Summer 7-23-2020

"A comparison of variable selection methods using bootstrap samples from environmental metal mixture data"

Paul-Yvann Djamen 4785403
University Of New Mexico

Paul-Yvann Djamen
University of New Mexico

Paul-Yvann Djamen

Candidate

Mathematics and Statistics

Department

This thesis is approved, and it is acceptable in quality and form for publication:

Approved by the Thesis Committee:

Dr. LI LI , Chairperson

Dr. James Degnan

Dr. Fletcher Christensen

Dr. Ronald Christensen

"A comparison of variable selection methods using bootstrap samples from environmental metal mixture data"

by

Paul-Yvann Djamen

B.S, Chemical engineering University of New Mexico, 2016

THESIS

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Master of Science

Statistics

The University of New Mexico

Albuquerque, New Mexico

December, 2018

Dedication

To my parents Gisele and Rene without whom I wouldn't be here, my advisor Dr. LI LI, my roommates as well my as biological family and church family for their helping me stay strong when the world seemed to fall apart. All these special individuals were brought in my life by GOD.

"I think therefore I am " – Rene Descartes

Acknowledgments

Special thanks to mom and dad, who helped shape the man I am. Hope to be the best example I can be someday as they were to me
Special thanks to my advisor Dr LI LI, for her support and firmness when she needed to be. She was also available outside of academia if need be.
Shout out thanks to all friends and my lady who helped me stay strong through unforeseen circumstances. Because I cannot thank you enough, I thank the Lord to have put you in my life.
I am also grateful to the UNM statistics program for their support and funding over the past semesters, as well as, the UNM METALS and UNM Comprehensive Cancer Center for their support and funding over the summer towards this project.

"A comparison of variable selection methods using bootstrap samples from environmental metal mixture data"

by

Paul-Yvann Djamen

B.S, Chemical engineering University of New Mexico, 2016

M.S., Statistics, University of New Mexico, 2018

Abstract

In this thesis, I studied a newly developed variable selection method SODA, and three customarily used variable selection methods: LASSO, Elastic net, and Random forest for environmental mixture data. The motivating datasets have neuro-developmental status as responses and metal measurements and demographic variables as covariates. The challenges for variable selections include (1) many measured metal concentrations are highly correlated, (2) there are many possible ways of modeling interactions among the metals, (3) the relationships between the outcomes and explanatory variables are possibly nonlinear, (4) the signal to noise ratio in the real data may be low. To compare these methods under the challenges, I simulated responses under various scenarios with covariates bootstrapped from real data and then compared the percentages of false positives and false negatives of these methods. I conclude that no method has the lowest percentage of false positives and false negatives at the same time across all scenarios. However, RF methods seem to have modest performances in both percentages, compared to SODA, LASSO, and Elastic net.

Contents

Contents

List of Figures

List of Tables

List of Tables

Chapter 1

Introduction

There has been increasing attention to identify and quantify risks from environmental exposures. Often exposures come from various sources, such as drugs, air pollutants, alcohol, tobacco, or even lifestyle factors. Among many studies, Environmental Influences on Child Health Outcomes (ECHO) focuses on how environmental factors may affect health outcomes around the time of birth as well as later in childhood or adolescence. As part of the ECHO group, Navajo Birth Cohort Study (NBCS), funded by NIEHS R25-ES013208, collected data from the Navajo nation to investigate how metal exposures affect children's neurodevelopmental conditions. Previous studies have reported evidence on the link between health outcomes and metal exposures related to Abandoned Uranium Mines(AUMs). The issue is of particular concern on the Navajo Nation where more than 500 AUMs remain as a legacy of Cold War mining. Exposure of community members to metal mixtures in AUM waste may contribute to diseases including hypertension, diabetes, and kidney disease. Considering the proximity of Navajo community members to AUMs, there is a need to systematically investigate the health impact of simultaneous exposures to multiple metals.

Environmental chemical exposures often occur in mixtures. Statistical methods for analysis of chemical mixtures that account for the complex characteristics of exposure profiles such as multicollinearity, high dimensionality, interactions, and non-linear dose response relationships are limited. More specifically, the challenges I face in analyzing environmental mixtures include, but are not limited to: (1) many measured metal concentrations were highly correlated; (2) there were many possible ways for metals to interact; (3) the relationships between the outcomes and explanatory variables were possibly nonlinear; (4) the signal to noise ratio in the real data may be low. Guidelines on the selection of appropriate methods have not been well established. Because current methods of analysis require a comprehensive assessment of environmental exposure profiles in search of disease risk factors, there is a need for identification and development of a toolbox of statistical approaches that account for the unique study design and complex analytical challenges.

In statistical analysis, regression techniques are widely used to explore relationships between response variables and covariates. For a continuous response, the statistical framework can be stated as

$$y = f(x_1, x_2, \ldots, x_p) + \epsilon,$$

where y is the response variable, $x_1, x_2, \ldots, x_p$ are covariates, and ϵ is an error term. For a binary response, often a logistic regression is used:

$$P(y = 1) = \frac{\exp(f(x_1, x_2, \ldots, x_p))}{1 + \exp(f(x_1, x_2, \ldots, x_p))}.$$

When multiple covariates are available ($p > n$) and especially when the number of covariates is more than the number of subjects in the observed data, variable selection is a classical way to reduce the dimension of covariates and to identify variables that are possibly related to the response.

There are many variable selection methods for either continuous response or binary response when the assumed response function, $f(\mathbf{x})$ is linear. Classical methods

include, but are not limited to, forward and backward stepwise selection using some information criteria. The backward stepwise selection method is not applicable when the number of covariates available is greater than the sample size ($p > n$). Penalized regression methods, for example, LASSO and its many variants can perform variable selection when $p > n$. They also perform model estimations and variable selections simultaneously, in contrast to separate steps in forward and backward selection methods. Since LASSO has been reported to perform poorly when the covariates are highly correlated, one of its variants, Elastic Net has been customarily recommended for highly correlated covariates. There are also methods that can perform variable selection without assuming a parametric form for $f(\mathbf{x})$. Among this category, random forest methods have been a popular method. Most random forest variable selection methods are based on variable importance measures. A recently published article [8] proposed SODA for binary response in the context of logistic regression and S-SODA for continuous response based on sliced responses and multi-category logistic regression. SODA and S-SODA are based on a combination of forward and backward procedure and information criteria EBIC. The acclaimed features are that these two methods can handle large $p > n$ and highly correlated covariates. In addition, S-SODA avoids assuming linearity in $f(\mathbf{x})$ which makes it applicable when the true $f(\mathbf{x})$ is nonlinear. Their simulation studies, as well as real-data applications, demonstrate superior performances of SODA in dealing with non-Gaussian design matrices in the logistic model for binary responses, linear and nonlinear models for continuous responses.

The objective of this thesis is to evaluate the performance of different variable selection methods in facing these challenges. Specifically, I performed a simulation study to compare SODA/S-SODA to LASSO, Elastic Net, and Random forest methods. The designed simulation used bootstrapped covariate samples to mimic the correlation and distribution structures in the data. To simulate the responses, I considered $f(\mathbf{x})$ functions that include linear and nonlinear scenarios for continuous

responses and linear for binary responses, terms that include main effects, quadratic effects, and interaction effects, and various noise to signal ratio. The rest of the thesis is organized as follows: chapter 2 outlines the methods I compare; chapter 3 presents the simulation studies; chapter 4 concludes the thesis.

Chapter 2

Methods in comparison

The methods under comparison were SODA, Elastic Net, Lasso, and Random Forest. In the following sections, I outline the main ideas of each method.

2.1 SODA

SODA [8] which stands for Stepwise Conditional Likelihood for Discriminant Analysis, is a tool that helps assess the performance of classification or regression by detecting the main effect and quadratic interaction terms under high dimensional settings. It is further characterized by stepwise model selection technique using forward and backward stages to tune and optimize the final model it selects based on a modified version of the BIC criterion called Extended BIC [8].

Let $((\mathbf{x}_i, y_i)), i = 1, \ldots, n$ denote n independent observations of $(\mathbf{X}, Y)$ where Y is categorical. Let p be the dimension of $\mathbf{x}$ and $\beta_k = (\beta_{k,1}, \ldots, \beta_{k,p})'$ and $A_k = A_{k,i,j}, i = 1, \ldots, p; j = 1, \ldots, p$ Consider the logistic model

$$P(Y = k|\mathbf{x}, \boldsymbol{\theta}) = \frac{e^{\delta_k(\mathbf{x}|\boldsymbol{\theta})}}{1 + \sum_{l=1}^{K-1} e^{\delta_l(\mathbf{x}|\boldsymbol{\theta})}}$$

where $\delta_k(\mathbf{x}|\boldsymbol{\theta})$ is the discriminant function for class k and $\boldsymbol{\theta}$ denotes the vector of parameters. Choosing class K as baseline and $\delta_K(\mathbf{x}|\boldsymbol{\theta}) = 0$, the method assumes that $\delta_k(\mathbf{x}|\boldsymbol{\theta}) = \alpha_k + \beta_k'\mathbf{x} + \mathbf{x}'A_k\mathbf{x}$ for $k = 1, \ldots, K-1$. Where A_k and β represent regression coefficients. When $K = 2$, it is a simple logistic regression.

Let $\mathcal{M}$ and $\mathcal{I}$ denote subsets of main effects and interaction pairs, respectively. Let $\mathcal{M}_0$ and $\mathcal{I}_0$ denote the corresponding true sets defined as

$$\mathcal{M}_0 = \{j : \exists k \text{ s.t. } \beta_{k,j} \neq 0\}, \text{ and } \mathcal{I}_0 = \{j : \exists k \text{ s.t. } A_{k,i,j} \neq 0\}$$

with k indicating the class label. Let $\mathcal{S} = \mathcal{M} \cup \mathcal{I}$ denote the set of all effects and let $\mathcal{A} = \mathcal{M}_0 \cup \mathcal{I}_0$ denote the true set of all effects. The true set of relevant predictors $\mathcal{P}$ is

$$\mathcal{P} = \mathcal{M}_0 \cup \{j : \exists i \text{ s.t. } (i, j) \in \mathcal{I}_0\}.$$

Let $\boldsymbol{\theta}_\mathcal{S}$ denote the collection of all coefficients in the model, whose 0's correspond to terms not in $\mathcal{S}$, and let $\boldsymbol{\theta}_{k,\mathcal{S}}$ denote the corresponding coefficients for class k. Let $\mathbf{Z} = (1, \mathbf{X}, \mathbf{X} \otimes \mathbf{X})$ be the augmented version of $\mathbf{X}$, containing 1, main effects, and all interaction terms of $\mathbf{X}$. Let $\mathbf{z}_i$ be the $i-$th observation of $\mathbf{Z}$. Then for a dataset $\{(x_i, y_i) : i = 1, \ldots, n\}$, the log-likelihood for $\boldsymbol{\theta}_\mathcal{S}$ is denoted as $l_n(\boldsymbol{\theta}_\mathcal{S})$ is

$$l_n(\boldsymbol{\theta}_\mathcal{S}) = \sum_{i=1}^{n} \left\{ \boldsymbol{\theta}_{y_i,\mathcal{S}}^T \mathbf{Z}_i - \log\left(1 + \sum_{l=1}^{K-1} \exp(\boldsymbol{\theta}_{l,\mathcal{S}}^T \mathbf{z}_i)\right) \right\}$$

Let $\tilde{\boldsymbol{\theta}}_\mathcal{S}$ denote the MLE of $\boldsymbol{\theta}_\mathcal{S}$. [2] proposed extended BIC (EBIC) and showed it to be consistent for linear regression models under high-dimensional settings. Following [2], and [8] shows that EBIC is consistent for the logistic regression models under some conditions. In particular, $\gamma > 2 - 1/2\kappa$ where $p < n^\kappa$

The EBIC for the set of $\mathcal{S}$ is defined as $\text{EBIC}_\gamma = -2l_n(\tilde{\boldsymbol{\theta}}_\mathcal{S}) + |\mathcal{S}|\log(n) + 2\gamma|\mathcal{S}|\log(p)$ where $|\mathcal{S}|$ is the the number of predictors that has at least main or interaction effect; p is the number of predictors; γ is a tuning parameter whit recommended values 0, 0.5 or 1. EBIC differs from regular BIC only by an extra penalty term, $2\gamma|\mathcal{S}|\log p$, which penalizes for having nonzero coefficients.

The authors proposed a stepwise procedure SODA, which consists of three stages: (1) preliminary forward main effect selection; (2) forward variable selection (considering both main and interaction effects); and (3) backward elimination.

1. Preliminary main effect selection: this step is the same as standard stepwise regression method. Let M_t denote set of main effects at step t. SODA starts with $M_1 = \emptyset$ and iterates below until termination.

   (a) For each predictor $j \notin$ in M_t, create a new candidate set $M_{t,j} = M_t \cup \{j\}$.

   (b) Find the predictor j with the lowest $\text{EBIC}_\gamma(M_{t,j})$. If $\text{EBIC}_\gamma(M_{t,j}) < \text{EBIC}_\gamma(M_t)$, continue with $(M_{t+1}) = M_{t,j}$, otherwise terminate with $\tilde{M}_f$ and go to stage 2.

2. Forward variable selection(interaction and main effect): Let $\{C_t\}$ denote selected set of predictors at step t, and let $S_t = \tilde{M}_f \cup C_t \cup C_t \times C_t$ denote the set of terms included by $\{C_t\}$. SODA starts with $C_1 = \emptyset$ and iterates operations below until termination.

   (a) For each $j \notin \{C_t\}$, set $C_{tj} = C_t \cup j$ and let $S_{t,j} = \tilde{M}_f \cup C_{tj} \cup C_{tj} \times C_{tj}$.

   (b) Find the predictor j with the lowest $\text{EBIC}_\gamma(S_{tj})$.
   If $\text{EBIC}_\gamma(S_{tj}) < \text{EBIC}_\gamma(S_t)$, continue with $C_{t+1} = C_{tj}$, otherwise terminate with $\tilde{C}_f$ and go to stage 3.

3. Backward elimination: Let $\{S_t\}$ denote selected set of individual terms at step t of the backward stage. SODA starts with $S_1 = (\tilde{M}_f \cup \tilde{C}_f \cup \tilde{C}_f \times \tilde{C}_f)$ and

iterates until termination below:

(a) For each main or interaction term j $\in S_t$, create a candidate set $S_{tj} = S_t/j$. SODA starts with $C_1 = \emptyset$ and iterates operations below until termination.

(b) Find the predictor j with the lowest $\text{EBIC}_\gamma(S_{tj})$.
If $\text{EBIC}_\gamma(S_{tj}) < \text{EBIC}_\gamma(S_t)$, remove j, otherwise terminate and retain $\tilde{S}$=S_t.

Suppose we have a total of four covariates before the variable selection process; that is consider $(X_1, \ldots, X_4)$. The preliminary stage starts from an empty set of covariates. It screens the EBICs of the models with the main effect of one covariate and finds a variable, for example, X_1 that results in the smallest EBIC. If the smallest EBIC is also smaller than that of the empty set, X_1 is retained at the preliminary stage. Otherwise, the preliminary stage is terminated, and the process proceeds to forward variable selection stage. If the preliminary stage is not terminated, it continues to screen the EBICs of the models with X_1 and another covariate from X_2, X_3, X_4. If X_1, X_3 results in the smallest EBIC and is less than the EBIC of the model with X_1 alone, X_3 is added at this stage. Otherwise, the process proceeds to forward variable selection stage. In summary, preliminary stage iteratively seeks a set of covariates that results in smaller EBIC.

Forward variable selection stage (stage 2) is akin to a forward variable selection step. This stage follows from the preliminary stage and starts with the set of covariates selected at the end of stage 1. It screens the EBICs of the models with the main effects of stage 1 and the main and quadratic effect of one covariate from $X_1, \ldots, X_4$. Duplicated terms are counted as one term. Suppose it finds adding the main and quadratic term of X_2 results in the lowest EBIC and it is lower than EBIC of the model with the main effects of stage 1 alone, the main and quadratic terms of X_2 are retained. Otherwise, stage 2

is terminated, and the process proceeds to backward selection or elimination stage(stage 3). If stage 2 is not terminated, it continues to screen the EBICs of the models with main effects of stage 1, the main and quadratic terms of X_2, and main, quadratic term of one covariate from X_1, X_3, X_4, plus its interaction term with X_2. If adding all the terms of one covariate results in the lowest EBIC and is lower than the EBIC from previous iteration of stage 2, all the terms of the covariates are retained. Otherwise, the process proceeds to stage 3. To summarize, forward stage iteratively seeks a set of covariates and their interaction that results in smaller EBIC.

Stage three is the backward elimination step. This stage starts from all the terms retained at stage 2. It screens the EBIC of the models with one term dropped from the final model of stage 2. If a model without the interaction term of X_2 achieves the lowest EBIC and it is lower than the EBIC of the final model of stage 2, the interaction term of X_2 is dropped. Otherwise, stage 3 is terminated and the final model is achieved. If stage 3 is not terminated, the algorithm continues to screen the EBIC of the models with one term dropped from the final model from previous iteration.

For continuous response, a slight modification of SODA, named sliced SODA, was proposed to perform variable selection. S-SODA involves sorting samples in ascending order of the response by partitioning them in equal slices before applying SODA. The paper by Li and Liu [8] demonstrated the efficiency and robustness of using S-SODA in variable selection for continuous response when the response function is nonlinear. The algorithm is summarized as:

1. Sort the samples in the ascending order of the response y_i and partition them into H equal slices. Each y_i is given a category label s_i.

2. Apply a SODA algorithm to the data $\{(s_i, \mathbf{x}_i)\}$. It outputs the main and

interaction effects, as well as the variables that

In practice, they proposed using a 10-fold cross-validation (CV) procedure for selecting γ from the set $\{0, 0.5, 1.0\}$. For simulation studies and real data analyses in their paper, they set $\gamma = 0.5$.

2.2 LASSO and Elastic Net

2.2.1 LASSO

LASSO is a regularization and variable selection algorithm which performs mostly better than Ridge regression whenever $p>n$. For continuous responses, the simple regression model assumed under LASSO is $y = \beta_0 + \sum_{j=1}^{p} \beta_j x_j + \epsilon$ where $\beta_0, \beta_1, \ldots, \beta_p$ are regression coefficients. Interaction terms and quadratic terms can be added into the regression. The LASSO estimates, $\hat{\beta}_\lambda^L$ minimizing the quantity

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{2.1}$$

where $\lambda \sum_{j=1}^{p} |\beta_j|$ is the penalty term.

It is important to note that LASSO is based on the L_1 penalty while ridge is based on the L_2 penalty on the coefficients. A graphic illustration is displayed in Figure 2.1. Here L_1 and L_2 penalty are based on L_1 NORM and L_2 NORM . For a two-dimensional $\boldsymbol{\beta}$,

$$L_1 \text{ NORM } : ||\boldsymbol{\beta}||_1 = |\beta_1| + |\beta_2|$$
$$L_2 \text{ NORM } : ||\boldsymbol{\beta}||_2 = \sqrt{\beta_1^2 + \beta_2^2}$$

With LASSO, the L_1 penalty tends to force some of the coefficient estimates to

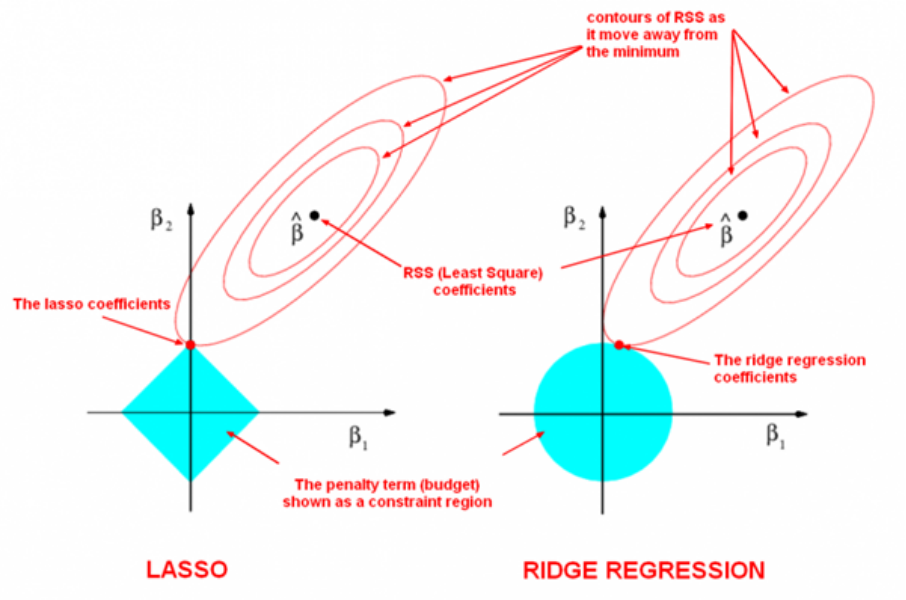

Figure 2.1: Graph for the Ridge Regression (right side) and LASSO (left side). The two areas are the constraint regions: the disk for the ridge regression, $\beta_1^2 + \beta_2^2 \leq t$, and the diamond for the LASSO, $|\beta_1| + |\beta_2| \leq t$. While the ellipses are the borders of the least squares error functions.

be exactly equal to zero when the tuning parameter λ is sufficiently large enough. Some noted features of LASSO include:

⇑ LASSO is advantageous in that it produces only a subset of the original variables.

⇓ LASSO will only select one variable if there are grouped which are highly correlated with each other.

For binary responses, the LASSO penalty can be applied in the logistic regression framework. For all simulations, I used LASSO methods for both continuous and binary responses. Moreover, quadratic terms and second order main effects were added to the linear/generalized linear model with first order main effects only.

2.2.2 Elastic Net

This algorithm overcomes the limitations of LASSO (selecting only one feature from a group of correlated features) and Ridge(interpretability problem since unimportant coefficients may be shrunken towards zero, but still in the model) by combining the use of L_1 and L_2 norms and hence increase the flexibility of selection and stabilize the selection of grouped variables. The Elastic Net estimates, $\hat{\beta}_\lambda^L$ minimizing the quantity

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\left[(1-\alpha)||\boldsymbol{\beta}||_2^2/2 + \alpha||\boldsymbol{\beta}||_1\right] \tag{2.2}$$

where $\lambda \geq 0$ is again a complexity parameter and $0 \leq \alpha \leq 1$ is a compromise between ridge ($\alpha = 0$) and LASSO ($\alpha = 1$).

LASSO and Elastic net variable selection can both be applied using the package 'glmnet'. The package allows a cross-validation procedure to determine the best tuning parameter λ. For Elastic net, I also added quadratic terms and second order main effects to the linear/generalized linear model.

2.3 Random Forest

Random Forest is a decision tree-based form of an algorithm which deals with problems related to regression, classification and various levels of complexities. It is based on the concept of building multiple trees whose combining effect later produces a single consensus. Since it does not require a functional relationship between the predictors and the response, it is a great candidate for model selection when considering high dimensional data.

2.3.1 Decision tree

Consider a learning set $L = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ where the input vector $X = (X_1, \ldots, X_p)$ and Y is either a class label for classification problems or a numerical response for regression ones. The CART (Classification and Regression Trees) method defined by Breiman et al. (1984) is a well-known way to design optimal single binary decision trees. A single tree represents recursive splitting of the covariate space where every node of interest corresponds to one region in the original space; two child nodes will occupy two different regions and if I put the two together, I get the same region as that of the parent node; in the end, every leaf node is assigned to a class. As an example, I show in Figure 2.2 a single fitted tree using data on Children who have had Corrective Spinal Surgery and 'rpart' package. The response variable is 'Kyphosis', which is a factor with two levels absent and present indicating if a kyphosis (a type of deformation) was present after the operation. Covariates used were 'Age' (age in months), 'Number' (the number of vertebrae involved), and 'Start' (the number of the first (topmost) vertebra operated on). The fitted tree shows that the first split uses variable 'Start' at value 8.5 and when Start' is less than 8.5, there are 8 kyphosis cases and 11 kyphosis absences. When Start' is greater than 8.5, the space is further split by variable 'Start' at value 14.5, and when Start' is greater than 14.5, there are 29 kyphosis cases and 0 kyphosis absences. The third and fourth split used the variable 'Age,' and the interpretations of the splits are similar to those of 'Start.' The construction of a single tree typically involves the selection of the splits, how to make splits and how to decide when to declare a node terminal and stop splitting.

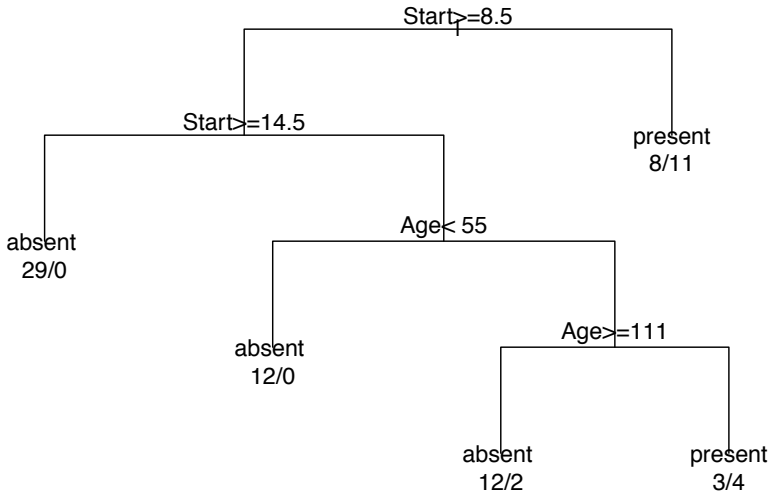

Figure 2.2: A single fitted tree for Kyphosis using data on Children who have had Corrective Spinal Surgery.

2.3.2 Random forest

Random forest (Breiman 2001) builds multiple decision trees and merges them to get a more accurate result. Let N be the number of observations and assume for now that the response variable is binary. The Random forest algorithm proceeds like the following:

1. Take a random sample of predictors without replacement. Typical setting set the number of predictors selected as $p/3$ for regression and $\sqrt{p}$ for classification.

2. Take a random sample without replacement of a fixed number of predictors.

This distinguishes random forests from bagging.

3. Construct a split by using predictors selected in Step 2.

4. Repeat Steps 2 and 3 for each subsequent split until the tree is as large as desired. Typically the splitting of a branch stops when the node has less than 5 observations for regression and 1 observation for classification.

5. Fit the out-of-bag data using the tree. Store the class assigned to each observation along with each observation's predictor values.

6. Repeat Steps 1-5 a large number of times. Typical setting is 500.

7. To obtain a classification, count the number of trees that it is classified in each category. Assign the observation to the category that has a majority vote.

For continuous step, the last step gives a fitted value by taking the average of the fitted values from each tree.

2.3.3 RF variable selection

The random forests variable selection procedures proposed by Genuer et al. (2010b) are based on prediction performance of RF focusing on the out-of-bag (OOB) error and the quantification of the variable importance. Out-of-bag data refers to data that are not included in the bootstrap sample that is used to construct the trees. The OOB error rate is defined by

$$errOOB = \begin{cases} \frac{1}{n}\text{Card}\{i \in \{1,\dots,n\} | y_i \neq \hat{y}_i\} & \text{in the classification framework} \\ \frac{1}{n}\sum_{\{i\in\{1,\dots,n\}}(y_i - \hat{y}_i)^2 & \text{in the regression framework} \end{cases}$$

where $\hat{y}_i$ is the aggregation of the predicted values by trees for which (x_i, y_i) belongs to the associated OOB sample. In other words, $errOOB$ is the mean prediction error on each training sample x_i, using only the trees that did not have x_i in their bootstrap sample.

Based on the definition of out-of-bag (OOB) error, the variable importance is outlined as the following. Define OOB_t as the out-of-bag data associated with tree t, which refers to sample data not included in the bootstrap sample used to construct tree t. Denote $errOOB_t$ as the error of a single tree t on the OOB_t sample. Let $\widetilde{OOB}_t^j$ be a perturbed sample by randomly permuting the values of X_j in OOB_t and $err\widetilde{OOB}_t^j$ be the associated error of tree t on the perturbed sample. Variable importance of X_j is then defined as

$$VI(X_j) = \frac{1}{ntree} \sum_t \left(err\widetilde{OOB}_t^j - err\widetilde{OOB}_t \right)$$

where the sum is over all trees t of the RF and 'ntree' denotes the number of trees of the RF. For the proposed variable selection methods, the authors propose running repetitions of forests and obtain a sample of variable importance measures.

Algorithm for variable selection (Genuer et al. 2015):

- Step 1: Preliminary elimination and ranking:
  - Rank the variables by sorting the VI (averaged over typically 50 RF runs) in descending order.
  - Eliminate the variables of small importance (let m be the number of remaining variables). The threshold is estimated as the minimum value given by a CART model where y are the sample standard deviation of the VI and X are the ranks of VI for the variables. Then only the variables with an averaged VI exceeding this threshold are retained. The

idea is based on the fact that variability of VI is larger for true variables compared to useless ones.

- Step 2. Variable selection:
  - For interpretation:Based on the collection of the variables retained by step 1 and the ranking of the those variables, construct the nested collection of RF models involving the k first variables, for $k = 1$ to m and select the set of variables involved in the model leading to the smallest OOB error. This leads to consider m' variables.
  - For prediction: starting with the ordered variables retained for interpretation, construct an ascending sequence of RF models, by invoking and testing the variables in a stepwise way. The test is that a variable is added only if the error decrease is larger than a threshold and the threshold is set to the mean of the absolute values of the first order differentiated OOB errors between the model with m' variables and the one with m variables:

$$\frac{1}{m-m'}\sum_{j=m'}^{m-1}|errOOB(j+1)-errOOB(j)|$$

where $errOOB(j)$ is the OOB error of the RF built using the j most important variables. The idea is that the OOB error decrease must be significantly greater than the average variation obtained by adding noisy variables.

The package 'VSURF' implements the variable selection methods for interpretation or prediction. The 'mtry' parameter is the number of variables randomly sampled as candidates at each split. Here it is set to p if $p < n$ and otherwise it is set to $p/3$. The default 'ntree' is 2000 in VSURF.

Chapter 3

Simulations

The motivating dataset examines the impact of uranium exposures on birth outcomes and early child development on Navajo Nation. The UNM team solicited community, Navajo and Federal agency input to inform a study design respectful of the Navajo culture and inclusive of appropriate measures to address community concerns. This ongoing study began recruitment in 2013 and is assessing exposures through quantification of 36 metals in blood and urine specimens. The NBCS enrolled pregnant women whose children were followed for growth and neuro-developmental outcomes for 12 months of age to determine the effect of legacy waste on both birth and overall health outcomes. For our simulations, I simulated responses under various scenarios with covariates bootstrapped from real data and then compared false positive percentages and false negative percentages of SODA/S-SODA, LASSO, Elastic Net, Random forest methods. Note that quadratic terms were included in models built to make a fair comparison between LASSO and SODA

Bootstrap covariates: From the preliminary data, if metals levels were below detection limits, the levels were replaced by the limit of detection (LOD) divided by a square root of 2 [10]. The levels of metals in urine were further divided by

the creatinine level to adjust for kidney function. Demographic variables recorded are mother's age (continuous), gestation age (continuous), number of people in the household (discrete), weight at birth (continuous), household income (categorical), mother's education (categorical), mother's marital status (categorical), and Mother's employment (categorical). Variables with more than 70% LOD or more than 40% missing were excluded. Observations with missing covariates were excluded. Covariates were created according to the following: (1)"standardization": all continuous variables were standardized about their means and standard deviations; (2) "dummy variables": dummy variables were created for all the categorical variables with the baseline excluded from the covariate matrix, since none of the methods need to create an intercept term; (3)"bootstrap samples": bootstrapped samples were created by randomly sampling a set of covariates from the data without replacement. There are 47 covariates in total with 33 that are continuous variables while 14 are binary variables.

Scenario 1: in this scenario the continuous responses were simulated from the model

$$\begin{aligned} y &= f(x) + \epsilon \\ f(x) &= 1 + x_1 - x_2^2 + x_3 + x_1 x_2 + z + z x_1 \\ \epsilon &\sim N(0, \sigma_\epsilon^2) \end{aligned}$$

where x_1, x_2, x_{i3} are three continuous covariates randomly selected from the set of continuous covariates, z_i is a discrete covariate randomly selected from the set of discrete variables; σ_ϵ^2 is determined so that the signal-to-noise ratio $\text{var}(f(x))/\sigma_\epsilon^2$ is 0.5, 2, or 4.

Scenario 2: in this scenario the continuous responses were simulated as shown beow;

$$\begin{aligned} y &= f(x) + \epsilon, \\ \log\left(\frac{f(x)}{1-f(x)}\right) &= 1 + x_1 - x_2^2 + x_3 + x_1x_2 + z + zx_1, \\ \epsilon &\sim N(0, \sigma_\epsilon^2). \end{aligned}$$

Covariates were selected in a similar way to Scenario 1 and σ_ϵ^2 is also determined so that the signal-to-noise ratio $\text{var}(f(x))/\sigma_\epsilon^2$ is 0.5, 2, or 4.

Scenario 3: in this scenario, binary responses were simulated from a logistic regression model. Let $P(y=1) = \log\left(\frac{f(x)}{1-f(x)}\right)$. Two sets of coefficients were considered for $f(x)$

$$\begin{aligned} \text{Set 1: } f(x) &= 1 + x_1 - x_2^2 + x_3 + x_1x_2 + z + zx_1, \\ \text{Set 2: } f(x) &= 1 + 0.5x_1 - 0.5x_2^2 + 0.5x_3 + 0.5x_1x_2 + 0.5z + 0.5zx_1, \end{aligned}$$

Hyper-parameter settings: for scenarios 1 and 2, S-SODA, LASSO, Elastic net, RF interpretation, and RF prediction methods were used to perform variable selection. For scenarios 3, SODA, LASSO, Elastic net, RF interpretation, and RF prediction methods were used to perform variable selection. For S-SODA, I consider six settings (1) $H = 5, \gamma = 0$, (2) $H = 5, \gamma = 0.5$, (3) $H = 5, \gamma = 1$, (4) $H = 2, \gamma = 0$, (5) $H = 2, \gamma = 0.5$, and (6) $H = 2, \gamma = 1$. For SODA, I consider three settings of

γ: $\{0, 0.5, 1\}$. For LASSO and Elastic net, a cross-validation procedure was used to determine the tuning parameter λ. For RF variable selection methods, 'mtry' is set as 12.

A seed of 53 was used for repeatability and reproducibility. Our measure of performance is based on the percent of false positives (the number of wrongly identified variables divided by the number of zero effect covariates) and percent of false negatives (the number of missed covariates divided by the number of nonzero effect covariates). Table 3.1 and Table 3.4 are for scenario 1; Table 3.2 and Table 3.5 are for scenario 2; Table 3.3 and Table 3.3 are for scenario 3.

Table 3.1: Percentages of false positives and false negatives selected by S-SODA, LASSO, Elastic Net, Random forest (interpretation), and Random forest (prediction). The response is continuous, and model is linear; S-SODA being compared here sets $H = 5$ and $\gamma = 0.5$.

		S-SODA		**Elastic net**		**LASSO**		**RF (interp.)**		**RF (pred.)**	
n	signal-to noise-ratio	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN
50	0.5	0.003	0.994	0.239	0.584	0.163	0.659	0.080	0.691	0.056	0.734
	2.0	0.002	0.956	0.375	0.322	0.279	0.395	0.062	0.582	0.042	0.621
	4.0	0.004	0.920	0.402	0.316	0.272	0.382	0.050	0.567	0.030	0.591
100	0.5	0.000	0.984	0.302	0.425	0.238	0.472	0.071	0.625	0.050	0.661
	2.0	0.000	0.875	0.476	0.215	0.389	0.251	0.043	0.500	0.026	0.511
	4.0	0.000	0.812	0.532	0.157	0.406	0.200	0.037	0.520	0.020	0.541
200	0.5	0.000	0.940	0.454	0.249	0.335	0.322	0.046	0.556	0.032	0.577
	2.0	0.000	0.742	0.577	0.112	0.478	0.134	0.026	0.472	0.015	0.500
	4.0	0.000	0.587	0.658	0.076	0.514	0.099	0.024	0.481	0.012	0.492

Based on the results in Table 3.1, S-SODA have the lowest FP percentage (the number of falsely selected variables divided by the total number of zero-effect variables) across all scenarios and the highest FN percentage (the number of missed selected variables divided by the total number of true variables). LASSO and Elastic Net have higher FP percentages than other methods but also lower FN percentages. Comparing Elastic Net with LASSO, LASSO has higher FP and lower FN. Random Forest methods have low FP rates and have high FN, but those FN percentages

are better than SODA. As expected, Random forest using interpretation step has higher FP percentage and lower FN percentage than Random forest using prediction step. As sample size increases or signal-to-noise ratio increases, the FN percentages decrease for all methods, FP percentages increase for LASSO and Elastic Net and decrease for S-SODA and Random forest methods. Elastic net performs better than LASSO regarding FP percentages since Elastic net handles correlated variables better than LASSO. Note that the number of true variables in the model is four and the number of zero-effect variables is 43. So LASSO and Elastic net include on average 9 to 26 zero-effect variables in the model.

Table 3.2: Percentages of false positives and false negatives selected by S-SODA, LASSO, Elastic Net, Random forest (interpretation), and Random forest (prediction). The response is continuous, and model is nonlinear; S-SODA being compared here sets $H = 5$ and $\gamma = 0.5$.

		S-SODA		**Elastic net**		**LASSO**		**RF (interp.)**		**RF (pred.)**	
n	signal-to noise-ratio	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN
50	0.5	0.000	1.000	0.432	0.500	0.163	0.750	0.079	0.650	0.070	0.650
	2.0	0.000	1.000	0.288	0.300	0.214	0.350	0.032	0.500	0.023	0.500
	4.0	0.000	0.950	0.316	0.550	0.349	0.600	0.037	0.400	0.023	0.450
100	0.5	0.000	0.950	0.288	0.450	0.302	0.350	0.051	0.500	0.042	0.500
	2.0	0.000	0.650	0.460	0.100	0.465	0.100	0.028	0.350	0.023	0.350
	4.0	0.000	0.600	0.749	0.500	0.530	0.250	0.023	0.400	0.023	0.400
200	0.5	0.000	0.800	0.293	0.250	0.367	0.300	0.023	0.400	0.018	0.500
	2.0	0.000	0.350	0.735	0.050	0.576	0.100	0.014	0.300	0.014	0.300
	4.0	0.000	0.250	0.837	0.050	0.725	0.050	0.005	0.300	0.000	0.350

Based on results in Table 3.2, I observe the same consistent observation regarding S-SODA across most if not all scenarios. Once more, Elastic Net and LASSO have the highest FP percentages and lowest FN percentages. Except for three scenarios, Elastic Net has higher FP percentages than and lower FN percentages when compared to LASSO. Random Forest methods have low FP percentages, and high FN percentages and FN percentages are better than S-SODA. As expected, Random forest using interpretation step has slightly higher FP percentage than Random

forest using prediction step whereas FN percentages are comparable. FP percentages increase for LASSO and Elastic Net and appear to decrease for Random forest methods. Elastic net performs better than LASSO regarding FP percentages since Elastic net handles correlated variables better than LASSO.

Table 3.3: Percentages of false positives and false negatives selected by SODA with a hyper-parameter of $\gamma = 0.5$, Elastic net, LASSO, Random forest (interpretation) and Random forest (prediction). The response here is categorical, and model is logistic; SODA being compared sets $\gamma = 0.5$.

		SODA		**Elastic Net**		**LASSO**		**RF (interp.)**		**RF (pred.)**	
n	coefficient	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN
50	1	0.040	0.819	0.198	0.706	0.119	0.805	0.064	0.782	0.042	0.834
	2	0.013	0.255	0.077	0.215	0.047	0.254	0.022	0.236	0.014	0.257
100	1	0.022	0.746	0.321	0.545	0.230	0.605	0.061	0.677	0.036	0.735
	2	0.021	0.571	0.457	0.327	0.303	0.441	0.60	0.552	0.031	0.646
200	1	0.014	0.591	0.465	0.311	0.350	0.376	0.043	0.626	0.024	0.695
	2	0.011	0.412	0.643	0.141	0.478	0.212	0.038	0.501	0.019	0.561

Observing results from Table 3.3, I notice once more SODA achieves lowest FP percentage compared to other methods while FN percentages are higher except for random forest predict method. Elastic net and LASSO show the highest values for FP percentages while their FN percentages are comparatively lower on average than other methods. Random Forest methods have high FP percentages accompanied by high FN percentages compared to other methods in this scenario. As expected, Random forest using interpretation step has higher FP percentages than Random forest using prediction step.

Based on the results in Table 3.4, S-SODA with parameters $H = 5$ and $\gamma = 1$ have the lowest FP percentages across all scenarios and the highest FN percentages. FP percentages for S-SODA parameters with $H = 2$ are comparatively higher than those with $H = 5$. Similarly, FN percentages for $H = 2$ are relatively lower than those with $H = 5$. This suggests S-SODA with few slices might be preferable. As

Table 3.4: Percentages of false positives and false negatives selected by S-SODA at different settings of its hyper-parameters of H and γ. Response is continuous, and model is linear; S-SODA being compared here are values from the set $H = (2, 5), \gamma = (0.0, 0.5, 1.0)$

		(H=2,γ=0.0)		**(H=2,γ=0.5)**		**(H=2,γ=1.0)**		**(H=5,γ=0.0)**		**(H=5,γ=0.5)**		**(H=5,γ=1.0)**	
n	signal-to noise-ratio	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN
50	0.5	0.112	0.781	0.038	0.871	0.013	0.947	0.054	0.861	0.003	0.994	0.000	1.000
	2.0	0.105	0.675	0.038	0.784	0.012	0.874	0.012	0.750	0.002	0.956	0.000	0.994
	4.0	0.099	0.642	0.038	0.704	0.014	0.810	0.055	0.666	0.004	0.920	0.000	0.986
100	0.5	0.078	0.741	0.023	0.824	0.005	0.920	0.015	0.876	0.000	0.984	0.000	0.997
	2.0	0.076	0.551	0.022	0.665	0.007	0.754	0.013	0.634	0.000	0.875	0.000	0.965
	4.0	0.070	0.507	0.018	0.589	0.007	0.686	0.014	0.556	0.000	0.812	0.000	0.915
200	0.5	0.044	0.361	0.011	0.415	0.004	0.476	0.003	0.427	0.000	0.587	0.000	0.725
	2.0	0.045	0.414	0.012	0.486	0.004	0.584	0.004	0.520	0.000	0.742	0.000	0.881
	4.0	0.044	0.361	0.011	0.415	0.004	0.476	0.003	0.427	0.000	0.587	0.000	0.725

γ increases, FP percentages decrease while FN percentages increase. This means S-SODA can detect more variables with its modified form of BIC as the hyper-parameter γ increases. Overall, it appears as sample size and signal to noise ratio increase, FP percentages decrease while FN percentages increase.

Table 3.5: Percentages of false positives and false negatives selected by S-SODA at different settings of its hyper-parameters of H and γ. Response is continuous, and model is non linear; S-SODA being compared here are values from the set $H = (2, 5), \gamma = (0.0, 0.5, 1.0)$

		(H=2,γ=0.0)		**(H=2,γ=0.5)**		**(H=2,γ=1.0)**		**(H=5,γ=0.0)**		**(H=5,γ=0.5)**		**(H=5,γ=1.0)**	
n	signal-to noise-ratio	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN
50	0.5	0.112	0.750	0.046	0.850	0.023	0.850	0.051	0.750	0.001	1.000	0.000	1.000
	2.0	0.093	0.350	0.014	0.450	0.009	0.600	0.018	0.450	0.000	0.650	0.000	0.950
	4.0	0.070	0.450	0.042	0.500	0.014	0.700	0.028	0.450	0.000	0.950	0.000	1.000
100	0.5	0.070	0.600	0.019	0.650	0.005	0.750	0.000	0.800	0.000	0.950	0.000	1.000
	2.0	0.093	0.350	0.014	0.450	0.009	0.600	0.019	0.450	0.000	0.650	0.000	0.950
	4.0	0.074	0.400	0.028	0.400	0.000	0.400	0.023	0.400	0.000	0.600	0.000	0.800
200	0.5	0.060	0.450	0.009	0.550	0.000	0.600	0.000	0.700	0.000	0.800	0.000	0.950
	2.0	0.032	0.200	0.014	0.200	0.000	0.300	0.000	0.300	0.000	0.350	0.000	0.450
	4.0	0.032	0.250	0.014	0.300	0.009	0.300	0.005	0.250	0.000	0.250	0.000	0.300

Based on the results in Table 3.5 As γ increases, FP percentages decrease while FN percentages increase. This S-SODA can detect more variables with its modified

form of BIC as γ changes. Overall, it appears as sample size and signal to noise ratio increase, FP percentages decrease while FN percentages increase.

Table 3.6: Percentages of false positives and false negatives selected by SODA at different settings of its hyper-parameter γ. The response is categorical, and model is logistic; SODA being compared here scearios with $\gamma = (0.0, 0.5, 1.0)$.

		($\gamma = 0.0$)		($\gamma = 0.5$)		($\gamma = 1.0$)	
n	coefficient	FP	FN	FP	FN	FP	FN
50	1	0.101	0.741	0.040	0.818	0.011	0.902
	2	0.033	0.226	0.013	0.255	0.004	0.278
100	1	0.076	0.630	0.022	0.746	0.007	0.829
	2	0.076	0.476	0.021	0.571	0.005	0.689
200	1	0.046	0.490	0.014	0.591	0.006	0.700
	2	0.040	0.331	0.011	0.412	0.004	0.466

Based on results from Table 3.6, S-SODA with parameter $\gamma = 1$ has the lowest FP percentage across all scenarios with the highest FN percentage for the categorical scenario. As γ increases, FP percentages decrease while FN percentages increase. Similar to my observations in 3.4 and 3.5 suggests SODA can detect more variables with its modified form of BIC as γ decreases. Comparing SODA with setting $\gamma = 0$ with other methods for binary responses, the performance of SODA lies somewhere in between that of Elastic net and Random Forest. When the response is continuous, a similar comparison is observed for S-SODA with setting $H = 2, \gamma = 0$ and other methods.

Chapter 4

Conclusion

Simulations reveal SODA/S-SODA is a technique which often misses or omits important variables while it wrongly identifies few. This observation improves slightly as sample size and signal to noise ratio increase. Setting $\gamma = 0$ increases the percentages of FPs slightly but decreases the percentages of FNs by large. For continuous responses, setting $H = 2$ will also increase the percentages of FPs slightly but decreases the percentages of FNs significantly. In most cases and scenarios, SODA/S-SODA, regardless of its setting of H and γ, performs better than LASSO, Elastic Net, and Random forest methods in terms of percentages of FPs. However, SODA/S-SODA performs much worse than other methods with large γ and/or large H in terms of percentages of FNs. With the low value of γ and/or low H, the percentages of FNs are comparable to Random forest methods but are still much higher than LASSO and Elastic Net. As I expected, Random Forest prediction step improves the interpretation step in all scenarios with better FP and FN percentages across all scenarios. LASSO and Elastic Net have high FP percentages and the percentages increase when sample size and/or signal to noise ratio increases. This is mostly due to the interaction terms of continuous and discrete variables which are often captured wrongly by LASSO and Elastic Net. Finally, my observations of SODA/S-SODA are different

from those in Li and Liu (2018) largely due to low signal-to-noise ratio in my settings and very high signal-to-noise ratio in their settings.

Bibliography

[1] Leo Breiman. *Classification and regression trees*. Routledge, 2017.

[2] Jiahua Chen and Zehua Chen. Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771, 2008.

[3] Jianqing Fan, Richard Samworth, and Yichao Wu. Ultrahigh dimensional feature selection: beyond the linear model. *Journal of machine learning research*, 10(Sep):2013–2038, 2009.

[4] Robin Genuer, Jean-Michel Poggi, and Christine Tuleau-Malot. Variable selection using random forests. *Pattern Recognition Letters*, 31(14):2225–2236, 2010.

[5] Robin Genuer, Jean-Michel Poggi, and Christine Tuleau-Malot. Vsurf: an r package for variable selection using random forests. *The R Journal*, 7(2):19–33, 2015.

[6] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

[7] Bo Jiang, Jun S Liu, et al. Variable selection for general index models via sliced inverse regression. *The Annals of Statistics*, 42(5):1751–1786, 2014.

[8] Yang Li and Jun S Liu. Robust variable and interaction selection for logistic regression and multiple index models. *arXiv preprint arXiv:1611.08649*, 2016.

[9] Joseph O Ogutu, Torben Schulz-Streeck, and Hans-Peter Piepho. Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. In *BMC proceedings*, volume 6, page S10. BioMed Central, 2012.

[10] Paul A Succop, Scott Clark, Mei Chen, and Warren Galke. Imputation of data values that are less than a detection limit. *Journal of occupational and environmental hygiene*, 1(7):436–441, 2004.

[11] Zhichao Sun, Yebin Tao, Shi Li, Kelly K Ferguson, John D Meeker, Sung Kyun Park, Stuart A Batterman, and Bhramar Mukherjee. Statistical strategies for constructing health risk models with multiple pollutants and their interactions: possible choices and comparisons. *Environmental Health*, 12(1):85, 2013.

[12] Kyla W Taylor, Bonnie R Joubert, Joe M Braun, Caroline Dilworth, Chris Gennings, Russ Hauser, Jerry J Heindel, Cynthia V Rider, Thomas F Webster, and Danielle J Carlin. Statistical approaches for assessing health effects of environmental chemical mixtures in epidemiology: lessons from an innovative workshop. *Environmental health perspectives*, 124(12):A227, 2016.

[13] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.