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 2, 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 linear and non-linear interactions between high dimensional data $\mathbf{X}$ and a single environmental factor E on a response vector Y.

1.1 Linear interaction model

We first consider a regression model for an outcome variable $Y = (Y_1, \ldots, Y_n)$ where n is the number of subjects. Let $E = (E_1, \ldots, E_n)$ be a binary or continuous environment vector and $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$ be the $n \times p$ matrix of high-dimensional data where $X_i = (X_{i1}, \ldots, X_{ij}, \ldots, X_{ip})$, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)$ a vector of errors. Consider the regression model with main effects and their interactions with E:

$$g(\boldsymbol{\mu}) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \beta_E E + \sum_{j=1}^{p} \alpha_j E X_j, \tag{1}$$

where $g(\cdot)$ is a known link function, $\boldsymbol{\mu} = \mathsf{E}\left[Y|\boldsymbol{X}, E, \boldsymbol{\beta}, \boldsymbol{\alpha}\right]$, and $\beta_0, \beta_j, \beta_E, \alpha_j$ are the true unknown model parameters for $j = 1, \ldots, p$. 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 $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$. For example, the lasso [2] penalizes the squared loss of the data with the L_1-norm of the regression coefficients resulting in a method that performs both model selection and estimation. A natural extension of the lasso to the interaction model (1) is given by:

$$\arg\min_{\beta_0, \boldsymbol{\beta}, \boldsymbol{\alpha}} \frac{1}{2} \|Y - g(\boldsymbol{\mu})\|_2^2 + \lambda \left(\|\boldsymbol{\beta}\|_1 + \|\boldsymbol{\alpha}\|_1 \right) \tag{2}$$

where $\|Y - g(\boldsymbol{\mu})\|_2^2 = \sum_i (y_i - g(\mu_i))^2$, $\|\boldsymbol{\beta}\|_1 = \sum_j |\beta_j|$, $\|\boldsymbol{\alpha}\|_1 = \sum_j |\alpha_j|$ and $\lambda \geq 0$ is a data driven tuning parameter that can set some of the coefficients to zero when sufficiently large. However, since no constraint is placed on the structure of the model in (2), it is possible that the estimated main effects are zero while the interaction term is not. This has motivated methods that produce structured sparsity [3]. Specifically, we are interested in imposing the strong heredity principle [4]:

$$\hat{\alpha}_j \neq 0 \qquad \Rightarrow \qquad \hat{\beta}_j \neq 0 \qquad \text{and} \qquad \hat{\beta}_E \neq 0 \tag{3}$$

In words, the interaction term will only have a non-zero estimate if its corresponding main effects are estimated to be non-zero. One benefit brought by hierarchy is that the number of measured variables can be reduced, referred to as practical sparsity [5, 6]. For example, a model involving $X_1, E, X_1 \cdot E$ is more parsimonious than a model involving $X_1, E, X_2 \cdot E$, because in the first model a researcher would only have to measure two variables compared to three in the second model. In order to address these issues, we propose to extend the model of [7] to simultaneously perform variable selection, estimation and impose the strong heredity principle in the context of high dimensional interactions with the environment (HD$\times E$). To do so, we follow Choi and reparametrize the coefficients for the interaction terms as $\alpha_j = \gamma_j \beta_j \beta_E$. Plugging this into (1):

$$g(\boldsymbol{\mu}) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \beta_E E + \sum_{j=1}^{p} \gamma_j \beta_j \beta_E E X_j \tag{4}$$

This reparametrization directly enforces the strong heredity principle (3), i.e., if either main effect estimates are 0, then $\hat{\alpha}_j$ will be zero and a non-zero interaction coefficient implies non-zero $\hat{\beta}_j$ and $\hat{\beta}_E$. To perform variable selection in this new parametrization, we follow [7] and penalize $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)$ instead of penalizing $\boldsymbol{\alpha}$ as in (2), leading to the following penalized least squares criterion:

$$\arg\min_{\beta_0,\boldsymbol{\beta},\boldsymbol{\gamma}} \frac{1}{2}\|Y - g(\boldsymbol{\mu})\|^2 + \lambda_\beta \sum_{j=1}^{p} w_j|\beta_j| + \lambda_\gamma \sum_{j=1}^{p} w_{jE}|\gamma_{jE}| \tag{5}$$

where $g(\boldsymbol{\mu})$ is from (10), λ_β and λ_γ are tuning parameters and $\mathbf{w} = (w_1, \ldots, w_q, w_{1E}, \ldots, w_{qE})$ are prespecified adaptive weights. The λ_β tuning parameter controls the amount of shrinkage applied to the main effects, while λ_γ controls the interaction estimates and allows for the possibility of excluding the interaction term from the model even if the corresponding main effects are non-zero. The adaptive weights serve as a way of allowing parameters to be penalized differently. Furthermore, adaptive weighting [8] has been shown to construct oracle procedures [9], i.e., asymptotically, it performs as well as if the true model were given in advance. The oracle property is achieved when the weights are a function of any root-n consistent estimator of the true parameters e.g. maximum likelihood (MLE) or ridge regression estimates. It can be shown that the procedure in (5) asymptotically possesses the oracle property [7], even when the number of parameters tends to ∞ as the sample size increases, if the weights are chosen such that

$$w_j = \left|\frac{1}{\hat{\beta}_j}\right|, \quad w_{jE} = \left|\frac{\hat{\beta}_j\hat{\beta}_E}{\hat{\alpha}_{jE}}\right| \qquad \text{for } j = 1, \ldots, q \tag{6}$$

where $\hat{\beta}_j$ and $\hat{\alpha}_j$ are the MLEs, from (1) or the ridge regression estimates when $p > n$. The rationale behind the data-dependent $\hat{\boldsymbol{w}}$ is that as the sample size grows, the weights for the truly zero predictors go to ∞ (which translates to a large penalty), whereas the weights for the truly non-zero predictors converge to a finite constant [8].

1.2 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 the model

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

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 `sail` method with cubic b-splines and 10-fold CV to choose the optimal value of λ. Default values were used for all other arguments. We plot the solution path for both main effects and interactions in Figure **??** and the estimated functions $\hat{f}_1$ and $\hat{f}_2$ in Figure **??**.

1.3 Existing Literature

Type	Model	Software
Linear	`CAP` (Zhao et al. 2009, *Ann. Stat*)	✗
	`SHIM` (Choi et al. 2009, *JASA*)	✗
	`hiernet` (Bien et al. 2013, *Ann. Stat*)	`hierNet(x, y)`
	`GRESH` (She and Jiang 2014, *JASA*)	✗
	`FAMILY` (Haris et al. 2014, *JCGS*)	`FAMILY(x, z, y)`
	`glinternet` (Lim and Hastie 2015, *JCGS*)	`glinternet(x, y)`
	`RAMP` (Hao et al. 2016, *JASA*)	`RAMP(x, y)`
Non-linear	`VANISH` (Radchenko and James 2010, *JASA*)	✗
	`sail` (Bhatnagar et al. 2017+)	`sail(x, e, y)`

2 Extension to Additive Models

Let $Y = (Y_1, \ldots, Y_n) \in \mathbb{R}^n$ be a continuous outcome variable, $X_E = (E_1, \ldots, E_n) \in \mathbb{R}^n$ a binary or continuous environment vector, $\boldsymbol{X} = (X_1, \ldots, X_p) \in \mathbb{R}^{n \times p}$ a matrix of predictors, and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n) \in \mathbb{R}^n$ a vector of i.i.d random variables with mean 0. Furthermore let $f_j : \mathbb{R} \rightarrow \mathbb{R}$ be a smoothing method for variable X_j by a projection on to a set of basis functions:

$$f_j(X_j) = \sum_{\ell=1}^{m_j} \psi_{j\ell}(X_j)\beta_{j\ell} \tag{8}$$

Here, the $\{\psi_{j\ell}\}_1^{m_j}$ are a family of basis functions in X_j [10]. 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

$$Y = \beta_0 \cdot \mathbf{1} + \sum_{j=1}^{p} \boldsymbol{\Psi}_j \boldsymbol{\theta}_j + \beta_E X_E + \sum_{j=1}^{p} (X_E \circ \boldsymbol{\Psi}_j)\boldsymbol{\alpha}_j + \varepsilon \tag{9}$$

where β_0 is the intercept, β_E is the coefficient for the environment variable, $\boldsymbol{\alpha}_j = (\alpha_{j1}, \ldots, \alpha_{jm_j}) \in \mathbb{R}^{m_j}$ 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 $\boldsymbol{\Psi}_j$. To enforce the strong heredity property, we reparametrize the coefficients for the interaction terms in (9) as $\boldsymbol{\alpha}_j = \gamma_j \beta_E \boldsymbol{\theta}_j$:

$$Y = \beta_0 \cdot \mathbf{1} + \sum_{j=1}^{p} \boldsymbol{\Psi}_j \boldsymbol{\theta}_j + \beta_E X_E + \sum_{j=1}^{p} \gamma_j \beta_E (X_E \circ \boldsymbol{\Psi}_j)\boldsymbol{\theta}_j + \varepsilon \tag{10}$$

For a continuous response, we use the squared-error loss:

$$\mathcal{L}(Y; \boldsymbol{\Theta}) = \frac{1}{2n} \left\| Y - \beta_0 \cdot \mathbf{1} - \sum_{j=1}^{p} \boldsymbol{\Psi}_j \boldsymbol{\theta}_j - \beta_E X_E - \sum_{j=1}^{p} \gamma_j \beta_E (X_E \circ \boldsymbol{\Psi}_j)\boldsymbol{\theta}_j \right\|_2^2 \tag{11}$$

where $\boldsymbol{\Theta} := (\beta_0, \beta_E, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p, \gamma_1, \ldots, \gamma_p)$.

We consider the following penalized least squares criterion for this problem:

$$\arg\min_{\boldsymbol{\Theta}} \mathcal{L}(Y; \boldsymbol{\Theta}) + \lambda(1 - \alpha) \left(w_E |\beta_E| + \sum_{j=1}^{p} w_j \|\boldsymbol{\theta}_j\|_2 \right) + \lambda\alpha \sum_{j=1}^{p} w_{jE} |\gamma_j| \tag{12}$$

where $\lambda > 0$ and $\alpha \in (0,1)$ are tuning parameters and 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. Furthermore, adaptive weighting [8] has been shown to construct oracle procedures [9], i.e., asymptotically, it performs as well as if the true model were given in advance. These weights are given by

$$w_E = \left| \frac{1}{\hat{\beta}_E} \right|, \quad w_j = \frac{1}{\left\| \hat{\boldsymbol{\theta}}_j \right\|_2}, \quad w_{jE} = \left| \frac{\hat{\beta}_E \left\| \hat{\boldsymbol{\theta}}_j \right\|_2}{\| \hat{\alpha}_j \|_2} \right| \quad \text{for } j = 1, \ldots, p \tag{13}$$

where $\hat{\beta}_E$, $\hat{\boldsymbol{\theta}}_j$ and $\hat{\boldsymbol{\alpha}}_j$ are the MLEs, from (9) or the ridge regression estimates when $p > n$.

3 Regularization Path

The `sail` model has the form

$$\hat{Y} = \beta_0 \cdot \mathbf{1} + \sum_{j=1}^{p} \boldsymbol{\Psi}_j \boldsymbol{\theta}_j + \beta_E X_E + \sum_{j=1}^{p} \gamma_j \beta_E (X_E \circ \boldsymbol{\Psi}_j) \boldsymbol{\theta}_j \tag{14}$$

The objective function is given by

$$Q(\boldsymbol{\Theta}) = \frac{1}{2n} \left\| Y - \hat{Y} \right\|_2^2 + \lambda(1-\alpha) \left(w_E |\beta_E| + \sum_{j=1}^{p} w_j \|\boldsymbol{\theta}_j\|_2 \right) + \lambda\alpha \sum_{j=1}^{p} w_{jE} |\gamma_j| \tag{15}$$

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

$$\frac{\partial Q}{\partial \beta_0} = \frac{1}{n}\left(Y - \beta_0 \cdot \mathbf{1} - \sum_{j=1}^{p} \mathbf{\Psi}_j \boldsymbol{\theta}_j - \beta_E X_E - \sum_{j=1}^{p} \gamma_j \beta_E (X_E \circ \mathbf{\Psi}_j)\boldsymbol{\theta}_j\right)^\top \mathbf{1} = 0 \tag{16}$$

$$\frac{\partial Q}{\partial \beta_E} = -\frac{1}{n}\left(X_E + \sum_{j=1}^{p} \gamma_j (X_E \circ \mathbf{\Psi}_j)\boldsymbol{\theta}_j\right)^\top R + \lambda(1-\alpha) w_E s_1 = 0 \tag{17}$$

$$\frac{\partial Q}{\partial \boldsymbol{\theta}_j} = -\frac{1}{n}\left(\mathbf{\Psi}_j + \gamma_j \beta_E (X_E \circ \mathbf{\Psi}_j)\right)^\top R + \lambda(1-\alpha) w_j s_2 = \mathbf{0} \tag{18}$$

$$\frac{\partial Q}{\partial \gamma_j} = -\frac{1}{n}\left(\beta_E (X_E \circ \mathbf{\Psi}_j)\boldsymbol{\theta}_j\right)^\top R + \lambda \alpha w_{jE} s_3 = 0 \tag{19}$$

where s_1 is in the subgradient of the ℓ_1 norm:

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

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:

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

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

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

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 (17), we see that $\beta_E = 0$ is a solution if

$$\frac{1}{w_E} \left| \frac{1}{n} \left(X_E + \sum_{j=1}^{p} \gamma_j (X_E \circ \boldsymbol{\Psi}_j) \boldsymbol{\theta}_j \right)^\top R_{(-E)} \right| \leq \lambda(1 - \alpha) \tag{20}$$

From the subgradient Equation (18), we see that $\boldsymbol{\theta}_j = \mathbf{0}$ is a solution if

$$\frac{1}{w_j} \left\| \frac{1}{n} \left(\boldsymbol{\Psi}_j + \gamma_j \beta_E (X_E \circ \boldsymbol{\Psi}_j) \right)^\top R_{(-j)} \right\|_2 \leq \lambda(1 - \alpha) \tag{21}$$

From the subgradient Equation (19), we see that $\gamma_j = 0$ is a solution if

$$\frac{1}{w_{jE}} \left| \frac{1}{n} \left(\beta_E (X_E \circ \boldsymbol{\Psi}_j) \boldsymbol{\theta}_j \right)^\top R_{(-jE)} \right| \leq \lambda \alpha \tag{22}$$

3.1 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:

$$\lambda_{max} = \frac{1}{n(1-\alpha)} \max \left\{ \frac{1}{w_E} \left(X_E + \sum_{j=1}^{p} \gamma_j (X_E \circ \mathbf{\Psi}_j) \boldsymbol{\theta}_j \right)^\top R_{(-E)}, \right.$$
$$\left. \max_j \frac{1}{w_j} \left\| \left(\mathbf{\Psi}_j + \gamma_j \beta_E (X_E \circ \mathbf{\Psi}_j) \right)^\top R_{(-j)} \right\|_2 \right\} \quad (23)$$

which reduces to

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

3.2 Optimization of Parameters

From the subgradient equations we see that

$$\hat{\beta}_0 = \left(Y - \sum_{j=1}^{p} \mathbf{\Psi}_j \hat{\boldsymbol{\theta}}_j - \hat{\beta}_E X_E - \sum_{j=1}^{p} \hat{\gamma}_j \hat{\beta}_E (X_E \circ \mathbf{\Psi}_j) \hat{\boldsymbol{\theta}}_j \right)^\top \mathbf{1} \quad (24)$$

$$\hat{\beta}_E = S\left(\frac{1}{n \cdot w_E} \left(X_E + \sum_{j=1}^{p} \hat{\gamma}_j (X_E \circ \mathbf{\Psi}_j) \hat{\boldsymbol{\theta}}_j \right)^\top R_{(-E)}, \lambda(1-\alpha) \right) \quad (25)$$

$$\lambda(1-\alpha) w_j \frac{\boldsymbol{\theta}_j}{\|\boldsymbol{\theta}_j\|_2} = \frac{1}{n} \left(\mathbf{\Psi}_j + \gamma_j \beta_E (X_E \circ \mathbf{\Psi}_j) \right)^\top R_{(-j)} \quad (26)$$

$$\hat{\gamma}_j = S\left(\frac{1}{n \cdot w_{jE}} \left(\beta_E (X_E \circ \mathbf{\Psi}_j) \boldsymbol{\theta}_j \right)^\top R_{(-jE)}, \lambda\alpha \right) \quad (27)$$

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

4 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` [11].

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 (15)

  $\Psi_j \leftarrow$ `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: $\left\| \boldsymbol{\Theta}^{(k)} - \boldsymbol{\Theta}^{(k-1)} \right\|_2^2 < \epsilon$

4.1 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

$$R_s = Y - \widetilde{Y}^{(k)}_{(-s)} \tag{28}$$

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

$$\widetilde{Y}^{(k)}_{(-s)} = \beta_0^{(k)} - \beta_E^{(k)} X_E - \sum_{\ell \neq s} \boldsymbol{\Psi}_\ell \boldsymbol{\theta}_\ell^{(k)} - \sum_{\ell \neq s} \gamma_\ell^{(k)} \beta_E^{(k)} \widetilde{\boldsymbol{\Psi}}_\ell \boldsymbol{\theta}_\ell^{(k)} \tag{29}$$

Using (29), (28) can be re-written as

$$\begin{aligned} R_s &= Y - \beta_0^{(k)} - \beta_E^{(k)} X_E - \sum_{j=1}^{p} (\boldsymbol{\Psi}_j + \gamma_j^{(k)} \beta_E^{(k)} \widetilde{\boldsymbol{\Psi}}_j) \boldsymbol{\theta}_j^{(k)} + (\boldsymbol{\Psi}_s + \gamma_s^{(k)} \beta_E^{(k)} \widetilde{\boldsymbol{\Psi}}_s) \boldsymbol{\theta}_s^{(k)} \\ &= R^* + (\boldsymbol{\Psi}_s + \gamma_s^{(k)} \beta_E^{(k)} \widetilde{\boldsymbol{\Psi}}_s) \boldsymbol{\theta}_s^{(k)} \end{aligned} \tag{30}$$

where

$$R^* = Y - \beta_0^{(k)} - \beta_E^{(k)} X_E - \sum_{j=1}^{p} (\boldsymbol{\Psi}_j + \gamma_j^{(k)} \beta_E^{(k)} \widetilde{\boldsymbol{\Psi}}_j) \boldsymbol{\theta}_j^{(k)} \tag{31}$$

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

$$\boldsymbol{\theta}_s^{(k)(new)} = \arg\min_{\boldsymbol{\theta}_j} \frac{1}{2n} \left\| R_s - (\boldsymbol{\Psi}_s + \gamma_s^{(k)} \beta_E^{(k)} \widetilde{\boldsymbol{\Psi}}_s) \boldsymbol{\theta}_j \right\|_2^2 + \lambda(1-\alpha) w_s \|\theta_j\|_2 \tag{32}$$

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

$$R_{s+1} = R^* + (\boldsymbol{\Psi}_{s+1} + \gamma_{s+1}^{(k)} \beta_E^{(k)} \widetilde{\boldsymbol{\Psi}}_{s+1}) \boldsymbol{\theta}_{s+1}^{(k)} + (\boldsymbol{\Psi}_s + \gamma_s^{(k)} \beta_E^{(k)} \widetilde{\boldsymbol{\Psi}}_s)(\boldsymbol{\theta}_s^{(k)} - \boldsymbol{\theta}_s^{(k)(new)}) \tag{33}$$

where R^* is given by (31), $\boldsymbol{\theta}_s^{(k)}$ is the parameter value prior to the update, and $\boldsymbol{\theta}_s^{(k)(new)}$ is the updated value given by (32). Taking the difference between (30) and (33) gives

$$\begin{aligned}
\Delta &= R_t - R_s \\
&= (\boldsymbol{\Psi}_t + \gamma_t^{(k)} \beta_E^{(k)} \widetilde{\boldsymbol{\Psi}}_t)\boldsymbol{\theta}_t^{(k)} + (\boldsymbol{\Psi}_s + \gamma_s^{(k)} \beta_E^{(k)} \widetilde{\boldsymbol{\Psi}}_s)(\boldsymbol{\theta}_s^{(k)} - \boldsymbol{\theta}_s^{(k)(new)}) - (\boldsymbol{\Psi}_s + \gamma_s^{(k)} \beta_E^{(k)} \widetilde{\boldsymbol{\Psi}}_s)\boldsymbol{\theta}_s^{(k)} \\
&= (\boldsymbol{\Psi}_t + \gamma_t^{(k)} \beta_E^{(k)} \widetilde{\boldsymbol{\Psi}}_t)\boldsymbol{\theta}_t^{(k)} - (\boldsymbol{\Psi}_s + \gamma_s^{(k)} \beta_E^{(k)} \widetilde{\boldsymbol{\Psi}}_s)\boldsymbol{\theta}_s^{(k)(new)} \qquad (34)
\end{aligned}$$

Therefore $R_t = R_s + \Delta$, and the partial residual for updating the next predictor can be computed by updating the previous partial residual by Δ, given by (34). This formulation can lead to computational speedups especially when $\Delta = 0$, meaning the partial residual does not need to be re-calculated.

5 Weak Heredity

We can also enforce the weak heredity property:

$$\hat{\alpha}_{jE} \neq 0 \qquad \Rightarrow \qquad \hat{\beta}_j \neq 0 \qquad \text{or} \qquad \hat{\beta}_E \neq 0 \qquad (35)$$

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 (9) 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 $\boldsymbol{\theta}_j$).

5.1 Regularization Path

The `sail` model with weak heredity has the form

$$\hat{Y} = \beta_0 \cdot \mathbf{1} + \sum_{j=1}^{p} \boldsymbol{\Psi}_j \boldsymbol{\theta}_j + \beta_E X_E + \sum_{j=1}^{p} \gamma_j (X_E \circ \boldsymbol{\Psi}_j)(\beta_E \cdot \mathbf{1}_{m_j} + \boldsymbol{\theta}_j) \tag{36}$$

The objective function is given by

$$Q(\boldsymbol{\Theta}) = \frac{1}{2n} \left\| Y - \hat{Y} \right\|_2^2 + \lambda(1-\alpha) \left(w_E |\beta_E| + \sum_{j=1}^{p} w_j \|\boldsymbol{\theta}_j\|_2 \right) + \lambda\alpha \sum_{j=1}^{p} w_{jE} |\gamma_j| \tag{37}$$

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

$$\frac{\partial Q}{\partial \beta_0} = \frac{1}{n} \left(Y - \beta_0 \cdot \mathbf{1} - \sum_{j=1}^{p} \boldsymbol{\Psi}_j \boldsymbol{\theta}_j - \beta_E X_E - \sum_{j=1}^{p} \gamma_j (X_E \circ \boldsymbol{\Psi}_j)(\beta_E \cdot \mathbf{1}_{m_j} + \boldsymbol{\theta}_j) \right)^\top \mathbf{1} = 0 \tag{38}$$

$$\frac{\partial Q}{\partial \beta_E} = -\frac{1}{n} \left(X_E + \sum_{j=1}^{p} \gamma_j (X_E \circ \boldsymbol{\Psi}_j) \mathbf{1}_{m_j} \right)^\top R + \lambda(1-\alpha) w_E s_1 = 0 \tag{39}$$

$$\frac{\partial Q}{\partial \boldsymbol{\theta}_j} = -\frac{1}{n} \left(\boldsymbol{\Psi}_j + \gamma_j (X_E \circ \boldsymbol{\Psi}_j) \right)^\top R + \lambda(1-\alpha) w_j s_2 = \mathbf{0} \tag{40}$$

$$\frac{\partial Q}{\partial \gamma_j} = -\frac{1}{n} \left((X_E \circ \boldsymbol{\Psi}_j)(\beta_E \cdot \mathbf{1}_{m_j} + \boldsymbol{\theta}_j) \right)^\top R + \lambda\alpha w_{jE} s_3 = 0 \tag{41}$$

where s_1 is in the subgradient of the ℓ_1 norm:

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

s_2 is in the subgradient of the ℓ_2 norm:

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

and s_3 is in the subgradient of the ℓ_1 norm:

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

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

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

the partial residual without X_E as

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

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

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

From the subgradient Equation (39), we see that $\beta_E = 0$ is a solution if

$$\frac{1}{w_E} \left| \frac{1}{n} \left(X_E + \sum_{j=1}^{p} \gamma_j (X_E \circ \boldsymbol{\Psi}_j) \mathbf{1}_{m_j} \right)^\top R_{(-E)} \right| \leq \lambda(1 - \alpha) \tag{42}$$

From the subgradient Equation (40), we see that $\boldsymbol{\theta}_j = \mathbf{0}$ is a solution if

$$\frac{1}{w_j} \left\| \frac{1}{n} \left(\boldsymbol{\Psi}_j + \gamma_j (X_E \circ \boldsymbol{\Psi}_j) \right)^\top R_{(-j)} \right\|_2 \leq \lambda(1 - \alpha) \tag{43}$$

From the subgradient Equation (41), we see that $\gamma_j = 0$ is a solution if

$$\frac{1}{w_{jE}} \left| \frac{1}{n} \left((X_E \circ \boldsymbol{\Psi}_j)(\beta_E \cdot \mathbf{1}_{m_j} + \boldsymbol{\theta}_j) \right)^\top R_{(-jE)} \right| \leq \lambda \alpha \tag{44}$$

5.2 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:

$$
\begin{aligned}
\lambda_{max} = \frac{1}{n} \max \Bigg\{ & \frac{1}{(1-\alpha) w_E} \left(X_E + \sum_{j=1}^{p} \gamma_j (X_E \circ \boldsymbol{\Psi}_j) \mathbf{1}_{m_j} \right)^\top R_{(-E)}, \\
& \max_j \frac{1}{(1-\alpha) w_j} \left\| \left(\boldsymbol{\Psi}_j + \gamma_j (X_E \circ \boldsymbol{\Psi}_j) \right)^\top R_{(-j)} \right\|_2, \\
& \max_j \frac{1}{\alpha w_{jE}} \left((X_E \circ \boldsymbol{\Psi}_j)(\beta_E \cdot \mathbf{1}_{m_j} + \boldsymbol{\theta}_j) \right)^\top R_{(-jE)} \Bigg\}
\end{aligned} \tag{45}
$$

which reduces to

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

5.3 Optimization of Parameters

From the subgradient equations we see that

$$
\hat{\beta}_0 = \left(Y - \sum_{j=1}^{p} \boldsymbol{\Psi}_j \hat{\boldsymbol{\theta}}_j - \hat{\beta}_E X_E - \sum_{j=1}^{p} \hat{\gamma}_j (X_E \circ \boldsymbol{\Psi}_j)(\hat{\beta}_E \cdot \mathbf{1}_{m_j} + \hat{\boldsymbol{\theta}}_j) \right)^\top \mathbf{1} \tag{46}
$$

$$
\hat{\beta}_E = S \left(\frac{1}{n \cdot w_E} \left(X_E + \sum_{j=1}^{p} \hat{\gamma}_j (X_E \circ \boldsymbol{\Psi}_j) \mathbf{1}_{m_j} \right)^\top R_{(-E)}, \lambda(1-\alpha) \right) \tag{47}
$$

$$
\lambda(1-\alpha) w_j \frac{\boldsymbol{\theta}_j}{\|\boldsymbol{\theta}_j\|_2} = \frac{1}{n} \left(\boldsymbol{\Psi}_j + \gamma_j (X_E \circ \boldsymbol{\Psi}_j) \right)^\top R_{(-j)} \tag{48}
$$

$$
\hat{\gamma}_j = S \left(\frac{1}{n \cdot w_{jE}} \left((X_E \circ \boldsymbol{\Psi}_j)(\beta_E \cdot \mathbf{1}_{m_j} + \boldsymbol{\theta}_j) \right)^\top R_{(-jE)}, \lambda\alpha \right) \tag{49}
$$

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

5.4 Algorithm

6 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 [12, 13]

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 (37)

$\Psi_j \leftarrow \texttt{splines::bs}(X_j,\ \texttt{df},\ \texttt{degree})$ 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 $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)$

  $\widetilde{X}_j \leftarrow \widetilde{\boldsymbol{\Psi}}_j (\beta_E^{(k)} \cdot \mathbf{1}_{m_j} + \boldsymbol{\theta}_j^{(k)}) \qquad$ 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

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

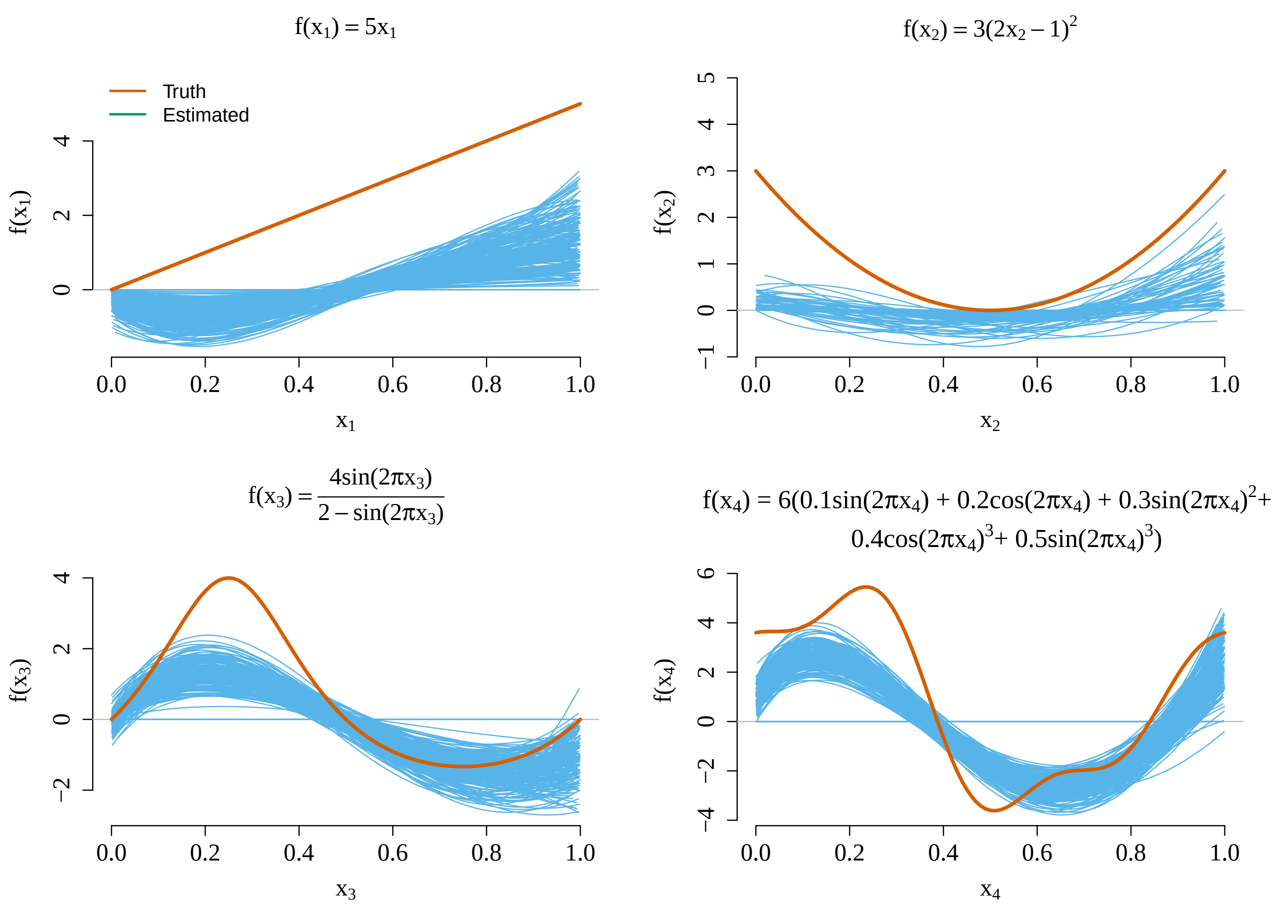

Figure 1: text

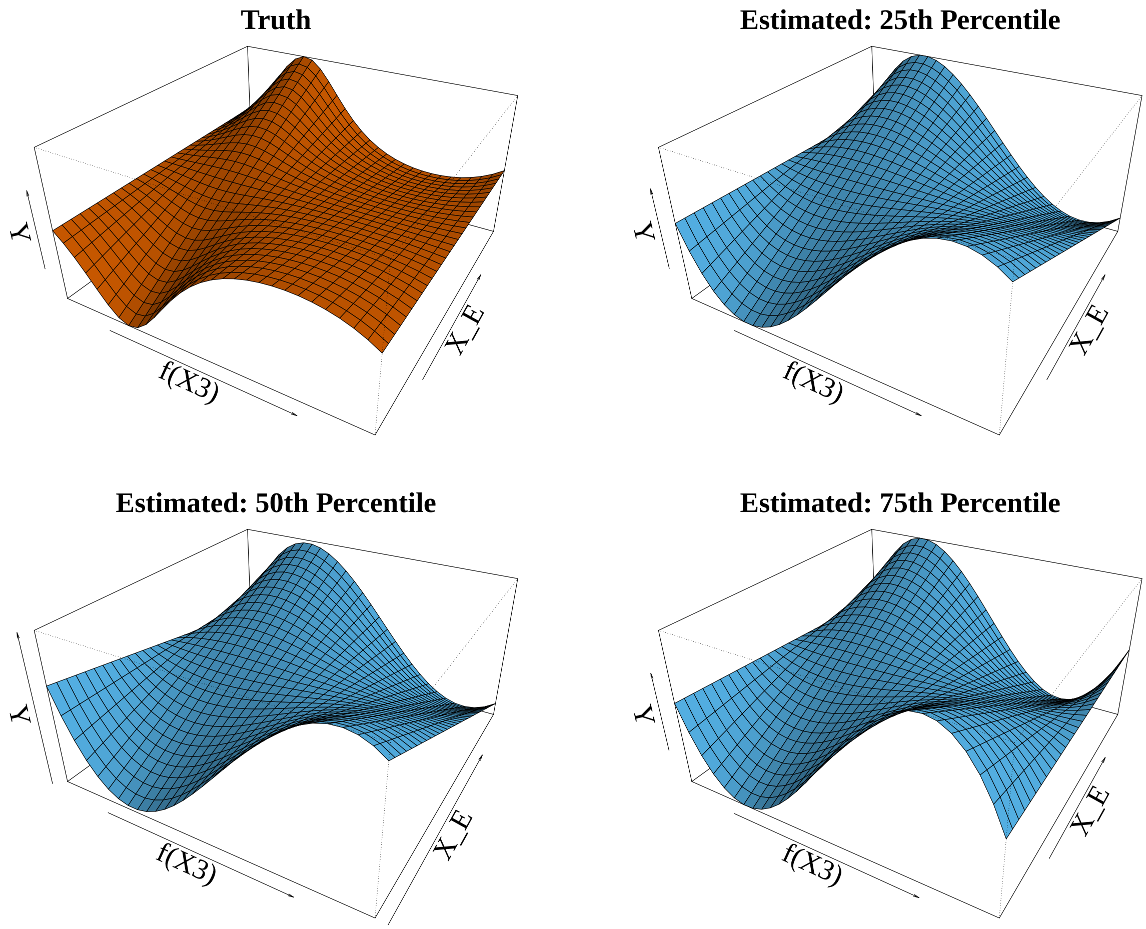

Figure 2: True and estimated interaction effects for $E \cdot f(X_3)$ in simulation scenario 1a).

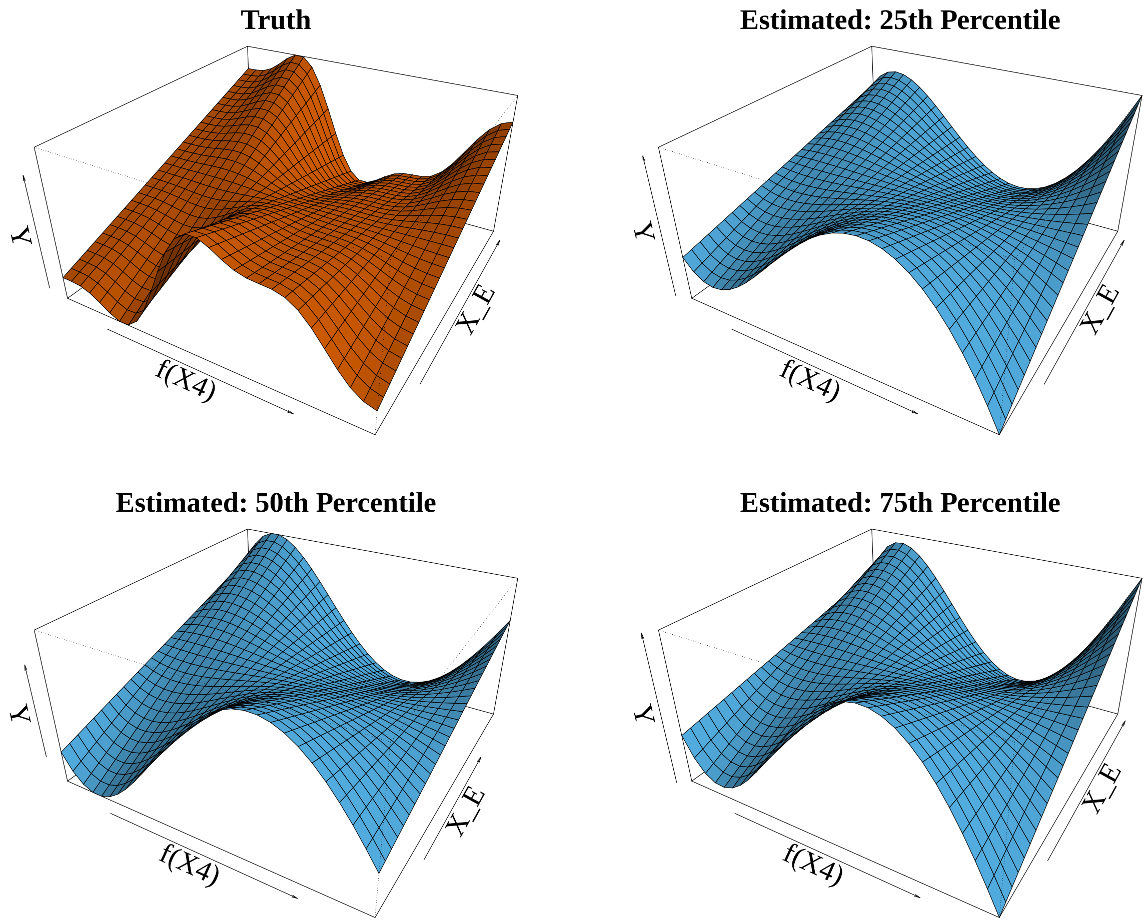

Figure 3: True and estimated interaction effects for $E \cdot f(X_4)$ in simulation scenario 1a).

7 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.

[2] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

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

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

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

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

[7] 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.

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

[9] 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.

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

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

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

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