

A Sparse Additive Model for High-Dimensional Interactions with an Exposure Variable

Supplementary Materials

Sahir R Bhatnagar^{1,2}, Tianyuan Lu^{3,4}, Amanda Lovato⁵, David L Olds⁶, Michael S Kobar⁷,
Michael J Meaney⁸, Kieran O'Donnell⁹, Yi Yang¹⁰, and Celia MT Greenwood^{1,3,5}

¹Department of Epidemiology, Biostatistics and Occupational Health, McGill University, ²Department
of Diagnostic Radiology, McGill University, ³Quantitative Life Sciences, McGill University, ⁴Lady
Davis Institute, Jewish General Hospital, Montréal, QC, ⁵Statistics Canada, Ottawa, ON, ⁶Department
of Pediatrics, University of Colorado School of Medicine, Denver, ⁷Department of Medical Genetics,
University of British Columbia, BC, ⁸Singapore Institute for Clinical Sciences, Singapore; McGill
University, ⁹Department of Psychiatry, McGill University, ¹⁰Department of Mathematics and Statis-
tics, McGill University, ¹¹Departments of Oncology and Human Genetics, McGill University

A Proofs

A.1 Regularity Conditions

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

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

and

$$\begin{aligned} \mathbf{I}_{j_1 k_1 j_2 k_2}(\Phi) &= E_{\Phi} \left[\frac{\partial}{\partial \phi_{j_1 k_1}} \log f(V, \Phi) \cdot \frac{\partial}{\partial \phi_{j_2 k_2}} \log f(V, \Phi) \right] \\ &= E_{\Phi} \left[-\frac{\partial^2}{\partial \phi_{j_1 k_1} \partial \phi_{j_2 k_2}} \log f(V, \Phi) \right], \end{aligned}$$

for any $j_1, j_2 = 1, \dots, 2p + 1$, and $k_1 = 1, \dots, p_{j_1}$, $k_2 = 1, \dots, p_{j_2}$, 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

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

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

(C3) There exists an open set ω of Ω that contains the true parameter point Φ^* such that

for almost all \mathbf{V} the density $f(\mathbf{V}, \Phi)$ admits all third derivatives $\frac{\partial^3 f(\mathbf{V}, \Phi)}{\partial \phi_{j_1 k_1} \partial \phi_{j_2 k_2} \partial \phi_{j_3 k_3}}$ for

all Φ in ω and any $j_1, j_2, j_3 = 1, \dots, 2p + 1$, and $k_1 = 1, \dots, p_{j_1}$, $k_2 = 1, \dots, p_{j_2}$ and

$k_3 = 1, \dots, p_{j_3}$. Furthermore, there exist functions $M_{j_1 k_1 j_2 k_2 j_3 k_3}$ such that

$$\left| \frac{\partial^3}{\partial \phi_{j_1 k_1} \partial \phi_{j_2 k_2} \partial \phi_{j_3 k_3}} \log f(\mathbf{V}, \Phi) \right| \leq M_{j_1 k_1 j_2 k_2 j_3 k_3}(\mathbf{V}) \quad \text{for all } \Phi \in \omega,$$

and $m_{j_1 k_1 j_2 k_2 j_3 k_3} = E_{\Phi^*} [M_{j_1 k_1 j_2 k_2 j_3 k_3}(\mathbf{V})] < \infty$.

A.2 Lemma (1) proof

Let $\eta_n = \frac{1}{\sqrt{n}} + a_n$ and $\{\Phi^* + \eta_n \delta : \|\delta\|_2 \leq C\}$ be the ball around Φ^* for $\delta \in \mathbb{R}^d$, 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 $\hat{\Phi}_n$ of $Q_n(\Phi)$ such that $\|\hat{\Phi}_n - \Phi^*\|_2 = O_p(\frac{1}{\sqrt{n}})$. For this proof, we adopt the approaches outlined in [8, 13, 23, 36] and extend it to our situation. Let $\eta_n = \frac{1}{\sqrt{n}} + a_n$ and $\{\Phi^* + \eta_n \delta : \|\delta\|_2 \leq C\}$ be the ball around Φ^* for $\delta = (\mathbf{u}_1^\top, \mathbf{u}_2^\top, \dots, \mathbf{u}_{p+1}^\top, \mathbf{u}_{p+2}^\top, \dots, \mathbf{u}_{2p+1}^\top)^\top \in \mathbb{R}^d$, where d is the dimension of the design matrix and C is some constant. The objective function is given by

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

Define

$$D_n(\delta) \equiv Q_n(\Phi^* + \eta_n \delta) - Q_n(\Phi^*).$$

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

$$\begin{aligned}
D_n(\boldsymbol{\delta}) &= -L_n(\boldsymbol{\Phi}^* + \eta_n \boldsymbol{\delta}) + L_n(\boldsymbol{\Phi}^*) + n \sum_{m=1}^{2p+1} \lambda_m (\|\boldsymbol{\theta}_m^* + \eta_n \mathbf{u}_m\|_2 - \|\boldsymbol{\theta}_m^*\|_2) \\
&\stackrel{(a)}{\geq} -L_n(\boldsymbol{\Phi}^* + \eta_n \boldsymbol{\delta}) + L_n(\boldsymbol{\Phi}^*) + n \sum_{m \in \mathcal{A}_1} \lambda_m^\theta (\|\boldsymbol{\theta}_m^* + \eta_n \mathbf{u}_m\|_2 - \|\boldsymbol{\theta}_m^*\|_2) \\
&\quad + n \sum_{m \in \mathcal{A}_2} \lambda_m^\theta (\|\boldsymbol{\theta}_m^* + \eta_n \mathbf{u}_m\|_2 - \|\boldsymbol{\theta}_m^*\|_2) \\
&\stackrel{(b)}{\geq} -L_n(\boldsymbol{\Phi}^* + \eta_n \boldsymbol{\delta}) + L_n(\boldsymbol{\Phi}^*) - n\eta_n \sum_{m \in \mathcal{A}_1} \lambda_m \|\mathbf{u}_m\|_2 - n\eta_n \sum_{m \in \mathcal{A}_2} \lambda_m \|\mathbf{u}_m\|_2 \\
&\stackrel{(c)}{\geq} -L_n(\boldsymbol{\Phi}^* + \eta_n \boldsymbol{\delta}) + L_n(\boldsymbol{\Phi}^*) - n\eta_n^2 \sum_{m \in \mathcal{A}_1} \|\mathbf{u}_m\|_2 - n\eta_n^2 \sum_{m \in \mathcal{A}_2} \|\mathbf{u}_m\|_2 \\
&\geq -L_n(\boldsymbol{\Phi}^* + \eta_n \boldsymbol{\delta}) + L_n(\boldsymbol{\Phi}^*) - n\eta_n^2 (|\mathcal{A}_1| + |\mathcal{A}_2|)C \\
&\stackrel{(d)}{=} -[\nabla L_n(\boldsymbol{\Phi}^*)]^\top (\eta_n \boldsymbol{\delta}) - \frac{1}{2} (\eta_n \boldsymbol{\delta})^\top [\nabla^2 L_n(\boldsymbol{\Phi}^*)] (\eta_n \boldsymbol{\delta}) (1 + o(1)) \\
&\quad - n\eta_n^2 (|\mathcal{A}_1| + |\mathcal{A}_2|)C
\end{aligned} \tag{13}$$

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:

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

628 We split (13) into three parts:

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

Then

$$\begin{aligned} D_1 &= -\eta_n [\nabla L_n(\Phi^*)]^\top \delta \\ &= -\sqrt{n}\eta_n \left(\frac{1}{\sqrt{n}} \nabla L_n(\Phi^*) \right)^\top \delta \\ &= -\sqrt{n}\eta_n \left(\sqrt{n} \frac{1}{n} \sum_{i=1}^n \nabla \log f(\mathbf{V}_i, \Phi)|_{\Phi=\Phi^*} \right)^\top \delta \\ &= -\sqrt{n}\eta_n \left(\sqrt{n} \left[\frac{1}{n} \sum_{i=1}^n \nabla \log f(\mathbf{V}_i, \Phi)|_{\Phi=\Phi^*} - \mathbf{0} \right] \right)^\top \delta \\ &= -\sqrt{n}\eta_n \left(\sqrt{n} \left[\frac{1}{n} \sum_{i=1}^n \nabla \log f(\mathbf{V}_i, \Phi)|_{\Phi=\Phi^*} - E_{\Phi^*} \nabla L(\Phi^*) \right] \right)^\top \delta \\ &= -\sqrt{n}\eta_n O_P(1) \delta \\ &= -O_P(n\eta_n^2) \delta \end{aligned} \tag{14}$$

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

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

$$\begin{aligned}
D_2 &= \frac{1}{2} n \eta_n^2 \left\{ \boldsymbol{\delta}^\top \left[-\frac{1}{n} \nabla^2 L_n(\boldsymbol{\Phi}^*) \right] \boldsymbol{\delta} \right\} (1 + o_p(1)) \\
&= \frac{1}{2} n \eta_n^2 \left\{ \boldsymbol{\delta}^\top [\mathbf{I}(\boldsymbol{\Phi}^*)] \boldsymbol{\delta} \right\} (1 + o_p(1)) \text{ by the weak law of large numbers.} \\
&= O_p(n \eta_n^2 \|\boldsymbol{\delta}\|_2^2)
\end{aligned} \tag{15}$$

Combining (14) and (15) with (13) gives:

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

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

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

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 $\hat{\boldsymbol{\Phi}}_n$ such that $\|\hat{\boldsymbol{\Phi}}_n - \boldsymbol{\Phi}^*\| \leq \eta_n \|\boldsymbol{\delta}\|_2 \leq \eta_n C = O_P(\eta_n) = O_P(\frac{1}{\sqrt{n}} + a_n) = O_P(\frac{1}{\sqrt{n}})$, since $a_n = o(\frac{1}{\sqrt{n}})$. Hence, $\left\| \hat{\boldsymbol{\Phi}}_n - \boldsymbol{\Phi}^* \right\|_2 = O_P\left(\frac{1}{\sqrt{n}}\right)$.

□

A.3 Theorem 1 proof

We first consider consistency for the main effects $P\left(\widehat{\Phi}_{\mathcal{A}_1^c} = \mathbf{0}\right) \rightarrow 1$. Following [8, 13], it is sufficient to show that for all $m \in \mathcal{A}_1^c$, $P\left(\widehat{\phi}_m = \mathbf{0}\right) \rightarrow 1$, which implies that $P\left(\widehat{\Phi}_{\mathcal{A}_1^c} = \mathbf{0}\right) \rightarrow 1$, i.e., the \sqrt{n} -consistent estimate $\widehat{\Phi}$ has oracle property $\widehat{\phi}_m = \mathbf{0}$ if $\phi_m^* = \mathbf{0}$. Denote

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

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

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

where $\tilde{\Phi}$ lies between $\widehat{\Phi}_n$ and Φ^* . By the regularity conditions and Lemma (1) that

$\|\widehat{\Phi}_n - \Phi^*\|_2 = O_P\left(\frac{1}{\sqrt{n}}\right)$, the first term is of the order $O_p(\sqrt{n})$

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

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

$$\frac{\partial Q_n(\hat{\Phi}_n)}{\partial \Phi_m} = \sqrt{n} \left\{ O_p(1) + \sqrt{n} \lambda_m \frac{\hat{\phi}_{mk}}{\|\hat{\Phi}_m\|_2} \right\}. \quad (16)$$

As $\sqrt{n} \lambda_m \geq \sqrt{n} b_n \rightarrow \infty$ 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 (16) with probability tending to one. This means that (16) 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 $\hat{\Phi}_m = (\hat{\phi}_{m1}, \dots, \hat{\phi}_{mp_m})$ must be in a position where $\hat{\Phi}_m$ is not differentiable. Hence $\hat{\Phi}_m = \mathbf{0}$ for all $m \in \mathcal{A}_1^c$. Hence $P(\hat{\Phi}_{\mathcal{A}_1^c} = \mathbf{0}) \rightarrow 1$. This completes the proof.

Next, we prove that for the interactions $P(\hat{\Phi}_{\mathcal{A}_2^c} = \mathbf{0}) \rightarrow 1$. For $m \in \mathcal{A}_2^c$ s.t. $\phi_m^* = \gamma_{jE}^* = 0$ but $\beta_E \neq 0$ and $\theta_j^* \neq \mathbf{0}$ ($1 \leq j \leq p$), we can prove $P(\hat{\Phi}_{\mathcal{A}_2^c} = \mathbf{0}) \rightarrow 1$ by a similar reasoning, which further implies that $P(\hat{\gamma}_{jE} = 0) \rightarrow 0$. For $m \in \mathcal{A}_2^c$ such that $\phi_m^* = \gamma_{jE}^* = 0$ and either $\beta_E = 0$ or $\theta_j^* = \mathbf{0}$ ($1 \leq j \leq p$): without loss of generality, assume that $\theta_j^* = \mathbf{0}$. Notice that $\hat{\theta}_j = \mathbf{0}$ 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 $P(\hat{\Phi}_{\mathcal{A}_1^c} = \mathbf{0}) \rightarrow 1$, therefore we get $P(\hat{\Phi}_{\mathcal{A}_2^c} = \mathbf{0}) \rightarrow 1$ as well for this case.

□

A.4 Theorem 2 proof

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

$$\left. \frac{\partial Q_n(\Phi_{\mathcal{A}})}{\partial \Phi_m} \right|_{\Phi = \begin{pmatrix} \hat{\Phi}_{\mathcal{A}} \\ 0 \end{pmatrix}} = 0, \quad \forall m \in \mathcal{A}, \quad (17)$$

that is

$$\left. \frac{\partial Q_n(\Phi_{\mathcal{A}})}{\partial \Phi_m} \right|_{\Phi_{\mathcal{A}} = \hat{\Phi}_{\mathcal{A}}} = 0, \quad \forall m \in \mathcal{A}, \quad (18)$$

where

$$\begin{aligned} Q_n(\Phi_{\mathcal{A}}) &= -L_n(\Phi_{\mathcal{A}}) + n \underbrace{\sum_{m \in \mathcal{A}_1} \lambda_m \|\phi_m\|_2 + \sum_{m \in \mathcal{A}_2} \lambda_m \|\phi_m\|_2}_{\triangleq nP(\Phi_{\mathcal{A}})} \\ &= -L_n(\Phi_{\mathcal{A}}) + nP(\Phi_{\mathcal{A}}). \end{aligned} \quad (19)$$

From (18) and (19) we have

$$\nabla_{\mathcal{A}} Q_n(\hat{\Phi}_{\mathcal{A}}) = -\nabla_{\mathcal{A}} L_n(\hat{\Phi}_{\mathcal{A}}) + n \nabla_{\mathcal{A}} P(\hat{\Phi}_{\mathcal{A}}) = \mathbf{0}, \quad (20)$$

with probability tending to 1.

667 Denote $\Sigma = \text{diag}\{o_p(1), \dots, o_p(1)\}$. We then expand $-\nabla_{\mathcal{A}} L_n(\Phi_{\mathcal{A}})$ at $\Phi_{\mathcal{A}} = \Phi_{\mathcal{A}}^*$ in (20):

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

668 The third line follows by

$$\frac{1}{n} \nabla_{\mathcal{A}}^2 L_n(\Phi_{\mathcal{A}}^*) = E \{ \nabla_{\mathcal{A}}^2 L(\Phi_{\mathcal{A}}^*) \} + \Sigma = -\mathbf{I}(\Phi_{\mathcal{A}}^*) + \Sigma.$$

669 Denote

$$\mathbf{b} = (\lambda_m \text{sgn}(\beta_m^*), \lambda_m \frac{\theta_m^*}{\|\theta_m^*\|_2}, \lambda_m \text{sgn}(\gamma_{mE}^*))^\top, \quad m \in \mathcal{A},$$

We also expand $n \nabla_{\mathcal{A}} P(\Phi_{\mathcal{A}})$ at $\Phi_{\mathcal{A}} = \Phi_{\mathcal{A}}^*$ in (20):

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

And due to the fact that $\sqrt{n} \lambda_m \leq \sqrt{n} a_n \rightarrow 0$ for $m \in \mathcal{A}$ and $\frac{\theta_{mk}^*}{\|\theta_m^*\|_2} \leq 1$ for any $1 \leq k \leq p_m$,

we know that $\sqrt{n} \mathbf{b} = (o_p(1), \dots, o_p(1))^\top$. Thus,

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

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

Therefore, by the central limit theorem, we know that

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

Hence,

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

□

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

$$\hat{Y} = \beta_0 \cdot \mathbf{1} + \sum_{j=1}^p \Psi_j \theta_j + \beta_E X_E + \sum_{j=1}^p \gamma_j \beta_E (X_E \circ \Psi_j) \theta_j \quad (21)$$

and the objective function is given by

$$Q(\Phi) = \frac{1}{2n} \|Y - \hat{Y}\|_2^2 + \lambda(1 - \alpha) \left(w_E |\beta_E| + \sum_{j=1}^p w_j \|\theta_j\|_2 \right) + \lambda \alpha \sum_{j=1}^p w_{jE} |\gamma_j| \quad (22)$$

Solving (22) 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

$$\frac{\partial Q}{\partial \beta_0} = \frac{1}{n} \left(Y - \beta_0 \cdot \mathbf{1} - \sum_{j=1}^p \Psi_j \theta_j - \beta_E X_E - \sum_{j=1}^p \gamma_j \beta_E (X_E \circ \Psi_j) \theta_j \right)^\top \mathbf{1} = 0 \quad (23)$$

$$\frac{\partial Q}{\partial \beta_E} = -\frac{1}{n} \left(X_E + \sum_{j=1}^p \gamma_j (X_E \circ \Psi_j) \theta_j \right)^\top R + \lambda(1 - \alpha) w_E s_1 = 0 \quad (24)$$

$$\frac{\partial Q}{\partial \theta_j} = -\frac{1}{n} (\Psi_j + \gamma_j \beta_E (X_E \circ \Psi_j))^\top R + \lambda(1 - \alpha) w_j s_2 = \mathbf{0} \quad (25)$$

$$\frac{\partial Q}{\partial \gamma_j} = -\frac{1}{n} (\beta_E (X_E \circ \Psi_j) \theta_j)^\top R + \lambda \alpha w_j s_3 = 0 \quad (26)$$

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:

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

and s_3 is in the subgradient of the ℓ_1 norm:

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

677 Define the partial residuals, without the j th predictor for $j = 1, \dots, p$, as

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

678 the partial residual without X_E as

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

679 and the partial residual without the j th interaction for $j = 1, \dots, p$, as

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

From the subgradient equations (23)–(26) we see that

$$\hat{\beta}_0 = \left(Y - \sum_{j=1}^p \Psi_j \hat{\boldsymbol{\theta}}_j - \hat{\beta}_E X_E - \sum_{j=1}^p \hat{\gamma}_j \hat{\beta}_E (X_E \circ \Psi_j) \hat{\boldsymbol{\theta}}_j \right)^\top \mathbf{1} \quad (27)$$

$$\hat{\beta}_E = \frac{S \left(\frac{1}{n \cdot w_E} \left(X_E + \sum_{j=1}^p \hat{\gamma}_j (X_E \circ \Psi_j) \hat{\boldsymbol{\theta}}_j \right)^\top R_{(-E)}, \lambda(1 - \alpha) \right)}{\left(X_E + \sum_{j=1}^p \hat{\gamma}_j (X_E \circ \Psi_j) \hat{\boldsymbol{\theta}}_j \right)^\top \left(X_E + \sum_{j=1}^p \hat{\gamma}_j (X_E \circ \Psi_j) \hat{\boldsymbol{\theta}}_j \right)} \quad (28)$$

$$\lambda(1 - \alpha) w_j \frac{\boldsymbol{\theta}_j}{\|\boldsymbol{\theta}_j\|_2} = \frac{1}{n} (\Psi_j + \gamma_j \beta_E (X_E \circ \Psi_j))^\top R_{(-j)} \quad (29)$$

$$\hat{\gamma}_j = \frac{S \left(\frac{1}{n \cdot w_{jE}} (\beta_E (X_E \circ \Psi_j) \boldsymbol{\theta}_j)^\top R_{(-jE)}, \lambda \alpha \right)}{(\beta_E (X_E \circ \Psi_j) \boldsymbol{\theta}_j)^\top (\beta_E (X_E \circ \Psi_j) \boldsymbol{\theta}_j)} \quad (30)$$

680 where $S(x, t) = \text{sign}(x)(|x| - t)$ is the soft-thresholding operator. We see from (27) and (28)

681 that there are closed form solutions for the intercept and β_E . From (30), each γ_j also has a

682 closed form solution and can be solved efficiently for $j = 1, \dots, p$ using a coordinate descent

procedure [14]. Since there is no closed form solution for β_j , we use a quadratic majorization technique [37] to solve (29). Furthermore, we update each θ_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., $\hat{\tau}_j = \hat{\gamma}_j \hat{\beta}_E \hat{\theta}_j$, $j = 1, \dots, p$ 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

```

1: function sail( $\mathbf{X}, Y, X_E, \text{basis}, \lambda, \alpha, w_j, w_E, w_{jE}, \epsilon$ ) ▷ Algorithm for solving (22)
2:    $\Psi_j \leftarrow \text{basis}(X_j), \tilde{\Psi}_j \leftarrow X_E \circ \Psi_j$  for  $j = 1, \dots, p$ 
3:   Initialize:  $\beta_0^{(0)} \leftarrow \bar{Y}, \beta_E^{(0)} = \boldsymbol{\theta}_j^{(0)} = \gamma_j^{(0)} \leftarrow 0$  for  $j = 1, \dots, p$ .
4:   Set iteration counter  $k \leftarrow 0$ 
5:    $R^* \leftarrow Y - \beta_0^{(k)} - \beta_E^{(k)} X_E - \sum_j (\Psi_j + \gamma_j^{(k)} \beta_E^{(k)} \tilde{\Psi}_j) \boldsymbol{\theta}_j^{(k)}$ 
6:   repeat
7:     • To update  $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_p)$ 
8:        $\tilde{X}_j \leftarrow \beta_E^{(k)} \tilde{\Psi}_j \boldsymbol{\theta}_j^{(k)}$  for  $j = 1, \dots, p$ 
9:        $R \leftarrow R^* + \sum_{j=1}^p \gamma_j^{(k)} \tilde{X}_j$ 
10:
11:       
$$\boldsymbol{\gamma}^{(k)(new)} \leftarrow \arg \min_{\boldsymbol{\gamma}} \frac{1}{2n} \left\| R - \sum_j \gamma_j \tilde{X}_j \right\|_2^2 + \lambda \alpha \sum_j w_{jE} |\gamma_j|$$

12:
13:        $\Delta = \sum_j (\gamma_j^{(k)} - \gamma_j^{(k)(new)}) \tilde{X}_j$ 
14:        $R^* \leftarrow R^* + \Delta$ 
15:     • To update  $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_p)$ 
16:        $\tilde{X}_j \leftarrow \Psi_j + \gamma_j^{(k)} \beta_E^{(k)} \tilde{\Psi}_j$  for  $j = 1, \dots, p$ 
17:       for  $j = 1, \dots, p$  do
18:          $R \leftarrow R^* + \tilde{X}_j \boldsymbol{\theta}_j^{(k)}$ 
19:
20:         
$$\boldsymbol{\theta}_j^{(k)(new)} \leftarrow \arg \min_{\boldsymbol{\theta}_j} \frac{1}{2n} \left\| R - \tilde{X}_j \boldsymbol{\theta}_j \right\|_2^2 + \lambda (1 - \alpha) w_j \|\boldsymbol{\theta}_j\|_2$$

21:
22:          $\Delta = \tilde{X}_j (\boldsymbol{\theta}_j^{(k)} - \boldsymbol{\theta}_j^{(k)(new)})$ 
23:          $R^* \leftarrow R^* + \Delta$ 
24:     • To update  $\beta_E$ 
25:        $\tilde{X}_E \leftarrow X_E + \sum_j \gamma_j^{(k)} \tilde{\Psi}_j \boldsymbol{\theta}_j^{(k)}$ 
26:        $R \leftarrow R^* + \beta_E^{(k)} \tilde{X}_E$ 
27:
28:       
$$\beta_E^{(k)(new)} \leftarrow \frac{1}{\tilde{X}_E^\top \tilde{X}_E} S \left( \frac{1}{n \cdot w_E} \tilde{X}_E^\top R, \lambda (1 - \alpha) \right)$$

29:
30:       
$$\triangleright S(x, t) = \text{sign}(x)(|x| - t)_+$$

31:
32:        $\Delta = (\beta_E^{(k)} - \beta_E^{(k)(new)}) \tilde{X}_E$ 
33:        $R^* \leftarrow R^* + \Delta$ 
34:     • To update  $\beta_0$ 
35:        $R \leftarrow R^* + \beta_0^{(k)}$ 
36:
37:       
$$\beta_0^{(k)(new)} \leftarrow \frac{1}{n} R \cdot \mathbf{1}$$

38:
39:        $\Delta = \beta_0^{(k)} - \beta_0^{(k)(new)}$ 
40:        $R^* \leftarrow R^* + \Delta$ 
41:
42:        $k \leftarrow k + 1$ 
43:   until convergence criterion is satisfied:  $|Q(\boldsymbol{\Phi}^{(k-1)}) - Q(\boldsymbol{\Phi}^{(k)})| / Q(\boldsymbol{\Phi}^{(k-1)}) < \epsilon$ 

```

692 B.2 Details on Update for θ

Here we discuss a computational speedup in the updates for the θ parameter. The partial residual (R_s) used for updating θ_s ($s \in 1, \dots, p$) at the k th iteration is given by

$$R_s = Y - \tilde{Y}_{(-s)}^{(k)} \quad (31)$$

where $\tilde{Y}_{(-s)}^{(k)}$ is the fitted value at the k th iteration excluding the contribution from Ψ_s :

$$\tilde{Y}_{(-s)}^{(k)} = \beta_0^{(k)} - \beta_E^{(k)} X_E - \sum_{\ell \neq s} \Psi_\ell \theta_\ell^{(k)} - \sum_{\ell \neq s} \gamma_\ell^{(k)} \beta_E^{(k)} \tilde{\Psi}_\ell \theta_\ell^{(k)} \quad (32)$$

Using (32), (31) can be re-written as

$$\begin{aligned} R_s &= Y - \beta_0^{(k)} - \beta_E^{(k)} X_E - \sum_{j=1}^p (\Psi_j + \gamma_j^{(k)} \beta_E^{(k)} \tilde{\Psi}_j) \theta_j^{(k)} + (\Psi_s + \gamma_s^{(k)} \beta_E^{(k)} \tilde{\Psi}_s) \theta_s^{(k)} \\ &= R^* + (\Psi_s + \gamma_s^{(k)} \beta_E^{(k)} \tilde{\Psi}_s) \theta_s^{(k)} \end{aligned} \quad (33)$$

693 where

$$R^* = Y - \beta_0^{(k)} - \beta_E^{(k)} X_E - \sum_{j=1}^p (\Psi_j + \gamma_j^{(k)} \beta_E^{(k)} \tilde{\Psi}_j) \theta_j^{(k)} \quad (34)$$

Denote $\theta_s^{(k)(\text{new})}$ the solution for predictor s at the k th iteration, given by:

$$\theta_s^{(k)(\text{new})} = \arg \min_{\theta_j} \frac{1}{2n} \left\| R_s - (\Psi_s + \gamma_s^{(k)} \beta_E^{(k)} \tilde{\Psi}_s) \theta_j \right\|_2^2 + \lambda(1 - \alpha) w_s \|\theta_j\|_2 \quad (35)$$

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

$$R_{s+1} = R^* + (\boldsymbol{\Psi}_{s+1} + \gamma_{s+1}^{(k)} \beta_E^{(k)} \tilde{\boldsymbol{\Psi}}_{s+1}) \boldsymbol{\theta}_{s+1}^{(k)} + (\boldsymbol{\Psi}_s + \gamma_s^{(k)} \beta_E^{(k)} \tilde{\boldsymbol{\Psi}}_s) (\boldsymbol{\theta}_s^{(k)} - \boldsymbol{\theta}_s^{(k)(new)}) \quad (36)$$

where R^* is given by (34), $\boldsymbol{\theta}_s^{(k)}$ is the parameter value prior to the update, and $\boldsymbol{\theta}_s^{(k)(new)}$ is the updated value given by (35). Taking the difference between (33) and (36) gives

$$\begin{aligned} \Delta &= R_t - R_s \\ &= (\boldsymbol{\Psi}_t + \gamma_t^{(k)} \beta_E^{(k)} \tilde{\boldsymbol{\Psi}}_t) \boldsymbol{\theta}_t^{(k)} + (\boldsymbol{\Psi}_s + \gamma_s^{(k)} \beta_E^{(k)} \tilde{\boldsymbol{\Psi}}_s) (\boldsymbol{\theta}_s^{(k)} - \boldsymbol{\theta}_s^{(k)(new)}) - (\boldsymbol{\Psi}_s + \gamma_s^{(k)} \beta_E^{(k)} \tilde{\boldsymbol{\Psi}}_s) \boldsymbol{\theta}_s^{(k)} \\ &= (\boldsymbol{\Psi}_t + \gamma_t^{(k)} \beta_E^{(k)} \tilde{\boldsymbol{\Psi}}_t) \boldsymbol{\theta}_t^{(k)} - (\boldsymbol{\Psi}_s + \gamma_s^{(k)} \beta_E^{(k)} \tilde{\boldsymbol{\Psi}}_s) \boldsymbol{\theta}_s^{(k)(new)} \end{aligned} \quad (37)$$

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 (37). 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 (24)–(26) can be used to determine the largest value of λ such that all coefficients are 0. From the subgradient Equation (24), we see that $\beta_E = 0$ is a solution if

$$\frac{1}{w_E} \left| \frac{1}{n} \left(X_E + \sum_{j=1}^p \gamma_j (X_E \circ \boldsymbol{\Psi}_j) \boldsymbol{\theta}_j \right)^\top R_{(-E)} \right| \leq \lambda(1 - \alpha) \quad (38)$$

702 From the subgradient Equation (25), we see that $\boldsymbol{\theta}_j = \mathbf{0}$ is a solution if

$$\frac{1}{w_j} \left\| \frac{1}{n} (\boldsymbol{\Psi}_j + \gamma_j \beta_E (X_E \circ \boldsymbol{\Psi}_j))^\top R_{(-j)} \right\|_2 \leq \lambda(1 - \alpha) \quad (39)$$

703 From the subgradient Equation (26), we see that $\gamma_j = 0$ is a solution if

$$\frac{1}{w_{jE}} \left| \frac{1}{n} (\beta_E (X_E \circ \boldsymbol{\Psi}_j) \boldsymbol{\theta}_j)^\top R_{(-jE)} \right| \leq \lambda \alpha \quad (40)$$

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

$$\lambda_{max} = \frac{1}{n(1 - \alpha)} \max \left\{ \frac{1}{w_E} \left(X_E + \sum_{j=1}^p \gamma_j (X_E \circ \boldsymbol{\Psi}_j) \boldsymbol{\theta}_j \right)^\top R_{(-E)}, \max_j \frac{1}{w_j} \left\| (\boldsymbol{\Psi}_j + \gamma_j \beta_E (X_E \circ \boldsymbol{\Psi}_j))^\top R_{(-j)} \right\|_2 \right\} \quad (41)$$

which reduces to

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

704 B.4 Least-Squares sail with Weak Heredity

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

$$\hat{Y} = \beta_0 \cdot \mathbf{1} + \sum_{j=1}^p \Psi_j \boldsymbol{\theta}_j + \beta_E X_E + \sum_{j=1}^p \gamma_j (X_E \circ \Psi_j) (\beta_E \cdot \mathbf{1}_{m_j} + \boldsymbol{\theta}_j) \quad (42)$$

706 The objective function is given by

$$Q(\Phi) = \frac{1}{2n} \|Y - \hat{Y}\|_2^2 + \lambda(1 - \alpha) \left(w_E |\beta_E| + \sum_{j=1}^p w_j \|\boldsymbol{\theta}_j\|_2 \right) + \lambda\alpha \sum_{j=1}^p w_{jE} |\gamma_j| \quad (43)$$

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

$$\frac{\partial Q}{\partial \beta_0} = \frac{1}{n} \left(Y - \beta_0 \cdot \mathbf{1} - \sum_{j=1}^p \Psi_j \boldsymbol{\theta}_j - \beta_E X_E - \sum_{j=1}^p \gamma_j (X_E \circ \Psi_j) (\beta_E \cdot \mathbf{1}_{m_j} + \boldsymbol{\theta}_j) \right)^\top \mathbf{1} = 0 \quad (44)$$

$$\frac{\partial Q}{\partial \beta_E} = -\frac{1}{n} \left(X_E + \sum_{j=1}^p \gamma_j (X_E \circ \Psi_j) \mathbf{1}_{m_j} \right)^\top R + \lambda(1 - \alpha) w_E s_1 = 0 \quad (45)$$

$$\frac{\partial Q}{\partial \boldsymbol{\theta}_j} = -\frac{1}{n} (\Psi_j + \gamma_j (X_E \circ \Psi_j))^\top R + \lambda(1 - \alpha) w_j s_2 = \mathbf{0} \quad (46)$$

$$\frac{\partial Q}{\partial \gamma_j} = -\frac{1}{n} ((X_E \circ \Psi_j) (\beta_E \cdot \mathbf{1}_{m_j} + \boldsymbol{\theta}_j))^\top R + \lambda\alpha w_{jE} s_3 = 0 \quad (47)$$

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:

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

and s_3 is in the subgradient of the ℓ_1 norm:

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

707 Define the partial residuals, without the j th predictor for $j = 1, \dots, p$, as

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

708 the partial residual without X_E as

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

709 and the partial residual without the j th interaction for $j = 1, \dots, p$

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

710 From the subgradient Equation (45), we see that $\beta_E = 0$ is a solution if

$$\frac{1}{w_E} \left| \frac{1}{n} \left(X_E + \sum_{j=1}^p \gamma_j (X_E \circ \Psi_j) \mathbf{1}_{m_j} \right)^\top R_{(-E)} \right| \leq \lambda(1 - \alpha) \quad (48)$$

711 From the subgradient Equation (46), we see that $\theta_j = \mathbf{0}$ is a solution if

$$\frac{1}{w_j} \left\| \frac{1}{n} (\Psi_j + \gamma_j (X_E \circ \Psi_j))^\top R_{(-j)} \right\|_2 \leq \lambda(1 - \alpha) \quad (49)$$

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

$$\frac{1}{w_{jE}} \left| \frac{1}{n} ((X_E \circ \Psi_j)(\beta_E \cdot \mathbf{1}_{m_j} + \theta_j))^\top R_{(-jE)} \right| \leq \lambda\alpha \quad (50)$$

From the subgradient equations we see that

$$\hat{\beta}_0 = \left(Y - \sum_{j=1}^p \Psi_j \hat{\theta}_j - \hat{\beta}_E X_E - \sum_{j=1}^p \hat{\gamma}_j (X_E \circ \Psi_j)(\hat{\beta}_E \cdot \mathbf{1}_{m_j} + \hat{\theta}_j) \right)^\top \mathbf{1} \quad (51)$$

$$\hat{\beta}_E = \frac{S \left(\frac{1}{n \cdot w_E} \left(X_E + \sum_{j=1}^p \hat{\gamma}_j (X_E \circ \Psi_j) \mathbf{1}_{m_j} \right)^\top R_{(-E)}, \lambda(1 - \alpha) \right)}{\left(X_E + \sum_{j=1}^p \hat{\gamma}_j (X_E \circ \Psi_j) \mathbf{1}_{m_j} \right)^\top \left(X_E + \sum_{j=1}^p \hat{\gamma}_j (X_E \circ \Psi_j) \mathbf{1}_{m_j} \right)} \quad (52)$$

$$\lambda(1 - \alpha) w_j \frac{\theta_j}{\|\theta_j\|_2} = \frac{1}{n} (\Psi_j + \gamma_j (X_E \circ \Psi_j))^\top R_{(-j)} \quad (53)$$

$$\hat{\gamma}_j = \frac{S \left(\frac{1}{n \cdot w_{jE}} ((X_E \circ \Psi_j)(\beta_E \cdot \mathbf{1}_{m_j} + \theta_j))^\top R_{(-jE)}, \lambda\alpha \right)}{((X_E \circ \Psi_j)(\beta_E \cdot \mathbf{1}_{m_j} + \theta_j))^\top ((X_E \circ \Psi_j)(\beta_E \cdot \mathbf{1}_{m_j} + \theta_j))} \quad (54)$$

713 where $S(x, t) = \text{sign}(x)(|x| - t)$ is the soft-thresholding operator. As was the case in the strong

714 heredity **sail** model, there are closed form solutions for the intercept and β_E , each γ_j also

715 has a closed form solution and can be solved efficiently for $j = 1, \dots, p$ using the coordinate

descent procedure implemented in the `glmnet` package [14], while we use the quadratic majorization technique implemented in the `gglasso` package [37] to solve (53). 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 $(\beta_E, \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_p, \gamma_1, \dots, \gamma_p)$ is $\mathbf{0}$ is:

$$\lambda_{max} = \frac{1}{n} \max \left\{ \frac{1}{(1-\alpha)w_E} \left(X_E + \sum_{j=1}^p \gamma_j (X_E \circ \boldsymbol{\Psi}_j) \mathbf{1}_{m_j} \right)^\top R_{(-E)}, \right. \\ \max_j \frac{1}{(1-\alpha)w_j} \left\| (\boldsymbol{\Psi}_j + \gamma_j (X_E \circ \boldsymbol{\Psi}_j))^\top R_{(-j)} \right\|_2, \\ \left. \max_j \frac{1}{\alpha w_{jE}} \left((X_E \circ \boldsymbol{\Psi}_j) (\beta_E \cdot \mathbf{1}_{m_j} + \boldsymbol{\theta}_j) \right)^\top R_{(-jE)} \right\} \quad (55)$$

which reduces to

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

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

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

```

1: function sail( $\mathbf{X}, Y, X_E, \text{basis}, \lambda, \alpha, w_j, w_E, w_{jE}, \epsilon$ ) ▷ Algorithm for solving (43)
2:    $\Psi_j \leftarrow \text{basis}(X_j), \tilde{\Psi}_j \leftarrow X_E \circ \Psi_j$  for  $j = 1, \dots, p$ 
3:   Initialize:  $\beta_0^{(0)} \leftarrow \bar{Y}, \beta_E^{(0)} = \theta_j^{(0)} = \gamma_j^{(0)} \leftarrow 0$  for  $j = 1, \dots, p$ .
4:   Set iteration counter  $k \leftarrow 0$ 
5:    $R^* \leftarrow Y - \beta_0^{(k)} - \beta_E^{(k)} X_E - \sum_j \Psi_j \theta_j^{(k)} - \sum_j \gamma_j^{(k)} \tilde{\Psi}_j (\beta_E^{(k)} \cdot \mathbf{1}_{m_j} + \theta_j^{(k)})$ 
6:   repeat
7:     • To update  $\gamma = (\gamma_1, \dots, \gamma_p)$ 
8:        $\tilde{X}_j \leftarrow \tilde{\Psi}_j (\beta_E^{(k)} \cdot \mathbf{1}_{m_j} + \theta_j^{(k)})$  for  $j = 1, \dots, p$ 
9:        $R \leftarrow R^* + \sum_{j=1}^p \gamma_j^{(k)} \tilde{X}_j$ 
10:
11:       
$$\gamma^{(k)(new)} \leftarrow \arg \min_{\gamma} \frac{1}{2n} \left\| R - \sum_j \gamma_j \tilde{X}_j \right\|_2^2 + \lambda \alpha \sum_j w_{jE} |\gamma_j|$$

12:        $\Delta = \sum_j (\gamma_j^{(k)} - \gamma_j^{(k)(new)}) \tilde{X}_j$ 
13:        $R^* \leftarrow R^* + \Delta$ 
14:     • To update  $\theta = (\theta_1, \dots, \theta_p)$ 
15:        $\tilde{X}_j \leftarrow \Psi_j + \gamma_j^{(k)} \tilde{\Psi}_j$  for  $j = 1, \dots, p$ 
16:       for  $j = 1, \dots, p$  do
17:          $R \leftarrow R^* + \tilde{X}_j \theta_j^{(k)}$ 
18:
19:         
$$\theta_j^{(k)(new)} \leftarrow \arg \min_{\theta_j} \frac{1}{2n} \left\| R - \tilde{X}_j \theta_j \right\|_2^2 + \lambda (1 - \alpha) w_j \|\theta_j\|_2$$

20:          $\Delta = \tilde{X}_j (\theta_j^{(k)} - \theta_j^{(k)(new)})$ 
21:          $R^* \leftarrow R^* + \Delta$ 
22:     • To update  $\beta_E$ 
23:        $\tilde{X}_E \leftarrow X_E + \sum_j \gamma_j^{(k)} \tilde{\Psi}_j \mathbf{1}_{m_j}$ 
24:        $R \leftarrow R^* + \beta_E^{(k)} \tilde{X}_E$ 
25:
26:       
$$\beta_E^{(k)(new)} \leftarrow \frac{1}{\tilde{X}_E^\top \tilde{X}_E} S \left( \frac{1}{n \cdot w_E} \tilde{X}_E^\top R, \lambda (1 - \alpha) \right)$$

27:
28:       ▷  $S(x, t) = \text{sign}(x)(|x| - t)_+$ 
29:        $\Delta = (\beta_E^{(k)} - \beta_E^{(k)(new)}) \tilde{X}_E$ 
30:        $R^* \leftarrow R^* + \Delta$ 
31:     • To update  $\beta_0$ 
32:        $R \leftarrow R^* + \beta_0^{(k)}$ 
33:
34:       
$$\beta_0^{(k)(new)} \leftarrow \frac{1}{n} R^* \cdot \mathbf{1}$$

35:
36:        $\Delta = \beta_0^{(k)} - \beta_0^{(k)(new)}$ 
37:        $R^* \leftarrow R^* + \Delta$ 
38:        $k \leftarrow k + 1$ 
39:   until convergence criterion is satisfied:  $|Q(\Phi^{(k-1)}) - Q(\Phi^{(k)})| / Q(\Phi^{(k-1)}) < \epsilon$ 

```

C Additional Results on PRS for Educational Attainment

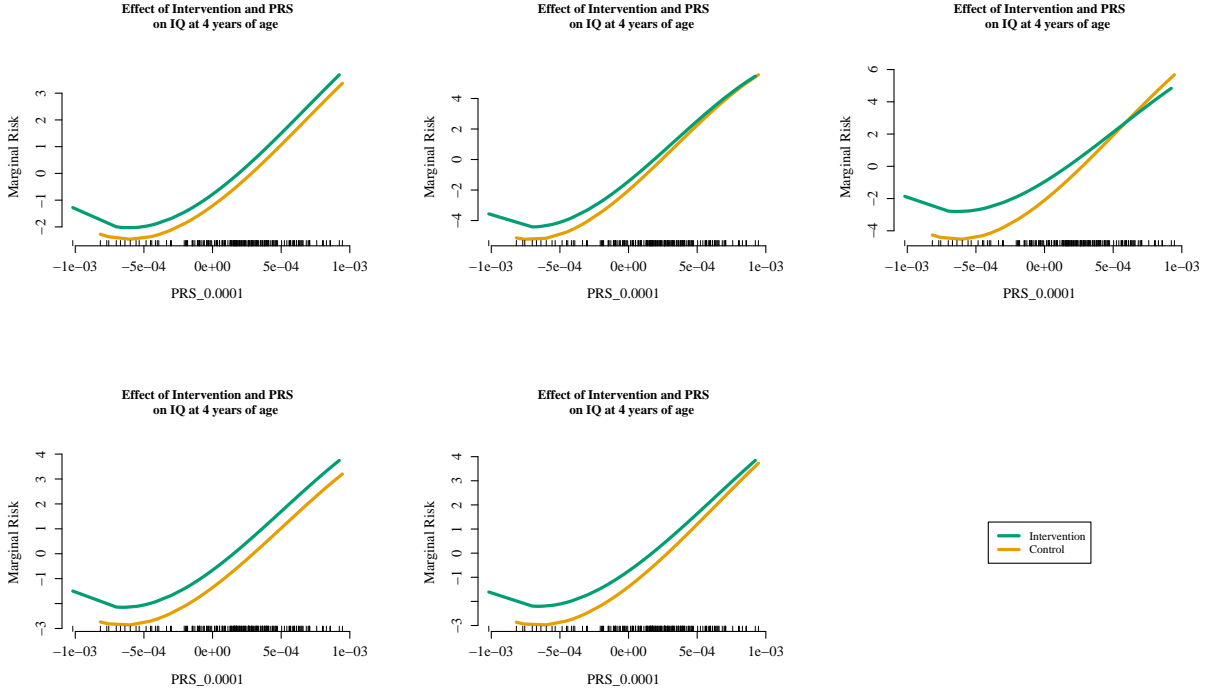


Figure C.1: Estimated interaction effect identified by the weak heredity `sail` using cubic B-splines and $\alpha = 0.1$ for the Nurse Family Partnership data for the 5 imputed datasets. Of the 189 subjects, 19 IQ scores were imputed using `mice` [5]. The selected model, chosen via 10-fold cross-validation, 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.

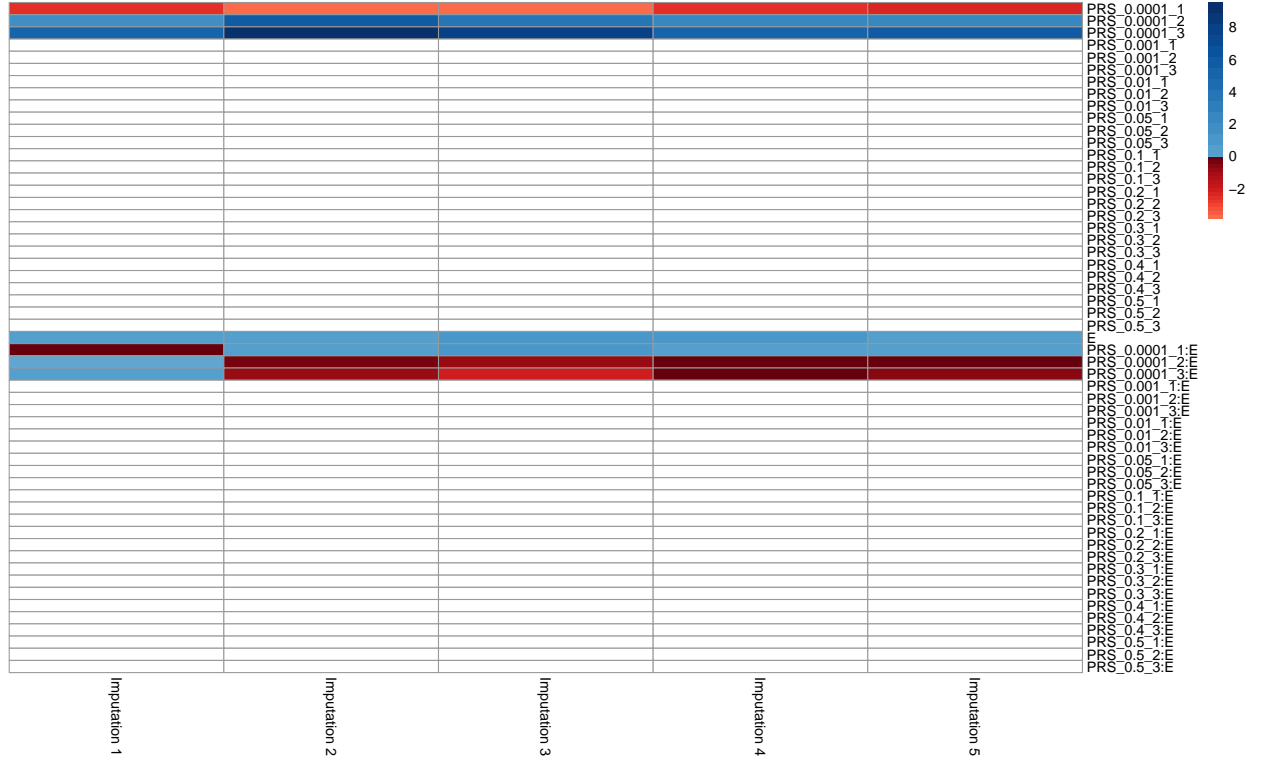


Figure C.2: Coefficient estimates obtained by the weak heredity `sail` using cubic B-splines and $\alpha = 0.1$ for the Nurse Family Partnership data for the 5 imputed datasets. Of the 189 subjects, 19 IQ scores were imputed using `mice` [5]. The selected model, chosen via 10-fold cross-validation, contained 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.

D Data Availability and Code to Reproduce Results

The R scripts used to simulate the data for the simulation studies in Section 4 are provided along with the code for each of the methods being compared. The data used for the two real data analyses in Section 5 are publicly available. The first dataset from the Nurse Family Partnership program is provided by one of the authors of the manuscript (David Olds). The second dataset from the Study to Understand Prognoses Preferences Outcomes and Risks of Treatment (SUPPORT) is publicly available from the Vanderbilt University Department of

Biostatistics website.

D.1 Datasets

The datasets are available at [https://github.com/sahirbhatnagar/sail/tree/jasa/manuscript/
raw_data](https://github.com/sahirbhatnagar/sail/tree/jasa/manuscript/raw_data)

1. Nurse Family Partnership program data consists of three files. They are merged together using the script [https://github.com/sahirbhatnagar/sail/blob/jasa/
manuscript/bin/PRS_bootstrap.R](https://github.com/sahirbhatnagar/sail/blob/jasa/manuscript/bin/PRS_bootstrap.R)

- Gen_3PC_scores.txt
- IQ_and_mental_development_variables_for_Sahir_with_study_ID.txt
- NFP_170614_INFO08_nodup_hard09_noambi_GWAS_EduYears_Pooled_beta_withaf_5000pruned_noambi_16Jan2018.score

2. The SUPPORT data consists of a single file:

- [https://github.com/sahirbhatnagar/sail/blob/jasa/manuscript/raw_data/
support2.csv](https://github.com/sahirbhatnagar/sail/blob/jasa/manuscript/raw_data/support2.csv)

All datasets are in .txt format. Code used to read in the datasets are provided in the section below. All output from this project published online is available according to the conditions of the Creative Commons License ([https://creativecommons.org/licenses/by-nc-sa/
2.0/](https://creativecommons.org/licenses/by-nc-sa/2.0/))

D.2 Code

The software which implements our algorithm is available in an R package published on CRAN (<https://cran.r-project.org/package=sail>) version 0.1.0 with MIT license. The paper itself is written in knitr format, and therefore includes both the code and text in the same `.Rnw` file.

The scripts and data used to produce the results in the manuscript are available at <https://github.com/sahirbhatnagar/sail/tree/jasa/manuscript>.

The knitr file which contains both the main text and code is available at: https://github.com/sahirbhatnagar/sail/blob/jasa/manuscript/source/sail_manuscript_v2.Rnw

The manuscript was compiled using R version 3.6.1 with knitr version 1.25.

The bootstrap analysis was run in parallel on a compute cluster with 40 cores. Though this is not necessary to reproduce the results, it definitely speeds up the computation time.

D.2.1 Instructions for Use

All tables and figures from the paper can be reproduced by compiling the knitr file. The easiest way to reproduce the results is to download the GitHub repository and compile the knitr file from within an R session as follows:

1. Download the GitHub repository <https://github.com/sahirbhatnagar/sail/archive/jasa.zip>
2. From within an R session, run the command: `knitr::knit2pdf('sail_manuscript_v2.Rnw')`

Note that to speed up compilation time, we have saved the simulation and bootstrap re-

sults in `.RData` files available at <https://github.com/sahirbhatnagar/sail/tree/jasa/manuscript/results>. These `.RData` files are called directly by the knitr file.

Note also that the R scripts used to generate the results are called from the knitr file using the ‘code externalization’ functionality of knitr (<https://yihui.org/knitr/demo/externalization/>). That is, the actual R code is stored in R scripts and not within the knitr file. These R scripts are available at <https://github.com/sahirbhatnagar/sail/tree/jasa/manuscript/bin>.

The expected run time to compile the manuscript is about 5 minutes on a standard desktop machine, assuming that you are using the pre-run simulation and bootstrap results.

D.2.2 R Package Vignette

A website with two vignettes has been created for our sail package available at <https://sahirbhatnagar.com/sail/>

The 2 vignettes are:

1. <https://sahirbhatnagar.com/sail/articles/introduction-to-sail.html>
2. <https://sahirbhatnagar.com/sail/articles/user-defined-design.html>