Statistica Sinica **30** (2020), 217-249
doi:https://doi.org/10.5705/ss.202017.0042

GENERALIZED EMPIRICAL LIKELIHOOD INFERENCES FOR NONSMOOTH MOMENT FUNCTIONS WITH NONIGNORABLE MISSING VALUES

Puying Zhao[1,2], Niansheng Tang[1] and Hongtu Zhu[3]

[1] *Yunnan University,* [2] *University of Waterloo and*
[3] *University of North Carolina at Chapel Hill*

Abstract: The main purpose of this study is to develop parameter identifiability and statistical inferences 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 uses an instrumental variable to address model identifiability in the presence of nonignorable missing data. A set of augmented inverse probability weighting moment functions is constructed as a basis for inferences performed using the generalized empirical likelihood method. Under some mild regularity conditions, we establish the large-sample properties of the resultant two-step generalized empirical likelihood estimators and generalized empirical likelihood ratio statistics for the case in which the propensity score is estimated parametrically using a correctly specified model. A derivative-free optimization method based on the simulated annealing algorithm is developed to implement the proposed methods. The methods are illustrated using simulations and an application to a data set on the serum-cholesterol levels of heart-attack patients.

Key words and phrases: Generalized empirical likelihood, identification, instrumental variable, nonignorable missing data, nonsmooth moment conditions, simulated annealing algorithm.

1. Introduction

Missing data are often a problem in clinical trials and survey studies. It is well known that performing complete case analyses using complete observations only results in a loss of information, leading to bias if the data are not missing completely at random. Consequently, various valid inferential alternatives to the complete-case approach have been developed for handling missing values, including the likelihood-based approach (Ibrahim, Lipsitz and Chen (1999); Little and Rubin (2002)), imputation approach (Rubin (1987); Cheng (1994)), and

augmented inverse probability weighting (AIPW) approach (Robins, Rotnitzky and Zhao (1994)). These methods have been applied successfully to analyze ignorable missing data (see Tstatis (2006) for a more detailed discussion). For a complete review of statistical analyses with missing data, see Little and Rubin (2002) and Kim and Shao (2013).

Most existing methods assume that an ignorable missing-data mechanism will fail to recover information from incomplete observed cases, or will fail to correct the bias when the mechanism is nonignorable. Handling nonignorable missing data is difficult, owing to several key challenges. Such challenges include developing a sensible model to characterize the nonignorable missing-data mechanism and its associated model identifiability and estimation (Robins and Ritov (1997); Kim and Yu (2011); Wang, Shao and Kim (2014); Tang, Zhao and Zhu (2014); Zhao and Shao (2015); Shao and Wang (2016); Miao and Tchetgen (2016); Zhao et al. (2017a)). The AIPW method is a popular semiparametric method used to handle missing data, including nonignorable missing data; see, for example, Scharfstein, Rotnitzky and Robins (1999); Rotnitzky, Robins and Scharfstein (1998), and Vansteelandt, Rotnitzky and Robins (2007), among others. However, existing AIPW methods focus only on smooth moment conditions and depend on identifiable nonignorable propensity score models with restrictions that are difficult to verify. Thus, the body of research on AIPW methods for nonignorable missing data is far from complete.

Statistical and econometric models defined using nonsmooth moment functions include the least absolute deviations, quantile regression models, and quantile treatment effects as special cases. Considerable effort has been devoted to estimating finite-dimensional parameters defined using nonsmooth moment functions in the presence of missing data; see, for example, Chen, Hong and Tarozzi (2008), Cattaneo (2010), Chen, Wan and Zhou (2015), and Chaudhuri and Guilkey (2016), as well as the references therein. However, most existing estimation procedures were developed to handle ignorable missing data based on the generalized method of moments (GMM, Hansen (1982)). Such GMM-based approaches lack the ability to generate likelihood ratio-based confidence regions with a shape that adapts to the support of the data. Moreover, the performance loss for the GMM can be substantial in the case of small samples. Thus, 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) methods, known as nonparametric maximum likelihood methods, have been shown to

be useful alternatives to the 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. In addition to improving their small-sample properties, this enables GMM estimators to compete with the bootstrap method (see, e.g., Owen (1990); Qin and Lawless (1994); Kitamura and Stutzer (1997); Imbens, Spady and Johnson (1998)). Here, we formulate 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, Van Keilegom and Veraverbeke (2009) and Parente and Smith (2011) developed EL and GEL methods, respectively, to make statistical inferences on nonsmooth moment functions.

This study makes three contributions to the literature. First, we suggest a more attractive imputation procedure that mitigates the effects of missing data and identifies the parameters in a 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." The parametric identification is based on the independence between a subset of the observed auxiliary variables, called nonrespondent instrumental variables (Wang, Shao and Kim (2014)), and the missing indicator, conditional on the missing variables and other observed auxiliary variables. A set of unbiased augmented inverse probability weighted moment functions (AIPW-MF) is constructed based on the proposed imputation approach. The use of a 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 estimated GEL ratio (GLR) statistics and develop a class of two-step AIPW-based GEL (AIPW-GEL) estimates for the parameters of interest. We systematically investigate the asymptotic properties of our proposed two-step AIPW-GEL estimators and GLR statistics for cases in which the propensity score is estimated parametrically under a correctly specified parametric model. The large-sample theories are established using the results of modern empirical process theories, including the uniform law of large numbers, stochastic equicontinuity, and Donsker class. The GEL confidence intervals for the parameters of interest are constructed using a 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, Gelatt and Vecchi (1983); Goffe, Ferrier and Rogers (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. The inner loop solves the optimization problem of Lagrange multipliers, which can be done using Newton-type methods. The outer loop implements the classical SA approach to minimize the concentrated GEL function in order to solve the optimization problem of unknown parameters defined using nonsmooth moment functions. The proposed algorithm is a sophisticated random search, which empirical studies have shown to be successful at locating global minima.

The rest of this paper is organized as follows. In Section 2, we discuss the identification of nonsmooth moment functions using a parametric nonignorable propensity score model and the semiparametric empirical likelihood estimation of the propensity. Here, we also 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 the modified SA algorithm. Our simulation studies are presented in Section 5, and a data example is discussed in Section 6. Section 7 concludes the paper. All technical details are presented in the Appendix and in the online Supplementary 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 $F(x, y) \in \mathscr{F}$, a class of distributions on a sample space. Let $\boldsymbol{\beta}$ be a vector of parameters of interest belonging to a compact subset $\mathcal{B}$ of $\mathcal{R}^p$, and let $g(X, Y, \boldsymbol{\beta})$ be a known vector-valued function with dimension $r \geq p$. On $\mathscr{F}$, there is some $\boldsymbol{\beta}_0 \in \mathcal{B} \subset \mathcal{R}^p$ such that

$$E\{g(X, Y, \boldsymbol{\beta}_0)\} = 0 \quad \text{w.p.1}, \tag{2.1}$$

where $E\{\cdot\}$ represents the expectation taken with respect to F, the notation "w.p.1" refers to "with probability one." Throughout this paper, we assume that the moment functions g are a class of nonsmooth functions with respect to $\boldsymbol{\beta}$. Thus, many parametric models, such as the quantile regression model (Koenker (2005)), copulas (Nelsen (1999)), receiver operating characteristic curves (Pepe

(2003)), and quantile treatment effects (Cattaneo (2010)), are special cases of model (2.1).

We consider the observed sample $\{(X_i, Y_i, \delta_i) : i = 1, \ldots, n\}$, which is an independently and identically distributed (i.i.d.) sample from (X, Y, δ). Here, δ is a dichotomous variable indicating whether or not Y is missing, and Y_i is observed if and only if $\delta_i = 1$, whereas X_i is always available for $i = 1, \ldots, n$. For simplicity, we suppose that the missing components have the same components across different individuals. Moverover, the missing Y may represent a 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 $\pi_i = \pi(X_i, Y_i)$ is a function that depends both on observed and missing variables. The main interest of this study is in making statistical inferences on the parameters $\boldsymbol{\beta}_0$ defined in (2.1) 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 the 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 the indicator δ 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:

$$\Pr(\delta = 1 \mid X, Y) = \Pr(\delta = 1 \mid U, Y) =: \pi(U, Y, \boldsymbol{\alpha}_0), \tag{2.2}$$

where π is a known smooth function with an l-dimensional unknown parameter $\boldsymbol{\alpha}_0 \in \mathcal{A}$, a compact subset of $\mathcal{R}^l$. It follows from (2.2) 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 a nonrespondent instrument (Wang, Shao and Kim (2014)), which means that it helps to identify parameters in the considered respondent probability model, but is not directly related to the response probability. In practice, the exclusion requirement (2.2) 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 (2.2), Wang, Shao and Kim (2014) proposed using the GMM approach to construct a root-n consistent estimator for the unknown $\boldsymbol{\alpha}_0$. In practical applications, we can specify (2.2) as cumulative distribution functions or logistic regression models.

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

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

for any measurable set B.

Under assumption (2.2), the conditional odds of missing data is $\Pr(\delta = 0 \mid X, Y)/\Pr(\delta = 1 \mid X, Y) = \Pr(\delta = 0 \mid U, Y)/\Pr(\delta = 1 \mid U, Y) = \pi^{-1}(U, Y, \boldsymbol{\alpha}_0) - 1 =: O(U, Y, \boldsymbol{\alpha}_0)$. Consequently, we obtain

$$f_0(Z, Y \mid U) = f_1(Z, Y \mid U) \times \frac{O(U, Y, \boldsymbol{\alpha}_0)}{E\{O(U, Y, \boldsymbol{\alpha}_0) \mid U, \delta = 1\}}, \tag{2.3}$$

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

$$f(Z, Y \mid U) = f_1(Z, Y \mid U)\Pr(\delta = 1 \mid U) + f_0(Z, Y \mid U)\Pr(\delta = 0 \mid U). \tag{2.4}$$

This, together with (2.3), implies that the unknown quantities in the joint density $f(X, Y)$ and conditional density $f(Z, Y \mid U)$ can be estimated based on the observed data distribution. For example, using identities (2.3) and (2.4), $E\{g(X, Y, \boldsymbol{\beta})\}$ is equal to the following functional of the observed data distribution:

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

where $m_g^\kappa(U, \boldsymbol{\beta}) = E\{g(X, Y, \boldsymbol{\beta}) \mid U, \delta = \kappa\}$, $\kappa = 0, 1$. If the response mechanism is ignorable, then we have $f_0(Z, Y \mid U) = f_1(Z, Y \mid U) = f(Z, Y \mid U)$ and $m_g^0(U, \boldsymbol{\beta}) = m_g^1(U, \boldsymbol{\beta}) = E\{g(X, Y, \boldsymbol{\beta}) \mid U\}$. The identity given in (2.3) is key to our methodology.

2.3. Augmented moment functions

From (2.3), it follows that we have

$$m_g^0(U, \boldsymbol{\beta}) = \frac{E\{\delta g(X, Y, \boldsymbol{\beta})O(U, Y, \boldsymbol{\alpha}_0) \mid U\}}{E\{\delta O(U, Y, \boldsymbol{\alpha}_0) \mid U\}} =: m_g^0(U, \boldsymbol{\beta}, \boldsymbol{\alpha}_0). \tag{2.5}$$

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

Our method is based on the following AIPW moment functions:

$$\widetilde{g}_i(\boldsymbol{\beta},\boldsymbol{\alpha}) = \frac{\delta_i g(X_i,Y_i,\boldsymbol{\beta})}{\pi(U_i,Y_i,\boldsymbol{\alpha})} - \frac{\delta_i - \pi(U_i,Y_i,\boldsymbol{\alpha})}{\pi(U_i,Y_i,\boldsymbol{\alpha})} m_g^0(U_i,\boldsymbol{\beta},\boldsymbol{\alpha}). \tag{2.6}$$

The imputation procedure proposed in (2.6) 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 $\widetilde{g}_i(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_0)$ are doubly robust when the propensity score (2.2) is of a special parametric model.

Proposition 1. *(i) Regardless of the choice of $m_g^0(U_i,\boldsymbol{\beta},\boldsymbol{\alpha})$, $\widetilde{g}_i(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_0)$ has mean zero, provided that the model for $\pi(U_i,Y_i,\boldsymbol{\alpha}_0)$ is specified correctly. (ii) Assume that the true response model is a parametric logistic model* $\text{logit}\{\pi(U_i,Y_i,\boldsymbol{\alpha}_0)\} = \varphi(U_i,\boldsymbol{\alpha}_0) + q(Y_i)$, *where $\varphi(\cdot)$ is a known smooth function in an unknown parameter vector $\boldsymbol{\alpha}_0$, and $q(\cdot)$ is an arbitrary user-specified (i.e., known) function that measures the departure from the ignorable missing-data mechanism assumption. Then, the AIPW moment function $\widetilde{g}_i(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_0)$ has mean zero, even if the model for $\varphi(U_i,\boldsymbol{\alpha}_0)$ is specified incorrectly.*

Remark 1. Despite enjoying the doubly robust property, Proposition 1 (ii) has a very limited application scope because an ad hoc sensitivity analysis (Vansteelandt, Rotnitzky and Robins (2007)) is required when we do 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 the motivation for our work.

It follows from Eq. (2.5) that the conditional expectation $m_g^0(U,\boldsymbol{\beta},\boldsymbol{\alpha})$ is estimable using the observed data set $\{(X_i,Y_i)$, for each $\delta_i = 1; i = 1,\dots,n\}$. To enhance the robustness against a potential model misspecification, we consider a nonparametric regression model for $f_1(Z,Y \mid U)$ that leads to a nonparametric kernel estimator of $m_g^0(U,\boldsymbol{\beta})$, given by

$$\widehat{m}_g^0(U,\boldsymbol{\beta},\boldsymbol{\alpha}_0) = \sum_{i=1}^{n} \mathcal{W}_{i0}(U,\boldsymbol{\alpha}_0) g(X_i,Y_i,\boldsymbol{\beta}),$$

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

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

Moreover, $\mathcal{W}_{i1}(U) = \delta_i \mathcal{K}_h(U-U_i) / \sum_{j=1}^n \delta_j \mathcal{K}_h(U-U_j)$, $\mathcal{K}_h(\cdot) = \text{diag}(K^{(1)}_{h_{(1)}}(\cdot),$

$\ldots, K^{(r)}_{h_{(r)}}(\cdot))$, $K^{(\nu)}_{h_{(\nu)}}(\cdot) = K^{(\nu)}(\cdot/h_{(\nu)})/h^{d_u}_{(\nu)}$, $K^{(\nu)}$ is a d_u-dimensional kernel function and $h_{(\nu)}$ is a bandwidth parameter for each $\nu \in \{1, \ldots, r\}$. Note that the kernel weight $\mathcal{W}_{i1}(U)$ represents a point mass assigned to $g(X_i, Y_i, \boldsymbol{\beta})$, for $i = 1, \ldots, n$, such that $E\{g(X, Y, \boldsymbol{\beta}) \mid U, \delta = 1\}$ can be approximated by the kernel-based regression estimator

$$\widehat{m}_g(U, \boldsymbol{\beta}) = \sum_{i=1}^{n} \mathcal{W}_{i1}(U) g(X_i, Y_i, \boldsymbol{\beta}),$$

which is widely used to deal with ignorable missing data; for example, Cheng (1994) proposed using $\widehat{m}_g(U, \boldsymbol{\beta})$ with $g(X_i, Y_i, \boldsymbol{\beta}) = Y_i - \boldsymbol{\beta}$ to develop a root-n consistent nonparametric imputation estimator for the mean response $\boldsymbol{\beta}_0 = E(Y)$.

Under the nonignorable missing-data mechanism (2.2), the set of propensity score-based and kernel-assisted nonsmooth functions for the ith individual is given by

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

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

2.4. Propensity score estimation

To make the modified moment functions $\widehat{g}_i(\boldsymbol{\beta}, \boldsymbol{\alpha})$ applicable, a consistent first-step estimator for $\boldsymbol{\alpha}_0$ should be specified in advance. Instead of using the GMM approach proposed in Wang, Shao and Kim (2014), we employ the semiparametric empirical likelihood (SEL) method (Qin, Leung and Shao (2002)) to estimate $\boldsymbol{\alpha}_0$. Because this method has been presented in the literature (Qin, Leung and Shao (2002); Zhao et al. (2017b)), we only outline the main steps of the method. We consider the following complete data likelihood function:

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

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

$$l(\boldsymbol{\omega}_0, \boldsymbol{\alpha}_0) = \sum_{i=1}^{n} \delta_i \log p_i + \sum_{i=1}^{n} \delta_i \log \pi(U_i, Y_i, \boldsymbol{\alpha}_0) + (n - n_1) \log(1 - \boldsymbol{\omega}_0), \quad (2.7)$$

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: $\sum_{i=1}^n \delta_i p_i = 1$, $p_i \geq 0$ for $i = 1, \ldots, n$, $\sum_{i=1}^n \delta_i p_i \{\pi(U_i, Y_i, \boldsymbol{\alpha}) - \boldsymbol{\omega}\} = 0$, and $\sum_{i=1}^n \delta_i p_i \phi(X_i, Y_i, \boldsymbol{\alpha}) = 0$. Here, $\phi(X_i, Y_i, \boldsymbol{\alpha})$ is an arbitrary user-specified κ-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 the efficiency improvement and is constructed from the auxiliary information in the observed data.

By introducing Lagrange multipliers λ_1 and λ_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 $i = 1, \ldots, n$. Substituting p_i into (2.7) yields

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

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

$$(\widehat{\lambda}_1, \widehat{\lambda}_2) = \arg\min_{\lambda_1, \lambda_2} l(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\omega}}, \lambda_1, \lambda_2).$$

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 $j \geq 1$. Similarly 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 recentered GEL criterion:

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

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

Given the SEL estimator $\widehat{\boldsymbol{\alpha}}$, the class of two-step AIPW-GEL estimators for $\boldsymbol{\beta}_0$ can be defined as the solution to the following saddle-point problem:

$$\widehat{\boldsymbol{\beta}}_{\scriptscriptstyle S} = \arg \inf_{\boldsymbol{\beta} \in \mathcal{B}} \sup_{\lambda \in \widehat{\Lambda}_n(\boldsymbol{\beta}, \widehat{\boldsymbol{\alpha}})} \widehat{P}_n(\boldsymbol{\beta}, \lambda, \widehat{\boldsymbol{\alpha}}), \tag{2.8}$$

where $\widehat{\Lambda}_n(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \{\lambda : \lambda^\top \widehat{g}_i(\boldsymbol{\beta}, \boldsymbol{\alpha}) \in \mathcal{V}, i = 1, \ldots, n\}$. For nonsmooth moment functions, AIPW-GEL estimators are no longer required to minimize (2.8), but

satisfy

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

where σ is nonnegative and $\widehat{\lambda}_s = \lambda(\widehat{\boldsymbol{\beta}}_s) = \arg\max_{\lambda\in\widehat{\Lambda}_n(\widehat{\boldsymbol{\beta}}_s,\widehat{\boldsymbol{\alpha}})} \widehat{P}_n(\widehat{\boldsymbol{\beta}}_s, \lambda, \widehat{\boldsymbol{\alpha}})$. 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

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

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

3. Main Results

In this section, using the 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 $\boldsymbol{\eta} = (\boldsymbol{\alpha}^\top, \boldsymbol{\omega}, \boldsymbol{\gamma}^\top)^\top$, with $\boldsymbol{\gamma} = \lambda_1(1-\boldsymbol{\omega})$, and let $\boldsymbol{\eta}_0 = (\boldsymbol{\alpha}_0^\top, \boldsymbol{\omega}_0, 0)^\top$ be the true value of $\boldsymbol{\eta}$. Denote $\widehat{\boldsymbol{\eta}} = (\widehat{\boldsymbol{\alpha}}^\top, \widehat{\boldsymbol{\omega}}, \widehat{\boldsymbol{\gamma}}^\top)^\top$ as the estimator of $\boldsymbol{\eta}$, where $\widehat{\boldsymbol{\gamma}} = \widehat{\lambda}_1(1-\widehat{\boldsymbol{\omega}})$. The following proposition shows that the SEL estimator $\widehat{\boldsymbol{\alpha}}$ proposed in Section 2.4 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)* $\boldsymbol{\eta}_0$ *is locally identified if and only if* $\operatorname{rank}(\mathbb{A}) = l + \kappa + 1$*; (ii)* $\widehat{\boldsymbol{\eta}} \xrightarrow{p} \boldsymbol{\eta}_0$ *and* $n^{1/2}(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathbb{A}^{-1}\mathbb{B}(\mathbb{A}^{-1})^\top)$*, where* $\mathbb{A}$ *and* $\mathbb{B}$ *are defined in the Supplementary Material.*

From Proposition 2, an asymptotic linear expansion for $\widehat{\boldsymbol{\alpha}}$ can be defined as $n^{1/2}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0) = n^{-1/2}\sum_{i=1}^n \Psi_i(\boldsymbol{\alpha}_0) + o_p(1)$, where $\Psi_i(\boldsymbol{\alpha}_0) = \Psi(X_i, Y_i, \boldsymbol{\alpha}_0)$ is an influence function, defined in the Supplementary Material. Let $V_1 = E\{\widetilde{g}_i(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0)\widetilde{g}_i(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0)^\top\}$, $V_2 = \operatorname{Var}\{\widetilde{g}_i(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0) - \Xi\Psi_i(\boldsymbol{\alpha}_0)\}$,

$$\Xi = \operatorname{Cov}\{\widetilde{g}_i(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0), \Delta(U, Y, \boldsymbol{\alpha}_0)\},$$

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

and $\Gamma = \partial E\{g(X, Y, \boldsymbol{\beta})\}/\partial \boldsymbol{\beta}^\top|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}$. Then, we have the following theorem.

Theorem 1. *Suppose that assumptions* (A1), (A2), (B1), (B2), *and C given in the Appendix hold. Then, the respondent probability model* $\pi(U, Y, \boldsymbol{\alpha}_0)$ *is correctly specified and the two-step AIPW-GEL estimator* $\widehat{\boldsymbol{\beta}}_s$ *is obtained by solving* (2.8), *with* $\widehat{\boldsymbol{\alpha}}$ *computed using the SEL approach. Thus, we have* $\widehat{\boldsymbol{\beta}}_s - \boldsymbol{\beta}_0 = o_p(1)$. *In addition, if assumptions* (A3), (A4), (B3), *and* (B4) *given in the Appendix hold, we obtain*

$$n^{1/2}(\widehat{\boldsymbol{\beta}}_s - \boldsymbol{\beta}_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma_s),$$

where $\Sigma_s = (\Gamma^\top V_1^{-1}\Gamma)^{-1}\Gamma^\top V_1^{-1}V_2V_1^{-1}\Gamma(\Gamma^\top V_1^{-1}\Gamma)^{-1}$.

Theorem 1 has some interesting implications. First, it indicates that the efficiency of the proposed estimators depends on the correlation between the efficient score function $\widetilde{g}_i(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0)$ and the influence function $\Psi_i(\boldsymbol{\alpha}_0)$. In particular, if $\Xi\mathrm{Var}\{\Psi_i(\boldsymbol{\alpha}_0)\}\Xi^\top - 2\Xi\mathrm{Cov}\{\widetilde{g}_i(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0), \Psi_i(\boldsymbol{\alpha}_0)\} \leq 0$, the proposed two-step AIPW-GEL estimators achieve an efficiency gain over the estimators computed using the known propensity score. This is a common phenomenon under a missing-at-random setup (Robins, Rotnitzky and Zhao (1994)). Second, if $\widetilde{g}_i(\boldsymbol{\beta}, \boldsymbol{\alpha}_0)$ is orthogonal to the score $\Delta(U, Y, \boldsymbol{\alpha}_0)$, that is, $\Xi = 0$, the limit distribution of $\widehat{\boldsymbol{\beta}}_s$ is invariant to that of $\widehat{\boldsymbol{\alpha}}$.

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, because the derivatives of the objective functions are no longer available. Hence, the Wald-type confidence regions for $\boldsymbol{\beta}_0$ 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* 1 *hold. As* $n \to \infty$, *we have* $2n\widehat{P}_n(\boldsymbol{\beta}_0, \lambda(\boldsymbol{\beta}_0), \widehat{\boldsymbol{\alpha}}) \xrightarrow{\mathcal{L}} Q^\top \Omega Q$ *under the null hypothesis* $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$, *where* $\Omega = V_2^{1/2}V_1^{-1}V_2^{1/2}$ *and* Q *is an* r*-dimensional standard normal random vector (i.e.,* $Q \sim \mathcal{N}(0, I_r)$*).*

Theorem 2 indicates that the proposed AIPW-GEL ratio converges to a linear combination of independent chi-square distributions. Despite the loss of Wilks' theorem, the confidence regions based on the AIPW-GEL ratio $2n\widehat{P}_n(\boldsymbol{\beta}_0, \lambda(\boldsymbol{\beta}_0), \widehat{\boldsymbol{\alpha}})$ are still appealing because they preserve the range and respect transformations, owing to their likelihood ratio-based nature. To construct GEL-based confidence regions for $\boldsymbol{\beta}_0$, we approximate the distribution of $2n\widehat{P}_n(\boldsymbol{\beta}_0, \lambda, \widehat{\boldsymbol{\alpha}})$ by

resampling. Let $\mathcal{X}_m^* = \{(X_i^*, Y_i^*, \delta_i^*) : i = 1, \ldots, m\}$ be a bootstrap sample from $\{\mathcal{X}_n = (X_j, Y_j, \delta_j) : j = 1, \ldots, n\}$. Based on $\mathcal{X}_m^*$, we compute the bootstrap estimator $\widehat{\boldsymbol{\alpha}}^*$ of $\boldsymbol{\alpha}_0$ using the aforementioned SEL approach. Then, the bootstrap version of $\widehat{P}_n(\boldsymbol{\beta}_0, \lambda, \widehat{\boldsymbol{\alpha}})$ is defined as

$$\widehat{P}_m^*(\widehat{\boldsymbol{\beta}}_s, \lambda^*, \widehat{\boldsymbol{\alpha}}^*) = \sum_{i=1}^{m} \frac{\{\rho(\lambda^{*\top} \widehat{g}(X_i^*, Y_i^*, \widehat{\boldsymbol{\beta}}_s, \widehat{\boldsymbol{\alpha}}^*)) - \rho_0\}}{m},$$

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

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

Let c_α^* be the $100(1-\alpha)\%$ quantile of the distribution of $2m\widehat{P}_m^*(\widehat{\boldsymbol{\beta}}_s, \lambda^*, \widehat{\boldsymbol{\alpha}}^*)$ evaluated using the resampling method. Then, it follows from Theorem 3 that the bootstrap empirical log-likelihood confidence region at the nominal coverage level $1-\alpha$ is given by $C_\alpha = \{\boldsymbol{\beta} : 2n\widehat{P}_n(\boldsymbol{\beta}, \lambda, \widehat{\boldsymbol{\alpha}}) \le c_\alpha^*\}$.

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

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

respectively, where $\mathcal{K}_h(\cdot)$ is defined based on a d_x-dimensional kernel function. That is, an alternative set of unbiased AIPW moment functions can be constructed to be structurally identical to $\widetilde{g}_i(\boldsymbol{\beta}, \boldsymbol{\alpha})$ in (2.6), except that $m_g^0(U_i, \boldsymbol{\beta}, \boldsymbol{\alpha})$ is replaced by $m_g^0(X_i, \boldsymbol{\beta}, \boldsymbol{\alpha})$.

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

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

4. Computation

Computing the proposed two-step AIPW-GEL estimators is computationally challenging because of the nonsmooth moment functions, which mean that none of the gradient functions are well defined. To overcome this difficulty, we develop

a derivative-free approach to implement the numerical optimization, based on the SA algorithm. This 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 global optimality. Assume that the estimator $\widehat{\boldsymbol{\alpha}}$ of $\boldsymbol{\alpha}_0$ has been obtained. The modified SA algorithm consists of an inner loop and an outer loop. The detailed steps are as follows.

Inner Loop. This step determines the Lagrange multiplier

$$\widehat{\lambda}_s = \arg\max_{\lambda} \widehat{P}_n(\boldsymbol{\beta}, \lambda, \widehat{\boldsymbol{\alpha}})$$

for a given $\boldsymbol{\beta}$. Let

$$\widehat{P}_\lambda(\boldsymbol{\beta}, \lambda) = \sum_{i=1}^{n} \rho_1(\lambda^\top \widehat{g}_i(\boldsymbol{\beta}, \widehat{\boldsymbol{\alpha}}))\widehat{g}_i(\boldsymbol{\beta}, \widehat{\boldsymbol{\alpha}}),$$

$$\widehat{P}_{\lambda\lambda}(\boldsymbol{\beta}, \lambda) = \sum_{i=1}^{n} \rho_2(\lambda^\top \widehat{g}_i(\boldsymbol{\beta}, \widehat{\boldsymbol{\alpha}}))\widehat{g}_i(\boldsymbol{\beta}, \widehat{\boldsymbol{\alpha}})\widehat{g}_i(\boldsymbol{\beta}, \widehat{\boldsymbol{\alpha}})^\top.$$

Then, the modified Newton–Raphson method for finding $\widehat{\lambda}_s$ is implemented using the following iterative equation:

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

where $\varrho > 0$ is a scalar step length, and $\lambda^{(t)}$ is the value of λ at the tth iteration. Here, the initial value $\lambda^{(0)}$ of λ is taken to be a zero vector. The iteration continues until the gradient function $\widehat{P}_\lambda(\boldsymbol{\beta}, \lambda_t)$ is smaller than some prespecified tolerance, such as 0.0001. Following Hansen (2015), we set ϱ to $\varrho = (\widehat{P}_2 + 3\widehat{P}_0 - 4\widehat{P}_1)/(4\widehat{P}_2 + 4\widehat{P}_0 - 8\widehat{P}_1)$, where $\widehat{P}_s = \widehat{P}_n(\boldsymbol{\beta}, \tilde{\lambda}^s, \widehat{\boldsymbol{\alpha}})$, in which $\tilde{\lambda}^s = \lambda^{(t)} - \varrho_s\{\widehat{P}_{\lambda\lambda}(\boldsymbol{\beta}, \lambda^{(t)})\}^{-1}\widehat{P}_\lambda(\boldsymbol{\beta}, \lambda^{(t)})$ 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 whether the condition $n\{1 - \lambda^\top \widehat{g}_i(\boldsymbol{\beta}, \widehat{\boldsymbol{\alpha}})\} \geq 1$ holds for all i. When the condition does not hold, we decrease the step length ϱ 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. Once $\widehat{\lambda}_s$ is obtained in the inner loop, we conduct the following minimization step: $\widehat{\boldsymbol{\beta}}_s = \arg\inf_{\boldsymbol{\beta}\in\mathcal{B}} \widehat{P}_n(\boldsymbol{\beta}, \widehat{\lambda}_s, \widehat{\boldsymbol{\alpha}})$. This can be done using the classical SA algorithm. Further details on the SA algorithm and its implementation can be found in Goffe, Ferrier and Rogers (1994). In the smooth case, this step can also be done using the Newton–Raphson method.

The proposed modified SA algorithm retains the main advantages of the classical SA method in that the objective function for implementing the SA algo-

rithm 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 general and useful because it can be used to solve various optimization problems, regardless of the linear or nonlinear relationship between the variables and the parameters, or the smooth or nonsmooth objective functions with or without missing data.

5. Simulation Studies

In this section, we present several simulation studies that 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:

$$\begin{aligned} Y &= \beta_1 + \zeta\beta_2 + \sigma(X_1, X_2)(\varepsilon - Q_\varepsilon(\tau)), \\ \zeta &= \frac{(X_1 + X_2)}{3} + \varepsilon + \varpi, \end{aligned} \tag{5.1}$$

where $\boldsymbol{\beta}_0 = (\beta_1, \beta_2)^\top = (1, 0.5)^\top$, $Q_\varepsilon(\tau)$ is the conditional τ-quantile of ε, $X_1 \sim \chi^2_1$, $X_2 \sim \chi^2_2$, $\varepsilon \sim \mathcal{N}(0,1)$, and $\varpi \sim \mathcal{N}(0,1)$, in which χ^2_k represents a 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 heteroscedasticity.

We assumed that $X = (X_1, X_2)^\top$ was completely observed, but that 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 the respondent indicator δ:

$$\Pr(\delta = 1|X, Y) = \frac{\exp(a + \alpha_1 X_2 + \alpha_2 Y)}{1 + \exp(a + \alpha_1 X_2 + \alpha_2 Y)} =: \pi(X_2, Y, \boldsymbol{\alpha}_0), \tag{5.2}$$

where $\boldsymbol{\alpha}_0 = (a, \alpha_1, \alpha_2)^\top = (a, 0.05, 0.01)^\top$, and a is set to 2.0, 1.5, 1.0, and 0.5, leading to different missing proportions. In Eq. (5.2), the respondent indicator depends on the missing variable Y, but is independent of the covariate X_1, given (X_2, Y). Thus, the covariate X_1 is treated as a nonrespondent instrument, which helps to make the parameter $\boldsymbol{\alpha}_0$ in (5.2) identifiable. Under the above settings, we have $U = X_2$ and $Z = X_1$.

We considered three quantile levels ($\tau = 0.25$, 0.5, and 0.75) for each of the four respondent probabilities. Newey and Smith (2004) showed that continuous updating is also a member of the GEL class. However, like the GMM approach,

continuous updating lacks the ability to generate likelihood ratio-based confidence regions with a shape that adapts to the support of the data. Thus, in this experiment, we focused only on EL and ET inferences. For each of the 12 combinations of three quantile levels and four respondent probabilities, we independently generated 1,000 data sets $\{(Y_i, X_i, \delta_i) : i = 1, \ldots, n\}$, with $n = 200$, according to the IVQR model (5.1) together with the propensity score model (5.2). Here, $X_i = (X_{1i}, X_{2i})^\top$. For a given data set $\{(X_i, Y_i, \delta_i) : i = 1, \ldots, n\}$, we computed the two-step AIPW-EL and AIPW-ET estimators of $\boldsymbol{\beta}_0$ based on the nonsmooth moment functions $g(X_i, Y_i, \boldsymbol{\beta}) = X_i\{I(Y_i \le \beta_1 + \zeta_i\beta_2) - \tau\}$ satisfying the restrictions $E\{g(X_i, Y_i, \boldsymbol{\beta}_0)\} = 0$, for $i = 1, \ldots, n$. In evaluating the nonparametric kernel estimation, that is, computing the nonparametric kernel estimator of the conditional expectation $m_g^0(X_{2i}, \boldsymbol{\beta}) = E\{g(X, Y, \boldsymbol{\beta}) \mid X_{2i}, \delta_i = 0\}$, for $i = 1, \ldots, n$, 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 $\boldsymbol{\beta}_0$ in the SA algorithm were computed using a complete-case analysis. The SEL approach discussed in Section 2.4 was employed to compute the estimator $\widehat{\boldsymbol{\alpha}}$ of $\boldsymbol{\alpha}_0$. The auxiliary information was defined as $\phi(X_i, Y_i, \boldsymbol{\alpha}) = \delta_i \pi^{-1}(X_{2i}, Y_i, \boldsymbol{\alpha})(X_i - \overline{X})$, where $\overline{X} = n^{-1}\sum_{i=1}^n X_i$. The initial values for $\boldsymbol{\alpha}_0$ were set to $(\tilde{\varsigma}, 0)$, where $\tilde{\varsigma} = (\tilde{a}, \tilde{\alpha}_1)$ is 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 $\operatorname{logit}\{p(X_2, \varsigma)\} = a + \alpha_1 X_2$.

The results for 1,000 repetitions in each of the 12 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 the 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 their true values, respectively, "AL" is the average length of 1,000 EL or ET-based 95% confidence intervals, and "CP" is the proportion of the 1,000 95% confidence intervals covered by the true value of the parameter. The proposed bootstrap calibration method with $B = 1,000$ bootstrap replications was used to compute the critical value of the limiting distribution of the AIPW-GEL ratio statistics.

Tables 1 and 2 reveal the following findings. (i) Under all considered circumstances, both the AIPW-EL and the AIPW-ET methods produce unbiased estimates of $\boldsymbol{\beta}_0$, 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 proba-

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

			EL				ET			
τ	$\boldsymbol{\beta}_0$	Statistic	$a=2$	$a=1.5$	$a=1$	$a=0.5$	$a=2$	$a=1.5$	$a=1$	$a=0.5$
0.25	β_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
	β_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	β_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
	β_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	β_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
	β_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

bilities of the resultant confidence intervals based on our proposed method are close to the prespecified nominal level of 95%. (iii) The missing rate improves the accuracy of the parameter estimate and the empirical coverage of the confidence interval. (iv) The AIPW-ET method has a consistently lower coverage probability and a shorter average length than those of the AIPW-EL method. (v) The AIPW-EL estimator has smaller RMS and SD values than those of the AIPW-ET estimator.

Experiment 2. In this experiment, we compared the proposed AIPW-GEL ap-

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

			EL				ET			
τ	$\boldsymbol{\beta}_0$	Statistic	$a=2$	$a=1.5$	$a=1$	$a=0.5$	$a=2$	$a=1.5$	$a=1$	$a=0.5$
0.25	β_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
	β_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	β_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
	β_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	β_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
	β_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

proach with two existing methods. The first is Tang and Qin's (2012) nonparametric multiple imputation, which assumes that missing data are ignorable. The second is the naive nonparametric imputation using follow-up data only. The other purpose of this experiment is 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 $\{(X_i, Y_i) : i = 1, \ldots, n\}$ with $n = 100$ from the two-dimensional multiplicative regression model $Y_i = \exp(X_i^\top \boldsymbol{\beta}_0)\varepsilon_i$,

where $\boldsymbol{\beta}_0 = (\beta_1, \beta_2)^\top = (0.5, 1)^\top$, $X_i = (X_{1i}, X_{2i})^\top$, and X_i is independently generated from the bivariate normal distribution $\mathcal{N}(0, \Sigma_x)$ with $\Sigma_x = (\sigma_x^{kj})$. Here, $\sigma_x^{kj} = 0.5^{|k-j|}$ for $1 \le k, j \le 2$, and ε_i is independently drawn from the following distribution assumptions: (A) $\log(\varepsilon_i) \sim \mathcal{N}(0, 1)$ and (B) $\log(\varepsilon_i) \sim$ Uniform$(0, 1)$. Similarly, we assumed that X_i is completely observed, but that Y_i may be subject to missingness. The respondent indicator δ_i for Y_i was generated from the Bernoulli distribution with probability $\pi_i(\boldsymbol{\alpha}_0)$, specified by

$$\pi_i(\boldsymbol{\alpha}_0) = \frac{\exp(a + \alpha_1 X_{2i} + \alpha_2 Y_i)}{1 + \exp(a + \alpha_1 X_{2i} + \alpha_2 Y_i)}, \tag{5.3}$$

where $\boldsymbol{\alpha}_0 = (a, \alpha_1, \alpha_2)^\top = (a, 0.05, 0.01)^\top$, with the true value of a being 1.0, 0.5, or 0.01, leading to different missing proportions. Clearly, the respondent indicator δ_i depends on missing variable Y_i, but is independent of the covariate X_{1i}, given (X_{2i}, Y_i). Thus, the covariate X_{1i} is treated as a nonrespondent instrument, which helps to make the parameter $\boldsymbol{\alpha}_0$ identifiable.

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

$$\text{LARE}_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left\{ \left| \frac{Y_i - \exp(X_i^\top \boldsymbol{\beta})}{Y_i} \right| + \left| \frac{Y_i - \exp(X_i^\top \boldsymbol{\beta})}{\exp(X_i^\top \boldsymbol{\beta})} \right| \right\}. \tag{5.4}$$

It is easily seen that the LARE function (5.4) is piecewisely differentiable with respect to $\boldsymbol{\beta}$. Let $\varepsilon_i(\boldsymbol{\beta}) = Y_i / \exp(X_i^\top \boldsymbol{\beta})$ and $g(X_i, Y_i, \boldsymbol{\beta}) = \{\varepsilon_i^{-1}(\boldsymbol{\beta}) + \varepsilon_i(\boldsymbol{\beta})\}$ $\text{sgn}\{\varepsilon_i(\boldsymbol{\beta}) - 1\} X_i$. These are a set of nonsmooth functions with respect to $\boldsymbol{\beta}$. Following Li et al. (2014), the solution to minimizing (5.4) is equivalent to the solution to estimating the equations $n^{-1} \sum_{i=1}^{n} g(X_i, Y_i, \boldsymbol{\beta}) = 0$. Thus, for a given combination of two error distribution assumptions and three respondent probabilities and each of 500 generated data sets, based on the objective functions $g(X_i, Y_i, \boldsymbol{\beta})$, we computed the following six types of GEL estimators for $\boldsymbol{\beta}_0$:

S1. the proposed AIPW-EL and AIPW-ET estimators using the correctly specified respondent probability model (5.3);

S2. the proposed AIPW-EL and AIPW-ET estimators using the misspecified respondent probability model: $\pi_i(\boldsymbol{\alpha}_0) = \Phi(a + \alpha_1 X_{2i} + \alpha_2 Y_i)$, where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution;

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

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

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

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

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

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

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

where

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

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

To evaluate estimators S1 and S2, the proposed nonparametric AIPW procedure was implemented using a one-dimensional Gaussian kernel function with bandwidth $h = 1.25 \widehat{\sigma}_{x_1} n^{-1/5}$, where $\widehat{\sigma}_{x_1}$ is the sample standard deviation of X_1. Moreover, the estimator of $\boldsymbol{\alpha}_0$ was computed using the same SEL procedure proposed in the first experiment. Estimators S3–S4 were computed 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 Experiment 1.

Table 3 reports the bias, RMS, and SD values of $6 \times 2 = 12$ estimators for $\boldsymbol{\beta}_0$. These results yield the following observations: (i) the proposed AIPW-EL and AIPW-ET estimators (i.e., S1–S4) consistently perform reasonably well, even when the working respondent probability model is misspecified; (ii) when $a = 1$, RMS and SD in S3 and S4 are smaller than those in S1 and S2; however, estimators S1 and S2 exhibit better performance in terms of RMS and SD when $a = 0.5$ and 0.01; (iii) as expected, Tang and Qin's estimator leads to considerable bias, because it depends heavily on the ignorable assumption of the respondent prob-

ability model; thus, it is sensitive to the selection of the respondent probability model; and (iv) the proposed AIPW-GEL estimators have smaller standard deviations than those of Tang and Qin's estimator and the naive estimator under the considered settings. The above findings indicate that the proposed AIPW-GEL method yields a significant improvement in the estimation efficiency over that of the naive estimation method. This is 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 using the parametrically estimated propensity score. However, 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 when the estimating equations might be of higher dimension. Here, we independently simulated 500 data sets $\{(X_i, Y_i) : i = 1, \ldots, n\}$, 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 $1 \le k, j \le 5$; 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 δ_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 using the SEL approach, incorporating the auxiliary information

$$\phi(X_i, Y_i, \boldsymbol{\alpha}) = \delta_i \pi^{-1}(X_{2i}, Y_i, \boldsymbol{\alpha})(X_i - \overline{X}),$$

where $\overline{X} = n^{-1} \sum_{i=1}^n X_i$. Assume that $\Pr(\delta_i = 1 \mid U_i) = \{1 - \exp(-\vartheta_1 - \vartheta_2^\top U_i)\}^{-1} =: \mathcal{S}(U_i, \boldsymbol{\vartheta}_0)$, where $U_i = (X_{2i}, X_{3i}, X_{4i}, X_{5i})$. We estimate $\boldsymbol{\vartheta}_0$ using the maximum likelihood estimation method and denote the estimate as $\widehat{\boldsymbol{\vartheta}} = (\widehat{\vartheta}_1, \widehat{\vartheta}_2^\top)^\top$. Let $\widehat{\mathcal{S}}_i = \mathcal{S}(U_i, \widehat{\boldsymbol{\vartheta}}) = \{1 - \exp(-\widehat{\vartheta}_1 - \widehat{\vartheta}_2^\top U_i)\}^{-1}$. We then computed the kernel dimension-reduction AIPW-GEL estimators for $\boldsymbol{\beta}_0$ using the one-dimensional kernel smoothers $\mathcal{K}_h(\widehat{\mathcal{S}} - \widehat{\mathcal{S}}_i)$. Table 4 reports the simulated bias, RMS, and SD. It can be seen that the proposed estimators have negligible bias, and that the values of RMS are quite close to those of SD.

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

				$a=1$			$a=0.5$			$a=0.01$		
ε_i	Type	GEL	$\boldsymbol{\beta}_0$	Bias	RMS	SD	Bias	RMS	SD	Bias	RMS	SD
A	S1	EL	β_1	0.008	0.148	0.148	0.010	0.154	0.154	0.022	0.174	0.173
			β_2	0.002	0.151	0.151	0.009	0.155	0.155	-0.008	0.183	0.183
		ET	β_1	0.009	0.149	0.149	0.009	0.154	0.154	0.024	0.174	0.173
			β_2	0.003	0.151	0.151	0.010	0.155	0.155	-0.008	0.182	0.182
	S2	EL	β_1	0.009	0.149	0.149	0.011	0.154	0.154	0.021	0.174	0.173
			β_2	0.004	0.152	0.152	0.010	0.156	0.156	-0.006	0.183	0.183
		ET	β_1	0.009	0.150	0.150	0.011	0.153	0.153	0.021	0.174	0.173
			β_2	0.004	0.152	0.152	0.011	0.155	0.155	-0.006	0.183	0.183
	S3	EL	β_1	0.008	0.146	0.146	0.005	0.163	0.163	0.014	0.186	0.186
			β_2	0.001	0.138	0.138	0.008	0.160	0.160	0.013	0.192	0.191
		ET	β_1	0.009	0.146	0.146	0.005	0.161	0.161	0.013	0.186	0.185
			β_2	0.000	0.139	0.139	0.009	0.161	0.161	0.014	0.190	0.190
	S4	EL	β_1	0.008	0.145	0.145	0.005	0.164	0.164	0.014	0.188	0.187
			β_2	0.001	0.139	0.139	0.009	0.160	0.160	0.011	0.191	0.191
		ET	β_1	0.008	0.145	0.145	0.005	0.163	0.163	0.014	0.188	0.188
			β_2	0.000	0.138	0.138	0.007	0.161	0.161	0.011	0.191	0.191
	S5	EL	β_1	-0.039	0.352	0.350	-0.037	0.396	0.395	-0.095	0.409	0.398
			β_2	-0.074	0.361	0.354	-0.118	0.440	0.424	-0.192	0.453	0.411
		ET	β_1	-0.024	0.251	0.250	-0.014	0.278	0.278	-0.054	0.290	0.285
			β_2	-0.030	0.272	0.271	-0.068	0.291	0.283	-0.097	0.309	0.294
	S6	EL	β_1	0.024	0.188	0.187	0.013	0.201	0.201	0.026	0.246	0.245
			β_2	0.000	0.192	0.193	0.005	0.201	0.201	-0.018	0.238	0.238
		ET	β_1	0.024	0.188	0.187	0.014	0.200	0.199	0.025	0.244	0.243
			β_2	0.000	0.193	0.193	0.004	0.200	0.200	-0.017	0.238	0.238
B	S1	EL	β_1	-0.003	0.103	0.103	-0.003	0.116	0.116	0.006	0.123	0.123
			β_2	0.005	0.095	0.095	0.008	0.099	0.099	0.003	0.114	0.114
		ET	β_1	-0.003	0.102	0.102	-0.003	0.114	0.114	0.006	0.120	0.120
			β_2	0.004	0.092	0.092	0.008	0.098	0.098	0.002	0.113	0.113
	S2	EL	β_1	-0.003	0.102	0.102	-0.003	0.117	0.117	0.005	0.123	0.123
			β_2	0.005	0.095	0.095	0.008	0.100	0.099	0.004	0.115	0.115
		ET	β_1	-0.003	0.101	0.101	-0.004	0.116	0.116	0.006	0.119	0.119
			β_2	0.004	0.094	0.094	0.009	0.097	0.097	0.002	0.112	0.112
	S3	EL	β_1	0.001	0.097	0.097	-0.009	0.108	0.108	-0.007	0.115	0.115
			β_2	-0.003	0.091	0.091	0.006	0.105	0.105	0.005	0.103	0.103
		ET	β_1	0.000	0.096	0.096	-0.008	0.107	0.107	-0.007	0.112	0.112
			β_2	-0.002	0.092	0.092	0.006	0.105	0.105	0.005	0.103	0.103
	S4	EL	β_1	0.001	0.097	0.097	-0.008	0.107	0.107	-0.007	0.114	0.114
			β_2	-0.003	0.091	0.091	0.006	0.105	0.105	0.005	0.103	0.103
		ET	β_1	0.001	0.096	0.096	-0.009	0.107	0.107	-0.009	0.114	0.114
			β_2	-0.002	0.090	0.090	0.007	0.105	0.105	0.006	0.102	0.102
	S5	EL	β_1	-0.041	0.262	0.259	-0.058	0.279	0.274	-0.102	0.319	0.303
			β_2	-0.101	0.310	0.293	-0.141	0.340	0.310	-0.173	0.370	0.327
		ET	β_1	-0.018	0.215	0.215	-0.031	0.229	0.227	-0.043	0.251	0.247
			β_2	-0.009	0.232	0.232	-0.046	0.240	0.236	-0.043	0.270	0.267
	S6	EL	β_1	0.002	0.145	0.145	-0.010	0.158	0.158	0.001	0.179	0.180
			β_2	0.018	0.136	0.135	0.024	0.145	0.143	0.034	0.165	0.162
		ET	β_1	0.001	0.146	0.146	-0.011	0.157	0.157	0.003	0.177	0.177
			β_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.

		$a = 0.7$					$a = 0.25$				
GEL	Statistic	β_1	β_2	β_3	β_4	β_5	β_1	β_2	β_3	β_4	β_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

6. A Real Example

We use a data set on the serum-cholesterol levels of heart-attack patients (Schafer (1997)) to illustrate the proposed methodologies. In this data set, the serum-cholesterol levels of 28 heart-attack patients at a Pennsylvania medical center were measured 2, 4, and 14 days after the heart attack. Let X_{1i} be the cholesterol level of the ith patient measured after two days, X_{2i} be the cholesterol level of the ith patient measured after four days, and Y_i be the cholesterol level of the ith patient measured after 14 days, for $i = 1, \ldots, 28$. We find that X_{1i} and X_{2i} were completely observed, but that Y_i is subject to missingness. Let δ_i be an indicator function taking the value one if Y_i is observed, and zero if Y_i is missing. The proportion of missing observations for Y_i is 32%. Let $X_i = (X_{1i}, X_{2i})$. Schafer (1997) analyzed this data set using an EM algorithm under a trivariate normal distribution assumption of (X_i, Y_i), together with an ignorable assumption of missing values for Y_i. Unlike Schafer (1997), we assume 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 following τth conditional quantile of Y_i: $Q_{Y_i}(\tau \mid X_i) = \mathbb{X}_i^\top \boldsymbol{\beta}_0$, $i = 1, \ldots, 28$, 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 $\boldsymbol{\beta}_0$ are defined as $g(X_i, Y_i, \boldsymbol{\beta}_0) = \mathbb{X}_i\{I(Y_i \le \mathbb{X}_i^\top \boldsymbol{\beta}_0) - \tau\}$, 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:

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

$$\text{Model 2: } \pi(U_i, Y_i, \boldsymbol{\alpha}_0) = \Phi(\alpha_1 + \alpha_2 U_i + \alpha_3 Y_i),$$

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

To determine U_i or Z_i, we follow Shao and Wang (2016) and consider the following criterion:

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

where $\widetilde{\boldsymbol{\alpha}}$ is an estimator of $\boldsymbol{\alpha}_0$ computed by the SEL approach with a candidate U_i. Note that D converges to zero if and only if $\pi(U_i, Y_i, \boldsymbol{\alpha}_0)$ is correctly specified and $\widetilde{\boldsymbol{\alpha}} = \boldsymbol{\alpha}_0 + o_p(1)$ and Z_i is the instrument. Hence, a 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.

Similarly to the simulation studies, we used a one-dimensional Gaussian kernel function with bandwidth $h = 1.25\hat{\sigma}_{x_1} n^{-1/5}$ to evaluate the nonparametric kernel estimation, where $\hat{\sigma}_{x_1}$ is the sample standard deviation of $\{X_{1i}, i = 1 \ldots, n\}$. For a given value of τ, we calculated the proposed two-step AIPW-EL and AIPW-ET estimators of $\boldsymbol{\beta}_0$. The proposed bootstrap calibration procedure with 200 bootstrap replications was adopted to estimate the standard errors (SE) of the proposed estimators, and to calculate the 95% bootstrap percentile-based confidence intervals. In addition, we calculated the 95% EL-based and ET-based confidence intervals for the parameters $\beta_1(\tau), \beta_2(\tau)$ and $\beta_3(\tau)$, respectively.

Table 5 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 5, we have the following observations. First, all of the considered parameter estimates yield the same conclusion that the cholesterol level measured after two days has a negative effect on the cholesterol level measured after 14 days. Furthermore, the cholesterol level measured after four days has a positive effect on the cholesterol level measured after 14 days, regardless of the quantile levels and the specified response probability models. Second, in general, the EL-based and ET-based methods lead to narrower confidence intervals than those generated by the bootstrap method.

7. Discussion

In an attempt to improve and refine existing methods for handling nonignorable missing data, we assume a parametric nonignorable propensity score model. Then, we propose a propensity score-based nonparametric imputation approach

Table 5. Results of the real-data analysis.

			EL			ET		
Model	$\boldsymbol{\beta}_0$	Statistic	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$
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^1	2.017	1.887	1.760	2.018	1.921	1.784
		length^2	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^1	0.893	0.919	0.836	0.921	0.934	0.855
		length^2	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^1	1.001	0.925	0.835	1.020	0.969	0.857
		length^2	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^1	2.176	1.875	1.767	2.036	1.842	1.744
		length^2	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^1	0.944	0.984	0.855	0.936	0.961	0.822
		length^2	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^1	1.020	0.912	0.899	1.009	0.961	0.862
		length^2	0.185	0.170	0.198	0.183	0.161	0.172

NOTE: length^1 denotes the lengths of approximate 95% bootstrap-based confidence intervals with 200 bootstrap replications, and length^2 represents the lengths of approximate 95% EL-based (or ET-based) confidence intervals.

that uses an instrument to address the potential model identifiability problem in the presence of nonignorable missing data. Moreover, we adopt the GEL together with the AIPW approach to make statistical inferences on the parameters of the nonsmooth moment functions, and the large-sample results are established under some fairly mild conditions.

Correctly specifying the propensity score model is critical to the proposed method. If the nonrespondent instrument is not selected appropriately, the propensity score model might be incorrectly specified, leading to misleading conclusions. 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 fully observed covariates; (b) it is not directly related

to the response mechanism, conditional on the fully observed covariates. If a nonrespondent instrument is manually selected as a subset or function of the auxiliary variables, the above two conditions are difficult to verify in practical applications.

Although the criterion D proposed in Shao and Wang (2016) is sensitive to the choice of instrument, it might not help us determine the best subset of instruments if multiple instruments are available. Choosing from among valid instruments is important when many are thought to be equally valid. Similarly to Donald, Imbens and Newey (2009), we could develop asymptotic mean square error (MSE)-based criteria, related to the efficiency of the resultant estimators, for the instrument selection in our estimation of the nonsmooth moment conditions with nonignorable missing data. An optimal instrument should simultaneously minimize the D and MSE criteria. The results are currently under investigation and are not discussed further here.

Supplementary Material

The online Supplementary Material contains detailed technical proofs of Propositions 1–2, Theorems 1–2, and the related lemmas.

Acknowledgements

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

Appendix

We first provide a few notation used in the rest of the paper. Define $\mathcal{N}_\varrho =: \{\boldsymbol{\beta} \in \mathcal{B} : |\boldsymbol{\beta} - \boldsymbol{\beta}_0| < \varrho\}$ 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 $\mathcal{H}(u, \boldsymbol{\beta})$, $\|\mathcal{H}(\boldsymbol{\beta})\|_\infty = \sup_{u \in \mathcal{U}} |\mathcal{H}(u, \boldsymbol{\beta})|$ for any $\boldsymbol{\beta} \in \mathcal{B}$, $\mathcal{U}$ is the support of random vectors U. For ease of presentation, we consider the nonparametric estimation of $m_g^0(U, \boldsymbol{\beta})$ using the same kernel function, that is, for each $\nu = 1, \ldots, r$, 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 $h_n \to 0$ as $n \to \infty$. Let $\mathcal{G}_n(\boldsymbol{\beta}, \boldsymbol{\alpha}) = n^{-1} \sum_{i=1}^n \widehat{g}_i(\boldsymbol{\beta}, \boldsymbol{\alpha})$, $\mathcal{G}(\boldsymbol{\beta}) = E\{g(X, Y, \boldsymbol{\beta})\}$, 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 $g(X, Y, \boldsymbol{\beta})$ satisfies:*

(A1) *(a) $\boldsymbol{\beta}_0 \in \mathcal{B}$ is the unique solution to $\mathcal{G}(\boldsymbol{\beta}) = 0$, and $\mathcal{B}$ is a compact subset of $\mathcal{R}^p$;*

(b) $E\{\sup_{\boldsymbol{\beta}\in\mathcal{B}} |g(X, Y, \boldsymbol{\beta})|\} < \infty$ and $E\{\sup_{\boldsymbol{\beta}\in\mathcal{N}_\varrho} |g(X, Y, \boldsymbol{\beta})|^2\} < \infty$;

(A2) *the class of functions $\{g(X, Y, \boldsymbol{\beta}) : \boldsymbol{\beta} \in \mathcal{B}\}$ is Glivenko-Cantelli;*

(A3) *for some $\varrho > 0$, $\{g(X, Y, \boldsymbol{\beta}) : \boldsymbol{\beta} \in \mathcal{N}_\varrho\}$ is Donsker;*

(A4) *for all $\boldsymbol{\beta} \in \mathcal{B}$ and all small positive value $\varrho = o(1)$,*

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

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

Remark A1. 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 $g(X, Y, \boldsymbol{\beta}) = X[I(Y \le X^\top\boldsymbol{\beta}) - \tau]$ 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 $E|X|^3 \le K_x$. Let X_j denote the jth coordinate of X, $j = 1, \ldots, r$, 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 $\mathcal{F} = \{I(Y \le X^\top\boldsymbol{\beta}), \boldsymbol{\beta} \in \mathcal{B}\}$ 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 $2\max_{j\in1,\ldots,r} |X|_j$ (see Theorem 2.10.6 in Van der Vaart and Wellner (1996)). Assumptions (A2) and (A3) are then verified since the functional class $\{X[I(Y \le X^\top\boldsymbol{\beta}) - \tau], \tau \in \mathcal{T}, \boldsymbol{\beta} \in \mathcal{B}\}$ is formed as $X(\mathcal{F} - \mathcal{T})$. For $j = 1, \ldots, r$, $|g_j(X, Y, \boldsymbol{\beta}') - g_j(X, Y, \boldsymbol{\beta})|^2 \le X_j^2 |I\{Y \le X^\top\boldsymbol{\beta}'\} - I\{Y \le X^\top\boldsymbol{\beta}\}|$. For small enough $\varrho > 0$, $E[\sup_{|\boldsymbol{\beta}-\boldsymbol{\beta}'|\le\varrho} X_j^2 |I\{Y \le X^\top\boldsymbol{\beta}'\} - I\{Y \le X^\top\boldsymbol{\beta}\}|] \le E[|X|^3\{F(X^\top\boldsymbol{\beta} + \varrho \mid X) - F(X^\top\boldsymbol{\beta} - \varrho \mid X)\}] \le \mathcal{C}\varrho$, 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 $m_g^0(U, \boldsymbol{\beta}, \boldsymbol{\alpha}_0)$ defined in* (2.5) *satisfies the following conditions:*

(B1) *the class of functions $\{m_g^0(U, \boldsymbol{\beta}, \boldsymbol{\alpha}_0) : \boldsymbol{\beta} \in \mathcal{B}\}$ is Glivenko-Cantelli;*

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

(a) $m_g^0(U, \boldsymbol{\beta}, \boldsymbol{\alpha}_0)$ *is continuously differentiable with derivative*

$$\partial_{\boldsymbol{\beta}} m_g^0(U, \boldsymbol{\beta}, \boldsymbol{\alpha}_0) =: \partial m_g^0(U, \boldsymbol{\beta}, \boldsymbol{\alpha}_0)/\partial \boldsymbol{\beta}^\top$$

in $\boldsymbol{\beta} \in \mathcal{N}_\varrho$*;*

(b) $E\{\sup_{\boldsymbol{\beta} \in \mathcal{N}_\varrho} |\partial_{\boldsymbol{\beta}} m_g^0(U, \boldsymbol{\beta}, \boldsymbol{\alpha}_0)|\} < \infty$*;*

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

$$\begin{aligned} &|\partial_{\boldsymbol{\beta}} \widetilde{m}_g^0(U, \boldsymbol{\beta}, \boldsymbol{\alpha}_0) - \partial_{\boldsymbol{\beta}} m_g^0(U, \boldsymbol{\beta}, \boldsymbol{\alpha}_0)| \\ &\leq b(U) \sup_{\boldsymbol{\beta} \in \mathcal{N}_\varrho} \|\widetilde{m}_g^0(\boldsymbol{\beta}, \boldsymbol{\alpha}_0) - m_g^0(\boldsymbol{\beta}, \boldsymbol{\alpha}_0)\|_\infty^\epsilon \end{aligned}$$

for all smooth functions $\widetilde{m}_g^0(U, \boldsymbol{\beta}, \boldsymbol{\alpha}_0) \in \mathcal{M}$ *with* $\sup_{\boldsymbol{\beta} \in \mathcal{N}_\varrho} \|\widetilde{m}_g^0(\boldsymbol{\beta}, \boldsymbol{\alpha}_0) - m_g^0(\boldsymbol{\beta}, \boldsymbol{\alpha}_0)\|_\infty < \varrho$*.*

Remark A2. Assumption B restricts the class of functions

$$\begin{aligned} \mathscr{G} = \Big\{ &\widetilde{m}_g^0(U, \boldsymbol{\beta}, \boldsymbol{\alpha}_0) : \widetilde{m}_g^0(U, \boldsymbol{\beta}, \boldsymbol{\alpha}_0) \in \mathcal{M}, \boldsymbol{\beta} \in \mathcal{N}_\varrho \text{ and} \\ &\|\widetilde{m}_g^0(\boldsymbol{\beta}, \boldsymbol{\alpha}_0) - m_g^0(\boldsymbol{\beta}, \boldsymbol{\alpha}_0)\|_\infty < \varrho \Big\}, \end{aligned}$$

where $m_g^0(U, \boldsymbol{\beta}, \boldsymbol{\alpha}_0) \in \mathscr{G}$ by construction. It is easy to verify Assumption (B1) because it is natural to assume that the conditional expectations $E\{\delta g(X, Y, \boldsymbol{\beta}) O(U, Y, \boldsymbol{\alpha}_0) \mid U\}$ are smooth in $\boldsymbol{\beta}$; Assumption (B2) is a usual dominance condition. Assumption (B3) is similar to Assumption 4 in Chen, Hong and Tamer (2005) and Assumption 7 in Cattaneo (2010). Assumption (B3) further restricts function class $\{g(X, Y, \boldsymbol{\beta}) : \boldsymbol{\beta} \in \mathcal{N}_\varrho\}$ 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* $\delta \perp\!\!\!\perp Z \mid (U, Y)$*, where* U *is continuously distributed with Lebesgue density* f*; (ii) The probability density function* $f(u)$ *is bounded away from* ∞ *in the support of* U *and the second derivative of* $f(u)$ *is continuous and bounded.*

(C2) *(a) For all* $\boldsymbol{\alpha}$ *in a neighborhood of* $\boldsymbol{\alpha}_0$*, the propensity model* $\pi(U, Y, \boldsymbol{\alpha})$ *is twice differentiable with respect to* $\boldsymbol{\alpha}$ *and* $E|\pi(U, Y, \boldsymbol{\alpha})|^3 < \infty$*; (b)* $\pi(U) = E\{\pi(U, Y, \boldsymbol{\alpha}_0) \mid U\} \neq 1$ *a.s.; (c) uniformly for all* $\boldsymbol{\alpha} \in \mathcal{A}$*,* $\pi(U_i, Y_i, \boldsymbol{\alpha}) \geq \mathcal{C} > 0$ *for all* $i = 1, \ldots, n$*, 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) $\int K(u_1, \ldots, u_{d_u}) du_1 \ldots du_{d_u} = 1$*;*

(iii) $\int u_s^l K(u_1, \ldots, u_{d_u}) du_1 \ldots du_{d_u} = 0$ *and* $\int u_s^q K(u_1, \ldots, u_{d_u}) du_1 \ldots du_{d_u} \neq 0$ *for any* $s = 1, \ldots, d_u$ *and* $1 \leq l < q$*.*

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

Remark A3. (C1) is conditional independence assumption, which is used to achieve identification with nonignorable missing data (e.g., Wang, Shao and Kim (2014) and Zhao and Shao (2015)). Assumptions (C2)–(C4) are commonly used in the missing data analysis and nonparametric regression inference.

To prove Theorems 1–3, we need the following lemmas, whose proofs can be found in the Supplementary Material.

Lemma A1. *Suppose that Assumption C holds. Then, we have*

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

Lemma A2. *Suppose that Assumption C holds; that the respondent probability model* $\pi(U, Y, \boldsymbol{\alpha}_0)$ *is correctly specified; and that* $\widehat{\boldsymbol{\alpha}}$ *is computed by the SEL approach. Then, we have*

$$\mathcal{G}_n(\boldsymbol{\beta}_0, \widehat{\boldsymbol{\alpha}}) = \frac{1}{n} \sum_{i=1}^n \widetilde{g}_i(\boldsymbol{\beta}, \boldsymbol{\alpha}_0) - \Xi \times (\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0) + o_p(n^{-1/2}),$$

where $\widetilde{g}_i(\boldsymbol{\beta}, \boldsymbol{\alpha})$ *is defined in* (2.6)*,* $\Xi = \mathrm{Cov}\{\widetilde{g}_i(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0), \Delta(U, Y, \boldsymbol{\alpha}_0)\}$ *with* $\Delta(U, Y, \boldsymbol{\alpha}) = \{\delta - \pi(U, Y, \boldsymbol{\alpha})\} \partial \mathrm{logit}\{\pi(U, Y, \boldsymbol{\alpha})\}/\partial \boldsymbol{\alpha}^\top$*.*

Lemma A3. *Suppose that Assumption C holds; that the respondent probability model* $\pi(U, Y, \boldsymbol{\alpha}_0)$ *is correctly specified; and that* $\widehat{\boldsymbol{\alpha}}$ *is computed by the SEL approach. Then, we have*

$$\frac{1}{n} \sum_{i=1}^n \widehat{g}_i(\boldsymbol{\beta}_0, \widehat{\boldsymbol{\alpha}}) \widehat{g}_i(\boldsymbol{\beta}_0, \widehat{\boldsymbol{\alpha}})^\top = V_1 + o_p(1),$$

where $V_1 = E\{\widetilde{g}_i(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0) \widetilde{g}_i(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0)^\top\}$*.*

Lemma A4. *Suppose that Assumptions A, B and C hold; that the respondent probability model $\pi(U, Y, \boldsymbol{\alpha}_0)$ is correctly specified; and that $\widehat{\boldsymbol{\alpha}}$ is computed by the SEL approach. Then, for all positive $\varrho_n = o_p(1)$, we have*

$$\sup_{|\boldsymbol{\beta}-\boldsymbol{\beta}_0|\leq\varrho_n} \frac{|\mathcal{G}_n(\boldsymbol{\beta}, \widehat{\boldsymbol{\alpha}}) - \mathcal{G}(\boldsymbol{\beta}) - \mathcal{G}_n(\boldsymbol{\beta}_0, \widehat{\boldsymbol{\alpha}}) + \mathcal{G}(\boldsymbol{\beta}_0)|}{1 + \mathcal{C}n^{1/2}|\boldsymbol{\beta} - \boldsymbol{\beta}_0|} = o_p(n^{-1/2}).$$

Lemma A5. *Suppose that Assumptions* (A1) *and C hold; that the respondent probability model $\pi(U, Y, \boldsymbol{\alpha}_0)$ is correctly specified; and that $\widehat{\boldsymbol{\alpha}}$ is computed by the SEL approach. Then, for $\Lambda_n = \{\lambda : |\lambda| \leq \mathcal{C}n^{-1/2}\}$, we obtain*

$$\sup_{\boldsymbol{\beta}\in\mathcal{B},\lambda\in\Lambda_n,1\leq i\leq n} |\lambda^\top \widehat{g}_i(\boldsymbol{\beta}, \widehat{\boldsymbol{\alpha}})| \xrightarrow{p} 0$$

and w.p.1, *$\Lambda_n \subseteq \widehat{\Lambda}_n(\boldsymbol{\beta}, \boldsymbol{\alpha})$ for all $\boldsymbol{\beta} \in \mathcal{B}$ and $\boldsymbol{\alpha} \in \mathcal{A}$.*

Lemma A6. *Suppose that Assumptions* (A1) *and C hold; that the respondent probability model $\pi(U, Y, \boldsymbol{\alpha}_0)$ is correctly specified; and that $\widehat{\boldsymbol{\alpha}}$ is computed by SEL approach. Then, we have $|\mathcal{G}_n(\widehat{\boldsymbol{\beta}}_s, \widehat{\boldsymbol{\alpha}})| = O_p(n^{-1/2})$.*

Proof of Theorem 3. The proof for Theorem 3 essentially involves establishing bootstrap version of Lemma A2 and Theorem 2. We only establish the bootstrap version of Lemma A2 here. Let X_i^*, Y_i^* and δ_i^* be the counterparts of X_i, Y_i, and δ_i in the bootstrap sample, respectively. Let $\widehat{\boldsymbol{\eta}}^* = (\widehat{\boldsymbol{\alpha}}^{*\top}, \widehat{\boldsymbol{\omega}}^*, \widehat{\boldsymbol{\gamma}}^{*\top})^\top$ be the bootstrap estimator of $\boldsymbol{\eta}_0 = (\boldsymbol{\alpha}_0^\top, \boldsymbol{\omega}_0, \boldsymbol{\gamma}_0^\top)^\top$. We use E_* to represent the conditional expectation given the original data. Define $\pi_i^*(\boldsymbol{\alpha}) = \pi(X_i^*, Y_i^*, \boldsymbol{\alpha})$, $\widehat{m}_g^*(U, \boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^m \mathcal{W}_{i0}^*(U, \boldsymbol{\alpha}) g(X_i^*, Y_i^*, \boldsymbol{\beta})$, and

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

in which $\mathcal{W}_{i0}^*(U, \boldsymbol{\alpha}) = \delta_i^* O_i^*(\boldsymbol{\alpha}) \mathcal{K}_h(U - U_i^*) / \{\sum_{k=1}^n \delta_k^* O_k^*(\boldsymbol{\alpha}) \mathcal{K}_h(U - U_k^*)\}$ with $O_i^*(\boldsymbol{\alpha}) = O(U_i^*, Y_i^*, \boldsymbol{\alpha})$. Then there exists $\boldsymbol{\alpha}^\dagger$ between $\widehat{\boldsymbol{\alpha}}^*$ and $\widehat{\boldsymbol{\alpha}}$ such that $\mathcal{G}_m^*(\widehat{\boldsymbol{\beta}}_s, \widehat{\boldsymbol{\alpha}}^*) - \mathcal{G}_n(\widehat{\boldsymbol{\beta}}_s, \widehat{\boldsymbol{\alpha}}) = \mathcal{G}_m^*(\widehat{\boldsymbol{\beta}}_s, \widehat{\boldsymbol{\alpha}}) - \mathcal{G}_n(\widehat{\boldsymbol{\beta}}_s, \widehat{\boldsymbol{\alpha}}) + \partial \mathcal{G}_m^*(\widehat{\boldsymbol{\beta}}_s, \boldsymbol{\alpha}^\dagger)/\partial \boldsymbol{\alpha}^\top (\widehat{\boldsymbol{\alpha}}^* - \widehat{\boldsymbol{\alpha}})$. Similar to the proof of Proposition 2 but replace the functions and the parameters with their corresponding bootstrap analogs, we can show $n^{1/2}(\widehat{\boldsymbol{\eta}}^* - \widehat{\boldsymbol{\eta}}) \xrightarrow{\mathcal{L}^*} \mathcal{N}(0, \mathbb{A}^{-1}\mathbb{B}(\mathbb{A}^{-1})^\top)$, 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 A2 by the following two steps.

Step 1. Show that $\partial \mathcal{G}_m^*(\widehat{\boldsymbol{\beta}}_s, \boldsymbol{\alpha}^\dagger)/\partial \boldsymbol{\alpha}^\top \xrightarrow{p} \Xi$. By Assumption C and similar arguments to those used in the proof of Lemma A2, as $n \to \infty$ and $m \to \infty$, we can obtain the result.

Step 2. Show that $n^{1/2}\{\mathcal{G}_m^*(\widehat{\boldsymbol{\beta}}_s, \widehat{\boldsymbol{\alpha}}) - \mathcal{G}_n(\widehat{\boldsymbol{\beta}}_s, \widehat{\boldsymbol{\alpha}})\} \xrightarrow{\mathcal{L}^*} \mathcal{N}(0, V_1)$. To prove this result, we only need to prove that $n^{1/2}\{\mathcal{G}_m^*(\widehat{\boldsymbol{\beta}}_s, \boldsymbol{\alpha}_0) - \mathcal{G}_n(\widehat{\boldsymbol{\beta}}_s, \boldsymbol{\alpha}_0)\} \xrightarrow{\mathcal{L}^*} \mathcal{N}(0, V_1)$ because $\partial \mathcal{G}_m^*(\widehat{\boldsymbol{\beta}}_s, \boldsymbol{\alpha}^{\dagger\dagger})/\partial \boldsymbol{\alpha}^\top = \partial \mathcal{G}_n(\widehat{\boldsymbol{\beta}}_s, \boldsymbol{\alpha}^{\dagger\dagger})/\partial \boldsymbol{\alpha}^\top + o_p(1)$ as $n \to \infty$ and $m \to \infty$ and $\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0 = O_p(n^{-1/2})$. Here $\boldsymbol{\alpha}^{\dagger\dagger}$ lies between $\widehat{\boldsymbol{\alpha}}$ and $\boldsymbol{\alpha}_0$. Next we note that $\mathcal{G}_m^*(\widehat{\boldsymbol{\beta}}_s, \boldsymbol{\alpha}_0) - \mathcal{G}_n(\widehat{\boldsymbol{\beta}}_s, \boldsymbol{\alpha}_0) = K_{mn1} + K_{mn2} + K_{mn3}$, where

$$
\begin{aligned}
K_{mn1} = \frac{1}{m}\sum_{i=1}^{m}\Bigg[& \frac{\delta_i^*}{\pi_i^*(\boldsymbol{\alpha}_0)} g(X_i^*, Y_i^*, \widehat{\boldsymbol{\beta}}_s) - \frac{\delta_i - \pi_i(\boldsymbol{\alpha}_0)}{\pi_i^*(\boldsymbol{\alpha}_0)} m_g^0(U_i^*, \widehat{\boldsymbol{\beta}}_s, \boldsymbol{\alpha}_0) \\
& - \frac{1}{n}\sum_{i=1}^{n}\left\{ \frac{\delta_i}{\pi_i(\boldsymbol{\alpha}_0)} g(X_i, Y_i, \widehat{\boldsymbol{\beta}}_s) - \frac{\delta_i - \pi_i(\boldsymbol{\alpha}_0)}{\pi_i(\boldsymbol{\alpha}_0)} m_g^0(U_i, \widehat{\boldsymbol{\beta}}_s, \boldsymbol{\alpha}_0) \right\}\Bigg], \\
K_{mn2} = \frac{1}{m}\sum_{i=1}^{m}\Bigg[& \left\{ 1 - \frac{\delta_i^*}{\pi_i^*(\boldsymbol{\alpha}_0)} \right\}\{\widehat{m}_g^0(U_i^*, \widehat{\boldsymbol{\beta}}_s, \boldsymbol{\alpha}_0) - m_g^0(U_i^*, \widehat{\boldsymbol{\beta}}_s, \boldsymbol{\alpha}_0)\} \\
& - \frac{1}{n}\sum_{i=1}^{n}\left\{ 1 - \frac{\delta_i}{\pi_i(\boldsymbol{\alpha}_0)} \right\}\{\widehat{m}_g^0(U_i, \widehat{\boldsymbol{\beta}}_s, \boldsymbol{\alpha}_0) - m_g^0(U_i, \widehat{\boldsymbol{\beta}}_s, \boldsymbol{\alpha}_0)\}\Bigg], \\
K_{mn3} = \frac{1}{m}\sum_{i=1}^{m} & \left\{ 1 - \frac{\delta_i^*}{\pi_i^*(\boldsymbol{\alpha}_0)} \right\}\{\widehat{m}_g^*(U_i^*, \widehat{\boldsymbol{\beta}}_s, \boldsymbol{\alpha}_0) - \widehat{m}_g^0(U_i^*, \widehat{\boldsymbol{\beta}}_s, \boldsymbol{\alpha}_0)\}.
\end{aligned}
$$

For K_{mn1}, we can apply the central limit theorem for bootstrap samples (Shao and Tu (1995)) to derive $n^{1/2}K_{mn1} \xrightarrow{\mathcal{L}} \mathcal{N}[0, E_*\{\widetilde{g}_i(\widehat{\boldsymbol{\beta}}_s, \boldsymbol{\alpha}_0)\widetilde{g}_i(\widehat{\boldsymbol{\beta}}_s, \boldsymbol{\alpha}_0)^\top\}]$. Use similar argument to I_{n2} in Lemma A2 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 $E_*\{\widetilde{g}_i(\widehat{\boldsymbol{\beta}}_s, \boldsymbol{\alpha}_0)\widetilde{g}_i(\widehat{\boldsymbol{\beta}}_s, \boldsymbol{\alpha}_0)^\top\} \to V_1$ as $n \to \infty$ and $m \to \infty$. The bootstrap version of Lemma A2 could be established by combining above arguments.

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. *The 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 (Statistical Methodology)* **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 University Press, Cambridge.

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 Edition), Wiley, New York.

Miao, W. and 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 confidece regions. *The 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 Statstical 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.* John Wiley, New York.

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.* Springer, New York.

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.

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

E-mail: pyzhao@live.cn

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

E-mail: nstang@ynu.edu.cn

Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599-7420, USA.

E-mail: hzhu@bios.unc.edu

(Received January 2017; accepted March 2018)