TOP (2020) 28:475–488
https://doi.org/10.1007/s11750-020-00538-1

ORIGINAL PAPER

Best subset selection via cross-validation criterion

Yuichi Takano[1] · **Ryuhei Miyashiro**[2]

Received: 13 January 2019 / Accepted: 18 January 2020 / Published online: 14 February 2020

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

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

Mathematics Subject Classification 62F07 · 62J05 · 90C11 · 90C90

✉ Yuichi Takano
ytakano@sk.tsukuba.ac.jp

[1] Faculty of Engineering, Information and Systems, University of Tsukuba, 1-1-1 Tennodai, Tsukuba-shi, Ibaraki 305-8577, Japan

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

1 Introduction

Subset selection (Miller 2002), also known as variable/feature/attribute selection, involves selecting a significant subset of explanatory variables from which to construct a regression model. Such selection aids in understanding causality between explanatory and response variables. It also reduces the costs of gathering data and the time required for estimating the values of model parameters. Moreover, the predictive performance of regression models can be improved by the use of subset selection because overfitting is mitigated by elimination of redundant explanatory variables.

Another approach to boosting the predictive performance is to use a shrinkage method. These include ridge regression (Hoerl and Kennard 1970), lasso (Tibshirani 1996), and elastic net (Zou and Hastie 2005). Shrinkage methods "shrink" the regression coefficients of the explanatory variables toward zero. Ridge regression has a theoretical advantage over other methods when dealing with multicollinearity. The lasso also works as a subset-selection method because it has the 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 adjusted R^2 (Wherry 1931), Mallows' C_p (Mallows 1973), the Akaike information criterion (AIC, Akaike 1974), and the Bayesian information criterion (BIC, Schwarz 1978). Both 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 the validity of these statistical criteria depend on some strict assumptions being met, they are not always suitable in practice.

This paper is focused on the cross-validation criterion (Allen 1974; Geisser 1975; Mosier 1951; Shao 1993; Stone 1974) for best-subset selection. Specifically, to evaluate the quality of a subset regression model, we split a set of given samples into a training set and a validation set. The training set is used for parameter-value estimation, and the prediction error is computed from applying the trained parameter values to the validation set. In contrast with the use of statistical criteria, the cross-validation only requires two mild assumptions: that samples are identically distributed and that the training and validation samples are independent (Arlot and Celisse 2010). Consequently, the cross-validation criterion can be applied to any predictive method, and it is expected to work well in most situations.

To accomplish best-subset selection via the cross-validation criterion, we adopt a mixed-integer optimization (MIO) approach. This approach was first proposed in the 1970s (Arthanari and Dodge 1981), and recently it has received renewed attention as advances in optimization algorithms and improved computer performance have made them feasible to execute (Bertsimas et al. 2016; Cozad et al. 2014; Konno and Yamamoto 2009; Park and Klabjan 2017; Ustun and Rudin 2016). Hastie et al. (2017) reported that when the signal-to-noise ratio (SNR) was high, an MIO approach resulted in better predictions than the lasso did. MIO approaches have been proposed for best-subset selection with respect to

adjusted R^2 (Miyashiro and Takano 2015b), Mallows' C_p (Miyashiro and Takano 2015a), and AIC/BIC (Kimura and Waki 2018; Miyashiro and Takano 2015b). Additionally, MIO-based subset selection has been applied to logit models (Bertsimas and King 2017; Naganuma et al. 2019; Sato et al. 2016, 2017), support vector machines (Maldonado et al. 2014), cluster analysis (Benati and García 2014), classification trees (Bertsimas and Dunn 2017), elimination of multicollinearity (Bertsimas and King 2016; Tamura et al. 2017, 2019), and statistical tests/diagnostics (Chung et al. 2017).

The aim of this paper is to establish a practicable MIO approach to selecting the best subset of explanatory variables via the cross-validation criterion used with ridge regression. This subset-selection problem can be posed as a bilevel MIO problem, but it is difficult to handle such bilevel optimization problems. To remedy this situation, we transform the problem into a single-level mixed-integer quadratic optimization (MIQO) problem. Optimization software can exactly solve the resultant MIQO problem. Algorithms for bilevel optimization (Colson et al. 2007; Sinha et al. 2018) have been employed for hyperparameter tuning in support vector regression (Bennett et al. 2006), support vector classification (Kunapuli et al. 2008), general supervised learning (Pedregosa 2016), and nonsmooth regularization (Okuno et al. 2018). To the best of our knowledge, however, we are the first to develop an effective method for best-subset selection using the cross-validation criterion.

The efficacy of our method is assessed through simulation experiments, following the method of previous studies (Bertsimas et al. 2016; Hastie et al. 2017). We compare the performance of our method with that of statistical-criterion-based exhaustive search algorithms (Miller 2002) and L_1-regularized regression (Tibshirani 1996). The simulation results demonstrate that when the SNR is low, our method is superior in terms of subset-selection accuracy and predictive performance.

The main contributions of this paper are the following.

- We derive a bilevel MIO formulation to select the best subset of explanatory variables via the cross-validation criterion for ridge regression. This opens up new possibilities of using the cross-validation criterion and (bilevel) MIO algorithms for best-subset selection.
- We devise a method of reformulating the bilevel MIO problem as a single-level MIQO problem that can be solved exactly by optimization software. This is the first computationally feasible method for best-subset selection via the cross-validation criterion.
- We carry out simulation experiments to examine the accuracy of subset selection and prediction when using various subset-selection methods. The simulation results reveal the advantages and disadvantages of the cross-validation criterion and also offer valuable insight for choosing an appropriate method of subset selection.
- We show through the simulation experiments that our method works well when the SNR is low. This highlights the usefulness of our method in cases where the SNR is low and enhances the potential for application of MIO approaches to best-subset selection.

2 Ridge regression

Let us assume that there is some relation between a response variable y and a vector $\boldsymbol{x} := (x_1, x_2, \ldots, x_p)^\top$ composed of p explanatory variables. Multiple linear regression involves estimating the following linear relation:

$$y = (\boldsymbol{a}^*)^\top \boldsymbol{x} + \varepsilon,$$

where $\boldsymbol{a}^* \in \mathbb{R}^p$ is a vector of (unknown) true coefficients, and ε is an error term. Here, we can regard $(\boldsymbol{a}^*)^\top \boldsymbol{x}$ as a predictable *signal* and ε as an unpredictable *noise*, both reflected in the observed responses.

The strength of the relation can be characterized by the SNR. For consistency with our simulation experiments, we assume that the explanatory variables and the error term are normally distributed; specifically, we assume that $\boldsymbol{x} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$ with covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$, and $\varepsilon \sim \mathrm{N}(0, \sigma^2)$ with standard deviation $\sigma \in \mathbb{R}$. The SNR is defined as the ratio of the signal variance to the noise variance:

$$\textbf{SNR} := \frac{\mathrm{Var}\left((\boldsymbol{a}^*)^\top \boldsymbol{x}\right)}{\mathrm{Var}\left(\varepsilon\right)} = \frac{(\boldsymbol{a}^*)^\top \boldsymbol{\Sigma} \boldsymbol{a}^*}{\sigma^2}. \tag{1}$$

When the SNR is high, the response variable is strongly associated with the explanatory variables, so it is easy to estimate the potential relation (i.e., $\boldsymbol{a}^*$) between the response and explanatory variables.

For the purpose of estimation, we are given n samples $(y_i, x_{i1}, x_{i2}, \ldots, x_{ip})$ for $i = 1, 2, \ldots, n$. Here, y_i is the 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 $\mathcal{P} := \{1, 2, \ldots, p\}$ and $\mathcal{N} := \{1, 2, \ldots, n\}$, respectively.

We assume without loss of generality that all explanatory and response variables are centered, meaning 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

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

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

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

Here, $\boldsymbol{a}$ is a vector of regression coefficients to be estimated, and $\boldsymbol{e}$ is a vector containing the prediction residuals.

This paper is focused on ridge regression for a multiple linear regression model. Specifically, we minimize the residual sum of squares (RSS) with the L_2 -regularization term to shrink the regression coefficients toward zero, expressed as

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

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

$$(\boldsymbol{X}^\top\boldsymbol{X}+\lambda\boldsymbol{I})\hat{\boldsymbol{a}}=\boldsymbol{X}^\top\boldsymbol{y}. \tag{3}$$

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* $\mathcal{V}_k$, respectively, as

$$\mathcal{T}_k := \mathcal{N}\setminus\mathcal{N}_k \quad\text{and}\quad \mathcal{V}_k := \mathcal{N}_k. \tag{4}$$

We also use the following notations to extract the parts of the response and explanatory variables corresponding to the 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 the value of the regularization parameter $\lambda \in \mathbb{R}_+$ and taking a subset $\mathcal{S} \subseteq \mathcal{P}$ of the explanatory variables. In the training phase, we compute the ridge estimator for each $k \in \mathcal{K}$ from the kth training set as follows:

$$\hat{\boldsymbol{a}}_{\mathcal{S}}^{(k)}\in\arg\min\{\,\|\boldsymbol{y}(\mathcal{T}_k)-\boldsymbol{X}(\mathcal{T}_k,\mathcal{S})\boldsymbol{a}_{\mathcal{S}}\|_2^2+\lambda\|\boldsymbol{a}_{\mathcal{S}}\|_2^2\mid\boldsymbol{a}_{\mathcal{S}}\in\mathbb{R}^{|\mathcal{S}|}\,\}. \tag{5}$$

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

$$\sum_{k\in\mathcal{K}}\|\boldsymbol{y}(\mathcal{V}_k)-\boldsymbol{X}(\mathcal{V}_k,\mathcal{S})\hat{\boldsymbol{a}}_{\mathcal{S}}^{(k)}\|_2^2. \tag{6}$$

Applying the *cross-validation criterion* involves selecting the subset $\mathcal{S}$ of explanatory variables that produces the lowest cross-validation error (6). To accomplish this, however, we must repeat 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 binary decision variables for subset selection; that is, $z_j = 1$ if $j \in \mathcal{S}$ and $z_j = 0$ otherwise. Here, we also introduce $\boldsymbol{a}^{(k)} := (a_j^{(k)})_{j\in\mathcal{P}}$, a vector of decision variables that correspond to the regression coefficients for the kth training set.

Best-subset selection via the cross-validation criterion can be posed as a bilevel MIO problem. Specifically, the subset-selection problem of minimizing the cross-validation error (6) can be formulated as the *upper-level* problem

$$\text{minimize} \quad \sum_{k\in\mathcal{K}} \|\boldsymbol{y}(\mathcal{V}_k) - \boldsymbol{X}(\mathcal{V}_k, \mathcal{P})\boldsymbol{a}^{(k)}\|_2^2 \tag{7}$$

$$\text{subject to} \quad \boldsymbol{a}^{(k)} \in \mathcal{A}^{(k)}(\boldsymbol{z}) \quad (k \in \mathcal{K}), \tag{8}$$

$$\boldsymbol{a}^{(k)} \in \mathbb{R}^p \quad (k \in \mathcal{K}), \quad \boldsymbol{z} \in \{0,1\}^p, \tag{9}$$

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

$$\mathcal{A}^{(k)}(\boldsymbol{z}) := \arg\min \|\boldsymbol{y}(\mathcal{T}_k) - \boldsymbol{X}(\mathcal{T}_k, \mathcal{P})\boldsymbol{a}^{(k)}\|_2^2 + \lambda\|\boldsymbol{a}^{(k)}\|_2^2 \tag{10}$$

$$\text{subject to} \quad z_j = 0 \Rightarrow a_j^{(k)} = 0 \quad (j \in \mathcal{P}), \tag{11}$$

$$\boldsymbol{a}^{(k)} \in \mathbb{R}^p. \tag{12}$$

Note that all of the decision variables are listed in the constraints (9) and (12).

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 (11). This logical implication can be imposed by using the indicator function implemented in modern optimization software. As a result of optimization, $\mathcal{A}^{(k)}(\boldsymbol{z})$ denotes a set containing the ridge estimator (5) 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 (10)–(12) from the training set, and the associated cross-validation error is minimized in the upper-level problem (7)–(9) for subset selection.

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

Theorem 1 *When $\lambda > 0$, the lower-level problem (10)–(12) has a unique optimal solution for each $\boldsymbol{z} \in \{0,1\}^p$.*

Proof Note that the lower-level problem (10)–(12) is equivalent to problem (5) when $\mathcal{S} = \{j \in \mathcal{P} \mid z_j = 1\}$. The Hessian matrix of the objective function in problem (5) is $\boldsymbol{X}(\mathcal{T}_k, \mathcal{S})^\top \boldsymbol{X}(\mathcal{T}_k, \mathcal{S}) + \lambda \boldsymbol{I}$, and it is positive definite. Hence, the objective function

is strongly convex, so problem (10)–(12) has a unique optimal solution (for details, see, e.g., Section 9.1.2 of Boyd and Vandenberghe 2004). □

Even when $\lambda > 0$, it is difficult to handle problem (7)–(12) due to its bilevel nature. To avoid this difficulty, we convert the bilevel MIO problem (7)–(12) into a single-level MIQO problem. For this purpose, we make use of the following constraint:

$$z_j = 1 \Rightarrow \boldsymbol{x}_j(\mathcal{T}_k)^\top \boldsymbol{X}(\mathcal{T}_k, \mathcal{P})\boldsymbol{a}^{(k)} + \lambda a_j^{(k)} = \boldsymbol{x}_j(\mathcal{T}_k)^\top \boldsymbol{y}(\mathcal{T}_k) \quad (j \in \mathcal{P}). \tag{13}$$

This is an extension of the normal-equation-based constraint (Cozad et al. 2014; Tamura et al. 2017) to the cross-validation criterion for ridge regression. We prove that imposing constraint (13) is equivalent to solving the lower-level problem (10)–(12) in the following sense.

Theorem 2 *Suppose that* $(\boldsymbol{a}^{(k)}, \boldsymbol{z}) \in \mathbb{R}^p \times \{0,1\}^p$ *satisfies constraint* (11). *Then,* $\boldsymbol{a}^{(k)} \in \mathcal{A}^{(k)}(\boldsymbol{z})$ *holds if and only if* $(\boldsymbol{a}^{(k)}, \boldsymbol{z})$ *satisfies constraint* (13).

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

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

due to constraint (11). Therefore, constraint (13) can be rewritten as

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

which corresponds to a system of linear equations (3) involving $(\mathcal{T}_k, \mathcal{S})$. It then follows that $\boldsymbol{a}_{\mathcal{S}}^{(k)}$ coincides with the ridge estimator (5) for the kth training set, or, equivalently, $\boldsymbol{a}^{(k)} \in \mathcal{A}^{(k)}(\boldsymbol{z})$ holds. □

From this theorem, we obtain the following single-level reformulation for the bilevel MIO problem (7)–(12):

$$\text{minimize} \quad \sum_{k \in \mathcal{K}} \|\boldsymbol{y}(\mathcal{V}_k) - \boldsymbol{X}(\mathcal{V}_k, \mathcal{P})\boldsymbol{a}^{(k)}\|_2^2 \tag{14}$$

$$\text{subject to} \quad z_j = 1 \Rightarrow \boldsymbol{x}_j(\mathcal{T}_k)^\top \boldsymbol{X}(\mathcal{T}_k, \mathcal{P})\boldsymbol{a}^{(k)} + \lambda a_j^{(k)} = \boldsymbol{x}_j(\mathcal{T}_k)^\top \boldsymbol{y}(\mathcal{T}_k) \quad (j \in \mathcal{P}, k \in \mathcal{K}), \tag{15}$$

$$z_j = 0 \Rightarrow a_j^{(k)} = 0 \quad (j \in \mathcal{P}, k \in \mathcal{K}), \tag{16}$$

$$\boldsymbol{a}^{(k)} \in \mathbb{R}^p \quad (k \in \mathcal{K}), \quad \boldsymbol{z} \in \{0,1\}^p. \tag{17}$$

This is an MIQO problem in which the convex quadratic function is minimized, subject to the logical implications and the linear constraints. Note that problem (14)–(17) can be rewritten as follows:

$$\begin{aligned} \text{minimize} \quad & \sum_{k\in\mathcal{K}} \big((\boldsymbol{a}^{(k)})^\top \boldsymbol{X}(\mathcal{V}_k, \mathcal{P})^\top \boldsymbol{X}(\mathcal{V}_k, \mathcal{P}) \boldsymbol{a}^{(k)} \\ & \qquad -2\boldsymbol{y}(\mathcal{V}_k)^\top \boldsymbol{X}(\mathcal{V}_k, \mathcal{P}) \boldsymbol{a}^{(k)} + \boldsymbol{y}(\mathcal{V}_k)^\top \boldsymbol{y}(\mathcal{V}_k) \big) \end{aligned} \tag{18}$$

$$\begin{aligned} \text{subject to} \quad & \boldsymbol{x}_j(\mathcal{T}_k)^\top \boldsymbol{X}(\mathcal{T}_k, \mathcal{P}) \boldsymbol{a}^{(k)} + \lambda a_j^{(k)} \\ & \qquad \geq \boldsymbol{x}_j(\mathcal{T}_k)^\top \boldsymbol{y}(\mathcal{T}_k) - M(1 - z_j) \quad (j \in \mathcal{P}, k \in \mathcal{K}), \end{aligned} \tag{19}$$

$$\begin{aligned} & \boldsymbol{x}_j(\mathcal{T}_k)^\top \boldsymbol{X}(\mathcal{T}_k, \mathcal{P}) \boldsymbol{a}^{(k)} + \lambda a_j^{(k)} \\ & \qquad \leq \boldsymbol{x}_j(\mathcal{T}_k)^\top \boldsymbol{y}(\mathcal{T}_k) + M(1 - z_j) \quad (j \in \mathcal{P}, k \in \mathcal{K}), \end{aligned} \tag{20}$$

$$-Mz_j \leq a_j^{(k)} \leq Mz_j \quad (j \in \mathcal{P}, k \in \mathcal{K}), \tag{21}$$

$$\boldsymbol{a}^{(k)} \in \mathbb{R}^p \quad (k \in \mathcal{K}), \quad \boldsymbol{z} \in \{0, 1\}^p, \tag{22}$$

where M is a sufficiently large positive constant.

Problem (14)–(17) can be handled by optimization software using a branch-and-bound procedure. Specifically, to solve the problem, the logical implications (15)–(16) are relaxed, and then the relaxed problem is solved while some components of $\boldsymbol{z} \in \{0, 1\}^p$ are fixed at each node of the enumeration tree. Suppose that $z_j = 0$ for $j \in \mathcal{J}_0$ and $z_j = 1$ for $j \in \mathcal{J}_1$ are fixed at a particular node of the enumeration tree. In this case, a basic strategy involves solving the relaxed problem with the following constraints:

$$\begin{aligned} & \boldsymbol{x}_j(\mathcal{T}_k)^\top \boldsymbol{X}(\mathcal{T}_k, \mathcal{P}) \boldsymbol{a}^{(k)} + \lambda a_j^{(k)} = \boldsymbol{x}_j(\mathcal{T}_k)^\top \boldsymbol{y}(\mathcal{T}_k) \quad (j \in \mathcal{J}_1, k \in \mathcal{K}), \\ & a_j^{(k)} = 0 \quad (j \in \mathcal{J}_0, k \in \mathcal{K}). \end{aligned}$$

These constraints result from the logical implications (15)–(16) and the fixed binary decision variables (see, e.g., Hooker and Osorio 1999 for the details).

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. We tested SNR $\in \{0.25, 1.00, 4.00\}$ using multiple values because Hastie et al. (2017) reported that the relative performance of subset-selection algorithms depends on the SNR.

Synthetic datasets. Following previous studies (Bertsimas et al. 2016; Hastie et al. 2017), we generated synthetic datasets according to the following steps.

1. First, we defined a vector of true coefficients having eight nonzero entries as
$$\boldsymbol{a}^* := (0, 0, 1, 0, 0, 1, 0, 0, 1, \ldots, 0, 0, 1, 0)^\top \in \mathbb{R}^p;$$
2. next, we obtained each row vector $\boldsymbol{x}^\top \in \mathbb{R}^p$ in the matrix $\boldsymbol{X}$ according to 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. finally, we generated a response $y := (\boldsymbol{a}^*)^\top \boldsymbol{x} + \varepsilon$ with the error term obtained according to a normal distribution $\varepsilon \sim \mathrm{N}(0, \sigma^2)$ with the standard deviation σ determined so as to produce SNR $\in \{0.25, 1.00, 4.00\}$, using Eq. (1).

Evaluation metrics. Let $\hat{\boldsymbol{z}} \in \{0, 1\}^p$ be a vector representing the selected explanatory variables, and let $\hat{\boldsymbol{a}} \in \mathbb{R}^p$ 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 the subset selection, we used the *F1 score* (van Rijsbergen 1979), which is the harmonic average of Recall $:= ((\boldsymbol{a}^*)^\top \hat{\boldsymbol{z}})/(\boldsymbol{1}^\top \boldsymbol{a}^*)$ and Precision $:= ((\boldsymbol{a}^*)^\top \hat{\boldsymbol{z}})/(\boldsymbol{1}^\top \hat{\boldsymbol{z}})$:

$$\textbf{F1 Score} := \frac{2 \cdot \text{ Recall } \cdot \text{ Precision}}{\text{Recall } + \text{ Precision}}.$$

We also computed the *relative test error* (Hastie et al. 2017), which characterizes the expected (out-of-sample) prediction error:

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

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

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

The results were averaged over five trials.

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

- **AR2**: exhaustive search based on adjusted R^2 (Wherry 1931);

- **MC**: exhaustive search based on Mallows' C_p (Mallows 1973);
- **BIC**: exhaustive search based on BIC (Schwarz 1978);
- **L1**: L_1-regularized regression (Tibshirani 1996); and
- **CV**: cross-validation-based MIQO formulation (14)–(17).

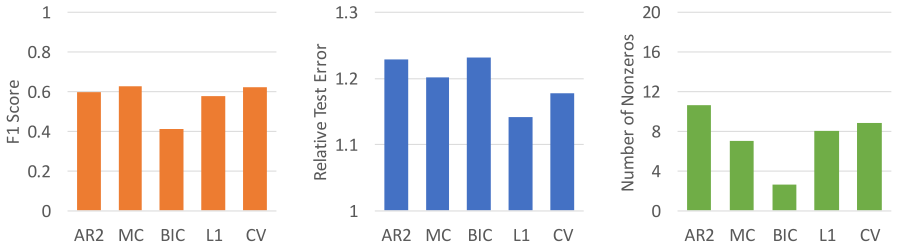
All computations were carried out on a Windows computer with an Intel Core i7-4790 CPU (3.60 GHz) and 16 GB of random-access memory. The exhaustive search for AR2, MC, and BIC was performed using the `leaps` 3.0 package (Miller 2002) with R 3.4.4. It is known that minimizing Mallows' C_p is approximately equivalent to minimizing the AIC (Akaike 1974) for a linear regression model (Miller 2002). L_1-regularized regression was performed using the `glmnet` 2.0-16 package (Friedman et al. 2010) with 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) required less than a few seconds during subset selection in our simulation. The MIQO problem (14)–(17) was solved using IBM ILOG CPLEX 12.8.0.0, and the indicator function implemented in CPLEX was used to impose logical implications (15)–(16). We used 10-fold cross-validation (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 (14) was minimized. Each of the MIQO computations was terminated if it did not finish by itself within 1,200 s. In these cases, the best feasible solution obtained within 1,200 s was taken as the result. The obtained regression coefficients 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, 2 and 3 show the simulation results for the various subset-selection methods. The F1 scores reflect the accuracy of subset selection, with higher scores indicating better performance. The relative test error corresponds to the expected prediction error, so lower scores are better.

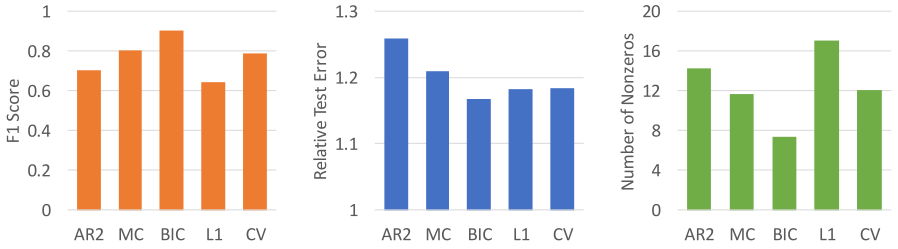
Figure 1 shows the results for SNR $= 0.25$. We can see that MC and CV resulted in almost the same F1 score, with these two obtaining a better score than the other methods. Meanwhile, the best relative test error was attained by L1, and the second-best was attained by CV. The main reason for this is that these two methods employ regularization terms to avoid overfitting the regression model to noisy datasets. It is worth noting that the number of explanatory variables selected by BIC was much smaller than eight (i.e., the true number of explanatory variables). For this reason, the performance of BIC was the worst of the five methods, scoring poorly on 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 on both the F1 score and the relative test error. MC and CV attained approximately the same F1 score, which was the second-best score by the methods. We note that L1 had the worst F1 score and over 16 variables, which is more than twice the true number of variables. The relative test errors of L1 and CV



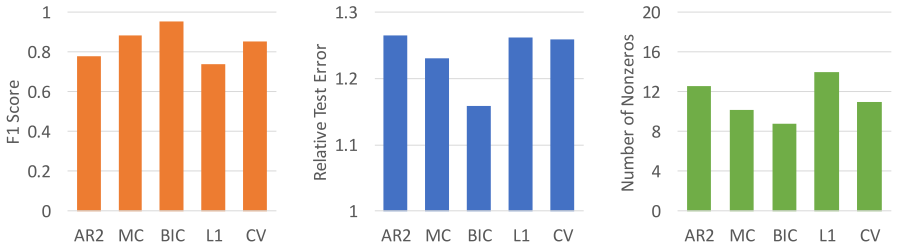
(a) F1 score **(b)** Relative test error **(c)** Number of nonzeros

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 $= 4.00$

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 on both the F1 score and the relative test error. A high SNR allows a regression model to fit the datasets very well, and so BIC was able to distinguish the true (i.e., explanatory) variables from the other variables. MC performed slightly better than CV on both the F1 score and the relative test error. In contrast, AR2 and L1 both received very low F1 scores, as seen in Figs. 2 and 3; this is

Table 1 Average computation times [s] for solving MIQO problems

SNR	λ					
	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

SNR	λ					
	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

because they are likely to select too many variables. These results also imply that the regularization terms did not work well when the SNR was very high, which is consistent with the simulation results reported by Hastie et al. (2017).

We conclude this section by examining the computational results of our MIQO formulation. Table 1 gives the average computation times (in seconds) required to solve the MIQO problems. We can see that the MIQO computations finished quickly when the SNR was high. The regularization parameter λ had little relation to computation time, but the MIQO computations were fast only for $\lambda = 1000$. Table 2 gives the frequency of regularization parameter values chosen by the MIQO formulation. The table shows 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 the SNR is low, one should shrink the regression coefficients more aggressively to avoid overfitting.

6 Conclusion

This paper dealt with the problem of selecting the best subset from a set of explanatory variables, to be applied in ridge regression, via the cross-validation criterion. This problem can be naturally posed as a bilevel MIO problem, but the bilevel optimization problem is quite hard. To decrease the computational difficulty, we reformulated the problem as a single-level MIQO problem by means of the optimality condition for ridge regression. The resultant MIQO problem can be solved exactly by using optimization software.

MIO approaches have been proposed for best-subset selection as judged by adjusted R^2 (Miyashiro and Takano 2015b), Mallows' C_p (Miyashiro and Takano 2015a), and AIC/BIC (Kimura and Waki 2018; Miyashiro and Takano 2015b). However, the assumptions underlying these statistical criteria cannot always be met

in applications. In contrast, the cross-validation criterion is theoretically valid but has assumptions that are easier to satisfy.

The simulation results confirmed that when the SNR was low, our method was effective, providing good subset-selection accuracy and predictive performance. Despite this, there is probably no algorithm that will perform best for all cases. In this sense, another contribution of this research is that the simulation experiments revealed the advantages and disadvantages of some algorithms commonly used for subset selection.

A future direction for this research 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 would be to apply bilevel optimization algorithms (Colson et al. 2007; Sinha et al. 2018) to our bilevel MIO formulation.

Acknowledgements The authors would like to thank two anonymous reviewers for their helpful comments. This work was partially supported by JSPS KAKENHI Grant Numbers JP17K01246 and JP17K12983.

References

Akaike H (1974) A new look at the statistical model identification. IEEE Trans Autom Control 19(6):716–723

Allen DM (1974) The relationship between variable selection and data augmentation and a method for prediction. Technometrics 16(1):125–127

Arthanari TS, Dodge Y (1981) Mathematical programming in statistics. Wiley, New York

Arlot S, Celisse A (2010) A survey of cross-validation procedures for model selection. Stat Surv 4:40–79

Benati S, García S (2014) A mixed integer linear model for clustering with variable selection. Comput Oper Res 43:280–285

Bennett KP, Hu J, Ji X, Kunapuli G, Pang JS (2006) Model selection via bilevel optimization. In: Proceedings of the 2006 IEEE international joint conference on neural networks, pp 1922–1929

Bertsimas D, King A (2016) OR forum—an algorithmic approach to linear regression. Oper Res 64(1):2–16

Bertsimas D, King A, Mazumder R (2016) Best subset selection via a modern optimization lens. Ann Stat 44(2):813–852

Bertsimas D, Dunn J (2017) Optimal classification trees. Mach Learn 106(7):1039–1082

Bertsimas D, King A (2017) Logistic regression: from art to science. Stat Sci 32(3):367–384

Boyd S, Vandenberghe L (2004) Convex optimization. Cambridge University Press, Cambridge

Chung S, Park YW, Cheong T (2017) A mathematical programming approach for integrated multiple linear regression subset selection and validation. arXiv preprint arXiv:1712.04543

Colson B, Marcotte P, Savard G (2007) An overview of bilevel optimization. Ann Oper Res 153(1):235–256

Cozad A, Sahinidis NV, Miller DC (2014) Learning surrogate models for simulation-based optimization. AIChE J 60(6):2211–2227

Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. J Stat Softw 33(1):1–22

Geisser S (1975) The predictive sample reuse method with applications. J Am Stat Assoc 70(350):320–328

Hastie T, Tibshirani R, Tibshirani RJ (2017) Extended comparisons of best subset selection, forward stepwise selection, and the lasso. arXiv preprint arXiv:1707.08692

Hoerl AE, Kennard RW (1970) Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12(1):55–67
Hooker JN, Osorio MA (1999) Mixed logical-linear programming. Discrete Appl Math 96–97:395–442
Kimura K, Waki H (2018) Minimization of Akaike's information criterion in linear regression analysis via mixed integer nonlinear program. Optim Methods Softw 33(3):633–649
Konno H, Yamamoto R (2009) Choosing the best set of variables in regression analysis using integer programming. J Glob Optim 44(2):273–282
Kunapuli G, Bennett KP, Hu J, Pang JS (2008) Classification model selection via bilevel programming. Optim Methods Softw 23(4):475–489
Maldonado S, Pérez J, Weber R, Labbé M (2014) Feature selection for support vector machines via mixed integer linear programming. Inf Sci 279:163–175
Mallows CL (1973) Some comments on C_p. Technometrics 15(4):661–675
Miller A (2002) Subset selection in regression. Chapman and Hall, Boca Raton
Miyashiro R, Takano Y (2015a) Subset selection by Mallows' C_p: a mixed integer programming approach. Expert Syst Appl 42(1):325–331
Miyashiro R, Takano Y (2015b) Mixed integer second-order cone programming formulations for variable selection in linear regression. Eur J Oper Res 247(3):721–731
Mosier CI (1951) I. Problems and designs of cross-validation. Educ Psychol Meas 11(1):5–11
Naganuma M, Takano Y, Miyashiro R (2019) Feature subset selection for ordered logit model via tangent-plane-based approximation. IEICE Tran Inf Syst E102-D(5), 1046–1053
Okuno T, Takeda A, Kawana A (2018) Hyperparameter learning for bilevel nonsmooth optimization. arXiv preprint arXiv:1806.01520
Park YW, Klabjan D (2017) Subset selection for multiple linear regression via optimization. arXiv preprint arXiv:1701.07920
Pedregosa F (2016) Hyperparameter optimization with approximate gradient. In: Proceedings of the 33rd international conference on machine learning, pp 737–746
Sato T, Takano Y, Miyashiro R, Yoshise A (2016) Feature subset selection for logistic regression via mixed integer optimization. Comput Optim Appl 64(3):865–880
Sato T, Takano Y, Miyashiro R (2017) Piecewise-linear approximation for feature subset selection in a sequential logit model. J Oper Res Soc Jpn 60(1):1–14
Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6(2):461–464
Shao J (1993) Linear model selection by cross-validation. J Am Stat Assoc 88(422):486–494
Sinha A, Malo P, Deb K (2018) A review on bilevel optimization: from classical to evolutionary approaches and applications. IEEE Trans Evolut Comput 22(2):276–295
Stone M (1974) Cross-validatory choice and assessment of statistical predictions. J R Stat Soc Ser B Methodol 36(2):111–147
Tamura R, Kobayashi K, Takano Y, Miyashiro R, Nakata K, Matsui T (2017) Best subset selection for eliminating multicollinearity. J Oper Res Soc Jpn 60(3):321–336
Tamura R, Kobayashi K, Takano Y, Miyashiro R, Nakata K, Matsui T (2019) Mixed integer quadratic optimization formulations for eliminating multicollinearity based on variance inflation factor. J Glob Optim 73(2):431–446
Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc Ser B Methodol 58:267–288
Ustun B, Rudin C (2016) Supersparse linear integer models for optimized medical scoring systems. Mach Learn 102(3):349–391
van Rijsbergen CJ (1979) Information retrieval, 2nd edn. Butterworth-Heinemann, Oxford
Wherry R (1931) A new formula for predicting the shrinkage of the coefficient of multiple correlation. Ann Math Stat 2(4):440–457
Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J R Stat Soc Ser B (Stat Methodol) 67(2):301–320