Statistica Sinica

Model-Assisted Inference for Covariate-Specific Treatment Effects with High-dimensional Data

Peng Wu[a], Zhiqiang Tan[b], Wenjie Hu[c] and Xiao-Hua Zhou[c,d]*

[a]*Beijing Technology and Business University*

[b]Rutgers University, [c]Peking University, [d]Pazhou Lab,

This Supplementary Material consists of Sections 1–7, where Section 1 presents several properties algebraically associated with $\hat{\pi}_{RCAL}(X)$ and $\hat{m}_{1,RWL}(X)$, Section 2 extends the proposed method to estimate $\mu^0(z)$ and $\tau(z)$, Section 3 presents the regularity assumptions and probability lemmas, Section 4 and Section 5 give the technical proofs of Theorem 1 and Theorem 2, Sections 6 and 7 contain additional numerical results from the simulation study and empirical application.

1. Properties algebraically associated with $\hat{\pi}_{RCAL}(X)$ and $\hat{m}_{1,RWL}(X)$

We present several interesting properties algebraically associated with $\hat{\pi}_{RCAL}(X)$ and $\hat{m}_{1,RWL}(X)$, part of which are also used in proving our results later.

*correspond to: azhou@math.pku.edu.cn

First, by the Karush-Kuhn-Tucker (KKT) condition for minimizing (3.7), the fitted propensity score $\hat{\pi}_{RCAL}(X)$ satisfies

$$\frac{1}{n}\sum_{i=1}^{n}\frac{T_i}{\hat{\pi}_{RCAL}(X_i)} = 1, \tag{S.1}$$

$$\frac{1}{n}\left|\sum_{i=1}^{n}\frac{T_i f_j(X_i)}{\hat{\pi}_{RCAL}(X_i)} - \sum_{i=1}^{n} f_j(X_i)\right| \leq \lambda, \quad j = 1, ..., p. \tag{S.2}$$

where equality holds in (S.2) for any j such that the j-th element of $\hat{\gamma}_{RCAL}$ is nonzero. Equation (S.1) shows that the sum of inverse probability weights $T/\hat{\pi}_{RCAL}(X)$ equals to sample size n, whereas equation (S.2) implies that the weighted average of each covariate $f_j(X_i)$ over the treated group may differ from the overall average of $f_j(X_i)$ by no more than λ. In addition, Tan (2020b) showed that, with possible model misspecification, calibrated estimation is better than maximum likelihood estimation in terms of controlling mean squared errors of inverse probability weighted estimators.

Second, by the KKT condition for minimizing (3.8), the fitted treatment regression function $\hat{m}_{1,RWL}(X)$ satisfies

$$n^{-1}\sum_{i=1}^{n} T_i w(X; \hat{\gamma}_{RCAL})\{Y_i - \hat{m}_{1,RWL}(X_i)\} = 0. \tag{S.3}$$

As a consequence of equation (S.3), the augmented IPW estimator for $E(Y^1)$, defined as $\hat{E}_{RCAL}(Y^1) = \tilde{E}\{\varphi(Y, T, X; \hat{m}_{1,RWL}, \hat{\pi}_{RCAL})\}$, can be reformulated as

$$\tilde{E}\left[\hat{m}_{1,RWL}(X) + \frac{T}{\hat{\pi}_{RCAL}(X)}\{Y - \hat{m}_{1,RWL}(X)\}\right] = \tilde{E}\{TY + (1-T)\hat{m}_{1,RWL}(X)\},$$

which implies that $\hat{E}_{RCAL}(Y^1)$ always fall within the range of the observed outcomes $\{Y_i : T_i = 1, i = 1, ..., n\}$ and the predicted values $\{\hat{m}_{1,RWL}(X_i) : T_i = 0, i = 1, ..., n\}$.

2. Estimations of $\mu^0(z)$ and $\tau(z)$

The results presented in Propositions 1 and 2 of the manuscript mainly focus on estimation of $\mu^1(z)$, but they can be directly extended for estimating $\mu^0(z)$ and $\tau(z)$. Similarly, we posit a marginal structural model for $\mu^0(z)$ based on basis functions $(1, \Phi(z))$. In addition to the propensity score model (2.3) and generalized linear outcome model (2.2) in the manuscript, consider the following outcome regression model in the untreated population,

$$E(Y \mid T = 0, X) = m_0(X; \alpha_0) = \psi\{\alpha_0^T g(X)\}, \tag{S.4}$$

where $g(X)$ is the same as in model (2.2) and α_0 is a vector of unknown parameters. Then for a given z_0, our point estimator of $\tau(z_0)$ is $\hat{\mu}^1(z_0; \hat{m}_1, \hat{\pi}) - \hat{\mu}^0(z_0; \hat{m}_0, \hat{\pi}_0)$, with

$$\hat{\mu}^0(z_0; \hat{m}_0, \hat{\pi}_0) = \Phi^\dagger(z_0)^T \tilde{E}^{-1}\left[\Phi^\dagger(Z)\Phi^\dagger(Z)^T\right] \tilde{E}\left[\Phi^\dagger(Z)\varphi(Y, 1-T, X; \hat{m}_0, 1-\hat{\pi}_0)\right],$$

where $\hat{\pi}_0 = \pi(X; \hat{\gamma}_{0,RCAL})$, $\hat{m}_0 = m_0(X; \hat{\alpha}_{0,RWL})$, and $\hat{\gamma}_{0,RCAL}$ is defined as a minimizer of (3.7) in the manuscript, but with the calibration loss function $L_{CAL}(\gamma)$ replaced by $L_{0,CAL}(\gamma) = \tilde{E}[(1 - T)\exp\{\gamma^T f(X)\} - T\gamma^T f(X)]$.

The estimator $\hat{\alpha}_{0,RWL}$ is defined as a minimizer of $L_{0,RWL}(\alpha_0;\hat{\gamma}_{0,RCAL}) = L_{0,WL}(\alpha_0;\hat{\gamma}_{0,RCAL}) + \lambda||(\alpha_0)_{1:q}||_1$, with

$$L_{0,WL}(\alpha_0;\hat{\gamma}_{0,RCAL}) = \tilde{E}[(1-T)w^{-1}(X;\hat{\gamma}_{0,RCAL})\{-Y\alpha_0^T g(X)+\Psi(\alpha_0^T g(X))\}].$$

Under similar conditions in Propositions 1 or 2, the estimator $\hat{\mu}^0(z_0;\hat{m}_0,\hat{\pi}_0)$ admits an asymptotic expansion

$$\hat{\mu}^0(z_0;\hat{m}_0,\hat{\pi}_0) = \hat{\mu}^0(z_0;\bar{m}_0,\bar{\pi}_0) + o_p(n^{-1/2}),$$

where $\bar{\pi}_0 = \pi(X;\bar{\gamma}_0)$, $\bar{m}_0 = m_0(X;\bar{\alpha}_0)$, $\bar{\gamma}_0$ and $\bar{\alpha}_0$ are defined as the minimizers of $E\{L_{0,CAL}(\gamma)\}$ and $E\{L_{0,WL}(\alpha_0;\hat{\gamma}_{0,RCAL})\}$, respectively. In particular, an asymptotic $(1-c)$ confidence interval for $\tau(z_0)$ can be given as

$$\hat{\mu}^1(z_0;\hat{m}_1,\hat{\pi}) - \hat{\mu}^0(z_0;\hat{m}_0,\hat{\pi}_0) \pm z_{c/2}(\hat{\mathbb{V}}(z_0)/n)^{1/2},$$

where $\hat{\mathbb{V}}(z_0) = \Phi^\dagger(z_0)^T M^{-1}\hat{\mathbb{G}}M^{-1}\Phi^\dagger(z_0)/n$, $\hat{\mathbb{G}} = \tilde{E}[\Phi^\dagger(Z)\Phi^\dagger(Z)^T\{\hat{\varphi}_\tau - \check{\beta}^T\Phi^\dagger(Z)\}^2]$, $\check{\beta} = \tilde{E}^{-1}\{\Phi^\dagger(Z)\Phi^\dagger(Z)^T\}\tilde{E}\{\Phi^\dagger(Z)\hat{\varphi}_\tau\}$, $\hat{\varphi}_\tau = \varphi(Y,T,X;\hat{m}_1,\hat{\pi}) - \varphi(Y,1-T,X;\hat{m}_0,1-\hat{\pi}_0)$, and $M = \tilde{E}\{\Phi^\dagger(Z)\Phi^\dagger(Z)^T\}$.

3. Regularity Assumptions and Probability Lemmas

For a matrix Σ with row indices $\{0,1,...,k\}$, a compatibility condition (Buhlmann and van de Geer, 2011) is said to hold with a subset $S \in$

$\{0, 1, ..., k\}$ and constants $\nu > 0$ and $\xi > 1$ if $\nu^2(\sum_{j \in S} |b_j|^2 \leq b^T \Sigma b)$ for any vector $b = (b_0, b_1, ..., b_k) \in \mathbb{R}^{k+1}$ satisfying $\sum_{j \notin S} |b_j| \leq \xi \sum_{j \in S} |b_j|$.

Assumption 1: Suppose that the following regularity conditions are satisfied:

(i) $max_{j=0,1,...,p} |f_j(X)| \leq C_0$ almost surely for a constant $C_0 \geq 1$;

(ii) $\bar{\gamma}^T f(X) \geq B_0$ almost surely for a constant B_0, that is, $\pi(X; \bar{\gamma}) \geq (1 + e^{-B_0})^{-1}$.

(iii) a compatibility condition holds for Σ_f with the subset $S_{\bar{\gamma}} = \{0\} \cup \{j : \bar{\gamma}_j \neq 0, j = 1, ..., p\}$ and some constants $\nu_0 > 0$ and $\xi_0 > 1$, where $\Sigma_f = E[Tw(X; \bar{\gamma}) f(X) f(X)^T]$ is the Hessian of $E\{L_{CAL}(\gamma)\}$ at $\gamma = \bar{\gamma}$ and $w(X; \bar{\gamma}) = e^{-\gamma^T f(X)}$.

(iv) $|S_{\bar{\gamma}}| \lambda_0 < \eta_0$ for a sufficiently small constant η_0, depending only on (A_0, C_0, ν_0, ξ_0).

Assumption 2: Suppose that the following regularity conditions are satisfied:

(i) $max_{j=0,1,...,q} |g_j(X)| \leq C_1$ almost surely for a constant $C_1 \geq 1$;

(ii) $\bar{\alpha}_1^T g(X)$ is bounded in absolute values by $B_1 > 0$ almost surely;

(iii) $\psi'(u) \leq \psi'(\tilde{u}) e^{C_2 |u - \tilde{u}|}$ for any $(u, \tilde{u})$, where C_2 is a constant.

(iv) $Y^1 - m_1(X; \bar{\alpha}_1)$ is uniformly sub-Gaussian given X:

$$D_0^2 E\big[\exp\{(Y^1 - m_1(X; \bar{\alpha}_1))^2 / D_0^2\} - 1 \big| X\big] \leq D_1^2$$

for some positive constants (D_0, D_1).

(v) a compatibility condition holds for Σ_g with the subset $S_{\bar\alpha_1} = \{0\} \cup \{j : \bar\alpha_{1,j} \neq 0, j = 1, ..., p\}$ and some constants $\nu_1 > 0$ and $\xi_1 > 0$, where $\Sigma_g = E[Tw(X; \bar\gamma) g(X) g(X)^T]$.

(vi) $|S_{\bar\gamma}|\lambda_0 + |S_{\bar\alpha_1}|\lambda_1 < \eta_1$ for a sufficiently small constant η_1, depending only on (A_1, C_1, ν_1, ξ_1).

The following Lemma 1 summarizes the results of Tan (2020a) related to $(\hat\gamma, \hat\alpha_1)$.

Lemma 1. *Suppose Assumptions 1 and 2 hold and $\lambda_0 \leq 1$. Then*

(a) If OR model (2.2) is used, $g(X)$ is specified as in (3.10), then we have with probability at least $1 - c_0\epsilon$,

$$||\hat\gamma - \bar\gamma||_1 \leq M_2 |S_{\bar\gamma}|\lambda_0, \quad (\hat\gamma - \bar\gamma)^T \tilde\Sigma_f (\hat\gamma - \bar\gamma) \leq M_2 |S_{\bar\gamma}|\lambda_0^2 \tag{S.5}$$

$$||\hat\alpha_1 - \bar\alpha_1||_1 \leq M_3(|S_{\bar\gamma}|\lambda_0 + |S_{\bar\alpha_1}|\lambda_1), \quad (\hat\alpha_1 - \bar\alpha_1)^T \tilde\Sigma_g (\hat\alpha_1 - \bar\alpha_1) \leq M_3(|S_{\bar\gamma}|\lambda_0^2 + |S_{\bar\alpha_1}|\lambda_1^2), \tag{S.6}$$

where c_0, M_2 and M_3 are positive constants, $\tilde\Sigma_f$ and $\tilde\Sigma_g$ are the sample versions of Σ_f and Σ_g, i.e., $\tilde\Sigma_f = \tilde E[Tw(X; \bar\gamma) f(X) f(X)^T]$ and $\tilde\Sigma_g = \tilde E[Tw(X; \bar\gamma) g(X) g(X)^T]$. Furthermore, if PS model (2.3) is correctly specified, we also have with probability at least $1 - c_0\epsilon$,

$$\left|\tilde E(\hat\varphi - \bar\varphi)\right| \leq M_4(|S_{\bar\gamma}|\lambda_0 + |S_{\bar\alpha_1}|\lambda_1)\lambda_1, \tag{S.7}$$

where M_4 is a positive constant.

(b) If linear OR model (3.12) is used, $f(X)$ and $g(X)$ are specified as in (3.13), then the results (S.5), (S.6) and (S.7) also hold.

□

Inequalities (S.5) and (S.6) lead directly to the convergence rates for $(\hat\gamma, \hat\alpha_1)$,

$$||\hat\gamma - \bar\gamma||_1 = O_p(1) \cdot |S_{\bar\gamma}|\{\log(p)/n\}^{1/2}, \ ||\hat\alpha_1 - \bar\alpha_1||_1 = O_p(1) \cdot (|S_{\bar\gamma}| + |S_{\bar\alpha_1}|)\{\log(q)/n\}^{1/2}.$$

Inequality (S.7) will be used to establish the inequalities (4.21) in Theorem 1(a) and (4.23) in Theorem 2(a),

Denote by Ω_0 the event that (S.5)–(S.7) hold. Then $P(\Omega_0) \geq 1 - c_0\epsilon$ under Lemma 1. The following Lemmas 2–3 will be used in the proof of Theorem 1, Lemmas 4–7 will be used in the proof of Theorem 2.

Recall that

$$\Sigma_f = E[Tw(X;\bar\gamma)f(X)f(X)^T]$$
$$\tilde\Sigma_f = \tilde E[Tw(X;\bar\gamma)f(X)f(X)^T]$$

and denote

$$\Sigma_{\alpha 2} = E[Tw(X;\bar\gamma)\{Y - m_1(X;\bar\alpha_1)\}^2 f(X)f(X)^T]$$
$$\tilde\Sigma_{\alpha 2} = \tilde E[Tw(X;\bar\gamma)\{Y - m_1(X;\bar\alpha_1)\}^2 f(X)f(X)^T]$$

Lemma 2. *Under Assumptions 1(i)–(ii), there exists a positive constant B_2, depending on (B_0, C_0), such that $P(\Omega_1) \geq 1 - 2\epsilon$, where Ω_1 denotes the event*

$$\sup_{j,k=0,1,\ldots,p} \left| (\tilde{\Sigma}_f)_{jk} - (\Sigma_f)_{jk} \right| \leq B_2 \lambda_0.$$

Proof. This can be shown similarly as Lemma 1(ii) in the Supplement of Tan (2020a). □

Lemma 3. *Under Assumptions 1(i)–(ii) and 2(iv), there exists a positive constant B_3, depending on (B_0, C_0, D_0, D_1), such that if $\lambda_0 \leq 1$, then $P(\Omega_2) \geq 1 - 2\epsilon$, where Ω_2 denotes the event*

$$\sup_{j,k=0,1,\ldots,p} \left| (\tilde{\Sigma}_{\alpha 2})_{jk} - (\Sigma_{\alpha 2})_{jk} \right| \leq B_3 \lambda_0.$$

Proof. This can be shown similarly as Lemma 3 in the Supplement of Tan (2020a). □

In the event $\Omega_1 \cap \Omega_2$, we have for any vector $b \in \mathbb{R}^{p+1}$,

$$\left| (\tilde{E} - E) \Big\{ Tw(X; \bar{\gamma}) \{ b^T f(X) \}^2 \Big\} \right| \leq B_2 \lambda_0 ||b||_1^2,$$
$$\left| (\tilde{E} - E) \Big\{ Tw(X; \bar{\gamma}) \{ Y - m_1(X; \bar{\alpha}_1) \}^2 \{ b^T f(X) \}^2 \Big\} \right| \leq B_3 \lambda_0 ||b||_1^2.$$

By Assumption 2(iv), $E[\{Y - m(X; \bar{\alpha}_1)\}^2 | X] \leq D_0^2 + D_1^2$ and it implies

$$E\big[Tw(X; \bar{\gamma}) \{ Y - m(X; \bar{\alpha}_1) \}^2 \{ b^T f(X) \}^2 \big] \leq (D_0^2 + D_1^2) E\big\{ Tw(X; \bar{\gamma}) \{ b^T f(X) \}^2 \big\}.$$

Take $b = (\hat{\gamma} - \bar{\gamma})$, then in the event $\Omega_0 \cap \Omega_1 \cap \Omega_2$,

$$
\begin{aligned}
&\tilde{E}\big[Tw(X;\bar{\gamma})\{Y - m_1(X;\bar{\alpha}_1)\}^2\{\hat{\gamma}^T f(X) - \bar{\gamma}^T f(X)\}^2\big] \\
&\le (D_0^2 + D_1^2)(\hat{\gamma} - \bar{\gamma})^T E\big\{Tw(X;\bar{\gamma}) f(X) f(X)^T\big\}(\hat{\gamma} - \bar{\gamma}) + B_3\lambda_0 ||\hat{\gamma} - \bar{\gamma}||_1^2 \\
&\le (D_0^2 + D_1^2)(\hat{\gamma} - \bar{\gamma})^T \tilde{E}\big\{Tw(X;\bar{\gamma}) f(X) f(X)^T\big\}(\hat{\gamma} - \bar{\gamma}) \\
&\quad + \{(D_0^2 + D_1^2)B_2 + B_3\}\lambda_0 ||\hat{\gamma} - \bar{\gamma}||_1^2 \\
&\le (D_0^2 + D_1^2)M_0|S_{\bar{\gamma}}|\lambda_0^2 + \{(D_0^2 + D_1^2)B_2 + B_3\}M_2^2|S_{\bar{\gamma}}|^2\lambda_0^3. \qquad \text{(S.8)}
\end{aligned}
$$

Lemma 4. *Suppose that Assumptions 1(ii) and 2(i)–(iii) hold, then* $P(\Omega_3) \ge 1 - 2\epsilon$ *for any* $r \ge 0$*, where* Ω_3 *denotes the event*

$$
\sup_{||\alpha_1 - \bar{\alpha}_1||_1 \le r;\, k=1,\ldots,K} \left| \tilde{E}\left(\big[\psi\{\alpha_1^T g(X)\} - \psi\{\bar{\alpha}_1^T g(X)\}\big] \left\{1 - \frac{T}{\pi^*(X)}\right\} \phi_k(Z) \right) \right| \le B_4 \lambda_0 r,
$$

where B_4 *is a positive constant, depending on* (B_0, B_1, C_1, C_2)*.*

Proof. This can be shown similarly as Lemma 13 in the Supplement of Tan (2020a). □

Lemma 5. *Under Assumptions 1(ii), 2(i) and 2(iv), there exists a positive constant* B_5*, depending on* (B_0, C_1, D_0, D_1)*, such that* $P(\Omega_4) \ge 1 - 2\epsilon$*, where* Ω_4 *denotes the event*

$$
\sup_{j=0,1,\ldots,p;\, k=1,\ldots,K} \big|\tilde{E}\big(Tw(X;\bar{\gamma})\{Y - \psi(\bar{\alpha}_1^T g(X))\} f_j(X)\phi_k(Z)\big)\big| \le B_5 \lambda_1.
$$

Proof. This can be shown similarly as Lemma 2 in the Supplement of Tan (2020a). □

Recall

$$\Sigma_g = E[Tw(X;\bar{\gamma})g(X)g(X)^T]$$

$$\tilde{\Sigma}_g = \tilde{E}[Tw(X;\bar{\gamma})g(X)g(X)^T]$$

and denote

$$\Sigma_{|\alpha|} = E[Tw(X;\bar{\gamma})|Y - \psi(\alpha_1^T g(X))|f(X)f(X)^T]$$

$$\tilde{\Sigma}_{|\alpha|} = \tilde{E}[Tw(X;\bar{\gamma})|Y - \psi(\alpha_1^T g(X))|f(X)f(X)^T]$$

Lemma 6. *Under Assumptions 1(i)–(ii) and 2(iv), there exists a positive constant B_6, depending on (B_0, C_0, D_0, D_1), such that $P(\Omega_5) \geq 1-2\epsilon$, where Ω_5 denotes the event*

$$\sup_{j,k=0,1,\ldots,p} \left| (\tilde{\Sigma}_{|\alpha|})_{jk} - (\Sigma_{|\alpha|})_{jk} \right| \leq B_6 \lambda_0$$

Proof. This can be shown similarly as Lemma 4 in the Supplement of Tan (2020a). □

Lemma 7. *Under Assumptions 1(ii) and 2(i), there exists a positive constant B_7, depending on (B_0, C_3), such that $P(\Omega_6) \geq 1-2\epsilon$, where Ω_6 denotes the event*

$$\sup_{j,k=0,1,\ldots,q} \left| (\tilde{\Sigma}_g)_{jk} - (\Sigma_g)_{jk} \right| \leq B_7 \lambda_1$$

Proof. This can be shown similarly as Lemma 1(ii) in the Supplement of Tan (2020a). □

4. Proof of Theorem 1

We focus on analyzing the case of binary Z, the other cases of discrete Z can be derived by a exact similar argument.

4.1 Proof of Theorem 1(a)

For a given z_0 of binary Z,

$$\hat{\mu}^1(z_0;\hat{m}_1,\hat{\pi}) = \bar{\mu}^1(z_0;\bar{m}_1,\bar{\pi}) + (1,z_0)\tilde{E}^{-1}\{\begin{pmatrix}1\\ Z\end{pmatrix}(1,Z)\}\cdot\tilde{E}\{(\hat{\varphi}-\bar{\varphi})\begin{pmatrix}1\\ Z\end{pmatrix}\}.$$

By the inequality (S.7) in Lemma 1, it suffices to show that

$$\left|\tilde{E}\{(\hat{\varphi}-\bar{\varphi})Z\}\right| \le M_0(|S_{\bar{\gamma}}|\lambda_0 + |S_{\bar{\alpha}_1}|\lambda_1)\lambda_1 \tag{S.9}$$

for a positive constant M_0. Consider the following decomposition,

$$\begin{aligned}\hat{\varphi}-\bar{\varphi} = \{\hat{m}_1(X)-\bar{m}_1(X)\}\Big\{1-\frac{T}{\bar{\pi}(X)}\Big\} + T\{Y-\bar{m}_1(X)\}\Big\{\frac{1}{\hat{\pi}(X)}-\frac{1}{\bar{\pi}(X)}\Big\}\\ + \{\hat{m}_1(X)-\bar{m}_1(X)\}\Big\{\frac{T}{\bar{\pi}(X)}-\frac{T}{\hat{\pi}(X)}\Big\},\end{aligned} \tag{S.10}$$

denoted as $\delta_1+\delta_2+\delta_3$, then $\tilde{E}\{(\hat{\varphi}-\bar{\varphi})Z\} = \Delta_1+\Delta_2+\Delta_3$ with $\Delta_1 = \tilde{E}(\delta_1 Z)$, $\Delta_2 = \tilde{E}(\delta_2 Z)$ and $\Delta_3 = \tilde{E}(\delta_3 Z)$. When $f(X)$ and $g(X)$ are specified as in

(3.13) and linear outcome model (3.12) is used,

$$\begin{aligned}
\Delta_1 &= (\hat\alpha_1 - \bar\alpha_1)^T \tilde E\Big[\Big\{1 - \frac{T}{\bar\pi(X)}\Big\} f(X) Z\Big], \\
\Delta_2 &= \tilde E\Big[T\{Y - \bar m_1(X)\}\Big\{\frac{1}{\hat\pi(X)} - \frac{1}{\bar\pi(X)}\Big\} Z\Big], \\
\Delta_3 &= (\hat\alpha_1 - \bar\alpha_1)^T \tilde E\Big[\Big\{\frac{T}{\bar\pi(X)} - \frac{T}{\hat\pi(X)}\Big\} f(X) Z\Big].
\end{aligned}$$

Since each element of $f(X)Z$ is included in $f(X)$ by our specification of $f(X)$, then according to the K.K.T. condition (S.2) and inequality (S.6) in Lemma 1, we have

$$\begin{aligned}
|\Delta_1 + \Delta_3| &= (\hat\alpha_1 - \bar\alpha_1)^T \tilde E\Big[\Big\{1 - \frac{T}{\hat\pi(X)}\Big\} f(X) Z\Big] \\
&\le \max_{j=1,\ldots,p} \Big|\tilde E\big[\big\{1 - \frac{T}{\hat\pi(X)}\big\} f_j(X) Z\big]\Big| \cdot ||\hat\alpha_1 - \bar\alpha_1||_1 \\
&\le A_0 \lambda_0 \cdot M_3(|S_{\bar\gamma}|\lambda_0 + |S_{\bar\alpha_1}|\lambda_1). \qquad \text{(S.11)}
\end{aligned}$$

By a Taylor expansion for Δ_2 yields for some $u \in (0,1)$ that

$$\begin{aligned}
\Delta_2 = & -(\hat\gamma - \bar\gamma)^T \tilde E\big[T\{Y - m_1(X;\bar\alpha_1)\} e^{-\bar\gamma^T f(X)} f(X) Z\big] \\
& + (\hat\gamma - \bar\gamma)^T \tilde E\big[T\{Y - m_1(X;\bar\alpha_1)\} e^{-u\hat\gamma^T f(X) - (1-u)\bar\gamma^T f(X)} f(X) f(X)^T Z\big](\hat\gamma - \bar\gamma)/2,
\end{aligned}$$

denoted as $\Delta_{21} + \Delta_{22}$. Note that $Z^2 = Z$ and the elements of $f(X)Z$ are included in $f(X)$ again, by a exact similar argument of Theorem 3 in the Supplement of Tan (2020a), there exists a positive constant C_3 such that

$$|\Delta_{21}| + |\Delta_{22}| \le C_3(|S_{\bar\gamma}|\lambda_0 + |S_{\bar\alpha_1}|\lambda_1)\lambda_1. \qquad \text{(S.12)}$$

Inequality (S.9) follows immediately from (S.11) with (S.12).

□

4.2 Proof of Theorem 1(b)

When Z is binary, recall that

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \tilde{E}^{-1}\{\begin{pmatrix} 1 \\ Z \end{pmatrix}(1, Z)\}\tilde{E}\{\hat{\varphi}\begin{pmatrix} 1 \\ Z \end{pmatrix}\}, \quad \begin{pmatrix} \bar{\beta}_0 \\ \bar{\beta}_1 \end{pmatrix} = \tilde{E}^{-1}\{\begin{pmatrix} 1 \\ Z \end{pmatrix}(1, Z)\}\tilde{E}\{\bar{\varphi}\begin{pmatrix} 1 \\ Z \end{pmatrix}\}$$

Let $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)^T$, $\bar{\beta} = (\bar{\beta}_0, \bar{\beta}_1)^T$, then for a given z_0, $\hat{\mu}^1(z_0; \hat{\pi}, \hat{m}_1) = (1, z_0)\hat{\beta}$, $\hat{\mu}^1(z_0; \bar{\pi}, \bar{m}_1) = (1, z_0)\bar{\beta}$. Define

$$\bar{V}(z_0) \equiv \text{var}\{n^{1/2}\hat{\mu}^1(z_0; \bar{\pi}, \bar{m}_1)\} = (1, z_0)\text{var}\{n^{1/2}\bar{\beta}\}(1, z_0)^T,$$

it is easy to see that $\bar{V}(z_0)$ is a consistent estimator of $V(z_0)$ defined in Proposition 2. Thus, it suffices to show that

$$\hat{V}(z_0) = \bar{V}(z_0) + o_p(1). \tag{S.13}$$

First, we claim that

$$\{\tilde{E}(1, Z)^T(1, Z)\}^{-1} \cdot \tilde{E}\{(1, Z)^T(1, Z)(\bar{\varphi} - (1, Z)\bar{\beta})^2\} \cdot \{\tilde{E}(1, Z)^T(1, Z)\}^{-1} \tag{S.14}$$

is a consistent estimator of $\text{var}\{n^{1/2}\bar{\beta}\}$. To show (S.14), note that $\beta^* = (\beta_0^*, \beta_1^*)^T$ (defined in (3.6)) satisfies the "population" version of estimating equation:

$$E[(1,Z)^T\{\bar{\varphi}-(1,Z)\beta^*\}]=0,$$

and least square estimator $\bar{\beta}$ satisfies the following sample estimating equation:

$$n^{-1}\sum_{i=1}^{n}(1,Z_i)^T\{\bar{\varphi}_i-(1,Z_i)\bar{\beta}\}=0.$$

By a Taylor expansion of the left hand side of the above equation around β^*, we have

$$\frac{1}{n}\sum_{i=1}^{n}(1,Z_i)^T\{\bar{\varphi}_i-(1,Z_i)\beta^*\}+\frac{1}{n}\sum_{i=1}^{n}(1,Z_i)^T(1,Z_i)(\bar{\beta}-\beta^*)=0.$$

Rearranging terms leads to

$$n^{1/2}(\bar{\beta}-\beta^*)=\Big(\frac{1}{n}\sum_{i=1}^{n}(1,Z_i)^T(1,Z_i)\Big)^{-1}n^{-1/2}\sum_{i=1}^{n}(1,Z_i)^T\{\bar{\varphi}_i-(1,Z_i)\beta^*\}.$$

By Central Limit Theorem and Slutsky Theorem, the asymptotic variance of $n^{1/2}(\bar{\beta}-\beta^*)$ can be given as

$$\{E(1,Z_i)^T(1,Z_i)\}^{-1}\cdot E\{(1,Z_i)^T(1,Z_i)(\bar{\varphi}_i-(1,Z_i)\beta^*)^2\}\cdot\{E(1,Z_i)^T(1,Z_i)\}^{-1}.$$

In addition, under standard regularity conditions and by a standard M-estimation theory ((van der Vaart, 1998)), $\bar{\beta}$ converges in probability to β^*, which implies (S.14) holds.

Second, we use $\hat\beta$ and $\hat\varphi$ to replace $\bar\beta$ and $\bar\varphi$ in (S.14) and then consider the difference between them. Denote

$$\tilde{E}\Big[(1,Z)^T(1,Z)\Big\{(\hat\varphi-(1,Z)\hat\beta)^2-(\bar\varphi-(1,Z)\bar\beta)^2\Big\}\Big] \equiv \tilde{E}\Big[(1,Z)^T(1,Z)(\hat\varphi_c^2-\bar\varphi_c^2)\Big].$$

Since $(1,Z)$ is bounded, it suffices to consider $\tilde{E}\{\hat\varphi_c^2-\bar\varphi_c^2\}$. Using the equality $a^2-b^2=2(a-b)b+(a-b)^2$ and the Cauchy-Schwartz inequality, we find that

$$\begin{aligned}\big|\tilde{E}\{\hat\varphi_c^2-\bar\varphi_c^2\}\big| &= \big|\tilde{E}\{2(\hat\varphi_c-\bar\varphi_c)\bar\varphi_c+(\hat\varphi_c-\bar\varphi_c)^2\}\big| \\ &\leq 2\tilde{E}^{1/2}(\bar\varphi_c^2)\tilde{E}^{1/2}\{(\hat\varphi_c-\bar\varphi_c)^2\}+\tilde{E}\{(\hat\varphi_c-\bar\varphi_c)^2\}.\end{aligned}$$

Using $(a-b)^2\leq 2(a^2+b^2)$,

$$\begin{aligned}\tilde{E}\{(\hat\varphi_c-\bar\varphi_c)^2\} &= \tilde{E}[\{\hat\varphi-\bar\varphi-(1,Z)^T(\hat\beta-\bar\beta)\}^2] \\ &\leq 2\tilde{E}\{(\hat\varphi-\bar\varphi)^2\}+2\tilde{E}\{(1,Z)^T(1,Z)(\hat\beta-\bar\beta)^2\}\end{aligned}$$

The term $\tilde{E}\{(1,Z)^T(1,Z)(\hat\beta-\bar\beta)^2\}=o_p(1)$ due to Theorem 2(a). Note that

$$\tilde{E}\{(\hat\varphi-\bar\varphi)^2\}=\tilde{E}\{(\delta_1+\delta_2+\delta_3)^2\}\leq 3\tilde{E}(\delta_1^2)+3\tilde{E}(\delta_2^2)+3\tilde{E}(\delta_3^2),$$

Therefore, to show (S.13), it is sufficient to show that

$$\tilde{E}(\delta_1^2)=o_p(1),\tilde{E}(\delta_2^2)=o_p(1),\tilde{E}(\delta_3^2)=o_p(1). \tag{S.15}$$

The following proof is divided into three steps, considering $\tilde{E}(\delta_1^2)$, $\tilde{E}(\delta_2^2)$ and $\tilde{E}(\delta_3^2)$, respectively.

Step 1. Using Assumptions 1(i)–(ii), $|1-T/\pi(X;\bar{\gamma})| \leq 1+1/\pi(X;\bar{\gamma}) \leq 1+e^{-B_0}$, and

$$\begin{aligned}\tilde{E}(\delta_1^2) &= \tilde{E}\Big[\{\hat{\alpha}_1^T f(X) - \bar{\alpha}_1^T f(X)\}^2\Big\{1-\frac{T}{\pi(X;\bar{\gamma})}\Big\}^2\Big] \\ &\leq (1+e^{-B_0})^2\tilde{E}\Big[\{\hat{\alpha}_1^T f(X) - \bar{\alpha}_1^T f(X)\}^2\Big] \\ &\leq (1+e^{-B_0})^2 C_0^2 ||\hat{\alpha}_1 - \bar{\alpha}_1||_1^2.\end{aligned}$$

Then in the event Ω_0,

$$\tilde{E}(\delta_1^2) \leq (1+e^{-B_0})^2 C_0^2 M_3^2 (|S_{\bar{\gamma}}|\lambda_0 + |S_{\bar{\alpha}_1}|\lambda_1)^2. \tag{S.16}$$

Step 2. By Assumptions 1(i)-(ii), $|w(X;\bar{\gamma})| \leq e^{-B_0}$. And by using the mean value theorem yields for some $u \in (0,1)$,

$$\begin{aligned}|\frac{1}{\pi(X;\hat{\gamma})} - \frac{1}{\pi(X;\bar{\gamma})}| &= |e^{-u\hat{\gamma}^T f(X) - (1-u)\bar{\gamma}^T f(X)}\{\hat{\gamma}-\bar{\gamma}\}^T f(X)| \\ &= |w(X;\bar{\gamma})e^{-u(\hat{\gamma}-\bar{\gamma})^T f(X)}\{\hat{\gamma}-\bar{\gamma}\}^T f(X)| \\ &\leq e^{C_0||\hat{\gamma}-\bar{\gamma}||_1}|w(X;\bar{\gamma})\{\hat{\gamma}-\bar{\gamma}\}^T f(X)|,\end{aligned} \tag{S.17}$$

which leads to that

$$\begin{aligned}\Big(\frac{1}{\pi(X;\hat{\gamma})} - \frac{1}{\pi(X;\bar{\gamma})}\Big)^2 &\leq e^{2C_0||\hat{\gamma}-\bar{\gamma}||_1}\Big[Tw^2(X;\bar{\gamma})\{\hat{\gamma}^T f(X) - \bar{\gamma}^T f(X)\}^2\Big] \\ &\leq e^{2C_0||\hat{\gamma}-\bar{\gamma}||_1 - B_0}\Big[Tw(X;\bar{\gamma})\{\hat{\gamma}^T f(X) - \bar{\gamma}^T f(X)\}^2\Big].\end{aligned}$$

Then in the event $\Omega_0 \cap \Omega_1 \cap \Omega_2$, we have

$$\begin{aligned}\tilde{E}(\delta_2^2) &= \tilde{E}\Big[T\{Y - m_1(X;\bar{\alpha}_1)\}^2\Big(\frac{1}{\pi(X;\hat{\gamma})} - \frac{1}{\pi(X;\bar{\gamma})}\Big)^2\Big] \\ &\leq e^{2C_0||\hat{\gamma}-\bar{\gamma}||_1 - B_0}\tilde{E}\Big[Tw(X;\bar{\gamma})\{Y - \bar{\alpha}_1^T f(X)\}^2\{\hat{\gamma}^T f(X) - \bar{\gamma}^T f(X)\}^2\Big] \\ &\leq e^{2C_0||\hat{\gamma}-\bar{\gamma}||_1 - B_0}\Big\{(D_0^2 + D_1^2)M_2|S_{\bar{\gamma}}|\lambda_0^2 + \{(D_0^2 + D_1^2)B_2 + B_3\}M_2^2|S_{\bar{\gamma}}|^2\lambda_0^3\Big\}\end{aligned} \tag{S.18}$$

The last inequality is due to (S.8).

Step 3. Writing $1/\pi(X;\hat{\gamma}) - 1/\pi(X;\bar{\gamma}) = e^{-\bar{\gamma}^T f(X)}\{e^{-\hat{\gamma}^T f(X) + \bar{\gamma}^T f(X)} - 1\}$ and by Assumption 1(i)-(ii),

$$\begin{aligned}\tilde{E}(\delta_3^2) &= \tilde{E}\Big[T\{\hat{\alpha}_1^T f(X) - \bar{\alpha}_1^T f(X)\}^2\Big(\frac{1}{\pi(X;\hat{\gamma})} - \frac{1}{\pi(X;\bar{\gamma})}\Big)^2\Big] \\ &= E\Big[T\{\hat{\alpha}_1^T f(X) - \bar{\alpha}_1^T f(X)\}^2 e^{-2\bar{\gamma}^T f(X)}\{e^{-\hat{\gamma}^T f(X) + \bar{\gamma}^T f(X)} - 1\}^2\Big] \\ &\leq e^{-B_0}\big(1 + e^{C_0||\hat{\gamma}-\bar{\gamma}||_1}\big)^2 E\Big[Tw(X;\bar{\gamma})\{\hat{\alpha}_1^T f(X) - \bar{\alpha}_1^T f(X)\}^2\Big].\end{aligned}$$

Then in the event Ω_0,

$$\tilde{E}(\delta_3^2) \leq e^{-B_0}\big(1 + e^{C_0||\hat{\gamma}-\bar{\gamma}||_1}\big)^2 M_3(|S_{\bar{\gamma}}|\lambda_0^2 + |S_{\bar{\alpha}_1}|\lambda_1^2) \tag{S.19}$$

Therefore, result (S.15) follows from (S.16), (S.18) and (S.19), provided that

$$|S_{\bar{\gamma}}|\lambda_0 + |S_{\bar{\alpha}_1}|\lambda_1 = o_p(1),$$

which is equivalent to

$$(|S_{\bar{\gamma}} + |S_{\bar{\alpha}_1}|)\sqrt{\log(q)} = o_p(n^{1/2}).$$

□

5. Proof of Theorem 2

We only consider generalized linear outcome model (2.2), in that linear outcome model (3.12) is a special case of it.

Proof of Theorem 2(a). Similar to the proof of Theorem 1(a), it suffices to show that

$$\sup_{k=1,\dots,K} \left|\tilde{E}\{(\hat{\varphi}-\bar{\varphi})\phi_k(Z)\}\right| \le M_4(|S_{\bar{\gamma}}|\lambda_0 + |S_{\bar{\alpha}_1}|\lambda_1)\lambda_1, \tag{S.20}$$

for a positive constant M_4. The decomposition of (S.10) yields that

$$\tilde{E}\{(\hat{\varphi}-\bar{\varphi})\phi_k(Z)\} = \Delta_{1,k} + \Delta_{2,k} + \Delta_{3,k},$$

where

$$\begin{aligned}
\Delta_{1,k} &= \tilde{E}\Big[\Big\{\psi\{\hat{\alpha}_1^T g(X)\} - \psi\{\bar{\alpha}_1^T g(X)\}\Big\}\Big\{1 - \frac{T}{\pi^*(X)}\Big\}\phi_k(Z)\Big],\\
\Delta_{2,k} &= \tilde{E}\Big[T\Big\{Y - \psi\{\bar{\alpha}_1^T g(X)\}\Big\}\Big\{\frac{1}{\hat{\pi}(X)} - \frac{1}{\pi^*(X)}\Big\}\phi_k(Z)\Big],\\
\Delta_{3,k} &= \tilde{E}\Big[\Big\{\psi\{\hat{\alpha}_1^T g(X)\} - \psi\{\bar{\alpha}_1^T g(X)\}\Big\}\Big\{\frac{T}{\pi^*(X)} - \frac{T}{\hat{\pi}(X)}\Big\}\phi_k(Z)\Big],
\end{aligned}$$

Under the condition that propensity score model (3) is correctly specified, i.e., $\pi(X;\bar{\gamma}) = \pi^*(X)$. Next we consider $\Delta_{1,k}$, $\Delta_{2,k}$ and $\Delta_{3,k}$ respectively.

Step 1. The term $\Delta_{1,k}$ can be decomposed as

$$\Delta_{1,k} = (\tilde{E} - E)\Big(\big[\psi\{\hat{\alpha}_1^T g(X)\} - \psi\{\bar{\alpha}_1^T g(X)\}\big]\Big\{1 - \frac{T}{\pi^*(X)}\Big\}\phi_k(Z)\Big)$$
$$+ E\Big(\big[\psi\{\hat{\alpha}_1^T g(X)\} - \psi\{\bar{\alpha}_1^T g(X)\}\big]\Big\{1 - \frac{T}{\pi^*(X)}\Big\}\phi_k(Z)\Big),$$

denoted as $\Delta_{1,k1} + \Delta_{1,k2}$. Take $r = M_3(|S_{\bar{\gamma}}|\lambda_0 + |S_{\bar{\alpha}_1}|\lambda_1)$ in Lemma 4, then in the event $\Omega_0 \cap \Omega_3$, we have $||\hat{\alpha}_1 - \bar{\alpha}_1||_1 \leq r$ and hence

$$|\Delta_{1,k1}| \leq B_4 M_3(|S_{\bar{\gamma}}|\lambda_0 + |S_{\bar{\alpha}_1}|\lambda_1)\lambda_0. \tag{S.21}$$

In addition, by the mean value theorem,

$$|\Delta_{1,k2}| \leq ||\hat{\alpha}_1 - \bar{\alpha}_1||_1 \sup_{j=0,1,\ldots,q} \Big| E\Big[\psi'\{\tilde{\alpha}_1^T g(X)\} g_j(X)\Big\{1 - \frac{T}{\pi^*(X)}\Big\}\phi_k(Z)\Big]\Big| = 0, \tag{S.22}$$

where $\tilde{\alpha}_1$ lies between $\hat{\alpha}_1$ and $\bar{\alpha}_1$. Combining (S.21) and (S.22) yields that

$$\sup_{k=1,\ldots,K} |\Delta_{1,k}| \leq B_4 M_3(|S_{\bar{\gamma}}|\lambda_0 + |S_{\bar{\alpha}_1}|\lambda_1)\lambda_0 \tag{S.23}$$

Step 2. By a Taylor expansion for $\Delta_{2,k}$ yields for some $u \in (0,1)$

$$\Delta_{2,k} = -(\hat{\gamma} - \bar{\gamma})^T \tilde{E}\big[T\{Y - \psi(\alpha_1^T g(X))\} e^{-\bar{\gamma}^T f(X)} f(X)\phi_k(Z)\big]$$
$$+ (\hat{\gamma} - \bar{\gamma})^T \tilde{E}\big[T\{Y - \psi(\alpha_1^T g(X))\} e^{-u\hat{\gamma}^T f(X) - (1-u)\bar{\gamma}^T f(X)} f(X) f(X)^T \phi_k^2(Z)\big](\hat{\gamma} - \bar{\gamma})/2,$$

denoted as $\Delta_{2,k1} + \Delta_{2,k2}$. In the event Ω_4, $\sup_{k=1,\ldots,K} |\Delta_{2,k1}| \leq B_5 \lambda_1 M_2 |S_{\bar{\gamma}}|\lambda_0$.

And in the event $\Omega_0 \cap \Omega_5$,

$$
\begin{aligned}
& \sup_{k=1,\ldots,K} |\Delta_{2,k2}| \\
\le\ & \sup_{k=1,\ldots,K} e^{C_0||\hat{\gamma}-\bar{\gamma}||_1} (\hat{\gamma}-\bar{\gamma})^T \tilde{E}\Big[Tw(X;\bar{\gamma})\big|Y-\psi(\alpha_1^T g(X))\big| f(X) f(X)^T \phi_k^2(Z)\Big](\hat{\gamma}-\bar{\gamma})/2 \\
\le\ & e^{C_0||\hat{\gamma}-\bar{\gamma}||_1} B_6 \lambda_0 \{M_2 |S_{\bar{\gamma}}| \lambda_0\}^2.
\end{aligned}
$$

Thus we have

$$
\sup_{k=1,\ldots,K} |\Delta_{2,k}| \le B_5 \lambda_1 M_2 |S_{\bar{\gamma}}| \lambda_0 + e^{C_0||\hat{\gamma}-\bar{\gamma}||_1} B_6 \lambda_0 \{M_2 |S_{\bar{\gamma}}| \lambda_0\}^2 \qquad \text{(S.24)}
$$

in the event $\Omega_0 \cap \Omega_4 \cap \Omega_5$.

Step 3. By Assumptions 2(ii)-(iii), there exists a positive constant C_4, such that

$$
\begin{aligned}
|\psi\{\hat{\alpha}_1^T g(X)\} - \psi\{\bar{\alpha}_1^T g(X)\}| &= |\psi'\{\tilde{\alpha}_1^T g(X)\}\{\hat{\alpha}_1 - \bar{\alpha}_1\}^T g(X)| \\
&\le C_4 |\{\hat{\alpha}_1 - \bar{\alpha}_1\}^T g(X)|, \qquad \text{(S.25)}
\end{aligned}
$$

where $\tilde{\alpha}_1$ lies between $\hat{\alpha}_1$ and $\bar{\alpha}_1$. Combing (S.17) with (S.25) and by using Cauchy-Schwartz inequality, we have in the event $\Omega_0 \cap \Omega_6$ that

$$
\begin{aligned}
\sup_{k=1,\ldots,K} |\Delta_{3,k}| \le\ & C_4 e^{C_0||\hat{\gamma}-\bar{\gamma}||_1} \tilde{E}^{1/2}\Big\{Tw(X;\bar{\gamma})\{\hat{\gamma}^T f(X) - \bar{\gamma}^T f(X)\}^2\Big\} \\
& \times \tilde{E}^{1/2}\Big\{Tw(X;\bar{\gamma})\{\hat{\alpha}_1^T g(X) - \bar{\alpha}_1^T g(X)\}^2\Big\} \\
\le\ & C_4 e^{C_0||\hat{\gamma}-\bar{\gamma}||_1} (M_2 |S_{\bar{\gamma}}| \lambda_0^2)^{1/2} M_3 (|S_{\bar{\gamma}}| \lambda_0 + |S_{\bar{\alpha}_1}| \lambda_1) B_7 \lambda_0.
\end{aligned}
\qquad \text{(S.26)}
$$

Inequality (S.20) follows immediately from (S.23), (S.24) and (S.26) □

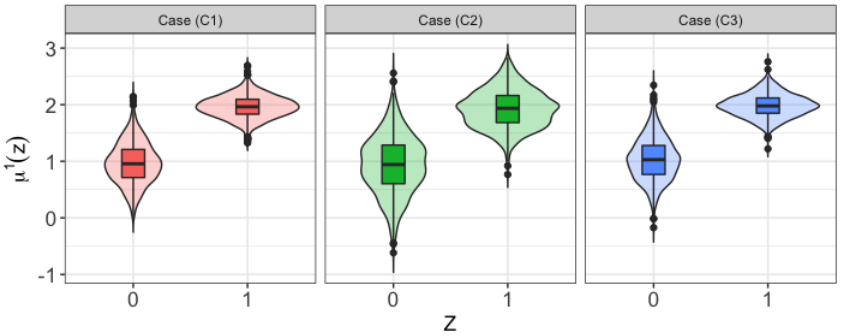

Figure 1: violin plots of the $\hat{\mu}^1(z)$ for cases (C1)-(C3) with $p = 400$.

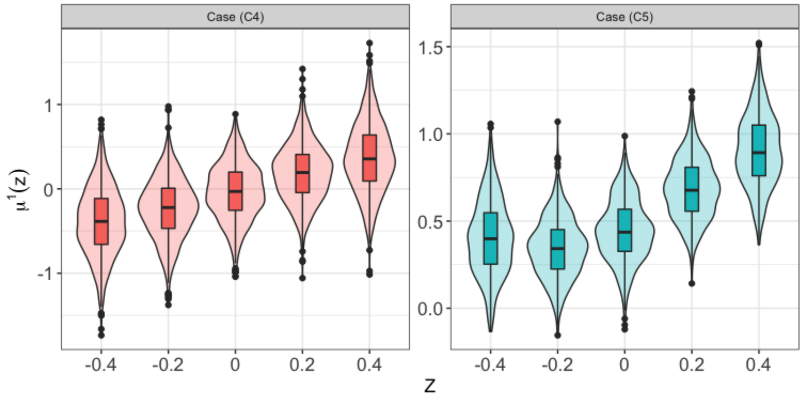

Figure 2: violin plots of $\hat{\mu}^1(z)$ for cases (C4) and (C5) with $p = 60, q = 420$.

Proof of Theorem 2(b). This can be shown similarly as Theorem 1(b). □

6. Additional results for simulation study

6.1 Results for simulation setup

Figures 1 and 2 display the violin plots of $\hat{\mu}^1(z)$ for cases (C1)-(C5) based on 1000 simulations of the points estimates. A violin plot is a blend of density plot and box-plot (Hintze and Nelson, 1998). As can be seen from Figures 1 and 2, all the density estimation curves (boundary of violin plot) seems to be well approximated by normal density function.

We compare the proposed method with competing AIPW methods of Fan et al. (2021) and Zimmert and Lechner (2019) discussed in Section 2 for continuous Z. We adopt the AIPW methods with full sample and four-fold cross-fitting as suggested in Fan et al. (2021). For PS and OR models, we set $f(X) = g(X) = (1, V^T, Z)^T$ and the associated tuning parameters are selected by 5-fold cross validation. As done in Fan et al. (2021), we use the Gaussian kernel throughout and the bandwidth is set as $h = \hat{h}_{opt} \times n^{1/5} \times n^{-2/7}$, where $\hat{h}_{opt}$ is a plug-in estimator of the optimal bandwidth that can be implemented using R package **KernSmooth** (Ruppert et al., 1995; Wand, 2015).

Table S1 summarizes the results of competing AIPW estimators for cases (C4) and (C5). When both PS and OR models are correctly specified, the two competing AIPW methods perform well in terms of Bias, CP90 and CP95. Nevertheless, when OR model is misspecified, their coverage proportions are away from the nominal values. This indicates that the competing approaches do not enjoy the property of doubly robust confidence intervals. As expected, the AIPW estimator with full sample tends to have larger Bias and smaller $\sqrt{\mathrm{Var}}$ than that of four-fold cross-fitting, in that sample-splitting may decrease bias and induce random errors in finite sample. In addition, by comparison of Table S1 with Table 2 in the manuscript, the proposed method has similar performance with competing AIPW methods when both PS and OR models are correctly specified. However, with a misspecified OR model, the proposed approach has smaller Bias and $\sqrt{\mathrm{Var}}$, and better coverage proportions.

6.2 Additional simulation: approximate sparsity

We explore the finite sample behaviors of the proposed estimators in approximate sparsity settings and consider three data generating scenarios:

(C6) Z is binary, $P(T=1|X)=\{1+\exp(-\sum_{i=1}^{d+1}X_i/i^2)\}^{-1}$, $Y^1=1+Z+\sum_{i=1}^{d}\{V_i+V_iZ+2V_i(1-Z)\}/i^2+\epsilon$.

(C7) Z consists of two binary variables, $P(T=1|X)=\{1+\exp(-\sum_{i=1}^{d+1}X_i/i^2)\}^{-1}$, $Y^1=1+Z_1-Z_1Z_2-\sum_{i=1}^{d}\{5V_i-V_iZ_1-V_iZ_1Z_2\}/i^2+\epsilon$.

(C8) Z is a continuous variable, $P(T=1|X)=\{1+\exp(\sum_{i=1}^{d+1}X_i/i^2)\}^{-1}$, $Y^1=Z(1-Z)\cos(Z)\log(Z+2)\exp(Z)+\sum_{i=1}^{d}(V_i+ZV_i)/i^2+\epsilon$,

where $X=(Z^T,V^T)^T$ and V_i be i-th element of V, the value of d is set to

6.2 Additional simulation: approximate sparsity

Table S1: Comparison of competing approaches for continuous Z

$\hat{\mu}^1(z)$	Bias	$\sqrt{\text{Var}}$	$\sqrt{\text{EVar}}$	CP90	CP95	Bias	$\sqrt{\text{Var}}$	$\sqrt{\text{EVar}}$	CP90	CP95
					Fan et al. (2021)'s AIPW with full sample					
	(C4) cor PS, cor OR					(C5) cor PS, mis OR				
$\hat{\mu}^1(-0.4)$	-0.018	0.406	0.359	0.858	0.930	-0.057	0.210	0.190	0.837	0.899
$\hat{\mu}^1(-0.2)$	-0.036	0.355	0.349	0.912	0.943	-0.037	0.191	0.184	0.861	0.917
$\hat{\mu}^1(0.0)$	-0.025	0.387	0.352	0.888	0.930	-0.038	0.207	0.195	0.856	0.925
$\hat{\mu}^1(0.2)$	-0.016	0.383	0.354	0.882	0.929	-0.050	0.198	0.197	0.862	0.914
$\hat{\mu}^1(0.4)$	-0.030	0.376	0.358	0.872	0.932	-0.050	0.218	0.209	0.828	0.881
	Fan et al. (2021) and Zimmert and Lechner (2019)'s AIPW with four-fold cross-fitting									
	(C4) cor PS, cor OR					(C5) cor PS, mis OR				
$\hat{\mu}^1(-0.4)$	-0.034	0.417	0.376	0.873	0.927	-0.047	0.296	0.300	0.826	0.899
$\hat{\mu}^1(-0.2)$	-0.001	0.441	0.399	0.875	0.931	-0.034	0.286	0.264	0.867	0.926
$\hat{\mu}^1(0.0)$	-0.016	0.417	0.397	0.869	0.928	-0.027	0.266	0.255	0.857	0.917
$\hat{\mu}^1(0.2)$	-0.032	0.388	0.377	0.892	0.942	-0.040	0.264	0.252	0.845	0.903
$\hat{\mu}^1(0.4)$	0.027	0.443	0.432	0.882	0.940	-0.009	0.266	0.242	0.848	0.908

ensure that the dimension of $g(X)$ is close to 200 or 400. $f(X)$ and $g(X)$ are specified as in (13), namely, $f(X) = g(X) = (1, V^T, Z, V^T Z)^T$ for case (C6), $f(X) = g(X) = (1, V^T, Z_1, Z_2, Z_1 Z_2, V^T Z_1, V^T Z_2, V^T Z_1 Z_2)^T$ for case (C7), $f(X) = (1, V^T, \Phi(Z)^T)^T$, $g(X) = (1, V^T, \Phi(Z)^T, V^T \phi_1(Z), ..., V^T \phi_K(Z), \Phi(Z) \otimes \Phi(Z))^T$ for case (C8) with $\Phi(Z)$ being cubic spline basis functions using three knots selected by the 25%, 50% and 75% sample quantiles of Z. The results of $\hat{\mu}^1(z)$ are summarized in Table S2 and Figure 3, from which one can see that the proposed estimators have a good performance for both binary variable Z, continuous variable Z and Z consisting of two binary variables, in a high-dimensional approximate sparsity setting. Moreover, under the scenario (C8), it is easy to see that the true CSTE curve $\mu^1(z) = z(1-z)\cos(z)\log(z+2)\exp(z)$ doesn't belong to the class $\{\beta_0 + \beta_1^T \Phi(z) : (\beta_0, \beta_1) \in \mathbb{R}^{K+1}\}$. The corresponding results of case (C8) shows that the approximation bias can be negligible and the proposed methods also perform well.

Table S2: Estimations of $\mu^1(z)$

$\hat{\mu}^1(z)$	Bias	$\sqrt{\text{Var}}$	$\sqrt{\text{EVar}}$	CP90	CP95	Bias	$\sqrt{\text{Var}}$	$\sqrt{\text{EVar}}$	CP90	CP95
					(C6) binary variable Z					
			$n = 500$, $p = 200$					$n = 500$, $p = 400$		
$\hat{\mu}^1(0)$	0.056	0.245	0.240	0.883	0.935	0.036	0.174	0.174	0.898	0.945
$\hat{\mu}^1(1)$	0.072	0.253	0.238	0.867	0.928	0.021	0.169	0.175	0.905	0.958
					(C7) two binary variables (Z_1, Z_2)					
			$n = 500$, $p = 200$					$n = 500$, $p = 400$		
$\hat{\mu}^1(0,0)$	0.043	0.563	0.551	0.885	0.937	0.047	0.548	0.547	0.879	0.944
$\hat{\mu}^1(0,1)$	0.025	0.556	0.549	0.895	0.952	0.021	0.535	0.550	0.912	0.954
$\hat{\mu}^1(1,0)$	-0.005	0.446	0.449	0.899	0.946	0.008	0.447	0.449	0.898	0.951
$\hat{\mu}^1(1,1)$	-0.048	0.341	0.342	0.896	0.939	-0.011	0.347	0.342	0.890	0.944
					(C8) continuous variable Z					
			$n = 500$, $p = 33$, $q = 210$					$n = 500$, $p = 63$, $q = 420$		
$\hat{\mu}^1(-0.4)$	0.019	0.219	0.197	0.851	0.922	0.016	0.209	0.195	0.857	0.923
$\hat{\mu}^1(-0.2)$	0.010	0.191	0.182	0.880	0.935	0.033	0.191	0.180	0.873	0.931
$\hat{\mu}^1(0.0)$	0.023	0.194	0.193	0.894	0.945	0.018	0.188	0.192	0.913	0.948
$\hat{\mu}^1(0.2)$	0.018	0.221	0.219	0.899	0.948	0.035	0.214	0.220	0.899	0.943
$\hat{\mu}^1(0.4)$	0.032	0.278	0.278	0.896	0.954	0.021	0.272	0.282	0.907	0.947

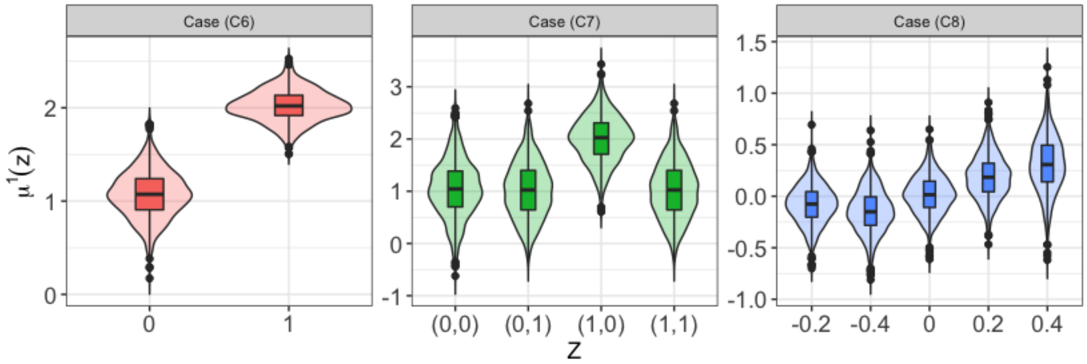

Figure 3: violin plots of $\hat{\mu}(z)$ for cases (C6)-(C8).

7. Additional results for Application

Table S3 presents the summary statistics of the variables in the dataset of Application. In addition, in the process of estimating a CSTE curve when Z is continuous, the proposed method uses cubic spline to approximate $\tau(z)$ and find the optimal number of knots by using grid search with Akaike information criterion (AIC) (Akaike, 1974) and Bayesian information criterion (BIC) (Schwarz, 2005). We conduct least squares by regressing $\varphi(Y, T, X; \hat{m}_1, \hat{\pi}) - \varphi(Y, 1-T, X; \hat{m}_0, 1-\hat{\pi}_0)$ on $\tilde{\Phi}(Z)$ to get the values of AIC and BIC, where $\tilde{\Phi}(Z)$ is cubic spline basis functions with number of knots ranging from 1 to 10. The corresponding results are given in Table S4.

Table S3: Descriptions of variables

Variable	**Name**	**Proportion/Mean**	**(Q1, Q3)**
Outcome	PASI 80		
	yes (%)	48.60%	
	no (%)	51.40%	
Treatment	Biologics		
	yes (%)	30.05%	
	no (%)	69.05%	
Covariates	Baseline PASI	11.99	(3.60, 17.83)
	Baseline BSA	20.35	(4.38, 30.00)
	Baseline DLQI	8.44	(3.00, 12.00)
	Age	41.70	(30.00, 53.00)
	BMI	24.45	(21.77, 26.12)
	Employment		
	Full-time (%)	61.29%	
	Part-time (%)	38.71%	
	Marital status		
	Married (%)	74.24%	
	Unmarried (%)	25.76%	
	Education		
	College and higher (%)	28.31%	
	High school and lower (%)	71.69%	
	Insurance		
	Free or commercial medical care (%)	8.28%	
	General government funded medical care (%)	91.72%	
	Nail involvement		
	yes (%)	6.75%	
	no (%)	93.25%	
	Sex		
	Female (%)	35.23%	
	Male (%)	64.77%	
	Smoking		
	Ex-smokers (%)	5.60%	
	Current smokers (%)	27.46%	
	Non-smokers (%)	66.94%	
	Comorbidity conditions		
	No comorbidity (%)	78.27%	
	Have comorbidity (%)	12.39%	
	Not clear (%)	9.34%	
	All the interactions between BMI and all other covariates.		
	All the interactions between Age and all other covariates.		
	All the interactions between Baseline PASI and all other covariates.		
	All the interactions between Baseline DLQI and all other covariates.		
	All the interactions between Baseline BSA and all other covariates.		

Note: DLQI and Baseline BSA denote self-reported Dermatology Life Quality Index and baseline Body Surface Area, respectively; Nail involvement refers to whether more than 5 nails are seriously affected.

Table S4: Values of AIC and BIC under different number of knots

	K	1	2	3	4	5	6	7	8	9	10
Baseline PASI	AIC	6894.3	**6893.8**	6896.4	6897.5	6898.4	6899.6	6901.1	6903.2	6904.7	6906.3
	BIC	6934.7	**6928.4**	6942.5	6949.4	6956.1	6963.0	6970.3	6978.2	6985.4	6992.7
Baseline DLQI	AIC	6874.4	6876.4	6876.6	**6872.4**	6873.1	6875.9	6874.6	6875.8	6876.5	6877.4
	BIC	6922.7	6916.8	**6909.0**	6918.5	6925.0	6933.5	6938.0	6939.2	6945.7	6952.3
Age	AIC	6893.9	6892.5	**6890.4**	6895.5	6897.1	6898.1	6898.6	6898.7	6898.0	6898.7
	BIC	6932.8	**6925.0**	6940.0	6947.3	6954.7	6961.5	6967.8	6973.7	6978.7	6985.2

Note: K is the number of knots for cubic spline.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control 19*, 716–723.

Buhlmann, P. and S. van de Geer (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications.* Springer Heidelberg.

Fan, Q., Y. C. Hsu, R. P. Lieli, and Y. Zhang (2021). Estimation of conditional average treatment effects with high-dimensional data. *Journal of Business and Economic Statistics 40*, 313–327.

Hintze, J. L. and R. D. Nelson (1998). Violin plots: a box plot-density trace synergism. *The American Statistician 52*, 181–184.

Ruppert, D., S. J. Sheather, and M. P. Wand (1995). An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association 90*, 1257–1270.

Schwarz, G. (2005). Estimating the dimension of a model. *Annals of Statistics 6*, 15–18.

Tan, Z. (2020a). Model-assisted inference for treatment effects using regularized calibrated estimation with high-dimensional data. *The Annals of Statistics 48*, 811–837.

Tan, Z. (2020b). Regularized calibrated estimation of propensity scores with model misspecification and high-dimensional data. *Biometrika 107*, 137–158.

van der Vaart, A. W. (1998). *Asymptotic statistics.* Cambridge University Press.

Wand, M. (2015). *KernSmooth: Functions for Kernel Smoothing Supporting Wand and Jones (1995).* https://CRAN.R-project.org/package=KernSmooth.

Zimmert, M. and M. Lechner (2019). Nonparametric estimation of causal heterogeneity under high-dimensional confounding. *arXiv:1908.08779v1*.