J Syst Sci Complex (2022) 35: 264–282

Model Averaging Estimation for Varying-Coefficient Single-Index Models*

LIU Yue · ZOU Jiahui · ZHAO Shangwei · YANG Qinglong

DOI: 10.1007/s11424-021-0158-5
Received: 15 July 2020 / Revised: 10 October 2020

Abstract The varying-coefficient single-index model (VCSIM) is widely used in economics, statistics and biology. A model averaging method for VCSIM based on a Mallows-type criterion is proposed to improve prodictive capacity, which allows the number of candidate models to diverge with sample size. Under model misspecification, the asymptotic optimality is derived in the sense of achieving the lowest possible squared errors. The authors compare the proposed model averaging method with several other classical model selection methods by simulations and the corresponding results show that the model averaging estimation has a outstanding performance. The authors also apply the method to a real dataset.

Keywords Asymptotic optimality, kernel-local smoothing method, Mallows-type criterion, model averaging, varying-coefficient single-index model.

1 Introduction

The varying-coefficient single-index model (VCSIM), a semiparametric model[1] has a powerful fitting ability for complex data since it is more flexible than parametric models. In particular, VCSIM has three important superiorities: Additivity of explanatory variables; coefficient function without stationary forms; and single-index threshold[1]. The additivity indicates the

LIU Yue
School of Statistics, Jiangxi University of Finance and Economics, Nanchang 330013*, China.*
ZOU Jiahui (Corresponding author)
School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing 100049*, China; Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing* 100190*, China.*
Email: zoujiahui16@mails.ucas.ac.cn.
ZHAO Shangwei
School of Science, Minzu University of China, Beijing 100081*, China.*
YANG Qinglong
School of Statistics and Mathematics, Zhongnan University of Economics and Law, Wuhan 430073*, China.*
*This research was supported by the National Nature Science Foundation of China, under Grant Nos. 12001559 and 11971324 and the Ministry of Education of Humanities and Social Science project, under Grant No. 19YJC910008.
◇*This paper was recommended for publication by Editor ZHANG Xinyu.*

effect of the marginal change of each explanatory variable[2], especially for modeling an economic system. The flexible coefficient function depicts more information from the data to avoid misspecified model forms well. Furthermore, a single-index threshold can overcome the "curse of dimensionality"[1, 3, 4]. When the dimension of the nonadditive explanatory variable typically in varying-coefficient models is large, it is subject to the "curse of dimensionality". A very large sample size is required to achieve the ideal estimation of the regression function. However, VCSIM assembles the multidimensional nonadditive explanatory variables into a "single-index term" in the varying-coefficient part to effectively avoid it. Moreover, there exist many mature methods to estimate parameters in VCSIM, such as semiparametric least square estimation[5, 6], averaging derivatives estimation[7, 8], sliced inverse regression estimation[9], smoothing splines[10–12], kernel smoothing[13, 14], local polynomial smoothing[15–17] and penalized splines[18].

Recently, prediction with multiple models has become increasingly prevalent such as ensemble learning in machine learning[19–21] and model averaging (MA) in statistics[22, 23]. Traditionally, the following problems will arise when we only use one model. First, we do not know which variables are admitted into the model and whether the admitted variables act as additive parts or single-index variables in practice for VCSIM. Second, we do not know whether the model form we used is true. A methodology to solve these problems is model selection by minimizing a criterion, such as AIC[24], BIC[25] and C_p[26]. However, the working model we choose through model selection approaches is likely to change when different training data are used, especially when sample size is small. Moreover, we will encounter the phenomenon whereby the criterion values of two or more candidate models are very similar and simply choosing one model may not be satisfactory. Hence, the aforementioned issues have given rise to model averaging in statistics and ensemble learning in machine learning, which are regarded as a continuous-type model selection. Compared with machine learning, the model averaging methodology has more theoretical support and is better suited for statistical models. Therefore, to improve the ability of VCSIM, we adopt a Mallows-type model averaging approach inspired by prior works, such as Hansen[22],Wan, et al.[27] and Zhu, et al.[28].

In the past decade, model averaging has become a hot topic in statistics. Bayesian model averaging (BMA) and frequentist model averaging (FMA) are two main streams of averaging techniques. Although BMA can be applied in many models since its flexibility, the choice of priors is often challenging[29]. Meanwhile, there are various FMA methods[30], and a partial list includes smoothed information criteria[31, 32], optimal weighting[33–35], adaptive weighting[36, 37], plug-in methods[38, 39] and marginal regression averaging[40]. We adopt the optimal model averaging method here.

We make contributions to VCSIM and model averaging mainly in two aspects. First, we propose a Mallows-type criterion to calculate the weights for VCSIM. Under this weight choice strategy, we also prove the model averaging estimator is asymptotically optimal in terms of minimum loss. Second, we allow the number of candidate models to diverge with the sample size. Therefore, it is more flexible in the preparation of candidate models.

The rest of the paper is organized as follows. The process for model averaging is introduced

in Section 2. In Section 3, the asymptotic optimality of the MA estimator is presented. Monte Carlo simulation is performed to investigate the performance of the MA estimator in Section 4, and we apply the MA method to a real dataset in Section 5. In addition, we offer some concluding remarks in Section 6. Finally, the technical proofs are provided in the Appendix.

2 Model Setup and Model Averaging Frame

This paper considers the varying-coefficient single-index model shown as

$$y_i = \mu_i + \varepsilon_i = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{g}(\boldsymbol{z}_i^{\mathrm{T}} \boldsymbol{\beta}) + \varepsilon_i, \quad i = 1,\ 2,\ \cdots,\ n, \tag{2.1}$$

where y_i is the response variable; $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})^{\mathrm{T}} \in \mathbb{R}^{p\times 1}$ and $\boldsymbol{z}_i = (z_{i1}, z_{i2}, \cdots, z_{iq})^{\mathrm{T}} \in \mathbb{R}^{q\times 1}$ are nonrandom explanatory variable vectors; $\boldsymbol{g}(\cdot) = (g_1(\cdot), g_2(\cdot), \cdots, g_p(\cdot))^{\mathrm{T}} \in \mathbb{R}^{p\times 1}$ is an unknown vector function mapping from $\mathbb{R}$ to $\mathbb{R}^p$; the random error ε_i satisfies $\mathbb{E}(\varepsilon_i|\boldsymbol{x}_i, \boldsymbol{z}_i) = 0$ and $\mathbb{E}(\varepsilon_i^2|\boldsymbol{x}_i, \boldsymbol{z}_i) = \sigma^2$; and $\boldsymbol{\beta}$ is a $q \times 1$ unknown parameter vector that satisfies $\|\boldsymbol{\beta}\| = 1$ and whose first component of $\boldsymbol{\beta}$ is positive for the model identifiability. For convenience in subsequent discussions, denote $\boldsymbol{y} = (y_1, y_2, \cdots, y_n)^{\mathrm{T}}$, $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n)^{\mathrm{T}}$, $\boldsymbol{Z} = (\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_n)^{\mathrm{T}}$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n)^{\mathrm{T}}$ and $\boldsymbol{\mu} = (\mu_1, \mu_2, \cdots, \mu_n)^{\mathrm{T}}$.

Before using the model averaging method, candidate models should be prepared. Suppose that there are s_n candidate varying-coefficient single-index models used to approximate the model (2.1), with s_n being allowed to diverge to infinity as $n \to \infty$. Among all candidate models, the sth one is set as

$$y_i = \mu_{(s),i} + \varepsilon_{(s),i} = \boldsymbol{x}_{(s),i}^{\mathrm{T}} \boldsymbol{g}_{(s)}(\boldsymbol{z}_{(s),i}^{\mathrm{T}} \boldsymbol{\beta}_{(s)}) + \varepsilon_{(s),i}, \qquad i = 1, 2, \cdots, n, \tag{2.2}$$

where $\boldsymbol{x}_{(s),i} \in \mathbb{R}^{p_s\times 1}$ and $\boldsymbol{z}_{(s),i} \in \mathbb{R}^{q_s\times 1}$ are the corresponding subset of $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$, respectively; $\boldsymbol{g}_{(s)}(\cdot) \in \mathbb{R}^{p_s\times 1}$ is the unknown vector function; $\boldsymbol{\beta}_{(s)} \in \mathbb{R}^{q_s\times 1}$ is the unknown parameter vector, which also satisfies $\|\boldsymbol{\beta}_{(s)}\| = 1$ and the first component of $\boldsymbol{\beta}_{(s)}$ is positive. In particular, it should be noted that $\varepsilon_{(s),i} = \varepsilon_i + \mu_i - \mu_{(s),i}$ is not from the identical distribution as ε. Finally, we also use $\boldsymbol{X}_{(s)}$, $\boldsymbol{Z}_{(s)}$, $\boldsymbol{\varepsilon}_{(s)}$ and $\boldsymbol{\mu}_{(s)}$ for the s^{th} candidate model as above.

Next, we will launch the MA estimation by the following two main steps.

Step 1 Fit every candidate model and obtain the estimator of $\boldsymbol{\mu}_{(s)}$ denoted as $\widehat{\boldsymbol{\mu}}_{(s)}$.

Step 2 Evaluate the weight for every candidate model and combine all the candidate models' estimators to achieve the MA estimator. Hence, the MA estimator for $\boldsymbol{\mu}$ has the form $\widehat{\boldsymbol{\mu}}(\boldsymbol{w}) = \sum_{s=1}^{s_n} w_s \widehat{\boldsymbol{\mu}}_{(s)}$, where $\boldsymbol{w} \in \mathcal{W} = \{[0,1]^{s_n} : \sum_{i=1}^{s_n} = 1\}$.

Specifically, there are three tasks to complete in Step 1. First, to obtain $\widehat{\boldsymbol{\beta}}_{(s)}$, we adopt the modified average derivative method inspired by Hristache, et al.[41], in which the derivative of $\boldsymbol{g}(\cdot)$ is derived from kernel-local polynomial smoothing method[42]. Second, regarding $\boldsymbol{z}_{(s),i}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_{(s)}$ as a scalar variable, we can obtain the estimator of $\boldsymbol{g}_{(s)}(\cdot)$, $\widehat{\boldsymbol{g}}_{(s)}(\cdot)$, based on the local constant least square method. Third, $\widehat{\boldsymbol{\mu}}_{(s),i}$, the estimator of the ith component of $\boldsymbol{\mu}_{(s)}$, is obtained by multiplying $\boldsymbol{x}_{(s),i}$ and $\widehat{\boldsymbol{g}}_{(s)}(\boldsymbol{z}_{(s),i}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_{(s)})$. In Step 2, we compute the weight vector $\widehat{\boldsymbol{w}} = (\widehat{w}_1, \widehat{w}_2, \cdots, \widehat{w}_{s_n})^{\mathrm{T}}$ through a Mallows-type criterion and combine the weights and the estimators from all candidate models to achieve the final MA estimator $\widehat{\boldsymbol{\mu}}(\widehat{\boldsymbol{w}})$.

2.1 Estimating for Each Candidate Model

It is complicated to estimate $\boldsymbol{\mu}$ for each candidate model, and there are three tasks to complete. First, to estimate $\boldsymbol{\beta}_{(s)}$, we should regard $g_{(s),j}(\boldsymbol{z}^{\mathrm{T}}\boldsymbol{\beta}_{(s)})$, the j^{th} component of $\boldsymbol{g}_{(s)}(\boldsymbol{z}^{\mathrm{T}}\boldsymbol{\beta}_{(s)})$, as a function with respect to $\boldsymbol{z}$ and denote it as $\varphi_{(s),j}(\boldsymbol{z})$, for $j = 1, 2, \cdots, p_s$. Accordingly, the derivative of $\varphi_{(s),j}(\boldsymbol{z})$ with respect to $\boldsymbol{z}$ can be written as $\dot{\varphi}_{(s),j}(\boldsymbol{z})$ which is a $q_s \times 1$ vector. For $s = 1, 2, \cdots, s_n$ and $\boldsymbol{z} \in \mathbb{R}^{q_s}$, denote $\boldsymbol{\varphi}_{(s)}(\boldsymbol{z}) = (\,\varphi_{(s),1}(\boldsymbol{z}),\ \varphi_{(s),2}(\boldsymbol{z}),\ \cdots,\ \varphi_{(s),p_s}(\boldsymbol{z})\,)^{\mathrm{T}} \in \mathbb{R}^{p_s\times 1}$, $\dot{\boldsymbol{\varphi}}_{(s)}(\boldsymbol{z}) = (\,\dot{\varphi}_{(s),1}(\boldsymbol{z}),\ \dot{\varphi}_{(s),2}(\boldsymbol{z}),\ \cdots,\ \dot{\varphi}_{(s),p_s}(\boldsymbol{z})\,)^{\mathrm{T}} \in \mathbb{R}^{p_s\times q_s}$. Then, $\boldsymbol{\beta}_{(s)}$ can be computed by

$$\boldsymbol{\beta}_{(s)} = \frac{\dot{\boldsymbol{\varphi}}_{(s),j}(\boldsymbol{z})}{\|\dot{\boldsymbol{\varphi}}_{(s),j}(\boldsymbol{z})\|},$$

which is identical for each j in theory.

By the denotation above, we rewrite the model (2.2) as

$$y_i = \boldsymbol{x}_{(s),i}^{\mathrm{T}}\boldsymbol{\varphi}_{(s)}(\boldsymbol{z}_{(s),i}) + \varepsilon_i, \quad i = 1, 2, \cdots, n, \quad s = 1, 2, \cdots, s_n. \tag{2.3}$$

By the kernel-local polynomial smoothing method, $\widehat{\boldsymbol{\varphi}}_{(s)}(\boldsymbol{z})$ and $\widehat{\dot{\boldsymbol{\varphi}}}_{(s)}(\boldsymbol{z})$, the estimators of $\boldsymbol{\varphi}_{(s)}(\boldsymbol{z})$ and $\dot{\boldsymbol{\varphi}}_{(s)}(\boldsymbol{z})$, are the solution to minimize

$$\begin{aligned} D(\boldsymbol{a}, \boldsymbol{B}) &= \sum_{i=1}^{n}\{y_i - \boldsymbol{x}_{(s),i}^{\mathrm{T}}\boldsymbol{a} - \boldsymbol{x}_{(s),i}^{\mathrm{T}}\boldsymbol{B}(\boldsymbol{z}_{(s),i} - \boldsymbol{z})\}^2 K_h(\|\boldsymbol{z}_{(s),i} - \boldsymbol{z}\|) \\ &= \left\| \boldsymbol{K}_{(s),h}^{1/2}(\boldsymbol{z})\boldsymbol{y} - \boldsymbol{K}_{(s),h}^{1/2}(\boldsymbol{z})\Gamma_s(\boldsymbol{z}) \begin{pmatrix} \boldsymbol{a} \\ \boldsymbol{b} \end{pmatrix} \right\|^2, \end{aligned} \tag{2.4}$$

where $\widehat{\boldsymbol{\varphi}}_{(s)}(\boldsymbol{z})$ and $\widehat{\dot{\boldsymbol{\varphi}}}_{(s)}(\boldsymbol{z})$ correspond to $\boldsymbol{a} \in \mathbb{R}^{p_s}$ and $\boldsymbol{B} \in \mathbb{R}^{p_s\times q_s}$, $\boldsymbol{b}$ is a $p_s q_s \times 1$ vector connecting all q_s-dimensional row vectors of $\boldsymbol{B}$ to a column, $K_h(\cdot) = \frac{1}{h}K(\cdot/h)$ with $h > 0$ called bandwidth and

$$\begin{aligned} \boldsymbol{K}_{(s),h}^{1/2}(\boldsymbol{z}) &= \mathrm{diag}\Big(\sqrt{K_h(\boldsymbol{z}_{(s),1} - \boldsymbol{z})},\ \sqrt{K_h(\boldsymbol{z}_{(s),2} - \boldsymbol{z})},\ \cdots,\ \sqrt{K_h(\boldsymbol{z}_{(s),n} - \boldsymbol{z})}\ \Big), \\ \Gamma_s(\boldsymbol{z}) &= \Big[\ \big\{\boldsymbol{x}_{(s),1}^{\mathrm{T}},\ \boldsymbol{x}_{(s),1}^{\mathrm{T}} \otimes (\,\boldsymbol{z}_{(s),1} - \boldsymbol{z})^{\mathrm{T}}\big\}^{\mathrm{T}},\ \big\{\,\boldsymbol{x}_{(s),2}^{\mathrm{T}},\ \boldsymbol{x}_{(s),2}^{\mathrm{T}} \otimes (\,\boldsymbol{z}_{(s),2} - \boldsymbol{z})^{\mathrm{T}}\,\big\}^{\mathrm{T}},\ \cdots, \\ &\qquad \big\{\boldsymbol{x}_{(s),n}^{\mathrm{T}}, \boldsymbol{x}_{(s),n}^{\mathrm{T}} \otimes (\,\boldsymbol{z}_{(s),n} - \boldsymbol{z})^{\mathrm{T}}\,\big\}^{\mathrm{T}}\ \Big]^{\mathrm{T}}. \end{aligned}$$

Fortunately, the optimal solution has an analytical form

$$\begin{pmatrix} \widehat{\boldsymbol{a}} \\ \widehat{\boldsymbol{b}} \end{pmatrix} = \big[\Gamma_s^{\mathrm{T}}(\boldsymbol{z})\boldsymbol{K}_{(s),h}(\boldsymbol{z})\Gamma_s(\boldsymbol{z})\big]^{-1}\Gamma_s^{\mathrm{T}}(\boldsymbol{z})\boldsymbol{K}_{(s),h}(\boldsymbol{z})\boldsymbol{y}.$$

For $j = 1, 2, \cdots, p_s$, the elements from $[(j-1)q_s + 1]^{\mathrm{th}}$ to jq_s^{th} in $\widehat{\boldsymbol{b}}$ construct a $q_s \times 1$ vector, i.e., $\widehat{\dot{\boldsymbol{\varphi}}}_{(s),j}(\boldsymbol{z})$. Finally, we can obtain the estimator of $\boldsymbol{\beta}_{(s)}$ based on $\widehat{\dot{\boldsymbol{\varphi}}}_{(s),j}(\boldsymbol{z}_{(s)})$ by

$$\widehat{\boldsymbol{\beta}}_{(s),ij} = \frac{\widehat{\dot{\boldsymbol{\varphi}}}_{(s),j}(\boldsymbol{z}_{(s),i})}{\|\widehat{\dot{\boldsymbol{\varphi}}}_{(s),j}(\boldsymbol{z}_{(s),i})\|} \quad \text{and} \quad \widehat{\boldsymbol{\beta}}_{(s)} = \frac{1}{np_s}\sum_{i=1}^{n}\sum_{j=1}^{p_s}\widehat{\boldsymbol{\beta}}_{(s),ij}. \tag{2.5}$$

As shown in Hristache, et al.[41], when $q_s > 4$, the convergence rate of $\widehat{\boldsymbol{\beta}}_{(s)}$ acquired in this way would be much worse. Therefore, to improve the estimation, we suggest refining $\widehat{\dot{\boldsymbol{\varphi}}}_{(s)}(\boldsymbol{z})$ by an iterative algorithm with elliptic windows, which are inspired by Hristache, et al.[41].

Second, we estimate the function $\boldsymbol{g}_{(s)}(\cdot)$ for each candidate model. By the kernel-local constant least squares method, $\widehat{\boldsymbol{g}}_{(s)}(v)$ is obtained as the solution by minimizing

$$\begin{aligned} J(\boldsymbol{a}; \widehat{\boldsymbol{\beta}}_{(s)}) &= \sum_{i=1}^{n} \{y_i - \boldsymbol{x}_{(s),i}^{\mathrm{T}} \boldsymbol{a}\}^2 K_h(\boldsymbol{z}_{(s),i}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_{(s)} - v) \\ &= \left\| \boldsymbol{F}_{(s),h}^{1/2}(v) \boldsymbol{y} - \boldsymbol{F}_{(s),h}^{1/2}(v) \boldsymbol{X}_{(s)} \boldsymbol{a} \right\|^2, \end{aligned}$$

where $\boldsymbol{a}$ is a $p_s \times 1$ vector, $\widehat{\boldsymbol{\beta}}_{(s)}$ is computed in the first process and

$$\boldsymbol{F}_{(s),h}^{1/2}(v) = \mathrm{diag}\Big(\sqrt{K_h(\boldsymbol{z}_{(s),1}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_{(s)} - v)}, \sqrt{K_h(\boldsymbol{z}_{(s),2}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_{(s)} - v)}, \cdots, \sqrt{K_h(\boldsymbol{z}_{(s),n}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_{(s)} - v)} \Big).$$

Then, we achieve an analytical solution for $\widehat{\boldsymbol{g}}_{(s)}(v)$,

$$\widehat{\boldsymbol{g}}_{(s)}(v) = \left\{ \boldsymbol{X}_{(s)}^{\mathrm{T}} \boldsymbol{F}_{(s),h}(v) \boldsymbol{X}_{(s)} \right\}^{-1} \boldsymbol{X}_{(s)} \boldsymbol{F}_{(s),h}(v) \boldsymbol{y}.$$

Third, combine the former two processes to estimate $\boldsymbol{\mu}_{(s)}$ for every candidate model. Recall that $\boldsymbol{\mu}_{(s)} = \Big(\boldsymbol{x}_{(s),1}^{\mathrm{T}} \boldsymbol{g}_{(s)}(\boldsymbol{z}_{(s),1}^{\mathrm{T}} \boldsymbol{\beta}_{(s)}),\ \boldsymbol{x}_{(s),2}^{\mathrm{T}} \boldsymbol{g}_{(s)}(\boldsymbol{z}_{(s),2}^{\mathrm{T}} \boldsymbol{\beta}_{(s)}), \cdots, \boldsymbol{x}_{(s),n}^{\mathrm{T}} \boldsymbol{g}_{(s)}(\boldsymbol{z}_{(s),n}^{\mathrm{T}} \boldsymbol{\beta}_{(s)}) \Big)^{\mathrm{T}}$, so the estimator of $\boldsymbol{\mu}_{(s)}$ can be written as

$$\begin{aligned} \widehat{\boldsymbol{\mu}}_{(s)} &= \Big(\boldsymbol{x}_{(s),1}^{\mathrm{T}} \widehat{\boldsymbol{g}}_{(s)}(\boldsymbol{z}_{(s),1}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_{(s)}),\ \boldsymbol{x}_{(s),2}^{\mathrm{T}} \widehat{\boldsymbol{g}}_{(s)}(\boldsymbol{z}_{(s),2}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_{(s)}),\ \cdots,\ \boldsymbol{x}_{(s),n}^{\mathrm{T}} \widehat{\boldsymbol{g}}_{(s)}(\boldsymbol{z}_{(s),n}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_{(s)}) \Big)^{\mathrm{T}} \\ &= \boldsymbol{X}_{(s)} \Big[\Big\{ \boldsymbol{X}_{(s)}^{\mathrm{T}} \boldsymbol{F}_{(s),h} \left(\boldsymbol{z}_{(s),1}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_{(s)} \right) \boldsymbol{X}_{(s)} \Big\}^{-1} \boldsymbol{X}_{(s)}^{\mathrm{T}} \boldsymbol{F}_{(s),h} \left(\boldsymbol{z}_{(s),1}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_{(s)} \right), \cdots, \\ &\qquad \Big\{ \boldsymbol{X}_{(s)}^{\mathrm{T}} \boldsymbol{F}_{(s),h} \left(\boldsymbol{z}_{(s),n}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_{(s)} \right) \boldsymbol{X}_{(s)} \Big\}^{-1} \boldsymbol{X}_{(s)}^{\mathrm{T}} \boldsymbol{F}_{(s),h} \left(\boldsymbol{z}_{(s),n}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_{(s)} \right) \Big] \boldsymbol{y} \\ &\equiv \boldsymbol{P}_{(s)}(\widehat{\boldsymbol{\beta}}_{(s)}) \boldsymbol{y}. \end{aligned} \tag{2.6}$$

2.2 Weight Choice Criterion for MA Estimation

After obtaining $\widehat{\boldsymbol{\beta}}_{(s)}$ from each candidate model, we can express the MA estimation for $\boldsymbol{\mu}$ as

$$\widehat{\boldsymbol{\mu}}(\boldsymbol{w}) = \sum_{s=1}^{s_n} w_s \widehat{\boldsymbol{\mu}}_{(s)} = \sum_{s=1}^{s_n} w_s \boldsymbol{P}_{(s)}(\widehat{\boldsymbol{\beta}}_{(s)}) \boldsymbol{y} = \mathcal{P}(\boldsymbol{w}) \boldsymbol{y}, \tag{2.7}$$

where

$$\mathcal{P}(\boldsymbol{w}) = \sum_{s=1}^{s_n} w_s \boldsymbol{P}_{(s)}(\widehat{\boldsymbol{\beta}}_{(s)}). \tag{2.8}$$

Using quadratic loss, we can assess the effectiveness of the MA estimator by

$$L_n(\boldsymbol{w}) = \|\widehat{\boldsymbol{\mu}}(\boldsymbol{w}) - \boldsymbol{\mu}\|^2.$$

However, $\boldsymbol{\mu}$ is unknown in $L_n(\boldsymbol{w})$, so we cannot calculate $\widehat{\boldsymbol{w}}$ by minimizing $L_n(\boldsymbol{w})$ directly. Thanks to the idea of Mallows-type criteria, we attempt to find an unbiased or an asymptotic unbiased estimator for the risk $R_n(\boldsymbol{w}) = \mathbb{E}\{L_n(\boldsymbol{w}) | \boldsymbol{X}, \boldsymbol{Z}\}$.

Let $\mathcal{P}^*(\boldsymbol{w}) = \sum_{s=1}^{s_n} w_s \boldsymbol{P}_{(s)}(\boldsymbol{\beta}^*_{(s)})$, $\widehat{\boldsymbol{\mu}}^* = \mathcal{P}^*(\boldsymbol{w})\boldsymbol{y}$, $L_n^*(\boldsymbol{w}) = \|\widehat{\boldsymbol{\mu}}^*(\boldsymbol{w}) - \boldsymbol{\mu}\|^2$ and $C_n^*(\boldsymbol{w}) = \|\mathcal{P}^*(\boldsymbol{w})\boldsymbol{y} - \boldsymbol{y}\|^2 + 2\sigma^2 \mathrm{tr}\big\{\mathcal{P}^*(\boldsymbol{w})\big\}$. It is seen that

$$\mathbb{E}\{C_n^*(\boldsymbol{w})|\boldsymbol{X}, \boldsymbol{Z}\} = \mathbb{E}\{L_n^*(\boldsymbol{w})|\boldsymbol{X}, \boldsymbol{Z}\} + n\sigma^2. \tag{2.9}$$

Consider that according to Corollary 1 in Xie and Li[1], under regular assumptions for any $s = 1, 2, \cdots, s_n$, there exists a $\boldsymbol{\beta}^*_{(s)}$ satisfying

$$\widehat{\boldsymbol{\beta}}_{(s)} \xrightarrow{\mathrm{P}} \boldsymbol{\beta}^*_{(s)}.$$

Simultaneously, under uniformly integrable conditions, we have

$$\mathbb{E}\{C_n(\boldsymbol{w})|\boldsymbol{X}, \boldsymbol{Z}\} - \mathbb{E}\{C_n^*(\boldsymbol{w})|\boldsymbol{X}, \boldsymbol{Z}\} \to 0, \tag{2.10}$$

$$\mathbb{E}\{L_n(\boldsymbol{w})|\boldsymbol{X}, \boldsymbol{Z}\} - \mathbb{E}\{L_n^*(\boldsymbol{w})|\boldsymbol{X}, \boldsymbol{Z}\} \to 0, \tag{2.11}$$

where $C_n(\boldsymbol{w}) = \|\mathcal{P}(\boldsymbol{w})\boldsymbol{y} - \boldsymbol{y}\|^2 + 2\sigma^2 \mathrm{tr}\big\{\mathcal{P}(\boldsymbol{w})\big\}$. Then, combining (2.9), (2.10) and (2.11), we have

$$\mathbb{E}\{C_n(\boldsymbol{w})|\boldsymbol{X}, \boldsymbol{Z}\} - \mathbb{E}\{L_n(\boldsymbol{w})|\boldsymbol{X}, \boldsymbol{Z}\} - n\sigma^2 \to 0, \tag{2.12}$$

which indicates that $C_n(\boldsymbol{w})$ is an asymptotic unbiased Mallow-type criterion for risk $\mathbb{E}\{L_n(\boldsymbol{w})|\boldsymbol{X}, \boldsymbol{Z}\}$. Finally, we can achieve the optimal weight vector by

$$\widehat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}\in\mathcal{W}} C_n(\boldsymbol{w}). \tag{2.13}$$

3 Asymptotic Optimality

The following regular conditions are required for the asymptotic optimality of the proposed MA estimator. Unless stated otherwise, all limiting processes below correspond to $n \to \infty$.

Condition 1 Suppose that for any $s = 1, 2, \cdots, s_n$ and $i = 1, 2, \cdots, q_s$, there exists a $\beta^*_{(s),i}$ satisfying

$$\widehat{\beta}_{(s),i} \xrightarrow{\mathrm{P}} \beta^*_{(s),i}. \tag{3.1}$$

This condition is a common solution in nonparametric or semiparametric models and usually holds when some regular assumptions are provided, such as Xia and Li[1]. Under this condition, (2.10) and (2.11) will be valid. Therefore, the aim of this condition is to guarantee that $C_n(\boldsymbol{w})$ is an asymptotic unbiased estimator of risk $R_n(\boldsymbol{w})$.

Condition 2 There exists a constant $\kappa_1 > 0$ such that

$$\limsup_{n\to\infty} \max_{1\le s\le s_n} \overline{\tau}\{\boldsymbol{P}_{(s)}(\boldsymbol{\beta}^*_{(s)})\} \le \kappa_1, \tag{3.2}$$

where $\overline{\tau}(\boldsymbol{M})$ denotes the largest singular value of matrix $\boldsymbol{M}$.

This condition states that the largest singular value of $\boldsymbol{P}_{(s)}(\boldsymbol{\beta}^*_{(s)})$ is boundary, and a similar condition is used in (A.3) and (A.4) of Hansen and Racine[43] as well as (8) and (9) of Zhang, et al.[44].

Define $R_n^*(\boldsymbol{w}) = \mathbb{E}\{L_n(\boldsymbol{w}, \mathcal{B}^*)|\boldsymbol{X},\ \boldsymbol{Z}\}$ and the infimum of $R_n^*(\boldsymbol{w})$ in set $\mathcal{W}$ as $\xi_n^* = \inf_{\boldsymbol{w}\in\mathcal{W}} R_n^*(\boldsymbol{w})$.

Condition 3 There exist a constant $\kappa_2 > 0$ and a fixed integer $G > 1$ such that

$$\max_{1\le i\le n} \mathbb{E}(\varepsilon_i^{4G}|\boldsymbol{x}_i, \boldsymbol{z}_i) \le \kappa_2 < \infty, \tag{3.3}$$

$$s_n \xi_n^{*-2G} \sum_{s=1}^{s_n} \{R_n^*(\boldsymbol{w}_s^0)\}^G \to 0 \tag{3.4}$$

and

$$\|\boldsymbol{\mu}\|^2 = O(n), \tag{3.5}$$

where $\boldsymbol{w}_s^0 \in \mathcal{W}$ is the weight vector with the sth component being 1 and all others being 0.

This is a common condition widely used in the literature on MA, such as Wan, et a.[27], Hansen and Racine[43], Zhang, et al.[44]. The equality (3.3) provides a constraint on the conditional expectation of ε_i, $i = 1, 2, \cdots, n$ and controls the noise in the data generating process. The equality (3.4) is the same as (13) in Zhang, et al.[44], and it is required for proving the asymptotic optimality of the Mallows model averaging estimator. Considering that s_n is allowed to be infinite, (3.4) offers a restriction on the order of ξ_n^* and s_n to diverge. Moreover, (3.5) is easy to satisfy if $|\mu_i| < \infty$ for $i = 1, 2, \cdots, n$.

Condition 4

$$n\rho_n \xi_n^{*-1} = o_p(1), \tag{3.6}$$

where $\rho_n = \max_{1\le s\le s_n} \overline{\tau}\left\{\boldsymbol{P}_{(s)}(\widehat{\boldsymbol{\beta}}_{(s)}) - \boldsymbol{P}_{(s)}(\boldsymbol{\beta}_{(s)}^*)\right\}$.

From Condition 1, it is easy to bring out $\rho_n \to 0$ because of the continuous of singular values. Furthermore, Condition 4 gives a stricter relation between ρ_n and ξ_n.

Theorem 3.1 *Under Conditions* 1–4, *we have*

$$\frac{L_n(\widehat{\boldsymbol{w}})}{\inf_{\boldsymbol{w}\in\mathcal{W}} L_n(\boldsymbol{w})} \xrightarrow{\mathrm{P}} 1. \tag{3.7}$$

This theorem shows that the MA estimator of $\boldsymbol{\mu}$ is asymptotically optimal in the sense that it will approach the "best estimator" under quadratic loss. The equality (3.7) is a measure widely used to evaluate the estimator in model averaging and model selection papers. In particular, this theorem allows candidate models to be misspecified in any form, unlike the local misspecification framework that restricts the degree of misspecification to decay as n increases[45].

4 Monte Carlo simulation

In this section, we will compare the performance of the following six estimators: AIC model selection method, see [24]; BIC model selection method, see [25, 31]; smoothed AIC model averaging method (SAIC), see [23, 45–47]; smoothed BIC model averaging method (SBIC), see [45–47]; Mallows model selection method (C_p), see [22, 48]; and Mallows model averaging

method (MMA), see [22, 27]. The AIC and BIC information criteria of the sth candidate model, see [49], can be computed as

$$\mathrm{AIC}_s = \log(\widehat{\sigma}^2_{(s)}) + 2n^{-1}\mathrm{tr}\{\boldsymbol{P}_{(s)}(\widehat{\boldsymbol{\beta}}_{(s)})\}, \tag{4.1}$$

$$\mathrm{BIC}_s = \log(\widehat{\sigma}^2_{(s)}) + n^{-1}\log(n)\mathrm{tr}\{\boldsymbol{P}_{(s)}(\widehat{\boldsymbol{\beta}}_{(s)})\}, \tag{4.2}$$

where $\widehat{\sigma}^2_{(s)} = n^{-1}\|\boldsymbol{y} - \boldsymbol{P}_{(s)}(\widehat{\boldsymbol{\beta}}_{(s)})\boldsymbol{y}\|$. In addition, the Mallows C_p criterion is shown as

$$\mathrm{MC}_s = \|\boldsymbol{P}_{(s)}(\widehat{\boldsymbol{\beta}}_{(s)})\boldsymbol{y} - \boldsymbol{y}\|^2 + 2\widehat{\sigma}^2_{(s)}\mathrm{tr}\{\boldsymbol{P}_{(s)}(\widehat{\boldsymbol{\beta}}_{(s)})\}. \tag{4.3}$$

For model selection approaches, the best candidate model is selected by minimizing these criteria. Furthermore, the SAIC model averaging estimation is obtained by assigning the weight to the sth candidate model as

$$w_s = \exp(-\mathrm{AIC}_s/2)\Big/\sum_{s=1}^{s_n}\exp(-\mathrm{AIC}_s/2), \quad s = 1, 2, \cdots, s_n, \tag{4.4}$$

and the SBIC model averaging estimation is

$$w_s = \exp(-\mathrm{BIC}_s/2)\Big/\sum_{s=1}^{s_n}\exp(-\mathrm{BIC}_s/2), \quad s = 1, 2, \cdots, s_n. \tag{4.5}$$

4.1 Simulation Designs

We attempt to determine the performance of the varying-coefficient single-index model in practice by simulation experiments. First, we generate $\{y_i\}$ from data generating process:

$$\begin{aligned} y_i &= \mu_i + \varepsilon_i \\ &= \sqrt{2\theta}\exp\left(\sum_{l=1}^{200} z_{il}\sqrt{2\phi}l^{-\phi-\frac{1}{2}}\right) + \sum_{j=1}^{200} x_{ij}\sqrt{2\theta}j^{-\theta-\frac{1}{2}}\exp\left(\sum_{l=1}^{200} z_{il}\sqrt{2\phi}l^{-\phi-\frac{1}{2}}\right) + \varepsilon_i, \end{aligned}$$

where $\{x_{ij}\}$ and $\{z_{il}\}$ are from an independent identical distribution (i.i.d.) $N(0,1)$; $\{\varepsilon_i\}$ is a random disturbance term from $N(0,\eta^2)$; the tunning parameters θ and ϕ are usually selected in $(0,1)$ and here they are set at 0.1 and 0.25 respectively. We alter η to compel $R^2 = \mathrm{var}(\mu_i)/\mathrm{var}(y_i)$ to vary between 0.1 and 0.9. Furthermore, taking the form of the model (2.1), the data generating process above can be reexpressed as

$$y_i = (1, \boldsymbol{x}_i^{\mathrm{T}})\boldsymbol{g}(\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\beta}) + \varepsilon_i,$$

where $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{i200})^{\mathrm{T}}$, $\boldsymbol{z}_i = (z_{i1}, z_{i2}, \cdots, z_{i200})^{\mathrm{T}}$ and $\boldsymbol{g}(\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\beta}) = \big(\boldsymbol{g}_0(\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\beta}), \boldsymbol{g}_1(\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\beta}), \cdots, \boldsymbol{g}_{200}(\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\beta})\big)^{\mathrm{T}}$ with

$$\boldsymbol{g}_0(\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\beta}) = \sqrt{2\theta}\exp\left(\sum_{l=1}^{200} z_{il}\sqrt{2\phi}l^{-\phi-\frac{1}{2}}\right)$$

and for $j = 1, 2, \cdots, 200$,

$$\boldsymbol{g}_j(\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\beta}) = \sqrt{2\theta}j^{-\theta-\frac{1}{2}}\exp\left(\sum_{l=1}^{200} z_{il}\sqrt{2\phi}l^{-\phi-\frac{1}{2}}\right).$$

In practice, we cannot collect all components of $\boldsymbol{x}_i$ or $\boldsymbol{z}_i$, and usually just the front terms can be observed. Here, we allow the number of the components observed, p_n, to grow with the sample size n. Let n be selected among 50, 100, 200 and 400; moreover, p_n is 14, 18 and 22 by $p_n = \text{round}(3n^{1/3})$, where round$(x)$ means the rounding of x. Therefore, all candidate models are misspecified. Next, we build a candidate model set in a nested manner like Hansen[22] so that s_n, the number of candidate models, is correspondingly determined to be 11, 14, 18 or 22. To estimate VCSIM, we use the method described in Subsection 2.1 and the bandwidth calculated by $0.75n^{-1/5}(\log n)^{-1/6}$ which is also used in Yu, et al.[50]. In addition, to eliminate the randomness of simulation, we perform $D = 100$ replications to check the performance of every method.

Finally, to evaluate the performance of these methods, we calculate the normalized mean square error (NMSE) through D repetitions,

$$\text{NMSE} = D^{-1}\sum_{d=1}^{D}\frac{\text{MSE}^{(d)}}{\text{MSE}^{(d)}_{\min}}, \tag{4.6}$$

where $\text{MSE}^{(d)} = n^{-1}\|\widehat{\boldsymbol{\mu}}^{(d)} - \boldsymbol{\mu}^{(d)}\|$, is the mean square error (MSE) of the estimator for the dth replication, and $\text{MSE}^{(d)}_{\min}$ represents the minimal MSE of all candidate models.

4.2 Simulation Results

The NMSE results are displayed in Figure 1 and the other results generated by different

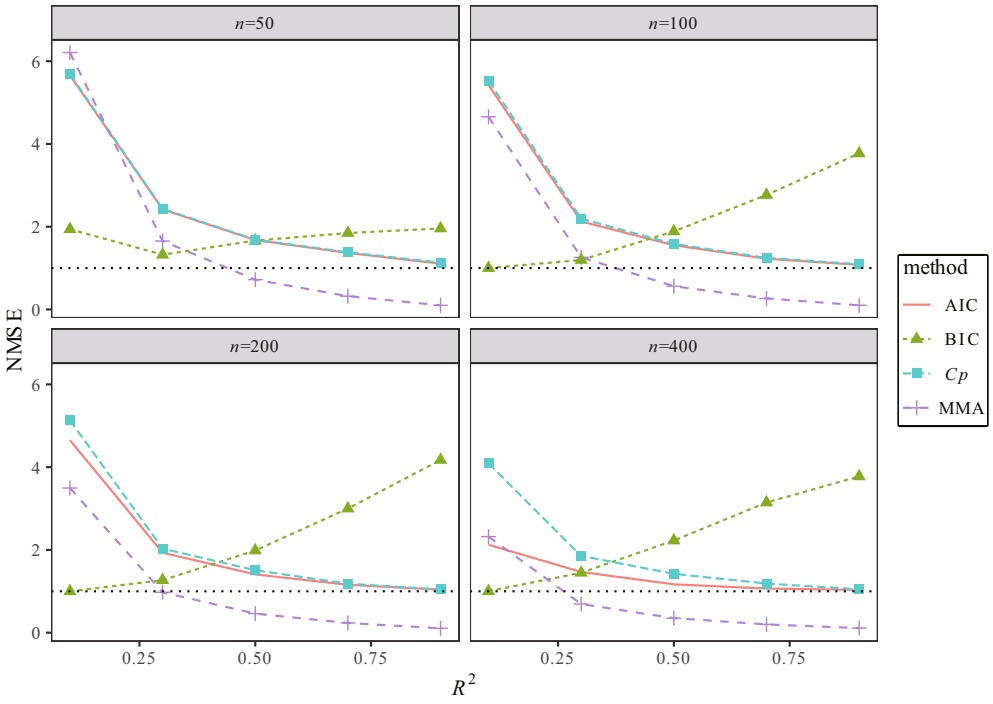

Figure 1 The NMSE of each method; the dotted black line denotes NMSE = 1 as the benchmark

θ and ϕ in $(0,1)$ are similar in our experiments. In each subfigure, the y axis means NMSE described in (4.6) and the x axis represents R^2, whose values are $0.1, 0.3, 0.5, 0.7$ and 0.9. From the results, AIC and Mallows C_p selection methods have quite similar NMSE, but AIC becomes better when n is large, and this phenomenon corresponds to Hansen[22]. Compared with the AIC method, BIC performs better when R^2 is small, but the opposite is true when R^2 is large. In particular, the turning point will move to the left as the sample size n grows. The MMA method performs best when R^2 is big, and the critical value of R^2 also moves to left when n is large. As described in Section 3, the new criterion we proposed is asymptotic unbiased estimation for the risk. So, when n is large enough, the random effects will be neglected and the good weights for every candidate model will be derived by this criterion.

5 Empirical Application

In this section, we run the six methods on a real dataset that contains observations on aged patients in 36 nursing homes in San Diego, CA, collected between 1980 and 1982. The same data were used by Xie, et al.[51] and Zhu, et al.[28], can be downloaded at http://www.stats.ox.ac.uk/pub/datasets/csb/. In this dataset, the response variable of interest, y, is the natural logarithm of the number of days the patient stayed in a nursing home. There are three indictor covariates, x_1, x_2 and x_3. x_1 is equal to 1 if the patient received medical treatment at the nursing home and 0 otherwise. x_2 conveys the gender information, equal to 1(0) if the patient is male (female), and x_3 equals 1(0) if the patient is married (not married), as an indicator for marital status. x_4 is a class variable that represents health status, with a larger x_4 indicating worse health conditions. $t = (\text{age} - 64)/(102 - 64)$, continuous but bounded, is the normalized age of the patients in the sample, with a range from 65 to 102. The original sample contains 1601 observations, including 332 censored observations, but our analysis is based on the remaining 1269 uncensored observations.

We place x_1, x_2 and x_3 in the coefficient part and the remaining x_4 and t into the covariate part because the indictor variables are suited to magnify the effect of covariates but not to work directly. Then, the candidate model set is constructed such that no fewer than one covariate exists in each part. Therefore, we can obtain $M_n = (3^2-1)\times(2^2-1) = 24$ candidate models. To evaluate the predictive performance, we divide the full sample into training set and testing set at random. The sample size of the training set, n_0, is selected from $\lfloor 70\%n \rfloor$, $\lfloor 75\%n \rfloor$, $\lfloor 80\%n \rfloor$, $\lfloor 85\%n \rfloor$ and $\lfloor 90\%n \rfloor$, where n denotes the full sample size, 1269. The remaining $n_{\text{test}} = n - n_0$ observations are used as the testing set. In operation, we first compute the weights from the training set and then obtain the estimate of $\boldsymbol{y}_{\text{test}}$ in the testing set, $\widehat{\boldsymbol{y}}$. Here, we also repeat $D = 100$ cycles to eliminate the random effects. Finally, we also define normalized mean squared prediction error (NMSPE) to evaluate the estimator for each method:

$$\text{MSPE}^{(d)} = n_{\text{test}}^{-1} \left\| \widehat{\boldsymbol{y}}^{(d)} - \boldsymbol{y}_{\text{test}}^{(d)} \right\|^2, \tag{5.1}$$

$$\text{NMSPE} = D^{-1} \sum_{d=1}^{D} \frac{\text{MSPE}^{(d)}}{\text{MSPE}_{\min}^{(d)}}, \tag{5.2}$$

where $\boldsymbol{y}_{\text{test}}^{(d)}$ means the response variable of the testing set in the dth repetition $\boldsymbol{y}^{(d)}$ is the corresponding estimation of $\boldsymbol{y}^{(d)}$, and $\text{MSPE}_{\text{min}}^{(d)}$ is the minimal $\text{MSPE}^{(d)}$ among all candidate models in the dth replication.

The numerical results are shown in Table 1 and Figure 2. The MMA method gains the

Table 1 The NMSPE of six methods in real data ($D = 100$)

n_0	Method	MMA	AIC	BIC	C_p
888 (70%)	Mean	**1.009**	1.024	1.030	1.022
	Median	**1.007**	1.021	1.029	1.018
	SD	**0.010**	0.014	0.012	0.015
951 (75%)	Mean	**1.010**	1.025	1.031	1.023
	Median	**1.008**	1.023	1.031	1.020
	SD	**0.011**	0.016	0.014	0.016
1015 (80%)	Mean	**1.012**	1.028	1.034	1.026
	Median	**1.012**	1.025	1.034	1.025
	SD	**0.012**	0.017	0.014	0.018
1078 (85%)	Mean	**1.018**	1.034	1.039	1.034
	Median	**1.016**	1.030	1.037	1.030
	SD	**0.015**	0.021	0.021	0.020
1142 (90%)	Mean	**1.026**	1.044	1.047	1.040
	Median	**1.024**	1.039	1.043	1.037
	SD	**0.019**	0.028	0.024	0.029

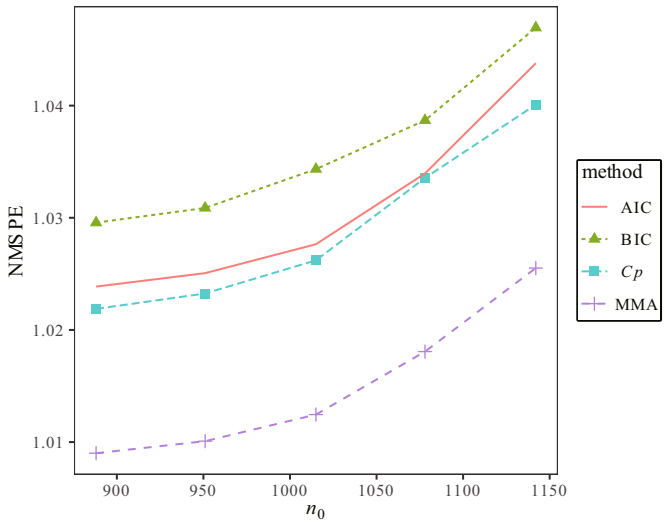

Figure 2 The NMSPE of each method in real data

lowest mean and median of NMSPE, with the median being notably lower than those of the alternatives. However, the AIC, BIC and Mallows C_p methods perform similarly but worse than the model averaging method. Note that the Mallows C_p performed best in the model selection methods. This is due to the Mallows criteria intending to diminish the forecast error by minimizing its unbiased estimation.

6 Conclusion

In this paper, we propose a Mallows-type model averaging method for VCSIM. Our averaging method allows the number of candidate models to be divergent with sample size. We also demonstrate the estimator achieved by our propose MA method contains the asymptotic optimality in the sense of achieving the lowest possible squared errors. In addition, there are at least three directions to extend this work. First, model screening could be introduced before model averaging. After removing poorly performing models, the MA estimator may be more efficient for prediction. Second, we do not allow the dimension of explanatory variables to grow with the sample size. Removing this restriction may be an interesting but difficult. Third, this paper only considers averaging among VCSIMs. Challenging work remains on averaging both parametric and semi/nonparametric models.

References

[1] Xia Y and Li W K, On single-index coefficient regression models, *Journal of the American Statistical Association*, 1999, **94**(448): 1275–1285.

[2] Härdle W, Müller M, Sperlich S, et al., *Nonparametric and Semiparametric Models*, Springer-Verlag, Berlin, 2004.

[3] Wong H, Wai C I, and Zhang R, Varying-coefficient single-index model, *Computational Statistics & Data Analysis*, 2008, **52**(3): 1458–1476.

[4] Zhao Y, Xue L, and Feng S, Semiparametric estimation of the single-index varying-coefficient model, *Communications in Statistics — Theory and Methods*, 2016, **46**(9): 4311–4326.

[5] Ichimura H, Semiparametric least squares and weighted sls estimation of single-index models, *Journal of Econometrics*, 1993, **58**(1): 71–120.

[6] Härdle W, Hall P, and Ichimura H, Optimal smoothing in single-index models, *The Annals of Statistics*, 1993, **21**(1): 157–178.

[7] Stoker T M, Consistent estimation of scaled coefficients, *Econometrica*, 1986, **54**(6): 1461–1481.

[8] Härdle W and Stoker T M, Investigating smooth multiple regression by the method of average derivatives, *Journal of the American Statistical Association*, 1989, **84**(408): 986–995.

[9] Li K C, Sliced inverse regression for dimension reduction, *Journal of the American Statistical Association*, 1991, **86**(414): 316–327.

[10] Wahba G, Smoothing noisy data with spline function, *Numerische Mathematik*, 1975, **24**: 383–393.

[11] Wahba G, Bayesian confidence-intervals for the cross-validated smoothing spline, *Journal of the Royal Statistical Society, Series B,* 1983, **45**(1): 133–150.

[12] Wahba G, A comparison of gcv and gml for choosing the smoothing parameter in the generalized spline smoothing problem, *The Annals of Mathematical Statistics*, 1985, **13**(4): 1378–1402.

[13] Nadaraya E A, On estimating regression, *Theory Probability and Its Application*, 1964, **9**(1): 141–142.

[14] Watson G S, Smooth regression analysis, *Sankhya: The Indian Journal of Statistics, Series A*, 1964, **26**(4): 359–372.

[15] Fan J, Design adaptive nonparametric regression, *Journal of the American Statistical Association*, 1991, **87**(420): 998–1004.

[16] Fan J, Local linear regression smoothers and their minimax efficiency, *The Annals of Statistics*, 1993, **21**(1): 196–216.

[17] Fan J and Gijbels I, *Local Polynomial Modeling and Its Application*, Chapman and Hall, London, 1996.

[18] Yu Y and Ruppert D, Penalized spline estimation for partially linear single-index models, *Journal of the American Statistical Association*, 2002, **97**(460): 1042–1054.

[19] Zhou Z H, When semi-supervised learning meets ensemble learning, *Frontiers of Electrical & Electronic Engineering in China*, 2011, **6**(1): 6–16.

[20] Fersini E, Messina E, and Pozzi F A, Sentiment analysis: Bayesian ensemble learning, *Decision Support Systems*, 2014, **68**: 26–38.

[21] Liu B, Wang S, Ren L, et al., irspot-el: Identify recombination spots with an ensemble learning approach, *Bioinformatics*, 2017, **33**(1): 35–41.

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

[23] Zhang X, Yu D, Zou G, et al., Optimal model averaging estimation for generalized linear models and generalized linear mixed-effects models, *Journal of the American Statistical Association*, 2016, **111**(516): 1775–1790.

[24] Akaike H, Information theory and an extension of the maximum likelihood principle, *Breakthroughs in Statistics*, 1973, **1**: 610–624.

[25] Schwarz G, Estimating the dimension of a model, *The Annals of Statistics*, 1978, **6**(2): 461–464.

[26] Mallows C L, Some comments on c_p, *Technometrics*, 1973, **42**(1): 87–94.

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

[28] Zhu R, Wan A T K, Zhang X, et al., A mallows-type model averaging estimator for the varying-coefficient partially linear model, *Journal of the American Statistical Association*, 2019, **114**(526): 882–892.

[29] Hoeting J A, Madigan D, Raftery A E, et al., Bayesian model averaging: A tutorial, *Statistical Science*, 1999, **14**: 382–417.

[30] Wang H, Zhang X, and Zou G, Frequentist model averaging estimation: A review, *Journal of Systems Science and Complexity*, 2009, **22**(4): 732–748.

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

[32] Claeskens G, Croux C, and van K J, Variable selection for logistic regression using a prediction-focused information criterion, *Biometrics*, 2006, **62**: 972–979.

[33] Liang H, Zou G, Wan A T K, et al., Optimal weight choice for frequentist model average estimators, *Journal of the American Statistical Association*, 2011, **106**: 1053–1066.

[34] Zhao Z H and Zou G H, Average estimation of semiparametric models for high-dimensional lon-

gitudinal data, *Journal of Systems Science and Complexity*, 2020, **33**(6): 2013–2047.

[35] Zhang X, Zheng Y, and Wang S, A demand forecasting method based on stochastic frontier analysis and model average: An application in air travel demand forecasting, *Journal of Systems Science and Complexity*, 2019, **32**(4): 615–633.

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

[37] Zhang X, Lu Z, and Zou G, Adaptively combined forecasting for discrete response time series, *Journal of Econometrics*, 2013, **176**: 80–91.

[38] Liu C, Distribution theory of the least squares averaging estimator, *Journal of Econometrics*, 2015, **186**: 142–159.

[39] Yin S, Liu C, and Lin C, Focused information criterion and model averaging for large panels with a multifactor error structure, *Journal of Business & Economic Statistics*, 2019, forthcoming.

[40] Li D, Linton O, and Lu Z, A flexible semiparametric forecasting model for time series, *Journal of Econometrics*, 2015, **187**: 345–357.

[41] Hristache M, Juditsky A, and Spokoiny V, Direct estimation of the index coefficient in a single-index model, *The Annals of Statistics*, 2001, **29**(3): 595-623.

[42] Fan J and Zhang W, Statistical estimation in varying coefficient models, *The Annals of Statistics*, 1999, **27**(5): 1491–1518.

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

[44] Zhang X, Wan A T K, and Zou G, Model averaging by jackknife criterion in models with dependent data, *Journal of Econometrics*, 2013, **174**(2): 82–94.

[45] Hjort N L and Claeskens G, Frequentist model average estimators, *Journal of the American Statistical Association*, 2003, **98**(464): 879–899.

[46] Hjort N L and Claeskens G, Rejoinder to the discussion of "frequentist model average estimators" and "focused information criterion", *Journal of the American Statistical Association*, 2003, **98**(464): 938–945.

[47] Hjort N L and Claeskens G, Focused information criteria and model averaging for the cox hazard regression model, *Journal of the American Statistical Association*, 2006, **101**(476): 1449–1464.

[48] Li K C, Asymptotic optimality for c_p, c_l, cross-validation and generalized cross-validation: Discrete dindex set, *The Annals of Statistics*, 1987, **15**(3): 958–975.

[49] Li C, Li Q, Racine J S, et al., Optimal model averaging of varying coefficient models, *Statistica Sinica*, 2018, **28**(4): 2795–2809.

[50] Yu Z, He B, and Chen M, Empirical likelihood for generalized partially linear single-index models, *Communications in Statistics*, 2014, **43**: 4156–4163.

[51] Xie S, Wan A T K, and Zhou Y, Quantile regression methods with varying-coefficient models for censored data, *Computational Statistics & Data Analysis*, 2015, **88**: 154–172.

[52] Gao Y, Zhang X, Wang S, et al., Frequentist model averaging for threshold models, *Annals of the Institute of Statistical Mathematics*, 2018, **71**(2): 275–306.

[53] Zhang X, Model averaging and its applications, PhD thesis, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, 2010.

Appendix

Proving Theorem 3.1

Proof It is seen that

$$C_n(\boldsymbol{w}) = L_n(\boldsymbol{w}) + \eta_n(\boldsymbol{w}) + \|\varepsilon\|^2, \tag{A.1}$$

where

$$\begin{aligned}\eta_n(\boldsymbol{w}) &= -2\varepsilon^{\mathrm{T}}\{\mathcal{P}(\boldsymbol{w})(\boldsymbol{\mu}+\varepsilon) - \boldsymbol{\mu}\} + 2\sigma^2\mathrm{tr}\{\mathcal{P}(\boldsymbol{w})\} \\ &= 2\varepsilon^{\mathrm{T}}\left[\boldsymbol{I} - \mathcal{P}^*(\boldsymbol{w}) - \{\mathcal{P}(\boldsymbol{w}) - \mathcal{P}^*(\boldsymbol{w})\}\right]\boldsymbol{\mu} \\ &\quad - 2\varepsilon^{\mathrm{T}}\left[\mathcal{P}^*(\boldsymbol{w}) + \{\mathcal{P}(\boldsymbol{w}) - \mathcal{P}^*(\boldsymbol{w})\}\right]\varepsilon \\ &\quad + 2\sigma^2\mathrm{tr}\left[\mathcal{P}^*(\boldsymbol{w}) + \{\mathcal{P}(\boldsymbol{w}) - \mathcal{P}^*(\boldsymbol{w})\}\right].\end{aligned} \tag{A.2}$$

Further, we have

$$L_n(\boldsymbol{w}) = L_n^*(\boldsymbol{w}) + \zeta_n(\boldsymbol{w}), \tag{A.3}$$

where

$$L_n^*(\boldsymbol{w}) = \|\widehat{\boldsymbol{\mu}}^*(\boldsymbol{w}) - \boldsymbol{\mu}\|^2, \tag{A.4}$$

$$\zeta_n(\boldsymbol{w}) = \|\widehat{\boldsymbol{\mu}}(\boldsymbol{w}) - \widehat{\boldsymbol{\mu}}^*(\boldsymbol{w})\|^2 + 2\left(\widehat{\boldsymbol{\mu}}^*(\boldsymbol{w}) - \boldsymbol{\mu}\right)^{\mathrm{T}}\left(\widehat{\boldsymbol{\mu}}(\boldsymbol{w}) - \widehat{\boldsymbol{\mu}}^*(\boldsymbol{w})\right). \tag{A.5}$$

As we know $\|\varepsilon\|^2$ is independent of $\boldsymbol{w}$, then

$$\begin{aligned}\widehat{\boldsymbol{w}} &= \arg\min_{\boldsymbol{w}} C_n(\boldsymbol{w}) \\ &= \arg\min_{\boldsymbol{w}}\{L_n^*(\boldsymbol{w}) + \zeta_n(\boldsymbol{w}) + \eta_n(\boldsymbol{w})\}.\end{aligned} \tag{A.6}$$

As for $\inf_{\boldsymbol{w}\in\mathcal{W}} L_n(\boldsymbol{w})$, by the definition of infimum, there exist two sequence, the positive sequence $\{\vartheta_n\}$ and the weight vector sequence $\{\boldsymbol{w}(n)\}$ such that

$$\lim_{n\to\infty}\vartheta_n = 0 \tag{A.7}$$

and

$$\inf_{\boldsymbol{w}\in\mathcal{W}} L_n(\boldsymbol{w}) = L_n(\boldsymbol{w}(n)) - \vartheta_n. \tag{A.8}$$

Next, note that $R_n^*(\boldsymbol{w}) = \mathbb{E}(L_n^*(\boldsymbol{w})|\boldsymbol{X}, \boldsymbol{Z})$. The for any $\delta > 0$, we have

$$\begin{aligned}
&\Pr\left\{\left|\frac{\inf_{\boldsymbol{w}\in\mathcal{W}} L_n(\boldsymbol{w})}{L_n(\widehat{\boldsymbol{w}})} - 1\right| > \delta\right\} \\
=&\Pr\left\{\left|\frac{\min_{\boldsymbol{w}\in\mathcal{W}}\{L_n(\boldsymbol{w}) + \eta_n(\boldsymbol{w})\} - \eta_n(\widehat{\boldsymbol{w}}) - \inf_{\boldsymbol{w}\in\mathcal{W}} L_n(\boldsymbol{w})}{L_n(\widehat{\boldsymbol{w}})}\right| > \delta\right\} \\
\le&\Pr\left\{\left|\frac{L_n(\boldsymbol{w}(n)) + \eta_n(\boldsymbol{w}(n)) - \eta_n(\widehat{\boldsymbol{w}}) - (L_n(\boldsymbol{w}(n)) - \vartheta_n)}{L_n(\widehat{\boldsymbol{w}})}\right| > \delta\right\} \\
\le&\Pr\left\{\frac{|\eta_n(\boldsymbol{w}(n))|}{\inf_{\boldsymbol{w}\in\mathcal{W}} L_n(\boldsymbol{w})} + \frac{|\eta_n(\widehat{\boldsymbol{w}})|}{L_n^*(\widehat{\boldsymbol{w}}) + \zeta_n(\widehat{\boldsymbol{w}})} + \frac{\vartheta_n}{L_n^*(\widehat{\boldsymbol{w}}) + \zeta_n(\widehat{\boldsymbol{w}})} > \delta\right\} \\
\le&\Pr\left[\sup_{\boldsymbol{w}\in\mathcal{W}} \frac{|\eta_n(\boldsymbol{w})|}{R_n^*(\boldsymbol{w})}\left\{\inf_{\boldsymbol{w}\in\mathcal{W}} \frac{|L_n^*(\boldsymbol{w}) + \zeta_n(\boldsymbol{w}) - \vartheta_n|}{R_n^*(\boldsymbol{w})}\right\}^{-1} > \frac{\delta}{3}\right] \\
&+\Pr\left[\sup_{\boldsymbol{w}\in\mathcal{W}} \frac{|\eta_n(\boldsymbol{w})|}{R_n^*(\boldsymbol{w})}\left\{\inf_{\boldsymbol{w}\in\mathcal{W}} \frac{L_n^*(\boldsymbol{w}) + \zeta_n(\boldsymbol{w})}{R_n^*(\boldsymbol{w})}\right\}^{-1} > \frac{\delta}{3}\right] \\
&+\Pr\left[\frac{\vartheta_n}{\inf_{\boldsymbol{w}\in\mathcal{W}} R_n^*(\boldsymbol{w})}\left\{\inf_{\boldsymbol{w}\in\mathcal{W}} \frac{L_n^*(\boldsymbol{w}) + \zeta_n(\boldsymbol{w})}{R_n^*(\boldsymbol{w})}\right\}^{-1} > \frac{\delta}{3}\right].
\end{aligned}$$

So to prove (3.7), we only need to verify the three formulas in the following, as Gao, et al.[52] and Zhang[53]:

$$\sup_{\boldsymbol{w}\in\mathcal{W}} \frac{|\eta_n(\boldsymbol{w})|}{R_n^*(\boldsymbol{w})}\left\{\inf_{\boldsymbol{w}\in\mathcal{W}} \frac{|L_n^*(\boldsymbol{w}) + \zeta_n(\boldsymbol{w}) - \vartheta_n|}{R_n^*(\boldsymbol{w})}\right\}^{-1} = o_p(1), \tag{A.9}$$

$$\sup_{\boldsymbol{w}\in\mathcal{W}} \frac{|\eta_n(\boldsymbol{w})|}{R_n^*(\boldsymbol{w})}\left\{\inf_{\boldsymbol{w}\in\mathcal{W}} \frac{L_n^*(\boldsymbol{w}) + \zeta_n(\boldsymbol{w})}{R_n^*(\boldsymbol{w})}\right\}^{-1} = o_p(1), \tag{A.10}$$

and

$$\frac{\vartheta_n}{\inf_{\boldsymbol{w}\in\mathcal{W}} R_n^*(\boldsymbol{w})}\left\{\inf_{\boldsymbol{w}\in\mathcal{W}} \frac{L_n^*(\boldsymbol{w}) + \zeta_n(\boldsymbol{w})}{R_n^*(\boldsymbol{w})}\right\}^{-1} = o_p(1). \tag{A.11}$$

Note that

$$\begin{aligned}
\inf_{\boldsymbol{w}\in\mathcal{W}} \frac{|L_n^*(\boldsymbol{w}) + \zeta_n(\boldsymbol{w}) - \vartheta_n|}{R_n^*(\boldsymbol{w})} \ge\ & 1 - \sup_{\boldsymbol{w}\in\mathcal{W}}\left|\frac{L_n^*(\boldsymbol{w})}{R_n^*(\boldsymbol{w})} - 1\right| - \sup_{\boldsymbol{w}\in\mathcal{W}}\left|\frac{\zeta_n(\boldsymbol{w})}{R_n^*(\boldsymbol{w})}\right| - \frac{\vartheta_n}{\xi_n^*}, \\
\inf_{\boldsymbol{w}\in\mathcal{W}} \frac{L_n^*(\boldsymbol{w}) + \zeta_n(\boldsymbol{w})}{R_n^*(\boldsymbol{w})} \ge 1 &- \sup_{\boldsymbol{w}\in\mathcal{W}}\left|\frac{L_n^*(\boldsymbol{w})}{R_n^*(\boldsymbol{w})} - 1\right| - \sup_{\boldsymbol{w}\in\mathcal{W}}\left|\frac{\zeta_n(\boldsymbol{w})}{R_n^*(\boldsymbol{w})}\right|.
\end{aligned} \tag{A.12}$$

Hence, we just need to verify

$$\sup_{\boldsymbol{w}\in\mathcal{W}} \frac{|\eta_n(\boldsymbol{w})|}{R_n^*(\boldsymbol{w})} = o_p(1), \tag{A.13}$$

$$\sup_{\boldsymbol{w}\in\mathcal{W}}\left|\frac{L_n^*(\boldsymbol{w})}{R_n^*(\boldsymbol{w})} - 1\right| = o_p(1), \tag{A.14}$$

and

$$\sup_{\boldsymbol{w}\in\mathcal{W}} \frac{|\zeta_n(\boldsymbol{w})|}{R_n^*(\boldsymbol{w})} = o_p(1). \tag{A.15}$$

First, the equality (A.13) will be proved. Employing the way to prove (2.38) and (2.39) in Zhang[53], by (3.3) and (3.4) in Condition 3, we can obtain the following equalities easily:

$$\sup_{\boldsymbol{w}\in\mathcal{W}}\frac{\left|\boldsymbol{\varepsilon}^{\mathrm{T}}(\boldsymbol{I}-\mathcal{P}^{*}(\boldsymbol{w}))\boldsymbol{\mu}\right|}{R_n^{*}(\boldsymbol{w})}=o_p(1), \tag{A.16}$$

$$\sup_{\boldsymbol{w}\in\mathcal{W}}\frac{\left|\sigma^2\mathrm{tr}(\mathcal{P}^{*}(\boldsymbol{w}))-\boldsymbol{\varepsilon}^{\mathrm{T}}\mathcal{P}^{*}(\boldsymbol{w})\boldsymbol{\varepsilon}\right|}{R_n^{*}(\boldsymbol{w})}=o_p(1). \tag{A.17}$$

Next, by Rayleight-Ritz Inequality and (3.5) in Condition 3, we have

$$\begin{aligned}
&\sup_{\boldsymbol{w}\in\mathcal{W}}\frac{\left|\boldsymbol{\varepsilon}^{\mathrm{T}}\left\{\mathcal{P}(\boldsymbol{w})-\mathcal{P}^{*}(\boldsymbol{w})\right\}\boldsymbol{\mu}\right|}{R_n^{*}(\boldsymbol{w})}\\
\leq&\xi_n^{*-1}\sup_{\boldsymbol{w}\in\mathcal{W}}\sum_{s=1}^{s_n}w_s\left|\boldsymbol{\varepsilon}^{\mathrm{T}}\left\{\boldsymbol{P}_{(s)}(\widehat{\boldsymbol{\beta}}_{(s)})-\boldsymbol{P}_{(s)}(\boldsymbol{\beta}_{(s)}^{*})\right\}\boldsymbol{\mu}\right|\\
\leq&\ \xi_n^{*-1}\max_{1\leq s\leq s_n}\left|\boldsymbol{\varepsilon}^{\mathrm{T}}\left\{\boldsymbol{P}_{(s)}(\widehat{\boldsymbol{\beta}}_{(s)})-\boldsymbol{P}_{(s)}(\boldsymbol{\beta}_{(s)}^{*})\right\}\boldsymbol{\mu}\right|\\
=&\xi_n^{*-1}\max_{1\leq s\leq s_n}\left[\boldsymbol{\varepsilon}^{\mathrm{T}}\left\{\boldsymbol{P}_{(s)}(\widehat{\boldsymbol{\beta}}_{(s)})-\boldsymbol{P}_{(s)}(\boldsymbol{\beta}_{(s)}^{*})\right\}\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}\left\{\boldsymbol{P}_{(s)}(\widehat{\boldsymbol{\beta}}_{(s)})-\boldsymbol{P}_{(s)}(\boldsymbol{\beta}_{(s)}^{*})\right\}^{\mathrm{T}}\boldsymbol{\varepsilon}\right]^{1/2}\\
\leq&\xi_n^{*-1}\max_{1\leq s\leq s_n}\left(\overline{\lambda}\left[\left\{\boldsymbol{P}_{(s)}(\widehat{\boldsymbol{\beta}}_{(s)})-\boldsymbol{P}_{(s)}(\boldsymbol{\beta}_{(s)}^{*})\right\}\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}\left\{\boldsymbol{P}_{(s)}(\widehat{\boldsymbol{\beta}}_{(s)})-\boldsymbol{P}_{(s)}(\boldsymbol{\beta}_{(s)}^{*})\right\}^{\mathrm{T}}\right]\right)^{1/2}\|\boldsymbol{\varepsilon}\|\\
=&\xi_n^{*-1}\max_{1\leq s\leq s_n}\left(\overline{\lambda}\left[\boldsymbol{\mu}^{\mathrm{T}}\left\{\boldsymbol{P}_{(s)}(\widehat{\boldsymbol{\beta}}_{(s)})-\boldsymbol{P}_{(s)}(\boldsymbol{\beta}_{(s)}^{*})\right\}^{\mathrm{T}}\left\{\boldsymbol{P}_{(s)}(\widehat{\boldsymbol{\beta}}_{(s)})-\boldsymbol{P}_{(s)}(\boldsymbol{\beta}_{(s)}^{*})\right\}\boldsymbol{\mu}\right]\right)^{1/2}\|\boldsymbol{\varepsilon}\|\\
=&\xi_n^{*-1}\max_{1\leq s\leq s_n}\left[\boldsymbol{\mu}^{\mathrm{T}}\left\{\boldsymbol{P}_{(s)}(\widehat{\boldsymbol{\beta}}_{(s)})-\boldsymbol{P}_{(s)}(\boldsymbol{\beta}_{(s)}^{*})\right\}^{\mathrm{T}}\left\{\boldsymbol{P}_{(s)}(\widehat{\boldsymbol{\beta}}_{(s)})-\boldsymbol{P}_{(s)}(\boldsymbol{\beta}_{(s)}^{*})\right\}\boldsymbol{\mu}\right]^{1/2}\|\boldsymbol{\varepsilon}\|\\
\leq&\xi_n^{*-1}\max_{1\leq s\leq s_n}\left(\overline{\lambda}\left[\left\{\boldsymbol{P}_{(s)}(\widehat{\boldsymbol{\beta}}_{(s)})-\boldsymbol{P}_{(s)}(\boldsymbol{\beta}_{(s)}^{*})\right\}^{\mathrm{T}}\left\{\boldsymbol{P}_{(s)}(\widehat{\boldsymbol{\beta}}_{(s)})-\boldsymbol{P}_{(s)}(\boldsymbol{\beta}_{(s)}^{*})\right\}\right]\right)^{1/2}\|\boldsymbol{\mu}\|\|\boldsymbol{\varepsilon}\|\\
=&\xi_n^{*-1}\max_{1\leq s\leq s_n}\overline{\tau}\left\{\boldsymbol{P}_{(s)}(\widehat{\boldsymbol{\beta}}_{(s)})-\boldsymbol{P}_{(s)}(\boldsymbol{\beta}_{(s)}^{*})\right\}nO_p(1)\\
\leq&\ \xi_n^{*-1}\rho_n nO_p(1)\\
=&o_p(1),
\end{aligned} \tag{A.18}$$

where the last equality is due to Condition 4. Similarly, we observe

$$\begin{aligned}
&\sup_{\boldsymbol{w}\in\mathcal{W}}\frac{\left|\boldsymbol{\varepsilon}^{\mathrm{T}}\left\{\mathcal{P}(\boldsymbol{w})-\mathcal{P}^{*}(\boldsymbol{w})\right\}\boldsymbol{\varepsilon}\right|}{R_n^{*}(\boldsymbol{w})}\\
\leq&\xi_n^{*-1}\max_{1\leq s\leq s_n}\left(\overline{\lambda}\left[\left\{\boldsymbol{P}_{(s)}(\widehat{\boldsymbol{\beta}}_{(s)})-\boldsymbol{P}_{(s)}(\boldsymbol{\beta}_{(s)}^{*})\right\}^{\mathrm{T}}\left\{\boldsymbol{P}_{(s)}(\widehat{\boldsymbol{\beta}}_{(s)})-\boldsymbol{P}_{(s)}(\boldsymbol{\beta}_{(s)}^{*})\right\}\right]\right)^{1/2}\|\boldsymbol{\varepsilon}\|^2\\
=&\xi_n^{*-1}\max_{1\leq s\leq s_n}\overline{\tau}\left\{\boldsymbol{P}_{(s)}(\widehat{\boldsymbol{\beta}}_{(s)})-\boldsymbol{P}_{(s)}(\boldsymbol{\beta}_{(s)}^{*})\right\}nO_p(1)\\
\leq&\ \xi_n^{*-1}\rho_n nO_p(1)\\
=&o_p(1).
\end{aligned} \tag{A.19}$$

Denoting $\lambda_i(M)$ as the ith eigenvalue of any matrix M, we have

$$\begin{aligned}
&\sup_{\boldsymbol{w}\in\mathcal{W}} \frac{\left|\sigma^2 \mathrm{tr}\left\{\mathcal{P}(\boldsymbol{w})-\mathcal{P}^*(\boldsymbol{w})\right\}\right|}{R_n^*(\boldsymbol{w})}\\
\leq&\xi_n^{*-1}\sigma^2 \sup_{\boldsymbol{w}\in\mathcal{W}} \sum_{s=1}^{s_n} w_s \left|\mathrm{tr}\left\{\boldsymbol{P}_{(s)}(\widehat{\boldsymbol{\beta}}_{(s)})-\boldsymbol{P}_{(s)}(\boldsymbol{\beta}^*_{(s)})\right\}\right|\\
\leq&\xi_n^{*-1}\sigma^2 \max_{1\leq s\leq s_n} \left|\mathrm{tr}\left\{\boldsymbol{P}_{(s)}(\widehat{\boldsymbol{\beta}}_{(s)})-\boldsymbol{P}_{(s)}(\boldsymbol{\beta}^*_{(s)})\right\}\right|\\
\leq&\xi_n^{*-1}\sigma^2 \max_{1\leq s\leq s_n} \left|\sum_{i=1}^{n}\lambda_i\left\{\boldsymbol{P}_{(s)}(\widehat{\boldsymbol{\beta}}_{(s)})-\boldsymbol{P}^*_{(s)}(\boldsymbol{\beta}^*_{(s)})\right\}\right|\\
\leq&\xi_n^{*-1}\sigma^2 n \max_{1\leq s\leq s_n} \left|\mathrm{Re}\left(\lambda\left\{\boldsymbol{P}_{(s)}(\widehat{\boldsymbol{\beta}}_{(s)})-\boldsymbol{P}_{(s)}(\boldsymbol{\beta}^*_{(s)})\right\}\right)\right|\\
\leq&\ \sigma^2\xi_n^{*-1}\rho_n n\\
=&o_P(1),
\end{aligned} \tag{A.20}$$

where the last equality is based on Condition 4. Combining with (A.16)–(A.20), we achieve (A.13).

Second, by Condition 2 and Condition 3, (A.14) is proved as the analogue of (2.40) in Zhang[53].

Third, we turn to prove (A.15). By Young inequality, Rayleight-Ritz inequality and (3.5) in Condition 3, it is seen that

$$\begin{aligned}
&\xi_n^{*-1}\sup_{\boldsymbol{w}\in\mathcal{W}} 2\boldsymbol{\mu}^{\mathrm{T}}\left\{\mathcal{P}(\boldsymbol{w})-\mathcal{P}^*(\boldsymbol{w})\right\}^{\mathrm{T}}\left\{\mathcal{P}(\boldsymbol{w})-\mathcal{P}^*(\boldsymbol{w})\right\}\boldsymbol{\mu}\\
=&\xi_n^{*-1}2\sup_{\boldsymbol{w}\in\mathcal{W}}\sum_{s=1}^{s_n}\sum_{t=1}^{s_n} w_s w_t \boldsymbol{\mu}^{\mathrm{T}}\left\{\boldsymbol{P}_{(s)}(\widehat{\boldsymbol{\beta}}_{(s)})-\boldsymbol{P}_{(s)}(\boldsymbol{\beta}^*_{(s)})\right\}^{\mathrm{T}}\left\{\boldsymbol{P}_{(t)}(\widehat{\boldsymbol{\beta}}_{(t)})-\boldsymbol{P}_{(t)}(\boldsymbol{\beta}^*_{(t)})\right\}\boldsymbol{\mu}\\
\leq&\xi_n^{*-1}\sup_{\boldsymbol{w}\in\mathcal{W}}\sum_{s=1}^{s_n}\sum_{t=1}^{s_n} w_s w_t \left[\left\|\left\{\boldsymbol{P}_{(s)}(\widehat{\boldsymbol{\beta}}_{(s)})-\boldsymbol{P}_{(s)}(\boldsymbol{\beta}^*_{(s)})\right)\boldsymbol{\mu}\right\|^2+\left\|\left\{\boldsymbol{P}_{(t)}(\widehat{\boldsymbol{\beta}}_{(t)})-\boldsymbol{P}_{(t)}(\boldsymbol{\beta}^*_{(t)})\right\}\boldsymbol{\mu}\right\|^2\right]\\
\leq&\xi_n^{*-1}\|\boldsymbol{\mu}\|^2\sup_{\boldsymbol{w}\in\mathcal{W}}\sum_{s=1}^{s_n}\sum_{t=1}^{s_n} w_s w_t \left[\overline{\tau}^2\left\{\boldsymbol{P}_{(s)}(\widehat{\boldsymbol{\beta}}_{(s)})-\boldsymbol{P}_{(s)}(\boldsymbol{\beta}^*_{(s)})\right\}+\overline{\tau}^2\left\{\boldsymbol{P}_{(t)}(\widehat{\boldsymbol{\beta}}_{(t)})-\boldsymbol{P}_{(t)}(\boldsymbol{\beta}^*_{(t)})\right\}\right]\\
\leq&\xi_n^{*-1}2\|\boldsymbol{\mu}\|^2\max_{1\leq s\leq s_n}\overline{\tau}^2\left\{\boldsymbol{P}_{(s)}(\widehat{\boldsymbol{\beta}}_{(s)})-\boldsymbol{P}_{(s)}(\boldsymbol{\beta}^*_{(s)})\right\}\\
=&O\left(n\rho_n^2\xi_n^{*-1}\right)\\
=&o_p(1),
\end{aligned} \tag{A.21}$$

where the last equality is based on the fact $\rho_n\to 0$. Similarly, we can obtain

$$\xi_n^{*-1}2\sup_{\boldsymbol{w}\in\mathcal{W}}\boldsymbol{\varepsilon}^{\mathrm{T}}\left\{\mathcal{P}(\boldsymbol{w})-\mathcal{P}^*(\boldsymbol{w})\right\}^{\mathrm{T}}\left\{\mathcal{P}(\boldsymbol{w})-\mathcal{P}^*(\boldsymbol{w})\right\}\boldsymbol{\varepsilon}=O\left(n\rho_n^2\xi_n^{*-1}\right)=o_p(1). \tag{A.22}$$

Combining with (A.21) and (A.22), we know

$$\begin{aligned}
&\sup_{\boldsymbol{w}\in\mathcal{W}} \frac{\|\widehat{\boldsymbol{\mu}}(\boldsymbol{w})-\widehat{\boldsymbol{\mu}}^*(\boldsymbol{w})\|^2}{R_n^*(\boldsymbol{w})}\\
\leq&\xi_n^{*-1} \sup_{\boldsymbol{w}\in\mathcal{W}} \left[2\boldsymbol{\mu}^{\mathrm{T}} \left\{\mathcal{P}(\boldsymbol{w})-\mathcal{P}^*(\boldsymbol{w})\right\}^{\mathrm{T}} \left\{\mathcal{P}(\boldsymbol{w})-\mathcal{P}^*(\boldsymbol{w})\right\} \boldsymbol{\mu}\right] +\\
&\xi_n^{*-1} \sup_{\boldsymbol{w}\in\mathcal{W}} \left[2\boldsymbol{\varepsilon}^{\mathrm{T}} \left\{\mathcal{P}(\boldsymbol{w})-\mathcal{P}^*(\boldsymbol{w})\right\}^{\mathrm{T}} \left\{\mathcal{P}(\boldsymbol{w})-\mathcal{P}^*(\boldsymbol{w})\right\} \boldsymbol{\varepsilon}\right]\\
=&o_p(1). \qquad (A.23)
\end{aligned}$$

By Cauchy-Schwarz inequality, we have

$$\begin{aligned}
&\sup_{\boldsymbol{w}\in\mathcal{W}} \frac{\left|(\widehat{\boldsymbol{\mu}}^*(\boldsymbol{w})-\boldsymbol{\mu})^{\mathrm{T}} (\widehat{\boldsymbol{\mu}}(\boldsymbol{w})-\widehat{\boldsymbol{\mu}}^*(\boldsymbol{w}))\right|}{R_n^*(\boldsymbol{w})}\\
\leq&\sqrt{\sup_{\boldsymbol{w}\in\mathcal{W}} \frac{L_n^*(\boldsymbol{w})}{R_n^*(\boldsymbol{w})}} \sqrt{\sup_{\boldsymbol{w}\in\mathcal{W}} \frac{\|\widehat{\boldsymbol{\mu}}(\boldsymbol{w})-\widehat{\boldsymbol{\mu}}^*(\boldsymbol{w})\|^2}{R_n^*(\boldsymbol{w})}}\\
\leq&\sqrt{\sup_{\boldsymbol{w}\in\mathcal{W}} \left|\frac{L_n^*(\boldsymbol{w})}{R_n^*(\boldsymbol{w})}-1\right|+1} \sqrt{\sup_{\boldsymbol{w}\in\mathcal{W}} \frac{\|\widehat{\boldsymbol{\mu}}(\boldsymbol{w})-\widehat{\boldsymbol{\mu}}^*(\boldsymbol{w})\|^2}{R_n^*(\boldsymbol{w})}}\\
=&o_p(1), \qquad (A.24)
\end{aligned}$$

where the last equality is based on (A.14) and (A.23). Finally, putting (A.5), (A.23) and (A.24), we achieve (A.15). ∎