Statistica Sinica **26** (2016), 691-719
doi:http://dx.doi.org/10.5705/ss.2014.030

DOUBLY ROBUST AND LOCALLY EFFICIENT ESTIMATION WITH MISSING OUTCOMES

Peisong Han[1], Lu Wang[2] and Peter X.-K. Song[2]

[1] *University of Waterloo and* [2] *University of Michigan*

Abstract: We consider parametric regression where the outcome is subject to missingness. To achieve the semiparametric efficiency bound, most existing estimation methods require the correct modeling of certain second moments of the data, which can be very challenging in practice. We propose an estimation procedure based on the conditional empirical likelihood (CEL) method. Our method does not require us to model any second moments. We study the CEL-based inverse probability weighted (CEL-IPW) and augmented inverse probability weighted (CEL-AIPW) estimators in detail. Under some regularity conditions and the missing at random (MAR) mechanism, the CEL-IPW estimator is consistent if the missingness mechanism is correctly modeled, and the CEL-AIPW estimator is consistent if either the missingness mechanism or the conditional mean of the outcome is correctly modeled. When both quantities are correctly modeled, the CEL-AIPW estimator attains the semiparametric efficiency bound without modeling any second moments. The asymptotic distributions are derived. Numerical implementation through nested optimization routines using the Newton-Raphson algorithm is discussed.

Key words and phrases: Augmented inverse probability weighting (AIPW), auxiliary variables, conditional empirical likelihood, mean regression, missing at random (MAR), surrogate outcome.

1. Introduction

We study the problem of parametric regression when the outcome is subject to missingness. The central interest is the estimation and inference of the regression coefficients. In practice there are many reasons that can lead to missing outcomes: budget or technique restrictions, subjects' failure to comply with the protocol, or simply the study design. Missing data usually bring big challenges to estimation and inference, as the application of statistical methods developed for data without missing values can lead to biased estimation and misleading conclusions.

In addition to the outcome and covariates, we assume that some auxiliary variables are available for all subjects. Although the auxiliary variables are not of direct statistical interest, they may help to explain the missingness mechanism, and thus reduce the impact of missing data on estimation and inference.

Data with auxiliary variables arise in many observational studies (e.g., Wang, Rotnitzky, and Lin (2010)) as well as two-stage design studies (e.g., Pepe (1992); Pepe, Reilly, and Fleming (1994)), where the second-stage outcome is not observed for all subjects, and the probability of observing this outcome depends on the first-stage outcome (the auxiliary variable) and covariates.

To fix notation, let Y denote the outcome, $\boldsymbol{X}$ the vector of covariates, $\boldsymbol{\beta}$ the p-dimensional vector of regression coefficients, and $\boldsymbol{S}$ the vector of auxiliary variables. Here $\boldsymbol{\beta}$ is the parameter of interest. Pepe (1992) proposed maximum likelihood estimation, which assumes the correct specification of the densities $f(Y|\boldsymbol{X})$ and $f(\boldsymbol{S}|Y,\boldsymbol{X})$. To reduce model assumptions, Pepe, Reilly, and Fleming (1994) proposed mean score estimation, which assumes the correct specification of density $f(Y|\boldsymbol{X})$. This latter assumption is still more than necessary, and is likely subject to model misspecification. In this paper, we only specify the mean regression model as

$$E(Y|\boldsymbol{X}) = \mu(\boldsymbol{X}^T\boldsymbol{\beta}) \quad \text{for some } \boldsymbol{\beta} = \boldsymbol{\beta}_0 \in \mathbb{R}^p, \tag{1.1}$$

where $\mu(\cdot)$ is some known link function, and the expectation is taken under the true density $f(Y|\boldsymbol{X})$. Let $R = 1$ if Y is observed and $R = 0$ if Y is missing. The observed data are $(R_i, R_iY_i, \boldsymbol{S}_i, \boldsymbol{X}_i)$, $i = 1, \cdots, N$, which are independent and identically distributed. We assume that the missingness of Y does not depend on Y itself given $\boldsymbol{X}$ and $\boldsymbol{S}$, the missing at random (MAR) mechanism (Little and Rubin (2002)):

$$P(R = 1|Y, \boldsymbol{S}, \boldsymbol{X}) = P(R = 1|\boldsymbol{S}, \boldsymbol{X}) \stackrel{\text{def}}{=} \pi(\boldsymbol{S}, \boldsymbol{X}) > \tau > 0, \tag{1.2}$$

where τ is a positive constant.

The model defined by (1.1) and (1.2) is embedded in a more general missing data setting that has been studied extensively by Robins, Rotnitzky, and their colleagues using semiparametric efficiency theory as in Bickel et al. (1993). Applying the theory developed by Robins, Rotnitzky, and Zhao (1994) and Robins and Rotnitzky (1995), Yu and Nan (2006) derived the semiparametric efficiency bound under this model. Estimators whose asymptotic variance attains such bound are efficient. Chen and Breslow (2004) independently derived the bound using the theory of estimating functions (Godambe (1960, 1991); Heyde (1988, 1997); Newey and McFadden (1994)).

Most existing estimation methods for missing outcome data rely on a set of estimating functions $\boldsymbol{U}(\boldsymbol{\beta}; Y, \boldsymbol{X})$ constructed from (1.1) satisfying the unbiasedness property $E\{\boldsymbol{U}(\boldsymbol{\beta}_0; Y, \boldsymbol{X})\} = \boldsymbol{0}$. While any function $\boldsymbol{D}(\boldsymbol{\beta}; \boldsymbol{X})$ depending only on $\boldsymbol{X}$ and $\boldsymbol{\beta}$ may be used to construct some unbiased estimating functions in the form $\boldsymbol{D}(\boldsymbol{\beta}; \boldsymbol{X})\{Y - \mu(\boldsymbol{X}^T\boldsymbol{\beta})\}$, the most typically used one is

$$\frac{\partial \mu(\boldsymbol{X}^T\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \mathrm{Var}(Y|\boldsymbol{X})^{-1}\{Y - \mu(\boldsymbol{X}^T\boldsymbol{\beta})\}.$$

Under the MAR mechanism (1.2), the augmented inverse probability weighted (AIPW) estimator (Robins, Rotnitzky, and Zhao (1994, 1995)); Robins and Rotnitzky (1995); Tsiatis (2006)) is the solution to the equation

$$\sum_{i=1}^{N}\left\{\frac{R_i}{\hat{\pi}(\boldsymbol{S}_i,\boldsymbol{X}_i)}\boldsymbol{U}(\boldsymbol{\beta};Y_i,\boldsymbol{X}_i)-\frac{R_i-\hat{\pi}(\boldsymbol{S}_i,\boldsymbol{X}_i)}{\hat{\pi}(\boldsymbol{S}_i,\boldsymbol{X}_i)}\boldsymbol{\sigma}(\boldsymbol{\beta};\boldsymbol{S}_i,\boldsymbol{X}_i)\right\}=\mathbf{0}, \tag{1.3}$$

where $\hat{\pi}(\boldsymbol{S},\boldsymbol{X})$ is the estimated value of $\pi(\boldsymbol{S},\boldsymbol{X})$, and $\boldsymbol{\sigma}(\boldsymbol{\beta};\boldsymbol{S},\boldsymbol{X})$ is an arbitrary function of $\boldsymbol{\beta}$, $\boldsymbol{S}$, and $\boldsymbol{X}$. When $\boldsymbol{\sigma}(\boldsymbol{\beta};\boldsymbol{S},\boldsymbol{X})=\mathbf{0}$, the AIPW estimator reduces to the inverse probability weighted (IPW) estimator (Horvitz and Thompson (1952)). The AIPW estimator possesses a double robustness property in the sense that it is consistent if either $\pi(\boldsymbol{S},\boldsymbol{X})$ is correctly modeled, or $\boldsymbol{\sigma}(\boldsymbol{\beta};\boldsymbol{S},\boldsymbol{X})$ is a correct model for $E\{\boldsymbol{U}(\boldsymbol{\beta};Y,\boldsymbol{X})|\boldsymbol{S},\boldsymbol{X}\}$. For a fixed $\boldsymbol{U}(\boldsymbol{\beta};Y,\boldsymbol{X})$, the smallest asymptotic variance of the AIPW estimator is achieved when both $\pi(\boldsymbol{S},\boldsymbol{X})$ and $E\{\boldsymbol{U}(\boldsymbol{\beta};Y,\boldsymbol{X})|\boldsymbol{S},\boldsymbol{X}\}$ are correctly modeled, and $\boldsymbol{\sigma}(\boldsymbol{\beta};\boldsymbol{S},\boldsymbol{X})$ is taken to be the correct model for the latter. However, this $\boldsymbol{U}(\boldsymbol{\beta};Y,\boldsymbol{X})$-dependent variance is usually larger than the semiparametric efficiency bound.

In recent literature, many alternative doubly robust estimators have been proposed. These include Tan (2006, 2008, 2010), Kang and Schafer (2007), Robins et al. (2007), Rubin and van der Laan (2008), Cao, Tsiatis, and Davidian (2009), Tsiatis, Davidian, and Cao (2011), Han (2012) and Rotnitzky et al. (2012). Most of them were proposed under the setting of estimating the population mean of a response variable. Han (2012) and Rotnitzky et al. (2012) considered the regression setting, and their estimators are referred to here as the HAN and RLSR estimators, respectively. Along the lines of Tan (2006, 2010), who considered estimating the population mean and tried to improve the efficiency of the AIPW estimator when the model in the augmentation term is incorrectly specified, the HAN estimator solves an estimating equation that employs a particular linear combination of the two terms in (1.3). When $\pi(\boldsymbol{S},\boldsymbol{X})$ is correctly modeled, this linear combination yields the residual of the projection of the first term on the second, which endows the HAN estimator with improved efficiency over both the IPW and the corresponding AIPW estimators, with the exception of when $\boldsymbol{\sigma}(\boldsymbol{\beta};\boldsymbol{S},\boldsymbol{X})$ is also a correct model for $E\{\boldsymbol{U}(\boldsymbol{\beta};Y,\boldsymbol{X})|\boldsymbol{S},\boldsymbol{X}\}$, in which case the HAN and the AIPW estimators have the same efficiency. The RLSR estimator, in addition to the efficiency improvement over both the IPW and the AIPW estimators, has the property that, for a given finite set of user-specified functions, each function evaluated at the RLSR estimator has asymptotic variance no larger than that of the same function evaluated at any AIPW estimator using the same model structure for $E\{\boldsymbol{U}(\boldsymbol{\beta};Y,\boldsymbol{X})|\boldsymbol{S},\boldsymbol{X}\}$. The RLSR estimator solves an outcome regression estimating equation that, unlike equation

(1.3), always has a solution if the estimated value of $E(Y|\boldsymbol{S},\boldsymbol{X})$ falls into the sample space of Y.

Empirical likelihood (EL) (Owen (1988, 1990, 2001); Qin and Lawless (1994); Kitamura (2007)) has become a popular tool in analyzing data with missing outcome. Under the MAR mechanism (1.2), Chen, Leung, and Qin (2008) proposed an estimator, referred to here as the CLQ estimator, by solving the estimating equation

$$\sum_{i=1}^{N} \hat{p}_i \frac{R_i}{\hat{\pi}(\boldsymbol{S}_i,\boldsymbol{X}_i)} \boldsymbol{U}(\boldsymbol{\beta}; Y_i, \boldsymbol{X}_i) = \mathbf{0},$$

where $\hat{p}_i$ is the EL probability mass assigned to the data point $(R_i = 1, Y_i, \boldsymbol{S}_i, \boldsymbol{X}_i)$ after incorporating the information carried by subjects with missing values. Qin, Zhang, and Leung (2009) proposed an estimator, referred to as the QZL estimator, by solving the over-identified estimating equation

$$\sum_{i=1}^{N} \left\{ \frac{R_i}{\hat{\pi}(\boldsymbol{S}_i,\boldsymbol{X}_i)} \boldsymbol{U}(\boldsymbol{\beta}; Y_i, \boldsymbol{X}_i)^T,\ \frac{R_i - \hat{\pi}(\boldsymbol{S}_i,\boldsymbol{X}_i)}{\hat{\pi}(\boldsymbol{S}_i,\boldsymbol{X}_i)} \boldsymbol{\sigma}(\boldsymbol{\beta}; \boldsymbol{S}_i, \boldsymbol{X}_i)^T \right\}^T = \mathbf{0},$$

where the EL is used to account for the over-identification. It has been shown that, when $\pi(\boldsymbol{S},\boldsymbol{X})$ is correctly modeled, the CLQ and the QZL estimators are both more efficient than the IPW estimator. In addition, when $\boldsymbol{\sigma}(\boldsymbol{\beta}; \boldsymbol{S},\boldsymbol{X})$ is a correct model for $E\{\boldsymbol{U}(\boldsymbol{\beta}; Y,\boldsymbol{X})|\boldsymbol{S},\boldsymbol{X}\}$, both estimators asymptotically coincide with the corresponding AIPW estimator using the same $\boldsymbol{\sigma}(\boldsymbol{\beta}; \boldsymbol{S},\boldsymbol{X})$. However, when $\pi(\boldsymbol{S},\boldsymbol{X})$ is incorrectly modeled, neither the CLQ nor the QZL estimator is consistent. Qin and Zhang (2007) and Qin, Shao, and Zhang (2008) proposed estimators that possess the double robustness property, but they only considered estimating the population mean. Other works that apply the EL method to missing data problems include Chen, Leung, and Qin (2003), Wang and Chen (2009), Tan (2006, 2010, 2011), Han and Wang (2013) and Han (2014a,b).

For all these methods, the estimating functions $\boldsymbol{U}(\boldsymbol{\beta}; Y,\boldsymbol{X})$ are essential and need to be explicitly specified priori. Different estimating functions yield different estimators with varying levels of efficiency, and their numerical performance can also dramatically differ from each other. The efficient influence function for $\boldsymbol{\beta}$ in our setting is given by (e.g., Chen and Breslow (2004); Yu and Nan (2006))

$$\left(\mathrm{Var}\left[\frac{\partial g(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \mathrm{Var}\{g(\boldsymbol{\beta})|\boldsymbol{X}\}^{-1} g(\boldsymbol{\beta}) \right] \right)^{-1} \frac{\partial g(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \mathrm{Var}\{g(\boldsymbol{\beta})|\boldsymbol{X}\}^{-1} g(\boldsymbol{\beta}), \tag{1.4}$$

where

$$g(\boldsymbol{\beta}) = \frac{R}{\pi(\boldsymbol{S},\boldsymbol{X})} \left\{ Y - \mu(\boldsymbol{X}^T\boldsymbol{\beta}) \right\} - \frac{R - \pi(\boldsymbol{S},\boldsymbol{X})}{\pi(\boldsymbol{S},\boldsymbol{X})} E\left\{ Y - \mu(\boldsymbol{X}^T\boldsymbol{\beta}) | \boldsymbol{S},\boldsymbol{X} \right\}.$$

To achieve semiparametric efficiency, these methods need to correctly model the second moment $\mathrm{Var}\{g(\boldsymbol{\beta})|\boldsymbol{X}\}$, which has a very complicated structure and is

difficult to model in practice. Hence, an estimator achieving the semiparametric efficiency without modeling $\mathrm{Var}\{g(\boldsymbol{\beta})|\boldsymbol{X}\}$ or any other second moments is desirable. For many EL-based estimators, including the CLQ and the QZL estimators, another drawback is that they are not robust against incorrect modeling of the missingness mechanism. Since model misspecification commonly occurs in practice, the double robustness property is important. We propose a conditional empirical likelihood (CEL) (Zhang and Gijbels (2003); Kitamura, Tripathi, and Ahn (2004)) based method. We study two CEL-based estimators, the CEL-IPW and the CEL-AIPW estimators. The CEL-IPW estimator is consistent if $\pi(\boldsymbol{S}, \boldsymbol{X})$ is correctly modeled; CEL-AIPW estimator enjoys the double robustness property, in the sense that it is consistent if either $\pi(\boldsymbol{S}, \boldsymbol{X})$ or $E(Y|\boldsymbol{S}, \boldsymbol{X})$ is correctly modeled. When both models are correct, the CEL-AIPW estimator attains the semiparametric efficiency bound without modeling any second moments of the data.

This paper is organized as follows. Section 2 describes the CEL estimation procedure and its numerical implementation. Section 3 concerns the large sample properties. Section 4 contains the results of simulation studies. Section 5 illustrates the data application. Section 6 consists of some further discussion. Technical assumptions and proofs are provided in the Appendix.

2. CEL Estimation

2.1. CEL-based estimators

Define the IPW residual and the AIPW residual as, respectively,

$$f(\boldsymbol{\beta}) = \frac{R\left\{Y - \mu(\boldsymbol{X}^T\boldsymbol{\beta})\right\}}{\pi(\boldsymbol{S}, \boldsymbol{X})},$$

$$g(\boldsymbol{\beta}) = \frac{R}{\pi(\boldsymbol{S}, \boldsymbol{X})}\left\{Y - \mu(\boldsymbol{X}^T\boldsymbol{\beta})\right\} - \frac{R - \pi(\boldsymbol{S}, \boldsymbol{X})}{\pi(\boldsymbol{S}, \boldsymbol{X})} E\left\{Y - \mu(\boldsymbol{X}^T\boldsymbol{\beta})|\boldsymbol{S}, \boldsymbol{X}\right\}.$$

Clearly we have that $E\{f(\boldsymbol{\beta}_0)|\boldsymbol{X}\} = 0$ and $E\{g(\boldsymbol{\beta}_0)|\boldsymbol{X}\} = 0$. This conditional mean zero property of both residuals serves as the foundation of the proposed estimation procedure. In this section we focus on describing the procedure based on the AIPW residual $g(\boldsymbol{\beta})$ that yields the CEL-AIPW estimator. Estimation based on the IPW residual $f(\boldsymbol{\beta})$ that yields the CEL-IPW estimator follows a similar procedure, with no need of modeling $E(Y|\boldsymbol{S}, \boldsymbol{X})$.

We first construct the empirical version of $E\{g(\boldsymbol{\beta}_0)|\boldsymbol{X}\} = 0$ based on the observed data $(R_i, R_iY_i, \boldsymbol{X}_i, \boldsymbol{S}_i)$, $i = 1, \ldots, N$. To do this, conditional on each $\boldsymbol{X}_i$, let p_{ij} denote the empirical probabilities defined by a discrete distribution that has support on $\{g_j(\boldsymbol{\beta}) : j = 1, \ldots, N\}$, where $g_j(\boldsymbol{\beta})$ is $g(\boldsymbol{\beta})$ evaluated at $(R_j, R_jY_j, \boldsymbol{X}_j, \boldsymbol{S}_j)$. Thus, for each i, $p_{ij} = dF\{g_j(\boldsymbol{\beta})|\boldsymbol{X}_i\}$ is the jump of the distribution $F\{g(\boldsymbol{\beta})|\boldsymbol{X}_i\}$ at the observed $g_j(\boldsymbol{\beta})$, $j = 1, \ldots, N$. We require

that $p_{ij} \geq 0$ and $\sum_{j=1}^{N} p_{ij} = 1$. The empirical version of $E\{g(\boldsymbol{\beta}_0)|\boldsymbol{X}\} = 0$ is $\sum_{j=1}^{N} p_{ij} g_j(\boldsymbol{\beta}) = 0$ for some $\boldsymbol{\beta}$. Second, we construct the log-likelihood localized at each subject i, $i = 1, \ldots, N$. With the conditional empirical probabilities p_{ij} available, $j = 1, \ldots, N$, a natural candidate for the localized log-likelihood takes the form of a weighted sum $\sum_{j=1}^{N} w_{ij} \log p_{ij}$, where w_{ij} are certain non-negative weights. Intuitively, $\sum_{j=1}^{N} w_{ij} \log p_{ij}$ serves as an empirical analogue of $\log dF\{g_i(\boldsymbol{\beta})|\boldsymbol{X}_i\}$, and w_{ij} represents how "likely" it is to observe the value $g_j(\boldsymbol{\beta})$ conditional on $\boldsymbol{X}_i$. Third, from the log-likelihood localized at each subject, the overall log-likelihood is given by $\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} \log p_{ij}$. Finally, applying the idea of maximum likelihood estimation, our CEL-AIPW estimator is defined by a constrained optimization:

$$\hat{\boldsymbol{\beta}}_{AIPW} = \arg\max_{\boldsymbol{\beta}} \max_{p_{ij}} \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} \log p_{ij} \quad \text{subject to}$$

$$p_{ij} \geq 0,\ \sum_{j=1}^{N} p_{ij} = 1,\ \sum_{j=1}^{N} p_{ij} g_j(\boldsymbol{\beta}) = 0 \quad (i,\ j = 1, \ldots, N). \tag{2.1}$$

A technique to carry out the localization in the second step is the nonparametric kernel method. Let $\boldsymbol{X}^c$ and $\boldsymbol{X}^d$ denote the continuous and categorical components of $\boldsymbol{X}$, respectively. Then one can calculate w_{ij} as

$$w_{ij} = \frac{\mathcal{K}\left\{(\boldsymbol{X}_i^c - \boldsymbol{X}_j^c)/b_N\right\} \mathcal{I}(\boldsymbol{X}_i^d = \boldsymbol{X}_j^d)}{\sum_{j=1}^{N} \mathcal{K}\left\{(\boldsymbol{X}_i^c - \boldsymbol{X}_j^c)/b_N\right\} \mathcal{I}(\boldsymbol{X}_i^d = \boldsymbol{X}_j^d)},$$

where $\mathcal{K}(\cdot)$ is a multivariate kernel function, b_N is the bandwidth parameter, and $\mathcal{I}(\cdot)$ is the indicator function. Here $\sum_{j=1}^{N} w_{ij} = 1$ for each i.

The AIPW residual $g(\boldsymbol{\beta})$ involves two possibly unknown quantities, $\pi(\boldsymbol{S}, \boldsymbol{X})$ and $E(Y|\boldsymbol{S}, \boldsymbol{X})$, which need to be estimated. When the missingness of Y is due to study design (e.g. two-stage design), $\pi(\boldsymbol{S}, \boldsymbol{X})$ is known. Otherwise, we postulate a parametric model $\pi(\boldsymbol{\alpha}; \boldsymbol{S}, \boldsymbol{X})$, with $\boldsymbol{\alpha}$ being an unknown finite-dimensional parameter whose true value is denoted by $\boldsymbol{\alpha}_0$. One example is the logistic model, $\text{logit}\{\pi(\boldsymbol{\alpha}; \boldsymbol{S}, \boldsymbol{X})\} = \boldsymbol{Z}^T \boldsymbol{\alpha}$, where $\boldsymbol{Z}^T = (\boldsymbol{S}^T, \boldsymbol{X}^T)$. An estimator $\hat{\boldsymbol{\alpha}}$ is given by

$$\hat{\boldsymbol{\alpha}} = \arg\max_{\boldsymbol{\alpha}} \prod_{i=1}^{N} \{\pi(\boldsymbol{\alpha}; \boldsymbol{S}_i, \boldsymbol{X}_i)\}^{R_i} \{1 - \pi(\boldsymbol{\alpha}; \boldsymbol{S}_i, \boldsymbol{X}_i)\}^{1-R_i}. \tag{2.2}$$

To estimate $E(Y|\boldsymbol{S}, \boldsymbol{X})$, we postulate a parametric model $h(\boldsymbol{\gamma}; \boldsymbol{S}, \boldsymbol{X})$, where $h(\cdot)$ is a known link function and $\boldsymbol{\gamma}$ is an unknown finite-dimensional parameter with true value $\boldsymbol{\gamma}_0$. Choices for this parametric model include the generalized linear

model (McCullagh and Nelder (1989)) and the quasi-likelihood model (Wedderburn (1974)). The MAR mechanism (1.2) is equivalent to $R \perp Y|(\boldsymbol{S}, \boldsymbol{X})$, which implies that $E(Y|\boldsymbol{S}, \boldsymbol{X}) = E(Y|\boldsymbol{S}, \boldsymbol{X}, R = 1)$. Therefore, an estimator $\hat{\boldsymbol{\gamma}}$ is obtained based on complete-case analysis by solving

$$\sum_{i=1}^{N} R_i \boldsymbol{Z}_i \dot{h}(\boldsymbol{Z}_i^T \boldsymbol{\gamma}) \left\{ Y_i - h(\boldsymbol{Z}_i^T \boldsymbol{\gamma}) \right\} = \mathbf{0}, \tag{2.3}$$

where $\dot{h}(\cdot)$ is the first order derivative function of $h(\cdot)$.

Given the estimators $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\gamma}}$, the estimated AIPW residual is

$$\begin{aligned} g(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) = & \frac{R}{\pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{S}, \boldsymbol{X})} \left\{ Y - \mu(\boldsymbol{X}^T \boldsymbol{\beta}) \right\} \\ & - \frac{R - \pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{S}, \boldsymbol{X})}{\pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{S}, \boldsymbol{X})} \left\{ h(\hat{\boldsymbol{\gamma}}; \boldsymbol{S}, \boldsymbol{X}) - \mu(\boldsymbol{X}^T \boldsymbol{\beta}) \right\}. \end{aligned}$$

The proposed CEL-AIPW estimator $\hat{\boldsymbol{\beta}}_{AIPW}$ is still defined by (2.1), but with $g_j(\boldsymbol{\beta})$ in the third constraint substituted by $g_j(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})$.

2.2. Numerical implementation

The calculation of $\hat{\boldsymbol{\beta}}_{AIPW}$ pertains to a constrained optimization problem. Using the Lagrange multipliers method, the Lagrangian is given by

$$\mathcal{L} = \sum_{i=1}^{N} \Big(\sum_{j=1}^{N} w_{ij} \log p_{ij} \Big) - \sum_{i=1}^{N} \varpi_i \Big(\sum_{j=1}^{N} p_{ij} - 1 \Big) - \sum_{i=1}^{N} \lambda_i \Big\{ \sum_{j=1}^{N} p_{ij} g_j(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) \Big\},$$

where scalars ϖ_i and λ_i are the Lagrange multipliers associated with the second and the third constraints in (2.1), respectively. With $\partial \mathcal{L} / \partial p_{ij} = 0$ and (2.1), it can be easily shown that, for a fixed $\boldsymbol{\beta}$,

$$p_{ij}(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) = \frac{w_{ij}}{1 + \hat{\lambda}_i(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) g_j(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})}, \quad i, j = 1, \cdots, N, \tag{2.4}$$

where $\hat{\lambda}_i(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})$ satisfies

$$\sum_{j=1}^{N} \frac{w_{ij} g_j(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})}{1 + \hat{\lambda}_i(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) g_j(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})} = 0.$$

It is easy to see that

$$\hat{\lambda}_i(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) = \arg\min_{\lambda_i} \Big[- \sum_{j=1}^{N} w_{ij} \log \left\{ 1 + \lambda_i g_j(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) \right\} \Big]. \tag{2.5}$$

If L denotes the objective function $\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}\log p_{ij}$ in (2.1) and

$$\Lambda_i(\lambda_i,\boldsymbol{\beta},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}}) = -\sum_{j=1}^{N} w_{ij}\log\left\{1+\lambda_i g_j(\boldsymbol{\beta},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})\right\},$$

then L can be rewritten as a function of $\boldsymbol{\beta}$ only:

$$L(\boldsymbol{\beta},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}}) = \sum_{i=1}^{N}\Lambda_i\left\{\hat{\lambda}_i(\boldsymbol{\beta},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}}),\boldsymbol{\beta},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}}\right\} + \sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}\log w_{ij}.$$

Therefore, the CEL-AIPW estimator can be equivalently defined through a nested optimization:

$$\hat{\boldsymbol{\beta}}_{AIPW} = \arg\max_{\boldsymbol{\beta}}\sum_{i=1}^{N}\left\{\min_{\lambda_i}\Lambda_i(\lambda_i,\boldsymbol{\beta},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})\right\}.$$

This definition of $\hat{\boldsymbol{\beta}}_{AIPW}$ essentially suggests a way of numerical implementation. The Newton-Raphson algorithm can be employed for the two optimizations. For convenience, we suppress $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\gamma}}$ in the following. For a fixed $\boldsymbol{\beta}$, given λ_i^{old}, the inner optimization updates λ_i by

$$\lambda_i^{new} = \lambda_i^{old} - \Lambda_{i,\lambda\lambda}^{-1}\left(\lambda_i^{old},\boldsymbol{\beta}\right)\Lambda_{i,\lambda}\left(\lambda_i^{old},\boldsymbol{\beta}\right),\ i=1,\cdots,N,$$

where

$$\Lambda_{i,\lambda}(\lambda_i,\boldsymbol{\beta}) = -\sum_{j=1}^{N} w_{ij}\frac{g_j(\boldsymbol{\beta})}{1+\lambda_i g_j(\boldsymbol{\beta})} \text{ and } \Lambda_{i,\lambda\lambda}(\lambda_i,\boldsymbol{\beta}) = \sum_{j=1}^{N} w_{ij}\frac{g_j(\boldsymbol{\beta})^2}{\{1+\lambda_i g_j(\boldsymbol{\beta})\}^2}.$$

For each i, an initial value can be taken as $\lambda_i=0$, and the converged value gives an estimate of $\hat{\lambda}_i(\boldsymbol{\beta})$. To guarantee the positivity of p_{ij}, the update should be restricted on the legitimate region $\{\lambda_i : 1+\lambda_i g_j(\beta)\geq w_{ij}\}$. Given $\boldsymbol{\beta}^{old}$ and the estimated $\hat{\lambda}_i(\boldsymbol{\beta}^{old})$ from the inner optimization, the outer optimization updates $\boldsymbol{\beta}$ by

$$\boldsymbol{\beta}^{new} = \boldsymbol{\beta}^{old} - \left\{\sum_{i=1}^{N}\boldsymbol{L}_{i,\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\beta}^{old})\right\}^{-1}\left\{\sum_{i=1}^{N}\boldsymbol{L}_{i,\boldsymbol{\beta}}(\boldsymbol{\beta}^{old})\right\},$$

where

$$\boldsymbol{L}_{i,\boldsymbol{\beta}}(\boldsymbol{\beta}) = -\hat{\lambda}_i(\boldsymbol{\beta})\sum_{j=1}^{N} w_{ij}\frac{\boldsymbol{G}_j(\boldsymbol{\beta})^T}{1+\hat{\lambda}_i(\boldsymbol{\beta})g_j(\boldsymbol{\beta})},$$

$$\boldsymbol{L}_{i,\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\beta}) = -\frac{\boldsymbol{\Lambda}_{i,\lambda\boldsymbol{\beta}}^T\left\{\hat{\lambda}_i(\boldsymbol{\beta}),\boldsymbol{\beta}\right\}\boldsymbol{\Lambda}_{i,\boldsymbol{\lambda}\boldsymbol{\beta}}\left\{\hat{\lambda}_i(\boldsymbol{\beta}),\boldsymbol{\beta}\right\}}{\Lambda_{i,\lambda\lambda}\left\{\hat{\lambda}_i(\boldsymbol{\beta}),\boldsymbol{\beta}\right\}},$$

$$\boldsymbol{\Lambda}_{i,\lambda\boldsymbol{\beta}}\left\{\hat{\lambda}_i(\boldsymbol{\beta}),\boldsymbol{\beta}\right\}=\sum_{j=1}^{N}w_{ij}\left[\frac{\hat{\lambda}_i(\boldsymbol{\beta})g_j(\boldsymbol{\beta})\boldsymbol{G}_j(\boldsymbol{\beta})}{\left\{1+\hat{\lambda}_i(\boldsymbol{\beta})g_j(\boldsymbol{\beta})\right\}^2}-\frac{\boldsymbol{G}_j(\boldsymbol{\beta})}{1+\hat{\lambda}_i(\boldsymbol{\beta})g_j(\boldsymbol{\beta})}\right],$$

and $\boldsymbol{G}_j(\boldsymbol{\beta}) = \partial g_j(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$. Iterate the inner and outer optimizations until a certain convergence criterion is satisfied. At convergence, the algorithm produces $\hat{\boldsymbol{\beta}}_{AIPW}$.

It is easy to see that (2.5) is a convex minimization problem. Therefore, for a fixed $\boldsymbol{\beta}$, the estimate from the inner optimization almost always converges to the global minimizer. A proof of this convergence can be given by following Chen, Sitter, and Wu (2002). The outer maximization is more complicated, and the convergence of the Newton-Raphson algorithm may not be guaranteed. See Owen (2001) for some detailed discussion on related issues in the setting of unconditional moment restrictions. Nonetheless, the nested optimization is widely used by many researchers to implement the EL (CEL) method. See, for example, Owen (2001), Kitamura (2007) and Hansen (2014). According to Kitamura (2007), the nested optimization appears to be "the most stable way to compute the EL estimator". In practice, a numerical complication may exist when 0 is not in the convex hull spanned by $\{g_j(\boldsymbol{\beta}) : j = 1,\ldots,N\}$, although this problem disappears when $N \to \infty$ and $\boldsymbol{\beta}$ is in a neighborhood of $\boldsymbol{\beta}_0$ because $E\{g(\boldsymbol{\beta}_0)|\boldsymbol{X}\} = 0$. This numerical complication exists for both the EL and the CEL methods, and more caution may be needed for the latter due to its subject-wise nature. To overcome this numerical issue, we follow the approach suggested by Kitamura (2007) and Hansen (2014). Specifically, in the inner loop of the numerical implementation, we restrict each updated value of λ_i to be in the legitimate region $\{\lambda_i : 1+\lambda_i g_j(\boldsymbol{\beta}) \geq w_{ij},\ j = 1,\ldots,N\}$ and to make $\Lambda_i(\lambda_i,\boldsymbol{\beta},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})$ decrease.

For the CEL method proposed by Kitamura, Tripathi, and Ahn (2004), the bandwidth parameter b_N was selected using the cross-validation criterion suggested by Newey (1993); these papers studied the same problem, namely estimation under conditional moment restrictions. Numerical studies in both papers demonstrated good performance of this criterion. We employ a similar criterion to select b_N. Specifically, take

$$CV(b_N)=\sum_{i=1}^{N}\frac{\left\{g_i(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})^2-\hat{\sigma}_{-i}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})^2\right\}^2}{\hat{\sigma}_{-i}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})^6}, \tag{2.6}$$

where $\hat{\sigma}_{-i}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})^2 = \sum_{j=1}^{N}\hat{w}_{ij}g_j(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})^2$, $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(b_N)$ is the CEL-AIPW estimator obtained with a given b_N, and

$$\hat{w}_{ii}=0,\quad \hat{w}_{ij}=\frac{\mathcal{K}\left\{(\boldsymbol{X}_i^c-\boldsymbol{X}_j^c)/b_N\right\}\mathcal{I}(\boldsymbol{X}_i^d=\boldsymbol{X}_j^d)}{\sum_{j=1,j\neq i}^{N}\mathcal{K}\left\{(\boldsymbol{X}_i^c-\boldsymbol{X}_j^c)/b_N\right\}\mathcal{I}(\boldsymbol{X}_i^d=\boldsymbol{X}_j^d)}\quad \text{for } j\neq i.$$

The optimal bandwidth b_N is chosen as the minimizer of $CV(b_N)$. The intuition behind (2.6) is to minimize the variance of $(\hat{\sigma}^{-2} - \sigma^{-2})g(\boldsymbol{\beta}_0)$ with the sampled data, where $\sigma^2 = \mathrm{Var}\{g(\boldsymbol{\beta}_0) \mid \boldsymbol{X}\}$ and $\hat{\sigma}^2$ is an estimator of σ^2. The linearization of $(\hat{\sigma}^{-2} - \sigma^{-2})g(\boldsymbol{\beta}_0)$ by ignoring the higher-order terms is $\sigma^{-4}(\hat{\sigma}^2 - \sigma^2)g(\boldsymbol{\beta}_0)$ (Newey (1993)). So the resulting variance may be approximated by $\sigma^{-6}(\hat{\sigma}^2 - \sigma^2)^2$. A leave-one-out estimation of this quantity with the sampled data then leads to (2.6). More discussion on bandwidth selection can be found in Newey (1993).

3. Large Sample Properties

For the large sample properties presented in this section, primary consideration is given to the CEL-AIPW estimator, and the corresponding results are summarized in a series of theorems. Regularity conditions and proofs are provided in the Appendix. Properties regarding the CEL-IPW estimator are listed as corollaries. The corresponding regularity conditions and proofs are omitted because they are trivially modified versions of those of the theorems.

Based on the results of White (1982), we can assume that $\hat{\boldsymbol{\alpha}} \xrightarrow{p} \boldsymbol{\alpha}_*$, $\hat{\boldsymbol{\gamma}} \xrightarrow{p} \boldsymbol{\gamma}_*$, $\sqrt{N}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_*) = O_p(1)$, and $\sqrt{N}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_*) = O_p(1)$. Here $\boldsymbol{\alpha}_*$ and $\boldsymbol{\gamma}_*$ are not necessarily equal to $\boldsymbol{\alpha}_0$ and $\boldsymbol{\gamma}_0$, and the corresponding equality is true only when the model for $\pi(\boldsymbol{S}, \boldsymbol{X})$ or $E(Y|\boldsymbol{S}, \boldsymbol{X})$ is correctly specified. Since $\hat{\boldsymbol{\alpha}}$ maximizes the binomial likelihood in (2.2), the asymptotic linear expansion for $\hat{\boldsymbol{\alpha}}$ is given by

$$\sqrt{N}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_*) = -\left[E\left\{\frac{\partial \boldsymbol{\psi}(\boldsymbol{\alpha}_*)}{\partial \boldsymbol{\alpha}}\right\}\right]^{-1} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \boldsymbol{\psi}_i(\boldsymbol{\alpha}_*) + o_p(1), \tag{3.1}$$

where $\boldsymbol{\psi}(\boldsymbol{\alpha}) = \boldsymbol{\psi}(\boldsymbol{\alpha}; \boldsymbol{S}, \boldsymbol{X}, R)$ is the score function. When a logistic regression model is assumed for $\pi(\boldsymbol{S}, \boldsymbol{X})$, we have

$$\boldsymbol{\psi}(\boldsymbol{\alpha}) = \left\{R - \frac{\exp(\boldsymbol{Z}^T\boldsymbol{\alpha})}{1 + \exp(\boldsymbol{Z}^T\boldsymbol{\alpha})}\right\}\boldsymbol{Z}.$$

Similarly, the asymptotic linear expansion for $\hat{\boldsymbol{\gamma}}$ is given by

$$\sqrt{N}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_*) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \boldsymbol{\phi}_i(\boldsymbol{\gamma}_*) + o_p(1), \tag{3.2}$$

where $\boldsymbol{\phi}(\boldsymbol{\gamma}) = \boldsymbol{\phi}(\boldsymbol{\gamma}; Y, \boldsymbol{S}, \boldsymbol{X}, R)$ denotes the influence function. When $\hat{\boldsymbol{\gamma}}$ is the solution to equation (2.3), we have

$$\boldsymbol{\phi}(\boldsymbol{\gamma}) = -\left[E\left\{\frac{\partial \boldsymbol{\zeta}(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}}\right\}\right]^{-1} \boldsymbol{\zeta}(\boldsymbol{\gamma}) \quad \text{with} \quad \boldsymbol{\zeta}(\boldsymbol{\gamma}) = R\boldsymbol{Z}\dot{h}(\boldsymbol{Z}^T\boldsymbol{\gamma})\left\{Y - h(\boldsymbol{Z}^T\boldsymbol{\gamma})\right\}.$$

Let us denote the CEL-IPW estimator by $\hat{\boldsymbol{\beta}}_{IPW}$.

Theorem 1. *For the model defined by* (1.1) *and* (1.2)*, under the assumptions in the Appendix, if either* $\boldsymbol{\alpha}_* = \boldsymbol{\alpha}_0$ *or* $\boldsymbol{\gamma}_* = \boldsymbol{\gamma}_0$*, we have* $\hat{\boldsymbol{\beta}}_{AIPW} \xrightarrow{p} \boldsymbol{\beta}_0$ *as* $N \to \infty$*.*

Corollary 1. *For the model defined by* (1.1) *and* (1.2)*, under assumptions similar to those in the Appendix, if* $\boldsymbol{\alpha}_* = \boldsymbol{\alpha}_0$*, we have* $\hat{\boldsymbol{\beta}}_{IPW} \xrightarrow{p} \boldsymbol{\beta}_0$ *as* $N \to \infty$*.*

From Theorem 1, $\hat{\boldsymbol{\beta}}_{AIPW}$ is doubly robust, in the sense that as long as one of $\pi(\boldsymbol{S}, \boldsymbol{X})$ and $E(Y|\boldsymbol{S}, \boldsymbol{X})$ is correctly modeled, $\hat{\boldsymbol{\beta}}_{AIPW}$ is a consistent estimator of $\boldsymbol{\beta}_0$.

To describe the asymptotic distribution of $\hat{\boldsymbol{\beta}}_{AIPW}$, let

$$
\begin{aligned}
V_{AIPW}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) &= E\left\{g(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma})^2 | \boldsymbol{X}\right\}, \\
\boldsymbol{G}_{\boldsymbol{\gamma}}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) &= E\left\{\frac{\partial g(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}} \middle| \boldsymbol{X}\right\}, \\
\boldsymbol{Q}_{AIPW}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) &= \frac{\partial \mu(\boldsymbol{X}^T\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T} V_{AIPW}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma})^{-1} g(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}), \\
\boldsymbol{I}_{AIPW}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) &= E\left\{\boldsymbol{Q}_{AIPW}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) \boldsymbol{Q}_{AIPW}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma})^T\right\}, \\
\boldsymbol{V}_{\boldsymbol{\alpha}, AIPW}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) &= \mathrm{Var}\left[\mathrm{Resid}\left\{\boldsymbol{Q}_{AIPW}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}), \boldsymbol{\psi}(\boldsymbol{\alpha})\right\}\right], \\
\boldsymbol{V}_{\boldsymbol{\gamma}}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) &= \mathrm{Var}\Bigg[\boldsymbol{Q}_{AIPW}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) \\
&\qquad + E\left\{\frac{\partial \mu(\boldsymbol{X}^T\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T} V_{AIPW}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma})^{-1} \boldsymbol{G}_{\boldsymbol{\gamma}}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma})\right\} \boldsymbol{\phi}(\boldsymbol{\gamma})\Bigg],
\end{aligned}
$$

where for any matrices $\boldsymbol{A}$ and $\boldsymbol{B}$,

$$
\mathrm{Resid}(\boldsymbol{A}, \boldsymbol{B}) = \boldsymbol{A} - E(\boldsymbol{A}\boldsymbol{B}^T)\{E(\boldsymbol{B}\boldsymbol{B}^T)\}^{-1}\boldsymbol{B}.
$$

Theorem 2. *For the model defined by* (1.1) *and* (1.2)*, under the assumptions in the Appendix, we have* $\sqrt{N}(\hat{\boldsymbol{\beta}}_{AIPW} - \boldsymbol{\beta}_0) \xrightarrow{d} \mathcal{N}\{\boldsymbol{0}, \boldsymbol{J}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_*, \boldsymbol{\gamma}_*)^{-1}\}$*, where*

$$
\boldsymbol{J}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) = \begin{cases}
\boldsymbol{I}_{AIPW}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) \boldsymbol{V}_{\boldsymbol{\alpha}, AIPW}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma})^{-1} \boldsymbol{I}_{AIPW}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) & \text{if } \boldsymbol{\alpha}_* = \boldsymbol{\alpha}_0, \\
\boldsymbol{I}_{AIPW}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) \boldsymbol{V}_{\boldsymbol{\gamma}}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma})^{-1} \boldsymbol{I}_{AIPW}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) & \text{if } \boldsymbol{\gamma}_* = \boldsymbol{\gamma}_0, \\
\boldsymbol{I}_{AIPW}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) & \text{if } \boldsymbol{\alpha}_* = \boldsymbol{\alpha}_0 \text{ and } \boldsymbol{\gamma}_* = \boldsymbol{\gamma}_0.
\end{cases}
$$

In the case that the data are collected based on a two-stage design, $\boldsymbol{\alpha}_0$ is known. When the known $\boldsymbol{\alpha}_0$ is used instead of $\hat{\boldsymbol{\alpha}}$, following the same arguments as that in the proof of Theorem 2, the asymptotic variance of $\hat{\boldsymbol{\beta}}_{AIPW}$ has the same structure as $\boldsymbol{J}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0, \boldsymbol{\gamma}_*)$, but with $\boldsymbol{V}_{\boldsymbol{\alpha}, AIPW}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0, \boldsymbol{\gamma}_*)$ in the middle replaced by $\mathrm{Var}\{\boldsymbol{Q}_{AIPW}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0, \boldsymbol{\gamma}_*)\} = \boldsymbol{I}_{AIPW}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0, \boldsymbol{\gamma}_*)$. This new asymptotic variance is no smaller than the old one, in the sense that the corresponding difference between the two asymptotic variance matrices is nonnegative-definite: $\boldsymbol{V}_{\boldsymbol{\alpha}, AIPW}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0, \boldsymbol{\gamma}_*)$ is the variance of the residual of the regression of

$\boldsymbol{Q}_{AIPW}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0, \boldsymbol{\gamma}_*)$ on $\boldsymbol{\psi}(\boldsymbol{\alpha}_0)$. Therefore, even if $\boldsymbol{\alpha}_0$ is known in practice, using the estimator $\hat{\boldsymbol{\alpha}}$ based on a correctly specified model for $\pi(\boldsymbol{S}, \boldsymbol{X})$ has the advantage of potential efficiency gain for the CEL-AIPW estimator. This counterintuitive phenomenon is well known in the literature of parametric regression with missing data (e.g., Robins, Rotnitzky, and Zhao (1995); Rotnitzky and Robins (1995)). For nonparametric regression with missing data, this does not hold any more (Wang, Rotnitzky, and Lin (2010)). When $E(Y|\boldsymbol{S}, \boldsymbol{X})$ is correctly modeled in addition to the known $\boldsymbol{\alpha}_0$, using $\hat{\boldsymbol{\alpha}}$ or $\boldsymbol{\alpha}_0$ will make no difference asymptotically, as in both cases the asymptotic variance of $\hat{\boldsymbol{\beta}}_{AIPW}$ is $\boldsymbol{I}_{AIPW}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0, \boldsymbol{\gamma}_0)^{-1}$. From Chen and Breslow (2004) and Yu and Nan (2006), $\boldsymbol{I}_{AIPW}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0, \boldsymbol{\gamma}_0)^{-1}$ is the semiparametric efficiency bound under the model defined by (1.1) and (1.2). Therefore, the CEL-AIPW estimator attains the semiparametric efficiency bound when both $\pi(\boldsymbol{S}, \boldsymbol{X})$ and $E(Y|\boldsymbol{S}, \boldsymbol{X})$ are correctly modeled, and thus is locally efficient.

To describe the asymptotic distribution of $\hat{\boldsymbol{\beta}}_{IPW}$, let

$$
\begin{aligned}
V_{IPW}(\boldsymbol{\beta}, \boldsymbol{\alpha}) &= E\left\{ f(\boldsymbol{\beta}, \boldsymbol{\alpha})^2 | \boldsymbol{X} \right\}, \\
\boldsymbol{Q}_{IPW}(\boldsymbol{\beta}, \boldsymbol{\alpha}) &= \frac{\partial \mu(\boldsymbol{X}^T\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T} V_{IPW}(\boldsymbol{\beta}, \boldsymbol{\alpha})^{-1} f(\boldsymbol{\beta}, \boldsymbol{\alpha}), \\
\boldsymbol{I}_{IPW}(\boldsymbol{\beta}, \boldsymbol{\alpha}) &= E\left\{ \boldsymbol{Q}_{IPW}(\boldsymbol{\beta}, \boldsymbol{\alpha}) \boldsymbol{Q}_{IPW}(\boldsymbol{\beta}, \boldsymbol{\alpha})^T \right\}, \\
\boldsymbol{V}_{\boldsymbol{\alpha}, IPW}(\boldsymbol{\beta}, \boldsymbol{\alpha}) &= \mathrm{Var}\left[\mathrm{Resid}\{ \boldsymbol{Q}_{IPW}(\boldsymbol{\beta}, \boldsymbol{\alpha}), \boldsymbol{\psi}(\boldsymbol{\alpha}) \} \right].
\end{aligned}
$$

Corollary 2. *For the model defined by* (1.1) *and* (1.2)*, under assumptions similar to those in the Appendix, if* $\boldsymbol{\alpha}_* = \boldsymbol{\alpha}_0$*, we have*

$$
\sqrt{N}(\hat{\boldsymbol{\beta}}_{IPW} - \boldsymbol{\beta}_0) \xrightarrow{d} \mathcal{N}\left\{ \boldsymbol{0}, \boldsymbol{I}_{IPW}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0)^{-1} \boldsymbol{V}_{\boldsymbol{\alpha}, IPW}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0) \boldsymbol{I}_{IPW}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0)^{-1} \right\}.
$$

If $\boldsymbol{\alpha}_0$ is known by design and is used instead of $\hat{\boldsymbol{\alpha}}$, the asymptotic variance of $\hat{\boldsymbol{\beta}}_{IPW}$ has the same structure as above, but with $\boldsymbol{V}_{\boldsymbol{\alpha}, IPW}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0)$ replaced by $\mathrm{Var}\{\boldsymbol{Q}_{IPW}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0)\} = \boldsymbol{I}_{IPW}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0)$. The new asymptotic variance is no smaller than that above. So using the estimator $\hat{\boldsymbol{\alpha}}$ is preferred for the CEL-IPW estimator as well even if $\boldsymbol{\alpha}_0$ is known.

Our Theorem 3 provides a consistent estimator of the asymptotic variance of the CEL-AIPW estimator of Theorem 2. Let

$$
\hat{V}_{i, AIPW}(\boldsymbol{\beta}) = \sum_{j=1}^{N} p_{ij}(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) g_j(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})^2,
$$

$$
\hat{\boldsymbol{G}}_{i, \boldsymbol{\alpha}}(\boldsymbol{\beta}) = \sum_{j=1}^{N} p_{ij}(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) \frac{\partial g_j(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})}{\partial \boldsymbol{\alpha}}, \quad \hat{\boldsymbol{G}}_{i, \boldsymbol{\gamma}}(\boldsymbol{\beta}) = \sum_{j=1}^{N} p_{ij}(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) \frac{\partial g_j(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})}{\partial \boldsymbol{\gamma}},
$$

$$
\hat{\boldsymbol{Q}}_{i, AIPW}(\boldsymbol{\beta}) = \frac{\partial \mu(\boldsymbol{X}_i^T\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T} \hat{V}_{i, AIPW}(\boldsymbol{\beta})^{-1} g_i(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}),
$$

$$\hat{\boldsymbol{I}}_{AIPW}(\boldsymbol{\beta}) = \frac{1}{N}\sum_{i=1}^{N} \hat{\boldsymbol{Q}}_{i,AIPW}(\boldsymbol{\beta})\hat{\boldsymbol{Q}}_{i,AIPW}(\boldsymbol{\beta})^T,$$

$$\hat{\boldsymbol{m}}_i(\boldsymbol{\beta}) = \hat{\boldsymbol{Q}}_{i,AIPW}(\boldsymbol{\beta}) - \left\{\frac{1}{N}\sum_{i=1}^{N} \frac{\partial \mu(\boldsymbol{X}_i^T\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\hat{V}_{i,AIPW}(\boldsymbol{\beta})^{-1}\hat{\boldsymbol{G}}_{i,\boldsymbol{\alpha}}(\boldsymbol{\beta})\right\}$$
$$\times\left\{\frac{1}{N}\sum_{i=1}^{N} \frac{\partial \boldsymbol{\psi}_i(\hat{\boldsymbol{\alpha}})}{\partial \boldsymbol{\alpha}}\right\}^{-1}\boldsymbol{\psi}_i(\hat{\boldsymbol{\alpha}})$$
$$+\left\{\frac{1}{N}\sum_{i=1}^{N} \frac{\partial \mu(\boldsymbol{X}_i^T\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\hat{V}_{i,AIPW}(\boldsymbol{\beta})^{-1}\hat{\boldsymbol{G}}_{i,\boldsymbol{\gamma}}(\boldsymbol{\beta})\right\}\boldsymbol{\phi}_i(\hat{\boldsymbol{\gamma}}).$$

Theorem 3. *Under the assumptions in the Appendix, we have that*

$$\left[\hat{\boldsymbol{I}}_{AIPW}(\boldsymbol{\beta})\left\{\frac{1}{N}\sum_{i=1}^{N}\hat{\boldsymbol{m}}_i(\boldsymbol{\beta})\hat{\boldsymbol{m}}_i(\boldsymbol{\beta})^T\right\}^{-1}\hat{\boldsymbol{I}}_{AIPW}(\boldsymbol{\beta})\right]\Bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_{AIPW}} \xrightarrow{p} \boldsymbol{J}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_*, \boldsymbol{\gamma}_*).$$

To consistently estimate the asymptotic variance of the CEL-IPW estimator given in Corollary 2, take

$$\hat{V}_{i,IPW}(\boldsymbol{\beta}) = \sum_{j=1}^{N} p_{ij}(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}) f_j(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}})^2,$$

$$\hat{\boldsymbol{Q}}_{i,IPW}(\boldsymbol{\beta}) = \frac{\partial \mu(\boldsymbol{X}_i^T\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\hat{V}_{i,IPW}(\boldsymbol{\beta})^{-1} f_i(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}),$$

$$\hat{\boldsymbol{I}}_{IPW}(\boldsymbol{\beta}) = \frac{1}{N}\sum_{i=1}^{N} \hat{\boldsymbol{Q}}_{i,IPW}(\boldsymbol{\beta})\hat{\boldsymbol{Q}}_{i,IPW}(\boldsymbol{\beta})^T,$$

$$\hat{\boldsymbol{t}}_i(\boldsymbol{\beta}) = \hat{\boldsymbol{Q}}_{i,IPW}(\boldsymbol{\beta}) - \left\{\frac{1}{N}\sum_{i=1}^{N} \hat{\boldsymbol{Q}}_{i,IPW}(\boldsymbol{\beta})\boldsymbol{\psi}_i(\hat{\boldsymbol{\alpha}})^T\right\}\left\{\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\psi}_i(\hat{\boldsymbol{\alpha}})\boldsymbol{\psi}_i(\hat{\boldsymbol{\alpha}})^T\right\}^{-1}\boldsymbol{\psi}_i(\hat{\boldsymbol{\alpha}}).$$

Here $p_{ij}(\boldsymbol{\beta}, \boldsymbol{\alpha})$ is defined similarly to that in (2.4), but is based on the IPW residual $f(\boldsymbol{\beta})$.

Corollary 3. *Under assumptions similar to those in the Appendix, if* $\boldsymbol{\alpha}_* = \boldsymbol{\alpha}_0$,

$$\left[\hat{\boldsymbol{I}}_{IPW}(\boldsymbol{\beta})^{-1}\left\{\frac{1}{N}\sum_{i=1}^{N}\hat{\boldsymbol{t}}_i(\boldsymbol{\beta})\hat{\boldsymbol{t}}_i(\boldsymbol{\beta})^T\right\}\hat{\boldsymbol{I}}_{IPW}(\boldsymbol{\beta})^{-1}\right]\Bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_{IPW}}$$
$$\xrightarrow{p} \boldsymbol{I}_{IPW}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0)^{-1}\boldsymbol{V}_{\boldsymbol{\alpha},IPW}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0)\boldsymbol{I}_{IPW}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0)^{-1}.$$

4. Simulation Experiments

We evaluated the finite sample performance of the proposed CEL estimators using simulation experiments. The simulation model contained two covariates,

$X_1 \sim \mathcal{N}(0, 2^2)$ and $X_2 \sim$ Bernoulli(0.5), as well as an auxiliary variable generated by $S = 1 + X_1 + X_2 + \epsilon_S$ with $\epsilon_S \sim \mathcal{N}(0, 2^2)$. The outcome of interest Y was generated as $Y = 1 + S + 0.6X_1 + 2X_2 + \epsilon_Y$, where $\epsilon_Y \sim \mathcal{N}(0, \sigma_Y^2)$ with covariate-dependent variance $\sigma_Y^2 = \exp(0.2 + 0.4S + 0.4X_1)$. A straightforward calculation shows that the conditional distribution of $Y|\boldsymbol{X}$ is Normal, with mean $E(Y|\boldsymbol{X}) = 2 + 1.6X_1 + 3X_2$ and variance $\mathrm{Var}(Y|\boldsymbol{X}) = 4 + \exp(0.92 + 0.8X_1 + 0.4X_2)$. The missingness mechanism was set to be logit $\{\pi(S, \boldsymbol{X})\} = 0.5 - 0.2S + 0.6X_1 - 0.2X_2$, under which approximately 50% of subjects had missing Y in our generated data. Therefore, the true parameter values used in our simulation were $\boldsymbol{\beta}_0 = (\beta_1, \beta_2, \beta_3)^T = (2, 1.6, 3)^T$, $\boldsymbol{\alpha}_0 = (0.5, -0.2, 0.6, -0.2)^T$, and $\boldsymbol{\gamma}_0 = (1, 1, 0.6, 2)^T$.

We compared the proposed CEL estimators with the IPW, AIPW, HAN, RLSR, CLQ, and QZL estimators under three scenarios: (i) only $\pi(S, \boldsymbol{X})$ is correctly modeled; (ii) only $E(Y|S, \boldsymbol{X})$ is correctly modeled; and (iii) both are correctly modeled. For the first scenario, $E(Y|S, \boldsymbol{X})$ is incorrectly modeled as $E(Y|S, \boldsymbol{X}) = \gamma_1 + \gamma_2 X_1$, and for the second scenario, $\pi(S, \boldsymbol{X})$ is incorrectly modeled as logit $\{\pi(S, \boldsymbol{X})\} = \alpha_1 + \alpha_2 S + \alpha_3 X_2$.

In each scenario, the six competitors were derived based on the estimating function $\boldsymbol{U}(\boldsymbol{\beta}; Y, \boldsymbol{X}) = \boldsymbol{X}\mathrm{Var}(Y|\boldsymbol{X})^{-1}(Y - \boldsymbol{X}^T\boldsymbol{\beta})$, where $\mathrm{Var}(Y|\boldsymbol{X})$ is modeled in two ways. The first assumes that $\mathrm{Var}(Y|\boldsymbol{X})$ is a constant V_1, and the second uses the true model $V_2 = \theta_1 + \exp(\theta_2 + \theta_3 X_1 + \theta_4 X_2)$, where $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3, \theta_4)^T$ is the vector of unknown nuisance parameters. To estimate $\boldsymbol{\theta}$, we first calculated the residual $\tilde{\epsilon} = Y - \boldsymbol{X}^T\tilde{\boldsymbol{\beta}}$ for subjects whose outcome was observed, where $\tilde{\boldsymbol{\beta}}$ is the IPW estimator based on $\tilde{\boldsymbol{U}}(\boldsymbol{\beta}; Y, \boldsymbol{X}) = \boldsymbol{X}(Y - \boldsymbol{X}^T\boldsymbol{\beta})$ with weight $R/\pi(\boldsymbol{\alpha}_0; S, \boldsymbol{X})$. The true value $\pi(S, \boldsymbol{X})$ was used here to ensure that $\tilde{\boldsymbol{\beta}}$ be a consistent estimator of $\boldsymbol{\beta}_0$. Generally, however, the six competitors cannot take this advantage, as $\pi(S, \boldsymbol{X})$ is usually unknown. We then minimized the least square objective function $\left[\log \tilde{\epsilon}^2 - \log\{\theta_1 + \exp(\theta_2 + \theta_3 X_1 + \theta_4 X_2)\}\right]^2$ with respect to $\boldsymbol{\theta}$ over all subjects whose residual had been calculated. Here the log transformation was used to ensure that the estimated value of $\mathrm{Var}(Y|\boldsymbol{X})$ be always positive.

To establish the benchmark for comparisons, we also include an estimator based on the full data. This estimator, IDEAL, is based on $\boldsymbol{U}(\boldsymbol{\beta}; Y, \boldsymbol{X}) = \boldsymbol{X}V_2^{-1}(Y - \boldsymbol{X}^T\boldsymbol{\beta})$, where V_2 is estimated following a procedure similar to before. We took sample sizes $N = 200$ and $N = 800$. The results are summarized in Tables 1 and 2, respectively, using 500 replications. For the proposed CEL estimators, the bandwidth was selected by minimizing the cross validation criterion (2.6).

When only $\pi(S, \boldsymbol{X})$ was correctly modeled, we have the following summary points. (i) All estimators under our comparison had ignorable bias. (ii) The CEL-IPW estimator had smaller total mean square error (TMSE) than the IPW estimator, even when the latter was based on the true model for $\mathrm{Var}(Y|\boldsymbol{X})$. (iii)

Table 1. Comparison of estimators ($N = 200$). For each estimator, we report the bias, the empirical standard error (the number in ( )), and the mean square error (the number in [ ]). For the CEL-AIPW and the CEL-IPW estimators, the number in { } is the mean of estimated standard error based on Theorem 3 and Corollary 3, respectively.

	correct $\pi(\boldsymbol{S}, \boldsymbol{X})$			correct $E(Y\|\boldsymbol{S}, \boldsymbol{X})$			both models correct		
estimator	β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3
IDEAL	-0.01	-0.02	-0.04	-0.01	-0.02	-0.04	-0.01	-0.02	-0.04
	(0.29)	(0.13)	(0.41)	(0.29)	(0.13)	(0.41)	(0.29)	(0.13)	(0.41)
	[0.09]	[0.02]	[0.17]	[0.09]	[0.02]	[0.17]	[0.09]	[0.02]	[0.17]
CEL-IPW	-0.01	0.00	-0.04	-0.49	0.07	-0.12	-0.01	0.00	-0.04
	(0.40)	(0.22)	(0.68)	(0.39)	(0.21)	(0.63)	(0.40)	(0.22)	(0.68)
	[0.16]	[0.05]	[0.47]	[0.39]	[0.05]	[0.41]	[0.16]	[0.05]	[0.47]
	{0.40}	{0.19}	{0.63}	{0.38}	{0.17}	{0.56}	{0.40}	{0.19}	{0.63}
IPW-V_1	-0.01	0.00	-0.05	-0.49	0.06	-0.12	-0.01	0.00	-0.05
	(0.44)	(0.28)	(0.76)	(0.44)	(0.32)	(0.76)	(0.44)	(0.28)	(0.76)
	[0.19]	[0.08]	[0.57]	[0.43]	[0.11]	[0.60]	[0.19]	[0.08]	[0.57]
IPW-V_2	-0.01	0.00	-0.05	-0.49	0.05	-0.13	-0.01	0.00	-0.05
	(0.42)	(0.23)	(0.70)	(0.39)	(0.21)	(0.60)	(0.42)	(0.23)	(0.70)
	[0.17]	[0.05]	[0.50]	[0.39]	[0.05]	[0.38]	[0.17]	[0.05]	[0.50]
CEL-AIPW	0.01	0.01	-0.07	-0.01	-0.01	-0.02	0.00	-0.02	-0.03
	(0.39)	(0.23)	(0.71)	(0.37)	(0.18)	(0.53)	(0.36)	(0.18)	(0.57)
	[0.15]	[0.05]	[0.51]	[0.14]	[0.03]	[0.29]	[0.13]	[0.03]	[0.33]
	{0.41}	{0.22}	{0.71}	{0.36}	{0.15}	{0.49}	{0.36}	{0.16}	{0.53}
AIPW-V_1	0.00	0.00	-0.07	-0.01	-0.01	-0.03	0.00	-0.02	-0.04
	(0.44)	(0.30)	(0.81)	(0.42)	(0.31)	(0.72)	(0.41)	(0.26)	(0.67)
	[0.19]	[0.09]	[0.66]	[0.18]	[0.10]	[0.52]	[0.17]	[0.07]	[0.45]
AIPW-V_2	-0.02	-0.02	-0.01	-0.04	-0.02	0.03	-0.05	-0.02	0.03
	(1.10)	(0.82)	(2.25)	(0.54)	(0.27)	(0.96)	(0.59)	(0.33)	(1.17)
	[1.20]	[0.67]	[5.07]	[0.29]	[0.07]	[0.92]	[0.35]	[0.11]	[1.38]
HAN-V_1	-0.01	0.00	-0.05	-0.01	-0.02	-0.04	0.00	-0.02	-0.03
	(0.44)	(0.28)	(0.74)	(0.44)	(0.34)	(0.82)	(0.41)	(0.27)	(0.68)
	[0.19]	[0.08]	[0.55]	[0.19]	[0.12]	[0.67]	[0.16]	[0.07]	[0.46]
HAN-V_2	0.01	0.02	-0.09	-0.05	-0.02	0.02	-0.04	-0.01	0.03
	(0.44)	(0.30)	(0.77)	(0.54)	(0.24)	(0.97)	(0.56)	(0.33)	(1.13)
	[0.20]	[0.09]	[0.60]	[0.29]	[0.06]	[0.94]	[0.32]	[0.11]	[1.27]
RLSR-V_1	-	-	-	0.00	-0.02	-0.03	0.00	-0.02	-0.02
	-	-	-	(0.45)	(0.31)	(0.75)	(0.42)	(0.25)	(0.63)
	-	-	-	[0.20]	[0.10]	[0.56]	[0.18]	[0.06]	[0.40]
RLSR-V_2	-	-	-	-0.05	-0.02	0.05	-0.05	-0.02	0.05
	-	-	-	(0.58)	(0.25)	(1.12)	(0.57)	(0.24)	(1.11)
	-	-	-	[0.34]	[0.06]	[1.26]	[0.33]	[0.06]	[1.23]
CLQ-V_1	-0.08	0.04	-0.07	-0.03	-0.02	0.01	0.01	-0.01	-0.04
	(1.04)	(0.52)	(1.47)	(1.28)	(0.44)	(1.82)	(0.43)	(0.28)	(0.72)
	[1.09]	[0.27]	[2.15]	[1.63]	[0.19]	[3.31]	[0.19]	[0.08]	[0.53]
CLQ-V_2	-0.02	0.02	-0.11	-0.11	-0.04	0.03	-0.05	-0.01	0.02
	(0.64)	(0.35)	(1.22)	(1.49)	(0.38)	(2.11)	(0.73)	(0.84)	(1.63)
	[0.41]	[0.12]	[1.49]	[2.22]	[0.15]	[4.44]	[0.54]	[0.70]	[2.65]
QZL-V_1	0.01	0.04	-0.08	-0.31	-0.45	-0.23	-0.03	-0.03	-0.04
	(0.45)	(0.32)	(0.77)	(0.46)	(0.31)	(0.83)	(0.42)	(0.30)	(0.72)
	[0.20]	[0.11]	[0.60]	[0.31]	[0.30]	[0.74]	[0.18]	[0.09]	[0.51]
QZL-V_2	-0.01	0.04	-0.06	-0.20	-0.21	-0.05	-0.03	-0.02	-0.04
	(0.43)	(0.26)	(0.70)	(0.59)	(0.32)	(0.89)	(0.36)	(0.20)	(0.58)
	[0.18]	[0.07]	[0.50]	[0.38]	[0.15]	[0.79]	[0.13]	[0.04]	[0.33]

Table 2. Comparison of estimators (N=800). For each estimator, we report the bias, the empirical standard error (the number in ( )), and the mean square error (the number in [ ]). For the CEL-AIPW and the CEL-IPW estimators, the number in { } is the mean of estimated standard error based on Theorem 3 and Corollary 3, respectively.

	correct $\pi(\boldsymbol{S}, \boldsymbol{X})$			correct $E(Y\|\boldsymbol{S}, \boldsymbol{X})$			both models correct		
estimator	β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3
IDEAL	0.01	0.00	-0.01	0.01	0.00	-0.01	0.01	0.00	-0.01
	(0.15)	(0.06)	(0.19)	(0.15)	(0.06)	(0.19)	(0.15)	(0.06)	(0.19)
	[0.02]	[0.00]	[0.03]	[0.02]	[0.00]	[0.03]	[0.02]	[0.00]	[0.03]
CEL-IPW	0.03	0.00	-0.03	-0.46	0.07	-0.12	0.03	0.00	-0.03
	(0.21)	(0.11)	(0.32)	(0.19)	(0.10)	(0.29)	(0.21)	(0.11)	(0.32)
	[0.04]	[0.01]	[0.10]	[0.24]	[0.02]	[0.10]	[0.04]	[0.01]	[0.10]
	{0.20}	{0.10}	{0.33}	{0.19}	{0.09}	{0.28}	{0.20}	{0.10}	{0.33}
IPW-V_1	0.03	0.01	-0.04	-0.46	0.08	-0.12	0.03	0.01	-0.04
	(0.23)	(0.14)	(0.37)	(0.22)	(0.16)	(0.37)	(0.23)	(0.14)	(0.37)
	[0.05]	[0.02]	[0.14]	[0.26]	[0.03]	[0.15]	[0.05]	[0.02]	[0.14]
IPW-V_2	0.03	0.00	-0.03	-0.46	0.07	-0.12	0.03	0.00	-0.03
	(0.22)	(0.12)	(0.35)	(0.19)	(0.10)	(0.29)	(0.22)	(0.12)	(0.35)
	[0.05]	[0.01]	[0.12]	[0.25]	[0.01]	[0.10]	[0.05]	[0.01]	[0.12]
CEL-AIPW	0.03	0.01	-0.04	0.02	0.00	-0.02	0.03	0.00	-0.03
	(0.20)	(0.11)	(0.33)	(0.18)	(0.09)	(0.24)	(0.18)	(0.09)	(0.26)
	[0.04]	[0.01]	[0.11]	[0.03]	[0.01]	[0.06]	[0.03]	[0.01]	[0.07]
	{0.20}	{0.11}	{0.35}	{0.18}	{0.07}	{0.23}	{0.18}	{0.08}	{0.25}
AIPW-V_1	0.03	0.01	-0.04	0.03	0.01	-0.03	0.03	0.01	-0.03
	(0.22)	(0.15)	(0.38)	(0.21)	(0.16)	(0.35)	(0.21)	(0.13)	(0.33)
	[0.05]	[0.02]	[0.15]	[0.05]	[0.02]	[0.12]	[0.04]	[0.02]	[0.11]
AIPW-V_2	0.00	-0.01	-0.01	0.02	0.00	-0.01	0.02	-0.01	-0.03
	(0.45)	(0.33)	(0.45)	(0.19)	(0.10)	(0.28)	(0.22)	(0.13)	(0.43)
	[0.21]	[0.11]	[0.20]	[0.04]	[0.01]	[0.08]	[0.05]	[0.02]	[0.19]
HAN-V_1	0.03	0.01	-0.04	0.03	0.01	-0.03	0.03	0.01	-0.03
	(0.22)	(0.14)	(0.36)	(0.22)	(0.18)	(0.40)	(0.21)	(0.13)	(0.33)
	[0.05]	[0.02]	[0.13]	[0.05]	[0.03]	[0.16]	[0.04]	[0.02]	[0.11]
HAN-V_2	0.03	0.01	-0.04	0.01	0.00	-0.01	0.01	0.00	-0.01
	(0.23)	(0.12)	(0.37)	(0.20)	(0.11)	(0.30)	(0.19)	(0.09)	(0.27)
	[0.05]	[0.01]	[0.14]	[0.04]	[0.01]	[0.09]	[0.04]	[0.01]	[0.08]
RLSR-V_1	-	-	-	0.03	0.01	-0.03	0.03	0.01	-0.02
	-	-	-	(0.22)	(0.16)	(0.36)	(0.21)	(0.13)	(0.30)
	-	-	-	[0.05]	[0.03]	[0.13]	[0.05]	[0.02]	[0.09]
RLSR-V_2	-	-	-	0.01	0.00	0.00	0.00	0.00	0.01
	-	-	-	(0.20)	(0.08)	(0.28)	(0.30)	(0.08)	(0.29)
	-	-	-	[0.04]	[0.01]	[0.08]	[0.09]	[0.01]	[0.09]
CLQ-V_1	0.03	0.00	-0.01	0.02	0.01	-0.02	0.03	0.01	-0.03
	(0.31)	(0.34)	(0.78)	(0.27)	(0.18)	(0.46)	(0.21)	(0.13)	(0.35)
	[0.09]	[0.11]	[0.62]	[0.07]	[0.03]	[0.21]	[0.05]	[0.02]	[0.12]
CLQ-V_2	0.02	0.00	-0.04	-0.06	-0.03	0.04	0.00	0.01	0.01
	(0.24)	(0.18)	(0.38)	(0.39)	(0.21)	(0.51)	(0.33)	(0.18)	(0.64)
	[0.06]	[0.03]	[0.15]	[0.15]	[0.05]	[0.26]	[0.11]	[0.03]	[0.41]
QZL-V_1	0.03	0.02	-0.04	-0.31	-0.50	-0.26	0.02	0.00	-0.03
	(0.22)	(0.14)	(0.36)	(0.25)	(0.17)	(0.44)	(0.21)	(0.14)	(0.33)
	[0.05]	[0.02]	[0.13]	[0.16]	[0.28]	[0.26]	[0.04]	[0.02]	[0.11]
QZL-V_2	0.03	0.02	-0.03	-0.12	-0.13	-0.01	0.01	0.00	-0.02
	(0.23)	(0.11)	(0.35)	(0.27)	(0.19)	(0.34)	(0.18)	(0.09)	(0.25)
	[0.05]	[0.01]	[0.12]	[0.08]	[0.05]	[0.12]	[0.03]	[0.01]	[0.06]

The CEL-AIPW estimator clearly had smaller TMSE than the AIPW and the CLQ estimators, and had smaller or similar TMSE compared to the HAN and the QZL estimators. Since the implementation of the RLSR estimator requires that the dimension of $\boldsymbol{\gamma}$ be no smaller than that of $\boldsymbol{\beta}$, not satisfied in the current scenario, the results are not reported.

When only $E(Y|S, \boldsymbol{X})$ was correctly modeled, we summarize the following points. (i) The IPW, CEL-IPW, and QZL estimators were clearly biased. In contrast, the AIPW, CEL-AIPW, HAN and RLSR estimators had ignorable biases; they are still consistent due to the double robustness property. The CLQ estimator seemed unbiased, although its double robustness property was not established in Chen, Leung, and Qin (2008). (ii) Compared to the other three doubly robust estimators, the CEL-AIPW estimator had apparent superior performance, judging from its smaller TMSE. When $N = 800$ and the true model for $\mathrm{Var}(Y|\boldsymbol{X})$ was used by the AIPW, HAN and RLSR estimators, the CEL-AIPW estimator still had smaller or similar TMSE.

When both $\pi(S, \boldsymbol{X})$ and $E(Y|S, \boldsymbol{X})$ were correctly modeled, the CEL-AIPW estimator was the only one that attained the semiparametric efficiency bound among the estimators under our comparison; from the efficient influence function (1.4), an estimation procedure needs to explicitly or implicitly estimate $\mathrm{Var}\{g(\boldsymbol{\beta})|\boldsymbol{X}\}$ to achieve the efficiency bound, but the estimators under our comparison modeled $\mathrm{Var}(Y|\boldsymbol{X})$ instead. Based on the numerical results, we have the following summary points. (i) In most cases, the TMSE of the CEL-AIPW estimator was significantly smaller compared to the other estimators. (ii) The HAN and QZL estimators based on the true model for $\mathrm{Var}(Y|\boldsymbol{X})$ had TMSE similar to that of the CEL-AIPW estimator when $N = 800$, although the former two do not attain the semiparametric efficiency bound. However, such an outstanding performance may not be achieved in practice, as the true model for $\mathrm{Var}(Y|\boldsymbol{X})$ is usually unknown. (iii) The CEL-AIPW estimator had smaller TMSE than the CEL-IPW estimator, since correctly modeling $E(Y|S, \boldsymbol{X})$ improves efficiency.

The superior performance of the CEL-AIPW estimator does not require one to model any second moments of the data. For some existing estimators, not only the efficiency, but also the numerical performance could be affected by modeling the second moments and estimating the unknown nuisance parameters. In the above simulation, with $N = 200$, the numerical performance of the AIPW, RLSR and CLQ estimators were worse when the correct model for $\mathrm{Var}(Y|\boldsymbol{X})$ was used. This observation does not contradict the fact that using the correct model for $\mathrm{Var}(Y|\boldsymbol{X})$ improves efficiency (when N goes to infinity), but rather reveals the numerical sensitivity of these three estimators to the estimation of nuisance parameters. For the special nonlinear structure of $\mathrm{Var}(Y|\boldsymbol{X})$ in our setting, obtaining an accurate estimate of the nuisance parameters under a small

sample size with missing data is not easy. This explains why using a constant to model $\mathrm{Var}(Y|\boldsymbol{X})$ leads to smaller TMSE. When $N = 800$ so that the nuisance parameters are better estimated, using the correct model for $\mathrm{Var}(Y|\boldsymbol{X})$ starts to yield similar or smaller TMSE compared to using the incorrect model.

It is also of interest to observe the numerical evidence on the convergence of the asymptotic variance estimators given in Theorem 3 and Corollary 3. Convergence is well demonstrated by the comparison across different sample sizes. When $N = 200$, both estimators had slight underestimation, but this disappeared when N increaseed to 800.

Since the proposed CEL procedure involves nonparametric calculation of the weight w_{ij}, one important question is whether increasing the number of covariates substantially affects the numerical performance. To assess this impact, we conducted the following simulation experiment. The simulation model involved four covariates, $X_1 \sim \mathcal{N}(0, 2^2)$, $X_2 \sim$ Bernoulli(0.5), $X_3 \sim \mathcal{N}(0, 1^2)$, and $X_4 \sim \mathcal{N}(0, 1^2)$. The auxiliary variable was generated by $S = 1 + X_1 + X_2 + X_3 + X_4 + \epsilon_S$ with $\epsilon_S \sim \mathcal{N}(0, 2^2)$, and the outcome Y was generated as $Y = 1 + S + 0.6X_1 + 2X_2 + 0.5X_3 + 0.5X_4 + \epsilon_Y$, where $\epsilon_Y \sim \mathcal{N}(0, \sigma_Y^2)$ with $\sigma_Y^2 = \exp(0.92 + 0.8X_1 + 0.4X_2)$. For this model, $Y|\boldsymbol{X}$ had a Normal distribution with mean $E(Y|\boldsymbol{X}) = 2 + 1.6X_1 + 3X_2 + 1.5X_3 + 1.5X_4$ and variance $\mathrm{Var}(Y|\boldsymbol{X}) = 4 + \exp(0.92 + 0.8X_1 + 0.4X_2)$. The missingness mechanism was set to be $\mathrm{logit}\{\pi(S, \boldsymbol{X})\} = 0.5 - 0.2S + 0.6X_1 - 0.2X_2 + 0.2X_3 + 0.2X_4$, under which approximately 48% of subjects had missing Y in the generated data. Compared to the previous simulation model, this new model had two extra continuous covariates X_3 and X_4, and had $\boldsymbol{\beta}_0 = (\beta_1, \cdots, \beta_5)^T = (2, 1.6, 3, 1.5, 1.5)^T$, $\boldsymbol{\alpha}_0 = (0.5, -0.2, 0.6, -0.2, 0.2, 0.2)^T$, and $\boldsymbol{\gamma}_0 = (1, 1, 0.6, 2, 0.5, 0.5)^T$. When $\pi(S, \boldsymbol{X})$ or $E(Y|S, \boldsymbol{X})$ was incorrectly modeled, they were incorrectly modeled as before. The numerical performance of $\hat{\boldsymbol{\beta}}_{AIPW}$ is summarized in Table 3 using 500 replications. The bandwidth was again selected by minimizing the cross validation criterion (2.6). The infeasible AIPW estimator based on $\boldsymbol{U}(\boldsymbol{\beta}; Y, \boldsymbol{X}) = \boldsymbol{X}\mathrm{Var}(Y|\boldsymbol{X})^{-1}(Y - \boldsymbol{X}^T\boldsymbol{\beta})$, where $\mathrm{Var}(Y|\boldsymbol{X})$ was given by its true value $4 + \exp(0.92 + 0.8X_1 + 0.4X_2)$, is included for comparison and is denoted by AIPW_{inf}. From Table 3, an increase in the number of covariates does not seem to have a dramatic impact on the numerical performance of $\hat{\boldsymbol{\beta}}_{AIPW}$.

5. Data Application

We applied the proposed method to an intervention study for adolescent children of parents with HIV (Rotheram-Borus et al. (2004)). In this study, a total of 307 parents having HIV, with adolescent children, were recruited from the Division of AIDS Services in New York City, and 423 adolescents from these families were eligible for study participation. After recruitment, each parent and each

Table 3. Numerical results for the model containing four covariates. For each estimator, we report the bias, the empirical standard error (the number in ( )), and the mean square error (the number in [ ]).

	$N = 200$					$N = 800$				
estimator	β_1	β_2	β_3	β_4	β_5	β_1	β_2	β_3	β_4	β_5
correct $\pi(\boldsymbol{S}, \boldsymbol{X})$										
AIPW$_{inf}$	0.02	-0.01	-0.08	0.00	-0.04	0.00	0.01	-0.02	-0.01	0.00
	(0.54)	(0.40)	(1.01)	(0.55)	(0.55)	(0.25)	(0.17)	(0.45)	(0.25)	(0.26)
	[0.29]	[0.16]	[1.02]	[0.30]	[0.30]	[0.06]	[0.03]	[0.20]	[0.06]	[0.07]
CEL-AIPW	0.02	0.01	-0.06	-0.03	-0.01	0.00	0.01	-0.02	-0.01	0.00
	(0.43)	(0.26)	(0.76)	(0.40)	(0.39)	(0.21)	(0.12)	(0.36)	(0.19)	(0.20)
	[0.19]	[0.07]	[0.58]	[0.16]	[0.15]	[0.05]	[0.02]	[0.13]	[0.04]	[0.04]
correct $E(Y\|\boldsymbol{S}, \boldsymbol{X})$										
AIPW$_{inf}$	-0.01	0.00	0.01	0.01	0.01	-0.01	0.00	0.00	0.00	0.00
	(0.37)	(0.21)	(0.57)	(0.28)	(0.28)	(0.18)	(0.10)	(0.25)	(0.12)	(0.12)
	[0.14]	[0.05]	[0.32]	[0.08]	[0.08]	[0.03]	[0.01]	[0.06]	[0.01]	[0.02]
CEL-AIPW	-0.01	-0.01	0.00	0.01	0.02	0.00	0.00	0.00	0.00	0.01
	(0.37)	(0.23)	(0.62)	(0.33)	(0.31)	(0.19)	(0.10)	(0.27)	(0.15)	(0.15)
	[0.14]	[0.05]	[0.39]	[0.11]	[0.10]	[0.04]	[0.01]	[0.07]	[0.02]	[0.02]
both models correct										
AIPW$_{inf}$	-0.01	0.00	0.01	0.03	0.02	0.00	0.00	0.00	0.00	0.00
	(0.38)	(0.22)	(0.59)	(0.28)	(0.29)	(0.17)	(0.07)	(0.24)	(0.11)	(0.11)
	[0.14]	[0.05]	[0.35]	[0.08]	[0.09]	[0.03]	[0.01]	[0.06]	[0.01]	[0.01]
CEL-AIPW	-0.01	0.00	0.00	0.02	0.02	0.00	0.00	0.00	0.01	0.01
	(0.37)	(0.21)	(0.60)	(0.30)	(0.30)	(0.18)	(0.09)	(0.25)	(0.13)	(0.13)
	[0.14]	[0.04]	[0.36]	[0.09]	[0.09]	[0.03]	[0.01]	[0.06]	[0.02]	[0.02]

adolescent received a baseline interview, which collected information on background characteristics as well as the measurements for adolescent assessment, such as emotional distress and somatic symptoms. At the end of the baseline interview, participant families were randomly assigned either to the intervention arm or to the control arm. The intervention in this study was designed using social learning theory and cognitive-behavioral principles (Bandura (1994)). Depending on the parents' phase of illness, families received the intervention in 3 different modules, which covered different aspects of information on the tasks for either parents or adolescents. The researchers followed up on the participants every 3 months for the first 2 years and every 6 months thereafter, until the end of 6 years. At each follow-up, measurements for adolescent assessment were collected.

In our analysis we used a subset of the data that contains the assessments on adolescents' emotional distress, which were collected using the Brief Symptom Inventory (BSI). BSI is a commonly used psychological survey consisting of 53 items that belong to 9 sub-groups. Each item is associated with a psychiatric symptom and has a 0-to-4 rate scale. Subjects report values to each item according to the level that they have been troubled by the corresponding symptom in the past week, with 0 meaning "having not been troubled at all" and 4 meaning "having been troubled a lot". One scientifically in-

teresting question is whether having parents with HIV has disparate impacts on the emotional distress between boys and girls during the delivery of intervention. Such a gender disparity, if it exists, may suggest the need for the development of gender-specific interventions that could result in more beneficial achievement. We try to answer this question using data collected at the end of the first year of intervention. The data are downloaded from "`http://rem.ph.ucla.edu/rob/mld/data/tabdelimiteddata/bsitotal.txt`", and a detailed description about the data can be found in Weiss (2005).

The outcome variable is the global severity index, which is the average rating score over all 53 items. Due to the skewed distribution of the global severity index and the possibility of occurrence of value 0, following the analysis instruction in Weiss (2005), we created a new outcome gsi by adding a small constant $1/53$ to the global severity index and then taking the log-transformation with base 2. We assumed

$$gsi = \beta_1 + \beta_2 age + \beta_3 girl + \beta_4 int + \epsilon,$$

where age is the age of adolescent at the end of the first year of intervention, $girl$ is gender indicator with $girl = 0$ for boys and $girl = 1$ for girls, int is the intervention indicator with $int = 0$ for the control arm and $int = 1$ for the intervention arm, and ϵ is the error term that has mean 0 conditional on all three covariates. However, scores on gsi are only available for about half of the adolescents at the end of the first year. On the other hand, almost all adolescents have their baseline gsi score observed. Therefore, we treated the baseline gsi, denoted by $bgsi$, as an auxiliary variable. To better model the missingness mechanism, we created two dummy variables, namely *winter* and *summer*, as indicators for the season (*winter* indicates November through February, *summer* indicates July through October, and the rest time of the calendar year is treated as the reference) when the measurements at the end of the first year were taken. These two dummy variables were considered as extra auxiliary variables. After removing adolescents who did not have scores on $bgsi$, we ended up with $N = 420$ subjects, 204 of whom did not have a score on gsi (the missing data proportion was 49%). There were in total 221 girls and 199 boys, and 211 were in the intervention arm and 209 were in the control arm. The average age was 16 years old, with a standard deviation 2 years.

To model the missingness mechanism, we fit a logistic regression model, and the results are presented in Table 4. It is seen there that having higher score on $bgsi$ significantly increases the probability of missing the interview conducted at the end of the first year. The season when the interview was conducted also plays a significant role, in the sense that subjects are more likely to take the interview during *winter* and *summer* seasons compared to the rest time of the year. A linear regression was employed to model $E(Y|\boldsymbol{S}, \boldsymbol{X})$.

Table 4. Results of modeling the missingness mechanism for the intervention study data ($N = 420$).

	est	se	z-value	p-value
constant	0.473	0.819	0.578	0.564
bgsi	-0.148	0.064	-2.316	0.021
winter	0.830	0.244	3.396	0.001
summer	0.513	0.245	2.089	0.037
age	-0.067	0.049	-1.353	0.176
girl	0.042	0.204	0.206	0.837
int	0.165	0.201	0.818	0.413

est: estimated value; se: standard error.

Table 5. Analysis results for the intervention study data ($N = 420$).

	CEL-AIPW			complete-case analysis			IPW			AIPW		
	est	se	p-value	est	se	p-value	est	se	p-value	est	se	p-value
constant	-5.531	1.098	< 0.000	-5.231	1.195	< 0.000	-5.476	1.220	< 0.000	-5.538	1.103	< 0.000
age	0.164	0.066	0.013	0.139	0.071	0.053	0.169	0.073	0.021	0.165	0.067	0.014
girl	0.745	0.265	0.005	0.634	0.294	0.032	0.627	0.295	0.034	0.734	0.266	0.006
int	0.277	0.266	0.297	0.200	0.296	0.500	0.175	0.297	0.556	0.267	0.267	0.317

est: estimated value; se: standard error.

Table 5 contains the final results of our analysis. The bandwidth was selected by minimizing the cross validation criterion (2.6). To make comparisons, we also include the results based on the complete-case analysis, the IPW and the AIPW estimators, for which a constant variance was used in the estimating functions. All methods conclude that gender had a significant effect on the global severity index, whereas the intervention did not. The age effect was significant based on the CEL-AIPW, the IPW and the AIPW methods, but was only marginally significant based on the complete-case analysis. The CEL-AIPW and the AIPW estimators produced very similar estimate for each covariate effect. On average, one year increase in age led to a $2^{0.164} - 1 = 12\%$ increase in the global severity index, and girls had their global severity index $2^{0.745} - 1 = 68\%$ higher than boys, where each effect was interpreted by holding all the others fixed. Having parents with HIV has different impacts on the emotional distress between boys and girls during the delivery of intervention, at least after one year of the delivery.

6. Discussion

We have investigated the CEL method for mean regression analysis when the outcome is missing at random, and studied the asymptotic properties of the CEL-IPW and the CEL-AIPW estimators. Our method does not model any second moments of the data distribution. As a result, achieving semiparametric

efficiency only requires the correct modeling of the missingness mechanism and the mean of the outcome.

Many researchers have focused on improving the efficiency of the AIPW estimator when only the missingness mechanism is correctly modeled. Most recent developments are under the setting of estimating the population mean of a variable that is missing at random. One idea is that (e.g., Tan (2006, 2008, 2010); Wang, Rotnitzky, and Lin (2010); Han (2012)), instead of simply taking the difference of the two terms in (1.3), one could use the residual of the projection of the first term on the second. Intuitively, the resulting estimator should be more efficient than the one based on the difference, the AIPW estimator. However, directly using the residual of the projection leads to an estimator without the double robustness property. For estimation of the population mean, Tan (2006, 2008, 2010) proposed a modification to the projection coefficient so that the resulting estimator is also doubly robust. Wang, Rotnitzky, and Lin (2010) and Han (2012) applied Tan's idea to the settings of nonparametric regression and estimating equations, respectively. The same idea can be applied to the CEL based estimation. Let

$$g^{\hat{\kappa}}(\boldsymbol{\beta},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})=\frac{R}{\pi(\hat{\boldsymbol{\alpha}};\boldsymbol{S},\boldsymbol{X})}\left\{Y-\mu(\boldsymbol{X}^T\boldsymbol{\beta})\right\} \\ -\hat{\kappa}(\boldsymbol{\beta},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})\frac{R-\pi(\hat{\boldsymbol{\alpha}};\boldsymbol{S},\boldsymbol{X})}{\pi(\hat{\boldsymbol{\alpha}};\boldsymbol{S},\boldsymbol{X})}\left\{h(\hat{\boldsymbol{\gamma}};\boldsymbol{S},\boldsymbol{X})-\mu(\boldsymbol{X}^T\boldsymbol{\beta})\right\},$$

where

$$\hat{\kappa}(\boldsymbol{\beta},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})=\frac{\hat{E}\left[\frac{R}{\pi(\hat{\boldsymbol{\alpha}};\boldsymbol{S},\boldsymbol{X})}\frac{R-\pi(\hat{\boldsymbol{\alpha}};\boldsymbol{S},\boldsymbol{X})}{\pi(\hat{\boldsymbol{\alpha}};\boldsymbol{S},\boldsymbol{X})}\left\{Y-\mu(\boldsymbol{X}^T\boldsymbol{\beta})\right\}\left\{h(\hat{\boldsymbol{\gamma}};\boldsymbol{S},\boldsymbol{X})-\mu(\boldsymbol{X}^T\boldsymbol{\beta})\right\}\mid\boldsymbol{X}\right]}{\hat{E}\left[\frac{R}{\pi(\hat{\boldsymbol{\alpha}};\boldsymbol{S},\boldsymbol{X})}\frac{1-\pi(\hat{\boldsymbol{\alpha}};\boldsymbol{S},\boldsymbol{X})}{\pi(\hat{\boldsymbol{\alpha}};\boldsymbol{S},\boldsymbol{X})}\left\{h(\hat{\boldsymbol{\gamma}};\boldsymbol{S},\boldsymbol{X})-\mu(\boldsymbol{X}^T\boldsymbol{\beta})\right\}^2\mid\boldsymbol{X}\right]}$$

and $\hat{E}(\cdot|\boldsymbol{X})$ is a consistent estimator of $E(\cdot|\boldsymbol{X})$. A new CEL-based estimator, CEL-AIPW-N, can be defined by using $g^{\hat{\kappa}}(\boldsymbol{\beta},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})$ instead of $g(\boldsymbol{\beta},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})$. When $\boldsymbol{\alpha}_*=\boldsymbol{\alpha}_0$, it is easy to show that $E\{g^{\kappa}(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_0,\boldsymbol{\gamma}_*)|\boldsymbol{X}\}=0$, where g^{κ} is $g^{\hat{\kappa}}$ by replacing $\hat{\kappa}(\boldsymbol{\beta},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})$ with $\kappa(\boldsymbol{\beta},\boldsymbol{\alpha}_0,\boldsymbol{\gamma}_*)$, and κ is $\hat{\kappa}$ by replacing $\hat{E}(\cdot|\boldsymbol{X})$ with $E(\cdot|\boldsymbol{X})$. When $\boldsymbol{\gamma}_*=\boldsymbol{\gamma}_0$, it is easy to show that the numerator and denominator of $\kappa(\boldsymbol{\beta},\boldsymbol{\alpha}_*,\gamma_0)$ are equal, and thus $\kappa(\boldsymbol{\beta},\boldsymbol{\alpha}_*,\gamma_0)=1$. Therefore, when $\boldsymbol{\gamma}_*=\boldsymbol{\gamma}_0$, we have $E\{g^{\kappa}(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_*,\boldsymbol{\gamma}_0)|\boldsymbol{X}\}=0$. These facts guarantee the double robustness and local efficiency of the CEL-AIPW-N estimator. When $\boldsymbol{\alpha}_*=\boldsymbol{\alpha}_0$, it is easy to show that the denominator of $\kappa(\boldsymbol{\beta},\boldsymbol{\alpha}_0,\boldsymbol{\gamma}_*)$ is

$$E\left(\left[\frac{R-\pi(\boldsymbol{\alpha}_0;\boldsymbol{S},\boldsymbol{X})}{\pi(\boldsymbol{\alpha}_0;\boldsymbol{S},\boldsymbol{X})}\left\{h(\boldsymbol{\gamma}_*;\boldsymbol{S},\boldsymbol{X})-\mu(\boldsymbol{X}^T\boldsymbol{\beta})\right\}\right]^2\mid\boldsymbol{X}\right),$$

and thus $g^{\hat{\kappa}}(\boldsymbol{\beta},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})$ is the sample version of

$$g^{\kappa}(\boldsymbol{\beta},\boldsymbol{\alpha}_0,\boldsymbol{\gamma}_*)=\frac{R}{\pi}\left\{Y-\mu(\boldsymbol{\beta})\right\}$$

$$-\frac{E\left[\frac{R}{\pi}\frac{R-\pi}{\pi}\{Y-\mu(\boldsymbol{\beta})\}\{h(\boldsymbol{\gamma}_*)-\mu(\boldsymbol{\beta})\}\mid \boldsymbol{X}\right]}{E\left(\left[\frac{R-\pi}{\pi}\{h(\boldsymbol{\gamma}_*)-\mu(\boldsymbol{\beta})\}\right]^2\mid \boldsymbol{X}\right)}\frac{R-\pi}{\pi}\{h(\boldsymbol{\gamma}_*)-\mu(\boldsymbol{\beta})\},$$

the residual of the projection of the first term in $g(\boldsymbol{\beta},\boldsymbol{\alpha}_0,\boldsymbol{\gamma}_*)$ on the second term conditional on $\boldsymbol{X}$. The asymptotic distribution of the CEL-AIPW-N estimator follows from Theorem 2 with $\boldsymbol{\alpha}_*=\boldsymbol{\alpha}_0$ and $g(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_0,\boldsymbol{\gamma}_*)$ substituted by $g^{\kappa}(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_0,\boldsymbol{\gamma}_*)$. Due to the projection structure of $g^{\kappa}(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_0,\boldsymbol{\gamma}_*)$, the asymptotic variance of the CEL-AIPW-N estimator is no larger than that of the CEL-AIPW estimator when $\boldsymbol{\alpha}_*=\boldsymbol{\alpha}_0$. When $\boldsymbol{\gamma}_*=\boldsymbol{\gamma}_0$ as well, since $\kappa(\boldsymbol{\beta},\boldsymbol{\alpha}_*,\boldsymbol{\gamma}_0)=1$, both estimators achieve the semiparametric efficiency bound. Despite its efficiency gain, the implementation of the CEL-AIPW-N estimator is difficult due to the extra term $\hat{\kappa}(\boldsymbol{\beta},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})$, which involves not only conditional expectations, but also the unknown parameter $\boldsymbol{\beta}$.

Another idea to improve the efficiency of the AIPW estimator when only the missingness mechanism is correctly modeled is that (e.g., Rubin and van der Laan (2008); Tan (2008); Cao, Tsiatis, and Davidian (2009)), instead of using $\hat{\boldsymbol{\gamma}}$ that solves (2.3) based on complete-case analysis, one treats the asymptotic variance of the AIPW estimator as a function of $\boldsymbol{\gamma}$ and use the value $\hat{\boldsymbol{\gamma}}$ that minimizes this function. The minimization is straightforwardly defined only when the parameter of interest is a scalar, such as in the setting of estimating the population mean, since only in this case is the asymptotic variance a scalar. In our setting, the asymptotic variance is a complicated matrix of functions of $\boldsymbol{\gamma}$, and the application and implementation of this idea is difficult, if not infeasible.

The proposed CEL method enjoys high estimation efficiency when moderate to high level of heteroscedasticity exists, especially when the heteroscedasticity is hard to model. When homoscedasticity is a more reasonable assumption, the proposed estimators may not outperform existing ones, such as the IPW or the AIPW estimators, due to the nonparametric nature. In many practical studies the main interest is to estimate the effect of one particular covariate adjusting for a few others. In this case the number of covariates is not very large, and the proposed method provides a useful alternative for efficiency improvement, especially if there exists unknown and hard-to-model heteroscedasticity. Our simulation results showed that the numerical performance of the proposed estimators is reasonable with several covariates. The cross-validation criterion for bandwidth selection is easy to implement, and numerical results in both Kitamura, Tripathi, and Ahn (2004) and our paper demonstrated the insensitivity of the proposed estimators to bandwidth selection. When the number of covariates is large, the proposed method will suffer from the curse of dimensionality, which exists for most nonparametric estimation procedures. The current theory of CEL was established based on weights calculated from the Kernel method. Whether

some other forms of weights, such as weights calculated based on splines, can lead to the same theoretical results is a challenging research topic, and deserves future investigation.

Acknowledgements

We thank the Editor, an associate editor and two referees for their valuable comments that have helped improve the quality of this paper. Support for this project was partially provided by the Natural Sciences and Engineering Research Council of Canada (NSERC) with grant number RGPIN-05055-2014 to the first author, and by the Natural Science Foundation (NSF) with grant number DMS-1208939 to the third author.

Appendix

Let $\mathcal{B}$, $\mathcal{A}$, $\mathcal{G}$, and $\mathcal{X}$ denote the domain of $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, $\boldsymbol{\gamma}$, and $\boldsymbol{X}$ respectively. Let $\mathcal{B}_0 \subseteq \mathcal{B}$ be some closed ball around $\boldsymbol{\beta}_0$, and take $\boldsymbol{G}_{\boldsymbol{\alpha}}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) = E\{\partial g(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma})/\partial \boldsymbol{\alpha} | \boldsymbol{X}\}$.

Assumption.

(i) (i) $\mathcal{B}$, $\mathcal{A}$, $\mathcal{G}$, and $\mathcal{X}$ are compact.

(ii) $\mu(\cdot)$, $\pi(\boldsymbol{\alpha})$, and $h(\boldsymbol{\gamma})$ are continuously differentiable.

(iii) For any $\boldsymbol{\beta} \neq \boldsymbol{\beta}_0$, there exists $\mathcal{X}_{\boldsymbol{\beta},\boldsymbol{\alpha}_*,\boldsymbol{\gamma}_*} \subseteq \mathcal{X}$ such that $P(\boldsymbol{x} \in \mathcal{X}_{\boldsymbol{\beta},\boldsymbol{\alpha}_*,\boldsymbol{\gamma}_*}) > 0$ and $E\{g(\boldsymbol{\beta}, \boldsymbol{\alpha}_*, \boldsymbol{\gamma}_*) | \boldsymbol{X} = \boldsymbol{x}\} \neq 0$ for every $\boldsymbol{x} \in \mathcal{X}_{\boldsymbol{\beta},\boldsymbol{\alpha}_*,\boldsymbol{\gamma}_*}$.

(iv) $E\left\{\sup_{\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\gamma}} |g(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma})|^m\right\} < \infty$ for some $m \geq 8$.

(v) $0 < \inf_{\boldsymbol{X},\boldsymbol{\beta}\in\mathcal{B}_0,\boldsymbol{\alpha},\boldsymbol{\gamma}} V(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) \leq \sup_{\boldsymbol{X},\boldsymbol{\beta}\in\mathcal{B}_0,\boldsymbol{\alpha},\boldsymbol{\gamma}} V(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) < \infty$, where $V(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) = E\left\{g(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma})^2 | \boldsymbol{X}\right\}$.

(vi) $b_N \to 0$, $N^{1-2\nu-2/m} b_N^{2q} \to \infty$, and $N^{1-2\nu} b_N^{5q/2} \to \infty$ as $N \to \infty$, where $\nu \in (0, 1/2)$, $m \geq 8$, and q is the dimension of $\boldsymbol{X}^c$.

(vii)$|\ \hat{\lambda}_i\ | \leq cN^{-1/m}$ for some $c > 0$.

Compactness of the parameter space in (i) is commonly imposed in large sample theory (e.g., Newey and McFadden (1994)). Differentiability in (ii) usually holds for the models used in practice. Assumption (iii) pertains to the identifiability of $\boldsymbol{\beta}_0$. Assumption (iv) is to ensure the uniform weak law of large numbers (e.g., Newey and McFadden (1994)). Assumption (v) guarantees that the variance-covariance matrix of the AIPW residual is invertible in a neighborhood of $\boldsymbol{\beta}_0$. The restrictions on b_N in (vi) follow those in Smith (2007); here the parameter $v \in (0, 1/2)$ appears due to the uniform convergence rate for kernel estimator (Kitamura, Tripathi, and Ahn (2004)). Assumption (vii) is similar to

Assumption 3.6 in Kitamura, Tripathi, and Ahn (2004), and can be established under some more elementary conditions (Lemma B.1 in Kitamura, Tripathi, and Ahn (2004)).

In the following proofs, we suppress the subscript "AIPW". Some technical details can be filled in by following the proofs in Kitamura, Tripathi, and Ahn (2004).

Proof of Theorem 1. With

$$H(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij} \log\left\{1 + \hat{\lambda}_i(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) g_j(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})\right\},$$

$\hat{\boldsymbol{\beta}}_{AIPW}$ is the maximizer of $H(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})$. It can be shown that $N^{1/m} H(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) \leq F(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) + o_p(1)$ for any $\boldsymbol{\beta} \in \mathcal{B}$, where

$$F(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) = -E\left[\frac{\mid E\left\{g(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) | \boldsymbol{X}\right\} \mid^2}{1 + \mid E\left\{g(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) | \boldsymbol{X}\right\} \mid}\right]$$

is continuous with respect to $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$. Therefore,

$$N^{1/m} H(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) \leq F(\boldsymbol{\beta}, \boldsymbol{\alpha}_*, \boldsymbol{\gamma}_*) + o_p(1) \quad \text{for any } \boldsymbol{\beta} \in \mathcal{B}. \tag{A.1}$$

On the other hand, from Assumption (iii), for any $\boldsymbol{\beta} \neq \boldsymbol{\beta}_0$, we have

$$F(\boldsymbol{\beta}, \boldsymbol{\alpha}_*, \boldsymbol{\gamma}_*) \leq -E\left[\mathbb{I}(\boldsymbol{X} \in \mathcal{X}_{\boldsymbol{\beta}, \boldsymbol{\alpha}_*, \boldsymbol{\gamma}_*}) \frac{\mid E\left\{g(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) | \boldsymbol{X}\right\} \mid^2}{1 + \mid E\left\{g(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) | \boldsymbol{X}\right\} \mid}\right] < 0.$$

Hence, from (A.1), the continuity of $F(\boldsymbol{\beta}, \boldsymbol{\alpha}_*, \boldsymbol{\gamma}_*)$ and the compactness of $\mathcal{B}$, for any $\delta > 0$, there exists $C(\delta) > 0$, such that

$$\sup_{\boldsymbol{\beta} \in \mathcal{B}/B(\boldsymbol{\beta}_0, \delta)} N^{1/m} H(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) \leq \sup_{\boldsymbol{\beta} \in \mathcal{B}/B(\boldsymbol{\beta}_0, \delta)} F(\boldsymbol{\beta}, \boldsymbol{\alpha}_*, \boldsymbol{\gamma}_*) + o_p(1) \leq -C(\delta) + o_p(1), \tag{A.2}$$

where $B(\boldsymbol{\beta}_0, \delta)$ is the ball centering at $\boldsymbol{\beta}_0$ with radius δ.

Assumption (vi) and (B.4) in Kitamura, Tripathi, and Ahn (2004) lead to $\max_{1 \leq i \leq N} \hat{\lambda}_i(\boldsymbol{\beta}_0, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) = o_p(N^{-1/m})$ if either $\boldsymbol{\alpha}_* = \boldsymbol{\alpha}_0$ or $\boldsymbol{\gamma}_* = \boldsymbol{\gamma}_0$. Therefore, since

$$H(\boldsymbol{\beta}_0, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) \geq -\frac{1}{N}\sum_{i=1}^{N}\left\{\hat{\lambda}_i(\boldsymbol{\beta}_0, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) \sum_{j=1}^{N} w_{ij} g_j(\boldsymbol{\beta}_0, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})\right\},$$

we have $N^{1/m} J(\boldsymbol{\beta}_0, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) \geq o_p(1)$. This, together with (A.2), gives the consistency of $\hat{\boldsymbol{\beta}}_{AIPW}$.

Proof of Theorem 2. Taking the Taylor expansion of $\partial L(\hat{\boldsymbol{\beta}}_{AIPW}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})/\partial \boldsymbol{\beta}^T = \mathbf{0}$ around $\boldsymbol{\beta}_0$ gives

$$\sqrt{N}(\hat{\boldsymbol{\beta}}_{AIPW} - \boldsymbol{\beta}_0) = \left\{ -\frac{1}{N}\frac{\partial^2 L(\tilde{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T} \right\}^{-1} \left\{ \frac{1}{\sqrt{N}}\frac{\partial L(\boldsymbol{\beta}_0, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})}{\partial \boldsymbol{\beta}^T} \right\},$$

where $\tilde{\boldsymbol{\beta}}$ is some point between $\hat{\boldsymbol{\beta}}_{AIPW}$ and $\boldsymbol{\beta}_0$. Following the proofs of Lemma C.1, (A.14), and (B.7) in Kitamura, Tripathi, and Ahn (2004) we have

$$-\frac{1}{N}\frac{\partial^2 L(\tilde{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T} \xrightarrow{p} I(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_*, \boldsymbol{\gamma}_*),$$

$$\begin{aligned}
&\frac{1}{\sqrt{N}}\frac{\partial L(\boldsymbol{\beta}_0, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})}{\partial \boldsymbol{\beta}^T} \\
&\quad = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} \boldsymbol{Q}_i(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_*, \boldsymbol{\gamma}_*) \\
&\qquad + E\left\{ \frac{\partial \mu(\boldsymbol{X}^T\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}^T} V(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_*, \boldsymbol{\gamma}_*)^{-1} \boldsymbol{G}_{\boldsymbol{\alpha}}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_*, \boldsymbol{\gamma}_*) \right\} \sqrt{N}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_*) \\
&\qquad + E\left\{ \frac{\partial \mu(\boldsymbol{X}^T\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}^T} V(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_*, \boldsymbol{\gamma}_*)^{-1} \boldsymbol{G}_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_*, \boldsymbol{\gamma}_*) \right\} \sqrt{N}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_*) + o_p(1).
\end{aligned}$$

When $\boldsymbol{\alpha}_* = \boldsymbol{\alpha}_0$,

$$\boldsymbol{G}_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0, \boldsymbol{\gamma}_*) = E\left[E\left\{ -\frac{R - \pi(\boldsymbol{\alpha}_0)}{\pi(\boldsymbol{\alpha}_0)}\frac{\partial h(\boldsymbol{\gamma}_*)}{\partial \boldsymbol{\gamma}} | \boldsymbol{X}, \boldsymbol{S} \right\} | \boldsymbol{X} \right] = \mathbf{0},$$

and when $\boldsymbol{\gamma}_* = \boldsymbol{\gamma}_0$,

$$\boldsymbol{G}_{\boldsymbol{\alpha}}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_*, \boldsymbol{\gamma}_0) = E\left\{ E\left[-\frac{R}{\pi(\boldsymbol{\alpha}_*)^2}\frac{\partial \pi(\boldsymbol{\alpha}_*)}{\partial \boldsymbol{\alpha}} \{Y - h(\boldsymbol{\gamma}_0)\} | \boldsymbol{X}, \boldsymbol{S} \right] | \boldsymbol{X} \right\} = \mathbf{0}.$$

Combining all these facts together with the linear expansion (3.1) and (3.2) and the information equalities $\boldsymbol{G}_{\boldsymbol{\alpha}}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0, \boldsymbol{\gamma}_*) = -E\{g(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0, \boldsymbol{\gamma}_*)\boldsymbol{\psi}(\boldsymbol{\alpha}_0)^T \mid \boldsymbol{X}\}$ and $E\{\partial \boldsymbol{\psi}(\boldsymbol{\alpha}_0)/\partial \boldsymbol{\alpha}\} = -E\{\boldsymbol{\psi}(\boldsymbol{\alpha}_0)\boldsymbol{\psi}(\boldsymbol{\alpha}_0)^T\}$, the Central Limit Theorem gives the desired results.

Proof of Theorem 3. Following the proof of Lemma D.2 in Kitamura, Tripathi, and Ahn (2004), we have $\max_{1\le i\le N} \sup_{\boldsymbol{\beta}\in\mathcal{B}_0} |g_i(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})| = o_p(N^{1/m})$. Then Assumption (vii) leads to $\max_{1\le i,j\le N} \sup_{\boldsymbol{\beta}\in\mathcal{B}_0} |\hat{\lambda}_i g_j(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})| = o_p(1)$. Therefore $p_{ij}(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) = w_{ij}\{1 + o_p(1)\}$, and the $o_p(1)$ term is independent of i, j and $\boldsymbol{\beta}$. This fact, together with the Weak Law of Large Numbers, gives the results.

References

Bandura, A. (1994). Social cognitive theory and exercise of control over HIV infection, chapter *Preventing AIDS: Theories and Methods of Behavioral Interventions*. Plenum Press, New York.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y. and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press, Baltimore.

Cao, W., Tsiatis, A. A. and Davidian, M. (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika* **96**, 723-734.

Chen, J. and Breslow, N. E. (2004). Semiparametric efficient estimation for the auxiliary outcome problem with conditional mean model. *Canad. J. Statist.* **32**, 359-372.

Chen, J., Sitter, R. R. and Wu, C. (2002). Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys. *Biometrika* **89**, 230-237.

Chen, S. X., Leung, D. H. Y. and Qin, J. (2003). Information recovery in a study with surrogate endpoints. *J. Amer. Statist. Assoc.* **98**, 1052-1062.

Chen, S. X., Leung, D. H. Y. and Qin, J. (2008). Improving semiparametric estimation by using surrogate data. *J. Roy. Statist. Soc. Ser. B* **70**, 803-823.

Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. *Ann. Math. Statist.* **31**, 1208-1212.

Godambe, V. P. (1991). *Estimating Functions*. Oxford University Press, Oxford.

Han, P. (2012). A note on improving the efficiency of inverse probability weighted estimator using the augmentation term. *Statist. Probab. Lett.* **82**, 2221-2228.

Han, P. (2014a). A further study of the multiply robust estimator in missing data analysis. *J. Statist. Plann. Inference* **148**, 101-110.

Han, P. (2014b). Multiply robust estimation in regression analysis with missing data. *J. Amer. Statist. Assoc.* **109**, 1159-1173.

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

Hansen, B. E. (2014). *Econometrics*. draft graduate textbook.

Heyde, C. C. (1988). Fixed sample and asymptotic optimality for classes of estimating functions. *Contemporary Math.* **80**, 241-247.

Heyde, C. C. (1997). *Quasi-likelihood and Its Application*. Springer-Verlag, New York.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.* **47**, 663-685.

Kang, J. D. Y. and Schafer, J. L. (2007). Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data (with discussion). *Statist. Sci.* **22**, 523-539.

Kitamura, Y. (2007). Empirical likelihood methods in econometrics: theory and practice. Chapter *Advances in Economics and Econometrics: Theory and Applications, NinthWorld Congress*, Vol. 3, 174-237. Cambridge University Press.

Kitamura, Y., Tripathi, G. and Ahn, H. (2004). Empirical likelihood-based inference in conditional moment restriction models. *Econometrica*, **72**, 1667-1714.

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

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman and Hall, London.

Newey, W. K. (1993). Efficient estimation of models with conditional moment restrictions. Chapter *Handbook of Statistics*, Vol 11, 419-454. North-Holland, Amsterdam.

Newey, W. K. and McFadden, D. L. (1994). Large sample estimation and hypothesis testing. Chapter *Handbook of Econometrics*, Vol 4, Elsevier Science, Amsterdam.

Owen, A. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, **75**, 237-249.

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

Owen, A. (2001). *Empirical Likelihood*. Chapman & Hall/CRC Press, New York.

Pepe, M. S. (1992). Inference using surrogate outcome data and a validation sample. *Biometrika*, **79**, 355-365.

Pepe, M. S., Reilly, M. and Fleming, T. R. (1994). Auxiliary outcome data and the mean score method. *J. Statist. Plann. Inference*, **42**, 137-160.

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

Qin, J., Shao, J. and Zhang, B. (2008). Efficient and doubly robust imputation for covariate-dependent missing responses. *J. Amer. Statist. Assoc.* **103**, 797-810.

Qin, J. and Zhang, B. (2007). Empirical-likelihood-based inference in missing response problems and its application in observational studies. *J. Roy. Statist. Soc. Ser. B* **69**, 101-122.

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

Robins, J. M. and Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regresion models with missing data. *J. Amer. Statist. Assoc.* **90**, 122-129.

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.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J. Amer. Statist. Assoc.* **90**, 106-121.

Robins, J. M., Sued, M., Gomez-Lei, Q. and Rotnitzky, A. (2007). Comment: performance of double-robust estimators when "inverse probability" weights are highly variable. *Statist. Sci.* **22**, 544-559.

Rotheram-Borus, M. J., Lee, M., Lin, Y. Y. and Lester, P. (2004). Six-year intervention outcomes for adolescent children of parents with the human immunodeficiency virus. *Archives of Pediatrics & Adolescent Medicine*, **158**, 742-748.

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

Rotnitzky, A. and Robins, J. M. (1995). Semiparametric regression estimation in the presence of dependent censoring. *Biometrika* **82**, 805-820.

Rubin, D. B. and van der Laan, M. J. (2008). Empirical efficiency maximization: improved locally efficient covariate adjustment in randomized experiments and survival analysis. *Internat. J. Biostatist.* **4**, article 5.

Smith, R. J. (2007). Efficient information theoretic inference for conditional moment restrictions. *J. Econom.* **138**, 430-460.

Tan, Z. (2006). A distributional approach for causal inference using propensity scores. *J. Amer. Statist. Assoc.* **101**, 1619-1637.

Tan, Z. (2008). Comment: Improved local efficiency and double robustness. *Internat. J. Biostatist.*, **4**, Article 10.

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

Tan, Z. (2011). Efficient restricted estimators for conditional mean models with missing data. *Biometrika* **98**, 663-684.

Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*. Springer, New York.

Tsiatis, A. A., Davidian, M. and Cao, W. (2011). Improved doubly robust estimation when data are monotonely coarsened, with application to longitudinal studies with dropout. *Biometrics*, **67**, 536-545.

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

Wang, L., Rotnitzky, A. and Lin, X. (2010). Nonparametric regression with missing outcomes using weighted kernel estimating equations. *J. Amer. Statist. Assoc.* **105**, 1135-1146.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the gauss-newton method. *Biometrika* **61**, 439-447.

Weiss, R. E. (2005). *Modeling Longitudinal Data*. Springer.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1-25.

Yu, M. and Nan, B. (2006). A revisit of semiparametric regression models with missing data. *Statist. Sinica* **16**, 1193-1212.

Zhang, J. and Gijbels, I. (2003). Sieve empirical likelihood and extensions of the generalized least squares. *Scand. J. Statist.* **30**, 1-24.

Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON N2L 3G1, Canada.

E-mail: peisonghan@uwaterloo.ca

Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA.

E-mail: luwang@umich.edu

Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA.

E-mail: pxsong@umich.edu

(Received January 2014; accepted June 2015)