Sparse Additive Interaction Learning

Sahir R Bhatnagar[1,2], Yi Yang[4], and Celia MT Greenwood[1,2,5]

[1]Department of Epidemiology, Biostatistics and Occupational Health, McGill University

[2]Lady Davis Institute, Jewish General Hospital, Montréal, QC

[4]Department of Mathematics and Statistics, McGill University

[5]Departments of Oncology and Human Genetics, McGill University

May 26, 2018

1 Introduction

Computational approaches to variable selection have become increasingly important with the advent of high-throughput technologies in genomics and brain imaging studies, where the data has become massive, yet where it is believed that the number of truly important variables is small relative to the total number of variables. Although many approaches have been developed for main effects, there is a consistent interest in powerful methods for estimating interactions, since interactions may reflect important modulation of a genomic system by an external factor. Accurate capture of interactions may hold the potential to better understand biological phenomena and improve prediction accuracy.

Furthermore, the manifestations of disease are often considered to be the result of changes in

entire biological networks whose states are affected by a complex interaction of genetic and environmental factors [1]. However, there is a general deficit of such replicated interactions in the literature [2]. Indeed, power to detect interactions is always lower than for main effects, and in high-dimensional settings ($p >> n$), this lack of power to detect interactions is exacerbated, since the number of possible interactions could be enormous and their effects may be non-linear. Hence, analytic methods that may improve power are essential.

Methylation
Brain imaging

You need a non-genetic example - i.e. gene expression or proteomic or ...

Results have on MSE but how you discuss variable effect estimation

Interactions may occur in numerous types and of varying complexities. In this paper, we consider one specific type of interaction models, where one (exposure) variable is involved in possibly non-linear interactions with a high-dimensional set of measures $\mathbf{X}$ leading to effects on a response variable, Y. We propose a multivariable penalization procedure for detecting non-linear interactions $\mathbf{X}$ and E.

Examples

1.1 A Sparse additive interaction model

Let $Y = (Y_1, \ldots, Y_n) \in \mathbb{R}^n$ be a continuous or binary outcome variable, $X_E = (E_1, \ldots, E_n) \in \mathbb{R}^n$ a binary or continuous environment vector, and $\boldsymbol{X} = (X_1, \ldots, X_p) \in \mathbb{R}^{n \times p}$ a matrix of predictors, possibly high-dimensional. Furthermore let $f_j : \mathbb{R} \to \mathbb{R}$ be a smoothing method for variable X_j by a projection on to a set of basis functions:

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

Here, the $\{\psi_{j\ell}\}_1^{m_j}$ are a family of basis functions in X_j [3]. Let $\boldsymbol{\Psi}_j$ be the $n \times m_j$ matrix of evaluations of the $\psi_{j\ell}$ and $\boldsymbol{\theta}_j = (\beta_{j1}, \ldots, \beta_{jm_j}) \in \mathbb{R}^{m_j}$ for $j = 1, \ldots, p$, i.e., $\boldsymbol{\theta}_j$ is a m_j-dimensional column vector of basis coefficients for the jth main effect. In this article we

$\theta_j = (\beta_{j1} \cdots \beta_{jm})$ be an ~~a~~ multidimensional

?

consider an additive interaction regression model of the form

$$g(\boldsymbol{\mu}) = \beta_0 \cdot \mathbf{1} + \sum_{j=1}^{p} \boldsymbol{\Psi}_j \boldsymbol{\theta}_j + \beta_E X_E + \sum_{j=1}^{p} (X_E \circ \boldsymbol{\Psi}_j)\boldsymbol{\tau}_j \tag{2}$$

where $g(\cdot)$ is a known link function, $\boldsymbol{\mu} = \mathsf{E}[Y|\boldsymbol{\Psi}, X_E]$, β_0 is the intercept, β_E is the coefficient for the environment variable, $\boldsymbol{\tau}_j = (\tau_{j1}, \ldots, \tau_{jm_j}) \in \mathbb{R}^{m_j}$ are the basis coefficients for the jth interaction term, and $(X_E \circ \boldsymbol{\Psi}_j)$ is the $n \times m_j$ matrix formed by the component-wise multiplication of the column vector X_E by each column of $\boldsymbol{\Psi}_j$. For a continuous response, we use the squared-error loss:

$$\mathcal{L}(\boldsymbol{\Theta}|\mathbf{D}) = \frac{1}{2n} \left\| Y - \beta_0 \cdot \mathbf{1} - \sum_{j=1}^{p} \boldsymbol{\Psi}_j \boldsymbol{\theta}_j - \beta_E X_E - \sum_{j=1}^{p} (X_E \circ \boldsymbol{\Psi}_j)\boldsymbol{\tau}_j \right\|_2^2 \tag{3}$$

and for binary response $Y_i \in \{-1, +1\}$ we use the logistic loss:

$$\mathcal{L}(\boldsymbol{\Theta}|\mathbf{D}) = \frac{1}{n} \sum_i \log \left(1 + \exp \left\{ -Y_i \left(\beta_0 \cdot \mathbf{1} - \sum_{j=1}^{p} \boldsymbol{\Psi}_j \boldsymbol{\theta}_j - \beta_E X_E - \sum_{j=1}^{p} (X_E \circ \boldsymbol{\Psi}_j)\boldsymbol{\tau}_j \right) \right\} \right) \tag{4}$$

where $\boldsymbol{\Theta} := (\beta_0, \beta_E, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p, \boldsymbol{\tau}_1, \ldots, \boldsymbol{\tau}_p)$ and $\mathbf{D} := (Y, \boldsymbol{\Psi}, X_E)$ is the working data.

Due to the large number of parameters to estimate with respect to the number of observations, one commonly-used approach 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. This reduced predictor set can lead to a more interpretable model with smaller prediction variance, albeit at the cost of having biased parameter estimates. In light of these goals, we consider the following objective function:

$$\arg\min_{\boldsymbol{\Theta}} \mathcal{L}(\boldsymbol{\Theta}|\mathbf{D}) + \lambda(1-\alpha) \left(w_E|\beta_E| + \sum_{j=1}^{p} w_j \|\boldsymbol{\theta}_j\|_2 \right) + \lambda\alpha \sum_{j=1}^{p} w_{jE} \|\boldsymbol{\tau}_j\|_2 \tag{5}$$

where $\|\boldsymbol{\theta}_j\|_2 = \sqrt{\sum_{k=1}^{m_j} \beta_{jk}^2}$, $\|\boldsymbol{\tau}_j\|_2 = \sqrt{\sum_{k=1}^{m_j} \tau_{jk}^2}$, $\lambda > 0$ and $\alpha \in (0,1)$ are tuning parameters, w_E, w_j, w_{jE} are adaptive weights for $j = 1, \ldots, p$. These weights serve as a way of allowing parameters to be penalized differently.

You need some sentences here on motivation for your objective function

An issue with (5) is that since no constraint is placed on the structure of the model, it is possible that an estimated interaction term is nonzero 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 [4]. This is known as the strong heredity principle [5]. Indeed, large main effects are more likely to lead to detectable interactions [6]. In the next section we discuss how a simple reparametrization of the model (5) can lead to this desirable property.

maybe this could be motivated with a diagram?

1.2 Strong and weak heredity

The strong heredity principle states that ~~the~~ an interaction term ~~can~~ only have a non-zero estimate if its corresponding main effects are estimated ~~to be non-zero~~ ~~and not forced to zero~~. The weak heredity principle allows for a non-zero interaction estimate as long as ~~one of~~ the corresponding main effects ~~are~~ is estimated to be non-zero [5]. In the context of penalized regression methods, these principles can be formulated as structured sparsity [7] problems. Several authors have proposed to modify the type of penalty in order to achieve the heredity principle [8, 9? ?]. We take an alternative approach. Following Choi et al. [10], we introduce a new set of parameters $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p) \in \mathbb{R}^p$ and reparametrize the coefficients for the interaction terms $\boldsymbol{\tau}_j$ in (2) as a function of γ_j and the main effect parameters $\boldsymbol{\theta}_j$ and β_E. This reparametrization for both strong and weak heredity is summarized in Table 1.

(weaker main)

To perform variable selection in this new parametrization, we penalize $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)$

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

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

instead of penalizing $\boldsymbol{\tau}$ as in (5), leading to the following objective function:

$$\arg\min_{\Theta} \mathcal{L}(\Theta|\mathbf{D}) + \lambda(1-\alpha)\left(w_E\beta_E + \sum_{j=1}^{p} w_j\|\boldsymbol{\theta}_j\|_2\right) + \lambda\alpha\sum_{j=1}^{p} w_{jE}|\gamma_j| \tag{6}$$

This penalty allows for the possibility of excluding the interaction term from the model even if the corresponding main effects are non-zero.

1.3 Toy example

We begin with a toy example to better illustrate our method. With a sample size of $n = 100$, we sample $p = 20$ covariates $X_1, \ldots X_p$ independently from a $N(0,1)$ truncated to the interval [0,1]. We generated data from a model which follows the strong heredity principle, but where only one covariates, X_2, is involved in the interaction with E:

$$Y = f_1(X_1) + f_2(X_2) + 1.75E + 1.5E \cdot f_2(X_2) + \varepsilon \tag{7}$$

Function $f_1(\cdot)$ is assumed to be linear, whereas function $f_2(\cdot)$ is non-linear: $f_1(x) = -3x$, $f_2(x) = 2(2x-1)^3$. The error term ε is generated from a normal distribution with variance chosen such that the signal-to-noise ratio (SNR) is 2. We generated a single simulated dataset and used the strong heredity `sail` method with cubic B-splines to estimate the functional forms. 10-fold CV was used to choose the optimal value of λ. We used $\alpha = 0.5$ and default values were used for all other arguments. We plot the solution path for both main effects

and interactions in Figure 1 ~~and we color~~ coloring lines to the lines corresponding ~~ing~~ to the selected model. We see that our method is able to correctly identify the true model. We can also visually see the effect of the penalty and strong heredity principle working in tandem, i.e., the interaction term $E \cdot f_2(X_2)$ (orange lines in the bottom panel) can only be nonzero if the main effects E and $f_2(X_2)$ (black and orange lines respectively in the top panel) are nonzero, while nonzero main effects doesn't necessarily imply a nonzero interaction.

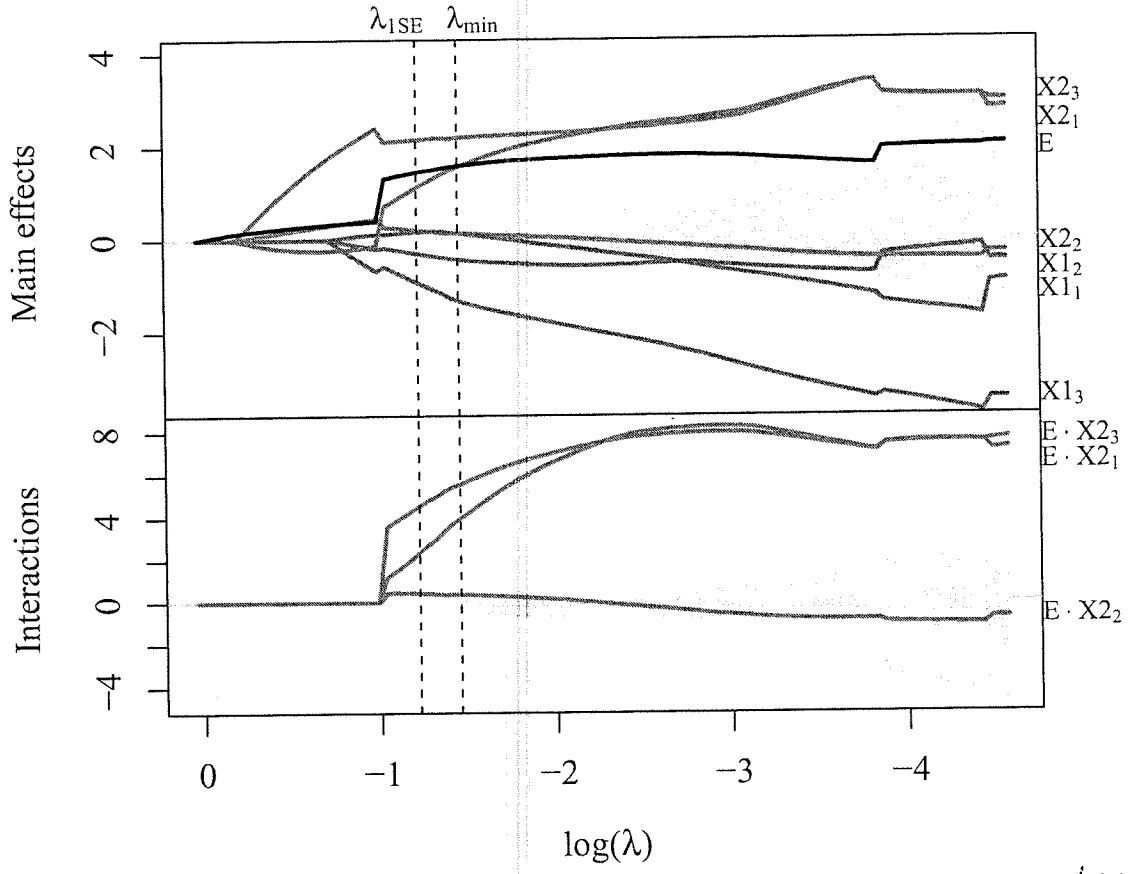

Figure 1: Toy example solution path ← Add explanation of notation, such as "X2₃"

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 their estimates ~~with~~ from this analysis with `sail`. We are able to capture the shape of the correct functional form, but ~~our~~ the means are not well aligned with the data. Lack-of-fit for $f_1(X_1)$ can be partially

explained by acknowledging that sail is trying to fit a cubic spline to a linear function. Nevertheless, this example demonstrates that it can still identify linear associations with reasonable sensitivity and specificity.

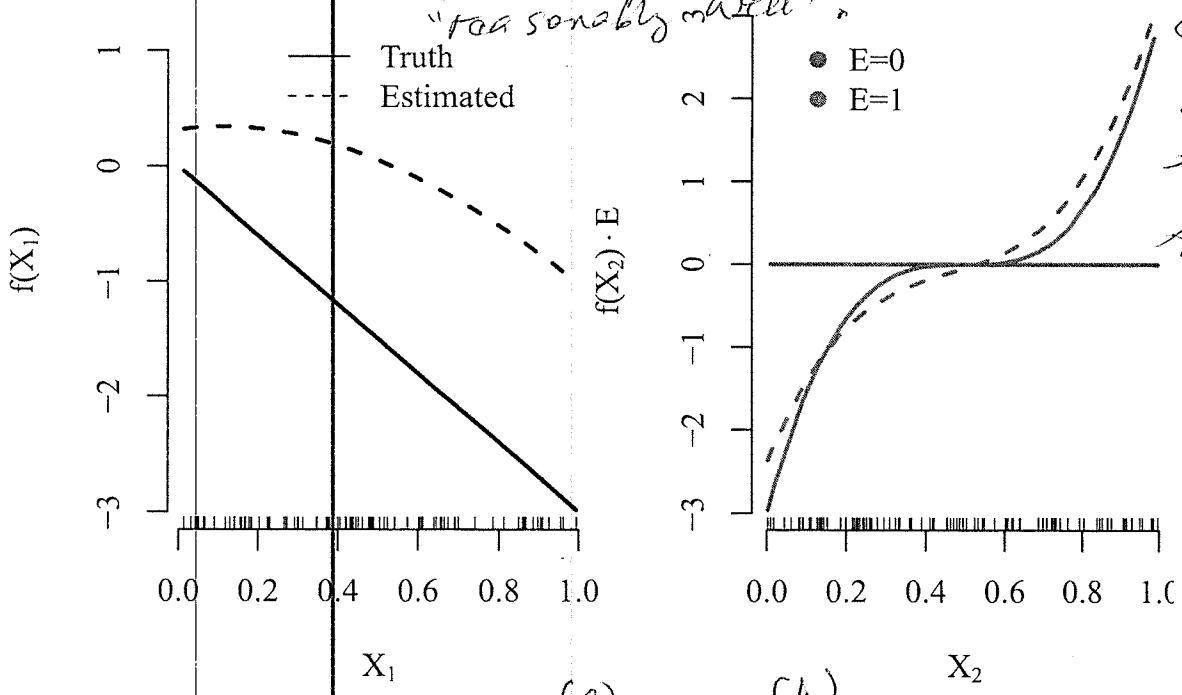

Figure 2: Estimated smooth functions by sail method based on λ_{min}.

1.4 Related Work

Methods for interaction selection can be broken down into two categories: linear and non-linear interaction effects. Many of the linear effect methods consider all pairwise interactions in $\boldsymbol{X}$ [9, 10, 11, 12] which can be computationally prohibitive when p is large. This is evidenced by the relatively small number of variables used in their simulations and real data analysis. More recent proposals allow the user to restrict the search space to interaction candidates [13, 14] which is useful when the researcher wants to impose some prior information on the model. Two-stage procedures, where interactions candidates are considered from

an original screen of main effects, have shown good performance when p is large [15, 16] in the linear setting. There are many fewer methods available for non-linear interactions. For example, Radchenko and James (2010) [8] proposed a model 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 component functions. This method is computationally expensive however, as it involves a complex penalty function and considers all pairwise interactions. Furthermore, it's effectiveness in both simulations and real-data applications is unknown as there is no software implementation.

While working on this paper, we were made aware of the recently proposed pliable lasso [17] which considers the interactions between $\boldsymbol{X}_{n\times p}$ and another matrix $\mathbf{Z}_{n\times K}$ and takes the form

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

where α_j is a K-dimensional vector. Our proposal is most closely related to this method with $\mathbf{Z}$ being a single column matrix; the key difference being the non-linearity of the predictor variables. As pointed out by the authors of the pliable lasso, these methods can be seen as a varying coefficient model, i.e., the effect of the predictors vary as a function of the exposure variable E.

The main contributions of this paper are fourfold. First, we develop a model for non-linear interactions with a key exposure variable following either the weak or strong heredity principle that is computationally efficient and scales to the high-dimensional setting ($n << p$). Second, through simulation studies, we show improved performance over existing methods that only consider linear interactions or additive main effects. Third, we show that our method possesses the oracle property [18], i.e., it performs as well as if the true model were known in advance. Fourth, all of our algorithms are implemented in the `sail` R package

hosted on GitHub with extensive documentation (`http://sahirbhatnagar.com/sail/`). 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 [19] idea.

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

2 Algorithm and Computational Details

In this section we describe a blockwise coordinate descent algorithm for fitting both the least-squares and logistic version of the `sail` model in (6). We fix the value for α and minimize the objective function over a decreasing sequence of λ values ($\lambda_{max} > \cdots > \lambda_{min}$). 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 $(\beta_E, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p)$ are zero. Following Friedman et al. [20], 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 $\lambda_{\ell+1}$.

2.1 Blockwise coordinate descent for least-squares loss

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

$$\hat{Y} = \beta_0 \cdot \mathbf{1} + \sum_{j=1}^{p} \mathbf{\Psi}_j \boldsymbol{\theta}_j + \beta_E X_E + \sum_{j=1}^{p} \gamma_j \beta_E (X_E \circ \mathbf{\Psi}_j)\boldsymbol{\theta}_j \tag{8}$$

and the objective function is given by

$$Q(\mathbf{\Theta}) = \frac{1}{2n}\left\|Y - \hat{Y}\right\|_2^2 + \lambda(1-\alpha)\left(w_E|\beta_E| + \sum_{j=1}^{p} w_j \|\boldsymbol{\theta}_j\|_2\right) + \lambda\alpha \sum_{j=1}^{p} w_{jE}|\gamma_j| \tag{9}$$

Solving (9) in a blockwise manner allows us to leverage computationally fast algorithms for ℓ_1 and ℓ_2 norm penalized regression. The objective function simplifies to a modified lasso problem when holding all θ_j fixed, and a modified group lasso problem when holding β_E and all γ_j fixed.

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

$$\frac{\partial Q}{\partial \beta_0} = \frac{1}{n}\left(Y - \beta_0 \cdot \mathbf{1} - \sum_{j=1}^{p} \mathbf{\Psi}_j \boldsymbol{\theta}_j - \beta_E X_E - \sum_{j=1}^{p} \gamma_j \beta_E (X_E \circ \mathbf{\Psi}_j)\boldsymbol{\theta}_j\right)^{\top} \mathbf{1} = 0 \tag{10}$$

$$\frac{\partial Q}{\partial \beta_E} = -\frac{1}{n}\left(X_E + \sum_{j=1}^{p} \gamma_j (X_E \circ \mathbf{\Psi}_j)\boldsymbol{\theta}_j\right)^{\top} R + \lambda(1-\alpha) w_E s_1 = 0 \tag{11}$$

$$\frac{\partial Q}{\partial \boldsymbol{\theta}_j} = -\frac{1}{n}\left(\mathbf{\Psi}_j + \gamma_j \beta_E (X_E \circ \mathbf{\Psi}_j)\right)^{\top} R + \lambda(1-\alpha) w_j s_2 = \mathbf{0} \tag{12}$$

$$\frac{\partial Q}{\partial \gamma_j} = -\frac{1}{n}\left(\beta_E (X_E \circ \mathbf{\Psi}_j)\boldsymbol{\theta}_j\right)^{\top} R + \lambda\alpha w_{jE} s_3 = 0 \tag{13}$$

where s_1 is in the subgradient of the ℓ_1 norm:

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

s_2 is in the subgradient of the ℓ_2 norm:

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

and s_3 is in the subgradient of the ℓ_1 norm:

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

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

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

the partial residual without X_E as

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

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

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

From the subgradient equations (10)–(13) we see that

$$\hat{\beta}_0 = \left(Y - \sum_{j=1}^{p} \boldsymbol{\Psi}_j \hat{\boldsymbol{\theta}}_j - \hat{\beta}_E X_E - \sum_{j=1}^{p} \hat{\gamma}_j \hat{\beta}_E (X_E \circ \boldsymbol{\Psi}_j) \hat{\boldsymbol{\theta}}_j \right)^\top \mathbf{1} \tag{14}$$

$$\hat{\beta}_E = S\left(\frac{1}{n \cdot w_E} \left(X_E + \sum_{j=1}^{p} \hat{\gamma}_j (X_E \circ \boldsymbol{\Psi}_j) \hat{\boldsymbol{\theta}}_j \right)^\top R_{(-E)}, \lambda(1-\alpha) \right) \tag{15}$$

$$\lambda(1-\alpha) w_j \frac{\boldsymbol{\theta}_j}{\|\boldsymbol{\theta}_j\|_2} = \frac{1}{n} \left(\boldsymbol{\Psi}_j + \gamma_j \beta_E (X_E \circ \boldsymbol{\Psi}_j) \right)^\top R_{(-j)} \tag{16}$$

$$\hat{\gamma}_j = S\left(\frac{1}{n \cdot w_{jE}} \left(\beta_E (X_E \circ \boldsymbol{\Psi}_j) \boldsymbol{\theta}_j \right)^\top R_{(-jE)}, \lambda\alpha \right) \tag{17}$$

where $S(x,t) = \text{sign}(x)(|x| - t)$ is the soft-thresholding operator. We see from (14) and (15) that there are closed form solutions for the intercept and β_E. From (17), each γ_j also has a closed form solution and can be solved efficiently for $j = 1, \ldots, p$ using the coordinate descent procedure implemented in the `glmnet` package [20]. While there is no closed form solution for β_j, we can use a quadratic majorization technique implemented in the `gglasso` package [21] to solve (16). From these estimates, we can compute the interaction effects using the reparametrizations presented in Table 1, e.g., $\hat{\boldsymbol{\tau}}_j = \hat{\gamma}_j \hat{\beta}_E \hat{\boldsymbol{\theta}}_j$, $j = 1, \ldots, p$ for the strong heredity `sail` model. We provide an overview of the computations in Algorithm 1. A more detailed version of this algorithm is given in Section A.1 of the Appendix.

in this section, you need to change the language to highlight innovation, e.g. Instead of "we can" something like "we show in the Appendix that by careful construction of subgradients, existing efficient algorithms can be used to estimate parameters" and pseudo versions

Your contribution need to be more visible

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

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

1. Initialize $\beta_0^{(0)}, \beta_E^{(0)}, \boldsymbol{\theta}_j^{(0)}, \gamma_j^{(0)}$ for $j = 1, \ldots, p$ and set iteration counter $k \leftarrow 0$.
2. Repeat the following until convergence:
   (a) update $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)$
      i. Compute the pseudo design: $\widetilde{X}_j \leftarrow \beta_E^{(k)}(X_E \circ \boldsymbol{\Psi}_j)\boldsymbol{\theta}_j^{(k)} \qquad$ for $j = 1, \ldots, p$
      ii. Compute the pseudo response $\widetilde{Y}$ by removing the contribution of every term not involving $\boldsymbol{\gamma}$ from Y
      iii. Solve:

$$\boldsymbol{\gamma}^{(k)(new)} \leftarrow \arg\min_{\boldsymbol{\gamma}} \frac{1}{2n}\left\|\widetilde{Y} - \sum_j \gamma_j \widetilde{X}_j\right\|_2^2 + \lambda\alpha \sum_j w_{jE}|\gamma_j| \tag{18}$$

      iv. Set $\boldsymbol{\gamma}^{(k)} = \boldsymbol{\gamma}^{(k)(new)}$
   (b) update $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p)$
      — for $j = 1, \ldots, p$
      i. Compute the pseudo design: $\widetilde{X}_j \leftarrow \boldsymbol{\Psi}_j + \gamma_j^{(k)}\beta_E^{(k)}(X_E \circ \boldsymbol{\Psi}_j)$
      ii. Compute the pseudo response $(\widetilde{Y})$ by removing the contribution of every term not involving $\boldsymbol{\theta}_j$ from Y
      iii. Solve:

$$\boldsymbol{\theta}_j^{(k)(new)} \leftarrow \arg\min_{\boldsymbol{\theta}_j} \frac{1}{2n}\left\|\widetilde{Y} - \widetilde{X}_j\boldsymbol{\theta}_j\right\|_2^2 + \lambda(1-\alpha)w_j\|\theta_j\|_2 \tag{19}$$

      iv. Set $\boldsymbol{\theta}_j^{(k)} \leftarrow \boldsymbol{\theta}_j^{(k)(new)}$
   (c) update β_E
      i. Compute the pseudo design: $\widetilde{X}_E \leftarrow X_E + \sum_j \gamma_j^{(k)}\widetilde{\boldsymbol{\Psi}}_j\boldsymbol{\theta}_j^{(k)}$
      ii. Compute the pseudo response $(\widetilde{Y})$ by removing the contribution of every term not involving β_E from Y
      iii. Soft-threshold update $(S(x,t) = \text{sign}(x)(|x| - t)_+)$:

$$\beta_E^{(k)(new)} \leftarrow S\left(\frac{1}{n \cdot w_E}\widetilde{X}_E^\top \widetilde{Y}, \lambda(1-\alpha)\right) \tag{20}$$

      iv. Set $\beta_E^{(k+1)} \leftarrow \beta_E^{(k)(new)}$, $k \leftarrow k+1$

maximum penalty parameter (Lambda max)

2.2 Lambda max

The subgradient equations (11)–(13) can be used to determine the largest value of λ such that all coefficients are 0. From the subgradient Equation (11), we see that $\beta_E = 0$ is a solution if

$$\frac{1}{w_E}\left|\frac{1}{n}\left(X_E + \sum_{j=1}^{p}\gamma_j(X_E \circ \mathbf{\Psi}_j)\boldsymbol{\theta}_j\right)^\top R_{(-E)}\right| \leq \lambda(1-\alpha) \tag{21}$$

From the subgradient Equation (12), we see that $\boldsymbol{\theta}_j = \mathbf{0}$ is a solution if

$$\frac{1}{w_j}\left\|\frac{1}{n}\left(\mathbf{\Psi}_j + \gamma_j\beta_E(X_E \circ \mathbf{\Psi}_j)\right)^\top R_{(-j)}\right\|_2 \leq \lambda(1-\alpha) \tag{22}$$

From the subgradient Equation (13), we see that $\gamma_j = 0$ is a solution if

$$\frac{1}{w_{jE}}\left|\frac{1}{n}\left(\beta_E(X_E \circ \mathbf{\Psi}_j)\boldsymbol{\theta}_j\right)^\top R_{(-jE)}\right| \leq \lambda\alpha \tag{23}$$

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

$$\lambda_{max} = \frac{1}{n(1-\alpha)}\max\left\{\frac{1}{w_E}\left(X_E + \sum_{j=1}^{p}\gamma_j(X_E \circ \mathbf{\Psi}_j)\boldsymbol{\theta}_j\right)^\top R_{(-E)}, \max_j \frac{1}{w_j}\left\|\left(\mathbf{\Psi}_j + \gamma_j\beta_E(X_E \circ \mathbf{\Psi}_j)\right)^\top R_{(-j)}\right\|_2\right\} \tag{24}$$

which reduces to

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

2.3 Weak Heredity

Our method can be easily adapted to enforce the weak heredity property:

$$\hat{\alpha}_{jE} \neq 0 \quad \Rightarrow \quad \hat{\beta}_j \neq 0 \quad \text{or} \quad \hat{\beta}_E \neq 0$$

That is, an interaction term can only be present if at least one of it's corresponding main effects is nonzero. To do so, we reparametrize the coefficients for the interaction terms in (2) as $\boldsymbol{\alpha}_j = \gamma_j(\beta_E \cdot \mathbf{1}_{m_j} + \boldsymbol{\theta}_j)$, where $\mathbf{1}_{m_j}$ is a vector of ones with dimension m_j (i.e. the length of $\boldsymbol{\theta}_j$). We defer the algorithm details for fitting the `sail` model with weak heredity in Section A.3 of the Appendix, as it is very similar to Algorithm 1 for the strong heredity `sail` model.

2.4 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 ~~way~~ 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 ~~can be applied mainly for~~ 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. ~~Adaptive~~ Allows users to apply the adaptive `sail`, similar to the adaptive lasso [19]

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 and provide an example in Section C.6 of the Appendix.

You need to define "adaptive" lasso

Algorithm 2 Adaptive `sail` algorithm

1. For a decreasing sequence $\lambda = \lambda_{max}, \ldots, \lambda_{min}$ 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: $\lambda^{[opt]} \in \{\lambda_{max}, \ldots, \lambda_{min}\}$
3. Let $\widehat{\beta}_E^{[opt]}, \widehat{\boldsymbol{\theta}}_j^{[opt]}$ and $\widehat{\boldsymbol{\tau}}_j^{[opt]}$ for $j = 1, \ldots, p$ be the coefficient estimates corresponding to the model at $\lambda^{[opt]}$
4. Set the weights to be
$$w_E = \left(\left|\widehat{\beta}_E^{[opt]}\right|\right)^{-1}, \; w_j = \left(\|\widehat{\boldsymbol{\theta}}_j^{[opt]}\|_2\right)^{-1}, \; w_{jE} = \left(\|\widehat{\boldsymbol{\tau}}_j^{[opt]}\|_2\right)^{-1} \text{ for } j = 1, \ldots, p$$
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

2.5 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 ~~want~~ desire. In the extreme case, ~~we can~~ 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 $\boldsymbol{X}$. This is a reasonable approach when all the variables are continuous. However, there are often ~~instances~~ situations when ~~our~~ the data contains a combination of categorical and continuous variables. In these ~~situations~~ 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 provide such an example in the `sail` package showcase in Section C.7 of the Appendix.

3 Simulation Study

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

3.1 Comparator Methods

Since there is no software that directly addresses our problem, we selected comparator methods based on the following criteria: 1) is a penalized regression method that can handle high-dimensional data ($n < p$); 2) considers linear, non-linear or interaction effects; and 3) has a software implementation in `R`. The selected methods can be grouped into three categories:

1. Linear main effects: `lasso` [22], `adaptive lasso` [19]
2. Linear interactions: `lassoBT` [16], `GLinternet` [13]
3. Non-linear main effects: `HierBasis` [23], `SPAM` [24], `gamsel` [25]

For `GLinternet` we specified the `interactionCandidates` argument as to only consider interactions between the environment and all other X variables. For all other methods we supplied $(\boldsymbol{X}, X_E)$ as the data matrix, 100 for the number of tuning parameters to fit, and used the default values otherwise[1]. `lassoBT` considers all pairwise interactions as there is no way for the user to restrict the search space. `SPAM` applies the same basis expansion to every column of the data matrix; we chose 5 basis spline functions. `HierBasis` and `gamsel` selects whether a term in an additive model is nonzero, linear, or a non-linear spline up to a specified max degrees of freedom per variable.

We compare the above listed methods with our main proposal, `sail`, as well as `adaptive sail` (Algorithm 2), `sail weak` which has the weak heredity property and `linear sail` as

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

described in Section 2.5. For each function f_j, we use a B-spline basis matrix with `degree=5` implemented in the `bs` function in `R` [26]. We center the environment variable and the basis functions before running the `sail` method.

3.2 Simulation Design

The covariates are simulated as follows. First, we generate $w_1, \ldots, w_p, u, v$ independently from a standard normal distribution truncated to the interval [0,1] for $i = 1, \ldots, n$. Then we set $x_j = (w_j + t \cdot u)/(1+t)$ for $j = 1, \ldots, 4$ and $x_j = (w_j + t \cdot v)/(1+t)$ for $j = 5, \ldots, p$, where the parameter t controls the amount of correlation among predictors. The first four variables are nonzero (i.e. active in the response), while the rest of the variables are zero (i.e. are noise variables). This leads to a compound symmetry correlation structure where $Corr(x_j, x_k) = t^2/(1+t^2)$, for $1 \leq j \leq 4, 1 \leq k \leq 4$, and $Corr(x_j, x_k) = t^2/(1+t^2)$, for $5 \leq j \leq p, 5 \leq k \leq p$, but the covariates of the nonzero and zero components are independent [27, 28]. We consider the case when $p = 1000$ and $t = 0$. 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.

1. Hierarchy

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

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

(b) Truth obeys weak hierarchy. Here, in addition to the interaction, the E variable

has its own main effect but the covariates X_3 and X_4 do not.

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

(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**

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

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

3. **Interactions**

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

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

The true component functions are the same as in [27, 28] and are given by $f_1(t) = 5t$, $f_2(t) = 3(2t-1)^2$, $f_3(t) = 4\sin(2\pi t)/(2 - \sin(2\pi t))$, $f_4(t) = 6(0.1\sin(2\pi t) + 0.2\cos(2\pi t) + 0.3\sin(2\pi t)^2 + 0.4\cos(2\pi t)^3 + 0.5\sin(2\pi t)^3)$. We set $\beta_E = 2$ and draw ε from a normal distribution with variance chosen such that the signal-to-noise ratio is 2. Using this setup, we generated 200 replications consisting of a training set of $n = 200$, a validation set of $n = 200$ and a test set of $n = 800$. The training set was used to fit the model and the

validation set was used to select the optimal tuning parameter corresponding to the minimum prediction mean squared error (MSE). Variable selection results including true positive rate, false positive rate and number of active variables (the number of variables with a non-zero coefficient estimate) were assessed on the training set and MSE was assessed on the test set.

3.3 Results

The test set MSE results for each of the five simulation scenarios are shown in Figure 3, while Figure 4 shows the mean true positive rate (TPR) vs. the mean false positive rate (FPR) ± 1 standard deviation (SD).

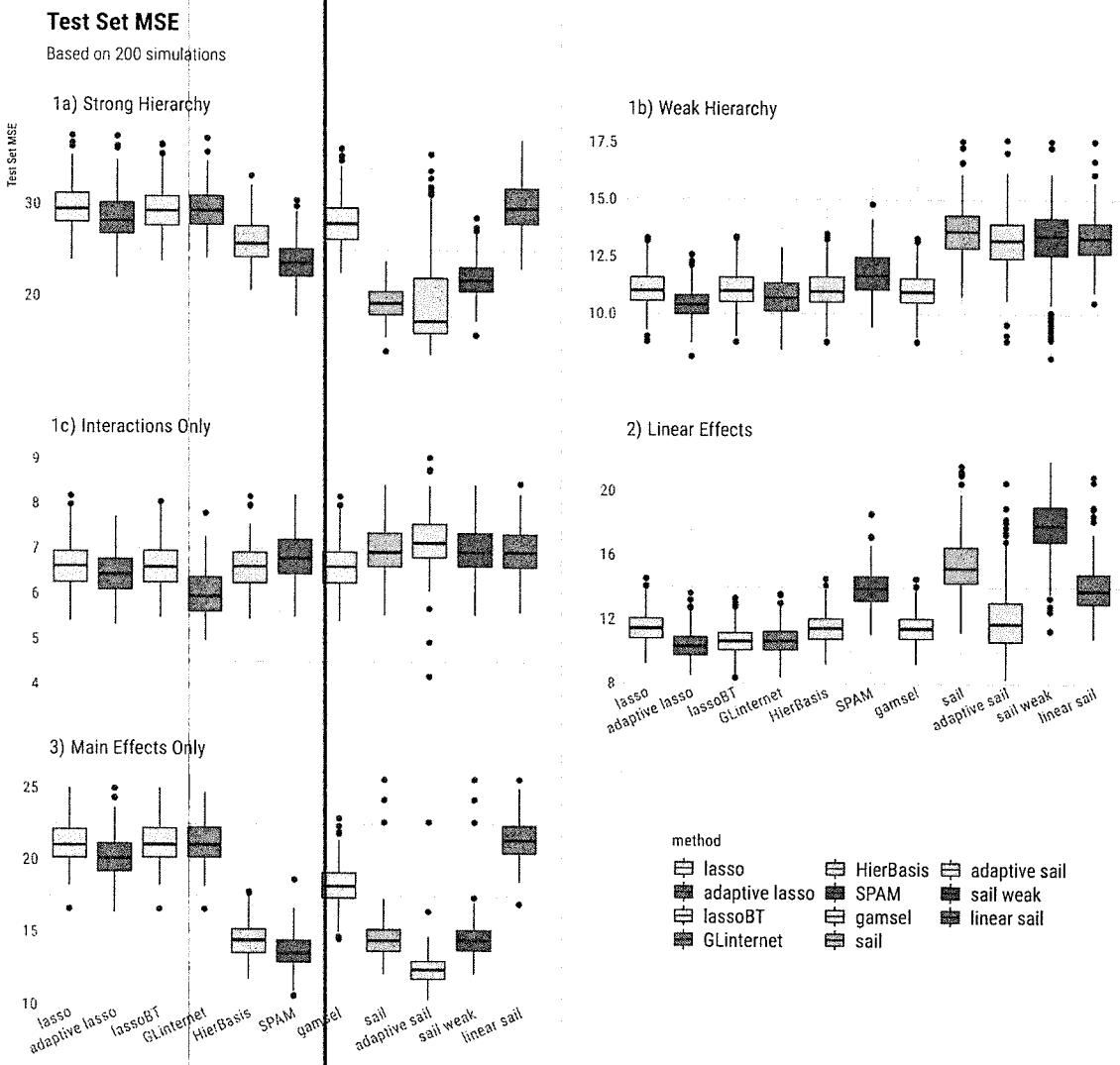

Figure 3: Boxplots of the test set mean squared error from 200 simulations for each of the five simulation scenarios.

We see that `sail`, `adaptive sail` and `sail weak` have the best performance in terms of both MSE and yielding correct sparse models when the truth follows strong hierarchy (scenario 1a), as we would expect, since this is exactly the scenario that our method is trying to target.

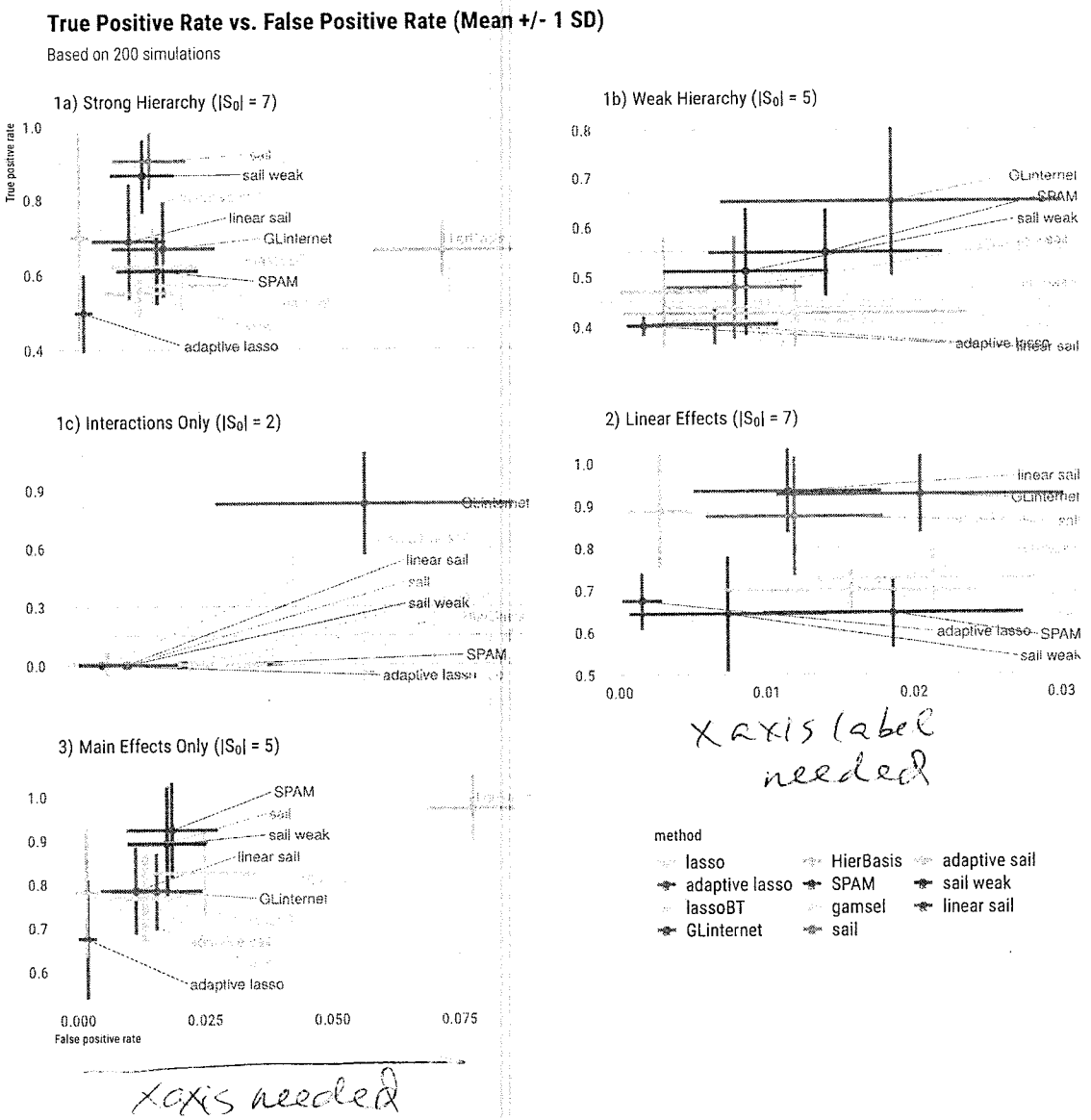

Figure 4: Means ± 1 standard deviation of true positive rate vs. false positive rate from 200 simulations for each of the five scenarios. $|S_0|$ is the number of truly associated variables.

Our method is also competitive when only main effects are present (scenario 3) and performs just as well as methods that only consider linear and non-linear main effects (`HierBasis`, `SPAM`), owing to the penalization applied to the interaction parameter. Due to the heredity property, our method is unable to capture any of the truly associated variables when only interactions are present (scenario 1c). However, the other methods also fail to capture any

signal, with the exception of `GLinternet` which has a high TPR and FPR. When only linear effects and interactions are present (scenario 2), we see that `linear sail` has a high TPR and low FPR as compared to the other linear interaction methods (`lassoBT` and `GLinternet`) though the test set MSE isn't as good. The `lasso` and `adaptive lasso` have good test set MSE performance but poor sensitivity. Additional results are available in Section B of the Appendix. Specifically, in Figure B.1 we plot the mean MSE against the mean number of active variables ± 1 standard deviation (SD). Figures B.2 and B.3 show the true positive and false positive rates, respectively. Figure B.4 shows the number of active variables.

We visually inspected whether our method could correctly capture the shape of the association between the predictors and the response for both main and interaction effects. To do so, we plotted the true and predicted curves for scenario 1a) only. Figure 5 shows each of the four main effects with the estimated curves from each of the 200 simulations along with the true curve. We can see the effect of the penalty on the parameters, i.e., decreasing prediction variance at the cost of increased bias. This is particularly well illustrated in the bottom right panel where `sail` smooths out the very wiggly component function $f_4(x)$.

To visualize the estimated interaction effects, we ordered the 200 simulation runs by the euclidean distance between the estimated and true regression functions. Following Radchenko et al. [8], we then identified the 25th, 50th, and 75th best simulations and plotted, in Figures 6 and 7, the interaction effects of X_E with $f_3(X_3)$ and $f_4(X_4)$, respectively. We see that `sail` does a good job at capturing the true interaction surface for $X_E \cdot f_3(X_3)$. Again, the smoothing and shrinkage effect is apparent when looking at the interaction surfaces for $X_E \cdot f_4(X_4)$

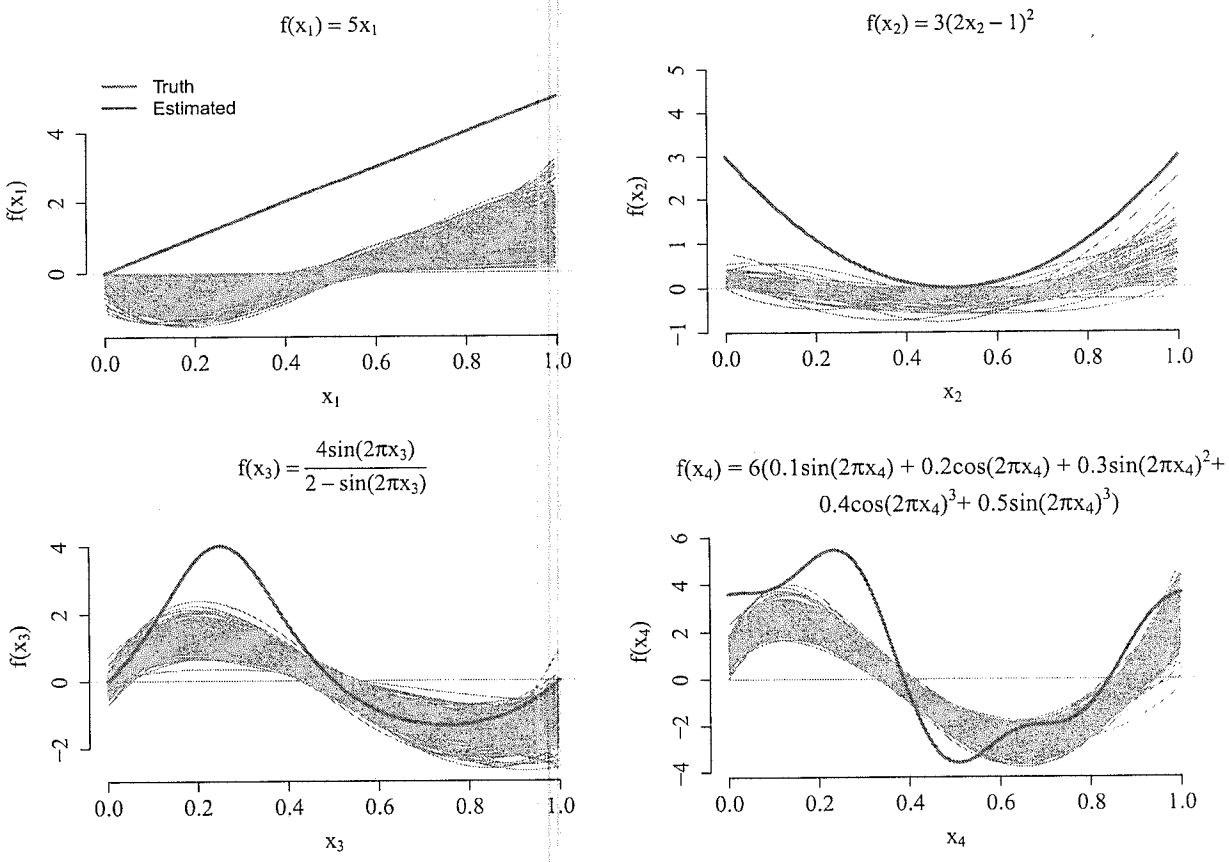

Figure 5: True and estimated main effect component functions for scenario 1a). The estimated curves represent the results from each one of the 200 simulations conducted.

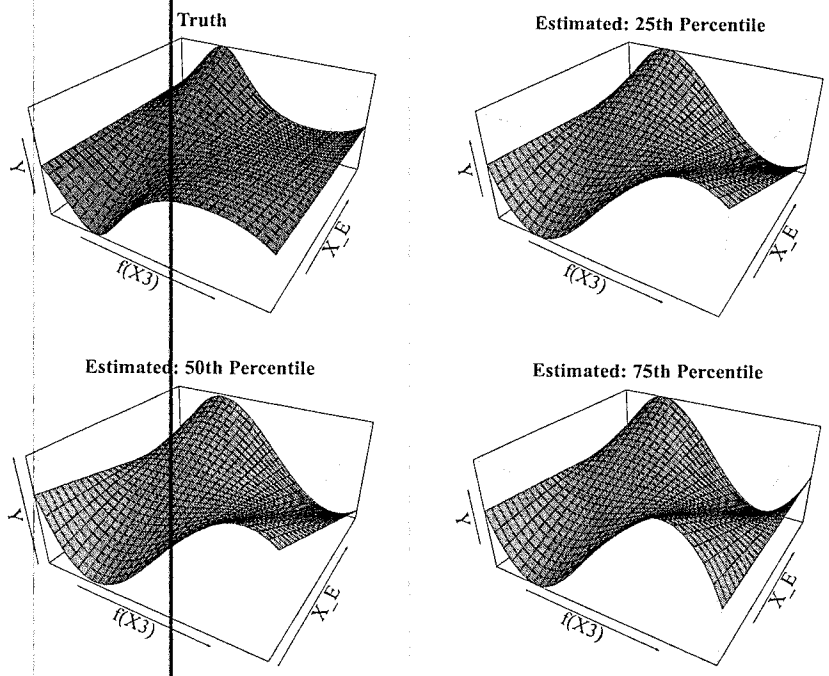

Figure 6: True and estimated interaction effects for $X_E \cdot f_3(X_3)$ in simulation scenario 1a).

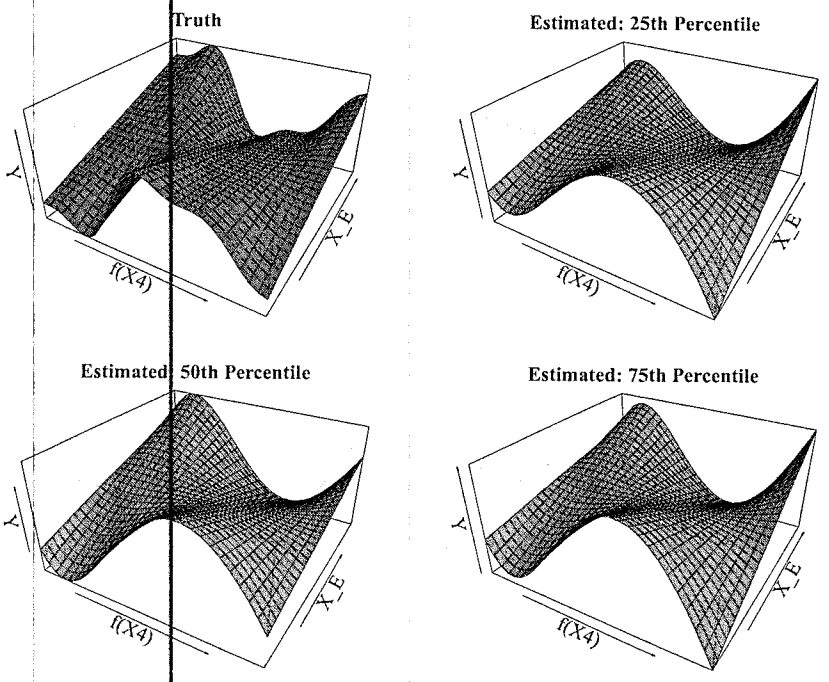

Figure 7: True and estimated interaction effects for $X_E \cdot f_4(X_4)$ in simulation scenario 1a).

4 Real Data Application

In this section we illustrate `sail` on several real data examples.

4.1 Alzheimer's Disease Neuroimaging Initiative

Alzheimer's is an irreversible neurodegenerative disease that results in a loss of mental function due to the deterioration of brain tissue. The overall goal of the Alzheimer's Disease Neuroimaging Initiative (ADNI) is to validate biomarkers for use in Alzheimer's disease clinical treatment trials [29]. The patients were selected into the study based on their clinical diagnosis: controls, mild cognitive impairment (MCI) or Alzheimer's disease (AD). PET amyloid imaging was used to asses amyloid beta ($A\beta$) protein load in 96 brain regions. The response ~~was~~ we use here is general cognitive decline, as measured by a continuous mini-mental state examination score. We applied `sail` to this data to see if there were any non-linear interactions between clinical diagnosis and $A\beta$ protein in the 96 brain regions on mini-mental state examination.

There were a total of 343 patients who we divided randomly into equal sized training/validation/test splits. We ran the strong heredity `sail` with cubic B-splines and $\alpha = 0.1$. We also applied the `lasso`, `lassoBT`, `HierBasis` and `GLinternet` to this data. Using the same default settings and strategy as the simulation study, we ran each method on the training data, determined the optimal tuning parameter on the validation data, and assessed MSE on the test data. We repeated this process 200 times.

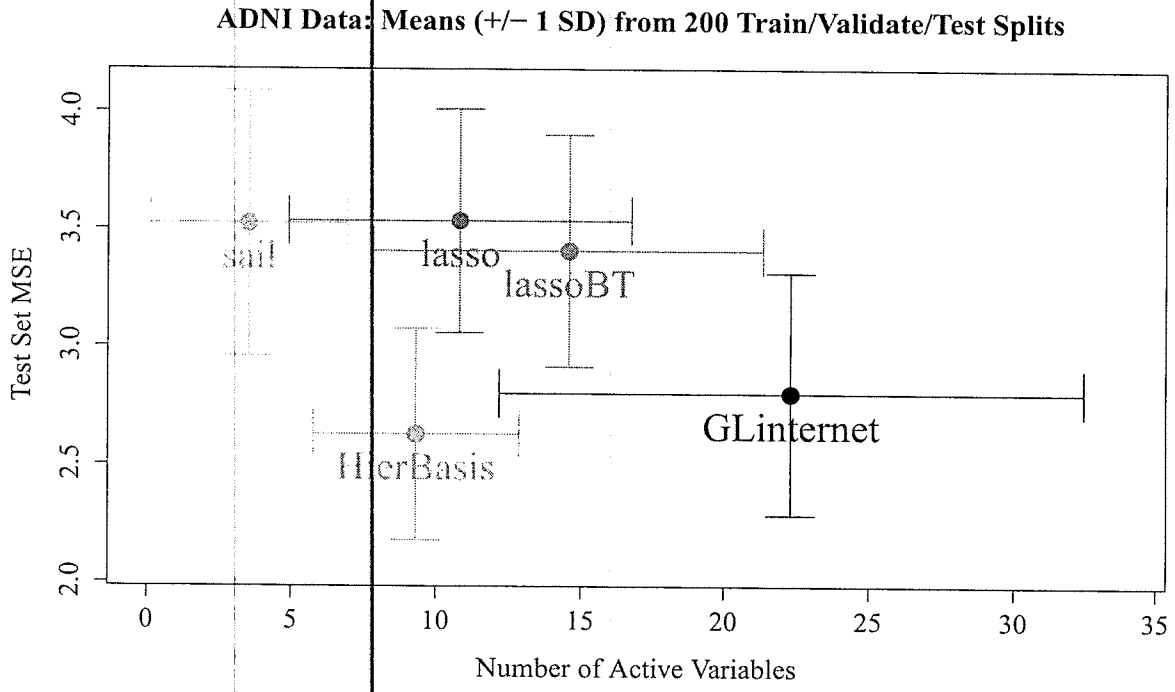

Figure 8: Mean test set MSE vs. mean number of active variables (±1 SD).

In Figure 8 we plot the mean test set MSE vs. the mean number of active variables ±1 SD. We see that `sail` produces the sparsest models but doesn't perform as well as `HierBasis` and `GLinternet` in terms of MSE. `sail` achieves a similar MSE to both the `lasso` and `lassoBT` with fewer variables on average. `GLinternet` produces the largest models and seems to be sensitive to the train/validate/test split as evidenced by the large standard deviations.

To visualize the results from the `sail` method, we chose the train/validate/test split which led to the best test set MSE, and then plotted the interaction effects in Figure 9. The left panel shows the middle occipital gyrus left region in the occipital lobe known for visual object perception. We see that more Aβ protein loads leads to a worse cognitive score for the MCI and AD group but not for the controls. The right panel shows the cuneus region known which is known to be involved in basic visual processing, and We see that more Aβ proteins leads to better cognitive scores for the MCI and AD group and poorer scores for

the controls.

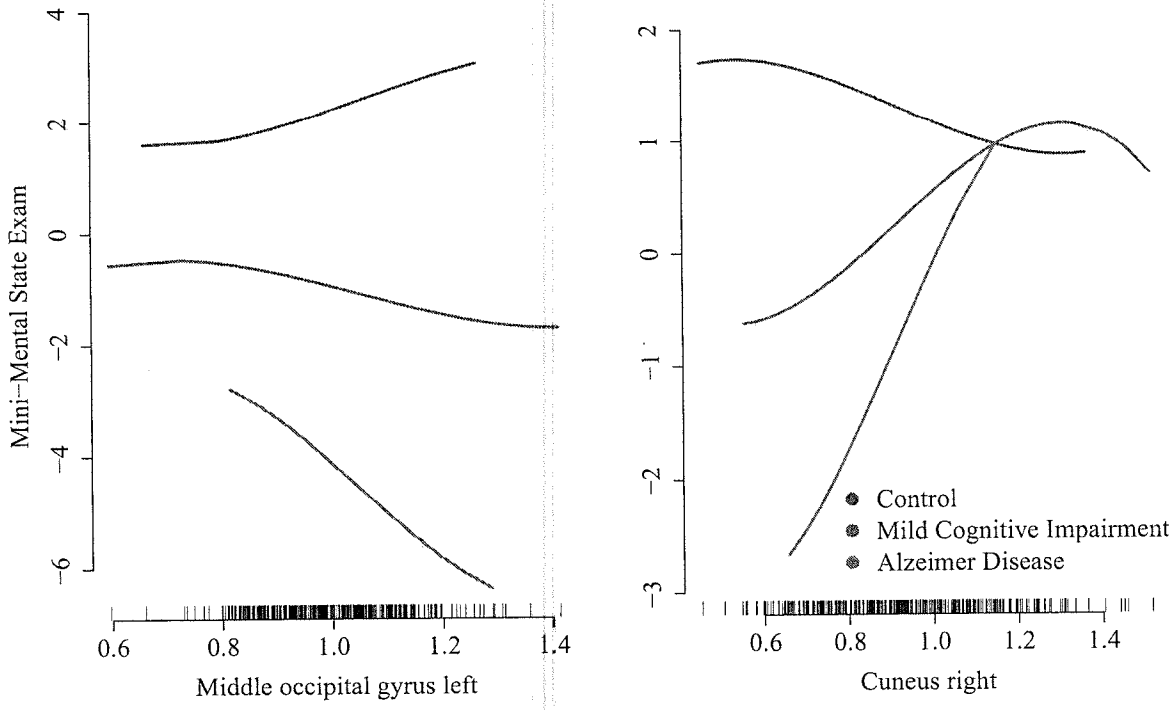

Figure 9: Estimated interaction effects by `sail` for ADNI data.

5 Discussion

In this article we introduce 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 the weak and strong heredity property without using a complex penalty function. We use a blockwise coordinate descent algorithm to solve the `sail` objective function for both least-squares and logistic loss functions. We show that the `adaptive sail` has the oracle property. All our algorithms are implemented in a computationally efficient, well-documented and freely available `R` package. Furthermore, our method is flexible 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 error, sensitivity and specificity when there are non-linear main effects only, as well as interactions with an exposure variable.

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.

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.