Statistica Sinica **31** (2021), 647-672
doi:https://doi.org/10.5705/ss.202019.0025

PROPENSITY MODEL SELECTION WITH NONIGNORABLE NONRESPONSE AND INSTRUMENT VARIABLE

Lei Wang[1], Jun Shao[2,3] and Fang Fang[2]

[1]*Nankai University,* [2]*East China Normal University and* [3]*University of Wisconsin-Madison*

Abstract: Handling data with nonignorable missing responses is difficult because of the identifiability issue caused by a nonignorable nonresponse. An effective approach described in the literature is to impose a parametric model on the nonresponse propensity (while the conditional distribution of the response, given covariates, is totally unspecified). Then, use a nonresponse instrument, which is a useful covariate vector that can be excluded from the propensity, given the response and other covariates. However, how to find a nonresponse instrument from a given set of covariates is not well addressed. In addition, we may want to select a parametric propensity model from a set of candidate models. Therefore, we propose a simultaneous propensity model and instrument selection criterion. In the presence of a nonignorable nonresponse, the proposed method consistently selects the most compact correct parametric propensity model and instrument from a group of candidate models, assuming one of these candidate models is correct and an instrument exists. Simulation results show that our proposed method works quite well. A real-data example is presented for illustration.

Key words and phrases: Generalized method of moments, identifiability, misspecified model, nonignorable propensity, nonresponse instrument, penalized validation criterion.

1. Introduction

Consider the problem in which a univariate outcome or response Y is subject to a nonresponse and a vector $\boldsymbol{X}$ of covariates is always observed. Here, we wish to estimate or infer unknown quantities in F_Y (i.e., the distribution of Y), or in $F_{Y|\boldsymbol{X}}$ (i.e., the conditional distribution of Y, given $\boldsymbol{X}$). The conditional probability $\Pr(\delta = 1|Y, \boldsymbol{X})$ is called the nonresponse propensity, or simply the propensity, where δ is the indicator of observing Y. When Y can be ex-

Corresponding author: Fang Fang, KLATASDS-MOE and School of Statistics, East China Normal University, 3663 North Zhong-shan Rd., Shanghai, China 200062. E-mail: ffang@sfs.ecnu.edu.cn.

cluded from the propensity $\Pr(\delta = 1|Y, \boldsymbol{X})$ such that the latter is a function of $\boldsymbol{X}$ only, the propensity is ignorable and missing data are at random (Little and Rubin (2002)). In this case, unknown quantities in F_Y or $F_{Y|\boldsymbol{X}}$ can be estimated using $F_{Y|\boldsymbol{X},\delta=1}$ and $F_{\boldsymbol{X}}$ because $F_{Y|\boldsymbol{X}} = F_{Y|\boldsymbol{X},\delta=1}$; see, for example, Robins (1987); Cheng (1994); Robins, Rotnitzky and Zhao (1994); Ibrahim et al. (2005); Kim and Shao (2013), and the references therein. When Y cannot be excluded from $\Pr(\delta = 1|Y, \boldsymbol{X})$, the propensity is nonignorable, and developing a valid estimation method is notoriously challenging. In this case, the population parameters are, in general, not identifiable, and estimates based on an assumption of ignorable nonresponses may have large biases (Fitzmaurice, Molenberghs and Lipsitz (1995); Wang, Shao and Kim (2014)). Thus, methods very different to those for an ignorable propensity have to be applied; see, for example, Scharfstein, Rotnitzkyand Robins (1999); Qin, Leung and Shao (2002); Tang, Little and Raghunathan (2003); Kim and Yu (2011); Xie, Qian and Qu (2011); Wang, Shao and Kim (2014); Tang, Zhao and Zhu (2014); Zhao and Shao (2015); Shao and Wang (2016); Guan and Qin (2017), and the references therein.

When the propensity is nonignorable, the distribution of $(\delta, Y, \boldsymbol{X})$ is typically not identifiable (Robins and Ritov (1997); Wang, Shao and Kim (2014)). Two general and sufficient conditions for the identifiability of the distribution are the following:

$$\begin{aligned} &\Pr(\delta = 1|Y, \boldsymbol{X}) = \pi(Y, \boldsymbol{U}), \quad \boldsymbol{X} = (\boldsymbol{U}, \boldsymbol{Z}), \\ &F_{Y|\boldsymbol{X}} \text{ depends on } \boldsymbol{Z}, \end{aligned} \tag{1.1}$$

and

$$\text{there is a parametric component in either } F_{Y|\boldsymbol{X}} \text{ or } \pi(Y, \boldsymbol{U}). \tag{1.2}$$

Condition (1.1) means that when Y cannot be excluded from the propensity, a sub-vector $\boldsymbol{Z}$ of $\boldsymbol{X}$ can be excluded, and $\boldsymbol{Z}$ is still a useful covariate for Y. Wang, Shao and Kim (2014) refer to such a $\boldsymbol{Z}$ as a nonresponse instrument. Excluding Y or $\boldsymbol{Z}$ simplifies the form of the propensity and enables us to identify it. Although (1.1) and (1.2) are sufficient conditions, either of them missing leads to a nonidentifiable distribution of $(\delta, Y, \boldsymbol{X})$; see Wang, Shao and Kim (2014) for (1.1), and Robins and Ritov (1997) for (1.2).

For condition (1.2), we can impose parametric models on both $\pi(Y, \boldsymbol{U})$ and $F_{Y|\boldsymbol{X}}$; see, for example, Molenberghs and Kenward (2007). Several studies have attempted to derive results under semiparametric models. Tang, Little and Raghunathan (2003) and Zhao and Shao (2015) studied a pseudo-likelihood method under a parametric model on $F_{Y|\boldsymbol{X}}$, but an unspecified propensity $\pi(Y, \boldsymbol{U})$, with a given instrument $\boldsymbol{Z}$. In contrast, as in this study, Wang, Shao and Kim (2014) derived estimators under a parametric model on the propensity $\pi(Y, \boldsymbol{U})$, but allowed an unspecified $F_{Y|\boldsymbol{X}}$; that is, (1.2) is replaced by

$$\pi(Y, \boldsymbol{U}) \text{ follows a parametric model but } F_{Y|\boldsymbol{X}} \text{ is unspecified.} \tag{1.2A}$$

The main technique in Wang, Shao and Kim (2014) is to use a given instrument $\boldsymbol{Z}$ to create sufficient estimating equations to enable the estimation of the parametric propensity $\pi(Y, \boldsymbol{U})$ in (1.2A); once the propensity is estimated, unknown quantities can be estimated using the inverse propensity weighting method. However, two important issues related to this approach have not been studied. The first is how to find an instrument, a sub-vector $\boldsymbol{Z}$ of $\boldsymbol{X}$, that satisfies (1.1). The second is how to select a parametric model for the propensity $\pi(Y, \boldsymbol{U})$. Although many works have examined model selections with ignorable missing responses, to the best of our knowledge, only two examine model selection with nonignorable missing responses. Fang and Shao (2016) and Zhao, Yang and Ning (2018) considered model/variable selection for $F_{Y|\boldsymbol{X}}$ when $F_{Y|\boldsymbol{X}}$ is parametric and $\pi(Y, \boldsymbol{U})$ is unspecified, which is different to (1.2A), which is the focus of our research. Furthermore, they do not consider how to search for an instrument.

This study proposes a method that simultaneously searches for an instrument satisfying (1.1) and selects a parametric model for the propensity from a set of available models. We formulate this search for an instrument and propensity model within a single model selection framework. Our key idea is to construct and compare two estimators of $F_{\boldsymbol{X}}$, the cumulative distribution function of the covariate vector $\boldsymbol{X}$. Because $\boldsymbol{X}$ is always observed, a simple consistent estimator that does not depend on a model and instrument is the empirical cumulative distribution function $\hat{F}_{\boldsymbol{X}}$, based on $\boldsymbol{X}$ data. On the other hand, for a candidate parametric propensity model k on $\pi(Y, \boldsymbol{U})$, with a possible instrument $\boldsymbol{Z}$, we construct an inverse propensity estimator $\hat{F}_k$ of $F_{\boldsymbol{X}}$ using Y data, $\boldsymbol{X}$ data, and the model information in the presence of nonignorable missing Y data. Because only a correct candidate model and a correct instrument can produce a consistent estimator $\hat{F}_k$ close to $\hat{F}_{\boldsymbol{X}}$, we select a model from a group of candidate models and an instrument by minimizing the distance between the two estimators $\hat{F}_{\boldsymbol{X}}$ and $\hat{F}_k$. Because some propensity models may be correct, but overfitted, we add a penalty term in our model selection criterion, following the general principle of the well-known BIC model selection.

When an instrument exists and the group of candidate models contains at least one correct propensity model, our theory shows that, with probability tending to one as the sample size increases to infinity, while the dimension of $\boldsymbol{X}$ remains fixed, the proposed method simultaneously selects the most compact correct parametric propensity model and a correct instrument. Consequently, parameter estimators using the inverse propensity weighting approach based on the selected model and instrument are consistent and asymptotically normal. Simulation studies and a real-data example demonstrate the effectiveness of the proposed method.

2. Methodology and Theory

Under conditions (1.1) and (1.2A), we would like to select sub-vectors $\boldsymbol{Z}$ and $\boldsymbol{U}$ such that $\boldsymbol{X} = (\boldsymbol{U}, \boldsymbol{Z})$, where $\boldsymbol{Z}$ is an instrument and $\pi(Y, \boldsymbol{U})$ is the propensity. Choosing different components of $\boldsymbol{X}$ as $\boldsymbol{Z}$ and $\boldsymbol{U}$ can be viewed as selecting different models. Thus, the instrument and propensity model selection can be combined into a general model selection problem.

To illustrate, consider three-dimensional $\boldsymbol{X} = (X_1, X_2, X_3)$ and $\pi(Y, \boldsymbol{U}, \theta)$, which are logistic in a linear combination of X_j and Y. Then, we have the following seven models:

$$\begin{aligned}
\pi_0(Y, \boldsymbol{U}_0, \theta_0) &= \frac{1}{\{1+\exp(\alpha_0+\gamma_0 Y)\}},\\
\pi_j(Y, \boldsymbol{U}_j, \theta_j) &= \frac{1}{\{1+\exp(\alpha_j+\beta_j X_j+\gamma_j Y)\}},\ j=1,2,3,\\
\pi_4(Y, \boldsymbol{U}_4, \theta_4) &= \frac{1}{\{1+\exp(\alpha_4+\beta_{41} X_1+\beta_{42} X_2+\gamma_4 Y)\}},\\
\pi_5(Y, \boldsymbol{U}_5, \theta_5) &= \frac{1}{\{1+\exp(\alpha_5+\beta_{51} X_1+\beta_{52} X_3+\gamma_5 Y)\}},\\
\pi_6(Y, \boldsymbol{U}_6, \theta_6) &= \frac{1}{\{1+\exp(\alpha_6+\beta_{61} X_2+\beta_{62} X_3+\gamma_6 Y)\}},
\end{aligned} \tag{2.1}$$

where $\boldsymbol{U}_0 = 0$; $\boldsymbol{U}_j = X_j$, for $j = 1, 2, 3$; $\boldsymbol{U}_4 = (X_1, X_2)$; $\boldsymbol{U}_5 = (X_1, X_3)$; and $\boldsymbol{U}_6 = (X_2, X_3)$. The model with $\boldsymbol{U} = \boldsymbol{X}$ is excluded because we assume the existence of an instrument. These seven models correspond to selecting a propensity and an instrument, because if model k is selected, then the selected instrument is $\boldsymbol{Z}_k$, which contains components in $\boldsymbol{X}$ in the propensity, but not in $\boldsymbol{U}_k$. If we need to select between a logistic and another model (e.g., a probit model), then replacing $1/\{1 + \exp(\cdot)\}$ with another function results in an additional seven models, and the total number of models becomes 14. Alternatively, we may want to add a nonlinear term, such as Y^2, to the linear combination of the logistic model, which results in a total of $3 \times 7 = 21$ models, because we may have a Y term only, a Y^2 term only, or both Y and Y^2 terms.

Let K be the total number of candidate models under all combinations of $\boldsymbol{U}$ and $\boldsymbol{Z}$ decompositions, and let

$$\mathcal{M} = \{\pi_k(Y, \boldsymbol{U}_k, \boldsymbol{\theta}_k),\ k = 1, \ldots, K\}$$

be the collection of all K parametric models, where $\boldsymbol{U}_k$ is the vector $\boldsymbol{U}$ under model k, π_k is a known function of $(Y, \boldsymbol{U}_k, \boldsymbol{\theta}_k)$, and $\boldsymbol{\theta}_k$ is an unknown parameter vector with dimension d_k under model k. If model k is selected, then $\boldsymbol{Z}_k$ with $\boldsymbol{X} = (\boldsymbol{U}_k, \boldsymbol{Z}_k)$ is selected as an instrument, and model $\pi_k(Y, \boldsymbol{U}_k, \boldsymbol{\theta}_k)$ is the selected propensity model. We say that model k is correct if and only if $\boldsymbol{Z}_k$ is an instrument satisfying (1.1) and $\pi_k(Y, \boldsymbol{U}_k, \boldsymbol{\theta}_k)$ is a correct propensity. Under this framework, finding an instrument and selecting a propensity model is the same as selecting a model from $\mathcal{M}$.

For simplicity, we now consider a fixed model k; note that we omit the subscript k in $\boldsymbol{U}$ and $\boldsymbol{Z}$ in the following discussion. Let $\boldsymbol{Z} = (\boldsymbol{Z}_c, \boldsymbol{Z}_d)$, where $\boldsymbol{Z}_c$ is a continuous covariate vector, and $\boldsymbol{Z}_d$ is a J_k-dimensional vector in which the jth component is the indicator of a discrete covariate, for $j = 1, \ldots, J_k$. Following Wang, Shao and Kim (2014), in order to estimate the parameter $\boldsymbol{\theta}_k$, we define the vector-valued function

$$\boldsymbol{g}_k(Y, \boldsymbol{X}, \delta, \boldsymbol{\theta}_k) = \boldsymbol{h}_k(\boldsymbol{X})\left\{\frac{\delta}{\pi_k(Y, \boldsymbol{U}, \boldsymbol{\theta}_k)} - 1\right\}, \tag{2.2}$$

where $\boldsymbol{h}_k(\boldsymbol{X})$ is a known vector-valued function of $\boldsymbol{X}$ with dimension $L_k \geq d_k$, which is the dimension of $\boldsymbol{\theta}_k$. For example, we can use $\boldsymbol{h}_k(\boldsymbol{X}) = (\boldsymbol{U}, \boldsymbol{Z}_c, \boldsymbol{Z}_d)$ when the dimension of $(\boldsymbol{U}, \boldsymbol{Z}_c)$ plus J_k is greater than or equal to d_k. If the dimension of $(\boldsymbol{U}, \boldsymbol{Z}_c)$ plus J_k is smaller than d_k, we add $\tilde{\boldsymbol{Z}}$ to $(\boldsymbol{U}, \boldsymbol{Z}_c, \boldsymbol{Z}_d)$, where the components of $\tilde{\boldsymbol{Z}}$ are higher moments of $\boldsymbol{Z}_c$, such that the dimension of $(\boldsymbol{U}, \boldsymbol{Z}_c, \boldsymbol{Z}_d, \tilde{\boldsymbol{Z}})$ is not smaller than d_k. The efficiency of the estimation based on (2.2) depends on the choice of $\boldsymbol{h}_k(\boldsymbol{X})$. Several approaches for choosing $\boldsymbol{h}_k(\boldsymbol{X})$ have been proposed by Morikawa, Kim and Kano (2017) and Ai, Linton and Zhang (2020). However, because we focus on model and instrument selection, we assume a fixed function $\boldsymbol{h}_k(\boldsymbol{X})$ in (2.2).

If model k is correct and $\boldsymbol{\theta}_k^0$ is the unique true parameter value of $\boldsymbol{\theta}_k$, then it can be verified that, under $\Pr(\delta = 1|Y, \boldsymbol{X}) = \pi(Y, \boldsymbol{U})$ and the first part of condition (1.1),

$$E\{\boldsymbol{g}_k(Y, \boldsymbol{X}, \delta, \boldsymbol{\theta}_k^0)\} = 0. \tag{2.3}$$

Thus, the function $\boldsymbol{g}_k$ in (2.2) provides an estimating equation for $\boldsymbol{\theta}_k$. The second part of condition (1.1) ensures that the estimation equations in (2.3) are not linearly dependent; thus, we have sufficient equations to estimate $\boldsymbol{\theta}_k$.

Throughout, model selection is based on a random sample of size n, $(\boldsymbol{X}_i, Y_i, \delta_i)$, for $i = 1, \ldots, n$, taken from the distribution of $(\boldsymbol{X}, Y, \delta)$, where $\boldsymbol{X}_i$ is always observed and Y_i is observed if and only if $\delta_i = 1$. Because L_k may be larger than d_k, we apply the generalized method of moments (GMM) to estimate $\boldsymbol{\theta}_k$, based on (2.2)–(2.3). Specifically, let $\bar{\boldsymbol{G}}_{kn}(\boldsymbol{\theta}_k) = n^{-1}\sum_{i=1}^n \boldsymbol{g}_k(Y_i, \boldsymbol{X}_i, \delta_i, \boldsymbol{\theta}_k)$, $\tilde{\boldsymbol{\theta}}_k = \operatorname{argmin}_{\boldsymbol{\theta}_k} \bar{\boldsymbol{G}}_{kn}(\boldsymbol{\theta}_k)^\top \bar{\boldsymbol{G}}_{kn}(\boldsymbol{\theta}_k)$, where $a^\top$ is the transpose of a, and let $\hat{\boldsymbol{W}}_{kn} = n^{-1}\sum_{i=1}^n \boldsymbol{g}_k(Y_i, \boldsymbol{X}_i, \delta_i, \tilde{\boldsymbol{\theta}}_k)\boldsymbol{g}_k(Y_i, \boldsymbol{X}_i, \delta_i, \tilde{\boldsymbol{\theta}}_k)^\top$. Then, the GMM estimator of $\boldsymbol{\theta}_k$ is

$$\hat{\boldsymbol{\theta}}_k = \operatorname*{argmin}_{\boldsymbol{\theta}_k} \bar{\boldsymbol{G}}_{kn}(\boldsymbol{\theta}_k)^\top \hat{\boldsymbol{W}}_{kn}^{-1} \bar{\boldsymbol{G}}_{kn}(\boldsymbol{\theta}_k). \tag{2.4}$$

For simplicity, we denote the cumulative distribution function of $\boldsymbol{X}$ by $F = F_{\boldsymbol{X}}$. Once we have $\hat{\boldsymbol{\theta}}_k$ in (2.4), an inverse propensity weighting estimator of $F(\boldsymbol{x})$ is given by

$$\hat{F}_k(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^n \frac{\delta_i \mathrm{I}(\boldsymbol{X}_i \le \boldsymbol{x})}{\pi_k(Y_i, \boldsymbol{U}_i, \hat{\boldsymbol{\theta}}_k)},$$

where $\mathrm{I}(\boldsymbol{X}_i \le \boldsymbol{x})$ is the indicator function of $\boldsymbol{X}_i \le \boldsymbol{x}$ and, for vectors $\boldsymbol{a}$ and $\boldsymbol{b}$, $\boldsymbol{a} \le \boldsymbol{b}$ means that all components of $\boldsymbol{b} - \boldsymbol{a}$ are nonnegative.

If model k is correct, then it can be shown that, as the sample size $n \to \infty$, $\hat{\boldsymbol{\theta}}_k$ is consistent for $\boldsymbol{\theta}_k^0$ and $\hat{F}_k(\boldsymbol{x})$ is consistent for $F(\boldsymbol{x})$. On the other hand, if either $\boldsymbol{Z}$ is not an instrument or $\pi_k(Y, \boldsymbol{U}, \boldsymbol{\theta}_k)$ is incorrect, then $\hat{F}_k(\boldsymbol{x})$ is inconsistent.

Without using a model, a consistent estimator of $F(\boldsymbol{x})$ is the empirical cumulative distribution function $\hat{F}(\boldsymbol{x}) = n^{-1}\sum_{i=1}^n \mathrm{I}(\boldsymbol{X}_i \le \boldsymbol{x})$. We then use the closeness between $\hat{F}$ and $\hat{F}_k$ to validate model k. Define the following model validation criterion:

$$\mathrm{VC}(k) = \frac{1}{n}\sum_{i=1}^n |\hat{F}_k(\boldsymbol{X}_i) - \hat{F}(\boldsymbol{X}_i)|. \tag{2.5}$$

As we show later, if model $\pi_k(Y, \boldsymbol{U}, \boldsymbol{\theta}_k)$ with the corresponding instrument is correct, then $\mathrm{VC}(k) \to 0$ in probability as $n \to \infty$. Otherwise, $\mathrm{VC}(k)$ does not converge to zero. Thus, correct and incorrect models can be detected using

VC(k).

A correct model may be an overfitted model that includes some redundant parameters. For example, suppose $\boldsymbol{X} = (\boldsymbol{S}, \boldsymbol{R}, \boldsymbol{T})$, $(\boldsymbol{R}, \boldsymbol{T})$ is an instrument and $\pi(\alpha + \gamma Y + \boldsymbol{\beta}^\top \boldsymbol{S})$ is a correct propensity, where α, γ, and $\boldsymbol{\beta}$ are unknown. Then, $\boldsymbol{T}$ is also an instrument and $\pi(\alpha + \gamma Y + \boldsymbol{\beta}^\top \boldsymbol{S} + \mathbf{0}^\top \boldsymbol{R})$ is a correct propensity containing a redundant $\boldsymbol{R}$, where $\mathbf{0}$ is a vector of zeros. A more compact propensity model may result in a propensity and other parameter estimators that are more efficient (see the simulation results in Section 3). Thus, we define the best model as the most compact correct propensity model in $\mathcal{M}$, and penalize the model dimension, following the well-known BIC; that is, we choose a model by minimizing the following penalized validation criterion (PVC):

$$\begin{aligned} \mathrm{PVC}_\lambda(k) &= \mathrm{VC}(k) + \lambda \log(d_k), \\ \hat{k} &= \operatorname*{argmin}_{1 \le k \le K} \mathrm{PVC}_\lambda(k), \end{aligned} \tag{2.6}$$

where d_k is the dimension of $\boldsymbol{\theta}_k$, and $\lambda \geq 0$ is a penalization factor that may depend on n and the sample data. The selected instrument is $\boldsymbol{Z}_{\hat{k}}$ with $\boldsymbol{X} = (\boldsymbol{U}_{\hat{k}}, \boldsymbol{Z}_{\hat{k}})$, and the selected model is $\pi_{\hat{k}}(Y, \boldsymbol{U}_{\hat{k}}, \boldsymbol{\theta}_{\hat{k}})$. Quantities of interest can be estimated using the inverse propensity weighting with the estimated propensity $\pi_{\hat{k}}(Y, \boldsymbol{U}_{\hat{k}}, \hat{\boldsymbol{\theta}}_{\hat{k}})$. For example, the population mean $\mu = E(Y)$ can be estimated by

$$\hat{\mu}_{pvc} = \frac{1}{n} \sum_{i=1}^{n} \frac{\delta_i Y_i}{\pi_{\hat{k}}(Y_i, \boldsymbol{U}_{\hat{k}i}, \hat{\boldsymbol{\theta}}_{\hat{k}})}. \tag{2.7}$$

We now present several asymptotic properties of the proposed method for instrument and model selection. If $\boldsymbol{Z}$ is an instrument and $\pi_k(Y, \boldsymbol{U}, \boldsymbol{\theta}_k)$ is correct, as shown in Wang, Shao and Kim (2014), $\hat{\boldsymbol{\theta}}_k$ obtained by (2.4) is consistent for $\boldsymbol{\theta}_k^0$, and is asymptotically normal under some regularity conditions. When either $\boldsymbol{Z}$ or $\pi_k(Y, \boldsymbol{U}, \boldsymbol{\theta}_k)$ is incorrect, the following lemma shows the property of $\hat{\boldsymbol{\theta}}_k$ under a misspecified model.

Lemma 1. *Assume the following regularity conditions:*

$C1$. (a) The dimension of $\boldsymbol{X}$, p, and the number of candidate models, K, remain fixed when the sample size $n \to \infty$; (b) The parameter space $\mathcal{A}$ for $\boldsymbol{\theta}_k$ is a compact set of $\mathcal{R}^{d_k}$, and $\boldsymbol{\theta}_k^*$ is the unique minimizer of $\|\boldsymbol{G}_k(\boldsymbol{\theta}_k)\|$ over $\boldsymbol{\theta}_k$, where $\boldsymbol{G}_k(\boldsymbol{\theta}_k) = E\{\boldsymbol{g}_k(Y, \boldsymbol{X}, \delta, \boldsymbol{\theta}_k)\}$ and $\| \cdot \|$ is the l_2-norm; (c) $\sup_{\boldsymbol{\theta}_k} \|\boldsymbol{g}_k(Y, \boldsymbol{X}, \delta, \boldsymbol{\theta}_k)\| < \infty$; (d) The matrix $\boldsymbol{\Gamma}_k(\boldsymbol{\theta}_k^*) = E\{\boldsymbol{h}_k(\boldsymbol{X})^\top \delta[\partial \pi_k^{-1}$ $(Y, \boldsymbol{U}, \boldsymbol{\theta}_k^*)/\partial \boldsymbol{\theta}_k\}$ is of full rank, and the matrix $\mathbf{W}_k(\boldsymbol{\theta}_k^*) = E\{\boldsymbol{g}_k(Y, \boldsymbol{X}, \delta, \boldsymbol{\theta}_k^*)$

$\boldsymbol{g}_k(Y, \boldsymbol{X}, \delta, \boldsymbol{\theta}_k^*)^\top\}$ is positive definite;

C2. (a) $\pi_k(Y, \boldsymbol{U}, \boldsymbol{\theta}_k)$ is twice differentiable with respect to $\boldsymbol{\theta}_k$; (b) $\pi_k(Y, \boldsymbol{U}, \boldsymbol{\theta}_k^*) \geq C > 0$, for $k = 1, \ldots, K$; (c) $\partial \pi_k(Y, \boldsymbol{U}, \boldsymbol{\theta}_k)/\partial \boldsymbol{\theta}_k$ is uniformly bounded.

Then, as $n \to \infty$,

$$n^{1/2}(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*) \to N(0, \{\boldsymbol{\Gamma}_k(\boldsymbol{\theta}_k^*)^\top \boldsymbol{W}_k^{-1}(\boldsymbol{\theta}_k^*)\boldsymbol{\Gamma}_k(\boldsymbol{\theta}_k^*)\}^{-1}) \text{ in distribution.}$$

In the presence of misspecification, the proposed $\hat{\boldsymbol{\theta}}_k$ consistently estimates $\boldsymbol{\theta}_k^*$ by minimizing the population version of the empirical generalized moment discrepancy. If model $\pi_k(Y, \boldsymbol{U}, \boldsymbol{\theta}_k)$ is correct, then $\hat{\boldsymbol{\theta}}_k$ is consistent for the true parameter vector $\boldsymbol{\theta}_k^0$.

Define

$$F_k(\boldsymbol{x}) = E\left\{\frac{\delta \mathrm{I}(\boldsymbol{X} \leq \boldsymbol{x})}{\pi_k(Y, \boldsymbol{U}, \boldsymbol{\theta}_k^*)}\right\} = E\left[E\left\{\frac{\pi(Y, \boldsymbol{U})}{\pi_k(Y, \boldsymbol{U}, \boldsymbol{\theta}_k^*)}\right\}\mathrm{I}(\boldsymbol{X} \leq \boldsymbol{x})\right].$$

Because $\hat{\boldsymbol{\theta}}_k \to \boldsymbol{\theta}_k^*$ in probability, it can be verified that $\hat{F}_k(\boldsymbol{x}) \to F_k(\boldsymbol{x})$ in probability. Define

$$\Delta_k = E|F_k(\boldsymbol{X}) - F(\boldsymbol{X})|.$$

Then, VC(k), defined in (2.5), converges in probability to Δ_k. If $\pi_k(Y, \boldsymbol{U}, \boldsymbol{\theta}_k)$ is a correct model and $\boldsymbol{Z}$ is an instrument, then $F_k(\boldsymbol{x}) = F(\boldsymbol{x})$ and $\Delta_k = 0$. If $\Delta_k > 0$ for any incorrect model k, then we can distinguish between an incorrect and a correct model.

The order of λ tending to 0 as $n \to \infty$ determines the asymptotic behavior of the proposed model selection procedure. Without loss of generality, we assume that the most compact correct model is $\pi_1(Y, \boldsymbol{U}, \boldsymbol{\theta}_1)$. The model selection procedure is consistent as $n \to \infty$ if and only if

$$\Pr\{\mathrm{PVC}_\lambda(k) > \mathrm{PVC}_\lambda(1)\} \to 1 \tag{2.8}$$

holds for any model k with $k > 1$. Suppose that $\Delta_k > 0$ when $\pi_k(Y, \boldsymbol{U}_k, \boldsymbol{\theta}_k)$ is an incorrect model. To achieve (2.8), we need λ satisfying $\lambda(\log d_1 - \log d_k) < \mathrm{VC}(k) - \mathrm{VC}(1)$; because $\mathrm{VC}(k) - \mathrm{VC}(1) \to \Delta_k > 0$ and d_k may be smaller than d_1, we need $\lambda \to 0$ as $n \to \infty$. Next, let $\pi_k(Y, \boldsymbol{U}_k, \boldsymbol{\theta}_k)$ be a correct model that is overfitted, such that $d_k > d_1$. In this case, we need to find λ such that $\lambda > \{\mathrm{VC}(1) - \mathrm{VC}(k)\}/(\log d_k - \log d_1)$ with probability tending to one. Because both VC(1) and VC(k) converge to zero under correct models, we need to choose λ that converges to zero at a rate slower than that of VC(1) − VC(k).

The following lemma, proved in the Appendix, provides the convergence rate of $\mathrm{VC}(1) - \mathrm{VC}(k)$.

Lemma 2. *Under the conditions in Lemma* 1*, if* $\pi_1(Y, \boldsymbol{U}, \boldsymbol{\theta}_1)$ *is the most compact correct model and* $\pi_k(Y, \boldsymbol{U}, \boldsymbol{\theta}_k)$ *is an overfitted correct model, then* $VC(1) - VC(k) = O_p(n^{-1/2})$.

This result and the previous discussion establish the following result about the consistency of the proposed propensity and instrument selection method.

Theorem 1. *Assume that* $\mathcal{M}$ *contains a correctly specified propensity model for* $\pi(Y, \boldsymbol{U})$*, with an instrument* $\boldsymbol{Z}$ *satisfying* (1.1)*. Under the regularity conditions in Lemmas* 1 *and* 2*, if* λ *in* (2.6) *is chosen such that* $\lambda \to 0$ *and* $n^{1/2}\lambda \to \infty$*, then* (2.8) *holds; that is,* $\Pr(\hat{k} = 1) \to 1$ *as* $n \to \infty$*, where model* 1 *is assumed to be the most compact correct model.*

In practice, we propose using $\lambda = Cn^{-1/2}(\log\log n)^{1/2}$, with a constant C chosen using cross-validation (CV). Specifically, we randomly split the set $\{1, \ldots, n\}$ into J nonoverlapping subsets $\{S_1, \ldots, S_J\}$ of roughly equal size, $n_1, \ldots, n_J$. For each $j = 1, \ldots, J$ and a given C, using all data from $i \notin S_j$, we compute

$$\mathrm{PVC}_{-j}(k) = (n - n_j)^{-1} \sum_{i \notin S_j} |\hat{F}_k(\boldsymbol{X}_i) - \hat{F}(\boldsymbol{X}_i)| + \lambda \log(d_k),$$
$$\hat{k}_{-j} = \underset{1 \le k \le K}{\operatorname{argmin}}\, \mathrm{PVC}_{-j}(k).$$

For a fixed C, we compute the error on the validation set S_j as

$$e_j(C) = \frac{1}{n_j} \sum_{i \in S_j} |\hat{F}_{\hat{k}_{-j}}(\boldsymbol{X}_i) - \hat{F}(\boldsymbol{X}_i)|,$$

and then choose the value C as $\hat{C}$ that minimizes the average error over all subsets; that is,

$$\hat{C} = \underset{C}{\operatorname{argmin}} \frac{1}{J} \sum_{j=1}^{J} e_j(C). \tag{2.9}$$

The collection $\mathcal{M}$ may include all possible decompositions of $\boldsymbol{X} = (\boldsymbol{U}, \boldsymbol{Z})$, which means we have at least $K = 2^p - 1$ models when the dimension of $\boldsymbol{X}$ is p. The grid search over all models may be computationally infeasible for a moderate p, for example $p \ge 7$. For the purpose of searching for a correct propensity and an instrument, however, it is not necessary to perform a grid search. Here, we propose a forward instrument selection procedure that can handle a moderate p. Consider the p-dimensional covariates $\boldsymbol{X} = (X_1, \ldots, X_p)$ and a fixed parametric function $\pi(Y, \boldsymbol{U}, \theta)$ (e.g., a logistic function). Then, we proceed as follows:

(i) Start with p models with $\boldsymbol{Z} = X_j$ and $\boldsymbol{U} = (X_t : t \neq j)$, for $j = 1, \ldots, p$. Select the model with the lowest PVC, yielding $\boldsymbol{Z}_1^* = X_{1^*}$ and $\boldsymbol{U}_1^* = (X_t : t \neq 1^*)$.

(ii) Consider the next $p-1$ models with $\boldsymbol{Z} = (X_{1^*}, X_j)$ and $\boldsymbol{U} = (X_t : t \neq j, t \neq 1^*)$, for $j = 1, \ldots, p$, $j \neq 1^*$. For these, select the model with the lowest PVC. If this PVC value is higher than that in step 1, then stop, and the model selected is $\pi(Y, \boldsymbol{U}_1^*, \theta)$. Otherwise, set $\boldsymbol{Z}_2^* = (X_{1^*}, X_{2^*})$ and $\boldsymbol{U}_2^* = (X_t : t \neq 1^*, t \neq 2^*)$, and continue to the next step.

(iii) At the kth step, consider $p-k+1$ models with $\boldsymbol{Z} = (X_{1^*}, \ldots, X_{(k-1)^*}, X_j)$ and $\boldsymbol{U} = (X_t : t \neq j, t \neq 1^*, \ldots, t \neq (k-1)^*)$, for $j = 1, \ldots, p$, $j \neq 1^*, \ldots, j \neq (k-1)^*$. For these, select the model with the lowest PVC. If this PVC value is higher than that in step $k-1$, then stop, and the model selected is $\pi(Y, \boldsymbol{U}_{(k-1)^*}, \theta)$. Otherwise, continue until $k = p$.

The number of models considered in this procedure is at most $p + (p-1) + \cdots + 2 + 1 = p(p+1)/2$. Furthermore, if we want select $\pi(Y, \boldsymbol{U}, \theta)$ between logistic and probit models or add a nonlinear term Y^2 to the linear combination of X_j and Y, we can apply the previous idea and establish a similar multi-step procedure. An asymptotic result similar to that in Theorem 1 can also be established.

3. Simulation Studies

Under assumptions (1.1) and (1.2A), we use a simulation to examine the finite-sample performance of the proposed method in terms of the rate of selecting the most compact correct model. We also examine the bias and root mean squared error (RMSE) of the resulting inverse propensity weighting estimator $\hat{\mu}_{pvc}$, defined in (2.7). All results are based on 1,000 simulation replications.

In simulation 1, we select a model from the seven models in (2.1). Here $\boldsymbol{X} = (X_1, X_2, X_3)$ is generated from a three-dimensional normal distribution with mean one and covariance $\mathrm{Cov}(X_j, X_{j'}) = 0.5$, for $1 \leq j < j' \leq 3$, and $\mathrm{Var}(X_j) = 1$ and $Y = X_1^2 + X_2^2 + X_3^2 + \varepsilon$, where ε is drawn from $N(0, 2)$, and is independent of $\boldsymbol{X}$. For convenience, we denote the seven models in (2.1) by M_0, $M_1(X_1)$, $M_1(X_2)$, $M_1(X_3)$, $M_2(X_1, X_2)$, $M_2(X_1, X_3)$, and $M_2(X_2, X_3)$, respectively, where the subscript s in $M_s(\boldsymbol{U})$ is the dimension of $\boldsymbol{U}$; for example, M_0 is the model with $\boldsymbol{U} = 0$ and $\boldsymbol{Z} = (X_1, X_2, X_3)$, and $M_2(X_1, X_3)$ is the model with a two-dimensional $\boldsymbol{U} = (X_1, X_3)$ and $\boldsymbol{Z} = X_2$.

Given $(Y, \boldsymbol{X})$, we generate δ from a Bernoulli distribution using the logistic function in (2.1) as the probability and the parameter vector $\boldsymbol{\theta}^0 = (-0.4, -0.3)$ for M_0; in addition, $\boldsymbol{\theta}^0 = (-0.8, 1.2, -0.3)$ for $M_1(X_j)$, with $j = 1$, 2, 3, and $\boldsymbol{\theta}^0 = (-0.8, 1.2, 1.2, -0.3)$ for $M_2(X_j, X_{j'})$, with $1 \leq j < j' \leq 3$. The coefficients in the propensity models are chosen such that the unconditional rates of missing data are between 20% and 40%. As in Wang, Shao and Kim (2014), we use $\boldsymbol{h}_k(\boldsymbol{X}) = (1, \boldsymbol{U}, \boldsymbol{Z})$ in (2.2) and (2.4) to obtain the GMM estimator $\hat{\boldsymbol{\theta}}_k$.

If the true propensity $\pi(Y, \boldsymbol{U}) = M_1(X_1)$, then $\boldsymbol{Z} = (X_2, X_3)$ is an instrument; models M_0, $M_1(X_2)$, $M_1(X_3)$, and $M_2(X_2, X_3)$ are incorrect; $M_2(X_1, X_2)$ and $M_2(X_1, X_3)$ are also correct propensity models, with $\boldsymbol{Z} = X_3$ and $\boldsymbol{Z} = X_2$ as instruments, respectively. Because both $M_2(X_1, X_2)$ and $M_2(X_1, X_3)$ are overfitted, the penalty term in (2.6) forces us to choose $M_1(X_1)$ more frequently. The discussion is similar if $M_1(X_2)$ or $M_1(X_3)$ is correct. If the true propensity $\pi(Y, \boldsymbol{U}) = M_2(X_1, X_2)$, then $M_2(X_1, X_2)$ is the only correct propensity model, and $\boldsymbol{Z} = X_3$ is the only correct instrument. Finally, if $\pi(Y, \boldsymbol{U}) = M_0$, then all models are correct, and M_0 is the most compact model with $\boldsymbol{Z} = \boldsymbol{X}$.

For n= 300, 500, and 1,000, we implement the PVC in (2.6), using a 10-fold CV method to determine the tuning parameter λ (see the end of Section 2), where the range for the minimization in (2.9) is $(0.1, 20)$. Table 1 reports the rates for 1,000 Monte Carlo replications, in which each model is selected using the proposed PVC under different best models (i.e., the most compact correct models). The results show that the proposed method selects the best model most of the time; that is, the simulation rates when selecting the best model are very high when the sample size $n = 300$, and are close to one when $n = 500$ or 1,000.

Following the model and instrument selection, we can estimate $\mu = E(Y)$, using the proposed estimator $\hat{\mu}_{pvc}$, defined in (2.7), based on the selected and estimated propensity. By way of comparison, we also include three other estimators: $\bar{Y} = n^{-1}\sum_{i=1}^n Y_i$, the sample mean when there is no missing data, which is used as a benchmark; $\hat{\mu}_{cc} = \sum_{i=1}^n \delta_i Y_i / \sum_{i=1}^n \delta_i$, the sample mean of observed Y data, which is a biased estimator; and the inverse propensity weighting estimators $\hat{\mu}_k = \sum_{i=1}^n \delta_i Y_i / \pi_k(Y_i, \boldsymbol{U}_i, \hat{\boldsymbol{\theta}}_k)$, for $k = 0, \ldots, 6$, which differs from $\hat{\mu}_{pvc}$ in (2.7) because $\hat{\mu}_k$ uses a fixed propensity without model selection, and may be biased when the propensity model is incorrect. The mean, μ, is 6 in all cases.

Owing to symmetry, the simulation results when $M_1(X_2)$ or $M_1(X_3)$ are best

Table 1. Simulated probability (÷1000) of selecting each model in simulation 1.

n	Best model	Selected model						
		M_0	$M_1(X_1)$	$M_1(X_2)$	$M_1(X_3)$	$M_2(X_1,X_2)$	$M_2(X_1,X_3)$	$M_2(X_2,X_3)$
300	M_0	996	2	1	1	0	0	0
	$M_1(X_1)$	33	945	1	0	11	10	0
	$M_1(X_2)$	29	0	956	0	9	0	6
	$M_1(X_3)$	37	1	0	942	0	11	9
	$M_2(X_1,X_2)$	0	22	20	0	958	0	0
	$M_2(X_1,X_3)$	0	23	0	19	0	959	1
	$M_2(X_2,X_3)$	0	0	18	22	0	0	960
500	M_0	996	1	2	1	0	0	0
	$M_1(X_1)$	1	975	0	0	11	13	0
	$M_1(X_2)$	0	0	980	0	15	0	5
	$M_1(X_3)$	0	0	0	976	0	9	15
	$M_2(X_1,X_2)$	0	1	2	0	997	0	0
	$M_2(X_1,X_3)$	0	3	0	2	0	995	0
	$M_2(X_2,X_3)$	0	0	4	3	0	0	993
1,000	M_0	1,000	0	0	0	0	0	0
	$M_1(X_1)$	0	988	0	0	4	8	0
	$M_1(X_2)$	0	0	987	0	7	0	6
	$M_1(X_3)$	0	0	0	989	0	8	3
	$M_2(X_1,X_2)$	0	0	0	0	1,000	0	0
	$M_2(X_1,X_3)$	0	0	0	0	0	1,000	0
	$M_2(X_2,X_3)$	0	0	0	1	0	0	999

are similar to those when $M_1(X_1)$ is best, and the results when $M_2(X_1,X_3)$ or $M_2(X_2,X_3)$ are best are similar to those when $M_2(X_1,X_2)$ is best. Hence, we present only those results when M_0, $M_1(X_1)$ and $M_2(X_1,X_2)$ are best. Table 2 shows the biases and RMSEs of the point estimators based on different methods. In terms of bias and RMSE, when M_0 is best, we find that the proposed PVC estimator and the estimator based on the seven propensity models are comparable. When $M_1(X_1)$ is the best, it can be seen that the proposed PVC estimator and the estimator based on $M_1(X_1)$ are comparable, both of which exhibit negligible bias and a slightly larger RMSE than that of $\bar{Y}$ in all cases; as expected, the estimators based on $M_1(X_1,X_2)$ and $M_1(X_1,X_3)$ are also unbiased, but less efficient. Lastly, the estimators based on observed Y values and other propensity models have larger biases and RMSEs, supporting our theory. Similar results are obtained when $M_2(X_1,X_2)$ is best. In this case, because only $M_2(X_1,X_2)$ is correct, we find that $\hat{\mu}_k$ based on the incorrect models have much larger bi-

Table 2. Simulated bias and RMSE when estimating $E(Y)$ in simulation 1.

Best model	Method	$n = 300$		$n = 500$		$n = 1{,}000$	
		Bias	RMSE	Bias	RMSE	Bias	RMSE
M_0	PVC	-0.022	0.357	-0.035	0.273	0.002	0.199
	FULL	-0.001	0.337	-0.022	0.261	0.006	0.188
	CC	0.815	0.899	0.785	0.837	0.814	0.841
	M_0	-0.021	0.357	-0.034	0.272	0.002	0.199
	$M_1(X_1)$	-0.017	0.376	-0.032	0.283	0.006	0.204
	$M_1(X_2)$	-0.008	0.380	-0.032	0.285	0.005	0.204
	$M_1(X_3)$	-0.020	0.376	-0.029	0.280	0.002	0.205
	$M_2(X_1, X_2)$	0.005	0.429	-0.032	0.319	0.010	0.224
	$M_2(X_1, X_3)$	-0.013	0.423	-0.027	0.322	0.004	0.226
	$M_2(X_2, X_3)$	-0.002	0.438	-0.024	0.322	0.002	0.226
$M_1(X_1)$	PVC	-0.004	0.368	-0.028	0.279	-0.006	0.206
	FULL	-0.001	0.337	-0.025	0.261	0.001	0.192
	CC	0.789	0.893	0.766	0.831	0.795	0.829
	M_0	0.096	0.382	0.090	0.296	0.083	0.231
	$M_1(X_1)$	-0.007	0.368	-0.027	0.279	-0.006	0.207
	$M_1(X_2)$	0.502	0.882	0.450	0.710	0.447	0.665
	$M_1(X_3)$	0.503	0.868	0.439	0.681	0.476	0.700
	$M_2(X_1, X_2)$	0.005	0.517	-0.014	0.443	-0.010	0.247
	$M_2(X_1, X_3)$	0.008	0.497	-0.035	0.360	-0.011	0.243
	$M_2(X_2, X_3)$	1.816	2.056	1.924	2.115	2.127	2.334
$M_2(X_1, X_2)$	PVC	0.018	0.453	-0.030	0.356	-0.008	0.251
	FULL	-0.001	0.337	-0.025	0.261	0.001	0.192
	CC	0.309	0.618	0.253	0.469	0.291	0.414
	M_0	0.611	0.923	0.626	0.840	0.692	0.804
	$M_1(X_1)$	0.458	0.783	0.477	0.668	0.507	0.585
	$M_1(X_2)$	0.478	0.884	0.465	0.648	0.501	0.592
	$M_1(X_3)$	2.633	2.826	2.756	2.912	2.922	3.072
	$M_2(X_1, X_2)$	-0.012	0.476	-0.033	0.364	-0.008	0.251
	$M_2(X_1, X_3)$	2.807	3.088	2.941	3.205	3.139	3.371
	$M_2(X_2, X_3)$	2.922	3.225	3.026	3.297	3.039	3.244

ases and RMSEs than those of the method based on observed Y values, which is consistent with the findings of Shao and Wang (2016), and is our motivation for studying model and instrument selection.

Simulation 2 evaluates the performance of the proposed method in selecting Y or Y^2 in the logistic propensity model; that is, in addition to the seven candidate models in (2.1), we include the following seven candidate models:

$$\begin{aligned}
\tilde{M}_0 &= \frac{1}{\{1+\exp(\alpha+\gamma Y^2)\}}, \ \boldsymbol{\theta}=(\alpha,\gamma)\\
\tilde{M}_1(X_j) &= \frac{1}{\{1+\exp(\alpha+\beta X_j+\gamma Y^2)\}}, \ \boldsymbol{\theta}=(\alpha,\beta,\gamma)\\
\tilde{M}_2(X_j,X_{j'}) &= \frac{1}{\{1+\exp(\alpha+\beta_1 X_j+\beta_2 X_{j'}+\gamma Y^2)\}}, \ \boldsymbol{\theta}=(\alpha,\beta_1,\beta_2,\gamma),
\end{aligned}$$

with $\boldsymbol{\theta}^0 = (0.8, -0.1)$ for $\tilde{M}_0$, $\boldsymbol{\theta}^0 = (-0.8, 2, -0.3)$ for $\tilde{M}_1(X_j)$, and $\boldsymbol{\theta}^0 = (-0.8, 1.5, 1.5, -0.1)$ for $\tilde{M}_2(X_j, X_{j'})$ in the simulation, for $1 \le j \le j' \le 3$.

If the true propensity $\pi(Y, \boldsymbol{U}) = M_0$, $\boldsymbol{X}$ is an instrument and $\boldsymbol{Z}$ = subvectors of $\boldsymbol{X}$ under all other models are correct instruments, even though M_0 is the most compact model. Furthermore, models $M_1(X_j)$ and $M_2(X_j, X_{j'})$ are correct, but $\tilde{M}_0$, $\tilde{M}_1(X_j)$ and $\tilde{M}_2(X_j, X_{j'})$ are incorrect, owing to their use of Y^2 instead of Y; the discussion is similar if $\tilde{M}_0$ is correct. If $\pi(Y, \boldsymbol{U}) = M_1(X_1)$, only models $M_1(X_1)$, $M_2(X_1, X_2)$, $M_2(X_1, X_3)$, $\tilde{M}_1(X_1)$, $\tilde{M}_2(X_1, X_2)$, and $\tilde{M}_2(X_1, X_3)$ give correct instruments $\boldsymbol{Z} = X_2$, X_3, or (X_2, X_3). However, $\tilde{M}_1(X_1)$, $\tilde{M}_2(X_1, X_2)$, and $\tilde{M}_2(X_1, X_3)$ are incorrect models, and $M_2(X_1, X_2)$ and $M_2(X_1, X_3)$ are correct, but overfitted. The discussion is similar if $M_1(X_j)$ or $\tilde{M}_1(X_j)$ is correct. If $\pi(Y, \boldsymbol{U}) = M_2(X_1, X_2)$, only $M_2(X_1, X_2)$ and $\tilde{M}_2(X_1, X_2)$ give correct instrument $\boldsymbol{Z} = X_3$, but $\tilde{M}_2(X_1, X_2)$ is incorrect. The discussion is similar if $M_2(X_j, X_{j'})$ or $\tilde{M}_2(X_j, X_{j'})$ is correct.

The model selection probabilities and estimation results for $\mu = E(Y)$ are shown in Figure 1 and Table 3, respectively. Figure 1 shows that our proposed method performs well in terms of selecting the best propensity model and instrument simultaneously. The results in Table 1 show that the selection rates for the best model decrease slightly when M_0 or $M_1(X_j)$ is best, but are close to one when $n = 1{,}000$.

For the seven propensity models using Y, the results are symmetric; thus, we present only those results when M_0, $M_1(X_1)$, or $M_2(X_1, X_2)$ are best. Table 3 shows the biases and RMSEs of the point estimators based on different methods. The results show that when M_0 is best, the proposed PVC estimator and the estimator based on the seven propensity models using Y are comparable. When $M_1(X_1)$ is best, the proposed PVC estimator and the estimators based on $M_1(X_1)$ and $\tilde{M}_1(X_1)$ are comparable, exhibiting negligible bias and slightly larger RMSEs than those of $\bar{Y}$ in all cases. The estimators based on $M_2(X_1, X_2)$, and $M_2(X_1, X_3)$ are also unbiased, but $\tilde{M}_2(X_1, X_3)$ and $\tilde{M}_2(X_1, X_2)$ have much larger RMSEs, owing to their use of Y^2. The estimators based on observed Y

Table 3. Simulated bias and RMSE when estimating $E(Y)$ in simulation 2.

Best model	Method	$n = 300$		$n = 500$		$n = 1{,}000$	
		Bias	RMSE	Bias	RMSE	Bias	RMSE
M_0	PVC	-0.029	0.349	0.000	0.285	0.003	0.195
	FULL	-0.020	0.336	0.003	0.274	0.005	0.188
	CC	0.791	0.878	0.814	0.869	0.819	0.846
	M_0	-0.037	0.349	-0.005	0.285	0.001	0.196
	$M_1(X_1)$	-0.021	0.359	-0.002	0.297	0.002	0.203
	$M_1(X_2)$	-0.036	0.359	-0.003	0.292	0.001	0.202
	$M_1(X_3)$	-0.034	0.364	-0.005	0.299	0.002	0.202
	$M_2(X_1, X_2)$	-0.006	0.412	0.002	0.342	0.001	0.223
	$M_2(X_1, X_3)$	-0.006	0.411	0.003	0.335	0.002	0.227
	$M_2(X_2, X_3)$	-0.013	0.415	-0.001	0.328	0.001	0.226
	$\tilde{M}_0$	-0.017	0.345	0.015	0.282	0.020	0.195
	$\tilde{M}_1(X_1)$	-0.014	0.356	0.018	0.291	0.023	0.200
	$\tilde{M}_1(X_2)$	-0.018	0.354	0.018	0.286	0.021	0.199
	$\tilde{M}_1(X_3)$	-0.013	0.359	0.016	0.294	0.022	0.199
	$\tilde{M}_2(X_1, X_2)$	0.011	0.399	0.020	0.312	0.015	0.223
	$\tilde{M}_2(X_1, X_3)$	0.014	0.397	0.011	0.330	0.017	0.232
	$\tilde{M}_2(X_2, X_3)$	-0.019	0.399	0.014	0.321	0.017	0.222
$M_1(X_1)$	PVC	-0.005	0.354	-0.007	0.287	0.000	0.205
	FULL	0.003	0.329	0.004	0.269	0.006	0.190
	CC	0.790	0.894	0.794	0.861	0.797	0.832
	M_0	0.097	0.382	0.101	0.308	0.114	0.240
	$M_1(X_1)$	-0.004	0.357	-0.004	0.288	0.002	0.206
	$M_1(X_2)$	0.516	0.891	0.492	0.815	0.464	0.684
	$M_1(X_3)$	0.510	0.866	0.455	0.679	0.468	0.648
	$M_2(X_1, X_2)$	0.003	0.490	-0.011	0.391	0.011	0.296
	$M_2(X_1, X_3)$	-0.009	0.444	0.000	0.351	-0.006	0.245
	$M_2(X_2, X_3)$	1.836	2.071	1.957	2.175	2.103	2.284
	$\tilde{M}_0$	0.096	0.397	0.092	0.323	0.109	0.265
	$\tilde{M}_1(X_1)$	-0.026	0.376	-0.029	0.304	-0.022	0.220
	$\tilde{M}_1(X_2)$	0.575	1.038	0.549	1.017	0.520	0.982
	$\tilde{M}_1(X_3)$	0.589	1.063	0.595	1.053	0.505	0.993
	$\tilde{M}_2(X_1, X_2)$	-0.026	0.527	-0.026	0.451	0.016	0.520
	$\tilde{M}_2(X_1, X_3)$	-0.011	0.560	-0.009	0.537	0.024	0.577
	$\tilde{M}_2(X_2, X_3)$	1.973	2.247	2.218	2.473	2.465	2.694

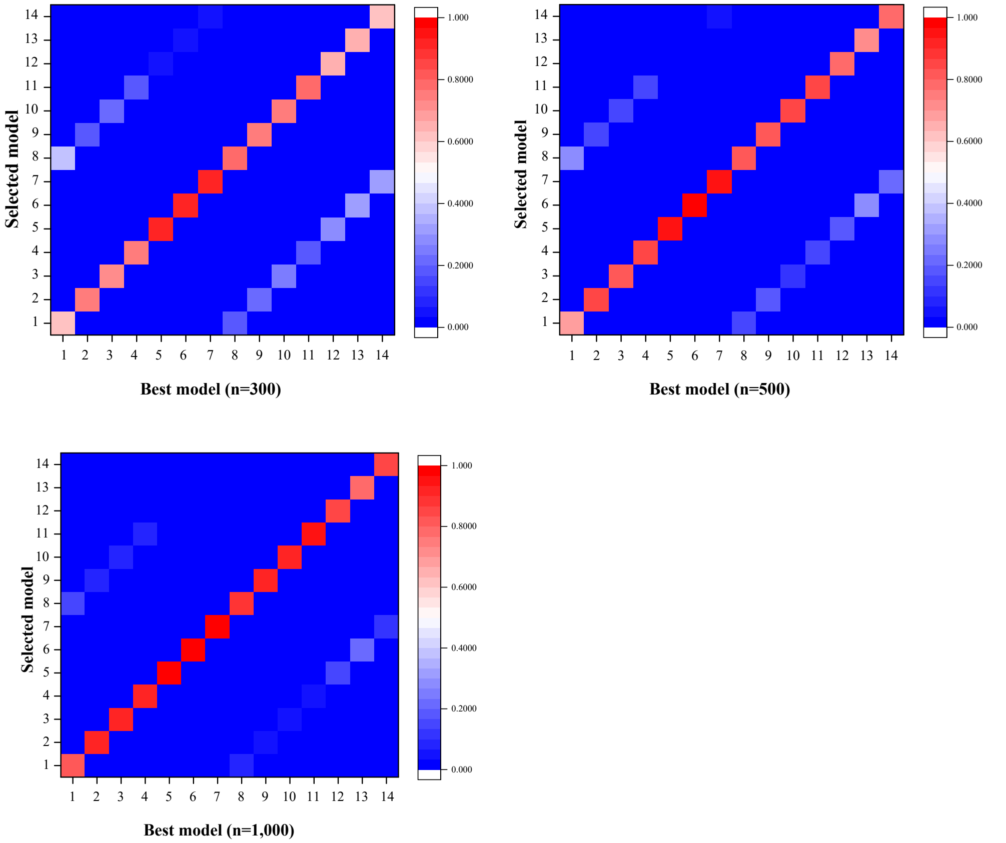

Figure 1. Heat map of true selection rates in simulation 2. The model numbers $\{1,2,\ldots,13,14\}$ denote models $\{M_0, M_1(X_1),\ldots,M_1(X_2,X_3),\tilde{M}_0,\tilde{M}_1(X_1),\ldots,\tilde{M}_1(X_2,X_3)\}$, respectively, in the second column in Table 3.

values and other propensity models have larger biases and RMSEs, supporting our theory. When $M_2(X_1,X_2)$ is best, the proposed PVC estimator and the estimator based on $M_2(X_1,X_2)$ are comparable, exhibiting negligible bias and slightly larger RMSEs than those of $\bar{Y}$ in all cases. Furthermore, the $\hat{\mu}_k$ based on the incorrect models have much larger biases and RMSEs than those of the method based on observed Y values. For the seven propensity models using Y^2, the conclusions are similar and, hence, the results are omitted.

Our final simulation examines the forward instrument selection procedure discussed in Section 3 when the dimension of $\boldsymbol{X}$ is 10. As in simulation 1, $\boldsymbol{X}=(X_1,X_2,\ldots,X_{10})$ is generated from a 10-dimensional normal distribution with mean one and covariance $\mathrm{Cov}(X_j,X_{j'})=0.5$ for $1\le j<j'\le 10$ and

Table 3. Continued.

Best model	Method	$n = 300$		$n = 500$		$n = 1{,}000$	
		Bias	RMSE	Bias	RMSE	Bias	RMSE
$M_2(X_1, X_2)$	PVC	-0.010	0.448	-0.023	0.353	-0.001	0.247
	FULL	0.005	0.338	-0.005	0.267	0.002	0.189
	CC	0.282	0.595	0.280	0.497	0.290	0.412
	M_0	0.599	0.894	0.651	0.826	0.698	0.818
	$M_1(X_1)$	0.473	0.838	0.479	0.689	0.534	0.811
	$M_1(X_2)$	0.487	0.859	0.475	0.655	0.513	0.594
	$M_1(X_3)$	2.662	2.864	2.853	3.076	2.918	3.055
	$M_2(X_1, X_2)$	-0.034	0.482	-0.033	0.378	0.000	0.248
	$M_2(X_1, X_3)$	2.863	3.122	3.018	3.301	3.074	3.336
	$M_2(X_2, X_3)$	2.914	3.231	3.109	3.423	3.099	3.321
	$\tilde{M}_0$	0.632	1.174	0.591	1.193	0.627	1.287
	$\tilde{M}_1(X_1)$	0.959	1.363	0.949	1.288	1.018	1.351
	$\tilde{M}_1(X_2)$	0.960	1.356	0.963	1.286	1.018	1.313
	$\tilde{M}_1(X_3)$	1.800	2.102	1.913	2.318	2.062	2.477
	$\tilde{M}_2(X_1, X_2)$	0.847	1.318	0.837	1.245	0.984	1.306
	$\tilde{M}_2(X_1, X_3)$	2.025	2.440	2.169	2.601	2.268	2.727
	$\tilde{M}_2(X_2, X_3)$	1.992	2.407	2.201	2.671	2.308	2.763

$\mathrm{Var}(X_j) = 1$; Y is generated from

$$Y = X_1^2 + X_2^2 + \cdots + X_{10}^2 + \varepsilon,$$

where ε is taken from $N(0, 2)$ and is independent of $\boldsymbol{X}$. We denote the 10 possible models in (2.1) by M_0 and $M_j(\boldsymbol{U}_j) = 1/\{1 + \exp(\alpha + \beta^T \boldsymbol{U}_j + \gamma Y)\}$, with $\boldsymbol{U}_j = (X_1, \ldots, X_j)$ for $j = 1, \ldots, 9$. In addition, we consider $\boldsymbol{\theta}^0 = (0.2, -0.2)$ for M_0, $\boldsymbol{\theta}^0 = (-0.8, 0.8, -0.2)$ for $M_1(U_1)$, $\boldsymbol{\theta}^0 = (-0.8, 0.8, 0.8, -0.2)$ for $M_2(\boldsymbol{U}_2)$, $\boldsymbol{\theta}^0 = (-0.8, 0.8, 0.8, 0.8, -0.4)$ for $M_3(\boldsymbol{U}_3)$, $\boldsymbol{\theta}^0 = (-0.8, 1, 1, 1, 1, -0.6)$ for $M_4(\boldsymbol{U}_4)$, $\boldsymbol{\theta}^0 = (-2, 0.8, \ldots, 0.8, -0.4)$ for $M_5(\boldsymbol{U}_5)$, $\boldsymbol{\theta}^0 = (-2, 0.8, \ldots, 0.8, -0.4)$ for $M_6(\boldsymbol{U}_6)$, $\boldsymbol{\theta}^0 = (-0.8, 0.8, \ldots, 0.8, -0.8)$ for $M_7(\boldsymbol{U}_7)$, $\boldsymbol{\theta}^0 = (-1, 0.8, \ldots, 0.8, -0.6)$ for $M_8(\boldsymbol{U}_8)$, and $\boldsymbol{\theta}^0 = (-2, 0.8, \ldots, 0.8, -0.6)$ for $M_9(\boldsymbol{U}_9)$. For $n = 1{,}000$ or $1{,}500$, we compute the simulation rates for the forward instrument selection procedure when selecting the correct, best (the most compact correct), and incorrect models. The results, shown in Table 4, indicate that the procedure performs well, especially when the dimension of $\boldsymbol{U}$ is small.

4. Real-Data Example

We consider a data set from the National Health and Nutrition Examination Survey (NHNES) conducted in 2005 by the United States Centers for Disease Control and Prevention. The survey was designed to assess the health and nutritional status of adults and children in the United States. The data are available at www.cdc.gov/nchs/nhanes.htm.

As in Fang and Shao (2016), we consider body fat percentage, measured by dual-energy X-ray absorptiometry (dxa), as the response variable Y; body mass index (bmi), gender, and age as covariates, that is, $\boldsymbol{X} =$ (bmi, gender, age); and middle-aged and older people (age ≥ 45). This yielded $n = 1{,}591$ subjects, 393 (24.7%) of which have missing Y data.

As in the first simulation study in Section 3, we consider the seven candidate propensity models in (2.1), based on the assumption that the underlying true propensity model is a logistic model linear in Y and $\boldsymbol{X}$. Then, we implement the proposed method to select an instrument $\boldsymbol{Z}$. Thus, we have seven choices of instrument: $\boldsymbol{Z} =$ bmi, $\boldsymbol{Z} =$ gender, $\boldsymbol{Z} =$ age, $\boldsymbol{Z} =$ (bmi, gender), $\boldsymbol{Z} =$ (bmi, age), $\boldsymbol{Z} =$ (gender, age), and $\boldsymbol{Z} =$ (bmi, gender, age). For each choice of $\boldsymbol{Z}$, the covariates not included in $\boldsymbol{Z}$ are treated as $\boldsymbol{U}$. We use the proposed PVC in (2.6), with a tuning parameter $\hat{\lambda}$ obtained from a 10-fold cross-validation. For each candidate propensity model, the values of the proposed PVC, estimate of the population mean of dxa, and its standard error are based on a bootstrap with 200 replications (see Table 4). The proposed PVC method selects M_1(bmi) (i.e., $\boldsymbol{Z}$ =(gender, age)), which is consistent with the results of Fang and Shao (2016), which were obtained under a different setting in which $F_{Y|\boldsymbol{X}}$ is parametric and the propensity $\pi(Y, \boldsymbol{U})$ is unspecified. Of the seven choices of instruments, the mean estimates based on $\boldsymbol{Z} =$ bmi, $\boldsymbol{Z} =$ (bmi, age), $\boldsymbol{Z} =$ (bmi, gender), and $\boldsymbol{Z} =$ (bmi, age, gender) differ from the proposed mean estimate based on $\boldsymbol{Z} =$ (gender, age), indicating that these are wrong choices of instruments. On the other hand, as mentioned in Section 2, $\boldsymbol{Z} =$ age and $\boldsymbol{Z} =$ gender are correct choices of instruments if $\boldsymbol{Z} =$ (gender, age) is an instrument; they provide similar mean estimates, but the estimates based on $\boldsymbol{Z} =$ age and $\boldsymbol{Z} =$ gender have much larger SEs. Therefore, to ensure an efficient mean estimator, we should select an instrument with the largest possible dimension.

As in the second simulation in Section 3, we further include the seven candidate propensity models that are logistically linear in $(Y^2, \boldsymbol{X})$. The results are also presented in Table 5. The proposed PVC method still selects M_1(bmi) from

Table 4. Instrument selection rates when the dimension of $\boldsymbol{X}$ is 10.

	$n = 1{,}000$			$n = 1{,}500$		
Best model	Correct	Best	Wrong	Correct	Best	Wrong
M_0	1.000	0.960	0.000	1.000	0.969	0.000
$M_1(X_1)$	0.995	0.955	0.005	1.000	0.988	0.000
$M_2(X_1, X_2)$	0.980	0.795	0.020	0.986	0.916	0.014
$M_3(X_1, X_2, X_3)$	0.975	0.815	0.025	0.978	0.856	0.022
$M_4(X_1, \ldots, X_4)$	0.963	0.563	0.038	0.992	0.711	0.008
$M_5(X_1, \ldots, X_5)$	0.918	0.664	0.092	0.950	0.745	0.050
$M_6(X_1, \ldots, X_6)$	0.900	0.342	0.100	0.915	0.375	0.085
$M_7(X_1, \ldots, X_7)$	0.693	0.288	0.317	0.909	0.424	0.091
$M_8(X_1, \ldots, X_8)$	0.739	0.320	0.261	0.865	0.351	0.135
$M_9(X_1, \ldots, X_9)$	0.530	0.530	0.470	0.733	0.733	0.267

all 14 candidate models, indicating that the propensity models with a term Y^2, but not Y, are incorrect.

The United States Centers for Disease Control and Prevention indicated that, in this problem, missing responses may not be ingorable, after examining missing items in the data files. To determine the effect of addressing nonignorable nonresponses, we computed estimates of $E(Y)$ by assuming ignorable nonresponses and excluding the Y term in the logistic propensity previously discussed. The resulting models are denoted by $M_s^{-Y}(\boldsymbol{U})$, and so on. For example, $M_1^{-Y}(\text{bmi}) = 1/\{1 + \exp(\alpha + \beta \times \text{bmi})\}$. The results are included in Table 5. Note that the estimate under M_0^{-Y} is equal to the sample mean of observed Y values, which is 34.44. Regardless of which ignorable propensity model is used, all estimates of $E(Y)$ are between 34.17 and 34.68, which are close to the sample mean of the observed Y data. Thus, in this example, we do see some effect of addressing nonignorable nonresponses, although the extent of this effect is unknown in a real data set.

5. Discussion

Handling nonignorable nonresponses is a challenging problem, mainly because of the identifiability of the nonresponse propensity. A nonresponse instrument plays a crucial role in identifiability, but is often assumed as given in the literature. Furthermore, to obtain consistent estimators, the imposed parametric propensity model must be verified. Thus, we have proposed a simultaneous propensity model and instrument selection criterion in the presence of nonignorable nonresponses. We showed that the proposed method consistently selects

Table 5. Values of PVC, $\hat{\mu}$, and standard error (SE) based on NHNES data

Model ($\boldsymbol{U}$)	$\boldsymbol{Z}$	PVC	$\hat{\mu}$	SE
M_0	bmi, age, gender	0.24	32.03	0.94
M_1(bmi)	age, gender	0.19	35.28	0.67
M_1(age)	bmi, gender	0.22	33.19	0.54
M_1(gender)	bmi, age	0.24	31.94	2.09
M_2(bmi, age)	gender	0.21	35.40	1.17
M_2(bmi, gender)	age	0.25	35.96	1.28
M_2(age, gender)	bmi	0.26	36.06	1.24
$\tilde{M}_0$	bmi, age, gender	0.25	31.92	0.74
$\tilde{M}_1$(bmi)	age, gender	1.38	31.52	0.58
$\tilde{M}_1$(age)	bmi, gender	1.37	31.49	0.79
$\tilde{M}_1$(gender)	bmi, age	1.39	31.36	0.49
$\tilde{M}_2$(bmi, age)	gender	0.22	35.40	1.09
$\tilde{M}_2$(bmi, gender)	age	1.40	31.50	0.67
$\tilde{M}_2$(age, gender)	bmi	0.22	36.02	1.30
M_0^{-Y}			34.44	0.95
M_1^{-Y}(bmi)			34.68	1.08
M_1^{-Y}(age)			34.46	0.95
M_1^{-Y}(gender)			34.16	1.13
M_2^{-Y}(bmi, age)			34.68	1.09
M_2^{-Y}(bmi, gender)			34.33	1.08
M_2^{-Y}(age, gender)			34.17	1.15
M_3^{-Y}(bmi,age, gender)			34.33	1.10

the most compact correct parametric propensity model and instrument from a group of candidate models, assuming one of these candidate models is correct and an instrument exists. The simulation studies and data analysis show that the proposed method performs well.

The proposed method based on (1.1), (1.2A), and (2.2)–(2.6) can be extended to the situation where Y is multivariate (with δ changed to a vector of indicators) or the situation where both Y and $\boldsymbol{X}$ have missing data. By way of illustration, we consider the situation where $\boldsymbol{Z}$ is always observed, and $\boldsymbol{U}$ and Y have missing values. Let δ_Y and δ_U be the indicators of observing Y and $\boldsymbol{U}$, respectively. Instead of (1.1), we assume that

$$\Pr(\delta_Y = t, \delta_U = s | Y, \boldsymbol{U}, \boldsymbol{Z}) = \Pr(\delta_Y = t, \delta_U = s | Y, \boldsymbol{U}),$$

where $t = 0, 1$ and s is a vector of zeros and ones values. Then, we consider the collection of all K parametric models $\mathcal{M} = \{\pi_k(Y, \boldsymbol{U}_k, \boldsymbol{\theta}_k),\ k = 1, \ldots, K\}$, for $\Pr(\delta_Y = t, \delta_U = s | Y, \boldsymbol{U})$. The proposed PVC can be adopted.

The proposed method has several limitations. First, our method is applicable for small or moderate p only. For high-dimensional covariates, the model selection and instrument search with nonignorable nonresponses is challenging. One possible solution is to first apply a proper variable/feature screening method to reduce the dimensionality of the covariates, and then to apply the proposed method to the reduced number of covariates. Second, the stepwise selection procedure has some limitations. For example, the order of covariate entry and the number of covariates may affect the selected model. Third, the proposed method relies on the assumption that one of the candidate models is correct and an instrument exists. In practice, we may consider a number of potential candidate models, and try to ensure that at least one is correct. When all candidate models are incorrect or no instrument exists, we may only derive some procedures that are approximately valid. Additional research on the case in which no correct candidate model or instrument exists is still interesting, although challenging, because no model is perfect in all practical applications. These issues will be explored in further research.

Acknowledgements

We are grateful to the editor, associate editor, and two anonymous referees for their insightful comments and suggestions. Our research was supported by the National Natural Science Foundation of China (11831008, 11871287, 11601156, 11771144, 11801359), Natural Science Foundation of Tianjin (18JCYBJC41100), Fundamental Research Funds for the Central Universities, Key Laboratory for Medical Data Analysis and Statistical Research of Tianjin, the Chinese 111 Project (B14019), and the U.S. National Science Foundation (DMS-1612873).

Appendix

Proof of Lemma 1: Recall

$$\bar{\boldsymbol{G}}_{kn}(\boldsymbol{\theta}_k) = n^{-1} \sum_{i=1}^{n} \boldsymbol{g}_k(Y_i, \boldsymbol{X}_i, \delta_i, \boldsymbol{\theta}_k) \text{ and } \boldsymbol{G}_k(\boldsymbol{\theta}_k) = E\{\boldsymbol{g}_k(Y, \boldsymbol{U}, \delta, \boldsymbol{\theta}_k)\}.$$

By the law of large number (LLN), it can be shown that $\bar{\boldsymbol{G}}_{kn}(\boldsymbol{\theta}_k) - \boldsymbol{G}_k(\boldsymbol{\theta}_k) = o_p(1)$ for all $\boldsymbol{\theta}_k \in \mathcal{A}$. Since both $\boldsymbol{g}_k(Y, \boldsymbol{U}, \delta, \boldsymbol{\theta}_k)$ and $\bar{\boldsymbol{G}}_{kn}(\boldsymbol{\theta}_k)$ are continuous at each

$\boldsymbol{\theta}_k \in \mathcal{A}$,

$$\sup_{\boldsymbol{\theta}_k \in \mathcal{A}} \|\bar{\boldsymbol{G}}_{kn}(\boldsymbol{\theta}_k) - \boldsymbol{G}_k(\boldsymbol{\theta}_k)\| = o_p(1).$$

This, coupled with GMM identification (i.e., Lemma 2.3 of Newey and McFadden (1994)), shows that the first-step estimator

$$\tilde{\boldsymbol{\theta}}_k = \boldsymbol{\theta}_k^* + o_p(1).$$

By the LLN, it can be shown that $\hat{\boldsymbol{W}}_{kn}^{-1} = \boldsymbol{W}_k^{-1}(\boldsymbol{\theta}_k^*) + o_p(1)$. Let

$$Q_k(\boldsymbol{\theta}_k) = \boldsymbol{G}_k(\boldsymbol{\theta}_k)^\top \boldsymbol{W}_k^{-1}(\boldsymbol{\theta}_k^*)\boldsymbol{G}_k(\boldsymbol{\theta}_k) \text{ and } \bar{Q}_k(\boldsymbol{\theta}_k) = \bar{\boldsymbol{G}}_{kn}(\boldsymbol{\theta}_k)^\top \hat{\boldsymbol{W}}_{kn}^{-1}\bar{\boldsymbol{G}}_{kn}(\boldsymbol{\theta}_k).$$

Based on Lemma 2.3 and Theorem 2.1 of Newey and McFadden (1994), to prove $\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^* = o_p(1)$, it is enough to show that

$$\sup_{\boldsymbol{\theta}_k \in \mathcal{A}} |\bar{Q}_{kn}(\boldsymbol{\theta}_k) - Q_k(\boldsymbol{\theta}_k)| = o_p(1).$$

Using the triangle and Cauchy-Schwartz inequalities, we have

$$\begin{aligned}
&\sup_{\boldsymbol{\theta}_k \in \mathcal{A}} |\bar{Q}_{kn}(\boldsymbol{\theta}_k) - Q_k(\boldsymbol{\theta}_k)| \\
&\le \sup_{\boldsymbol{\theta}_k \in \mathcal{A}} \left|\{\bar{\boldsymbol{G}}_{kn}(\boldsymbol{\theta}_k) - \boldsymbol{G}_k(\boldsymbol{\theta}_k)\}^\top \hat{\boldsymbol{W}}_{kn}^{-1}\{\bar{\boldsymbol{G}}_{kn}(\boldsymbol{\theta}_k) - \boldsymbol{G}_k(\boldsymbol{\theta}_k)\}\right| \\
&\quad + \sup_{\boldsymbol{\theta}_k \in \mathcal{A}} \left|\boldsymbol{G}_k(\boldsymbol{\theta}_k)^\top (\hat{\boldsymbol{W}}_{kn}^{-1} + (\hat{\boldsymbol{W}}_{kn}^{-1})^\top)\{\bar{\boldsymbol{G}}_{kn}(\boldsymbol{\theta}_k) - \boldsymbol{G}_k(\boldsymbol{\theta}_k)\}\right| \\
&\quad + \sup_{\boldsymbol{\theta}_k \in \mathcal{A}} \left|\boldsymbol{G}_k(\boldsymbol{\theta}_k)^\top (\hat{\boldsymbol{W}}_{kn}^{-1} - \boldsymbol{W}_k^{-1}(\boldsymbol{\theta}_k^*))\boldsymbol{G}_k(\boldsymbol{\theta}_k)\right| \\
&\le \sup_{\boldsymbol{\theta}_k \in \mathcal{A}} \left\|\bar{\boldsymbol{G}}_{kn}(\boldsymbol{\theta}_k) - \boldsymbol{G}_k(\boldsymbol{\theta}_k)\}\right\|^2 \|\hat{\boldsymbol{W}}_{kn}^{-1}\| \\
&\quad + 2\sup_{\boldsymbol{\theta}_k \in \mathcal{A}} \left\|\boldsymbol{G}_k(\boldsymbol{\theta}_k)\right\|\left\|\bar{\boldsymbol{G}}_{kn}(\boldsymbol{\theta}_k) - \boldsymbol{G}_k(\boldsymbol{\theta}_k)\right\|\|\boldsymbol{W}_k^{-1}(\boldsymbol{\theta}_k^*)\| \\
&\quad + \sup_{\boldsymbol{\theta}_k \in \mathcal{A}} \left\|\boldsymbol{G}_k(\boldsymbol{\theta}_k)\right\|^2 \|\hat{\boldsymbol{W}}_{kn}^{-1} - \boldsymbol{W}_k^{-1}(\boldsymbol{\theta}_k^*)\| = o_p(1).
\end{aligned}$$

Thus, we prove that $\hat{\boldsymbol{\theta}}_k = \boldsymbol{\theta}_k^* + o_p(1)$.

Next, we derive the asymptotic normality of $\hat{\boldsymbol{\theta}}_k$. With probability approaching one, we have the first-order condition

$$2\boldsymbol{\Gamma}_k(\hat{\boldsymbol{\theta}}_k)\hat{\boldsymbol{W}}_{kn}^{-1}\bar{\boldsymbol{G}}_{kn}(\hat{\boldsymbol{\theta}}_k) = 0,$$

where $\boldsymbol{\Gamma}_k(\boldsymbol{\theta}_k) = \partial\bar{\boldsymbol{G}}_{kn}(\boldsymbol{\theta}_k)/\partial\boldsymbol{\theta}_k$. Expanding $\bar{\boldsymbol{G}}_{kn}(\hat{\boldsymbol{\theta}}_k)$ around $\boldsymbol{\theta}_{k^*}$, we have

$$n^{1/2}(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*) = -[\boldsymbol{\Gamma}_k^\top(\hat{\boldsymbol{\theta}}_k)\hat{\boldsymbol{W}}_{kn}^{-1}\boldsymbol{\Gamma}_k(\check{\boldsymbol{\theta}}_k)]^{-1}\boldsymbol{\Gamma}_k^\top(\hat{\boldsymbol{\theta}}_k)\hat{\boldsymbol{W}}_{kn}^{-1}n^{1/2}\bar{\boldsymbol{G}}_{kn}(\boldsymbol{\theta}_k^*),$$

where $\check{\boldsymbol{\theta}}_k$ is between $\hat{\boldsymbol{\theta}}_k$ and $\boldsymbol{\theta}_k^0$. By simple calculation and the LLN, for all $\boldsymbol{\theta}_k \in \mathcal{A}$,

$$\boldsymbol{\Gamma}_k(\boldsymbol{\theta}_k) = E\left\{h_k(\boldsymbol{X})^\top\delta\frac{\partial\pi_k(Y, \boldsymbol{U}, \boldsymbol{\theta}_k)^{-1}}{\partial\boldsymbol{\theta}_k}\right\} + o_p(1).$$

This, together with $\hat{\boldsymbol{W}}_{kn}^{-1} = \boldsymbol{W}_k^{-1}(\boldsymbol{\theta}_k^*) + o_p(1)$ and $\hat{\boldsymbol{\theta}}_k = \boldsymbol{\theta}_k^* + o_p(1)$, implies that

$$\begin{aligned}&[\boldsymbol{\Gamma}_k^\top(\hat{\boldsymbol{\theta}}_k)\hat{\boldsymbol{W}}_{kn}^{-1}\boldsymbol{\Gamma}_k(\check{\boldsymbol{\theta}}_k)]^{-1}\boldsymbol{\Gamma}_k^\top(\hat{\boldsymbol{\theta}}_k)\hat{\boldsymbol{W}}_{kn}^{-1}\\ &= [\boldsymbol{\Gamma}_k(\boldsymbol{\theta}_k^*)^\top\boldsymbol{W}_k^{-1}(\boldsymbol{\theta}_k^*)\boldsymbol{\Gamma}_k(\boldsymbol{\theta}_k^*)]^{-1}\Gamma_k(\boldsymbol{\theta}_k^*)^\top\boldsymbol{W}_k^{-1}(\boldsymbol{\theta}_k^*) + o_p(1).\end{aligned}$$

By the Slutzky theorem, we can show

$$n^{1/2}(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*) \overset{\mathcal{L}}{\to} N(0, (\boldsymbol{\Gamma}_k(\boldsymbol{\theta}_k^*)^\top\boldsymbol{W}_k^{-1}(\boldsymbol{\theta}_k^*)\boldsymbol{\Gamma}_k(\boldsymbol{\theta}_k^*))^{-1}).$$

Particularly, when the intermittent propensity model is correctly specified, $\pi_k(Y, \boldsymbol{U}, \boldsymbol{\theta}_k^*) = \pi_k(Y, \boldsymbol{U}, \boldsymbol{\theta}_k^0)$, $\hat{\boldsymbol{W}}_{kn} = \boldsymbol{W}_k(\boldsymbol{\theta}_k^0) + o_p(1)$ and $\hat{\boldsymbol{\theta}}_k = \boldsymbol{\theta}_k^0 + o_p(1)$.

Proof of Lemma 2: Notice

$$|\mathrm{VC}(1) - \mathrm{VC}(k)| \le \frac{1}{n}\sum_{i=1}^n |\hat{F}_1(\boldsymbol{X}_i) - \hat{F}_k(\boldsymbol{X}_i)|.$$

We just need to show $n^{-1/2}\sum_{i=1}^n |\hat{F}_1(\boldsymbol{X}_i) - \hat{F}_k(\boldsymbol{X}_i)| = O_p(1)$. Note that

$$\begin{aligned}&n^{-1/2}\sum_{i=1}^n |\hat{F}_k(\boldsymbol{X}_i) - \hat{F}_1(\boldsymbol{X}_i)|\\ &= n^{-1/2}\sum_{i=1}^n \left|\frac{1}{n}\sum_{j=1}^n \delta_j \mathrm{I}(\boldsymbol{X}_j \le \boldsymbol{X}_i)\left\{\frac{1}{\pi_k(Y_j, \boldsymbol{U}_j, \hat{\boldsymbol{\theta}}_k)} - \frac{1}{\pi_1(Y_j, \boldsymbol{U}_j, \hat{\boldsymbol{\theta}}_1)}\right\}\right|\end{aligned}$$

Let

$$\begin{aligned}A_{ni}^{(1)} &= \frac{1}{n}\sum_{j=1}^n \delta_j \mathrm{I}(\boldsymbol{X}_j \le \boldsymbol{X}_i)\left\{\frac{1}{\pi_k(Y_j, \boldsymbol{U}_j, \hat{\boldsymbol{\theta}}_k)} - \frac{1}{\pi_k(Y_j, \boldsymbol{U}_j, \boldsymbol{\theta}_k^*)}\right\},\\ A_{ni}^{(2)} &= -\frac{1}{n}\sum_{j=1}^n \delta_j \mathrm{I}(\boldsymbol{X}_j \le \boldsymbol{X}_i)\left\{\frac{1}{\pi_1(Y_j, \boldsymbol{U}_j, \hat{\boldsymbol{\theta}}_1)} - \frac{1}{\pi_1(Y_j, \boldsymbol{U}_j, \boldsymbol{\theta}_1^*)}\right\},\end{aligned}$$

$$A_{ni}^{(3)} = \frac{1}{n}\sum_{j=1}^{n} \delta_j \mathrm{I}(\boldsymbol{X}_j \leq \boldsymbol{X}_i)\left\{\frac{1}{\pi_k(Y_j, \boldsymbol{U}_j, \boldsymbol{\theta}_k^*)} - \frac{1}{\pi_1(Y_j, \boldsymbol{U}_j, \boldsymbol{\theta}_1^*)}\right\}.$$

We have

$$\begin{aligned} n^{-1/2}\sum_{i=1}^{n} |\hat{F}_k(\boldsymbol{X}_i) - \hat{F}_1(\boldsymbol{X}_i)| &= n^{-1/2}\sum_{i=1}^{n} |A_{ni}^{(1)} + A_{ni}^{(2)} + A_{ni}^{(3)}| \\ &\leq n^{-1/2}\left(\sum_{i=1}^{n}|A_{ni}^{(1)}| + \sum_{i=1}^{n}|A_{ni}^{(2)}| + \sum_{i=1}^{n}|A_{ni}^{(3)}|\right). \end{aligned}$$

For $n^{-1/2}\sum_{i=1}^{n}|A_{ni}^{(1)}|$, we have

$$\begin{aligned} & n^{-1/2}\sum_{i=1}^{n}|A_{ni}^{(1)}| \\ & \leq n^{-3/2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left|\delta_j \mathrm{I}(\boldsymbol{X}_j \leq \boldsymbol{X}_i)\left\{\frac{1}{\pi_k(Y_j, \boldsymbol{U}_j, \hat{\boldsymbol{\theta}}_k)} - \frac{1}{\pi_k(Y_j, \boldsymbol{U}_j, \boldsymbol{\theta}_k^*)}\right\}\right| \\ & \leq n^{-3/2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left|\frac{1}{\pi_k(Y_j, \boldsymbol{U}_j, \hat{\boldsymbol{\theta}}_k)} - \frac{1}{\pi_k(Y_j, \boldsymbol{U}_j, \boldsymbol{\theta}_k^*)}\right| \\ & = n^{-3/2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left|\frac{\partial \pi_k^{-1}(Y_j, \boldsymbol{U}_j, \boldsymbol{\theta}_k^*)}{\partial \boldsymbol{\theta}_k}(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*) + o_p(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*)\right| \\ & \leq |\sqrt{n}(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*)| \times \frac{1}{n}\sum_{j=1}^{n}\left|\frac{\partial \pi_k^{-1}(Y_j, \boldsymbol{U}_j, \boldsymbol{\theta}_k^*)}{\partial \boldsymbol{\theta}_k}\right| + o_p(1) \\ & = |\sqrt{n}(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*)| \times E\left|\frac{\partial \pi_k^{-1}(Y, \boldsymbol{U}, \boldsymbol{\theta}_k^*)}{\partial \boldsymbol{\theta}_k}\right| + o_p(1) \\ & = O_p(1). \end{aligned}$$

Similarly, we can show that $n^{-1/2}\sum_{i=1}^{n}|A_{ni}^{(2)}| = O_p(1)$ and $n^{-1/2}\sum_{i=1}^{n}|A_{ni}^{(3)}| = O_p(1)$.

References

Ai, C., Linton, O. and Zhang, Z. (2020). A simple and efficient estimation method for models with nonignorable missing data. *Statistica Sinica*, to appear.

Cheng, P. E. (1994). Nonparametric estimation of mean functionals with data missing at random. *Journal of American Statistical Association* **89**, 81–87.

Fang, F. and Shao, J. (2016). Model selection with nonignorable nonresponse. *Biometrika* **103**, 861–874.

Fitzmaurice, G. M., Molenberghs, G. and Lipsitz, S. R. (1995). Regression models for longitudinal binary responses with informative drop-outs. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **57**, 691–704.

Guan, Z. and Qin, J. (2017). Empirical likelihood method for non-ignorable missing data problems. *Lifetime Data Analysis* **23**, 113–135.

Ibrahim, J. G., Chen, M. H., Lipsitz, S. R. and Herring, A. H. (2005). Missing data methods for generalized linear models: A comparative review. *Journal of American Statistical Association* **100**, 332–346.

Kim, J. K. and Shao, J. (2013). *Statistical Methods for Handling Incomplete Data.* Chapman & Hall/CRC, London.

Kim, J. K. and Yu, C. L. (2011). A semiparametric estimation of mean functionals with nonignorable missing data. *Journal of American Statistical Association* **106**, 157–165.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data.* 2nd Edition. Wiley, New York.

Molenberghs, G. and Kenward, M. G. (2007). *Missing Data in Clinical Studies.* Wiley, New York.

Morikawa, K., Kim, J. K. and Kano, Y. (2017). Semiparametric maximum likelihood estimation with data missing not at random. *Canadian Journal of Statistics* **45**, 393–409.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of Econometrics* **4**, 2111–2245.

Qin, J., Leung, D. and Shao, J. (2002). Estimation with survey data under nonignorable nonresponse or informative sampling. *Journal of American Statistical Association* **97**, 193–200.

Robins, J. M. (1987). Inference and missing data. *Biometrika* **63**, 581–592.

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.

Robins, J. 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.

Scharfstein, D. O., Rotnitzky, A. and Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of American Statistical Association* **94**, 1096–1120.

Shao, J. and Wang, L. (2016). Semiparametric inverse propensity weighting for nonignorable missing data. *Biometrika* **103**, 175–187.

Tang, G., Little, R. J. A. and Raghunathan, T. E. (2003). Analysis of multivariate missing data with nonignorable nonresponse. *Biometrika* **90**, 747–764.

Tang, N., Zhao, P. and Zhu, H. (2014). Empirical likelihood for estimating equations with nonignorably missing data. *Statistica Sinica* **24**, 723–747.

Wang, S., Shao, J. and Kim, J. K. (2014). An instrumental variable approach for identification and estimation with nonignorable nonresponse. *Statistica Sinica* **24**, 1097–1116.

Xie, H., Qian, Y. and Qu, L. (2011). Semiparametric approach for analyzing nonignorable missing data. *Statistica Sinica* **21**, 1881–1899.

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, J., Yang, Y. and Ning, Y. (2018). Penalized pairwise pseudo likelihood for variable selection with nonignorable missing data. *Statistica Sinica* **28**, 2125–2148.

Lei Wang

School of Statistics and Data Science & LPMC, Nankai University, Tianjin 300071, China.

E-mail: lwangstat@nankai.edu.cn

Jun Shao

KLATASDS-MOE and School of Statistics, East China Normal University, 3663 North Zhongshan Rd., Shanghai, China 200062.

E-mail: shao@stat.wisc.edu

Fang Fang

KLATASDS-MOE and School of Statistics, East China Normal University, 3663 North Zhongshan Rd., Shanghai, China 200062.

E-mail: ffang@sfs.ecnu.edu.cn

(Received March 2018; accepted June 2019)