Statistica Sinica **27** (2017), 89-113
doi:http://dx.doi.org/10.5705/ss.202015.0052

SEMIPARAMETRIC ESTIMATING EQUATIONS INFERENCE WITH NONIGNORABLE MISSING DATA

Puying Zhao[1], Niansheng Tang[1], Annie Qu[2] and Depeng Jiang[3]

[1] *Yunnan University,* [2] *University of IIIinois at Urbana-Champaign and* [3] *University of Manitoba*

Abstract: Handling data with the missing not at random (MNAR) mechanism is still a challenging problem in statistics. In this article, we propose a nonparametric imputation method based on the propensity score in a general class of semiparametric models for nonignorable missing data. Compared with the existing imputation methods, the proposed imputation method is more flexible as it does not require any model specification for the propensity score but rather a general parametric model involving an unknown parameter which can be estimated consistently. To obtain a consistent estimator of the parametric propensity score, two approaches are proposed. One is based on a validation sample. The other is a semi-empirical likelihood (SEL) method. By incorporating auxiliary information from some calibration conditions under the MNAR assumption, we gain significant efficiency with the SEL-based estimator. We investigate the asymptotic properties of the proposed estimators based on either known or estimated propensity scores. Our empirical studies show that that the resultant estimator is robust against the misspecified response model. Simulation studies and data analysis are provided to evaluate the finite sample performance of the proposed method.

Key words and phrases: Generalized method of moments, imputation, not missing at random, propensity score, semi-empirical likelihood, semi-parametric estimation.

1. Introduction

Missing data arises frequently in surveys, social science, and biomedical research. The commonly used methods for handling missing data include complete case analysis, which can lead to a biased estimator and information loss (Little and Rubin (2002); Kim and Shao (2013)), the imputation method (Rubin (1987); Cheng (1994); Wang and Chen (2009)), and the augmented inverse probability weighted (AIPW) method (Robins, Rotnitzky and Zhao (1994)). The implementation of the latter two methods is applicable for missingness at random (MAR), but not suitable if the missing mechanism is missing not at random (MNAR), nonignorable missingness.

Nonignorable missing models have been mainly studied using maximum likelihood, empirical likelihood (EL), and Bayesian approaches. For example, see

Troxel, Lipsitz and Brennan's (1997) weighted estimating equations for nonignorable nonresponse, Lipsitz et al.'s (1999) generalized linear models for nonignorable missing covariates, Tang, Little and Raghunathan's (2003) multivariate regression analysis with nonignorable nonresponse, and Lee and Tang's (2006) nonlinear structural equation models for nonignorable missing data. Recently, Kim and Yu (2011) proposed the exponential tilting model and developed a semiparametric estimation procedure for nonignorable missing data. Tang, Zhao and Zhu (2014) further extended the idea of Kim and Yu (2011) and Zhou, Wan and Wang's (2008) imputing estimating functions under MAR, and developed the EL inference procedure through estimating equations for nonignorable missing data. Zhao and Shao (2015) proposed a pseudo likelihood approach to generalized linear models in the presence of nonignorable missing data, and presented a two-step iteration algorithm to implement the numerical maximization of the pseudo likelihood.

The semiparametric estimating equations (SEEs) approach has been investigated for missing data in recent years (Robins and Ritov (1997); Graham (2011)). For example, Chen, Hong and Tarozzi (2008) provided semiparametric efficiency bounds under missing data, Graham (2011) employed the AIPW approach to study efficiency bounds under the semiparametric framework for ignorable missing, Chen and Van Keilegom (2013) discussed SEEs using the nonparametric imputation method for response/covariates with MAR, and Wang, Cui and Li (2013) presented an EL-based AIPW in SEEs. However, these approaches cannot be used to make statistical inference directly for nonginorable missing data due to the complexity of the nonignorable missing mechanism. To the best of our knowledge, consistent estimation under the SEE framework for nonignorable missing data has not been investigated.

In this article, we develop a general SEE approach for nonignorable missing data, and provide consistent estimators for finite-dimensional parameters in the presence of the nonparametric function with dimensions of nuisance parameters infinite. The SEEs estimator is well known for its desirable unbiasedness when the data is complete. To achieve unbiasedness of the SEEs estimation when observations are missing not at random, we propose a propensity-score-based and kernel-assisted SEE imputation scheme for parameter estimations. The construction of the proposed imputation approach is motivated by the exponential tilting model-based imputation procedure proposed by Kim and Yu (2011). To make the propensity-score-based imputation applicable, it is important to estimate the propensity score consistently. Once a consistent propensity score estimator is obtained, we can formulate a basis for inferences using the imputed SEEs via the generalized method of moments (GMM) (Hansen (1982)).

Specifically, a general parametric model is proposed for the response probability. To estimate the parametric propensity scores consistently, we first propose

an estimation procedure based on validation samples, and apply the theory of GMM to improve the efficiency. However, as the method of independent survey using validation samples is often not realistic since budget or technical limitations often restrict researchers to design studies that collect follow-up samples to gain more information for estimating the response probability model. In addition, it is important to utilize the calibration conditions presented in missing data problems. To this end, we employ a semi-empirical likelihood (SEL) procedure (Qin, Leung and Shao (2002)) to estimate parametric propensity scores because of its properties that its implementation needs only complete observations, and that one can incorporate some auxiliary information from the calibration conditions under nonignorable missing mechanism to gain a more efficient estimator for the response probability.

Compared with the exponential tilting model-based imputation of Kim and Yu (2011), our proposed parametric-propensity score-based imputation method has the advantages of being more flexible so that one can develop an appropriate approach for the estimation of parametric propensity score to improve the efficiency of the resultant estimator. In addition, the use of parametrically estimated propensity scores can alleviate the dimensionality issue to a certain extent. Another advantage of our approach is that consistency and asymptotic normality of the estimators are established under fairly mild conditions. In particular, we do not require the criterion function to satisfy standard smoothness conditions.

The parameter identifiability issue (Robins and Ritov (1997)) is crucial and challenging in nonignorable missing data analysis. Many authors have studied this issue, see Tang, Little and Raghunathan (2003), Wang, Shao and Kim (2014), and Zhao and Shao (2015). Their methods can be applied to our model. Hence, throughout this article, we assume that the models considered are identifiable.

The rest of this article is organized as follows. We introduce the propensity-score-based nonparametric imputation in Section 2. We construct a class of GMM estimators of parameters defined via SEEs with nonignorable missing data, and we show the consistency and asymptotic normality of the proposed estimators in Section 3. A bootstrap procedure for approximating asymptotic variance of the proposed estimators and a simple dimensionality reduction technique in relation to the proposed kernel procedure are proposed in Section 4. An example illustrating the proposed method and some extensions are discussed in Section 5. Simulation studies conducted to investigate the finite sample performance of the proposed estimators are presented in Section 6. Data from a Workplace Safety & Insurance Board research study in Canada is used to illustrate the proposed method in Section 7. Some concluding remarks are given in Section 8. Technical conditions and proofs of the theorems are reported in the Supplementary Material.

2. Propensity Score-based Nonparametric Imputation

Let $Z_i = (X_i^\top, Y_i^\top)^\top$, $i = 1, \ldots, n$, be a set of $(s+d)$-dimensional independent and identically distributed (i.i.d.) random variables with the cumulative distribution function $F(z)$. Let Θ be a finite-dimensional parameter set (a compact subset of $\mathcal{R}^p$) and $\mathcal{H}$ be an infinite dimensional parameter set. The function in $\mathcal{H}$ is allowed to depend on θ. Suppose that $\psi(Y, X, \theta, h)$ is a vector of q estimating equations, known up to the finite dimensional parameter $\theta \in \Theta$ and the infinite dimensional nuisance function $h \in \mathcal{H}$. The only prior restriction on $F(z)$ is that $E\{\psi(Y, X, \theta_0, h_0)\} = 0$ for some $\theta_0 \in \Theta \subset \mathcal{R}^p$ and $h_0 \in \mathcal{H}$. Here, θ_0 and h_0 are the true value of θ and the true function of h, respectively. That $q > p$ implies that $\psi(Y, X, \theta, h)$ is an over-identified system. Similar to Chen, Linton and Van Keilegom (2003), it is assumed that the function h_0 depends on θ and the data X and/or Y. For simplicity, we write $(\theta, h) =: (\theta, h_\theta)$, $(\theta, h_0) =: (\theta, h_{0\theta})$, and $(\theta_0, h_0) =: (\theta_0, h_{0\theta_0})$.

We assume that Y_i is subject to missingness, whereas X_i is always available. Generally, the missing components may vary across different individuals. For simplicity, we suppose that the missing components have the same dimensions for $Z_1, \ldots, Z_n$. Further, a missing variable Y_i may represent a response or covariate. Let $\delta_i = 1$ if Y_i is observed and $\delta_i = 0$ if Y_i is missing. We assume that δ_i is independent of δ_j for any $i \neq j$ and that $\Pr(\delta_i = 1|X_i, Y_i) =: \pi(X_i, Y_i)$, which allows that the missingness mechanism is MNAR. Let $\mathcal{G}(\theta, h) = E\{\psi(Y, X, \theta, h)\}$, a non-random vector-valued function $\mathcal{G}$: $\Theta \times \mathcal{H} \to \mathcal{R}^q$ such that $\mathcal{G}(\theta_0, h_0) = 0$. The issue is to estimate θ in the presence of nonignorable missing data.

Let $\{(X_i, Y_i, \delta_i), i = 1, \cdots, n\}$ be i.i.d. random vectors having the same distribution as (X, Y, δ). To incorporate the nonignorable missing data set, we consider a set of semi-parametric estimating functions given by

$$\check{\psi}(Y_i, X_i, \theta, h) = \delta_i \psi(Y_i, X_i, \theta, h) + (1 - \delta_i) m_\psi^0(X_i, \theta, h),$$

where $m_\psi^0(X_i, \theta, h) = E\{\psi(Y_i, X_i, \theta, h)|X_i, \delta_i = 0\}$. Let $f_1(Y_i|X_i)$ be the conditional probability density of Y_i given X_i and $\delta_i = 1$, and $f_0(Y_i|X_i)$ the conditional probability density of Y_i given X_i and $\delta_i = 0$. We further assume that the response probability model has the parametric form

$$\pi(X_i, Y_i) = \Pr(\delta_i = 1|X_i, Y_i) =: \pi(X_i, Y_i, \alpha_0), \tag{2.1}$$

where $\pi(\cdot)$ is a known smooth function in the finite-dimensional response model parameter α_0.

Following the reasoning of Kim and Yu (2011), we can obtain

$$f_0(Y_i|X_i) = f_1(Y_i|X_i) \times \frac{O(X_i, Y_i, \alpha_0)}{E\{O(X_i, Y_i, \alpha_0)|X_i, \delta_i = 1\}}, \tag{2.2}$$

where $O(X_i, Y_i, \alpha_0) = \Pr(\delta_i = 0|X_i, Y_i)/\Pr(\delta_i = 1|X_i, Y_i) = \pi^{-1}(X_i, Y_i, \alpha_0) - 1$ is the conditional odds of nonresponse. Using (2.2), we obtain

$$m^0_\psi(X_i, \theta, h) =: m^0_\psi(X_i, \theta, h, \alpha_0) = \frac{E\{\delta_i \psi(Y_i, X_i, \theta, h) O(X_i, Y_i, \alpha_0)|X_i\}}{E\{\delta_i O(X_i, Y_i, \alpha_0)|X_i\}}.$$

Then, SEEs $\check{\psi}(Y_i, X_i, \theta, h)$ can be rewritten as

$$\tilde{\psi}(Y_i, X_i, \theta, h, \alpha_0) = \delta_i \psi(Y_i, X_i, \theta, h) + (1 - \delta_i) m^0_\psi(X_i, \theta, h, \alpha_0). \tag{2.3}$$

If the true response probability follows the parametric model given in (2.1), we can show that $E\{\tilde{\psi}(Y_i, X_i, \theta_0, h_0, \alpha_0)\} = 0$. Thus (2.3) is unbiased, the key idea of our approach.

Let $K(\cdot)$ be an s-dimensional kernel function of the mth order satisfying $\int K(u_1, \ldots, u_s) du_1 \ldots du_s = 1$, $\int u^l_k K(u_1, \ldots, u_s) du_1 \ldots du_s = 0$ for any $k = 1, \ldots, s$ and $1 \le l < m$, and $\int u^m_k K(u_1, \ldots, u_s) du_1 \ldots du_s \neq 0$. Then, a nonparametric regression estimator of $m^0_\psi(X, \theta, h)$ can be expressed as

$$\hat{m}^0_\psi(X, \theta, h, \alpha_0) = \frac{\sum_{i=1}^n \delta_i O(X_i, Y_i, \alpha_0) K_a(X - X_i) \psi(Y_i, X_i, \theta, h)}{\sum_{i=1}^n \delta_i O(X_i, Y_i, \alpha_0) K_a(X - X_i)}, \tag{2.4}$$

where the weight $\delta_i O(X_i, Y_i, \alpha_0) K_a(X - X_i)/\sum_{j=1}^n \delta_j O(X_j, Y_j, \alpha_0) K_a(X - X_j)$ represents the point mass assigned to $\psi(Y_i, X_i, \theta, h)$ when $m^0_\psi(X, \theta, h)$ is approximated by $\hat{m}^0_\psi(X, \theta, h, \alpha_0)$, $K_a(u) = a^{-s} K(u/a)$ and a is a bandwidth sequence. Using the arguments of Devroye and Wagner (1980), we can show that, under the true response model (2.1) and some regularity conditions, $\lim_{n\to\infty} \hat{m}^0_\psi(X, \theta, h, \alpha_0) = m^0_\psi(X, \theta, h)$. Therefore, a set of the modified SEEs for the ith observation is given by $\hat{\psi}(Y_i, X_i, \theta_0, h_0, \alpha_0) = \delta_i \psi(Y_i, X_i, \theta_0, h_0) + (1 - \delta_i) \hat{m}^0_\psi(X_i, \theta_0, h_0, \alpha_0)$.

3. Generalized Method of Moments Estimation

3.1. Nonparametric estimation

Let $\mathcal{G}_n(\theta, h, \alpha) = n^{-1} \sum_{i=1}^n \hat{\psi}(Y_i, X_i, \theta, h, \alpha)$. Given an estimator $\hat{h}$ of h and a known propensity score, we define the nonparametric estimator of θ by

$$\hat{\theta}_{NP} = \arg\min_{\theta \in \Theta} \|\mathcal{G}_n(\theta, \hat{h}, \alpha_0)\|_W,$$

where $\|A\|_W = \{\mathrm{tr}(A^\top W A)\}^{1/2}$ for any q-dimensional vector A and some fixed symmetric $q \times q$ positive definite matrix W; here $\mathrm{tr}(\cdot)$ stands for the trace of a matrix.

Theorem 1. *If the conditions* (A1)−(A3) *and* (C1)−(C4) *given in the Supplementary Material hold and the response probability* $\pi(X,Y)$ *is known, then* $\hat{\theta}_{NP}-\theta_0=o_p(1)$. *If the conditions* (B1)−(B5) *given in the Supplementary Material also hold, we have*

$$n^{1/2}(\hat{\theta}_{NP}-\theta_0)\xrightarrow{\mathcal{L}}\mathcal{N}(0,\Sigma_1),$$

where $\Sigma_1=(\Lambda^\top W\Lambda)^{-1}\Lambda^\top W\Gamma_1 W\Lambda(\Lambda^\top W\Lambda)^{-1}$, $\Gamma_1=\mathrm{Var}\{\mathcal{S}(X,Y,\theta_0,h_0)\}$, $\mathcal{S}(X,Y,\theta,h)=\delta\{\pi(X,Y)\}^{-1}\{\psi(Y,X,\theta,h)-m_\psi^0(X,\theta,h)\}+m_\psi^0(X,\theta,h)+\nabla(X,Y,\delta)$, *the function* $\nabla(\cdot)$ *is defined in the condition* (B4), *and* $\Lambda=\Lambda(\theta_0,h_0)$ *with*

$$\Lambda(\theta,h_0)=\frac{\partial}{\partial\theta}\mathcal{G}(\theta,h_0)=\lim_{\kappa\to0}\frac{1}{\kappa}\{\mathcal{G}(\theta+\kappa,h_{0,\theta+\kappa})-\mathcal{G}(\theta,h_{0\theta})\}.$$

Theorem 1 shows that using the nonparametric regression estimator of $m_\psi^0(X,\theta,h)$ with known response model can lead to an efficient influence function of estimator $\hat{\theta}_{NP}$, which has the AIPW form (Robins, Rotnitzky and Zhao (1994); Graham (2011)). However, the nonparametric regression methods are impeded by the curse of dimensionality.

3.2. Semiparametric estimation

Although $\hat{\theta}_{NP}$ is theoretically attractive, it is practically useless because the parameter vector α_0 in (2.1) is unknown in many applications. While we should estimate α_0 consistently before making inference on θ, it is difficult to obtain a suitable estimator of α_0 under the MNAR assumption because Y_i is unobserved in the set of nonrespondents. We use $\hat{\alpha}$ to denote a suitable estimator of α_0. Then, given estimators $\hat{\alpha}$ and $\hat{h}$, a semiparametric estimator of θ can be obtained as

$$\hat{\theta}_{SP}=\arg\min_{\theta\in\Theta}\|\mathcal{G}_n(\theta,\hat{h},\hat{\alpha})\|_W. \tag{3.1}$$

In what follows, we try to find some reasonable estimators $\hat{\alpha}$ to make the estimating equations $\hat{\psi}(Y_i,X_i,\theta_0,\hat{h},\hat{\alpha})$ applicable.

(1) Independent Survey and Validation Sample

Motivated by Kim and Yu (2011), we consider two approaches to compute $\hat{\alpha}$: independent survey and a validation sample.

Theorem 2. *If the conditions* (A1)−(A3) *and* (C1)−(C4) *of the Supplementary Material hold, the response probability model* $\pi(X,Y,\alpha_0)$ *is correctly specified and* $\hat{\alpha}$ *is consistent, then* $\hat{\theta}_{SP}-\theta_0=o_p(1)$. *If the conditions* (B1)−(B5) *given in the Supplementary Material also hold,* $n^{1/2}(\hat{\alpha}-\alpha_0)\xrightarrow{\mathcal{L}}\mathcal{N}(0,V_\alpha)$, *and* $\hat{\alpha}$ *is independent of* $\hat{\psi}(Y_i,X_i,\theta_0,h_0,\alpha_0)$, *we have*

$$n^{1/2}(\hat{\theta}_{SP}-\theta_0)\xrightarrow{\mathcal{L}}\mathcal{N}(0,\Sigma_2),$$

where $\Sigma_2 = (\Lambda^\top W\Lambda)^{-1}\Lambda^\top W\Gamma_2 W\Lambda(\Lambda^\top W\Lambda)^{-1}$, $\Gamma_2 = \mathrm{Var}\{\mathcal{S}(X,Y,\theta_0,h_0)\}+HV_\alpha H^\top$, $H = E[(1-\delta)\{\psi(Y,X,\theta_0,h_0)-m^0_\psi(X,\theta_0,h_0,\alpha_0)\}\{z(X,Y,\alpha_0)-m^0_z(X,\alpha_0)\}^\top]$ *with* $z(X,Y,\alpha) = \partial\mathrm{logit}\{\pi(X,Y,\alpha)\}/\partial\alpha$ *and* $m^0_z(X,\alpha)=E\{z(X,Y,\alpha)|X,\delta=0\}$, *and* $\mathrm{logit}(p) = \log\{p/(1-p)\}$.

It follows from Theorems 1 and 2 that $\hat\theta_{SP}$ has larger asymptotic variance than $\hat\theta_{NP}$ due to estimating α_0.

We consider then that $\hat\alpha$ is obtained from a validation sample, randomly selected from the set of the nonrespondents. For clarity, we take $\mathcal{Q}(\alpha,\theta,h) = E[(1-\delta_i)\{\psi(Y_i,X_i,\theta,h)-m^0_\psi(X_i,\theta,h,\alpha)\}]$. Under the MNAR assumption,

$$\begin{aligned}E\{(1-\delta_i)\psi(Y_i,X_i,\theta,h)\} &= E[E\{(1-\delta_i)\psi(Y_i,X_i,\theta,h)|X_i,\delta_i=0\}]\\ &= E\{(1-\delta_i)m^0_\psi(X_i,\theta,h,\alpha_0)\},\end{aligned}$$

which leads to $\mathcal{Q}(\alpha_0,\theta,h) = 0$ for any $\theta\in\Theta\subset\mathcal{R}^p$ and $h\in\mathcal{H}$. Therefore, a consistent estimator $\hat\alpha$ of α_0 can be obtained by solving

$$\mathcal{Q}_n(\alpha_0,\theta,h) = \frac{1}{n}\sum_{i=1}^n(1-\delta_i)\frac{r_i}{\nu}\{\psi(Y_i,X_i,\theta,h)-\hat m^0_\psi(X_i,\theta,h,\alpha_0)\} = 0$$

for α_0, where $\nu = E(r_i|\delta_i=0)$ and r_i is 1 if individual i belongs to the follow-up sample, and 0 otherwise.

It is noteworthy that a set of SEEs $\mathcal{Q}_n(\alpha_0,\theta,h)$ may also be an over-identified system with respect to α_0. Hence, the GMM approach can again be used to compute α_0 to improve efficiency. Given an appropriate and symmetric $q\times q$ positive definite matrix $\widetilde W$, an estimator of α_0 can be obtained as

$$\hat\alpha_{\mathrm{v}} = \underset{\theta\in\Theta,\alpha\in\mathcal{B}}{\arg\min}\,\|\mathcal{Q}_n(\alpha,\theta,\hat h)\|_{\widetilde W}. \tag{3.2}$$

Proposition 1. *Suppose the conditions of the Supplementary Material hold, and the response probability model* $\pi(X,Y,\alpha_0)$ *is correctly specified. Then,* $\hat\alpha_{\mathrm{v}}-\alpha_0 = o_p(1)$, *and* $\hat\alpha_{\mathrm{v}}-\alpha_0 = -(H^\top\widetilde WH)^{-1}H^\top\widetilde Wn^{-1}\sum_{i=1}^n\mathcal{D}_i(\theta_0,h_0,\alpha_0)+o_p(n^{-1/2})$, *equivalently* $n^{1/2}(\hat\alpha_{\mathrm{v}}-\alpha_0)\xrightarrow{\mathcal{L}}\mathcal{N}(0,\Sigma_{\mathrm{v}})$, *where*

$$\Sigma_{\mathrm{v}} = (H^\top\widetilde WH)^{-1}H^\top\widetilde WV_{\mathrm{v}}\widetilde WH(H^\top\widetilde WH)^{-1},\quad V_{\mathrm{v}} = \mathrm{Var}\{\mathcal{D}(\theta_0,h_0,\alpha_0)\},\quad\text{and}$$

$$\mathcal{D}(\theta_0,h_0,\alpha_0) = [(1-\delta)r/\nu-\delta\{\pi^{-1}(X,Y)-1\}]\{\psi(Y,X,\theta_0,h_0)-m^0_\psi(X,\theta_0,h_0,\alpha_0)\}.$$

Corollary 1. *If* $q=\dim(\alpha_0)$ *or* $\widetilde W = V_{\mathrm{v}}^{-1}$, *then* $\Sigma_{\mathrm{v}} = (H^\top V_{\mathrm{v}}^{-1}H)^{-1}$ *and* $\hat\alpha_{\mathrm{v}}$ *is efficient among the class of GMM validation sample estimators.*

Using the estimated parameter $\hat\alpha_{\mathrm{v}}$, we can construct $\mathcal{G}_n(\theta,\hat h_\theta,\hat\alpha_{\mathrm{v}})$. Then, we can obtain a semiparametric estimator $\hat\theta_{SP}$ of θ from (3.1).

Theorem 3. *If the conditions* (A1)−(A3) *and* (C1)−(C4) *of the Supplementary Material hold, the response probability model* $\pi(X, Y, \alpha_0)$ *is correctly specified, and* $\hat{\theta}_{SP}$ *is the estimator of* θ *obtained by solving* (3.1) *with* $\hat{\alpha} = \hat{\alpha}_{\rm v}$*, then* $\hat{\theta}_{SP} - \theta_0 = o_p(1)$*. If the conditions* (B1)−(B5) *of the Supplementary Material also hold, we have*

$$n^{1/2}(\hat{\theta}_{SP} - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma_3),$$

where $\Sigma_3 = (\Lambda^\top W \Lambda)^{-1}\Lambda^\top W \Gamma_3 W \Lambda (\Lambda^\top W \Lambda)^{-1}$, $\Gamma_3 = \mathrm{Var}\{\mathcal{O}(X, Y, \delta)\}$, $\mathcal{O}(X, Y, \delta) = \mathcal{S}(X, Y, \theta_0, h_0) + H(H^\top \widetilde{W} H)^{-1} H^\top \widetilde{W} \mathcal{D}(\theta_0, h_0, \alpha_0)$.

Corollary 2. *If* $q = p$ *or* $W = \Gamma_3^{-1}$*, we have* $\Sigma_3 = (\Lambda^\top \Gamma_3^{-1} \Lambda)^{-1}$*, then* $\hat{\theta}_{SP}$ *is efficient among the class of semiparametric estimators for* θ *based on the method of validation sample.*

Remark 1. If $q = \dim(\alpha_0)$, we have $\mathcal{O}(X, Y, \delta) = \{(r/\nu)(1-\delta) + \delta\}\{\psi(Y, X, \theta_0, h_0) - m_\psi^0(X, \theta_0, h_0, \alpha_0)\} + m_\psi^0(X, \theta_0, h_0, \alpha_0) + \nabla(X, Y, \delta)$, and $\mathrm{Var}\{\mathcal{O}(X, Y, \delta)\} = \mathrm{Var}\{\psi(Y, X, \theta_0, h_0)\} + \mathrm{Var}\{\nabla(X, Y, \delta)\} + 2\mathrm{Cov}(\psi(Y, X, \theta_0, h_0), \nabla(X, Y, \delta)) + (\nu^{-1} - 1)E[(1-\delta)\{\psi(Y, X, \theta_0, h_0) - m_\psi^0(X, \theta_0, h_0, \alpha_0)\}^{\otimes 2}]$, where $a^{\otimes 2} = aa^\top$ for any vector a. Using standard kernel regression theory, for any $\alpha^* \in \mathcal{B}$ (α^* may be the probability limit of $\hat{\alpha}_{\rm v}$ when (2.1) is misspecified), $m_\psi^0(X, \theta_0, h_0, \alpha^*) = \lim_{n\to\infty} \hat{m}_\psi^0(X, \theta_0, h_0, \alpha^*)$. If $\alpha^* = \alpha_0$, (2.1) is correctly specified, then $m_\psi^0(X, \theta_0, h_0, \alpha^*) = E\{\psi(Y, X, \theta_0, h_0) | X, \delta = 0\} = m_\psi^0(X, \theta_0, h_0)$. Here $E[(1-\delta)\{\psi(Y, X, \theta_0, h_0) - m_\psi^0(X, \theta_0, h_0, \alpha^*)\}^{\otimes 2}] \geq E[(1-\delta)\{\psi(Y, X, \theta_0, h_0) - m_\psi^0(X, \theta_0, h_0)\}^{\otimes 2}]$ indicates that Σ_3 attains its minimum for this scenario. Thus, (2.1) can be used to improve estimation efficiency (Kim and Yu (2011)).

Remark 2. Our asymptotic results are obtained under the correctly specified response model. It is challenging to establish asymptotic properties of the proposed semiparametric estimators when the response model is misspecified because of the non-smoothness of the underlying SEEs and the infinite-dimensional nuisance function involved. To the best of our knowledge, this issue has not been addressed for the estimation of over-identified moment conditions even under the MAR assumption (e.g., Chen, Hong and Tarozzi (2008)). For two special cases, including the population mean (Kim and Yu (2011)) and the distribution function of the response variable, we investigate the robustness of the proposed imputation approach to the selection of response model in the Supplementary Material.

(2) Semi-empirical Likelihood Estimation

The advantages of using a validation sample to estimate α_0 include robustness properties and the parametric rate of convergence for the resulting $\hat{\alpha}_{\rm v}$, but

budget or technical limitations may restrict researchers to design studies that collect follow-up samples to evaluate $\hat{\alpha}_{\rm v}$. To overcome such difficulties, we employ an approach of semiparametric likelihood (Qin, Leung and Shao (2002)) based on complete observations to obtain an efficient estimator for the response probability function. To this end, let $F(X,Y)$ be the unconditional joint distribution of (X,Y) and $\mathcal{A}=\{j:\delta_j=1\}$ be the set of respondents in the sample $\{(X_j,Y_j):j=1,\ldots,n\}$; $n_1=|\mathcal{A}|$ denotes the size of the set $\mathcal{A}$. The likelihood of (α_0,F) based on complete observations $\{(X_j,Y_j):j\in\mathcal{A}\}$ is given by

$$\prod_{j\in\mathcal{A}}\pi(X_j,Y_j,\alpha_0)dF(X_j,Y_j)\prod_{j\notin\mathcal{A}}\iint\{1-\pi(X_j,Y_j,\alpha_0)\}dF(X_j,Y_j),$$

which can be rewritten as

$$\left\{\prod_{j\in\mathcal{A}}\frac{\pi(X_j,Y_j,\alpha_0)dF(X_j,Y_j)}{\omega}\right\}\omega^{n_1}(1-\omega)^{n-n_1}, \tag{3.3}$$

where $\omega=\Pr(\delta=1)=\iint\pi(X,Y,\alpha_0)dF(X,Y)$ is the unconditional response rate. The first term in (3.3) is the likelihood conditional on $\delta=1$, and the term $\omega^{n_1}(1-\omega)^{n-n_1}$ is the binomial likelihood of δ.

Some auxiliary information on X of the form $E\{g(X)\}=0$ is often available, where $g(X)=(g_1(X),\ldots,g_l(X))^\top$ is a known $l\geq 1$ vector (or scalar) function. Based on the auxiliary information from X, and without assuming any specific form for $F(X,Y)$, we can maximize the semiparametric likelihood (3.3) subject to the contraints

$$p_j\geq 0,\quad \sum_{j\in\mathcal{A}}p_j=1,\quad \sum_{j\in\mathcal{A}}p_j\{\pi(X_j,Y_j,\alpha_0)-\omega\}=0,\quad \sum_{j\in\mathcal{A}}p_jg(X_j)=0,$$

where p_j is the jump of F at $\{(X_j,Y_j):j\in\mathcal{A}\}$. By introducing Langrange multipliers λ_1 and λ_2, the log-likelihood with respect to α_0 and ω is

$$\begin{aligned}l(\alpha_0,\omega,\lambda_1,\lambda_2)=&\sum_{j\in\mathcal{A}}\log\pi(X_j,Y_j,\alpha_0)+(n-n_1)\log(1-\omega)\\&-\sum_{j\in\mathcal{A}}\log\{1+\lambda_1g(X_j)+\lambda_2(\pi(X_j,Y_j,\alpha_0)-\omega)\}.\end{aligned}\tag{3.4}$$

The solution of the constrained maximum likelihood can be obtained by maximizing the log-likelihood $l(\alpha_0,\omega,\lambda_1,\lambda_2)$. Denote the solution by $(\hat{\alpha}_{\rm s},\hat{\omega},\hat{\lambda}_1,\hat{\lambda}_2)$, and take $\zeta=\lambda_1(1-\omega)$, $\eta_0=(\alpha_0,\omega_0,0)^\top$, $\hat{\zeta}=\hat{\lambda}_1(1-\hat{\omega})$, and $\hat{\eta}=(\hat{\alpha}_{\rm s},\hat{\omega},\hat{\zeta})^\top$.

Computing the semiparametric likelihood estimator $\hat{\eta}$ is computationally challenging because too many constraints are involved. To address this, we adopt the algorithm of Qin, Leung and Shao (2002). *Step 1.* Given (α,ω), compute $(\lambda_1(\alpha,\omega),\lambda_2(\alpha,\omega))=\arg\min_{\lambda_1,\lambda_2}l(\alpha,\omega,\lambda_1,\lambda_2)$; *Step 2.* Compute $(\hat{\alpha}_s,\hat{\omega})=\arg\max_{\alpha,\omega}l(\alpha,\omega,\lambda_1(\alpha,\omega),\lambda_2(\alpha,\omega))$.

Remark 3. One can also obtain a consistent estimator of α_0 by solving

$$\sum_{i=1}^{n} g(X_i, Y_i, \alpha_0) = \sum_{i=1}^{n} \left\{ \frac{\delta_i}{\pi(X_i, Y_i, \alpha_0)} - 1 \right\} \tau(X_i) = 0, \tag{3.5}$$

where $\tau(\cdot)$ is a user-specified vector function with the same dimension as α_0 or with the dimension being greater than that of α_0. Condition (3.5) is often called the calibration condition, and has been widely used in survey sampling as well as in nonignorable missing problem (e.g., Chang and Kott (2008); Wang, Shao and Kim (2014); Kim and Shao (2013); Riddles, Kim and Im (2015)). Based on (3.5), an alternative semiparametric empirical likelihood function can be constructed by using $g(X_j, Y_j, \alpha_0)$ to replace $g(X_j)$ at (3.4).

Proposition 2. *If the conditions in the Supplementary Material hold and the matrix U defined in the Appendix is nonsingular, then $\hat{\eta} \xrightarrow{p} \eta_0$ and $n^{1/2}(\hat{\eta} - \eta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, U^{-1}V(U^{-1})^\top)$, where V is defined in the Supplementary Material.*

Remark 4. Compared with the GMM-based validation sample method, the SEL method is more efficient because it easily incorporates the auxiliary information and is easy to implement as it uses only the complete observations. The auxiliary information $g(X)$ should be carefully selected such that the EL procedure works because U, defined in Proposition 2, will fail to be invertible if the dimension of the parameter α_0 is too high whilst the dimension of $g(X)$ is too low. This issue can be addressed by utilizing a nonresponse instrumental variable that does not relate to the response mechanism but can be used to identify the parameters in the nonignorable response mechanism. More details on nonresponse instrumental variables can be found in Wang, Shao and Kim (2014) and Zhao and Shao (2015).

From Proposition 2, we can obtain an asymptotic linear expansion for $\hat{\alpha}_{\mathrm{s}}$: $n^{1/2}(\hat{\alpha}_{\mathrm{s}} - \alpha_0) = n^{-1/2} \sum_{i=1}^{n} \Psi_i(\alpha_0) + o_p(1)$, where $\Psi_i(\alpha_0) := \Psi(X_i, Y_i, \alpha_0)$ is an influence function that is defined in the Supplementary Material.

Theorem 4. *If the conditions* (A1)−(A3) *and* (C1)−(C4) *of the Supplementary Material hold, the response probability model $\pi(X, Y, \alpha_0)$ is correctly specified, the solution $\hat{\alpha}_s$ for maximizing* (3.4) *exists almost everywhere, and $\hat{\theta}_{SP}$ is the estimator of θ obtained by solving equation* (3.1) *with $\hat{\alpha} = \hat{\alpha}_{\mathrm{s}}$, then $\hat{\theta}_{SP} - \theta_0 = o_p(1)$. If the conditions* (B1)−(B5) *of the Supplementary Material also hold,*

$$n^{1/2}(\hat{\theta}_{SP} - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma_4),$$

where $\Sigma_4 = (\Lambda^\top W \Lambda)^{-1} \Lambda^\top W \Gamma_4 W \Lambda (\Lambda^\top W \Lambda)^{-1}$, $\Gamma_4 = \mathrm{Var}\{\mathcal{S}(X, Y, \theta_0, h_0) - H\Psi(\alpha_0)\}$ *and H is given in Theorem 2.*

Corollary 3. *If $q = p$ or $W = \Gamma_4^{-1}$, we have $\Sigma_4 = (\Lambda^\top \Gamma_4^{-1} \Lambda)^{-1}$, and the estimator $\hat{\theta}_{SP}$ is efficient among the class of semiparametric estimator of θ using the approach of semiparametric likelihood.*

Remark 5. It is easy to show that $\Gamma_4 = \mathrm{Var}\{\mathcal{S}(X, Y, \theta_0, h_0)\} + H\mathrm{Var}\{\Psi(\alpha_0)\}H^\top - 2H\mathrm{Cov}\{\mathcal{S}(X, Y, \theta_0, h_0), \Psi(\alpha_0)\}$, so the efficiency of the proposed semiparametric estimator in Theorem 4 depends on the correlation between the score function $\mathcal{S}(\cdot)$ and the influence function $\Psi(\cdot)$. Particularly, if $H\mathrm{Var}\{\Psi(\alpha_0)\}H^\top < 2H\mathrm{Cov}\{\mathcal{S}(X, Y, \theta_0, h_0), \Psi(\alpha_0)\}$, the semiparametric estimator achieves efficiency gain over the nonparametric estimator.

4. Asymptotic Variance Estimation and Dimension Reduction

In Theorems 1-4, the asymptotic covariance matrices of the proposed nonparametric/semiparametric estimators have complicated forms, so it is difficult to directly estimate them. We adopt a bootstrap procedure to approximate their asymptotic variances.

1. Let $\mathcal{X}_n^* = \{(X_i^*, Y_i^*, \delta_i^*) : i = 1, \ldots, n\}$ be a bootstrap sample drawn from $\{(X_j, Y_j, \delta_j) : j = 1, \ldots, n\}$. Based on the bootstrap sample $\mathcal{X}_n^*$, compute the bootstrap estimators $\hat{h}_\theta^*$ and $\hat{\alpha}^*$ via the proposed approaches.
2. Let $\hat{\psi}(Y_i^*, X_i^*, \theta, \hat{h}_\theta^*, \hat{\alpha}^*)$ be the bootstrap version of $\hat{\psi}(Y_i, X_i, \theta, \hat{h}_\theta, \hat{\alpha})$. Define the recentered SEEs

$$\hat{\psi}^c(Y_i^*, X_i^*, \theta, \hat{h}_\theta^*, \hat{\alpha}^*) = \hat{\psi}(Y_i^*, X_i^*, \theta, \hat{h}_\theta^*, \hat{\alpha}^*) - \hat{\psi}(Y_i, X_i, \hat{\theta}_{SP}, \hat{h}_\theta, \hat{\alpha})$$

   and $\mathcal{G}_n^*(\theta, \hat{h}_\theta^*, \hat{\alpha}^*) = n^{-1} \sum_{i=1}^n \hat{\psi}^c(Y_i^*, Z_i^*, \theta, \hat{h}_\theta^*, \hat{\alpha}^*)$. Obtain the bootstrap $\hat{\theta}^* = \arg\min_{\theta \in \Theta} ||\mathcal{G}_n^*(\theta, \hat{h}_\theta^*, \hat{\alpha}^*)||_W$.
3. Repeat the two steps B times to get $\hat{\theta}^{*1}, \ldots, \hat{\theta}^{*B}$. Take $\widehat{\mathrm{var}(\hat{\theta}_{SP})} = B^{-1} \sum_{j=1}^B (\hat{\theta}^{*j} - \bar{\theta}^*)(\hat{\theta}^{*j} - \bar{\theta}^*)^\top$ with $\bar{\theta}^* = B^{-1} \sum_{j=1}^B \hat{\theta}^{*j}$, and the $100(1-\alpha)\%$ confidence interval for θ to be $(\hat{\theta}^*_{([B\alpha/2])}, \hat{\theta}^*_{([B(1-\alpha/2)])})$, where $\hat{\theta}^*_{(1)}, \ldots, \hat{\theta}^*_{(B)}$ denote the ordered values of $\hat{\theta}^{*1}, \ldots, \hat{\theta}^{*B}$ and $[d]$ represents the integer part of d.

When the dimension of variate X is high, it is difficult to get an accurate estimator of $m_\psi^0(X_i, \theta, h, \alpha_0)$ by a kernel-smoothing procedure. Here we propose a dimension reduction technique such that our method is still effective for high-dimensional data.

Let S be a continuous function from $\mathcal{R}^s$ to $\mathcal{R}$, such that $E\{\psi(Y_i, X_i, \theta, h)|S_i, \delta_i = 0\} = E\{\psi(Y_i, X_i, \theta, h)|X_i, \delta_i = 0\}$ with $S_i = S(X_i)$. Then $E\{\delta_i \psi(Y_i, X_i, \theta, h) + (1 - \delta_i) m_\psi^0(S_i, \theta, h, \alpha)\} = 0$, where $m_\psi^0(S_i, \theta, h, \alpha) = E\{\delta_i \psi(Y_i, X_i, \theta, h) O(X_i, Y_i, \alpha)|S_i\} / E\{\delta_i O(X_i, Y_i, \alpha)|S_i\}$. Consequently, the kernel-assisted SEEs can be constructed as $\hat{\psi}_R(Y_i, X_i, \theta, h, \alpha) = \delta_i \psi(Y_i, X_i, \theta, h) + (1 - \delta_i) \hat{m}_\psi^0(S_i, \theta, h, \alpha)$, where

$\hat{m}^0_\psi(S_i, \theta, h, \alpha)$ is structurally identical to $\hat{m}^0_\psi(X_i, \theta, h, \alpha)$ at (2.4) except that X is replaced by S. Given $\hat{\alpha}$, one can obtain a semiparametric dimension reduction GMM estimator $\hat{\theta}_R = \arg\min_{\theta\in\Theta} \|n^{-1}\sum_{i=1}^n \hat{\psi}_R(Y_i, X_i, \theta, \hat{h}_\theta, \hat{\alpha})\|_W$.

In many applications, we assume that the working index $S = S(X, \gamma)$ involves an unknown parameter vector γ. Given an estimator $\hat{\gamma}$ of γ, a set of semiparametric dimension reduction kernel-assisted SEEs can be constructed as $\hat{\psi}_S(Y_i, X_i, \theta, h, \alpha) = \delta_i\psi(Y_i, X_i, \theta, h) + (1-\delta_i)\hat{m}^0_\psi(\hat{S}_i, \theta, h, \alpha)$ with $\hat{S}_i = S(X, \hat{\gamma})$. Using the arguments of Hu, Follmann and Qin (2010), we can show that the resultant GMM estimator based on $\hat{\psi}_S$ is asymptotically equivalent to $\hat{\theta}_R$ when $\hat{\gamma} - \gamma = O_p(n^{-1/2})$.

5. An Example and Some Extensions

In the Supplemental Material, we consider an example of a partial linear regression model to illustrate the proposed strategy for dealing with non-ignorable missing values. We discuss some extensions. Without considering the infinite-dimensional nuisance function $h \in \mathcal{H}$, the prior restriction on $F(z)$ reduces to $E\{\psi(Y, X, \theta_0)\} = 0$ for some $\theta_0 \in \Theta \subset \mathcal{R}^p$. Under an MNAR assumption such as (2.1), and given an estimator $\hat{\alpha}$ of α, a GMM estimator of θ can be obtained as $\hat{\theta}_{SP} = \arg\min_{\theta\in\Theta} \|\mathcal{G}_n(\theta, \hat{\alpha})\|_W$, where $\mathcal{G}_n(\theta, \alpha) = n^{-1}\sum_{i=1}^n \hat{\psi}(Y_i, X_i, \theta, \alpha)$ and $\hat{\psi}(Y_i, X_i, \theta_0, \alpha_0) = \delta_i\psi(Y_i, X_i, \theta_0) + (1-\delta_i)\hat{m}^0_\psi(X_i, \theta_0, \alpha_0)$, $\hat{m}^0_\psi(X_i, \theta_0, \alpha_0)$ is identical to $\hat{m}^0_\psi(X_i, \theta_0, h_0, \alpha_0)$ at (2.4) except that $\psi(Y_i, X_i, \theta_0, h_0)$ is replaced by $\psi(Y_i, X_i, \theta_0)$.

Let $m^0_\psi(X_i, \theta) = E\{\psi(Y_i, X_i, \theta)|X_i, \delta_i = 0\}$, $\tilde{\psi}(Y_i, X_i, \theta, \alpha_0) = \delta_i\psi(Y_i, X_i, \theta)$ $+(1-\delta_i)m^0_\psi(X_i, \theta, \alpha_0)$, $\widetilde{\mathcal{G}}_n(\theta, \alpha) = n^{-1}\sum_{i=1}^n \tilde{\psi}(Y_i, X_i, \theta, \alpha)$, and $\Lambda(\theta)$ be the partial derivative of $\mathcal{G}(\theta) = E\{\psi(Y, X, \theta)\}$ with respect to θ. Define a generic neighborhood $\Theta_\varrho =: \{\theta \in \Theta : \|\theta - \theta_0\| \le \varrho\}$ of θ_0 for some constant $\varrho > 0$.

Theorem 5. *Suppose the conditions* (C1)−(C4) *of the Supplementary Material hold, except that* $\psi(Y, X, \theta_0, h_0)$ *and* $m^0_\psi(X_i, \theta_0, h_0, \alpha_0)$ *are, respectively, replaced by* $\psi(Y, X, \theta_0)$ *and* $m^0_\psi(X_i, \theta_0, \alpha_0)$*, and that the response probability model* (2.1) *is correctly specified.*

(a) *If the function class* $\{\psi(Y, X, \theta) : \theta \in \Theta\}$ *is Glivenko-Cantelli,* $\hat{\alpha}$ *is a consistent estimator of* α *and* $\sup_{\theta\in\Theta} \|\mathcal{G}_n(\theta, \hat{\alpha}) - \widetilde{\mathcal{G}}_n(\theta, \alpha_0)\| = o_p(1)$*, so* $\hat{\theta}_{SP} - \theta_0 = o_p(1)$.

(b) *Assume that* $n^{1/2}(\hat{\alpha} - \alpha_0) = n^{-1/2}\sum_{i=1}^n \mathscr{C}(X_i, Y_i, \alpha_0) + o_p(1)$*, where* $\mathscr{C}(\cdot)$ *is an influence function. If the function class* $\{\psi(Y, X, \theta) : \theta \in \Theta_\varrho\}$ *is Donsker, for some constant* $K > 0$ *and* $\varsigma \in (0, 1]$ *each component of* $\psi(Y_i, X_i, \theta)$ *is uniformly* $L_2(P)$*-continuous with respect to* θ *in the sense that* $E\{\sup_{\theta,\theta'\in\Theta_\varrho} |\psi_j(Y, X, \theta) - \psi_j(Y, X, \theta')|^2\} \le K\varrho^{2\varsigma}$*; for all sequences* $\varrho_n =$

$o_p(1)$, $\sup_{\|\theta-\theta_0\|\le\varrho_n} \|\mathcal{G}_n(\theta,\hat{\alpha})-\widetilde{\mathcal{G}}_n(\theta,\alpha_0)-\mathcal{G}_n(\theta_0,\hat{\alpha})+\widetilde{\mathcal{G}}_n(\theta_0,\alpha_0)\| = o_p(n^{-1/2})$, *then*

$$n^{1/2}(\hat{\theta}_{SP}-\theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0,\Sigma_5),$$

where $\Sigma_5=(\Lambda^\top W\Lambda)^{-1}\Lambda^\top W\Gamma_5 W\Lambda(\Lambda^\top W\Lambda)^{-1}$, $\Gamma_5=\mathrm{Var}\{\mathcal{S}(X,Y,\theta)-H\mathscr{C}(X,Y,\alpha_0)\}$, $\mathcal{S}(X,Y,\theta) = \delta\{\pi(X,Y)\}^{-1}\{\psi(Y,X,\theta) - m^0_\psi(X,\theta)\} + m^0_\psi(X,\theta)$, $H = E[(1-\delta)\{\psi(Y,X,\theta_0) - m^0_\psi(X,\theta_0,\alpha_0)\}\{z(X,Y,\alpha_0) - m^0_z(X,\alpha_0)\}^\top]$, $z(X,Y,\alpha) = \partial\mathrm{logit}\{\pi(X,Y,\alpha)\}/\partial\alpha$, $m^0_z(X,\alpha) = E\{z(X,Y,\alpha)|X,\ \delta=0\}$, *and* $\mathrm{logit}(p)=\log\{p/(1-p)\}$.

It follows from Theorem 4 and Theorem 5 that the estimation of the infinite-dimensional nuisance function leads to the limiting distributions of estimators of parameters of interest defined via SEEs depending on the used nuisance parameter estimator.

For the parametric models $E\{\psi(Y,X,\theta_0)\}=0$ with ignorable missing covariates, Qin, Zhang and Leung (2009) presented an EL procedure to estimate unknown parameter θ_0 when the response probability model is known or parametrically estimated. We can extend this approach to our semiparametric models with nonignorable missing data by letting

$$\begin{aligned}\xi_1(Y_i,X_i,\theta,h,\alpha) &= \frac{\delta_i\psi(Y_i,X_i,\theta,h)}{\pi(X_i,Y_i,\alpha)},\\ \xi_2(Y_i,X_i,\theta,h,\alpha) &= \frac{\delta_i-\pi(X_i,Y_i,\alpha)}{\pi(X_i,Y_i,\alpha)}m^0_\psi(X_i,\theta,h,\alpha).\end{aligned} \tag{5.1}$$

In this case, we have $E\{\xi_j(Y_i,X_i,\theta_0,h_0,\alpha_0)\}=0$ for $j=1$ and 2. In particular, the unbiasedness of the second equation does not depend on the selection of $m^0_\psi(X_i,\theta,h,\alpha)$. Here, ξ_2 can be regarded as auxiliary information, used to improve upon the Horvitz-Thompson estimating function ξ_1. Also, the number of the SEEs $\xi=(\xi_1^\top,\xi_2^\top)^\top$ is greater than the dimension of parameter vector θ regardless of whether SEEs $\psi(Y_i,X_i,\theta,h)$ is just-identified or over-identified. Thus, the EL method (Owen (1990); Qin and Lawless (1994)) can be used to combine these over-identified unbiased SEEs to obtain an improved inference.

6. Simulation Studies

We used two simulation studies, including a partial nonlinear regression model, and a partial linear regression model, to evaluate the finite sample performance of the proposed methodologies.

Experiment 1 (Partial nonlinear regression model) In this experiment, the data were generated from the model $Y_i=\exp(X_i^\top\theta)+h(T_i)+\varepsilon_i$ for $i=1,\ldots,n$, where $h(t)=\cos(4\pi t)$, $X_i=(1,X_{i1},X_{i2})^\top$, $(X_{1i},X_{2i})^\top$, and T_i were

independently generated as $\mathcal{N}(0, \Sigma_x)$ and $\mathcal{U}(0,1)$, respectively, and the ε_i were independently generated as $\mathcal{N}(0,1)$ and $\mathcal{U}(0,1)$. We took the true values of $\theta = (\theta_1, \theta_2, \theta_3)^\top$ and $\Sigma_x = (\sigma_{xij})$ to be $\theta = (1, 1.5, 0.5)^\top$ and $\sigma_{xij} = 0.5^{|i-j|}$ for $1 \le i, j \le 2$, respectively. We assumed $Z_i = (X_{1i}, X_{2i}, T_i)^\top$'s were completely observed, but Y_i's were subject to missingness. With $\delta = 1$ if Y was observed and $\delta = 0$ if Y was missing, δ_i of Y_i was Bernoulli with probability $\pi_i(\alpha) := \pi(Z_i, Y_i, \alpha)$,

$$\pi_i(\alpha) = \frac{\exp(0.5 + 0.01X_{1i} + 0.01X_{2i} + 0.25T_i + 0.01Y_i)}{1 + \exp(0.5 + 0.01X_{1i} + 0.01X_{2i} + 0.25T_i + 0.01Y_i)}.$$

The response rate was about 67% for the above missingness data mechanism. We took sample size $n = 200$, and simulated 1,000 datasets. To estimate the propensity score, we considered a correctly specified model (**C**)

$$\pi_i(\alpha) = \frac{\exp(\alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \alpha_3 T_i + \alpha_4 Y_i)}{1 + \exp(\alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \alpha_3 T_i + \alpha_4 Y_i)},$$

and a misspecified model (**M**)

$$\pi_i(\alpha) = \Phi(\alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \alpha_3 T_i + \alpha_4 Y_i),$$

where $\Phi(\cdot)$ is the cumulative density function of the standard normal. Model (**M**) was used to investigate the robustness of the proposed *Propensity-Score-Based Nonparametric Imputation* procedure to the misspecified response probability model.

To illustrate the proposed methods, we constructed the SEEs

$$\begin{aligned}
&\mathbf{E1}: \psi(Y_i, Z_i, \theta, h) = \tilde{\mathcal{X}}_i \left\{ Y_i - \exp(X_i^\top \theta) - h(T_i) \right\}, \\
&\mathbf{E2}: \psi(Y_i, Z_i, \theta, h) = \tilde{X}_i \left\{ Y_i - \exp(X_i^\top \theta) - h(T_i) \right\},
\end{aligned}$$

where $\tilde{\mathcal{X}}_i = \tilde{X}_i - E(\tilde{X}_i | T_i)$, and $\tilde{X}_i = X_i \exp(X_i^\top \theta)$. Here, SEEs **E1** was constructed by using the first-order-condition for minimizing the objective function $Q_1(\theta) = \sum_{i=1}^n (Y_i - \exp(X_i^\top \theta) - h_\theta(T_i))^2$ in which $h_\theta(T) = E\{Y - \exp(X^\top \theta) | T\}$, whilst SEEs **E2** was constructed by using the first-order-condition for minimizing the objective function $Q_2(\theta) = \sum_{i=1}^n (Y_i - \exp(X_i^\top \theta) - h(T_i))^2$. Clearly, $E\{\psi(Y_i, Z_i, \theta, h)\} = 0$ for **E1** and **E2**.

Let $O(Z_i, Y_i, \alpha) = \pi^{-1}(Z_i, Y_i, \alpha) - 1$ and $\mathcal{K}_b(\cdot)$ be a univariate kernel function. We considered as estimator of $h(t)$,

$$\hat{h}_\theta(t) = \frac{\sum_{i=1}^n \delta_i O(Z_i, Y_i, \alpha) \mathcal{K}_b(t - T_i) \{Y_i - \exp(X_i^\top \theta)\}}{\sum_{i=1}^n \delta_i O(Z_i, Y_i, \alpha) \mathcal{K}_b(t - T_i)}, \tag{6.1}$$

a nonparametric regression estimator of

$$\tilde{h}_\theta^0(t) = \frac{E\{\delta(Y - \exp(X^\top\theta))O(Z,Y,\alpha)|T=t\}}{E\{\delta O(Z,Y,\alpha)|T=t\}}.$$

Here $\hat{h}_\theta(t)$ is not a consistent estimator of $h(t)$ because $\lim_{n\to\infty}\hat{h}_\theta(T) = \tilde{h}_\theta^0(T) \neq E\{(Y - \exp(X^\top\theta))|T\} = h(T)$ a.s. The conditional expectation $E(\tilde{X}_i|T_i)$ can be estimated via the same nonparametric method. Let

$$\hat{m}_\psi^0(Z,\theta,\hat{h}_\theta,\alpha) = \frac{\sum_{i=1}^n \delta_i O(Z_i,Y_i,\alpha)K_a(Z-Z_i)\psi(Y_i,Z_i,\theta,\hat{h}_\theta(T_i))}{\sum_{i=1}^n \delta_i O(Z_i,Y_i,\alpha)K_a(Z-Z_i)}, \tag{6.2}$$

where $K_a(\cdot)$ is a d_z-dimensional kernel function. Then, the modified SEEs for θ is given by

$$\mathcal{G}_n(\theta,\hat{h}_\theta,\alpha) = \frac{1}{n}\sum_{i=1}^n\{\delta_i\psi(Y_i,Z_i,\theta,\hat{h}_\theta(T_i)) + (1-\delta_i)\hat{m}_\psi^0(Z_i,\theta,\hat{h}_\theta,\alpha)\}.$$

Two approaches are employed to estimate α: a GMM-based validation sample method with 25% follow-up rate; an SEL method by incorporating the following auxiliary information

$$g(Z_i,Y_i,\alpha) = \begin{pmatrix} \delta_i\pi^{-1}(Z_i,Y_i,\alpha)(X_{1i}-\bar{X}_1) \\ \delta_i\pi^{-1}(Z_i,Y_i,\alpha)(X_{2i}-\bar{X}_2) \end{pmatrix},$$

where $\bar{X}_1 = n^{-1}\sum_{i=1}^n X_{1i}$, and $\bar{X}_2 = n^{-1}\sum_{i=1}^n X_{2i}$.

Given an estimator of α, we considered three estimators of θ: a validation sample-based estimator (*vse*), an SEL-based estimator (*sel*), Chen and Van Keilegom's (2013) estimator under MAR assumption (*mar*). The kernel function was taken to be the Gaussian kernel, and $\kappa = 20$ observations were imputed for each of missing Y_i's in computing Chen and Van Keilegom's (2013) estimator. The bandwidths b relating to (6.1) and a relating to (6.2) were taken to be $a = b = n^{-1/5}$.

Results are reported in Table 1, where 'Bias' denotes the absolute difference between the true value and the mean of the estimates based on 1,000 replications, 'RMS' is the root mean square between the estimates based on 1,000 replications and its true value, 'Std' is the standard deviation of estimates based on 1,000 replications. Examination of Table 1 reveals the following findings: under the considered settings, our proposed semiparametric estimators *vse* and *sel* perform well in the sense that their corresponding Biases are quite close to zero and their corresponding values of RMS are relatively close to those of Std; the performances of our proposed semiparametric estimators computed using the misspecified response probability model (**M**) do not differ much from that of using the correctly

specified response probability model (**C**), our proposed estimators are robust to the misspecified response probability model; the proposed semiparametric estimator *sel* has a slight advantage over the estimator *vse* because *sel* provides smaller RMS and Std than *vse* in most cases; the *mar* estimator has larger values of Bias and RMS than of the *vse* and *sel* estimators.

We also investigated the performance of the proposed estimators for the response model parameter α under the correctly specified response probability model (**C**). Their corresponding values of Bias, RMS, and Std are in Table 2. Inspection of Table 2 indicates that the proposed two estimators are nearly unbiased, and the semi-empirical likelihood method outperforms the GMM-based validation sample method in terms of RMS and Std.

Experiment 2 (Partial linear regression model) To investigate the performance of the proposed bootstrapping approach to approximate variance estimation of our estimators, we conducted a second simulation study. Here, the data were generated from the model $Y_i = X_i\theta + h(T_i) + \varepsilon_i$ for $i = 1, \ldots, n$, where $h(t) = \cos(4\pi t)$, X_i's were independently $\mathcal{N}(0,1)$, T_i's were independently $\mathcal{U}(0,1)$ and then sorted in ascending order, ε_i's were independently $\mathcal{N}(0,1)$ and $\mathcal{U}(0,1)$. The true value of θ was set to 1. We assumed the $Z_i = (X_i, T_i)^\top$'s were completely observed, but the Y_i's subject to missingness. With $\delta_i = 1$ if Y_i was observed, and $\delta_i = 0$ if not. The δ_i were independently Bernoulli with probability $\pi_i(\alpha) := \pi(Z_i, Y_i, \alpha)$ specified by

Model I: $\pi_i(\alpha) = \exp(\tilde{\alpha}_0 + \tilde{\alpha}_1 Y_i)/(1 + \exp(\tilde{\alpha}_0 + \tilde{\alpha}_1 Y_i))$, where $\alpha = (\tilde{\alpha}_0, \tilde{\alpha}_1)^\top$, and the true value of $\tilde{\alpha}_1$ was $\tilde{\alpha}_1 = 0.2$;

Model II: $\pi_i(\alpha) = \exp(\tilde{\alpha}_0 + \tilde{\alpha}_1 X_i + \tilde{\alpha}_2 Y_i)/(1 + \exp(\tilde{\alpha}_0 + \tilde{\alpha}_1 X_i + \tilde{\alpha}_2 Y_i))$, where $\alpha = (\tilde{\alpha}_0, \tilde{\alpha}_1, \tilde{\alpha}_2)^\top$, and the true values of $\tilde{\alpha}_1$ and $\tilde{\alpha}_2$ were $\tilde{\alpha}_1 = 0.5$ and $\tilde{\alpha}_2 = 0.2$.

For these models, considered in Qin, Leung and Shao (2002), we took the true value of $\tilde{\alpha}_0$ to be $2.5, 1.5, 1.0, 0.5, 0.01$, leading to the average missing proportions 7%, 17%, 25%, 36% and 47% for Model I, and 10%, 21%, 29%, 39% and 50% for Model II, respectively.

The proposed SEL method was adopted to estimate α in Model I and Model II by incorporating the auxiliary information $g(X_i, T_i) = (X_i - \bar{X}, T_i - \bar{T})$ with $\bar{X} = n^{-1}\sum_{i=1}^n X_i$ and $\bar{T} = n^{-1}\sum_{i=1}^n T_i$. Under the MNAR assumption, $h(t) = E\{(\delta Y + (1-\delta)m_Y^0(Z) - X\theta)|T = t\}$, where $m_Y^0(Z) = E(Y|X, T, \delta = 0)$, and a nonparametric estimator of $m_Y^0(Z)$ is $\hat{m}_Y^0(Z) = \sum_{j=1}^n \delta_j O(Z_j, Y_j, \alpha) K_a(Z - Z_j) Y_j / \sum_{\ell=1}^n \delta_\ell O(Z_\ell, Y_\ell, \alpha) K_a(Z - Z_\ell)$, where $O(Z_j, Y_j, \alpha) = \pi^{-1}(Z_j, Y_j, \alpha) - 1$. Then, a consistent nonparametric estimator of $h(t)$ is

$$\hat{h}_\theta(t) = \sum_{i=1}^n W_{ni}(t)\{\delta_i Y_i + (1-\delta_i)\hat{m}_Y^0(Z_i) - X_i\theta\}, \tag{6.3}$$

Table 1. Performance of various estimators in the simulation study: Experiment 1.

				$\varepsilon \sim \mathcal{N}(0,1)$			$\varepsilon \sim \mathcal{U}(0,1)$		
SEEs	Model	Methods	Est.	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$
E1	**C**	*vse*	Bias	0.000	0.000	0.000	0.000	0.000	0.000
			RMS	0.022	0.009	0.006	0.014	0.006	0.004
			Std	0.022	0.009	0.006	0.014	0.006	0.004
		sel	Bias	0.000	0.000	0.000	0.000	0.000	0.000
			RMS	0.022	0.009	0.006	0.013	0.006	0.004
			Std	0.022	0.009	0.006	0.013	0.006	0.004
		mar	Bias	-0.850	0.248	0.087	-0.862	0.216	0.125
			RMS	1.397	0.417	0.300	2.645	0.370	1.139
			Std	1.110	0.335	0.287	2.503	0.301	1.133
	M	*vse*	Bias	0.001	0.000	0.000	0.000	0.000	0.000
			RMS	0.024	0.010	0.006	0.014	0.006	0.004
			Std	0.024	0.010	0.006	0.014	0.006	0.004
		sel	Bias	0.001	0.000	0.000	0.000	0.000	0.000
			RMS	0.023	0.009	0.006	0.013	0.006	0.004
			Std	0.023	0.009	0.006	0.013	0.006	0.004
E2	**C**	*vse*	Bias	0.006	-0.002	-0.001	0.004	-0.001	-0.001
			RMS	0.034	0.013	0.006	0.019	0.008	0.004
			Std	0.033	0.013	0.006	0.018	0.008	0.004
		sel	Bias	0.003	-0.001	0.000	0.002	0.000	0.000
			RMS	0.022	0.009	0.006	0.014	0.006	0.004
			Std	0.022	0.009	0.006	0.014	0.006	0.004
		mar	Bias	-0.387	0.057	0.078	-0.385	0.057	0.077
			RMS	2.725	0.332	1.223	2.688	0.330	1.207
			Std	2.700	0.328	1.222	2.663	0.326	1.205
	M	*vse*	Bias	0.008	-0.003	-0.001	0.009	-0.003	-0.001
			RMS	0.086	0.034	0.012	0.069	0.026	0.010
			Std	0.086	0.034	0.012	0.069	0.025	0.010
		sel	Bias	0.013	-0.004	-0.002	0.012	-0.004	-0.002
			RMS	0.083	0.033	0.012	0.070	0.027	0.010
			Std	0.082	0.033	0.012	0.069	0.026	0.010

where $W_{ni}(t) = \mathcal{K}_b(t - T_i)/\sum_{j=1}^{n} \mathcal{K}_b(t - T_j)$, and $\mathcal{K}_b(\cdot)$ is a univariate kernel function. To use the method to estimate θ, we consider SEEs: $\psi(Y, Z, \theta, h) = X(Y - X\theta - h(t))$.

For comparison, we considered two approaches to estimate θ:

(i) the *Propensity-Score-Based Nonparametric Imputation* estimator $\hat{\theta}_{sp}$: the so-

Table 2. Performance of $\hat{\alpha}$ under model **C** in the simulation study: Experiment 1.

		$\varepsilon \sim \mathcal{N}(0,1)$		$\varepsilon \sim \mathcal{U}(0,1)$	
Parameter	Method	*vse*	*sel*	*vse*	*sel*
$\hat{\alpha}_0$	Bias	0.080	0.080	0.089	0.089
	RMS	0.315	0.315	0.302	0.302
	Std	0.305	0.305	0.289	0.289
$\hat{\alpha}_1$	Bias	0.018	0.038	0.004	0.026
	RMS	0.354	0.186	0.251	0.152
	Std	0.354	0.182	0.251	0.150
$\hat{\alpha}_2$	Bias	0.005	0.002	-0.003	0.010
	RMS	0.232	0.101	0.304	0.098
	Std	0.232	0.101	0.305	0.098
$\hat{\alpha}_3$	Bias	-0.013	-0.018	0.014	0.009
	RMS	0.427	0.390	0.386	0.366
	Std	0.427	0.390	0.386	0.366
$\hat{\alpha}_4$	Bias	0.013	-0.005	0.014	-0.005
	RMS	0.102	0.018	0.091	0.016
	Std	0.101	0.017	0.090	0.015

lution to $\mathcal{G}_n(\theta, \hat{h}_\theta, \hat{\alpha}) = 0$, where

$$\mathcal{G}_n(\theta, \hat{h}_\theta, \alpha) = \frac{1}{n}\sum_{i=1}^{n}\{\delta_i \psi(Y_i, Z_i, \theta, \hat{h}_\theta(T_i)) + (1-\delta_i)\hat{m}_\psi^0(Z_i, \theta, \hat{h}_\theta, \alpha)\} \tag{6.4}$$

in which $\hat{m}_\psi^0(Z_i, \theta, \hat{h}_\theta, \alpha)$ is defined at (6.2);

(ii) Qin, Zhang and Leung's (2009) estimator $\hat{\theta}_{el}$, an EL estimator of θ found by maximizing the EL ratio function

$$\hat{\ell}_e(\theta) = -\sum_{i=1}^{n} \log\left\{1 + \lambda^\top \xi(Y_i, Z_i, \theta, \hat{h}_\theta, \hat{\alpha})\right\},$$

where $\xi(Y_i, Z_i, \theta, \hat{h}_\theta, \hat{\alpha}) = (\xi_1(Y_i, Z_i, \theta, \hat{h}_\theta, \hat{\alpha}), \xi_2(Y_i, Z_i, \theta, \hat{h}_\theta, \hat{\alpha}))^\top$, in which ξ_1 and ξ_2 are defined at (5.1) except that $m_\psi^0(X_i, \theta, h, \alpha)$ is replaced by $\hat{m}_\psi^0(Z_i, \theta, \hat{h}_\theta, \hat{\alpha})$ and $\pi(X_i, Y_i, \alpha)$ is replaced by $\pi(X_i, Y_i, \hat{\alpha})$, and $\lambda = \lambda(\theta)$ is a solution to

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\xi(Y_i, Z_i, \theta, \hat{h}_\theta, \hat{\alpha})}{1 + \lambda^\top \xi(Y_i, Z_i, \theta, \hat{h}_\theta, \hat{\alpha})} = 0.$$

Here, we considered $n = 200$ together with 500 replications, the Gaussian kernel, and bandwidths $a = 0.3$ and $b = 0.2$. To approximate asymptotic variances and evaluate confidence intervals of estimators $\hat{\theta}_{sp}$ and $\hat{\theta}_{el}$ we used the bootstrap sample $B = 100$.

Table 3 presents the values of Bias, standard deviation (Std), and standard error (SE) evaluated by using the bootstrap method described in Section 4, and coverage probability (CP) of the approximate 95% confidence intervals for $\hat{\theta}_{sp}$ and $\hat{\theta}_{el}$. Examination of Table 3 shows that $\hat{\theta}_{sp}$ and $\hat{\theta}_{el}$ behave well in that their absolute values of Bias and Std are less than 0.07 and 0.17, respectively; the coverage probabilities are quite close to the pre-specified confidence level 95% when the nonresponse proportion is not too high, the values of SE are rather close to those of Std, indicating that our presented bootstrap method performs well; the smaller $\tilde{\alpha}_0$ is, the larger the value of SE or Std is; the higher the nonresponse proportion is, the larger the difference between the empirical coverage probability and the pre-specified confidence level is; our proposed estimator $\hat{\theta}_{sp}$ performs better than does Qin, Zhang and Leung's (2009) EL estimator because the SEs/Stds of the former are less than those of the latter; increasing the mean response rates improves the accuracy of parameter estimate and the empirical coverage of confidence interval, as expected.

7. An Example

The Mobility Program Clinical Research Unit of St. Michael's Hospital, affiliated with the University of Toronto, conducted a study aimed at understanding prognostic factors associated with more successful outcomes (return to work or with higher at work productivity) after an upper limb injury. As an illustration of the proposed methods, we focused on a subset of samples with missing values on response variable only, but completely observed on the selected predictor variables. The response variable is the work productivity after one year of injury. The Work Limitations Questionnaire (WLQ) (Amick et al. (2004); Lerner et al. (2001); Lerner et al. (2002)) was used as a measure of at-work productivity loss. The reverse of this scale was used as a measure of work productivity (response variable (Y)). The response measure was only completed by persons who were working at the time of assessment and missing for those not at work. The potential work productivity for those away from work was much lower if they were at work. Therefore, the missing data mechanism of Y was not at random. About 48% of participants was missing on this response variable. In this analysis, we were interested in the predictor variables pain disorder score (x_1), better mental health factor score (x_2), and supervisor support (x_3). We also wanted to know whether supervisor support moderated the effects of level of pain disorder and mental health status on work productivity, interactions x_1x_3 and x_2x_3 were included in the regression model. The controlled covariate was participants' age (t). A sample of 347 was used for the current analysis.

Figure 1 shows the scatter plot of work limitation questionnaire index score (Y) against age (t) and the fitted smoothing spline, ignoring those with missing values on Y. Clearly, the relationship between work productivity and age is

Table 3. Performance of $\hat{\theta}_{sp}$ and $\hat{\theta}_{el}$ in the simulation study: Experiment 2.

	Missing			$\tilde{\alpha}_0$				
ε_i	mechanism	Estimator		2.5	1.5	1.0	0.5	0.01
$\mathcal{N}(0,1)$	ModelI	$\hat{\theta}_{sp}$	Bias	0.002	0.004	0.005	-0.001	-0.003
			SE	0.090	0.097	0.104	0.113	0.127
			Std	0.088	0.098	0.104	0.114	0.134
			CP	0.946	0.946	0.940	0.942	0.924
		$\hat{\theta}_{el}$	Bias	-0.015	-0.013	-0.015	-0.023	-0.033
			SE	0.110	0.111	0.115	0.122	0.136
			Std	0.103	0.105	0.111	0.121	0.142
			CP	0.968	0.966	0.954	0.942	0.924
	Model II	$\hat{\theta}_{ps}$	Bias	-0.004	-0.009	-0.009	-0.011	-0.016
			SE	0.093	0.103	0.112	0.125	0.139
			Std	0.094	0.105	0.113	0.134	0.155
			CP	0.946	0.944	0.938	0.930	0.904
		$\hat{\theta}_{el}$	Bias	-0.019	-0.032	-0.037	-0.051	-0.067
			SE	0.109	0.113	0.121	0.136	0.154
			Std	0.102	0.113	0.123	0.142	0.166
			CP	0.958	0.942	0.920	0.902	0.874
$\mathcal{U}(0,1)$	ModelI	$\hat{\theta}_{sp}$	Bias	-0.001	0.000	0.000	0.000	0.001
			SE	0.054	0.057	0.061	0.067	0.074
			Std	0.054	0.057	0.062	0.068	0.078
			CP	0.934	0.942	0.940	0.926	0.922
		$\hat{\theta}_{el}$	Bias	-0.009	-0.010	-0.010	-0.014	-0.017
			SE	0.066	0.066	0.068	0.073	0.082
			Std	0.061	0.062	0.064	0.069	0.080
			CP	0.950	0.964	0.954	0.930	0.926
	Model II	$\hat{\theta}_{ps}$	Bias	0.002	0.004	0.005	0.012	0.008
			SE	0.056	0.064	0.069	0.077	0.087
			Std	0.056	0.064	0.072	0.082	0.090
			CP	0.934	0.942	0.934	0.940	0.932
		$\hat{\theta}_{el}$	Bias	-0.013	-0.016	-0.018	-0.020	-0.035
			SE	0.066	0.072	0.077	0.084	0.095
			Std	0.063	0.070	0.077	0.087	0.098
			CP	0.956	0.948	0.928	0.910	0.888

not linear. Therefore, we considered the semi-parametric linear regression models for the response process $Y_i = X_i^\top \theta + h(t_i) + \varepsilon_i$ for $i = 1, \ldots, 347$, where $X_i = (x_{1i}, x_{2i}, x_{3i}, x_{1i}x_{3i}, x_{2i}x_{3i})^\top$ and $\theta = (\theta_1, \theta_2, \theta_3, \theta_4, \theta_5)^\top$. This model assumes that work productivity depends linearly on predictor variables in X_i but nonlinearly on age (t). Motivated by Lee and Tang (2006), we considered the nonignorable missingness data mechanism models

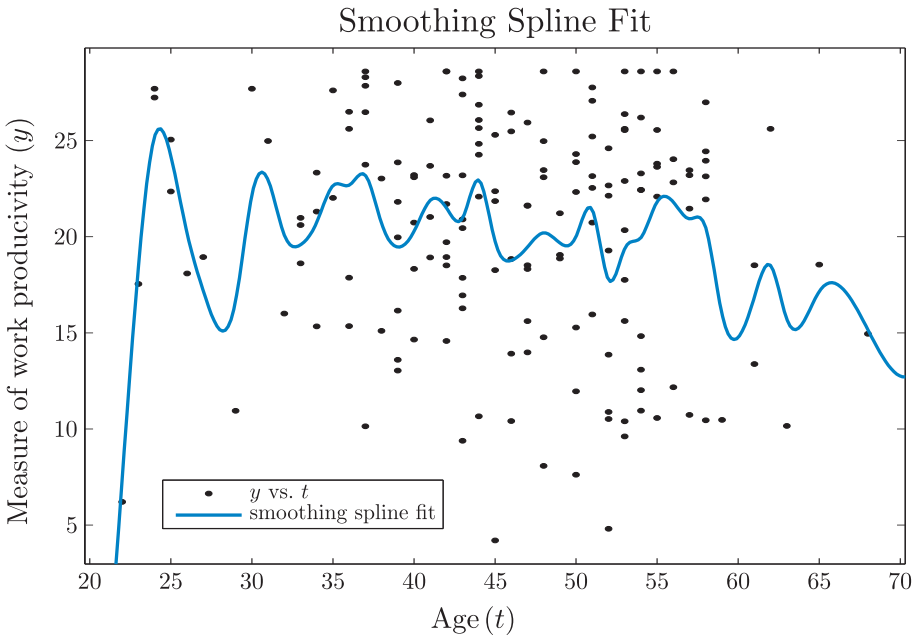

Figure 1. Canada WSIB data, plot of observations, and a smoothing spline fit using complete-case analysis.

Model 1: $\pi(Z_i, Y_i, \alpha) = \dfrac{\exp(\alpha_0 + \alpha_1^\top X_i + \alpha_2 t_i + \alpha_3 Y_i)}{1 + \exp(\alpha_0 + \alpha_1^\top X_i + \alpha_2 t_i + \alpha_3 Y_i)}$,

Model 2: $\pi(Z_i, Y_i, \alpha) = \Phi(\alpha_0 + \alpha_1^\top X_i + \alpha_2 t_i + \alpha_3 Y_i)$, where $\alpha = (\alpha_0, \alpha_1^\top, \alpha_2, \alpha_3)^\top$ and $Z_i = (X_i^\top, t_i)^\top$. Models 1 and 2 were adopted to investigate the sensitivity of the proposed procedures to the potentially misspecified response probability models. To illustrate our methods, we considered the SEEs $\psi(Y_i, Z_i, \theta, h) = \tilde{X}_i \{Y_i - X_i^\top \theta - h(t_i)\}$, where $\tilde{X}_i = X_i - E(X_i|t_i)$ and $Z_i = (X_i^\top, t_i)^\top$. A nonparametric estimator $\hat{h}_\theta(t)$ of $h(t)$ was constructed by using $X_i^\top \theta$ to replace $X_i \theta$ at (6.3), and the kernel function was the Gaussian with the bandwidths $\hat{\sigma}_t n^{-1/5}$, where $\hat{\sigma}_t$ was the standard deviation of observations $\{t_i : i = 1, \ldots, 347\}$.

The GMM-based validation sample method with 25% follow-up rate and the semi-parametric empirical likelihood method were employed to estimate unknown parameters in α. To implement the semi-parametric empirical likelihood method, we considered the auxiliary information $g_k(Z_i, Y_i, \alpha) = \delta_i \pi^{-1}(Z_i, Y_i, \alpha)(X_{ki} - \bar{X}_k)$ for $k = 1, 2, 3, 4$ and 5, where $\bar{X}_k = n^{-1} \sum_{i=1}^n X_{ki}$ and X_{ki} is the kth component of X_i. Then, the semi-parametric estimators of θ, *vse* and *sel*, were obtained from a set of the imputed SEEs as given at (6.4). The standard errors (SE) of the proposed estimators were evaluated by using the bootstrap approach. The results are given in Table 4. Examination of Table 4 indicates that the assumed

Table 4. Estimated parameters and standard errors in illustrative example.

Model	Methods	Statistic	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{\theta}_5$
Model 1	*vse*	EST	-1.595	1.280	-0.353	0.153	0.665
		SE	0.474	0.535	0.618	0.495	0.570
	sel	EST	-1.663	1.151	-0.118	1.006	1.662
		SE	0.524	0.480	0.545	0.586	0.584
Model 2	*vse*	EST	-1.781	1.458	-0.555	0.209	0.205
		SE	0.687	0.878	0.827	0.654	0.738
	sel	EST	-1.783	1.424	-0.300	0.125	0.643
		SE	0.508	0.546	0.549	0.576	0.564

response probability models led to quite similar parameter estimates, which suggests that our proposed estimators are insensitive to the choice of response probability models; the proposed semiparametric estimator *sel* has a slight advantage over estimator *vse* because of a smaller SE.

We computed $\hat{h}_\theta(t)$ via (6.3) with different parametrically estimated propensity scores, and present the corresponding estimated curves for $h(t)$ in Figure 2. From Figure 2, we find that the relationship between age and work productivity is negative, but non-linear (ignoring those at two ends of age range because of fewer data points before age 20 and after age 65). The results indicate that the level pain disorder is negatively related to work productivity and better mental health is positively related to the work productivity. This together with the results given in Table 4 imply an interesting finding that the social supports enhance the positive relationship between mental health and work productivity.

8. Discussion

Under MNAR, we have developed kernel-assisted SEE imputation based on propensity scores to estimate parameters of interest for a general class of semiparametric models. The proposed method is applicable if the propensity scores can be estimated parametrically. To obtain a consistent estimator of the propensity score, we consider a validation-sample-based method and a semi-empirical likelihood approach using available observations. The semi-empirical likelihood method is promising since it allows one to incorporate auxiliary information from the calibration constraints for the data with MNAR mechanism, and is able to achieve high efficiency. also promising because it can achieve both good robustness and efficiency.

There are some related research topics that require further investigation. For example, it is of interest to generalize the proposed propensity-score-based and kernel-assisted SEEs imputation approach from a cross-sectional study to a

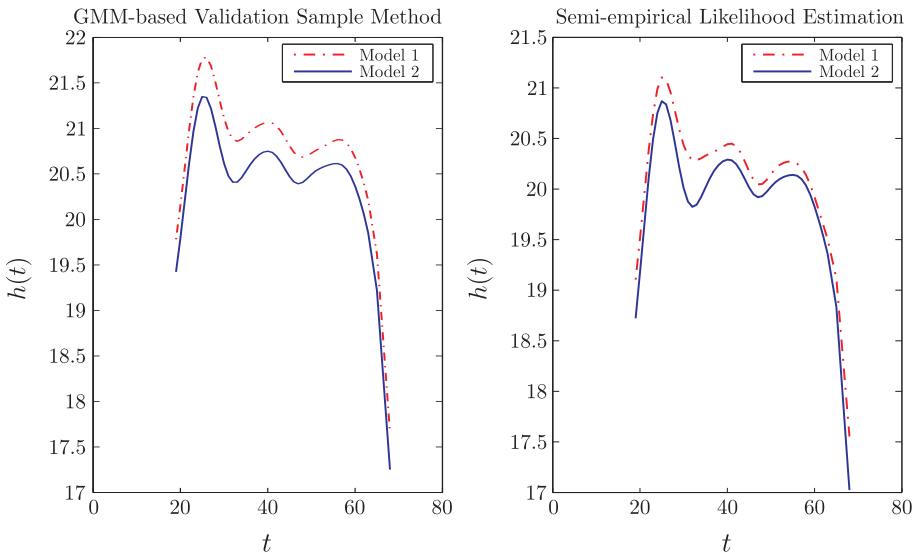

Figure 2. Canada WSIB data, estimated curve of $h(t)$ using the proposed propensity score based nonparametric imputation.

longitudinal study (Qu, Lindsay and Li (2000)), and to explore doubly robust estimation using SEEs inference under the MNAR mechanism. It is also important to develop diagnostic measures for the GMM or EL approach using SEEs (Zhu et al. (2008)), in additional to constructing EL confidence regions for parameters and EL confidence bands for the nonparametric functional component under the general class of SEEs for nonignorable missing data.

Supplementary Material

Supplementary Materials available in the attached file include technical conditions and proofs.

Acknowledgement

The authors are grateful to the Editor, an associate editor, and two referees for their valuable suggestions and comments that greatly improved the manuscript. This research was supported by grants from the National Science Fund for Distinguished Young Scholars of China (11225103), National Nature Science Foundation of China (11301464, 11301465), the US National Science Foundation (DMS-1308227), and research funds from Manitoba Health Research Council (MHRC).

References

Amick, B. C., Habeck, R. V., Ossmann, J., Fossel, A. H., Keller, R. and Katz, J. N. (2004). Predictors of successful work role functioning after carpal tunnel release surgery. *J. Occup. Environ. Med.* **46**, 490-500.

Chang, T. and Kott, P. (2008). Using calibration weighting to adjust for nonresponse under a plausible model. *Biometrika* **95**, 555-571.

Chen, X., Hong, H. and Tarozzi, A. (2008). Semiparametric efficiency in GMM models with auxiliary data. *Ann. Statist.* **36**, 808-843.

Chen, X., Linton, O. B. and Van Keilegom, I. (2003). Estimation of semiparametric models when the criterion function is not smooth. *Econometrica* **71**, 1591-1608.

Chen, S. X. and Van Keilegom, I. (2013). Estimation in semiparametric models with missing data. *Ann. Inst. Statist. Math.* **65**, 785-805.

Cheng, P. E. (1994). Nonparametric estimation of mean functionals with data missing at random. *J. Amer. Statist. Assoc.* **89**, 81-87.

Devroye, L. P. and Wagner, T. J. (1980). Distribution-free consistency results in nonparametric discrimination and regression function estimation. *Ann. Statist.* **8**, 231-239.

Graham, B. S. (2011). Efficiency bounds for missing data models with semiparametric restrictions. *Econometrica* **79**, 437-452.

Hansen, L. (1982). Large sample properties of generalized method of moments estimators. *Econometrica* **50**, 1029-1054.

Hu, Z., Follmann, D. A. and Qin, J. (2010). Semiparametric dimension reduction estimation for mean response with missing data. *Biometrika* **97**, 305-319.

Kim, J. K. and Yu, C. L. (2011). A semiparametric estimation of mean functionals with nonignorable missing data. *J. Amer. Statist. Assoc.* **106**, 157-165.

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

Lee, S. Y. and Tang, N. S. (2006). Bayesian analysis of nonlinear structural equation models with nonignorable missing data. *Psychometrika* **71**, 541-564.

Lipsitz, S. R., Ibrahim, J. G., Chen, M. and Peterson, H. (1999). Non-ignorable missing covariates in generalized linear models. *Statist. Medicine* **18**, 2435-2448.

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

Lerner, D., Amick III, B. C., Rogers, W. H., Malspeis, S., Bungay, K. and Cynn, D. (2001). The work limitations questionnaire. *Medical Care* **39**, 72-85.

Lerner, D., Reed, J. I., Massarotti, E., Wester, L. M. and Burke, T. A. (2002). The work limitations questionnaire's validity and reliability among patients with osteoarthritis. *J. Clin. Epidemi.* **55**, 197-208.

Owen, A. B. (1990). Empirical likelihood confidence regions. *Ann. Statist.* **18**, 90-120.

Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *Ann. Statist.* **22**, 300-325.

Qin, J., Leung, D. and Shao, J. (2002). Estimation with survey data under nonignorable nonresponse or informative sampling. *J. Amer. Statist. Assoc.* **97**, 193-200.

Qin, J., Zhang, B. and Leung, D. (2009). Empirical likelihood in missing data problems. *J. Amer. Statist. Assoc.* **104**, 1492-1503.

Qu, A., Lindsay, B. G. and Li, B. (2000). Improving generalised estimating equations using quadratic inference functions. *Biometrika* **87**, 823-836.

Riddles, M. K., Kim, J. K. and Im, J. (2015). Propensity score adjustment for nonignorable nonresponse. *J. Survey Statist. Method.*, Under 2nd revision.

Robins, J. M. and Ritov, Y. (1997). Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Statist. Medicine* **16**, 285-319.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.* **89**, 846-866.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York.

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

Tang, N. S., Zhao, P. Y. and Zhu, H. T. (2014). Empirical likelihood for estimating equations with nonignorably missing data. *Statist. Sinica* **24**, 723-748.

Troxel, A. B., Lipsitz, S. R. and Brennan, T. A. (1997). Weighted estimating equations with nonignorably missing response data. *Biometrics* **53**, 857-869.

Wang, D. and Chen, S. X. (2009). Empirical likelihood for estimating equations with missing values. *Ann. Statist.* **37**, 490-517.

Wang, S., Cui, H. and Li, R. (2013). Empirical likelihood inference for semi-parametric estimating equations. *Sci. China Math.* **56**, 1247-1262.

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

Zhao, J. and Shao, J. (2015). Semiparametric pseudo likelihoods in generalized linear models with nonignorable missing data. *J. Amer. Statist. Assoc.*, DOI: 10.1080/01621459.2014.983234.

Zhou, Y., Wan, A. T. and Wang, X. (2008). Estimating equations inference with missing data. *J. Amer. Statist. Assoc.* **103**, 1187-1199.

Zhu, H., Ibrahim, J. G., Tang, N. S. and Zhang, H. (2008). Diagnostic measures for empirical likelihood of general estimating equations. *Biometrika* **95**, 489-507.

Department of Statistics, Yunnan University, Kunming 650091, China.

E-mail: pyzhao@live.cn

Department of Statistics, Yunnan University, Kunming 650091, China.

E-mail: nstang@ynu.edu.cn

Department of Statistics, University of Illinois at Urbana-Champaign, 725 S. Wright Street Champaign, IL 61820 USA.

E-mail: anniequ@illinois.edu

Department of Community Health Sciences, University of Manitoba, Winnipeg, Manitoba, R3E 0W3 Canada.

E-mail: depeng.jiang@umanitoba.ca

(Received February 2015; accepted January 2016)