



# Robust model selection with covariables missing at random

Zhongqi Liang<sup>1</sup> · Qihua Wang<sup>1,2</sup> · Yuting Wei<sup>3</sup>

Received: 30 November 2020 / Revised: 21 July 2021 / Accepted: 3 August 2021 /

Published online: 25 August 2021

© The Institute of Statistical Mathematics, Tokyo 2021

## Abstract

Let  $f_{Y|X,Z}(y|x, z)$  be the conditional probability function of  $Y$  given  $(X, Z)$ , where  $Y$  is the scalar response variable, while  $(X, Z)$  is the covariable vector. This paper proposes a robust model selection criterion for  $f_{Y|X,Z}(y|x, z)$  with  $X$  missing at random. The proposed method is developed based on a set of assumed models for the selection probability function. However, the consistency of model selection by our proposal does not require these models to be correctly specified, while it only requires that the selection probability function is a function of these assumed selective probability functions. Under some conditions, it is proved that the model selection by the proposed method is consistent and the estimator for population parameter vector is consistent and asymptotically normal. A Monte Carlo study was conducted to evaluate the finite-sample performance of our proposal. A real data analysis was used to illustrate the practical application of our proposal.

**Keywords** Model selection · Inverse probability weight · Model misspecification · Missing at random · Kullback–Leibler divergence · Robust

---

✉ Qihua Wang  
qhwang@amss.ac.cn

Zhongqi Liang  
2945155436@qq.com

Yuting Wei  
ytwei@mail.usc.edu.cn

<sup>1</sup> School of Statistics and Mathematics, Zhejiang Gongshang University, Hangzhou 310018, Zhejiang, China

<sup>2</sup> Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China

<sup>3</sup> Department of Statistics and Finance, University of Science and Technology of China, Hefei 230026, China

## 1 Introduction

Model selection has been a hot spot in statistical analysis for a long time. Significant practicality and theoretical advances have been made in the field of model selection over the past five decades. The main advances, however, focus on the case where data are observed completely (see, e.g., Mallow 1973; Claeskens and Hjort, 2003; Jiang et al., 2008; Rolling and Yang, 2014; Shao and Yang, 2017; Zhang et al., 2017), while less attention has been paid to the case of missing data (see, e.g., Hens et al., 2006; Claeskens and Consentino, 2008; Ibrahim et al., 2008; Jiang et al., 2015; Wei et al., 2021). Missing data occur commonly in the study of many practical problems such as socioeconomic research, medical research, observational research and so on. Therefore, it is of great practical importance to develop model selection strategies which are applicable to missing data. This paper considers the model selection problem in the presence of missing data.

In many scientific areas, a basic task is to assess the simultaneous influence of several factors (covariates) on a quantity of interest (response variable) (Wang and Rao, 2002a). In this paper, we are interested in the model selection problem for the conditional probability function of  $Y$  given  $(X, Z)$ , denoted by  $f_{Y|X,Z}(y|x, z)$ , where  $Y$  is the scalar response variable, while  $(X, Z)$  is the covariable vector. Throughout this paper, a generic notation  $f_{V_1|V_2}(v_1|v_2)$  is utilized to denote the conditional probability function of the variable  $V_1$  given the variable  $V_2$ . Missing data mechanism plays an important role in the study of missing data problems. In this paper, we consider the case where  $X$  is missing at random (MAR) which has been commonly assumed in the analysis of missing data. See, e.g., Little and Rubin (2002). In short, this paper considers the model selection problem for  $f_{Y|X,Z}(y|x, z)$  with  $X$  missing at random.

There are several alternative model selection approaches that can be applied for the considered model selection problem. These approaches can be roughly classified into the following four types. The first type of approach is developed based on Bayesian point of view (see, e.g., Celeux et al., 2006; Gelman et al., 2005). A drawback of this type is that it needs to set prior distributions for parameters in the candidate models. In general, it is not easy to obtain reasonable priors and meanwhile make sure that these priors are not in conflict with each other. The second type of approach is established based on the expectation-maximization (EM) algorithm (see, e.g., Claeskens and Consentino, 2008; Ibrahim et al., 2008; Jiang et al., 2015). A drawback of this type is that it requires a correctly specified parametric model for the condition probability function  $f_{X|Z}(x|z)$ . In practice, it may be hard to get a correct specification for  $f_{X|Z}(x|z)$ , not mention  $X$  is missing. The third type is the bias corrected model selection criterion of Wei et al. (2021), and it does not need that  $f_{X|Y,Z}(x|y, z)$  is specified correctly. Hence, it improves the model selection approaches based on EM algorithm. However, it requires a correctly specified parametric model for the selection probability function. The final type of approach is the weighted AIC (Hens et al., 2006) which is developed based on the inverse probability weighted (IPW) method (Horvitz and Thompson, 1952; Robins et al., 1994), a widely used method in the study of missing data

problems. Comparing to the previous three types of approaches, the advantages of the weighted AIC are that it is a frequentist model selection approach so that it does not need to set prior distributions, and its calculation is irrelevant to  $f_{X|Z}(x|z)$  or  $f_{X|Y,Z}(x|y, z)$ , so that it does not need to specify parametric model for  $f_{X|Z}(x|z)$  or  $f_{X|Y,Z}(x|y, z)$ . However, a drawback of the weighted AIC is that it requires a consistent estimate of the selection probability function. If the selection probability function is estimated parametrically, then a correctly specified parametric model for the selection probability function is required. And if the selection probability function is estimated nonparametrically, the problem of “curse of dimension” occurs. Unfortunately, the theoretical properties of the weighted AIC are lacked in Hens et al. (2006).

This paper proposes a robust inverse probability weighting Kullback–Leibler divergence (RIPW-KL) criterion-based method by estimating the weight function semiparametrically. The main idea of this method can be described as follows. We first postulate several possible parametric models for the selection probability function and then combine the information contained in these estimated parametric models through a nonparametric smoothing method to get a semiparametric estimator of a conditional probability function conditional on the assumed models. With the inverse of this estimator as the weight, an inverse probability weighted estimator of the Kullback–Leibler divergence can be obtained immediately. And further, the proposed criterion can be established by minimizing the weighted Kullback–Leibler divergence with a suitable penalty term. Under some conditions, we prove that the model selection by our proposed criterion is consistent and the estimator of population parameter vector corresponding to the selected model is asymptotically normal. Unlike the weighted AIC which makes a choice between parametric and nonparametric methods, our semiparametric weighting criterion makes a balance between these two kinds of methods and thus alleviates the problems of model misspecification and “curse of dimension” simultaneously. Concretely, under certain conditions, it is shown that the consistency of model selection by our proposal is valid as long as the selection probability function is a function of its assumed models. Hence, comparing to the parametrically weighted AIC, our proposed RIPW-KL criterion is more robust to misspecification of the selection probability function. This alleviates the problem of model misspecification. Besides, for the purpose of weights, our proposal applies the kernel smoothing method to estimate regression on the assumed models for the selection probability function, while the nonparametrically weighted AIC applies the kernel smoothing method to estimate the selection probability regression function on the observed variables. Usually, the number of these assumed models is far smaller than the number of the observed variables, and thus, the proposed method also alleviates the problem of “curse of dimension”. That is, the proposed method has following significant advantages over the existing methods:

1. The RIPW-KL criterion does not need to set prior distributions as Bayesian methods;
2. The RIPW-KL criterion does not need a correct specification for  $f_{X|Z}(x|z)$  or  $f_{X|Y,Z}(x|y, z)$  comparing to EM algorithm.

3. The RIPW-KL criterion does not need a correct specification of the selection probability function as the parametrically weighted AIC method due to Hens et al. (2006) and the corrected method due to Wei et al. (2021).
4. The RIPW-KL criterion alleviates the problem of “curse of dimension” comparing to the nonparametrically weighted AIC method due to Hens et al. (2006).

The rest of this paper is organized as follows. In Sect. 2, we describe the model framework and then establish the RIPW-KL criterion. In Sect. 3, the theoretical properties of our proposal are presented, and the finite-sample performance of our proposal is investigated through a Monte Carlo study in Sect. 4. And a real data analysis is implemented in Sect. 5. And all the technical details are relegated in the “Appendix”.

## 2 Methodology

Let  $\{(Y_i, X_i, Z_i, \delta_i), 1 \leq i \leq n\}$  be the independent and identically distributed sample from  $(Y, X, Z, \delta)$ , where  $Y_i$  and  $Z_i$  are observed completely, and  $\delta_i = 0$  if  $X_i$  is missing, otherwise  $\delta_i = 1$ . Throughout this paper, we assume that  $X$  is missing at random (MAR), that is,

$$P(\delta = 1|Y, X, Z) = P(\delta = 1|Y, Z) := \pi(Y, Z).$$

Suppose that a finite set of candidate parametric models can be obtained for  $f_{Y|X,Z}(y|x, z)$ . For the candidate model  $M$ , we have a parametric model  $g_M(y|x, z; \theta_M)$ , where  $\theta_M$  is an unknown parameter vector, while  $g_M(\cdot|x, z; \theta_M)$  is a known function. It’s well known that the Kullback–Leibler (KL) divergence is a widely used measure on the closeness between the assumed parametric model and the true model that generates data. For the candidate model  $M$ , the KL divergence from  $f_{Y|X,Z}(y|x, z)$  to  $g_M(y|x, z; \theta_M)$  is

$$KL(M, \theta_M) = E\{\log f_{Y|X,Z}(Y|X, Z)\} - E\{\log g_M(Y|X, Z; \theta_M)\}. \tag{1}$$

Clearly, we take  $E\{\log g_M(Y|X, Z; \theta_M)\}$  to measure the closeness of  $f_{Y|X,Z}(y|x, z)$  to  $g_M(y|x, z; \theta_M)$  and denote it by  $D(M, \theta_M)$ , since  $E\{\log f_{Y|X,Z}(Y|X, Z)\}$  is irrelevant to the candidate models. Clearly, the larger  $D(M, \theta_M)$  is, the smaller KL divergence is.

We assume a parametric model  $\pi(y, z; \bar{\alpha})$  for the selection probability function  $P(\delta = 1|y, z)$ , where  $\bar{\alpha}$  is the unknown parameter vector. Inspired by Hens et al. (2006), an inverse probability weighting estimator of  $D(M, \theta_M)$  is given by

$$\hat{D}(M, \hat{\alpha}, \theta_M) = \frac{1}{n} \sum_{i=1}^n \frac{\delta_i}{\pi(Y_i, Z_i; \hat{\alpha})} \log g_M(Y_i|X_i, Z_i; \theta_M),$$

where  $\hat{\alpha}$  is the maximum likelihood estimate (MLE) of  $\bar{\alpha}$ .

For writing convenience, we abbreviate  $\pi(y, z; \bar{\alpha})$  to  $\pi(\bar{\alpha})$ . Obviously, if  $\pi(\bar{\alpha})$  is misspecified,  $\hat{D}(M, \hat{\alpha}, \theta_M)$  defines an inconsistent estimator for  $D(M, \theta_M)$ . In this case, one may use the augmented inverse probability weighting approach. This

method, however, needs that  $f_{X|Y,Z}(x|y, z)$  is specified correctly when  $\pi(\bar{\alpha})$  is misspecified. Nevertheless, it is impractical to specify  $f_{X|Y,Z}(x|y, z)$  correctly since  $f_{Y|X,Z}(y|x, z)$  is unknown, not to mention that  $X$  is missing. Another strategy is to use the nonparametric kernel method to estimate the selecting probability function. However, this approach leads to ‘‘curse of dimension’’ problem if the dimension of  $(Y, Z)$  is large. These motivate us to develop a robust model selection approach based on the inverse probability weighting KL divergence with a semiparametric weight.

Specify multiple possible parametric models  $\{\pi_j(y, z; \alpha_j), j = 1, \dots, J\}$  for  $\pi(y, z)$ . Define  $\phi_\pi(y, z; \alpha) = (\pi_1(y, z; \alpha_1), \dots, \pi_J(y, z; \alpha_J))^\top$ , where  $\alpha = (\alpha_1^\top, \dots, \alpha_J^\top)^\top$ . Usually, the number of the assumed models  $J$  is less than the dimension of the observed variables.

Let

$$\hat{p}_{\alpha,n}(v) = \frac{\sum_{l=1}^n K\left(\frac{\{v - \phi_\pi(Y_l, Z_l; \alpha)\}}{h_n}\right) \delta_l}{\sum_{l=1}^n K\left(\frac{\{v - \phi_\pi(Y_l, Z_l; \alpha)\}}{h_n}\right)}, \tag{2}$$

and

$$\hat{r}_{\alpha,n}(v) = \frac{1}{nh_n^J} \sum_{l=1}^n K\left(\frac{\{v - \phi_\pi(Y_l, Z_l; \alpha)\}}{h_n}\right) \delta_l, \tag{3}$$

where  $h_n$  is a scalar bandwidth,  $K(\cdot)$  is the multivariate kernel function. Then, a robust estimator of  $D(M, \theta_M)$  can be defined as follows:

$$\hat{D}_{IP}(M, \theta_M, \hat{\alpha}_n) = \frac{1}{n} \sum_{i=1}^n \left\{ \frac{\delta_i}{\hat{p}_{\hat{\alpha}_n, b_n}(\phi_\pi(Y_i, Z_i; \hat{\alpha}_n))} \log g_M(Y_i|X_i, Z_i; \theta_M) \right\}, \tag{4}$$

where  $\hat{\alpha}_n$  is the MLE of  $\alpha$ ,  $\hat{p}_{\alpha, b_n}(v) = \hat{p}_{\alpha,n}(v) \hat{r}_{\alpha, b_n}(v) / \hat{r}_{\alpha,n}(v)$ ,  $b_n$  is a positive constant sequence tending to 0, and  $\hat{r}_{\alpha, b_n}(v) = \max\{\hat{r}_{\alpha,n}(v), b_n\}$ . It is obvious that  $\hat{D}_{IP}(M, \theta_M, \hat{\alpha}_n)$  is a dimension reduction estimation comparing to the nonparametric inverse probability weighting approach since  $J$  is usually taken less than the dimension of  $(Y, Z)$ . Specially,  $J$  is taken to be 1 if one can specify a correct model for  $\pi(y, z)$ .

Further, let us denote the maximizer of  $\hat{D}_{IP}(M, \theta_M, \hat{\alpha}_n)$  with respect to  $\theta_M$  as  $\hat{\theta}_M^{IP}$ , that is,

$$\hat{\theta}_M^{IP} = \arg \max_{\theta_M \in \Theta_M} \hat{D}_{IP}(M, \theta_M, \hat{\alpha}_n), \tag{5}$$

where  $\Theta_M$  is the parameter space of  $\theta_M$ . Then, the proposed RIPW-KL criterion is given by

$$IC_{IP}(M) = -\hat{D}_{IP}(M, \hat{\theta}_M^{IP}, \hat{\alpha}_n) + \lambda_n d_M, \tag{6}$$

where  $d_M$  is the dimension of the unknown parameter vector  $\theta_M$  and  $\lambda_n$  is a positive tuning parameter tending to zero. Define

$$\hat{M}_{IP} = \arg \min_{M \in \mathcal{M}} IC_{IP}(M), \tag{7}$$

where  $\mathcal{M}$  is the set of all the candidate models. Then,  $\hat{M}_{IP}$  is the selected model and the parameter vector can be estimated using  $\hat{\theta}_{\hat{M}_{IP}}^{IP}$ .

### 3 Theoretical properties

Before giving the main results, we first give some notations and some required conditions. To prove the consistency of model selection by the proposed criterion, firstly, we need to prove the asymptotic normality of the estimator  $\hat{\theta}_M^{IP}$  in (5) for each candidate model  $M$ . Let  $f_\alpha(\cdot)$  be the density function of  $\phi_\pi(y, z; \alpha)$ . Clearly,  $\hat{r}_{\alpha, n}(v)$  in (3) is the  $J$ -dimensional kernel density estimate of  $f_\alpha(v)$ . And denote  $r_\alpha(v) = p_\alpha(v)f_\alpha(v)$ , where  $p_\alpha(v) = E\{\delta \mid \phi_\pi(Y, Z; \alpha) = v\}$ . Then,  $\hat{p}_{\alpha, n}(v)$  in (2) is the nonparametric regression estimate of  $p_{\alpha^*}(v)$  where  $\alpha^*$  is the probability limit of  $\hat{\alpha}_n$ . And denote  $\|\cdot\|$  as Euclidean norm. Now, we present the required conditions as follows.

- (C.1)  $\Theta_M$  is a compact set. And the KL divergence  $KL(M, \theta_M)$  has a unique minimum point at  $\theta_M^*$ , where  $\theta_M^*$  is an inner point of  $\Theta_M$ .
- (C.2)  $E\left\{-\frac{\partial^2 \log g_M(Y|X, Z; \theta_M)}{\partial \theta_M \partial \theta_M^T} \Big|_{\theta_M = \theta_M^*}\right\}$  is positive definite.  $\frac{\partial^3 \log g_M(y|x, z; \theta_M)}{\partial \theta_M^3}$  is continuous with respect to  $\theta_M$ . And  $E\left\{\sup_{\theta_M \in \Theta_M} \log^2 g_M(Y|X, Z; \theta_M)\right\} < \infty$ .
- (C.3)  $p_\alpha(v), f_\alpha(v)$  have bounded partial derivatives up to order  $k(k > J)$ .
- (C.4)  $\inf_{y, z} p(\delta = 1|Y = y, Z = z) > 0$ .
- (C.5) (i)  $\|\phi_\pi(y, z; \alpha) - \phi_\pi(y, z; \alpha')\| \leq l(y, z)\|\alpha - \alpha'\|$ , with  $El(Y, Z) \leq \infty$ . (ii)  $E[\sup_\alpha \|\nabla_\alpha p_\alpha(\phi_\pi(Y, Z; \alpha))\|] < \infty$  and  $E[\sup_\alpha \|\nabla_\alpha r_\alpha(\phi_\pi(Y, Z; \alpha))\|] < \infty$ . (iii)  $\sup_{z, x} E[Y^2|Z = z, X = x] < \infty$ .
- (C.6) The multivariate function  $K(v)$  is bounded and continuous kernel function of order  $k(k > J)$  defined on the compact support.
- (C.7)  $b_n$  is a constant sequence satisfying  $\sqrt{nh_n^{\kappa+1}b_n^2}/\log n \rightarrow \infty, \sqrt{nh_n^\kappa/b_n^2} \rightarrow 0, \kappa > J + 1$ .
- (C.8)  $\sqrt{n}E\left[\left|\frac{1}{p_{\alpha^*}(\phi_\pi(Y, Z; \alpha^*))}\right|I[f_{\alpha^*}(\phi_\pi(Y, Z; \alpha^*)) < b_n]\right] \rightarrow 0$ , where  $I$  is the indicator function.

**Remark** Condition (C.1) is the same as Condition 3 and Condition 5 of Fang and Shao (2016), which is an identifiability condition in model selection. Condition (C.2) defines the smoothness of  $\log g_M(Y|X, Z; \theta_M)$ . Condition (C.5)(i) controls the complexity of the multiple possible models  $\phi_\pi(y, z; \alpha)$  for  $\pi(y, z)$ , which is identical to the Condition (C.3) in Wang et al. (2021). Condition (C.7) is a regular condition in nonparametric regression (see, Condition (C.h<sub>n</sub>) in Wang and Rao (2002a) and Conditions (C.5), (C.6) in Wang et al. (2021). Condition (C.8) is similar to Condition (C.gmb<sub>n</sub>) in Wang and Rao (2002a), which controls the rates of  $b_n$  tending to zero.

**Theorem 1** *Under Conditions (C.1)–(C.8), we have  $\sqrt{n}(\hat{\theta}_M^{\text{IP}} - \theta_M^*)$  is asymptotically mean-zero normal if  $\pi(y, z)$  is a function of  $\phi_\pi(y, z; \alpha^*)$ .*

Before establishing the consistency of model selection of the proposed method, we first present the definition of correct model. If  $g_M(y|x, z; \theta_M)$  is a correctly specified model for  $f_{Y|X,Z}(y|x, z)$ , then we say that the model  $M$  is a correct model. Otherwise, we say that the model  $M$  is an incorrect one. Let  $M_{\text{opt}}$  be the optimal model with the smallest dimension among all correct models in the class of candidate models. To prove the consistency of model selection by our proposal, we assume that there is at least one correct model among the class of candidate models. Such an assumption guarantees the existence of  $M_{\text{opt}}$  and is commonly used in the model selection literature such as Jiang et al. (2015); Fang and Shao (2016) and Wei et al. (2021).

**Theorem 2** *Under Conditions (C.1)–(C.8), if  $\lambda_n \rightarrow 0$ ,  $\sqrt{n}\lambda_n \rightarrow \infty$  and satisfies that  $\pi(y, z)$  is a function of  $\phi_\pi(y, z; \alpha^*)$ , then we have  $P(\hat{M}_{\text{IP}} = M_{\text{opt}}) \rightarrow 1$  as  $n \rightarrow \infty$ .*

Theorem 2 indicates that the model selection by our proposed criterion is consistent. Our proposed criterion is robust to the model specifications for  $\pi(y, z)$  as long as  $\pi(y, z)$  is a function of  $\phi_\pi(y, z; \alpha^*)$ . In addition, according to Theorem 2, it is clear that Theorem 1 is still valid if  $M$  in Theorem 1 is replaced by  $\hat{M}_{\text{IP}}$ .

There are many choices of  $\lambda_n$  that meet its restrictions given in Theorem 2. This arouses the interest of a referee in the question of how to select  $\lambda_n$ . As pointed out by Fang and Shao (2016), no optimal solution has been derived for this question in the model selection literature even for the case of no missing data. In this article, we mainly put our efforts into developing model selection methods with consistency of model selection rather than finding the best ways of choosing  $\lambda_n$ , because the latter is to some extent outside the scope of this article and we leave it for future study. At last, it should be pointed out that, following the strategy taken in Fang and Shao (2016), the tuning parameter  $\lambda_n$  was set to be  $\{0.25 \log \log n\}^{0.25} n^{-1/2}$  in the following simulation studies and the real data analysis. And based on our simulation results, we believe that this setting of  $\lambda_n$  is, if not the best, quite competitive.

## 4 Simulations

In this section, we conduct a Monte Carlo study with two designs to investigate the finite-sample performance of the proposed RIPW-KL criterion. The first design is concerned with the case where the assumed models  $\{\pi_j(y, z; \alpha_j), j = 1, \dots, J\}$  are all misspecified, but  $\pi(y, z)$  is approximately a function of  $\phi_\pi(y, z; \alpha^*)$ . And the second design is considered with the case where one of the assumed models  $\{\pi_j(y, z; \alpha_j), j = 1, \dots, J\}$  is correctly specified. A comparison between our proposal and two related existing model selection strategies based on KL divergence are made. The first model selection strategy uses classical Bayesian information

criterion (BIC; Schwartz, 1978) with complete-case analysis which just ignores all the individuals that contain the missing. And we term this method as CC-KL. The second model selection strategy is the weighted AIC whose weight is selected by Akaike’s information criterion (Claeskens and Hjort, 2008) from the assumed models  $\{\pi_j(y, z; \alpha_j), j = 1, \dots, J\}$ . We term this method as wAIC-KL. We also consider the classical BIC based on the complete data as a gold standard and term this method as CD-KL. The details and results of the Monte Carlo study are given below.

**Design 1:** In this design, we consider the case where  $\{\pi_j(y, z; \alpha_j), j = 1, \dots, J\}$  are all misspecified, while  $\pi(y, z)$  is a function of  $\phi_\pi(y, z; \alpha^*)$ . The data-generating process is

$$Y_i = \vartheta_0 + \vartheta_1 Z_{i,1} + \vartheta_2 Z_{i,2} + \vartheta_3 X_i + e_i, \quad (\vartheta_0, \vartheta_1, \vartheta_2, \vartheta_3)^\top = (0, -0.5, 0, 1)^\top, \tag{8}$$

where  $\{Z_{i,1}, Z_{i,2}, X_i, e_i\}$  are independent standard normal random variables for  $i = 1, 2, \dots, n$ . The selection probability function is

$$\pi(y, z) = \Phi(\tilde{\alpha}_0 + \tilde{\alpha}_1 z_1 y + \tilde{\alpha}_2 z_2 y), \tag{9}$$

where  $\Phi(\cdot)$  is the cumulative distribution function of the standard normal distribution. We consider the following two settings of  $\tilde{\alpha} = (\tilde{\alpha}_0, \tilde{\alpha}_1, \tilde{\alpha}_2)^\top$

$$\text{Case 1 : } \tilde{\alpha} = (0.9, 0.5, -0.5)^\top, \quad \text{Case 2 : } \tilde{\alpha} = (0.2, 0.5, -0.5)^\top.$$

The corresponding average missing rates are approximated 30% and 50%, respectively. For the conditional probability function  $f_{Y|X,Z}(y|x, z)$ , we consider the following four candidate models:

$$\begin{aligned} M_1 : Y_i &= \theta_0 + \theta_3 X_i + \sigma e_i, \\ M_2 : Y_i &= \theta_0 + \theta_1 Z_{i,1} + \theta_3 X_i + \sigma e_i, \\ M_3 : Y_i &= \theta_0 + \theta_2 Z_{i,2} + \theta_3 X_i + \sigma e_i, \\ M_4 : Y_i &= \theta_0 + \theta_1 Z_{i,1} + \theta_2 Z_{i,2} + \theta_3 X_i + \sigma e_i. \end{aligned}$$

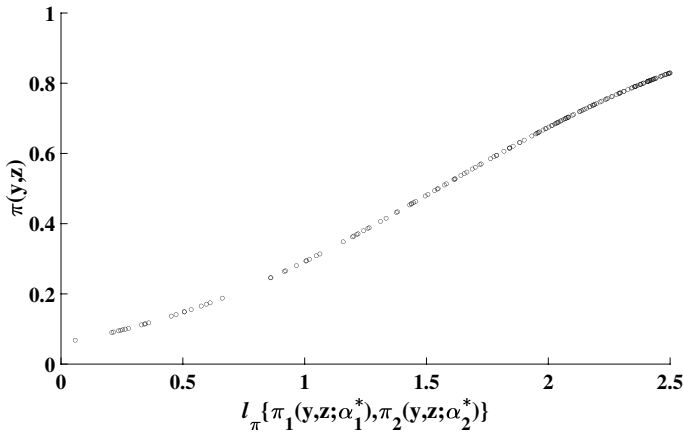
According to the data-generating process (8), it is easy to see that  $M_2$  is the optimal model. For the selection probability function  $\pi(y, z)$ , we consider the following two parametric models,

$$\pi_1(y, z; \alpha_1) = \frac{\exp(\alpha_{10} + \alpha_{11} z_1 y)}{1 + \exp(\alpha_{10} + \alpha_{11} z_1 y)}, \quad \pi_2(y, z; \alpha_2) = \frac{\exp(\alpha_{20} + \alpha_{22} z_2 y)}{1 + \exp(\alpha_{20} + \alpha_{22} z_2 y)}.$$

According to (9), it is easy to see that both  $\pi_1(y, z; \alpha_1)$  and  $\pi_2(y, z; \alpha_2)$  are misspecified, and  $\pi(y, z)$  is neither a function of  $\pi_1(y, z; \alpha_1)$  nor a function of  $\pi_2(y, z; \alpha_2)$ . Consider the following function of  $\pi_1(y, z; \alpha_1)$  and  $\pi_2(y, z; \alpha_2)$ ,

$$l_\pi(\pi_1(y, z; \alpha_1), \pi_2(y, z; \alpha_2)) = c_1^{-1} \log \left\{ \frac{\pi_1(y, z; \alpha_1)}{1 - \pi_1(y, z; \alpha_1)} \right\} + c_2^{-1} \log \left\{ \frac{\pi_2(y, z; \alpha_2)}{1 - \pi_2(y, z; \alpha_2)} \right\},$$





**Fig. 1** Relationship between the selection probability function  $\pi(y, z)$  and the function  $l_\pi(\pi_1(y, z; \alpha_1^*), \pi_2(y, z; \alpha_2^*))$  based on 500 simulated sample points

where  $c_1$  and  $c_2$  are some suitable constants. We analyze the relationship between  $\pi(y, z)$  and  $l_\pi(\pi_1(y, z; \alpha_1^*), \pi_2(y, z; \alpha_2^*))$  based on 500 simulated sample points. And the result is displayed in Fig. 1.

From Fig. 1, it can be seen that  $\pi(y, z)$  is a function of  $l_\pi(\pi_1(y, z; \alpha_1^*), \pi_2(y, z; \alpha_2^*))$ , and hence a function of  $\{\pi_1(y, z; \alpha_1^*), \pi_2(y, z; \alpha_2^*)\}$ . This means that the assumed models  $\{\pi_1(y, z; \alpha_1^*), \pi_2(y, z; \alpha_2^*)\}$  for the selection probability function can approximately recover information of  $\pi(y, z)$ . This actually presents an example that the function is nonlinear since  $l_\pi(\cdot, \cdot)$  is nonlinear.

In order to implement our proposal, following Wang et al. (2021), we use the Epanechnikov kernel function of order 4 which is

$$K(t) = \frac{3}{4\sqrt{5}} \left( \frac{15}{8} - \frac{7t^2}{8} \right) \left( 1 - \frac{t^2}{5} \right) 1_{\{|t| \leq \sqrt{5}\}},$$

and set  $h_n = 1.5n^{-1/(J+4+1)}, b_n = cn^{-1/4(J+4+1)} \log n$ , where  $c$  is a constant. The sample size is considered to be  $n = 100, 200$  and  $400$ . Table 1 reports the number of times that a model is selected in 1000 simulation runs for the considered four methods.

From Table 1, some simulation results can be summarized as follows. First, our proposed RIPW-KL criterion is significantly better than wAIC-KL and CC-KL for all the combinations of simulation conditions. In addition, under the same average missing rate, the performance of our proposed criterion gets better and gets closer to the performance of CD-KL as the sample size increases. This is a significant advantage of our proposal over other model selection methods. Finally, it is worth noting that when the average missing rate is high, both wAIC-KL and CC-KL perform poorly even if the sample size is large. The reason may be that wAIC-KL depends on a correct specification of  $\pi(y, z)$  and CC-KL may

**Table 1** Number of times that each candidate model was selected by the considered four methods in 1000 simulation runs under Design 1

	Method	Case 1				Case 2			
		M1	M2	M3	M4	M1	M2	M3	M4
$n = 100$	CD-KL	2	956	0	42	4	962	1	33
	RIPW-KL	338	569	45	48	377	435	118	70
	wAIC-KL	127	564	55	254	223	298	161	318
	CC-KL	302	407	154	137	343	140	417	100
$n = 200$	CD-KL	0	975	0	25	0	971	0	29
	RIPW-KL	84	864	20	32	185	685	44	86
	wAIC-KL	36	481	11	472	156	255	83	506
	CC-KL	70	454	106	370	127	130	529	214
$n = 400$	CD-KL	0	982	0	18	0	990	0	10
	RIPW-KL	6	972	1	21	27	887	7	79
	wAIC-KL	7	336	1	656	75	196	54	675
	CC-KL	1	273	22	704	14	38	447	501

lead to inconsistency of model selection. The simulation results suggest that our proposed criterion is robust to the misspecified models for  $\pi(y, z)$  as long as the selection probability function is a function of its assumed models.

**Design 2:** In this design, we consider the case where one of the assumed models  $\{\pi_j(y, z; \alpha_j), j = 1, \dots, J\}$  is correctly specified. In fact, it is almost impossible to correctly specify one of the assumed models for the selection probability function  $\pi(y, z)$ . However, in order to make a comparison with Design 1 and fully understand the effect of misspecified parametric model for  $\pi(y, z)$  on our proposed criterion, we take this special case into consideration. The current design is the same as Design 1 except for the following three changes. The selection probability function is now changed to the following

$$\pi(y, z) = \frac{\exp(\tilde{\alpha}_0 + \tilde{\alpha}_1 z_1 y + \tilde{\alpha}_2 z_2 y)}{1 + \exp(\tilde{\alpha}_0 + \tilde{\alpha}_1 z_1 y + \tilde{\alpha}_2 z_2 y)}. \tag{10}$$

And the settings of  $\tilde{\alpha}$  are changed to the following two cases

$$\text{Case 1 : } \tilde{\alpha} = (1.2, 0.5, -0.5)^\top, \quad \text{Case 2 : } \tilde{\alpha} = (0.2, 0.5, -0.5)^\top,$$

such that the corresponding average missing rates are also approximated 30% and 50%, respectively. Furthermore, we change  $\pi_1(y, z; \alpha_1)$  to

$$\pi_1(y, z; \alpha_1) = \frac{\exp(\alpha_{10} + \alpha_{11} z_1 y + \alpha_{12} z_2 y)}{1 + \exp(\alpha_{10} + \alpha_{11} z_1 y + \alpha_{12} z_2 y)},$$

and keep  $\pi_2(y, z; \alpha_2)$  unchanged, so that  $\pi_1(y, z; \alpha_1)$  now becomes a correct specification for the selection probability function  $\pi(y, z)$  in (10), and  $\pi_2(y, z; \alpha_2)$  is misspecified. Table 2 reports the number of times that a model is selected in 1000 simulation runs with the considered four methods.

**Table 2** Number of times that each candidate model was selected by the considered four methods in 1000 simulation runs under Design 2

	Method	Case 1				Case 2			
		M1	M2	M3	M4	M1	M2	M3	M4
$n = 100$	CD-KL	5	955	0	40	9	947	0	44
	RIPW-KL	253	672	19	56	321	565	32	82
	wAIC-KL	33	848	6	113	79	721	20	180
	CC-KL	215	590	62	133	386	312	192	110
$n = 200$	CD-KL	0	986	0	14	0	975	0	25
	RIPW-KL	79	883	3	35	107	830	10	53
	wAIC-KL	1	900	1	98	3	836	2	159
	CC-KL	30	726	16	228	162	375	182	281
$n = 400$	CD-KL	0	988	0	12	0	982	0	18
	RIPW-KL	5	982	0	13	5	959	0	36
	wAIC-KL	0	919	0	81	0	852	0	148
	CC-KL	0	596	0	404	16	304	79	601

From Table 2, some results can be summarized as follows. Under this special case, the performances of both RIPW-KL and wAIC-KL perform well except for CC-KL. It is what we expect since the model selection proposed by wAIC-KL is also consistent when  $\pi_1(y, z; \alpha_1)$  is correctly specified for the selection probability function  $\pi(y, z)$ . As the sample size increases, the performance of our proposed RIPW-KL gets better. Note that, the performance of our proposed RIPW-KL is superior to wAIC-KL when  $n = 400$ . Thus, our proposed RIPW-KL is still comparable in this case. Generally, it is impossible to correctly specify the models for the selection probability function  $\pi(y, z)$  in real life as in Design 2. However, it is relatively easy to assume that the selection probability function  $\pi(y, z)$  is a function of its assumed models. Hence, the RIPW-KL criterion we propose is more robust and more feasible.

All in all, the proposed RIPW-KL criterion performs competitively in the presence of missing data.

### 5 Real data analysis

In this section, we apply the proposed criterion to the automobile data set from the Machine Learning Repository at the University of California Irvine (<http://mlr.cs.umass.edu/mldatasets/Automobile>). The raw data contain 205 sample points. And we analyze the relationship between the price of the car and its corresponding car attributes. Following Wei et al. (2021), we take the logarithm of the price of the car as the response variable  $Y$  and choose normalized-losses ( $X$ ), wheel-base ( $Z_1$ ), length ( $Z_2$ ), width ( $Z_3$ ), height ( $Z_4$ ), curb-weight ( $Z_5$ ), engine-size ( $Z_6$ ), bore ( $Z_7$ ), stroke ( $Z_8$ ), compression-ratio ( $Z_9$ ), horsepower ( $Z_{10}$ ), peak-rpm ( $Z_{11}$ ), city-mpg ( $Z_{12}$ ), highway-mpg ( $Z_{13}$ ), symboling ( $Z_{14}$ ), fuel-type ( $Z_{15} = 1$ , diesel;  $Z_{15} = 2$ , gas), aspiration ( $Z_{16} = 1$ , std;  $Z_{16} = 2$ , turbo), num-of-doors ( $Z_{17}$ ), and num-of-cylinders ( $Z_{18}$ ) as predictors.

Among the 18 predictors, the  $Z_{14}, Z_{15}, Z_{16}, Z_{17},$  and  $Z_{18}$  are discrete predictors and the remaining are continuous predictors. Denote  $Z = (Z_1, Z_2, \dots, Z_{18})$ . Note that, there are 4 sample points in  $Y$  and 8 samples points in  $Z$  contain missing values, and we simply delete these sample points. Thus, the response variable  $Y$  and covariates  $Z$  are observed completely, while there are 34 missing values for covariates  $X$ . Further, we standardize each continuous predictors based on complete cases. Thus, for the conditional probability linear function  $f_{Y|X,Z}(y|x, z)$ , we have  $2^{19} - 1$  candidate models. To the candidate model  $M$ , we have

$$Y = \theta_0 + \sum_{i \in Y_M} \theta_i T_i + e,$$

where  $T = (Z, X)$ ,  $Y_M \subset \{1, 2, \dots, 19\}$ ,  $\theta_i$  is the unknown parameter vector, and  $e$  is a normal error vector which is independent of  $T$  with mean zero and unknown variance  $\sigma^2$ . Clearly, the sheer number of models is not conducive to implementing our proposed criterion. Thus, to obtain a series of candidate models, we take full advantage of FW/BW procedure in Jiang et al. (2015). Similar to Jiang et al. (2015), due to missing data, we calculate  $RSS(M, T, \theta)$  based on the complete cases and denote it as  $RSS_{cc}(M, T, \theta)$ , where  $RSS(M, T, \theta) = \sum_{i=1}^n (Y_i - \theta_0 - \sum_{i \in Y_M} \theta_i T_i)^2$  and  $RSS_{cc}(M, T, \theta) = \sum_{i=1}^n \delta_i (Y_i - \theta_0 - \sum_{i \in Y_M} \theta_i T_i)^2$ . Then, taking FW/BW procedure, we can obtain the candidate models for  $f_{Y|X,Z}(y|x, z)$  based on  $RSS_{cc}(M, T, \theta)$ .

Similarly, we assume that  $X$  is missing at random (MAR). As for the selection probability function  $\pi(y, z)$ , we consider the following two cases:

Case1:

$$\begin{aligned} \pi_1(y, z; \alpha_1) &= \exp(\alpha_{10} + \alpha_{11}y + \alpha_{11}z) / \{1 + \exp(\alpha_{10} + \alpha_{11}y + \alpha_{11}z)\}, \\ \pi_2(y, z; \alpha_2) &= \exp(\alpha_{21}y + \alpha_{22}z) / \{1 + \exp(\alpha_{21}y + \alpha_{22}z)\}, \end{aligned}$$

where  $\alpha_1, \alpha_2$  are the corresponding parameter vectors.

Case2:

$$\begin{aligned} \pi_1(y, z; \beta_1) &= \exp(\beta_{10} + \beta_{11}y + \beta_{12}z + \beta_{13}zy) / \{1 + \exp(\beta_{10} + \beta_{11}y + \beta_{12}z + \beta_{13}zy)\}, \\ \pi_2(y, z; \beta_2) &= \exp(\beta_{20} + \beta_{21}y + \beta_{22}z) / \{1 + \exp(\beta_{20} + \beta_{21}y + \beta_{22}z)\}, \end{aligned}$$

where  $\beta_1, \beta_2$  are the corresponding parameter vectors.

Table 3 reports the models selected by the two methods, respectively.

From Table 3, it can be observed that RIPW-KL criterion is more robust than wAIC-KL since our proposal selects the same model under two different specifications

**Table 3** The models selected by RIPW-KL criterion, wAIC-KL, respectively, for the automobile data set

method	Case 1	Case 2
RIPW-KL	$Z_5, Z_{10}, X$	$Z_5, Z_{10}, X$
wAIC-KL	$Z_5, Z_{10}, X$	$Z_3, Z_5, Z_{10}, X$

of  $\pi(y, z)$ . Clearly, in real data analysis, it is hard to select the true model for the selection probability function. However, it is relatively reliable to assume some models for the selection probability function. Similar to Wang et al. (2021), we can see that the spirit of RIPW-KL criterion is like the idea of model averaging that can combine the information of each assumed models for the selection probability function, while the wAIC-KL simply uses the information of one model in the end. This may be the reason that the RIPW-KL criterion is more robust than wAIC-KL.

### Appendix

We present proofs of theorems as follows.

**Proof of Theorem 1** In order to prove Theorem 1, we need to prove the existence and consistency of  $\hat{\theta}_M^{IP}$  defined in (5) firstly. Based on Property 24.1 in Gourieroux and Monfort (1995), the existence of  $\hat{\theta}_M^{IP}$  can be guaranteed under (C.1), (C.2), (C.5) and (C.6). Recalling the definition of  $\hat{D}_{IP}(M, \theta_M, \hat{\alpha}_n)$  given in (4), based on Theorem 2.1 in Newey and Mcfadden (1994) and (C.1), (C.2), it suffices to prove the consistency of the  $\hat{\theta}_M^{IP}$  by verifying the following equation:

$$\sup_{\theta_M \in \Theta_M} |\hat{D}_{IP}(M, \theta_M, \hat{\alpha}_n) - D(M, \theta_M)| = o_p(1).$$

Note that,

$$\begin{aligned} &|\hat{D}_{IP}(M, \theta_M, \hat{\alpha}_n) - D(M, \theta_M)| \\ &\leq |\hat{D}_{IP}(M, \theta_M, \hat{\alpha}_n) - \tilde{D}_{IP}(M, \theta_M, \alpha^*)| + |\tilde{D}_{IP}(M, \theta_M, \alpha^*) - D(M, \theta_M)|, \end{aligned}$$

where  $\tilde{D}_{IP}(M, \theta_M, \alpha^*) = n^{-1} \sum_{i=1}^n \left\{ \frac{\delta_i}{p_{\alpha^*}(\phi_\pi(Y_i, Z_i; \alpha^*))} \log g_M(Y_i | X_i, Z_i; \theta_M) \right\}$ , in which  $p_{\alpha^*}(\phi_\pi(Y, Z; \alpha^*))$  is defined in the first paragraph of Sect. 3. We need only to prove,

$$\sup_{\theta_M \in \Theta_M} |\hat{D}_{IP}(M, \theta_M, \hat{\alpha}_n) - \tilde{D}_{IP}(M, \theta_M, \alpha^*)| = o_p(1), \tag{11}$$

and

$$\sup_{\theta_M \in \Theta_M} |\tilde{D}_{IP}(M, \theta_M, \alpha^*) - D(M, \theta_M)| = o_p(1). \tag{12}$$

According to Lemma 2.4 in Newey and Mcfadden (1994), with (C.1) and (C.2), it is direct to prove (12) by noting

$$E\{\delta|\phi_\pi(y, z; \alpha^*)\} = \pi(y, z). \tag{13}$$

Note that,

$$\begin{aligned}
 & \hat{D}_{IP}(M, \theta_M, \hat{\alpha}_n) - \tilde{D}_{IP}(M, \theta_M, \alpha^*) \\
 &= \frac{1}{n} \sum_{i=1}^n \delta_i \log g_M(Y_i|X_i, Z_i; \theta_M) \{ \hat{q}_{\hat{\alpha}_n, b_n}(\phi_\pi(Y_i, Z_i; \hat{\alpha}_n)) - \hat{q}_{\alpha^*, b_n}(\phi_\pi(Y_i, Z_i; \alpha^*)) \} \\
 &+ \frac{1}{n} \sum_{i=1}^n \delta_i \log g_M(Y_i|X_i, Z_i; \theta_M) \{ \hat{q}_{\alpha^*, b_n}(\phi_\pi(Y_i, Z_i; \alpha^*)) - q_{\alpha^*, b_n}(\phi_\pi(Y_i, Z_i; \alpha^*)) \} \\
 &+ \frac{1}{n} \sum_{i=1}^n \delta_i \log g_M(Y_i|X_i, Z_i; \theta_M) \{ q_{\alpha^*, b_n}(\phi_\pi(Y_i, Z_i; \alpha^*)) - q_{\alpha^*}(\phi_\pi(Y_i, Z_i; \alpha^*)) \} \\
 &:= Q_{n1} + Q_{n2} + Q_{n3},
 \end{aligned} \tag{14}$$

where  $\hat{q}_{\alpha, n}(u) = 1/\hat{p}_{\alpha, n}(u)$ ,  $\hat{q}_{\alpha, b_n}(u) = 1/\hat{p}_{\alpha, b_n}(u) = \hat{q}_{\alpha, n}(u)\hat{r}_{\alpha, n}(u)/\hat{r}_{\alpha, b_n}(u)$ ,  $q_\alpha(u) = 1/p_\alpha(u)$ ,  $q_{\alpha, b_n}(u) = q_\alpha(u)r_\alpha(u)/r_{\alpha, b_n}(u)$ ,  $r_{\alpha, b_n}(u) = \max\{r_\alpha(u), b_n\}$ , in which  $p_\alpha(v)$  and  $r_\alpha(v)$  are given at the first paragraph of Sect. 3. By (2) and (3), we know that the definition of  $\hat{q}_{\alpha, n}(u)$ ,  $\hat{q}_{\alpha, b_n}(u)$ ,  $q_\alpha(u)$  and  $q_{\alpha, b_n}(u)$  are similar to  $\hat{a}_{\gamma, n}(v)$ ,  $\hat{a}_{\gamma, b_n}(u)$ ,  $a_\gamma(u)$  and  $a_{\gamma, b_n}(u)$  in Wang et al. (2021), respectively. By (C.2), we have  $E\{\sup_{\theta_M \in \Theta_M} \delta \log g_M(Y|X, Z; \theta_M)\} < \infty$ . This together with conditions (C.3)-(C.7) proves  $\sup_{\theta_M \in \Theta_M} Q_{n1} = o_p(1)$  and  $\sup_{\theta_M \in \Theta_M} Q_{n2} = o_p(1)$ , respectively, using the similar arguments to that of Lemma S1 and S2 in Wang et al. (2021). Clearly,

$$|Q_{n3}| \leq \frac{2}{n} \sum_{i=1}^n |\delta_i \log g_M(Y_i|X_i, Z_i; \theta_M) q_{\alpha^*}(\phi_\pi(Y_i, Z_i; \alpha^*))| I[r_{\alpha^*}(\phi_\pi(Y_i, Z_i; \alpha^*)) < b_n].$$

By (C.2), (C.3) and (C.4),  $\forall \epsilon > 0$ , we then have

$$\begin{aligned}
 & P(|\sup_{\theta_M \in \Theta_M} Q_{n3}| > \epsilon) \\
 & \leq \frac{2}{\epsilon} E\{|\sup_{\theta_M \in \Theta_M} \delta \log g_M(Y|X, Z; \theta_M) q_{\alpha^*}(\phi_\pi(Y, Z; \alpha^*))| I[r_{\alpha^*}(\phi_\pi(Y, Z; \alpha^*)) < b_n]\} \\
 & \rightarrow 0.
 \end{aligned}$$

This yields

$$\begin{aligned}
 & \sup_{\theta_M \in \Theta_M} |\hat{D}_{IP}(M, \theta_M, \hat{\alpha}_n) - \tilde{D}_{IP}(M, \theta_M, \alpha^*)| \\
 & \leq \left| \sup_{\theta_M \in \Theta_M} Q_{n1} \right| + \left| \sup_{\theta_M \in \Theta_M} Q_{n2} \right| + \left| \sup_{\theta_M \in \Theta_M} Q_{n3} \right| = o_p(1).
 \end{aligned}$$

This completes the proof of (11).

Now, in what follows, we prove the asymptotically normality of  $\hat{\theta}_M^{IP}$ . By (14), we know that

$$\hat{D}_{IP}(M, \theta_M, \hat{\alpha}_n) = \tilde{D}_{IP}(M, \theta_M, \alpha^*) + Q_{n1} + Q_{n2} + Q_{n3}.$$

Lemma 1 in the supplementary material proves that

$$Q_{n1} = n^{-1} \sum_{i=1}^n \{ \delta_i \log g_M(Y_i | X_i, Z_i; \theta_M) \} \{ \partial q_{\alpha^*}(\phi_\pi(Y_i, Z_i; \alpha^*)) / \partial \alpha \} (\hat{\alpha}_n - \alpha^*) + o_p(n^{-1/2}). \tag{15}$$

Further, according to Lemma 2 in the supplementary material, we have

$$Q_{n2} = n^{-1} \sum_{i=1}^n \{ 1 - \delta_i q_{\alpha^*}(\phi_\pi(Y_i, Z_i; \alpha^*)) \} q_{\alpha^*}(\phi_\pi(Y_i, Z_i; \alpha^*)) \times \delta_i \log g_M(Y_i | X_i, Z_i; \theta_M) + o_p(n^{-1/2}). \tag{16}$$

For  $Q_{n3}$ , by (C.2), (C.8) and Markov’s inequality, we have

$$\begin{aligned} &P(n^{1/2} |Q_{n3}| > \epsilon) \\ &\leq P(n^{-1/2} \sum_{i=1}^n 2 |C_i q_{\alpha^*}(\phi_\pi(Y_i, Z_i; \alpha^*))| I[r_{\alpha^*}(\phi_\pi(Y_i, Z_i; \alpha^*)) < b_n] > \epsilon) \\ &\leq 2\epsilon^{-1} E\{ \sqrt{n} |C q_{\alpha^*}(\phi_\pi(Y, Z; \alpha^*))| I[r_{\alpha^*}(\phi_\pi(Y, Z; \alpha^*)) < b_n] \} \rightarrow 0, \end{aligned} \tag{17}$$

where  $C_i = \delta_i \log g_M(Y_i | X_i, Z_i; \theta_M)$ . Then, we have  $Q_{n3} = o_p(n^{-1/2})$ . Thus, we have

$$\hat{D}_{IP}(M, \theta_M, \hat{\alpha}_n) = \tilde{D}_{IP}(M, \theta_M, \alpha^*) + Q_{n1} + Q_{n2} + o_p(n^{-1/2}). \tag{18}$$

Let  $\Psi(u, v)$  be a general vector-valued or matrix-valued function, and we denote

$$\Psi'_{\{u\}}(u, v) = \frac{\partial \Psi(u, v)}{\partial u},$$

$$K_M(\theta_M, \hat{\alpha}_n) = \hat{D}'_{IP(\theta_M)}(M, \theta_M, \hat{\alpha}_n).$$

Then, by (4), we have

$$K_M(\theta_M, \hat{\alpha}_n) = n^{-1} \sum_{i=1}^n \left\{ \frac{\delta_i}{\hat{p}_{\hat{\alpha}_n, b_n}(\phi_\pi(Y_i, Z_i; \hat{\alpha}_n))} t_{M,i}(\theta_M) \right\},$$

where  $t_{M,i}(\theta_M) = \frac{\partial \log g_M(Y_i | X_i, Z_i; \theta_M)}{\partial \theta_M}$ , for  $i = 1, 2, \dots, n$ . With the same technique of (18), under (C.2)-(C.8), we can obtain that

$$\begin{aligned} K_M(\hat{\theta}_M^{IP}, \hat{\alpha}_n) &= n^{-1} \sum_{i=1}^n q_{\alpha^*}(\phi_\pi(Y_i, Z_i; \alpha^*)) \delta_i t_{M,i}(\hat{\theta}_M^{IP}) \\ &+ n^{-1} \sum_{i=1}^n \delta_i t_{M,i}(\hat{\theta}_M^{IP}) \{ \partial q_{\alpha^*}(\phi_\pi(Y_i, Z_i; \alpha^*)) / \partial \alpha \} (\hat{\alpha}_n - \alpha^*) \\ &+ n^{-1} \sum_{i=1}^n \{ 1 - \delta_i q_{\alpha^*}(\phi_\pi(Y_i, Z_i; \alpha^*)) \} q_{\alpha^*}(\phi_\pi(Y_i, Z_i; \alpha^*)) \delta_i t_{M,i}(\hat{\theta}_M^{IP}) + o_p(n^{-1/2}). \end{aligned} \tag{19}$$

From (5), it follows that  $K_M(\hat{\theta}_M^{\text{IP}}, \hat{\alpha}_n) = 0$ . Thus, applying Taylor expansion to  $K_M(\hat{\theta}_M^{\text{IP}}, \hat{\alpha}_n)$  around the point  $(\theta_M^*, \alpha^*)$ , we have

$$\begin{aligned}
 0 &= n^{-1} \sum_{i=1}^n \left\{ q_{\alpha^*}(\phi_\pi(Y_i, Z_i; \alpha^*)) \delta_i t_{M,i}(\theta_M^*) + [1 - \delta_i q_{\alpha^*}(\phi_\pi(Y_i, Z_i; \alpha^*))] \right. \\
 &\quad \times q_{\alpha^*}(\phi_\pi(Y_i, Z_i; \alpha^*)) \delta_i t_{M,i}(\theta_M^*) \left. \right\} + n^{-1} \sum_{i=1}^n \left\{ q_{\alpha^*}(\phi_\pi(Y_i, Z_i; \alpha^*)) \delta_i t'_{M,i}(\theta_M^*) \right. \\
 &\quad \left. + [1 - \delta_i q_{\alpha^*}(\phi_\pi(Y_i, Z_i; \alpha^*))] q_{\alpha^*}(\phi_\pi(Y_i, Z_i; \alpha^*)) \delta_i t'_{M,i}(\theta_M^*) \right\} (\hat{\theta}_M^{\text{IP}} - \theta_M^*) \\
 &\quad + n^{-1} \sum_{i=1}^n \left\{ \delta_i t_{M,i}(\theta_M^*) \partial q_{\alpha^*}(\phi_\pi(Y_i, Z_i; \alpha^*)) / \partial \alpha \right\} (\hat{\alpha}_n - \alpha^*) + o_p(n^{-1/2}) \\
 &:= H_{n1}(M, \theta_M^*, \alpha^*) + H_{n2}(M, \theta_M^*, \alpha^*) (\hat{\theta}_M^{\text{IP}} - \theta_M^*) + H_{n3}(M, \theta_M^*, \alpha^*) (\hat{\alpha}_n - \alpha^*) \\
 &\quad + o_p(n^{-1/2}),
 \end{aligned} \tag{20}$$

where  $\alpha^*$  is the probability limit of  $\hat{\alpha}_n$  and  $\theta_M^*$  is given in Condition (C.1). Let

$$t_\pi(\alpha) = \frac{\delta - \phi_\pi(y, z; \alpha)}{\phi_\pi(y, z; \alpha) \{1 - \phi_\pi(y, z; \alpha)\}} \cdot \phi'_{\pi(\alpha)}(y, z; \alpha).$$

And by a standard argument, one can easily obtain the following equation:

$$\sqrt{n}(\hat{\alpha}_n - \alpha^*) = I_{\alpha^*}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n t_{\pi,i}(\alpha^*) + o_p(1), \tag{21}$$

where  $I_{\alpha^*} = -E[t'_{\pi}(\alpha)_{\{\alpha^T\}}(\alpha^*)]$ .

Similarly, by the law of larger numbers, we can obtain that

$$H_{n2}(M, \theta_M^*, \alpha^*) = -I_{\theta_M^*} + o_p(1), \quad H_{n3}(M, \theta_M^*, \alpha^*) = -A_{M,\alpha^*} + o_p(1), \tag{22}$$

where

$$\begin{aligned}
 I_{\theta_M^*} &= -E \left[ \delta q_{\alpha^*}(\phi_\pi(Y, Z; \alpha^*)) t'_{M\{\theta_M^*\}}(\theta_M^*) + \{1 - \delta q_{\alpha^*}(\phi_\pi(Y, Z; \alpha^*))\} \right. \\
 &\quad \left. \times q_{\alpha^*}(\phi_\pi(Y, Z; \alpha^*)) \delta t'_{M\{\theta_M^*\}}(\theta_M^*) \right], \\
 A_{M,\alpha^*} &= -E \left\{ \delta \frac{\partial q_{\alpha^*}(\phi_\pi(Y, Z; \alpha^*))}{\partial \alpha^*} t_M(\theta_M^*) \right\}.
 \end{aligned}$$

Thus, (20) together with (21) and (22), we prove

$$\sqrt{n}(\hat{\theta}_M^{\text{IP}} - \theta_M^*) = \frac{1}{\sqrt{n}} \sum_{i=1}^n R_{M,i} + o_p(1),$$

where



$$R_{M,i} = I_{\theta_M^*}^{-1} \left\{ \delta_i q_{\alpha^*}(\phi_\pi(Y_i, Z_i; \alpha^*)) t_{M,i}(\theta_M^*) + [1 - \delta_i q_{\alpha^*}(\phi_\pi(Y_i, Z_i; \alpha^*))] \right. \\ \left. \times q_{\alpha^*}(\phi_\pi(Y_i, Z_i; \alpha^*)) \delta_i t_{M,i}(\theta_M^*) - A_{M,\alpha} I_{\alpha^*}^{-1} t_{\pi,i}(\alpha^*) \right\}.$$

By the central limit theorem and the above prove process, we prove that  $\hat{\theta}_M^{\text{IP}}$  is asymptotically normal with zero mean, as  $n \rightarrow \infty$ . This completes the proof of Theorem 1. □

**Proof of Theorem 2** Obviously, in order to prove Theorem 2, it suffices to prove the following equation:

$$P\{\text{IC}_{\text{IP}}(\hat{M}_{\text{IP}}) = \text{IC}_{\text{IP}}(M_{\text{opt}})\} \rightarrow 1 \quad (n \rightarrow \infty). \tag{23}$$

Recalling the definition of  $\hat{M}_{\text{IP}}$  in (7),  $\hat{M}_{\text{IP}}$  is the minimizer of  $\text{IC}_{\text{IP}}(M)$  with respect to  $M$ . So, it is obvious that we have  $\text{IC}_{\text{IP}}(\hat{M}_{\text{IP}}) \leq \text{IC}_{\text{IP}}(M_{\text{opt}})$ . Thus, to prove (23), we only need to prove that

$$P\{\text{IC}_{\text{IP}}(\hat{M}_{\text{IP}}) \geq \text{IC}_{\text{IP}}(M_{\text{opt}})\} \rightarrow 1 \quad (n \rightarrow \infty). \tag{24}$$

To prove (24), it's enough to prove the following equation:

$$P\{\text{IC}_{\text{IP}}(M) \geq \text{IC}_{\text{IP}}(M_{\text{opt}})\} \rightarrow 1 \quad (n \rightarrow \infty), \tag{25}$$

for each candidate model  $M$ . By the definition of  $\text{IC}_{\text{IP}}(M)$  defined in (6), we know that in order to prove (25), it's equivalent to prove the following equation, when  $n \rightarrow \infty$ ,

$$P(\hat{D}_{\text{IP}}(M_{\text{opt}}, \hat{\theta}_{M_{\text{opt}}}^{\text{IP}}, \hat{\alpha}_n) - \hat{D}_{\text{IP}}(M, \hat{\theta}_M^{\text{IP}}, \hat{\alpha}_n) + \lambda_n(d_M - d_{M_{\text{opt}}}) \geq 0) \rightarrow 1. \tag{26}$$

If  $M = M_{\text{opt}}$ , (26) is clearly true. Thus, we consider the case where the model  $M$  is not  $M_{\text{opt}}$  only. By Theorem 1, applying Taylor-expansion to  $\hat{D}_{\text{IP}}(M, \hat{\theta}_M^{\text{IP}}, \hat{\alpha}_n)$ , we have

$$\hat{D}_{\text{IP}}(M, \hat{\theta}_M^{\text{IP}}, \hat{\alpha}_n) = \hat{D}_{\text{IP}}(M, \theta_M^*, \alpha^*) + \hat{D}'_{\text{IP}\{\theta_M^*\}}(M, \theta_M^*, \alpha^*)(\hat{\theta}_M^{\text{IP}} - \theta_M^*) \\ + \hat{D}'_{\text{IP}\{\alpha^*\}}(M, \theta_M^*, \alpha^*)(\hat{\alpha}_n - \alpha^*) + o_p(n^{-1/2}),$$

where the definition of  $\hat{\theta}_{M_{\text{opt}}}^{\text{IP}}$  is similar to  $\hat{\theta}_M^{\text{IP}}$ . By (C.2), (C.4) and (C.5)(ii) as well as the root  $n$  consistency of  $\hat{\theta}_M^{\text{IP}}$  and  $\hat{\alpha}_n$ , we have

$$\hat{D}'_{\text{IP}\{\theta_M^*\}}(M, \theta_M^*, \alpha^*)(\hat{\theta}_M^{\text{IP}} - \theta_M^*) = O_p(n^{-1/2}), \\ \hat{D}'_{\text{IP}\{\alpha^*\}}(M, \theta_M^*, \alpha^*)(\hat{\alpha}_n - \alpha^*) = O_p(n^{-1/2}).$$

Thus, we have

$$\hat{D}_{\text{IP}}(M, \hat{\theta}_M^{\text{IP}}, \hat{\alpha}_n) = \hat{D}_{\text{IP}}(M, \theta_M^*, \alpha^*) + O_p(n^{-1/2}). \tag{27}$$

Similarly, we have

$$\hat{D}_{\text{IP}}(M, \hat{\theta}_{M_{\text{opt}}}^{\text{IP}}, \hat{\alpha}_n) = \hat{D}_{\text{IP}}(M, \theta_{M_{\text{opt}}}^*, \alpha^*) + O_p(n^{-1/2}), \quad (28)$$

where the definition of  $\theta_{M_{\text{opt}}}^*$  is similar to  $\theta_M^*$ . Note that,  $\pi(y, z)$  is a function of  $\phi_x(y, z; \alpha^*)$ , together with (13) and (18) and the law of large numbers, we have

$$\hat{D}_{\text{IP}}(M, \hat{\theta}_M^{\text{IP}}, \hat{\alpha}_n) = D(M, \theta_M^*) + O_p(n^{-1/2}). \quad (29)$$

Obviously, (29) is also true for  $M = M_{\text{opt}}$ , we then have

$$\begin{aligned} & \hat{D}_{\text{IP}}(M_{\text{opt}}, \hat{\theta}_{M_{\text{opt}}}^{\text{IP}}, \hat{\alpha}_n) - \hat{D}_{\text{IP}}(M, \hat{\theta}_M^{\text{IP}}, \hat{\alpha}_n) \\ &= D(M_{\text{opt}}, \theta_{M_{\text{opt}}}^*) - D(M, \theta_M^*) + O_p(n^{-1/2}). \end{aligned} \quad (30)$$

Recalling the definition of  $D(M, \theta_M)$  given below (1), it follows that  $D(M_{\text{opt}}, \theta_{M_{\text{opt}}}^*) - D(M, \theta_M^*)$  is non-negative in probability. Recalling that  $\lambda_n$  is a positive tuning parameter tending to zero as  $n \rightarrow \infty$ . By Fang and Shao (2016), we consider the following three cases to prove (26):

*Case 1.*  $M$  is an incorrect model and  $d_{M_{\text{opt}}} < d_M$ . In this case, we then have  $\hat{D}_{\text{IP}}(M, \hat{\theta}_{M_{\text{opt}}}^{\text{IP}}, \hat{\alpha}_n) - \hat{D}_{\text{IP}}(M, \hat{\theta}_M^{\text{IP}}, \hat{\alpha}_n) > 0$  in probability, and hence (26) is clearly true.

*Case 2.*  $M$  is an incorrect model but  $d_{M_{\text{opt}}} \geq d_M$ . Similar to *Case 1*, we have  $\hat{D}_{\text{IP}}(M, \hat{\theta}_{M_{\text{opt}}}^{\text{IP}}, \hat{\alpha}_n) - \hat{D}_{\text{IP}}(M, \hat{\theta}_M^{\text{IP}}, \hat{\alpha}_n) > 0$  in probability, (26) then holds by noting  $\lambda_n \rightarrow 0$ .

*Case 3.*  $M$  is a correct model but  $d_{M_{\text{opt}}} < d_M$ . In this case, we then have  $\hat{D}_{\text{IP}}(M, \hat{\theta}_M^{\text{IP}}, \hat{\alpha}_n) - \hat{D}_{\text{IP}}(M, \hat{\theta}_{M_{\text{opt}}}^{\text{IP}}, \hat{\alpha}_n) = O_p(n^{-\frac{1}{2}})$ , and hence (26) is true as long as  $\sqrt{n}\lambda_n \rightarrow \infty$ .

This completes the proof of (26) and hence the proof of Theorem 2.  $\square$

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1007/s10463-021-00806-2>.

**Acknowledgements** Wang's research was supported by the National Natural Science Foundation of China (General program 11871460, Key program 11331011 and program for Innovative Research Group Project 61621003), a grant from the Key Lab of Random Complex Structure and Data Science, CAS.

## References

- Celeux, G., Forbes, F., Robert, C. P., Titterton, D. M. (2006). Deviance information criteria for missing data models. *Bayesian Analysis*, 1(4), 651–673.
- Claeskens, G., Consentino, F. (2008). Variable selection with incomplete covariate data. *Biometrics*, 64(4), 1062–1069.
- Claeskens, G., Hjort, N. L. (2003). The focused information criterion. *Journal of the American Statistical Association*, 98(464), 900–916.
- Claeskens, G., Hjort, N. L. (2008). *Model selection and model averaging*. Cambridge University Press.

- Fang, F., Shao, J. (2016). Model selection with nonignorable nonresponse. *Biometrika*, 103(4), 861–874.
- Gelman, A., Van Mechelen, I., Verbeke, G., Heitjan, D. F., Meulders, M. (2005). Multiple imputation for model checking: Completed-data plots with missing and latent data. *Biometrics*, 61(1), 74–85.
- Gourieroux, C., Monfort, A. (1995). *Statistics and econometric models* (Vol. 2). Cambridge University Press.
- Hens, N., Aerts, M., Molenberghs, G. (2006). Model selection for incomplete and design-based samples. *Statistics in Medicine*, 25(14), 2502–2520.
- Horvitz, D. G., Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260), 663–685.
- Ibrahim, J. G., Zhu, H., Tang, N. (2008). Model selection criteria for missing data problems using the EM algorithm. *Journal of the American Statistical Association*, 103(484), 1648–1658.
- Jiang, J., Rao, J. S., Gu, Z., Nguyen, T. (2008). Fence methods for mixed model selection. *The Annals of Statistics*, 36(4), 1669–1692.
- Jiang, J., Nguyen, T., Rao, J. S. (2015). The E-MS algorithm: Model selection with incomplete data. *Journal of the American Statistical Association*, 110(511), 1136–1147.
- Little, R. J. A., Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Wiley.
- Mallow, C. L. (1973). Some comments on  $C_p$ . *Technometrics*, 15(4), 661–675.
- Newey, W. K., Mcfadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of Econometrics*, 4(05), 2111–2245.
- Robins, J. M., Rotnitzky, A., Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427), 846–866.
- Rolling, C. A., Yang, Y. (2014). Model selection for estimating treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(4), 749–769.
- Schwartz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 461–464.
- Shao, Q., Yang, L. (2017). Oracally efficient estimation and consistent model selection for auto-regressive moving average time series with trend. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(2), 507–524.
- Wang, Q., Rao, J. N. K. (2002a). Empirical likelihood-based inference under imputation for missing response data. *The Annals of Statistics*, 30(3), 896–924.
- Wang, Q., Su, M., Wang, R. (2021). A beyond multiple robust approach for missing response problem. *Computational Statistics & Data Analysis*, 155, 107111.
- Wei, Y., Wang, Q., Duan, X., Qin, J. (2021). Bias-corrected Kullback-Leibler distance criterion based model selection with covariables missing at random. *Computational Statistics & Data Analysis*, 160.
- Zhang, X., Wang, H., Ma, Y., Carroll, R. J. (2017). Linear model selection when covariates contain errors. *Journal of the American Statistical Association*, 112(520), 1553–1561.

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.