Noname manuscript No.
(will be inserted by the editor)

Best subset selection via cross-validation criterion

Yuichi Takano · Ryuhei Miyashiro

Received: date / Accepted: date

Abstract This paper is concerned with the cross-validation criterion for best subset selection in a linear regression model. In contrast with the use of statistical criteria (e.g., Mallows' C_p, AIC, BIC, and various information criteria), the cross-validation only requires the mild assumptions, namely, samples are identically distributed, and training and validation samples are independent. For this reason, the cross-validation criterion is expected to work well in most situations for any predictive methods. The purpose of this paper is to establish a mixed-integer optimization (MIO) approach to selecting the best subset of explanatory variables via the cross-validation criterion. This subset selection problem can be formulated as a bilevel MIO problem. We then reduce it to a mixed-integer quadratic optimization problem, which can be solved exactly using optimization software. The efficacy of our method is evaluated through simulation experiments by comparison with statistical-criterion-based exhaustive search algorithms and the L_1-regularized regression. Simulation results demonstrate that our method delivered good performance in both the subset selection accuracy and the predictive performance when the signal-to-noise ratio was low.

Keywords Integer programming · Subset selection · Cross-validation · Ridge regression · Statistics

Yuichi Takano (corresponding author)
Faculty of Engineering, Information and Systems, University of Tsukuba, 1-1-1 Tennodai, Tsukuba-shi, Ibaraki 305-8577, Japan
Tel.: +81-29-853-5087
E-mail: ytakano@sk.tsukuba.ac.jp

Ryuhei Miyashiro
Institute of Engineering, Tokyo University of Agriculture and Technology, 2-24-16 Naka-cho, Koganei-shi, Tokyo 184-8588, Japan

1 Introduction

Subset selection, also known as variable/feature/attribute selection, involves selecting a significant subset of explanatory variables with which to construct a regression model [24]. It aids in understanding causality between explanatory and response variables. It also reduces data-gathering cost and the time required for estimating model parameters. Moreover, the predictive performance of a regression model can be improved because overfitting is mitigated by elimination of redundant explanatory variables.

Another approach to boosting the predictive performance is the shrinkage method, which includes ridge regression [18], lasso [41], and elastic net [44]. This approach shrinks regression coefficients of explanatory variables toward zero. The ridge regression has a theoretical advantage of dealing with multicollinearity, whereas the lasso has the subset selection property of setting unnecessary regression coefficients to exactly zero.

To assess the quality of a subset regression model, various statistical criteria are commonly used; these include the adjusted R^2 [43], Mallows' C_p [23], Akaike information criterion (AIC) [1], and Bayesian information criterion (BIC) [35]. It is known that C_p and AIC are derived from estimating the out-of-sample predictive performance, whereas BIC is aimed at identifying the "true model." However, since these statistical criteria depend on some strict assumptions, they are not suitable when such assumptions are violated.

This paper is focused on the cross-validation criterion [2,16,27,38] for best subset selection. Specifically, to evaluate the quality of a subset regression model, we split a set of given samples into training and validation sets; the training set is used for parameter estimation, and the prediction error is computed from the validation set. In contrast with the use of statistical criteria, the cross-validation only requires the mild assumptions, namely, samples are identically distributed, and training and validation samples are independent [4]. Consequently, the cross-validation criterion can be applied to any predictive methods, and it is expected to work well in most situations.

To accomplish best subset selection via the cross-validation criterion, we adopt the mixed-integer optimization (MIO) approach. One of these approaches was first proposed in the 1970s [3], and recently they have received renewed attention due to advances in optimization algorithms and computer performance [9,14,20,30,42]. Hastie et al. [17] reported that when the signal-to-noise ratio (SNR) was high, the MIO approach achieved superior predictive performance compared with the lasso. The MIO approaches have been proposed for best subset selection with respect to the adjusted R^2 [26], Mallows' C_p [25], and AIC/BIC [19,26]. Additionally, MIO-based subset selection has been applied to logit models [11,28,33,34], support vector machines [22], cluster analysis [6], classification trees [10], eliminating multicollinearity [8,39,40], and statistical tests/diagnostics [12].

The aim of this paper is to establish a computationally tractable MIO approach to selecting the best subset of explanatory variables via the cross-validation criterion for ridge regression. This subset selection problem can be posed as a bilevel MIO problem, but it is difficult to handle such a bilevel optimization problem. To remedy this situation, we transform the problem into a single-level mixed-integer quadratic optimization (MIQO) problem. We can exactly solve the resultant MIQO problem by means of optimization software. Algorithms for bilevel optimization [13,37] have been employed for hyperparameter tuning in support vector regression [7], support vector classification [21], general supervised learning [31], and nonsmooth regularization [29]. To the best of our knowledge, however, we are the first to develop an effective method for best subset selection via the cross-validation criterion.

The effectiveness of our method is assessed through simulation experiments following the previous studies [9,17]. We compare our method with statistical-criterion-based exhaustive search algorithms [24] and the L_1-regularized regression [41]. The simulation results demonstrate that when SNR was low, our method was superior in terms of the subset selection accuracy and the predictive performance.

2 Ridge Regression

Let us suppose that we are given n samples $(y_i; x_{i1}, x_{i2}, \dots, x_{ip})$ for $i = 1, 2, \dots, n$. Here, y_i is a response variable and x_{ij} is the jth explanatory variable for the ith sample. The index sets of explanatory variables and samples are denoted by $\mathcal{P} := \{1, 2, \dots, p\}$ and $\mathcal{N} := \{1, 2, \dots, n\}$, respectively.

We assume that all explanatory and response variables are centered such that

$$\sum_{i \in \mathcal{N}} y_i = 0, \quad \sum_{i \in \mathcal{N}} x_{ij} = 0 \quad (j \in \mathcal{P}).$$

The multiple linear regression model is then formulated as follows:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{a} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{y} := (y_i)_{i \in \mathcal{N}}$, $\boldsymbol{a} := (a_j)_{j \in \mathcal{P}}$, and $\boldsymbol{\varepsilon} := (\varepsilon_i)_{i \in \mathcal{N}}$ are all column vectors, and $\boldsymbol{X}$ is a matrix composed of explanatory variables,

$$\boldsymbol{X} := (\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_p) := (x_{ij})_{(i,j) \in \mathcal{N} \times \mathcal{P}}.$$

Here, $\boldsymbol{a}$ is a vector of regression coefficients to be estimated, and $\boldsymbol{\varepsilon}$ is a vector formed from prediction residuals.

We focus on the ridge regression for the multiple linear regression model. Specifically, we minimize the residual sum of squares (RSS) with the L_2-regularization term to shrink regression coefficients toward zero,

$$\underbrace{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{a}\|_2^2}_{\text{RSS}} + \underbrace{\lambda \|\boldsymbol{a}\|_2^2}_{\text{regularization}} = \boldsymbol{y}^\top \boldsymbol{y} - 2\boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{a} + \boldsymbol{a}^\top (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})\boldsymbol{a}, \tag{1}$$

where $\lambda \in \mathbb{R}_+$ is a regularization parameter. After partial differentiation, this is equivalent to solving a system of linear equations for $\hat{\boldsymbol{a}}$,

$$\left(\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}\right) \hat{\boldsymbol{a}} = \boldsymbol{X}^\top \boldsymbol{y}. \tag{2}$$

The solution $\hat{\boldsymbol{a}}$ is called the *ridge estimator*.

3 Cross-validation Criterion

Let us partition the index set $\mathcal{N}$ of samples into K subsets of (almost) the same size as follows:

$$\mathcal{N} = \bigcup_{k \in \mathcal{K}} \mathcal{N}_k, \quad \mathcal{N}_k \cap \mathcal{N}_{k'} = \emptyset \quad (k \neq k'), \quad |\mathcal{N}_k| \approx \frac{|\mathcal{N}|}{K} \quad (k \in \mathcal{K}),$$

where $\mathcal{K} := \{1, 2, \ldots, K\}$. For each $k \in \mathcal{K}$, we define the *training set* $\mathcal{T}_k$ and the *validation set* $\mathcal{V}_k$ as follows:

$$\mathcal{T}_k := \mathcal{N} \setminus \mathcal{N}_k, \quad \mathcal{V}_k := \mathcal{N}_k. \tag{3}$$

We also use the following notations to extract the parts of response and explanatory variables corresponding to subsets $\mathcal{M} \subseteq \mathcal{N}$ and $\mathcal{S} \subseteq \mathcal{P}$:

$$\begin{aligned} &\boldsymbol{y}(\mathcal{M}) := (y_i)_{i \in \mathcal{M}}, \quad \boldsymbol{x}_j(\mathcal{M}) := (x_{ij})_{i \in \mathcal{M}} \quad (j \in \mathcal{P}), \\ &\boldsymbol{X}(\mathcal{M}, \mathcal{S}) := (x_{ij})_{(i,j) \in \mathcal{M} \times \mathcal{S}}. \end{aligned}$$

We are now in a position to formulate the procedure of *K-fold cross-validation* for a ridge regression model. We begin by setting a value of the regularization parameter $\lambda \in \mathbb{R}_+$ and a subset $\mathcal{S} \subseteq \mathcal{P}$ of explanatory variables. In the training phase, we compute the ridge estimator for each $k \in \mathcal{K}$ from the kth training set as follows:

$$\hat{\boldsymbol{a}}_{\mathcal{S}}^{(k)} \in \arg\min\{\|\boldsymbol{y}(\mathcal{T}_k) - \boldsymbol{X}(\mathcal{T}_k, \mathcal{S})\boldsymbol{a}_{\mathcal{S}}\|_2^2 + \lambda \|\boldsymbol{a}_{\mathcal{S}}\|_2^2 \mid \boldsymbol{a}_{\mathcal{S}} \in \mathbb{R}^{|\mathcal{S}|}\}. \tag{4}$$

In the validation phase, we use the validation sets to compute the *cross-validation error* from the obtained ridge estimator as follows:

$$\sum_{k \in \mathcal{K}} \|\boldsymbol{y}(\mathcal{V}_k) - \boldsymbol{X}(\mathcal{V}_k, \mathcal{S})\hat{\boldsymbol{a}}_{\mathcal{S}}^{(k)}\|_2^2. \tag{5}$$

The *cross-validation criterion* involves selecting the best subset $\mathcal{S}$ of explanatory variables in terms of the cross-validation error (5). To accomplish this, however, we must repeatedly perform the cross-validation procedure for all possible subsets $\mathcal{S} \subseteq \mathcal{P}$.

4 Mixed-integer Optimization Formulations

This section presents our MIO formulations for best subset selection via the cross-validation criterion. Let $\boldsymbol{z} := (z_j)_{j \in \mathcal{P}}$ be a vector of 0–1 decision variables for subset selection; that is, $z_j = 1$ if $j \in \mathcal{S}$; otherwise, $z_j = 0$. We also introduce $\boldsymbol{a}^{(k)} := (a_j^{(k)})_{j \in \mathcal{P}}$, a vector of decision variables that correspond to regression coefficients for the kth training set.

The best subset selection via the cross-validation criterion can be posed as a bilevel MIO problem. Specifically, the subset selection problem for minimizing the cross-validation error (5) is formulated as the *upper-level problem*:

$$\begin{aligned} \text{minimize} \quad & \sum_{k \in \mathcal{K}} \|\boldsymbol{y}(\mathcal{V}_k) - \boldsymbol{X}(\mathcal{V}_k, \mathcal{P})\boldsymbol{a}^{(k)}\|_2^2 && (6) \\ \text{subject to} \quad & \boldsymbol{a}^{(k)} \in \mathcal{A}^{(k)}(\boldsymbol{z}) \quad (k \in \mathcal{K}), && (7) \\ & \boldsymbol{a}^{(k)} \in \mathbb{R}^p \quad (k \in \mathcal{K}), \quad \boldsymbol{z} \in \{0,1\}^p, && (8) \end{aligned}$$

where the training phase of the cross-validation is expressed as the *lower-level problem*:

$$\begin{aligned} \mathcal{A}^{(k)}(\boldsymbol{z}) := \quad \mathop{\arg\min} \quad & \|\boldsymbol{y}(\mathcal{T}_k) - \boldsymbol{X}(\mathcal{T}_k, \mathcal{P})\boldsymbol{a}^{(k)}\|_2^2 + \lambda \|\boldsymbol{a}^{(k)}\|_2^2 && (9) \\ \text{subject to} \quad & z_j = 0 \ \Rightarrow \ a_j^{(k)} = 0 \quad (j \in \mathcal{P}), && (10) \\ & \boldsymbol{a}^{(k)} \in \mathbb{R}^p. && (11) \end{aligned}$$

If $z_j = 0$, then the jth explanatory variable is eliminated from the regression model because its coefficient must be zero by the logical implication (10). This logical implication can be imposed using the indicator function offered by modern optimization software. As a result, $\mathcal{A}^{(k)}(\boldsymbol{z})$ denotes a set of the ridge estimator (4) with $\mathcal{S} = \{j \in \mathcal{P} \mid z_j = 1\}$ for the kth training set. In this bilevel MIO formulation, the ridge estimator is computed in the lower-level problem (9)–(11) from the training set, and its cross-validation error is minimized in the upper-level problem (6)–(8) for subset selection.

When the regularization parameter λ is positive, the lower-level problem has the following desirable property.

Theorem 1 *When $\lambda > 0$, the lower-level problem* (9)–(11) *has a unique optimal solution for each $\boldsymbol{z} \in \{0,1\}^p$.*

Proof Note that the lower-level problem (9)–(11) is equivalent to problem (4) when $\mathcal{S} = \{j \in \mathcal{P} \mid z_j = 1\}$. The Hessian matrix of the objective function in problem (4) is $\boldsymbol{X}(\mathcal{T}_k, \mathcal{S})^\top \boldsymbol{X}(\mathcal{T}_k, \mathcal{S}) + \lambda \boldsymbol{I}$, which is positive definite when $\lambda > 0$. Hence, the objective function is strictly convex, and problem (9)–(11) has a unique optimal solution (see, e.g., Sections 3.3.7 and 3.4.2 [5]). □

Even though $\lambda > 0$, it is difficult to handle problem (6)–(11) due to its bilevel nature. To avoid this difficulty, we convert the bilevel MIO problem (6)–(11) into a single-level MIO problem. For this purpose, we make use of the following constraint:

$$z_j = 1 \;\Rightarrow\; \boldsymbol{x}_j(\mathcal{T}_k)^\top \boldsymbol{X}(\mathcal{T}_k, \mathcal{P})\boldsymbol{a}^{(k)} + \lambda a_j^{(k)} = \boldsymbol{x}_j(\mathcal{T}_k)^\top \boldsymbol{y}(\mathcal{T}_k) \quad (j \in \mathcal{P}). \tag{12}$$

This is an extension of the normal-equation-based constraint [14,39] to the cross-validation for ridge regression. We prove that imposing constraint (12) is equivalent to solving problem (9)–(11) in the following sense.

Theorem 2 *Suppose that* $(\boldsymbol{a}^{(k)}, \boldsymbol{z}) \in \mathbb{R}^p \times \{0,1\}^p$ *satisfies constraint* (10)*. Then,* $\boldsymbol{a}^{(k)} \in \mathcal{A}^{(k)}(\boldsymbol{z})$ *holds if and only if* $(\boldsymbol{a}^{(k)}, \boldsymbol{z})$ *satisfies constraint* (12)*.*

Proof Without loss of generality, we may partition $\boldsymbol{a}^{(k)}$ as

$$\boldsymbol{a}^{(k)} = \begin{pmatrix} \boldsymbol{a}_{\mathcal{S}}^{(k)} \\ \boldsymbol{0} \end{pmatrix}, \quad \boldsymbol{a}_{\mathcal{S}}^{(k)} := (a_j^{(k)})_{j \in \mathcal{S}},$$

due to constraint (10). Therefore, constraint (12) is rewritten as

$$\left(\boldsymbol{X}(\mathcal{T}_k, \mathcal{S})^\top \boldsymbol{X}(\mathcal{T}_k, \mathcal{S}) + \lambda \boldsymbol{I}\right) \boldsymbol{a}_{\mathcal{S}}^{(k)} = \boldsymbol{X}(\mathcal{T}_k, \mathcal{S})^\top \boldsymbol{y}(\mathcal{T}_k),$$

which corresponds to a system of linear equations (2) with $(\mathcal{T}_k, \mathcal{S})$. It then follows that $\boldsymbol{a}_{\mathcal{S}}^{(k)}$ coincides with the ridge estimator (4) for the kth training set, or equivalently, $\boldsymbol{a}^{(k)} \in \mathcal{A}^{(k)}(\boldsymbol{z})$ holds. □

Consequently, we obtain the following single-level reformulation for the bilevel MIO problem (6)–(11):

$$\begin{aligned}
\text{minimize} \quad & \sum_{k \in \mathcal{K}} \|\boldsymbol{y}(\mathcal{V}_k) - \boldsymbol{X}(\mathcal{V}_k, \mathcal{P})\boldsymbol{a}^{(k)}\|_2^2 && (13)\\
\text{subject to} \quad & z_j = 1 \;\Rightarrow\; \boldsymbol{x}_j(\mathcal{T}_k)^\top \boldsymbol{X}(\mathcal{T}_k, \mathcal{P})\boldsymbol{a}^{(k)} + \lambda a_j^{(k)} \\
& \qquad\qquad = \boldsymbol{x}_j(\mathcal{T}_k)^\top \boldsymbol{y}(\mathcal{T}_k) \quad (j \in \mathcal{P},\ k \in \mathcal{K}), && (14)\\
& z_j = 0 \;\Rightarrow\; a_j^{(k)} = 0 \quad (j \in \mathcal{P},\ k \in \mathcal{K}), && (15)\\
& \boldsymbol{a}^{(k)} \in \mathbb{R}^p \quad (k \in \mathcal{K}), \quad \boldsymbol{z} \in \{0,1\}^p. && (16)
\end{aligned}$$

This is an MIQO problem, where the convex quadratic function is minimized subject to the logical implications and the linear constraints. Hence, it can be handled by optimization software using a branch-and-bound procedure.

5 Simulation Experiments

This section evaluates the effectiveness of our subset selection method through simulation experiments.

5.1 Experimental Design

We set the numbers of candidate explanatory variables and samples as $p := 25$ and $n := 100$, respectively. As formally defined later in Eq. (17), SNR corresponds to the goodness of fit of a regression model. We tested SNR $\in \{0.25, 1.00, 4.00\}$ because Hastie et al. [17] reported that the relative performance of subset selection algorithms was dependent on SNR.

Synthetic Datasets. In reference to the previous studies [9,17], we generated synthetic datasets according to the following steps:

1. we defined a vector of "true coefficients" having eight nonzero entries as
$$\boldsymbol{a}^* := (a_j^*)_{j \in \mathcal{P}} := (0, 0, 1, 0, 0, 1, 0, 0, 1, \dots, 0, 0, 1, 0)^\top \in \mathbb{R}^p;$$
2. we drew each row vector $\boldsymbol{x}^\top \in \mathbb{R}^p$ in the matrix $\boldsymbol{X}$ from a normal distribution $\boldsymbol{x} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} := (\sigma_{ij})_{(i,j) \in \mathcal{P} \times \mathcal{P}}$ is the covariance matrix with $\sigma_{ij} := 0.35^{|i-j|}$;
3. we generated a response $y := (\boldsymbol{a}^*)^\top \boldsymbol{x} + \varepsilon$, where each residual was drawn from a normal distribution $\varepsilon \sim \mathrm{N}(0, \sigma^2)$, and the standard deviation σ was determined to meet SNR $\in \{0.25, 1.00, 4.00\}$, which is expressed as
$$\mathbf{SNR} := \frac{\mathrm{Var}((\boldsymbol{a}^*)^\top \boldsymbol{x})}{\mathrm{Var}(\varepsilon)} = \frac{(\boldsymbol{a}^*)^\top \boldsymbol{\Sigma} \boldsymbol{a}^*}{\sigma^2}. \tag{17}$$

Evaluation Metrics. Let $\hat{\boldsymbol{z}} \in \{0, 1\}^p$ be a vector representing selected explanatory variables, and $\hat{\boldsymbol{a}} \in \mathbb{R}^p$ be the associated regression coefficients. Then, the number of correctly selected variables is $(\boldsymbol{a}^*)^\top \hat{\boldsymbol{z}}$, whereas the numbers of selected variables and true variables are $\boldsymbol{1}^\top \hat{\boldsymbol{z}}$ and $\boldsymbol{1}^\top \boldsymbol{a}^*$, respectively. To evaluate the accuracy of subset selection, we used the *F1 score* [32], which is the harmonic average of Recall $:= ((\boldsymbol{a}^*)^\top \hat{\boldsymbol{z}})/(\boldsymbol{1}^\top \boldsymbol{a}^*)$ and Precision $:= ((\boldsymbol{a}^*)^\top \hat{\boldsymbol{z}})/(\boldsymbol{1}^\top \hat{\boldsymbol{z}})$,

$$\mathbf{F1\ Score} := \frac{2 \cdot \mathrm{Recall} \cdot \mathrm{Precision}}{\mathrm{Recall} + \mathrm{Precision}}.$$

We also computed the *relative test error* [17], which represents the expected (our-of-sample) prediction error,

$$\mathbf{Relative\ Test\ Error} := \frac{\mathbb{E}[(y - \hat{\boldsymbol{a}}^\top \boldsymbol{x})^2]}{\mathrm{Var}(\varepsilon)} = \frac{(\boldsymbol{a}^* - \hat{\boldsymbol{a}})^\top \boldsymbol{\Sigma} (\boldsymbol{a}^* - \hat{\boldsymbol{a}}) + \sigma^2}{\sigma^2},$$

where its perfect score is 1 when $\hat{\boldsymbol{a}} = \boldsymbol{a}^*$, and its null score is SNR + 1 when $\hat{\boldsymbol{a}} = \boldsymbol{0}$. Note also that we refer to the number of selected variables as

$$\mathbf{Number\ of\ Nonzeros} := \boldsymbol{1}^\top \hat{\boldsymbol{z}}.$$

These results were averaged over five repetitions.

Subset Selection Methods. We compare the performance of the following subset selection methods:

- **AR2**: Exhaustive search based on the adjusted R^2 [43],
- **MC**: Exhaustive search based on Mallows' C_p [23],
- **BIC**: Exhaustive search based on BIC [35],
- **L1**: L_1-regularized regression [41],
- **CV**: Cross-validation-based MIQO formulation (13)–(16).

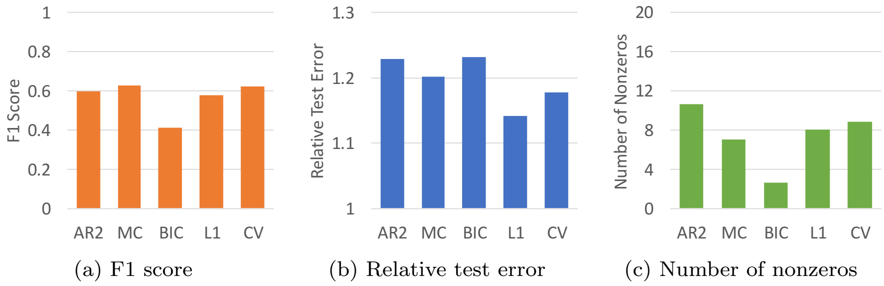
All computations were carried out on a Windows computer with an Intel Core i7-4790 MCU (3.60 GHz) and 16 GB memory. The exhaustive search for AR2, MC, and BIC was performed using the `leaps` 3.0 package [24] in R 3.4.4. It is known that minimizing Mallows' C_p is approximately equivalent to minimizing AIC [1] for a linear regression model [24]. The L_1-regularized regression was performed using the `glmnet` 2.0-16 package [15] in R 3.4.4, where the regularization parameter was tuned based on the mean cross-validation error. These algorithms (i.e., AR2, MC, BIC, and L1) spent less than a few seconds on subset selection in our simulation. The MIQO problem (13)–(16) was solved using IBM ILOG CPLEX 12.8.0.0, and the indicator function implemented in CPLEX was used to impose logical implications (14)–(15). Here, the 10-fold cross-validation was employed (i.e., $K := 10$). A sequence of MIQO problems with $\lambda \in \{0, 0.1, 1, 10, 100, 1000\}$ were solved and then λ was chosen such that the corresponding optimal value of the objective function (13) was the smallest. Each of the MIQO computations was terminated if it did not finish by itself within 1,200 seconds. In these cases, the best feasible solution obtained within 1,200 seconds was taken as the result. The obtained regression coefficients $\hat{\boldsymbol{a}}^{(k)}$ were averaged as $\hat{\boldsymbol{a}} = (\sum_{k \in \mathcal{K}} \hat{\boldsymbol{a}}^{(k)})/K$ to compute the relative test error.

5.2 Simulation Results

Figures 1–3 show the simulation results of the subset selection methods. The F1 score reflects the accuracy of subset selection, so the higher the better. The relative test error corresponds to the expected prediction error, so the lower the better.

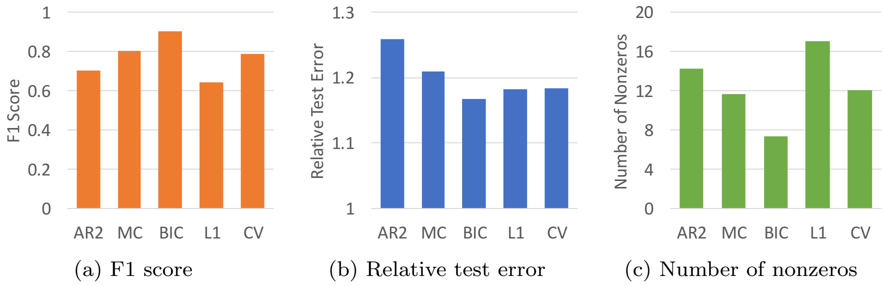
Figure 1 shows the results for SNR = 0.25. We can see that MC and CV provided almost the same F1 score, which is better than those obtained by the other methods. Meanwhile, the best relative test error was attained by L1, and the second best was given by CV. The main reason for this is that these two methods contain the regularization terms to avoid overly fitting a regression model to noisy datasets. It is noteworthy that the number of explanatory variables selected by BIC was much smaller than eight (i.e., the number of true variables). For this reason, the performance of BIC was the worst of the five methods in both the F1 score and the relative test error.

Figure 2 shows the results for SNR = 1.00. In this case, BIC achieved the best performance in both the F1 score and the relative test error. MC and CV



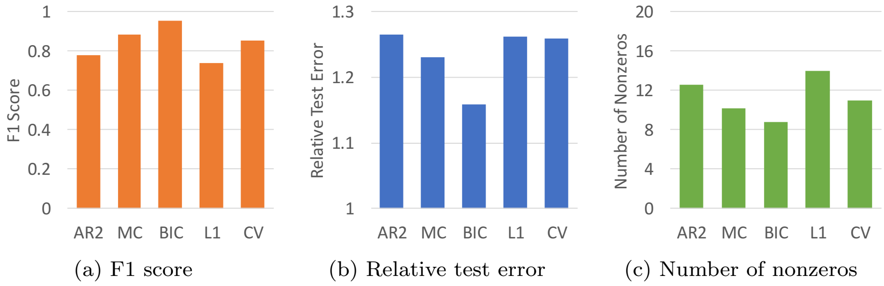
(a) F1 score (b) Relative test error (c) Number of nonzeros

Fig. 1 Simulation results: SNR = 0.25



(a) F1 score (b) Relative test error (c) Number of nonzeros

Fig. 2 Simulation results: SNR = 1.00



(a) F1 score (b) Relative test error (c) Number of nonzeros

Fig. 3 Simulation results: SNR = 4.00

attained approximately the same F1 score, which was the second best of the five methods. We can also find that L1 had the worst F1 score and that L1 selected over 16 variables, which is twice the number of true variables. The relative test errors of L1 and CV were very similar and slightly worse than the best one obtained by BIC. The relative test error of AR2 was by far the worst of the five methods.

Figure 3 shows the results for SNR = 4.00. As in Fig. 2, BIC had the best performance in both the F1 score and the relative test error. High SNR allows a regression model to fit the datasets very well, and thus BIC was

Table 1 Average computation times (s) of solving MIQO problems

	λ					
SNR	0	0.1	1	10	100	1000
0.25	>1200	>1200	>1200	>1200	>1200	1057.6
1.00	>1200	>1200	>1200	>1200	>1200	857.4
4.00	702.8	704.9	715.2	717.0	>1200	241.9

Table 2 Frequency of regularization parameter values chosen by MIQO formulation

	λ					
SNR	0	0.1	1	10	100	1000
0.25	0	0	0	5	0	0
1.00	0	0	0	5	0	0
4.00	0	1	4	0	0	0

able to distinguish true variables from other variables. MC performed slightly better than CV in both the F1 score and the relative test error. On the other hand, AR2 and L1 provided very low F1 scores in Figs. 2–3; this is because they are likely to select too many variables. These results also imply that the regularization terms did not work well when SNR was very high, which is consistent with the simulation results reported by Hastie et al. [17].

We conclude this section by examining the computational results of our MIQO formulation. Table 1 gives the average computation times (in seconds) required for solving MIQO problems. We can see that MIQO computations finished early when SNR was high. The regularization parameter λ had little association with the computation time, but the MIQO computations were fast only for $\lambda = 1000$. Table 2 gives the frequency of regularization parameter values chosen by MIQO formulation. It reveals that $\lambda = 10$ was always chosen when SNR $\in \{0.25, 1.00\}$, whereas $\lambda = 1$ worked well when SNR $= 4.00$. This means that when SNR is low, one should shrink regression coefficients more considerably to avoid overfitting.

6 Conclusion

This paper dealt with the problem of selecting the best subset of explanatory variables for ridge regression via the cross-validation criterion. This problem can naturally be posed as a bilevel MIO problem, but the bilevel optimization problem is very hard to handle. To make the problem computationally tractable, we derived a single-level MIQO reformulation by means of the optimality condition for ridge regression.

MIO approaches have been proposed for best subset selection with respect to the adjusted R^2 [26], Mallows' C_p [25], and AIC/BIC [19,26]. However, these statistical criteria are not always valid in a variety of applications because they are heavily dependent on some assumptions. In contrast, the cross-validation criterion works in theory without such strong assumptions.

The simulation results confirmed that when SNR was low, our method was effective in terms of the subset selection accuracy and the predictive performance. There is probably no algorithm that always performs best in all situations. In this sense, another contribution of this research is to reveal the advantages and disadvantages of commonly used algorithms for subset selection through the simulation experiments.

A future direction of this study will be to extend our MIO formulation to various regression and classification models. Another direction will be to devise an MIO formulation for selecting λ and $\mathcal{S}$ simultaneously. It is also necessary to speed up the computation for subset selection. One way to do this will be to apply bilevel optimization algorithms [13, 37] to our bilevel MIO formulation.

Acknowledgements This work was partially supported by JSPS KAKENHI Grant Numbers JP17K01246 and JP17K12983.

References

1. Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19(6), 716–723.
2. Allen, D. M. (1974). The relationship between variable selection and data augmentation and a method for prediction. Technometrics, 16(1), 125–127.
3. Arthanari, T.S., & Dodge, Y. (1981). Mathematical Programming in Statistics. Wiley, New York.
4. Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. Statistics Surveys, 4, 40–79.
5. Bazaraa, M. S., Sherali, H. D., & Shetty, C. M. (2013). Nonlinear Programming: Theory and Algorithms. John Wiley & Sons.
6. Benati, S., & García, S. (2014). A mixed integer linear model for clustering with variable selection. Computers & Operations Research, 43, 280–285.
7. Bennett, K. P., Hu, J., Ji, X., Kunapuli, G., & Pang, J. S. (2006). Model selection via bilevel optimization. Proceedings of the 2006 IEEE International Joint Conference on Neural Networks (pp. 1922–1929).
8. Bertsimas, D., & King, A. (2016). OR Forum—An algorithmic approach to linear regression. Operations Research, 64(1), 2–16.
9. Bertsimas, D., King, A., & Mazumder, R. (2016). Best subset selection via a modern optimization lens. The Annals of Statistics, 44(2), 813–852.
10. Bertsimas, D., & Dunn, J. (2017). Optimal classification trees. Machine Learning, 106(7), 1039–1082.
11. Bertsimas, D., & King, A. (2017). Logistic regression: From art to science. Statistical Science, 32(3), 367–384.
12. Chung, S., Park, Y. W., & Cheong, T. (2017). A mathematical programming approach for integrated multiple linear regression subset selection and validation. arXiv preprint arXiv:1712.04543.
13. Colson, B., Marcotte, P., & Savard, G. (2007). An overview of bilevel optimization. Annals of Operations Research, 153(1), 235–256.
14. Cozad, A., Sahinidis, N. V., & Miller, D. C. (2014). Learning surrogate models for simulation-based optimization. AIChE Journal, 60(6), 2211–2227.
15. Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software, 33(1), 1–22.
16. Geisser, S. (1975). The predictive sample reuse method with applications. Journal of the American statistical Association, 70(350), 320–328.

17. Hastie, T., Tibshirani, R., & Tibshirani, R. J. (2017). Extended comparisons of best subset selection, forward stepwise selection, and the lasso. arXiv preprint arXiv:1707.08692.
18. Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1), 55–67.
19. Kimura, K., & Waki, H. (2018). Minimization of Akaike's information criterion in linear regression analysis via mixed integer nonlinear program. Optimization Methods and Software, 33(3), 633–649.
20. Konno, H., & Yamamoto, R. (2009). Choosing the best set of variables in regression analysis using integer programming. Journal of Global Optimization, 44(2), 273–282.
21. Kunapuli, G., Bennett, K. P., Hu, J., & Pang, J. S. (2008). Classification model selection via bilevel programming. Optimization Methods & Software, 23(4), 475–489.
22. Maldonado, S., Pérez, J., Weber, R., & Labbé, M. (2014). Feature selection for support vector machines via mixed integer linear programming. Information Sciences, 279, 163–175.
23. Mallows, C. L. (1973). Some comments on C_p. Technometrics, 15(4), 661–675.
24. Miller, A. (2002). Subset Selection in Regression. Chapman and Hall/CRC.
25. Miyashiro, R., & Takano, Y. (2015). Subset selection by Mallows' C_p: A mixed integer programming approach. Expert Systems with Applications, 42(1), 325–331.
26. Miyashiro, R., & Takano, Y. (2015). Mixed integer second-order cone programming formulations for variable selection in linear regression. European Journal of Operational Research, 247(3), 721–731.
27. Mosier, C. I. (1951). I. Problems and designs of cross-validation. Educational and Psychological Measurement, 11(1), 5–11.
28. Naganuma, M., Takano, Y., & Miyashiro, R. (in press). Feature subset selection for ordered logit model via tangent-plane-based approximation. IEICE Transactions on Information and Systems.
29. Okuno, T., Takeda, A., & Kawana, A. (2018). Hyperparameter learning for bilevel nonsmooth optimization. arXiv preprint arXiv:1806.01520.
30. Park, Y. W., & Klabjan, D. (2017). Subset selection for multiple linear regression via optimization. arXiv preprint arXiv:1701.07920.
31. Pedregosa, F. (2016). Hyperparameter optimization with approximate gradient. Proceedings of the 33rd International Conference on Machine Learning (pp. 737–746).
32. van Rijsbergen, C. J. (1979). Information Retrieval, 2nd Edition. Butterworth-Heinemann.
33. Sato, T., Takano, Y., Miyashiro, R., & Yoshise, A. (2016). Feature subset selection for logistic regression via mixed integer optimization. Computational Optimization and Applications, 64(3), 865–880.
34. Sato, T., Takano, Y., & Miyashiro, R. (2017). Piecewise-linear approximation for feature subset selection in a sequential logit model. Journal of the Operations Research Society of Japan, 60(1), 1–14.
35. Schwarz, G. (1978). Estimating the dimension of a model. The Annals of Statistics, 6(2), 461–464.
36. Shao, J. (1993). Linear model selection by cross-validation. Journal of the American statistical Association, 88(422), 486–494.
37. Sinha, A., Malo, P., & Deb, K. (2018). A review on bilevel optimization: From classical to evolutionary approaches and applications. IEEE Transactions on Evolutionary Computation, 22(2), 276–295.
38. Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. Journal of the Royal Statistical Society, Series B (Methodological), 111–147.
39. Tamura, R., Kobayashi, K., Takano, Y., Miyashiro, R., Nakata, K., & Matsui, T. (2017). Best subset selection for eliminating multicollinearity. Journal of the Operations Research Society of Japan, 60(3), 321–336.
40. Tamura, R., Kobayashi, K., Takano, Y., Miyashiro, R., Nakata, K., & Matsui, T. (in press). Mixed integer quadratic optimization formulations for eliminating multicollinearity based on variance inflation factor. Journal of Global Optimization.
41. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B (Methodological), 267–288.
42. Ustun, B., & Rudin, C. (2016). Supersparse linear integer models for optimized medical scoring systems. Machine Learning, 102(3), 349–391.

43. Wherry, R. (1931). A new formula for predicting the shrinkage of the coefficient of multiple correlation. The Annals of Mathematical Statistics, 2(4), 440–457.
44. Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society, Series B (Statistical Methodology), 67(2), 301–320.