
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
¹⁴ result of changes in entire biological networks whose states are affected by a complex
¹⁵ interaction of genetic and environmental factors. However, when modelling a relevant
¹⁶ phenotype as a function of high dimensional measurements, power to estimate inter-
¹⁷ actions is low, the number of possible interactions could be enormous and their effects
¹⁸ may be non-linear. In this work, we introduce a method called **sail** for detecting non-
¹⁹ linear interactions with a key environmental or exposure variable in high-dimensional
²⁰ settings which respects the strong or weak heredity constraints. We prove that asymp-
²¹totically, our method possesses the oracle property, i.e., it performs as well as if the true
²² model were known in advance. We develop a computationally efficient fitting algorithm
²³ with automatic tuning parameter selection, which scales to high-dimensional datasets.
²⁴ Through an extensive simulation study, we show that **sail** outperforms existing pe-
²⁵nalized regression methods in terms of prediction accuracy and support recovery when
²⁶ there are non-linear interactions with an exposure variable. We apply **sail** to detect
²⁷ non-linear interactions between genes and a prenatal psychosocial intervention program

28 on cognitive performance in children at 4 years of age. Results show that individuals
29 who are genetically predisposed to lower educational attainment are those who stand
30 to benefit the most from the intervention. Our algorithms are implemented in an R
31 package available on CRAN (<https://cran.r-project.org/package=sail>).

32 *Keywords:* Blockwise coordinate descent, Gene-environment interaction, Hierarchical inter-
33 action, High-dimensional data, Penalized regression, Variable selection

34

1 Introduction

35 Computational approaches to variable selection have become increasingly important with
36 the advent of high-throughput technologies in genomics and brain imaging studies, where
37 the data has become massive, yet where it is believed that the number of truly important
38 variables is small relative to the total number of variables. Although many approaches
39 have been developed for main effects, there is an enduring interest in powerful methods for
40 estimating interactions, since interactions may reflect important modulation of a genomic
41 system by an external factor and vice versa [2]. Accurate capture of interactions may hold the
42 potential to better understanding biological phenomena and improving prediction accuracy.
43 For example, a model that considered interactions between brain imaging data and genetic
44 features had better classification accuracy compared to a model that considered the main
45 effects only [24]. Furthermore, the manifestations of disease are often considered to be
46 the result of changes in entire biological networks whose states are affected by a complex
47 interaction of genetic and environmental factors [31]. However, there is a general deficit of
48 such replicated interactions in the literature [35]. Indeed, power to detect interactions is
49 always lower than for main effects, and in high-dimensional settings ($p >> n$), this lack of
50 power to detect interactions is exacerbated, since the number of possible interactions could
51 be enormous and their effects may be non-linear. Hence, analytic methods that may improve
52 power are essential. Furthermore, methods capable of detecting non-linear interactions are

53 uncommon.

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

68 1.1 A sparse additive interaction model

69 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-
70 nary or continuous environment/exposure vector of known importance, and $\mathbf{X} = (X_1, \dots, X_p) \in \mathbb{R}^{n \times p}$
71 a matrix of additional predictors, possibly high-dimensional. Furthermore let $f_j : \mathbb{R} \rightarrow \mathbb{R}$ be
72 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.

An issue with (3) is that since no constraint is placed on the structure of the model, it is

possible that an estimated interaction term is non-zero while the corresponding main effects are zero. While there may be certain situations where this is plausible, statisticians have generally argued that interactions should only be included if the corresponding main effects are also in the model [22]. This is known as the strong heredity principle [7]. Indeed, large main effects are more likely to lead to detectable interactions [11]. In the next section we discuss how a simple reparametrization of the model (3) can lead to this desirable property.

1.2 Strong and weak heredity

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

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

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

An estimate of the regression parameters is given by $\widehat{\boldsymbol{\Phi}} = \arg \min_{\boldsymbol{\Phi}} Q(\boldsymbol{\Phi})$. This penalty allows for the possibility of excluding the interaction term from the model even if the corresponding main effects are non-zero. Furthermore, smaller values for α would lead to more interactions being included in the final model while values approaching 1 would favor main effects. Similar

116 to the elastic net [40], we fix α and obtain a solution path over a sequence of λ values.

117 **1.3 Toy example**

118 We present here a toy example to better illustrate the methods proposed in this paper.

119 With a sample size of $n = 100$, we sample $p = 20$ covariates X_1, \dots, X_p independently from

120 a $N(0, 1)$ distribution truncated to the interval $[0, 1]$. Data were generated from a model

121 which follows the strong heredity principle, but where only one covariate, X_2 , is involved in

122 an interaction with a binary exposure variable (E):

$$Y = f_1(X_1) + f_2(X_2) + 1.75E + 1.5E \cdot f_2(X_2) + \varepsilon.$$

123 For illustration, function $f_1(\cdot)$ is assumed to be linear, whereas function $f_2(\cdot)$ is non-linear:

124 $f_1(x) = -3x$, $f_2(x) = 2(2x - 1)^3$. The error term ε is generated from a normal distribution

125 with variance chosen such that the signal-to-noise ratio (SNR) is 2. We generated a single

126 simulated dataset and used the strong heredity **sail** method (described below) with cubic B-

127 splines to estimate the functional forms. 10-fold cross-validation (CV) was used to choose the

128 optimal value of penalization. We used $\alpha = 0.5$ and default values for all other arguments.

129 We plot the solution path for both main effects and interactions in Figure 1, coloring lines to

130 correspond to the selected model. We see that our method is able to correctly identify the true

131 model. We can also visually see the effect of the penalty and strong heredity principle working

132 in tandem, i.e., the interaction term $E \cdot f_2(X_2)$ (orange lines in the bottom panel) can only

133 be non-zero if the main effects E and $f_2(X_2)$ (black and orange lines respectively in the top

134 panel) are non-zero, while non-zero main effects does not imply a non-zero interaction.

135 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

136 their estimates from this analysis with **sail**. We are able to capture the shape of the correct

137 functional form, but the means are not well aligned with the data. Lack-of-fit for $f_1(X_1)$

138 can be partially explained by acknowledging that **sail** is trying to fit a cubic spline to a

139 linear function. Nevertheless, this example demonstrates that `sail` can still identify trends
140 reasonably well.

141 **1.4 Related work**

142 Methods for variable selection of interactions can be broken down into two categories: linear
143 and non-linear interaction effects. Many of the linear effect methods consider all pairwise
144 interactions in \mathbf{X} [3, 8, 33, 38] which can be computationally prohibitive when p is large.
145 More recent proposals for selection of interactions allow the user to restrict the search space
146 to interaction candidates [17, 20]. This is useful when the researcher wants to impose prior
147 information on the model. Two-stage procedures, where interaction candidates are con-
148 sidered from an original screen of main effects, have shown good performance when p is
149 large [15, 32] in the linear setting. There are many fewer methods available for estimating
150 non-linear interactions. For example, Radchenko and James (2010) [28] proposed a model
151 of the form

$$Y = \beta_0 + \sum_{j=1}^p f_j(X_j) + \sum_{j>k} f_{jk}(X_j, X_k) + \varepsilon,$$

152 where $f(\cdot)$ are smooth component functions. This method is more computationally expensive
153 than `sail` since it considers all pairwise interactions between the basis functions, and its
154 effectiveness in simulations or real-data applications is unknown as there is no software
155 implementation.

156 The main contributions of this paper are five-fold. First, we develop a model for non-
157 linear interactions with a key exposure variable, following either the weak or strong hered-
158 ity principle, that is computationally efficient and scales to the high-dimensional setting
159 ($n \ll p$). Second, through simulation studies, we show improved performance in terms of
160 prediction accuracy and support recovery over existing methods that only consider linear
161 interactions or additive main effects. Third, we show that our method possesses the oracle
162 property [13], i.e., it performs as well as if the true model were known in advance. Fourth,

we demonstrate the performance of our method in two applications: 1) gene-environment interactions in a prenatal psychosocial intervention program [26] and 2) a study aimed at identifying which clinical variables influence mortality rates amongst seriously ill hospitalized patients [10]. Fifth, we implement our algorithms in the **sail** R package on CRAN (<https://cran.r-project.org/package=sail>), along with extensive documentation. In particular, our implementation also allows for linear interaction models, user-defined basis expansions, a cross-validation procedure for selecting the optimal tuning parameter, and differential shrinkage parameters to apply the adaptive lasso idea [39].

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

2 Computation

In this section we describe a blockwise coordinate descent algorithm for fitting the least-squares version of the **sail** model in (4). We fix the value for α and minimize the objective function over a decreasing sequence of λ values ($\lambda_{max} > \dots > \lambda_{min}$). We use the subgradient equations to determine the maximal value λ_{max} such that all estimates are zero. Due to the heredity principle, this reduces to finding the largest λ such that all main effects $(\beta_E, \theta_1, \dots, \theta_p)$ are zero. Following Friedman et al. [14], we construct a λ -sequence of 100 values decreasing from λ_{max} to $0.001\lambda_{max}$ on the log scale, and use the warm start strategy where the solution for λ_ℓ is used as a starting value for $\lambda_{\ell+1}$.

186 **2.1 Blockwise coordinate descent for least-squares loss**

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

$$\hat{Y} = \beta_0 \cdot \mathbf{1} + \sum_{j=1}^p \Psi_j \boldsymbol{\theta}_j + \beta_E X_E + \sum_{j=1}^p \gamma_{jE} \beta_E (X_E \circ \Psi_j) \boldsymbol{\theta}_j, \quad (5)$$

188 and the objective function is given by

$$Q(\Phi) = \frac{1}{2n} \left\| Y - \hat{Y} \right\|_2^2 + \lambda(1-\alpha) \left(w_E |\beta_E| + \sum_{j=1}^p w_j \|\boldsymbol{\theta}_j\|_2 \right) + \lambda\alpha \sum_{j=1}^p w_{jE} |\gamma_{jE}|. \quad (6)$$

189 Solving (6) in a blockwise manner allows us to leverage computationally fast algorithms for
190 ℓ_1 and ℓ_2 norm penalized regression. We show in Supplemental Section B that by careful
191 construction of pseudo responses and pseudo design matrices, existing efficient algorithms can
192 be used to estimate the parameters. Indeed, the objective function simplifies to a modified
193 lasso problem when holding all $\boldsymbol{\theta}_j$ fixed, and a modified group lasso problem when holding
194 β_E and all γ_{jE} fixed. We provide an overview of the computations in Algorithm 1.

195 **2.2 Weak Heredity**

196 Our method can be easily adapted to enforce the weak heredity property. That is, an
197 interaction term can only be present if at least one of its corresponding main effects is
198 non-zero. To do so, we reparametrize the coefficients for the interaction terms in (2) as
199 $\boldsymbol{\alpha}_j = \gamma_{jE} (\beta_E \cdot \mathbf{1}_{m_j} + \boldsymbol{\theta}_j)$, where $\mathbf{1}_{m_j}$ is a vector of ones with dimension m_j (i.e. the length of $\boldsymbol{\theta}_j$).
200 We defer the algorithm details for fitting the **sail** model with weak heredity in Supplemental
201 Section B.4, as it is very similar to Algorithm 1 for the strong heredity **sail** model.

202 **2.3 Adaptive sail**

203 The weights for the environment variable, main effects and interactions are given by w_E, w_j
204 and w_{jE} respectively. These weights serve as a means of allowing a different penalty to be

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:
$$\boldsymbol{\gamma}^{(k)(new)} \leftarrow \arg \min_{\boldsymbol{\gamma}} \frac{1}{2n} \left\| \tilde{Y} - \sum_j \gamma_{jE} \tilde{X}_j \right\|_2^2 + \lambda \alpha \sum_j w_{jE} |\gamma_{jE}| \quad (7)$$
 - 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:
$$\boldsymbol{\theta}_j^{(k)(new)} \leftarrow \arg \min_{\boldsymbol{\theta}_j} \frac{1}{2n} \left\| \tilde{Y} - \tilde{X}_j \boldsymbol{\theta}_j \right\|_2^2 + \lambda(1 - \alpha) w_j \|\boldsymbol{\theta}_j\|_2 \quad (8)$$
 - 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 ($S(x, t) = \text{sign}(x)(|x| - t)_+$):
$$\beta_E^{(k)(new)} \leftarrow \frac{1}{\tilde{X}_E^\top \tilde{X}_E} S \left(\frac{1}{n \cdot w_E} \tilde{X}_E^\top \tilde{Y}, \lambda(1 - \alpha) \right) \quad (9)$$
 - iv. Set $\beta_E^{(k+1)} \leftarrow \beta_E^{(k)(new)}$, $k \leftarrow k + 1$
-

205 applied to each variable. In particular, any variable with a weight of zero is not penalized
206 at all. This feature is usually selected for one of two reasons:

- 207 1. Prior knowledge about the importance of certain variables is known. Larger weights
208 will penalize the variable more, while smaller weights will penalize the variable less
209 2. Allows users to apply the adaptive `sail`, similar to the adaptive lasso [39]

210 We describe the adaptive `sail` in Algorithm 2. This is a general procedure that can be
211 applied to the weak and strong heredity settings, as well as both least squares and logistic
212 loss functions. We provide this capability in the `sail` package using the `penalty.factor`
213 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
-

214 **2.4 Flexible design matrix**

215 The definition of the basis expansion functions in (1) is very flexible, in the sense that our
216 algorithms are independent of this choice. As a result, the user can apply any basis expansion
217 they desire. In the extreme case, one could apply the identity map, i.e., $f_j(X_j) = X_j$ which
218 leads to a linear interaction model (referred to as `linear sail`). When little information is
219 known a priori about the relationship between the predictors and the response, by default, we
220 choose to apply the same basis expansion to all columns of \mathbf{X} . This is a reasonable approach
221 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 $\widehat{\Phi}$, 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 [39]. 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

$$\widehat{\Phi}_n = \arg \min_{\Phi} Q_n(\Phi) = -L_n(\Phi) + n\lambda_m \sum_{m=1}^{2p+1} \|\boldsymbol{\phi}_m\|_2, \quad (10)$$

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 $m = p + 2, \dots, 2p + 1$. Define

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

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

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

245

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

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

250 **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
 251 $\|\widehat{\Phi}_n - \Phi^*\|_2 = O_p(\frac{1}{\sqrt{n}})$

252 Lemma (1) states that if the tuning parameters corresponding to the non-zero coefficients
 253 converge to 0 at a speed faster than $\frac{1}{\sqrt{n}}$, then there exists a local minimizer of $Q_n(\Phi)$ which
 254 is \sqrt{n} -consistent [8, 36].

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

$$P \left(\widehat{\Phi}_{\mathcal{A}_1^c} = \mathbf{0} \right) \rightarrow 1 \quad \text{and} \quad P \left(\widehat{\Phi}_{\mathcal{A}_2^c} = \mathbf{0} \right) \rightarrow 1 \quad (11)$$

256 Theorem (1) shows that **sail** can consistently remove the main effects and interaction terms
 257 which are not associated with the response with high probability. Together with Lemma (1),
 258 we see that the asymptotic behaviour of the penalty terms for the zero and non-zero predic-
 259 tors must be different to satisfy the model selection consistency property (11) [23]. Specif-
 260 ically, when the tuning parameters for the non-zero coefficients converge to 0 faster than
 261 $1/\sqrt{n}$ (i.e. $\sqrt{n}a_n \rightarrow 0$) and those for zero coefficients are large enough (i.e. $\sqrt{n}b_n \rightarrow$
 262 ∞), the Lemma (1) and Theorem (1) imply that the \sqrt{n} -consistent estimator $\widehat{\Phi}_n$ satisfies

263 $P(\widehat{\Phi}_{\mathcal{A}_2^c} = \mathbf{0}) \rightarrow 1.$

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

265 **Theorem 2** (Asymptotic normality). Denote $\mathcal{A} = \mathcal{A}_1 \cup \mathcal{A}_2$. Assume that $\sqrt{n}a_n \rightarrow 0$ and
266 $\sqrt{n}b_n \rightarrow \infty$. Under the regularity conditions, the subvector $\widehat{\Phi}_{\mathcal{A}}$ of the local minimizer $\widehat{\Phi}_n$
267 given in Lemma (1) satisfies

$$\sqrt{n}(\widehat{\Phi}_{\mathcal{A}} - \Phi_{\mathcal{A}}^*) \xrightarrow{d} N(\mathbf{0}, \mathbf{I}^{-1}(\Phi_{\mathcal{A}}^*)), \quad (12)$$

268 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
269 in advance.

270 Together, Theorems (1) and (2) establish that if the tuning parameters satisfy the conditions
271 $\sqrt{n}a_n \rightarrow 0$ and $\sqrt{n}b_n \rightarrow \infty$, then as the sample size grows large, `sail` has the oracle
272 property [13]. In order for the conditions on the tuning parameters to be satisfied, we follow
273 the strategies outlined for the adaptive Lasso [39], the adaptive group Lasso [23] and the
274 adaptive elastic-net [41]. That is, we define the adaptive weights as $w_m = \|\widehat{\phi}_m^{\text{init}} + 1/n\|_2^{-\xi}$
275 for $m = 1, \dots, 2p + 1$, where ξ is a positive constant and $\widehat{\phi}_m^{\text{init}}$ is an initial \sqrt{n} -consistent
276 estimate of ϕ_m^* . Here, the $1/n$ is to avoid division by zero.

277 4 Simulation Study

278 In this section, we use simulated data to understand the performance of `sail` in different
279 scenarios.

280 4.1 Comparator Methods

281 Since there are no other packages that directly address our chosen problem, we selected
282 comparator methods based on the following criteria: 1) penalized regression methods that
283 can handle high-dimensional data ($n < p$), 2) allowing at least one of linear effects, non-

284 linear effects or interaction effects, and 3) having a software implementation in R. The selected
285 methods can be grouped into three categories:

- 286 1. Linear main effects: `lasso` [34], `adaptive lasso` [39]
287 2. Linear interactions: `lassoBT` [32], `GLinternet` [20]
288 3. Non-linear main effects: `HierBasis` [16], `SPAM` [29], `gamsel` [9]

289 For `GLinternet` we specified the `interactionCandidates` argument so as to only consider
290 interactions between the environment and all other X variables. For all other methods we
291 supplied (X, X_E) as the data matrix, 100 for the number of tuning parameters to fit, and
292 used the default values otherwise (R code for each method available at <https://github.com/>
293 [sahirbhatnagar/sail/blob/master/my_sims/method_functions.R](https://github.com/sahirbhatnagar/sail/blob/master/my_sims/method_functions.R)). `lassoBT` considers
294 all pairwise interactions as there is no way for the user to restrict the search space. `SPAM`
295 applies the same basis expansion to every column of the data matrix; we chose 5 basis spline
296 functions. `HierBasis` and `gamsel` selects whether a term in an additive model is non-zero,
297 linear, or a non-linear spline up to a specified max degrees of freedom per variable.

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

303 **4.2 Simulation Design**

304 To make the comparisons with other methods as fair as possible, we followed a simulation
305 framework that has been previously used for variable selection methods in additive mod-
306 els [19, 21]. We extend this framework to include interaction effects as well. The covariates
307 are simulated as follows. First, we generate x_1, \dots, x_{1000} independently from a standard

normal distribution truncated to the interval $[0,1]$ for $i = 1, \dots, n$. The first four variables are non-zero (i.e. active in the response), while the rest of the variables are zero (i.e. are noise variables). The outcome Y is then generated following one of the models and assumptions described below. We evaluate the performance of our method on three of its defining characteristics: 1) the strong heredity property, 2) non-linearity of predictor effects and 3) interactions. Simulation scenarios are designed specifically to test the performance of these characteristics.

1. Heredity simulation

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

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

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

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

Scenario (c) Truth only has interactions. In this simulation, the covariates involved in interactions do not have main effects as well.

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

2. Non-linearity simulation scenario

Truth is linear. `sail` is designed to model non-linearity; here we assess its per-

324 formance if the true model is completely linear.

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

325 3. Interactions simulation scenario

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

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

328 The true component functions are the same as in [19, 21] and are given by $f_1(t) = 5t$,
329 $f_2(t) = 3(2t - 1)^2$, $f_3(t) = 4\sin(2\pi t)/(2 - \sin(2\pi t))$, $f_4(t) = 6(0.1\sin(2\pi t) + 0.2\cos(2\pi t) +$
330 $0.3\sin(2\pi t)^2 + 0.4\cos(2\pi t)^3 + 0.5\sin(2\pi t)^3)$. We set $\beta_E = 2$ and draw ε from a normal
331 distribution with variance chosen such that the signal-to-noise ratio is 2. Using this setup,
332 we generated 200 replications consisting of a training set of $n = 200$, a validation set of
333 $n = 200$ and a test set of $n = 800$. The training set was used to fit the model and the
334 validation set was used to select the optimal tuning parameter corresponding to the minimum
335 prediction mean squared error (MSE). Variable selection results including true positive rate,
336 false positive rate and number of active variables (the number of variables with a non-zero
337 coefficient estimate) were assessed on the training set, and MSE was assessed on the test
338 set.

339 4.3 Results

340 The prediction accuracy and variable selection results for each of the five simulation scenarios
341 are shown in Figure 3 and Table 2, respectively. We see that `sail`, `adaptive sail` and `sail`
342 `weak` have the best performance in terms of both MSE and yielding correct sparse models
343 when the truth follows a strong heredity (scenario 1a), as we would expect, since this is

344 exactly the scenario that our method is trying to target. Our method is also competitive
345 when only main effects are present (scenario 3) and performs just as well as methods that
346 only consider linear and non-linear main effects (`HierBasis`, `SPAM`), owing to the penalization
347 applied to the interaction parameter. Due to the heredity property being violated in scenario
348 1c), no method can identify the correct model with the exception of `GLinternet`. When only
349 linear effects and interactions are present (scenario 2), we see that `adaptive sail` has similar
350 MSE compared to the other linear interaction methods (`lassoBT` and `GLinternet`) with a
351 better TPR and FPR. Overall, our simulation study results suggests that `sail` outperforms
352 existing methods when the true model contains non-linear interactions, and is competitive
353 even when the truth only has either linear or additive main effects.

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

362 To visualize the estimated interaction effects, we ordered the 200 simulation runs by the Eu-
363 clidean distance between the estimated and true regression functions. Following Radchenko
364 et al. [28], we then identified the 25th, 50th, and 75th best simulations and plotted, in Fig-
365 ures 5 and 6, the interaction effects of X_E with $f_3(X_3)$ and $f_4(X_4)$, respectively. We see
366 that `sail` does a good job at capturing the true interaction surface for $X_E \cdot f_3(X_3)$. Again,
367 the smoothing and shrinkage effect is apparent when looking at the interaction surfaces for
368 $X_E \cdot f_4(X_4)$

369 **5 Real data applications**

370 **5.1 Gene-environment interactions in the Nurse Family Partner-
371 ship program**

372 It is well known that environmental exposures can have an important impact on academic
373 achievement. Indeed, early intervention in young children has been shown to positively im-
374 pact intellectual abilities [6]. More recent studies have shown that cognitive performance,
375 a trait that measures the ability to learn, reason and solve problems, is also strongly influ-
376 enced by genetic factors. Genome-wide association studies (GWAS) suggest that 20% of the
377 variance in educational attainment (years of education) may be accounted for by common
378 genetic variation [25, 30]. Unsurprisingly, there is significant overlap in the SNPs that predict
379 educational attainment and measures of cognitive function. An interesting query that arises
380 is how the environment interacts with these genetics variants to predict measures of cognitive
381 function. To address this question, we analyzed data from the Nurse Family Partnership
382 (NFP), a psychosocial intervention program that begins in pregnancy and targets maternal
383 health, parenting and mother-infant interactions [26]. The Stanford Binet IQ scores at 4
384 years of age were collected for 189 subjects (including 19 imputed using `mice` [5]) born to
385 women randomly assigned to control ($n = 100$) or nurse-visited intervention groups ($n =$
386 89). For each subject, we calculated a polygenic risk score (PRS) for educational attainment
387 at different p-value thresholds using weights from the GWAS conducted in Okbay et al. [25].
388 In this context, individuals with a higher PRS have a propensity for higher educational at-
389 tainment. The goal of this analysis was to determine if there was an interaction between
390 genetic predisposition to educational attainment (X) and maternal participation in the NFP
391 program (E) on child IQ at 4 years of age (Y). We applied the weak heredity `sail` with cubic
392 B-splines and $\alpha = 0.1$ to encourage interactions, and selected the optimal tuning parameter
393 using 10-fold cross-validation. Our method identified an interaction between the intervention
394 and PRS which included genetic variants at the 0.0001 level of significance. This interaction

395 is shown in Figure 7. We see that the intervention has a much larger effect on IQ for lower
396 PRS compared to a higher PRS. In other words, perinatal home visitation by nurses can im-
397 pact IQ scores in children who are genetically predisposed to lower educational attainment.
398 Similar results were obtained for the other imputed datasets (Supplemental Section C).

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

405 **5.2 Study to Understand Prognoses Preferences Outcomes and Risks**
406 **of Treatment**

407 The Study to Understand Prognoses Preferences Outcomes and Risks of Treatment (SUP-
408 PORT) aimed at identifying which clinical variables influence medium-term (half-year) mor-
409 tality rate amongst seriously ill hospitalized patients and improving clinical decision mak-
410 ing [10]. With a relatively large sample size of 9,105 and detailed documentation of clinical
411 variables, the SUPPORT dataset allows detection of potential interactions using the strategy
412 implemented in **sail**. We applied **sail** to test for non-linear interactions between acute renal
413 failure or multiple organ system failure (ARF/MOSF), an important predictor for survival
414 rate, and 13 other variables that were deemed clinically relevant. These variables included
415 the number of comorbidities (excluding ARF/MOSF), age, sex, as well as multiple physio-
416 logical and blood biochemical indices. The response was whether a patient survived after
417 six months since hospitalization.

418 A total of 8,873 samples had complete data on all variables of interest. We randomly divided
419 these samples into equal sized training/validation/test splits and ran **lassoBT**, **GLinternet**,

and the weak heredity **sail** with cubic B-splines and $\alpha = 0.1$ (as was done in the Nurse Family Partnership program case study). A binomial distribution family was specified for **GLinternet**, whereas **lassoBT** had the same default settings as the simulation study since it did not support a specialized implementation for binary outcomes. We again ran each method on the training data, determined the optimal tuning parameter on the validation data based on the area under the receiver operating characteristic curve (AUC), and assessed AUC on the test data. We repeated this process 200 times and report the results in Table 3. We found that **sail** achieved similar prediction accuracy to **lassoBT** and **GLinternet**. However, the predictive performance of **lassoBT** and **GLinternet** relied on models which included many more variables. In Figure 8, we visualize the two strongest interaction effects associated with the number of comorbidities and age, respectively. For those having undergone ARF/MOSF, an increased number of comorbidities decreases their chance of survival, while there seems to be no such relationship for non-ARF/MOSF patients. The interaction between ARF/MOSF and age shows the risk incurred by ARF/MOSF is most distinguishing among patients between the ages of 70 and 80.

6 Discussion

In this article we have introduced the sparse additive interaction learning model **sail** for detecting non-linear interactions with a key environmental or exposure variable in high-dimensional settings. Using a simple reparametrization, we are able to achieve either the weak or strong heredity property without using a complex penalty function. We developed 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 main effects models [16]. To our knowledge, our proposal is the first to allow for non-linear interactions with a key exposure variable following the weak or strong heredity property in high-dimensional settings. We also provide a first software implementation for these models.

472 Description of Supplementary Materials

473 The reader is referred to the on-line Supplementary Materials for:

- 474 A **Proofs** - Regularity conditions and proofs for Lemma 1, Theorem 1 and Theorem 2
- 475 B **Algorithm Details** - Detailed description of the algorithms used to solve the strong
476 and weak heredity sail objective function.
- 477 C **Additional Results on PRS for Educational Attainment** - Estimated coefficient
478 estimates and visualization of interaction effects for the Nurse Family Partnership data
479 for the 5 imputed datasets.
- 480 D **Data Availability and Code to Reproduce Results** - Detailed description of the
481 materials required (code, datasets) to reproduce the results in the manuscript.

482 Acknowledgments

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

488 References

- 489 [1] F. Bach, R. Jenatton, J. Mairal, G. Obozinski, et al. Structured sparsity through convex
490 optimization. *Statistical Science*, 27(4):450–468, 2012.
- 491 [2] S. R. Bhatnagar, Y. Yang, B. Khundrakpam, A. C. Evans, M. Blanchette, L. Bouchard,
492 and C. M. Greenwood. An analytic approach for interpretable predictive models in high-
493 dimensional data in the presence of interactions with exposures. *Genetic epidemiology*,
494 42(3):233–249, 2018.
- 495 [3] J. Bien, J. Taylor, R. Tibshirani, et al. A lasso for hierarchical interactions. *The Annals
496 of Statistics*, 41(3):1111–1141, 2013.
- 497 [4] P. Bühlmann and S. Van De Geer. *Statistics for high-dimensional data: methods, theory
498 and applications*. Springer Science & Business Media, 2011.
- 499 [5] S. v. Buuren and K. Groothuis-Oudshoorn. mice: Multivariate imputation by chained
500 equations in r. *Journal of statistical software*, pages 1–68, 2010.
- 501 [6] F. A. Campbell and C. T. Ramey. Effects of early intervention on intellectual and
502 academic achievement: a follow-up study of children from low-income families. *Child
503 development*, 65(2):684–698, 1994.
- 504 [7] H. Chipman. Bayesian variable selection with related predictors. *Canadian Journal of
505 Statistics*, 24(1):17–36, 1996.
- 506 [8] N. H. Choi, W. Li, and J. Zhu. Variable selection with the strong heredity constraint and
507 its oracle property. *Journal of the American Statistical Association*, 105(489):354–364,
508 2010.

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

- 554 for Statistical Computing, Vienna, Austria, 2017.
- 555 [28] P. Radchenko and G. M. James. Variable selection using adaptive nonlinear interac-
556 tion structures in high dimensions. *Journal of the American Statistical Association*,
557 105(492):1541–1553, 2010.
- 558 [29] P. Ravikumar, J. Lafferty, H. Liu, and L. Wasserman. Sparse additive models. *Journal*
559 *of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5):1009–1030,
560 2009.
- 561 [30] C. A. Rietveld, S. E. Medland, J. Derringer, J. Yang, T. Esko, N. W. Martin, H.-
562 J. Westra, K. Shakhbazov, A. Abdellaoui, A. Agrawal, et al. Gwas of 126,559 in-
563 dividuals identifies genetic variants associated with educational attainment. *science*,
564 340(6139):1467–1471, 2013.
- 565 [31] E. E. Schadt. Molecular networks as sensors and drivers of common human diseases.
566 *Nature*, 461(7261):218–223, 2009.
- 567 [32] R. D. Shah. Modelling interactions in high-dimensional data with backtracking. *Journal*
568 *of Machine Learning Research*, 17(207):1–31, 2016.
- 569 [33] Y. She and H. Jiang. Group regularized estimation under structural hierarchy. *arXiv*
570 *preprint arXiv:1411.4691*, 2014.
- 571 [34] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal*
572 *Statistical Society. Series B (Methodological)*, pages 267–288, 1996.
- 573 [35] N. J. Timpson, C. M. Greenwood, N. Soranzo, D. J. Lawson, and J. B. Richards. Genetic
574 architecture: the shape of the genetic contribution to human traits and disease. *Nature*
575 *Reviews Genetics*, 19(2):110, 2018.
- 576 [36] H. Wang, G. Li, and C.-L. Tsai. Regression coefficient and autoregressive order shrinkage
577 and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical*
578 *Methodology)*, 69(1):63–78, 2007.
- 579 [37] Y. Yang and H. Zou. A fast unified algorithm for solving group-lasso penalize learning
580 problems. *Statistics and Computing*, 25(6):1129–1141, 2015.
- 581 [38] P. Zhao, G. Rocha, and B. Yu. The composite absolute penalties family for grouped
582 and hierarchical variable selection. *The Annals of Statistics*, pages 3468–3497, 2009.
- 583 [39] H. Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical*
584 *association*, 101(476):1418–1429, 2006.
- 585 [40] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal*
586 *of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.
- 587 [41] H. Zou and H. H. Zhang. On the adaptive elastic-net with a diverging number of
588 parameters. *Annals of statistics*, 37(4):1733, 2009.

589 7 Tables and Figures

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

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

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}} $
sail	3.5 (0.6)	4 (3)	0.66 (0.01)	25 (3)
lassoBT	3.53 (0.477)	11 (6)	0.65 (0.009)	49 (14)
GLinternet ^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.

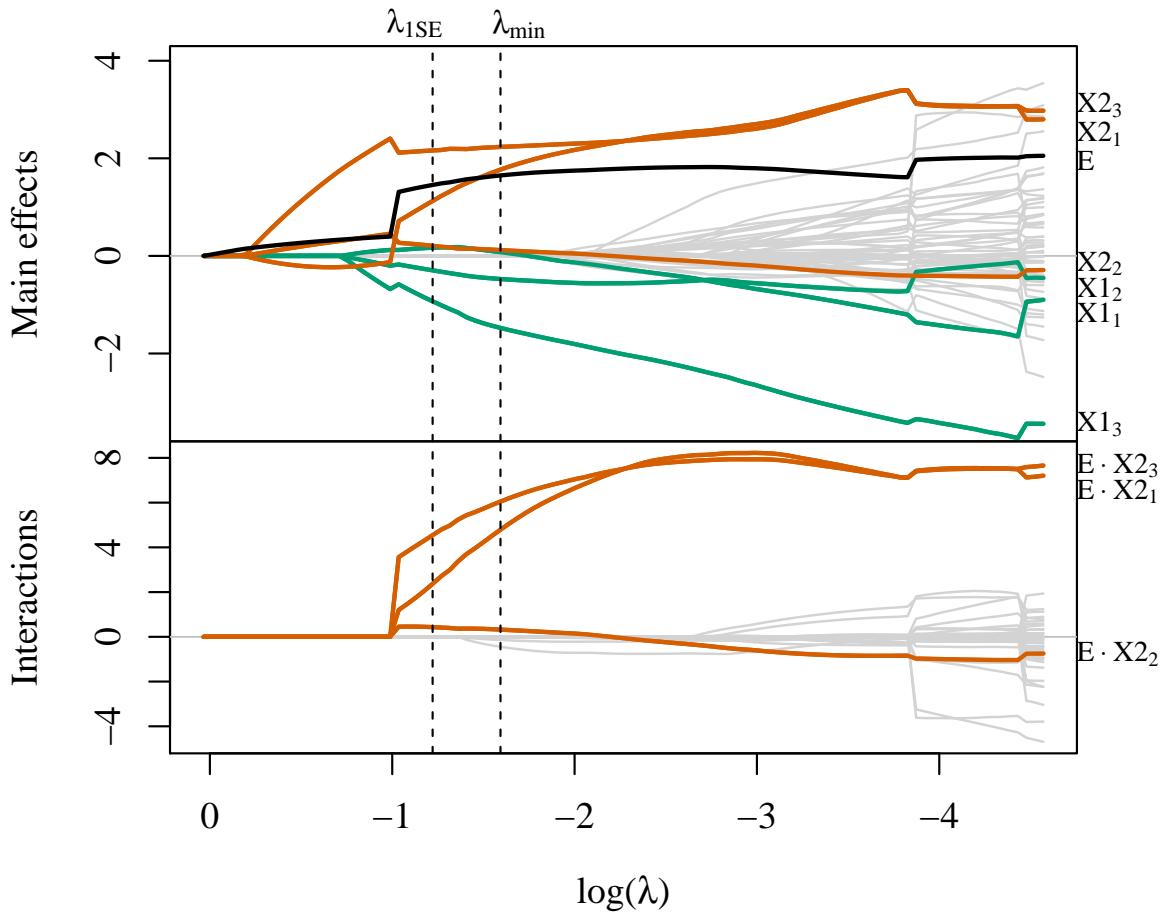


Figure 1: Toy example solution path for main effects (top) and interactions (bottom). $\{X1_1, X1_2, X1_3\}$ and $\{X2_1, X2_2, X2_3\}$ are the three basis coefficients for X_1 and X_2 , respectively. λ_{1SE} is the largest value of penalization for which the CV error is within one standard error of the minimizing value λ_{min} .

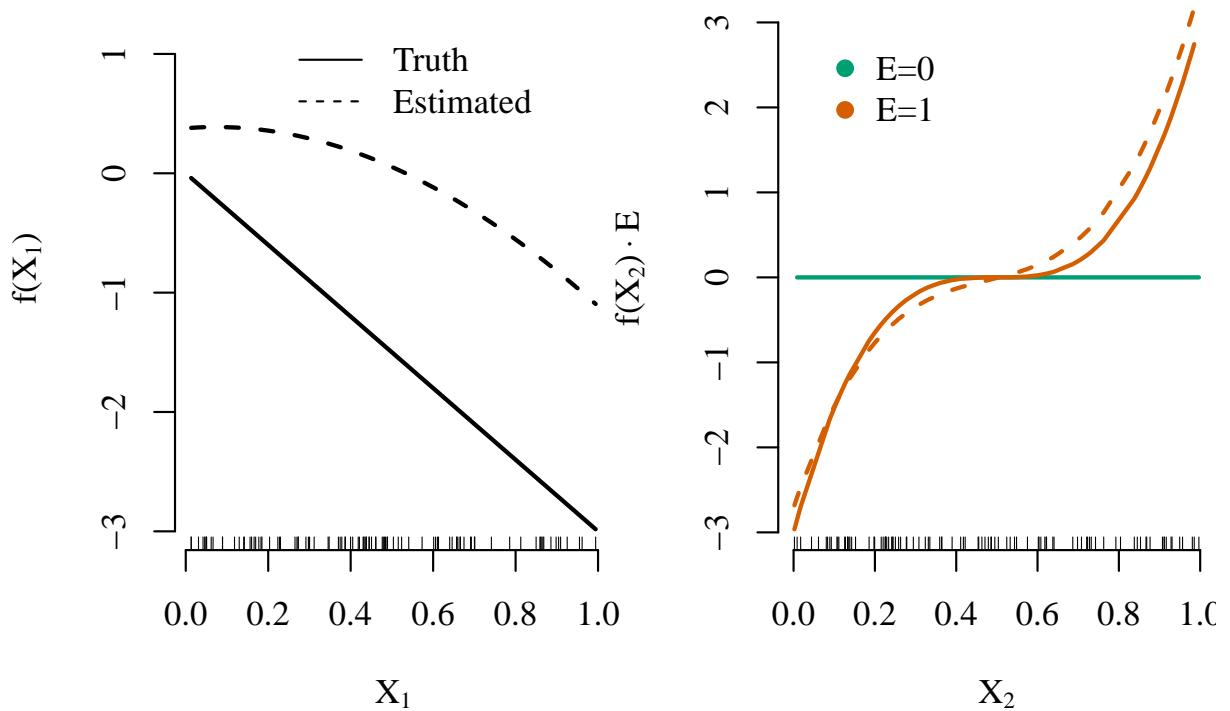


Figure 2: Estimated smooth functions for X_1 and the $X_2 \cdot E$ interaction by the `sail` method based on λ_{min} .

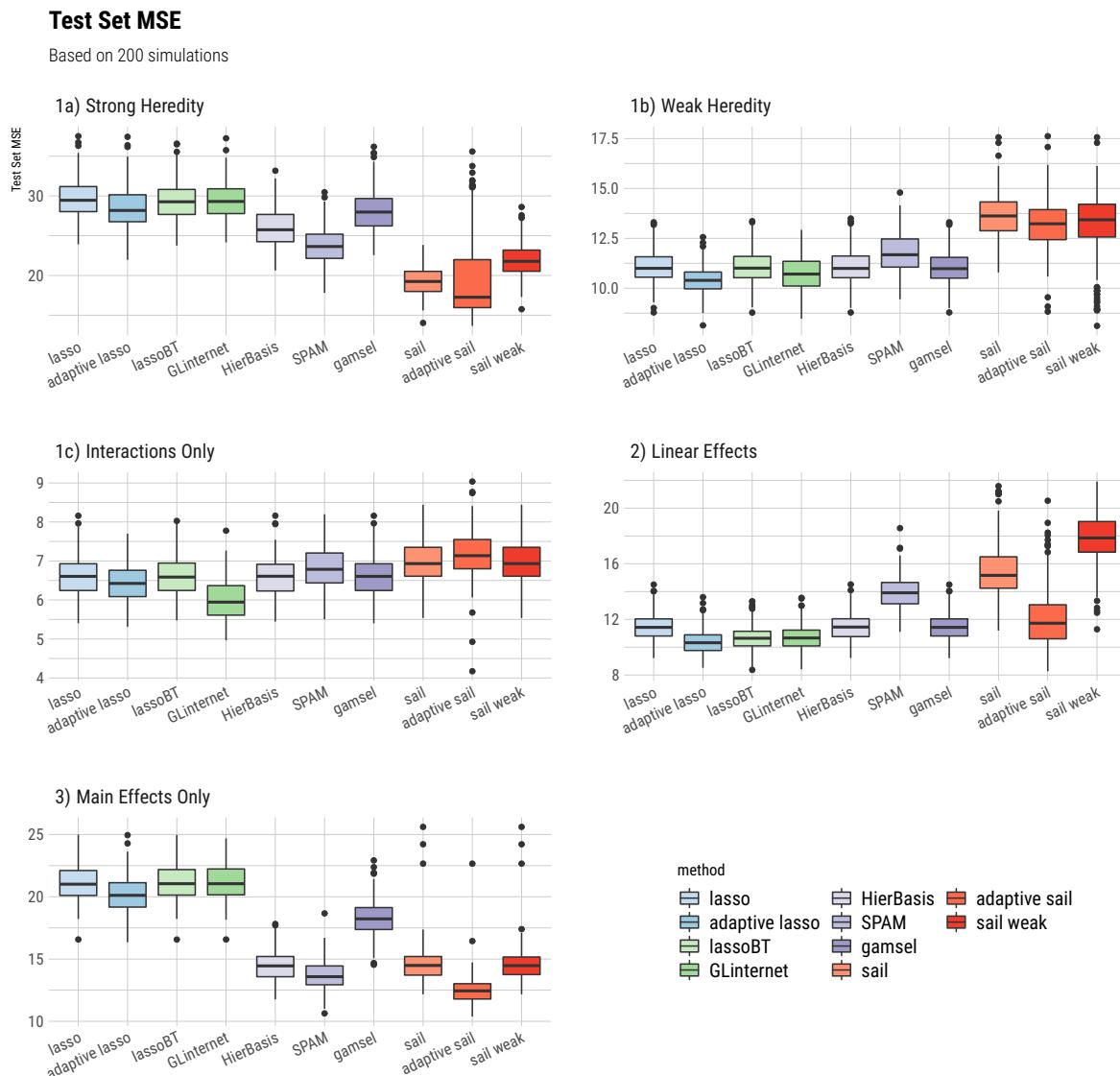


Figure 3: Boxplots of the test set mean squared error from 200 simulations for each of the five simulation scenarios.

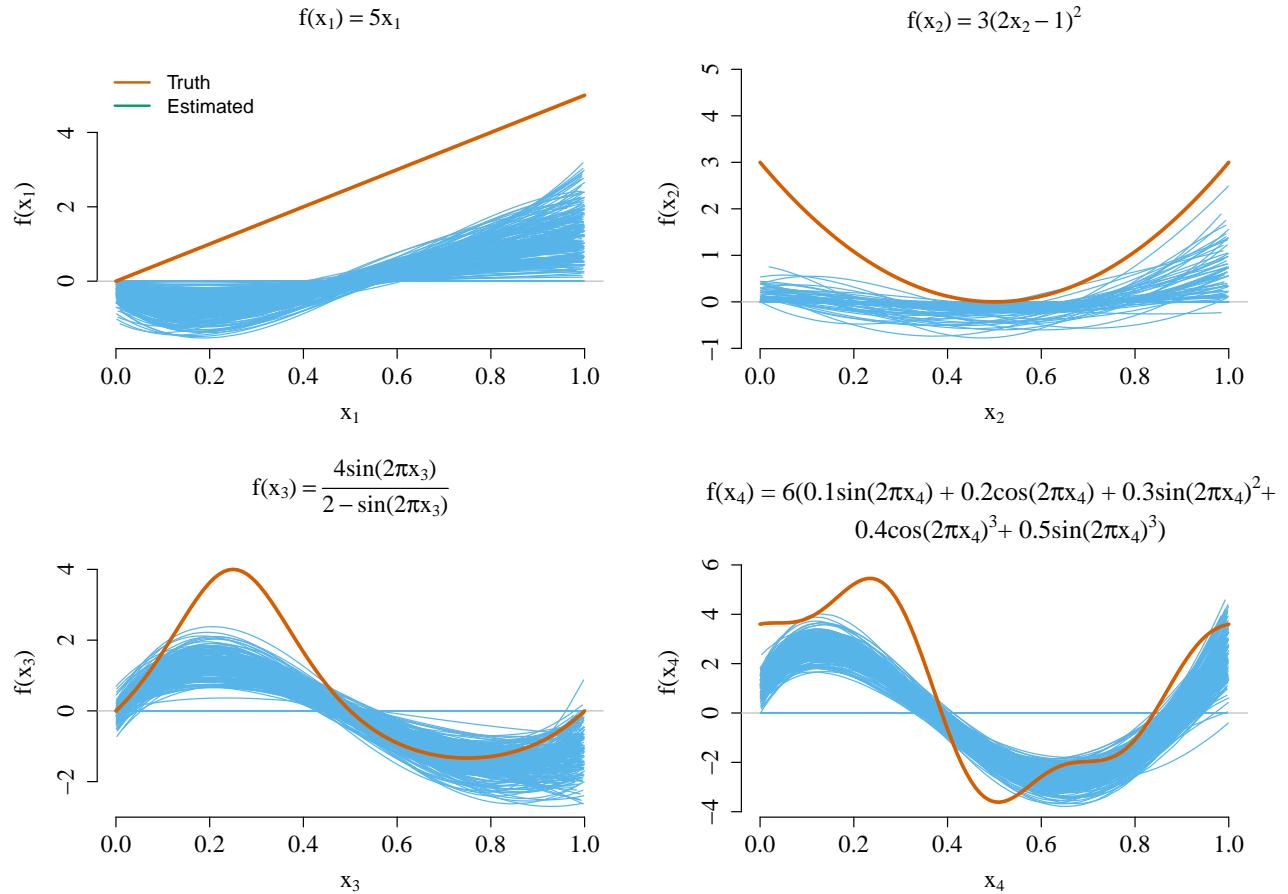


Figure 4: True and estimated main effect component functions for scenario 1a). The estimated curves represent the results from each one of the 200 simulations conducted.

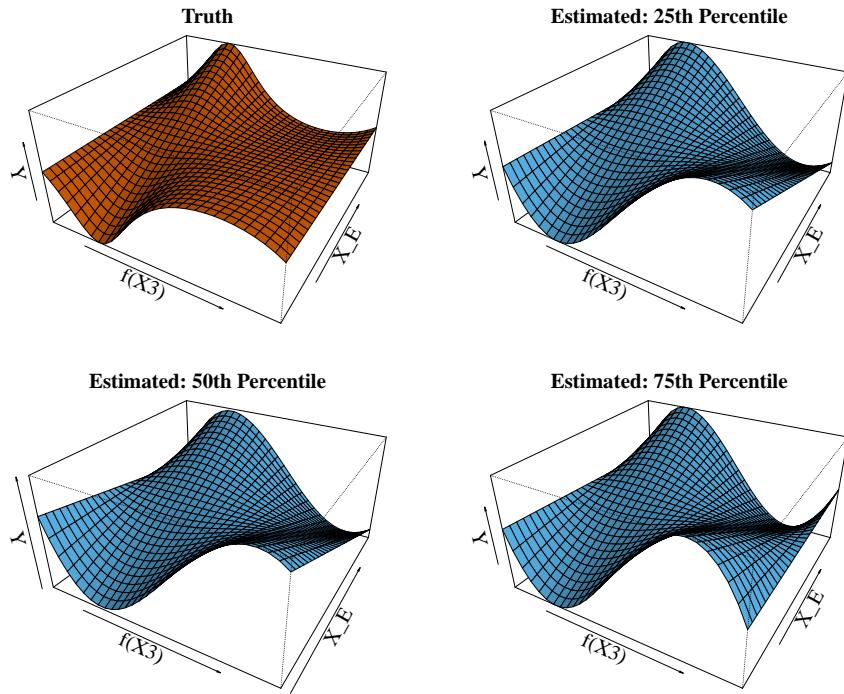


Figure 5: True and estimated interaction effects for $X_E \cdot f_3(X_3)$ in simulation scenario 1a).

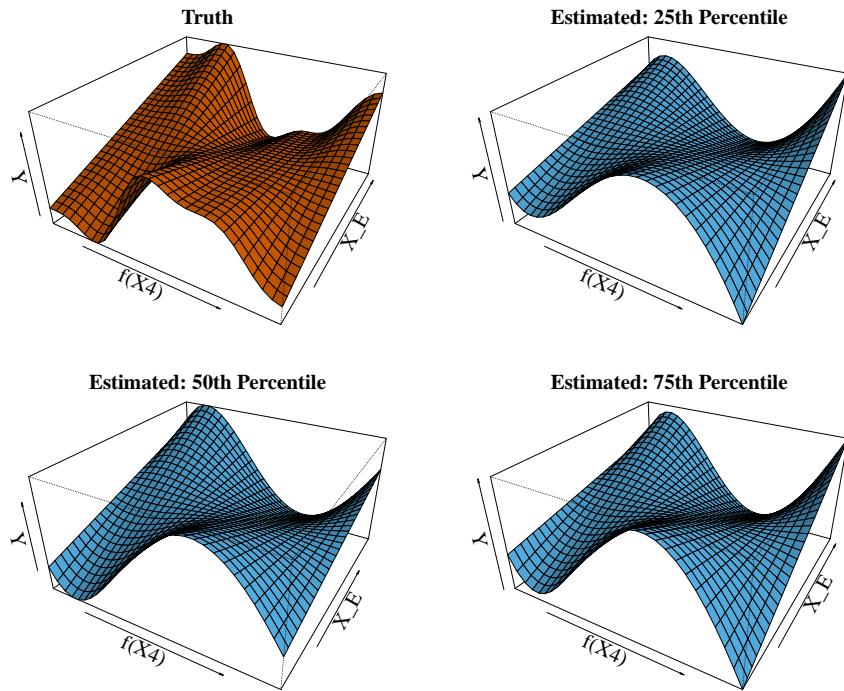


Figure 6: True and estimated interaction effects for $X_E \cdot f_4(X_4)$ in simulation scenario 1a).

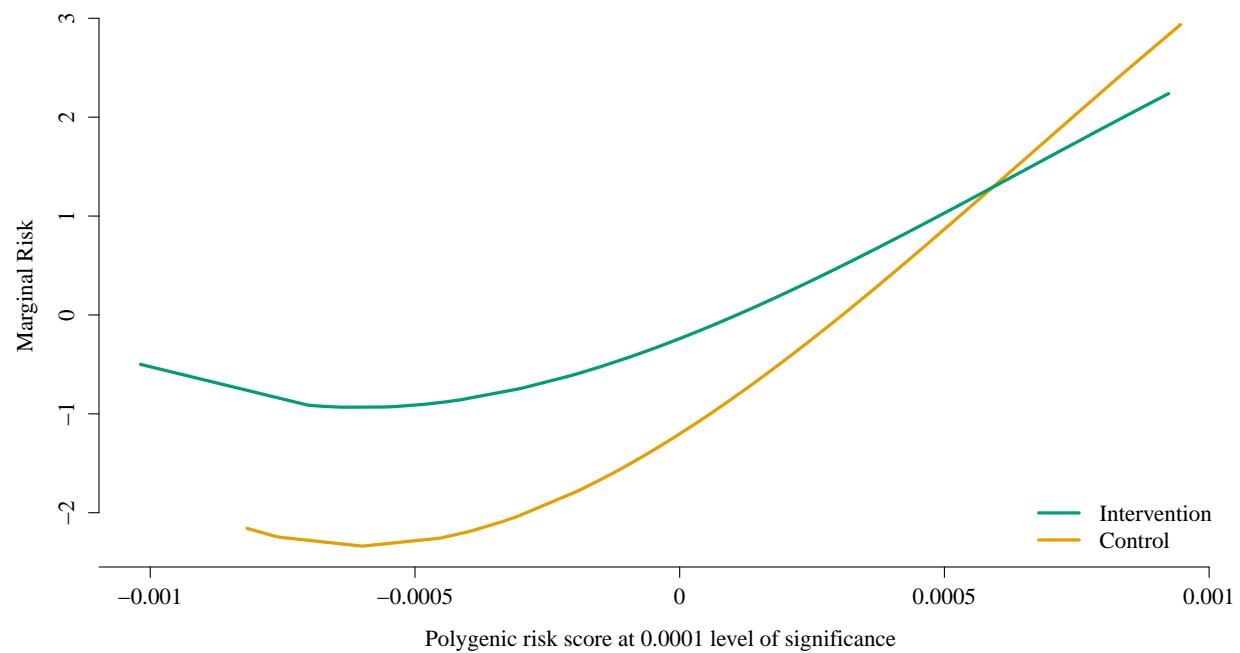


Figure 7: Estimated interaction effect identified by the weak heredity `sail` using cubic B-splines and $\alpha = 0.1$ for the Nurse Family Partnership data. The selected model, chosen via 10-fold cross-validation, contained three variables: the main effects for the intervention and the PRS for educational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

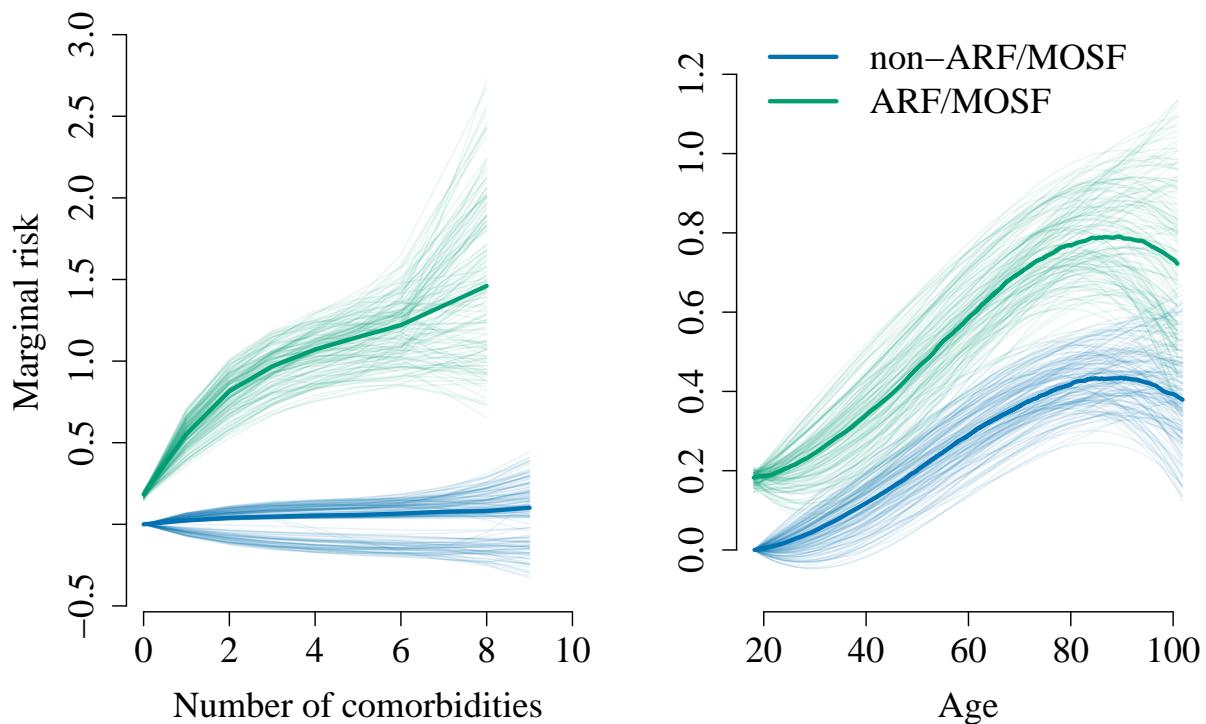


Figure 8: Illustration of estimated interaction effects identified by `sail` for the SUPPORT data. Median prediction curves in dark colors based on 200 train/validate/test splits represent the estimated marginal interaction effects. Coefficients estimated in each of the 200 train/validate/test splits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