Supplementary Information for

A Polynomial Algorithm for Best Subset Selection Problem

Junxian Zhu, Canhong Wen, Jin Zhu, Heping Zhang and Xueqin Wang

Heping Zhang and Xueqin Wang.
E-mail: heping.zhang@yale.edu; wangxq88@mail.sysu.edu.cn

This PDF file includes:

Supplementary text
Figs. S1 to S2
Tables S1 to S3
SI References

Supporting Information Text

Computation

We build an R package called ABESS that implements our algorithms. Two major strategies to accelerate the computation are noteworthy.

Warm start : A good initial active set can shorten the computation time. For ABESS, we can use the previous solution as an initial set for the next step.

Max splicing size : Splicing takes the most computation in ABESS. It requires the order of $k_{\max}$ operations in each iteration. We set the max splicing size $k_{max} = s$ in Algorithm 1. Based on our experience, we suggest to use a splicing size in $k_{\max}, k_{\max}/2, \ldots, 1$.

Additional simulation studies

High-dimensional case. Here, we intend to compare ABESS with more variable selection methods under the same setting in our body content. The variable selection methods take into consideration include state-of-the-art variable selection algorithms (i.e., LASSO, SCAD, and MCP) and recent proposed best subset selection solvers (i.e., SDAR, coordinate-wise minimizer, and partial swap inescapable minimizer). Analogous to LASSO/SCAD/MCP, coordinate-wise minimizer (CWM) solves ℓ_0 regularization ordinary least squares estimator via the coordinate descent algorithm (1). In regard to partial swap inescapable minimizer (PSIM) and SDAR, their solve best subset selection by local combinatorial search algorithm and primal-dual active set algorithm (1, 2). We use their *R* implementations, *ncvreg*, *glmnet*, *L0Learn* and *BeSS* (3–6), in our simulation studies. As for tuning parameters selection, we employ CV, SIC, EBIC and BIC to pick out the optimal one from l pre-specified tunning parameters, which are default values given by their implementations. All methods to be compared are summarized in Table S1. The synthetic datasets are generated following the same procedure in body content. The simulation result are exhibited in Tables S2 and S3.

We first investigate the results in the uncorrelated-covariate regime (See Tables S2). In this regime, LASSO or the best subset selection solvers (BSSS) are the best true positive detectors, and MCP or SCAD or information criterion (IC) based BSSS have the least the false positive discovery. The notable advantage of BSSS is the balance of the true and false discoveries, especially in $p \gg n$ case. As regard to parameter estimation, ABESS/SDAR/CWM/PSIM gives the most accurate estimations among all of the methods, and specifically, the SDAR-EBIC have the most precise estimation when $p \leq 1500$ and ABESS-SIC (or SDAR-SIC) takes the first place when p increases to 2500. Moreover, it can be seen that the the model size of the IC based BSSS is closed to the ground truth. In terms of computational time, the most efficient method is IC based LASSO estimator, and the second place is IC based SCAD/MCP, follow by ABESS, which is the fastest algorithms among all BSSS. Since CV and local combinatorial search are two time-consuming techniques, PSIM and CV based methods generally require much more time to conduct algorithms. It is also worthy to note that, in the uncorrelated case, ABESS, SDAR-SIC, CWM-SIC, PSIM-SIC have a quite similar performance because global optimal, coordinate-wise optimal, (i.e., SDAR-SIC and CWM-SIC) and partial swap inescapable optimal (PSIM) are likely to be equivalent in this scenario. Next, we study the simulation result under the high-correlation regime (See Tables S3). By contrast with the uncorrelated case, both TPR and TNR of almost all methods decrease in the context of high correlation, but LASSO/BSSS still have the highest TPR and MCP/SCAD/BSSS still have the highest TNR. Notably, among SIC base BSSS, ABESS has a preferable variable selection in this scenario. As for the parameter estimation, ABESS or SDAR-EBIC takes a dominant role due to the virtue of both algorithm and information criterion. Besides, the best subset selection solvers equipped with SIC or EBIC are capable of detecting a proper model size. Computationally, ABESS is the fastest best subset selection solvers and LASSO is the fastest variable selection solvers.

Comparison with SDAR and IHT. In this part, we study the convergence problems of SDAR and IHT algorithms. Consider the classical linear model setting in our main body, the rows of $\boldsymbol{X}_{n\times p}$ are *i.i.d* sampled from the multivariate normal distribution with covariance matrix Σ, where $\Sigma = (\sigma_{ij})_{p\times p}$ follows the three structures: uncorrelated (i.e., $\sigma_{ij} = \mathrm{I}(i = j)$), decayed correlation (i.e., $\sigma_{ij} = 0.8^{|i-j|}$) and constant correlation (i.e., $\sigma_{ij} = 0.8^{I(i\neq j)}$). For coefficient vector $\boldsymbol{\beta}$, it has 10 non-zero elements, half of which are 1 and half -1. The n error terms are *i.i.d* drawn from the normal distribution $N(0, \sigma^2)$, where $\sigma^2 = \boldsymbol{\beta}^\top \Sigma \boldsymbol{\beta}$ such that the signal-noise ratio is fixed at 1.

We compare IHT, SDAR and ABESS for the ℓ_0 constraint problem with a fixed support size in terms of the number of iterations. The maximum number of iterations is set to 100. We generate 100 synthetic datasets containing $n = 300$ observations with $p = 1500$ covariates. The IHT algorithm is implemented in *R* and the step size in each iteration is determined by the line search strategy. It stops when the ℓ_2-norm of two consecutive estimations is smaller than 10^{-6}. We employ the SDAR algorithm implemented in the *BeSS* package (2).

From Figure S1, the TPR and TNR of ABESS is significantly superior to the other algorithms in all of the settings, which implies a better support recovery performance of ABESS, and in the meantime, ABESS also possesses the best estimation performance in all situations. We note from the right-bottom panel of Figure S1 that SDAR may not converge in all of the cases due to the potential periodic iterative issue, and even worse, it cannot converge with a high probability when the correlation structure decays. For the IHT algorithm, its convergence rate declines as the global correlation increases. However, ABESS gets rid of both problems and generally converges to the solution within 5 iterations.

Faster sparsity level selection strategy. Golden section (GS) search (11) is a strategy for finding an extremum of a function inside a specified interval without regularity conditions such as continuity and derivative. It is a computationally cheap strategy since it avoids some redundant computation steps. By applying GS on the search of the best sparsity level, we

Table S1. All methods considered in the simulation study .

Method	Problem	Algorithm	Selector	R-package	Reference
ABESS	Best subset selection	ABESS	SIC	splicing	this paper
SDAR-EBIC	Best subset selection	SDAR	EBIC	BeSS	(2)
SDAR-SIC	Best subset selection	SDAR	SIC	BeSS	(7)
CWM-CV	Best subset selection	CWM	CV	L0Learn	(1)
CWM-SIC	Best subset selection	CWM	SIC	L0Learn	this paper
PSIM-CV	Best subset selection	PSIM	CV	L0Learn	(1)
PSIM-SIC	Best subset selection	PSIM	SIC	L0Learn	this paper
MCP-CV	Noncave regularization	CWM	CV	ncvreg	(4)
MCP-SIC	Noncave regularization	CWM	SIC	ncvreg	(8)
MCP-BIC	Noncave regularization	CWM	BIC	ncvreg	(9)
MCP-EBIC	Noncave regularization	CWM	EBIC	ncvreg	(9)
SCAD-CV	Noncave regularization	CWM	CV	ncvreg	(4)
SCAD-SIC	Noncave regularization	CWM	SIC	ncvreg	(8)
SCAD-BIC	Noncave regularization	CWM	BIC	ncvreg	(9)
SCAD-EBIC	Noncave regularization	CWM	EBIC	ncvreg	(9)
LASSO-CV	ℓ_1 regularization	CWM	CV	glmnet	(3)
LASSO-SIC	ℓ_1 regularization	CWM	SIC	glmnet	this paper
LASSO-BIC	ℓ_1 regularization	CWM	BIC	glmnet	(10)
LASSO-EBIC	ℓ_1 regularization	CWM	EBIC	glmnet	(9)

Table S2. TPR, TNR, specificities, relative errors and sparse level errors of nineteen variable selection methods for uncorrelated correlation structure.

	Method	TPR	TNR	ReErr	SLE	Runtime
$p = 500$	ABESS	0.959 (0.064)	0.999 (0.001)	0.125 (0.112)	**0.060** (0.908)	0.145 (0.029)
	SDAR-EBIC	0.959 (0.064)	0.999 (0.001)	**0.110** (**0.107**)	-0.120 (**0.795**)	0.950 (0.166)
	SDAR-SIC	0.959 (0.064)	0.999 (0.001)	0.125 (0.112)	**0.060** (0.908)	0.919 (0.167)
	CWM-CV	0.965 (0.059)	0.997 (0.003)	0.187 (0.173)	1.020 (1.531)	1.423 (0.209)
	CWM-SIC	0.959 (0.064)	0.999 (0.001)	0.125 (0.112)	**0.060** (0.908)	0.176 (0.039)
	PSIM-CV	0.965 (0.059)	0.997 (0.002)	0.181 (0.165)	0.950 (1.359)	7.990 (1.186)
	PSIM-SIC	0.959 (0.064)	0.999 (0.001)	0.125 (0.112)	**0.060** (0.908)	0.907 (0.174)
	MCP-CV	0.830 (0.137)	**1.000** (**0.000**)	1.823 (1.598)	-1.700 (1.374)	0.530 (0.134)
	MCP-SIC	0.830 (0.137)	**1.000** (**0.000**)	1.823 (1.598)	-1.700 (1.374)	0.074 (**0.018**)
	MCP-EBIC	0.830 (0.137)	**1.000** (**0.000**)	1.823 (1.598)	-1.700 (1.374)	0.078 (0.022)
	MCP-BIC	0.830 (0.137)	**1.000** (**0.000**)	1.823 (1.598)	-1.700 (1.374)	0.074 (**0.018**)
	SCAD-CV	0.831 (0.138)	**1.000** (**0.000**)	3.108 (2.488)	-1.690 (1.383)	0.511 (0.120)
	SCAD-SIC	0.831 (0.138)	**1.000** (**0.000**)	3.108 (2.488)	-1.690 (1.383)	0.076 (0.019)
	SCAD-EBIC	0.831 (0.138)	**1.000** (**0.000**)	3.108 (2.488)	-1.690 (1.383)	0.077 (0.021)
	SCAD-BIC	0.831 (0.138)	**1.000** (**0.000**)	3.108 (2.488)	-1.690 (1.383)	0.075 (0.019)
	LASSO-CV	0.968 (0.058)	0.973 (0.031)	0.534 (0.183)	12.710 (15.330)	0.446 (0.076)
	LASSO-SIC	0.966 (0.057)	0.997 (0.003)	0.736 (0.476)	1.360 (1.685)	0.048 (0.023)
	LASSO-EBIC	0.966 (0.057)	0.996 (0.004)	0.712 (0.458)	1.660 (1.871)	**0.047** (0.020)
	LASSO-BIC	**0.969** (**0.056**)	0.991 (0.007)	0.601 (0.408)	4.150 (3.436)	0.048 (0.021)
$p = 1500$	ABESS	0.972 (0.049)	**1.000** (0.001)	0.154 (0.153)	**0.260** (0.949)	0.342 (0.051)
	SDAR-EBIC	0.972 (0.049)	**1.000** (0.001)	**0.145** (**0.151**)	0.170 (**0.900**)	1.913 (0.337)
	SDAR-SIC	0.972 (0.049)	**1.000** (0.001)	0.154 (0.153)	**0.260** (0.949)	1.894 (0.332)
	CWM-CV	0.974 (**0.046**)	0.999 (0.001)	0.201 (0.157)	0.920 (1.116)	4.637 (0.611)
	CWM-SIC	0.972 (0.049)	**1.000** (0.001)	0.155 (0.153)	0.290 (1.018)	0.536 (0.074)
	PSIM-CV	0.974 (**0.046**)	0.999 (0.001)	0.202 (0.159)	0.920 (1.116)	33.876 (4.131)
	PSIM-SIC	0.972 (0.049)	**1.000** (0.001)	0.155 (0.153)	0.290 (1.018)	3.750 (0.636)
	MCP-CV	0.857 (0.128)	**1.000** (**0.000**)	1.930 (1.539)	-1.430 (1.281)	1.440 (0.223)
	MCP-SIC	0.857 (0.128)	**1.000** (**0.000**)	1.930 (1.539)	-1.430 (1.281)	0.224 (0.040)
	MCP-EBIC	0.857 (0.128)	**1.000** (**0.000**)	1.930 (1.539)	-1.430 (1.281)	0.219 (0.037)
	MCP-BIC	0.857 (0.128)	**1.000** (**0.000**)	1.930 (1.539)	-1.430 (1.281)	0.216 (0.040)
	SCAD-CV	0.858 (0.129)	**1.000** (**0.000**)	3.281 (2.197)	-1.420 (1.288)	1.456 (0.276)
	SCAD-SIC	0.856 (0.130)	**1.000** (**0.000**)	3.289 (2.206)	-1.440 (1.297)	0.220 (0.038)
	SCAD-EBIC	0.858 (0.129)	**1.000** (**0.000**)	3.281 (2.197)	-1.420 (1.288)	0.221 (0.038)
	SCAD-BIC	0.858 (0.129)	**1.000** (**0.000**)	3.281 (2.197)	-1.420 (1.288)	0.221 (0.035)
	LASSO-CV	**0.976** (0.047)	0.987 (0.014)	0.737 (0.284)	19.430 (21.005)	1.307 (0.179)
	LASSO-SIC	0.969 (0.056)	0.999 (0.001)	1.099 (0.605)	1.100 (1.534)	0.117 (0.029)
	LASSO-EBIC	0.971 (0.050)	0.999 (0.001)	1.068 (0.600)	1.460 (1.772)	0.117 (0.022)
	LASSO-BIC	0.975 (0.048)	0.998 (0.002)	0.947 (0.501)	3.250 (3.255)	**0.114** (**0.020**)
$p = 2500$	ABESS	**0.957** (**0.067**)	**1.000** (**0.000**)	**0.139** (**0.177**)	**0.000** (**0.953**)	0.523 (0.077)
	SDAR-EBIC	**0.957** (**0.067**)	**1.000** (**0.000**)	0.152 (0.188)	0.090 (0.996)	2.869 (0.538)
	SDAR-SIC	**0.957** (**0.067**)	**1.000** (**0.000**)	**0.139** (**0.177**)	**0.000** (**0.953**)	2.837 (0.533)
	CWM-CV	**0.957** (**0.067**)	**1.000** (**0.000**)	0.207 (0.252)	0.720 (1.138)	7.191 (0.833)
	CWM-SIC	**0.957** (**0.067**)	**1.000** (**0.000**)	**0.139** (**0.177**)	**0.000** (**0.953**)	0.816 (0.102)
	PSIM-CV	**0.957** (**0.067**)	**1.000** (**0.000**)	0.206 (0.252)	0.710 (1.140)	48.601 (5.930)
	PSIM-SIC	**0.957** (**0.067**)	**1.000** (**0.000**)	**0.139** (**0.177**)	**0.000** (**0.953**)	5.248 (0.963)
	MCP-CV	0.836 (0.131)	**1.000** (**0.000**)	2.029 (1.566)	-1.640 (1.314)	2.426 (0.440)
	MCP-SIC	0.835 (0.133)	**1.000** (**0.000**)	2.030 (1.567)	-1.650 (1.329)	0.344 (0.062)
	MCP-EBIC	0.835 (0.133)	**1.000** (**0.000**)	2.030 (1.567)	-1.650 (1.329)	0.342 (0.059)
	MCP-BIC	0.836 (0.131)	**1.000** (**0.000**)	2.029 (1.566)	-1.640 (1.314)	0.342 (0.059)
	SCAD-CV	0.836 (0.131)	**1.000** (**0.000**)	3.563 (2.717)	-1.640 (1.314)	2.363 (0.386)
	SCAD-SIC	0.834 (0.132)	**1.000** (**0.000**)	3.567 (2.716)	-1.660 (1.320)	0.336 (0.061)
	SCAD-EBIC	0.836 (0.131)	**1.000** (**0.000**)	3.563 (2.717)	-1.640 (1.314)	0.333 (0.058)
	SCAD-BIC	0.836 (0.131)	**1.000** (**0.000**)	3.563 (2.717)	-1.640 (1.314)	0.335 (0.055)
	LASSO-CV	0.956 (0.069)	0.991 (0.010)	0.811 (0.487)	23.020 (26.117)	2.019 (0.281)
	LASSO-SIC	0.951 (0.069)	**1.000** (0.001)	1.212 (0.887)	0.710 (1.526)	**0.174** (**0.027**)
	LASSO-EBIC	0.953 (0.069)	0.999 (0.001)	1.188 (0.877)	0.910 (1.688)	0.177 (0.029)
	LASSO-BIC	0.954 (0.069)	0.999 (0.002)	1.034 (0.709)	3.130 (4.007)	0.177 (0.032)

Table S3. TPR, TNR, relative errors and sparse level errors of nineteen variable selection methods for constant correlation structure.

	Method	TPR	TNR	ReErr	SLE	Runtime
$p = 500$	ABESS	0.907 (0.091)	0.999 (0.001)	0.669 (0.615)	-0.430 (1.121)	0.098 (0.026)
	SDAR-EBIC	0.900 (0.097)	0.999 (0.001)	**0.623** (**0.599**)	-0.670 (**1.092**)	0.511 (0.172)
	SDAR-SIC	0.906 (0.092)	0.999 (0.001)	0.674 (0.612)	-0.430 (1.121)	0.507 (0.162)
	CWM-CV	0.912 (0.089)	0.991 (0.009)	1.602 (1.798)	3.500 (4.629)	3.371 (0.765)
	CWM-SIC	0.898 (0.097)	0.999 (0.002)	0.686 (0.657)	**-0.420** (1.273)	0.319 (0.086)
	PSIM-CV	0.914 (0.090)	0.993 (0.008)	1.459 (1.653)	2.470 (4.279)	24.206 (5.539)
	PSIM-SIC	0.900 (0.097)	0.999 (0.002)	0.691 (0.638)	-0.430 (1.257)	2.406 (0.614)
	MCP-CV	0.542 (0.227)	**1.000** (**0.000**)	52.418 (61.560)	-4.580 (2.270)	0.316 (0.055)
	MCP-SIC	0.540 (0.227)	**1.000** (**0.000**)	52.419 (61.560)	-4.600 (2.270)	0.042 (**0.014**)
	MCP-EBIC	0.540 (0.227)	**1.000** (**0.000**)	52.419 (61.560)	-4.600 (2.270)	**0.041** (0.016)
	MCP-BIC	0.540 (0.227)	**1.000** (**0.000**)	52.419 (61.560)	-4.600 (2.270)	0.042 (0.015)
	SCAD-CV	0.552 (0.227)	**1.000** (**0.000**)	72.049 (85.035)	-4.480 (2.267)	0.331 (0.066)
	SCAD-SIC	0.542 (0.220)	**1.000** (**0.000**)	72.101 (85.009)	-4.580 (2.198)	**0.041** (0.015)
	SCAD-EBIC	0.544 (0.223)	**1.000** (**0.000**)	72.091 (85.016)	-4.560 (2.231)	0.046 (0.017)
	SCAD-BIC	0.550 (0.227)	**1.000** (**0.000**)	72.053 (85.033)	-4.500 (2.267)	0.043 (**0.014**)
	LASSO-CV	0.847 (0.119)	0.987 (0.023)	11.228 (10.250)	4.910 (11.914)	0.440 (0.123)
	LASSO-SIC	0.901 (0.090)	0.991 (0.008)	4.484 (3.158)	3.270 (4.080)	0.098 (0.066)
	LASSO-EBIC	0.905 (0.088)	0.989 (0.009)	4.089 (2.978)	4.350 (4.635)	0.102 (0.073)
	LASSO-BIC	**0.915** (**0.083**)	0.985 (0.009)	3.437 (2.350)	6.550 (4.751)	0.098 (0.070)
$p = 1500$	ABESS	0.930 (0.088)	**1.000** (0.001)	0.768 (**0.851**)	-0.240 (1.084)	0.266 (0.065)
	SDAR-EBIC	0.929 (0.089)	**1.000** (0.001)	**0.764** (0.855)	-0.270 (**1.081**)	1.239 (0.266)
	SDAR-SIC	0.930 (0.088)	**1.000** (0.001)	0.768 (**0.851**)	-0.240 (1.084)	1.229 (0.261)
	CWM-CV	0.933 (0.082)	0.997 (0.004)	1.541 (1.503)	3.420 (5.919)	3.525 (0.727)
	CWM-SIC	0.920 (0.086)	**1.000** (0.001)	0.850 (0.878)	-0.120 (1.305)	0.351 (0.067)
	PSIM-CV	**0.935** (**0.081**)	0.997 (0.004)	1.885 (2.363)	3.290 (6.243)	54.868 (12.254)
	PSIM-SIC	0.920 (0.086)	**1.000** (0.001)	0.881 (0.902)	**-0.080** (1.331)	6.325 (1.944)
	MCP-CV	0.540 (0.204)	**1.000** (**0.000**)	53.214 (63.374)	-4.600 (2.045)	1.103 (0.232)
	MCP-SIC	0.540 (0.204)	**1.000** (**0.000**)	53.214 (63.374)	-4.600 (2.045)	0.161 (0.043)
	MCP-EBIC	0.540 (0.204)	**1.000** (**0.000**)	53.214 (63.374)	-4.600 (2.045)	0.162 (0.039)
	MCP-BIC	0.540 (0.204)	**1.000** (**0.000**)	53.214 (63.374)	-4.600 (2.045)	0.163 (0.042)
	SCAD-CV	0.540 (0.208)	**1.000** (**0.000**)	81.637 (87.852)	-4.600 (2.084)	1.107 (0.208)
	SCAD-SIC	0.533 (0.207)	**1.000** (**0.000**)	81.785 (87.959)	-4.670 (2.070)	0.160 (0.041)
	SCAD-EBIC	0.533 (0.207)	**1.000** (**0.000**)	81.785 (87.959)	-4.670 (2.070)	0.165 (0.036)
	SCAD-BIC	0.540 (0.208)	**1.000** (**0.000**)	81.637 (87.853)	-4.600 (2.084)	0.163 (0.041)
	LASSO-CV	0.861 (0.136)	0.994 (0.011)	13.456 (9.583)	7.230 (16.650)	1.265 (0.306)
	LASSO-SIC	0.852 (0.138)	0.999 (0.002)	14.339 (8.904)	0.140 (3.260)	**0.112** (0.033)
	LASSO-EBIC	0.855 (0.137)	0.999 (0.002)	14.135 (9.017)	0.530 (3.878)	0.113 (0.034)
	LASSO-BIC	0.859 (0.136)	0.998 (0.003)	13.747 (9.322)	1.620 (5.065)	**0.112** (**0.032**)
$p = 2500$	ABESS	**0.909** (0.087)	**1.000** (**0.000**)	**0.885** (0.871)	-0.190 (**1.253**)	0.460 (0.093)
	SDAR-EBIC	**0.909** (0.087)	**1.000** (**0.000**)	0.958 (1.006)	**-0.080** (1.316)	1.917 (0.334)
	SDAR-SIC	**0.909** (0.087)	**1.000** (**0.000**)	0.887 (**0.870**)	-0.180 (1.258)	1.908 (0.335)
	CWM-CV	0.907 (0.088)	0.998 (0.003)	1.502 (1.325)	3.580 (7.087)	4.492 (0.785)
	CWM-SIC	0.897 (0.094)	**1.000** (**0.000**)	0.930 (0.937)	-0.240 (1.379)	0.461 (0.088)
	PSIM-CV	0.908 (**0.085**)	0.999 (0.002)	1.402 (1.430)	1.620 (4.456)	92.834 (21.810)
	PSIM-SIC	0.904 (0.091)	**1.000** (**0.000**)	0.943 (0.950)	-0.090 (1.386)	11.009 (2.951)
	MCP-CV	0.512 (0.199)	**1.000** (**0.000**)	55.665 (58.164)	-4.880 (1.991)	2.172 (0.429)
	MCP-SIC	0.512 (0.199)	**1.000** (**0.000**)	55.665 (58.164)	-4.880 (1.991)	0.291 (0.061)
	MCP-EBIC	0.512 (0.199)	**1.000** (**0.000**)	55.665 (58.164)	-4.880 (1.991)	0.289 (0.074)
	MCP-BIC	0.512 (0.199)	**1.000** (**0.000**)	55.665 (58.164)	-4.880 (1.991)	0.282 (0.064)
	SCAD-CV	0.512 (0.195)	**1.000** (**0.000**)	84.011 (87.386)	-4.880 (1.945)	2.203 (0.427)
	SCAD-SIC	0.501 (0.196)	**1.000** (**0.000**)	84.259 (87.703)	-4.990 (1.962)	0.292 (0.075)
	SCAD-EBIC	0.504 (0.194)	**1.000** (**0.000**)	84.074 (87.383)	-4.960 (1.943)	0.290 (0.063)
	SCAD-BIC	0.507 (0.195)	**1.000** (**0.000**)	84.037 (87.386)	-4.930 (1.950)	0.282 (0.062)
	LASSO-CV	0.851 (0.134)	0.996 (0.007)	12.895 (8.866)	7.900 (17.620)	2.310 (0.584)
	LASSO-SIC	0.843 (0.131)	0.999 (0.002)	14.099 (8.597)	0.900 (4.800)	**0.202** (**0.057**)
	LASSO-EBIC	0.845 (0.132)	0.999 (0.002)	13.649 (8.647)	1.600 (6.145)	0.206 (0.062)
	LASSO-BIC	0.848 (0.134)	0.998 (0.003)	13.448 (8.778)	2.300 (6.927)	0.206 (0.062)

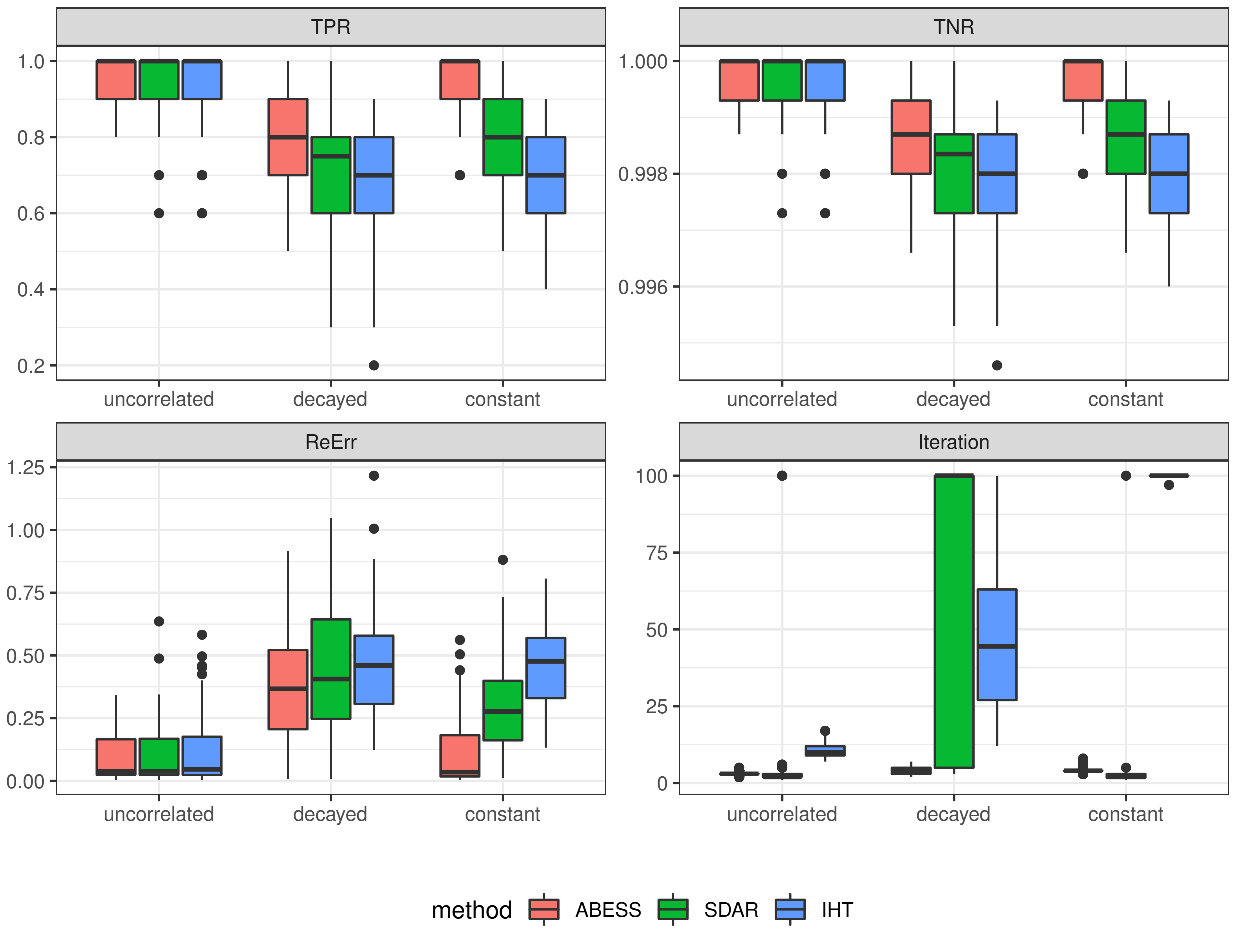

Fig. S1. TNR, TPR, relative error, and the number of iterations in the fixed support size best subset selection problem.

can efficiently speed up our algorithm. We conduct numerical simulation studies to explore the utility of the GS strategy. The simulation settings are completely adopted from Section Simulation. The numerical results are displayed in Figure S2. As we can see in Figure S2, by combining golden section search and SIC (or CV), we can accelerate the best sparsity level selection procedure while maintaining the sparsity level selection performance.

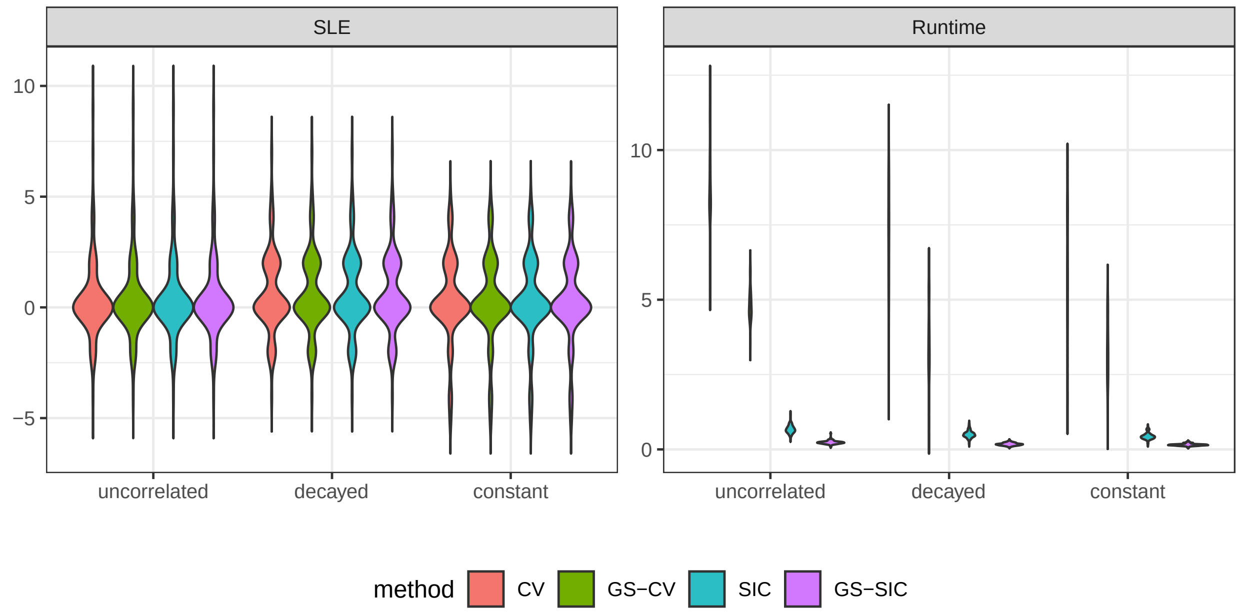

Fig. S2. Sparsity level error and runtime comparison in the sparsity level selection problem. The methods with prefix "GS" is golden section based the sparsity level selection strategy.

Proofs of Main results

Proof of Lemma 1.

proof 1 *Let*

$$\mathcal{A}_1 = \hat{\mathcal{A}} \cap \mathcal{A}^*, \mathcal{A}_2 = \hat{\mathcal{A}} \cap \mathcal{I}^*$$
$$\mathcal{I}_1 = \hat{\mathcal{I}} \cap \mathcal{A}^*, \mathcal{I}_2 = \hat{\mathcal{I}} \cap \mathcal{I}^*$$

We assume $\mathcal{I}_1 \neq \varnothing$ and show that it will lead to a contradiction.

Let $k = |\mathcal{I}_1|$. Denote the splicing set in the active and inactive sets, respectively, as

$$\hat{\mathcal{A}}_k = \{j \in \hat{\mathcal{A}} : \sum_{i \in \hat{\mathcal{A}}} \mathrm{I}(|\hat{\boldsymbol{\beta}}_j| \geq |\hat{\boldsymbol{\beta}}_i|) \leq k\},$$
$$\hat{\mathcal{I}}_k = \{j \in \hat{\mathcal{I}} : \sum_{i \in \hat{\mathcal{I}}} \mathrm{I}(|\hat{\mathrm{d}}_j| \leq |\hat{\mathrm{d}}_i|) \leq k\}.$$

And denote

$$\mathcal{A}_{11} = \mathcal{A}_1 \cap (\hat{\mathcal{A}}_k)^c, \mathcal{A}_{12} = \mathcal{A}_1 \cap \hat{\mathcal{A}}_k,$$
$$\mathcal{A}_{21} = \mathcal{A}_2 \cap (\hat{\mathcal{A}}_k)^c, \mathcal{A}_{22} = \mathcal{A}_2 \cap \hat{\mathcal{A}}_k,$$

and

$$\mathcal{I}_{11} = \mathcal{I}_1 \cap \hat{\mathcal{I}}_k,\ \mathcal{I}_{12} = \mathcal{I}_1 \cap (\hat{\mathcal{I}}_k)^c,$$
$$\mathcal{I}_{21} = \mathcal{I}_2 \cap \hat{\mathcal{I}}_k,\ \mathcal{I}_{22} = \mathcal{I}_2 \cap (\hat{\mathcal{I}}_k)^c.$$

Consider the following four cases:

$$(1)\ \mathcal{I}_{12} \neq \varnothing, \mathcal{A}_{12} \neq \varnothing,\ \ (2)\ \mathcal{I}_{12} \neq \varnothing, \mathcal{A}_{12} = \varnothing,$$
$$(3)\ \mathcal{I}_{12} = \varnothing, \mathcal{A}_{12} \neq \varnothing,\ \ (4)\ \mathcal{I}_{12} = \varnothing, \mathcal{A}_{12} = \varnothing.$$

We provide the details for the first case as the other cases follow similarly.

Let $\boldsymbol{H}_A = \boldsymbol{X}_A\left(\boldsymbol{X}_A'\boldsymbol{X}_A\right)^{-1}\boldsymbol{X}_A'$, *for any index set* $A \subseteq \{1,\ldots,p\}$. *The estimator* $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{d}})$ *can be expressed as*

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{\mathcal{A}_1} &= \left(\boldsymbol{X}_{\mathcal{A}_1}'(\boldsymbol{I}-\boldsymbol{H}_{\mathcal{A}_2})\boldsymbol{X}_{\mathcal{A}_1}\right)^{-1}\boldsymbol{X}_{\mathcal{A}_1}'(\boldsymbol{I}-\boldsymbol{H}_{\mathcal{A}_2})\boldsymbol{y},\\
\hat{\boldsymbol{\beta}}_{\mathcal{A}_2} &= \left(\boldsymbol{X}_{\mathcal{A}_2}'(\boldsymbol{I}-\boldsymbol{H}_{\mathcal{A}_1})\boldsymbol{X}_{\mathcal{A}_2}\right)^{-1}\boldsymbol{X}_{\mathcal{A}_2}'(\boldsymbol{I}-\boldsymbol{H}_{\mathcal{A}_1})\boldsymbol{y},\\
\hat{\boldsymbol{d}}_{\mathcal{I}_1} &= \boldsymbol{X}_{\mathcal{I}_1}'(\boldsymbol{I}-\boldsymbol{H}_{\hat{\mathcal{A}}})\boldsymbol{y}/n,\\
\hat{\boldsymbol{d}}_{\mathcal{I}_2} &= \boldsymbol{X}_{\mathcal{I}_2}'(\boldsymbol{I}-\boldsymbol{H}_{\hat{\mathcal{A}}})\boldsymbol{y}/n.
\end{aligned}$$

First of all, Assume that events $\{c_-(s)\|\boldsymbol{\beta}^*_{\mathcal{I}_{12}}\|_2 \le 2(1+\Delta)\Big(\theta_{s,s}+\frac{\theta_{s,s}^2}{c_-(s)}\Big)\|\boldsymbol{\beta}^*_{\mathcal{I}_1}\|_2\}$ *and* $\{\|\boldsymbol{\beta}^*_{\mathcal{A}_{12}}\|_2 \le 2(1+\Delta)\frac{\theta_{s,s}}{c_-(s)}\|\boldsymbol{\beta}^*_{\mathcal{I}_1}\|_2\}$ *hold. We will show that these two events hold with probability* $1-\frac{1}{3}\gamma_1-\frac{1}{3}\gamma_2$ *later.*

Now, we splicing $\hat{\mathcal{A}}_k = \mathcal{A}_{12}\cup\mathcal{A}_{22}$ *and* $\hat{\mathcal{I}}_k = \mathcal{I}_{11}\cup\mathcal{I}_{21}$, *and then the new active set is* $\tilde{\mathcal{A}} = (\hat{\mathcal{A}}\backslash\hat{\mathcal{A}}_k)\cup\hat{\mathcal{I}}_k$ *and the inactive set is* $\tilde{\mathcal{I}} = (\tilde{\mathcal{A}})^c$. *Let* $\tilde{\boldsymbol{\beta}} = \arg\min_{\beta_{\tilde{\mathcal{I}}}=0}\frac{1}{2n}\|\boldsymbol{y}-\boldsymbol{X}\beta\|_2^2$. *The loss function of* $\tilde{\boldsymbol{\beta}}$ *is*

$$\begin{aligned}
2n\mathcal{L}(\tilde{\beta}) =& \|\boldsymbol{y}-\boldsymbol{X}\tilde{\beta}\|_2^2 = \boldsymbol{y}'(\boldsymbol{I}-\boldsymbol{H}_{\tilde{\mathcal{A}}})y\\
=&(\boldsymbol{X}_{\mathcal{A}_{12}\cup\mathcal{I}_{12}}\boldsymbol{\beta}^*_{\mathcal{A}_{12}\cup\mathcal{I}_{12}}+\boldsymbol{\epsilon})'(\boldsymbol{I}-\boldsymbol{H}_{\tilde{\mathcal{A}}})(X_{\mathcal{A}_{12}\cup\mathcal{I}_{12}}\boldsymbol{\beta}^*_{\mathcal{A}_{12}\cup\mathcal{I}_{12}}+\boldsymbol{\epsilon})\\
\le& nc_+(s)(\|\boldsymbol{\beta}^*_{\mathcal{A}_{12}}\|_2^2+\|\boldsymbol{\beta}^*_{\mathcal{I}_{12}}\|_2^2)+\boldsymbol{\epsilon}'\boldsymbol{\epsilon}+\\
&\qquad 2|\boldsymbol{\epsilon}'(\boldsymbol{I}-\boldsymbol{H}_{\tilde{\mathcal{A}}})\boldsymbol{X}_{\mathcal{A}_{12}\cup\mathcal{I}_{12}}\boldsymbol{\beta}^*_{\mathcal{A}_{12}\cup\mathcal{I}_{12}}|+|\boldsymbol{\epsilon}'\boldsymbol{H}_{\tilde{\mathcal{A}}}\boldsymbol{\epsilon}|\\
\le& nc_+(s)\Big[\Big(2(1+\Delta)\frac{\theta_{s,s}}{c_-(s)}\Big)^2+\Big(2(1+\Delta)\frac{\theta_{s,s}}{c_-(s)}\big(1+\frac{\theta_{s,s}}{c_-(s)}\big)\Big)^2\Big]\|\boldsymbol{\beta}^*_{\mathcal{I}_1}\|_2^2+\\
&\qquad \boldsymbol{\epsilon}'\boldsymbol{\epsilon}+2|\boldsymbol{\epsilon}'(\boldsymbol{I}-\boldsymbol{H}_{\tilde{\mathcal{A}}})\boldsymbol{X}_{\mathcal{A}_{12}\cup\mathcal{I}_{12}}\boldsymbol{\beta}^*_{\mathcal{A}_{12}\cup\mathcal{I}_{12}}|+|\boldsymbol{\epsilon}'\boldsymbol{H}_{\tilde{\mathcal{A}}}\boldsymbol{\epsilon}|\\
\le& 8nc_+(s)\Big((1+\Delta)\frac{\theta_{s,s}}{c_-(s)}\big(1+\frac{\theta_{s,s}}{c_-(s)}\big)\Big)^2\|\boldsymbol{\beta}^*_{\mathcal{I}_1}\|_2^2+\boldsymbol{\epsilon}'\boldsymbol{\epsilon}+f_1(\boldsymbol{\epsilon}),
\end{aligned} \tag{1}$$

where $f_1(\boldsymbol{\epsilon}) = 2|\boldsymbol{\epsilon}'(\boldsymbol{I}-\boldsymbol{H}_{\tilde{\mathcal{A}}})\boldsymbol{X}_{\mathcal{A}_{12}\cup\mathcal{I}_{12}}\boldsymbol{\beta}^*_{\mathcal{A}_{12}\cup\mathcal{I}_{12}}|+|\boldsymbol{\epsilon}'\boldsymbol{H}_{\tilde{\mathcal{A}}}\boldsymbol{\epsilon}|$. *In addition, the loss function of* $\hat{\boldsymbol{\beta}}$ *is*

$$\begin{aligned}
2n\mathcal{L}(\hat{\beta}) =& \|\boldsymbol{y}-\boldsymbol{X}\hat{\beta}\|_2^2 = \boldsymbol{y}'(\boldsymbol{I}-\boldsymbol{H}_{\hat{\mathcal{A}}})\boldsymbol{y}\\
=&(\boldsymbol{X}_{\mathcal{I}_1}\boldsymbol{\beta}^*_{\mathcal{I}_1}+\boldsymbol{\epsilon})'(\boldsymbol{I}-\boldsymbol{H}_{\hat{\mathcal{A}}})(\boldsymbol{X}_{\mathcal{I}_1}\boldsymbol{\beta}^*_{\mathcal{I}_1}+\boldsymbol{\epsilon})\\
\ge& n(c_-(s)-\frac{\theta_{s,s}^2}{c_-(s)})\|\boldsymbol{\beta}^*_{\mathcal{I}_1}\|_2^2+\boldsymbol{\epsilon}'\boldsymbol{\epsilon}-2|\boldsymbol{\epsilon}'(\boldsymbol{I}-\boldsymbol{H}_{\hat{\mathcal{A}}})\boldsymbol{X}_{\mathcal{I}_1}\boldsymbol{\beta}^*_{\mathcal{I}_1}|-|\boldsymbol{\epsilon}'\boldsymbol{H}_{\hat{\mathcal{A}}}\boldsymbol{\epsilon}|\\
=& n(c_-(s)-\frac{\theta_{s,s}^2}{c_-(s)})\|\boldsymbol{\beta}^*_{\mathcal{I}_1}\|_2^2+\boldsymbol{\epsilon}'\boldsymbol{\epsilon}-f_2(\boldsymbol{\epsilon}),
\end{aligned} \tag{2}$$

where $f_2(\boldsymbol{\epsilon}) = 2|\boldsymbol{\epsilon}'(\boldsymbol{I}-\boldsymbol{H}_{\hat{\mathcal{A}}})\boldsymbol{X}_{\mathcal{I}_1}\boldsymbol{\beta}^*_{\mathcal{I}_1}|+|\boldsymbol{\epsilon}'\boldsymbol{H}_{\hat{\mathcal{A}}}\boldsymbol{\epsilon}|$. *The conditions of this lemma assure that*

$$(1-\Delta)n\big(c_-(s)-\frac{\theta_{s,s}^2}{c_-(s)}\big)\|\boldsymbol{\beta}^*_{\mathcal{I}_1}\|_2^2 > 8nc_+(s)\Big((1+\Delta)\frac{\theta_{s,s}}{c_-(s)}\big(1+\frac{\theta_{s,s}}{c_-(s)}\big)\Big)^2\|\boldsymbol{\beta}^*_{\mathcal{I}_1}\|_2^2.$$

Let $\boldsymbol{u}_j' = e_j'(\boldsymbol{X}_{\mathcal{A}}'\boldsymbol{X}_{\mathcal{A}})^{-\frac{1}{2}}\boldsymbol{X}_{\mathcal{A}}'$, *where the* j*th element of* $\boldsymbol{e}_j$ *is 1 and otherwise is 0. Note that* $\|u_j\|_2^2 \le 1$, *we have*

$$\begin{aligned}
\mathrm{P}\left(\|(\boldsymbol{X}_{\mathcal{A}}'\boldsymbol{X}_{\mathcal{A}})^{-\frac{1}{2}}\boldsymbol{X}_{\mathcal{A}}'\boldsymbol{\epsilon}\|_2 \ge t\right) \le& \sum_{j\in\mathcal{A}}\mathrm{P}\left(\|\boldsymbol{e}_j'(\boldsymbol{X}_{\mathcal{A}}'\boldsymbol{X}_{\mathcal{A}})^{-\frac{1}{2}}\boldsymbol{X}_{\mathcal{A}}'\boldsymbol{\epsilon}\|_2 \ge \frac{t}{\sqrt{|\mathcal{A}|}}\right)\\
\le& \sum_{j\in\mathcal{A}}\mathrm{P}\left(\|\boldsymbol{u}_j'\epsilon\|_2 \ge \frac{t}{\sqrt{|\mathcal{A}|}}\right)\\
\le& 2p\exp\{-\frac{t^2}{\sigma^2|\mathcal{A}|}\}.
\end{aligned}$$

Thus

$$\begin{aligned}
\mathrm{P}\big(\|\boldsymbol{H}_{\tilde{\mathcal{A}}}\boldsymbol{\epsilon}\|_2 \ge \sqrt{\frac{\Delta}{4}n(c_-(s)-\frac{\theta_{s,s}^2}{c_-(s)})}\|\boldsymbol{\beta}^*_{\mathcal{I}_1}\|_2\big) \le \frac{\gamma_3}{6},\\
\mathrm{P}\big(\|\boldsymbol{H}_{\hat{\mathcal{A}}}\boldsymbol{\epsilon}\|_2 \ge \sqrt{\frac{\Delta}{4}n(c_-(s)-\frac{\theta_{s,s}^2}{c_-(s)})}\|\boldsymbol{\beta}^*_{\mathcal{I}_1}\|_2\big) \le \frac{\gamma_3}{6}
\end{aligned}$$

where $\gamma_3 = 12\exp\{\log p - \frac{K_{s,3}nb^*}{s}\}$, $K_{s,3} = \frac{\Delta(c_-(s)^2-\theta_{s,s}^2)}{4c_-(s)\sigma^2}$. *Similarly, we can show that,*

$$\mathrm{P}\Big(2|\boldsymbol{\epsilon}'(\boldsymbol{I}-\boldsymbol{H}_{\hat{\mathcal{A}}})\boldsymbol{X}_{\mathcal{A}_{12}\cup\mathcal{I}_{12}}\boldsymbol{\beta}^*_{\mathcal{A}_{12}\cup\mathcal{I}_{12}}| \geq \frac{\Delta}{4}n(c_-(s)-\frac{\theta_{s,s}^2}{c_-(s)})\|\boldsymbol{\beta}^*_{\mathcal{I}_1}\|_2^2\Big) \leq \frac{\gamma_4}{2},$$
$$\mathrm{P}\Big(2|\boldsymbol{\epsilon}'(\boldsymbol{I}-\boldsymbol{H}_{\hat{\mathcal{A}}})\boldsymbol{X}_{\mathcal{I}_1}\boldsymbol{\beta}^*_{\mathcal{I}_1}| \geq \frac{\Delta}{4}n(c_-(s)-\frac{\theta_{s,s}^2}{c_-(s)})\|\boldsymbol{\beta}^*_{\mathcal{I}_1}\|_2^2\Big) \leq \frac{\gamma_4}{2},$$

where $\gamma_4 = 4\exp\{\log p - \frac{K_{s,4}nb^*}{s^*}\}$, $K_{s,4} = \min\{\frac{\big(\Delta(c_-(s)-\frac{\theta_{s,s}^2}{c_-(s)})/8\big)^2}{c_+(s)\sigma^2}/\Big(4(1+\Delta)\frac{\theta_{s,s}}{c_-(s)}\big(1+\frac{\theta_{s,s}}{c_-(s)}\big)\Big)^2, \frac{\big(\Delta(c_-(s)-\frac{\theta_{s,s}^2}{c_-(s)})/8\big)^2}{c_+(s)\sigma^2}\}$. *Thus,*

$$\mathrm{P}\left(f_1(\boldsymbol{\epsilon})+f_2(\boldsymbol{\epsilon}) \geq \Delta n(c_-(s)-\frac{\theta_{s,s}^2}{c_-(s)})\|\boldsymbol{\beta}^*_{\mathcal{I}_1}\|_2^2\right) \leq \frac{1}{3}\gamma_3+\gamma_4.$$

Now, it follows from Conditions (4) and (6) that

$$\begin{aligned}\mathcal{L}(\hat{\boldsymbol{\beta}})-\mathcal{L}(\tilde{\boldsymbol{\beta}}) \geq& \frac{(1-\delta_s)(1-\Delta)\big(c_-(s)-\frac{\theta_{s,s}^2}{c_-(s)}\big)}{2}\|\boldsymbol{\beta}^*_{\mathcal{I}_1}\|_2^2\\ \geq& \frac{(1-\delta_s)(1-\Delta)\big(c_-(s)-\frac{\theta_{s,s}^2}{c_-(s)}\big)}{2}b^*\\ >&\tau_s.\end{aligned}$$

Consequently,

$$\mathrm{P}\left(\mathcal{L}(\hat{\boldsymbol{\beta}})-\mathcal{L}(\tilde{\boldsymbol{\beta}})>\tau_s\right) \geq 1-\frac{1}{3}\gamma_1-\frac{1}{3}\gamma_2-\frac{1}{3}\gamma_3-\gamma_4,$$

which leads to a contradiction with $\mathcal{I}_1 \neq \varnothing$. *Therefore,*

$$\mathrm{P}(\hat{\mathcal{A}} \supseteq \mathcal{A}^*) \geq 1-\gamma(s,n,p,b^*),$$

where $\gamma(s,n,p,b^*) = 16\exp\{\log p - \frac{K_s nb^*}{s}\}$ *and* $K_s = \min\{K_{s,1},K_{s,2},K_{s,3},K_{s,4}\}$. *By Condition (6),*

$$\lim_{n\to\infty}\mathrm{P}(\hat{\mathcal{A}} \supseteq \mathcal{A}^*) = 1.$$

Especially, if $s=s^*$, *we can show that*

$$\mathrm{P}(\hat{\mathcal{A}} = \mathcal{A}^*) = 1.$$

It remains to show that

$$\mathrm{P}\left(c_-(s)\|\boldsymbol{\beta}^*_{\mathcal{I}_{12}}\|_2 \leq 2(1+\Delta)\Big(\theta_{s,s}+\frac{\theta_{s,s}^2}{c_-(s)}\Big)\|\boldsymbol{\beta}^*_{\mathcal{I}_1}\|_2\right) \geq 1-\frac{1}{3}\gamma_1. \tag{3}$$

and

$$\mathrm{P}\left(\|\boldsymbol{\beta}^*_{\mathcal{A}_{12}}\|_2 \leq 2(1+\Delta)\frac{\theta_{s,s}}{c_-(s)}\|\boldsymbol{\beta}^*_{\mathcal{I}_1}\|_2\right) \geq 1-\frac{1}{3}\gamma_2. \tag{4}$$

On one hand, it follows from $|\hat{\mathcal{I}}_k| = |\mathcal{I}_{11}|+|\mathcal{I}_{21}| = |\mathcal{I}_1|$, $|\mathcal{I}_{12}| = |\mathcal{I}_{21}|$ *and the definition of* $\hat{\mathcal{I}}_k$ *that*

$$\min_{j\in\mathcal{I}_{21}}|\hat{d}_j| \geq \max_{j\in\mathcal{I}_{12}}|\hat{d}_j|.$$

Note that,

$$\begin{aligned}n\|\hat{\boldsymbol{d}}_{\mathcal{I}_{12}}\|_2 &= \|\boldsymbol{X}'_{\mathcal{I}_{12}}(\boldsymbol{I}-\boldsymbol{H}_{\hat{\mathcal{A}}})\boldsymbol{y}\|_2\\ &= \|\boldsymbol{X}'_{\mathcal{I}_{12}}(\boldsymbol{I}-\boldsymbol{H}_{\hat{\mathcal{A}}})\left(\boldsymbol{X}_{\mathcal{I}_1}\boldsymbol{\beta}^*_{\mathcal{I}_1}+\boldsymbol{\epsilon}\right)\|_2\\ &\geq \|\boldsymbol{X}'_{\mathcal{I}_{12}}\boldsymbol{X}_{\mathcal{I}_{12}}\boldsymbol{\beta}^*_{\mathcal{I}_{12}}\|_2-\|\boldsymbol{X}'_{\mathcal{I}_{12}}\boldsymbol{X}_{\mathcal{I}_{11}}\boldsymbol{\beta}^*_{\mathcal{I}_{11}}\|_2-\\ &\qquad\qquad \|\boldsymbol{X}'_{\mathcal{I}_{12}}\boldsymbol{H}_{\hat{\mathcal{A}}}\boldsymbol{X}_{\mathcal{I}_1}\boldsymbol{\beta}^*_{\mathcal{I}_1}\|_2-\|\boldsymbol{X}'_{\mathcal{I}_{12}}(\boldsymbol{I}-\boldsymbol{H}_{\hat{\mathcal{A}}})\boldsymbol{\epsilon}\|_2\\ &\geq nc_-(s)\|\boldsymbol{\beta}^*_{\mathcal{I}_{12}}\|_2-n\theta_{s,s}\|\boldsymbol{\beta}^*_{\mathcal{I}_{11}}\|_2-n\frac{\theta_{s,s}^2}{c_-(s)}\|\boldsymbol{\beta}^*_{\mathcal{I}_1}\|_2-\|\boldsymbol{X}'_{\mathcal{I}_{12}}(\boldsymbol{I}-\boldsymbol{H}_{\hat{\mathcal{A}}})\boldsymbol{\epsilon}\|_2\end{aligned} \tag{5}$$

and

$$\begin{aligned}n\|\hat{\boldsymbol{d}}_{\mathcal{I}_{21}}\|_2 &= \|\boldsymbol{X}'_{\mathcal{I}_{21}}(\boldsymbol{I}-\boldsymbol{H}_{\hat{\mathcal{A}}})\boldsymbol{y}\|_2\\ &= \|\boldsymbol{X}'_{\mathcal{I}_{21}}(\boldsymbol{I}-\boldsymbol{H}_{\hat{\mathcal{A}}})\left(\boldsymbol{X}_{\mathcal{I}_1}\boldsymbol{\beta}^*_{\mathcal{I}_1}+\boldsymbol{\epsilon}\right)\|_2\\ &\leq \|\boldsymbol{X}'_{\mathcal{I}_{21}}(\boldsymbol{I}-\boldsymbol{H}_{\hat{\mathcal{A}}})\boldsymbol{X}_{\mathcal{I}_1}\boldsymbol{\beta}^*_{\mathcal{I}_1}\|_2+\|\boldsymbol{X}'_{\mathcal{I}_{21}}(\boldsymbol{I}-\boldsymbol{H}_{\hat{\mathcal{A}}})\boldsymbol{\epsilon}\|_2\\ &\leq n\Big(\theta_{s,s}+\frac{\theta_{s,s}^2}{c_-(s)}\Big)\|\boldsymbol{\beta}^*_{\mathcal{I}_1}\|_2+\|\boldsymbol{X}'_{\mathcal{I}_{21}}(\boldsymbol{I}-\boldsymbol{H}_{\hat{\mathcal{A}}})\boldsymbol{\epsilon}\|_2.\end{aligned} \tag{6}$$

Since $|\mathcal{I}_{12}| = |\mathcal{I}_{21}|$, $\|\hat{\boldsymbol{d}}_{\mathcal{I}_{21}}\|_2 \geq \|\hat{\boldsymbol{d}}_{\mathcal{I}_{12}}\|_2$, *combined with (5) and (6), we have*

$$\begin{aligned}&2n\Big(\theta_{s,s} + \frac{\theta_{s,s}^2}{c_-(s)}\Big)\|\boldsymbol{\beta}^*_{\mathcal{I}_1}\|_2 + \|\boldsymbol{X}'_{\mathcal{I}_{21}}(\boldsymbol{I} - \boldsymbol{H}_{\hat{\mathcal{A}}})\boldsymbol{\epsilon}\|_2 \\ \geq & nc_-(s)\|\boldsymbol{\beta}^*_{\mathcal{I}_{12}}\|_2 - \|\boldsymbol{X}'_{\mathcal{I}_{12}}(\boldsymbol{I} - \boldsymbol{H}_{\hat{\mathcal{A}}})\boldsymbol{\epsilon}\|_2.\end{aligned}$$

Let $\boldsymbol{h}'_j = \boldsymbol{X}'_j(\boldsymbol{I} - \boldsymbol{H}_{\hat{\mathcal{A}}})$,

$$\|\boldsymbol{h}_j\|_2^2 = |\boldsymbol{X}'_j(\boldsymbol{I} - \boldsymbol{H}_{\hat{\mathcal{A}}})\boldsymbol{X}_j| \leq nc_+(s).$$

By Condition (1) and $|\mathcal{I}_1| \geq |\mathcal{I}_{21}|$, *for some* $\Delta > 0$, *we have*

$$\begin{aligned}&\mathrm{P}\left(\|\boldsymbol{X}'_{\mathcal{I}_{21}}(\boldsymbol{I} - \boldsymbol{H}_{\hat{\mathcal{A}}})\boldsymbol{\epsilon}\|_2 > n\Delta\Big(\theta_{s,s} + \frac{\theta_{s,s}^2}{c_-(s)}\Big)\|\boldsymbol{\beta}^*_{\mathcal{I}_1}\|_2\right) \\ \leq & \sum_{j\in\mathcal{I}_{21}} \mathrm{P}\left(|\boldsymbol{h}'_j\epsilon| > \frac{n\Delta\Big(\theta_{s,s} + \frac{\theta_{s,s}^2}{c_-(s)}\Big)}{\sqrt{|\mathcal{I}_{21}|}}\|\boldsymbol{\beta}^*_{\mathcal{I}_1}\|_2\right) \\ \leq & 2p\exp\left\{-\frac{1}{nc_+(s)\sigma^2}\frac{n^2\Delta^2\Big(\theta_{s,s} + \frac{\theta_{s,s}^2}{c_-(s)}\Big)^2}{|\mathcal{I}_{21}|}\|\boldsymbol{\beta}^*_{\mathcal{I}_1}\|_2^2\right\} \\ \leq & 2p\exp\left\{-\frac{n^2\Delta^2\Big(\theta_{s,s} + \frac{\theta_{s,s}^2}{c_-(s)}\Big)^2}{nc_+(s)\sigma^2}(\min_{j\in\mathcal{A}^*}|\beta_j^*|)^2\right\} \\ = & \frac{1}{6}\gamma_1,\end{aligned}$$

where $\gamma_1 = 12\exp\{\log p - K_{s,1}nb^*\}$ *and* $K_{s,1} = \frac{\Delta^2\left(\theta_{s,s} + \frac{\theta_{s,s}^2}{c_-(s)}\right)^2}{c_+(s)\sigma^2}$. *Similarly,*

$$\begin{aligned}&\mathrm{P}\left(\|\boldsymbol{X}'_{\mathcal{I}_{12}}(\boldsymbol{I} - \boldsymbol{H}_{\hat{\mathcal{A}}})\boldsymbol{\epsilon}\|_2 > n\Delta\Big(\theta_{s,s} + \frac{\theta_{s,s}^2}{c_-(s)}\Big)\|\boldsymbol{\beta}^*_{\mathcal{I}_1}\|_2\right) \\ \leq & 2p\exp\left\{-\frac{1}{2nc_+(s)\sigma^2}\frac{n^2\Delta^2\Big(\theta_{s,s} + \frac{\theta_{s,s}^2}{c_-(s)}\Big)^2}{|\mathcal{I}_{12}|}\|\boldsymbol{\beta}^*_{\mathcal{I}_1}\|_2^2\right\} \\ \leq & \frac{1}{6}\gamma_1.\end{aligned}$$

Consequently,

$$\mathrm{P}\left(c_-(s)\|\boldsymbol{\beta}^*_{\mathcal{I}_{12}}\|_2 \leq 2(1+\Delta)\Big(\theta_{s,s} + \frac{\theta_{s,s}^2}{c_-(s)}\Big)\|\boldsymbol{\beta}^*_{\mathcal{I}_1}\|_2\right) \geq 1 - \frac{1}{3}\gamma_1.$$

On the other hand,

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\mathcal{A}_1} &= \boldsymbol{\beta}^*_{\mathcal{A}_1} + \left(\boldsymbol{X}'_{\mathcal{A}_1}(\boldsymbol{I} - \boldsymbol{H}_{\mathcal{A}_2})\boldsymbol{X}_{\mathcal{A}_1}\right)^{-1}\boldsymbol{X}'_{\mathcal{A}_1}(\boldsymbol{I} - \boldsymbol{H}_{\mathcal{A}_2})(\boldsymbol{X}_{\mathcal{I}_1}\boldsymbol{\beta}^*_{\mathcal{I}_1} + \boldsymbol{\epsilon}), \\ \hat{\boldsymbol{\beta}}_{\mathcal{A}_2} &= 0 + \left(\boldsymbol{X}'_{\mathcal{A}_2}(\boldsymbol{I} - \boldsymbol{H}_{\mathcal{A}_1})\boldsymbol{X}_{\mathcal{A}_2}\right)^{-1}\boldsymbol{X}'_{\mathcal{A}_2}(\boldsymbol{I} - \boldsymbol{H}_{\mathcal{A}_1})(\boldsymbol{X}_{\mathcal{I}_1}\boldsymbol{\beta}^*_{\mathcal{I}_1} + \boldsymbol{\epsilon}).\end{aligned}$$

Then $\hat{\mathcal{A}}_k = \mathcal{A}_{12} \cup \mathcal{A}_{22}$ *and* $|\mathcal{A}_{12}| + |\mathcal{A}_{22}| = |\mathcal{I}_1| \leq |\mathcal{A}_2| = |\mathcal{A}_{21}| + |\mathcal{A}_{22}|$, *and we have*

$$\max_{j\in\mathcal{A}_{12}\cup\mathcal{A}_{22}} |\hat{\boldsymbol{\beta}}_j| \leq \min_{j\in\mathcal{A}_{11}\cup\mathcal{A}_{21}} |\hat{\boldsymbol{\beta}}_j|.$$

Considering the case when $\mathcal{A}_{12} \neq \varnothing$, *we have*

$$\frac{1}{\sqrt{|\mathcal{A}_{12}|}}\|\hat{\boldsymbol{\beta}}_{\mathcal{A}_{12}}\|_2 \leq \frac{1}{\sqrt{|\mathcal{A}_{21}|}}\|\hat{\boldsymbol{\beta}}_{\mathcal{A}_{21}}\|_2.$$

Denote $\boldsymbol{E}_{\mathcal{A}_{12}}$ *as a* $|\mathcal{A}_{12}| \times |\mathcal{A}_1|$ *matrix and its* j*th row is a* $|\mathcal{A}_1|$*-dimensional vector* $\boldsymbol{e}_j$, *where the* j*th element of* $\boldsymbol{e}_j$ *is 1 and otherwise 0.* $\boldsymbol{E}_{\mathcal{A}_{21}}$ *can be defined analogously. Then*

$$\begin{aligned}\|\hat{\boldsymbol{\beta}}_{\mathcal{A}_{12}}\|_2 \geq & \|\boldsymbol{\beta}^*_{\mathcal{A}_{12}}\|_2 - \|\boldsymbol{E}_{\mathcal{A}_{12}}\left(\boldsymbol{X}'_{\mathcal{A}_1}(\boldsymbol{I} - \boldsymbol{H}_{\mathcal{A}_2})\boldsymbol{X}_{\mathcal{A}_1}\right)^{-1}\boldsymbol{X}'_{\mathcal{A}_1}(\boldsymbol{I} - \boldsymbol{H}_{\mathcal{A}_2})(\boldsymbol{X}_{\mathcal{I}_1}\boldsymbol{\beta}^*_{\mathcal{I}_1} + \boldsymbol{\epsilon})\|_2 \\ \geq & \|\boldsymbol{\beta}^*_{\mathcal{A}_{12}}\|_2 - \left[nc_-(s)\right]^{-1}n\theta_{s,s}\|\boldsymbol{\beta}^*_{\mathcal{I}_1}\|_2 - \\ & \|\boldsymbol{E}_{\mathcal{A}_{12}}\left(\boldsymbol{X}'_{\mathcal{A}_1}(\boldsymbol{I} - \boldsymbol{H}_{\mathcal{A}_2})\boldsymbol{X}_{\mathcal{A}_1}\right)^{-1}\boldsymbol{X}'_{\mathcal{A}_1}(\boldsymbol{I} - \boldsymbol{H}_{\mathcal{A}_2})\boldsymbol{\epsilon}\|_2,\end{aligned} \tag{7}$$

and

$$\begin{aligned}\|\hat{\boldsymbol\beta}_{\mathcal{A}_{21}}\|_2 &\le \|\boldsymbol{E}_{\mathcal{A}_{21}}\big(\boldsymbol{X}'_{\mathcal{A}_2}(\boldsymbol{I}-\boldsymbol{H}_{\mathcal{A}_1})\boldsymbol{X}_{\mathcal{A}_2}\big)^{-1}\boldsymbol{X}'_{\mathcal{A}_2}(\boldsymbol{I}-\boldsymbol{H}_{\mathcal{A}_1})(\boldsymbol{X}_{\mathcal{I}_1}\boldsymbol\beta^*_{\mathcal{I}_1}+\boldsymbol\epsilon)\|_2\\ &\le \big[nc_-(s)\big]^{-1}n\theta_{s,s}\|\boldsymbol\beta^*_{\mathcal{I}_1}\|_2+\|\boldsymbol{E}_{\mathcal{A}_{21}}\big(\boldsymbol{X}'_{\mathcal{A}_2}(\boldsymbol{I}-\boldsymbol{H}_{\mathcal{A}_1})\boldsymbol{X}_{\mathcal{A}_2}\big)^{-1}\boldsymbol{X}'_{\mathcal{A}_2}(\boldsymbol{I}-\boldsymbol{H}_{\mathcal{A}_1})\boldsymbol\epsilon\|_2.\end{aligned}\tag{8}$$

Combining (7) and (8), we have

$$\begin{aligned}\sqrt{\frac{|\mathcal{A}_{21}|}{|\mathcal{A}_{12}|}}\|\boldsymbol\beta^*_{\mathcal{A}_{12}}\|_2 \le \left(1+\sqrt{\frac{|\mathcal{A}_{21}|}{|\mathcal{A}_{12}|}}\right)\frac{\theta_{s,s}}{c_-(s)}\|\boldsymbol\beta^*_{\mathcal{I}_1}\|_2+&\\ \sqrt{\frac{|\mathcal{A}_{21}|}{|\mathcal{A}_{12}|}}\|\boldsymbol{E}_{\mathcal{A}_{12}}\big(\boldsymbol{X}'_{\mathcal{A}_1}(\boldsymbol{I}-\boldsymbol{H}_{\mathcal{A}_2})\boldsymbol{X}_{\mathcal{A}_1}\big)^{-1}\boldsymbol{X}'_{\mathcal{A}_1}(\boldsymbol{I}-\boldsymbol{H}_{\mathcal{A}_2})\boldsymbol\epsilon\|_2+&\\ \|\boldsymbol{E}_{\mathcal{A}_{21}}\big(\boldsymbol{X}'_{\mathcal{A}_2}(\boldsymbol{I}-\boldsymbol{H}_{\mathcal{A}_1})\boldsymbol{X}_{\mathcal{A}_2}\big)^{-1}\boldsymbol{X}'_{\mathcal{A}_2}(\boldsymbol{I}-\boldsymbol{H}_{\mathcal{A}_1})\boldsymbol\epsilon\|_2.&\end{aligned}$$

Denote $\boldsymbol{h}'_j=\boldsymbol{e}'_j\big(\boldsymbol{X}'_{\mathcal{A}_1}(\boldsymbol{I}-\boldsymbol{H}_{\mathcal{A}_2})\boldsymbol{X}_{\mathcal{A}_1}\big)^{-1}\boldsymbol{X}'_{\mathcal{A}_1}(\boldsymbol{I}-\boldsymbol{H}_{\mathcal{A}_2})$,

$$\|\boldsymbol{h}_j\|_2^2=\|\boldsymbol{e}'_j\big(\boldsymbol{X}'_{\mathcal{A}_1}(\boldsymbol{I}-\boldsymbol{H}_{\mathcal{A}_2})\boldsymbol{X}_{\mathcal{A}_1}\big)^{-1}\boldsymbol{e}_j\|_2^2\le\big[nc_-(s)\big]^{-1}.$$

It follows that, for some constant $\Delta>0$,

$$\begin{aligned}&\mathrm{P}\left(\sqrt{\frac{|\mathcal{A}_{21}|}{|\mathcal{A}_{12}|}}\|\boldsymbol{E}_{\mathcal{A}_{12}}\big(\boldsymbol{X}'_{\mathcal{A}_1}(\boldsymbol{I}-\boldsymbol{H}_{\mathcal{A}_2})\boldsymbol{X}_{\mathcal{A}_1}\big)^{-1}\boldsymbol{X}'_{\mathcal{A}_1}(\boldsymbol{I}-\boldsymbol{H}_{\mathcal{A}_2})\boldsymbol\epsilon\|_2>\Delta\sqrt{\frac{|\mathcal{A}_{21}|}{|\mathcal{A}_{12}|}}\ \frac{\theta_{s,s}}{c_-(s)}\|\boldsymbol\beta^*_{\mathcal{I}_1}\|_2\right)\\ \le&\sum_{j\in\mathcal{A}_{12}}\mathrm{P}\left(|\boldsymbol{h}'_j\boldsymbol\epsilon|>\frac{\Delta}{\sqrt{|\mathcal{A}_{12}|}}\frac{\theta_{s,s}}{c_-(s)}\|\boldsymbol\beta^*_{\mathcal{I}_1}\|_2\right)\\ \le&2p\exp\Big\{-\frac{nc_-(s)}{\sigma^2}\big(\frac{1}{\sqrt{|\mathcal{A}_{12}|}}\frac{\Delta\theta_{s,s}}{c_-(s)}\|\boldsymbol\beta^*_{\mathcal{I}_1}\|_2\big)^2\Big\}\\ \le&2p\exp\Big\{-\frac{nc_-(s)}{\sigma^2}\big(\sqrt{\frac{|\mathcal{I}_1|}{|\mathcal{A}_{12}|}}\frac{\Delta\theta_{s,s}}{c_-(s)}\min_{j\in\mathcal{A}^*}|\beta^*_j|\big)^2\Big\}\\ =&\frac{1}{6}\gamma_2,\end{aligned}$$

where $\gamma_2=12\exp\{\log p-K_{s,2}nb^*\}$ *and* $K_{s,2}=\frac{(\Delta\theta_{s,s})^2}{\sigma^2c_-(s)}$. *Similarly, we have*

$$\begin{aligned}&\mathrm{P}\left(\|\boldsymbol{E}_{\mathcal{A}_{21}}\big(\boldsymbol{X}'_{\mathcal{A}_2}(\boldsymbol{I}-\boldsymbol{H}_{\mathcal{A}_1})\boldsymbol{X}_{\mathcal{A}_2}\big)^{-1}\boldsymbol{X}'_{\mathcal{A}_2}(\boldsymbol{I}-\boldsymbol{H}_{\mathcal{A}_1})\boldsymbol\epsilon\|_2>\sqrt{\frac{|\mathcal{A}_{21}|}{|\mathcal{A}_{12}|}}\Delta\frac{\theta_{s,s}}{c_-(s)}\|\boldsymbol\beta^*_{\mathcal{I}_1}\|_2\right)\\ \le&2p\exp\Big\{-\frac{nc_-(s)}{\sigma^2}\big(\frac{\Delta}{\sqrt{|\mathcal{A}_{12}|}}\frac{\theta_{s,s}}{c_-(s)}\|\boldsymbol\beta^*_{\mathcal{I}_1}\|_2\big)^2\Big\}\\ \le&\frac{1}{6}\gamma_2.\end{aligned}$$

Therefore,

$$\begin{aligned}&\mathrm{P}\left(\|\boldsymbol\beta^*_{\mathcal{A}_{12}}\|_2\le2(1+\Delta)\frac{\theta_{s,s}}{c_-(s)}\|\boldsymbol\beta^*_{\mathcal{I}_1}\|_2\right)\\ \ge&\mathrm{P}\left(\|\boldsymbol\beta^*_{\mathcal{A}_{12}}\|_2\le\big(1+2\Delta+\frac{\sqrt{|\mathcal{A}_{12}|}}{\sqrt{|\mathcal{A}_{21}|}}\big)\frac{\theta_{s,s}}{c_-(s)}\|\boldsymbol\beta^*_{\mathcal{I}_1}\|_2\right)\\ \ge&1-\frac{1}{3}\gamma_2.\end{aligned}$$

Proof of Thoerem 3.

proof 2 *It follows from the proof of lemma 1, and inequalities (1) and (2) that*

$$\begin{aligned}2n\mathcal{L}_n(\boldsymbol\beta^{m+1})-2n\mathcal{L}_n(\boldsymbol\beta^*)&\le2n\mathcal{L}_n(\tilde{\boldsymbol\beta})-2n\mathcal{L}_n(\boldsymbol\beta^*)\\ &\le8nc_+(s)\Big((1+\Delta)\frac{\theta_{s,s}}{c_-(s)}\big(1+\frac{\theta_{s,s}}{c_-(s)}\big)\Big)^2\|\boldsymbol\beta^*_{\mathcal{I}_1^m}\|_2^2+f_1(\boldsymbol\epsilon)\\ &=\delta_s(1-\Delta)n(c_-(s)-\frac{\theta^2_{s,s}}{c_-(s)})\|\boldsymbol\beta^*_{\mathcal{I}_1^m}\|_2^2+f_1(\boldsymbol\epsilon)\\ &\le\delta_s(2n\mathcal{L}_n(\boldsymbol\beta^m)-2n\mathcal{L}_n(\boldsymbol\beta^*)).\end{aligned}\tag{9}$$

With probability $1-\gamma(s,n,p,b^*)$. Letting $\mathcal{A}^0=\varnothing$ and using (9) repeatedly, we have

$$2n\mathcal{L}_n(\boldsymbol{\beta}^m)-2n\mathcal{L}_n(\boldsymbol{\beta}^*)\leq\delta_s^m(2n\mathcal{L}_n(\boldsymbol{\beta}^0)-2n\mathcal{L}_n(\boldsymbol{\beta}^*)). \quad [10]$$

Now we prove part (i) of the theorem. If $m>\log_{\frac{1}{\delta_s}}\frac{\|\boldsymbol{y}\|_2^2}{n(1-\Delta)\left(c_-(s)-\frac{\theta_{s,s}^2}{c_-(s)^2}\right)b^*}>\log_{\frac{1}{\delta_s}}\left[\frac{\left|2n\mathcal{L}_n(\boldsymbol{\beta}^0)-2n\mathcal{L}_n(\boldsymbol{\beta}^*)\right|}{n(1-\Delta)\left(c_-(s)-\frac{\theta_{s,s}^2}{c_-(s)^2}\right)b^*}\right]$,

$$2n\mathcal{L}_n(\boldsymbol{\beta}^m)-2n\mathcal{L}_n(\boldsymbol{\beta}^*)\leq\delta_s^m(2n\mathcal{L}_n(\boldsymbol{\beta}^0)-2n\mathcal{L}_n(\boldsymbol{\beta}^*))<n(1-\Delta)\Big(c_-(s)-\frac{\theta_{s,s}^2}{c_-(s)^2}\Big)b^*. \quad [11]$$

Assume $\mathcal{A}^m\not\supseteq\mathcal{A}^*$. ,

$$\begin{aligned}2n\mathcal{L}_n(\boldsymbol{\beta}^m)-2n\mathcal{L}_n(\boldsymbol{\beta}^*)&=\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}^m\|_2^2-\boldsymbol{\epsilon}'\boldsymbol{\epsilon}=\|(\boldsymbol{I}-\boldsymbol{H}_{\mathcal{A}^m})\boldsymbol{y}\|_2^2-\boldsymbol{\epsilon}'\boldsymbol{\epsilon}\\&=\|(\boldsymbol{I}-\boldsymbol{H}_{\mathcal{A}^m})(\boldsymbol{X}_{\mathcal{I}^m}\boldsymbol{\beta}_{\mathcal{I}^m}^m+\boldsymbol{\epsilon})\|_2^2-\boldsymbol{\epsilon}'\boldsymbol{\epsilon}\\&\geq n(1-\Delta)\Big(c_-(s)-\frac{\theta_{s,s}^2}{c_-(s)^2}\Big)b^*,\end{aligned} \quad [12]$$

where $\boldsymbol{H}_{\mathcal{A}^m}=\boldsymbol{X}_{\mathcal{A}^m}(\boldsymbol{X}_{\mathcal{A}^m}^\top\boldsymbol{X}_{\mathcal{A}^m})^{-1}\boldsymbol{X}_{\mathcal{A}^m}^\top$, $b^*=\min_{j\in\mathcal{A}^*}(\beta_j^*)^2$. Combining (11) and (12) leads to a contradiction. Therefore $\mathcal{A}^m\supseteq\mathcal{A}^*$.

To prove part (ii) of the theorem, note that

$$\begin{aligned}\|\boldsymbol{\beta}^m-\boldsymbol{\beta}^*\|_2\leq&\|\boldsymbol{\beta}_{\mathcal{A}^m}^m-\boldsymbol{\beta}_{\mathcal{A}^m}^*\|_2+\|\boldsymbol{\beta}_{\mathcal{I}_1^m}^*\|_2\\\leq&\|(\boldsymbol{X}_{\mathcal{A}^m}'\boldsymbol{X}_{\mathcal{A}^m})^{-1}\boldsymbol{X}_{\mathcal{A}^m}'(\boldsymbol{X}\boldsymbol{\beta}^*+\boldsymbol{\epsilon}-\boldsymbol{X}_{\mathcal{A}^m}\boldsymbol{\beta}_{\mathcal{A}^m}^*)\|_2+\|\boldsymbol{\beta}_{\mathcal{I}_1^m}^*\|_2\\\leq&\|(\boldsymbol{X}_{\mathcal{A}^m}'\boldsymbol{X}_{\mathcal{A}^m})^{-1}\boldsymbol{X}_{\mathcal{A}^m}'\boldsymbol{X}_{\mathcal{I}_1^m}\boldsymbol{\beta}_{\mathcal{I}_1^m}^*\|_2+\|(\boldsymbol{X}_{\mathcal{A}^m}'\boldsymbol{X}_{\mathcal{A}^m})^{-1}\boldsymbol{X}_{\mathcal{A}^m}'\boldsymbol{\epsilon}\|_2+\|\boldsymbol{\beta}_{\mathcal{I}_1^m}^*\|_2\\\leq&(1+\Delta+\frac{\theta_{s,s}}{c_-(s)})\|\boldsymbol{\beta}_{\mathcal{I}_1^m}^*\|_2\\\leq&\frac{1+\Delta+\frac{\theta_{s,s}}{c_-(s)}}{(1-\Delta)n(c_-(s)-\frac{\theta_{s,s}^2}{c_-(s)})}\delta_s^m\left|2n\mathcal{L}_n(\boldsymbol{\beta}^0)-2n\mathcal{L}_n(\boldsymbol{\beta}^*)\right|\\\leq&\frac{1+\Delta+\frac{\theta_{s,s}}{c_-(s)}}{(1-\Delta)n(c_-(s)-\frac{\theta_{s,s}^2}{c_-(s)})}\delta_s^m\|\boldsymbol{y}\|_2^2.\end{aligned}$$

where the last second inequality follows from (9) and (10) .

Proof of Theorem 4.

proof 3 *For simplicity, denote $\mathcal{L}_{\mathcal{A}}=min_{\beta_{\mathcal{I}}=0}\mathcal{L}_n(\boldsymbol{\beta})$, where $\mathcal{I}=(\mathcal{A})^c$. We need to bound $\log\mathcal{L}_{\mathcal{A}}-\log\mathcal{L}_{\mathcal{B}}$. It follows from $1-\frac{1}{x}\leq\log(x)\leq x-1$ for $x>0$, that*

$$\frac{\mathcal{L}_{\mathcal{A}}-\mathcal{L}_{\mathcal{B}}}{\mathcal{L}_{\mathcal{A}}}\leq\log\frac{\mathcal{L}_{\mathcal{A}}}{\mathcal{L}_{\mathcal{B}}}\leq\frac{\mathcal{L}_{\mathcal{A}}-\mathcal{L}_{\mathcal{B}}}{\mathcal{L}_{\mathcal{B}}}. \quad [13]$$

Let $(\hat{\boldsymbol{\beta}}^s,\hat{\boldsymbol{d}}^s,\hat{\mathcal{A}}^s,\hat{\mathcal{I}}^s)$ be the output of Algorithm 1 with support size s. Firstly, consider the case when $s>s^*$. By Lemma 1, for any $s\geq s^*$,

$$\begin{aligned}\mathrm{P}(\hat{\mathcal{A}}^s\supseteq\mathcal{A}^*)\geq&1-\gamma(s,n,p,b^*)\\\geq&1-O(\exp\{\log p-K_s\log p\log\log n\})\\\geq&1-O(p^{-\alpha}),\end{aligned}$$

for some constant $\alpha>0$, and the last second inequality uses Condition (6) and Condition (7). Let $\hat{\mathcal{A}}^s=\mathcal{A}^*\cup\mathcal{B}^s$,

$$\begin{aligned}\mathcal{L}_{\mathcal{A}^*}-\mathcal{L}_{\hat{\mathcal{A}}^s}&=\frac{1}{2n}\boldsymbol{y}'(\boldsymbol{I}-\boldsymbol{H}_{\mathcal{A}^*})\boldsymbol{H}_{\hat{\mathcal{A}}^s}(\boldsymbol{I}-\boldsymbol{H}_{\mathcal{A}^*})\boldsymbol{y}\\&=\frac{1}{2n}\boldsymbol{y}'(\boldsymbol{I}-\boldsymbol{H}_{\mathcal{A}^*})\boldsymbol{X}_{\mathcal{B}^s}(\boldsymbol{X}_{\mathcal{B}^s}'(\boldsymbol{I}-\boldsymbol{H}_{\mathcal{A}^*})\boldsymbol{X}_{\mathcal{B}^s})^{-1}\boldsymbol{X}_{\mathcal{B}^s}'(\boldsymbol{I}-\boldsymbol{H}_{\mathcal{A}^*})\boldsymbol{y}\\&=\frac{1}{2n}\boldsymbol{\epsilon}'(\boldsymbol{I}-\boldsymbol{H}_{\mathcal{A}^*})\boldsymbol{X}_{\mathcal{B}^s}(\boldsymbol{X}_{\mathcal{B}^s}'(\boldsymbol{I}-\boldsymbol{H}_{\mathcal{A}^*})\boldsymbol{X}_{\mathcal{B}^s})^{-1}\boldsymbol{X}_{\mathcal{B}^s}'(\boldsymbol{I}-\boldsymbol{H}_{\mathcal{A}^*})\boldsymbol{\epsilon}.\end{aligned}$$

Note that,

$$
\begin{aligned}
&\mathrm{P}\left(\frac{1}{\sqrt{2n}}\|\left(\boldsymbol{X}'_{\mathcal{B}^s}(\boldsymbol{I}-\boldsymbol{H}_{\mathcal{A}^*})\boldsymbol{X}_{\mathcal{B}^s}\right)^{-\frac{1}{2}}\boldsymbol{X}'_{\mathcal{B}^s}(\boldsymbol{I}-\boldsymbol{H}_{\mathcal{A}^*})\boldsymbol{\epsilon}\|_2 \geq t\right)\\
\leq&\sum_{j\in\mathcal{B}^s}\mathrm{P}\left(\frac{1}{\sqrt{2n}}\|\boldsymbol{e}'_j\left(\boldsymbol{X}'_{\mathcal{B}^s}(\boldsymbol{I}-\boldsymbol{H}_{\mathcal{A}^*})\boldsymbol{X}_{\mathcal{B}^s}\right)^{-\frac{1}{2}}\boldsymbol{X}'_{\mathcal{B}^s}(\boldsymbol{I}-\boldsymbol{H}_{\mathcal{A}^*})\boldsymbol{\epsilon}\|_2 \geq \frac{t}{\sqrt{|\mathcal{B}^s|}}\right)\\
\leq&2p\exp\left\{-\frac{2nt^2}{|\mathcal{B}^s|\sigma^2}\right\}.
\end{aligned}
$$

Then with probability $1-(2p)^{-\alpha}$, we have

$$
\frac{1}{\sqrt{2n}}\|(\boldsymbol{X}'_{\mathcal{B}^s}(\boldsymbol{I}-\boldsymbol{H}_{\mathcal{A}^*})\boldsymbol{X}_{\mathcal{B}^s})^{-\frac{1}{2}}\boldsymbol{X}'_{\mathcal{B}^s}(\boldsymbol{I}-\boldsymbol{H}_{\mathcal{A}^*})\boldsymbol{\epsilon}\|_2 \leq \sqrt{\frac{\sigma^2|\mathcal{B}^s|}{2n}(1+\alpha)\log(2p)}. \tag{14}
$$

Therefore, with probability $1-O(p^{-\alpha})$,

$$
\mathcal{L}_{\mathcal{A}^*}-\mathcal{L}_{\hat{\mathcal{A}}^s} \leq \frac{\sigma^2|\mathcal{B}^s|}{2n}(1+\alpha)\log(2p). \tag{15}
$$

Now turn to $\mathcal{L}_{\hat{\mathcal{A}}^s}$. Similar to (14), with probability $1-O(p^{-\alpha})$

$$
\begin{aligned}
\mathcal{L}_{\hat{\mathcal{A}}^s} =&\frac{1}{2n}\|\boldsymbol{y}-\boldsymbol{X}_{\hat{\mathcal{A}}^s}\hat{\boldsymbol{\beta}}_{\hat{\mathcal{A}}^s}\|_2^2\\
\geq&\frac{1}{2n}\|\boldsymbol{\epsilon}\|_2^2-\frac{1}{2n}\|\boldsymbol{X}_{\hat{\mathcal{A}}^s}(\hat{\boldsymbol{\beta}}_{\hat{\mathcal{A}}^s}-\boldsymbol{\beta}^*_{\hat{\mathcal{A}}^s})\|_2^2\\
\geq&\frac{\sigma^2}{2}\left(1-\frac{1}{2n}\alpha\log(2p)\right)-\frac{\sigma^2 s}{2n}(1+\alpha)\log(2p)\\
>&0,
\end{aligned} \tag{16}
$$

where the last second inequality uses the fact that $\|\boldsymbol{\epsilon}\|_2^2$ is sub-exponential random variables and the last inequality is induced by Condition (7).

It follows from (13), (15) and (16) that with probability $1-O(p^{-\alpha})$,

$$
\log\frac{\mathcal{L}_{\mathcal{A}^*}}{\mathcal{L}_{\hat{\mathcal{A}}^s}} \leq \frac{\frac{\sigma^2|\mathcal{B}^s|}{n}(1+\alpha)\log(2p)}{\frac{\sigma^2}{2}(1-\frac{1}{2n}\alpha\log(2p))-\frac{\sigma^2 s}{2n}(1+\alpha)\log(2p)}.
$$

Consequently,

$$
\begin{aligned}
\mathrm{SIC}(\mathcal{A}^*)-\mathrm{SIC}(\hat{\mathcal{A}}^s) &= n\log\frac{\mathcal{L}_{\mathcal{A}^*}}{\mathcal{L}_{\hat{\mathcal{A}}^s}}-|\mathcal{B}^s|\log(p)\log\log n\\
&\leq O(|\mathcal{B}^s|\log(2p))-|\mathcal{B}^s|\log(p)\log\log n\\
&<0
\end{aligned}
$$

for a sufficiently large n.

Next, consider the case when $s<s^*$. Denote $\mathcal{I}_1^s=\hat{\mathcal{I}}^s\cap\mathcal{A}^*$. Similar to the (2) in Lemma 1, with probability $1-O(p^{-\alpha})$, for some constants $0<\Delta<\frac{1}{2}$, we have

$$
\begin{aligned}
2\mathcal{L}_{\hat{\mathcal{A}}^s} \geq&(1-\Delta)(c_-(s)-\frac{\theta^2_{s,s}}{c_-(s)})\|\boldsymbol{\beta}^*_{\mathcal{I}_1^s}\|_2^2+\sigma^2,\\
2\mathcal{L}_{\hat{\mathcal{A}}^s} \leq&(1+\Delta)c_+(s)\|\boldsymbol{\beta}^*_{\mathcal{I}_1^s}\|_2^2+\sigma^2.
\end{aligned} \tag{17}
$$

Denote $\boldsymbol{H}_A=\boldsymbol{X}_A(\boldsymbol{X}'_A\boldsymbol{X}_A)^{-1}\boldsymbol{X}'_A$. We have

$$
2n(\mathcal{L}_{\hat{\mathcal{A}}^s}-\mathcal{L}_{\mathcal{A}^*}) \geq n(1-\Delta)(c_-(s)-\frac{\theta^2_{s,s}}{c_-(s)})\|\boldsymbol{\beta}^*_{\mathcal{I}_1^s}\|_2^2-\|\boldsymbol{\epsilon}^\top\boldsymbol{H}_{\mathcal{A}^*}\boldsymbol{\epsilon}\|_2^2.
$$

We also have

$$
\begin{aligned}
&\mathrm{P}\left(\|(\boldsymbol{X}'_{\mathcal{A}^*}\boldsymbol{X}_{\mathcal{A}^*})^{-\frac{1}{2}}\boldsymbol{X}'_{\mathcal{A}^*}\boldsymbol{\epsilon}\|_2\geq t\right)\\
\leq&\sum_{j\in\mathcal{A}^*}\mathrm{P}\left(\|\boldsymbol{e}'_j(\boldsymbol{X}'_{\mathcal{A}^*}\boldsymbol{X}_{\mathcal{A}^*})^{-\frac{1}{2}}\boldsymbol{X}'_{\mathcal{A}^*}\boldsymbol{\epsilon}\|_2\geq\frac{t}{\sqrt{s^*}}\right)\\
\leq&2p\exp\{-\frac{t^2}{\sigma^2 s^*}\}.
\end{aligned}
$$

Thus,

$$
\begin{aligned}
&\mathrm{P}\left(\|\boldsymbol{\epsilon}^\top \boldsymbol{H}_{\mathcal{A}^*}\boldsymbol{\epsilon}\|_2^2 \geq n\Delta(c_-(s) - \frac{\theta_{s,s}^2}{c_-(s)})\|\boldsymbol{\beta}^*_{\mathcal{I}_1^s}\|_2^2\right)\\
\leq& 2\exp\{\log p - \frac{nK_{s,5}b^*}{s^*}\}\\
\leq& 2\exp\{\log p - K_{s,5}\log p \log\log n\}\\
\leq& O(p^{-\alpha}),
\end{aligned} \tag{18}
$$

where $K_{s,5} = \frac{\Delta(c_-(s)^2-\theta_{s,s}^2)}{c_-(s)\sigma^2}$ *and the second inequality uses Condition (6).*

Therefore, with probability $1 - O(p^{-\alpha})$*, we have*

$$
2(\mathcal{L}_{\hat{\mathcal{A}}^s} - \mathcal{L}_{\mathcal{A}^*}) \geq (1-2\Delta)(c_-(s) - \frac{\theta_{s,s}^2}{c_-(s)})\|\boldsymbol{\beta}^*_{\mathcal{I}_1^s}\|_2^2. \tag{19}
$$

It follows from inequalities (13), (17) and (19) that with probability $1 - O(p^{-\alpha})$*,*

$$
\log\frac{\mathcal{L}_{\hat{\mathcal{A}}^s}}{\mathcal{L}_{\hat{\mathcal{A}}^*}} \geq \frac{(1-2\Delta)(c_-(s) - \frac{\theta_{s,s}^2}{c_-(s)})\|\boldsymbol{\beta}^*_{\mathcal{I}_1^s}\|_2^2}{(1+\Delta)c_+(s)\|\boldsymbol{\beta}^*_{\mathcal{I}_1^s}\|_2^2 + \sigma^2}.
$$

Consequently, for sufficiently large n*,*

$$
\begin{aligned}
\mathrm{SIC}(\hat{\mathcal{A}}^s) - \mathrm{SIC}(\mathcal{A}^*) &= n\log\frac{\mathcal{L}_{\hat{\mathcal{A}}^s}}{\mathcal{L}_{\mathcal{A}^*}} - (s^* - |\hat{\mathcal{A}}^s|)\log(p)\log\log n\\
&\geq nO(\min\{1, |\mathcal{I}_1^s|b^*\}) - |\mathcal{I}_1^s|\log(p)\log\log n\\
&> 0,
\end{aligned}
$$

based on Condition (6) and Condition (7).

Therefore, information criterion $\mathrm{SIC}(\hat{\mathcal{A}})$ *attains the minimum at* $\mathcal{A}^*$ *with probability* $1 - O(p^{-\alpha})$*.*

Proof of Theorem 2.

proof 4 *Denote* $\mathcal{L}_{\mathcal{A}} = min_{\boldsymbol{\beta}_{\mathcal{I}}=0}\mathcal{L}_n(\boldsymbol{\beta})$*, where* $\mathcal{I} = (\mathcal{A})^c$*. First of all, consider* $0 \leq s < s^*$*. Since the loss function decreases by at least* τ_s *at each iteration, Algorithm 1 stops before* $O(\frac{\|\boldsymbol{y}\|_2^2}{\tau_s})$ *iterations.*

Next, consider $s^* \leq s \leq s_{max}$*, for a fix* $s > 0$*. Denote* $\mathcal{A}^m$ *as the active set output by Algorithm 1 in* m*th iteration. Assuming* $\mathcal{A}^m \not\supseteq \mathcal{A}^*$*, by (9), we have*

$$
\mathcal{L}_{\mathcal{A}^m} - \mathcal{L}_{\mathcal{A}^{m+1}} \geq (1-\delta_s)(\mathcal{L}_{\mathcal{A}^m} - \mathcal{L}_n(\boldsymbol{\beta}^*)) \geq (1-\delta_s)(1-\Delta)(c_-(s) - \frac{\theta_{s,s}^2}{c_-(s)})\|\boldsymbol{\beta}^*_{\mathcal{I}_1^m}\|_2^2 > \tau_s,
$$

so the difference between the m*th loss function and the* $(m+1)$*th loss function is bigger than the threshold* τ_s*. Thus, from Theorem 3, we have* $\mathcal{A}^m \supseteq \mathcal{A}^*$ *after* $O(\log\frac{\|\boldsymbol{y}\|_2^2}{s\log p\log\log n})$ *iterations.*

Now $\mathcal{A}^{m+1} \supseteq \mathcal{A}^*$*, then* $\mathcal{L}_{\mathcal{A}^{m+1}} = \boldsymbol{\epsilon}'\boldsymbol{H}_{\mathcal{A}^{m+1}}\boldsymbol{\epsilon}$ *and* $\mathcal{L}_{\mathcal{A}^m} = \boldsymbol{\epsilon}'\boldsymbol{H}_{\mathcal{A}^m}\boldsymbol{\epsilon}$*. By (16), with probability* $1 - O(p^{-\alpha})$*, we have*

$$
\begin{aligned}
\mathcal{L}_{\mathcal{A}^m} &\leq \frac{\sigma^2}{2} + \frac{\sigma^2}{4n}\alpha\log(2p) + \frac{\sigma^2 s}{2n}(1+\alpha)\log(2p),\\
\mathcal{L}_{\mathcal{A}^{m+1}} &\geq \frac{\sigma^2}{2} - \frac{\sigma^2}{4n}\alpha\log(2p) - \frac{\sigma^2 s}{2n}(1+\alpha)\log(2p).
\end{aligned}
$$

Thus, for sufficient large n*, we have*

$$
\mathcal{L}_{\mathcal{A}^m} - \mathcal{L}_{\mathcal{A}^{m+1}} \leq \frac{\sigma^2}{2n}\alpha\log(2p) + \frac{\sigma^2 s}{n}(1+\alpha)\log(2p) \leq \tau_s.
$$

Therefore, for $s^* \leq s \leq s_{max}$*, Algorithm 1 terminates after* $O(\log\frac{\|\boldsymbol{y}\|_2^2}{s\log p\log\log n})$ *iterations.*

Now we analyze the computational complexity of Algorithm 2 for a given active set size s*. Computing* $\boldsymbol{\xi}$ *and* $\boldsymbol{\zeta}$ *takes* $O(np + ns)$ *steps. Finding the smallest (largest)* s *values in* $\boldsymbol{\xi}$ *(*$\boldsymbol{\zeta}$*) takes* $O(p)$ *steps via Hoare's selection algorithm (12). Because the procedure repeats* k_{max}*,* $O(k_{max}p)$ *steps are demanded. In Algorithm 2, the splicing method iterates at most* s *times. Therefore the computational complexity of Algorithm 1 is*

$$
O\Big(\log\frac{\|\boldsymbol{y}\|_2^2}{s\log p\log\log n}\mathbb{I}(s^* \leq s) + \frac{\|\boldsymbol{y}\|_2^2}{\tau_s}\mathbb{I}(s^* > s)\Big)\cdot O(nsp + ns^2 + k_{max}sp),
$$

where $\mathbb{I}(\cdot)$ *is a indicator function. Since* s *varies from 1 to* s_{max} *in Algorithm 3, the total computational complexity is*

$$
\begin{aligned}
&O\Big(\log\frac{\|\boldsymbol{y}\|_2^2}{\log p\log\log n}\big(nps_{max}^2 + ns_{max}^3 + k_{max}ps_{max}^2\big) + \frac{n\|\boldsymbol{y}\|_2^2}{\log p\log\log n}\big(nps^* + n(s^*)^2 + k_{max}ps^*\big)\Big)\\
\leq& O\Big(\big(s_{max}\log\frac{\|\boldsymbol{y}\|_2^2}{\log p\log\log n} + \frac{n\|\boldsymbol{y}\|_2^2}{\log p\log\log n}\big)\big(nps_{max} + ns_{max}^2 + k_{max}ps_{max}\big)\Big).
\end{aligned}
$$

References

1. H Hazimeh, R Mazumder, Fast best subset selection: Coordinate descent and local combinatorial optimization algorithms. *Oper. Res.* (2020).
2. C Wen, A Zhang, S Quan, X Wang, Bess: An r package for best subset selection in linear, logistic and cox proportional hazards models. *J. Stat. Softw.* **94**, 1–24 (2020).
3. J Friedman, T Hastie, R Tibshirani, Regularization paths for generalized linear models via coordinate descent. *J. Stat. Software, Articles* **33**, 1–22 (2010).
4. P Breheny, J Huang, Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Ann. Appl. Stat.* **5**, 232–253 (2011).
5. H Hazimeh, R Mazumder, *L0Learn: Fast Algorithms for Best Subset Selection*, (2019) R package version 1.2.0.
6. C Wen, A Zhang, S Quan, X Wang, *BeSS: Best Subset Selection for Sparse Generalized Linear Model and Cox Model*, (2017) R package version 1.0.2.
7. J Huang, Y Jiao, Y Liu, X Lu, A constructive approach to l_0 penalized regression. *J. Mach. Learn. Res.* **19**, 1–37 (2018).
8. L Wang, Y Kim, R Li, Calibrating non-convex penalized regression in ultra-high dimension. *Annals statistics* **41**, 2505 (2013).
9. DF Saldana, Y Feng, Sis: an r package for sure independence screening in ultrahigh dimensional statistical models. *J. Stat. Softw.* **83**, 1–25 (2018).
10. H Zou, T Hastie, R Tibshirani, , et al., On the "degrees of freedom" of the lasso. *The Annals Stat.* **35**, 2173–2192 (2007).
11. J Kiefer, Sequential minimax search for a maximum. *Proc. Am. mathematical society* **4**, 502–506 (1953).
12. CAR Hoare, Algorithm 65: Find. *Commun. ACM* **4**, 321–322 (1961).