to 1$ as $T \to \infty$.*

Theorem 5 (Asymptotic Optimality)**.** *Suppose Conditions (C.1)-(C.6) and (C.12) hold. Then, we have*

$$\frac{L_T(u_0, \widehat{\mathbf{w}}_{u_0}^*)}{\inf_{\mathbf{w}\in\mathcal{H}_T} L_T(u_0, \mathbf{w})} \xrightarrow{P} 1.$$

Theorem 5 states that under Condition (C.12) together with other conditions, the proposed model averaging estimator for the ultra-high dimensional case is still asymptotically optimal based on the model set $\mathcal{H}_T^*$.

4 Simulation Studies

In this section, we conducted simulations to compare the in-sample prediction and out of sample forecasting performance of different model averaging methods for horizons $h = 1, 2$ and 4:

- FVMA: the proposed method in the paper;
- FVMASA: the forward-validation model averaging estimator with uniform weights;
- AIC_c: the nonparametric version of bias-corrected AIC model selection by Cai & Tiwari (2010);
- SAICc, the smoothed AICc model averaging;
- SAIC: the smoothed Akaike information criterion model averaging as in Buckland et al. (1997);

- SBIC: the smoothed Bayesian information criterion model averaging;
- JMA: the jackknife model averaging initiated by Hansen & Racine (2012a).

In the following examples, we use 1000 replicates and for each replication, we draw a sample of size $T = 200$ and 600 from the data generating process (DGP, hereafter), respectively. Our simulation study is based on the DGP framework:

$$Y_{t+h} = \sum_{j=1}^{p} \alpha_j(X_{t,1}) X_{t,j} + \varepsilon_{t+h}, \quad t = 1, \cdots, T, \tag{7}$$

and comparisons between model averaging methods above are presented by considering different setting of the DGP. The Epanechnikov kernel function $K(u) = 0.75(1-u^2)\mathbb{I}_{|u|\leq 1}$ is employed in all simulation examples, and it down-weights more distant observations within the subsample. We also ran simulations with the Gaussian kernel, the results were similar and thus, omitted to save space. For each replication, we compute the mean squared error of model risk by $\text{MSE}^{(k)} = \frac{1}{T-1}\sum_{t=1}^{T-1}(\widehat{Y}_{t+h}^{(k)} - \mu_t^{(k)})^2$, where $\{\widehat{Y}_{t+h}^{(k)}, t = 1, \cdots, T-1\}$ is the in-sample prediction and $\mu_t^{(k)}$ is the conditional mean of $Y_{t+h}^{(k)}$, we report the $\text{MSE} = \frac{1}{1000}\sum_{k=1}^{1000}\text{MSE}^{(k)}$ for all methods. For the out of sample forecast error, denote $\widehat{Y}_{T+h}^{(k)}$ be the out of sample forecast of $Y_{T+h}^{(k)}$ in the kth simulation, then, MSFE can be given by $\text{MSFE} = \frac{1}{1000}\sum_{k=1}^{1000}(\widehat{Y}_{T+h}^{(k)} - Y_{T+h}^{(k)})^2$ for all methods.

Example 1: We consider model (7) with $\alpha_j(u) = [1+\exp(-\frac{cu}{j})]^{-1}$ and $\varepsilon_t \sim N(0, 0.3^2)$. We generate the predicting variables from ARMA processes, to be specific, $X_{t,1} = 0.8X_{t-1,1} + v_{t,1}$, $v_{t,1} \sim N(0,1)$ and $X_{t,2} = X_{t-1,1}$. $X_{t,3} = 0.6X_{t-1,3} + 0.3v_{t-1,3} + v_{t,3}$, $v_{t,3} \sim N(0,1)$, and $X_{t,4} = X_{t-1,3}$, $X_{t,5} = X_{t-2,3}$. That is, we have some predictors are lag variables. Furthermore, we consider $X_{t,j} = (-0.3 + 0.1j)X_{t-1,j} + v_{t,j}$, $v_{t,j} \sim N(0, s_j^2)$ for $j > 5$, $\{s_j\}$ is a random sample generated from the Chi-square distribution χ_1^2, then the conditional variances of these predictors are not identical in general. We set $c = 2$ and it is assumed the number of the predictors $p = 10$.

The proposed data generating process in this example includes functional coefficients $\alpha_j(\cdot)$, which are dependent on the value of j and are modeled as *logit* functions. This design

allows for the contribution of different predictor variables to the conditional mean to vary. Specifically, the functional form of $\alpha_j(\cdot)$ is such that the curvature of the *logit* function tends to flatten as j increases. As a result, both linear and nonlinear types of functional coefficients are incorporated into the model. This approach enables the model to capture relationships between the predictors and the response variable, including both linear and nonlinear associations.

Example 2: The setup for this simulation example is adapted from Example 1, but with functional coefficients $\alpha_j(u) = [1 + \exp((-1)^j \cdot \frac{cu}{j})]^{-1}$ for $j = 1, \cdots, p$. The key innovation of this design is that the functional coefficients are dependent on both the state variable u and the index j. Specifically, the parity of j determines the direction of the relationship between the predicting variable and the response variable, such that 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 nearly identical to that of Example 1, with the exception of the functional coefficients $\alpha_j(u)$, which are defined as follows:

$$\alpha_j(u) = \begin{cases} \sqrt{2c} j^{-c} \exp(-3u^2) & j = 1, 2; \\ (1-\alpha)\alpha^j u & j = 3, 4, 5; \\ (u^2 - ju)/3j & j > 5. \end{cases}$$

Here, $c = 1$, $\alpha = 0.95$, and we have a total of $p = 8$ predictors. The functional coefficients in the proposed design are defined in a way that the underlying relationships between the predictors and the response variable may be complex in practice. To model the complex relationships between the predictors and the response variable, we partition the predictors into three groups based on their characteristics, and design functional coefficients for each group. To be more specific, the first two functional coefficients ($j = 1, 2$) are formulated as a combination of a decaying exponential function and a power function of j. These coefficients correspond to the predictors $X_{t,1} = 0.8X_{t-1,1} + v_{t,1}$ and $X_{t,2} = X_{t-1,1}$. This functional form facilitates the capture of nonlinear relationships between the predictors and the response variable. The functional coefficients for the third, fourth, and fifth predictors

$(j = 3, 4, 5)$ are structured as linear functions of u. These coefficients correspond to the predictors $X_{t,3} = 0.6X_{t-1,3} + 0.3v_{t-1,3} + v_{t,3}$, $X_{t,4} = X_{t-1,3}$, and $X_{t,5} = X_{t-2,3}$. Furthermore, for predictors with indices greater than 5 $(j > 5)$, denoted by $X_{t,j} = (-0.3+0.1j)X_{t-1,j}+v_{t,j}$, the functional coefficients adopt a quadratic form in terms of u, where j serves as the scaling factor.

Example 4: In this example, we consider a scenario of diverging dimensionality where the number of predicting variables p grows without bound as the sample size T converges to infinity. Specifically, we use $p = \lfloor 3T^{1/3} \rceil$ where $\lfloor x \rceil$ denotes the rounding of x to the nearest integer. This choice of p ensures that the growth rate of p is not too fast relative to the sample size T, which is a common consideration in the literature on high-dimensional statistical modeling. This choice also ensures that there is enough sample size to estimate the functional coefficients in our model, which is crucial for obtaining accurate and reliable estimates.

We employ the same setting of predicting variables as in Example 1, and for the functional form of coefficients, we use $\alpha_j(u) = \sqrt{2}j^{-1}\exp(-3u^2)$. One notable feature of these functional coefficients is that their values shrink to zero as j increases, indicating that the contribution of the corresponding predicting variable $X_{t,j}$ becomes increasingly insignificant as j grows larger. This phenomenon is analogous to a situation that a regressor with nuisance parameter in high-dimensional linear regression model. It is common to encounter situations where some predictors are only weakly associated with the response variable in high-dimensional setting. Their inclusion in the model may lead to increased variance and reduced efficiency in estimating the coefficients of interest. Such predictors are often referred to as "nuisance" predictors, as they add noise to the model but do not contribute much to the estimation.

The numerical results are reported in Tables 1 for Example 1 in the top panel and Example 2 in the bottom panel and 2 for Example 3 in the top panel and Example 4 in the bottom panel. It is observed that the FVMA method demonstrates significantly smaller MSE and MSFE than the FVMASA method. This outcome can be attributed to the

asymptotic optimality of the proposed weight estimator in Section 3, which indicates that its local squared loss converges asymptotically to that of the best possible model averaging estimator. In comparison to the other methods, the results consistently indicate that the FVMA approach outperforms. It exhibits the smallest MSE in nearly all examples and the smallest MSFE in every case. This is within our expectation since that the proposed model averaging methodology is designed for selecting optimal weights in out-of-sample forecasts.

5 An Empirical Example

This section is devoted to an empirical application of the proposed method to illustrate its practical usefulness. The dataset used in this analysis includes monthly observations of the S&P 500 stock price index, Federal funds rate, industrial production (IP), and the US government budget deficit (or surplus). The data range from October 1980, which is the first available observation of the Federal deficit on the FRED database, to December 2020. The following variables are utilized based on the transformation of the original data: stock return (SR_t), the growth in industrial production ($\mathrm{IPG}_t = \ln(\mathrm{IP}_t) - \ln(\mathrm{IP}_{t-1})$), the first differences of the effective federal fund rate ($\mathrm{DFF}_t = \mathrm{FF}_t - \mathrm{FF}_{t-1}$) and the change in fiscal deficits ($\mathrm{CFD}_t = \mathrm{FD}_t - \mathrm{FD}_{t-1}$). Deficits are denoted as positive values of FD, while surpluses are represented as negative values.

In line with much of the existing literature, we regard federal deficits as an indicator of constraints on monetary policy actions. Federal deficits can limit the ability of the government to implement monetary policy measures, such as interest rate adjustments, to stabilize the economy. By employing the framework proposed by Jansen et al. (2008), we obtain the candidate model

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

where $\mathbf{X}_t^{(m)} = (X_{t1}, \cdots, X_{tq_m})'$ is a $q_m \times 1$ vector of regressors. We construct the candidate pool Ω using the variables $\{SR_t, IPG_t, DFF_t\}$ and their respective lags. The inclusion

Table 1: Simulation Results of Examples 1-2.

Example 1	h=1		h=2		h=4	
p=10, T=200	MSE	MSFE	MSE	MSFE	MSE	MSFE
FVMA	0.0992	0.3482	0.1006	0.2885	0.1012	0.4068
FVMASA	4.2854	5.2842	4.2594	4.7695	4.2814	5.1670
AICc	0.1240	0.3846	0.1246	0.3335	0.1248	0.5028
SAICc	1.1547	1.7474	1.1517	1.5681	1.1570	1.7285
SAIC	5.7805	7.0876	5.4713	6.8454	5.7341	6.6753
SBIC	5.7916	7.1159	5.7552	6.9497	5.7457	6.7322
JMA	5.7910	7.1065	5.7514	6.8713	5.7442	6.6996
p=10, T=600						
FVMA	0.0422	0.1469	0.0421	0.2035	0.0415	0.1792
FVMASA	4.5039	5.1484	4.4941	4.8370	4.5035	5.4351
AICc	0.0544	0.1570	0.0542	0.2891	0.0535	0.1905
SAICc	0.9785	1.2254	0.9778	1.2744	0.9741	1.3128
SAIC	6.3233	6.7646	6.3409	6.6130	6.3607	7.1993
SBIC	6.3233	6.7646	6.3409	6.6130	6.3607	7.1993
JMA	6.3244	6.7813	6.3420	6.6124	6.3618	7.1992
Example 2	h=1		h=2		h=4	
p=10, T=200	MSE	MSFE	MSE	MSFE	MSE	MSFE
FVMA	0.0547	0.2045	0.0549	0.1869	0.0592	0.1854
FVMASA	4.0519	4.8070	4.0545	4.5621	4.0531	4.7398
AICc	0.0648	0.4359	0.0648	0.2022	0.0648	0.1917
SAICc	0.9697	1.4180	0.9700	1.2952	0.9724	1.3452
SAIC	0.8406	1.1613	0.8371	1.1156	0.8376	1.0962
SBIC	0.8406	1.1613	0.8371	1.1156	0.8376	1.0962
JMA	0.8409	1.1650	0.8373	1.1167	0.8378	1.0991
p=10, T=600						
FVMA	0.0257	0.1226	0.0258	0.1334	0.0259	0.1196
FVMASA	4.2499	4.3865	4.2430	4.6761	4.2606	4.3674
AICc	0.0311	0.1334	0.0312	0.1411	0.0313	0.1291
SAICc	0.8887	0.9911	0.8873	1.0920	0.8886	1.0012
SAIC	0.9122	1.0664	0.9124	1.1783	0.9114	1.1308
SBIC	0.9122	1.0664	0.9124	1.1783	0.9114	1.1308
JMA	0.9122	1.0659	0.9124	1.1786	0.9114	1.1311

Table 2: Simulation Results of Examples 3-4.

Example 3	h=1		h=2		h=4	
p=8, T=200	MSE	MSFE	MSE	MSFE	MSE	MSFE
FVMA	0.0578	0.2517	0.0579	0.1874	0.0582	0.1959
FVMASA	0.1541	0.3302	0.1530	0.2955	0.1523	0.3023
AICc	0.0546	0.2810	0.0545	0.2053	0.0544	0.2127
SAICc	0.1164	0.2950	0.1159	0.2620	0.1149	0.2600
SAIC	0.7351	0.8789	0.7292	0.8006	0.7321	0.8447
SBIC	0.7542	0.8673	0.7476	0.8041	0.7508	0.8326
JMA	0.7395	0.8712	0.7335	0.7968	0.7366	0.8326
p=8, T=600						
FVMA	0.0351	0.1260	0.0355	0.1399	0.0356	0.1437
FVMASA	0.1517	0.2668	0.1513	0.2896	0.1517	0.2788
AICc	0.0339	0.1278	0.0340	0.1897	0.0340	0.1643
SAICc	0.1037	0.2123	0.1035	0.2368	0.1034	0.2253
SAIC	0.7888	0.8919	0.7857	0.9231	0.7915	0.9044
SBIC	0.7974	0.8953	0.7940	0.9266	0.8000	0.9066
JMA	0.7904	0.8918	0.7872	0.9200	0.7931	0.9044
Example 4	h=1		h=2		h=4	
T=200	MSE	MSFE	MSE	MSFE	MSE	MSFE
FVMA	0.2503	0.4421	0.2502	0.4889	0.2512	0.4694
FVMASA	0.3900	0.6042	0.3907	0.6671	0.3904	0.6204
AICc	0.3490	0.6150	0.3490	0.6233	0.3493	0.5982
SAICc	0.3421	0.5607	0.3426	0.6135	0.3423	0.5770
SAIC	2.2009	2.2984	2.1926	2.5668	2.1903	2.3763
SBIC	2.3115	2.3015	2.3039	2.5265	2.2992	2.3291
JMA	2.2081	2.2800	2.1990	2.5462	2.1968	2.3562
T=600						
FVMA	0.1384	0.2914	0.1382	0.2876	0.2512	0.4694
FVMASA	0.2426	0.4267	0.2415	0.9009	0.3904	0.6204
AICc	0.1992	0.3680	0.1988	0.3644	0.3493	0.5982
SAICc	0.1986	0.3806	0.1977	1.0072	0.3423	0.5770
SAIC	2.3181	2.6062	2.3067	2.6924	2.1903	2.3763
SBIC	2.3830	2.6497	2.3711	2.7310	2.2992	2.3291
JMA	2.3140	2.6087	2.3027	2.6926	2.1968	2.3562

of lagged variables allows us to incorporate feedback over time and capture the dynamic relationships. Specifically, we consider

$$\Omega = \{\mathrm{SR}_t, \cdots, \mathrm{SR}_{t-k_1}, \mathrm{IPG}_t, \cdots, \mathrm{IPG}_{t-k_1}, \mathrm{DFF}_t, \cdots, \mathrm{DFF}_{t-k_1}\}$$

and $X_{ti} \in \Omega$ and $k_1 = 5$. Note that our methodology does not require the candidate model to be parsimonious. While a larger lag order k_1 could be considered to capture more complex dynamics, this would result in an increase in computational load. Additionally, it is worth noting that when $h = 1$, setting the lag order k_1 to 5 is consistent with the maximum lag order used in previous studies such as Jansen et al. (2008) and Tu & Wang (2020). The bandwidth is selected via $l = 2.34 S_{\mathrm{CFD}} T^{-1/5}$, where S_{CFD} is the sample standard deviation of $\{\mathrm{CFD}_t\}$. The number of candidate models is determined by the rule $M = min(\lfloor 3T^{1/3} \rceil, q)$, where q is the number predicting variables in Ω.

In our analysis, we evaluate the forecasting performance of the methods used in simulation section, and we rely on plots of relative MSFE to illustrate our findings. These plots provide a visual representation of the forecasting performance of the different methods under consideration. Specifically, let T_1 denote the start forecast date and vary it from 2019:6 until 2020:1, the MSFE for each method i is computed via

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

Next, we evaluate the relative percentage gains in mean squared forecast errors of the forecasts produced by the first six methods compared to the forecast produced by the JMA method. A negative value suggests that the corresponding method produces more a accurate forecast than the JMA method. This allows us to assess the relative performance of these methods in terms of their ability to reduce forecast errors.

As shown in Figure 1, the performance of the SAIC and SBIC methods are similar for $h = 1$ and $h = 2$, but both are inferior to the JMA method. Figure 1 reveals that the FVMA method strongly outperforms the FVMASA method over the forecasting period. This suggests that the proposed time-varying weighting scheme $\widehat{\mathbf{w}}_{u_0}$ is not equivalent to a simple

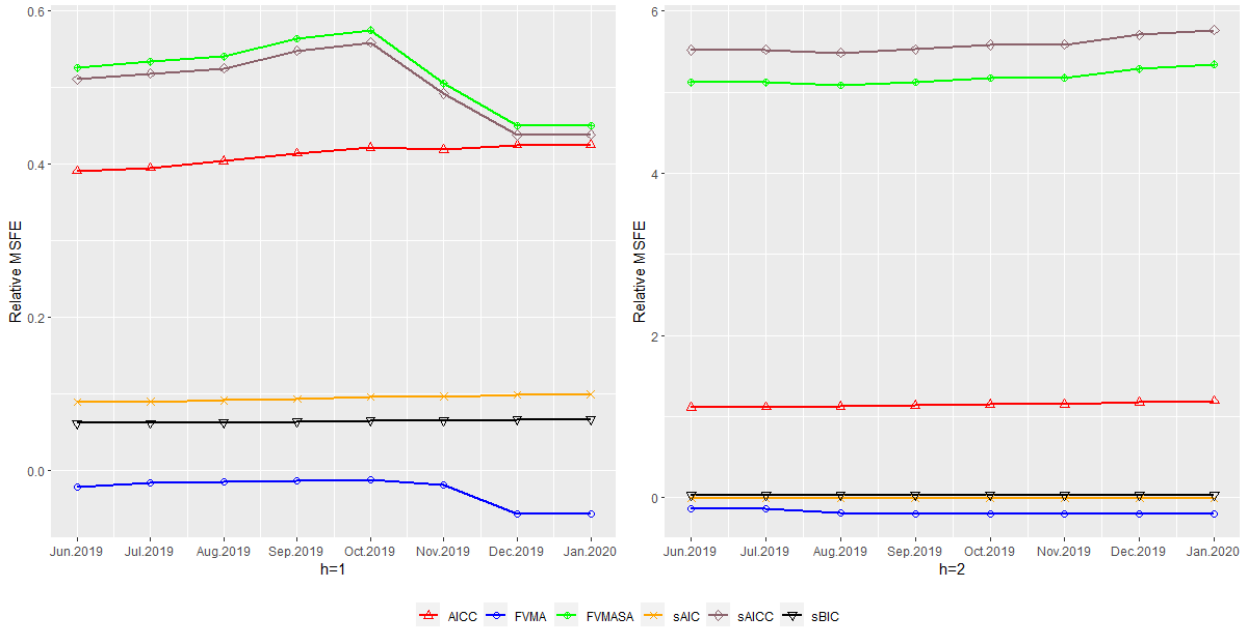

Figure 1: MSFE plots: the left panel for $h = 1$ and the right panel for $h = 2$.

averaging weighting scheme. Furthermore, the results indicate that the proposed method delivers out-of-sample forecasts that are no worse than existing model-average methods.

6 Conclusion

In this article, we have proposed a novel model averaging method for high-dimensional functional-coefficient regression models, which allows the selected weights to change over state variables. We have established the asymptotic optimality of the proposed estimator and the rate of the selected varying weights converging to the optimal weight, even 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 context has been investigated. Numerical analysis and empirical application strongly favor the proposed model averaging in comparison with the existing conventional methods.

Some relevant issues deserve further research. First, this paper has reduced model uncertainty mainly caused by high-dimensional covariates. It would be interesting to study local optimal averaging for functional-coefficient models with various state variables. For example, we could follow the spirit of Cai et al. (2015b) to select state variables prior to model averaging. In addition, one extension is to generalize the proposed method to functional-coefficient models for nonstationary time series data, which covers more applications in economics and finance (Cai et al., 2009; Xiao, 2009).

References

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

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

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

Cai, Z. (2007). Trending time-varying coefficient time series models with serially correlated errors. *Journal of Econometrics* **136**, 163–188.

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

Cai, Z., Das, M., Xiong, H. & Wu, X. (2006). Functional coefficient instrumental variables models. *Journal of Econometrics* **133**, 207–241.

Cai, Z., Fan, J. & Li, R. (2000a). Efficient estimation and inferences for varying-coefficient models. *Journal of the American Statistical Association* **95**, 888–902.

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

Cai, Z., Fang, Y. & Xu, Q. (2022). Testing capital asset pricing models using functional-coefficient panel data models with cross-sectional dependence. *Journal of Econometrics* **227**, 114–133.

Cai, Z., Juhl, T. & Yang, B. (2015a). Functional index coefficient models with variable selection. *Journal of Econometrics* **189**, 272–284.

Cai, Z., Li, Q. & Park, J. Y. (2009). Functional-coefficient models for nonstationary time series data. *Journal of Econometrics* **148**, 101–113.

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

Cai, Z. & Tiwari, R. C. (2010). Application of a local linear autoregressive model to bod time series. *Envirometrics* **11**, 341–350.

Cai, Z. & Xu, X. (2008). Nonparametric quantile estimations for dynamic smooth coefficient models. *Journal of the American Statistical Association* **103**, 1595–1608.

Card, D. (2001). Estimating the return to schooling: Progress on some persistent econometric problems. *Econometrica* **69**, 1127–1160.

Chan, K.-S. (1993). Consistency and limiting distribution of the least squares estimator of a threshold autoregressive model. *Annals of statistics* , 520–533.

Chen, B. & Hong, Y. (2012). Testing for smooth structural changes in time series models via nonparametric regression. *Econometrica* **80**, 1157–1183.

Chen, J., Li, D., Linton, O. & Lu, Z. (2018). Semiparametric ultra-high dimensional model averaging of nonlinear dynamic time series. *Journal of American Statistical Association* **113**, 919–932.

Chen, R. & Tsay, R. S. (1993). Functional-coefficient autoregressive models. *Journal of the American Statistical Association* **88**, 298–308.

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

Claeskens, G. & Hjort, N. L. (2008). *Model Selection and Model Averaging*. Cambridge: Cambridge University Press.

Doukhan, P. (2012). *Mixing: Properties and Examples*, vol. 85. Springer Science & Business Media.

Fan, J. & Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Annals of Statistics* **32**, 928–961.

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

Gao, Y., Zhang, X., Wang, S. & Zou, G. (2016). Model averaging based on leave-subject-out cross validation. *Journal of Econometrics* **192**, 139–151.

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

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

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

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

Hong, Y. & Lee, T.-H. (2003). Inference on predictability of foreign exchange rates via generalized spectrum and nonlinear time series models. *Review of Economics and Statistics* **85**, 1048–1062.

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

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

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

Li, K.-C. (1987). Asymptotic optimality for C_p, C_l, cross-validation and generalized cross-validation: Discrete index set. *Annals of Statistics* **15**, 958–975.

Liao, J., Zong, X., Zhang, x. & Zou, G. (2019). Model averaging based on leave-subject-out cross-validation for vector autoregressions. *Journal of Econometrics* **209**, 35–60.

Phillips, P. C. & Wang, Y. (2022). Functional coefficient panel modeling with communal smoothing covariates. *Journal of Econometrics* **227**, 371–407.

Roussanov, N. (2014). Composition of wealth, conditioning information, and the cross-section of stock returns. *Journal of Financial Economics* **111**, 352–380.

Steel, M. F. (2020). Model averaging and its use in economics. *Journal of Economic Literature* **58**, 644–719.

Stock, J. H. & Watson, M. W. (2012). Disentangling the channels of the 2007-2009 recession. Tech. rep., National Bureau of Economic Research.

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

Sun, Y., Hong, Y., Wang, S. & Zhang, X. (2022). Penalized time-varying model averaging. *Journal of Econometrics, Forthcoming* .

Teräsvirta, T. (1994). Specification, estimation, and evaluation of smooth transition autoregressive models. *Journal of American Statistical Association* **89**, 208–218.

Tong, H. (1978). *On A Threshold Model*. No. 29. Sijthoff & Noordhoff.

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

Tu, Y. & Wang, Y. (2022). Spurious functional-coefficient regression models and robust inference with marginal integration. *Journal of Econometrics* **229**, 396–421.

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

Yuan, Z. & Yang, Y. (2005). Combining linear regression models: When and how? *Journal of the American Statistical Association* **100**, 1202–1214.

Zhang, X. & Liu, C.-A. (2022). Model averaging prediction by k-fold cross-validation. *Journal of Econometrics* .

Zhang, X., Yu, D., Zou, G. & Liang, H. (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. & Zhang, X. (2022). Optimal model averaging based on forward-validation. *Journal of Econometrics* .

Zhao, S., Zhou, J. & Yang, G. (2019). Averaging estimators for discrete choice by M-fold cross-validation. *Economics Letters* **174**, 65–69.

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

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101**, 1418–1429.

Appendix

Before we embrace on providing the detailed proof to the main theorems, some lemmas are needed, presented as follows.

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

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

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. (2000b). □

Lemma 2. *Suppose Conditions (C.1)-(C.3) hold. Then, we have*

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

and

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

if q grows at some rate of T.

Proof. We have

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

and

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

Lemma 2 is valid if the following holds:

$$||\mathbf{X}'\mathbf{K}(u_0)\boldsymbol{\epsilon}|| = O_p(\sqrt{qTl}), \tag{A.1}$$

and

$$||\mathbf{X}'\mathbf{K}(u_0)\mathbf{Y}|| = O_p(\sqrt{q}Tl). \tag{A.2}$$

First, with Conditions (C.1)-(C.3), we have

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

for some positive C and non-stochastic $\mathbf{X}$. We have similar results for random $\mathbf{X}$. Thus, (A.1) holds.

Next, with (A.1) and Conditions (C.1)-(C.3), we have

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

for some positive constant C. Thus, the proof of (A.2) is completed. □

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

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

Proof. With Conditions (C.1)-(C.5), we have

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

$$
\begin{aligned}
&= \Pr\left[\left|\left|\frac{T^{1/2}l^{1/2}}{q^{1/2}}\left(\mathbf{X}^{(m)'}\mathbf{K}(u_0)\mathbf{X}^{(m)}\right)^{-1}\mathbf{X}^{(m)'}\mathbf{K}(u_0)(\mathbf{Y}-\boldsymbol{\mu})\right|\right| \le \delta\right] \\
&\ge \Pr\left(C_0^{-1}\left|\left|\frac{T^{-1/2}l^{-1/2}}{q^{1/2}}\mathbf{X}^{(m)'}\mathbf{K}(u_0)(\mathbf{Y}-\boldsymbol{\mu})\right|\right| \le \delta\right) \\
&\ge 1-\frac{\mathrm{var}(\mathbf{X}^{(m)'}\mathbf{K}(u_0)\boldsymbol{\epsilon})}{C_0^2\delta^2 Tlq} \ge 1-\frac{C_1}{C_0^2\delta^2}
\end{aligned}
$$

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

Lemma 4. *Suppose that Conditions (C.1)-(C.2) hold. We have*

$$FV_T(u_0,\mathbf{w}) = \mathbf{Y}'(\mathbf{A}(\mathbf{w},\mathbf{X})+\mathbf{Q}(\mathbf{w},\mathbf{X}))'\mathbf{K}(u_0)(\mathbf{A}(\mathbf{w},\mathbf{X})+\mathbf{Q}(\mathbf{w},\mathbf{X}))\mathbf{Y},$$

where $\mathbf{A}(\mathbf{w},\mathbf{X}) \equiv \sum_{m=1}^M w^m\mathbf{A}^{(m)}(\mathbf{X})$, $\mathbf{Q}(\mathbf{w},\mathbf{X}) \equiv \sum_{m=1}^M w^m\mathbf{Q}^{(m)}(\mathbf{X})$, $\mathbf{A}^{(m)}(\mathbf{X}) = \mathbf{I} - \mathbf{P}^{(m)}(\mathbf{X})$, $\mathbf{Q}^{(m)}(\mathbf{X}) \equiv \boldsymbol{\pi}\mathbf{D}^{(m)}\boldsymbol{\phi}\mathbf{A}^{(m)}(\mathbf{X})$, $\boldsymbol{\phi} = (\boldsymbol{\phi}_1',\cdots,\boldsymbol{\phi}_T')'$, $(\mathbf{I}-\mathbf{P}_{tt}^{(m)})^{-1} = \mathbf{I}+\sum_{j=1}^\infty(\mathbf{P}_{tt}^{(m)})^j \equiv \mathbf{I}+\mathbf{D}_t^{(m)}$, *and* $\boldsymbol{\pi}$ *and* $\mathbf{D}^{(m)}$ *are block diagonal matrices with the t-th diagonal block being* $\boldsymbol{\pi}_t$ *and* $\mathbf{D}_t^{(m)}$ *(*$1 \le t \le T$*).*

Proof. It can be shown easily that

$$
\begin{aligned}
\widetilde{\mu}_t^{(m)} &= \boldsymbol{\pi}_t\left[\mathbf{Y}_{[t+h]} - (\mathbf{I}-\mathbf{P}_{tt}^{(m)})^{-1}(\mathbf{Y}_{[t+h]} - \widehat{\boldsymbol{\mu}}_{[t]}^{(m)})\right] \\
&= \boldsymbol{\pi}_t\boldsymbol{\phi}_t\mathbf{P}^{(m)}(\mathbf{X})\mathbf{Y} - \boldsymbol{\pi}_t\mathbf{D}_t^{(m)}\boldsymbol{\phi}_t\mathbf{A}^{(m)}(\mathbf{X})\mathbf{Y},
\end{aligned}
$$

where $\widehat{\boldsymbol{\mu}}_{[t]}^{(m)} \equiv \boldsymbol{\phi}_t\mathbf{P}^{(m)}(\mathbf{X})\mathbf{Y}$ and $\mathbf{P}_{tt}^{(m)} \equiv \boldsymbol{\phi}_t\mathbf{P}^{(m)}(\mathbf{X})\boldsymbol{\phi}_t'$. Then, we have that

$$\widetilde{\boldsymbol{\mu}}^{(m)} = \mathbf{P}^{(m)}(\mathbf{X})\mathbf{Y} - \mathbf{Q}^{(m)}(\mathbf{X})\mathbf{Y},$$

and

$$\mathbf{Y} - \widetilde{\boldsymbol{\mu}}(\mathbf{w}) = \left(\mathbf{I} - \sum_{m=1}^M w^m\mathbf{P}^{(m)}(\mathbf{X}) + \sum_{m=1}^M w^m\mathbf{Q}^{(m)}(\mathbf{X})\right)\mathbf{Y} = (\mathbf{A}(\mathbf{w},\mathbf{X})+\mathbf{Q}(\mathbf{w},\mathbf{X}))\mathbf{Y},$$

which implies that Lemma 4 is obtained. □

Lemma 5. *Suppose Conditions (C.8)-(C.9) hold. Then,*

$$\zeta_{\max}(\mathbf{Q}^{(m)}(\mathbf{X})) = O_p(M_T q^{1/2}T^{-1}l^{-1}).$$

Proof. With Condition (C.8), we have $\mathrm{tr}(\mathbf{P}_{tt}^{(m)}) = \mathrm{tr}(\boldsymbol{\phi}_t \mathbf{P}^{(m)}(\mathbf{X})\boldsymbol{\phi}_t') \leq (h+1)\max_{1\leq t\leq T}\mathbf{P}_t^{(m)} = O_p(M_T T^{-1} l^{-1})$ uniformly in m. Then, it follows that

$$\begin{aligned}
\zeta_{\max}(\mathbf{D}^{(m)}) &\leq \max_{1\leq t\leq T}\mathrm{tr}(\mathbf{D}_t^{(m)}) = \max_{1\leq t\leq T}\mathrm{tr}\{\sum_{j=1}^{\infty}(\mathbf{P}_{tt}^{(m)})^j\} \\
&\leq \max_{1\leq t\leq T}\sum_{j=1}^{\infty}\{\zeta_{\max}[\mathbf{P}_{tt}^{(m)}]^{j-1}\mathrm{tr}[\mathbf{P}_{tt}^{(m)}]\} \\
&\leq \max_{1\leq t\leq T}\sum_{j=1}^{\infty}\mathrm{tr}[\mathbf{P}_{tt}^{(m)}]^j \\
&= \max_{1\leq t\leq T}\mathrm{tr}\mathbf{P}_{tt}^{(s)}/(1-\mathrm{tr}\mathbf{P}_{tt}^{(s)}) = O_p(M_T T^{-1} l^{-1}),
\end{aligned}$$

which implies that $\zeta_{\max}(\boldsymbol{\pi}) = 1$, so that

$$\begin{aligned}
\zeta_{\max}(\mathbf{Q}^{(m)}(\mathbf{X})) &= \zeta_{\max}\{\boldsymbol{\pi}\mathbf{D}^{(m)}\boldsymbol{\phi}\mathbf{A}^{(m)}(\mathbf{X})\} \\
&\leq \zeta_{\max}(\boldsymbol{\pi})\zeta_{\max}(\mathbf{D}^{(m)})\zeta_{\max}(\boldsymbol{\phi})\zeta_{\max}(\mathbf{A}^{(m)}(\mathbf{X})) \\
&= O_p(M_T q^{1/2} T^{-1} l^{-1})
\end{aligned}$$

with Condition (C.9). This proves the lemma. □

Proof of Theorem 1. For any given u_0, we first do the following decomposition

$$\begin{aligned}
L_T(u_0,\mathbf{w}) &= (\boldsymbol{\mu}^*(\mathbf{w}) - \boldsymbol{\mu} + \widehat{\boldsymbol{\mu}}(\mathbf{w}) - \boldsymbol{\mu}^*(\mathbf{w}))'\mathbf{K}(u_0)(\boldsymbol{\mu}^*(\mathbf{w}) - \boldsymbol{\mu} + \widehat{\boldsymbol{\mu}}(\mathbf{w}) - \boldsymbol{\mu}^*(\mathbf{w})) \\
&= L_T^*(u_0,\mathbf{w}) + 2(\widehat{\boldsymbol{\mu}}(\mathbf{w}) - \boldsymbol{\mu}^*(\mathbf{w}))'\mathbf{K}(u_0)(\boldsymbol{\mu}^*(\mathbf{w}) - \boldsymbol{\mu}) \\
&\quad +(\widehat{\boldsymbol{\mu}}(\mathbf{w}) - \boldsymbol{\mu}^*(\mathbf{w}))'\mathbf{K}(u_0)(\widehat{\boldsymbol{\mu}}(\mathbf{w}) - \boldsymbol{\mu}^*(\mathbf{w})) \\
&\equiv L_T^*(u_0,\mathbf{w}) + \Gamma_T(u_0,\mathbf{w}) \\
&= \mathbb{E}L_T^*(u_0,\mathbf{w}) + \Gamma_T(u_0,\mathbf{w}) + (L_T^*(u_0,\mathbf{w}) - \mathbb{E}L_T^*(u_0,\mathbf{w})),
\end{aligned}$$

and

$$\begin{aligned}
&|\mathrm{FV}_T(u_0,\mathbf{w}) - L_T(u_0,\mathbf{w})| \\
= &\left|[\mathbf{Y} - \widetilde{\boldsymbol{\mu}}(\mathbf{w})]'\,\mathbf{K}(u_0)\,[\mathbf{Y} - \widetilde{\boldsymbol{\mu}}(\mathbf{w})] - (\widehat{\boldsymbol{\mu}}(\mathbf{w}) - \boldsymbol{\mu})'\mathbf{K}(u_0)(\widehat{\boldsymbol{\mu}}(\mathbf{w}) - \boldsymbol{\mu})\right| \\
= &\left|[\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}(\mathbf{w}) - (\widetilde{\boldsymbol{\mu}}(\mathbf{w}) - \boldsymbol{\mu}^*(\mathbf{w})) + (\widehat{\boldsymbol{\mu}}(\mathbf{w}) - \boldsymbol{\mu}^*(\mathbf{w})) + (\mathbf{Y} - \boldsymbol{\mu})]'\,\mathbf{K}(u_0)\right.
\end{aligned}$$

$$
\begin{aligned}
&\times [\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}(\mathbf{w}) - (\widetilde{\boldsymbol{\mu}}(\mathbf{w}) - \boldsymbol{\mu}^*(\mathbf{w})) + (\widehat{\boldsymbol{\mu}}(\mathbf{w}) - \boldsymbol{\mu}^*(\mathbf{w})) + (\mathbf{Y} - \boldsymbol{\mu})] \\
&- (\widehat{\boldsymbol{\mu}}(\mathbf{w}) - \boldsymbol{\mu})' \mathbf{K}(u_0)(\widehat{\boldsymbol{\mu}}(\mathbf{w}) - \boldsymbol{\mu})| \\
\leq\ & (\widetilde{\boldsymbol{\mu}}(\mathbf{w}) - \boldsymbol{\mu}^*(\mathbf{w}))' \mathbf{K}(u_0)(\widetilde{\boldsymbol{\mu}}(\mathbf{w}) - \boldsymbol{\mu}^*(\mathbf{w})) + |(\widehat{\boldsymbol{\mu}}(\mathbf{w}) - \boldsymbol{\mu}^*(\mathbf{w}))' \mathbf{K}(u_0)(\widehat{\boldsymbol{\mu}}(\mathbf{w}) - \boldsymbol{\mu}^*(\mathbf{w}))| \\
&+2|(\widehat{\boldsymbol{\mu}}(\mathbf{w}) - \boldsymbol{\mu}^*(\mathbf{w}))' \mathbf{K}(u_0)(\widetilde{\boldsymbol{\mu}}(\mathbf{w}) - \boldsymbol{\mu}^*(\mathbf{w}))| + 2|(\boldsymbol{\mu}^*(\mathbf{w}) - \boldsymbol{\mu})' \mathbf{K}(u_0)(\widetilde{\boldsymbol{\mu}}(\mathbf{w}) - \boldsymbol{\mu}^*(\mathbf{w}))| \\
&+2|(\widehat{\boldsymbol{\mu}}(\mathbf{w}) - \boldsymbol{\mu}^*(\mathbf{w}))' \mathbf{K}(u_0)(\widehat{\boldsymbol{\mu}}(\mathbf{w}) - \boldsymbol{\mu}^*(\mathbf{w}))| + 2|(\boldsymbol{\mu}^*(\mathbf{w}) - \boldsymbol{\mu})' \mathbf{K}(u_0)(\widehat{\boldsymbol{\mu}}(\mathbf{w}) - \boldsymbol{\mu}^*(\mathbf{w}))| \\
&+2|(\widehat{\boldsymbol{\mu}}(\mathbf{w}) - \boldsymbol{\mu}^*(\mathbf{w}))' \mathbf{K}(u_0)(\boldsymbol{\mu} - \mathbf{Y})| + 2|(\boldsymbol{\mu}^*(\mathbf{w}) - \boldsymbol{\mu}(\mathbf{w}))' \mathbf{K}(u_0)(\boldsymbol{\mu} - \mathbf{Y})| \\
&+2|(\widetilde{\boldsymbol{\mu}}(\mathbf{w}) - \boldsymbol{\mu}^*(\mathbf{w}))' \mathbf{K}(u_0)(\widehat{\boldsymbol{\mu}}(\mathbf{w}) - \boldsymbol{\mu}^*(\mathbf{w}))| + 2|(\widetilde{\boldsymbol{\mu}} - \boldsymbol{\mu}^*)' \mathbf{K}(u_0)(\boldsymbol{\mu} - \mathbf{Y})| \\
&+2|(\widehat{\boldsymbol{\mu}}(\mathbf{w}) - \boldsymbol{\mu}^*(\mathbf{w}))' \mathbf{K}(u_0)(\boldsymbol{\mu} - \mathbf{Y})| + |(\boldsymbol{\mu} - \mathbf{Y})' \mathbf{K}(u_0)(\boldsymbol{\mu} - \mathbf{Y})| \\
\equiv\ & \Lambda_T(u_0, \mathbf{w}) + |(\boldsymbol{\mu} - \mathbf{Y})' \mathbf{K}(u_0)(\boldsymbol{\mu} - \mathbf{Y})|,
\end{aligned}
$$

where the second term is unrelated to $\mathbf{w}$. From Theorem 1 of Zhao et al. (2019), Theorem 1 is valid if

$$
\sup_{\mathbf{w} \in \mathcal{H}_T} \frac{|\Gamma_T(u_0, \mathbf{w})|}{\mathbb{E} L_T^*(u_0, \mathbf{w})} = o_p(1), \tag{A.3}
$$

$$
\sup_{\mathbf{w} \in \mathcal{H}_T} \frac{|\Lambda_T(u_0, \mathbf{w})|}{\mathbb{E} L_T^*(u_0, \mathbf{w})} = o_p(1), \tag{A.4}
$$

and

$$
\sup_{\mathbf{w} \in \mathcal{H}_T} \frac{|L_T^*(u_0, \mathbf{w}) - \mathbb{E} L_T^*(u_0, \mathbf{w})|}{\mathbb{E} L_T^*(u_0, \mathbf{w})} = o_p(1). \tag{A.5}
$$

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

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

Then, we have

$$
\begin{aligned}
\widehat{\boldsymbol{\alpha}}(u_0, \mathbf{w}) - \boldsymbol{\alpha}^*(u_0, \mathbf{w}) &= \sum_{m=1}^{M_T} w^m \left[\widehat{\boldsymbol{\alpha}}^{(m)}(u_0) - \boldsymbol{\alpha}^{(m)*}(u_0) \right] \\
&= O_p(M_T^{1/2} q^{1/2} T^{-1/2} l^{-1/2}).
\end{aligned} \tag{A.6}
$$

Thus,

$$
\sup_{\mathbf{w} \in \mathcal{H}_T} \left[\widehat{\boldsymbol{\mu}}(\mathbf{w}) - \boldsymbol{\mu}^*(\mathbf{w}) \right]' \mathbf{K}(u_0) \left[\widehat{\boldsymbol{\mu}}(\mathbf{w}) - \boldsymbol{\mu}^*(\mathbf{w}) \right]
$$

$$
\begin{aligned}
&= \sup_{\mathbf{w}\in\mathcal{H}_T} [\mathbf{X}'(\widehat{\boldsymbol{\alpha}}(u_0,\mathbf{w}) - \boldsymbol{\alpha}^*(u_0,\mathbf{w}))]' \mathbf{K}(u_0) [\mathbf{X}'(\widehat{\boldsymbol{\alpha}}(u_0,\mathbf{w}) - \boldsymbol{\alpha}^*(u_0,\mathbf{w}))] \\
&= O_p(Tlq) * O_p(M_T q T^{-1} l^{-1}) = O_p(M_T q^2) \qquad \text{(A.7)}
\end{aligned}
$$

with Conditions (C.2)-(C.4) and (A.6). Similarly, we have

$$
\sup_{\mathbf{w}\in\mathcal{H}_T} [\widetilde{\boldsymbol{\mu}}(\mathbf{w}) - \boldsymbol{\mu}^*(\mathbf{w})]' \mathbf{K}(u_0) [\widetilde{\boldsymbol{\mu}}(\mathbf{w}) - \boldsymbol{\mu}^*(\mathbf{w})] = O_p(M_T q^2). \qquad \text{(A.8)}
$$

Also, with Condition (C.3), it is shown that

$$
\sup_{\mathbf{w}\in\mathcal{H}_T} |(\boldsymbol{\mu}^*(\mathbf{w}) - \boldsymbol{\mu})' \mathbf{K}(u_0) (\widehat{\boldsymbol{\mu}}(\mathbf{w}) - \boldsymbol{\mu}^*(\mathbf{w}))| = O_p(M_T^{1/2} q T^{1/2} l^{1/2}), \qquad \text{(A.9)}
$$

$$
\sup_{\mathbf{w}\in\mathcal{H}_T} |(\boldsymbol{\mu}^*(\mathbf{w}) - \boldsymbol{\mu})' \mathbf{K}(u_0) (\widetilde{\boldsymbol{\mu}}(\mathbf{w}) - \boldsymbol{\mu}^*(\mathbf{w}))| = O_p(M_T^{1/2} q T^{1/2} l^{1/2}), \qquad \text{(A.10)}
$$

and

$$
(\boldsymbol{\mu} - \mathbf{Y})' \mathbf{K}(u_0) (\boldsymbol{\mu} - \mathbf{Y}) = \boldsymbol{\epsilon}' \mathbf{K}(u_0) \boldsymbol{\epsilon} = O_p(Tl). \qquad \text{(A.11)}
$$

From (A.7) and (A.9), we have $\sup_{\mathbf{w}\in\mathcal{H}_T} |\Gamma_T| = O_p(M_T^{1/2} q T^{1/2} l^{1/2})$ and thus, (A.3) is obtained.

By the same token, we have

$$
\begin{aligned}
\sup_{\mathbf{w}\in\mathcal{H}_T} |(\widehat{\boldsymbol{\mu}}(\mathbf{w}) - \boldsymbol{\mu}^*(\mathbf{w}))' \mathbf{K}(u_0) (\widetilde{\boldsymbol{\mu}}(\mathbf{w}) - \boldsymbol{\mu}^*(\mathbf{w}))| &= O_p(M_T q^2), \\
\sup_{\mathbf{w}\in\mathcal{H}_T} |(\boldsymbol{\mu}^*(\mathbf{w}) - \boldsymbol{\mu})' \mathbf{K}(u_0) (\boldsymbol{\mu} - \mathbf{Y})| &= O_p(Tl), \\
\sup_{\mathbf{w}\in\mathcal{H}_T} |(\widehat{\boldsymbol{\mu}}(\mathbf{w}) - \boldsymbol{\mu}^*(\mathbf{w}))' \mathbf{K}(u_0) (\boldsymbol{\mu} - \mathbf{Y})| &= O_p(M_T^{1/2} q T^{1/2} l^{1/2}),
\end{aligned}
$$

and

$$
\sup_{\mathbf{w}\in\mathcal{H}_T} |(\widetilde{\boldsymbol{\mu}}(\mathbf{w}) - \boldsymbol{\mu}^*(\mathbf{w}))' \mathbf{K}(u_0) (\boldsymbol{\mu} - \mathbf{Y})| = O_p(M_T^{1/2} q T^{1/2} l^{1/2}).
$$

Besides, we will verify that

$$
\sup_{\mathbf{w}\in\mathcal{H}_T} \frac{|(\boldsymbol{\mu}^*(\mathbf{w}) - \boldsymbol{\mu})' \mathbf{K}(u_0) (\boldsymbol{\mu} - \mathbf{Y})|}{\mathbb{E} L_T^*(u_0,\mathbf{w})} = o_p(1). \qquad \text{(A.12)}
$$

For any $\delta > 0$, we have

$$
\Pr\left\{ \sup_{\mathbf{w}\in\mathcal{H}_T} \xi_T(u_0)^{-1} |(\boldsymbol{\mu}^*(\mathbf{w}) - \boldsymbol{\mu})' \mathbf{K}(u_0) (\boldsymbol{\mu} - \mathbf{Y})| > \delta \right\}
$$

$$
\begin{aligned}
&\leq \Pr\left\{\sup_{\mathbf{w}\in\mathcal{H}_T}\xi_T(u_0)^{-1}\sum_{m=1}^{M}w^m|(\boldsymbol{\mu}^{(m)*}-\boldsymbol{\mu})'\mathbf{K}(u_0)(\boldsymbol{\mu}-\mathbf{Y})|>\delta\right\}\\
&= \Pr\left\{\max_{m}|(\boldsymbol{\mu}^{(m)*}-\boldsymbol{\mu})'\mathbf{K}(u_0)(\boldsymbol{\mu}-\mathbf{Y})|>\xi_T(u_0)\delta\right\}\\
&\leq \sum_{m=1}^{M_T}\Pr\{|(\boldsymbol{\mu}^{(m)*}-\boldsymbol{\mu})'\mathbf{K}(u_0)(\boldsymbol{\mu}-\mathbf{Y})|>\xi_T(u_0)\delta\}\\
&\leq \xi_T^{-2}(u_0)\delta^{-2}\sum_{m=1}^{M_T}\sum_{s=1}^{T}\mathbb{E}\{\mathbb{E}[((\mu_s^{(m)*}-\mu_s)'(\mu_s-Y_{s+h})k_s)^2|I_s]\}\\
&\leq \xi_T^{-2}(u_0)\delta^{-2}\sum_{m=1}^{M_T}\sum_{s=1}^{T}k_s^2\mathbb{E}\{\mathbb{E}(\epsilon_{s+h}^2|I_s)(\mu_s^{(m)*}-\mu_s)^2)]\}\\
&= O(q^2M_TTl\xi_T^{-2}(u_0))=o(1),
\end{aligned}
$$

where the last step of the above is obtained from Conditions (C.3), (C.4) and (C.6). Thus, (A.12) is completed. Thus, with (A.7)-(A.12) and Condition (C.6), we have $\sup_{\mathbf{w}\in\mathcal{H}_T}|\Lambda_T| = O_p(qM_T^{1/2}T^{1/2}l^{1/2})$, so that (A.4) is derived.

Finally, with Conditions (C.1)-(C.2), for any $\delta>0$ and uniformly for any u_0, we have

$$\Pr\{\sum_{\mathbf{w}\in\mathcal{H}_T}\xi_T^{-1}(u_0)|L_T^*(u_0,\mathbf{w})-\mathbb{E}L_T^*(u_0,\mathbf{w})|>\delta\}=o(1)$$

based on the law of large numbers of the mixing-process as in Doukhan (2012), and so (A.5) is obtained. Therefore, the proof of Theorem 1 is completed. □

Proof of Theorem 2. Following Fan & Peng (2004) and Chen et al. (2018), we need only to verify that for any given u_0, there is a constant c_0 such that

$$\lim_{T\to\infty}\Pr\left(\inf_{||\mathbf{v}||=c_0,(\mathbf{w}^0(u_0)+\eta_T(u_0)\mathbf{v})\in\mathcal{H}_T}\mathrm{FV}_T(u_0,\mathbf{w}^0(u_0)+\eta_T(u_0)\mathbf{v})>\mathrm{FV}_T(u_0,\mathbf{w}^0(u_0))\right)=1$$

with $\mathbf{v}=(v^1,\cdots,v^{M_T})'$ and $\eta_T(u_0)\equiv M_TqT^{-1/2+\delta}l^{-1/2+\delta}$, which implies that for any given u_0, $||\widehat{\mathbf{w}}_{u_0}-\mathbf{w}^0(u_0)||=O_p(\eta_T(u_0))$ with a minimum $\widehat{\mathbf{w}}_{u_0}$ in the set $\{\mathbf{w}^0(u_0)+\eta_T(u_0)\mathbf{v}:||\mathbf{v}||\leq c_0,\mathbf{w}^0(u_0)+\eta_T(u_0)\mathbf{v}\in\mathcal{H}_T\}$.

First, we decompose $\mathrm{FV}_T(u_0,\mathbf{w}^0(u_0)+\eta_T(u_0)\mathbf{v})-\mathrm{FV}_T(u_0,\mathbf{w}^0(u_0))$ in the following four parts:

$$\mathrm{FV}_T(u_0,\mathbf{w}^0(u_0)+\eta_T(u_0)\mathbf{v})-\mathrm{FV}_T(u_0,\mathbf{w}^0(u_0))$$

$$
\begin{aligned}
&= \mathbf{Y}'\left[\mathbf{A}(\mathbf{w}^0(u_0)+\eta_T(u_0)\mathbf{v},\mathbf{X})+\mathbf{Q}(\mathbf{w}^0(u_0)+\eta_T(u_0)\mathbf{v},\mathbf{X})\right]'\mathbf{K}(u_0)\left[\mathbf{A}(\mathbf{w}^0(u_0)+\eta_T(u_0)\mathbf{v},\mathbf{X})\right.\\
&\quad \left.+\mathbf{Q}(\mathbf{w}^0(u_0)+\eta_T(u_0)\mathbf{v},\mathbf{X})\right]\mathbf{Y}-\mathbf{Y}'\left[\mathbf{A}(\mathbf{w}^0(u_0),\mathbf{X})+\mathbf{Q}(\mathbf{w}^0(u_0),\mathbf{X})\right]'\mathbf{K}(u_0)\\
&\quad \times\left[\mathbf{A}(\mathbf{w}^0(u_0),\mathbf{X})+\mathbf{Q}(\mathbf{w}^0(u_0),\mathbf{X})\right]\mathbf{Y}\\
&= \mathbf{Y}'\mathbf{A}'(\mathbf{w}^0(u_0)+\eta_T(u_0)\mathbf{v},\mathbf{X})\mathbf{K}(u_0)\mathbf{A}(\mathbf{w}^0(u_0)+\eta_T(u_0)\mathbf{v},\mathbf{X})\mathbf{Y}+\mathbf{Y}'\mathbf{M}(\mathbf{w}^0(u_0)+\eta_T(u_0)\mathbf{v},\mathbf{X})\mathbf{Y}\\
&\quad -\mathbf{Y}'\mathbf{A}'(\mathbf{w}^0(u_0),\mathbf{X})\mathbf{K}(u_0)\mathbf{A}(\mathbf{w}^0(u_0),\mathbf{X})\mathbf{Y}-\mathbf{Y}'\mathbf{M}(\mathbf{w}^0(u_0),\mathbf{X})\mathbf{Y}\\
&= \widehat{\boldsymbol{\mu}}(\eta_T(u_0)\mathbf{v})'\mathbf{K}(u_0)\widehat{\boldsymbol{\mu}}(\eta_T(u_0)\mathbf{v})-2(\boldsymbol{\mu}-\widehat{\boldsymbol{\mu}}(\mathbf{w}^0(u_0)))'\mathbf{K}(u_0)\widehat{\boldsymbol{\mu}}(\eta_T(u_0)\mathbf{v})-2\boldsymbol{\varepsilon}'\mathbf{K}(u_0)\widehat{\boldsymbol{\mu}}(\eta_T(u_0)\mathbf{v})\\
&\quad +\left[\mathbf{Y}'\mathbf{M}(\mathbf{w}^0(u_0)+\eta_T(u_0)\mathbf{v},\mathbf{X})\mathbf{Y}-\mathbf{Y}'\mathbf{M}(\mathbf{w}^0(u_0),\mathbf{X})\mathbf{Y}\right]\\
&\equiv \Xi_1-2\Xi_2-2\Xi_3+\Xi_4,
\end{aligned}
$$

where

$$
\begin{aligned}
\mathbf{M}(\mathbf{w},\mathbf{X}) \equiv\ & \mathbf{Q}'(\mathbf{w},\mathbf{X})\mathbf{K}(u_0)+\mathbf{K}(u_0)\mathbf{Q}(\mathbf{w},\mathbf{X})-\mathbf{P}'(\mathbf{w},\mathbf{X})\mathbf{K}(u_0)\mathbf{Q}(\mathbf{w},\mathbf{X})\\
&-\mathbf{Q}'(\mathbf{w},\mathbf{X})\mathbf{K}(u_0)\mathbf{P}(\mathbf{w},\mathbf{X})+\mathbf{Q}'(\mathbf{w},\mathbf{X})\mathbf{K}(u_0)\mathbf{Q}(\mathbf{w},\mathbf{X}),
\end{aligned}
$$

$\Xi_1 \equiv \widehat{\boldsymbol{\mu}}(\eta_T(u_0)\mathbf{v})'\mathbf{K}(u_0)\widehat{\boldsymbol{\mu}}(\eta_T(u_0)\mathbf{v})$, $\Xi_2 \equiv (\boldsymbol{\mu}-\widehat{\boldsymbol{\mu}}(\mathbf{w}^0(u_0)))'\mathbf{K}(u_0)\widehat{\boldsymbol{\mu}}(\eta_T(u_0)\mathbf{v})$, $\Xi_3 \equiv \boldsymbol{\varepsilon}'\mathbf{K}(u_0)\widehat{\boldsymbol{\mu}}(\eta_T(u_0)\mathbf{v})$, and $\Xi_4 \equiv \mathbf{Y}'\mathbf{M}(\mathbf{w}^0(u_0)+\eta_T(u_0)\mathbf{v},\mathbf{X})\mathbf{Y}-\mathbf{Y}'\mathbf{M}(\mathbf{w}^0(u_0),\mathbf{X})\mathbf{Y}$.

To verify Theorem 2, it is equivalent to showing that $\Xi_1 > 0$ in probability converges to 1, and Ξ_1 asymptotically dominates $\{\Xi_2,\Xi_3,\Xi_4\}$, respectively. Based on Conditions (C.4)-(C.5) and (C.7), it is derived that

$$
\Xi_1=\sum_{s=1}^{T}(\sum_{m=1}^{M_T}\eta_T(u_0)v^m\widehat{\mu}_s^{(m)})^2k_s=\sum_{s=1}^{T}\eta_T^2(u_0)(\sum_{m=1}^{M_T}v^m\widehat{\mu}_s^{(m)})^2k_s\geq\kappa_1\eta_T^2(u_0)||\mathbf{v}||^2Tl>0
$$

in probability approaching to 1. Next, given

$$
\widetilde{\xi}_T(u_0)=\inf_{\mathbf{w}\in\mathcal{H}_T}\mathbb{E}L_{t,T}(\mathbf{w})=\mathbb{E}\left[(\boldsymbol{\mu}-\widehat{\boldsymbol{\mu}}(\mathbf{w}^0(u_0)))'\mathbf{K}(u_0)(\boldsymbol{\mu}-\widehat{\boldsymbol{\mu}}(\mathbf{w}^0(u_0)))\right],
$$

we have $||\sqrt{\mathbf{K}(u_0)}\left[\boldsymbol{\mu}-\widehat{\boldsymbol{\mu}}(\mathbf{w}^0(u_0))\right]|| = O_p(\widetilde{\xi}_T^{1/2}(u_0))$. Then, from Conditions (C.4)-(C.5) and (C.7), similar to (A.5) in Li et al. (2022), it is shown that

$$
\begin{aligned}
|\Xi_2| &\leq ||\sqrt{\mathbf{K}(u_0)}\left[\boldsymbol{\mu}-\widehat{\boldsymbol{\mu}}(\mathbf{w}^0(u_0))\right]||\times||\sqrt{\mathbf{K}(u_0)}\widehat{\boldsymbol{\mu}}(\eta_T(u_0)\mathbf{v})||\\
&= O_p(\widetilde{\xi}_T^{1/2}(u_0)q^{1/2}T^{1/2}l^{1/2}\eta_T(u_0))||\mathbf{v}||.
\end{aligned}
$$

With Condition (C.10), it is shown that $|\Xi_2|$ is dominated by Ξ_1 asymptotically. Next, similar to (A.54) of Liao et al. (2019), it is observed that

$$\begin{aligned}
|\Xi_3| &= |\mathrm{tr}\{\boldsymbol{\varepsilon}'\mathbf{K}(u_0)\widehat{\boldsymbol{\mu}}(\eta_T(u_0)\mathbf{v})\}| \\
&= |\mathrm{tr}\{\mathbf{K}(u_0)\widehat{\boldsymbol{\mu}}(\eta_T(u_0)\mathbf{v})\boldsymbol{\varepsilon}'\}| \\
&= |\{\mathrm{vec}(\boldsymbol{\varepsilon}\widehat{\boldsymbol{\mu}}(\eta_T(u_0)\mathbf{v})')\}'\mathrm{vec}(\mathbf{K}(u_0))| \\
&= |\eta_T(u_0)\mathbf{v}'\{\mathrm{vec}(\boldsymbol{\varepsilon}\widehat{\boldsymbol{\mu}}^{(1)\prime}), \cdots, \mathrm{vec}(\boldsymbol{\varepsilon}\widehat{\boldsymbol{\mu}}^{(M_T)\prime})\}'\mathrm{vec}(\mathbf{K}(u_0))| \\
&\leq \eta_T(u_0)||\mathbf{v}||\mathrm{tr}^{1/2}\left[\sum_{m=1}^{M_T}\{\mathrm{vec}(\boldsymbol{\varepsilon}\widehat{\boldsymbol{\mu}}^{(m)\prime})\}\{\mathrm{vec}(\boldsymbol{\varepsilon}\widehat{\boldsymbol{\mu}}^{(m)\prime})\}'\right]||\mathbf{K}(u_0)|| \\
&= O_p(M_T q^{1/2}T^{1/2}l^{1/2}\eta_T(u_0))||\mathbf{v}||,
\end{aligned}$$

which is dominated by Ξ_1 asymptotically based on Condition (C.10). Furthermore, Ξ_4 can be decomposed as

$$\begin{aligned}
\Xi_4 &= \mathbf{Y}'\mathbf{M}(\mathbf{w}^0(u_0) + \eta_T(u_0)\mathbf{v}, \mathbf{X})\mathbf{Y} - \mathbf{Y}'\mathbf{M}(\mathbf{w}^0(u_0), \mathbf{X})\mathbf{Y} \\
&= \mathbf{Y}'\mathbf{Q}'(\eta_T(u_0)\mathbf{v}, \mathbf{X})\mathbf{K}(u_0)\mathbf{Y} + \mathbf{Y}'\mathbf{K}(u_0)\mathbf{Q}(\eta_T(u_0)\mathbf{v}, \mathbf{X})\mathbf{Y} \\
&\quad -\Xi_{41} - \Xi_{42} + \Xi_{43},
\end{aligned}$$

where

$$\begin{aligned}
\Xi_{41} &\equiv \mathbf{Y}'\mathbf{P}'(\mathbf{w}^0(u_0) + \eta_T(u_0)\mathbf{v}, \mathbf{X})\mathbf{K}(u_0)\mathbf{Q}(\mathbf{w}^0(u_0) + \eta_T(u_0)\mathbf{v}, \mathbf{X})\mathbf{Y} \\
&\quad -\mathbf{Y}'\mathbf{P}'(\mathbf{w}^0(u_0), \mathbf{X})\mathbf{K}(u_0)\mathbf{Q}(\mathbf{w}^0(u_0), \mathbf{X})\mathbf{Y},
\end{aligned}$$

$$\begin{aligned}
\Xi_{42} &\equiv \mathbf{Y}'\mathbf{Q}'(\mathbf{w}^0(u_0) + \eta_T(u_0)\mathbf{v}, \mathbf{X})\mathbf{K}(u_0)\mathbf{P}(\mathbf{w}^0(u_0) + \eta_T(u_0)\mathbf{v}, \mathbf{X})\mathbf{Y} \\
&\quad -\mathbf{Y}'\mathbf{Q}'(\mathbf{w}^0(u_0), \mathbf{X})\mathbf{K}(u_0)\mathbf{P}(\mathbf{w}^0(u_0), \mathbf{X})\mathbf{Y},
\end{aligned}$$

and

$$\begin{aligned}
\Xi_{43} &\equiv \mathbf{Y}'\mathbf{Q}'(\mathbf{w}^0(u_0) + \eta_T(u_0)\mathbf{v}, \mathbf{X})\mathbf{K}(u_0)\mathbf{Q}(\mathbf{w}^0(u_0) + \eta_T(u_0)\mathbf{v}, \mathbf{X})\mathbf{Y} \\
&\quad -\mathbf{Y}'\mathbf{Q}'(\mathbf{w}^0(u_0), \mathbf{X})\mathbf{K}(u_0)\mathbf{Q}(\mathbf{w}^0(u_0), \mathbf{X})\mathbf{Y}.
\end{aligned}$$

Denote

$$\mathbf{F}_1 \equiv (\sqrt{\mathbf{K}(u_0)}\mathbf{P}^{(1)}(\mathbf{X})\mathbf{Y}, \cdots, \sqrt{\mathbf{K}(u_0)}\mathbf{P}^{(M_T)}(\mathbf{X})\mathbf{Y}),$$

and

$$\mathbf{F}_2 \equiv (\sqrt{\mathbf{K}(u_0)}\mathbf{Q}^{(1)}(\mathbf{X})\mathbf{Y}, \cdots, \sqrt{\mathbf{K}(u_0)}\mathbf{Q}^{(M_T)}(X)\mathbf{Y}).$$

Then, we have

$$\begin{aligned} ||\mathbf{F}_1|| &= \mathrm{tr}^{1/2}(\mathbf{F}_1\mathbf{F}_1') = \mathrm{tr}^{1/2}\{\sum_{m=1}^{M_T}\mathbf{Y}'\mathbf{P}^{(m)'}(\mathbf{X})\mathbf{K}(u_0)\mathbf{P}^{(m)}(\mathbf{X})\mathbf{Y}\} \\ &\leq \{\sum_{m=1}^{M_T}\zeta_{\max}(\mathbf{K}(u_0))||\mathbf{Y}k_t||^2\zeta^2_{\max}(\mathbf{P}^{(m)}(\mathbf{X}))\}^{1/2} = O_p(\sqrt{qTlM_T}), \end{aligned}$$

and similarly

$$\begin{aligned} ||\mathbf{F}_2|| &= \{\sum_{m=1}^{M_T}||\sqrt{\mathbf{K}(u_0)}\mathbf{Q}^{(m)}(\mathbf{X})\mathbf{Y}||^2\}^{1/2} \\ &\leq \{\sum_{m=1}^{M_T}\zeta_{\max}(\mathbf{K}(u_0))\zeta^2_{\max}(\mathbf{Q}^{(m)}(\mathbf{X}))||\mathbf{Y}||^2\}^{1/2} \\ &= O_p(q^{1/2}T^{-1/2}l^{-1/2}M_T^{3/2}). \end{aligned}$$

Then, we have

$$\begin{aligned} & |\Xi_{41} + \Xi_{42} + \Xi_{43}| \\ \leq\ & 2\eta_T(u_0)||\mathbf{v}||||\mathbf{F}_1||||\mathbf{F}_2||||\mathbf{w}^0(u_0)|| + 2\eta_T(u_0)||\mathbf{w}^0(u_0)||||\mathbf{F}_1||||\mathbf{F}_2||||\mathbf{v}|| \\ & +2\eta_T^2(u_0)||\mathbf{F}_1||||\mathbf{F}_2||||\mathbf{v}||^2 + 2\eta_T(u_0)||\mathbf{v}||||\mathbf{F}_2||^2||\mathbf{w}^0(u_0)|| + \eta_T^2(u_0)||\mathbf{v}||^2||\mathbf{F}_2||^2 \\ =\ & O_p(\eta_T(u_0)qM_T^2)||\mathbf{v}|| + O_p(\eta_T^2(u_0)qM_T^2)||\mathbf{v}||^2 \\ & +O_p(T^{-1}l^{-1}\eta_T(u_0)qM_T^3)||\mathbf{v}|| + O_p(T^{-1}l^{-1}\eta_T^2(u_0)qM_T^3)||\mathbf{v}||^2. \end{aligned}$$

In addition, it is seen that

$$\begin{aligned} & |\mathbf{Y}'\mathbf{Q}'(\eta_T(u_0)\mathbf{v}, \mathbf{X})\mathbf{K}(u_0)\mathbf{Y} + \mathbf{Y}'\mathbf{K}(u_0)\mathbf{Q}(\eta_T(u_0)\mathbf{v}, \mathbf{X})\mathbf{Y}| \\ =\ & |2\mathbf{Y}'(\sum_{m=1}^{M_T}\eta_T(u_0)v^m\mathbf{Q}^{(m)'}(\mathbf{X}))\mathbf{K}(u_0)\mathbf{Y}| \\ \leq\ & 2\eta_T(u_0)||\mathbf{v}||||\mathbf{F}_2'\sqrt{\mathbf{K}(u_0)}\mathbf{Y}|| \\ =\ & O_p(\eta_T(u_0)q^{1/2}{M_T}^{3/2})||\mathbf{v}||. \end{aligned}$$

This shows that Ξ_4 is dominated by Ξ_1 asymptotically based on Conditions (C.6) and (C.10). Therefore, Theorem 2 is verified. □

Proof of Theorem 3. With Theorem (3.a) in Andrew (1992) and Theorem 2, it is easily obtained that

$$\begin{aligned}
& ||\widehat{\boldsymbol{\alpha}}(u_0, \widehat{\mathbf{w}}_{u_0}) - \boldsymbol{\alpha}^*(u_0, \mathbf{w}^0(u_0))|| \\
\leq\ & ||\widehat{\boldsymbol{\alpha}}(u_0, \widehat{\mathbf{w}}_{u_0}) - \widehat{\boldsymbol{\alpha}}(u_0, \mathbf{w}^0(u_0))|| + ||\widehat{\boldsymbol{\alpha}}(u_0, \mathbf{w}^0(u_0)) - \boldsymbol{\alpha}^*(u_0, \mathbf{w}^0(u_0))|| \\
=\ & ||\widehat{\boldsymbol{\alpha}}(u_0, \widehat{\mathbf{w}}_{u_0} - \mathbf{w}^0(u_0))|| + || \sum_{m=1}^{M_T} w_m^0 (\widehat{\boldsymbol{\alpha}}^{(m)}(u_0) - \boldsymbol{\alpha}^{(m)*}(u_0))|| \\
\leq\ & ||\widehat{\mathbf{w}}_{u_0} - \mathbf{w}^0(u_0)||||\widehat{\mathbf{A}}(u_0)|| + ||\mathbf{w}^0(u_0)||||\widehat{\mathbf{A}}(u_0) - \mathbf{A}^*(u_0)|| \\
=\ & O_p(T^{-1/2+\delta} l^{-1/2+\delta} M_T^{3/2} q^{3/2}) + O_p(M_T^{1/2} q^{1/2} T^{-1/2} l^{-1/2}) = o_p(1),
\end{aligned}$$

where $\widehat{\mathbf{A}}(u_0) = (\widehat{\boldsymbol{\alpha}}^{(1)}(u_0), \cdots, \widehat{\boldsymbol{\alpha}}^{(M_T)}(u_0))'$ and $\mathbf{A}^*(u_0) = (\boldsymbol{\alpha}^{(1)*}(u_0), \cdots, \boldsymbol{\alpha}^{(M_T)*}(u_0))'$. Thus, the proof of Theorem 3 is completed. □

Proof of Theorem 4. Let $\mathrm{FV}_T^*(u_0, \mathbf{w}) = \mathrm{FV}_T(u_0, \mathbf{w}) - |(\boldsymbol{\mu} - \mathbf{Y})' \mathbf{K}(u_0)(\boldsymbol{\mu} - \mathbf{Y})|$, where the last term is unrelated to $\mathbf{w}$, and thus, $\widehat{\mathbf{w}}_{u_0} = \arg\min_{\mathbf{w} \in \mathcal{H}_T} \mathrm{FV}_T(u_0, \mathbf{w}) = \arg\min_{\mathbf{w} \in \mathcal{H}_T} \mathrm{FV}_T^*(u_0, \mathbf{w})$. Similar to the proof of Theorem 1, it is obtained that

$$\mathrm{FV}_T^*(u_0, \widehat{\mathbf{w}}_{u_0}) = \mathbb{E} L_{t,T}^*(\widehat{\mathbf{w}}_{u_0}) + O_p(q M_T^{1/2} T^{1/2} l^{1/2}). \tag{A.13}$$

Denote $\tau = \sum_{m \in \mathcal{D}} w^m$. Let $\widetilde{\mathbf{w}}$ be a weight vector with $w^m = 0$ for the correctly specified models (i.e., $m \in \mathcal{D}$) and $\widetilde{w}^m = w^m/(1-\tau)$ 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

$$\mu_t^{(m)} - \mu_t = 0 \ \text{ for } m \in \mathcal{D}. \tag{A.14}$$

Then, we obtain that

$$\begin{aligned}
\mathbb{E} L_T^*(u_0, \mathbf{w}) &= \mathbb{E}\left[(\boldsymbol{\mu}^*(\mathbf{w}) - \boldsymbol{\mu})' \mathbf{K}(u_0)(\boldsymbol{\mu}^*(\mathbf{w}) - \boldsymbol{\mu})\right] \\
&= \mathbb{E}\left[\sum_{s=1}^{T} (\sum_{m=1}^{M_T} w^m \{\mu_s^{(m)*} - \mu_s\})^2 k_{st}\right] \\
&= \mathbb{E}\left[\sum_{s=1}^{T} (\sum_{m=1}^{M_0} w^m \{\mu_s^{(m)*} - \mu_s\})^2 k_{st}\right] \\
&= (1-\tau)^2 \mathbb{E}\left[\sum_{s=1}^{T} \left(\sum_{m=1}^{M_0} (1 - w^{M_0+1})^{-1} w^m \{\mu_s^{(m)*} - \mu_s\}\right)^2 k_{st}\right]
\end{aligned}$$

$$= \quad (1-\tau)^2 L_T^*(u_0, \widetilde{\mathbf{w}}). \tag{A.15}$$

By replacing $\mathbf{w}$ with its estimator and based on (A.13) and (A.15), we have

$$\mathrm{FV}_T^*(u_0, \widehat{\mathbf{w}}_{u_0}) = (1-\tau)^2 \mathbb{E} L_T^*(u_0, \widehat{\widetilde{\mathbf{w}}}_{u_0}) + O_p(qM_T^{1/2}T^{1/2}l^{1/2}).$$

Let $\boldsymbol{\lambda}$ be a weight vector with $\sum_{m\in\mathcal{D}} \lambda^m$. Also, we obtain $\mathbb{E}L_T^*(u_0, \boldsymbol{\lambda}) = 0$ based on (A.14). Then, it is shown that

$$\mathrm{FV}_T^*(u_0, \boldsymbol{\lambda}) = O_p(qM_T^{1/2}T^{1/2}l^{1/2}),$$

and

$$(1-\tau)^2 \mathbb{E} L_T^*(u_0, \widehat{\widetilde{\mathbf{w}}}_{u_0}) + O_p(qM_T^{1/2}T^{1/2}l^{1/2}) \leq \mathrm{FV}_T^*(u_0, \boldsymbol{\lambda}) = O_p(qM_T^{1/2}T^{1/2}l^{1/2})$$

based on the fact that $\widehat{\mathbf{w}}_{u_0} = \arg\min_{\mathbf{w}\in\mathcal{H}_T} \mathrm{FV}_T^*(u_0, \mathbf{w})$. Therefore, we derive that

$$(1-\widehat{\tau}_{u_0})^2 \inf_{\mathbf{w}\in\widetilde{\mathcal{H}}_T} \mathbb{E} L_T^*(u_0, \mathbf{w}) + O_p(qM_T^{1/2}T^{1/2}l^{1/2}) \leq O_p(qM_T^{1/2}T^{1/2}l^{1/2}),$$

where $\widehat{\tau}_{u_0}$ is the estimator of τ for any given u_0, i.e., $\widehat{\tau}_{u_0} = \sum_{j\in\mathcal{D}} \widehat{w}_{u_0}^j$. Combined with Condition (C.11), for any given u_0 we obtain $\widehat{\tau}_{u_0} \xrightarrow{P} 1$, and thus, the proof of Theorem 4 is completed. □

Proof of Theorem 5. To verify Theorem 5, it is equivalent to showing that for any given u_0 and for any $\delta > 0$,

$$\Pr\left\{ \left| \frac{\inf_{\mathbf{w}\in\mathcal{H}_T} L_T(u_0, \mathbf{w})}{L_T(u_0, \widehat{\mathbf{w}}_{u_0}^*)} - 1 \right| > \delta \right\} \xrightarrow{P} 0. \tag{A.16}$$

First, we define $DF_T(u_0, \mathbf{w}) = FV_T(u_0, \mathbf{w}) - L_T(u_0, \mathbf{w}) - |(\boldsymbol{\mu} - \mathbf{Y})'\mathbf{K}(u_0)(\boldsymbol{\mu} - \mathbf{Y})|$. Based on (A.4) and (A.5), $\sup_{\mathbf{w}\in\mathcal{H}_T} |DF_T(u_0, \mathbf{w})/L_T^*(u_0, \mathbf{w})| = o_p(1)$ for any given u_0. With (A.5), Conditions (C.6) and (C.12), it is observed that for any given u_0

$$\sup_{\mathbf{w}\in\mathcal{H}_T} \left| \frac{v_T(u_0)}{L_T^*(u_0, \mathbf{w})} \right| = o_p(1),$$

and

$$\sup_{\mathbf{w}\in\mathcal{H}_T} \left| \frac{L_T^*(u_0, \mathbf{w})}{L_T(u_0, \mathbf{w})} \right|$$

$$
\begin{aligned}
&= \left\{ \inf_{\mathbf{w}\in\mathcal{H}_T} \left| \frac{L_T(u_0,\mathbf{w})}{L_T^*(u_0,\mathbf{w})} \right| \right\}^{-1} \\
&\leq \left\{ 1 - \sup_{\mathbf{w}\in\mathcal{H}_T} \left| \frac{L_T(u_0,\mathbf{w}) - L_T^*(u_0,\mathbf{w})}{L_T^*(u_0,\mathbf{w})} \right| \right\}^{-1} \\
&\xrightarrow{P} 1,
\end{aligned}
$$

where the last step is obtained from (A.3) and (A.5). Similarly, we have

$$
\begin{aligned}
&\sup_{\mathbf{w}\in\mathcal{H}_T} \left| \frac{L_T^*(u_0,\mathbf{w})}{L_T(u_0,\mathbf{w}) - v_T(u_0)} \right| \\
&\leq \left\{ 1 - \sup_{\mathbf{w}\in\mathcal{H}_T} \left| \frac{L_T(u_0,\mathbf{w}) - L_T^*(u_0,\mathbf{w})}{L_T^*(u_0,\mathbf{w})} \right| - \sup_{\mathbf{w}\in\mathcal{H}_T} \left| \frac{v_T(u_0)}{L_T^*(u_0,\mathbf{w})} \right| \right\}^{-1} \\
&\xrightarrow{P} 1.
\end{aligned}
$$

Finally, we will prove (A.16). It is observed that

$$
\begin{aligned}
&\Pr\left\{ \left| \frac{\inf_{\mathbf{w}\in\mathcal{H}_T} L_T(u_0,\mathbf{w})}{L_T(u_0,\widehat{\mathbf{w}}_{u_0}^*)} - 1 \right| > \delta \right\} \\
&= \Pr\left\{ \left| \frac{L_T(u_0,\widehat{\mathbf{w}}_{u_0}^*) - \inf_{\mathbf{w}\in\mathcal{H}_T} L_T(u_0,\mathbf{w})}{L_T(u_0,\widehat{\mathbf{w}}_{u_0}^*)} \right| > \delta \right\} \\
&= \Pr\left\{ \left| \frac{FV_T(u_0,\widehat{\mathbf{w}}_{u_0}^*) - DF_T(u_0,\widehat{\mathbf{w}}_{u_0}^*) - \inf_{\mathbf{w}\in\mathcal{H}_T} L_T(u_0,\mathbf{w})}{L_T(u_0,\widehat{\mathbf{w}}_{u_0}^*)} \right| > \delta \right\} \\
&= \Pr\left\{ \left| \frac{\inf_{\mathbf{w}\in\mathcal{H}_T^*}[L_T(u_0,\mathbf{w}) + DF_T(u_0,\mathbf{w})] - DF_T(u_0,\widehat{\mathbf{w}}_{u_0}^*) - \inf_{\mathbf{w}\in\mathcal{H}_T} L_T(u_0,\mathbf{w})}{L_T(u_0,\widehat{\mathbf{w}}_{u_0}^*)} \right| > \delta \right\} \\
&\leq \Pr\left\{ \left| \frac{\inf_{\mathbf{w}\in\mathcal{H}_T^*}[L_T(u_0,\mathbf{w}) + DF_T(u_0,\mathbf{w})] - DF_T(u_0,\widehat{\mathbf{w}}_{u_0}^*) - \inf_{\mathbf{w}\in\mathcal{H}_T} L_T(u_0,\mathbf{w})}{L_T(u_0,\widehat{\mathbf{w}}_{u_0}^*)} \right| > \delta | \mathbf{w}_T \in \mathcal{H}_T^* \right\} \\
&\quad \times \Pr(\mathbf{w}_T \in \mathcal{H}_T^*) + \Pr(\mathbf{w}_T \notin \mathcal{H}_T^*) \\
&\leq \Pr\left\{ \left| \frac{DF_T(u_0,\mathbf{w}_T) - DF_T(u_0,\widehat{\mathbf{w}}_{u_0}^*) + v_T(u_0)}{L_T(u_0,\widehat{\mathbf{w}}_{u_0}^*)} \right| > \delta \right\} + \Pr(\mathbf{w}_T \notin \mathcal{H}_T^*) \\
&\leq \Pr\left\{ \left| \frac{DF_T(u_0,\mathbf{w}_T)}{\inf_{\mathbf{w}\in\mathcal{H}_T} L_T(u_0,\mathbf{w})} \right| + \sup_{\mathbf{w}\in\mathcal{H}_T} \left| \frac{DF_T(u_0,\mathbf{w})}{L_T(u_0,\mathbf{w})} \right| + \sup_{\mathbf{w}\in\mathcal{H}_T} \left| \frac{v_T(u_0)}{L_T(u_0,\mathbf{w})} \right| > \delta \right\} + \Pr(\mathbf{w}_T \notin \mathcal{H}_T^*) \\
&\leq \Pr\left\{ \sup_{\mathbf{w}\in\mathcal{H}_T} \left| \frac{DF_T(u_0,\mathbf{w})}{L_T^*(u_0,\mathbf{w})} \right| \sup_{\mathbf{w}\in\mathcal{H}_T} \left| \frac{L_T^*(u_0,\mathbf{w})}{L_T(u_0,\mathbf{w}) - v_T(u_0)} \right| + \sup_{\mathbf{w}\in\mathcal{H}_T} \left| \frac{DF_T(u_0,\mathbf{w})}{L_T^*(u_0,\mathbf{w})} \right| \sup_{\mathbf{w}\in\mathcal{H}_T} \left| \frac{L_T^*(u_0,\mathbf{w})}{L_T(u_0,\mathbf{w})} \right| \right. \\
&\quad \left. + \sup_{\mathbf{w}\in\mathcal{H}_T} \left| \frac{v_T(u_0)}{L_T^*(u_0,\mathbf{w})} \right| \sup_{\mathbf{w}\in\mathcal{H}_T} \left| \frac{L_T^*(u_0,\mathbf{w})}{L_T(u_0,\mathbf{w})} \right| > \delta \right\} + \Pr(\mathbf{w}_T \notin \mathcal{H}_T^*) \xrightarrow{P} 0.
\end{aligned}
$$

Thus, (A.16) is obtained and therefore, the proof of Theorem 5 is completed. □