Web Supplementary Materials for "Regularized Outcome Weighted Subgroup Identification for Differential Treatment Effects" by

Yaoyao Xu[1], Menggang Yu[2,*], Ying-Qi Zhao[2], Quefeng Li[3], Sijian Wang[1,2], and Jun Shao[1]

[1]Department of Statistics, University of Wisconsin, Madison, Wisconsin, U.S.A

[2]Department of Biostatistics & Medical Informatics, University of Wisconsin, Madison, Wisconsin, U.S.A

[3]Department of Operation Research and Financial Engineering, Princeton University, Princeton, New Jersey, U.S.A.

*email: meyu@biostat.wisc.edu

Web Appendix A: Proofs

Proof of Proposition 1

For almost surely any fixed $\boldsymbol{x}$, the optimal rule

$$\begin{aligned}\mathcal{D}^*(x) &= \text{sign}\{\text{E}[Y|\boldsymbol{X}=\boldsymbol{x}, T=1] - \text{E}[Y|\boldsymbol{X}=\boldsymbol{x}, T=-1]\} \\ &= \text{sign}\{h(\boldsymbol{x}, \boldsymbol{x}'\boldsymbol{\beta}^\dagger) - h(\boldsymbol{x}, -\boldsymbol{x}'\boldsymbol{\beta}^\dagger)\}.\end{aligned}$$

Then under the assumptions of Proposition 1, $\mathcal{D}^*(\boldsymbol{x}) = 1$ if and only if $\boldsymbol{x}'\boldsymbol{\beta}^\dagger \geq 0$. Hence, $D^*(\boldsymbol{x}) = \text{sign}(\boldsymbol{x}'\boldsymbol{\beta}^\dagger)$.

Theoretical properties of $\hat{\boldsymbol{\beta}}$

We present the theoretical properties of our proposed method under the framework that both n and p diverge to infinity with $p \gg n$ but $n^{-1}\log p \to 0$. For the ease of presentation, we first give the main result without the second penalty $\eta(\boldsymbol{\beta})$ in (4). We then address the case when $\eta(\boldsymbol{\beta})$ is present.

We begin with some notations. Let $\mathcal{M}_{\tilde{\boldsymbol{\beta}}} = \{j : \tilde{\beta}_j \neq 0\}$ be the set of non-zero components of $\tilde{\boldsymbol{\beta}}$, and correspondingly $\mathcal{M}_{\hat{\boldsymbol{\beta}}} = \{j : \hat{\beta}_j \neq 0\}$. Denote $\tilde{\boldsymbol{\beta}}_1$ and $\tilde{\boldsymbol{\beta}}_0$ the subvector of $\tilde{\boldsymbol{\beta}}$ with indices in and out of $\mathcal{M}_{\tilde{\boldsymbol{\beta}}}$, respectively. Denote $\boldsymbol{X}_1$ and $\boldsymbol{X}_0$ the subvector of $\boldsymbol{X}$ with indices in and out of $\mathcal{M}_{\tilde{\boldsymbol{\beta}}}$, respectively. Let $s_p = |\mathcal{M}_{\tilde{\boldsymbol{\beta}}}|$, the cardinality of $\mathcal{M}_{\tilde{\boldsymbol{\beta}}}$; $d_p = \min_{j\in\mathcal{M}_{\tilde{\boldsymbol{\beta}}}} |\tilde{\beta}_j|$, the minimal signal. Finally, let $\mu(x) = e^x/(1+e^x)$. We write several quantities in terms of an order of n. We let $\log p \asymp n^{1-2\alpha_p}$ and $s_p \asymp n^{\alpha_s}$, where $0 < \alpha_p < 1/2$ and $\alpha_s > 0$. For any two sequence a_n and b_n, $a_n \asymp b_n$ means $a_n = O(b_n)$ and $b_n = O(a_n)$. Without loss of generality, we assume $\text{E}(\boldsymbol{X}) = \boldsymbol{0}$.

The following results show that $\hat{\boldsymbol{\beta}}$ possesses the "weak oracle property" (Fan and Lv, 2011), namely with large property, $\hat{\boldsymbol{\beta}}$ identifies all zero components of $\tilde{\boldsymbol{\beta}}$ and give consistent estimate to

non-zero components of $\tilde{\boldsymbol{\beta}}$ in a rate slower than $\sqrt{n}$.

Theorem 1 (weak oracle property) *Under the following conditions,*

(C1) $\max_{1\le j\le p} \mathrm{E}e^{tX_j} \le e^{ct^2/2}$, *for any real number* t, *where* c *is a constant.*

(C2) $|Y| \le M$ *a.s.*

(C3) $0 < \alpha_s < \gamma < \alpha_p < 1/2$.

(C4) $0 < \alpha_d < \gamma$.

(C5) $\|[\mathrm{E}\{W\mu'(\boldsymbol{X}_1'\tilde{\boldsymbol{\beta}}_1)\boldsymbol{X}_1\boldsymbol{X}_1'\}]^{-1}\|_\infty = O(b_n)$, *where* $W = Y/(\pi T + (1-T)/2)$, $\mu'(\cdot)$ *is the derivative of* $\mu(x)$, *and* $b_n = o(\min\{n^{1/2-\gamma}/\sqrt{\log n}, n^{\gamma-\alpha_s}\})$.

(C6) $\|[\mathrm{E}\{W\mu'(\boldsymbol{X}_1'\tilde{\boldsymbol{\beta}}_1)\boldsymbol{X}_0\boldsymbol{X}_1'\}][\mathrm{E}\{W\mu'(\boldsymbol{X}_1'\tilde{\boldsymbol{\beta}}_1)\boldsymbol{X}_1\boldsymbol{X}_1'\}]^{-1}\|_\infty < 1$.

(C7) $\max_{1\le j\le p} \lambda_{\max}(\mathrm{E}|X_j\boldsymbol{X}_1\boldsymbol{X}_1'|) = O(1)$, *where* $\lambda_{\max}(\boldsymbol{A})$ *is the maximal eigenvalue of* $\boldsymbol{A}$.

If we choose $\lambda_{1n} \asymp n^{-\alpha_{\lambda_1}}$ *such that*

$$0 < \alpha_{\lambda_1} < \min\{2\gamma - \alpha_s, \alpha_p\}, \text{ and } \lambda_{1n} b_n = o(n^{-\gamma}), \tag{A.1}$$

and $\lambda_{2n} = 0$, *then for sufficiently large* n, *with probability greater than* $1 - 4\{s_p n^{-1} + (p - s_p)e^{-n^{1-2\alpha_p}\log n}\}$, *the solution to* (4), $\hat{\boldsymbol{\beta}}$ *satisfies:*

(a) (sparsity) $\mathcal{M}_{\hat{\boldsymbol{\beta}}} = \mathcal{M}_{\tilde{\boldsymbol{\beta}}}$;

(b) (L_∞ *consistency)* $\|\hat{\boldsymbol{\beta}}_1 - \tilde{\boldsymbol{\beta}}_1\|_\infty \le n^{-\gamma}$.

Proof. Let $p_1(\boldsymbol{\beta}) = \lambda_{1n}\sum_{j=1}^p |\beta_j|$. Then, the subgradient of $p_1(\boldsymbol{\beta})$ is $\partial p_1(\beta_j) = s(\beta_j)$, where $s(\beta_j)$ is a set-valued function such that

$$s(\beta_j) = \begin{cases} \mathrm{sign}(\beta_j), & \text{if } \beta_j \ne 0; \\ c, & \text{if } \beta_j = 0. \end{cases} \tag{A.2}$$

where $0 < c < 1$.

Then, by classical optimization theory, any vector $\hat{\boldsymbol{\beta}} \in \mathcal{R}^p$ satisfying the following KKT conditions is a solution to (4).

$$\frac{1}{n}\tilde{\mathbf{X}}_1'[\boldsymbol{\mu}(\tilde{\mathbf{X}}_1\hat{\boldsymbol{\beta}}_1) \circ \boldsymbol{W} - \boldsymbol{W}] + \lambda_{1n}\,\mathrm{sign}(\hat{\boldsymbol{\beta}}_1) = 0, \tag{A.3}$$

$$\|\frac{1}{n}\tilde{\mathbf{X}}_0'[\boldsymbol{\mu}(\tilde{\mathbf{X}}_1\hat{\boldsymbol{\beta}}_1) \circ \boldsymbol{W} - \boldsymbol{W}]\|_\infty < \lambda_{1n}, \tag{A.4}$$

where $\mathbf{X} = (X_1, \ldots, X_n)'$ is the design matrix, $\boldsymbol{W} = (\frac{Y_1}{\pi T_1 + (1-T_1)/2}, \ldots, \frac{Y_n}{\pi T_n + (1-T_n)/2})'$, $\boldsymbol{T} = (T_1, \ldots, T_n)'$, $\tilde{\mathbf{X}} = (T_1X_1, \ldots, T_nX_n)'$, $\tilde{\mathbf{X}}_1$ is the submatrix of $\tilde{\mathbf{X}}$ with columns in $\mathcal{M}_{\hat{\boldsymbol{\beta}}}$ and $\tilde{\mathbf{X}}_0$ is the submatrix of $\tilde{\mathbf{X}}$ with columns not in $\mathcal{M}_{\hat{\boldsymbol{\beta}}}$, $\hat{\boldsymbol{\beta}}_1$ is the subvector of $\hat{\boldsymbol{\beta}}$ with indices in $\mathcal{M}_{\hat{\boldsymbol{\beta}}}$. $\boldsymbol{\mu}(\boldsymbol{x}) : \mathcal{R}^n \to \mathcal{R}^n$ is a function such that the ith element $\mu(x_i) = e^{x_i}/(1+e^{x_i})$, and $\circ$ denotes componentwise product. In the following, we show that within a neighborhood of $\tilde{\boldsymbol{\beta}}$, such a vector exists and satisfies (a) and (b). Since the original problem (4) is convex, it has a unique solution. Then, the theorem follows.

Let $\boldsymbol{\epsilon}_1 = \frac{1}{n}\tilde{\mathbf{X}}_1'\boldsymbol{W} - \mathrm{E}(W\tilde{\boldsymbol{X}}_1)$, $\boldsymbol{\epsilon}_0 = \frac{1}{n}\tilde{\mathbf{X}}_0'\boldsymbol{W} - \mathrm{E}(W\tilde{\boldsymbol{X}}_0)$, where $W = \frac{Y}{\pi T + (1-T)/2}$, $\tilde{\boldsymbol{X}} = T\boldsymbol{X}$, $\tilde{\boldsymbol{X}}_1$ and $\tilde{\boldsymbol{X}}_0$ are subvectors of $\boldsymbol{X}$ with indices in and out of $\mathcal{M}_{\hat{\boldsymbol{\beta}}}$, respectively. Let $\boldsymbol{\xi}_1 = \frac{1}{n}\tilde{\mathbf{X}}_1'[\boldsymbol{\mu}(\tilde{\mathbf{X}}\boldsymbol{\beta}) \circ \boldsymbol{W}] - \mathrm{E}[\mu(\tilde{\boldsymbol{X}}'\boldsymbol{\beta})W\tilde{\boldsymbol{X}}_1]$, $\boldsymbol{\xi}_0 = \frac{1}{n}\tilde{\mathbf{X}}_0'[\boldsymbol{\mu}(\tilde{\mathbf{X}}\boldsymbol{\beta}) \circ \boldsymbol{W}] - \mathrm{E}[\mu(\tilde{\boldsymbol{X}}'\boldsymbol{\beta})W\tilde{\boldsymbol{X}}_0]$.

Let event $E_1 = \{\|\boldsymbol{\epsilon}_1\|_\infty \le C_1\sqrt{\log n/n}\}$, $E_2 = \{\|\boldsymbol{\epsilon}_0\|_\infty \le C_1 n^{-\alpha_p}\sqrt{\log n}\}$, $E_3 = \{\|\boldsymbol{\xi}_1\|_\infty \le C_2\sqrt{\log n/n}\}$ and $E_4 = \{\|\boldsymbol{\xi}_0\|_\infty \le C_2 n^{-\alpha_p}\sqrt{\log n}\}$, where C_1 and C_2 are constants depending on c and M.

Condition (C1) ensures that X_j is a sub-Gaussian random variable. It then follows from (C2) that $\tilde{X}_jW$ is also sub-Gaussian with mean zero, i.e. there exists a constant c_1 depending on c and M that

$$\max_{1\le j\le p} \mathrm{E}e^{t\tilde{X}_jW} \le e^{c_1t^2/2}.$$

By the Hoeffding's bound for sub-Gaussian random variables, it holds that

$$\max_{1\le j\le s_p} P\left(|\frac{1}{n}\sum_{i=1}^{n}\tilde{X}_{ij}W_i - \mathrm{E}(\tilde{X}_jW)| > \sqrt{2c_1\log n/n}\right) \le 2\exp(-\log n) = 2/n.$$

Let $C_1 = \sqrt{2c_1}$, it follows from Bonferroni inequality that

$$\begin{aligned}
&P\left(\|\boldsymbol{\epsilon}_1\|_\infty > C_1\sqrt{\log n/n}\right)\\
&\le s_p \max_{1\le j\le s_p} P\left(|\frac{1}{n}\sum_{i=1}^{n}\tilde{X}_{ij}W_i - \mathrm{E}(\tilde{X}_jW)| \ge 2C_1\sqrt{\log n/n}\right)\\
&\le 2s_p/n.
\end{aligned}$$

Similarly, we can show that

$$P\left(\|\boldsymbol{\epsilon}_0\|_\infty > C_1 n^{-\alpha_p}\sqrt{\log n}\right) \le 2(p-s_p)e^{-n^{1-2\alpha_p}\log n}.$$

Since $|\mu(x)| \le 1$, following the same technique as in the above, we can show that

$$\begin{aligned}
P\left(\|\boldsymbol{\xi}_1\|_\infty > C_2\sqrt{\log n/n}\right) &\le 2s_p/n,\\
P\left(\|\boldsymbol{\xi}_0\|_\infty > C_2 n^{-\alpha_p}\sqrt{\log n}\right) &\le 2(p-s_p)e^{-n^{1-2\alpha_p}\log n}.
\end{aligned}$$

Therefore,

$$P(E_1\cap E_2\cap E_3\cap E_4) \ge 1 - 4\{s_p/n + (p-s_p)e^{-n^{1-2\alpha_p}\log n}\}.$$

Next, we show that within event $E_1\cap E_2\cap E_3\cap E_4$, there exists a solution to (A.3) and (A.4) and satisfies (a) and (b).

Step 1: Solution to (A.3). First, we prove that, when n is sufficiently large, there exists a solution to (A.3) in the hypercube

$$\mathcal{N} = \{\boldsymbol{\delta}\in\mathcal{R}^{s_p} : \|\boldsymbol{\delta} - \tilde{\boldsymbol{\beta}}_1\|_\infty = n^{-\gamma}\}.$$

Since $\tilde{\boldsymbol{\beta}}$ is the minimizer of (3), and $|\frac{\partial}{\partial\boldsymbol{\beta}}W\log(1+e^{-\tilde{\boldsymbol{X}}'\boldsymbol{\beta}})| = |\{\mu(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{\beta}}) - 1\}W\tilde{\boldsymbol{X}}| \le C|\boldsymbol{X}|$, which is integrable, it follows that

$$\mathrm{E}\left[\{\mu(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{\beta}}) - 1\}W\tilde{\boldsymbol{X}}\right] = \frac{\partial}{\partial\boldsymbol{\beta}}\mathrm{E}\left[W\log(1+e^{-\tilde{\boldsymbol{X}}'\boldsymbol{\beta}})\right]\Big|_{\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}} = \mathbf{0}. \tag{A.5}$$

Then, (A.3) is equivalent to

$$\frac{1}{n}\tilde{\mathbf{X}}_1'[\boldsymbol{\mu}(\tilde{\mathbf{X}}_1\boldsymbol{\delta})\circ\boldsymbol{W}]-\frac{1}{n}\tilde{\mathbf{X}}_1'\boldsymbol{W}-\mathrm{E}\left[\{\mu(\tilde{\boldsymbol{X}}_1'\tilde{\boldsymbol{\beta}}_1)-1\}W\tilde{\boldsymbol{X}}_1\right]=-\lambda_{1n}\,\mathrm{sign}(\boldsymbol{\delta}).$$

It is further equivalent to

$$\mathrm{E}\left[\mu(\tilde{\boldsymbol{X}}_1'\boldsymbol{\delta})W\tilde{\boldsymbol{X}}_1\right]-\mathrm{E}\left[\mu(\tilde{\boldsymbol{X}}_1'\tilde{\boldsymbol{\beta}}_1)W\tilde{\boldsymbol{X}}_1\right]=\boldsymbol{\epsilon}_1-\boldsymbol{\xi}_1-\lambda_{1n}\,\mathrm{sign}(\boldsymbol{\delta}). \tag{A.6}$$

By Taylor expansion,

$$\mathrm{E}\left[\mu(\tilde{\boldsymbol{X}}_1'\boldsymbol{\delta})W\tilde{\boldsymbol{X}}_1\right]-\mathrm{E}\left[\mu(\tilde{\boldsymbol{X}}_1'\tilde{\boldsymbol{\beta}}_1)W\tilde{\boldsymbol{X}}_1\right]=\mathrm{E}\left[W\tilde{\boldsymbol{X}}_1\mu'(\tilde{\boldsymbol{X}}_1'\tilde{\boldsymbol{\beta}})\tilde{\boldsymbol{X}}_1'\right](\boldsymbol{\delta}-\tilde{\boldsymbol{\beta}}_1)+\boldsymbol{r}, \tag{A.7}$$

where the jth component of the reminder term $\boldsymbol{r}$,

$$r_j=(\boldsymbol{\delta}-\tilde{\boldsymbol{\beta}}_1)'\mathrm{E}\left[2^{-1}W\tilde{X}_j\mu''(\tilde{\boldsymbol{X}}'\bar{\boldsymbol{\delta}})\tilde{\boldsymbol{X}}_1\tilde{\boldsymbol{X}}_1'\right](\boldsymbol{\delta}-\tilde{\boldsymbol{\beta}}_1),$$

where $\bar{\boldsymbol{\delta}}$ lies on the line segment connecting $\boldsymbol{\delta}$ and $\tilde{\boldsymbol{\beta}}_1$. Since $\mu''(\tilde{\boldsymbol{X}}'\bar{\boldsymbol{\delta}})=O(1)$ for all $\boldsymbol{\delta}\in\mathcal{N}$, by (C2),

$$\lambda_{\max}(\mathrm{E}|2^{-1}W\tilde{X}_j\mu''(\tilde{\boldsymbol{X}}'\bar{\boldsymbol{\delta}})\tilde{\boldsymbol{X}}_1\tilde{\boldsymbol{X}}_1'|)=O(\lambda_{\max}(\mathrm{E}|WTX_j\tilde{\boldsymbol{X}}_1\tilde{\boldsymbol{X}}_1'|))=O(\lambda_{\max}(\mathrm{E}|X_j\boldsymbol{X}_1\boldsymbol{X}_1'|)). \tag{A.8}$$

Then, by (C7), $\|\boldsymbol{r}\|_\infty=O(||\boldsymbol{\delta}-\tilde{\boldsymbol{\beta}}_1||^2)=O(s_pn^{-2\gamma})$. It follows from (A.6) and (A.7) that (A.3) is equivalent to

$$\begin{aligned}&\mathrm{E}\left[W\tilde{\boldsymbol{X}}_1\mu'(\tilde{\boldsymbol{X}}_1'\tilde{\boldsymbol{\beta}}_1)\tilde{\boldsymbol{X}}_1'\right](\boldsymbol{\delta}-\tilde{\boldsymbol{\beta}}_1)-\boldsymbol{\epsilon}_1+\boldsymbol{\xi}_1+\boldsymbol{r}+\lambda_{1n}\,\mathrm{sign}(\boldsymbol{\delta})\\=&\,\mathrm{E}\left[W\boldsymbol{X}_1\mu'(\boldsymbol{X}_1'\tilde{\boldsymbol{\beta}}_1)\boldsymbol{X}_1'\right](\boldsymbol{\delta}-\tilde{\boldsymbol{\beta}}_1)-\boldsymbol{\epsilon}_1+\boldsymbol{\xi}_1+\boldsymbol{r}+\lambda_{1n}\,\mathrm{sign}(\boldsymbol{\delta})\\=&\,\mathbf{0}.\end{aligned}$$

Let $\boldsymbol{\Psi}(\boldsymbol{\delta})=\boldsymbol{\delta}-\tilde{\boldsymbol{\beta}}_1-\left\{\mathrm{E}\left[W\boldsymbol{X}_1\mu'(\boldsymbol{X}_1'\tilde{\boldsymbol{\beta}}_1)\boldsymbol{X}_1'\right]\right\}^{-1}(\boldsymbol{\epsilon}_1-\boldsymbol{\xi}_1-\boldsymbol{r}-\lambda_{1n}\,\mathrm{sign}(\boldsymbol{\delta}))$. Then, if $\boldsymbol{\delta}$ solves $\boldsymbol{\Psi}(\boldsymbol{\delta})=\mathbf{0}$, it also solves (A.3).

It follows from (C5) and the choice of λ_{1n} that

$$\begin{aligned}
&\|\Big\{\mathrm{E}\left[W\boldsymbol{X}_1\mu'(\boldsymbol{X}_1'\tilde{\boldsymbol{\beta}}_1)\boldsymbol{X}_1'\right]\Big\}^{-1}(\boldsymbol{\epsilon}_1-\boldsymbol{\xi}_1-\boldsymbol{r}-\lambda_{1n}\,\mathrm{sign}(\boldsymbol{\delta}))\|_\infty \\
\le\ &\|\Big\{\mathrm{E}\left[W\boldsymbol{X}_1\mu'(\boldsymbol{X}_1'\tilde{\boldsymbol{\beta}}_1)\boldsymbol{X}_1'\right]\Big\}^{-1}\|_\infty(\|\boldsymbol{\epsilon}_1\|_\infty+\|\boldsymbol{\xi}_1\|_\infty+\|\boldsymbol{r}\|_\infty+\lambda_{1n}) \\
=\ &o(n^{-\gamma}).
\end{aligned}$$

Then, for sufficiently large n, if $\delta_j-\tilde{\beta}_j=n^{-\gamma}$, $\boldsymbol{\Psi}(\delta_j)>0$; if $\delta_j-\tilde{\beta}_j=-n^{-\gamma}$, $\boldsymbol{\Psi}(\delta_j)<0$. By continuity of $\Psi(\boldsymbol{\delta})$, an application of Miranda's existence theorem shows that $\boldsymbol{\Psi}(\boldsymbol{\delta})=\mathbf{0}$ has a solution in $\mathcal{N}$, which is also the solution to (A.3).

Step 2: verify (A.4). Let $\hat{\boldsymbol{\beta}}=(\hat{\boldsymbol{\beta}}_1,\mathbf{0})'$, where $\hat{\boldsymbol{\beta}}_1$ is the solution to (A.3) as shown above. Next, we prove that $\hat{\boldsymbol{\beta}}$ satisfies (A.4).

Since, by (A.5), $\mathrm{E}\left[\{\mu(\tilde{\boldsymbol{X}}_1'\tilde{\boldsymbol{\beta}}_1)-1\}W\tilde{\boldsymbol{X}}_0\right]=\mathbf{0}$. Then, it follows that

$$\begin{aligned}
\frac{1}{n}\tilde{\mathbf{X}}_0'[\boldsymbol{\mu}(\tilde{\mathbf{X}}_1\hat{\boldsymbol{\beta}}_1)\circ\boldsymbol{W}-\boldsymbol{W}]&=\frac{1}{n}\tilde{\mathbf{X}}_0'[\boldsymbol{\mu}(\tilde{\mathbf{X}}_1\hat{\boldsymbol{\beta}}_1)\circ\boldsymbol{W}-\boldsymbol{W}]-\mathrm{E}\left[\{\mu(\tilde{\boldsymbol{X}}_1'\tilde{\boldsymbol{\beta}}_1)-1\}W\tilde{\boldsymbol{X}}_0\right] \\
&=\mathrm{E}\left[\mu(\tilde{\boldsymbol{X}}_1'\hat{\boldsymbol{\beta}}_1)W\tilde{\boldsymbol{X}}_0\right]-\mathrm{E}\left[\mu(\tilde{\boldsymbol{X}}_1'\tilde{\boldsymbol{\beta}}_1)W\tilde{\boldsymbol{X}}_0\right]-\boldsymbol{\xi}_0+\boldsymbol{\epsilon}_0
\end{aligned}\tag{A.9}$$

By Taylor expansion,

$$\begin{aligned}
\mathrm{E}\left[\mu(\tilde{\boldsymbol{X}}_1'\hat{\boldsymbol{\beta}}_1)W\tilde{\boldsymbol{X}}_0\right]-\mathrm{E}\left[\mu(\tilde{\boldsymbol{X}}_1'\tilde{\boldsymbol{\beta}}_1)W\tilde{\boldsymbol{X}}_0\right]&=\mathrm{E}\left[W\tilde{\boldsymbol{X}}_0\mu'(\tilde{\boldsymbol{X}}_1'\tilde{\boldsymbol{\beta}})\tilde{\boldsymbol{X}}_1'\right](\hat{\boldsymbol{\beta}}_1-\tilde{\boldsymbol{\beta}}_1)+\boldsymbol{u}, \\
&=\mathrm{E}\left[W\boldsymbol{X}_0\mu'(\boldsymbol{X}_1'\tilde{\boldsymbol{\beta}}_1)\boldsymbol{X}_1'\right](\hat{\boldsymbol{\beta}}_1-\tilde{\boldsymbol{\beta}}_1)+\boldsymbol{u},
\end{aligned}\tag{A.10}$$

where the jth component of $\boldsymbol{u}$ is

$$u_j=(\hat{\boldsymbol{\beta}}_1-\tilde{\boldsymbol{\beta}}_1)'\mathrm{E}\left[2^{-1}\tilde{X}_j\mu''(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{\delta}})\tilde{\boldsymbol{X}}_1\tilde{\boldsymbol{X}}_1'\right](\hat{\boldsymbol{\beta}}_1-\tilde{\boldsymbol{\beta}}_1),$$

where $\tilde{\boldsymbol{\delta}}$ lies on the line segment connecting $\hat{\boldsymbol{\beta}}_1$ and $\tilde{\boldsymbol{\beta}}_1$. In analogous to (A.8), under (C7), we can show that $\|\boldsymbol{u}\|_\infty=O(s_pn^{-2\gamma})$.

Since $\hat{\boldsymbol{\beta}}_1$ is the solution to $\boldsymbol{\Psi}(\boldsymbol{\delta})=\mathbf{0}$, it holds that

$$\hat{\boldsymbol{\beta}}_1-\tilde{\boldsymbol{\beta}}_1=\Big\{\mathrm{E}\left[W\boldsymbol{X}_1\mu'(\boldsymbol{X}_1'\tilde{\boldsymbol{\beta}}_1)\boldsymbol{X}_1'\right]\Big\}^{-1}(\boldsymbol{\epsilon}_1-\boldsymbol{\xi}_1-\boldsymbol{r}-\lambda_{1n}\,\mathrm{sign}(\boldsymbol{\delta})).$$

Then, by (A.9) and (A.10),

$$\begin{aligned}
&\frac{1}{n\lambda_{1n}}\tilde{\mathbf{X}}_0'[\boldsymbol{\mu}(\tilde{\mathbf{X}}_1\hat{\boldsymbol{\beta}}_1)\circ\boldsymbol{W}-\boldsymbol{W}] \\
&= \frac{1}{\lambda_{1n}}\mathrm{E}\left[W\boldsymbol{X}_0\mu'(\boldsymbol{X}_1'\tilde{\boldsymbol{\beta}}_1)\boldsymbol{X}_1'\right]\left\{\mathrm{E}\left[W\boldsymbol{X}_1\mu'(\boldsymbol{X}_1'\tilde{\boldsymbol{\beta}}_1)\boldsymbol{X}_1'\right]\right\}^{-1}(\boldsymbol{\epsilon}_1-\boldsymbol{\xi}_1-\boldsymbol{r}-\lambda_{1n}\operatorname{sign}(\boldsymbol{\delta})) \\
&\quad+\frac{1}{\lambda_{1n}}\boldsymbol{\epsilon}_0-\frac{1}{\lambda_{1n}}\boldsymbol{\xi}_0+\frac{1}{\lambda_{1n}}\boldsymbol{u}
\end{aligned}$$

In the event $E_1\cap E_2\cap E_3\cap E_4$, by the choice of λ_{1n},

$$\|\lambda_{1n}^{-1}\boldsymbol{\epsilon}_0\|_\infty=o(1),\ \|\lambda_{1n}^{-1}\boldsymbol{\xi}_0\|_\infty=o(1),\ \|\lambda_{1n}^{-1}\boldsymbol{u}\|_\infty=o(1),$$

By (C6),

$$\frac{1}{\lambda_{1n}}\left\|\mathrm{E}\left[W\boldsymbol{X}_0\mu'(\boldsymbol{X}_1'\tilde{\boldsymbol{\beta}}_1)\boldsymbol{X}_1'\right]\left\{\mathrm{E}\left[W\boldsymbol{X}_1\mu'(\boldsymbol{X}_1'\tilde{\boldsymbol{\beta}}_1)\boldsymbol{X}_1'\right]\right\}^{-1}(\boldsymbol{\epsilon}_1-\boldsymbol{\xi}_1-\boldsymbol{r})\right\|_\infty<\frac{1}{\lambda_{1n}}\|\boldsymbol{\epsilon}_1-\boldsymbol{\xi}_1-\boldsymbol{r}\|_\infty=o(1).$$

Finally, by (C6),

$$\frac{1}{\lambda_{1n}}\left\|\mathrm{E}\left[W\boldsymbol{X}_0\mu'(\boldsymbol{X}_1'\tilde{\boldsymbol{\beta}}_1)\boldsymbol{X}_1'\right]\left\{\mathrm{E}\left[W\boldsymbol{X}_1\mu'(\boldsymbol{X}_1'\tilde{\boldsymbol{\beta}}_1)\boldsymbol{X}_1'\right]\right\}^{-1}\lambda_{1n}\operatorname{sign}(\hat{\boldsymbol{\beta}}_1)\right\|_\infty<1.$$

Therefore, $\hat{\boldsymbol{\beta}}$ satisfies (A.4). This completes the proof.

If $\eta(\boldsymbol{\beta})$ in (4) is the fused LASSO penalty as described in Section 2, then Theorem 1 is still valid, given that $\lambda_{2n}/\lambda_{1n}=o(1)$. In fact, in the presence of fused LASSO penalty, the original KKT conditions (A.3) and (A.4) becomes

$$\begin{aligned}
\frac{1}{n}\tilde{\mathbf{X}}_1'[\boldsymbol{\mu}(\tilde{\mathbf{X}}_1\hat{\boldsymbol{\beta}}_1)\circ\boldsymbol{W}-\boldsymbol{W}]+\lambda_{1n}\operatorname{sign}(\hat{\boldsymbol{\beta}}_1)+\lambda_{2n}\partial\eta(\hat{\boldsymbol{\beta}}_1)=0,\\
\|\frac{1}{n}\tilde{\mathbf{X}}_0'[\boldsymbol{\mu}(\tilde{\mathbf{X}}_1\hat{\boldsymbol{\beta}}_1)\circ\boldsymbol{W}-\boldsymbol{W}]\|_\infty+\lambda_{2n}\|\partial\eta(\hat{\boldsymbol{\beta}}_0)\|_\infty<\lambda_{1n},
\end{aligned}$$

where the subgradient of $\eta(\boldsymbol{\beta})$ (Lemma 6.1 of Rinaldo (2009)) is any vector such that

$$\partial\eta(\beta_{j,m})=\begin{cases}
-s(\beta_{j,2}-\beta_{j,1}), & \text{if } m=1;\\
s(\beta_{j,m}-\beta_{j,m-1})-s(\beta_{j,m+1}-\beta_{j,m}), & \text{if } 2<m<C_j-2;\\
s(\beta_{j,C_j-1}-\beta_{j,C_j-2}), & \text{if } m=C_j-1;
\end{cases}$$

where $s(\beta)$ is defined in (A.2).

Since $\|\partial\eta(\boldsymbol{\beta})\|_\infty \leq 2$, if we choose $\lambda_{2n}/\lambda_{1n} = o(1)$, following the same procedure as the above proof, we can still show that $\hat{\boldsymbol{\beta}}$ processes the weak oracle property as in (a) and (b). The choice of λ_{2n} indicates that adding penalty on adjacent levels of ordinal variables won't change the asymptotic property of $\hat{\boldsymbol{\beta}}$ as long as it is dominated by the main L_1 penalty.

Web Appendix B: Additional simulations for time-to-event outcomes

We studied our method for time-to-event outcomes, where the survival time Y and censoring time C were generated using the following five models. The survival and censoring times were conditionally independent given covariates. Each model was examined through three different sample sizes: $n = 500$, 1000 and 2000 and repeated for 200 simulations. The results are shown in Table A.1.

(A) Accelerated failure time (AFT) model with normal error

$$\log(Y) = 2X_{A1} - 2X_{B1} + (5X_{a2} + 5X_{a3} - X_{Ca}) * T + 0.5\epsilon_1,$$

$$\log(C) = 3X_{A1} - 3X_{B1} + (4X_{a2} + 4X_{a3} - 4X_{b3}) * T + 0.5\epsilon_2, \text{ where } \epsilon_1, \epsilon_2 \sim N(0,1);$$

(B) AFT model with logistic error

$$\log(Y) = 2X_{A1} - 2X_{B1} + (5X_{a2} + 5X_{a3} - X_{Ca}) * T + 0.5\epsilon_1$$

$$\log(C) = 3X_{A1} - 3X_{B1} + (4X_{a2} + 4X_{a3} - 4X_{b3}) * T + 0.5\epsilon_2, \text{ where } \epsilon_1, \epsilon_2 \sim \text{logistic}(0,1);$$

(C) Cox Weibull model

$$\lambda_Y(Y|\boldsymbol{X}) = \lambda_{Y_0}(Y)\exp(2X_{A1} - 2X_{B1} + (5X_{a2} + 5X_{a3} - X_{Ca}) * T), \text{ where } \lambda_{Y_0}(y) = 2y,$$

$$\lambda_C(C|\boldsymbol{X}) = \lambda_{C_0}(C)\exp(3X_{A1} - 3X_{B1} + (4X_{a2} + 4X_{a3} - 4X_{b3}) * T), \text{ where } \lambda_{C_0}(c) = 2c$$

(D) Cox Gompertz model

$$\lambda_Y(Y|\boldsymbol{X}) = \lambda_{Y_0}(Y)\exp(2X_{A1} - 2X_{B1} + (5X_{a2} + 5X_{a3} - X_{Ca}) * T), \text{ where } \lambda_{Y_0}(y) = e^y$$

$$\lambda_Y(C|\boldsymbol{X}) = \lambda_{C_0}(C)\exp(3X_{A1} - 3X_{B1} + (4X_{a2} + 4X_{a3} - 4X_{b3}) * T), \text{ where } \lambda_{C_0}(c) = e^c$$

(E) Cox Exponential model

$$\lambda_Y(Y|\boldsymbol{X}) = \lambda_{Y_0}(Y)\exp(2X_{A1} - 2X_{B1} + (5X_{a2} + 5X_{a3} - X_{Ca}) * T), \text{ where } \lambda_{Y_0}(y) = 1$$

$$\lambda_C(C|\boldsymbol{X}) = \lambda_{C_0}(C)\exp(3X_{A1} - 3X_{B1} + (4X_{a2} + 4X_{a3} - 4X_{b3}) * T), \text{ where } \lambda_{C_0}(c) = 1$$

Model (A)	Sensitivity	Specificity	Misclassification rate
$n = 500$	1.00	0.90	0.16
$n = 1000$	1.00	0.90	0.13
$n = 2000$	1.00	0.90	0.14
Model (B)	Sensitivity	Specificity	Misclassification rate
$n = 500$	1.00	0.90	0.15
$n = 1000$	1.00	0.90	0.14
$n = 2000$	1.00	0.90	0.14
Model (C)	Sensitivity	Specificity	Misclassification rate
$n = 500$	0.99	0.91	0.12
$n = 1000$	1.00	0.91	0.12
$n = 2000$	1.00	0.90	0.12
Model (D)	Sensitivity	Specificity	Misclassification rate
$n = 500$	0.99	0.90	0.13
$n = 1000$	1.00	0.90	0.13
$n = 2000$	0.99	0.90	0.13
Model (E)	Sensitivity	Specificity	Misclassification rate
$n = 500$	1.00	0.90	0.13
$n = 1000$	1.00	0.90	0.12
$n = 2000$	1.00	0.90	0.12

Table A.1: Results of the five time-to-event models

Web Appendix C: Values of the target function from simulation

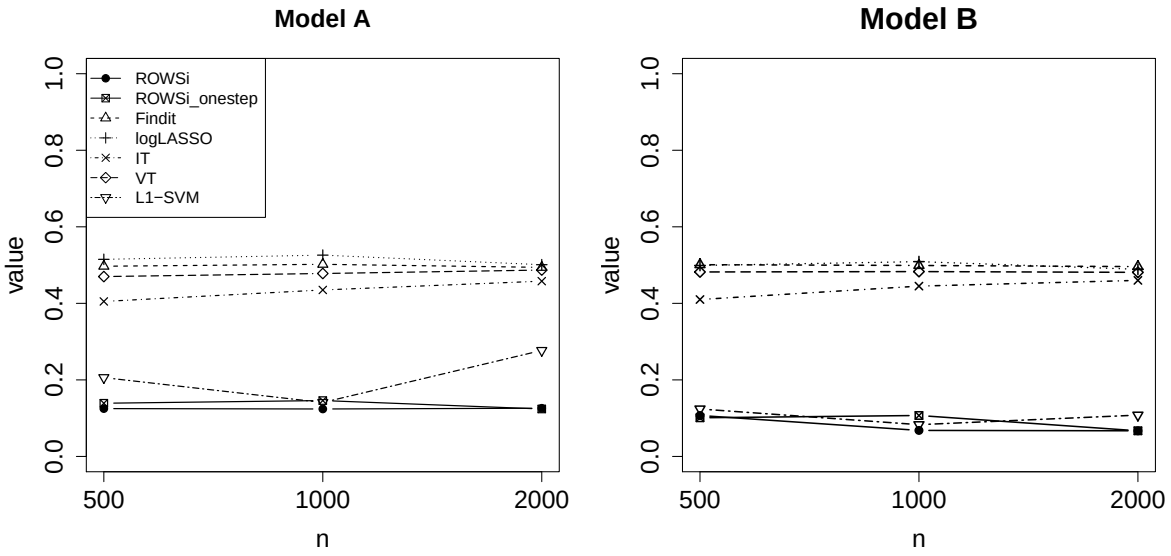
Figure A.1 plots values of the target function

$$\mathrm{E}\left[\frac{I(T \neq \mathcal{D}(\boldsymbol{X}))}{T\pi + (1-T)/2} Y\right]$$

in (1) under various methods. Smaller values indicate better performance. The target function is directly related to the value function

$$\mathrm{E}^{\mathcal{D}}(Y) = \int Y \frac{dP^{\mathcal{D}}}{dP} dP = \mathrm{E}\left[\frac{I(T = \mathcal{D}(\boldsymbol{X}))}{T\pi + (1-T)/2} Y\right] = \mathrm{E}\left[\frac{Y}{T\pi + (1-T)/2}\right] - \mathrm{E}\left[\frac{I(T \neq \mathcal{D}(\boldsymbol{X}))}{T\pi + (1-T)/2} Y\right].$$

We can see that our method reaches the smallest value among all methods in most cases.



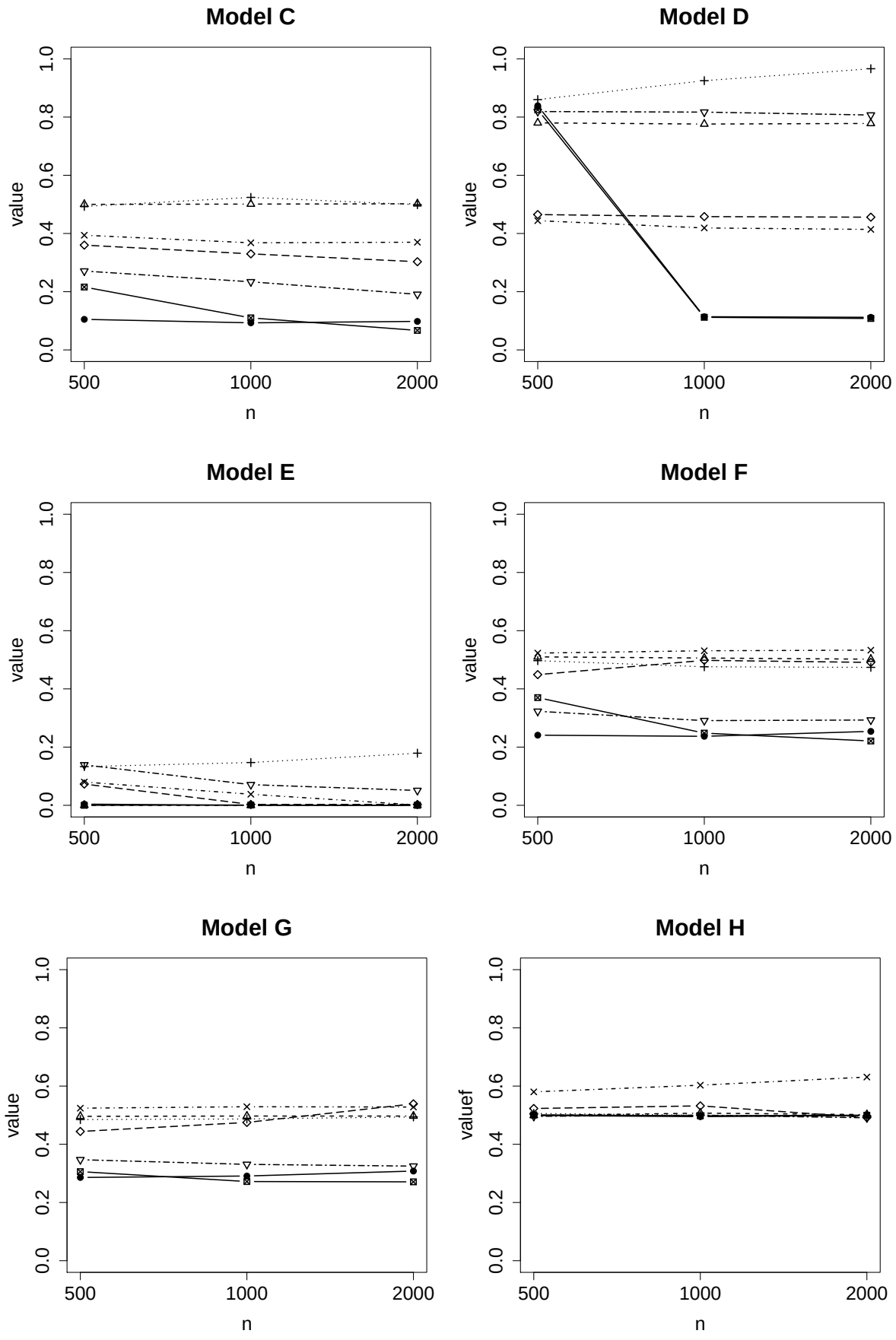


Figure A.1: Comparison of values of target function (number of simulations = 200).

Web Appendix D: Sensitivity analysis on the choice of α in the m-out-of-n bootstrap

To investigate the effects of different choices for α on the m-out-of-n bootstrap, we applied our method to Model (A) in Section 4 with $\alpha = 0.7, 0.8$, and 0.9. Results are shown in Table A.2. It's seen that the coverage of the confidence intervals was not satisfactory when $\alpha = 0.9$, but was quite good for the other two choices.

Table A.2: Confidence intervals of enhanced comparative treatment effects in Models (A) for different choice of α. For notational simplicity, we denote $d_+ \equiv E_{\boldsymbol{X}}\{d(\boldsymbol{X}, 1, \hat{\boldsymbol{\beta}})\}$ and $d_- \equiv E_{\boldsymbol{X}}\{d(\boldsymbol{X}, -1, \hat{\boldsymbol{\beta}})\}$

	$n = 500$		$n = 1000$		$n = 2000$	
	d_+	d_-	d_+	d_-	d_+	d_-
$\alpha = 0.7$						
CI length MEAN	0.67	0.66	0.52	0.50	0.35	0.36
coverage	0.93	1.00	0.99	1.00	0.98	0.98
$\alpha = 0.8$						
CI length MEAN	0.55	0.57	0.41	0.34	0.24	0.24
coverage	0.93	0.98	0.97	0.99	0.95	0.96
$\alpha = 0.9$						
CI length MEAN	0.39	0.31	0.19	0.20	0.17	0.17
coverage	0.90	0.97	0.81	0.84	0.88	0.90

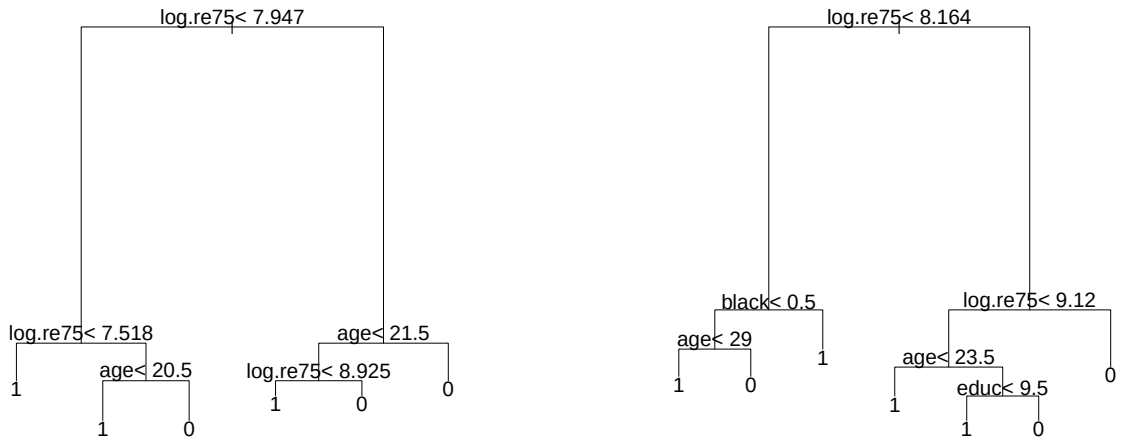
Web Appendix E: Further data analysis results

Table A.3: Rules for the National Supported Work Study

Method	Rules
ROWSi	$\hat{\mathcal{D}} = \text{sign}(0.1093 + 0.1888I\{\text{married}\} + 0.1152I\{\text{no high school degree \& low earning}\}$ $-0.2234I\{\text{Hispanics}\} - 0.2976I\{\text{employed}\} - 0.1600\text{age} * \log(\text{real earning in 1975}))$
FindIt	$\hat{\mathcal{D}} = \text{sign}(0.0213 + 0.1022I\{\text{black}\} + 0.0634I\{\text{no high school degree}\}$ $-0.2018I\{\text{Hispanics}\} * \log(\text{real earning in 1975}))$
VT	Figure A.2 (a)
IT	Figure A.2 (b)

Table A.4: Rules for the Mammography Screening Study

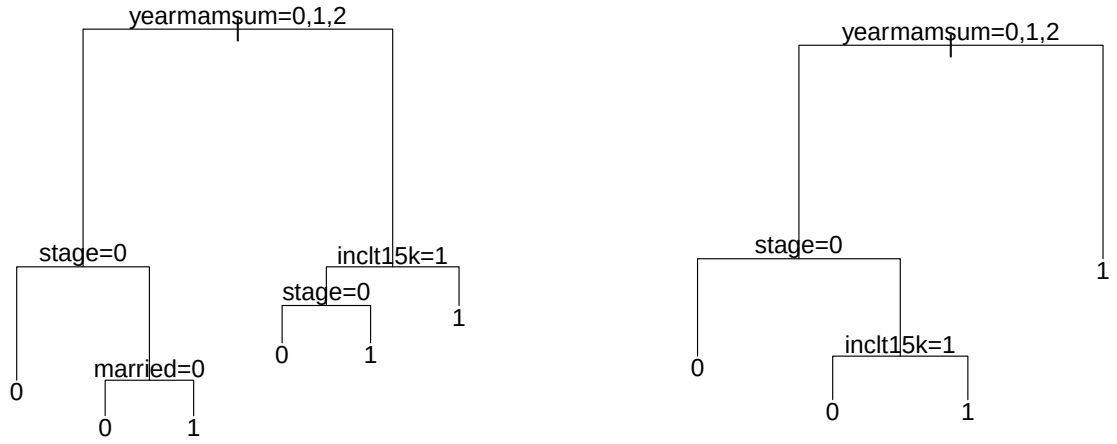
Method	Rules
ROWSi	$\hat{\mathcal{D}} = \text{sign}(0.2350 + 0.6403I\{\text{yearmamsum} = 3 \text{ or } 4\} + 0.3060I\{\text{se1tot} > 40\}$ $-0.6011I\{\text{inclt15k}\} - 0.2025I\{\text{bar1tot} > 30\})$
VT	Figure A.3 (a)
IT	Figure A.3 (b)
Remark	**yearmamsum:** number of years had a mammogram in the past 2 to 5 years; **inclt15k:** household income less than \$15,000; **se1tot:** self efficacy score; **bar1tot:** barriers scale score



(a) VT

(b) IT

Figure A.2: Results for the National Supported Work Study. The variable **log.re75** stands for log(real earnings in 1975), and **educ** stands for years of schooling. For each split the left child represents observations that meet the condition associated with the split. Terminal nodes are labeled "1" if proportion of subjects in the node with binary outcome=1 exceeded cutoff of 0.5. The final subgroup is formed as the union of terminal nodes labeled as "1".



(a) VT (b) IT

Figure A.3: Results for the Mammography Screening Study. The variable **stage** stands for baseline stage of mammography screening behavior, **yearmamsum** stands for number of years having a mammogram in past 2 to 5 years, and **incle15k** stands for household income $\leq$ \$15,000. For each split the left child represents observations that meet the condition associated with the split. Terminal nodes are labeled "1" if proportion of subjects in the node with binary outcome=1 exceeded cutoff of 0.5. The final subgroup is formed as the union of terminal nodes labeled as "1".

Web Appendix F: Sample codes

The following five R files are available from the Biometrics website on Wiley Online Library.

- `generate.R` is for data generation.
- `our.R` is the key function for our method. To keep it easier for the readers to implement, we have presented codes that use LASSO and grouped LASSO. We are developing an R package that will incorporate the fussed LASSO penalty.
- `output.R` provides the outputs including sensitivity, specificity, miss-classification rate, and target function value.
- `ncbCI.R` is the non-centered bootstrap for generating confidence intervals for $E_{\boldsymbol{X}}\{d(X, 1, \hat{\boldsymbol{\beta}})\}$ and $E_{\boldsymbol{X}}\{d(X, -1, \hat{\boldsymbol{\beta}})\}$.
- `main.R` generates data with `generate.R` and applies our method.

References

Fan, J. and Lv, J. (2011). Nonconcave penalized likelihood with np-dimensionality. *Information Theory, IEEE Transactions* **57,** 5467–5484.

Rinaldo, A. (2009). Properties and refinements of the fused lasso. *The Annals of Statistics* **37,** 2922–2952.