

**Statistica Sinica Preprint No: SS-2017-0042**

<b>Title</b>	Generalized empirical likelihood inference for nonsmooth moment functions with non-ignorable missing values
<b>Manuscript ID</b>	SS-2017-0042
<b>URL</b>	<a href="http://www.stat.sinica.edu.tw/statistica/">http://www.stat.sinica.edu.tw/statistica/</a>
<b>DOI</b>	10.5705/ss.202017.0042
<b>Complete List of Authors</b>	Puying Zhao Niansheng Tang and Hongtu Zhu
<b>Corresponding Author</b>	Niansheng Tang
<b>E-mail</b>	nstang@ynu.edu.cn
Notice: Accepted version subject to English editing.	

# Generalized empirical likelihood inference for nonsmooth moment functions with non-ignorable missing values

Puying Zhao<sup>1,2</sup>, Niansheng Tang<sup>1</sup>, and Hongtu Zhu<sup>3</sup>

<sup>1</sup>*Key Lab of Statistical Modeling and Data Analysis of Yunnan Province, Yunnan University, China.*

<sup>2</sup>*Department of Statistics and Actuarial Science, University of Waterloo, Canada.*

<sup>3</sup>*Department of Biostatistics, The University of Texas MD Anderson Cancer Center, U.S.A.*

*Abstract:* The main purpose of this paper is to develop parameter identifiability and statistical inference for a class of possibly over-identified nonsmooth moment functions with nonignorable missing data. Assuming a parametric model on the respondent probability, we propose a propensity score based nonparametric imputation approach that makes use of an instrumental variable to address the model identifiability in the presence of nonignorable missing data. A set of the augmented inverse probability weighting moment functions is constructed as a basis for inference by the generalized empirical likelihood. Under some mild regularity conditions, we establish large sample properties of the resultant two-step generalized empirical likelihood estimators and generalized empirical likelihood ratio statistics for the case that the propensity score is parametrically estimated using a correctly specified model. A derivative-free optimization method based on the simulated annealing algorithm is developed to implement the proposed methods. The proposed methods are illustrated by some simulation results and an application to a data set on the serum-cholesterol levels of heart-attack patients.

*Keywords:* Generalized empirical likelihood; Identification; Instrumental variable; Nonignorable missing data; Nonsmooth moment conditions; Simulated annealing algorithm.

# 1 Introduction

Missing data are usually encountered in many studies such as clinical trials and survey studies. It is well-known that complete case analysis that using only complete observations loses information that is available in the data, and thus may lead to bias if the data is not missing completely at random. Over the last decades, various valid inferential alternatives to complete-case approach have been developed for handling missing values, that include the likelihood-based approach (Ibrahim et al., 1999; Little and Rubin, 2002); the imputation approach (Rubin, 1987; Cheng, 1994) and the augmented inverse probability weighting (AIPW) approach (Robins et al., 1994). All these methods have been successfully applied to analyze ignorable missing data (see Tstatis 2006 for more detailed discussions). For a complete review on statistical analysis with missing data, see Little and Rubin (2002) and Kim and Shao (2013).

Most of the existing methods assuming an ignorable missing-data mechanism will fail to recover information from the incomplete observed cases and/or to correct bias that may occur when the missing-data mechanism is nonignorable. Handling nonignorable missing data is difficult due to several key challenges. Such challenges include a sensible model for characterizing the non-ignorable missing data mechanism and its associated model identifiability and estimation (Robins and Ritov, 1997; Kim and Yu, 2011; Wang et al., 2014; Tang et al., 2014; Zhao and Shao, 2015; Shao and Wang, 2016; Miao and Tchetgen, 2016; Zhao et al., 2017a). The AIPW method is a prominent semiparametric method dealing with missing data, which has been well developed for handling non-ignorable missing data in the literature; see, for example, Scharfstein et al. (1999), Rotnitzky et al. (1998), Vansteelandt et al. (2007), among others. However, all these existing AIPW methods focus only on smooth moment conditions and depend on identifiable nonignorable propensity score models with certain restrictions that are hard to verify. The research on the topic of AIPW for nonignorable missing data is far from complete.

Statistical and econometric models defined via nonsmooth moment functions are rather common, that include the least absolute deviations, the quantile regression models, and the quantile treatment effects as special cases. Considerable effort has been devoted to the estimation of the finite dimensional parameters defined via nonsmooth moment functions in the presence of missing data; see, for example, Chen et al. (2008), Cattaneo (2010), Chen et al. (2015), and Chaudhuri and Guilkey (2016) and the references therein. However, most existing estimation procedures were primarily developed to handle ignorable missing data based on the generalized method of moments (GMM, Hansen, 1982). Such GMM-based approaches have disadvantages of being lacking the ability to generate likelihood ratio-based confidence regions whose shape adapts to the support of the data. Moreover, the performance loss for the GMM can be substantial in small samples. The development of a likelihood ratio-based approach for nonsmooth models with missing data is well motivated.

The empirical likelihood (EL) and exponentially tilted likelihood (ET), known as non-parametric maximum likelihood methods, have been shown to be useful alternatives to GMM for finding estimators, constructing confidence regions, and testing hypotheses. Newey and Smith (2004) showed that the EL and ET estimators are members of a class of generalized empirical likelihood (GEL) estimators, which not only can improve the small sample properties of the GMM estimators but also compete very convincingly with the bootstrap (see, e.g., Owen, 1990; Qin and Lawless, 1994; Kitamura and Stutzer, 1997; Imbens et al., 1998). In this paper, we formulate the GEL and AIPW procedures for parameter identification and estimation in a collection of possibly over-identified nonsmooth moment functions in the presence of nonignorable missing data. There is surprisingly little discussion in the literature on this topic. Without considering missing values, Molanes-Lopez et al. (2009) and Parente and Smith (2011), respectively, developed the EL and GEL methods to make statistical inference on nonsmooth moment functions.

We have three contributions in this paper. First, we suggest a more attractive imputation procedure that not only can mitigate the effects of missing data, but also has ability to identify the parameters in the nonignorable propensity score model. The proposed imputation procedure is applicable under a general and easily-verified parametric model assumption on the respondent probability, and is developed based on a “kernel-assisted moment function imputation scheme”. Parametric identification is based on independence between a subset of the observed auxiliary variables called nonrespondent instrumental variables (Wang et al., 2014) and the missing indicator, conditional on the missing variables and other observed auxiliary variables. A set of the unbiased augmented inverse probability weighted moment functions (AIPW-MF) is constructed based on the proposed imputation approach. The use of nonparametric kernel approach makes the AIPW procedure robust against possible model misspecification.

Second, by applying the theory of GEL to the AIPW-MF, we construct a class of the estimated GEL ratio (GLR) statistics, and develop a class of two-step AIPW-based GEL (AIPW-GEL) estimates for parameters of interest. We systematically investigate asymptotic properties of our proposed two-step AIPW-GEL estimators and GLR statistics for the cases that the propensity score is parametrically estimated under a correctly specified parametric model. The large sample theories are established by using the results of modern empirical process theories including the uniform law of large numbers, stochastic equicontinuity and the Donsker class. The GEL confidence intervals for parameters of interest are constructed by using the bootstrap approximation to the distribution of the proposed GLR statistics.

Third, we propose a derivative-free optimization method based on the simulated annealing (SA, Kirkpatrick et al., 1983; Goffe et al., 1994) algorithm for the numerical implementation of the proposed two-step AIPW-GEL estimators. The proposed algorithm consists of an inner loop and an outer loop. Specifically, the inner loop solves the optimization problem

of Lagrange multipliers, which can be done by using the Newton-type methods; and the outer loop conducts the classical SA approach to minimize the concentrated GEL function for solving the optimization problem of unknown parameters defined via nonsmooth moment functions. The proposed algorithm is a sophisticated random search, which has been found in empirical studies to be successful at locating the global minima.

The rest of this paper is organized as follows. In Section 2, we study the identification of nonsmooth moment functions with parametric nonignorable propensity score model, discuss semi-parametric empirical likelihood estimation of propensity, and outline the formulation of the GEL procedure. In Section 3, we present the asymptotic results for the proposed method, and introduce a bootstrap calibration procedure. Section 4 discusses a modified SA algorithm. Simulation studies are conducted in Section 5. A data example for illustration is contained in Section 6. Section 7 concludes. Technical details are presented in the Appendix and a supplement material.

## 2 Methodology

### 2.1 Basic Setup

Let  $(X^\top, Y^\top)^\top$  be a  $(d_x + d_y)$ -dimensional vector of variables jointly distributed as a cumulative distribution function  $F(x, y)$ , where  $X \in R^{d_x}$ ,  $Y \in R^{d_y}$ , and  $F(x, y) \in \mathcal{F}$ , a class of distributions on a sample space. Let  $\beta$  be a vector of parameters of interest belonging to a compact subset  $\mathcal{B}$  of  $\mathcal{R}^p$ , and let  $g(X, Y, \beta)$  be a known vector-valued function with the dimension  $r \geq p$ . On  $\mathcal{F}$ , there is some  $\beta_0 \in \mathcal{B} \subset \mathcal{R}^p$  such that

$$E\{g(X, Y, \beta_0)\} = 0 \quad \text{w.p.1,} \quad (1)$$

where  $E\{\cdot\}$  represents the expectation taken with respect to  $F$ , the notation ‘w.p.1’ is the abbreviation of “with probability one”. Throughout this paper, it is assumed that the moment functions  $g$  are a class of nonsmooth functions with respect to  $\beta$ . Thus, many parametric models, such as the quantile regression model (Koenker, 2005), copulas (Nelsen, 1999), receiver operating characteristic curves (Pepe, 2003), quantile treatment effects (Cattaneo, 2010) are special cases of model (??).

We consider the observed sample  $\{(X_i, Y_i, \delta_i) : i = 1, \dots, n\}$ , which is an independently and identically distributed sample from  $(X, Y, \delta)$ . Here,  $\delta$  is a dichotomous variable denoting whether  $Y$  is missing or not, and  $Y_i$  is observed if and only if  $\delta_i = 1$ , whereas  $X_i$  is always available for  $i = 1, \dots, n$ . For the sake of simplicity, we suppose that the missing components have the same components across different individuals. Moreover, missing  $Y$  may represent response or covariates in a regression setting. We assume that the missing data mechanism is nonignorable in the sense that  $\delta_i \mid (X_i, Y_i) \sim \text{Bernoulli}(\pi_i)$ , where  $\pi_i = \pi(X_i, Y_i)$ , a function depends both on observed and missing variables. The main

interest of this paper is to make statistical inference on parameters  $\beta_0$  defined in (??) under the nonignorable missing data mechanism.

## 2.2 Identification

When nonignorable missing data are involved, model identification can be a crucial issue even if a fully parametric approach is adopted. To make unknown parameters under study identifiable, some additional assumptions on the missing data mechanism are required. Assume that the observable variables  $X$  can be decomposed as  $X = (U, Z) \in \mathcal{U} \times \mathcal{Z} \subset \mathcal{R}^{d_u} \times \mathcal{R}^{d_z}$  with  $0 < d_u < d_x$ , where  $U$  is continuously distributed, and  $Z$  can be continuous, discrete, or mixed. We assume that indicator  $\delta$  is independent of  $Z$  conditional on  $(U, Y)$ , that is,  $\delta \perp\!\!\!\perp Z \mid (U, Y)$ . In this case, we consider the following fully parametric propensity score model

$$\Pr(\delta = 1 \mid X, Y) = \Pr(\delta = 1 \mid U, Y) =: \pi(U, Y, \alpha_0), \quad (2)$$

where  $\pi$  is a known and smooth function with a  $l$ -dimensional unknown parameter  $\alpha_0 \in \mathcal{A}$ , a compact subset of  $\mathcal{R}^l$ . It follows from (??) that the respondent probability is not independent of the missing variable  $Y$  even after adjusting for the auxiliary variables  $X$ . In this case, the missing data mechanism is nonignorable (Little and Rubin, 2002). The excluded variable  $Z$  is referred to as nonrespondent instrument (Wang et al., 2014), which means that it helps to identify parameters in the considered respondent probability model, but it is not directly related with the response probability. In practice, the exclusion requirement (??) is reasonable because the conditional independence  $\delta \perp\!\!\!\perp Z \mid (U, Y)$  is more likely to hold when the dimension of the auxiliary variables  $X$  increases or  $Z$  is determined by the experimental design. Assuming (??), Wang et al. (2014) proposed using GMM approach to construct a root- $n$  consistent estimator for the unknown  $\alpha_0$ . In practical applications, we can specify (??) as some cumulative distribution functions or logistic regression models.

Denote  $W = (Z, Y)$ . Using some algebraic manipulations, we can show the following relationship

$$\frac{\Pr(W \in B \mid U, \delta = 0)}{\Pr(W \in B \mid U, \delta = 1)} = \frac{\Pr(\delta = 0 \mid W \in B, U) / \Pr(\delta = 1 \mid W \in B, U)}{\Pr(\delta = 0 \mid U) / \Pr(\delta = 1 \mid U)} \quad (3)$$

for any measurable set  $B$ .

Under assumption (??), the conditional odds of missing data is  $\Pr(\delta = 0 \mid X, Y) / \Pr(\delta = 1 \mid X, Y) = \Pr(\delta = 0 \mid U, Y) / \Pr(\delta = 1 \mid U, Y) = \pi^{-1}(U, Y, \alpha_0) - 1 =: O(U, Y, \alpha_0)$ . Following the argument of Kim and Yu (2011), we obtain

$$f_0(Z, Y \mid U) = f_1(Z, Y \mid U) \times \frac{O(U, Y, \alpha_0)}{E\{O(U, Y, \alpha_0) \mid U, \delta = 1\}}, \quad (4)$$

where  $f_\kappa(Z, Y \mid U) = f(Z, Y \mid U, \delta = \kappa)$  is the conditional density of  $(Z, Y)$  given  $U$  and

$\delta = \kappa$  for  $\kappa = 0$  and 1. Note that

$$f(Z, Y | U) = f_1(Z, Y | U)\Pr(\delta = 1 | U) + f_0(Z, Y | U)\Pr(\delta = 0 | U). \quad (5)$$

This, together with (??), implies that the unknown quantities in the joint density  $f(X, Y)$  and conditional density  $f(Z, Y | U)$  can be estimated based on the observed data distribution. For example, using identities (4) and (5),  $E\{g(X, Y, \beta)\}$  is equal to the following functional of the observed data distribution:

$$\begin{aligned} E\{g(X, Y, \beta)\} &= E\{E[g(X, Y, \beta) | U]\} \\ &= E\left\{\Pr(\delta = 1|U)m_g^1(U, \beta) + \Pr(\delta = 0 | U)m_g^0(U, \beta)\right\} \\ &= E\{\delta g(X, Y, \beta) + (1 - \delta)m_g^0(U, \beta)\}, \end{aligned}$$

where  $m_g^\kappa(U, \beta) = E\{g(X, Y, \beta) | U, \delta = \kappa\}$ ,  $\kappa = 0, 1$ . If the response mechanism is ignorable, then we have  $f_0(Z, Y | U) = f_1(Z, Y | U) = f(Z, Y | U)$  and  $m_g^0(U, \beta) = m_g^1(U, \beta) = E\{g(X, Y, \beta) | U\}$ . The identity (??) is a key to our methodology.

### 2.3 Augmented Moment Functions

From (??), it follows that we have

$$m_g^0(U, \beta) = \frac{E\{\delta g(X, Y, \beta)O(U, Y, \alpha_0) | U\}}{E\{\delta O(U, Y, \alpha_0) | U\}} =: m_g^0(U, \beta, \alpha_0). \quad (6)$$

It is assumed that  $m_g^0(U, \beta, \alpha) \in \mathcal{M}$  for all  $\beta \in \mathcal{B}$  and  $\alpha \in \mathcal{A}$ , where  $\mathcal{M}$  represents a subspace of smooth functions on  $\mathcal{U}$ .

Our method is based on the following AIPW moment functions

$$\tilde{g}_i(\beta, \alpha) = \frac{\delta_i g(X_i, Y_i, \beta)}{\pi(U_i, Y_i, \alpha)} - \frac{\delta_i - \pi(U_i, Y_i, \alpha)}{\pi(U_i, Y_i, \alpha)} m_g^0(U_i, \beta, \alpha). \quad (7)$$

The imputation procedure proposed in (??) is based on a projection of the moment functions  $g$  with nonignorable missing values onto the space generated by the non-excluded auxiliary variables  $U$  of nonrespondents. The following proposition shows that the proposed moment functions  $\tilde{g}_i(\beta_0, \alpha_0)$  are double robust when the propensity score (??) is of a special parametric model.

**PROPOSITION 1.** (i) Regardless of the choice of  $m_g^0(U_i, \beta, \alpha)$ ,  $\tilde{g}_i(\beta_0, \alpha_0)$  has mean zero provided that model for  $\pi(U_i, Y_i, \alpha_0)$  is correctly specified. (ii) Assume that the true response model is taken to be a parametric logistic model as  $\text{logit}\{\pi(U_i, Y_i, \alpha_0)\} = \varphi(U_i, \alpha_0) + q(Y_i)$ , where  $\varphi(\cdot)$  is a known smooth function in an unknown parameter vector  $\alpha_0$ , and  $q(\cdot)$  is an arbitrary user-specified (i.e., known) function that measures the amount of departure from the ignorable missing-data mechanism assumption. Then, the AIPW moment function  $\tilde{g}_i(\beta_0, \alpha_0)$  has mean zero even if model for  $\varphi(U_i, \alpha_0)$  is incorrectly specified.

**REMARK 1.** *Despite enjoying the attraction of doubly robust property, Proposition ?? (ii) has a very limited application scope because an ad hoc sensitivity analysis (Vansteelandt et al., 2007) should be proposed to deal with the situation where one does not know the information from the known part  $q(Y)$ . The nonrespondent instrument can successfully handle the identifiability issue of a fully parametric nonignorable propensity without any sensitivity analysis techniques, which is just the motivation of our work.*

It follows from Eq. (??) that the conditional expectation  $m_g^0(U, \beta, \alpha)$  is estimable using the observed data set  $\{(X_i, Y_i) \text{ for each } \delta_i = 1; i = 1, \dots, n\}$ . To enhance robustness against the potential model misspecification, we consider a nonparametric regression model for  $f_1(Z, Y | U)$  that leads to a nonparametric kernel estimator of  $m_g^0(U, \beta)$  given by

$$\hat{m}_g^0(U, \beta, \alpha_0) = \sum_{i=1}^n \mathcal{W}_{i0}(U, \alpha_0) g(X_i, Y_i, \beta),$$

where  $\mathcal{W}_{i0}(U, \alpha)$  is a point mass assigned to  $g(X_i, Y_i, \theta)$  and given by

$$\mathcal{W}_{i0}(U, \alpha) = \frac{\delta_i O(U_i, Y_i, \alpha) \mathcal{K}_h(U - U_i)}{\sum_{j=1}^n \delta_j O(U_j, Y_j, \alpha) \mathcal{K}_h(U - U_j)} = \frac{\mathcal{W}_{i1}(U) O(U_i, Y_i, \alpha)}{\sum_{j=1}^n \mathcal{W}_{j1}(U) O(U_j, Y_j, \alpha)}.$$

Moreover,  $\mathcal{W}_{i1}(U) = \delta_i \mathcal{K}_h(U - U_i) / \sum_{j=1}^n \delta_j \mathcal{K}_h(U - U_j)$ ,  $\mathcal{K}_h(\cdot) = \text{diag}(K_{h^{(1)}}^{(1)}(\cdot), \dots, K_{h^{(r)}}^{(r)}(\cdot))$ ,  $K_{h^{(\nu)}}^{(\nu)}(\cdot) = K^{(\nu)}(\cdot/h^{(\nu)})/h^{d_u^{(\nu)}}$ ,  $K^{(\nu)}$  is a  $d_u$ -dimensional kernel function and  $h^{(\nu)}$  is a bandwidth parameter for each  $\nu \in \{1, \dots, r\}$ . Note that kernel weight  $\mathcal{W}_{i1}(U)$  represents a point mass assigned to  $g(X_i, Y_i, \beta)$  for  $i = 1, \dots, n$  such that  $E\{g(X, Y, \beta) | U, \delta = 1\}$  can be approximated by kernel-based regression estimator  $\hat{m}_g(U, \beta) = \sum_{i=1}^n \mathcal{W}_{i1}(U) g(X_i, Y_i, \beta)$ , which has been commonly used to deal with ignorable missing data; for example, Cheng (1994) proposed using  $\hat{m}_g(U, \beta)$  with  $g(X_i, Y_i, \beta) = Y_i - \beta$  to develop a root- $n$  consistent nonparametric imputation estimator for mean response  $\beta_0 = E(Y)$ .

Under nonignorable missing data mechanism (2), a set of the propensity score-based and kernel-assisted nonsmooth functions for the  $i$ th individual is given by

$$\hat{g}_i(\beta, \alpha) = \frac{\delta_i g(X_i, Y_i, \beta)}{\pi(U_i, Y_i, \alpha)} - \frac{\delta_i - \pi(U_i, Y_i, \alpha)}{\pi(U_i, Y_i, \alpha)} \hat{m}_g^0(U_i, \beta, \alpha).$$

Note that the above introduced nonparametric AIPW procedure simultaneously achieves the identifiability of parameters and robustness against the potential model misspecification.

## 2.4 Propensity Score Estimation

To make the modified moment functions  $\hat{g}_i(\beta, \alpha)$  applicable, a consistent first-step estimator for  $\alpha_0$  should be given in advance. Instead of using GMM approach proposed in Wang et al. (2014), we in this paper employ the semiparametric empirical likelihood (SEL) method (Qin et al., 2002) to estimate  $\alpha_0$ . Since such method has been presented in existing literature

(Qin et al., 2002; Zhao et al., 2017b), we will only outline the main steps of the method. We consider a complete data likelihood function as follows:

$$L(\boldsymbol{\alpha}_0) = \prod_{i=1}^n \left\{ \pi(U_i, Y_i, \boldsymbol{\alpha}_0) dF(X_i, Y_i) \right\}^{\delta_i} \left[ \iint \{1 - \pi(U, Y, \boldsymbol{\alpha}_0)\} dF(X, Y) \right]^{1-\delta_i}.$$

Define  $p_i = dF(X_i, Y_i)$  and  $\boldsymbol{\omega}_0 = \iint \pi(U, Y, \boldsymbol{\alpha}_0) dF(X, Y)$ . We could obtain the following complete data log-likelihood function

$$l(\boldsymbol{\omega}_0, \boldsymbol{\alpha}_0) = \sum_{i=1}^n \delta_i \log p_i + \sum_{i=1}^n \delta_i \log \pi(U_i, Y_i, \boldsymbol{\alpha}_0) + (n - n_1) \log(1 - \boldsymbol{\omega}_0), \quad (8)$$

where  $n_1 = \sum_{i=1}^n \delta_i$ . An estimator of  $\boldsymbol{\alpha}_0$  can be obtained by maximizing  $l(\boldsymbol{\omega}, \boldsymbol{\alpha})$  subject to four constraints including  $\sum_{i=1}^n \delta_i p_i = 1$ ,  $p_i \geq 0$  for  $i = 1, \dots, n$ ,  $\sum_{i=1}^n \delta_i p_i \{\pi(U_i, Y_i, \boldsymbol{\alpha}) - \boldsymbol{\omega}\} = 0$  and  $\sum_{i=1}^n \delta_i p_i \phi(X_i, Y_i, \boldsymbol{\alpha}) = 0$ , where  $\phi(X_i, Y_i, \boldsymbol{\alpha})$  is an arbitrary user-specified  $\kappa$ -dimensional vector function satisfying  $E\{\phi(X_i, Y_i, \boldsymbol{\alpha}_0)\} = 0$ . The third constraint reflects the feature of missing not at random which is necessary. The fourth constraint is required for efficiency improvement and is constructed based on the auxiliary information from observed data.

By introducing Lagrange multipliers  $\lambda_1$  and  $\lambda_2$ , we obtain the optimal value of  $p_i$  as  $p_i = \delta_i n_1^{-1} \{1 + \lambda_1^\top \phi(X_i, Y_i, \boldsymbol{\alpha}_0) + \lambda_2 [\pi(U_i, Y_i, \boldsymbol{\alpha}_0) - \boldsymbol{\omega}_0]\}^{-1}$  for  $i = 1, \dots, n$ . Substituting  $p_i$ 's into (??) yields

$$\begin{aligned} l(\boldsymbol{\alpha}_0, \boldsymbol{\omega}_0, \lambda_1, \lambda_2) &= \sum_{i=1}^n \delta_i \log \pi(U_i, Y_i, \boldsymbol{\alpha}_0) + (n - n_1) \log(1 - \boldsymbol{\omega}) \\ &\quad - \sum_{i=1}^n \delta_i \log \{1 + \lambda_1^\top \phi(X_i, Y_i, \boldsymbol{\alpha}_0) + \lambda_2 [\pi(U_i, Y_i, \boldsymbol{\alpha}_0) - \boldsymbol{\omega}_0]\}. \end{aligned}$$

The consistent estimators of  $(\boldsymbol{\alpha}_0^\top, \boldsymbol{\omega}_0, \lambda_1^\top, \lambda_2)$ , say  $(\hat{\boldsymbol{\alpha}}^\top, \hat{\boldsymbol{\omega}}, \hat{\lambda}_1^\top, \hat{\lambda}_2)^\top$ , are defined as  $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\omega}}) = \arg \max_{\boldsymbol{\alpha}, \boldsymbol{\omega}} \inf_{\lambda_1, \lambda_2} l(\boldsymbol{\alpha}, \boldsymbol{\omega}, \lambda_1, \lambda_2)$  and  $(\hat{\lambda}_1, \hat{\lambda}_2) = \arg \min_{\lambda_1, \lambda_2} l(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\omega}}, \lambda_1, \lambda_2)$ .

## 2.5 Two-Step Generalized Empirical Likelihood

Let  $\rho(v)$  be a concave function of the scalar  $v \in \mathcal{V}$  (e.g., an open interval  $\mathcal{V}$  containing zero) and let  $\rho_j(v) = \partial^j \rho(v) / \partial v^j$  and  $\rho_j = \rho_j(0)$  for  $j \geq 1$ . Similar to Newey and Smith (2004), we impose a normalization on  $\rho(v)$  such that  $\rho_1 = \rho_2 = -1$ . For any given  $\boldsymbol{\alpha}$ , we construct the following recentred GEL criterion:

$$\hat{P}_n(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\alpha}) = \sum_{i=1}^n \{\rho(\boldsymbol{\lambda}^\top \hat{g}_i(\boldsymbol{\beta}, \boldsymbol{\alpha})) - \rho_0\} / n, \quad (9)$$

where  $\boldsymbol{\lambda}$  is an  $r$ -vector of auxiliary parameters. The factor  $\rho_0 = \rho(0)$  in the definition of  $\hat{P}_n(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\alpha})$  is for the convenience of asymptotic development and can be dropped for computational purposes.

Given SEL estimator  $\hat{\alpha}$ , a class of two-step AIPW-GEL estimators for  $\beta_0$  can be defined as a solution to the following saddle-point problem

$$\hat{\beta}_s = \arg \inf_{\beta \in \mathcal{B}} \sup_{\lambda \in \hat{\Lambda}_n(\beta, \hat{\alpha})} \hat{P}_n(\beta, \lambda, \hat{\alpha}), \quad (10)$$

where  $\hat{\Lambda}_n(\beta, \alpha) = \{\lambda : \lambda^\top \hat{g}_i(\beta, \alpha) \in \mathcal{V}, i = 1, \dots, n\}$ . For nonsmooth moment functions, AIPW-GEL estimators are no longer required to minimize (??), but satisfy

$$\hat{P}_n(\hat{\beta}_s, \hat{\lambda}_s, \hat{\alpha}) \leq \arg \inf_{\beta \in \mathcal{B}} \sup_{\lambda \in \hat{\Lambda}_n(\beta, \hat{\alpha})} \hat{P}_n(\beta, \lambda, \hat{\alpha}) + o_p(n^{-\sigma}),$$

where  $\sigma$  is nonnegative and  $\hat{\lambda}_s = \lambda(\hat{\beta}_s) = \arg \max_{\lambda \in \hat{\Lambda}_n(\hat{\beta}_s, \hat{\alpha})} \hat{P}_n(\hat{\beta}_s, \lambda, \hat{\alpha})$ . The empirical likelihood-based AIPW (AIPW-EL) estimator is obtained by taking  $\rho(v) = \log(1 - v)$  and  $\mathcal{V} = (-\infty, 1)$ , whereas the exponential tilting-based AIPW (AIPW-ET) estimator is constructed by setting  $\rho(v) = -\exp(v)$ . In addition, the implied GEL empirical probabilities associated with each AIPW-GEL estimator are given by

$$\hat{p}_i = \frac{\rho_1(\hat{\lambda}_s^\top \hat{g}_i(\hat{\beta}_s, \hat{\alpha}))}{\sum_{j=1}^n \rho_1(\hat{\lambda}_s^\top \hat{g}_j(\hat{\beta}_s, \hat{\alpha}))}, \quad i = 1, \dots, n.$$

The empirical conditional probabilities  $\hat{p}_i$  ( $i = 1, \dots, n$ ) sum to one by construction and satisfy the sample moment condition  $\sum_{i=1}^n \hat{p}_i \hat{g}_i(\hat{\beta}_s, \hat{\alpha}) = 0$ .

### 3 Main Results

In this section, using the modern empirical process theory for statistics (see, e.g., Pakes and Pollard, 1989; Van der Vaart and Wellner, 1996), we investigate the large sample properties of the proposed two-step AIPW-GEL estimators given in Section 2. We use  $\xrightarrow{\mathcal{L}}$  to denote convergence in distribution.

Let  $\eta = (\alpha^\top, \omega, \gamma^\top)^\top$  with  $\gamma = \lambda_1(1 - \omega)$ , and  $\eta_0 = (\alpha_0^\top, \omega_0, 0)^\top$  be the true value of  $\eta$ . Denote  $\hat{\eta} = (\hat{\alpha}^\top, \hat{\omega}, \hat{\gamma}^\top)^\top$  as the estimator of  $\eta$ , where  $\hat{\gamma} = \hat{\lambda}_1(1 - \hat{\omega})$ . The following proposition shows that the SEL estimator  $\hat{\alpha}$  proposed in Section 2.5 is consistent and asymptotically normal.

**PROPOSITION 2.** (Zhao et al., 2017b) *Suppose that Assumptions (C1)-(C2) given in the Appendix hold. We have the following results: (i)  $\eta_0$  is locally identified if and only if  $\text{rank}(\mathbb{A}) = l + \kappa + 1$ ; (ii)  $\hat{\eta} \xrightarrow{P} \eta_0$  and  $n^{1/2}(\hat{\eta} - \eta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathbb{A}^{-1}\mathbb{B}(\mathbb{A}^{-1})^\top)$ , where  $\mathbb{A}$  and  $\mathbb{B}$  are defined in the Supplementary Material.*

From Proposition ??, an asymptotic linear expansion for  $\hat{\alpha}$  could be defined as:  $n^{1/2}(\hat{\alpha} - \alpha_0) = n^{-1/2} \sum_{i=1}^n \Psi_i(\alpha_0) + o_p(1)$ , where  $\Psi_i(\alpha_0) = \Psi(X_i, Y_i, \alpha_0)$  is an influence function

that is defined in the Supplementary Material. Let  $V_1 = E\{\tilde{g}_i(\beta_0, \alpha_0)\tilde{g}_i(\beta_0, \alpha_0)^\top\}$ ,  $V_2 = \text{Var}\{\tilde{g}_i(\beta_0, \alpha_0) - \Xi\Psi_i(\alpha_0)\}$ ,  $\Xi = \text{Cov}\{\tilde{g}_i(\beta_0, \alpha_0), \Delta(U, Y, \alpha_0)\}$ ,

$$\Delta(U, Y, \alpha) = \{\delta - \pi(U, Y, \alpha)\} \frac{\partial \text{logit}\{\pi(U, Y, \alpha)\}}{\partial \alpha^\top},$$

and  $\Gamma = \partial E\{g(X, Y, \beta)\} / \partial \beta^\top |_{\beta=\beta_0}$ . Then, we have the following theorem.

**THEOREM 1.** *Suppose that assumptions (A1), (A2), (B1), (B2) and C given in the Appendix hold, the respondent probability model  $\pi(U, Y, \alpha_0)$  is correctly specified and two-step AIPW-GEL estimator  $\hat{\beta}_s$  is obtained by solving (??) with  $\hat{\alpha}$  computed by SEL approach. Then, we have  $\hat{\beta}_s - \beta_0 = o_p(1)$ . If in addition, assumptions (A3), (A4), (B3) and (B4) given in the Appendix hold, thus we obtain*

$$n^{1/2}(\hat{\beta}_s - \beta_0) \xrightarrow{L} \mathcal{N}(0, \Sigma_s),$$

where  $\Sigma_s = (\Gamma^\top V_1^{-1} \Gamma)^{-1} \Gamma^\top V_1^{-1} V_2 V_1^{-1} \Gamma (\Gamma^\top V_1^{-1} \Gamma)^{-1}$ .

Theorem ?? has some interesting implications. First, Theorem ?? indicates that the efficiency of the proposed estimators depends on the correlation between the efficient score function  $\tilde{g}_i(\beta_0, \alpha_0)$  and the influence function  $\Psi_i(\alpha_0)$ . In particular, if  $\Xi \text{Var}\{\Psi_i(\alpha_0)\} \Xi^\top - 2\Xi \text{Cov}\{\tilde{g}_i(\beta_0, \alpha_0), \Psi_i(\alpha_0)\} \leq 0$ , the proposed two-step AIPW-GEL estimators achieve efficiency gain over the estimators computed by using the known propensity score. This is a common phenomenon under missing at random setup (Robins et al. 1994). Second, if  $\tilde{g}_i(\beta, \alpha_0)$  is orthogonal to the score  $\Delta(U, Y, \alpha_0)$ , that is,  $\Xi = 0$ , thus the limit distribution of  $\hat{\beta}_s$  is invariant to that of  $\hat{\alpha}$ .

Note that the asymptotic variance of the proposed AIPW-GEL estimators contain derivative and variance terms. In the nonsmooth case, the derivative terms are not easy to estimate since derivatives of the objective functions are no longer available. Hence the Wald-type confidence regions for  $\beta_0$  are difficult to establish. The following theorems show that the proposed AIPW-GEL ratio statistics provide a convenient framework for developing confidence regions.

**THEOREM 2.** *Assume that the conditions given in Theorem ?? hold. As  $n \rightarrow \infty$ , we have  $2n\hat{P}_n(\beta_0, \lambda(\beta_0), \hat{\alpha}) \xrightarrow{L} Q^\top \Omega Q$  under the null hypothesis  $H_0 : \beta = \beta_0$ , where  $\Omega = V_2^{1/2} V_1^{-1} V_2^{1/2}$  and  $Q$  is a  $r$ -dimensional standard normal random vector (i.e.,  $Q \sim \mathcal{N}(0, I_r)$ ).*

Theorem ?? indicates that the proposed AIPW-GEL ratio converges to a linear combination of independent chi-square distributions. Despite the loss of Wilks' theorem, confidence regions based on the AIPW-GEL ratio  $2n\hat{P}_n(\beta_0, \lambda(\beta_0), \hat{\alpha})$  are still appealing in that they are range preserving and transformation respecting due to their likelihood ratio based nature. To construct GEL-based confidence regions for  $\beta_0$ , we may approximate the distribution of  $2n\hat{P}_n(\beta_0, \lambda, \hat{\alpha})$  via resampling. Let  $\mathcal{X}_m^* = \{(X_i^*, Y_i^*, \delta_i^*) : i = 1, \dots, m\}$  be a bootstrap sample from  $\{\mathcal{X}_n = (X_j, Y_j, \delta_j) : j = 1, \dots, n\}$ . Based on  $\mathcal{X}_m^*$ , we compute

the bootstrap estimator  $\hat{\alpha}^*$  of  $\alpha_0$  via the previously mentioned SEL approach. Then, the bootstrap version of  $\hat{P}_n(\beta_0, \lambda, \hat{\alpha})$  is defined as

$$\hat{P}_m^*(\hat{\beta}_S, \lambda^*, \hat{\alpha}^*) = \sum_{i=1}^m \{\rho(\lambda^{*\top} \hat{g}(X_i^*, Y_i^*, \hat{\beta}_S, \hat{\alpha}^*)) - \rho_0\} / m,$$

where  $\lambda^* = \arg \max_{\lambda} \hat{P}_m^*(\hat{\beta}_S, \lambda, \hat{\alpha}^*)$ . The following theorem justifies the bootstrap procedure.

**THEOREM 3.** *Assume that the conditions given in Theorem ?? hold. Then, the conditional distribution of  $2m\hat{P}_m^*(\hat{\beta}_S, \lambda^*, \hat{\alpha}^*)$  given the original sample  $\mathcal{X}_n$  converges to the distribution of  $Q^\top \Omega Q$ , w. p. 1, as  $n \rightarrow \infty$  and  $m \rightarrow \infty$ .*

Let  $c_\alpha^*$  be the  $100(1 - \alpha)\%$  quantile of the distribution of  $2m\hat{P}_m^*(\hat{\beta}_S, \lambda^*, \hat{\alpha}^*)$  evaluated by resampling method. Then, it follows from Theorem ?? that the bootstrap empirical log-likelihood confidence region at the nominal coverage level  $1 - \alpha$  is given by  $C_\alpha = \{\beta : 2n\hat{P}_n(\beta, \lambda, \hat{\alpha}) \leq c_\alpha^*\}$ .

**REMARK 2.** *The results obtained here are still valid if we respectively replace  $m_g^0(U, \beta, \alpha)$  and  $\mathcal{W}_{i0}(U, \alpha)$  with  $m_g^0(X, \beta, \alpha)$  and*

$$\mathcal{W}_{i0}(X, \alpha) = \frac{\delta_i O(U_i, Y_i, \alpha) \mathcal{K}_h(X - X_i)}{\sum_{j=1}^n \delta_j O(U_j, Y_j, \alpha) \mathcal{K}_h(X - X_j)},$$

in which  $\mathcal{K}_h(\cdot)$  is defined based on a  $d_x$ -dimensional kernel function. That is, an alternative set of unbiased AIPW moment functions could be constructed structurally identical to  $\tilde{g}_i(\beta, \alpha)$  in (??) except that  $m_g^0(U_i, \beta, \alpha)$  is replaced by  $m_g^0(X_i, \beta, \alpha)$ .

Assume that  $Z$  is the vector of discrete components that take on at most a finite number of values (see, e.g., Wang, et al., 2014). Then using the arguments of Andrews (1995), the results obtained above continue to be valid if we redefine  $\mathcal{W}_{i0}(U_j, \alpha)$  as

$$\mathcal{W}_{i0}(U_j, \alpha) = \frac{\delta_i O(U_i, Y_i, \alpha) \mathcal{K}_h(U_i - U_j) I\{Z_i = Z_j\}}{\sum_{j=1}^n \delta_j O(U_j, Y_j, \alpha) \mathcal{K}_h(U_i - U_j) I\{Z_i = Z_j\}}.$$

## 4 Computation

Computing the proposed two-step AIPW-GEL estimators is computationally challenging because of the nonsmooth moment functions involved, which leads to the fact that all gradient functions are not well defined. To overcome this difficulty, we develop a derivative-free approach to implement the numerical optimization based on the simulated annealing (SA) algorithm. The SA algorithm is a kind of calculation precision of the random search algorithm, which has advantages over other local search methods because of its flexibility and its ability to achieve the global optimality. Assume that the estimator  $\hat{\alpha}$  of  $\alpha_0$  has

been obtained. The modified SA algorithm consists of an inner loop and an outer loop. The detailed steps are as follows.

*Inner Loop* step. This step is to find the Lagrange multiplier  $\hat{\lambda}_s = \arg \max_{\lambda} \hat{P}_n(\boldsymbol{\beta}, \lambda, \hat{\boldsymbol{\alpha}})$  for a given  $\boldsymbol{\beta}$ . Let

$$\begin{aligned}\hat{P}_{\lambda}(\boldsymbol{\beta}, \lambda) &= \sum_{i=1}^n \rho_1(\lambda^{\top} \hat{g}_i(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}})) \hat{g}_i(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}), \\ \hat{P}_{\lambda\lambda}(\boldsymbol{\beta}, \lambda) &= \sum_{i=1}^n \rho_2(\lambda^{\top} \hat{g}_i(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}})) \hat{g}_i(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}) \hat{g}_i(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}})^{\top}.\end{aligned}$$

Then, the modified Newton-Raphson method for finding  $\hat{\lambda}_s$  is implemented via the following iterative equation:

$$\lambda^{(t+1)} = \lambda^{(t)} - \varrho \{ \hat{P}_{\lambda\lambda}(\boldsymbol{\beta}, \lambda^{(t)}) \}^{-1} \hat{P}_{\lambda}(\boldsymbol{\beta}, \lambda^{(t)}),$$

where  $\varrho > 0$  is a scalar steplength, and  $\lambda^{(t)}$  is the value of  $\lambda$  at the  $t$ -th iteration. Here, the initial value  $\lambda^{(0)}$  of  $\lambda$  is taken to be zero vector. Implementing the iteration considered above until the gradient function  $\hat{P}_{\lambda}(\boldsymbol{\beta}, \lambda_t)$  is smaller than some prespecified tolerance, such as 0.0001. Following Hansen (2015), we may set  $\varrho$  to be  $\varrho = (\hat{P}_2 + 3\hat{P}_0 - 4\hat{P}_1) / (4\hat{P}_2 + 4\hat{P}_0 - 8\hat{P}_1)$ , where  $\hat{P}_s = \hat{P}_n(\boldsymbol{\beta}, \tilde{\lambda}^s, \hat{\boldsymbol{\alpha}})$  in which  $\tilde{\lambda}^s = \lambda^{(t)} - \varrho_s \{ \hat{P}_{\lambda\lambda}(\boldsymbol{\beta}, \lambda^{(t)}) \}^{-1} \hat{P}_{\lambda}(\boldsymbol{\beta}, \lambda^{(t)})$  for  $s = 0, 1, 2$ , and  $\varrho_0 = 0$ ,  $\varrho_1 = 1/2$  and  $\varrho_2 = 1$ . For the AIPW-EL approach, at each iteration step, we need to check if the condition  $n\{1 - \lambda^{\top} \hat{g}_i(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}})\} \geq 1$  holds for all  $i$ . When the condition does not hold, we decrease the steplength  $\varrho$  until this condition is satisfied. A detailed discussion of the convergence problem of the above algorithm can be found in Chen, Sitter and Wu (2002).

*Outer Loop* step. Once  $\hat{\lambda}_s$  is obtained in the inner loop, we conduct the following minimization step:  $\hat{\boldsymbol{\beta}}_s = \arg \inf_{\boldsymbol{\beta} \in \mathcal{B}} \hat{P}_n(\boldsymbol{\beta}, \hat{\lambda}_s, \hat{\boldsymbol{\alpha}})$ . This can be done by using the classical SA algorithm. More details on the SA algorithm and its implementation can refer to Goffe et al. (1994). In the smooth case, this step can also be done by using the Newton-Raphson method.

The modified SA algorithm introduced above retains the main advantages of the classical SA method in that the objective function for implementing the SA algorithm only is a profile function of  $\boldsymbol{\beta}$ , although nuisance parameters are involved. Thus the global convergence property of the *Outer Loop* step is guaranteed by the classical SA algorithm. Moreover, the proposed algorithm is rather general and useful because it can deal with various optimization problem regardless of linear or nonlinear relationship between variables and parameters, and smooth or nonsmooth objective functions with or without missing data.

## 5 Simulation Studies

In this section, several simulation studies were conducted to evaluate the finite sample performance of the proposed methodologies.

**Experiment 1.** In this experiment, we simulated data from the following instrumental variable quantile regression (IVQR) model

$$\begin{aligned} Y &= \beta_1 + \zeta\beta_2 + \sigma(X_1, X_2)(\varepsilon - Q_\varepsilon(\tau)), \\ \zeta &= (X_1 + X_2)/3 + \varepsilon + \varpi, \end{aligned} \quad (11)$$

where  $\beta_0 = (\beta_1, \beta_2)^\top = (1, 0.5)^\top$ ,  $Q_\varepsilon(\tau)$  is the conditional  $\tau$ -quantile of  $\varepsilon$ ,  $X_1 \sim \chi_1^2$ ,  $X_2 \sim \chi_2^2$ ,  $\varepsilon \sim \mathcal{N}(0, 1)$  and  $\varpi \sim \mathcal{N}(0, 1)$ , in which  $\chi_k^2$  represents the chi-squared distribution with  $k$  degrees of freedom. Following Parente and Smith (2011), we considered two scenarios for  $\sigma(X_1, X_2)$ : (i)  $\sigma(X_1, X_2) = 1$ , and (ii)  $\sigma(X_1, X_2) = \sqrt{3/14}\{1 + (X_1 + X_2)/3\}$ , which is used to investigate the effect of heteroskedasticity.

We assumed that  $X = (X_1, X_2)^\top$  was completely observed, but  $Y$  might be subject to missingness. Denote  $\delta = 1$  if  $Y$  is observed, and  $\delta = 0$  if  $Y$  is missing. The following model was used to generate respondent indicator  $\delta$ :

$$\Pr(\delta = 1|X, Y) = \frac{\exp(a + \alpha_1 X_2 + \alpha_2 Y)}{1 + \exp(a + \alpha_1 X_2 + \alpha_2 Y)} =: \pi(X_2, Y, \alpha_0), \quad (12)$$

where  $\alpha_0 = (a, \alpha_1, \alpha_2)^\top = (a, 0.05, 0.01)^\top$ , and  $a$  was set to be 2.0, 1.5, 1.0 and 0.5, respectively, leading to different missing proportions. In Eq. (??), the respondent indicator depends on missing variable  $Y$ , but is independent of covariate  $X_1$  given  $(X_2, Y)$ . Thus, covariate  $X_1$  was treated as a nonrespondent instrument, which helps to make parameter  $\alpha_0$  in (??) identifiable. Under the above settings, we have  $U = X_2$  and  $Z = X_1$ .

We considered three quantile levels (e.g.,  $\tau = 0.25, 0.5$  and  $0.75$ ) for each of the above four respondent probabilities. Newey and Smith (2004) showed that continuous updating is also a member of the GEL class. However, like GMM approach, continuous updating has disadvantages of lacking the ability of generating likelihood ratio-based confidence regions whose shape adapts to the support of the data. Thus, in this experiment, we focused only on EL and ET inferences. For each of twelve combinations of three quantile levels and four respondent probabilities, we independently generated 1,000 data sets  $\{(Y_i, X_i, \delta_i) : i = 1, \dots, n\}$  with  $n = 200$  according to the IVQR model (??) together with the propensity score model (??). Here  $X_i = (X_{1i}, X_{2i})^\top$ . For a given data set  $\{(X_i, Y_i, \delta_i) : i = 1, \dots, n\}$ , we computed the two-step AIPW-EL and AIPW-ET estimators of  $\beta_0$  based on the nonsmooth moment functions:  $g(X_i, Y_i, \beta) = X_i\{I(Y_i \leq \beta_1 + \zeta_i\beta_2) - \tau\}$  satisfying the restrictions  $E\{g(X_i, Y_i, \beta_0)\} = 0$  for  $i = 1, \dots, n$ . In evaluating nonparametric kernel estimation, i.e., computing the nonparametric kernel estimator of conditional expectation  $m_g^0(X_{2i}, \beta) = E\{g(X, Y, \beta) | X_{2i}, \delta_i = 0\}$  for  $i = 1, \dots, n$ , we took the one-dimensional Gaussian kernel function with a fixed bandwidth  $h = 1.25\hat{\sigma}_{x_2}n^{-1/5}$ , where  $\hat{\sigma}_{x_2}$  is the sample standard deviation of  $X_2$ . The initial values for  $\beta_0$  in the SA algorithm were computed by using the complete-case analysis. The SEL approach discussed in Section 2.5 was employed to compute estimator  $\hat{\alpha}$  of  $\alpha_0$ . The auxiliary information was defined as  $\phi(X_i, Y_i, \alpha) = \delta_i\pi^{-1}(X_{2i}, Y_i, \alpha)(X_i - \bar{X})$ , where  $\bar{X} = n^{-1}\sum_{i=1}^n X_i$ . The initial values for

$\alpha_0$  were set to be  $(\tilde{\zeta}, 0)$ , where  $\tilde{\zeta} = (\tilde{a}, \tilde{\alpha}_1)$  was obtained by maximizing the log-binomial likelihood given by  $\log\{\prod_{i=1}^n p(X_{2i}, \varsigma)^{\delta_i} (1 - p(X_{2i}, \varsigma))^{1-\delta_i}\}$  with  $\text{logit}\{p(X_2, \varsigma)\} = a + \alpha_1 X_2$ .

Results for 1,000 repetitions under the considered twelve cases are presented in Tables 1 and 2 for homoscedasticity (i.e., case (i) of  $\sigma(X_1, X_2)$ ) and heteroscedasticity (i.e., case (ii) of  $\sigma(X_1, X_2)$ ), respectively. In tables, ‘Bias’ represents the difference between the true value and the mean of 1,000 estimates, ‘SD’ and ‘RMS’ denote the standard deviation of 1,000 estimates and the root mean square between the estimates of 1,000 repetitions and its true value, respectively, ‘AL’ is the average length of 1,000 EL or ET-based 95% confidence intervals, and ‘CP’ is the proportion among 1,000 95% confidence intervals that covered the true value of parameter. The proposed bootstrap calibration method with  $B = 1,000$  bootstrap replications was used to compute the critical value of the limiting distribution of AIPW-GEL ratio statistics.

Examining Tables 1 and 2 reveals the following findings. (i) Under all the considered circumstances, both AIPW-EL and AIPW-ET methods produce unbiased estimates of  $\beta_0$  in that the absolute values of their Biases are less than 0.07 and their RMS values are quite close to their corresponding SD values, which are consistent with our established theoretical properties. (ii) The coverage probabilities of the resultant confidence intervals based on our proposed method are rather close to the pre-specified nominal level 95%. (iii) The missing rate improves the accuracy of parameter estimate and the empirical coverage of confidence interval. (v) The AIPW-ET method has consistently lower coverage probability and shorter average length than the AIPW-EL method. (vi) The AIPW-EL estimator has smaller RMS and SD values than the AIPW-ET estimator.

**Experiment 2** In this experiment, we compared the proposed AIPW-GEL approach with two existing methods, including Tang and Qin’s (2012) nonparametric multiple imputation assuming that missing data are ignorable, and the naive nonparametric imputation using only the follow-up data. The other purpose of this experiment was to examine the robustness of the proposed two-step AIPW-GEL estimators to the misspecified nonignorable parametric propensity score model.

We independently simulated 500 data sets  $\{(X_i, Y_i) : i = 1, \dots, n\}$  with  $n = 100$  from the two-dimensional multiplicative regression model  $Y_i = \exp(X_i^\top \beta_0) \varepsilon_i$ , where  $\beta_0 = (\beta_1, \beta_2)^\top = (0.5, 1)^\top$ ,  $X_i = (X_{1i}, X_{2i})^\top$  and  $X_i$ ’s were independently generated from the bivariate normal distribution  $\mathcal{N}(0, \Sigma_x)$  with  $\Sigma_x = (\sigma_x^{kj})$ , in which  $\sigma_x^{kj} = 0.5^{|k-j|}$  for  $1 \leq k, j \leq 2$ , and  $\varepsilon_i$ ’s were independently drawn from the following distribution assumptions: (A)  $\log(\varepsilon_i) \sim \mathcal{N}(0, 1)$  and (B)  $\log(\varepsilon_i) \sim \text{Uniform}(0, 1)$ . Similarly, we assumed that  $X_i$ ’s are completely observed, but  $Y_i$ ’s may be subject to missingness. The respondent indicator  $\delta_i$  for  $Y_i$  was generated from the Bernoulli distribution with probability  $\pi_i(\alpha_0)$  specified by

$$\pi_i(\alpha_0) = \frac{\exp(a + \alpha_1 X_{2i} + \alpha_2 Y_i)}{1 + \exp(a + \alpha_1 X_{2i} + \alpha_2 Y_i)}, \quad (13)$$

Table 1: Simulation results for the IVQR model with homoscedasticity assumption.

$\tau$	$\beta_0$	Statistic	EL				ET			
			$a = 2$	$a = 1.5$	$a = 1$	$a = 0.5$	$a = 2$	$a = 1.5$	$a = 1$	$a = 0.5$
0.25	$\beta_1$	Bias	0.004	0.006	0.011	0.010	0.009	0.006	0.008	0.008
		RMS	0.144	0.142	0.153	0.165	0.181	0.175	0.187	0.190
		SD	0.144	0.142	0.153	0.165	0.181	0.175	0.187	0.190
		CP	0.952	0.942	0.934	0.930	0.935	0.924	0.921	0.915
		AL	0.355	0.375	0.418	0.484	0.319	0.331	0.349	0.418
	$\beta_2$	Bias	-0.039	-0.047	-0.049	-0.050	-0.059	-0.055	-0.051	-0.065
		RMS	0.138	0.147	0.152	0.158	0.201	0.196	0.200	0.209
		SD	0.132	0.139	0.144	0.150	0.193	0.189	0.193	0.199
		CP	0.962	0.947	0.940	0.936	0.943	0.930	0.928	0.926
		AL	0.363	0.383	0.426	0.492	0.327	0.339	0.357	0.426
0.5	$\beta_1$	Bias	0.000	0.003	-0.002	-0.003	0.004	0.008	0.001	0.006
		RMS	0.139	0.150	0.147	0.155	0.193	0.193	0.205	0.201
		SD	0.139	0.150	0.147	0.155	0.193	0.193	0.205	0.201
		CP	0.949	0.958	0.963	0.936	0.942	0.948	0.947	0.928
		AL	0.197	0.203	0.209	0.240	0.189	0.184	0.190	0.221
	$\beta_2$	Bias	-0.009	-0.014	-0.009	-0.010	-0.009	-0.016	-0.009	-0.019
		RMS	0.098	0.104	0.100	0.106	0.129	0.132	0.135	0.140
		SD	0.098	0.103	0.100	0.105	0.129	0.131	0.135	0.139
		CP	0.952	0.960	0.966	0.939	0.945	0.949	0.951	0.931
		AL	0.205	0.211	0.217	0.248	0.197	0.192	0.198	0.229
0.75	$\beta_1$	Bias	-0.009	-0.005	-0.009	0.010	-0.006	0.000	-0.014	-0.002
		RMS	0.160	0.172	0.184	0.186	0.214	0.207	0.224	0.225
		SD	0.160	0.172	0.184	0.186	0.214	0.207	0.223	0.225
		CP	0.939	0.946	0.928	0.934	0.936	0.935	0.911	0.908
		AL	0.158	0.158	0.173	0.204	0.158	0.147	0.158	0.198
	$\beta_2$	Bias	-0.002	-0.003	-0.002	-0.012	-0.005	-0.007	0.000	-0.008
		RMS	0.088	0.088	0.090	0.098	0.106	0.107	0.108	0.113
		SD	0.088	0.088	0.090	0.097	0.106	0.107	0.108	0.113
		CP	0.945	0.950	0.931	0.938	0.943	0.936	0.915	0.911
		AL	0.166	0.166	0.181	0.212	0.166	0.155	0.166	0.206

Table 2: Simulation results for the IVQR model with heteroscedasticity assumption.

$\tau$	$\beta_0$	Statistic	EL				ET			
			$a = 2$	$a = 1.5$	$a = 1$	$a = 0.5$	$a = 2$	$a = 1.5$	$a = 1$	$a = 0.5$
0.25	$\beta_1$	Bias	0.006	0.003	0.005	-0.002	0.003	0.005	0.008	0.000
		RMS	0.144	0.149	0.152	0.160	0.174	0.175	0.180	0.184
		SD	0.144	0.149	0.152	0.160	0.175	0.175	0.180	0.184
		CP	0.947	0.948	0.941	0.927	0.933	0.937	0.929	0.914
		AL	0.466	0.497	0.539	0.587	0.417	0.422	0.469	0.510
	$\beta_2$	Bias	-0.048	-0.051	-0.044	-0.057	-0.055	-0.053	-0.069	-0.064
		RMS	0.158	0.154	0.152	0.170	0.214	0.206	0.214	0.221
		SD	0.151	0.145	0.146	0.161	0.207	0.199	0.203	0.212
		CP	0.953	0.950	0.944	0.931	0.939	0.941	0.934	0.916
		AL	0.474	0.505	0.547	0.595	0.425	0.430	0.477	0.518
0.5	$\beta_1$	Bias	0.001	0.008	0.000	0.004	0.007	0.009	0.007	0.012
		RMS	0.132	0.140	0.139	0.152	0.184	0.188	0.179	0.194
		SD	0.132	0.140	0.139	0.152	0.184	0.188	0.179	0.194
		CP	0.952	0.953	0.943	0.943	0.946	0.946	0.930	0.933
		AL	0.246	0.254	0.252	0.274	0.219	0.229	0.218	0.254
	$\beta_2$	Bias	-0.014	-0.018	-0.013	-0.022	-0.018	-0.020	-0.020	-0.029
		RMS	0.114	0.117	0.117	0.121	0.150	0.157	0.147	0.157
		SD	0.113	0.116	0.117	0.119	0.149	0.155	0.146	0.155
		CP	0.952	0.954	0.944	0.947	0.946	0.948	0.932	0.937
		AL	0.254	0.262	0.260	0.282	0.227	0.237	0.226	0.262
0.75	$\beta_1$	Bias	-0.012	-0.009	-0.002	0.006	0.007	0.007	0.010	-0.012
		RMS	0.160	0.168	0.168	0.184	0.204	0.210	0.218	0.216
		SD	0.159	0.168	0.168	0.184	0.204	0.210	0.217	0.215
		CP	0.944	0.952	0.937	0.929	0.935	0.937	0.922	0.909
		AL	0.206	0.205	0.209	0.233	0.187	0.190	0.198	0.212
	$\beta_2$	Bias	-0.002	-0.004	-0.006	-0.011	-0.012	-0.011	-0.013	-0.002
		RMS	0.095	0.101	0.105	0.109	0.117	0.120	0.124	0.124
		SD	0.095	0.101	0.104	0.108	0.117	0.120	0.123	0.125
		CP	0.949	0.954	0.941	0.934	0.940	0.940	0.925	0.916
		AL	0.214	0.213	0.217	0.241	0.195	0.198	0.206	0.220

where  $\alpha_0 = (a, \alpha_1, \alpha_2)^\top = (a, 0.05, 0.01)^\top$  with the true value of  $a$  being 1.0, 0.5 and 0.01, respectively, leading to different missing proportions. Clearly, the respondent indicator  $\delta_i$  depended on missing variable  $Y_i$ , but was independent of covariate  $X_{1i}$  given  $(X_{2i}, Y_i)$ . Thus, covariate  $X_{1i}$  was treated as a nonrespondent instrument, that helps to make parameter  $\alpha_0$  identifiable.

Without missing data, following the argument of Chen et al. (2010), parameters in multiplicative regression model can be estimated by minimizing the following least absolute relative error (LARE)

$$\text{LARE}_n(\beta) = \sum_{i=1}^n \left\{ \left| \frac{Y_i - \exp(X_i^\top \beta)}{Y_i} \right| + \left| \frac{Y_i - \exp(X_i^\top \beta)}{\exp(X_i^\top \beta)} \right| \right\}. \quad (14)$$

It is easily seen that the LARE function (??) is piecewisely differentiable with respect to  $\beta$ . Let  $\varepsilon_i(\beta) = Y_i / \exp(X_i^\top \beta)$  and  $g(X_i, Y_i, \beta) = \{\varepsilon_i^{-1}(\beta) + \varepsilon_i(\beta)\} \text{sgn}\{\varepsilon_i(\beta) - 1\} X_i$ , which are a set of nonsmooth functions with respect to  $\beta$ . Following Li et al. (2014), the solution to minimizing (??) is equivalent to the solution to estimating equations  $n^{-1} \sum_{i=1}^n g(X_i, Y_i, \beta) = 0$ . Thus, for a given combination of two error distribution assumptions and three respondent probabilities, and each of 500 generated data sets, based on objective functions  $g(X_i, Y_i, \beta)$ , we computed the following six types of GEL estimators for  $\beta_0$ :

S1. the proposed AIPW-EL and AIPW-ET estimators via the correctly specified respondent probability model (??);

S2. the proposed AIPW-EL and AIPW-ET estimators via the misspecified respondent probability model:  $\pi_i(\alpha_0) = \Phi(a + \alpha_1 X_{2i} + \alpha_2 Y_i)$ , where  $\Phi(\cdot)$  is the cumulative distribution function of the standard normal distribution;

S3. the same as S1 except that kernel weight  $\mathcal{W}_{i0}(X_{2j}, \alpha)$  is replaced with  $\mathcal{W}_{i0}(X_j, \alpha)$ , which is defined in remark ??;

S4. the same as S2 except that kernel weight  $\mathcal{W}_{i0}(X_{2j}, \alpha)$  is replaced with  $\mathcal{W}_{i0}(X_j, \alpha)$ ;

S5. EL and ET estimators under the ignorable assumption of missing response, which were evaluated by using the following inverse probability weighted estimating functions with multiple imputation (Tang and Qin, 2012):

$$\widehat{g}_i^M(\beta) = \frac{\delta_i}{\widehat{\pi}(X_i)} g(X_i, Y_i, \beta) + \left\{ 1 - \frac{1}{\widehat{\pi}(X_i)} \right\} \frac{1}{\kappa} \sum_{j=1}^{\kappa} g(X_i, \widetilde{Y}_{ij}, \beta),$$

where  $\widehat{\pi}(X_i) = \sum_{j=1}^n \delta_j \mathcal{K}_h(X_i - X_j) / \sum_{j=1}^n \mathcal{K}_h(X_i - X_j)$ ,  $\widetilde{Y}_{ij} \sim \widehat{F}(y|X_i)$ , and  $\widehat{F}(y|X) = \sum_{j=1}^n \delta_j \times \mathcal{K}_h(X_i - X_j) I(Y_j \leq y) / \sum_{j=1}^n \delta_j \mathcal{K}_h(X_i - X_j)$  with  $\kappa = 20$ . Here  $\mathcal{K}_h(\cdot)$  is a two-dimensional product kernel, i.e.,  $\mathcal{K}_h(X_i - X_j) = h^{-2} K(h^{-1}(X_{1i} - X_{1j})) K(h^{-1}(X_{2i} - X_{2j}))$ , where  $K(\cdot)$  denotes the Gaussian kernel, and  $h$  is a bandwidth parameter.

S6. EL and ET estimators based on a naive nonparametric imputation:

$$\widehat{g}_i^N(\beta) = \delta_i g(X_i, Y_i, \beta) - (1 - \delta_i) \widetilde{m}_g^0(X_{2i}, \beta),$$

where

$$\tilde{m}_g^0(X_{2i}, \boldsymbol{\beta}) = \frac{\sum_{j=1}^n (1 - \delta_j) r_j K(h^{-1}(X_{2i} - X_{2j})) g(X_j, Y_j, \boldsymbol{\beta})}{\sum_{j=1}^n (1 - \delta_j) r_j K(h^{-1}(X_{2i} - X_{2j}))},$$

in which  $r_j$  was an indicator function taking 1 if unit  $j$  belongs to the follow-up sample and 0 otherwise (Kim and Yu, 2011). Here, the kernel function  $K(\cdot)$  was defined similarly to that in S5, and the used follow-up rate was 0.25.

In evaluating the above estimators S1 and S2, the proposed nonparametric AIPW procedure was implemented by using a one-dimensional Gaussian kernel function with the bandwidth  $h = 1.25\hat{\sigma}_{x_1} n^{-1/5}$ , where  $\hat{\sigma}_{x_1}$  is the sample standard deviation of  $X_1$ . Moreover, the estimator of  $\boldsymbol{\alpha}_0$  was computed by using the same SEL procedure proposed in the first experiment. Estimators S3–S4 were computed by using a two-dimensional product Gaussian kernel with the same bandwidth as in S1 and S2. The bandwidths for S5 and S6 were chosen in the same way as those in S1–S4. The initial values for  $\boldsymbol{\beta}_0$  and  $\boldsymbol{\alpha}_0$  were chosen in the same way as those in the Experiment 1.

Table ?? reports Bias, RMS and SD values of  $6 \times 2 = 12$  estimators for  $\boldsymbol{\beta}_0$ . Inspection of Table ?? leads to the following observations: (i) under the considered settings, the proposed AIPW-EL and AIPW-ET estimators (i.e., S1–S4) consistently perform reasonably well even in situations where the working respondent probability model was misspecified; (ii) when  $a = 1$ , the RMS and SD in S3 and S4 are smaller than those in S1 and S2, while estimators S1 and S2 have better performance in terms of RMS and SD when  $a = 0.5$  and 0.01; (iii) Tang and Qin’s estimator as expected leads to a considerable bias because Tang and Qin’s estimator heavily depends on the ignorable assumption of respondent probability model, which shows that Tang and Qin’s estimator is sensitive to the selection of the respondent probability model; (iv) the proposed AIPW-GEL estimators have smaller standard deviations than Tang and Qin’s estimator and the naive estimator under the considered settings. The above findings indicate that the proposed AIPW-GEL method can significantly improve the estimation efficiency over the naive estimation method because the AIPW-GEL method uses the respondent data to estimate  $m_g^0(X_{2i}, \boldsymbol{\beta}) = E\{g(X, Y, \boldsymbol{\beta}) \mid X_{2i}, \delta = 0\}$ , and the nonparametric kernel regression estimator of  $m_g^0(X_{2i}, \boldsymbol{\beta})$  was computed via the parametrically estimated propensity score, but the naive estimator only utilized the follow-up data to estimate  $m_g^0(X_{2i}, \boldsymbol{\beta})$ .

We further conducted a simple simulation study to evaluate the finite sample performance of the proposed approach for the case when the estimating equations might be of higher dimension. To do this, we independently simulated 500 data sets  $\{(X_i, Y_i) : i = 1, \dots, n\}$  with  $n = 100$  from a five-dimensional multiplicative regression model  $Y_i = \exp(X_i^\top \boldsymbol{\beta}_0) \varepsilon_i$ , where  $\boldsymbol{\beta}_0 = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5)^\top = (1, 0.5, 1, 1.5, 1)^\top$ ,  $X_i = (X_{1i}, X_{2i}, X_{3i}, X_{4i}, X_{5i})^\top \sim \mathcal{N}(0, \Sigma_x)$  with  $\Sigma_x = (\sigma_x^{kj})$  and  $\sigma_x^{kj} = 0.5^{|k-j|}$  for  $1 \leq k, j \leq 5$ , and  $\log(\varepsilon_i) \sim \mathcal{N}(0, 1)$ . To estimate  $\boldsymbol{\beta}_0$ , we considered the same moment functions as in the two-dimensional case. The

response indicator  $\delta_i$  for  $Y_i$  was generated from the following logistic regression model

$$\pi_i(\boldsymbol{\alpha}_0) = \frac{\exp(a + \alpha_1 X_{2i} + \alpha_2 X_{3i} + \alpha_3 X_{4i} + \alpha_4 X_{5i} + \alpha_5 Y_i)}{1 + \exp(a + \alpha_1 X_{2i} + \alpha_2 X_{3i} + \alpha_3 X_{4i} + \alpha_4 X_{5i} + \alpha_5 Y_i)},$$

where  $\boldsymbol{\alpha}_0 = (a, \alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5)^\top = (a, 0.05, 0.01, 0.025, 0.01, 0.01)^\top$  with  $a = 0.7$  and  $0.25$ . The estimator of  $\boldsymbol{\alpha}_0$  was computed by the SEL approach incorporating auxiliary information  $\phi(X_i, Y_i, \boldsymbol{\alpha}) = \delta_i \pi^{-1}(X_{2i}, Y_i, \boldsymbol{\alpha})(X_i - \bar{X})$ , where  $\bar{X} = n^{-1} \sum_{i=1}^n X_i$ . Assume that  $\Pr(\delta_i = 1 | U_i) = \{1 - \exp(-\vartheta_1 - \vartheta_2^\top U_i)\}^{-1} =: \mathcal{S}(U_i, \boldsymbol{\vartheta}_0)$ , where  $U_i = (X_{2i}, X_{3i}, X_{4i}, X_{5i})$ . We estimated  $\boldsymbol{\vartheta}_0$  by using the maximum likelihood estimation method and denoted the estimate  $\hat{\boldsymbol{\vartheta}} = (\hat{\vartheta}_1, \hat{\vartheta}_2^\top)^\top$ . Let  $\hat{\mathcal{S}}_i = \mathcal{S}(U_i, \hat{\boldsymbol{\vartheta}}) = \{1 - \exp(-\hat{\vartheta}_1 - \hat{\vartheta}_2^\top U_i)\}^{-1}$ . We then computed the kernel dimension reduction AIPW-GEL estimators for  $\boldsymbol{\beta}_0$  by using the one-dimensional kernel smoothers  $\mathcal{K}_h(\hat{\mathcal{S}} - \hat{\mathcal{S}}_i)$ . Table ?? reports the simulated Bias, RMS and SD. It can be seen that the proposed estimators have negligible biases and the values of RMS are quite close to those of SD.

## 6 A Real Example

We used a dataset on the serum-cholesterol levels of heart-attack patients (Schafer, 1997) to illustrate the proposed methodologies. In this dataset, the serum-cholesterol levels of 28 heart-attack patients at a Pennsylvania medical center were measured on 2, 4 and 14 days after the heart attack. Let  $X_{1i}$  be the cholesterol level of the  $i$ th patient measured on 2 days,  $X_{2i}$  be the cholesterol level of the  $i$ th patient measured on 4 days, and  $Y_i$  be the cholesterol level of the  $i$ th patient measured on 14 days for  $i = 1, \dots, 28$ . We observed that  $X_{1i}$  and  $X_{2i}$  were completely observed, whilst  $Y_i$ 's were subject to missingness. Let  $\delta_i$  be an indicator function taking 1 if  $Y_i$  is observed and 0 if  $Y_i$  is missing. The proportion of missing observations for  $Y_i$  was 32%. Let  $X_i = (X_{1i}, X_{2i})$ . Schafer (1997) had even analyzed the dataset via an EM algorithm under a trivariate normal distribution assumption of  $(X_i, Y_i)$  together with ignorable assumption of missing value for  $Y_i$ . Unlike Schafer (1997), we here assumed that the missing data mechanism is nonignorable.

The main purpose of our study was to investigate whether  $Y_i$  was related to  $X_i$ . To this end, we considered a quantile regression model. Specifically, we assumed the  $\tau$ -th conditional quantile of  $Y_i$  as follows  $Q_{Y_i}(\tau | X_i) = \mathbb{X}_i^\top \boldsymbol{\beta}_0$ ,  $i = 1, \dots, 28$ , where  $\mathbb{X}_i = (1, X_{1i}, X_{2i})^\top$  and  $\boldsymbol{\beta}_0 = (\beta_1(\tau), \beta_2(\tau), \beta_3(\tau))^\top$ . The moment functions for  $\boldsymbol{\beta}_0$  were defined as  $g(X_i, Y_i, \boldsymbol{\beta}_0) = \mathbb{X}_i \{I(Y_i \leq \mathbb{X}_i^\top \boldsymbol{\beta}_0) - \tau\}$  satisfying  $E\{g(X_i, Y_i, \boldsymbol{\beta}_0)\} = 0$ . We considered three different quantile levels:  $\tau = 0.25, 0.5$  and  $0.75$ . Moreover, we assumed that  $\Pr(\delta_i = 1 | X_i, Y_i) = \pi(U_i, Y_i, \boldsymbol{\alpha}_0)$ , where  $U_i = X_{1i}$  or  $X_{2i}$ , and considered the following two respondent probability models:

$$\begin{aligned} \text{Model 1: } \pi(U_i, Y_i, \boldsymbol{\alpha}_0) &= \frac{\exp(\alpha_1 + \alpha_2 U_i + \alpha_3 Y_i)}{1 + \exp(\alpha_1 + \alpha_2 U_i + \alpha_3 Y_i)}, \\ \text{Model 2: } \pi(U_i, Y_i, \boldsymbol{\alpha}_0) &= \Phi(\alpha_1 + \alpha_2 U_i + \alpha_3 Y_i), \end{aligned}$$

Table 3: Simulation results of the two-dimensional multiplicative regression model.

$\varepsilon_i$	Type	GEL	$\beta_0$	$a = 1$			$a = 0.5$			$a = 0.01$		
				Bias	RMS	SD	Bias	RMS	SD	Bias	RMS	SD
A	S1	EL	$\beta_1$	0.008	0.148	0.148	0.010	0.154	0.154	0.022	0.174	0.173
			$\beta_2$	0.002	0.151	0.151	0.009	0.155	0.155	-0.008	0.183	0.183
		ET	$\beta_1$	0.009	0.149	0.149	0.009	0.154	0.154	0.024	0.174	0.173
			$\beta_2$	0.003	0.151	0.151	0.010	0.155	0.155	-0.008	0.182	0.182
	S2	EL	$\beta_1$	0.009	0.149	0.149	0.011	0.154	0.154	0.021	0.174	0.173
			$\beta_2$	0.004	0.152	0.152	0.010	0.156	0.156	-0.006	0.183	0.183
		ET	$\beta_1$	0.009	0.150	0.150	0.011	0.153	0.153	0.021	0.174	0.173
			$\beta_2$	0.004	0.152	0.152	0.011	0.155	0.155	-0.006	0.183	0.183
	S3	EL	$\beta_1$	0.008	0.146	0.146	0.005	0.163	0.163	0.014	0.186	0.186
			$\beta_2$	0.001	0.138	0.138	0.008	0.160	0.160	0.013	0.192	0.191
		ET	$\beta_1$	0.009	0.146	0.146	0.005	0.161	0.161	0.013	0.186	0.185
			$\beta_2$	0.000	0.139	0.139	0.009	0.161	0.161	0.014	0.190	0.190
	S4	EL	$\beta_1$	0.008	0.145	0.145	0.005	0.164	0.164	0.014	0.188	0.187
			$\beta_2$	0.001	0.139	0.139	0.009	0.160	0.160	0.011	0.191	0.191
		ET	$\beta_1$	0.008	0.145	0.145	0.005	0.163	0.163	0.014	0.188	0.188
			$\beta_2$	0.000	0.138	0.138	0.007	0.161	0.161	0.011	0.191	0.191
	S5	EL	$\beta_1$	-0.039	0.352	0.350	-0.037	0.396	0.395	-0.095	0.409	0.398
			$\beta_2$	-0.074	0.361	0.354	-0.118	0.440	0.424	-0.192	0.453	0.411
		ET	$\beta_1$	-0.024	0.251	0.250	-0.014	0.278	0.278	-0.054	0.290	0.285
			$\beta_2$	-0.030	0.272	0.271	-0.068	0.291	0.283	-0.097	0.309	0.294
	S6	EL	$\beta_1$	0.024	0.188	0.187	0.013	0.201	0.201	0.026	0.246	0.245
			$\beta_2$	0.000	0.192	0.193	0.005	0.201	0.201	-0.018	0.238	0.238
		ET	$\beta_1$	0.024	0.188	0.187	0.014	0.200	0.199	0.025	0.244	0.243
			$\beta_2$	0.000	0.193	0.193	0.004	0.200	0.200	-0.017	0.238	0.238
B	S1	EL	$\beta_1$	-0.003	0.103	0.103	-0.003	0.116	0.116	0.006	0.123	0.123
			$\beta_2$	0.005	0.095	0.095	0.008	0.099	0.099	0.003	0.114	0.114
		ET	$\beta_1$	-0.003	0.102	0.102	-0.003	0.114	0.114	0.006	0.120	0.120
			$\beta_2$	0.004	0.092	0.092	0.008	0.098	0.098	0.002	0.113	0.113
	S2	EL	$\beta_1$	-0.003	0.102	0.102	-0.003	0.117	0.117	0.005	0.123	0.123
			$\beta_2$	0.005	0.095	0.095	0.008	0.100	0.099	0.004	0.115	0.115
		ET	$\beta_1$	-0.003	0.101	0.101	-0.004	0.116	0.116	0.006	0.119	0.119
			$\beta_2$	0.004	0.094	0.094	0.009	0.097	0.097	0.002	0.112	0.112
	S3	EL	$\beta_1$	0.001	0.097	0.097	-0.009	0.108	0.108	-0.007	0.115	0.115
			$\beta_2$	-0.003	0.091	0.091	0.006	0.105	0.105	0.005	0.103	0.103
		ET	$\beta_1$	0.000	0.096	0.096	-0.008	0.107	0.107	-0.007	0.112	0.112
			$\beta_2$	-0.002	0.092	0.092	0.006	0.105	0.105	0.005	0.103	0.103
	S4	EL	$\beta_1$	0.001	0.097	0.097	-0.008	0.107	0.107	-0.007	0.114	0.114
			$\beta_2$	-0.003	0.091	0.091	0.006	0.105	0.105	0.005	0.103	0.103
		ET	$\beta_1$	0.001	0.096	0.096	-0.009	0.107	0.107	-0.009	0.114	0.114
			$\beta_2$	-0.002	0.090	0.090	0.007	0.105	0.105	0.006	0.102	0.102
	S5	EL	$\beta_1$	-0.041	0.262	0.259	-0.058	0.279	0.274	-0.102	0.319	0.303
			$\beta_2$	-0.101	0.310	0.293	-0.141	0.340	0.310	-0.173	0.370	0.327
		ET	$\beta_1$	-0.018	0.215	0.215	-0.031	0.229	0.227	-0.043	0.251	0.247
			$\beta_2$	-0.009	0.232	0.232	-0.046	0.240	0.236	-0.043	0.270	0.267
	S6	EL	$\beta_1$	0.002	0.145	0.145	-0.010	0.158	0.158	0.001	0.179	0.180
			$\beta_2$	0.018	0.136	0.135	0.024	0.145	0.143	0.034	0.165	0.162
		ET	$\beta_1$	0.001	0.146	0.146	-0.011	0.157	0.157	0.003	0.177	0.177
			$\beta_2$	0.019	0.136	0.134	0.024	0.145	0.143	0.032	0.164	0.161

Table 4: Simulation results of the five-dimensional multiplicative regression model.

GEL	Statistic	$a = 0.7$					$a = 0.25$				
		$\beta_1$	$\beta_2$	$\beta_3$	$\beta_4$	$\beta_5$	$\beta_1$	$\beta_2$	$\beta_3$	$\beta_4$	$\beta_5$
EL	Bias	0.021	-0.004	0.006	0.005	0.017	0.008	0.003	0.013	0.019	0.022
	RMS	0.148	0.167	0.173	0.161	0.145	0.164	0.189	0.188	0.185	0.165
	SD	0.147	0.167	0.173	0.161	0.144	0.164	0.189	0.187	0.184	0.164
ET	Bias	0.020	-0.003	0.006	0.008	0.015	0.006	0.007	0.012	0.021	0.018
	RMS	0.148	0.168	0.171	0.165	0.144	0.167	0.192	0.189	0.187	0.165
	SD	0.147	0.168	0.171	0.165	0.144	0.167	0.193	0.189	0.186	0.165

where  $\Phi(\cdot)$  is the cumulative probability density function of the standard normal distribution.

To determine  $U_i$  or  $Z_i$ , similar to Shao and Wang (2016), we considered the following criterion

$$D = \left\| \frac{1}{n} \sum_{i=1}^n \frac{\delta_i X_i}{\pi(U_i, Y_i, \tilde{\alpha})} - \frac{1}{n} \sum_{i=1}^n X_i \right\|,$$

where  $\tilde{\alpha}$  is an estimator of  $\alpha_0$  computed by the SEL approach with a candidate  $U_i$ . Note that  $D$  converges to zero if and only if  $\pi(U_i, Y_i, \alpha_0)$  is correctly specified and  $\tilde{\alpha} = \alpha_0 + o_p(1)$  and  $Z_i$  is the instrument. Hence, the valid  $U_i$  or  $Z_i$  can be selected by minimizing  $D$  over a set of candidate variables. By calculation, we obtained  $U_i = X_{1i}$  and  $Z_i = X_{2i}$  for both respondent probability models considered above.

Similar to simulation studies, we used a one-dimensional Gaussian kernel function with the bandwidth  $h = 1.25\hat{\sigma}_{x_1} n^{-1/5}$  in evaluating nonparametric kernel estimation, where  $\hat{\sigma}_{x_1}$  is the sample standard deviation of  $\{X_{1i}, i = 1, \dots, n\}$ . For a given value of  $\tau$ , we calculated the proposed two-step AIPW-EL and AIPW-ET estimators of  $\beta_0$ . The proposed bootstrap calibration procedure with 200 bootstrap replications was adopted to estimate standard errors (SE) of the proposed estimators, and to calculate the 95% bootstrap-percentile-based confidence intervals. Also, we calculated the 95% EL-based and ET-based confidence intervals of parameters  $\beta_1(\tau)$ ,  $\beta_2(\tau)$  and  $\beta_3(\tau)$ , respectively.

Table ?? presents the point estimates (Est) of parameters  $\beta_1(\tau)$ ,  $\beta_2(\tau)$  and  $\beta_3(\tau)$ , and the corresponding SE and lengths of various 95% confidence intervals. From Table ??, we have the following observations. First, all the considered parameter estimates yield the same conclusion that the cholesterol levels measured on 2 days has a negative effect on the cholesterol level measured on 14 days, and the cholesterol levels measured on 4 days has a positive effect on the cholesterol level measured on 14 days, regardless of the quantile levels and the specified response probability models. Second, the EL-based and ET-based methods generally lead to narrower confidence intervals than the bootstrap method.

Table 5: Results of real data analysis.

Model	$\beta_0$	Statistic	EL			ET		
			$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$
1	$\beta_1(\tau)$	Est	0.584	0.654	0.776	0.768	0.636	0.984
		SE	0.515	0.481	0.449	0.515	0.490	0.455
		length <sup>1</sup>	2.017	1.887	1.760	2.018	1.921	1.784
		length <sup>2</sup>	0.084	0.078	0.087	0.065	0.077	0.093
	$\beta_2(\tau)$	Est	-0.311	-0.263	-0.249	-0.373	-0.204	-0.161
		SE	0.228	0.234	0.213	0.235	0.238	0.218
		length <sup>1</sup>	0.893	0.919	0.836	0.921	0.934	0.855
		length <sup>2</sup>	0.088	0.082	0.091	0.069	0.081	0.097
	$\beta_3(\tau)$	Est	0.965	0.943	0.962	0.949	0.876	0.820
		SE	0.255	0.236	0.213	0.260	0.247	0.219
		length <sup>1</sup>	1.001	0.925	0.835	1.020	0.969	0.857
		length <sup>2</sup>	0.184	0.178	0.187	0.165	0.177	0.193
2	$\beta_1(\tau)$	Est	0.461	0.465	0.687	0.392	0.768	0.856
		SE	0.555	0.478	0.451	0.519	0.470	0.445
		length <sup>1</sup>	2.176	1.875	1.767	2.036	1.842	1.744
		length <sup>2</sup>	0.085	0.070	0.098	0.083	0.061	0.072
	$\beta_2(\tau)$	Est	-0.210	-0.160	-0.105	-0.198	-0.180	-0.229
		SE	0.241	0.251	0.218	0.239	0.245	0.210
		length <sup>1</sup>	0.944	0.984	0.855	0.936	0.961	0.822
		length <sup>2</sup>	0.089	0.074	0.102	0.087	0.065	0.076
	$\beta_3(\tau)$	Est	0.872	0.897	0.877	0.878	0.813	0.906
		SE	0.260	0.233	0.229	0.257	0.245	0.220
		length <sup>1</sup>	1.020	0.912	0.899	1.009	0.961	0.862
		length <sup>2</sup>	0.185	0.170	0.198	0.183	0.161	0.172

NOTE: length<sup>1</sup> stands for the lengths of approximate 95% bootstrap-based confidence intervals with 200 bootstrap replications, length<sup>2</sup> represents the lengths of approximate 95% EL-based (or ET-based) confidence intervals.

## 7 Discussion

In an attempt to improve and refine existing methods for handling nonignorable missing data, we assume in this paper a parametric nonignorable propensity score model, and propose a propensity-score-based nonparametric imputation approach that makes use of an instrument to address the potential model identifiability problem in the presence of nonignorable missing data. Moreover, the GEL together with the AIPW approach is adopted to make statistical inference on parameters in nonsmooth moment functions, and the large sample results are established under some fairly mild conditions.

Correctly specifying propensity score model is critical to the proposed method. If the nonrespondent instrument is inappropriately selected, the propensity score model might be incorrectly specified such that conclusions may be potentially misleading. A valid nonrespondent instrument is known to satisfy the following conditions: (a) it has to be related to the outcome in the underlying population, conditional on a set of the fully observed covariates; (b) it is not directly related with the response mechanism, conditional on the fully observed covariates. If a nonrespondent instrument is manually selected as a subset or function of auxiliary variables, the above two conditions are hard to be verified in practical applications.

Although the criterion  $D$  proposed in Shao and Wang (2016) is sensitive to the choice of instrument, it may not help us find out which subset of instruments is the best if multiple instruments are available. Choosing among valid instruments is important when there are many thought to be equally valid. Similar to Donald et al. (2009), we could develop asymptotic mean square error (MSE) based criteria, related to the efficiency of the resultant estimators, for instrument selection in estimation of nonsmooth moment conditions with nonignorable missing data. An optimal instrument should simultaneously minimize the  $D$  and MSE criteria. Results are currently under investigation and are not discussed further here.

## Supplementary Material

Supplementary Materials available in the attached file contains detailed technical proofs of Propositions 1–2, Theorems 1–2 and the following Lemmas.

## Acknowledgement

The authors are grateful to the Editor, an associate editor, and two referees for their valuable suggestions and comments that greatly improved the manuscript. This research was supported by grants from the National Nature Science Foundation of China (Grant No.: 11671349, 11731011).

## Appendix

We first provide a few notation used in the rest of the paper. Define  $\mathcal{N}_\varrho = \{\beta \in \mathcal{B} : \|\beta - \beta_0\| < \varrho\}$  for some small  $\varrho > 0$ . Let  $\|\cdot\|$  denote the matrix norm given by  $\|\mathcal{H}\| = \sqrt{\text{trace}(\mathcal{H}^\top \mathcal{H})}$  for any  $q \times m$  matrix  $\mathcal{H}$  (including  $q = 1$  or  $q = m = 1$ ). For any  $q \times m$  matrix  $\mathcal{H}(u, \beta)$ ,  $\|\mathcal{H}(\beta)\|_\infty = \sup_{u \in \mathcal{U}} |\mathcal{H}(u, \beta)|$  for any  $\beta \in \mathcal{B}$ ,  $\mathcal{U}$  is the support of random vectors  $U$ . For ease of presentation, we consider the nonparametric estimation of  $m_g^0(U, \beta)$  using the same kernel function, that is, for each  $\nu = 1, \dots, r$ , we set  $K_{h^{(\nu)}}^{(\nu)}(\cdot) = K_h(\cdot) = K(\cdot/h)/h^{d_u}$ , in which  $K(\cdot)$  is a  $d_u$ -dimensional kernel function and  $h = h_n$  is a bandwidth sequence satisfying  $h_n \rightarrow 0$  as  $n \rightarrow \infty$ . Let  $\mathcal{G}_n(\beta, \alpha) = n^{-1} \sum_{i=1}^n \hat{g}_i(\beta, \alpha)$ ,  $\mathcal{G}(\beta) = E\{g(X, Y, \beta)\}$ , and let  $a^{\otimes 2} = aa^\top$  for any vector  $a$ . Throughout the appendix,  $\mathcal{C}$  represents a generic positive constant which may vary depending on the context.

**Assumption A:** The moment function  $g(X, Y, \beta)$  satisfies:

- (A1) (a)  $\beta_0 \in \mathcal{B}$  is the unique solution to  $\mathcal{G}(\beta) = 0$ , and  $\mathcal{B}$  is a compact subset of  $\mathcal{R}^p$ ;  
 (b)  $E\{\sup_{\beta \in \mathcal{B}} |g(X, Y, \beta)|\} < \infty$  and  $E\{\sup_{\beta \in \mathcal{N}_\varrho} |g(X, Y, \beta)|^2\} < \infty$ ;  
 (A2) the class of functions  $\{g(X, Y, \beta) : \beta \in \mathcal{B}\}$  is Glivenko-Cantelli;  
 (A3) for some  $\varrho > 0$ ,  $\{g(X, Y, \beta) : \beta \in \mathcal{N}_\varrho\}$  is Donsker;  
 (A4) for all  $\beta \in \mathcal{B}$  and all small positive value  $\varrho = o(1)$ ,

$$E\left\{ \sup_{\beta, \beta' \in \mathcal{N}_\varrho} \left| g_j(X, Y, \beta') - g_j(X, Y, \beta) \right|^2 \right\} \leq \mathcal{C}\varrho^{2s}$$

for some constants  $s \in (0, 1]$ , where  $g_j(\cdot)$  denotes the  $j$ th coordinates of  $g(\cdot)$  and  $j = 1, \dots, r$ .

**REMARK 3.** Assumption A has been used extensively in econometrics and statistics (see, e.g., Cattaneo, 2010; Chaudhuri and Guilkey, 2016). Consider the quantile regression model, where  $g(X, Y, \beta) = X[I(Y \leq X^\top \beta) - \tau]$  for  $\tau \in \mathcal{T} \subset [\tau_L, \tau_U]$  and  $0 < \tau_L < \tau_U < 1$ . Assume that there exists a constant  $K_x$  such that  $E|X|^3 \leq K_x$ . Let  $X_j$  denote the  $j$ th coordinates of  $X$ ,  $j = 1, \dots, r$ , let  $F_{Y|x}$  be the conditional distribution function at evaluation point  $X = x$ . Assumption (A1) is satisfied with some additional regularity conditions on quantile regression model. Note that the functional class  $\mathcal{F} = \{I(Y \leq X^\top \beta), \beta \in \mathcal{B}\}$  is a VC subgraph class and hence a bounded Donsker class. Hence  $\mathcal{F} - \mathcal{T}$  is also bounded Donsker and  $X(\mathcal{F} - \mathcal{T})$  is, therefore Donsker with a square-integrable envelope  $2 \max_{j \in 1, \dots, r} |X|_j$  (see Theorem 2.10.6 in Van der Vaart and Wellner, 1996). Assumptions (A2) and (A3) are then verified since the functional class  $\{X[I(Y \leq X^\top \beta) - \tau], \tau \in \mathcal{T}, \beta \in \mathcal{B}\}$  is formed as  $X(\mathcal{F} - \mathcal{T})$ . For  $j = 1, \dots, r$ ,  $|g_j(X, Y, \beta') - g_j(X, Y, \beta)|^2 \leq X_j^2 |I\{Y \leq X^\top \beta'\} - I\{Y \leq X^\top \beta\}|$ . For small enough  $\varrho > 0$ ,  $E[\sup_{|\beta - \beta'| \leq \varrho} X_j^2 |I\{Y \leq X^\top \beta'\} - I\{Y \leq X^\top \beta\}|] \leq$

$E[|X|^3\{F(X^\top\beta + \varrho | X) - F(X^\top\beta - \varrho | X)\}] \leq C\varrho$ , where the last inequality follows provided  $F_{Y|x}$  is Lipschitz in  $Y$  uniformly in  $x$ . This verifies Assumption (A4).

**Assumption B:** The conditional expectation  $m_g^0(U, \beta, \alpha_0)$  defined in (??) satisfies the following conditions:

(B1) the class of functions  $\{m_g^0(U, \beta, \alpha_0) : \beta \in \mathcal{B}\}$  is Glivenko-Cantelli;

(B2) for all  $U \in \mathcal{U}$ , and for some  $\varrho > 0$ ,

(a)  $m_g^0(U, \beta, \alpha_0)$  is continuously differentiable with derivative  $\partial_\beta m_g^0(U, \beta, \alpha_0) =: \partial m_g^0(U, \beta, \alpha_0)/\partial\beta^\top$  in  $\beta \in \mathcal{N}_\varrho$ ;

(b)  $E\{\sup_{\beta \in \mathcal{N}_\varrho} |\partial_\beta m_g^0(U, \beta, \alpha_0)|\} < \infty$ ;

(B3) there exist  $\epsilon \in (0, 1]$  and a measurable function  $b(U)$  with  $E\{|b(U)|\} < \infty$  such that

$$|\partial_\beta \tilde{m}_g^0(U, \beta, \alpha_0) - \partial_\beta m_g^0(U, \beta, \alpha_0)| \leq b(U) \sup_{\beta \in \mathcal{N}_\varrho} \|\tilde{m}_g^0(\beta, \alpha_0) - m_g^0(\beta, \alpha_0)\|_\infty^\epsilon$$

for all smooth functions  $\tilde{m}_g^0(U, \beta, \alpha_0) \in \mathcal{M}$  with  $\sup_{\beta \in \mathcal{N}_\varrho} \|\tilde{m}_g^0(\beta, \alpha_0) - m_g^0(\beta, \alpha_0)\|_\infty < \varrho$ .

**REMARK 4.** Assumption B restricts the class of functions

$$\mathcal{G} = \left\{ \tilde{m}_g^0(U, \beta, \alpha_0) : \tilde{m}_g^0(U, \beta, \alpha_0) \in \mathcal{M}, \beta \in \mathcal{N}_\varrho \text{ and } \|\tilde{m}_g^0(\beta, \alpha_0) - m_g^0(\beta, \alpha_0)\|_\infty < \varrho \right\},$$

where  $m_g^0(U, \beta, \alpha_0) \in \mathcal{G}$  by construction. It is easy to verify Assumption (B1) because it is nature to assume that the conditional expectations  $E\{\delta g(X, Y, \beta)O(U, Y, \alpha_0) | U\}$  are smooth in  $\beta$ ; Assumption (B2) is a usual dominance condition. Assumption (B3) is similar to Assumption 4 in Chen et al. (2005) and Assumption 7 in Cattaneo (2010). Assumption (B3) further restricts function class  $\{g(X, Y, \beta) : \beta \in \mathcal{N}_\varrho\}$  by requiring that functions are uniformly close and also their derivatives close. Assumptions A and B are necessary in order to establish the uniform convergence of the proposed estimators and derive the stochastic equicontinuity for guaranteeing the resulting estimators are still root- $n$  consistent and asymptotically normally distributed under nonsmooth moment conditions in the presence of nonignorable missing data.

**Assumption C:** Regularity conditions:

(C1) (i) The random vector  $X$  can be decomposed as  $X = (U, Z) \in \mathcal{U} \times \mathcal{Z} \subset \mathcal{R}^{d_u} \times \mathcal{R}^{d_z}$ , and  $\delta \perp\!\!\!\perp Z | (U, Y)$ , where  $U$  is continuously distributed with Lebesgue density  $f$ ; (ii) The probability density function  $f(u)$  is bounded away from  $\infty$  in the support of  $U$  and the second derivative of  $f(u)$  is continuous and bounded.

(C2) (a) For all  $\alpha$  in a neighborhood of  $\alpha_0$ , the propensity model  $\pi(U, Y, \alpha)$  is twice differentiable with respect to  $\alpha$  and  $E|\pi(U, Y, \alpha)|^3 < \infty$ ; (b)  $\pi(U) = E\{\pi(U, Y, \alpha_0) | U\} \neq 1$  a.s.; (c) uniformly for all  $\alpha \in \mathcal{A}$ ,  $\pi(U_i, Y_i, \alpha) \geq \mathcal{C} > 0$  for all  $i = 1, \dots, n$ , uniformly in  $n$ .

(C3) The kernel function  $K(\cdot)$  of the  $q$ -th order satisfies the following conditions

- (i)  $K(\cdot)$  is bounded and has compact support;
- (ii)  $\int K(u_1, \dots, u_{d_u}) du_1 \dots du_{d_u} = 1$ ;
- (iii)  $\int u_s^l K(u_1, \dots, u_{d_u}) du_1 \dots du_{d_u} = 0$  and  $\int u_s^q K(u_1, \dots, u_{d_u}) du_1 \dots du_{d_u} \neq 0$  for any  $s = 1, \dots, d_u$  and  $1 \leq l < q$ .

(C4) The data-dependent bandwidth  $h$  satisfies  $nh^{d_u}/\log n \rightarrow \infty$  and  $nh^{2q} \rightarrow 0$ .

**REMARK 5.** Assumption (C1) is conditional independence assumptions, which are used to achieve identification with nonignorable missing data (e.g., Wang et al., 2014 and Zhao and Shao, 2015). Assumptions (C2)-(C4) are commonly used in the missing data analysis and nonparametric regression inference.

To prove Theorems ??-??, we need the following Lemmas, whose proofs can be found in the Supplementary Material.

**LEMMA 1.** Suppose that Assumption C holds. Then, we have

$$\sup_{\beta \in \mathcal{B}, \alpha \in \mathcal{A}} \|\widehat{m}_g^0(\beta, \alpha) - m_g^0(\beta, \alpha)\|_\infty = o_p(n^{-1/4}).$$

**LEMMA 2.** Suppose that Assumption C holds; the respondent probability model  $\pi(U, Y, \alpha_0)$  is correctly specified and  $\widehat{\alpha}$  is computed by the SEL approach. Then, we have

$$\mathcal{G}_n(\beta_0, \widehat{\alpha}) = \frac{1}{n} \sum_{i=1}^n \widetilde{g}_i(\beta, \alpha_0) - \Xi \times (\widehat{\alpha} - \alpha_0) + o_p(n^{-1/2}),$$

where  $\widetilde{g}_i(\beta, \alpha)$  is defined in (6),  $\Xi = \text{Cov}\{\widetilde{g}_i(\beta_0, \alpha_0), \Delta(U, Y, \alpha_0)\}$  with  $\Delta(U, Y, \alpha) = \{\delta - \pi(U, Y, \alpha)\} \partial \text{logit}\{\pi(U, Y, \alpha)\} / \partial \alpha^\top$ .

**LEMMA 3.** Suppose that Assumption C holds; the respondent probability model  $\pi(U, Y, \alpha_0)$  is correctly specified and  $\widehat{\alpha}$  is computed by the SEL approach. Then, we have

$$\frac{1}{n} \sum_{i=1}^n \widehat{g}_i(\beta_0, \widehat{\alpha}) \widehat{g}_i(\beta_0, \widehat{\alpha})^\top = V_1 + o_p(1),$$

where  $V_1 = E\{\widetilde{g}_i(\beta_0, \alpha_0) \widetilde{g}_i(\beta_0, \alpha_0)^\top\}$ .

**LEMMA 4.** Suppose that Assumptions A, B and C hold; the respondent probability model  $\pi(U, Y, \alpha_0)$  is correctly specified and  $\widehat{\alpha}$  is computed by the SEL approach. Then, for all positive  $\varrho_n = o_p(1)$ , we have

$$\sup_{|\beta - \beta_0| \leq \varrho_n} \frac{|\mathcal{G}_n(\beta, \widehat{\alpha}) - \mathcal{G}(\beta) - \mathcal{G}_n(\beta_0, \widehat{\alpha}) + \mathcal{G}(\beta_0)|}{1 + \mathcal{C}n^{1/2}|\beta - \beta_0|} = o_p(n^{-1/2}).$$

**LEMMA 5.** Suppose that Assumptions (A1) and C hold; the respondent probability model  $\pi(U, Y, \alpha_0)$  is correctly specified and  $\hat{\alpha}$  is computed by the SEL approach. Then, for  $\Lambda_n = \{\lambda : |\lambda| \leq Cn^{-1/2}\}$ , we obtain  $\sup_{\beta \in \mathcal{B}, \lambda \in \Lambda_n, 1 \leq i \leq n} |\lambda^\top \hat{g}_i(\beta, \hat{\alpha})| \xrightarrow{P} 0$  and w.p.1,  $\Lambda_n \subseteq \hat{\Lambda}_n(\beta, \alpha)$  for all  $\beta \in \mathcal{B}$  and  $\alpha \in \mathcal{A}$ .

**LEMMA 6.** Suppose that Assumptions (A1) and C hold; the respondent probability model  $\pi(U, Y, \alpha_0)$  is correctly specified and  $\hat{\alpha}$  is computed by SEL approach. Then, we have  $|\mathcal{G}_n(\hat{\beta}_S, \hat{\alpha})| = O_p(n^{-1/2})$ .

*Proof of Theorem ??.* The proof for Theorem ?? essentially involves establishing bootstrap version of Lemma ?? and Theorem ??. We only establish the bootstrap version of Lemma ?? here. Let  $X_i^*, Y_i^*$  and  $\delta_i^*$  be the counterparts of  $X_i, Y_i$ , and  $\delta_i$  in the bootstrap sample, respectively. Let  $\hat{\eta}^* = (\hat{\alpha}^{*\top}, \hat{\omega}^*, \hat{\gamma}^{*\top})^\top$  be the bootstrap estimator of  $\eta_0 = (\alpha_0^\top, \omega_0, \gamma_0^\top)^\top$ . We use  $E_*$  to represent the conditional expectation given the original data. Define  $\pi_i^*(\alpha) = \pi(X_i^*, Y_i^*, \alpha)$ ,  $\hat{m}_g^*(U, \beta, \alpha) = \sum_{i=1}^m \mathcal{W}_{i0}^*(U, \alpha) g(X_i^*, Y_i^*, \beta)$ , and

$$\mathcal{G}_m^*(\beta, \alpha) = \frac{1}{m} \sum_{i=1}^m \left\{ \frac{\delta_i^*}{\pi_i^*(\alpha)} g(X_i^*, Y_i^*, \beta) - \frac{\delta_i^* - \pi_i^*(\alpha)}{\pi_i^*(\alpha)} \hat{m}_g^*(U_i^*, \beta, \alpha) \right\},$$

in which  $\mathcal{W}_{i0}^*(U, \alpha) = \delta_i^* O_i^*(\alpha) \mathcal{K}_h(U - U_i^*) / \{\sum_{k=1}^n \delta_k^* O_k^*(\alpha) \mathcal{K}_h(U - U_k^*)\}$  with  $O_i^*(\alpha) = O(U_i^*, Y_i^*, \alpha)$ . Then there exists  $\alpha^\dagger$  between  $\hat{\alpha}^*$  and  $\hat{\alpha}$  such that  $\mathcal{G}_m^*(\hat{\beta}_S, \hat{\alpha}^*) - \mathcal{G}_n(\hat{\beta}_S, \hat{\alpha}) = \mathcal{G}_m^*(\hat{\beta}_S, \hat{\alpha}) - \mathcal{G}_n(\hat{\beta}_S, \hat{\alpha}) + \partial \mathcal{G}_m^*(\hat{\beta}_S, \alpha^\dagger) / \partial \alpha^\top (\hat{\alpha}^* - \hat{\alpha})$ . Similar to the proof of Proposition ?? but replace the functions and the parameters with their corresponding bootstrap analogs, we can show  $n^{1/2}(\hat{\eta}^* - \hat{\eta}) \xrightarrow{\mathcal{L}^*} \mathcal{N}(0, \mathbb{A}^{-1} \mathbb{B} (\mathbb{A}^{-1})^\top)$ , where  $\mu_n^* \xrightarrow{\mathcal{L}^*} \mu$  means  $\Pr_*(\mu_n^* \in B) - \Pr(\mu \in B) \xrightarrow{P} 0$  for any Borel set  $B$ , and  $\Pr_*$  denotes a probability under the bootstrap distribution conditional on the original data set. We establish the bootstrap version of Lemma 2 by the following two steps.

**Step 1.** Show that  $\partial \mathcal{G}_m^*(\hat{\beta}_S, \alpha^\dagger) / \partial \alpha^\top \xrightarrow{P} \Xi$ . By Assumption C and similar arguments to those used in the proof of Lemma 2, as  $n \rightarrow \infty$  and  $m \rightarrow \infty$ , we can obtain the result.

**Step 2.** Show that  $n^{1/2} \{\mathcal{G}_m^*(\hat{\beta}_S, \hat{\alpha}) - \mathcal{G}_n(\hat{\beta}_S, \hat{\alpha})\} \xrightarrow{\mathcal{L}^*} \mathcal{N}(0, V_1)$ . To prove this result, we only need to prove that  $n^{1/2} \{\mathcal{G}_m^*(\hat{\beta}_S, \alpha_0) - \mathcal{G}_n(\hat{\beta}_S, \alpha_0)\} \xrightarrow{\mathcal{L}^*} \mathcal{N}(0, V_1)$  because  $\partial \mathcal{G}_m^*(\hat{\beta}_S, \alpha^\dagger) / \partial \alpha^\top = \partial \mathcal{G}_n(\hat{\beta}_S, \alpha^\dagger) / \partial \alpha^\top + o_p(1)$  as  $n \rightarrow \infty$  and  $m \rightarrow \infty$  and  $\hat{\alpha} - \alpha_0 = O_p(n^{-1/2})$ . Here  $\alpha^\dagger$  lies between  $\hat{\alpha}$  and  $\alpha_0$ . Next we note that  $\mathcal{G}_m^*(\hat{\beta}_S, \alpha_0) - \mathcal{G}_n(\hat{\beta}_S, \alpha_0) = K_{mn1} + K_{mn2} + K_{mn3}$ , where

$$\begin{aligned} K_{mn1} &= \frac{1}{m} \sum_{i=1}^m \left[ \frac{\delta_i^*}{\pi_i^*(\alpha_0)} g(X_i^*, Y_i^*, \hat{\beta}_S) - \frac{\delta_i - \pi_i(\alpha_0)}{\pi_i^*(\alpha_0)} m_g^0(U_i^*, \hat{\beta}_S, \alpha_0) \right. \\ &\quad \left. - \frac{1}{n} \sum_{i=1}^n \left\{ \frac{\delta_i}{\pi_i(\alpha_0)} g(X_i, Y_i, \hat{\beta}_S) - \frac{\delta_i - \pi_i(\alpha_0)}{\pi_i(\alpha_0)} m_g^0(U_i, \hat{\beta}_S, \alpha_0) \right\} \right], \\ K_{mn2} &= \frac{1}{m} \sum_{i=1}^m \left[ \left\{ 1 - \frac{\delta_i^*}{\pi_i^*(\alpha_0)} \right\} \{ \hat{m}_g^0(U_i^*, \hat{\beta}_S, \alpha_0) - m_g^0(U_i^*, \hat{\beta}_S, \alpha_0) \} \right. \\ &\quad \left. - \frac{1}{n} \sum_{i=1}^n \left\{ 1 - \frac{\delta_i}{\pi_i(\alpha_0)} \right\} \{ \hat{m}_g^0(U_i, \hat{\beta}_S, \alpha_0) - m_g^0(U_i, \hat{\beta}_S, \alpha_0) \} \right], \\ K_{mn3} &= \frac{1}{m} \sum_{i=1}^m \left\{ 1 - \frac{\delta_i^*}{\pi_i^*(\alpha_0)} \right\} \{ \hat{m}_g^*(U_i^*, \hat{\beta}_S, \alpha_0) - \hat{m}_g^0(U_i^*, \hat{\beta}_S, \alpha_0) \}. \end{aligned}$$

For  $K_{mn1}$ , we can apply the central limit theorem for bootstrap samples (Shao and Tu, 1994) to derive  $n^{1/2}K_{mn1} \xrightarrow{L} \mathcal{N}[0, E_*\{\tilde{g}_i(\hat{\beta}_S, \alpha_0)\tilde{g}_i(\hat{\beta}_S, \alpha_0)^\top\}]$ . Use similar argument to  $I_{n2}$  in Lemma 2 to show  $K_{mn2} = o_p(n^{-1/2})$ . Also it can be shown  $K_{mn3} = o_p(n^{-1/2})$ . Then the desired result is obtained by noting  $E_*\{\tilde{g}_i(\hat{\beta}_S, \alpha_0)\tilde{g}_i(\hat{\beta}_S, \alpha_0)^\top\} \rightarrow V_1$  as  $n \rightarrow \infty$  and  $m \rightarrow \infty$ . The bootstrap version of Lemma 2 could be established by combining above arguments.  $\square$

## References

- Andrews, D. W. (1995). Nonparametric Kernel Estimation for Semiparametric Models. *Econometric Theory* **11**, 560–596.
- Cattaneo, M. (2010). Efficient semiparametric estimation of multi-valued treatment effects under ignorability. *Journal of Econometrics* **155**, 138–154.
- Chaudhuri, S. and Guilkey, D. K. (2016). GMM with multiple missing variables. *Journal of Applied Econometrics* **31**, 678–706
- Chen, J., Sitter, R. R. and Wu, C. (2002). Using empirical likelihood method to obtain range restricted weights in regression estimators for surveys. *Biometrika* **89**, 230–237.
- Chen, X., Hong, H. and Tarozzi, A. (2008). Semiparametric efficiency in GMM models with auxiliary data. *Annals of Statistics* **36**, 808–843.
- Chen, X., Hong, H. and Tamer, E. (2005). Measurement error models with auxiliary data. *The Review of Economic Studies* **72**, 343–366.
- Chen, K., Guo, S., Lin, Y and Ying, Z. (2010). Least absolute relative error estimation. *Journal of the American Statistical Association* **105**, 1104–1112.
- Cheng, P. E. (1994). Nonparametric estimation of mean functionals with data missing at random. *Journal of the American Statistical Association* **89**, 81–87.
- Chen, X. R, Wan, A. T. K., and Zhou, Y. (2015). Efficient quantile regression analysis with missing observations. *Journal of the American Statistical Association* **110**, 723–741.
- Donald, S. G., Imbens, G. W. and Newey, W. K. (2009). Choosing instrumental variables in conditional moment restriction models. *Journal of Econometrics* **152**, 28–36.
- Goffe, W. L., Ferrier, G. D. and Rogers, J. (1994). Global optimization of statistical functions with simulated annealing. *Journal of Econometrics* **60**, 65–99.
- Hansen, B. E. (2015). *Econometrics (online draft of graduate textbook)*. University of Wisconsin.

- Hansen, L. (1982). Large sample properties of generalized method of moments estimators. *Econometrica* **50**, 1029–1054.
- Imbens, G. W., Spady, R. H. and Johnson, P. (1998). Information theoretic approaches to inference in moment condition models. *Econometrica* **66**, 333–357.
- Ibrahim, J. G., Lipsitz, S. R. and Chen, M. (1999). Missing covariates in generalized linear models when the missing data mechanism is non-ignorable. *Journal of the Royal Statistical Society: Series B* **61**, 173–190.
- Kim, J. K. and Yu, C. L. (2011). A semiparametric estimation of mean functionals with nonignorable missing data. *Journal of the American Statistical Association* **106**, 157–165.
- Kim, J. K. and Shao, J. (2013). *Statistical Methods for Handling Incomplete Data*. Chapman & Hall / CRC.
- Kitamura, Y. and Stutzer, M. (1997). An information-theoretic alternative to generalized method of moments estimation. *Econometrica* **65**, 861–874.
- Kirkpatrick, S., Gelatt, C. D. and Vecchi, M. P. (1983). Optimization by simulated annealing. *Science* **220**, 671–680.
- Koenker, R. (2005). *Quantile Regression*. Cambridge: Cambridge University Press.
- Li, Z., Lin, Y., Zhou, G. and Zhou, W. (2014). Empirical likelihood for least absolute relative error regression. *Test* **23**, 86–99.
- Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis With Missing Data (2nd ed.)*, New York: Wiley.
- Miao, W. and Tchetgen Tchetgen, E. J. (2016). On varieties of doubly robust estimators under missingness not at random with a shadow variable. *Biometrika* **103**, 475–482.
- Molanes-Lopez, E. M., Van Keilegom, I. and Veraverbeke, N. (2009). Empirical likelihood for non-smooth criterion functions. *Scandinavian Journal of Statistics* **36**, 413–432.
- Newey, W. and Smith, R. J. (2004). Higher-order properties of GMM and generalized empirical likelihood estimators. *Econometrica* **72**, 219–255.
- Nelsen, R. B. (1999). *An introduction to copulas*. Springer, New York.
- Owen, A. (1990) Empirical likelihood ratio confidence regions. *Annals of Statistics* **18**, 90–120.
- Parente, P. M. and Smith, R. J. (2011). GEL methods for nonsmooth moment indicators. *Econometric Theory* **27**, 74–113.

- Pakes, A. and Pollard, D. (1989). Simulation and the asymptotics of optimization estimators. *Econometrica* **57**, 1027-1057.
- Pepe, M. S. (2003). *The statistical evaluation of medical tests for classification and prediction*. Oxford University Press, New York.
- Qin, J., Leung, D. and Shao, J. (2002). Estimation with survey data under nonignorable nonresponse or informative sampling. *Journal of the American Statistical Association* **97**, 193–200.
- Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics* **22**, 300–325.
- Rotnitzky, A., Robins, J. M. and Scharfstein, D. (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association* **93**, 1321–1339
- Robins, M., Rotnitzky, A. and Zhao, L. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of American Statistical Association* **89**, 846–866.
- Robins, J. M. and Ritov, Y. (1997). Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Statistics in Medicine* **16**, 285–319.
- Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley.
- Scharfstein, D. O., Rotnitzky, A. and Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association* **94**, 1096–1120.
- Shao, J. and Tu, D. (1995). *The Jackknife and Bootstrap*. Springer Verlag, New York.
- Shao, J. and Wang, L. (2016). Semiparametric inverse propensity weighting for nonignorable missing data. *Biometrika* **103**, 175–187.
- Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman and Hall, London.
- Tstatis, A. A. (2006). *Semiparametric Theory and Missing Data*. New York: Springer.
- Tang, N. S., Zhao, P. Y. and Zhu, H. T. (2014). Empirical likelihood for estimating equations with nonignorably missing data. *Statistica Sinica* **24**, 723–747.
- Tang, C. Y. and Qin, Y. (2012). An efficient empirical likelihood approach for estimating equations with missing data. *Biometrika* **99**, 1001–1007.

- Van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer-Verlag, New York.
- Vansteelandt, S., Rotnitzky, A. and Robins, J. (2007). Estimation of regression models for the mean of repeated outcomes under nonignorable nonmonotone nonresponse. *Biometrika* **94**, 841–860.
- Wang, S., Shao, J. and Kim, J. K. (2014). An instrument variable approach for identification and estimation with nonignorable nonresponse. *Statistica Sinica* **24**, 1097–1116.
- Zhao, J. and Shao, J. (2015). Semiparametric pseudo likelihoods in generalized linear models with nonignorable missing data. *Journal of the American Statistical Association* **110**, 1577–1590.
- Zhao, P. Y., Tang, N. S., Qu, A. and Jiang, D. P. (2017a). Semiparametric estimating equations inference with nonignorable missing data. *Statistica Sinica* **27**, 89–113.
- Zhao, P. Y., Zhao, H., Tang, N. S. and Li, Z. H. (2017b). Weighted composite quantile regression analysis for nonignorable missing data using nonresponse instrument. *Journal of Nonparametric Statistics* **29**, 189–212.

Key Lab of Statistical Modeling and Data Analysis of Yunnan Province,  
Yunnan University, Kunming 650091, China.

Department of Statistics and Actuarial Science, University of Waterloo, Canada.

E-mail: pyzhao@live.cn

Key Lab of Statistical Modeling and Data Analysis of Yunnan Province,  
Yunnan University, Kunming 650091, China.

E-mail: nstang@ynu.edu.cn

Department of Biostatistics, The University of Texas MD Anderson Cancer Center,  
Houston, TX 77030, U.S.A.

E-mail: hzhu@bios.unc.edu