Statistica Sinica **30** (2020), 1949-1970
doi:https://doi.org/10.5705/ss.202018.0107

# A SIMPLE AND EFFICIENT ESTIMATION METHOD FOR MODELS WITH NON-IGNORABLE MISSING DATA

Chunrong Ai, Oliver Linton and Zheng Zhang

*Chinese University of Hong Kong, Shenzhen, University of Cambridge and Renmin University of China*

*Abstract:* This paper proposes a simple and efficient generalized method of moments (GMM) estimation for a model with non-ignorable missing data. In contrast to the existing the GMM estimation with a fixed number of moments, we allow the number of moments to grow with the sample size and use optimal weighting. Hence, our estimator is efficient, attaining the semiparametric efficiency bound derived in the literature. Existing semiparametric estimators estimate an efficient score. However, this approach is either locally efficient, or it suffers from the curse of dimensionality and the bandwidth selection problem. In contrast, our estimator does not suffer from these problems. Moreover, the proposed estimator and its consistent covariance matrix are easily computed using commercially available GMM packages. We propose two data-driven methods to select the number of moments. A small-scale simulation study reveals that the proposed estimator outperforms existing alternatives in finite samples.

*Key words and phrases:* Generalized method of moments, non-ignorable nonresponse, semiparametric efficiency.

## 1. Introduction

Missing data are common in many fields. One way to resolve the problem is to delete observations containing such data. However, in doing so, we may produce biased estimates and erroneous conclusions, depending on the missing data mechanism. If data are missing completely at random, standard estimation and inference procedures remain consistent when the missing data observations are ignored; see Heitjan and Basu (1996) and Little (1988), among others. If data are missing at random (MAR), in the sense that the propensity of missingness depends only on the observed covariates, a consistent estimation can still be obtained using covariate balancing; see Rubin (1976a,b), Little and Rubin (1989), Robins and Rotnitzky (1995), Robins, Rotnitzky and Zhao (1995), Bang and Robins (2005), Qin and Zhang (2007), Chen, Hong and Tarozzi (2008), Tan (2010), Rotnitzky et al. (2012) and Little and Rubin (2014), among others. In

many applications, data are missing not at random (MNAR). For example, the income question in sample surveys is often not answered by people at the top end of the distribution; that is, their response frequency depends on an outcome variable that is often the key focus. Alternatively, consider an investigator who examines the effect of sleep on pain by calling subjects each day to ask them about their previous night's sleep and their pain that day. Patients who are experiencing severe pain are less likely to answer the phone, resulting in data missing for that particular day; again, this violates the MAR assumption. From political science, roll-call votes, which measure legislatures' ideological positions, are subject to non-ignorable nonresponses because politicians behave strategically. In the MNAR case, the parameter of interest may not even be identified (e.g., Robins and Ritov (1997)), let alone be consistently estimated. Specifically, let $T \in \{0,1\}$ denote a binary random variable indicating the missing status of the outcome variable $Y$: if $Y$ is observed, $T$ takes the value one, and if $Y$ is not observed, $T$ takes the value zero. Let $\boldsymbol{X}$ denote a vector of explanatory variables, $\pi(\boldsymbol{x}, y) = P(T = 1|\boldsymbol{X} = \boldsymbol{x}, Y = y)$ denote the propensity score function, and $f_{Y|\boldsymbol{X}}(y|\boldsymbol{x})$ denote the conditional density function of $Y$, given $\boldsymbol{X}$. Robins and Ritov (1997) shows that if both the propensity score function and the conditional density function are completely unknown, the joint distribution of $(T, Y)$, given $\boldsymbol{X}$, is not identifiable. In this case, a necessary identification condition is the parameterization of one of the two functions. Molenberghs and Kenward (2007) propose parameterizing both functions as an identification strategy, whereas Sverchkov (2008) and Riddles, Kim and Im (2016) parameterize the propensity score function and only a component of the conditional density function: $f_{Y|\boldsymbol{X},T}(y|\boldsymbol{x}, T = 1)$.

If the joint distribution is not the parameter of interest, the aforementioned identification strategy can be modified. For example, if the parameter of interest is the conditional density of $Y$, given $\boldsymbol{X}$ (i.e., $f_{Y|\boldsymbol{X}}(y|\boldsymbol{x})$), the parameterization of the propensity score function is not needed. However, the parameterization of $f_{Y|\boldsymbol{X}}(y|\boldsymbol{x})$ in this case is not sufficient for identification, owing to missing data. Tang et al. (2003) suggest parameterizing the marginal density $f_{\boldsymbol{X}}(\boldsymbol{x})$ as well, and Zhao and Shao (2015) impose an exclusion retriction. In both studies, $f_{Y|\boldsymbol{X}}(y|\boldsymbol{x})$ is identified and consistently estimated.

We estimate the parameter $\theta_0 = E[U(\boldsymbol{X}, Y)]$, where $U(\cdot)$ is any known function. We suppose that the propensity score $\pi$ is parameterized, but do not restrict the conditional density function of the outcome variable. In earlier works that adopt this framework, either the coefficients in the propensity score function

are known or are consistently estimated from an external sample (Kim and Yu (2011)), or an exclusion restriction is imposed (Wang, Shao and Kim (2014)) and Shao and Wang (2016)). Wang, Shao and Kim (2014) propose a generalized method of moments estimation for $\theta_0$. However, their estimator is not efficient because their moments are not optimal. Morikawa and Kim (2016) estimate $\theta_0$ by deriving the efficient score function (and, hence, the semiparametric efficiency bound) for $\theta_0$. They propose estimating this function by estimating $f_{Y|\boldsymbol{X},T}(y|,\boldsymbol{x},1)$, using either a working parametric model (MK1) or a kernel nonparametric method (MK2). However, the MK1 approach is not efficient unless the working parametric model is correct, although it is consistent. The MK2 method suffers from the curse of dimensionality, because their smoothness conditions depend on the dimensionality of the covariates (see their condition C14). Furthermore, MK2 suffers from the bandwidth selection problem; unfortunately, no guidance is provided on how to resolve this problem.

We study the same estimation problem as those in Wang, Shao and Kim (2014) and Morikawa and Kim (2016), but propose a simpler, yet equally efficient estimation procedure. Our proposed method does not require an explicit nonparametric estimation and, hence, does not suffer from the curse of dimensionality. The proposed estimator is motivated by the key insight that the model parameter satisfies a parametric conditional moment restriction, of which the semiparametric efficiency bound is identical to the bound derived in Morikawa and Kim (2016). The conditional moment restriction is then turned into an expanding set of unconditional moment restrictions, and the parameter of interest is estimated by applying the widely available and easy to compute GMM estimation (see Hansen (1982)). Under sufficient conditions, we establish that the proposed estimator is consistent and asymptotically normally distributed, even if the set of unconditional moment restrictions does not expand. This resolves the curse of dimensionality and bandwidth selection problems, because when the set does expand, the proposed estimator attains the semiparametric efficiency bound.

The remainder of the paper is organized as follows. Section 2 describes the estimation, and Section 3 derives the large-sample properties of the estimator. Section 4 provides a consistent asymptotic variance estimator, and Section 5 suggests two data-driven approaches to determine the number of unconditional moment restrictions. Section 6 presents on a small-scale simulation study. Section 7 concludes the paper. All technical proofs are relegated to the Supplementary Material Ai, Linton and Zhang (2018).

## 2. Basic Framework and Estimation

We begin by setting up the basic framework. Denote $\boldsymbol{Z} = (\boldsymbol{X}^\top, Y)^\top$. The following assumption is maintained throughout the paper:

**Assumption 2.1.** *(i) Parameterization of missing data mechanism:* $P(T = 1|Y, \boldsymbol{X}) = \pi(Y, \boldsymbol{X}; \gamma_0) = \pi(\boldsymbol{Z}; \gamma_0)$ *holds for some known function* $\pi(\cdot;\cdot)$*, where* $\gamma_0 \in \mathbb{R}^p$*, for some known* $p \in \mathbb{N}$*, is the true (unknown) value; (ii) exclusion restriction: there exist nonresponse instrument variables* $\boldsymbol{X}_1$ *in* $\boldsymbol{X} = (\boldsymbol{X}_1^\top, \boldsymbol{X}_2^\top)^\top$ *such that* $\boldsymbol{X}_2$ *is independent of* $T$*, given both* $\boldsymbol{X}_1$ *and* $Y$*; and (iii) the parameter of interest is* $\theta_0 = E[U(\boldsymbol{Z})]$*, for some known function* $U(\cdot)$*.*

Under Assumption 2.1 and by applying the law of iterated expectations, we obtain the following conditional moment restrictions:

$$E\left[1 - \frac{T}{\pi(\boldsymbol{Z}; \gamma_0)}\middle|\boldsymbol{X}\right] = 0, \tag{2.1}$$

$$E\left[\theta_0 - \frac{T}{\pi(\boldsymbol{Z}; \gamma_0)}U(\boldsymbol{Z})\right] = 0, \tag{2.2}$$

on which the proposed estimation is based. Note that the parameters of interest in (2.1)–(2.2) are finite dimensional (and there is no explicit infinite-dimensional nuisance parameter), and can be easily estimated using a GMM estimation.

The (nuisance) parameter $\gamma_0$ is identified by (2.1), and the parameter of interest $\theta_0$ is identified by (2.2). The following condition is also maintained throughout the paper:

**Assumption 2.2.** *The parameter space* $\Gamma$ *is a compact subset of* $\mathbb{R}^p$*. The true value* $\gamma_0$ *lies in the interior of* $\Gamma$*, and is the only solution to* (2.1)*. The parameter space* $\Theta$ *is a compact subset of* $\mathbb{R}$*, and the true value* $\theta_0$ *lies in the interior of* $\Theta$*.*

To estimate model (2.1)–(2.2), we first turn it into a set of unconditional moment restrictions. We work with a set of known basis functions: for each integer $K \in \mathbb{N}$, with $K \geq p$, let $u_K(\boldsymbol{X}) = (u_{1K}(\boldsymbol{X}), \ldots, u_{KK}(\boldsymbol{X}))^\top$. A discussion on the choice of $u_K(\boldsymbol{X})$ and its properties can be found in Chen (2007), in Wei, Sun and Hu (2018), and in Section 2.2 of the Supplementary Material. Model (2.1)–(2.2) implies the following unconditional moment restrictions:

$$E\left[\left(1 - \frac{T}{\pi(\boldsymbol{Z}; \gamma_0)}\right)u_K(\boldsymbol{X})\right] = 0, \tag{2.3}$$

$$E\left[\theta_0 - \frac{T}{\pi(\boldsymbol{Z};\gamma_0)}U(\boldsymbol{Z})\right] = 0. \tag{2.4}$$

To avoid redundant moment restrictions, we require that $E\left[u_K(\boldsymbol{X})u_K(\boldsymbol{X})^\top\right]$ be nonsingular for every $K$. The following, somewhat stronger identification condition is maintained throughout the paper:

**Assumption 2.2′.** *The parameter space* $\Gamma$ *is a compact subset of* $\mathbb{R}^p$*. The true value* $\gamma_0$ *lies in the interior of* $\Gamma$*, and is the only solution to* (2.3)*. The parameter space* $\Theta$ *is a compact subset of* $\mathbb{R}$*, and the true value* $\theta_0$ *lies in the interior of* $\Theta$*.*

We can estimate the parameter of interest using the GMM. Let $\{T_i, \boldsymbol{Z}_i\}_{i=1}^N$ denote an independent and identically distributed (i.i.d.) sample drawn from the joint distribution of $(T, \boldsymbol{Z})$. Denote $\boldsymbol{G}_K(\gamma,\theta) := \sum_{i=1}^N g_K(T_i, \boldsymbol{Z}_i;\gamma,\theta)$, where $g_K(T,\boldsymbol{Z};\gamma,\theta) := \left(\left[1 - T\pi(\boldsymbol{Z};\gamma)^{-1}\right]u_K(\boldsymbol{X})^\top, \theta - T\pi(\boldsymbol{Z};\gamma)^{-1}U(\boldsymbol{Z})\right)^\top$. The GMM estimator of $\gamma_0$ and $\theta_0$ is defined as

$$(\check{\gamma},\check{\theta}) = \arg\min_{\gamma\in\Gamma,\theta\in\Theta} \boldsymbol{G}_K(\gamma,\theta)^\top \cdot \mathbf{W} \cdot \boldsymbol{G}_K(\gamma,\theta),$$

where $\mathbf{W}$ is a $(K+1)\times(K+1)$ symmetric weighting matrix. For every fixed $K \geq p$, Hansen (1982) shows that, under some regularity conditions, the estimator

$$(\check{\gamma} - \gamma_0, \check{\theta} - \theta_0) = O_p(N^{-1/2}) \tag{2.5}$$

is asymptotically normally distributed, but performs best only when the best weighting matrix is used. The later matrix is the inverse of $\boldsymbol{D}_{(K+1)\times(K+1)} := E\left[g_K(T,\boldsymbol{Z};\gamma_0,\theta_0)g_K(T,\boldsymbol{Z};\gamma_0,\theta_0)^\top\right]$. The best estimator (within the class defined by the specific unconditional moments) is defined as

$$(\overline{\gamma},\overline{\theta}) = \arg\min_{\gamma\in\Gamma,\theta\in\Theta} \boldsymbol{G}_K(\gamma,\theta)^\top \cdot \boldsymbol{D}^{-1}_{(K+1)\times(K+1)} \cdot \boldsymbol{G}_K(\gamma,\theta).$$

Suppose that the propensity score function is differentiable with respect to $\gamma$. Denote

$$\boldsymbol{B}_{(K+1)\times(p+1)} = \nabla_{\gamma,\theta} E\left[\frac{1}{N}\boldsymbol{G}_K(\gamma_0,\theta_0)\right] = \begin{pmatrix} E\left[u_K(\boldsymbol{X})\frac{\nabla_\gamma \pi(\boldsymbol{Z};\gamma_0)^\top}{\pi(\boldsymbol{Z};\gamma_0)}\right], \mathbf{0}_{K\times 1} \\ E\left[U(\boldsymbol{Z})\frac{\nabla_\gamma \pi(\boldsymbol{Z};\gamma_0)^\top}{\pi(\boldsymbol{Z};\gamma_0)}\right], \quad 1 \end{pmatrix}$$

and

$$\boldsymbol{V}_K = \left\{ \left(\boldsymbol{B}_{(K+1)\times(p+1)}\right)^\top \boldsymbol{D}^{-1}_{(K+1)\times(K+1)} \left(\boldsymbol{B}_{(K+1)\times(p+1)}\right) \right\}^{-1}.$$

Hansen (1982) shows that, for every fixed $K \geq p$,

$$\boldsymbol{V}_K^{-1/2} \begin{pmatrix} \sqrt{N}(\overline{\gamma} - \gamma_0) \\ \sqrt{N}(\overline{\theta} - \theta_0) \end{pmatrix} \to N\left(0, I_{(p+1)\times(p+1)}\right) \text{ in distribution.} \tag{2.6}$$

Because the best weighting matrix depends on the unknown parameter value, the best estimator $(\overline{\gamma}, \overline{\theta})$ is infeasible. Hansen (1982) suggests the following two-step procedure:

**Step I.** Compute the initial $\sqrt{N}$-consistent estimator

$$\widehat{\boldsymbol{W}}_0 := \begin{pmatrix} \frac{1}{N}\sum_{i=1}^N u_K(\mathbf{X}_i) u_K(\mathbf{X}_i)^\top & \mathbf{0}_{K\times 1} \\ \mathbf{0}^\top_{K\times 1} & 1 \end{pmatrix},$$
$$(\check{\gamma}, \check{\theta}) = \arg \min_{(\gamma,\theta)\in\Gamma\times\Theta} \boldsymbol{G}_K(\gamma,\theta)^T \cdot \widehat{\boldsymbol{W}}_0^{-1} \cdot \boldsymbol{G}_K(\gamma,\theta).$$

**Step II.** Compute the best weighting matrix and the best estimator

$$\hat{\boldsymbol{D}}_{(K+1)\times(K+1)} := \frac{1}{N}\sum_{i=1}^N g_K(T_i, \boldsymbol{Z}_i; \check{\gamma}, \check{\theta}) g_K(T_i, \boldsymbol{Z}_i; \check{\gamma}, \check{\theta})^\top,$$

$$(\widehat{\gamma}, \widehat{\theta}) = \arg \min_{\gamma\in\Gamma,\theta\in\Theta} \boldsymbol{G}_K(\gamma,\theta)^T \cdot \widehat{\boldsymbol{D}}^{-1}_{(K+1)\times(K+1)} \cdot \boldsymbol{G}_K(\gamma,\theta),$$

respectively. Hansen (1982) establishes that, for every fixed $K \geq p$,

$$\boldsymbol{V}_K^{-1/2} \begin{pmatrix} \sqrt{N}(\widehat{\gamma} - \gamma_0) \\ \sqrt{N}(\widehat{\theta} - \theta_0) \end{pmatrix} \to N\left(0, I_{(p+1)\times(p+1)}\right) \text{ in distribution.} \tag{2.7}$$

Moreover, denote

$$\widehat{\boldsymbol{B}}_{(K+1)\times(p+1)} := \begin{pmatrix} N^{-1}\sum_{i=1}^N u_K(\boldsymbol{X}_i) \frac{\nabla_\gamma \pi(\boldsymbol{Z}_i;\widehat{\gamma})^\top}{\pi(\boldsymbol{Z}_i;\widehat{\gamma})}, \mathbf{0}_{K\times 1} \\ N^{-1}\sum_{i=1}^N U(\boldsymbol{Z}_i) \frac{\nabla_\gamma \pi(\boldsymbol{Z}_i;\widehat{\gamma})^\top}{\pi(\boldsymbol{Z}_i;\widehat{\gamma})}, \quad 1 \end{pmatrix}$$

and

$$\widehat{\boldsymbol{V}}_K := \left\{ \left(\widehat{\boldsymbol{B}}_{(K+1)\times(p+1)}\right)^\top \widehat{\boldsymbol{D}}^{-1}_{(K+1)\times(K+1)} \left(\widehat{\boldsymbol{B}}_{(K+1)\times(p+1)}\right) \right\}^{-1}.$$

Hansen (1982) proves that, for every fixed $K \geq p$,

$$\widehat{\boldsymbol{V}}_K \to \boldsymbol{V}_K \text{ in probability.} \tag{2.8}$$

**Remark 1.** Despite the popularity and theoretical appeal of the inverse propensity score weighting, a major practical weakness of this method is the parameterization of the propensity score. It is well established that a slight misspecification of the propensity score function can lead to substantial bias (Kang and Schafer, 2007). In practice, applied researchers can apply diagnostic procedures, such as the "propensity score tautology" proposed in Imai, King and Stuart (2008), to select a particular parametric function from a prespecified set of functions. This procedure selects a particular propensity score if it balances the covariates. Specifically, for a fixed $K \in \mathbb{N}$ larger than $p$, Hansen's $J$-statistic,

$$J = N \left\{ \frac{1}{N} \boldsymbol{G}_K(\hat{\gamma}, \hat{\theta})^\top \hat{\boldsymbol{D}}^{-1}_{(K+1)\times(K+1)} \frac{1}{N} \boldsymbol{G}_K(\hat{\gamma}, \hat{\theta}) \right\} \xrightarrow{d} \chi^2_{K-p},$$

can be employed to test the null hypothesis of the propensity score being correctly specified. If it is correctly specified, the deviation of this statistic from zero should be within the range of the sampling error. For details, see Kosuke and Marc (2015). For an assessment of other models used in missing data analyses, see Ibrahim and Molenberghs (2009).

**Remark 2.** There are two approaches to dealing with non-ignorable missing data in a semiparametric estimation: the moment-based approach, and the empirical likelihood approach. Morikawa and Kim (2018) establish the equivalence between the empirical likelihood estimator (Owen (2004)) and the moment-based estimator. To describe the empirical likelihood estimation for our model, suppose that the first $n$ $Y_i$ are observed, and the remaining $(N-n)$ $Y_i$ are missing; that is, $T_i = 1$ for $i = 1, \ldots, n$, and $T_i = 0$ for $i = n+1, \ldots, N$. Qin, Leung and Shao (2002) construct the likelihood using the data with $T_i = 1$, as follows:

$$\prod_{i=1}^{n} \pi(\boldsymbol{Z}_i; \gamma) dF(\boldsymbol{Z}_i) \prod_{i=n+1}^{N} \int \{1 - \pi(\boldsymbol{z}; \gamma)\} dF(\boldsymbol{z}), \tag{2.9}$$

and discretize the distribution $F$ by $w_i$ $(i = 1, \ldots, n)$. The discretized distribution $w_i$ can be estimated by maximizing $\prod_{i=1}^{n} w_i$, subject to the following

constraints:

$$\begin{cases} w_i \geq 0,\ \sum_{i=1}^n w_i = 1,\ \sum_{i=1}^n w_i\{\pi(\boldsymbol{Z}_i;\gamma) - \mu_T\} = 0, \\ \sum_{i=1}^n w_i\{u_K(\boldsymbol{X}_i) - \bar{u}_K\} = 0,\ \sum_{i=1}^n w_i\{U(\boldsymbol{Z}_i) - \theta\} = 0, \end{cases}$$

where $\mu_T := E[T] = \int \pi(\boldsymbol{z};\gamma)dF(\boldsymbol{z})$ and $\bar{u}_K := N^{-1}\sum_{i=1}^N u_K(\boldsymbol{X}_i)$. The approximating basis functions $u_K(\boldsymbol{X})$ increase the estimation efficiency. With $\lambda_1$, $\lambda_2$, and $\lambda_3$ as Lagrange multipliers, the solution to the above optimization problem is $\hat{w}_i^{-1} = n[1+\lambda_1^\top\{u_K(\boldsymbol{X}_i)-\bar{u}_K\}+\lambda_2\{U(\boldsymbol{Z}_i)-\theta\}+\lambda_3\{\pi(\boldsymbol{Z}_i;\gamma)-\mu_T\}]$. Profiling out the unknown $F$ using the estimates $\hat{w}_i$ $(i = 1,\ldots,n)$ in (2.9) and taking the logarithm, we obtain the profile pseudo-loglikelihood

$$\begin{aligned} \ell(\gamma,\theta,\mu_T,\lambda_1,\lambda_2) &= \sum_{i=1}^n \log\pi(\boldsymbol{Z}_i;\gamma) \\ &\quad - \sum_{i=1}^n \log[1+\lambda_1^\top\{u_K(\boldsymbol{X}_i)-\bar{u}_K\}+\lambda_2\{U(\boldsymbol{Z}_i)-\theta\}+\lambda_3\{\pi(\boldsymbol{Z}_i;\gamma)-\mu_T\}] \\ &\quad + (N-n)\log(1-\mu_T), \end{aligned} \tag{2.10}$$

where $\lambda_3 = (N/n-1)/(1-\mu_T)$. Morikawa and Kim (2018) show that the empirical likelihood estimator is the same as the GMM estimator. Thus, we choose the GMM method, owing to its computational simplicity.

The best GMM estimator (within the class defined by the specific unconditional moments) is, in general, not semiparametrically efficient. To obtain an efficient estimator, we allow $K$ to increase with the sample size at the rate $o(N^{1/3})$, such that $\{u_K(\boldsymbol{X})\}$ spans the space of measureable functions (see also Geman and Hwang (1982) and Newey (1997)). In the next two sections, we establish that the results in (2.5)–(2.8) still hold with increasing $K = o(N^{1/3})$.

An advantage of the proposed estimator over existing estimators is that it does not require that we estimate $f_{Y|\boldsymbol{X},T}(y|\boldsymbol{x},1)$, relying instead on the moment conditions. Because the number of the unknown parameters is fixed and finite, and is independent of the number of covariates, the proposed estimator is always consistent, as long as the number of moment conditions exceeds the number of the unknown parameters (i.e., $K \geq p$). Further increasing the moment conditions only improves efficiency. Therefore, the classical trade-off between bias and variance in nonparametric estimations does not apply here. This is in constrast to the estimators proposed by Riddles, Kim and Im (2016) and Morikawa and Kim (2016), which do require an estimation of $f_{Y|\boldsymbol{X},T}(y|\boldsymbol{x},1)$. The estimator of

Morikawa and Kim (2016) is consistent, even if the parametric specification of $f_{Y|\boldsymbol{X},T}(y|\boldsymbol{x},1)$ is incorrect, but is inefficient. If $f_{Y|\boldsymbol{X},T}(y|\boldsymbol{x},1)$ is estimated nonparametrically (e.g., kernel estimation), the resulting estimator suffers from the curse of dimensionality and the bandwidth selection problem.

## 3. Asymptotic Theory

In this section, we show that the results in (2.5)–(2.7) still hold with increasing $K$. All technical proofs can be found in the Supplementary Material (Ai, Linton and Zhang (2018)). First, we establish the convergence rate of the first-step estimator $(\check{\gamma},\check{\theta})$.

**Theorem 1.** *Under Assumptions* 2.1–2.2′ *and Assumptions* 1, 2, 4, 6, *and* 7 *listed in the Appendix, with* $K = o(N^{1/3})$, *the first-step estimator satisfies* $(\check{\gamma} - \gamma_0, \check{\theta} - \theta_0) = O_p\left(N^{-1/2}\right)$.

Next, we establish the large-sample properties of the infeasible best estimator $(\overline{\gamma},\overline{\theta})$, without imposing the smoothness Assumptions 3 and 5 listed in the Appendix.

**Theorem 2.** *Under Assumptions* 2.1–2.2′ *and Assumptions* 1, 2, 4, 6, *and* 7 *listed in the Appendix, with* $K = o(N^{1/3})$, *the infeasible best estimator satisfies*

$$\boldsymbol{V}_K^{-1/2}\begin{pmatrix}\sqrt{N}(\overline{\gamma}-\gamma_0)\\ \sqrt{N}(\overline{\theta}-\theta_0)\end{pmatrix} \to N\left(0, I_{(p+1)\times(p+1)}\right) \text{ in distribution.}$$

If, in addition, the smoothness Assumptions 3 and 5 are satisfied, the next result shows that $\boldsymbol{V}_K \to \boldsymbol{V}_{eff}$, where $\boldsymbol{V}_{eff} := E[\boldsymbol{S}_{eff}\boldsymbol{S}_{eff}^\top]^{-1}$ is the semiparametric efficiency bound of $(\gamma_0,\theta_0)$ derived in Morikawa and Kim (2016), $\boldsymbol{S}_{eff} = (\boldsymbol{S}_1^\top, S_2)^\top$, and $\boldsymbol{S}_1$, $S_2$ are defined in (A.1) and (A.2), respectively.

**Theorem 3.** *Under Assumptions* 2.1–2.2′ *and Assumptions* 1–7 *listed in the Appendix, with* $K = o(N^{1/3})$, *we obtain* $\boldsymbol{V}_K \to \boldsymbol{V}_{eff}$.

By Theorem 1–3, the infeasible best estimator attains the semiparametric efficiency bound. The next result establishes the equivalence between the best estimator $(\widehat{\gamma},\widehat{\theta})$ and the infeasible best estimator $(\overline{\gamma},\overline{\theta})$, implying that the best estimator also attains the semiparametric efficiency bound.

**Theorem 4.** *Under Assumptions* 2.1–2.2′ *and Assumptions* 1–7 *listed in the Appendix, with* $K = o(N^{1/3})$, *we obtain* $(\sqrt{N}(\overline{\gamma}-\widehat{\gamma}), \sqrt{N}(\overline{\theta}-\widehat{\theta})) = o_p(1)$.

## 4. Variance Estimation

In order to conduct a statistical inference, we need a consistent covariance estimator. Note that (2.5) implies that $\widehat{\boldsymbol{V}}_K$ is a consistent estimator of $\boldsymbol{V}_K$, for every fixed $K \geq p$. We now show that this result still holds with increasing $K$, thereby providing a consistent covariance estimator.

**Theorem 5.** *Under Assumptions* 2.1–2.2′ *and Assumptions* 1–7 *listed in the Appendix, with* $K = o(N^{1/3})$*, we obtain* $\widehat{\boldsymbol{V}}_K \to \boldsymbol{V}_K$ *in probability.*

Note that our covariance estimator is much simpler and more natural than that suggested by Morikawa and Kim (2016), which requires a nonparametric estimation of $f_{Y|\boldsymbol{X},T}(y|x,1)$, and tends to exhibit poor performance in finite samples. Our covariance estimator is a GMM covariance estimator and is easily computed using existing statistical packages.

## 5. Selection of $K$

The large-sample properties of the proposed estimator established in the previous sections allow for a wide range of values for $K$. As a result, theoretically, the sensitivity of the estimator to the choice of $K$ is not as pronounced, affecting higher-order terms in a way that does not affect the consistency and asymptotic normality. Nevertheless, there may be some higher-order effect of the choice of $K$ on perfomance. In this section, we present two data-driven approaches to select $K$.

**Covariate balancing approach**. The first approach attempts to balance the distribution of the covariates between the whole population and the nonmissing population using weighting. Note that

$$E\left[\frac{T}{\pi(\boldsymbol{Z};\gamma_0)}I(X_j \leq x_j)\right] = E[I(X_j \leq x_j)],\ j \in \{1,\ldots,r\},$$

where $X_j$ is the $j^{\text{th}}$ component of $\boldsymbol{X}$, and $I(X_j \leq x_j)$ is the indicator function. Obviously, the propensity score function $\pi(\boldsymbol{Z};\gamma_0)$ plays the role of balancing. Note that the estimator $\hat{\gamma}$ depends on $K$. For a given $K$, we compute

$$\hat{F}^j_{N,K}(x_j) := \frac{1}{N}\sum_{i=1}^{N}\frac{T_i}{\pi(\boldsymbol{X}_i;\hat{\gamma})}I(X_{ij} \leq x_j),\ j \in \{1,\ldots,r\}.$$

We compute the empirical distributions of the covariates

$$\tilde{F}_N^j(x_j) := \frac{1}{N}\sum_{i=1}^{N} I(X_{ij} \le x_j),\ j \in \{1,\dots,r\}.$$

We choose the lowest $K$ such that the difference between $\{\hat{F}_{N,K}^j\}_{j=1}^r$ and $\{\hat{F}_N^j\}_{j=1}^r$ is small. Denote the upper bound of $K$ by $\bar{K}$ (e.g., $\bar{K}=7$ in our simulation studies). We choose $K \in \{1,\dots,\bar{K}\}$ to minimize the aggregate Kolmogorov–Smirnov distance between $\{\hat{F}_{N,K}^j\}_{j=1}^r$ and $\{\hat{F}_N^j\}_{j=1}^r$:

$$\hat{K} = \arg\min_{K\in\{1,\dots,\bar{K}\}} D_N(K) := \sum_{j=1}^{r} \sup_{x_j\in\mathbb{R}} \left|\tilde{F}_N^j(x_j) - \hat{F}_{N,K}^j(x_j)\right|.$$

**Higher-order mean squared error (MSE) approach**. The second approach chooses $K$ to minimize the MSE of the estimator. Donald, Imbens and Newey (2009) derive the higher-order asymptotic MSE of a linear combination $\mathbf{t}^\top\hat{\gamma}$, for some fixed $\mathbf{t}\in\mathbb{R}^p$. Let $\check{\gamma}$ be some preliminary estimator. Define:

$$\begin{aligned}
\widehat{\Pi}(K;\boldsymbol{t}) &= \sum_{i=1}^{N} \hat{\xi}_{ii}\rho(T_i,\boldsymbol{X}_i,Y_i;\check{\gamma})\cdot(\mathbf{t}^\top\hat{\Omega}_{p\times p}^{-1}\tilde{\eta}_i),\\
\hat{\Phi}(K;\boldsymbol{t}) &= \sum_{i=1}^{N} \hat{\xi}_{ii}\left\{\mathbf{t}^\top\hat{\Omega}_{p\times p}^{-1}\left[\widehat{\mathbf{D}}_i^*\rho(T_i,\boldsymbol{X}_i,Y_i;\check{\gamma})^2 - \nabla_\gamma\rho(T_i,\boldsymbol{X}_i,Y_i;\check{\gamma})\right]\right\}^2\\
&\quad - \mathbf{t}^\top\hat{\Omega}_{p\times p}^{-1}(\hat{\Gamma}_{K\times p})^\top\hat{\Upsilon}_{K\times K}^{-1}\hat{\Gamma}_{K\times p}\hat{\Omega}_{p\times p}^{-1}\boldsymbol{t},
\end{aligned}$$

where $\rho(T_i,\boldsymbol{X}_i,Y_i;\check{\gamma})$, $\hat{\Omega}_{p\times p}$, $\tilde{\eta}_i$, $\hat{\xi}_{ii}$, $\hat{\boldsymbol{D}}_i^*$, $\hat{\Gamma}_{K\times p}$, and $\hat{\Upsilon}_{K\times K}$ are defined in Section A of the Appendix. Note that $\widehat{\Pi}(K;\boldsymbol{t})^2/N$ is an estimate of the squared bias term derived in Newey and Smith (2004), and $\hat{\Phi}(K;\boldsymbol{t})$ is the asymptotic variance. The second approach chooses $K$ to minimize the following higher-order MSEs of $\hat{\gamma}_j$, for $j=1,\dots,p$:

$$S_{GMM}(K) = \sum_{j=1}^{p}\left\{\frac{1}{N}\widehat{\Pi}(K;e_j)^2 + \hat{\Phi}(K;e_j)\right\}, \tag{5.1}$$

where $e_j$ is the $j^{\text{th}}$ column of the $p$-dimensional identity matrix. In practice, we set the upper bound $\bar{K}$, and then choose $K\in\{1,2,\dots,\bar{K}\}$ to minimize the criteria defined in (5.1) .

## 6. Simulations

After establishing the large-sample properties of the proposed estimator, we now evaluate its finite-sample performance using a small-scale simulation study. We consider four scenarios. In all scenarios, the parameter of interest is $\theta_0 = E[Y]$, and the sample size is set to $N = 200$, 500, and $1,000$.

- **Scenario I**: $X$ is generated from the normal distribution $N(0,1)$, and the outcome $Y$ is generated from the normal distribution with mean $X+1$ and unit variance; that is, $Y \sim N(X+1,1)$. The relationship between the outcome variable and the covariate is linear, and the distribution of the outcome is normal. The missing mechanism is modeled by $P(T=1|Y,X) = [1+\exp(\alpha_0+\beta_0 Y)]^{-1}$, with the true value $(\alpha_0,\beta_0) = (0,-1.2)$. The true value of the parameter of interest is $\theta_0 = E[Y] = 1$.

- **Scenario II**: $X$ is generated from the normal distribution $N(0,1)$, and the outcome $Y$ is generated from the normal distribution with mean $X^2+1$ and unit variance; that is, $Y \sim N(X^2+1,1)$. Thus, the relationship between the outcome variable and the covariate is nonlinear, and the distribution of the outcome is nonnormal. The missing mechanism is modeled as $P(T=1|Y,X) = [1+\exp(\alpha_0+\beta_0 Y)]^{-1}$, with the true value $(\alpha_0,\beta_0) = (1.25,-1.2)$. The true value of the parameter of interest is $\theta_0 = E[Y] = 2$.

- **Scenario III**. The design follows Qin, Leung and Shao (2002). We generate the outcome from $Y = 0.1X^2 + ZX^{1/2}/5$, where $Z$ and $X$ are independent, $Z$ is a standard normal random variable, and $X$ follows the $\chi^2_{(6)}/2$ distribution. The missing mechanism is modeled as $P(T=1|Y,X) = [1+\exp(\alpha_0+\beta_0 Y)]^{-1}$, with the true value $(\alpha_0,\beta_0) = (3,-1)$. The true value of the target parameter is $\theta_0 = E[Y] = 1.2$.

- **Scenario IV**. The design is similar to that in Kang and Schafer (2007). $\boldsymbol{Z} = (Z_1, Z_2)$ is generated from the standard bivariate normal distribution, and $Y$ is generated from the normal distribution with mean $2+Z_1$ and unit variance. The missing mechanism is modeled as $P(T=1|Y,X_1,X_2) = [1+\exp(\alpha_0 Z_1+\beta_0 Y)]^{-1}$, with $(\alpha_0,\beta_0) = (1,-1)$. The true value of the parameter of interest is $\theta_0 = E[Y] = 2$. Instead of observing the covariates $\boldsymbol{Z}$ directly, we observe a nonlinear transformation of $\boldsymbol{Z}$: $X_1 = \exp(Z_1/2)$ and $X_2 = Z_2/(1+\exp(Z_1))$.

In all scenarios, we generate $J = 500$ random samples, and for each sample, we compute the following three estimators:

1. Naive estimator. We compute the MAR estimator $(\tilde{\alpha}_{MAR}, \tilde{\beta}_{MAR}, \tilde{\theta}_{MAR})$ as $\tilde{\theta}_{MAR} = N^{-1}\sum_{i=1}^{N} T_i Y_i / \pi(\boldsymbol{X}_i; \tilde{\alpha}_{MAR}, \tilde{\beta}_{MAR})$, where $\pi(\boldsymbol{X}_i; \tilde{\alpha}_{MAR}, \tilde{\beta}_{MAR})$ is an estimated response model. In Scenarios I, II, and III, $\pi(\boldsymbol{X}_i; \tilde{\alpha}_{MAR}, \tilde{\beta}_{MAR}) = [1+\exp(\tilde{\alpha}_{MAR}+\tilde{\beta}_{MAR}X_i)]^{-1}$, and in Scenario IV, $\pi(\boldsymbol{X}_i; \tilde{\alpha}_{MAR}, \tilde{\beta}_{MAR}) = \left[1+\exp(\tilde{\alpha}_{MAR}Z_{1i}+\tilde{\beta}_{MAR}X_{2i})\right]^{-1}$, where $(\tilde{\alpha}_{MAR}, \tilde{\beta}_{MAR})$ are estimated using the GMM.

2. MK2 estimator. We compute $(\hat{\alpha}_{MK}, \hat{\beta}_{MK}, \hat{\theta}_{MK})$ using the approach of Morikawa and Kim (2016); that is, $(\hat{\alpha}_{MK}, \hat{\beta}_{MK}, \hat{\theta}_{MK})$ is the solution to

$$\sum_{i=1}^{N} \left(\hat{\boldsymbol{S}}_1(T_i, \boldsymbol{Z}_i; \alpha, \beta)^\top, \hat{S}_2(T_i, \boldsymbol{Z}_i; \alpha, \beta, \theta)\right)^\top = 0,$$

where

$$\hat{\boldsymbol{S}}_1(T, \boldsymbol{Z}; \alpha, \beta) = -\left(1 - \frac{T}{\pi(\boldsymbol{Z}; \alpha, \beta)}\right) E^\star\left[\frac{\nabla_\gamma \pi(\boldsymbol{Z}; \alpha, \beta)}{1-\pi(\boldsymbol{Z}; \alpha, \beta)}\middle| \boldsymbol{X}\right],$$
$$\hat{S}_2(T, \boldsymbol{Z}; \alpha, \beta, \theta) = -\frac{T}{\pi(\boldsymbol{Z}; \alpha, \beta)} U(\boldsymbol{Z}) + \theta - \left(1 - \frac{T}{\pi(\boldsymbol{Z}; \alpha, \beta)}\right) E^\star\left[U(\boldsymbol{Z})|\boldsymbol{X}\right],$$

and for any function $g(\boldsymbol{Z})$, the quantity $E^\star[g(\boldsymbol{Z})|\boldsymbol{X}]$ is defined by

$$E^\star[g(\boldsymbol{Z})|\boldsymbol{X}=\boldsymbol{x}] := \frac{\sum_{j=1}^{N} T_j K_h(\boldsymbol{x}-\boldsymbol{X}_j) T_j \pi(\boldsymbol{Z}_j; \alpha, \beta)^{-1} O(\boldsymbol{x}, Y_j; \alpha, \beta) g(\boldsymbol{x}, Y_j)}{\sum_{j=1}^{N} K_h(\boldsymbol{x}-\boldsymbol{X}_j) T_j \pi(\boldsymbol{Z}_j; \alpha, \beta)^{-1} O(\boldsymbol{x}, Y_j; \alpha, \beta)},$$

where $O(\boldsymbol{z}; \alpha, \beta) = \dfrac{1-\pi(\boldsymbol{z}; \alpha, \beta)}{\pi(\boldsymbol{z}; \alpha, \beta)}$, $K_h(\boldsymbol{x}-\boldsymbol{w}) = K\left(\boldsymbol{x} - \dfrac{\boldsymbol{w}}{h}\right)$,

$K(\cdot)$ is a Gaussian kernel function, and $h$ is the bandwidth. Because Morikawa and Kim (2016) do not describe how to select the bandwidth, we choose $h = 0.1$ in Scenarios I, II, and III, and $h = 0.2$ in Scenario IV (the numeric computation fails if $h = 0.1$ in Scenario IV, perhaps due to overfitting in the multivariate case).

3. WSK Eestimator. We compute $(\hat{\alpha}_{WSK}, \hat{\beta}_{WSK}, \hat{\theta}_{WSK})$ using the approach of Wang, Shao and Kim (2014); that is, $(\hat{\alpha}_{WSK}, \hat{\beta}_{WSK}, \hat{\theta}_{WSK})$ is a GMM estimator from the following moments:

$$E\left[\begin{array}{c} \frac{T}{\pi(\boldsymbol{X}, Y; \gamma)} - 1 \\ \left\{\frac{T}{\pi(\boldsymbol{X}, Y; \gamma)} - 1\right\} \boldsymbol{X} \\ \frac{T}{\pi(\boldsymbol{X}, Y; \gamma)} Y - \theta \end{array}\right] = 0.$$

4. The proposed GMM estimator. We compute $(\hat{\alpha}, \hat{\beta}, \hat{\theta})$ using the proposed approach and the covariate balancing approach to select $K$, with $\bar{K} = 7$ in Scenarios I, II, and III, and $\bar{K} = 10$ in Scenario IV. Here, $\bar{K}$ is the maximal number of candidate moments to be considered.

The simulation results show the bias, standard deviation (Stdev), MSE, and coverage probability (CP) (for significance level $\alpha = 0.05$) of the point estimates for the four scenarios; see Tables 1, 2, 3, and 4, respectively. A histogram of the selected $K$ (based on 500 Monte Carlo samples) in all scenarios is reported in Figure 1. The results are as follows:

1. In all scenarios, the naive estimator using the MAR assumption has a large bias, because this assumption does not hold.

2. In all scenarios, the proposed estimator of $E[Y]$ outperforms the MK estimator.

3. In all scenarios, the proposed estimator of $E[Y]$ is always consistent, in the sense that its bias decreases to zero as the sample size increases. However, in Scenarios II and IV, the WSK estimator has a large bias that does not decrease to zero as the sample size increases.

4. In all scenarios, the proposed estimators of the nuisance parameters $\alpha_0$ and $\beta_0$ in the response model are always consistent. The MK estimators of the nuisance parameters demonstrate reasonable performance in Scenarios I, III, and IV, but have a large bias in Scenario II. The WSK estimators of the nuisance parameters exhibit relatively poor performance in Scenarios II and IV.

5. In all scenarios, the proposed variance estimator has coverage probability close to 95%, even when the sample size is small. The MK variance estimator performs well in Scenario IV, but poorly in the other scenarios: in Scenario I, the coverage probability using the MK approach converges to 90%, rather than 95%; in Scenario II, the CP values are far from 95% in Scenario 2, regardless of the sample size; in Scenario III, the MK variance estimator is consistent only when the sample size is large.

6. When the sample size is small, the optimal $K$ tends to be two, with large probability. When the sample size is large, the optimal $K$ tends to be three, with large probability. The growth rate of $K$ is extremely slow compared

Table 1. Simulation results under Scenario I.

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | $N = 200$ | | | | | | |
| | $\hat{\alpha}$ | $\hat{\beta}$ | $\hat{\theta}$ | $\hat{\alpha}_{MK}$ | $\hat{\beta}_{MK}$ | $\hat{\theta}_{MK}$ | $\hat{\alpha}_{WSK}$ | $\hat{\beta}_{WSK}$ | $\hat{\theta}_{WSK}$ | $\tilde{\alpha}_{MAR}$ | $\tilde{\beta}_{MAR}$ | $\tilde{\theta}_{MAR}$ |
| Bias | 0.028 | -0.125 | 0.039 | 0.040 | 0.202 | 0.134 | -0.140 | -0.448 | -0.031 | -0.997 | 0.167 | 0.301 |
| Stdev | 0.254 | 0.413 | 0.129 | 0.229 | 0.256 | 0.118 | 0.910 | 1.183 | 0.150 | 0.197 | 0.266 | 0.101 |
| MSE | 0.065 | 0.186 | 0.018 | 0.054 | 0.106 | 0.032 | 0.849 | 1.601 | 0.023 | 1.033 | 0.099 | 0.101 |
| CP | — | — | 0.906 | — | — | 0.860 | — | — | 0.946 | — | — | 0.220 |
| | | | | | | $N = 500$ | | | | | | |
| | $\hat{\alpha}$ | $\hat{\beta}$ | $\hat{\theta}$ | $\hat{\alpha}_{MK}$ | $\hat{\beta}_{MK}$ | $\hat{\theta}_{MK}$ | $\hat{\alpha}_{WSK}$ | $\hat{\beta}_{WSK}$ | $\hat{\theta}_{WSK}$ | $\tilde{\alpha}_{MAR}$ | $\tilde{\beta}_{MAR}$ | $\tilde{\theta}_{MAR}$ |
| Bias | 0.011 | -0.067 | 0.016 | 0.050 | 0.097 | 0.083 | -0.029 | -0.147 | -0.014 | -0.966 | 0.220 | 0.299 |
| Stdev | 0.161 | 0.282 | 0.090 | 0.152 | 0.173 | 0.075 | 0.207 | 0.403 | 0.108 | 0.126 | 0.160 | 0.063 |
| MSE | 0.026 | 0.084 | 0.008 | 0.025 | 0.039 | 0.013 | 0.044 | 0.184 | 0.012 | 0.949 | 0.074 | 0.093 |
| CP | — | — | 0.928 | — | — | 0.844 | — | — | 0.948 | — | — | 0.034 |
| | | | | | | $N = 1,000$ | | | | | | |
| | $\hat{\alpha}$ | $\hat{\beta}$ | $\hat{\theta}$ | $\hat{\alpha}_{MK}$ | $\hat{\beta}_{MK}$ | $\hat{\theta}_{MK}$ | $\hat{\alpha}_{WSK}$ | $\hat{\beta}_{WSK}$ | $\hat{\theta}_{WSK}$ | $\tilde{\alpha}_{MAR}$ | $\tilde{\beta}_{MAR}$ | $\tilde{\theta}_{MAR}$ |
| Bias | 0.005 | -0.040 | 0.008 | 0.041 | 0.048 | 0.053 | -0.002 | -0.057 | -0.002 | -0.962 | 0.235 | 0.298 |
| Stdev | 0.103 | 0.187 | 0.065 | 0.103 | 0.128 | 0.055 | 0.1118 | 0.218 | 0.073 | 0.078 | 0.099 | 0.045 |
| MSE | 0.010 | 0.036 | 0.004 | 0.012 | 0.018 | 0.006 | 0.012 | 0.051 | 0.005 | 0.932 | 0.065 | 0.091 |
| CP | — | — | 0.934 | — | — | 0.864 | — | — | 0.956 | — | — | 0.012 |

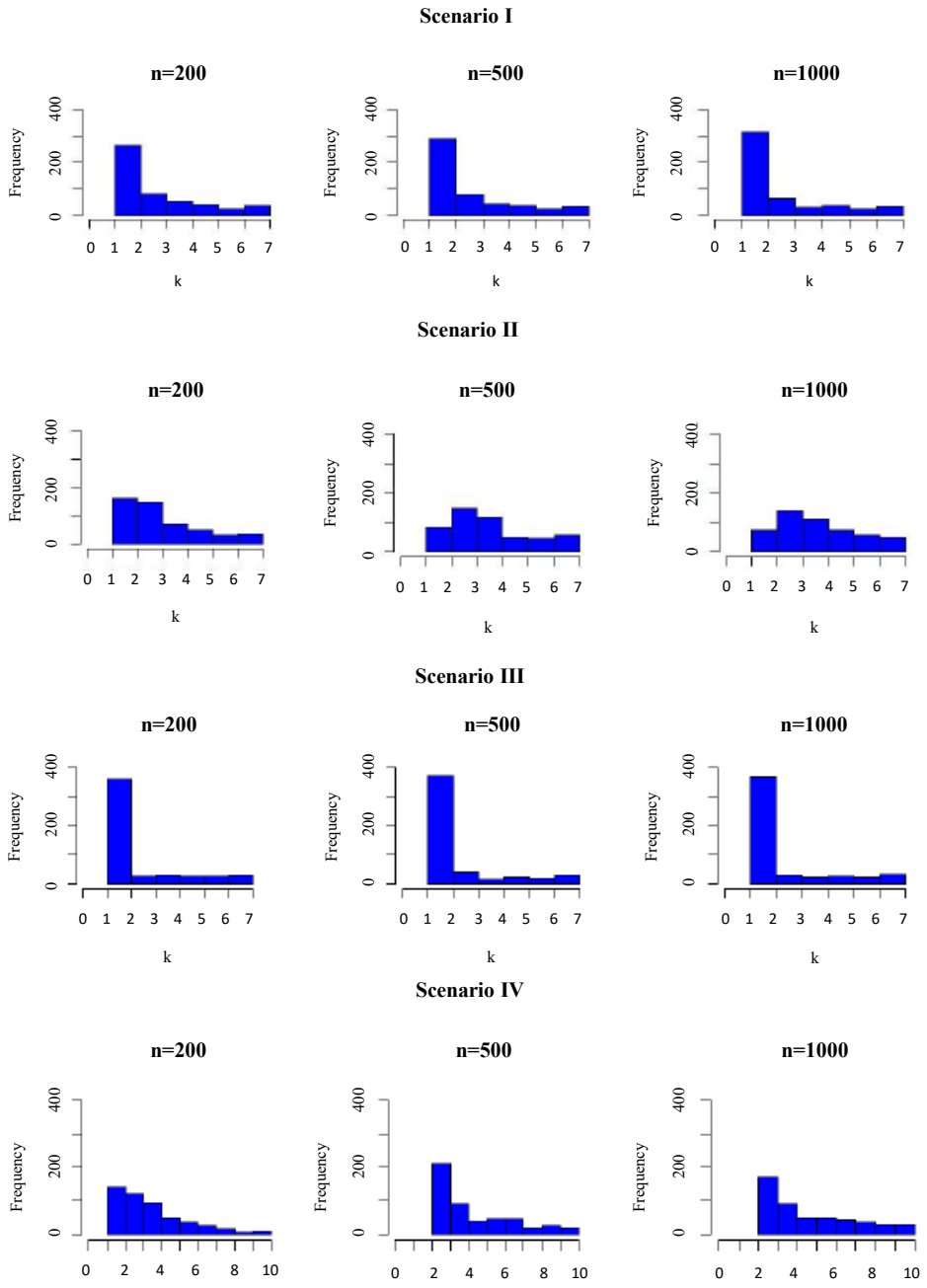
Stdev: standard deviation; MSE: mean squared error; CP: coverage probability. The bandwidth used to compute the nonparametric kernel estimators $(\hat{\alpha}_{MK}, \hat{\beta}_{MK}, \hat{\theta}_{MK})$ is $h = 0.1$.

with that of the sample size $N$, which is consistent with our theoretical Assumption 7.

These results clearly show that the proposed approach exhibits better finite-sample performance than that of its competitors.

## 7. Discussion

Many practical applications suffer from the data MNAR problem. Morikawa and Kim (2016) propose two efficient estimators for a class of MNAR problems, where they parametrically and nonparametrically, respectively, estimate $f_{Y|\boldsymbol{X},T}(y|\boldsymbol{x}, 1)$. If the working model of $f_{Y|\boldsymbol{X},T}(y|\boldsymbol{x}, 1)$ is misspecified, the parametric estimator is consistent, but not efficient. In this study, we examine the same class of MNAR problems, but present a much simpler and more natural efficient estimator. Our approach is based on a parametric moment restriction model that does not require a nonparametric estimation and, hence, does not suffer from the curse of dimensionality problem or the bandwidth selection problem. The simulation results confirm that the proposed approach outperforms its competitors in finite samples. The GMM approach is also easy to adapt to




The Monte Carlo sample size used to plot the histogram of $K$ is $J = 500$.

Figure 1. Histogram of $K$.

Table 2. Simulation results under Scenario II.

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | $N = 200$ | | | | | | |
| | $\hat{\alpha}$ | $\hat{\beta}$ | $\hat{\theta}$ | $\hat{\alpha}_{MK}$ | $\hat{\beta}_{MK}$ | $\hat{\theta}_{MK}$ | $\hat{\alpha}_{WSK}$ | $\hat{\beta}_{WSK}$ | $\hat{\theta}_{WSK}$ | $\tilde{\alpha}_{MAR}$ | $\tilde{\beta}_{MAR}$ | $\tilde{\theta}_{MAR}$ |
| Bias | -0.208 | 0.096 | 0.084 | -0.254 | 0.254 | 0.114 | -0.849 | -2.158 | 0.003 | -2.053 | 1.215 | 0.530 |
| Stdev | 0.646 | 0.555 | 0.201 | 0.381 | 0.252 | 0.134 | 8.939 | 16.509 | 0.367 | 0.809 | 0.148 | 0.205 |
| MSE | 0.462 | 0.318 | 0.047 | 0.210 | 0.128 | 0.031 | 80.632 | 277.217 | 0.134 | 4.873 | 1.498 | 0.323 |
| CP | — | — | 0.950 | — | — | 0.910 | — | — | 0.890 | — | — | 0.138 |
| | | | | | | $N = 500$ | | | | | | |
| | $\hat{\alpha}$ | $\hat{\beta}$ | $\hat{\theta}$ | $\hat{\alpha}_{MK}$ | $\hat{\beta}_{MK}$ | $\hat{\theta}_{MK}$ | $\hat{\alpha}_{WSK}$ | $\hat{\beta}_{WSK}$ | $\hat{\theta}_{WSK}$ | $\tilde{\alpha}_{MAR}$ | $\tilde{\beta}_{MAR}$ | $\tilde{\theta}_{MAR}$ |
| Bias | -0.081 | 0.040 | 0.044 | -0.119 | 0.140 | 0.073 | -0.648 | -1.532 | -0.077 | -1.924 | 1.203 | 0.583 |
| Stdev | 0.406 | 0.363 | 0.131 | 0.262 | 0.188 | 0.096 | 8.289 | 10.832 | 0.349 | 0.175 | 0.064 | 0.132 |
| MSE | 0.171 | 0.134 | 0.019 | 0.083 | 0.055 | 0.014 | 69.132 | 119.697 | 0.128 | 3.732 | 1.451 | 0.357 |
| CP | — | — | 0.932 | — | — | 0.894 | — | — | 0.910 | — | — | 0.06 |
| | | | | | | $N = 1,000$ | | | | | | |
| | $\hat{\alpha}$ | $\hat{\beta}$ | $\hat{\theta}$ | $\hat{\alpha}_{MK}$ | $\hat{\beta}_{MK}$ | $\hat{\theta}_{MK}$ | $\hat{\alpha}_{WSK}$ | $\hat{\beta}_{WSK}$ | $\hat{\theta}_{WSK}$ | $\tilde{\alpha}_{MAR}$ | $\tilde{\beta}_{MAR}$ | $\tilde{\theta}_{MAR}$ |
| Bias | -0.036 | 0.019 | 0.019 | -0.078 | 0.093 | 0.047 | -0.111 | -0.856 | -0.131 | -1.900 | 1.201 | 0.590 |
| Stdev | 0.260 | 0.225 | 0.086 | 0.184 | 0.142 | 0.068 | 1.092 | 1.547 | 0.311 | 0.086 | 0.044 | 0.078 |
| MSE | 0.069 | 0.051 | 0.007 | 0.040 | 0.029 | 0.007 | 1.206 | 3.127 | 0.114 | 3.618 | 1.445 | 0.354 |
| CP | — | — | 0.932 | — | — | 0.900 | — | — | 0.902 | — | — | 0.018 |

Stdev: standard deviation; MSE: mean squared error; CP: coverage probability. The bandwidth used to compute the nonparametric kernel estimators ($\hat{\alpha}_{MK}, \hat{\beta}_{MK}, \hat{\theta}_{MK}$) is $h = 0.1$.

stratified sampling and other sampling schemes common in survey data.

Both approaches require the correct parameterization of the propensity score function. If this function is misspecified, then both approaches yield inconsistent estimates. Several attempts have been made to resolve this problem. For instance, Zhao and Shao (2015) introduce a partial linear index to model the missing mechanism. The proposed approach can be extended in this direction, which we leave to future research.

## Supplementary Material

The online Supplementary Material contains technical proofs for Theorems 1, 2, 3, 4, and 5.

## Acknowledgments

We thank a referee, an associate editor, and the editor Yuan-chin Ivan Chang for their constructive comments. Shigeyuki Hamori, Lukang Huang, Shih-Hao Huang, Kaiji Motegi, Shuenn-ji Sheu, Shang-Yuan Shiu, and Li-Hsien Sun provided helpful comments. Chunrong Ai acknowledges financial support from the

Table 3. Simulation results under Scenario III.

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | $N = 200$ | | | | | | |
| | $\hat{\alpha}$ | $\hat{\beta}$ | $\hat{\theta}$ | $\hat{\alpha}_{MK}$ | $\hat{\beta}_{MK}$ | $\hat{\theta}_{MK}$ | $\hat{\alpha}_{WSK}$ | $\hat{\beta}_{WSK}$ | $\hat{\theta}_{WSK}$ | $\tilde{\alpha}_{MAR}$ | $\tilde{\beta}_{MAR}$ | $\tilde{\theta}_{MAR}$ |
| Bias | 0.155 | -0.171 | 0.003 | 0.047 | 0.015 | 0.071 | 0.184 | -0.205 | -0.005 | -2.794 | 0.954 | -1.146 |
| Stdev | 0.584 | 0.585 | 0.155 | 0.376 | 0.190 | 0.131 | 0.753 | 0.922 | 0.149 | 1.395 | 0.396 | 0.263 |
| MSE | 0.365 | 0.372 | 0.024 | 0.144 | 0.036 | 0.022 | 0.602 | 0.892 | 0.022 | 9.758 | 1.069 | 1.384 |
| CP | — | — | 0.934 | — | — | 0.884 | — | — | 0.942 | — | — | 0.032 |
| | | | | | | $N = 500$ | | | | | | |
| | $\hat{\alpha}$ | $\hat{\beta}$ | $\hat{\theta}$ | $\hat{\alpha}_{MK}$ | $\hat{\beta}_{MK}$ | $\hat{\theta}_{MK}$ | $\hat{\alpha}_{WSK}$ | $\hat{\beta}_{WSK}$ | $\hat{\theta}_{WSK}$ | $\tilde{\alpha}_{MAR}$ | $\tilde{\beta}_{MAR}$ | $\tilde{\theta}_{MAR}$ |
| Bias | 0.034 | -0.036 | 0.000 | 0.012 | 0.012 | 0.034 | 0.037 | -0.036 | -0.002 | 0.782 | 0.355 | 0.123 |
| Stdev | 0.305 | 0.224 | 0.103 | 0.250 | 0.128 | 0.085 | 0.304 | 0.216 | 0.100 | 0.433 | 0.113 | 0.101 |
| MSE | 0.094 | 0.051 | 0.010 | 0.062 | 0.016 | 0.008 | 0.094 | 0.048 | 0.010 | 0.799 | 0.139 | 0.025 |
| CP | — | — | 0.902 | — | — | 0.894 | — | — | 0.912 | — | — | 0.698 |
| | | | | | | $N = 1,000$ | | | | | | |
| | $\hat{\alpha}$ | $\hat{\beta}$ | $\hat{\theta}$ | $\hat{\alpha}_{MK}$ | $\hat{\beta}_{MK}$ | $\hat{\theta}_{MK}$ | $\hat{\alpha}_{WSK}$ | $\hat{\beta}_{WSK}$ | $\hat{\theta}_{WSK}$ | $\tilde{\alpha}_{MAR}$ | $\tilde{\beta}_{MAR}$ | $\tilde{\theta}_{MAR}$ |
| Bias | 0.009 | -0.010 | 0.002 | 0.002 | 0.009 | 0.017 | 0.007 | -0.008 | 0.001 | 0.728 | 0.372 | 0.126 |
| Stdev | 0.215 | 0.157 | 0.069 | 0.167 | 0.083 | 0.056 | 0.213 | 0.156 | 0.069 | 0.302 | 0.078 | 0.067 |
| MSE | 0.046 | 0.024 | 0.004 | 0.028 | 0.007 | 0.003 | 0.045 | 0.024 | 0.004 | 0.621 | 0.144 | 0.020 |
| CP | — | — | 0.932 | — | — | 0.934 | — | — | 0.93 | — | — | 0.454 |

Stdev: standard deviation; MSE: mean squared error; CP: coverage probability. The bandwidth used to compute the nonparametric kernel estimators $(\hat{\alpha}_{MK}, \hat{\beta}_{MK}, \hat{\theta}_{MK})$ is $h = 0.1$.

National Natural Science Foundation of China through project 71873138. Oliver Linton acknowledges Cambridge INET for their financial support. Zheng Zhang acknowledges the financial support provided by Renmin University of China through project 297517501221 and the fund for building world-class universities (disciplines) of Renmin University of China.

## Appendix

## A. Notation

The following notation are needed for presenting the efficiency bound:

$$O(\boldsymbol{Z}) := \frac{1-\pi(\boldsymbol{Z};\gamma_0)}{\pi(\boldsymbol{Z};\gamma_0)}, \quad \boldsymbol{S}_0(\boldsymbol{Z}) := -\frac{\nabla_\gamma \pi(\boldsymbol{Z};\gamma_0)}{1-\pi(\boldsymbol{Z};\gamma_0)},$$
$$m(\boldsymbol{X}) := \frac{E[O(\boldsymbol{Z})\boldsymbol{S}_0(\boldsymbol{Z})|\boldsymbol{X}]}{E[O(\boldsymbol{Z})|\boldsymbol{X}]}, \quad R(\boldsymbol{X}) := \frac{E[O(\boldsymbol{Z})U(\boldsymbol{Z})|\boldsymbol{X}]}{E[O(\boldsymbol{Z})|\boldsymbol{X}]},$$
$$\boldsymbol{S}_1(T,\boldsymbol{Z};\gamma_0) := \left(1-\frac{T}{\pi(\boldsymbol{Z};\gamma_0)}\right) m(\boldsymbol{X}), \tag{A.1}$$
$$S_2(T,\boldsymbol{Z};\gamma_0,\theta_0) := -\frac{T}{\pi(\boldsymbol{Z};\gamma_0)}U(\boldsymbol{Z}) + \theta_0 - \left(1-\frac{T}{\pi(\boldsymbol{Z};\gamma_0)}\right) R(\boldsymbol{X}). \tag{A.2}$$

Table 4. Simulation results under Scenario IV.

| | | | | | | $N = 200$ | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\hat{\alpha}$ | $\hat{\beta}$ | $\hat{\theta}$ | $\hat{\alpha}_{MK}$ | $\hat{\beta}_{MK}$ | $\hat{\theta}_{MK}$ | $\hat{\alpha}_{WSK}$ | $\hat{\beta}_{WSK}$ | $\hat{\theta}_{WSK}$ | $\tilde{\alpha}_{MAR}$ | $\tilde{\beta}_{MAR}$ | $\tilde{\theta}_{MAR}$ |
| Bias | 0.097 | -0.114 | 0.005 | -0.018 | 0.027 | 0.043 | 0.738 | -0.748 | -0.011 | -1.002 | 1.003 | 0.136 |
| Stdev | 1.140 | 0.721 | 0.118 | 0.308 | 0.185 | 0.103 | 3.241 | 3.079 | 0.134 | 0.081 | 0.139 | 0.348 |
| MSE | 1.310 | 0.533 | 0.014 | 0.095 | 0.035 | 0.013 | 11.055 | 10.046 | 0.0183 | 1.011 | 1.026 | 0.139 |
| CP | — | — | 0.914 | — | — | 0.920 | — | — | 0.882 | — | — | 0.998 |
| | | | | | | $N = 500$ | | | | | | |
| | $\hat{\alpha}$ | $\hat{\beta}$ | $\hat{\theta}$ | $\hat{\alpha}_{MK}$ | $\hat{\beta}_{MK}$ | $\hat{\theta}_{MK}$ | $\hat{\alpha}_{WSK}$ | $\hat{\beta}_{WSK}$ | $\hat{\theta}_{WSK}$ | $\tilde{\alpha}_{MAR}$ | $\tilde{\beta}_{MAR}$ | $\tilde{\theta}_{MAR}$ |
| Bias | -0.001 | -0.026 | 0.003 | -0.042 | 0.041 | 0.022 | 0.537 | -0.564 | -0.024 | -1.003 | 1.000 | 0.146 |
| Stdev | 0.203 | 0.139 | 0.071 | 0.172 | 0.100 | 0.067 | 1.759 | 1.773 | 0.111 | 0.048 | 0.088 | 0.199 |
| MSE | 0.041 | 0.020 | 0.005 | 0.031 | 0.011 | 0.005 | 3.384 | 3.464 | 0.0131 | 1.010 | 1.009 | 0.061 |
| CP | — | — | 0.944 | — | — | 0.946 | — | — | 0.842 | — | — | 1.000 |
| | | | | | | $N = 1,000$ | | | | | | |
| | $\hat{\alpha}$ | $\hat{\beta}$ | $\hat{\theta}$ | $\hat{\alpha}_{MK}$ | $\hat{\beta}_{MK}$ | $\hat{\theta}_{MK}$ | $\hat{\alpha}_{WSK}$ | $\hat{\beta}_{WSK}$ | $\hat{\theta}_{WSK}$ | $\tilde{\alpha}_{MAR}$ | $\tilde{\beta}_{MAR}$ | $\tilde{\theta}_{MAR}$ |
| Bias | 0.010 | -0.034 | -0.001 | -0.027 | 0.024 | 0.011 | 0.580 | -0.586 | -0.040 | -1.000 | 0.997 | 0.134 |
| Stdev | 0.262 | 0.264 | 0.052 | 0.122 | 0.070 | 0.048 | 1.565 | 1.509 | 0.107 | 0.035 | 0.065 | 0.148 |
| MSE | 0.068 | 0.070 | 0.002 | 0.015 | 0.005 | 0.002 | 2.787 | 2.623 | 0.013 | 1.003 | 1.000 | 0.039 |
| CP | — | — | 0.936 | — | — | 0.932 | — | — | 0.786 | — | — | 1.000 |

Stdev: standard deviation; MSE: mean squared error; CP: coverage probability. The bandwidth used to compute the nonparametric kernel estimators $(\hat{\alpha}_{MK}, \hat{\beta}_{MK}, \hat{\theta}_{MK})$ is $h = 0.2$.

The following notation are needed to describe the higher-order MSE criteria proposed by Donald, Imbens and Newey (2009) :

$$\rho(T_i, \boldsymbol{X}_i, Y_i; \check{\gamma}) = 1 - \frac{T_i}{\pi(\boldsymbol{X}_i, Y_i; \check{\gamma})}, \quad \hat{\Upsilon}_{K\times K} = \frac{1}{N}\sum_{i=1}^{N} \rho(T_i, \boldsymbol{X}_i, Y_i; \check{\gamma})^2 u_K(\boldsymbol{X}_i)^{\otimes 2},$$

$$\widehat{\Gamma}_{K\times p} = \frac{1}{N}\sum_{i=1}^{N} u_K(\boldsymbol{X}_i)\nabla_\gamma \rho(T_i, \boldsymbol{X}_i, Y_i; \check{\gamma})^\top, \quad \hat{\Omega}_{p\times p} = (\hat{\Gamma}_{K\times p})^\top \hat{\Upsilon}_{K\times K}^{-1} \hat{\Gamma}_{K\times p},$$

$$\tilde{\mathbf{d}}_i = (\hat{\Gamma}_{K\times p})^\top \left( \frac{1}{N}\sum_{j=1}^{N} u_K(\boldsymbol{X}_j)^{\otimes 2} \right)^{-1} u_K(\boldsymbol{X}_i), \quad \tilde{\eta}_i = \nabla_\gamma \rho(T_i, \boldsymbol{X}_i, Y_i; \check{\gamma}) - \tilde{\mathbf{d}}_i,$$

$$\hat{\xi}_{ij} = \frac{1}{N} u_K(\boldsymbol{X}_i)^\top \hat{\Upsilon}_{K\times K}^{-1} u_K(\boldsymbol{X}_j), \quad \widehat{\mathbf{D}}_i^* = (\hat{\Gamma}_{K\times p})^\top \hat{\Upsilon}_{K\times K}^{-1} u_K(\boldsymbol{X}_i).$$

## B. Assumptions

The following assumptions are maintained in this paper:

**Assumption 1.** *There exists a nonresponse instrumental variable* $\mathbf{X_2}$*, i.e.,* $\mathbf{X} =$

$(\mathbf{X}_1^\intercal, \mathbf{X}_2^\intercal)^\intercal$, *such that* $\boldsymbol{X}_2$ *is independent of* $T$, *given* $\boldsymbol{X}_1$ *and* $Y$; *furthermore,* $\boldsymbol{X}_2$ *is correlated with* $Y$.

**Assumption 2.** *The support of* $\mathbf{X}$, *which is denoted by* $\mathcal{X}$, *is a Cartesian product of* $r$*-compact intervals, and we denote* $\boldsymbol{X} = (X_1, \ldots, X_r)^\top$.

**Assumption 3.** $E[O(\boldsymbol{Z})S_0(\boldsymbol{Z})|\mathbf{X} = \boldsymbol{x}]$, $E[O(\boldsymbol{Z})U(\boldsymbol{Z})|\mathbf{X} = \boldsymbol{x}]$ *and* $E[O(\boldsymbol{Z})|\boldsymbol{X} = \boldsymbol{x}]$ *are* $s$*-smooth in* $\boldsymbol{x}$ *(the definition is given in page* 5569 *of Chen* (2007) *and Definition* 1.1 *of the Supplementary Material), where* $s > 0$.

**Assumption 4.** *There exist two finite positive constants* $\underline{a}$ *and* $\overline{a}$ *such that the smallest (resp. largest) eigenvalue of* $E[u_K(\boldsymbol{X})u_K^\top(\boldsymbol{X})]$ *is bounded away from* $\underline{a}$ *(resp.* $\overline{a}$*) uniformly in* $K$, *i.e.,*

$$0 < \underline{a} \leq \lambda_{\min}(E[u_K(\boldsymbol{X})u_K(\boldsymbol{X})^\top]) \leq \lambda_{\max}(E[u_K(\boldsymbol{X})u_K(\boldsymbol{X})^\top]) \leq \overline{a} < \infty \ .$$

**Assumption 5.** *(i) The parameter spaces* $\Gamma$ *and* $\Theta$ *are compact; (ii) The efficient score function* $\boldsymbol{S}_{eff}(T, \boldsymbol{Z}; \gamma, \theta) := (\boldsymbol{S}_1^\intercal(T, \boldsymbol{Z}; \gamma), S_2(T, \boldsymbol{Z}; \gamma, \theta))^\intercal$ *is continuously differentiable at each* $(\gamma, \theta) \in \Gamma \times \Theta$, *and* $\mathbb{E}\left[\partial \boldsymbol{S}_{eff}(\gamma, \theta)/\partial(\gamma^\top, \theta)\right]$ *is nonsingular at* $(\gamma_0, \theta_0)$.

**Assumption 6.** *(i) There exist two positive constants* $\bar{c}$ *and* $\underline{c}$ *such that* $0 < \underline{c} \leq \pi(\boldsymbol{x}, y; \gamma) \leq \bar{c} < 1$ *for all* $\gamma \in \Gamma$ *and* $(\boldsymbol{x}, y) \in \mathcal{X} \times \mathbb{R}$; *(ii)* $\pi(\boldsymbol{x}, y; \gamma)$ *is twice continuously differentiable in* $\gamma \in \Gamma$, *and the derivatives are uniformly bounded.*

**Assumption 7.** *Suppose* $K \to \infty$ *and* $K^3/N \to 0$.

Assumption 1 is a sufficient condition for model identification, which is proposed by Wang, Shao and Kim (2014). Assumptions 2 and 3 are required for $L^2$ approximations. Assumption 4 is a standard assumption for sieve basis, see also Newey (1997). Assumptions 5 and 6 ensure the convergence of the proposed estimator as well as the finiteness of the asymptotic variance. Assumption 7 is required for controlling the asymptotic variance, which is desirable in practice because $K$ grows very slowly with $N$ so a relatively small number of moment conditions is sufficient for the method proposed to perform well.

## References

Ai, C., Linton, O. and Zhang, Z. (2018). *Supplemental Material for "A Simple and Efficient Estimation Method for Models with Nonignorable Missing Data"*. Technical report.

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61**, 962–973.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. *Handbook of Econometrics* **6**, 5549–5632.

Chen, X., Hong, H. and Tarozzi, A. (2008). Semiparametric efficiency in GMM models with auxiliary data. *The Annals of Statistics* **36**, 808–843.

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.

Geman, S. and Hwang, C.-R. (1982). Nonparametric maximum likelihood estimation by the method of sieves. *The Annals of Statistics* **10**, 401–414.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society* **50**, 1029–1054.

Heitjan, D. F. and Basu, S. (1996). Distinguishing "missing at random" and "missing completely at random". *The American Statistician* **50**, 207–213.

Ibrahim, J. G. and Molenberghs, G. (2009). Missing data methods in longitudinal studies: a review. *Test* **18**, 1–43.

Imai, K., King, G. and Stuart, E. A. (2008). Misunderstandings among experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society* **171**, 481–502.

Kang, J. and Schafer, J. (2007). Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science* **22**, 523–539.

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.

Kosuke, I. and Marc, R. (2015). Covariate balancing propensity score. *Journal of the Royal Statistical Society* **76**, 243–263.

Little, R. J. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association* **83**, 1198–1202.

Little, R. J. and Rubin, D. B. (1989). The analysis of social science data with missing values. *Sociological Methods & Research* **18**, 292–326.

Little, R. J. and Rubin, D. B. (2014). *Statistical Analysis with Missing Data*. John Wiley & Sons.

Morikawa, K. and Kim, J. K. (2016). Semiparametric adaptive estimation with nonignorable nonresponse data. *arXiv preprint arXiv:1612.09207*.

Molenberghs, G. and Kenward, M.(2007). *Missing Data in Clinical Studies*. John Wiley & Sons.

Morikawa, K. and Kim, J. K. (2018). A note on the equivalence of two semiparametric estimation methods for nonignorable nonresponse. *Statistics and Probability Letters* **140**, 1–6.

Newey, W. K. (1997). Convergence rates and asymptotic normality for series estimators. *Journal of econometrics* **79**, 147–168.

Newey, W. K. and Smith, R. J. (2004). Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica* **72**, 219–255.

Owen, A. B. (2004). *Empirical Likelihood*. Chapman & Hall/CRC.

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 Zhang, B. (2007). Empirical-likelihood-based inference in missing response problems

and its application in observational studies. *Journal of the Royal Statistical Society. Series B. (Statistical Methodology)* **69**, 101–122.

Riddles, M. K., Kim, J. K. and Im, J. (2016). A propensity-score-adjustment method for nonignorable nonresponse. *Journal of Survey Statistics and Methodology* **4**, 215–245.

Robins, J. M. and Ritov, Y. (1997). Toward a curse of dimensionality appropriate (coda) asymptotic theory for semiparametric models. *Statistics in Medicine* **16**, 285–319.

Robins, J. M. and Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association* **90**, 122–129.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association* **90**, 106–121.

Rotnitzky, A., Lei, Q., Sued, M. and Robins, J. M. (2012). Improved double-robust estimation in missing data and causal inference models. *Biometrika* **99**, 439–456.

Rubin, D. B. (1976a). Comparing regressions when some predictor values are missing. *Technometrics* **18**, 201–205.

Rubin, D. B. (1976b). Inference and missing data. *Biometrika* **63**, 581–592.

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

Sverchkov, M. (2008). A new approach to estimation of response probabilities when missing data are not missing at random. In *Proceedings of the Survey Research Methods Section*, 867–874.

Tan, Z. (2010). Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika* **97**, 661–682.

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

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.

Wei, H., Sun, Y. and Hu, M. (2018). Model selection in spatial autoregressive models with varying coefficients. *Frontiers of Economics in China* **13**, 559–576.

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.

School of Management and Economics, Chinese University of Hong Kong, Shenzhen, China.

E-mail: chunrongai@hotmail.com

Department of Economics, University of Cambridge, Cambridge, CB2 1TN, UK.

E-mail: obl20@cam.ac.uk

Institute of Statistics and Big Data, Renmin University of China, Beijing, 100872, China

E-mail: zhengzhang@ruc.edu.cn

(Received March 2018; accepted November 2018)