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 20, 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 are several applications where interaction models can reflect biological phenomena and improve prediction accuracy. For example, genome wide association studies (GWAS) have been unable to explain a large proportion of heritability (the variance in phenotype attributable to genetic variants) and it has been suggested that this missing heritability may in part be due to gene-environment interactions [1]. Furthermore, diseases are now thought to be the result of changes in entire biological networks

whose states are affected by a complex interaction of genetic and environmental factors. In high-dimensional settings ($p >> n$), power to estimate interactions is low, the number of possible interactions could be enormous and their effects may be non-linear. In this paper, we propose a multivariable penalization procedure for detecting non-linear interactions between high dimensional data $\mathbf{X}$ and a single environmental factor E on a response vector Y.

1.1 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. 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 [2]. 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 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 $\boldsymbol{\Theta} \coloneqq (\beta_0, \beta_E, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p, \boldsymbol{\tau}_1, \ldots, \boldsymbol{\tau}_p)$ and $\mathbf{D} \coloneqq (Y, \boldsymbol{\Psi}, X_E)$ 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 penalized least squares criterion:

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

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.

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 [3]. This is known as the strong heredity principle [4]. Indeed, large main effects are more likely to lead to detectable interactions [5]. In the next section we discuss how a simple reparametrization of the model (5) can lead to this desirable property.

1.2 Strong and weak heredity

The strong heredity principle states that the interaction term can only have a non-zero estimate if its corresponding main effects are estimated to be non-zero. The weak heredity principle allows for a non-zero interaction estimate as long as one of the corresponding main effects are estimated to be non-zero [4]. In the context of penalized regression methods, these principles can be formulated as structured sparsity [6] problems. Several authors have proposed to modify the type of penalty in order to achieve the heredity principle [7, 8**? ?**]. We take an alternative approach. Following Choi et al. [9], we introduce a parameter $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p) \in \mathbb{R}^p$ 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.

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

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

To perform variable selection in this new parametrization, we penalize $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)$ instead of penalizing $\boldsymbol{\tau}$ as in (5), leading to the following penalized least squares criterion:

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

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. We sample $p = 20$ covariates independently from a $N(0, 1)$ truncated to the interval [0,1] and sample size $n = 100$. We generated data from a model which follows the strong heredity principle:

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

where $f_1(x) = -3x$, $f_2(x) = 2(2x - 1)^3$ and the error term ε is generated from a normal distribution with variance chosen such that the signal-to-noise ratio (SNR) is 2. We run the strong heredity `sail` method with cubic b-splines and 10-fold CV 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 **??** and we color the lines corresponding 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.

In Figure **??**, we plot the true and estimated component functions $\hat{f}_1(X_1)$ and $E \cdot \hat{f}_2(X_2)$. Our method is able to capture the correct functional form for the nonlinear interaction. The lack-of-fit for $f_1(X_1)$ can be explained by the fact that `sail` is trying to fit a cubic spline to a linear function. Nevertheless, this example demonstrates that it can still identify linear associations.

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 $\boldsymbol{X}$ [8, 9, 10, 11] 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 [12, 13] 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 [14, 15] in the linear setting. There are many fewer methods avaliable for non-linear interactions. For example, Radchenko and James (2010) [7] 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 [16] 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 [17], 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 [18] 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 Regularization Path

The `sail` model has the form

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

The objective function is given by

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

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

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

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

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

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

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:

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

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

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

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 $j = 1, \ldots, p$

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

From the subgradient Equation (11), we see that $\beta_E = 0$ is a solution if

$$\frac{1}{w_E} \left| \frac{1}{n} \left(X_E + \sum_{j=1}^{p} \gamma_j (X_E \circ \boldsymbol{\Psi}_j) \boldsymbol{\theta}_j \right)^\top R_{(-E)} \right| \leq \lambda(1 - \alpha) \tag{14}$$

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

$$\frac{1}{w_j} \left\| \frac{1}{n} \left(\boldsymbol{\Psi}_j + \gamma_j \beta_E (X_E \circ \boldsymbol{\Psi}_j) \right)^\top R_{(-j)} \right\|_2 \leq \lambda(1 - \alpha) \tag{15}$$

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

$$\frac{1}{w_{jE}} \left| \frac{1}{n} \left(\beta_E (X_E \circ \boldsymbol{\Psi}_j) \boldsymbol{\theta}_j \right)^\top R_{(-jE)} \right| \leq \lambda\alpha \tag{16}$$

2.1 Lambda Max

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 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:

$$\lambda_{max} = \frac{1}{n(1-\alpha)} \max \left\{ \frac{1}{w_E} \left(X_E + \sum_{j=1}^{p} \gamma_j (X_E \circ \mathbf{\Psi}_j) \boldsymbol{\theta}_j \right)^\top R_{(-E)}, \max_j \frac{1}{w_j} \left\| (\mathbf{\Psi}_j + \gamma_j \beta_E (X_E \circ \mathbf{\Psi}_j))^\top R_{(-j)} \right\|_2 \right\} \quad (17)$$

which reduces to

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

2.2 Optimization of Parameters

From the subgradient equations we see that

$$\hat{\beta}_0 = \left(Y - \sum_{j=1}^{p} \mathbf{\Psi}_j \hat{\boldsymbol{\theta}}_j - \hat{\beta}_E X_E - \sum_{j=1}^{p} \hat{\gamma}_j \hat{\beta}_E (X_E \circ \mathbf{\Psi}_j) \hat{\boldsymbol{\theta}}_j \right)^\top \mathbf{1} \quad (18)$$

$$\hat{\beta}_E = S \left(\frac{1}{n \cdot w_E} \left(X_E + \sum_{j=1}^{p} \hat{\gamma}_j (X_E \circ \mathbf{\Psi}_j) \hat{\boldsymbol{\theta}}_j \right)^\top R_{(-E)}, \lambda(1-\alpha) \right) \quad (19)$$

$$\lambda(1-\alpha) w_j \frac{\boldsymbol{\theta}_j}{\|\boldsymbol{\theta}_j\|_2} = \frac{1}{n} (\mathbf{\Psi}_j + \gamma_j \beta_E (X_E \circ \mathbf{\Psi}_j))^\top R_{(-j)} \quad (20)$$

$$\hat{\gamma}_j = S \left(\frac{1}{n \cdot w_{jE}} (\beta_E (X_E \circ \mathbf{\Psi}_j) \boldsymbol{\theta}_j)^\top R_{(-jE)}, \lambda\alpha \right) \quad (21)$$

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

3 Algorithm

For each function f_j, we use a cubic B-spline parameterization with 5 degrees of freedom implemented in the `bs` function in `R` [19].

3.1 Details on update for $\boldsymbol{\theta}$

Here we discuss a computational speedup in the updates for the $\boldsymbol{\theta}$ parameter. The partial residual (R_s) used for updating $\boldsymbol{\theta}_s$ ($s \in 1, \ldots, p$) at the kth iteration is given by

$$R_s = Y - \widetilde{Y}^{(k)}_{(-s)} \tag{22}$$

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

$$\widetilde{Y}^{(k)}_{(-s)} = \beta_0^{(k)} - \beta_E^{(k)} X_E - \sum_{\ell \neq s} \boldsymbol{\Psi}_\ell \boldsymbol{\theta}_\ell^{(k)} - \sum_{\ell \neq s} \gamma_\ell^{(k)} \beta_E^{(k)} \widetilde{\boldsymbol{\Psi}}_\ell \boldsymbol{\theta}_\ell^{(k)} \tag{23}$$

Using (23), (22) can be re-written as

$$\begin{aligned} R_s &= Y - \beta_0^{(k)} - \beta_E^{(k)} X_E - \sum_{j=1}^{p} (\boldsymbol{\Psi}_j + \gamma_j^{(k)} \beta_E^{(k)} \widetilde{\boldsymbol{\Psi}}_j) \boldsymbol{\theta}_j^{(k)} + (\boldsymbol{\Psi}_s + \gamma_s^{(k)} \beta_E^{(k)} \widetilde{\boldsymbol{\Psi}}_s) \boldsymbol{\theta}_s^{(k)} \\ &= R^* + (\boldsymbol{\Psi}_s + \gamma_s^{(k)} \beta_E^{(k)} \widetilde{\boldsymbol{\Psi}}_s) \boldsymbol{\theta}_s^{(k)} \end{aligned} \tag{24}$$

where

$$R^* = Y - \beta_0^{(k)} - \beta_E^{(k)} X_E - \sum_{j=1}^{p} (\boldsymbol{\Psi}_j + \gamma_j^{(k)} \beta_E^{(k)} \widetilde{\boldsymbol{\Psi}}_j) \boldsymbol{\theta}_j^{(k)} \tag{25}$$

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

$$\boldsymbol{\theta}_s^{(k)(new)} = \arg\min_{\boldsymbol{\theta}_j} \frac{1}{2n} \left\| R_s - (\boldsymbol{\Psi}_s + \gamma_s^{(k)} \beta_E^{(k)} \widetilde{\boldsymbol{\Psi}}_s) \boldsymbol{\theta}_j \right\|_2^2 + \lambda(1-\alpha) w_s \|\theta_j\|_2 \tag{26}$$

Algorithm 1 Coordinate descent for least-squares `sail` with strong heredity

function `sail`$(Y, \boldsymbol{X}, X_E, \texttt{df}, \texttt{degree}, \epsilon)$ ▷ Algorithm for solving (9)

  $\Psi_j \leftarrow$ `splines::bs`$(X_j,$ `df`, `degree`$)$ for $j = 1, \ldots, p$

  $\widetilde{\Psi}_j \leftarrow X_E \circ \Psi_j$ for $j = 1, \ldots, p$

  Initialize: $\beta_0^{(0)} \leftarrow \bar{Y}$, $\beta_E^{(0)} = \boldsymbol{\theta}_j^{(0)} \leftarrow 0$ for $j = 1, \ldots, p$.

  Set iteration counter $k \leftarrow 0$

  $R^* \leftarrow Y - \beta_0^{(k)} - \beta_E^{(k)} X_E - \sum_j (\boldsymbol{\Psi}_j + \gamma_j^{(k)} \beta_E^{(k)} \widetilde{\boldsymbol{\Psi}}_j) \boldsymbol{\theta}_j^{(k)}$

  repeat

    • To update $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)$

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

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

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

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

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

    • To update $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p)$

      $\widetilde{X}_j \leftarrow \boldsymbol{\Psi}_j + \gamma_j^{(k)} \beta_E^{(k)} \widetilde{\boldsymbol{\Psi}}_j$ for $j = 1, \ldots, p$

      for $j = 1, \ldots, p$ **do**

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

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

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

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

    • To update β_E

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

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

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

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

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

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

    • To update β_0

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

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

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

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

    $k \leftarrow k + 1$

  until convergence criterion is satisfied: $\left\| \boldsymbol{\Theta}^{(k)} - \boldsymbol{\Theta}^{(k-1)} \right\|_2^2 < \epsilon$

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

$$R_{s+1} = R^* + (\boldsymbol{\Psi}_{s+1} + \gamma_{s+1}^{(k)} \beta_E^{(k)} \widetilde{\boldsymbol{\Psi}}_{s+1}) \boldsymbol{\theta}_{s+1}^{(k)} + (\boldsymbol{\Psi}_s + \gamma_s^{(k)} \beta_E^{(k)} \widetilde{\boldsymbol{\Psi}}_s)(\boldsymbol{\theta}_s^{(k)} - \boldsymbol{\theta}_s^{(k)(new)}) \tag{27}$$

where R^* is given by (25), $\boldsymbol{\theta}_s^{(k)}$ is the parameter value prior to the update, and $\boldsymbol{\theta}_s^{(k)(new)}$ is the updated value given by (26). Taking the difference between (24) and (27) gives

$$\begin{aligned} \Delta &= R_t - R_s \\ &= (\boldsymbol{\Psi}_t + \gamma_t^{(k)} \beta_E^{(k)} \widetilde{\boldsymbol{\Psi}}_t) \boldsymbol{\theta}_t^{(k)} + (\boldsymbol{\Psi}_s + \gamma_s^{(k)} \beta_E^{(k)} \widetilde{\boldsymbol{\Psi}}_s)(\boldsymbol{\theta}_s^{(k)} - \boldsymbol{\theta}_s^{(k)(new)}) - (\boldsymbol{\Psi}_s + \gamma_s^{(k)} \beta_E^{(k)} \widetilde{\boldsymbol{\Psi}}_s) \boldsymbol{\theta}_s^{(k)} \\ &= (\boldsymbol{\Psi}_t + \gamma_t^{(k)} \beta_E^{(k)} \widetilde{\boldsymbol{\Psi}}_t) \boldsymbol{\theta}_t^{(k)} - (\boldsymbol{\Psi}_s + \gamma_s^{(k)} \beta_E^{(k)} \widetilde{\boldsymbol{\Psi}}_s) \boldsymbol{\theta}_s^{(k)(new)} \end{aligned} \tag{28}$$

Therefore $R_t = R_s + \Delta$, and the partial residual for updating the next predictor can be computed by updating the previous partial residual by Δ, given by (28). This formulation can lead to computational speedups especially when $\Delta = 0$, meaning the partial residual does not need to be re-calculated.

4 Weak Heredity

We can also enforce the weak heredity property:

$$\hat{\alpha}_{jE} \neq 0 \qquad \Rightarrow \qquad \hat{\beta}_j \neq 0 \qquad \text{or} \qquad \hat{\beta}_E \neq 0 \tag{29}$$

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 $\boldsymbol{\theta}_j$).

4.1 Regularization Path

The `sail` model with weak heredity has the form

$$\hat{Y} = \beta_0 \cdot \mathbf{1} + \sum_{j=1}^{p} \mathbf{\Psi}_j \boldsymbol{\theta}_j + \beta_E X_E + \sum_{j=1}^{p} \gamma_j (X_E \circ \mathbf{\Psi}_j)(\beta_E \cdot \mathbf{1}_{m_j} + \boldsymbol{\theta}_j) \tag{30}$$

The objective function is given by

$$Q(\mathbf{\Theta}) = \frac{1}{2n} \left\| Y - \hat{Y} \right\|_2^2 + \lambda(1-\alpha) \left(w_E |\beta_E| + \sum_{j=1}^{p} w_j \|\boldsymbol{\theta}_j\|_2 \right) + \lambda\alpha \sum_{j=1}^{p} w_{jE} |\gamma_j| \tag{31}$$

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

$$\frac{\partial Q}{\partial \beta_0} = \frac{1}{n} \left(Y - \beta_0 \cdot \mathbf{1} - \sum_{j=1}^{p} \mathbf{\Psi}_j \boldsymbol{\theta}_j - \beta_E X_E - \sum_{j=1}^{p} \gamma_j (X_E \circ \mathbf{\Psi}_j)(\beta_E \cdot \mathbf{1}_{m_j} + \boldsymbol{\theta}_j) \right)^\top \mathbf{1} = 0 \tag{32}$$

$$\frac{\partial Q}{\partial \beta_E} = -\frac{1}{n} \left(X_E + \sum_{j=1}^{p} \gamma_j (X_E \circ \mathbf{\Psi}_j) \mathbf{1}_{m_j} \right)^\top R + \lambda(1-\alpha) w_E s_1 = 0 \tag{33}$$

$$\frac{\partial Q}{\partial \boldsymbol{\theta}_j} = -\frac{1}{n} \left(\mathbf{\Psi}_j + \gamma_j (X_E \circ \mathbf{\Psi}_j) \right)^\top R + \lambda(1-\alpha) w_j s_2 = \mathbf{0} \tag{34}$$

$$\frac{\partial Q}{\partial \gamma_j} = -\frac{1}{n} \left((X_E \circ \mathbf{\Psi}_j)(\beta_E \cdot \mathbf{1}_{m_j} + \boldsymbol{\theta}_j) \right)^\top R + \lambda\alpha w_{jE} s_3 = 0 \tag{35}$$

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

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

the partial residual without X_E as

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

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

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

From the subgradient Equation (33), we see that $\beta_E = 0$ is a solution if

$$\frac{1}{w_E} \left| \frac{1}{n} \left(X_E + \sum_{j=1}^{p} \gamma_j (X_E \circ \mathbf{\Psi}_j) \mathbf{1}_{m_j} \right)^\top R_{(-E)} \right| \leq \lambda(1 - \alpha) \tag{36}$$

From the subgradient Equation (34), we see that $\boldsymbol{\theta}_j = \mathbf{0}$ is a solution if

$$\frac{1}{w_j} \left\| \frac{1}{n} \left(\mathbf{\Psi}_j + \gamma_j (X_E \circ \mathbf{\Psi}_j) \right)^\top R_{(-j)} \right\|_2 \leq \lambda(1 - \alpha) \tag{37}$$

From the subgradient Equation (35), we see that $\gamma_j = 0$ is a solution if

$$\frac{1}{w_{jE}} \left| \frac{1}{n} \left((X_E \circ \mathbf{\Psi}_j)(\beta_E \cdot \mathbf{1}_{m_j} + \boldsymbol{\theta}_j) \right)^\top R_{(-jE)} \right| \leq \lambda\alpha \tag{38}$$

4.2 Lambda Max

The smallest value of λ for which the entire parameter vector $(\beta_E, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p, \gamma_1, \ldots, \gamma_p)$ is $\mathbf{0}$ is:

$$\lambda_{max} = \frac{1}{n} \max \left\{ \frac{1}{(1-\alpha)w_E} \left(X_E + \sum_{j=1}^{p} \gamma_j (X_E \circ \boldsymbol{\Psi}_j) \mathbf{1}_{m_j} \right)^\top R_{(-E)}, \right.$$
$$\max_j \frac{1}{(1-\alpha)w_j} \left\| (\boldsymbol{\Psi}_j + \gamma_j (X_E \circ \boldsymbol{\Psi}_j))^\top R_{(-j)} \right\|_2,$$
$$\left. \max_j \frac{1}{\alpha w_{jE}} \left((X_E \circ \boldsymbol{\Psi}_j)(\beta_E \cdot \mathbf{1}_{m_j} + \boldsymbol{\theta}_j) \right)^\top R_{(-jE)} \right\} \quad (39)$$

which reduces to

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

4.3 Optimization of Parameters

From the subgradient equations we see that

$$\hat{\beta}_0 = \left(Y - \sum_{j=1}^{p} \boldsymbol{\Psi}_j \hat{\boldsymbol{\theta}}_j - \hat{\beta}_E X_E - \sum_{j=1}^{p} \hat{\gamma}_j (X_E \circ \boldsymbol{\Psi}_j)(\hat{\beta}_E \cdot \mathbf{1}_{m_j} + \hat{\boldsymbol{\theta}}_j) \right)^\top \mathbf{1} \quad (40)$$

$$\hat{\beta}_E = S \left(\frac{1}{n \cdot w_E} \left(X_E + \sum_{j=1}^{p} \hat{\gamma}_j (X_E \circ \boldsymbol{\Psi}_j) \mathbf{1}_{m_j} \right)^\top R_{(-E)}, \lambda(1-\alpha) \right) \quad (41)$$

$$\lambda(1-\alpha) w_j \frac{\boldsymbol{\theta}_j}{\|\boldsymbol{\theta}_j\|_2} = \frac{1}{n} (\boldsymbol{\Psi}_j + \gamma_j (X_E \circ \boldsymbol{\Psi}_j))^\top R_{(-j)} \quad (42)$$

$$\hat{\gamma}_j = S \left(\frac{1}{n \cdot w_{jE}} \left((X_E \circ \boldsymbol{\Psi}_j)(\beta_E \cdot \mathbf{1}_{m_j} + \boldsymbol{\theta}_j) \right)^\top R_{(-jE)}, \lambda\alpha \right) \quad (43)$$

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

4.4 Algorithm

5 Simulations

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. 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 [20, 21]

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.

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

   (b) Truth obeys weak hierarchy.

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

   (c) Truth only has interactions.

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

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

function `sail`$(Y, \boldsymbol{X}, X_E, \texttt{df}, \texttt{degree}, \epsilon)$ ▷ Algorithm for solving (31)

  $\Psi_j \leftarrow$ `splines::bs`$(X_j,$ `df`, `degree`$)$ for $j = 1, \ldots, p$

  $\widetilde{\Psi}_j \leftarrow X_E \circ \Psi_j$ for $j = 1, \ldots, p$

  Initialize: $\beta_0^{(0)} \leftarrow \bar{Y}$, $\beta_E^{(0)} = \boldsymbol{\theta}_j^{(0)} \leftarrow 0$ for $j = 1, \ldots, p$.

  Set iteration counter $k \leftarrow 0$

  $R^* \leftarrow Y - \beta_0^{(k)} - \beta_E^{(k)} X_E - \sum_j \boldsymbol{\Psi}_j \boldsymbol{\theta}_j^{(k)} - \sum_j \gamma_j^{(k)} \widetilde{\boldsymbol{\Psi}}_j (\beta_E^{(k)} \cdot \mathbf{1}_{m_j} + \boldsymbol{\theta}_j^{(k)})$

  repeat

    • To update $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)$

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

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

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

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

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

    • To update $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p)$

      $\widetilde{X}_j \leftarrow \boldsymbol{\Psi}_j + \gamma_j^{(k)} \widetilde{\boldsymbol{\Psi}}_j$ for $j = 1, \ldots, p$

      for $j = 1, \ldots, p$ **do**

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

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

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

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

    • To update β_E

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

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

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

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

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

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

    • To update β_0

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

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

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

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

    $k \leftarrow k + 1$

  until convergence criterion is satisfied: $\left\| \boldsymbol{\Theta}^{(k)} - \boldsymbol{\Theta}^{(k-1)} \right\|_2^2 < \epsilon$

2. **Non-linearity**

   (a) Truth is linear

$$Y = \sum_{j=1}^{4} \beta_j X_j + \beta_E \cdot X_E + X_E \cdot X_3 + X_E \cdot X_4 + \varepsilon$$

3. **Interactions**

   (a) Truth only has main effects

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

The true component functions are the same as in [20, 21]. Using this setup, we generated 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. Performance was assessed on the test set. Results are shown in Figure **??**.

6 Real Data Application

References

[1] Teri A Manolio, Francis S Collins, Nancy J Cox, David B Goldstein, Lucia A Hindorff, David J Hunter, Mark I McCarthy, Erin M Ramos, Lon R Cardon, Aravinda Chakravarti, et al. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, 2009. 1

[2] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC Press, 2015. 2

[3] Peter McCullagh and John A Nelder. *Generalized linear models*, volume 37. CRC press, 1989. 4

[4] Hugh Chipman. Bayesian variable selection with related predictors. *Canadian Journal of Statistics*, 24(1):17–36, 1996. 4

[5] David R Cox. Interaction. *International Statistical Review/Revue Internationale de Statistique*, pages 1–24, 1984. 4

[6] Francis Bach, Rodolphe Jenatton, Julien Mairal, Guillaume Obozinski, et al. Structured sparsity through convex optimization. *Statistical Science*, 27(4):450–468, 2012. 4

[7] Peter Radchenko and Gareth M James. Variable selection using adaptive nonlinear interaction structures in high dimensions. *Journal of the American Statistical Association*, 105(492):1541–1553, 2010. 4, 6

[8] Jacob Bien, Jonathan Taylor, Robert Tibshirani, et al. A lasso for hierarchical interactions. *The Annals of Statistics*, 41(3):1111–1141, 2013. 4, 6

[9] Nam Hee Choi, William Li, and Ji Zhu. Variable selection with the strong heredity constraint and its oracle property. *Journal of the American Statistical Association*, 105(489):354–364, 2010. 4, 6

[10] Peng Zhao, Guilherme Rocha, and Bin Yu. The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, pages 3468–3497, 2009. 6

[11] Yiyuan She and He Jiang. Group regularized estimation under structural hierarchy. *arXiv preprint arXiv:1411.4691*, 2014. 6

[12] Michael Lim and Trevor Hastie. Learning interactions via hierarchical group-lasso regularization. *Journal of Computational and Graphical Statistics*, 24(3):627–654, 2015. 6

[13] Asad Haris, Daniela Witten, and Noah Simon. Convex modeling of interactions with strong heredity. *Journal of Computational and Graphical Statistics*, 25(4):981–1004, 2016. 6

[14] Ning Hao, Yang Feng, and Hao Helen Zhang. Model selection for high-dimensional quadratic regression via regularization. *Journal of the American Statistical Association*, pages 1–11, 2018. 6

[15] Rajen D Shah. Modelling interactions in high-dimensional data with backtracking. *Journal of Machine Learning Research*, 17(207):1–31, 2016. 6

[16] Robert Tibshirani and Jerome Friedman. A pliable lasso. *arXiv preprint arXiv:1712.00484*, 2017.

[17] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

[18] Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.

[19] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017. 11

[20] Yi Lin, Hao Helen Zhang, et al. Component selection and smoothing in multivariate nonparametric regression. *The Annals of Statistics*, 34(5):2272–2297, 2006. 17, 19

[21] Jian Huang, Joel L Horowitz, and Fengrong Wei. Variable selection in nonparametric additive models. *Annals of statistics*, 38(4):2282, 2010. 17, 19