1 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

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

and the objective function is given by

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

Solving (2) 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

$$\frac{\partial Q}{\partial \beta_0} = \frac{1}{n} \left(Y - \beta_0 \cdot \mathbf{1} - \sum_{j=1}^{p} \mathbf{\Psi}_j \boldsymbol{\theta}_j - \beta_E X_E - \sum_{j=1}^{p} \gamma_j \beta_E (X_E \circ \mathbf{\Psi}_j)\boldsymbol{\theta}_j \right)^T \boldsymbol{W}\, \mathbf{1} = 0 \tag{3}$$

$$\frac{\partial Q}{\partial \beta_E} = -\frac{1}{n} \left(X_E + \sum_{j=1}^{p} \gamma_j (X_E \circ \mathbf{\Psi}_j)\boldsymbol{\theta}_j \right)^{\top} \boldsymbol{W} R + \lambda(1-\alpha) w_E s_1 = 0 \tag{4}$$

$$\frac{\partial Q}{\partial \boldsymbol{\theta}_j} = -\frac{1}{n} \left(\mathbf{\Psi}_j + \gamma_j \beta_E (X_E \circ \mathbf{\Psi}_j) \right)^{\top} \boldsymbol{W} R + \lambda(1-\alpha) w_j s_2 = \mathbf{0} \tag{5}$$

$$\frac{\partial Q}{\partial \gamma_j} = -\frac{1}{n} \left(\beta_E (X_E \circ \mathbf{\Psi}_j)\boldsymbol{\theta}_j \right)^{\top} \boldsymbol{W} R + \lambda\alpha w_{jE} s_3 = 0 \tag{6}$$

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:

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

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

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

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 (3)–(6) we see that

$$\hat{\beta}_0 = \left(Y - \sum_{j=1}^{p} \boldsymbol{\Psi}_j \hat{\boldsymbol{\theta}}_j - \hat{\beta}_E X_E - \sum_{j=1}^{p} \hat{\gamma}_j \hat{\beta}_E (X_E \circ \boldsymbol{\Psi}_j) \hat{\boldsymbol{\theta}}_j \right)^{\top} \boldsymbol{W} \mathbf{1} \tag{7}$$

$$\hat{\beta}_E = \frac{S\left(\left(X_E + \sum_{j=1}^{p} \hat{\gamma}_j (X_E \circ \boldsymbol{\Psi}_j) \hat{\boldsymbol{\theta}}_j \right)^{\top} \boldsymbol{W} R_{(-E)}, n \cdot w_E \lambda (1-\alpha) \right)}{\left(X_E + \sum_{j=1}^{p} \hat{\gamma}_j (X_E \circ \boldsymbol{\Psi}_j) \hat{\boldsymbol{\theta}}_j \right)^{\top} \boldsymbol{W} \left(X_E + \sum_{j=1}^{p} \hat{\gamma}_j (X_E \circ \boldsymbol{\Psi}_j) \hat{\boldsymbol{\theta}}_j \right)} \tag{8}$$

$$\hat{\gamma}_j = \frac{S\left((\beta_E (X_E \circ \boldsymbol{\Psi}_j) \boldsymbol{\theta}_j)^{\top} \boldsymbol{W} R_{(-jE)}, n \cdot w_{jE} \lambda \alpha \right)}{(\beta_E (X_E \circ \boldsymbol{\Psi}_j) \boldsymbol{\theta}_j)^{\top} \boldsymbol{W} (\beta_E (X_E \circ \boldsymbol{\Psi}_j) \boldsymbol{\theta}_j)} \tag{9}$$

$$\boldsymbol{\theta}_j = \begin{cases} \mathbf{0} & \text{if } \boldsymbol{A}^T \boldsymbol{W} R_{(-j)} = n\lambda(1-\alpha) w_j u \\ \left(\boldsymbol{A}^T \boldsymbol{W} \boldsymbol{A} + \frac{n\lambda(1-\alpha) w_j}{\|\boldsymbol{\theta}_j\|_2} \right)^{-1} \boldsymbol{A}^T \boldsymbol{W} R_{-j} & \text{if } \boldsymbol{\theta}_j \neq 0. \end{cases} \tag{10}$$

where $S(x,t) = \text{sign}(x)(|x| - t)$ is the soft-thresholding operator and

$$\boldsymbol{A} = \boldsymbol{\Psi}_j + \gamma_j \beta_E (X_E \circ \boldsymbol{\Psi}_j)$$

Now we give some details of the derivations from equation (6) to (10), other derivations are similar. Let us started from a simple lasso problem: the model only contain single variable.

$$\frac{\partial Q}{\partial \beta} = \frac{1}{n} \sum_{i=1}^{n} w_i (y_i - \beta x_i) x_i - \lambda s \tag{11}$$

Set equation (11) to 0 we have $\hat{\beta} = \frac{S\left(\sum_{i=1}^{n} w_i x_i y_i, n\lambda\right)}{\sum_{i=1}^{n} w_i x_i^2}$

Then we go to the lasso problem. Recall that the first derivative of a lasso objective function

problem is

$$\frac{\partial Q}{\partial \beta_j} = \frac{1}{n} w_i \sum_{i=1}^{n} (y_i - \sum_{k=1}^{p} \beta_k x_{ik}) x_{ij} - \lambda s \tag{12}$$

$$= \frac{1}{n} \sum_{i=1}^{n} w_i (y_i - \sum_{k \neq j}^{p} \beta_k x_{ik} - \beta_j x_{ij}) x_{ij} - \lambda s \tag{13}$$

Set equation (12) to 0, we can solve β_j. If we fix other β's, then solve equation (13) is just the same as solve lasso problem with single predictor: $\hat{\beta}_j = \frac{S\left(\sum_{i=1}^{n} w_i x_{ij} R_{-j}, n\lambda\right)}{\sum_{i=1}^{n} w_i x_{ij}^2}$

Now go to our specific problem. To solve equation (6), if all the parameters except for γ_j are fixed, then it is equivalent to solve a lasso problem. Similar to the methods we have used before, it can be shown that

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

.

We see from (7) and (8) that there are closed form solutions for the intercept and β_E. From (9), each γ_j also has a closed form solution and can be solved efficiently for $j = 1, \ldots, p$ using a coordinate descent procedure (**?**). Since there is no closed form solution for β_j, we use a quadratic majorization technique (**?**) to solve (**??**). 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 1.2. From these estimates, we compute the interaction effects using the reparametrizations presented in Table **??**, 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.

1.1 Least-Squares `sail` with Strong Heredity

A more detailed algorithm for fitting the least-squares `sail` model with strong heredity is given in Algorithm 1.

Algorithm 1 Blockwise Coordinate Descent for Least-Squares `sail` with Strong Heredity

function `sail`$(\boldsymbol{X}, Y, \boldsymbol{W}, X_E, \texttt{basis}, \lambda, \alpha, w_j, w_E, w_{jE}, \epsilon)$ ▷ Algorithm for solving (2)

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

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

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

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

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

  • 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**

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

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

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

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

  • To update β_E

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

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

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

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

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

  • To update β_0

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

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

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

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

  $k \leftarrow k + 1$

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

1.2 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

$$R_s = Y - \widetilde{Y}^{(k)}_{(-s)} \tag{14}$$

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

$$\widetilde{Y}^{(k)}_{(-s)} = \beta_0^{(k)} + \beta_E^{(k)} X_E + \sum_{\ell \neq s} \boldsymbol{\Psi}_\ell \boldsymbol{\theta}_\ell^{(k)} + \sum_{\ell \neq s} \gamma_\ell^{(k)} \beta_E^{(k)} \widetilde{\boldsymbol{\Psi}}_\ell \boldsymbol{\theta}_\ell^{(k)} \tag{15}$$

Using (15), (14) can be re-written as

$$\begin{aligned} R_s &= Y - \beta_0^{(k)} - \beta_E^{(k)} X_E - \sum_{j=1}^{p} (\boldsymbol{\Psi}_j + \gamma_j^{(k)} \beta_E^{(k)} \widetilde{\boldsymbol{\Psi}}_j) \boldsymbol{\theta}_j^{(k)} + (\boldsymbol{\Psi}_s + \gamma_s^{(k)} \beta_E^{(k)} \widetilde{\boldsymbol{\Psi}}_s) \boldsymbol{\theta}_s^{(k)} \\ &= R^* + (\boldsymbol{\Psi}_s + \gamma_s^{(k)} \beta_E^{(k)} \widetilde{\boldsymbol{\Psi}}_s) \boldsymbol{\theta}_s^{(k)} \end{aligned} \tag{16}$$

where

$$R^* = Y - \beta_0^{(k)} - \beta_E^{(k)} X_E - \sum_{j=1}^{p} (\boldsymbol{\Psi}_j + \gamma_j^{(k)} \beta_E^{(k)} \widetilde{\boldsymbol{\Psi}}_j) \boldsymbol{\theta}_j^{(k)} \tag{17}$$

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

$$\boldsymbol{\theta}_s^{(k)(new)} = \arg\min_{\boldsymbol{\theta}_j} \frac{1}{2n} \left\| \sqrt{\boldsymbol{W}} (R_s - (\boldsymbol{\Psi}_s + \gamma_s^{(k)} \beta_E^{(k)} \widetilde{\boldsymbol{\Psi}}_s) \boldsymbol{\theta}_j) \right\|_2^2 + \lambda(1-\alpha) w_s \|\theta_j\|_2 \tag{18}$$

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

$$R_{s+1} = R^* + (\boldsymbol{\Psi}_{s+1} + \gamma_{s+1}^{(k)} \beta_E^{(k)} \widetilde{\boldsymbol{\Psi}}_{s+1}) \boldsymbol{\theta}_{s+1}^{(k)} + (\boldsymbol{\Psi}_s + \gamma_s^{(k)} \beta_E^{(k)} \widetilde{\boldsymbol{\Psi}}_s)(\boldsymbol{\theta}_s^{(k)} - \boldsymbol{\theta}_s^{(k)(new)}) \tag{19}$$

where R^* is given by (17), $\boldsymbol{\theta}_s^{(k)}$ is the parameter value prior to the update, and $\boldsymbol{\theta}_s^{(k)(new)}$ is the updated value given by (18). Taking the difference between (16) and (19) gives

$$\begin{aligned}
\Delta &= R_t - R_s \\
&= (\boldsymbol{\Psi}_t + \gamma_t^{(k)}\beta_E^{(k)}\widetilde{\boldsymbol{\Psi}}_t)\boldsymbol{\theta}_t^{(k)} + (\boldsymbol{\Psi}_s + \gamma_s^{(k)}\beta_E^{(k)}\widetilde{\boldsymbol{\Psi}}_s)(\boldsymbol{\theta}_s^{(k)} - \boldsymbol{\theta}_s^{(k)(new)}) - (\boldsymbol{\Psi}_s + \gamma_s^{(k)}\beta_E^{(k)}\widetilde{\boldsymbol{\Psi}}_s)\boldsymbol{\theta}_s^{(k)} \\
&= (\boldsymbol{\Psi}_t + \gamma_t^{(k)}\beta_E^{(k)}\widetilde{\boldsymbol{\Psi}}_t)\boldsymbol{\theta}_t^{(k)} - (\boldsymbol{\Psi}_s + \gamma_s^{(k)}\beta_E^{(k)}\widetilde{\boldsymbol{\Psi}}_s)\boldsymbol{\theta}_s^{(k)(new)} \qquad (20)
\end{aligned}$$

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 (20). This formulation can lead to computational speedups especially when $\Delta = 0$, meaning the partial residual does not need to be re-calculated.

1.3 Least-Squares `sail` with Weak Heredity

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

$$\hat{Y} = \beta_0 \cdot \mathbf{1} + \sum_{j=1}^{p} \boldsymbol{\Psi}_j\boldsymbol{\theta}_j + \beta_E X_E + \sum_{j=1}^{p} \gamma_j(X_E \circ \boldsymbol{\Psi}_j)(\beta_E \cdot \mathbf{1}_{m_j} + \boldsymbol{\theta}_j) \qquad (21)$$

The objective function is given by

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

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

$$\frac{\partial Q}{\partial \beta_0} = \frac{1}{n}\left(Y - \beta_0 \cdot \mathbf{1} - \sum_{j=1}^{p} \boldsymbol{\Psi}_j \boldsymbol{\theta}_j - \beta_E X_E - \sum_{j=1}^{p} \gamma_j (X_E \circ \boldsymbol{\Psi}_j)(\beta_E \cdot \mathbf{1}_{m_j} + \boldsymbol{\theta}_j)\right)^\top \boldsymbol{W}\mathbf{1} = 0 \tag{23}$$

$$\frac{\partial Q}{\partial \beta_E} = -\frac{1}{n}\left(X_E + \sum_{j=1}^{p} \gamma_j (X_E \circ \boldsymbol{\Psi}_j)\mathbf{1}_{m_j}\right)^\top \boldsymbol{W}R + \lambda(1-\alpha)w_E s_1 = 0 \tag{24}$$

$$\frac{\partial Q}{\partial \boldsymbol{\theta}_j} = -\frac{1}{n}\left(\boldsymbol{\Psi}_j + \gamma_j (X_E \circ \boldsymbol{\Psi}_j)\right)^\top \boldsymbol{W}R + \lambda(1-\alpha)w_j s_2 = \mathbf{0} \tag{25}$$

$$\frac{\partial Q}{\partial \gamma_j} = -\frac{1}{n}\left((X_E \circ \boldsymbol{\Psi}_j)(\beta_E \cdot \mathbf{1}_{m_j} + \boldsymbol{\theta}_j)\right)^\top \boldsymbol{W}R + \lambda\alpha w_{jE} s_3 = 0 \tag{26}$$

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

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

the partial residual without X_E as

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

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

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

From the subgradient Equation (24), we see that $\beta_E = 0$ is a solution if

$$\frac{1}{w_E} \left| \frac{1}{n} \left(X_E + \sum_{j=1}^{p} \gamma_j (X_E \circ \boldsymbol{\Psi}_j) \mathbf{1}_{m_j} \right)^\top \boldsymbol{W} R_{(-E)} \right| \leq \lambda(1 - \alpha) \tag{27}$$

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

$$\frac{1}{w_j} \left\| \frac{1}{n} \left(\boldsymbol{\Psi}_j + \gamma_j (X_E \circ \boldsymbol{\Psi}_j) \right)^\top \boldsymbol{W} R_{(-j)} \right\|_2 \leq \lambda(1 - \alpha) \tag{28}$$

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

$$\frac{1}{w_{jE}} \left| \frac{1}{n} \left((X_E \circ \boldsymbol{\Psi}_j)(\beta_E \cdot \mathbf{1}_{m_j} + \boldsymbol{\theta}_j) \right)^\top \boldsymbol{W} R_{(-jE)} \right| \leq \lambda \alpha \tag{29}$$

From the subgradient equations we see that

$$\hat{\beta}_0 = \left(Y - \sum_{j=1}^{p} \boldsymbol{\Psi}_j \hat{\boldsymbol{\theta}}_j - \hat{\beta}_E X_E - \sum_{j=1}^{p} \hat{\gamma}_j (X_E \circ \boldsymbol{\Psi}_j)(\hat{\beta}_E \cdot \mathbf{1}_{m_j} + \hat{\boldsymbol{\theta}}_j) \right)^\top \boldsymbol{W} \mathbf{1} \tag{30}$$

$$\hat{\beta}_E = \frac{S\left(\left(X_E + \sum_{j=1}^{p} \hat{\gamma}_j (X_E \circ \boldsymbol{\Psi}_j) \mathbf{1}_{m_j} \right)^\top \boldsymbol{W} R_{(-E)}, n \cdot w_E \lambda (1-\alpha) \right)}{\left(X_E + \sum_{j=1}^{p} \hat{\gamma}_j (X_E \circ \boldsymbol{\Psi}_j) \mathbf{1}_{m_j} \right)^\top \boldsymbol{W} \left(X_E + \sum_{j=1}^{p} \hat{\gamma}_j (X_E \circ \boldsymbol{\Psi}_j) \mathbf{1}_{m_j} \right)} \tag{31}$$

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

$$\boldsymbol{\theta}_j = \begin{cases} \mathbf{0}, & if\ \boldsymbol{A}^T \boldsymbol{W} (R_{(-j)} - \gamma_j (X_E \circ \boldsymbol{\Psi}_j) \beta_E \mathbf{1}_{m_j}) = n\lambda(1-\alpha) w_j u \\ \left(\boldsymbol{A}^T \boldsymbol{W} \boldsymbol{A} + \frac{n\lambda(1-\alpha) w_j}{\|\boldsymbol{\theta}_j\|_2} \right)^{-1} \boldsymbol{A}^T \boldsymbol{W} (R_{(-j)} - \gamma_j (X_E \circ \boldsymbol{\Psi}_j) \beta_E \mathbf{1}_{m_j}) & \text{if } \boldsymbol{\theta}_j \neq 0. \end{cases} \tag{33}$$

where $S(x,t) = \text{sign}(x)(|x| - t)$ is the soft-thresholding operator and

$$\boldsymbol{A} = \boldsymbol{\Psi}_j + \gamma_j (X_E \circ \boldsymbol{\Psi}_j)$$

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 (**?**), while we use the quadratic majorization technique implemented in the `gglasso` package (**?**) to solve (**??**). Algorithm 2 details the procedure used to fit the least-squares weak heredity `sail` model.

Algorithm 2 Coordinate descent for least-squares `sail` with weak heredity

function `sail`$(\boldsymbol{X}, Y, \boldsymbol{W}, X_E, \texttt{basis}, \lambda, \alpha, w_j, w_E, w_{jE}, \epsilon)$ ▷ Algorithm for solving (22)

  $\Psi_j \leftarrow \texttt{basis}(X_j)$, $\widetilde{\Psi}_j \leftarrow X_E \circ \Psi_j$ 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 $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)$

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

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

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

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

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

    • 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**

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

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

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

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

    • To update β_E

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

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

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

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

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

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

    • To update β_0

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

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

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

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

    $k \leftarrow k + 1$

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