
¹ **A Sparse Additive Model for High-Dimensional Interactions with
2 an Exposure Variable**

³ Sahir R Bhatnagar^{1,2}, Tianyuan Lu^{3,4}, Amanda Lovato⁵, David L Olds⁶, Michael S Kobor⁷,
⁴ Michael J Meaney⁸, Kieran O'Donnell⁹, Yi Yang¹⁰, and Celia MT Greenwood^{1,3,5}

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

¹² **Abstract**

¹³ A conceptual paradigm for onset of a new disease is often considered to be the result
¹⁴ of changes in entire biological networks whose states are affected by a complex interac-
¹⁵ tion of genetic and environmental factors. However, when modelling a relevant pheno-
¹⁶ type as a function of high dimensional measurements, power to estimate interactions
¹⁷ is low, the number of possible interactions could be enormous and their effects may be
¹⁸ non-linear. Existing approaches for high dimensional modelling such as the lasso might
¹⁹ keep an interaction but remove a main effect, which is problematic for interpretation.
²⁰ In this work, we introduce a method called **sail** for detecting non-linear interactions
²¹ with a key environmental or exposure variable in high-dimensional settings which re-
²² spects either the strong or weak heredity constraints. We prove that asymptotically, our
²³ method possesses the oracle property, i.e., it performs as well as if the true model were
²⁴ known in advance. We develop a computationally efficient fitting algorithm with auto-
²⁵ matic tuning parameter selection, which scales to high-dimensional datasets. Through
²⁶ an extensive simulation study, we show that **sail** outperforms existing penalized re-
²⁷ gression methods in terms of prediction accuracy and support recovery when there are

28 non-linear interactions with an exposure variable. We then apply `sail` to detect non-
29 linear interactions between genes and a prenatal psychosocial intervention program on
30 cognitive performance in children at 4 years of age. Results from our method show that
31 individuals who are genetically predisposed to lower educational attainment are those
32 who stand to benefit the most from the intervention. Our algorithms are implemented
33 in an R package available on CRAN (<https://cran.r-project.org/package=sail>).

34 *Keywords:* Variable selection, Gene-environment interaction, Penalized regression, High-
35 dimensional data, Hierarchical interaction, Blockwise coordinate descent

36 1 Introduction

37 Computational approaches to variable selection have become increasingly important with
38 the advent of high-throughput technologies in genomics and brain imaging studies, where
39 the data has become massive, yet where it is believed that the number of truly important
40 variables is small relative to the total number of variables. Although many approaches
41 have been developed for main effects, there is an enduring interest in powerful methods for
42 estimating interactions, since interactions may reflect important modulation of a genomic
43 system by an external factor and vice versa [2]. Accurate capture of interactions may hold the
44 potential to better understanding biological phenomena and improving prediction accuracy.
45 For example, a model that considered interactions between brain imaging data and genetic
46 features had better classification accuracy compared to a model that considered the main
47 effects only [24]. Furthermore, the manifestations of disease are often considered to be
48 the result of changes in entire biological networks whose states are affected by a complex
49 interaction of genetic and environmental factors [31]. However, there is a general deficit of
50 such replicated interactions in the literature [36]. Indeed, power to detect interactions is
51 always lower than for main effects, and in high-dimensional settings ($p \gg n$), this lack of
52 power to detect interactions is exacerbated, since the number of possible interactions could

53 be enormous and their effects may be non-linear. Hence, analytic methods that may improve
54 power are essential. Furthermore, methods capable of detecting non-linear interactions are
55 uncommon. Interactions may occur in numerous types and of varying complexities. In
56 this paper, we consider one specific type of interaction model, where one exposure variable
57 E is involved in possibly non-linear interactions with a high-dimensional set of measures
58 \mathbf{X} leading to effects on a response variable, Y . We propose a multivariable penalization
59 procedure for detecting non-linear interactions between \mathbf{X} and E . Our method is motivated
60 by the Nurse Family Partnership (NFP); a program of prenatal and infancy home visiting
61 by nurses for low-income mothers and their children [26]. In this intervention, NFP nurses
62 guided pregnant women and parents of young children to improve the outcomes of pregnancy,
63 their children's health and development, and their economic self-sufficiency, with the goal
64 of reducing disparities over the life-course. Early intervention in young children has been
65 shown to positively impact intellectual abilities [6], and more recent studies have shown that
66 cognitive performance is also strongly influenced by genetic factors [30]. Given the important
67 role of both environment and genetics, we are interested in finding interactions between these
68 two components on cognitive function in children.

69 1.1 A sparse additive interaction model

70 Let $Y = (Y_1, \dots, Y_n) \in \mathbb{R}^n$ be a continuous outcome variable, $X_E = (E_1, \dots, E_n) \in \mathbb{R}^n$ a bi-
71 nary or continuous environment/exposure vector of known importance, and $\mathbf{X} = (X_1, \dots, X_p) \in \mathbb{R}^{n \times p}$
72 a matrix of additional predictors, possibly high-dimensional. Furthermore let $f_j : \mathbb{R} \rightarrow \mathbb{R}$ be
73 a smoothing method for variable X_j by a projection on to a set of basis functions:

$$f_j(X_j) = \sum_{\ell=1}^{m_j} \psi_{j\ell}(X_j) \beta_{j\ell} \quad (1)$$

Here, the $\{\psi_{j\ell}\}_1^{m_j}$ are a family of basis functions in X_j [18]. Let Ψ_j be the $n \times m_j$ matrix of
evaluations of the $\psi_{j\ell}$ and $\boldsymbol{\theta}_j = (\beta_{j1}, \dots, \beta_{jm_j}) \in \mathbb{R}^{m_j}$ for $j = 1, \dots, p$ ($\boldsymbol{\theta}_j$ is a m_j -dimensional

column vector of basis coefficients for the j th main effect). In this article we consider an additive interaction regression model of the form

$$Y = \beta_0 \cdot \mathbf{1}_n + \sum_{j=1}^p \Psi_j \boldsymbol{\theta}_j + \beta_E X_E + \sum_{j=1}^p (X_E \circ \Psi_j) \boldsymbol{\tau}_j + \varepsilon \quad (2)$$

where $\beta_0 \in \mathbb{R}$ is the intercept, $\beta_E \in \mathbb{R}$ is the coefficient for the environment variable, $\boldsymbol{\tau}_j = (\tau_{j1}, \dots, \tau_{jm_j}) \in \mathbb{R}^{m_j}$ are the basis coefficients for the j th interaction term, $(X_E \circ \Psi_j)$ is the $n \times m_j$ matrix formed by the component-wise multiplication of the column vector X_E by each column of Ψ_j , and $\varepsilon \in \mathbb{R}^n$ is a vector of i.i.d errors with mean zero and finite variance.

Here we assume that p is large relative to n , and particularly that $\sum_{j=1}^p m_j/n$ is large. Due to the large number of parameters to estimate with respect to the number of observations, one commonly-used approach in the penalization literature is to shrink the regression coefficients by placing a constraint on the values of $(\beta_E, \boldsymbol{\theta}_j, \boldsymbol{\tau}_j)$. Certain constraints have the added benefit of producing a sparse model in the sense that many of the coefficients will be set exactly to 0 [4]. Such a reduced predictor set can lead to a more interpretable model with smaller prediction variance, albeit at the cost of having biased parameter estimates [12]. In light of these goals, consider the following penalized objective function:

$$Q(\Phi) = -L(\Phi) + \lambda(1 - \alpha) \left(w_E |\beta_E| + \sum_{j=1}^p w_j \|\boldsymbol{\theta}_j\|_2 \right) + \lambda \alpha \sum_{j=1}^p w_{jE} \|\boldsymbol{\tau}_j\|_2 \quad (3)$$

where $\Phi = (\beta_0, \beta_E, \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_p, \boldsymbol{\tau}_1, \dots, \boldsymbol{\tau}_p)$, $L(\Phi)$ is the log-likelihood function of the observations $\mathbf{V}_i = (Y_i, \Psi_i, X_{iE})$ for $i = 1, \dots, n$, $\|\boldsymbol{\theta}_j\|_2 = \sqrt{\sum_{k=1}^{m_j} \beta_{jk}^2}$, $\|\boldsymbol{\tau}_j\|_2 = \sqrt{\sum_{k=1}^{m_j} \tau_{jk}^2}$, $\lambda > 0$ and $\alpha \in (0, 1)$ are adjustable tuning parameters, w_E, w_j, w_{jE} are non-negative penalty factors for $j = 1, \dots, p$ which serve as a way of allowing parameters to be penalized differently. The first term in the penalty penalizes the main effects while the second term penalizes the interactions. The parameter α controls the relative weight on the two penalties. Note that we do not penalize the intercept.

93 An issue with (3) is that since no constraint is placed on the structure of the model, it is
 94 possible that an estimated interaction term is non-zero while the corresponding main effects
 95 are zero. While there may be certain situations where this is plausible, statisticians have gen-
 96 erally argued that interactions should only be included if the corresponding main effects are
 97 also in the model [22]. This is known as the strong heredity principle [7]. Indeed, large main
 98 effects are more likely to lead to detectable interactions [11]. In the next section we discuss
 99 how a simple reparametrization of the model (3) can lead to this desirable property.

100 1.2 Strong and weak heredity

101 The strong heredity principle states that an interaction term can only have a non-zero es-
 102 timate if its corresponding main effects are estimated to be non-zero, whereas the weak
 103 heredity principle allows for a non-zero interaction estimate as long as one of the corre-
 104 sponding main effects is estimated to be non-zero [7]. In the context of penalized regression
 105 methods, these principles can be formulated as structured sparsity [1] problems. Several
 106 authors have proposed to modify the type of penalty in order to achieve the heredity princi-
 107 ple [3, 17, 20, 28]. We take an alternative approach. Following Choi et al. [8], we introduce
 108 a new set of parameters $\boldsymbol{\gamma} = (\gamma_{1E}, \dots, \gamma_{pE}) \in \mathbb{R}^p$ and reparametrize the coefficients for the
 109 interaction terms $\boldsymbol{\tau}_j$ in (2) as a function of γ_{jE} and the main effect parameters $\boldsymbol{\theta}_j$ and β_E .
 110 This reparametrization for both strong and weak heredity is summarized in Table 1.

111 To perform variable selection in this new parametrization, we penalize $\boldsymbol{\gamma} = (\gamma_{1E}, \dots, \gamma_{pE})$
 112 instead of penalizing $\boldsymbol{\tau}$ as in (3), leading to the following penalized objective function:

$$Q(\boldsymbol{\Phi}) = -L(\boldsymbol{\Phi}) + \lambda(1 - \alpha) \left(w_E |\beta_E| + \sum_{j=1}^p w_j \|\boldsymbol{\theta}_j\|_2 \right) + \lambda\alpha \sum_{j=1}^p w_{jE} |\gamma_{jE}| \quad (4)$$

113 An estimate of the regression parameters is given by $\widehat{\boldsymbol{\Phi}} = \arg \min_{\boldsymbol{\Phi}} Q(\boldsymbol{\Phi})$. This penalty allows
 114 for the possibility of excluding the interaction term from the model even if the correspond-
 115 ing main effects are non-zero. Furthermore, smaller values for α would lead to more interactions

116 being included in the final model while values approaching 1 would favor main effects. Similar
117 to the elastic net [42], we fix α and obtain a solution path over a sequence of λ values.

118 **1.3 Toy example**

119 We present here a toy example to better illustrate the methods proposed in this paper. With
120 a sample size of $n = 100$, we sample $p = 20$ covariates X_1, \dots, X_p independently from a $N(0, 1)$
121 distribution truncated to the interval $[0, 1]$. Data were generated from a model which follows
122 the strong heredity principle, but where only one covariate, X_2 , is involved in an interaction
123 with a binary exposure variable (E): $Y = f_1(X_1) + f_2(X_2) + 1.75E + 1.5E \cdot f_2(X_2) + \varepsilon$.
124 For illustration, function $f_1(\cdot)$ is assumed to be linear, whereas function $f_2(\cdot)$ is non-linear:
125 $f_1(x) = -3x$, $f_2(x) = 2(2x - 1)^3$. The error term ε is generated from a normal distribution
126 with variance chosen such that the signal-to-noise ratio (SNR) is 2. We generated a single
127 simulated dataset and used the strong heredity **sail** method (described below) with cubic B-
128 splines to estimate the functional forms. 10-fold cross-validation (CV) was used to choose the
129 optimal value of penalization. We used $\alpha = 0.5$ and default values for all other arguments.
130 We plot the solution path for both main effects and interactions in Figure 1, coloring lines to
131 correspond to the selected model. We see that our method is able to correctly identify the true
132 model. We can also visually see the effect of the penalty and strong heredity principle working
133 in tandem, i.e., the interaction term $E \cdot f_2(X_2)$ (orange lines in the bottom panel) can only
134 be non-zero if the main effects E and $f_2(X_2)$ (black and orange lines respectively in the top
135 panel) are non-zero, while non-zero main effects does not imply a non-zero interaction.

136 In Figure 2, we plot the true and estimated component functions $\hat{f}_1(X_1)$ and $E \cdot \hat{f}_2(X_2)$, and
137 their estimates from this analysis with **sail**. We are able to capture the shape of the correct
138 functional form, but the means are not well aligned with the data. Lack-of-fit for $f_1(X_1)$
139 can be partially explained by acknowledging that **sail** is trying to fit a cubic spline to a
140 linear function. Nevertheless, this example demonstrates that **sail** can still identify trends
141 reasonably well.

142 1.4 Related work

143 Methods for variable selection of interactions can be broken down into two categories: linear
 144 and non-linear interaction effects. Many of the linear effect methods consider all pairwise
 145 interactions in \mathbf{X} [3, 8, 33, 40] which can be computationally prohibitive when p is large.
 146 More recent proposals for selection of interactions allow the user to restrict the search space
 147 to interaction candidates [17, 20]. This is useful when the researcher wants to impose prior
 148 information on the model. Two-stage procedures, where interaction candidates are con-
 149 sidered from an original screen of main effects, have shown good performance when p is
 150 large [15, 32] in the linear setting. There are many fewer methods available for estimating
 151 non-linear interactions. For example, Radchenko and James (2010) [28] proposed a model
 152 of the form $Y = \beta_0 + \sum_{j=1}^p f_j(X_j) + \sum_{j>k} f_{jk}(X_j, X_k) + \varepsilon$, where $f(\cdot)$ are smooth compo-
 153 nent functions. This method is more computationally expensive than **sail** since it considers
 154 all pairwise interactions between the basis functions, and its effectiveness in simulations or
 155 real-data applications is unknown as there is no software implementation.

156 While working on this paper, we were made aware of the recently proposed pliable lasso [35]
 157 which considers the interactions between $\mathbf{X}_{n \times p}$ and another matrix $\mathbf{Z}_{n \times K}$ and takes the
 158 form

$$Y = \beta_0 + \sum_{j=1}^p \beta_j X_j + \sum_{j=1}^K \theta_j Z_j + \sum_{j=1}^p (X_j \circ \mathbf{Z}) \boldsymbol{\alpha}_j + \varepsilon \quad (5)$$

159 where $\boldsymbol{\alpha}_j$ is a K -dimensional vector. Our proposal is most closely related to this method
 160 with \mathbf{Z} being a single column matrix; the key difference being the non-linearity effects of our
 161 predictor variables. As pointed out by the authors of the pliable lasso, either their method
 162 or ours can be seen as a varying coefficient model, i.e., the effect of X varies as a function
 163 of the exposure variable E or \mathbf{Z} in (5).

164 The main contributions of this paper are five-fold. First, we develop a model for non-
 165 linear interactions with a key exposure variable, following either the weak or strong hered-
 166 ity principle, that is computationally efficient and scales to the high-dimensional setting

167 ($n << p$). Second, through simulation studies, we show improved performance in terms of
168 prediction accuracy and support recovery over existing methods that only consider linear
169 interactions or additive main effects. Third, we show that our method possesses the oracle
170 property [13], i.e., it performs as well as if the true model were known in advance. Fourth,
171 we demonstrate the performance of our method in two applications: 1) gene-environment
172 interactions in a prenatal psychosocial intervention program [26] and 2) a study aimed at
173 identifying which clinical variables influence mortality rates amongst seriously ill hospital-
174 ized patients [10]. Fifth, we implement our algorithms in the **sail** R package on CRAN
175 (<https://cran.r-project.org/package=sail>), along with extensive documentation. In
176 particular, our implementation also allows for linear interaction models, user-defined basis
177 expansions, a cross-validation procedure for selecting the optimal tuning parameter, and
178 differential shrinkage parameters to apply the adaptive lasso idea [41].

179 The rest of the paper is organized as follows. Section 2 describes our optimization procedure
180 and some details about the algorithm used to fit the **sail** model for the least squares case.
181 Theoretical results are given in Section 3. In Section 4, through simulation studies we
182 compare the performance of our proposed approach and demonstrate the scenarios where it
183 can be advantageous to use **sail** over existing methods. Section 5 contains two real data
184 examples and Section 6 discusses some limitations and future directions.

185 2 Computation

186 In this section we describe a blockwise coordinate descent algorithm for fitting the least-
187 squares version of the **sail** model in (4). We fix the value for α and minimize the objective
188 function over a decreasing sequence of λ values ($\lambda_{max} > \dots > \lambda_{min}$). We use the subgradi-
189 ent equations to determine the maximal value λ_{max} such that all estimates are zero. Due
190 to the heredity principle, this reduces to finding the largest λ such that all main effects
191 ($\beta_E, \theta_1, \dots, \theta_p$) are zero. Following Friedman et al. [14], we construct a λ -sequence of 100

192 values decreasing from λ_{max} to $0.001\lambda_{max}$ on the log scale, and use the warm start strategy
193 where the solution for λ_ℓ is used as a starting value for $\lambda_{\ell+1}$.

194 2.1 Blockwise coordinate descent for least-squares loss

195 The strong heredity **sail** model with least-squares loss has the form

$$\hat{Y} = \beta_0 \cdot \mathbf{1} + \sum_{j=1}^p \Psi_j \boldsymbol{\theta}_j + \beta_E X_E + \sum_{j=1}^p \gamma_{jE} \beta_E (X_E \circ \Psi_j) \boldsymbol{\theta}_j \quad (6)$$

196 and the objective function is given by

$$Q(\boldsymbol{\Phi}) = \frac{1}{2n} \left\| Y - \hat{Y} \right\|_2^2 + \lambda(1 - \alpha) \left(w_E |\beta_E| + \sum_{j=1}^p w_j \|\boldsymbol{\theta}_j\|_2 \right) + \lambda \alpha \sum_{j=1}^p w_{jE} |\gamma_{jE}| \quad (7)$$

197 Solving (7) in a blockwise manner allows us to leverage computationally fast algorithms for
198 ℓ_1 and ℓ_2 norm penalized regression. We show in Supplemental Section B that by careful
199 construction of pseudo responses and pseudo design matrices, existing efficient algorithms can
200 be used to estimate the parameters. Indeed, the objective function simplifies to a modified
201 lasso problem when holding all $\boldsymbol{\theta}_j$ fixed, and a modified group lasso problem when holding
202 β_E and all γ_{jE} fixed. We provide an overview of the computations in Algorithm 1.

203 2.2 Weak Heredity

204 Our method can be easily adapted to enforce the weak heredity property. That is, an
205 interaction term can only be present if at least one of its corresponding main effects is
206 non-zero. To do so, we reparametrize the coefficients for the interaction terms in (2) as
207 $\boldsymbol{\alpha}_j = \gamma_{jE} (\beta_E \cdot \mathbf{1}_{m_j} + \boldsymbol{\theta}_j)$, where $\mathbf{1}_{m_j}$ is a vector of ones with dimension m_j (i.e. the length of $\boldsymbol{\theta}_j$).
208 We defer the algorithm details for fitting the **sail** model with weak heredity in Supplemental
209 Section B.4, as it is very similar to Algorithm 1 for the strong heredity **sail** model.

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

For a decreasing sequence $\lambda = \lambda_{max}, \dots, \lambda_{min}$ and fixed α :

1. Initialize $\beta_0^{(0)}, \beta_E^{(0)}, \boldsymbol{\theta}_j^{(0)}, \gamma_{jE}^{(0)}$ for $j = 1, \dots, p$ and set iteration counter $k \leftarrow 0$.
 2. Repeat the following until convergence:
 - (a) update $\boldsymbol{\gamma} = (\gamma_{1E}, \dots, \gamma_{pE})$
 - i. Compute the pseudo design: $\tilde{X}_j \leftarrow \beta_E^{(k)}(X_E \circ \boldsymbol{\Psi}_j)\boldsymbol{\theta}_j^{(k)}$ for $j = 1, \dots, p$
 - ii. Compute the pseudo response \tilde{Y} by removing the contribution of every term not involving $\boldsymbol{\gamma}$ from Y
 - iii. Solve:
$$\boldsymbol{\gamma}^{(k)(new)} \leftarrow \arg \min_{\boldsymbol{\gamma}} \frac{1}{2n} \left\| \tilde{Y} - \sum_j \gamma_{jE} \tilde{X}_j \right\|_2^2 + \lambda \alpha \sum_j w_{jE} |\gamma_{jE}| \quad (8)$$
 - iv. Set $\boldsymbol{\gamma}^{(k)} = \boldsymbol{\gamma}^{(k)(new)}$
 - (b) update $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_p)$
 - for $j = 1, \dots, p$
 - i. Compute the pseudo design: $\tilde{X}_j \leftarrow \boldsymbol{\Psi}_j + \gamma_{jE}^{(k)} \beta_E^{(k)}(X_E \circ \boldsymbol{\Psi}_j)$
 - ii. Compute the pseudo response (\tilde{Y}) by removing the contribution of every term not involving $\boldsymbol{\theta}_j$ from Y
 - iii. Solve:
$$\boldsymbol{\theta}_j^{(k)(new)} \leftarrow \arg \min_{\boldsymbol{\theta}_j} \frac{1}{2n} \left\| \tilde{Y} - \tilde{X}_j \boldsymbol{\theta}_j \right\|_2^2 + \lambda(1 - \alpha) w_j \|\boldsymbol{\theta}_j\|_2 \quad (9)$$
 - iv. Set $\boldsymbol{\theta}_j^{(k)} \leftarrow \boldsymbol{\theta}_j^{(k)(new)}$
 - (c) update β_E
 - i. Compute the pseudo design: $\tilde{X}_E \leftarrow X_E + \sum_j \gamma_{jE}^{(k)}(X_E \circ \boldsymbol{\Psi}_j)\boldsymbol{\theta}_j^{(k)}$
 - ii. Compute the pseudo response (\tilde{Y}) by removing the contribution of every term not involving β_E from Y
 - iii. Soft-threshold update ($S(x, t) = \text{sign}(x)(|x| - t)_+$):
$$\beta_E^{(k)(new)} \leftarrow \frac{1}{\tilde{X}_E^\top \tilde{X}_E} S \left(\frac{1}{n \cdot w_E} \tilde{X}_E^\top \tilde{Y}, \lambda(1 - \alpha) \right) \quad (10)$$
 - iv. Set $\beta_E^{(k+1)} \leftarrow \beta_E^{(k)(new)}$, $k \leftarrow k + 1$
-

2.3 Adaptive sail

The weights for the environment variable, main effects and interactions are given by w_E, w_j and w_{jE} respectively. These weights serve as a means of allowing a different penalty to be applied to each variable. In particular, any variable with a weight of zero is not penalized at all. This feature is usually selected for one of two reasons:

1. Prior knowledge about the importance of certain variables is known. Larger weights will penalize the variable more, while smaller weights will penalize the variable less
2. Allows users to apply the adaptive `sail`, similar to the adaptive lasso [41]

We describe the adaptive `sail` in Algorithm 2. This is a general procedure that can be applied to the weak and strong heredity settings, as well as both least squares and logistic loss functions. We provide this capability in the `sail` package using the `penalty.factor` argument.

Algorithm 2 Adaptive `sail` algorithm

1. For a decreasing sequence $\lambda = \lambda_{max}, \dots, \lambda_{min}$ and fixed α run the `sail` algorithm
 2. Use cross-validation or a data splitting procedure to determine the optimal value for the tuning parameter: $\lambda^{[opt]} \in \{\lambda_{max}, \dots, \lambda_{min}\}$
 3. Let $\widehat{\beta}_E^{[opt]}, \widehat{\boldsymbol{\theta}}_j^{[opt]}$ and $\widehat{\boldsymbol{\tau}}_j^{[opt]}$ for $j = 1, \dots, p$ be the coefficient estimates corresponding to the model at $\lambda^{[opt]}$
 4. Set the weights to be

$$w_E = \left(\left| \widehat{\beta}_E^{[opt]} \right| + 1/n \right)^{-1}, w_j = \left(\| \widehat{\boldsymbol{\theta}}_j^{[opt]} \|_2 + 1/n \right)^{-1}, w_{jE} = \left(\| \widehat{\boldsymbol{\tau}}_j^{[opt]} \|_2 + 1/n \right)^{-1}$$
for $j = 1, \dots, p$
 5. Run the `sail` algorithm with the weights defined in step 4), and use cross-validation or a data splitting procedure to choose the optimal value of λ
-

2.4 Flexible design matrix

The definition of the basis expansion functions in (1) is very flexible, in the sense that our algorithms are independent of this choice. As a result, the user can apply any basis expansion they desire. In the extreme case, one could apply the identity map, i.e., $f_j(X_j) = X_j$ which

leads to a linear interaction model (referred to as `linear sail`). When little information is known a priori about the relationship between the predictors and the response, by default, we choose to apply the same basis expansion to all columns of \mathbf{X} . This is a reasonable approach when all the variables are continuous. However, there are often situations when the data contains a combination of categorical and continuous variables. In these cases it may be sub-optimal to apply a basis expansion to the categorical variables. Owing to the flexible nature of our algorithm, we can handle this scenario in our implementation by allowing a user-defined design matrix. The only extra information needed is the group membership of each column in the design matrix. We illustrate such an example in a vignette of the `sail R` package.

3 Theory

In this section we study the asymptotic behaviour of the `sail` estimator $\hat{\Phi}$, defined as the minimizer of (4), as well as the model selection properties. We show that `sail` possesses the oracle property when the sample size approaches infinity and the number of predictors is fixed. That is, under certain regularity conditions, it performs as well as if the true model were known in advance and has the optimal estimation rate [41]. The regularity conditions and proofs are given in Supplemental Section A.

Let $\Phi^* = (\beta_E^*, \boldsymbol{\theta}_1^{*\top}, \dots, \boldsymbol{\theta}_p^{*\top}, \gamma_{1E}^*, \dots, \gamma_{pE}^*)^\top$ denote the unknown vector of true coefficients in (4). To simplify the notation, we use the representation $\Phi^* = (\boldsymbol{\phi}_1^{*\top}, \boldsymbol{\phi}_2^{*\top}, \dots, \boldsymbol{\phi}_{p+1}^{*\top}, \boldsymbol{\phi}_{p+2}^{*\top}, \dots, \boldsymbol{\phi}_{2p+1}^{*\top})^\top$, where $\boldsymbol{\phi}_1^* = \beta_E^*$, $\boldsymbol{\phi}_2^* = \boldsymbol{\theta}_1^*, \dots, \boldsymbol{\phi}_{p+1}^* = \boldsymbol{\theta}_p^*$, and $\boldsymbol{\phi}_{p+2}^* = \gamma_{1E}^*, \dots, \boldsymbol{\phi}_{2p+1}^* = \gamma_{pE}^*$. Denote by $\mathcal{A} = \{m : \boldsymbol{\phi}_m^* \neq \mathbf{0}\}$ the unknown sparsity pattern of Φ^* , and $\widehat{\mathcal{A}} = \left\{m : \widehat{\boldsymbol{\phi}}_m \neq \mathbf{0}\right\}$ the estimated `sail` model selector. We can rewrite the penalty terms in (4), and consider the `sail` estimates $\widehat{\Phi}_n$ given b

$$\widehat{\Phi}_n = \arg \min_{\Phi} Q_n(\Phi) = -L_n(\Phi) + n\lambda_m \sum_{m=1}^{2p+1} \|\boldsymbol{\phi}_m\|_2, \quad (11)$$

249 where $\lambda_1 = \lambda(1 - \alpha)w_E$, $\lambda_m = \lambda(1 - \alpha)w_m$ for $m = 2, \dots, p + 1$, and $\lambda_m = \lambda\alpha w_{mE}$ for
 250 $m = p + 2, \dots, 2p + 1$. Define

$$\mathcal{A}_1 = \{m : \phi_m^* \neq \mathbf{0} (1 \leq m \leq p+1)\}, \quad \mathcal{A}_2 = \{m : \phi_m^* \neq \mathbf{0} (p+2 \leq m \leq 2p+1)\}, \quad \mathcal{A} = \mathcal{A}_1 \cup \mathcal{A}_2$$

251 that is, \mathcal{A}_1 contains the indices for main effects whose true coefficients are non-zero, and \mathcal{A}_2
 252 contains the indices for interaction terms whose true coefficients are non-zero. Let

$$a_n = \max \{\lambda_m, \lambda_{m'} : m \in \mathcal{A}_1, m' \in \mathcal{A}_2\}$$

253

$$b_n = \min \{\lambda_m, \lambda_{m'} : m \in \mathcal{A}_1^c, m' \in \mathcal{A}_2^c \text{ s.t. } \phi_{m'}^* = \gamma_{jE}^* = 0 \text{ but } \beta_E \neq 0 \text{ and } \theta_j^* \neq \mathbf{0} \quad (1 \leq j \leq p)\}$$

254 Note that our asymptotic results are stated for the main effects and interaction terms only,
 255 even though our formulation includes an unpenalized intercept. Consistency results imme-
 256 diately follow for β_0 since we assume the data has been centered, leading to a closed form
 257 solution for the intercept in the least-squares setting.

258 **Lemma 1.** [Existence of a local minimizer] If $a_n = o(\frac{1}{\sqrt{n}})$ as $n \rightarrow \infty$, i.e. $\sqrt{n}a_n \rightarrow 0$, then

$$259 \quad \|\widehat{\Phi}_n - \Phi^*\|_2 = O_p(\frac{1}{\sqrt{n}})$$

260 Lemma (1) states that if the tuning parameters corresponding to the non-zero coefficients
 261 converge to 0 at a speed faster than $\frac{1}{\sqrt{n}}$, then there exists a local minimizer of $Q_n(\Phi)$ which
 262 is \sqrt{n} -consistent [8, 37].

263 **Theorem 1** (Model selection consistency). If $\sqrt{n}a_n \rightarrow 0$ and $\sqrt{n}b_n \rightarrow \infty$, then

$$P \left(\widehat{\Phi}_{\mathcal{A}_1^c} = \mathbf{0} \right) \rightarrow 1 \quad \text{and} \quad P \left(\widehat{\Phi}_{\mathcal{A}_2^c} = \mathbf{0} \right) \rightarrow 1 \quad (12)$$

264 Theorem (1) shows that **sail** can consistently remove the main effects and interaction terms
 265 which are not associated with the response with high probability. Together with Lemma (1),

we see that the asymptotic behaviour of the penalty terms for the zero and non-zero predictors must be different to satisfy the model selection consistency property (12) [23]. Specifically, when the tuning parameters for the non-zero coefficients converge to 0 faster than $1/\sqrt{n}$ (i.e. $\sqrt{n}a_n \rightarrow 0$) and those for zero coefficients are large enough (i.e. $\sqrt{nb_n} \rightarrow \infty$), the Lemma (1) and Theorem (1) imply that the \sqrt{n} -consistent estimator $\widehat{\Phi}_n$ satisfies $P(\widehat{\Phi}_{\mathcal{A}_c^c} = \mathbf{0}) \rightarrow 1$.

Next, we obtain the asymptotic distribution of the `sail` estimator.

Theorem 2 (Asymptotic normality). *Denote $\mathcal{A} = \mathcal{A}_1 \cup \mathcal{A}_2$. Assume that $\sqrt{n}a_n \rightarrow 0$ and $\sqrt{nb_n} \rightarrow \infty$. Under the regularity conditions, the subvector $\widehat{\Phi}_{\mathcal{A}}$ of the local minimizer $\widehat{\Phi}_n$ given in Lemma (1) satisfies*

$$\sqrt{n} (\widehat{\Phi}_{\mathcal{A}} - \Phi_{\mathcal{A}}^*) \xrightarrow{d} N(\mathbf{0}, \mathbf{I}^{-1}(\Phi_{\mathcal{A}}^*)) , \quad (13)$$

where $\mathbf{I}(\Phi_{\mathcal{A}}^*)$ is the Fisher information matrix for $\Phi_{\mathcal{A}}$ at $\Phi_{\mathcal{A}} = \Phi_{\mathcal{A}}^*$, assuming \mathcal{A}_c is known in advance.

Together, Theorems (1) and (2) establish that if the tuning parameters satisfy the conditions $\sqrt{n}a_n \rightarrow 0$ and $\sqrt{nb_n} \rightarrow \infty$, then as the sample size grows large, `sail` has the oracle property [13]. In order for the conditions on the tuning parameters to be satisfied, we follow the strategies outlined for the adaptive Lasso [41], the adaptive group Lasso [23] and the adaptive elastic-net [43]. That is, we define the adaptive weights as $w_m = \|\widehat{\phi}_m^{\text{init}} + 1/n\|_2^{-\xi}$ for $m = 1, \dots, 2p + 1$, where ξ is a positive constant and $\widehat{\phi}_m^{\text{init}}$ is an initial \sqrt{n} -consistent estimate of ϕ_m^* . Here, the $1/n$ is to avoid division by zero.

285 4 Simulation Study

286 4.1 Comparator Methods

287 Since there are no other packages that directly address our chosen problem, we selected
 288 comparator methods based on the following criteria: 1) penalized regression methods that
 289 can handle high-dimensional data ($n < p$), 2) allowing at least one of linear effects, non-
 290 linear effects or interaction effects, and 3) having a software implementation in R. The selected
 291 methods can be grouped into three categories:

- 292 1. Linear main effects: `lasso` [34], `adaptive lasso` [41]
- 293 2. Linear interactions: `lassoBT` [32], `GLinternet` [20]
- 294 3. Non-linear main effects: `HierBasis` [16], `SPAM` [29], `gamsel` [9]

295 For `GLinternet` we specified the `interactionCandidates` argument so as to only consider
 296 interactions between the environment and all other X variables. For all other methods we
 297 supplied $(\mathbf{X}, \mathbf{X}_E)$ as the data matrix, 100 for the number of tuning parameters to fit, and
 298 used the default values otherwise¹. `lassoBT` considers all pairwise interactions as there is
 299 no way for the user to restrict the search space. `SPAM` applies the same basis expansion to
 300 every column of the data matrix; we chose 5 basis spline functions. `HierBasis` and `gamsel`
 301 selects whether a term in an additive model is non-zero, linear, or a non-linear spline up to
 302 a specified max degrees of freedom per variable.

303 We compare the above listed methods with our main proposal method `sail`, as well as
 304 with `adaptive sail` (Algorithm 2) and `sail weak` which has the weak heredity property.
 305 For each function f_j , we use a B-spline basis matrix with `degree=5` implemented in the `bs`
 306 function in R [27]. We center the environment variable and the basis functions before running
 307 the `sail` method.

¹R code for each method available at https://github.com/sahirbhatnagar/sail/blob/master/my_sims/method_functions.R

308 **4.2 Simulation Design**

309 To make the comparisons with other methods as fair as possible, we followed a simulation
310 framework that has been previously used for variable selection methods in additive mod-
311 els [19, 21]. We extend this framework to include interaction effects as well. The covariates
312 are simulated as follows. First, we generate x_1, \dots, x_{1000} independently from a standard
313 normal distribution truncated to the interval $[0,1]$ for $i = 1, \dots, n$. The first four variables
314 are non-zero (i.e. active in the response), while the rest of the variables are zero (i.e. are
315 noise variables). The outcome Y is then generated following one of the models and assump-
316 tions described below. We evaluate the performance of our method on three of its defining
317 characteristics: 1) the strong heredity property, 2) non-linearity of predictor effects and 3)
318 interactions. Simulation scenarios are designed specifically to test the performance of these
319 characteristics.

320 **1. Heredity simulation**

321 Scenario (a) Truth obeys strong heredity. In this situation, the true model for Y
322 contains main effect terms for all covariates involved in interactions.

$$Y = \sum_{j=1}^4 f_j(X_j) + \beta_E \cdot X_E + X_E \cdot f_3(X_3) + X_E \cdot f_4(X_4) + \varepsilon$$

323 Scenario (b) Truth obeys weak heredity. Here, in addition to the interaction, the
324 E variable has its own main effect but the covariates X_3 and X_4 do not.

$$Y = f_1(X_1) + f_2(X_2) + \beta_E \cdot X_E + X_E \cdot f_3(X_3) + X_E \cdot f_4(X_4) + \varepsilon$$

325 Scenario (c) Truth only has interactions. In this simulation, the covariates in-

326 involved in interactions do not have main effects as well.

$$Y = X_E \cdot f_3(X_3) + X_E \cdot f_4(X_4) + \varepsilon$$

327 **2. Non-linearity simulation scenario**

328 Truth is linear. `sail` is designed to model non-linearity; here we assess its per-
329 formance if the true model is completely linear.

$$Y = 5X_1 + 3(X_2 + 1) + 4X_3 + 6(X_4 - 2) + \beta_E \cdot X_E + X_E \cdot 4X_3 + X_E \cdot 6(X_4 - 2) + \varepsilon$$

330 **3. Interactions simulation scenario**

331 Truth only has main effects. `sail` is designed to capture interactions; here we
332 assess its performance when there are none in the true model.

$$Y = \sum_{j=1}^4 f_j(X_j) + \beta_E \cdot X_E + \varepsilon$$

333 The true component functions are the same as in [19, 21] and are given by $f_1(t) = 5t$,
334 $f_2(t) = 3(2t - 1)^2$, $f_3(t) = 4\sin(2\pi t)/(2 - \sin(2\pi t))$, $f_4(t) = 6(0.1\sin(2\pi t) + 0.2\cos(2\pi t) +$
335 $0.3\sin(2\pi t)^2 + 0.4\cos(2\pi t)^3 + 0.5\sin(2\pi t)^3)$. We set $\beta_E = 2$ and draw ε from a normal
336 distribution with variance chosen such that the signal-to-noise ratio is 2. Using this setup,
337 we generated 200 replications consisting of a training set of $n = 200$, a validation set of
338 $n = 200$ and a test set of $n = 800$. The training set was used to fit the model and the
339 validation set was used to select the optimal tuning parameter corresponding to the minimum
340 prediction mean squared error (MSE). Variable selection results including true positive rate,
341 false positive rate and number of active variables (the number of variables with a non-zero
342 coefficient estimate) were assessed on the training set, and MSE was assessed on the test
343 set.

344 4.3 Results

345 The prediction accuracy and variable selection results for each of the five simulation scenarios
346 are shown in Figure 3 and Table 2, respectively. We see that `sail`, `adaptive sail` and `sail`
347 `weak` have the best performance in terms of both MSE and yielding correct sparse models
348 when the truth follows a strong heredity (scenario 1a), as we would expect, since this is
349 exactly the scenario that our method is trying to target. Our method is also competitive
350 when only main effects are present (scenario 3) and performs just as well as methods that
351 only consider linear and non-linear main effects (`HierBasis`, `SPAM`), owing to the penalization
352 applied to the interaction parameter. Due to the heredity property being violated in scenario
353 1c), no method can identify the correct model with the exception of `GLinternet`. When only
354 linear effects and interactions are present (scenario 2), we see that `adaptive sail` has similar
355 MSE compared to the other linear interaction methods (`lassoBT` and `GLinternet`) with a
356 better TPR and FPR. Overall, our simulation study results suggests that `sail` outperforms
357 existing methods when the true model contains non-linear interactions, and is competitive
358 even when the truth only has either linear or additive main effects.

359 We visually inspected whether our method could correctly capture the shape of the associa-
360 tion between the predictors and the response for both main and interaction effects. To do
361 so, we plotted the true and predicted curves for scenario 1a) only. Figure 4 shows each of the
362 four main effects with the estimated curves from each of the 200 simulations along with the
363 true curve. We can see the effect of the penalty on the parameters, i.e., decreasing prediction
364 variance at the cost of increased bias. This is particularly well illustrated in the bottom right
365 panel where `sail` smooths out the very wiggly component function $f_4(x)$. Nevertheless, the
366 primary shapes are clearly being captured.

367 To visualize the estimated interaction effects, we ordered the 200 simulation runs by the Eu-
368 clidean distance between the estimated and true regression functions. Following Radchenko
369 et al. [28], we then identified the 25th, 50th, and 75th best simulations and plotted, in Fig-

370 ures 5 and 6, the interaction effects of X_E with $f_3(X_3)$ and $f_4(X_4)$, respectively. We see
371 that **sail** does a good job at capturing the true interaction surface for $X_E \cdot f_3(X_3)$. Again,
372 the smoothing and shrinkage effect is apparent when looking at the interaction surfaces for
373 $X_E \cdot f_4(X_4)$

374

5 Real data applications

375

5.1 Gene-environment interactions in the Nurse Family Partnership 376 program

377 It is well known that environmental exposures can have an important impact on academic
378 achievement. Indeed, early intervention in young children has been shown to positively im-
379 pact intellectual abilities [6]. More recent studies have shown that cognitive performance,
380 a trait that measures the ability to learn, reason and solve problems, is also strongly influ-
381 enced by genetic factors. Genome-wide association studies (GWAS) suggest that 20% of the
382 variance in educational attainment (years of education) may be accounted for by common
383 genetic variation [25, 30]. Unsurprisingly, there is significant overlap in the SNPs that predict
384 educational attainment and measures of cognitive function. An interesting query that arises
385 is how the environment interacts with these genetics variants to predict measures of cognitive
386 function. To address this question, we analyzed data from the Nurse Family Partnership
387 (NFP), a psychosocial intervention program that begins in pregnancy and targets maternal
388 health, parenting and mother-infant interactions [26]. The Stanford Binet IQ scores at 4
389 years of age were collected for 189 subjects (including 19 imputed using **mice** [5]) born to
390 women randomly assigned to control ($n = 100$) or nurse-visited intervention groups ($n =$
391 89). For each subject, we calculated a polygenic risk score (PRS) for educational attainment
392 at different p-value thresholds using weights from the GWAS conducted in Okbay et al. [25].
393 In this context, individuals with a higher PRS have a propensity for higher educational at-
394 tainment. The goal of this analysis was to determine if there was an interaction between

395 genetic predisposition to educational attainment (X) and maternal participation in the NFP
396 program (E) on child IQ at 4 years of age (Y). We applied the weak heredity **sail** with
397 cubic B-splines and $\alpha = 0.1$, and selected the optimal tuning parameter using 10-fold cross-
398 validation. Our method identified an interaction between the intervention and PRS which
399 included genetic variants at the 0.0001 level of significance. This interaction is shown in Fig-
400 ure 7. We see that the intervention has a much larger effect on IQ for lower PRS compared
401 to a higher PRS. In other words, perinatal home visitation by nurses can impact IQ scores
402 in children who are genetically predisposed to lower educational attainment. Similar results
403 were obtained for the other imputed datasets (Supplemental Section C).

404 We also compared **sail** with two other interaction selection methods, **lassoBT** and **GLinternet**
405 with default settings, on 200 bootstrap samples of the data. The average and standard de-
406 viation of the MSE and size of the active set ($|\hat{\mathcal{J}}|$) across the 200 bootstrap samples are
407 given in Table 3. We see that **sail** tends to select sparser models while maintaining similar
408 prediction performance compared to **lassoBT**. The **GLinternet** statistics are omitted here
409 since the algorithm did not converge for many of the 200 simulations.

410 **5.2 Study to Understand Prognoses Preferences Outcomes and Risks**
411 **of Treatment**

412 The Study to Understand Prognoses Preferences Outcomes and Risks of Treatment (SUP-
413 PORT) aimed at identifying which clinical variables influence medium-term (half-year) mor-
414 tality rate amongst seriously ill hospitalized patients and improving clinical decision mak-
415 ing [10]. With a relatively large sample size of 9,105 and detailed documentation of clinical
416 variables, the SUPPORT dataset allows detection of potential interactions using the strategy
417 implemented in **sail**. We applied **sail** to test for non-linear interactions between acute renal
418 failure or multiple organ system failure (ARF/MOSF), an important predictor for survival
419 rate, and 13 other variables that were deemed clinically relevant. These variables included

420 the number of comorbidities (excluding ARF/MOSF), age, sex, as well as multiple physio-
421 logical and blood biochemical indices. The response was whether a patient survived after
422 six months since hospitalization.

423 A total of 8,873 samples had complete data on all variables of interest. We randomly divided
424 these samples into equal sized training/validation/test splits and ran `lassoBT`, `GLinternet`,
425 and the weak heredity `sail` with cubic B-splines and $\alpha = 0.1$. A binomial distribution
426 family was specified for `GLinternet`, whereas `lassoBT` had the same default settings as the
427 simulation study since it did not support a specialized implementation for binary outcomes.
428 We again ran each method on the training data, determined the optimal tuning parameter
429 on the validation data based on the area under the receiver operating characteristic curve
430 (AUC), and assessed AUC on the test data. We repeated this process 200 times and report
431 the results in Table 3. We found that `sail` achieved similar prediction accuracy to `lassoBT`
432 and `GLinternet`. However, the predictive performance of `lassoBT` and `GLinternet` relied
433 on models which included many more variables. In Figure 8, we visualize the two strongest
434 interaction effects associated with the number of comorbidities and age, respectively. For
435 those having undergone ARF/MOSF, an increased number of comorbidities decreases their
436 chance of survival, while there seems to be no such relationship for non-ARF/MOSF patients.
437 The interaction between ARF/MOSF and age shows the risk incurred by ARF/MOSF is most
438 distinguishing among patients between the ages of 70 and 80.

439

6 Discussion

440 In this article we have introduced the sparse additive interaction learning model `sail` for
441 detecting non-linear interactions with a key environmental or exposure variable in high-
442 dimensional settings. Using a simple reparametrization, we are able to achieve either the
443 weak or strong heredity property without using a complex penalty function. We developed
444 a blockwise coordinate descent algorithm to solve the `sail` objective function for the least-

squares loss. We further studied the asymptotic properties of our method and showed that under certain conditions, it possesses the oracle property. All our algorithms have been implemented in a computationally efficient, well-documented and freely available R package on CRAN. Furthermore, our method is flexible enough to handle any type of basis expansion including the identity map, which allows for linear interactions. Our implementation allows the user to selectively apply the basis expansions to the predictors, allowing for example, a combination of continuous and categorical predictors. An extensive simulation study shows that **sail**, **adaptive sail** and **sail weak** outperform existing penalized regression methods in terms of prediction accuracy, sensitivity and specificity when there are non-linear main effects only, as well as interactions with an exposure variable. We then demonstrated the utility of our method to identify non-linear interactions in both biological and epidemiological data. In the NFP program, we showed that individuals who are genetically predisposed to lower educational attainment are those who stand to benefit the most from the intervention. Analysis of the SUPPORT data revealed that those having undergone ARF/MOSF, an increased number of comorbidities decreased their chances of survival, while there seemed to be no such relationship for non-ARF/MOSF patients. In a bootstrap analysis of both datasets, we observed that **sail**tended to select sparser models while maintaining similar prediction performance compared to other interaction selection methods.

Our method however does have its limitations. **sail** can currently only handle $X_E \cdot f(X)$ or $f(X_E) \cdot X$ and does not allow for $f(X, X_E)$, i.e., only one of the variables in the interaction can have a non-linear effect and we do not consider the tensor product. The reparametrization leads to a non-convex optimization problem which makes convergence rates difficult to assess, though we did not experience any major convergence issues in our simulations and real data analysis. The memory footprint can also be an issue depending on the degree of the basis expansion and the number of variables. Furthermore, the functional form of the covariate effects is treated as known in our method. Being able to automatically select for example, linear vs. nonlinear components, is currently an active area of research in

472 main effects models [16]. To our knowledge, our proposal is the first to allow for non-linear
473 interactions with a key exposure variable following the weak or strong heredity property
474 in high-dimensional settings. We also provide a first software implementation for these
475 models.

476 **Acknowledgments**

477 SRB was supported by the Ludmer Centre for Neuroinformatics and Mental Health and
478 the Canadian Institutes for Health Research PJT 148620. This research was enabled in
479 part by support provided by Calcul Québec (www.calculquebec.ca) and Compute Canada
480 (www.computecanada.ca). The funders had no role in study design, data collection and
481 analysis, decision to publish, or preparation of the manuscript.

482 **Description of Supplementary Materials**

- 483 A **Proofs** - Regularity conditions and proofs for Lemma 1, Theorem 1 and Theorem 2
- 484 B **Algorithm Details** - Detailed description of the algorithms used to solve the strong
485 and weak heredity `sail` objective function.
- 486 C **Additional Results on PRS for Educational Attainment** - Estimated coefficient
487 estimates and visualization of interaction effects for the Nurse Family Partnership data
488 for the 5 imputed datasets.

489 **References**

- 490 [1] F. Bach, R. Jenatton, J. Mairal, G. Obozinski, et al. Structured sparsity through convex
491 optimization. *Statistical Science*, 27(4):450–468, 2012.
- 492 [2] S. R. Bhatnagar, Y. Yang, B. Khundrakpam, A. C. Evans, M. Blanchette, L. Bouchard,
493 and C. M. Greenwood. An analytic approach for interpretable predictive models in high-
494 dimensional data in the presence of interactions with exposures. *Genetic epidemiology*,
495 42(3):233–249, 2018.
- 496 [3] J. Bien, J. Taylor, R. Tibshirani, et al. A lasso for hierarchical interactions. *The Annals
497 of Statistics*, 41(3):1111–1141, 2013.
- 498 [4] P. Bühlmann and S. Van De Geer. *Statistics for high-dimensional data: methods, theory
499 and applications*. Springer Science & Business Media, 2011.
- 500 [5] S. v. Buuren and K. Groothuis-Oudshoorn. mice: Multivariate imputation by chained
501 equations in r. *Journal of statistical software*, pages 1–68, 2010.
- 502 [6] F. A. Campbell and C. T. Ramey. Effects of early intervention on intellectual and
503 academic achievement: a follow-up study of children from low-income families. *Child
504 development*, 65(2):684–698, 1994.

- 505 [7] H. Chipman. Bayesian variable selection with related predictors. *Canadian Journal of Statistics*, 24(1):17–36, 1996.
506
507 [8] N. H. Choi, W. Li, and J. Zhu. Variable selection with the strong heredity constraint and
508 its oracle property. *Journal of the American Statistical Association*, 105(489):354–364,
509 2010.
510 [9] A. Chouldechova and T. Hastie. Generalized additive model selection. *arXiv preprint arXiv:1506.03850*, 2015.
511
512 [10] A. F. Connors, N. V. Dawson, N. A. Desbiens, W. J. Fulkerson, L. Goldman, W. A.
513 Knaus, J. Lynn, R. K. Oye, M. Bergner, A. Damiano, et al. A controlled trial to
514 improve care for seriously ill hospitalized patients: The study to understand prognoses
515 and preferences for outcomes and risks of treatments (support). *Jama*, 274(20):1591–
516 1598, 1995.
517 [11] D. R. Cox. Interaction. *International Statistical Review/Revue Internationale de Statistique*, pages 1–24, 1984.
518
519 [12] J. Fan, F. Han, and H. Liu. Challenges of big data analysis. *National science review*,
520 1(2):293–314, 2014.
521 [13] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle
522 properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.
523 [14] J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear
524 models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.
525 [15] N. Hao, Y. Feng, and H. H. Zhang. Model selection for high-dimensional quadratic
526 regression via regularization. *Journal of the American Statistical Association*, pages
527 1–11, 2018.
528 [16] A. Haris, A. Shojaie, and N. Simon. Nonparametric regression with adaptive truncation
529 via a convex hierarchical penalty. *arXiv preprint arXiv:1611.09972*, 2016.
530 [17] A. Haris, D. Witten, and N. Simon. Convex modeling of interactions with strong
531 heredity. *Journal of Computational and Graphical Statistics*, 25(4):981–1004, 2016.
532 [18] T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC Press, 2015.
533
534 [19] J. Huang, J. L. Horowitz, and F. Wei. Variable selection in nonparametric additive
535 models. *Annals of statistics*, 38(4):2282, 2010.
536 [20] M. Lim and T. Hastie. Learning interactions via hierarchical group-lasso regularization.
537 *Journal of Computational and Graphical Statistics*, 24(3):627–654, 2015.
538 [21] Y. Lin, H. H. Zhang, et al. Component selection and smoothing in multivariate non-
539 parametric regression. *The Annals of Statistics*, 34(5):2272–2297, 2006.
540 [22] P. McCullagh and J. A. Nelder. *Generalized linear models*, volume 37. CRC press, 1989.
541 [23] Y. Nardi, A. Rinaldo, et al. On the asymptotic properties of the group lasso estimator
542 for linear models. *Electronic Journal of Statistics*, 2:605–633, 2008.
543 [24] K. Ning, B. Chen, F. Sun, Z. Hobel, L. Zhao, W. Matloff, A. W. Toga, A. D. N.
544 Initiative, et al. Classifying alzheimer’s disease with brain imaging and genetic data
545 using a neural network framework. *Neurobiology of aging*, 2018.
546 [25] A. Okbay, J. P. Beauchamp, M. A. Fontana, J. J. Lee, T. H. Pers, C. A. Rietveld,
547 P. Turley, G.-B. Chen, V. Emilsson, S. F. W. Meddens, et al. Genome-wide association
548 study identifies 74 loci associated with educational attainment. *Nature*, 533(7604):539,
549 2016.

- [26] D. Olds, C. R. Henderson Jr, R. Cole, J. Eckenrode, H. Kitzman, D. Luckey, L. Pettitt, K. Sidora, P. Morris, and J. Powers. Long-term effects of nurse home visitation on children's criminal and antisocial behavior: 15-year follow-up of a randomized controlled trial. *Jama*, 280(14):1238–1244, 1998.
- [27] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017.
- [28] P. Radchenko and G. M. James. Variable selection using adaptive nonlinear interaction structures in high dimensions. *Journal of the American Statistical Association*, 105(492):1541–1553, 2010.
- [29] P. Ravikumar, J. Lafferty, H. Liu, and L. Wasserman. Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5):1009–1030, 2009.
- [30] C. A. Rietveld, S. E. Medland, J. Derringer, J. Yang, T. Esko, N. W. Martin, H.-J. Westra, K. Shakhbazov, A. Abdellaoui, A. Agrawal, et al. Gwas of 126,559 individuals identifies genetic variants associated with educational attainment. *science*, 340(6139):1467–1471, 2013.
- [31] E. E. Schadt. Molecular networks as sensors and drivers of common human diseases. *Nature*, 461(7261):218–223, 2009.
- [32] R. D. Shah. Modelling interactions in high-dimensional data with backtracking. *Journal of Machine Learning Research*, 17(207):1–31, 2016.
- [33] Y. She and H. Jiang. Group regularized estimation under structural hierarchy. *arXiv preprint arXiv:1411.4691*, 2014.
- [34] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.
- [35] R. Tibshirani and J. Friedman. A pliable lasso. *arXiv preprint arXiv:1712.00484*, 2017.
- [36] N. J. Timpson, C. M. Greenwood, N. Soranzo, D. J. Lawson, and J. B. Richards. Genetic architecture: the shape of the genetic contribution to human traits and disease. *Nature Reviews Genetics*, 19(2):110, 2018.
- [37] H. Wang, G. Li, and C.-L. Tsai. Regression coefficient and autoregressive order shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(1):63–78, 2007.
- [38] Y. Xie. *Dynamic Documents with R and knitr*, volume 29. CRC Press, 2015.
- [39] Y. Yang and H. Zou. A fast unified algorithm for solving group-lasso penalize learning problems. *Statistics and Computing*, 25(6):1129–1141, 2015.
- [40] P. Zhao, G. Rocha, and B. Yu. The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, pages 3468–3497, 2009.
- [41] H. Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.
- [42] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.
- [43] H. Zou and H. H. Zhang. On the adaptive elastic-net with a diverging number of parameters. *Annals of statistics*, 37(4):1733, 2009.

Table 1: Reparametrization for strong and weak heredity principle for **sail** model

Type	Feature	Reparametrization
Strong heredity	$\hat{\boldsymbol{\tau}}_j \neq 0$ only if $\hat{\boldsymbol{\theta}}_j \neq 0$ and $\hat{\beta}_E \neq 0$	$\boldsymbol{\tau}_j = \gamma_{jE} \beta_E \boldsymbol{\theta}_j$
Weak heredity	$\hat{\boldsymbol{\tau}}_j \neq 0$ only if $\hat{\boldsymbol{\theta}}_j \neq 0$ or $\hat{\beta}_E \neq 0$	$\boldsymbol{\tau}_j = \gamma_{jE} (\beta_E \cdot \mathbf{1}_{m_j} + \boldsymbol{\theta}_j)$

Table 2: Mean (standard deviation) of the number of selected variables ($|\widehat{\mathcal{J}}|$), true positive rate (TPR) and false positive rate (FPR) as a percentage from 200 simulations for each of the five scenarios. $|\mathcal{J}|$ is the number of truly associated variables.

Linear		Linear		Non-linear			Non-linear										
Main Effects		Interactions		Main Effects			Interactions										
lasso	adaptive	lassoBT	GLinternet	HierBasis	SPAM	gamsel	sail	adaptive	sail								
lasso																	
1a) Strong heredity ($\mathcal{J} = 7$)																	
$ \widehat{\mathcal{J}} $	30 (14)	8 (4)	37 (17)	41 (21)	152 (28)	38 (17)	47 (19)	37 (15)	8 (5)	34 (13)							
TPR	54.9 (7.4)	49.7 (10.4)	62.0 (10.4)	66.7 (12.8)	66.2 (7.6)	60.9 (9.0)	57.1 (6.5)	90.6 (7.7)	69.7 (28.8)	86.4 (10.1)							
FPR	1.3 (0.7)	0.2 (0.2)	1.6 (0.8)	1.8 (1.0)	7.4 (1.4)	1.7 (0.8)	2.2 (0.9)	1.5 (0.7)	1.1 (9.7)	1.4 (0.6)							
1b) Weak heredity ($\mathcal{J} = 5$)																	
$ \widehat{\mathcal{J}} $	19 (12)	4 (2)	20 (13)	37 (22)	23 (22)	28 (15)	22 (15)	16 (9)	7 (6)	17 (11)							
TPR	41.0 (4.5)	40.2 (1.9)	41.0 (4.5)	65.1 (15.2)	42.6 (6.7)	54.8 (8.8)	43.8 (7.9)	47.8 (10.4)	46.9 (11.2)	51.0 (12.8)							
FPR	0.8 (0.6)	0.1 (0.1)	0.9 (0.7)	1.7 (1.1)	1.1 (1.1)	1.3 (0.7)	1.0 (0.8)	0.7 (0.4)	0.2 (0.3)	0.7 (0.5)							
1c) Interactions Only ($\mathcal{J} = 2$)																	
$ \widehat{\mathcal{J}} $	14 (13)	3 (2)	15 (14)	42 (21)	14 (14)	14 (12)	14 (13)	6 (7)	3 (5)	6 (7)							
TPR	0.0 (0.0)	0.0 (0.0)	0.2 (3.5)	82.6 (26.3)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.7 (5.9)	0.0 (0.0)							
FPR	0.7 (0.6)	0.6 (6.9)	0.8 (0.7)	2.0 (1.1)	0.7 (0.7)	0.7 (0.6)	0.7 (0.6)	0.3 (0.4)	0.2 (0.2)	0.3 (0.4)							
2) Linear Effects ($\mathcal{J} = 7$)																	
$ \widehat{\mathcal{J}} $	36 (16)	8 (3)	48 (17)	47 (20)	36 (17)	42 (18)	36 (16)	30 (12)	12 (4)	19 (14)							
TPR	69.9 (4.7)	67.4 (6.7)	72.7 (6.6)	92.6 (9.1)	69.9 (4.6)	64.6 (8.4)	69.9 (4.7)	87.4 (14.1)	88.6 (13.5)	64.3 (13.6)							
FPR	1.6 (0.8)	0.2 (0.1)	2.1 (0.8)	2.1 (1.0)	1.6 (0.9)	1.9 (0.9)	1.6 (0.8)	1.2 (0.6)	0.3 (0.2)	0.7 (0.7)							
3) Main Effects Only ($\mathcal{J} = 5$)																	
$ \widehat{\mathcal{J}} $	30 (15)	7 (4)	31 (15)	35 (18)	160 (17)	42 (18)	54 (20)	40 (16)	8 (5)	40 (16)							
TPR	76.6 (10.0)	67.4 (13.6)	77.0 (10.1)	78.3 (8.8)	97.0 (7.5)	92.3 (10.9)	82.4 (10.0)	89.3 (13.0)	78.0 (14.8)	89.1 (13.0)							
FPR	1.3 (0.7)	0.2 (0.2)	1.4 (0.8)	1.6 (0.9)	7.8 (0.8)	1.9 (0.9)	2.5 (1.0)	1.8 (0.8)	0.2 (0.2)	1.8 (0.8)							

Table 3: Comparison of analytic methods for selecting interactions using the Nurse Family Partnership program and the SUPPORT datasets. Averages (standard deviations in parentheses) are based on 200 bootstrap samples.

Method	Nurse Family Partnership		SUPPORT	
	Mean Squared Error	$ \hat{\mathcal{J}} $	AUC	$ \hat{\mathcal{H}} $
<code>sail</code>	3.5 (0.6)	4 (3)	0.66 (0.01)	25 (3)
<code>lassoBT</code>	3.53 (0.477)	11 (6)	0.65 (0.009)	49 (14)
<code>GLinternet</code> ^a	—	—	0.65 (0.009)	58 (7)

^a `GLinternet` results not reported for NFP data since the algorithm did not converge in many of the bootstrap samples.

^b $|\hat{\mathcal{J}}|$ is the number of variables selected by the method.

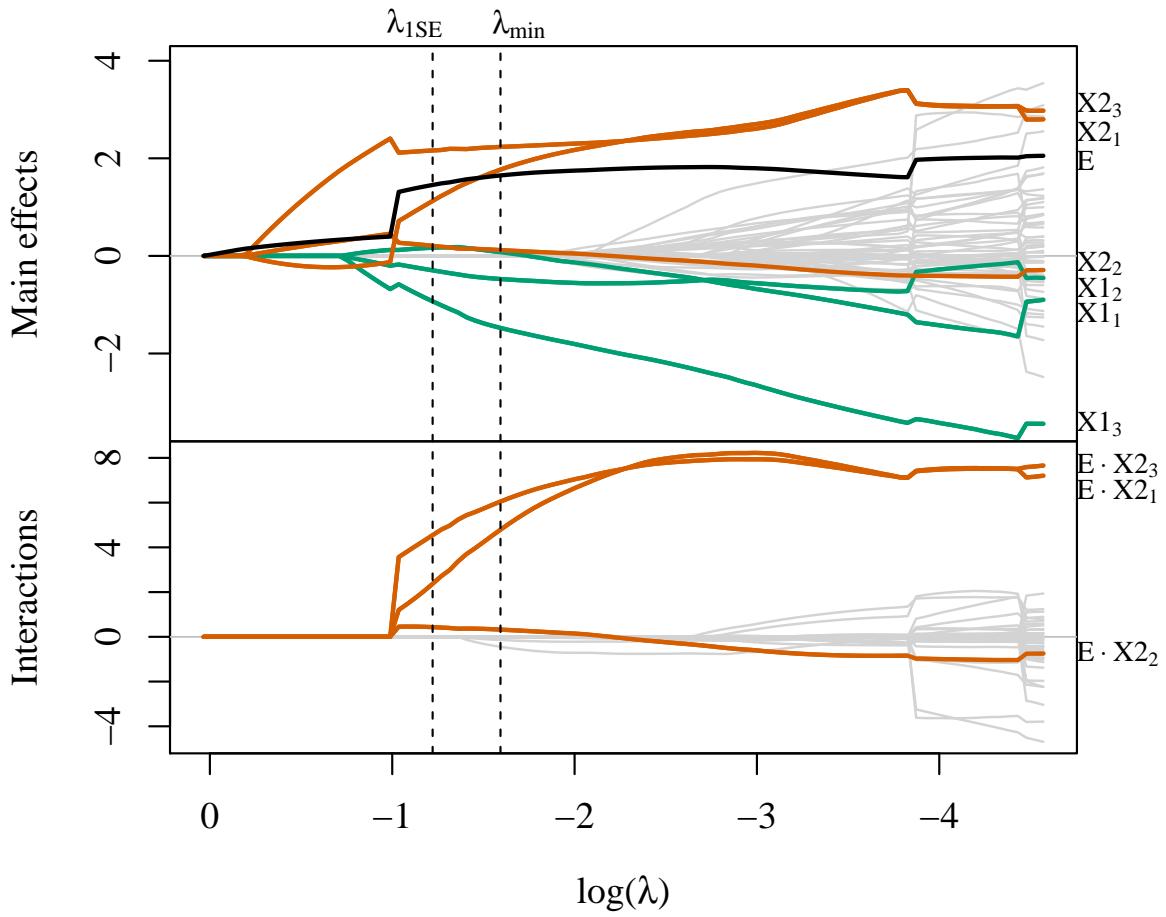


Figure 1: Toy example solution path for main effects (top) and interactions (bottom). $\{X_{11}, X_{12}, X_{13}\}$ and $\{X_{21}, X_{22}, X_{23}\}$ are the three basis coefficients for X_1 and X_2 , respectively. λ_{1SE} is the largest value of penalization for which the CV error is within one standard error of the minimizing value λ_{min} .

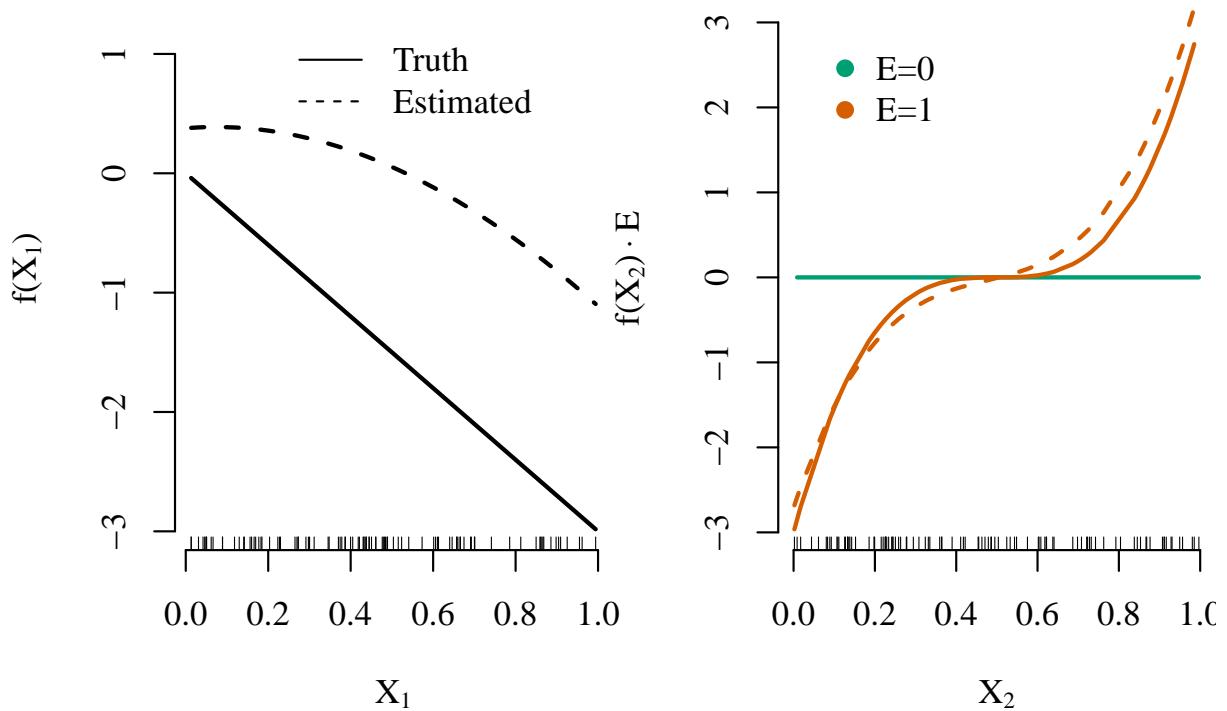


Figure 2: Estimated smooth functions for X_1 and the $X_2 \cdot E$ interaction by the **sail** method based on λ_{min} .

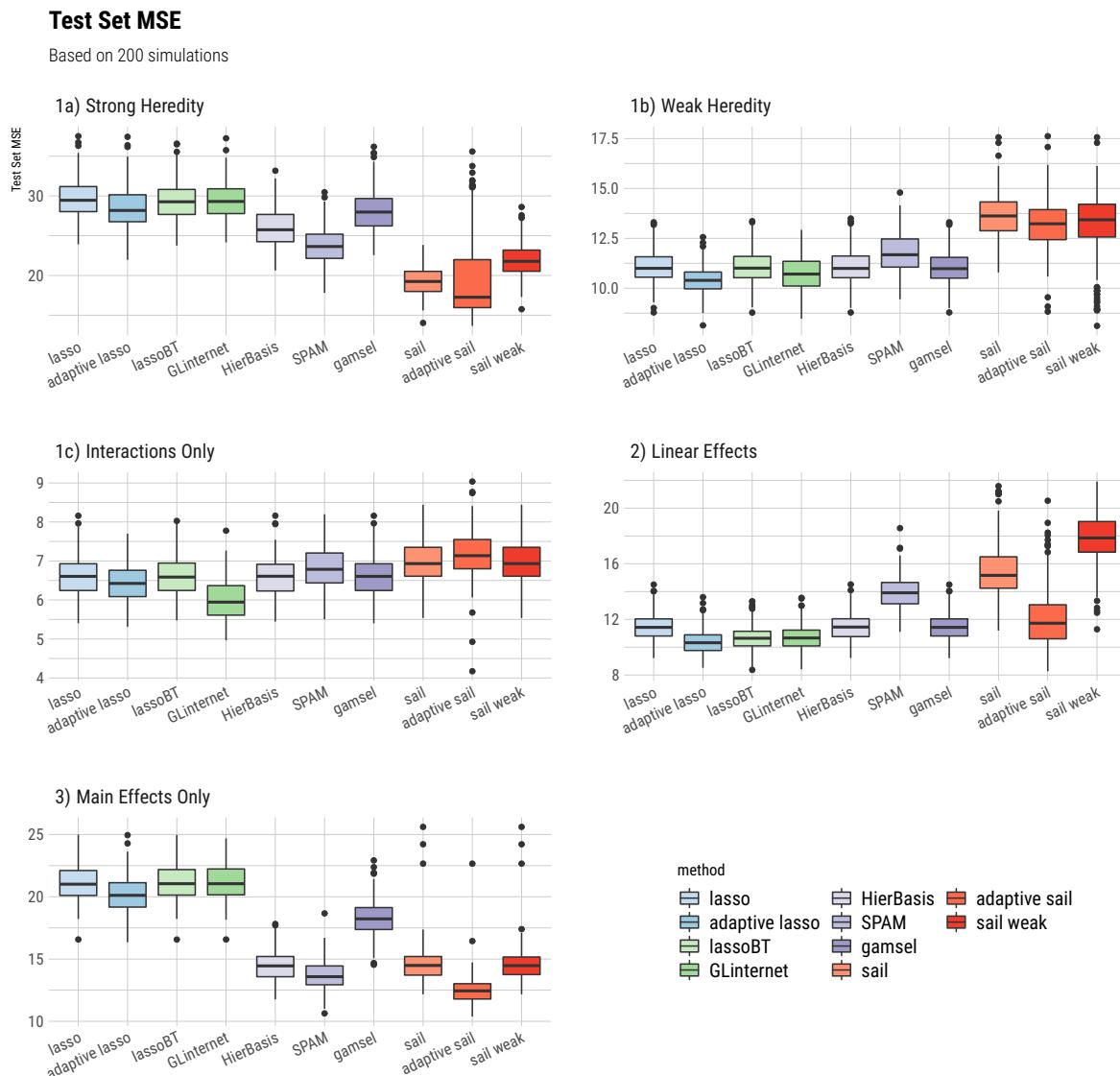


Figure 3: Boxplots of the test set mean squared error from 200 simulations for each of the five simulation scenarios.

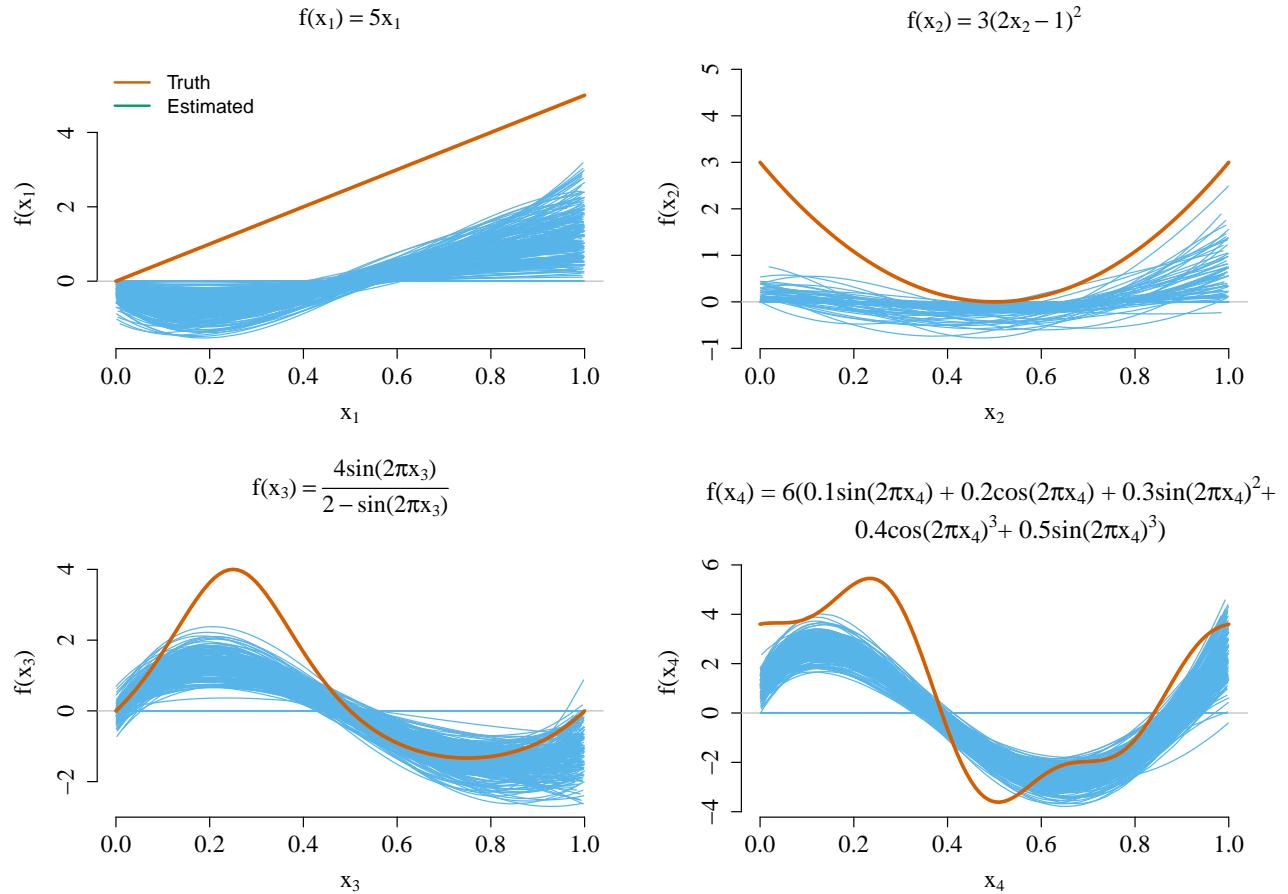


Figure 4: True and estimated main effect component functions for scenario 1a). The estimated curves represent the results from each one of the 200 simulations conducted.

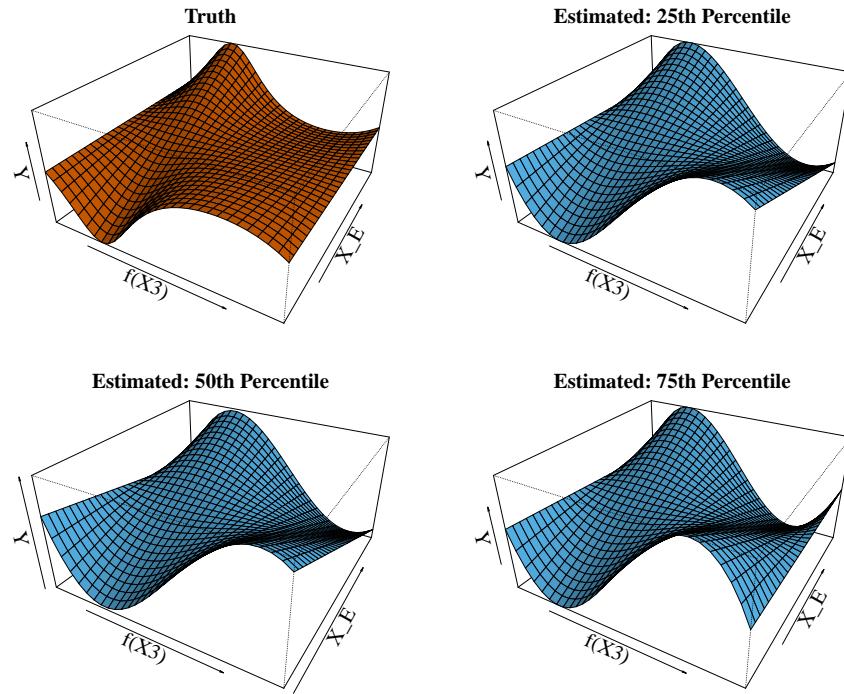


Figure 5: True and estimated interaction effects for $X_E \cdot f_3(X_3)$ in simulation scenario 1a).

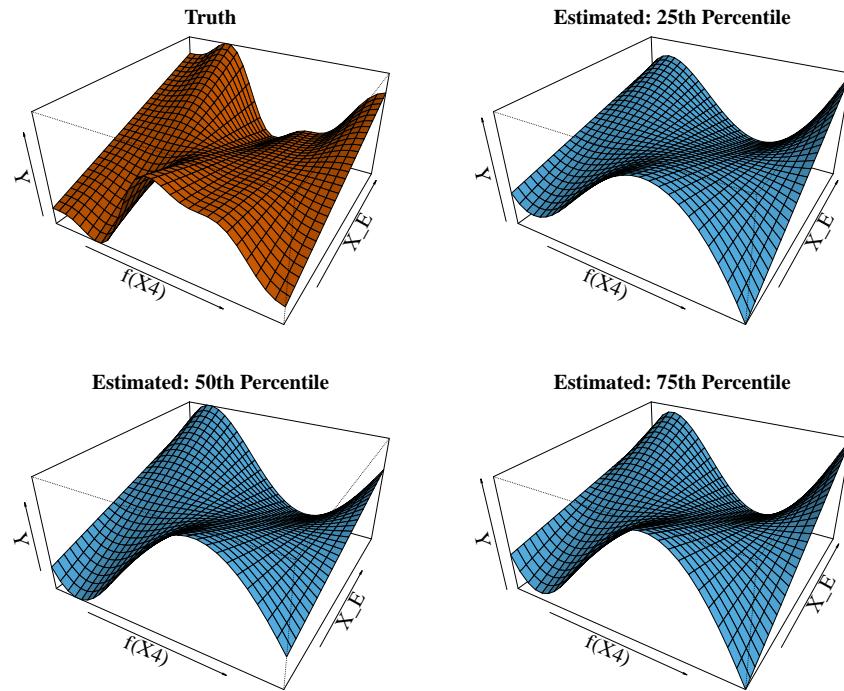


Figure 6: True and estimated interaction effects for $X_E \cdot f_4(X_4)$ in simulation scenario 1a).

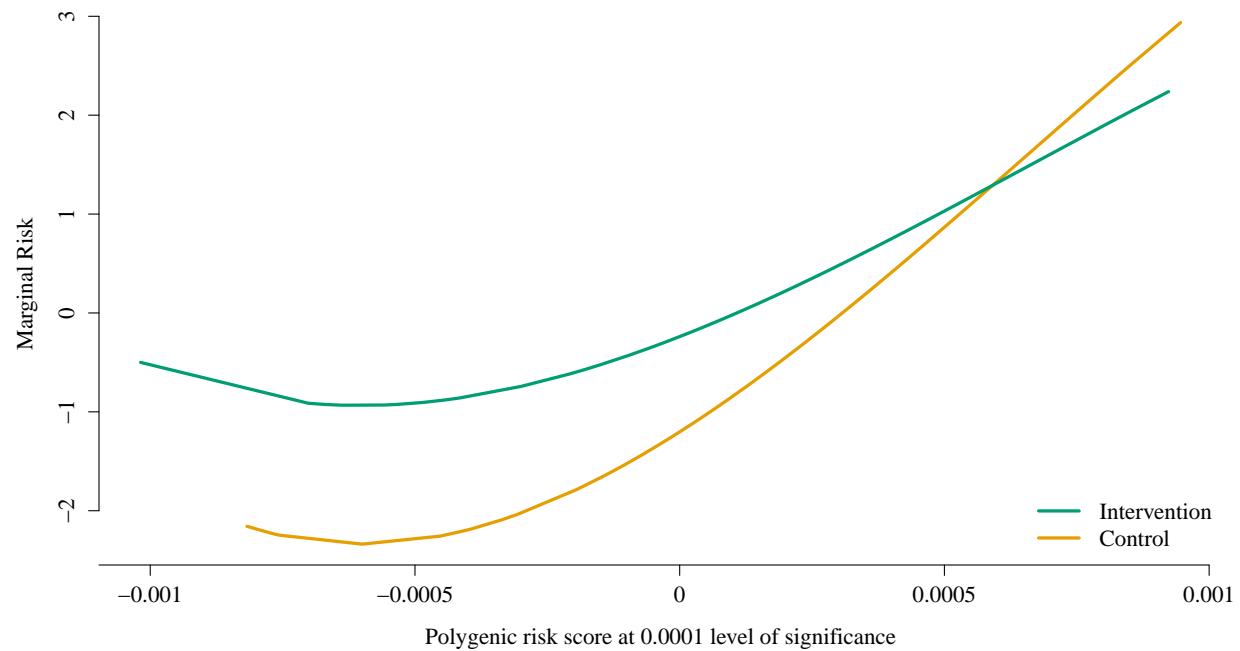


Figure 7: Estimated interaction effect identified by the weak heredity `sail` using cubic B-splines and $\alpha = 0.1$ for the Nurse Family Partnership data. The selected model, chosen via 10-fold cross-validation, contained three variables: the main effects for the intervention and the PRS for educational attainment using genetic variants significant at the 0.0001 level, as well as their interaction.

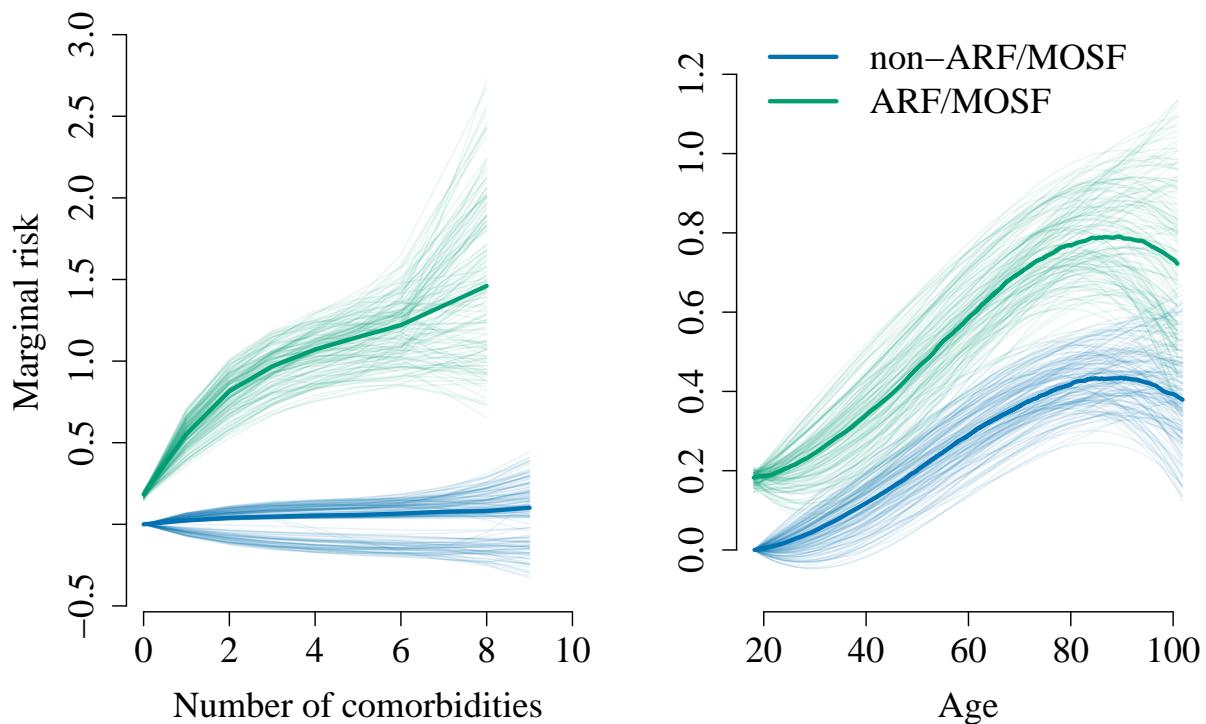


Figure 8: Illustration of estimated interaction effects identified by `sail` for the SUPPORT data. Median prediction curves in dark colors based on 200 train/validate/test splits represent the estimated marginal interaction effects. Coefficients estimated in each of the 200 train/validate/test splits were used to generate prediction curves representing a 90% confidence interval colored in corresponding light colors.

592 A Sparse Additive Model for High-Dimensional Interactions with 593 an Exposure Variable

594 Supplementary Materials

595 Sahir R Bhatnagar^{1,2}, Tianyuan Lu^{3,4}, Amanda Lovato⁵, David L Olds⁶, Michael S Kobor⁷,
596 Michael J Meaney⁸, Kieran O'Donnell⁹, Yi Yang¹⁰, and Celia MT Greenwood^{1,3,5}

⁵⁹⁷ ¹Department of Epidemiology, Biostatistics and Occupational Health, McGill University, ²Department
⁵⁹⁸ of Diagnostic Radiology, McGill University, ³Quantitative Life Sciences, McGill University, ⁴Lady
⁵⁹⁹ Davis Institute, Jewish General Hospital, Montréal, QC, ⁵Statistics Canada, Ottawa, ON, ⁶Department
⁶⁰⁰ of Pediatrics, University of Colorado School of Medicine, Denver, ⁷Department of Medical Genetics,
⁶⁰¹ University of British Columbia, BC, ⁸Singapore Institute for Clinical Sciences, Singapore; McGill
⁶⁰² University, ⁹Department of Psychiatry, McGill University, ¹⁰Department of Mathematics and Statis-
⁶⁰³ tics, McGill University, ¹¹Departments of Oncology and Human Genetics, McGill University

604 A Proofs

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

619 **A.2 Lemma (1) proof**

620 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
621 dimension of the design matrix and C is some constant. Under the regularity assumptions,
622 we show that there exists a local minimizer $\widehat{\Phi}_n$ of $Q_n(\Phi)$ such that $\|\widehat{\Phi}_n - \Phi^*\|_2 = O_p(\frac{1}{\sqrt{n}})$.
623 For this proof, we adopt the approaches outlined in [8, 13, 23, 37] and extend it to our
624 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
625 $\delta = (\mathbf{u}_1^\top, \mathbf{u}_2^\top, \dots, \mathbf{u}_{p+1}^\top, \mathbf{u}_{p+2}^\top, \dots, \mathbf{u}_{2p+1}^\top)^\top \in \mathbb{R}^d$, where d is the dimension of the design
626 matrix and C is some constant. The objective function is given by

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

627 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

$$\begin{aligned}
D_n(\boldsymbol{\delta}) &= -L_n(\boldsymbol{\Phi}^* + \eta_n \boldsymbol{\delta}) + L_n(\boldsymbol{\Phi}^*) + n \sum_{m=1}^{2p+1} \lambda_m (\|\boldsymbol{\theta}_m^* + \eta_n \mathbf{u}_m\|_2 - \|\boldsymbol{\theta}_m^*\|_2) \\
&\stackrel{(a)}{\geq} -L_n(\boldsymbol{\Phi}^* + \eta_n \boldsymbol{\delta}) + L_n(\boldsymbol{\Phi}^*) + n \sum_{m \in \mathcal{A}_1} \lambda_m^\theta (\|\boldsymbol{\theta}_m^* + \eta_n \mathbf{u}_m\|_2 - \|\boldsymbol{\theta}_m^*\|_2) \\
&\quad + n \sum_{m \in \mathcal{A}_2} \lambda_m^\theta (\|\boldsymbol{\theta}_m^* + \eta_n \mathbf{u}_m\|_2 - \|\boldsymbol{\theta}_m^*\|_2) \\
&\stackrel{(b)}{\geq} -L_n(\boldsymbol{\Phi}^* + \eta_n \boldsymbol{\delta}) + L_n(\boldsymbol{\Phi}^*) - n\eta_n \sum_{m \in \mathcal{A}_1} \lambda_m \|\mathbf{u}_m\|_2 - n\eta_n \sum_{m \in \mathcal{A}_2} \lambda_m \|\mathbf{u}_m\|_2 \\
&\stackrel{(c)}{\geq} -L_n(\boldsymbol{\Phi}^* + \eta_n \boldsymbol{\delta}) + L_n(\boldsymbol{\Phi}^*) - nn\eta_n^2 \sum_{m \in \mathcal{A}_1} \|\mathbf{u}_m\|_2 - nn\eta_n^2 \sum_{m \in \mathcal{A}_2} \|\mathbf{u}_m\|_2 \\
&\geq -L_n(\boldsymbol{\Phi}^* + \eta_n \boldsymbol{\delta}) + L_n(\boldsymbol{\Phi}^*) - nn\eta_n^2 (|\mathcal{A}_1| + |\mathcal{A}_2|)C \\
&\stackrel{(d)}{=} -[\nabla L_n(\boldsymbol{\Phi}^*)]^\top (\eta_n \boldsymbol{\delta}) - \frac{1}{2} (\eta_n \boldsymbol{\delta})^\top [\nabla^2 L_n(\boldsymbol{\Phi}^*)] (\eta_n \boldsymbol{\delta}) (1 + o(1)) \\
&\quad - nn\eta_n^2 (|\mathcal{A}_1| + |\mathcal{A}_2|)C
\end{aligned} \tag{14}$$

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 (14) into three parts:

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

Then

$$\begin{aligned} D_1 &= -\eta_n [\nabla L_n(\Phi^*)]^\top \boldsymbol{\delta} \\ &= -\sqrt{n}\eta_n \left(\frac{1}{\sqrt{n}} \nabla L_n(\Phi^*) \right)^\top \boldsymbol{\delta} \\ &= -\sqrt{n}\eta_n \left(\sqrt{n} \frac{1}{n} \sum_{i=1}^n \nabla \log f(\mathbf{V}_i, \Phi)|_{\Phi=\Phi^*} \right)^\top \boldsymbol{\delta} \\ &= -\sqrt{n}\eta_n \left(\sqrt{n} \left[\frac{1}{n} \sum_{i=1}^n \nabla \log f(\mathbf{V}_i, \Phi)|_{\Phi=\Phi^*} - \mathbf{0} \right] \right)^\top \boldsymbol{\delta} \\ &= -\sqrt{n}\eta_n \left(\sqrt{n} \left[\frac{1}{n} \sum_{i=1}^n \nabla \log f(\mathbf{V}_i, \Phi)|_{\Phi=\Phi^*} - E_{\Phi^*} \nabla L(\Phi^*) \right] \right)^\top \boldsymbol{\delta} \\ &= -\sqrt{n}\eta_n O_P(1) \boldsymbol{\delta} \\ &= -O_P(n\eta_n^2) \boldsymbol{\delta} \end{aligned} \tag{15}$$

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

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

$$\begin{aligned}
D_2 &= \frac{1}{2}n\eta_n^2 \left\{ \boldsymbol{\delta}^\top \left[-\frac{1}{n}\nabla^2 L_n(\Phi^*) \right] \boldsymbol{\delta} \right\} (1 + o_p(1)) \\
&= \frac{1}{2}n\eta_n^2 \{ \boldsymbol{\delta}^\top [\mathbf{I}(\Phi^*)] \boldsymbol{\delta} \} (1 + o_p(1)) \text{ by the weak law of large numbers.} \\
&= O_p(n\eta_n^2 \|\boldsymbol{\delta}\|_2^2)
\end{aligned} \tag{16}$$

629 Combining (15) and (16) with (14) gives:

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

630 We can see that the first term D_1 is linear in $\boldsymbol{\delta}$ and the second term D_2 is quadratic in $\boldsymbol{\delta}$.

631 We can conclude that for a large enough constant $C = \|\boldsymbol{\delta}\|_2$, D_2 dominates D_1 and D_3 . Note

632 that this is a positive term since $I(\Phi)$ is positive definite at $\Phi = \Phi^*$ by regularity condition

633 (C2). Therefore, for each $\varepsilon > 0$, there exists a large enough constant C such that, for large

634 enough n

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

635 This implies with probability at least $1 - \varepsilon$ that the empirical likelihood Q_n has a local

636 minimizer in the ball $\{\Phi^* + \eta_n \boldsymbol{\delta} : \|\boldsymbol{\delta}\|_2 \leq C\}$ (since Q_n is bounded and $\{\Phi^* + \alpha_n \boldsymbol{\delta} : \|\boldsymbol{\delta}\|_2 \leq C\}$

637 is closed). In other words, there exists a local solution $\widehat{\Phi}_n$ such that $\|\widehat{\Phi}_n - \Phi^*\| \leq \eta_n \|\boldsymbol{\delta}\|_2 \leq$

638 $\eta_n C = O_P(\eta_n) = O_P(\frac{1}{\sqrt{n}} + a_n) = O_p(\frac{1}{\sqrt{n}})$, since $a_n = o(\frac{1}{\sqrt{n}})$. Hence, $\|\widehat{\Phi}_n - \Phi^*\|_2 = O_P\left(\frac{1}{\sqrt{n}}\right)$.

639 \square

640 A.3 Theorem 1 proof

641 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
642 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 $P\left(\widehat{\Phi}_{\mathcal{A}_1^c} = \mathbf{0}\right) \rightarrow$
643 1, i.e., the \sqrt{n} -consistent estimate $\widehat{\Phi}$ has oracle property $\widehat{\phi}_m = \mathbf{0}$ if $\phi_m^* = \mathbf{0}$. Denote

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

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

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

644 where $\widetilde{\Phi}$ lies between $\widehat{\Phi}_n$ and Φ^* . By the regularity conditions and Lemma (1) that
645 $\left\| \widehat{\Phi}_n - \Phi^* \right\|_2 = O_P\left(\frac{1}{\sqrt{n}}\right)$, the first term is of the order $O_p(\sqrt{n})$

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

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

$$\frac{\partial Q_n\left(\widehat{\Phi}_n\right)}{\partial \Phi_m} = \sqrt{n} \left\{ O_p(1) + \sqrt{n} \lambda_m \frac{\widehat{\phi}_{mk}}{\|\widehat{\phi}_m\|_2} \right\}. \quad (17)$$

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 (17) with probability tending to one. This means that (17) 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 $\widehat{\phi}_m = (\widehat{\phi}_{m1}, \dots, \widehat{\phi}_{mp_m})$ must be in a position where $\widehat{\phi}_m$ is not differentiable. Hence $\widehat{\phi}_m = \mathbf{0}$ for all $m \in \mathcal{A}_1^c$. Hence $P\left(\widehat{\Phi}_{\mathcal{A}_1^c} = \mathbf{0}\right) \rightarrow 1$. This completes the proof.

Next, we prove that for the interactions $P\left(\widehat{\Phi}_{\mathcal{A}_2^c} = \mathbf{0}\right) \rightarrow 1$. 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 $P\left(\widehat{\Phi}_{\mathcal{A}_2^c} = \mathbf{0}\right) \rightarrow 1$ by a similar reasoning, which further implies that $P(\widehat{\gamma}_{jE} = 0) \rightarrow 0$. 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 $\widehat{\theta}_j = \mathbf{0}$ implies $\widehat{\gamma}_{jE} = 0$, since if $\widehat{\gamma}_{jE} \neq 0$, the value of the loss function does not change but the value of the penalty function will increase. Because we already prove $P\left(\widehat{\Phi}_{\mathcal{A}_1^c} = \mathbf{0}\right) \rightarrow 1$, therefore we get $P\left(\widehat{\Phi}_{\mathcal{A}_2^c} = \mathbf{0}\right) \rightarrow 1$ as well for this case.

□

660 **A.4 Theorem 2 proof**

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

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

664 that is

$$\left. \frac{\partial Q_n(\Phi_{\mathcal{A}})}{\partial \Phi_m} \right|_{\Phi_{\mathcal{A}}=\widehat{\Phi}_{\mathcal{A}}} = 0, \quad \forall m \in \mathcal{A}, \quad (19)$$

where

$$\begin{aligned} Q_n(\Phi_{\mathcal{A}}) &= -L_n(\Phi_{\mathcal{A}}) + n \underbrace{\sum_{m \in \mathcal{A}_1} \lambda_m \|\phi_m\|_2 + n \sum_{m \in \mathcal{A}_2} \lambda_m \|\phi_m\|_2}_{\triangleq nP(\Phi_{\mathcal{A}})} \\ &= -L_n(\Phi_{\mathcal{A}}) + nP(\Phi_{\mathcal{A}}). \end{aligned} \quad (20)$$

665 From (19) and (20) we have

$$\nabla_{\mathcal{A}} Q_n\left(\widehat{\Phi}_{\mathcal{A}}\right) = -\nabla_{\mathcal{A}} L_n\left(\widehat{\Phi}_{\mathcal{A}}\right) + n \nabla_{\mathcal{A}} P\left(\widehat{\Phi}_{\mathcal{A}}\right) = \mathbf{0}, \quad (21)$$

666 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 (21):

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

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

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

We also expand $n\nabla_{\mathcal{A}} P(\Phi_{\mathcal{A}})$ at $\Phi_{\mathcal{A}} = \Phi_{\mathcal{A}}^*$ in (21):

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

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

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

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

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

670 Therefore, by the central limit theorem, we know that

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

671 Hence,

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

672 \square

673 B Algorithm Details

674 In this section we provide more specific details about the algorithms used to solve the **sail** ob-

675 jective function. The strong heredity **sail** model with least-squares loss has the form

$$\hat{Y} = \beta_0 \cdot \mathbf{1} + \sum_{j=1}^p \Psi_j \boldsymbol{\theta}_j + \beta_E X_E + \sum_{j=1}^p \gamma_j \beta_E (X_E \circ \Psi_j) \boldsymbol{\theta}_j \quad (22)$$

676 and the objective function is given by

$$Q(\Phi) = \frac{1}{2n} \|Y - \hat{Y}\|_2^2 + \lambda(1 - \alpha) \left(w_E |\beta_E| + \sum_{j=1}^p w_j \|\boldsymbol{\theta}_j\|_2 \right) + \lambda \alpha \sum_{j=1}^p w_{jE} |\gamma_j| \quad (23)$$

Solving (23) 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

$$\frac{\partial Q}{\partial \beta_0} = \frac{1}{n} \left(Y - \beta_0 \cdot \mathbf{1} - \sum_{j=1}^p \Psi_j \boldsymbol{\theta}_j - \beta_E X_E - \sum_{j=1}^p \gamma_j \beta_E (X_E \circ \Psi_j) \boldsymbol{\theta}_j \right)^\top \mathbf{1} = 0 \quad (24)$$

$$\frac{\partial Q}{\partial \beta_E} = -\frac{1}{n} \left(X_E + \sum_{j=1}^p \gamma_j (X_E \circ \Psi_j) \boldsymbol{\theta}_j \right)^\top R + \lambda(1-\alpha) w_E s_1 = 0 \quad (25)$$

$$\frac{\partial Q}{\partial \boldsymbol{\theta}_j} = -\frac{1}{n} (\Psi_j + \gamma_j \beta_E (X_E \circ \Psi_j))^\top R + \lambda(1-\alpha) w_j s_2 = \mathbf{0} \quad (26)$$

$$\frac{\partial Q}{\partial \gamma_j} = -\frac{1}{n} (\beta_E (X_E \circ \Psi_j) \boldsymbol{\theta}_j)^\top R + \lambda \alpha w_{jE} s_3 = 0 \quad (27)$$

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:

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

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 (24)–(27) we see that

$$\hat{\beta}_0 = \left(Y - \sum_{j=1}^p \Psi_j \hat{\boldsymbol{\theta}}_j - \hat{\beta}_E X_E - \sum_{j=1}^p \hat{\gamma}_j \hat{\beta}_E (X_E \circ \Psi_j) \hat{\boldsymbol{\theta}}_j \right)^\top \mathbf{1} \quad (28)$$

$$\hat{\beta}_E = \frac{S \left(\frac{1}{n \cdot w_E} \left(X_E + \sum_{j=1}^p \hat{\gamma}_j (X_E \circ \Psi_j) \hat{\boldsymbol{\theta}}_j \right)^\top R_{(-E)}, \lambda(1-\alpha) \right)}{\left(X_E + \sum_{j=1}^p \hat{\gamma}_j (X_E \circ \Psi_j) \hat{\boldsymbol{\theta}}_j \right)^\top \left(X_E + \sum_{j=1}^p \hat{\gamma}_j (X_E \circ \Psi_j) \hat{\boldsymbol{\theta}}_j \right)} \quad (29)$$

$$\lambda(1-\alpha) w_j \frac{\boldsymbol{\theta}_j}{\|\boldsymbol{\theta}_j\|_2} = \frac{1}{n} (\Psi_j + \gamma_j \beta_E (X_E \circ \Psi_j))^\top R_{(-j)} \quad (30)$$

$$\hat{\gamma}_j = \frac{S \left(\frac{1}{n \cdot w_{jE}} (\beta_E (X_E \circ \Psi_j) \boldsymbol{\theta}_j)^\top R_{(-jE)}, \lambda \alpha \right)}{(\beta_E (X_E \circ \Psi_j) \boldsymbol{\theta}_j)^\top (\beta_E (X_E \circ \Psi_j) \boldsymbol{\theta}_j)} \quad (31)$$

680 where $S(x, t) = \text{sign}(x)(|x| - t)$ is the soft-thresholding operator. We see from (28) and (29)

681 that there are closed form solutions for the intercept and β_E . From (31), 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 [39] to solve (30). 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.

689 B.1 Least-Squares sail with Strong Heredity

690 A more detailed algorithm for fitting the least-squares sail model with strong heredity is
691 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 (23)
 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: $\Delta = \sum_j (\gamma_j^{(k)} - \gamma_j^{(k)(new)}) \tilde{X}_j$
 13: $R^* \leftarrow R^* + \Delta$
 14: • To update $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_p)$
 15: $\tilde{X}_j \leftarrow \Psi_j + \gamma_j^{(k)} \beta_E^{(k)} \tilde{\Psi}_j$ for $j = 1, \dots, p$
 16: **for** $j = 1, \dots, p$ **do**
 17: $R \leftarrow R^* + \tilde{X}_j \boldsymbol{\theta}_j^{(k)}$
 18:
$$\boldsymbol{\theta}_j^{(k)(new)} \leftarrow \arg \min_{\boldsymbol{\theta}_j} \frac{1}{2n} \left\| R - \tilde{X}_j \boldsymbol{\theta}_j \right\|_2^2 + \lambda(1 - \alpha) w_j \|\boldsymbol{\theta}_j\|_2$$

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

▷ $S(x, t) = \text{sign}(x)(|x| - t)_+$
 24: $\Delta = (\beta_E^{(k)} - \beta_E^{(k)(new)}) \tilde{X}_E$
 25: $R^* \leftarrow R^* + \Delta$
 26: • To update β_0
 27: $R \leftarrow R^* + \beta_0^{(k)}$
 28:
 29:
$$\beta_0^{(k)(new)} \leftarrow \frac{1}{n} R \cdot \mathbf{1}$$

 30: $\Delta = \beta_0^{(k)} - \beta_0^{(k)(new)}$
 31: $R^* \leftarrow R^* + \Delta$
 32:
 33: **until** convergence criterion is satisfied: $|Q(\Phi^{(k-1)}) - Q(\Phi^{(k)})| / Q(\Phi^{(k-1)}) < \epsilon$

692 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 $\boldsymbol{\theta}_s$ ($s \in 1, \dots, p$) at the k th iteration is given by

$$R_s = Y - \tilde{Y}_{(-s)}^{(k)} \quad (32)$$

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

$$\tilde{Y}_{(-s)}^{(k)} = \beta_0^{(k)} - \beta_E^{(k)} X_E - \sum_{\ell \neq s} \Psi_\ell \boldsymbol{\theta}_\ell^{(k)} - \sum_{\ell \neq s} \gamma_\ell^{(k)} \beta_E^{(k)} \tilde{\Psi}_\ell \boldsymbol{\theta}_\ell^{(k)} \quad (33)$$

Using (33), (32) can be re-written as

$$\begin{aligned} R_s &= Y - \beta_0^{(k)} - \beta_E^{(k)} X_E - \sum_{j=1}^p (\Psi_j + \gamma_j^{(k)} \beta_E^{(k)} \tilde{\Psi}_j) \boldsymbol{\theta}_j^{(k)} + (\Psi_s + \gamma_s^{(k)} \beta_E^{(k)} \tilde{\Psi}_s) \boldsymbol{\theta}_s^{(k)} \\ &= R^* + (\Psi_s + \gamma_s^{(k)} \beta_E^{(k)} \tilde{\Psi}_s) \boldsymbol{\theta}_s^{(k)} \end{aligned} \quad (34)$$

693 where

$$R^* = Y - \beta_0^{(k)} - \beta_E^{(k)} X_E - \sum_{j=1}^p (\Psi_j + \gamma_j^{(k)} \beta_E^{(k)} \tilde{\Psi}_j) \boldsymbol{\theta}_j^{(k)} \quad (35)$$

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

$$\boldsymbol{\theta}_s^{(k)(\text{new})} = \arg \min_{\boldsymbol{\theta}_j} \frac{1}{2n} \left\| R_s - (\Psi_s + \gamma_s^{(k)} \beta_E^{(k)} \tilde{\Psi}_s) \boldsymbol{\theta}_j \right\|_2^2 + \lambda(1-\alpha) w_s \|\boldsymbol{\theta}_j\|_2 \quad (36)$$

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

$$R_{s+1} = R^* + (\Psi_{s+1} + \gamma_{s+1}^{(k)} \beta_E^{(k)} \tilde{\Psi}_{s+1}) \boldsymbol{\theta}_{s+1}^{(k)} + (\Psi_s + \gamma_s^{(k)} \beta_E^{(k)} \tilde{\Psi}_s) (\boldsymbol{\theta}_s^{(k)} - \boldsymbol{\theta}_s^{(k)(new)}) \quad (37)$$

where R^* is given by (35), $\boldsymbol{\theta}_s^{(k)}$ is the parameter value prior to the update, and $\boldsymbol{\theta}_s^{(k)(new)}$ is the updated value given by (36). Taking the difference between (34) and (37) gives

$$\begin{aligned} \Delta &= R_t - R_s \\ &= (\Psi_t + \gamma_t^{(k)} \beta_E^{(k)} \tilde{\Psi}_t) \boldsymbol{\theta}_t^{(k)} + (\Psi_s + \gamma_s^{(k)} \beta_E^{(k)} \tilde{\Psi}_s) (\boldsymbol{\theta}_s^{(k)} - \boldsymbol{\theta}_s^{(k)(new)}) - (\Psi_s + \gamma_s^{(k)} \beta_E^{(k)} \tilde{\Psi}_s) \boldsymbol{\theta}_s^{(k)} \\ &= (\Psi_t + \gamma_t^{(k)} \beta_E^{(k)} \tilde{\Psi}_t) \boldsymbol{\theta}_t^{(k)} - (\Psi_s + \gamma_s^{(k)} \beta_E^{(k)} \tilde{\Psi}_s) \boldsymbol{\theta}_s^{(k)(new)} \end{aligned} \quad (38)$$

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

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

699 The subgradient equations (25)–(27) can be used to determine the largest value of λ such
700 that all coefficients are 0. From the subgradient Equation (25), we see that $\beta_E = 0$ is a
701 solution if

$$\frac{1}{w_E} \left| \frac{1}{n} \left(X_E + \sum_{j=1}^p \gamma_j (X_E \circ \Psi_j) \boldsymbol{\theta}_j \right)^\top R_{(-E)} \right| \leq \lambda(1 - \alpha) \quad (39)$$

⁷⁰² From the subgradient Equation (26), we see that $\boldsymbol{\theta}_j = \mathbf{0}$ is a solution if

$$\frac{1}{w_j} \left\| \frac{1}{n} (\boldsymbol{\Psi}_j + \gamma_j \beta_E(X_E \circ \boldsymbol{\Psi}_j))^{\top} R_{(-j)} \right\|_2 \leq \lambda(1 - \alpha) \quad (40)$$

⁷⁰³ From the subgradient Equation (27), we see that $\gamma_j = 0$ is a solution if

$$\frac{1}{w_{jE}} \left| \frac{1}{n} (\beta_E(X_E \circ \boldsymbol{\Psi}_j) \boldsymbol{\theta}_j)^{\top} R_{(-jE)} \right| \leq \lambda\alpha \quad (41)$$

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:

$$\begin{aligned} \lambda_{max} &= \frac{1}{n(1 - \alpha)} \max \left\{ \frac{1}{w_E} \left(X_E + \sum_{j=1}^p \gamma_j (X_E \circ \boldsymbol{\Psi}_j) \boldsymbol{\theta}_j \right)^{\top} R_{(-E)}, \right. \\ &\quad \left. \max_j \frac{1}{w_j} \left\| (\boldsymbol{\Psi}_j + \gamma_j \beta_E(X_E \circ \boldsymbol{\Psi}_j))^{\top} R_{(-j)} \right\|_2 \right\} \end{aligned} \quad (42)$$

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

$$\hat{Y} = \beta_0 \cdot \mathbf{1} + \sum_{j=1}^p \Psi_j \boldsymbol{\theta}_j + \beta_E X_E + \sum_{j=1}^p \gamma_j (X_E \circ \Psi_j) (\beta_E \cdot \mathbf{1}_{m_j} + \boldsymbol{\theta}_j) \quad (43)$$

⁷⁰⁶ The objective function is given by

$$Q(\Phi) = \frac{1}{2n} \left\| Y - \hat{Y} \right\|_2^2 + \lambda(1-\alpha) \left(w_E |\beta_E| + \sum_{j=1}^p w_j \|\boldsymbol{\theta}_j\|_2 \right) + \lambda\alpha \sum_{j=1}^p w_{jE} |\gamma_j| \quad (44)$$

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

$$\frac{\partial Q}{\partial \beta_0} = \frac{1}{n} \left(Y - \beta_0 \cdot \mathbf{1} - \sum_{j=1}^p \Psi_j \boldsymbol{\theta}_j - \beta_E X_E - \sum_{j=1}^p \gamma_j (X_E \circ \Psi_j) (\beta_E \cdot \mathbf{1}_{m_j} + \boldsymbol{\theta}_j) \right)^\top \mathbf{1} = 0 \quad (45)$$

$$\frac{\partial Q}{\partial \beta_E} = -\frac{1}{n} \left(X_E + \sum_{j=1}^p \gamma_j (X_E \circ \Psi_j) \mathbf{1}_{m_j} \right)^\top R + \lambda(1-\alpha) w_E s_1 = 0 \quad (46)$$

$$\frac{\partial Q}{\partial \boldsymbol{\theta}_j} = -\frac{1}{n} (\Psi_j + \gamma_j (X_E \circ \Psi_j))^\top R + \lambda(1-\alpha) w_j s_2 = \mathbf{0} \quad (47)$$

$$\frac{\partial Q}{\partial \gamma_j} = -\frac{1}{n} ((X_E \circ \Psi_j) (\beta_E \cdot \mathbf{1}_{m_j} + \boldsymbol{\theta}_j))^\top R + \lambda\alpha w_{jE} s_3 = 0 \quad (48)$$

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

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

⁷⁰⁸ the partial residual without X_E as

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

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

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

710 From the subgradient Equation (46), we see that $\beta_E = 0$ is a solution if

$$\frac{1}{w_E} \left| \frac{1}{n} \left(X_E + \sum_{j=1}^p \gamma_j (X_E \circ \Psi_j) \mathbf{1}_{m_j} \right)^\top R_{(-E)} \right| \leq \lambda(1 - \alpha) \quad (49)$$

711 From the subgradient Equation (47), we see that $\boldsymbol{\theta}_j = \mathbf{0}$ is a solution if

$$\frac{1}{w_j} \left\| \frac{1}{n} (\Psi_j + \gamma_j (X_E \circ \Psi_j))^\top R_{(-j)} \right\|_2 \leq \lambda(1 - \alpha) \quad (50)$$

712 From the subgradient Equation (48), we see that $\gamma_j = 0$ is a solution if

$$\frac{1}{w_{jE}} \left| \frac{1}{n} ((X_E \circ \Psi_j)(\beta_E \cdot \mathbf{1}_{m_j} + \boldsymbol{\theta}_j))^\top R_{(-jE)} \right| \leq \lambda\alpha \quad (51)$$

From the subgradient equations we see that

$$\hat{\beta}_0 = \left(Y - \sum_{j=1}^p \Psi_j \hat{\boldsymbol{\theta}}_j - \hat{\beta}_E X_E - \sum_{j=1}^p \hat{\gamma}_j (X_E \circ \Psi_j)(\hat{\beta}_E \cdot \mathbf{1}_{m_j} + \hat{\boldsymbol{\theta}}_j) \right)^\top \mathbf{1} \quad (52)$$

$$\hat{\beta}_E = \frac{S \left(\frac{1}{n \cdot w_E} \left(X_E + \sum_{j=1}^p \hat{\gamma}_j (X_E \circ \Psi_j) \mathbf{1}_{m_j} \right)^\top R_{(-E)}, \lambda(1 - \alpha) \right)}{\left(X_E + \sum_{j=1}^p \hat{\gamma}_j (X_E \circ \Psi_j) \mathbf{1}_{m_j} \right)^\top \left(X_E + \sum_{j=1}^p \hat{\gamma}_j (X_E \circ \Psi_j) \mathbf{1}_{m_j} \right)} \quad (53)$$

$$\lambda(1 - \alpha) w_j \frac{\boldsymbol{\theta}_j}{\|\boldsymbol{\theta}_j\|_2} = \frac{1}{n} (\Psi_j + \gamma_j (X_E \circ \Psi_j))^\top R_{(-j)} \quad (54)$$

$$\hat{\gamma}_j = \frac{S \left(\frac{1}{n \cdot w_{jE}} \left((X_E \circ \Psi_j)(\beta_E \cdot \mathbf{1}_{m_j} + \boldsymbol{\theta}_j) \right)^\top R_{(-jE)}, \lambda\alpha \right)}{\left((X_E \circ \Psi_j)(\beta_E \cdot \mathbf{1}_{m_j} + \boldsymbol{\theta}_j) \right)^\top \left((X_E \circ \Psi_j)(\beta_E \cdot \mathbf{1}_{m_j} + \boldsymbol{\theta}_j) \right)} \quad (55)$$

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

716 descent procedure implemented in the `glmnet` package [14], while we use the quadratic
 717 majorization technique implemented in the `gglasso` package [39] to solve (54). Algorithm 4
 718 details the procedure used to fit the least-squares weak heredity `sail` model.

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

720 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 **0**
 721 is:

$$\begin{aligned}
 \lambda_{max} = & \frac{1}{n} \max \left\{ \frac{1}{(1-\alpha)w_E} \left(X_E + \sum_{j=1}^p \gamma_j (X_E \circ \boldsymbol{\Psi}_j) \mathbf{1}_{m_j} \right)^\top R_{(-E)}, \right. \\
 & \max_j \frac{1}{(1-\alpha)w_j} \left\| (\boldsymbol{\Psi}_j + \gamma_j (X_E \circ \boldsymbol{\Psi}_j))^\top R_{(-j)} \right\|_2, \\
 & \left. \max_j \frac{1}{\alpha w_{jE}} \left((X_E \circ \boldsymbol{\Psi}_j) (\beta_E \cdot \mathbf{1}_{m_j} + \boldsymbol{\theta}_j) \right)^\top R_{(-jE)} \right\} \quad (56)
 \end{aligned}$$

which reduces to

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

722 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( $\mathbf{X}, Y, X_E, \mathbf{basis}, \lambda, \alpha, w_j, w_E, w_{jE}, \epsilon$ )           ▷ Algorithm for solving (44)
2:    $\Psi_j \leftarrow \mathbf{basis}(X_j)$ ,  $\tilde{\Psi}_j \leftarrow X_E \circ \Psi_j$  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:    $R^* \leftarrow Y - \beta_0^{(k)} - \beta_E^{(k)} X_E - \sum_j \Psi_j \boldsymbol{\theta}_j^{(k)} - \sum_j \gamma_j^{(k)} \tilde{\Psi}_j (\beta_E^{(k)} \cdot \mathbf{1}_{m_j} + \boldsymbol{\theta}_j^{(k)})$ 
6:   repeat
7:     • To update  $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_p)$ 
8:        $\tilde{X}_j \leftarrow \tilde{\Psi}_j (\beta_E^{(k)} \cdot \mathbf{1}_{m_j} + \boldsymbol{\theta}_j^{(k)})$       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:       $\Delta = \sum_j (\gamma_j^{(k)} - \gamma_j^{(k)(new)}) \tilde{X}_j$ 
13:       $R^* \leftarrow R^* + \Delta$ 
14:      • To update  $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_p)$ 
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:           $R \leftarrow R^* + \tilde{X}_j \boldsymbol{\theta}_j^{(k)}$ 
18:          
$$\boldsymbol{\theta}_j^{(k)(new)} \leftarrow \arg \min_{\boldsymbol{\theta}_j} \frac{1}{2n} \left\| R - \tilde{X}_j \boldsymbol{\theta}_j \right\|_2^2 + \lambda(1 - \alpha) w_j \|\boldsymbol{\theta}_j\|_2$$

19:           $\Delta = \tilde{X}_j (\boldsymbol{\theta}_j^{(k)} - \boldsymbol{\theta}_j^{(k)(new)})$ 
20:           $R^* \leftarrow R^* + \Delta$ 
21:          • To update  $\beta_E$ 
22:             $\tilde{X}_E \leftarrow X_E + \sum_j \gamma_j^{(k)} \tilde{\Psi}_j \mathbf{1}_{m_j}$ 
23:            
$$\beta_E^{(k)(new)} \leftarrow \frac{1}{\tilde{X}_E^\top \tilde{X}_E} S \left( \frac{1}{n \cdot w_E} \tilde{X}_E^\top R, \lambda(1 - \alpha) \right)$$
           ▷  $S(x, t) = \text{sign}(x)(|x| - t)_+$ 
24:             $\Delta = (\beta_E^{(k)} - \beta_E^{(k)(new)}) \tilde{X}_E$ 
25:             $R^* \leftarrow R^* + \Delta$ 
26:            • To update  $\beta_0$ 
27:               $R \leftarrow R^* + \beta_0^{(k)}$ 
28:
29:              
$$\beta_0^{(k)(new)} \leftarrow \frac{1}{n} R^* \cdot \mathbf{1}$$

30:               $\Delta = \beta_0^{(k)} - \beta_0^{(k)(new)}$ 
31:               $R^* \leftarrow R^* + \Delta$ 
32:               $k \leftarrow k + 1$ 
33: until convergence criterion is satisfied:  $|Q(\Phi^{(k-1)}) - Q(\Phi^{(k)})| / Q(\Phi^{(k-1)}) < \epsilon$ 

```

723 C Additional Results on PRS for Educational Attainment

724

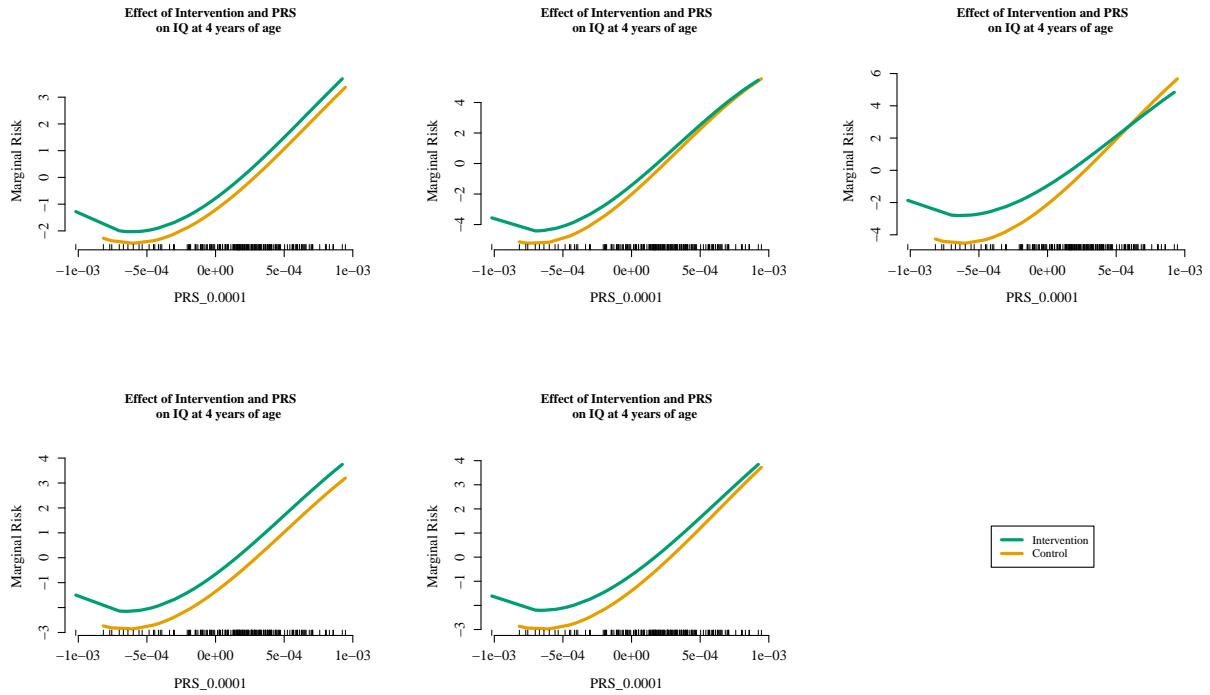


Figure C.1: Estimated interaction effect identified by the weak heredity `sail` using cubic B-splines and $\alpha = 0.1$ for the Nurse Family Partnership data for the 5 imputed datasets. Of the 189 subjects, 19 IQ scores were imputed using `mice` [5]. The selected model, chosen via 10-fold cross-validation,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

C ADDITIONAL RESULTS ON PRS FOR EDUCATIONAL ATTAINMENT

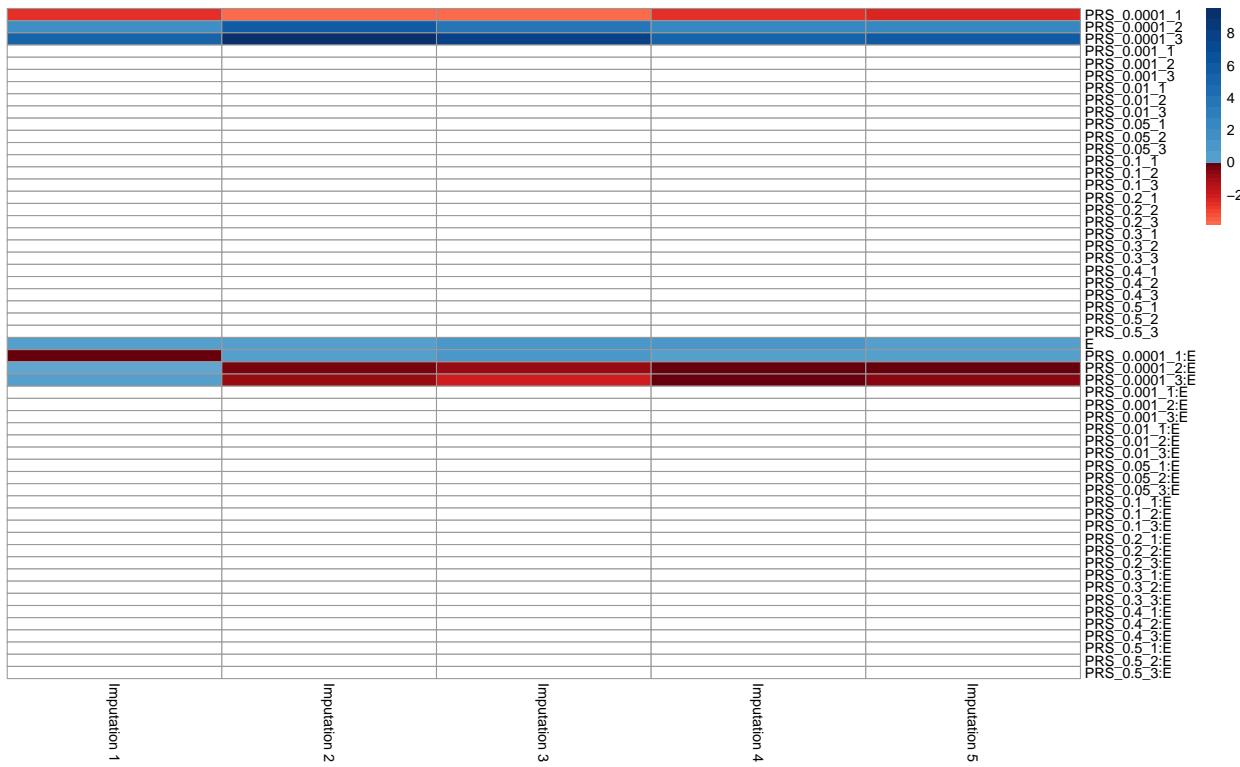


Figure C.2: Coefficient estimates obtained by the weak heredity `sail` using cubic B-splines and $\alpha = 0.1$ for the Nurse Family Partnership data for the 5 imputed datasets. Of the 189 subjects, 19 IQ scores were imputed using `mice` [5]. The selected model, chosen via 10-fold cross-validation, contained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