FULL-SEMIPARAMETRIC-LIKELIHOOD-BASED INFERENCE FOR NON-IGNORABLE MISSING DATA

Yukun Liu [a], Pengfei Li [b], and Jing Qin [c]

[a] *KLATASDS - MOE, School of Statistics, East China Normal University, China*

[b] *Department of Statistics and Actuarial Science, University of Waterloo, Canada*

[c] *National Institute of Allergy and Infectious Diseases, National Institutes of Health, USA*

Supplementary Material

This supplementary material consists of seven sections. Section S1 reviews the results of the main paper and provides the regularity conditions needed in Theorems 1-3. Section S2 presents a proof of Corollary 1. Section S3 presents preliminary work for our proofs of Theorems 1–3, which are provided in Sections S4–S6, respectively. Section S7 contains additional simulation results.

S1 Main results of the paper

S1.1 Model setup and identifiability

Suppose $\{(y_i, \mathbf{x}_i, d_i), i = 1, \ldots, n\}$ are n independent and identically distributed copies of $(Y, \mathbf{X}, D)$, where the covariates $\mathbf{x}_i$ are always observed, and y_i is observed if and only if $d_i = 1$. Suppose there are n_1 completely observed data and n_2 partially observed data. Without loss of generality, we assume that $d_i = 1$, $i = 1, \dots, n_1$ and $d_i = 0$, $i = n_1 + 1, \dots, n$.

We assume that the missing probability satisfies the logistic regression model,

$$\mathrm{pr}(D = 0|\mathbf{x}, y) = \frac{\exp(\alpha^* + \mathbf{x}^\top\beta + y\gamma)}{1 + \exp(\alpha^* + \mathbf{x}^\top\beta + y\gamma)}, \tag{S1.1}$$

and $\mathrm{pr}(y|\mathbf{x}, D = 1) = f(y|\mathbf{x}, \xi)$. Let $\eta = \mathrm{pr}(D = 1)$ and $\alpha = \alpha^* + \log\{\eta/(1-\eta)\}$.

The parametric model for the conditional distribution of Y given $(\mathbf{X} = \mathbf{x}, D = 1)$ and (S1.1) together imply two DRMs:

$$\begin{aligned} \mathrm{pr}(y|\mathbf{x}, D = 0) &= \exp\{\gamma y - c(\mathbf{x}, \gamma, \xi)\} f(y|\mathbf{x}, \xi), & \text{(S1.2)} \\ \mathrm{pr}(\mathbf{x}|D = 0) &= \exp\{\alpha + \mathbf{x}^\top\beta + c(\mathbf{x}, \gamma, \xi)\} \mathrm{pr}(\mathbf{x}|D = 1), & \text{(S1.3)} \end{aligned}$$

where

$$c(\mathbf{x}, \gamma, \xi) = \ln\left\{\int \exp(y\gamma) f(y|\mathbf{x}, \xi) dy\right\}. \tag{S1.4}$$

Based on (S1.2)–(S1.3) and the fact that ξ is identifiable, we have the following lemma regarding the identifiability of (α, β, γ).

Proposition 1 of the main paper. *Let S be the common support of $\mathrm{pr}(\mathbf{x}|D = 0)$ and $\mathrm{pr}(\mathbf{x}|D = 1)$, and*

$\Omega = \{h(\mathbf{x}) : S \mapsto \mathbb{R} \mid \exists(\alpha, \beta, \gamma)$ *such that* $\forall\, \mathbf{x} \in S, h(\mathbf{x}) = \alpha + \mathbf{x}^\top\beta + c(\mathbf{x}, \gamma, \xi)\}$.

If for any $h(\mathbf{x}) \in \Omega$, there exists a unique (α, β, γ) such that $h(\mathbf{x}) = \alpha + \mathbf{x}^\top \beta + c(\mathbf{x}, \gamma, \xi)$, then (α, β, γ) is identifiable.

Applying the above proposition, we find that (α, β, γ) is identifiable in two specific cases.

Corollary 1 of the main paper. *Suppose the logistic regression model in* (S1.1) *holds and that the density function of Y given $(\mathbf{X} = \mathbf{x}, D = 1)$ is $f(y|\mathbf{x}, \xi)$.*

(a) *If there exists an instrument variable z in $\mathbf{x}$, then (α, β, γ) is identifiable.*

(b) *Assume that the set S in Proposition 1 of the main paper contains an open set, and $c(\mathbf{x}, \gamma, \xi)$ can be expressed as*

$$c(\mathbf{x}, \gamma, \xi) = \sum_{i=1}^{k} a_i(\gamma) g_i(\mathbf{x}) + a_{k+1}(\gamma) + \mathbf{x}^\top a_{k+2}(\gamma)$$

for some positive integer k, and continuous functions $a_i(\gamma)$ $(i = 1, \ldots, k+2)$ and $g_i(\mathbf{x})$ $(i = 1, \ldots, k)$, where $1, \mathbf{x}, g_1(\mathbf{x}), \ldots, g_k(\mathbf{x})$ are linearly independent, and $a_j(\gamma)$ $(j = 1, \ldots, k)$ are not equal to the zero functions. If $\big(a_1(\gamma_1), \ldots, a_k(\gamma_1)\big) \neq \big(a_1(\gamma_2), \ldots, a_k(\gamma_2)\big)$ for any $\gamma_1 \neq \gamma_2$, then (α, β, γ) is identifiable.

S1.2 Empirical likelihood

Let $\theta = (\alpha, \beta^\top, \gamma, \xi^\top)^\top$ and $t(\mathbf{x}, \theta) = \alpha + \mathbf{x}^\top\beta + c(\mathbf{x}, \gamma, \xi)$. In the main paper, we showed that the profile log-likelihood of (η, θ) is

$$\ell(\eta, \theta) = \ell_1(\eta) + \ell_2(\theta), \tag{S1.5}$$

where

$$\ell_1(\eta) = n_1 \log(\eta) + n_2 \log(1 - \eta) \tag{S1.6}$$

and

$$\ell_2(\theta) \;=\; \sum_{i=1}^{n_1} \log\{f(y_i|\mathbf{x}_i, \xi)\} + \sum_{i=n_1+1}^{n} \{t(\mathbf{x}_i, \theta)\} - \sum_{i=1}^{n} \log\{1 + \lambda[\exp\{t(\mathbf{x}_i, \theta)\} - 1]\}$$

with λ being the solution to

$$\sum_{i=1}^{n} \frac{\exp\{t(\mathbf{x}_i, \theta)\} - 1}{1 + \lambda[\exp\{t(\mathbf{x}_i, \theta)\} - 1]} = 0. \tag{S1.7}$$

The MLE of (η, θ) is defined as

$$(\hat{\eta}, \hat{\theta}) = \arg\max_{\eta, \theta} \ell(\eta, \theta). \tag{S1.8}$$

Equivalently, $\hat{\eta}$ maximizes $\ell_1(\eta)$, which gives $\hat{\eta} = n_1/n$, and

$$\hat{\theta} = \arg\max_{\theta} \ell_2(\theta).$$

The likelihood ratio function of θ is defined as

$$R(\theta) = 2\{\max_{\eta,\theta} \ell(\eta, \theta) - \max_{\eta} \ell(\eta, \theta)\} = 2\{\ell_2(\hat{\theta}) - \ell_2(\theta)\}.$$

Further, denote the truth of (η, θ) by (θ_0, η_0) with $\theta_0 = (\alpha_0, \beta_0^\top, \gamma_0, \xi_0^\top)^\top$ and $\eta_0 \in (0, 1)$. We assume that the proposed models satisfy the following regularity conditions on $f(y|\mathbf{x}, \xi)$, which mimic those for the consistency and asymptotic normality of the MLE under a regular parametric model on pp. 144–145 of Serfling (1980).

Regularity Conditions:

(A1) In a neighbourhood of ξ_0, $\log\{f(y|\mathbf{x}, \xi)\}$ is three-times differentiable with respect to ξ for any $(y, \mathbf{x})$.

(A2) For (γ, ξ) in a neighbourhood of (γ_0, ξ_0) and any $\mathbf{x}$ on S, the inequality $\int e^{y\gamma} f(y|\mathbf{x}, \xi) dy < \infty$ holds.

(A3) The matrix V defined below in (S1.9) is well defined and nonsingular.

(A4) There exists a function $M(\mathbf{x})$ not depending on (γ, ξ) such that $\mathbb{E}\{M(\mathbf{X})\} < \infty$ and

$$\left\| \int e^{y\gamma} \nabla_\xi f(y|\mathbf{x}, \xi) dy \right\| + \left\| \int e^{y\gamma} \nabla_{\xi,\xi} f(y|\mathbf{x}, \xi) dy \right\| \\ + \left\| \int e^{y\gamma} \nabla_{\xi,\xi,\xi} f(y|\mathbf{x}, \xi) dy \right\| < M(\mathbf{x})$$

uniformly for (γ, ξ) in a neighbourhood of (γ_0, ξ_0) and a neighbourhood of $(0, \xi_0)$. Here $\nabla_\xi f(y|\mathbf{x}, \xi)$, $\nabla_{\xi,\xi} f(y|\mathbf{x}, \xi)$, and $\nabla_{\xi,\xi,\xi} f(y|\mathbf{x}, \xi)$ are the first, second, and third derivatives of $f(y|\mathbf{x}, \xi)$ with respect to ξ.

We now give the formal definition of the matrix V, which is closely related to the asymptotic variance matrix of $\hat{\theta}$. Define

$$\pi(\mathbf{x}; \theta, \eta) \quad = \quad \frac{(1-\eta)\exp\{t(\mathbf{x}, \theta)\}}{\eta + (1-\eta)\exp\{t(\mathbf{x}, \theta)\}}$$

and we write $\pi(\mathbf{x}) = \pi(\mathbf{x}; \theta_0, \eta_0)$ for abbreviation. Let d_θ denote the dimension of θ and $\mathbf{e}_1$ be a $d_\theta \times 1$ vector with the first component being 1 and the remaining components 0. Finally define

$$V = \mathbb{E}[\{1-\pi(\mathbf{X})\}\pi(\mathbf{X})\{\nabla_\theta t(\mathbf{X}, \theta)\}^{\otimes 2}] + \mathbb{E}[D I_e\{\nabla_\xi f(Y|\mathbf{X}, \xi)\}^{\otimes 2} I_e^\top], \quad \text{(S1.9)}$$

where ∇_θ is the differentiation operator with respect to θ, $I_e^\top = (0_{d_\xi \times (2+d_\beta)}, I_{d_\xi \times d_\xi})$, and $B^{\otimes 2} = BB^\top$ for any matrix or vector B.

Theorem 1 of the main paper. *Suppose Conditions (A1)–(A4) hold. Assume that the logistic regression model in* (S1.1) *holds with* $(\alpha_0, \beta_0, \gamma_0)$ *in place of* (α, β, γ), *and that the density function of* Y *given* $(\mathbf{X} = \mathbf{x}, D = 1)$ *is* $f(y|\mathbf{x}, \xi_0)$. *Further, assume that* θ *is identifiable. Then as* $n \to \infty$,

(1) $\sqrt{n}(\hat{\theta} - \theta_0) \to N\big(0,\ V^{-1} - \{\eta_0(1-\eta_0)\}^{-1}\mathbf{e}_1\mathbf{e}_1^\top\big)$ *in distribution with* V *defined in* (S1.9);

(2) $R(\theta_0) \to \chi^2_{d_\theta}$ *in distribution.*

In the main paper, we define the MLE of μ as

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \frac{\int_y y\{\hat{\eta} + (1-\hat{\eta})\exp(\hat{\alpha} + \mathbf{x}_i^\top\hat{\beta} + \hat{\gamma}y)\} f(y|\mathbf{x}_i, \hat{\xi})dy}{\hat{\eta} + (1-\hat{\eta})\exp\{\hat{\alpha} + \mathbf{x}_i^\top\hat{\beta} + c(\mathbf{x}_i, \hat{\gamma}, \hat{\xi})\}}. \quad \text{(S1.10)}$$

The next theorem studies the asymptotic normality of $\hat{\mu}$ in (S1.10).

Theorem 2 of the main paper. *Under the conditions of Theorem 1 of the main paper, as* n *goes to infinity,* $\sqrt{n}(\hat{\mu}-\mu) \to N(0,\sigma^2)$ *in distribution, where* $\sigma^2 = \mathbb{V}\mathrm{ar}\{K(\mathbf{X};\theta_0,\eta_0)\} + A^\top V^{-1} A$ *with* V *given in (S1.9),*

$$K(\mathbf{x};\theta,\eta) \quad = \quad \frac{\int y\{\eta + (1-\eta)\exp(\alpha + \mathbf{x}^\top\beta + \gamma y)\} f(y|\mathbf{x},\xi)dy}{\eta + (1-\eta)\exp\{\alpha + \mathbf{x}^\top\beta + c(\mathbf{x},\gamma,\xi)\}},$$

and $A = \mathbb{E}\left\{\nabla_\theta K(\mathbf{X};\theta_0,\eta_0)\right\}$.

Theorem 3 of the main paper. *Under the conditions of Theorem 1 of the main paper, the MLEs* $(\hat{\theta},\hat{\eta})$ *in* (S1.8) *and* $\hat{\mu}$ *in* (S1.10) *are both semiparametric efficient, in sense that their asymptotic variances both attains the semiparametric efficiency lower bounds.*

S2 Proof of Corollary 1 in the main paper

We first consider Part (a). Since z is an instrument variable, (S1.3) becomes

$$\mathrm{pr}(\mathbf{x}|D=0) \quad = \quad \exp\{\alpha + u^\top\beta + c(\mathbf{x},\gamma,\xi)\}\mathrm{pr}(\mathbf{x}|D=1).$$

The identification of (α,β,γ) is equivalent to the identification of (α,β,γ) in $\alpha + u^\top\beta + c(\mathbf{x},\gamma,\xi)$.

Recall that $f(y|\mathbf{x},\xi)=\mathrm{pr}(y|\mathbf{x},D=1)$. Then

$$\begin{aligned} f(y|\mathbf{x},\xi) &= \mathrm{pr}(y|\mathbf{x})\mathrm{pr}(D=1|\mathbf{x},y)\Big/\int \mathrm{pr}(y|\mathbf{x})\mathrm{pr}(D=1|\mathbf{x},y)dy \\ &= \mathrm{pr}(y|z,u)\mathrm{pr}(D=1|u,y)\Big/\int \mathrm{pr}(y|z,u)\mathrm{pr}(D=1|u,y)dy. \end{aligned}$$

Since z is an instrument variable, it follows that $f(y|\mathbf{x},\xi)$ must depend on z, and so must $c(\mathbf{x},\gamma,\xi)$. Suppose $(\alpha_1,\beta_1,\gamma_1)$ and $(\alpha_2,\beta_2,\gamma_2)$ satisfy

$$\alpha_1+u^\top\beta_1+c(\mathbf{x},\gamma_1,\xi)=\alpha_2+u^\top\beta_2+c(\mathbf{x},\gamma_2,\xi)$$

for all $\mathbf{x}$, which implies

$$c(\mathbf{x},\gamma_1,\xi_0)-c(\mathbf{x},\gamma_2,\xi_0)=(\alpha_2-\alpha_1)+u^\top(\beta_2-\beta_1).$$

Since the left-hand side depends on z, while the right-hand side does not, we must have $\gamma_1=\gamma_2$, which further implies that $\alpha_1=\alpha_2$ and $\beta_1=\beta_2$. This indicates that the parameters (α,β,γ) are identifiable, which completes the proof of Part (a).

We next consider Part (b). Suppose $(\alpha_1,\beta_1,\gamma_1)$ and $(\alpha_2,\beta_2,\gamma_2)$ satisfy

$$\alpha_1+\beta_1^\top\mathbf{x}+c(\mathbf{x},\gamma_1,\xi_0)=\alpha_2+\beta_2^\top\mathbf{x}+c(\mathbf{x},\gamma_2,\xi_0)$$

for all $\mathbf{x}\in S$. According to the expression for $c(\mathbf{x},\gamma,\xi_0)$, this implies that

$$\begin{aligned} & \sum_{i=1}^k a_i(\gamma_1)g_i(\mathbf{x})+\{\alpha_1+a_{k+1}(\gamma_1)\}+\mathbf{x}^\top\{\beta_1+a_{k+2}(\gamma_1)\} \\ = & \sum_{i=1}^k a_i(\gamma_2)g_i(\mathbf{x})+\{\alpha_2+a_{k+1}(\gamma_2)\}+\mathbf{x}^\top\{\beta_2+a_{k+2}(\gamma_2)\}. \end{aligned}$$

Since $1, \mathbf{x}, g_1(\mathbf{x}), \dots, g_k(\mathbf{x})$ are linearly independent, it follows that

$$\begin{aligned}\big(a_1(\gamma_1), \dots, a_k(\gamma_1)\big) &= \big(a_1(\gamma_2), \dots, a_k(\gamma_2)\big), \\ \alpha_1 + a_{k+1}(\gamma_1) &= \alpha_2 + a_{k+1}(\gamma_2), \\ \beta_1 + a_{k+2}(\gamma_1) &= \beta_2 + a_{k+2}(\gamma_2)\end{aligned}$$

hold simultaneously. Because $\big(a_1(\gamma_1), \dots, a_k(\gamma_1)\big) \neq \big(a_1(\gamma_2), \dots, a_k(\gamma_2)\big)$ for any $\gamma_1 \neq \gamma_2$, the first equation implies $\gamma_1 = \gamma_2$. Then the last two equations lead to $\alpha_1 = \alpha_2$ and $\beta_1 = \beta_2$. This completes the proof of Part (b) and that of Corollary 1 in the main paper.

S3 Preparation for proving Theorems 1–3 in the main paper

S3.1 Re-expression

It can be verified that

$$\ell_2(\theta) = h(\theta, \lambda_\theta),$$

where

$$\begin{aligned}h(\theta, \lambda) &= \sum_{i=1}^{n_1} \log\{f(y_i|\mathbf{x}_i, \xi)\} + \sum_{i=n_1+1}^{n} t(\mathbf{x}_i, \theta) \\ &\quad - \sum_{i=1}^{n} \log\{1 + \lambda[\exp\{t(\mathbf{x}_i, \theta)\} - 1]\} \qquad \text{(S3.1)}\end{aligned}$$

and λ_θ is the solution to $\nabla_\lambda h = 0$.

Let $\hat{\lambda}$ be the solution to (S1.7) with $\hat{\theta}$ in place of θ. We first discuss some properties of $\hat{\lambda}$. It can be verified that $(\hat{\theta}, \hat{\lambda})$ satisfy

$$\nabla_\alpha h(\hat{\theta}, \hat{\lambda}) = 0, \quad \nabla_\lambda h(\hat{\theta}, \hat{\lambda}) = 0.$$

Note that

$$\begin{aligned} \nabla_\lambda h(\theta, \lambda) &= -\sum_{i=1}^{n} \frac{\exp\{t(\mathbf{x}_i, \theta)\} - 1}{1 + \lambda[\exp\{t(\mathbf{x}_i, \theta)\} - 1]} = 0 \quad \text{and} \\ \nabla_\alpha h(\theta, \lambda) &= n_2 - \lambda \sum_{i=1}^{n} \frac{\exp\{t(\mathbf{x}_i, \theta)\}}{1 + \lambda[\exp\{t(\mathbf{x}_i, \theta)\} - 1]} = 0 \end{aligned}$$

together imply that

$$\hat{\lambda} = n_2/n, \tag{S3.2}$$

which converges in probability to $\lambda_0 = 1 - \eta_0$.

For convenience of presentation, let $\omega = (\theta^\top, \lambda)^\top$. It can be verified that $\hat{\omega} = (\hat{\theta}^\top, \hat{\lambda})^\top$ is the solution to $\partial h(\theta, \lambda)/\partial\omega = 0$. To investigate the asymptotic properties of $(\hat{\theta}, \hat{\lambda})$, we need their approximations, which can be obtained via the second-order Taylor expansion of $h(\theta, \lambda)$ around $\omega = \omega_0 \equiv (\theta_0^\top, \lambda_0)^\top$. In the next subsection, we derive the forms of $\partial h(\theta_0, \lambda_0)/\partial\omega$ and $\partial^2 h(\theta_0, \lambda_0)/(\partial\omega\partial\omega^\top)$ and study their properties.

S3.2 First and second derivatives of $h(\theta, \lambda)$ at $\omega = \omega_0$

For convenience of presentation, we write $\pi_i = \pi(\mathbf{x}_i)$. Denote

$$u_n = (u_{n1}, u_{n2}^\top)^\top , \tag{S3.3}$$

where

$$\begin{aligned} u_{n1} &= \nabla_\theta h(\theta_0, \lambda_0) = \sum_{i=1}^n \left[(1 - d_i - \pi_i) \nabla_\theta t(\mathbf{x}_i, \theta_0) + d_i I_e \nabla_\xi \log\{f(y_i|\mathbf{x}_i, \xi_0)\} \right] , \\ u_{n2} &= \nabla_\lambda h(\theta_0, \lambda_0) = \frac{1}{\lambda_0(1-\lambda_0)} \sum_{i=1}^n (\lambda_0 - \pi_i) . \end{aligned}$$

After some calculation, it can be verified that the second derivatives of $h(\theta, \lambda)$ at (θ_0, λ_0) are

$$\begin{aligned} \nabla_{\theta\theta^\top} h(\theta_0, \lambda_0) &= V_n = \sum_{i=1}^n d_i I_e \nabla_{\xi\xi} \log\{f(y_i|\mathbf{x}_i, \xi_0)\} I_e^\top \\ &\quad + \sum_{i=1}^n (1 - d_i - \pi_i) \nabla_{\theta\theta} t(\mathbf{x}_i, \theta_0) \\ &\quad - \sum_{i=1}^n \pi_i (1 - \pi_i) \{\nabla_\theta t(\mathbf{x}_i, \theta_0)\}^{\otimes 2}, \\ \nabla_{\theta\lambda} h(\theta_0, \lambda_0) &= \frac{1}{\lambda_0(1-\lambda_0)} V_n \mathbf{e}_1, \\ \nabla_{\lambda\lambda} h(\theta_0, \lambda_0) &= \frac{1}{\lambda_0^2(1-\lambda_0)^2} \sum_{i=1}^n (\lambda_0 - \pi_i)^2. \end{aligned}$$

S3.3 Some useful technical lemmas

When deriving the asymptotic distribution of $\hat{\theta}$, we need to use $\mathbb{E}\{\nabla_{\theta\theta^\top} h(\theta_0, \lambda_0)\}$, $\mathbb{E}\{\nabla_{\theta\lambda} h(\theta_0, \lambda_0)\}$, $\mathbb{E}\{\nabla_{\lambda\lambda} h(\theta_0, \lambda_0)\}$, and the expectation and variance of u_n

defined in (S3.3). We need the following lemma to simplify our calculation.

Lemma 1. *The following equations hold:*

$$\begin{aligned}
\mathbb{E}[d_i\nabla_{\xi}\log\{f(y_i|\mathbf{x}_i,\xi_0)\}] &= 0, & \text{(S3.4)}\\
\mathbb{E}[d_i\nabla_{\xi\xi}\log\{f(y_i|\mathbf{x}_i,\xi_0)\}] &= -\mathbb{E}\{d_i[\nabla_{\xi}\log\{f(y_i|\mathbf{x}_i,\xi_0)\}]^{\otimes 2}\}, & \text{(S3.5)}\\
-\frac{1}{n}\mathbb{E}\{\nabla_{\theta\theta^\top}h(\theta_0,\lambda_0)\} &= V, & \text{(S3.6)}\\
-\frac{1}{n}\mathbb{E}\{\nabla_{\theta\lambda}h(\theta_0,\lambda_0)\} &= \frac{1}{\lambda_0(1-\lambda_0)}V\mathbf{e}_1, & \text{(S3.7)}\\
-\frac{1}{n}\mathbb{E}\{\nabla_{\lambda\lambda}h(\theta_0,\lambda_0)\} &= \frac{\mathbf{e}_1^\top V\mathbf{e}_1-\lambda_0(1-\lambda_0)}{\lambda_0^2(1-\lambda_0)^2}. & \text{(S3.8)}
\end{aligned}$$

Proof. By the fact $f(y|\mathbf{x},\xi)=\mathrm{pr}(Y=y|\mathbf{X}=\mathbf{x},D=1)$, it can be verified that

$$\begin{aligned}
\mathbb{E}[\nabla_{\xi}\log\{f(y_i|\mathbf{x}_i,\xi_0)\}|\mathbf{x}_i,d_i=1] &= 0,\\
\mathbb{E}[\nabla_{\xi\xi}\log\{f(y_i|\mathbf{x}_i,\xi_0)\}|\mathbf{x}_i,d_i=1] &= -\mathbb{E}\{[\nabla_{\xi}\log\{f(y_i|\mathbf{x}_i,\xi_0)\}]^{\otimes 2}|\mathbf{x}_i,d_i=1\},
\end{aligned}$$

which imply respectively Equations(S3.4) and (S3.5) by conditioning on $(\mathbf{x}_i,d_i=1)$.

Equations (S3.6) and (S3.7) follows immediately from (S3.5). To prove (S3.8), by noticing

$$\lambda_0=1-\eta_0=\mathrm{pr}(D=0) \quad \text{and} \quad \pi(\mathbf{x})=\mathrm{pr}(D=0|\mathbf{x}),$$

we have $\mathbb{E}\{\pi(\mathbf{X})\}=\lambda_0$ and

$$\frac{1}{n}\mathbb{E}\{\nabla_{\lambda\lambda}h(\theta_0,\lambda_0)\} = \mathbb{E}\{\lambda_0-\pi(\mathbf{x}_i)\}^2=\mathbb{E}[\{\pi(\mathbf{X})\}^2]-\lambda_0^2. \quad \text{(S3.9)}$$

Since $\mathbf{e}_1^\top V \mathbf{e}_1 = \mathbb{E}[\pi(\mathbf{X})\{1-\pi(\mathbf{X})\}] = -\mathbb{E}[\{\pi(\mathbf{X})\}^2] + \lambda_0$, Equation (S3.8) follows by comparing this equation with (S3.9). This finishes the proof. □

The final lemma presents the expectation and variance of u_n.

Lemma 2. *With u_n defined in (S3.3), we have $\mathbb{E}(u_n) = 0$ and*

$$\frac{1}{n}\mathbb{V}\text{ar}(u_n) = U = \begin{pmatrix} V & 0 \\ 0 & \frac{-\mathbf{e}_1^\top V \mathbf{e}_1 + \lambda_0(1-\lambda_0)}{\lambda_0^2(1-\lambda_0)^2} \end{pmatrix}.$$

Proof. The result $\mathbb{E}(u_{n2}) = 0$ follows from $\pi(\mathbf{x}) = \text{pr}(D=0|\mathbf{X}=\mathbf{x})$ and Equation (S3.4).

For $\mathbb{V}\text{ar}(u_n)$, we first calculate $\mathbb{V}\text{ar}(u_{n2})$. It can be seen that

$$\frac{1}{n}\mathbb{V}\text{ar}(u_{n2}) \;=\; \frac{1}{\lambda_0^2(1-\lambda_0)^2}\mathbb{E}[\{\lambda_0 - \pi(\mathbf{x}_i)\}^2] = \frac{1}{\lambda_0^2(1-\lambda_0)^2}[\mathbb{E}\{\pi(\mathbf{x}_i)\}^2 - \lambda_0^2].$$

We have shown that $\mathbf{e}_1^\top V \mathbf{e}_1 = -\mathbb{E}[\{\pi(\mathbf{X})\}^2] + \lambda_0$ in the proof of Lemma 1. Therefore

$$\frac{1}{n}\mathbb{V}\text{ar}(u_{n2}) = \frac{-\mathbf{e}_1^\top V \mathbf{e}_1 + \lambda_0(1-\lambda_0)}{\lambda_0^2(1-\lambda_0)^2}.$$

It remains to calculate $\mathbb{V}\text{ar}(u_{n1})$. Re-write

$$u_{n1} = \sum_{i=1}^n (u_{n11,i} + u_{n12,i}),$$

where

$$\begin{aligned} u_{n11,i} &= (1 - d_i - \pi_i)\nabla_\theta t(\mathbf{x}_i, \theta_0) \\ u_{n12,i} &= d_i I_e \nabla_\xi \log\{f(y_i|\mathbf{x}_i, \xi_0)\}. \end{aligned}$$

Since both $u_{n11,i}$ and $u_{n12,i}$ have mean zero, it follows from equality (S3.4) that

$$\frac{1}{n}\mathbb{V}\mathrm{ar}(u_{n1}) \;=\; \mathbb{E}\{(u_{n11,i}+u_{n12,i})^{\otimes 2}\} = \mathbb{E}(u_{n11,i}^{\otimes 2}) + \mathbb{E}(u_{n12,i}^{\otimes 2}).$$

Because $\mathbb{E}(d_i = 0|\mathbf{x}_i) = \mathrm{pr}(D = 0|\mathbf{X} = \mathbf{x}_i) = \pi(\mathbf{x}_i)$, by conditioning on $\mathbf{x}_i$, we have

$$\begin{aligned}\mathbb{E}(u_{n11,i}^{\otimes 2}) &= \mathbb{E}[\{(1-d_i-\pi_i)\nabla_\theta t(\mathbf{x}_i,\theta_0)\}^{\otimes 2}] \\ &= \mathbb{E}[(\pi(\mathbf{x}_i)\{1-\pi(\mathbf{x}_i)\}\{\nabla_\theta t(\mathbf{x}_i,\theta_0)\}^{\otimes 2}].\end{aligned}$$

Clearly $\mathbb{E}(u_{n12,i}^2) = \mathbb{E}[d_i I_e \nabla_\xi \log\{f(y_i|\mathbf{x}_i,\xi_0)\}]^2$. This proves $\frac{1}{n}\mathbb{V}\mathrm{ar}(u_{n1}) = V$ by comparing the expression of V with $\frac{1}{n}\mathbb{V}\mathrm{ar}(u_{n1})$. □

S4 Proof of Theorem 1 in the main paper

We start with Part (a). Using a similar argument to that used in the proofs of Lemma 1 and Theorem 1 of Qin and Lawless (1994), we have $\hat\theta = \theta_0 + O_p(n^{-1/2})$ and $\hat\lambda - \lambda_0 = O_p(n^{-1/2})$. Next we investigate the asymptotic approximation of $\hat\theta$.

The maximum likelihood estimator $\hat\theta$ of θ and the associated Lagrange multiplier $\hat\lambda$ must satisfy

$$\begin{pmatrix}\nabla_\theta h(\hat\theta,\hat\lambda) \\ \nabla_\lambda h(\hat\theta,\hat\lambda)\end{pmatrix} = 0.$$

Applying a first-order expansion to the left-hand side of the above equation gives

$$0=\begin{pmatrix}\nabla_\theta h(\theta_0,\lambda_0)\\ \nabla_\lambda h(\theta_0,\lambda_0)\end{pmatrix}+\begin{pmatrix}\nabla_{\theta\theta^\top}h(\theta_0,\lambda_0) & \nabla_{\theta\lambda}h(\theta_0,\lambda_0)\\ \nabla_{\lambda\theta^\top}h(\theta_0,\lambda_0) & \nabla_{\lambda\lambda}h(\theta_0,\lambda_0)\end{pmatrix}\begin{pmatrix}\hat\theta-\theta_0\\ \hat\lambda-\lambda_0\end{pmatrix}+o_p(n^{1/2}). \tag{S4.10}$$

By Lemma 1,

$$\begin{pmatrix}\nabla_{\theta\theta^\top}h(\theta_0,\lambda_0) & \nabla_{\theta\lambda}h(\theta_0,\lambda_0)\\ \nabla_{\lambda\theta^\top}h(\theta_0,\lambda_0) & \nabla_{\lambda\lambda}h(\theta_0,\lambda_0)\end{pmatrix}=-nW+o_p(n), \tag{S4.11}$$

where

$$W=\begin{pmatrix}V & \frac{1}{\lambda_0(1-\lambda_0)}V\mathbf{e}_1\\ \frac{1}{\lambda_0(1-\lambda_0)}\mathbf{e}_1^\top V & \frac{\mathbf{e}_1^\top V\mathbf{e}_1-\lambda_0(1-\lambda_0)}{\lambda_0^2(1-\lambda_0)^2}\end{pmatrix}.$$

Recall that $u_n=(\nabla_\theta h(\theta_0,\lambda_0),\nabla_\lambda h(\theta_0,\lambda_0))$. Combining (S4.10) and (S4.11), we get

$$\begin{pmatrix}\hat\theta-\theta_0\\ \hat\lambda-\lambda_0\end{pmatrix}=\frac{1}{n}W^{-1}u_n+o_p(n^{-1/2}). \tag{S4.12}$$

Note that $|W|=-|V|\cdot|\lambda_0(1-\lambda_0)|=-|V|\cdot|\eta_0(1-\eta_0)|$ and we have assumed that $\eta_0\in(0,1)$ and V is nonsingular, the matrix W is nonsingular and its inverse W^{-1} is well defined. Since

$$W^{-1}=\begin{pmatrix}V^{-1}-\frac{1}{\lambda_0(1-\lambda_0)}\mathbf{e}_1\mathbf{e}_1^\top & \mathbf{e}_1\\ \mathbf{e}_1^\top & -\lambda_0(1-\lambda_0)\end{pmatrix}, \tag{S4.13}$$

we have

$$\hat{\theta} - \theta_0 = n^{-1} \left(\begin{array}{cc} V^{-1} - \frac{1}{\lambda_0(1-\lambda_0)}\mathbf{e}_1\mathbf{e}_1^\top & \mathbf{e}_1 \end{array} \right) u_n + o_p(n^{-1/2}). \qquad \text{(S4.14)}$$

With Lemma 2, we can verify that

$$\mathbb{V}\mathrm{ar}\left\{ n^{-1/2} \left(\begin{array}{cc} V^{-1} - \frac{1}{\lambda_0(1-\lambda_0)}\mathbf{e}_1\mathbf{e}_1^\top & \mathbf{e}_1 \end{array} \right) u_n \right\} = V^{-1} - \frac{1}{\lambda_0(1-\lambda_0)}\mathbf{e}_1\mathbf{e}_1^\top.$$

Note that u_n is the sum of independent and identically distributed random vectors. Hence,

$$\sqrt{n}(\hat{\theta} - \theta_0) \to N\Big(0, \quad V^{-1} - \frac{1}{\lambda_0(1-\lambda_0)}\mathbf{e}_1\mathbf{e}_1^\top\Big)$$

in distribution. This completes the proof of Part (a).

Next, we consider Part (b). Recall that $R(\theta) = 2\{\ell_2(\hat{\theta}) - \ell_2(\theta)\}$. Then $R(\theta_0) = 2\{h(\hat{\theta}, \hat{\lambda}) - h(\theta_0, \lambda_{\theta_0})\}$, where λ_{θ_0} is the solution to $\partial h(\theta_0, \lambda)/\partial\lambda = 0$.

Applying a second-order Taylor expansion to $h(\hat{\theta}, \hat{\lambda})$ and using (S4.12), we have

$$h(\hat{\theta}, \hat{\lambda}) = \frac{n}{2} u_n^\top W^{-1} u_n + o_p(1). \qquad \text{(S4.15)}$$

Following a similar argument to that for (S4.15), we get

$$h(\theta_0, \lambda_{\theta_0}) = -\frac{n}{2} u_{n2}^2 \frac{\lambda_0^2(1-\lambda_0)^2}{\lambda_0(1-\lambda_0) - \mathbf{e}_1^\top V \mathbf{e}_1} + o_p(1). \qquad \text{(S4.16)}$$

Combining (S4.15) and (S4.16) gives

$$R(\theta_0) = n u_n^\top \begin{pmatrix} V^{-1} - \frac{\mathbf{e}_1 \mathbf{e}_1^\top}{\lambda_0(1-\lambda_0)} & \mathbf{e}_1 \\ \mathbf{e}_1^\top & \frac{\lambda_0(1-\lambda_0)\mathbf{e}_1^\top V \mathbf{e}_1}{\lambda_0(1-\lambda_0) - \mathbf{e}_1^\top V \mathbf{e}_1} \end{pmatrix} u_n + o_p(1).$$

Since W^{-1} are invertible, the matrix $V^{-1} - \{\lambda_0(1-\lambda_0)\}^{-1}\mathbf{e}_1\mathbf{e}_1^\top$ is also invertible. Let

$$v_n = u_{n1} + \left[V^{-1} - \{\lambda_0(1-\lambda_0)\}^{-1}\mathbf{e}_1\mathbf{e}_1^\top\right]^{-1} \mathbf{e}_1 u_{n2}.$$

After some algebra, $R(\theta_0)$ can be written as

$$R(\theta_0) = n v_n^\top \left[V^{-1} - \{\lambda_0(1-\lambda_0)\}^{-1}\mathbf{e}_1\mathbf{e}_1^\top\right] v_n + o_p(1).$$

With Lemma 2, we can further verify that $\mathbb{E}(v_n) = 0$ and

$$\mathbb{V}\text{ar}\left(n^{-1/2} v_n\right) = V + \frac{V \mathbf{e}_1 \mathbf{e}_1^\top V}{\lambda_0(1-\lambda_0) - \mathbf{e}_1^\top V \mathbf{e}_1} = \left[V^{-1} - \{\lambda_0(1-\lambda_0)\}^{-1}\mathbf{e}_1\mathbf{e}_1^\top\right]^{-1}.$$

Hence, $R(\theta_0) \to \chi^2_{d_\theta}$ in distribution. This completes the proof of Theorem 1 in the main paper.

S5 Proof of Theorem 2 in the main paper

Recall that $\hat{\eta} = n_1/n = 1 - \hat{\lambda}$ with $\eta_0 = \mathrm{pr}(D = 1)$. Then $\hat{\mu}$ in (S1.10) can be rewritten as

$$\begin{aligned}
\hat{\mu} &= \sum_{i=1}^{n} \hat{p}_i \left[\int y\{\hat{\eta} + (1-\hat{\eta})\exp(\hat{\alpha} + \mathbf{x}_i^\top \hat{\beta} + \hat{\gamma} y)\} f(y|\mathbf{x}_i, \hat{\xi}) dy \right] \\
&= \frac{1}{n} \sum_{i=1}^{n} \frac{\int y\{\hat{\eta} + (1-\hat{\eta})\exp(\hat{\alpha} + \mathbf{x}_i^\top \hat{\beta} + \hat{\gamma} y)\} f(y|\mathbf{x}_i, \hat{\xi}) dy}{\hat{\eta} + (1-\hat{\eta})\exp\{\hat{\alpha} + \mathbf{x}_i^\top \hat{\beta} + c(\mathbf{x}_i, \hat{\gamma}, \hat{\xi})\}} \\
&= n^{-1} \sum_{i=1}^{n} K(\mathbf{x}_i; \hat{\theta}, \hat{\eta}),
\end{aligned}$$

where

$$K(\mathbf{x}; \theta, \eta) = \frac{\int y\{\eta + (1-\eta)\exp(\alpha + \mathbf{x}^\top \beta + \gamma y)\} f(y|\mathbf{x}, \xi) dy}{\eta + (1-\eta)\exp\{\alpha + \mathbf{x}^\top \beta + c(\mathbf{x}, \gamma, \xi)\}}.$$

Applying the first-order Taylor expansion and the law of large numbers, we have

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} K(\mathbf{x}_i; \theta_0, \eta_0) + A^\top(\hat{\theta} - \theta_0) - B\{(1-\hat{\eta}) - (1-\eta_0)\} + o_p(n^{-1/2}),$$

where $A = \mathbb{E}\{\nabla_\theta K(\mathbf{X}; \theta_0, \eta_0)\}$ and $B = \mathbb{E}\{\nabla_\eta K(\mathbf{X}; \theta_0, \eta_0)\}$. Hence, with Equation (S4.12) and $\hat{\eta} = 1 - \hat{\lambda}$, we have

$$\hat{\mu} - \mu = \frac{1}{n} \sum_{i=1}^{n} \{K(\mathbf{x}_i; \theta_0, \eta_0) - \mu\} + n^{-1}(A^\top, -B) W^{-1} u_n + o_p(n^{-1/2}).$$

We first argue that $E\{K(\mathbf{X}; \theta_0, \eta_0)\} = \mu$. By (S1.3), we have

$$\mathrm{pr}(\mathbf{x}) = \{\eta_0 + (1-\eta_0)\exp\{\alpha_0 + \mathbf{x}^\top \beta_0 + c(\mathbf{x}, \gamma_0, \xi_0)\}\mathrm{pr}(\mathbf{x}|D = 1). \quad \text{(S5.17)}$$

It then follows that

$$
\begin{aligned}
& \mathbb{E}\{K(\mathbf{X};\theta_0,\eta_0)\} \\
= & \int_{\mathbf{X}} \frac{\int_y y\{\eta_0+(1-\eta_0)\exp(\alpha_0+\mathbf{x}^\top\beta_0+\gamma_0 y)\}f(y|\mathbf{x},\xi_0)dy}{\eta_0+(1-\eta_0)\exp\{\alpha_0+\mathbf{x}^\top\beta_0+c(\mathbf{x},\gamma_0,\xi_0)\}}\mathrm{pr}(\mathbf{x})d\mathbf{x} \\
= & \int_{\mathbf{X}}\int_y y\{\eta_0+(1-\eta_0)\exp(\alpha_0+\mathbf{x}^\top\beta_0+\gamma_0 y)\}f(y|\mathbf{x},\xi_0)\mathrm{pr}(\mathbf{x}|D=1)dyd\mathbf{x} \\
= & \int_{\mathbf{X}}\int_y y\{\eta_0+(1-\eta_0)\exp(\alpha_0+\mathbf{x}^\top\beta_0+\gamma_0 y)\}\mathrm{pr}(y,\mathbf{x}|D=1)dyd\mathbf{x} \\
= & \int_{\mathbf{X},y} y\mathrm{pr}(y,\mathbf{x})dyd\mathbf{x}=\mu.
\end{aligned}
$$

After some calculus, we found that $-B=A^\top\mathbf{e}_1/\{(1-\eta_0)\eta_0\}$. With (S4.13), we have

$$
\begin{aligned}
& (A^\top,-B)W^{-1}u_n \\
= & A^\top\left(I,\frac{\mathbf{e}_1}{\lambda_0(1-\lambda_0)}\right)\begin{pmatrix} V^{-1}-\frac{1}{\lambda_0(1-\lambda_0)}\mathbf{e}_1\mathbf{e}_1^\top & \mathbf{e}_1 \\ \mathbf{e}_1^\top & -\lambda_0(1-\lambda_0)\end{pmatrix}\begin{pmatrix} u_{n1} \\ u_{n2}\end{pmatrix} \\
= & A^\top V^{-1}u_{n1}.
\end{aligned}
$$

Since $\mathbb{E}(u_{n1}|\mathbf{x}_1,\ldots,\mathbf{x}_n)=0$, we arrive at

$$
\begin{aligned}
& \mathbb{C}\mathrm{ov}\left(\sum_{i=1}^n K(\mathbf{x}_i;\theta_0,\eta_0),(A^\top,-B)W^{-1}u_n\right) \\
= & \mathbb{C}\mathrm{ov}\left(\sum_{i=1}^n K(\mathbf{x}_i;\theta_0,\eta_0),A^\top V^{-1}u_{n1}\right) \\
= & 0.
\end{aligned}
$$

Finally, By Lemma 2 and the central limit theorem and Slutsky's theorem, we have

$$\sqrt{n}(\hat{\mu} - \mu) \to N\left(0, \sigma^2\right),$$

where $\sigma^2 = \mathbb{V}\mathrm{ar}\{K(\mathbf{X}; \theta_0, \eta_0)\} + A^\top V^{-1} A$. This proves Theorem 2 of the main paper.

S6 Proof of Theorem 3 in the main paper

S6.1 Preparations

The observed data are $(d_i = 1, \mathbf{x}_i, y_i)$ $(i = 1, \ldots, n_1)$ and $(d_i = 0, \mathbf{x}_i)$ $(i = n_1 + 1, \ldots, n)$. We make two parametric assumptions:

$$\begin{aligned} \mathrm{pr}(D = 1 | \mathbf{X} = \mathbf{x}, Y = y) &= \frac{1}{1 + \exp\{\alpha^* + \mathbf{x}^\top \beta + \gamma y\}}, \\ \mathrm{pr}(Y = y | \mathbf{X} = \mathbf{x}, D = 1) &= f(y | \mathbf{x}, \xi). \end{aligned}$$

Recall that $\eta = \mathrm{pr}(D = 1)$ and $c(\mathbf{x}, \gamma, \xi) = \log \int e^{\gamma y} f(y|\mathbf{x}, \xi) dy$. Let $\vartheta = (\beta, \gamma, \xi)$ and $r(\mathbf{x}, \vartheta) = \mathbf{x}^\top \beta + c(\mathbf{x}, \gamma, \xi)$, so that $\theta = (\alpha, \vartheta^\top)^\top$ and $t(\mathbf{x}, \theta) = \alpha + r(\mathbf{x}, \vartheta)$. We have shown

$$\begin{aligned} \mathrm{pr}(y, \mathbf{x} | D = 0) &= \exp\{\alpha + \mathbf{x}^\top \beta + \gamma y\} \mathrm{pr}(y, \mathbf{x} | D = 1), \\ \mathrm{pr}(\mathbf{X} = \mathbf{x} | D = 0) &= \exp\{\alpha + r(\mathbf{x}, \vartheta)\} \mathrm{pr}(\mathbf{X} = \mathbf{x} | D = 1), \end{aligned}$$

where $\alpha = \alpha^* + \log\{\eta/(1 - \eta)\}$.

In addition,

$$\mathbb{E}(1-D|\mathbf{X}=\mathbf{x}) = \text{pr}(D=0|\mathbf{X}=\mathbf{x}) = \pi(\mathbf{x};\alpha,\vartheta,\eta),$$

where we have defined

$$\pi(\mathbf{x};\alpha,\vartheta,\eta) = \frac{(1-\eta)\exp\{t(\mathbf{x},\theta)\}}{\eta+(1-\eta)\exp\{t(\mathbf{x},\theta)\}} = \frac{(1-\eta)\exp\{\alpha+r(\mathbf{x},\vartheta)\}}{\eta+(1-\eta)\exp\{\alpha+r(\mathbf{x},\vartheta)\}}$$

with $\pi(\mathbf{x})$ abbreviation for $\pi(\mathbf{x};\theta_0,\eta_0)$.

The observed data are iid from $(D,\mathbf{X},\tilde{Y})$, where $\tilde{Y}$ is empty when $D=0$, and $\tilde{Y}=Y$ when $D=1$. The joint distribution of $(D,\mathbf{X},\tilde{Y})$ is

$$\begin{aligned}
& \{\text{pr}(Y=y|\mathbf{X}=\mathbf{x},D=1)\text{pr}(\mathbf{X}=\mathbf{x}|D=1)\text{pr}(D=1)\}^d \\
& \qquad \times\{\text{pr}(\mathbf{X}=\mathbf{x}|D=0)\text{pr}(D=0)\}^{1-d} \\
= & \{\text{pr}(Y=y|\mathbf{X}=\mathbf{x},D=1)\text{pr}(D=1)\}^d \\
& \qquad \times\{\exp(t(\mathbf{x},\theta))\text{pr}(D=0)\}^{1-d}\times\text{pr}(\mathbf{x}|D=1) \\
= & \{f(y|\mathbf{x},\xi)\eta\}^d\times\{\exp(\alpha+r(\mathbf{x},\vartheta))(1-\eta)\}^{1-d}\times\text{pr}(\mathbf{x}|D=1).
\end{aligned}$$

Here all except $\text{pr}(\mathbf{X}=\mathbf{x}|D=1)$ are completely parametric, and we regard $\text{pr}(\mathbf{X}=\mathbf{x}|D=1)$ as an infinite-dimensional parameter, or simply

$$\text{pr}(\mathbf{X}=\mathbf{x}|D=1)\geq 0, \qquad \int \text{pr}(\mathbf{X}=\mathbf{x}|D=1)d\mathbf{x}=1.$$

Therefore our imposed model is clearly semi-parametric.

Throughout this section, we use $g(\mathbf{x},\zeta)$ to denote a parametric submodel for $\text{pr}(\mathbf{X}=\mathbf{x}|D=1)$ with $g(\mathbf{x},\zeta_0)$ being the true model. The joint density function of $(D, \mathbf{X}, \tilde{Y})$ is

$$h(d, \mathbf{x}, y; \alpha, \vartheta, \eta, \zeta) \;=\; \{f(y|\mathbf{x}, \xi)\eta\}^d \times \{\exp(\alpha + r(\mathbf{x}, \vartheta))(1-\eta)\}^{1-d} \times g(\mathbf{x}, \zeta).$$

It is worth noting that α is not free but is a function of (ϑ, ζ) and determined by

$$1 = \int \exp\{\alpha + r(\mathbf{x}, \vartheta)\} g(\mathbf{x}, \zeta) d\mathbf{x}. \tag{S6.18}$$

The following three functions will be useful in our proof:

$$\begin{aligned}
B_1(d, \mathbf{x}, y) &= \frac{\partial \log h(d, \mathbf{x}, y; \alpha_0, \vartheta_0, \eta_0, \zeta_0)}{\partial \vartheta} \\
&= (1-d)\{\nabla_\vartheta \alpha(\vartheta_0) + \nabla_\vartheta r(\mathbf{x}, \vartheta_0)\} + d I_{e,-1} \nabla_\xi \log f(y|\mathbf{x}, \xi_0), \\
B_2(d, \mathbf{x}, y) &= \frac{\partial \log h(d, \mathbf{x}, y; \alpha_0, \vartheta_0, \eta_0, \zeta_0)}{\partial \eta} = \frac{D - \eta_0}{\eta_0(1-\eta_0)}, \\
B_3(d, \mathbf{x}, y) &= \frac{\partial \log h(d, \mathbf{x}, y; \alpha_0, \vartheta_0, \eta_0, \zeta_0)}{\partial \zeta} = \nabla_\zeta \log g(\mathbf{x}, \zeta_0),
\end{aligned}$$

where $I_{e,-1}$ is I_e without the first row.

S6.2 Semiparametric efficiency of $(\hat{\theta}, \hat{\eta})$

We have shown that

$$\begin{aligned}
\hat{\theta} - \theta_0 &= n^{-1} \begin{pmatrix} V^{-1} - \tfrac{1}{\lambda_0(1-\lambda_0)} \mathbf{e}_1 \mathbf{e}_1^\top & \mathbf{e}_1 \end{pmatrix} u_n + o_p(n^{-1/2}), \\
\hat{\eta} - \eta_0 &= \frac{1}{n} \sum_{i=1}^n (d_i - \eta_0),
\end{aligned}$$

where $u_n = (u_{n1}^\top, u_{n2})^\top$ with

$$\begin{aligned} u_{n1} &= \sum_{i=1}^n \left[(1-d_i-\pi_i)\nabla_\theta t(\mathbf{x}_i,\theta_0) + d_i I_e \nabla_\xi \log\{f(y_i|\mathbf{x}_i,\xi_0)\}\right], \\ u_{n2} &= \frac{1}{\lambda_0(1-\lambda_0)}\sum_{i=1}^n (\lambda_0 - \pi_i). \end{aligned}$$

Therefore

$$\begin{aligned} \hat\theta - \theta_0 &= n^{-1}\{V^{-1}u_{n1} - \frac{1}{\lambda_0(1-\lambda_0)}\mathbf{e}_1\mathbf{e}_1^\top u_{n1} + \mathbf{e}_1 u_{n2}\} + o_p(n^{-1/2}) \\ &= n^{-1}\{V^{-1}u_{n1} + \frac{1}{\eta_0(1-\eta_0)}\mathbf{e}_1\sum_{i=1}^n (d_i-\eta_0)\} + o_p(n^{-1/2}) \\ &= n^{-1}\sum_{i=1}^n \Big[V^{-1}(1-d_i-\pi_i)\nabla_\theta t(\mathbf{x}_i,\theta_0) + V^{-1}d_i I_e\nabla_\xi \log\{f(y_i|\mathbf{x}_i,\xi_0)\} \\ &\quad + \frac{1}{\eta_0(1-\eta_0)}\mathbf{e}_1(d_i-\eta_0)\}\Big] + o_p(n^{-1/2}). \end{aligned}$$

Then the respective influence functions of $\hat\theta$ and $\hat\eta$ are

$$\begin{aligned} \varphi_\theta(D,\mathbf{X},Y) &= V^{-1}(1-D-\pi(\mathbf{X}))\nabla_\theta t(\mathbf{X},\theta_0) + V^{-1}DI_e\nabla_\xi \log\{f(Y|\mathbf{X},\xi_0)\} \\ &\quad + \frac{(D-\eta_0)}{\eta_0(1-\eta_0)}\mathbf{e}_1 \end{aligned}$$

and $\varphi_\eta(D,\mathbf{X},Y) = D-\eta_0$. We prove only the semiparametric efficiency of $\hat\theta$; the semiparametric efficiency of $\hat\eta$ can be proved in the same way with less algebra.

Referring to the established theory for the semiparametric efficiency bound, for example Chapter 3 of Bickel et al (1992) and Newey (1990), we need to show only the following two results to establish the semiparametric efficiency of $\hat\theta$:

(a) $\hat{\theta}$ is a regular estimator of θ_0;

(b) there exists a parametric submodel with $h_\psi(d, \mathbf{x}, \tilde{y})$ the joint density of $(D, \mathbf{X}, \tilde{Y})$ such that the true model is $h_0(d, \mathbf{x}, \tilde{y})$ and

$$\varphi_\theta(d, \mathbf{x}, y) = \frac{\partial \log h_\psi(d, \mathbf{x}, \tilde{y})}{\partial \psi}\Big|_{\psi=0}.$$

Proof of (a)

By Theorem 2 in Newey (1990), arguing $\hat{\theta}$ is a regular estimator of θ_0 is equivalent to showing that

$$\begin{aligned}
Z_1 &\equiv \mathbb{E}\{\varphi_\theta(D, \mathbf{X}, Y) B_1^\top(D, \mathbf{X}, Y)\} \\
&= \frac{\partial \theta}{\partial \vartheta^\top}\Big|_{(\theta_0, \eta_0, \zeta_0)} = (\nabla_\vartheta \alpha,\ I_{d_\vartheta})^\top, && \text{(S6.19)} \\
Z_2 &\equiv \mathbb{E}\{\varphi_\theta(D, \mathbf{X}, Y) B_2(D, \mathbf{X}, Y)\} = \frac{\partial \theta}{\partial \eta}\Big|_{(\theta_0, \eta_0, \zeta_0)} = 0, && \text{(S6.20)} \\
Z_3 &\equiv \mathbb{E}\{\varphi_\theta(D, \mathbf{X}, Y) B_3^\top(D, \mathbf{X}, Y)\} = \frac{\partial \theta}{\partial \zeta^\top}\Big|_{(\theta_0, \eta_0, \zeta_0)} = 0, && \text{(S6.21)}
\end{aligned}$$

where throughout this section $\mathbb{E}$ takes expectation with respect to $h(d, \mathbf{x}, y; \theta_0, \eta_0, \zeta_0)$.

(1) Proof of Equality (S6.19)

Since $\mathbb{E}\{D\nabla_\xi \log f(Y|\mathbf{X}, \xi_0)|\mathbf{X}\} = 0$, it follows that

$$\begin{aligned}
Z_1 &= \mathbb{E}\{\varphi_\theta(D, \mathbf{X}, Y) B_1^\top(D, \mathbf{X}, Y)\} \\
&= \mathbb{E}[(1 - D - \pi(\mathbf{X}))V^{-1}\nabla_\theta t(\mathbf{X}, \theta_0)(1 - D)\{\nabla_\vartheta \alpha(\vartheta_0) + \nabla_\vartheta r(\mathbf{X}, \theta_0)\}^\top] \\
&\quad + \mathbb{E}[(1 - D) r(\mathbf{X}, \theta_0)\frac{(D - \eta_0)}{\eta_0(1 - \eta_0)}\mathbf{e}_1\{\nabla_\vartheta \alpha(\vartheta_0) + \nabla_\vartheta\}^\top] \\
&\quad + \mathbb{E}[DV^{-1} I_e \nabla_\xi\{\log f(Y|\mathbf{X}, \xi_0)\}^{\otimes 2} I_{e,-1}^\top]
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}[\pi(\mathbf{X})(1-\pi(\mathbf{X}))V^{-1}\nabla_\theta t(\mathbf{X},\theta_0)\{\nabla_\vartheta\alpha(\vartheta_0)+\nabla_\vartheta r(\mathbf{X},\theta_0)\}^\top] \\
&\quad -\mathbb{E}[\frac{(1-D)\eta_0}{\eta_0(1-\eta_0)}\mathbf{e}_1\{\nabla_\vartheta\alpha(\vartheta_0)+\nabla_\vartheta r(\mathbf{X},\theta_0)\}^\top\}] \\
&\quad +\mathbb{E}[DI_e\nabla_\xi\{\log f(Y|\mathbf{X},\xi_0)\}^{\otimes 2}I_{e,-1}^\top V^{-1}] \\
&= \mathbf{e}_1\{\nabla_\vartheta\alpha(\vartheta_0)\}^\top+\mathbb{E}[\frac{\pi(\mathbf{X})}{1-\eta_0}\mathbf{e}_1\{\nabla_\vartheta\alpha(\vartheta_0)+\nabla_\vartheta r(\mathbf{X},\theta_0)\}^\top\}]+I_eI_{e,-1}^\top,
\end{aligned}
$$

where we have used the definition

$$V=\mathbb{E}[\{1-\pi(\mathbf{X})\}\pi(\mathbf{X})\{\nabla_\theta t(\mathbf{X},\theta)\}^{\otimes 2}]+\mathbb{E}[DI_e\{\nabla_\xi f(Y|\mathbf{X},\xi)\}^{\otimes 2}I_e^\top].$$

Taking derivative with respect to ϑ on both sides of (S6.18) gives

$$0 = \int\{\nabla_\vartheta\alpha(\vartheta_0)+\nabla_\vartheta r(\mathbf{x},\vartheta_0)\}\exp\{\alpha(\vartheta_0)+r(\mathbf{x},\vartheta_0)\}g(\mathbf{x},\zeta_0)d\mathbf{x}.$$

This together with $g(\mathbf{x},\zeta_0)d\mathbf{x}=dF(\mathbf{x}|D=1)$ leads to

$$
\begin{aligned}
&\frac{1}{1-\eta_0}\mathbf{e}_1^\top\mathbb{E}[\{\nabla_\vartheta\alpha(\vartheta_0)+\nabla_\vartheta r(\mathbf{X},\theta_0)\}\pi(\mathbf{X})\}] \\
=\ &\frac{\eta_0}{1-\eta_0}\mathbf{e}_1^\top\int\{\nabla_\vartheta\alpha(\vartheta_0)+\nabla_\vartheta r(\mathbf{x},\vartheta_0)\}\exp\{\alpha(\vartheta_0)+r(\mathbf{x},\vartheta_0)\}g(\mathbf{x},\zeta_0)d\mathbf{x} \\
=\ &0.
\end{aligned}
$$

Therefore, we have

$$Z_1 = \nabla_\vartheta\alpha(\vartheta_0)e_1^\top+I_eI_{e,-1}^\top=(\nabla_\vartheta\alpha,I_{d_\vartheta}^\top)^\top.$$

This proves (S6.19).

(2) Proof of Equality (S6.20)

Since $\mathbb{E}\varphi_\theta(D,\mathbf{X},Y)=0$, we have

$$\begin{aligned}
Z_2 &= \mathbb{E}\{B_2(D,\mathbf{X},Y)\varphi_\theta(D,\mathbf{X},Y)\}\\
&= \frac{1}{\eta_0(1-\eta_0)}\mathbb{E}\{D\varphi_\theta(D,\mathbf{X},Y)\}\\
&= \frac{1}{\eta_0(1-\eta_0)}\mathbb{E}\{-DV^{-1}\pi(\mathbf{X})\nabla_\theta t(\mathbf{X},\theta_0)+D\frac{1}{\eta_0(1-\eta_0)}\mathbf{e}_1(1-\eta_0)\}\\
&\quad +DV^{-1}I_e\nabla_\xi\log\{f(Y|\mathbf{X},\xi_0)\}\}\\
&= \frac{1}{\eta_0(1-\eta_0)}\mathbb{E}\{-V^{-1}(1-\pi(\mathbf{X}))\pi(\mathbf{X})\nabla_\theta t(\mathbf{X},\theta_0)+\mathbf{e}_1\}\\
&= 0,
\end{aligned}$$

where the last equality holds because

$$\mathbb{E}\{\pi(\mathbf{X})(1-\pi(\mathbf{X}))\nabla_\theta t(\mathbf{X},\theta_0)\}=Ve_1.$$

This proves Equality (S6.20).

(3) Proof of Equality (S6.21)

Since

$$\mathbb{E}\{\varphi_\theta(D,\mathbf{X},Y)|\mathbf{x}\}=\frac{1}{\eta_0(1-\eta_0)}\mathbf{e}_1\{1-\eta_0-\pi(\mathbf{X})\},$$

we have

$$\begin{aligned}
Z_3 &= \mathbb{E}\{\varphi_\theta(D,\mathbf{X},Y)B_3^\top(D,\mathbf{X},Y)\}\\
&= \mathbb{E}[\frac{1}{\eta_0(1-\eta_0)}\mathbf{e}_1\{1-\eta_0-\pi(\mathbf{X})\}\nabla_{\zeta^\top}\log g(\mathbf{X},\zeta_0)]\\
&= -\frac{1}{\eta_0(1-\eta_0)}\mathbb{E}[\mathbf{e}_1\pi(\mathbf{X})\nabla_{\zeta^\top}\log g(\mathbf{X},\zeta_0)].
\end{aligned}$$

Taking derivative with respect to ζ on both sides of Eq (S6.18) gives

$$\begin{aligned} 0 &= \int \exp\{\alpha_0 + r(\mathbf{x}, \vartheta_0)\}\{\nabla_\zeta \log g(\mathbf{x}, \zeta_0)\} g(\mathbf{x}, \zeta_0) d\mathbf{x} \\ &= \int \exp\{\alpha_0 + r(\mathbf{x}, \vartheta_0)\}\{\nabla_\zeta \log g(\mathbf{x}, \zeta_0)\} \frac{1 - \pi(\mathbf{x})}{\eta_0} \mathrm{pr}(\mathbf{x}) d\mathbf{x} \\ &= \int \{\nabla_\zeta \log g(\mathbf{x}, \zeta_0)\} \frac{\pi(\mathbf{x})}{\eta_0(1 - \eta_0)} \mathrm{pr}(\mathbf{x}) d\mathbf{x} \\ &= \mathbb{E}[\{\nabla_\zeta \log g(\mathbf{X}, \zeta_0)\} \frac{\pi(\mathbf{X})}{\eta_0(1 - \eta_0)}], \end{aligned}$$

which means $Z_3 = 0$. This proves Equality (S6.21) and also completes the proof of (a).

Proof of (b)

Consider the following function

$$\begin{aligned} h_\psi(d, \mathbf{x}, \tilde{y}) &= \{1 + \psi \varphi_\theta(d, \mathbf{x}, y)\} \times \{f(y|\mathbf{x}, \xi_0)\eta_0\}^d \\ & \quad \times \{\exp(\alpha_0 + r(\mathbf{x}, \vartheta_0))(1 - \eta_0)\}^{1-d} g(\mathbf{x}, \zeta_0). \end{aligned}$$

Suppose the support of $(\mathbf{X}, Y)$ is compact, then it can be verified that the function

$$\begin{aligned} \varphi_\theta(D, \mathbf{X}, Y) &= V^{-1}(1 - D - \pi(\mathbf{X}))\nabla_\theta t(\mathbf{X}, \theta_0) + \frac{1}{\eta_0(1 - \eta_0)} \mathbf{e}_1 (D - \eta_0) \\ & \quad + V^{-1} D I_e \nabla_\xi \log\{f(Y|\mathbf{X}, \xi_0)\} \end{aligned}$$

is bounded. Because $\mathbb{E}\{\varphi_\theta(D, \mathbf{X}, Y)\} = 0$ where $\mathbb{E}$ takes expectation with respect to $h(d, \mathbf{x}, y; \alpha_0, \vartheta_0, \eta_0, \zeta_0)$, the function $h_\psi(d, \mathbf{x}, \tilde{y})$ is a density function when ψ is small enough. When $\psi = 0$, it reduces to the true joint density function $h(d, \mathbf{x}, y; \alpha_0, \vartheta_0, \eta_0, \zeta_0)$. It is easy to check that $h_\psi(d, \mathbf{x}, \tilde{y})$ with small enough ψ is a parametric submodel and

$$\nabla_\psi h_\psi(d, \mathbf{x}, \tilde{y})\Big|_{\psi=0} = \varphi_\theta(d, \mathbf{x}, y).$$

This proves (b), and hence proves the semiparametric efficiency of $\hat{\theta}$.

S6.3 Semiparametric efficiency of $\hat{\mu}$

The population mean can be expressed as

$$\begin{aligned}
\mu &= \int_y \int_{\mathbf{X}} y\mathrm{pr}(y|\mathbf{x}, D=1)\mathrm{pr}(\mathbf{x}|D=1)\mathrm{pr}(D=1)d\mathbf{x}dy \\
&\quad + \int_y \int_{\mathbf{X}} y\mathrm{pr}(y|\mathbf{x}, D=0)\mathrm{pr}(\mathbf{x}|D=0)\mathrm{pr}(D=0)d\mathbf{x}dy \\
&= \int_y \int_{\mathbf{X}} y\mathrm{pr}(y|\mathbf{x}, D=1)\mathrm{pr}(\mathbf{x}|D=1)\eta d\mathbf{x}dy \\
&\quad + \int_y \int_{\mathbf{X}} y\exp(\alpha + \mathbf{x}^\top\beta + \gamma y)\mathrm{pr}(y|\mathbf{x}, D=1)\mathrm{pr}(\mathbf{x}|D=1)(1-\eta)d\mathbf{x}dy \\
&= \int_{\mathbf{X}} \left[\int_y y\{\eta + (1-\eta)\exp(\alpha + \mathbf{x}^\top\beta + \gamma y)\}f(y|\mathbf{x}, \xi)dy\right] dF(\mathbf{x}|D=1).
\end{aligned}$$

The proposed mean estimator is

$$\begin{aligned}
\hat{\mu} &= \frac{1}{n}\sum_{i=1}^n \frac{\int_y y\{\hat{\eta} + (1-\hat{\eta})\exp(\hat{\alpha} + \mathbf{x}_i^\top\hat{\beta} + \hat{\gamma}y)\}f(y|\mathbf{x}_i, \hat{\xi})dy}{\hat{\eta} + (1-\hat{\eta})\exp\{\hat{\alpha} + \mathbf{x}_i^\top\hat{\beta} + c(\mathbf{x}_i, \hat{\gamma}, \hat{\xi})\}} \\
&= n^{-1}\sum_{i=1}^n K(\mathbf{x}_i; \hat{\theta}, \hat{\eta}),
\end{aligned}$$

where

$$K(\mathbf{x}; \theta, \eta) = \frac{\int y\{\eta + (1-\eta)\exp(\alpha + \mathbf{x}^\top\beta + \gamma y)\}f(y|\mathbf{x}, \xi)dy}{\eta + (1-\eta)\exp\{\alpha + \mathbf{x}^\top\beta + c(\mathbf{x}, \gamma, \xi)\}}.$$

Recall that $A = \mathbb{E}\left\{\nabla_\theta K(\mathbf{X};\theta_0,\eta_0)\right\}$, $\pi_i = \pi(\mathbf{x}_i)$ and $I_e^\top = (0_{d_\xi \times (2+d_\beta)}, I_{d_\xi \times d_\xi})$. We have shown in the proof of Theorem 2 that

$$\hat{\mu} - \mu = \frac{1}{n}\sum_{i=1}^n \{K(\mathbf{x}_i;\theta_0,\eta_0) - \mu\} + n^{-1}A^\top V^{-1} u_{n1} + o_p(n^{-1/2}),$$

where $u_{n1} = \sum_{i=1}^n \left[(1-d_i-\pi_i)\nabla_\theta t(\mathbf{x}_i,\theta_0) + d_i I_e \nabla_\xi \log\{f(y_i|\mathbf{x}_i,\xi_0)\}\right]$. Equivalently

$$\begin{aligned}\hat{\mu} - \mu &= \frac{1}{n}\sum_{i=1}^n \{K(\mathbf{x}_i;\theta_0,\eta_0) - \mu_0 + (1-d_i-\pi_i)A^\top V^{-1}\nabla_\theta t(\mathbf{x}_i,\theta_0) \\ &\quad + d_i A^\top V^{-1} I_e \nabla_\xi \log f(y_i|\mathbf{x}_i,\xi_0)\} + o_p(n^{-1/2}),\end{aligned}$$

which implies that the influence function of $\hat{\mu}$ is

$$\begin{aligned}\varphi_\mu(D,\mathbf{X},Y) &= K(\mathbf{X};\theta_0,\eta_0) - \mu_0 + \{1-D-\pi(\mathbf{X})\}A^\top V^{-1}\nabla_\theta t(\mathbf{X},\theta_0) \\ &\quad + DA^\top V^{-1} I_e \nabla_\xi \log f(Y|\mathbf{X},\xi_0).\end{aligned}$$

Similar to the proof of the semiparametric efficiency of $\hat{\theta}$, we need to show only the following two results to establish the semiparametric efficiency of $\hat{\mu}$:

(a1) $\hat{\mu}$ is a regular estimator of μ_0;

(b1) there exists a parametric submodel with $h_\psi^*(d,\mathbf{x},\tilde{y})$ the joint density of $(D,\mathbf{X},\tilde{Y})$ such that the true model is $h_0^*(d,\mathbf{x},\tilde{y})$ and

$$\varphi_\mu(d,\mathbf{x},y) = \left.\frac{\partial \log h_\psi^*(d,\mathbf{x},\tilde{y})}{\partial \psi}\right|_{\psi=0}.$$

Proof of (a1)

Under the submodel $g(\mathbf{x}, \zeta)$ for $\mathrm{pr}(\mathbf{X} = \mathbf{x}|D = 1)$, we can write μ as

$$\mu = \mu(\theta, \eta, \zeta) \equiv \int_{\mathbf{X}} \left[\int_y y\{\eta + (1-\eta)\exp(\alpha + \mathbf{x}^\top \beta + \gamma y)\} f(y|\mathbf{x}, \xi) dy \right] g(\mathbf{x}, \zeta) d\mathbf{x}.$$

Define $w(\mathbf{x}, y) = (\mathbf{x}^\top, y, \nabla_{\xi^\top} \log f(y|\mathbf{x}, \xi_0))^\top$. The partial derivative of μ is

$$\begin{aligned} \nabla_\vartheta \mu(\theta_0, \eta_0) &= \int_{\mathbf{X}} \int_y y\{\eta_0 + (1-\eta_0)\exp(\alpha_0 + \mathbf{x}^\top \beta_0 + \gamma_0 y)\} \\ &\quad \times\{\nabla_\vartheta \alpha(\vartheta_0) + w(\mathbf{x}, y)\} f(y|\mathbf{x}, \xi_0) dy g(\mathbf{x}, \zeta) d\mathbf{x}. \end{aligned}$$

By Theorem 2 in Newey (1990), arguing $\hat{\mu}$ is a regular estimator of μ_0 is equivalent to showing that

$$C_1 \equiv \mathbb{E}\{\varphi_\mu(D, \mathbf{X}, Y) B_1(D, \mathbf{X}, Y)\} = \frac{\partial \mu(\theta_0, \eta_0, \zeta_0)}{\partial \vartheta}, \quad \text{(S6.22)}$$

$$C_2 \equiv \mathbb{E}\{\varphi_\mu(D, \mathbf{X}, Y) B_2(D, \mathbf{X}, Y)\} = \frac{\partial \mu(\theta_0, \eta_0, \zeta_0)}{\partial \eta}, \quad \text{(S6.23)}$$

$$C_3 \equiv \mathbb{E}\{\varphi_\mu(D, \mathbf{X}, Y) B_3(D, \mathbf{X}, Y)\} = \frac{\partial \mu(\theta_0, \eta_0, \zeta_0)}{\partial \zeta}, \quad \text{(S6.24)}$$

where $\mathbb{E}$ takes expectation with respect to $h(d, \mathbf{x}, y; \theta_0, \eta_0, \zeta_0)$. Keep in mind that α is a function of ϑ and ζ.

(1) Proof of Equality (S6.22)

Since $\mathbb{E}\{D\nabla_\xi \log f(Y|\mathbf{X}, \xi_0)|\mathbf{X}\} = 0$, it follows that

$$\begin{aligned} C_1 &= \mathbb{E}\{\varphi_\mu(D, \mathbf{X}, Y) B_1(D, \mathbf{X}, Y)\} \\ &= \mathbb{E}[(1-D)\nabla_\theta\{\nabla_\vartheta \alpha(\vartheta_0) + \nabla_\vartheta r(\mathbf{X}, \theta_0)\}\{K(\mathbf{X}; \theta_0, \eta_0) - \mu_0\}] \\ &\quad +\mathbb{E}[(1-D)\{1 - D - \pi(\mathbf{X})\}\{\nabla_\vartheta \alpha(\vartheta_0) + \nabla_\vartheta r(\mathbf{X}, \theta_0)\}\nabla_{\theta^\top} t(\mathbf{X}, \theta_0) V^{-1} A] \end{aligned}$$

$$
\begin{aligned}
&+\mathbb{E}[D\nabla_\xi I_{e,-1}\{\log f(Y|\mathbf{X},\xi_0)\}^{\otimes 2} I_e^\top V^{-1}A] \\
= \;& \mathbb{E}[\pi(\mathbf{X})\{\nabla_\vartheta \alpha(\vartheta_0) + \nabla_\vartheta r(\mathbf{X},\theta_0)\}\{K(\mathbf{X};\theta_0,\eta_0) - \mu_0\}] \\
&+\mathbb{E}[\pi(\mathbf{X})\{1-\pi(\mathbf{X})\}\{\nabla_\vartheta t(\mathbf{X},\theta_0)\}^{\otimes 2} V^{-1}A] \\
&+\mathbb{E}[D\nabla_\xi I_{e,-1}\{\log f(Y|\mathbf{X},\xi_0)\}^{\otimes 2} I_e^\top V^{-1}A] \\
= \;& \mathbb{E}[\pi(\mathbf{X})\{\nabla_\vartheta \alpha(\vartheta_0) + \nabla_\vartheta r(\mathbf{X},\theta_0)\}\{K(\mathbf{X};\theta_0,\eta_0) - \mu_0\}] + A_{-1},
\end{aligned}
$$

where A_{-1} is A without its first component and we have used the definition of V.

Because

$$
\begin{aligned}
& \nabla_\vartheta K(\mathbf{x};\theta_0,\lambda_0) \\
= \;& \frac{\int y\{\nabla_\vartheta \alpha(\vartheta_0) + w(\mathbf{x},y)\}\lambda_0 \exp(\alpha_0 + \mathbf{x}^\top\beta_0 + \gamma_0 y) f(y|\mathbf{x},\xi_0)dy}{(1-\lambda_0) + \lambda_0\exp\{t(\mathbf{x},\theta_0)\}} \\
& -K(\mathbf{x};\theta_0,\lambda_0)\frac{\lambda_0\exp\{t(\mathbf{x},\theta_0)\}}{(1-\lambda_0)+\lambda_0\exp\{t(\mathbf{x},\theta_0)\}}\{\nabla_\vartheta \alpha(\vartheta_0) + \nabla_\vartheta r(\mathbf{x},\theta_0)\} \\
= \;& \frac{1-\eta_0}{\eta_0}\{1-\pi(\mathbf{x})\}\cdot\int y\{\nabla_\vartheta \alpha(\vartheta_0) + w(\mathbf{x},y)\}\exp(\alpha_0 + \mathbf{x}^\top\beta_0 + \gamma_0 y) f(y|\mathbf{x},\xi_0)dy \\
& -K(\mathbf{x};\theta_0,\lambda_0)\pi(\mathbf{x})\{\nabla_\vartheta \alpha(\vartheta_0) + \nabla_\vartheta r(\mathbf{x},\theta_0)\},
\end{aligned}
$$

we have

$$
\begin{aligned}
A_{-1} \;=\; & \mathbb{E}\{\nabla_\vartheta K(\mathbf{x};\theta_0,\lambda_0)\} \\
= \;& \mathbb{E}[\frac{1-\eta_0}{\eta_0}\{1-\pi(\mathbf{x})\}\int y\{\nabla_\vartheta \alpha(\vartheta_0) + w(\mathbf{x},y)\}\exp(\alpha_0 + \mathbf{x}^\top\beta_0 + \gamma_0 y) f(y|\mathbf{x},\xi_0)dy] \\
& -\mathbb{E}[K(\mathbf{x};\theta_0,\lambda_0)\pi(\mathbf{x})\{\nabla_\vartheta \alpha(\vartheta_0) + \nabla_\vartheta r(\mathbf{x},\theta_0)\}]
\end{aligned}
$$

$$
\begin{aligned}
= \; & \int (1-\eta_0) \int y\{\nabla_\vartheta \alpha(\vartheta_0) + w(\mathbf{x}, y)\} \exp(\alpha_0 + \mathbf{x}^\top \beta_0 + \gamma_0 y) f(y|\mathbf{x}, \xi_0) dy] dF(\mathbf{x}|D=1) \\
& - \int K(\mathbf{x}; \theta_0, \lambda_0)(1-\eta_0) \exp\{t(\mathbf{x}, \theta_0)\}\{\nabla_\vartheta \alpha(\vartheta_0) + \nabla_\vartheta r(\mathbf{x}, \theta_0)\} dF(\mathbf{x}|D=1),
\end{aligned}
$$

where we have used $t(\mathbf{x}, \theta) = \alpha + r(\mathbf{x}, \vartheta)$ and

$$
\mathrm{pr}(\mathbf{x}) = \frac{\eta_0}{1-\pi(\mathbf{x})} \mathrm{pr}(\mathbf{x}|D=1).
$$

It follows that

$$
\begin{aligned}
C_1 \;=\; & \mathbb{E}[\pi(\mathbf{X})\{\nabla_\vartheta \alpha(\vartheta_0) + \nabla_\vartheta r(\mathbf{X}, \vartheta_0)\}\{K(\mathbf{X}; \theta_0, \eta_0) - \mu_0\}] + A_{-1} \\
= \; & \int (1-\eta_0) \exp\{t(\mathbf{x}, \theta_0)\}\{\nabla_\beta \alpha(\vartheta_0) + \nabla_\vartheta r(\mathbf{x}, \vartheta_0\}\{K(\mathbf{x}; \theta_0, \lambda_0) - \mu_0\} dF(\mathbf{x}|D=1) + \\
& + \int (1-\eta_0) \int y\{\nabla_\vartheta \alpha(\vartheta_0) + w(\mathbf{x}, y)\} \exp(\alpha_0 + \mathbf{x}^\top \beta_0 + \gamma_0 y) f(y|\mathbf{x}, \xi_0) dy] dF(\mathbf{x}|D=1) \\
& - \int K(\mathbf{x}; \theta_0, \lambda_0)(1-\eta_0) \exp\{t(\mathbf{x}, \theta_0)\}\{\nabla_\vartheta \alpha(\vartheta_0) + \nabla_\vartheta r(\mathbf{x}, \theta_0)\} dF(\mathbf{x}|D=1) \\
= \; & -\mu_0 \int (1-\eta_0) \exp\{t(\mathbf{x}, \theta_0)\}\{\nabla_\beta \alpha(\vartheta_0) + \nabla_\vartheta r(\mathbf{x}, \vartheta_0\} dF(\mathbf{x}|D=1) + \\
& + \int (1-\eta_0) \int y\{\nabla_\vartheta \alpha(\vartheta_0) + w(\mathbf{x}, y)\} \exp(\alpha_0 + \mathbf{x}^\top \beta_0 + \gamma_0 y) f(y|\mathbf{x}, \xi_0) dy] dF(\mathbf{x}|D=1).
\end{aligned}
$$

Taking derivative with respect to ϑ on both sides of (S6.18) gives

$$
0 \;=\; \int \{\nabla_\vartheta \alpha(\vartheta_0) + \nabla_\vartheta r(\mathbf{x}, \vartheta_0)\} \exp\{\alpha(\vartheta_0) + r(\mathbf{x}, \vartheta_0)\} g(\mathbf{x}, \zeta_0) d\mathbf{x}.
$$

Since $g(\mathbf{x}, \zeta_0) d\mathbf{x} = \mathrm{pr}(\mathbf{x}|D=1) d\mathbf{x} = dF(\mathbf{x}|D=1)$, it follows that

$$
C_1 \;=\; \int (1-\eta_0) \int y\{\nabla_\vartheta \alpha(\vartheta_0) + w(\mathbf{x}, y)\} \exp(\alpha_0 + \mathbf{x}^\top \beta_0 + \gamma_0 y) f(y|\mathbf{x}, \xi_0) dy] dF(\mathbf{x}|D=1),
$$

which is exactly $\nabla_\vartheta \mu(\theta_0, \eta_0, \zeta_0)$.

(2) Proof of Equality (S6.23)

Since $\mathbb{E}\varphi_\mu(D,\mathbf{X},Y)=0$, we have

$$\begin{aligned}
C_2 &= \mathbb{E}\{B_2(D,\mathbf{X},Y)\varphi_\mu(D,\mathbf{X},Y)\}\\
&= \frac{1}{\eta_0(1-\eta_0)}\mathbb{E}\{D\varphi_\mu(D,\mathbf{X},Y)\}\\
&= \frac{1}{\eta_0(1-\eta_0)}\mathbb{E}\{DK(\mathbf{X};\theta_0,\eta_0)-\mu_0 D-\pi(\mathbf{X})DA^\top V^{-1}\nabla_\theta t(\mathbf{X},\theta_0)\\
&\quad +DA^\top V^{-1}I_e\nabla_\xi \log f(Y|\mathbf{X},\xi_0)\}\\
&= \frac{1}{\eta_0(1-\eta_0)}\mathbb{E}\{(1-\pi(\mathbf{X}))K(\mathbf{X};\theta_0,\eta_0)-\mu_0(1-\pi(\mathbf{X}))\\
&\quad -\pi(\mathbf{X})(1-\pi(\mathbf{X}))A^\top V^{-1}\nabla_\theta t(\mathbf{X},\theta_0)\}\\
&= \frac{1}{\eta_0(1-\eta_0)}[\mathbb{E}\{(1-\pi(\mathbf{X}))K(\mathbf{X};\theta_0,\eta_0)\}-\mu_0\eta_0-A^\top \mathbf{e}_1],
\end{aligned}$$

where the last equality holds because $V\mathbf{e}_1=\mathbb{E}\{\pi(\mathbf{X})(1-\pi(\mathbf{X}))\nabla_\theta t(\mathbf{X},\theta_0)\}$.

Meanwhile because

$$\begin{aligned}
A^\top \mathbf{e}_1 &= \mathbb{E}\{\nabla_\alpha K(\mathbf{x};\theta_0,\lambda_0)\}\\
&= \mathbb{E}[\frac{1-\eta_0}{\eta_0}\{1-\pi(\mathbf{X})\}\int y\exp(\alpha_0+\mathbf{X}^\top\beta_0+\gamma_0 y)f(y|\mathbf{X},\xi_0)dy]-\mathbb{E}\{K(\mathbf{x};\theta_0,\lambda_0)\pi(\mathbf{x})\},
\end{aligned}$$

we further have

$$\begin{aligned}
C_2\eta_0(1-\eta_0) &= \mathbb{E}\{(1-\pi(\mathbf{X}))K(\mathbf{X};\theta_0,\eta_0)\}-\mu_0\eta_0\\
&\quad -\mathbb{E}[\frac{1-\eta_0}{\eta_0}\{1-\pi(\mathbf{X})\}\int y\exp(\alpha_0+\mathbf{X}^\top\beta_0+\gamma_0 y)f(y|\mathbf{X},\xi_0)dy]\\
&\quad +\mathbb{E}[K(\mathbf{X};\theta_0,\eta_0)\pi(\mathbf{X})]
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}\{K(\mathbf{X};\theta_0,\eta_0)\} - \mu_0\eta_0 \\
&\quad -\mathbb{E}[\frac{1-\eta_0}{\eta_0}\{1-\pi(\mathbf{X})\}\int y\exp(\alpha_0+\mathbf{X}^\top\beta_0+\gamma_0 y)f(y|\mathbf{X},\xi_0)dy] \\
&= \mu_0(1-\eta_0) - (1-\eta_0)\int\int y\exp(\alpha_0+\mathbf{X}^\top\beta_0+\gamma_0 y)f(y|\mathbf{X},\xi_0)dydF(\mathbf{x}|D=1).
\end{aligned}
$$

Using the definition of μ_0, we have

$$
\begin{aligned}
&C_2(1-\eta_0)\eta_0 \\
&= (1-\eta_0)\int_{\mathbf{X}}\left[\int_y y\{\eta+(1-\eta)\exp(\alpha+\mathbf{x}^\top\beta+\gamma y)\}f(y|\mathbf{x},\xi)dy\right]dF(\mathbf{x}|D=1) \\
&\quad -(1-\eta_0)\int\int y\exp(\alpha_0+\mathbf{X}^\top\beta_0+\gamma_0 y)f(y|\mathbf{X},\xi_0)dy\mathrm{pr}(\mathbf{x}|D=1)d\mathbf{x} \\
&= (1-\eta_0)\int_{\mathbf{X}}\left[\int_y y\{\eta_0-\eta_0\exp(\alpha_0+\mathbf{x}^\top\beta_0+\gamma_0 y)\}f(y|\mathbf{x},\xi)dy\right]dF(\mathbf{x}|D=1) \\
&= (1-\eta_0)\eta_0\int_{\mathbf{X}}\left[\int_y y\{1-\exp(\alpha_0+\mathbf{x}^\top\beta_0+\gamma_0 y)\}f(y|\mathbf{x},\xi)dy\right]dF(\mathbf{x}|D=1).
\end{aligned}
$$

Since

$$
\nabla_\eta\mu = \int_{\mathbf{X}}\left[\int_y y\{1-\exp(\alpha+\mathbf{x}^\top\beta+\gamma y)\}f(y|\mathbf{x},\xi)dy\right]dF(\mathbf{x}|D=1),
$$

we arrive at

$$
C_2(1-\eta_0)\eta_0 = (1-\eta_0)\eta_0\nabla_\eta\mu(\theta_0,\eta_0,\zeta_0) \Longleftrightarrow C_2 = \nabla_\eta\mu(\theta_0,\eta_0,\zeta_0).
$$

This proves Equality (S6.23).

(3) Proof of Equality (S6.24)

Since $\mathbb{E}\{\varphi_\mu(D,\mathbf{X},Y)|D=1\}=K(\mathbf{X};\theta_0,\eta_0)-\mu_0$, we have

$$\begin{aligned} C_3 &= \mathbb{E}\{B_3(D,\mathbf{X},Y)\varphi_\mu(D,\mathbf{X},Y)\} \\ &= \mathbb{E}[\nabla_\zeta \log g(\mathbf{X},\zeta_0)\{K(\mathbf{X};\theta_0,\eta_0)-\mu_0\}] \\ &= \mathbb{E}[\nabla_\zeta \log g(\mathbf{X},\zeta_0)K(\mathbf{X};\theta_0,\eta_0)]. \end{aligned}$$

Note that

$$\mu=\mathbb{E}\{K(\mathbf{X};\theta_0,\eta_0)\}=\int_{\mathbf{X}} K(\mathbf{x},\theta_0,\eta_0)[\eta_0+(1-\eta_0)\exp\{t(\mathbf{x},\theta_0)\}]dF(\mathbf{x}|D=1),$$

which implies

$$\begin{aligned} \nabla_\zeta\mu &= \int_{\mathbf{X}} K(\mathbf{x},\theta_0,\lambda_0)[\eta_0+(1-\eta_0)\exp\{t(\mathbf{x},\theta_0)\}]\{\nabla_\zeta \log g(\mathbf{x},\zeta_0)\}dF(\mathbf{x}|D=1) \\ &= \int_{\mathbf{X}} K(\mathbf{x},\theta_0,\lambda_0)\{\nabla_\zeta \log g(\mathbf{x},\zeta_0)\}\mathrm{pr}(\mathbf{x})d\mathbf{x} \\ &= \mathbb{E}[\nabla_\zeta \log g(\mathbf{X},\zeta_0)K(\mathbf{X};\theta_0,\eta_0)] \\ &= C_3. \end{aligned}$$

This proves Equality (S6.24) and also completes the proof of (a1).

Proof of (b1)

Consider the following function

$$\begin{aligned} h^*_\psi(d,\mathbf{x},\tilde{y}) &= \{1+\psi\varphi_\mu(d,\mathbf{x},y)\}\times\{f(y|\mathbf{x},\xi_0)\eta_0\}^d\{\exp(\alpha_0+r(\mathbf{x},\vartheta_0))(1-\eta_0)\}^{1-d} \\ &\quad \times g(\mathbf{x},\zeta_0). \end{aligned}$$

Suppose the support of $(\mathbf{X},Y)$ is compact, then it can be verified that the function

$$\begin{aligned}\varphi_\mu(D,\mathbf{X},Y) &= K(\mathbf{X};\theta_0,\eta_0)-\mu_0+\{1-D-\pi(\mathbf{X})\}A^\top V^{-1}\nabla_\theta t(\mathbf{X},\theta_0)\\ &\quad +DA^\top V^{-1}I_e\nabla_\xi \log f(Y|\mathbf{X},\xi_0)\end{aligned}$$

is bounded. Because $\mathbb{E}\{\varphi_\mu(D,\mathbf{X},Y)\}=0$ where $\mathbb{E}$ takes expectation with respect to $h(d,\mathbf{x},y;\alpha_0,\vartheta_0,\eta_0,\zeta_0)$, the function $h^*_\psi(d,\mathbf{x},\tilde y)$ is a density function when ψ is small enough. When $\psi=0$, it reduces to the true joint density function $h(d,\mathbf{x},y;\alpha_0,\vartheta_0,\eta_0,\zeta_0)$. It is easy to check that $h^*_\psi(d,\mathbf{x},\tilde y)$ with small enough ψ is a parametric submodel and

$$\nabla_\psi h^*_\psi(d,\mathbf{x},\tilde y)\Big|_{\psi=0}=\varphi_\mu(d,\mathbf{x},y).$$

This proves (b1), and hence the semiparametric efficiency of $\hat\mu$.

S7 Additional simulation results

In all four examples, the missing probability model involves three parameters, the intercept (α^*), the coefficient for the covariate (β), and the tilting parameter (γ). In this section, we present the simulation results for estimating (α^*,β,γ). We compare the proposed estimator $(\hat\alpha^*,\hat\beta,\hat\gamma)$ with two others: (1) Morikawa and Kim (2016)'s adaptive estimator $(\tilde\alpha^*_t,\tilde\beta_t,\tilde\gamma_t)$ with correctly specified parametric form for $\mathrm{pr}(y|\mathbf{x},D=1)$; (2) Morikawa and Kim (2016)'s adaptive estimator $(\tilde\alpha^*_{np},\tilde\beta_{np},\tilde\gamma_{np})$ without specifying a parametric form for $\mathrm{pr}(y|\mathbf{x}, D = 1)$. We evaluate the performance of the three estimators in terms of relative bias (RB) and mean square error (MSE). The simulation results are summarized in Tables 1–4.

As we can see, the proposed estimator $(\hat{\alpha}^*, \hat{\beta}, \hat{\gamma})$ has small relative bias in all examples. It further has the smallest MSEs in almost all examples. The only exception is Example 1 with $\sigma^2 = 4$ and $n = 500$. In this situation, $(\tilde{\alpha}^*_{np}, \tilde{\gamma}_{np})$ has smaller MSE but much larger relative bias than $(\hat{\alpha}^*, \hat{\gamma})$.

Table 1: Relative bias (RB; ×100) and mean square error (MSE; ×100) of three estimates of $(\alpha^*, \beta, \gamma)$ in Example 1.

	n	σ^2	$\hat{\alpha}^*$	$\hat{\beta}$	$\hat{\gamma}$	$\tilde{\alpha}^*_t$	$\tilde{\beta}_t$	$\tilde{\gamma}_t$	$\tilde{\alpha}^*_{np}$	$\tilde{\beta}_{np}$	$\tilde{\gamma}_{np}$
RB	500	1	0.81	-0.79	0.97	1.12	-1.30	1.50	-0.63	2.71	0.06
MSE	500	1	7.84	6.59	0.71	8.15	6.73	0.75	7.94	6.68	0.76
RB	2000	1	0.47	-0.51	0.47	0.53	-0.60	0.58	-0.27	1.18	-0.14
MSE	2000	1	1.88	1.56	0.17	1.92	1.59	0.17	1.91	1.59	0.17
RB	500	4	2.92	-1.64	1.86	10.52	-9.75	7.12	-15.45	24.09	-10.27
MSE	500	4	20.46	6.62	1.08	101.49	18.28	3.46	18.68	7.79	1.01
RB	2000	4	0.92	-0.89	0.46	2.47	-1.95	1.82	-8.26	11.77	-5.70
MSE	2000	4	4.61	1.59	0.24	9.97	2.23	0.44	5.84	2.00	0.31

Table 2: Relative bias (RB; ×100) and mean square error (MSE; ×100) of three estimates of $(\alpha^*, \beta, \gamma)$ in Example 2.

	n	σ^2	$\hat{\alpha}^*$	$\hat{\beta}$	$\hat{\gamma}$	$\tilde{\alpha}^*_t$	$\tilde{\beta}_t$	$\tilde{\gamma}_t$	$\tilde{\alpha}^*_{np}$	$\tilde{\beta}_{np}$	$\tilde{\gamma}_{np}$
RB	500	1	0.60	2.46	0.71	0.85	2.56	1.10	-12.31	19.49	-10.06
MSE	500	1	9.07	1.94	0.72	9.48	2.03	0.77	11.85	2.54	0.89
RB	2000	1	0.46	-0.04	0.37	0.48	0.00	0.41	-8.45	11.68	-6.99
MSE	2000	1	2.37	0.49	0.19	2.43	0.50	0.20	4.13	0.71	0.30
RB	500	4	2.64	0.36	1.90	5.90	-0.73	4.35	-46.54	44.33	-33.08
MSE	500	4	18.58	2.20	0.78	45.84	4.11	1.57	69.82	5.07	3.13
RB	2000	4	0.89	-0.34	0.50	1.54	-0.55	0.98	-38.76	35.54	-26.94
MSE	2000	4	4.76	0.56	0.20	7.00	0.76	0.28	45.63	2.54	1.93

Table 3: Relative bias (RB; ×100) and mean square error (MSE; ×100) of three estimates of $(\alpha^*, \beta, \gamma)$ in Example 3.

	n	σ^2	$\hat{\alpha}^*$	$\hat{\beta}$	$\hat{\gamma}$	$\tilde{\alpha}^*_t$	$\tilde{\beta}_t$	$\tilde{\gamma}_t$	$\tilde{\alpha}^*_{np}$	$\tilde{\beta}_{np}$	$\tilde{\gamma}_{np}$
RB	500	1	0.62	2.05	0.85	1.53	0.21	2.34	1.85	0.24	3.84
MSE	500	1	11.98	1.65	0.85	13.97	1.82	1.02	15.48	2.01	1.20
RB	2000	1	0.25	0.93	0.22	0.43	0.54	0.51	1.27	-2.14	2.32
MSE	2000	1	2.78	0.41	0.20	3.02	0.45	0.22	3.42	0.49	0.26
RB	500	$e^{0.7}$	0.79	2.26	0.84	3.47	-1.34	4.52	1.40	-3.71	3.43
MSE	500	$e^{0.7}$	15.61	1.75	0.86	25.30	2.12	1.44	20.84	2.19	1.27
RB	2000	$e^{0.7}$	0.30	0.93	0.24	0.89	0.08	1.06	1.50	-5.56	2.83
MSE	2000	$e^{0.7}$	3.47	0.41	0.19	4.50	0.46	0.26	4.69	0.57	0.29

Table 4: Relative bias (RB; ×100) and mean square error (MSE; ×100) of three estimates of $(\alpha^*, \beta, \gamma)$ in Example 4.

	n	σ^2	$\hat{\alpha}^*$	$\hat{\beta}$	$\hat{\gamma}$	$\tilde{\alpha}_t^*$	$\tilde{\beta}_t$	$\tilde{\gamma}_t$	$\tilde{\alpha}_{np}^*$	$\tilde{\beta}_{np}$	$\tilde{\gamma}_{np}$
RB	500	3	0.23	2.96	0.43	0.52	3.12	0.75	-13.91	21.20	-11.36
MSE	500	3	9.32	1.98	0.73	9.82	2.08	0.79	13.04	2.67	0.96
RB	2000	3	0.45	-0.06	0.54	0.74	0.02	0.83	-9.59	12.95	-7.72
MSE	2000	3	2.40	0.49	0.18	2.64	0.53	0.20	4.72	0.77	0.32
RB	500	6	-0.05	2.81	0.65	1.19	2.81	1.34	-17.11	23.58	-13.65
MSE	500	6	9.34	1.94	0.71	13.71	2.43	0.90	15.41	2.81	1.04
RB	2000	6	-0.18	-0.11	0.44	0.70	-0.05	0.70	-12.98	15.58	-10.22
MSE	2000	6	2.52	0.49	0.18	3.17	0.59	0.23	6.86	0.88	0.42

Bibliography

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

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

Newey, W. K. (1990). Semiparametric efficiency bounds. *Journal of Applied Econometrics*, **5**, pp. 99–135.

Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *Annals of Statistics 22*, pp. 300–325.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics.* New York: Wiley.