

Solar: a least-angle regression for accurate and stable variable selection in high-dimensional data

Ning Xu

November 12, 2019

Overview

1 Motivation

- Dimensions, computataion load and complicated structure
- Motivating examples

2 Solar algorithm

- Pseudo code
- Computation flow

3 Simulation results

- p/n approaching 0
- p/n approaching 1
- p/n does not change
- irrepresentable condition simulation

The curse of dimensionality

- Dimensionality : p (number of variables) for natural language processing can easily go beyond 10 million;
- Greater dimensions requires
 - ① improvements of the variable selection algorithm (e.g., better sparsity and accuracy);
 - ② restraining the growth of computation load for variable selection
- More complicated dependence structures in datasets (more severe multicollinearity if the structure is linear);

These challenges compel variable-selection algorithms to enhance the accuracy, stability and robustness of variable-selection in high dimensional spaces.

Bayesian Network (Koller and Friedman, 2009)

Directed acyclic graphs (DAG) are used to describe dependence structure of the data.

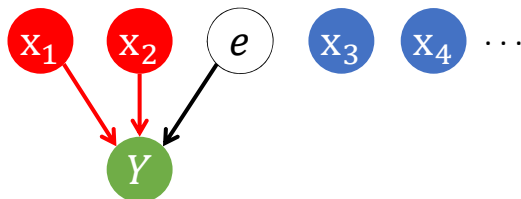


Figure: the DAG (referred to as the *standard structure*) typically assumed in the regression analysis, where $\{x_1, x_2\}$ are the 'parents' of Y ; Y is the 'child' of $\{x_1, x_2\}$ and x_1 is a 'spouse' of x_2 .

Example 1 (Lim and Yu, 2016)

Assuming the standard strcture, consider the lars-lasso algorithm, which uses the L_1 -norm fraction $t \in [0, 1]$ as a tuning parameter. In each CV training-validation split, for all β on the solution path,

$$t = \frac{\|\beta\|_1}{\|\beta_{\max}\|_1}, \quad (1)$$

where $\|\beta_{\max}\|_1$ is defined as the L_1 norm of the non-shrunked solution on the solution path.

Example 1 (Lim and Yu, 2016)

The solution path of lars-lasso is unstable in high dimensions (also applies to coordinate descent of lasso).

- When $p > n$, β_{\max} is the traditional forward regression solution with n selected variables.
- β_{\max} uses up all n degrees of freedom (a saturated fit)
- Hence, due to the resampling randomness in CV, $\|\beta_{\max}\|_1$ may vary substantially across validation sets.
- As a result, the same value of t may correspond to different amounts of shrinkage (or λ) on the solution paths of different CV training-validation splits.

Example 1 (Lim and Yu, 2016)

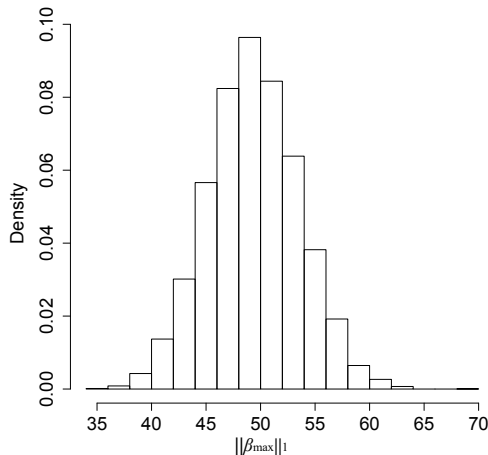


Figure: Histogram of $\|\beta_{\max}\|_1$ from 10,000 bootstrap lasso estimates from a Gaussian simulation, where the variance of each variable is 1, pairwise correlations are 0.5, $n = 100$ and $p = 150$ (Lim and Yu, 2016, Section 3.1.1.).

Example 2. (Zhao and Yu, 2006)

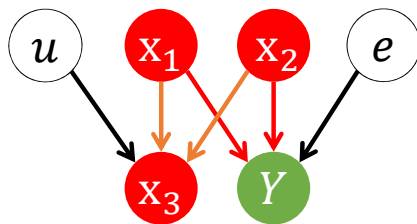


Figure: Example 2 dependence structure.

$$\begin{cases} \mathbf{x}_3 = \omega_1 \mathbf{x}_1 + \omega_2 \mathbf{x}_2 + \sqrt{1 - 2\omega^2} \cdot u \\ Y = \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \delta e \end{cases} \quad (2)$$

irrepresentable condition: for variable selection consistency of any lasso-type estimator, $\sum_i |\omega_i| < 1$

Example 2. (Zhao and Yu, 2006)

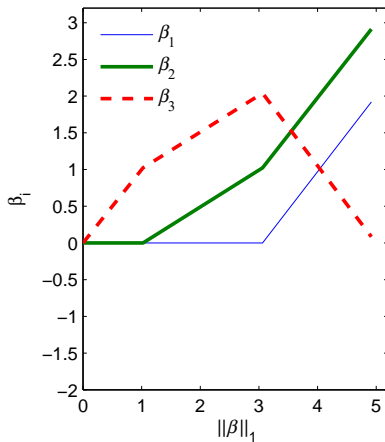


Figure: $\sum_i |\omega_i| > 1$: the irrepresentable condition fails and lasso selects redundant variables

Solar algorithm

Algorithm 1: average L_0 solution path estimation

input : (Y, X) .

```
1 generate  $K$  subsamples  $\{(Y^k, X^k)\}_{k=1}^K$ ;  
2 set  $\tilde{p} = \min\{n_{\text{sub}}, p\}$ ;  
3 for  $k := 1$  to  $K$ , stepsize = 1 do  
4   run an unrestricted lars (Algorithm 3.2 in Friedman et al. (2001)) on  
    $(Y^k, X^k)$  and record the order of variable inclusion at each stage;  
5   define  $\hat{q}^k = \mathbf{0} \in \mathbb{R}^p$ ;  
6   for all  $i$  and  $l$ , if  $\mathbf{x}_i$  is included at stage  $l$ , set  $\hat{q}_i^k = (\tilde{p} + 1 - l)/\tilde{p}$ , where  
    $\hat{q}_i^k$  is the  $i^{\text{th}}$  entry of  $\hat{q}^k$ ;  
7 end  
8  $\hat{q} := \frac{1}{K} \sum_{k=1}^K \hat{q}^k$ ;  
9 return  $\hat{q}$ 
```

Solar algorithm

Algorithm 2: Subsample-ordered least-angle regression (solar)

input : (Y, X)

- 1 Randomly select 20% of the points in (Y, X) to be the validation set (Y_v, X_v) ; denote the remaining points (Y_r, X_r) ;
 - 2 estimate \hat{q} using Algorithm 1 on (Y_r, X_r) ;
 - 3 **for** $c := 1$ to 0, *stepsize* = -0.02 **do**
 - 4 set $Q(c) = \{\mathbf{x}_j \mid \hat{q}_j \geq c, \forall j\}$ and add all variables in $Q(c)$ into an OLS model;
 - 5 **if** *sample size of (Y_r, X_r) is not less than $|Q(c)|$* **then**
 - 6 train the OLS model on (Y_r, X_r) and compute its validation error on (Y_v, X_v) ;
 - 7 **else**
 - 8 **break** the if-else statement and for loop
 - 9 **end**
 - 10 **end**
 - 11 find c^* , the value of c associated with the minimal validation error on (Y_v, X_v) ; find $Q(c^*)$;
 - 12 **return** $\hat{q}, \beta(Q(c^*), Y)$
-

Solar algorithm demo

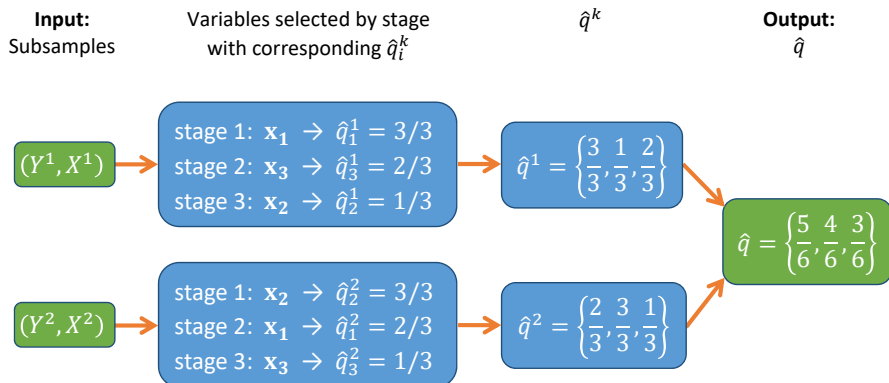


Figure: Computation of \hat{q} on 2 subsamples, where $\{\mathbf{x}_1, \mathbf{x}_2\}$ are informative and \mathbf{x}_3 is redundant.

Solar algorithm demo

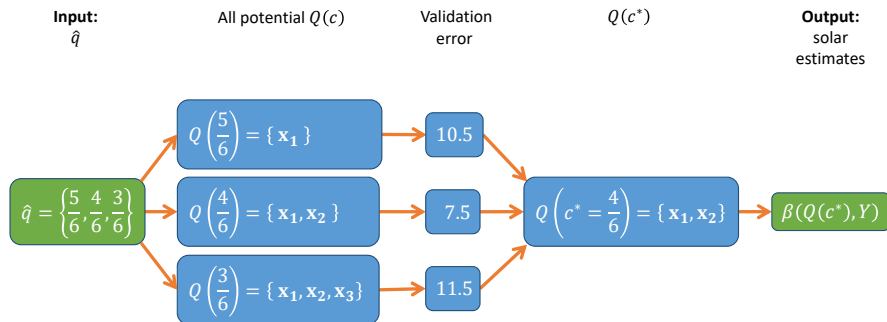


Figure: Computation flow for Algorithm 2 (continued from Figure 5).

Solar algorithm demo

- Same computation load as CV-lars-lasso: one least-angle regression on each subsample to compute \hat{q} ; Another one for selecting variable.
- less computational expensive than coordinate descend
- My goal: outperform lasso-type estimators without increasing the computation load

Comparison simulation

vs CV-lars-lasso and cross-validated, cyclic pathwise coordinate descent with warm start (CV-cd, for short)

- when p/n approaches 0 ;
- when p/n approaches 1 ;
- when n and p both increase rapidly in high-dimensional space;
- when the dependence structure gets complicated and violates the irrerepresentable condition;

Simulation settings

$$Y = X\beta + e = 2\mathbf{x}_0 + 3\mathbf{x}_1 + 4\mathbf{x}_2 + 5\mathbf{x}_3 + 6\mathbf{x}_4 + e \quad (3)$$

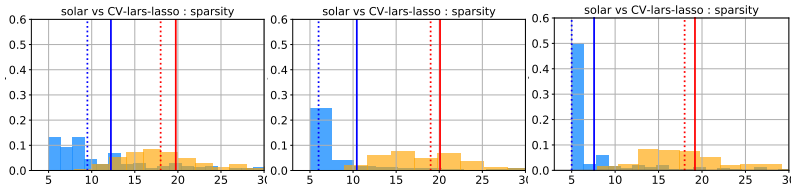
- $X \in \mathbb{R}^{n \times p}$ is generated from a zero-mean, multivariate Gaussian distribution with covariance matrix with 1 on the main diagonal and 0.5 for the off-diagonal elements.
- All data points are identically and independently distributed. Each \mathbf{x}_j is independent from the noise term e , which is standard Gaussian.
- n is the sample size and p is the dimension.
- we generate 200 samples, on each of which we run lasso solvers and solar. We average the performance of each algorithm on 200 samples.

Simulation result : p/n approaching 0

Table: Number of variables selected

| | | p/n | | |
|---|---------------|--------|---------|---------|
| | | 100/50 | 100/100 | 100/200 |
| mean | solar | 12.33 | 10.43 | 7.58 |
| | CV-lars-lasso | 19.75 | 20.09 | 19.19 |
| | CV-cd | 20.77 | 20.46 | 20.00 |
| median | solar | 9.5 | 6 | 5 |
| | CV-lars-lasso | 18 | 19 | 18 |
| | CV-cd | 18 | 19 | 18 |
| Pr(only select $\{\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4\}$) | solar | 0.025 | 0.305 | 0.560 |
| | CV-lars-lasso | 0 | 0 | 0 |
| | CV-cd | 0 | 0 | 0 |

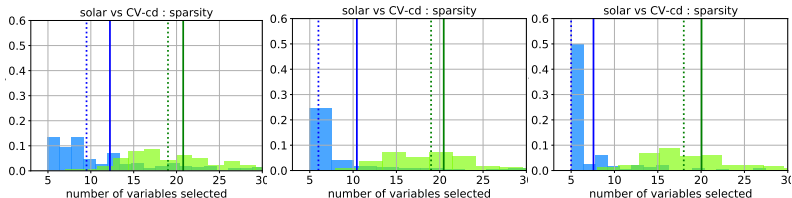
Histogram of the number of variables selected by solar



(a) $p/n = 100/50$, vs CV-lars-lasso

(b) $p/n = 100/100$, vs CV-lars-lasso

(c) $p/n = 100/200$, vs CV-lars-lasso



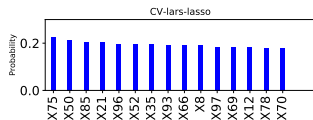
(d) $p/n = 100/50$, vs CV-cd

(e) $p/n = 100/100$, vs CV-cd

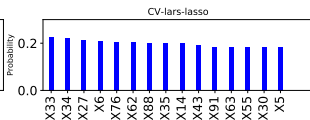
(f) $p/n = 100/200$, vs CV-cd



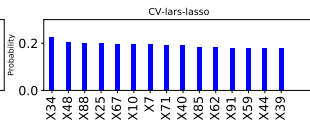
Probability of selecting redundant variables (top 15).



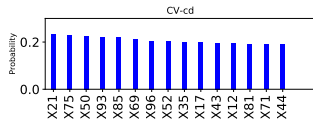
(a) $p/n = 100/50$,
CV-lars-lasso



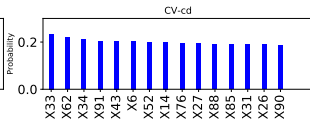
(b) $p/n = 100/100$,
CV-lars-lasso



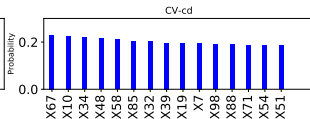
(c) $p/n = 100/200$,
CV-lars-lasso



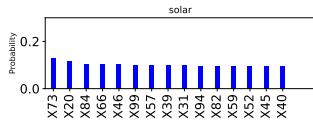
(d) $p/n = 100/50$, CV-cd



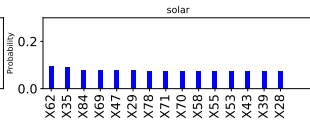
(e) $p/n = 100/100$, CV-cd



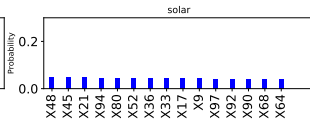
(f) $p/n = 100/200$, CV-cd



(g) $p/n = 100/50$, solar

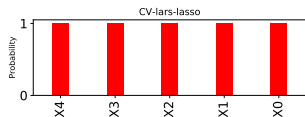


(h) $p/n = 100/100$, solar

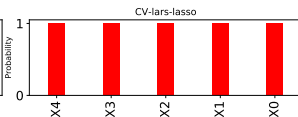


(i) $p/n = 100/200$, solar

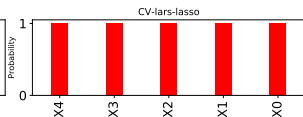
Probability of selecting informative variables.



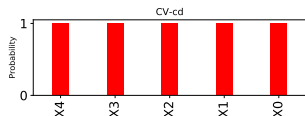
(a) $p/n = 100/50$,
CV-lars-lasso



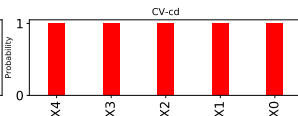
(b) $p/n = 100/100$,
CV-lars-lasso



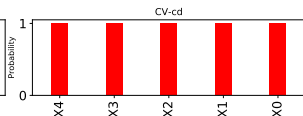
(c) $p/n = 100/200$,
CV-lars-lasso



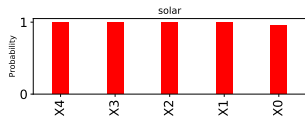
(d) $p/n = 100/50$, CV-cd



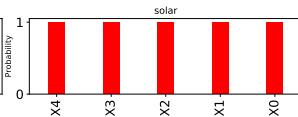
(e) $p/n = 100/100$, CV-cd



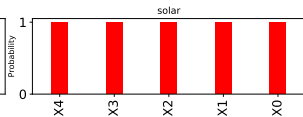
(f) $p/n = 100/200$, CV-cd



(g) $p/n = 100/50$, solar
(95% for x_0 and 99.5% for
 x_1)

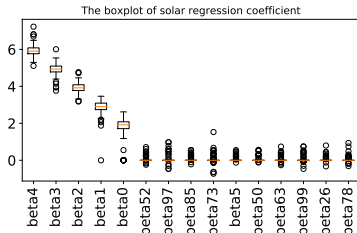


(h) $p/n = 100/100$, solar

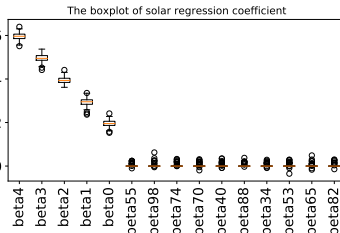


(i) $p/n = 100/200$, solar

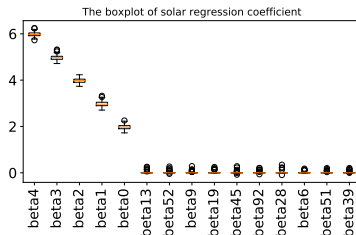
Solar regression coefficient boxplots (top 15 by mean)



(a) $p/n = 100/50$



(b) $p/n = 100/100$



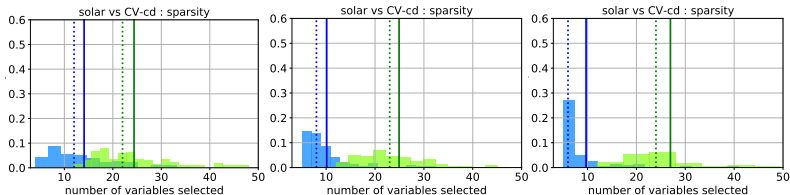
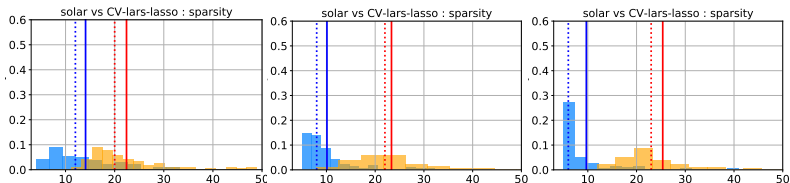
(c) $p/n = 100/200$

Simulation result: p/n approaching 1

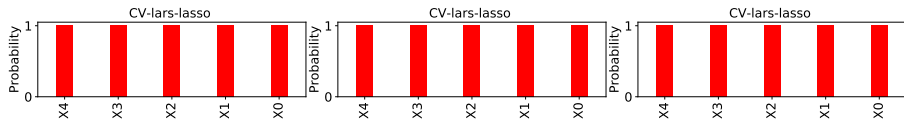
Table: Number of variables selected

| | | p/n | | |
|---|---------------|--------|---------|---------|
| | | 150/50 | 200/100 | 250/150 |
| mean | solar | 14.07 | 10.10 | 9.7 |
| | CV-lars-lasso | 22.41 | 23.34 | 25.37 |
| | CV-cd | 24.37 | 24.92 | 26.96 |
| median | solar | 12 | 8 | 6 |
| | CV-lars-lasso | 20 | 22 | 23 |
| | CV-cd | 20 | 22 | 23 |
| Pr(only select $\{\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4\}$) | solar | 0.015 | 0.115 | 0.445 |
| | CV-lars-lasso | 0 | 0 | 0 |
| | CV-cd | 0 | 0 | 0 |

Histogram of the number of variables selected by solar



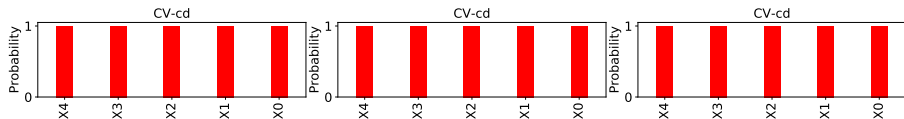
Probability of selecting informative variables



(a) $p/n = 150/50$,
CV-lars-lasso

(b) $p/n = 200/100$,
CV-lars-lasso

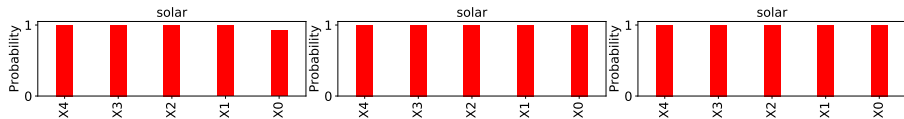
(c) $p/n = 250/150$,
CV-lars-lasso



(d) $p/n = 150/50$, CV-cd

(e) $p/n = 200/100$, CV-cd

(f) $p/n = 250/150$, CV-cd

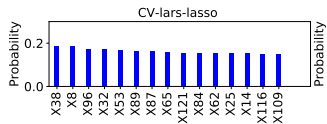


(g) $p/n = 150/50$, solar
(92.01% for x_0 and 99.5%
for x_1)

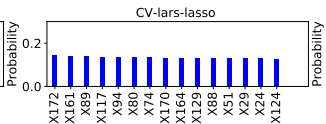
(h) $p/n = 200/100$, solar

(i) $p/n = 250/150$, solar

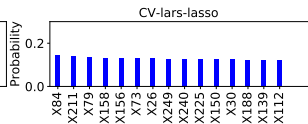
Probability of selecting redundant variables (top 15)



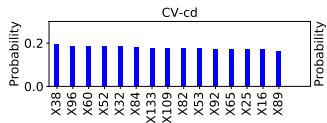
(a) $p/n = 150/50$,
CV-lars-lasso



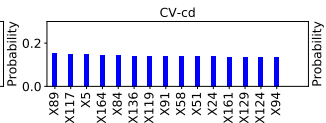
(b) $p/n = 200/100$,
CV-lars-lasso



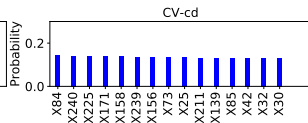
(c) $p/n = 250/150$,
CV-lars-lasso



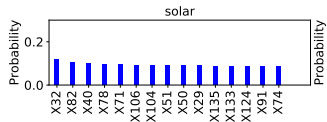
(d) $p/n = 150/50$, CV-cd



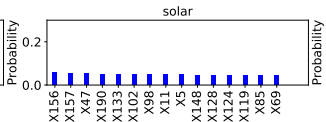
(e) $p/n = 200/100$, CV-cd



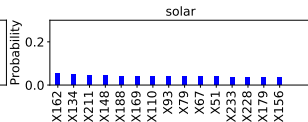
(f) $p/n = 250/150$, CV-cd



(g) $p/n = 150/50$, solar

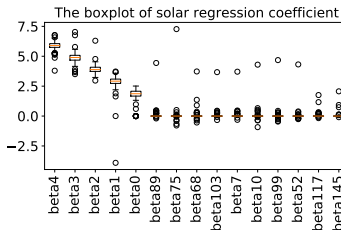


(h) $p/n = 200/100$, solar

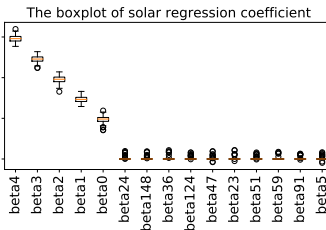


(i) $p/n = 250/150$, solar

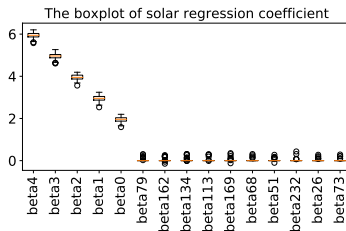
Solar regression coefficient boxplots (top 15 by mean)



(a) $p/n = 150/50$



(b) $p/n = 200/100$



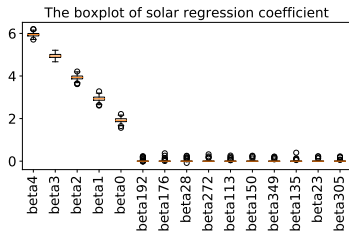
(c) $p/n = 250/150$

Simulation result: p/n does not change

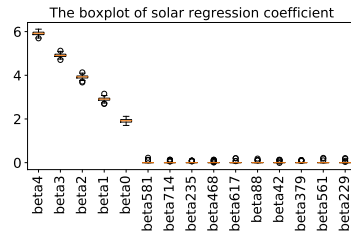
Table: Number of variables selected.

| | | p/n | | |
|---|---------------|---------|---------|----------|
| | | 400/200 | 800/400 | 1200/600 |
| mean | solar | 10.88 | 13.80 | 14.85 |
| | CV-lars-lasso | 28.17 | 33.13 | 36.90 |
| | CV-cd | 29.58 | 35.07 | 38.76 |
| median | solar | 7 | 11 | 13 |
| | CV-lars-lasso | 26 | 31 | 33 |
| | CV-cd | 26 | 31 | 33 |
| $\Pr(\text{active set} = \{\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4\})$ | solar | 0.150 | 0.010 | 0 |
| | CV-lars-lasso | 0 | 0 | 0 |
| | CV-cd | 0 | 0 | 0 |

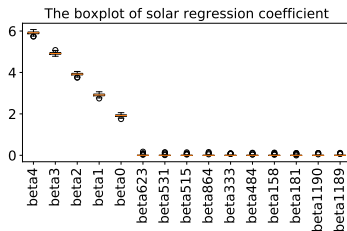
Solar regression coefficient boxplots (top 15 by mean)



(a) $p/n = 400/200$

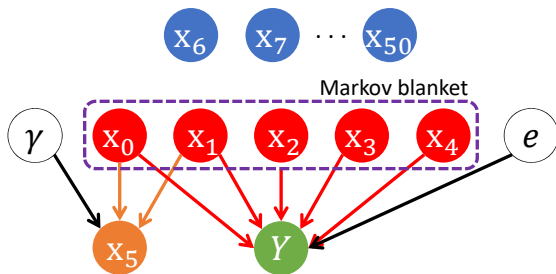


(b) $p/n = 800/400$



(c) $p/n = 1200/600$

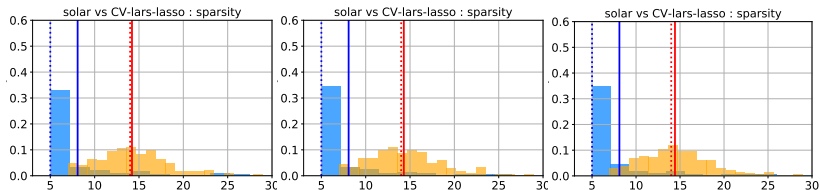
simulation setting about irrerepresentable condition



$$\begin{cases} \mathbf{x}_5 = \omega_1 \mathbf{x}_0 + \omega_2 \mathbf{x}_1 + \gamma \cdot \sqrt{1 - 2\omega^2} \\ Y = 2\mathbf{x}_0 + 3\mathbf{x}_1 + 4\mathbf{x}_2 + 5\mathbf{x}_3 + 6\mathbf{x}_4 + e \end{cases} \quad (4)$$

where $n = 150$, $p = 50$, $\omega_i \in \mathbb{R}$ and γ , e are both standard Gaussian noise terms, independent from each other and all the other variables in the simulation. By setting ω_i to either $1/4$, $1/3$ or $1/2$, the population value of $\sum_i |\omega_i|$ changes, respectively, to either $1/2$, $2/3$ or 1 .

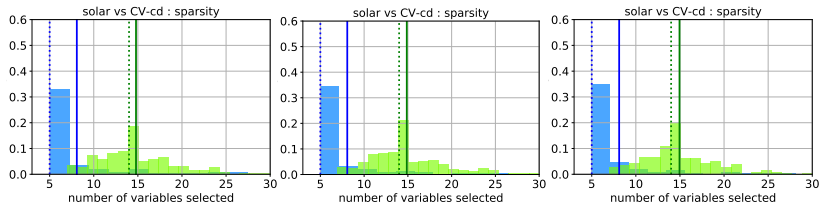
Histogram of the number of variables selected by solar



(a) $\omega = 1/4$, vs
CV-lars-lasso

(b) $\omega = 1/3$, vs
CV-lars-lasso

(c) $\omega = 1/2$, vs
CV-lars-lasso



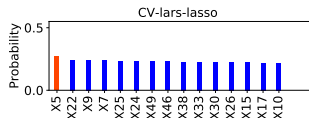
(d) $\omega = 1/4$, vs CV-cd

(e) $\omega = 1/3$, vs CV-cd

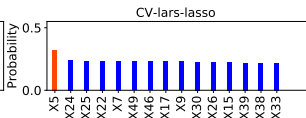
(f) $\omega = 1/2$, vs CV-cd



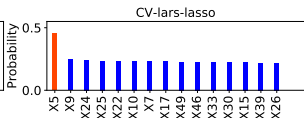
Probability of selecting redundant variables (x_5 in orange, top 15 by probability)



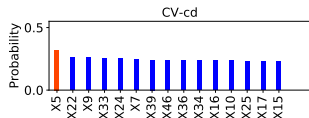
(a) $\omega = 1/4$, CV-lars-lasso



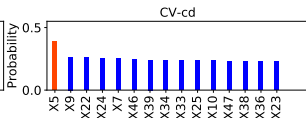
(b) $\omega = 1/3$, CV-lars-lasso



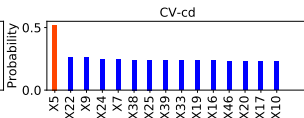
(c) $\omega = 1/2$, CV-lars-lasso



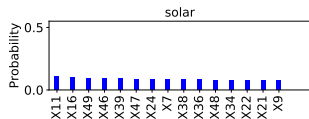
(d) $\omega = 1/4$, CV-cd



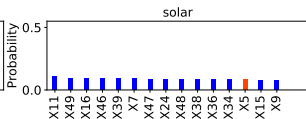
(e) $\omega = 1/3$, CV-cd



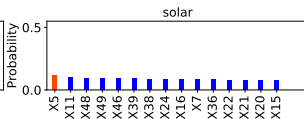
(f) $\omega = 1/2$, CV-cd



(g) $\omega = 1/4$, solar

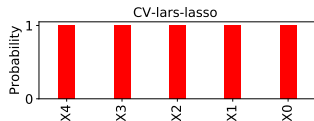


(h) $\omega = 1/3$, solar

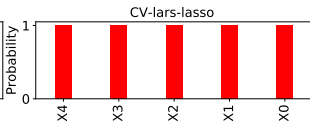


(i) $\omega = 1/2$, solar

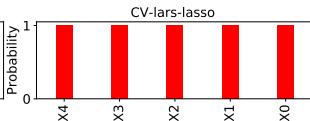
Probability of selecting informative variables



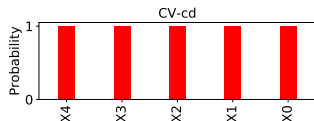
(a) $\omega = 1/4$, CV-lars-lasso



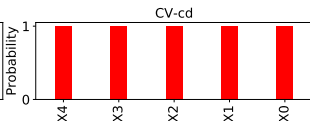
(b) $\omega = 1/3$, CV-lars-lasso



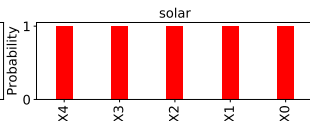
(c) $\omega = 1/2$, CV-lars-lasso



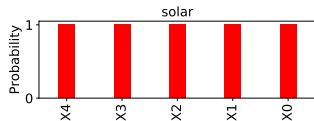
(d) $\omega = 1/4$, CV-cd



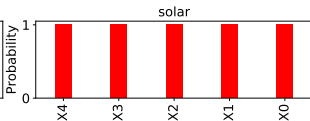
(e) $\omega = 1/3$, CV-cd



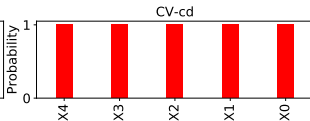
(f) $\omega = 1/2$, CV-cd



(g) $\omega = 1/4$, solar



(h) $\omega = 1/3$, solar



(i) $\omega = 1/2$, solar

Reference

- Friedman, J., Hastie, T., Tibshirani, R., 2001. The elements of statistical learning. Vol. 1 of Springer Series in Statistics. Springer-Verlag New York.
- Koller, D., Friedman, N., 2009. Probabilistic graphical models: principles and techniques. MIT press.
- Lim, C., Yu, B., 2016. Estimation stability with cross-validation (ESCV). *Journal of Computational and Graphical Statistics* 25 (2), 464–492.
- Zhao, P., Yu, B., 2006. On model selection consistency of Lasso. *Journal of Machine Learning Research* 7, 2541–2563.