SUPPLEMENT TO "CONCORDANCE AND VALUE INFORMATION CRITERIA FOR OPTIMAL TREATMENT DECISION"

By Chengchun Shi, Rui Song and Wenbin Lu

North Carolina State University

This supplement is organized as follows. In Section 10, we derive the consistencies of our doubly-robust information criteria. In Section 11, we extend our information criteria to multiple stages. In Section 12, we investigate the theoretical and numerical properties of our information criteria under the settings where the contrast function does not take the monotonic linear index form. Conditions (A5') and (A6') are given in Section A. Section B contains discussion of some technical conditions. Proofs of Theorem 3.4, Theorem 7.1, Lemma 7.1 and Theorem 10.1 are given in Section C-G. We omit the proof of Theorem 3.1, since it is similar to the proof of Theorem 11.1. Section H summarizes some technical lemmas used in our proofs. Additional simulation results are reported in Section I. Some additional details regarding the cross-validation procedure in Section 5.1.2 are presented in Section J.

10. Consistencies of doubly-robust information criteria. Let

$$\widehat{\mathcal{M}}_V^{DR} = \arg\max_{\mathcal{M}\in\Omega} \mathrm{VIC}^{DR}(\hat{\theta}_{\mathcal{M}}), \quad \widehat{\mathcal{M}}_C^{DR} = \arg\max_{\mathcal{M}\in\Omega} \mathrm{CIC}^{DR}(\hat{\beta}_{\mathcal{M}}).$$

We use a shorthand and write $\pi_\alpha(x) = \pi(x,\alpha)$ and $h_\eta(x) = h(x,\eta)$. Define function $g(o,\beta,\alpha,\eta)$ as

$$\frac{1}{2}\mathrm{E}\left\{\frac{\{A_0-\pi_\alpha(X_0)\}\{Y_0-h_\eta(X_0)\}a}{\pi_\alpha(X_0)\{1-\pi_\alpha(X_0)\}\pi_\alpha(x)} - \frac{\{a-\pi_\alpha(x)\}\{y-h_\eta(x)\}A_0}{\pi_\alpha(x)\{1-\pi_\alpha(x)\}\pi_\alpha(X_0)}\right\}\mathbb{I}(X_0^T\beta > x^T\beta)$$
$$+\frac{1}{2}\mathrm{E}\left\{\frac{\{a-\pi_\alpha(x)\}\{y-h_\eta(x)\}A_0}{\pi_\alpha(x)\{1-\pi_\alpha(x)\}\pi_\alpha(X_0)} - \frac{\{A_0-\pi_\alpha(X_0)\}\{Y_0-h_\eta(X_0)\}a}{\pi_\alpha(X_0)\{1-\pi_\alpha(X_0)\}\pi_\alpha(x)}\right\}\mathbb{I}(x^T\beta > X_0^T\beta).$$

For $\theta = (c,\beta^T)^T$, let

$$\begin{aligned} V(\theta,\alpha,\eta) &= \mathrm{E}\left\{\frac{A_0\mathbb{I}(X_0^T\beta > -c)}{\pi(X_0,\alpha)} + \frac{(1-A_0)\mathbb{I}(X_0^T\beta \le -c)}{1-\pi(X_0,\alpha)}\right\}Y_0 \\ &- \mathrm{E}\left\{\frac{A_0\mathbb{I}(X_0^T\beta > -c)}{\pi(X_0,\alpha)} + \frac{(1-A_0)\mathbb{I}(X_0^T\beta \le -c)}{1-\pi(X_0,\alpha)} - 1\right\}h(X_0,\eta). \end{aligned}$$

We write $g(o,\beta,\alpha,\eta)$ as $g(o,\zeta)$ where ζ represents the $(p+q_1+q_2)$-dimensional parameter $(\beta^T,\alpha^T,\eta^T)^T$. Similarly, we write $V(\theta,\alpha,\eta)$ as $V(\bar{\zeta})$ where $\bar{\zeta}$ represents the $(p+1+q_1+q_2)$-dimensional parameter $(\theta^T,\alpha^T,\eta^T)^T$. Let $\partial_j g(o,\zeta)$ and $\partial_{jk} g(o,\zeta)$ stand for the derivatives

$$\frac{\partial g(o,\zeta)}{\partial \zeta^j} \text{ and } \frac{\partial^2 g(o,\zeta)}{\partial \zeta^j \partial \zeta^k}.$$

Let $\Delta_2 g(o,\zeta)$ be the Hessian matrix of $g(o,\zeta)$. Define $\zeta_0 = (\beta_0^T,(\alpha^*)^T,(\eta^*)^T)^T$ and $\bar{\zeta}_0 = (c_0,\zeta_0^T)^T$. We impose the following conditions.

(A3'.)(i) Assume there exists some constants $0 < c_1, c_2 < 1$ such that $c_1 \le \pi(x,\alpha) \le c_2$ for all x and α in a small neighborhood of α^*.
(ii) Assume $\sup_x \mathrm{E}(Y_0^2|X_0 = x) = O(1)$, $\sup_x \sup_{\eta:\|\eta-\eta^*\|_2\le\epsilon} |h_\eta(x)| = O(1)$ for some sufficiently small $\epsilon > 0$.
(A7.)(i)Assume

$$\|\hat{\alpha} - \alpha^*\|_2 = O_p(n^{-1/2}) \text{ and } \|\hat{\eta} - \eta^*\|_2 = O_p(n^{-1/2}).$$

(ii) Assume for all α and η in a neighborhood of α^* and η^*,

$$\mathrm{E}|\pi(X_0,\alpha) - \pi(X_0,\alpha^*)|_2^2 = O(\|\alpha-\alpha^*\|_2^2),$$
$$\mathrm{E}|h(X_0,\eta) - h(X_0,\eta^*)|_2^2 = O(\|\eta-\eta^*\|_2^2).$$

(A8.)Assume the class of functions $\{\pi(x,\alpha) : \alpha \in \mathbb{R}^{q_1}\}$, $\{h(x,\eta) : \eta \in \mathbb{R}^{q_2}\}$ belong to the VC subgraph class (cf. Section 2.6, van der Vaart and Wellner, 1996) with finite VC indexes.
(A9.)(i) Assume $V^{DR}(\theta_0) > V^{DR}(0), V^{DR}(\theta_0) > \sup_{\theta\in\tilde{N}_{\varepsilon_0}^c\cap\tilde{S}(\theta_0)} V^{DR}(\theta) > 0$ for some constant $0 < \varepsilon_0 \le \delta$.
(ii) Assume

$$\mathrm{E}\Big\{ \sup_{\substack{\|\theta-\theta_0\|_2\le\varepsilon \\ \theta=(c,\beta^T)^T}} \big|\mathbb{I}(X_0^T\beta > -c) - \mathbb{I}(X_0^T\beta_0 > -c_0)\big| \Big\} = O(\varepsilon),$$

as $\varepsilon \to 0$.
(iii) There exist some constants $\bar{c}_1, \bar{c}_2 > 0$ such that

$$\bar{c}_1\|\theta_0-\theta\|_2^2 \le V^{DR}(\theta_0) - V^{DR}(\theta) \le \bar{c}_2\|\theta_0-\theta\|_2^2, \quad \forall\theta \in \tilde{N}_{\varepsilon_0}\cap\tilde{S}(\theta_0).$$

(iv) Assume V is uniformly continuous. Besides, for any $\bar{\zeta}$ in a small neighborhood of $\bar{\zeta}_0$,

$$V(\bar{\zeta}) = V(\bar{\zeta}_0) + \frac{\partial V(\bar{\zeta}_0)}{\partial\bar{\zeta}}(\bar{\zeta}-\bar{\zeta}_0) + \frac{1}{2}(\bar{\zeta}-\bar{\zeta}_0)^T\Delta_2 V(\bar{\zeta}_0)(\bar{\zeta}-\bar{\zeta}_0) + o(1)\|\bar{\zeta}-\bar{\zeta}_0\|_2^2.$$

(A10.)(i) Assume $C^{DR}(\beta_0) > C^{DR}(0)$, $C^{DR}(\beta_0) > \sup_{\beta \in N^c_{\varepsilon_0} \cap S(\beta_0)} C^{DR}(\beta)$ for some $0 < \varepsilon_0 \leq \delta$.
(ii) There exist some constants $\bar{c}_1, \bar{c}_2 > 0$ such that

$$\bar{c}_1 \|\beta_0 - \beta\|_2^2 \leq C^{DR}(\beta_0) - C^{DR}(\beta) \leq \bar{c}_2 \|\beta_0 - \beta\|_2^2, \quad \forall \beta \in N_{\varepsilon_0} \cap S(\beta_0).$$

(iii) Assume function $g(o, \zeta)$ is twice continuously differentiable for all ζ in a small neighborhood of ζ_0.
(iv) Assume there's an integrable function $K(o)$ such that for all o and ζ in a small neighborhood of ζ_0,

$$\|\Delta_2 g(o, \zeta) - \Delta_2 g(o, \zeta_0)\|_2 \leq K(o) \|\zeta - \zeta_0\|_2.$$

(v) $\max_j \mathrm{E}|\partial_j g(O_0, \zeta_0)| < \infty$, $\max_{ij} \mathrm{E}|\partial_{ij} g(O_0, \zeta_0)| < \infty$.

Assumptions (A7) and (A8) are not restrictive. Under certain regularity conditions, Assumption (A7)(i) holds for misspecified models (White, 1982). Assumption (A7)(ii) holds if π and h are Lipschitz continuous in α and η. Assumption (A8) holds when π and h are fitted by generalized linear models. Assumptions (A3')(i), (A9) and (A10) are similar to (A3), (A5) and (A6). The following theorem states the consistencies of our doubly-robust information criteria.

THEOREM 10.1. *Suppose either π or h is correctly specified. Set $\kappa_n = c_n \max(nR_n^2, \sqrt{nR_n}, n^{1/3})$ for some $c_n \to \infty$. If $\kappa_n = o(n)$, under Assumptions (A1),(A2), (A3'), (A4), (A7)-(A9), we have*

$$Pr(\widehat{\mathcal{M}}_V^{DR} = \mathcal{M}_{\beta_0}) \to 1.$$

Set $\kappa_n = n(R_n^{(1)})^2 \log(n)$. If $\kappa_n = o(n)$, under Assumptions (A1),(A2), (A3'),(A4),(A7),(A8) and (A10), we have

$$Pr(\widehat{\mathcal{M}}_C^{DR} = \mathcal{M}_{\beta_0}) \to 1.$$

11. Extensions to multiple stages.

11.1. *Two-stage study.* To illustrate the idea, we first consider a two-stage study. Let $O_0 = (X_0^{(1)}, A_0^{(1)}, X_0^{(2)}, A_0^{(2)}, Y_0)$ where Y_0 denotes the final response, $A_0^{(1)}$ and $A_0^{(2)}$ refer to the treatments patient receives at time point t_1 and t_2 respectively, $X_0^{(1)} \in \mathbb{R}^{p_1}$ stands for the baseline covariates and $X_0^{(2)}$ denotes some intermediate covariates collected on the patient between t_1 and t_2. Let $\overline{X}_0^{(2)} = \{(X_0^{(1)})^T, A_0^{(1)}, (X_0^{(2)})^T\}^T$. Assume $A_0^{(1)}$ and $A_0^{(2)}$ are binary

treatments. For any $a^{(1)}, a^{(2)} \in \{0,1\}$, let $X_0^{(2)*}(a^{(1)})$ and $Y_0^*(a^{(1)}, a^{(2)})$ be the potential outcomes of the patient if he/she receives treatment a_1 at t_1 and treatment a_2 at t_2. For a given treatment regime $\bar{d} = (d_1, d_2)$, define the potential outcome

$$Y^*(\bar{d}) = \sum_{a^{(1)}, a^{(2)} \in \{0,1\}} Y_0^*(a^{(1)}, a^{(2)}) \mathbb{I}\{d_1(X_0^{(1)}) = a^{(1)}, d_2(\overline{X}_0^{(2)*}) = a^{(2)}\},$$

where $\overline{X}_0^{(2)*}$ is a shorthand for $[(X_0^{(1)})^T, a^{(1)}, \{X_0^{(2)*}(a^{(1)})^T\}]^T$. Let $\bar{d}^{opt} = (d_1^{opt}, d_2^{opt}) = \arg\max_{\bar{d}} \mathrm{E} Y^*(\bar{d})$. Define the Q-function and the contrast function as

$$\begin{aligned}
&Q^{(2)}(\bar{x}^{(2)}, a^{(2)}) = \mathrm{E}(Y | \overline{X}_0^{(2)} = \bar{x}^{(2)}, A_0^{(2)} = a^{(2)}),\\
&\tau^{(2)}(\bar{x}^{(2)}) = Q^{(2)}(\bar{x}^{(2)}, 1) - Q^{(2)}(\bar{x}^{(2)}, 0),\\
&Q^{(1)}(x^{(1)}, a^{(1)}) = \mathrm{E}\left(Q^{(2)}\{\overline{X}_0^{(2)}, d_2^{opt}(\overline{X}_0^{(2)})\} \mid X_0^{(1)} = x^{(1)}, A_0^{(1)} = a^{(1)}\right),\\
&\tau^{(1)}(x^{(1)}) = Q^{(1)}(x^{(1)}, 1) - Q^{(1)}(x^{(1)}, 0).
\end{aligned}$$

Under the following three assumptions,

(C1.) $Y_0 = \sum_{a^{(1)}, a^{(2)}} Y_0^*(a^{(1)}, a^{(2)}) \mathbb{I}(A_0^{(1)} = a^{(1)}, A_0^{(2)} = a^{(2)})$,
$X_0^{(2)} = \sum_{a^{(1)}} X_0^{(2)*}(a^{(1)}) \mathbb{I}(A_0^{(1)} = a^{(1)})$,
(C2.) For any $a^{(1)}, a^{(2)} \in \{0,1\}$, $A_0^{(2)} \perp\!\!\!\perp \{Y_0^*(a^{(1)}, a^{(2)}), X_0^{(2)*}(a_1)\} | \{\overline{X}_0^{(2)}, A_0^{(1)}\}$, and $A_0^{(1)} \perp\!\!\!\perp \{Y_0^*(a^{(1)}, a^{(2)}), X_0^{(2)*}(a^{(1)})\} | X_0^{(1)}$,
(C3.) There exists some constant $0 < c_1 < 1$ such that $\Pr(A_0^{(2)} = a^{(2)} | \overline{X}_0^{(2)} = \bar{x}^{(2)}) > c_1$ and $\Pr(A_0^{(1)} = a^{(1)} | X_0^{(1)} = x^{(1)}) > c_1$, for any $a^{(1)}, a^{(2)}, x^{(1)}, x^{(2)}$, we have

$$d_1^{opt}(x^{(1)}) = \mathbb{I}\{\tau^{(1)}(x^{(1)}) > 0\} \quad \text{and} \quad d_2^{opt}(\bar{x}^{(2)}) = \mathbb{I}\{\tau^{(2)}(\bar{x}^{(2)}) > 0\}.$$

Assumption (C2) is the sequential randomization assumption, which automatically holds in the sequential multiple assignment randomized trial (SMART) studies.

For simplicity, we focus on the fixed-p scenario. The case where the dimension of the covariates grows much faster than the sample size can be similarly discussed. The observed data are summarized as

$$\left\{O_i = \left(X_i^{(1)}, A_i^{(1)}, X_i^{(2)}, A_i^{(2)}, Y_i\right)\right\}_{i=1}^n.$$

We focus on the class of linear decision rules $d_1 = \mathbb{I}(\beta_1^T x^{(1)} + c_1 > 0)$, $d_2 = \mathbb{I}(\beta_2^T \bar{x}^{(2)} + c_2 > 0)$ and select the support of β_1 and β_2 via backward

induction. Assume

$$\tau^{(2)}(\bar{x}^{(2)}) = G^{(2)}(\beta_{0,2}^T \bar{x}^{(2)} + c_{0,2}),$$

for some $\beta_{0,2} \in \mathbb{R}^{p_2}$, $c_{0,2} \in \mathbb{R}$ and some monotonically increasing function $G^{(2)}$ with $G^{(2)}(0) = 0$. Let Ω_2 be the space of all candidate models for β_2. For a given model $\mathcal{M}_2 \in \Omega_2$, let $\hat{\theta}_{2,\mathcal{M}_2} = (\hat{c}_{2,\mathcal{M}_2}, \hat{\beta}_{2,\mathcal{M}_2}^T)^T$ be some estimator for $\theta_{0,2} = (c_{0,2}, \beta_{0,2}^T)^T$ on the restricted model space. These estimators can be obtained via robust learning or CAL. Let $\pi_0^{(2)}(\bar{x}^{(2)}) = \Pr(A_0^{(2)} = 1 | \overline{X}_0^{(2)} = \bar{x}^{(2)})$ and $\pi_0^{(1)}(x^{(1)}) = \Pr(A_0^{(1)} = 1 | X_0^{(1)} = x^{(1)})$. For any $\theta_2 = (c_2, \beta_2^T)^T$, define

$$\begin{aligned}
\widehat{V}^{(2)}(\theta_2) &= \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{A_i^{(2)} \mathbb{I}\{\beta_2^T \overline{X}_i^{(2)} > -c_2\}}{\pi_0^{(2)}(\overline{X}_i^{(2)})} + \frac{(1 - A_i^{(2)}) \mathbb{I}\{\beta_2^T \overline{X}_i^{(2)} \le -c_2\}}{1 - \pi_0^{(2)}(\overline{X}_i^{(2)})} \right\} Y_i, \\
\widehat{C}^{(2)}(\beta_2) &= \frac{1}{n(n-1)} \sum_{i \neq j} \left\{ \omega_i^{(2)} - \omega_j^{(2)} \right\} \mathbb{I}(\beta_2^T \overline{X}_i^{(2)} > \beta_2^T \overline{X}_j^{(2)}),
\end{aligned}$$

where

$$\omega_i^{(2)} = \left\{ \frac{A_i^{(2)}}{\pi_0^{(2)}(\overline{X}_i^{(2)})} - \frac{1 - A_i^{(2)}}{1 - \pi_0^{(2)}(\overline{X}_i^{(2)})} \right\} Y_i.$$

Let

$$\mathrm{VIC}^{(2)}(\theta_2) = n\widehat{V}^{(2)}(\theta_2) - \kappa_n^{(2)} \|\theta_2\|_0, \quad \mathrm{CIC}^{(2)}(\theta_2) = n\widehat{C}^{(2)}(\theta_2) - \kappa_n^{(2)} \|\theta_2\|_0.$$

We first use $\mathrm{VIC}^{(2)}$ and $\mathrm{CIC}^{(2)}$ to estimate $\mathcal{M}_{0,2}$, the support of $\theta_{0,2}$. Define

$$\begin{aligned}
\widehat{\mathcal{M}}_2^V &= \arg\max_{\mathcal{M}_2 \in \Omega_2} \mathrm{VIC}^{(2)}(\hat{\theta}_{2,\mathcal{M}_2}), \\
\widehat{\mathcal{M}}_2^C &= \arg\max_{\mathcal{M}_2 \in \Omega_2} \mathrm{CIC}^{(2)}(\hat{\theta}_{2,\mathcal{M}_2}).
\end{aligned}$$

Similar to Theorem 3.1, we can show $\mathrm{VIC}^{(2)}$ and $\mathrm{CIC}^{(2)}$ are consistent under certain conditions.

Let Ω_1 be the space of all candidate models for β_1. For any $\mathcal{M}_1 \in \Omega_1$, let $\hat{\theta}_{1,\mathcal{M}_1} = (\hat{c}_{1,\mathcal{M}_1}, \hat{\beta}_{1,\mathcal{M}_1}^T)^T$ be some estimator on the restricted model space. To introduce our information criteria, we define the pseudo outcomes

$$\begin{aligned}
Y_i^{(1),V} &= \left[\frac{A_i^{(2)} \hat{d}_2^V(\overline{X}_i^{(2)})}{\pi_0^{(2)}(\overline{X}_i^{(2)})} + \frac{(1 - A_i^{(2)})\{1 - \hat{d}_2^V(\overline{X}_i^{(2)})\}}{1 - \pi_0^{(2)}(\overline{X}_i^{(2)})} \right] Y_i, \\
Y_i^{(1),C} &= \left[\frac{A_i^{(2)} \hat{d}_2^C(\overline{X}_i^{(2)})}{\pi_0^{(2)}(\overline{X}_i^{(2)})} + \frac{(1 - A_i^{(2)})\{1 - \hat{d}_2^C(\overline{X}_i^{(2)})\}}{1 - \pi_0^{(2)}(\overline{X}_i^{(2)})} \right] Y_i,
\end{aligned}$$

where

$$\hat{d}_2^V(\bar{x}^{(2)}) = \mathbb{I}(\hat{\beta}^T_{2,\widehat{\mathcal{M}}_V^{(2)}}\bar{x}^{(2)} > -\hat{c}_{2,\widehat{\mathcal{M}}_V^{(2)}}), \quad \hat{d}_2^C(\bar{x}^{(2)}) = \mathbb{I}(\hat{\beta}^T_{2,\widehat{\mathcal{M}}_C^{(2)}}\bar{x}^{(2)} > -\hat{c}_{2,\widehat{\mathcal{M}}_C^{(2)}}).$$

For any $\theta_1 = (c_1, \beta_1^T)^T$, let

$$\begin{aligned}
\widehat{V}^{(1)}(\theta_1) &= \frac{1}{n}\sum_{i=1}^n \left\{ \frac{A_i^{(1)}\mathbb{I}\{\beta_1^T X_i^{(1)} > -c_1\}}{\pi_0^{(1)}(X_i^{(1)})} + \frac{(1-A_i^{(1)})\mathbb{I}\{\beta_1^T X_i^{(1)} \le -c_1\}}{1-\pi_0^{(1)}(X_i^{(1)})} \right\} Y_i^{(1),V}, \\
\widehat{C}^{(1)}(\beta_1) &= \frac{1}{n(n-1)}\sum_{i\neq j}\left\{\omega_i^{(1)} - \omega_j^{(1)}\right\}\mathbb{I}(\beta_1^T X_i^{(1)} > \beta_1^T X_j^{(1)}),
\end{aligned}$$

where

$$\omega_i^{(1)} = \left\{ \frac{A_i^{(1)}}{\pi_0^{(1)}(X_i^{(1)})} - \frac{1-A_i^{(1)}}{1-\pi_0^{(1)}(X_i^{(1)})} \right\} Y_i^{(1),C}.$$

Define

$$\mathrm{VIC}^{(1)}(\theta_1) = n\widehat{V}^{(1)}(\theta_1) - \kappa_n^{(1)}\|\theta_1\|_0, \quad \mathrm{CIC}^{(1)}(\theta_1) = n\widehat{C}^{(1)}(\theta_1) - \kappa_n^{(1)}\|\theta_1\|_0,$$

and

$$\begin{aligned}
\widehat{\mathcal{M}}_1^V &= \arg\max_{\mathcal{M}_1\in\Omega_1} \mathrm{VIC}^{(1)}(\hat{\theta}_{1,\mathcal{M}_1}), \\
\widehat{\mathcal{M}}_1^C &= \arg\max_{\mathcal{M}_1\in\Omega_1} \mathrm{CIC}^{(1)}(\hat{\theta}_{1,\mathcal{M}_1}).
\end{aligned}$$

When $\tau^{(1)}(x^{(1)}) = G^{(1)}(\beta_{0,1}^T x^{(1)} + c_{0,1})$ for some $\beta_{0,1}$, $c_{0,1}$ and monotonically increasing function $G^{(1)}$ with $G^{(1)}(0) = 0$, then $\widehat{\mathcal{M}}_1^V$ and $\widehat{\mathcal{M}}_1^C$ are consistent to $\mathcal{M}_{0,1} = \mathrm{support}(\beta_{0,1})$ under proper choice of $\kappa_n^{(1)}$, provided that $(\hat{c}_{1,\mathcal{M}_{0,1}}, \hat{\beta}^T_{1,\mathcal{M}_{0,1}})^T$ are consistent to $(c_{0,1}, \beta_{0,1}^T)^T$. Otherwise, $\mathrm{VIC}^{(1)}$ and $\mathrm{VIC}^{(1)}$ may select different models. More specifically, let

$$\begin{aligned}
\theta_1^V &= \{c_1^V, (\beta_1^V)^T\}^T = \arg\max_{\theta_1=(c_1,\beta_1^T)^T} \mathrm{E}\left\{\tau^{(1)}(X_0^{(1)})\mathbb{I}(\beta_1^T X_0^{(1)} + c_1 > 0)\right\}, \\
\beta_1^C &= \arg\max_{\beta_1} \mathrm{E}\left\{\tau^{(1)}(X_i^{(1)}) - \tau^{(1)}(X_j^{(1)})\right\}\mathbb{I}(\beta_1^T X_i^{(1)} > \beta_1^T X_j^{(1)}), \quad i \neq j.
\end{aligned}$$

Define $\mathcal{M}_1^V$ and $\mathcal{M}_1^C$ to be the support of θ_1^V and β_1^C, respectively. In the following, we show that $\widehat{\mathcal{M}}_1^V$ and $\widehat{\mathcal{M}}_1^C$ are consistent to $\mathcal{M}_1^V$ and $\mathcal{M}_1^C$.

Let $V(\theta_1,\theta_2)$ be the average potential outcome of patients following the regime $d_1(x^{(1)}) = \mathbb{I}(\theta_1^T x^{(1)} > 0)$, $d_2(\bar{x}^{(2)}) = \mathbb{I}(\theta_2^T \bar{x}^{(2)} > 0)$, and $V^{(1)}(\theta_1) = V(\theta_1, \theta_{0,2})$. Define

$$C^{(1)}(\beta_1) = \mathrm{E}\left\{\tau^{(1)}(X_i^{(1)}) - \tau^{(1)}(X_j^{(1)})\right\}\mathbb{I}(\beta_1^T X_i^{(1)} > \beta_1^T X_j^{(1)}), \quad i \neq j.$$

Let $\delta_V = \min_{j\in\mathcal{M}_1^V} |\theta_1^{V,j}|$ and $\delta_C = \min_{j\in\mathcal{M}_1^C} |\beta_1^{V,j}|$. For any $\delta > 0$, define

$$N_\delta^V = \{\theta_1 \in \mathbb{R}^{p_1+1} : \|\theta_1 - \theta_1^V\|_2 \le \delta\}, \quad N_\delta^C = \{\beta_1 \in \mathbb{R}^{p_1} : \|\beta_1 - \beta_1^C\|_2 \le \delta\},$$

and

$$S^V = \{\theta_1 \in \mathbb{R}^{p_1+1} : \|\theta_1\|_2 = \|\theta_1^V\|_2\}, \quad S^C = \{\beta_1 \in \mathbb{R}^{p_1} : \|\beta_1\|_2 = \|\beta_1^C\|_2\}.$$

For any $o = (x^{(1)}, a^{(1)}, x^{(2)}, a^{(2)}, y)$, define $g(o, \beta_1)$ as

$$\begin{aligned} & \frac{1}{2}\mathrm{E}\left\{\frac{\{A_0^{(1)} - \pi_0^{(1)}(x^{(1)})\}Y_0^{(1)}}{\pi_0^{(1)}(x^{(1)})\{1-\pi_0^{(1)}(x^{(1)})\}} - \frac{\{a^{(1)} - \pi_0^{(1)}(x^{(1)})\}y^{(1)}}{\pi_0^{(1)}(x^{(1)})\{1-\pi_0^{(1)}(x^{(1)})\}}\right\}\mathbb{I}(\beta_1^T X_0^{(1)} > \beta_1^T x^{(1)}) \\ + \quad & \frac{1}{2}\mathrm{E}\left\{\frac{(a^{(1)} - \pi_0^{(1)}(x^{(1)}))y^{(1)}}{\pi_0^{(1)}(x^{(1)})\{1-\pi_0^{(1)}(x^{(1)})\}} - \frac{\{A_0^{(1)} - \pi_0^{(1)}(x^{(1)})\}Y_0^{(1)}}{\pi_0^{(1)}(x^{(1)})\{1-\pi_0^{(1)}(x^{(1)})\}}\right\}\mathbb{I}(\beta_1^T x^{(1)} > \beta_1^T X_0^{(1)}), \end{aligned}$$

where

$$\begin{aligned} y_1^{(1)} &= \left\{\frac{a^{(2)}}{\pi_0^{(2)}(\bar{x}^{(2)})}d_2^{opt}(\bar{x}^{(2)}) + \frac{1-a^{(2)}}{1-\pi_0^{(2)}(\bar{x}^{(2)})}\{1 - d_2^{opt}(\bar{x}^{(2)})\}\right\} y, \\ Y_0^{(1)} &= \left\{\frac{A_0^{(2)}}{\pi_0^{(2)}(\overline{X}_0^{(2)})}d_2^{opt}(\overline{X}_0^{(2)}) + \frac{1-A_0^{(2)}}{1-\pi_0^{(2)}(\overline{X}_0^{(2)})}\{1 - d_2^{opt}(\overline{X}_0^{(2)})\}\right\} Y_0, \end{aligned}$$

Let

$$\phi_1(x^{(1)}, \beta_1) = \Pr(\beta_1^T X_0^{(1)} > \beta_1^T x^{(1)}), \qquad \phi_2(x^{(1)}, \beta_1) = \Pr(\beta_1^T X_0^{(1)} < \beta_1^T x^{(1)}).$$

We first introduce some conditions.

(C4.) Assume $\sup_{x^{(1)}} \mathrm{E}(Y_0^2 | X_0^{(1)} = x^{(1)}) < \infty$.
(C5.) Assume $\|\hat{\theta}_{2,\mathcal{M}_{0,2}} - \theta_{0,2}\|_2 = O_p(R_{n,2})$ for some $n^{-1/2} \le R_{n,2} \to 0$. Besides, assume

$$\mathrm{E}\Big\{\sup_{\substack{\|\theta_2 - \theta_{0,2}\|_2 \le \varepsilon \\ \theta_2 = (c_2, \beta_2^T)^T}} |\mathbb{I}(\beta_2^T X_0^{(2)} > -c_2) - \mathbb{I}(\beta_{0,2}^T X_0^{(2)} > -c_{0,2})|\Big\} = O(\varepsilon),$$

as $\varepsilon \to 0$.
(C6.) Assume $\|\hat{\theta}_{1,\mathcal{M}_1^V} - \theta_1^V\|_2 = O_p(R_{n,1})$ for some $n^{-1/2} \leq R_{n,1} \to 0$.
(C7.) (i)Assume $V(\theta_1, \theta_2)$ is twice continuously differentiable in a small neighborhood around $(\theta_1^V, \theta_{0,2})$.
(ii)Assume $V^{(1)}(\theta_1^V) > V^{(1)}(0)$ and $V^{(1)}(\theta_1^V) > \sup_{\theta_1 \in (N_{\varepsilon_0}^V)^c \cap S^V} V^{(1)}(\theta_1)$ for some $0 < \varepsilon_0 \leq \delta^V$.
(iii)Assume there exist some constants $\bar{c}_1, \bar{c}_2 > 0$ such that for any $\theta_1 \in N_{\varepsilon_0}^V \cap S^V$,

$$\bar{c}_1 \|\theta_1 - \theta_1^V\|_2^2 \leq V^{(1)}(\theta_1^V) - V^{(1)}(\theta_1) \leq \bar{c}_2 \|\theta_1 - \theta_1^V\|_2^2.$$

(iv)Assume

$$\mathrm{E}\Big\{ \sup_{\substack{\|\theta_1 - \theta_1^V\|_2 \leq \varepsilon \\ \theta_1 = (c_1, \beta_1^T)^T}} |\mathbb{I}(\beta_1^T X_0^{(1)} > -c_1) - \mathbb{I}\{(\beta_1^V)^T X_0^{(1)} > -c_1^V\}| \Big\} = O(\varepsilon),$$

as $\varepsilon \to 0$.
(C8.) Assume $\|\hat{\beta}_{1,\mathcal{M}_1^C} - \beta_1^C\|_2 = O_p(R_{n,1}^{(1)})$ for some $n^{-1/2} \leq R_{n,1}^{(1)} \to 0$.
(C9.) (i)Assume $\mathrm{E}|G^{(2)}(\beta_2^T X_0^{(2)} + c_2) - G^{(2)}(\beta_{0,2}^T X_0^{(2)} + c_{0,2})|^2 = O(\|\beta_2 - \beta_{0,2}\|_2^2) + O(|c_2 - c_{0,2}|_2^2)$ for any $\theta_2 = (c_2, \beta_2^T)^T$ in a small neighborhood of $\theta_{2,0}$.
(ii) $C^{(1)}(\beta_1^C) > C^{(1)}(0)$ and $C^{(1)}(\beta_1^C) > \sup_{\beta_1 \in (N_{\varepsilon_0}^C)^c \cap S^C} C^{(1)}(\beta_1)$ for some constants $0 < \varepsilon_0 \leq \delta^C$.
(iii) There exist some constants $\bar{c}_1, \bar{c}_2 > 0$ such that

$$\bar{c}_1 \|\beta_1^C - \beta_1\|_2^2 \leq C^{(1)}(\beta_1^C) - C^{(1)}(\beta_1) \leq \bar{c}_2 \|\beta_1^C - \beta_1\|_2^2, \quad \forall \beta_1 \in N_{\varepsilon_0}^C \cap S^C.$$

(iv) There exist some ψ_1, ψ_2 such that $\mathrm{E}\psi_1^2(X_0^{(1)}), \mathrm{E}\psi_2^2(X_0^{(1)}) < \infty$ and $|\phi_j(X_0^{(1)}, \beta_1) - \phi_j(X_0^{(1)}, \beta_1^C)| \leq \psi_j(X_0^{(1)}) \|\beta_1 - \beta_1^C\|_2$ for all $\beta_1 \in N_{\varepsilon_0}^C$, $j = 1, 2$.
(v) Function $g(o, \beta_1)$ is twice continuously differentiable for all $\beta_1 \in N_{\varepsilon_0}^C$.
(vi) There is an integrable function $K(o)$ such that for all o and $\beta_1 \in N_{\varepsilon_0}^C$,

$$\|\Delta_2 g(o, \beta_1) - \Delta_2 g(o, \beta_1^C)\|_2 \leq K(o) \|\beta_1 - \beta_1^C\|_2.$$

(vii) $\mathrm{E}|\partial_i g(O_0, \beta_1^C)|^2 < \infty$, $\mathrm{E}|\partial_{ij} g(O_0, \beta_1^C)| < \infty$.

THEOREM 11.1. *Assume (C1)-(C7) hold. If* $\kappa_n^{(1)} = o(n)$, *and*

$$\kappa_n^{(1)} \gg \max(nR_{n,1}^2, \sqrt{nR_{n,1}}, nR_{n,2}^2, \sqrt{nR_{n,2}}, n^{-1/3}).$$

Then conditional on the event $\widehat{\mathcal{M}}_2^V = \mathcal{M}_{0,2}$, *we have*

$$Pr\left(\widehat{\mathcal{M}}_1^V = \mathcal{M}_1^V\right) \to 1.$$

Assume (C1)-(C5), (C8)-(C9) hold. If $\kappa_n^{(1)} = o(n)$, *and*

$$\kappa_n^{(1)} \gg \max(n(R_{n,1}^{(1)})^2, \sqrt{nR_{n,2}}, nR_{n,2}^2),$$

Then conditional on the event $\widehat{\mathcal{M}}_2^C = \mathcal{M}_{0,2}$, *we have*

$$Pr\left(\widehat{\mathcal{M}}_1^C = \mathcal{M}_1^C\right) \to 1.$$

REMARK 11.1. *For a given model* $\mathcal{M}_1$, *if we obtain* $\hat{\theta}_{1,\mathcal{M}_1}$ *by maximizing* $\widehat{V}^{(1)}(\theta_1)$, *then Condition (C6) automatically holds with* $R_{n,1} = n^{-1/3}$. *Similarly, Condition (C8) holds with* $R_{n,1}^{(1)} = n^{-1/2}$ *when we obtain* $\hat{\beta}_{1,\mathcal{M}_1}$ *by maximizing* $\widehat{C}^{(1)}(\beta_1)$ *for any candidate model* $\mathcal{M}_1$. *For simplicity, we assume the propensity scores are known for each patient. A doubly-robust version of VIC and CIC can be similarly constructed.*

REMARK 11.2. *Compared to Theorem 3.1, we can see that conditions on* $\kappa_n^{(1)}$ *are strengthened in the backward induction algorithm, due to the variability of* $\hat{\beta}_{2,\widehat{\mathcal{M}}_2^V}$ *and* $\hat{\beta}_{2,\widehat{\mathcal{M}}_2^C}$. *Take CIC as an example, when estimating* $\hat{\beta}_{2,\mathcal{M}_2}$ *and* $\hat{c}_{2,\mathcal{M}_2}$ *via CAL for any* $\mathcal{M}_2$, *we have* $\hat{\beta}_{2,\mathcal{M}_{0,2}} = \beta_{0,2} + O_p(n^{-1/2})$ *and* $\hat{c}_{2,\mathcal{M}_{0,2}} = c_{0,2} + O_p(n^{-1/3})$. *Thus, CIC*$^{(2)}$ *is consistent when* $\kappa_n^{(2)} \to \infty$ *and* $\kappa_n^{(2)} = o(n)$ *while consistency of CIC*$^{(1)}$ *requires* $n^{1/3} \ll \kappa_n^{(1)} = o(n)$.

11.2. *Multi-stage study.* We generalize our information criteria to multi-stage studies. Assume treatment decisions are made at a finite number of time points $t_1, \ldots, t_K$. Data are summarized as

$$\left\{O_i = \left(X_i^{(1)}, A_i^{(1)}, \ldots, X_i^{(K)}, A_i^{(K)}, Y_i\right)\right\}_{i=1}^n,$$

where Y_i denotes the ith patient's response, $X_i^{(1)}$ stands for the baseline covariates, $X_i^{(j)}$ stands for the covariates collected between t_{j-1} and t_j for $2 \le j \le K$, and $A_i^{(j)}$ is the treatment received at time point t_j for $1 \le j \le K$.

Let $\overline{X}_i^{(k)} = (X_i^{(1)}, A_i^{(1)}, \ldots, X_i^{(k)})$, $\pi_{0,i}^{(k)}(\bar{x}^{(k)}) = \Pr(A_i^{(k)} = 1 \mid \overline{X}_i^{(k)} = \bar{x}^{(k)})$ for $1 \le k \le K$. Similar to the two-stage study, we select models via backward induction. For any $\theta_K = (c_K, \beta_K^T)^T$, define

$$\mathrm{VIC}^{(K)}(\theta_K) = \widehat{V}^{(K)}(\theta_K) - \kappa_n^{(K)}\|\theta_K\|_0, \mathrm{CIC}^{(K)}(\theta_K) = \widehat{C}^{(K)}(\theta_K) - \kappa_n^{(K)}\|\theta_K\|_0,$$

where

$$\widehat{V}^{(K)}(\theta_K) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{A_i^{(K)}\mathbb{I}\{\beta_K^T\overline{X}_i^{(K)} > -c_K\}}{\pi_0^{(K)}(\overline{X}_i^{(K)})} + \frac{(1-A_i^{(K)})\mathbb{I}\{\beta_K^T\overline{X}_i^{(K)} \le -c_K\}}{1-\pi_0^{(K)}(\overline{X}_i^{(K)})}\right] Y_i,$$

$$\widehat{C}^{(K)}(\beta_K) = \frac{1}{n(n-1)}\sum_{i\neq j}\left\{\omega_i^{(K)} - \omega_j^{(K)}\right\}\mathbb{I}(\beta_K^T\overline{X}_i^{(K)} > \beta_K^T\overline{X}_j^{(K)}),$$

$$\omega_i^{(K)} = \left\{\frac{A_i^{(K)}}{\pi_0^{(K)}(\overline{X}_i^{(K)})} - \frac{1-A_i^{(K)}}{1-\pi_0^{(K)}(\overline{X}_i^{(K)})}\right\} Y_i.$$

Based on $\mathrm{VIC}^{(K)}$ and $\mathrm{CIC}^{(K)}$, we can select models for the contrast function on the last stage, which we denoted by $\widehat{\mathcal{M}}_K^V$ and $\widehat{\mathcal{M}}_K^C$, accordingly.

Assume for now, we have $\widehat{\mathcal{M}}_j^V$ and $\widehat{\mathcal{M}}_j^C$ for $k+1 \le j \le K$. To obtain $\widehat{\mathcal{M}}_k^V$ and $\widehat{\mathcal{M}}_k^C$, we iteratively define the pseudo response $Y_i^{(k),V}$ and $Y_i^{(k),C}$ by

$$Y_i^{(K),V} = Y_i^{(K),C} = Y_i,$$

$$Y_i^{(j),V} = \left[\frac{A_i^{(j+1)}\hat{d}_{j+1}^V(\overline{X}_i^{(j+1)})}{\pi_0^{(j+1)}(\overline{X}_i^{(j+1)})} + \frac{(1-A_i^{(j+1)})\{1-\hat{d}_{j+1}^V(\overline{X}_i^{(j+1)})\}}{1-\pi_0^{(j+1)}(\overline{X}_i^{(j+1)})}\right] Y_i^{(j+1),V},$$

$$Y_i^{(j),C} = \left[\frac{A_i^{(j+1)}\hat{d}_{j+1}^C(\overline{X}_i^{(j+1)})}{\pi_0^{(j+1)}(\overline{X}_i^{(j+1)})} + \frac{(1-A_i^{(j+1)})\{1-\hat{d}_{j+1}^C(\overline{X}_i^{(j+1)})\}}{1-\pi_0^{(j+1)}(\overline{X}_i^{(j+1)})}\right] Y_i^{(j+1),C},$$

for $j = K-1, K-2, \ldots, k$, where

$$\begin{aligned}\hat{d}_{j+1}^V(\bar{x}^{(j+1)}) &= \mathbb{I}(\hat{\beta}^T_{j+1,\widehat{\mathcal{M}}_{j+1}^V}\bar{x}^{(j+1)} > -\hat{c}_{j+1,\widehat{\mathcal{M}}_{j+1}^V}),\\ \hat{d}_{j+1}^C(\bar{x}^{(j+1)}) &= \mathbb{I}(\hat{\beta}^T_{j+1,\widehat{\mathcal{M}}_{j+1}^C}\bar{x}^{(j+1)} > -\hat{c}_{j+1,\widehat{\mathcal{M}}_{j+1}^C}),\end{aligned}$$

where $\hat{c}_{j+1,\mathcal{M}_{j+1}}$ and $\hat{\beta}_{j+1,\mathcal{M}_{j+1}}$ correspond to some estimators on the restricted model space.

Our information criteria on the kth stage is defined by

$$\mathrm{VIC}^{(k)}(\theta_k) = \widehat{V}^{(k)}(\theta_k) - \kappa_n^{(k)}\|\theta_k\|_0, \mathrm{CIC}^{(k)}(\theta_k) = \widehat{C}^{(k)}(\theta_k) - \kappa_n^{(k)}\|\theta_k\|_0,$$

where

$$\widehat{V}^{(k)}(\theta_k) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{A_i^{(k)}\mathbb{I}\{\beta_k^T\overline{X}_i^{(k)} > -c_k\}}{\pi_0^{(k)}(\overline{X}_i^{(k)})} + \frac{(1-A_i^{(k)})\mathbb{I}\{\beta_k^T\overline{X}_i^{(k)} \le -c_k\}}{1-\pi_0^{(k)}(\overline{X}_i^{(k)})}\right] Y_i^{(k),V},$$

$$\widehat{C}^{(k)}(\beta_k) = \frac{1}{n(n-1)}\sum_{i\neq j}\left\{\omega_i^{(k)} - \omega_j^{(k)}\right\}\mathbb{I}(\beta_k^T\overline{X}_i^{(k)} > \beta_k^T\overline{X}_j^{(k)}),$$

$$\omega_i^{(k)} = \left\{\frac{A_i^{(k)}}{\pi_0^{(k)}(\overline{X}_i^{(k)})} - \frac{1-A_i^{(k)}}{1-\pi_0^{(k)}(\overline{X}_i^{(k)})}\right\} Y_i^{(k),C}.$$

Our information criteria are consistent when the monotonic linear index assumption holds for the contrast function on each stage.

12. Model misspecification.

12.1. *Consistency under model misspecification.* In this section, we investigate the statistical properties of our proposed information criteria if the contrast function does not take the monotonic linear index form. We focus on the fixed-p scenario. Results in the high-dimensional setting can be similarly obtained. Due to the model misspecification, VIC (2.2) and CIC (2.3) may select different models. Specifically, define

$$\beta^C = \underset{\beta:\|\beta\|_2=1}{\arg\max}\, C(\beta) \quad \text{and} \quad \theta^V = \{c^V, (\beta^V)^T\}^T = \underset{\substack{\theta=(c,\beta^T)^T\\ \|\theta\|_2=1}}{\arg\max}\, V(\theta).$$

Let $\mathcal{M}_C$ and $\mathcal{M}_V$ denote the support of β^C and β^V, respectively. For each candidate model $\mathcal{M} \subseteq \{1,\ldots,p\}$, let $\hat{\theta}_{\mathcal{M}} = (\hat{c}_{\mathcal{M}}, \hat{\beta}_{\mathcal{M}}^T)^T$ denote some estimator on the restricted model space. We define

$$\widehat{\mathcal{M}}_V = \underset{\mathcal{M}\in\{1,\ldots,p\}}{\arg\max}\ \mathrm{VIC}(\hat{\theta}_{\mathcal{M}}) \quad \text{and} \quad \widehat{\mathcal{M}}_C = \underset{\mathcal{M}\in\{1,\ldots,p\}}{\arg\max}\ \mathrm{CIC}(\hat{\beta}_{\mathcal{M}}).$$

In the following, we will show that $\widehat{\mathcal{M}}_C$ and $\widehat{\mathcal{M}}_V$ are consistent to $\mathcal{M}_C$ and $\mathcal{M}_V$, respectively. Let δ^C and δ^V be some positive constants such that $\delta^C < \min_{j\in\mathcal{M}_C}|\beta^{C,j}|$, $\delta^V < \min_{j\in\mathcal{M}_V}|\beta^{V,j}|$. For any $\varepsilon > 0$, define

$$N_\delta^V = \{\theta \in \mathbb{R}^{p+1} : \|\theta - \theta^V\|_2 \le \delta\}, \quad N_\delta^C = \{\beta \in \mathbb{R}^p : \|\beta - \beta^C\|_2 \le \delta\},$$
$$S^V = \{\theta \in \mathbb{R}^{p+1} : \|\theta\|_2 = \|\theta^V\|_2\}, \quad S^C = \{\beta \in \mathbb{R}^p : \|\beta\|_2 = \|\beta^C\|_2\}.$$

We introduce the following conditions.

(A4*) Assume $\|\hat{\theta}_{\mathcal{M}_V} - \theta^V\| = O_p(R_n^V)$ for some $n^{-1/2} \le R_n^V \to 0$.
(A5*) (i) Assume $V(\theta^V) > V(0)$ and $\liminf_n V(\theta^V) - \sup_{\theta\in(N_{\varepsilon_0}^V)^c\cap S^V} V(\theta) >$

0 for some constant $0 < \varepsilon_0 \le \delta^V$.
(ii) Assume

$$\mathrm{E} \sup_{\substack{\|\theta-\theta_0\|_2 \le \varepsilon \\ \theta=(c,\beta^T)^T}} |\mathbb{I}(X_0^T\beta > -c) - \mathbb{I}(X_0^T\beta^V > -c^V)| = O(\varepsilon),$$

as $\varepsilon \to 0$.
(iii) Assume there exist some constants $\bar{c}_1, \bar{c}_2 > 0$ such that

$$\bar{c}_1\|\theta^V - \theta\|_2^2 \le V(\theta^V) - V(\theta) \le \bar{c}_2\|\theta^V - \theta\|_2^2, \quad \forall \theta \in N_{\varepsilon_0}^V \cap S^V.$$

(A6*) Assume $\|\hat{\beta}_{\mathcal{M}_C} - \beta^C\| = O_p(R_n^C)$ for some $n^{-1/2} \le R_n^C \to 0$.
(A7*)(i) Assume $C(\beta^C) > C(0)$ and $\liminf_n\{C(\beta^C) - \sup_{\beta \in (N_{\varepsilon_0}^C)^c \cap S^C} C(\beta)\} > 0$ for some constant $0 < \varepsilon_0 \le \delta^C$.
(ii) Assume there exist some constants $\bar{c}_1, \bar{c}_2 > 0$ such that

$$\bar{c}_1\|\beta^C - \beta\|_2^2 \le C(\beta^C) - C(\beta) \le \bar{c}_2\|\beta^C - \beta\|_2^2, \quad \forall \beta \in N_{\varepsilon_0}^C \cap S^C.$$

(iii) Function $g(o, \beta)$ defined in (3.1) is twice continuously differentiable for all $\beta \in N_{\varepsilon_0}^C$.
(iv) There is an integrable function $K(o)$ such that for all o and $\beta \in N_{\varepsilon_0}^C$,

$$\|\Delta_2 g(o, \beta) - \Delta_2 g(o, \beta^C)\|_2 \le K(o)\|\beta - \beta^C\|_2.$$

(v) $\mathrm{E}|\partial_i g(O_0, \beta^C)|^2 < \infty$, $\mathrm{E}|\partial_{ij} g(O_0, \beta^C)| < \infty$.

For a given model $\mathcal{M}$, if we obtain $\hat{\theta}_{\mathcal{M}}$ by maximizing $\widehat{V}(\theta)$ on the restricted model space, then Condition (A4*) automatically holds with $R_n^V = n^{-1/3}$. Similarly, Condition (A6*) holds with $R_n^C = n^{-1/2}$ when we obtain $\hat{\beta}_{\mathcal{M}}$ by maximizing $\widehat{C}(\beta)$ on the restricted model space. Condition (A5*) and (A7*) are very similar to (A5) and (A6) introduced in Section 3.1.

THEOREM 12.1. *Suppose* $\sup_x E(Y_0^2|X_0 = x) \le \bar{C}$ *for some constant* $\bar{C} > 0$. *Set* $\kappa_n = c_n \max((nR_n^V)2, \sqrt{nR_n^V}, n^{1/3})$ *for some* $c_n \to \infty$, *if* $\kappa_n = o(n)$, *under Assumptions (A1)-(A3), (A4*) and (A5*), we have*

$$Pr(\widehat{\mathcal{M}}_V = \mathcal{M}^V) \to 1.$$

Set $\kappa_n = n(R_n^{(1)})^2 \log(n)$, *if* $\kappa_n = o(n)$, *under Assumptions (A1)-(A3), (A6*) and (A7*), we have*

$$Pr(\widehat{\mathcal{M}}_C = \mathcal{M}^C) \to 1.$$

It is worth mentioning that when $\mathcal{M}^V \neq \mathcal{M}^C$, CIC and VIC will choose different models. Proof of Theorem 12.1 is very similar to that of Theorem 11.1 and is hence omitted for brevity.

12.2. *Numerical studies.* We generate the data from the following model

$$Y_i = h_0(X_i^1, X_i^3) + A_i \left(X_i^1 + X_i^2 - \frac{(X_i^2)^2}{4} \right) + \varepsilon_i,$$

where $A_i \overset{i.i.d}{\sim} \text{Bernoulli}(0.5)$, $X_i \overset{i.i.d}{\sim} N_p(0, I_p)$, $\varepsilon_i \overset{i.i.d}{\sim} N(0, 0.5^2)$. We fix $p = 8$, and consider two choices of the sample size, corresponding to $n = 100$ and $n = 200$. We also consider two choices of h_0, $h_0(x^1, x^3) = 1 + x^1 - x^3$ and $h_0(x^1, x^3) = 1 + x^1 x^3$, respectively.

Notice that

$$\tau(x) = x^1 + x^2 + \frac{(x^2)^2}{4}.$$

The contrast function does not take the monotonic index form. In Section 12.2.2 and Section 12.2.3, we show that

$$\mathcal{M}^V = \{1, 2\} \text{ and } \mathcal{M}^C = \{1, 2\}.$$

We use CIC^{DR} and VIC^{DR} for model selection. To apply CIC^{DR}, for each candidate model $\mathcal{M} \in \{1, \ldots, 8\}$, we compute the parameter $\hat{\beta}_{\mathcal{M}}$ by maximizing $\widehat{C}^{DR}$ on the restricted model space. Condition (A7*) thus holds with $R_n^C = n^{-1/2}$. To implement VIC^{DR}, we compute $\hat{\theta}_{\mathcal{M}}$ by maximizing $\widehat{V}^{DR}$ on the restricted model space. Condition (A5*) is thus satisfied with $R_n^V = n^{-1/3}$. The baseline and propensity score functions are estimated via penalized linear and logistic regression with SCAD penalty function, respectively. We set $\kappa_n = \log(n)$ in CIC^{DR} and $\kappa_n = n^{-1/3} \log(\log(n))$ in VIC^{DR}.

12.2.1. *Results.* Reported in Table 1 are the percentage of selecting the true models, the false negatives rate, the false positives rate, the average error rate and the average value ratio aggregated over 100 simulations. Similar to the settings where the contrast function takes the monotonic linear index form, CIC^{DR} perform much better than VIC^{DR}. TP's of CIC^{DR} are higher than VIC^{DR} in all cases. CIC^{DR} also achieves a smaller ER and a higher VR compared to VIC^{DR}. In addition, all results get improved with increased sample sizes. This backs up our findings in Theorem 12.1.

Table 1: Simulation results (%, standard deviations in parenthesis)

		$h_0(x^1,x^3)=1+x^1-x^3$		$h_0(x^1,x^3)=1+x^1x^3$	
	n	100	200	100	200
CIC^{DR}	TP	100.00(0.00)	100.00(0.00)	68.00(4.69)	81.00(3.94)
	FN	0.00(0.00)	0.00(0.00)	4.50(1.44)	0.00(0.00)
	FP	0.00(0.00)	0.00(0.00)	6.67(1.18)	3.67(0.81)
	ER	9.70(0.50)	7.33(0.33)	14.98(0.85)	10.91(0.50)
	VR	98.10(0.18)	98.92(0.07)	94.89(0.57)	97.59(0.21)
VIC^{DR}	TP	87.00(3.38)	99.00(1.00)	53.00(5.17)	70.00(4.61)
	FN	6.50(1.69)	0.50(0.50)	17.50(2.50)	6.00(1.63)
	FP	0.00(0.00)	0.00(0.00)	7.67(1.17)	4.50(0.82)
	ER	11.01(0.68)	7.65(0.35)	19.71(1.07)	13.26(0.71)
	VR	96.90(0.40)	98.63(0.13)	91.31(0.82)	95.98(0.40)

12.2.2. *Additional details regarding* $\mathcal{M}^V$. We prove $\mathcal{M}^V=\{1,2\}$. It suffices to show there exist some $\beta^{V,1}$, $\beta^{V,2}>0$ and $c^V\in\mathbb{R}$ such that

$$\begin{aligned}(12.1)\qquad &\mathrm{E}\left(X_0^1+X_0^2-\frac{(X_0^2)^2}{4}\right)\mathbb{I}(\beta^{V,1}X_0^1+\beta^{V,2}X_0^2>c^V)\\ >\ &\mathrm{E}\left(X_0^1+X_0^2-\frac{(X_0^2)^2}{4}\right)\mathbb{I}(X_0^T\beta>c),\end{aligned}$$

for any $\beta\in\mathbb{R}^p$ and $c\in\mathbb{R}$ such that either $\beta^1=0$, or $\beta^2=0$, or there exists some $j\in\{3,4,5,6,7,8\}$, $\beta^j\neq 0$.

We first show there exists some $\beta^{V,1}$, $\beta^{V,2}\geq 0$ and $c^V\in\mathbb{R}$ such that

$$\begin{aligned}(12.2)\qquad &\mathrm{E}\left(X_0^1+X_0^2-\frac{(X_0^2)^2}{4}\right)\mathbb{I}(\beta^{V,1}X_0^1+\beta^{V,2}X_0^2>c^V)\\ >\ &\mathrm{E}\left(X_0^1+X_0^2-\frac{(X_0^2)^2}{4}\right)\mathbb{I}(X_0^T\beta>c),\end{aligned}$$

for any $\beta\in\mathbb{R}^p$ and $c\in\mathbb{R}$ such that there exists some $j\in\{3,4,5,6,7,8\}$, $\beta^j\neq 0$. Since the treatment regime is scale free, it suffices to show (12.2) holds for any $\beta\in\mathbb{R}^p$ and $c\in\mathbb{R}$ such that $\|\beta\|_2=1$ and there exists some $j\in\{3,4,5,6,7,8\}$, $\beta^j\neq 0$.

Let $Z_0=X_0^T\beta$, $Z_1=(X_0^1-\beta^1Z_0)/\sqrt{1-(\beta^1)^2}$ and

$$Z_2=\frac{X_0^2-\beta^2Z_0+\beta^1\beta^2Z_1/\sqrt{1-(\beta^1)^2}}{\sqrt{1-(\beta^2)^2/\{1-(\beta^1)^2\}}}.$$

Since $\|\beta\|_2=1$, Z_0,Z_1 and Z_2 are independent standard normal random

variables. It follows that

$$
\begin{aligned}
&\mathrm{E}\left(X_0^1 + X_0^2 - \frac{(X_0^2)^2}{4}\right)\mathbb{I}(X_0^T\beta > c) \qquad (12.3)\\
&= \mathrm{E}\left\{\sqrt{1-(\beta^1)^2}Z_1 + \beta^1 Z_0 + \sqrt{1-\frac{(\beta^2)^2}{1-(\beta^1)^2}}Z_2 + \beta^2 Z_0 - \frac{\beta^1\beta^2 Z_1}{\sqrt{1-(\beta^1)^2}}\right.\\
&- \left.\frac{1}{4}\left(\sqrt{1-\frac{(\beta^2)^2}{1-(\beta^1)^2}}Z_2 + \beta^2 Z_0 - \frac{\beta^1\beta^2 Z_1}{\sqrt{1-(\beta^1)^2}}\right)^2\right\}\mathbb{I}(Z_0 > c)\\
&= (\beta^1+\beta^2)\mathrm{E}Z_0\mathbb{I}(Z_0 > c) - \frac{(\beta^2)^2}{4}\mathrm{E}Z_0^2\mathbb{I}(Z_0 > c) - \frac{1-(\beta^2)^2}{4}\mathrm{E}\mathbb{I}(Z_0 > c),
\end{aligned}
$$

where the last equality is due to the independence between Z_0, Z_1 and Z_2. With some calculations, we can show

$$
\begin{aligned}
&\mathrm{E}\left(X_0^1 + X_0^2 - \frac{(X_0^2)^2}{4}\right)\mathbb{I}(X_0^T\beta > c)\\
&= \frac{\beta^1+\beta^2}{\sqrt{2\pi}}\exp\left(-\frac{c^2}{2}\right) - \frac{1}{4}\{1-\Phi(c)\} - \frac{(\beta^2)^2 c}{4\sqrt{2\pi}}\exp\left(-\frac{c^2}{2}\right).
\end{aligned}
$$

Under the constraint $\|\beta\|_2 = 1$, it is immediate to see that the maximum is achieved at some β^V that satisfies $\beta^{V,1}, \beta^{V,2} \geq 0$, $(\beta^{V,1})^2 + (\beta^{V,2})^2 = 1$. In other words, we have proven (12.2).

We now show there exist some $\beta^{V,1}, \beta^{V,2} > 0$ and $c^V \in \mathbb{R}$ that satisfy $(\beta^{V,1})^2 + (\beta^{V,2})^2 = 1$ such that

$$
\begin{aligned}
&\mathrm{E}\left(X_0^1 + X_0^2 - \frac{(X_0^2)^2}{4}\right)\mathbb{I}(\beta^{V,1}X_0^1 + \beta^{V,2}X_0^2 > c^V) \qquad (12.4)\\
&> \mathrm{E}\left(X_0^1 + X_0^2 - \frac{(X_0^2)^2}{4}\right)\mathbb{I}(\beta^1 X_0^1 + \beta^2 X_0^2 > c),
\end{aligned}
$$

for any $\beta^1, \beta^2, c \in \mathbb{R}$ that satisfy $(\beta^1)^2 + (\beta^2)^2 = 1$ and either $\beta^1 = 0$, or $\beta^2 = 0$. Under the constraint $(\beta^1)^2 + (\beta^2)^2 = 1$, we have

$$
\begin{aligned}
&\mathrm{E}\left(X_0^1 + X_0^2 - \frac{(X_0^2)^2}{4}\right)\mathbb{I}(\beta^1 X_0^1 + \beta^2 X_0^2 > c)\\
&= \frac{\sqrt{1-(\beta^2)^2}+\beta^2}{\sqrt{2\pi}}\exp\left(-\frac{c^2}{2}\right) - \frac{1}{4}\{1-\Phi(c)\} - \frac{(\beta^2)^2 c}{4}\exp\left(-\frac{c^2}{2}\right).
\end{aligned}
$$

For any c, the derivative of the second line with respect to β^2 evaluated at $\beta^2 = 0$ is positive. Similarly, the derivative of

$$
\frac{\sqrt{1-(\beta^1)^2}+\beta^1}{\sqrt{2\pi}}\exp\left(-\frac{c^2}{2}\right) - \frac{1}{4}\{1-\Phi(c)\} - \frac{c-(\beta^1)^2 c}{4}\exp\left(-\frac{c^2}{2}\right)
$$

with respect to β^1 evaluated at $\beta^1 = 0$ is positive. This proves (12.4).

Notice that

$$\sup_{c\in\mathbb{R}} \mathrm{E}\left(X_0^1 + X_0^2 - \frac{(X_0^2)^2}{4}\right)\mathbb{I}(0 > c) = 0,$$

and

$$\begin{aligned}
&\sup_{\substack{\beta^1,\beta^2>0,c\in\mathbb{R}\\(\beta^1)^2+(\beta^2)^2=1}} \mathrm{E}\left(X_0^1 + X_0^2 - \frac{(X_0^2)^2}{4}\right)\mathbb{I}(\beta^1 X_0^1 + \beta^2 X_0^2 > c)\\
\geq\ & \mathrm{E}\left(X_0^1 + X_0^2 - \frac{(X_0^2)^2}{4}\right)\mathbb{I}(X_0^1 + X_0^2 > 0) = \frac{1}{\sqrt{\pi}} - \frac{1}{8} > 0.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\sup_{\substack{\beta^1,\beta^2>0,c\in\mathbb{R}\\(\beta^1)^2+(\beta^2)^2=1}} \mathrm{E}\left(X_0^1 + X_0^2 - \frac{(X_0^2)^2}{4}\right)\mathbb{I}(\beta^1 X_0^1 + \beta^2 X_0^2 > c)\\
>\ & \sup_{c\in\mathbb{R}} \mathrm{E}\left(X_0^1 + X_0^2 - \frac{(X_0^2)^2}{4}\right)\mathbb{I}(0 > c).
\end{aligned}$$

Combining this together with (12.2) and (12.4) yields (12.1). Hence, we've shown $\mathcal{M}^V = \{1, 2\}$.

12.2.3. *Additional details regarding* $\mathcal{M}^C$. We prove $\mathcal{M}^C = \{1, 2\}$. Similar to (12.3), we can show that for any $\beta \in \mathbb{R}^p$ that satisfies $\|\beta\|_2 = 1$, the concordance function takes the form

$$\begin{aligned}
C(\beta) &= \mathrm{E}\left(X_1^1 + X_1^2 - \frac{(X_1^2)^2}{4} - X_2^1 - X_2^2 + \frac{(X_2^2)^2}{4}\right)\mathbb{I}(X_1^T\beta > X_2^T\beta)\\
&= (\beta^1 + \beta^2)\mathrm{E}(X_1^T\beta - X_2^T\beta)\mathbb{I}(X_1^T\beta > X_2^T\beta)\\
&- \frac{(\beta^2)^2}{4}\mathrm{E}\{(X_1^T\beta)^2 - (X_2^T\beta)^2\}\mathbb{I}(X_1^T\beta > X_2^T\beta) = 2(\beta^1 + \beta^2)\mathrm{E}X_0^T\beta\Phi(X_0^T\beta),
\end{aligned}$$

where the last equality is due to that $X_1^T\beta + X_2^T\beta$ is independent of $X_1^T\beta - X_2^T\beta$. The expectation $\mathrm{E}X_0^T\beta\Phi(X_0^T\beta)$ is independent of β for any $\beta \in \mathbb{R}^p$ that satisfies $\|\beta\|_2 = 1$. Therefore, it is immediate to see that $C(\beta)$ is maximized at

$$\beta^C = (1/\sqrt{2}, 1/\sqrt{2}, 0, 0, 0, 0, 0, 0)^T,$$

with the constraint $\|\beta\|_2 = 1$. In addition, it is easy to show $C(\beta^C) > C(0)$. Therefore, we have $\beta^C = \arg\max_{\beta\in\mathbb{R}^p} C(\beta)$. This shows $\mathcal{M}^C = \{1, 2\}$.

APPENDIX A: SOME ADDITIONAL TECHNICAL CONDITIONS

(A5'.)(i) Assume $V(\theta_0) > V(0)$ and $\liminf_n V(\theta_0) - \sup_{\theta \in \tilde{N}^c_{\varepsilon_0} \cap \tilde{S}(\theta_0)} V(\theta) > 0$ for some constant $0 < \varepsilon_0 \leq \delta$.
(ii) Assume

$$\sup_{\substack{\mathcal{M} \in \Omega \\ |\mathcal{M}| \leq s_n}} \mathrm{E} \sup_{\substack{\theta = (c, \beta^T)^T, \beta^{\mathcal{M}^c} = 0 \\ \|\theta - \theta_0\|_2 \leq \varepsilon}} |\mathbb{I}(X_0^T \beta > -c) - \mathbb{I}(X_0^T \beta_0 > -c_0)| = O(\sqrt{|\mathcal{M}|}\varepsilon),$$

as $\varepsilon \to 0$.
(iii) Assume there exist some constants $\bar{c}_1, \bar{c}_2 > 0$ such that

$$\bar{c}_1 \|\theta_0 - \theta\|_2^2 \leq V(\theta_0) - V(\theta) \leq \bar{c}_2 \|\theta_0 - \theta\|_2^2, \quad \text{for all } \theta \in \tilde{N}_{\varepsilon_0} \cap \tilde{S}(\theta_0), \|\beta\|_0 \leq s_n.$$

(A6'.)(i) Assume $C(\beta_0) > C(0)$ and $\liminf_n \{C(\beta_0) - \sup_{\beta \in N^c_{\varepsilon_0} \cap S(\beta_0)} C(\beta)\} > 0$ for some constant $0 < \varepsilon_0 \leq \delta$.
(ii) Assume there exist some constants $\bar{c}_1, \bar{c}_2 > 0$ such that

$$\bar{c}_1 \|\beta_0 - \beta\|_2^2 \leq C(\beta_0) - C(\beta) \leq \bar{c}_2 \|\beta_0 - \beta\|_2^2, \quad \text{for all } \beta \in N_{\varepsilon_0} \cap S(\beta_0), \|\beta\|_0 \leq s_n.$$

(iii) Function $g(o, \beta)$ is twice continuously differentiable for all $\beta \in N_{\varepsilon_0}$.
(iv) Assume there exists an function $K(o)$ with $\|K(O_0)\|_{\psi_1} = O(1)$ such that for all o and $\beta \in N_{\varepsilon_0}$, $\|\beta\|_0 \leq s_n$,

$$\|\Delta_2 g(o, \beta) - \Delta_2 g(o, \beta_0)\|_2 \leq K(o) \|\beta - \beta_0\|_2,$$

where ψ_p is the Orlicz norm defined in Section 3.2.
(v) $\max_j \|\partial_j g(o, \beta_0)\|_{\psi_1} = O(1)$, $\max_{ij} \|\partial_{ij} g(o, \beta_0)\|_{\psi_1} = O(1)$, $|\nu^T \mathrm{E} \Delta_2 g(O_0, \beta_0) \nu| = O(\|\nu\|_2^2)$ for any $\nu \in \mathbb{R}^p$ that satisfies $\|\nu\|_0 \leq s_n$.

APPENDIX B: DISCUSSION OF TECHNICAL CONDITIONS

B.1. Discussion of (A5)(ii) and (A5')(ii).

B.1.1. *Relation to the margin assumption.* Condition (A5)(ii) and (A5')(ii) are related to the margin assumption (Qian and Murphy, 2011; Luedtke and van der Laan, 2016) which requires $\Pr(0 < |Q(X_0^T \beta_0)| < \varepsilon) = O(\varepsilon^\alpha)$ for some $\alpha > 0$, as $\varepsilon \to 0$. Notice that under (A5)(ii) or (A5')(ii), we have

$$\mathrm{E}\left(\sup_{|c - c_0| \leq \varepsilon} |\mathbb{I}(X_0^T \beta_0 > -c) - \mathbb{I}(X_0^T \beta_0 > -c_0)| \right) = O(\varepsilon), \quad \text{as } \varepsilon \to 0.$$

It follows that

$$\mathrm{E}\left(\sup_{0 < t \leq \varepsilon} \mathbb{I}(-t \leq |X_0^T \beta_0 + c_0| \leq t) \right) = O(\varepsilon), \quad \text{as } \varepsilon \to 0, \tag{B.1}$$

and hence

$$\Pr(0 < |X_0^T\beta_0 + c_0| \le \varepsilon) = O(\varepsilon), \quad \text{as } \varepsilon \to 0.$$

Notice that the function $Q(\cdot)$ satisfies $Q(-c_0) = 0$. Assume

$$\left.\frac{dQ(z-c_0)}{dz}\right|_{z=0} \neq 0.$$

Then for sufficiently small $\varepsilon > 0$, the event $0 < |Q(X_0^T\beta_0)| \le \varepsilon$ is contained in the event $0 < |X_0^T\beta_0 + c_0| \le c_*\varepsilon$ for some constant $c_* > 0$. As a result, we have

$$\Pr(0 < |Q(X_0^T\beta_0)| \le \varepsilon) \le \Pr(0 < |X_0^T\beta_0 + c_0| \le c_*\varepsilon) = O(\varepsilon), \quad \text{as } \varepsilon \to 0.$$

Therefore, the margin assumption holds with $\alpha = 1$.

B.1.2. *Nonregular cases.* In this section, we show (A5)(ii) and (A5')(ii) are violated in the nonregular cases where $\Pr(\tau(X_0) = 0) > 0$. Since c_0 is the unique constant that satisfied $Q(c_0) = 0$, we have $\Pr(X_0^T\beta_0 + c_0 = 0) > 0$. This apparently violates (B.1). The proof is hence completed.

B.2. Discussion of (A6)(iii) and (A6')(iii). In (A6)(iii) and (A6')(iii), we require $g(o,\beta)$ to be twice continuously differentiable for any $\beta \in N_{\varepsilon_0}$. This condition is likely to be violated under treatment effect homogeneity, i.e, $\beta_0 = 0$, $c_0 = 0$, or under the nonregular scenarios where $\Pr\{\tau(X_0) = 0\} > 0$. In this paper, we assume $\beta_0 \neq 0$. Without loss of generality, assume $\beta_0^1 > 0$. Since $\varepsilon_0 < \beta_0^1$, we have $\beta^1 > 0$, $\forall \beta \in N_{\varepsilon_0}$. Notice that $\tau(x) = Q(x^T\beta_0)$. It follows from (A1) and (A2) that

$$\begin{aligned} g(o,\beta) &= \frac{1}{2}\mathrm{E}Q(X_0^T\beta_0)\{\mathbb{I}(X_0^T\beta > x^T\beta) - \mathbb{I}(X_0^T\beta < x^T\beta)\} \\ &- \frac{\{a-\pi_0(x)\}y}{2\pi_0(x)\{1-\pi_0(x)\}}\{\Pr(X_0^T\beta > x^T\beta) - \Pr(X_0^T\beta < x^T\beta)\}. \end{aligned} \tag{B.2}$$

For any $\beta \in \mathbb{R}^p$, let $\beta^{(-1)}$ denote the last $p-1$ components of β. Let $F_{X_0^{(-1)}}(\cdot)$ denote the cumulative distribution function of the last $p-1$ components of X_0. Let $f_{X_0^{(1)}}(\cdot|z^{(-1)})$ denote the conditional density function of $X_0^{(1)}$ given $X_0^{(-1)} = z^{(-1)}$. With some calculations, we can show

$$\begin{aligned} g(o,\beta) = \int_{\mathbb{R}^{p-1}} \int_{\frac{\beta^T x - \beta^{(-1)T} z^{(-1)}}{\beta^1}}^{+\infty} \left(Q(z^1\beta_0^1 + z^{(-1)}\beta_0^{(-1)}) - \frac{\{a-\pi_0(x)\}y}{2\pi_0(x)\{1-\pi_0(x)\}} \right) \\ \times f_{X_0^1}(z^1|z^{(-1)}) dz^1 dF_{X_0^{(-1)}}(z^{(-1)}) - \frac{1}{2}\left(\mathrm{E}Q(X_0^T\beta_0) - \frac{\{a-\pi_0(x)\}y}{\pi_0(x)\{1-\pi_0(x)\}} \right). \end{aligned}$$

Therefore, (A6)(iii) and (A6')(iii) hold under certain regularity conditions on $f_{X_0^1}$, $F_{X_0^{(-1)}}$ and their derivatives.

In Section B.2.1, we provide two examples assuming the covariates are jointly normal and show (A6)(iii) and (A6')(iii) hold under both examples. In Section B.2.2, we show both conditions are violated under the treatment effect homogeneity. In Section B.2.3, we show both conditions are violated in the nonregular cases.

B.2.1. *Gaussian covariates.* We assume $X_0 \sim N(0, \Sigma)$ for some positive definite covariance matrix Σ.

Example 1 (Linear contrast function) Assume $Q(z) = a_0 z + b_0$ for some $a_0, b_0 \in \mathbb{R}$. It follows from (B.2) that

$$\begin{aligned} g(o,\beta) &= \frac{1}{2}\mathrm{E}(a_0 X_0^T\beta_0 + b_0)\{\mathbb{I}(X_0^T\beta > x^T\beta) - \mathbb{I}(X_0^T\beta < x^T\beta)\} \\ &- \frac{\{a - \pi_0(x)\}y}{2\pi_0(x)\{1-\pi_0(x)\}}\{\Pr(X_0^T\beta > x^T\beta) - \Pr(X_0^T\beta < x^T\beta)\}. \end{aligned}$$

To show (A6)(iii) and (A6')(iii) hold, it suffices to show functions

$$\begin{aligned} g_1(x,\beta) &= \mathrm{E}X_0^T\beta_0\{\mathbb{I}(X_0^T\beta > x^T\beta) - \mathbb{I}(X_0^T\beta < x^T\beta)\}, \\ g_2(x,\beta) &= \Pr(X_0^T\beta > x^T\beta) - \Pr(X_0^T\beta < x^T\beta), \end{aligned}$$

are twice continuously differentiable for all $\beta \in N_{\varepsilon_0}$. Since $\beta_0 \neq 0$ and $\varepsilon_0 < \min_{j\in\mathcal{M}_\beta} |\beta_0^j|$, we have $\beta \neq 0$, for any $\beta \in N_{\varepsilon_0}$. The random variable $X_0^T\beta$ is normally distributed with variance $\beta^T\Sigma\beta > 0$. As a result,

$$g_2(x,\beta) = 1 - 2\Phi\left(\frac{x^T\beta}{\sqrt{\beta^T\Sigma\beta}}\right).$$

It is immediate to see that g_2 is twice continuously differentiable with respect to $\beta \in N_{\varepsilon_0}$.

As for g_1, we have

$$\begin{aligned} g_1(x,\beta) &= \mathrm{E}\left(X_0^T\beta_0 - \frac{\beta_0^T\Sigma\beta}{\beta^T\Sigma\beta}X_0^T\beta + \frac{\beta_0^T\Sigma\beta}{\beta^T\Sigma\beta}X_0^T\beta\right)\{\mathbb{I}(X_0^T\beta > x^T\beta) - \mathbb{I}(X_0^T\beta < x^T\beta)\} \\ &= \frac{\beta_0^T\Sigma\beta}{\beta^T\Sigma\beta}\mathrm{E}X_0^T\beta\{\mathbb{I}(X_0^T\beta > x^T\beta) - \mathbb{I}(X_0^T\beta < x^T\beta)\} = 2\frac{\beta_0^T\Sigma\beta}{\beta^T\Sigma\beta}\mathrm{E}X_0^T\beta\mathbb{I}(X_0^T\beta > x^T\beta) \\ &= \frac{\sqrt{2}\beta_0^T\Sigma\beta}{\sqrt{\pi\beta^T\Sigma\beta}}\int_{\frac{x^T\beta}{\sqrt{\beta^T\Sigma\beta}}}^{+\infty} z\exp(-z^2/2)dz = \frac{\sqrt{2}\beta_0^T\Sigma\beta}{\sqrt{\pi\beta^T\Sigma\beta}}\exp\left(-\frac{(x^T\beta)^2}{2\beta^T\Sigma\beta}\right), \end{aligned}$$

where the second equality is due to that $X_0^T\beta_0 - (\beta_0^T\Sigma\beta)(X_0^T\beta)/(\beta^T\Sigma\beta)$ is independent of $X_0^T\beta$. Hence, g_1 is twice continuously differentiable with respect to $\beta \in N_{\varepsilon_0}$.

Example 2 (Nonlinear contrast function) Assume $Q(z) = \exp(a_0 z + b_0) - 1$ for some $a_0, b_0 \in \mathbb{R}$. It follows from (B.2) that

$$\begin{aligned} g(o,\beta) &= \frac{1}{2}\mathrm{E}\{\exp(a_0 X_0^T\beta_0 + b_0) - 1\}\{\mathbb{I}(X_0^T\beta > x^T\beta) - \mathbb{I}(X_0^T\beta < x^T\beta)\} \\ &- \frac{\{a - \pi_0(x)\}y}{2\pi_0(x)\{1-\pi_0(x)\}}\{\Pr(X_0^T\beta > x^T\beta) - \Pr(X_0^T\beta < x^T\beta)\}. \end{aligned}$$

To show (A6)(iii) and (A6')(iii) hold, it suffices to show functions

$$\begin{aligned} g_2(x,\beta) &= \Pr(X_0^T\beta > x^T\beta) - \Pr(X_0^T\beta < x^T\beta), \\ g_3(x,\beta) &= \mathrm{E}\exp(a_0 X_0^T\beta_0)\{\mathbb{I}(X_0^T\beta > x^T\beta) - \mathbb{I}(X_0^T\beta < x^T\beta)\}, \end{aligned}$$

are twice continuously differentiable for all $\beta \in N_{\varepsilon_0}$. We've shown in Example 1 that g_2 is twice continuously differentiable with respect to $\beta \in N_{\varepsilon_0}$. As for g_3, since $X_0^T\beta_0 - (\beta_0^T\Sigma\beta)(X_0^T\beta)/(\beta^T\Sigma\beta)$ is independent of $X_0^T\beta$, we have

$$\begin{aligned} g_3(x,\beta) = \mathrm{E}\exp\left(a_0\frac{\beta_0^T\Sigma\beta}{\beta^T\Sigma\beta}X_0^T\beta\right)\{\mathbb{I}(X_0^T\beta > x^T\beta) - \mathbb{I}(X_0^T\beta < x^T\beta)\} \\ \times \mathrm{E}\exp\left(a_0 X_0^T\beta_0 - a_0\frac{\beta_0^T\Sigma\beta}{\beta^T\Sigma\beta}X_0^T\beta\right) \end{aligned}$$

With some calculations, we can show

$$g_3(x,\beta) = \exp\left(\frac{a_0^2\beta_0^T\Sigma\beta_0}{2}\right)\left\{1 - 2\Phi\left(\frac{x^T\beta - a_0\beta_0^T\Sigma\beta}{\sqrt{\beta^T\Sigma\beta}}\right)\right\}.$$

It is immediate to see that g_3 is twice continuously differentiable with respect to $\beta \in N_{\varepsilon_0}$.

B.2.2. *Treatment effect homogeneity.* Under the treatment effect homogeneity, we have $\beta_0 = 0$, $c_0 = 0$, and hence $Q(x^T\beta_0) = 0, \forall x$. It follows from (B.2) that

$$g(o,\beta) = \frac{\{a - \pi_0(x)\}y}{2\pi_0(x)(1-\pi_0(x))}\{\Pr(X_0^T\beta < x^T\beta) - \Pr(X_0^T\beta > x^T\beta)\}.$$

Let $\beta^{(i)}(\gamma) = (\underbrace{0,\ldots,0}_{i-1},\gamma,\underbrace{0,\ldots,0}_{p-i})^T$, we have

$$\lim_{\gamma\to 0^+} g(o,\beta^{(i)}(\gamma)) = \frac{\{a - \pi_0(x)\}y}{2\pi_0(x)(1-\pi_0(x))}\{\Pr(X_0^i < x^i) - \Pr(X_0^i > x^i)\},$$

and

$$\lim_{\gamma\to 0^-} g(o,\beta^{(i)}(\gamma)) = \frac{\{a-\pi_0(x)\}y}{2\pi_0(x)(1-\pi_0(x))}\{\Pr(X_0^i > x^i) - \Pr(X_0^i < x^i)\}.$$

As a result, $g(o,\beta)$ is even not continuous when $\Pr(X_0^i < x^i) \neq \Pr(X_0^i > x^i)$ for some $i \in \{1,\ldots,p\}$. Hence, $g(o,\beta)$ is not twice differentiable. Condition (A6)(iii) and (A6')(iii) are thus violated.

B.2.3. *Nonregular cases.* We assume the p-dimensional covariates X_0 consist of p independent left-censored Gaussian random variables. Specifically, for any $j \in \{1,\ldots,p\}$,

$$\Pr(X_0^j = 0) = \frac{1}{2} \quad \text{and} \quad \Pr(X_0^j \le z) = \frac{1}{2} + \frac{1}{\sqrt{2\pi}}\int_0^z \exp\left(-\frac{z^2}{2}\right) dz, \quad \forall z > 0.$$

In addition, the contrast function takes the form $\tau(X_0) = X_0^1 - X_0^2$. It is immediate to see that

$$\Pr\{\tau(X_0) = 0\} = \Pr(X_0^1 = X_0^2) = \frac{1}{4}.$$

Let $\beta_0^{(3)}(\gamma) = \beta_0 + (0,0,\gamma,0,0,\ldots,0)^T$ and $x^{(1)} = x^{(2)} = x^{(3)} = 0$. It follows that $x^T\beta_0^{(3)}(\gamma) = 0$. Notice that

$$\Pr(X_0^T\beta_0^{(3)}(\gamma) = 0) \ge \Pr\left(\bigcap_{j=1}^{3}\{X_0^j = 0\}\right) = \frac{1}{8}.$$

Below, we show

$$\begin{aligned}\lim_{\gamma\to 0} &\mathrm{E}(X_0^1 - X_0^2)\{\mathbb{I}(X_0^1 - X_0^2 + \gamma X_0^3 > 0) - \mathbb{I}(X_0^1 - X_0^2 + \gamma X_0^3 < 0)\}\\ &= \mathrm{E}(X_0^1 - X_0^2)\{\mathbb{I}(X_0^1 - X_0^2 > 0) - \mathbb{I}(X_0^1 - X_0^2 < 0)\},\end{aligned} \tag{B.3}$$

and

$$\begin{aligned}&\lim_{\gamma\to 0^+}\left(\Pr(X_0^1 - X_0^2 + \gamma X_0^3 > 0) - \Pr(X_0^1 - X_0^2 + \gamma X_0^3 < 0)\right)\\ \neq\ &\lim_{\gamma\to 0^-}\left(\Pr(X_0^1 - X_0^2 + \gamma X_0^3 > 0) - \Pr(X_0^1 - X_0^2 + \gamma X_0^3 < 0)\right).\end{aligned} \tag{B.4}$$

Condition (A6)(iii) and (A6')(iii) are thus violated.

With some calculations, we have for sufficiently small $\gamma > 0$,

$$\begin{aligned}
&\mathrm{E}(X_0^1 - X_0^2)\{\mathbb{I}(X_0^1 - X_0^2 + \gamma X_0^3 > 0) - \mathbb{I}(X_0^1 - X_0^2 + \gamma X_0^3 < 0)\} \\
= \;& \frac{1}{2\sqrt{2\pi}} \int_0^{+\infty} x \exp\left(-\frac{x^2}{2}\right) dx - \frac{1}{2\sqrt{2\pi}} \int_0^{+\infty} x \exp\left(-\frac{x^2}{2}\right) \left\{1 - 2\Phi\left(\frac{x}{\gamma}\right)\right\} dx \\
- \;& \frac{1}{2\pi} \int_0^{+\infty} \int_0^{+\infty} (x-y) \exp\left(-\frac{x^2+y^2}{2}\right) \left\{1 - 2\Phi\left(\frac{y-x}{\gamma}\right) \mathbb{I}(y > x)\right\} dxdy, \\
&\mathrm{E}(X_0^1 - X_0^2)\{\mathbb{I}(X_0^1 - X_0^2 - \gamma X_0^3 > 0) - \mathbb{I}(X_0^1 - X_0^2 - \gamma X_0^3 < 0)\} \\
= \;& \frac{1}{2\sqrt{2\pi}} \int_0^{+\infty} x \exp\left(-\frac{x^2}{2}\right) \left\{2\Phi\left(\frac{x}{\gamma}\right) - 1\right\} dx + \frac{1}{2\sqrt{2\pi}} \int_0^{+\infty} x \exp\left(-\frac{x^2}{2}\right) dx \\
- \;& \frac{1}{2\pi} \int_0^{+\infty} \int_0^{+\infty} (x-y) \exp\left(-\frac{x^2+y^2}{2}\right) \left\{2\Phi\left(\frac{x-y}{\gamma}\right) \mathbb{I}(x > y) - 1\right\} dxdy,
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{E}(X_0^1 - X_0^2)\{\mathbb{I}(X_0^1 - X_0^2 > 0) - \mathbb{I}(X_0^1 - X_0^2 < 0)\} = \;& \frac{1}{\sqrt{2\pi}} \int_0^{+\infty} \int_0^{+\infty} x \exp\left(-\frac{x^2}{2}\right) dx \\
& - \frac{1}{2\pi} \int_0^{+\infty} (x-y) \exp\left(-\frac{x^2+y^2}{2}\right) \{\mathbb{I}(x > y) - \mathbb{I}(x < y)\} dxdy.
\end{aligned}$$

It follows from the dominated convergence theorem that

$$\begin{aligned}
&\lim_{\gamma \to 0} \mathrm{E}(X_0^1 - X_0^2)\{\mathbb{I}(X_0^1 - X_0^2 + \gamma X_0^3 > 0) - \mathbb{I}(X_0^1 - X_0^2 + \gamma X_0^3 < 0)\} \\
&\qquad = \mathrm{E}(X_0^1 - X_0^2)\{\mathbb{I}(X_0^1 - X_0^2 > 0) - \mathbb{I}(X_0^1 - X_0^2 < 0)\}.
\end{aligned}$$

This proves (B.3).

Similarly, we can show

$$\begin{aligned}
&\lim_{\gamma \to 0^+} \left(\Pr(X_0^1 - X_0^2 + \gamma X_0^3 > 0) - \Pr(X_0^1 - X_0^2 + \gamma X_0^3 < 0)\right) \\
= \;& \Pr(X_0^1 = X_0^2 = 0, X_0^3 > 0) + \Pr(X_0^1 > X_0^2) - \Pr(X_0^1 < X_0^2) = \frac{1}{8},
\end{aligned}$$

and

$$\begin{aligned}
&\lim_{\gamma \to 0^-} \left(\Pr(X_0^1 - X_0^2 + \gamma X_0^3 > 0) - \Pr(X_0^1 - X_0^2 + \gamma X_0^3 < 0)\right) \\
= \;& \Pr(X_0^1 > X_0^2) - \Pr(X_0^1 < X_0^2) - \Pr(X_0^1 = X_0^2 = 0, X_0^3 > 0) = -\frac{1}{8}.
\end{aligned}$$

This proves (B.4).

APPENDIX C: PROOF OF THEOREM 3.4

Similar to the proof of Theorem 3.3, in the following, we provide tail inequalities for

$$\Pr\left(\mathrm{CIC}(\hat{\beta}_{\mathcal{M}(\lambda_0)}) \le \sup_{\lambda\in\Omega_-} \mathrm{CIC}(\hat{\beta}_{\mathcal{M}(\lambda)})\right), \tag{C.1}$$

and

$$\Pr\left(\mathrm{CIC}(\hat{\beta}_{\mathcal{M}(\lambda_0)}) \le \sup_{\lambda\in\Omega_+} \{n\widehat{C}(\tilde{\beta}_{\mathcal{M}(\lambda)}) - \kappa_n\|\hat{\beta}_{\mathcal{M}(\lambda)}\|_0\}\right). \tag{C.2}$$

C.1. Underfitted model space. Under Assumption (A6')(i) and (ii), using similar arguments in Section 9.1, we can show

$$C(\hat{\beta}_{\mathcal{M}(\lambda_0)}) - \sup_{\lambda\in\Omega_-} C(\hat{\beta}_{\mathcal{M}(\lambda)}) > 5\xi, \tag{C.3}$$

for some constant $\xi > 0$. This together with

$$\kappa_n(\|\hat{\beta}_{\mathcal{M}(\lambda_0)}\|_0 - \inf_{\lambda\in\Omega_-} \|\hat{\beta}_{\mathcal{M}(\lambda)}\|_0) \le O(1)\kappa_n,$$

and the condition $\kappa_n = o(n)$ suggests that for sufficiently large n, we have

$$C(\hat{\beta}_{\mathcal{M}(\lambda_0)}) - \kappa_n\|\hat{\beta}_{\mathcal{M}(\lambda_0)}\|_0 - \sup_{\lambda\in\Omega_-}\left\{C(\hat{\beta}_{\mathcal{M}(\lambda)}) - \kappa_n\|\hat{\beta}_{\mathcal{M}(\lambda)}\|_0\right\} \ge 4\xi.$$

Therefore, the event defined in (C.1) happens if

$$\sup_{\lambda\in\Omega_-}\left|\left\{\widehat{C}(\hat{\beta}_{\mathcal{M}(\lambda)}) - C(\hat{\beta}_{\mathcal{M}(\lambda)}) - \widehat{C}(\hat{\beta}_{\mathcal{M}(\lambda_0)}) + C(\hat{\beta}_{\mathcal{M}(\lambda_0)})\right\}\right| \ge 4\xi,$$

or

$$\sup_{\|\beta\|_0\le s_n}\left|\widehat{C}(\beta) - C(\beta)\right| \ge 2\xi.$$

This means (C.1) is smaller than

$$\Pr\left(\sup_{\|\beta\|_0\le s_n}\left|\widehat{C}(\beta) - C(\beta)\right| \ge 2\xi\right). \tag{C.4}$$

Define

$$\begin{aligned} h(O_i, O_j, \beta) \quad &= \quad \frac{1}{2}\left(\frac{(A_i - \pi_{0,i})Y_i}{\pi_{0,i}(1-\pi_{0,i})} - \frac{(A_j - \pi_{0,j})Y_j}{\pi_{0,j}(1-\pi_{0,j})}\right)\mathbb{I}(X_i^T\beta > X_j^T\beta) \\ &+ \quad \frac{1}{2}\left(\frac{(A_j - \pi_{0,j})Y_j}{\pi_{0,j}(1-\pi_{0,j})} - \frac{(A_i - \pi_{0,i})Y_i}{\pi_{0,i}(1-\pi_{0,i})}\right)\mathbb{I}(X_j^T\beta > X_i^T\beta), \end{aligned} \tag{C.5}$$

and

$$D(O_i, O_j, \beta) = h(O_i, O_j, \beta) - g(O_i, \beta) - g(O_j, \beta) + C(\beta). \tag{C.6}$$

Function D is symmetric. Besides, it satisfies

$$\mathrm{E}D(O_i, o, \beta) = \mathrm{E}D(o, O_j, \beta) = 0.$$

The process $\sum_{i \neq j} D(O_i, O_j, \beta)$ is a degenerate U-process. Define the event

$$\bar{\mathcal{A}} = \left\{ \left| \frac{1}{n(n-1)} \sum_{i \neq j} D(O_i, O_j, \beta) \right| \leq \xi \right\}.$$

Since

$$m_C(\beta) = \widehat{C}(\beta) - C(\beta) = \frac{1}{n(n-1)} \sum_{i \neq j} h(O_i, O_j, \beta) - C(\beta)$$

$$= \left\{ \frac{1}{n(n-1)} \sum_{i \neq j} D(O_i, O_j, \beta) \right\} + \left\{ \frac{2}{n} \sum_i g(O_i, \beta) - 2C(\beta) \right\},$$

under the event defined in $\bar{\mathcal{A}}$, (C.4) can be bounded by

$$\Pr\left(\sup_{\|\beta\|_0 \leq s_n} \left| \frac{1}{n} \sum_i \{ g(O_i, \beta) - C(\beta) \} \right| \geq \frac{\xi}{2} \right),$$

or

$$\sum_{\mathcal{M} \in \Omega^*} \Pr\left(\sup_{\beta \in B_{\mathcal{M}}} \left| \frac{1}{n} \sum_i \{ g(O_i, \beta) - C(\beta) \} \right| \geq \frac{\xi}{2} \right). \tag{C.7}$$

Let $O_{n+1}, \ldots, O_{2n}$ as i.i.d copies of O_0, independent of $\{O_1, \ldots, O_n\}$. Observe that $g(O_i, \beta) = \mathrm{E}\{h(O_i, O_{n+i}, \beta) | O_i\}$. Here the expectation is taken with respect to O_{n+i}. It follows from Jensen's inequality that

$$\begin{aligned} &\sup_{\mathcal{M} \in \Omega^*} \mathrm{E}\left(\sup_{\beta \in B_{\mathcal{M}}} \left| \frac{1}{n} \sum_i \{ g(O_i, \beta) - C(\beta) \} \right| \right) \\ \leq\ &\sup_{\mathcal{M} \in \Omega^*} \mathrm{E}\left(\sup_{\beta \in B_{\mathcal{M}}} \left| \frac{1}{n} \sum_i \{ h(O_i, O_{n+i}, \beta) - C(\beta) \} \right| \right). \end{aligned} \tag{C.8}$$

Similar to (9.15), RHS of (C.8) is of the order $O(\sqrt{s_n/n})$. This together with (C.8) and the assumption $s_n = o(n)$ yields

$$\sup_{\mathcal{M}\in\Omega^*} \mathrm{E}\left(\sup_{\beta\in B_{\mathcal{M}}}\left|\frac{1}{n}\sum_i\{g(O_i,\beta)-C(\beta)\}\right|\right) \le \frac{\xi}{6}, \tag{C.9}$$

for sufficiently large n. Under the condition $\|Y_0\|_{\psi_1} = O(1)$, we can show the ψ_1 Orlicz norm of the envelope function of the class $\{g(o,\beta)\}_\beta$ is $O(1)$. Hence, it follows from Lemma H.4 that

$$\begin{aligned}
&\sup_{\mathcal{M}\in\Omega^*} \Pr\left(\sup_{\beta\in B_{\mathcal{M}}}\left|\frac{1}{n}\sum_i\{g(O_i,\beta)-C(\beta)\}\right|\right.\\
\ge\ & \left.\frac{3}{2}\mathrm{E}\sup_{\beta\in B_{\mathcal{M}}}\left|\frac{1}{n}\sum_i\{g(O_i,\beta)-C(\beta)\}\right| + \frac{\xi}{4}\right) \le \exp\left(-\frac{\bar{c}n}{\log(n)}\right),
\end{aligned}$$

for some constant $\bar{c}$. This together with (C.9) suggests

$$\sup_{\mathcal{M}\in\Omega^*} \Pr\left(\sup_{\beta\in B_{\mathcal{M}}}\left|\frac{1}{n}\sum_i\{g(O_i,\beta)-C(\beta)\}\right| \ge \frac{\xi}{2}\right) \le \exp\left(-\frac{\bar{c}n}{\log(n)}\right). \tag{C.10}$$

Since $n/\log(n) \gg s_n\log(p)$, for sufficiently large n, (C.7) is bounded by

$$O(p^{s_n})\exp\left(-\frac{\bar{c}n}{\log(n)}\right) \le \exp\left(-\frac{\bar{c}n}{2\log(n)}\right).$$

Therefore, we've shown (C.4) can be bounded by

$$\Pr(\bar{\mathcal{A}}^c) + \exp\left(-\frac{\bar{c}n}{2\log(n)}\right).$$

Besides, similar to (C.19)-(C.21) (appear a few pages later), we have

$$\Pr(\bar{\mathcal{A}}^c) \le \exp\left(-\frac{\bar{k}n}{2\log(n)}\right), \tag{C.11}$$

for some constant $\bar{k}$. This gives the probability bound that CIC chooses an underfitted model.

C.2. Overfitted model space. Let $R^C_{\mathcal{M}} = t_0 n^{-1/2}|\mathcal{M}|^{1/2}\log^{1/2}(p)$, and $B^C_{\mathcal{M}} = \{\beta : \beta \in B_{\mathcal{M}}, \|\beta - \beta_0\|_2 \le R^C_{\mathcal{M}}\}$. Since $|\mathcal{M}| \le s_n$ and $n \gg s_n \log(p)\log(n)$, we obtain that $\sup_{\mathcal{M}\in\Omega^*_+} R^C_{\mathcal{M}} \to 0$. Similar to (9.22), conditional on the event

$$\bigcap_{\lambda\in\Omega_+} \left\{ \|\tilde{\beta}_{\mathcal{M}(\lambda)} - \beta_0\|_2 \le R^C_{\mathcal{M}} \right\}, \tag{C.12}$$

we can show that (C.2) is bounded by

$$\begin{aligned} &\Pr\left(\sup_{\mathcal{M}\in\Omega^*_+} \sup_{\beta\in B^C_{\mathcal{M}}} \frac{n}{|\mathcal{M}|} |\widehat{m}_C(\beta) - m_C(\beta)| \ge 4\bar{c}\kappa_n \right) \\ + \quad &\Pr\left(n|\widehat{m}_C(\hat{\beta}_{\mathcal{M}(\lambda_0)}) - m_C(\hat{\beta}_{\mathcal{M}(\lambda_0)})| \ge 4\bar{c}\kappa_n \right), \end{aligned} \tag{C.13}$$

for some constant $\bar{c} > 0$, where $\widehat{m}_C(\beta) = \widehat{C}(\beta) - \widehat{C}(\beta_0)$, $m_C(\beta) = C(\beta) - C(\beta_0)$. Notice that

$$\begin{aligned} \widehat{C}(\beta) - C(\beta) &= \frac{1}{n(n-1)} \sum_{i\ne j} h(O_i, O_j, \beta) - C(\beta) \\ &= \left\{ \frac{1}{n(n-1)} \sum_{i\ne j} D(O_i, O_j, \beta) \right\} + \left\{ \frac{2}{n} \sum_i g(O_i, \beta) - 2C(\beta) \right\}. \end{aligned}$$

The first term of (C.13) can be bounded by

$$\begin{aligned} &\Pr\left(\sup_{\mathcal{M}\in\Omega^*_+} \sup_{\beta\in B_{\mathcal{M}}} \frac{n}{|\mathcal{M}|} \left| \frac{1}{n(n-1)} \sum_{i\ne j} D(O_i, O_j, \beta) \right| \ge \bar{c}\kappa_n \right) \\ + \quad &\Pr\left(\sup_{\mathcal{M}\in\Omega^*_+} \sup_{\beta\in B^C_{\mathcal{M}}} \frac{n}{|\mathcal{M}|} \left| \frac{1}{n} \sum_i \{g(O_i, \beta) - g(O_i, \beta_0) - C(\beta) + C(\beta_0)\} \right| \ge \bar{c}\kappa_n \right) \\ \le \quad &\sum_{\mathcal{M}\in\Omega^*_+} \Pr\left(\sup_{\beta\in B_{\mathcal{M}}} \frac{n}{|\mathcal{M}|} \left| \frac{1}{n(n-1)} \sum_{i\ne j} D(O_i, O_j, \beta) \right| \ge \bar{c}\kappa_n \right) \\ + \quad &\Pr\left(\sup_{\mathcal{M}\in\Omega^*_+} \sup_{\beta\in B^C_{\mathcal{M}}} \frac{n}{|\mathcal{M}|} \left| \frac{1}{n} \sum_i \{g(O_i, \beta) - g(O_i, \beta_0) - C(\beta) + C(\beta_0)\} \right| \ge \bar{c}\kappa_n \right). \end{aligned}$$

We begin by providing an upper bound for

$$\sum_{\mathcal{M}\in\Omega^*_+} \Pr\left(\sup_{\beta\in B_{\mathcal{M}}} \frac{n}{|\mathcal{M}|} \left| \frac{1}{n(n-1)} \sum_{i\ne j} D(O_i, O_j, \beta) \right| \ge \bar{c}\kappa_n \right). \tag{C.14}$$

Under Assumption (A3), we have

$$\begin{aligned}\sup_\beta |h(O_i,O_j,\beta)| &= \frac{1}{2}\sup_\beta \left|\left(\frac{(A_i-\pi_{0,i})Y_i}{\pi_{0,i}(1-\pi_{0,i})}-\frac{(A_j-\pi_{0,j})Y_j}{\pi_{0,j}(1-\pi_{0,j})}\right)\mathbb{I}(X_i^T\beta > X_j^T\beta)\right| \\ &+ \frac{1}{2}\sup_\beta \left|\left(\frac{(A_j-\pi_{0,j})Y_j}{\pi_{0,j}(1-\pi_{0,j})}-\frac{(A_i-\pi_{0,i})Y_i}{\pi_{0,i}(1-\pi_{0,i})}\right)\mathbb{I}(X_j^T\beta > X_i^T\beta)\right| \\ &\le \frac{|Y_i|+|Y_j|}{c_1(1-c_2)} \equiv H(O_i,O_j).\end{aligned}$$

Therefore, H is the envelope function of h. Under the assumption $\|Y_i\|_{\psi_1} = O(1)$, this also implies

$$\|\sup_\beta |h(O_i,O_j,\beta)|\|_{\psi_1} \le \|H(O_i,O_j)\|_{\psi_1} = O(1).$$

Hence, it follows from Lemma 2.2.2 in van der Vaart and Wellner (1996) that

$$\|\max_{i,j}\sup_\beta |h(O_i,O_j,\beta)|\|_{\psi_1} = O(\log(n)). \tag{C.15}$$

By Jensen's inequality, we have

$$\begin{aligned}&\|\sup_\beta |D(O_i,O_j,\beta)|\|_{\psi_1} \le \|\sup_\beta |h(O_i,O_j,\beta)|\|_{\psi_1} + \|\sup_\beta |g(O_i,\beta)|\|_{\psi_1} \\ +\ &\|\sup_\beta |g(O_j,\beta)|\|_{\psi_1} + \|\sup_\beta |C(\beta)|\|_{\psi_1} \le 4\|\sup_\beta |h(O_i,O_j,\beta)|\|_{\psi_1}.\end{aligned}$$

This together with (C.15) implies that

$$\omega_n \equiv \left\|\max_{i\neq j}\sup_\beta |D(O_i,O_j,\beta)|\right\|_{\psi_1} = O(\log(n)). \tag{C.16}$$

We can also show $4H$ is the envelope function of D. Define $\epsilon_1,\ldots,\epsilon_n$ to be i.i.d Radamacher random variables independent of $\{O_1,\ldots,O_n\}$. It follows from Jensen's inequality and the degeneracy of D that

$$\begin{aligned}&\mathrm{E}\sup_\beta |\sum_{i\neq j} D(O_i,O_j,\beta)\mathbb{I}(4H(O_i,O_j)\le 8\omega_n)| \\ \le\ &\mathrm{E}\sup_\beta |\sum_{i\neq j} \epsilon_i\epsilon_j D(O_i,O_j,\beta)\mathbb{I}(4H(O_i,O_j)\le 8\omega_n)| \\ \le\ &4\mathrm{E}\sup_\beta |\sum_{i\neq j} \epsilon_i\epsilon_j h(O_i,O_j,\beta)\mathbb{I}(4H(O_i,O_j)\le 8\omega_n)|.\end{aligned}$$

The class of functions $\{h(O_i, O_j, \beta) : \beta \in B_{\mathcal{M}}\}$ has VC index $|\mathcal{M}| + 2$, so is the class of functions $\{h(O_i, O_j, \beta)\mathbb{I}(H(O_i, O_j) \le 8\omega_n) : \beta \in B_{\mathcal{M}}\}$. Using similar arguments in the proofs of Lemma 5 and Theorem 6 in Nolan and Pollard (1987), we can show

$$\text{(C.17)} \qquad \mathrm{E}Z_\varepsilon \equiv \mathrm{E}\sup_{\beta} \Big| \sum_{i \neq j} D(O_i, O_j, \beta)\mathbb{I}(4H(O_i, O_j) \le 8\omega_n) \Big| = O(|\mathcal{M}|n).$$

Let $t = \bar{c}(n-1)|\mathcal{M}|\kappa_n/2$. Define

$$U_\varepsilon = \sup_{\beta \in B_{\mathcal{M}}} \sup_{\alpha : \|\alpha\|_2 \le 1} \sum_{i,j} \epsilon_i \alpha_j D(O_i, O_j, \beta)\mathbb{I}(4H(O_i, O_j) \le 8\omega_n),$$

$$M_\varepsilon = \sup_{\beta \in B_{\mathcal{M}}} \sup_{k=1,\ldots,n} \left| \sum_{i} \epsilon_i D(O_i, O_k, \beta)\mathbb{I}(4H(O_i, O_k) \le 8\omega_n) \right|.$$

Since $\kappa_n \to \infty$, by (C.16) and (C.17), for sufficiently large n, Theorem 7.1 applies and we have

$$\begin{aligned} \text{(C.18)} \qquad & \Pr\left(\sup_{\beta \in B_{\mathcal{M}}} \frac{n}{|\mathcal{M}|} \left| \frac{1}{n(n-1)} \sum_{i \neq j} D(O_i, O_j, \beta) \right| \ge \bar{c}\kappa_n \right) \\ = \; & \Pr\left(\sup_{\beta \in B_{\mathcal{M}}} \left| \sum_{i \neq j} D(O_i, O_j, \beta) \right| \ge \bar{c}(n-1)|\mathcal{M}|\kappa_n \right) \\ \le \; & \Pr\left(\sup_{\beta \in B_{\mathcal{M}}} \left| \sum_{i \neq j} D(O_i, O_j, \beta) \right| \ge \bar{k}\mathrm{E}Z_\varepsilon + t \right) \\ \le \; & 3\exp\left(-\frac{1}{\bar{k}} \min\left(\frac{t^2}{(\mathrm{E}U_\varepsilon)^2}, \frac{t}{\mathrm{E}M_\varepsilon}, \frac{t}{n\omega_n}, \left(\frac{t}{\omega_n\sqrt{n}} \right)^{2/3}, \sqrt{\frac{t}{\omega_n}} \right) \right), \end{aligned}$$

for some constant $\bar{k} > 0$, since $t = \bar{c}(n-1)|\mathcal{M}|\kappa_n - \bar{k}\mathrm{E}Z_\varepsilon \gg \bar{c}(n-1)|\mathcal{M}|\kappa_n/2$.

Using similar arguments in the proof of Corollary 6 in Clémençon, Lugosi and Vayatis (2008), we can show $\mathrm{E}U_\varepsilon = O(\sqrt{|\mathcal{M}|n})$. Besides, by definition, it is immediate to see $\mathrm{E}M_\varepsilon = O(n\omega_n)$. Hence, it follows from (C.16), (C.18)

and $n \gg \kappa_n \gg \log(n)$ that for sufficiently large n,

$$\begin{aligned}
(\text{C.19}) \quad & \Pr\left(\sup_{\beta\in B_{\mathcal{M}}} \frac{n}{|\mathcal{M}|}\left|\frac{1}{n(n-1)}\sum_{i\neq j} D(O_i,O_j,\beta)\right| \geq \bar{c}\kappa_n\right) \\
\leq\ & 3\exp\left(-K\min\left(|\mathcal{M}|\kappa_n^2, \frac{|\mathcal{M}|\kappa_n}{\log(n)}, \left(\frac{|\mathcal{M}|\sqrt{n}\kappa_n}{\log(n)}\right)^{2/3}, \sqrt{\frac{|\mathcal{M}|n\kappa_n}{\log(n)}}\right)\right) \\
\leq\ & 3\exp\left(-K\min\left(\frac{|\mathcal{M}|\kappa_n}{\log(n)}, \left(\frac{|\mathcal{M}|\sqrt{n}\kappa_n}{\log(n)}\right)^{2/3}, \sqrt{\frac{|\mathcal{M}|n\kappa_n}{\log(n)}}\right)\right),
\end{aligned}$$

for some constant $K > 0$. Recall $\Omega_s^* = \{\mathcal{M} \in \Omega_+^*, |\mathcal{M}| = s\}$. The number of elements $|\Omega_s^*|$ is bounded by $O(p^s)$. This with (C.19) suggests

$$\begin{aligned}
(\text{C.20}) \quad & \sum_{\mathcal{M}\in\Omega_s^*} \Pr\left(\sup_{\beta\in B_{\mathcal{M}}} \frac{n}{s}\left|\frac{1}{n(n-1)}\sum_{i\neq j} D(O_i,O_j,\beta)\right| \geq \bar{c}\kappa_n\right) \\
\leq\ & O(p^s)\exp\left(-K\min\left(\frac{s\kappa_n}{\log(n)}, \left(\frac{s\sqrt{n}\kappa_n}{\log(n)}\right)^{2/3}, \sqrt{\frac{sn\kappa_n}{\log(n)}}\right)\right) \\
\leq\ & \exp\left(-K\min\left(\frac{s\kappa_n}{\log(n)}, \left(\frac{s\sqrt{n}\kappa_n}{\log(n)}\right)^{2/3}, \sqrt{\frac{sn\kappa_n}{\log(n)}}\right) + O(s\log(p))\right).
\end{aligned}$$

Under the given condition $\kappa_n \gg \log(p)\log(n)$, $n \gg s\log(p)$ for all $s \leq s_n$, we obtain

$$\frac{s\kappa_n}{\log(n)} \gg s\log(p), \ \left(\frac{s\sqrt{n}\kappa_n}{\log(n)}\right)^{2/3} \gg s\log(p), \ \sqrt{\frac{sn\kappa_n}{\log(n)}} \gg s\log(p).$$

Together with the condition $\kappa_n = o(n)$, (C.20) is bounded by

$$\begin{aligned}
& \exp\left(-\frac{K}{2}\min\left(\frac{s\kappa_n}{\log(n)}, \left(\frac{s\sqrt{n}\kappa_n}{\log(n)}\right)^{2/3}, \sqrt{\frac{sn\kappa_n}{\log(n)}}\right)\right) \\
\leq\ & \exp\left(-\frac{K}{2}\min\left(\frac{\kappa_n}{\log(n)}, \left(\frac{\sqrt{n}\kappa_n}{\log(n)}\right)^{2/3}, \sqrt{\frac{n\kappa_n}{\log(n)}}\right)\right) \leq \exp\left(-\frac{K\kappa_n}{2\log(n)}\right),
\end{aligned}$$

for sufficiently large n. Since $\Omega_+^* = \cup_{s=1}^{s_n}\Omega_s^*$, this together with (C.20) suggests (C.14) is bounded by

$$\begin{aligned}
(\text{C.21}) |s_n|\exp\left(-\frac{K\kappa_n}{2\log(n)}\right) \ &\leq\ \exp\left(-\frac{K\kappa_n}{2\log(n)} + \log(n)\right) \\
&\leq\ \exp\left(-\frac{K\kappa_n}{3\log(n)}\right) \leq \exp(-K\log(p)),
\end{aligned}$$

since $s_n = o(n)$, $\kappa_n \gg \log(p)\log(n)$. Thus, we've establish the upper bound for (C.14). Now, we provide an upper bound for

$$\Pr\left(\sup_{\substack{\mathcal{M}\in\Omega_+^* \\ \beta\in B_{\mathcal{M}}^C}} \frac{1}{|\mathcal{M}|}\left|\sum_{i=1}^n \{g(O_i,\beta) - g(O_i,\beta_0) - C(\beta) + C(\beta_0)\}\right| \geq \bar{c}\kappa_n\right). \tag{C.22}$$

It follows from Assumption (A6')(ii) that

$$\sup_{\beta\in B_{\mathcal{M}}^C} \frac{1}{|\mathcal{M}|}|C(\beta) - C(\beta_0)| \leq \bar{c}_2 \frac{(R_{|\mathcal{M}|}^C)^2}{|\mathcal{M}|} \leq O(1)\frac{\log(p)\log(n)}{n}, \quad \forall M \in \Omega_+^*.$$

Since $\kappa_n \gg \log(p)\log(n)$, (C.22) is bounded by

$$\Pr\left(\sup_{\substack{\mathcal{M}\in\Omega_+^* \\ \beta\in B_{\mathcal{M}}^C}} \frac{1}{|\mathcal{M}|}\left|\sum_{i=1}^n \{g(O_i,\beta) - g(O_i,\beta_0)\}\right| \geq \frac{\bar{c}}{2}\kappa_n\right), \tag{C.23}$$

for sufficiently large n.

By Assumption (A6')(iii), g is twice continuously differentiable around β_0. For any $\beta \in B_{\mathcal{M}}^C$, a second-order Taylor expansion gives

$$\begin{aligned} g(O_i,\beta) &= g(O_i,\beta_0) + \frac{\partial g(O_i,\beta_0)}{\partial\beta}(\beta_0 - \beta) \\ &+ \frac{1}{2}(\beta_0 - \beta)^T \Delta_2 g(O_i,\beta^*)(\beta_0 - \beta), \end{aligned} \tag{C.24}$$

for some β^* lying on the line segment joining β and β_0. It follows from $\sup_{\mathcal{M}\in\Omega_+^*} R_{|\mathcal{M}|}^C \to 0$ that $\|\beta^* - \beta_0\|_2 \to 0$. Therefore, by Assumption (A6')(iv),

$$\begin{aligned} &\|\Delta_2 g(O_i,\beta^*) - \Delta_2 g(O_i,\beta_0)\|_2 \leq K(O_i)\|\beta^* - \beta_0\|_2 \\ \leq\ & K(O_i)\|\beta - \beta_0\|_2 \leq R_{|\mathcal{M}|}^C K(O_i) = o(1)K(O_i), \end{aligned} \tag{C.25}$$

where the $o(1)$ term is uniform in $i = 1,\ldots,n$. Combining (C.24) with (C.25) gives

$$\begin{aligned} g(O_i,\beta) &= g(O_i,\beta_0) + \frac{\partial g(O_i,\beta_0)}{\partial\beta}(\beta_0 - \beta) \\ &+ \frac{1}{2}(\beta_0 - \beta)^T \Delta_2 g(O_i,\beta_0)(\beta_0 - \beta) + o(1)\|\beta - \beta_0\|_2^2 K(O_i). \end{aligned} \tag{C.26}$$

For any $\beta \in B_{\mathcal{M}}^C$, it follows from Assumption (A6')(v) that

$$|(\beta_0-\beta)^T \mathrm{E}\Delta_2 g(O_i,\beta_0)(\beta_0-\beta)| = O(\|\beta-\beta_0\|_2^2), \tag{C.27}$$

Notice that

$$\mathrm{E}K(O_0) \le \|K(O_i)\|_{\psi_1} = O(1). \tag{C.28}$$

Define the events

$$\mathcal{E}_0 = \left\{ \sum_i K(O_i) \le 2n\mathrm{E}K(O_0) \right\},$$
$$\mathcal{E}_1 = \left\{ \max_{kj} \left| \sum_i \{\partial_{kj} g(O_i,\beta_0) - \mathrm{E}\partial_{kj} g(O_i,\beta_0)\} \right| \le \sqrt{n\log(p)\log(n)} \right\}.$$

Assume $\mathcal{E}_0 \cup \mathcal{E}_1$ holds. It follows that

$$\begin{aligned}
&\left| \sum_{i=1}^n (\beta_0-\beta)^T \{\Delta_2 g(O_i,\beta_0) - \mathrm{E}\Delta_2 g(O_i,\beta_0)\}(\beta_0-\beta) \right| \\
= &\left| \sum_{k_1,k_2\in\mathcal{M}} (\beta_0^{k_1}-\beta^{k_1})(\beta_0^{k_2}-\beta^{k_2}) \left[\sum_i \{\partial_{k_1k_2} g(O_i,\beta_0) - \mathrm{E}\partial_{kj} g(O_i,\beta_0)\} \right] \right| \\
\le &\ |\mathcal{M}| \|\beta_0-\beta\|_2^2 \max_{k_1k_2} \left| \sum_i \{\partial_{k_1k_2} g(O_i,\beta_0) - \mathrm{E}\partial_{k_1k_2} g(O_i,\beta_0)\} \right| \\
= &\ O\left(|\mathcal{M}| \sqrt{n\log(p)\log(n)} \|\beta_0-\beta\|_2^2 \right) = (n\|\beta_0-\beta\|_2^2),
\end{aligned}$$

where the last equality is due to the condition that $n \gg s_n^2 \log(p)\log(n)$. This together with (C.26) and (C.28) implies

$$\sum_i g(O_i,\beta) - g(O_i,\beta_0) = \sum_i \frac{\partial g(O_i,\beta_0)}{\partial\beta}(\beta-\beta_0) + O(1)n\|\beta-\beta_0\|_2^2.$$

Since $\kappa_n \gg n(R_{|\mathcal{M}|}^C)^2/|\mathcal{M}| = O(\log(p)\log(n))$, this implies under $\mathcal{E}_0, \mathcal{E}_1$, (C.23) is bounded by

$$\Pr\left(\sup_{\mathcal{M}\in\Omega_+^*} \sup_{\beta\in B_{\mathcal{M}}^C} \left| \sum_i \frac{1}{|\mathcal{M}|} \frac{\partial g(O_i,\beta_0)}{\partial\beta}(\beta-\beta_0)^T \right| \ge \frac{\bar{c}\kappa_n}{4} \right),$$

or equivalently,

$$\Pr\left(\sup_{\mathcal{M}\in\Omega_+^*} \sup_{\beta\in B_{\mathcal{M}}^C} \sup_{j\in\mathcal{M}} \left| \sum_i \partial_j g(O_i,\beta_0) \right|_2 \ge \frac{\bar{c}\sqrt{|\mathcal{M}|}\kappa_n}{4R_{|\mathcal{M}|}^C} \right). \tag{C.29}$$

Define

$$\mathcal{E}_2 = \left\{ \sup_j \left| \sum_i \partial_j g(O_i, \beta_0) \right| \geq \frac{\bar{c}\sqrt{|\mathcal{M}|}\kappa_n}{4R^C_{|\mathcal{M}|}} \right\}. \tag{C.30}$$

The probability defined in (C.29) is bounded by $\Pr(\mathcal{E}_2)$. From the above discussion, we've shown (C.23) is bounded by $\Pr(\mathcal{E}_0 \cup \mathcal{E}_1 \cup \mathcal{E}_2)$.

Under Assumption (A6')(iii) and (v), we have

$$\|K(O_i)\|_{\psi_1} = O(1),\ \sup_j \|\partial_j g(O_i, \beta_0)\|_{\psi_1} = O(1),\ \sup_{jk} \|\partial_{jk} g(O_i, \beta_0)\|_{\psi_1} = O(1).$$

Since $C(\beta_0)$ maximizes C, we have $\partial C(\beta_0)/\partial\beta = 0$ and thus

$$\mathrm{E}\frac{\partial g(O_i, \beta_0)}{\partial \beta} = 0.$$

It follows from Lemma H.2 that there exists some constants $\bar{c}_0 > 0$ that

$$\begin{aligned} &\sup_j \Pr\left(\left| \sum_i \partial_j g(O_i, \beta_0) \right| \geq \frac{\bar{c}\sqrt{|\mathcal{M}|}\kappa_n}{4R^C_{|\mathcal{M}|}} \right) \\ \leq\ & 2\exp\left(-\frac{\bar{c}_0 \kappa_n^2}{\log(p)\log(n)} \right) + 2\exp\left(-\frac{\bar{c}_0 \sqrt{n}\kappa_n}{\sqrt{\log(p)\log(n)}} \right) \leq 4\exp(-\bar{c}_0 \log(p)\log(n)), \end{aligned} \tag{C.31}$$

where the last inequality is due to $\kappa_n \gg \log(p)\log(n)$ and $n \gg \log(p)\log(n)$. Similarly we can show

$$\begin{aligned} &\Pr\left(\left| \sum_i K(O_i) \geq 2n\mathrm{E}K(O_0) \right| \right) \leq 2\exp(-\bar{c}_0 \log(p)\log(n)) \\ &\sup_{jk} \Pr\left(\left| \sum_i \{\partial_{jk} g(O_i, \beta_0) - \mathrm{E}\partial_{jk} g(O_i, \beta_0)\} \right| \geq \sqrt{n\log(p)\log(n)} \right) \\ &\qquad \leq 2\exp(-\bar{c}_0 \log(p)\log(n)). \end{aligned} \tag{C.32}$$

Using Bonferroni's inequality, it follows from (C.31) and (C.32) that $\Pr(\mathcal{E}_0 \cup \mathcal{E}_1 \cup \mathcal{E}_2)$ is bounded by

$$\begin{aligned} &\Pr(\mathcal{E}_0 \cup \mathcal{E}_1 \cup \mathcal{E}_2) \leq 2(p^2+1)\exp\left(-\bar{c}_0 \log(p)\log(n)\right) + 4p\exp(-\bar{c}_0 \log(p)\log(n)) \\ \leq\ & \exp(-\bar{c}_0 \log(p)\log(n) + \log(p) + \log 4) + \exp(-\bar{c}_0 \log(p)\log(n) + \log(2p^2+2)) \\ \leq\ & 2\exp(-\bar{c}_0 \log(p)\log(n)/2) \leq \exp(-\bar{c}_0 \log(p)/4). \end{aligned}$$

Combining this together with (C.21), we've shown the first term of (C.13) is bounded by

$$\exp(-K\log(p)) + \exp(-\bar{c}_0\log(p)/4) \le 2\exp(-\bar{K}\log(p)) \le \exp(-\bar{K}\log(p)/2),$$

for some constant $\bar{K} > 0$ and sufficiently large n.

Similarly, we can show that conditional on (C.12), the second term of (C.13) is bounded by

$$\exp\left(-\frac{\bar{K}_0\kappa_n^2}{n(R_n^{(1)})^2}\right) + \exp(-\bar{K}_0\log(p)),$$

for some constant $\bar{K}_0 > 0$. By Lemma 7.1, the event in (C.12) happens with probability at least $1 - \exp(-K_0\log(p))$ for some $K_0 > 0$. Thus, (C.2) can be bounded by

$$\begin{aligned} & \exp\left(-\frac{\bar{K}_0\kappa_n^2}{n(R_n^{(1)})^2}\right) + \exp(-\bar{K}_0\log(p)) + \exp(-\bar{K}\log(p)/2) \\ \le\ & \exp(-\bar{c}\log(p)) + \exp\left(-\frac{\bar{c}\kappa_n^2}{n(R_n^{(1)})^2}\right), \end{aligned}$$

for some constant $\bar{c} > 0$. The proof is hence completed.

APPENDIX D: PROOF OF THEOREM 7.1

Let F be the envelope of the class of functions $\mathcal{F}$, i.e,

$$F(x,y) = \sup_{f\in\mathcal{F}} |f(x,y)|.$$

For a given $\rho > 0$, define $\mathcal{F}_1$ and $\mathcal{F}_2$ to be the following class of functions

$$\begin{aligned} \mathcal{F}_1 &= \{f(x,y)\mathbb{I}(F(x,y) \le \rho), f \in \mathcal{F}\}, \\ \mathcal{F}_2 &= \{f(x,y)\mathbb{I}(F(x,y) > \rho), f \in \mathcal{F}\}. \end{aligned}$$

It is immediate to see that

$$\begin{aligned} Z \ \le\ & \sup_{f_1\in\mathcal{F}_1} \left| \sum_{i\neq j} \{f_1(X_i,X_j) - \mathrm{E}f_1(X_i,X_j)\} \right| \\ +\ & \sup_{f_2\in\mathcal{F}_2} \left| \sum_{i\neq j} \{f_2(X_i,X_j) - \mathrm{E}f_2(X_i,X_j)\} \right| \triangleq Z_1 + Z_2. \end{aligned} \tag{D.1}$$

Besides, it follows from Jensen's inequality that

$$\mathrm{E}Z_2 \le 2\mathrm{E}\sup_{f_2\in\mathcal{F}_2}\left|\sum_{i\neq j}^{n} f_2(X_i,X_j)\right| \triangleq 2\mathrm{E}I. \tag{D.2}$$

Combining (D.1) with (D.2), we obtain

$$\begin{aligned}&\Pr\left(Z \ge C\mathrm{E}Z_\varepsilon + 2t\right)\\ \le\ &\Pr\left(Z_1 \ge C\mathrm{E}Z_\varepsilon + t\right) + \Pr\left(Z_2 \ge t\right).\end{aligned} \tag{D.3}$$

Hence, it suffices to provide upper bounds on the probabilities in the second line of (D.3). We first bound $\Pr(Z_2 \ge t)$.

Observe that for any function f, we can express the U-statistic as average of sums of i.i.d blocks,

$$\frac{1}{n(n-1)}\sum_{i\neq j} f(X_i,X_j) = \frac{1}{n!}\sum_{\pi}\frac{1}{\lfloor n/2\rfloor}\sum_{i=1}^{\lfloor n/2\rfloor} f(X_{\pi(i)},X_{\pi(i+\lfloor n/2\rfloor)}). \tag{D.4}$$

where π stands for all permutations over $[1,\ldots,n]$ and $\lfloor x\rfloor$ stands for the largest integer y such that $y \le x$.

Let $\rho = 8\omega_n$. It follows from the definition of ω_n that

$$8\mathrm{E}\max_{1\le i\le\lfloor n/2\rfloor} F(X_i,X_{i+\lfloor n/2\rfloor}) \le 8\|\max_i F(X_i,X_{i+\lfloor n/2\rfloor})\|_{\psi_1} \le \rho, \tag{D.5}$$

By Chebyshev inequality, we have

$$\begin{aligned}&\Pr\left(\max_{k\le\lfloor n/2\rfloor}\sup_{f_2\in\mathcal{F}_2}\left|\sum_{i=1}^{k} f_2(X_i,X_{i+\lfloor n/2\rfloor})\right| > 0\right)\\ \le\ &\Pr\left(\max_{1\le i\le\lfloor n/2\rfloor} F(X_i,X_{i+\lfloor n/2\rfloor}) > \rho\right) \le \frac{1}{8}.\end{aligned}$$

Combining this with the Hoffmann-Jørgensen inequality (cf. Ledoux and Talagrand, 2011, Proposition 6.8), we have

$$\mathrm{E}\sup_{f_2\in\mathcal{F}_2}\left|\sum_{i=1}^{\lfloor n/2\rfloor} f_2(X_i,X_{i+\lfloor n/2\rfloor})\right| \le 8\mathrm{E}\max_{1\le i\le\lfloor n/2\rfloor} F(X_i,X_{i+\lfloor n/2\rfloor}). \tag{D.6}$$

Apply the decomposition (D.4) to the class of functions in $\mathcal{F}_2$, it follows from Lemma A.1 in Clémençon, Lugosi and Vayatis (2008) that

$$\begin{aligned} \mathrm{E}I &\leq \frac{1}{n!}\sum_{\pi}\mathrm{E}\sup_{f_2\in\mathcal{F}_2}\left|\frac{1}{\lfloor n/2\rfloor}\sum_{i=1}^{\lfloor n/2\rfloor} f_2(X_{\pi(i)}, X_{\pi(i+\lfloor n/2\rfloor)})\right| \\ &\leq \frac{1}{\lfloor n/2\rfloor}\mathrm{E}\sup_{f_2\in\mathcal{F}_2}\left|\sum_{i=1}^{\lfloor n/2\rfloor} f_2(X_i, X_{i+\lfloor n/2\rfloor})\right|. \end{aligned}$$

Consequently, together with (D.6) and (D.5), we obtain

$$\begin{aligned} \mathrm{E}I &\leq 8\mathrm{E}\max_{1\leq i\leq\lfloor n/2\rfloor} F(X_i, X_{i+\lfloor n/2\rfloor}) \\ &\leq 8\left\|\max_{1\leq i\leq\lfloor n/2\rfloor} F(X_i, X_{i+\lfloor n/2\rfloor})\right\|_{\psi_1} \leq \rho. \end{aligned} \tag{D.7}$$

Due to the decomposition in (D.4), we have

$$\begin{aligned} \|Z_2\|_{\psi_1} &\leq \frac{1}{n!}\sum_{\pi}\frac{n(n-1)}{\lfloor n/2\rfloor}\left\|\sup_{f_2\in\mathcal{F}_2}\left|\sum_{i=1}^{\lfloor n/2\rfloor}\{f_2(X_{\pi(i)}, X_{\pi(i+\lfloor n/2\rfloor)}) - \mathrm{E}f_2\}\right|\right\|_{\psi_1} \\ &\leq 2n\left\|\sup_{f_2\in\mathcal{F}_2}\left|\sum_{i=1}^{\lfloor n/2\rfloor}\{f_2(X_i, X_{i+\lfloor n/2\rfloor}) - \mathrm{E}f_2(X_i, X_{i+\lfloor n/2\rfloor})\}\right|\right\|_{\psi_1}. \end{aligned} \tag{D.8}$$

It follows from Theorem 6.21 in Ledoux and Talagrand (2011) that

$$\begin{aligned} &\left\|\sup_{f_2\in\mathcal{F}_2}\left|\sum_{i=1}^{\lfloor n/2\rfloor}\{f_2(X_i, X_{i+\lfloor n/2\rfloor}) - \mathrm{E}f_2(X_i, X_{i+\lfloor n/2\rfloor})\}\right|\right\|_{\psi_1} \\ \leq\ &\mathrm{E}\sup_{f_2\in\mathcal{F}_2}\left|\sum_{i=1}^{\lfloor n/2\rfloor}\{f_2(X_i, X_{i+\lfloor n/2\rfloor}) - \mathrm{E}f_2(X_i, X_{i+\lfloor n/2\rfloor})\}\right| + 2\left\|\max_{i=1}^{\lfloor n/2\rfloor} F(X_i, X_{i+\lfloor n/2\rfloor})\right\|_{\psi_1}. \end{aligned}$$

This together with (D.2), (D.5), (D.6) and (D.7) yields that

$$\left\|\sup_{f_2\in\mathcal{F}_2}\left|\sum_{i=1}^{\lfloor n/2\rfloor}\{f_2(X_i, X_{i+\lfloor n/2\rfloor}) - \mathrm{E}f_2(X_i, X_{i+\lfloor n/2\rfloor})\}\right|\right\|_{\psi_1} \leq K'\omega_n/2,$$

for some constant K'. Combining this with (D.8), we obtain

$$\|Z_2\|_{\psi_1} \leq nK'\omega_n.$$

Therefore, by Markov's inequality, we obtain

$$\Pr(Z_2 \geq t) \leq 2\exp\left(-\frac{t}{nK'\omega_n}\right). \tag{D.9}$$

It remains to bound

$$\Pr(Z_1 \geq C\mathrm{E}Z_\varepsilon + t).$$

Since each function in $\mathcal{F}_1$ is bounded, it follows from Lemma (H.5) that

$$\begin{aligned}\Pr(Z_1 \geq C\mathrm{E}Z_1 + t) \\ \leq \exp\left(-\frac{1}{C}\min\left(\frac{t^2}{(\mathrm{E}U_\varepsilon)^2}, \frac{t}{\mathrm{E}M_\varepsilon}, \frac{t}{8n\omega_n}, \left(\frac{t}{8\sqrt{n}\omega_n}\right)^{2/3}, \sqrt{\frac{t}{8\omega_n}}\right)\right).\end{aligned} \tag{D.10}$$

The proof is hence completed by combining (D.9) with (D.10).

APPENDIX E: PROOF OF LEMMA 7.2

E.1. Uniform convergence of empirical maximizer of $\widehat{V}$. Recall that $\Omega_+^* = \{\mathcal{M} \subseteq \{1, \ldots, p\} : \mathcal{M}_{\beta_0} \subsetneq \mathcal{M}, |\mathcal{M}| \leq s_n\}$. Hence, we have

$$\{\mathcal{M}(\lambda) : \lambda \in \Omega_+\} \subseteq \Omega_+^*. \tag{E.1}$$

Since $\tilde{\theta}_{\mathcal{M}(\lambda_1)} = \tilde{\theta}_{\mathcal{M}(\lambda_2)}$ whenever $\mathcal{M}(\lambda_1) = \mathcal{M}(\lambda_2)$, we have

$$\begin{aligned}&\bigcap_{\lambda\in\Omega_+}\left\{\|\tilde{\theta}_{\mathcal{M}(\lambda)} - \theta_0\|_2 \geq tn^{-1/3}|\mathcal{M}(\lambda)|^{1/3}\log^{1/3} p\right\} \\ \subseteq &\bigcap_{\mathcal{M}\in\Omega_+^*}\left\{\|\tilde{\theta}_{\mathcal{M}} - \theta_0\|_2 \geq tn^{-1/3}|\mathcal{M}(\lambda)|^{1/3}\log^{1/3} p\right\}.\end{aligned} \tag{E.2}$$

Define

$$\mathcal{A}_0 = \bigcap_{\mathcal{M}\in\Omega_+^*}\left\{\tilde{\theta}_{\mathcal{M}} \in \tilde{N}_{\varepsilon_0}\right\},$$

where ε_0 is defined in (A5')(i). It follows from (E.2) that

$$
\begin{aligned}
(\text{E.3}) \quad & \Pr\left(\bigcap_{\lambda\in\Omega_+}\left\{\|\tilde{\theta}_{\mathcal{M}(\lambda)}-\theta_0\|_2 \geq tn^{-1/3}|\mathcal{M}(\lambda)|^{1/3}\log^{1/3}p\right\}\right) \\
\leq\ & \Pr\left(\bigcap_{\mathcal{M}\in\Omega_+^*}\left\{\|\tilde{\theta}_{\mathcal{M}}-\theta_0\|_2 \geq tn^{-1/3}|\mathcal{M}|^{1/3}\log^{1/3}p\right\}\bigcap\mathcal{A}_0^c\right) \\
+\ & \Pr\left(\bigcap_{\mathcal{M}\in\Omega_+^*}\left\{\|\tilde{\theta}_{\mathcal{M}}-\theta_0\|_2 \geq tn^{-1/3}|\mathcal{M}|^{1/3}\log^{1/3}p\right\}\bigcap\mathcal{A}_0\right) \\
\leq\ & \Pr(\mathcal{A}_0^c)+\Pr\left(\bigcap_{\mathcal{M}\in\Omega_+}\left\{\|\tilde{\theta}_{\mathcal{M}}-\theta_0\|_2 \geq \frac{t|\mathcal{M}|^{1/3}\log^{1/3}p}{n^{1/3}}\right\}\bigcap\mathcal{A}_0\right).
\end{aligned}
$$

In view of (E.1) and (E.3), it suffices to bound

$$
\Pr(\mathcal{A}_0^c)+\Pr\left(\bigcap_{\mathcal{M}\in\Omega_+^*}\left\{\|\tilde{\theta}_{\mathcal{M}}-\theta_0\|_2 \geq tn^{-1/3}|\mathcal{M}|^{1/3}\log^{1/3}p\right\}\bigcap\mathcal{A}_0\right).
$$

In the following, we break the proof into two steps. In the first step, we show

$$
\Pr(\mathcal{A}_0^c) \leq \exp\left(-\frac{\bar{c}_0 n}{\log(n)}\right), \tag{E.4}
$$

for some constant $\bar{c}_0 > 0$. In the second step, we show

$$
\begin{aligned}
(\text{E.5}) \quad & \Pr\left(\bigcap_{\mathcal{M}\in\Omega_+^*}\left\{\|\tilde{\theta}_{\mathcal{M}}-\theta_0\|_2 \geq tn^{-1/3}|\mathcal{M}|^{1/3}\log^{1/3}p\right\}\bigcap\mathcal{A}_0\right) \\
\leq\ & \exp\left(-c_0^* t^3\log(p)\right)+\exp\left(-\frac{c_0^* t^2 n^{1/3}\log^{2/3}p}{\log(n)}\right),
\end{aligned}
$$

for some constant $c_0^* > 0$ and all $t \geq t_0$, (7.7) thus holds by setting $\bar{c} = \min(\bar{c}_0, c_0^*)$.

E.1.1. *Proof of* (E.4). For any $\mathcal{M} \in \Omega_+^*$, if $\tilde{\theta}_{\mathcal{M}} \notin \tilde{N}_{\varepsilon_0}$, by definition, we have

$$
\sup_{\substack{\theta=(c,\beta^T)^T\in\tilde{S}(\theta_0)\cap\tilde{N}_{\varepsilon_0}^c \\ \beta^{\mathcal{M}^c}=0}} \widehat{V}(\theta) \geq \sup_{\substack{\theta=(c,\beta^T)^T\in\tilde{S}(\theta_0)\cap\tilde{N}_{\varepsilon_0} \\ \beta^{\mathcal{M}^c}=0}} \widehat{V}(\theta),
$$

and hence

$$\sup_{\substack{\theta=(c,\beta^T)^T\in\tilde{S}(\theta_0)\cap\tilde{N}^c_{\varepsilon_0}\\ \beta^{\mathcal{M}^c}=0}} \widehat{V}(\theta) \geq \widehat{V}(\theta_0).$$

This further implies

$$\sup_{\substack{\theta=(c,\beta^T)^T\in\tilde{S}(\theta_0)\cap\tilde{N}^c_{\varepsilon_0}\\ \beta^{\mathcal{M}^c}=0}} \{\widehat{m}_V(\theta) - m_V(\theta)\} \geq V(\theta_0) - \sup_{\substack{\theta=(c,\beta^T)^T\in\tilde{S}(\theta_0)\cap\tilde{N}^c_{\varepsilon_0}\\ \beta^{\mathcal{M}^c}=0}} V(\theta). \tag{E.6}$$

By Assumption (A5'), there exists some constant $\delta_0 > 0$ such that for sufficiently large n,

$$V(\theta_0) - \sup_{\substack{\theta=(c,\beta^T)^T\in\tilde{S}(\theta_0)\cap\tilde{N}^c_{\varepsilon_0}\\ \beta^{\mathcal{M}^c}=0}} V(\theta) \geq 2\delta_0.$$

Combining this together with (E.6) gives

$$\sup_{\substack{\theta=(c,\beta^T)^T\in\tilde{S}(\theta_0)\cap\tilde{N}^c_{\varepsilon_0}\\ \beta^{\mathcal{M}^c}=0}} \{\widehat{m}_V(\theta) - m_V(\theta)\} \geq 2\delta_0,$$

which further implies

$$\sup_{\substack{\theta=(c,\beta^T)^T\\ \beta^{\mathcal{M}^c}=0}} \left|\widehat{V}(\theta) - V(\theta)\right| \geq \delta_0 \quad \text{or} \quad |\widehat{V}(\theta_0) - V(\theta_0)| \geq \delta_0.$$

To summarize, we've shown

$$\begin{aligned}
\Pr(\tilde{\theta}_M \notin \tilde{N}_{\varepsilon_0}) &\leq \Pr\left(\sup_{\substack{\theta=(c,\beta^T)^T\\ \beta^{\mathcal{M}^c}=0}} |m_V(\theta)| \geq \delta_0\right) + \Pr(|m_V(\theta_0)| \geq \delta_0) \\
&\leq 2\Pr\left(\sup_{\substack{\theta=(c,\beta^T)^T\\ \beta^{\mathcal{M}^c}=0}} |m_V(\theta)| \geq \delta_0\right),
\end{aligned}$$

where the last inequality is due to that $\beta_0^{\mathcal{M}^c} = 0$. It follows from Bonferroni's inequality that

$$2\Pr(\mathcal{A}_0^c) \leq 2\sum_{\mathcal{M}\in\Omega_+^*} \left(\sup_{\substack{\theta=(c,\beta^T)^T\\ \beta^{\mathcal{M}^c}=0}} |m_V(\theta)| \geq \delta_0\right). \tag{E.7}$$

Similar to (9.10), we can show that RHS of (E.7) is upper bounded by $\exp(-\bar{c}n/\log(n))$. This proves (E.4).

E.1.2. *Proof of* (E.5). On the set $\mathcal{A}_0$, it follows from Assumption (A5')(iii) that for all $\mathcal{M} \in \Omega_+^*$, we have

$$V(\theta_0) - V(\tilde{\theta}_{\mathcal{M}}) \geq \bar{c}_1 \|\theta_0 - \tilde{\theta}_{\mathcal{M}}\|_2^2.$$

When $\|\theta_0 - \tilde{\theta}_{\mathcal{M}}\|_2 \geq t n^{-1/3} |\mathcal{M}|^{1/3} \log^{1/3} p$, we have

$$V(\theta_0) - V(\tilde{\theta}_{\mathcal{M}}) \geq \bar{c}_1 t^2 n^{-2/3} |\mathcal{M}|^{2/3} \log^{2/3} p. \tag{E.8}$$

It follows from the definition of $\tilde{\theta}_{\mathcal{M}}$ that $\widehat{V}(\tilde{\theta}_{\mathcal{M}}) \geq \widehat{V}(\theta_0)$, which together with (E.8) gives

$$\widehat{V}(\tilde{\theta}_{\mathcal{M}}) - V(\tilde{\theta}_{\mathcal{M}}) - \widehat{V}(\theta_0) + V(\theta_0) \geq \bar{c}_1 t^2 n^{-2/3} |\mathcal{M}|^{2/3} \log^{2/3} p,$$

or equivalently,

$$\widehat{m}_V(\tilde{\theta}_{\mathcal{M}}) - m_V(\tilde{\theta}_{\mathcal{M}}) \geq \bar{c}_1 t^2 n^{-2/3} |\mathcal{M}|^{2/3} \log^{2/3} p.$$

Hence, LHS of (E.5) is bounded by

$$\Pr\left(\bigcap_{\mathcal{M} \in \Omega_+^*} \left| \widehat{m}_V(\tilde{\theta}_{\mathcal{M}}) - m_V(\tilde{\theta}_{\mathcal{M}}) \right| \geq \bar{c}_1 t^2 n^{-2/3} |\mathcal{M}|^{2/3} \log^{2/3} p \right).$$

For sufficiently large t_0 and any $t \geq t_0$, (E.5) follows using similar arguments in proving (9.21). This completes the proof.

E.2. Uniform convergence of empirical maximizer of $\widehat{C}$. Define

$$\mathcal{A}_0 = \bigcap_{\mathcal{M} \in \Omega_+^*} \left\{ \tilde{\beta}_{\mathcal{M}} \in N_{\varepsilon_0} \right\}.$$

Similar to Section E.1, it suffices to show

$$\Pr(\mathcal{A}_0^c) \leq \exp\left(-\frac{\bar{c}n}{\log(n)} \right), \tag{E.9}$$

and

$$\begin{aligned} &\Pr\left(\bigcap_{\mathcal{M} \in \Omega_+^*} \left\{ \|\tilde{\beta}_{\mathcal{M}} - \beta_0\|_2 \geq \frac{t |\mathcal{M}|^{1/2} \log^{1/2}(p) \log^{1/2}(n)}{\sqrt{n}} \right\} \bigcap \mathcal{A}_0 \right) \\ \leq\ &\exp\left(-\bar{c} t^2 \log(p) \right) + \exp\left(-\bar{c} t \sqrt{n \log(p)} \right), \end{aligned} \tag{E.10}$$

for some constant $\bar{c} > 0$.

E.2.1. *Proof of* (E.9). When $\mathcal{A}_0^c$ holds, it follows from Assumption (A6')(i) that

$$\sup_{\mathcal{M}\in\Omega_+^*}\left\{C(\beta_0)-C(\tilde{\beta}_{\mathcal{M}})\right\}\geq\xi_0,$$

for some constant $\xi_0>0$. By the definition of $\tilde{\beta}_{\mathcal{M}}$, this further implies

$$\sup_{\mathcal{M}\in\Omega_+^*}\left\{\widehat{C}(\tilde{\beta}_{\mathcal{M}})-\widehat{C}(\beta_0)-C(\tilde{\beta}_{\mathcal{M}})+C(\beta_0)\right\}\geq\xi_0.$$

Therefore, we have

$$\Pr(\mathcal{A}_0^c)\leq\Pr\left[\sup_{\mathcal{M}\in\Omega_+^*}\left\{\widehat{C}(\tilde{\beta}_{\mathcal{M}})-\widehat{C}(\beta_0)-C(\tilde{\beta}_{\mathcal{M}})+C(\beta_0)\right\}\geq\xi_0\right],$$

the RHS of which can be bounded in a similar manner as (C.4). Assertion (E.9) is thus proven.

E.2.2. *Proof of* (E.10). When the event defined on the LHS of (E.10) holds, we have by Assumption (A6')(iii) that

$$\bigcup_{\mathcal{M}\in\Omega_+^*}\left\{\left(C(\beta_0)-C(\tilde{\beta}_{\mathcal{M}})\right)\geq\frac{\bar{c}_1|\mathcal{M}|t^2\log(p)\log(n)}{n}\right\}.$$

By the definition of $\tilde{\beta}_{\mathcal{M}}$, this further implies

$$\bigcup_{\mathcal{M}\in\Omega_+^*}\left\{\left(\widehat{C}(\tilde{\beta}_{\mathcal{M}})-\widehat{C}(\beta_0)-C(\tilde{\beta}_{\mathcal{M}})+C(\beta_0)\right)\geq\frac{\bar{c}_1|\mathcal{M}|t^2\log(p)\log(n)}{n}\right\}.$$

Hence, LHS of (E.10) is bounded by

(E.11)
$$\Pr\left(\bigcup_{\mathcal{M}\in\Omega_+^*}\left\{\left(\widehat{C}(\tilde{\beta}_{\mathcal{M}})-\widehat{C}(\beta_0)-C(\tilde{\beta}_{\mathcal{M}})+C(\beta_0)\right)\geq\frac{\bar{c}_1|\mathcal{M}|t^2\log(p)\log(n)}{n}\right\}\right).$$

Using similar arguments in bounding (C.13), we can show (E.11) is bounded by

$$\begin{aligned}&\frac{1}{2}\exp\left(-\bar{c}t^2\log(p)\right)+\frac{1}{2}\exp\left(-\bar{c}t\sqrt{n\log(p)}\right)\\+\;&\frac{1}{2}\exp\left(-\bar{c}n^{1/3}t^{4/3}\log^{2/3}p\right)+\exp\left(\frac{\bar{c}n}{\log(n)}\right),\end{aligned}$$

for some constant $\bar{c} > 0$.

Since $n^{1/3}t^{4/3}\log^{2/3} p = (t^2\log(p))^{1/3}(t\sqrt{n\log(p)})^{2/3}$, we have

$$n^{1/3}t^{4/3}\log^{2/3} p \geq \min(t^2\log(p), t\sqrt{n\log(p)}),$$

and therefore

$$\exp\left(-\bar{c}n^{1/3}t^{4/3}\log^{2/3} p\right) \leq \exp\left(-\bar{c}t^2\log(p)\right) + \exp\left(-\bar{c}t\sqrt{n\log(p)}\right).$$

This further implies (E.11) is bounded by

$$\exp\left(-\bar{c}t^2\log(p)\right) + \exp\left(-\bar{c}t\sqrt{n\log(p)}\right) + \exp\left(\frac{\bar{c}n}{\log(n)}\right).$$

The proof is hence completed.

APPENDIX F: PROOF OF THEOREM 10.1

F.1. Consistency of VIC^{DR}. For any $\mathcal{M}$, let

$$\tilde{\theta}_{\mathcal{M}} = \underset{\substack{\theta=(c,\beta^T)^T:\beta^{\mathcal{M}^c}=0 \\ \|\theta\|_2=\|\theta_0\|_2}}{\arg\max} \widehat{V}^{DR}(\theta).$$

It suffices to show with probability tending to 1,

$$\text{VIC}^{DR}(\hat{\theta}_{\mathcal{M}_{\beta_0}}) > \sup_{\mathcal{M}\in\Omega_-} \text{VIC}^{DR}(\hat{\theta}_{\mathcal{M}}), \tag{F.1}$$

$$\text{VIC}^{DR}(\hat{\theta}_{\mathcal{M}_{\beta_0}}) > \sup_{\mathcal{M}\in\Omega_+} \{n\widehat{V}^{DR}(\tilde{\theta}_{\mathcal{M}}) - \kappa_n\|\hat{\beta}_{\mathcal{M}}\|_0\}. \tag{F.2}$$

F.1.1. *Underfitted model space.* Similar to (9.8), it follows from Assumption (A4) and (A9)(i) that there exists some constant $\xi > 0$,

$$V^{DR}(\hat{\theta}_{\mathcal{M}_{\beta_0}}) \geq \sup_{\mathcal{M}\in\Omega_-} V^{DR}(\hat{\theta}_{\mathcal{M}}) + 3\xi, \tag{F.3}$$

for sufficiently large n. Since $\kappa_n = o(n)$, for sufficiently large n, (F.3) further implies

$$\begin{aligned} &nV^{DR}(\hat{\theta}_{\mathcal{M}_{\beta_0}}) - \kappa_n\|\hat{\beta}_{\mathcal{M}_{\beta_0}}\|_0 \\ - &\sup_{\mathcal{M}\in\Omega_-}\left\{nV^{DR}(\hat{\theta}_{\mathcal{M}}) - \kappa_n\|\hat{\beta}_{\mathcal{M}}\|_0\right\} \geq 2n\xi. \end{aligned} \tag{F.4}$$

Observe that $V(\theta, \alpha^*, \eta^*) = V^{DR}(\theta)$. Under the events defined in (A7)(i), it follows from (A9)(iii) that V is uniformly continuous and hence

$$\sup_{\theta}\left|V(\theta, \hat{\alpha}, \hat{\eta}) - V^{DR}(\theta)\right| = o(1).$$

Combining this together with (F.4) implies

$$\begin{aligned} & nV(\hat{\theta}_{\mathcal{M}_{\beta_0}}, \hat{\alpha}, \hat{\eta}) - \kappa_n \|\hat{\beta}_{\mathcal{M}_{\beta_0}}\|_0 \\ - & \sup_{\mathcal{M} \in \Omega_-} \left\{ nV(\hat{\theta}_{\mathcal{M}}, \hat{\alpha}, \hat{\eta}) - \kappa_n \|\hat{\beta}_{\mathcal{M}}\|_0 \right\} \geq n\xi, \end{aligned} \tag{F.5}$$

for sufficiently large n. In view of (F.5), the event defined in (F.1) holds when

$$\sup_{\theta} \sup_{\substack{\|\hat{\alpha}-\alpha^*\|_2 \leq \epsilon \\ \|\hat{\eta}-\eta^*\|_2 \leq \epsilon}} \left| \widehat{V}(\theta, \alpha, \eta) - V(\theta, \alpha, \eta) \right| \geq \frac{\xi}{2},$$

for some small $\epsilon > 0$. Therefore, it suffices to show

$$\Pr \left(\sup_{\theta} \sup_{\substack{\|\hat{\alpha}-\alpha^*\|_2 \leq \epsilon \\ \|\hat{\eta}-\eta^*\|_2 \leq \epsilon}} \left| \widehat{V}(\theta, \alpha, \eta) - V(\theta, \alpha, \eta) \right| \geq \frac{\xi}{2} \right) \to 0. \tag{F.6}$$

We now show the class of functions

$$\begin{aligned} f(o) \quad = \quad & \left\{ \frac{a\mathbb{I}(x^T\beta > -c)}{\pi(x, \alpha)} + \frac{(1-a)\mathbb{I}(x^T\beta \leq -c)}{1 - \pi(x, \alpha)} \right\} y \\ - \quad & \left\{ \frac{a\mathbb{I}(x^T\beta > -c)}{\pi(x, \alpha)} + \frac{(1-a)\mathbb{I}(x^T\beta \leq -c)}{1 - \pi(x, \alpha)} - 1 \right\} h(x, \eta), \end{aligned}$$

indexed by c, β, α, η belongs to the VC type class. That means, there exist some measurable envelope function F and positive constants K, $1 \leq v < \infty$ such that

$$\sup_{Q} N(\varepsilon \|F\|_{Q,2}, \mathcal{F}, L_2(Q)) \leq (K/\varepsilon)^v, \forall 0 < \varepsilon \leq 1, \tag{F.7}$$

where $\mathcal{F} = \{f(o) : c \in \mathbb{R}, \beta \in \mathbb{R}^p, \alpha \in \mathbb{R}^{q_1}, \eta \in \mathbb{R}^{q_2}\}$, $N(\cdot, \cdot, \cdot)$ stands for the entropy function (cf. Definition 2.2.3, van der Vaart and Wellner, 1996). The supremum in (F.7) is taken over all discrete measure Q such that $0 < QF^2 < \infty$, and $L_2(Q)$ is the norm on $\mathcal{F}$ defined as $\|f\|_{Q,2} = (\int |f|^2 dQ)^{1/2}$.

To prove this, we first show the class of functions

$$f_1(o) = \frac{a\mathbb{I}(x^T\beta > -c)}{\pi(x, \alpha)} y$$

indexed by c, β, α belongs to the VC type class with the envelope function

$$F_1(o) = \frac{\sqrt{2|y|^2 + 2}}{c_1^2}.$$

Define

$$\mathcal{G}_1 = \{ay\mathbb{I}(x^T\beta > -c) : c \in \mathbb{R}, \beta \in \mathbb{R}^p\}, \quad \mathcal{G}_2 = \{\pi(x,\alpha) : \alpha \in \mathbb{R}^{q_1}\}.$$

Class of functions $\mathcal{G}_1$ has finite VC index with the envelope function $G_1(o) = |y|$. By Assumption (A8), $\mathcal{G}_2$ also has VC index. Besides, $\mathcal{G}_2$ is uniformly bounded by 1. Consider function

$$\phi(x_1, x_2) = \frac{x_1}{x_2}.$$

Note that the class of function $\mathcal{F}_1 = \{f_1(o) : c \in \mathbb{R}, \beta \in B_{\mathcal{M}_0}, \alpha \in A_{\mathcal{M}_1}\}$ can be represented as $\{\phi(g_1, g_2) : g_1 \in \mathcal{G}_1, g_2 \in \mathcal{G}_2\}$. Besides, for any $g_1, g_3 \in \mathcal{G}_1$, $g_2, g_4 \in \mathcal{G}_2$, we have

$$\begin{aligned} |\phi(g_1,g_2) - \phi(g_3,g_4)|^2 &\le \frac{|g_1g_4 - g_2g_3|^2}{|g_2g_4|^2} \le \frac{1}{c_1^4}|g_1g_4 - g_2g_3|^2 \\ &\le \frac{2}{c_1^4}|g_1 - g_3|^2 + \frac{2}{c_1^4}|g_2 - g_4|^2, \end{aligned} \tag{F.8}$$

by Assumption (A3)(i). It follows from (F.8) and Lemma A.6 in Chernozhukov, Chetverikov and Kato (2014) that

$$\sup_Q N(\varepsilon\|F_1\|_2, \mathcal{F}_1, L_2(Q)) \le \sup_Q N(\varepsilon\|G_1\|_2, \mathcal{G}_1, L_2(Q)) \sup_Q N(\varepsilon, \mathcal{G}_2, L_2(Q)).$$

The above entropy is bounded by $(K_0/\varepsilon)^{v_0}$ for some constants K_0 and $1 \le v_0 < \infty$. This shows $\mathcal{F}_1$ belongs to the VC type class with VC index bounded by v_0. Similarly one can show the following classes of functions

$$\begin{aligned} f_2(o) &= \frac{(1-a)\mathbb{I}(x^T\beta \le -c)}{1-\pi(x,\alpha)}y, \\ f_3(o) &= \frac{a\mathbb{I}(x^T\beta > -c)}{\pi(x,\alpha)}h(x,\eta), \\ f_4(o) &= \frac{(1-a)\mathbb{I}(x^T\beta \le -c)}{1-\pi(x,\alpha)}h(x,\eta), \end{aligned}$$

belong to the VC type class. Then repeated applications of Lemma H.3 imply that $\mathcal{F}$ belongs to the VC type class.

Therefore, similar to (9.15), we can show

$$\mathrm{E}\left(\sup_\theta \sup_{\substack{\|\hat\alpha-\alpha^*\|_2\le\epsilon \\ \|\hat\eta-\eta^*\|_2\le\epsilon}} \left|\widehat{V}(\theta,\alpha,\eta) - V(\theta,\alpha,\eta)\right|\right) \to 0.$$

Assertion (F.6) thus follows. This proves (F.1).

F.1.2. *Overfitted model space.* Similar to Lemma 7.1, we can show that for any overfitted model space $\mathcal{M}$, $\tilde{\theta}_{\mathcal{M}}$ satisfies

$$\tilde{\theta}_{\mathcal{M}} = \theta_0 + O_p(n^{-1/3}). \tag{F.9}$$

The class of functions

$$\begin{aligned} &\left\{\{\mathbb{I}(x^T\beta > -c) - \mathbb{I}(x^T\beta_0 > -c_0)\}\left(\frac{a}{\pi_\alpha(x)} + \frac{(1-a)}{1-\pi_\alpha(x)}\right)\right\} y \\ - &\left\{\{\mathbb{I}(x^T\beta > -c) - \mathbb{I}(x^T\beta_0 > -c_0)\}\left(\frac{a}{\pi_\alpha(x)} + \frac{(1-a)}{1-\pi_\alpha(x)}\right) - 1\right\} h_\eta(x), \end{aligned}$$

indexed by $\{\|\theta - \theta_0\|_2 = O\{\max(R_n, n^{-1/3})\}, \|\alpha - \alpha^*\|_2 = O(n^{-1/2}), \|\eta - \eta^*\|_2 = O(n^{-1/2})\}$, belongs to the VC type class with finite VC index. Similar to (9.23), we can show its envelope function Ψ_V satisfies

$$\mathrm{E}|\Psi_V|^2 = O\{R_n, n^{-1/3}\}.$$

Therefore, similar to (9.24), we can show that conditional on $\hat{\alpha}$ and $\hat{\eta}$,

$$\begin{aligned} &\mathrm{E}\left(\sup_{\substack{\|\theta-\theta_0\|_2 = O\{\max(R_n, n^{-1/3})\} \\ \|\alpha-\alpha^*\|_2 = O(n^{-1/2}), \|\eta-\eta^*\|_2 = O(n^{-1/2})}} n\,|\widehat{m}_V(\theta,\alpha,\eta) - m_V(\theta,\alpha,\eta)|\right) \\ = \;& O\left(\sqrt{n\max(R_n, n^{-1/3})}\right), \end{aligned}$$

where $\widehat{m}_V(\theta,\alpha,\eta) = \widehat{V}(\theta,\alpha,\eta) - \widehat{V}(\theta_0,\alpha,\eta)$, $m_V(\theta,\alpha,\eta) = V(\theta,\alpha,\eta) - V(\theta_0,\alpha,\eta)$.

By Markov's inequality and Assumption (A7)(i), this means

$$\begin{aligned} &\sup_{\|\theta-\theta_0\|_2 = O\{\max(R_n, n^{-1/3})\}} n\,|m_V(\theta,\hat{\alpha},\hat{\eta}) - m_V(\theta_0,\hat{\alpha},\hat{\eta})| \\ = \;& O_p\left(\sqrt{n\max(R_n, n^{-1/3})}\right). \end{aligned} \tag{F.10}$$

We now show that for any θ that satisfies $\|\theta - \theta_0\|_2 = O\{\max(R_n, n^{-1/3})\}$,

$$\begin{aligned} &|V(\theta,\hat{\alpha},\hat{\eta}) - V(\theta_0,\hat{\alpha},\hat{\eta}) - V(\theta,\alpha^*,\theta^*) + V(\theta_0,\alpha^*,\theta^*)| \\ = \;& O_p\{\max(R_n^2, n^{-2/3})\}. \end{aligned} \tag{F.11}$$

Let $\bar{\zeta} = (\theta^T, \alpha^T, \eta^T)^T$ and $\bar{\zeta}_0 = (\theta_0^T, (\alpha^*)^T, (\eta^*)^T)^T$. It follows from Assumptions (A9)(iii) that

$$\begin{aligned} V(\theta, \alpha, \eta) &= V(\theta_0, \alpha^*, \eta^*) + \frac{\partial V}{\partial \theta_0}(\theta - \theta_0) + \frac{\partial V}{\partial \alpha^*}(\alpha - \alpha^*) + \frac{\partial V}{\partial \eta^*}(\eta - \eta^*) \\ &+ \frac{1}{2}(\bar{\zeta} - \bar{\zeta}_0)^T \Delta_2 V(\bar{\zeta}_0)(\bar{\zeta} - \bar{\zeta}_0) + o(\|\bar{\zeta} - \bar{\zeta}_0\|_2^2). \end{aligned} \tag{F.12}$$

Specifically, take $\theta = \theta_0$ in (F.12), we obtain

$$\begin{aligned} V(\theta_0, \alpha, \eta) &= V(\theta_0, \alpha^*, \eta^*) + \frac{\partial V}{\partial \alpha^*}(\alpha - \alpha^*) + \frac{\partial V}{\partial \eta^*}(\eta - \eta^*) \\ &+ O(\|\alpha - \alpha^*\|_2^2 + \|\eta - \eta^*\|_2^2), \end{aligned} \tag{F.13}$$

Besides, set $\alpha = \alpha^*$ and $\eta = \eta^*$ in (F.12), we have

$$V(\theta, \alpha^*, \eta^*) = V(\theta_0, \alpha^*, \eta^*) + \frac{\partial V}{\partial \theta_0}(\theta - \theta_0) + O(\|\theta - \theta_0\|_2^2). \tag{F.14}$$

The linear term $\partial V/\partial \theta_0$ in (F.12) and (F.14) is equal to 0 since θ_0 maximizes $V^{DR}(\theta) = V(\theta, \alpha^*, \eta^*)$. Combining (F.14) together with (F.12), (F.13) and Assumption (A7)(i), we obtain (F.11).

Under Assumption (A9)(iii), we have for any θ that satisfies $\|\theta - \theta_0\|_2 = O\{\max(R_n, n^{-1/3})\}$,

$$|V(\theta_0, \alpha^*, \eta^*) - V(\theta, \alpha^*, \eta^*)| = O\{\max(R_n^2, n^{-2/3})\}.$$

This together with (F.11) implies that

$$|V(\theta, \hat{\alpha}, \hat{\eta}) - V(\theta_0, \hat{\alpha}, \hat{\eta})| = O_p\{\max(R_n^2, n^{-2/3})\}, \tag{F.15}$$

for any θ such that $\|\theta - \theta_0\|_2 = O\{\max(R_n, n^{-1/3})\}$.

Observe that $\widehat{V}^{DR}(\theta) = \widehat{V}(\theta, \hat{\alpha}, \hat{\eta})$ and $V^{DR}(\theta) = V(\theta, \alpha^*, \theta^*)$. Combining (F.10) with (F.15), we have

$$\begin{aligned} \sup_{\|\theta - \theta_0\|_2 = O\{\max(R_n, n^{-1/3})\}} & n \left| \widehat{V}^{DR}(\theta) - V^{DR}(\theta) - \widehat{V}^{DR}(\theta_0) + V^{DR}(\theta_0) \right| \\ & = O_p\left(nR_n^2 + \sqrt{nR_n} + n^{1/3} \right). \end{aligned}$$

Under the given condition, we have $\kappa_n \gg \max(nR_n^2, \sqrt{nR_n}, n^{1/3})$. Under Assumption (A4) and (F.9), this further implies

$$\sup_{\mathcal{M} \in \Omega_+} n \left| \widehat{V}^{DR}(\hat{\theta}_{\mathcal{M}}) - V^{DR}(\hat{\theta}_M) - \widehat{V}^{DR}(\hat{\theta}_{\mathcal{M}_{\beta_0}}) + V^{DR}(\hat{\theta}_{\mathcal{M}_{\beta_0}}) \right| \ll \kappa_n, \tag{F.16}$$

with probability tending to 1. Recall that $\mathrm{VIC}^{DR}(\theta) = n\widehat{V}^{DR}(\theta) - \kappa_n\|\beta\|_0$ for $\theta = (c, \beta^T)^T$. For any $\mathcal{M} \in \Omega_+$, we have $\kappa_n\|\hat{\beta}_{\mathcal{M}}\|_0 - \kappa_n\|\hat{\beta}_{\mathcal{M}_{\beta_0}}\|_0 \geq \kappa_n$. Under the event defined in (F.16), this implies (F.2). The proof is hence completed.

F.2. Consistency of CIC^{DR}. Similar to the proof of Theorem 3.4, it suffices to show with probability tending to 1,

$$\mathrm{CIC}^{DR}(\hat{\beta}_{\mathcal{M}_{\beta_0}}) > \sup_{\mathcal{M}\in\Omega_-} \mathrm{CIC}^{DR}(\hat{\beta}_{\mathcal{M}}), \tag{F.17}$$

$$\mathrm{CIC}^{DR}(\hat{\beta}_{\mathcal{M}_{\beta_0}}) > \sup_{\mathcal{M}\in\Omega_+} \{n\widehat{C}^{DR}(\tilde{\beta}_{\mathcal{M}}) - \kappa_n\|\hat{\beta}_{\mathcal{M}}\|_0\}, \tag{F.18}$$

where $\tilde{\beta}_{\mathcal{M}}$ denotes the empirical maximizer of $\widehat{C}^{DR}$ on the restricted model space with $\|\tilde{\beta}_{\mathcal{M}}\|_2 = \|\beta_0\|_2$. (F.17) can be proven using similar arguments in Section C.2. In the following, we focus on (F.18).

Using similar arguments in the proof of Theorem 4 in Fan et al. (2017), we can show that for any overfitted model $\mathcal{M}$,

$$\tilde{\beta}_{\mathcal{M}} = \beta_0 + O_p(n^{-1/2}). \tag{F.19}$$

The class of functions

$$\left\{ \frac{\{A_i - \pi(X_i,\alpha)\}\{Y_i - h(X_i,\eta)\}A_j}{\pi(X_i,\alpha)\{1-\pi(X_i,\alpha)\}\pi(X_j,\alpha)} - \frac{\{A_j - \pi(X_j,\alpha)\}\{Y_j - h(X_j,\eta)\}A_i}{\pi(X_j,\alpha)\{1-\pi(X_j,\alpha)\}\pi(X_i,\alpha)} \right\} \mathbb{I}(X_i^T\beta > X_j^T\beta),$$

belongs to the VC type class. Let $\tilde{\beta}_{\mathcal{M}_{\beta_0}} = \hat{\beta}_{\mathcal{M}_{\beta_0}}$. The maximal inequality for degenerate U-process implies for any model $\mathcal{M} \in \Omega_+ \cup \{\mathcal{M}_{\beta_0}\}$,

$$\widehat{C}^{DR}(\tilde{\beta}_{\mathcal{M}}) = \frac{2}{n}\sum_{i=1}^{n} g(O_i, \tilde{\beta}_{\mathcal{M}}, \hat{\alpha}, \hat{\eta}) - C(\tilde{\beta}_{\mathcal{M}}, \hat{\alpha}, \hat{\eta}) + O_p\left(\frac{1}{n}\right). \tag{F.20}$$

Besides, it follows from Assumption (A10)(ii) and (F.19) that for any $\mathcal{M} \in \Omega_+ \cup \{\mathcal{M}_{\beta_0}\}$,

$$\begin{aligned} C^{DR}(\tilde{\beta}_{\mathcal{M}}) &= C^{DR}(\beta_0) + O_p\left((R_n^{(1)})^2\right) + O_p\left(\frac{1}{n}\right) \\ &= C^{DR}(\beta_0) + O_p\left((R_n^{(1)})^2\right), \end{aligned} \tag{F.21}$$

where the last equation is due to that $R_n^{(1)} \geq n^{-1/2}$. Since $\kappa_n \gg n(R_n^{(1)})^2$, it follows from (F.21) that with probability tending to 1,

$$n|C^{DR}(\tilde{\beta}_{\mathcal{M}}) - C^{DR}(\tilde{\beta}_{\mathcal{M}_{\beta_0}})| \ll \kappa_n. \tag{F.22}$$

Assume for now we can show

$$\left|\sum_i g(O_i,\tilde{\beta}_{\mathcal{M}},\hat{\alpha},\hat{\eta})-\sum_i g(O_i,\tilde{\beta}_{\mathcal{M}_{\beta_0}},\hat{\alpha},\hat{\eta})\right|\ll\kappa_n, \tag{F.23}$$

with probability tending to 1. Combining (F.22), (F.23) together with (F.20) suggests that for any $\mathcal{M}\in\Omega_+$,

$$n|\widehat{C}^{DR}(\tilde{\beta}_{\mathcal{M}})-\widehat{C}^{DR}(\hat{\beta}_{\mathcal{M}_{\beta_0}})|\ll\kappa_n. \tag{F.24}$$

By the definition of CIC^{DR}, (F.24) implies (F.18). Therefore, it remains to show (F.23).

Let $\hat{\zeta}_{\mathcal{M}}=(\tilde{\beta}_{\mathcal{M}}^T,\hat{\alpha}^T,\hat{\eta}^T)^T$ and $\zeta_0=\{\beta_0^T,(\alpha^*)^T,(\eta^*)^T\}^T$. Under Assumption (A10)(iii), a second order Taylor expansion around ζ_0 implies

$$\begin{aligned}\sum_i g(O_i,\hat{\zeta}_{\mathcal{M}}) &= \sum_i g(O_i,\zeta_0)+\sum_i\frac{\partial g(O_i,\zeta_0)}{\partial\zeta_0}(\hat{\zeta}_{\mathcal{M}}-\zeta_0)\\ &+ \frac{1}{2}(\hat{\zeta}_{\mathcal{M}}-\zeta_0)^T\sum_i\Delta_2 g(O_i,\zeta_M^*)(\hat{\zeta}_{\mathcal{M}}-\zeta_0),\end{aligned} \tag{F.25}$$

for some $\zeta_{\mathcal{M}}^*$ lying on the line segment between ζ_0 and $\hat{\zeta}_{\mathcal{M}}$. Besides, it follows from Assumption (A10)(iv) that

$$\left\|\sum_i\Delta_2 g(O_i,\zeta_{\mathcal{M}}^*)-\sum_i\Delta_2 g(O_i,\zeta_0)\right\|_2\le\sum_i K(O_i)\|\zeta_{\mathcal{M}}^*-\zeta_0\|_2. \tag{F.26}$$

Note that function $K(O_0)$ is integrable, and each element in the Hessian matrix $\Delta_2 g(O_i,\zeta_0)$ is integrable. Combining (F.25) with (F.26) suggests

$$\begin{aligned}\sum_i g(O_i,\hat{\zeta}_{\mathcal{M}}) &= \sum_i g(O_i,\zeta_0)+\sum_i\frac{\partial g(O_i,\zeta_0)}{\partial\zeta_0}(\hat{\zeta}_{\mathcal{M}}-\zeta_0)\\ &+ O_p(n\|\hat{\zeta}_{\mathcal{M}}-\zeta_0\|_2^2).\end{aligned} \tag{F.27}$$

Observe $\sum_i\partial_j g(O_i,\zeta_0)$ is the summation of mean zero i.i.d random variable and hence we have $\sum_i\partial_j g(O_i,\zeta_0)=O_p(\sqrt{n})$. Under Assumptions (A4) and (A7), this together with (F.27) implies

$$\begin{aligned}\sum_i g(O_i,\hat{\zeta}_{\mathcal{M}}) &= \sum_i g(O_i,\zeta_0)+O_p(n^{1/2}(R_n^{(1)}+n^{-1/2}))+O_p(n(R_n^{(1)}+n^{-1/2})^2)\\ &= \sum_i g(O_i,\zeta_0)+O_p(n(R_n^{(1)})^2).\end{aligned}$$

Under the Assumption that $\kappa_n\gg n(R_n^{(1)})^2$, this proves (F.23). The proof is hence completed.

APPENDIX G: PROOF OF THEOREM 11.1

G.1. Consistency of VIC$^{(1)}$. It suffices to show that with probability tending to 1,

$$\text{VIC}^{(1)}(\hat{\theta}_{1,\mathcal{M}_1^V}) > \sup_{\mathcal{M}_1 \in \Omega_-^V} \text{VIC}^{(1)}(\hat{\theta}_{1,\mathcal{M}_1}), \tag{G.1}$$

$$\text{VIC}^{(1)}(\hat{\theta}_{1,\mathcal{M}_1^V}) > \sup_{\mathcal{M}_1 \in \Omega_+^V} \{n\widehat{V}^{(1)}(\tilde{\theta}_{1,\mathcal{M}_1}) - \kappa_n^{(1)} \|\hat{\beta}_{1,\mathcal{M}_1}\|\}, \tag{G.2}$$

where $\tilde{\theta}_{1,\mathcal{M}_1}$ denotes the empirical maximizer of $\widehat{V}^{(1)}$ on the restricted model space with $\|\tilde{\theta}_{1,\mathcal{M}_1}\|_2 = \|\theta_1^V\|_2$, and

$$\Omega_-^V = \{\mathcal{M}_1 \in \Omega_1 : \mathcal{M}_1^V \not\subset \mathcal{M}_1\} \quad \text{and} \quad \Omega_+^V = \{\mathcal{M}_1 \in \Omega_1 : \mathcal{M}_1^V \subsetneq \mathcal{M}_1\}.$$

We focus on (G.2). (G.1) can be similarly proven.

Under Condition (C7)(i), we have that for any $(\theta_1^T, \theta_2^T)^T \to \{(\theta_1^V)^T, \theta_{0,2}^T\}^T$,

$$\begin{aligned} V(\theta_1, \theta_2) &= V(\theta_1^V, \theta_{0,2}) + \frac{1}{2}(\theta_1 - \theta_1^V)^T \frac{\partial^2 V(\theta_1^V, \theta_{0,2})}{\partial \theta_1^V \partial (\theta_1^V)^T} (\theta_1 - \theta_1^V)^T \\ &+ \frac{1}{2}(\theta_2 - \theta_{0,2})^T \frac{\partial^2 V(\theta_1^V, \theta_{0,2})}{\partial \theta_{0,2} \partial \theta_{0,2}^T} (\theta_2 - \theta_{0,2})^T + (\theta_1 - \theta_1^V)^T \frac{\partial^2 V(\theta_1^V, \theta_{0,2})}{\partial \theta_1 \partial \theta_2^T} (\theta_2 - \theta_{0,2})^T \\ &+ o(\|\theta_1 - \theta_1^V\|_2^2) + o(\|\theta_2 - \theta_{0,2}\|_2^2). \end{aligned} \tag{G.3}$$

Note that the first-order terms vanish in (G.3) since $\{(\theta_1^V)^T, (\theta_{0,2})^T\}^T$ maximizes $V(\theta_1, \theta_2)$.

Set $\theta_2 = \theta_{0,2}$ in (G.3), we obtain

$$V(\theta_1, \theta_{0,2}) - V(\theta_1^V, \theta_{0,2}) = \frac{1}{2}(\theta_1 - \theta_1^V)^T \frac{\partial^2 V(\theta_1^V, \theta_{0,2})}{\partial (\theta_1^V)^2} (\theta_1 - \theta_1^V)^T + o(\|\theta_1 - \theta_1^V\|_2^2).$$

Combining this together with (G.3), we obtain that for any $(\theta_1^T, \theta_2^T)^T \to \{(\theta_1^V)^T, (\theta_{0,2})^T\}^T$,

$$\begin{aligned} V(\theta_1, \theta_2) &= V(\theta_1, \theta_{0,2}) + \frac{1}{2}(\theta_2 - \theta_{0,2})^T \frac{\partial^2 V(\theta_1^V, \theta_{0,2})}{\partial (\theta_{0,2})^2} (\theta_2 - \theta_{0,2})^T \\ &+ (\theta_1 - \theta_1^V)^T \frac{\partial^2 V(\theta_1^V, \theta_{0,2})}{\partial \theta_1 \partial \theta_2^T} (\theta_2 - \theta_{0,2})^T + o(\|\theta_1 - \theta_1^V\|_2^2) + o(\|\theta_2 - \theta_{0,2}\|_2^2). \end{aligned}$$

Conditional on the event that $\widehat{\mathcal{M}}_2^V = \mathcal{M}_{0,2}$, we have for any $\theta_1 \to \theta_1^V$,

$$V(\theta_1, \hat{\theta}_{2,\widehat{\mathcal{M}}_2^V}) = V(\theta_1, \theta_{0,2}) + O_p(R_{n,2}^2) + O_p(R_{n,2})\|\theta_1 - \theta_1^V\|_2 + o(\|\theta_1 - \theta_1^V\|_2^2). \tag{G.4}$$

Besides, under Condition (C5), using similar arguments in (9.24), we can show that

$$\sup_{\theta_1} \sup_{\|\theta_2-\theta_{0,2}\|_2=O(R_{n,2})} n\left|\widehat{V}(\theta_1,\theta_2)-\widehat{V}(\theta_1,\theta_{0,2})-V(\theta_1,\theta_2)+V(\theta_1,\theta_{0,2})\right| = O_p(\sqrt{nR_{n,2}}),$$

where

$$\widehat{V}(\theta_1,\theta_2)=\frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{I}\{A_i^{(1)}=d_1^{\theta_1}(X_i^{(1)}),A_i^{(2)}=d_2^{\theta_2}(\overline{X}_i^{(2)})\}}{\Pr(A_i^{(1)}=d_1^{\theta_1}(X_i^{(1)})|X_i^{(1)})\Pr(A_i^{(2)}=d_2^{\theta_2}(\overline{X}_i^{(2)})|\overline{X}_i^{(2)})}Y_i,$$

where

$$d_1^{\theta_1}(x^{(1)})=\mathbb{I}(\beta_1^T x^{(1)}>-c_1) \quad \text{and} \quad d_2^{\theta_2}(\bar{x}^{(2)})=\mathbb{I}(\beta_2^T \bar{x}^{(2)}>-c_2).$$

Conditional on the event that $\widehat{\mathcal{M}}_2^V=\mathcal{M}_{0,2}$, it follows from Condition (C5) that

(G.5)
$$\sup_{\theta_1} n\left|\widehat{V}^{(1)}(\theta_1)-\widehat{V}(\theta_1,\theta_{0,2})-V(\theta_1,\theta_{2,\widehat{\mathcal{M}}_2^V})+V(\theta_1,\theta_{0,2})\right| = O_p(\sqrt{nR_{n,2}}).$$

Based on (G.4) and (G.5), using similar arguments in the proof of Lemma 7.1, we can show that $\tilde{\theta}_{1,\mathcal{M}_1}$ satisfies $\tilde{\theta}_{1,\mathcal{M}_1}=\theta_1^V+O_p(R_{n,2})+O_p(n^{-1/4}R_{n,2}^{1/4})+O_p(n^{-1/3})=\theta_1^V+O_p(R_{n,2})+O_p(n^{-1/3})$. (G.2) can thus be proven using similar arguments in the proof of Theorem 3.3.

G.2. Consistency of CIC$^{(1)}$. It suffices to show with probability tending to 1,

$$\text{(G.6)} \qquad \text{CIC}^{(1)}(\hat{\beta}_{1,\mathcal{M}_1^C})>\sup_{\mathcal{M}_1\in\Omega_-^C}\text{CIC}^{(1)}(\hat{\beta}_{1,\mathcal{M}_1}),$$

$$\text{(G.7)} \qquad \text{CIC}^{(1)}(\hat{\beta}_{1,\mathcal{M}_1^C})>\sup_{\mathcal{M}_1\in\Omega_+^C}\{n\widehat{C}^{(1)}(\tilde{\beta}_{1,\mathcal{M}_1})-\kappa_n^{(1)}\|\hat{\beta}_{1,\mathcal{M}_1}\|\},$$

where $\tilde{\beta}_{1,\mathcal{M}_1}$ denotes the empirical maximizer of $\widehat{C}^{(1)}$ on the restricted model space with $\|\tilde{\beta}_{1,\mathcal{M}_1}\|_2=\|\beta_1^C\|_2$, and

$$\Omega_-^C=\{\mathcal{M}_1\in\Omega_1:\mathcal{M}_1^C\not\subset\mathcal{M}_1\}, \qquad \Omega_+^C=\{\mathcal{M}_1\in\Omega_1:\mathcal{M}_1^C\subsetneq\mathcal{M}_1\}.$$

In the following, we focus on proving (G.7). (G.6) can be similarly proven.

For $\theta_2 = (c_2, \beta_2^T)^T$, define

$$Y_i^{(1)}(\theta_2) = \left\{ \frac{A_i^{(2)}}{\pi_0^{(2)}(\overline{X}_i^{(2)})} \mathbb{I}(\beta_2^T \overline{X}_i^{(2)} > -c_2) + \frac{1 - A_i^{(2)}}{1 - \pi_0^{(2)}(\overline{X}_i^{(2)})} \mathbb{I}(\beta_2^T \overline{X}_i^{(2)} \le -c_2) \right\} Y_i,$$

$$y^{(1)}(\theta_2) = \left\{ \frac{a^{(2)}}{\pi_0^{(2)}(\bar{x}^{(2)})} \mathbb{I}(\beta_2^T \bar{x}^{(2)} > -c_2) + \frac{1 - a^{(2)}}{1 - \pi_0^{(2)}(\bar{x}^{(2)})} \mathbb{I}(\beta_2^T \bar{x}^{(2)} \le -c_2) \right\} y.$$

For any $o = (x^{(1)}, a^{(1)}, x^{(2)}, a^{(2)}, y)$, define $g(o, \beta_1, \theta_2)$ as

$$\frac{1}{2}\mathrm{E}\left\{ \frac{\{A_0^{(1)} - \pi_0^{(1)}(X_0^{(1)})\} Y_0^{(1)}(\theta_2)}{\pi_0^{(1)}(X_0^{(1)})\{1 - \pi_0^{(1)}(X_0^{(1)})\}} - \frac{\{a^{(1)} - \pi^{(1)}(x^{(1)})\} y^{(1)}(\theta_2)}{\pi^{(1)}(x^{(1)})\{1 - \pi^{(1)}(x^{(1)})\}} \right\} \mathbb{I}(\beta_1^T X_0^{(1)} > \beta_1^T x^{(1)})$$

$$+\frac{1}{2}\mathrm{E}\left\{ \frac{\{a^{(1)} - \pi^{(1)}(x^{(1)})\} y^{(1)}(\theta_2)}{\pi_0^{(1)}(x^{(1)})\{1 - \pi_0^{(1)}(x^{(1)})\}} - \frac{\{A_0^{(1)} - \pi_0^{(1)}(X_0^{(1)})\} Y_0^{(1)}(\theta_2)}{\pi_0^{(1)}(X_0^{(1)})\{1 - \pi_0^{(1)}(X_0^{(1)})\}} \right\} \mathbb{I}(\beta_1^T x^{(1)} > \beta_1^T X_0^{(1)}).$$

It follows from the maximal inequality for U-process that

$$\sup_{\theta_2} \sup_{\beta_1} \left| \widehat{C}(\beta_1, \theta_2) - \frac{2}{n} \sum_{i=1}^{n} g(O_i, \beta_1, \theta_2) + C(\beta_1, \theta_2) \right| = O_p\left(\frac{1}{n}\right),$$

where

$$\widehat{C}(\beta_1, \theta_2) = \frac{1}{n(n-1)} \sum_{i \neq j} \left\{ \omega_i^{(2)}(\theta_2) - \omega_j^{(2)}(\theta_2) \right\} \mathbb{I}(\beta_2^T \overline{X}_i^{(2)} > \beta_2^T \overline{X}_j^{(2)}),$$

$$C(\beta_1, \theta_2) = \mathrm{E}\widehat{C}(\beta_1, \theta_2), \quad \omega_i^{(2)}(\theta_2) = \left\{ \frac{A_i^{(2)}}{\pi_0^{(2)}(\overline{X}_i^{(2)})} - \frac{1 - A_i^{(2)}}{1 - \pi_0^{(2)}(\overline{X}_i^{(2)})} \right\} Y_i^{(1)}(\theta_2).$$

This further implies that

$$\sup_{\beta_1} \left| \widehat{C}^{(1)}(\beta_1) - \frac{2}{n} \sum_{i=1}^{n} g(O_i, \beta_1, \hat{\theta}_{2,\widehat{\mathcal{M}}_2^C}) + C(\beta_1, \hat{\theta}_{2,\widehat{\mathcal{M}}_2^C}) \right| = O_p\left(\frac{1}{n}\right). \tag{G.8}$$

In the following, we focus on proving

$$\mathrm{E} \sup_{\substack{\|\beta_1 - \beta_1^C\|_2 \le \varepsilon_n \\ \|\theta_2 - \theta_{0,2}\|_2 = O(R_{n,2})}} \left| \sum_{i=1}^{n} \{ g(O_i, \beta_1, \theta_2) - g(O_i, \beta_1, \theta_{0,2}) - g(O_i, \beta_1^C, \theta_2) + g(O_i, \beta_1^C, \theta_{0,2}) \} \right|$$

$$= O\left(\sqrt{n R_{n,2}} + n R_{n,2} \varepsilon_n \right), \tag{G.9}$$

for any sequence $\varepsilon_n \to 0$. When this holds, it follows from Jensen's inequality that

$$\sup_{\substack{\|\beta_1-\beta_1^C\|_2\le\varepsilon_n \\ \|\theta_2-\theta_{0,2}\|_2=O(R_{n,2})}} \left|C(\beta_1,\theta_2)-C(\beta_1,\theta_2^C)-C(\beta_1^C,\theta_2)+C(\beta_1^C,\theta_2^C)\right|$$

$$= O\left(\sqrt{nR_{n,2}}+nR_{n,2}\varepsilon_n\right). \tag{G.10}$$

Conditional on $\widehat{\mathcal{M}}_2^C = \mathcal{M}_2^C$, it follows from (G.9), (G.10) and Condition (C5) that

$$\sup_{\|\beta_1-\beta_1^C\|_2\le\varepsilon_n} \left|\sum_{i=1}^n \{g(O_i,\beta_1,\hat\theta_{2,\widehat{\mathcal{M}}_2^C})-g(O_i,\beta_1,\theta_{0,2})-g(O_i,\beta_1^C,\hat\theta_{2,\widehat{\mathcal{M}}_2^C})+g(O_i,\beta_1^C,\theta_{0,2})\}\right|$$

$$= O_p\left(\sqrt{nR_{n,2}}+nR_{n,2}\varepsilon_n\right), \tag{G.11}$$

and

$$\sup_{\|\beta_1-\beta_1^C\|_2\le\varepsilon_n} n\left|C^{(1)}(\beta_1)-C(\beta_1,\theta_{0,2})-C^{(1)}(\beta_1^C)+C(\beta_1^C,\theta_{0,2})\right|$$

$$= O_p\left(\sqrt{nR_{n,2}}+nR_{n,2}\varepsilon_n\right). \tag{G.12}$$

Based on (G.8),(G.11) and (G.12), using similar arguments in the proof of Lemma 7.1, we can show that $\tilde\beta_{1,\mathcal{M}_1}$ converges at a rate of $O_p(R_{n,2}) + O_p(n^{-1/4}R_{n,2}^{1/4}) = O_p(R_{n,2}) + O_p(n^{-1/3})$ for any $\mathcal{M}_1 \in \Omega_+^C$. Similar to the proof of Theorem 3.4, we can show (G.7) holds. It remains to show (G.9).

For any $o = (x^{(1)}, a^{(1)}, x^{(2)}, a^{(2)}, y)$, observe that $\{g(o,\beta_1,\theta_2) - g(o,\beta_1^C,\theta_2) - g(o,\beta_1,\theta_{0,2}) + g(o,\beta_1^C,\theta_{0,2})\}$ can be decomposed as

$$\sum_{j=1}^4 \{g_j(o,\beta_1,\theta_2) - g_j(o,\beta_1^C,\theta_2)\},$$

where

$$g_1(o,\beta_1,\theta_2) = \frac{1}{2}\mathrm{E}\frac{\{A_0^{(1)}-\pi_0^{(1)}(X_0^{(1)})\}\{Y_0^{(1)}(\theta_2)-Y_0^{(1)}\}}{\pi_0^{(1)}(X_0^{(1)})\{1-\pi_0^{(1)}(X_0^{(1)})\}}\mathbb{I}(\beta_1^T X_0^{(1)} > \beta_1^T x^{(1)}),$$

$$g_2(o,\beta_1,\theta_2) = -\frac{1}{2}\frac{\{a^{(1)}-\pi_0^{(1)}(x^{(1)})\}\{y^{(1)}(\theta_2)-y^{(1)}\}}{\pi_0^{(1)}(x^{(1)})\{1-\pi_0^{(1)}(x^{(1)})\}}\phi_1(x^{(1)},\beta_1),$$

$$g_3(o,\beta_1,\theta_2) = \frac{1}{2}\frac{\{a^{(1)}-\pi_0^{(1)}(x^{(1)})\}\{y^{(1)}(\theta_2)-y^{(1)}\}}{\pi_0^{(1)}(x^{(1)})\{1-\pi_0^{(1)}(x^{(1)})\}}\phi_2(x^{(1)},\beta_1),$$

$$g_4(o,\beta_1,\theta_2) = -\frac{1}{2}\mathrm{E}\frac{\{A_0^{(1)}-\pi_0^{(1)}(X_0^{(1)})\}\{Y_0^{(1)}(\theta_2)-Y_0^{(1)}\}}{\pi_0^{(1)}(X_0^{(1)})\{1-\pi_0^{(1)}(X_0^{(1)})\}}\mathbb{I}(\beta_1^T x^{(1)} > \beta_1^T X_0^{(1)}).$$

Therefore, it suffices to show for $j = 1, 2, 3, 4$,

$$\mathrm{E} \sup_{\substack{\|\beta_1-\beta_1^C\|_2\le\varepsilon_n \\ \|\theta_2-\theta_{0,2}\|_2=O(R_{n,2})}} \left|\sum_{i=1}^n \{g_j(O_i,\beta_1,\theta_2) - g_j(O_i,\beta_1^C,\theta_2)\}\right| = O\left(\sqrt{nR_{n,2}} + nR_{n,2}\varepsilon_n\right),$$

We first show

$$\mathrm{E} \sup_{\substack{\|\beta_1-\beta_1^C\|_2\le\varepsilon_n \\ \|\theta_2-\theta_{0,2}\|_2=O(R_{n,2})}} \left|\sum_{i=1}^n \{g_1(O_i,\beta_1,\theta_2) - g_1(O_i,\beta_1^C,\theta_2)\}\right| = O\left(\sqrt{nR_{n,2}} + nR_{n,2}\varepsilon_n\right). \tag{G.13}$$

Note that

$$\begin{aligned}
&\left|\mathrm{E}\{g_1(O_i,\beta_1,\theta_2) - g_1(O_i,\beta_1^C,\theta_2)\}\right| \\
=\ & \frac{1}{2}\left|\mathrm{E}\left\{\frac{\{A_0^{(1)} - \pi_0^{(1)}(X_0^{(1)})\}\{Y_0^{(1)}(\theta_2) - Y_0^{(1)}\}}{\pi_0^{(1)}(X_0^{(1)})\{1-\pi_0^{(1)}(X_0^{(1)})\}}\{\phi_2(X_0^{(1)},\beta_1) - \phi_2(X_0^{(1)},\beta_1^C)\}\right\}\right| \\
\le\ & \frac{\varepsilon_n}{2}\mathrm{E}\left|\frac{\{A_0^{(1)} - \pi_0^{(1)}(X_0^{(1)})\}G^{(2)}(\beta_{0,2}^T X_0^{(2)} + c_{0,2})}{\pi_0^{(1)}(X_0^{(1)})\{1-\pi_0^{(1)}(X_0^{(1)})\}\psi_2^{-1}(X_0^{(2)})}\right| |\mathbb{I}(\beta_2^T X_0^{(2)} > -c_2) - \mathbb{I}(\beta_{0,2}^T X_0^{(2)} > -c_{0,2})| \\
\le\ & \frac{\varepsilon_n}{2c_1(1-c_1)}\mathrm{E}|G^{(2)}(\beta_{0,2}^T X_0^{(2)} + c_{0,2})||\mathbb{I}(\beta_2^T X_0^{(2)} > -c_2) - \mathbb{I}(\beta_{0,2}^T X_0^{(2)} > -c_{0,2})|\psi_2(X_0^{(2)}) \\
\le\ & O(\varepsilon_n)\sqrt{\mathrm{E}\{G^{(2)}(\beta_{0,2}^T X_0^{(2)} + c_{0,2})\}^2|\mathbb{I}(\beta_2^T X_0^{(2)} > -c_2) - \mathbb{I}(\beta_{0,2}^T X_0^{(2)} > -c_{0,2})|}\sqrt{\mathrm{E}\psi_2^2(X_0^{(2)})} \\
\le\ & O(\varepsilon_n)\sqrt{\mathrm{E}\{G^{(2)}(\beta_{0,2}^T X_0^{(2)} + c_{0,2})\}^2|\mathbb{I}(\beta_2^T X_0^{(2)} > -c_2) - \mathbb{I}(\beta_{0,2}^T X_0^{(2)} > -c_{0,2})|}.
\end{aligned}$$

where the first inequality is due to Condition (C9)(iv), the second inequality is due to Condition (C4), the third inequality is due to Cauchy-Schwarz inequality and the last inequality is due to Condition (C9)(iv).

Now we claim for any $\theta_2 = (c_2, \beta_2^T)^T$ such that $\|\theta_2 - \theta_{0,2}\|_2 = O(R_{n,2})$,

$$\mathrm{E}\{G^{(2)}(\beta_{0,2}^T X_0^{(2)})\}^2|\mathbb{I}(\beta_2^T X_0^{(2)} > -c_2) - \mathbb{I}(\beta_{0,2}^T X_0^{(2)} > -c_{0,2})| = O(R_{n,2}^2). \tag{G.14}$$

This implies

$$\left|\mathrm{E}\{g_1(O_i,\beta_1,\theta_2) - g_1(O_i,\beta_1^C,\theta_2)\}\right| = O(\varepsilon_n R_{n,2}). \tag{G.15}$$

Note that

$$\mathbb{I}(\beta_2^T X_0^{(2)} > -c_2) = \mathbb{I}(G^{(2)}(\beta_2^T X_0^{(2)} + c_2) > 0),$$

and

$$\mathbb{I}(\beta_{0,2}^T X_0^{(2)} > -c_{0,2}) = \mathbb{I}(G^{(2)}(\beta_{0,2}^T X_0^{(2)} + c_{0,2}) > 0).$$

To prove (G.14), it suffices to show

$$\mathrm{E}\{G^{(2)}(\beta_{0,2}^T X_0^{(2)} + c_{0,2})\}^2 |\mathbb{I}(G^{(2)}(\beta_2^T X_0^{(2)} + c_2) > 0) - \mathbb{I}(G^{(2)}(\beta_{0,2}^T X_0^{(2)} + c_{0,2}) > 0)| = O(R_{n,2}^2).$$

Note that when $\mathbb{I}(G^{(2)}(\beta_2^T X_0^{(2)} + c_2) > 0) \neq \mathbb{I}(G^{(2)}(\beta_{0,2}^T X_0^{(2)} + c_{0,2}) > 0)$, we have

$$|G^{(2)}(\beta_2^T X_0^{(2)} + c_2) - G^{(2)}(\beta_{0,2}^T X_0^{(2)} + c_{0,2})| \geq |G^{(2)}(\beta_{0,2}^T X_0^{(2)} + c_{0,2})|.$$

By Markov's inequality, This implies

$$\begin{aligned}\mathrm{E}\{G^{(2)}(\beta_{0,2}^T X_0^{(2)})\}^2 &|\mathbb{I}(G^{(2)}(\beta_2^T X_0^{(2)} + c_2) > 0) - \mathbb{I}(G^{(2)}(\beta_{0,2}^T X_0^{(2)} + c_{0,2}) > 0)| \\ &\leq \mathrm{E}|G^{(2)}(\beta_2^T X_0^{(2)} + c_2) - G^{(2)}(\beta_{0,2}^T X_0^{(2)} + c_{0,2})|^2 = O(R_{n,2}^2),\end{aligned}$$

for any $\theta_2 = (c_2, \beta_2^T)^T$ such that $\|\theta_2 - \theta_{0,2}\|_2 = O(R_{n,2})$, where the last equality is due to Condition (C9)(i). This proves (G.14). To show (G.13), in view of (G.15), it suffices to show

$$\begin{aligned}\mathrm{E} \sup_{\substack{\|\beta_1 - \beta_1^C\|_2 \leq \varepsilon_n \\ \|\theta_2 - \theta_{0,2}\|_2 = O(R_{n,2})}} &\left| \sum_{i=1}^n \{g_1(O_i, \beta_1, \theta_2) - g_1(O_i, \beta_1^C, \theta_2) \right. \\ &\left. - \mathrm{E}g_1(O_i, \beta_1, \theta_2) + \mathrm{E}g_1(O_i, \beta_1^C, \theta_2)\} \right| = O\left(\sqrt{nR_{n,2}}\right),\end{aligned}$$

or

$$\mathrm{E} \sup_{\beta_1} \sup_{\|\theta_2 - \theta_{0,2}\|_2 = O(R_{n,2})} \left| \sum_{i=1}^n \{g_1(O_i, \beta_1, \theta_2) - \mathrm{E}g_1(O_i, \beta_1, \theta_2)\} \right| = O\left(\sqrt{nR_{n,2}}\right).$$

This can be proven in a similar manner as (C.9).

Similarly, we can show that for $j = 2, 3, 4$,

$$\begin{aligned}\mathrm{E} \sup_{\substack{\|\beta_1 - \beta_1^C\|_2 \leq \varepsilon_n \\ \|\theta_2 - \theta_{0,2}\|_2 = O(R_{n,2})}} &\left| \sum_{i=1}^n \{g_j(O_i, \beta_1, \theta_2) - g_j(O_i, \beta_1^C, \theta_2)\} \right| \\ &= O\left(\sqrt{nR_{n,2}} + nR_{n,2}\varepsilon_n\right).\end{aligned}$$

This proves (G.9). The proof is thus completed.

APPENDIX H: TECHNICAL LEMMAS

LEMMA H.1. *For any random variable X, if $\omega = \|X\|_{\psi_1} < \infty$, then for any integer $p \geq 1$, $E|X|^p \leq p!\omega^p$.*

Proof: It follows by the definition of Orlicz norm that

$$1 + \frac{\mathrm{E}|X|^p}{\omega^p} \leq \mathrm{E} \exp\left(\frac{|X|}{\omega}\right) \leq 2.$$

The assertion thus follows.

LEMMA H.2 (Bernstein inequality). *Let $X_1, \ldots, X_n$ be independent mean zero random variables, if $\omega = \max_i \|X_i\|_{\psi_1} < \infty$, there exists some constant c such that for all $t > 0$,*

$$Pr\left(|\sum_{i=1}^{n} X_i| > t\right) \leq 2\exp\left(-c\min\left(\frac{t^2}{n\omega^2}, \frac{t}{\omega}\right)\right).$$

Proof: See Theorem 3.1 in Klartag and Mendelson (2005).

LEMMA H.3. *Let $\mathcal{F}$ and $\mathcal{G}$ be class of measurable functions $S \to \mathbb{R}$, to which measurable envelopes F and G are attached, respectively. Assume there exists some constant K and $v \geq 1$ such that*

$$\sup_Q N(\varepsilon\|F\|_{Q,2}, \mathcal{F}, L_2(Q)) \leq (K/\varepsilon)^v, \quad \sup_Q N(\varepsilon\|G\|_{Q,2}, \mathcal{G}, L_2(Q)) \leq (K/\varepsilon)^v,$$

for all $0 < \varepsilon \leq 1$. Denoted by $\mathcal{F} + \mathcal{G}$ the pointwise sum of $\mathcal{F}$ and $\mathcal{G}$. Then

$$\sup_Q N(\varepsilon\|\sqrt{2F^2 + 2G^2}\|_{Q,2}, \mathcal{F}, L_2(Q)) \leq (K/\varepsilon)^{2v}.$$

Proof: For any $f_1, f_2 \in \mathcal{F}$ and $g_1, g_2 \in \mathcal{G}$, by Cauchy-Schwarz inequality, we have

$$|f_1 + g_1 - f_2 - g_2|^2 \leq 2|f_1 - f_2|_2^2 + 2|g_1 - g_2|_2^2.$$

The assertion then follows from application of Lemma A.6 in Chernozhukov, Chetverikov and Kato (2014).

LEMMA H.4. *Let $X_1, \ldots, X_n$ be i.i.d random variables with values in a measurable space $(\mathcal{S}, \mathcal{B})$, and let $\mathcal{F}$ be a countable class of measurable functions $f : \mathcal{S} \to \mathbb{R}$. Assume $Ef(X_i) = 0$ for all f, and $\omega = \|\sup_{f\in\mathcal{F}} |f(X_i)|\|_{\psi_1} <$*

∞. *Then for all* $0 < \eta < 1$ *and* $\delta > 0$, *there exists some constant* $C = C(\eta, \delta)$ *such that for all* $t \geq 0$,

$$Pr\left(\sup_{f\in\mathcal{F}} |\sum_{i=1}^{n} f(X_i)| \geq (1+\eta)E\|\sum_{i=1}^{n} f(X_i)\|_{\mathcal{F}} + t\right)$$
$$\leq \exp\left(-\frac{t^2}{2(1+\delta)n\sigma^2}\right) + 3\exp\left(-\frac{t}{C\omega\log(n)}\right),$$

where $\sigma^2 = \sup_{f\in\mathcal{F}} Ef^2(X_i)$.

Proof: It follows from Lemma 2.2.2 in van der Vaart and Wellner (1996) that

$$\|\max_i \sup_{f\in\mathcal{F}} |f(X_i)|\|_{\psi_1} \leq K\log(n)\max_i \|\sup_{f\in\mathcal{F}} |f(X_i)|\|_{\psi_1},$$

for some constant K. The assertion follows by Theorem 4 in Adamczak (2008).

LEMMA H.5. *Assume* $X_1, \ldots, X_n$ *are i.i.d random variables. Assume function* f *is symmetric and satisfies* $Ef(X_i, x) = Ef(x, X_i) = 0$, $f(x, x) = 0, \forall x$, $\sup_f |f(x, y)| \leq F, \forall x, y$. *Define the following degenerate U-process,*

$$Z = \sup_{f\in\mathcal{F}} |\sum_{i\neq j} f(X_i, X_j)|$$

Let $\varepsilon_1, \ldots, \varepsilon_n$ *be i.i.d Rademacher random variables independent of* $\{X_1, \ldots, X_n\}$, *and introduce the random variables:*

$$Z_\varepsilon = \sup_{f\in\mathcal{F}} \left|\sum_{i,j} \varepsilon_i\varepsilon_j f(X_i, X_j)\right|,$$
$$U_\varepsilon = \sup_{f\in\mathcal{F}} \sup_{\alpha:\|\alpha\|_2\leq 1} \sum_{i,j} \varepsilon_i\alpha_j f(X_i, X_j),$$
$$M_\varepsilon = \sup_{f\in\mathcal{F}} \sup_{k=1,\ldots,n} \left|\sum_i \varepsilon_i f(X_i, X_k)\right|.$$

Then there exists some constants $C > 0$ *such that for all* n *and* $t > 0$,

$$Pr(Z > CEZ_\varepsilon + t)$$
$$\leq \exp\left(-\frac{1}{C}\min\left(\frac{t^2}{(EU_\varepsilon)^2}, \frac{t}{EM_\varepsilon}, \frac{t}{nF}, \left(\frac{t}{F\sqrt{n}}\right)^{2/3}, \sqrt{\frac{t}{F}}\right)\right). \tag{H.1}$$

Proof: It follows from Theorem 11 in Clémençon, Lugosi and Vayatis (2008) that (H.1) is bounded by

$$\text{(H.2)}\ \exp\left(-\frac{1}{C'}\min\left(\frac{t^2}{(\mathrm{E}U_\varepsilon)^2}, \frac{t}{nF+\mathrm{E}M_\varepsilon}, \left(\frac{t}{F\sqrt{n}}\right)^{2/3}, \sqrt{\frac{t}{F}}\right)\right),$$

for some constant C'. Note that (H.2) is smaller than

$$\exp\left(-\frac{1}{C'}\min\left(\frac{t^2}{(\mathrm{E}U_\varepsilon)^2}, \frac{t}{2nF}, \frac{t}{2\mathrm{E}M_\varepsilon}, \left(\frac{t}{F\sqrt{n}}\right)^{2/3}, \sqrt{\frac{t}{F}}\right)\right).$$

Lemma H.5 thus holds by setting $C = 2C'$.

APPENDIX I: ADDITIONAL SIMULATIONS

I.1. Nonregular cases. In this section, we examine the finite sample performance of our proposed information criteria under the settings where the contrast function has a positive probability of equal to 0. We consider a two-stage study (see Section 11.1) and the observations $\{X_i^{(1)}, A_i^{(1)}, X_i^{(2)}, A_i^{(2)}, Y_i\}$, $i = 1, \ldots, n$ is a sample from the following model:

$$\text{(I.1)}\quad Y_i = 1 + A_i^{(1)}\beta_*^T X_i^{(1)} + A_i^{(2)}\beta_*^T X_i^{(1)} + \varepsilon_i \text{ and } X_i^{(2)} = A_i^{(1)} + \nu_i,$$

where $A_i^{(1)}, A_i^{(2)} \sim$ Bernoulli(0.5), $\varepsilon_i \sim N(0, 0.25)$, $\nu_i \sim N_p(0, 1)$ and $X_i^{(1)}$ is a vector consisting of p independent left-censored Gaussian random variables. Specifically, for any $j \in \{1, \ldots, p\}$, the distribution function for the j-th element of X_0 is given by

$$\Pr(X_0^{(1),j} = 0) = \frac{1}{2} \quad \text{and} \quad \Pr(X_0^{(1),j} \le z) = \frac{1}{2} + \frac{1}{\sqrt{2\pi}}\int_0^z \exp\left(-\frac{z^2}{2}\right) dz, \quad \forall z > 0.$$

Variables $A_i^{(1)}, A_i^{(2)}, X_i^{(1)}, \varepsilon_i$ and ν_i are all independent. We set $p = 8$,

$$\beta_* = (1, -1, \underbrace{0, \ldots, 0}_{6})^T,$$

and consider two choices of n, $n = 300$ and $n = 500$.

It can be seen from (I.1) that the response Y_i doesn't depend on the intermediate outcome $X_i^{(2)}$. Moreover, the contrast function at each stage takes the form:

$$\tau^{(2)}(\overline{X}_0^{(2)}) = \beta_*^T X_0^{(1)} \text{ and } \tau^{(1)}(X_0^{(1)}) = \beta_*^T X_0^{(1)}.$$

As shown in Section B.2.3, we have

$$\Pr\{\tau^{(2)}(\overline{X}_0^{(2)}) = 0\} = \Pr\{\tau^{(1)}(X_0^{(1)}) = 0\} = \Pr(X_0^{(1),1} = X_0^{(1),2}) = \frac{1}{4}.$$

We use backward induction (See Section 11.1 for details) to select important variables involved in the optimal dynamic treatment regime. At each stage, we apply concordance-assisted learning to estimate the parameters in the contrast function on the restricted model space. For $l = 1, \ldots, L$, $j = 1, 2$, denoted by $\widehat{\mathcal{M}}_j^{(l)}$ the set of variables selected by our information criteria at the j-th stage, in the l-th simulation. Let $\hat{d}_j^{(l)}$ denote the corresponding estimated optimal treatment regime based on the set of selected variables in $\widehat{\mathcal{M}}_j^{(l)}$. Define $\mathcal{M}_{0,j}$ to be the true support of important variables involved in the j-th stage contrast function.

Table 2: Simulation results (%, standard deviations in parenthesis)

n		CIC	VIC		CIC	VIC
300	TP_1	73.00(4.44)	37.00(4.83)	TP_2	87.00(3.36)	51.00(5.00)
	FN_1	0.50(0.50)	26.50(2.51)	FN_2	0.00(0.00)	23.50(2.51)
	FP_1	18.00(0.45)	12.33(0.81)	FP_2	1.63(0.42)	0.38(0.21)
	ER_1	6.84(0.88)	13.58(1.40)	ER_2	6.48(0.90)	10.89(1.14)
	VR_1	98.57(0.12)	95.63(0.28)	VR_2	99.31(0.01)	97.60(0.23)
500	TP_1	66.00(4.74)	53.00(4.99)	TP_2	86.00(3.67)	71.00(4.54)
	FN_1	0.00(0.00)	15.50(2.32)	FN_2	0.00(0.00)	11.50(2.11)
	FP_1	18.33(0.50)	15.50(0.68)	FP_2	1.75(0.44)	0.75(0.30)
	ER_1	5.60(0.74)	10.74(1.15)	ER_2	4.15(0.64)	8.18(1.09)
	VR_1	99.12(0.07)	97.55(0.22)	VR_2	99.66(0.04)	98.74(0.18)

In Table 2, we report the false positives rate,

$$\mathrm{FP}_j = \frac{1}{L}\sum_{l=1}^{L} \frac{|\mathcal{M}_{0,j}^c \cap \widehat{\mathcal{M}}_{0,j}^{(l)}|}{|\mathcal{M}_{0,j}^c|},$$

the false negatives rate,

$$\mathrm{FN}_j = \frac{1}{L}\sum_{l=1}^{L} \frac{|\mathcal{M}_{0,j} \cap (\widehat{\mathcal{M}}_{0,j}^{(l)})^c|}{|\mathcal{M}_{0,j}|},$$

the percentage of selecting the true models,

$$\mathrm{TP}_j = \frac{1}{L}\sum_{l=1}^{L} \mathbb{I}(\mathcal{M}_{0,j} = \widehat{\mathcal{M}}_0^{(l)}),$$

the average error rate of the estimated optimal treatment regime at each stage,

$$\text{ER}_1 = \frac{1}{L}\sum_{l=1}^{L}\frac{\text{E}|\hat{d}_1^{(l)}(X_0^{(1)}) - d_1^{opt}(X_0^{(1)})|}{\Pr\{d_1^{opt}(X_0^{(1)}) = 0\}}, \tag{I.2}$$

$$\text{ER}_2 = \frac{1}{L}\sum_{l=1}^{L}\frac{\text{E}|\hat{d}_2^{(l)}(\overline{X}_0^{(2)}) - d_2^{opt}(\overline{X}_0^{(2)})|}{\Pr\{d_2^{opt}(X_0^{(2)}) = 0\}}, \tag{I.3}$$

the average value ratio of the estimated optimal dynamic treatment regime and the average value ratio of estimated optimal treatment regime at the second stage,

$$\text{VR}_1 = \frac{1}{L}\sum_{l=1}^{L}\frac{\text{E}Y_0^*(\hat{d}_1^{(l)}, \hat{d}_2^{(l)})}{\text{E}Y_0^*(d_1^{opt}, d_2^{opt})}, \tag{I.4}$$

$$\text{VR}_2 = \frac{1}{L}\sum_{l=1}^{L}\frac{\text{E}Y_0^*(A_0^{(1)}, \hat{d}_2^{(l)})}{\text{E}Y_0^*(A_0^{(1)}, d_2^{opt})} \tag{I.5}$$

where L is the total number of simulations. We set $L = 100$. The expectation in (I.2)-(I.5) are approximated based on 1000 Monte Carlo samples.

It can see from Table 2 that in the nonregular cases, CIC also achieves better numerical performance when compared to VIC. For example, TP's of CIC are much larger than those of VIC especially with small sample size. However, we notice that increasing the sample size does not improve the model selection results for CIC. For examples, TP's of CIC with $n = 500$ are even smaller than those with $n = 300$. This suggests our information criteria might not be consistent in the noregular cases.

I.2. Other choice of $\boldsymbol{\beta_0}$. In this section, we examine the numerical performance of our information criteria under the settings when the nonzero components of β_0 are not of the same order of magnitude. Specifically, we consider the following contrast function

$$\tau(x) = 2.5x^1 + x^2 + 0.4x^3. \tag{I.6}$$

Hence, we have $\beta_0 = (2.5, 1, 0.4, \underbrace{0, \ldots, 0}_{p-3})$. The corresponding optimal treatment regime is given by

$$d^{opt}(x) = \mathbb{I}(2.5x^1 + x^2 + 0.4x^3 > 0).$$

In Section I.2.1, we design a fixed-p setting and estimate the OTR by CAL. In Section I.2.2, we design a high-dimensional setting and estimate the OTR by PAL.

I.2.1. *Concordance-assisted learning.* We set $n = 400$, $p = 8$ and generate the response from the following model:

$$Y_i = h_0(X_i^1, X_i^3) + A_i\tau(X_i) + \varepsilon_i,$$

where $A_i \overset{i.i.d}{\sim}$ Bernoulli(0.5), $X_i \overset{i.i.d}{\sim} N_p(0, I_p)$, $\varepsilon_i \overset{i.i.d}{\sim} N(0, 0.5^2)$, and the function $\tau(\cdot)$ is given in (I.6). We consider two choices of h_0, corresponding to $h_0(x^1, x^3) = 1 + x^1 - x^3$ and $h_0(x^1, x^3) = 1 + x^1x^3$, respectively.

To apply CIC^{DR} and VIC^{DR} for model selection, we compute the parameter $\hat{\beta}_{\mathcal{M}}$ for each candidate model $\mathcal{M}$ by maximizing $\widehat{C}^{DR}$ on the restricted model space. The detailed implementation can be found in Section 6.1. We set $\kappa_n = \log(n)$ in CIC^{DR} and $\kappa_n = n^{-1/3}\log(\log(n))$ in VIC^{DR}.

The number of simulation replications is set to be $L = 100$. For any $l = 1, \ldots, L$, we denoted by $\widehat{\mathcal{M}}^{(l)}$ the set of important variables selected by our information criteria in the l-th simulation. In Table 4, we report the percentages that the first three important variables being selected by our information criteria, i.e,

$$\frac{1}{L}\sum_{l=1}^{L}\mathbb{I}(j \in \widehat{\mathcal{M}}^{(l)}), \quad j = 1, 2, 3,$$

and the average ratio of value (VR) of the estimated OTR.

Table 3: Model selection results and the average value ratio of the estimated OTR

	$\sum_{l=1}^{L}\frac{\mathbb{I}(1\in\widehat{\mathcal{M}}^{(l)})}{L}$	$\sum_{l=1}^{L}\frac{\mathbb{I}(2\in\widehat{\mathcal{M}}^{(l)})}{L}$	$\sum_{l=1}^{L}\frac{\mathbb{I}(3\in\widehat{\mathcal{M}}^{(l)})}{L}$	VR
$h_0(x) = 1 + x^1 - x^3$				
CIC^{DR}	100.00(0.00)	100.00(0.00)	61.00(4.88)	99.62(0.03)
VIC^{DR}	100.00(0.00)	92.00(2.71)	1.00(0.99)	99.39(0.06)
$h_0(x) = 1 + x^1x^3$				
CIC^{DR}	100.00(0.00)	100.00(0.00)	56.00(4.96)	99.43(0.04)
VIC^{DR}	100.00(0.00)	71.00(4.54)	27.00(4.44)	99.32(0.04)

We make the following observations. First, variables with larger coefficients (in absolute value) are more likely to be selected by our information criteria. For example, CIC^{DR} and VIC^{DR} always select the first variable. For the third variable however, its probabilities of being selected are at most 61% in all cases. Second, CIC^{DR} is more capable of selecting variables with small coefficients than VIC^{DR}. VIC^{DR} fails to identify the third variable.

Moreover, despite that variables with small coefficients might be missed, the average value ratio of the estimated OTR are close to 1 in all cases.

I.2.2. *Penalized A-learning.* We set $p = 1000$ and generate the response from the following model:

$$Y_i = 1 + X_i^1 - X_i^3 + A_i\tau(X_i) + \varepsilon_i,$$

where $X_i \overset{i.i.d}{\sim} N_p(0, I_p)$, $A_i \overset{i.i.d}{\sim}$ Bernoulli(0.5), $\varepsilon_i \overset{i.i.d}{\sim} N(0, 0.5^2)$ and the function $\tau(\cdot)$ is given in (I.6).

We estimate the OTR by PAL. The detailed implementation can be found in Section 6.2. Tuning parameters in the estimating procedure are selected by CIC^{DR} and VIC^{DR}. As in Section 6.2, we set

$$\kappa_n = \log(p)\log_{10}(n)\log(\log_{10}(n)),$$

in CIC^{DR} and set

$$\kappa_n = n^{1/3}\log^{2/3}(p)\log(\log(n))/\kappa,$$

in VIC^{DR} where κ is a constant from a set $\{3, 4, 5\}$. For each κ, we denote the corresponding information criterion as VIC_κ^{DR}.

We set the sample size $n = 2000$ and the number of simulation replications $L = 100$. In Table 4, we report the percentages that the first three important variables being selected by our information criteria and the average ratio of value (VR) of the estimated OTR. Findings are very similar to Section I.2.1.

Table 4: Model selection results and the average value ratio of the estimated OTR

	$\sum_{l=1}^{L} \frac{\mathbb{I}(1\in\widehat{\mathcal{M}}^{(l)})}{L}$	$\sum_{l=1}^{L} \frac{\mathbb{I}(2\in\widehat{\mathcal{M}}^{(l)})}{L}$	$\sum_{l=1}^{L} \frac{\mathbb{I}(3\in\widehat{\mathcal{M}}^{(l)})}{L}$	VR
CIC^{DR}	100.00(0.00)	100.00(0.00)	90.00(3.00)	99.94(0.02)
VIC_3^{DR}	100.00(0.00)	100.00(0.00)	20.00(4.00)	99.54(0.02)
VIC_4^{DR}	100.00(0.00)	100.00(0.00)	46.00(4.98)	99.71(0.03)
VIC_5^{DR}	100.00(0.00)	100.00(0.00)	65.00(4.77)	99.80(0.03)

APPENDIX J: ADDITIONAL DETAILS REGARDING THE CROSS-VALIDATION PROCEDURE IN SECTION 5.1.2

Given a candidate set of tuning parameters $\{\kappa_{n,j}\}_{1\le j\le J}$, one can randomly divide $\mathcal{I}_0 \equiv \{1, \ldots, n\}$ into non-overlapping sets $\{\mathcal{I}_k\}_{k=1,\ldots,K}$ that satisfy $\mathcal{I}_0 = \cup_{k=1}^K \mathcal{I}_k$. Let $\mathcal{I}_{(-k)} = \mathcal{I}_0 - \mathcal{I}_k$, for $k = 1, \ldots, K$. For an arbitrary set $\mathcal{I}$, let $|\mathcal{I}|$ denote the cardinality of $\mathcal{I}$.

For any k and j, we choose $\kappa_{|\mathcal{I}_{(-k)}|,j}$ as the tuning parameter and use VIC^{DR} (or CIC^{DR}) based on observations in $\mathcal{I}_{(-k)}$ to select the model $\widehat{\mathcal{M}}^{(-k,j)}$. Then we compute the estimated value (or concordance) function under the decision rule $\mathbb{I}(x^T\hat{\beta}_{\widehat{\mathcal{M}}^{(-k,j)}} > -\hat{c}_{\widehat{\mathcal{M}}^{(-k,j)}})$ using observations in $\mathcal{I}_k$. Finally, we choose j_0 among $\{1,\ldots,J\}$ that maximizes the estimated value (or concordance) aggregated over $k = 1,\ldots,K$ and set $\kappa_n = \kappa_{n,j_0}$.

REFERENCES

Adamczak, R. (2008). A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. *Electron. J. Probab.* **13** no. 34, 1000–1034.

Chernozhukov, V., Chetverikov, D. and Kato, K. (2014). Gaussian approximation of suprema of empirical processes. *Ann. Statist.* **42** 1564–1597.

Clémençon, S., Lugosi, G. and Vayatis, N. (2008). Ranking and empirical minimization of U-statistics. *Ann. Statist.* **36** 844–874.

Fan, C., Lu, W., Song, R. and Zhou, Y. (2017). Concordance-assisted learning for estimating optimal individualized treatment regimes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 1565–1582.

Klartag, B. and Mendelson, S. (2005). Empirical processes and random projections. *J. Funct. Anal.* **225** 229–245.

Ledoux, M. and Talagrand, M. (2011). *Probability in Banach spaces. Classics in Mathematics*. Springer-Verlag, Berlin Isoperimetry and processes, Reprint of the 1991 edition.

Luedtke, A. R. and van der Laan, M. J. (2016). Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Ann. Statist.* **44** 713–742.

Nolan, D. and Pollard, D. (1987). U-processes: rates of convergence. *Ann. Statist.* **15** 780–799.

Qian, M. and Murphy, S. A. (2011). Performance guarantees for individualized treatment rules. *Ann. Statist.* **39** 1180–1210.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes. Springer Series in Statistics*. Springer-Verlag, New York With applications to statistics.

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

Chengchun Shi
Department of Statistics,
North Carolina State University,
Raleigh NC, U.S.A.
E-mail: cshi4@ncsu.edu

Rui Song
Department of Statistics,
North Carolina State University,
Raleigh NC, U.S.A.
E-mail: rsong@ncsu.edu

Wenbin Lu
Department of Statistics,
North Carolina State University,
Raleigh NC, U.S.A.
E-mail: wlu4@ncsu.edu