Data Science in Kyoto Women's University 1 (2026) 61 − 78

Cardinality-Specified Sparse Regression for Variable Selection: Comparisons Against Lasso and All Possible Regression

Kohei Adachi*

Abstract

Variable selection in regression analysis is considered in this paper, with the selection purposed to find the best subset of predictors. The number of all possible subsets is $2^p - 1$ for the full set of p predictors. The best one among $2^p - 1$ subsets can be found through all possible regression (APREG). For avoiding the difficulty that APREG is time-consuming, heuristic variable selection procedures (such as variable increasing/decreasing and stepwise ones) used to be prevalent, but can be considered as useless now, for the following reasons: As described in the first half of this paper, APREG can be performed instantly when $p \leq 13$, and we can consider that the variable selection is achieved efficiently by sparse regression procedures even if $p > 13$. Among those procedures, the one called cardinality-specified regression (CSREG) is focused on and compared with well-known lasso, in this paper.

CSREG differs from lasso as follows: The ordinary regression function is minimized subject to the cardinality of regression coefficients being specified in CSREG, while the regression function plus a penalty function is minimized in lasso. The results following from the difference are shown mathematically and numerically in this paper. The mathematical results are summarized as follows: [1] The CSREG solution is scale invariant and equivalent to the APREG counterpart, if the iterative algorithm for CSREG provides the global minimizers; [2] LASSO cannot provide the estimates of coefficients that are same as the APREG counterparts. These results are numerically demonstrated by a real data example and ascertained in a simulation study.

In the simulation study, CSREG and lasso are performed for the data sets generated from the true subsets in some conditions. We present the detailed results for how CSREG and lasso differ in variable selection. Additionally, we report and discuss how well the true subsets can be recovered in CSREG and lasso.

Keywords: Sparse regression, Variable selection, Best subset selection, All possible regression, Cardinality-specified regression, Lasso, Scale invariance

* Faculty of Data Science, Kyoto Women's University, Japan
E-mail: adachik@kyoto-wu.ac.jp

1. Introduction

Let [**y, X**] be the n-observations × (1+p)-variables block data matrix whose column averages are zeros, with the columns of $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_p]$ ($n \times p$) the vectors of predictors for $\mathbf{y}$ ($n \times 1$). Then, regression analysis is modeled as $\mathbf{y} = \mathbf{X\beta} + \mathrm{e}$, where $\boldsymbol{\beta} = [\beta_1, \dots, \beta_p]'$ (p × 1) and $\mathbf{e}$ ($n \times 1$) contain regression coefficients and errors, respectively. In the regression analysis for [$\mathbf{y}$, $\mathbf{X}$], $f(\boldsymbol{\beta}) = \|\mathbf{e}\|^2 = \|\mathbf{y} - \mathbf{X\beta}\|^2$ is minimized over $\boldsymbol{\beta}$. Theoretical details of regression analysis can be seen in Sawa (1979) and Fahrmeir, Kneib, Lang, and Marx (2021), for example.

Let $\mathbf{S}_{kl}$, $l = 1, \cdots, {}_p\mathrm{C}_k$, be the subsets of the full set $\{\mathbf{x}_1, \cdots, \mathbf{x}_p\}$ with ${}_p\mathrm{C}_k = p!/\{k!(p-k)!\}$. For example, if $p = 3$ and $k = 2$ with the full set $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$ and ${}_p\mathrm{C}_k = 3$, then $\mathbf{S}_{21} = \{\mathbf{x}_1, \mathbf{x}_2\}$, $\mathbf{S}_{22} = \{\mathbf{x}_1, \mathbf{x}_3\}$, $\mathbf{S}_{23} = \{\mathbf{x}_2, \mathbf{x}_3\}$. The regression of $\mathbf{y}$ onto the predictor vectors in $\mathbf{S}_{kl}$ is modeled as

$$\mathbf{y} = \mathbf{X}_{[kl]}\boldsymbol{\beta}_{[kl]} + \mathbf{e}_{kl} = \mathbf{X}\boldsymbol{\beta}_{kl} + \mathbf{e}_{kl}. \tag{1}$$

Here, $\mathbf{e}_{kl}$ ($n \times 1$) contains errors, the k columns of $\mathbf{X}_{[kl]}$ ($n \times k$) are the vectors in $\mathbf{S}_{kl}$, $\boldsymbol{\beta}_{[kl]}$ ($k \times 1$) contains the k elements of $\boldsymbol{\beta}$ corresponding to the columns of $\mathbf{X}_{[kl]}$, and $\boldsymbol{\beta}_{kl}$ is the $p \times 1$ vector obtained by substituting zeros for the elements of $\boldsymbol{\beta}$ that are not included is $\boldsymbol{\beta}_{kl}$. For example, when $p = 3$ and $\mathbf{S}_{23} = \{\mathbf{x}_2, \mathbf{x}_3\}$, $\boldsymbol{\beta}_{[23]} = [\beta_2, \beta_3]'$ and $\boldsymbol{\beta}_{23} = [0, \beta_2, \beta_3]'$. The regression analysis for (1) can be formulated as

$$\text{minimize } f(\boldsymbol{\beta}_{kl}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_{kl}\|^2 \text{ over } \boldsymbol{\beta}_{kl}\,. \tag{2}$$

Let $\hat{\boldsymbol{\beta}}_{kl}$ ($p \times 1$) denote the solution of (2) and $\phi(\boldsymbol{\beta}_{kl})$ denote a model selection criterion, whose relative smallness of the $\phi(\boldsymbol{\beta}_{kl})$ value stands for the goodness of model (1). As $\phi(\boldsymbol{\beta}_{kl})$, we can consider information criterions (Konishi & Kitagawa, 2007) and a C_P statistic (Mallows, 1973). Then, the best subset for predicting $\mathbf{y}$ can be expressed as $\mathbf{S}_{\hat{k}\hat{l}}$. Here, $[\hat{k}, \hat{l}]$ is the $[k, l]$ that minimizes $\phi(\hat{\boldsymbol{\beta}}_{kl})$ among $k = 0, 1, \cdots, p$, and $l = 1, \cdots, {}_p\mathrm{C}_k$. Only for $k = 0$ with ${}_p\mathrm{C}_k = 1$ and $\mathbf{S}_{kl} = \mathbf{S}_{k1} = \varnothing$, $\boldsymbol{\beta}_{kl}$ equals the $p \times 1$ zero vector $\mathbf{0}$ and (2) may not be run.

Obviously, the best subset can be selected through all possible regression (APREG), i.e., performimg (2) over all possible k and l. This can be formally listed as follows:

[AR1] For each of $k = 1, \dots p$, perform (2) to obtain $\phi(\hat{\boldsymbol{\beta}}_{kl})$ over $l = 1, \dots, {}_p\mathrm{C}_k$ and select

$$\hat{\boldsymbol{\beta}}_k = \underset{l=1,\dots,{}_p C_k}{\arg\min}\ \phi(\hat{\boldsymbol{\beta}}_{kl})$$

[AR2] Select $\hat{\boldsymbol{\beta}} = \underset{k=0,1,\dots,p}{\arg\min}\ \phi(\hat{\boldsymbol{\beta}}_k)$.

The resulting $\hat{\boldsymbol{\beta}}$ shows the best subset; the nonzero elements in $\hat{\boldsymbol{\beta}}$ correspond to the predictors in the best subset. Let us note that (2) is performed

$${}_p\mathrm{C}_1 + \dots + {}_p\mathrm{C}_p = 2^p - 1 \tag{3}$$

times in [AR1] (e.g., Sawa, 1979, Eq. (7.1)).

For example, when $p = 13$ and (3) equals $2^{13} - 1 = 8{,}191$, APREG can be run to provide the best subset instantly (with its FORTRAN code) as described in Section 5. But, greater p gives a very great value of (3); for example, $p = 30$ leads to $2^p - 1 = 1{,}073{,}741{,}823$. These facts show

that APREG should be used for $p \leq 13$, but is considered very time-consuming for p being great. For avoiding the difficulty in APREG being time-consuming, heuristic variable selection procedures (such as variable increasing/decreasing and stepwise ones) used to be prevalent but are considered as useless now. It is because sparse regression procedures have been presented that perform variable selection not heuristically but logically and optimally.

Among the sparse regression procedures, we consider using two sparse regression procedures in place of APREG. One of the two procedures is lasso (Tibshirani, 1996) and the other is referred to as cardinality-specified regression (CSREG) in this paper. CSREG was proposed mutually independently in Adachi and Kiers (2017) and Xiong (2014). In contrast to lasso being well-known, CSREG is not so. Further, properties of CSREG in the best subset selection have not been discussed in Adachi and Kiers (2017) and Xiong (2014). In this paper, we will show how CSRER is useful for the selection

The rest of the paper is organized as follows: In Section 2, we describe CSREG, lasso, and their-based best subset selection. In Section 3, we contrast CSREG and lasso in how each procedure is advantageous. In Section 4, we describe the details of CSREG- and lasso-based best subset selection. In Sections 5 and 6, we compare the CSREG- and lasso-based selection with real and simulated data sets, which is followed by the discussion in Section 7.

2. Sparse Regression Procedures

First, we introduce cardinality-specified regression (CSREG) and discuss the best subset selection (BSS) based on CSREG. Then, we describe lasso and its based BSS.

2.1. Cardinality-Specified Regression

Let Card($\mathbf{\beta}$) denote the cardinality of $\mathbf{\beta}$, i.e., the number of the nonzero elements in $\mathbf{\beta}$. CSREG is formulated as

$$\text{minimize } f(\mathbf{\beta}) = \|\mathbf{y} - \mathbf{X}\mathbf{\beta}\|^2 \text{ over } \mathbf{\beta} \text{ subject to Card}(\mathbf{\beta}) = k\,, \tag{4}$$

(Adachi and Kiers, 2017; Xiong, 2014); (4) is the constrained regressin analysis with Card($\mathbf{\beta}$) specified to be an integer k, as the name CSREG shows. Let $\hat{\mathbf{\beta}}_k^{\mathrm{C}}$ denote $\mathbf{\beta}$ resulting in (4). The CSREG-based BSS can be listed as follows:

[CR1] For each of $k = 1, \ldots p$, perform (4) and obtain $\phi(\hat{\mathbf{\beta}}_k^{\mathrm{C}})$.

[CR2] Select $\hat{\mathbf{\beta}}^{\mathrm{C}} = \underset{k=0,1,\ldots,p}{\arg\min}\, \phi(\hat{\mathbf{\beta}}_k^{\mathrm{C}})$.

If the global minimizer for (4) is found by an algorithm for CSREG (to be introduced later), $\hat{\mathbf{\beta}}^{\mathrm{C}}$ resulting in [CR2] is equal to $\hat{\mathbf{\beta}}$ given by APREG, as explained in the next paragraph.

The definitions of the information criterions and C_{P} statistic, which are considered as model selection criterion $\phi(\mathbf{\beta})$, can be written in the form

$$\phi(\mathbf{\beta}) = c_1 g(\|\mathbf{y} - \mathbf{X}\mathbf{\beta}\|^2) + c_2 k + c_3\,. \tag{5}$$

Here, $g(y)$ is a function satisfying $g(y_1) < g(y_2)$ for $y_1 < y_2$ with $c_1 - c_3$ positive scalars independent of $\boldsymbol{\beta}$ and k. Thus, the minimization of $f(\boldsymbol{\beta})$ amount to that of $\phi(\boldsymbol{\beta})$ for k being kept fixed. Further, Card($\boldsymbol{\beta}$) = k in (4) is equivalent to $\boldsymbol{\beta} \in \{\boldsymbol{\beta}_{kl};\ l = 1, \ldots, {}_pC_k\}$. These facts allow (4) to be rewritten as

$$\text{minimize } \phi(\boldsymbol{\beta}) \text{ over } \boldsymbol{\beta} \text{ subject to } \boldsymbol{\beta} \in \{\boldsymbol{\beta}_{kl};\ l = 1 \ldots, {}_pC_k\}\ . \tag{4′}$$

This shows that $\hat{\boldsymbol{\beta}}_k^{\mathrm{C}}$ in [CR1] equals $\hat{\boldsymbol{\beta}}_k = \arg\min_{l=1\cdots pCk} \phi(\hat{\boldsymbol{\beta}}_{kl})$ in [AR1]; $\hat{\boldsymbol{\beta}}_k^{\mathrm{C}} = \hat{\boldsymbol{\beta}}_k$. This equality equalizes [CR2] to [AR2]. Thus, the CSREG-based BSS can provide the solution of APREG.

Regression analysis must be performed $2^p - 1$ times in APREG, while CSREG (4) is run p times in CSREG-based BSS. Simply comparing p against $2^p - 1$ allows us to consider that APREG is much more time-consuming than the CSREG-based BSS for $p << 2^p - 1$, though CSREG needs iterative algorithm descrbed next.

Adachi and Kiers (2017) and Xiong (2014) mutually independently have presented the same algorithm for CSREG (4). This algorithm, which is described in Appendix 1, can give a local minimizer. If it is given by the CSREG algorithm, then (4) is not attained and $\hat{\boldsymbol{\beta}}^{\mathrm{C}}$ resulting in [CR2] is not equal to $\hat{\boldsymbol{\beta}}$. However, this possibility of $\hat{\boldsymbol{\beta}}^{\mathrm{C}} \neq \hat{\boldsymbol{\beta}}$ can be reduced by running the CSREG algorithm multiple times to have multiple solutions and select the best solution among the multiple ones. We take this multiple-runs procedure. Thus, the above [CR1] will be modified in Section 4.2.

2.2. Lasso

To the best of our knowledge, the loss function is not simply regression function $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$, but rather the sum of the function and the penalty term that penalizes $\boldsymbol{\beta}$ having nonzero elements (e.g., Hastie, Tibshirani & Wainwright, 2015; Fan & Li, 2001; Mazumider, Friedman & Hastie, 2011), in almost all of sparse regression procedures. Among them, a prevalent one is Tibshirani's (1996) lasso, in which the penalty term is $\lambda\|\boldsymbol{\beta}\|_1$ with $\lambda > 0$ and $\|\boldsymbol{\beta}\|_1$ the L_1 norm of $\boldsymbol{\beta}$: lasso is formulated as

$$\text{minimize } g(\boldsymbol{\beta}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_1 \text{ over } \boldsymbol{\beta} \text{ for specified } \lambda > 0\ , \tag{6}$$

where λ is the tuning parameter that specifies the weight for penalty function $\|\boldsymbol{\beta}\|_1$. Let $\hat{\boldsymbol{\beta}}_\lambda^{\mathrm{L}}$ be the minimizer of (6). A larger λ leads to gives $\hat{\boldsymbol{\beta}}_\lambda^{\mathrm{L}}$ with a smaller Card($\hat{\boldsymbol{\beta}}_\lambda^{\mathrm{L}}$). But, what Card($\hat{\boldsymbol{\beta}}_\lambda^{\mathrm{L}}$) follows from λ is unknown before (6); λ is specified before (6), but Card($\hat{\boldsymbol{\beta}}_\lambda^{\mathrm{L}}$) is found after (6).

Tuning parameter λ is a real number with $\lambda \subseteq [0, \infty]$. Thus, it is impossible to consider the solutions for all possible values of λ. However, we suppose that $\boldsymbol{\Lambda} = \{\lambda_1, \ldots, \lambda_L\}$ is a suitable set of the candidates for λ. Then, a lasso-based BSS procedure can be listed as

[LA1] For each of $\lambda \in \boldsymbol{\Lambda}$, perform (6) to obtain $\phi(\hat{\boldsymbol{\beta}}_\lambda^{\mathrm{L}})$.

[LA2] Select $\hat{\boldsymbol{\beta}}^{\mathrm{L}} = \arg\min_{\lambda \in \boldsymbol{\Lambda}} \phi(\hat{\boldsymbol{\beta}}_\lambda^{\mathrm{L}})$.

The resulting $\hat{\boldsymbol{\beta}}^{\mathrm{L}}$ *can suggest* the best subset. Here, we have used "*can suggest*" rather than "show", as $\hat{\boldsymbol{\beta}}^{\mathrm{L}}$ is not guaranteed to show the best subset, since the loss function in (6) differs

from that in (2) and (4), i.e., the regression function. Further, this difference leads to

$$\hat{\boldsymbol{\beta}}^{\mathrm{L}}_{\lambda} \neq \hat{\boldsymbol{\beta}}_{\mathrm{Card}(\hat{\boldsymbol{\beta}}^{\mathrm{L}}_{\lambda}),l(\hat{\boldsymbol{\beta}}^{\mathrm{L}}_{\lambda})} \tag{7}$$

in general with typically $\left|(\hat{\boldsymbol{\beta}}^{\mathrm{L}}_{\lambda})_j\right| \leq \left|(\hat{\boldsymbol{\beta}}_{\mathrm{Card}(\hat{\boldsymbol{\beta}}^{\mathrm{L}}_{\lambda}),l(\hat{\boldsymbol{\beta}}^{\mathrm{L}}_{\lambda})})_j\right|$ (e.g., Adachi, 2020, p. 353). Here, $\hat{\boldsymbol{\beta}}_{\mathrm{Card}(\hat{\boldsymbol{\beta}}^{\mathrm{L}}_{\lambda}),l(\hat{\boldsymbol{\beta}}^{\mathrm{L}}_{\lambda})}$ is the regression solution corresponding to the lasso $\hat{\boldsymbol{\beta}}^{\mathrm{L}}_{\lambda}$, in more exactness, the solution $\hat{\boldsymbol{\beta}}_{kl}$ for (2) when $[k, l]$ is set to $[\mathrm{Card}(\hat{\boldsymbol{\beta}}^{\mathrm{L}}), \mathrm{l}(\hat{\boldsymbol{\beta}}^{\mathrm{L}}_{\lambda})]$ so that the nonzero elements of $\hat{\boldsymbol{\beta}}_{kl}$ are those of $\hat{\boldsymbol{\beta}}^{\mathrm{L}}_{\lambda}$, with $(\mathbf{b})_j$ the jth element of vector $\mathbf{b}$. Thus, even if the nonzero elements in $\hat{\boldsymbol{\beta}}^{\mathrm{L}}$ exactly correspond to the predictors in the best subset, $\hat{\boldsymbol{\beta}}^{\mathrm{L}}$ can only *approximate*, but not equal $\hat{\boldsymbol{\beta}}$ in the model with the best subset.

One advantage of lasso over CSREG is that $\boldsymbol{\beta}$ is continuous and the loss function in (6) is convex, as this property prevents an algorithm for lasso from giving a local minimizer.

3. Contrasting CSREG and Lasso

In this section, we summarize the properties of CSREG and lasso so that they are contrasted in their advantages/disadvantages. Here, the properties have already been mentioned in Section 2, except a property to be presented by Theorem 1 in Section 3.2.

3.1. One Advantage of Lasso

In CSREG, $\boldsymbol{\beta}$ is constrained as $\mathrm{Card}(\boldsymbol{\beta}) = k$ and discontinuous: for example, when $p = 3$ and $k = 2$, the vectors other than $[\beta_1, \beta_2, 0]'$, $[\beta_1, 0, \beta_3]'$, and $[\beta_1, \beta_2, 0]'$ are excluded from the domain of the definition of $\boldsymbol{\beta}$. Thus, the CSREG algorithm can provide a local minimizer and must be run multiple times so that we select the best one among the resulting multiple solutions. Such a remedy for avoiding the selection of a local minimizer is unnecessary in lasso.

3.2. Three Advantages of CSREG

The tunning parameter in CSREG (4) is k, which is restricted to an integer in the range $[1, p]$. Thus, we can consider all possible k in the CSREG-based best subset selection (BSS) [CR1]−[CR2]. On the other hand, the tunning parameter in lasso (4) is λ, which is a nonnegative real values $\subseteq [0, \infty]$. Thus, we cannot cosider the solutions for all possible λ values. Further, what $\mathrm{Card}(\hat{\boldsymbol{\beta}}^{\mathrm{L}}_{\lambda})$ follows from λ is unknown beforehand. Thus, we must take efforts to find $\boldsymbol{\Lambda} = \{\lambda_1, \ldots, \lambda_L\}$ (the suitable set of the candidates for λ) to construct the lasso-based BSS [LA1]-[LA2].

As (7) shows, the lasso solution differs from the corresponding regression solution. Thus, lasso-based BSS cannot give the solution equal to $\hat{\boldsymbol{\beta}}_{\hat{k}\hat{l}}$ given by APREG. In contrast, the CSREG-based BSS gives $\hat{\boldsymbol{\beta}}_{\hat{k}\hat{l}}$, unless a local minimizer is selected as the best solution.

The lasso solution is scale-variant; the solution for an unstandardized data whose predictors have heterogeneous variances is essentially different from that for the standardized counterpart with variances homogeneous among variables. Thus, lasso is recommended to be performed for

standardized data (Hastie, Tibshirani, & Wainwright, 2015). But, the CSREG solution is scale-invariant, as is the ordinary regression solution. This fact is proved as follows:

Theorem 1. *Let* **D** *be a p* × *p positive definite* (*p.d.*) *diagonal matrix, v be a positive scalar, and* $\mathbf{D}v = v^{-1}\mathbf{D}$. *When the columns of* $\mathbf{XD}^{-1}$ *are the predictors for* $v^{-1}\mathbf{y}$, *the loss function of CSREG can be expressed as*

$$f^*(\boldsymbol{\beta}^*) = v^2 \left\| \frac{1}{v}\mathbf{y} - \mathbf{XD}^{-1}\boldsymbol{\beta}^* \right\|^2 = \|\mathbf{y} - \mathbf{XD}_v^{-1}\boldsymbol{\beta}^*\|^2 , \tag{8}$$

and CSREG is formulated as

$$\text{minimize } f^*(\boldsymbol{\beta}^*) = \|\mathrm{y} - \mathbf{XD}_v^{-1}\boldsymbol{\beta}^*\|^2 \text{ over } \boldsymbol{\beta}^* \text{ subject to Card}(\boldsymbol{\beta}^*) = k . \tag{9}$$

This solution is given by $\hat{\boldsymbol{\beta}}_k^* = \mathbf{D}_v\hat{\boldsymbol{\beta}}_k^{\mathrm{C}}$ with $f^*(\hat{\boldsymbol{\beta}}_k^*) = f(\hat{\boldsymbol{\beta}}_k^{\mathrm{C}})$. *Here,* $\hat{\boldsymbol{\beta}}_k^{\mathrm{C}}$ *the solution of* (4) *and f*(**β**) *is the loss function in* (4).

Proof. Since both $\boldsymbol{\beta}^*$ and $\mathbf{D}_v\boldsymbol{\beta}$ stand for an arbitrary $p \times 1$ vector, we can set $\boldsymbol{\beta}^* = \mathbf{D}v\boldsymbol{\beta}$ in (9), i.e., relace its $\boldsymbol{\beta}^*$by $\mathbf{D}_v\boldsymbol{\beta}$; (9) can be rewritten as

$$\text{minimize } f^*(\mathbf{D}_v\boldsymbol{\beta}) = \|\mathbf{y} - \mathbf{XD}_v^{-1}(\mathbf{D}_v\boldsymbol{\beta})\|^2 \text{ over } \mathbf{D}v\boldsymbol{\beta} \text{ subject to Card}(\mathbf{D}_v\boldsymbol{\beta}) = k . \tag{9′}$$

This is equivalent to (4) since of the following three facts: [1] $f^*(\mathbf{D}v\boldsymbol{\beta}) = \|\mathbf{y} - \mathbf{XD}_v^{-1}(\mathbf{D}_v\boldsymbol{\beta})\|^2 = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 = f(\boldsymbol{\beta})$; [2] Card($\mathbf{D}_v\boldsymbol{\beta}$) = *k* is equivalent to Card(**β**) = *k* since $\mathbf{D}_v$ is p.d. diagonal; [3] "over $\mathbf{D}_v\boldsymbol{\beta}$" may be rewritten as "over **β**" since $\mathbf{D}_v\boldsymbol{\beta}$ and **β** stand for an arbitrary $p \times 1$ vector.

Equation $\boldsymbol{\beta}^* = \mathbf{D}_v\boldsymbol{\beta}$ and the equivalence of (9′) and (4) with their solutions being $\hat{\boldsymbol{\beta}}_k^*$ and $\hat{\boldsymbol{\beta}}_k^{\mathrm{C}}$ lead to $\hat{\boldsymbol{\beta}}_k^* = \mathbf{D}_v\hat{\boldsymbol{\beta}}_k^{\mathrm{C}}$. This equation and $f^*(\mathbf{D}v\boldsymbol{\beta}) = f(\boldsymbol{\beta})$ in [1] leads to $f^*(\hat{\boldsymbol{\beta}}_k^*) = f^*(\mathbf{D}_v\hat{\boldsymbol{\beta}}_k^{\mathrm{C}}) = f(\hat{\boldsymbol{\beta}}_k^{\mathrm{C}})$. ■

If the *j*th diagonal element of **D** in (8) is the standard deviation (SD) of the corresponding column of **X**, then (9) stands for CSREG for the standardized predictors with homogeneous variances. Further, if *v* being SD of the elements in **y** is added to the above specification of **D**, (9) is CSREG the standardized scores of [**X**, **y**].

4. Details of Procedures for Best Subset Selection

We describe the details of the procedures in APREG, CSREG-based best subset selection (BSS), and lasso-based BSS. The procedures will be used for the computations in Sections 5 and 6.

4.1. Model Selection Criterion

As model selection criterion $\phi(\boldsymbol{\beta})$, we use Bayesian information criterion (Schwarz, 1978), which is expressed as BIC = $n\log\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + n\log\{\text{Card}(\boldsymbol{\beta})\}$ and is a special case of (5).

4.2. Multiple Runs of CSREG

The CSREG algorithm is started 30 times with mutually different initial values of **β**. Here, **β** is initialized as the ordinary regression estimate $\hat{\boldsymbol{\beta}}^{\mathrm{O}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ in one of the 30 starts, while the *j*th element of **β** is initialized as $\hat{\beta}_j + 0.5\left|\hat{\beta}_j\right| r$ in the other starts, with *r* drawn from the uniform

distribution defined for the real numbers within range [−1, 1] and $\hat{\beta}_j$ being the jth element of $\hat{\boldsymbol{\beta}}$. Thus, [CR1] in Section 2.1 must be rewritten with [CR2] unchanged as follows:

[CR1′] For each of $k = 1, \ldots p$, run the CSREG algorithm 30 times with $\boldsymbol{\beta}$ initialized as above. Among the resulting 30 $\boldsymbol{\beta}$'s, select $\boldsymbol{\beta}$ with the least $f(\boldsymbol{\beta})$ as $\hat{\boldsymbol{\beta}}_k^{\mathrm{C}}$ and obtain $\phi(\hat{\boldsymbol{\beta}}_k^{\mathrm{C}})$.

[CR2] Select $\hat{\boldsymbol{\beta}}^{\mathrm{C}} = \underset{k=0,1,\ldots,p}{\arg\min}\, \phi(\hat{\boldsymbol{\beta}}_k^{\mathrm{C}})$.

From here, we refer to the procedure of [CR1′] follwed by [CR2] simply as CSREG.

4.3. Details of the Lasso-Based Best Subset Selection

A coordinate descending method (e.g., Adachi, 2020, pp. 344-348) is used as the iterative lasso algorithm for (6), with $\boldsymbol{\beta}$ initialized as $\hat{\boldsymbol{\beta}}^{\mathrm{O}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. The candidates for λ are formed as explained next.

The proportion of penalty term $\lambda\|\boldsymbol{\beta}\|^1$ in the lasso loss function, whose $\boldsymbol{\beta}$ is set to $\hat{\boldsymbol{\beta}}^{\mathrm{O}}$, is expressed as

$$\mathrm{PP}(\lambda) = \frac{\lambda\left\|\hat{\boldsymbol{\beta}}^{\mathrm{O}}\right\|_1}{h(\hat{\boldsymbol{\beta}}^{\mathrm{O}})} = \frac{\lambda\left\|\hat{\boldsymbol{\beta}}^{\mathrm{O}}\right\|_1}{\left\|\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}^{\mathrm{O}}\right\|^2 + \lambda\left\|\hat{\boldsymbol{\beta}}^{\mathrm{O}}\right\|_1}$$

with $0 \le \mathrm{PP}(\lambda) \le 1$. Let us consider lasso (6) with λ satisfying $\mathrm{PP}(\lambda) = l/10000$ for $l = 1, 2, \ldots, 9999$. We can find such λ to be expressed as

$$\lambda(l) = \frac{\left\|\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}^{\mathrm{O}}\right\|^2}{\left\|\hat{\boldsymbol{\beta}}^{\mathrm{O}}\right\|_1} \times \frac{l/10000}{1-l/10000} \tag{10}$$

by rewriting $\mathrm{PP}(\lambda) = l/10000$. When $l = 1$ and $\mathrm{PP}(\lambda) = 1/10000$, lasso is expected to give $\hat{\boldsymbol{\beta}}_\lambda^{\mathrm{L}}$ with $\mathrm{Card}(\hat{\boldsymbol{\beta}}_\lambda^{\mathrm{L}})$ close to the upper limit p. In contrast, when $l = 9999$ and $\mathrm{PP}(\lambda) = 9999/10000$, lasso is expected to give $\hat{\boldsymbol{\beta}}_\lambda^{\mathrm{L}}$ with $\mathrm{Card}(\hat{\boldsymbol{\beta}}_\lambda^{\mathrm{L}})$ close to 1 or 0. Thus, (10) for l = 1, 2, ... , 9999 leads to a wide range of $\mathrm{Card}(\hat{\boldsymbol{\beta}}_\lambda^{\mathrm{L}})$. Further, $\mathrm{Card}(\hat{\boldsymbol{\beta}}_\lambda^{\mathrm{L}})$ decreases monotonically with the increase in l. Thus, [LA1] in Section 2.2 is modified into the procedure consisitng of the following steps:

Step 1. Initialize l as 1.

Step 2. Increase l by one. Run the lasso algorithm with λ set to (10) and obtain $\phi(\hat{\boldsymbol{\beta}}_\lambda^{\mathrm{L}})$.

Step 3. If $\mathrm{Card}(\hat{\boldsymbol{\beta}}_\lambda^{\mathrm{L}}) = 1$ or $\mathrm{PP}(\lambda) = 9999/10000$ finish; otherwise, go back to Step 2.

The above procedure allows the set $\boldsymbol{\Lambda}$ containing the candidates λ to be expressed as $\boldsymbol{\Lambda} = \{\lambda(l); l = 1, \ldots, L\}$. Here, $L = 9999$ if Step 2 does not lead to $\mathrm{Card}(\hat{\boldsymbol{\beta}}_\lambda^{\mathrm{L}}) = 1$; otherwise L is the integer leading to $\mathrm{Card}(\hat{\boldsymbol{\beta}}_\lambda^{\mathrm{L}}) = 1$. Thus, [LA1] in Section 2.1 must be rewritten with [LA2] unchanged as follows:

[LA1′] Perform the procedure consisting of Steps 1−3.

[LA2] Select $\hat{\boldsymbol{\beta}}^{\mathrm{L}} = \underset{\lambda\in\boldsymbol{\Lambda}}{\arg\min}\, \phi(\hat{\boldsymbol{\beta}}_\lambda^{\mathrm{L}})$.

From here, we refer to the procedure of [LA1′] followed by [LA2] simply as lasso.

5. Real Data Illustration

For illustration, we use the 252 × 14 column-centered data matrix [**y**, **X**], whose raw data version is included in the file "bodyfat.txt" available at https://lib.stat.cmu.edu/datasets/bodyfat. The rows of [**y**, **X**] correspond to persons, and the raw data version of **y** contains the bodyfat percentages of the persons, while their ages and body circumference measurements are contained in the raw version of **X** (252 × 13). Those percentages and measurements are detailed in "bodyfat.txt".

Table 1. Results for bodyfat data. Blanks stand for zeros.

Dataset		Standardized			Unstandardized		
Procedure		APREG	CSREG	Lasso	APREG	CSREG	Lasso
age	β_1			0.091			0.001
weight	β_2	-0.477	-0.477		-0.136	-0.123	-0.123
height	β_3			-0.088			
neck	β_4			-0.101		-0.365	
chest	β_5						
abdomen	β_6	1.283	1.283	0.939	0.996	1.008	0.918
hip	β_7						
thigh	β_8						
knee	β_9						
ankle	β_{10}						
biceps	β_{11}						
forearm	β_{12}	0.114	0.114	0.072	0.473	0.527	
wrist	β_{13}	-0.168	-0.168	-0.170	-1.504	-1.245	
BIC		-301.55	-301.55	-288.73	-301.55	-298.79	-290.51
Comput. Time		0.250	0.984	0.000	0.250	13.953	0.016

Table 2. Results for bodyfat data with noise predictors. Blanks stand for zeros

Dataset		Standardized			Unstandardized		
Procedure		APREG	CSREG	Lasso	APREG	CSREG	Lasso
age	β_1			0.079			0.011
weight	β_2	-0.477	-0.477		-0.136		-0.083
height	β_3			-0.083			
neck	β_4			-0.054		-0.603	
chest	β_5						
abdomen	β_6	1.283	1.283	0.896	0.996	0.974	0.797
hip	β_7					-0.332	
thigh	β_8						
knee	β_9						
ankle	β_{10}						
biceps	β_{11}						
forearm	β_{12}	0.114	0.114	0.039	0.473	0.409	
wrist	β_{13}	-0.168	-0.168	-0.149	-1.504	-1.612	
noise 1	β_{14}						0.005
noise 2	β_{15}						0.007
noise 3	β_{16}						
noise 4	β_{17}						
noise 5	β_{18}						
noise 6	β_{19}						
noise 7	β_{20}						
noise 8	β_{21}			-0.021			0.025
noise 9	β_{22}						
noise 10	β_{23}						
noise 11	β_{24}						
noise 12	β_{25}						
noise 13	β_{26}						
BIC		-301.55	-301.55	-278.17	-301.55	-293.34	-273.40
Comput. Time		1622.234	8.719	0.016	1491.375	159.719	0.094

We performed APREG, CSREG, and lasso for [**y**, **X**] and its standardized version, whose results are presented in Table 1. We also performed those procedures for 252 × 27 matrix [**y**, **X**, **N**] and its standardized version, with 252 × 26 [**X**, **N**] treated as the predictor matrix for **y**. Here **N** (252 × 13) contains artificial noises independent of **y**; **N** is randomly generated as described in Appendix 2, with the variance of the elements in the *j*th column of **N** equaling the variance for the *j*th column of **X** and the columns of **N** being mutually correlated. The results for [**y**, **X**, **N**] are shown in Table 2.

First, let us note the computation times in the bottom rows of Tables 1−2. Table 1 shows that APREG is faster than CSREG with the former time 0.250 second. This clearly shows that APREG should be used for $p \leq 13$. However, Table 2 shows that APREG is far more time- consuming than CSREG for $p = 26$. It suggests that APREG is unusable for data with too many predictors. On the other hand, lasso is very fast and the fastest of the three procedures in every case

The left panels of Tables 1−2 show that the CSREG and APREG solutions for the standardized versions of data sets are identical, without noise variables selected. It demonstrates that CSREG for standardized data is useful for finding the best subset. However, the right panels of the tables show that the CSREG solutions for the unstandardized data sets differ from the APREG counterparts. This difference caused from that the CSREG solutions selected were local minimizers. This suggests that CSREG for unstandardized data is sensitive to local minima. Further, the computation times of CSREG for the unstandardized data are longer than those for the standardized data. However, the optimal solution (i.e., global minimizer) of CSREG for unstandardized data can be transformed from the counterpart for standardized data, as seen in Theorem 1. This property and the above results show that CSREG should be performed for standardized data.

Let us note the lasso solutions in Tables 1−2. lasso has failed to find the best subset for all cases. It suggests the inferiority of lasso to CSREG in best subset selection. Further, lasso has selected a noise variable for both standardized and unstandardized versions of the noise-contaminated data (Table 2). This also suggests the inferiority of lasso in selecting the true subset. Further, we can find that the variables selected by lasso differ between standardized and unstandardized versions. This demonstrates that lasso should be performed for standardized data, as recommended in the lasso literature (e.g., Hastie et al., 2015).

What the above results show for APREG, CSREG, and lasso can be summarized as follows:

[APREG] It is to be used when p is not greater than 13.

[CSREG] It is to be performed for standardized data, which can be easily transformed into the counterpart for unstandardized data.

[lasso]. It should be performed for standardized data, though the resulting solution may not show the best subset. However, lasso is very fast.

6. Simulation Study

In this section, we explore whether the results in the last section are found for most of the data sets in practice. It is not efficient to make such assessments with real data sets. We thus resort to a simulation study. After its procedures are described in Section 6.1, the results are reported and discussed: We first report computation times and the cardinality of estimated coefficients in Section 6.2. Then, in Sections 6.3−6.4, we describe and discuss the following two subjects, [1] how accurately the best subset is found by CSREG and lasso; [2] how well the coefficients in the true model are recovered by APREG, CSREG, and lasso. Here, it should be noticed that the true model differs from the model with the best subset found by APREG. In Sections 6.2−6.4, $\hat{\boldsymbol{\beta}}^{*}$ is used for $\boldsymbol{\beta}$ estimated in CSREG or lasso; $\hat{\boldsymbol{\beta}}^{*}$ refers to $\hat{\boldsymbol{\beta}}^{\mathrm{C}}$ resulting in CSREG or $\hat{\boldsymbol{\beta}}^{\mathrm{L}}$ in lasso, with SD an abbreviation for standard deviation.

6.1. Procedures

Let us express an error level and the cardinality of $\boldsymbol{\beta}$ as

$$\varepsilon = \frac{\| \mathbf{e} \|^2}{\| \mathbf{X}\boldsymbol{\beta} \|^2 + \| \mathbf{e} \|^2} \quad \text{and} \quad c = \mathrm{Card}(\boldsymbol{\beta}) , \tag{11}$$

respectively. For each of the combinations of nine ε values and nine c ones, we synthesize a 200 × 21 data matrix [$\mathbf{y}$, $\mathbf{X}$] from model (1) (with $p = 21 - 1 = 20$) through the following procedures.

[1] Let $\mathbf{1}_n$ be the $n \times 1$ vector of ones and $\mathbf{J} = \mathbf{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_{n'}$. Each row of $\mathbf{X}_0$ ($n \times p$), which leads to $\mathbf{X} = \mathbf{J}\mathbf{X}_0$, is sampled from $N_p(\mathbf{0}_p, \mathbf{R})$, i.e., the p-variate normal distribution whose mean is $\mathbf{0}_p$ and covariance matrix is $\mathbf{R} = \boldsymbol{\Psi}^{-1/2}\mathbf{V}\boldsymbol{\Delta}\mathbf{V}'\boldsymbol{\Psi}^{-1/2}$. Here, $\mathbf{V}$ is a $p \times p$ orthonormal matrices selected randomly; $\boldsymbol{\Delta} = (\delta_{jk})$ is the $p \times p$ diagonal matrix with $\delta_{11} = 10$, $\delta_{pp} = 2$, and each of $\delta_{22}, \ldots, \delta_{p-1,\,p-1}$ drawn from $U(2, 10)$; $\boldsymbol{\Psi}$ is the $p \times p$ diagonal matrix whose diagonal elements are those of $\mathbf{V}\boldsymbol{\Delta}\mathbf{V}'(p \times p)$, i.e., $\mathbf{R}$ is a correlation matrix.

[2] Card($\boldsymbol{\beta}$) is set to each of the nine even numbers 2, 4, 6, … , 16, 18 and the value of a nonzero element of $\boldsymbol{\beta}$ is drawn from $U(0.1, 0.9)$ or U(−0.9, −0.1). Here, $U(a, b)$ denotes the uniform distribution defined for the real numbers ∈ [a, b].

[3] Each element of $\mathbf{e}_0$ ($n \times 1$), which leads to $\mathbf{e} = t\mathbf{J}\mathbf{e}_0$, is drawn from $N_1(0, 1)$, and $\mathbf{y}$ is synthesized with (1). Here, t is chosen so that the error level ε in (11) equals each of the nine values 0.1, 0.2, 0.3, … , 0.8, 0.9.

Procedures [1]−[3] provides the 81 = 9 (ε: error levels) × 9 (c: cardinality) data matrices [$\mathbf{y}$, $\mathbf{X}$] that are based on the identical $\mathbf{X}$ and $\mathbf{e}_0$. We replicate [1]−[3] 50 times to have a total of 4050 = 9 (cardinalities) × 9 (error levels) × 50 (replications) [$\mathbf{y}$, $\mathbf{X}$], where $\mathbf{y}$ and $\mathbf{X}$ vary randomly across the replications.

APREG, CSREG, and lasso were performed for the standardized versions of the 4050 [$\mathbf{y}$, $\mathbf{X}$]

for the reasons described in Section 5. Thus, the true coefficient vector, which the estimated counterpart should approximate, expressed as $\boldsymbol{\beta}_{\text{true}} = s^{-1}\mathbf{D}_s\boldsymbol{\beta}$. Here, $s = (n^{-1}\mathbf{y}'\mathbf{J}\mathbf{y})^{1/2}$ and $\mathbf{D}_s = \text{diag}(n^{-1}\mathbf{X}'\mathbf{J}\mathbf{X})^{1/2}$, with $\boldsymbol{\beta}$, $\mathbf{y}$, and $\mathbf{X}$ being those generated in the above [1]−[3] and diag($\mathbf{S}$) denoting the diagonal matrix whose diagonal elements are those of square $\mathbf{S}$. The procedures were run by FORTARN codes

Table 3. Average (Ave), SD, and $P = N_< / N_{\neq}$ of the cardinality proportions (CP) for each of error (ε) and true cardinality (c) levels. Here, $N_<$ and $N_{\neq}$ stand for how often $\text{CP}_\text{C} < \text{CP}_\text{L}$ and $\text{CP}_\text{C} \neq \text{CP}_\text{L}$ occurred, respectively, with CP_C and CP_L the CP values in CSREG and lasso, respectively.

Level	CSREG		Lasso		CSREG < Lasso			
	Ave	SD	Ave	SD	$N_<$	$N_{\neq}$		P
ε=.1	**.302**	.316	**.354**	.361	350	352	=	**.994**
ε=.2	**.295**	.306	**.353**	.360	355	360	=	**.986**
ε=.3	**.284**	.291	**.349**	.356	365	368	=	**.992**
ε=.4	**.273**	.278	**.344**	.351	369	374	=	**.987**
ε=.5	**.257**	.258	**.336**	.343	378	383	=	**.987**
ε=.6	**.239**	.238	**.320**	.327	367	373	=	**.984**
ε=.7	**.215**	.212	**.299**	.306	352	360	=	**.978**
ε=.8	**.176**	.169	**.248**	.266	323	362	=	**.892**
ε=.9	**.108**	.103	**.087**	.130	157	358	=	**.439**
c = 2	**.065**	.056	**.083**	.077	196	221	=	**.887**
c = 4	**.116**	.098	**.149**	.135	274	288	=	**.951**
c = 6	**.168**	.144	**.207**	.190	308	329	=	**.936**
c = 8	**.209**	.181	**.254**	.231	333	355	=	**.938**
c=10	**.247**	.219	**.321**	.298	382	414	=	**.923**
c=12	**.281**	.253	**.363**	.341	384	417	=	**.921**
c=14	**.322**	.292	**.402**	.377	381	420	=	**.907**
c=16	**.362**	.331	**.440**	.418	384	427	=	**.899**
c=18	**.382**	.355	**.472**	.442	374	419	=	**.893**
Total	**.239**	.257	**.299**	.329	3016	3290	=	**.917**

6.2. Computation times and the Cardinality of Estimated Coefficients

The computation time was measured by the second for each data set. Its average (with SD) over the 4050 data sets were 9.0517 (7.7842) for APREG, 0.0248 (0.0278) for CSREG, and 0.0007 (0.0039) for lasso. This shows that APREG is far more time-consuming than CSREG and lasso, with lasso faster than CSREG.

We refer to $\text{Card}(\hat{\boldsymbol{\beta}}^*)/p$ as the cardinality proportions (CP). The left part of Table 3 shows the average and SD of CP for each of the error levels (ε) and cardinality (c). On the other hand, the right part of Table 3 shows the proportion $P = N_< / N_{\neq}$. Here, $N_<$ and $N_{\neq}$ stand for the number of the cases with $\text{CP}_\text{C} < \text{CP}_\text{L}$ and that of the cases with $\text{CP}_\text{C} \neq \text{CP}_\text{L}$, respectively, with CP_C and CP_L the CP values in CSREG and lasso, respectively.

The bottom row labeled “total” in Table 3 shows that the CP values in CSREG are smaller than the lasso counterparts, i.e., CSREG tends for provide more sparse coefficients. But the reverse result is seen only for ε=.9. i.e., the greatest error level. Then, lasso provided more sparse coefficients. This reverse result will be referred to in Section 6.5. The reverse is not very surprising, in that the CP_L is found to decrease with the increases in ε. This observation suggests that it remains to study CP for the ε values between 0.8 and 0.9.

6.4. Deviations from the Best Subset

Let $\boldsymbol{\beta}_{\text{best}}$ refer to $\hat{\boldsymbol{\beta}}$ resulting in APREG, as the model with the best subset is found by APREG.

For each of $\hat{\boldsymbol{\beta}}^*$, how it deviates from $\boldsymbol{\beta}_{\text{best}}$ can be indicated by the mean absolute error (MAE), the mis-removal (MisR) rate (i.e the proportion of mistakenly removed predictors), and the mis-selection (MisS) rate (i.e., the proportion of mis-selected predictors). These are defined as

$$\text{MAE}(\hat{\boldsymbol{\beta}}^*, \boldsymbol{\beta}_{\text{best}}) = \frac{1}{p} \| \hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}_{\text{best}} \|_1 , \tag{11}$$

$$\text{MisR}(\hat{\boldsymbol{\beta}}^*, \boldsymbol{\beta}_{\text{best}}) = \frac{1}{p} \sum_{j=1}^{p} I\{(\boldsymbol{\beta}_{\text{best}})_j \neq 0 , (\hat{\boldsymbol{\beta}}^*)_j = 0\} \quad , \tag{12}$$

$$\text{MisS}(\hat{\boldsymbol{\beta}}^*, \boldsymbol{\beta}_{\text{best}}) = \frac{1}{p} \times \sum_{j=1}^{p} I\{(\boldsymbol{\beta}_{\text{best}})_j = 0 , (\hat{\boldsymbol{\beta}}^*)_j \neq 0\} \tag{13}$$

Here, $(\boldsymbol{\beta})_j$ denotes the jth element of $\boldsymbol{\beta}$, and $I\{\} = 1$ if the two formulas in {} hold, but $I\{\} = 0$ otherwise.

Table 4. Deviations from the best subset: Averages (Ave) and SD of MAE, MisR rate, and MisS rate for each of ε and c values, with the total Ave and SD over all cases.

	CSREG						**Lasso**					
Level	**MAE**		**MisR**		**MisS**		**MAE**		**MisR**		**MisS**	
	Ave	SD	Ave	SD	Ave	SD	Ave	SD	Ave	SD	Ave	SD
ε=.1	**.0000**	.0000	**.0000**	.0000	**.0000**	.0000	**.004**	.004	**.000**	.003	**.052**	.071
ε=.2	**.0000**	.0000	**.0000**	.0000	**.0000**	.0000	**.006**	.006	**.000**	.003	**.058**	.079
ε=.3	**.0000**	.0008	**.0001**	.0024	**.0002**	.0047	**.008**	.007	**.000**	.004	**.065**	.086
ε=.4	**.0001**	.0009	**.0002**	.0033	**.0002**	.0033	**.010**	.009	**.000**	.005	**.071**	.093
ε=.5	**.0002**	.0017	**.0003**	.0041	**.0006**	.0070	**.011**	.010	**.000**	.004	**.080**	.102
ε=.6	**.0003**	.0025	**.0007**	.0057	**.0008**	.0091	**.013**	.012	**.001**	.007	**.082**	.109
ε=.7	**.0004**	.0032	**.0013**	.0099	**.0010**	.0084	**.013**	.013	**.001**	.008	**.085**	.112
ε=.8	**.0007**	.0049	**.0016**	.0093	**.0030**	.0242	**.014**	.014	**.009**	.049	**.082**	.120
ε=.9	**.0012**	.0062	**.0040**	.0202	**.0047**	.0263	**.012**	.012	**.044**	.083	**.024**	.056
c = 2	**.0001**	.0012	**.0002**	.0047	**.0002**	.0047	**.004**	.004	**.002**	.011	**.020**	.038
c = 4	**.0000**	.0005	**.0001**	.0024	**.0001**	.0024	**.007**	.006	**.001**	.009	**.035**	.054
c = 6	**.0000**	.0009	**.0001**	.0024	**.0002**	.0047	**.009**	.008	**.004**	.021	**.042**	.067
c = 8	**.0000**	.0005	**.0002**	.0033	**.0000**	.0000	**.010**	.009	**.004**	.026	**.049**	.075
c=10	**.0001**	.0009	**.0004**	.0047	**.0002**	.0033	**.011**	.011	**.006**	.030	**.080**	.103
c=12	**.0004**	.0027	**.0016**	.0109	**.0013**	.0115	**.012**	.011	**.007**	.039	**.090**	.106
c=14	**.0006**	.0043	**.0018**	.0136	**.0026**	.0188	**.012**	.012	**.009**	.042	**.090**	.107
c=16	**.0005**	.0035	**.0012**	.0077	**.0020**	.0159	**.013**	.014	**.013**	.056	**.092**	.112
c=18	**.0010**	.0065	**.0026**	.0152	**.0038**	.0268	**.013**	.014	**.011**	.049	**.102**	.126
Total	**.0003**	.0031	**.0009**	.0086	**.0012**	.0130	**.010**	.011	**.006**	.035	**.067**	.096

Table 5. Proportions $N_< / N_{\neq} = P$. Here, $N_<$ and $N_{\neq}$ stand for how often $V_C < V_L$ and $V_C \neq V_L$ occurred, respectively, with V_C and V_L the values in CSREG and lasso, respectively.

Level	MAE			MisR			MisS		
	$N_<$ / $N_{\neq}$	=	P	$N_<$ / $N_{\neq}$	=	P	$N_<$ / $N_{\neq}$	=	P
ε=.1	449 / 450	=	**.998**	3 / 4	=	**.750**	349 / 349	=	**1.000**
ε=.2	450 / 450	=	**1.000**	5 / 5	=	**1.000**	355 / 355	=	**1.000**
ε=.3	449 / 450	=	**.998**	5 / 7	=	**.714**	364 / 364	=	**1.000**
ε=.4	449 / 450	=	**.998**	6 / 9	=	**.667**	370 / 370	=	**1.000**
ε=.5	450 / 450	=	**1.000**	7 / 13	=	**.538**	379 / 379	=	**1.000**
ε=.6	448 / 450	=	**.996**	10 / 21	=	**.476**	368 / 368	=	**1.000**
ε=.7	448 / 450	=	**.996**	14 / 31	=	**.452**	358 / 358	=	**1.000**
ε=.8	443 / 450	=	**.984**	48 / 71	=	**.676**	331 / 333	=	**.994**
ε=.9	439 / 450	=	**.976**	202 / 208	=	**.971**	167 / 187	=	**.893**
c = 2	449 / 450	=	**.998**	32 / 32	=	**1.000**	199 / 200	=	**.995**
c = 4	450 / 450	=	**1.000**	14 / 16	=	**.875**	274 / 274	=	**1.000**
c = 6	450 / 450	=	**1.000**	26 / 26	=	**1.000**	312 / 313	=	**.997**
c = 8	450 / 450	=	**1.000**	24 / 24	=	**1.000**	335 / 335	=	**1.000**
c=10	449 / 450	=	**.998**	36 / 39	=	**.923**	389 / 389	=	**1.000**
c=12	448 / 450	=	**.996**	34 / 42	=	**.810**	387 / 389	=	**.995**
c=14	445 / 450	=	**.989**	41 / 55	=	**.745**	383 / 391	=	**.980**
c=16	444 / 450	=	**.987**	44 / 63	=	**.698**	385 / 388	=	**.992**
c=18	440 / 450	=	**.978**	49 / 72	=	**.681**	377 / 384	=	**.982**
Total	4025 / 4050	=	**.994**	300 / 369	=	**.813**	3041 / 3063	=	**.993**

Table 4 shows the averages and SD of (11)−(13) over the solutions for each of error levels (ε) and true cardinality (c), with those over all solutions presented in the bottom row. Table 5 shows the proportions $P = N_</N_{\neq}$ of the cases with $V_C < V_L$. Here, V_C stands for the CS-REG value of (11), (12), or (13) and V_L stands for the lasso counterpart, $N_<$ is the number of the cases with $V_C < V_L$, and $N_{\neq}$ is the number of the cases with $V_C \neq V_L$. In the remainder, we compare the CSREG and lasso to discuss their differences in each of (11)−(13), with referring to Tables 4−5.

6.3.1. Mean Absolute Error

As described in Section 2, CSREG provides the best subset with MAE being zero, if the CSREG algorithm gives the global minimizers. Table 4 shows that the resulting MAE are not zero but close to zero, except for some values of ε and c. This implies that the CSREG algorithm gave local minimizers but are not sensitive to local minima. We can also find that the sensitivity to local minima tends to increase with ε and c values being larger.

Comparing the results in Tables 4 and 5 between CSREG and lasso, we find that MAE resulting in CSREG were far less than the lasso counterpart, and the former value almost always were less than the latter. However, the overall average of MAE for lasso in Table 4 is not very large. This allows us to consider that the deviation of the lasso solutions from the APREG ones are not serious.

6.3.2. Mis-Removal and Mis-Selection Rates

In Tables 2 and 3, both MisR and MisS rates in CSREG are found to be lower than the lasso counterparts. It is to be noted in Table 5 that MisS resulted more frequently in lasso than in CSREG. This allows us to consider that the deviation of the lasso solutions from the best subset follows from mis-selection rather than mis-removal.

6.4. Deviation from the True Subset

The best model, whose coefficient vector is denoted $\boldsymbol{\beta}_{best}$ in the last subsection, differs from the true model, whose coefficient vector $\boldsymbol{\beta}$ is synthesized as in Section 5.1. Let $\boldsymbol{\beta}_{true}$ denote the (true) $\boldsymbol{\beta}$ synthesized as in Section 5.1 and $\hat{\boldsymbol{\beta}}$ denote the estimate of $\boldsymbol{\beta}$ obtained in APREG, CSREG or lasso. In this section, we study the deviation of $\hat{\boldsymbol{\beta}}$ from $\boldsymbol{\beta}_{true}$. This study might be tentative, in that we focused on finding the best model so far, thus the true subsets were beyond the scope of our discussion. However, we can assess the deviation and should study it.

We obtained MAE($\hat{\boldsymbol{\beta}}^*$, $\boldsymbol{\beta}_{true}$), MisR($\hat{\boldsymbol{\beta}}^*$, $\boldsymbol{\beta}_{true}$), and MisS($\hat{\boldsymbol{\beta}}^*$, $\boldsymbol{\beta}_{true}$), i.e., the versions of (11)−(13) with $\boldsymbol{\beta}_{best}$ replaced by $\boldsymbol{\beta}_{true}$. The statistics of those versions are presented in Tables 6 and 7, which have been obtained by replacing $\boldsymbol{\beta}_{best}$ by $\boldsymbol{\beta}_{true}$ from Tables 4 and 5, respectively.

As seen in Tables 4−5, the CSREG and APREG solutions were almost equivalent. Thus, the performances of APREG for exploring the true subset were almost equivalent to the CSREG counterparts. From this equivalence, the results of APREG are not shown in Tables 6−7.

6.4.1. Mean Absolute Error

Tables 6 and 7 shows that the total average of MAE for CSREG is a little smaller than the lasso counterpart, but the difference cannot be considered as substantial. Inspections of the details in Tables 6 and 7 show the following facts: The averages of MAE for CSREG are smaller than the lasso counterparts for $\varepsilon \leq .8$ and $c \leq 14$, but the average for CSREG is greater than the lasso counterpart for $\varepsilon = 0.9$ and c=18. That is, $\boldsymbol{\beta}_{true}$ was recovered better in lasso than in CSREG for ε and c being the greatest values.

Table 6. Deviations from the true parameters: Averages (Ave) and SD of MAE, MisR rate, and MisS rate for each of ε and c values, with the total Ave and SD over all cases.

	CSREG						**Lasso**					
Level	**MAE**		**MisR**		**MisS**		**MAE**		**MisR**		**MisS**	
	Ave	SD	Ave	SD	Ave	SD	Ave	SD	Ave	SD	Ave	SD
ε=.1	**.004**	.004	**.001**	.006	**.003**	.013	**.006**	.006	**.000**	.000	**.054**	.073
ε=.2	**.006**	.007	**.008**	.023	**.003**	.013	**.008**	.008	**.001**	.006	**.054**	.073
ε=.3	**.008**	.009	**.019**	.040	**.003**	.013	**.010**	.010	**.004**	.015	**.053**	.073
ε=.4	**.009**	.011	**.031**	.057	**.004**	.014	**.012**	.012	**.007**	.020	**.051**	.072
ε=.5	**.011**	.014	**.047**	.077	**.004**	.014	**.013**	.013	**.013**	.031	**.049**	.070
ε=.6	**.013**	.015	**.065**	.098	**.004**	.014	**.015**	.015	**.024**	.047	**.045**	.068
ε=.7	**.015**	.017	**.088**	.127	**.004**	.013	**.016**	.016	**.040**	.070	**.040**	.064
ε=.8	**.017**	.020	**.127**	.171	**.004**	.013	**.018**	.020	**.084**	.142	**.032**	.055
ε=.9	**.019**	.021	**.196**	.241	**.005**	.016	**.021**	.022	**.224**	.293	**.011**	.029
c = 2	**.002**	.003	**.001**	.007	**.006**	.017	**.004**	.004	**.001**	.005	**.024**	.042
c = 4	**.004**	.005	**.008**	.021	**.003**	.016	**.007**	.007	**.007**	.022	**.036**	.057
c = 6	**.006**	.007	**.020**	.044	**.008**	.020	**.011**	.011	**.019**	.051	**.045**	.069
c = 8	**.009**	.009	**.035**	.066	**.004**	.014	**.012**	.012	**.026**	.070	**.040**	.069
c=10	**.011**	.012	**.056**	.096	**.003**	.011	**.014**	.014	**.041**	.100	**.063**	.088
c=12	**.013**	.014	**.080**	.125	**.001**	.008	**.017**	.016	**.059**	.132	**.062**	.084
c=14	**.016**	.017	**.100**	.151	**.002**	.010	**.017**	.017	**.069**	.164	**.050**	.066
c=16	**.018**	.019	**.120**	.178	**.002**	.009	**.018**	.019	**.082**	.197	**.042**	.060
c=18	**.023**	.023	**.162**	.216	**.003**	.013	**.019**	.019	**.095**	.215	**.027**	.038
Total	**.011**	.015	**.065**	.132	**.004**	.014	**.013**	.015	**.044**	.132	**.043**	.067

Table 7. Proportions $N_< / N_{\neq} = P$. Here, $N_<$ and $N_{\neq}$ stand for how often $V_C < V_L$ and $V_C \neq V_L$ occurred, respectively, with V_C and V_L the values in CSREG and lasso, respectively.

Level	MAE				MisR				MisS			
	$N_<$	$N_{\neq}$		P	$N_<$	$N_{\neq}$		P	$N_<$	$N_{\neq}$		P
ε=.1	416	447	=	**.931**	0	14	=	**.000**	347	**349**	=	.994
ε=.2	396	449	=	**.882**	0	82	=	**.000**	342	**347**	=	.986
ε=.3	373	449	=	**.831**	0	156	=	**.000**	338	**342**	=	.988
ε=.4	358	450	=	**.796**	1	186	=	**.005**	328	**333**	=	.985
ε=.5	325	450	=	**.722**	3	236	=	**.013**	323	**326**	=	.991
ε=.6	310	450	=	**.689**	3	254	=	**.012**	304	**309**	=	.984
ε=.7	295	450	=	**.656**	7	264	=	**.027**	268	**271**	=	.989
ε=.8	284	447	=	**.635**	35	293	=	**.119**	220	**228**	=	.965
ε=.9	308	449	=	**.686**	199	314	=	**.634**	99	**123**	=	.805
c = 2	405	445	=	**.910**	0	7	=	**.000**	193	**219**	=	.881
c = 4	437	450	=	**.971**	9	38	=	**.237**	267	**275**	=	.971
c = 6	391	449	=	**.871**	22	82	=	**.268**	294	**299**	=	.983
c = 8	396	450	=	**.880**	23	166	=	**.139**	266	**268**	=	.993
c=10	377	450	=	**.838**	34	226	=	**.150**	343	**346**	=	.991
c=12	360	449	=	**.802**	33	273	=	**.121**	325	**328**	=	.991
c=14	313	450	=	**.696**	39	296	=	**.132**	334	**339**	=	.985
c=16	225	448	=	**.502**	43	347	=	**.124**	302	**305**	=	.990
c=18	161	450	=	**.358**	45	364	=	**.124**	245	**249**	=	.984
Total	3065	4041	=	**.758**	248	1799	=	**.138**	2569	**2628**	=	.978

6.4.2. Mis-Removal and Mis-Selection Rates

The bottom rows in Tables 6 and 7 show the following results: [1] The overall average of the MisR rates for CSREG is larger than the lasso counterpart, which shows that CSREG tends to remove the predictors included in the true subsets, as compared to lasso. [2] The overall average of the MisS rates for CSREG is smaller than the lasso counterpart, which shows that lasso tends to fail the selection of the predictors included in the true subsets, as compared to CSREG.

It is to be noted in Tables 6 and 7 that the average of the MisS rates for CSERG rather decreases with the increases in ε. This shows that the mis-selection of predictors in CSREG less likely occurs with the increases in error levels. We can consider that this result is related to the averages of cardinality proportions (CP) in Table 3; it shows that the average of CP for lasso decreases with the increases in ε. That is, the number of the predictors selected in lasso decreases with the increases in error levels, which lead to the decrease in the mis-selection.

7. Discussion

Cardinality-specified regression (CSREG) has not been treated in the studies for best subset selection. In this paper, we considered using CSREG or lasso to find the best subset of predictors, which can be found by all possible regression (APREG).

First, we discussed the following advantages of CSREG/lasso: CSREG can provide the best subset of predictors and the estimates of their coefficients resulting in APREG, but lasso cannot provide the APREG estimates of coefficients. But lasso does not provide a local minimizer, which can be given by CSREG. We also proved that CSREG is scale-invariant, though lasso is not so: The CSREG solution for a standardized data set can be transformed into the solution for the unstandardized counterpart, as is the solution of ordinary regression. But such a transformation is not possible in lasso.

Then, we presented the procedures of the CSREG- and lasso-based best subset selections. In CSREG, its algorithm is run 30 times, in order to reduce the possibility of selecting a local minimizer as the optimal one. Among the resulting 30 solutions, the one with the lowest loss function value is selected as the optimal solution. In lasso, penalty weight λ is set to at most 9,999 values, so that the resulting cardinalities of $\boldsymbol{\beta}$ are as various integers as possible. In all procedures, BIC is used for selecting the model with the best subset.

Using the above procedure, we compared CSREG and lasso numerically. The real data examples in Section 5 suggest the following: [1] CSREG for unstandardized data may be more sensitive to local minima and time-consuming than CSREG for the standardized counterparts, but the former CSREG solutions can be obtained from the latter CSREG solutions, thanks to the scale-invariance of CSREG; [2] lasso may tend to select the variables irrelevant to a predicted variable as its predictors. The results of the simulation study in Sub-section 6.3 showed that [1] CSREG can almost always find the best subset; [2] lasso is more inaccurate than CSREG in the best subset selection and tends to miss the predictors included in the best subset, but the lasso inaccuracy cannot be considered serious; [3] CSREG can also find the true subset more accurately than lasso; [4] lasso is much faster than CSREG.

From the above discussions, we can conclude as follows: Obviously, APREG is to be performed, if the number of the predictors in the full set is small enough; otherwise, CSRER is to be run for the standardized data sets. The resulting solutions can be transformed into the counterparts for unstandardized ones. However, lasso should be used if the subsets approximating the best ones are wished to be found very quickly.

Additionally, we compared CSREG and lasso in the recovery of the true coefficients in Section 6.4. This is beyond the problem of variable selection that is the subject of this paper. However, the comparison gave the results to be further studied. Among the results, to be noted is that lasso outperformed CSREG only in the condition of the highest error level with $\varepsilon = .9$, i.e., when the ninety percents of the variations in a predicted variable are occupied by errors. This result suggests that lasso may be useful particularly in social sciences for the following reasons: We can consider that the variables observed in social sciences are underlain by a number of factors which cannot be included in predictors.

Appendix 1. Algorithm for CSREG

The CSREG algorithm was derived through orthogonalizing predictors in Xiong (2014), but derived through Kiers' (1990, 2002) majorization algorithm in Adachi and Kiers (2017). Here, the CSREG algorithm is introduced based on the latter derivation.

Using $s_y = n^{-1}\mathbf{y}'\mathbf{y}$, $\mathbf{s}_{Xy} = n^{-1}\mathbf{X}'\mathbf{y}$, and $\mathbf{S}_{XX} = n^{-1}\mathbf{X}'\mathbf{X}$, the CSREG loss function divided by n, i.e., $f^{\#}(\boldsymbol{\beta}) = n^{-1}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|$, can be rewritten as $f^{\#}(\boldsymbol{\beta}) = s_y - 2\mathbf{s}_{XY}'\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbf{S}_{XX}\boldsymbol{\beta}$. Using $\boldsymbol{\beta}_c$ for the current

value of $\boldsymbol{\beta}$ and $\mathbf{d} = \boldsymbol{\beta} - \boldsymbol{\beta}_c$ or $\boldsymbol{\beta} = \boldsymbol{\beta}_c + \mathbf{d}$, further $f^{\#}(\boldsymbol{\beta})$ can be rewritten as

$$f^{\#}(\boldsymbol{\beta}) = f^{\#}(\boldsymbol{\beta}_c + \mathbf{d}) = f^{\#}(\boldsymbol{\beta}_c) + 2(\boldsymbol{\beta}_c'\mathbf{S}_{xx} - \mathbf{s}_{xy}')\mathbf{d} + \mathbf{d}'\mathbf{S}_{xx}\mathbf{d} \,. \tag{A.1}$$

Now, let us use α for the largest eigenvalue of $\mathbf{S}_{XX}$ and substitute $\alpha\mathbf{d}'\mathbf{d}$ for $\mathbf{d}'\mathbf{S}_{xx}\mathbf{d}$ in (A.1) to have function

$$m(\boldsymbol{\beta}) = f^{\#}(\boldsymbol{\beta}_c) + 2(\boldsymbol{\beta}_c'\mathbf{S}_{xx} - \mathbf{s}_{xy}')\mathbf{d} + \alpha\mathbf{d}'\mathbf{d} = f^{\#}(\boldsymbol{\beta}_c) + \alpha(\|\mathbf{d} - \mathbf{h}\|^2 - \|\mathbf{h}\|^2) \tag{A.2}$$

with $\mathbf{h} = -\alpha^{-1}(\mathbf{S}_{xx}\boldsymbol{\beta}_c - \mathbf{s}_{xy})$. By comparing (A.1) and (A.2), we can find that $\mathbf{d} = 0$ leads to $f^{\#}(\boldsymbol{\beta}_c) = m(\boldsymbol{\beta}_c)$ and the Rayleigh quotient $\alpha \geq \mathbf{d}'\mathbf{S}_{xx}\mathbf{d}/\mathbf{d}'\mathbf{d}$ leads to $m(\boldsymbol{\beta}) \geq f^{\#}(\boldsymbol{\beta})$. Thus, we have $f^{\#}(\boldsymbol{\beta}_c) = m(\boldsymbol{\beta}_c) \geq \min_{\beta} m(\boldsymbol{\beta}) \geq f^{\#}(\boldsymbol{\beta})$, which the update of $\boldsymbol{\beta}_c$ to $\boldsymbol{\beta}$ satisfying $f^{\#}(\boldsymbol{\beta}_c) \geq f^{\#}(\boldsymbol{\beta})$ can be attained the minimization of (A.2) over $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_p]'$.

Since only $\|\mathbf{d} - \mathbf{h}\|^2$ is a function of (A.2), the minimization of (A.2) amounts to that of

$$m^{\#}(\boldsymbol{\beta}) = \|\mathbf{d} - \mathbf{h}\|^2 = \|\boldsymbol{\beta} - \boldsymbol{\beta}_c + \alpha^{-1}(\mathbf{S}_{xx}\boldsymbol{\beta}_c - \mathbf{s}_{xy})\|^2 = \|\boldsymbol{\beta} - \mathbf{b}\|^2 = \sum_{j=1}^{p}(\beta_j - b_j)^2 \tag{A.3}$$

with $\mathbf{b} = [b_1, \ldots, b_p]' = \boldsymbol{\beta}_c\lambda^{-1}(\mathbf{S}_{xx}\boldsymbol{\beta}_c + \mathbf{s}_{xy})$. Using $\mathbf{O}$ for the set containing the subscripts of β_j to be 0, (A.3) can be rewritten as

$$m^{\#}(\boldsymbol{\beta}) = \sum_{j\in\mathbf{O}}(\beta_j - b_j)^2 + \sum_{j\notin\mathbf{O}}(\beta_j - b_j)^2 = \sum_{j\in\mathbf{O}} b_j^2 + \sum_{j\notin\mathbf{O}}(\beta_j - b_j)^2 \geq \sum_{j\in\mathbf{O}} b_j^2 \quad . \tag{A.4}$$

This shows that (A.3) is minimized subject to Card($\boldsymbol{\beta}$) = k for

$$\beta_j = \begin{cases} b_j & \text{if } |b_j| \geq |b_{[k]}| \\ 0 & \text{otherwise} \end{cases} \tag{A.5}$$

with $|b_{[k]}|$ the kth largest one among $|b_1|, \ldots, |b_p|$, since $m^{\#}(\boldsymbol{\beta})$ attains its lower limit $\Sigma_{j\in\mathbf{O}} b_j^2$ and this is minimal for (A.5).

Thus, the CSREG algorithm is listed as follows:

Step 1. Initialize $\boldsymbol{\beta}$

Step 2. Update the elements of $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_p]'$ as (A.5).

Step 3. Finish if convergence is reached; otherwise, go back to Step 2.

The convergence in Step 3 is defined as the decrease of the (A.1) value being less than 0.1^{12} in this paper.

Appendix 2. Generation of Noise Predictors

Let $n = 235$, $p = 13$, $\mathbf{I}_n$ be the $n \times n$ identity matrix, and $\mathbf{1}_n$ be the $n \times 1$ vector of ones. We set $\mathbf{N}$ ($n \times p$) to $(\mathbf{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n')\mathbf{N}_0\mathbf{D}_v$. Here, $\mathbf{D}_v$ ($p \times p$) is the diagonal matrix whose jth diagonal element is the variance of the elements in the jth column of $\mathbf{X}$, and each row of $\mathbf{N}_0$ ($n \times p$) is sampled from the p-variate normal distribution, whose mean is $\mathbf{0}_p$ and covariance matrix is $\mathbf{V}\boldsymbol{\Delta}\mathbf{V}'$ with $\mathbf{V}$ and $\boldsymbol{\Delta}$ defined as follows: $\mathbf{V}$ ($p \times p$) is an orthonormal matrices selected randomly, and $\boldsymbol{\Delta}$ ($p \times p$) is the diagonal matrix, whose first diagonal element is 10, final one is 2, and other diagonal elements are drawn from the uniform distribution defined for the real numbers $\in$ [2, 10].

References

Adachi, K. (2020). *Matrix-based introduction to multivariate data analysis* (*2nd ed*.). Singapore: Springer.

Adachi, K. & Kiers, H. A. L. (2017). Sparse regression without using a penalty function. *Proceedings of the Japanese Joint Statistical Meeting 2017*, p. 123.

Fan, J., & Li, R. (2011). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **96**, 1348-1860.

Fahrmeir, L., Kneib, T., Lang, S., & Marx, B. D. (2021). *Regression: Models, methods, and applications* (Second Edition). Berlin: Springer.

Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical learning with sparsity: The lasso and generalizations*. Boca Raton, FL: CRC Press/Taylor & Francis Group.

Kiers, H. A. L. (1990). Majorization as a tool for optimizing a class of matrix functions. *Psychometrika*, **55**, 417-428.

Kiers, H. A. L. (2002). Setting up alternating least squares and iterative majorization algorithms for solving various matrix optimization problems. *Computational Statistics and Data Analysis*, **41**, 157-170.

Konishi, S., & Kitagawa, G. (2007). *Information criteria and statistical modeling*. New York: Springer.

Mallows, C. L. (1973). Some comments on Cp. *Technometrics*, **15**, 661-675.

Mazumuder, R., Friedman, J. H., & Hastie, T. (2011). SparseNet: Coordinate descent with nonconvex penalties. *Journal of the American Statistical Association*, **106**, 1125-1138.

Sawa, T. (1979). *Regression analysis*. Tokyo: Asakura Shoten (in Japanese).

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

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

Xiong, S. (2014). Better subset regression. *Biometrika*, **101**, 71-84.

Received: 18 February 2026