

Cross Validation Based Model Averaging for Varying-Coefficient Models with Response Missing at Random

Huixin Li, Xiuli Wang*

School of Mathematics and Statistics, Shandong Normal University, Jinan, China

Email: *wxlmath@163.com

How to cite this paper: Li, H.X., and Wang, X.L. (2024) Cross Validation Based Model Averaging for Varying-Coefficient Models with Response Missing at Random. *Journal of Applied Mathematics and Physics*, 12, 764-777.

<https://doi.org/10.4236/jamp.2024.123047>

Received: February 20, 2024

Accepted: March 22, 2024

Published: March 25, 2024

Copyright © 2024 by author(s) and Scientific Research Publishing Inc.

This work is licensed under the Creative Commons Attribution International

License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

Abstract

In this paper, a model averaging method is proposed for varying-coefficient models with response missing at random by establishing a weight selection criterion based on cross-validation. Under certain regularity conditions, it is proved that the proposed method is asymptotically optimal in the sense of achieving the minimum squared error.

Keywords

Response Missing at Random, Model Averaging, Asymptotic Optimality, B-Spline Approximation

1. Introduction

In the past two decades, there have been more and more studies on model averaging and model selection. Model selection is to select a final model before estimation, but due to the uncertainty of the model, there will be a deviation caused by the selection of wrong model. Model averaging is to estimate a large number of candidate models, and then give greater weight to the sub-models with better prediction results, that is, to select the model after estimation, and the candidate model does not require the inclusion of the real model, which can effectively overcome the shortcomings of the model selection method. Therefore, model averaging is increasingly used in various fields.

Most of the model averaging methods have been devoted to the frequency model averaging, and mainly in parametric models. For example, Hansen (2007) [1], Zhang *et al.* (2008) [2], Wan *et al.* (2010) [3] proposed the Mallows-type model averaging (MMA) method. To further improve the method, Zhang *et al.*

(2016) [4] proposed a weight selection criterion based on the Kullback-Leibler loss with penalty term for generalized linear models and generalized linear mixed-effects models. Using nonparametric kernel smoothing method to estimate nonparametric functions, Li *et al.* (2018) [5], Zhang and Wang (2019) [6] extended the Mallows-type model averaging method to semiparametric models. Recently, Xia (2021) [7] extended the Mallows-type model averaging method to varying-coefficient models using B-spline to estimate the nonparametric functions. However, the above mentioned literatures focus on the case where the data is completely observed. In statistical studies such as economics, market research and medical research, missing data is a common phenomenon. According to the different missing mechanisms, it can be divided into three types, namely Missing completely at random (MCAR), Missing at random (MAR) and Missing not ant random (MNAR). Little attention has been paid to model averaging with missing data. One of the reasons is that the complexity of missing data makes it very challenging to extend existing methods with complete data to missing data. The study of model averaging with missing data is still a relatively new topic and needs further development.

In this paper, we focus on model averaging for varying-coefficient models with response missing at random (MAR), which has been widely studied in the field of missing data. As far as we know, there are few literatures on the model averaging with missing response. For example, Sun *et al.* (2014) [8] developed a model average estimation approach for linear regression models with response missing at random under a local misspecification framework, called smoothed focused vector information criteria (SFVIC). Zeng *et al.* (2018) [9] extended the SFVIC to varying-coefficient partially linear models with response missing at random by employing the profile least-squares estimation process and inverse probability weighted method to estimate regression coefficients of the models under the framework of local misspecification. However, these researches are limited to the local misspecification framework, and the model averaging without local misspecification framework is relatively small. For example, based on the weighted delete-one cross-validation (WDCV) criterion, Xie *et al.* (2021) [10] proposed a two-step model averaging procedure for high-dimensional linear regression models with missing response without local misspecification framework, and under certain conditions, they proved the WDCV criterion asymptotically achieved the lowest possible prediction loss. But their criterion required the covariates used in different candidate models cannot overlap. Based on the Mallows-type model averaging method, Wei *et al.* (2021) [11] established the HRC_p criterion of linear models with missing response, which improved the shortcomings of MMA in estimating the error covariance matrix. Wei and Wang (2021) [12] developed a linear models averaging procedure with response missing at random by establishing a cross-validation-based weight choice criterion, and proved its asymptotic optimality. The three literatures mentioned above are limited to parametric linear models.

Inspired by Xia (2021) [7], Wei and Wang (2021) [12], and Wei *et al.* (2021) [11], in this paper we shall extend the Jackknife model averaging (JMA) method (e.g. Hansen and Racine (2012) [13] and Zhang *et al.* (2013) [14]) to varying-coefficient models with response missing at random and heteroscedasticity error. The varying-coefficient model is a common type of non-parametric model. It has no fixed regression function and is more flexible in form. It can further fit complex data and is more widely used in various fields. Following Xia (2021) [7], we use B-spline to estimate the nonparametric varying-coefficient functions, and use the inverse probability weighted (IPW) method to deal with the missing data. In order to avoid the “curse of dimensionality”, we assume a parametric model for propensity score function. Compared with previous articles, our method has the following advantages. Firstly, we use B-spline to estimate the varying-coefficient functions, which results in less computational burden compared with the local kernel function method. Secondly, we choose the weight vector based on JMA method. Compared with the MMA method, our method does not need to estimate the covariance matrix of the error and is more suitable for the heteroscedasticity models, which makes it easier to deal with real data. In the theoretical proof, we do not need to consider the relationship between the estimated error covariance matrix and the real error covariance matrix, which makes our theoretical proof easier. Finally, our method can deal with three cases for candidate models (Nested case, Marginal case and Full case), not just the case where the candidate models are nested. Under certain conditions, our proposal is asymptotically optimal in the sense of achieving the lowest possible squared error, which shows that our method is feasible. The finite sample performance of our proposal is studied by numerical simulations and compared with some related methods. The simulation results display that our proposal is feasible.

The rest of this paper is organized as follows. Section 2 and Section 3 describe the method of model averaging estimator and weight selection respectively. Some conditions and theoretical result for asymptotic optimality are given in Section 4. The proofs of the main results are given in Section 5.

2. Model Framework and Model Averaging Estimation Method

We consider the following varying-coefficient model:

$$\begin{cases} Y_i = \mu_i + \varepsilon_i \\ \mu_i = \sum_{j=1}^p X_{ij} \beta_j(U_i) \end{cases}, \quad i = 1, 2, \dots, n, \quad (1)$$

where Y_i is response variable, the index variable U_i is a scalar and ε_i is a random error with $E(\varepsilon_i | \mathbf{X}_i, U_i) = 0$ and $Var(\varepsilon_i | \mathbf{X}_i, U_i) = \sigma_i^2$. We assume $\mu_i = E(Y_i | \mathbf{X}_i, U_i) = \sum_{j=1}^p X_{ij} \beta_j(U_i) = \mathbf{X}_i^T \boldsymbol{\beta}(U_i)$, where $\mathbf{X}_i = (X_{i1}, \dots, X_{ip})^T$ is a p -dimensional covariate, $\boldsymbol{\beta}(U_i) = (\beta_1(U_i), \dots, \beta_p(U_i))^T$ corresponds to the vector of the unknown coefficient functions.

Suppose that X_i and U_i are completely observable, and Y_i may be missing, where $\delta_i = 1$ if Y_i is observed, $\delta_i = 0$ otherwise. In this paper, we assume that Y_i is missing at random (MAR), that is

$$P(\delta_i = 1 | Y_i, \mathbf{X}_i, U_i) = P(\delta_i = 1 | \mathbf{X}_i, U_i) := \pi(\mathbf{X}_i, U_i). \tag{2}$$

Let $Z_{\pi,i} = \{\pi(\mathbf{X}_i, U_i)\}^{-1} \delta_i Y_i, i = 1, 2, \dots, n$, using model (1) and MAR assumption, we can obtain

$$\begin{cases} Z_{\pi,i} = \mu_i + \varepsilon_{\pi,i} \\ \mu_i = E(Z_{\pi,i} | \mathbf{X}_i, U_i) \end{cases}, \quad i = 1, 2, \dots, n, \tag{3}$$

where $E(\varepsilon_{\pi,i} | \mathbf{X}_i, U_i) = 0, \text{Var}(\varepsilon_{\pi,i} | \mathbf{X}_i, U_i) = \sigma_{\pi,i}^2$,

$$\sigma_{\pi,i}^2 = \left[\{\pi(\mathbf{X}_i, U_i)\}^{-1} - 1 \right] \mu_i^2 + \{\pi(\mathbf{X}_i, U_i)\}^{-1} \sigma_i^2.$$

We consider a series of approximate models. In particular, we use M_n candidate model $\{\mathcal{M}_1, \dots, \mathcal{M}_{M_n}\}$ to approximate the real data generating process of Y . The m th candidate model \mathcal{M}_m contains k_m covariates, that is, $k_m = |\mathcal{M}_m|$ is the cardinality of \mathcal{M}_m . The total number of all possible candidate models is $M_n = 2^p$, which contains an empty model that excludes all covariates. From model (3), under the m th model, we have

$$Z_{\pi,i} = \sum_{j=1}^{k_m} X_{ij(m)} \beta_{j(m)}(U_i) + \varepsilon_{\pi,i}. \tag{4}$$

The purpose of this paper is to construct an asymptotically optimal model average estimator of the conditional mean $\mu = (\mu_1, \mu_2, \dots, \mu_n)^T$ with missing response under MAR assumption.

Note that $\{Z_{\pi,i}, \mathbf{X}_i, U_i\}$ is fully observed, if the selection probability function $\pi(X_i, U_i)$ is known, the estimator of $\beta_{j(m)}(u)$ can be obtained by using B-spline to approximate the coefficient function for $\{Z_{\pi,i}, \mathbf{X}_i, U_i\}$. Specifically, we first approximate each $\beta_{j(m)}(\cdot)$ with a function in a polynomial spline space. Without losing generality, we assume that U has a compact set $\mathcal{U} = [0, 1]$. Let $\mathbf{B}(u) = (B_1(u), \dots, B_{L_n}(u))^T$ is a B-spline function basis of order r , where $L_n = J_n + r + 1, J_n$ is the number of interior knots and increases with the increase of sample size n . According to the B-spline theory in De Boor (2001) [15], there exists a parameter vector $\gamma_{j(m)}$ satisfying

$$\beta_{j(m)}(u) \approx \mathbf{B}(u)^T \gamma_{j(m)},$$

where $\beta_{j(m)}(u)$ is the j th component of $\beta_{(m)}(u)$. So the m th model can be re-expressed as

$$Z_{\pi,i} = \sum_{j=1}^{k_m} X_{ij(m)} \mathbf{B}(U_i)^T \gamma_{j(m)} + \varepsilon_{\pi,i} \approx \sum_{j=1}^{k_m} \mathbf{W}_{i(m)} \gamma_{j(m)} + \varepsilon_{\pi,i},$$

where $\mathbf{W}_{i(m)} = (X_{i1} \mathbf{B}(U_i)^T, X_{i2} \mathbf{B}(U_i)^T, \dots, X_{ik_m} \mathbf{B}(U_i)^T)^T, i = 1, 2, \dots, n$ is a $k_m L_n$ dimensional column vector, $\mathbf{B}(U_i) = (B_1(U_i), \dots, B_{L_n}(U_i))^T$.

Therefore, $\beta_{j(m)}(u)$ can be obtained by solving the following least squares optimization problem

$$\min_{\gamma_{j(m)}} \sum_{i=1}^n \left[Z_{\pi,i} - \sum_{j=1}^{k_m} \mathbf{W}_{i(m)} \gamma_{j(m)} \right]^2. \tag{5}$$

Let $Z_{\pi} = (Z_{\pi,1}, Z_{\pi,2}, \dots, Z_{\pi,n})^T$, $\gamma_{(m)} = (\gamma_{j(m)}^T, j \in \mathcal{M}_m)^T$ is a vector of length $k_m L_n$, then the solution of (5) is given by

$$\hat{\gamma}_{(m)} = [\mathbf{W}_{(m)}^T \mathbf{W}_{(m)}]^{-1} \mathbf{W}_{(m)}^T Z_{\pi},$$

where $\mathbf{W}_{(m)} = (\mathbf{W}_{1(m)}, \dots, \mathbf{W}_{n(m)})^T$ is a $n \times k_m L_n$ -dimensional matrix. Thus, the estimator of $\beta_{(m)}(u)$ is $\hat{\beta}_{(m)}(u) = (\hat{\beta}_{j(m)}(u), j \in \mathcal{M}_m)^T$ with $\hat{\beta}_{j(m)}(u) = \mathbf{B}(u)^T \hat{\gamma}_{j(m)}$.

For the m th candidate model, the estimator of μ is then obtained as

$$\hat{\mu}_{\pi(m)} = \mathbf{W}_{(m)} [\mathbf{W}_{(m)}^T \mathbf{W}_{(m)}]^{-1} \mathbf{W}_{(m)}^T Z_{\pi} = \mathbf{P}_{(m)} Z_{\pi}, \tag{6}$$

where $\mathbf{P}_{(m)} = \mathbf{W}_{(m)} [\mathbf{W}_{(m)}^T \mathbf{W}_{(m)}]^{-1} \mathbf{W}_{(m)}^T$. From (6), we know that $\hat{\mu}_{\pi(m)}$ is linearly related to Z_{π} , we can define the model average estimator of μ as the weighted average of $\hat{\mu}_{\pi(m)}$, that is

$$\hat{\mu}_{\pi}(\omega) = \sum_{m=1}^{M_n} \omega_m \hat{\mu}_{\pi(m)} = \sum_{m=1}^{M_n} \omega_m \mathbf{P}_{(m)} Z_{\pi} = \mathbf{P}(\omega) Z_{\pi}, \tag{7}$$

where $\mathbf{P}(\omega) = \sum_{m=1}^{M_n} \omega_m \mathbf{P}_{(m)}$, $\omega = (\omega_1, \dots, \omega_{M_n})^T$ is the weight vector, which satisfies

$$\mathcal{H}_n = \left\{ \omega \in [0,1]^{M_n} : \sum_{m=1}^{M_n} \omega_m = 1 \right\}.$$

3. Weight Selection

According to Hansen and Racine (2012) [13], we give the details for the jackknife selection (also known as leave-one-out cross validation) of ω as follows.

The jackknife estimator of the m th model is denoted by

$$\tilde{\mu}_{\pi(m)} = (\tilde{\mu}_{\pi(m)}^1, \tilde{\mu}_{\pi(m)}^2, \dots, \tilde{\mu}_{\pi(m)}^n)^T,$$

where $\tilde{\mu}_{\pi(m)}^i = \mathbf{W}_{(m)}^i [\mathbf{W}_{(m)}^{(-i)T} \mathbf{W}_{(m)}^{(-i)}]^{-1} \mathbf{W}_{(m)}^{(-i)T} Z_{\pi}^{(-i)}$, and $\mathbf{W}_{(m)}^{(-i)}$ and $Z_{\pi}^{(-i)}$ are defined as matrices $\mathbf{W}_{(m)}$ and column vectors Z_{π} with the i th row deleted. After some calculations, we can derive the following relationship:

$$\tilde{\mathbf{P}}_{(m)} = \mathbf{D}_{(m)} (\mathbf{P}_{(m)} - \mathbf{I}_n) + \mathbf{I}_n,$$

where $\mathbf{D}_{(m)}$ is an $n \times n$ diagonal matrix and the i th diagonal element is

$$(1 - P_{(m),ii})^{-1} \text{ with } P_{(m),ii} \text{ the } i\text{-th diagonal element of } \mathbf{P}_{(m)}, \text{ then}$$

$\tilde{\mu}_{\pi}(\omega) = \tilde{\mathbf{P}}(\omega) Z_{\pi}$, $\tilde{\mathbf{P}}(\omega) = \sum_{m=1}^{M_n} \omega_m \tilde{\mathbf{P}}_{(m)}$. Therefore, we obtain the following JMA criterion for selecting the weight vector ω :

$$CV_{\pi}(\omega) = \|Z_{\pi} - \tilde{\mu}_{\pi}(\omega)\|^2. \tag{8}$$

In fact, $\pi(X_i, U_i)$ in (8) is unknown and need to be modeled. Following Wei

and Wang (2021) [12], we assume that $\pi(\mathbf{X}, U; \alpha)$ is a parametric model of $\pi(\mathbf{X}, U)$, where $\pi(\cdot; \alpha)$ is a known function and can be correctly specified while α is an unknown parameter vector. Denote $\hat{\alpha}_n$ as the maximum likelihood estimate (MLE) of α , $\hat{\pi}(\mathbf{X}, U) = \pi(\mathbf{X}, U; \hat{\alpha}_n)$. Replacing $\pi(\mathbf{X}, U)$ in $CV_\pi(\omega)$ with $\hat{\pi}(\mathbf{X}, U)$, we can get the following criterion:

$$CV_{\hat{\pi}}(\omega) = \|Z_{\hat{\pi}} - \tilde{\mu}_{\hat{\pi}}(\omega)\|^2. \tag{9}$$

Then the optimal weight vector $\hat{\omega}_{\hat{\pi}}$ is defined as

$$\hat{\omega}_{\hat{\pi}} = \arg \min_{\omega \in \mathcal{X}_n} CV_{\hat{\pi}}(\omega). \tag{10}$$

Thus, the model average estimator of μ is $\hat{\mu}_{\hat{\pi}}(\hat{\omega}_{\hat{\pi}})$, and named as the varying-coefficient jackknife model average estimator (VC-JMA).

4. Asymptotically Optimality

In this section, we investigate the asymptotic optimality of the proposed estimator $\hat{\mu}_{\hat{\pi}}(\hat{\omega}_{\hat{\pi}})$. We first give some necessary symbols. Define

$$L_\pi(\omega) = \|\mu - \hat{\mu}_\pi(\omega)\|^2, R_\pi(\omega) = E\{L_\pi(\omega) | \mathbf{X}, U\}, \tag{11}$$

as the loss function and risk function of $\hat{\mu}_\pi(\omega)$, respectively.

$\Omega_{e_\pi} = \text{diag}(\sigma_{\pi,1}^2, \dots, \sigma_{\pi,n}^2)$. Let $\xi_\pi = \inf_{\omega \in \mathcal{X}_n} R_\pi(\omega)$, ω_m^0 be a M_n -dimensional unit vector whose the m th element is one and the other elements are zeros and $\mathbf{X}_u = (\mathbf{X}_{(1)}, \mathbf{X}_{(2)}, \dots, \mathbf{X}_{(M_n)})^\top$.

Next, we give some conditions which are required for proving the asymptotic optimality of $\hat{\mu}_{\hat{\pi}}(\hat{\omega}_{\hat{\pi}})$. All the limiting processes discussed here and throughout the paper are with respect to $n \rightarrow \infty$.

(C1) The MLE $\hat{\alpha}_n$ of α is \sqrt{n} consistent and satisfies the regularity conditions of asymptotic normality. $\pi(\mathbf{X}, U; \alpha)$ is bounded away from 0 and twice continuously differentiable with respect to α . For all α 's in a neighborhood of

$$\alpha_0, \max_{1 \leq i \leq n} \left| \frac{\partial \pi(\mathbf{X}_i, U_i; \alpha)}{\partial \alpha} \right| = O_p(1), \text{ where } \alpha_0 \text{ is the true value of } \hat{\alpha}_n.$$

(C2) $\lambda_{\max}\{\Omega_{e_\pi}\} \leq C_e$, $p_{(m),ii} \leq C_p k_m n^{-1}$, a.s., where C_e and C_p are constant.

(C3) $r_u \xi_\pi^{-1} \xrightarrow{\text{a.s.}} 0$, where r_u is the rank of \mathbf{X}_u .

(C4) $M_n \xi_\pi^{-2G} \sum_{m=1}^{M_n} (R_\pi(\omega_m^0))^G \xrightarrow{\text{a.s.}} 0$, $\mu^\top \mu n^{-1} = O(1)$, $E(\varepsilon_{\pi,i}^{4G} | \mathbf{X}_i, U_i) \leq C_v < \infty$,

a.s., where G is an integer satisfying $1 \leq G < \infty$.

(C5) Functions $\beta_j(u), j = 1, \dots, p$ belong to a class of functions \mathcal{B} , whose r th derivatives $\beta_j^{(r)}(u)$ s exist and are Lipschitz of order β_0 .

(C6) The density f_U of U is bounded away from 0 and infinity on its support.

Remark 1 As Zhang *et al.* (2013) [14] pointed out, condition (C2) is commonly used in the study of asymptotic optimality of cross-validation methods. The condition (C3) is based on Zhang *et al.* (2013) [14]'s condition (22). For a

detailed explanation of the condition, see Zhang *et al.* (2013) [14]. Conditions (C4) is commonly used in model averaging literature, see for example Wan *et al.* (2010) [3], Zhang *et al.* (2013) [14] and Zhu *et al.* (2019) [16].

Remark 2 Conditions (C5) and (C6) are two common assumptions in approximating nonparametric coefficient functions with B-spline basis functions, which can be seen in Fan *et al.* (2014) [17].

With these conditions, the following theorem states the asymptotic optimality of $\hat{\mu}_{\hat{\omega}_{\pi}}(\hat{\omega}_{\pi})$.

Theorem 1 Suppose that conditions (C1) - (C6) hold, then

$$\frac{L_{\hat{\omega}_{\pi}}(\hat{\omega}_{\pi})}{\inf_{\omega \in \mathcal{X}_n} L_{\hat{\omega}_{\pi}}(\omega)} \xrightarrow{p} 1, \tag{12}$$

where $L_{\hat{\omega}_{\pi}}(\omega) = \|\mu - \hat{\mu}_{\hat{\omega}_{\pi}}(\omega)\|^2$.

Theorem 1 implies that our proposed estimator $\hat{\mu}_{\hat{\omega}_{\pi}}(\hat{\omega}_{\pi})$ is asymptotically optimal because the selected weight vector $\hat{\omega}_{\pi}$ yields a squared error that is asymptotically identical to that of the infeasible optimal weight vector.

5. Theorem Proof

Before proving Theorem 1, we first introduce some symbols used in the proof process. We use c to represent a universal positive constant, which can take different values in different situations. For any matrix A , we define $\lambda_{\max}\{A\}$ as the maximum singular value of A . Let $\tilde{\xi}_{\pi} = \inf_{\omega \in \mathcal{X}_n} \tilde{R}_{\pi}(\omega)$. According to the definitions of $L_{\pi}(\omega)$ and $R_{\pi}(\omega)$ in (11), we define the loss function and risk function of $\tilde{\mu}_{\pi}(\omega)$ in (8) by

$$\tilde{L}_{\pi}(\omega) = \|\mu - \tilde{\mu}_{\pi}(\omega)\|^2, \quad \tilde{R}_{\pi}(\omega) = E\{\tilde{L}_{\pi}(\omega) | \mathbf{X}, U\},$$

respectively, and then obtain

$$\tilde{R}_{\pi}(\omega) = \|\tilde{\mathbf{A}}(\omega)\mu\|^2 + \text{tr}[\tilde{\mathbf{P}}(\omega)^T \Omega_{e_{\pi}} \tilde{\mathbf{P}}(\omega)],$$

where $\tilde{\mathbf{A}} = \mathbf{I} - \tilde{\mathbf{P}}(\omega)$. Let $Q_{(m)}$ be an $n \times n$ diagonal matrix and the i th diagonal element is $P_{(m),ii} (1 - P_{(m),ii})^{-1}$. Then according to the definition of $D_{(m)}$ given in Section 2.2, we obtain

$$\tilde{\mathbf{P}}_{(m)} = \mathbf{P}_{(m)} - \mathbf{Q}_m \mathbf{A}_{(m)},$$

where $\mathbf{A}_{(m)} = \mathbf{I}_n - \mathbf{P}_{(m)}$.

Next, give the four Lemmas needed to prove Theorem 1.

Lemma 1 Let

$$\omega^* = \arg \min_{\omega \in \mathcal{X}_n} (\tilde{L}_{\pi}(\omega) + a_{\pi}(\omega)),$$

if, as $n \rightarrow \infty$,

$$\sup_{\omega \in \mathcal{X}_n} \frac{|a_{\pi}(\omega)|}{\tilde{R}_{\pi}(\omega)} \xrightarrow{p} 0, \quad \sup_{\omega \in \mathcal{X}_n} \left| \frac{\tilde{L}_{\pi}(\omega)}{\tilde{R}_{\pi}(\omega)} - 1 \right| \xrightarrow{p} 0,$$

and there exists a constant c such that

$$\inf_{\omega \in \mathcal{Z}_n} \tilde{R}_\pi(\omega) \geq c > 0,$$

almost surely, then

$$\frac{\tilde{L}_\pi(\omega^*)}{\inf_{\omega \in \mathcal{Z}_n} \tilde{L}_\pi(\omega)} \xrightarrow{p} 1.$$

The proof of Lemma 1 refers to Zhang *et al.* (2013) [14].

Lemma 2 If (C4) is satisfied, we have

$$\sup_{\omega \in \mathcal{Z}_n} \left| \frac{\tilde{L}_\pi(\omega)}{\tilde{R}_\pi(\omega)} - 1 \right| = o_p(1). \tag{13}$$

Proof of Lemma 2. Note that

$$\begin{aligned} |\tilde{L}_\pi(\omega) - \tilde{R}_\pi(\omega)| &= \left| \|\mu - \tilde{\mu}_\pi(\omega)\|^2 - \|\tilde{\mathbf{A}}(\omega)\mu\|^2 - \text{tr}\{\tilde{\mathbf{P}}(\omega)^\top \Omega_{e_\pi} \tilde{\mathbf{P}}(\omega)\} \right| \\ &= \left| \|\tilde{\mathbf{P}}(\omega)\varepsilon_\pi\|^2 - \text{tr}\{\tilde{\mathbf{P}}(\omega)^\top \Omega_{e_\pi} \tilde{\mathbf{P}}(\omega)\} - 2\mu^\top \tilde{\mathbf{A}}(\omega)^\top \tilde{\mathbf{P}}(\omega)\varepsilon_\pi \right|. \end{aligned}$$

Therefore, to prove (13), it suffices to prove that

$$\sup_{\omega \in \mathcal{Z}_n} \frac{\left| \|\tilde{\mathbf{P}}(\omega)\varepsilon_\pi\|^2 - \text{tr}\{\tilde{\mathbf{P}}(\omega)^\top \Omega_{e_\pi} \tilde{\mathbf{P}}(\omega)\} \right|}{\tilde{R}_\pi(\omega)} = o_p(1), \tag{14}$$

$$\sup_{\omega \in \mathcal{Z}_n} \frac{\left| \mu^\top \tilde{\mathbf{A}}(\omega)^\top \tilde{\mathbf{P}}(\omega)\varepsilon_\pi \right|}{\tilde{R}_\pi(\omega)} = o_p(1). \tag{15}$$

By Zhang *et al.* (2013) [14], under the condition (C4), we have

$$M_n \tilde{\zeta}_\pi^{-2G} \sum_{m=1}^{M_n} \left(\tilde{R}_\pi(\omega_m^0) \right)^G \xrightarrow{\text{a.s.}} 0.$$

We use conditions (C4), Chebyshev inequality and Theorem 2 of Whittle (1960) [18] to prove (14) and (15), respectively, as follows:

$$\begin{aligned} &P \left\{ \sup_{\omega \in \mathcal{Z}_n} \left| \|\tilde{\mathbf{P}}(\omega)\varepsilon_\pi\|^2 - \text{tr}\{\tilde{\mathbf{P}}(\omega)^\top \Omega_{e_\pi} \tilde{\mathbf{P}}(\omega)\} \right| / \tilde{R}_\pi(\omega) > \delta \mid \mathbf{X}, U \right\} \\ &\leq P \left\{ \sup_{\omega \in \mathcal{Z}_n} \sum_{t=1}^{M_n} \sum_{m=1}^{M_n} \omega_t \omega_m \left| \varepsilon_\pi^\top \tilde{\mathbf{P}}_{(t)}^\top \tilde{\mathbf{P}}_{(t)} \varepsilon_\pi - \text{tr}\{\tilde{\mathbf{P}}_{(t)}^\top \Omega_{e_\pi} \tilde{\mathbf{P}}_{(m)}\} \right| > \delta \tilde{\zeta}_\pi \mid \mathbf{X}, U \right\} \\ &\leq P \left\{ \max_{1 \leq t \leq M_n} \max_{1 \leq m \leq M_n} \left| \tilde{\mathbf{P}}_{(t)}^\top \tilde{\mathbf{P}}_{(t)} \varepsilon_\pi - \text{tr}\{\tilde{\mathbf{P}}_{(t)}^\top \Omega_{e_\pi} \tilde{\mathbf{P}}_{(m)}\} \right| > \delta \tilde{\zeta}_\pi \mid \mathbf{X}, U \right\} \\ &= P \left\{ \left| \varepsilon_\pi^\top \tilde{\mathbf{P}}(\omega_1^0)^\top \tilde{\mathbf{P}}(\omega_1^0) \varepsilon_\pi - \text{tr}\{\tilde{\mathbf{P}}(\omega_1^0)^\top \Omega_{e_\pi} \tilde{\mathbf{P}}(\omega_1^0)\} \right| > \delta \tilde{\zeta}_\pi \mid \mathbf{X}, U \right\} \\ &\quad \cup \left\{ \left| \varepsilon_\pi^\top \tilde{\mathbf{P}}(\omega_1^0)^\top \tilde{\mathbf{P}}(\omega_2^0) \varepsilon_\pi - \text{tr}\{\tilde{\mathbf{P}}(\omega_1^0)^\top \Omega_{e_\pi} \tilde{\mathbf{P}}(\omega_2^0)\} \right| > \delta \tilde{\zeta}_\pi \mid \mathbf{X}, U \right\} \\ &\quad \cup \dots \cup \left\{ \left| \varepsilon_\pi^\top \tilde{\mathbf{P}}(\omega_1^0)^\top \tilde{\mathbf{P}}(\omega_{M_n}^0) \varepsilon_\pi - \text{tr}\{\tilde{\mathbf{P}}(\omega_1^0)^\top \Omega_{e_\pi} \tilde{\mathbf{P}}(\omega_{M_n}^0)\} \right| > \delta \tilde{\zeta}_\pi \mid \mathbf{X}, U \right\} \\ &\quad \cup \left\{ \left| \varepsilon_\pi^\top \tilde{\mathbf{P}}(\omega_2^0)^\top \tilde{\mathbf{P}}(\omega_1^0) \varepsilon_\pi - \text{tr}\{\tilde{\mathbf{P}}(\omega_2^0)^\top \Omega_{e_\pi} \tilde{\mathbf{P}}(\omega_1^0)\} \right| > \delta \tilde{\zeta}_\pi \mid \mathbf{X}, U \right\} \\ &\quad \cup \left\{ \left| \varepsilon_\pi^\top \tilde{\mathbf{P}}(\omega_2^0)^\top \tilde{\mathbf{P}}(\omega_2^0) \varepsilon_\pi - \text{tr}\{\tilde{\mathbf{P}}(\omega_2^0)^\top \Omega_{e_\pi} \tilde{\mathbf{P}}(\omega_2^0)\} \right| > \delta \tilde{\zeta}_\pi \mid \mathbf{X}, U \right\} \end{aligned}$$

$$\begin{aligned}
 & \cup \dots \cup \left\{ \left| \varepsilon_\pi^T \tilde{\mathbf{P}}(\omega_2^0)^T \tilde{\mathbf{P}}(\omega_{M_n}^0) \varepsilon_\pi - \text{tr} \left\{ \tilde{\mathbf{P}}(\omega_2^0)^T \Omega_{e_\pi} \tilde{\mathbf{P}}(\omega_{M_n}^0) \right\} \right| > \delta \tilde{\xi}_\pi \mid \mathbf{X}, U \right\} \\
 & \cup \dots \cup \left\{ \left| \varepsilon_\pi^T \tilde{\mathbf{P}}(\omega_{M_n}^0)^T \tilde{\mathbf{P}}(\omega_{M_n}^0) \varepsilon_\pi - \text{tr} \left\{ \tilde{\mathbf{P}}(\omega_{M_n}^0)^T \Omega_{e_\pi} \tilde{\mathbf{P}}(\omega_{M_n}^0) \right\} \right| > \delta \tilde{\xi}_\pi \mid \mathbf{X}, U \right\} \\
 & \leq \sum_{t=1}^{M_n} \sum_{m=1}^{M_n} P \left\{ \left| \varepsilon_\pi^T \tilde{\mathbf{P}}(\omega_t^0)^T \tilde{\mathbf{P}}(\omega_m^0) \varepsilon_\pi - \text{tr} \left\{ \tilde{\mathbf{P}}(\omega_t^0)^T \Omega_{e_\pi} \tilde{\mathbf{P}}(\omega_m^0) \right\} \right| > \delta \tilde{\xi}_\pi \mid \mathbf{X}, U \right\} \\
 & \leq \sum_{t=1}^{M_n} \sum_{m=1}^{M_n} \frac{E \left(\left\| \varepsilon_\pi^T \tilde{\mathbf{P}}(\omega_t^0)^T \tilde{\mathbf{P}}(\omega_m^0) \varepsilon_\pi - \text{tr} \left\{ \tilde{\mathbf{P}}(\omega_t^0)^T \Omega_{e_\pi} \tilde{\mathbf{P}}(\omega_m^0) \right\} \right\|^2 \mid \mathbf{X}, U \right)}{\delta^2 \tilde{\xi}_\pi^2} \\
 & \leq c \delta^{-2G} \tilde{\xi}_\pi^{-2G} \sum_{t=1}^{M_n} \sum_{m=1}^{M_n} \left(\text{tr} \left\{ \tilde{\mathbf{P}}(\omega_t^0)^T \tilde{\mathbf{P}}(\omega_m^0) \right\} \right)^G \\
 & \leq c \delta^{-2G} \tilde{\xi}_\pi^{-2G} \sum_{t=1}^{M_n} \sum_{m=1}^{M_n} \left(\text{tr} \left\{ \tilde{\mathbf{P}}(\omega_m^0)^T \tilde{\mathbf{P}}(\omega_m^0) \right\} \right)^G \\
 & \leq c \delta^{-2G} \tilde{\xi}_\pi^{-2G} M_n \sum_{m=1}^{M_n} \left(\tilde{R}_\pi(\omega_m^0) \right)^G \xrightarrow{\text{a.s.}} 0.
 \end{aligned}$$

Therefore, we have

$$\begin{aligned}
 & P \left\{ \sup_{\omega \in \mathcal{F}_n} \left| \left\| \tilde{\mathbf{P}}(\omega) \varepsilon_\pi \right\|^2 - \text{tr} \left\{ \tilde{\mathbf{P}}(\omega)^T \Omega_{e_\pi} \tilde{\mathbf{P}}(\omega) \right\} \right| / \tilde{R}_\pi(\omega) > \delta \right\} \\
 & = E \left\{ P \left\{ \sup_{\omega \in \mathcal{F}_n} \left| \left\| \tilde{\mathbf{P}}(\omega) \varepsilon_\pi \right\|^2 - \text{tr} \left\{ \tilde{\mathbf{P}}(\omega)^T \Omega_{e_\pi} \tilde{\mathbf{P}}(\omega) \right\} \right| / \tilde{R}_\pi(\omega) > \delta \mid \mathbf{X}, U \right\} \right\} = o_p(1).
 \end{aligned}$$

(14) is proven. The proof process of (15) is similar to that of (14), and hence the proof of the Lemma 2.

Lemma 3 If (C1) - (C4) are satisfied, then

$$\sup_{\omega \in \mathcal{F}_n} \left| \frac{L_{\hat{\pi}}(\omega)}{R_\pi(\omega)} - 1 \right| = o_p(1). \tag{16}$$

Proof of Lemma 3: By Lemma 1 of Wei and Wang (2021) [12] and Cauchy-Schwarz inequality, we have

$$\begin{aligned}
 \left| \frac{L_{\hat{\pi}}(\omega)}{R_\pi(\omega)} - 1 \right| & \leq \left| \frac{L_\pi(\omega)}{R_\pi(\omega)} - 1 \right| + \frac{2 \{L_\pi(\omega)\}^{1/2} \cdot \|\hat{\mu}_\pi(\omega) - \hat{\mu}_{\hat{\pi}}(\omega)\|}{R_\pi(\omega)} \\
 & \quad + \frac{\|\hat{\mu}_\pi(\omega) - \hat{\mu}_{\hat{\pi}}(\omega)\|^2}{R_\pi(\omega)}, \\
 \|\hat{\mu}_\pi(\omega) - \hat{\mu}_{\hat{\pi}}(\omega)\|^2 & = \|\mathbf{P}(\omega) Z_\pi - \mathbf{P}(\omega) Z_{\hat{\pi}}\|^2 \\
 & \leq \left[\lambda_{\max} \{ \mathbf{P}(\omega) \} \right]^2 \|Z_\pi - Z_{\hat{\pi}}\|^2 \leq \|Z_\pi - Z_{\hat{\pi}}\|^2.
 \end{aligned}$$

Therefore, to prove (16), it suffices to show that

$$\sup_{\omega \in \mathcal{F}_n} \left| \frac{L_\pi(\omega)}{R_\pi(\omega)} - 1 \right| = o_p(1), \tag{17}$$

$$\sup_{\omega \in \mathcal{F}_n} \frac{\|Z_\pi - Z_{\hat{\pi}}\|^2}{R_\pi(\omega)} = o_p(1). \tag{18}$$

Note that

$$\begin{aligned} \left| \frac{L_\pi(\omega)}{R_\pi(\omega)} - 1 \right| &= \left| \frac{\|\mathbf{P}(\omega)\varepsilon_\pi\|^2 - \text{tr}\{\mathbf{P}(\omega)^2\Omega_{\varepsilon_\pi}\} - 2\mu^\top\mathbf{A}(\omega)\mathbf{P}(\omega)\varepsilon_\pi}{R_\pi(\omega)} \right| \\ &\leq \left| \frac{\|\mathbf{P}(\omega)\varepsilon_\pi\|^2 - \text{tr}\{\mathbf{P}(\omega)^2\Omega_{\varepsilon_\pi}\}}{R_\pi(\omega)} \right| + 2 \left| \frac{\mu^\top\mathbf{A}(\omega)\mathbf{P}(\omega)\varepsilon_\pi}{R_\pi(\omega)} \right|. \end{aligned}$$

Therefore, to prove (17), it suffices to show that

$$\sup_{\omega \in \mathcal{X}_n} \left| \frac{\mu^\top\mathbf{A}(\omega)\mathbf{P}(\omega)\varepsilon_\pi}{R_\pi(\omega)} \right| = o_p(1), \quad (19)$$

$$\sup_{\omega \in \mathcal{X}_n} \left| \frac{\|\mathbf{P}(\omega)\varepsilon_\pi\|^2 - \text{tr}\{\mathbf{P}(\omega)^2\Omega_{\varepsilon_\pi}\}}{R_\pi(\omega)} \right| = o_p(1). \quad (20)$$

The proof steps of (19) and (20) are similar to those of (14) and (15), respectively. The detailed process is omitted here, so it is established under (C4) condition, *i.e.* (17) is proven, and next prove (18).

According to (C1) and Cauchy-Schwarz inequality, we have

$$\begin{aligned} \sup_{\omega \in \mathcal{X}_n} \frac{\|Z_\pi - Z_{\hat{\pi}}\|^2}{R_\pi(\omega)} &\leq \xi_\pi^{-1} \sum_{i=1}^n \left\{ \frac{1}{\pi(\mathbf{X}_i, U_i; \hat{\alpha}_n)} - \frac{1}{\pi(\mathbf{X}_i, U_i)} \right\}^2 \cdot (\mu_i + \varepsilon_i)^2 \\ &\leq \xi_\pi^{-1} \left\{ \max_{1 \leq i \leq n} \left| \frac{1}{\pi(\mathbf{X}_i, U_i; \hat{\alpha}_n)} - \frac{1}{\pi(\mathbf{X}_i, U_i)} \right| \right\}^2 \cdot \sum_{i=1}^n (\mu_i + \varepsilon_i)^2 \\ &\leq \xi_\pi^{-1} \left\{ \sqrt{n} \cdot \max_{1 \leq i \leq n} \left| \frac{1}{\pi(\mathbf{X}_i, U_i; \hat{\alpha}_n)} - \frac{1}{\pi(\mathbf{X}_i, U_i)} \right| \right\}^2 \cdot c \cdot \left(\frac{1}{n} \mu^\top \mu + \frac{1}{n} \varepsilon^\top \varepsilon \right). \end{aligned}$$

According to (C3) and (C4), to prove (18), it suffices to prove that

$$(n\xi_\pi)^{-1} \varepsilon^\top \varepsilon = o_p(1), \quad (21)$$

$$\sqrt{n} \cdot \max_{1 \leq i \leq n} \left| \frac{1}{\pi(\mathbf{X}_i, U_i; \hat{\alpha}_n)} - \frac{1}{\pi(\mathbf{X}_i, U_i)} \right| = O_p(1). \quad (22)$$

By (C2) - (C4) and Markov inequality, for any $\forall \delta > 0$, we have

$$\begin{aligned} &P\left\{(n\xi_\pi)^{-1} \varepsilon^\top \varepsilon > \delta \mid \mathbf{X}, U\right\} \\ &\leq (n\xi_\pi \delta)^{-1} E\left\{\varepsilon^\top \varepsilon \mid \mathbf{X}, U\right\} \leq (n\xi_\pi \delta)^{-1} \text{tr}\{\mathbf{W}\} \\ &\leq \xi_\pi^{-1} \delta^{-1} \lambda_{\max}\{\Omega_{\varepsilon_\pi}\} \leq C_e \cdot \xi_\pi^{-1} \delta^{-1} \xrightarrow{\text{a.s.}} 0. \end{aligned}$$

This together with the dominated convergence theorem indicates $(n\xi_\pi)^{-1} \varepsilon^\top \varepsilon = o_p(1)$.

We perform Taylor expansion of $\{\pi(\mathbf{X}_i, U_i; \hat{\alpha}_n)\}^{-1}$ around the true value α_0 and then obtain

$$\begin{aligned} & \sqrt{n} \cdot \max_{1 \leq i \leq n} \left| \frac{1}{\pi(\mathbf{X}_i, U_i; \hat{\alpha}_n)} - \frac{1}{\pi(\mathbf{X}_i, U_i)} \right| \\ &= \sqrt{n} \cdot \max_{1 \leq i \leq n} \left| \frac{\left(\frac{\partial \pi(\mathbf{X}_i, U_i; \alpha)}{\partial \alpha} \Big|_{\alpha=\alpha_{n, X_i}} \right)^T}{\pi(\mathbf{X}_i, U_i, \hat{\alpha}_n) \cdot \pi(\mathbf{X}_i, U_i)} \cdot (\hat{\alpha}_n - \alpha_0) \right| \\ &\leq c \cdot \max_{1 \leq i \leq n} \left| \frac{\partial \pi(\mathbf{X}_i, U_i; \alpha)}{\partial \alpha} \Big|_{\alpha=\alpha_{n, X_i}} \right| \cdot \sqrt{n} \cdot \|\hat{\alpha}_n - \alpha_0\|, \end{aligned}$$

where the last inequality is due to (C1) and Cauchy-Schwarz inequality, and α_{n, X_i} is a linear combination between $\hat{\alpha}_n$ and α_0 . Since $\hat{\alpha}_n$ is MLE, by (C1) we have $\sqrt{n} \cdot \|\hat{\alpha}_n - \alpha_0\| = O_p(1)$. Because of the consistency of $\hat{\alpha}_n$ and (C1), we have $\max_{1 \leq i \leq n} \left| \frac{\partial \pi(\mathbf{X}_i, U_i; \alpha)}{\partial \alpha} \Big|_{\alpha=\alpha_{n, X_i}} \right| = O_p(1)$. These results imply (22), this completes the proof of (18) and hence the proof of the Lemma 3.

Lemma 4 If (C2) - (C4) are satisfied, then

$$\frac{\tilde{L}_\pi(\hat{\omega}_{\hat{\pi}})}{\inf_{\omega \in \mathcal{X}_n} \tilde{L}_\pi(\omega)} - 1 = o_p(1). \tag{23}$$

Proof of Lemma 4: After some careful calculations, we have

$$CV_{\hat{\pi}}(\omega) = \tilde{L}_\pi(\omega) + 2a_{\hat{\pi}}(\omega) + \|Z_{\hat{\pi}} - \mu\|^2,$$

where

$$\begin{aligned} a_{\hat{\pi}}(\omega) &= \varepsilon_\pi^T \tilde{\mathbf{A}}(\omega) \mu - \varepsilon_\pi^T \tilde{\mathbf{P}}(\omega) \varepsilon_\pi + \frac{1}{2} \|\tilde{\mu}_\pi(\omega) - \tilde{\mu}_{\hat{\pi}}(\omega)\|^2 \\ &\quad + (Z_{\hat{\pi}} - Z_\pi)^T \{ \mu - \tilde{\mu}_\pi(\omega) \} + (Z_{\hat{\pi}} - Z_\pi)^T \{ \tilde{\mu}_\pi(\omega) - \tilde{\mu}_{\hat{\pi}}(\omega) \} \\ &\quad + \{ \mu - \tilde{\mu}_\pi(\omega) \}^T \{ \tilde{\mu}_\pi(\omega) - \tilde{\mu}_{\hat{\pi}}(\omega) \} + \varepsilon_\pi^T \tilde{\mathbf{P}}(\omega) (Z_\pi - Z_{\hat{\pi}}). \end{aligned}$$

Because $\|Z_{\hat{\pi}} - \mu\|^2$ is independent of ω , then

$$\hat{\omega}_{\hat{\pi}} = \arg \min_{\omega \in \mathcal{X}_n} \{ \tilde{L}_\pi(\omega) + 2a_{\hat{\pi}}(\omega) \}.$$

By Lemma 1, we know that, to prove (23) it suffices to prove (13) and

$$\sup_{\omega \in \mathcal{X}_n} \frac{|a_{\hat{\pi}}(\omega)|}{\tilde{R}_\pi(\omega)} = o_p(1). \tag{24}$$

Since (13) has been proved, the following only need to prove (24). By Cauchy-Schwarz inequality, we have

$$\begin{aligned} |a_{\hat{\pi}}(\omega)| &\leq \left| \varepsilon_\pi^T \tilde{\mathbf{A}}(\omega) \mu \right| + \left| \varepsilon_\pi^T \tilde{\mathbf{P}}(\omega) \varepsilon_\pi \right| + \frac{1}{2} \|\tilde{\mu}_\pi(\omega) - \tilde{\mu}_{\hat{\pi}}(\omega)\|^2 \\ &\quad + \|Z_{\hat{\pi}} - Z_\pi\| \{ \tilde{L}_\pi(\omega) \}^{1/2} + \|Z_{\hat{\pi}} - Z_\pi\| \cdot \|\tilde{\mu}_\pi(\omega) - \tilde{\mu}_{\hat{\pi}}(\omega)\| \\ &\quad + \{ \tilde{L}_\pi(\omega) \}^{1/2} \|\tilde{\mu}_\pi(\omega) - \tilde{\mu}_{\hat{\pi}}(\omega)\| + \left\| \{ \tilde{\mathbf{P}}(\omega) \}^T \varepsilon_\pi \right\| \cdot \|Z_{\hat{\pi}} - Z_\pi\|. \end{aligned}$$

So to prove (24), it suffices to verify

$$\sup_{\omega \in \mathcal{X}_n} \frac{|\varepsilon_\pi^T \tilde{\mathbf{A}}(\omega) \mu|}{\tilde{R}_\pi(\omega)} = o_p(1), \tag{25}$$

$$\sup_{\omega \in \mathcal{X}_n} \frac{|\varepsilon_\pi^T \tilde{\mathbf{P}}(\omega) \varepsilon_\pi|}{\tilde{R}_\pi(\omega)} = o_p(1), \tag{26}$$

$$\sup_{\omega \in \mathcal{X}_n} \frac{\varepsilon_\pi^T \tilde{\mathbf{P}}(\omega) \{\tilde{\mathbf{P}}(\omega)\}^T \varepsilon_\pi}{\tilde{R}_\pi(\omega)} = o_p(1), \tag{27}$$

$$\sup_{\omega \in \mathcal{X}_n} \frac{\|Z_{\hat{\pi}} - Z_\pi\|^2}{\tilde{R}_\pi(\omega)} = o_p(1) \tag{28}$$

$$\sup_{\omega \in \mathcal{X}_n} \frac{\|\tilde{\mu}_{\hat{\pi}}(\omega) - \tilde{\mu}_\pi(\omega)\|^2}{\tilde{R}_\pi(\omega)} = o_p(1). \tag{29}$$

The proofs of (25), (26) and (27) are similar to (14), and hold under the condition (C4).

$$\begin{aligned} \sup_{\omega \in \mathcal{X}_n} \frac{\|Z_{\hat{\pi}} - Z_\pi\|^2}{\tilde{R}_\pi(\omega)} &= \sup_{\omega \in \mathcal{X}_n} \frac{\|Z_{\hat{\pi}} - Z_\pi\|^2}{R_\pi(\omega)} \left\{ \frac{R_\pi(\omega)}{\tilde{R}_\pi(\omega)} - 1 + 1 \right\} \\ &\leq \sup_{\omega \in \mathcal{X}_n} \frac{\|Z_{\hat{\pi}} - Z_\pi\|^2}{R_\pi(\omega)} \cdot \sup_{\omega \in \mathcal{X}_n} \left| \frac{R_\pi(\omega)}{\tilde{R}_\pi(\omega)} - 1 \right| + \sup_{\omega \in \mathcal{X}_n} \frac{\|Z_{\hat{\pi}} - Z_\pi\|^2}{R_\pi(\omega)}. \end{aligned} \tag{30}$$

By Zhang *et al.* (2013) [14], under (C2) - (C4), we have

$$\sup_{\omega \in \mathcal{X}_n} \left| \frac{R_\pi(\omega)}{\tilde{R}_\pi(\omega)} - 1 \right| = o_p(1),$$

which together with (18) implies (28), and next prove (29).

By Lemma 1 of Wei and Wang (2021) [12], we have

$$\begin{aligned} \|\tilde{\mu}_{\hat{\pi}}(\omega) - \tilde{\mu}_\pi(\omega)\|^2 &= \|\tilde{\mathbf{P}}(\omega)(Z_{\hat{\pi}} - Z_\pi)\|^2 \leq \lambda_{\max} \left\{ \tilde{\mathbf{P}}(\omega) \tilde{\mathbf{P}}(\omega)^T \right\} \|Z_{\hat{\pi}} - Z_\pi\|^2 \\ &\leq \left[\lambda_{\max} \left\{ \tilde{\mathbf{P}}(\omega) \right\} \right]^2 \|Z_{\hat{\pi}} - Z_\pi\|^2 \leq (1 + \tilde{p})^2 \|Z_{\hat{\pi}} - Z_\pi\|^2. \end{aligned} \tag{31}$$

By Zhang *et al.* (2013) [14], we know that, under (C2), $\tilde{p} \xrightarrow{\text{a.s.}} 0$. This together with (31) as well as (28) implies (29), and (25) - (29) together prove (24). This completes the proof of the Lemma 4.

Proof of Theorem 1: Firstly, we note that

$$\begin{aligned} \frac{L_{\hat{\pi}}(\hat{\omega}_{\hat{\pi}})}{\inf_{\omega \in \mathcal{X}_n} L_{\hat{\pi}}(\omega)} - 1 &= \sup_{\omega \in \mathcal{X}_n} \left\{ \frac{L_{\hat{\pi}}(\hat{\omega}_{\hat{\pi}})}{L_{\hat{\pi}}(\omega)} - 1 \right\} \\ &= \sup_{\omega \in \mathcal{X}_n} \left\{ \frac{L_{\hat{\pi}}(\hat{\omega}_{\hat{\pi}}) R_\pi(\hat{\omega}_{\hat{\pi}}) \tilde{R}_\pi(\hat{\omega}_{\hat{\pi}}) \tilde{L}_\pi(\hat{\omega}_{\hat{\pi}}) \tilde{L}_\pi(\omega) \tilde{R}_\pi(\omega) R_\pi(\omega)}{R_\pi(\hat{\omega}_{\hat{\pi}}) \tilde{R}_\pi(\hat{\omega}_{\hat{\pi}}) \tilde{L}_\pi(\hat{\omega}_{\hat{\pi}}) \tilde{L}_\pi(\omega) \tilde{R}_\pi(\omega) R_\pi(\omega) L_{\hat{\pi}}(\omega)} - 1 \right\} \\ &\leq \sup_{\omega \in \mathcal{X}_n} \left(\frac{L_{\hat{\pi}}(\omega)}{R_\pi(\omega)} \right) \sup_{\omega \in \mathcal{X}_n} \left(\frac{R_\pi(\omega)}{\tilde{R}_\pi(\omega)} \right) \sup_{\omega \in \mathcal{X}_n} \left(\frac{\tilde{R}_\pi(\omega)}{\tilde{L}_\pi(\omega)} \right) \sup_{\omega \in \mathcal{X}_n} \left(\frac{\tilde{L}_\pi(\omega)}{\tilde{R}_\pi(\omega)} \right) \\ &\quad \times \sup_{\omega \in \mathcal{X}_n} \left(\frac{\tilde{R}_\pi(\omega)}{R_\pi(\omega)} \right) \sup_{\omega \in \mathcal{X}_n} \left(\frac{R_\pi(\omega)}{L_{\hat{\pi}}(\omega)} \right) \frac{\tilde{L}_\pi(\hat{\omega}_{\hat{\pi}})}{\inf_{\omega \in \mathcal{X}_n} \tilde{L}_\pi(\omega)} - 1. \end{aligned}$$

Therefore, to prove Theorem 1, it suffices to prove that

$$\sup_{\omega \in \mathcal{X}_n} \left| \frac{R_\pi(\omega)}{\tilde{R}_\pi(\omega)} - 1 \right| = o_p(1), \quad (32)$$

$$\sup_{\omega \in \mathcal{X}_n} \left| \frac{\tilde{L}_\pi(\omega)}{\tilde{R}_\pi(\omega)} - 1 \right| = o_p(1), \quad (33)$$

$$\sup_{\omega \in \mathcal{X}_n} \left| \frac{L_{\hat{\pi}}(\omega)}{R_\pi(\omega)} - 1 \right| = o_p(1), \quad (34)$$

$$\frac{\tilde{L}_\pi(\hat{\omega}_{\hat{\pi}})}{\inf_{\omega \in \mathcal{X}_n} \tilde{L}_\pi(\omega)} - 1 = o_p(1). \quad (35)$$

According to Zhang *et al.* (2013) [14], we know that (32) holds under (C2) - (C4). Then according to Lemma 2, Lemma 3 and Lemma 4, we know that (33) - (35) hold, so the Theorem 1 is proved.

6. Conclusions

In this paper, we extend the JMA method to the nonparametric varying-coefficient models with response missing at random. Firstly, we use the inverse probability weighted method to deal with the missing data, then use B-spline to estimate the nonparametric functions, and finally use jackknife method to select weight ω . Under certain conditions, the asymptotic optimality of our method is proved.

In this paper, we only consider the varying-coefficient models. Undoubtedly, it is meaningful to extend our ideas to more complex models, such as partially linear models and semiparametric varying-coefficient models. However, this is a very challenging research. Second, for this article, we assume that the parametric model of the selection probability function is correctly specified, and further research is needed to develop a model averaging method that is robust against the misspecification of the selection probability function. These will be interesting future research directions.

Acknowledgements

The research is supported by NSF projects (ZR2021MA077) of Shandong Province of China.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

- [1] Hansen, B. (2007) Least Squares Model Averaging. *Econometrica*, **75**, 1175-1189. <https://doi.org/10.1111/j.1468-0262.2007.00785.x>
- [2] Zhang, X., Wan, A.T.K. and Zou, G. (2008) Least Squares Model Combining by Mallows Criterion. <https://doi.org/10.2139/ssrn.1272288>

- [3] Wan, A.T.K, Zhang, X. and Zou, G. (2010) Least Squares Model Averaging by Mallows Criterion. *Journal of Econometrics*, **156**, 277-283. <https://doi.org/10.1016/j.jeconom.2009.10.030>
- [4] Zhang, X., Yu, D., Zou, G. and 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. <https://doi.org/10.1080/01621459.2015.1115762>
- [5] Li, J., Xia, X., Wong, W.K. and Nott, D. (2018) Varying-Coefficient Semiparametric Model Averaging Prediction. *Biometrics*, **74**, 1417-1426. <https://doi.org/10.1111/biom.12904>
- [6] Zhang, X. and Wang, W. (2019) Optimal Model Averaging Estimation for Partially Linear Models. *Statistica Sinica*, **29**, 693-718. <https://doi.org/10.5705/ss.202015.0392>
- [7] Xia, X. (2021) Model Averaging Prediction for Nonparametric Varying-Coefficient Models with B-Spline Smoothing. *Statistical Papers*, **62**, 2885-2905. <https://doi.org/10.1007/s00362-020-01218-9>
- [8] Sun, Z., Su, Z. and Ma, J. (2014) Focused Vector Information Criterion Model Selection and Model Averaging Regression with Missing Response. *Metrika*, **77**, 415-432. <https://doi.org/10.1007/s00184-013-0446-8>
- [9] Zeng, J., Cheng, W., Hu, G. and Rong, Y. (2018) Model Averaging Procedure for Varying-Coefficient Partially Linear Models with Missing Responses. *Journal of the Korean Statistical Society*, **47**, 379-394. <https://doi.org/10.1016/j.jkss.2018.04.004>
- [10] Xie, J., Yan, X. and Tang, N. (2021) A Model-Averaging Method for High-Dimensional Regression with Missing Responses At Random. *Statistica Sinica*, **31**, 1005-1026. <https://doi.org/10.5705/ss.202018.0297>
- [11] Wei, Y., Wang, Q. and Liu, W. (2021) Model Averaging for Linear Models with Responses Missing at Random. *Annals of the Institute of Statistical Mathematics*, **73**, 535-553. <https://doi.org/10.1007/s10463-020-00759-y>
- [12] Wei, Y. and Wang, Q. (2021) Cross-Validation-Based Model Averaging in Linear Models with Response Missing at Random. *Statistics and Probability Letters*, **171**, 108990. <https://doi.org/10.1016/j.spl.2020.108990>
- [13] Hansen, B. and Racine, J. (2012) Jackknife Model Averaging. *Journal of Econometrics*, **167**, 38-46. <https://doi.org/10.1016/j.jeconom.2011.06.019>
- [14] Zhang, X., Wan, A.T.K. and Zou, G. (2013) Model Averaging by Jackknife Criterion in Models with Dependent Data. *Journal of Econometrics*, **174**, 82-94. <https://doi.org/10.1016/j.jeconom.2013.01.004>
- [15] De Door, C. (2001) A Practical Guide to Splines. Springer, New York.
- [16] Zhu, R., Wan, A.T.K., Zhang, X. and 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. <https://doi.org/10.1080/01621459.2018.1456936>
- [17] Fan, J., Ma, Y. and Dai, W. (2014) Nonparametric Independent Screening in Sparse Ultra-High Dimensional Varying Coefficient Models. *American Statistical Association*, **109**, 1270-1284. <https://doi.org/10.1080/01621459.2013.879828>
- [18] Whittle, P. (1960) Bounds for the Moments of Linear and Quadratic Forms in Independent Variables. *Theory of Probability & Its Applications*, **5**, 302-305. <https://doi.org/10.1137/1105028>