Variable Selection in Presence of Strong Collinearity with Application to Environmental Mixtures

BY

Jiyeong Jang
B.H.S., Korea University, 2012
B.Ec., Korea University, 2012
M.S., Seoul National University, 2014

THESIS

Submitted as partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Public Health Sciences
in the Graduate College of the
University of Illinois at Chicago, 2020

Chicago, Illinois

Defense Committee:
Sanjib Basu, Chair and Advisor
Hua Yun Chen, Biostatistics
Mary Turyk, Epidemiology
Runa Bhaumik, Computer Science
Saria Awadalla, Biostatistics

ACKNOWLEDGMENTS

I have been helped by many people to complete this work. First of all, I would like to express deep gratitude to my advisor, Dr. Sanjib Basu for his kind supports for me. Thanks to his extensive knowledge in statistics and care for his student, I have been able to have great opportunities to broaden my perspective in studying statistics and in conducting statistical research. Besides my advisor, I appreciate my committee members Drs. Hua Yun Chen, Mary Turyk, Runa Bhaumik, and Saria Awadalla for providing valuable comments and suggestions on my work. I am especially grateful for Drs. Turyk and Chen, the principal investigators of the NIEHS (National Institute of Environmental Health Sciences) funded project "Innovative methodologic advances for mixtures research in epidemiology" for offering helpful resources that have contributed to the completion of this work (grant number: R01ES028790).

Without the efforts and sacrifices of my parents, Mr. Hyunjun Jang and Ms. Yeonhui Park, I would not be the person I am today. I would like to express my special thanks to them for their continuous love and support. My husband Sanghoon deserves my genuine thanks for his enormous help and care for me to finish this race. I appreciate him having been my trustworthy family and friend standing by my side for about one and a half decades. My particular thanks goes as well to my beloved brothers Changhoon and Myeonghun who have consistently rooted for me. I also give my heartfelt thanks to my grandparents, relatives, in-laws, and many of my friends and acquaintances for all they have done for me.

Lastly, I give my sincere thanks and glory to the Lord, who has been my refuge and hope.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

CHAPTER **PAGE**

TABLE OF CONTENTS (Continued)

CHAPTER **PAGE**

LIST OF TABLES

LIST OF TABLES (Continued)

LIST OF TABLES (Continued)

LIST OF FIGURES

LIST OF FIGURES (Continued)

LIST OF FIGURES (Continued)

LIST OF ABBREVIATIONS

AHR	Aryl Hydrocarbon Receptor
AIC	Akaike Information Criterion
AICc	Akaike Information Corrected Criterion
BIC	Bayesian Information Criterion
BKMR	Bayesian Kernel Machine Regression
BMI	Body Mass Index
COLRNS	Correlation Learning for Variable Selection
FDR	False Discovery Rate
FNR	False Negative Rate
GLFCS	Great Lakes Fish Consumer Study
GLM	Generalized Linear Model
HORSES	Hexagonal Operator for Regression with Shrinkage and Equality Selection
ISIS	Iterative Sure Independence Screening
KMR	Kernel Machine Regression
LASSO	Least Absolute Shrinkage and Selection Operator
LOD	Limit of Detection
LTL	Leukocyte Telomere Length

LIST OF ABBREVIATIONS (Continued)

MCP	Minimax Concave Penalty
MMLE	Maximum Marginal Likelihood Estimator
MSE	Mean Squared Error
NHANES	National Health and Nutrition Examination Survey
OLS	Ordinary Least Squares
OSCAR	Octagonal Shrinkage and Clustering Algorithm for Regression
PACS	Pairwise Absolute Clustering and Sparsity
PCBs	Polychlorinated Biphenyls
PE	Prediction Error
POPs	Persistent Organic Pollutants
ROC	Receiver Operating Characteristic
RMSE	Root Mean Squared Error
SCAD	Smoothly Clipped Absolute Deviation
SRIG	Sparse Regression Incorporating Graphical Structure Among Predictors
SIS	Sure Independence Screening
TEQ	Toxic Equivalent
VIF	Variance Inflation Factor

SUMMARY

Variable selection has become an essential element of high dimensional statistical modeling to yield parsimonious models while keeping high prediction accuracy. High dimensionality often induces collinearity problems. For instance, studies of environmental mixtures include a large number of pollutants which are strongly inter-correlated. Regularized variable selection methods such as LASSO are popular for statistical variable selection, however these methods often not perform well in presence of strong collinearity in terms of selection and prediction. To address these challenges a novel method, namely COrrelation LeaRNing for variable Selection (COLRNS), is developed that is based on iterative correlation learning for cluster detection and variable selection. The COLRNS is further extended to COLRNS Generalized Linear Model (COLRNS-GLM) to be applicable in a generalized linear regression setting. The performance of the methods is evaluated through an extensive set of simulations and real-world applications to environmental mixtures data. The results show that the methods effectively identify a set of influential predictors, improve prediction accuracy, and reduce error in parameter estimation in most simulation scenarios and data applications under strong collinearity in high dimensional data.

CHAPTER 1

INTRODUCTION

1.1 Challenges of High Dimensional Data Analysis

The rapid advances of technology have allowed us to produce and collect massive amount of data with relatively low cost in a short time. High dimensional data are characterized by high dimensionality of variables in the data set. The number of variables can often far exceed the number of observations. High dimensional data are frequently been collected in a variety of areas such as health science, biology, geology, economics, and finance. For instance, many precision medicine studies seek risk factors in diverse types of high dimensional data such as genomic, clinical, and protein data for complex diseases [Fan and Li, 2006, JingYuan et al., 2015, Pan et al., 2019].

Analysis of high dimensional data poses many statistical challenges. The classical ordinary least squares (OLS) estimates that are used for linear regression are not unique, hence no longer applicable because of the lack of degree of freedom when the number of variables is higher than the number of observations in the data set [Wang and Leng, 2016]. In addition, two primary goals of data analysis, prediction and interpretation, are not often satisfied with the OLS estimates in high dimensional data analysis [Tibshirani, 1996]. For prediction, the OLS estimate often has low bias but large variance which causes hardships in accurate prediction of the future observation [Hastie et al., 2015]. It also makes it difficult to interpret and gain

insight into the relationships between the predictors and response because of a large number of predictors in the regression model. Furthermore, the high dimensionality often induces collinearity problem [Zou and Zhang, 2009]. The OLS estimate would perform poorly when collinearity is high among predictors. High collinearity may degrade accurate estimation of regression coefficients by inflating the standard errors of the coefficient estimators. It induces larger confidence intervals in statistical inference and also deflates t-test statistics causing false nonsignificant p-values in hypothesis. Moreover, it reduces accuracy of prediction for future observations which degrades generalization ability of the model [Hoerl and Kennard, 1970,Hoerl and Kennard, 1988, Shen and Gao, 2008]. Hence, high dimensional data analysis calls for new statistical methodologies and theories.

1.2 Challenges of Statistical Modeling of Environmental Mixtures

Environmental mixtures include a large number of environmental pollutants which potentially interact and affect health as mixture components. There are growing efforts to examine risks from these pollutant mixtures as people are simultaneously exposed to multitude of environmental contaminants. Mixtures of concern include air pollution [Kioumourtzoglou et al., 2013, Billionnet et al., 2012], mixtures of toxic waste [Hu et al., 2007], and mixtures of persistent organic chemicals [Gennings et al., 2010]. Studying chemical mixtures requires one to identify important individual components of the mixture that are responsible for the health effects of the mixture [Bobb et al., 2015]. Not only identification of critical pollutants, but precise estimation of their health effects is also important to examine their harmful effects. In

addition, a parsimonious well-structured model can be instrumental for health risk prediction or disease classification.

However, understanding health risks from environmental mixtures presents statistical challenges. As seen in the correlation matrix of environmental exposures that we use in our data application (Figure 1), environmental mixtures include a large number of pollutants. Each pollutant may have weak individual effect that contributes to the overall heath effect of the mixture. Further, the pollutant measurements are often extremely correlated at levels that are not generally seen in other areas of science [Bobb et al., 2015]. This is illustrated in Figure 1.

These challenges of high dimensional data analysis may lead to a poorly fitted regression model [Bobb et al., 2018]. Moreover, in presence of high collinearity, popular variable selection methods such as LASSO may not perform well when we conduct variable selection to identify a subset of the mixture components that is responsible for health effects. Hastie et al. [2015] show that LASSO estimates exhibit erratic behavior not reflecting the importance of the individual variables when there are groups of highly correlated variables. Further, when we have a high dimensional vector of exposures with weak signals, the sparsity principle which is important in high dimensional data analysis may fail. Under the principle, regression parameters are frequently assumed to be sparse with only a small number of predictors contributing to the response [Fan and Lv, 2010]. With sparsity, variable selection can enhance accurate estimation of parameters by effectively identifying influential predictors and can also improve model interpretability by yielding a parsimonious model [Fan and Lv, 2010], however, the principle does not hold in environmental mixture studies.

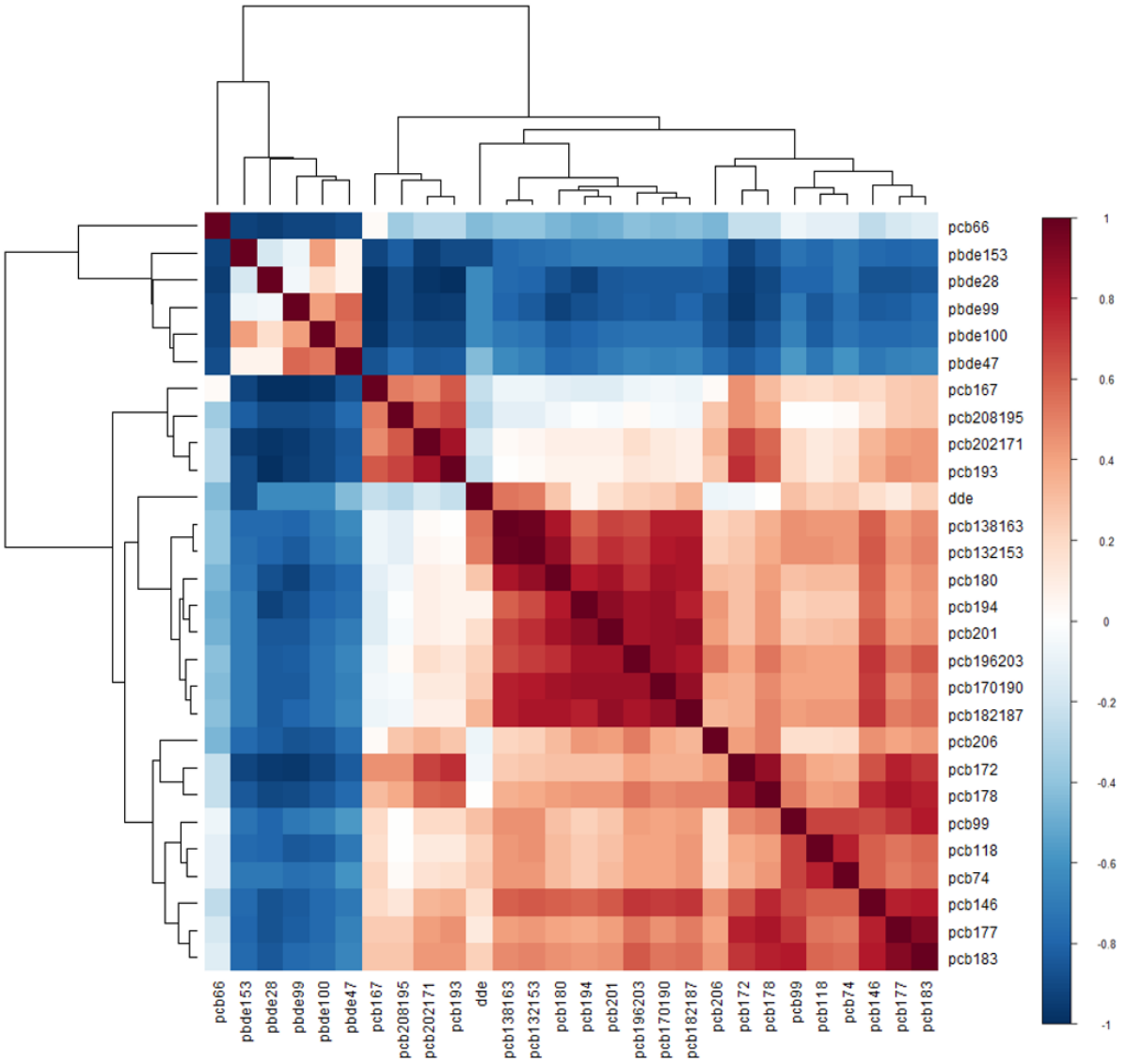


Figure 1: Pearsons's correlation matrix of log-transformed exposures

1.3 Proposed Work

We develop a COrrelation LeaRNing for variable Selection (COLRNS) method to address the problems of high dimensional variable selection when there is strong collinearity among variables. The method is based on iterative correlation learning for cluster detection and selection with the objectives of improved variable selection, accurate parameter estimation, and precise prediction of future outcomes. The COLRNS and COLRNS-GLM are suited to be applied to outcomes measured in continuous and binary scales in the setting of linear and generalized linear models.

This study is organized as follows. In Chapter 2, we introduce various existing methods for variable selections and their pros and cons. In Chapter 3, we describe the main procedures and rationales of COLRNS. An extensive set of simulation studies is conducted to evaluate the performance of the method in three areas: prediction, variable selection, and parameter estimation. We apply this method to real-world environmental mixtures data and perform a data-driven nested cross-validation study to examine prediction performance of the method based on real-world data. In Chapter 4, we introduce COLRNS-GLM, a generalization of COLRNS in the setting of generalized linear model, that is motivated by the binary outcome of diabetes incidence in the Great Lakes fish consumer study. A nested cross-validation study that is based on the data is performed to explore the prediction performance of the method under dichotomous responses. Concluding remarks and directions for future research are discussed in Chapter 5. The Appendix includes supplementary matreials such as ROC (receiver operating characteristic) curves.

CHAPTER 2

VARIABLE SELECTION POSSIBLE SOLUTIONS

2.1 Introduction

Dimension reduction and feature selection play fundamental roles in knowledge discovery from massive data as an effective way in dealing with high dimensionality [Fan and Li, 2006, JingYuan et al., 2015]. In reduced dimension, more accurate estimation can be achieved by incorporating some well-developed lower dimensional methods [Fan and Lv, 2008]. In this chapter, we introduce four categories of variable selection methods: criterion based, penalty-based, screening, and Bayesian methods. We discuss properties, advantages, and disadvantages of these widely applied feature selection methods. In Section 2.3 for penalty-based methods, we introduce the popular LASSO method, and various extensions of the method. This chapter also includes the penalized selection methods that are proposed to deal with multiple groups of correlated predictors. These methods are compared with the method that we propose in Section 3.5. In Section 2.4 for screening methods, we introduce sure independence screening [Fan and Lv, 2008]. Iterative sure independence screening is also introduced as an extension of sure independence screening. In section 2.5, Bayesian kernel machine regression is introduced as a Bayesian variable selection method.

2.2 Criterion Based Variable Selection Methods

Criterion based variable selection methods are classical ways of performing model selection in statistical models. One of the most well-known and widely used method is stepwise regression [Breaux, 1967]. It performs variable selection by sequentially adding the predictors or eliminating them from the model one at a time based on certain criterion. Stepwise regression can be broadly categorized into forward selection, backward elimination, and stepwise method. The forward selection adds significant predictors, the backward elimination deletes insignificant features, and the stepwise method is similar to the forward selection, but it also considers to remove insignificant predictors at each step as in the backward elimination method [Chong and Jun, 2005]. These methods can be implemented with various criteria such as p-value, adjusted R^2, AIC (Akaike Information Criterion) [Akaike, 1973], AICc (Akaike Information Corrected Criterion) [Hurvich and Tsai, 1989], BIC (Bayesian Information Criterion) [Schwarz, 1978], Mallow's C_p [Mallows, 1973], or prediction error, etc. The idea of stepwise regression is simple and it is easy to implement, however, it faces many issues and criticisms. Issues related to biased estimation and inconsistent selection of stepwise regression have been discussed in many literatures [Steyerberg et al., 1999, Whittingham and Stephens, 2006, Flom and Cassell, 2007].

2.3 Penalized Variable Selection Methods

Sparsity, which assumes that only a few predictors importantly contribute to the response [JingYuan et al., 2015], is frequently assumed and used in high dimensional data analysis . Following this general principle, many penalized variable selection methods have been introduced to estimate the parameters and at the same time to conduct variable selection by penalizing

loss functions through sparsity inducing penalties [Wang and Leng, 2016]. The penalization shrinks the values of the regression coefficients, and may set some of them close to zero. This introduces some bias but decreases the variance of the predicted outcomes, and thus may improve the overall performance of prediction with respect to prediction accuracy measured by the mean-squared error [Hastie et al., 2015]. Moreover, since the penalized variable selection methods estimate a sparse model by identifying a smaller subset of predictors that exhibit strong effects on the response, it enhances interpretability of the model with regard to the relationship between features and response [Hastie et al., 2015, Pan et al., 2019].

One of the most popular penalized variable selection method is the *Least Absolute Shrinkage and Selection Operator (LASSO)* [Tibshirani, 1996] which is inspired from the non-negative garrote method [Breiman, 1995]. The LASSO uses the l_1 penalty for the penalization of the regression coefficients. Given that we have n samples $\{(x_i, y_i)\}_{i=1}^n$ in the regression setting, where each $x_i = (x_{i1}, \ldots, x_{ip})$ is a p-dimensional vector of features or predictors, let $X = (x_1, x_2, \ldots, x_n)^T$ be an $n \times p$ random design matrix of predictors, and let $y_i \in \mathbb{R}$ be the response variable. The LASSO solution $\hat{\beta}$ to the optimization problem is given by

$$\begin{aligned} \hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \Big\{ \sum_{i=1}^{n} (y_i - x_{ij}\beta_j)^2 \Big\} \\ \text{subject to} \sum_{j=1}^{p} |\beta_j| \leq t. \end{aligned} \tag{2.1}$$

The constraint $\sum_{j=1}^{p} |\beta_j| \leq t$ can be written as the l_1 norm constraint $\|\beta\|_1 \leq t$. Typically, the predictors are standardized so the each column of the design matrix is centered ($\frac{1}{n}\sum_{i=1}^{n} x_{ij} = 0$)

and has unit variance ($\frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 = 1$). We also assume that the response variable y_i is centered meaning that $\frac{1}{n}\sum_{i=1}^{n} y_{ij} = 0$. Under these conditions, we can rewrite the constrained problem (2.1) to the following Lagrangian form

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \Big\{ \sum_{i=1}^{n} (y_i - x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \Big\} \tag{2.2}$$

for some $\lambda \geq 0$. By Lagrangian duality, for each value of t in the range where the constraint $\|\beta\|_1 \leq t$ is valid, there is a corresponding value of λ that gives the same solution for the Lagrangian from (2.2) [Hastie et al., 2015]. The l_1 constraint shrinks the small coefficient to zero while large coefficients are also shrunken, but remain nonzero. Thus, a key property of the l_1 penalty is its ability to generate a sparse solution which makes it useful in analysis of "wide" data where number of predictors is higher than that of observations [Hastie et al., 2015].

In contrast to the selection property of the LASSO approach, the solution of *Ridge regression* [Hoerl and Kennard, 1970] is not sparse. It solves the following optimization criterion

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \Big\{ \sum_{i=1}^{n} (y_i - x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \Big\}. \tag{2.3}$$

The ridge regression addresses collinearity problems in high dimensional data analysis, a topic that is covered in detail in Section 3.4.1.

There have been many extensions and improvements of the LASSO method to the situations where LASSO does not perform well. They all share two essential features of original LASSO which are shrinkage of regression coefficients and selection of variables or groups of variables.

Zou and Hastie [2005] propose *Elastic net* by combining the l_1 penalty and l_2 penalty. It solves the problem

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \Big\{ \sum_{i=1}^{n} (y_i - x_{ij}\beta_j)^2 + \lambda \Big[(1-\alpha) \sum_{j=1}^{p} \beta_j^2 + \alpha \sum_{j=1}^{p} |\beta_j| \Big] \Big\}.$$

Elastic net are designed to address the case where there are groups of correlated variables. By combining the LASSO and ridge penalties, it selects the correlated features together and they approximately share similar values.

Yuan and Lin [2006] propose *Group LASSO* to deal with the regression problems in which the features have a group structure. Suppose that J groups of covariates are involved in the regression model where the vector $Z_j \in \mathbb{R}^{p_j}$ represents the covariates in the group j, $j = 1, \ldots, J$. Then a linear model for the function $E(Y|Z)$ takes the form $\sum_{j=1}^{J} Z_j^T \theta_j$ where $\theta_j \in \mathbb{R}^{p_j}$ denotes a group of p_j regression coefficients. Given the n samples $\{(y_i, z_{i1}, z_{i2}, \ldots, z_{iJ})\}_{i=1}^{n}$, Group LASSO solves the problem

$$\hat{\theta} = \underset{\theta_j \in \mathbb{R}^{p_j}}{\arg\min} \Big\{ \sum_{i=1}^{n} (y_i - \sum_{j=1}^{J} z_{ij}\theta_j)^2 + \lambda \sum_{j=1}^{J} \|\theta_j\|_2 \Big\}$$

where $\|\theta_j\|$ is the Euclidean norm of the vector θ_j. It enables to omit the features within a group together by shrinking the entire elements of the vector $\hat{\theta}_j$ to zero.

In some applications, covariates are measured over contiguous time points or adjacent regions. *Fused LASSO* is introduced to account for *spatial* correlation of predictor [Tibshirani

et al., 2005]. It allows the value of neighboring coefficients to be the same or similar by solving the following optimization problem

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \Big\{ \sum_{i=1}^{n} (y_i - x_{ij}\beta_j)^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=2}^{p} |\beta_j - \beta_{j-1}| \Big\}.$$

Here the covariates x_{ij} and their coefficients β_j are indexed along a certain sequence j where the neighboring clumping makes sense [Hastie et al., 2015]. The first penalty is the l_1 penalty of the LASSO regression which shrinks the β_j towards zero. The second constraint is called a *fusion penalty* in that it encourages the neighboring coefficients to be identical or similar, hence it enhances sparsity in the differences of adjacent coefficients.

Frank and Friedman [1993] propose a generalization of penalized l_q regression, called *bridge regression*. It involves l_q penalty of β which is defined as [Liu et al., 2007]

$$\big\|\beta\big\|_q^q = \sum_{j=1}^{p} |\beta_j|^q. \tag{2.4}$$

For a fixed real number $q \geq 0$, the bridge regression solves the constrained least square optimization problem

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \Big\{ \sum_{i=1}^{n} (y_i - x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q \Big\}. \tag{2.5}$$

This creates a bridge between *best subset* regression ($q = 0$) [Beale et al., 1967, Hocking and Leslie, 1967] and the ridge regression ($q = 2$). For $q = 0$, the term $\sum_{j=1}^{p} |\beta_j|^q$ counts the number of nonzero components in β. It is known that classical variable selection criteria

such as AIC, AICc, BIC, Mallow's C_p, or the adjusted R^2 are variations of the l_0 penalized regression [Desboulets, 2018, Fan and Lv, 2010].

TABLE I: ESTIMATORS FROM L_q PENALTY UNDER AN ORTHONORMAL DESIGN MATRIX

q	Estimator	Formula
0	Best subset	$\tilde{\beta}_j \cdot \mathbb{I}[\lvert\tilde{\beta}_j\rvert > \sqrt{2\lambda}]$
1	LASSO	$\text{sign}(\tilde{\beta}_j)\ (\lvert\tilde{\beta}_j\rvert - \lambda)_+$
2	Ridge	$\tilde{\beta}_j/(1+\lambda)$

(Source: [Hastie et al., 2015])

As illustrated in Table I, the best subset selection leaves the coefficient as it is if its absolute value is larger than $\sqrt{2\lambda}$, otherwise shrinks it to zero. This is refered to as hard thresholding. The ridge regression yields proportaional shrinkage, and LASSO applies soft thresholding which translates the coefficient by a constant factor λ and truncates at zero. Figure 2 presents the constraint regions where $\sum_{j=1}^{p} |\beta_j|^q \leq 1$ is satisfied when there are two predictors [Hastie et al., 2015]. For $q < 1$, the constrained region is nonconvex, hence LASSO is the case in that it has the smallest value of $q = 1$ that leads a convex optimization problem. When $q > 2$, the bridge

regression tends to shrink less for the small coefficients and more for the large ones, thus the bridge regression does not capture the large signals as well as LASSO does [Fu, 1998].

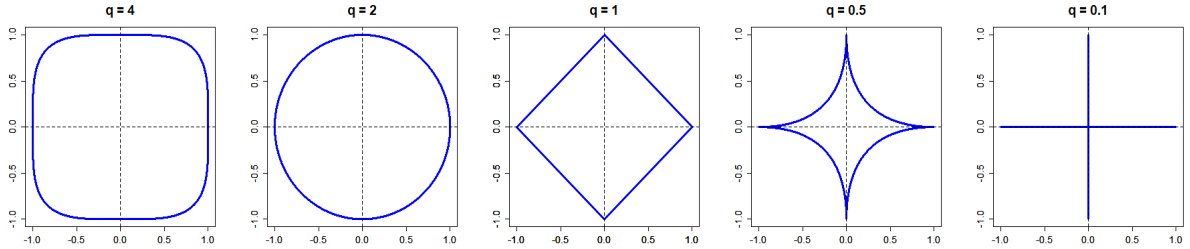


Figure 2: Constraint regions of $\sum_{i=1}^{p} |\beta_j|^q \leq 1$ for different values of q

The *Adaptive LASSO* proposed by Zou [2006] assigns weights to coefficients in the penalty function. Given a pilot estimate $\tilde{\beta}$, it solves

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \left\{ \sum_{i=1}^{n} (y_i - x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} w_j |\beta_j| \right\} \tag{2.6}$$

where $w_j = 1/|\tilde{\beta}_j|^v$. Adaptive LASSO performs sparser selection than LASSO. It can be viewed as an approximation of the l_q penalization with $q = 1 - v$. One merit of Adaptive LASSO is that the optimization problem (2.5) is convex given that the pilot estimates. Moreover, it is shown by Zou [2006] that if the pilot estimates are $\sqrt{n}$ consistent such as the ordinary least

square estimator, Adaptive LASSO works better in recovering the true model than the LASSO approach [Qiao, 2014, Hastie et al., 2015].

There have been selection methods that involve l_q penalties ($0 \leq q \leq 1$) as solutions to the over-shrinking problems of the LASSO regression that may arise in some sparse conditions [Hastie et al., 2015]. However, the nonconvexity brings computational complexity for which alternative nonconvex penalties have been introduced such as *Smoothly Clipped Absolute Deviation (SCAD)* penalty [Fan and Li, 2001] and *Minimax Concave Penalty (MCP)* [Zhang, 2010]. The general form of the nonconvex optimization problem can be written as

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \left\{ \sum_{i=1}^{n} (y_i - x_{ij}\beta_j)^2 + \sum_{j=1}^{p} P(\beta_j) \right\}$$

where the penalty function $P(\cdot)$ is specific to each method. The SCAD penalty on each coordinate is defined by

$$P_{\lambda,\alpha}(\beta) = \lambda \left\{ I(\beta \leq \lambda) + \frac{(\alpha\lambda - \beta)_+}{(\alpha - 1)\lambda} I(\beta > \lambda) \right\}$$

where for some $\alpha > 2$. The MCP penalty is given by

$$P_{\lambda,\gamma}(\beta) := \int_0^{|\beta|} \left(1 - \frac{x}{\lambda\gamma}\right)_+ dx$$

where the nonconvexityy parameter $\gamma \in (0, \infty)$. The MCP bridges between LASSO ($\gamma = \infty$) and best subset ($\gamma = 1_+$). The SCAD penalty does not excessively penalize the large coefficients,

so the estimates are nearly unbiased for such coefficients [Fan and Li, 2001]. Both concave optimization problems can be solved by the local quadratic approximation introduced by Fan and Li [2001], however, the computation is challenging compared to the LASSO type approaches because fast computation methods such as convex programming approach can not be directly used to computing the estimator.

Other than Elastic net, Group LASSO, and Fused LASSO, there have also been other penalized regression methods proposed for grouped predictors. One of possible options is first to cluster the features based on their correlation structure and to take the averages of the features in each group as a new set of predictors to perform variable selection on. Park et al. [2007] introduce that approach by using hierarchical clustering to define the groups and LASSO to conduct a subset selection of averaged predictors. However, it is sometimes more desirable to keep all variables in analysis rather than using the averaged predictors in terms of better prediction performance [Jang et al., 2013]. There are other variable selection methods that incorporate graphical structure among predictors such as *SRIG* (Sparse Regression Incorporating Graphical structure among predictors) [Yu and Liu, 2016].

The *OSCAR* (Octagonal Shrinkage and Clustering Algorithm for Regression) [Bondell and Reich, 2008] and *PACS* (Pairwise Absolute Clustering and Sparsity) [Sharma et al., 2010] select groups of correlated features to deal with multicollineairty [Xie et al., 2015]. The methods

involve novel penalty functions to encourage correlated variables to have identical coefficient estimates. The OSCAR solves the optimization criterion

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \Big\{ \sum_{i=1}^{n} (y_i - x_{ij}\beta_j)^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j<k} max\{|\beta_j|, |\beta_k|\} \Big\}.$$

The second penalty, a pairwise L_∞ norm, can also be represented as $\sum_{j<k} max\{|\beta_j|, |\beta_k|\} = \sum_{j<k} \frac{1}{2}\{|\beta_k - \beta_j| + |\beta_j + \beta_k|\}$. It makes OSCAR a special case of PACS in that OSCAR assigns the same weights 0.5 to the difference of pairs of coefficients and the sums of pairs of coefficients. By adopting the penalties, OSCAR and PACS allow grouping of not only positively, but also negatively correlated predictors with the octagonal shape of constraint region.

Jang et al. [2013] propose *HORSES* (Hexagonal Operator for Regression with Shrinkage and Equality Selection) It solves the following problem

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \Big\{ \sum_{i=1}^{n} (y_i - x_{ij}\beta_j)^2 + \lambda \Big[\alpha \sum_{j=1}^{p} |\beta_j| + (1-\alpha) \sum_{j<k} |\beta_j - \beta_k| \Big] \Big\}$$

where $0 < \alpha \leq 1$. The penalty of HORSES combines the l_1 penalty and another l_1 penalty for pairwise differences of coefficients. By accommodating a penalty similar to Fused LASSO the constraint region of HORSES is of hexagonal shape, hence it encourages the coefficients of positively correlated predictors to be identical rather than grouping both positively and negatively correlated features.

2.4 Feature Screening Methods

While the penalized model selection methods have been widely applied in high dimensional data analyses, the dimension of data may grow exponentially with the sample size in many areas. The problem arise frequently in many areas such as genomics, proteomics, tomography, and biomedical imaging studies where the number of features p can be much more than the sample size n [Fan and Lv, 2010]. These data are often called ultrahigh dimensional data in the literature. In this setting, penalized approaches may not work well suffering from high variance and overfitting caused by the curse of dimensionality [Wang and Leng, 2016, Pan et al., 2019]. To overcome the challenges associated with ultrahigh dimensionality, many feature screening methods have been proposed. The main idea of screening methods is to decrease the dimensionality from huge to moderate scale by filtering out many uninformative features [Fan and Lv, 2008, Fan and Lv, 2010]. Such reduction of the dimension of candidate features is an important aspect of variable selection in dealing with ultrahigh dimensional features space [Desboulets, 2018]. Subsequent to screening, refined penalized methods such as LASSO, SCAD, MCP, or Adaptive LASSO can be applied to the reduced feature space [Pan et al., 2019].

2.4.1 Sure Independence Screening

Many commonly used techniques for features screening is based on *independence screening*. Independence screening rank features with respect to marginal utility meaning that each feature is independently used as a predictor to assess its usefulness in predicting the response [Fan et al., 2009]. *Sure Independence Screening (SIS)* proposed by Fan and Lv [2008] is one of the most well-

known independence screening methods and many other methods have been derived from it. It uses marginal regression coefficients from the linear regression as their marginal utility, and rank the importance of features in terms of the magnitude of the coefficients [Fan et al., 2009, Fan and Song, 2010]. In the setting of linear regression, the marginal regression coefficients turn out to be the sample marginal correlations with the response when the predictors are standardized and the response is centered. That is, independence screening recruits predictors which have high marginal utility that corresponds to large marginal correlation with the response [Fan et al., 2009]. It is also called *correlation learning* or *correlation screening*. More precisely, let w be the p-vector obtained by componentwise regression

$$w = (w_1, \ldots, w_p)^T = X^T y$$

where each column of $n \times p$ design matrix X is assumed to be standardized with mean zero and variance one. Then w is a vector of marginal correlation between features and the response which is rescaled by the standard deviation of the response. For any given d_n, we take the submodel given by

$$\hat{M}_d = \{1 \leq j \leq p : |w_j| \text{ is among the first largest } d_n \text{ of all}\}.$$

The model size of the full model p, which could be lager than the sample size n, is reduced to d_n which can be less the n. The method screens predictors that have weak marginal correlations, hence can rapidly decrease the dimension of the parameter space. Fan and Lv [2008] show that

feature screening using correlation ranking possesses a sure independence screening property which means, with asymptotic probability one, the technique recruits all of the important variables in the model under certain regularity conditions [Fan et al., 2009].

2.4.2 Iterative Sure Independence Screening

Sure independence screening only uses marginal information of predictors and the sure independence screening property can be violated when the regularity conditions fail [Fan and Lv, 2010]. One important condition made for sure independence screening is that the marginal correlation of the important variables must be bounded away from zero. However, this assumption often fail as features are often correlated in high dimensional data sets [Wang and Leng, 2016]. Three potential problems that can arise with SIS are listed in Fan and Lv [2010] as follows.

- **False Positive:** Unimportant features that are strongly correlated with important predictors can have higher chance to be selected than important variables that are weakly correlated with the response.
- **False Negative:** Important variables that are jointly correlated with the response can be screened out because of their low marginal correlations with the response.
- **Collinearity:** Collinearity among the features adds difficulty to the problem of variable selection.

For these reasons, Fan and Lv [2008] propose *Iterative Sure Independence Screening (ISIS)* as an extension of SIS to cover the cases where the regularity conditions may fail. The ISIS is an iterative procedure that repeatedly applies SIS using the working residuals. Roughly, ISIS

performs correlation screening followed by a lower dimensional model selection method, and find the residual based on the fitted model with the selected variables. By updating the residual as the new response variable, it continues the procedures until the dimension of the recruited variable is less than the sample size.

More precisely, we first apply SIS screening by using correlation ranking of predictors to have the $[n/log(n)]$ screened variables. Then we perform penalized variable selection such as LASSO or SCAD on the set of screened features. Let the set of selected k_1 variables be denoted by $A_1 = \{x_{i_1}, \ldots, x_{ik_1}\}$. We find the n-vector of residuals by regressing the response y over $x_{i_1}, \ldots, x_{ik_1}$. We treat these residuals as the new responses and repeat the same procedure as in the previous steps to the remaining $p - k_1$ features. Let the subset of selected k_2 variables from this stage be denoted as $A_2 = \{x_{j_1}, \ldots, x_{jk_2}\}$. Fan and Lv [2008] note that by using the residuals as the new responses, ISIS helps to weaken the priority of unimportant variables that are highly correlated with the outcome through important variables because when the response is regressed on the set A_1, the residuals are not correlated with the selected features in A_1, which addresses the false positive problem mentioned before. In addition, ISIS reconsiders the missed important features since those important variables that have weak marginal correlations with the response only because of the presence of the variables in set A_1, should now be related with the residuals. This addresses the false negative problem. The ISIS allows deletion of variables as well during the iterative process which can also deal with the false positive issue [Fan and Lv, 2010]. The procedure is continued until disjoint l subsets $A_1, \ldots, A_l$ are obtained which satisfy the size d of the union $A = \bigcup_{i=1}^{l} A_i$ is less than the sample size n.

2.5 Bayesian Variable Selection Methods

In this section, we consider *Bayesian Kernel Machine Regression (BKMR)* among several Bayesian variable selection methods that have been proposed. The BKMR is an extended approach from *Kernel Machine Regression (KMR)* methods that concentrate on variable selection and prediction and is developed specifically focusing on estimating health effects of environmental mixtures. It also involves a hierarchical variable selection approach that can account for the structure of correlated mixture components [Bobb et al., 2015,Bobb et al., 2018]. The BKMR considers the model given by

$$y_i = h(z_i) + x_i^T \beta + \epsilon_i$$

where y_i is a health outcome, $z_i = (z_{i1}, z_{i2}, \ldots, z_{iM})^T$ is a vector of environmental exposures, x_i is a set of confounders, and $\epsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$. The high dimensional exposure-response function is characterized by $h(\cdot)$ which is represented by a kernel function. A kernel function $K(\mathbf{z}, \mathbf{z}^{'})$ involves two arguments: $\mathbf{z} = (z_1, \ldots, z_M)^T$, the vector for environmental mixtures for one subject, and $\mathbf{z}^{'} = (z_1^{'}, \ldots, z_M^{'})^T$, the profile of the exposures for a second subject. The function $h(\cdot)$ can be represented with a positive-definite kernel function $K(\cdot, \cdot)$ which is named *dual form*, with $h(\mathbf{z}) = \sum_{i=1}^{n} K(\mathbf{z}_i, \mathbf{z})\alpha_i$ for some set of coefficients $\{\alpha_i\}_{i=1}^{n}$. Examples include the linear kernel $K(\mathbf{z}, \mathbf{z}^{'}) = 1 + z_1 z_1^{'} + \cdots + z_M z_M^{'}$, the quadratic kernel $K(\mathbf{z}, \mathbf{z}^{'}) = (1 + z_1 z_1^{'} + \cdots + z_M z_M^{'})^2$, and Gaussian kernel $K(\mathbf{z}, \mathbf{z}^{'}) = exp\{-\sum_{m=1}^{M}(z_m - z_m^{'})^2/\rho\}$ with ρ a tuning parameter [Bobb et al., 2015].

2.5.1 Componentwise Variable Selection

To incorporate Bayesian variable selection approach, the augmented Gaussian kernel function is defined which is represented as $K(\mathbf{z}, \mathbf{z}^{'}; \mathbf{r}) = exp\{-\sum_{m=1}^{M} r_m(z_m - z_m^{'})^2\}$, where $\mathbf{r} = (r_1, \ldots, r_M)^T$. On the basis of the approaches of Bayesian variable selection for multiple regression problems [George and Mcculloch, 1993], a *slab-and-spike* prior is assumed on auxiliary parameters as follows

$$r_m \,|\, \delta_m \;\sim\; \delta_m f_1(r_m) + (1 - \delta_m) P_0, \quad m = 1, \ldots, M,$$

$$\delta_m \sim Bernoulli(\pi)$$

where $f_1(\cdot)$ is a density with support on $\mathbb{R}^+$ and P_0 denotes the density with point mass at 0 [Bobb et al., 2015]. The posteria mean of the indicator δ_m can be interpreted as the posterior probability that the exposure m is an important component of the mixture, or as the posterior *inclusion probability* of the exposure m [Bobb et al., 2015] which indicates the measure of importance for each exposure [Bobb et al., 2018].

2.5.2 Hierarchical Variable Selection

The componentwise variable selection may fail when mixture pollutants are strongly correlated since the method treats components exchangeably. In this situation, a hierarchical variable selection approach is considered that is proposed to incorporate information of the structure of the environmental mixture into the model [Bobb et al., 2015]. Suppose the mixture components $z_a, \ldots, z_M$ are correlated in multiple groups G_k $(k = 1, \ldots, q)$ with high within-group corre-

lation and low across-group correlation. Then the indicator variables from the slab-and-spike prior are distributed as follows

$$\delta_{G_k} \,|\, \omega_k \;\sim\; Multinomial(\omega_k, \pi_{G_k}), \quad k = 1, \ldots, q,$$

$$\omega_k \sim Bernoulli(\pi)$$

where $\delta_{G_k} = (\delta_m)_{z_m \in G_k}$ is the vector of indicator variables and π_{G_k} is the vector of prior probabilities for the mixture elements z_m in group G_k. This method enables BKMR to estimate the posterior inclusion probability for each group of exposures. It provides the posterior inclusion probability for each pollutant within each group as well given that the group was entered into the model [Bobb et al., 2018]. This approach allows most of the mixture components in the same group to be included in the model together at a time [Bobb et al., 2015].

CHAPTER 3

HIGH DIMENSIONAL VARIABLE SELECTION IN PRESENCE OF COLLINEARITY: LINEAR MODEL

3.1 Introduction

In this chapter, we propose a method for high dimensional variable selection in presence of collinearity among predictors in the setting of linear models. We conduct simulation studies to evaluate the performance of the method with respect to multiple performance criteria. We also apply the method to the National Health and Nutrition Examination Survey (NHANES) persistent organic pollutants dataset.

3.2 Background

Suppose that we observe $(x_1, y_1), \ldots, (x_n, y_n)$ where $x_i = (x_{i1}, \ldots, x_{ip})^T$ is a p-dimensional predictor and y_i is the response variable. We consider a linear regression model

$$y = X\beta + \epsilon$$

where $y = (y_1, y_2, \ldots, y_n)^T$ is an n-vector of responses, $X = (x_1, x_2, \ldots, x_n)^T$ is an $n \times p$ design matrix of predictors, $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^T$ is a p-vector of parameters, and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$ is an n-vector of errors. We assume the column vectors of the predictor matrix are standardized

to have mean 0 and unit variance and the response variable is centered to have mean 0 such that

$$\sum_{i=1}^{n} y_i = 0, \qquad \sum_{i=1}^{n} x_{ij} = 0, \qquad \frac{1}{n-1}\sum_{i=1}^{n} x_{ij}^2 = 1, \quad j = 1, 2, \ldots, p.$$

3.3 Major Concepts and Methods

3.3.1 Two Main Procedures

The method consists of two main procedures. We first screen variables by constructing the lead cluster based on the lead variable which has the highest marginal correlation with the outcome. Secondly, variable selection is performed in the lead cluster which contains the variables highly correlated with the lead variable.

These two procedures are iteratively conducted where at each iteration, the outcome is updated by the residual from the model resulted at the previous iteration. That is, we find the second lead variable which is the most correlated with the outcome, here the residual, and construct the second lead cluster with the variables strongly correlated with the second lead variable. Then variable selection is performed on the set of the selected variables from the previous selection and the variables contained in the second lead cluster. We repeat the procedures so that we can choose the model which has the minimum cross-validation error among the iterations. The detailed steps of the method are introduced in the following section.

3.3.2 Correlation Learning with the Response

We consider high dimensional data which mean the number of predictors p is high or relatively high in the data set. Feature screening is regarded as an effective strategy for dimension reduction. We propose to use correlation learning using $w = (w_1, w_2, \ldots, w_p)^T$ by componentwise regression,

$$w = X^T y$$

where X is $n \times p$ design matrix of predictors $X_1, \ldots, X_p$ and y is a $n \times 1$ vector of responses. Each column of X is standardized, hence w is a vector of marginal correlations between predictors and the response which is rescaled by the standard deviation of the response.

These p components are ranked based on the magnitudes of absolute value of $|w_j|,\ j = 1, 2, \ldots, p$. We choose the feature with the largest value of this absolute magnitude, and call it *lead variable*. The absolute magnitudes indicate the importance of features in terms of their marginal correlations with the response. That is, the feature of higher value is regarded as more important according to the ability of predicting the response. Hence, the lead variable chosen at first out of all features can be considered as the most useful variable among all features with regard to the marginal relationship with the outcome variable. The 1^{st} lead variable which is denoted by $X^{[1]}$ can be written as

$$X^{[1]} = \arg\max_{X_j} \{|w_j|,\ j = 1, 2, \ldots, p\}$$

where w_j is the marginal correlation of X_j and y.

3.3.3 Correlation Learning with Predictors

After identifying the lead variable, we construct a group of features that are highly correlated with the lead variable. Since the lead variable is the most marginally related feature with the response, the features of high correlation with the lead variable can also be reasonably considered as important features in terms of the relationship with the response. Thus, we call this group of features *lead cluster*.

In order to identify the features that are included in the lead cluster, we use the correlation learning method in the same way we use it to select the lead variable. However, we replace the response by the lead variable because the relationship that is considered is between the lead variable and the remaining features. We find a vector $\rho = (\rho_1, \rho_2, \ldots, \rho_p)$ by componentwise regression

$$\rho = X^T X^{[1]}$$

where $X^{[1]}$ is the 1^{st} lead variable. Each component ρ_j represents the correlation between X_j and the lead variable because all the features are standardized in the beginning. The vector ρ can be easily found by taking the j^{th} column or row of the $p \times p$ correlation matrix $X^T X$ when X_j is the lead variable, $j = 1, 2, \ldots, p$.

We sort the p components of ρ in a decreasing order and define the first lead cluster C_1 with the submodel

$$C_1 = \{X_j, j = 1, \ldots, p : |\rho_j| > \delta\}$$

where δ is a cutoff value. This screening procedure can be viewed as a clustering problem which is one of typical unsupervised learning methods where we need to set a cutoff for a distance measure to split data into similar groups. Since we consider features that are moderately or strongly correlated, we follow a similar approach as in the paper [Xie and Zeng, 2010] that the cutoff value can be the 0.75^{th} percentile of all pairwise correlations between features or the correlation coefficient that indicates moderate to high correlation such as the value in the range of 0.5 to 0.9. We can use the cutoff value that is initially chosen based on the correlation structure of the data for the whole procedures, but we also devise a cutoff adjustment method that can be implemented as an option for more efficient selection of features which is described in Section 3.3.7.

3.3.4 Selection of Predictors

After we screen features by identifying the lead cluster, we perform variable selection on the screened features in the cluster. As introduced in Chapter 2, ridge regression has been widely used under conditions of collinearity since it handles collinearity problems by adding a constant to the diagonal of $X^T X$ to improve its condition number [García et al., 2015, Liu, 2003, Hoerl

and Kennard, 1970]. The detail of its rationale is addressed in Section 3.4. The ridge estimator is decided as follows

$$\hat{\beta}^{Ridge} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

where some $\lambda \geq 0$. However, it does not turn out an easily interpretable model because it only shrinks the value of coefficients not setting any of them to 0 [Tibshirani, 1996]. As an alternative, we consider the method that combines the l_2 penalization of ridge with another penalty which allows selection of predictors.

The elastic net is one of famous methods among them which combined the ridge and the LASSO penalties [Zou and Hastie, 2005]. The estimator of the elastic net is given by

$$\hat{\beta}^{Elnet} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \Big[(1-\alpha) \sum_{j=1}^{p} \beta_j^2 + \alpha \sum_{j=1}^{p} |\beta_j| \Big] \right\}$$

where $\lambda \geq 0$ and $\alpha \in [0, 1]$ is a parameter that can be varied. The l_1 penalization of LASSO shows poor performance when predictors are strongly correlated. As shown in Zou and Hastie [2005], the solution paths of LASSO become unstable as the level of collinearity among predictors increases. Since it is more often to encounter collinearity problems in high dimensional data, Zou and Hastie[2005] propose the elastic net to improve the performance of LASSO in high dimensional data analysis [Zou and Zhang, 2009].

The elastic net combines the merits of LASSO shrinkage and the quadratic regularization of ridge. The l_1 penalty of lasso performs variable selection that derives a more simpler model with

a subset of features by forcing some of weak coefficients to be set to 0 [Tibshirani, 1996, Zou and Zhang, 2009]. The component of the ridge penalty controls for strong within-group correlations [Hastie et al., 2015] by stabilizing the solution paths. Hence, when the level of collinearity becomes high, the elastic net makes a significant improvement in the prediction performance of LASSO [Zou and Zhang, 2009].

Other than the LASSO penalty, we can also consider other penalties that can be combined with the ridge penalty. The SCAD [Fan and Li, 2001] penalty or MCP that is introduced in Chapter 2 can be another option to apply since these penalties are known to have oracle property when the penalization parameters are carefully chosen [Fan and Li, 2001, Zhang, 2010].

In the implementation of selection methods, there are tuning parameters α and λ that should be chosen. There are several methods available to decide tuning parameters such as cross-validation, generalized cross-validation, AIC, and BIC [Zou and Zhang, 2009]. We use cross-validation method as the book Hastie et al. [2015] suggest for choosing the tuning the parameters, which is described in more detail in Section 3.3.6.

3.3.5 Iterative Cluster Based Selection: Iterative Correlation Learning

The variables selected in the above approach are representative of only a subset of predictors, the screened features included in the lead cluster. There are potentially other important predictors that are not considered in the selection because they are not included in the lead cluster. To fully consider the joint information from all of the predictors, we use an iterative approach. The key is to iteratively perform correlation screening followed by variable selection that can tackle collinearity among the screened variables. By doing this, we expect the method

to find the model which has improved performance in variable selection and prediction. The conceptual diagram of iterative cluster based selection can be described as in Figure 3.

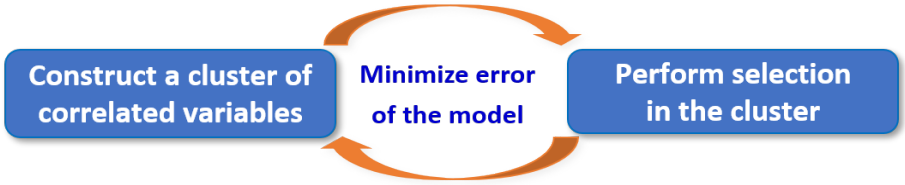


Figure 3: Conceptual diagram: iterative cluster based selection

The iterative correlation learning and variable selection work as follows. Let us define the active set A_k that contains variables that can be considered to be a lead variable. In the beginning, the active set A_1 includes all of the predictors in the data set as $A_1 = \{X_1, X_2, \ldots, X_p\}$. Let C_1 and S_1 denote the 1^{st} lead cluster and the set of selected variables, respectively. Then, the set S_1 is given by

$$S_1 = \{X_{1j}, j = 1, \ldots, p_1 : |\hat{\beta}_{1j}| > 0\}$$

where $C_1 = \{X_{11}, \ldots, X_{1p_1}\}$ and $\hat{\beta}_{1j}$ is the coefficient estimate of the variable X_{1j} in C_1, $j = 1, \ldots, p_1$ using a variable selection method such as elastic net. The lead cluster includes the predictors that are selected by correlation screening with the lead variable, that is, the

variable of the largest correlation with the response. The variables included in S_1 are selected by applying the selection method in the lead cluster C_1.

We define the residuals of the selected model from this stage as

$$r^{(2)} = r^{(1)} - X_{S_1}\hat{\beta}_{S_1}$$

where $r^{(1)}$ is the n-vector of original response variable y, X_{S_1} is the design matrix of variables in S_1, and $\hat{\beta}_{S_1}$ is the vector of estimated coefficients. Then we have an n-vector of residuals by regressing the response over the selected variables. In this way, the response is adjusted for the selected variables which are representative of the lead cluster. We treat the residuals $r^{(2)}$ as the new responses for stage 2 and conduct the same method as in the previous procedures. However, we find the next lead variable over the variables in the updated active set defined as

$$A_2 = A_1 \setminus C_1.$$

Since the original response is adjusted for the effects of the lead cluster C_1, the variables in C_1 are excluded when we perform correlation learning with the new response. By allowing the next lead variable to be found out of the lead cluster that was just considered for variable selection, it blocks the possibility to get stuck in a certain cluster in performing selection of predictors. Hence, it enables the method to explore the domain of all predictors in term of their relationship with the response.

We perform correlation learning with the new response $r^{(2)}$ and the predictors contained in the new active set A_2, which turns out the next lead variable $X^{[2]}$ that has the largest correlation with the new response. Then we construct the next lead cluster C_2 of the features that are highly correlated with the lead variable $X^{[2]}$ following the same previous method. Since the new response is not correlated with the selected variables in S_1, it can weaken the priority of unimportant variables that are correlated with the original response through the selected variables in S_1 [Fan and Lv, 2008]. It also enhances the priority of an important variable that has weak marginal correlations but high joint correlation with the selected predictors [Fan and Lv, 2008] when we find the next lead variable $X^{[2]}$. This makes the lead cluster C_2 that is constructed based on $X^{[2]}$ less correlated with the previous lead cluster C_1. Now we have a set of selected variables S_1 and the new lead cluster C_2, so we conduct selection on the variables in the union of $(S_1 \cup C_2)$.

In a more generalized notation, the k^{th} variable selection is performed on the variables included in the set $(S_{k-1} \cup C_k)$ which produces a set of selected variables S_k where C_k denotes k^{th} lead cluster, and $S_0 = \emptyset$. Then we update the $(k+1)^{th}$ response by

$$r^{(k+1)} = r^{(k)} - X_{S_k}\hat{\beta}_{S_k}$$

where $r^{(1)}$ is the original response variable y, S_k is the set of selected variables from the k^{th} variable selection, X_{S_k} is the design matrix of variables in S_k, and $\hat{\beta}_{S_k}$ is the vector of estimated

coefficients. The $(k+1)^{th}$ lead variable $X^{[k+1]}$ is found among the variables in the active set defined by

$$A_{k+1} = A_1 \setminus (S_k \cup C_k).$$

As noted before, updating the response and the active set in this manner permits the next lead variable chosen outside the union of already selected features or just considered ones for variable selection. It enables the following lead cluster to be composed of the variables that are less correlated with the predictors in the union. Additionally, we expect the method to deal with two issues described in Fan and Lv [2008] in the correlation learning searching for the next lead variable: It enhances the priority of an important predictor if it is marginally weakly correlated but jointly correlated with the response through the variables in the union; It also weakens the priority of an unimportant variable that is jointly correlated with the response through the variables in the union. Another critical benefit of the iterative method is that it makes the important variables that are missed in the previous procedures possible to be included [Fan and Lv, 2008] and allows the unimportant variables that are chosen in the model to leave the selected set of predictors.

3.3.6 Tuning Parameter Selection

In the implementation of the selection method, the tuning parameters α and λ need to be chosen. The α represents proportion between ridge and another penalty. For example, in the case of the elastic net, when $\alpha = 1$, it reduces to the l_1-norm corresponding the LASSO penalty,

and with $\alpha = 0$, it reduces to the the square l_2-norm equivalent to the ridge penalty. Hastie et al. [2015] recommend that in practice, α can be regarded as a higher-level parameter, and can be chosen on subjective grounds. Alternatively, a (coarse) gird of values of α can be used in a cross-validation scheme.

Cross-validation is the method that can be used to create artificial training and test data sets by splitting up the given data into roughly equal folds and evaluating the repeated performance on the test data set. Hastie et al. [2015] describe the details of conducting cross-validation. For k-folds cross-validation, we first split the data into k folds, $k > 1$, and designate one fold as the test set with the remaining $k - 1$ folds used as the training set. We then fit the model on the training set at each combination of (α, λ), and let each fitted model predict the response on the test set. Performance can be evaluated in terms of mean squared prediction errors given by

$$E_k(\alpha, \lambda) = \sum_{i \in k^{th} fold} (y_i - x_i^T \beta^{-k}(\alpha, \lambda))^2, \quad k = 1, \ldots, K$$

where $\beta^{-k}(\alpha, \lambda)$ is the vector of regression coefficients from the fitted model using the training set over a range of each combination of parameters. This process is repeated K times by letting each fold play the role of the test set. Then we have the K estimates for prediction error, and the cross-validation error can be obtained by averaging these values such as

$$CV(\alpha, \lambda) = \frac{1}{K} \sum_{k=1}^{K} E_k(\alpha, \lambda).$$

We choose (α, λ) which yields minimum cross-validation error. Regression coefficients can be estimated at these parameters or at the "one-standard-error rule" choice [Hastie et al., 2015] which means selecting the most parsimonious model which allows its error no more than one standard error of the cross-validation estimates from the minimum value of cross-validation error [Hastie et al., 2009, Curto and Pinto, 2014, Hastie et al., 2015]. In the numerical studies, we use $K = 10$ as $K = 5$ or 10 are typical choices [Hastie et al., 2015].

3.3.7 Adaptive Cutoff

In this proposed correlation learning method, the lead cluste is formed based on a cutoff for the correlation coefficients between the lead variable and the remaining predictors. As mentioned in Section 3.3.5, k^{th} feature selection is performed on the set $(S_{k-1} \cup C_k)$, the union of predictors that are selected at previous iteration and the ones that are recruited in the current lead cluster. However, if few variables are contained in the cluster, the set of data on which we perform selection is barely changed from the set that was considered at previous iteration. In an extreme, when there is no predictor that is correlated with the lead variable more than the cutoff value, the lead cluster cannot recruit any other predictor except the lead variable itself. From this reason, we propose an adaptive cutoff adjustment method that allows to include at least a certain amount of predictors in the lead cluster. More precisely, when the lead cluster doesn't include a certain percentage of predictors, we adaptively update the cutoff so that the cluster can recruit at least the specified percentage of predictors in the lead cluster.

The SIS method filters d features having the largest marginal correlation with the response. Fan and Lv [2008] mention that d can be $n - 1$ or $n/log(n)$ so that it can be less than the

sample size n, but it can also be greater than the sample size because as it is larger, there is higher chance the true model can be covered. Zhong and Zhu [2015] suggest $d = 2n/log(n)$ for its correlation learning based iterative approach. We can adopt either of criterion for the minimum number of recruited predictors in the lead cluster. For instance, in the simulation studies where we generate data sets of $n = 150$ and $p = 200$, the cutoff is adjusted until we contain at least $2n/log(n) \approx 60$ predictors in the lead cluster which take 30 percent of the total number of features. Using this approach, as the predictors are closer to orthogonal, the proposed method tends to original selection approach such as Elastic net, or SCAD or MCP combined with the ridge penalty on the entire data set. When the sample size is high, $n/log(n)$ or $2n/log(n)$ can be large, even can be greater than p. We generally recommend to use values such as 20 to 40 percentage so that the method can create various sets of predictors on which selection is performed while we guarantee the minimum size of the lead cluster.

3.4 Rationales

3.4.1 Measures of Collinearity

High collinearity in the set of predictors induces many problems on inference and prediction that is based on the regression analysis. When there is complete absence of linear relation among the predictors, they are called to be *orthogonal* [Chatterjee, 2012]. When the predictors are orthogonal, $X^T X$ is diagonal. The condition of severe departure from orthogonality is referred to *collinearity* or *multicollinearity*, that is, there exist strong linear relationships among the predictors in the data. When at least one eigenvalue deviates from 1, especially towards the value very close to 0, that indicates nonorthogonality exists meaning that multicollinearity is

present [Vinod and Ullah, 1981, Walker, 1989, Greene, 1993]. When there is a complete linear relationship among the predictors, the rank of the design matrix X and X^TX is below the number of regression parameters p, that is, $rank(X) = rank(X^TX) < p$. It means the nullity of the determinant of X^TX and the OLS estimators

$$\hat{\beta} = (X^TX)^{-1}X^Ty, \quad \hat{Var}(\hat{\beta}) = s^2(X^TX)^{-1}, \quad s^2 = \frac{e^Te}{n-p-1}, \quad e = y - X\hat{\beta}$$

do not exist [Curto and Pinto, 2007].

There are several measures to investigate collinearity. One quantity is *variance inflation factor (VIF)*. The variance inflation for the predictor X_j is given by

$$VIF_j = \frac{1}{1-R_j^2}, \quad j = 1, \ldots, p$$

when R_j^2 is the coefficient of determination when the predictor X_j is regressed by all the other predictors. In absence of any linear relation between X_j and the remaining variables, VIF_j is 0. In contrasts, if there is a strong linear relationship, the value of VIF_j is large with R_j^2 close to 1. Hence, as the value of VIF_j deviates from 0, there is more departure from orthogonality and higher tendency toward collinearity. It is often regarded as a signal of collinearity problems if the value of the variance inflation factor is larger than 10 [Mardikyan and Çetin, 2008, Chatterjee, 2012].

There are also several measures regarding the *condition number* which indicates degree of multicollinearity. Vinod and Ullah [1981] suggest the condition number given by

$$\phi_1 = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}}$$

where λ_{min} is the smallest eigenvalue and λ_{max} is the largest eigenvalue of $X^T X$. Montgomery et al. [2012] propose the condition number with the ratio of two eigen values given by

$$\phi_2 = \frac{\lambda_{max}}{\lambda_{min}}.$$

If the $\lambda_{min} = 0$, then the condition numbers ϕ_1 and ϕ_2 are infinite which means complete multicollinearity among predictors. When the λ_{min} and λ_{max} are equal, the condition numbers are 1, which indicates predictors are orthogonal. Pagel and Lunneborg [1985] introduce another version of condition number which is defined by

$$\phi_3 = \sum_{j=1}^{p} \frac{1}{\lambda_j}.$$

El-Dereny and Rashwan [2011] mention that if the condition number lies between 5 and 30, it is conventionally taken as moderate to high collinearity.

The concept of condition number is introduced by Atkinson [1989] as a measure of stability for solving a mathematical problem. Let the mathematical problem is given as the form of an equation

$$F(x, y) = 0.$$

The variable x is the unknown quantity that is being sought, and the variable y is data where the solution x depends on. Atkinson [1989] describes the problem to be *stable* when small changes in y lead correspondingly to small changes in x, which is also called *well-posed*. Otherwise, it is called *unstable* or *ill-posed*. The condition number seeks to measure the possible worst impact on the solution x when y is perturbed by a small quantity. Let δy denote a small perturbation of y, and let $x + \delta x$ be the solution of the perturbed equation

$$F(x + \delta x, y + \delta y) = 0.$$

Then the condition number is defined by

$$K(x) = \sup_{\delta y} \frac{\|\delta x\|/\|x\|}{\|\delta y\|/\|y\|}$$

where the notation of vector norm $\|\cdot\|$ denotes a measure of size. $K(x)$ is a measure of sensitivity of the solution x to a small amount of changes in the data y. If $K(x)$ is large, there exist relatively small changes δy in y that cause relatively large changes δx in x. Such

problems are *ill-conditioned* and generally very hard to be solved accurately. If $K(x)$ is small, say $K(x) \leq 10$, small relative perturbations in the data y always induce correspondingly relative small changes in the solution x.

Atkinson [1989] also describes the stability of a linear system $Ax = b$ following this general schemata. Let $Ax = b$, of order n, be uniquely solvable and consider the solution $\tilde{x}$ when r is a small change in b

$$A\tilde{x} = b + r.$$

Let $e = \tilde{x} - x$, then $Ae = r$ and $e = A^{-1}r$. To examine the stability of $Ax = b$ with the condition number, we seek to bound the measure

$$\frac{\|e\|}{\|x\|} \div \frac{\|r\|}{\|b\|}.$$

Take norms to obtain $\|r\| \leq \|A\|\|e\|$ and $\|e\| \leq \|A^{-1}\|\|r\|$, where the matrix norm is the operator matrix norm induced by the vector norm. Divide by $\|A\|\|x\|$ in the first inequality and by $\|x\|$ in the second one to derive

$$\frac{\|r\|}{\|A\|\|x\|} \leq \frac{\|e\|}{\|x\|} \leq \frac{\|A^{-1}\|\|r\|}{\|x\|}.$$

Using the bounds $\|b\| \leq \|A\|\|x\|$ and $\|x\| \leq \|A^{-1}\|\|b\|$, we obtain

$$\frac{1}{\|A\|\|A^{-1}\|} \cdot \frac{\|r\|}{\|b\|} \leq \frac{\|e\|}{\|x\|} \leq \|A\|\|A^{-1}\| \cdot \frac{\|r\|}{\|b\|}.$$

Considering the measure we want to bound, it justifies to put the condition number of A as follows

$$cond(A) = \|A\|\|A^{-1}\|.$$

The condition number $cond(A)$ can vary according to the the norm being used, however it is always bounded by one because

$$1 \leq \|I\| = \|AA^{-1}\| \leq \|A\|\|A^{-1}\| = cond(A).$$

When the condition number is close to 1, as in many common mathematical problems, small relative changes in b will make correspondingly small perturbations in the solution x. If the the value of condition number is large, there may exist small perturbations that cause large changes in x. Since the condition number varies with the choice of norm, we often use another definition of condition number which is independent of the norm. For an arbitrary square matrix A, it is known that

$$\max_{\lambda \in \sigma(A)} |\lambda| \leq \|A\|$$

for any operator matrix norm, where $\sigma(A)$ denotes the set of all eigenvalues of A. We thus have the result

$$cond(A) \geq \frac{\max\limits_{\lambda \in \sigma(A)} |\lambda|}{\min\limits_{\lambda \in \sigma(A)} |\lambda|} \equiv cond(A)_*$$

since the eigenvalues of A^{-1} are the reciprocals of those of A. A linear system is *ill-conditioned* if the solution x is unstable according to the slight changes in b. The condition numbers $cond(A)$ and $cond(A)_*$ are fairly good predictors of ill-conditioning. In general, if $cond(A)_*$ is large, the linear system $Ax = b$ will have the value b for which the system is sensitive to changes r in b.

3.4.2 Rationale of Ridge Regression

Hoerl and Kennard [1970] introduce ridge regression and described the properties of its estimator in comparison with the ordinary least square estimator. In this section, we review ridge regression and its rationale. In a general linear regression model $y = X\beta + \epsilon$ described in Section 3.2, when X is $n \times p$ matrix of rank p, the ordinary least square (OLS) estimate is the minimum variance unbiased linear estimate and is given by

$$\hat{\beta} = (X^T X)^{-1} X^T y.$$

Consider the distance $\hat{\beta}$ from β. The mean and variance of the squared distance are given by

$$E[(\hat{\beta} - \beta)^T (\hat{\beta} - \beta)] = \sigma^2 tr(X^T X)^{-1}$$

$$Var[(\hat{\beta} - \beta)^T (\hat{\beta} - \beta)] = 2\sigma^4 tr(X^T X)^{-2}.$$

The average of the squared distance between the estimator and the value of the parameter is also called mean squared error (MSE). Using the eigenvalues of X^TX, they can be described as follows

$$E[(\hat{\beta}-\beta)^T(\hat{\beta}-\beta)] = \sigma^2 \sum_{j=1}^{p}(1/\lambda_j)$$
$$Var[(\hat{\beta}-\beta)^T(\hat{\beta}-\beta)] = 2\sigma^4 \sum_{j=1}^{p}(1/\lambda_j)^2$$

where the p eigenvalues are denoted by $\lambda_{max} = \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p = \lambda_{min} > 0$. It thus follows that the distance from $\hat{\beta}$ to β is large when X^TX has one or more small eigenvalues. That is, as X^TX deviates from a unit matrix inducing *nonorthogonality*, it tends to have small eigenvalues that cause higher condition number and less possibility of a short distance between $\hat{\beta}$ and β. Moreover, the least square estimator $\hat{\beta} = (X^TX)^{-1}X^Ty$ is not well defined when $p > n$ because the rank of the $p \times p$ matrix X^TX is at most n, hence it is singular and not invertible. It makes the equation $X^TX\beta = X^Ty$ that derives the estimator $\hat{\beta}$ do not have a unique solution for β [Zou and Zhang, 2009].

To tackle the problems associated with the least square estimates, ridge regression was suggested by Hoerl and Kennard [1970]. The rational of the ridge regression is to add a constant to the diagonal of X^TX to improve its condition number [Liu, 2003]. In linear algebra, this

is known as Tikhonov regularization [Tikhonov, 1943]. It uses l_2 penalization of regression coefficients in the constrained least square optimization problem

$$\hat{\beta}^{Ridge} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \left\{ \sum_{i=1}^{n} (y_i - x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

for some $\lambda > 0$. This gives

$$\hat{\beta}^{Ridge} = (X^T X + \lambda I)^{-1} X^T y.$$

It involves an invertible matrix $X^T X + \lambda I$ even when $X^T X$ is singular, hence is more stable. When X is composed of orthonormal variables such that $X^T X = I_p$, the ridge regression estimator is a down-weighted version of the OLS estimator given by

$$\hat{\beta}^{Ridge} = ((1 + \lambda)I)^{-1} X^T y = \frac{1}{1 + \lambda} \hat{\beta}^{OLS}.$$

The effect of the quardratic penalization is to constrain the size of regression coefficients by shrinking them toward zero. In the condition where there are groups of predictors and the level of collinearity is high within each group, the ridge estimation shrinks the coefficients toward each other as well as toward zero [Zou and Zhang, 2009].

3.4.3 Mean, Variance, and MSE of the Ridge Regression Estimator

The expected value of the ridge regression estimator is given by

$$\begin{aligned} E[\hat{\beta}^{Ridge}] &= (X^T X + \lambda I)^{-1} X^T X \beta \\ &= [I - \lambda (X^T X + \lambda I)^{-1}] \beta. \end{aligned}$$

By adding the positive constant to the diagonal of $X^T X$, ridge estimation involves a biased estimator of the true parameter β. However, ridge estimator has smaller variance than the OLS estimator. Two quantities can be compared by finding the trace of the variance matrix of a vector of the estimator which is often called *total variance*. The total variance of the two estimators is given by

$$\begin{aligned} tr(Var[\hat{\beta}^{OLS}]) &= tr(\sigma^2 (X^T X)^{-1}) = \sigma^2 \sum_{j=1}^{p} \frac{1}{\lambda_j} \\ tr(Var[\hat{\beta}^{Ridge}]) &= tr(\sigma^2 (X^T X + \lambda I)^{-1} X^T X (X^T X + \lambda I)^{-1}) \\ &= \sigma^2 \sum_{j=1}^{p} \frac{\lambda_j}{(\lambda_j + \lambda)^2} \end{aligned}$$

where $\lambda_j, i = 1, \ldots, p$ are the eigenvalues of $X^T X$. It shows that the ridge regression estimator has less total variance than the OLS estimator

$$tr(Var[\hat{\beta}^{OLS}]) \geq tr(Var[\hat{\beta}^{Ridge}]).$$

The MSE is the sum of the variance and the squared bias of the estimator. The MSE of the ridge and OLS estimator are obtained by

$$MSE(\hat{\beta}^{OLS}) = E[(\hat{\beta}^{OLS} - \beta)^T(\hat{\beta}^{OLS} - \beta)] = \sigma^2 \sum_{j=1}^{p}(1/\lambda_j)$$

$$\begin{aligned} MSE(\hat{\beta}^{Ridge}) &= E[(\hat{\beta}^{Ridge} - \beta)^T(\hat{\beta}^{Ridge} - \beta)] \\ &= \sigma^2 tr(X^TX + \lambda I)^{-1}X^TX(X^TX + \lambda I)^{-1} \\ &\quad + \lambda^2\beta^T(X^TX + \lambda I)^{-2}\beta \\ &= \sigma^2 \sum_{j=1}^{p} \frac{\lambda_j}{(\lambda_j + \lambda)^2} + \lambda^2\beta^T(X^TX + \lambda I)^{-2}\beta. \end{aligned}$$

Hoerl and Kennard [1970] prove that there always exists a $\lambda > 0$ such that

$$MSE(\hat{\beta}^{Ridge}) < MSE(\hat{\beta}^{OLS}).$$

That is, the mean squared error of the ridge regression estimator is less than than that of the OLS estimator.

3.4.4 Rationale of Correlation Learning

As introduced before, correlation learning uses ranking of marginal correlations between the response and individual predictor to screen important variables. It has a close relationship with

the ridge regression estimator. Let $w^{\lambda} = (w_1^{\lambda}, \ldots, w_p^{\lambda})^T$ be a p-vector that is given by ridge regression

$$w^{\lambda} = (X^T X + \lambda I)^{-1} X^T y$$

where the regularization parameter $\lambda > 0$. According to the size of the regularization parameter, it shows that if $\lambda \to 0$, then the ridge estimator tends to the OLS estimator. When $\lambda \to \infty$, $\lambda\hat{\beta}$ tends to $X^T y$ such that

$$w^{\lambda} \to \hat{\beta}^{OLS} \qquad as\ \lambda \to 0$$

$$\lambda w^{\lambda} \to X^T y \qquad as\ \lambda \to \infty.$$

When the each column of X and the response y are standardized with the mean 0 and unit variance, $\frac{1}{n} X^T y$ becomes the vector of sample correlation coefficients between the individual predictor and the response. This is the rationale of using correlation learning with Pearson correlation as a marginal utility for feature screening. Especially, when both the j^{th} predictor X_j and the response y are standardized, the sample correlation between the j^{th} predictor and the response which is written by

$$w_j = \frac{1}{n} X_j^T y, \quad for\ j = 1, \ldots, p$$

is used as a marginal utility for the importance of each predictor [Fan and Lv, 2008, JingYuan et al., 2015].

3.4.5 Rationale of Clustering Correlated Predictors

As introduced in Chapter 1, high dimensional data in areas such as genomics and health sciences tend to have high level of collinearity and its variance-covariance matrix often possess block-diagonal structure. In the covariance matrix, there are multiple groups where features are strongly correlated among themselves within each group and the features are weakly correlated between groups as in the simplified illustration in Figure 4. Based on this, we describe the insight of conducting feature selection using these clusters of highly correlated predictors rather than using the set of entire variables. We show advantages of utilizing the clusters in terms of colinearity level by comparing the condition number of the block versus the whole covariance matrix.

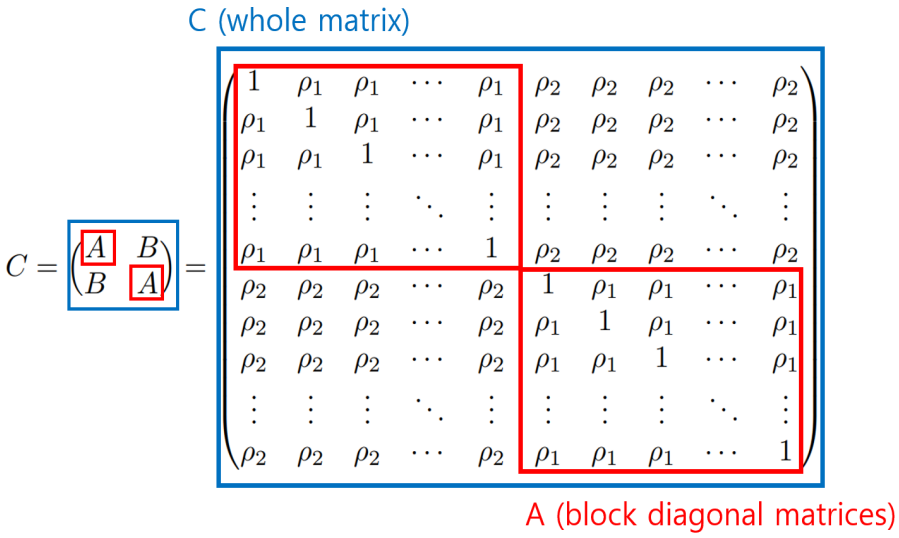


Figure 4: Illustration of block-diagonal variance-covariance matrix

Consider the $k \times k$ block matrix

$$C = \begin{pmatrix} A & B & B & \cdots & B \\ B & A & B & \cdots & B \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ B & B & B & \cdots & A \end{pmatrix} = I_k \otimes (A - B) + \mathbf{1}_k \mathbf{1}_k^{'} \otimes B$$

where $k \geq 2$, $p \geq 1$, A and B are $p \times p$ matrices, and $\mathbf{1}_k$ is a $1 \times k$ vector of all ones. We first find the eigenvalues of the matrix. The determinant of the matrix C can be given by the theorem

$$det(C) = det(A - B)^{k-1} det(A + (k-1)B). \tag{3.1}$$

Using this theorem, we can find the eigenvalues of C with the characteristic equation

$$det(C - \lambda I_k) = det((A - \lambda I_p) - B)^{k-1} det((A - \lambda I_p) + (k-1)B) = 0.$$

From this, the eigenvalues of C are the eigenvalues of the matrices $A - B$ and $A + (k-1)B$.

We assume the structure $p \times p$ matrices A and B such that

$$A = \begin{pmatrix} 1 & \rho_1 & \rho_1 & \cdots & \rho_1 \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_1 \\ \rho_1 & \rho_1 & 1 & \cdots & \rho_1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_1 & \rho_1 & \rho_1 & \cdots & 1 \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} \rho_2 & \rho_2 & \rho_2 & \cdots & \rho_2 \\ \rho_2 & \rho_2 & \rho_2 & \cdots & \rho_2 \\ \rho_2 & \rho_2 & \rho_2 & \cdots & \rho_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_2 & \rho_2 & \rho_2 & \cdots & \rho_2 \end{pmatrix}$$

where $1 > \rho_1 \gg |\rho_2| \geq 0$ is assumed for the structure of block-diagonal covariance structure. If we assume $p = 1$ for the simplest case, the matrix C is given by

$$C = \begin{pmatrix} 1 & \rho_2 \\ \rho_2 & 1 \end{pmatrix}.$$

The eigenvalues of C can be obtained from the characteristic equation from the theorem (3.1)

$$det(C - \lambda I_2) = det(1 - \lambda - \rho_2) det(1 - \lambda + \rho_2).$$

It turns out the eigenvalues $1 - \rho_2$ and $1 + \rho_2$. If we assume $p = 2$, then the matrices $A - B$ and $A + (k-1)B$ are written as

$$A - B = \begin{pmatrix} 1 - \rho_2 & \rho_1 - \rho_2 \\ \rho_1 - \rho_2 & 1 - \rho_2 \end{pmatrix} \quad \text{and} \quad A + (k-1)B = \begin{pmatrix} 1 + (k-1)\rho_2 & \rho_1 + (k-1)\rho_2 \\ \rho_1 + (k-1)\rho_2 & 1 + (k-1)\rho_2 \end{pmatrix}.$$

Using the theorem (3.1), the eigenvalues of two matrices can be obtained from the characteristic equations

$$
\begin{aligned}
&det(A - B - \lambda I_2) \\
&= det((1 - \rho_2 - \lambda) - (\rho_1 - \rho_2))det((1 - \rho_2 - \lambda) + (\rho_1 - \rho_2)) = 0
\end{aligned}
$$

$$
\begin{aligned}
&det(A + (k-1)B - \lambda I_2) \\
&= det((1 + (k-1)\rho_2 - \lambda) - (\rho_1 + (k-1)\rho_2))det((1 + (k-1)\rho_2 - \lambda) + (\rho_1 + (k-1)\rho_2)) = 0.
\end{aligned}
$$

This gives the eigenvalues $1 - \rho_1$ and $1 + \rho_1 - 2\rho_2$ for $A + B$, and $1 - \rho_1$ and $1 + \rho_1 + (k-1)\rho_2$ for $A + (k-1)B$, respectively. For the whole matrix C, it has the same $k - 1$ number of each eigenvalue of $A+B$, and the two eigenvalues of $A+(k-1)B$ from the theorem (3.1). In general when $p \geq 2$, $A + B$ has $p - 1$ number of $1 - \rho_1$ and $1 + (p-1)\rho_1 - p\rho_2$, and $A + (k-1)B$ has $1 - \rho_1$ and $1 + (p-1)\rho_1 + p(k-1)\rho_2$ as their eigenvalues, respectively. In a similar way, we can obtain the eigenvalues for the matrix C as well.

Now we compare the condition numbers for the diagonal matrix A and the whole matrix C. For the condition number which is given by the ratio of the maximum to the minimum eigenvalue is described as

$$
\begin{aligned}
cond(C)_* = \left(\frac{\lambda_{max}}{\lambda_{min}}\right)_C &= \frac{1 + (p-1)\rho_1 + p(k-1)\rho_2}{1 - \rho_1} \\
&\geq \frac{1 + (p-1)\rho_1}{1 - \rho_1} = \left(\frac{\lambda_{max}}{\lambda_{min}}\right)_A = cond(A)_*
\end{aligned}
$$

where equality holds when $\rho_2 = 0$, hence it shows the level of collinearity is higher in the whole matrix than in the block matrix. Moreover, another version of condition number which is the sum of inverse eigenvalues is given by

$$\left(\sum_{j=1}^{kp}\frac{1}{\lambda_j}\right)_C = \frac{k(p-1)}{1-\rho_1} + \frac{k-1}{1+(p-1)\rho_1 - p\rho_2} + \frac{1}{1+(p-1)\rho_1 + p(k-1)\rho_2}$$
$$\geq \frac{k(p-1)}{1-\rho_1} + \frac{k}{1+(p-1)\rho_1} = \left(\sum_{j=1}^{p}\frac{1}{\lambda_j}\right)_A \times k.$$

The condition number of A is multiplied by k since we compare the collinearity level from all k block matrices A with that of the whole matrix C. This can be proven by showing the following

$$\frac{k-1}{1+(p-1)\rho_1 - p\rho_2} + \frac{1}{1+(p-1)\rho_1 + p(k-1)\rho_2} - \frac{k}{1+(p-1)\rho_1} \geq 0.$$

The equality holds when $\rho_2 = 0$. This shows that this version of condition number of the whole matrix C is also greater than the condition number of k number of block matrices A. Otherwise $\rho_2 = 0$, the condition number is greater in the whole matrix than in the block matrix. This gives us insight into advantages of performing feature selection on the cluster that has a subset of predictors rather than on the complete set of features.

3.5 Simulations for Performance Evaluation

We evaluate and compare the performance of the proposed method and other variable selection methods: LASSO, Elastic net, SCAD, MCP, SIS-LASSO, SIS-SCAD, SIS-MCP, Group LASSO, and OSCAR in an extensive set of simulation studies. A variety of high-dimensional

settings are considered in the simulations to create diverse scenarios varying parameter values referring to the literatures [Xie and Zeng, 2010, Bondell and Reich, 2008, Jang et al., 2013, Ročková and George, 2018]. We consider multiple evaluation criteria to measure the performance of model chosen by the methods: prediction, parameter estimation, and selection of variables of true signal.

3.5.1 Data Generation Model

For the simulation studies, the data are generated from the regression model described in Section 3.2

$$y = X\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

where we generate predictors $x_i = (x_{i1}, \ldots, x_{ip})^T$ from a multivariate normal distribution with mean 0 and covariance $\sum$ where $\sum_{j,j} = 1$ for $j = 1, 2, \ldots, p$. The standard deviation of random errors is set as $\sigma = 15$. The number of observations in the data set $n = 150$, the number of predictors in each data set $p = 200$.

We generate various scenarios by differing the parameters in the data generation model: the number of groups of strongly correlated predictors, correlation coefficients in the covariance matrix $\sum$, the size of true coefficient values β varying the level of *signal-to-noise ratio* (SNR), and the sparsity of the true coefficient values. The signal-to-noise ratio is given by

$$SNR = \frac{\beta^T \sum \beta}{\sigma^2}$$

where $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^T$, $\sum$ is the covariance matrix for X, and σ is standard deviation of random errors. When there are q number of groups which are denoted by $G_1, \ldots, G_q$, we assume that predictors are strongly correlated within a group with pairwise correlation $Corr(X_i, X_j) = 0.8$ for $i, j \in G_k, k = 1, \ldots, q$. Otherwise, they are weakly correlated between groups with pairwise correlation $Corr(X_i, X_j) = 0.2$. We describe the details of the scenarios for simulations below.

- **Scenario 1:** There are twenty groups with each group having 10 variables within a group. The data generating regression coefficients are sparsely set as

$$\beta = (5, \underbrace{0, \cdots, 0}_{9}) \qquad \text{for} \quad G_1, \ldots, G_{10}$$

$$\beta = (\underbrace{0, \cdots, 0}_{10}) \qquad \text{for} \quad G_{11}, \ldots, G_{20}.$$

  The SNR is estimated as 3.11.

- **Scenario 2:** There are twenty groups as in Scenario 3. The data generating regression coefficients are smaller and less sparse and are set as

$$\beta = (\underbrace{1, \cdots, 1}_{5}, \underbrace{0, \cdots, 0}_{5}) \qquad \text{for} \quad G_1, \ldots, G_{15}$$

$$\beta = (\underbrace{0, \cdots, 0}_{10}) \qquad \text{for} \quad G_{16}, \ldots, G_{20}.$$

  The SNR is 6.07.

- **Scenario 3:** There are twenty groups. The data generating regression coefficients are less sparse and are set as

$$\beta = (\underbrace{1, \cdots, 1}_{5}, \underbrace{0, \cdots, 0}_{5}) \qquad \text{for} \quad G_1, \ldots, G_{10}$$

$$\beta = (\underbrace{0, \cdots, 0}_{10}) \qquad \text{for} \quad G_{11}, \ldots, G_{20}.$$

  The SNR is 2.93.

- **Scenario 4:** There are five groups of strongly correlated predictors with each group, G_k, $k = 1, \ldots, 5$ having 40 variables within a group. The data generating regression coefficients are set as

$$\beta = (\underbrace{1, \cdots, 1}_{15}, \underbrace{0, \cdots, 0}_{25}) \qquad \text{for} \quad G_1, G_2, G_3$$

$$\beta = (\underbrace{0, \cdots, 0}_{40}) \qquad \text{for} \quad G_4, G_5.$$

  The SNR is estimated as 3.64.

- **Scenario 5:** There is one group of strongly equicorrelated predictors. The data generating regression coefficients are as follows

$$\beta = (5, \underbrace{0, \cdots, 0}_{9}) \times 15 \qquad \text{for} \quad j = 1, \ldots, 150$$

$$\beta = (\underbrace{0, \cdots, 0}_{10}) \times 5 \qquad \text{for} \quad j = 151, \ldots, 200.$$

  The SNR is 20.33.

- **Scenario 6:** There is one group of strongly equicorrelated predictors. The data generating regression coefficients are smaller and less sparse which are given by

$$\beta = (\underbrace{1, \cdots, 1}_{5}, \underbrace{0, \cdots, 0}_{5}) \times 5 \qquad \text{for} \quad j = 1, \ldots, 50$$

$$\beta = (\underbrace{0, \cdots, 0}_{10}) \times 15 \qquad \text{for} \quad j = 151 \ldots, 200.$$

  The SNR is 2.24.

- **Scenario 7:** The data generating regression coefficients are the same as Scenario 1, but here we have one group of strongly equicorrelated predictors, which are as follows

$$\beta = (\underbrace{1, \cdots, 1}_{15}, \underbrace{0, \cdots, 0}_{25}) \times 3 \qquad \text{for} \quad j = 1, \ldots, 120$$

$$\beta = (\underbrace{0, \cdots, 0}_{40}) \times 2 \qquad \text{for} \quad j = 121 \ldots, 200.$$

  The SNR is 7.24.

3.5.2 Evaluation Criteria

3.5.2.1 Prediction

We consider the prediction performance measures described in [Tibshirani, 1996] which are mean squared error (MSE) and prediction error (PE). Suppose that

$$Y = \eta(X) + \epsilon$$

where $E(\epsilon) = 0$ and $var(\epsilon) = \sigma^2$. The definition of mean squared error of an estimate $\hat{\eta}(X)$ is given by

$$MSE = E\{\hat{\eta}(X) - \eta(X)\}^2.$$

A similar measure is prediction error which is defined by

$$PE = E\{Y - \eta(\hat{X})\}^2.$$

The relationship between the two measures can be written as

$$PE = MSE + \sigma^2.$$

We report simulation results with MSE as in the paper [Tibshirani, 1996] which has the form

$$MSE = (\hat{\beta} - \beta)^T V (\hat{\beta} - \beta)$$

for the linear models $\eta(X) = X\hat{\beta}$ in our context where V is the population covariance matrix for X.

3.5.2.2 Parameter Estimation

We also assess the performance for the estimation of the parameters $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^T$ in the regression model. The root mean squared error (RMSE) which is the square root of mean squared error (MSE) is the most widely used measure of estimation accuracy [Li and Zhao, 2001]. The RMSE of the estimator $\hat{\theta}$ is defined by

$$RMSE(\hat{\theta}) = \sqrt{MSE(\hat{\theta})} = \sqrt{E(\hat{\theta} - \theta)^2}.$$

It is known that MSE is the sum of the variance and the squared bias of the estimator [Lebanon, 2010]

$$MSE(\hat{\theta}) = Var(\hat{\theta}) + Bias^2(\hat{\theta}).$$

In the case of multivariate estimators $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_d)^T$, MSE is can be written as

$$E(\|\hat{\theta} - \theta\|^2) = trace(Var(\hat{\theta})) + \|Bias^2(\hat{\theta})\|^2$$

where $Var(\hat{\theta})$ is the covariance matrix of $\hat{\theta}$, so its trace is $\sum_{j=1}^{p} Var(\hat{\theta}_j)$. From this, we measure RMSE for the estimators of the parameters $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p)^T$ in the context of our simulation study which is given by

$$RMSE(\hat{\beta}) = \sqrt{\sum_{j=1}^{p} [Var(\hat{\beta}_j) + \{E(\hat{\beta}_j) - \beta_j\}^2]}.$$

3.5.2.3 Variable Selection

For the evaluation of the performance of variable selection, we adopt the confusion matrix which describes information about true and predicted classes [Chong and Jun, 2005]. Table XXXI shows the confusion matrix and the meanings of its entries under the context of our simulation study.

TABLE II: CONFUSION MATRIX

		Selection Classes	
		Selected predictor	Not-selected predictor
True Classes	Relevant predictor	*True positive* [a]	*False negative* [b]
	Irrelevant predictor	*False positive* [c]	*True negative* [d]

[a] Relevant predictors identified correctly; [b] Relevant predictors identified incorrectly;
[c] Irrelevant predictors identified incorrectly; [d] Irrelevant predictors identified correctly

In our simulation, the predictor x_j is truly relevant when its corresponding true parameter $\beta_j \neq 0$ and it is irrelevant when $\beta_j = 0$ in the data generation model. For the selection classes, the predictor x_j is regarded to be selected when its estimated coefficient $\hat{\beta}_j \neq 0$ and not-selected if $\hat{\beta}_j = 0$ in the chosen model. The ideal variable selection would correctly select predictors of non-zero signals with 100 percent of accuracy.

We find the number of predictors corresponding to each cell of the confusion matrix from the final model chosen by each of the variable selection methods. Then, we assess the measures known as *sensitivity*, *specificity*, *false discovery rate (FDR)*, and *false negative rate (FNR)* which can be obtained by

$$Sensitivity = \frac{Number\ of\ true\ positive}{Number\ of\ true\ positive\ +\ Number\ of\ false\ negative}$$

$$Specificity = \frac{Number\ of\ true\ negative}{Number\ of\ false\ positive\ +\ Number\ of\ true\ negative}$$

$$False\ discovery\ rate = \frac{Number\ of\ false\ positive}{Number\ of\ false\ positive\ +\ Number\ of\ true\ positive}$$

$$False\ negative\ rate = \frac{Number\ of\ false\ negative}{Number\ of\ false\ negative\ +\ Number\ of\ true\ positive}.$$

The *sensitivity* is the proportion of relevant predictors that are correctly selected by a selection method. It is a measure of ability to correctly identify the predictors of non-zero value of parameters. The *specificity* is the proportion of irrelevant predictors that are correctly classified. It shows the ability of a method to correctly identify predictors that are not truly relevant. The *false discovery rate* is the proportion of the predictors that are not truly relevant, but selected in the model. The *false negative rate* is the fraction of the truly relevant predictors that are not selected.

3.5.3 Simulation Results

The COLRNS consistently shows the best prediction performance with the lowest Q1, median, mean, and Q3 values of MSE in all the 7 scenarios. It demonstrates robust prediction performance with the smallest IQR among all methods in Scenarios 1, 2, 3, and 4 where there are multiple groups of strongly correlated predictors and also lower IQR compared to Elastic net in Scenario 5 with one group of equi-correlated predictors of sparse and strong signals.

In Scenarios 1, 2, and 3 with 20 groups, OSCAR presents the worst prediction performance with the highest descriptive statistics of MSE. The Group Lasso also shows the worst level performance together with OSCAR in Scenario 4 having 5 groups with similar high median and Q3 values of prediction error. In Scenarios 1, 2, 3, and 4, SIS-SCAD and SIS-MCP show higher prediction error compared to the same SIS combined method SIS-LASSO. In Scenario 5, 6, and 7 of equi-correlated one group, the Group LASSO didn't perform well with very high prediction error with extremely large variance compared to the other methods.

The COLRNS also depicts good performance in terms of accurate estimation of parameters in all scenarios by having the lowest level of RMSE. It has the smallest RMSE in Scenario 4 with 5 groups and Scenario 5 having one group with sparse and strong signals. In the other scenarios, COLRNS has almost the same RMSE as the Elastic net which indicates comparable performance. In all scenarios except the Scenario 1, COLRNS demonstrates better estimation performance than LASSO.

In Scenarios 1, 2, 3, and 4 having multiple groups, OSCAR gives the worst performance in accurate parameter estimation followed by SIS-SCAD and SIS-MCP. In Scenarios 5, 6, and 7

with 1 group, Group LASSO has the largest error followed by OSCAR. In Scenario 5, MCP also indicates poor performance in estimation together with Group LASSO and OSCAR.

For the selection performance, OSCAR tends to select all predictors in multiple groups regardless of the importance of the predictors in terms of their true signals in all scenarios. The methods even select all the members of the groups where there is no true signal at all, which results in absolute 1 for sensitivity and 0 for specificity and FNR in every simulated data sets.

Group LASSO chooses all predictors within a group if there is any member with true signal, which turns out perfect sensitivity 1 without an outlier and 0 for specificity and FNR in Scenarios 1, 2, 3, and 4 with multiple groups. In one group scenarios, it shows erratic selection performance by choosing all or none of 200 predictors, which leads excessively high variance for all of the four selection criteria.

Excluding the two methods of OSCAR and Group LASSO which demonstrate poor selection performances, COLRNS yields the highest Q1, median, mean, and Q3 of sensitivity in comparison to the other methods in Scenario 1 and 5 where predictors are sparse with strong signals. In other scenarios, it gives good sensitivity overall by having higher Q1 and median sensitivity with smaller IQR than Elastic net in multiple group scenarios, and better median and Q3 sensitivity in one group scenarios. It can be seen that COLRNS effectively selects predictors with true signal.

The COLRNS also improves specificity in multiple group scenarios where there are many predictors of small signal. More specifically, in comparison to Elastic net, it has higher Q1, median, and mean in Scenario 2, better Q1 in Scenario 3, and superior Q1 and mean specificity

in Scenario 4. In overall, it improves robustness in ability of filtering out unimportant variables with smaller IQR level than Elastic net when there are multiple groups of strongly correlated predictors.

For FNR, COLRNS shows the lowest level of FNR overall. It performs the best with the lowest Q1, median, mean, and Q3 of FNR when there are predictors of sparse and strong signals in Scenario 1 and 5. In other multiple group scenarios with numerous small signals, it also gives lower Q3 FNR than Elastic net in Scenario 2, 3, and 4 and median as well in Scenario 3. In other one group scenarios with many weak signals, it turns out to have lower median and mean in Scenarios 6 and 7 and Q1 as well in Scenario 6. Based on the results of the improved FNR, it can be viewed that COLRNS has lower chance of missing out important predictors. For FDR, COLRNS has comparable performance across the scenarios.

TABLE III: MSE OF PREDICTION IN SCENARIO 1

	COLRNS	LASSO	ELNET	SCAD	MCP	S-SCAD	S-MCP	S-LASSO	G-LASSO	OSCAR
Q1	61.2	73.8	74.3	80.1	77.6	160.4	162.5	65.6	120.8	245.5
Median	69.2	90.9	86.9	94.3	96.5	19.01	200.6	87.7	134.3	278.0
Mean	76.6	95.6	93.0	100.1	100.3	190.7	194.6	90.5	137.5	280.4
Q3	86.9	107.1	105.0	117.2	118.5	223.7	229.8	109.0	149.6	311.7

TABLE IV: RMSE OF PARAMETER ESTIMATION IN SCENARIO 1

	COLRNS	LASSO	ELNET	SCAD	MCP	S-SCAD	S-MCP	S-LASSO	G-LASSO	OSCAR
RMSE	13.75	13.36	13.42	16.15	17.63	15.60	27.00	27.30	19.48	33.23

TABLE V: VARIABLE SELECTION MEASUREMENTS IN SCENARIO 1

	COLRNS	LASSO	ELNET	SCAD	MCP	S-SCAD	S-MCP	S-LASSO	G-LASSO	OSCAR
					Sensitivity					
Q1	0.800	0.600	0.700	0.500	0.300	0.500	0.500	0.500	1.000	1.000
Median	0.800	0.700	0.800	0.600	0.500	0.600	0.600	0.600	1.000	1.000
Mean	0.829	0.726	0.804	0.608	0.460	0.587	0.574	0.610	0.992	1.000
Q3	0.900	0.800	0.900	0.700	0.600	0.700	0.700	0.700	1.000	1.000
					Specificity					
Q1	0.774	0.889	0.784	0.914	0.947	0.874	0.874	0.888	0.105	0.000
Median	0.800	0.905	0.858	0.926	0.958	0.879	0.879	0.921	0.184	0.000
Mean	0.806	0.902	0.826	0.927	0.955	0.886	0.887	0.914	0.191	0.000
Q3	0.858	0.921	0.889	0.942	0.963	0.884	0.884	0.937	0.263	0.000
					False Discovery Rate					
Q1	0.770	0.667	0.737	0.624	0.538	0.759	0.759	0.667	0.933	0.950
Median	0.819	0.708	0.794	0.688	0.643	0.793	0.793	0.724	0.941	0.950
Mean	0.803	0.712	0.783	0.687	0.642	0.777	0.774	0.717	0.939	0.950
Q3	0.843	0.769	0.834	0.750	0.750	0.828	0.828	0.783	0.944	0.950
					False Negative Rate					
Q1	0.100	0.200	0.100	0.300	0.400	0.300	0.300	0.300	0.000	0.000
Median	0.200	0.300	0.200	0.400	0.500	0.400	0.400	0.400	0.000	0.000
Mean	0.171	0.274	0.196	0.392	0.540	0.413	0.426	0.390	0.008	0.000
Q3	0.200	0.400	0.300	0.500	0.700	0.500	0.500	0.500	0.000	0.000

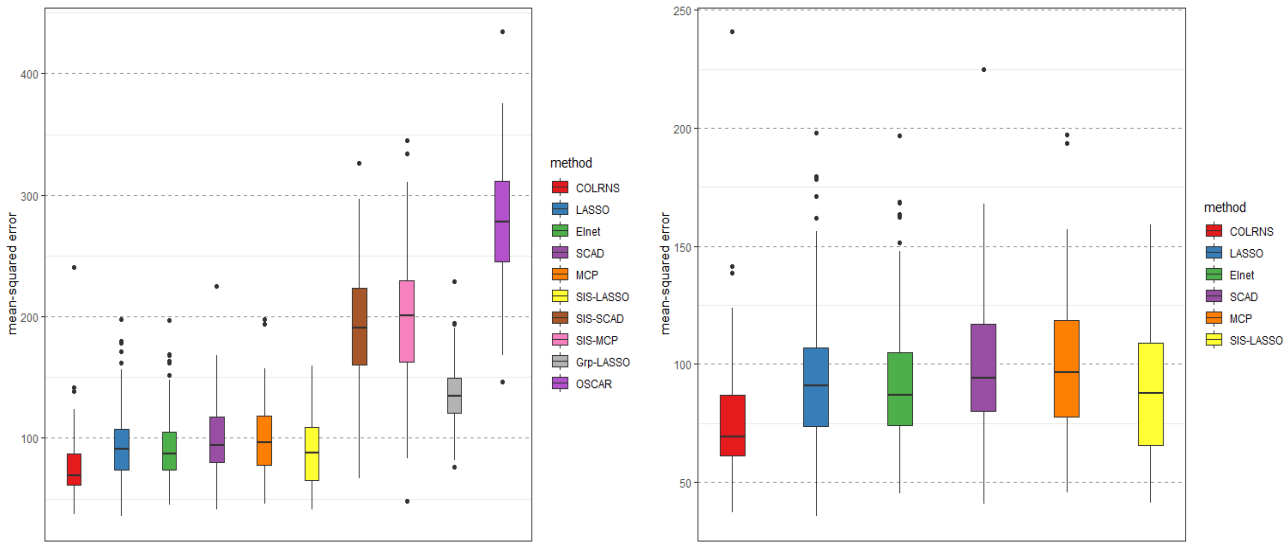


Figure 5: MSE of prediction in scenario 1

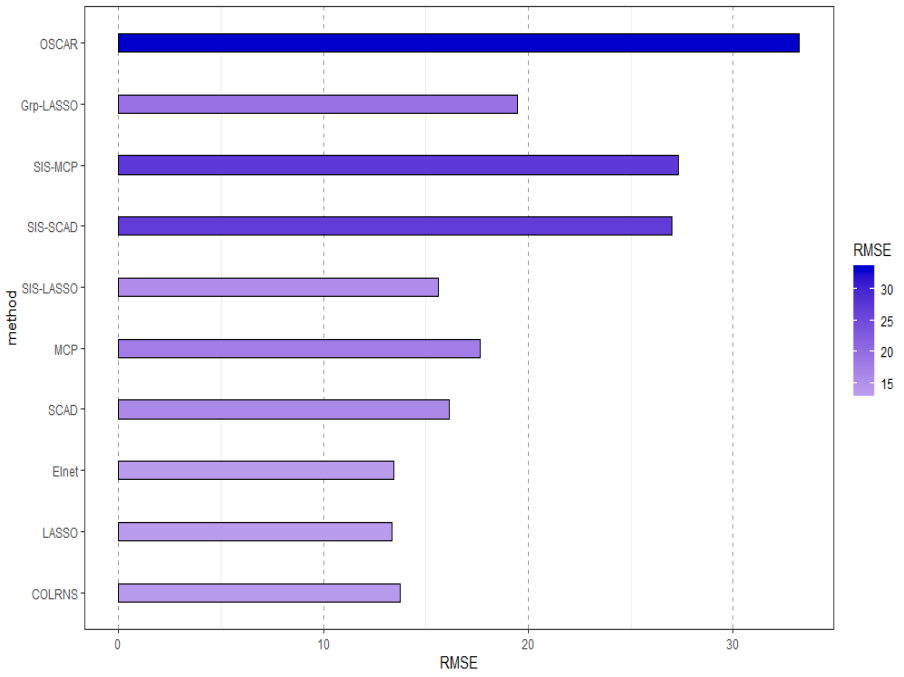
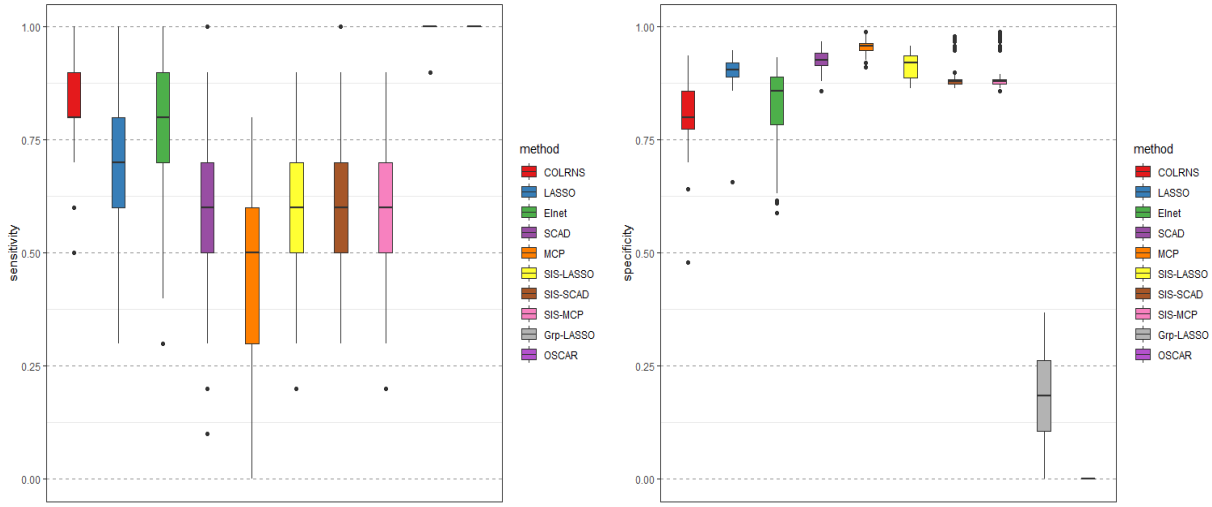
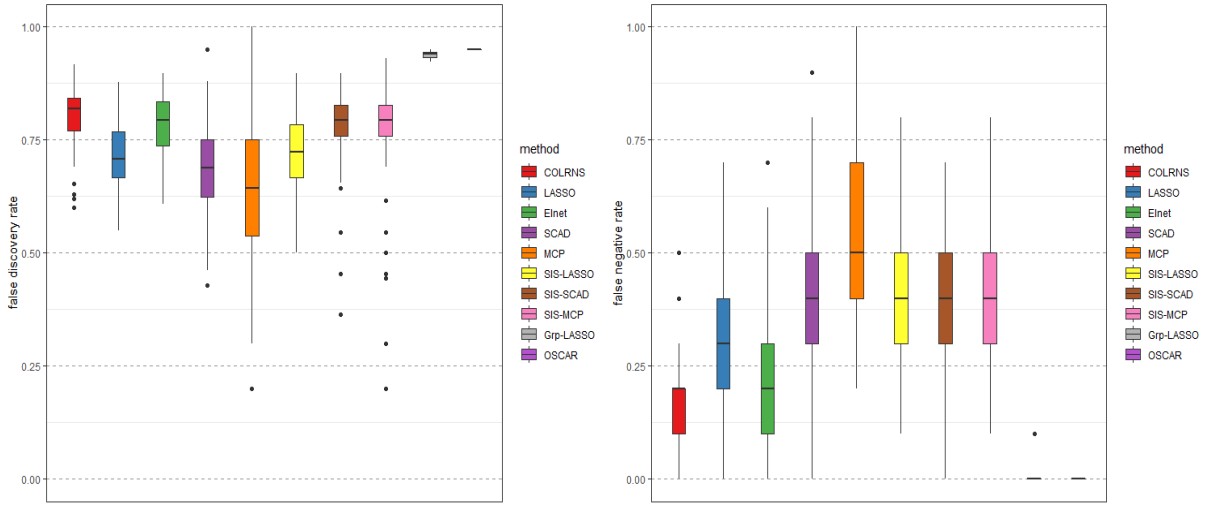


Figure 6: RMSE of parameter estimation in scenario 1



(a) Sensitivity

(b) Specificity



(c) False discovery rate

(d) False negative rate

Figure 7: Variable selection measurements in scenario 1

TABLE VI: MSE OF PREDICTION IN SCENARIO 2

	COLRNS	LASSO	ELNET	SCAD	MCP	S-SCAD	S-MCP	S-LASSO	G-LASSO	OSCAR
Q1	62.4	79.4	67.5	126.0	120.4	183.0	189.1	116.0	154.8	235.0
Median	74.7	95.7	82.2	145.7	132.4	206.2	219.4	142.9	170.9	260.9
Mean	76.6	101.2	85.5	149.6	139.1	216.7	225.4	144.2	173.9	267.5
Q3	86.4	120.3	99.2	169.4	152.2	246.7	260.1	164.3	190.7	288.9

TABLE VII: RMSE OF PARAMETER ESTIMATION IN SCENARIO 2

	COLRNS	LASSO	ELNET	SCAD	MCP	S-SCAD	S-MCP	S-LASSO	G-LASSO	OSCAR
RMSE	10.160	12.530	9.900	19.630	20.530	16.080	28.00	28.77	21.79	32.49

TABLE VIII: VARIABLE SELECTION MEASUREMENTS IN SCENARIO 2

	COLRNS	LASSO	ELNET	SCAD	MCP	S-SCAD	S-MCP	S-LASSO	G-LASSO	OSCAR
					Sensitivity					
Q1	0.427	0.280	0.397	0.130	0.107	0.173	0.173	0.173	1.000	1.000
Median	0.480	0.313	0.560	0.153	0.120	0.187	0.200	0.187	1.000	1.000
Mean	0.473	0.311	0.538	0.173	0.121	0.193	0.193	0.190	0.995	1.000
Q3	0.520	0.347	0.680	0.213	0.147	0.213	0.213	0.213	1.000	1.000
					Specificity					
Q1	0.688	0.816	0.598	0.888	0.928	0.872	0.872	0.896	0.080	0.000
Median	0.724	0.848	0.676	0.912	0.936	0.880	0.888	0.908	0.080	0.000
Mean	0.725	0.842	0.693	0.905	0.938	0.885	0.886	0.910	0.094	0.000
Q3	0.768	0.866	0.784	0.928	0.952	0.896	0.904	0.928	0.160	0.000
					False Discovery Rate					
Q1	0.448	0.403	0.457	0.400	0.373	0.448	0.447	0.375	0.583	0.625
Median	0.489	0.456	0.491	0.476	0.456	0.517	0.483	0.435	0.605	0.625
Mean	0.489	0.456	0.481	0.483	0.459	0.498	0.496	0.439	0.602	0.625
Q3	0.519	0.500	0.511	0.550	0.551	0.552	0.552	0.500	0.611	0.625
					False Negative Rate					
Q1	0.480	0.653	0.320	0.787	0.853	0.787	0.787	0.787	0.000	0.000
Median	0.520	0.687	0.440	0.847	0.880	0.813	0.800	0.813	0.000	0.000
Mean	0.527	0.689	0.462	0.827	0.879	0.807	0.807	0.810	0.005	0.000
Q3	0.573	0.720	0.603	0.870	0.893	0.827	0.827	0.827	0.000	0.000

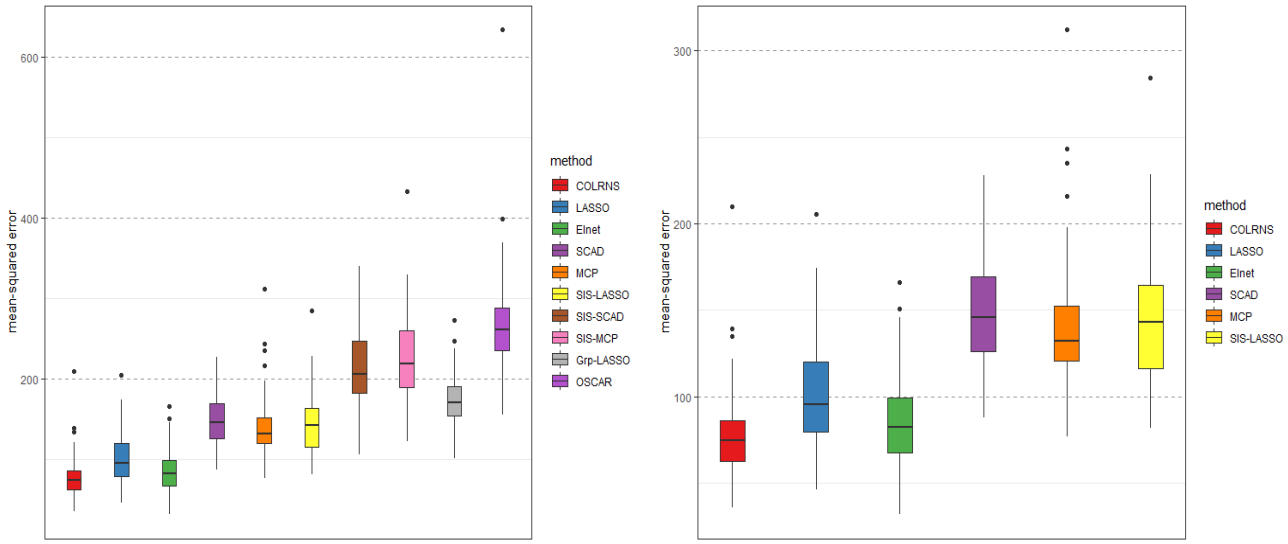


Figure 8: MSE of prediction in scenario 2

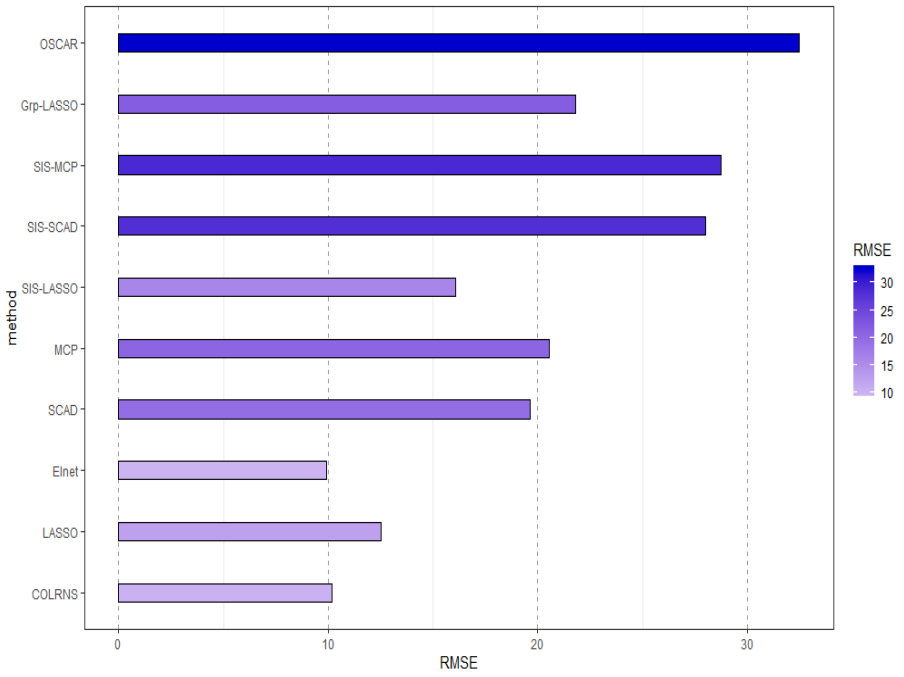
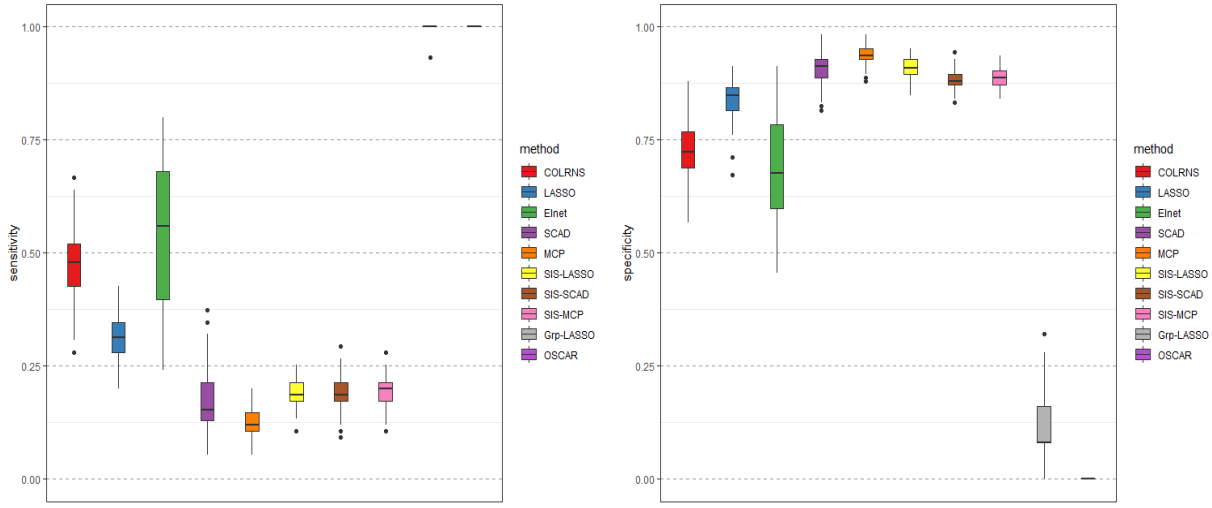
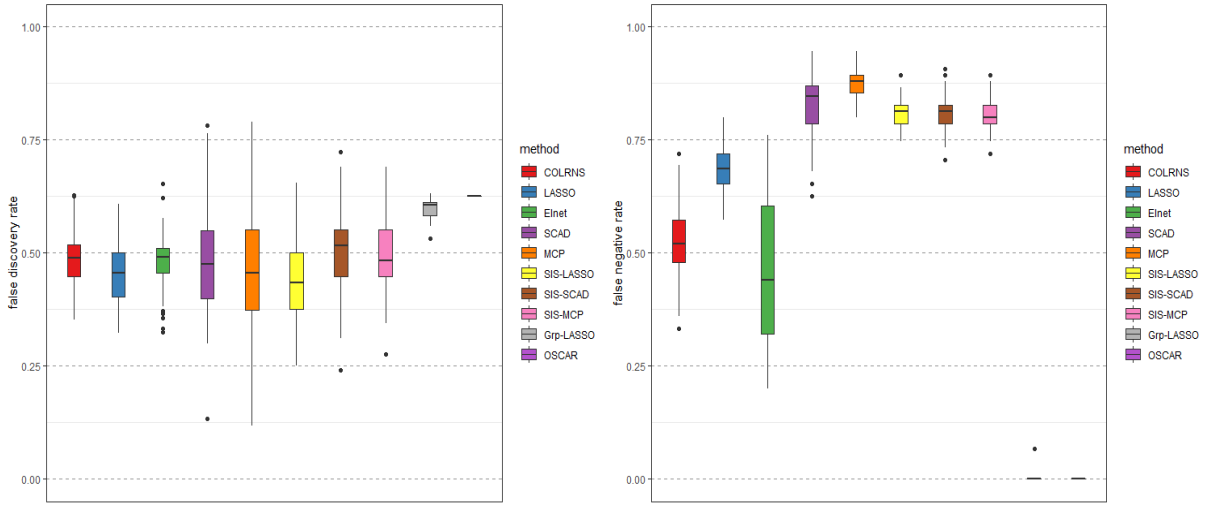


Figure 9: RMSE of parameter estimation in scenario 2



(a) Sensitivity

(b) Specificity



(c) False discovery rate

(d) False negative rate

Figure 10: Variable selection measurements in scenario 2

TABLE IX: MSE OF PREDICTION IN SCENARIO 3

	COLRNS	LASSO	ELNET	SCAD	MCP	S-SCAD	S-MCP	S-LASSO	G-LASSO	OSCAR
Q1	47.6	65.4	57.9	77.9	75.2	161.7	152.7	60.6	106.4	261.8
Median	56.4	76.8	66.9	91.0	86.7	183.8	190.6	76.6	119.2	261.1
Mean	60.2	80.2	71.9	95.0	90.3	188.3	189.2	79.2	121.0	237.8
Q3	65.2	91.2	83.4	108.3	98.1	214.5	223.1	92.5	133.4	287.7

TABLE X: RMSE OF PARAMETER ESTIMATION IN SCENARIO 3

	COLRNS	LASSO	ELNET	SCAD	MCP	S-SCAD	S-MCP	S-LASSO	G-LASSO	OSCAR
RMSE	8.54	9.71	8.08	14.4	16.1	13.34	26.78	26.69	16.72	32.02

TABLE XI: VARIABLE SELECTION MEASUREMENTS IN SCENARIO 3

	COLRNS	LASSO	ELNET	SCAD	MCP	S-SCAD	S-MCP	S-LASSO	G-LASSO	OSCAR
					Sensitivity					
Q1	0.415	0.280	0.360	0.140	0.100	0.200	0.180	0.200	1.000	1.000
Median	0.480	0.300	0.440	0.210	0.120	0.240	0.220	0.240	1.000	1.000
Mean	0.474	0.298	0.499	0.212	0.121	0.229	0.212	0.238	0.996	1.000
Q3	0.540	0.340	0.640	0.280	0.140	0.260	0.260	0.280	1.000	1.000
					Specificity					
Q1	0.780	0.893	0.740	0.907	0.940	0.873	0.873	0.913	0.133	0.000
Median	0.807	0.907	0.827	0.920	0.953	0.887	0.887	0.927	0.200	0.000
Mean	0.802	0.903	0.809	0.920	0.953	0.888	0.894	0.926	0.235	0.000
Q3	0.833	0.920	0.880	0.933	0.967	0.900	0.900	0.947	0.333	0.000
					False Discovery Rate					
Q1	0.518	0.439	0.486	0.464	0.417	0.517	0.517	0.400	0.667	0.750
Median	0.550	0.482	0.531	0.525	0.500	0.586	0.586	0.480	0.706	0.750
Mean	0.555	0.490	0.524	0.542	0.526	0.592	0.591	0.478	0.694	0.750
Q3	0.592	0.548	0.563	0.637	0.643	0.655	0.655	0.550	0.722	0.750
					False Negative Rate					
Q1	0.460	0.660	0.360	0.720	0.860	0.740	0.740	0.720	0.000	0.000
Median	0.520	0.700	0.560	0.790	0.880	0.760	0.780	0.760	0.000	0.000
Mean	0.526	0.702	0.501	0.788	0.879	0.771	0.788	0.762	0.004	0.000
Q3	0.585	0.720	0.640	0.860	0.900	0.800	0.820	0.800	0.000	0.000

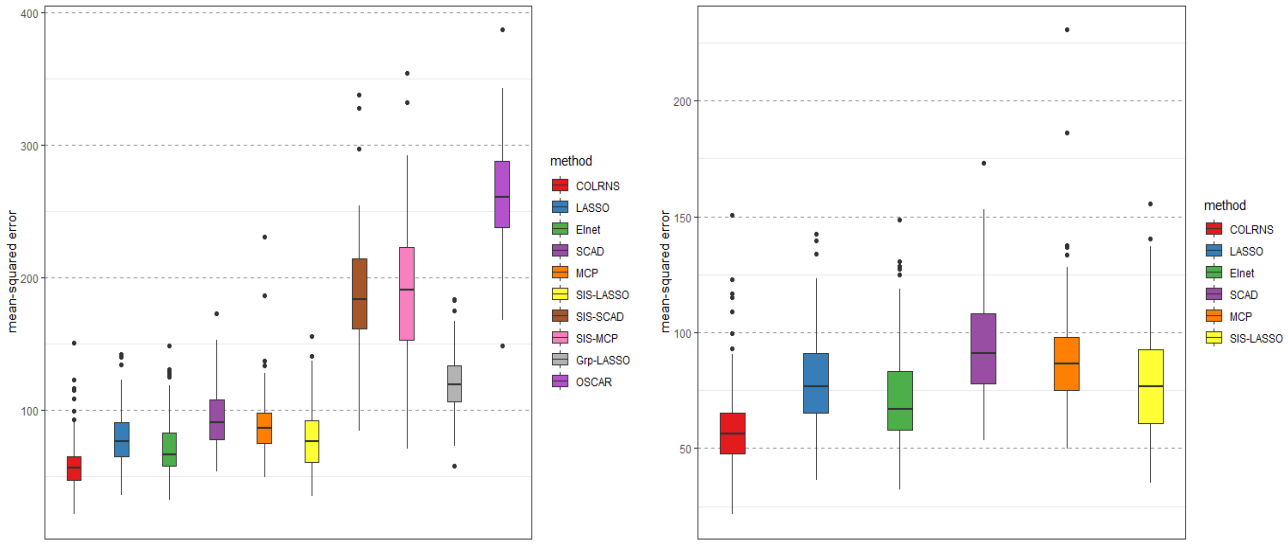


Figure 11: MSE of prediction in scenario 3

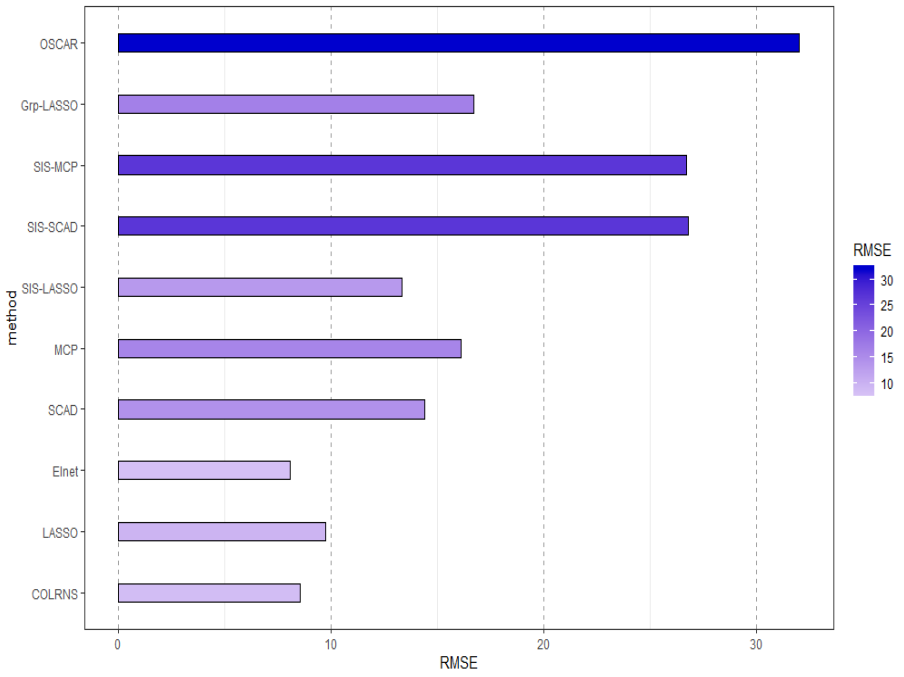
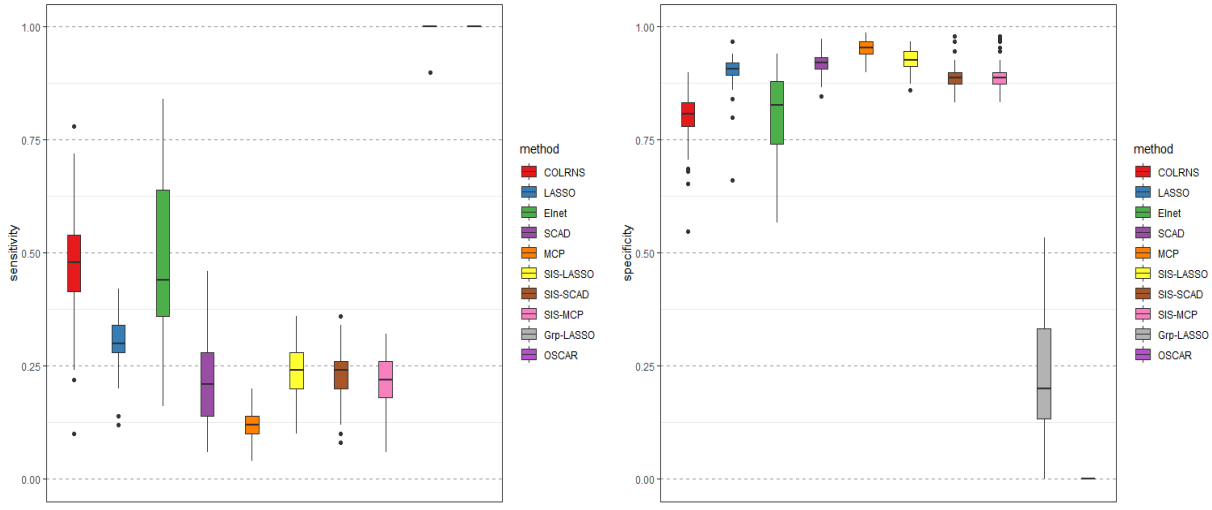
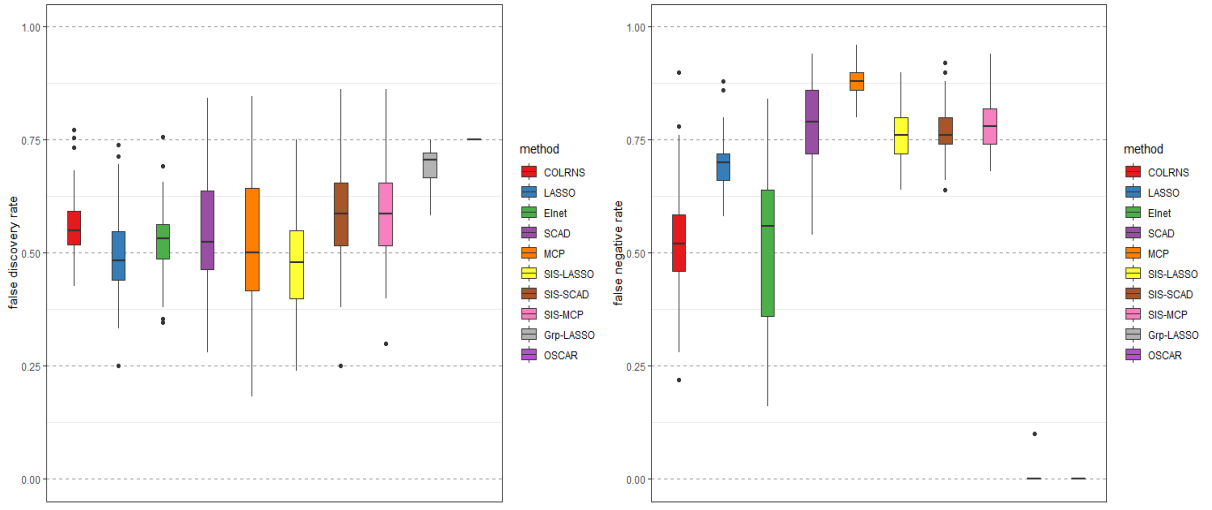


Figure 12: RMSE of parameter estimation in scenario 3



(a) Sensitivity

(b) Specificity



(c) False discovery rate

(d) False negative rate

Figure 13: Variable selection measurements in scenario 3

TABLE XII: MSE OF PREDICTION IN SCENARIO 4

	COLRNS	LASSO	ELNET	SCAD	MCP	S-SCAD	S-MCP	S-LASSO	G-LASSO	OSCAR
Q1	32.1	42.2	39.5	75.5	69.9	145.8	144.8	40.7	218.9	215.2
Median	38.1	54.0	48.0	86.3	86.9	173.2	175.2	50.7	230.1	235.6
Mean	44.0	58.5	52.7	91.2	91.5	170.1	173.3	55.7	240.3	239.9
Q3	52.4	66.5	61.1	106.2	111.3	195.3	203.5	69.2	259.6	263.4

TABLE XIII: RMSE OF PARAMETER ESTIMATION IN SCENARIO 4

	COLRNS	LASSO	ELNET	SCAD	MCP	S-SCAD	S-MCP	S-LASSO	G-LASSO	OSCAR
RMSE	7.74	10.15	7.88	17.76	20.13	14.49	28.51	28.41	25.29	33.69

TABLE XIV: VARIABLE SELECTION MEASUREMENTS IN SCENARIO 4

	COLRNS	LASSO	ELNET	SCAD	MCP	S-SCAD	S-MCP	S-LASSO	G-LASSO	OSCAR
					Sensitivity					
Q1	0.394	0.222	0.333	0.067	0.044	0.200	0.172	0.178	1.000	1.000
Median	0.467	0.244	0.467	0.111	0.067	0.244	0.222	0.222	1.000	1.000
Mean	0.464	0.256	0.500	0.138	0.063	0.236	0.209	0.227	1.000	1.000
Q3	0.556	0.289	0.689	0.200	0.089	0.289	0.267	0.267	1.000	1.000
					Specificity					
Q1	0.785	0.903	0.715	0.935	0.968	0.877	0.877	0.897	0.000	0.000
Median	0.816	0.916	0.816	0.948	0.974	0.890	0.884	0.929	0.000	0.000
Mean	0.814	0.918	0.803	0.950	0.976	0.894	0.904	0.921	0.080	0.000
Q3	0.847	0.935	0.890	0.963	0.987	0.897	0.910	0.942	0.258	0.000
					False Discovery Rate					
Q1	0.535	0.471	0.537	0.475	0.400	0.552	0.517	0.450	0.719	0.775
Median	0.576	0.523	0.575	0.556	0.571	0.586	0.621	0.552	0.775	0.775
Mean	0.574	0.522	0.562	0.560	0.543	0.593	0.577	0.538	0.756	0.775
Q3	0.609	0.583	0.598	0.667	0.714	0.655	0.655	0.621	0.775	0.775
					False Negative Rate					
Q1	0.444	0.711	0.311	0.800	0.911	0.711	0.733	0.733	0.000	0.000
Median	0.533	0.756	0.533	0.889	0.933	0.756	0.778	0.778	0.000	0.000
Mean	0.536	0.744	0.500	0.862	0.937	0.764	0.791	0.773	0.000	0.000
Q3	0.606	0.778	0.667	0.933	0.956	0.800	0.828	0.822	0.000	0.000

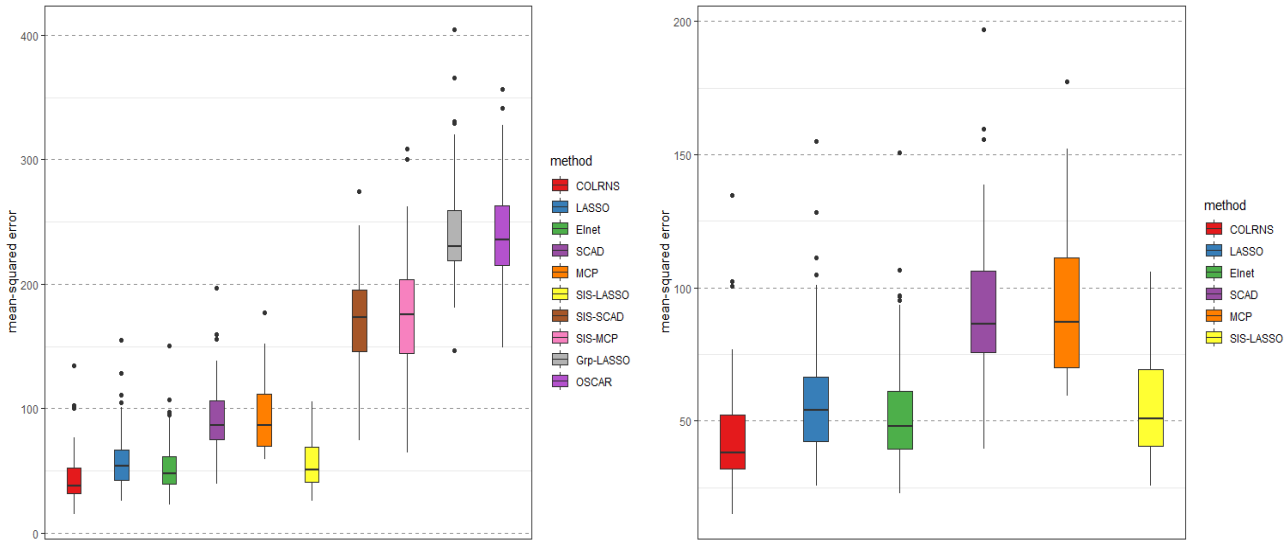


Figure 14: MSE of prediction in scenario 4

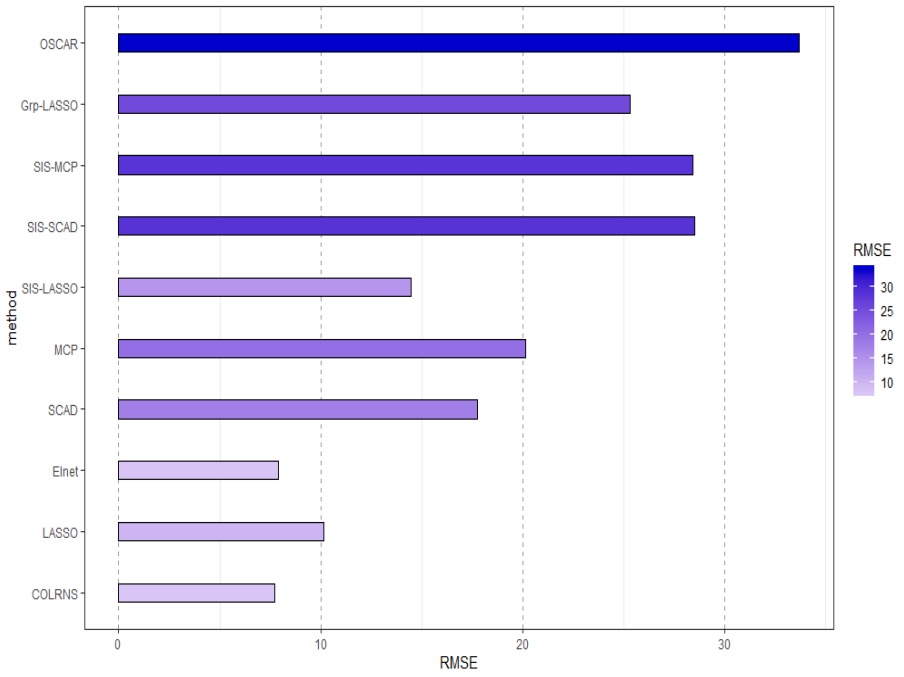
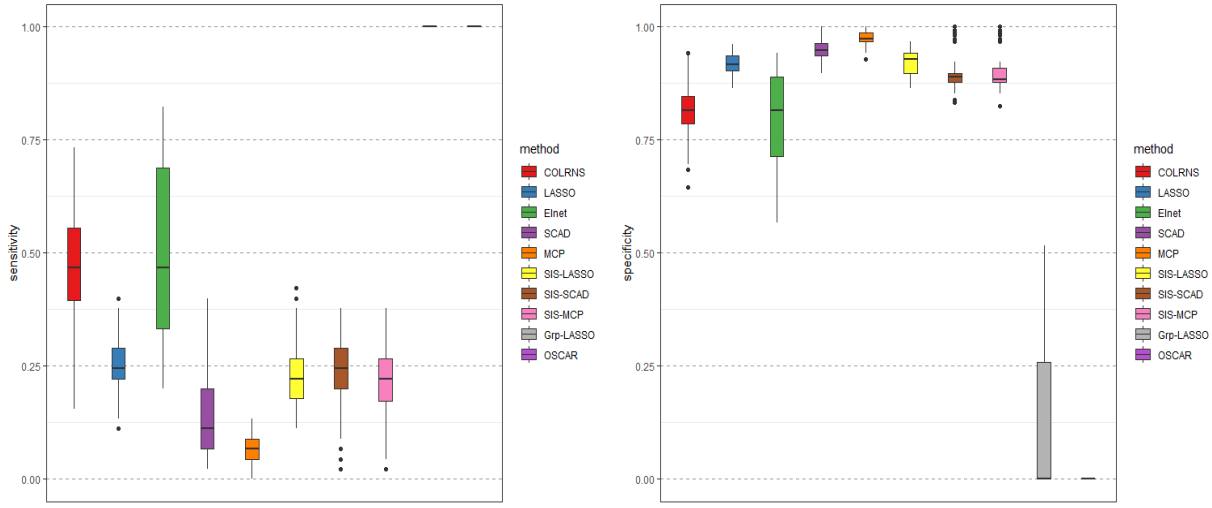
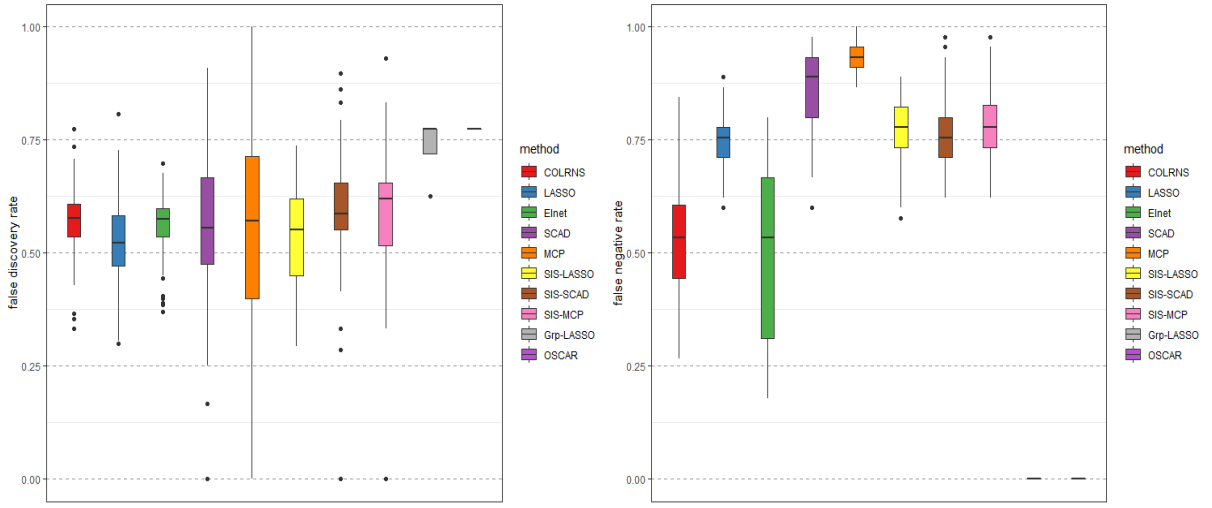


Figure 15: RMSE of parameter estimation in scenario 4



(a) Sensitivity

(b) Specificity



(c) False discovery rate

(d) False negative rate

Figure 16: Variable selection measurements in scenario 4

TABLE XV: MSE OF PREDICTION IN SCENARIO 5

	COLRNS	LASSO	ELNET	SCAD	MCP	S-SCAD	S-MCP	S-LASSO	G-LASSO	OSCAR
Q1	67.2	80.3	76.2	133.7	226.9	183.4	192.7	95.3	571.0	234.4
Median	78.4	90.2	89.5	161.9	247.6	224.6	224.8	115.0	4575.0	254.2
Mean	80.5	94.5	90.8	174.5	254.9	220.2	225.6	118.4	2707.3	269.0
Q3	91.5	104.9	106.3	189.2	279.1	248.8	252.0	141.7	4575.0	288.5

TABLE XVI: RMSE OF PARAMETER ESTIMATION IN SCENARIO 5

	COLRNS	LASSO	ELNET	SCAD	MCP	S-SCAD	S-MCP	S-LASSO	G-LASSO	OSCAR
RMSE	16.56	17.52	16.67	21.40	35.47	23.91	33.11	33.55	39.51	36.68

TABLE XVII: VARIABLE SELECTION MEASUREMENTS IN SCENARIO 5

	COLRNS	LASSO	ELNET	SCAD	MCP	S-SCAD	S-MCP	S-LASSO	G-LASSO	OSCAR
					Sensitivity					
Q1	0.800	0.600	0.733	0.533	0.133	0.400	0.400	0.400	0.000	1.000
Median	0.867	0.667	0.800	0.667	0.133	0.467	0.467	0.467	1.000	1.000
Mean	0.853	0.686	0.829	0.601	0.149	0.475	0.476	0.487	0.640	1.000
Q3	0.933	0.733	0.933	0.733	0.200	0.533	0.550	0.600	1.000	1.000
					Specificity					
Q1	0.524	0.854	0.561	0.870	0.984	0.876	0.876	0.880	0.000	0.000
Median	0.662	0.870	0.754	0.886	0.989	0.881	0.881	0.886	0.000	0.000
Mean	0.661	0.867	0.695	0.893	0.988	0.882	0.883	0.889	0.360	0.000
Q3	0.811	0.881	0.832	0.908	0.991	0.886	0.892	0.892	1.000	0.000
					False Discovery Rate					
Q1	0.749	0.664	0.734	0.630	0.312	0.724	0.690	0.690	0.000	0.925
Median	0.831	0.694	0.798	0.677	0.500	0.759	0.759	0.724	0.925	0.925
Mean	0.805	0.700	0.791	0.668	0.492	0.754	0.750	0.733	0.592	0.925
Q3	0.864	0.750	0.859	0.738	0.600	0.793	0.793	0.793	0.925	0.925
					False Negative Rate					
Q1	0.067	0.267	0.067	0.267	0.800	0.467	0.450	0.400	0.000	0.000
Median	0.133	0.333	0.200	0.333	0.867	0.533	0.533	0.533	0.000	0.000
Mean	0.147	0.314	0.171	0.399	0.851	0.525	0.524	0.513	0.360	0.000
Q3	0.200	0.400	0.267	0.467	0.867	0.600	0.600	0.600	1.000	0.000

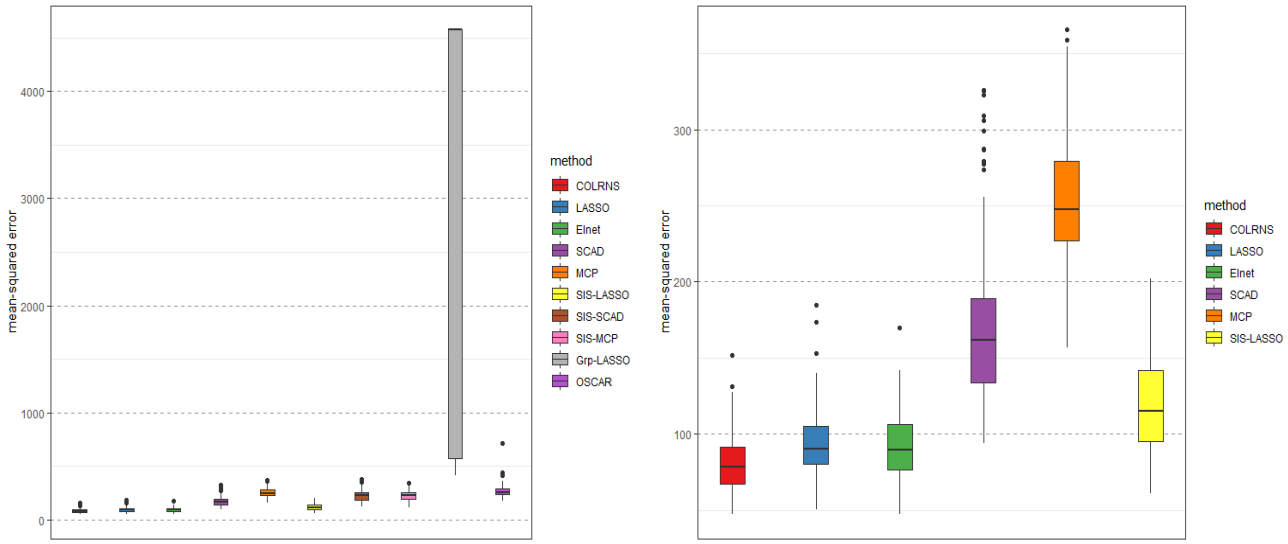


Figure 17: MSE of prediction in scenario 5

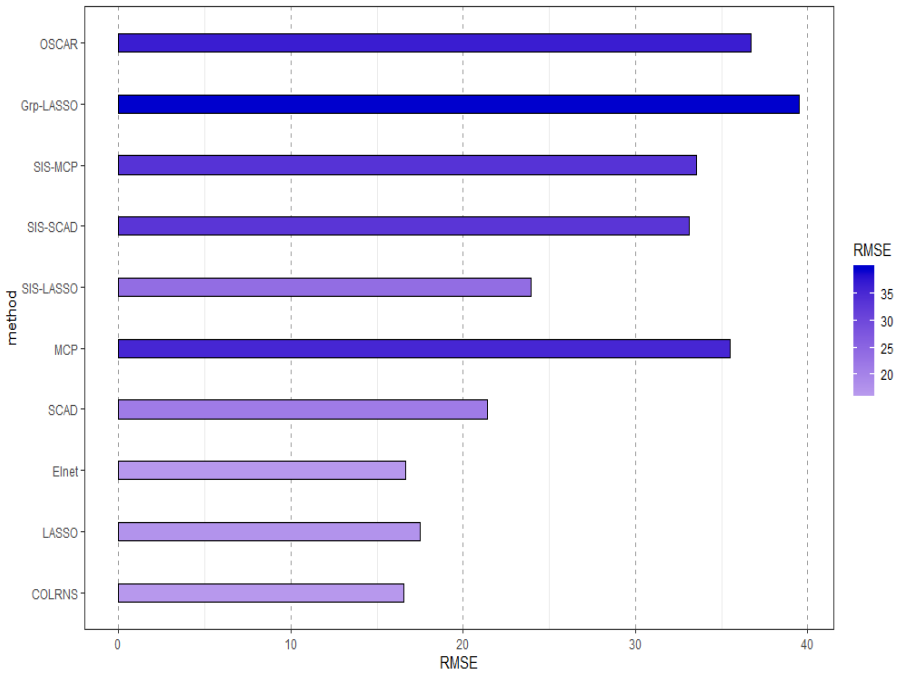
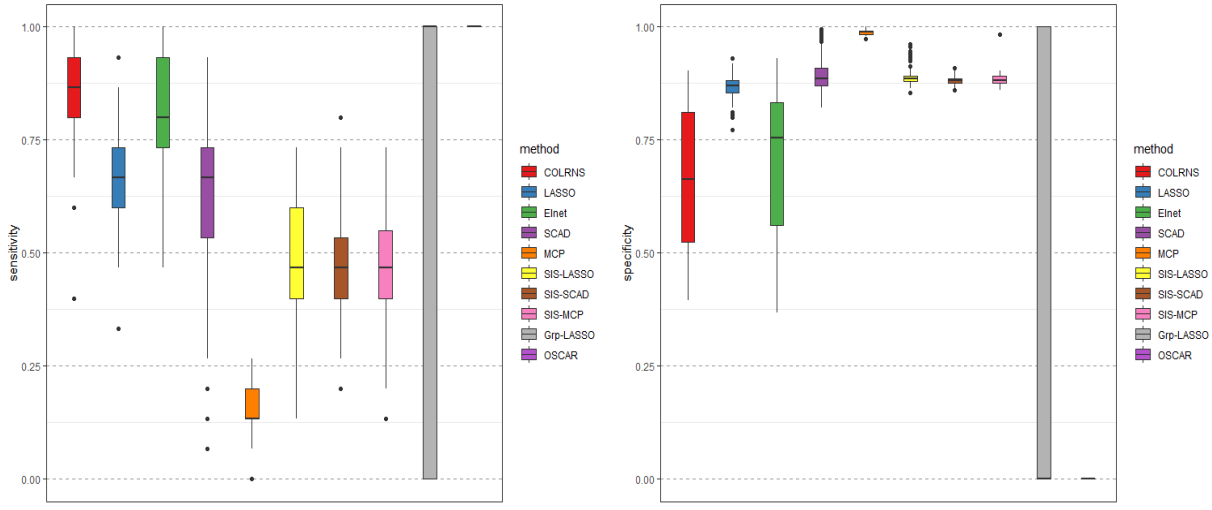
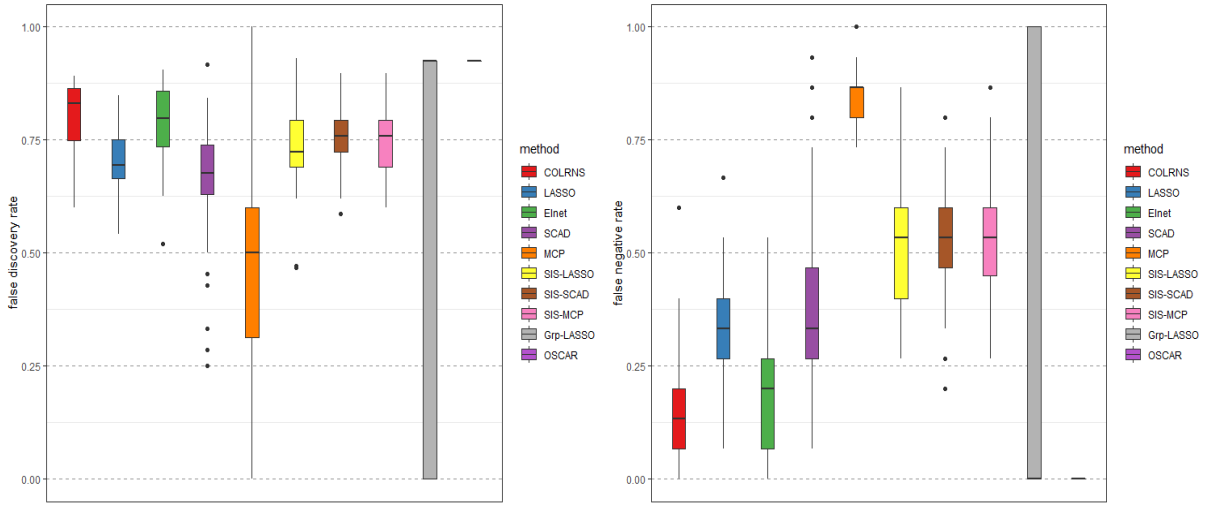


Figure 18: RMSE of parameter estimation in scenario 5



(a) Sensitivity

(b) Specificity



(c) False discovery rate

(d) False negative rate

Figure 19: Variable selection measurements in scenario 5

TABLE XVIII: MSE OF PREDICTION IN SCENARIO 6

	COLRNS	LASSO	ELNET	SCAD	MCP	S-SCAD	S-MCP	S-LASSO	G-LASSO	OSCAR
Q1	21.0	31.7	27.6	31.7	43.6	139.9	116.5	45.4	505.0	203.1
Median	31.4	40.0	36.0	47.9	48.0	171.9	178.9	70.9	505.0	227.3
Mean	32.9	43.5	38.2	48.9	56.5	172.6	165.2	63.9	526.1	228.6
Q3	40.4	53.5	43.2	63.2	63.9	210.8	205.4	78.5	512.8	250.6

TABLE XIX: RMSE OF PARAMETER ESTIMATION IN SCENARIO 6

	COLRNS	LASSO	ELNET	SCAD	MCP	S-SCAD	S-MCP	S-LASSO	G-LASSO	OSCAR
RMSE	6.10	7.61	6.01	13.19	16.53	17.43	29.15	28.38	35.92	33.80

TABLE XX: VARIABLE SELECTION MEASUREMENTS IN SCENARIO 6

	COLRNS	LASSO	ELNET	SCAD	MCP	S-SCAD	S-MCP	S-LASSO	G-LASSO	OSCAR
					Sensitivity					
Q1	0.200	0.110	0.200	0.040	0.000	0.120	0.040	0.120	0.000	1.000
Median	0.360	0.160	0.320	0.120	0.040	0.160	0.160	0.160	1.000	1.000
Mean	0.368	0.141	0.342	0.110	0.032	0.165	0.148	0.164	0.700	1.000
Q3	0.520	0.160	0.480	0.160	0.040	0.200	0.200	0.200	1.000	1.000
					Specificity					
Q1	0.606	0.920	0.634	0.930	0.983	0.851	0.857	0.857	0.000	0.000
Median	0.749	0.931	0.803	0.951	0.989	0.863	0.863	0.863	0.000	0.000
Mean	0.746	0.930	0.775	0.951	0.985	0.874	0.894	0.880	0.300	0.000
Q3	0.903	0.943	0.904	0.977	0.989	0.869	0.914	0.881	1.000	0.000
					False Discovery Rate					
Q1	0.793	0.714	0.783	0.667	0.667	0.793	0.784	0.793	0.000	0.875
Median	0.822	0.782	0.816	0.750	0.750	0.828	0.828	0.828	0.875	0.875
Mean	0.815	0.774	0.806	0.758	0.762	0.825	0.791	0.829	0.612	0.875
Q3	0.857	0.833	0.851	0.833	1.000	0.897	0.897	0.862	0.875	0.875
					False Negative Rate					
Q1	0.480	0.840	0.520	0.840	0.960	0.800	0.800	0.800	0.000	0.000
Median	0.640	0.840	0.680	0.880	0.960	0.840	0.840	0.840	0.000	0.000
Mean	0.632	0.859	0.658	0.890	0.968	0.835	0.852	0.836	0.300	0.000
Q3	0.800	0.890	0.800	0.960	1.000	0.880	0.960	0.880	1.000	0.000

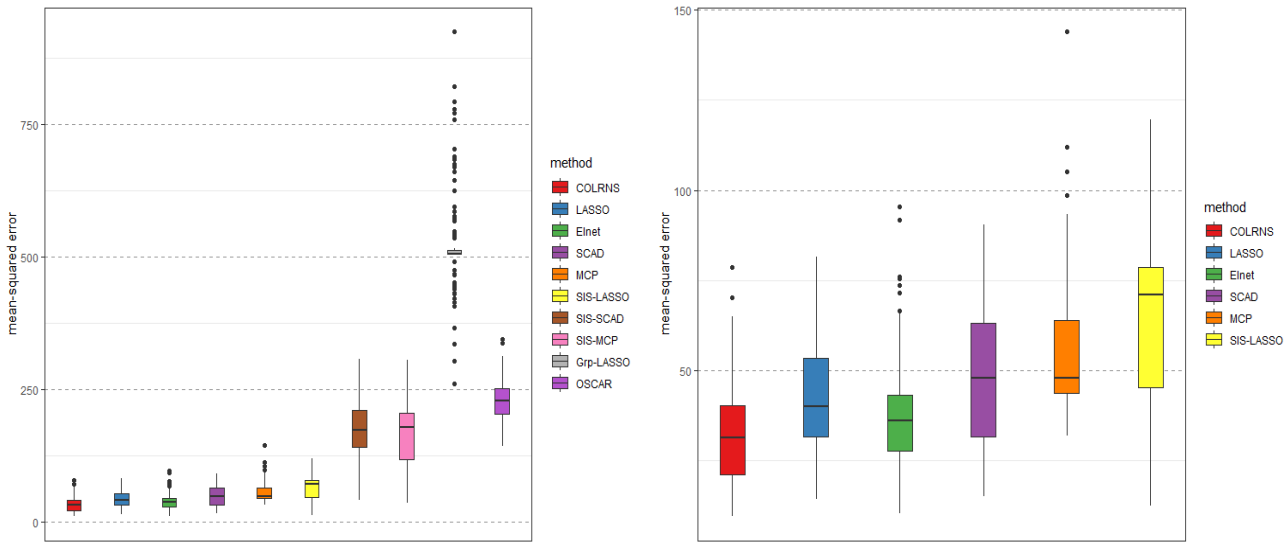


Figure 20: MSE of prediction in scenario 6

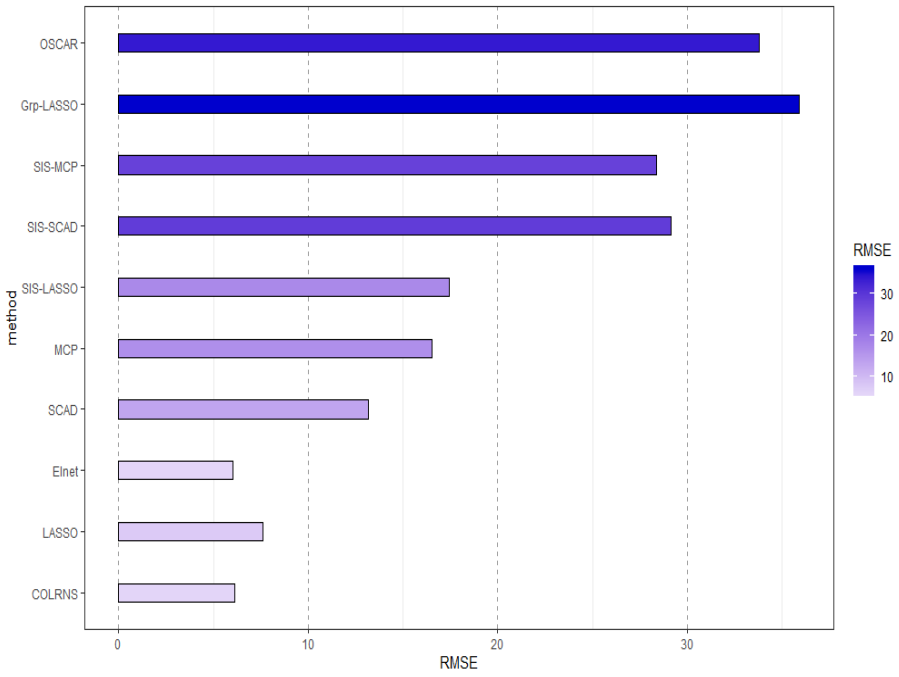
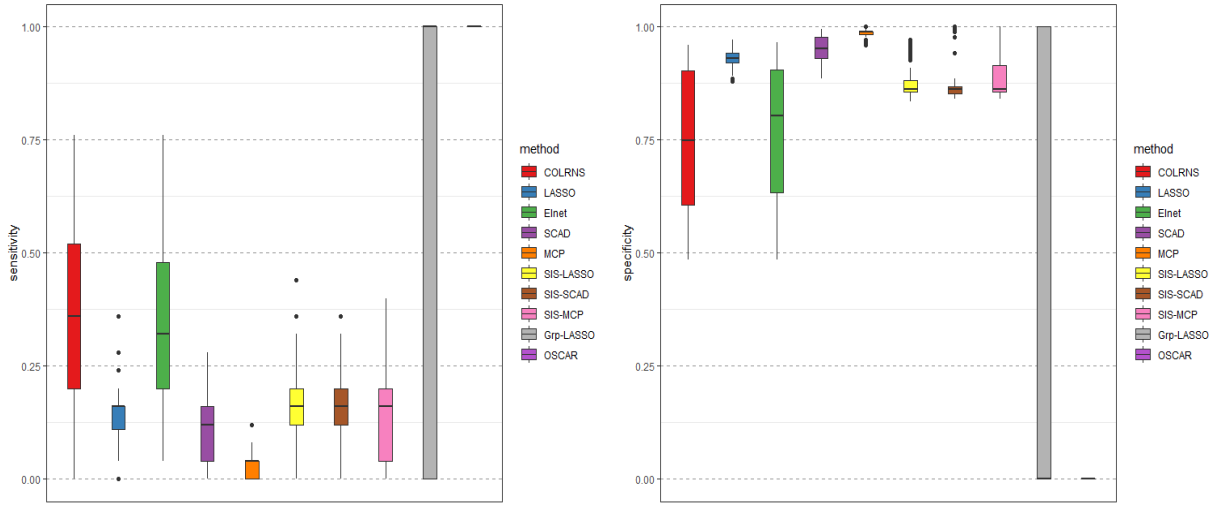
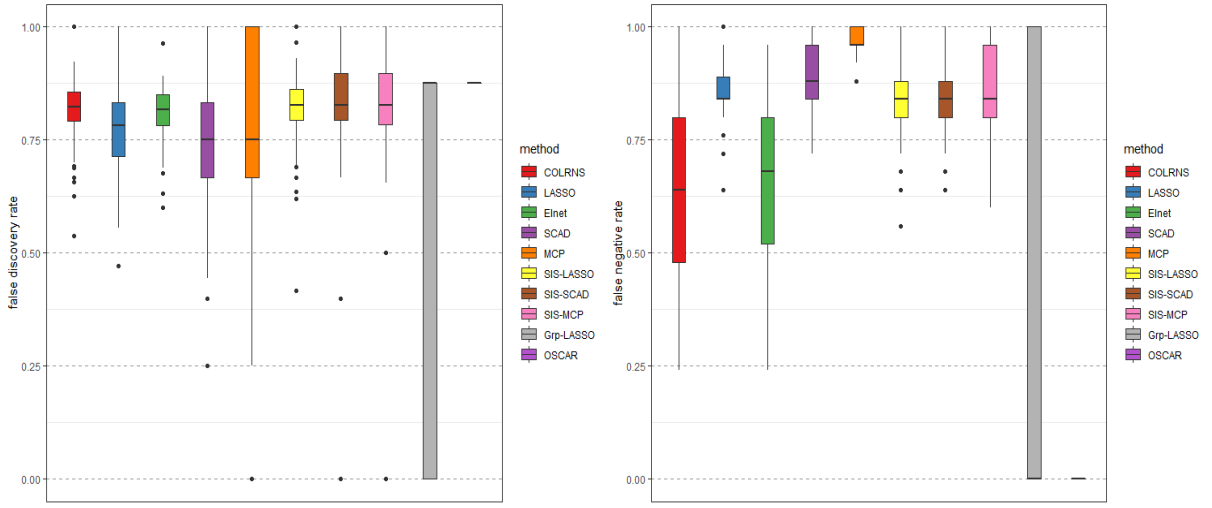


Figure 21: RMSE of parameter estimation in scenario 6



(a) Sensitivity

(b) Specificity



(c) False discovery rate

(d) False negative rate

Figure 22: Variable selection measurements in scenario 6

TABLE XXI: MSE OF PREDICTION IN SCENARIO 7

	COLRNS	LASSO	ELNET	SCAD	MCP	S-SCAD	S-MCP	S-LASSO	G-LASSO	OSCAR
Q1	29.6	43.3	35.5	58.3	92.8	162.4	167.7	45.3	552.1	202.6
Median	37.7	51.0	43.5	76.0	100.7	193.4	201.2	74.1	1629.0	228.1
Mean	40.7	55.8	46.0	84.2	100.4	193.0	201.9	71.2	1142.7	230.6
Q3	51.1	68.7	54.0	107.9	106.2	217.7	230.3	92.9	1629.0	256.8

TABLE XXII: RMSE OF PARAMETER ESTIMATION IN SCENARIO 7

	COLRNS	LASSO	ELNET	SCAD	MCP	S-SCAD	S-MCP	S-LASSO	G-LASSO	OSCAR
RMSE	7.96	10.84	7.92	15.42	22.20	17.76	30.97	31.73	36.49	33.96

TABLE XXIII: VARIABLE SELECTION MEASUREMENTS IN SCENARIO 7

	COLRNS	LASSO	ELNET	SCAD	MCP	S-SCAD	S-MCP	S-LASSO	G-LASSO	OSCAR
					Sensitivity					
Q1	0.311	0.156	0.328	0.111	0.022	0.133	0.128	0.111	0.000	1.000
Median	0.533	0.200	0.511	0.156	0.044	0.156	0.156	0.133	1.000	1.000
Mean	0.488	0.200	0.483	0.160	0.040	0.162	0.161	0.140	0.690	1.000
Q3	0.644	0.222	0.644	0.200	0.044	0.200	0.200	0.178	1.000	1.000
					Specificity					
Q1	0.497	0.871	0.497	0.877	0.981	0.852	0.852	0.858	0.000	0.000
Median	0.587	0.884	0.665	0.897	0.987	0.858	0.861	0.884	0.000	0.000
Mean	0.639	0.884	0.647	0.909	0.985	0.863	0.866	0.894	0.310	0.000
Q3	0.794	0.903	0.802	0.934	0.987	0.871	0.871	0.929	1.000	0.000
					False Discovery Rate					
Q1	0.685	0.612	0.690	0.570	0.400	0.690	0.690	0.647	0.000	0.775
Median	0.721	0.679	0.717	0.667	0.550	0.759	0.741	0.724	0.775	0.775
Mean	0.708	0.666	0.704	0.629	0.562	0.746	0.732	0.710	0.535	0.775
Q3	0.739	0.726	0.740	0.731	0.750	0.793	0.793	0.788	0.775	0.775
					False Negative Rate					
Q1	0.356	0.778	0.356	0.800	0.956	0.800	0.800	0.822	0.000	0.000
Median	0.467	0.800	0.489	0.844	0.956	0.844	0.844	0.867	0.000	0.000
Mean	0.512	0.800	0.517	0.840	0.960	0.838	0.839	0.860	0.310	0.000
Q3	0.689	0.844	0.672	0.889	0.978	0.867	0.872	0.889	1.000	0.000

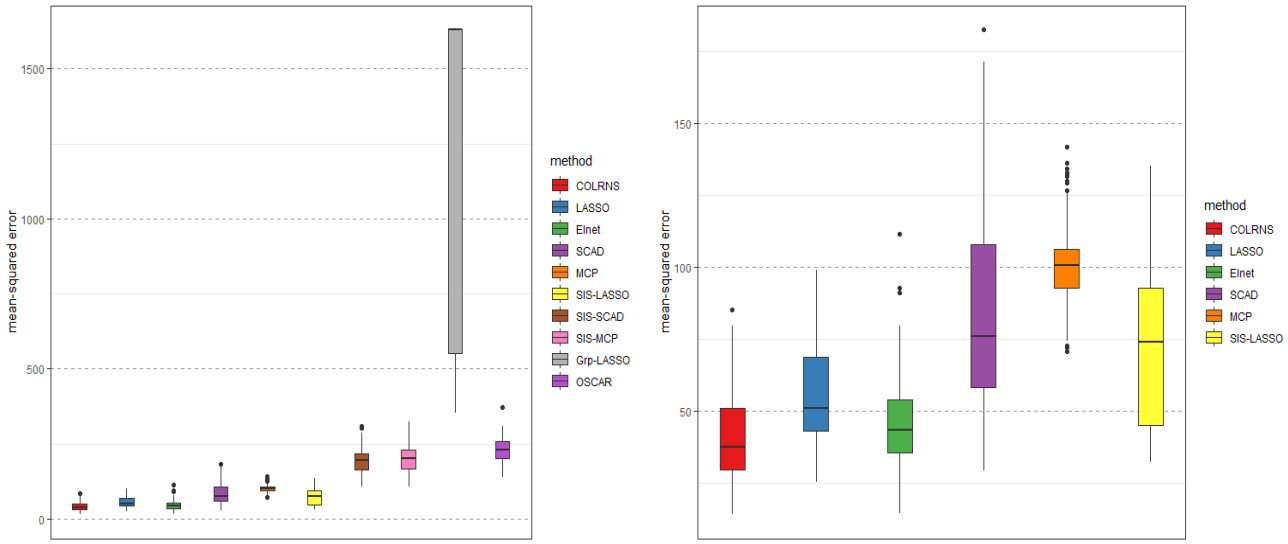


Figure 23: MSE of prediction in scenario 7

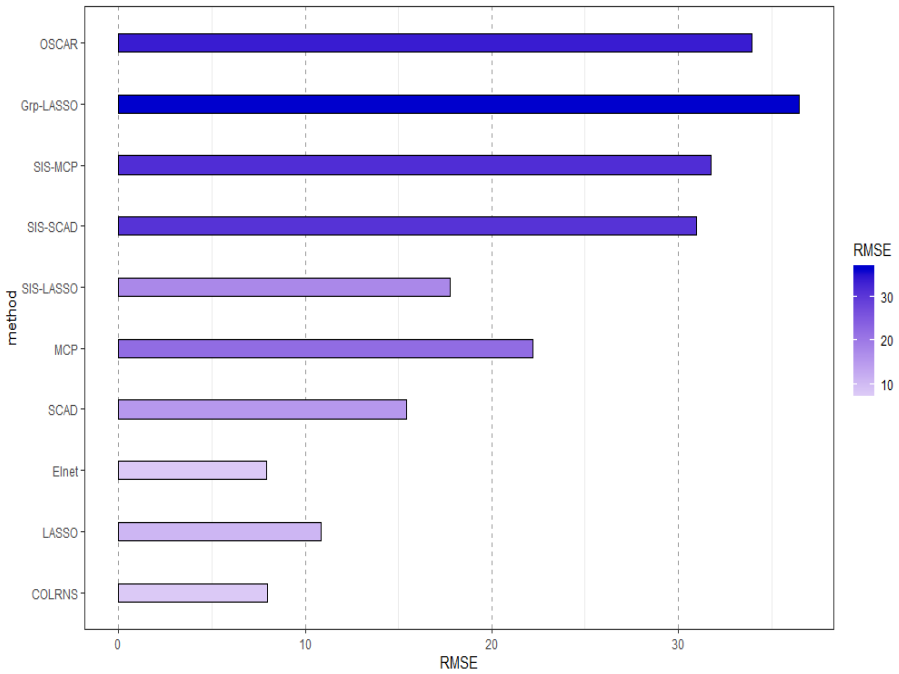
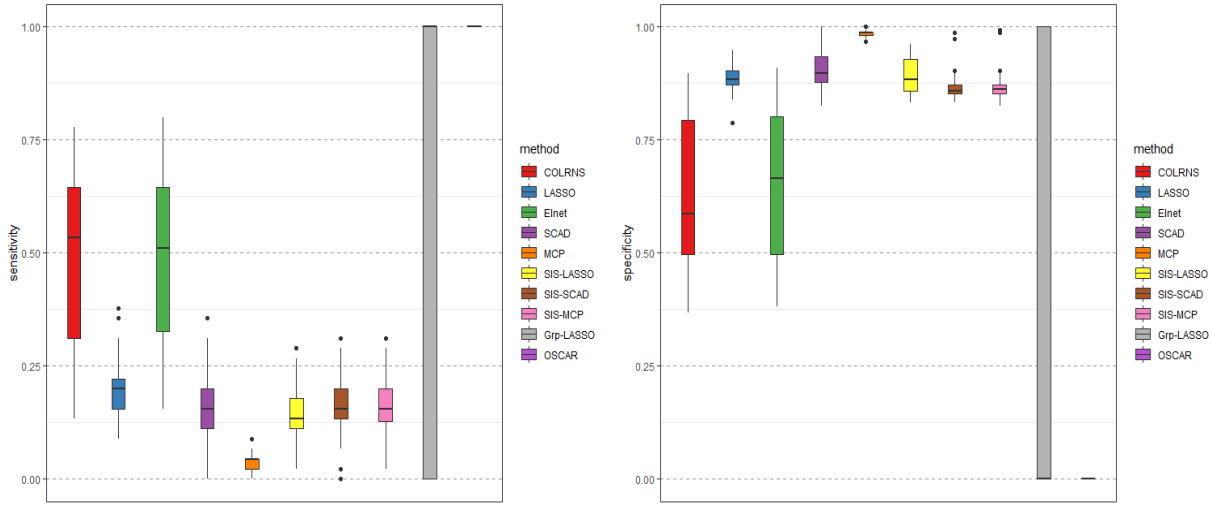
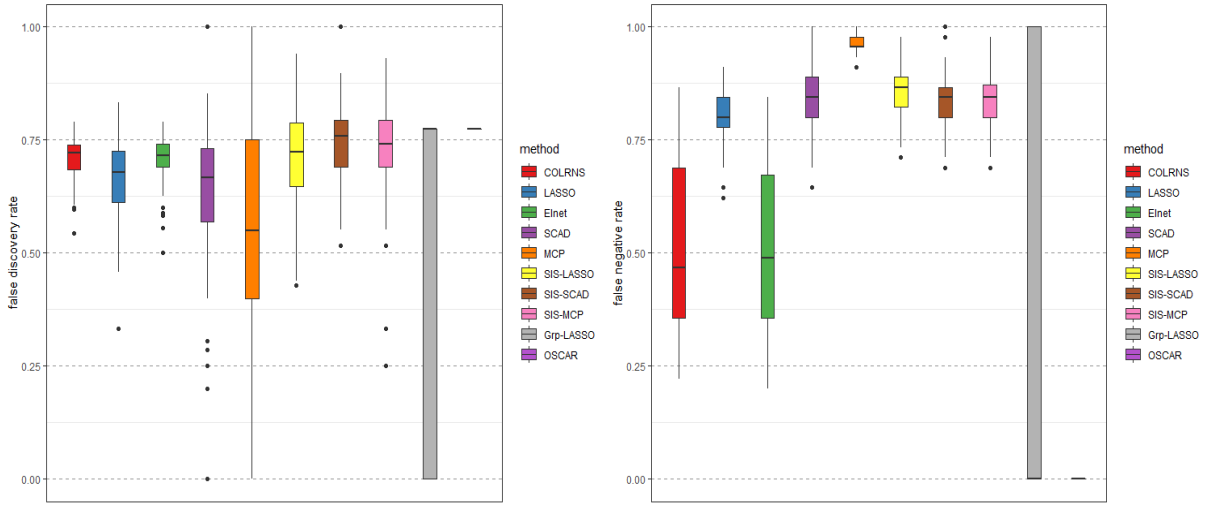


Figure 24: RMSE of parameter estimation in scenario 7



(a) Sensitivity

(b) Specificity



(c) False discovery rate

(d) False negative rate

Figure 25: Variable selection measurements in scenario 7

3.6 Data Application: Leukocyte Telomere Length and Exposure to Pollutants

3.6.1 Introduction

We apply the proposed feature selection method to the 2001-2002 NHANES data [Mitro et al., 2016, Gibson et al., 2019] used to investigate the association between exposures to persistent organic pollutants (POPs) and leukocyte telomere length (LTL), a biomarker associated with chronic disease. The authors hypothesize that exposures to POPs such as polychlorinated biphenyls (PCBs), dioxins, and furans would have association with longer LTL. We analyze the data to illustrate the application of the method and to demonstrate its performance in real-world data. By applying the method, we examine the relationship between the exposures and LTL and also evaluate the performance of the resulting regression model.

3.6.2 Data Description

We use the NHANES data in the 2001-2002 cycle where $11,039$ people were interviewed. Among the people over 20 years of age who provided a blood sample and consented to the usage of their DNA, biological samples were available for $4,260$ participants to process telomere length. We excluded those for whom environmental chemical analysis data were not available ($n = 2,850$) or individuals with missing values on covariates (body mass index (BMI) $n = 70$, education (n=2), serum cotinine ($n = 8$). The participants were further excluded who had missing information for individual PCBs, dioxins, or furans ($n = 327$). This resulted in $n = 1,003$ participants in the analysis dataset [Gibson et al., 2019].

The data includes 18 POPs (Table XXIV) exposures which were measured using high-resolution gas chromatography/isotope-dilution high resolution mass spectrometry. All POPs

measurements were lipid-adjusted by the U.S. Centers for Disease Control and Prevention (CDC) [Needham et al., 1989, Cawthon, 2002, Lin et al., 2010, Needham et al., 2013, Mitro et al., 2016]. We include the congeners whose concentrations were detected more than 60 percentage of samples. The limit of detection (LOD) were reported for each sample. The typical detection limits were 2 ng/g while it was as high as 10.5 ng/g. The samples below the LOD were replaced by the sample-specific LOD divided by the square root of 2 following the guideline of CDC. The details of exposure assessment are described in Mitro et al. [2016]. Figure 30 shows Pearson correlation matrix of 18 POPs variables.

Telomere length relative to standard reference DNA (T/S ratio) was measured with the quantitative polymerase chain reaction (qPCR) method [Cawthon, 2002,Lin et al., 2010]. Each sample was assayed in duplicate wells for three times producing six data points. The mean T/S ratio was calculated by averaging them [Needham et al., 2013]. The CDC performed a quality review before linking the telemore measurements to the NHANES 1999-2002 public usage data files.

3.6.3 Statistical Analysis

We apply variable selection methods in a multiple regression setting to examine associations between 18 POPs exposure measurements and relative LTL. Variable selection methods such as LASSO, Elastic net, and Group LASSO are applied to compare the results from Gibson et al. [2019] with the results derived from COLRNS. The models are adjusted by covariates that are included in Mitro et al. [2016] to control for confounding bias. The covariates are age, squared age, sex, race/ethnicity, educational attainment, BMI, and blood cell count and distribution

(white blood cell count, percent lymphocytes, percent monocytes, percent neutrophils, percent eosinophils, percent basophils). Covariates for blood cell count and distribution are included because LTL is measured in immune cells and the covariates are associated with serum PCBs [Mitro et al., 2016, Serdar et al., 2014].

The LTL and POPs, and serum cotinine are natural log-transformed for analyses accounting for their nonnormal distributions [Mitro et al., 2016]. Figure 26 shows the distribution of log-transformed relative LTL. Figure 27 presents the histogram of log-transformed POPs measurements. We use dummy variables representing each category of categorical covariates for selection purpose. Table XXIV shows the labels of POPs exposures, and Table XXV and Table XXVI present the labels for continuous and categorical covariates. The distribution of these variables are illustrated in Figure 28 and Figure 31.

In our application of variable selection methods, we only penalize regression coefficients of POPs exposures and always include other covariates in the model to control for confounding bias as in Gibson et al. [2019]. We use the grid of values $0.1, \ldots, 0.9$ for α, and use 10-fold cross-validation and minimum cross-validation prediction error for choosing λ and α. The mean squared error of the resulting model is calculated as

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^T\hat{\beta}(\alpha,\lambda))^2$$

where y_i is log-transformed LTL and x_i is a vector of POPs exposures and covariate values for i^{th} participants, $i = 1, \ldots, n$. We use the package 'glmnet' [Friedman et al., 2019] for implementing LASSO, Elastic net, and COLRNS, and 'grpreg' [Breheny, 2020] for Group LASSO.

For analysis of Group LASSO, we divide POPs variables into three groups corresponding to the grouping method of Mitro et al. [2016] based on toxic equivalency factor (TEF) that is assigned according to the evidence of toxicity by World Health Organization (WHO). The three groups are non-dioxin-like PCBs with no TEFs, non-ortho PCBs with high TEFs and AhR affinity, and toxic equivalent (TEQ) POPs with high TEFs and AhR affinity including furans, dioxins, and a mono-ortho PCB. The grouping infomration for each exposure is listed in Table XXIV.

All predictors are standardized before implementing the methods in order to apply the penalization across all regressors fairly [Tibshirani, 2007]. Before applying all the selection methods, we first examine the degree of collineariy in the data set and also fit a typical linear regression model with the same exposures and confounders for purpose of comparison with the models given by the penalized selection methods.

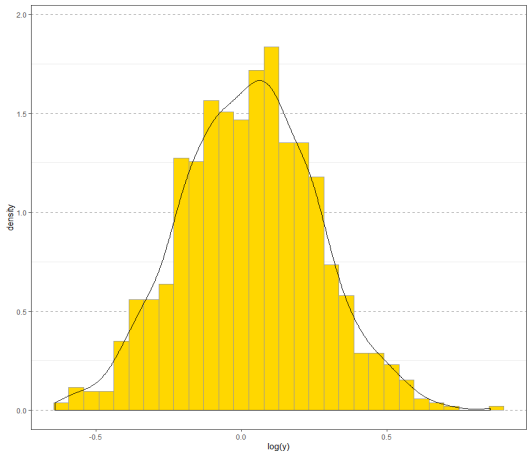


Figure 26: Density plot of log-transformed outcome

TABLE XXIV: GROUPS OF ENVIRONMENTAL EXPOSURES

Group	Variable
Non-dioxin-like PCBs	PCB74 (ng/g)
	PCB99 (ng/g)
	PCB138 (ng/g)
	PCB153 (ng/g)
	PCB170 (ng/g)
	PCB180 (ng/g)
	PCB187 (ng/g)
	PCB194 (ng/g)
Non-ortho PCBs	PCB126 (ng/g), 3,3',4,4',5-Pentachlorobiphenyl (PnCB)
	PCB169 (ng/g), 3,3',4,4',5,5'-hexachlorobiphenyl (HxCB)
Toxic equivalent POPs	PCB118 (ng/g)
	Dioxin (pg/g), 1,2,3,6,7,8-Hexachlorodibenzo-p-dioxin (HxCDD)
	Dioxin (pg/g), 1,2,3,4,6,7,8-Heptachlororodibenzo-p-dioxin (HpCDD)
	Dioxin (pg/g), 1,2,3,4,6,7,8,9-Octachlorodibenzo-p-dioxin (OCDD)
	Furan (pg/g), 2,3,4,7,8-Pentachlorodibenzofuran (PnCDF)
	Furan (pg/g), 1,2,3,4,7,8-Hexachlorodibenzofuran (HxCDF)
	Furan (pg/g), 1,2,3,6,7,8-Hexachlorodibenzofuran (HxCDF)
	Furan (pg/g), 1,2,3,4,6,7,8-Heptachlorodibenzofuran (HpCDF)

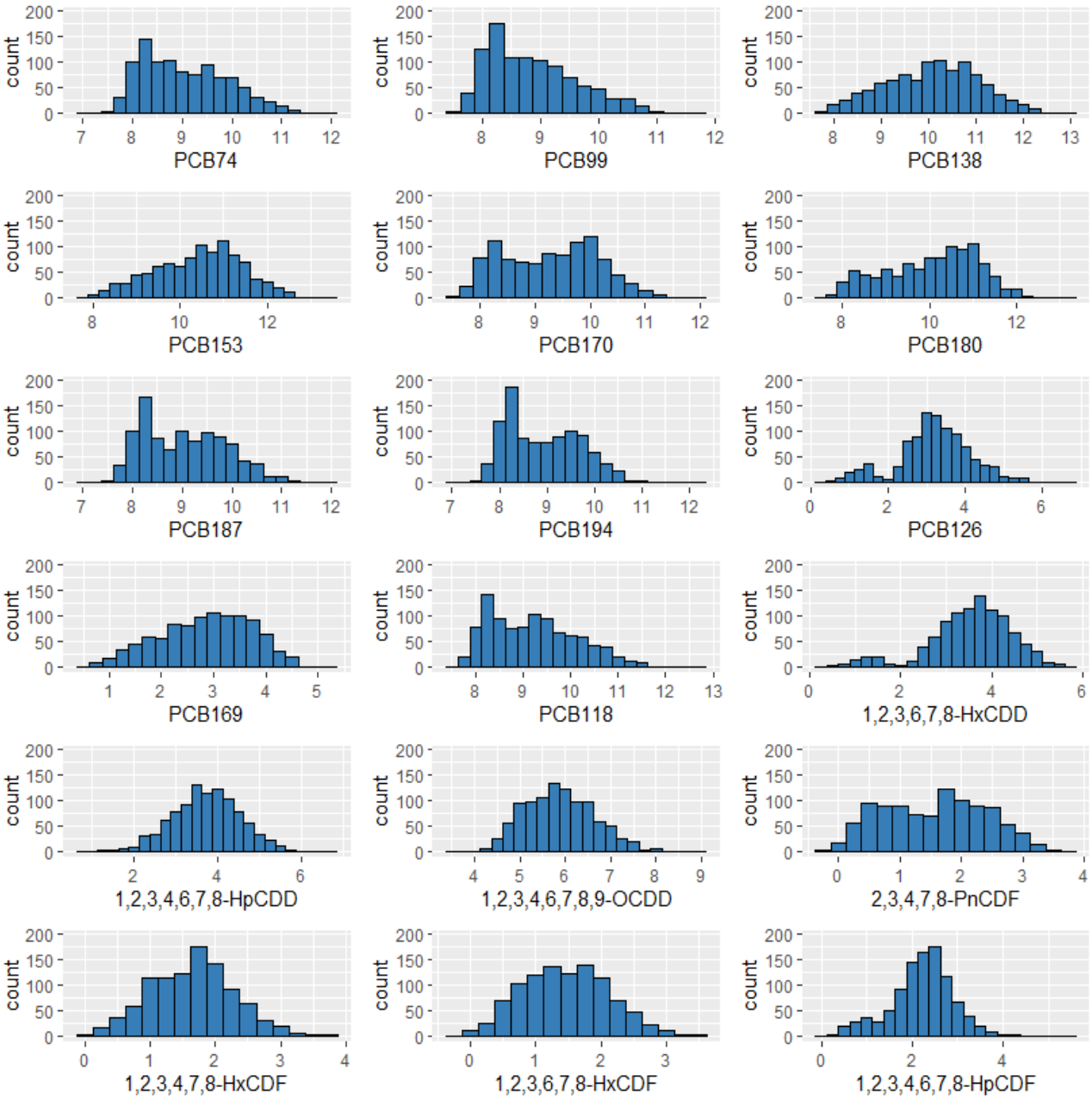


Figure 27: Histogram of log-transformed POPs exposures

TABLE XXV: CONTINUOUS COVARIATES

Variable	Label
Age_centered	Age at screening, centered
Age_sq	Age at screening, squared
Cotinine	Cotinine (ng/mL), log-transformed
White blood cell	White blood cell count (SI)
Lymphocyte	Lymphocyte percent (%)
Monocyte	Monocyte percent (%)
Eosinophils	Eosinophils percent (%)
Basophils	Basophils percent (%)
Neutrophils	Segmented neutrophils percent (%)

TABLE XXVI: CATEGORICAL COVARIATES

Category	Variable	Label
Gender	Male1	Male
	Otherwise	Female
Race/Ethnicity	Race_cat2	Mexican American
	Race_cat3	Non-Hispanic black
	Race_cat4	Non-Hispanic white
	Otherwise	Other Hispanic or Race including multi-racial
Body Mass Index	BMI_cat2	$25 \leq$ BMI ≤ 30 (kg/m^2)
	BMI_cat3	BMI > 30 (kg/m^2)
	Otherwise	BMI < 25 (kg/m^2)
Education Level	Edu_cat2	High school grad/GED or equivalent
	Edu_cat3	Some college or AA degree
	Edu_cat4	College graduate or above
	Otherwise	Less than 9th grade

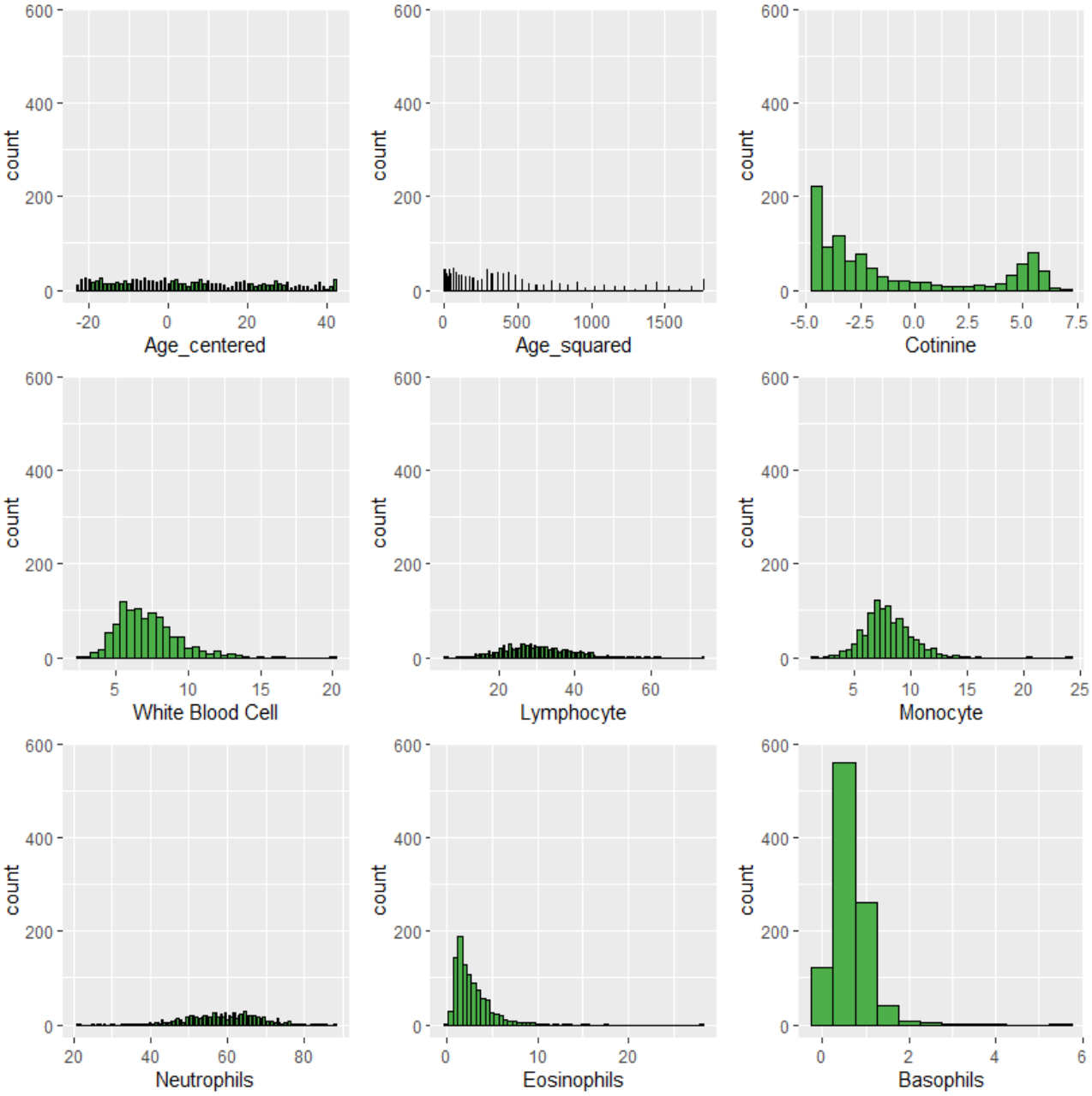


Figure 28: Histogram of continuous confounders

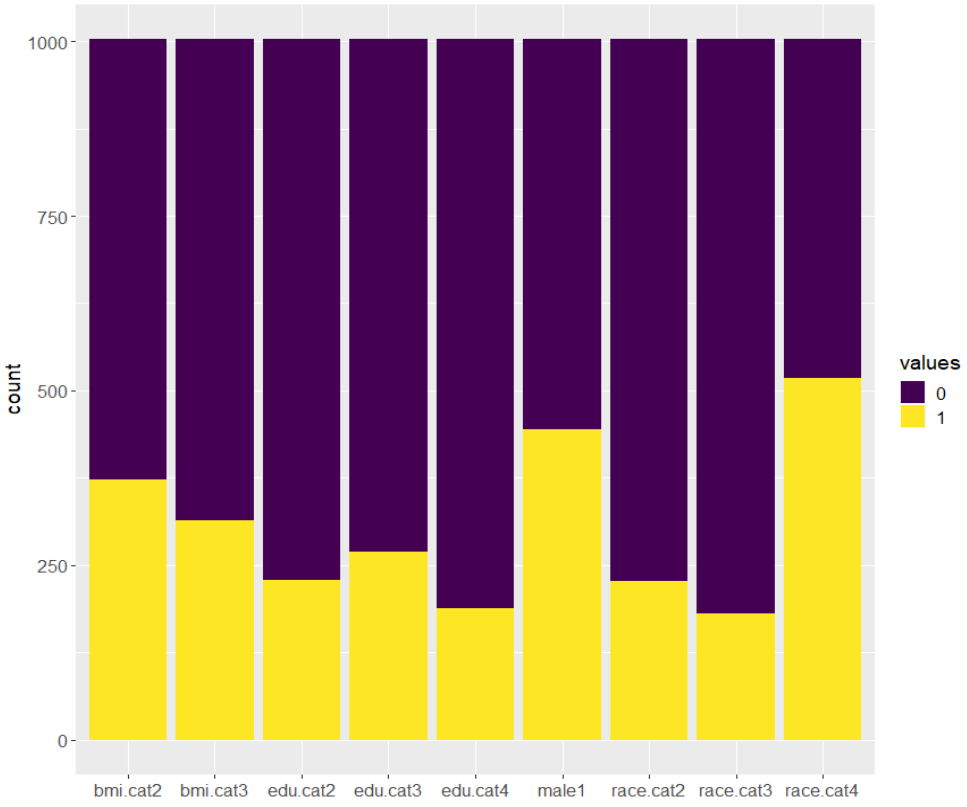


Figure 29: Bar chart of categorical confounders

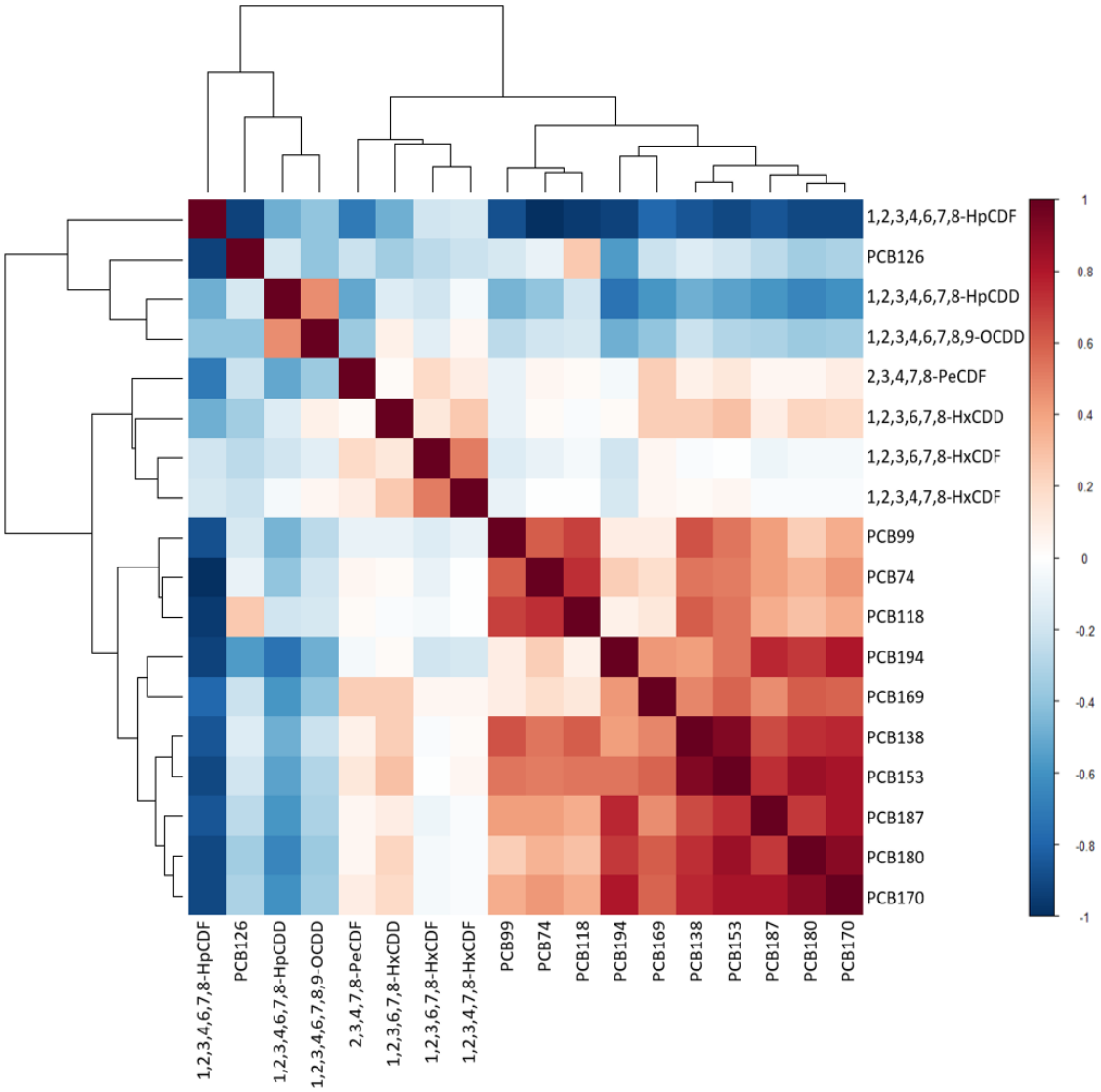


Figure 30: Pearsons's correlation matrix of POPs exposures

3.6.4 Analysis Results

We first examine the condition number by computing the square root of the ratio of the maximum to the minimum eigenvalues from the correlation matrix of 18 POPs exposures. The condition number equals 27.524 which indicates high degree of collinearity among the exposures. As mentioned in Chapter 3, when there exists strong collineariy in predictors, the standard errors of the predictors tend to be inflated, thus it often leads to false nonsignificant p-values by deflating the t-test statistics.

Table XXVII shows results of fitting a linear regression model with 18 POPs exposures with adjustment of 18 confounders. Among 18 POPs exposures, only Furan, 2,3,4,7,8-PnCDF is significantly associated with longer log-LTL under a 5% significance level ($\beta = 0.02866$). It can be interpreted as one unit increase in Furan, 2,3,4,7,8-PnCDF is associated with 0.02866 unit increase in log-LTL. In addition, PCB180 presents marginal association with shorter log-LTL under a significance level of 10% ($\beta = -0.05121$). Overall, the regression coefficients of 10 exposures are estimated as negative and 8 as positive, which is different from the results obtained by the selection methods: LASSO, Elastic net, and COLRNS in Table XLVI where all of the selected exposures have positive coefficients.

For LASSO, the best λ is estimated as 0.0034 by cross-validation. Under this parameter, the LASSO selects 4 POPs as important exposures: PCB99 ($\beta = 0.00139$), PCB126 ($\beta = 0.01033$), PCB118 ($\beta = 0.00340$), and Furan, 2,3,4,7,8-PnCDF ($\beta = 0.02205$).

The Elastic net chooses the same 4 exposures with similar coefficients as LASSO estimates, however, it selects one additional exposure Furan, 1,2,3,4,6,7,8-HpCDF ($\beta = 0.00009$) in the

group of toxic equivalent POPs. The model is estimated under the parameter $\alpha = 0.8$ and $\lambda = 0.0039$.

The proposed COLRNS approach selects all of the predictors selected by LASSO and Elastic net and one additional PCB: PCB99 ($\beta = 0.00327$) in the group of non-dioxin-like-PCBs, PCB126 ($\beta = 0.01155$) and PCB169 ($\beta = 0.00075$) contained in the group of non-ortho PCBs, PCB118 ($\beta = 0.00559$), Furan, 2,3,4,7,8-PnCDF ($\beta = 0.02419$), and Furan, 1,2,3,4,6,7,8-HpCDF ($\beta = 0.000221$) from the group of toxic equivalent POPs. The size of coefficients is slightly larger than that estimated by LASSO and Elastic net.

The Group LASSO chooses all exposures in the two groups: the non-ortho PCBs and toxic equivalent POPs, while screening the other exposures in the group of non-dioxin-like-PCBs to have zero coefficients. Increases in the two non-orthos PCBs are associated with longer log-LTL (PCB126 $\beta = 0.01149$, PCB 169 $\beta = 0.00756$). Increased exposure to five toxic equivalent POPs are also associated with longer log-LTL (PCB118 $\beta = 0.00331$, Dioxin, 1,2,3,4,6,7,8-HpCDD $\beta = 0.00186$, Furan, 2,3,4,7,8-PnCDF $\beta = 0.00870$, Furan, 1,2,3,4,7,8-HxCDF $\beta = 0.00248$, Furan, 1,2,3,4,6,7,8-HpCDF $\beta = 0.00544$). Increases of the other three toxic equivalent POPs are associated with shorter log-LTL.

Among the confounders in Table XXVII, male gender, BMI, ethnicity of Mexican American and Non-Hispanic white, white blood cell count, the distribution of four blood cells (lymphocyte, monocyte, eosinophils, neutrophils) and increase in age are negatively associated with longer log-LTL. Higher level of education, ethnicity of non-Hispanic black, increased serum cotinine and basophils are positively associated with longer log-LTL.

The mean sqaured error of the resulting models by the penalized selection methods are shown in Table XXIX. The model resulted from COLRNS has the least error among the four models derived by the penalized selection methods.

TABLE XXVII: RESULTS OF LINEAR REGRESSION

Variable	Estimate (10^{-3})	t-value	p-value
Intercept	4233.00	0.52	0.604
PCB74	-1.51	-0.08	0.940
PCB99	4.02	0.18	0.860
PCB138	-19.40	-0.55	0.585
PCB153	47.39	1.10	0.271
PCB170	17.08	0.46	0.645
PCB180	-51.21	-1.86	0.064
PCB187	-7.75	-0.30	0.765
PCB194	-0.22	-0.01	0.994
Non-ortho PCBs			
PCB126	12.57	1.17	0.241
PCB169	19.23	1.18	0.239
Toxic equivalent POPs			
PCB118	13.23	0.58	0.564
Dioxin, 1,2,3,6,7,8-HxCDD	-9.76	-0.77	0.441
Dioxin, 1,2,3,4,6,7,8-HpCDD	-15.11	-0.86	0.390
Dioxin, 1,2,3,4,6,7,8,9-OCDD	-21.82	-1.31	0.190
Furan, 2,3,4,7,8-PnCDF	28.66	2.36	0.019
Furan, 1,2,3,4,7,8-HxCDF	11.09	0.57	0.568
Furan, 1,2,3,6,7,8-HxCDF	-15.11	-0.86	0.390
Furan, 1,2,3,4,6,7,8-HpCDF	16.89	1.38	0.169
Age (years)			
Age	-5.87	-7.13	0.000
Age^2	-0.03	-1.11	0.266
Sex			
Male	-40.25	-2.46	0.014
BMI (kg/m^2)			
25-29.9	-7.67	-0.45	0.650
$\geq$ 30	-22.47	-1.18	0.239
Education			
High school graduate	19.55	1.05	0.295
Some college	37.63	2.04	0.042
$\geq$ College graduate	12.84	0.59	0.556
Race/ethnicity			
Mexican American	-21.44	-0.77	0.442
Non-Hispanic black	13.13	0.45	0.653
Non-Hispanic white	-33.50	-1.29	0.197
Serum cotinine (ng/mL)	4.14	1.91	0.057
White blood cell count (SI)	-5.64	-1.52	0.128
Blood cell distribution (%)			
Lymphocyte	-41.71	-0.51	0.609
Monocyte	-46.72	-0.57	0.567
Eosinophils	-40.56	-0.50	0.619
Basophils	-24.69	-0.30	0.765
Neutrophils	-40.65	-0.50	0.618

TABLE XXVIII: REGRESSION COEFFICIENTS FROM THE PENALIZED VARIABLE SELECTION METHODS (10^{-3})

Pollutants	G-LASSO	LASSO	ELNET	COLRNS
Intercept	157.34	117.02	108.12	90.55
Non-dioxin-like-PCBs				
PCB74	.	.	.	.
PCB99	.	1.39	1.86	3.27
PCB138	.	.	.	.
PCB153	.	.	.	.
PCB170	.	.	.	.
PCB180	.	.	.	.
PCB187	.	.	.	.
PCB194	.	.	.	.
Non-ortho PCBs				
PCB126	11.49	10.33	10.56	11.55
PCB169	7.56	.	.	0.75
Toxic equivalent POPs				
PCB118	3.31	3.40	3.85	5.59
Dioxin, 1,2,3,6,7,8-HxCDD	-3.53	.	.	.
Dioxin, 1,2,3,4,6,7,8-HpCDD	1.86	.	.	.
Dioxin, 1,2,3,4,6,7,8,9-OCDD	-6.35	.	.	.
Furan, 2,3,4,7,8-PnCDF	8.70	22.05	22.38	24.19
Furan, 1,2,3,4,7,8-HxCDF	2.48	.	.	.
Furan, 1,2,3,6,7,8-HxCDF	-4.56	.	.	.
Furan, 1,2,3,4,6,7,8-HpCDF	5.44	.	0.09	2.21
Age (years)				
Age	-6.10	-6.39	-6.43	-6.60
Age^2	-0.01	-0.01	-0.01	-0.01
Sex				
Male	-40.33	-35.04	-34.78	-34.07
BMI (kg/m^2)				
25-29.9	-4.20	-5.80	-5.83	-5.75
$\geq$ 30	-18.07	-18.98	-19.16	-19.53
Education				
High school graduate	19.55	20.13	20.07	19.95
Some college	36.15	36.34	36.25	36.17
$\geq$ College graduate	15.27	16.48	16.41	16.26
Race/ethnicity				
Mexican American	-30.20	-33.08	-32.80	-30.89
Non-Hispanic black	9.88	4.94	4.63	2.94
Non-Hispanic white	-34.85	-39.39	-39.42	-39.45
Serum cotinine (ng/mL)	3.38	3.32	3.34	3.39
White blood cell count (SI)	-6.37	-6.27	-6.25	-6.10
Blood cell distribution (%)				
Lymphocyte	-1.11	-1.09	-1.10	-1.34
Monocyte	-6.03	-6.42	-6.45	-6.80
Eosinophils	-0.27	0.07	0.07	-0.15
Basophils	18.55	17.80	17.72	17.27
Neutrophils	-0.26	-0.20	-0.20	-0.42

TABLE XXIX: MEAN SQUARED ERROR OF THE MODELS (10^{-5})

	COLRNS	LASSO	ELNET	G-LASSO
MSE	3928.7	3937.9	3936.1	3947.3

3.7 Cross-validation Study

3.7.1 Methods

In this section, we considered data-driven cross-validation to evaluate and compare the prediction performance of the selection methods: COLRNS, LASSO, Elastic net, and BKMR. We apply nested cross-validation to the same NHANES data considerd in section 3.6 which include 18 environmental exposures and 18 confounders of $11,039$ people. We randomly divide the data into 10 folds. 9 folds are used as a training set to fit the model and remaining one fold is set as a test set. In this study, since we focus on evaluating prediction performance, we penalize not only 18 environmental exposures, but also 18 confounders not forcing them to be included in the resulting models to increase prediction performance. We evaluate the prediction performance of the resulting model in the test set by estimating the mean squared error (MSE)

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y_i})^2$$

where m is the number of observation in the test set. Here, $\hat{y_i} = x_i^T\hat{\beta}$ where x_i is the vector of environmental exposures and confounders of the i^{th} observation in the test set. For BKMR, $\hat{y_i} = h(\hat{z_i}) + x_i^T\hat{\beta}$ where $h(\hat{z_i})$ is the posterior mean of the exposure-response function $h(\cdot)$, z_i

is the vector of environmental exposures and x_i is a set of confounders of the i^{th} observation in the test set. The same procedure is repeated to measure MSE by rotating the role of the training and test set. The 10 values of MSE estimated from each test set are averaged for the average prediction error across the 10 folds. We conduct the simulation 20 times independently in the same manner. The descriptive statistics are calculated and box plots are drawn with the average prediction errors for each method.

3.7.2 Results

Table XXX shows that the descriptive statistics of the average prediction errors from the 20 repetitions of the nested cross-validation procedure. The COLRNS approach presents the best performance in prediction with the lowest values for all descriptive statistics including Q1, median, mean, and Q3 of average prediction errors in comparison to other methods. The LASSO and Elastic net have the similar values, but Elastic net shows higher values for all descriptive statistics and also larger variability in Figure 31. The prediction performance of BKMR is the worst among the methods with the highest Q1, median, mean, and Q3 values and the largest variance of the average prediction errors.

TABLE XXX: AVERAGE PREDICTION ERROR (10^{-4})

	COLRNS	LASSO	ELNET	BKMR
Q1	425.9	429.8	430.0	440.1
Median	427.0	430.3	431.1	444.2
Mean	426.8	430.5	430.8	442.9
Q3	428.0	431.1	431.8	447.6

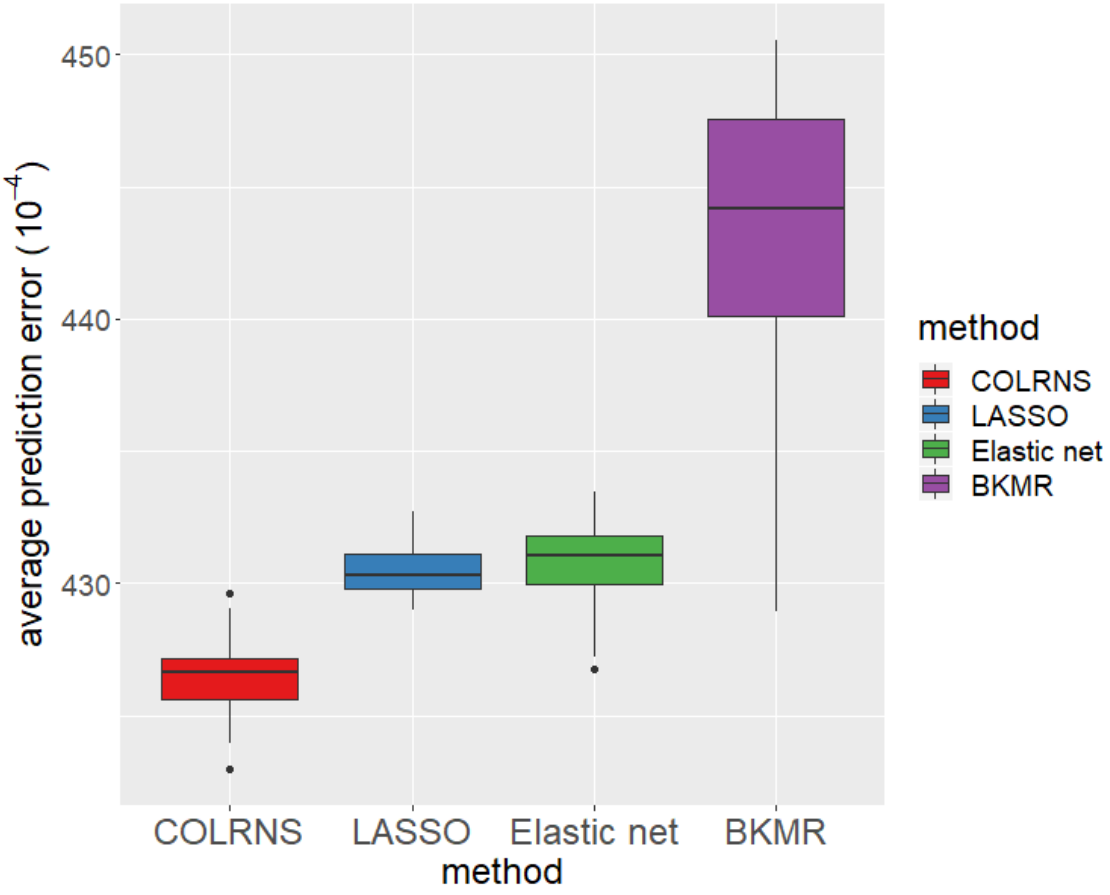


Figure 31: Box plot of average prediction error

CHAPTER 4

HIGH DIMENSIONAL VARIABLE SELECTION IN PRESENCE OF COLLINEARITY: GENERALIZED LINEAR MODEL

4.1 Introduction

In this chapter, we extend COLRNS to be applicable in the setting of generalized linear model. In the specific setting of binary outcomes, logistic regression is a popular generalized linear model using logit-tansformation of a binomial parameter. Fitting a logistic regression model is a popular approach for addressing classification problems which assume that class labels take values 0 or 1 [Fan et al., 2009]. Hence, the variable selection for binary outcomes can also be treated as a classification problem aiming at finding a discriminant function that accurately classifies future observations [Fan and Lv, 2010]. As a traditional variable selection problem, it also aims to identify all important variables and to precisely estimate the coefficients of those variables [Fan and Lv, 2010].

4.2 COLRNS-GLM

4.2.1 Generalized Linear Models

We assume that the random variable y is from an exponential family. The probability density function taking the canonical from is given by

$$f_Y(y;\theta) = exp\left\{ \frac{y\theta - b(\theta)}{\phi} + c(y,\phi) \right\}$$

for some known functions $b(\cdot), c(\cdot)$, and unknown function θ. We assume the dispersion parameter $\phi = 1$ without loss of generality and each predictor is standardized with mean 0 and variance 1 as before. We consider the problem of estimating $\beta = (\beta_1, \ldots, \beta_p)$ from the following generalized linear model

$$E(y|\boldsymbol{x}) = b'(\theta(\boldsymbol{x})) = g^{-1}\left(\sum_{j=1}^{p} \beta_j x_j\right)$$

where $\boldsymbol{x} = (x_1, \ldots, x_p)^T$ represents p covariates. When g is canonical link meaning $g = (b')^{-1}$, then $\theta(\boldsymbol{x}) = \sum_{j=1}^{p} \beta_j x_j$.

Fan et al. [2009] mention that from the form of the likelihood function of a generalized linear model, it is obvious that modeling the relationship between Y and $(X_1, \ldots, X_p)^T$ amounts to minimizing a negative pseudo-likelihood function which has the form

$$Q(\beta_0, \boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} l(y_i, \boldsymbol{x}_i^T \boldsymbol{\beta})$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$. We have in general that

$$l(y_i, \boldsymbol{x}_i^T \boldsymbol{\beta}) = \{b(\theta(g^{-1}(\boldsymbol{x}_i^T \boldsymbol{\beta})) - y_i \theta(g^{-1}(\boldsymbol{x}_i^T \boldsymbol{\beta}))\}.$$

If we have the canonical link function, $g(\cdot) = \theta(\cdot)$, then it simplifies to

$$l(y_i, \boldsymbol{x}_i^T \boldsymbol{\beta}) = \{b(\boldsymbol{x}_i^T \boldsymbol{\beta}) - y_i \theta(\boldsymbol{x}_i^T \boldsymbol{\beta})\}.$$

In the logistic regression model with the response taking 0 or 1, we have the form

$$l(y_i, \boldsymbol{x}_i^T\boldsymbol{\beta}) = \{log(1 + e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}) - y_i\boldsymbol{x}_i^T\boldsymbol{\beta}\}.$$

4.2.2 COLRNS-GLM: Learning with Maximum Marginal Likelihood Estimator

We propose to use learning with the maximum marginal likelihood estimator (MMLE) [Fan et al., 2009]. The MMLE is defined as the minimizer of the component-wise regression

$$\hat{\beta}_{Mj} = \arg\min_{\beta_o,\beta_j} \frac{1}{n}\sum_{i=1}^{n} l(\beta_0 + x_{ij}\beta_j, y_i)$$

where $l(y;\theta) = -[y\theta - b(\theta) - logc(y)]$ and $x_i = (x_{i1}, \ldots, x_{ip})^T$. Here, l can be regarded as the loss of adopting $\beta_0 + x_i^T\beta$ to predict y_i. These estimators can be ranked according to the size of MMLE. As the size is smaller, the corresponding variable is more important to predict the response. We use the method to figure out the marginal utilities in the generalized linear model with binary responses. The feature that has the least value of MMLE is chosen as the lead variable, which is corresponding to the selection of the lead variable through correlation learning under continuous outcomes in Section 3.3.2.

4.2.3 COLRNS-GLM: Selection

Once we identify the lead variable with the least MMLE, correlation learning with the lead variable and the remaining predictors are performed to detect the lead cluster. The procedures are exactly the same as the case of continuous outcomes in Section 3.3.3. After we screen features by identifying the lead cluster, we perform variable selection on the screened features in the

cluster using penalized likelihood [Fan et al., 2009]. Let $x_1, \ldots, x_d$ denote a set of variables on which we perform selection. In the penalized likelihood approach, we aim to minimize

$$l(\beta_0, \boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} l(\beta_0 + \boldsymbol{x}_{i,d}^T \boldsymbol{\beta}, y_i) + \sum_{j=1}^{d} p_\lambda(|\beta_j|)$$

where $p_\lambda(\cdot)$ is a penalty function, $\lambda > 0$ is a regularization parameter, $\boldsymbol{x_{i,d}} = (x_{i1}, \ldots, x_{id})^T$, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)^T$. As introduced in Section 3.3.4, we consider combining the l_2 penalty with another penalty. The penalty function $p_\lambda(\cdot)$ can be written as

$$p_\lambda(|\beta_j|) = \lambda(1-\alpha)\beta_j^2 + \alpha q_\lambda(|\beta_j|)$$

where $\alpha \in [0, 1]$ is a parameter that can be varied. The other penalty function can be l_1 penalty $q_\lambda(|\beta_j|) = \lambda|\beta|$ [Tibshirani, 1996, Park and Hastie, 2007], SCAD penalty [Fan and Li, 2001], which is symmetric with $q_\lambda(0) = 0$ and a quadratic spline whose first order derivative is given by

$$q'_\lambda(|\beta|) = \lambda \left\{ \mathbb{I}_{\{|\beta| \leq \lambda\}} + \frac{(a\lambda - |\beta|)_+}{(a-1)\lambda} \mathbb{I}_{\{|\beta| > \lambda\}} \right\}$$

for some $\alpha > 2$ and $|\beta| > 0$, or MCP penalty [Zhang, 2007], $q'_\lambda(|\beta|) = (\lambda - |\beta|/\alpha)_+$.

4.2.4 COLRNS-GLM: Iterative Feature Selection

We apply iterative cluster based selection that is introduced in Section 3.3.5 to generalized linear models. For iterative application of cluster based selection in linear model, we utilize residuals based on the fitted model using variables selected in previous steps. However, such a

residual based approach is not immediately generalized in the GLM setting. Let S_k denote a set of selected variables from the k^{th} implementation of selection, C_k be the k^{th} lead cluster, and A_{k+1} be the active set on which we search for the $(k+1)^{th}$ lead variable as described in Section 3.3.5. Instead of residuals, in COLRNS-GLM we consider

$$\hat{\beta}_{Mj}^{(k+1)} = \underset{\beta_0,\beta_j}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} l(\beta_0 + \boldsymbol{x}_{i,S_k}^T \boldsymbol{\beta}_{S_k} + x_{ij}\beta_j, y_i)$$

for $j \in A_{k+1} = \{1, \ldots, p\} \setminus (S_k \cup C_k)$, where $\boldsymbol{x}_{i,S_k}$ is the sub-vector of X_i corresponding to the variables in S_k. The estimated coefficient $\hat{\beta}_{Mj}^{(k+1)}$ can be regarded as the additional contribution of the j^{th} predictor X_j given the existence of the set S_k which contains previously selected variables. We use the coefficients as marginal utilities and choose the lead variable which has the smallest marginal utility. Once the new lead variable is identified, we iterate the procedures of cluster detection and selection as described in the conceptual diagram in Figure 3.

4.3 Simulations for Performance Evaluation

We evaluate and compare performance of COLRNS-GLM with LASSO, Elastic net, SCAD, MCP, SIS-LASSO, SIS-SCAD, SIS-MCP, and Group LASSO in simulated data sets. The performance of the model resulted by the methods is evaluated after 100 simulations in three different areas as in simulations in Chapter 3: prediction, parameter estimation, and selection of variables of true signal.

4.3.1 Data Generation Model

For simulations, the data are generated from the generalized linear model described in Section 4.2.1

$$E(y|\boldsymbol{x}) = b'(\theta(\boldsymbol{x})) = g^{-1}\left(\sum_{j=1}^{p}\beta_j x_j\right)$$

where $\boldsymbol{x} = (x_1, \ldots, x_p)^T$ is p-dimensional covariate and each predictor is standardized with mean 0 and variacne 1. When g is canonical link meaning $g = (b')^{-1}$, then $\theta(\boldsymbol{x}) = \sum_{j=1}^{p}\beta_j x_j$. As the outcome is binary, we use the logit link for the canonical link function, that is, $g(\mu) = log(\mu/(1-\mu))$ where $\mu = E(y|\boldsymbol{x})$.

The number of observations in the data set is $n = 500$ which are equally divided into training and test sets. The training set is used to fit the model and the fitted model is evaluated on the test set. We consider four scenarios by varying the number of predictors p, the number of groups of strongly correlated predictors q, and the true coefficient values $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^T$. When there are q number of groups which are denoted by $G_1, \ldots, G_q$, we assume that predictors are strongly correlated within a group. The within group pairwise correlation $Corr(X_i, X_j)$ is taken as 0.6 in Scenario 1, 2, and 4, and 0.8 in Scenario 3. Two predictors from different groups are weakly correlated with the correlation coefficient 0.2. The details of each simulation scenario are described as follows.

- **Scenario 1:** The number of predictors $p = 30$. There are three groups where each group has 10 strongly correlated predictors. The true regression coefficients are given by

$$\beta = (\underbrace{0.5, 0.5, 0.5}_{3}, \underbrace{0, \cdots, 0}_{7}) \qquad \text{for} \quad G_1,\, G_2,\, G_3.$$

- **Scenario 2:** We have three groups of 10 strongly correlated predictors as in Scenario 1. The data generating regression coefficients are set as

$$\beta = (0.5, 0, 0, 0, 0, 0.5, 0, 0, 0, 0) \qquad \text{for} \quad G_1,\, G_2,\, G_3.$$

- **Scenario 3:** There are 30 predictors with the three groups of 10 strongly correlated predictors as the same as in the previous two scenarios. The data generating regression coefficients are smaller and not sparse which are described as

$$\beta = (\underbrace{0.05, \cdots, 0.05}_{5}, \underbrace{0, \cdots, 0}_{5}) \qquad \text{for} \quad G_1,\, G_2,\, G_3.$$

- **Scenario 4:** The number of predictors p is 100. There are five groups of strongly correlated predictors with each group having 20 variables within a group. The data generating true regression coefficients are set as

$$\beta = (\underbrace{0.06, 0.06, 0.06}_{3}, \underbrace{0, \cdots, 0}_{17}) \qquad \text{for} \quad G_1, G_2, \ldots, G_5.$$

4.3.2 Evaluation Criteria

4.3.2.1 Prediction

As the outcome is binary taking value either 0 or 1, we can create the confusion matrix Table XXXI by comparing the true value of outcome to the predicted one. Based on this confusion matrix, we compute prediction accuracy to access prediction performance. Accuracy is a typical performance metric for classifiers [Hernandez-Orallo et al., 2012]. It indicates the percentage of the cases that are correctly predicted among all cases. The measurement is described as follows

$$\textit{Prediction Accuracy} = \frac{\textit{Number of true positive + Number of true negative}}{\textit{Number of all cases}}.$$

TABLE XXXI: CONFUSION MATRIX

		Predicted Class	
		1	0
True Class	1	*True positive*	*False negative*
	0	*False positive*	*True negative*

4.3.2.2 Parameter Estimation

We use the RMSE as descibed in section 3.5.2.2 to assess estimation accuracy of the parameters $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^T$ in the generalized linear model.

4.3.2.3 Variable Selection

The same metrics described in section 3.5.2.3 are used to evaluate the performance of variable selection: *sensitivity*, *specificity*, *false discovery rate*, *false negative rate*. These measures are computed based on the confusion matrix Table XXXI by comparing the true versus predicted class of each variable.

4.3.3 Simulation Results

The COLRNS-GLM performs the best in terms of prediction performance by having the highest level of prediction accuracy across all scenarios. It shows the highest Q1, median, mean, and Q3 of prediction accuracy in Scenario 1 with 3 groups of correlated predictors of relatively large signals, and also in Scenario 3 with the predictors of less sparse and weak signals. The results are similar with the largest median, mean, and Q3 in Scenario 2, and the greatest Q3 of prediction accuracy in Scenario 3.

The SIS-MCP illustrates poor prediction performance with the lowest Q1, median, mean, and Q3 of prediction accuracy in Scenarios 1, 2, and 3 with 30 predictors in 3 groups, and the smallest median and Q3 in Scenario 4 having 100 predictors in 5 groups. The LASSO also has bad performance in prediction with the least Q3 in Scenario 2, and the worst Q1 and mean accuracy in Scenario 4.

The COLRNS-GLM demonstrates the best ability in accurate estimation of parameters by obtaining the least RMSE in Scenario 1 and 2 with relatively large sparse signals in comparison to the other selection methods. In Scenarios 3 and 4 with weak signals, the estimation error is slightly larger than LASSO, Elastic net, and SIS combined methods, however, it is far less than MCP, SCAP and Group LASSO which show poor performance in precise parameter estimation.

For variable selection performance, COLRNS-GLM yields high sensitivity across all scenarios. When the true signals are sparse and large, it shows as good performance as Elastic net in Scenarios 1 and 2. In Scenario 1, it gives the largest Q1, median, and Q3 of sensitivity, and the greatest Q1 and Q3 in Scenario 2. It also demonstrates its ability of detecting true signals even when the true signals are weak by having the largest values for all the descriptive statistics of sensitivity in Scenarios 3 and 4.

In Scenarios 1 and 2, COLRNS-GLM and Elastic net have similar results for sensitivity as mentioned before, however, COLRNS improves specificity with the larger Q1, median, mean, and Q3 values compared to Elastic net resulting in more robust outcomes with less IQR values in both scenarios. Thus, COLRNS-GLM has better ability of screening out unimportant variables than Elastic net when there are correlated predictors of sparse and large signals.

In Scenarios 3 and 4, there are some cases where none of important predictors are chosen because of their weak signals. Table XXXII presents the number of cases in which none of the predictors are selected for each method out of 100 simulations. The LASSO has the highest tendency to filter out influential predictors when their signals are weak, which is followed by Elastic net. There is no case in which the SIS combined methods choose no predictors, however, their sensitivity measures are not as good as those of COLRNS in both scenarios. Group LASSO always selects some of variables in Scenario 3, but when the signals of critical predictors are more sparse and weaker, none are chosen with high probability in Scenario 4. It results in the largest variability in its variable selection performance.

We remove the cases where no variables are chosen for drawing the plots of FDR since those cases turn out invalid values with the denominator equal to zero. The COLRNS has the least level of FNR as Elastic net in Scenarios 1 and 2, however, it improves the Q3 of FDR compared to Elastic net. In Scenarios 3 and 4 having the predictors of weak signals, COLRNS still presents the smallest FNR measures while showing comparable FDR values with the similar median and mean of FDR in comparison to Elastic net, SCAD, and MCP.

TABLE XXXII: THE CASES OF NONE OF THE PREDICTORS ARE CHOSEN

	Scenario 3								
	COLRNS	LASSO	ELNET	SCAD	MCP	S-LASSO	S-SCAD	S-MCP	G-LASSO
Frequency	4	34	31	4	6	0	0	0	0
	Scenario 4								
	COLRNS	LASSO	ELNET	SCAD	MCP	S-LASSO	S-SCAD	S-MCP	G-LASSO
Frequency	6	52	38	12	20	0	0	0	33

TABLE XXXIII: PREDICTION ACCURACY IN SCENARIO 1

	COLRNS	LASSO	ELNET	SCAD	MCP	S-SCAD	S-MCP	S-LASSO	G-LASSO
Q1	0.788	0.792	0.784	0.780	0.772	0.780	0.763	0.788	0.776
Median	0.804	0.804	0.804	0.796	0.790	0.794	0.780	0.800	0.792
Mean	0.806	0.802	0.804	0.795	0.790	0.797	0.778	0.801	0.792
Q3	0.824	0.816	0.821	0.816	0.808	0.812	0.800	0.820	0.809

TABLE XXXIV: RMSE OF PARAMETER ESTIMATION IN SCENARIO 1

	COLRNS	LASSO	ELNET	SCAD	MCP	S-SCAD	S-MCP	S-LASSO	G-LASSO
RMSE	1.05	1.08	1.09	1.38	1.50	1.08	1.14	1.27	1.21

TABLE XXXV: VARIABLE SELECTION MEASUREMENTS IN SCENARIO 1

	COLRNS	LASSO	ELNET	SCAD	MCP	S-SCAD	S-MCP	S-LASSO	G-LASSO
					Sensitivity				
Q1	0.889	0.778	0.889	0.667	0.556	0.667	0.444	0.444	1.000
Median	0.889	0.889	0.889	0.778	0.667	0.778	0.444	0.444	1.000
Mean	0.902	0.837	0.924	0.797	0.622	0.787	0.498	0.498	1.000
Q3	1.000	0.889	1.000	0.889	0.778	0.889	0.556	0.556	1.000
					Specificity				
Q1	0.524	0.714	0.381	0.762	0.857	0.762	0.905	0.714	0.000
Median	0.619	0.786	0.571	0.810	0.905	0.810	0.952	0.810	0.000
Mean	0.601	0.777	0.541	0.796	0.884	0.812	0.940	0.782	0.000
Q3	0.714	0.857	0.714	0.857	0.952	0.857	1.000	0.810	0.000
					False Discovery Rate				
Q1	0.435	0.333	0.429	0.273	0.167	0.273	0.000	0.333	0.700
Median	0.500	0.364	0.513	0.364	0.293	0.364	0.200	0.364	0.700
Mean	0.494	0.372	0.508	0.352	0.275	0.343	0.211	0.368	0.700
Q3	0.556	0.442	0.602	0.442	0.400	0.417	0.333	0.429	0.700
					False Negative Rate				
Q1	0.000	0.111	0.000	0.111	0.222	0.111	0.444	0.111	0.000
Median	0.111	0.111	0.111	0.222	0.333	0.222	0.556	0.222	0.000
Mean	0.098	0.163	0.076	0.203	0.378	0.213	0.502	0.176	0.000
Q3	0.111	0.222	0.111	0.333	0.444	0.333	0.556	0.222	0.000

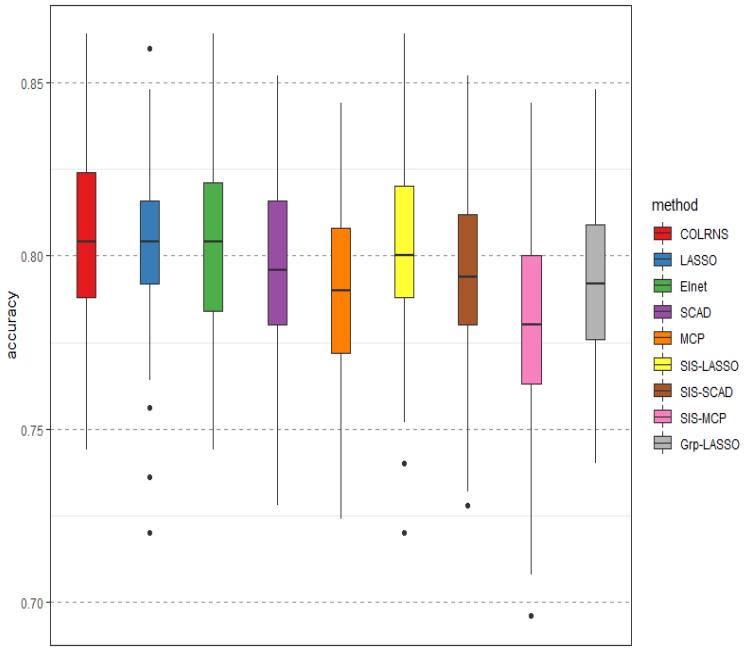


Figure 32: Prediction accuracy in scenario 1

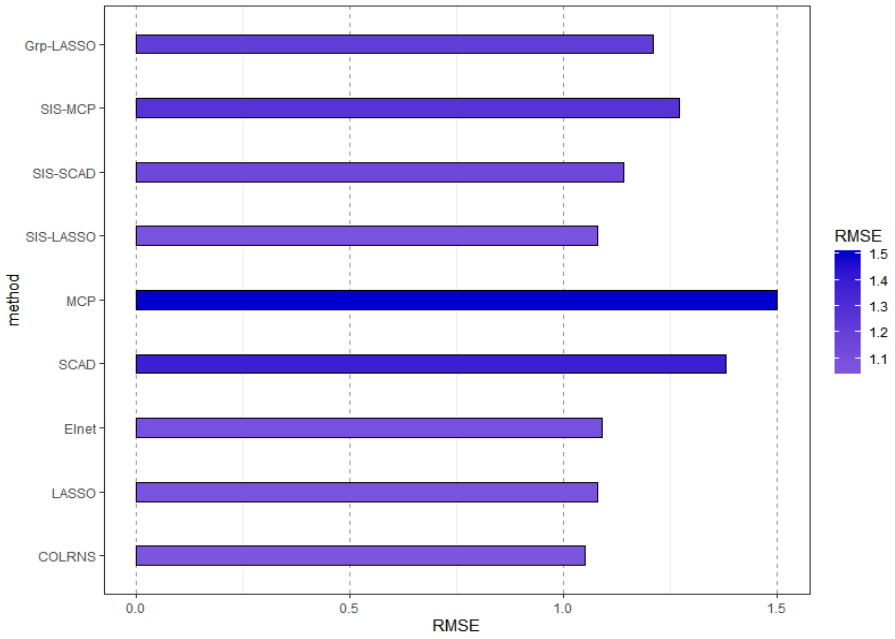
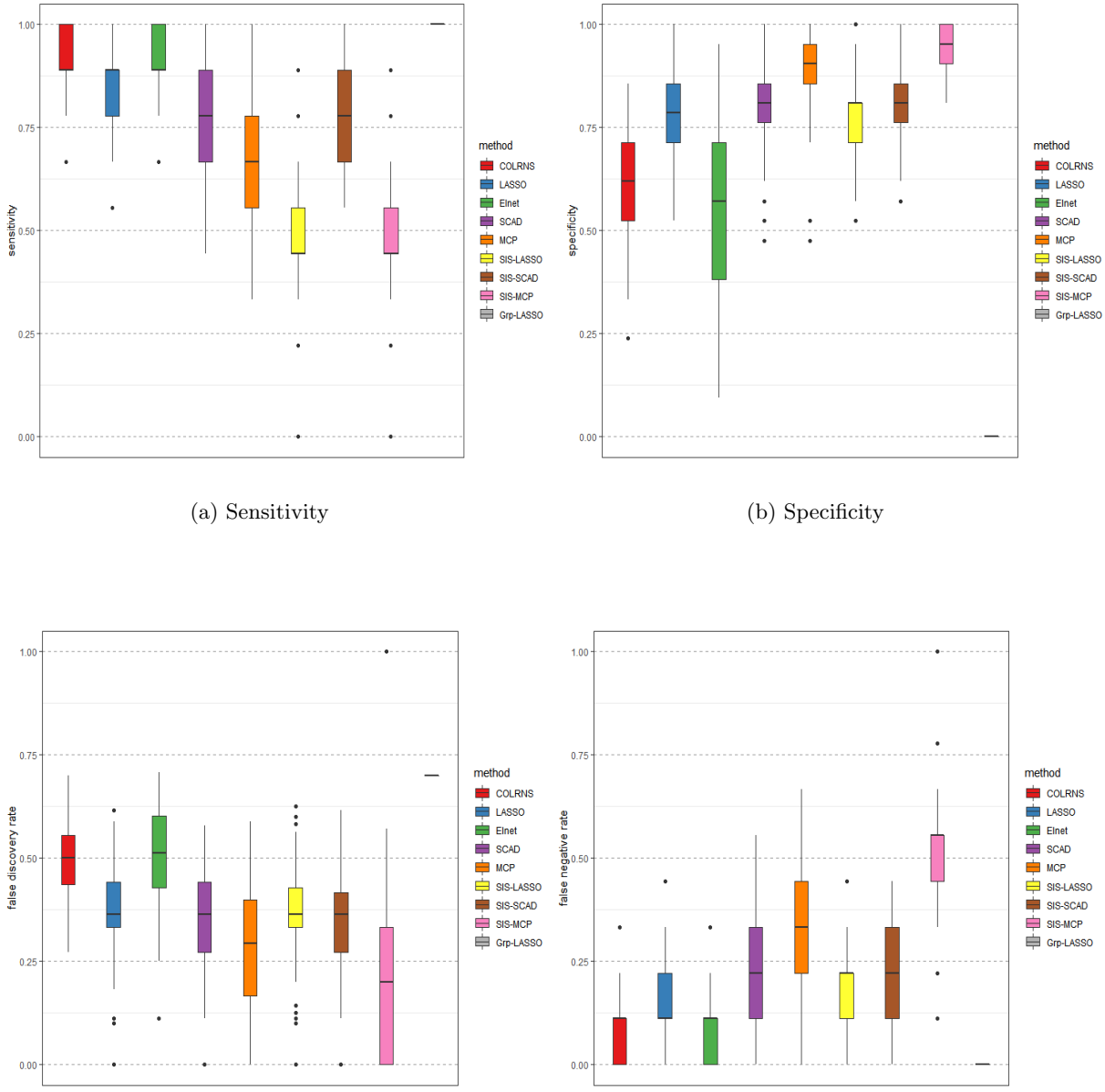


Figure 33: RMSE of parameter estimation in scenario 1



(a) Sensitivity

(b) Specificity

(c) False discovery rate

(d) False negative rate

Figure 34: Variable selection measurements in scenario 1

TABLE XXXVI: PREDICTION ACCURACY IN SCENARIO 2

	COLRNS	LASSO	ELNET	SCAD	MCP	S-SCAD	S-MCP	S-LASSO	G-LASSO
Q1	0.724	0.724	0.728	0.724	0.716	0.719	0.707	0.727	0.720
Median	0.744	0.740	0.744	0.736	0.732	0.736	0.724	0.740	0.732
Mean	0.746	0.741	0.745	0.741	0.736	0.737	0.725	0.740	0.735
Q3	0.764	0.748	0.761	0.757	0.753	0.749	0.748	0.752	0.753

TABLE XXXVII: RMSE OF PARAMETER ESTIMATION IN SCENARIO 2

	COLRNS	LASSO	ELNET	SCAD	MCP	S-SCAD	S-MCP	S-LASSO	G-LASSO
RMSE	0.89	0.90	0.92	1.04	1.12	0.89	0.93	0.98	1.02

TABLE XXXVIII: VARIABLE SELECTION MEASUREMENTS IN SCENARIO 2

	COLRNS	LASSO	ELNET	SCAD	MCP	S-SCAD	S-MCP	S-LASSO	G-LASSO
					Sensitivity				
Q1	0.833	0.833	0.833	0.792	0.500	0.667	0.500	0.500	1.000
Median	0.917	0.833	1.000	0.833	0.667	0.833	0.500	0.500	1.000
Mean	0.907	0.852	0.947	0.828	0.675	0.808	0.537	0.537	1.000
Q3	1.000	1.000	1.000	1.000	0.833	0.833	0.667	0.667	1.000
					Specificity				
Q1	0.583	0.792	0.500	0.750	0.875	0.792	0.917	0.792	0.000
Median	0.667	0.833	0.646	0.833	0.917	0.833	0.958	0.833	0.000
Mean	0.674	0.825	0.615	0.821	0.912	0.842	0.957	0.820	0.000
Q3	0.750	0.875	0.750	0.875	0.958	0.917	1.000	0.875	0.000
					False Discovery Rate				
Q1	0.538	0.375	0.500	0.375	0.167	0.333	0.000	0.375	0.800
Median	0.583	0.444	0.600	0.444	0.333	0.444	0.250	0.455	0.800
Mean	0.572	0.435	0.578	0.440	0.309	0.422	0.224	0.440	0.800
Q3	0.647	0.500	0.671	0.538	0.429	0.500	0.400	0.500	0.800
					False Negative Rate				
Q1	0.000	0.000	0.000	0.000	0.167	0.167	0.333	0.000	0.000
Median	0.083	0.167	0.000	0.167	0.333	0.167	0.500	0.167	0.000
Mean	0.093	0.148	0.053	0.172	0.325	0.192	0.463	0.147	0.000
Q3	0.167	0.167	0.167	0.208	0.500	0.333	0.500	0.167	0.000

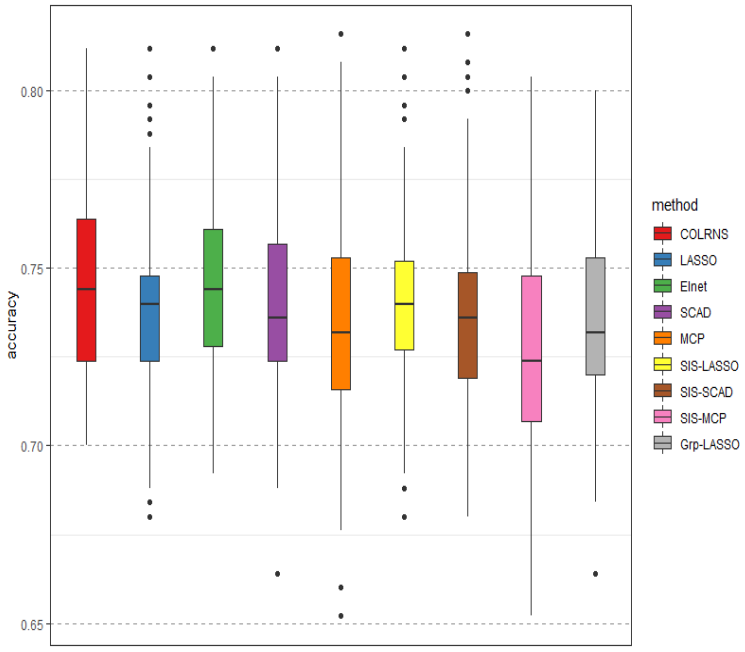


Figure 35: Prediction accuracy in scenario 2

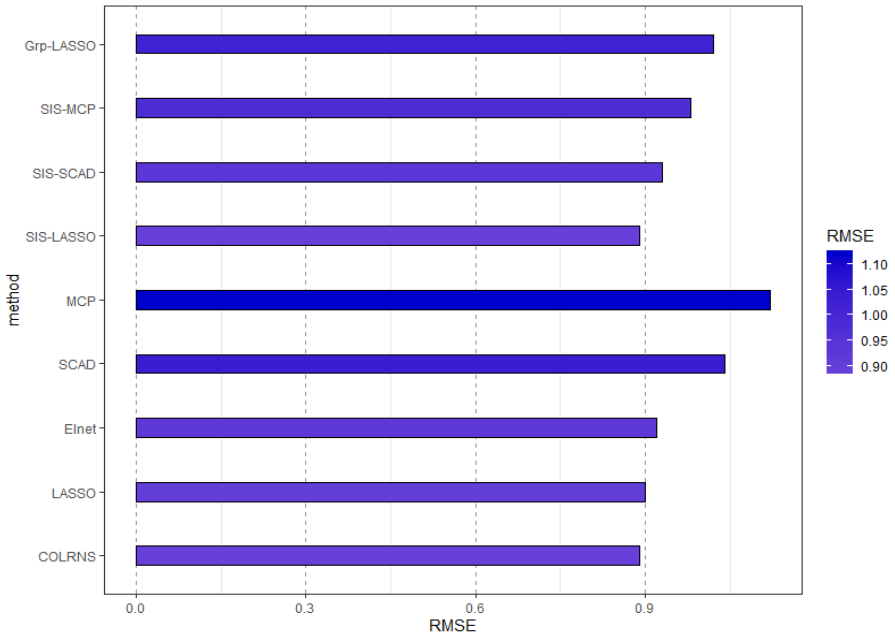
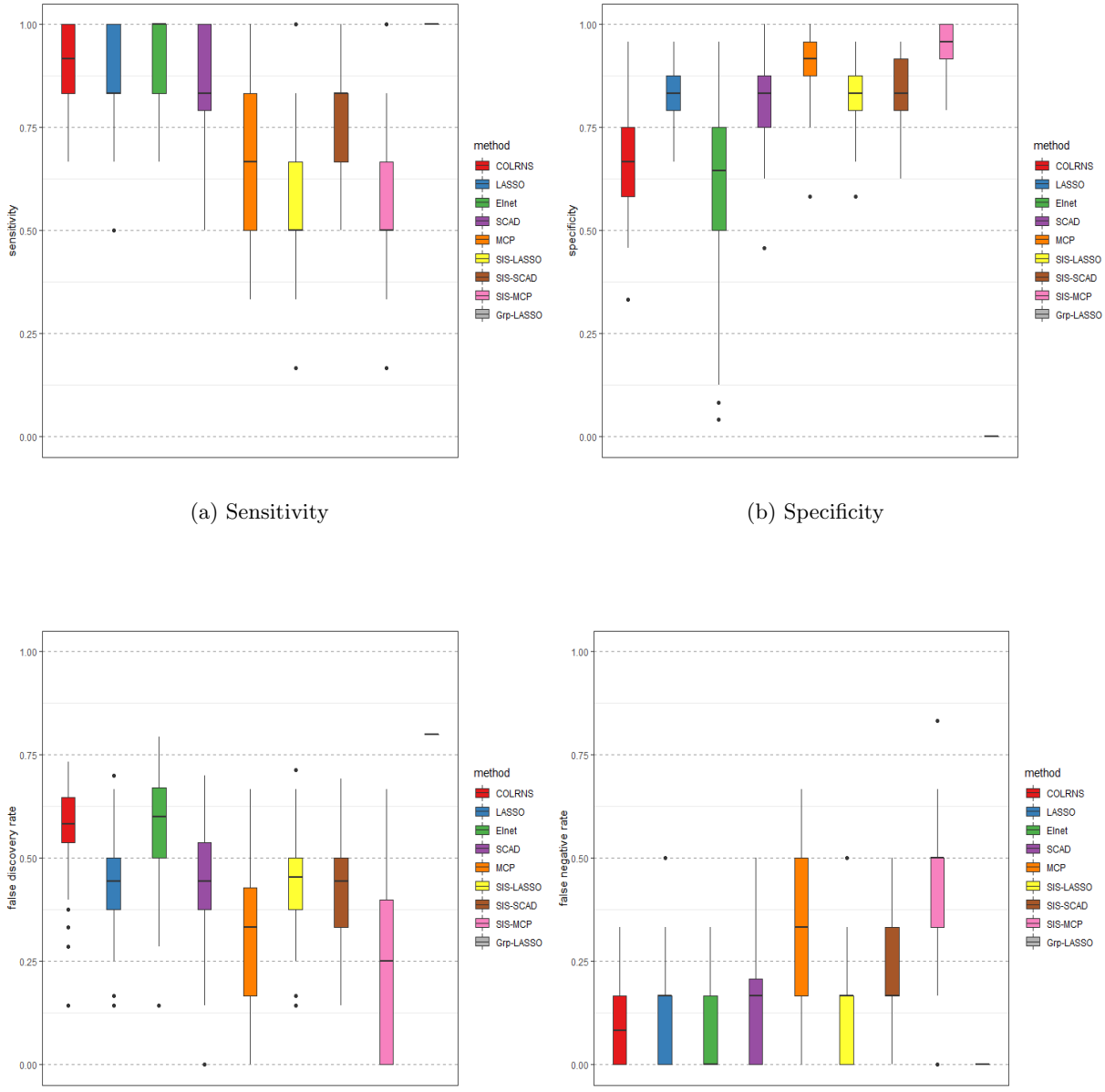


Figure 36: RMSE of parameter estimation in scenario 2



(a) Sensitivity

(b) Specificity

(c) False discovery rate

(d) False negative rate

Figure 37: Variable selection measurements in scenario 2

TABLE XXXIX: PREDICTION ACCURACY IN SCENARIO 3

	COLRNS	LASSO	ELNET	SCAD	MCP	S-SCAD	S-MCP	S-LASSO	G-LASSO
Q1	0.544	0.512	0.516	0.544	0.535	0.516	0.504	0.516	0.516
Median	0.566	0.540	0.544	0.564	0.564	0.548	0.534	0.548	0.542
Mean	0.566	0.541	0.543	0.564	0.559	0.542	0.533	0.547	0.540
Q3	0.585	0.568	0.576	0.584	0.584	0.568	0.564	0.577	0.572

TABLE XL: RMSE OF PARAMETER ESTIMATION IN SCENARIO 3

	COLRNS	LASSO	ELNET	SCAD	MCP	S-SCAD	S-MCP	S-LASSO	G-LASSO
RMSE	0.26	0.21	0.20	0.62	0.70	0.23	0.21	0.22	0.52

TABLE XLI: VARIABLE SELECTION MEASUREMENTS IN SCENARIO 3

	COLRNS	LASSO	ELNET	SCAD	MCP	S-SCAD	S-MCP	S-LASSO	G-LASSO
					Sensitivity				
Q1	0.133	0.000	0.000	0.133	0.067	0.000	0.000	0.000	0.333
Median	0.133	0.067	0.133	0.133	0.067	0.067	0.067	0.067	0.667
Mean	0.199	0.071	0.176	0.162	0.099	0.080	0.044	0.044	0.670
Q3	0.267	0.133	0.267	0.200	0.133	0.133	0.067	0.067	1.000
					Specificity				
Q1	0.783	0.933	0.800	0.800	0.867	0.933	0.933	0.867	0.000
Median	0.867	1.000	0.933	0.867	0.933	0.933	1.000	0.933	0.333
Mean	0.837	0.947	0.855	0.873	0.929	0.945	0.979	0.927	0.330
Q3	0.933	1.000	1.000	0.933	1.000	1.000	1.000	1.000	0.667
					False Discovery Rate				
Q1	0.333	0.050	0.333	0.277	0.000	0.000	0.000	0.000	0.500
Median	0.453	0.333	0.462	0.437	0.500	0.333	0.000	0.333	0.500
Mean	0.428	0.400	0.441	0.430	0.390	0.321	0.202	0.342	0.500
Q3	0.556	0.500	0.545	0.600	0.593	0.500	0.375	0.518	0.500
					False Negative Rate				
Q1	0.733	0.867	0.733	0.800	0.867	0.867	0.933	0.867	0.000
Median	0.867	0.933	0.867	0.867	0.933	0.933	0.933	0.933	0.333
Mean	0.801	0.929	0.824	0.838	0.901	0.920	0.956	0.907	0.330
Q3	0.867	1.000	1.000	0.867	0.933	1.000	1.000	1.000	0.667

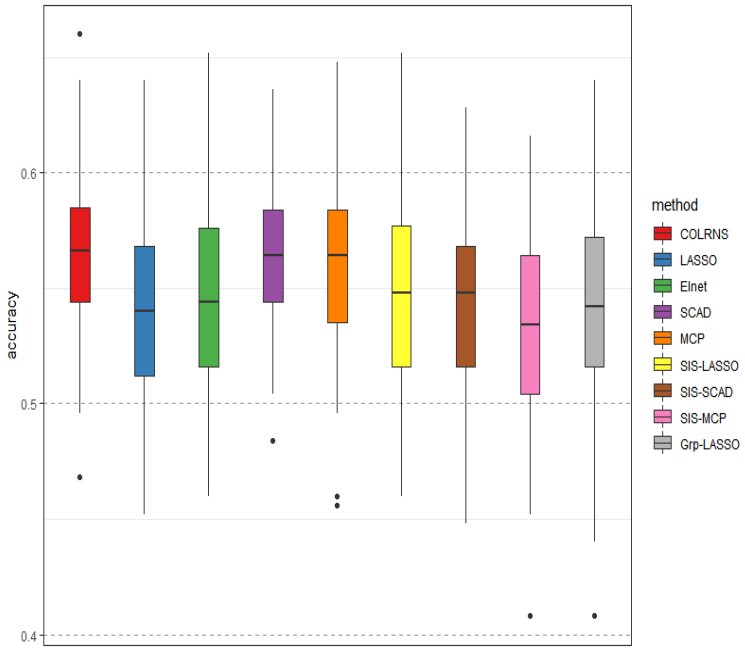


Figure 38: Prediction accuracy in scenario 3

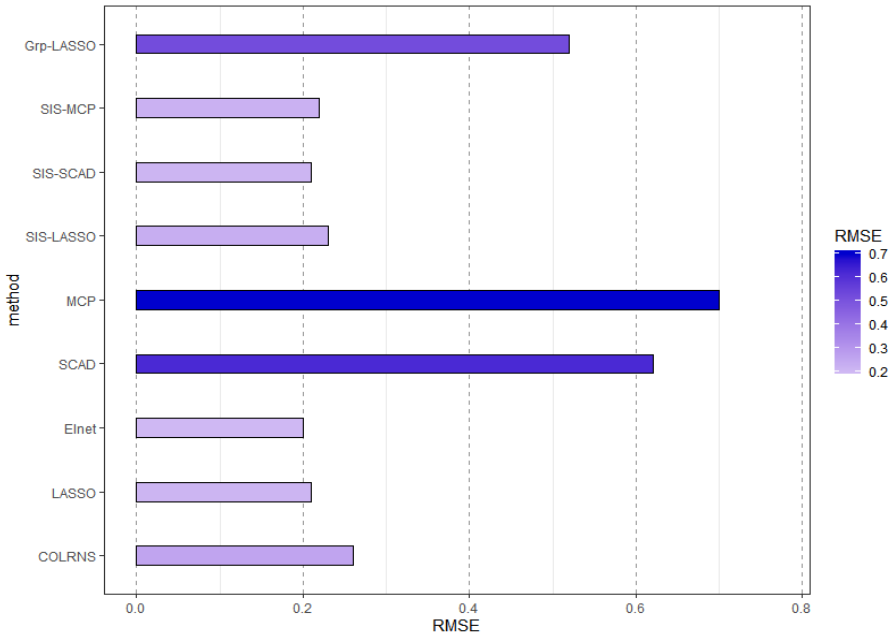
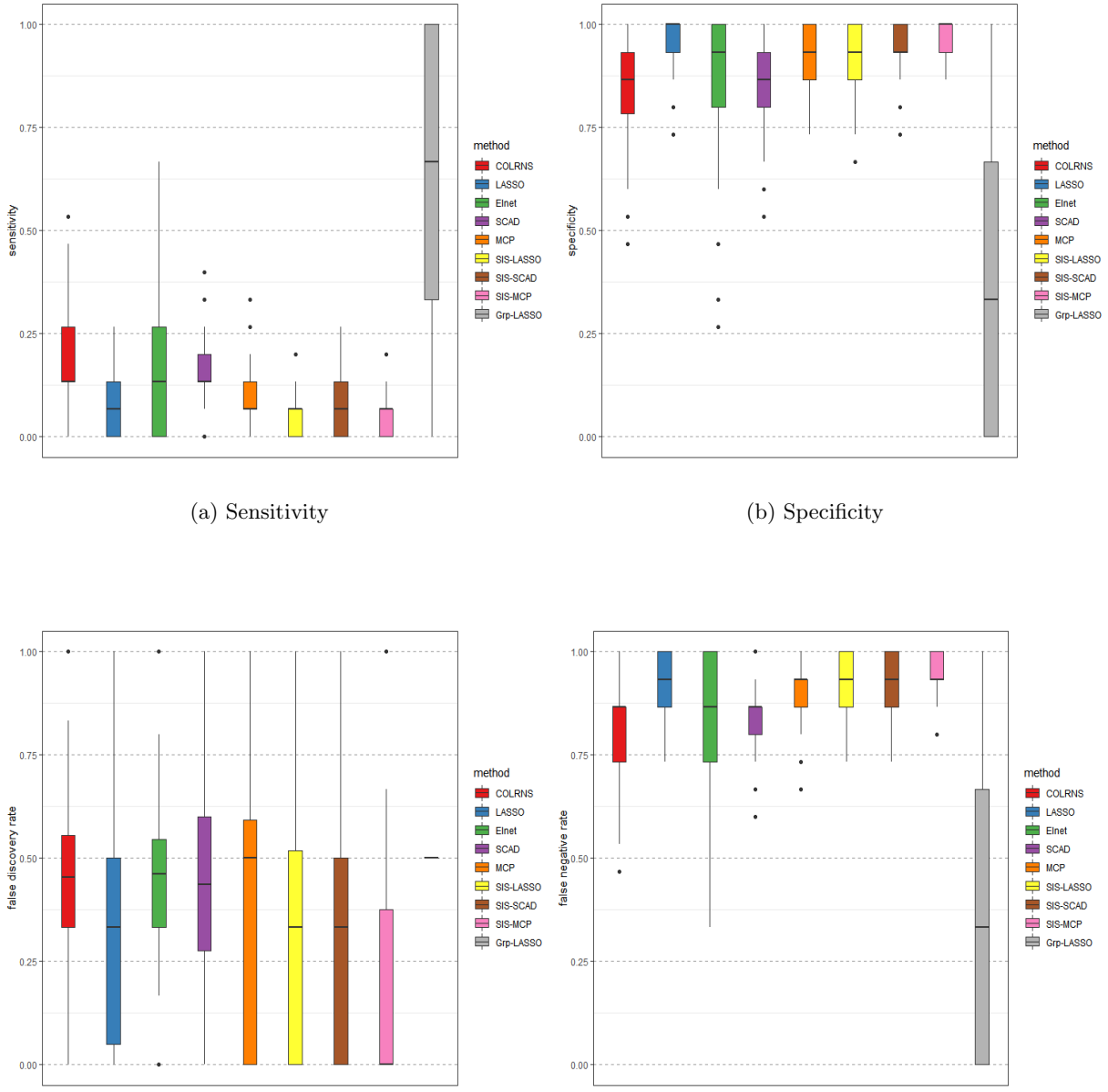


Figure 39: RMSE of parameter estimation in scenario 3



(a) Sensitivity

(b) Specificity

(c) False discovery rate

(d) False negative rate

Figure 40: Variable selection measurements in scenario 3

TABLE XLII: PREDICTION ACCURACY IN SCENARIO 4

	COLRNS	LASSO	ELNET	SCAD	MCP	S-SCAD	S-MCP	S-LASSO	G-LASSO
Q1	0.520	0.488	0.496	0.532	0.516	0.495	0.492	0.496	0.503
Median	0.556	0.520	0.528	0.560	0.548	0.520	0.518	0.522	0.524
Mean	0.556	0.521	0.533	0.558	0.546	0.524	0.522	0.528	0.523
Q3	0.584	0.552	0.568	0.581	0.572	0.546	0.544	0.553	0.548

TABLE XLIII: RMSE OF PARAMETER ESTIMATION IN SCENARIO 4

	COLRNS	LASSO	ELNET	SCAD	MCP	S-SCAD	S-MCP	S-LASSO	G-LASSO
RMSE	0.28	0.25	0.24	0.37	0.41	0.25	0.25	0.25	0.38

TABLE XLIV: VARIABLE SELECTION MEASUREMENTS IN SCENARIO 4

	COLRNS	LASSO	ELNET	SCAD	MCP	S-SCAD	S-MCP	S-LASSO	G-LASSO
					Sensitivity				
Q1	0.067	0.000	0.000	0.067	0.000	0.000	0.000	0.000	0.000
Median	0.133	0.000	0.067	0.100	0.000	0.000	0.000	0.000	0.400
Mean	0.129	0.036	0.085	0.103	0.047	0.038	0.017	0.017	0.408
Q3	0.200	0.067	0.133	0.133	0.067	0.067	0.000	0.000	0.650
					Specificity				
Q1	0.847	0.976	0.906	0.906	0.965	0.976	0.988	0.965	0.350
Median	0.924	1.000	0.976	0.929	0.976	0.988	1.000	0.988	0.600
Mean	0.902	0.984	0.945	0.935	0.974	0.983	0.992	0.978	0.592
Q3	0.965	1.000	1.000	0.965	0.988	1.000	1.000	1.000	1.000
					False Discovery Rate				
Q1	0.701	0.500	0.704	0.692	0.600	0.000	0.000	0.000	0.850
Median	0.800	0.667	0.778	0.778	0.750	0.500	0.000	0.612	0.850
Mean	0.775	0.701	0.756	0.772	0.758	0.475	0.349	0.486	0.850
Q3	0.887	1.000	0.851	0.875	1.000	0.839	1.000	0.800	0.850
					False Negative Rate				
Q1	0.800	0.933	0.867	0.867	0.933	0.933	1.000	0.933	0.350
Median	0.867	1.000	0.933	0.900	1.000	1.000	1.000	1.000	0.600
Mean	0.871	0.964	0.915	0.897	0.953	0.962	0.983	0.956	0.592
Q3	0.933	1.000	1.000	0.933	1.000	1.000	1.000	1.000	1.000

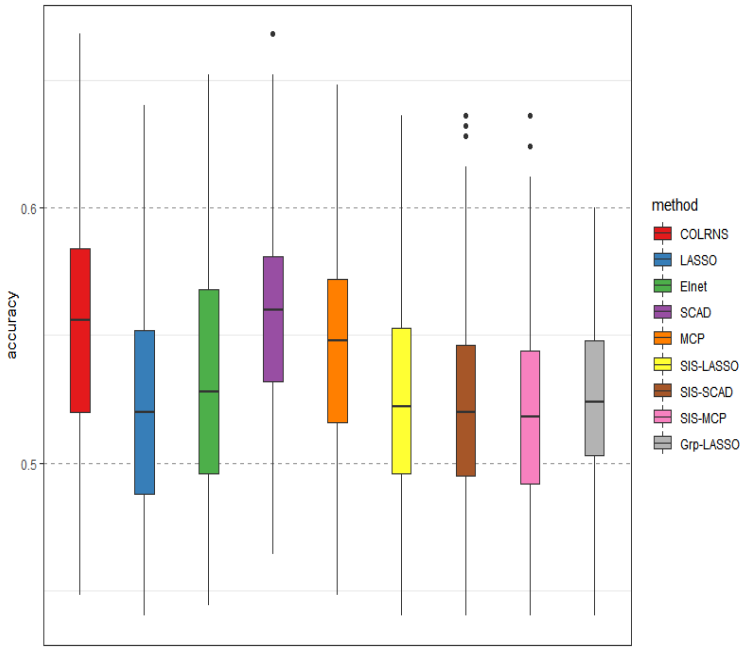


Figure 41: Prediction accuracy in scenario 4

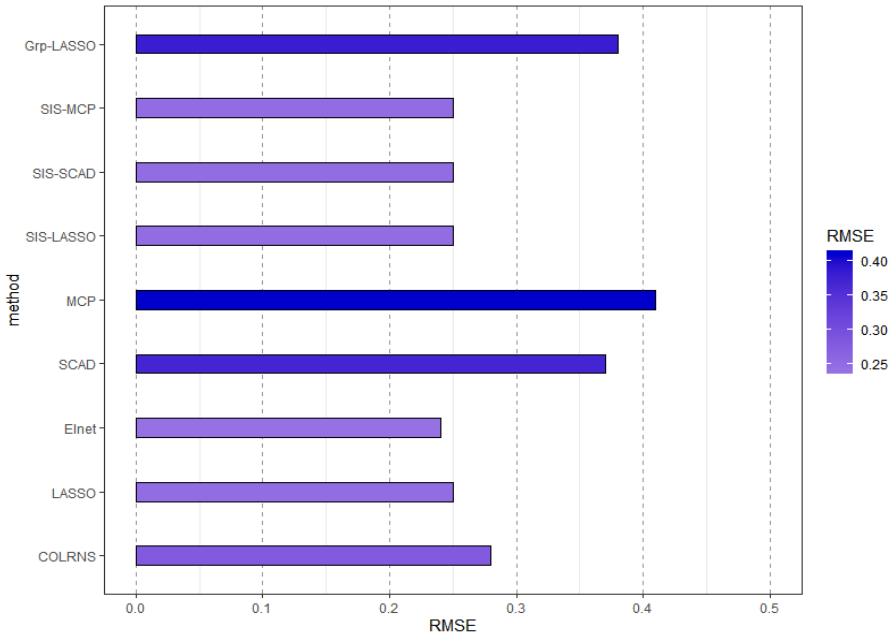
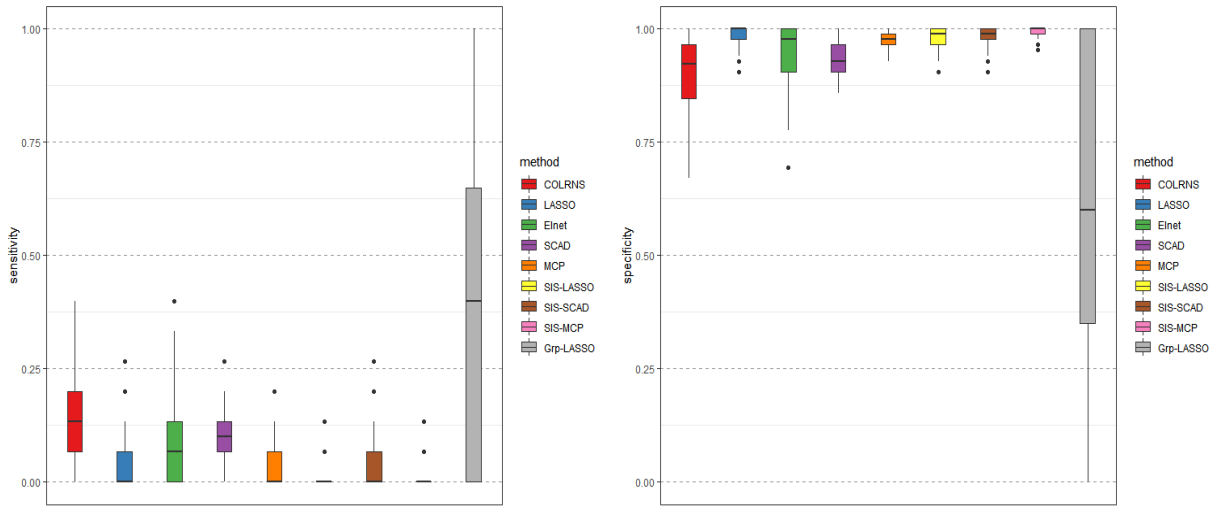
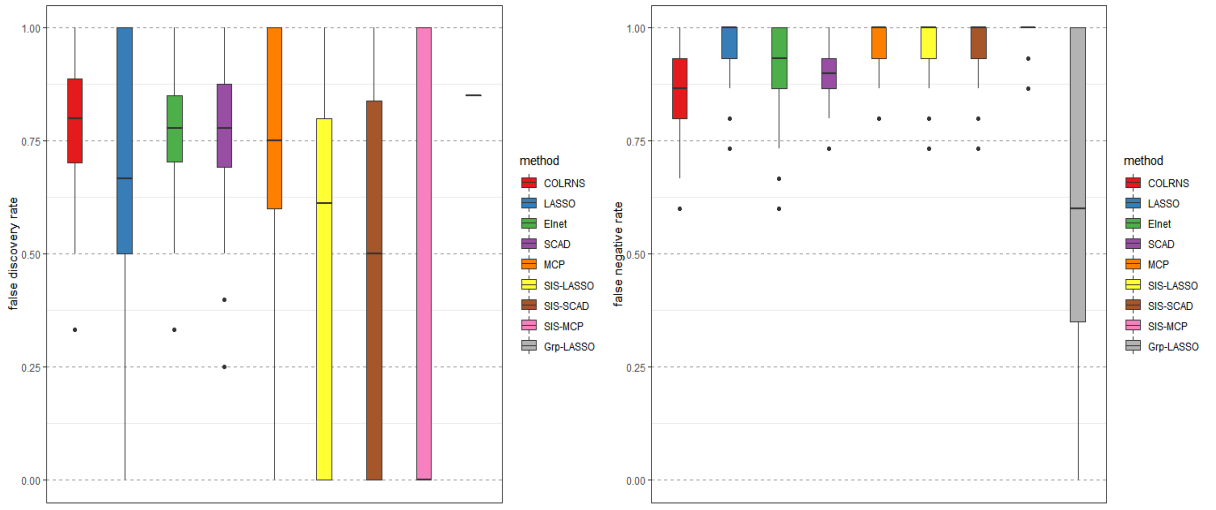


Figure 42: RMSE of parameter estimation in scenario 4



(a) Sensitivity

(b) Specificity



(c) False discovery rate

(d) False negative rate

Figure 43: Variable selection measurements in scenario 4

4.4 Data Application: Great Lakes Fish Consumer Study

4.4.1 Introduction

Consumption of fish is a source of exposure to environmental mixtures including persistent organic pollutants such as *polychlorinated biphenyls* (PCBs), *p,p'-diphenyldichloroethene* (DDE), and *polybrominated diphenyl ethers* (PBDEs). These pollutants have been shown to potentially cause carcinogenic effects increase, neurologic disorders, and endocrine homeostasis disruption [Hanrahan et al., 1999]. The Great Lakes Fish Consumer Study (GLFCS) follows a cohort of $4,206$ frequent and infrequent sport fish consumers collecting longitudinal data since 1994 as part of a health assessment consortium of the Health Departments of Wisconsin, Illinois, Indiana, Ohio, and Michigan [Hanrahan et al., 1999]. The data include information of PCB, DDE, and PBDE biomarkers, fish consumption, health status and extensive measures of endocrine and metabolic function of fish consumers. These data have been analyzed using methods that employ summation of the exposures by chemical group, i.e., PCBs or PBDEs, as well as analysis of some individual chemicals [Turyk et al., 2009].

4.4.2 Statistical Analysis

We investigate associations between environmental mixtures and diabetes incidence in the GLFCS cohort. The 598 GLFCS participants donated blood samples: 91 participants in 2001-2003 and 507 in 2004-2005. We study the endpoint of diagnosed diabetes which is recorded as a binary indicator 0 (No diabetes) or 1 (Diabetes). The available exposure measurements in the serum samples contain PCBs (24 congeners), DDE, and PBDEs (6 congeners). The environmental exposure measurements that are below the Level of Detection (LOD) are imputed

as LOD divided by the square root of 2 in accordance with the approach taken by the NHANES. The details of exposure assessment are described in Hanrahan et al. [1999].

We consider 508 participants with non-misssing data on environmental exposure variables. We exclude three exposures (PBDE49, PBDE85, and PCB128) which show strongly skewed distributions. The remaining exposure measurements are (natural) log-transformed. The distribution of the log-transformed exposure measurements are shown in Figure 45 and Figure 46. We observed moderate to strong positive pairwise correlations among the PBDEs and PCBs, however, the correlation between the PBDE and the PCB group is found to be negative (Figure 1). The reason for the negative correlation between the two groups may be caused by different exposures sources: PCBs from food and fish, and PBDEs from consumer products in the home.

Among the 508 participants, 64 people (13%) have diagnosed diabetes. Almost 70% of the participants are male (352 male, 156 female). We analyze these data by applying the following comparator methods: COLRNS-GLM, LASSO, Elastic net, SCAD, and MCP. The following variables are taken as confounders adjusted in the analysis model: participants' gender, age, BMI, and the level of serum lipids. The distributions of the continuous confounders are shown in Figure 44. The predictors and confounders are scaled to have mean 0 and variance 1 for analysis. All 28 environmental exposures (22 PCBs, 1 DDE, 5 PBDEs) and 4 confounders are penalized by the selection methods to fit the model. We also fit a generalized linear model (with regularization) for comparison purpose.

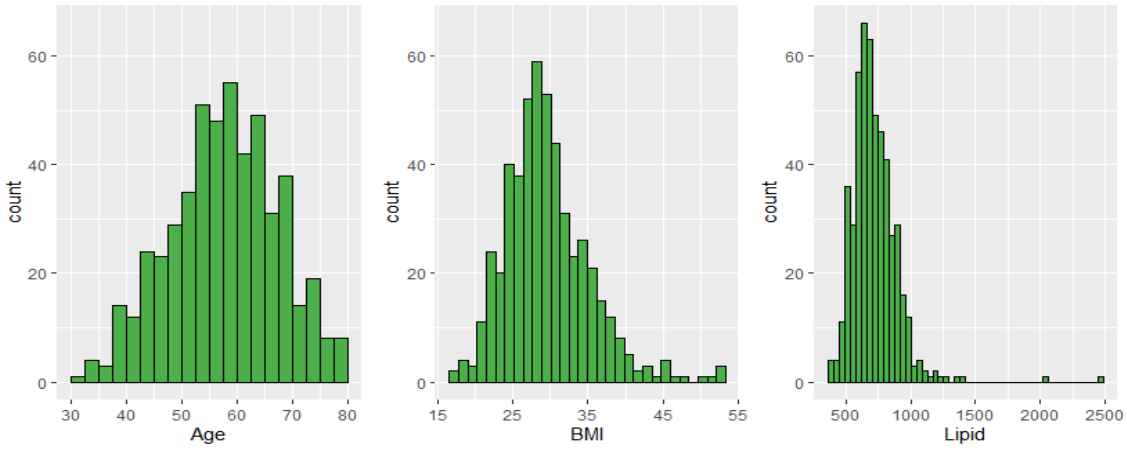


Figure 44: Histogram of continuous confounders

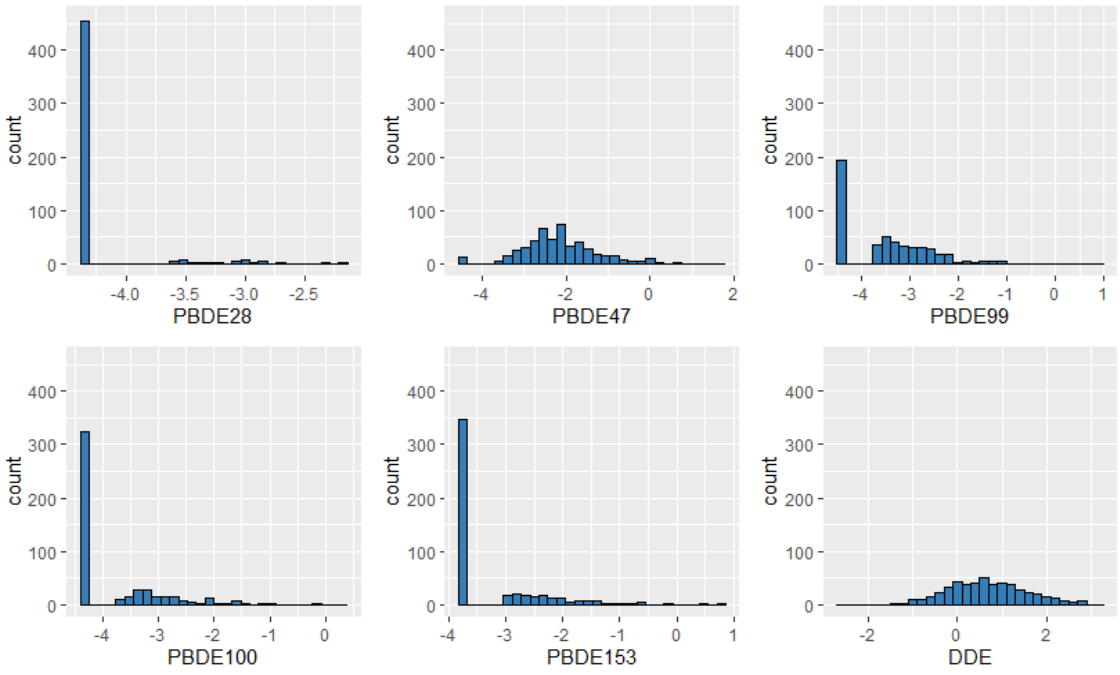


Figure 45: Histogram of log-transformed PBDEs and DDE exposures

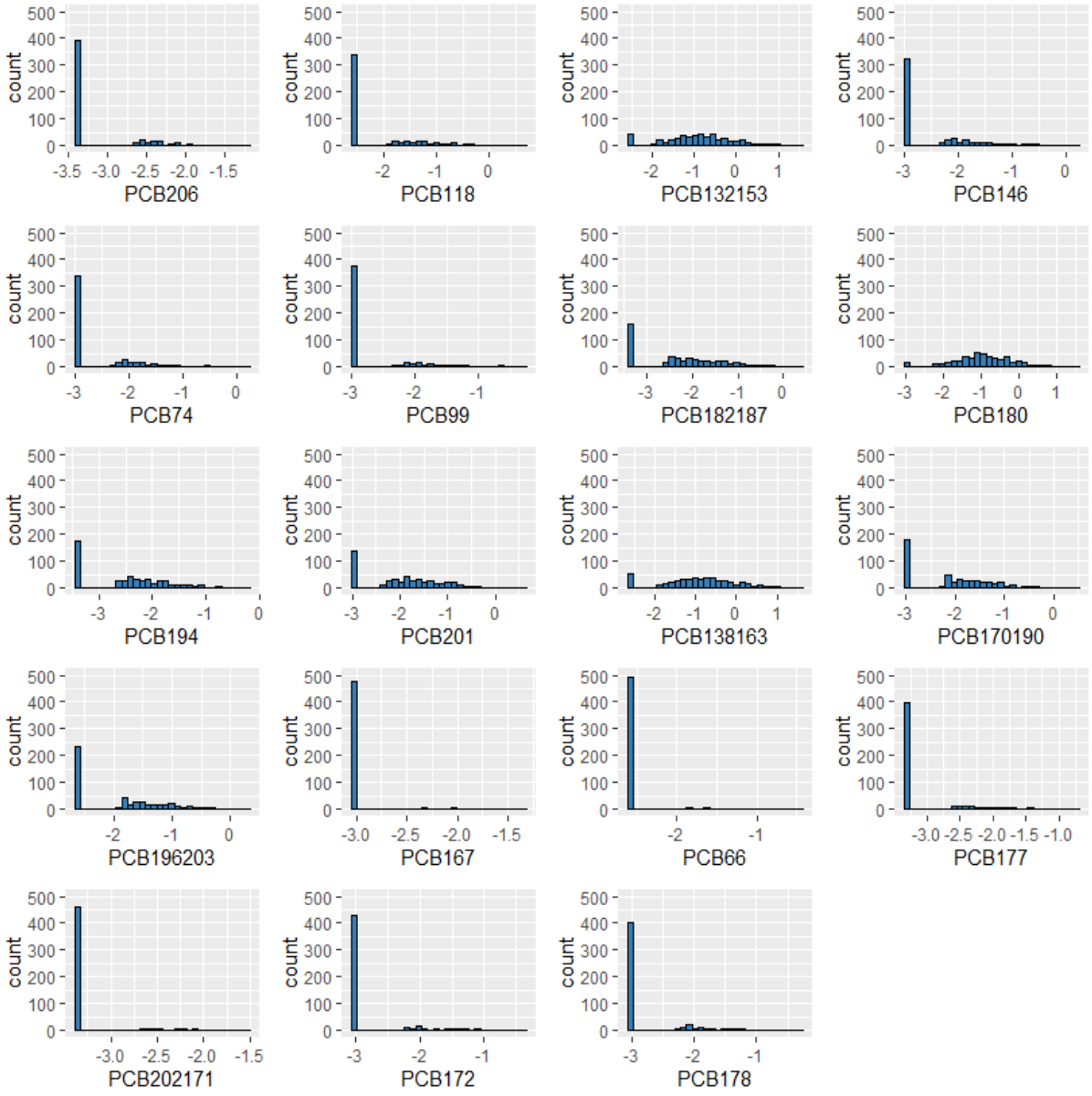


Figure 46: Histogram of log-transformed PCBs exposures

4.4.3 Analysis Results

We investigate the degree of collinearity in the 28 environmental exposures by estimating the condition number from the correlation matrix of the exposures. The condition number obtained by the square root of the ratio of the maximum to minimum eigenvalues gives 42.526. It indicates that there exists strong collinearity among the predictors in the data. While usual (not regularized) generalized linear model analysis in Table XLV finds that among the 28 environmental exposures, only PBDE100 is significantly associated with diabetes incidence, this analysis may have been affected by this high collinearity and the resulting inflated standard errors.

Table XLV shows that regularized LASSO, Elastic net, and SCAD methods all select the same set of exposures: PBDE153, PCB167, DDE. Increase of PBDE153 is negatively associated with diabetes and increases to PCB167 and DDE are associated positively with diabetes. The absolute values of the estimated coefficients are highest for Elastic net (PBDE153 $\beta = -0.0878$, PCB167 $\beta = 0.1117$, DDE $\beta = 0.1824$) followed by LASSO (PBDE153 $\beta = -0.0807$, PCB167 $\beta = 0.1108$, DDE $\beta = 0.1710$) and SCAD (PBDE153 $\beta = -0.0097$, PCB167 $\beta = 0.1069$, DDE $\beta = 0.0191$). The MCP only chooses PCB167 as a significant exposure having positive association with diabetes (PCB167 $\beta = 0.1291$).

The COLRNS-GLM selects the three exposures chosen by LASSO, Elastic net, and SCAD (PBDE153 $\beta = -0.1301$, PCB167 $\beta = 0.1353$, DDE $\beta = 0.2221$) and one additional exposure (PCB196203 $\beta = -0.0424$). The absolute coefficients of the three common exposures are greater than those from Elastic net. The PCB167 and DDE are positively associated with diabetes

by having odds ratios $e^{0.1353} = 1.145$ and $e^{0.2221} = 1.249$, respectively. That is, one standard deviation increase to PCB167 raises odds of diabetes about 15% and it becomes higher about 25% for the increase of DDE.

For confounders, COLRNS-GLM, LASSO, Elastic net choose all four confounders age, sex female, BMI, serum lipids, however, SCAD and MCP only select age and BMI. The directions of association are the same across the five selection methods. Increased age and BMI show positive associations with diabetes, however, being female and higher serum lipids are negatively associated with diabetes. For COLRNS, the odds ratio of diabetes is $e^{0.5579} = 1.747$ and $e^{0.5837} = 1.793$ for one standard deviation increase of age and BMI, respectively. For being female, the odds ratio is $e^{-0.0986} = 0.906$ which means the odds of diabetes is about 9% decreased for female than male.

TABLE XLV: RESULTS OF GENERALIZED LINEAR REGRESSION

Variable	Estimate	z-value	p-value
Intercept	-2.61	-11.89	0.000
Age	0.64	3.08	0.002
Sex (female)	-0.23	-1.06	0.290
BMI	0.73	4.12	0.000
Lipid	-0.08	-0.39	0.699
PBDE28	-0.24	-1.17	0.241
PBDE47	-0.48	-1.57	0.115
PBDE99	0.02	0.09	0.925
PBDE100	0.61	1.82	0.069
PBDE153	-0.31	-1.26	0.207
PCB206	0.00	-0.01	0.991
PCB118	0.33	0.94	0.348
PCB132153	-1.96	-1.08	0.282
PCB146	-0.45	-0.87	0.385
PCB74	-0.17	-0.48	0.635
PCB99	-0.03	-0.07	0.948
PCB182187	-0.54	-0.85	0.396
PCB180	0.58	0.60	0.549
PCB194	-0.17	-0.26	0.793
PCB201	1.05	1.49	0.136
PCB138163	1.99	1.22	0.223
PCB170190	-0.63	-1.00	0.319
PCB196203	-0.38	-0.70	0.487
PCB167	0.39	1.40	0.161
PCB66	-0.08	-0.56	0.576
PCB177	0.19	0.31	0.755
PCB202171	0.05	0.12	0.904
PCB172	-0.24	-0.45	0.653
PCB178	0.28	0.50	0.618
PCB183	0.07	0.09	0.927
PCB193	0.17	0.36	0.722
PCB208195	-0.37	-1.39	0.164
DDE	0.40	1.47	0.143

TABLE XLVI: REGRESSION COEFFICIENTS FROM THE PENALIZED VARIABLE SELECTION METHODS

	COLRNS	LASSO	ELNET	SCAD	MCP
Intercept	-2.2779	-2.2170	-2.2067	-2.3496	-2.3509
Age	0.5579	0.5208	0.4908	0.7906	0.7941
Sex (female)	-0.0986	-0.0445	-0.0635	.	.
BMI	0.5837	0.5503	0.5313	0.7246	0.7266
Lipid	-0.0777	-0.0376	-0.0502	.	.
PBDE28	.	.	.	.	.
PBDE47	.	.	.	.	.
PBDE99	.	.	.	.	.
PBDE100	.	.	.	.	.
PBDE153	-0.1301	-0.0807	-0.0878	-0.0097	.
PCB206	.	.	.	.	.
PCB118	.	.	.	.	.
PCB132153	.	.	.	.	.
PCB146	.	.	.	.	.
PCB74	.	.	.	.	.
PCB99	.	.	.	.	.
PCB182187	.	.	.	.	.
PCB180	.	.	.	.	.
PCB194	.	.	.	.	.
PCB201	.	.	.	.	.
PCB138163	.	.	.	.	.
PCB170190	.	.	.	.	.
PCB196203	-0.0424	.	.	.	.
PCB167	0.1353	0.1108	0.1117	0.1069	0.1291
PCB66	.	.	.	.	.
PCB177	.	.	.	.	.
PCB202171	.	.	.	.	.
PCB172	.	.	.	.	.
PCB178	.	.	.	.	.
PCB183	.	.	.	.	.
PCB193	.	.	.	.	.
PCB208195	.	.	.	.	.
DDE	0.2221	0.1710	0.1824	0.0191	.

4.5 Cross-validation Study

4.5.1 Methods

We conduct cross-validation study with the GLFCS data following an approach similar to section 3.7. The goal is to evaluate and compare the prediction performance of the selection methods when the methods are applied in the data with binary outcomes.

We consider nested cross-validation in this study. The data of 508 participants are randomly split into 5 folds with 4 folds being used as training set and the remaining 1 fold used as a test set. Penalized selection methods, COLRNS-GLM, LASSO, Elastic net, SCAD, and MCP are applied in the training set to fit models. The confounders are not forced to be included in the models so that whether to select each of 32 covariate (28 environmental exposures, 4 confounders) is depending on each method. We do not forcibly control for confounders to increase the ability of prediction rather than interpretation of the model in the simulations. The prediction performance of the resulting model is evaluated in the test set by estimating the measure of prediction accuracy we use in Section 4.3.2. The same simulation is repeated by alternating the test and training set. Then the five prediction accuracy measures are averaged. After performing the same simulation procedures 20 times independently, the descriptive statistics of the average prediction accuracy are found for each penalized selection method.

4.5.2 Results

Table XXX presents the descriptive statistics of the average prediction accuracy measures from the simulations. The SCAD and MCP show poor prediction performance by showing the lowest Q1, mean, median, and Q3 values of the average prediction accuracy compared to

the remaining three methods: COLRNS-GLM, LASSO, and Elastic net. The two methods also show very large variance of the average accuracy measures as shown in Figure 47. The COLRNS-GLM has the highest average prediction accuracy as an outlier across all comparing methods. Without the outlier, COLRNS-GLM, LASSO, and Elasitc net indicate comparable performance in prediction, however, COLRNS-GLM still shows the higher mean and Q3 of the averaged prediction accuracy than LASSO and Elastic net.

TABLE XLVII: AVERAGE PREDICTION ACCURACY

	COLRNS	LASSO	ELNET	SCAD	MCP
Q1	0.873966	0.873966	0.873966	0.866084	0.866143
Median	0.874005	0.874005	0.874005	0.868152	0.868132
Mean	0.874099	0.874000	0.874000	0.868582	0.868287
Q3	0.874063	0.874034	0.874034	0.872030	0.870127

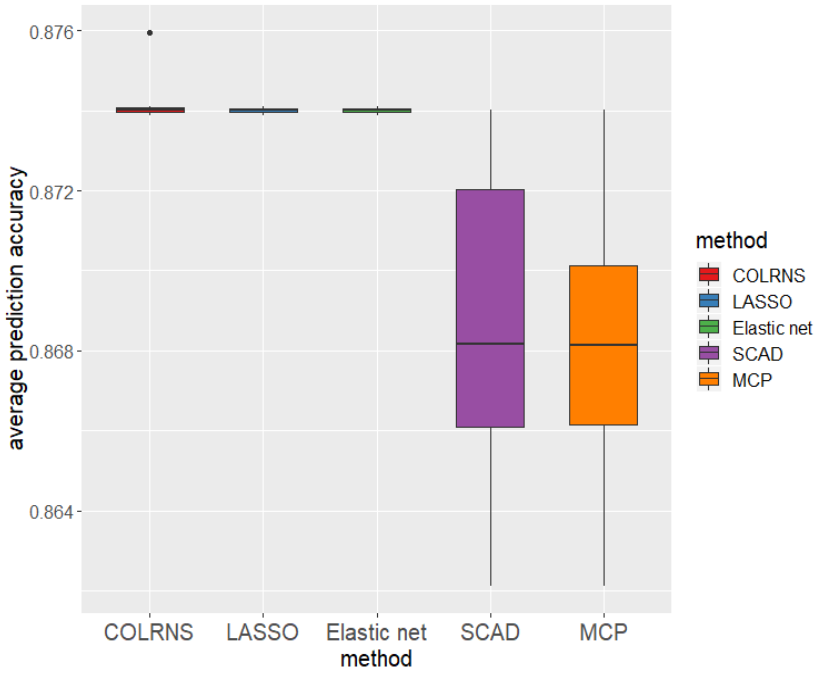


Figure 47: Box plot of average prediction error

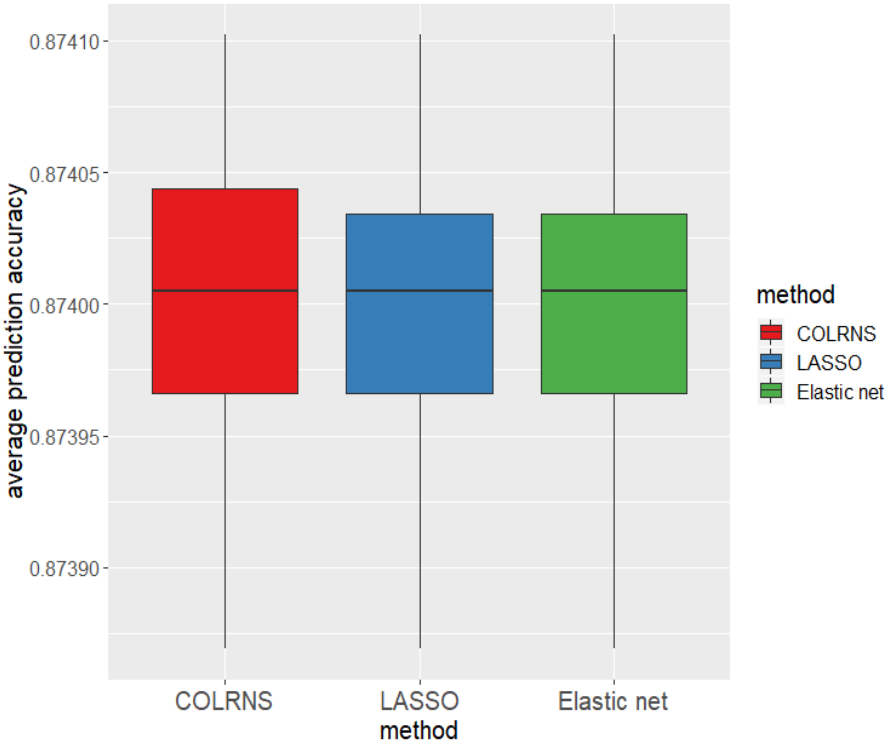


Figure 48: Box plot of average prediction error without an outlier

CHAPTER 5

CONCLUSION AND DISCUSSION

We develop COrrelation LeaRNing for variable Selection (COLRNS) to handle many statistical challenges from unstable estimates of regression parameters caused by high dimensional data analysis that frequently involves strong collinearity among predictors. It performs variable selection using the cluster of correlated predictors that is identified by correlation learning. The cluster based selection is to leverage the advantage of reduced degree of collinearity in the cluster in comparison to that in the entire predictors. We especially use the l_2 penalization which is effective for stabilizing the estimates under strong collinearity by combining it with another penalty to select a subset of variables. The cluster identification and variable selection are iterated to improve selection performance and to decrease squared error loss of the resulting model. As an extension of COLRNS for linear regression setting, COLRNS-GLM is also developed to apply the method to the generalized linear regression setting, especially to deal with the data of dichotomous outcomes.

We investigate the performance of COLRNS and COLRNS-GLM in an extensive set of simulations according to three areas: prediction, variable selection, and parameter estimation. The methods show better performance in predicting of future outcomes in comparison to other penalized variable selection methods. They demonstrate great ability of selecting predictors with true signals by showing high sensitivity and low false negative rate. They also improve the chance of screening out unimportant variables compared to Elastic net when predictors of many

weak effects are strongly correlated in multiple groups, which increases robustness of variable selection. They further show good performance in accurate estimation of regression parameters with low level of error compared to other methods.

The methods COLRNS and COLRNS-GLM are applied to two real-world environmental data sets which are characterized by having strongly correlated high dimensional environmental exposures of weak individual effect. The results indicate that the methods are more selective in choosing key environmental pollutants that are associated with health outcomes compared to other penalized selection methods with lower error of the resulting model. The selection can also be applied to confounders. The results indicate that the confounders considered to affect health outcomes are effectively chosen by the methods. We additionally conduct nested cross-validation studies with the two real-world data sets to examine the out-of-sample prediction performance of the methods. Both COLRNS and COLRNS-GLM present greater performance of prediction than other compared methods by having lower prediction error and higher prediction accuracy, respectively.

Considering that the risk assessment of multichemical exposures such as hazard identification or dose-response assessment are being more widely conducted to address health concerns [Choudhury et al., 2000], the methods can be effectively applied to mixtures studies where we want to identify risk chemicals and assess their relative associations with health outcomes. The estimated models from the methods can also be usefully deployed in prediction or classification of health risk in public health studies. Moreover, the methods can be broadly utilized in other areas of science that involve high dimensional data given that the methods show

improved performance on the major issues in high dimensional statistical learning: improving the accuracy of estimated model parameters and reducing the expected loss of the estimated model [Fan and Li, 2006, Fan and Lv, 2010].

For potential future studies, COLRNS-GLM can be extended to deal with various types of multicategorical responses in a generalized linear regression setting. As longitudinal data are more commonly collected in diverse areas, we can also consider the extensions of COLRNS and COLRNS-GLM that are applicable to longitudinal data.

APPENDIX

ROC CURVES FOR VARIABLE SELECTION

We evaluated the performance of COLRNS regarding variable selection by examining the sensitivity, specificity, false discovery rate, and false negative rate in simulations. In this Appendix, ROC curves are shown for the same scenarios of the simulations in Chapter 3 to further explore how COLRNS behaves for variable selection in comparison to Elastic net.

As explained before, the key part of COLRNS is that it yields the best model with the smallest cross-validation error among the models given after iteratively performing the procedures: identification of the lead cluster followed by selection of predictors. In each selection using the lead cluster, the model is resulted by the two parameters α and λ that are determined to give the least cross-validation error.

To fully reflect this gist of COLRNS, we adopt a stochastic approach to generate the parameter values that are used in each selection step. We use the uniform distribution to generate the parameters $\alpha \sim U(0.1, 0.9)$ and $\lambda \sim U(\lambda_{min}, \lambda_{max})$ where the minimum and maximum parameters of the uniform distribution vary according to each scenario. For the parameters of the uniform distribution to generate λ, we refer to the value λ^* that results in the final model when COLRNS applies to each scenario. We set the values of the minimum and maximum parameters to cover λ^*, but not to have a too wide interval between them so that the parameter λ is randomly generated and not very far from the value of λ^* at the same time. The values of λ^*, λ_{min}, and λ_{max} are provided in Table XLVIII. We apply COLRNS $10,000$ times to the

APPENDIX (Continued)

data set for each scenario and plot the true positive rate versus false positive rate of variable selection from the $10,000$ resulting final models. The true positive rate and false positive rate are obtained by calculating sensitivity and $1-$ specificity referring to the definition of criteria in Section 3.5.2.

TABLE XLVIII: PARAMETER INFORMATION FOR GENERATING λ

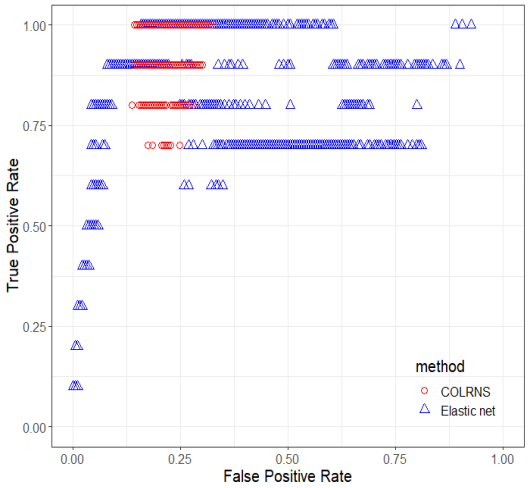
Scenario	λ^*	λ_{min}	λ_{max}
1	17.2	5	20
2	17.4	10	25
3	22.0	15	30
4	27.6	15	30
5	8.0	5	20
6	6.0	0.1	15
7	5.3	0.1	15

We investigate the performance of Elastic net over a large domain of parameters α and λ to use the results as reference. For the parameter values, two dimensional grid is created with 10 values from 0.1 to 0.9 for α and $1,000$ values from 0.1 to λ_{max} for λ according to each scenario with even intervals. In the same way, the pairs of true positive rate and false positive rate are plotted from the $10,000$ resulting models which are given by the $10,000$ combinations of the two parameters on the grid.

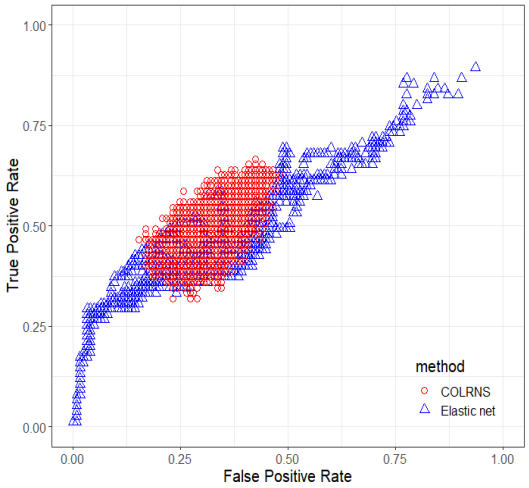
APPENDIX (Continued)

The ROC curves from COLRNS and Elastic net for the 7 scenarios are presented in Figure 49 and Figure 50. In Scenarios 1, 2, 3, and 4 with multiple groups of correlated predictors, the models from COLRNS tend to have high true positive rate in overall and lower false positive rate compared to the models given by Elastic net that have the same true positive rate. There are some models from Elastic net showing high true positive rate which are caused by small λ values, however, the models have large false positive rate at the same time. In Scenario 5 with one equi-correlated predictors of sparse and strong signals, COLRNS shows similiar pattern as in Scenario 1 with high true positive rate and relatively low false positive rate. In Scenario 6 and 7 having equi-correlated predictors with many weak signals, COLRNS and Elastic net almost overlap in the plots, however, COLRNS still presents relatively large true positive rate. The points of COLRNS in the plots are located close each other occupying narrow areas. In overall, the pattern of ROC curves corresponds to the results of simulations that COLRNS generally shows good performance for detecting important predictors with true signals improving robustness in variable selection.

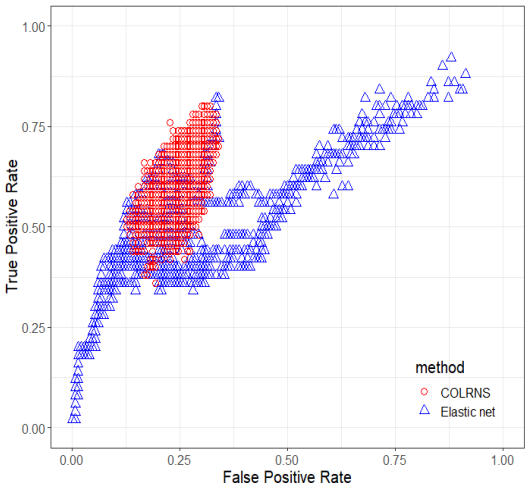
APPENDIX (Continued)



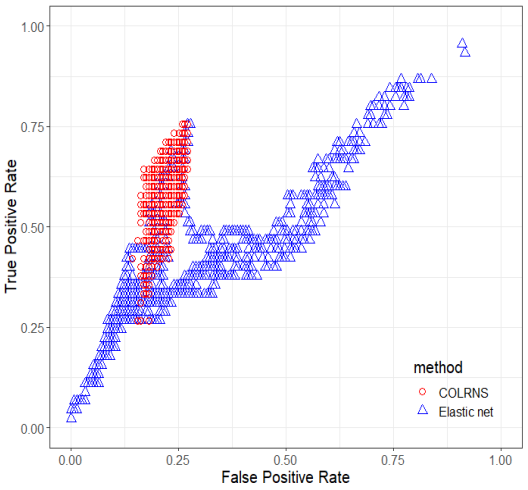
(a) Scenario 1



(b) Scenario 2



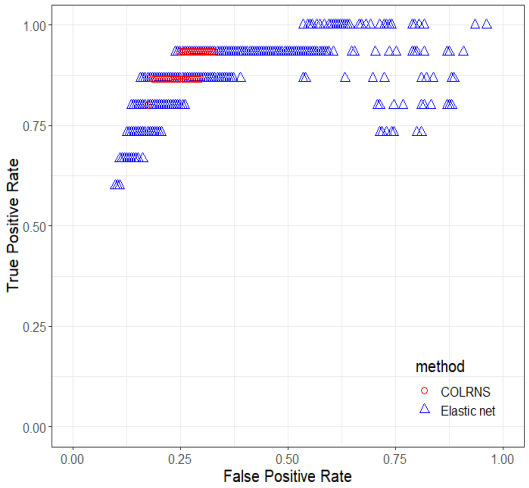
(c) Scenario 3



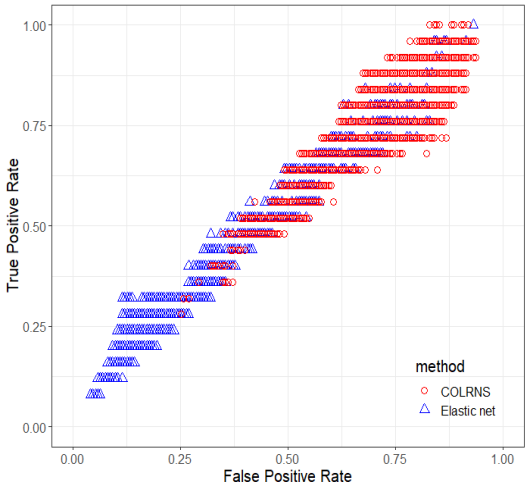
(d) Scenario 4

Figure 49: ROC curves for scenarios 1, 2, 3, and 4

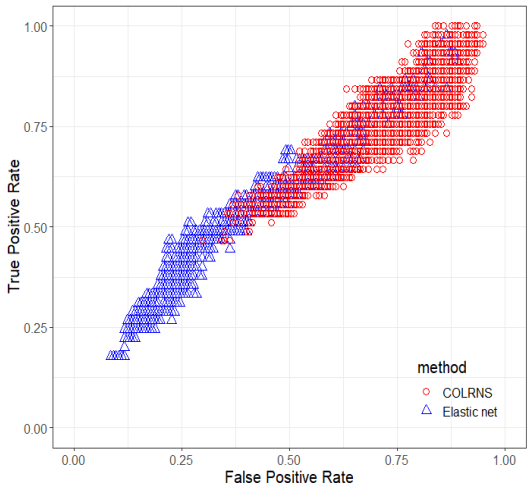
APPENDIX (Continued)



(a) Scenario 5



(b) Scenario 6



(c) Scenario 7

Figure 50: ROC curves for scenarios 5, 6, and 7

CITED LITERATURE

Akaike, H. (1973). Information theory and an extension of maximum likelihood principle. The 2nd International Symposium on Information Theory, pages 267–281.

Atkinson, K. (1989). An Introduction to Numerical Analysis, 2nd Edition. John Wiley Sons, Inc.

Beale, E. M. L., Kendall, M. G., and Mann, D. W. (1967). The discarding of variables in multivariate analysis. Biometrika, 54(3/4):357–366.

Billionnet, C., Sherrill, D., Annesi-Maesano, I., and study, G. (2012). Estimating the health effects of exposure to multi-pollutant mixture. Annals of Epidemiology, 22(2):126–141.

Bobb, J. F., Henn, B. C., Valeri, L., and Coull, B. A. (2018). Statistical software for analyzing the health effects of multiple concurrent exposures via bayesian kernel machine regression. Environmental Health, 17(67).

Bobb, J. F., Valeri, L., Henn, B. C., Christiani, D. C., Wright, R. O., Mazumdar, M., Godleski, J. J., and Coull, B. A. (2015). Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. Biostatistics, 16(3).

Bondell, H. D. and Reich, B. J. (2008). Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with oscar. Biometrics, 64(1):115–123.

Breaux, H. J. (1967). On Stepwise Multiple Linear Regression. Technical Report. Aberdeen: Army Ballistic Research Lab Aberdeen Proving Ground MD.

Breheny, P. (2020). Regularization paths for SCAD and MCP penalized regression models.

Breiman, L. (1995). Better subset regression using the nonnegative garrote. Technometrics, 37:373–384.

Cawthon, R. M. (2002). Telomere measurement by quantitative pcr. Nucleic Acids Res, 30(10):47.

Chatterjee, S. (2012). Regression Analysis by Example, 5th Edition. Wiley.

Chong, I.-G. and Jun, C.-H. (2005). Performance of some variable selection methods when multicollinearity is present. Chemometrics and Intelligent Laboratory Systems, 78:103–112.

Choudhury, H., Hertzberg, R., Rice, G., Doyle, E., Woo, Y., and Schoeny, R. (2000). Supplementary guidance for conducting health risk assessment of chemical mixtures. U.S. Environmental Protection Agency.

Curto, J. D. and Pinto, J. C. (2007). New multicollinearity indicators in linear regression models. International Statistical Review, 75(1):114–121.

Curto, J. D. and Pinto, J. C. (2014). Cross-validation pitfalls when selecting and assessing regression and classification models. Journal of Cheminformatics, 6(10).

Desboulets, L. D. D. (2018). A review on variable selection in regression analysis. Econometrics, 6(4).

El-Dereny, M. and Rashwan, N. I. (2011). Solving multicollinearity problem using ridge regression models. International Journal of Contemporary Mathematical Sciences, 6(12):585–600.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. American Statistical Association, 96(456):1348–1360.

Fan, J. and Li, R. (2006). Statistical challenges with high dimensionality: Feature selection in knowledge discovery. arXiv:math/0602133.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. Journal of the Royal Statistical Society: Series B (Methodology), 70(5):849–911.

Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. Statistica Sinica, 20(1):101–148.

Fan, J., Samworth, R., and Wu, Y. (2009). Ultrahigh dimensional variable selection: beyond the linear model. Journal of Machine Learning Research, 10(70):2013–2038.

Fan, J. and Song, R. (2010). Sure independence screening in generalized linear models with np-dimensionality. The Annals of Statistics, 38(6):3567–3640.

Flom, P. L. and Cassell, D. L. (2007). Stopping stepwise: Why stepwise and similar selection methods are bad, and what you should use. NorthEast SAS Users Group Inc 20th Annual Conference.

Frank, I. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. Technometrics, 35:109–135.

Friedman, J., Hastie, T., Tibshirani, R., Narasimhan, B., Simon, N., and Qian, J. (2019). Lasso and elastic-net regularized generalized linear models.

Fu, W. (1998). The bridge versus the lasso. Journal of Computational and Graphical Statistics, 7(3):397–416.

García, C. B., García, J., Martín, M. M. L., and Salmerón, R. (2015). Collinearity: revisiting the variance inflation factor in ridge regression. Journal of Applied Statistics, 42(3):648–661.

Gennings, C., Sabo, R., and Carney, E. (2010). Identifying subsets of complex mixtures most associated with complex diseases: polychlorinated biphenyls and endometriosis as a case study. Epidemiology, 21(Suppl 4):77–84.

Genovese, C. and Wasserman, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. Journal of the Royal Statistical Society: Series B (Methodology), 64:499–517.

George, E. and Mcculloch, R. (1993). Variable selection via gibbs sampling. Journal of the American Statistical Association, 88(423):881–889.

Gibson, E. A., Yanelli Nunez and, A. A., Zota, A. R., Renzetti, S., Devick, K. L., Gennings, C., Goldsmith, J., Coull, B. A., and Kioumourtzoglou, M.-A. (2019). An overview of methods to address distinct research questions on environmental mixtures: an application to persistent organic pollutants and leukocyte telomere length. Environmental Health, 18(76).

Greene, W. H. (1993). Econometric Analysis, 2nd Edition. Macmillan, New York.

Hanrahan, L. P., Falk, C., Anderson, H. A., Draheim, L., Kanarek, M. S., Olson, J., and The Great Lakes Consortium (1999). Serum pcb and dde levels of frequent great lakes sport fish consumers—a first look. Environmental Research Section A, 80:26–37.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition. Springer, New York.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). Statistical Learning with Sparsityy: the LASSO and Generalizations. Boca Raton, FL, USA: CRC Press.

Hernandez-Orallo, J., Flach, P., and Ferri, C. (2012). A unified view of performance metrics: Translating threshold choice into expected classification loss. Journal of Machine Learning Research, 13:2813–2869.

Hocking, R. R. and Leslie, R. N. (1967). Selection of the best subset in regression analysis. Technometrics, 9(4):531–540.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1):55–67.

Hoerl, A. E. and Kennard, R. W. (1988). Ridge regression. Encyclopedia of Statistical Sciences, 8:129–136.

Hu, H., Shine, J., and Wright, R. O. (2007). The challenge posed to children's health by mixtures of toxic waste: the tar creek superfund site as a case-study. Pediatric Clinics of North America, 54(1):155–175.

Hurvich, C. M. and Tsai, C.-L. (1989). Regression and time series model selection in small samples. Biometrika, 76:297–307.

Jang, W., Lim, J., Lazar, N. A., Loh, J. M., and Yu, D. (2013). Regression shrinkage and grouping of highly correlated predictors with horses. arXiv:1302.0256.

JingYuan, L., Wei, Z., and RunZe, L. (2015). A selective overview of feature screening for ultra high dimensional data. Science China Mathematics, 58(10):2033–2054.

Kioumourtzoglou, M.-A., Zanobetti, A., Schwartz, J. D., Coull, B. A., Dominici, F., and Suh, H. H. (2013). The effect of primary organic particles on emergency hospital admissions among the elderly in 3 us cities. Environmental Health, 12(19):20.

Lebanon, G. (2010). Bias, variance, and mse of estimators. `http://theanalysisofdata.com/notes/estimators1.pdf`.

Li, X. R. and Zhao, Z. (2001). Measures of performance for evaluation of estimators and filters. International Society for Optics and Photonics, 4473:530 – 541.

Lin, J., Epel, E., Cheon, J., Kroenke, C., Sinclair, E., Bigos, M., Wolkowitz, O., Mellon, S., and Blackburn, E. (2010). Analyses and comparisons of telomerase activity and telomere length in human t and b cells: insights for epidemiology of telomere maintenance. Journal of Immunological Methods, 352(1-2):71–80.

Liu, K. (2003). Using liu type estimator to combat collinearity. Communications in Statistics - Theory and Methods, 32(5):1009–1020.

Liu, Y., Zhang, H. H., Park, C., and Ahn, J. (2007). Support vector machines with adaptive l_q penalty. Computational Statistics Data Analysis, 51(12):6380–6394.

Mallows, C. L. (1973). Some comments on cp. Technometrics, 15:661–675.

Mardikyan, S. and Çetin, E. (2008). Efficient choice of biasing constant. International Journal of Contemporary Mathematical Sciences, 3(11):527–536.

Mitro, S. D., Birnbaum, L. S., Needham, B. L., , and Zota, A. R. (2016). Cross-sectional associations between exposure to persistent organic pollutants and leukocyte telomere length among u.s. adults in NHANES, 2001–2002. Environmental Health Perspectives, 124(5):651–658.

Montgomery, D. C., Peck, E. A., and Vining, G. (2012). Introduction to Linear Regression Analysis, 5th Edition. John Wiley and Sons, Inc.

Needham, B. L., Adler, N., Gregorich, S., Rehkopf, D., Lin, J., Blackburn, E., and Epel, E. S. (1989). The estimation of total serum lipids by a completely enzymatic 'summation' method. Clinica Chimica Acta, 184(3):219–226.

Needham, B. L., Adler, N., Gregorich, S., Rehkopf, D., Lin, J., Blackburn, E., and Epel, E. S. (2013). Socioeconomic status, health behavior, and leukocyte telomere length in the national health and nutrition examination survey, 1999–2002. Social Science Medicine, 85:1–8.

Pagel, M. D. and Lunneborg, C. E. (1985). Empirical evaluation of ridge regression. Psychological Bulletin, 97(2):342–355.

Pan, W., Wang, X., Xiao, W., and Zhu, H. (2019). A generic sure independence screening procedure. Journal of the American Statistical Association, 114(526):928–937.

Park, M. Y. and Hastie, T. (2007). l_1-regularization path algorithm for generalized linear models. Journal of the Royal Statistical Society: Series B (Methodology), 69:659–677.

Park, M. Y., Hastie, T., and Tibshirani, R. (2007). Averaged gene expressions for regression. Biostatistics, 8(2):212–227.

Qiao, X. (2014). Variable selection using l_q penalties. Wiley Interdisciplinary Reviews: Computational Statistics, 6(3):177–184.

Ročková, V. and George, E. I. (2018). The spike-and-slab lasso. Journal of the American Statistical Association, 113(521):431–444.

Schwarz, G. (1978). Estimating the dimension of a model. The Annals of Statistics, 6:461–464.

Serdar, B., LeBlanc, W. G., Norris, J. M., and Dickinson, L. M. (2014). Potential effects of polychlorinated biphenyls (PCBs) and selected organochlorine pesticides (OCPs) on immune cells and blood biochemistry measures: a cross-sectional assessment of the NHANES 2003–2004 data. Environmental Health, 13:114.

Sharma, D. B., Bondell, H. D., and Zhang, H. H. (2010). Consistent group identification and variable selection in regression with correlated predictors. Journal of Computational and Graphical Statistics, 20(1):101–148.

Shen, J. and Gao, S. (2008). A solution to separation and multicollinearity in multiple logistic regression. Journal of Data Science, 6(4):515–531.

Steyerberg, E. W., Eijkemans, M. J. C., and Habbema, J. D. F. (1999). Stepwise selection in small data sets: A simulation study of bias in logistic regression analysis. Journal of Clinical Epidemiology, 52:935–942.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodology), 58(1):267–288.

Tibshirani, R. (2007). The lasso method for variable selection in the cox model. Statistics in Medicine, 16:385–395.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. Journal of the Royal Statistical Society: Series B (Methodology), 67:91–108.

Tikhonov, A. (1943). `https://en.wikipedia.org/wiki/Tikhonov_regularization`.

Turyk, M., Anderson, H. A., Knobeloch, L., Imm, P., and Persky, V. W. (2009). Prevalence of diabetes and body burdens of polychlorinated biphenyls, polybrominated diphenyl ethers, and p,p'-diphenyldichloroethene in Great Lakes sport fish consumers. Chemosphere, 75(5):674–679.

Vinod, H. D. and Ullah, A. (1981). Recent Advances in Regression Models. Marcel Dekker Incorporated, New York.

Walker, E. (1989). Detection of collinearity-influential observations. Communications in Statistics - Theory and Methods, 18(5):1675–1690.

Wang, X. and Leng, C. (2016). High dimensional ordinary least squares projection for screening variables. Journal of the Royal Statistical Society: Series B (Methodology), 78(3):589–611.

Whittingham, M. J. and Stephens, P. A. (2006). Why do we still use stepwise modelling in ecology and behaviour? Journal of Animal Ecology, 75:1182–1189.

Wong, H. B. and Lim, G. H. (2011). Measures of diagnostic accuracy: Sensitivity, specificity, ppv and npv. Proceedings of Singapore Healthcare, 20(4).

Xie, J. and Zeng, L. (2010). Group Variable Selection Methods and Their Applications in Analysis of Genomic Data, volume 15. Springer, London.

Xie, Y., Wang, X., and Jr., J. A. S. (2015). Deciduous forest responses to temperature, precipitation, and drought imply complex climate change impacts. Proceedings of the National Academy of Sciences of the United States of America, 112(44):13585–13590.

Yu, G. and Liu, Y. (2016). Sparse regression incorporating graphical structure among predictors. Journal of the American Statistical Association, 111(514):707–720.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Methodology), 68(1):49–67.

Zhang, C.-H. (2007). Penalized linear unbiased selection. Manuscript.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. The Annals of Statistics, 38(2):894–942.

Zhong, W. and Zhu, L. (2015). An iterative approach to distance correlationbased sure independence screening. Journal of Statistical Computation and Simulation, 85(1):2331–2345.

Zou, H. (2006). The adaptive lasso and its oracle properties. Journal of the American Statistical Association, 101:1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society. Series B (Statistical Methodology), 67(2):301–320.

Zou, H. and Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. The Annals of Statistics, 37(4):1733–1751.

VITA

JIYEONG JANG

Education

- **Doctor of Philosophy**, 2020
  Biostatistics
  University of Illinois at Chicago, Chicago, Illinois, USA
- **Master of Science**, 2014
  Statistics
  Seoul National University, Seoul, South Korea
- **Bachelor of Economics**, 2012
  Statistics
  Korea University, Seoul, South Korea
- **Bachelor of Health Science**, 2012
  Environmental Health
  Korea University, Seoul, South Korea

Research Experiences

- **The National Institute of Environmental Health Sciences (NIEHS) grant: Innovative Methodologic Advances for Mixtures Research in Epidemiology**
  Graduate Research Assistant, Chicago, Illinois, June 2019–October 2020
- **Global Medical Affairs Statistics, AbbVie**
  Data and Statistical Sciences Extern, Chicago, Illinois, August 2018 - May 2019
- **Janssen Pharmaceutical Companies of Johnson & Johnson**
  Quantitative Sciences Summer Intern, Titusville, New Jersey, May 2018 - August 2018
- **The US Department of Health and Human Services (HHS) grant: Coordination of Healthcare for Complex Kids (CHECK)**
  Graduate Research Assistant, Chicago, Illinois, August 2017 - May 2018
- **Campus Research Board Pilot Project: Phthalate Exposures and Metabolic Dysfunction in Chicago Adults**
  Graduate Research Assistant, Chicago, Illinois, May 2017 - August 2017
- **The National Institutes of Health (NIH) grant: Asthma Action at Erie Trial**
  Graduate Research Assistant, Chicago, Illinois, August 2016 - August 2017

- **Methodology Research Core, Institute for Health Research and Policy**
  Graduate Research Assistant, Chicago, Illinois, August 2015 - July 2016
- **The National Research Foundation of Korea (NRF) grant:**
  Association between predicted particulate matter air pollution and health
  Graduate Research Assistant, Seoul, South Korea, September 2014 - July 2015

Teaching Experiences

- **Division of Epidemiology and Biostatistics, University of Illinois at Chicago: Biostatistics I and Biostatistics II**
  Teaching Assistant, 2015 – 2017
- **Korea National Open University: Elementary Mathematics (Calculus)**
  Tutor, 2014
- **Department of Statistics, Seoul National University: Statistics Lab**
  Instructor, 2013

Honors

- Graduate Assistantship, University of Illinois at Chicago, 2015 - 2020
- Conference Travel Award, University of Illinois at Chicago, 2019
- Teaching and Research Assistant Scholarship, Seoul National University, 2013
- Superior Academic Performance Scholarship, Seoul National University, 2012
- Top Honor Student, Korea University, 2007 and 2008
- Honor Student, Korea University, 2008, 2009, and 2010
- Academic Excellence Scholarship, Korea University, 2007, 2008, 2009, and 2011

Presentations

- **J. Jang**, S. Basu, H. Y. Chen, and M. Turyk (2020). "Selecting Key Agents from Environmental Mixtures: A Method for High-Dimensional Feature Selection in Presence of Strong Collinearity". Powering Research Through Innovative Methods for Mixtures in Epidemiology (PRIME) Program NIEHS Grantee Meeting, Virtual
- S. Basu, **J. Jang**, H. Y. Chen, and M. Turyk (2020). "Selecting Key Agents from Environmental Mixtures: A Method for High-Dimensional Feature Selection in Presence of Strong Collinearity". Annual Conference of the International Society for Environmental Epidemiology, Virtual

- S. Basu and **J. Jang** (2020). "Modeling Risks from Environmental Mixtures". Joint Statistical Meetings, Virtual
- **J. Jang** and S. Basu (2019). "A method for high-dimensional variable selection in presence of collinearity". Joint Statistical Meetings, Denver
- Y. Sun, **J. Jang**, X. Huang, H. Wang, and W. He (2019) "Leveraging free text data for decision making in drug development". Joint Statistical Meetings, Denver
- **J. Jang**, S. Basu, H. Y. Chen, M. Daviglus, and M. Turyk (2019). "SEEM: Selecting key environmental exposures in mixtures". Powering Research Through Innovative Methods for Mixtures in Epidemiology (PRIME) Program NIEHS Grantee Meeting, Raleigh
- **J. Jang**, S. Basu, J. A. Borgia, and M. J. Fiddler (2019). "Selecting key cancer biomarkers: a method for high-dimensional feature selection in presence of collinearity". Dr. Gary Kruh Cancer Research Symposium, Chicago
- M. A. Martin, R. Caskey, A. E. Glassgow, A. Pappalardo, L. L. Hsu, **J. Jang**, S. Basu, M. Minier, K. Fox, G. Munoz, B. V. Voorhee (2019). "Trends in school attendance for low-income children with chronic health conditions". Pediatric Academic Society Meeting, Baltimore
- J. Buscemi, O. Pugach, **J. Jang**, S. Springfield, L. Schiffer, M. R. Stolley, and M. L. Fitzgibbon (2016). "Relations between fiber intake and body mass index among African-American women participating in a randomized controlled weight loss trial". Society of Behavioral Medicine Annual Meeting, Washington, D.C.
- **J. Jang** and S. Y. Kim (2015). "Long-term exposure to PM10 and risk of low birth weight in Seoul, Korea". Korean Society of Environmental Health and Toxicology Spring Conference, Jeju, South Korea
- **J. Jang**, S. J. Yi, Y. S. Eum, and S. Y. Kim (2014). "PM10 concentrations and infant mortality across 25 districts in Seoul, South Korea". Korean Society of Environmental Health and Toxicology Fall Conference, Jeju, South Korea

Publications

- S. Choe, **J. Jang**, M. Kim, Y. Jun, and S.Y. Kim (2019). "Association between ambient particulate matter concentration and fetal growth restriction stratified by maternal employment". BMC Pregnancy and Childbirth, 19(1), 246.
- J. Buscemi, O. Pugach, S. Springfield, **J. Jang**, L. Tussing-Humphreys, L. Schiffer, M. R. Stolley, and M. L. Fitzgibbon (2018). "Associations between fiber intake and Body Mass Index (BMI) among African-American women participating in a randomized weight loss and maintenance trial". Eating Behaviors, 29, 48-53.

- M. Kim, M. C. Paik, **J. Jang**, Y. K. Cheung, J. Willey, M. S. V. Elkind, and R. L. Sacco (2017). "Cox proportional hazards models with left truncation and time-varying coefficient: Application of age at event as outcome in cohort studies". Biometrical Journal, 59(3), 405-419.

Professional Membership

- American Statistical Association
- Korean International Statistics Society