CausalSelect: An R Package for Variable Selection Methods for High dimensional Causal Inference

true true

2025-06-19

Abstract

Estimating causal effects in high-dimensional settings poses major challenges, especially when the number of covariates far exceeds the number of observations. The `CausalSelect` R package implements a suite of recent variable selection methods tailored for causal inference, including the Outcome-Adaptive Lasso (OAL), its generalized version (GOAL), and Sure Independence Screening (SIS) extensions. Additionally, the package introduces the CBS method, a screening approach coupled with doubly robust ATE estimation. This article presents the methodological background of these approaches, outlines the structure of the `CausalSelect` package, and demonstrates its use through simulated examples.

Introduction

Causal inference aims to estimate the effect of a treatment, exposure, or intervention on an outcome by comparing what actually happened with what would have happened under an alternative scenario. This kind of reasoning, known as counterfactual reasoning, is formalized in the widely used potential outcomes framework.

One major challenge in observational studies is the presence of confounding variables (covariates that influence both the treatment assignment and the outcome). Failing to adjust for these confounders can bias the estimation of causal effects. However, selecting the right variables to adjust for is a difficult task. Including irrelevant variables can reduce precision and increase variance, while omitting true confounders may lead to biased conclusions.

This challenge is further complicated in high-dimensional settings, where the number of covariates (p) is very large, often exceeding the number of observations (n). Such situations frequently occur in modern biomedical research, genomics, and medical imaging. Traditional methods are often ill-suited to handle these settings due to issues like multicollinearity and computational limits. To overcome these challenges, recent methodological advances have proposed variable selection techniques tailored for causal inference in high dimension. The Outcome Adaptive Lasso (OAL) adjusts penalization based on the strength of association with the outcome, improving confounder selection compared to standard Lasso. The Generalized Outcome Adaptive Lasso (GOAL) extends this approach by introducing a smoothed penalty function that enhances stability and convergence.

When p is much larger than n, screening techniques such as Sure Independence Screening (SIS) are used as a preliminary step to reduce the dimensionality before applying OAL or GOAL. Additionally, the Causal Ball Screening (CBS) method provides an alternative approach by selecting variables based on their conditional correlations and estimating the average treatment effect through a doubly robust estimator, which combines outcome and treatment models for greater efficiency and bias reduction.

The CausalSelect R package integrates these recent methods into a unified framework. It allows researchers to efficiently select relevant variables and estimate causal effects in complex, high dimensional data settings. Accurate variable selection not only reduces bias and variance but also facilitates interpretability, which is essential in critical fields such as epidemiology, oncology and personalized medicine.

Variable Selection in High-Dimensional Causal Inference

When working with high-dimensional observational data, identifying the right variables to adjust for is one of the most important steps in causal analysis. Including true confounders helps reduce bias, but adding too many irrelevant variables can inflate variance and hurt interpretability.

To address this issue, several recent methods have been developed specifically for variable selection in causal inference. The `CausalSelect` package brings together some of the most promising approaches, beginning with the Outcome-Adaptive Lasso.

Outcome-Adaptive Lasso (OAL)

We present a brief overview of the Outcome-Adaptive Lasso **(OAL)** approach proposed by Shortreed and Ertefaie (2017), which is designed to estimate the average treatment effect (ATE) in observational studies. The method models the propensity score (PS) using a logistic regression of the form:

$$\text{logit}\{\pi(X,\alpha)\} = \text{logit}\{P(A=1 \mid X)\} = \sum_{j=1}^{p} \alpha_j X_j$$

To perform variable selection, OAL applies an adaptive lasso penalty that encourages the inclusion of variables that are either confounders (X_C) or predictors of the outcome (X_P), while down-weighting predictors of the treatment only (X_I) or irrelevant variables (X_S). The fitted model becomes:

$$\text{logit}\{\pi(X,\widehat{\alpha})\} = \text{logit}\{\widehat{P}(A=1 \mid X)\} = \sum_{j\in C} \widehat{\alpha}_j X_j + \sum_{j\in P} \widehat{\alpha}_j X_j$$

The OAL estimator is obtained by solving the following optimization problem:

$$\widehat{\alpha}(\text{OAL}) = \arg\min_{\alpha} \left[\sum_{i=1}^{n} \left(-a_i(x_i^T\alpha) + \log(1+e^{x_i^T\alpha}) \right) + \lambda \sum_{j=1}^{p} \widehat{w}_j |\alpha_j| \right]$$

$$\text{where } \lambda > 0, \widehat{w}_j = |\hat{\beta}_j^{\text{ols}}|^{-\gamma} \text{ with } \gamma > 1, \text{and } (\hat{\beta}_A^{\text{ols}}, \hat{\beta}^{\text{ols}}) = \arg\min_{(\beta_A,\beta)} \|\mathbf{Y} - \beta_A A - \mathbf{X}\beta\|_2^2 .$$

Using the estimated propensity scores, inverse probability of treatment (IPT) weights are calculated as:

$$\hat{\tau}_i^{\lambda} = \frac{A_i}{\pi_i^{\lambda}(X_i, \widehat{\alpha}(\text{OAL}))} + \frac{1-A_i}{1-\pi_i^{\lambda}(X_i, \widehat{\alpha}(\text{OAL}))}$$

The optimal λ is selected from the set:

$$S_{\lambda} = \{n^{-10}, n^{-5}, n^{-2}, n^{-1}, n^{-0.75}, n^{-0.5}, n^{-0.25}, n^{0.25}, n^{0.49}\}$$

by minimizing the weighted absolute mean difference (wAMD):

$$\text{wAMD}(\lambda) = \sum_{j=1}^{p} |\hat{\beta}_j^{\text{ols}}| \cdot \left| \frac{\sum_{i=1}^{n} \hat{\tau}_i^{\lambda} X_{ij} A_i}{\sum_{i=1}^{n} \hat{\tau}_i^{\lambda} A_i} - \frac{\sum_{i=1}^{n} \hat{\tau}_i^{\lambda} X_{ij}(1-A_i)}{\sum_{i=1}^{n} \hat{\tau}_i^{\lambda}(1-A_i)} \right|$$

Finally, the ATE is estimated using the IPT weights:

$$\widehat{\text{ATE}} = \frac{\sum_{i=1}^{n} \hat{\tau}_i^{\widehat{\lambda}} Y_i A_i}{\sum_{i=1}^{n} \hat{\tau}_i^{\widehat{\lambda}} A_i} - \frac{\sum_{i=1}^{n} \hat{\tau}_i^{\widehat{\lambda}} Y_i (1-A_i)}{\sum_{i=1}^{n} \hat{\tau}_i^{\widehat{\lambda}} (1-A_i)}$$

- **Limitations of OAL**

Although the OAL method improves variable selection by focusing on outcome-related covariates, it has some important limitations. When covariates are highly correlated (which often happens in high-dimensional datasets) OAL can become unstable and may select irrelevant variables or miss important confounders. It also relies on solving a penalized logistic regression, which becomes computationally challenging when the number of covariates greatly exceeds the sample size. Finally, because OAL depends on weights derived from an initial outcome model, its performance is sensitive to the quality of that model. If the outcome regression is poorly specified or noisy, the adaptive penalty may misguide the selection process.

Generalized Outcome-Adaptive Lasso (GOAL):

While the Outcome-Adaptive Lasso (OAL) introduced by Shortreed and Ertefaie (2017) was an important step forward for variable selection in causal inference, it has known limitations, particularly in the presence of correlated covariates. In practice, OAL tends to perform poorly when variables are highly collinear, which is a common feature in high-dimensional datasets. To address this, M. Ismaila Baldé proposed the Generalized Outcome-Adaptive Lasso (GOAL). GOAL improves upon OAL by incorporating an elastic net-type penalty into the OAL framework. This hybrid penalty helps stabilize the estimation and improves variable selection, especially in settings with strongly correlated covariates.

Assuming a propensity score model parameterized by α:

$$\text{logit}\{P(A=1 \mid X)\} = \sum_{j=1}^{p} \alpha_j X_j$$

Let $\ell_n(\alpha; A, \mathbf{X}) = \sum_{i=1}^{n} \left\{ -a_i(x_i^\top \alpha) + \log\left(1 + e^{x_i^\top \alpha}\right) \right\}$ adenote the negative log-likelihood. OAL solves the following penalized logistic regression:

$$\hat{\alpha}(\text{OAL}) = \arg\min_{\alpha} \left[\sum_{i=1}^{n} \ell_n(\alpha; A, \mathbf{X}) + \lambda_n \sum_{j=1}^{p} \widehat{w}_j |\alpha_j| \right]$$

$$\text{where } \widehat{w}_j = |\hat{\beta}_j^{\text{ols}}|^{-\gamma} \text{ with } \gamma > 1, \text{and } (\hat{\beta}_A^{\text{ols}}, \hat{\beta}^{\text{ols}}) = \arg\min_{(\beta_A, \beta)} \|\mathbf{Y} - \beta_A A - \mathbf{X}\beta\|_2^2 .$$

To tune the penalty parameter λ_n, OAL minimizes the weighted absolute mean difference (wAMD) between treated and untreated groups:

$$S_{\lambda_n} = \{n^{-10}, n^{-5}, n^{-2}, n^{-1}, n^{-0.75}, n^{-0.5}, n^{-0.25}, n^{0.25}, n^{0.49}\}$$

$$\widehat{\lambda_n} = \arg\min_{\lambda_n \in S_{\lambda_n}} \text{wAMD}(\lambda_n, \mathbf{X}, A)$$

where

$$\text{wAMD}(\lambda_n, \mathbf{X}, A) = \sum_{i=1}^{p} |\hat{\beta}_j^{\text{ols}}| \left| \frac{\sum_{i=1}^{n} \hat{\tau}_i^{\lambda_n} X_{ij} A_i}{\sum_{i=1}^{n} \hat{\tau}_i^{\lambda_n} A_i} - \frac{\sum_{i=1}^{n} \hat{\tau}_i^{\lambda} X_{ij} (1 - A_i)}{\sum_{i=1}^{n} \hat{\tau}_i^{\lambda} (1 - A_i)} \right|$$

$\hat{\tau}_i^{\lambda_n}$ is the inverse probability of treatment weight for individual i constructed using the PS model fitted from OAL.

Building upon this, GOAL estimator is defined through the following optimization problem :

$$\hat{\alpha}(\text{GOAL}) = \arg\min_{\alpha} \left[\ell_n(\alpha; A, \mathbf{X}) + \lambda_1 \sum_{j=1}^{p} \widehat{w}_j |\alpha_j| + \lambda_2 \sum_{j=1}^{p} \alpha_j^2 \right]$$

where $\widehat{w}_j$ is the same as in the OAL formulation.

In practice, when covariates are independent, GOAL and OAL yield similar results. However, in both low and high dimensional settings with correlated covariates, GOAL provides more robust and accurate variable selection than OAL.

Causal Ball Screening (CBS):

Causal Ball Screening (CBS) is a two-step procedure developed for variable selection and average treatment effect (ATE) estimation in ultra-high dimensional settings, where traditional outcome-model-based methods like OAL and GOAL become unstable or inefficient. CBS addresses these challenges by first applying a model-free screening step based on the conditional ball covariance, an extension of ball covariance (Pan et al., 2018b, 2020) that measures associations between covariates and the outcome conditional on treatment. This is followed by a refined selection and estimation step, ensuring both validity and double robustness in causal inference.

Let $w = P(D = 1)$ be the probability of receiving treatment. Let $X^{(d)},\ Y^{(d)},\ d = 0, 1$ be random variables such that $(X^{(d)}, Y^{(d)}) = (X, Y \mid D = d), d = 0, 1$ The conditional ball covariance between a covariate X and the outcome Y given D is defined as the square root of :

$$BCov^2(X, Y \mid D) = \ w\ BCov^2(X^{(1)}, Y^{(1)}) + (1 - \ w)\ Bcov^2(X^{(0)}, Y^{(0)})$$

Let $n_1 = \sum_{i=1}^{n} D_i$ be the number of subject who receive treatment ando $n_0 = n - n_1$. And $\widehat{w} = \frac{n_1}{n}$ be the empirical estimator of w

The empirical conditional ball covariance $BCov_n(X, Y|D)$ is defined as the square root of:

$$BCov_n^2(X, Y \mid D) = \frac{\widehat{w}}{n_1^6} \sum_{(i,j,k,l,s,t):D_i,D_j,D_k,D_l,D_s,D_t=1} \xi^X_{ij,klst}\, \xi^Y_{ij,klst} + \frac{1-\widehat{w}}{n_0^6} \sum_{(i,j,k,l,s,t):D_i,D_j,D_k,D_l,D_s,D_t=0} \xi^X_{ij,klst}\, \xi^Y_{ij,klst}$$

This measure satisfies a key property:

$$BCov(X, Y \mid D) = 0 \quad \Leftrightarrow \quad X \perp Y \mid D$$

To perform screening, we compute $Bcov_n(X^{(j)}, Y \mid D)$ for each covariate $j = 1, ..., p$, and retain the q variables with the largest values to form the set $K = \{1, 2, ..., q\}$

In the second step, CBS applies a refined selection procedure. For the outcome model, the Lasso is used to estimate parameters on the selected variables in K

$$\hat{\alpha}_K^{(d)} = \arg\min_{\alpha_K} \left\{ \sum_{i:D_i=d} (Y_i - X_{i,K}^\top \alpha_K)^2 + \lambda_Y^{(d)} \|\alpha_K\|_1 \right\}, d = 0, 1$$

where $\lambda_Y^{(d)}$ is chosen via 10-fold cross validation . For the propensity score model, CBS uses a weighted adaptive Lasso that avoids outcome model dependence. Let $e(X;\ \beta) = expit(X^T\ \beta) = \frac{exp(X^T\ \beta)}{1+exp(X^T\ \beta)}$ The estimator is defined as:

$$\hat{\beta}_K = \arg\min_{\beta_K} \left(\sum_{i=1}^{n} \left[D_i \log \left\{ \frac{1 - e(X_{i,K}; \beta_K)}{e(X_{i,K}; \beta_K)} \right\} - \log\{1 - e(X_{i,K}; \beta_K)\} \right] + \lambda_D \sum_{i=1}^{q} \frac{1}{\widehat{w}_j} |\beta_j| \right)$$

The weights $w_j$are derived from the conditional ball covariance as follows:

$$\widehat{w}_j = \left| \hat{z}\, Bcov_n^2(X^{(j)}, Y \mid D) \right|^\gamma \quad \text{with} \quad \hat{z} = \frac{1}{max_j \left| Bcov_n^2(X^{(j)}, Y \mid D) \right|}$$

The pair of parameters (γ, λ_D) that minimize the weighted absolute mean difference (Shortreed and Ertefaie, 2017) is selected

$$wAMD(\lambda_D, \gamma) = \sum_{i=1}^{q} |\beta_j| \times \left| \frac{\sum_{i=1}^{n} \widehat{\tau}_i^{\lambda_D,\gamma} X_i^{(j)} D_i}{\sum_{i=1}^{n} \widehat{\tau}_i^{\lambda_D,\gamma} D_i} - \frac{\sum_{i=1}^{n} \widehat{\tau}_i^{\lambda_D,\gamma} X_i^{(j)} (1 - D_i)}{\sum_{i=1}^{n} \widehat{\tau}_i^{\lambda_D,\gamma} (1 - D_i)} \right|$$

This two-step structure allows CBS to efficiently select relevant confounders while preserving the double robustness property of the average treatment effect estimator, that is, consistency is guaranteed as long as either the outcome model or the treatment assignment model is correctly specified.

Sure Independance Screening Combined with OAL and GOAL:

Besides CBS, which addresses the limitations of OAL and GOAL in ultra-high dimensional settings, another solution consists in combining a screening step with adaptive penalization. As an alternative solution, we present an extension of OAL and GOAL that also targets ultra-high dimensional settings. Specifically, we combine Sure Independence Screening (SIS), as introduced by Tang et al. (2022), with either OAL or GOAL. These combined procedures, referred to as SIS+OAL and SIS+GOAL, aim to reduce the dimensionality before applying adaptive penalized estimation.The GOAL variant used in this extension is GOALi, as recommended by Baldé et al. (2023) for high-dimensional data. We assume the following propensity score model:

$$\text{logit}\{\pi(X,\alpha)\} = \text{logit}\{P(A=1 \mid X)\} = \sum_{j=1}^{p} \alpha_j X_j$$

Let:

- C: the set of confounders (associated with both treatment and outcome)
- P: the set of precision variables (associated only with the outcome)
- I: the set of instrumental variables (associated only with the treatment)
- S: the set of irrelevant variables (associated with neither outcome nor treatment)

Define $A = C \cup P$ and $A^c = I \cup S$. The goal of the proposed procedures is to estimate the following reduced propensity score model:

$$\text{logit}\{\pi(X,\alpha)\} = \sum_{j \in A} \alpha_j X_j$$

Each procedure starts with a SIS step to reduce the covariate dimension from p to $q < n$ yielding the selected index set $K = \{1, 2, \ldots, q\}$. Then, either OAL or GOAL is applied to the reduced dataset (X_K, A). The negative log-likelihood is defined as:

$$\ell_n(\alpha_K; A, \mathbf{X}_K) = \sum_{i=1}^{n} \left\{ -a_i(\mathbf{x}_{i,K}^\top \alpha_K) + \log\left(1 + e^{\mathbf{x}_{i,K}^\top \alpha_K}\right) \right\}$$

The SIS+OAL estimator is defined as:

$$\hat{\alpha}_K(\text{SIS+OAL}) = \arg\min_{\alpha_K} \left[\ell_n(\alpha; A, \mathbf{X}_K) + \lambda_1 \sum_{j=1}^{q} \widehat{w}_j |\alpha_{j,K}| \right]$$

and the SIS+GOAL estimates are :

$$\hat{\alpha}_K(\text{SIS+GOAL}) = \arg\min_{\alpha_K} \left[\ell_n(\alpha; A, \mathbf{X}_K) + \lambda_1 \sum_{j=1}^{q} \widehat{w}_j |\alpha_{j,K}| + \lambda_2 \sum_{j=1}^{q} \alpha_{j,K}^2 \right]$$

The adaptive weights are defined as $\widehat{w}_j = |\hat{\beta}_j|^{-\gamma}$ with $\gamma > 1$, for $j = 1, \ldots, q$. The estimates $(\hat{\beta}_A, \hat{\beta})$ are obtained by solving the following optimization problem:

$$(\hat{\beta}_A, \hat{\beta}) = \arg\min_{(\beta_A, \beta)} \ \mathcal{L}_n^Y(\beta_A, \beta; Y, A, \mathbf{X}_K)$$

Where $\mathcal{L}_n^Y$ denotes the negative log-likelihood of the outcome Y, conditional on the treatment and the design matrix $\mathbf{X}_K$ for a simple of size n. In this formulation $\hat{\beta}_A$ represents the estimated coefficient for the treatment variable, and $\hat{\beta}$ contains the estimated coefficients for the q selected covariates. # Illustrative Examples: After introducing the methods, we now turn to practical examples using both simulated and radiomic datasets to showcase how the CausalSelect package performs in different settings.

Simulated Data Example:

We simulate high-dimensional data using the Data_G() function, developed jointly by M. Baldé and Rime Naaman , to test and illustrate the methods implemented in the package.

OAL and GOAL methods:

```
rm(list=ls())
set.seed(2015)
## Generate a multivariate normal X matrix
# set information for simulating coviariates
mean_x = 0
sig_x = 1

# pairwise correlation between covariates
rho = 0.5

# set number of monte carlo (MC) simulation
S=50

# sample size
n = naug = 300

# total number of predictors
p = 100

# note:  pC, pP and pI are number of confounders,
#pure predictors of outcome and pure predictors of exposure, respectively
pC = pP = pI  = 2

# pS number of spurious covariates
pS = p - (pC+pP+pI)

# list of all p variables
var.list = c(paste("Xc",1:pC,sep=""),
             paste("Xp",1:pP,sep=""),
             paste("Xi",1:pI,sep=""),
             paste("Xs",1:pS,sep=""))

# list of threshold variables
#var.list_Ball = c(paste("X",1:threshold,sep=""))

# set strength of relationship between covariates and outcome
beta_v =  c( 0.6, 0.6, 0.6, 0.6, 0, 0, rep(0,p-6) )

# Set strength of relationship between covariates and treatment
alpha_v = c( 1, 1, 0, 0, 1, 1,  rep(0,p-6) )
names(beta_v) = names(alpha_v) = var.list

# set true average treatment effect (taken from Shortreed and Ertefaie (2017))
bA = 0


Data=NULL
```

```r
### simulate data
Sigma_x = matrix(rho*sig_x^2,nrow=length(var.list),ncol=length(var.list))
diag(Sigma_x) = sig_x^2
Mean_x = rep(mean_x,length(var.list))
Data = as.data.frame(MASS::mvrnorm(n = n,mu=Mean_x,Sigma = Sigma_x,empirical = FALSE))
names(Data) = var.list

X=Data



## Data generation setting
## alpha: Xc's scale is 0.2 0.2 and Xi's scale is 0.3 0.3
## so this refers that there is 2 Xc and Xi
## beta: Xc's scale is 2 2 and Xp's scale is 2 2
## so this refers that there is 2 Xc and Xp
## rest with following setup
Data_fun <- Data_G(X, alpha_v = c( 1, 1, 0, 0, 1, 1,  rep(0,p-6) )
, beta_v = c( 0.6, 0.6, 0.6, 0.6, 0, 0, rep(0,p-6) )
, bA = 0, sig_x=sig_x, linearY=TRUE,pC=2,pP=2,pI=2)

X=Data_fun$X
A=Data_fun$A
Y=Data_fun$Y
OAL(X,A,Y)
```

```
## Selected variables are:
##  [1] "Xc1"  "Xc2"  "Xp1"  "Xp2"  "XS14" "XS17" "XS20" "XS29" "XS31" "XS32"
## [11] "XS40" "XS49" "XS56" "XS58" "XS59" "XS61" "XS75" "XS76" "XS82" "XS86"
## [21] "XS88" "XS89"
```

```
##   [1] -0.02424637  1.00000000  1.00000000  1.00000000  1.00000000  0.00000000
##   [7]  0.00000000  0.00000000  0.00000000  0.00000000  0.00000000  0.00000000
##  [13]  0.00000000  0.00000000  0.00000000  0.00000000  0.00000000  0.00000000
##  [19]  0.00000000  0.00000000  1.00000000  0.00000000  0.00000000  1.00000000
##  [25]  0.00000000  0.00000000  1.00000000  0.00000000  0.00000000  0.00000000
##  [31]  0.00000000  0.00000000  0.00000000  0.00000000  0.00000000  1.00000000
##  [37]  0.00000000  1.00000000  1.00000000  0.00000000  0.00000000  0.00000000
##  [43]  0.00000000  0.00000000  0.00000000  0.00000000  1.00000000  0.00000000
##  [49]  0.00000000  0.00000000  0.00000000  0.00000000  0.00000000  0.00000000
##  [55]  0.00000000  1.00000000  0.00000000  0.00000000  0.00000000  0.00000000
##  [61]  0.00000000  0.00000000  1.00000000  0.00000000  1.00000000  1.00000000
##  [67]  0.00000000  1.00000000  0.00000000  0.00000000  0.00000000  0.00000000
##  [73]  0.00000000  0.00000000  0.00000000  0.00000000  0.00000000  0.00000000
##  [79]  0.00000000  0.00000000  0.00000000  1.00000000  1.00000000  0.00000000
##  [85]  0.00000000  0.00000000  0.00000000  0.00000000  1.00000000  0.00000000
##  [91]  0.00000000  0.00000000  1.00000000  0.00000000  1.00000000  1.00000000
##  [97]  0.00000000  0.00000000  0.00000000  0.00000000  0.00000000
```

```r
GOAL(X,A,Y)
```

```
## Selected variables are:
##  [1] "Xc1"  "Xc2"  "Xp1"  "Xp2"  "XS14" "XS17" "XS20" "XS29" "XS31" "XS32"
## [11] "XS40" "XS45" "XS47" "XS49" "XS56" "XS58" "XS59" "XS61" "XS67" "XS75"
## [21] "XS76" "XS82" "XS86" "XS88" "XS89"
```

```
##   [1] -0.2329178  1.0000000  1.0000000  1.0000000  1.0000000  0.0000000
##   [7]  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000
##  [13]  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000
##  [19]  0.0000000  0.0000000  1.0000000  0.0000000  0.0000000  1.0000000
##  [25]  0.0000000  0.0000000  1.0000000  0.0000000  0.0000000  0.0000000
##  [31]  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  1.0000000
##  [37]  0.0000000  1.0000000  1.0000000  0.0000000  0.0000000  0.0000000
##  [43]  0.0000000  0.0000000  0.0000000  0.0000000  1.0000000  0.0000000
##  [49]  0.0000000  0.0000000  0.0000000  1.0000000  0.0000000  1.0000000
##  [55]  0.0000000  1.0000000  0.0000000  0.0000000  0.0000000  0.0000000
##  [61]  0.0000000  0.0000000  1.0000000  0.0000000  1.0000000  1.0000000
##  [67]  0.0000000  1.0000000  0.0000000  0.0000000  0.0000000  0.0000000
##  [73]  0.0000000  1.0000000  0.0000000  0.0000000  0.0000000  0.0000000
##  [79]  0.0000000  0.0000000  0.0000000  1.0000000  1.0000000  0.0000000
##  [85]  0.0000000  0.0000000  0.0000000  0.0000000  1.0000000  0.0000000
##  [91]  0.0000000  0.0000000  1.0000000  0.0000000  1.0000000  1.0000000
##  [97]  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000
```

Methods Tailored to Ultra-High Dimensional Scenarios:

```r
rm(list=ls())
set.seed(2015)
## Generate a multivariate normal X matrix
# set information for simulating coviariates
mean_x = 0
sig_x = 1

# pairwise correlation between covariates
rho = 0.5

# set number of monte carlo (MC) simulation
S=50

# sample size
n = naug = 600

# total number of predictors
p = 1000
# note:  pC, pP and pI are number of confounders,
#pure predictors of outcome and pure predictors of exposure, respectively
pC = pP = pI  = 2

# pS number of spurious covariates
pS = p - (pC+pP+pI)

# list of all p variables
var.list = c(paste("Xc",1:pC,sep=""),
             paste("Xp",1:pP,sep=""),
             paste("Xi",1:pI,sep=""),
             paste("Xs",1:pS,sep=""))

# list of threshold variables
#var.list_Ball = c(paste("X",1:threshold,sep=""))
```

```r
# set strength of relationship between covariates and outcome
beta_v =  c( 0.6, 0.6, 0.6, 0.6, 0, 0, rep(0,p-6) )

# Set strength of relationship between covariates and treatment
alpha_v = c( 1, 1, 0, 0, 1, 1,  rep(0,p-6) )
names(beta_v) = names(alpha_v) = var.list

# set true average treatment effect (taken from Shortreed and Ertefaie (2017))
bA = 0


Data=NULL
### simulate data
Sigma_x = matrix(rho*sig_x^2,nrow=length(var.list),ncol=length(var.list))
diag(Sigma_x) = sig_x^2
Mean_x = rep(mean_x,length(var.list))
Data = as.data.frame(MASS::mvrnorm(n = n,mu=Mean_x,Sigma = Sigma_x,empirical = FALSE))
names(Data) = var.list

X=Data



## Data generation setting
## alpha: Xc's scale is 0.2 0.2 and Xi's scale is 0.3 0.3
## so this refers that there is 2 Xc and Xi
## beta: Xc's scale is 2 2 and Xp's scale is 2 2
## so this refers that there is 2 Xc and Xp
## rest with following setup
Data_fun <- Data_G(X, alpha_v = c( 1, 1, 0, 0, 1, 1,  rep(0,p-6) )
, beta_v = c( 0.6, 0.6, 0.6, 0.6, 0, 0, rep(0,p-6) )
, bA = 0, sig_x=sig_x, linearY=TRUE,pC=2,pP=2,pI=2)

X=Data_fun$X
A=Data_fun$A
Y=Data_fun$Y
CBS(X,A,Y)
```

```
## Selected variables are:
##  [1] "Xc1"   "Xc2"   "Xp1"   "Xp2"   "XS31"  "XS66"  "XS200" "XS447" "XS472"
## [10] "XS493" "XS514" "XS576" "XS723" "XS857" "XS868" "XS930" "XS931" "XS989"
```

```
##  [1] -0.2504782  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000
##  [7]  0.0000000  0.0000000  1.0000000  0.0000000  0.0000000  1.0000000
## [13]  0.0000000  0.0000000  0.0000000  1.0000000  1.0000000  1.0000000
## [19]  0.0000000  1.0000000  1.0000000  0.0000000  1.0000000  0.0000000
## [25]  1.0000000  1.0000000  0.0000000  0.0000000  1.0000000  1.0000000
## [31]  1.0000000
```

```r
SIS_OAL(X,A,Y)
```

```
## Selected variables are:
## [1] "Xc1"   "Xc2"   "Xp1"   "Xp2"   "XS40"  "XS200" "XS752" "XS874" "XS885"
```

```
##  [1] 1.093364 1.000000 1.000000 1.000000 1.000000 0.000000 1.000000 0.000000
##  [9] 0.000000 0.000000 0.000000 1.000000 0.000000 0.000000 0.000000 0.000000
```

```
## [17] 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 1.000000
## [25] 0.000000 0.000000 1.000000 1.000000 0.000000 0.000000 0.000000
```

```
SIS_GOAL(X,A,Y)
```

```
## Selected variables are:
##  [1] "Xc1"   "Xc2"   "Xp1"   "Xp2"   "XS40"  "XS200" "XS264" "XS348" "XS499"
## [10] "XS501" "XS502" "XS523" "XS550" "XS569" "XS573" "XS586" "XS621" "XS644"
## [19] "XS678" "XS710" "XS719" "XS752" "XS868" "XS874" "XS880" "XS885" "XS908"
## [28] "XS930" "XS932" "XS939" "XS948" "XS951"
```

```
##  [1] -0.1630836  1.0000000  1.0000000  1.0000000  1.0000000  0.0000000
##  [7]  0.0000000  1.0000000  0.0000000  0.0000000  0.0000000  0.0000000
## [13]  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000
## [19]  0.0000000  0.0000000  1.0000000  0.0000000  0.0000000  0.0000000
## [25]  0.0000000  1.0000000  0.0000000  0.0000000  0.0000000  0.0000000
## [31]  1.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000
## [37]  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000
## [43]  1.0000000  1.0000000  1.0000000  0.0000000  1.0000000  0.0000000
## [49]  0.0000000  1.0000000  0.0000000  0.0000000  1.0000000  1.0000000
## [55]  0.0000000  1.0000000  0.0000000  0.0000000  0.0000000  1.0000000
## [61]  0.0000000  1.0000000  1.0000000  1.0000000  1.0000000  0.0000000
## [67]  0.0000000  0.0000000  0.0000000  1.0000000  0.0000000  0.0000000
## [73]  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  1.0000000
## [79]  0.0000000  1.0000000  1.0000000  1.0000000  1.0000000  0.0000000
## [85]  0.0000000  1.0000000  0.0000000  1.0000000  0.0000000  1.0000000
## [91]  1.0000000  1.0000000  0.0000000  0.0000000
```

Radiomic Data Example:

- Gliosarcoma :

```
rm(list=ls())
set.seed(2015)
library(readxl)
set.seed(2015)
GData <- read_excel("/Users/rimenaaman/Downloads/GBM_GBS.xlsx")
Gdat <- GData[3:186,]
l=t(GData[3,])
var.listGSM=l[2:1303]
Y = c(rep(0, 100), rep(1, 83))
A = as.numeric(Gdat[2:184,6] == "Y")
Conf <- Gdat[2:184,7:1309]
char_columns <- sapply(Conf, is.character)
data_chars_as_num <- Conf
data_chars_as_num[,char_columns]<-as.data.frame(apply(data_chars_as_num[,char_columns],2,as.numeric))
invisible(sapply(data_chars_as_num,class))
X <- data_chars_as_num
colnames(X)=var.listGSM
CBS(X,A,Y)
```

```
## Selected variables are:
## [1] "Necrosis"
## [2] "original_shape_Elongation"
## [3] "original_shape_Maximum2DDiameterColumn"
## [4] "log-sigma-5-0-mm-3D_firstorder_InterquartileRange"
```

```
##  [1] 0.2502798 1.0000000 1.0000000 0.0000000 0.0000000 1.0000000 0.0000000
##  [8] 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000
## [15] 1.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000
## [22] 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000
## [29] 0.0000000 0.0000000 0.0000000
```

```
SIS_GOAL(X, A, Y)
```

```
## Selected variables are:
##  [1] "original_shape_Elongation"
##  [2] "original_shape_LeastAxis"
##  [3] "original_shape_Maximum2DDiameterColumn"
##  [4] "original_glszm_HighGrayLevelZoneEmphasis"
##  [5] "original_glszm_SmallAreaLowGrayLevelEmphasis"
##  [6] "log-sigma-3-0-mm-3D_firstorder_Entropy"
##  [7] "log-sigma-3-0-mm-3D_firstorder_Mean"
##  [8] "log-sigma-3-0-mm-3D_glszm_HighGrayLevelZoneEmphasis"
##  [9] "log-sigma-3-0-mm-3D_glszm_SmallAreaLowGrayLevelEmphasis"
## [10] "log-sigma-5-0-mm-3D_firstorder_Entropy"
## [11] "log-sigma-5-0-mm-3D_firstorder_InterquartileRange"
## [12] "log-sigma-5-0-mm-3D_firstorder_Mean"
## [13] "wavelet-LHL_gldm_LargeDependenceHighGrayLevelEmphasis"
## [14] "wavelet-LHL_gldm_SmallDependenceHighGrayLevelEmphasis"
## [15] "wavelet-LHH_glcm_Imc1"
## [16] "wavelet-HLH_firstorder_Kurtosis"
## [17] "wavelet-HHL_firstorder_Median"
```

```
##  [1] 0.235497 0.000000 1.000000 0.000000 1.000000 1.000000 0.000000 0.000000
##  [9] 0.000000 1.000000 1.000000 1.000000 0.000000 1.000000 1.000000 1.000000
## [17] 1.000000 1.000000 1.000000 0.000000 0.000000 0.000000 0.000000 0.000000
## [25] 0.000000 1.000000 1.000000 0.000000 0.000000 1.000000 1.000000 0.000000
## [33] 1.000000 0.000000 0.000000 0.000000
```

```
SIS_OAL(X,A,Y)
```

```
## Selected variables are:
## [1] "original_shape_Elongation"
## [2] "original_shape_LeastAxis"
## [3] "original_glszm_HighGrayLevelZoneEmphasis"
## [4] "original_glszm_SmallAreaLowGrayLevelEmphasis"
## [5] "log-sigma-5-0-mm-3D_firstorder_Entropy"
## [6] "log-sigma-5-0-mm-3D_firstorder_Mean"
## [7] "wavelet-LHL_gldm_SmallDependenceHighGrayLevelEmphasis"
```

```
##  [1] 0.2720792 0.0000000 1.0000000 0.0000000 1.0000000 0.0000000 0.0000000
##  [8] 0.0000000 1.0000000 1.0000000 0.0000000 0.0000000 0.0000000 1.0000000
## [15] 0.0000000 1.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000
## [22] 0.0000000 1.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000
## [29] 0.0000000 0.0000000 0.0000000
```

- Osteosarcoma :

```
rm(list=ls())
set.seed(2015)
RData <- read.csv("/Users/rimenaaman/Downloads/osteosarcoma.csv")

# Data loading
```

```r
Y = as.numeric(RData$tumor.volume == 'effective group')
A = as.numeric(RData$Surgical.staging == " ")

# Convert character columns to numeric
Conf = RData[,10:1418]
char_columns <- sapply(Conf, is.character)
Conf[ , char_columns] <- as.data.frame(apply(Conf[ , char_columns], 2, as.numeric))
X = as.matrix(Conf)

# STABILIZATION: add small noise BEFORE calling the functions
X <- X + matrix(rnorm(nrow(X) * ncol(X), mean = 0, sd = 1e-8), nrow = nrow(X))

CBS(X,A,Y)
```

```
## Selected variables are:
## [1] "Uniformity.2"                    "LowGrayLevelRunEmphasis.2"
## [3] "ZoneVariance.2"                  "SmallAreaLowGrayLevelEmphasis.2"
## [5] "Correlation.6"                   "SizeZoneNonUniformity.6"
## [7] "ShortRunLowGrayLevelEmphasis.10" "LongRunLowGrayLevelEmphasis.10"
```

```
##  [1] -0.2327393  1.0000000  1.0000000  1.0000000  0.0000000  0.0000000
##  [7]  1.0000000  0.0000000  0.0000000  1.0000000  1.0000000  0.0000000
## [13]  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000
## [19]  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000
## [25]  1.0000000  0.0000000  0.0000000  1.0000000  0.0000000  0.0000000
## [31]  0.0000000
```

```r
SIS_OAL(X, A, Y)
```

```
## Selected variables are:
##  [1] "ZoneVariance.2"
##  [2] "GrayLevelNonUniformity.8"
##  [3] "JointAverage.8"
##  [4] "SumAverage.8"
##  [5] "SmallDependenceEmphasis.10"
##  [6] "SmallDependenceHighGrayLevelEmphasis.10"
##  [7] "LargeDependenceLowGrayLevelEmphasis.10"
##  [8] "SmallDependenceLowGrayLevelEmphasis.10"
##  [9] "ShortRunLowGrayLevelEmphasis.10"
## [10] "ShortRunEmphasis.10"
```

```
##  [1] -0.3344088  0.0000000  0.0000000  1.0000000  1.0000000  0.0000000
##  [7]  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000
## [13]  1.0000000  1.0000000  0.0000000  0.0000000  0.0000000  0.0000000
## [19]  0.0000000  0.0000000  1.0000000  1.0000000  1.0000000  1.0000000
## [25]  1.0000000  1.0000000  0.0000000  0.0000000  0.0000000  0.0000000
## [31]  0.0000000
```

```r
SIS_GOAL(X,A,Y)
```

```
## Selected variables are:
##  [1] "ZoneVariance.2"
##  [2] "GrayLevelNonUniformity.8"
##  [3] "JointAverage.8"
##  [4] "SumAverage.8"
##  [5] "ClusterShade.8"
```

```
##  [6] "SmallDependenceEmphasis.10"
##  [7] "LargeDependenceLowGrayLevelEmphasis.10"
##  [8] "SmallDependenceLowGrayLevelEmphasis.10"
##  [9] "ShortRunLowGrayLevelEmphasis.10"
## [10] "ShortRunEmphasis.10"
## [11] "LongRunLowGrayLevelEmphasis.10"
```

```
##  [1] -0.3043722  0.0000000  0.0000000  1.0000000  1.0000000  0.0000000
##  [7]  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  1.0000000
## [13]  1.0000000  1.0000000  0.0000000  1.0000000  1.0000000  1.0000000
## [19]  1.0000000  1.0000000  1.0000000  0.0000000  0.0000000
```

References

- In Song Kim, Phil Martin, Nina McMurry, Andy Halterman (2018). *Instructions for Creating Your own R Package*.
- Bai, Y., Yang, Y., & Baldé, I. (2023). *Generalized outcome adaptive lasso for high dimensional causal inference*. arXiv:2310.06315.
- Acharya, A., Blackwell, M., & Sen, M. (2016). *Explaining causal findings without bias: Detecting and assessing direct effects. American Political Science Review*, **110**(3), 512–529.
- Baldé, I., Yang, Y. A., & Lefebvre, G. (2022). *Reader reaction to "Outcome-adaptive lasso: Variable selection for causal inference" by Shortreed and Ertefaie (2017). Biometrics*, 1–7. https://doi.org/10.1111/biom.13683
- Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., & Stürmer, T. (2006). *Variable selection for propensity score models. American Journal of Epidemiology*, 163(12), **1149–1156**.
- Baldé, I., Yang, A. Y. & Lefebvre, G. (2023). *Reader Reaction to "Outcome-adaptive lasso: Variable selection for causal inference" by Shortreed and Ertefaie (2017). Biometrics*, 79(1), **514–520**.
- De Luna, X., Waernbaum, I., & Richardson, T. (2011). *Covariate selection for the nonparametric estimation of an average treatment effect. Biometrika*, 98(4), **861–875**.
- Fan, J., & Lv, J. (2008). *Sure independence screening for ultrahigh dimensional feature space. Journal of the Royal Statistical Society: Series B*, 70(5), **849–911**.
- Islam, M. S., & Noor-E-Alam, M. (2021). *Feature selection for causal inference from high dimensional observational data with outcome adaptive elastic net*. https://arxiv.org/abs/2111.13800
- Koch, B., Vock, D. M., & Wolfson, J. (2018). *Covariate selection with group lasso and doubly robust estimation of causal effects. Biometrics*, **74**(1), 8–17.
- Pearl, J. (2009). *Causality: Models, Reasoning, and Inference* (2nd ed.). Cambridge University Press.
- Rosenbaum, P. R., & Rubin, D. B. (1983). *The central role of the propensity score in observational studies for causal effects. Biometrika*, 70(1), **41–55.**
- Rubin, D. B. (1974). *Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology*, 66(5), **688–701**.
- Shortreed, S. M., & Ertefaie, A. (2017). *Outcome-adaptive lasso: Variable selection for causal inference. Biometrics*, 73(4), **1111–1122**.
- Tibshirani, R. (1996). *Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B*, **58**(1), 267–288.
- Zou, H. (2006). *The adaptive lasso and its oracle properties. Journal of the American Statistical Association*, 101(476), **1418–1429**.
- Zou, H., & Hastie, T. (2005). *Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B*, 67(2), **301–320**.
- Barut, E., Fan, J. and Verhasselt, A. (2016). *Conditional sure independence screening. Journal of the American Statistical Association 111*, **1266–1277**.
- Brookhart, M., Schneeweiss, S., Rothman, K., Glynn, R., Avorn, J. and Sturmer, T. (2006). *Variable selection for propensity score models. American Journal of Epidemiology 163*, **1149–1156**.
- De Luna, X., Waernbaum, I. and Richardson, T. S. (2011). *Covariate selection for the nonparametric estimation of an average treatment effect. Biometrika 98(4)*, **861–875**.
- Ertefaie, A., Asgharian, M. and Stephens, D. A. (2018). *Variable Selection in Causal Inference using a Simultaneous Penalization Method. Journal of Causal Inference 6 (1)*, **20170010**.

- Fan, J. and Lv, J. (2008). *Sure independence screening for ultrahigh dimensional feature space. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70,***849–911**.
- Ghosh, D., Mastej, E., Jain, R. and Choi, Y. S. (2022). *Causal Inference in Radiomics: Framework, Mechanisms, and Algorithms. Frontiers in Neuroscience* , **16:884708**.
- Miller, C. R. and Perry, A. (2007). *Glioblastoma: Morphologic and Molecular Genetic Diversity. Archives of Pathology and Laboratory Medicine 131 (3),***397–406**.
- Mirchia, K., Mahoney, M. T., Christie, O., Fuller, C. E., Mirchia, K. and Fuller, C. (2023). *A Rare Tumor in a Rare Location: Radiology and Pathology Findings With a Literature. Review on Intraventricular Gliosarcoma. Cureus 15 (2).*
- Ohgaki, H. (2009). *Epidemiology of Brain Tumors. Cancer Epidemiology 472*, **323–342**.
- Pan, W., Wang, X., Xiao, W. and Zhu, H. (2018). *A generic sure independence screening procedure. Journal of the American Statistical Association 114*, **928–937**.
- Pan, W., Wang, X., Zhang, H., Zhu, H. and Zhu, J. (2020). *Ball covariance: a generic measure of dependence in Banach space. Journal of the American Statistical Association 115*, **307–317**.
- Patrick, A., Schneeweiss, S., Brookhart, M., Glynn, R., Rothman, K., Avorn, J., et al. (2011).*The implications of propensity score variable selection strategies in pharmacoepidemiology: An empirical illustration. Pharmacoepidemiology and Drug Safety 20*, **551–559**.
- Pearl, J. and Mackenzie, D. (2018). *The book of why: The new science of cause and effect. New York: Basic Books* .
- Qian Z, Zhang L, Hu J, Chen S, Chen H, Shen H, Zheng F, Zang Y and Chen X. (2021). *Machine Learning Based Analysis of Magnetic Resonance Radiomics for the Classification of Gliosarcoma and Glioblastoma.* **Frontiers in Oncology 11: 699789.**
- Rosenbaum, P. R. and Rubin, D. B. (1983). *The central role of the propensity score in observational studies for causal effects. Biometrika 70*, **41–55**.
- Shortreed, S. M. and Ertefaie, A. (2017). *Outcome-adaptive lasso: Variable selection for causal inference. Biometrics 73(4)*, **1111–1122**.
- Tamimi, A. F. and Juweid, M. (2017). *Epidemiology and outcome of glioblastoma. Glioblas- toma, Codon Publications: Brisbane, Australia; Chapter 8* ,**143–153**.
- Tang, D., Kong, D., Pan, W. and Wang, L. (2022). *Ultra-high dimensional variable selection for doubly robust causal inference. Biometrics* , **1–12**.
- Van Timmeren, J. E., Cester, D., Tanadini-Lang, S., Alkadhi, H. and Baessler, B. (2020). *Radiomics in medical imaging –"how – to" guide and critical reflection. Insights Imaging 11 (1): 91.*
- Zigler, C. M. and Dominici, F. (2014). *Uncertainty in Propensity Score Estimation: Bayesian. Methods for Variable Selection and Model Averaged Causal Effects. Journal of the American Statistical 109*, **95–107**.
- Zou, H. (2006). *The adaptive lasso and its oracle properties. Journal of the American Statistical Association: Series B 101*, **1418–1429**.
- Zou, H. and Zhang, H. H. (2009). *On the adaptive elastic-net with a diverging number of parameters. The Annals of Statistics 37*, **1733–1751**.