Ann Inst Stat Math (2019) 71:1201–1232
https://doi.org/10.1007/s10463-018-0672-y

Semiparametric estimation in regression with missing covariates using single-index models

Zhuoer Sun[1] · **Suojin Wang**[1]

Received: 16 September 2017 / Revised: 4 April 2018 / Published online: 19 June 2018

Abstract We investigate semiparametric estimation of regression coefficients through generalized estimating equations with single-index models when some covariates are missing at random. Existing popular semiparametric estimators may run into difficulties when some selection probabilities are small or the dimension of the covariates is not low. We propose a new simple parameter estimator using a kernel-assisted estimator for the augmentation by a single-index model without using the inverse of selection probabilities. We show that under certain conditions the proposed estimator is as efficient as the existing methods based on standard kernel smoothing, which are often practically infeasible in the case of multiple covariates. A simulation study and a real data example are presented to illustrate the proposed method. The numerical results show that the proposed estimator avoids some numerical issues caused by estimated small selection probabilities that are needed in other estimators.

Keywords Asymptotic efficiency · Generalized estimating equation · Kernel estimation · Missing at random · Regression · Single-index model

1 Introduction

Standard methods for regression generally require fully observed data. In practice, however, for the regression analysis of a response, Y, on a set of covariates (X, Z), the covariate X may be missing for some subjects. This is common in many areas such as biomedical sciences and clinical trials due to different reasons, including unavailability of covariate measurements, loss of data and survey non-response. For example, in

✉ Suojin Wang
sjwang@stat.tamu.edu

[1] Department of Statistics, Texas A&M University, College Station, TX 77843, USA

a subset of Canada 2010/2011 Youth Smoking Survey (YSS) data as described in Sect. 6, 144 students (total $n = 493$) had their BMI (body mass index) missing. Here we consider the linear regression model $Y = W\beta + \varepsilon$, where $W = (1, X, Z^\top)^\top$ is the covariate vector, and $E(\varepsilon|W) = 0$. The objective of this regression analysis is to estimate the regression coefficients β when the scalar covariate X is assumed to be missing at random (MAR) in the sense of Rubin (1976).

Much work using the maximum likelihood has been developed in the area of missing covariate regression analysis. This model-based method is flexible and clear for inference, and the asymptotic properties can be obtained via the second derivatives of the log-likelihood. However, the observed likelihood is generally difficult to get in a closed form for most missing data problems, which needs factorization and reparameterization of the likelihood. These can be achieved with multivariate normal model (Hartley and Hocking 1971) or specific monotone patterns of missing (Little and Rubin 2014). When likelihood factorization is not available, the EM algorithm is a popular technique for obtaining the maximum likelihood estimate (MLE) with ignorable missing categorical or continuous covariates (Fuchs 1982; Schluchter and Jackson 1989; Ibrahim 1990).

Another likelihood-based approach is Bayesian methods, which are straightforward in terms of concepts and inferences. On the other hand, it can be challenging to correctly specify the conditional covariate distribution and the joint priors over parameters. Ibrahim et al. (2002) considered Bayesian methods for MAR covariates in GLMs with informative prior based on historical data.

The most popular approach in industry and software packages might be multiple imputation (MI). The basic idea of MI is to impute missing data to create a new "complete" sample and then analyze it as if it were a complete data set. MI methods for MAR covariates in linear regression models are discussed in Rubin (2004) and Little and Rubin (2014). In terms of specifying the conditional covariate distribution, MI works similarly to the EM algorithm and the motivation is Bayesian, but the idea of MI itself is quite general and can be applied to other methods (Ibrahim et al. 2005). Hsu et al. (2014) proposed a nearest neighbor-based nonparametric multiple imputation approach for missing covariate data by using the distance calculated from a two-dimensional summary score of information about the missing covariate and the missingness indicator.

In most situations, all the above three methods depend on the specification of the likelihood. When the distributional assumptions are correct, they are optimal. However, they can give biased estimation when the assumptions are violated. To have more robust estimation with less likelihood assumptions, some semiparametric methods such as weighted estimating equations (WEEs) are proposed. Robins et al. (1994) proposed a class of semiparametric estimators based on inverse-probability WEE

$$\Delta(\beta, \psi) = n^{-1/2} \sum_{i=1}^{n} \left\{ \frac{R_i}{\pi_i} D_i(W_i)(y_i - W_i^\top \beta) + \left(1 - \frac{R_i}{\pi_i}\right) \psi_i(Q_i) \right\} = 0,$$

where n is the total sample size (including incomplete cases), $Q_i = (y_i, Z_i^\top)^\top$, $D_i(W_i)$ is a function satisfying a local identification condition as non-singular $E\{D_i(W_i)W_i^\top\}$,

$\psi_i(Q_i)$ is an arbitrary function of Q_i, R_i is the binary indicator such that $R_i = 1$ if the covariate X_i is observed and $R_i = 0$ otherwise, $\pi_i = E(R_i|Q_i)$ is the probability of observing X_i. Robins et al. (1994) discussed the choices of D_i and ψ_i and showed that there exist unique $D_i^{(\text{eff})}$ and corresponding $\psi_i^{(\text{eff})}(Q_i) = E\{D_i^{(\text{eff})}(W_i)(y_i - W_i^\top\beta)|Q_i\}$ that can achieve the semiparametric efficiency bound of $\hat{\beta}$. However, $D_i^{(\text{eff})}$ does not always have a closed form, so we choose a convenient one as $D_i = \partial\mu_i/\partial\beta = W_i$, where $\mu_i = E(y_i|X_i, Z_i)$. Then $\psi_i(Q_i) = E(T_i|Q_i)$ with $T_i = W_i(y_i - W_i^\top\beta)$ the regular score function. If we let $\psi_i \equiv 0$, the resulting estimator is the inverse-probability weighted estimator (IPW). We usually fit a logistic regression model of R_i on Q_i to estimate π_i's in the equations. When the logistic regression model is incorrect for π_i, the estimation can be biased. To overcome this difficulty, Wang et al. (1997) proposed a nonparametric kernel smoother for the selection probabilities and developed asymptotic theory for the estimator, including the optimal bandwidth rate. When $\psi_i \neq 0$, the estimator is the augmented inverse-probability weighted estimator (AIPW). This estimator is generally more efficient than IPW because it also incorporates the incomplete cases. The most important advantage of AIPW is that it has double robustness (DR) in the sense that the estimator will be consistent when either the selection probability model (π_i) or the augmentation model ($\psi_i(Q_i)$) is correctly specified. There are many works discussing the DR estimator and extensions, such as Bang and Robins (2005), Kang and Schafer (2007) and Robins and Ritov (1997). AIPW can still fail when both models are misspecified, and it always needs distributional assumptions on $p(x_i|y_i, z_i)$ to estimate $\psi_i(Q_i)$.

Although Robins et al. (1994) restricted π_i to be bounded away from 0, in practice you may still get some positive but near-zero values for the estimates $\hat{\pi}_i$, which can make the inverse probabilities highly variable and skewed (Kang and Schafer 2007; Robins et al. 2007). Extending Wang et al. (1997), Wang and Wang (2001) considered kernel estimation for both π_i and $\psi_i(Q_i)$, developed several kernel-assisted estimators and showed their asymptotic equivalence. However, when the dimension of the continuous part in Q_i increases, we need multivariate kernel functions and the estimation procedure suffers from the "curse of dimensionality." This motivates us to propose using a single-index model on $E(T_i|Q_i)$. Then, we can apply a univariate kernel function on the single-index $Q_i^\top\gamma$. Han and Wang (2013), Han (2014) and Han (2016) recently developed some methods to improve the robustness of AIPW estimators by allowing multiple working models for both the selection probability and the augmentation. Thus, multiple robustness can be gained, which means the estimators are consistent if any of the working models is correctly specified. In addition, these estimators are not sensitive to near-zero selection probabilities. Our method, from another perspective, is numerically stable with near-zero selection probabilities simply by not including them in the point estimation procedure. Based on Wang and Wang (2001), we develop asymptotic distribution theory for the resulting estimator and compare it with IPW and AIPW. We also conduct simulation studies to investigate the finite-sample performance of the proposed estimator in comparison with existing methods.

The rest of the paper is organized as follows: In Sect. 2, we briefly review IPW and AIPW. We then describe our new estimator in Sect. 3. In Sect. 4, we present

asymptotic theory of the above estimators of β and compare their asymptotic efficiency. In particular, we show that under certain conditions and assumptions, our proposed estimator is as efficient as the existing methods based on standard kernel smoothing, which are often practically infeasible in the case of multiple covariates. In Sect. 5, we provide the results of our simulation studies. In Sect. 6, we apply our methods to Canada 2010/2011 Youth Smoking Survey (YSS) data. We make some concluding remarks in Sect. 7. Technical proofs are given in "Appendix."

2 Brief review of existing methods

2.1 Inverse-probability weighted estimator

Robins et al. (1994) proposed a class of estimators based on weighted estimating equations. One of them is IPW through the estimating equation

$$\Delta_1(\beta, \pi) = n^{-1/2} \sum_{i=1}^{n} \frac{R_i}{\pi_i} W_i \left(y_i - W_i^\top \beta \right) = 0. \tag{1}$$

The selection probabilities π_i's are usually unknown in observational studies. One can assume a parametric model for π_i, for example, a logistic regression model under MAR

$$\begin{aligned} \pi_i(\alpha) &= \mathrm{P}\left(R_i = 1 | y_i, Z_i, \alpha\right) = \left\{1 + \exp\left(-\alpha_0 - \alpha_1 y_i - \alpha_2^\top Z_i\right)\right\}^{-1} \\ &= \left\{1 + \exp\left(-\alpha^\top Q_i\right)\right\}^{-1}, \end{aligned}$$

where α is unknown. In our estimation problem, α is a nuisance parameter and can be estimated by maximum likelihood estimator $\hat{\alpha}$. We denote the solution of $\Delta_1(\beta, \pi(\hat{\alpha})) = 0$ as $\hat{\beta}_{\mathrm{PIP}}$ in the rest of the paper.

Another approach is to estimate π_i nonparametrically. Wang et al. (1997) considered nonparametric kernel smoothers for the selection probabilities. Let d be the number of continuous components of Q, K be an rth-order kernel function, h_1 be the bandwidth parameter, and define $K_{h_1}(\cdot) = K(\cdot/h_1)$. Then the kernel estimator of $\pi(q)$ is given by

$$\hat{\pi}(q) = \frac{\sum_{i=1}^{n} R_i K_{h_1}(q - Q_i)}{\sum_{i=1}^{n} K_{h_1}(q - Q_i)}. \tag{2}$$

The resulting estimator is consistent, but is difficult to implement when Q is multi-dimensional.

A complete-case (CC) analysis is to use the observed data only treating the partial data set as a completely observed data set. This approach generally not only leads to inconsistent estimates when the missing mechanism is not MCAR, but also loses efficiency due to discarding information from incomplete cases. We will illustrate these points through simulations in Sect. 5.

2.2 Augmented inverse-probability weighted estimator

The IPW does not incorporate the incomplete cases, which generally leads to inefficient estimates. Robins et al. (1994) proposed the AIPW by solving the following equations:

$$\Delta_2(\beta, \pi, \psi) = n^{-1/2} \sum_{i=1}^{n} \left\{ \frac{R_i}{\pi_i} T_i + \left(1 - \frac{R_i}{\pi_i}\right) \psi_i \right\} = 0, \tag{3}$$

where $\psi_i = E(T_i|Q_i)$. Further, it also obtains the DR property. We can still estimate the selection probability π_i using a parametric model or a kernel smoother such as (2). To estimate ψ_i, Wang and Wang (2001) proposed to use a kernel estimator similar to that for π_i given by

$$\hat{\psi}(q) = \frac{\sum_{i=1}^{n} R_i T_i K_{h_2}(q - Q_i)}{\sum_{i=1}^{n} R_i K_{h_2}(q - Q_i)} \tag{4}$$

for another kernel bandwidth h_2. It can share the same order of kernel function and same rate of bandwidth with $\hat{\pi}(q)$ in (2). It is possible to use a parametric model on ψ or specify the conditional distribution $p(x_i|Q_i)$, but then it will fall into the same area of techniques as EM algorithm, Bayesian or MI.

2.3 Mean-score estimator

The mean-score estimator (MS) solves

$$\Delta_3(\beta, \psi) = n^{-1/2} \sum_{i=1}^{n} \{R_i T_i + (1 - R_i)\psi_i\} = 0. \tag{5}$$

Reilly and Pepe (1995) proposed this estimator with all components of Q discrete, where ψ_i is estimated by

$$\hat{\psi}_i = \frac{1}{n^{(o)}_{y_i, Z_i}} \sum_{j \in V^{(o)}_{y_i, Z_i}} T_j(X_j; \beta | y_i, Z_i)$$

with $V^{(o)}_{y_i, Z_i}$ denoting the subset of complete cases for $y = y_i$, $Z = Z_i$, $n^{(o)}_{y_i, Z_i}$ the size of $V^{(o)}_{y_i, Z_i}$, $T_j(X_j; \beta | y_i, Z_i)$ the score function for the samples in $V^{(o)}_{y_i, Z_i}$. It simply uses the averaged score of the complete cases with the same Q_i as the estimate of ψ_i. Wang and Wang (2001) extended it to the setting where some components are continuous by (4). Unlike IPW or AIPW, MS does not need to estimate the selection probabilities π_i.

3 Proposed methodology

3.1 The issues of small selection probabilities and curse of dimensionality

Although theoretically the IPW and AIPW estimators are unbiased estimators when either the model for selection probabilities (π) or for augmentation (ψ) is correctly specified, they may encounter numerical problems if some π_i's are small so that the inverse-probability weights are highly variable. In this case, some subjects may have very large weights to significantly influence the weighted averages, and the sampling distribution of a locally semiparametric efficient estimator (IPW, AIPW) can be markedly skewed and highly variable, leading to biased estimation. We will illustrate this point through simulations in Sect. 5. This phenomenon is observed and discussed by Kang and Schafer (2007) and Robins et al. (2007), existing at least when parametrically modeling π_i. In this sense, the mean-score estimator has its advantage as it does not need to model and use inverse-probability weights.

However, all the kernel-estimation-based estimators mentioned above, including the MS estimator, have the same problem: curse of dimensionality. If the dimension d of the continuous part in Q is more than one, the performance of kernel functions can be unsatisfying.

3.2 Single-index model and the proposed estimator

To overcome the problem discussed above, we consider a single-index model on ψ. Notice that

$$\begin{aligned}\psi_i &= E(T_i|Q_i) = E\left\{W_i\left(y_i - W_i^\top\beta\right)\Big|Q_i\right\} = E\left(W_i|Q_i\right)y_i - E\left(W_iW_i^\top|Q_i\right)\beta \\ &= \begin{pmatrix} 1 \\ E\left(X_i|Q_i\right) \\ Z_i \end{pmatrix} y_i - \begin{pmatrix} 1 & E\left(X_i|Q_i\right) & Z_i^\top \\ E\left(X_i|Q_i\right) & E\left(X_i^2|Q_i\right) & E\left(X_i|Q_i\right)Z_i^\top \\ Z_i & Z_iE\left(X_i|Q_i\right) & Z_iZ_i^\top \end{pmatrix}\beta. \end{aligned} \tag{6}$$

Thus we only need to model $E(X_i|Q_i)$ and $E(X_i^2|Q_i)$. Assume a single-index model (SIM)

$$X_i = g\left(Q_i^\top\gamma\right) + e_i, \tag{7}$$

where g is an unknown smooth univariate function, γ is the parameter of the model with the same dimension of Q_i, e_i's are random errors with zero mean. To guarantee identifiability, we assume the first nonzero element of γ is positive 1. If the number of complete cases is n_1, one estimator of $g(\cdot)$ based only on the complete cases is

$$\hat{g}(u|\gamma) = \frac{\sum_{j=1}^{n_1} X_j^{(o)} K_h\left(u - Q_j^{(o)\top}\gamma\right)}{\sum_{j=1}^{n_1} K_h\left(u - Q_j^{(o)\top}\gamma\right)} = \frac{\sum_{k=1}^{n} R_kX_kK_h\left(u - Q_k^\top\gamma\right)}{\sum_{k=1}^{n} R_kK_h\left(u - Q_k^\top\gamma\right)},$$

where $(X_j^{(o)}, Q_j^{(o)})$ are pairs of the complete cases. Then under the SIM condition, we have

$$\hat{E}\left(X_i|Q_i\right) = \hat{E}\left(X_i|Q_i^\top\gamma\right) = \hat{g}\left(Q_i^\top\gamma\right) = \frac{\sum_{k=1}^n R_k X_k K_h\left((Q_i - Q_k)^\top\gamma\right)}{\sum_{k=1}^n R_k K_h\left((Q_i - Q_k)^\top\gamma\right)}. \quad (8)$$

We can also apply this model to get an estimate of $E(X_i^2|Q_i)$ as

$$\hat{E}\left(X_i^2|Q_i\right) = \frac{\sum_{k=1}^n R_k X_k^2 K_h\left((Q_i - Q_k)^\top\gamma\right)}{\sum_{k=1}^n R_k K_h\left((Q_i - Q_k)^\top\gamma\right)}. \quad (9)$$

We can construct

$$\hat{\pi}_i^*(\gamma) = \hat{E}\left(R_i|Q_i^\top\gamma\right) = \frac{\sum_{k=1}^n R_k K_h\left((Q_i - Q_k)^\top\gamma\right)}{\sum_{k=1}^n K_h\left((Q_i - Q_k)^\top\gamma\right)} \quad (10)$$

as the estimated selection probabilities modeled by the SIM using the same (γ, h). Notice that (8) and (10) have the same forms as (4) and (2) (NW-estimators). But the former two are conditional on the single-index and thus can just use univariate kernel functions with the additional parameter γ. Due to this similarity, we can extend the asymptotic results by Wang and Wang (2001) to the single-index models. The details will be shown in Sect. 4. On the other hand, compared to Wang and Wang (2001), here we only estimate the first two moments of X_i given Q_i by using the local average when estimating ψ_i but keep the original y_i, Z_i since they are always observed, instead of using the local average of the whole score function like (4).

Let

$$\hat{\psi}_i(\gamma) = \frac{\sum_{k=1}^n R_k T_{i,k} K_h\left((Q_i - Q_k)^\top\gamma\right)}{\sum_{k=1}^n R_k K_h\left((Q_i - Q_k)^\top\gamma\right)},$$

where $T_{i,k} = W_{i,k}(y_i - W_{i,k}^\top\beta)$ with $W_{i,k} = (1, X_k, Z_i^\top)^\top$. This $\hat{\psi}_i(\gamma)$ is a kernel estimate of ψ_i by estimating only $E(X_i|Q_i)$ and $E(X_i^2|Q_i)$ with a kernel smoother via (6). Then, the AIPW (3) and MS (5) estimators in the previous section can be extended using the SIM as follows:

(a) AIPW with a parametric model on selection probabilities:

$$\Delta_2\left(\beta, \pi(\hat{\alpha}), \hat{\psi}(\gamma)\right) = n^{-1/2}\sum_{i=1}^n\left\{\frac{R_i}{\pi_i\left(\hat{\alpha}\right)}T_i + \left(1 - \frac{R_i}{\pi_i(\hat{\alpha})}\right)\hat{\psi}_i(\gamma)\right\} = 0; \quad (11)$$

(b) MS without inverse-probability weights:

$$\Delta_3(\beta, \hat{\psi}(\gamma)) = n^{-1/2}\sum_{i=1}^n\left\{R_i T_i + (1 - R_i)\hat{\psi}_i(\gamma)\right\} = 0. \quad (12)$$

We use $\hat{\beta}_{\mathrm{PIP}A}$ and $\hat{\beta}_A$ to denote the solutions of Eqs. (11) and (12), respectively.

Generally, besides the main parameter β, γ is an unknown nuisance parameter which also needs to be estimated. However, in our case with the linear relationship between Y and $(X, Z^\top)^\top$, we have a special form of γ as $\gamma = (1, -\beta_Z^\top)^\top$, where β_Z is the regression coefficient of Z as $E(Y|X, Z) = \beta_0 + \beta_1 X + \beta_Z^\top Z$. In that sense, the single index is $u_i = Q_i^\top \gamma = y_i - \beta_Z^\top Z_i$ and γ is a part of β so that we do not need to estimate γ separately. Note also that the choice of bandwidth h is crucial. Technical details about bandwidth selection will be discussed in Sect. 4.

Under certain conditions, we can show that these two estimators are asymptotically equivalent (see Corollary 1 below), and they are both as efficient as the existing estimators using standard kernel smoothing, which are often practically infeasible in the case of multi-covariates. In practice, we prefer $\hat{\beta}_A$ because the estimation procedure of $\hat{\beta}_A$ has the clear advantage of not involving inverse of selection probabilities to avoid modeling π_i's; thus, it is simpler. Moreover, it is important to note that unlike all inverse probability weighted estimators, $\hat{\beta}_A$ is not sensitive to the positive near-zero π_i's since we do not use them in the point estimation procedure.

Of course, $\hat{\beta}_A$ no longer has the property of double robustness and thus needs a consistent estimator of ψ. In this setting, the performance of $\hat{\beta}_A$ depends on whether the single-index model (7) is reasonable. Since the relationship between the response and covariates is assumed to be linear, it is not unreasonable to assume this model. Actually, it is valid when (y_i, X_i, Z_i) jointly follows a multivariate normal distribution. More generally, it can still give reasonably robust results under other distributions, as is to be shown in our numerical studies in Sect. 5.

4 Asymptotic properties

In this section, we will show the asymptotic behavior of the proposed estimator $\hat{\beta}_A$, and its asymptotic equivalence to some other estimators described above under certain conditions. For simplicity, we define $\pi_i^*(\gamma) = E(R_i|Q_i^\top \gamma)$ as the selection probabilities conditional on the single-index $Q_i^\top \gamma$ with parameter γ, $\pi_i(\alpha)$ as the selection probabilities based on a parametric model with parameter α. We need some regularity conditions to establish the asymptotic theory, which can be found in "Appendix A.1." Recall that r is the order of the kernel function used in the estimation. From regularity condition (i), r is related to the rate of the bandwidth h. Since we are considering a SIM for estimation, a standard 2nd-order ($r = 2$) univariate kernel function seems reasonable in practice.

Let $\eta_n = \{nh^{2r} + (nh^2)^{-1}\}^{1/2}$. The following lemmas are important to prove our main theorems.

Lemma 1 *Under regularity conditions* (i)–(vii) *given in* "Appendix A.1" *and assuming that the single-index model* (7) *is true, we have*

$$n^{-1/2}\sum_{i=1}^{n}(1-R_i)\left\{\hat{\psi}_i(\gamma)-\psi_i(\gamma)\right\}$$
$$=n^{-1/2}\sum_{i=1}^{n}R_i\left\{T_i^0-\psi_i^0(\gamma)\right\}a\left(Q_i^\top\gamma\right)+O_p(\eta_n),$$

where $a(Q_i^\top\gamma)=\{1-\pi_i^*(\gamma)\}/\pi_i^*(\gamma)$, $T_i^0=E_{Z_i|u_i,R_i=0}(T_i)=\int T_i f(Z_i|u_i,R_i=0)dZ_i$, $\psi_i^0(\gamma)=E_{Z_i|u_i,R_i=0}\{\psi_i(\gamma)\}=\int\psi_i(\gamma)f(Z_i|u_i,R_i=0)dZ_i$ *with* $u_i=Q_i^\top\gamma$ *as the single index.*

This is an extension of Lemma 1 in Wang and Wang (2001). The proof of this lemma is given in "Appendix A.2."

Note that with the SIM and the single index $u_i=y_i-\beta_Z^\top Z_i$, we can write T_i and $\psi_i(\gamma)$ as

$$T_i=\begin{pmatrix}u_i-\beta_0-\beta_1X_i\\ u_iX_i-\beta_0X_i-\beta_1X_i^2\\ Z_i\,(u_i-\beta_0-\beta_1X_i)\end{pmatrix},$$
$$\psi_i(\gamma)=\begin{pmatrix}u_i-\beta_0-\beta_1E\,(X_i|u_i)\\ u_iE\,(X_i|u_i)-\beta_0E\,(X_i|u_i)-\beta_1E\left(X_i^2|u_i\right)\\ Z_i\,\{u_i-\beta_0-\beta_1E\,(X_i|u_i)\}\end{pmatrix}.$$

Since MAR implies $(X_i\perp R_i)|Q_i$, we also have

$$T_i^0=\begin{pmatrix}u_i-\beta_0-\beta_1X_i\\ u_iX_i-\beta_0X_i-\beta_1X_i^2\\ Z_i^{u|0}\,(u_i-\beta_0-\beta_1X_i)\end{pmatrix},$$
$$\psi_i^0(\gamma)=\begin{pmatrix}u_i-\beta_0-\beta_1E(X_i|u_i)\\ u_iE(X_i|u_i)-\beta_0E(X_i|u_i)-\beta_1E(X_i^2|u_i)\\ Z_i^{u|0}\{u_i-\beta_0-\beta_1E(X_i|u_i)\}\end{pmatrix}$$

with $Z_i^{u|0}=E(Z_i|u_i,R_i=0)$.

Lemma 1 is useful because it converts asymptotically a sum of dependent random variables to a sum of independent and identically distributed (i.i.d.) random variables. Then it is easier to be dealt with by applying standard asymptotic theory.

Lemma 2 *Under the same conditions as those in Lemma 1, we have*

(a) $n^{-1/2}\sum_{i=1}^{n}R_i\left\{\hat{\psi}_i(\gamma)-\psi_i(\gamma)\right\}=n^{-1/2}\sum_{i=1}^{n}R_i\left\{T_i^1-\psi_i^1(\gamma)\right\}+O_p(\eta_n)$;

(b) $n^{-1/2}\sum_{i=1}^{n}\frac{R_i}{\hat{\pi}_i^*(\gamma)}\left\{\hat{\psi}_i(\gamma)-\psi_i(\gamma)\right\}=n^{-1/2}\sum_{i=1}^{n}\frac{R_i}{\pi_i^*(\gamma)}\left\{T_i^1-\psi_i^1(\gamma)\right\}+O_p(\eta_n)$;

(c) *In addition, if the parametric model for selection probabilities is correctly specified and has a single-index model form with the same single index* $u_i = Q_i^\top \gamma$ *as the augmentation, which means* $\pi_i(\alpha) = \pi_i = \pi_i^*(\gamma)$*, then*

$$n^{-1/2} \sum_{i=1}^{n} \frac{R_i}{\pi_i(\hat{\alpha})} \left\{ \hat{\psi}_i(\gamma) - \psi_i(\gamma) \right\} = n^{-1/2} \sum_{i=1}^{n} \frac{R_i}{\pi_i} \left\{ T_i^1 - \psi_i^1(\gamma) \right\} + O_p(\eta_n),$$

where $T_i^1 = E_{Z_i|u_i, R_i=1}(T_i)$, $\psi_i^1(\gamma) = E_{Z_i|u_i, R_i=1}\{\psi_i(\gamma)\}$.

The proof of Lemma 2 is given in "Appendix A.3." The idea of the proof is analogous to that of Lemma 1.

Define

$$U_i = R_i T_i + (1 - R_i)\psi_i(\gamma) + R_i\{T_i^0 - \psi_i^0(\gamma)\}a(Q_i^\top \gamma).$$

Based on Lemmas 1 and 2, we have the following main theorems.

Theorem 1 *Under the regularity conditions given in* "Appendix A.1" *and assuming that the single-index model* (7) *is true,* $\hat{\beta}_A$ *is asymptotically equivalent to the solution of the following estimating equation:*

$$n^{-1/2} \sum_{i=1}^{n} U_i = 0.$$

Furthermore, we have

$$n^{1/2} \left(\hat{\beta}_A - \beta \right) \xrightarrow{\mathcal{D}} N_p \left(0, \boldsymbol{\Sigma}_A \right),$$

where $\boldsymbol{\Sigma}_A = \boldsymbol{D}^{-1} \boldsymbol{\mathcal{M}} \boldsymbol{D}^{-1}$ *with* $\boldsymbol{D} = -n^{-1} E(\partial T_1 / \partial \beta^\top) = E(W_1 W_1^\top)$ *and* $\boldsymbol{\mathcal{M}} = \mathrm{cov}(U_1) = \boldsymbol{\mathcal{A}} + \boldsymbol{\mathcal{B}} + 2\boldsymbol{\mathcal{C}}$ *for*

$$\begin{aligned}
\boldsymbol{\mathcal{A}} &= E\left(\pi_1 T_1 T_1^\top \right) + E\left\{ (1 - \pi_1)\psi_1 \psi_1^\top \right\}, \\
\boldsymbol{\mathcal{B}} &= E\left\{ \pi_1 (T_1^0 - \psi_1^0)(T_1^0 - \psi_1^0)^\top a^2(Q_1^\top \gamma) \right\}, \\
\boldsymbol{\mathcal{C}} &= E\left\{ \pi_1 T_1 \left(T_1^0 - \psi_1^0 \right)^\top a(Q_1^\top \gamma) \right\}.
\end{aligned}$$

The main step of the proof is to obtain the asymptotic equivalence between our estimating equation $\Delta_3(\beta, \hat{\psi}(\gamma)) = 0$, and $n^{-1/2} \sum_{i=1}^{n} U_i = 0$ with true $\pi^*(\gamma)$ and $\psi(\gamma)$. The details can be found in "Appendix A.4."

The asymptotic covariance matrix Σ_A of $\hat{\beta}_A$ can be estimated by first estimating $\boldsymbol{\mathcal{A}}$, $\boldsymbol{\mathcal{B}}$ and $\boldsymbol{\mathcal{C}}$ separately. However, this approach cannot guarantee the necessary property of nonnegative definiteness of the resulting covariance estimate, and it might lead to

numerically unstable results. For this reason, we propose to estimate Σ_A directly as follows:

$$\hat{\boldsymbol{\Sigma}}_A = \hat{\boldsymbol{D}}_n^{-1}(\hat{\gamma}) \left(\frac{1}{n} \sum_{i=1}^{n} \hat{U}_i \hat{U}_i^\top \right) \hat{\boldsymbol{D}}_n^{-1}(\hat{\gamma}), \tag{13}$$

where

$$\hat{U}_i = R_i T_i \left(\hat{\beta}_A \right) + (1 - R_i)\hat{\psi}_i + R_i \left\{ \hat{T}_i^0 \left(\hat{\beta}_A \right) - \hat{\psi}_i^0 \right\} a \left(Q_i^\top \hat{\gamma} \right)$$

with $\hat{\psi}_i = \hat{\psi}_i(\hat{\beta}_A, \hat{\gamma})$, $\hat{\psi}_i^0 = \hat{\psi}_i^0(\hat{\beta}_A, \hat{\gamma})$ being the estimates of ψ_i, ψ_i^0 based on $\hat{\beta}_A$ and

$$\hat{\boldsymbol{D}}_n \left(\hat{\gamma} \right) = n^{-1} \sum_{i=1}^{n} \left\{ R_i W_i W_i^\top + (1 - R_i) \hat{E} \left(W_i W_i^\top | Q_i^\top \hat{\gamma} \right) \right\}.$$

Here $\hat{T}_i^0$ is T_i^0 with $Z_i^{u|0}$ estimated by

$$\hat{Z}_i^{u|0} = \hat{E}(Z_i | \hat{u}_i, R_i = 0) = \frac{\sum_{k=1}^{n} (1 - R_k) Z_k K_h((Q_i - Q_k)^\top \hat{\gamma})}{\sum_{k=1}^{n} (1 - R_k) K_h((Q_i - Q_k)^\top \hat{\gamma})}$$

with $\hat{u}_i = Q_i^\top \hat{\gamma}$.

Note that to get $\hat{E}(W_i W_i^\top | Q_i^\top \hat{\gamma})$, we only need to calculate $\hat{E}(X_i | Q_i)$ and $\hat{E}(X_i^2 | Q_i)$ through (8) and (9) because of the structure in (6).

Theorem 2 *Under the same conditions as in Theorem 1 and the additional conditions for Lemma 2*(c)*, we have*

$$n^{1/2} \left(\hat{\beta}_{\mathrm{PIP}A} - \beta \right) \xrightarrow{\mathcal{D}} N_p(0, \boldsymbol{\Sigma}_{PA}),$$

where $\boldsymbol{\Sigma}_{PA} = \boldsymbol{D}^{-1}(\boldsymbol{S} - \boldsymbol{S^*} + \boldsymbol{V})\boldsymbol{D}^{-1}$, *with* $\boldsymbol{D} = E(W_1 W_1^\top)$, $\boldsymbol{S} = E\{T_1 T_1^\top / \pi_1^*(\gamma)\}$, $\boldsymbol{S^*} = E\{\psi_1 \psi_1^\top / \pi_1^*(\gamma)\}$ *and* $\boldsymbol{V} = E(\psi_1 \psi_1^\top)$.

It is readily seen that a consistent covariance matrix estimate of $\hat{\beta}_{\mathrm{PIP}A}$ is given by

$$\hat{\boldsymbol{\Sigma}}_{PA} = \hat{\boldsymbol{D}}_n^{-1}(\hat{\gamma}) \left[\frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{R_i}{\hat{\pi}_i^{*2}(\hat{\gamma})} \left(\hat{T}_i - \hat{\psi}_i \right) \left(\hat{T}_i - \hat{\psi}_i \right)^\top + \frac{R_i}{\hat{\pi}_i^*(\hat{\gamma})} \hat{\psi}_i \hat{\psi}_i^\top \right\} \right] \hat{\boldsymbol{D}}_n^{-1}(\hat{\gamma}) \tag{14}$$

with $\hat{T}_i = \hat{T}_i(\hat{\beta}_{\mathrm{PIP}A}, \hat{\gamma})$, $\hat{\psi}_i = \hat{\psi}_i(\hat{\beta}_{\mathrm{PIP}A}, \hat{\gamma})$ being estimates of T_i, ψ_i based on $\hat{\beta}_{\mathrm{PIP}A}$. Since $Pr(R_i = 1 | u_i, Z_i) = Pr(R_i = 1 | Q_i) = \pi_i = \pi_i^*(\gamma) = Pr(R_i = 1 | u_i)$, the additional condition for Lemma 2(c) implies that $(Z_i \perp R_i) | u_i$. Then $T_i^0 = T_i^1 = E_{Z_i | u_i}(T_i)$, $\psi_i^0(\gamma) = \psi_i^1(\gamma) = E_{Z_i | u_i}\{\psi_i(\gamma)\}$.

Although the relationship between $\boldsymbol{\Sigma}_A$ and $\boldsymbol{\Sigma}_{PA}$ is generally not clear even under the conditions of Theorem 2, numerically the SE of $\hat{\beta}_A$ is competitive (see Sect. 5) and $\hat{\beta}_A$ does not have the potential danger of having exceedingly high inverse-probability weights. The theorems also demonstrate the asymptotic normality of the above estimators, which helps us to make inferences with the estimators.

Corollary 1 *Under the same conditions as in Theorem 2 and further assuming* $E(Z_i|u_i) = Z_i$*, we have*

(a) $\hat{\beta}_A$ *and* $\hat{\beta}_{\text{PIP}A}$ *are asymptotically equivalent and are both more efficient than* $\hat{\beta}_{\text{PIP}}$*;*
(b) *The estimators* $\hat{\beta}_A$ *and* $\hat{\beta}_{\text{PIP}A}$ *based on a single-index model are as efficient as those based on a standard multivariate kernel smoother.*

It is intuitive to see that these estimators are asymptotically more efficient than $\hat{\beta}_{\text{PIP}}$ because they incorporate the incomplete cases. However, when the conditions are satisfied, it is surprising to see that the estimators based on the SIM can keep the efficiency of the standard kernel smoothers [such as (2), (4) proposed by Wang and Wang (2001)] with a lower dimension of information. Since IPW is asymptotically equivalent to AIPW and MS estimators with both selection probabilities and augmentation estimated by a standard kernel smoother (Wang and Wang 2001), the proof of the corollary also shows that IPW using a standard kernel smoother is more efficient than $\hat{\beta}_{\text{PIP}}$, which was not discussed by Wang et al. (1997).

We define $\boldsymbol{\Sigma}_P$ as the asymptotic covariance matrix of $\hat{\beta}_{\text{PIP}}$ and $\tilde{\boldsymbol{\Sigma}}$ for $\hat{\beta}_A$, $\hat{\beta}_{\text{PIP}A}$ based on a standard kernel smoother. For two positive semi-definite covariance matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, we define $\boldsymbol{A} \succeq \boldsymbol{B}$ if $\boldsymbol{A} - \boldsymbol{B}$ is positive semi-definite. From the proof in "Appendix A.6," we see that $\boldsymbol{\Sigma}_P \succeq \tilde{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}_A = \boldsymbol{\Sigma}_{PA}$ under the conditions in Corollary 1.

The performance of the estimator $\hat{\beta}_A$ depends on the choice of the bandwidth h used in the kernel function $K_h(\cdot)$. In the regularity conditions in "Appendix A.1," we require $nh^2 \to \infty$ and $nh^{2r} \to 0$, as $n \to \infty$. Therefore, the classical optimal rate of the bandwidth $O(n^{-1/5})$ does not work in our situation, as indicated in Sepanski et al. (1994). A reasonable choice is $h = Cn^{-1/3}$, where C is a constant. A plug-in method can be applied to estimate C. For simplicity, we can use $C = \hat{\sigma}_u$ as suggested by Wang et al. (1997) and Zhou et al. (2008), where $\hat{\sigma}_u$ is the sample standard deviation of the single index u_i. We use this formula to choose the bandwidth in our following numerical studies.

5 Simulations

In this section, we investigate the performance of the proposed estimator $\hat{\beta}_A$ compared to other estimators, in terms of bias and standard error. We also examine the covariance estimation using the sandwich formula (13), by comparing the asymptotic standard error with the empirical standard error. The empirical standard error is obtained from 1000 estimates through independent Monte Carlo simulations under the same data-generating conditions. The asymptotic normality of the estimators is examined by calculating the 95% coverage probabilities. We also use this numerical study as an

example to illustrate the phenomenon of highly variable inverse probabilities, as well as the robustness of our estimator under non-normal distributions.

There are two main scenarios in our simulations. For both of them, we consider $n = 250$ and 500. In the first scenario, we have (y_i, X_i, Z_i) generated from a multivariate normal distribution with $X_i \sim N(0, 1)$, $Z_i \sim N_3(0, \boldsymbol{I}_3)$, $\varepsilon_i \sim N(0, 1)$, $i = 1, 2, \ldots, n$. Thus we have $p = 4$. The true regression coefficients $\beta = (0, 0.5, 1, -1, -0.5)^\top$, and $y_i = W_i\beta + \varepsilon_i$ with $W_i = (1, X_i, Z_i^\top)^\top$. The selection probabilities for observing X_i are $\pi_i = \{1 + \exp(-\alpha_0 - \alpha_1 y_i - \alpha_2 Z_{i1} - \alpha_3 Z_{i2} - \alpha_4 Z_{i3})\}^{-1}$, which satisfy MAR on X. In this setting, the single-index model on the augmentation is easily seen to be valid. We have three different choices for the values of α. On average, there are about 20%, 40% and 60% of the cases that have X missing. We choose to use a second-order Gaussian kernel function ($r = 2$). The bandwidth selection has been discussed in the previous section. In practice, since X_i^2 is more variable than X_i, we use $h = 0.4\hat{\sigma}_u n^{-1/3}$ when estimating $E(X_i^2|Q_i)$. The coefficient parameters are estimated through the estimating equations (12) by iterations in R.

We use a logistic regression model to model the missing process parametrically for $\hat{\beta}_{\mathrm{PIP}}$ and $\hat{\beta}_{\mathrm{PIP}A}$. In this setting, this model is correctly specified so that theoretically they are unbiased estimators. However, the (estimated) selection probabilities can be positive but near zero, which may lead to numerically biased estimates as we indicated earlier. Our empirical experience suggests that, since we only use the information in incomplete cases when estimating $\hat{E}(Z_i|u_i, R_i = 0)$, it would be helpful to include a correction factor matrix in the sandwich formula (13) for small-to-moderate sample sizes, such as those in our simulation studies, especially when the percentage of missingness is high and the data are believed to be skewed. For example, we may replace the estimated asymptotic covariance by $\hat{\boldsymbol{\Sigma}}_A^* = \boldsymbol{F}_c \cdot \hat{\boldsymbol{\Sigma}}_A$, where $\boldsymbol{F}_c = \mathrm{diag}\{a, \ldots, a, b, a, \ldots, a\}^{-1}$, $a = 1 - 0.3 \times \mathrm{miss}\%$, $b = (1 - 0.7 \times \mathrm{miss}\%) \cdot \min(\exp\{(n - 500)/5000\}, 1)$, and miss% means the percentage of missingness of X in the data set. The position of b matches the position of the coefficient of the missing covariate. This is what we used for $\hat{\beta}_A$ in our numerical results. For simulation purposes, we also show the results of the full data $\hat{\beta}_F$ as a benchmark for comparison.

The results for the first scenario are displayed in Tables 1, 2 and 3. For each estimator, the first line displays the bias, the second line is the empirical standard error/averaged asymptotic standard error by formula over 1000 replications and the third line is the 95% coverage probability. Since the conditions for Theorem 2 are not satisfied in the simulation studies, we do not have a closed form for $\hat{\Sigma}_{PA}$. Thus, we put an "$*$" in the places of averaged asymptotic SE of $\hat{\beta}_{\mathrm{PIP}A}$ and use the 1% trimmed empirical SE to calculate the 95% coverage probabilities. The reason to use the trimmed SE is that we have some extremely "bad" results caused by the near-zero selection probabilities, and these few extreme values make the empirical SE too large compared to other estimators. As expected, the CC analysis produces biased estimates in this scenario. We also observe that $\hat{\beta}_{\mathrm{PIP}}$ always has significant bias for each parameter, and $\hat{\beta}_{\mathrm{PIP}A}$ has bias at least for β_1, the coefficient of X, even with $n = 500$. Moreover, the above two estimators have much larger standard errors than $\hat{\beta}_A$. Given the multivariate normal data and correctly specified logistic model for the selection probabilities in Tables 1,

Table 1 Simulation results of 1000 replications for the normal data, $X_i \sim N(0, 1)$, $Z_i \sim N_3(0, \boldsymbol{I}_3)$, $\varepsilon_i \sim N(0, 1)$, with $\alpha = (2.2, -0.9, -0.7, 0, 0)$, about 20% missing at random on average

	$\hat{\beta}_0 - \beta_0$	$\hat{\beta}_1 - \beta_1$	$\hat{\beta}_2 - \beta_2$	$\hat{\beta}_3 - \beta_3$	$\hat{\beta}_4 - \beta_4$
$n = 250$					
$\hat{\beta}_F$	−0.0024	−0.0014	−0.0001	0.0006	0.0011
	0.0020/0.0020	0.0021/0.0020	0.0020/0.0020	0.0020/0.0020	0.0020/0.0020
	0.943	0.936	0.953	0.944	0.943
$\hat{\beta}_{\mathrm{CC}}$	−0.1527	−0.0307	−0.1051	0.0575	0.0309
	0.0023/0.0022	0.0023/0.0022	0.0024/0.0024	0.0023/0.0023	0.0022/0.0022
	0.436	0.917	0.712	0.870	0.924
$\hat{\beta}_{\mathrm{PIP}}$	−0.0120	−0.0026	−0.0150	0.0072	0.0068
	0.0026/0.0022	0.0029/0.0024	0.0033/0.0025	0.0029/0.0024	0.0028/0.0024
	0.909	0.904	0.854	0.905	0.910
$\hat{\beta}_{\mathrm{PIP}A}$	−0.0048	0.0100	0.0002	0.0074	0.0019
	0.0021/∗	0.0025/∗	0.0022/∗	0.0021/∗	0.0021/∗
	0.926	0.928	0.927	0.931	0.934
$\hat{\beta}_A$	−0.0009	−0.0059	0.0010	−0.0002	0.0004
	0.0021/0.0021	0.0023/0.0024	0.0021/0.0020	0.0021/0.0020	0.0021/0.0020
	0.949	0.956	0.934	0.944	0.940
$n = 500$					
$\hat{\beta}_F$	0.0004	−0.0014	0.0004	−0.0016	0.0024
	0.0015/0.0014	0.0015/0.0014	0.0013/0.0014	0.0014/0.0014	0.0015/0.0014
	0.941	0.938	0.960	0.948	0.941
$\hat{\beta}_{\mathrm{CC}}$	−0.1512	−0.0299	−0.1045	0.0573	0.0317
	0.0017/0.0016	0.0016/0.0015	0.0016/0.0017	0.0016/0.0016	0.0016/0.0016
	0.148	0.893	0.501	0.779	0.891
$\hat{\beta}_{\mathrm{PIP}}$	−0.0053	−0.0031	−0.0084	0.0036	0.0044
	0.0019/0.0016	0.0021/0.0018	0.0024/0.0020	0.0022/0.0019	0.0021/0.0018
	0.903	0.911	0.891	0.904	0.919
$\hat{\beta}_{\mathrm{PIP}A}$	0.0028	0.0294	0.0090	−0.0154	0.0007
	0.0015/∗	0.0018/∗	0.0016/∗	0.0016/∗	0.0015/∗
	0.927	0.934	0.927	0.927	0.928
$\hat{\beta}_A$	0.0009	−0.0037	0.0005	−0.0018	0.0018
	0.0015/0.0015	0.0017/0.0017	0.0014/0.0014	0.0015/0.0014	0.0015/0.0014
	0.945	0.953	0.946	0.941	0.940

For each entry, the first line displays the bias, the second line is the empirical standard error/averaged asymptotic standard error and the third line is the 95% coverage probability. An "∗" indicates the asymptotic standard error formula unavailable

Table 2 Simulation results of 1000 replications for the normal data, $X_i \sim N(0,1)$, $Z_i \sim N_3(0, \boldsymbol{I}_3)$, $\varepsilon_i \sim N(0,1)$, with $\alpha = (0.5, -1, -0.5, 0, 0)$, about 40% missing at random on average

	$\hat{\beta}_0 - \beta_0$	$\hat{\beta}_1 - \beta_1$	$\hat{\beta}_2 - \beta_2$	$\hat{\beta}_3 - \beta_3$	$\hat{\beta}_4 - \beta_4$
$n = 250$					
$\hat{\beta}_F$	−0.0024	−0.0014	−0.0001	0.0006	0.0011
	0.0020/0.0020	0.0021/0.0020	0.0020/0.0020	0.0020/0.0020	0.0020/0.0020
	0.943	0.936	0.953	0.944	0.943
$\hat{\beta}_{\mathrm{CC}}$	−0.3499	−0.0578	−0.1642	0.1086	0.0536
	0.0028/0.0028	0.0025/0.0025	0.0029/0.0028	0.0027/0.0026	0.0026/0.0025
	0.027	0.884	0.543	0.727	0.880
$\hat{\beta}_{\mathrm{PIP}}$	−0.0462	−0.0166	−0.0432	0.0258	0.0151
	0.0035/0.0028	0.0038/0.0030	0.0044/0.0032	0.0043/0.0031	0.0038/0.0030
	0.814	0.869	0.793	0.822	0.861
$\hat{\beta}_{\mathrm{PIP}A}$	−0.0037	0.0197	−0.0075	0.0127	0.0018
	0.0026/∗	0.0040/∗	0.0028/∗	0.0026/∗	0.0025/∗
	0.922	0.918	0.926	0.928	0.934
$\hat{\beta}_A$	−0.0002	−0.0127	0.0005	−0.0006	−0.0003
	0.0022/0.0023	0.0028/0.0029	0.0021/0.0020	0.0021/0.0021	0.0021/0.0021
	0.956	0.959	0.935	0.938	0.941
$n = 500$					
$\hat{\beta}_F$	0.0004	−0.0014	0.0004	−0.0016	0.0024
	0.0015/0.0014	0.0015/0.0014	0.0013/0.0014	0.0014/0.0014	0.0015/0.0014
	0.941	0.938	0.960	0.948	0.941
$\hat{\beta}_{\mathrm{CC}}$	−0.3458	−0.0548	−0.1631	0.1080	0.0558
	0.0020/0.0020	0.0018/0.0018	0.0020/0.0020	0.0019/0.0019	0.0018/0.0018
	0.000	0.824	0.257	0.554	0.831
$\hat{\beta}_{\mathrm{PIP}}$	−0.0278	−0.0111	−0.0292	0.0203	0.0148
	0.0027/0.0022	0.0029/0.0024	0.0035/0.0026	0.0033/0.0025	0.0029/0.0024
	0.852	0.907	0.819	0.844	0.880
$\hat{\beta}_{\mathrm{PIP}A}$	−0.0088	−0.0210	−0.0023	0.0216	0.0121
	0.0018/∗	0.0027/∗	0.0019/∗	0.0018/∗	0.0018/∗
	0.927	0.925	0.928	0.931	0.934
$\hat{\beta}_A$	0.0020	−0.0065	0.0005	−0.0024	0.0012
	0.0016/0.0016	0.0020/0.0021	0.0015/0.0014	0.0015/0.0015	0.0015/0.0015
	0.954	0.950	0.947	0.936	0.935

For each entry, the first line displays the bias, the second line is the empirical standard error/averaged asymptotic standard error and the third line is the 95% coverage probability. An "∗" indicates the asymptotic standard error formula unavailable

Table 3 Simulation results of 1000 replications for the normal data, $X_i \sim N(0, 1)$, $Z_i \sim N_3(0, \boldsymbol{I}_3)$, $\varepsilon_i \sim N(0, 1)$, with $\alpha = (-0.5, -0.5, -0.5, 0, 0)$, about 60% missing at random on average

	$\hat{\beta}_0 - \beta_0$	$\hat{\beta}_1 - \beta_1$	$\hat{\beta}_2 - \beta_2$	$\hat{\beta}_3 - \beta_3$	$\hat{\beta}_4 - \beta_4$
$n = 250$					
$\hat{\beta}_F$	−0.0024	−0.0014	−0.0001	0.0006	0.0011
	0.0020/0.0020	0.0021/0.0020	0.0020/0.0020	0.0020/0.0020	0.0020/0.0020
	0.943	0.936	0.953	0.944	0.943
$\hat{\beta}_{\mathrm{CC}}$	−0.2830	−0.0227	−0.0902	0.0447	0.0243
	0.0036/0.0036	0.0032/0.0030	0.0035/0.0033	0.0032/0.0031	0.0033/0.0031
	0.288	0.929	0.832	0.917	0.915
$\hat{\beta}_{\mathrm{PIP}}$	−0.0269	−0.0037	−0.0174	0.0098	0.0085
	0.0037/0.0030	0.0042/0.0035	0.0049/0.0036	0.0046/0.0036	0.0042/0.0035
	0.861	0.872	0.827	0.862	0.897
$\hat{\beta}_{\mathrm{PIP}A}$	−0.0063	0.0290	−0.0041	0.0021	< 0.0001
	0.0026/∗	0.0040/∗	0.0028/∗	0.0028/∗	0.0026/∗
	0.933	0.922	0.930	0.929	0.922
$\hat{\beta}_A$	−0.0023	−0.0083	0.0003	−0.0006	−0.0003
	0.0023/0.0025	0.0032/0.0035	0.0021/0.0021	0.0022/0.0021	0.0022/0.0021
	0.964	0.968	0.936	0.937	0.934
$n = 500$					
$\hat{\beta}_F$	0.0004	−0.0014	0.0004	−0.0016	0.0024
	0.0015/0.0014	0.0015/0.0014	0.0013/0.0014	0.0014/0.0014	0.0015/0.0014
	0.941	0.938	0.960	0.948	0.941
$\hat{\beta}_{\mathrm{CC}}$	−0.2805	−0.0226	−0.0907	0.0428	0.0261
	0.0026/0.0025	0.0022/0.0022	0.0024/0.0024	0.0023/0.0022	0.0023/0.0022
	0.062	0.923	0.774	0.888	0.915
$\hat{\beta}_{\mathrm{PIP}}$	−0.0107	−0.0022	−0.0093	0.0033	0.0079
	0.0026/0.0023	0.0030/0.0027	0.0036/0.0029	0.0033/0.0028	0.0031/0.0027
	0.896	0.908	0.876	0.890	0.908
$\hat{\beta}_{\mathrm{PIP}A}$	0.0005	0.0107	0.0001	−0.0040	0.0034
	0.0019/∗	0.0026/∗	0.0020/∗	0.0019/∗	0.0018/∗
	0.919	0.933	0.926	0.932	0.933
$\hat{\beta}_A$	0.0015	−0.0055	0.0004	−0.0022	0.0015
	0.0017/0.0018	0.0022/0.0025	0.0015/0.0015	0.0015/0.0015	0.0016/0.0015
	0.944	0.975	0.954	0.933	0.932

For each entry, the first line displays the bias, the second line is the empirical standard error/averaged asymptotic standard error and the third line is the 95% coverage probability. An "∗" indicates the asymptotic standard error formula unavailable

2 and 3, $\hat{\beta}_{\text{PIP}}$ and $\hat{\beta}_{\text{PIP}A}$ should be consistent. The deviation from the expectation arises from the estimated positive but near-zero selection probabilities. These inverse-probability weights make $\hat{\beta}_{\text{PIP}}$ and $\hat{\beta}_{\text{PIP}A}$ unstable and skewed distributed, resulting in large standard errors and biases. The near-zero selection probabilities also have influence on the sandwich formula of the asymptotic covariance, making the averaged asymptotic standard error very different from the empirical standard error and resulting in low coverages. On the other hand, our proposed estimator $\hat{\beta}_A$ performs well on bias and standard error. Its asymptotic standard error is also close to the empirical standard error. The 95% coverage probabilities of $\hat{\beta}_A$ are reasonable.

In the second scenario, we use non-normal distributions to generate data. Specifically, we generate X_i from a standardized gamma distribution $(\text{Gamma}(5, 1) - 5)/\sqrt{5}$, $Z_i \sim N_3(0, \boldsymbol{I}_3)$ and ε_i from a standardized t distribution with $df = 5$ as $t_5/\sqrt{5/3}$. We keep the same settings for the parameters. The results for the second scenario are displayed in Tables 4, 5 and 6. In this setting, the parametric model for selection probabilities is still valid, but the single-index model on the augmentation is not. However, we get conclusions similar to those from the first scenario. Estimators $\hat{\beta}_{\text{PIP}}$ and $\hat{\beta}_{\text{PIP}A}$ still have large biases and standard errors. Our proposed estimator has a slightly low coverage for β_1, but is much better compared to other estimators in terms of bias, standard errors and coverage probabilities.

These simulation results illustrate our point on the numerical issues in $\hat{\beta}_{\text{PIP}}$ and $\hat{\beta}_{\text{PIP}A}$ that are mainly caused by estimated positive but near-zero selection probabilities. Our proposed estimator $\hat{\beta}_A$ has its advantage of not only the simplicity but also that it does not need to make parametric model assumption on the selection probabilities or the conditional covariate distribution $p(X|y, Z)$. It is not sensitive to the near-zero selection probabilities and gives pretty robust estimates even when the single-index model is misspecified.

Although both scenarios have a continuous missing covariate X, our method can also be applied to the situations with a categorical missing covariate. The parallel theory should still be valid as long as the single-index model $E(X_i|Q_i) = g(Q_i^\top \gamma)$ is still true (e.g., GLMs). When X is a binary variable, the estimation procedure can even be simpler because $E(X_i^2|Q_i) = E(X_i|Q_i)$.

6 Illustrative example of data analysis

In this section, we apply our proposed method to the data collected from the Canada 2010/2011 Youth Smoking Survey (YSS). The 2010/2011 Youth Smoking Survey (YSS) is a Health Canada sponsored pan-Canadian, classroom-based survey of a representative sample of students in Grades 6 through 12. The 2010/2011 YSS was implemented in schools between October 2010 and June 2011 by provincial level teams located in the 9 participating provinces in Canada. More details can be found in *2010/2011 Youth Smoking Survey Microdata User Guide*, or from https://uwaterloo.ca/canadian-student-tobacco-alcohol-drugs-survey.

We focus on data collected from Asian students (Grade 6 through 8). The main interest is to explore the correlation between the students' self-esteem scores and

Table 4 Simulation results of 1000 replications for the normal data, $X_i \sim (\text{Gamma}(5,1)-5)/\sqrt{5}$, $Z_i \sim N_3(0, \boldsymbol{I}_3)$, $\varepsilon_i \sim t_5/\sqrt{5/3}$, with $\alpha = (2.2, -0.9, -0.7, 0, 0)$, about 20% missing at random on average

	$\hat{\beta}_0 - \beta_0$	$\hat{\beta}_1 - \beta_1$	$\hat{\beta}_2 - \beta_2$	$\hat{\beta}_3 - \beta_3$	$\hat{\beta}_4 - \beta_4$
$n = 250$					
$\hat{\beta}_F$	−0.0012	0.0009	−0.0018	0.0005	0.0001
	0.0020/0.0020	0.0021/0.0020	0.0020/0.0020	0.0020/0.0020	0.0020/0.0020
	0.950	0.940	0.956	0.939	0.949
$\hat{\beta}_{\text{CC}}$	−0.1481	−0.0284	−0.0988	0.0570	0.0294
	0.0022/0.0022	0.0023/0.0022	0.0024/0.0024	0.0023/0.0022	0.0022/0.0022
	0.447	0.918	0.725	0.859	0.921
$\hat{\beta}_{\text{PIP}}$	−0.0160	−0.0065	−0.0209	0.0091	0.0075
	0.0029/0.0022	0.0033/0.0024	0.0032/0.0025	0.0034/0.0024	0.0029/0.0024
	0.883	0.884	0.877	0.879	0.906
$\hat{\beta}_{\text{PIP}A}$	−0.0019	0.0029	0.0002	−0.0045	−0.0003
	0.0021/∗	0.0029/∗	0.0023/∗	0.0021/∗	0.0021/∗
	0.925	0.930	0.929	0.923	0.929
$\hat{\beta}_A$	−0.0005	−0.0012	0.0001	−0.0008	0.0005
	0.0021/0.0021	0.0025/0.0024	0.0021/0.0020	0.0021/0.0020	0.0020/0.0020
	0.961	0.945	0.941	0.937	0.945
$n = 500$					
$\hat{\beta}_F$	−0.0009	−0.0015	−0.0006	−0.0012	−0.0006
	0.0014/0.0014	0.0014/0.0014	0.0013/0.0014	0.0015/0.0014	0.0014/0.0014
	0.957	0.948	0.956	0.943	0.954
$\hat{\beta}_{\text{CC}}$	−0.1479	−0.0317	−0.0980	0.0544	0.0279
	0.0016/0.0016	0.0016/0.0016	0.0016/0.0017	0.0016/0.0016	0.0015/0.0015
	0.146	0.886	0.558	0.807	0.908
$\hat{\beta}_{\text{PIP}}$	−0.0103	−0.0089	−0.0138	0.0063	0.0041
	0.0022/0.0017	0.0023/0.0019	0.0025/0.0020	0.0025/0.0019	0.0023/0.0018
	0.895	0.902	0.885	0.914	0.918
$\hat{\beta}_{\text{PIP}A}$	−0.0033	0.0083	−0.0040	−0.0012	0.0007
	0.0015/∗	0.0021/∗	0.0016/∗	0.0015/∗	0.0015/∗
	0.930	0.934	0.932	0.923	0.926
$\hat{\beta}_A$	−0.0009	−0.0017	−0.0002	−0.0017	−0.0003
	0.0015/0.0015	0.0018/0.0017	0.0014/0.0014	0.0015/0.0014	0.0014/0.0014
	0.959	0.930	0.943	0.933	0.943

For each entry, the first line displays the bias, the second line is the empirical standard error/averaged asymptotic standard error and the third line is the 95% coverage probability. An "∗" indicates the asymptotic standard error formula unavailable

Table 5 Simulation results of 1000 replications for the normal data, $X_i \sim (\text{Gamma}(5,1)-5)/\sqrt{5}$, $Z_i \sim N_3(0, \boldsymbol{I}_3)$, $\varepsilon_i \sim t_5/\sqrt{5/3}$, with $\alpha = (0.5, -1, -0.5, 0, 0)$, about 40% missing at random on average

	$\hat{\beta}_0 - \beta_0$	$\hat{\beta}_1 - \beta_1$	$\hat{\beta}_2 - \beta_2$	$\hat{\beta}_3 - \beta_3$	$\hat{\beta}_4 - \beta_4$
$n = 250$					
$\hat{\beta}_F$	−0.0012	0.0009	−0.0018	0.0005	0.0001
	0.0020/0.0020	0.0021/0.0020	0.0020/0.0020	0.0020/0.0020	0.0020/0.0020
	0.950	0.940	0.956	0.939	0.949
$\hat{\beta}_{\text{CC}}$	−0.3292	−0.0527	−0.1561	0.1030	0.0542
	0.0031/0.0029	0.0028/0.0027	0.0029/0.0029	0.0027/0.0027	0.0026/0.0025
	0.048	0.899	0.594	0.755	0.877
$\hat{\beta}_{\text{PIP}}$	−0.0546	−0.0183	−0.0467	0.0282	0.0166
	0.0037/0.0027	0.0040/0.0031	0.0041/0.0030	0.0040/0.0030	0.0038/0.0029
	0.758	0.860	0.794	0.842	0.869
$\hat{\beta}_{\text{PIP}A}$	0.0002	0.0300	0.0035	−0.0077	−0.0020
	0.0026/∗	0.0044/∗	0.0027/∗	0.0026/∗	0.0024/∗
	0.931	0.918	0.937	0.930	0.933
$\hat{\beta}_A$	< 0.0001	−0.0012	0.0003	−0.0001	0.0002
	0.0022/0.0023	0.0032/0.0030	0.0021/0.0020	0.0022/0.0021	0.0021/0.0021
	0.947	0.925	0.942	0.932	0.936
$n = 500$					
$\hat{\beta}_F$	−0.0009	−0.0015	−0.0006	−0.0012	−0.0006
	0.0014/0.0014	0.0014/0.0014	0.0013/0.0014	0.0015/0.0014	0.0014/0.0014
	0.957	0.948	0.956	0.943	0.954
$\hat{\beta}_{\text{CC}}$	−0.3306	−0.0570	−0.1562	0.1014	0.0526
	0.0021/0.0021	0.0020/0.0019	0.0020/0.0020	0.0019/0.0019	0.0018/0.0018
	0.000	0.838	0.288	0.592	0.837
$\hat{\beta}_{\text{PIP}}$	−0.0317	−0.0188	−0.0291	0.0189	0.0154
	0.0037/0.0023	0.0034/0.0026	0.0040/0.0026	0.0033/0.0025	0.0034/0.0024
	0.792	0.863	0.824	0.863	0.882
$\hat{\beta}_{\text{PIP}A}$	0.0004	0.0306	−0.0015	−0.0140	−0.0108
	0.0019/∗	0.0032/∗	0.0019/∗	0.0018/∗	0.0018/∗
	0.924	0.922	0.930	0.924	0.925
$\hat{\beta}_A$	−0.0008	−0.0033	−0.0001	−0.0017	−0.0005
	0.0016/0.0016	0.0024/0.0021	0.0015/0.0014	0.0016/0.0015	0.0015/0.0015
	0.949	0.921	0.942	0.921	0.942

For each entry, the first line displays the bias, the second line is the empirical standard error/averaged asymptotic standard error and the third line is the 95% coverage probability. An "∗" indicates the asymptotic standard error formula unavailable

Table 6 Simulation results of 1000 replications for the normal data, $X_i \sim (\text{Gamma}(5,1)-5)/\sqrt{5}$, $Z_i \sim N_3(0, \boldsymbol{I}_3)$, $\varepsilon_i \sim t_5/\sqrt{5/3}$, with $\alpha = (-0.5, -0.5, -0.5, 0, 0)$, about 60% missing at random on average

	$\hat{\beta}_0 - \beta_0$	$\hat{\beta}_1 - \beta_1$	$\hat{\beta}_2 - \beta_2$	$\hat{\beta}_3 - \beta_3$	$\hat{\beta}_4 - \beta_4$
$n = 250$					
$\hat{\beta}_F$	−0.0012	0.0009	−0.0018	0.0005	0.0001
	0.0020/0.0020	0.0021/0.0020	0.0020/0.0020	0.0020/0.0020	0.0020/0.0020
	0.950	0.940	0.956	0.939	0.949
$\hat{\beta}_{\text{CC}}$	−0.2777	−0.0279	−0.0964	0.0449	0.0266
	0.0040/0.0037	0.0034/0.0033	0.0035/0.0034	0.0032/0.0032	0.0033/0.0031
	0.329	0.924	0.842	0.918	0.916
$\hat{\beta}_{\text{PIP}}$	−0.0314	−0.0137	−0.0327	0.0068	0.0100
	0.0038/0.0029	0.0043/0.0036	0.0047/0.0036	0.0046/0.0034	0.0043/0.0034
	0.839	0.879	0.849	0.877	0.886
$\hat{\beta}_{\text{PIP}A}$	−0.0477	0.1346	−0.1189	0.0309	0.0571
	0.0026/∗	0.0042/∗	0.0027/∗	0.0025/∗	0.0025/∗
	0.931	0.924	0.924	0.925	0.923
$\hat{\beta}_A$	0.0013	−0.0029	−0.0005	−0.0005	0.0003
	0.0023/0.0025	0.0035/0.0035	0.0022/0.0021	0.0022/0.0021	0.0022/0.0021
	0.960	0.947	0.937	0.927	0.929
$n = 500$					
$\hat{\beta}_F$	−0.0009	−0.0015	−0.0006	−0.0012	−0.0006
	0.0014/0.0014	0.0014/0.0014	0.0013/0.0014	0.0015/0.0014	0.0014/0.0014
	0.957	0.948	0.956	0.943	0.954
$\hat{\beta}_{\text{CC}}$	−0.2790	−0.0256	−0.0958	0.0447	0.0284
	0.0027/0.0026	0.0023/0.0023	0.0024/0.0024	0.0022/0.0023	0.0024/0.0022
	0.067	0.924	0.768	0.897	0.906
$\hat{\beta}_{\text{PIP}}$	−0.0168	−0.0104	−0.0196	0.0041	0.0101
	0.0030/0.0023	0.0034/0.0028	0.0037/0.0029	0.0036/0.0028	0.0033/0.0027
	0.877	0.902	0.868	0.908	0.893
$\hat{\beta}_{\text{PIP}A}$	0.0016	0.0118	0.0073	−0.0031	−0.0012
	0.0018/∗	0.0029/∗	0.0019/∗	0.0018/∗	0.0019/∗
	0.928	0.933	0.922	0.929	0.926
$\hat{\beta}_A$	−0.0002	< 0.0001	−0.0002	−0.0019	−0.0003
	0.0016/0.0017	0.0025/0.0025	0.0015/0.0015	0.0016/0.0015	0.0016/0.0015
	0.962	0.963	0.938	0.921	0.927

For each entry, the first line displays the bias, the second line is the empirical standard error/averaged asymptotic standard error and the third line is the 95% coverage probability. An "∗" indicates the asymptotic standard error formula unavailable

smoking status, controlling other covariates as sex, marks and BMI. The variables used are displayed below:

1. *esteem* a 0–12 score measuring the student's overall self-esteem;
2. *sex* a binary variable indicating the student's gender (0 for female and 1 for male);
3. *marks* a categorical variable with five levels describing the student's marks during the past year: mostly A's (1), mostly A's and B's (2), mostly B's and C's (3), mostly C's (4) and mostly below C's (5);
4. *smoke* originally a categorical variable with 3 levels: currently smokes, formerly smoked and never smoked. In this data set of Asian students from Grade 6 to 8, we do not have students in status of "formerly smoked." Thus, we can regard this variable as binary for smoking (smoke $= 1$) or not (smoke $= 0$);
5. *BMI* a continuous variable that measures the respondent's body mass index.

We take a subset with size $n = 493$, which has complete observations on esteem, sex, marks and smoke. In this data set, there are 121, 160 and 212 students in Grades 6 through 8, respectively. There are 252 female students and 241 male students, and only 9 smokers and 484 non-smokers. But 29.2% (144 out of 493) students have BMI missing. We consider a linear model on the self-esteem score as

$$\text{esteem} = \beta_0 + \beta_1 \text{BMI} + \beta_2 \text{sex} + \beta_3 \text{marks} + \beta_4 \text{smoke} + \varepsilon.$$

Assume that the missing mechanism is MAR and that the parametric model for selection probabilities is

$$\text{logit}(\pi) = \alpha_0 + \alpha_1 \text{esteem} + \alpha_2 \text{sex} + \alpha_3 \text{marks} + \alpha_4 \text{smoke}.$$

After fitting a logistic model, we find the p values for α_1, α_2 and α_3 as 0.0321, 0.0431 and 0.0159, respectively, so that esteem, sex and marks are significant at the significance level of 0.05.

Before estimating the regression coefficients β, we first look at the self-esteem scores of the eight smokers: {0, 0, 0, 2, 7, 8, 9, 9, 12}. We find most of them are lower than the average self-esteem score of the non-smokers of 9.24, and 4 of them have extremely low scores. This in some sense implies that the smokers among the students have a lower self-esteem score compared to the non-smokers. The results of the analysis can be found in Table 7. In the estimating procedure of $\hat{\beta}_{\text{PIP}A}$, we find $\pi_i(\hat{\alpha})$ and $\hat{\pi}_i(\hat{\gamma})$ are very close. This indicates that the assumptions in Theorem 2 might be reasonable in this situation. The values in brackets are the standard errors of the corresponding estimators using the sandwich formulas (13) and (14).

In Table 7, we observe that all methods conclude that BMI and marks are significant in the linear model, while sex is insignificant. The significant effects show that higher body mass index and worse marks lead to lower self-esteem scores. The main difference lies in the effect of smoke. The complete-case analysis and inverse-probability method give insignificant results, but the rest of the methods conclude significance. The difference is caused mainly by the smaller absolute value of the point estimates of the first two methods, compared to $\hat{\beta}_{\text{PIP}A}$ and $\hat{\beta}_A$. Since the missing mechanism is not completely missing at random, the results of $\hat{\beta}_{\text{CC}}$ are likely to be biased. Combined

Table 7 2010/2011 YSS data analysis focusing on Asian students ($n = 493$)

	$\hat{\beta}_0$ (intercept)	$\hat{\beta}_1$ (BMI)	$\hat{\beta}_2$ (sex)	$\hat{\beta}_3$ (marks)	$\hat{\beta}_4$ (smoke)
$\hat{\beta}_{\mathrm{CC}}$	12.1476(0.6649)	−0.0975(0.0291)	−0.0177(0.2322)	−0.6646(0.1747)	−1.1455(0.9920)
p value	< 0.0001	0.0009	0.9392	0.0002	0.2488
$\hat{\beta}_{\mathrm{PIP}}$	12.0189(0.7695)	−0.0966(0.0314)	0.0388(0.2251)	−0.5395(0.1858)	−1.4591(1.1036)
p value	< 0.0001	0.0022	0.8633	0.0039	0.1867
$\hat{\beta}_{\mathrm{PIP}A}$	12.4119(0.6205)	−0.1092(0.0338)	0.0038(0.1976)	−0.6034(0.1572)	−3.0881(1.2228)
p value	< 0.0001	0.0133	0.9846	0.0001	0.0119
$\hat{\beta}_A$	12.2657(0.6332)	−0.0991(0.0314)	−0.0178(0.2046)	−0.6241(0.1512)	−3.2420(1.2059)
p value	< 0.0001	0.0017	0.9305	< 0.0001	0.0074

with the comparison of the self-esteem scores between smokers and non-smokers mentioned before, we believe that the results of significance are more reliable. The performance of $\hat{\beta}_{\mathrm{PIP}}$ might be explained by the misspecification of the model on selection probabilities. Based on the analysis above, we conclude that Asian students in Grade 6 through 8 who smoke have a significantly lower self-esteem score compared to the non-smokers, controlling other covariates, BMI, sex and marks during the past year.

7 Concluding remarks

In this paper, we have proposed an unweighted mean-score-form estimator of regression coefficients through GEE with a single-index model when some covariates are missing at random. This is a semiparametric estimation approach since we only assume a single-index model on augmentation without making any distribution assumptions. We do not even specify a parametric model such as a logistic model for the missingness mechanism. We have also introduced the standard doubly robust estimator $\hat{\beta}_{\mathrm{PIP}A}$ with the same single-index model on augmentation and parametrically modeled selection probabilities. We have presented the asymptotic distribution for $\hat{\beta}_{\mathrm{PIP}A}$ and $\hat{\beta}_A$ in Theorems 1 and 2, along with the sandwich formulas of the asymptotic covariances and the choice of the bandwidth. We also have shown the asymptotic equivalence between the two augmented estimators under certain conditions. However, one important advantage of our proposed estimator over the (augmented) inverse-probability weighted estimators is that it does not include selection probabilities in the point estimation procedure so that it does not need to model π_i's and avoids the situation of having highly variable inverse-probability weights, as described in Robins et al. (2007). In this sense, numerically our proposed estimator is not sensitive to positive but near-zero selection probabilities, while the performance of the inverse-probability weighted estimators is highly influenced by those near-zero π_i's. Furthermore, compared to using a standard multivariate kernel function, the SIM we use on augmentation not only avoids the curse of dimensionality, but also keeps the efficiency of standard kernel smoothing

in some particular situations. The R code used in our simulations and the example can be found on the following website: https://github.com/zhuoersun/Missing-Data.

In this work, we only considered a single univariate covariate X in simulation studies and the real data example. The results can be easily extended to the particular case of a multivariate X when $R_i = 0$ means that all the covariates in X_i are missing at the same time. It would be interesting but more challenging to consider more complex missingness patterns such as monotone or non-monotone missingness in covariates. One can refer to Chen (2004) and Sinha et al. (2014) for more information. It would be natural to extend the proposed methodology to generalized linear models. However, generalized linear models have a more complicated score function and do not have the simple form of augmentation like (6). Further investigation will be required in this important problem. Yet as another future research problem, it would also be interesting to apply this idea to longitudinal data with some covariates partially missing.

Acknowledgements The authors thank the Associate Editor and two referees for their helpful comments and suggestions that have led to much improvement of this paper. This research was supported in part by the Simons Foundation Mathematics and Physical Sciences—Collaboration Grants for Mathematicians Program Award No. 499650.

Appendix A

A.1 Regularity conditions

To establish the asymptotic theory in this work, we first assume the following general regularity conditions:

(i) The smoothing parameter h satisfies $nh^2 \to \infty$ and $nh^{2r} \to 0$, as $n \to \infty$.
(ii) All the selection probabilities π_i's are bounded away from zero.
(iii) The selection probability function on the single-index $\pi^*(\gamma)$ has r continuous and bounded partial derivatives a.e.
(iv) The density function $f(u)$ of U and the conditional density function $f_{U|R}(u)$ of $U|R$ have r continuous and bounded partial derivatives a.e.
(v) The conditional distributions $f_{U|R=0}(u)$ and $f_{U|R=1}(u)$ have the same support, and $b(u) = f_{U|R=0}(u)/f_{U|R=1}(u)$ is bounded over the support.
(vi) The conditional expectations $\psi(u|\gamma) = E(T|Q^\top\gamma = u)$ and $E(TT^\top|Q^\top\gamma)$ exist and have r continuous and bounded partial derivatives a.e.
(vii) For score T, $E(TT^\top)$ and $E\{(\partial/\partial\beta)T\}$ exist and are positive definite, and $(\partial^2/\partial\beta\partial\beta^\top)T$ exists and is continuous with respect to β a.e.

A.2 Proof of Lemma 1

Proof The idea in the proof is similar to that in the proof of Lemma 1 in Wang and Wang (2001). Recall that $u_i = Q_i^\top\gamma = y_i - \beta_Z^\top Z_i$ is the single index and that n_1 is the number of complete cases. Let

$$\hat{f}_{U|R=1}(u) = \frac{1}{n_1 h}\sum_{k=1}^{n} R_k K_h(u-u_k), \quad E_n(u) = \hat{f}_{U|R=1}(u) - f_{U|R=1}(u),$$

$$V_{ni} = \hat{f}_{U|R=1}(u_i), \quad W_{ni} = \frac{1}{n_1 h}\sum_{k=1}^{n} R_k T_{i,k} K_h(u_i - u_k).$$

Under the regularity conditions, we have $E\{E_n(u)\} = O(h^r)$ and $\text{var}\{E_n(u)\} = O\{(nh)^{-1}\}$ by the Taylor expansions. Then by the Chebyshev inequality, $E_n(u) - E\{E_n(u)\} = O_p\{(nh)^{-1/2}\}$, which implies $E_n(u) = O_p\{h^r + (nh)^{-1/2}\}$, and thus $E_n(u_i) = O_p\{h^r + (nh)^{-1/2}\}$. Similarly, we have $W_{ni} - \psi_i V_{ni} = O_p\{h^r + (nh)^{-1/2}\}$.

Define $\delta_n = h^{2r} + (nh)^{-1}$. Under the SIM condition,

$$\begin{aligned}\hat{\psi}_i - \psi_i &= \frac{W_{ni} - \psi_i V_{ni}}{f_{U|R=1}(u_i)} - \frac{(W_{ni} - \psi_i V_{ni}) E_n(u_i)}{V_{ni} f_{U|R=1}(u_i)} \\ &= \frac{W_{ni} - \psi_i V_{ni}}{f_{U|R=1}(u_i)} + O_p(\delta_n). \end{aligned} \tag{A.1}$$

Let $Q_i^* = R_i Q_i$, $X_i^* = R_i X_i$ for $i = 1, \ldots, n$ as the values of the complete cases. Then

$$\begin{aligned}
&E\left\{ \frac{W_{ni} - \psi_i V_{ni}}{f_{U|R=1}(u_i)} \middle| R_i = 0, \text{all } (R, Q^*, X^*) \right\} \\
&= \frac{1}{n_1}\sum_{k=1}^{n} R_k \int \frac{(T_{i,k} - \psi_i) K_h(u_i - u_k)}{h f_{U|R=1}(u_i)} f_{Q|R=0}(Q_i) \mathrm{d}Q_i \\
&= \frac{1}{n_1}\sum_{k=1}^{n} R_k \iint \frac{(T_{i,k} - \psi_i) K_h(u_i - u_k)}{h f_{U|R=1}(u_i)} f_{U,Z|R=0}(u_i, Z_i) \mathrm{d}Z_i \mathrm{d}u_i \\
&= \frac{1}{n_1}\sum_{k=1}^{n} R_k \int \left\{ \int \frac{(T_{i,k} - \psi_i) K_h(u_i - u_k)}{h f_{U|R=1}(u_i)} f_{Z|U,R=0}(Z_i) \mathrm{d}Z_i \right\} f_{U|R=0}(u_i) du_i \\
&= \frac{1}{n_1}\sum_{k=1}^{n} R_k \int \frac{\left(T_{i,k}^0 - \psi_i^0\right) K_h(u_i - u_k)}{h f_{U|R=1}(u_i)} f_{U|R=0}(u_i) \mathrm{d}u_i \\
&= \frac{1}{n_1}\sum_{k=1}^{n} R_k \left(T_k^0 - \psi_k^0\right) b(u_k) + O_p(h^r),
\end{aligned}$$

where $T_{i,k}^0 = E_{Z_i|u_i, R_i=0}(T_{i,k}) = \int T_{i,k} f(Z_i|u_i, R_i = 0) \mathrm{d}Z_i$, $b(u)$ is defined in regularity condition (iv). The last step is because of the concentration of u_i on u_k. Using the same idea and $\{\cdots\}$ to denote a repeat of the preceding term, we also have

$$
\begin{aligned}
&\operatorname{var}\left\{\frac{W_{ni}-\psi_i V_{ni}}{f_{U|R=1}(u_i)}\middle| R_i=0, \text{all } (R, Q^*, X^*)\right\}\\
&=\frac{1}{n_1^2}\sum_{k=1}^{n} R_k\Bigg[\int\left\{\frac{(T_{i,k}-\psi_i)K_h(u_i-u_k)}{hf_{U|R=1}(u_i)}\right\}\{\cdots\}^{\top} f_{Q|R=0}(Q_i)\mathrm{d}Q_i\\
&\quad-\left\{\sum_{k=1}^{n} R_k\int\frac{(T_{i,k}-\psi_i)K_h(u_i-u_k)}{hf_{U|R=1}(u_i)} f_{Q|R=0}(Q_i)\mathrm{d}Q_i\right\}\{\cdots\}^{\top}\Bigg]+O_p\left(\frac{1}{nh}\right)\\
&=O_p\left(\frac{1}{nh}\right).
\end{aligned}
$$

Let

$$
S_n=n^{-1/2}\sum_{i=1}^{n}(1-R_i)\left\{\frac{W_{ni}-\psi_i V_{ni}}{f_{U|R=1}(u_i)}-\frac{1}{n_1}\sum_{k=1}^{n} R_k\left(T_k^0-\psi_k^0\right)b(u_k)\right\}.
$$

Then the summations with $R_i=0$ in S_n are i.i.d. random variables conditioning on all (R, Q^*, X^*). Thus, we have

$$
\begin{aligned}
\operatorname{var}\{S_n|\text{all } (R, Q^*, X^*)\}&=\frac{n-n_1}{n}\operatorname{var}\left\{\frac{W_{n1}-\psi_1 V_{n1}}{f_{U|R=1}(u_1)}\middle|\text{ all } (R, Q^*, X^*)\right\}\\
&=O_p\left(h^{2r}+\frac{1}{nh}\right).
\end{aligned}
$$

Then $E(S_n)=O(h^r)$ and $\operatorname{var}(S_n)=O(h^{2r}+(nh)^{-1})$ imply $S_n=O_p(\eta_n)$. Back to (A.1), we have

$$
\begin{aligned}
n^{-1/2}\sum_{i=1}^{n}(1-R_i)\left(\hat{\psi}_i-\psi_i\right)&=n^{-1/2}\sum_{i=1}^{n}\left\{(1-R_i)\frac{1}{n_1}\sum_{k=1}^{n} R_k\left(T_k^0-\psi_k^0\right)b(u_k)\right\}+O_p(\eta_n)\\
&=n^{-1/2}\sum_{k=1}^{n} R_k\left(T_k^0-\psi_k^0\right)a(u_k)+O_p(\eta_n).
\end{aligned}
$$

□

A.3 Proof of Lemma 2

Proof (a) The proof is analogous to that of Lemma 1. The main difference is that this is the summation of the complete cases. Thus we need to condition on $R_i=1$. Then

$$
\begin{aligned}
&E\left\{\frac{W_{ni}-\psi_i V_{ni}}{f_{U|R=1}(u_i)}\middle| R_i=1, \text{all } (R, Q^*, X^*)\right\}\\
&=\frac{1}{n_1}\sum_{k=1}^{n} R_k \int \frac{(T_{i,k}-\psi_i)K_h(u_i-u_k)}{hf_{U|R=1}(u_i)} f_{Q|R=1}(Q_i)\mathrm{d}Q_i\\
&=\frac{1}{n_1}\sum_{k=1}^{n} R_k \iint \frac{(T_{i,k}-\psi_i)K_h(u_i-u_k)}{hf_{U|R=1}(u_i)} f_{U,Z|R=1}(u_i, Z_i)\mathrm{d}Z_i\mathrm{d}u_i\\
&=\frac{1}{n_1}\sum_{k=1}^{n} R_k \int \left\{\int \frac{(T_{i,k}-\psi_i)K_h(u_i-u_k)}{hf_{U|R=1}(u_i)} f_{Z|U,R=1}(Z_i)\mathrm{d}Z_i\right\} f_{U|R=1}(u_i)du_i\\
&=\frac{1}{n_1}\sum_{k=1}^{n} R_k \int \frac{\left(T^1_{i,k}-\psi^1_i\right)K_h(u_i-u_k)}{hf_{U|R=1}(u_i)} f_{U|R=1}(u_i)\mathrm{d}u_i\\
&=\frac{1}{n_1}\sum_{k=1}^{n} R_k \left(T^1_k-\psi^1_k\right)+O_p(h^r),
\end{aligned}
$$

where $T^1_{i,k}=E_{Z_i|u_i,R_i=1}(T_{i,k})=\int T_{i,k}f(Z_i|u_i, R_i=1)\mathrm{d}Z_i$. The rest of the proof follows in the same manner as in the proof of Lemma 1.

(b) Similarly to the proof of (a), we have

$$
n^{-1/2}\sum_{i=1}^{n}\frac{R_i}{\pi^*_i(\gamma)}\left\{\hat{\psi}_i(\gamma)-\psi_i(\gamma)\right\}=n^{-1/2}\sum_{i=1}^{n}\frac{R_i}{\pi^*_i(\gamma)}\left\{T^1_i-\psi^1_i(\gamma)\right\}+O_p(\eta_n).
$$

According to the Hölder inequality for the sum of the product terms in the second term below, we have

$$
\begin{aligned}
&n^{-1/2}\sum_{i=1}^{n}\frac{R_i}{\hat{\pi}^*_i(\gamma)}\left\{\hat{\psi}_i(\gamma)-\psi_i(\gamma)\right\}\\
&=n^{-1/2}\sum_{i=1}^{n}\frac{R_i}{\pi^*_i(\gamma)}\left\{\hat{\psi}_i(\gamma)-\psi_i(\gamma)\right\}\\
&\quad+n^{-1/2}\sum_{i=1}^{n}\frac{R_i}{\hat{\pi}^*_i(\gamma)\pi^*_i(\gamma)}\{\pi^*_i(\gamma)-\hat{\pi}^*_i(\gamma)\}\left\{\hat{\psi}_i(\gamma)-\psi_i(\gamma)\right\}\\
&=n^{-1/2}\sum_{i=1}^{n}\frac{R_i}{\pi^*_i(\gamma)}\left\{T^1_i-\psi^1_i(\gamma)\right\}+O_p(\eta_n).
\end{aligned}
$$

(c) The proof can be obtained analogously as in (b). □

A.4 Proof of Theorem 1

Proof Based on the conclusion of Lemma 1,

$$\begin{aligned}\Delta_3\left(\beta, \hat{\psi}(\gamma)\right) &= n^{-1/2}\sum_{i=1}^{n} R_i T_i + (1-R_i)\psi_i(\gamma) + (1-R_i)\left\{\hat{\psi}_i(\gamma) - \psi_i(\gamma)\right\} \\ &= n^{-1/2}\sum_{i=1}^{n} R_i T_i + (1-R_i)\psi_i(\gamma) \\ &\quad + R_i\left\{T_i^0 - \psi_i^0(\gamma)\right\} a\left(Q_i^\top \gamma\right) + O_p(\eta_n) \\ &= n^{-1/2}\sum_{i=1}^{n} U_i + O_p(\eta_n).\end{aligned}$$

Since $\Delta_3(\beta, \hat{\psi}(\gamma))$ is asymptotically equivalent to a sum of i.i.d. random variables, $\hat{\beta}_A$ is asymptotically normally distributed and has the asymptotic covariance $\boldsymbol{\Sigma}_A = \boldsymbol{D}^{-1}\boldsymbol{\mathcal{M}}\boldsymbol{D}^{-1}$ with

$$\begin{aligned}\boldsymbol{\mathcal{M}} &= \operatorname{cov}\left(n^{-1/2}\sum_{i=1}^{n} U_i\right) = \operatorname{cov}(U_1) \\ &= \operatorname{cov}\{R_1 T_1 + (1-R_1)\psi_i\} + \operatorname{cov}\left[R_i\left\{T_i^0 - \psi_i^0(\gamma)\right\} a(Q_i^\top \gamma)\right] \\ &\quad + 2\operatorname{cov}\left(R_1 T_1 + (1-R_1)\psi_i, R_i\left\{T_i^0 - \psi_i^0(\gamma)\right\} a\left(Q_i^\top \gamma\right)\right) \\ &= \boldsymbol{\mathcal{A}} + \boldsymbol{\mathcal{B}} + 2\boldsymbol{\mathcal{C}}.\end{aligned}$$

□

A.5 Proof of Theorem 2

Proof We first consider the first part, $\Delta_1(\beta, \pi(\hat{\alpha}))$, of its estimating Eq. (11). By assumption, a correctly specified parametric model for the selection probabilities with parameter α is given by

$$\pi_i = \pi_i(\alpha) = E(R_i|Q_i) = \pi(\alpha|Q_i).$$

The log-likelihood is

$$l(\alpha) = \sum_{i=1}^{n} R_i \log\{\pi_i(\alpha)\} + (1-R_i)\log\{1-\pi_i(\alpha)\}.$$

The corresponding estimating equation for MLE $\hat{\alpha}$ is given by

$$n^{-1/2}\sum_{i=1}^{n}\frac{\pi_i'(\alpha)}{\pi_i(\alpha)\{1-\pi_i(\alpha)\}}\{R_i-\pi_i(\alpha)\}=0.$$

Then we have

$$n^{1/2}(\hat{\alpha}-\alpha)=\left[E\left\{\frac{\pi_1'(\alpha)\pi_1'(\alpha)^\top}{\pi_1(\alpha)\{1-\pi_1(\alpha)\}}\right\}\right]^{-1}\left\{n^{-1/2}\sum_{i=1}^{n}\frac{\pi_i'(\alpha)}{\pi_i(\alpha)\{1-\pi_i(\alpha)\}}\{R_i-\pi_i(\alpha)\}\right\}+O_p(n^{-1/2}).$$

Moreover,

$$\begin{aligned}\Delta_1(\beta,\pi(\hat{\alpha}))&=n^{-1/2}\sum_{i=1}^{n}\frac{R_i}{\pi_i(\hat{\alpha})}T_i\\&=n^{-1/2}\sum_{i=1}^{n}\frac{R_i}{\pi_i(\alpha)}T_i-n^{-1/2}\sum_{i=1}^{n}\frac{R_i}{\pi_i^2(\alpha)}T_i\pi_i'(\alpha)^\top(\hat{\alpha}-\alpha)+O_p(n^{-1/2})\\&=n^{-1/2}\sum_{i=1}^{n}\frac{R_i}{\pi_i(\alpha)}T_i-E\left\{\frac{1}{\pi_1(\alpha)}\psi_1\pi_1'(\alpha)^\top\right\}n^{1/2}(\hat{\alpha}-\alpha)+O_p(n^{-1/2})\\&=n^{-1/2}\sum_{i=1}^{n}\frac{R_i}{\pi_i}T_i-E\left\{\frac{1}{\pi_1(\alpha)}\psi_1\pi_1'(\alpha)^\top\right\}\left[E\left\{\frac{\pi_1'(\alpha)\pi_1'(\alpha)^\top}{\pi_1(\alpha)\{1-\pi_1(\alpha)\}}\right\}\right]^{-1}\\&\quad\left\{n^{-1/2}\sum_{i=1}^{n}\frac{\pi_i'(\alpha)}{\pi_i(\alpha)\{1-\pi_i(\alpha)\}}\{R_i-\pi_i(\alpha)\}\right\}+O_p(n^{-1/2})\\&=\Delta_1(\beta,\pi)-\boldsymbol{F}(\alpha)\boldsymbol{C}^{-1}(\alpha)P_n(\alpha)+O_p(n^{-1/2}),\end{aligned}$$

where $\boldsymbol{F}(\alpha)=E\left\{\frac{1}{\pi_1(\alpha)}\psi_1\pi_1'(\alpha)^\top\right\}$, $\boldsymbol{C}(\alpha)=E\left\{\frac{\pi_1'(\alpha)\pi_1'(\alpha)^\top}{\pi_1(\alpha)\{1-\pi_1(\alpha)\}}\right\}$, $P_n(\alpha)=n^{-1/2}\sum_{i=1}^{n}\frac{\pi_i'(\alpha)}{\pi_i(\alpha)\{1-\pi_i(\alpha)\}}\{R_i-\pi_i(\alpha)\}$.

We now consider the second part of the estimating equation. By Lemmas 1 and 2(a), we obtain that

$$\begin{aligned}n^{-1/2}\sum_{i=1}^{n}\left\{\hat{\psi}_i(\gamma)-\psi_i(\gamma)\right\}&=n^{-1/2}\sum_{i=1}^{n}R_i\left\{\hat{\psi}_i(\gamma)-\psi_i(\gamma)\right\}\\&\quad+n^{-1/2}\sum_{i=1}^{n}(1-R_i)\left\{\hat{\psi}_i(\gamma)-\psi_i(\gamma)\right\}\end{aligned}$$

$$= n^{-1/2} \sum_{i=1}^{n} R_i \left\{ T_i^1 - \psi_i^1(\gamma) \right\}$$
$$+ n^{-1/2} \sum_{i=1}^{n} R_i \left\{ T_i^0 - \psi_i^0(\gamma) \right\} a\left(Q_i^\top \gamma \right) + O_p(\eta_n).$$

Recall that the additional condition for Lemma 2(c) requires $\pi_i = \pi_i^*(\gamma)$. This implies that $T_i^0 = T_i^1 = E_{Z_i|u_i}(T_i)$, $\psi_i^0(\gamma) = \psi_i^1(\gamma) = E_{Z_i|u_i}\{\psi_i(\gamma)\}$. Let $T_i^* = E_{Z_i|u_i}(T_i)$, $\psi_i^*(\gamma) = E_{Z_i|u_i}\{\psi_i(\gamma)\}$. Then

$$n^{-1/2} \sum_{i=1}^{n} \left\{ \hat{\psi}_i(\gamma) - \psi_i(\gamma) \right\} = n^{-1/2} \sum_{i=1}^{n} \frac{R_i}{\pi_i} \left\{ T_i^* - \psi_i^*(\gamma) \right\} + O_p(\eta_n). \quad \text{(A.2)}$$

Equation (A.2) and Lemma 2(c) imply that

$$n^{-1/2} \sum_{i=1}^{n} \left\{ 1 - \frac{R_i}{\pi_i(\hat{\alpha})} \right\} \{ \hat{\psi}_i(\gamma) - \psi_i(\gamma) \} = O_p(\eta_n).$$

Then

$$n^{-1/2} \sum_{i=1}^{n} \left\{ 1 - \frac{R_i}{\pi_i(\hat{\alpha})} \right\} \hat{\psi}_i(\gamma) = n^{-1/2} \sum_{i=1}^{n} \left\{ 1 - \frac{R_i}{\pi_i(\hat{\alpha})} \right\} \psi_i(\gamma) + O_p(\eta_n).$$

As in the proof for the first part $\Delta_1(\beta, \pi(\hat{\alpha}))$, we can show that

$$n^{-1/2} \sum_{i=1}^{n} \frac{R_i}{\pi_i(\hat{\alpha})} \psi_i(\gamma) = n^{-1/2} \sum_{i=1}^{n} \frac{R_i}{\pi_i} \psi_i(\gamma) - \boldsymbol{F}(\alpha) \boldsymbol{C}^{-1}(\alpha) P_n(\alpha) + O_p(n^{-1/2}).$$

Finally we have

$$\begin{aligned}
\Delta_2\left(\beta, \pi(\hat{\alpha}), \hat{\psi}(\gamma)\right) &= n^{-1/2} \sum_{i=1}^{n} \frac{R_i}{\pi_i(\hat{\alpha})} T_i + \left\{ 1 - \frac{R_i}{\pi_i(\hat{\alpha})} \right\} \hat{\psi}_i(\gamma) \\
&= n^{-1/2} \sum_{i=1}^{n} \frac{R_i}{\pi_i} T_i - \boldsymbol{F}(\alpha) \boldsymbol{C}^{-1}(\alpha) P_n(\alpha) + n^{-1/2} \sum_{i=1}^{n} \psi_i(\gamma) \\
&\quad - n^{-1/2} \sum_{i=1}^{n} \frac{R_i}{\pi_i} \psi_i(\gamma) + \boldsymbol{F}(\alpha) \boldsymbol{C}^{-1}(\alpha) P_n(\alpha) + O_p(\eta_n) \\
&= n^{-1/2} \sum_{i=1}^{n} \frac{R_i}{\pi_i} T_i + n^{-1/2} \sum_{i=1}^{n} \left(1 - \frac{R_i}{\pi_i} \right) \psi_i(\gamma) + O_p(\eta_n)
\end{aligned}$$

$$= n^{-1/2} \sum_{i=1}^{n} \frac{R_i}{\pi_i^*(\gamma)} T_i + n^{-1/2} \sum_{i=1}^{n} \left\{ 1 - \frac{R_i}{\pi_i^*(\gamma)} \right\} \psi_i(\gamma) + O_p(\eta_n)$$
$$= \Delta_2 \left(\beta, \pi^*(\gamma), \psi \right) + O_p(\eta_n).$$

In summary, we have shown that $\Delta_2(\beta, \pi(\hat{\alpha}), \hat{\psi}(\gamma))$ is asymptotically equivalent to $\Delta_2(\beta, \pi^*(\gamma), \psi)$, which is a sum of i.i.d. terms. Hence, $\hat{\beta}_{\mathrm{PIP}A}$ is asymptotically equivalent to the solution of $\Delta_2(\beta, \pi^*(\gamma), \psi) = 0$, having asymptotic normality with asymptotic covariance

$$\boldsymbol{\Sigma}_{PA} = \boldsymbol{D}^{-1}(\boldsymbol{S} - \boldsymbol{S}^* + \boldsymbol{V})\boldsymbol{D}^{-1}.$$

□

A.6 Proof of Corollary 1

Proof By the fact that

$$\Delta_1 \left(\beta, \pi(\hat{\alpha}) \right) = \Delta_1(\beta, \pi) - \boldsymbol{F}(\alpha)\boldsymbol{C}^{-1}(\alpha) P_n(\alpha) + O_p(n^{-1/2}),$$

where $\boldsymbol{F}(\alpha)$ and $\boldsymbol{C}(\alpha)$ are given in the proof of Theorem 1, and by (A.1) in Wang et al. (1997), with an extension to a general parametric model, we have the asymptotic covariance for $\hat{\beta}_{\mathrm{PIP}}$ as

$$\boldsymbol{\Sigma}_P = \boldsymbol{D}^{-1} \left\{ \tilde{\boldsymbol{S}} - \boldsymbol{F}(\alpha)\boldsymbol{C}^{-1}(\alpha)\boldsymbol{F}(\alpha)^\top \right\} \boldsymbol{D}^{-1},$$

where $\tilde{\boldsymbol{S}} = E(T_1 T_1^\top / \pi_1)$. By Wang and Wang (2001),

$$\tilde{\boldsymbol{\Sigma}} = \boldsymbol{D}^{-1}(\tilde{\boldsymbol{S}} - \tilde{\boldsymbol{S}}^* + \boldsymbol{V})\boldsymbol{D}^{-1}$$

is the asymptotic covariance matrix for $\hat{\beta}$ when $\hat{\psi}$ is based on a standard kernel smoother, where $\tilde{\boldsymbol{S}}^* = E(\psi_1 \psi_1^\top / \pi_1)$.

First we show that $\boldsymbol{\Sigma}_P \succeq \tilde{\boldsymbol{\Sigma}}$. By the construction of the covariances, we only need to show that $\tilde{\boldsymbol{S}}^* - \boldsymbol{V} \succeq \boldsymbol{F}(\alpha)\boldsymbol{C}^{-1}(\alpha)\boldsymbol{F}(\alpha)^\top$. Define $\xi = \left(\sqrt{\frac{1-\pi_1}{\pi_1}} \psi_1, \frac{\pi_1'(\alpha)}{\sqrt{(1-\pi_1)\pi_1}} \right)^\top$. Then we have

$$E(\xi\xi^\top) = \begin{pmatrix} E\left(\frac{1-\pi_1}{\pi_1} \psi_1 \psi_1^\top \right) & E\left\{ \frac{1}{\pi_1} \psi_1 \pi_1'(\alpha)^\top \right\} \\ E\left\{ \frac{1}{\pi_1} \pi_1'(\alpha) \psi_1^\top \right\} & E\left\{ \frac{\pi_1'(\alpha)\pi_1'(\alpha)^\top}{(1-\pi_1)\pi_1} \right\} \end{pmatrix}$$
$$= \begin{pmatrix} \tilde{\boldsymbol{S}}^* - \boldsymbol{V} & \boldsymbol{F}(\alpha) \\ \boldsymbol{F}(\alpha)^\top & \boldsymbol{C}(\alpha) \end{pmatrix} \succeq 0.$$

By the Schur complement condition of the matrix above, we have

$$(\tilde{\boldsymbol{S}}^* - \boldsymbol{V}) - \boldsymbol{F}(\alpha)\boldsymbol{C}^{-1}(\alpha)\boldsymbol{F}(\alpha)^\top \succeq 0.$$

Therefore, $\tilde{\boldsymbol{S}}^* - \boldsymbol{V} \succeq \boldsymbol{F}(\alpha)\boldsymbol{C}^{-1}(\alpha)\boldsymbol{F}(\alpha)^\top$, which implies that $\boldsymbol{\Sigma}_P \succeq \tilde{\boldsymbol{\Sigma}}$.

Next, we show that $\tilde{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}_A = \boldsymbol{\Sigma}_{PA}$ and thus the asymptotic equivalence between $\hat{\beta}_A$ and $\hat{\beta}_{\mathrm{PIP}A}$. Based on the results of Theorem 1, we can rewrite $\Delta_3(\beta, \hat{\psi}(\gamma))$ as

$$\begin{aligned}\Delta_3\left(\beta, \hat{\psi}(\gamma)\right) &= n^{-1/2}\sum_{i=1}^{n} U_i + O_p(\eta_n) \\ &= n^{-1/2}\sum_{i=1}^{n}\left[\frac{R_i}{\pi_i^*(\gamma)}T_i + \left\{1 - \frac{R_i}{\pi_i^*(\gamma)}\right\}\psi_i \right. \\ &\quad \left. + R_i a(Q_i^\top \gamma)\left\{(T_i^0 - \psi_i^0) - (T_i - \psi_i)\right\}\right] + O_p(\eta_n).\end{aligned}$$

The condition $E(Z_i|u_i) = Z_i$ implies that $T_i^0 = T_i^1 = T_i$ and $\psi_i^0(\gamma) = \psi_i^1(\gamma) = \psi_i(\gamma)$. By Theorem 2, both $\Delta_2(\beta, \pi(\hat{\alpha}), \hat{\psi}(\gamma))$ and $\Delta_3(\beta, \hat{\psi}(\gamma))$ are asymptotically equivalent to $\Delta_2(\beta, \pi^*(\gamma), \psi)$ and thus have the same asymptotic covariance matrix as

$$\boldsymbol{\Sigma}_A = \boldsymbol{\Sigma}_{PA} = \boldsymbol{D}^{-1}(\boldsymbol{S} - \boldsymbol{S}^* + \boldsymbol{V})\boldsymbol{D}^{-1}.$$

Recall the condition of Lemma 2(c) that $\pi_i = \pi_i^*(\gamma)$. Then $\boldsymbol{S} = \tilde{\boldsymbol{S}}$, $\boldsymbol{S}^* = \tilde{\boldsymbol{S}}^*$. Thus, we finally have

$$\boldsymbol{\Sigma}_P \succeq \tilde{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}_A = \boldsymbol{\Sigma}_{PA}.$$

□

References

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

Chen, H. Y. (2004). Nonparametric and semiparametric models for missing covariates in parametric regression. *Journal of the American Statistical Association*, *99*, 1176–1189.

Fuchs, C. (1982). Maximum likelihood estimation and model selection in contingency tables with missing data. *Journal of the American Statistical Association*, *77*, 270–278.

Han, P. (2014). Multiply robust estimation in regression analysis with missing data. *Journal of the American Statistical Association*, *109*, 1159–1173.

Han, P. (2016). Combining inverse probability weighting and multiple imputation to improve robustness of estimation. *Scandinavian Journal of Statistics*, *43*, 246–260.

Han, P., Wang, L. (2013). Estimation with missing data: Beyond double robustness. *Biometrika*, *100*, 417–430.

Hartley, H., Hocking, R. (1971). The analysis of incomplete data. *Biometrics*, *27*, 783–823.

Hsu, C.-H., Long, Q., Li, Y., Jacobs, E. (2014). A nonparametric multiple imputation approach for data with missing covariate values with application to colorectal adenoma data. *Journal of Biopharmaceutical Statistics*, *24*, 634–648.

Ibrahim, J. G. (1990). Incomplete data in generalized linear models. *Journal of the American Statistical Association*, *85*, 765–769.
Ibrahim, J. G., Chen, M.-H., Lipsitz, S. R. (2002). Bayesian methods for generalized linear models with covariates missing at random. *Canadian Journal of Statistics*, *30*, 55–78.
Ibrahim, J. G., Chen, M.-H., Lipsitz, S. R., Herring, A. H. (2005). Missing-data methods for generalized linear models: A comparative review. *Journal of the American Statistical Association*, *100*, 332–346.
Kang, J. D., Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, *22*, 523–539.
Little, R. J., Rubin, D. B. (2014). *Statistical analysis with missing data*. New Jersey: Wiley.
Reilly, M., Pepe, M. S. (1995). A mean score method for missing and auxiliary covariate data in regression models. *Biometrika*, *82*, 299–314.
Robins, J. M., Ritov, Y. (1997). Toward a curse of dimensionality appropriate(coda) asymptotic theory for semi-parametric models. *Statistics in Medicine*, *16*, 285–319.
Robins, J. M., Rotnitzky, A., Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, *89*, 846–866.
Robins, J., Sued, M., Lei-Gomez, Q., Rotnitzky, A. (2007). Comment: Performance of double-robust estimators when "inverse probability" weights are highly variable. *Statistical Science*, *22*, 544–559.
Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*, 581–592.
Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys*. New Jersey: Wiley.
Schluchter, M. D., Jackson, K. L. (1989). Log-linear analysis of censored survival data with partially observed covariates. *Journal of the American Statistical Association*, *84*, 42–52.
Sepanski, J., Knickerbocker, R., Carroll, R. (1994). A semiparametric correction for attenuation. *Journal of the American Statistical Association*, *89*, 1366–1373.
Sinha, S., Saha, K. K., Wang, S. (2014). Semiparametric approach for non-monotone missing covariates in a parametric regression model. *Biometrics*, *70*, 299–311.
Wang, C., Wang, S., Zhao, L.-P., Ou, S.-T. (1997). Weighted semiparametric estimation in regression analysis with missing covariate data. *Journal of the American Statistical Association*, *92*, 512–525.
Wang, S., Wang, C. (2001). A note on kernel assisted estimators in missing covariate regression. *Statistics & Probability Letters*, *55*, 439–449.
Zhou, Y., Wan, A. T. K., Wang, X. (2008). Estimating equations inference with missing data. *Journal of the American Statistical Association*, *103*, 1187–1199.