Computational Statistics and Data Analysis 179 (2023) 107624

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

A sparse additive model for high-dimensional interactions with an exposure variable

Sahir R. Bhatnagar [a,b,*], Tianyuan Lu [c,d], Amanda Lovato [e], David L. Olds [f], Michael S. Kobor [g], Michael J. Meaney [h,i], Kieran O'Donnell [j], Archer Y. Yang [k], Celia M.T. Greenwood [a,c,d,l]

[a] Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montréal, Canada
[b] Department of Diagnostic Radiology, McGill University, Montréal, Canada
[c] Quantitative Life Sciences, McGill University, Canada
[d] Lady Davis Institute, Jewish General Hospital, Montréal, QC, Canada
[e] Statistics Canada, Ottawa, ON, Canada
[f] Department of Pediatrics, University of Colorado School of Medicine, Denver, United States of America
[g] Department of Medical Genetics, University of British Columbia, BC, Canada
[h] Singapore Institute for Clinical Sciences, Singapore
[i] Departments of Psychiatry and Neurology & Neurosurgery, McGill University, Canada
[j] Department of Obstetrics, Gynecology and Reproductive Sciences, Yale School of Medicine, United States of America
[k] Department of Mathematics and Statistics, McGill University, Canada
[l] Departments of Oncology and Human Genetics, McGill University, Canada

ARTICLE INFO

Article history:
Received 13 October 2020
Received in revised form 16 September 2022
Accepted 17 September 2022
Available online 24 September 2022

Keywords:
Gene-environment interaction
Strong heredity property
Blockwise coordinate descent
High-dimensional data
Variable selection

ABSTRACT

A conceptual paradigm for onset of a new disease is 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. However, when modeling a relevant phenotype as a function of high dimensional measurements, power to estimate interactions is low, the number of possible interactions could be enormous and their effects may be non-linear. A method called `sail` for detecting non-linear interactions with a key environmental or exposure variable in high-dimensional settings which respects the strong or weak heredity constraints is proposed. It is proven that asymptotically, `sail` possesses the oracle property, i.e., it performs as well as if the true model were known in advance. A computationally efficient fitting algorithm with automatic tuning parameter selection, which scales to high-dimensional datasets is proposed. Simulation results show that `sail` outperforms existing penalized regression methods in terms of prediction accuracy and support recovery when there are non-linear interactions with an exposure variable. `sail` is applied to detect non-linear interactions between genes and a prenatal psychosocial intervention program on cognitive performance in children at 4 years of age. Results show that individuals who are genetically predisposed to lower educational attainment are those who stand to benefit the most from the intervention. The proposed algorithms are implemented in an R package available on CRAN (https://cran.r-project.org/package=sail).

* Corresponding author at: Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montréal, Canada.
E-mail address: sahir.bhatnagar@mcgill.ca (S.R. Bhatnagar).

https://doi.org/10.1016/j.csda.2022.107624
0167-9473/

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 an enduring interest in powerful methods for estimating interactions, since interactions may reflect important modulation of a genomic system by an external factor and vice versa (Bhatnagar et al., 2018).

Interactions may occur in numerous types and of varying complexities. In this paper, we consider one specific type of interaction model, where one exposure variable E 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 between $\mathbf{X}$ and E. Our method is motivated by the Nurse Family Partnership (NFP); a program of prenatal and infancy home visiting by nurses for low-income mothers and their children (Olds et al., 1998). In this intervention, NFP nurses guided pregnant women and parents of young children to improve the outcomes of pregnancy, their children's health and development, and their economic self-sufficiency, with the goal of reducing disparities over the life-course. Early intervention in young children has been shown to positively impact intellectual abilities (Campbell and Ramey, 1994), and more recent studies have shown that cognitive performance is also strongly influenced by genetic factors (Rietveld et al., 2013). Given the important role of both environment and genetics, we are interested in finding interactions between these two components on cognitive function in children.

1.1. A sparse additive interaction model

Let $Y \in \mathbb{R}$ be a continuous outcome variable, $E \in \mathbb{R}$ a binary or continuous environment/exposure vector of known importance, and $X \in \mathbb{R}^p$ a vector of additional predictors, possibly high-dimensional. Assume that we have n observations of each quantity denoted by $Y = (Y_1, \ldots, Y_n) \in \mathbb{R}^n$, $X_E = (E_1, \ldots, E_n) \in \mathbb{R}^n$, and $\boldsymbol{X} = (X_1^\top, \ldots, X_p^\top) \in \mathbb{R}^{n \times p}$. 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 (Hastie et al., 2015). 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 $j = 1, \ldots, p$ ($\boldsymbol{\theta}_j$ 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

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

where $\beta_0 \in \mathbb{R}$ is the intercept, $\beta_E \in \mathbb{R}$ is the coefficient for the environment variable, $\boldsymbol{\tau}_j = (\tau_{j1}, \ldots, \tau_{jm_j}) \in \mathbb{R}^{m_j}$ are the basis coefficients for the jth interaction term, $(X_E \circ \boldsymbol{\Psi}_j)$ is the $n \times m_j$ matrix formed by the component-wise multiplication of the column vector X_E by each column of $\boldsymbol{\Psi}_j$, and $\varepsilon \in \mathbb{R}^n$ is a vector of i.i.d. errors with mean zero and finite variance. Here we assume that p is large relative to n, and particularly that $\sum_{j=1}^{p} m_j/n$ is large. Due to the large number of parameters to estimate with respect to the number of observations, one commonly-used approach in the penalization literature is to shrink the regression coefficients by placing a constraint on the values of $(\beta_E, \boldsymbol{\theta}_j, \boldsymbol{\tau}_j)$. Certain constraints have the added benefit of producing a sparse model in the sense that many of the coefficients will be set exactly to 0 (Bühlmann and Van De Geer, 2011). Such a reduced predictor set can lead to a more interpretable model with smaller prediction variance, albeit at the cost of having biased parameter estimates (Fan et al., 2014). In light of these goals, consider the following penalized objective function:

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

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

An issue with (3) is that since no constraint is placed on the structure of the model, it is possible that an estimated interaction term is non-zero while the corresponding main effects are zero. While there may be certain situations where

this is plausible, statisticians have generally argued that interactions should only be included if the corresponding main effects are also in the model (McCullagh and Nelder, 1989). This is known as the strong heredity principle (Chipman, 1996). Indeed, large main effects are more likely to lead to detectable interactions (Cox, 1984).

The strong heredity principle states that an interaction term can only have a non-zero estimate if its corresponding main effects are estimated to be non-zero, whereas the weak heredity principle allows for a non-zero interaction estimate as long as one of the corresponding main effects is estimated to be non-zero (Chipman, 1996). In the context of penalized regression methods, these principles can be formulated as structured sparsity (Bach et al., 2012) problems. Several authors have proposed to modify the type of penalty in order to achieve the heredity principle (Radchenko and James, 2010; Bien et al., 2013; Lim and Hastie, 2015; Haris et al., 2016). We take an alternative approach. In Section 2 we discuss how a simple reparametrization of the model (3) can lead to this desirable property.

1.2. Related work

Methods for variable selection of interactions can be broken down into two categories: linear and non-linear interaction effects. Many of the linear effect methods consider all pairwise interactions in $\boldsymbol{X}$ (Zhao et al., 2009; Choi et al., 2010; Bien et al., 2013; She and Jiang, 2014) which can be computationally prohibitive when p is large. More recent proposals for selection of interactions allow the user to restrict the search space to interaction candidates (Lim and Hastie, 2015; Haris et al., 2016). This is useful when the researcher wants to impose prior information on the model. Two-stage procedures, where interaction candidates are considered from an original screen of main effects, have shown good performance when p is large (Hao et al., 2018; Shah, 2016) in the linear setting. There are many fewer methods available for estimating non-linear interactions. For example, Radchenko and James (2010) 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 more computationally expensive than `sail` since it considers all pairwise interactions between the basis functions, and its effectiveness in simulations or real-data applications is unknown as there is no software implementation.

1.3. Our contributions

The main contributions of this paper are five-fold. 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 in terms of prediction accuracy and support recovery over existing methods that only consider linear interactions or additive main effects. Third, we show that our method possesses the oracle property (Fan and Li, 2001), i.e., it performs as well as if the true model were known in advance. Fourth, we demonstrate the performance of our method in two applications: 1) gene-environment interactions in a prenatal psychosocial intervention program Olds et al. (1998) and 2) a study aimed at identifying which clinical variables influence mortality rates amongst seriously ill hospitalized patients (Connors et al., 1995). Fifth, we implement our algorithms in the `sail` R package on CRAN (https://cran.r-project.org/package=sail), along with extensive documentation. In particular, our implementation also allows for linear interaction models, user-defined basis expansions, a cross-validation procedure for selecting the optimal tuning parameter, and differential shrinkage parameters to apply the adaptive lasso idea (Zou, 2006).

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

2. Model and algorithm

2.1. Strong and weak heredity

Following Choi et al. 2010, we introduce a new set of parameters $\boldsymbol{\gamma} = (\gamma_{1E}, \ldots, \gamma_{pE}) \in \mathbb{R}^p$ and reparametrize the coefficients for the interaction terms $\boldsymbol{\tau}_j$ in (2) as a function of γ_{jE} and the main effect parameters $\boldsymbol{\theta}_j$ and β_E. This reparametrization for both strong and weak heredity is summarized in Table 1.

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

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

An estimate of the regression parameters is given by $\widehat{\boldsymbol{\Phi}} = \arg\min_{\boldsymbol{\Phi}} Q(\boldsymbol{\Phi})$. This penalty allows for the possibility of excluding the interaction term from the model even if the corresponding main effects are non-zero. Furthermore, smaller values for

Table 1
Summary of reparametrization and penalty terms for strong and weak heredity `sail` model. Note that the penalty terms are identical for both model types, i.e., the reparametrization only affects the likelihood term of the objective function.

Model	Reparametrization	Penalty
Strong heredity	$\boldsymbol{\tau}_j = \gamma_{jE}\beta_E\boldsymbol{\theta}_j$	$\lambda(1-\alpha)\left(w_E\lvert\beta_E\rvert + \sum_{j=1}^p w_j\lVert\boldsymbol{\theta}_j\rVert_2\right) + \lambda\alpha\sum_{j=1}^p w_{jE}\lvert\gamma_{jE}\rvert$
Weak heredity	$\boldsymbol{\tau}_j = \gamma_{jE}(\beta_E\cdot\mathbf{1}_{m_j} + \boldsymbol{\theta}_j)$	$\lambda(1-\alpha)\left(w_E\lvert\beta_E\rvert + \sum_{j=1}^p w_j\lVert\boldsymbol{\theta}_j\rVert_2\right) + \lambda\alpha\sum_{j=1}^p w_{jE}\lvert\gamma_{jE}\rvert$

α would lead to more interactions being included in the final model while values approaching 1 would favor main effects. Similar to the elastic net (Zou and Zhang, 2009), we fix α and obtain a solution path over a sequence of λ values.

2.2. Toy example

We present here a toy example to better illustrate the methods proposed in this paper. With a sample size of $n = 100$, we sample $p = 20$ covariates $X_1, \dots X_p$ independently from a $N(0,1)$ distribution truncated to the interval [0,1]. Data were generated from a model which follows the strong heredity principle, but where only one covariate, X_2, is involved in an interaction with a binary exposure variable (E):

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

For illustration, 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 (described below) with B-splines (df=5) to estimate the functional forms. 10-fold cross-validation (CV) was used to choose the optimal value of penalization. We used $\alpha = 0.5$ and default values for all other arguments. We plot the solution path for both main effects and interactions in Fig. 1 (top panel), coloring lines to correspond 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 non-zero if the main effects E and $f_2(X_2)$ (black and orange lines respectively in the top panel) are non-zero, while non-zero main effects do not imply a non-zero interaction.

In Fig. 1 (bottom panel), we plot the true and estimated component functions $\hat{f}_1(X_1)$ and $E\cdot\hat{f}_2(X_2)$, and their estimates from this analysis with `sail`. We are able to capture the shape of the correct functional form. Lack-of-fit for $f_1(X_1)$ can be partially explained by acknowledging that `sail` is trying to fit a spline to a linear function. Nevertheless, this example demonstrates that `sail` can still identify trends reasonably well.

2.3. Blockwise coordinate descent for least-squares loss

Here we describe a blockwise coordinate descent algorithm for fitting the least-squares version of the `sail` model in (4). We fix the value for α and minimize the objective function over a decreasing sequence of λ values ($\lambda_{max} > \cdots > \lambda_{min}$). We use the subgradient equations to determine the maximal value λ_{max} such that all estimates are zero (the derivation of λ_{max} is provided in Supplemental Section B.3). Due to the heredity principle, this reduces to finding the largest λ such that all main effects $(\beta_E, \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_p)$ are zero. Following Friedman et al. 2010, 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 $\lambda_{\ell+1}$.

We assume that Y, $\boldsymbol{\Psi}_j$, X_E and $X_E\circ\boldsymbol{\Psi}_j$ have been centered by their sample means $\overline{Y}$, $\overline{\boldsymbol{\Psi}}_j$, $\overline{X}_E$, and $\overline{X_E\circ\boldsymbol{\Psi}_j}$, respectively. Here, $\overline{\boldsymbol{\Psi}}_j \in \mathbb{R}^{m_j}$ and $\overline{X_E\circ\boldsymbol{\Psi}_j} \in \mathbb{R}^{m_j}$ represent the column means of $\boldsymbol{\Psi}_j$ and $X_E\circ\boldsymbol{\Psi}_j$, respectively. Since the intercept (β_0) is not penalized and all variables have been centered, we can omit it from the loss function and compute it once the algorithm has converged for all other parameters. The strong heredity `sail` model with least-squares loss has the form:

$$\hat{Y} = \sum_{j=1}^p \boldsymbol{\Psi}_j\boldsymbol{\theta}_j + \beta_E X_E + \sum_{j=1}^p \gamma_{jE}\beta_E(X_E\circ\boldsymbol{\Psi}_j)\boldsymbol{\theta}_j\,, \tag{5}$$

and the objective function is given by

$$Q(\boldsymbol{\Phi}) = \frac{1}{2n}\lVert Y - \hat{Y}\rVert_2^2 + \lambda(1-\alpha)\left(w_E\lvert\beta_E\rvert + \sum_{j=1}^p w_j\lVert\boldsymbol{\theta}_j\rVert_2\right) + \lambda\alpha\sum_{j=1}^p w_{jE}\lvert\gamma_{jE}\rvert. \tag{6}$$

Solving (6) in a blockwise manner allows us to leverage computationally fast algorithms for ℓ_1 and ℓ_2 norm penalized regression. Indeed, by careful construction of pseudo responses and pseudo design matrices, existing efficient algorithms

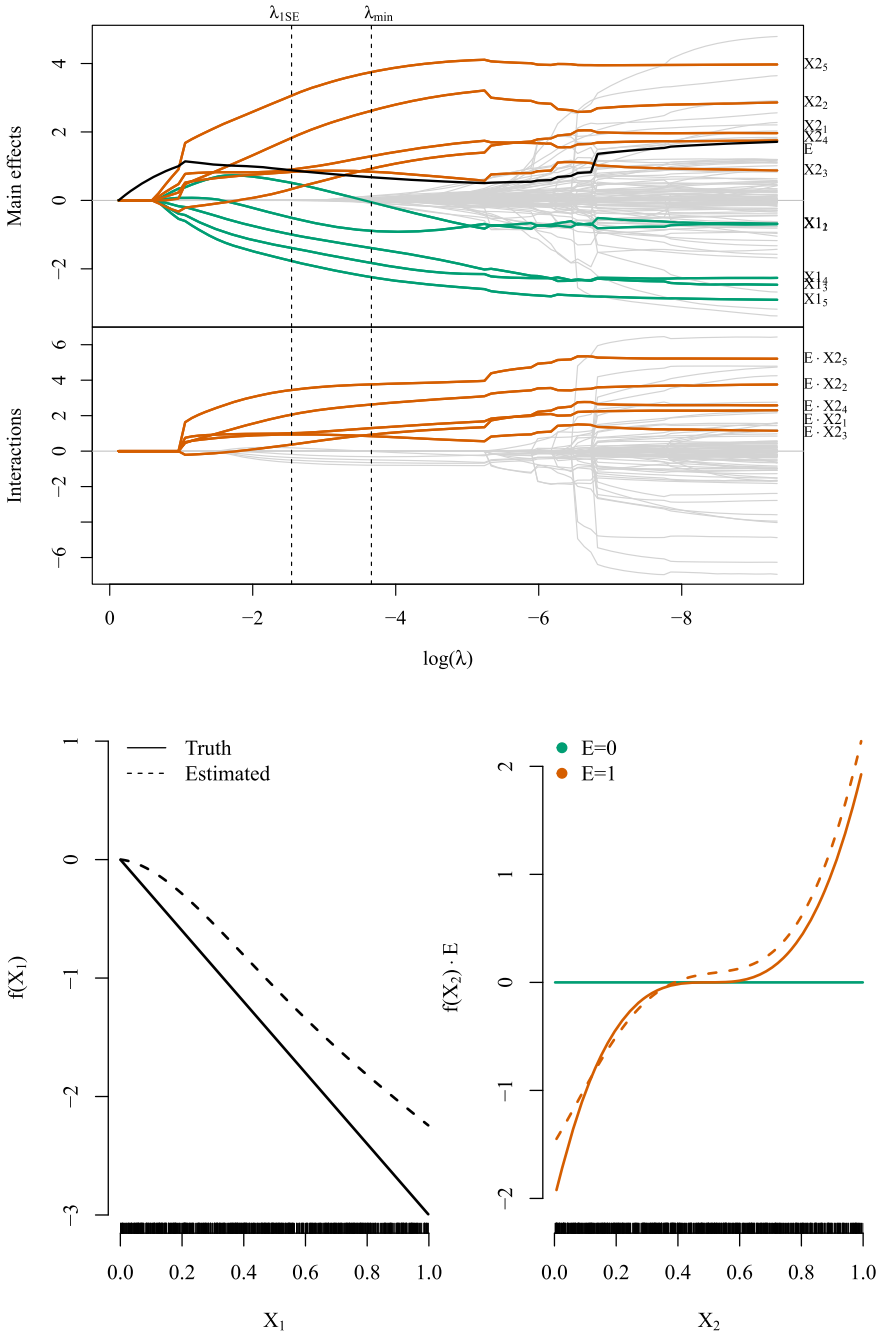

Fig. 1. Top: Toy example solution path for main effects (top) and interactions (bottom). $\{X1_1, X1_2, X1_3, X1_4, X1_5\}$ and $\{X2_1, X2_2, X2_3, X2_4, X2_5\}$ are the five basis coefficients for X_1 and X_2, respectively. λ_{1SE} is the largest value of penalization for which the CV error is within one standard error of the minimizing value λ_{min}. **Bottom**: Estimated smooth functions for X_1 and the $X_2 \cdot E$ interaction by the `sail` method based on λ_{min}. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

can be used to estimate the parameters. 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_{jE} fixed. The main computations are provided in Algorithm 1. A more detailed version of the derivations is given in Supplemental Section B.1. The sequence of tuning parameters $(\lambda_{max} > \cdots > \lambda_{min})$ is automatically chosen by our software package based on the data inputs $\boldsymbol{X}$, Y, and X_E. The user may also choose to supply their own decreasing sequence. We recommend B-splines with 5 degrees of freedom for the *basis* function, and $\alpha = 0.5$ to provide similar penalties to both main effects and interactions. Smaller values of α will favor the inclusion of more interaction terms. The weights for the environment variable (w_E), main effects (w_j) and interactions

(w_{jE}) can be chosen via the adaptive `sail` (Algorithm 2), or be left to their default values of 1. Smaller weights will penalize the corresponding variables less. The default value for the convergence threshold (ϵ) is 1×10^{-4}.

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

function `sail`$(\boldsymbol{X}, Y, X_E, \texttt{basis}, \lambda, \alpha, w_j, w_E, w_{jE}, \epsilon)$ ▷ Algorithm for solving (6)
  $\boldsymbol{\Psi}_j \leftarrow \texttt{basis}(X_j)$, $\widetilde{\boldsymbol{\Psi}}_j \leftarrow X_E \circ \boldsymbol{\Psi}_j$ for $j = 1, \ldots, p$
  Center all variables by their sample means
  Initialize: $\beta_E^{(0)} = \boldsymbol{\theta}_j^{(0)} = \gamma_j^{(0)} \leftarrow 0$ for $j = 1, \ldots, p$.
  Set iteration counter $k \leftarrow 0$
  $R^* \leftarrow Y - \beta_E^{(k)} X_E - \sum_j (\boldsymbol{\Psi}_j + \gamma_j^{(k)} \beta_E^{(k)} \widetilde{\boldsymbol{\Psi}}_j) \boldsymbol{\theta}_j^{(k)}$
  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 $j = 1, \ldots, p$
      $R \leftarrow R^* + \sum_{j=1}^p \gamma_j^{(k)} \widetilde{X}_j$

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

      $\Delta = \sum_j (\gamma_j^{(k)} - \gamma_j^{(k)(new)}) \widetilde{X}_j$
      $R^* \leftarrow R^* + \Delta$
    • 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**
        $R \leftarrow R^* + \widetilde{X}_j \boldsymbol{\theta}_j^{(k)}$

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

        $\Delta = \widetilde{X}_j (\boldsymbol{\theta}_j^{(k)} - \boldsymbol{\theta}_j^{(k)(new)})$
        $R^* \leftarrow R^* + \Delta$
    • To update β_E
      $\widetilde{X}_E \leftarrow X_E + \sum_j \gamma_j^{(k)} \widetilde{\boldsymbol{\Psi}}_j \boldsymbol{\theta}_j^{(k)}$
      $R \leftarrow R^* + \beta_E^{(k)} \widetilde{X}_E$

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

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

      $\Delta = (\beta_E^{(k)} - \beta_E^{(k)(new)}) \widetilde{X}_E$
      $R^* \leftarrow R^* + \Delta$
    $k \leftarrow k + 1$
  until convergence criterion is satisfied: $|Q(\boldsymbol{\Phi}^{(k-1)}) - Q(\boldsymbol{\Phi}^{(k)})| / Q(\boldsymbol{\Phi}^{(k-1)}) < \epsilon$
  Compute the intercept β_0
    $\beta_0 \leftarrow \overline{Y} - \sum_{j=1}^p \overline{\boldsymbol{\Psi}}_j \hat{\boldsymbol{\theta}}_j - \hat{\beta}_E \overline{X}_E - \sum_{j=1}^p \hat{\gamma}_j \hat{\beta}_E (\overline{X_E \circ \boldsymbol{\Psi}_j}) \hat{\boldsymbol{\theta}}_j$

2.4. 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

$$R_s = Y - \widetilde{Y}_{(-s)}^{(k)}, \tag{7}$$

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

$$\widetilde{Y}_{(-s)}^{(k)} = \beta_E^{(k)} X_E + \sum_{\ell \neq s} \boldsymbol{\Psi}_\ell \boldsymbol{\theta}_\ell^{(k)} + \sum_{\ell \neq s} \gamma_\ell^{(k)} \beta_E^{(k)} \widetilde{\boldsymbol{\Psi}}_\ell \boldsymbol{\theta}_\ell^{(k)}. \tag{8}$$

Using (8), (7) can be re-written as

$$\begin{aligned} R_s &= Y - \beta_E^{(k)} X_E - \sum_{j=1}^{p} (\boldsymbol{\Psi}_j + \gamma_j^{(k)} \beta_E^{(k)} \widetilde{\boldsymbol{\Psi}}_j) \boldsymbol{\theta}_j^{(k)} + (\boldsymbol{\Psi}_s + \gamma_s^{(k)} \beta_E^{(k)} \widetilde{\boldsymbol{\Psi}}_s) \boldsymbol{\theta}_s^{(k)} \\ &= R^* + (\boldsymbol{\Psi}_s + \gamma_s^{(k)} \beta_E^{(k)} \widetilde{\boldsymbol{\Psi}}_s) \boldsymbol{\theta}_s^{(k)}, \end{aligned} \tag{9}$$

where

$$R^* = Y - \beta_E^{(k)} X_E - \sum_{j=1}^{p} (\boldsymbol{\Psi}_j + \gamma_j^{(k)} \beta_E^{(k)} \widetilde{\boldsymbol{\Psi}}_j) \boldsymbol{\theta}_j^{(k)}. \tag{10}$$

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

$$\boldsymbol{\theta}_s^{(k)(new)} = \arg\min_{\boldsymbol{\theta}_j} \frac{1}{2n} \| R_s - (\boldsymbol{\Psi}_s + \gamma_s^{(k)} \beta_E^{(k)} \widetilde{\boldsymbol{\Psi}}_s) \boldsymbol{\theta}_j \|_2^2 + \lambda (1-\alpha) w_s \| \theta_j \|_2. \tag{11}$$

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

$$R_{s+1} = R^* + (\boldsymbol{\Psi}_{s+1} + \gamma_{s+1}^{(k)} \beta_E^{(k)} \widetilde{\boldsymbol{\Psi}}_{s+1}) \boldsymbol{\theta}_{s+1}^{(k)} + (\boldsymbol{\Psi}_s + \gamma_s^{(k)} \beta_E^{(k)} \widetilde{\boldsymbol{\Psi}}_s)(\boldsymbol{\theta}_s^{(k)} - \boldsymbol{\theta}_s^{(k)(new)}), \tag{12}$$

where R^* is given by (10), $\boldsymbol{\theta}_s^{(k)}$ is the parameter value prior to the update, and $\boldsymbol{\theta}_s^{(k)(new)}$ is the updated value given by (11). Taking the difference between (9) and (12) gives

$$\begin{aligned} \Delta &= R_t - R_s \\ &= (\boldsymbol{\Psi}_t + \gamma_t^{(k)} \beta_E^{(k)} \widetilde{\boldsymbol{\Psi}}_t) \boldsymbol{\theta}_t^{(k)} + (\boldsymbol{\Psi}_s + \gamma_s^{(k)} \beta_E^{(k)} \widetilde{\boldsymbol{\Psi}}_s)(\boldsymbol{\theta}_s^{(k)} - \boldsymbol{\theta}_s^{(k)(new)}) - (\boldsymbol{\Psi}_s + \gamma_s^{(k)} \beta_E^{(k)} \widetilde{\boldsymbol{\Psi}}_s) \boldsymbol{\theta}_s^{(k)} \\ &= (\boldsymbol{\Psi}_t + \gamma_t^{(k)} \beta_E^{(k)} \widetilde{\boldsymbol{\Psi}}_t) \boldsymbol{\theta}_t^{(k)} - (\boldsymbol{\Psi}_s + \gamma_s^{(k)} \beta_E^{(k)} \widetilde{\boldsymbol{\Psi}}_s) \boldsymbol{\theta}_s^{(k)(new)}. \end{aligned} \tag{13}$$

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

2.5. Weak heredity

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

2.6. Adaptive `sail`

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

1. Prior knowledge about the importance of certain variables is known. Larger weights will penalize the variable more, while smaller weights will penalize the variable less
2. Allows users to apply the adaptive `sail`, similar to the adaptive lasso (Zou, 2006)

We describe the adaptive `sail` in Algorithm 2. This is a general procedure that can be applied to the weak and strong heredity settings. We provide this capability in the `sail` package using the `penalty.factor` argument.

2.7. Flexible design matrix

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

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
$$w_E = \left(|\widehat{\beta}_E^{[opt]}| + 1/n\right)^{-1},\ w_j = \left(\|\widehat{\boldsymbol{\theta}}_j^{[opt]}\|_2 + 1/n\right)^{-1},\ w_{jE} = \left(\|\widehat{\boldsymbol{\tau}}_j^{[opt]}\|_2 + 1/n\right)^{-1} \text{ for } j = 1, \ldots, p$$
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

are often situations when the data contains a combination of categorical and continuous variables. In these cases it may be sub-optimal to apply a basis expansion to the categorical variables. Owing to the flexible nature of our algorithm, we can handle this scenario in our implementation by allowing a user-defined design matrix. The only extra information needed is the group membership of each column in the design matrix. We illustrate such an example in a vignette of the `sail` R package.

3. Theory

In this section we study the asymptotic behavior of the `sail` estimator $\widehat{\boldsymbol{\Phi}}$, defined as the minimizer of (4), as well as the model selection properties. We show that `sail` possesses the oracle property when the sample size approaches infinity and the number of predictors is fixed. That is, under certain regularity conditions, it performs as well as if the true model were known in advance and has the optimal estimation rate (Zou, 2006). The regularity conditions and proofs are given in Supplemental Section 1.

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

$$\widehat{\boldsymbol{\Phi}}_n = \arg\min_{\boldsymbol{\Phi}} Q_n(\boldsymbol{\Phi}) = -L_n(\boldsymbol{\Phi}) + n\lambda_m \sum_{m=1}^{2p+1} \left\|\boldsymbol{\phi}_m\right\|_2, \tag{14}$$

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

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

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

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

and

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

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

Lemma 1. *[Existence of a local minimizer] If $a_n = o(\frac{1}{\sqrt{n}})$ as $n \to \infty$, i.e. $\sqrt{n}a_n \to 0$, then $\|\widehat{\boldsymbol{\Phi}}_n - \boldsymbol{\Phi}^*\|_2 = O_p(\frac{1}{\sqrt{n}})$.*

Lemma 1 states that if the tuning parameters corresponding to the non-zero coefficients converge to 0 at a speed faster than $\frac{1}{\sqrt{n}}$, then there exists a local minimizer of $Q_n(\boldsymbol{\Phi})$ which is $\sqrt{n}$-consistent (Wang et al., 2007; Choi et al., 2010).

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

$$P\left(\widehat{\boldsymbol{\Phi}}_{\mathcal{A}_1^c} = \mathbf{0}\right) \to 1 \quad \textit{and} \quad P\left(\widehat{\boldsymbol{\Phi}}_{\mathcal{A}_2^c} = \mathbf{0}\right) \to 1. \tag{15}$$

Theorem 1 shows that `sail` can consistently remove the main effects and interaction terms which are not associated with the response with high probability. Together with Lemma 1, we see that the asymptotic behavior of the penalty terms for the zero and non-zero predictors must be different to satisfy the model selection consistency property (15) (Nardi and Rinaldo, 2008). Specifically, when the tuning parameters for the non-zero coefficients converge to 0 faster than $1/\sqrt{n}$ (i.e.

$\sqrt{n}a_n \to 0$) and those for zero coefficients are large enough (i.e. $\sqrt{n}b_n \to \infty$), the Lemma 1 and Theorem 1 imply that the $\sqrt{n}$-consistent estimator $\widehat{\mathbf{\Phi}}_n$ satisfies $P\left(\widehat{\mathbf{\Phi}}_{\mathcal{A}_2^c} = \mathbf{0}\right) \to 1$.

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

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

$$\sqrt{n}\left(\widehat{\mathbf{\Phi}}_{\mathcal{A}} - \mathbf{\Phi}^*_{\mathcal{A}}\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \mathbf{I}^{-1}\left(\mathbf{\Phi}^*_{\mathcal{A}}\right)\right), \tag{16}$$

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

Together, Theorems 1 and 2 establish that if the tuning parameters satisfy the conditions $\sqrt{n}a_n \to 0$ and $\sqrt{n}b_n \to \infty$, then as the sample size grows large, `sail` has the oracle property (Fan and Li, 2001). In order for the conditions on the tuning parameters to be satisfied, we follow the strategies outlined for the adaptive Lasso (Zou, 2006), the adaptive group Lasso (Nardi and Rinaldo, 2008) and the adaptive elastic-net (Zou and Zhang, 2009). That is, we define the adaptive weights as $w_m = \|\widehat{\boldsymbol{\phi}}_m^{\text{init}} + 1/n\|_2^{-\xi}$ for $m = 1, \ldots, 2p+1$, where ξ is a positive constant and $\widehat{\boldsymbol{\phi}}_m^{\text{init}}$ is an initial $\sqrt{n}$-consistent estimate of $\boldsymbol{\phi}^*_m$. Here, the $1/n$ is to avoid division by zero.

4. Simulation study

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

4.1. Comparator methods

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

1. Linear main effects: `lasso` (Tibshirani, 1996), `adaptive lasso` (Zou, 2006)
2. Linear interactions: `lassoBT` (Shah, 2016), `GLinternet` (Lim and Hastie, 2015)
3. Non-linear main effects: `HierBasis` (Haris et al., 2019), `SPAM` (Ravikumar et al., 2009), `gamsel` (Chouldechova and Hastie, 2015)

For `GLinternet` we specified the `interactionCandidates` argument so 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 (`R` code for each method available at https://github.com/sahirbhatnagar/sail/blob/master/my_sims/method_functions.R). `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` select whether a term in an additive model is non-zero, 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 method `sail`, as well as with `adaptive sail` (Algorithm 2) and `sail weak` which has the weak heredity property. For each function f_j, we use a B-spline basis matrix with `degree=5` implemented in the `bs` function in `R` (R Core Team, 2017). We center the environment variable and the basis functions before running the `sail` method.

4.2. Simulation design

To make the comparisons with other methods as fair as possible, we followed a simulation framework that has been previously used for variable selection methods in additive models (Lin and Zhang, 2006; Huang et al., 2010). We extend this framework to include interaction effects as well. The covariates are simulated as follows. First, we generate $x_1, \ldots, x_{1000}$ independently from a standard normal distribution truncated to the interval [0,1] for $i = 1, \ldots, n$. The first four variables are non-zero (i.e. active in the response), while the rest of the variables are zero (i.e. are noise variables). The exposure variable (X_E) is generated from a standard normal distribution truncated to the interval [-1,1]. The outcome Y is then generated following one of the models and assumptions described below. We evaluate the performance of our method on three of its defining characteristics: 1) the strong heredity property, 2) non-linearity of predictor effects and 3) interactions. Simulation scenarios are designed specifically to test the performance of these characteristics.

1. **Heredity simulation**

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

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

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

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

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

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

2. **Non-linearity simulation scenario**

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

   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 (Lin and Zhang, 2006; Huang et al., 2010) 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.

4.3. Results

The prediction accuracy and variable selection results for each of the five simulation scenarios are shown in Fig. 2 and Table 2, respectively. 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 a strong heredity (scenario 1a), as we would expect, since this is exactly the scenario that our method is trying to target. 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 being violated in scenario 1c), no method can identify the correct model with the exception of `GLinternet`. When only linear effects and interactions are present (scenario 2), we see that `adaptive sail` has similar MSE compared to the other linear interaction methods (`lassoBT` and `GLinternet`) with a better TPR and FPR. It is important to note that the variable selection performance of `sail` is highly dependent on being able to correctly select the exposure variable (X_E). In Supplemental Section C, we show the selection rates of X_E. We see that `sail` is able to consistently identify the exposure variable across all simulation scenarios and replications. Overall, our simulation study results suggest that `sail` outperforms existing methods when the true model contains non-linear interactions, and is competitive even when the truth only has either linear or additive main effects.

We also plotted the true and predicted curves for scenario 1a) in Supplemental Section C, to visually inspect whether our method could correctly capture the shape of the association between the predictors and the response for both main and interaction effects. In general, we see the non-linear effects are clearly being captured by `sail`.

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

	Linear Main Effects		Linear Interactions		Non-linear Main Effects			Non-linear Interactions		
	lasso	adaptive lasso	lassoBT	GLinternet	HierBasis	SPAM	gamsel	sail	adaptive sail	sail weak
1a) Strong heredity ($\|\mathcal{J}\| = 7$)										
$\|\widehat{\mathcal{J}}\|$	28 (15)	8 (4)	35 (18)	40 (20)	133 (48)	42 (19)	46 (21)	37 (15)	8 (3)	21 (3)
TPR	53.9 (8.4)	49.3 (10.1)	61.7 (11.5)	66.4 (14.0)	65.2 (8.1)	60.9 (8.5)	56.9 (7.7)	89.5 (8.2)	81.4 (13.0)	82.1 (10.9)
FPR	1.2 (0.7)	0.2 (0.2)	1.5 (0.9)	1.8 (1.0)						
1b) Weak heredity ($\|\mathcal{J}\| = 5$)										
$\|\widehat{\mathcal{J}}\|$	19 (12)	4 (2)	20 (13)	38 (23)	24 (23)	28 (16)	21 (15)	24 (19)	5 (3)	14 (10)
TPR	40.7 (3.6)	40.1 (1.4)	40.8 (3.8)	64.1 (14.9)	42.2 (6.3)	53.9 (9.4)	42.7 (6.8)	52.4 (11.4)	46.4 (10.1)	55.0 (13.7)
FPR	0.9 (0.6)	0.1 (0.1)	0.9 (0.7)	1.7 (1.1)	1.1 (1.1)	1.2 (0.8)	1.0 (0.7)	1.0 (0.9)	0.2 (0.1)	0.6 (0.5)
1c) Interactions Only ($\|\mathcal{J}\| = 2$)										
$\|\widehat{\mathcal{J}}\|$	12 (12)	3 (2)	14 (13)	38 (21)	12 (13)	13 (12)	12 (12)	10 (18)	2 (2)	26 (30)
TPR	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	81.4 (27.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	1.0 (6.9)	0.0 (0.0)	22.9 (36.9)
FPR	0.6 (0.6)	0.6 (6.9)	0.7 (0.7)	1.8 (1.0)	0.6 (0.7)	0.7 (0.6)	0.6 (0.6)	0.5 (0.9)	0.1 (0.1)	1.3 (1.5)
2) Linear Effects ($\|\mathcal{J}\| = 7$)										
$\|\widehat{\mathcal{J}}\|$	37 (17)	8 (3)	48 (19)	51 (23)	37 (19)	42 (19)	37 (16)	34 (18)	11 (4)	20 (4)
TPR	70.4 (3.7)	67.2 (6.7)	72.3 (6.3)	93.4 (8.5)	70.3 (3.8)	65.0 (8.1)	70.4 (3.7)	93.9 (9.9)	86.0 (18.5)	68.1 (14.9)
FPR	1.6 (0.8)	0.2 (0.2)	2.2 (1.0)	2.2 (1.2)	1.6 (0.9)	1.9 (0.9)	1.6 (0.8)	1.4 (0.9)	0.2 (0.2)	0.7 (0.2)
3) Main Effects Only ($\|\mathcal{J}\| = 5$)										
$\|\widehat{\mathcal{J}}\|$	29 (14)	7 (4)	31 (15)	34 (18)	154 (17)	46 (21)	56 (20)	44 (19)	9 (2)	22 (2)
TPR	75.9 (10.9)	66.5 (15.3)	76.0 (10.9)	77.0 (9.5)	97.5 (6.6)	93.1 (10.7)	81.3 (9.5)	91.5 (10.3)	84.1 (9.2)	85.2 (12.1)
FPR	1.3 (0.7)	0.2 (0.2)	1.3 (0.8)	1.5 (0.9)	7.5 (0.9)	2.1 (1.0)	2.6 (1.0)	2.0 (0.9)	0.2 (0.1)	0.9 (0.1)

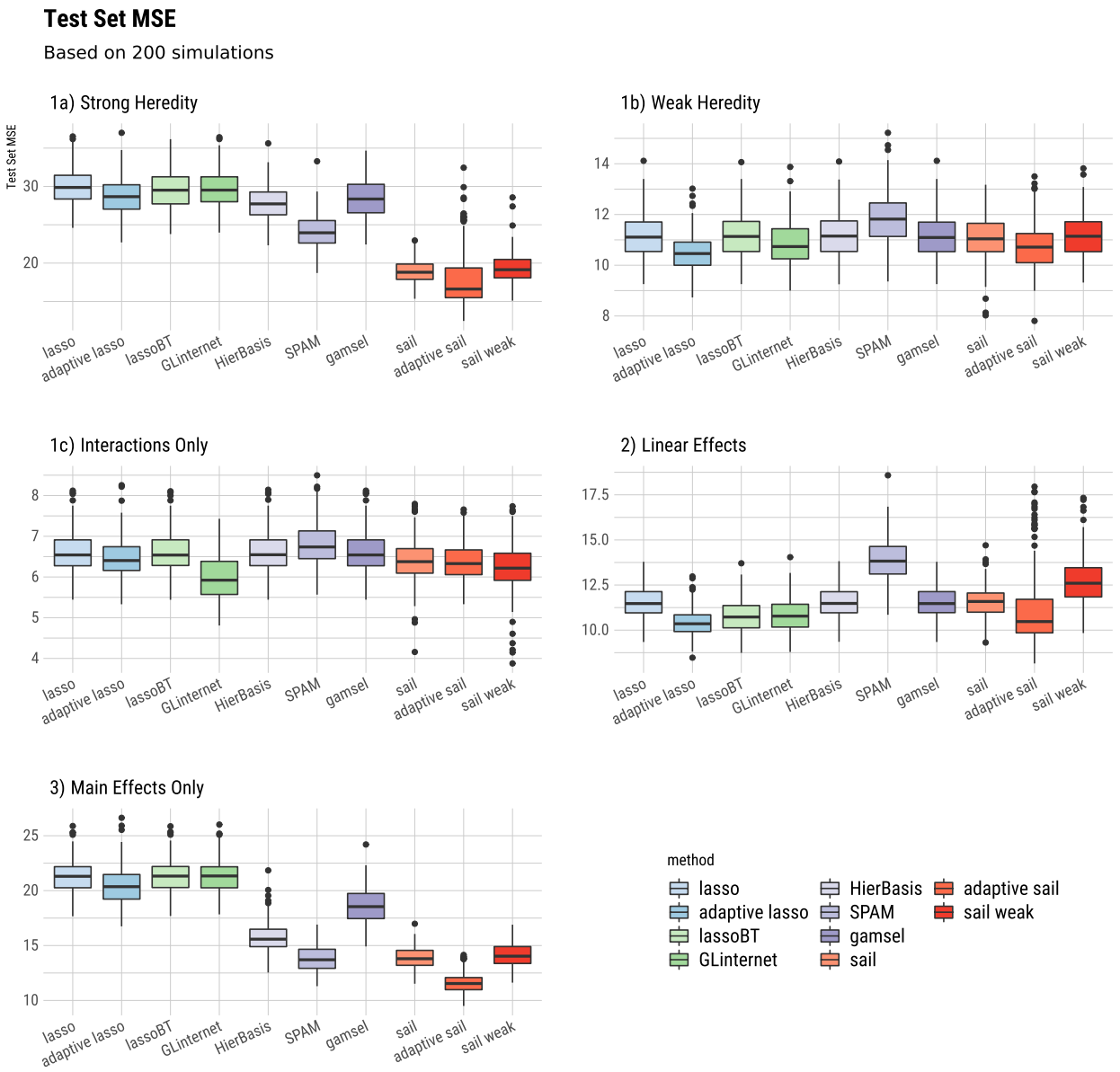

Fig. 2. Boxplots of the test set mean squared error from 200 replications for each of the five simulation scenarios.

5. Real data applications

5.1. Gene-environment interactions in the nurse family partnership program

It is well known that environmental exposures can have an important impact on academic achievement. Indeed, early intervention in young children has been shown to positively impact intellectual abilities (Campbell and Ramey, 1994). More recent studies have shown that cognitive performance, a trait that measures the ability to learn, reason and solve problems, is also strongly influenced by genetic factors. Genome-wide association studies (GWAS) suggest that 20% of the variance in educational attainment (years of education) may be accounted for by common genetic variation (Rietveld et al., 2013; Okbay et al., 2016). Unsurprisingly, there is significant overlap in the SNPs that predict educational attainment and measures of cognitive function. An interesting query that arises is how the environment interacts with these genetics variants to predict measures of cognitive function. To address this question, we analyzed data from the Nurse Family Partnership (NFP), a psychosocial intervention program that begins in pregnancy and targets maternal health, parenting and mother-infant interactions (Olds et al., 1998). The Stanford Binet IQ scores at 4 years of age were collected for 189 subjects (including 19 imputed using `mice` (Buuren and Groothuis-Oudshoorn, 2010)) born to women randomly assigned to control ($n = 100$) or nurse-visited intervention groups ($n = 89$). For each subject, we calculated a polygenic risk score (PRS) for educational

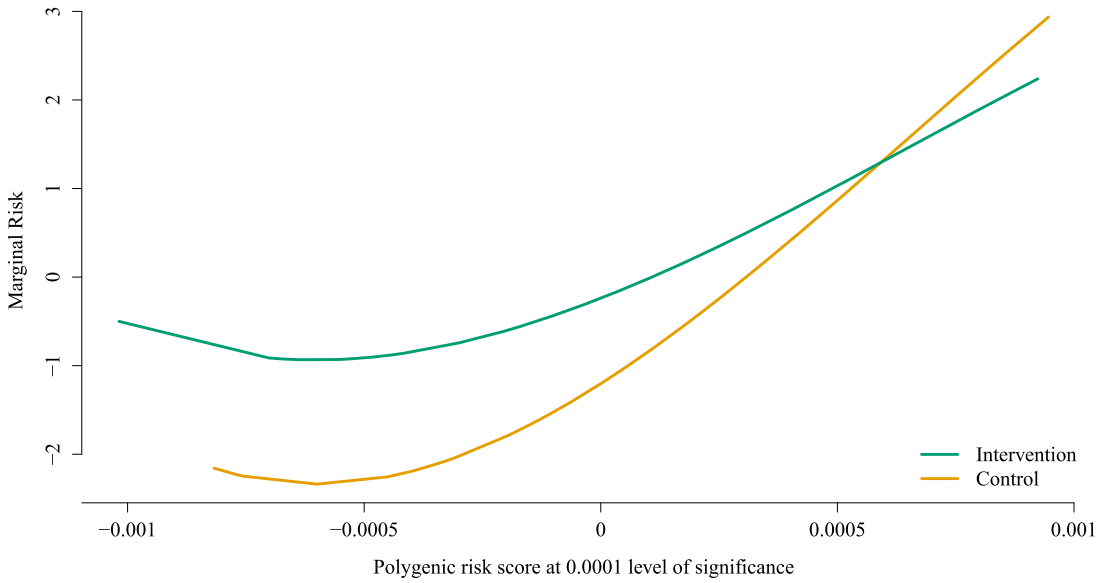

Fig. 3. Estimated interaction effect identified by the weak heredity `sail` using cubic B-splines and $\alpha = 0.1$ for the Nurse Family Partnership data. The selected model, chosen via 10-fold cross-validation, contained three variables: the main effects for the intervention and the PRS for educational attainment using genetic variants significant at the 0.0001 level, as well as their interaction.

attainment at nine different p-value thresholds using weights from the GWAS conducted in Okbay et al. 2016. We applied the weak heredity `sail` with cubic B-splines and $\alpha = 0.1$ to encourage interactions, and selected the optimal tuning parameter using 10-fold cross-validation. This resulted in a total of 55 parameters to estimate. In this context, individuals with a higher PRS have a propensity for higher educational attainment. The goal of this analysis was to determine if there was an interaction between genetic predisposition to educational attainment (X) and maternal participation in the NFP program (E) on child IQ at 4 years of age (Y). Our method identified an interaction between the intervention and PRS which included genetic variants at the 0.0001 level of significance. This interaction is shown in Fig. 3. We see that the intervention has a much larger effect on IQ for lower PRS compared to a higher PRS. In other words, perinatal home visitation by nurses can impact IQ scores in children who are genetically predisposed to lower educational attainment. Similar results were obtained for the other imputed datasets (Supplemental Section D). We also compared `sail` with two other interaction selection methods, `lassoBT` and `GLinternet` with default settings, on 200 bootstrap samples of the data. The average and standard deviation of the MSE and size of the active set ($|\widehat{\mathcal{J}}|$) across the 200 bootstrap samples are given in Table 3. We see that sail tends to select sparser models while maintaining similar prediction performance compared to `lassoBT`. The `GLinternet` statistics are omitted here since the algorithm did not converge for many of the 200 simulations.

5.2. Study to understand prognoses preferences outcomes and risks of treatment

The Study to Understand Prognoses Preferences Outcomes and Risks of Treatment (SUPPORT) aimed at identifying which clinical variables influence medium-term (half-year) mortality rate amongst seriously ill hospitalized patients and improving clinical decision making (Connors et al., 1995). With a relatively large sample size of 9,105 and detailed documentation of clinical variables, the SUPPORT dataset allows detection of potential interactions using the strategy implemented in `sail`. We applied `sail` to test for non-linear interactions between acute renal failure or multiple organ system failure (ARF/MOSF), an important predictor for survival rate, and 13 other variables that were deemed clinically relevant. These variables included the number of comorbidities (excluding ARF/MOSF), age, sex, as well as multiple physiological and blood biochemical indices. The response was whether a patient survived after six months since hospitalization.

A total of 8,873 samples had complete data on all variables of interest. We randomly divided these samples into equal sized training/validation/test splits and ran `lassoBT`, `GLinternet`, and the weak heredity `sail` with cubic B-splines and $\alpha = 0.1$ (as was done in the Nurse Family Partnership program case study). A binomial distribution family was specified for `GLinternet`, whereas `lassoBT` had the same default settings as the simulation study since it did not support a specialized implementation for binary outcomes. We again ran each method on the training data, determined the optimal tuning parameter on the validation data based on the area under the receiver operating characteristic curve (AUC), and assessed AUC on the test data. We repeated this process 200 times and report the results in Table 3. We found that `sail` achieved similar prediction accuracy to `lassoBT` and `GLinternet`. However, the predictive performance of `lassoBT` and `GLinternet` relied on models which included many more variables. In Fig. 4, we visualize the two strongest interaction effects associated with the number of comorbidities and age, respectively. For those having undergone ARF/MOSF, an increased number of comorbidities decreases their chance of survival, while there seems to be no such relationship for non-ARF/MOSF

Table 3
Comparison of analytic methods for selecting interactions using the Nurse Family Partnership program and the SUPPORT datasets. Averages (standard deviations in parentheses) are based on 200 bootstrap samples. $|\widehat{\mathcal{J}}|$ is the number of variables selected by the method. `GLinternet` results not reported for NFP data since the algorithm did not converge in many of the bootstrap samples.

Method	Nurse Family Partnership		SUPPORT	
	Mean Squared Error	$\lvert\widehat{\mathcal{J}}\rvert$	AUC	$\lvert\widehat{\mathcal{H}}\rvert$
`sail`	3.5 (0.6)	4 (3)	0.66 (0.01)	25 (3)
`lassoBT`	3.53 (0.477)	11 (6)	0.65 (0.009)	49 (14)
`GLinternet`	–	–	0.65 (0.009)	58 (7)

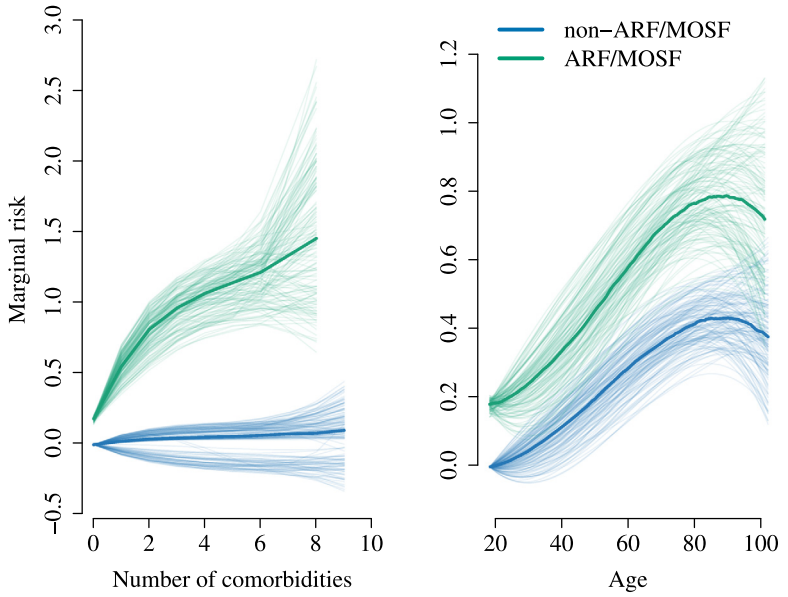

Fig. 4. Illustration of estimated interaction effects identified by `sail` for the SUPPORT data. Median prediction curves in dark colors based on 200 train/validate/test splits represent the estimated marginal interaction effects. Coefficients estimated in each of the 200 train/validate/test splits were used to generate prediction curves representing a 90% confidence interval colored in corresponding light colors.

patients. The interaction between ARF/MOSF and age shows the risk incurred by ARF/MOSF is most distinguishing among patients between the ages of 70 and 80.

6. Discussion

In this article we have introduced the sparse additive interaction learning model `sail` for detecting non-linear interactions with a key environmental or exposure variable in high-dimensional settings. Using a simple reparametrization, we are able to achieve either the weak or strong heredity property without using a complex penalty function. We developed a blockwise coordinate descent algorithm to solve the `sail` objective function for the least-squares loss. We further studied the asymptotic properties of our method and showed that under certain conditions, it possesses the oracle property. All our algorithms have been implemented in a computationally efficient, well-documented and freely available `R` package on CRAN. Furthermore, our method is flexible enough to handle any type of basis expansion including the identity map, which allows for linear interactions. Our implementation allows the user to selectively apply the basis expansions to the predictors, allowing for example, a combination of continuous and categorical predictors. An extensive simulation study shows that `sail`, `adaptive sail` and `sail weak` outperform existing penalized regression methods in terms of prediction accuracy, sensitivity and specificity when there are non-linear main effects only, as well as interactions with an exposure variable. We then demonstrated the utility of our method to identify non-linear interactions in both biological and epidemiological data. In the NFP program, we showed that individuals who are genetically predisposed to lower educational attainment are those who stand to benefit the most from the intervention. Analysis of the SUPPORT data revealed that those having undergone ARF/MOSF, an increased number of comorbidities decreased their chances of survival, while there seemed to be no such relationship for non-ARF/MOSF patients. In a bootstrap analysis of both datasets, we observed that `sail`tended to select sparser models while maintaining similar prediction performance compared to other interaction selection methods.

Our method however does have its limitations. `sail` can currently only handle $X_E \cdot f(X)$ or $f(X_E) \cdot X$ and does not allow for $f(X, X_E)$, i.e., only one of the variables in the interaction can have a non-linear effect and we do not consider the

tensor product. The reparametrization leads to a non-convex optimization problem which makes convergence rates difficult to assess, though we did not experience any major convergence issues in our simulations and real data analysis. The memory footprint can also be an issue depending on the degree of the basis expansion and the number of variables. Furthermore, the functional form of the covariate effects is treated as known in our method. Being able to automatically select for example, linear vs. nonlinear components, is currently an active area of research in main effects models (Haris et al., 2019). 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.

Acknowledgements

SRB and CMTG were supported by the Ludmer Centre for Neuroinformatics and Mental Health and the Canadian Institutes for Health Research PJT 148620. SRB acknowledges the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), RGPIN-2020-05133. This research was enabled in part by support provided by Calcul Québec (www.calculquebec.ca) and Compute Canada (www.computecanada.ca). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.csda.2022.107624.

References

Bach, F., Jenatton, R., Mairal, J., Obozinski, G., et al., 2012. Structured sparsity through convex optimization. Stat. Sci. 27, 450–468.
Bhatnagar, S.R., Yang, Y., Khundrakpam, B., Evans, A.C., Blanchette, M., Bouchard, L., Greenwood, C.M., 2018. An analytic approach for interpretable predictive models in high-dimensional data in the presence of interactions with exposures. Genet. Epidemiol. 42, 233–249.
Bien, J., Taylor, J., Tibshirani, R., et al., 2013. A lasso for hierarchical interactions. Ann. Stat. 41, 1111–1141.
Bühlmann, P., Van De Geer, S., 2011. Statistics for High-Dimensional Data: Methods, Theory and Applications. Springer Science & Business Media.
Buuren, S.v., Groothuis-Oudshoorn, K., 2010. mice: multivariate imputation by chained equations in r. J. Stat. Softw., 1–68.
Campbell, F.A., Ramey, C.T., 1994. Effects of early intervention on intellectual and academic achievement: a follow-up study of children from low-income families. Child Dev. 65, 684–698.
Chipman, H., 1996. Bayesian variable selection with related predictors. Can. J. Stat. 24, 17–36.
Choi, N.H., Li, W., Zhu, J., 2010. Variable selection with the strong heredity constraint and its oracle property. J. Am. Stat. Assoc. 105, 354–364.
Chouldechova, A., Hastie, T., 2015. Generalized additive model selection. arXiv preprint. arXiv:1506.03850.
Connors, A.F., Dawson, N.V., Desbiens, N.A., Fulkerson, W.J., Goldman, L., Knaus, W.A., Lynn, J., Oye, R.K., Bergner, M., Damiano, A., et al., 1995. A controlled trial to improve care for seriously iii hospitalized patients: the study to understand prognoses and preferences for outcomes and risks of treatments (support). JAMA 274, 1591–1598.
Cox, D.R., 1984. Interaction. Int. Stat. Rev., 1–24.
Fan, J., Han, F., Liu, H., 2014. Challenges of big data analysis. Nat. Sci. Rev. 1, 293–314.
Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. J. Am. Stat. Assoc. 96, 1348–1360.
Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. 33, 1.
Hao, N., Feng, Y., Zhang, H.H., 2018. Model selection for high-dimensional quadratic regression via regularization. J. Am. Stat. Assoc. 113, 615–625.
Haris, A., Shojaie, A., Simon, N., 2019. Nonparametric regression with adaptive truncation via a convex hierarchical penalty. Biometrika 106, 87–107.
Haris, A., Witten, D., Simon, N., 2016. Convex modeling of interactions with strong heredity. J. Comput. Graph. Stat. 25, 981–1004.
Hastie, T., Tibshirani, R., Wainwright, M., 2015. Statistical Learning with Sparsity: The Lasso and Generalizations. CRC Press.
Huang, J., Horowitz, J.L., Wei, F., 2010. Variable selection in nonparametric additive models. Ann. Stat. 38, 2282–2313.
Lim, M., Hastie, T., 2015. Learning interactions via hierarchical group-lasso regularization. J. Comput. Graph. Stat. 24, 627–654.
Lin, Y., Zhang, H.H., 2006. Component selection and smoothing in multivariate nonparametric regression. Ann. Stat. 34, 2272–2297.
McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models, vol. 37. CRC Press.
Nardi, Y., Rinaldo, A., 2008. On the asymptotic properties of the group lasso estimator for linear models. Electron. J. Stat. 2, 605–633.
Okbay, A., Beauchamp, J.P., Fontana, M.A., Lee, J.J., Pers, T.H., Rietveld, C.A., Turley, P., Chen, G.B., Emilsson, V., Meddens, S.F.W., et al., 2016. Genome-wide association study identifies 74 loci associated with educational attainment. Nature 533, 539.
Olds, D., Henderson Jr, C.R., Cole, R., Eckenrode, J., Kitzman, H., Luckey, D., Pettitt, L., Sidora, K., Morris, P., Powers, J., 1998. Long-term effects of nurse home visitation on children's criminal and antisocial behavior: 15-year follow-up of a randomized controlled trial. JAMA 280, 1238–1244.
R Core Team, 2017. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.
Radchenko, P., James, G.M., 2010. Variable selection using adaptive nonlinear interaction structures in high dimensions. J. Am. Stat. Assoc. 105, 1541–1553.
Ravikumar, P., Lafferty, J., Liu, H., Wasserman, L., 2009. Sparse additive models. J. R. Stat. Soc., Ser. B, Stat. Methodol. 71, 1009–1030.
Rietveld, C.A., Medland, S.E., Derringer, J., Yang, J., Esko, T., Martin, N.W., Westra, H.J., Shakhbazov, K., Abdellaoui, A., Agrawal, A., et al., 2013. Gwas of 126,559 individuals identifies genetic variants associated with educational attainment. Science 340, 1467–1471.
Shah, R.D., 2016. Modelling interactions in high-dimensional data with backtracking. J. Mach. Learn. Res. 17, 1–31.
She, Y., Jiang, H., 2014. Group regularized estimation under structural hierarchy. arXiv preprint. arXiv:1411.4691.
Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. B, 267–288.
Wang, H., Li, G., Tsai, C.L., 2007. Regression coefficient and autoregressive order shrinkage and selection via the lasso. J. R. Stat. Soc., Ser. B, Stat. Methodol. 69, 63–78.
Zhao, P., Rocha, G., Yu, B., 2009. The composite absolute penalties family for grouped and hierarchical variable selection. Ann. Stat., 3468–3497.
Zou, H., 2006. The adaptive lasso and its oracle properties. J. Am. Stat. Assoc. 101, 1418–1429.
Zou, H., Zhang, H.H., 2009. On the adaptive elastic-net with a diverging number of parameters. Ann. Stat. 37, 1733–1751.