A Proofs

A.1 Regularity Conditions

(C1) The observation $\{\mathbf{V}_i : i = 1, \ldots, n\}$ are independent and identically distributed with a probability density $f(\mathbf{V}, \mathbf{\Phi})$, which has a common support. We assume the density f satisfies the following equations:

$$E_{\mathbf{\Phi}}\left[\nabla_{\boldsymbol{\phi}_j} \log f\left(\boldsymbol{V}, \mathbf{\Phi}\right)\right] = \mathbf{0} \quad \text{for } j = 1, \ldots, 2p+1.$$

and

$$\begin{aligned}\mathbf{I}_{j_1k_1j_2k_2}(\mathbf{\Phi}) &= E_{\mathbf{\Phi}}\left[\frac{\partial}{\partial\phi_{j_1k_1}} \log f(V, \mathbf{\Phi}) \cdot \frac{\partial}{\partial\phi_{j_2k_2}} \log f(V, \mathbf{\Phi})\right] \\ &= E_{\mathbf{\Phi}}\left[-\frac{\partial^2}{\partial\phi_{j_1k_1}\phi_{j_2k_2}} \log f(V, \mathbf{\Phi})\right],\end{aligned}$$

for any $j_1, j_2 = 1, \ldots, 2p+1$, and $k_1 = 1, \ldots, p_{j1}$, $k_2 = 1, \ldots, p_{j2}$, where j_1, j_2 are the index of group, k_1, k_2 be the index of elements within the corresponding group, p_{j_1}, p_{j_2} are the group size of j_1, j_2 respectively.

(C2) The Fisher information matrix

$$\mathbf{I}\left(\mathbf{\Phi}\right) = E\left[\left(\frac{\partial}{\partial\mathbf{\Phi}} \log f(V, \mathbf{\Phi})\right)\left(\frac{\partial}{\partial\mathbf{\Phi}} \log f(V, \mathbf{\Phi})\right)^{\top}\right]$$

is finite and positive definite at $\mathbf{\Phi} = \mathbf{\Phi}^*$.

(C3) There exists an open set ω of Ω that contains the true parameter point $\mathbf{\Phi}^*$ such that for almost all $\mathbf{V}$ the density $f(\mathbf{V}, \mathbf{\Phi})$ admits all third derivatives $\frac{\partial^3 f(\mathbf{V}, \mathbf{\Phi})}{\partial\phi_{j_1k_1}\partial\phi_{j_2k_2}\partial\phi_{j_3k_3}}$ for all $\mathbf{\Phi}$ in ω and any $j_1, j_2, j_3 = 1, \ldots, 2p+1$, and $k_1 = 1, \ldots, p_{j1}$, $k_2 = 1, \ldots, p_{j2}$ and

$k_3 = 1, \ldots, p_{j3}$. Furthermore, there exist functions $M_{j_1k_1j_2k_2j_3k_3}$ such that

$$\left| \frac{\partial^3}{\partial \phi_{j_1k_1} \partial \phi_{j_2k_2} \partial \phi_{j_3k_3}} \log f(\mathbf{V}, \mathbf{\Phi}) \right| \leq M_{j_1k_1j_2k_2j_3k_3}(\mathbf{V}) \quad \text{for all } \mathbf{\Phi} \in \omega,$$

and $m_{j_1k_1j_2k_2j_3k_3} = E_{\mathbf{\Phi}^*}[M_{j_1k_1j_2k_2j_3k_3}(\mathbf{V})] < \infty$.

A.2 Lemma (1) proof

Let $\eta_n = \frac{1}{\sqrt{n}} + a_n$ and $\{\mathbf{\Phi}^* + \eta_n \boldsymbol{\delta} : \|\boldsymbol{\delta}\|_2 \leq C\}$ be the ball around $\mathbf{\Phi}^*$ for $\boldsymbol{\delta} \in \mathbb{R}^d$, where d is the dimension of the design matrix and C is some constant. Under the regularity assumptions, we show that there exists a local minimizer $\widehat{\mathbf{\Phi}}_n$ of $Q_n(\mathbf{\Phi})$ such that $\|\widehat{\mathbf{\Phi}}_n - \mathbf{\Phi}^*\|_2 = O_p(\frac{1}{\sqrt{n}})$. For this proof, we adopt the approaches outlined in [8, 13, 23, 37] and extend it to our situation. Let $\eta_n = \frac{1}{\sqrt{n}} + a_n$ and $\{\mathbf{\Phi}^* + \eta_n \boldsymbol{\delta} : \|\boldsymbol{\delta}\|_2 \leq C\}$ be the ball around $\mathbf{\Phi}^*$ for $\boldsymbol{\delta} = (\mathbf{u}_1^\top, \mathbf{u}_2^\top, , \ldots, \mathbf{u}_{p+1}^\top, \mathbf{u}_{p+2}^\top, \ldots, \mathbf{u}_{2p+1}^\top)^\top \in \mathbb{R}^d$, where d is the dimension of the design matrix and C is some constant. The objective function is given by

$$Q_n(\mathbf{\Phi}) = -L_n(\mathbf{\Phi}) + n\lambda_m \sum_{m=1}^{2p+1} \|\boldsymbol{\phi}_m\|_2 \,,$$

Define

$$D_n(\boldsymbol{\delta}) \equiv Q_n(\mathbf{\Phi}^* + \eta_n \boldsymbol{\delta}) - Q_n(\mathbf{\Phi}^*).$$

Then for $\boldsymbol{\delta}$ that satisfies $\|\boldsymbol{\delta}\|_2 = C$, we have

$$
\begin{aligned}
D_n(\boldsymbol{\delta}) &= -L_n(\boldsymbol{\Phi}^* + \eta_n \boldsymbol{\delta}) + L_n(\boldsymbol{\Phi}^*) + n \sum_{m=1}^{2p+1} \lambda_m (\|\boldsymbol{\theta}_m^* + \eta_n \mathbf{u}_m\|_2 - \|\boldsymbol{\theta}_m^*\|_2) \\
&\overset{(a)}{\geq} -L_n(\boldsymbol{\Phi}^* + \eta_n \boldsymbol{\delta}) + L_n(\boldsymbol{\Phi}^*) + n \sum_{m \in \mathcal{A}_1} \lambda_m^\theta (\|\boldsymbol{\theta}_m^* + \eta_n \mathbf{u}_m\|_2 - \|\boldsymbol{\theta}_m^*\|_2) \\
&\quad + n \sum_{m \in \mathcal{A}_2} \lambda_m^\theta (\|\boldsymbol{\theta}_m^* + \eta_n \mathbf{u}_m\|_2 - \|\boldsymbol{\theta}_m^*\|_2) \\
&\overset{(b)}{\geq} -L_n(\boldsymbol{\Phi}^* + \eta_n \boldsymbol{\delta}) + L_n(\boldsymbol{\Phi}^*) - n\eta_n \sum_{m \in \mathcal{A}_1} \lambda_m \|\mathbf{u}_m\|_2 - n\eta_n \sum_{m \in \mathcal{A}_2} \lambda_m \|\mathbf{u}_m\|_2 \\
&\overset{(c)}{\geq} -L_n(\boldsymbol{\Phi}^* + \eta_n \boldsymbol{\delta}) + L_n(\boldsymbol{\Phi}^*) - n\eta_n^2 \sum_{m \in \mathcal{A}_1} \|\mathbf{u}_m\|_2 - n\eta_n^2 \sum_{m \in \mathcal{A}_2} \|\mathbf{u}_m\|_2 \\
&\geq -L_n(\boldsymbol{\Phi}^* + \eta_n \boldsymbol{\delta}) + L_n(\boldsymbol{\Phi}^*) - n\eta_n^2 (|\mathcal{A}_1| + |\mathcal{A}_2|) C \\
&\overset{(d)}{=} -[\nabla L_n(\boldsymbol{\Phi}^*)]^\top (\eta_n \boldsymbol{\delta}) - \frac{1}{2} (\eta_n \boldsymbol{\delta})^\top [\nabla^2 L_n(\boldsymbol{\Phi}^*)] (\eta_n \boldsymbol{\delta}) (1 + o(1)) \\
&\quad - n\eta_n^2 (|\mathcal{A}_1| + |\mathcal{A}_2|) C
\end{aligned} \tag{15}
$$

Inequality (a) is by the fact that $\sum_{m \notin \mathcal{A}_1} \|\boldsymbol{\phi}_m^*\|_2 = 0$ and $\sum_{m \notin \mathcal{A}_2} \|\boldsymbol{\phi}_m^*\|_2 = 0$. Inequality (b) is due to the reverse triangle inequality $\|a\|_2 - \|b\|_2 \geq -\|a - b\|_2$. Inequality (c) is by $\lambda_m \leq a_n \leq \eta_n$ for $m \in \mathcal{A}_1$ and $m \in \mathcal{A}_2$. Equality (d) is by the standard argument on the Taylor expansion of the loss function:

$$
\begin{aligned}
L_n(\boldsymbol{\Phi}^* + \eta_n \boldsymbol{\delta}) &= L_n(\boldsymbol{\Phi}^* + \eta_n \cdot \mathbf{0}) + \eta_n \nabla L_n(\boldsymbol{\Phi}^* + \eta_n \cdot \mathbf{0})^\top (\boldsymbol{\delta} - \mathbf{0}) \\
&\qquad + \frac{1}{2} (\boldsymbol{\delta} - \mathbf{0})^\top \nabla^2 L_n(\boldsymbol{\Phi}^* + \eta_n \cdot \mathbf{0}) (\boldsymbol{\delta} - \mathbf{0}) \{1 + o(1)\} \\
&= L_n(\boldsymbol{\Phi}^*) + \eta_n \nabla L_n(\boldsymbol{\Phi}^*)^\top \boldsymbol{\delta} + \frac{1}{2} \boldsymbol{\delta}^\top \nabla^2 L_n(\boldsymbol{\Phi}^*) \boldsymbol{\delta} \eta_n^2 \{1 + o(1)\}
\end{aligned}
$$

We split (15) into three parts:

$$
\begin{aligned}
D_1 &= -\left[\nabla L_n\left(\boldsymbol{\Phi}^*\right)\right]^{\mathrm{T}}\left(\eta_n \boldsymbol{\delta}\right) \\
D_2 &= -\frac{1}{2}\left(\eta_n \boldsymbol{\delta}\right)^{\top}\left[\nabla^2 L_n\left(\boldsymbol{\Phi}^*\right)\right]\left(\eta_n \boldsymbol{\delta}\right)(1+o(1)) \\
D_3 &= -n\eta_n^2(|\mathcal{A}_1|+|\mathcal{A}_2|)C
\end{aligned}
$$

Then

$$
\begin{aligned}
D_1 &= -\eta_n\left[\nabla L_n\left(\boldsymbol{\Phi}^*\right)\right]^{\top} \boldsymbol{\delta} \\
&= -\sqrt{n}\eta_n\left(\frac{1}{\sqrt{n}} \nabla L_n\left(\boldsymbol{\Phi}^*\right)\right)^{\top} \boldsymbol{\delta} \\
&= -\sqrt{n}\eta_n\left(\sqrt{n}\frac{1}{n} \sum_{i=1}^n \nabla \log f\left(\boldsymbol{V}_i, \boldsymbol{\Phi}\right)\big|_{\boldsymbol{\Phi}=\boldsymbol{\Phi}^*}\right)^{\top} \boldsymbol{\delta} \\
&= -\sqrt{n}\eta_n\left(\sqrt{n}\left[\frac{1}{n} \sum_{i=1}^n \nabla \log f\left(\boldsymbol{V}_i, \boldsymbol{\Phi}\right)\big|_{\boldsymbol{\Phi}=\boldsymbol{\Phi}^*}-\mathbf{0}\right]\right)^{\top} \boldsymbol{\delta} \\
&= -\sqrt{n}\eta_n\left(\sqrt{n}\left[\frac{1}{n} \sum_{i=1}^n \nabla \log f\left(\boldsymbol{V}_i, \boldsymbol{\Phi}\right)\big|_{\boldsymbol{\Phi}=\boldsymbol{\Phi}^*}-E_{\boldsymbol{\Phi}^*} \nabla L\left(\boldsymbol{\Phi}^*\right)\right]\right)^{\top} \boldsymbol{\delta} \\
&= -\sqrt{n}\eta_n O_P(1)\, \boldsymbol{\delta} \\
&= -O_P\left(n\eta_n^2\right) \boldsymbol{\delta}
\end{aligned} \tag{16}
$$

The last equation is by $a_n = o(\frac{1}{\sqrt{n}})$ and

$$
\begin{aligned}
O_P(n\eta_n^2) &= O_P(n(n^{-1/2}+a_n)^2) = O_P(1+2n^{1/2}a_n+na_n^2)) \\
&= O_P(1+n^{1/2}a_n+(n^{1/2}a_n)^2) = O_P(1+n^{1/2}a_n+o(1)) \\
&= O_p(n^{1/2}(n^{-1/2}+a_n)) = O_p(n^{1/2}\eta_n)
\end{aligned}
$$

$$
\begin{aligned}
D_2 &= \frac{1}{2} n\eta_n^2 \left\{ \boldsymbol{\delta}^\top \left[-\frac{1}{n} \nabla^2 L_n\left(\boldsymbol{\Phi}^*\right) \right] \boldsymbol{\delta} \right\} (1 + o_p(1)) \\
&= \frac{1}{2} n\eta_n^2 \left\{ \boldsymbol{\delta}^\top \left[\mathbf{I}\left(\boldsymbol{\Phi}^*\right) \right] \boldsymbol{\delta} \right\} (1 + o_p(1)) \text{ by the weak law of large numbers.} \\
&= O_p(n\eta_n^2 \|\boldsymbol{\delta}\|_2^2) && (17)
\end{aligned}
$$

Combining (16) and (17) with (15) gives:

$$
\begin{aligned}
D_n(\boldsymbol{\delta}) &\geq D_1 + D_2 + D_3 \\
&= -O_P\left(n\eta_n^2\right) \boldsymbol{\delta} + O_p(n\eta_n^2 \|\boldsymbol{\delta}\|_2^2) - n\eta_n^2 (|\mathcal{A}_1| + |\mathcal{A}_2|) C
\end{aligned}
$$

We can see that the first term D_1 is linear in $\boldsymbol{\delta}$ and the second term D_2 is quadratic in $\boldsymbol{\delta}$. We can conclude that for a large enough constant $C = \|\boldsymbol{\delta}\|_2$, D_2 dominates D_1 and D_3. Note that this is a positive term since $I(\boldsymbol{\Phi})$ is positive definite at $\boldsymbol{\Phi} = \boldsymbol{\Phi}^*$ by regularity condition (C2). Therefore, for each $\varepsilon > 0$, there exists a large enough constant C such that, for large enough n

$$
P\left\{ \inf_{\|\boldsymbol{\delta}\|_2 = C} D_n\left(\boldsymbol{\delta}\right) > 0 \right\} \geq 1 - \varepsilon
$$

This implies with probability at least $1 - \varepsilon$ that the empirical likelihood Q_n has a local minimizer in the ball $\{\boldsymbol{\Phi}^* + \eta_n \boldsymbol{\delta} : \|\boldsymbol{\delta}\|_2 \leq C\}$ (since Q_n is bounded and $\{\boldsymbol{\Phi}^* + \alpha_n \boldsymbol{\delta} : \|\boldsymbol{\delta}\|_2 \leq C\}$ is closed). In other words, there exists a local solution $\widehat{\boldsymbol{\Phi}}_n$ such that $\|\widehat{\boldsymbol{\Phi}}_n - \boldsymbol{\Phi}^*\| \leq \eta_n \|\boldsymbol{\delta}\|_2 \leq \eta_n C = O_P(\eta_n) = O_P(\frac{1}{\sqrt{n}} + a_n) = O_p(\frac{1}{\sqrt{n}})$, since $a_n = o(\frac{1}{\sqrt{n}})$. Hence, $\left\| \widehat{\boldsymbol{\Phi}}_n - \boldsymbol{\Phi}^* \right\|_2 = O_P\left(\frac{1}{\sqrt{n}}\right)$.

□

A.3 Theorem 1 proof

We first consider consistency for the main effects $P\left(\widehat{\boldsymbol{\Phi}}_{\mathcal{A}_1^c} = \mathbf{0}\right) \to 1$. Following [8, 13], it is sufficient to show that for all $m \in \mathcal{A}_1^c$, $P\left(\widehat{\boldsymbol{\phi}}_m = \mathbf{0}\right) \to 1$, which implies that $P\left(\widehat{\boldsymbol{\Phi}}_{\mathcal{A}_1^c} = \mathbf{0}\right) \to$

1, i.e., the $\sqrt{n}$-consistent estimate $\widehat{\mathbf{\Phi}}$ has oracle property $\widehat{\boldsymbol{\phi}}_m = \mathbf{0}$ if $\boldsymbol{\phi}_m^* = \mathbf{0}$. Denote

$$\widehat{\boldsymbol{\phi}}_m = (\hat{\phi}_{m1}, \ldots, \hat{\phi}_{mp_m}),$$

where p_m is the group size of $\widehat{\boldsymbol{\phi}}_m$. Let $\hat{\phi}_{mk}$ be the k-th entry of $\widehat{\boldsymbol{\phi}}_m$. Note that if $\widehat{\boldsymbol{\phi}}_m \neq \mathbf{0}$, then $\hat{\phi}_{mk} \neq 0$ for $k = 1, \ldots, p_m$, then penalty function $\|\widehat{\boldsymbol{\phi}}_m\|_2$ becomes differentiable. Therefore ϕ_{mk} for $k = 1, \ldots, p_m$ must satisfy the following normal equation

$$\begin{aligned}
\frac{\partial Q_n\left(\widehat{\mathbf{\Phi}}_n\right)}{\partial \phi_{mk}} &= -\frac{\partial L_n\left(\widehat{\mathbf{\Phi}}_n\right)}{\partial \phi_{mk}} + n\lambda_m \frac{\hat{\phi}_{mk}}{\|\widehat{\boldsymbol{\phi}}_m\|_2} \\
&= -\frac{\partial L_n\left(\mathbf{\Phi}^*\right)}{\partial \phi_{mk}} - \sum_{j_1=1}^{2p+1} \sum_{k_1=1}^{p_{j_1}} \frac{\partial^2 L_n\left(\mathbf{\Phi}^*\right)}{\partial \phi_{mk} \partial \phi_{j_1 k_1}} \left(\hat{\phi}_{j_1 k_1} - \phi^*_{j_1 k_1}\right) \\
&\quad - \frac{1}{2} \sum_{j_1=1}^{2p+1} \sum_{k_1=1}^{p_{j_1}} \sum_{j_2=1}^{2p+1} \sum_{k_2=1}^{p_{j_2}} \frac{\partial^3 L_n(\widetilde{\mathbf{\Phi}})}{\partial \phi_{mk} \partial \phi_{j_1 k_1} \partial \phi_{j_2 k_2}} \left(\hat{\phi}_{j_1 k_1} - \phi^*_{j_1 k_1}\right) \left(\hat{\phi}_{j_2 k_2} - \phi^*_{j_2 k_2}\right) \\
&\quad + n\lambda_m \frac{\hat{\phi}_{mk}}{\|\widehat{\boldsymbol{\phi}}_m\|_2} \triangleq I_1 + I_2 + I_3 + I_4 = 0
\end{aligned}$$

where $\widetilde{\mathbf{\Phi}}$ lies between $\widehat{\mathbf{\Phi}}_n$ and $\mathbf{\Phi}^*$. By the regularity conditions and Lemma (1) that $\left\|\widehat{\mathbf{\Phi}}_n - \mathbf{\Phi}^*\right\|_2 = O_P\left(\frac{1}{\sqrt{n}}\right)$, the first term is of the order $O_p(\sqrt{n})$

$$I_1 = -\frac{\partial L_n\left(\widehat{\mathbf{\Phi}}_n\right)}{\partial \phi_{mk}} = -\sqrt{n}\sqrt{n}\frac{1}{n}\frac{\partial L_n\left(\widehat{\mathbf{\Phi}}_n\right)}{\partial \phi_{mk}} = \sqrt{n} O_p(1) = O_p(\sqrt{n}).$$

Then the second is of the order $O_P\left(\frac{1}{\sqrt{n}}\right)$ and the third term is of the order $O_P\left(\frac{1}{n}\right)$. Hence

$$\frac{\partial Q_n\left(\widehat{\mathbf{\Phi}}_n\right)}{\partial \mathbf{\Phi}_m} = \sqrt{n}\left\{O_p(1) + \sqrt{n}\lambda_m \frac{\hat{\phi}_{mk}}{\|\widehat{\boldsymbol{\phi}}_m\|_2}\right\}. \tag{18}$$

As $\sqrt{n}\lambda_m \geq \sqrt{n} b_n \to \infty$ for $m \in \mathcal{A}_1^c$ from the assumption, therefore we know that I_4 dominates I_1, I_2 and I_3 in (18) with probability tending to one. This means that (18) cannot be true as long as the sample size is sufficiently large. As a result, we can conclude that with probability tending to one, the estimate $\widehat{\boldsymbol{\phi}}_m = (\hat{\phi}_{m1}, \ldots, \hat{\phi}_{mp_m})$ must be in a position where $\widehat{\boldsymbol{\phi}}_m$ is not differentiable. Hence $\widehat{\boldsymbol{\phi}}_m = \mathbf{0}$ for all $m \in \mathcal{A}_1^c$. Hence $P\left(\widehat{\boldsymbol{\Phi}}_{\mathcal{A}_1^c} = \mathbf{0}\right) \to 1$. This completes the proof.

Next, we prove that for the interactions $P\left(\widehat{\boldsymbol{\Phi}}_{\mathcal{A}_2^c} = \mathbf{0}\right) \to 1$. For $m \in \mathcal{A}_2^c$ s.t. $\boldsymbol{\phi}_m^* = \gamma_{jE}^* = 0$ but $\beta_E \neq 0$ and $\boldsymbol{\theta}_j^* \neq \mathbf{0} \quad (1 \leq j \leq p)$, we can prove $P\left(\widehat{\boldsymbol{\Phi}}_{\mathcal{A}_2^c} = \mathbf{0}\right) \to 1$ by a similar reasoning, which further implies that $P(\hat{\gamma}_{jE} = 0) \to 0$. For $m \in \mathcal{A}_2^c$ such that $\boldsymbol{\phi}_m^* = \gamma_{jE}^* = 0$ and either $\beta_E = 0$ or $\boldsymbol{\theta}_j^* = \mathbf{0} \quad (1 \leq j \leq p)$: without loss of generality, assume that $\boldsymbol{\theta}_j^* = \mathbf{0}$. Notice that $\hat{\boldsymbol{\theta}}_j = \mathbf{0}$ implies $\hat{\gamma}_{jE} = 0$, since if $\hat{\gamma}_{jE} \neq 0$, the value of the loss function does not change but the value of the penalty function will increase. Because we already prove $P\left(\widehat{\boldsymbol{\Phi}}_{\mathcal{A}_1^c} = \mathbf{0}\right) \to 1$, therefore we get $P\left(\widehat{\boldsymbol{\Phi}}_{\mathcal{A}_2^c} = \mathbf{0}\right) \to 1$ as well for this case.

□

A.4 Theorem 2 proof

By Lemma (1) and Theorem (1), there exists a $\widehat{\boldsymbol{\Phi}}_{\mathcal{A}}$ that is a $\sqrt{n}$-consistent local minimizer of $Q(\boldsymbol{\Phi}_{\mathcal{A}})$, therefore $\left\|\widehat{\boldsymbol{\Phi}}_{\mathcal{A}} - \boldsymbol{\Phi}_{\mathcal{A}}^*\right\|_2 = O_P\left(\frac{1}{\sqrt{n}}\right)$ and $P\left(\widehat{\boldsymbol{\Phi}}_{\mathcal{A}^c} = \mathbf{0}\right) \to 1$. Thus satisfies (with probability tending to 1):

$$\left.\frac{\partial Q_n\left(\boldsymbol{\Phi}_{\mathcal{A}}\right)}{\partial \boldsymbol{\Phi}_m}\right|_{\boldsymbol{\Phi}=\begin{pmatrix}\widehat{\boldsymbol{\Phi}}_{\mathcal{A}} \\ 0\end{pmatrix}} = 0, \quad \forall m \in \mathcal{A}, \tag{19}$$

that is

$$\left.\frac{\partial Q_n\left(\boldsymbol{\Phi}_{\mathcal{A}}\right)}{\partial \boldsymbol{\Phi}_m}\right|_{\boldsymbol{\Phi}_{\mathcal{A}}=\widehat{\boldsymbol{\Phi}}_{\mathcal{A}}} = 0, \quad \forall m \in \mathcal{A}, \tag{20}$$

where

$$Q_n(\boldsymbol{\Phi}_{\mathcal{A}}) = -L_n(\boldsymbol{\Phi}_{\mathcal{A}}) + \underbrace{n\sum_{m\in\mathcal{A}_1}\lambda_m\|\boldsymbol{\phi}_m\|_2 + n\sum_{m\in\mathcal{A}_2}\lambda_m\|\boldsymbol{\phi}_m\|_2}_{\triangleq nP(\boldsymbol{\Phi}_{\mathcal{A}})}$$
$$= -L_n(\boldsymbol{\Phi}_{\mathcal{A}}) + nP(\boldsymbol{\Phi}_{\mathcal{A}}). \tag{21}$$

655 From (20) and (21) we have

$$\nabla_{\mathcal{A}}Q_n\left(\widehat{\boldsymbol{\Phi}}_{\mathcal{A}}\right) = -\nabla_{\mathcal{A}}L_n\left(\widehat{\boldsymbol{\Phi}}_{\mathcal{A}}\right) + n\nabla_{\mathcal{A}}P\left(\widehat{\boldsymbol{\Phi}}_{\mathcal{A}}\right) = \mathbf{0}, \tag{22}$$

656 with probability tending to 1.

657 Denote $\boldsymbol{\Sigma} = \operatorname{diag}\{o_p(1),\ldots,o_p(1)\}$. We then expand $-\nabla_{\mathcal{A}}L_n\left(\boldsymbol{\Phi}_{\mathcal{A}}\right)$ at $\boldsymbol{\Phi}_{\mathcal{A}} = \boldsymbol{\Phi}^*_{\mathcal{A}}$ in (22):

$$\begin{aligned}
-\nabla_{\mathcal{A}}L_n\left(\widehat{\boldsymbol{\Phi}}_{\mathcal{A}}\right) &= -\nabla_{\mathcal{A}}L_n\left(\boldsymbol{\Phi}^*_{\mathcal{A}}\right) - \left[\nabla^2_{\mathcal{A}}L_n\left(\boldsymbol{\Phi}^*_{\mathcal{A}}\right) + \boldsymbol{\Sigma}\right]\left(\widehat{\boldsymbol{\Phi}}_{\mathcal{A}} - \boldsymbol{\Phi}^*_{\mathcal{A}}\right)\\
&= \sqrt{n}\left[-\frac{1}{\sqrt{n}}\nabla_{\mathcal{A}}L_n\left(\boldsymbol{\Phi}^*_{\mathcal{A}}\right) + \left(-\frac{1}{n}\nabla^2_{\mathcal{A}}L_n\left(\boldsymbol{\Phi}^*_{\mathcal{A}}\right) - \boldsymbol{\Sigma}\right)\sqrt{n}\left(\widehat{\boldsymbol{\Phi}}_{\mathcal{A}} - \boldsymbol{\Phi}^*_{\mathcal{A}}\right)\right]\\
&= \sqrt{n}\left[-\frac{1}{\sqrt{n}}\nabla_{\mathcal{A}}L_n\left(\boldsymbol{\Phi}^*_{\mathcal{A}}\right) + \left(\mathbf{I}\left(\boldsymbol{\Phi}^*_{\mathcal{A}}\right) - \boldsymbol{\Sigma}\right)\sqrt{n}\left(\widehat{\boldsymbol{\Phi}}_{\mathcal{A}} - \boldsymbol{\Phi}^*_{\mathcal{A}}\right)\right].
\end{aligned}$$

658 The third line follows by

$$\frac{1}{n}\nabla^2_{\mathcal{A}}L_n\left(\boldsymbol{\Phi}^*_{\mathcal{A}}\right) = E\left\{\nabla^2_{\mathcal{A}}L\left(\boldsymbol{\Phi}^*_{\mathcal{A}}\right)\right\} + \boldsymbol{\Sigma} = -\mathbf{I}\left(\boldsymbol{\Phi}^*_{\mathcal{A}}\right) + \boldsymbol{\Sigma}.$$

659 Denote

$$\mathbf{b} = \left(\lambda_m\operatorname{sgn}\left(\beta^*_m\right), \lambda_m\frac{\boldsymbol{\theta}^*_m}{\|\boldsymbol{\theta}^*_m\|_2}^{\top}, \lambda_m\operatorname{sgn}(\gamma^*_{mE})\right)^{\top}, \qquad m\in\mathcal{A},$$

We also expand $n\nabla_{\mathcal{A}}P\left(\boldsymbol{\Phi}_{\mathcal{A}}\right)$ at $\boldsymbol{\Phi}_{\mathcal{A}} = \boldsymbol{\Phi}^*_{\mathcal{A}}$ in (22):

$$n\nabla_{\mathcal{A}}P\left(\widehat{\boldsymbol{\Phi}}_{\mathcal{A}}\right) = n\left[\mathbf{b} + \boldsymbol{\Sigma}\left(\widehat{\boldsymbol{\Phi}}_{\mathcal{A}} - \boldsymbol{\Phi}^*_{\mathcal{A}}\right)\right].$$

And due to the fact that $\sqrt{n}\lambda_m \le \sqrt{n}a_n \to 0$ for $m \in \mathcal{A}$ and $\frac{\theta^*_{mk}}{\|\boldsymbol{\theta}^*_m\|_2} \le 1$ for any $1 \le k \le p_m$, we know that $\sqrt{n}\mathbf{b} = (o_p(1), \ldots, o_p(1))^\top$ Thus,

$$\begin{aligned}\nabla_{\mathcal{A}} Q_n\left(\widehat{\boldsymbol{\Phi}}_{\mathcal{A}}\right) &= \sqrt{n}\left[-\frac{1}{\sqrt{n}}\nabla_{\mathcal{A}} L_n\left(\boldsymbol{\Phi}^*_{\mathcal{A}}\right) + \left(\mathbf{I}\left(\boldsymbol{\Phi}^*_{\mathcal{A}}\right) + \boldsymbol{\Sigma}\right)\sqrt{n}\left(\widehat{\boldsymbol{\Phi}}_{\mathcal{A}} - \boldsymbol{\Phi}^*_{\mathcal{A}}\right)\right] \\ &\quad + \sqrt{n}\left[\sqrt{n}\mathbf{b} + \boldsymbol{\Sigma}\sqrt{n}\left(\widehat{\boldsymbol{\Phi}}_{\mathcal{A}} - \boldsymbol{\Phi}^*_{\mathcal{A}}\right)\right] \\ &= \sqrt{n}\left[-\frac{1}{\sqrt{n}}\nabla_{\mathcal{A}} L_n\left(\boldsymbol{\Phi}^*_{\mathcal{A}}\right) + \sqrt{n}\mathbf{b} + \left(\mathbf{I}\left(\boldsymbol{\Phi}^*_{\mathcal{A}}\right) + \boldsymbol{\Sigma}\right)\sqrt{n}\left(\widehat{\boldsymbol{\Phi}}_{\mathcal{A}} - \boldsymbol{\Phi}^*_{\mathcal{A}}\right)\right] \\ &= \mathbf{0}.\end{aligned}$$

$$\left(\mathbf{I}\left(\boldsymbol{\Phi}^*_{\mathcal{A}}\right) + \boldsymbol{\Sigma}\right)\sqrt{n}(\widehat{\boldsymbol{\Phi}}_{\mathcal{A}} - \boldsymbol{\Phi}^*_{\mathcal{A}}) = \sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\nabla_{\mathcal{A}}\log f\left(\boldsymbol{V}_i, \boldsymbol{\Phi}^*_{\mathcal{A}}\right) + o_p(1).$$

Therefore, by the central limit theorem, we know that

$$\sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n}\nabla_{\mathcal{A}}\log f(V_i, \boldsymbol{\Phi}^*_{\mathcal{A}})\right] \to N(\mathbf{0}, \mathbf{I}(\boldsymbol{\Phi}^*_{\mathcal{A}})).$$

Hence,

$$\sqrt{n}\left(\widehat{\boldsymbol{\Phi}}_{\mathcal{A}} - \boldsymbol{\Phi}^*_{\mathcal{A}}\right) \xrightarrow{d} N\left(\mathbf{0}, \mathbf{I}^{-1}\left(\boldsymbol{\Phi}^*_{\mathcal{A}}\right)\right).$$

$\square$

B Algorithm Details

In this section we provide more specific details about the algorithms used to solve the `sail` objective function. The strong heredity `sail` model with least-squares loss has the form

$$\hat{Y} = \beta_0 \cdot \mathbf{1} + \sum_{j=1}^{p}\boldsymbol{\Psi}_j\boldsymbol{\theta}_j + \beta_E X_E + \sum_{j=1}^{p}\gamma_j\beta_E(X_E \circ \boldsymbol{\Psi}_j)\boldsymbol{\theta}_j \tag{23}$$

and the objective function is given by

$$Q(\mathbf{\Phi}) = \frac{1}{2n}\left\|Y - \hat{Y}\right\|_2^2 + \lambda(1-\alpha)\left(w_E|\beta_E| + \sum_{j=1}^{p} w_j \|\boldsymbol{\theta}_j\|_2\right) + \lambda\alpha \sum_{j=1}^{p} w_{jE}|\gamma_j| \tag{24}$$

Solving (24) in a blockwise manner allows us to leverage computationally fast algorithms for ℓ_1 and ℓ_2 norm penalized regression. Denote the n-dimensional residual column vector $R = Y - \hat{Y}$. The subgradient equations are given by

$$\frac{\partial Q}{\partial \beta_0} = \frac{1}{n}\left(Y - \beta_0 \cdot \mathbf{1} - \sum_{j=1}^{p} \mathbf{\Psi}_j \boldsymbol{\theta}_j - \beta_E X_E - \sum_{j=1}^{p} \gamma_j \beta_E (X_E \circ \mathbf{\Psi}_j)\boldsymbol{\theta}_j\right)^\top \mathbf{1} = 0 \tag{25}$$

$$\frac{\partial Q}{\partial \beta_E} = -\frac{1}{n}\left(X_E + \sum_{j=1}^{p} \gamma_j (X_E \circ \mathbf{\Psi}_j)\boldsymbol{\theta}_j\right)^\top R + \lambda(1-\alpha) w_E s_1 = 0 \tag{26}$$

$$\frac{\partial Q}{\partial \boldsymbol{\theta}_j} = -\frac{1}{n}\left(\mathbf{\Psi}_j + \gamma_j \beta_E (X_E \circ \mathbf{\Psi}_j)\right)^\top R + \lambda(1-\alpha) w_j s_2 = \mathbf{0} \tag{27}$$

$$\frac{\partial Q}{\partial \gamma_j} = -\frac{1}{n}\left(\beta_E (X_E \circ \mathbf{\Psi}_j)\boldsymbol{\theta}_j\right)^\top R + \lambda\alpha w_{jE} s_3 = 0 \tag{28}$$

where s_1 is in the subgradient of the ℓ_1 norm:

$$s_1 \in \begin{cases} \operatorname{sign}(\beta_E) & \text{if } \beta_E \neq 0 \\ [-1, 1] & \text{if } \beta_E = 0, \end{cases}$$

s_2 is in the subgradient of the ℓ_2 norm:

$$s_2 \in \begin{cases} \dfrac{\boldsymbol{\theta}_j}{\|\boldsymbol{\theta}_j\|_2} & \text{if } \boldsymbol{\theta}_j \neq \mathbf{0} \\ u \in \mathbb{R}^{m_j} : \|u\|_2 \leq 1 & \text{if } \boldsymbol{\theta}_j = \mathbf{0}, \end{cases}$$

and s_3 is in the subgradient of the ℓ_1 norm:

$$s_3 \in \begin{cases} \operatorname{sign}(\gamma_j) & \text{if } \gamma_j \neq 0 \\ [-1, 1] & \text{if } \gamma_j = 0. \end{cases}$$

Define the partial residuals, without the jth predictor for $j = 1, \ldots, p$, as

$$R_{(-j)} = Y - \beta_0 \cdot \mathbf{1} - \sum_{\ell \neq j} \boldsymbol{\Psi}_\ell \boldsymbol{\theta}_\ell - \beta_E X_E - \sum_{\ell \neq j} \gamma_\ell \beta_E (X_E \circ \boldsymbol{\Psi}_\ell) \boldsymbol{\theta}_\ell$$

the partial residual without X_E as

$$R_{(-E)} = Y - \beta_0 \cdot \mathbf{1} - \sum_{j=1}^{p} \boldsymbol{\Psi}_j \boldsymbol{\theta}_j$$

and the partial residual without the jth interaction for $j = 1, \ldots, p$, as

$$R_{(-jE)} = Y - \beta_0 \cdot \mathbf{1} - \sum_{j=1}^{p} \boldsymbol{\Psi}_j \boldsymbol{\theta}_j - \beta_E X_E - \sum_{\ell \neq j} \gamma_\ell \beta_E (X_E \circ \boldsymbol{\Psi}_\ell) \boldsymbol{\theta}_\ell$$

From the subgradient equations (25)–(28) we see that

$$\hat{\beta}_0 = \left(Y - \sum_{j=1}^{p} \boldsymbol{\Psi}_j \hat{\boldsymbol{\theta}}_j - \hat{\beta}_E X_E - \sum_{j=1}^{p} \hat{\gamma}_j \hat{\beta}_E (X_E \circ \boldsymbol{\Psi}_j) \hat{\boldsymbol{\theta}}_j \right)^{\top} \mathbf{1} \tag{29}$$

$$\hat{\beta}_E = \frac{S\left(\frac{1}{n \cdot w_E} \left(X_E + \sum_{j=1}^{p} \hat{\gamma}_j (X_E \circ \boldsymbol{\Psi}_j) \hat{\boldsymbol{\theta}}_j \right)^{\top} R_{(-E)}, \lambda(1-\alpha) \right)}{\left(X_E + \sum_{j=1}^{p} \hat{\gamma}_j (X_E \circ \boldsymbol{\Psi}_j) \hat{\boldsymbol{\theta}}_j \right)^{\top} \left(X_E + \sum_{j=1}^{p} \hat{\gamma}_j (X_E \circ \boldsymbol{\Psi}_j) \hat{\boldsymbol{\theta}}_j \right)} \tag{30}$$

$$\lambda(1-\alpha) w_j \frac{\boldsymbol{\theta}_j}{\|\boldsymbol{\theta}_j\|_2} = \frac{1}{n} \left(\boldsymbol{\Psi}_j + \gamma_j \beta_E (X_E \circ \boldsymbol{\Psi}_j) \right)^{\top} R_{(-j)} \tag{31}$$

$$\hat{\gamma}_j = \frac{S\left(\frac{1}{n \cdot w_{jE}} \left(\beta_E (X_E \circ \boldsymbol{\Psi}_j) \boldsymbol{\theta}_j \right)^{\top} R_{(-jE)}, \lambda\alpha \right)}{\left(\beta_E (X_E \circ \boldsymbol{\Psi}_j) \boldsymbol{\theta}_j \right)^{\top} \left(\beta_E (X_E \circ \boldsymbol{\Psi}_j) \boldsymbol{\theta}_j \right)} \tag{32}$$

where $S(x, t) = \operatorname{sign}(x)(|x| - t)$ is the soft-thresholding operator. We see from (29) and (30) that there are closed form solutions for the intercept and β_E. From (32), each γ_j also has a closed form solution and can be solved efficiently for $j = 1, \ldots, p$ using a coordinate descent procedure [14]. Since there is no closed form solution for β_j, we use a quadratic majorization technique [38] to solve (31). Furthermore, we update each $\boldsymbol{\theta}_j$ in a coordinate wise fashion and leverage this to implement further computational speedups which are detailed in Supplemental Section B.2. From these estimates, we compute the interaction effects using the reparametrizations presented in Table 1, e.g., $\hat{\boldsymbol{\tau}}_j = \hat{\gamma}_j \hat{\beta}_E \hat{\boldsymbol{\theta}}_j$, $j = 1, \ldots, p$ for the strong heredity `sail` model.

B.1 Least-Squares `sail` with Strong Heredity

A more detailed algorithm for fitting the least-squares `sail` model with strong heredity is given in Algorithm 3.

Algorithm 3 Blockwise Coordinate Descent for Least-Squares `sail` with Strong Heredity

function `sail`$(\boldsymbol{X}, Y, X_E, \texttt{basis}, \lambda, \alpha, w_j, w_E, w_{jE}, \epsilon)$ ▷ Algorithm for solving (24)

$\Psi_j \leftarrow \texttt{basis}(X_j)$, $\widetilde{\Psi}_j \leftarrow X_E \circ \Psi_j$ for $j = 1, \ldots, p$

Initialize: $\beta_0^{(0)} \leftarrow \bar{Y}$, $\beta_E^{(0)} = \boldsymbol{\theta}_j^{(0)} = \gamma_j^{(0)} \leftarrow 0$ for $j = 1, \ldots, p$.

Set iteration counter $k \leftarrow 0$

$R^* \leftarrow Y - \beta_0^{(k)} - \beta_E^{(k)} X_E - \sum_j (\boldsymbol{\Psi}_j + \gamma_j^{(k)} \beta_E^{(k)} \widetilde{\boldsymbol{\Psi}}_j) \boldsymbol{\theta}_j^{(k)}$

repeat

- To update $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)$

$\widetilde{X}_j \leftarrow \beta_E^{(k)} \widetilde{\boldsymbol{\Psi}}_j \boldsymbol{\theta}_j^{(k)}$ for $j = 1, \ldots, p$

$R \leftarrow R^* + \sum_{j=1}^{p} \gamma_j^{(k)} \widetilde{X}_j$

$$\boldsymbol{\gamma}^{(k)(new)} \leftarrow \arg\min_{\boldsymbol{\gamma}} \frac{1}{2n} \left\| R - \sum_j \gamma_j \widetilde{X}_j \right\|_2^2 + \lambda\alpha \sum_j w_{jE} |\gamma_j|$$

$\Delta = \sum_j (\gamma_j^{(k)} - \gamma_j^{(k)(new)}) \widetilde{X}_j$

$R^* \leftarrow R^* + \Delta$

- To update $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p)$

$\widetilde{X}_j \leftarrow \boldsymbol{\Psi}_j + \gamma_j^{(k)} \beta_E^{(k)} \widetilde{\boldsymbol{\Psi}}_j$ for $j = 1, \ldots, p$

for $j = 1, \ldots, p$ **do**

$R \leftarrow R^* + \widetilde{X}_j \boldsymbol{\theta}_j^{(k)}$

$$\boldsymbol{\theta}_j^{(k)(new)} \leftarrow \arg\min_{\boldsymbol{\theta}_j} \frac{1}{2n} \left\| R - \widetilde{X}_j \boldsymbol{\theta}_j \right\|_2^2 + \lambda(1-\alpha) w_j \|\theta_j\|_2$$

$\Delta = \widetilde{X}_j (\boldsymbol{\theta}_j^{(k)} - \boldsymbol{\theta}_j^{(k)(new)})$

$R^* \leftarrow R^* + \Delta$

- To update β_E

$\widetilde{X}_E \leftarrow X_E + \sum_j \gamma_j^{(k)} \widetilde{\boldsymbol{\Psi}}_j \boldsymbol{\theta}_j^{(k)}$

$R \leftarrow R^* + \beta_E^{(k)} \widetilde{X}_E$

$$\beta_E^{(k)(new)} \leftarrow \frac{1}{\widetilde{X}_E^\top \widetilde{X}_E} S\left(\frac{1}{n \cdot w_E} \widetilde{X}_E^\top R, \lambda(1-\alpha) \right)$$

▷ $S(x, t) = \text{sign}(x)(|x| - t)_+$

$\Delta = (\beta_E^{(k)} - \beta_E^{(k)(new)}) \widetilde{X}_E$

$R^* \leftarrow R^* + \Delta$

- To update β_0

$R \leftarrow R^* + \beta_0^{(k)}$

$$\beta_0^{(k)(new)} \leftarrow \frac{1}{n} R^* \cdot \mathbf{1}$$

$\Delta = \beta_0^{(k)} - \beta_0^{(k)(new)}$

$R^* \leftarrow R^* + \Delta$

$k \leftarrow k + 1$

until convergence criterion is satisfied: $\left| Q(\boldsymbol{\Phi}^{(k-1)}) - Q(\boldsymbol{\Phi}^{(k)}) \right| / Q(\boldsymbol{\Phi}^{(k-1)}) < \epsilon$

683 B.2 Details on Update for $\boldsymbol{\theta}$

Here we discuss a computational speedup in the updates for the $\boldsymbol{\theta}$ parameter. The partial residual (R_s) used for updating $\boldsymbol{\theta}_s$ ($s \in 1, \ldots, p$) at the kth iteration is given by

$$R_s = Y - \widetilde{Y}^{(k)}_{(-s)} \tag{33}$$

where $\widetilde{Y}^{(k)}_{(-s)}$ is the fitted value at the kth iteration excluding the contribution from $\boldsymbol{\Psi}_s$:

$$\widetilde{Y}^{(k)}_{(-s)} = \beta_0^{(k)} - \beta_E^{(k)} X_E - \sum_{\ell \neq s} \boldsymbol{\Psi}_\ell \boldsymbol{\theta}_\ell^{(k)} - \sum_{\ell \neq s} \gamma_\ell^{(k)} \beta_E^{(k)} \widetilde{\boldsymbol{\Psi}}_\ell \boldsymbol{\theta}_\ell^{(k)} \tag{34}$$

Using (34), (33) can be re-written as

$$\begin{aligned} R_s &= Y - \beta_0^{(k)} - \beta_E^{(k)} X_E - \sum_{j=1}^{p} (\boldsymbol{\Psi}_j + \gamma_j^{(k)} \beta_E^{(k)} \widetilde{\boldsymbol{\Psi}}_j) \boldsymbol{\theta}_j^{(k)} + (\boldsymbol{\Psi}_s + \gamma_s^{(k)} \beta_E^{(k)} \widetilde{\boldsymbol{\Psi}}_s) \boldsymbol{\theta}_s^{(k)} \\ &= R^* + (\boldsymbol{\Psi}_s + \gamma_s^{(k)} \beta_E^{(k)} \widetilde{\boldsymbol{\Psi}}_s) \boldsymbol{\theta}_s^{(k)} \end{aligned} \tag{35}$$

684 where

$$R^* = Y - \beta_0^{(k)} - \beta_E^{(k)} X_E - \sum_{j=1}^{p} (\boldsymbol{\Psi}_j + \gamma_j^{(k)} \beta_E^{(k)} \widetilde{\boldsymbol{\Psi}}_j) \boldsymbol{\theta}_j^{(k)} \tag{36}$$

Denote $\boldsymbol{\theta}_s^{(k)(\mathrm{new})}$ the solution for predictor s at the kth iteration, given by:

$$\boldsymbol{\theta}_s^{(k)(new)} = \arg\min_{\boldsymbol{\theta}_j} \frac{1}{2n} \left\| R_s - (\boldsymbol{\Psi}_s + \gamma_s^{(k)} \beta_E^{(k)} \widetilde{\boldsymbol{\Psi}}_s) \boldsymbol{\theta}_j \right\|_2^2 + \lambda(1-\alpha) w_s \|\theta_j\|_2 \tag{37}$$

Now we want to update the parameters for the next predictor $\boldsymbol{\theta}_{s+1}$ ($s+1 \in 1, \ldots, p$) at the kth iteration. The partial residual used to update $\boldsymbol{\theta}_{s+1}$ is given by

$$R_{s+1} = R^* + (\boldsymbol{\Psi}_{s+1} + \gamma_{s+1}^{(k)} \beta_E^{(k)} \widetilde{\boldsymbol{\Psi}}_{s+1}) \boldsymbol{\theta}_{s+1}^{(k)} + (\boldsymbol{\Psi}_s + \gamma_s^{(k)} \beta_E^{(k)} \widetilde{\boldsymbol{\Psi}}_s)(\boldsymbol{\theta}_s^{(k)} - \boldsymbol{\theta}_s^{(k)(new)}) \tag{38}$$

where R^* is given by (36), $\boldsymbol{\theta}_s^{(k)}$ is the parameter value prior to the update, and $\boldsymbol{\theta}_s^{(k)(new)}$ is the updated value given by (37). Taking the difference between (35) and (38) gives

$$\begin{aligned}\Delta &= R_t - R_s \\ &= (\boldsymbol{\Psi}_t + \gamma_t^{(k)}\beta_E^{(k)}\widetilde{\boldsymbol{\Psi}}_t)\boldsymbol{\theta}_t^{(k)} + (\boldsymbol{\Psi}_s + \gamma_s^{(k)}\beta_E^{(k)}\widetilde{\boldsymbol{\Psi}}_s)(\boldsymbol{\theta}_s^{(k)} - \boldsymbol{\theta}_s^{(k)(new)}) - (\boldsymbol{\Psi}_s + \gamma_s^{(k)}\beta_E^{(k)}\widetilde{\boldsymbol{\Psi}}_s)\boldsymbol{\theta}_s^{(k)} \\ &= (\boldsymbol{\Psi}_t + \gamma_t^{(k)}\beta_E^{(k)}\widetilde{\boldsymbol{\Psi}}_t)\boldsymbol{\theta}_t^{(k)} - (\boldsymbol{\Psi}_s + \gamma_s^{(k)}\beta_E^{(k)}\widetilde{\boldsymbol{\Psi}}_s)\boldsymbol{\theta}_s^{(k)(new)}\end{aligned} \tag{39}$$

Therefore $R_t = R_s + \Delta$, and the partial residual for updating the next predictor can be computed by updating the previous partial residual by Δ, given by (39). This formulation can lead to computational speedups especially when $\Delta = 0$, meaning the partial residual does not need to be re-calculated.

B.3 Maximum penalty parameter (λ_{max}) for strong heredity

The subgradient equations (26)–(28) can be used to determine the largest value of λ such that all coefficients are 0. From the subgradient Equation (26), we see that $\beta_E = 0$ is a solution if

$$\frac{1}{w_E}\left|\frac{1}{n}\left(X_E + \sum_{j=1}^{p}\gamma_j(X_E \circ \boldsymbol{\Psi}_j)\boldsymbol{\theta}_j\right)^\top R_{(-E)}\right| \leq \lambda(1-\alpha) \tag{40}$$

From the subgradient Equation (27), we see that $\boldsymbol{\theta}_j = \mathbf{0}$ is a solution if

$$\frac{1}{w_j}\left\|\frac{1}{n}\left(\boldsymbol{\Psi}_j + \gamma_j\beta_E(X_E \circ \boldsymbol{\Psi}_j)\right)^\top R_{(-j)}\right\|_2 \leq \lambda(1-\alpha) \tag{41}$$

From the subgradient Equation (28), we see that $\gamma_j = 0$ is a solution if

$$\frac{1}{w_{jE}}\left|\frac{1}{n}\left(\beta_E(X_E \circ \boldsymbol{\Psi}_j)\boldsymbol{\theta}_j\right)^\top R_{(-jE)}\right| \leq \lambda\alpha \tag{42}$$

Due to the strong heredity property, the parameter vector $(\beta_E, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p, \gamma_1, \ldots, \gamma_p)$ will be entirely equal to $\mathbf{0}$ if $(\beta_E, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p) = \mathbf{0}$. Therefore, the smallest value of λ for which the entire parameter vector (excluding the intercept) is $\mathbf{0}$ is:

$$\begin{aligned}\lambda_{max} = \frac{1}{n(1-\alpha)} \max \Bigg\{ & \frac{1}{w_E} \left(X_E + \sum_{j=1}^{p} \gamma_j (X_E \circ \boldsymbol{\Psi}_j) \boldsymbol{\theta}_j \right)^\top R_{(-E)}, \\ & \max_j \frac{1}{w_j} \left\| \left(\boldsymbol{\Psi}_j + \gamma_j \beta_E (X_E \circ \boldsymbol{\Psi}_j) \right)^\top R_{(-j)} \right\|_2 \Bigg\} \end{aligned} \tag{43}$$

which reduces to

$$\lambda_{max} = \frac{1}{n(1-\alpha)} \max \left\{ \frac{1}{w_E} \left(X_E \right)^\top R_{(-E)}, \max_j \frac{1}{w_j} \left\| \left(\boldsymbol{\Psi}_j \right)^\top R_{(-j)} \right\|_2 \right\}$$

B.4 Least-Squares `sail` with Weak Heredity

The least-squares `sail` model with weak heredity has the form

$$\hat{Y} = \beta_0 \cdot \mathbf{1} + \sum_{j=1}^{p} \boldsymbol{\Psi}_j \boldsymbol{\theta}_j + \beta_E X_E + \sum_{j=1}^{p} \gamma_j (X_E \circ \boldsymbol{\Psi}_j)(\beta_E \cdot \mathbf{1}_{m_j} + \boldsymbol{\theta}_j) \tag{44}$$

The objective function is given by

$$Q(\boldsymbol{\Phi}) = \frac{1}{2n} \left\| Y - \hat{Y} \right\|_2^2 + \lambda(1-\alpha) \left(w_E |\beta_E| + \sum_{j=1}^{p} w_j \|\boldsymbol{\theta}_j\|_2 \right) + \lambda\alpha \sum_{j=1}^{p} w_{jE} |\gamma_j| \tag{45}$$

Denote the n-dimensional residual column vector $R = Y - \hat{Y}$. The subgradient equations are given by

$$\frac{\partial Q}{\partial \beta_0} = \frac{1}{n}\left(Y - \beta_0 \cdot \mathbf{1} - \sum_{j=1}^{p} \boldsymbol{\Psi}_j \boldsymbol{\theta}_j - \beta_E X_E - \sum_{j=1}^{p} \gamma_j (X_E \circ \boldsymbol{\Psi}_j)(\beta_E \cdot \mathbf{1}_{m_j} + \boldsymbol{\theta}_j)\right)^{\top} \mathbf{1} = 0 \quad (46)$$

$$\frac{\partial Q}{\partial \beta_E} = -\frac{1}{n}\left(X_E + \sum_{j=1}^{p} \gamma_j (X_E \circ \boldsymbol{\Psi}_j) \mathbf{1}_{m_j}\right)^{\top} R + \lambda(1-\alpha) w_E s_1 = 0 \quad (47)$$

$$\frac{\partial Q}{\partial \boldsymbol{\theta}_j} = -\frac{1}{n}\left(\boldsymbol{\Psi}_j + \gamma_j (X_E \circ \boldsymbol{\Psi}_j)\right)^{\top} R + \lambda(1-\alpha) w_j s_2 = \mathbf{0} \quad (48)$$

$$\frac{\partial Q}{\partial \gamma_j} = -\frac{1}{n}\left((X_E \circ \boldsymbol{\Psi}_j)(\beta_E \cdot \mathbf{1}_{m_j} + \boldsymbol{\theta}_j)\right)^{\top} R + \lambda \alpha w_{jE} s_3 = 0 \quad (49)$$

where s_1 is in the subgradient of the ℓ_1 norm:

$$s_1 \in \begin{cases} \text{sign}\,(\beta_E) & \text{if } \beta_E \neq 0 \\ [-1, 1] & \text{if } \beta_E = 0, \end{cases}$$

s_2 is in the subgradient of the ℓ_2 norm:

$$s_2 \in \begin{cases} \dfrac{\boldsymbol{\theta}_j}{\|\boldsymbol{\theta}_j\|_2} & \text{if } \boldsymbol{\theta}_j \neq \mathbf{0} \\ u \in \mathbb{R}^{m_j} : \|u\|_2 \leq 1 & \text{if } \boldsymbol{\theta}_j = \mathbf{0}, \end{cases}$$

and s_3 is in the subgradient of the ℓ_1 norm:

$$s_3 \in \begin{cases} \text{sign}\,(\gamma_j) & \text{if } \gamma_j \neq 0 \\ [-1, 1] & \text{if } \gamma_j = 0. \end{cases}$$

Define the partial residuals, without the jth predictor for $j = 1, \ldots, p$, as

$$R_{(-j)} = Y - \beta_0 \cdot \mathbf{1} - \sum_{\ell \neq j} \boldsymbol{\Psi}_\ell \boldsymbol{\theta}_\ell - \beta_E X_E - \sum_{\ell \neq j} \gamma_\ell (X_E \circ \boldsymbol{\Psi}_\ell)(\beta_E \cdot \mathbf{1}_{m_\ell} + \boldsymbol{\theta}_\ell)$$

the partial residual without X_E as

$$R_{(-E)} = Y - \beta_0 \cdot \mathbf{1} - \sum_{j=1}^{p} \mathbf{\Psi}_j \boldsymbol{\theta}_j - \sum_{j=1}^{p} \gamma_j (X_E \circ \mathbf{\Psi}_j) \boldsymbol{\theta}_j$$

and the partial residual without the jth interaction for $j = 1, \ldots, p$

$$R_{(-jE)} = Y - \beta_0 \cdot \mathbf{1} - \sum_{j=1}^{p} \mathbf{\Psi}_j \boldsymbol{\theta}_j - \beta_E X_E - \sum_{\ell \neq j} \gamma_\ell (X_E \circ \mathbf{\Psi}_\ell)(\beta_E \cdot \mathbf{1}_{m_\ell} + \boldsymbol{\theta}_\ell)$$

From the subgradient Equation (47), we see that $\beta_E = 0$ is a solution if

$$\frac{1}{w_E} \left| \frac{1}{n} \left(X_E + \sum_{j=1}^{p} \gamma_j (X_E \circ \mathbf{\Psi}_j) \mathbf{1}_{m_j} \right)^\top R_{(-E)} \right| \leq \lambda(1 - \alpha) \tag{50}$$

From the subgradient Equation (48), we see that $\boldsymbol{\theta}_j = \mathbf{0}$ is a solution if

$$\frac{1}{w_j} \left\| \frac{1}{n} \left(\mathbf{\Psi}_j + \gamma_j (X_E \circ \mathbf{\Psi}_j) \right)^\top R_{(-j)} \right\|_2 \leq \lambda(1 - \alpha) \tag{51}$$

From the subgradient Equation (49), we see that $\gamma_j = 0$ is a solution if

$$\frac{1}{w_{jE}} \left| \frac{1}{n} \big((X_E \circ \mathbf{\Psi}_j)(\beta_E \cdot \mathbf{1}_{m_j} + \boldsymbol{\theta}_j) \big)^\top R_{(-jE)} \right| \leq \lambda \alpha \tag{52}$$

From the subgradient equations we see that

$$\hat{\beta}_0 = \left(Y - \sum_{j=1}^{p} \boldsymbol{\Psi}_j \hat{\boldsymbol{\theta}}_j - \hat{\beta}_E X_E - \sum_{j=1}^{p} \hat{\gamma}_j (X_E \circ \boldsymbol{\Psi}_j)(\hat{\beta}_E \cdot \mathbf{1}_{m_j} + \hat{\boldsymbol{\theta}}_j)\right)^\top \mathbf{1} \tag{53}$$

$$\hat{\beta}_E = \frac{S\left(\frac{1}{n \cdot w_E}\left(X_E + \sum_{j=1}^{p} \hat{\gamma}_j (X_E \circ \boldsymbol{\Psi}_j)\mathbf{1}_{m_j}\right)^\top R_{(-E)}, \lambda(1-\alpha)\right)}{\left(X_E + \sum_{j=1}^{p} \hat{\gamma}_j (X_E \circ \boldsymbol{\Psi}_j)\mathbf{1}_{m_j}\right)^\top \left(X_E + \sum_{j=1}^{p} \hat{\gamma}_j (X_E \circ \boldsymbol{\Psi}_j)\mathbf{1}_{m_j}\right)} \tag{54}$$

$$\lambda(1-\alpha) w_j \frac{\boldsymbol{\theta}_j}{\|\boldsymbol{\theta}_j\|_2} = \frac{1}{n}\left(\boldsymbol{\Psi}_j + \gamma_j (X_E \circ \boldsymbol{\Psi}_j)\right)^\top R_{(-j)} \tag{55}$$

$$\hat{\gamma}_j = \frac{S\left(\frac{1}{n \cdot w_{jE}}\left((X_E \circ \boldsymbol{\Psi}_j)(\beta_E \cdot \mathbf{1}_{m_j} + \boldsymbol{\theta}_j)\right)^\top R_{(-jE)}, \lambda\alpha\right)}{\left((X_E \circ \boldsymbol{\Psi}_j)(\beta_E \cdot \mathbf{1}_{m_j} + \boldsymbol{\theta}_j)\right)^\top \left((X_E \circ \boldsymbol{\Psi}_j)(\beta_E \cdot \mathbf{1}_{m_j} + \boldsymbol{\theta}_j)\right)} \tag{56}$$

where $S(x,t) = \text{sign}(x)(|x| - t)$ is the soft-thresholding operator. As was the case in the strong heredity `sail` model, there are closed form solutions for the intercept and β_E, each γ_j also has a closed form solution and can be solved efficiently for $j = 1, \ldots, p$ using the coordinate descent procedure implemented in the `glmnet` package [14], while we use the quadratic majorization technique implemented in the `gglasso` package [38] to solve (55). Algorithm 4 details the procedure used to fit the least-squares weak heredity `sail` model.

B.4.1 Maximum penalty parameter (λ_{max}) for weak heredity

The smallest value of λ for which the entire parameter vector $(\beta_E, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p, \gamma_1, \ldots, \gamma_p)$ is $\mathbf{0}$ is:

$$\begin{aligned}\lambda_{max} = \frac{1}{n} \max \Bigg\{ & \frac{1}{(1-\alpha) w_E}\left(X_E + \sum_{j=1}^{p} \gamma_j (X_E \circ \boldsymbol{\Psi}_j)\mathbf{1}_{m_j}\right)^\top R_{(-E)}, \\ & \max_j \frac{1}{(1-\alpha) w_j}\left\|\left(\boldsymbol{\Psi}_j + \gamma_j (X_E \circ \boldsymbol{\Psi}_j)\right)^\top R_{(-j)}\right\|_2, \\ & \max_j \frac{1}{\alpha w_{jE}}\left((X_E \circ \boldsymbol{\Psi}_j)(\beta_E \cdot \mathbf{1}_{m_j} + \boldsymbol{\theta}_j)\right)^\top R_{(-jE)} \Bigg\}\end{aligned} \tag{57}$$

Algorithm 4 Coordinate descent for least-squares `sail` with weak heredity

function `sail`$(\boldsymbol{X}, Y, X_E, \texttt{basis}, \lambda, \alpha, w_j, w_E, w_{jE}, \epsilon)$ ▷ Algorithm for solving (45)

  $\Psi_j \leftarrow \texttt{basis}(X_j)$, $\widetilde{\Psi}_j \leftarrow X_E \circ \Psi_j$ for $j = 1, \ldots, p$

  Initialize: $\beta_0^{(0)} \leftarrow \bar{Y}$, $\beta_E^{(0)} = \boldsymbol{\theta}_j^{(0)} = \gamma_j^{(0)} \leftarrow 0$ for $j = 1, \ldots, p$.

  Set iteration counter $k \leftarrow 0$

  $R^* \leftarrow Y - \beta_0^{(k)} - \beta_E^{(k)} X_E - \sum_j \boldsymbol{\Psi}_j \boldsymbol{\theta}_j^{(k)} - \sum_j \gamma_j^{(k)} \widetilde{\boldsymbol{\Psi}}_j (\beta_E^{(k)} \cdot \mathbf{1}_{m_j} + \boldsymbol{\theta}_j^{(k)})$

  repeat

    • To update $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)$

      $\widetilde{X}_j \leftarrow \widetilde{\boldsymbol{\Psi}}_j (\beta_E^{(k)} \cdot \mathbf{1}_{m_j} + \boldsymbol{\theta}_j^{(k)}) \qquad$ for $j = 1, \ldots, p$

      $R \leftarrow R^* + \sum_{j=1}^{p} \gamma_j^{(k)} \widetilde{X}_j$

$$\boldsymbol{\gamma}^{(k)(new)} \leftarrow \arg\min_{\boldsymbol{\gamma}} \frac{1}{2n} \left\| R - \sum_j \gamma_j \widetilde{X}_j \right\|_2^2 + \lambda\alpha \sum_j w_{jE} |\gamma_j|$$

      $\Delta = \sum_j (\gamma_j^{(k)} - \gamma_j^{(k)(new)}) \widetilde{X}_j$

      $R^* \leftarrow R^* + \Delta$

    • To update $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p)$

      $\widetilde{X}_j \leftarrow \boldsymbol{\Psi}_j + \gamma_j^{(k)} \widetilde{\boldsymbol{\Psi}}_j$ for $j = 1, \ldots, p$

      for $j = 1, \ldots, p$ **do**

        $R \leftarrow R^* + \widetilde{X}_j \boldsymbol{\theta}_j^{(k)}$

$$\boldsymbol{\theta}_j^{(k)(new)} \leftarrow \arg\min_{\boldsymbol{\theta}_j} \frac{1}{2n} \left\| R - \widetilde{X}_j \boldsymbol{\theta}_j \right\|_2^2 + \lambda(1-\alpha) w_j \|\theta_j\|_2$$

        $\Delta = \widetilde{X}_j (\boldsymbol{\theta}_j^{(k)} - \boldsymbol{\theta}_j^{(k)(new)})$

        $R^* \leftarrow R^* + \Delta$

    • To update β_E

      $\widetilde{X}_E \leftarrow X_E + \sum_j \gamma_j^{(k)} \widetilde{\boldsymbol{\Psi}}_j \mathbf{1}_{m_j}$

      $R \leftarrow R^* + \beta_E^{(k)} \widetilde{X}_E$

$$\beta_E^{(k)(new)} \leftarrow \frac{1}{\widetilde{X}_E^\top \widetilde{X}_E} S\left(\frac{1}{n \cdot w_E} \widetilde{X}_E^\top R, \lambda(1-\alpha) \right)$$

▷ $S(x,t) = \text{sign}(x)(|x| - t)_+$

      $\Delta = (\beta_E^{(k)} - \beta_E^{(k)(new)}) \widetilde{X}_E$

      $R^* \leftarrow R^* + \Delta$

    • To update β_0

      $R \leftarrow R^* + \beta_0^{(k)}$

$$\beta_0^{(k)(new)} \leftarrow \frac{1}{n} R^* \cdot \mathbf{1}$$

      $\Delta = \beta_0^{(k)} - \beta_0^{(k)(new)}$

      $R^* \leftarrow R^* + \Delta$

    $k \leftarrow k + 1$

  until convergence criterion is satisfied: $\left| Q(\boldsymbol{\Phi}^{(k-1)}) - Q(\boldsymbol{\Phi}^{(k)}) \right| / Q(\boldsymbol{\Phi}^{(k-1)}) < \epsilon$

which reduces to

$$\lambda_{max} = \frac{1}{n(1-\alpha)} \max \left\{ \frac{1}{w_E} \left(X_E\right)^\top R_{(-E)}, \max_j \frac{1}{w_j} \left\| \left(\boldsymbol{\Psi}_j\right)^\top R_{(-j)} \right\|_2 \right\}$$

This is the same λ_{max} as the least-squares strong heredity `sail` model.

C Additional Results on PRS for Educational Attainment

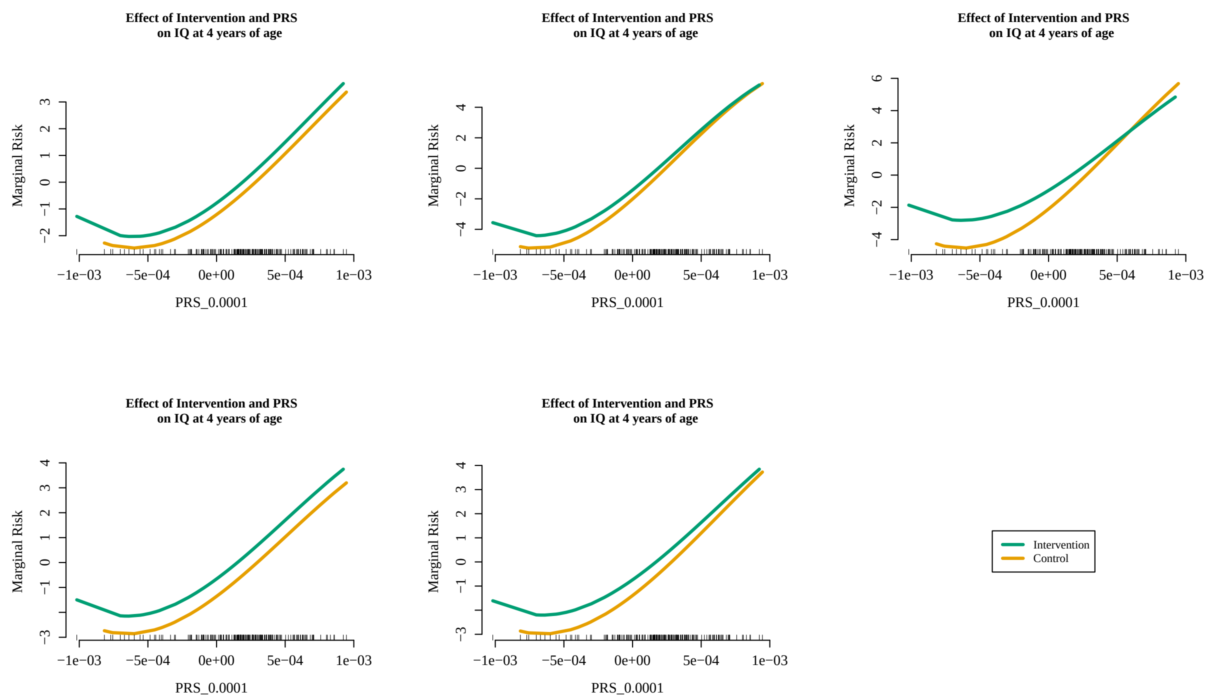

Figure C.1: Estimated interaction effect identified by the weak heredity `sail` using cubic B-splines and $\alpha = 0.1$ for the Nurse Family Partnership data for the 5 imputed datasets. Of the 189 subjects, 19 IQ scores were imputed using `mice` [5]. The selected model, chosen via 10-fold cross-validation, contained three variables: the main effects for the intervention and the PRS for educational attainment using genetic variants significant at the 0.0001 level, as well as their interaction.

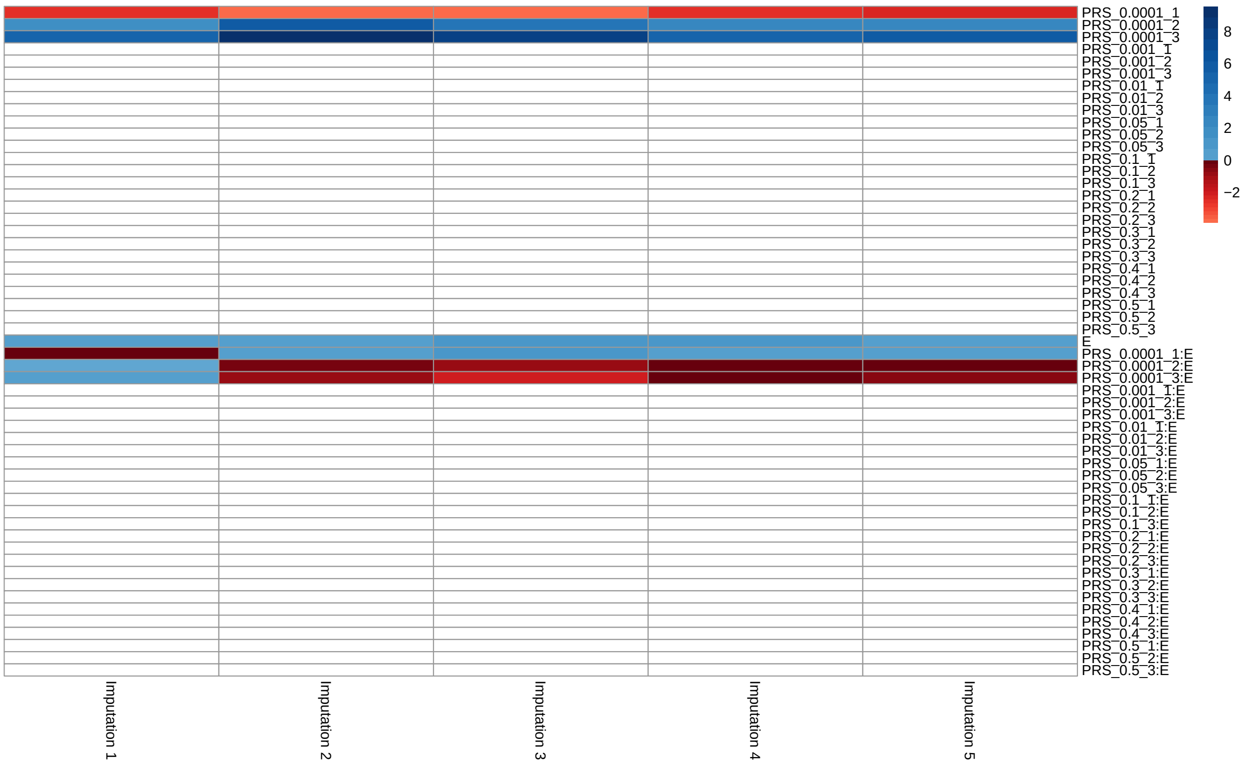

Figure C.2: Coefficient estimates obtained by the weak heredity `sail` using cubic B-splines and $\alpha = 0.1$ for the Nurse Family Partnership data for the 5 imputed datasets. Of the 189 subjects, 19 IQ scores were imputed using `mice` [5]. The selected model, chosen via 10-fold cross-validation, contained three variables: the main effects for the intervention and the PRS for educational attainment using genetic variants significant at the 0.0001 level, as well as their interaction. This results was consistent across all 5 imputed datasets. The white boxes indicate a coefficient estimate of 0.