SEMIPARAMETRIC ESTIMATION OF PROBABILISTIC INDEX MODELS: EFFICIENCY AND BIAS

Karel Vermeulen, Jan De Neve, Gustavo Amorim,

Olivier Thas and Stijn Vansteelandt

Ghent University, Vanderbilt University,

Hasselt University, London School of Hygiene and Tropical Medicine

1. Class of influence functions for $\boldsymbol{\beta}$

Let $\mathbf{Z}_i^T = (Y_i, \mathbf{X}_i^T)$, $i = 1, \ldots, n$, denote a sample of n independent and identically distributed (IID) random vectors, where Y_i denotes the outcome of interest associated with the p-dimensional vector of covariates $\mathbf{X}_i$. A PIM is defined by the constraint

$$\mathrm{P}\left(Y_i \preccurlyeq Y_j \mid \mathbf{X}_i, \mathbf{X}_j\right) = m(\mathbf{X}_i, \mathbf{X}_j; \boldsymbol{\beta}_0), \tag{1.1}$$

with $\mathrm{P}\left(Y_i \preccurlyeq Y_j \mid \mathbf{X}_i, \mathbf{X}_j\right) := \mathrm{P}\left(Y_i < Y_j \mid \mathbf{X}_i, \mathbf{X}_j\right) + 0.5\mathrm{P}\left(Y_i = Y_j \mid \mathbf{X}_i, \mathbf{X}_j\right)$. The function $m(\cdot)$ is a known function with range $[0, 1]$, smooth in the p-

dimensional parameter vector $\boldsymbol{\beta}$ and satisfying the antisymmetry condition $m(\mathbf{X}_i, \mathbf{X}_j; \boldsymbol{\beta}) = 1 - m(\mathbf{X}_j, \mathbf{X}_i; \boldsymbol{\beta})$. We let $\boldsymbol{\beta}_0$ denote the unknown true value of $\boldsymbol{\beta}$ that generated the data. Let $\mathcal{M}_{\text{PIM}}$ denote the semiparametric model induced by (1.1), i.e., the set of all proper joint density functions $f_{\mathbf{Z}}(\mathbf{z}; \boldsymbol{\beta}, \boldsymbol{\eta}) = f_{Y|\mathbf{X}}(y \mid \mathbf{x}; \boldsymbol{\beta}, \boldsymbol{\eta}_1) f_{\mathbf{X}}(\mathbf{x}; \boldsymbol{\eta}_2)$ satisfying the key restriction that

$$\int \{\mathrm{I}\,(y \preccurlyeq y^*) - m(\mathbf{x}, \mathbf{x}^*; \boldsymbol{\beta})\} f_{Y|\mathbf{X}}(y \mid \mathbf{x}; \boldsymbol{\beta}, \boldsymbol{\eta}_1) f_{Y|\mathbf{X}}(y^* \mid \mathbf{x}^*; \boldsymbol{\beta}, \boldsymbol{\eta}_1) dy dy^* = 0, \tag{1.2}$$

with $\boldsymbol{\beta}$ the p-dimensional parameter of interest and where $\boldsymbol{\eta} = (\boldsymbol{\eta}_1^T, \boldsymbol{\eta}_2^T)^T$ are infinite-dimensional variation independent nuisance parameters. We denote the truth as $f_0(\mathbf{z}) = f_{\mathbf{Z}}(\mathbf{z}; \boldsymbol{\beta}_0, \boldsymbol{\eta}_0) = f_{Y|\mathbf{X}}(y \mid \mathbf{x}; \boldsymbol{\beta}_0, \boldsymbol{\eta}_{10}) f_{\mathbf{X}}(\mathbf{x}; \boldsymbol{\eta}_{20})$.

We denote by $\mathcal{H}$ the Hilbert space of p-dimensional measurable random functions $\mathbf{h}(\mathbf{Z})$ of $\mathbf{Z}$ satisfying (i) $\mathrm{E}\{\mathbf{h}(\mathbf{Z})\} = \mathbf{0}$ (mean zero) and (ii) $\mathrm{E}\{\mathbf{h}^T(\mathbf{Z})\mathbf{h}(\mathbf{Z})\} < \infty$ (square integrable) equipped with the covariance inner product $\mathrm{E}\{\mathbf{h}_1^T(\mathbf{Z})\mathbf{h}_2(\mathbf{Z})\}$ for $\mathbf{h}_1(\mathbf{Z}), \mathbf{h}_2(\mathbf{Z}) \in \mathcal{H}$. The aim is to find the set of all influence functions $\boldsymbol{\varphi}(Y, \mathbf{X}; \boldsymbol{\beta}_0, \boldsymbol{\eta}_0) \in \mathcal{H}$ of RAL estimators for the p-dimensional parameter of interest $\boldsymbol{\beta}$ under model $\mathcal{M}_{\text{PIM}}$. For this, we need to find the orthogonal complement $\Lambda^{\perp}$ of the nuisance tangent space Λ of model $\mathcal{M}_{\text{PIM}}$ in $\mathcal{H}$.

Derivation of the model nuisance tangent space: Λ. The semiparametric nuisance tangent space Λ is the direct sum of the semiparametric nuisance tangent spaces Λ_1 and Λ_2 corresponding to $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ respectively: $\Lambda = \Lambda_1 \bigoplus \Lambda_2$. Because the marginal distribution $f_{\mathbf{X}}(\mathbf{x}; \boldsymbol{\eta}_2)$ of the covariates $\mathbf{X}$ is completely left unspecified, the sole restriction is that the scores for $\boldsymbol{\eta}_2$, which is a function of $\mathbf{X}$ only, must have mean zero so that $\Lambda_2 = \{\mathbf{s}_2(\mathbf{X}) \in \mathcal{H} \mid \mathrm{E}\{\mathbf{s}_2(\mathbf{X})\} = \mathbf{0}\}$. Because the set Λ_1 corresponds to all scores of the possible conditional distribution of Y given $\mathbf{X}$, all $\mathbf{s}_1(Y, \mathbf{X}) \in \Lambda_1$ must satisfy $\mathrm{E}\{\mathbf{s}_1(Y, \mathbf{X})|\mathbf{X}\} = \mathbf{0}$. The sole other restriction is implied by (1.2) following which all $\mathbf{s}_1(Y, \mathbf{X}) \in \Lambda_1$ must also satisfy $\mathrm{E}[\{\mathbf{s}_1(Y, \mathbf{X}) + \mathbf{s}_1(Y^*, \mathbf{X}^*)\}\{\mathrm{I}(Y \preccurlyeq Y^*) - m(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta}_0)\}|\mathbf{X}, \mathbf{X}^*] = \mathbf{0}$ for $(Y, \mathbf{X}^T) \perp\!\!\!\perp (Y^*, \mathbf{X}^{*,T})$, where $\perp\!\!\!\perp$ denotes statistical independence. It follows that $\Lambda_1 = \{\mathbf{s}_1(Y, \mathbf{X}) \in \mathcal{H}|\mathrm{E}\{\mathbf{s}_1(Y, \mathbf{X})|\mathbf{X}\} = \mathbf{0}$ and $\mathrm{E}[\{\mathbf{s}_1(Y, \mathbf{X}) + \mathbf{s}_1(Y^*, \mathbf{X}^*)\}\{\mathrm{I}(Y \preccurlyeq Y^*) - m(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta}_0)\}|\mathbf{X}, \mathbf{X}^*] = \mathbf{0}\}$. Next, it is easy to see that $\Lambda_1 \perp \Lambda_2$ because for $\mathbf{s}_1(Y, \mathbf{X}) \in \Lambda_1$ and $\mathbf{s}_2(\mathbf{X}) \in \Lambda_2$, $\mathrm{E}\{\mathbf{s}_1^T(Y, \mathbf{X})\mathbf{s}_2(\mathbf{X})\} = \mathrm{E}[\mathrm{E}\{\mathbf{s}_1^T(Y, \mathbf{X})|\mathbf{X}\}\mathbf{s}_2(\mathbf{X})] = 0$. Now note that $\Lambda_1 = \Lambda_{1a} \cap \Lambda_{1b}$ with $\Lambda_{1a} = \{\mathbf{s}_1(Y, \mathbf{X}) \in \mathcal{H}|\mathrm{E}\{\mathbf{s}_1(Y, \mathbf{X})|\mathbf{X}\} = \mathbf{0}\}$ and $\Lambda_{1b} = \{\mathbf{s}_1(Y, \mathbf{X}) \in \mathcal{H}|\mathrm{E}[\{\mathbf{s}_1(Y, \mathbf{X}) + \mathbf{s}_1(Y^*, \mathbf{X}^*)\}\{\mathrm{I}(Y \preccurlyeq Y^*) - m(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta}_0)\}|\mathbf{X}, \mathbf{X}^*] = \mathbf{0}\}$. We find that $\Lambda = (\Lambda_{1a} \cap \Lambda_{1b}) \oplus \Lambda_2$, i.e., every $\mathbf{s}(Y, \mathbf{X}) \in \Lambda$ can be written as $\mathbf{s}_1(Y, \mathbf{X}) + \mathbf{s}_2(\mathbf{X})$ with $\mathbf{s}_1(Y, \mathbf{X}) \in \Lambda_{1a} \cap \Lambda_{1b}$ and $\mathbf{s}_2(\mathbf{X}) \in \Lambda_2$.

Lemma 1. *We have the following relations:* (i) $\Lambda_{1a} = \Lambda_2^{\perp}$, (ii) $\Lambda_2 \subset \Lambda_{1b}$ *and* (iii) $\Lambda = \Lambda_{1b}$.

Proof. (i) Take arbitrary elements $\mathbf{s}_1(Y, \mathbf{X}) \in \Lambda_{1a}$ and $\mathbf{s}_2(\mathbf{X}) \in \Lambda_2$, then $\mathrm{E}\{\mathbf{s}_1^T(Y, \mathbf{X})\mathbf{s}_2(\mathbf{X})\} = \mathrm{E}[\mathrm{E}\{\mathbf{s}_1^T(Y, \mathbf{X})|\mathbf{X}\}\mathbf{s}_2(\mathbf{X})] = 0$ so that $\Lambda_{1a} \subset \Lambda_2^{\perp}$. Now take an arbitrary $\mathbf{h}(Y, \mathbf{X}) \in \mathcal{H}$. Define $\mathbf{s}_1(Y, \mathbf{X}) = \mathbf{h}(Y, \mathbf{X}) - \mathrm{E}\{\mathbf{h}(Y, \mathbf{X})|\mathbf{X}\}$ and $\mathbf{s}_2(\mathbf{X}) = \mathrm{E}\{\mathbf{h}(Y, \mathbf{X})|\mathbf{X}\}$. It follows that $\mathbf{h}(Y, \mathbf{X}) = \mathbf{s}_1(Y, \mathbf{X}) + \mathbf{s}_2(\mathbf{X})$ with $\mathbf{s}_1(Y, \mathbf{X}) \in \Lambda_{1a}$ and $\mathbf{s}_2(\mathbf{X}) \in \Lambda_2$ and we can conclude that $\Lambda_{1a} = \Lambda_2^{\perp}$. (ii) Take an arbitrary $\mathbf{s}_2(\mathbf{X}) \in \Lambda_2$. We find that $\mathrm{E}[\{\mathbf{s}_2(\mathbf{X}) + \mathbf{s}_2(\mathbf{X}^*)\}\{\mathrm{I}(Y \preccurlyeq Y^*) - m(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta}_0)\}|\mathbf{X}, \mathbf{X}^*] = \{\mathbf{s}_2(\mathbf{X}) + \mathbf{s}_2(\mathbf{X}^*)\}[\mathrm{E}\{\mathrm{I}(Y \preccurlyeq Y^*)|\mathbf{X}, \mathbf{X}^*\} - m(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta}_0)] = \mathbf{0}$ with $(Y, \mathbf{X}^T) \perp\!\!\!\perp (Y^*, \mathbf{X}^{*,T})$ so that $\mathbf{s}_2(\mathbf{X}) \in \Lambda_{1b}$. (iii) We know that $\Lambda = (\Lambda_{1a} \cap \Lambda_{1b}) \oplus \Lambda_2$. Take any $\mathbf{s}_1(Y, \mathbf{X}) \in \Lambda_{1a} \cap \Lambda_{1b}$ and $\mathbf{s}_2(\mathbf{X}) \in \Lambda_2$. By definition, $\mathbf{s}_1(Y, \mathbf{X}) \in \Lambda_{1b}$ and by part (ii) $\mathbf{s}_2(\mathbf{X}) \in \Lambda_{1b}$ and because Λ_{1b} is a linear subspace of $\mathcal{H}$, $\mathbf{s}_1(Y, \mathbf{X}) + \mathbf{s}_2(\mathbf{X}) \in \Lambda_{1b}$ so that $\Lambda \subset \Lambda_{1b}$. Now consider an arbitrary function $\mathbf{s}(Y, \mathbf{X}) \in \Lambda_{1b}$. Because $\mathbf{s}(Y, \mathbf{X}) \in \mathcal{H}$, $\mathbf{0} = \mathrm{E}\{\mathbf{s}(Y, \mathbf{X})\} = \mathrm{E}[\mathrm{E}\{\mathbf{s}(Y, \mathbf{X})|\mathbf{X}\}]$, $\mathrm{E}\{\mathbf{s}(Y, \mathbf{X})|\mathbf{X}\} \in \Lambda_2$. From part (ii), we also know that $\mathrm{E}\{\mathbf{s}(Y, \mathbf{X})|\mathbf{X}\} \in \Lambda_{1b}$ and because Λ_{1b} is a linear space, $\mathbf{s}(Y, \mathbf{X}) - \mathrm{E}\{\mathbf{s}(Y, \mathbf{X})|\mathbf{X}\} \in \Lambda_{1b}$. Since $\mathbf{s}(Y, \mathbf{X}) - \mathrm{E}\{\mathbf{s}(Y, \mathbf{X})|\mathbf{X}\} \perp \Lambda_2$, $\mathbf{s}(Y, \mathbf{X}) - \mathrm{E}\{\mathbf{s}(Y, \mathbf{X})|\mathbf{X}\} \in \Lambda_{1a}$. It follows that $\mathbf{s}(Y, \mathbf{X}) = [\mathbf{s}(Y, \mathbf{X}) - \mathrm{E}\{\mathbf{s}(Y, \mathbf{X})|\mathbf{X}\}] + \mathrm{E}\{\mathbf{s}(Y, \mathbf{X})|\mathbf{X}\}$ with $\mathbf{s}(Y, \mathbf{X}) - \mathrm{E}\{\mathbf{s}(Y, \mathbf{X})|\mathbf{X}\} \in \Lambda_{1a} \cap \Lambda_{1b}$ and $\mathrm{E}\{\mathbf{s}(Y, \mathbf{X})|\mathbf{X}\} \in \Lambda_2$ so that $\Lambda_{1b} \subset \Lambda$. □

We conclude with the following proposition:

Proposition 1. *The semiparametric nuisance tangent space* Λ *of model* $\mathcal{M}_{\text{PIM}}$ *equals*

$$\Lambda = \{\mathbf{s}(Y, \mathbf{X}) \in \mathcal{H} \,|\, \mathrm{E}[\{\mathbf{s}(Y, \mathbf{X}) + \mathbf{s}(Y^*, \mathbf{X}^*)\}\{\mathrm{I}(Y \preccurlyeq Y^*) - m(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta}_0)\} | \mathbf{X}, \mathbf{X}^*] = \mathbf{0}\}, \tag{1.3}$$

for $(Y, \mathbf{X}^T) \perp\!\!\!\perp (Y^*, \mathbf{X}^{*,T})$.

Derivation of the orthogonal complement of the model nuisance tangent space: $\Lambda^{\perp}$. From semiparametric theory, we know that for any influence function $\boldsymbol{\varphi}(Y, \mathbf{X}; \boldsymbol{\beta}_0, \boldsymbol{\eta}_0)$ of a RAL estimator of $\boldsymbol{\beta}$, it holds that $\boldsymbol{\varphi}(Y, \mathbf{X}; \boldsymbol{\beta}_0, \boldsymbol{\eta}_0) \perp \Lambda$. Before deriving $\Lambda^{\perp}$, we prove the following lemma. Throughout, $(Y, \mathbf{X}^T) \perp\!\!\!\perp (Y^*, \mathbf{X}^{*,T})$.

Lemma 2. *Define the sets of p-dimensional functions*

$$\begin{aligned}\mathcal{A} &= \{\mathbf{a}(\mathbf{X}, \mathbf{X}^*) - \mathbf{a}(\mathbf{X}^*, \mathbf{X}) \mid \mathbf{a}(\mathbf{X}, \mathbf{X}^*) \textit{ is square integrable } \},\\ \mathcal{B} &= \{\mathbf{b}(\mathbf{X}, \mathbf{X}^*) \mid \mathbf{b}(\mathbf{X}, \mathbf{X}^*) \textit{ is square integrable and } \mathbf{b}(\mathbf{X}, \mathbf{X}^*) + \mathbf{b}(\mathbf{X}^*, \mathbf{X}) = \mathbf{0}\}.\end{aligned}$$

It holds that $\mathcal{A} = \mathcal{B}$.

Proof. (i) $\mathcal{A} \subset \mathcal{B}$. Take $\mathbf{b}^{\dagger}(\mathbf{X}, \mathbf{X}^*) := \mathbf{a}(\mathbf{X}, \mathbf{X}^*) - \mathbf{a}(\mathbf{X}^*, \mathbf{X}) \in \mathcal{A}$. It follows that $\mathbf{b}^{\dagger}(\mathbf{X}, \mathbf{X}^*) + \mathbf{b}^{\dagger}(\mathbf{X}^*, \mathbf{X}) = \mathbf{0}$ so that $\mathbf{b}^{\dagger}(\mathbf{X}, \mathbf{X}^*) \in \mathcal{B}$. (ii) $\mathcal{B} \subset \mathcal{A}$. Take $\mathbf{b}(\mathbf{X}, \mathbf{X}^*) \in \mathcal{B}$. Define $\mathbf{a}^{\dagger}(\mathbf{X}, \mathbf{X}^*) = \mathbf{b}(\mathbf{X}, \mathbf{X}^*)/2$. Because $\mathbf{b}(\mathbf{X}, \mathbf{X}^*) = \mathbf{a}^{\dagger}(\mathbf{X}, \mathbf{X}^*) - \mathbf{a}^{\dagger}(\mathbf{X}^*, \mathbf{X})$, $\mathbf{b}(\mathbf{X}, \mathbf{X}^*) \in \mathcal{A}$.

□

Proposition 2. *The orthogonal complement of the semiparametric nuisance tangent space $\Lambda^{\perp}$ of model $\mathcal{M}_{PIM}$ equals*

$$\begin{aligned}\Lambda^{\perp} = \{\mathbf{s}^{\perp}(Y, \mathbf{X}) \in \mathcal{H} \mid \mathbf{s}^{\perp}(Y, \mathbf{X}) = \mathrm{E}[\mathbf{b}(\mathbf{X}, \mathbf{X}^*)\{\mathrm{I}\,(Y \preccurlyeq Y^*) - m(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta}_0)\} \mid Y, \mathbf{X}], \quad (1.4)\\ \mathbf{b}(\mathbf{X}, \mathbf{X}^*) \in \mathcal{B}\},\end{aligned}$$

for $(Y, \mathbf{X}^T) \perp\!\!\!\perp (Y^*, \mathbf{X}^{*,T})$.

Proof. From Lemma 2 it is sufficient to show that $\Lambda^\perp$ consists of all elements $\mathbf{h}(Y, \mathbf{X}) \in \mathcal{H}$ of the form $\mathbf{s}^\perp(Y, \mathbf{X}) = \mathrm{E}[\{\mathbf{a}(\mathbf{X}, \mathbf{X}^*) - \mathbf{a}(\mathbf{X}^*, \mathbf{X})\}\{\mathrm{I}\,(Y \preccurlyeq Y^*) - m(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta}_0)\} \mid Y, \mathbf{X}]$. Take an arbitrary $\mathbf{s}(Y, \mathbf{X}) \in \Lambda$. We have that

$$
\begin{aligned}
&\mathrm{E}\{\mathbf{s}^T(Y, \mathbf{X})\mathbf{s}^\perp(Y, \mathbf{X})\} \\
=&\mathrm{E}(\mathbf{s}^T(Y, \mathbf{X})\mathrm{E}[\{\mathbf{a}(\mathbf{X}, \mathbf{X}^*) - \mathbf{a}(\mathbf{X}^*, \mathbf{X})\}\{\mathrm{I}\,(Y \preccurlyeq Y^*) - m(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta}_0)\} \mid Y, \mathbf{X}]) \\
=&\mathrm{E}(\mathbf{s}^T(Y, \mathbf{X})\mathrm{E}[\mathbf{a}(\mathbf{X}, \mathbf{X}^*)\{\mathrm{I}\,(Y \preccurlyeq Y^*) - m(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta}_0)\} \mid Y, \mathbf{X}]) \\
&+ \mathrm{E}(\mathbf{s}^T(Y, \mathbf{X})\mathrm{E}[\mathbf{a}(\mathbf{X}^*, \mathbf{X})\{\mathrm{I}\,(Y^* \preccurlyeq Y) - m(\mathbf{X}^*, \mathbf{X}; \boldsymbol{\beta}_0)\} \mid Y, \mathbf{X}]) \\
=&\mathrm{E}[\mathbf{s}^T(Y, \mathbf{X})\mathbf{a}(\mathbf{X}, \mathbf{X}^*)\{\mathrm{I}\,(Y \preccurlyeq Y^*) - m(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta}_0)\}] \\
&+ \mathrm{E}[\mathbf{s}^T(Y^*, \mathbf{X}^*)\mathbf{a}(\mathbf{X}, \mathbf{X}^*)\{\mathrm{I}\,(Y \preccurlyeq Y^*) - m(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta}_0)\}] \\
=&\mathrm{E}[\{\mathbf{s}(Y, \mathbf{X}) + \mathbf{s}(Y^*, \mathbf{X}^*)\}^T\mathbf{a}(\mathbf{X}, \mathbf{X}^*)\{\mathrm{I}\,(Y \preccurlyeq Y^*) - m(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta}_0)\}]
\end{aligned}
$$

Because $\mathbf{s}(Y, \mathbf{X}) \in \Lambda$,

$$
\begin{aligned}
&\mathrm{E}\{\mathbf{s}^T(Y, \mathbf{X})\mathbf{s}^\perp(Y, \mathbf{X})\} \\
&= \mathrm{E}\left(\mathrm{E}[\{\mathbf{s}(Y, \mathbf{X}) + \mathbf{s}(Y^*, \mathbf{X}^*)\}^T\{\mathrm{I}\,(Y \preccurlyeq Y^*) - m(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta}_0)\} \mid \mathbf{X}, \mathbf{X}^*]\mathbf{a}^T(\mathbf{X}, \mathbf{X}^*)\right) = 0,
\end{aligned}
$$

showing that indeed $\mathbf{s}^\perp(Y, \mathbf{X}) \perp \Lambda$. To show that the space $\Lambda^\perp$ indeed equals the orthogonal complement of Λ, we need to show that any arbitrary $\mathbf{h}(Y, \mathbf{X}) \in \mathcal{H}$ can be written as $\mathbf{h}(Y, \mathbf{X}) = \mathbf{s}_\mathbf{h}(Y, \mathbf{X}) + \mathbf{s}_\mathbf{h}^\perp(Y, \mathbf{X})$ for $\mathbf{s}_\mathbf{h}(Y, \mathbf{X}) \in \Lambda$ and $\mathbf{s}_\mathbf{h}^\perp(Y, \mathbf{X}) \in \Lambda^\perp$. This is equivalent to saying that for each $\mathbf{h}(Y, \mathbf{X}) \in$

$\mathcal{H}$, there exists a function $\mathbf{b_h}(\mathbf{X}, \mathbf{X}^*)$ satisfying $\mathbf{b_h}(\mathbf{X}, \mathbf{X}^*) + \mathbf{b_h}(\mathbf{X}^*, \mathbf{X}) = \mathbf{0}$ such that $\mathbf{s_h}(Y, \mathbf{X}) := \mathbf{h}(Y, \mathbf{X}) - \mathrm{E}[\mathbf{b_h}(\mathbf{X}, \mathbf{X}^*)\{\mathrm{I}(Y \preccurlyeq Y^*) - m(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta}_0)\} \mid Y, \mathbf{X}] \in \Lambda$. For this, we need that $\mathrm{E}[\{\mathbf{s_h}(Y, \mathbf{X}) + \mathbf{s_h}(Y^*, \mathbf{X}^*)\}\{\mathrm{I}(Y \preccurlyeq Y^*) - m(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta}_0)\}|\mathbf{X}, \mathbf{X}^*] = \mathbf{0}$. For this to be fulfilled, we need the function $\mathbf{b_h}(Y, \mathbf{X})$ to satisfy the integral equation

$$
\begin{aligned}
&\mathrm{E}[\{\mathbf{h}(Y, \mathbf{X}) + \mathbf{h}(Y^*, \mathbf{X}^*)\}\{\mathrm{I}(Y \preccurlyeq Y^*) - m(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta}_0)\}|\mathbf{X}, \mathbf{X}^*] \\
&= \mathrm{E}(\mathrm{E}[\mathbf{b_h}(\mathbf{X}, \mathbf{X}^\dagger)\{\mathrm{I}(Y \preccurlyeq Y^\dagger) - m(\mathbf{X}, \mathbf{X}^\dagger; \boldsymbol{\beta}_0)\} \mid Y, \mathbf{X}]\{\mathrm{I}(Y \preccurlyeq Y^*) - m(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta}_0)\}|\mathbf{X}, \mathbf{X}^*) \\
&\quad + \mathrm{E}(\mathrm{E}[\mathbf{b_h}(\mathbf{X}^*, \mathbf{X}^\dagger)\{\mathrm{I}(Y^* \preccurlyeq Y^\dagger) - m(\mathbf{X}^*, \mathbf{X}^\dagger; \boldsymbol{\beta}_0)\} \mid Y^*, \mathbf{X}^*]\{\mathrm{I}(Y \preccurlyeq Y^*) - m(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta}_0)\}|\mathbf{X}, \mathbf{X}^*) \\
&= \mathrm{E}([\mathbf{b_h}(\mathbf{X}, \mathbf{X}^\dagger)\{\mathrm{I}(Y \preccurlyeq Y^\dagger) - m(\mathbf{X}, \mathbf{X}^\dagger; \boldsymbol{\beta}_0)\} + \mathbf{b_h}(\mathbf{X}^*, \mathbf{X}^\dagger)\{\mathrm{I}(Y^* \preccurlyeq Y^\dagger) - m(\mathbf{X}^*, \mathbf{X}^\dagger; \boldsymbol{\beta}_0)\}] \\
&\quad \times \{\mathrm{I}(Y \preccurlyeq Y^*) - m(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta}_0)\}|\mathbf{X}, \mathbf{X}^*) \\
&= \mathrm{E}\{\mathbf{b_h}(\mathbf{X}, \mathbf{X}^\dagger)\mathrm{E}[\{\mathrm{I}(Y \preccurlyeq Y^\dagger) - m(\mathbf{X}, \mathbf{X}^\dagger; \boldsymbol{\beta}_0)\}\{\mathrm{I}(Y \preccurlyeq Y^*) - m(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta}_0)\}|\mathbf{X}, \mathbf{X}^*, \mathbf{X}^\dagger]|\mathbf{X}, \mathbf{X}^*\} \\
&\quad + \mathrm{E}\{\mathbf{b_h}(\mathbf{X}^*, \mathbf{X}^\dagger)\mathrm{E}[\{\mathrm{I}(Y^* \preccurlyeq Y^\dagger) - m(\mathbf{X}^*, \mathbf{X}^\dagger; \boldsymbol{\beta}_0)\}\{\mathrm{I}(Y \preccurlyeq Y^*) - m(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta}_0)\}|\mathbf{X}, \mathbf{X}^*, \mathbf{X}^\dagger]|\mathbf{X}, \mathbf{X}^*\} \\
&= \mathrm{E}\{\mathbf{b_h}(\mathbf{X}, \mathbf{X}^\dagger)V(\mathbf{X}, \mathbf{X}^\dagger, \mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta}_0) + \mathbf{b_h}(\mathbf{X}^*, \mathbf{X}^\dagger)V(\mathbf{X}^*, \mathbf{X}^\dagger, \mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta}_0)|\mathbf{X}, \mathbf{X}^*\},
\end{aligned}
\tag{1.5}
$$

for IID copies $(Y, \mathbf{X}^T)$, $(Y^*, \mathbf{X}^{*,T})$ and $(Y^\dagger, \mathbf{X}^{\dagger,T})$ with $V(\mathbf{X}, \mathbf{X}^\dagger, \mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta}_0)$ the covariance of $\mathrm{I}(Y \preccurlyeq Y^\dagger)$ and $\mathrm{I}(Y \preccurlyeq Y^*)$ conditional on the covariates (and similar for $V(\mathbf{X}^*, \mathbf{X}^\dagger, \mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta}_0)$). To show that the latter equation allows for a solution $\mathbf{b_h}(\mathbf{X}, \mathbf{X}^*)$, we use the results of Chamberlain (1987). For this purpose, suppose that $\mathbf{X}$ is discrete, taking values $(\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_r)$ and

has a multinomial distribution, $P(\mathbf{X} = \boldsymbol{\psi}_i) = \pi_{0i}$ $(i = 1, \ldots, r)$. For every ℓ-th component of $\mathbf{b_h}$ $(\ell = 1, \ldots, p)$, equation (1.5) then boils down to a linear system of r^2 equations

$$a_{ij}^{(\ell)}(\boldsymbol{\beta}_0) = \sum_{k=1}^{r} \left\{ b_{ik}^{(\ell)} \pi_{0k} V_{ikij}(\boldsymbol{\beta}_0) + b_{jk}^{(\ell)} \pi_{0k} V_{jkij}^*(\boldsymbol{\beta}_0) \right\}$$

for any (i, j) with $i, j \in \{1, \ldots, r\}$, where $a_{ij}^{(\ell)}(\boldsymbol{\beta}_0) = \mathrm{E}[\{h^{(\ell)}(Y, \mathbf{X}) + h^{(\ell)}(Y^*, \mathbf{X}^*)\}\{\mathrm{I}(Y \preccurlyeq Y^*) - m(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta}_0)\} | \mathbf{X} = \boldsymbol{\psi}_i, \mathbf{X}^* = \boldsymbol{\psi}_j]$ ($h^{(\ell)}$ is the ℓ-th component of $\mathbf{h}$), $b_{ik}^{(\ell)} = b_{\mathbf{h}}^{(\ell)}(\mathbf{X} = \boldsymbol{\psi}_i, \mathbf{X}^\dagger = \boldsymbol{\psi}_k)$ ($b_{\mathbf{h}}^{(\ell)}$ is the ℓ-th component of $\mathbf{b_h}$ and a similar definition holds for $b_{jk}^{(\ell)}$) and finally, $V_{ikij}(\boldsymbol{\beta}_0) = V(\mathbf{X} = \boldsymbol{\psi}_i, \mathbf{X}^\dagger = \boldsymbol{\psi}_k, \mathbf{X} = \boldsymbol{\psi}_i, \mathbf{X}^* = \boldsymbol{\psi}_j; \boldsymbol{\beta}_0)$ and $V_{jkij}^*(\boldsymbol{\beta}_0) = V(\mathbf{X}^* = \boldsymbol{\psi}_j, \mathbf{X}^\dagger = \boldsymbol{\psi}_k, \mathbf{X} = \boldsymbol{\psi}_i, \mathbf{X}^* = \boldsymbol{\psi}_j; \boldsymbol{\beta}_0)$. For discrete $\mathbf{X}$, the result follows from solving this set of r^2 linear equations. The result for arbitrary $\mathbf{X}$ then follows from Lemma 3 of Chamberlain (1987).

□

Set of all influence function of RAL estimator of $\boldsymbol{\beta}$. Any influence function of a RAL estimator of $\boldsymbol{\beta}$ must satisfy $\boldsymbol{\varphi}(Y, \mathbf{X}; \boldsymbol{\beta}_0, \boldsymbol{\eta}_0) \in \Lambda^\perp$. It should however also be properly normalized so that $\mathrm{E}\{\boldsymbol{\varphi}^T(Y, \mathbf{X}; \boldsymbol{\beta}_0, \boldsymbol{\eta}_0)\mathbf{s}_{\boldsymbol{\beta}}(Y, \mathbf{X}; \boldsymbol{\beta}_0, \boldsymbol{\eta}_0)\} = \mathbf{I}_p$ with score function for $\boldsymbol{\beta}$ equal to $\mathbf{s}_{\boldsymbol{\beta}}(Y, \mathbf{X}; \boldsymbol{\beta}_0, \boldsymbol{\eta}_0) = \partial \log f_{\mathbf{Z}}(\mathbf{z}; \boldsymbol{\beta}, \boldsymbol{\eta}_0)/\partial \boldsymbol{\beta}|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}$ and $\mathbf{I}_p$ a $p \times p$ identity matrix. Consequently, any influence function $\boldsymbol{\varphi}(Y, \mathbf{X}; \boldsymbol{\beta}_0, \boldsymbol{\eta}_0)$

of a RAL estimator of $\boldsymbol{\beta}$ must equal

$$\boldsymbol{\varphi}(Y, \mathbf{X}; \boldsymbol{\beta}_0, \boldsymbol{\eta}_0) = \boldsymbol{C}_0 \mathrm{E}\left[\mathbf{b}(\mathbf{X}, \mathbf{X}^*)\{\mathrm{I}\left(Y \preccurlyeq Y^*\right) - m(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta}_0)\} \mid Y, \mathbf{X}\right], \tag{1.6}$$

with $\boldsymbol{C}_0 = \mathrm{E}[\mathbf{b}(\mathbf{X}, \mathbf{X}^*)\{\mathrm{I}\left(Y \preccurlyeq Y^*\right) - m(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta}_0)]\mathbf{s}^T_{\boldsymbol{\beta}}(Y, \mathbf{X}; \boldsymbol{\beta}_0, \boldsymbol{\eta}_0)]^{-1}$. We next derive an expression for the normalisation constant $\boldsymbol{C}_0$ that does not depend on $\mathbf{s}_{\boldsymbol{\beta}}(Y, \mathbf{X}; \boldsymbol{\beta}_0, \boldsymbol{\eta}_0)$.

Lemma 3. *The score function for $\boldsymbol{\beta}$ satisfies* $\mathrm{E}[\{\mathbf{s}^T_{\boldsymbol{\beta}}(Y, \mathbf{X}; \boldsymbol{\beta}_0, \boldsymbol{\eta}_0) + \mathbf{s}^T_{\boldsymbol{\beta}}(Y^*, \mathbf{X}^*; \boldsymbol{\beta}_0, \boldsymbol{\eta}_0)\}\{\mathrm{I}(Y \preccurlyeq Y^*) - m(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta}_0)\} \mid \mathbf{X}, \mathbf{X}^*] = \partial m(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta}) / \partial \boldsymbol{\beta}^T|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}$.

Proof. From (1.2) it follows that

$$\frac{\partial}{\partial \boldsymbol{\beta}^T}\left[\int \{\mathrm{I}(y \preccurlyeq y^*) - m(\mathbf{x}, \mathbf{x}^*; \boldsymbol{\beta})\} f_{Y|\mathbf{X}}(y|\mathbf{x}; \boldsymbol{\beta}, \boldsymbol{\eta}_{10}) f_{Y|\mathbf{X}}(y^*|\mathbf{x}^*; \boldsymbol{\beta}, \boldsymbol{\eta}_{10}) dy dy^*\right]\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} = \mathbf{0}.$$

Assuming we can interchange derivation and integration, we find

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\beta}^T} m(\mathbf{x}, \mathbf{x}^*; \boldsymbol{\beta})\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} = \int &\{\mathrm{I}(y \preccurlyeq y^*) - m(\mathbf{x}, \mathbf{x}^*; \boldsymbol{\beta}_0)\}\{\mathbf{s}^T_{\boldsymbol{\beta}}(y, \mathbf{x}; \boldsymbol{\beta}_0, \boldsymbol{\eta}_0) + \mathbf{s}^T_{\boldsymbol{\beta}}(y^*, \mathbf{x}^*; \boldsymbol{\beta}_0, \boldsymbol{\eta}_0)\} \\ &\times f_{Y|\mathbf{X}}(y|\mathbf{x}; \boldsymbol{\beta}_0, \boldsymbol{\eta}_{10}) f_{Y|\mathbf{X}}(y^*|\mathbf{x}^*; \boldsymbol{\beta}_0, \boldsymbol{\eta}_{10}) dy dy^*,\end{aligned}$$

proving the lemma. □

Lemma 4. *For an influence function $\boldsymbol{\varphi}(Y, \mathbf{X}; \boldsymbol{\beta}_0, \boldsymbol{\eta}_0)$ given in (1.6), the normalization constant* $\boldsymbol{C}_0 = -2\mathrm{E}(\partial[\mathbf{b}(\mathbf{X}, \mathbf{X}^*)\{\mathrm{I}(Y \preccurlyeq Y^*) - m(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta})\}] / \partial \boldsymbol{\beta}^T|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0})^{-1}$.

Proof. We have that $\mathrm{E}\{\mathbf{b}(\mathbf{X},\mathbf{X}^*)\partial m(\mathbf{X},\mathbf{X}^*;\boldsymbol{\beta})/\partial\boldsymbol{\beta}^T|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\} = -\mathrm{E}(\partial[\mathbf{b}(\mathbf{X},\mathbf{X}^*)\{\mathrm{I}(Y \preccurlyeq Y^*) - m(\mathbf{X},\mathbf{X}^*;\boldsymbol{\beta})\}]/\partial\boldsymbol{\beta}^T|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0})$. Next, we find that

$$
\begin{aligned}
&\mathrm{E}\left(\mathbf{b}(\mathbf{X},\mathbf{X}^*)\mathrm{E}[\{\mathbf{s}^T_{\boldsymbol{\beta}}(Y,\mathbf{X};\boldsymbol{\beta}_0,\boldsymbol{\eta}_0) + \mathbf{s}^T_{\boldsymbol{\beta}}(Y^*,\mathbf{X}^*;\boldsymbol{\beta}_0,\boldsymbol{\eta}_0)\}\{\mathrm{I}(Y \preccurlyeq Y^*) - m(\mathbf{X},\mathbf{X}^*;\boldsymbol{\beta}_0)\} \mid \mathbf{X},\mathbf{X}^*]\right)\\
&= \mathrm{E}\left[\mathbf{b}(\mathbf{X},\mathbf{X}^*)\{\mathrm{I}(Y \preccurlyeq Y^*) - m(\mathbf{X},\mathbf{X}^*;\boldsymbol{\beta}_0)\}\mathbf{s}^T_{\boldsymbol{\beta}}(Y,\mathbf{X};\boldsymbol{\beta}_0,\boldsymbol{\eta}_0)\right]\\
&\qquad + \mathrm{E}\left[\mathbf{b}(\mathbf{X},\mathbf{X}^*)\{\mathrm{I}(Y \preccurlyeq Y^*) - m(\mathbf{X},\mathbf{X}^*;\boldsymbol{\beta}_0)\}\mathbf{s}^T_{\boldsymbol{\beta}}(Y^*,\mathbf{X}^*;\boldsymbol{\beta}_0,\boldsymbol{\eta}_0)\right]\\
&= \mathrm{E}\left[\mathbf{b}(\mathbf{X},\mathbf{X}^*)\{\mathrm{I}(Y \preccurlyeq Y^*) - m(\mathbf{X},\mathbf{X}^*;\boldsymbol{\beta}_0)\}\mathbf{s}^T_{\boldsymbol{\beta}}(Y,\mathbf{X};\boldsymbol{\beta}_0,\boldsymbol{\eta}_0)\right]\\
&\qquad + \mathrm{E}\left[\mathbf{b}(\mathbf{X}^*,\mathbf{X})\{\mathrm{I}(Y^* \preccurlyeq Y) - m(\mathbf{X}^*,\mathbf{X};\boldsymbol{\beta}_0)\}\mathbf{s}^T_{\boldsymbol{\beta}}(Y,\mathbf{X};\boldsymbol{\beta}_0,\boldsymbol{\eta}_0)\right]\\
&= \mathrm{E}\left[\mathbf{b}(\mathbf{X},\mathbf{X}^*)\{\mathrm{I}(Y \preccurlyeq Y^*) - m(\mathbf{X},\mathbf{X}^*;\boldsymbol{\beta}_0)\}\mathbf{s}^T_{\boldsymbol{\beta}}(Y,\mathbf{X};\boldsymbol{\beta}_0,\boldsymbol{\eta}_0)\right]\\
&\qquad - \mathrm{E}\left[\mathbf{b}(\mathbf{X},\mathbf{X}^*)\{1 - \mathrm{I}(Y \preccurlyeq Y^*) - 1 + m(\mathbf{X},\mathbf{X}^*;\boldsymbol{\beta}_0)\}\mathbf{s}^T_{\boldsymbol{\beta}}(Y,\mathbf{X};\boldsymbol{\beta}_0,\boldsymbol{\eta}_0)\right]\\
&= 2\mathrm{E}\left[\{\mathbf{b}(\mathbf{X},\mathbf{X}^*)\{\mathrm{I}(Y \preccurlyeq Y^*) - m(\mathbf{X},\mathbf{X}^*;\boldsymbol{\beta}_0)\}\mathbf{s}^T_{\boldsymbol{\beta}}(Y,\mathbf{X};\boldsymbol{\beta}_0,\boldsymbol{\eta}_0)\right],
\end{aligned}
$$

so that $\mathrm{E}[\{\mathbf{b}(\mathbf{X},\mathbf{X}^*)\{\mathrm{I}(Y \preccurlyeq Y^*) - m(\mathbf{X},\mathbf{X}^*;\boldsymbol{\beta}_0)\}\mathbf{s}^T_{\boldsymbol{\beta}}(Y,\mathbf{X};\boldsymbol{\beta}_0,\boldsymbol{\eta}_0)] = -\mathrm{E}(\partial[\mathbf{b}(\mathbf{X},\mathbf{X}^*)\{\mathrm{I}(Y \preccurlyeq Y^*) - m(\mathbf{X},\mathbf{X}^*;\boldsymbol{\beta})\}]/\partial\boldsymbol{\beta}^T|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0})/2$. Hence, we obtain that the normalization constant is given by $\boldsymbol{C}_0 = \mathrm{E}[\{\mathbf{b}(\mathbf{X},\mathbf{X}^*)\{\mathrm{I}(Y \preccurlyeq Y^*) - m(\mathbf{X},\mathbf{X}^*;\boldsymbol{\beta}_0)\}\mathbf{s}^T_{\boldsymbol{\beta}}(Y,\mathbf{X};\boldsymbol{\beta}_0,\boldsymbol{\eta}_0)]^{-1} = -2\mathrm{E}(\partial[\mathbf{b}(\mathbf{X},\mathbf{X}^*)\{\mathrm{I}(Y \preccurlyeq Y^*) - m(\mathbf{X},\mathbf{X}^*;\boldsymbol{\beta})\}]/\partial\boldsymbol{\beta}^T|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0})^{-1}$. □

Conclusion. Any influence function of a RAL estimator of $\boldsymbol{\beta}$ is of the form

$$
\boldsymbol{\varphi}(Y,\mathbf{X};\boldsymbol{\beta}_0,\boldsymbol{\eta}_0) = -2\mathrm{E}\{\partial\mathbf{u}(\mathbf{Z},\mathbf{Z}^*;\boldsymbol{\beta})/\partial\boldsymbol{\beta}|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\}^{-1}\mathrm{E}\{\mathbf{u}(\mathbf{Z},\mathbf{Z}^*;\boldsymbol{\beta}_0) \mid Y,\mathbf{X}\}
$$

with $\mathbf{u}(\mathbf{Z}, \mathbf{Z}^*; \boldsymbol{\beta}_0) = \mathbf{b}(\mathbf{X}, \mathbf{X}^*)\{\mathrm{I}(Y \preccurlyeq Y^*) - m(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta}_0)\}$ and for any arbitrary $\mathbf{b}(\mathbf{X}, \mathbf{X}^*) \in \mathcal{B}$.

2. Asymptotic distribution of $\widehat{\boldsymbol{\beta}}$

Consider the estimator $\widehat{\boldsymbol{\beta}}$ that solves the system of estimating equations $\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{U}_{ij}(\widehat{\boldsymbol{\beta}})=\mathbf{0}$ with $\mathbf{U}_{ij}(\boldsymbol{\beta})=\mathbf{B}_{ij}(\boldsymbol{\beta})\{\mathrm{I}_{ij}-M_{ij}(\boldsymbol{\beta})\}$, with $\mathbf{U}_{ij}(\boldsymbol{\beta})$ a shorthand for $\mathbf{u}(\mathbf{Z}_i,\mathbf{Z}_j;\boldsymbol{\beta})$, $\mathbf{B}_{ij}(\boldsymbol{\beta})=\mathbf{b}(\mathbf{X}_i,\mathbf{X}_j;\boldsymbol{\beta})$, $\mathrm{I}_{ij}=\mathrm{I}\left(Y_i\preccurlyeq Y_j\right)$ and $M_{ij}(\boldsymbol{\beta})=m(\mathbf{X}_i,\mathbf{X}_j;\boldsymbol{\beta})$ and $\mathbf{B}_{ij}(\boldsymbol{\beta})$ satisfying the antisymmetry condition $\mathbf{B}_{ij}(\boldsymbol{\beta})+\mathbf{B}_{ji}(\boldsymbol{\beta})=\mathbf{0}$. Throughout, $\mathbf{Z}^T=(Y,\mathbf{X}^T)$ and $\mathbf{Z}^{*,T}=(Y^*,\mathbf{X}^{*,T})$ denote two IID copies. Define the U-statistic (van der Vaart, 1998) $\mathbf{U}_n(\boldsymbol{\beta})=\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{U}_{ij}(\boldsymbol{\beta})/n^2$ and with $\mathrm{E}\{\mathbf{U}_{ij}(\boldsymbol{\beta}_0)\}=\mathbf{0}$. Note that $\mathbf{U}_n(\boldsymbol{\beta})$ is permutation symmetric in $\mathbf{X}_i$ and $\mathbf{X}_j$ since $\mathbf{U}_{ij}(\boldsymbol{\beta})=\mathbf{U}_{ji}(\boldsymbol{\beta})$, which follows by the antisymmetry of $\mathbf{B}_{ij}(\boldsymbol{\beta})$.

Regularity conditions. To show that $\widehat{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}$, to derive the asymptotic distribution of $\widehat{\boldsymbol{\beta}}$ and to show consistency of sandwich estimator $\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}})$, we assume the following regularity conditions. (R1) the parameter space $\Theta\subset\mathbb{R}^p$ of $\boldsymbol{\beta}$ is a compact set; (R2) $\boldsymbol{\beta}_0$ lies in the interior of Θ; (R3) $\mathrm{E}\{\mathbf{u}(\mathbf{Z},\mathbf{Z}^*;\boldsymbol{\beta})\}\neq\mathbf{0}$ if $\boldsymbol{\beta}\neq\boldsymbol{\beta}_0$ and $\mathrm{E}\{\mathbf{u}(\mathbf{Z},\mathbf{Z}^*;\boldsymbol{\beta}_0)\}=\mathbf{0}$ (that is, $\mathrm{E}\{\mathbf{u}(\mathbf{Z},\mathbf{Z}^*;\boldsymbol{\beta})\}$ has a unique root $\boldsymbol{\beta}_0\in\Theta$); (R4) $\mathbf{u}(\mathbf{Z},\mathbf{Z}^*;\boldsymbol{\beta})$ is continuous in every $\boldsymbol{\beta}\in\Theta$ wp1; (R5) $\mathrm{E}\{\sup_{\boldsymbol{\beta}\in\Theta}\|\mathbf{u}(\mathbf{Z},\mathbf{Z}^*;\boldsymbol{\beta})\|\}<\infty$; (R6) $\partial\mathbf{u}(\mathbf{Z},\mathbf{Z}^*;\boldsymbol{\beta})/\partial\boldsymbol{\beta}^T$ is continuous in every $\boldsymbol{\beta}\in\Theta$ wp1; (R7) $\mathrm{E}\{\sup_{\boldsymbol{\beta}\in\Theta}\|\partial\mathbf{u}(\mathbf{Z},\mathbf{Z}^*;\boldsymbol{\beta})/\partial\boldsymbol{\beta}^T\|\}<\infty$; (R8) $\mathbf{J}(\boldsymbol{\beta}_0)=\mathrm{E}\{\partial\mathbf{u}(\mathbf{Z},\mathbf{Z}^*;\boldsymbol{\beta})/\partial\boldsymbol{\beta}^T|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\}$ is invertible; (R9) $\mathrm{E}\{\sup_{\boldsymbol{\beta}\in\Theta}\|\mathbf{u}(\mathbf{Z},\mathbf{Z}^*;\boldsymbol{\beta})\|^2\}<$

∞; and (R10) $\mathrm{E}\{\sup_{\boldsymbol{\beta}\in\Theta} \|\partial \mathbf{u}(\mathbf{Z}, \mathbf{Z}^*; \boldsymbol{\beta})/\partial \boldsymbol{\beta}^T\|^2\} < \infty$, with $\|\cdot\|$ denoting the Euclidean norm so that $\|\mathbf{a}\| = (\sum_{i=1}^p a_i^2)^{1/2}$ for $\mathbf{a} = (a_1, \ldots, a_p)^T \in \mathbb{R}^p$ and $\|\mathrm{A}\| = (\sum_{i=1}^p \sum_{j=1}^p a_{ij}^2)^{1/2}$ for $\mathrm{A} = (a_{ij})_{i,j=1\ldots p} \in \mathbb{R}^{p\times p}$.

Consistency of $\widehat{\boldsymbol{\beta}}$. To show that $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_0$, we will apply Theorem 2.1 of Newey and McFadden (1994). Define the function $\mathbb{U}_n(\boldsymbol{\beta}) = -[\mathbf{U}_n^T(\boldsymbol{\beta})\mathbf{U}_n(\boldsymbol{\beta})]$ so that $\widehat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}\in\Theta}\{\mathbb{U}_n(\boldsymbol{\beta})\}$ with maximal value 0. Next define the function $\mathfrak{u}(\boldsymbol{\beta}) = -[\mathrm{E}^T\{\mathbf{u}(\mathbf{Z}, \mathbf{Z}^*; \boldsymbol{\beta})\}\mathrm{E}\{\mathbf{u}(\mathbf{Z}, \mathbf{Z}^*; \boldsymbol{\beta})\}]$ where $\mathfrak{u}(\boldsymbol{\beta}_0) = \mathbf{0}$. Consistency will follow if the subsequent conditions are satisfied: (C1) $\mathfrak{u}(\boldsymbol{\beta})$ is uniquely maximized at $\boldsymbol{\beta}_0$, (C2) Θ is compact, (C3) $\mathfrak{u}(\boldsymbol{\beta})$ is continuous, and (C4) $\mathbb{U}_n(\boldsymbol{\beta})$ converges uniformly in probability to $\mathfrak{u}(\boldsymbol{\beta})$. (C1) is satisfied by (R2) and (R3). (C2) is satisfied by (R1). To show (C3) and (C4), we make use of Lemma 8.5 of Newey and McFadden (1994), which guarantees that $\mathrm{E}\{\mathbf{u}(\mathbf{Z}, \mathbf{Z}^*; \boldsymbol{\beta})\}$ is continuous in $\boldsymbol{\beta} \in \Theta$ and that $\sup_{\boldsymbol{\beta}\in\Theta} \|\mathbf{U}_n(\boldsymbol{\beta}) - \mathrm{E}\{\mathbf{u}(\mathbf{Z}, \mathbf{Z}^*; \boldsymbol{\beta})\| \xrightarrow{p} 0$ if (i) $\mathbf{u}(\mathbf{Z}, \mathbf{Z}^*; \boldsymbol{\beta})$ is continuous at each $\boldsymbol{\beta} \in \Theta$ wp1, which is satisfied by (R4), and (ii) $\mathrm{E}\{\sup_{\boldsymbol{\beta}\in\Theta} \|\mathbf{u}(\mathbf{Z}, \mathbf{Z}^*; \boldsymbol{\beta})\|\} < \infty$, which is satisfied by (R5). It follows that (C3) is satisfied since $\mathfrak{u}(\boldsymbol{\beta})$ is a continuous transformation of $\mathrm{E}\{\mathbf{u}(\mathbf{Z}, \mathbf{Z}^*; \boldsymbol{\beta})\}$. For (C4), we need to show that $\sup_{\boldsymbol{\beta}\in\Theta} |\mathbb{U}_n(\boldsymbol{\beta}) - \mathfrak{u}(\boldsymbol{\beta})| \xrightarrow{p} 0$. We find that (using the triangle inequality and the Cauchy-Schwarz inequality) $|\mathbb{U}_n(\boldsymbol{\beta}) - \mathfrak{u}(\boldsymbol{\beta})| = |[\mathbf{U}_n^T(\boldsymbol{\beta})\mathbf{U}_n(\boldsymbol{\beta})] -$

$[\mathrm{E}^T\{\mathbf{u}(\mathbf{Z},\mathbf{Z}^*;\boldsymbol{\beta})\}\mathrm{E}\{\mathbf{u}(\mathbf{Z},\mathbf{Z}^*;\boldsymbol{\beta})\}]| \leq |[\mathbf{U}_n(\boldsymbol{\beta}) - \mathrm{E}\{\mathbf{u}(\mathbf{Z},\mathbf{Z}^*;\boldsymbol{\beta})\}]^T[\mathbf{U}_n(\boldsymbol{\beta}) - \mathrm{E}\{\mathbf{u}(\mathbf{Z},\mathbf{Z}^*;\boldsymbol{\beta})\}]|+2|\mathrm{E}^T\{\mathbf{u}(\mathbf{Z},\mathbf{Z}^*;\boldsymbol{\beta})\}[\mathbf{U}_n(\boldsymbol{\beta})-\mathrm{E}\{\mathbf{u}(\mathbf{Z},\mathbf{Z}^*;\boldsymbol{\beta})\}]| \leq \|\mathbf{U}_n(\boldsymbol{\beta})-\mathrm{E}\{\mathbf{u}(\mathbf{Z},\mathbf{Z}^*;\boldsymbol{\beta})\}\|^2+2\|\mathrm{E}\{\mathbf{u}(\mathbf{Z},\mathbf{Z}^*;\boldsymbol{\beta})\}\|\|\mathbf{U}_n(\boldsymbol{\beta})-\mathrm{E}\{\mathbf{u}(\mathbf{Z},\mathbf{Z}^*;\boldsymbol{\beta})\}\|$, so $\sup_{\boldsymbol{\beta}\in\Theta}|\mathbb{U}_n(\boldsymbol{\beta})-\mathfrak{u}(\boldsymbol{\beta})| \xrightarrow{p} 0$ because $\sup_{\boldsymbol{\beta}\in\Theta}\|\mathbf{U}_n(\boldsymbol{\beta})-\mathrm{E}\{\mathbf{u}(\mathbf{Z},\mathbf{Z}^*;\boldsymbol{\beta})\}\| \xrightarrow{p} 0$ and $\sup_{\boldsymbol{\beta}\in\Theta}\|\mathrm{E}\{\mathbf{u}(\mathbf{Z},\mathbf{Z}^*;\boldsymbol{\beta})\}\| < \infty$ by the continuity of $\mathrm{E}\{\mathbf{u}(\mathbf{Z},\mathbf{Z}^*;\boldsymbol{\beta})\}$ over the compact set Θ. We can conclude that $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_0$.

Asymptotic distribution of $\widehat{\boldsymbol{\beta}}$. Consider the first-order Taylor expansion $\mathbf{0} = \sqrt{n}\mathbf{U}_n(\widehat{\boldsymbol{\beta}}) = \sqrt{n}\mathbf{U}_n(\boldsymbol{\beta}_0)+\partial\mathbf{U}_n(\boldsymbol{\beta})/\partial\boldsymbol{\beta}^T|_{\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}}\sqrt{n}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)$, with $\|\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\| \leq \|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|$. Because $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_0$, we also have that $\tilde{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_0$. From this first-order Taylor expansion, we find that $\sqrt{n}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0) = -\{\partial\mathbf{U}_n(\boldsymbol{\beta})/\partial\boldsymbol{\beta}^T|_{\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}}\}^{-1}\sqrt{n}\mathbf{U}_n(\boldsymbol{\beta}_0)$.

First we show that $\partial\mathbf{U}_n(\boldsymbol{\beta})/\partial\boldsymbol{\beta}^T|_{\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}} \xrightarrow{p} \mathrm{E}\{\partial\mathbf{u}(\mathbf{Z},\mathbf{Z}^*;\boldsymbol{\beta})/\partial\boldsymbol{\beta}^T|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\}$. In doing so, we again employ Lemma 8.5 of Newey and McFadden (1994), which guarantees that $\mathrm{E}\{\partial\mathbf{u}(\mathbf{Z},\mathbf{Z}^*;\boldsymbol{\beta})/\partial\boldsymbol{\beta}^T\}$ is continuous in $\boldsymbol{\beta} \in \Theta$ and that $\sup_{\boldsymbol{\beta}\in\Theta}\|\partial\mathbf{U}_n(\boldsymbol{\beta})/\partial\boldsymbol{\beta}^T-\mathrm{E}\{\partial\mathbf{u}(\mathbf{Z},\mathbf{Z}^*;\boldsymbol{\beta})/\partial\boldsymbol{\beta}^T\}\| \xrightarrow{p} 0$ if (i) $\partial\mathbf{u}(\mathbf{Z},\mathbf{Z}^*;\boldsymbol{\beta})/\partial\boldsymbol{\beta}^T$ is continuous at each $\boldsymbol{\beta} \in \Theta$ wp1, which is satisfied by (R6), and (ii) $\mathrm{E}\{\sup_{\boldsymbol{\beta}\in\Theta}\|\partial\mathbf{u}(\mathbf{Z},\mathbf{Z}^*;\boldsymbol{\beta})/\partial\boldsymbol{\beta}^T\|\} < \infty$, which is satisfied by (R7). Next, from the triangle inequality, it follows that $\|\partial\mathbf{U}_n(\boldsymbol{\beta})/\partial\boldsymbol{\beta}^T|_{\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}}-\mathrm{E}\{\partial\mathbf{u}(\mathbf{Z},\mathbf{Z}^*;\boldsymbol{\beta})/\partial\boldsymbol{\beta}^T|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\}\| \leq \|\partial\mathbf{U}_n(\boldsymbol{\beta})/\partial\boldsymbol{\beta}^T|_{\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}}-\mathrm{E}\{\partial\mathbf{u}(\mathbf{Z},\mathbf{Z}^*;\boldsymbol{\beta})/\partial\boldsymbol{\beta}^T|_{\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}}\}\|+\|\mathrm{E}\{\partial\mathbf{u}(\mathbf{Z},\mathbf{Z}^*;\boldsymbol{\beta})/\partial\boldsymbol{\beta}^T|_{\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}}\}-\mathrm{E}\{\partial\mathbf{u}(\mathbf{Z},\mathbf{Z}^*;\boldsymbol{\beta})/\partial\boldsymbol{\beta}^T|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\}\|$. The first term of the rhs is bounded by $\sup_{\boldsymbol{\beta}\in\Theta}\|\partial\mathbf{U}_n(\boldsymbol{\beta})/\partial\boldsymbol{\beta}^T-$

$\mathrm{E}\{\partial \mathbf{u}(\mathbf{Z}, \mathbf{Z}^*; \boldsymbol{\beta})/\partial \boldsymbol{\beta}^T\}\| \xrightarrow{p} 0$ and the second term of the rhs is $o_p(1)$ by the continuous mapping theorem since $\mathrm{E}\{\partial \mathbf{u}(\mathbf{Z}, \mathbf{Z}^*; \boldsymbol{\beta})/\partial \boldsymbol{\beta}^T\}$ is continuous in $\boldsymbol{\beta} \in \Theta$ and $\tilde{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_0$ so that the lhs is $o_p(1)$. We conclude that the Jacobian $\partial \mathbf{U}_n(\boldsymbol{\beta})/\partial \boldsymbol{\beta}^T|_{\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}} = \mathrm{E}\{\partial \mathbf{u}(\mathbf{Z}, \mathbf{Z}^*; \boldsymbol{\beta})/\partial \boldsymbol{\beta}^T|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\} + o_p(1) = \mathbf{J}(\boldsymbol{\beta}_0) + o_p(1)$.

From the uniform convergence of the Jacobian-term, and assuming it is invertible (R8), we find that $\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = -\mathbf{J}(\boldsymbol{\beta}_0)^{-1}\sqrt{n}\mathbf{U}_n(\boldsymbol{\beta}_0) + o_p(1)$. Next, consider the Hájek-projection of $\mathbf{U}_n(\boldsymbol{\beta}_0)$: $\widehat{\mathbf{U}}_n(\boldsymbol{\beta}_0) = \sum_{i=1}^n \mathrm{E}\{\mathbf{U}_n(\boldsymbol{\beta}_0)|Y_i, \mathbf{X}_i\}/n = 2\sum_{i=1}^n \mathrm{E}\{\mathbf{U}_{ij}(\boldsymbol{\beta}_0)|Y_i, \mathbf{X}_i\}/n$. Assuming (R5) and (R9), it follows from Theorem 12.3 of van der Vaart (1998) that $\sqrt{n}\mathbf{U}_n(\boldsymbol{\beta}_0) = \sqrt{n}\widehat{\mathbf{U}}_n(\boldsymbol{\beta}_0) + o_p(1)$, so that $\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = \sum_{i=1}^n -2\mathbf{J}(\boldsymbol{\beta}_0)^{-1}\mathrm{E}\{\mathbf{U}_{ij}(\boldsymbol{\beta}_0)|Y_i, \mathbf{X}_i\}/\sqrt{n} + o_p(1)$. This shows that $\widehat{\boldsymbol{\beta}}$ is an asymptotically linear estimator of $\boldsymbol{\beta}_0$ with influence function $\boldsymbol{\varphi}(Y_i, \mathbf{X}_i; \boldsymbol{\beta}_0, \boldsymbol{\eta}_0) = -2\mathbf{J}(\boldsymbol{\beta}_0)^{-1}\mathrm{E}\{\mathbf{U}_{ij}(\boldsymbol{\beta}_0)|Y_i, \mathbf{X}_i\}$.

Consistency of $\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}})$. We show that the sandwich estimator $\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}}) = 4\widehat{\mathbf{J}}(\widehat{\boldsymbol{\beta}})^{-1}\widehat{\mathbf{K}}(\widehat{\boldsymbol{\beta}})\widehat{\mathbf{J}}(\widehat{\boldsymbol{\beta}})^{-T} \xrightarrow{p} 4\mathbf{J}(\boldsymbol{\beta}_0)^{-1}\mathrm{cov}[\mathrm{E}\{\mathbf{U}_{ij}(\boldsymbol{\beta}_0) \mid Y_i, \mathbf{X}_i\}]\mathbf{J}(\boldsymbol{\beta}_0)^{-T} = \boldsymbol{\Sigma}_0$, $\widehat{\mathbf{J}}(\widehat{\boldsymbol{\beta}}) = \sum_{i=1}^n \sum_{j=1}^n \partial \mathbf{U}_{ij}(\boldsymbol{\beta})/\partial \boldsymbol{\beta}^T|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}/n^2$ and $\widehat{\mathbf{K}}(\widehat{\boldsymbol{\beta}}) = \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n \mathbf{U}_{ij}(\widehat{\boldsymbol{\beta}})\mathbf{U}_{ik}^T(\widehat{\boldsymbol{\beta}})/n^3$. We already know that $\widehat{\mathbf{J}}(\widehat{\boldsymbol{\beta}}) \xrightarrow{p} \mathbf{J}(\boldsymbol{\beta}_0)$. From (R10), Lemma 8.7 and Lemma 8.3 of Newey and McFadden (1994), it also follows that $\widehat{\mathbf{K}}(\widehat{\boldsymbol{\beta}}) \xrightarrow{p} \mathrm{cov}[\mathrm{E}\{\mathbf{U}_{ij}(\boldsymbol{\beta}_0) \mid Y_i, \mathbf{X}_i\}]$. The continuous mapping theorem then guarantees that $\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}}) = 4\widehat{\mathbf{J}}(\widehat{\boldsymbol{\beta}})^{-1}\widehat{\mathbf{K}}(\widehat{\boldsymbol{\beta}})\widehat{\mathbf{J}}(\widehat{\boldsymbol{\beta}})^{-T} \xrightarrow{p} 4\mathbf{J}(\boldsymbol{\beta}_0)^{-1}\mathrm{cov}[\mathrm{E}\{\mathbf{U}_{ij}(\boldsymbol{\beta}_0) \mid Y_i, \mathbf{X}_i\}]\mathbf{J}(\boldsymbol{\beta}_0)^{-T} = \boldsymbol{\Sigma}_0$.

Regularity conditions, revisited. We try to make (R4)-(R7), (R9) and (R10) more explicit. We have $\mathbf{u}(\mathbf{Z}, \mathbf{Z}^*; \boldsymbol{\beta}) = \mathbf{b}(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta})\{\mathrm{I}(Y \preceq Y^*) - m(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta})\}$ and thus $\partial \mathbf{u}(\mathbf{Z}, \mathbf{Z}^*; \boldsymbol{\beta})/\partial \boldsymbol{\beta}^T = \partial \mathbf{b}(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta})/\partial \boldsymbol{\beta}^T \{\mathrm{I}(Y \preceq Y^*) - m(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta})\} - \mathbf{b}(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta}) \partial m(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta})/\partial \boldsymbol{\beta}^T$. It follows that (R4) and (R6) will be satisfied if both $\mathbf{b}(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta})$ and $m(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta})$ are sufficiently smooth, e.g., they are both continuously differentiable wp1 with respect to $\boldsymbol{\beta}$. Next, we find that $\mathrm{E}\{\sup_{\boldsymbol{\beta} \in \Theta} \|\mathbf{u}(\mathbf{Z}, \mathbf{Z}^*; \boldsymbol{\beta})\|\} = \mathrm{E}[\sup_{\boldsymbol{\beta} \in \Theta}\{\|\mathbf{b}(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta})\| |\mathrm{I}(Y \preceq Y^*) - m(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta})|\}]$, which is bounded by $2\mathrm{E}\{\sup_{\boldsymbol{\beta} \in \Theta} \|\mathbf{b}(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta})\|\}$. For the derivative, from the triangle inequality it follows that $\mathrm{E}\{\sup_{\boldsymbol{\beta} \in \Theta} \|\partial \mathbf{u}(\mathbf{Z}, \mathbf{Z}^*; \boldsymbol{\beta})/\partial \boldsymbol{\beta}^T\|\}$
$\leq \mathrm{E}[\sup_{\boldsymbol{\beta} \in \Theta}\{\|\partial \mathbf{b}(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta})/\partial \boldsymbol{\beta}^T\| |\mathrm{I}(Y \preceq Y^*) - m(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta})|\}]$
$+ \mathrm{E}[\sup_{\boldsymbol{\beta} \in \Theta}\{\|\mathbf{b}(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta})\| \|\partial m(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta})/\partial \boldsymbol{\beta}^T\|\}]$. The second term of the rhs can be bounded by $\mathrm{E}[\{\sup_{\boldsymbol{\beta} \in \Theta} \|\mathbf{b}(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta})\| \}^2]^{1/2} \mathrm{E}[\{\sup_{\boldsymbol{\beta} \in \Theta} \|\partial m(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta})/\partial \boldsymbol{\beta}^T\|\}^2]^{1/2}$ using the Cauchy-Schwarz inequality. If follows that (R5) and (R7) will be satisfied if the classes of functions $\{\mathbf{b}(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta}) : \boldsymbol{\beta} \in \Theta\}$, $\{\partial \mathbf{b}(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta})/\partial \boldsymbol{\beta}^T : \boldsymbol{\beta} \in \Theta\}$ and $\{\partial m(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta})/\partial \boldsymbol{\beta}^T : \boldsymbol{\beta} \in \Theta\}$ (with Θ compact) has square integrable envelopes (wrt the the true probability measure), which is not a strong condition since these are standard smoothness and moment conditions. These moment conditions are also standard in order to obtain a proper influence function with finite variance. A similar reasoning can be made for (R9)-(R10).

3. Proof of Theorem 2

Because the asymptotic variance of a RAL estimator is dictated by the variance of its influence function, we want to identify the influence function with the smallest possible variance under model $\mathcal{M}_{\text{PIM}}$, the semiparametric efficiency bound. This influence function is called the *efficient influence function*. From semiparametric theory, we know that the efficient influence function is given by $\boldsymbol{\varphi}^{\text{EFF}}(Y, \mathbf{X}; \boldsymbol{\beta}_0, \boldsymbol{\eta}) = \text{E}\{\mathbf{s}^{\text{EFF}}(Y, \mathbf{X}; \boldsymbol{\beta}_0, \boldsymbol{\eta}_0)\mathbf{s}^{\text{EFF},T}(Y, \mathbf{X}; \boldsymbol{\beta}_0, \boldsymbol{\eta}_0)\}^{-1}\mathbf{s}^{\text{EFF}}(Y, \mathbf{X}; \boldsymbol{\beta}_0, \boldsymbol{\eta}_0)$ with efficient score

$$\mathbf{s}^{\text{EFF}}(Y, \mathbf{X}; \boldsymbol{\beta}_0, \boldsymbol{\eta}_0) = \mathbf{s}_{\boldsymbol{\beta}}(Y, \mathbf{X}; \boldsymbol{\beta}_0, \boldsymbol{\eta}_0) - \Pi\{\mathbf{s}_{\boldsymbol{\beta}}(Y, \mathbf{X}; \boldsymbol{\beta}_0, \boldsymbol{\eta}_0) \mid \Lambda\} = \Pi\{\mathbf{s}_{\boldsymbol{\beta}}(Y, \mathbf{X}; \boldsymbol{\beta}_0, \boldsymbol{\eta}_0) \mid \Lambda^{\perp}\}$$

and $\Pi(\cdot \mid \cdot)$ the orthogonal projection operator. In order to find the efficient score, we thus need to find the function $\mathbf{b}^{\text{EFF}}(\mathbf{X}, \mathbf{X}^*) \in \mathcal{B}$ so that $\mathbf{s}^{\text{EFF}}(Y, \mathbf{X}; \boldsymbol{\beta}_0, \boldsymbol{\eta}_0) = \text{E}[\mathbf{b}^{\text{EFF}}(\mathbf{X}, \mathbf{X}^*)\{\text{I}(Y \preccurlyeq Y^*) - m(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta}_0)\} \mid Y, \mathbf{X}]$. From the proof of Proposition 2, it follows that this boils down to solving the integral equation

$$\begin{aligned}
&\text{E}[\{\mathbf{s}_{\boldsymbol{\beta}}(Y, \mathbf{X}; \boldsymbol{\beta}_0, \boldsymbol{\eta}_0) + \mathbf{s}_{\boldsymbol{\beta}}(Y^*, \mathbf{X}^*; \boldsymbol{\beta}_0, \boldsymbol{\eta}_0)\}\{\text{I}(Y \preccurlyeq Y^*) - m(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta}_0)\}|\mathbf{X}, \mathbf{X}^*] \\
&= \text{E}([\mathbf{b}^{\text{EFF}}(\mathbf{X}, \mathbf{X}^\dagger)\{\text{I}(Y \preccurlyeq Y^\dagger) - m(\mathbf{X}, \mathbf{X}^\dagger; \boldsymbol{\beta}_0)\} + \mathbf{b}^{\text{EFF}}(\mathbf{X}^*, \mathbf{X}^\dagger)\{\text{I}(Y^* \preccurlyeq Y^\dagger) - m(\mathbf{X}^*, \mathbf{X}^\dagger; \boldsymbol{\beta}_0)\}] \\
&\quad \times \{\text{I}(Y \preccurlyeq Y^*) - m(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta}_0)\}|\mathbf{X}, \mathbf{X}^*).
\end{aligned}$$

Define $\mathbf{d}(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta}_0) = \partial m(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta})/\partial \boldsymbol{\beta}|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}$ and $V(\mathbf{X}, \mathbf{X}^*, \mathbf{X}^\dagger, \mathbf{X}'; \boldsymbol{\beta}_0) = \text{cov}\{\text{I}(Y \preccurlyeq Y^*), \text{I}(Y^\dagger \preccurlyeq Y') \mid \mathbf{X}, \mathbf{X}^*, \mathbf{X}^\dagger, \mathbf{X}'\} = \text{E}[\{\text{I}(Y \preccurlyeq Y^*) - m(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta}_0)\}\{\text{I}(Y^\dagger \preccurlyeq$

$Y') - m(\mathbf{X}^\dagger, \mathbf{X}'; \boldsymbol{\beta}_0)\} \mid \mathbf{X}, \mathbf{X}^*, \mathbf{X}^\dagger, \mathbf{X}']$ for IID copies $(Y, \mathbf{X}^T)$, $(Y^*, \mathbf{X}^{*,T})$, $(Y^\dagger, \mathbf{X}^{\dagger,T})$ and $(Y', \mathbf{X}^{',T})$. From Lemma 3, we know that $\mathrm{E}[\{\mathbf{s}_{\boldsymbol{\beta}}(Y, \mathbf{X}; \boldsymbol{\beta}_0, \boldsymbol{\eta}_0) + \mathbf{s}_{\boldsymbol{\beta}}(Y^*, \mathbf{X}^*; \boldsymbol{\beta}_0, \boldsymbol{\eta}_0)\}\{\mathrm{I}(Y \preccurlyeq Y^*) - m(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta}_0)\} | \mathbf{X}, \mathbf{X}^*] = \mathbf{d}(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta}_0)$.

The function $\mathbf{b}^{\mathrm{EFF}}(\mathbf{X}, \mathbf{X}^*)$ is thus the solution to the integral equation

$$\mathbf{d}(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta}_0) \tag{3.1}$$
$$= \mathrm{E}\{\mathbf{b}^{\mathrm{EFF}}(\mathbf{X}, \mathbf{X}^\dagger; \boldsymbol{\beta}_0) V(\mathbf{X}, \mathbf{X}^*, \mathbf{X}, \mathbf{X}^\dagger; \boldsymbol{\beta}_0) + \mathbf{b}^{\mathrm{EFF}}(\mathbf{X}^*, \mathbf{X}^\dagger; \boldsymbol{\beta}_0) V(\mathbf{X}, \mathbf{X}^*, \mathbf{X}^*, \mathbf{X}^\dagger; \boldsymbol{\beta}_0) \mid \mathbf{X}, \mathbf{X}^*\}.$$

This proves Theorem 2.

4. Equivalence equation (8) and (9) from Section 2.2

Using the same notation as before, we show that

$$\mathbf{D}_{ij}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{k=1}^{n}\left\{\widehat{\mathbf{B}}_{ik}^{\mathrm{EFF}}(\boldsymbol{\beta})V_{ijik}(\boldsymbol{\beta}) + \widehat{\mathbf{B}}_{jk}^{\mathrm{EFF}}(\boldsymbol{\beta})V_{ijjk}(\boldsymbol{\beta})\right\}, \qquad i,j = 1,\ldots,n \tag{4.1}$$

can be written as

$$\mathbf{D}_{ij}(\boldsymbol{\beta}) = \frac{1}{n}\left[\sum_{k=1}^{n-1}\sum_{\ell=k+1}^{n}\left\{\widehat{\mathbf{B}}_{k\ell}^{\mathrm{EFF}}(\boldsymbol{\beta})V_{ijk\ell}(\boldsymbol{\beta})\right\} + \widehat{\mathbf{B}}_{ij}^{\mathrm{EFF}}(\boldsymbol{\beta})V_{ijij}(\boldsymbol{\beta})\right], \qquad i<j. \tag{4.2}$$

In Section 2.2, we already argued that (4.1) reduces to those $n(n-1)/2$ equations for which $i<j$. The two terms of the right hand side of equation equation (4.1) can be rewritten as

$$\begin{aligned}
\frac{1}{n}\sum_{k=1}^{n}\widehat{\mathbf{B}}_{ik}^{\mathrm{EFF}}(\boldsymbol{\beta})V_{ijik}(\boldsymbol{\beta}) &= \frac{1}{n}\sum_{k=1}^{i-1}\widehat{\mathbf{B}}_{ik}^{\mathrm{EFF}}(\boldsymbol{\beta})V_{ijik}(\boldsymbol{\beta}) + \frac{1}{n}\sum_{k=i+1}^{n}\widehat{\mathbf{B}}_{ik}^{\mathrm{EFF}}(\boldsymbol{\beta})V_{ijik}(\boldsymbol{\beta}) \\
&= \frac{1}{n}\sum_{k=1}^{i-1}\widehat{\mathbf{B}}_{ik}^{\mathrm{EFF}}(\boldsymbol{\beta})V_{ijik}(\boldsymbol{\beta}) + \frac{1}{n}\sum_{\ell=i+1}^{n}\widehat{\mathbf{B}}_{i\ell}^{\mathrm{EFF}}(\boldsymbol{\beta})V_{iji\ell}(\boldsymbol{\beta}) \\
\frac{1}{n}\sum_{k=1}^{n}\widehat{\mathbf{B}}_{jk}^{\mathrm{EFF}}(\boldsymbol{\beta})V_{ijjk}(\boldsymbol{\beta}) &= \frac{1}{n}\sum_{k=1}^{j-1}\widehat{\mathbf{B}}_{jk}^{\mathrm{EFF}}(\boldsymbol{\beta})V_{ijjk}(\boldsymbol{\beta}) + \frac{1}{n}\sum_{k=j+1}^{n}\widehat{\mathbf{B}}_{jk}^{\mathrm{EFF}}(\boldsymbol{\beta})V_{ijjk}(\boldsymbol{\beta}) \\
&= \frac{1}{n}\sum_{k=1}^{j-1}\widehat{\mathbf{B}}_{jk}^{\mathrm{EFF}}(\boldsymbol{\beta})V_{ijjk}(\boldsymbol{\beta}) + \frac{1}{n}\sum_{\ell=j+1}^{n}\widehat{\mathbf{B}}_{j\ell}^{\mathrm{EFF}}(\boldsymbol{\beta})V_{ijj\ell}(\boldsymbol{\beta}).
\end{aligned}$$

The terms where $k=i$ and $k=j$ do not contribute to these sums since $\widehat{\mathbf{B}}_{ii}^{\mathrm{EFF}}(\boldsymbol{\beta}) = \widehat{\mathbf{B}}_{jj}^{\mathrm{EFF}}(\boldsymbol{\beta}) = \mathbf{0}$. For the sums going from $i+1$ to n and $j+1$ to n, the summation index is changed from k to ℓ. Equation (4.1) can thus be

written as

$$\begin{aligned}\mathbf{D}_{ij}(\boldsymbol{\beta}) &= \frac{1}{n}\left\{\sum_{\ell=i+1}^{n}\widehat{\mathbf{B}}_{i\ell}^{\mathrm{EFF}}(\boldsymbol{\beta})V_{iji\ell}(\boldsymbol{\beta})+\sum_{\ell=j+1}^{n}\widehat{\mathbf{B}}_{j\ell}^{\mathrm{EFF}}(\boldsymbol{\beta})V_{ijj\ell}(\boldsymbol{\beta})\right.\\ &\qquad\left.+\sum_{k=1}^{i-1}\widehat{\mathbf{B}}_{ik}^{\mathrm{EFF}}(\boldsymbol{\beta})V_{ijik}(\boldsymbol{\beta})+\sum_{k=1}^{j-1}\widehat{\mathbf{B}}_{jk}^{\mathrm{EFF}}(\boldsymbol{\beta})V_{ijjk}(\boldsymbol{\beta})\right\}\\ &= \frac{1}{n}\left\{\sum_{\ell=i+1}^{n}\widehat{\mathbf{B}}_{i\ell}^{\mathrm{EFF}}(\boldsymbol{\beta})V_{iji\ell}(\boldsymbol{\beta})+\sum_{\ell=j+1}^{n}\widehat{\mathbf{B}}_{j\ell}^{\mathrm{EFF}}(\boldsymbol{\beta})V_{ijj\ell}(\boldsymbol{\beta})\right.\\ &\qquad\left.+\sum_{k=1}^{i-1}\widehat{\mathbf{B}}_{ki}^{\mathrm{EFF}}(\boldsymbol{\beta})V_{ijki}(\boldsymbol{\beta})+\sum_{k=1}^{j-1}\widehat{\mathbf{B}}_{kj}^{\mathrm{EFF}}(\boldsymbol{\beta})V_{ijkj}(\boldsymbol{\beta})\right\}\end{aligned}$$

since $\widehat{\mathbf{B}}_{ik}^{\mathrm{EFF}}(\boldsymbol{\beta}) = -\widehat{\mathbf{B}}_{ki}^{\mathrm{EFF}}(\boldsymbol{\beta})$ and $V_{ijik}(\boldsymbol{\beta}) = -V_{ijki}(\boldsymbol{\beta})$. A similar argument holds for j.

Next we show that

$$\begin{aligned}&\sum_{k=1}^{i-1}\widehat{\mathbf{B}}_{ki}^{\mathrm{EFF}}(\boldsymbol{\beta})V_{ijki}(\boldsymbol{\beta})+\sum_{k=1}^{j-1}\widehat{\mathbf{B}}_{kj}^{\mathrm{EFF}}(\boldsymbol{\beta})V_{ijkj}(\boldsymbol{\beta})\\ &\qquad=\sum_{k\in K}\sum_{\ell=k+1}^{n}\left\{\widehat{\mathbf{B}}_{k\ell}^{\mathrm{EFF}}(\boldsymbol{\beta})V_{ijk\ell}(\boldsymbol{\beta})\right\}+\widehat{\mathbf{B}}_{ij}^{\mathrm{EFF}}(\boldsymbol{\beta})V_{ijij}(\boldsymbol{\beta})\end{aligned}$$

with $K = \{1,\ldots,n-1\}\backslash\{i,j\}$. Observe that when $\{i,j\}\cap\{k,\ell\} = \emptyset$, the pseudo-observations I_{ij} and $\mathrm{I}_{k\ell}$ are uncorrelated so that $V_{ijk\ell}(\boldsymbol{\beta}) = 0$. Now take $k = 1,\ldots,i-1$. It follows that $V_{ijk\ell}(\boldsymbol{\beta}) = 0$ for $k < \ell < i < j$, $k < i < \ell < j$ and $k < i < j < \ell$. Only when $k < i = \ell < j$ and $k < i < j = \ell$, $V_{ijk\ell}(\boldsymbol{\beta}) \neq 0$, so that

$$\widehat{\mathbf{B}}_{ki}^{\mathrm{EFF}}(\boldsymbol{\beta})V_{ijki}(\boldsymbol{\beta})+\widehat{\mathbf{B}}_{kj}^{\mathrm{EFF}}(\boldsymbol{\beta})V_{ijkj}(\boldsymbol{\beta}) = \sum_{\ell=k+1}^{n}\widehat{\mathbf{B}}_{k\ell}^{\mathrm{EFF}}(\boldsymbol{\beta})V_{ijk\ell}(\boldsymbol{\beta}).$$

Summing up for $k = 1, \ldots, i-1$, we find

$$\sum_{k=1}^{i-1} \left\{ \widehat{\mathbf{B}}_{ki}^{\mathrm{EFF}}(\boldsymbol{\beta}) V_{ijki}(\boldsymbol{\beta}) + \widehat{\mathbf{B}}_{kj}^{\mathrm{EFF}}(\boldsymbol{\beta}) V_{ijkj}(\boldsymbol{\beta}) \right\} = \sum_{k=1}^{i-1} \sum_{\ell=k+1}^{n} \widehat{\mathbf{B}}_{k\ell}^{\mathrm{EFF}}(\boldsymbol{\beta}) V_{ijk\ell}(\boldsymbol{\beta}).$$

Next take $k = i+1, \ldots, j-1$. Now we have that $V_{ijk\ell}(\boldsymbol{\beta}) = 0$ for $i < k < \ell < j$ and $i < k < j < \ell$ and only when $i < k < j = \ell$, $V_{ijk\ell}(\boldsymbol{\beta}) \neq 0$. This implies that

$$\widehat{\mathbf{B}}_{kj}^{\mathrm{EFF}}(\boldsymbol{\beta}) V_{ijkj}(\boldsymbol{\beta}) = \sum_{\ell=k+1}^{n} \widehat{\mathbf{B}}_{k\ell}^{\mathrm{EFF}}(\boldsymbol{\beta}) V_{ijk\ell}(\boldsymbol{\beta}).$$

Summing up for $k = i+1, \ldots, j-1$, we find

$$\sum_{k=i+1}^{j-1} \widehat{\mathbf{B}}_{kj}^{\mathrm{EFF}}(\boldsymbol{\beta}) V_{ijkj}(\boldsymbol{\beta}) = \sum_{k=i+1}^{j-1} \sum_{\ell=k+1}^{n} \widehat{\mathbf{B}}_{k\ell}^{\mathrm{EFF}}(\boldsymbol{\beta}) V_{ijk\ell}(\boldsymbol{\beta}).$$

Finally, for $k = j+1, \ldots, n-1$, $i < j < k < \ell$ so that $V_{ijk\ell}(\boldsymbol{\beta}) = 0$ and hence

$$\sum_{k=j+1}^{n-1} \sum_{\ell=k+1}^{n} \widehat{\mathbf{B}}_{k\ell}^{\mathrm{EFF}}(\boldsymbol{\beta}) V_{ijk\ell}(\boldsymbol{\beta}) = \mathbf{0}.$$

Putting everything together, we find that

$$\begin{aligned} &\sum_{k=1}^{i-1} \widehat{\mathbf{B}}_{ki}^{\mathrm{EFF}}(\boldsymbol{\beta}) V_{ijki}(\boldsymbol{\beta}) + \sum_{k=1}^{j-1} \widehat{\mathbf{B}}_{kj}^{\mathrm{EFF}}(\boldsymbol{\beta}) V_{ijkj}(\boldsymbol{\beta}) \\ &\qquad = \sum_{k \in K} \sum_{\ell=k+1}^{n} \left\{ \widehat{\mathbf{B}}_{k\ell}^{\mathrm{EFF}}(\boldsymbol{\beta}) V_{ijk\ell}(\boldsymbol{\beta}) \right\} + \widehat{\mathbf{B}}_{ij}^{\mathrm{EFF}}(\boldsymbol{\beta}) V_{ijij}(\boldsymbol{\beta}) \end{aligned}$$

with $K = \{1, \ldots, n-1\} \backslash \{i, j\}$. Using this, we find that

$$
\begin{aligned}
&\frac{1}{n}\left\{ \sum_{\ell=i+1}^{n} \widehat{\mathbf{B}}_{i\ell}^{\mathrm{EFF}}(\boldsymbol{\beta}) V_{iji\ell}(\boldsymbol{\beta}) + \sum_{\ell=j+1}^{n} \widehat{\mathbf{B}}_{j\ell}^{\mathrm{EFF}}(\boldsymbol{\beta}) V_{ijj\ell}(\boldsymbol{\beta}) \right. \\
&\qquad \left. + \sum_{k=1}^{i-1} \widehat{\mathbf{B}}_{ki}^{\mathrm{EFF}}(\boldsymbol{\beta}) V_{ijki}(\boldsymbol{\beta}) + \sum_{k=1}^{j-1} \widehat{\mathbf{B}}_{kj}^{\mathrm{EFF}}(\boldsymbol{\beta}) V_{ijkj}(\boldsymbol{\beta}) \right\} \\
&= \frac{1}{n}\left[\sum_{\ell=i+1}^{n} \widehat{\mathbf{B}}_{i\ell}^{\mathrm{EFF}}(\boldsymbol{\beta}) V_{iji\ell}(\boldsymbol{\beta}) + \sum_{\ell=j+1}^{n} \widehat{\mathbf{B}}_{j\ell}^{\mathrm{EFF}}(\boldsymbol{\beta}) V_{ijj\ell}(\boldsymbol{\beta}) \right. \\
&\qquad \left. + \sum_{k \in K} \sum_{\ell=k+1}^{n} \left\{ \widehat{\mathbf{B}}_{k\ell}^{\mathrm{EFF}}(\boldsymbol{\beta}) V_{ijk\ell}(\boldsymbol{\beta}) \right\} + \widehat{\mathbf{B}}_{ij}^{\mathrm{EFF}}(\boldsymbol{\beta}) V_{ijij}(\boldsymbol{\beta}) \right] \\
&= \frac{1}{n}\left[\sum_{k=1}^{n-1} \sum_{\ell=k+1}^{n} \left\{ \widehat{\mathbf{B}}_{k\ell}^{\mathrm{EFF}}(\boldsymbol{\beta}) V_{ijk\ell}(\boldsymbol{\beta}) \right\} + \widehat{\mathbf{B}}_{ij}^{\mathrm{EFF}}(\boldsymbol{\beta}) V_{ijij}(\boldsymbol{\beta}) \right],
\end{aligned}
$$

leading to (4.2).

5. Derivation of the second order bias

Let $\widehat{\boldsymbol{\beta}}^*$ be the solution for $\boldsymbol{\beta}$ of $\mathbf{U}^*(\boldsymbol{\beta}, \boldsymbol{\beta}^*) = \mathbf{0}$ with $\boldsymbol{\beta}^*$ fixed, where

$$\mathbf{U}^*(\boldsymbol{\beta}, \boldsymbol{\beta}^*) = \mathbf{D}^T(\boldsymbol{\beta})\mathbf{V}_{\text{eff}}^{-1}(\boldsymbol{\beta}^*)\{\mathbf{I} - \mathbf{M}(\boldsymbol{\beta})\}.$$

The matrix $\mathbf{V}_{\text{eff}}(\boldsymbol{\beta}^*) = \mathbf{V}(\boldsymbol{\beta}^*) + \mathbf{V}_{\text{indep}}(\boldsymbol{\beta}^*)$ is an $n(n-1)/2 \times n(n-1)/2$ matrix with diagonal elements of the form $2m(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta}^*)\{1-m(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta}^*)\}$ and off-diagonal elements of the form $\text{cov}\{\text{I}(Y \preccurlyeq Y^*), \text{I}(Y \preccurlyeq Y^\dagger)|\mathbf{X}, \mathbf{X}^*, \mathbf{X}^\dagger; \boldsymbol{\beta}^*\}$ and $\text{cov}\{\text{I}(Y \preccurlyeq Y^*), \text{I}(Y^* \preccurlyeq Y^\dagger) \mid \mathbf{X}, \mathbf{X}^*, \mathbf{X}^\dagger; \boldsymbol{\beta}^*\}$, see also Section 2.2 of the main paper. $\mathbf{I}$ and $\mathbf{M}(\boldsymbol{\beta})$ are $n(n-1)/2$-dimensional vectors so that the $[(i-1)(2n-i)/2+j-i]$-th element is given by I_{ij} and $M_{ij}(\boldsymbol{\beta})$ respectively. The conditional covariance $\text{cov}\{\text{I}(Y \preccurlyeq Y^*), \text{I}(Y \preccurlyeq Y^\dagger) \mid \mathbf{X}, \mathbf{X}^*, \mathbf{X}^\dagger; \boldsymbol{\beta}^*\}$ can further be rewritten as

$$h_1^{-1}\{(\mathbf{X}-\mathbf{X}^*)^T\boldsymbol{\beta}^*, (\mathbf{X}-\mathbf{X}^\dagger)^T\boldsymbol{\beta}^*\} - F_1\{(\mathbf{X}-\mathbf{X}^*)^T\boldsymbol{\beta}^*\}F_1\{(\mathbf{X}-\mathbf{X}^\dagger)^T\boldsymbol{\beta}^*\}$$

and $\text{cov}\{\text{I}(Y \preccurlyeq Y^*), \text{I}(Y^* \preccurlyeq Y^\dagger) \mid \mathbf{X}, \mathbf{X}^*, \mathbf{X}^\dagger; \boldsymbol{\beta}\}$ as

$$h_2^{-1}\{(\mathbf{X}-\mathbf{X}^*)^T\boldsymbol{\beta}^*, (\mathbf{X}-\mathbf{X}^\dagger)^T\boldsymbol{\beta}^*\} - F_1\{(\mathbf{X}-\mathbf{X}^*)^T\boldsymbol{\beta}^*\}F_1\{(\mathbf{X}-\mathbf{X}^\dagger)^T\boldsymbol{\beta}^*\}.$$

Here, $F_1(u) = \int F(w-u)dF(w)$, $F(\cdot)$ and $f(\cdot)$ are the conditional cumulative distribution function and probability density function of Y given $\mathbf{X}$, respectively, and the functions $h_1(\cdot)$ and $h_2(\cdot)$ are given by $h_1^{-1}(u, v) =$

$\int\{1-F(w+u)\}\{1-F(w+v)\}dF(w)$ and $h_2^{-1}(u,v) = \int F(w-u)\{1-F(w+v)\}dF(w)$.

The bias-reduced estimator $\widehat{\boldsymbol{\beta}}^{\mathrm{BR}}$ (recall that this is the estimator $\widehat{\boldsymbol{\beta}}^*$ by choosing $\boldsymbol{\beta}^* = \mathbf{0}$) is motivated by the results of Paul and Zhang (2014), who derived bias-corrected estimators in the generalized estimation equations (GEE) context. As GEE can be regarded as quasi-likelihood score equations, Paul and Zhang (2014) treat the generalized estimating functions as if they were likelihood functions and apply the bias correction technique of Cox and Snell (1968) to obtain a closed-form expression for the second-order bias. Here we follow a similar idea. That is, as in Cox and Snell (1968), by taking a higher-order expansion of $\mathbf{U}^*(\boldsymbol{\beta},\boldsymbol{\beta}^*)$, we can derive a closed-form expression for the second-order bias of $\widehat{\boldsymbol{\beta}}^*$. This second-order bias will be a function of $\boldsymbol{\beta}^*$ and our goal is to find the value of $\boldsymbol{\beta}^*$ that minimizes this second-order bias. To this end, let $\mathbf{U}^*(\boldsymbol{\beta},\boldsymbol{\beta}^*) = [U_1^*(\boldsymbol{\beta},\boldsymbol{\beta}^*),\ldots,U_p^*(\boldsymbol{\beta},\boldsymbol{\beta}^*)]^T$ with p the dimension of $\boldsymbol{\beta}$ and let $\kappa_{rk}(\boldsymbol{\beta},\boldsymbol{\beta}^*) = \mathrm{E}\{\partial U_r^*(\boldsymbol{\beta},\boldsymbol{\beta}^*)/\partial\beta_k|\mathbf{X}_1,\ldots,\mathbf{X}_n\}$, $\kappa_{rk\ell}(\boldsymbol{\beta},\boldsymbol{\beta}^*) = \mathrm{E}\{\partial^2 U_r^*(\boldsymbol{\beta},\boldsymbol{\beta}^*)/\partial\beta_k\partial\beta_\ell|\mathbf{X}_1,\ldots,\mathbf{X}_n\}$ and $\kappa_{rk}^{(\ell)}(\boldsymbol{\beta},\boldsymbol{\beta}^*) = \partial\kappa_{rk}(\boldsymbol{\beta},\boldsymbol{\beta}^*)/\partial\beta_\ell$, for $r,k,\ell = 1,\ldots,p$. We further denote by $\mathbf{K}(\boldsymbol{\beta},\boldsymbol{\beta}^*)$ the Fisher information matrix analogue with $-\kappa_{rk}(\boldsymbol{\beta},\boldsymbol{\beta}^*)$ its (r,k)-th element and finally, we let $\kappa^{rk}(\boldsymbol{\beta},\boldsymbol{\beta}^*)$ denote the (r,k)-th element of $\mathbf{K}(\boldsymbol{\beta},\boldsymbol{\beta}^*)^{-1}$, the inverse of the Fisher information matrix analogue. Fol-

lowing Cox and Snell (1968), the second-order bias $b_s(\boldsymbol{\beta}, \boldsymbol{\beta}^*)$ $(s = 1, \ldots, p)$ of $\widehat{\boldsymbol{\beta}}^*$ can be written as

$$b_s(\boldsymbol{\beta}, \boldsymbol{\beta}^*) = \sum_{r=1}^{p} \kappa^{rs}(\boldsymbol{\beta}, \boldsymbol{\beta}^*) \sum_{k=1}^{p} \sum_{\ell=1}^{p} \left\{ \kappa_{rk}^{(\ell)}(\boldsymbol{\beta}, \boldsymbol{\beta}^*) - \frac{1}{2} \kappa_{rk\ell}(\boldsymbol{\beta}, \boldsymbol{\beta}^*) \right\} \kappa^{k\ell}(\boldsymbol{\beta}, \boldsymbol{\beta}^*), \tag{5.1}$$

for $s = 1, \ldots, p$. Now, since $\mathbf{V}_{\text{eff}}^{-1}(\boldsymbol{\beta}^*)$ does not depend on $\boldsymbol{\beta}$, we have that $\partial U_r^*(\boldsymbol{\beta}, \boldsymbol{\beta}^*)/\partial \beta_k = \{\partial \mathbf{D}_r^T(\boldsymbol{\beta})/\partial \beta_k\} \mathbf{V}_{\text{eff}}^{-1}(\boldsymbol{\beta}^*) \{\mathbf{I} - \mathbf{M}(\boldsymbol{\beta})\} - \mathbf{D}_r^T(\boldsymbol{\beta}) \mathbf{V}_{\text{eff}}^{-1}(\boldsymbol{\beta}^*) \{\partial \mathbf{M}(\boldsymbol{\beta})/\partial \beta_k\}$. Since we have $\mathrm{E}\{\mathbf{I} - \mathbf{M}(\boldsymbol{\beta}) \mid \mathbf{X}_1, \ldots, \mathbf{X}_n\} = \mathbf{0}$, we have that $\kappa_{rk}(\boldsymbol{\beta}, \boldsymbol{\beta}^*) = -\mathbf{D}_r^T(\boldsymbol{\beta}) \mathbf{V}_{\text{eff}}^{-1}(\boldsymbol{\beta}^*) \{\partial \mathbf{M}(\boldsymbol{\beta})/\partial \beta_k\}$. Next, $\kappa_{rk\ell}(\boldsymbol{\beta}, \boldsymbol{\beta}^*) = -\{\partial \mathbf{D}_r^T(\boldsymbol{\beta})/\partial \beta_k\} \mathbf{V}_{\text{eff}}^{-1}(\boldsymbol{\beta}^*) \times \{\partial \mathbf{M}(\boldsymbol{\beta})/\partial \beta_\ell\} - \{\partial \mathbf{D}_r^T(\boldsymbol{\beta})/\partial \beta_\ell\} \mathbf{V}_{\text{eff}}^{-1}(\boldsymbol{\beta}^*) \{\partial \mathbf{M}(\boldsymbol{\beta})/\partial \beta_k\} - \mathbf{D}_r^T(\boldsymbol{\beta}) \mathbf{V}_{\text{eff}}^{-1}(\boldsymbol{\beta}^*) \{\partial^2 \mathbf{M}(\boldsymbol{\beta})/\partial \beta_k \beta_\ell\}$ and $\kappa_{rk}^{(\ell)}(\boldsymbol{\beta}, \boldsymbol{\beta}^*) = -\{\partial \mathbf{D}_r^T(\boldsymbol{\beta})/\partial \beta_\ell\} \mathbf{V}_{\text{eff}}^{-1}(\boldsymbol{\beta}^*) \{\partial \mathbf{M}(\boldsymbol{\beta})/\partial \beta_k\} - \mathbf{D}_r^T(\boldsymbol{\beta}) \mathbf{V}_{\text{eff}}^{-1}(\boldsymbol{\beta}^*) \{\partial^2 \mathbf{M}(\boldsymbol{\beta})/\partial \beta_k \beta_\ell\}$. Note that as $\mathbf{V}_{\text{eff}}^{-1}(\boldsymbol{\beta}^*)$ does not depend on $\boldsymbol{\beta}$, it will remain constant while calculating $\kappa_{rk}(\boldsymbol{\beta}, \boldsymbol{\beta}^*)$, $\kappa_{rkl}(\boldsymbol{\beta}, \boldsymbol{\beta}^*)$ and $\kappa_{rk}^{(\ell)}(\boldsymbol{\beta}, \boldsymbol{\beta}^*)$, for $r, k, \ell = 1, \ldots, p$, and will therefore be present in all of them.

Our goal is to minimize $b_s^2(\boldsymbol{\beta}, \boldsymbol{\beta}^*)$ with respect to $\boldsymbol{\beta}^*$, for $s = 1, \ldots, p$. Note, however, that a clear characteristic of $\kappa_{rk}(\boldsymbol{\beta}, \boldsymbol{\beta}^*)$, $\kappa_{rk\ell}(\boldsymbol{\beta}, \boldsymbol{\beta}^*)$ and $\kappa_{rk}^{(\ell)}(\boldsymbol{\beta}, \boldsymbol{\beta}^*)$ is that they all depend on $\boldsymbol{\beta}^*$ via $\mathbf{V}_{\text{eff}}^{-1}(\boldsymbol{\beta}^*)$ only. For example, $\partial \kappa_{rk}(\boldsymbol{\beta}, \boldsymbol{\beta}^*)/\partial \beta_s^* = -\mathbf{D}_r^T(\boldsymbol{\beta}) \{\partial \mathbf{V}_{\text{eff}}^{-1}(\boldsymbol{\beta}^*)/\partial \beta_s^*\} \{\partial \mathbf{M}(\boldsymbol{\beta})/\partial \beta_k\}$, so if we set each element of $\partial \mathbf{V}_{\text{eff}}^{-1}(\boldsymbol{\beta}^*)/\partial \beta_s^*$ to zero, $\partial \kappa_{rk}(\boldsymbol{\beta}, \boldsymbol{\beta}^*)/\partial \beta_s^*$ will also be equal to zero. The same argument holds for $\partial \kappa_{rk\ell}(\boldsymbol{\beta}, \boldsymbol{\beta}^*)/\partial \beta_s^*$ and $\partial \kappa_{rk}^{(\ell)}(\boldsymbol{\beta}, \boldsymbol{\beta}^*)/\partial \beta_s^*$. A trivial solution of $\partial b_s^2(\boldsymbol{\beta}, \boldsymbol{\beta}^*)/\partial \beta_s^* = 0$ is therefore obtained by setting

the derivative of all elements of $\mathbf{V}_{\text{eff}}^{-1}(\boldsymbol{\beta}^*)$ with respect to β_s^* to zero. Our goal is thus to find the value of $\boldsymbol{\beta}^*$ that solves $\partial \mathbf{V}_{\text{eff}}^{-1}(\boldsymbol{\beta}^*)/\partial \beta_s^* = \mathbf{0}$, for $s = 1, \ldots, p$. Finally, as $\partial \mathbf{V}_{\text{eff}}^{-1}(\boldsymbol{\beta}^*)/\partial \beta_s^* = \mathbf{V}_{\text{eff}}^{-1}(\boldsymbol{\beta}^*)\{\partial \mathbf{V}_{\text{eff}}(\boldsymbol{\beta}^*)/\partial \beta_s^*\}\mathbf{V}_{\text{eff}}^{-1}(\boldsymbol{\beta}^*)$, it is enough to set the elements of $\partial \mathbf{V}_{\text{eff}}(\boldsymbol{\beta}^*)/\partial \beta_s^*$ to zero instead. We first consider the diagonal elements of $\mathbf{V}_{\text{eff}}(\boldsymbol{\beta}^*)$. Equating the partial derivatives to zero delivers $\partial[m\left(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta}^*\right)\{1 - m\left(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta}^*\right)\}]/\partial \beta_s^* = \mathbf{0}$ which yields $\partial\left\{m\left(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta}^*\right)\right\}/\partial \beta_s^* \times \{1 - 2m\left(\mathbf{X}, \mathbf{X}^*; \boldsymbol{\beta}^*\right)\} = \mathbf{0}$, $s = 1, \ldots, p$. Thus, for a symmetric link function around zero (e.g., probit or logit), this implies that $(\mathbf{X}^* - \mathbf{X})^T \boldsymbol{\beta}^* = \mathbf{0}$. Now let $\mathbf{Z}_1 = \mathbf{X} - \mathbf{X}^*$ and $\mathbf{Z}_2 = \mathbf{X} - \mathbf{X}^\dagger$. For the off-diagonal elements, we first take the derivative with respect to $\boldsymbol{\beta}^*$ and then show that at $\boldsymbol{\beta}^* = \mathbf{0}$ they are zero. That is, we want to show that at $\beta_s^* = 0$, for $s = 1, \ldots, p$,

$$\frac{\partial}{\partial \beta_s^*} \text{cov}\left\{\text{I}\left(Y \preccurlyeq Y^*\right), \text{I}\left(Y \preccurlyeq Y^\dagger\right) \mid \mathbf{X}, \mathbf{X}^*, \mathbf{X}^\dagger\right\} = 0 \tag{5.2}$$

and

$$\frac{\partial}{\partial \beta_s^*} \text{cov}\left\{\text{I}\left(Y \preccurlyeq Y^*\right), \text{I}\left(Y^* \preccurlyeq Y^\dagger\right) \mid \mathbf{X}, \mathbf{X}^*, \mathbf{X}^\dagger\right\} = 0, \tag{5.3}$$

which is equivalent to

$$\frac{\partial}{\partial \beta_s^*}\left[h_1^{-1}(\mathbf{Z}_1^T \boldsymbol{\beta}^*, \mathbf{Z}_2^T \boldsymbol{\beta}^*)\right] = \frac{\partial}{\partial \beta_s^*}\left[F_1(\mathbf{Z}_1^T \boldsymbol{\beta}^*) F_1(\mathbf{Z}_2^T \boldsymbol{\beta}^*)\right] \tag{5.4}$$

and

$$\frac{\partial}{\partial \beta_s^*}\left[h_2^{-1}(\mathbf{Z}_1^T \boldsymbol{\beta}^*, \mathbf{Z}_2^T \boldsymbol{\beta}^*)\right] = \frac{\partial}{\partial \beta_s^*}\left[F_1(\mathbf{Z}_1^T \boldsymbol{\beta}^*) F_1(\mathbf{Z}_2^T \boldsymbol{\beta}^*)\right]. \tag{5.5}$$

First we take the derivative with respect to β_s^* (for $s = 1, \ldots, p$):

$$\begin{aligned}
\frac{\partial}{\partial \beta_s^*}\left[h_1^{-1}(\mathbf{Z}_1^T\boldsymbol{\beta}^*, \mathbf{Z}_2^T\boldsymbol{\beta}^*)\right] &= -Z_{1,s}\int f(w + \mathbf{Z}_1^T\boldsymbol{\beta}^*)\{1 - F(w + \mathbf{Z}_2^T\boldsymbol{\beta}^*)\}dF(w) \\
&\quad -Z_{2,s}\int \{1 - F(w + \mathbf{Z}_1^T\boldsymbol{\beta}^*)\}f(w + \mathbf{Z}_2^T\boldsymbol{\beta}^*)dF(w);
\end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial \beta_s^*}\left[h_2^{-1}(\mathbf{Z}_1^T\boldsymbol{\beta}^*, \mathbf{Z}_2^T\boldsymbol{\beta}^*)\right] &= -Z_{1,s}\int f(w - \mathbf{Z}_1^T\boldsymbol{\beta}^*)\{1 - F(w + \mathbf{Z}_2^T\boldsymbol{\beta}^*)\}dF(w) \\
&\quad -Z_{2,s}\int \{F(w - \mathbf{Z}_1^T\boldsymbol{\beta}^*)\}f(w + \mathbf{Z}_2^T\boldsymbol{\beta}^*)dF(w);
\end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial \beta_s^*}\left[F_1(\mathbf{Z}_1^T\boldsymbol{\beta}^*)F_1(\mathbf{Z}_2^T\boldsymbol{\beta}^*)\right] &= -Z_{1,s}\int f(w - \mathbf{Z}_1^T\boldsymbol{\beta}^*)dF(w)\int F(w - \mathbf{Z}_2^T\boldsymbol{\beta}^*)dF(w) \\
&\quad -Z_{2,s}\int F(w - \mathbf{Z}_1^T\boldsymbol{\beta}^*)dF(w)\int f(w - \mathbf{Z}_2^T\boldsymbol{\beta}^*)dF(w),
\end{aligned}$$

where $Z_{q,s}$ corresponds to the sth component of $\mathbf{Z}_q$, for $q = 1, 2$. Now let $f(\cdot)$ be an even function. Then,

$$\int f(w)\left\{1 - F(w)\right\}dF(w) = \int f(w)F(-w)dF(w) = \int f(w)F(w)dF(w).$$

Evaluating all previous derivatives at $\boldsymbol{\beta}^* = 0$ and using the equality above leads to:

$$\begin{aligned}
\frac{\partial}{\partial \beta_s^*}\left[h_1^{-1}(\mathbf{Z}_1^T\boldsymbol{\beta}^*, \mathbf{Z}_2^T\boldsymbol{\beta}^*)\right]\Big|_{\boldsymbol{\beta}^*=0} &= -(Z_{1,s} + Z_{2,s})\int f(w)\{1 - F(w)\}dF(w); \\
\frac{\partial}{\partial \beta_s^*}\left[h_2^{-1}(\mathbf{Z}_1^T\boldsymbol{\beta}^*, \mathbf{Z}_2^T\boldsymbol{\beta}^*)\right]\Big|_{\boldsymbol{\beta}^*=0} &= -(Z_{1,s} + Z_{2,s})\int f(w)\{1 - F(w)\}dF(w); \\
\frac{\partial}{\partial \beta_s^*}\left[F_1(\mathbf{Z}_1^T\boldsymbol{\beta}^*)F_1(\mathbf{Z}_2^T\boldsymbol{\beta}^*)\right]\Big|_{\boldsymbol{\beta}^*=0} &= -(Z_{1,s} + Z_{2,s})\int f(w)dF(w)\int F(w)dF(w) \\
&= -\frac{(Z_{1,s} + Z_{2,s})}{2}\int f(w)dF(w).
\end{aligned}$$

Finally, since

$$\int f(w)dF(w) = 2\int f(w)F(w)dF(w),$$

we have that

$$\frac{\partial}{\partial\beta_s^*}\left[F_1(\mathbf{Z}_1^T\boldsymbol{\beta}^*)F_1(\mathbf{Z}_2^T\boldsymbol{\beta}^*)\right]\Big|_{\boldsymbol{\beta}^*=0} = -\left(Z_{1,s}+Z_{2,s}\right)\int f(w)F(w)dF(w)$$

and equations (5.4) and (5.5) are satisfied. We therefore conclude that

$$\frac{\partial}{\partial\beta_s^*}\mathbf{V}_{\text{eff}}(\boldsymbol{\beta}^*)\Big|_{\boldsymbol{\beta}^*=\mathbf{0}} = \mathbf{0}, \quad (s=1,\ldots,p)$$

and so that $\boldsymbol{\beta}^* = \mathbf{0}$ minimizes the bias (5.1).

6. Computational issues

Table 1 demonstrates the performance (computation time in seconds) of the three estimators. We fit a probit-PIM with 5 predictors to subsets of sizes ranging from $n = 25$ to $n = 200$ of data of Section 4 of the main paper. All computations are performed on two devices: a MacBook pro (mid 2014) 2,6 GHz Intel Core i5 with 8 GB 1600 MHz DDR3 RAM and a Dell PowerEdge R900 2.66GHz Intel Xeon CPU X7460 with 132GB 667MHz DDR2 RAM.

Table 1: Computation time (in seconds) as a function of the sample size n.

n	$\widehat{\boldsymbol{\beta}}^{\mathrm{ST}}$	$\widehat{\boldsymbol{\beta}}^{\mathrm{BR}}$	$\widehat{\boldsymbol{\beta}}^{\mathrm{EFF}}$	$\widehat{\boldsymbol{\beta}}^{\mathrm{ST}}$	$\widehat{\boldsymbol{\beta}}^{\mathrm{BR}}$	$\widehat{\boldsymbol{\beta}}^{\mathrm{EFF}}$
		8 GB RAM			132 GB RAM	
			time (seconds)			
25	0.005	0.055	1.585	0.011	0.301	8.627
50	0.007	1.800	23.199	0.020	2.787	45.546
75	0.011	12.996	183.303	0.035	13.852	215.664
100	0.017	64.086	821.156	0.096	33.394	379.219
150	0.039	720.514	10 483.180	0.075	124.264	1 780.878
200	0.071	4 353.420	58 751.820	0.119	404.274	4 321.588

Computation of $\widehat{\boldsymbol{\beta}}^{\mathrm{ST}}$ remains below one second since it only requires

inverting a diagonal matrix. The computation times of $\widehat{\boldsymbol{\beta}}^{\mathrm{BR}}$ and $\widehat{\boldsymbol{\beta}}^{\mathrm{EFF}}$ increase dramatically with the sample size. For $n \geq 100$, the 132 GB RAM machine is substantially faster since it does not need to use virtual memory.

It is clear that for sample sizes exceeding 150 $\widehat{\boldsymbol{\beta}}^{\mathrm{BR}}$ is practically unusable, while for $\widehat{\boldsymbol{\beta}}^{\mathrm{EFF}}$ this is already the case for sample sizes exceeding 100 when there is only 8 GB of memory. With 132 GB of memory and $n = 200$ $\widehat{\boldsymbol{\beta}}^{\mathrm{BR}}$ is still feasible, while $\widehat{\boldsymbol{\beta}}^{\mathrm{EFF}}$ becomes very slow when $n \geq 150$.

7. Partition estimator: proof of Theorem 3

Let S_i denote the set of indices corresponding to the ith part of the partition of the full dataset where $m_i = |S_i|$ and $\sum_{i=1}^{k} m_i = n$ with $m_i \to \infty$ as $n \to \infty$, with k the number of subsets which satisfies $k \to \infty$ and $k/\sqrt{n} \to 0$ as $n \to \infty$. Under model $\mathcal{M}_{\mathrm{PIM}}$, $\widetilde{\boldsymbol{\beta}}$ is a consistent estimator for $\boldsymbol{\beta}_0$ since for every i, $\widehat{\boldsymbol{\beta}}_i = \boldsymbol{\beta}_0 + o_p(1)$ and thus $\widetilde{\boldsymbol{\beta}} = n^{-1} \sum_{i=1}^{k} m_i \widehat{\boldsymbol{\beta}}_i = n^{-1} \sum_{i=1}^{k} m_i \{\boldsymbol{\beta}_0 + o_p(1)\} = \beta_0 + o_p(1)$ since $\sum_{i=1}^{k} m_i = n$. It further holds under $\mathcal{M}_{\mathrm{PIM}}$ that

$$\sqrt{m_i}(\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_0) = \frac{1}{\sqrt{m_i}} \sum_{j \in S_i} \boldsymbol{\varphi}(Y_j, \mathbf{X}_j; \boldsymbol{\beta}_0, \boldsymbol{\eta}_0) + o_p(1) = \frac{1}{\sqrt{m_i}} \sum_{j \in S_i} \boldsymbol{\varphi}(Y_j, \mathbf{X}_j; \boldsymbol{\beta}_0, \boldsymbol{\eta}_0) + O_p(1/\sqrt{m_i}).$$

By the definition of the partition estimator, it now follows that

$$\begin{aligned}
\sqrt{n}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) &= \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{k} m_i \widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_0\right) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{k} m_i \left(\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_0\right) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{k} \sqrt{m_i}\left(\frac{1}{\sqrt{m_i}}\sum_{j\in S_i} \boldsymbol{\varphi}(Y_j, \mathbf{X}_j; \boldsymbol{\beta}_0, \boldsymbol{\eta}_0) + o_p(1)\right) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{k} \sqrt{m_i}\left(\frac{1}{\sqrt{m_i}}\sum_{j\in S_i} \boldsymbol{\varphi}(Y_j, \mathbf{X}_j; \boldsymbol{\beta}_0, \boldsymbol{\eta}_0) + O_p(1/\sqrt{m_i})\right) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{k} \left(\sum_{j\in S_i} \boldsymbol{\varphi}(Y_j, \mathbf{X}_j; \boldsymbol{\beta}_0, \boldsymbol{\eta}_0) + \sqrt{m_i} O_p(1/\sqrt{m_i})\right) \\
&= \frac{1}{\sqrt{n}}\sum_{j=1}^{n} \boldsymbol{\varphi}(Y_j, \mathbf{X}_j; \boldsymbol{\beta}_0, \boldsymbol{\eta}_0) + \sum_{i=1}^{k} \frac{1}{\sqrt{n}} O_p(1) \\
&= \frac{1}{\sqrt{n}}\sum_{j=1}^{n} \boldsymbol{\varphi}(Y_j, \mathbf{X}_j; \boldsymbol{\beta}_0, \boldsymbol{\eta}_0) + o_p(1)
\end{aligned}$$

where the last equality follows since we assume that $k/\sqrt{n} \to 0$ as $n \to \infty$, which is satisfied when $k = o(n^{0.5-\delta})$ with $0 < \delta < 0.5$. From this derivation we see that $\sqrt{n}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ and $\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$, with $\widehat{\boldsymbol{\beta}}$ the estimator applied to the entire dataset, have the same asymptotic distribution.

8. Simulation scenarios

The simulation scenarios used in Section 3 of the main paper, more specifically the values of α and u used to generate data under model (2.11), were carefully chosen to empirically study settings where efficiency gains, when the efficient PIM estimator $\widehat{\boldsymbol{\beta}}^{\text{EFF}}$ is used instead of the simpler estimator $\widehat{\boldsymbol{\beta}}^{\text{ST}}$, can potentially be found. We consider specific choices of α and u to obtain two scenarios: in the first, α and u are chosen so that $\widehat{\boldsymbol{\beta}}^{\text{EFF}}$ is expected to outperform $\widehat{\boldsymbol{\beta}}^{\text{ST}}$ in terms of mean squared error (MSE), while in the second scenario α and u are chosen so that both estimators have a more similar performance. These scenarios are obtained by maximizing (or minimizing) the difference, with respect to the Frobenius norm, between the estimating functions (6) $\widehat{\mathbf{B}}^{\text{ST},T}(\boldsymbol{\beta})\{\mathbf{I}-\mathbf{M}(\boldsymbol{\beta})\}$ with $\mathbf{B}^{\text{ST}}(\boldsymbol{\beta}) = \mathbf{D}^T(\boldsymbol{\beta})\mathbf{V}_{\text{indep}}^{-1}(\boldsymbol{\beta})$, and the estimating functions $\widehat{\mathbf{B}}^{\text{EFF},T}(\boldsymbol{\beta})\{\mathbf{I}-\mathbf{M}(\boldsymbol{\beta})\}$ with $\widehat{\mathbf{B}}^{\text{EFF}}(\boldsymbol{\beta}) = n\mathbf{D}^T(\boldsymbol{\beta})\{\mathbf{V}(\boldsymbol{\beta})+\mathbf{V}_{\text{indep}}(\boldsymbol{\beta})\}^{-1}$, associated to $\widehat{\boldsymbol{\beta}}^{\text{ST}}$ and $\widehat{\boldsymbol{\beta}}^{\text{EFF}}$, respectively. As $\mathbf{X}$ is fixed by design, depending only on the upper value u, the difference between these two estimating functions can be calculated at fixed values of $\boldsymbol{\beta}$ and u and we expect this difference is related to the performance of the two estimators. That is, we expect that gains in efficiency, if any, will increase as the difference between the two estimating functions (and thus between $\widehat{\boldsymbol{\beta}}^{\text{ST}}$ and $\widehat{\boldsymbol{\beta}}^{\text{EFF}}$) increases.

Next we describe the choices of α and u that maximize or minimize the difference between the two estimating functions, when the error term follows either a normal or Gumbel distribution. These values were later used to generate data for the simulation studies described in Section 3 of the main paper.

8.1 Normally distributed data

8.1.1 Rationale

Recall that for the probit-PIM, $\mathbf{B}^{\mathrm{ST}}(\boldsymbol{\beta}) = (\mathbf{X}^*-\mathbf{X})\phi\{(\mathbf{X}^*-\mathbf{X})^T\boldsymbol{\beta}\}\mathbf{V}_{\mathrm{indep}}^{-1}(\boldsymbol{\beta})$ and $\mathbf{B}^{\mathrm{EFF}}(\boldsymbol{\beta}) = (\mathbf{X}^* - \mathbf{X})\phi\{(\mathbf{X}^* - \mathbf{X})^T\boldsymbol{\beta}\}\{\mathbf{V}(\boldsymbol{\beta}) + \mathbf{V}_{\mathrm{indep}}(\boldsymbol{\beta})\}^{-1}$. Thus, the two estimating functions differ by the structure of $\mathbf{V}_{\mathrm{eff}}(\boldsymbol{\beta}) = \mathbf{V}(\boldsymbol{\beta}) + \mathbf{V}_{\mathrm{indep}}(\boldsymbol{\beta})$: $\widehat{\boldsymbol{\beta}}^{\mathrm{ST}}$ ignores the correlation between the pseudo-observations when estimating $\boldsymbol{\beta}$ while $\widehat{\boldsymbol{\beta}}^{\mathrm{EFF}}$ accounts for it. Maximizing this difference is, therefore, equivalent to maximizing the cross-correlation between the pseudo-observations. That is, we must maximize the off-diagonal elements of $\mathbf{V}(\boldsymbol{\beta})$: $\mathrm{cov}\{\mathrm{I}(Y \preccurlyeq Y^*), \mathrm{I}(Y \preccurlyeq Y^\dagger) \mid \mathbf{X}, \mathbf{X}^*, \mathbf{X}^\dagger; \boldsymbol{\beta}\}$ and $\mathrm{cov}\{\mathrm{I}(Y \preccurlyeq Y^*), \mathrm{I}(Y^* \preccurlyeq Y^\dagger) \mid \mathbf{X}, \mathbf{X}^*, \mathbf{X}^\dagger; \boldsymbol{\beta}\}$.

This maximization is a by-product of minimizing the second-order bias of $\widehat{\boldsymbol{\beta}}^{\mathrm{BR}}$ (see Section 5, equations (5.2) and (5.3)) and is achieved at $(\mathbf{X}^* - \mathbf{X})^T\boldsymbol{\beta} = \mathbf{0}$. Therefore, when $\boldsymbol{\beta}_0 = \mathbf{0}$, which is obtained by setting $\boldsymbol{\alpha} = \mathbf{0}$ in (11), we

expect the largest difference in performance (i.e., in MSEs) between $\widehat{\boldsymbol{\beta}}^{\text{EFF}}$ and $\widehat{\boldsymbol{\beta}}^{\text{ST}}$, while when $\boldsymbol{\beta}_0 \neq \mathbf{0}$ we expect that both estimators have similar MSEs.

8.1.2 Simulation results

Table 2 shows, based on 1000 Monte Carlo simulation runs, the average of the estimates, empirical variances, the average of sandwich variance estimates, coverage of 95% CIs and the relative efficiency based on the MSE (relative to $\widehat{\beta}^{\text{EFF}}$, where RE > 1 indicates that $\widehat{\boldsymbol{\beta}}^{\text{EFF}}$ has a lower MSE) of all three estimates, where

$$\text{MSE}(\widehat{\boldsymbol{\beta}}) = \{\text{Av}(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\beta}_0\}^2 + \text{Var}(\widehat{\boldsymbol{\beta}}), \tag{8.1}$$

with $\text{Av}(\cdot)$ and $\text{Var}(\cdot)$ the empirical Monte Carlo mean and variance of the estimators.

Table 3 shows the results when we consider a binary covariate $\mathbf{X} = (X_1, X_2)$, with X_1 and X_2 independent, where X_1 takes equidistant values between 0 and an upper value denoted by u and $X_{2i} \sim U(0.1, 0.8)$ for $i = 1, \ldots, \lfloor n/2 \rfloor$ and $X_{2i} \sim U(0.2, 1)$ for $i = \lfloor n/2 \rfloor + 1, \ldots, n$. The minor differences in MSE between the estimators when a single covariate was considered, are now practically vanished when the covariate is bivariate.

Table 2: Simulation results based on 1000 Monte Carlo runs, for model (2.11) with a standard normal error distribution.

Estimators	$\mathrm{Av}(\widehat{\beta})$	$\mathrm{Var}(\widehat{\beta})$	$\mathrm{Av}\{\widehat{\Sigma}(\widehat{\beta})\}$	Cov(%)	$\mathrm{RE}_{\mathrm{eff}}$	$\mathrm{Av}(\widehat{\beta})$	$\mathrm{Var}(\widehat{\beta})$	$\mathrm{Av}\{\widehat{\Sigma}(\widehat{\beta})\}$	Cov(%)	$\mathrm{RE}_{\mathrm{eff}}$
	$\beta_0 = 0$					$\beta_0 = \sqrt{2} = 1.414$				
					$n = 25$					
$\widehat{\beta}^{\mathrm{ST}}$	0.005	0.068	0.057	92.0	1.004	1.497	0.135	0.092	88.2	1.031
$\widehat{\beta}^{\mathrm{BR}}$	0.005	0.067	0.058	91.3	0.994	1.427	0.153	0.132	89.5	1.114
$\widehat{\beta}^{\mathrm{EFF}}$	0.005	0.068	0.057	91.3	1.000	1.489	0.132	0.095	88.5	1.000
					$n = 50$					
$\widehat{\beta}^{\mathrm{ST}}$	0.002	0.030	0.029	94.5	1.002	1.449	0.060	0.048	92.6	1.020
$\widehat{\beta}^{\mathrm{BR}}$	0.002	0.029	0.029	94.6	0.998	1.413	0.071	0.063	94.4	1.192
$\widehat{\beta}^{\mathrm{EFF}}$	0.002	0.030	0.029	94.5	1.000	1.444	0.059	0.048	93.9	1.000
					$n = 100$					
$\widehat{\beta}^{\mathrm{ST}}$	0.000	0.015	0.014	95.4	1.001	1.443	0.030	0.027	93.8	1.024
$\widehat{\beta}^{\mathrm{BR}}$	0.000	0.015	0.014	95.3	0.995	1.415	0.036	0.030	94.3	1.207
$\widehat{\beta}^{\mathrm{EFF}}$	0.000	0.015	0.015	95.2	1.000	1.430	0.030	0.027	94.4	1.000

NOTE: $\mathrm{Av}(\widehat{\beta})$: average of the β estimates; $\mathrm{Var}(\widehat{\beta})$: Monte Carlo variance of $\widehat{\beta}$; $\mathrm{Av}(\widehat{\Sigma}(\widehat{\beta}))$: average of the sandwich variance estimates; Cov(%): coverage of 95% CIs; $\mathrm{RE}_{\mathrm{eff}}$: relative efficiency compared to $\widehat{\beta}^{\mathrm{EFF}}$ in terms of MSE.

Table 3: Simulation results based on 1000 Monte Carlo runs, for model (2.11) with a standard normal error distribution and $u = 2$.

Estimators	$\mathrm{Av}(\widehat{\boldsymbol{\beta}})$	$\mathrm{Var}(\widehat{\boldsymbol{\beta}})$	$\mathrm{Av}\{\widehat{\Sigma}(\widehat{\boldsymbol{\beta}})\}$	Cov(%)	$\mathrm{RE}_{\mathrm{eff}}$	$\mathrm{Av}(\widehat{\boldsymbol{\beta}})$	$\mathrm{Var}(\widehat{\boldsymbol{\beta}})$	$\mathrm{Av}\{\widehat{\Sigma}(\widehat{\boldsymbol{\beta}})\}$	Cov(%)	$\mathrm{RE}_{\mathrm{eff}}$
	$\boldsymbol{\beta}_0 = (0, 0)$					$\boldsymbol{\beta}_0 = (0.707, 0.707)$				
	$n = 40$									
$\widehat{\boldsymbol{\beta}}^{\mathrm{ST}}$	0.008	0.220	0.178	91.7	1.010	0.739	0.209	0.173	91.0	1.009
	0.005	0.307	0.280	93.7	1.017	0.759	0.261	0.261	95.5	1.007
$\widehat{\boldsymbol{\beta}}^{\mathrm{BR}}$	0.007	0.220	0.181	92.7	1.008	0.734	0.210	0.184	93.5	1.013
	0.008	0.301	0.283	94.1	0.999	0.754	0.266	0.270	95.5	1.024
$\widehat{\boldsymbol{\beta}}^{\mathrm{EFF}}$	0.008	0.218	0.186	92.5	1.000	0.737	0.207	0.173	92.0	1.000
	0.007	0.302	0.294	94.2	1.000	0.757	0.259	0.262	94.5	1.000
	$n = 80$									
$\widehat{\boldsymbol{\beta}}^{\mathrm{ST}}$	0.006	0.103	0.089	92.4	0.998	0.728	0.092	0.083	93.9	0.993
	0.010	0.156	0.138	93.6	0.992	0.753	0.135	0.128	95.6	1.001
$\widehat{\boldsymbol{\beta}}^{\mathrm{BR}}$	0.007	0.103	0.089	92.9	1.005	0.730	0.101	0.090	94.4	1.098
	0.010	0.158	0.139	93.4	1.006	0.749	0.149	0.136	94.6	1.099
$\widehat{\boldsymbol{\beta}}^{\mathrm{EFF}}$	0.007	0.103	0.091	92.1	1.000	0.730	0.092	0.084	93.9	1.000
	0.010	0.157	0.142	93.4	1.000	0.756	0.136	0.129	94.8	1.000

NOTE: $\mathrm{Av}(\widehat{\boldsymbol{\beta}})$: average of the β estimates; $\mathrm{Var}(\widehat{\boldsymbol{\beta}})$: Monte Carlo variance of $\widehat{\boldsymbol{\beta}}$; $\mathrm{Av}(\widehat{\Sigma}(\widehat{\boldsymbol{\beta}}))$: average of the sandwich variance estimates; Cov(%): coverage of 95% CIs; $\mathrm{RE}_{\mathrm{eff}}$: relative efficiency compared to $\widehat{\boldsymbol{\beta}}^{\mathrm{EFF}}$ in terms of MSE.

8.2 Gumbel distributed data

8.2.1 Rationale

Recall that for the logit-PIM, $\mathbf{B}^{\mathrm{ST}}(\boldsymbol{\beta}) = (\mathbf{X}^* - \mathbf{X})$ and $\mathbf{B}^{\mathrm{EFF}}(\boldsymbol{\beta}) = (\mathbf{X}^* - \mathbf{X})\mathrm{expit}\{(\mathbf{X}^* - \mathbf{X})^T\boldsymbol{\beta}\}[1 - \mathrm{expit}\{(\mathbf{X}^* - \mathbf{X})^T\boldsymbol{\beta}\}]\{\mathbf{V}(\boldsymbol{\beta}) + \mathbf{V}_{\mathrm{indep}}(\boldsymbol{\beta})\}^{-1}$. The difference between the two estimating functions depends not only on the structure of $\mathbf{V}(\boldsymbol{\beta})$, as in the probit-PIM, but also on $\mathrm{expit}\{(\mathbf{X}^* - \mathbf{X})^T\boldsymbol{\beta}\}$. As this is substantially more complicated than before, the difference between them was maximized, with respect to $\boldsymbol{\beta}$ and u, the upper limit of the $\mathbf{X}$-values, numerically. Figure 1 shows that, for $n = 25$ and 50, the difference between the two estimating functions, and therefore between the two estimators $\widehat{\boldsymbol{\beta}}^{\mathrm{ST}}$ and $\widehat{\boldsymbol{\beta}}^{\mathrm{EFF}}$, increases as the product $(\mathbf{X}^* - \mathbf{X})^T\boldsymbol{\beta}$ deviates from zero. That is, we expect that regions away from the origin are more likely to lead to more efficient estimates when the efficient PIM estimator $\widehat{\boldsymbol{\beta}}^{\mathrm{EFF}}$ is used instead of $\widehat{\boldsymbol{\beta}}^{\mathrm{ST}}$. When $\boldsymbol{\alpha}$ (and thus $\boldsymbol{\beta}_0$) is zero, gains in efficiency are expected to be smaller, if any.

8.2.2 Simulation results

Table 4 shows the average of the estimates, empirical variances, the average of sandwich variance estimates, coverage of 95% CIs and the relative efficiency (relative to $\widehat{\boldsymbol{\beta}}^{\mathrm{EFF}}$) for different values of α (where $\widehat{\boldsymbol{\beta}}^{\mathrm{EFF}}$ is expected to

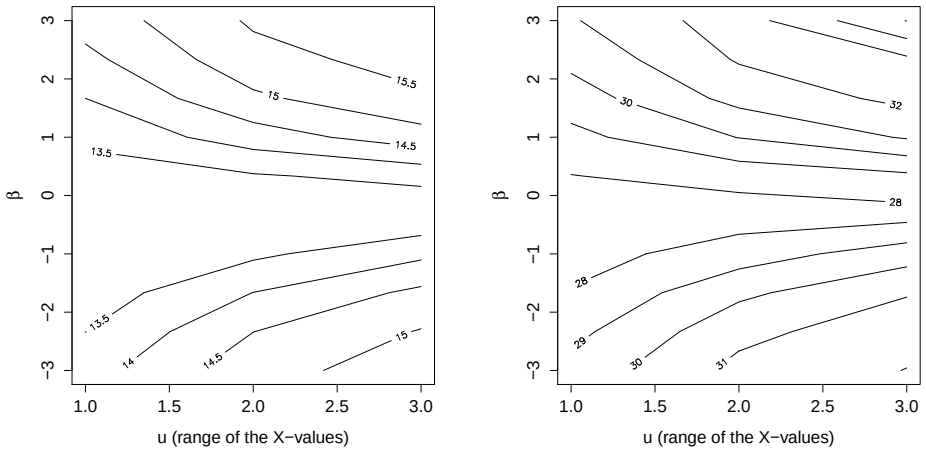

Figure 1: Contour plots for the Frobenius norm of the difference between the estimating functions of $\widehat{\boldsymbol{\beta}}^{\mathrm{ST}}$ and $\widehat{\boldsymbol{\beta}}^{\mathrm{EFF}}$, with respect to $\boldsymbol{\beta}$ and u, the range of the $\mathbf{X}$-values, for $n = 25$ (left panel) and $n = 50$ (right panel) for the logit-PIM (Gumbel distributed data).

perform better than $\widehat{\boldsymbol{\beta}}^{\mathrm{ST}}$ when $\alpha \neq 0$) and sample sizes, after 1000 Monte Carlo runs and with u fixed at 2.

Table 4: Simulation results based on 1000 Monte Carlo runs, for model (2.11) with a Gumbel distributed error and u set to 2.

Estimators	$\mathrm{Av}(\widehat{\beta})$	$\mathrm{Var}(\widehat{\beta})$	$\mathrm{Av}\{\widehat{\Sigma}(\widehat{\beta})\}$	Cov(%)	$\mathrm{RE}_{\mathrm{eff}}$	$\mathrm{Av}(\widehat{\beta})$	$\mathrm{Var}(\widehat{\beta})$	$\mathrm{Av}\{\widehat{\Sigma}(\widehat{\beta})\}$	Cov(%)	$\mathrm{RE}_{\mathrm{eff}}$
			$\beta_0 = 0$					$\beta_0 = 2$		
					$n = 25$					
$\widehat{\beta}^{\mathrm{ST}}$	0.030	0.180	0.151	93.0	0.996	2.169	0.456	0.304	88.4	1.195
$\widehat{\beta}^{\mathrm{BR}}$	0.030	0.179	0.154	93.0	0.989	2.018	0.397	0.362	89.4	0.981
$\widehat{\beta}^{\mathrm{EFF}}$	0.032	0.181	0.152	92.6	1.000	2.113	0.393	0.309	90.5	1.000
$\widehat{\beta}^{\mathrm{PH}}$	−0.021	0.153	0.137	95.5	0.844	2.115	0.340	0.319	95.8	0.872
					$n = 50$					
$\widehat{\beta}^{\mathrm{ST}}$	0.000	0.083	0.075	95.0	1.001	2.077	0.192	0.161	91.3	1.207
$\widehat{\beta}^{\mathrm{BR}}$	−0.001	0.083	0.075	95.2	1.000	1.995	0.185	0.173	92.5	1.129
$\widehat{\beta}^{\mathrm{EFF}}$	0.000	0.083	0.075	94.8	1.000	2.037	0.163	0.150	92.6	1.000
$\widehat{\beta}^{\mathrm{PH}}$	0.001	0.064	0.062	95.2	0.777	2.046	0.132	0.134	94.9	0.816
					$n = 100$					
$\widehat{\beta}^{\mathrm{ST}}$	0.010	0.040	0.037	94.2	0.995	2.016	0.088	0.083	93.4	1.205
$\widehat{\beta}^{\mathrm{BR}}$	0.010	0.041	0.037	93.7	1.005	1.987	0.083	0.087	95.0	1.137
$\widehat{\beta}^{\mathrm{EFF}}$	0.010	0.041	0.037	94.4	1.000	1.985	0.073	0.074	94.1	1.000
$\widehat{\beta}^{\mathrm{PH}}$	0.006	0.031	0.029	94.2	0.765	2.035	0.059	0.062	94.9	0.822

NOTE: $\mathrm{Av}(\widehat{\beta})$: average of the β estimates; $\mathrm{Var}(\widehat{\beta})$: Monte Carlo variance of $\widehat{\beta}$; $\mathrm{Av}(\widehat{\Sigma}(\widehat{\beta}))$: average of the sandwich variance estimates; Cov(%): coverage of 95% CIs; $\mathrm{RE}_{\mathrm{eff}}$: relative efficiency compared to $\widehat{\beta}^{\mathrm{EFF}}$ in terms of MSE.

Results for different values of $\boldsymbol{\beta}$ for a bivariate covariate are displayed in Table 5. Similar as for the probit PIMs, the difference between the PIM estimators practically vanishes when a bivariate covariate is considered.

8.3 Partition estimator

Table 6 shows the results when the data generating model of Section 2.4 is considered for which $k = \lfloor n^{0.25} \rfloor$ partitions are used for $\widetilde{\boldsymbol{\beta}}^{\mathrm{ST}}$. For $n \geq 500$ the partition estimator $\widetilde{\boldsymbol{\beta}}^{\mathrm{ST}}$ is almost as efficient as $\widehat{\boldsymbol{\beta}}^{\mathrm{ST}}$. The partition variance estimator exhibits a slight underestimation, but this reduces with increasing sample size. Overall we can say that the distributions of $\widetilde{\boldsymbol{\beta}}^{\mathrm{ST}}$ and $\widehat{\boldsymbol{\beta}}^{\mathrm{ST}}$ are approximately equal for $n \geq 500$.

Table 5: Simulation results based on 1000 Monte Carlo runs, for model (2.11) with a Gumbel error and $u = 1$.

Estimators	$\mathrm{Av}(\widehat{\boldsymbol{\beta}})$	$\mathrm{Var}(\widehat{\boldsymbol{\beta}})$	$\mathrm{Av}\{\widehat{\Sigma}(\widehat{\boldsymbol{\beta}})\}$	Cov(%)	$\mathrm{RE}_{\mathrm{eff}}$	$\mathrm{Av}(\widehat{\boldsymbol{\beta}})$	$\mathrm{Var}(\widehat{\boldsymbol{\beta}})$	$\mathrm{Av}\{\widehat{\Sigma}(\widehat{\boldsymbol{\beta}})\}$	Cov(%)	$\mathrm{RE}_{\mathrm{eff}}$
	$\boldsymbol{\beta}_0 = (0,0)$					$\boldsymbol{\beta}_0 = (1,1)$				
					$n = 40$					
$\widehat{\boldsymbol{\beta}}^{\mathrm{ST}}$	−0.012	0.521	0.458	93.5	1.039	1.041	0.572	0.495	94.0	1.044
	0.013	0.793	0.717	93.7	1.062	1.048	0.860	0.774	94.1	1.067
$\widehat{\boldsymbol{\beta}}^{\mathrm{BR}}$	−0.007	0.502	0.456	93.7	1.001	1.029	0.567	0.494	93.3	1.033
	0.011	0.746	0.708	93.6	0.999	1.019	0.811	0.764	93.4	1.004
$\widehat{\boldsymbol{\beta}}^{\mathrm{EFF}}$	−0.008	0.502	0.468	92.6	1.000	1.032	0.549	0.503	93.5	1.000
	0.010	0.747	0.727	93.6	1.000	1.024	0.807	0.775	93.9	1.000
$\widehat{\boldsymbol{\beta}}^{\mathrm{PH}}$	0.013	0.447	0.403	95.1	0.891	1.038	0.460	0.428	95.1	0.840
	−0.006	0.660	0.637	95.1	0.883	1.044	0.696	0.666	95.2	0.864
					$n = 80$					
$\widehat{\boldsymbol{\beta}}^{\mathrm{ST}}$	0.022	0.257	0.228	93.6	0.996	1.059	0.283	0.247	93.8	1.004
	−0.055	0.359	0.353	94.9	1.027	0.969	0.401	0.380	94.8	1.083
$\widehat{\boldsymbol{\beta}}^{\mathrm{BR}}$	0.022	0.260	0.228	94.2	1.009	1.065	0.314	0.254	93.2	1.116
	−0.055	0.352	0.352	94.6	1.007	0.957	0.387	0.374	94.7	1.047
$\widehat{\boldsymbol{\beta}}^{\mathrm{EFF}}$	0.022	0.258	0.234	93.8	1.000	1.057	0.282	0.253	93.6	1.000
	−0.057	0.350	0.357	94.5	1.000	0.954	0.369	0.360	95.0	1.000
$\widehat{\boldsymbol{\beta}}^{\mathrm{PH}}$	−0.015	0.203	0.185	94.7	0.787	1.054	0.217	0.197	94.2	0.773
	0.044	0.277	0.285	96.1	0.789	0.984	0.305	0.299	95.3	0.822

NOTE: $\mathrm{Av}(\widehat{\boldsymbol{\beta}})$: average of the $\boldsymbol{\beta}$ estimates; $\mathrm{Var}(\widehat{\boldsymbol{\beta}})$: Monte Carlo variance of $\widehat{\boldsymbol{\beta}}$; $\mathrm{Av}(\widehat{\Sigma}(\widehat{\boldsymbol{\beta}}))$: average of the sandwich variance estimates; Cov(%): coverage of 95% CIs; $\mathrm{RE}_{\mathrm{eff}}$: relative efficiency compared to $\widehat{\boldsymbol{\beta}}^{\mathrm{EFF}}$ in terms of MSE.

Table 6: Comparison of $\widehat{\beta}^{\mathrm{ST}}$ and $\widetilde{\beta}^{\mathrm{ST}}$ when $\beta_0 = 2$ and based on 1000 Monte-Carlo simulations with $k = \lfloor n^{0.25} \rfloor$ partitions.

n	Av($\widehat{\beta}^{\mathrm{ST}}$)	Var($\widehat{\beta}^{\mathrm{ST}}$)	Av{$\widehat{\Sigma}(\widehat{\beta}^{\mathrm{ST}})$}	Cov(%)	Av($\widetilde{\beta}^{\mathrm{ST}}$)	Var($\widetilde{\beta}^{\mathrm{ST}}$)	Av{$\widehat{\Sigma}(\widetilde{\beta}^{\mathrm{ST}})$}	Cov(%)	$\mathrm{RE}_{\mathrm{eff}}$
250	2.022	0.0346	0.0343	94.80	2.052	0.0374	0.0335	92.90	1.1447
500	2.005	0.0172	0.0172	94.00	2.025	0.0176	0.0169	94.30	1.0604
1000	1.996	0.0085	0.0086	95.20	2.010	0.0087	0.0085	94.60	1.0321
2000	2.001	0.0044	0.0043	94.90	2.009	0.0044	0.0043	94.10	1.0345
5000	1.999	0.0017	0.0017	95.90	2.004	0.0017	0.0017	95.80	1.0211

NOTE: Av($\widehat{\beta}$): average of the β estimates; Var($\widehat{\beta}$): Monte Carlo variance of $\widehat{\beta}$; Av($\widehat{\Sigma}(\widehat{\beta})$): average of the sandwich variance estimates; Cov(%): coverage of 95% CIs; $\mathrm{RE}_{\mathrm{eff}}$: relative efficiency compared to $\widehat{\beta}^{\mathrm{ST}}$ in terms of MSE.

9. Cox partial likelihood estimator

To better understand the restrictions imposed by the probabilistic index models, we also added $\widehat{\boldsymbol{\beta}}^{\mathrm{PH}}$, the Cox partial likelihood estimator, to the simulation study. As noted in Section 3.2 of the main paper, a semiparametric transformation model with Gumbel error (location parameter zero and scale parameter one),

$$H(Y) = \mathbf{X}^T\boldsymbol{\beta} + \varepsilon,$$

is equivalent to the Cox proportional hazard model,

$$\lambda(y \mid \mathbf{X}) = \lambda_0(y) \times \exp(\mathbf{X}^T\boldsymbol{\beta}),$$

where $\lambda(y \mid \mathbf{X})$ is the hazard rate function and $\lambda_0(y)$ is the baseline hazard rate, in which case $\widehat{\boldsymbol{\beta}}^{\mathrm{PH}}$ is the efficient estimator under this more restrictive semiparametric transformation model.

To see that the semiparametric transformation model with Gumbel error is more restrictive than the logit-PIM $\mathrm{logit}\{\mathrm{P}(Y \preceq Y^* \mid \mathbf{X}, \mathbf{X}^*)\} = (\mathbf{X}^* - \mathbf{X})^T\boldsymbol{\beta}$, we consider the estimating equation of $\widehat{\boldsymbol{\beta}}^{\mathrm{PH}}$ for the Cox model (in the absence of censoring), see Tsiatis (2006), p.126:

$$\sum_{i=1}^{n}\left[\mathbf{X}_i - \frac{\sum_{j=1}^{n}\mathbf{X}_j \exp(\mathbf{X}_j^T\boldsymbol{\beta})\mathrm{I}(Y_i < Y_j)}{\sum_{j=1}^{n}\exp(\mathbf{X}_j^T\boldsymbol{\beta})\mathrm{I}(Y_i < Y_j)}\right] = \mathbf{0},$$

which can be rewritten as

$$\sum_{i=1}^{n} \frac{\sum_{j=1}^{n} \exp(\mathbf{X}_j^T \boldsymbol{\beta}) \mathrm{I}(Y_i < Y_j)(\mathbf{X}_i - \mathbf{X}_j)}{\sum_{j=1}^{n} \exp(\mathbf{X}_j^T \boldsymbol{\beta}) \mathrm{I}(Y_i < Y_j)} = \mathbf{0}.$$

This estimating function is unbiased under the Cox proportional hazard model, but is not guaranteed to be unbiased under the PIM. One can see this because the above form suggests that evaluation of the mean of the above estimating function demands knowledge about the joint distribution of the pseudo-observations $\mathrm{I}(Y_i < Y_j)$. This is information not subsumed by the PIM, suggesting that $\widehat{\boldsymbol{\beta}}^{\mathrm{PH}}$ is not contained in the class of RAL estimators for $\boldsymbol{\beta}$ under the PIM. It is at least as efficient as $\widehat{\boldsymbol{\beta}}^{\mathrm{EFF}}$ by relying on more stringent restrictions on the observed data law.

10. Illustration

Figure 2 shows the residual and normal QQ-plot when analysing the data according to least squares regression. Since the residual plot shows a non-constant variance and the QQ-plot indicates a violation of the normality assumption, a Box–Cox transformation is considered. The upper panel of Figure 3 displays the Box–Cox transformation. Figure 3 (lower panels) indicates that the Box–Cox transformation stabilizes the variance and that the residuals are approximately normally distributed.

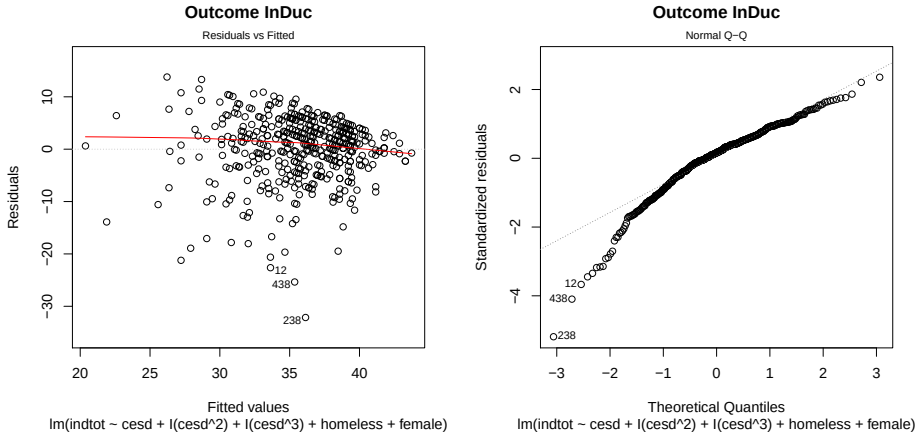

Figure 2: Residual and normal QQ-plot of least-squares regression of the original outcome.

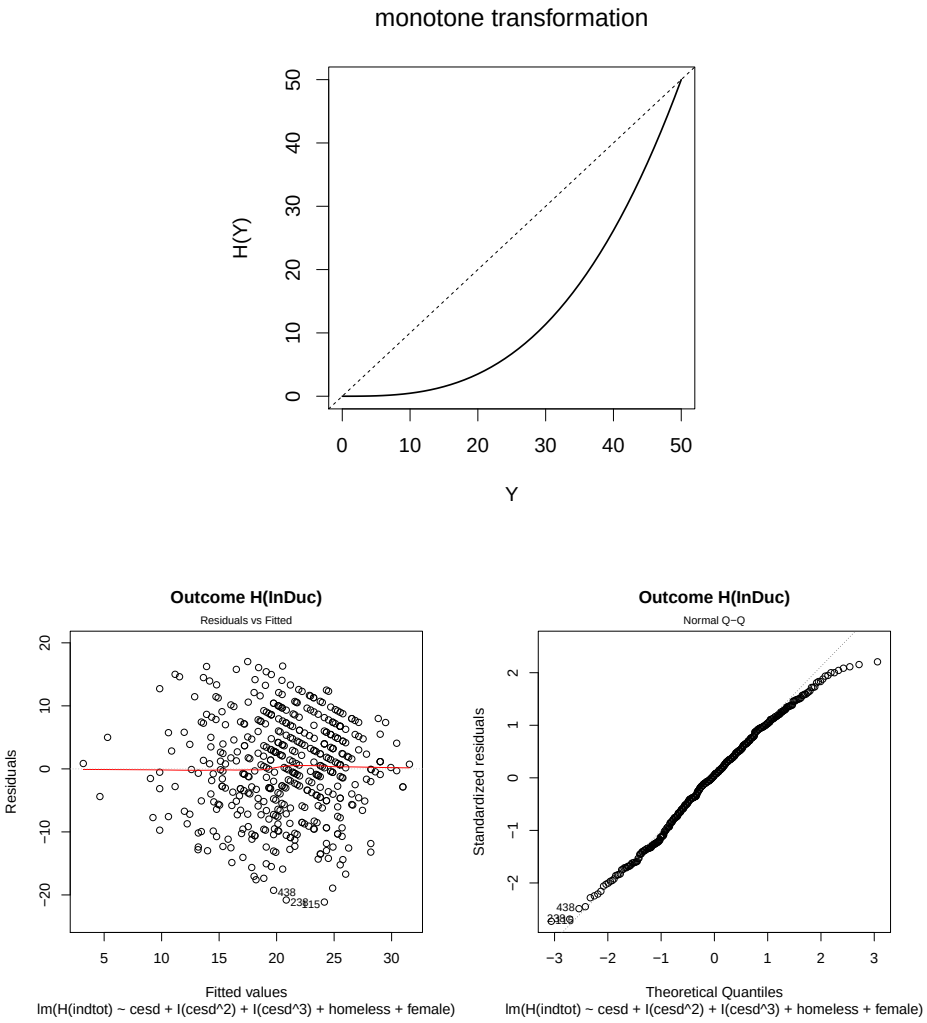

Figure 3: Residual and normal QQ-plot of least-squares regression of the transformed outcome (lower panels) according to a Box–Cox power transformation (upper panel).

11. Working covariance structure versus true covariance structure of pseudo-observations

In the paper, we showed that $\widehat{\boldsymbol{\beta}}^{\mathrm{ST}}$ can be obtained from $\mathbf{B}^{\mathrm{EFF}}(\boldsymbol{\beta})$ by setting $\mathbf{V}(\boldsymbol{\beta}) = \mathbf{0}$ and $\mathbf{B}^{\mathrm{BR}}(\boldsymbol{\beta})$ is obtained by considering $V_{ijkl}(\boldsymbol{\beta})$ with a fixed $\boldsymbol{\beta} = \boldsymbol{\beta}^*$ chosen to minimize the second-order finite sample bias. In some of the simulations when $\boldsymbol{\beta}_0 \neq \mathbf{0}$, $\widehat{\boldsymbol{\beta}}^{\mathrm{BR}}$ is less efficient than $\widehat{\boldsymbol{\beta}}^{\mathrm{ST}}$. As the parameter of interest $\boldsymbol{\beta}_0$ deviates from zero, the bias-reduced PIM estimator $\widehat{\boldsymbol{\beta}}^{\mathrm{BR}}$ behaves worse than $\widehat{\boldsymbol{\beta}}^{\mathrm{ST}}$ which assumes an independence covariance structure. This is because as we deviate more from the null ($\boldsymbol{\beta}_0 = \mathbf{0}$), the covariance structure used in the construction of $\widehat{\boldsymbol{\beta}}^{\mathrm{BR}}$ becomes more *misspecified* than that used in the construction of $\widehat{\boldsymbol{\beta}}^{\mathrm{ST}}$.

More intuition into this behavior can be garnered as follows. We focus on the following semiparametric transformation model: $H(Y) = \alpha X + \varepsilon$ and consider the X-values to be fixed and $\alpha > 0$, delivering the PIM in equation (12) of the revised manuscript. Recall that the off-diagonal elements of the covariance matrix $\mathbf{V}(\beta)$ are given by

$$\begin{aligned} V_{ijik}(\beta) &= \mathrm{P}(Y_i < \min(Y_j, Y_k) \mid X_i, X_j, X_k) - M_{ij}(\beta) M_{ik}(\beta), \\ V_{ijjk}(\beta) &= \mathrm{P}(Y_i < Y_j < Y_k \mid X_i, X_j, X_k) - M_{ij}(\beta) M_{jk}(\beta). \end{aligned}$$

Using the semiparametric transformation model, we obtained model-based

expressions for $\mathrm{P}(Y_i < \min(Y_j, Y_k) \mid X_i, X_j, X_k)$ and $\mathrm{P}(Y_i < Y_j < Y_k \mid X_i, X_j, X_k)$, see equations (13) and (14) of the manuscript. When α (and thus β) becomes larger and the X-values are fixed, the Y-values also become more separated and their order becomes completely determined by the order of the X-values, since the error-term becomes less important for larger α-values. This implies that for instance $\mathrm{P}(Y_i < \min(Y_j, Y_k) \mid X_i, X_j, X_k) \to 1$ when $X_i < \min(X_j, X_k)$ since the distance between the Y-values increases (since the error-term becomes less important as argued above). Similarly, as we deviate more from the null, we also have that $\mathrm{P}(Y_i < Y_j \mid X_i, X_j) \to 1$ for $X_i < X_j$. We thus find that in this case $V_{ijik}(\beta) \to 1 - 1 = 0$. If we are not in this case, then both $\mathrm{P}(Y_i < \min(Y_j, Y_k) \mid X_i, X_j, X_k)$ and $\mathrm{P}(Y_i < Y_j \mid X_i, X_j)$ go to zero. We thus find that $V_{ijik}(\beta)$ approaches zero as we deviate more from the null. A similar reasoning can be made for $V_{ijjk}(\beta)$.

In conclusion: the off-diagonal elements of the covariance matrix go to zero as we deviate more from the null ($\beta_0 = 0$). This is why $\widehat{\beta}^{\mathrm{ST}}$ (assuming independence covariance structure) turns out to be more efficient than $\widehat{\beta}^{\mathrm{BR}}$ in such cases (indeed, for the bias-reduced PIM estimator, the covariance matrix is evaluated at $\beta^* = 0$ so that the off-diagonal elements get values $1/12$ or $-1/12$).

We illustrate the above arguments by means of some simulation experiments. We consider a sample size $n = 25$ and we generate data from a normal linear model $Y_i = \alpha X_i + \varepsilon_i$ with $\varepsilon_i \sim N(0, 1)$ for $i = 1, \ldots, 25$ with X_i (fixed throughout the simulations) taking equally spaced values between 0 and 1 for varying values of α, and thus for varying values of $\beta = \alpha/\sqrt{2}$. Specifically, we calculate the off-diagonal elements $V_{ijik}(\beta)$ and $V_{ijjk}(\beta)$ with $i = 1$, $j = 4$ and $k = 25$. The results of these off-diagonal elements are shown in Figure 4. The solid line corresponds to the *true* covariance structure, in which case $\widehat{\beta}^{\mathrm{EFF}}$ is used and this depends on the true value of β. The dashed line corresponds to $\widehat{\beta}^{\mathrm{BR}}$, in which case $V_{ijjk}(\beta)$ (left) and $V_{ijik}(\beta)$ (right) are constant ($1/12$ or $-1/12$) since the covariance does not change here as a function of β. The dotted line corresponds to $\widehat{\beta}^{\mathrm{ST}}$, in which case $V_{ijjk}(\beta)$ (left) and $V_{ijik}(\beta)$ (right) are constant and both equal to zero. We focus on the left panel of the figure. We observe that when the true parameter β_0 is close to zero (so we are close to the null), the covariance structure used for $\widehat{\beta}^{\mathrm{BR}}$, is close to the true covariance structure used for $\widehat{\beta}^{\mathrm{EFF}}$ since the covariance structure used by $\widehat{\beta}^{\mathrm{BR}}$ then closely resembles the true covariance structure used by $\widehat{\beta}^{\mathrm{EFF}}$. Next, when the true β_0 deviates from zero, the off-diagonal element of the true covariance structure decreases as β deviates more from zero. In this case, the

independence covariance structure used by $\widehat{\beta}^{\mathrm{ST}}$ now better approximates the true covariance structure (since the latter approaches the zero line). A similar reasoning can be made using the right panel.

To get an overall idea of how well the different covariance structures compare to each other on the whole, in Figure 5, we compare the Frobenius norm of the inverse of the corresponding covariance structure $\mathbf{V}(\beta)$ as a function of β for $\widehat{\beta}^{\mathrm{EFF}}$ (solid line), $\widehat{\beta}^{\mathrm{BR}}$ (dotted line) and $\widehat{\beta}^{\mathrm{ST}}$ (dashed line). When the true parameter β_0 is close to zero, we see close correspondence between the Frobenius norm of the inverse covariance matrix when $\widehat{\beta}^{\mathrm{BR}}$ or $\widehat{\beta}^{\mathrm{EFF}}$ is used and where that of $\widehat{\beta}^{\mathrm{ST}}$ is deviating. As the true β_0 deviates from zero, the Frobenius norm of the inverse of the covariance matrices obtained from the $\widehat{\beta}^{\mathrm{ST}}$ and $\widehat{\beta}^{\mathrm{EFF}}$ become more similar than that obtained from $\widehat{\beta}^{\mathrm{BR}}$. This suggests that when $\widehat{\beta}^{\mathrm{BR}}$ is used, estimation of the covariance matrix is done poorly when the true β deviates from zero, leading to less efficient estimates, explaining the behavior seen in the simulations.

References

Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics 34*, 305–334.

Cox, D. R. and E. J. Snell (1968). A general definition of residuals. *Journal of the Royal*

Statistical Society. Series B (Methodological), 248–275.

Newey, W. K. and D. McFadden (1994). Large sample estimation and hypothesis testing. *Handbook of Econometrics* **4**, 2111–2245.

Paul, S. and X. Zhang (2014). Small sample GEE estimation of regression parameters for longitudinal data. *Statistics in Medicine 33*(22), 3869–3881.

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

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics.

Department of Data Analysis and Mathematical Modelling, Ghent University, Ghent, Belgium

E-mail: Karelb.Vermeulen@UGent.be

Department of Data Analysis, Ghent University, Ghent, Belgium

E-mail: Jan.DeNeve@UGent.be

Department of Data Analysis and Mathematical Modelling, Ghent University, Ghent, Belgium and Department of Biostatistics, Vanderbilt University Medical Center, Nashville, US.

E-mail: ggca@outlook.com

Department of Data Analysis and Mathematical Modelling, Ghent University, Ghent, Belgium and Center for Statistics, Hasselt University, Hasselt, Belgium and National Institute for Applied Statistics Research Australia (NIASRA), School of Mathematics and Applied Statistics, University of Wollongong, Wollongong, Australia

E-mail: Olivier.Thas@UGent.be

Department of Applied Mathematics, Computer Sciences and Statistics, Ghent University, Ghent, Belgium and Department of Medical Statistics, London School of Hygiene and Tropical Medicine, London, United Kingdom

E-mail: Stijn.Vansteelandt@UGent.be

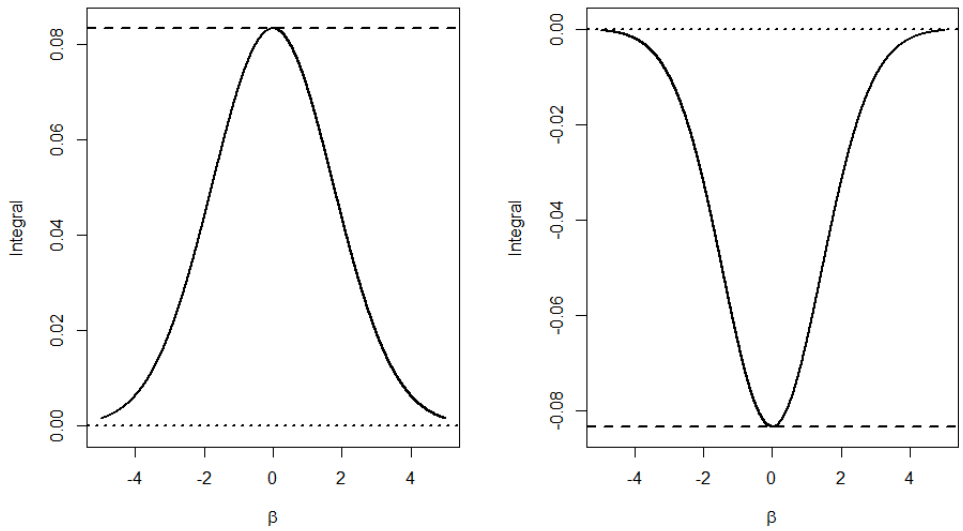

Figure 4: Off-diagonal elements $V_{ijjk}(\beta)$ (left panel) and $V_{ijik}(\beta)$ (right panel), for $i = 1$, $j = 4$ and $k = 25$; computed for several values of β. The solid line corresponds to the case when the semiparametric efficient estimator $\widehat{\beta}^{\mathrm{EFF}}$ is used, the dashed line corresponds to the case the bias-reduced estimator $\widehat{\beta}^{\mathrm{BR}}$ is used and the dotted line corresponds to the case where $\widehat{\beta}^{\mathrm{ST}}$ is used.

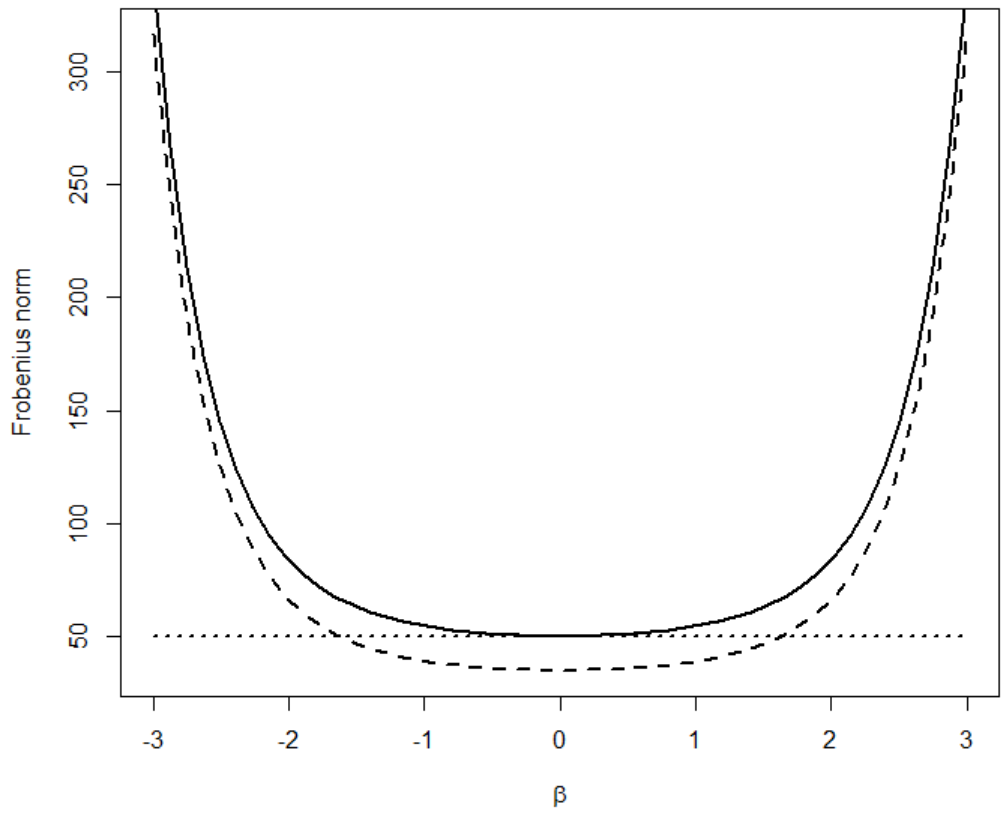

Figure 5: Frobenius norm for the inverse of the three covariance matrices used for the different estimators $\widehat{\beta}^{\mathrm{ST}}$ (dashed line), $\widehat{\beta}^{\mathrm{BR}}$ (dotted line) and $\widehat{\beta}^{\mathrm{EFF}}$ (solid line).