AIMS Mathematics, 8(9): 21683–21699.
DOI: 10.3934/math.20231106
Received: 14 March 2023
Revised: 28 May 2023
Accepted: 31 May 2023
Published: 07 July 2023

http://www.aimspress.com/journal/Math

Theory article

Model averaging based on weighted generalized method of moments with missing responses

Zhongqi Liang[1,2] **and Yanqiu Zhou**[3,*]

[1] School of Mathematical Sciences, Zhejiang University, Hangzhou 310058, China
[2] School of Computer and Computing Science, Hangzhou City University, Hangzhou 310015, China
[3] School of Science, Guangxi University of Science and Technology, Liuzhou 545006, China

* **Correspondence:** Email: zhouyanqiunihao@163.com.

Abstract: Model averaging based on the least squares estimator or the maximum likelihood estimator has been widely followed, while model averaging based on the generalized method of moments is almost rarely addressed. This paper is concerned with a model averaging method based on the weighted generalized method of moments for missing responses problem. The weight vector for model averaging is obtained via minimizing the leave-one-out cross validation criterion. With some mild conditions, the asymptotic optimality of the proposed method in the sense that it can achieve the lowest squared error asymptotically is proved. Some numerical experiments are conducted to evaluate the proposed method with the existing related ones, and the results suggest that the proposed method performs relatively well.

Keywords: asymptotic optimality; generalized method of moments; missing responses; model averaging
Mathematics Subject Classification: 62F12, 62F99, 62J05

1. Introduction

Due to the development of modern statistics, fitting multiple uncertain candidate models to a given data set is no longer a challenging job. With multiple candidate models, it becomes clear that some reasonable methods are needed to summarize the information of these candidate models in some way. Model selection and model averaging, two methods specifically designed to deal with scenarios with multiple fitted candidate models, are increasingly popular among statistics. Model selection aims to pick one model from multiple candidate models, and thus it is like putting all the eggs in one basket. Therefore, model selection is relatively risky and has a lot of undesirable drawbacks in some situations compared to model averaging [13]. Model averaging utilizes fully the useful information among each

candidate model and then yields a robust weighting estimator. Thus, model averaging is more likely to hedge the risk associated with putting all the eggs in one basket. A substantial amount of research has been devoted to frequentist model averaging over the past decades. See e.g. [1,2,4,5,9,11,12,17,23, 25–28] among others. This paper also focuses on the frequentist model averaging method. However, all of the above model averaging literature is based on the least squares estimation or the maximum likelihood estimation. It is well known that the generalized method of moments (GMM) only needs to know some moment functions, and hence GMM is more flexible and applicable compared to the least squares estimation method and the maximum likelihood estimation method in some situations. Since [6] proposed the two-step GMM and verified the large sample properties of GMM estimators, GMM has become a fundamental estimation method and is widely applied in the field of statistical prediction and statistical inference. However, the application of GMM for model averaging is very rare. Under the local misspecified framework, [3] proposed a simulation-based averaging procedure, from which the GMM estimators of candidate models were obtained from different moment condition sets. It should be pointed out that the assumption of local misspecified framework is somewhat unrealistic since it requires sufficient knowledge of the true model [15].

Under the fixed parameter framework, [24] proposed a model averaging method to combine the GMM estimator of the interesting parameter of each candidate model, and they proved that the proposed method is asymptotically optimal. On the basis of [24], [20] considered the model averaging methods for the conditional mean of the responses based on GMM by using J-fold cross validation criterion. Similarly, they also proved that their method is asymptotically optimal under some certain conditions. However, all of the above work was done with fully observed data. So far, no one has considered the model averaging methods based on GMM with missing data. The purpose of this paper is to fill the gap in this area.

Missing data occurs commonly in questionnaires, socioeconomic studies and medical studies and so on [10]. The complexity of missing data often invalidates the original model averaging approach with complete data. Therefore, it is necessary to study model averaging methods with missing data. The missing pattern is very important in the study of missing data. In this paper, we consider the missing pattern that responses are missing at random (MAR) [18]. In fact, there is already some literature on model averaging with this missing pattern. Under the local misspecified framework, [16] proposed model averaging methods for linear models with responses are missing at random. Under the fixed parameter framework, [21] developed a model averaging scheme for linear models with responses are missing at random. [22] extended the linear models to high-dimensional linear regression models and established a novel model averaging criterion for high dimensional linear models with responses are missing at random. However, the core of the above literature is based on the least squares estimator rather than the GMM estimator.

In this paper, we suggest a novel model averaging method based on GMM for regression models with responses are missing at random under the fixed parameter framework. With an assumed parametric model for the selection probability function, through the inverse probability weighting method and GMM, the inverse probability weighted GMM estimator of the parameter of interest under each candidate model is obtained. Then, the model averaging estimator of the conditional mean of responses based on GMM can be derived directly. The optimal weight vector for model averaging is obtained via minimizing the leave-one-out cross validation criterion. Under certain mild conditions, we prove that the proposed model averaging method is asymptotically optimal in the sense that the

mean squared loss obtained by our proposal is asymptotically identical to that of the infeasible optimal weight vector.

The rest of this paper is organized as follows. In Section 2, we describe the methodology and then give the theoretical property of the proposed method. Some Monte Carlo simulation studies are presented in Section 3. A short conclusion is made in Section 4. All the technical details are delayed to the Appendix.

2. Material and method

Consider the following regression model

$$Y_i = \mu_i + \epsilon_i = \varphi(X_i, \theta) + \epsilon_i, \quad i = 1, 2, ..., n, \tag{2.1}$$

where $X_i = (X_{i1}, X_{i2}, \cdots, X_{ip})^\intercal$ is a p-dimensional vector of regressors, $\varphi(\cdot)$ is a known function of X_i and θ, θ is the corresponding $p \times 1$ parametric vector, and " $^\intercal$ " denotes the transpose of a vector or matrix. ϵ_i are the random errors with $E(\epsilon_i|X_i) = 0$ and $E(\epsilon_i^2|X_i) = \sigma^2$ for $i = 1, 2, \cdots n$. Let $Y = (Y_1, Y_2, \cdots, Y_n)^\intercal$, $X = (X_1, X_2, \cdots, X_n)^\intercal$, $\mu = (\mu_1, \mu_2, \cdots, \mu_n)^\intercal$ and $\epsilon = (\epsilon_1, \epsilon_2, \cdots, \epsilon_n)^\intercal$. Assume that $\mu = \mu(\theta)$ is a continuous function with respect to θ under (2.1). In this paper, we consider the case that some responses Y_i are missing, while X_i is fully observed. That is, the data comes from $\{(Y_i, X_i, \delta_i), i = 1, 2, \cdots, n\}$, where δ_i is the missing indicator of Y_i, $\delta_i = 1$ if Y_i is observed, $\delta_i = 0$ otherwise. Assume that Y is missing at random (MAR), a commonly used missing mechanism, that is,

$$Pr(\delta = 1|Y, X) = Pr(\delta = 1|X) := \pi(X), \tag{2.2}$$

where X may be a sub-vector of itself X, a slight abuse of notation here.

Under MAR assumption, it is easy to show that $E[Y_{\pi,i}|X_i] = E[Y_i|X_i] = \mu_i$, where $Y_{\pi,i} = \delta_i Y_i/\pi(X_i)$ for $i = 1, 2, \cdots, n$. We then have

$$Y_{\pi,i} = \mu_i + \epsilon_{\pi,i} = \varphi(X_i, \theta) + \epsilon_{\pi,i}, \quad i = 1, 2, ..., n, \tag{2.3}$$

where $\epsilon_{\pi,i}$ are the corresponding random errors with $E(\epsilon_{\pi,i}|X_i) = 0$ and $E(\epsilon_{\pi,i}^2|X_i) = \sigma_{\pi,i}^2$ with $\sigma_{\pi,i}^2 = \{\pi(X_i)\}^{-1}(\mu_i^2 + \sigma^2) - \mu_i^2$ for $i = 1, 2, \cdots n$. Let $Y_\pi = (Y_{\pi,1}, Y_{\pi,2}, \cdots, Y_{\pi,n})^\intercal$ and $\epsilon_\pi = (\epsilon_{\pi,1}, \epsilon_{\pi,2}, \cdots, \epsilon_{\pi,n})^\intercal$. In this paper, we consider an optimal model averaging method to estimate μ with responses missing at random.

In what follows, we give the details of how to obtain the optimal model averaging estimator of μ. Suppose that we have finite M candidate models to approximate model (2.1). Accordingly, there are also M candidate models to approximate model (2.3). Naturally, under the mth candidate model, we have $\mu_{(m)} = \mu(\theta_{(m)})$, where $\theta_{(m)}$ is the unknown $p_m \times 1$ sub-vector. Correspondingly, for M candidate models, we have M estimators $\{\hat{\mu}_{(1)}, \hat{\mu}_{(2)}, \cdots, \hat{\mu}_{(M)}\}$ of μ. Let $w = (w_1, w_2, \cdots, w_M)^\intercal$ be the weight vector corresponding to the candidate models, which belongs to the set

$$W_n = \left\{ w \in [0, 1]^M : \sum_{m=1}^{M} w_m = 1 \right\}.$$

Then the model averaging estimator of μ can be given as follows,

$$\hat{\mu}(w) = \sum_{m=1}^{M} w_m \hat{\mu}_{(m)} = \hat{\mu} w, \tag{2.4}$$

where $\hat{\mu} = (\hat{\mu}_{(1)}, \hat{\mu}_{(2)}, \cdots, \hat{\mu}_{(M)})$ is an $n \times M$ matrix, and $\hat{\mu}_{(m)} = \mu(\hat{\theta}_{(m)})$ with $\hat{\theta}_{(m)}$ is a feasible estimator of $\theta_{(m)}$ for $m = 1, 2, \cdots, M$.

However, the selection probability function $\pi(X)$ and $\hat{\theta}_{(m)}$ are still unknown so far for $m = 1, 2, \cdots, M$. Thus, $\hat{\mu}(w)$ in (2.4) still can not be used directly. We first consider the estimation method of $\pi(X)$. Following the common missing data literature [10], we assume a parametric model $\pi(X; \alpha)$ for the selection probability function $\pi(X)$, where $\pi(\cdot; \alpha)$ is a known function and α is an unknown parametric vector. Denote $\hat{\alpha}_n$ by the maximum likelihood estimator (MLE) of α, which is obtained by maximizing the following binomial log-likelihood,

$$l(\alpha) = \sum_{i=1}^{n} [\delta_i \log(\pi(X_i; \alpha)) + (1 - \delta_i) \log(1 - \pi(X_i; \alpha))].$$

Then a consistent estimator $\pi(X_i; \hat{\alpha}_n)$ of $\pi(X_i; \alpha)$ can be obtained immediately for $i = 1, 2, \cdots, n$. In the following, for notation convenience, we write $\hat{\pi}_i(\hat{\alpha}_n) = \pi(X_i; \hat{\alpha}_n)$ for $i = 1, 2, \cdots, n$.

Next, we consider how to obtain the inverse probability weighted GMM estimator $\hat{\theta}_{(m)}$ of $\theta_{(m)}$ for $m = 1, 2, \cdots, M$. Denote $g_{(m)}(T_i, \theta)$ by the q_m-vector of moment functions of the mth candidate model for an integer $q_m \geq p_m$, where $T_i = (X_i^\top, Y_i)^\top$, $i = 1, 2, \cdots, n$. Let $\theta_{(m)}^0$ be the unique true parametric value of $\Theta_{(m)} \subset \mathbb{R}^{p_m}$ such that $E[g_{(m)}(T_i, \theta_{(m)}^0)] = 0$ for the mth candidate model. Then, the inverse probability weighted GMM estimator of $\theta_{(m)}$ can be given by

$$\hat{\theta}_{(m)} = \arg\min_{\Theta_{(m)}} \left[\sum_{i=1}^{n} \frac{\delta_i}{\hat{\pi}_i(\hat{\alpha}_n)} g_{(m)}(T_i, \theta) \right]^\top \hat{\Omega}_m \left[\sum_{i=1}^{n} \frac{\delta_i}{\hat{\pi}_i(\hat{\alpha}_n)} g_{(m)}(T_i, \theta) \right], \tag{2.5}$$

where $\hat{\Omega}_m$ is a positive definite weighting matrix. Similar to [24], we set $\hat{\Omega}_m = (X_{(m)}^\top X_{(m)}/n)^{-1}$, where $X_{(m)}$ is the regression matrix in the mth candidate model for $m = 1, 2, \cdots, M$. Thus, the feasible weighted GMM model averaging estimator of μ is

$$\hat{\mu}(w) = \sum_{m=1}^{M} w_m \hat{\mu}_{(m)} = \sum_{m=1}^{M} w_m \mu(\hat{\theta}_{(m)}). \tag{2.6}$$

To obtain the optimal weight vector w, we adopt the leave-one-out cross validation criterion. Denote $\tilde{\theta}_{(m)}^{(-i)}$ by the leave-one-out weight GMM estimator, which is obtained when the ith observation is deleted when computing $\hat{\theta}_{(m)}$. Let $\tilde{\theta}_{(m)} = (\tilde{\theta}_{(m)}^{(-1)}, \tilde{\theta}_{(m)}^{(-2)}, \cdots, \tilde{\theta}_{(m)}^{(-n)})^\top$ and $\tilde{\mu}_{(m)} = (\tilde{\mu}_{(m)}^{(-1)}, \tilde{\mu}_{(m)}^{(-2)}, \cdots, \tilde{\mu}_{(m)}^{(-n)})^\top$, where $\tilde{\mu}_{(m)}^{(-i)} = \mu(\tilde{\theta}_{(m)}^{(-i)})$. Similar to [5], we adopt the following model averaging criterion,

$$C_{GMM-\hat{\pi}}(w) = \|Y_{\hat{\pi}} - \tilde{\mu}(w)\|^2 = w^\top \tilde{\epsilon}_{\hat{\pi}}^\top \tilde{\epsilon}_{\hat{\pi}} w, \tag{2.7}$$

where $Y_{\hat{\pi}} = \{Y_{\hat{\pi},1}, Y_{\hat{\pi},2}, \cdots, Y_{\hat{\pi},n}\}^\top$ with $Y_{\hat{\pi},i} = \delta_i Y_i / \{\hat{\pi}_i(\hat{\alpha}_n)\}$, $\tilde{\mu}(w) = \sum_{m=1}^{M} w_m \tilde{\mu}_{(m)}$, $\tilde{\epsilon}_{\hat{\pi}} = (\tilde{\epsilon}_{\hat{\pi},(1)}, \tilde{\epsilon}_{\hat{\pi},(2)}, \cdots, \tilde{\epsilon}_{\hat{\pi},(M)})$, $\tilde{\epsilon}_{\hat{\pi},(m)} = Y_{\hat{\pi}} - \tilde{\mu}_{(m)}$ and $\|\cdot\|$ represents the Euclidean norm.

Denote $\hat{w}$ by the optimal weight vector obtained by minimizing $C_{GMM-\hat{\pi}}(w)$ among W_n, that is

$$\hat{w} = \arg\min_{w \in W_n} C_{GMM-\hat{\pi}}(w). \tag{2.8}$$

Then the corresponding model averaging estimator of μ is $\hat{\mu}(\hat{w})$. For convenience, hereinafter this estimator is referred to as weighted averaging GMM estimator (WAGMM), the corresponding model averaging method above is abbreviated as WAGMM method.

In what follows, we give the theoretical property of this paper. We give some notations and regularity conditions at first. Let α_0 be the true value of α, $\lambda_{\max}(A)$ the largest singular value of matrix A, w_m^0 the special weight vector whose mth element is one and the others are zero. For $m = 1, 2, \cdots, M$, denote $\mu_{(m)}^0 = \mu(\theta_{(m)}^0)$, $\mu^0(w) = \sum_{m=1}^M w_m \mu_{(m)}^0$. Define the mean square loss function of $\hat{\mu}(w)$ as $L(w) = \|\hat{\mu}(w) - \mu\|^2$, and denote by $L^0(w) = \|\mu^0(w) - \mu\|^2$ and $\xi_\pi = \inf_{w \in W_n} L^0(w)$. The required regularity conditions are given as follows.

(C.1) For some fixed $1 \le G < \infty$, $\max_{1 \le i \le n} E[\epsilon_i^{4G} | X_i] < c_1 < \infty$.

(C.2) There exists a positive constant c_2 such that $\max_{1 \le i \le n} |\mu_i| \le c_2$.

(C.3) $\inf_x \pi(x; \alpha) > c_\pi > 0$, its first three-order partial derivatives with respect to α are bounded and continuous in the neighborhood α_0, and $E\left[\|\partial\pi(X; \alpha)/\partial\alpha\|^2\right] < \infty$.

(C.4) $E[g_{(m)}(T_i, \theta_{(m)}^0) g_{(m)}^\intercal(T_i, \theta_{(m)}^0)]$ is positive definite, $\partial g_{(m)}(T_i, \theta_{(m)})/\partial\theta$ is continuous and bounded in the neighborhood of $\theta_{(m)}^0$, and $E\{\sup_{\theta \in \Theta} \|g_{(m)}(T_i, \theta_{(m)})\|^3\} < \infty$ for $m = 1, 2 \cdots, M$ and $i = 1, 2, \cdots, n$.

(C.5) The partial derivatives $A_{(m)}^{(i)} = \partial\mu^{(i)}(\theta_{(m)})/\partial\theta_{(m)}^{(i)}$ and $\tilde{A}_{(m)}^{(-i)} = \partial\tilde{\mu}^{(-i)}(\theta_{(m)})/\partial\theta_{(m)}^{(-i)}$ exist, where $\mu_{(m)}^{(i)}$ is the ith component of $\mu_{(m)}$ and $\tilde{\mu}_{(m)}^{(i)}$ is the ith component of $\tilde{\mu}_{(m)}$. $\max_{1 \le i \le n} \lambda_{\max}([A_{(m)}^{(i)}]^\intercal A_{(m)}^{(i)}) = O(1)$ as well as $\max_{1 \le i \le n} \lambda_{\max}([\tilde{A}_{(m)}^{(i)}]^\intercal \tilde{A}_{(m)}^{(i)}) = O(1)$, a.s., for each $\theta_{(m)}$ and $i = 1, 2, \cdots, n$.

(C.6) $\xi_\pi^{-1} n^{1/2} = o(1)$.

Remark. Condition (C.1) is a mild conditional moment condition of ϵ, which places a common bound on ϵ. It is common in model averaging literature, such as [4, 17]. Condition (C.2) imposes some restrictions on the mean value function μ, which is also common in regression analysis and model averaging literature. See [11, 20] and so on. Condition (C.3) is necessary for the missing data, the lower bound ensures that the weights for the inverse probability weighting can not go to infinity as $n \to \infty$. Similar conditions can be seen in [18, 19, 22] and so on. Condition (C.4) is similar to Condition (C4) in [14], which is just a regular condition of the moment function vector. Condition (C.5) imposes some restrictions on the largest singular value of two partial derivatives, which is the same as Condition (C.2) in [20]. Condition (C.6) requires that $\xi_\pi \to \infty$, which can be deduced from the conditions of much of the existing literature such as Condition (7) in [1], Condition (A3) in [2] and Conditions (C.3) and (C.6) in [27].

Theorem 1. *Under Conditions (C.1) to (C.6), as $n \to \infty$, it holds that*

$$\frac{L(\hat{w})}{\inf_{w \in W_n} L(w)} = 1 + o_p(1). \tag{2.9}$$

Theorem 1 indicates that the selected weight vector $\hat{w}$ is asymptotically optimal. That is, the mean squared loss obtained by the proposed WAGMM estimator $\hat{\mu}(\hat{w})$ is asymptotically identical to that of the infeasible best model averaging estimator.

3. Simulation

In this section, some Monte Carlo simulation studies were conducted to examine the finite-sample performance of the proposed WAGMM method. Referring to the idea of [24], the two methods we compare are all based on the GMM estimator and do not compare with other existing model selection and model averaging methods. The first method is the weighted AIC [7] model selection method based

on the GMM estimator. We abbreviated it as wAIC-GMM. The second comparison method is the leave-one-out model averaging method based on the GMM estimator with complete case (CC), which simply ignores the missing individuals in handling missing data. We abbreviated it as CC-GMM.

The data generating process is similar to that of [20], as follows specifically,

$$Y_i = \beta_0 X_i + \epsilon_i, \quad X_i = (1, h_i)^\tau, \quad h_i = \eta^\tau Z_i + e_i, \quad i = 1, 2, \cdots, n,$$

where the interest parameter β_0 was set to $\beta_0 = (1, -0.5)$, $Z_i = \{Z_{i1}, Z_{i2}, \cdots, Z_{i10}\}$ were simulated from the independent normal distribution with mean 0 and variance 1. η is the 10×1 regression coefficient. Similar to [20], we considered the following two cases,

$$\begin{aligned} &\text{Case 1:} \quad \eta = \sqrt{Q_f^2/(10(1 - Q_f^2))}, \\ &\text{Case 2:} \quad \eta = \sqrt{Q_f^2/(t(1 - Q_f^2))}, \quad t = 1, 2, \cdots, 10, \end{aligned}$$

where $Q_f = 0.8$. Obviously, Case 1 means that all Z_{ij} are equally important, while Case 2 means that the importance of Z_{ij} is gradually decreasing, $j = 1, 2, \cdots, n$. And $(\epsilon_i, e_i)^\tau \overset{i.i.d}{\sim} Normal(0_{2\times 1}, \Sigma)$, where

$$\Sigma = \begin{pmatrix} \sigma^2 & 0.5\sigma \\ 0.5\sigma & 1 \end{pmatrix}.$$

We controlled the value of σ such that $R^2 = var(\beta_0 X_i)/var(Y_i)$ varied in the set $\{0.1, 0.2, \cdots, 0.9\}$ for $i = 1, 2, \cdots, n$. The variables in Z_i were added sequentially to obtain the candidate models. Thus, in the above setting, we have, in total, $M = 10$ candidate models. The estimation equation was set as the unconditional moment $E[Z_i(Y_i - \beta_0 X_i)] = 0$ for $i = 1, 2, \cdots, n$. For generating the missing values of $\{Y_i\}_{i=1}^n$, the selection probability function was set to

$$Pr(\delta_i = 1 | Y_i, X_i) = \Phi(\alpha_0 + \alpha_1 X_i), \quad i = 1, 2, \cdots, n, \tag{3.1}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal function, and $\alpha = (\alpha_0, \alpha_1)^\tau = (0.6, 0.1)^\tau$, which makes the corresponding average missing rate is approximately 30%. It is obvious that the selection probability function (3.1) satisfies the assumption of MAR.

To evaluate the performance of these methods, we considered the following mean squared error (MSE), which is defined as,

$$\text{MSE} = \frac{1}{K} \sum_{k=1}^{K} \|\mu^{(k)}(\hat{w}) - \mu^{(k)}\|^2, \tag{3.2}$$

where K represents the number of repetitions, $\mu^{(k)}$ denotes μ in kth repetition and $\hat{\mu}^{(k)}(\hat{w})$ is the model averaging or selection estimator of μ in the kth repetition. The repetition number K was set to be 300. The sample size n we considered was $n = 50, 100, 200$ and 400. For a more intuitive comparison, we calculated the ratio of the MSE of each method divided by the MSE of the proposed WAGGM method separately. The results of the simulation studies are presented below (see Figures 1 and 2).

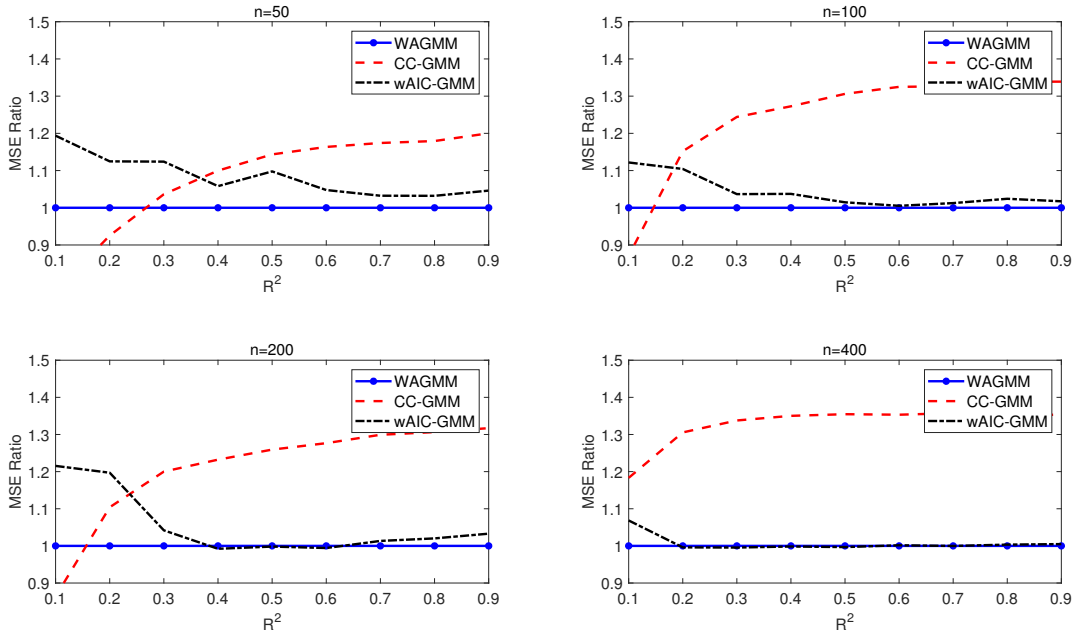

Figure 1. The MSE ratio under Case 1.

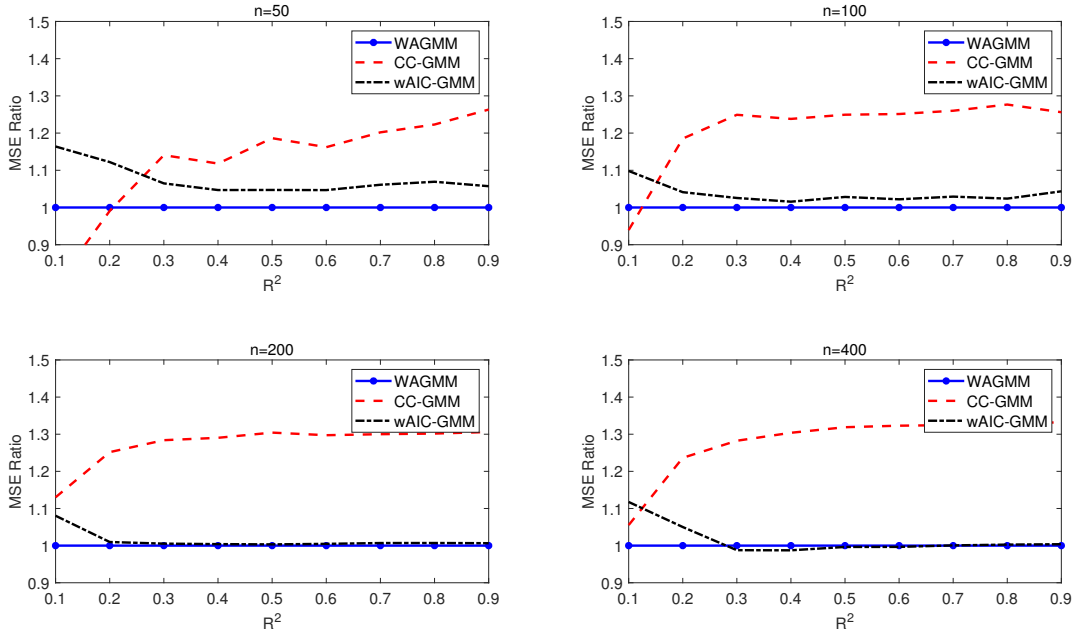

Figure 2. The MSE ratio under Case 2.

It can be seen from Figures 1 and 2 that the MSE of the proposed WAGMM method is the smallest in most cases, no matter in Case 1 or Case 2, and is significantly better than the wAIC-GMM method

and the CC-GMM method. It is not difficult to find that the MSE of the wAIC-GMM method gradually decreases and tends to one as the sample size increases. The reason is that the wAIC-GMM method is also consistent when the selection probability function is correctly specified [7]. Nonetheless, the proposed WAGMM method still outperforms the wAIC-GMM method in most combinations we consider, especially when the sample size is small enough. The performance of the CC-GMM is very poor in all cases. This is because the CC-GMM method is inconsistent under the assumption of MAR [10].

In order to illustrate the asymptotic optimality of the proposed WAGMM method more intuitively, we separately calculated the mean of GR in Case 1 and Case 2 when $R^2 = 0.5$ based on 1000 simulation repetitions. In this simulation experiment, the sample size N was set to $N = 100, 200, 400$ and 800, respectively. The expression for GR is defined as follows,

$$\text{GR} = \frac{L(\hat{w})}{\inf_{w \in W_n} L(w)}.$$

The results are shown in Figures 3 and 4 below.

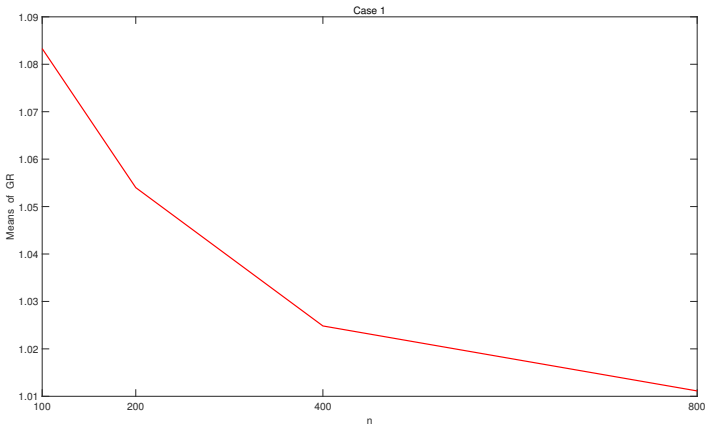

Figure 3. Asymptotic optimality evaluation of the WAGMM method in Case 1.

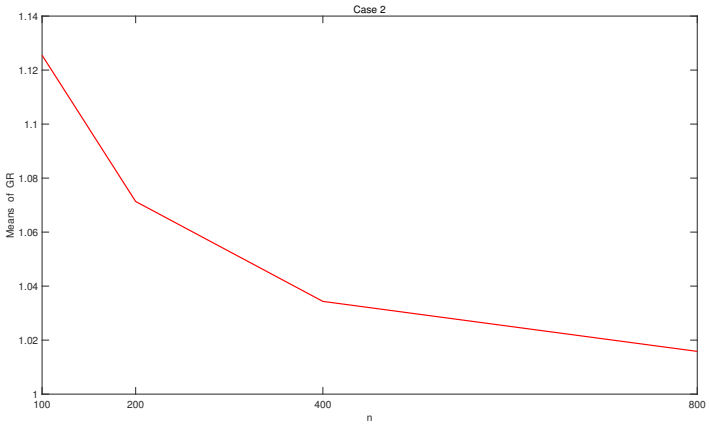

Figure 4. Asymptotic optimality evaluation of the WAGMM method in Case 2.

Figures 3 and 4 present the mean curves of GR for Case 1 and Case 2, respectively. It can be clearly seen from Figures 3 and 4 that no matter in Case 1 or Case 2, the mean curves of GR gradually decrease and gradually approach 1 as the sample size increases. The simulation results in Figures 3 and 4 above numerically and intuitively verify the asymptotic optimality stated in Theorem 1.

4. Conclusions

To the best of our knowledge, the current paper is the first work to develop the model averaging method based on GMM for regression models with missing data. Nevertheless, there are still some significant issues left for future study. At first, as [20] said, the selection of moment conditions is very important in GMM. Thus, it is important to choose the optimal moment conditions to obtain the corresponding GMM estimators in each candidate model before performing model averaging method with missing data. However, the current work also does not address this issue. Secondly, the number of candidate models M is finite in this paper. The scenario when the M is divergent to infinite as the sample size tends to infinite is also common. All of these issues mentioned above deserve further consideration in the future.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

The authors thank editor, associate editor and two anonymous reviewers for their constructive comments and helpful suggestions that help greatly improve the manuscript.

Zhou's research was supported by the Guangxi University Young and Middle-aged Teachers' Basic Research Ability Improvement Project (2021KY0343), and the Guangxi Science and Technology Base and Talent Project (2020ACI9151).

Conflict of interest

All authors declare no conflicts of interest in this paper.

Appendix

In this Appendix, we will give the technical proof of Theorem 1. To facilitate the proof of Theorem 1, we would give some lemmas at first. Throughout Appendix, C denotes a generic positive constant depending on the context, which may take different values in different scenarios. The norm of the matrix is the Euclidean norm, that is, the largest singular value of the matrix.

Lemma 1. *Under Conditions (C.1) to (C.4), for any $\hat{\theta}_{(m)}$, as $n \to \infty$, it holds that*

$$\hat{\theta}_{(m)} - \theta^0_{(m)} = O_p(n^{-1/2}), \quad m = 1, 2, \cdots, M. \tag{A1}$$

Proof of Lemma 1. Denote

$$R_n(\theta_{(m)}, \hat{\Omega}_m) = \left[\sum_{i=1}^{n} \hat{r}_i g_{(m)}(T_i, \theta)\right]^{\tau} \hat{\Omega}_m \left[\sum_{i=1}^{n} \hat{r}_i g_{(m)}(T_i, \theta)\right],$$

where $\hat{r}_i = \delta_i/\pi_i(\hat{\alpha}_n)$, $i = 1, 2, \cdots, n$. Then by (2.5), it can be seen that $\hat{\theta}_{(m)} = \arg\min_{\Theta_{(m)}} R_n(\theta_{(m)}, \hat{\Omega}_m)$. Under Conditions (C.1)–(C.4), by similar proving statements of Theorem 1 in [14], we have

$$\hat{\theta}_{(m)} \xrightarrow{p} \theta^0_{(m)}, \quad m = 1, 2, \cdots, M, \tag{A2}$$

where $\xrightarrow{p}$ denotes convergence in probability. Immediately following, by Taylor expansion, for any $m = 1, 2, \cdots, M$, we have

$$0 = \frac{\partial R_n(\theta_{(m)}, \hat{\Omega}_m)}{\partial \theta_{(m)}}\bigg|_{\theta_{(m)}=\theta^0_{(m)}} + \frac{\partial^2 R_n(\theta_{(m)}, \hat{\Omega}_m)}{\partial \theta_{(m)} \partial \theta^{\tau}_{(m)}}\bigg|_{\theta_{(m)}=\theta^*_{(m)}} (\hat{\theta}_{(m)} - \theta^0_{(m)}),$$

where $\theta^*_{(m)}$ is a real mean value between $\theta_{(m)}$ and $\theta^0_{(m)}$, and

$$\frac{\partial R_n(\theta_{(m)}, \hat{\Omega}_m)}{\partial \theta_{(m)}} = \left[\sum_{i=1}^{n} \hat{r}_i \frac{\partial g_{(m)}(T_i, \theta)}{\partial \theta_{(m)}}\right]^{\tau} \hat{\Omega}_m \left[\sum_{i=1}^{n} \hat{r}_i g_{(m)}(T_i, \theta)\right],$$

$$\begin{aligned}\frac{\partial^2 R_n(\theta_{(m)}, \hat{\Omega}_m)}{\partial \theta_{(m)} \theta^{\tau}_{(m)}} =& \left[\sum_{i=1}^{n} \hat{r}_i \frac{\partial g_{(m)}(T_i, \theta)}{\partial \theta_{(m)}}\right]^{\tau} \hat{\Omega}_m \left[\sum_{i=1}^{n} \hat{r}_i \frac{\partial g_{(m)}(T_i, \theta)}{\partial \theta_{(m)}}\right] \\ &+ \left[\sum_{i=1}^{n} \hat{r}_i \frac{\partial^2 g_{(m)}(T_i, \theta)}{\partial \theta_{(m)} \partial \theta^{\tau}_{(m)}}\right]^{\tau} \hat{\Omega}_m \left[\sum_{i=1}^{n} \hat{r}_i g_{(m)}(T_i, \theta)\right].\end{aligned}$$

By simply shifting the terms, it can be seen that

$$\begin{aligned}&\sqrt{n}(\hat{\theta}_{(m)} - \theta^0_{(m)}) \\ &= -\left\{\left[\frac{\partial^2 R_n(\theta_{(m)}, \hat{\Omega}_m)}{\partial \theta_{(m)} \partial \theta^{\tau}_{(m)}}\bigg|_{\theta_{(m)}=\theta^*_{(m)}}\right]^{-1} - \left[\frac{\partial^2 R_n(\theta_{(m)}, \hat{\Omega}_m)}{\partial \theta_{(m)} \partial \theta^{\tau}_{(m)}}\bigg|_{\theta_{(m)}=\theta^0_{(m)}}\right]^{-1}\right\} \\ &\times \sqrt{n}\frac{\partial R_n(\theta_{(m)}, \hat{\Omega}_m)}{\partial \theta_{(m)}}\bigg|_{\theta_{(m)}=\theta^0_{(m)}} - \left[\frac{\partial^2 R_n(\theta_{(m)}, \hat{\Omega}_m)}{\partial \theta_{(m)} \partial \theta^{\tau}_{(m)}}\bigg|_{\theta_{(m)}=\theta^0_{(m)}}\right]^{-1} \\ &\times \sqrt{n}\frac{\partial R_n(\theta_{(m)}, \hat{\Omega}_m)}{\partial \theta_{(m)}}\bigg|_{\theta_{(m)}=\theta^0_{(m)}}.\end{aligned} \tag{A3}$$

Then, under Conditions (C.3) and (C.4), by invoking similar proving arguments of Lemma (A.1) in [14], for $m = 1, 2, \cdots, M$, we have

$$\sum_{i=1}^{n} \hat{r}_i \frac{\partial g_{(m)}(T_i, \theta)}{\partial \theta_{(m)}}\bigg|_{\theta_{(m)}=\theta^0_{(m)}} \xrightarrow{p} E\left[\frac{\partial g_{(m)}(T, \theta)}{\partial \theta}\bigg|_{\theta_{(m)}=\theta^0_{(m)}}\right], \tag{A4}$$

and

$$\sqrt{n}\sum_{i=1}^{n}\hat{r}_i g_{(m)}(T_i,\theta^0_{(m)}) = O_p(1), \quad E\left[\sum_{i=1}^{n}\hat{r}_i g_{(m)}(T_i,\theta^0_{(m)})\right] = 0. \tag{A5}$$

Then, combining with (A2)–(A5) for $m = 1, 2, \cdots, M$, we have

$$\hat{\theta}_{(m)} - \theta^0_{(m)} = O_p(n^{-1/2}). \tag{A6}$$

Thus, the proof of Lemma 1 is completed.

Lemma 2. *Under Conditions (C.1) to (C.3) and (C.6), as $n \to \infty$, it holds that*

$$\sup_{w\in W_n}\frac{\|Y_{\hat{\pi}} - Y_\pi\|^2}{L^0(w)} = o_p(1). \tag{A7}$$

Proof. By Conditions (C.1) and (C.2) and the law of large numbers, it is easy to show that

$$\frac{1}{n}\mu^\top\mu = O_p(1), \quad \frac{1}{n}\epsilon^\top\epsilon = O_p(1). \tag{A8}$$

Note that $Y_{\hat{\pi},i} = \delta_i Y_i/\{\pi(X_i;\hat{\alpha}_n)\}$ for $i = 1, 2, \cdots, n$. By Cauchy-Schwarz inequality and Taylor expansion, we have

$$\begin{aligned}
&\sup_{w\in W_n}\frac{\|Y_{\hat{\pi}} - Y_\pi\|^2}{L^0(w)}\\
\le&\xi_\pi^{-1}\sum_{i=1}^{n}\left\{\frac{\delta_i}{\hat{\pi}(X_i;\hat{\alpha}_n)} - \frac{\delta_i}{\pi(X_i;\alpha_0)}\right\}^2 Y_i^2\\
\le&C\xi_\pi^{-1}\left\{\sqrt{n}\max_{1\le i\le n}\left|\frac{1}{\hat{\pi}(X_i;\hat{\alpha}_n)} - \frac{1}{\pi(X_i;\alpha_0)}\right|\right\}^2\left\{\frac{1}{n}\mu^\top\mu + \frac{1}{n}\epsilon^\top\epsilon\right\}\\
\le&C\xi_\pi^{-1}\left\{\sqrt{n}\max_{1\le i\le n}\left|\left\{\frac{1}{\pi(X_i;\alpha)^2}\frac{\partial\pi(X_i;\alpha)}{\partial\alpha^\top}\Big|_{\alpha=\alpha_0}\right\}(\hat{\alpha}_n - \alpha_0) + o(\|\hat{\alpha}_n - \alpha_0\|)\right|\right\}^2 O_p(1)\\
\le&C\xi_\pi^{-1}\left\{\max_{1\le i\le n}\left\|\frac{\partial\pi(X_i;\alpha)}{\partial\alpha^\top}\Big|_{\alpha=\alpha_0}\right\|\sqrt{n}\|\hat{\alpha}_n - \alpha_0\| + o(\sqrt{n}\|\hat{\alpha}_n - \alpha_0\|)\right\}^2 O_p(1)\\
\le&o_p(1)\{O_p(1)O_p(1) + o_p(1)\}^2 O_p(1)\\
=&o_p(1),
\end{aligned}$$

where the third inequality is due to (A8), and the fifth inequality is obtained from Condition (C.6) and fact that $\sqrt{n}\|\hat{\alpha}_n - \alpha_0\| = O_p(1)$. And hence, the proof of Lemma 2 is completed. □

Lemma 3. *Under Conditions (C.2) to (C.6), by Lemma 1, as $n \to \infty$, it holds that*

$$\sup_{w\in W_n}\frac{\|\hat{\mu}(w) - \mu^0(w)\|^2}{L^0(w)} = o_p(1), \tag{A9}$$

$$\sup_{w\in W_n}\frac{\|\tilde{\mu}(w) - \hat{\mu}(w)\|^2}{L^0(w)} = o_p(1). \tag{A10}$$

Proof. Firstly, we begin to proof (A9). Combining with the result of Lemma 1 and Conditions (C.5) and (C.6), we have

$$\begin{aligned}
&\sup_{w\in W_n}\frac{\|\hat{\mu}(w)-\mu^0(w)\|^2}{L^0(w)}\\
\leq&\xi_\pi^{-1}\sup_{w\in W_n}\left\|\sum_{m=1}^{M}w_m\left\{\mu(\hat{\theta}_{(m)})-\mu(\theta^0_{(m)})\right\}\right\|^2\\
\leq&\xi_\pi^{-1}\max_{1\leq m\leq M}\left\|\mu(\hat{\theta}_{(m)})-\mu(\theta^0_{(m)})\right\|^2\\
\leq&\xi_\pi^{-1}\max_{1\leq m\leq M}\lambda_{\max}\left(n^{-1}\sum_{j=1}^{n}[A^{(j)}_{(m)}]^\intercal A^{(j)}_{(m)}\right)n\|\hat{\theta}_{(m)}-\theta^0_{(m)}\|^2\\
\leq&\xi_\pi^{-1}\max_{1\leq m\leq M}\max_{1\leq j\leq n}\lambda_{\max}\left([A^{(j)}_{(m)}]^\intercal A^{(j)}_{(m)}\right)n\|\hat{\theta}_{(m)}-\theta^0_{(m)}\|^2\\
=&\xi_\pi^{-1}\max_{1\leq m\leq M}\max_{1\leq j\leq n}\lambda_{\max}\left([A^{(j)}_{(m)}]^\intercal A^{(j)}_{(m)}\right)n\|\hat{\theta}_{(m)}-\theta^0_{(m)}\|^2\\
=&o_p(1),
\end{aligned}$$

where $A^{(j)}_{(m)}$ is given in Condition (C.5), and the last equation is due to Lemma 1 and Conditions (C.5) and (C.6).

Next, we prove (A10). Under Conditions (C.5) and (C.6), with the similar statements of (A9), we know that

$$\begin{aligned}
&\sup_{w\in W_n}\frac{\|\tilde{\mu}(w)-\hat{\mu}(w)\|^2}{L^0(w)}\\
\leq&\xi_\pi^{-1}\max_{1\leq m\leq M}tr([\tilde{A}_{(m)}]^\intercal\tilde{A}_{(m)})\sum_{j=1}^{n}\|\tilde{\theta}^{(-j)}_{(m)}-\theta^0_{(m)}\|^2\\
\leq&\xi_\pi^{-1}\max_{1\leq m\leq M}p_m\max_{1\leq j\leq n}\lambda_{\max}([\tilde{A}^{(j)}_{(m)}]^\intercal\tilde{A}^{(j)}_{(m)})n\|\tilde{\theta}^{(-j)}_{(m)}-\theta^0_{(m)}\|^2\\
=&o_p(1),
\end{aligned}$$

where $tr(\cdot)$ denotes the trace of a matrix, $\tilde{A}_{(m)} = diag\left(\tilde{A}^{(1)}_{(m)},\tilde{A}^{(2)}_{(m)},\cdots,\tilde{A}^{(n)}_{(m)}\right)$, and $A^{(j)}_{(m)}$ is given in Condition (C.5).

Thus, we complete the proof of Lemma 3. □

Proof of Theorem 1. By the definition of $C_{GMM-\hat{\pi}}(w)$ in (2.7), we have

$$\begin{aligned}
C_{GMM-\hat{\pi}}(w)&=\|Y_{\hat{\pi}}-\tilde{\mu}(w)\|^2\\
&=\|Y_{\hat{\pi}}-Y_\pi+Y_\pi-\tilde{\mu}(w)\|^2\\
&\leq\|Y_{\hat{\pi}}-Y_\pi\|^2+\|Y_\pi-\tilde{\mu}(w)\|^2+2\sqrt{\|Y_{\hat{\pi}}-Y_\pi\|^2\|Y_\pi-\tilde{\mu}(w)\|^2}.
\end{aligned}$$

Thus, by Lemma 2 and the idea in [8], Theorem 1 is valid if

$$\sup_{w\in W_n}\frac{\|Y_\pi-\tilde{\mu}(w)\|^2}{L^0(w)}=1+o_p(1). \tag{A11}$$

By some careful matrix manipulation and Cauchy-Schwartz inequality, we can obtain

$$\begin{aligned}
&||Y_\pi - \tilde{\mu}(w)||^2 \\
=&||(Y_\pi - \mu) - (\hat{\mu}(w) - \mu) - (\tilde{\mu}(w) - \hat{\mu}(w))|| \\
=&||Y_\pi - \mu||^2 + ||\hat{\mu}(w) - \mu||^2 + ||\tilde{\mu}(w) - \hat{\mu}(w)||^2 - 2(\hat{\mu}(w) - \mu)^\intercal (Y_\pi - \mu) \\
&+ 2(\hat{\mu}(w) - \mu)^\intercal (\tilde{\mu}(w) - \hat{\mu}(w)) - 2(\tilde{\mu}(w) - \hat{\mu}(w))^\intercal (Y_\pi - \mu) \\
\leq&||Y_\pi - \mu||^2 + ||\hat{\mu}(w) - \mu||^2 + ||\tilde{\mu}(w) - \hat{\mu}(w)||^2 + 2\sqrt{||\hat{\mu}(w) - \mu||^2 ||\tilde{\mu}(w) - \hat{\mu}(w)||^2} \\
&+ 2\sqrt{||\hat{\mu}(w) - \mu||^2 ||Y_\pi - \mu||^2} + 2\sqrt{||\tilde{\mu}(w) - \hat{\mu}(w)||^2 ||Y_\pi - \mu||^2}.
\end{aligned}$$

Remove the item that is not related to w, to prove (A11), we only need to prove that

$$\sup_{w \in W_n} \frac{||\hat{\mu}(w) - \mu||^2}{L^0(w)} = 1 + o_p(1), \tag{A12}$$

$$\sup_{w \in W_n} \frac{||\tilde{\mu}(w) - \hat{\mu}(w)||^2}{L^0(w)} = o_p(1), \tag{A13}$$

$$\sup_{w \in W_n} \frac{|(\hat{\mu}(w) - \mu)^\intercal \epsilon_\pi|}{L^0(w)} = o_p(1), \tag{A14}$$

$$\sup_{w \in W_n} \frac{|(\tilde{\mu}(w) - \hat{\mu}(w))^\intercal \epsilon_\pi|}{L^0(w)} = o_p(1), \tag{A15}$$

where $\epsilon_\pi = Y_\pi - \mu$. We begin to prove (A12) at first. By Cauchy-Schwarz inequality, it is easy to show that

$$\begin{aligned}
||\hat{\mu}(w) - \mu||^2 =& ||\hat{\mu}(w) - \mu^0(w) + \mu^0(w) - \mu||^2 \\
=& L^0(w) + ||\hat{\mu}(w) - \mu^0(w)||^2 + 2(\hat{\mu}(w) - \mu^0(w))^\intercal (\mu^0(w) - \mu) \\
\leq& L^0(w) + ||\hat{\mu}(w) - \mu^0(w)||^2 + 2\sqrt{L^0(w) ||\hat{\mu}(w) - \mu^0(w)||^2},
\end{aligned}$$

where $L^0(w) = ||\mu^0(w) - \mu||^2$. Thus, by condition (C.6) and (A9) in Lemma 3, we have

$$\begin{aligned}
&\sup_{w \in W_n} \frac{||\hat{\mu}(w) - \mu||^2}{L^0(w)} \\
\leq& 1 + \sup_{w \in W_n} \frac{||\hat{\mu}(w) - \mu^0(w)||^2}{L^0(w)} + 2\sqrt{\sup_{w \in W_n} \frac{L^0(w)}{L^0(w)} \sup_{w \in W_n} \frac{||\hat{\mu}(w) - \mu^0(w)||^2}{L^0(w)}} \\
=& 1 + o_p(1) + 2\sqrt{o_p(1)}.
\end{aligned}$$

Thus, we complete the proof of (A12).

As for (A13), note that

$$||\tilde{\mu}(w) - \hat{\mu}(w)||^2$$

$$
\begin{aligned}
&=\|\tilde{\mu}(w)-\mu^0(w)-(\hat{\mu}(w)-\mu^0(w))\|^2\\
&\leq\|\hat{\mu}(w)-\mu^0(w)\|^2+\|\tilde{\mu}(w)-\mu^0(w)\|^2+2\sqrt{\|\tilde{\mu}(w)-\mu^0(w)\|^2\|\hat{\mu}(w)-\mu^0(w)\|^2},
\end{aligned}
$$

with similar steps of proving (A12), by Lemma 3, we can obtain that (A13) holds.

Next, we begin to prove (A14) and (A15). Recalling the definition of ϵ_π in (2.3), together with Condition (C.3) and the triangle inequality, for $i = 1, 2, \cdots, n$, we have

$$
\begin{aligned}
&|\epsilon_{\pi,i}|\\
=&\left|\frac{\delta_i Y_i}{\pi(X_i;\alpha_0)}-\mu_i\right|\\
=&\left|\frac{\delta_i Y_i}{\pi(X_i;\alpha_0)}-\frac{\delta_i\mu_i}{\pi(X_i;\alpha_0)}+\frac{\delta_i\mu_i}{\pi(X_i;\alpha_0)}-\mu_i\right|\\
\leq&c_\pi^{-1}|\epsilon_i|+(c_\pi^{-1}+1)|\mu_i|.
\end{aligned}
$$

Then, by Conditions (C.1) and (C.2) and the C_r inequality, for a fixed $1 \leq G < \infty$, it is clear that

$$
\max_{1\leq i\leq n} E[\epsilon_{\pi,i}^{2G}|X_i] \leq C, \quad \max_{1\leq i\leq n} E[\epsilon_{\pi,i}^{2G}] \leq C. \tag{A16}
$$

Further, it is easy to show that $\|\epsilon_\pi\| = O_p(n^{1/2})$.

Note that,

$$
\begin{aligned}
&\sup_{w\in W_n}\frac{|(\hat{\mu}(w)-\mu)^\tau\epsilon_\pi|}{L^0(w)}\\
\leq&\sup_{w\in W_n}\frac{|(\hat{\mu}(w)-\mu^0(w))^\tau\epsilon_\pi|}{L^0(w)}+\sup_{w\in W_n}\frac{|(\mu^0(w)-\mu)^\tau\epsilon_\pi|}{L^0(w)}.
\end{aligned}
$$

Thus, to prove (A14), it suffices to verify that

$$
\sup_{w\in W_n}\frac{|(\mu^0(w)-\mu)^\tau\epsilon_\pi|}{L^0(w)} = o_p(1), \tag{A17}
$$

$$
\sup_{w\in W_n}\frac{|(\hat{\mu}(w)-\mu^0(w))^\tau\epsilon_\pi|}{L^0(w)} = o_p(1). \tag{A18}
$$

We firstly prove (A17). By Chebyshev's inequality, Markov inequality, for any $\kappa > 0$ and $1 \leq G < \infty$, we have

$$
\begin{aligned}
&\Pr\left\{\sup_{w\in W_n}\frac{|(\mu^0(w)-\mu)^\tau\epsilon_\pi|}{L^0(w)}\geq\kappa\right\}\\
\leq&\Pr\left\{\sup_{w\in W_n}|(\mu^0(w)-\mu)^\tau\epsilon_\pi|\geq\kappa\xi_\pi\right\}\\
\leq&\Pr\left\{\max_{1\leq m\leq M}|(\mu^0_{(m)}-\mu)^\tau\epsilon_\pi|\geq\kappa\xi_\pi\right\}\\
\leq&\sum_{m=1}^{M}\Pr\left\{|(\mu^0_{(m)}-\mu)^\tau\epsilon_\pi|\geq\kappa\xi_\pi\right\}
\end{aligned}
$$

$$
\begin{aligned}
&\leq \sum_{m=1}^{M} E\left\{\frac{[(\mu_{(m)}^0-\mu)^\tau \epsilon_\pi]^{2G}}{\kappa^{2G}\xi_\pi^{2G}}\right\}\\
&\leq C\kappa^{-2G}\xi_\pi^{-2G}\sum_{m=1}^{M}\left\{\sum_{i=1}^{n}|\mu_{(m),i}^0-\mu_i|^2\left[E(\epsilon_{\pi,i}^{2G})\right]^{1/G}\right\}^G\\
&\leq C\kappa^{-2G}\xi_\pi^{-2G}\max_{1\leq i\leq n}\left[E(\epsilon_{\pi,i}^{2G})\right]\sum_{m=1}^{M}\|\mu_{(m)}^0-\mu\|^{2G}\\
&\leq C\kappa^{-2G}M\xi_\pi^{-2G}O(n^G),
\end{aligned}
$$

where the forth inequality is due to the Chebyshev's inequality and the seventh inequality is because of (A16) and Condition (C.2). Then, according to Condition (C.6), we arrive at the result of (A17).

Next, we turn to (A18). By the proof of Lemma 3, it is not hard to find that

$$\sup_{w\in W_n}\|\hat{\mu}(w)-\mu^0(w)\|^2 = O_p(1), \tag{A19}$$

$$\sup_{w\in W_n}\|\tilde{\mu}(w)-\mu^0(w)\|^2 = O_p(1). \tag{A20}$$

In fact, the results of equations (A19) and (A20) are also given in [20]. Then, under Condition (C.6), combining with (A16), (A19) and Cauchy-Schwarz inequality, we can obtain

$$
\begin{aligned}
&\sup_{w\in W_n}\frac{|(\hat{\mu}(w)-\mu^0(w))^\tau\epsilon_\pi|}{L^0(w)}\\
\leq&\xi_\pi^{-1}\sup_{w\in W_n}\|\hat{\mu}(w)-\mu^0(w)\|\|\epsilon_\pi\|\\
\leq&\xi_\pi^{-1}O_p(1)O_p(n^{1/2})\\
=&o_p(1).
\end{aligned} \tag{A21}
$$

Finally, we consider the proof of (A18). Similarly, under Condition (C.6), by (A20) and (A21), we have

$$
\begin{aligned}
&\sup_{w\in W_n}\frac{|(\tilde{\mu}(w)-\hat{\mu}(w))^\tau\epsilon_\pi|}{L^0(w)}\\
=&\sup_{w\in W_n}\frac{|\{(\tilde{\mu}(w)-\mu^0(w))-(\hat{\mu}(w)-\mu^0(w))\}^\tau\epsilon_\pi|}{L^0(w)}\\
\leq&\sup_{w\in W_n}\frac{|(\tilde{\mu}(w)-\mu^0(w))^\tau\epsilon_\pi|}{L^0(w)}+\sup_{w\in W_n}\frac{|(\hat{\mu}(w)-\mu^0(w))^\tau\epsilon_\pi|}{L^0(w)}\\
\leq&\xi_\pi^{-1}\sup_{w\in W_n}\|\tilde{\mu}(w)-\mu^0(w)\|\|\epsilon_\pi\|+o_p(1)\\
=&o_p(1).
\end{aligned}
$$

Until now, the proof of Theorem 1 is completed.

References

1. T. Ando, K. C. Li, A model-averaging approach for high-dimensional regression, *J. Am. Stat. Assoc.*, **109** (2014), 254–265. https://doi.org/10.1080/01621459.2013.838168
2. T. Ando, K. C. Li, A weight-relaxed model averaging approach for high-dimensional generalized linear models, *Ann. Stat.*, **45** (2017), 2654–2679. https://doi.org/10.1214/17-aos1538
3. F. J. DiTraglia, J. Francis, Using invalid instruments on purpose: Focused moment selection and averaging for GMM, *J. Econ.*, **195** (2016), 187–208. https://doi.org/10.1016/j.jeconom.2016.07.006
4. B. E. Hansen, Least squares model averaging, *Econometrica*, **75** (2007), 1175–1189. https://doi.org/10.1111/j.1468-0262.2007.00785.x
5. B. E. Hansen, J. S. Racine, Jackknife model averaging, *J. Econ.*, **167** (2012), 38–46. https://doi.org/10.1111/j.1747-4477.2012.00344.x
6. L. P. Hansen, Large sample properties of generalized method of moments estimators, *Econometrica*, **50** (1982), 1029–1054. https://doi.org/10.2307/1912775
7. N. Hens, M. Aerts, G. Molenberghs, Model selection for incomplete and design-based samples, *Stat. Med.*, **25** (2006), 2502–2520. https://doi.org/10.1002/sim.2559
8. K. C. Li, Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: Discrete index set, *Ann. Stat.*, **15** (1987), 958–975. https://doi.org/10.1214/aos/1176350486
9. J. Liao, G. Zou, Corrected mallows citerion for model averaging, *Comput. Stat. Data Anal.*, **144** (2020), 106902. https://doi.org/10.1016/j.csda.2019.106902
10. R. J. A. Little, D. B. Rubin, *Statistical analysis with missing data (2Eds.)*, 2002, Hoboken, NJ: Wiley. https://doi.org/10.1002/9781119013563
11. Q. Liu, R. Okui, Heteroscedasticity-robust C_p model averaging, *Econ. J.*, **16** (2013), 463–472. https://doi.org/10.1111/ectj.12009
12. Q. Liu, R. Okui, A. Yoshimura, Generalized least squares model averaging, *Econ. Rev.*, **35** (2016), 1692–1752. https://doi.org/10.1080/07474938.2015.1092817
13. N. T. Longford, Model Selection and efficiency: is 'which model...?' the right question? *J. R. Stat. Soc. A Stat.*, **168** (2005), 469–472. https://doi.org/10.1111/j.1467-985x.2005.00366.x
14. R. Luo, Q. Wang, Empirical likelihood based weighted GMM estimation with missing response at random, *J. Stat. Plan. Infer.*, **156** (2015), 64–79. https://doi.org/10.1016/j.jspi.2014.07.011
15. A. Raftery, Y. Zheng, Discussion: Performance of bayesian model averaging, *J. Am. Stat. Assoc.*, **98** (2003), 931–938. https://doi.org/10.1198/016214503000000891
16. Z. Sun, Z. Su, J. Ma, Focused vector information criterion model selection and model averaging regression with missing response, *Metrika*, **77** (2014), 415–432. https://doi.org/10.1007/s00184-013-0446-8

17. A. T. K. Wan, X. Zhang, G. Zou, Least squares model averaging by mallows criterion, *J. Econ.*, **156** (2010), 277–283. https://doi.org/10.1016/j.jeconom.2009.10.030

18. Q. Wang, J. N. K. Rao, Empirical likelihood-based inference under imputation for missing response data, *Ann. Stat.*, **30** (2002), 896–924. https://doi.org/10.1214/aos/1028674845

19. Q. Wang, Z. Sun, Estimation in partially linear models with missing responses at random, *J. Multivariate Anal.*, **98** (2007), 1470–1493. https://doi.org/10.1016/j.jmva.2006.10.003

20. W. Wang, Q. Zhang, X. Zhang, X. Li, Model averaging based on generalized method of moments, *Econ. Lett.*, **200** (2021), 109735. https://doi.org/10.1016/j.econlet.2021.109735

21. Y. Wei, Q. Wang, Cross-validation-based model averaging in linear models with response missing at random, *Stat. Prob. Lett.*, **171** (2021), 108990. https://doi.org/10.1016/j.spl.2020.108990

22. J. Xie, X. Yan, N. Tang, A model averaging method for high-dimensional regression with missing responses at random, *Stat. Sinica*, **31** (2021), 1005–1026. https://doi.org/10.5705/ss.202018.0297

23. X. Zhang, Model averaging by least squares approximation, *Sci. Sinica Math.*, **51** (2021), 535–548. https://doi.org/10.1360/n012019-00137

24. X. Zhang, Optimal model averaging based on generalized method of moments, *Stat. Sinica*, **31** (2021), 2103–2122. https://doi.org/10.5705/ss.202019.0230

25. X. Zhang, H. Liang, Focused information criterion and model averaging for generalized additive partial linear models, *Ann. Stat.*, **39** (2011), 174–200. https://doi.org/10.1214/10-aos832

26. X. Zhang, W. Wang, Optimal model averaging estimation for partially linear models, *Stat. Sinica*, **29** (2019), 693–718. https://doi.org/10.5705/ss.202015.0392

27. X. Zhang, D. Yu, G. Zou, H. Liang, Optimal model averaging estimation for generalized linear models and generalized linear mixed-effects models, *J. Am. Stat. Assoc.*, **111** (2016), 1775–1790. https://doi.org/10.1080/01621459.2015.1115762

28. R. Zhu, A. T. K. Wan, X. Zhang, G. Zou, A mallows-type model averaging estimator for the varying-coefficient partially linear model, *J. Am. Stat. Assoc.*, **114** (2019), 882–892. https://doi.org/10.1080/01621459.2018.1456936

AIMS Press