# Enhancing Treatment Effect Estimation: A Model Robust Approach Integrating Randomized Experiments and External Controls using the Double Penalty Integration Estimator (Supplementary Material)

**Yuwen Cheng**[1] **Lili Wu**[2] **Shu Yang**[3]

[1]Statistics Dept. North Carolina State University
[2]Microsoft Research NYC
[3]Statistics Dept. North Carolina State University

The supplementary material is structured as follows: Section S1 includes additional assumptions for the likelihood and density functions. Section S2 provides proofs for the main theorems. Section S3 presents the toy example mentioned in Section 2. Section S4 displays additional figures for the first simulation study in this section. The simulation codes are in the `https://github.com/yuwen997/simulation-codes`.

## S1 REGULARITY CONDITIONS

In this section, we provide the same assumptions on likelihood functions and density functions as those in Fan and Peng [2004], White [1982]. For a better understanding of these conditions, an informal summary of assumptions S1–S9 is provided here. Assumptions S1–S7 align with A1–A7 in White [1982], ensuring MLE estimator consistency and asymptotic normality in both misspecified and correct models. S8–S9 resemble F–G in Fan and Peng [2004], bounding $f$ moments.

**Assumption S1** *The independent random vectors $(p_i, Y_i)$, $i = 1, \ldots, N$, have common joint distribution function $G$ on $\Upsilon$, a measurable Euclidean space, with measurable Radon-Nikodym density $g = dG/d\mu$.*

**Assumption S2** *The family of distribution functions $F(Y_1, p, \theta)$ has Radon-Nikodym densities $f(y, p, \theta) = dF(y, p, \theta)/d\mu$ which are measurable in $(y, p)$ for every $\theta$ in $\Theta$, a compact subset of Euclidean space, and continuous in $\theta$ for every $(y, p)$ in $\Upsilon$.*

**Assumption S3** *(a) $\mathbb{E}\{\log g(Y, p)\}$ exists and $|\log f(y, p, \theta)| \leq m(y, p)$ for all $\theta$ in $\Theta$, where $m$ is integrable with respect to $G$; (b) $KLIC(g : f, \theta)$ has a unique minimum at $\theta_*$ in $\Theta$.*

**Assumption S4** *$\partial \log f(y, p, \theta)/\partial\theta_j, j = 1, \ldots, K$, are measurable functions of $(y, p)$ for each $\theta$ in $\Theta$ and continuously differentiable functions of $\theta$ for each $(y, p)$ in $\Upsilon$.*

**Assumption S5** *$|\partial^2 \log f(y, p, \theta)/\partial\theta_i\partial\theta_j|$ and $|\partial \log f(y, p, \theta)/\partial\theta_i \cdot \partial \log f(y, p, \theta)/\partial\theta_j|$, $i, j = 1, \ldots, K$ are dominated by functions integrable with respect to $G$ for each $\theta$ in $\Theta$ and $(y, p)$ in $\Upsilon$.*

**Assumption S6** *Define matrix*

$$A(\theta) = -\mathbb{E}\left\{\frac{\partial^2 \log f(Y_1, p_1, \theta)}{\partial\theta_j\partial\theta_k}\right\} > 0,$$
$$B(\theta) = \mathbb{E}\left[\left\{\frac{\partial \log f(Y_1, p_1, \theta)}{\partial\theta}\right\}\left\{\frac{\partial \log f(Y_1, p_1, \theta)}{\partial\theta}\right\}^{\mathrm{T}}\right],$$

*and (a) $\theta_*$ is interior to $\Theta$; (b) $B(\theta_*)$ is nonsingular; (c) $\theta_*$ is a regular point of $A(\theta)$.*

*Accepted for the 39th Conference on Uncertainty in Artificial Intelligence* (UAI 2023).

**Assumption S7** *$|\partial\{\partial f(y,p,\theta)/\partial\theta_i \cdot f(y,p,\theta)\}/\partial\theta_j|$, $i,j=1,\ldots,K$ are dominated by functions integrable with respect to $\mu$ for all $\theta$ in $\Theta$ and the minimal support of $f(y,p,\theta)$ does not depend on $\theta$.*

**Assumption S8** *Define matrix*

$$C(\theta) = A(\theta)B^{-1}(\theta)A(\theta).$$

*Assume matrix $A(\theta)$ and $B(\theta)$ satisfy conditions*

$$0 < C_1 < \lambda_{\min}\{A(\theta)\} \leq \lambda_{\max}\{A(\theta)\} < C_2 < \infty \text{ for all } N,$$
$$0 < C_1^* < \lambda_{\min}\{B(\theta)\} \leq \lambda_{\max}\{B(\theta)\} < C_2^* < \infty \text{ for all } N,$$

*and for $j,k=1,\ldots,K$,*

$$\mathbb{E}_\theta\left\{\frac{\partial \log f(Y_1,p_1,\theta)}{\partial\theta_j}\frac{\partial \log f(Y_1,p_1,\theta)}{\partial\theta_k}\right\}^2 < C_3 < \infty$$

*and*

$$\mathbb{E}_\theta\left\{\frac{\partial^2 \log f(Y_1,p_1,\theta)}{\partial\theta_j\partial\theta_k}\right\}^2 < C_4 < \infty.$$

**Assumption S9** *There is a large enough open subset $\omega_N$ of $\Theta \in R^K$ which contains the parameter point $\theta_*$, such that for almost all $(p_i,Y_i)$ the density admits all third derivatives $\partial f(Y_i,p_i,\theta)/\partial\theta_j\theta_k\theta_l$ for all $\theta \in \omega_N$. Furthermore, there are functions $M_{jkl}$ such that*

$$\left|\frac{\partial \log f(Y_i,p_i,\theta)}{\partial\theta_j\partial\theta_k\partial\theta_l}\right| \leq M_{jkl}(Y_i,p_i)$$

*for all $\theta \in \omega_N$, and*

$$\mathbb{E}_\theta\left\{M_{jkl}^2(Y_i,p_i)\right\} < C_5 < \infty$$

*for all $K, N, j, k$ and $l$.*

**Assumption S10** *Define $h(A,X,S \mid \beta,\delta) = \beta_{\text{int}} + \beta_A A + \beta_X^{\mathrm{T}} p_\mu(X) + (1-S)\delta^{\mathrm{T}} p_b(X)$. $f, g$ satisfy that the minimizers $\beta_*,\delta_*$ of $\mathbb{E}\{Y - h(A,X,S \mid \beta,\delta)\}^2$ are also the minimizers $\beta_*,\delta_*$ of $KLIC(g:f,\beta,\delta) = \mathbb{E}\left[\log\{g(Y,p)/f(Y,p,\beta,\delta)\}\right]$.*

## S2 PROOF

In this section, we provide proof of Theorems 1-4. Define $L(\theta) = L(\beta,\delta) = \sum_{i=1}^N \{\ln f(Y_i,p_i,\beta,\delta)\}$. Subsequently, the penalized likelihood function is $Q(\theta) = L(\theta) - N\sum_{i=1}^{K_1} P_{\lambda_{1,i}}(|\beta|) - N\sum_{i=1}^{K_2} P_{\lambda_{2,i}}(|\delta|)$. Assume $\beta_* = \begin{pmatrix}\beta_{*1}\\ \beta_{*2}\end{pmatrix}$, $\delta_* = \begin{pmatrix}\delta_{*1}\\ \delta_{*2}\end{pmatrix}$ where $\beta_{*1} \neq 0$ with $s_1$ dimensions, $\delta_{*1} \neq 0$ with $s_2$ dimensions, $\beta_{*2} = 0$ with $K_1 - s_1$ dimensions and $\delta_{*2} = 0$ with $K_2 - s_2$ dimensions. Further, let $\theta_* = \begin{pmatrix}\theta_{*1}\\ \theta_{*2}\end{pmatrix}$, where $\theta_{*1} = \begin{pmatrix}\beta_{*1}\\ \delta_{*1}\end{pmatrix} \neq 0$ with $s = s_1 + s_2$ dimensions and $\theta_{*2} = \begin{pmatrix}\beta_{*2}\\ \delta_{*2}\end{pmatrix} = 0$ with $K - s$ dimensions.

### S2.1 THEOREM S1

Theorem S1 was previously demonstrated in Lorentz [1966] and Chen [2007], and we restate it here.

**Theorem S1** *For any unknown function $f: \mathbb{R}^d \to \mathbb{R}$, assuming function $f(\cdot)$ is $t$ times continuously differentiable. Let $K = (q+1)^d$ where $x_1,\ldots,x_d$ are at least up to power $q$, and let $r^K(x)$ be the $K$-dimension power series basis function, $R(x) = A_K r^K(x)$ where $A_K$ is the matrix such that $\mathbb{E}\left\{R(X)R^{\mathrm{T}}(X)\right\} = \mathcal{I}$ where $\mathcal{I}$ is the identity matrix. Then there is a $K$-vector $\theta$ such that on the compact set $\mathcal{X}$, $\sup_{x\in\mathcal{X}}|f(x) - R^{\mathrm{T}}(x)\theta| = O\left(K^{-t/d}\right)$.*

## S2.2 PROOF FOR THEOREM 1

Under the combined dataset, the ANCOVA working model is

$$Y = \beta_{\text{int}} + \beta_A A + \beta_X^{\text{T}} p_\mu(X) + (1-S) b_0(X).$$

We can rewrite the models as

$$\begin{aligned}\tilde{Y} &= \beta_{\text{int}} + \beta_A A + \beta_X^{\text{T}} p_\mu(X) \\ &= \beta^{\text{T}} p_\mu \\ &= \bar{\mu}_{A,1}(X;\beta),\end{aligned}$$

where $\tilde{Y} := Y - (1-S)b_0(X)$. Note, $\bar{\mu}_{A,1}(X;\beta) = \beta^{\text{T}} p_\mu$. Remember $\bar{\mu}_{A,1}(X) = \beta_0 + \beta_A A + \beta_X^{\text{T}} p_\mu(X)$. The same in Proof of Theorem for Linear Models in Rosenblum and Van Der Laan [2009], the ordinary least squares estimate of $\beta$ is asymptotically normal and converges in probability to the minimizer $\tilde{\beta}_*$ of $\mathbb{E}\{\tilde{Y} - \bar{\mu}_{A,1}(X;\beta)\}^2$ . Then

$$\begin{aligned}\mathbb{E}\{\tilde{Y} - \bar{\mu}_{A,1}(X;\beta)\}^2 &= \mathbb{E}\{\tilde{Y} - \mathbb{E}\left(\tilde{Y} \mid A, X, S\right) + \mathbb{E}\left(\tilde{Y} \mid A, X, S\right) - \bar{\mu}_{A,1}(X;\beta)\}^2 \\ &= \mathbb{E}\{\tilde{Y} - \mathbb{E}\left(\tilde{Y} \mid A, X, S\right)\}^2 + \mathbb{E}\left\{\mathbb{E}\left(\tilde{Y} \mid A, X, S\right) - \bar{\mu}_{A,1}(X;\beta)\right\}^2 \\ &= \mathbb{E}\{\tilde{Y} - \mathbb{E}\left(\tilde{Y} \mid A, X, S\right)\}^2 + \mathbb{E}\left[\mathbb{E}\left\{\mathbb{E}\left(\tilde{Y} \mid A, X, S\right) - \bar{\mu}_{A,1}(X;\beta)\right\}^2 \mid S\right] \\ &= \mathbb{E}\{\tilde{Y} - \mathbb{E}\left(\tilde{Y} \mid A, X, S\right)\}^2 + \mathbb{E}\left\{\mathbb{E}\left(\tilde{Y} \mid A, X, S=1\right) - \bar{\mu}_{A,1}(X;\beta)\right\}^2 \mathbb{P}(S=1) \\ &+ \mathbb{E}\left\{\mathbb{E}\left(\tilde{Y} \mid A=0, X, S=0\right) - \bar{\mu}_{0,1}(X;\beta)\right\}^2 \mathbb{P}(S=0).\end{aligned}$$

Under HCs, by the definition of $b_0(X)$, we have

$$\begin{aligned}\mathbb{E}\left(\tilde{Y} \mid A=0, X, S=0\right) - \bar{\mu}_{0,1}(X;\beta) &= \mathbb{E}\left\{Y - b_0(X) \mid A=0, X, S=0\right\} - \bar{\mu}_{0,1}(X;\beta) \\ &= \mathbb{E}\left\{Y - \mathbb{E}\left(Y \mid A=0, X, S=0\right) + \bar{\mu}_{0,1}(X;\beta) \mid A=0, X, S=0\right\} - \bar{\mu}_{0,1}(X) \\ &= \mathbb{E}\left(Y \mid A=0, X, S=0\right) - \mathbb{E}\left(Y \mid A=0, X, S=0\right) + \bar{\mu}_{0,1}(X) - \bar{\mu}_{0,1}(X) \\ &= 0.\end{aligned}$$

Similar in [Wang et al., 2023], the $\tilde{\beta}_*$ minimizing

$$\begin{aligned}\mathbb{E}\left\{\mathbb{E}\left(\tilde{Y} \mid A, X, S=1\right) - \bar{\mu}_{A,1}(X;\beta)\right\}^2 &= \mathbb{E}\left\{\mathbb{E}\left(Y \mid A, X, S=1\right) - \bar{\mu}_{A,1}(X;\beta)\right\}^2 \\ &= \mathbb{E}\left\{\mu_{A,1}(X) - \bar{\mu}_{A,1}(X;\beta)\right\}^2.\end{aligned}$$

On the other hand, $\beta_*$ minimizes

$$\begin{aligned}\mathbb{E}[\{Y - \bar{\mu}_{A,1}(X;\beta)\}^2 \mid S=1] &= \mathbb{E}[\{Y - \mu_{A,1}(X) + \mu_{A,1}(X) - \bar{\mu}_{A,1}(X;\beta)\}^2 \mid S=1] \\ &= \mathbb{E}[\{Y - \mu_{A,1}(X)\}^2 \mid S=1] + \mathbb{E}[\{\mu_{A,1}(X) - \bar{\mu}_{A,1}(X;\beta)\}^2 \mid S=1] \\ &= \mathbb{E}[\{Y - \mu_{A,1}(X)\}^2 \mid S=1] + \mathbb{E}\{\mu_{A,1}(X) - \bar{\mu}_{A,1}(X;\beta)\}^2.\end{aligned}$$

Therefore $\tilde{\beta}_* = \beta_*$,which both minimizes $\mathbb{E}\left\{\mu_{A,1}(X) - \bar{\mu}_{A,1}(X;\beta)\right\}^2$. Similar in [Wang et al., 2023], by the first formula of taking the first derivative of $\mathbb{E}\left\{\mu_{A,1}(X) - \bar{\mu}_{A,1}(X;\beta)\right\}^2$, $\beta_*$ satisfies $\mathbb{E}\left\{\mu_{A,1}(X) - \bar{\mu}_{A,1}(X;\beta)\right\} = 0$. That is, $\beta_*$ satisfies $\tau = \mathbb{E}\left\{\mu_{1,1}(X) - \mu_{0,1}(X)\right\} = \mathbb{E}\left\{\bar{\mu}_{1,1}(X) - \bar{\mu}_{0,1}(X)\right\} = \beta_{A*}$. Similarly, in the REs, we have $\tau = \beta_{A*}$.

## S2.3 PROOF FOR THEOREM 2

Assume $g(Y,p)$ is the true density function, $f(Y,p,\theta)$ is our working density function. From Assumption S10, the selected $f$ makes minimizing the KLIC equivalent to minimizing the least square to $\mathbb{E}\{Y - h(A,X,S \mid \beta,\delta)\}^2$, therefore, $\theta_* = (\beta_*^{\text{T}}, \delta_*^{\text{T}})^{\text{T}}$ is also the parameter which minimizes the Kullback-Leibler Information Criterion (KLIC),

$$KLIC(g:f,\theta) = \mathbb{E}\left[\log\left\{g(Y,p)/f(Y,p,\theta)\right\}\right].$$

We follow the similar proofs inFan and Peng [2004], let $a_N = \sqrt{K}\left(N^{-1/2} + \alpha_N\right)$ and set $\|\mathbf{u}\|_2 = C$, where $C$ is a large enough constant, our aim is to show that for any given $\epsilon$ there is a large constant $C$ such that, for large $N$ we have

$$\mathbb{P}\left\{\sup_{\|\mathbf{u}\|_2=C} Q(\theta_* + a_N\mathbf{u}) < Q(\theta_*)\right\} \geq 1-\epsilon.$$

This implies that with probability tending to 1 there is a local maximum $\hat{\theta}$ in the call $\{\theta_* + a_N\mathbf{u} : \|\mathbf{u}\|_2 \leq C\}$ such that $\|\hat{\theta} - \theta_*\|_2 = O_p\left(a_N\right)$. Because $P_{\lambda_1}(0) = P_{\lambda_2}(0) = 0$, We have

$$\begin{aligned}
D(\mathbf{u}) &= Q(\theta_* + a_N\mathbf{u}) - Q(\theta_*) \\
&\leq \underbrace{L(\theta_* + a_N\mathbf{u}) - L(\theta_*)}_{(I)} \\
&\underbrace{-N\sum_{j=1}^{s_1}\left\{P_{\lambda_1}\left(|\beta_{*,j} + a_N u_{1j}|\right) - P_{\lambda_1}\left(||\beta_{*,j}|\right)\right\}}_{(II)} \\
&\underbrace{-N\sum_{j=1}^{s_2}\left\{P_{\lambda_2}\left(|\delta_{*,j} + a_N u_{2j}|\right) - P_{\lambda_2}\left(|\delta_{*,j}|\right)\right\}}_{(III)} \\
&:= (I) + (II) + (III),
\end{aligned}$$

where $\mathbf{u}^{\mathrm{T}} = (\mathbf{u_1^{\mathrm{T}}}, \mathbf{u_2^{\mathrm{T}}})$ with $\mathbf{u_1}$ as $K_1$ dimensions and $\mathbf{u_2}$ as $K_2$ dimensions. First for $(II)$ we have

$$\begin{aligned}
(II) &= -\sum_{j=1}^{s_1}\left[Na_N P_{\lambda_1}^{'}\left(|\beta_{*,j}|\right)\mathrm{sgn}(\beta_{*,j})u_{1j} + Na_N^2 P_{\lambda_1}^{''}\left(\beta_{*,j}\right)u_{1j}^2\left\{1+o(1)\right\}\right] \\
&:= I_1 + I_2, \\
|I_1| &\leq \sum_{j=1}^{s_1}|Na_N P_{\lambda_1}^{'}\left(|\beta_{*,j}|\right)\mathrm{sgn}(\beta_{*,j})u_{1j}| \leq \sqrt{s_1}Na_N\alpha_N\|\mathbf{u}_1\|_2 \leq Na_N^2\|\mathbf{u}\|_2, \\
|I_2| &= \sum_{j=1}^{s_1}Na_N^2 P_{\lambda_1}^{''}(|\beta_{*,j}|)u_{1j}^2\left\{1+o(1)\right\} \leq 2\max_{1\leq j\leq s_1} P_{\lambda_1}^{''}(|\beta_{*,j}|)Na_N^2\|\mathbf{u}\|_2^2.
\end{aligned}$$

Similarly for $(III)$. Then for $(I)$ we have

$$\begin{aligned}
(I) &= a_N\nabla^{\mathrm{T}}L(\theta_*)\mathbf{u} + \frac{1}{2}\mathbf{u}^{\mathrm{T}}\nabla^2 L(\theta_*)\mathbf{u}a_N^2 + \frac{1}{6}\nabla^{\mathrm{T}}\left\{\mathbf{u}^{\mathrm{T}}\nabla^2 L(\theta^*)\mathbf{u}\right\}\mathbf{u}a_N^3 \\
&:= I_3 + I_4 + I_5,
\end{aligned}$$

with the same proof in Theorem 1 in Fan and Peng [2004], by condition S8, we have

$$\begin{aligned}
|I_3| &= |a_N\nabla^{\mathrm{T}}L(\theta_*)\mathbf{u}| \leq a_N\|\nabla^{\mathrm{T}}L(\theta_*)\|_2\|\mathbf{u}\|_2 = O_p(a_N^2 N)\|\mathbf{u}\|_2. \\
I_4 &= \frac{1}{2}\mathbf{u}^{\mathrm{T}}\left\{\frac{1}{N}\left(\left[\nabla^2 L(\theta_*) - \mathbb{E}\left\{\nabla^2 L(\theta_*)\right\}\right]\right)\right\}\mathbf{u}Na_N^2 \\
&\quad - \frac{1}{2}\mathbf{u}^{\mathrm{T}}A(\theta_*)\mathbf{u}Na_N^2 \\
&= -\frac{Na_N^2}{2}\mathbf{u}^{\mathrm{T}}A(\theta_*)\mathbf{u} + o_p(1)Na_N^2\|\mathbf{u}\|_2^2.
\end{aligned}$$

By condition S9 and $K^4/N \to 0$ and $K^2\alpha_N \to 0$ as $N \to \infty$, we have

$$\begin{aligned}
|I_5| &= |\frac{1}{6}\sum_{i,j,k=1}^{K}\frac{\partial L(\theta^*)}{\partial\theta_i\partial\theta_j\partial\theta_k}u_iu_ju_ka_N^3| \\
&\leq \frac{1}{6}\sum_{l=1}^{N}\left\{\sum_{i,j,k=1}^{K}M_{ijk}^2(Y_i,p_i)\right\}^{1/2}\|\mathbf{u}\|_2^3a_N^3 \\
&= o_p(Na_N^2)\|\mathbf{u}\|_2^2.
\end{aligned}$$

Therefore, by Assumption 5 and allowing $\|\mathbf{u}\|_2$ to be large enough, all $I_1, I_2, I_3, I_5$ and $(III)$ are dominated by $I_4$, which is negative, therefore proves $\|\hat{\theta}-\theta_*\|_2 = O_p\left\{\sqrt{K}\left(N^{-1/2}+a_N\right)\right\}$. Further we have $\max\left\{\|\hat{\beta}-\beta_*\|_2, \|\hat{\delta}-\delta_*\|_2\right\} \leq \|\hat{\theta}-\theta_*\|_2 = O_p\left\{\sqrt{K}\left(N^{-1/2}+a_N\right)\right\}$. For the SCAD penalty, it is clear that $a_N = O_p\left(N^{-1/2}\right)$, therefore $\hat{\beta}$ and $\hat{\delta}$ are root-$(N/K)$-consistent penalized likelihood estimators exist with probability tending to 1, and no requirements are imposed on the convergence rate of $\lambda_1$ and $\lambda_2$.

### S2.4 PROOF FOR THEOREM 3

We follow the similar proofs in Fan and Peng [2004]. we first show that the nonconcave penalized estimator possesses the sparsity property $\hat{\theta}_2 = 0$ by the following lemma.

**Lemma 1** *Assume Assumption 5, Assumption S1–S9 are satisfied, if* $\lambda_1, \lambda_2 \to 0$, $\sqrt{N/K}\lambda_1 \to \infty$, $\sqrt{N/K}\lambda_2 \to \infty$, *and* $K^5/N \to 0$ *as* $N \to \infty$, *then first show that with probability tending to* 1, *for any given* $\theta_1$ *satisfying* $\|\theta_1-\theta_{*1}\|_2 = O_p\left(\sqrt{K/N}\right)$ *and any constant* $C$,

$$Q\left\{(\theta_1^{\mathrm{T}},0)^{\mathrm{T}}\right\} = \max_{\|\theta_2\|_2\leq C(K/N)^{1/2}} Q\left\{(\theta_1^{\mathrm{T}},\theta_2^{\mathrm{T}})^{\mathrm{T}}\right\}.$$

Proof: Let $\epsilon = C\sqrt{K/N}$. It is sufficient to show that with probability tending to 1 as $N \to \infty$, for any $\theta_1 - \theta_{*1} = O_p\left(\sqrt{K/N}\right)$ we have for $j = s+1, \ldots, K$,

$$\begin{array}{ll}
\frac{\partial Q(\theta)}{\partial\theta_j} & < 0 \quad \text{for } 0 < \theta_j < \epsilon, \\
\frac{\partial Q(\theta)}{\partial\theta_j} & > 0 \quad \text{for } -\epsilon < \theta_j < 0.
\end{array}$$

By Taylor expansion,

$$\begin{aligned}
\frac{\partial Q(\theta)}{\partial\theta_j} &= \frac{\partial L(\theta)}{\partial\theta_j} - NP_\lambda^{'}(|\theta_j|)\mathrm{sgn}\left(\theta_j\right) \\
&= \frac{\partial L(\theta_*)}{\partial\theta_j} + \sum_{l=1}^{K}\frac{\partial^2 L(\theta_*)}{\partial\theta_j\partial\theta_l}(\theta_l-\theta_{*,l}) \\
&+ \sum_{l,k=1}^{K}\frac{\partial^3 L(\theta^*)}{\partial\theta_j\partial\theta_l\partial\theta_k}(\theta_l-\theta_{*,l})\left(\theta_k-\theta_{*,k}\right) \\
&- NP_\lambda^{'}(|\theta_j|)\mathrm{sgn}\left(\theta_j\right) \\
&:= I_1 + I_2 + I_3 + I_4,
\end{aligned}$$

where $\theta^*$ lies between $\theta$ and $\theta_*$, and $P_\lambda^{'}(|\theta_j|) = P_{\lambda_1}^{'}(|\beta_j|)$ for $j = s+1, \ldots, K_1 - s_1 + s$, and $P_\lambda^{'}(|\theta_j|) = P_{\lambda_2}^{'}(|\delta_j|)$ for $j = K_1 - s_1 + s + 1, \ldots, K$.

Following the same proof in Fan and Peng [2004], we prove $I_1 + I_2 + I_3 = O_p\left(\sqrt{NK}\right)$. First, $I_1 = O_p\left(\sqrt{N}\right) = O_p\left(\sqrt{NK}\right)$. Also,

$$\begin{aligned} I_2 &= \sum_{l=1}^{K}\left(\frac{\partial^2 L(\theta_*)}{\partial\theta_j\partial\theta_l} - \mathbb{E}\left\{\frac{\partial^2 L(\theta_*)}{\partial\theta_j\partial\theta_l}\right\}\right)(\theta_l - \theta_{*,l}) \\ &+ \sum_{l=1}^{K}\mathbb{E}\left\{\frac{\partial^2 L(\theta_*)}{\partial\theta_j\partial\theta_l}\right\}(\theta_l - \theta_{*,l}) \\ &:= S_1 + S_2. \end{aligned}$$

Using the Cauchy-Schwarz inequality and $\|\theta - \theta_*\|_2 = O_p\left(K/N\right)$, we have

$$\begin{aligned} |S_2| &= |N\sum_{l=1}^{K} A(\theta_*)(j,l)(\theta_l - \theta_{*,l})| \\ &\leq NO_p\left(\sqrt{\frac{K}{N}}\right)\left\{\sum_{l=1}^{K} A^2(\theta_*)(j,l)\right\}^{1/2}. \end{aligned}$$

By Assumption S8, as the eigenvalues of the $A(\theta)$ are bounded, we have $S_2 = O_p\left(\sqrt{NK}\right)$. On the other hand,

$$|S_1| \leq \|\theta - \theta_*\|_2\left(\sum_{l=1}^{K}\left[\frac{\partial^2 L(\theta_*)}{\partial\theta_j\partial\theta_l} - \mathbb{E}\left\{\frac{\partial^2 L(\theta_*)}{\partial\theta_j\partial\theta_l}\right\}\right]^2\right)^{1/2}.$$

By Assumption S8, we have

$$\left(\sum_{l=1}^{K}\left[\frac{\partial^2 L(\theta_*)}{\partial\theta_j\partial\theta_l} - \mathbb{E}\left\{\frac{\partial^2 L(\theta_*)}{\partial\theta_j\partial\theta_l}\right\}\right]^2\right)^{1/2} = O_p\left(\sqrt{NK}\right).$$

Therefore $S_1 = O_p\left(\sqrt{NK}\right)$ and $I_2 = O_p\left(\sqrt{NK}\right)$. Further,

$$\begin{aligned} I_3 &= \sum_{l,k=1}^{K}\left[\frac{\partial^3 L(\theta^*)}{\partial\theta_j\partial\theta_l\partial\theta_k} - \mathbb{E}\left\{\frac{\partial^3 L(\theta^*)}{\partial\theta_j\partial\theta_l\partial\theta_k}\right\}\right](\theta_l - \theta_{*,l})(\theta_k - \theta_{*,k}) \\ &+ \sum_{l,k=1}^{K}\mathbb{E}\left\{\frac{\partial^3 L(\theta^*)}{\partial\theta_j\partial\theta_l\partial\theta_k}\right\}(\theta_l - \theta_{*,l})(\theta_k - \theta_{*,k}) \\ &:= S_3 + S_4. \end{aligned}$$

By Assumption S9, $|S_4| \leq C_5^{1/2} NK\|\theta - \theta_*\|_2^2 = O_p(K^2) = o_p\left(\sqrt{NK}\right)$. Further,

$$S_3^2 \leq \sum_{l,k=1}^{K}\left[\frac{\partial^3 L(\theta^*)}{\partial\theta_j\partial\theta_l\partial\theta_k} - \mathbb{E}\left\{\frac{\partial^3 L(\theta^*)}{\partial\theta_j\partial\theta_l\partial\theta_k}\right\}\right]^2\|\theta - \theta_*\|_2^4,$$

where under the Assumption S9 and Assumption 5, $S_3 = O_p\left\{\left(NK^2\frac{K^2}{N^2}\right)^{1/2}\right\} = o_p\left(\sqrt{NK}\right)$. Then

$$I_1 + I_2 + I_3 = O_p\left(\sqrt{NK}\right).$$

Because we focus on the SCAD penalty, Fan and Peng [2004] illustrates that under Assumption 5, the SCAD penalty satisfies that

$$\begin{aligned} \lim\inf_{N\to+\infty}\lim\inf_{\beta\to 0+} P_{\lambda_1}^{'}(\beta)/\lambda_1 > 0 \\ \lim\inf_{N\to+\infty}\lim\inf_{\delta\to 0+} P_{\lambda_2}^{'}(\delta)/\lambda_2 > 0, \end{aligned}$$

therefore from

$$\frac{\partial Q(\theta)}{\partial \theta_j} = N\lambda \left\{ -\frac{P_\lambda^{'}(|\theta_j|)}{\lambda}\mathrm{sgn}(\theta_j) + O_p\left(\sqrt{\frac{K}{N}}/\lambda\right)\right\},$$

where $\lambda = \lambda_1$ if $j = s+1, \ldots, K_1 - s_1 + s$ and $\lambda = \lambda_2$ if $j = K_1 - s_1 + s + 1, \ldots, K$, and $P_\lambda^{'}(|\theta_j|) = P_{\lambda_1}^{'}(|\beta_j|)$ for $j = s+1, \ldots, K_1 - s_1 + s$, and $P_\lambda^{'}(|\theta_j|) = P_{\lambda_2}^{'}(|\delta_j|)$ for $j = K_1 - s_1 + s + 1, \ldots, K$, the sign of $\theta_j$ completely determines the sign of $\partial Q(\theta)/\partial\theta_j$. We complete the proof of Lemma 1.

By Lemma 1 we prove $\hat{\theta}_2 = \begin{pmatrix}\hat{\beta}_2 \\ \hat{\delta}_2\end{pmatrix} = 0$. Then we prove the part 2.

Let

$$\begin{aligned}\Sigma &= \mathrm{diag}\left\{P_\lambda^{''}(\theta_{*,1}), \ldots, P_\lambda^{''}(\theta_{*,s})\right\} \\ &= \mathrm{diag}\left\{P_{\lambda_1}^{''}(\beta_{*,1}), \ldots, P_{\lambda_1}^{''}(\beta_{*,s_1}), P_{\lambda_2}^{''}(\delta_{*,1}), \ldots, P_{\lambda_2}^{''}(\delta_{*,s_2})\right\}\end{aligned}$$

and

$$\begin{aligned}b &= \left\{P_\lambda^{'}(|\theta_{*,1}|)\,\mathrm{sgn}(\theta_{*,1}), \ldots, P_\lambda^{'}(|\theta_{*,s}|)\,\mathrm{sgn}(\theta_{*,s})\right\}^{\mathrm{T}} \\ &= \left\{P_{\lambda_1}^{'}(|\beta_{*,1}|)\,\mathrm{sgn}(\beta_{*,1}), \ldots, P_{\lambda_1}^{'}(|\beta_{*,s_1}|)\,\mathrm{sgn}(\beta_{*,s_1}),\ P_{\lambda_2}^{'}(|\delta_{*,1}|)\,\mathrm{sgn}(\delta_{*,1}), \ldots, P_{\lambda_2}^{'}(|\delta_{*,s_2}|)\,\mathrm{sgn}(\delta_{*,s_2})\right\}^{\mathrm{T}}.\end{aligned}$$

If we can show that

$$\{A(\theta_{*1}) + \Sigma\}\left(\hat{\theta}_1 - \theta_{*1}\right) + b = \frac{1}{N}\nabla L(\theta_{*1}) + o_p(N^{-1/2}),$$

then

$$\begin{aligned}&\sqrt{N}WA^{-1/2}(\theta_{*1})\left\{A(\theta_{*1}) + \Sigma\right\}\left[\hat{\theta}_1 - \theta_{*1} + \left\{A(\theta_{*1}) + \Sigma\right\}^{-1} b\right] \\ =&\frac{1}{\sqrt{N}}WA^{-1/2}(\theta_{*1})\nabla L(\theta_{*1}) + o_p\left\{WA^{-1/2}(\theta_{*1})\right\} \\ =&\frac{1}{\sqrt{N}}WA^{-1/2}(\theta_{*1})\nabla L(\theta_{*1}) + o_p(1).\end{aligned}$$

Let $R_i = \frac{1}{\sqrt{N}}WA^{-1/2}(\theta_{*1})\nabla L_i(\theta_{*1})$, $i = 1, \ldots, N$. Following the same proof in Fan and Peng [2004], for any $\epsilon$, we have

$$\begin{aligned}\sum_{i=1}^{N}\mathbb{E}\|R_i\|_2^2\mathbf{1}\left\{\|R_i\|_2 > \epsilon\right\} &= N\mathbb{E}\|R_1\|_2^2\mathbf{1}\left\{\|R_1\|_2 > \epsilon\right\}, \\ &\le N(\mathbb{E}\|R_1\|_2^4)^{1/2}\left\{\mathbb{P}\left(\|R_1\|_2 > \epsilon\right)\right\}^{1/2}.\end{aligned}$$

By Assumption S8 and $WW^{\mathrm{T}} \to G$, we obtain

$$\mathbb{P}\left(\|R_1\|_2 > \epsilon\right) \le \frac{\mathbb{E}\|WA^{-1/2}(\theta_{*1})\nabla L_1(\theta_{*1})\|_2^2}{N\epsilon^2} = O(N^{-1})$$

and

$$\begin{aligned}\mathbb{E}\|R_1\|_2^4 &= \frac{1}{N^2}\mathbb{E}\|WA^{-1/2}(\theta_{*1})\nabla L_1(\theta_{*1})\|_2^4 \\ &\le \frac{1}{N^2}\lambda_{\max}(WW^{\mathrm{T}})\lambda_{\max}\left\{A^{-1}(\theta_{*1})\right\}\mathbb{E}\|\nabla^{\mathrm{T}}L_1(\theta_{*1})\nabla L_1(\theta_{*1})\|_2^2 \\ &\le O(\frac{K^2}{N^2}).\end{aligned}$$

Thus, we have

$$\sum_{i=1}^{N}\mathbb{E}\|R_i\|_2^2\mathbf{1}\left\{\|R_i\|_2 > \epsilon\right\} = O(N\frac{K}{N}\frac{1}{\sqrt{N}}) = o(1).$$

and

$$\sum_{i=1}^{N} \operatorname{cov}(R_i) = \operatorname{cov}\left\{ WA^{-1/2}(\theta_{*1}) \nabla L_1(\theta_{*1}) \right\} = WA^{-1/2}(\theta_{*1}) B(\theta_{*1}) A^{-1/2}(\theta_{*1}) W^{\mathrm{T}},$$

so that the $R_i$ satisfies the conditions of the Lindeberg-Feller central limit theorem. Further, using the Taylor expansion on $\nabla Q(\hat{\theta}_1)$ at the point $\theta_{*1}$, we have

$$\begin{aligned} &\frac{1}{N}\left[ \left\{ \nabla^2 L(\theta_{*1}) - \nabla^2 P_\lambda(\theta_1^{**}) \right\} \left( \hat{\theta}_1 - \theta_{*1} \right) - \nabla P_\lambda \left( \theta_{*1} \right) \right] \\ =& -\frac{1}{N}\left[ \nabla L(\theta_{*1}) + \frac{1}{2} \left( \hat{\theta}_1 - \theta_{*1} \right)^{\mathrm{T}} \nabla^2 \left\{ \nabla L\left( \theta_1^* \right) \left( \hat{\theta}_1 - \theta_{*1} \right) \right\} \right], \end{aligned}$$

where $\theta_1^*$ and $\theta_1^{**}$ lie between $\hat{\theta}_1$ and $\theta_{*1}$. Now define

$$\mathcal{L} := \nabla^2 L(\theta_{*1}) - \nabla^2 P_\lambda(\theta_1^{**})$$

and

$$C := \frac{1}{2} \left( \hat{\theta}_1 - \theta_{*1} \right)^{\mathrm{T}} \nabla^2 \left\{ \nabla L\left( \theta_1^* \right) \left( \hat{\theta}_1 - \theta_{*1} \right) \right\}.$$

Following the proof in Fan and Peng [2004], under Assumption S9 and Assumption 5 and by the Cauchy–Schwarz inequality, we have $\|1/N\mathcal{C}\|_2^2 = o_p\left(1/N\right)$. Further, we have

$$\lambda_i \left\{ \frac{1}{N}\mathcal{L} + A(\theta_{*1}) + \Sigma \right\} = o_p\left( \frac{1}{\sqrt{K}} \right), i = 1, \ldots, s,$$

where $\lambda_i(M)$ is the $i$th eigenvalue of a symmetric matrix $M$. Therefore,

$$\left\{ \frac{1}{N}\mathcal{L} + A(\theta_{*1}) + \Sigma \right\} \left( \hat{\theta}_1 - \theta_{*1} \right) = o_p\left( \frac{1}{\sqrt{N}} \right).$$

Then, we have $\left\{ A(\theta_{*1}) + \Sigma \right\} \left( \hat{\theta}_1 - \theta_{*1} \right) + b = \frac{1}{N} \nabla L(\theta_{*1}) + o_p(N^{-1/2})$, and finally we have

$$\begin{aligned} &\sqrt{N} WA^{-1/2}(\theta_{*1}) \left\{ A(\theta_{*1}) + \Sigma \right\} \left[ \hat{\theta}_1 - \theta_{*1} + \left\{ A(\theta_{*1}) + \Sigma \right\}^{-1} b \right] \\ &\to \mathcal{N}(0, WA^{-1/2}(\theta_{*1}) \left[ B(\theta_{*1}) - \mathbb{E}\left\{ \nabla L_1(\theta_{*1}) \right\} \right] A^{-1/2}(\theta_{*1}) W^{\mathrm{T}}. \end{aligned}$$

Further, based on the SCAD penalty, $\Sigma = 0$ and $b = 0$, therefore, we have

$$\sqrt{N} WA^{1/2}\left( \theta_{*1} \right) \left( \hat{\theta}_1 - \theta_{*1} \right) \to \mathcal{N}\left( 0, WA^{-1/2}(\theta_{*1}) B(\theta_{*1}) A^{-1/2}(\theta_{*1}) W^{\mathrm{T}} \right).$$

If the model is correctly specified, i.e., $g(Y, p_i) = f(Y, p_i, \theta)$ for some $\theta \in \Theta$, then $\theta_0 = \theta_*$, and

$$\sqrt{N} WI^{1/2}\left( \theta_{01} \right) \left( \hat{\theta}_1 - \theta_{01} \right) \to \mathcal{N}\left( 0, WW^{\mathrm{T}} \right).$$

We finish the second part.

## S2.5 COMMENTS ON THEOREM 2 AND THEOREM 3

It's important to clarify that if our focus isn't primarily on the least square loss, we can adjust the assumptions on $f$ and $g$ from S1–S10 to only satisfy Assumptions S1–S9 for Theorem 2 and Theorem 3. The critical point is that for any other $f$ and $g$ fulfilling Assumptions S1–S9, a local maximizer $\hat{\theta}$ of the $Q(\theta)$ exists such that $\hat{\theta}$ converges to $\theta_*$, where $\theta_*$ minimizes the KLIC between $f$ and $g$, and also exhibits the oracle property and asymptotic normality. However, in the main paper, our primary concern is the least square loss, and we further constrain $f$ to meet assumption S10 in addition to assumptions S1–S9. Assumption S10 implies that the parameters minimizing the KLIC also minimize a specific form: the least square form. This is because in the main paper, we are particularly interested in ANCOVA least square estimates.

### S2.6 PROOF FOR THEOREM 4

The consistency and the asymptotic normality of the estimator $\hat{\tau}$ are from Theorem 2 and Theorem 3. Here we only prove the calculation of the variance. If $\mathbb{V}(\epsilon) = \sigma^2$, then the influence function of $\theta$ is

$$\phi_\theta(x, y, s) = \mathbb{E}\left\{ \begin{pmatrix} p_\mu \\ (1-S)p_b(X) \end{pmatrix} \left(p_\mu^{\mathrm{T}}, \quad (1-S)p_b^{\mathrm{T}}(X)\right) \right\}^{-1} \begin{pmatrix} p_\mu \\ (1-s)p_b(x) \end{pmatrix} (y - p^{\mathrm{T}}\theta).$$

Then the asymptotic variance of $\hat{\beta}$ is

$$\begin{aligned} \mathbb{V}(\hat{\beta}) &= \mathbb{E}\left\{\phi_\theta(X, Y, S)\phi_\theta^{\mathrm{T}}(X, Y, S)\right\}_{11} \\ &= \sigma^2\mathbb{E}\left\{ \begin{pmatrix} p_\mu p_\mu^{\mathrm{T}} & (1-S)p_\mu p_b^{\mathrm{T}}(X) \\ (1-S)p_b(X)p_\mu^{\mathrm{T}} & (1-S)p_b(X)p_b^{\mathrm{T}} \end{pmatrix} \right\}^{-1}_{11} \\ &= \sigma^2\mathbb{E}\left[ Sp_\mu p_\mu^{\mathrm{T}} + (1-S)p_\mu p_\mu^{\mathrm{T}} - (1-S)\left\{p_\mu p_b^{\mathrm{T}}(X)\right\}\left\{p_b(X)p_b^{\mathrm{T}}(X)\right\}^{-1}\left\{p_b(X)p_\mu^{\mathrm{T}}\right\} \right]^{-1}. \end{aligned}$$

On the other hand, under the RE data, the ANCOVA working model is

$$Y = \beta_{\mathrm{int}} + \beta_A A + \beta_X^{\mathrm{T}} p_\mu(X) + \epsilon, \ \epsilon \sim (0, \sigma^2),$$

Similarly, the asymptotic variance of $\hat{\beta}_{\mathrm{RE}}$ is

$$\mathbb{V}(\hat{\beta}_{\mathrm{RE}}) = \sigma^2\mathbb{E}\left(Sp_\mu p_\mu^{\mathrm{T}}\right)^{-1}.$$

Then by Holder inequality,

$$\begin{aligned} &\mathbb{E}\left\{(1-S)p_\mu p_\mu^{\mathrm{T}}\right\} \\ &- \mathbb{E}\left[(1-S)\left\{p_\mu p_b^{\mathrm{T}}(X)\right\}\left\{p_b(X)p_b^{\mathrm{T}}(X)\right\}^{-1}\left\{p_b(X)p_\mu^{\mathrm{T}}\right\}\right] \geq 0, \end{aligned}$$

where the inequality holds iff $p_\mu = Mp_b(X)$ for some matrix $M$. Therefore, we have $\mathbb{V}\left\{\hat{\beta}\right\} \leq \mathbb{V}\left\{\hat{\beta}_{\mathrm{RE}}\right\}$, and the inequality holds iff $p_\mu = Mp_b(X)$ for some matrix $M$. Therefore $\mathbb{V}(\hat{\tau}_{\mathrm{RE}}) \geq \mathbb{V}(\hat{\tau})$.

## S3 TOY EXAMPLE

Consider the case $y = x^{\mathrm{T}}\beta + (1-s)x^{\mathrm{T}}\delta + \epsilon$, $x = (x_1, \ldots, x_{50})^{\mathrm{T}} \in \mathbb{R}^{50}$, $\beta^{\mathrm{T}} = (1, \ldots, 50)/10$, $\delta^{\mathrm{T}} = (1, \ldots, 50) \times 30$, $\epsilon \sim \mathcal{N}(0, 1)$, where $s$ denotes the zero-one indicator variable that determines whether the observation belongs to the REs, for simplicity, we assume $x^{\mathrm{T}}\beta$ is the correct outcome mean function of the REs, and $x^{\mathrm{T}}\delta$ is the bias function reflecting the difference between ECs and REs, i.e., if $\delta = 0$, then the observed covariates capture all confounders in the ECs and REs and thus the exchangeability assumption is valid. For didactic purposes, the magnitude of $\delta$ is much larger than the magnitude of $\beta$. Using the same regularization parameter appears to assign the same weight for $\beta$ and $\delta$, thus any penalty regularization methods tend to omit small signals $\beta$ and only pick up big signals $\delta$. Therefore, in order to make penalizations between different parameters comparable, it is crucial to add regularizations to $\beta$ and $\delta$ separately. Figure S1 shows the smoothed linear regression between $\hat{\beta}$ and $\beta$ after applying the single-penalty regularization (denoted as "Single") and the double-penalty regularizations (denoted as "Double") to select variables and refitting the model using selected variables, where double penalties make $\hat{\beta}$ more accurate than the single penalties.

## S4 PLOTS

We present the extra figures for the first simulation study in this section. Figure S2 shows the results for the case $\|\beta_0\|_1 > \|\delta_0\|_1$ when setting half of parameters in $\delta_0$ equal to zero: $\|\beta_0\|_1 = c\|\delta_0\|_1, c = 1, 3, 5, 7, 9$; and Figure S3 shows the results for varying the sparsity level of $\delta_0$ when setting $\|\beta_0\|_1 = \|\delta_0\|_1$ with the x axis as the the number of variables in $\delta_0$ equal to zero. Each figure shows the MSE results and the percentage of Under-select and Over-select. Figure S2 shows the SPIE has a larger MSE compared to the DPIE, which is consistent with the theoretical results. On the other hand, the changes in the sparsity of $\delta_0$ affect little on the results. This finding also consists of the theoretical result, where we only need to restrain the magnitude of different parameters to guarantee the consistency and oracle properties.

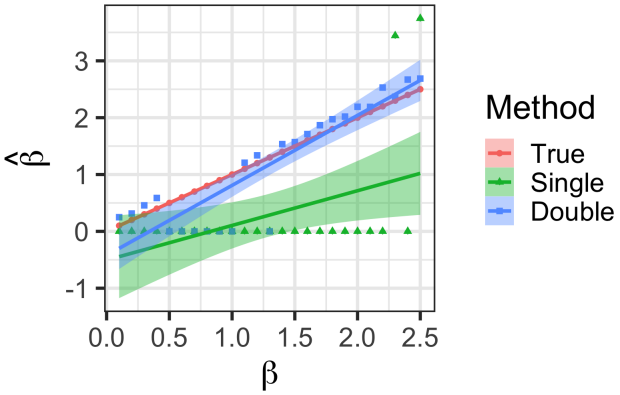


Figure S1: The smoothed linear regression between $\hat{\beta}$ and $\beta$ with the 95% confidence intervals as the shaded area and $(\beta, \hat{\beta})$ as the points.

**References**

Xiaohong Chen. Large sample sieve estimation of semi-nonparametric models. *Handbook of Econometrics*, 6:5549–5632, 2007.

Jianqing Fan and Heng Peng. Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, 32(3):928–961, 2004.

G.G. Lorentz. *Approximation of Functions*. Holt, Rinehart and Winston, 1966.

Michael Rosenblum and Mark J Van Der Laan. Using regression models to analyze randomized trials: Asymptotically valid hypothesis tests despite incorrectly specified models. *Biometrics*, 65(3):937–945, 2009.

Bingkai Wang, Ryoko Susukida, Ramin Mojtabai, Masoumeh Amin-Esmaeili, and Michael Rosenblum. Model-robust inference for clinical trials that improve precision by stratified randomization and covariate adjustment. *Journal of the American Statistical Association*, 118(542):1152–1163, 2023.

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

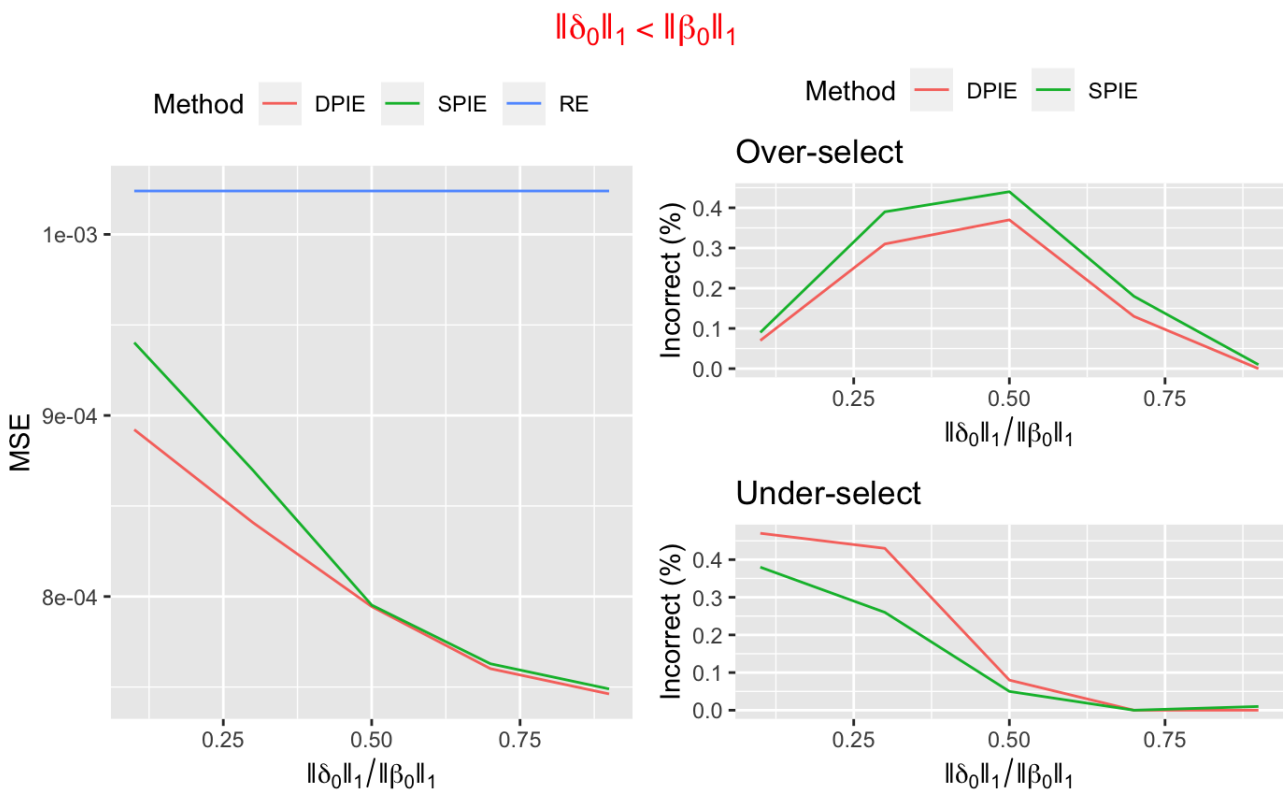


Figure S2: Simulation results based on 100 Monte Carlo times. The left panel shows the MSE versus the magnitude ratio between $\delta$ and $\beta$. The right panel shows the percentage of wrongly choosing more and less parameters, separately.

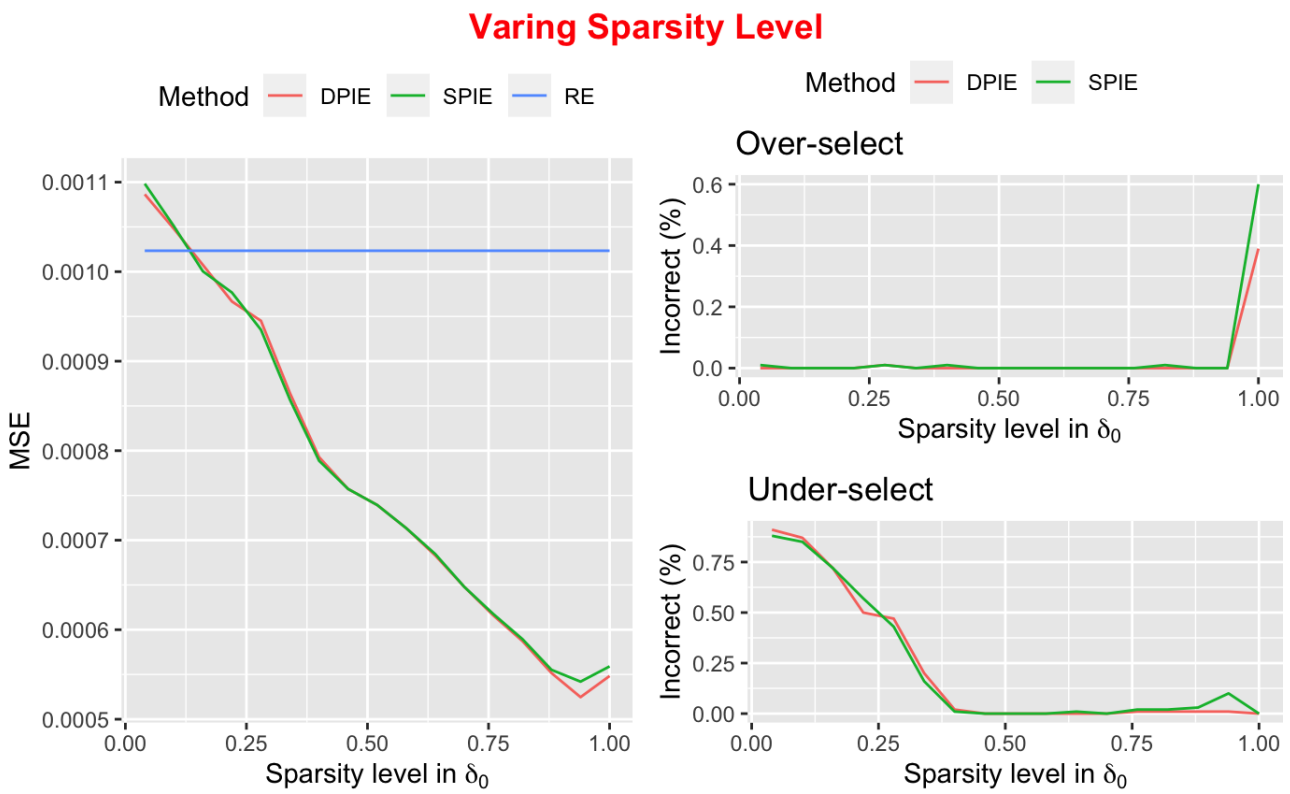


Figure S3: Simulation results based on 100 Monte Carlo times. The left panel shows the MSE versus the sparsity level in $\delta$. The right panel shows the percentage of wrongly choosing more and less parameters, separately.