Journal of Global Optimization (2019) 73:431–446
https://doi.org/10.1007/s10898-018-0713-3

Mixed integer quadratic optimization formulations for eliminating multicollinearity based on variance inflation factor

Ryuta Tamura[1,2] · Ken Kobayashi[3] · Yuichi Takano[4,5] · Ryuhei Miyashiro[6] · Kazuhide Nakata[7] · Tomomi Matsui[7]

Received: 7 October 2017 / Accepted: 11 October 2018 / Published online: 22 October 2018

Abstract
Multicollinearity exists when some explanatory variables of a multiple linear regression model are highly correlated. High correlation among explanatory variables reduces the reliability of the analysis. To eliminate multicollinearity from a linear regression model, we consider how to select a subset of significant variables by means of the variance inflation factor (VIF), which is the most common indicator used in detecting multicollinearity. In particular, we adopt the mixed integer optimization (MIO) approach to subset selection. The MIO approach was proposed in the 1970s, and recently it has received renewed attention due to advances in algorithms and hardware. However, none of the existing studies have developed a computationally tractable MIO formulation for eliminating multicollinearity on the basis of VIF. In this paper, we propose mixed integer quadratic optimization (MIQO) formulations for selecting the best subset of explanatory variables subject to the upper bounds on the VIFs of selected variables. Our two MIQO formulations are based on the two equivalent definitions of VIF. Computational results illustrate the effectiveness of our MIQO formula-

✉ Ryuhei Miyashiro
r-miya@cc.tuat.ac.jp

1 Graduate School of Engineering, Tokyo University of Agriculture and Technology, 2-24-16 Naka-cho, Koganei-shi, Tokyo 184-8588, Japan

2 Present Address: October Sky Co., Ltd., Zelkova Bldg., 1-25-12 Fuchucho, Fuchu-shi, Tokyo 183-0055, Japan

3 Artificial Intelligence Laboratory, Fujitsu Laboratories Ltd., 4-1-1 Kamikodanaka, Nakahara-ku, Kawasaki-shi, Kanagawa 211-8588, Japan

4 School of Network and Information, Senshu University, 2-1-1 Higashimita, Tama-ku, Kawasaki-shi, Kanagawa 214-8580, Japan

5 Present Address: Faculty of Engineering, Information and Systems, University of Tsukuba, 1-1-1 Tennodai, Tsukuba-shi, Ibaraki 305-8577, Japan

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

7 School of Engineering, Tokyo Institute of Technology, 2-12-1 Ookayama, Meguro-ku, Tokyo 152-8552, Japan

tions by comparison with conventional local search algorithms and MIO-based cutting plane algorithms.

Keywords Integer programming · Subset selection · Multicollinearity · Variance inflation factor · Multiple linear regression · Statistics

1 Introduction

Multiple regression analysis is a statistical process for estimating the relationship between explanatory and response variables. When some of explanatory variables are highly correlated, the reliability of the analysis is decreased because of the low quality of the resultant estimates. This problem is known as *multicollinearity* [4,11,12].

There are several approaches to avoiding the deleterious effects of multicollinearity, such as principal component regression [21,28], partial least squares regression [39,40], ridge regression [17], and subset selection [16,30]. This paper is focused on subset selection, a commonly used method for eliminating multicollinearity. Conventionally in this method, explanatory variables are removed one at a time on the basis of indicators for detecting multicollinearity, such as condition number of the correlation matrix and variance inflation factor (VIF) [11]. On the other hand, the potential disadvantage of this iterative procedure is that the best (e.g., in the least-squares sense) subset of variables is not necessarily found. More precisely, the iterative procedure may fail to provide an optimal solution to the following problem: Find a subset of variables that minimizes the residual sum of squares under the constraint that multicollinearity measured by the VIF is undetected when using that subset.

Multiple regression analysis has two primary purposes: prediction and description [18]. Subset selection is particularly beneficial for description purposes because it promotes better understanding of the causal relationships between explanatory and response variables. Various computational algorithms have been proposed for subset selection [10,14,23,26], and many of them are categorized as heuristic algorithms. However, these heuristic algorithms are often unsuitable for description purposes because they can yield low-quality solutions that lead to incorrect conclusions about causality.

For this reason, we adopt a mixed integer optimization (MIO) approach to subset selection. This approach was first proposed in the 1970s [1], and recently it has received renewed attention due to advances in algorithms and hardware [9,15,24,37,38]. In contrast to heuristic algorithms, the MIO approach has the potential to provide the best subset of variables with respect to several criterion functions, which include Mallows' C_p [31], adjusted R^2 [32], discrete Dantzig selector [29], and some information criteria [22,32]. Due to its usefulness and good performance, MIO-based subset selection has extended the range of applications to areas such as logistic regression [8,34], sequential logit models [35], support vector machines [27], cluster analysis [5], and classification trees [6].

To avoid multicollinearity in MIO-based subset selection, Bertsimas and King [7] suggested the use of a cutting plane algorithm, which iteratively adds valid inequalities to cut off sets of collinear variables. These valid inequalities can be strengthened by means of a local search algorithm [36]; however, the cutting plane algorithm must solve a series of MIO problems, each of which is NP-hard.

Meanwhile, Tamura et al. [36] devised a mixed integer semidefinite optimization (MISDO) formulation for subset selection to eliminate multicollinearity. In contrast with the cutting plane algorithm, this approach merely needs to solve a single MISDO problem. In this MISDO

formulation, however, only the condition number can be adopted as an indicator for detecting multicollinearity. Although VIF is better-grounded in statistical theory [4,11], to the best of our knowledge, none of the existing studies have developed a computationally tractable MIO formulation for eliminating multicollinearity on the basis of VIF.

The purpose of this paper is to devise mixed integer quadratic optimization (MIQO) formulations that can be used to select the best subset of explanatory variables subject to the upper bounds on the VIFs of selected variables. Our two MIQO formulations are based on two equivalent definitions of VIF. The effectiveness of our MIQO formulations is assessed through computational experiments using several datasets from the UCI Machine Learning Repository [25].

The main contributions of the present paper are as follows.

- We obtain computationally tractable MIQO formulations for best subset selection under the upper-bound constraints on VIF. Although the cutting plane algorithm was the only way to exactly solve this subset selection problem, we successfully reformulated the problem as a convex MIQO problem, which can be handled using standard MIO software.
- Our MIQO formulations are capable of verifying the optimality of the selected subset of variables. We verify through computational experiments that when the number of candidate explanatory variables is less than 30, our MIQO problems can be solved to optimality within a few tens of seconds.
- The proposed MIQO formulations provide solutions of good quality even if the computation is terminated before verifying optimality. The computational results demonstrate that even when the number of candidate explanatory variables is more than 30, our MIQO formulations with some preprocessing can find, within a time limit of 10,000 s, better subsets of variables than those obtained using local search algorithms.

2 Multiple linear regression and variance inflation factor

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

For simplicity of explanation, in Sects. 2 and 3 we assume that all explanatory and response variables are centered and scaled for unit length; that is,

$$\sum_{i=1}^{n} x_{ij} = \sum_{i=1}^{n} y_i = 0 \quad \text{and} \quad \sum_{i=1}^{n} (x_{ij})^2 = \sum_{i=1}^{n} (y_i)^2 = 1 \tag{1}$$

for all $j \in P$. The multiple linear regression model is then formulated as follows:

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

where $\boldsymbol{y} := (y_1, y_2, \ldots, y_n)^\top$, $\boldsymbol{a} := (a_1, a_2, \ldots, a_p)^\top$, $\boldsymbol{\varepsilon} := (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^\top$, and

$$\boldsymbol{X} := \left(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p\right) = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}.$$

Here, $\boldsymbol{a}$ is a vector of regression coefficients to be estimated, and $\boldsymbol{\varepsilon}$ is a vector composed of a prediction residual for each sample $i = 1, 2, \ldots, n$.

In what follows, we consider selecting a subset $S \subseteq P$ of explanatory variables to mitigate the negative influence of multicollinearity on regression estimates. On account of assumption (1), the correlation matrix of selected variables is calculated as

$$\boldsymbol{R}_S := (r_{j\ell})_{(j,\ell)\in S\times S} = \boldsymbol{X}_S^\top \boldsymbol{X}_S,$$

where $\boldsymbol{X}_S := (\boldsymbol{x}_j)_{j\in S}$ is the submatrix of $\boldsymbol{X}$ corresponding to the set S.

The variance inflation factor, VIF, for detecting multicollinearity is defined for each $\ell \in S$. Specifically, VIF of the ℓth explanatory variable is defined as the ℓth diagonal entry of the inverse of $\boldsymbol{R}_S$, that is,

$$\mathrm{VIF}(\ell, S) := [\boldsymbol{R}_S^{-1}]_{\ell\ell}. \tag{2}$$

When some of the selected variables are highly correlated, $\boldsymbol{R}_S$ is close to singular, and the corresponding VIF value is very large. Hence, the following upper-bound constraints should be imposed on the set S:

$$\mathrm{VIF}(\ell, S) \le \alpha \quad (\ell \in S), \tag{3}$$

where α is a user-defined parameter larger than one.

On the other hand, VIF has another easily interpretable definition. To describe this, we consider a linear regression model that explains the relationship between the ℓth explanatory variable and other variables in the set S,

$$\boldsymbol{x}_\ell = \boldsymbol{X}_{S\setminus\{\ell\}}\boldsymbol{a}^{(\ell,S)} + \boldsymbol{\varepsilon}^{(\ell,S)}, \tag{4}$$

where $\boldsymbol{a}^{(\ell,S)} \in \mathbb{R}^{|S|-1}$ and $\boldsymbol{\varepsilon}^{(\ell,S)} \in \mathbb{R}^n$ are vectors of regression coefficients and residuals.

To estimate the regression coefficients, $\boldsymbol{a}^{(\ell,S)}$, the ordinary least squares (OLS) method minimizes the residual sum of squares (RSS),

$$\|\boldsymbol{x}_\ell - \boldsymbol{X}_{S\setminus\{\ell\}}\boldsymbol{a}^{(\ell,S)}\|_2^2 = (\boldsymbol{x}_\ell - \boldsymbol{X}_{S\setminus\{\ell\}}\boldsymbol{a}^{(\ell,S)})^\top(\boldsymbol{x}_\ell - \boldsymbol{X}_{S\setminus\{\ell\}}\boldsymbol{a}^{(\ell,S)}). \tag{5}$$

This is equivalent to solving the well-known normal equation:

$$\boldsymbol{X}_{S\setminus\{\ell\}}^\top \boldsymbol{X}_{S\setminus\{\ell\}}\hat{\boldsymbol{a}}^{(\ell,S)} = \boldsymbol{X}_{S\setminus\{\ell\}}^\top \boldsymbol{x}_\ell, \tag{6}$$

where $\hat{\boldsymbol{a}}^{(\ell,S)}$ is called the OLS estimator.

The goodness-of-fit of regression model (4) is measured by the coefficient of determination. Due to assumption (1), it is calculated based on the OLS estimator as follows:

$$R^2(\ell, S) := 1 - \|\boldsymbol{x}_\ell - \boldsymbol{X}_{S\setminus\{\ell\}}\hat{\boldsymbol{a}}^{(\ell,S)}\|_2^2.$$

When $R^2(\ell, S)$ is close to one, the ℓth explanatory variable has a strong linear relationship with other variables in the set S. It is known that VIF of the ℓth explanatory variable can also be defined as follows [4,12]:

$$\mathrm{VIF}(\ell, S) := \frac{1}{1 - R^2(\ell, S)} = \frac{1}{\|\boldsymbol{x}_\ell - \boldsymbol{X}_{S\setminus\{\ell\}}\hat{\boldsymbol{a}}^{(\ell,S)}\|_2^2}. \tag{7}$$

Here we briefly explain an advantage to using VIF over using the condition number as an indicator of multicollinearity. To eliminate multicollinearity, we select explanatory variables and construct a model in which VIF or the condition number does not exceed a specified upper bound; 10 and 225, respectively, are frequently used as upper bounds for VIF and the condition number [11]. From a modeling perspective, we can directly control the upper bound of the degree of multicollinearity by using VIF. For example, when we need to tighten

the upper bound of $R^2(\ell, S)$ from 0.9 to 0.8, we change the upper bound of $\mathrm{VIF}(\ell, S)$ from $1/(1-0.9) = 10$ to $1/(1-0.8) = 5$ due to the definition (7). On the other hand, such control is not obvious when using the condition number.

3 Mixed integer quadratic optimization formulations

In this section, we consider minimizing RSS of a subset regression model under the upper-bound constraints (3) on VIFs. Let $\boldsymbol{z} := (z_1, z_2, \ldots, z_p)^\top$ be a vector of 0–1 decision variables for subset selection. Accordingly, $S(\boldsymbol{z}) := \{j \in P \mid z_j = 1\}$ is a selected subset of explanatory variables. The subset selection problem for eliminating multicollinearity based on VIF is posed as an MIO problem:

$$\text{minimize} \quad \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{a}\|_2^2 \tag{8}$$

$$\text{subject to} \quad z_j = 0 \;\Rightarrow\; a_j = 0 \quad (j \in P), \tag{9}$$

$$z_\ell = 1 \;\Rightarrow\; \mathrm{VIF}(\ell, S(\boldsymbol{z})) \le \alpha \quad (\ell \in P), \tag{10}$$

$$\boldsymbol{a} \in \mathbb{R}^p,\ \boldsymbol{z} \in \{0,1\}^p. \tag{11}$$

If $z_j = 0$, then the jth explanatory variable is deleted from the regression model because its coefficient is set to zero by the logical implications (9). The VIF constraints (3) are imposed in the form of logical implications (10). It is known that these logical implications can be represented by using a big-M method or a special ordered set type 1 (SOS 1) constraint [2,3].

However, other than logical implications, constraints (10) still contain difficulty to be handled in an MIO formulation. In the following two subsections, we derive two tractable formulations of these VIF constraints.

3.1 Normal-equation-based formulation

We first propose an MIQO formulation based on the definition (7) of VIF. Let us introduce a vector of decision variables, $\boldsymbol{a}^{(\ell)} := (a_j^{(\ell)})_{j \in P\setminus\{\ell\}} \in \mathbb{R}^{p-1}$ $(\ell \in P)$. To convert the VIF constraints (10) into a set of linear constraints, we exploit the normal-equation-based constraints proposed by Tamura et al. [36]:

$$z_j = 1 \;\Rightarrow\; \boldsymbol{x}_j^\top \boldsymbol{X}_{P\setminus\{\ell\}} \boldsymbol{a}^{(\ell)} = \boldsymbol{x}_j^\top \boldsymbol{x}_\ell \quad (j \in P \setminus \{\ell\}), \tag{12}$$

$$z_j = 0 \;\Rightarrow\; a_j^{(\ell)} = 0 \quad (j \in P \setminus \{\ell\}). \tag{13}$$

Theorem 1 *Suppose that $(\boldsymbol{a}^{(\ell)}, \boldsymbol{z}) \in \mathbb{R}^{p-1} \times \{0,1\}^p$ satisfies constraints (12)–(13). Then, we have*

$$\mathrm{VIF}(\ell, S(\boldsymbol{z})) = \frac{1}{\|\boldsymbol{x}_\ell - \boldsymbol{X}_{P\setminus\{\ell\}} \boldsymbol{a}^{(\ell)}\|_2^2} \quad (\ell \in S(\boldsymbol{z})).$$

Proof Let s be the number of nonzero elements of $\boldsymbol{z}$. Without loss of generality, we may assume that $S(\boldsymbol{z}) = \{1, 2, \ldots, s\}$. According to $S(\boldsymbol{z})$, we partition $\boldsymbol{a}^{(\ell)}$ as

$$\boldsymbol{a}^{(\ell)} = \begin{pmatrix} \boldsymbol{a}_1^{(\ell)} \\ \boldsymbol{a}_2^{(\ell)} \end{pmatrix},\ \boldsymbol{a}_1^{(\ell)} \in \mathbb{R}^{s-1},\ \boldsymbol{a}_2^{(\ell)} \in \mathbb{R}^{p-s},$$

where $\boldsymbol{a}_2^{(\ell)} = \boldsymbol{0}$ due to constraints (13). Therefore constraints (12) correspond to the normal equation (6) for $S = S(\boldsymbol{z})$; that is,

$$\boldsymbol{X}_{S(z)\setminus\{\ell\}}^\top \boldsymbol{X}_{S(z)\setminus\{\ell\}} \boldsymbol{a}_1^{(\ell)} = \boldsymbol{X}_{S(z)\setminus\{\ell\}}^\top \boldsymbol{x}_\ell. \tag{14}$$

Since the OLS estimator provides the minimum value of RSS (5), it follows that

$$\|\boldsymbol{x}_\ell - \boldsymbol{X}_{S(z)\setminus\{\ell\}} \hat{\boldsymbol{a}}^{(\ell,S(z))}\|_2^2 = \|\boldsymbol{x}_\ell - \boldsymbol{X}_{S(z)\setminus\{\ell\}} \boldsymbol{a}_1^{(\ell)}\|_2^2 = \|\boldsymbol{x}_\ell - \boldsymbol{X}_{P\setminus\{\ell\}} \boldsymbol{a}^{(\ell)}\|_2^2.$$

Hence, the definition (7) of VIF completes the proof. □

The VIF constraints (10) can be rewritten by Theorem 1 as follows:

$$z_\ell \le \alpha \|\boldsymbol{x}_\ell - \boldsymbol{X}_{P\setminus\{\ell\}} \boldsymbol{a}^{(\ell)}\|_2^2 \quad (\ell \in P).$$

However, these are reverse convex constraints, which are very difficult to handle in an MIO formulation. To resolve this reverse-convexity, we exploit the normal equation (14) as follows:

$$\begin{aligned}
&\|\boldsymbol{x}_\ell - \boldsymbol{X}_{P\setminus\{\ell\}} \boldsymbol{a}^{(\ell)}\|_2^2 \\
&= \boldsymbol{x}_\ell^\top \boldsymbol{x}_\ell - 2\boldsymbol{x}_\ell^\top \boldsymbol{X}_{P\setminus\{\ell\}} \boldsymbol{a}^{(\ell)} + (\boldsymbol{a}_1^{(\ell)})^\top \boldsymbol{X}_{S(z)\setminus\{\ell\}}^\top \boldsymbol{X}_{S(z)\setminus\{\ell\}} \boldsymbol{a}_1^{(\ell)} \\
&= \boldsymbol{x}_\ell^\top \boldsymbol{x}_\ell - 2\boldsymbol{x}_\ell^\top \boldsymbol{X}_{P\setminus\{\ell\}} \boldsymbol{a}^{(\ell)} + (\boldsymbol{a}_1^{(\ell)})^\top \boldsymbol{X}_{S(z)\setminus\{\ell\}}^\top \boldsymbol{x}_\ell \\
&= \boldsymbol{x}_\ell^\top \boldsymbol{x}_\ell - \boldsymbol{x}_\ell^\top \boldsymbol{X}_{P\setminus\{\ell\}} \boldsymbol{a}^{(\ell)}.
\end{aligned} \tag{15}$$

Consequently, the subset selection problem (8)–(11) can be formulated as an MIQO problem, which we call the normal-equation-based formulation:

$$\begin{aligned}
\text{minimize} \quad & \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{a}\|_2^2 && (16)\\
\text{subject to} \quad & z_j = 0 \;\Rightarrow\; a_j = 0 \quad (j \in P), && (17)\\
& z_\ell \le \alpha(\boldsymbol{x}_\ell^\top \boldsymbol{x}_\ell - \boldsymbol{x}_\ell^\top \boldsymbol{X}_{P\setminus\{\ell\}} \boldsymbol{a}^{(\ell)}) \quad (\ell \in P), && (18)\\
& z_j = 1 \;\Rightarrow\; \boldsymbol{x}_j^\top \boldsymbol{X}_{P\setminus\{\ell\}} \boldsymbol{a}^{(\ell)} = \boldsymbol{x}_j^\top \boldsymbol{x}_\ell \quad (\ell \in P,\ j \in P\setminus\{\ell\}), && (19)\\
& z_j = 0 \;\Rightarrow\; a_j^{(\ell)} = 0 \quad (\ell \in P,\ j \in P\setminus\{\ell\}), && (20)\\
& \boldsymbol{a} \in \mathbb{R}^p,\ \boldsymbol{z} \in \{0,1\}^p,\ \boldsymbol{a}^{(\ell)} \in \mathbb{R}^{p-1} \quad (\ell \in P). && (21)
\end{aligned}$$

3.2 Inverse-matrix-based formulation

We next propose another MIQO formulation based on the definition (2) of VIF. Let us introduce square matrices of decision variables, $\boldsymbol{Q} := (q_{\ell j})_{(\ell,j)\in P\times P}$ and $\boldsymbol{U} := (u_{\ell j})_{(\ell,j)\in P\times P}$. To compute the inverse of the correlation matrix $\boldsymbol{R}_{S(z)}$, we make use of the following constraints:

$$\begin{aligned}
& \boldsymbol{Q}\boldsymbol{R}_P + \boldsymbol{U} = \boldsymbol{I}_p, && (22)\\
& z_j = 1 \;\Rightarrow\; u_{\ell j} = 0 \quad (\ell \in P,\ j \in P), && (23)\\
& z_j = 0 \;\Rightarrow\; q_{\ell j} = q_{j\ell} = 0 \quad (\ell \in P,\ j \in P), && (24)
\end{aligned}$$

where $\boldsymbol{I}_p$ is the identity matrix of size p. To promote an understanding of constraints (22), we consider a case in which all candidate explanatory variables are selected (i.e., $z_j = 1$ $(j \in P)$). In this case, $\boldsymbol{U}$ becomes the zero matrix because of constraints (23). Consequently, we have $\boldsymbol{Q}\boldsymbol{R}_P = \boldsymbol{I}_p$; that is, $\boldsymbol{Q}$ is the inverse of the correlation matrix of all candidate explanatory variables.

Theorem 2 *Suppose that* $(\boldsymbol{Q}, \boldsymbol{U}, \boldsymbol{z}) \in \mathbb{R}^{p\times p} \times \mathbb{R}^{p\times p} \times \{0,1\}^p$ *satisfies constraints (22)–(24). Then, we have* $q_{\ell\ell} = [\boldsymbol{R}_{S(\boldsymbol{z})}^{-1}]_{\ell\ell}$ *for* $\ell \in S(\boldsymbol{z})$, *and* $q_{\ell\ell} = 0$ *for* $\ell \notin S(\boldsymbol{z})$.

Proof Similarly to Theorem 1, we may assume without loss of generality that $S(\boldsymbol{z}) = \{1, 2, \ldots, s\}$. We partition $\boldsymbol{Q}, \boldsymbol{R}_P, \boldsymbol{U}$ and $\boldsymbol{I}_p$ according to $S(\boldsymbol{z})$ and rewrite constraints (22) as follows:

$$\begin{pmatrix} \boldsymbol{Q}_1 & \boldsymbol{Q}_2 \\ \boldsymbol{Q}_3 & \boldsymbol{Q}_4 \end{pmatrix} \begin{pmatrix} \boldsymbol{R}_1 & \boldsymbol{R}_2 \\ \boldsymbol{R}_3 & \boldsymbol{R}_4 \end{pmatrix} + \begin{pmatrix} \boldsymbol{U}_1 & \boldsymbol{U}_2 \\ \boldsymbol{U}_3 & \boldsymbol{U}_4 \end{pmatrix} = \begin{pmatrix} \boldsymbol{I}_s & \boldsymbol{O} \\ \boldsymbol{O}^\top & \boldsymbol{I}_{p-s} \end{pmatrix},$$

where $\boldsymbol{Q}_1, \boldsymbol{R}_1, \boldsymbol{U}_1 \in \mathbb{R}^{s\times s}$, $\boldsymbol{Q}_2, \boldsymbol{R}_2, \boldsymbol{U}_2 \in \mathbb{R}^{s\times(p-s)}$, $\boldsymbol{Q}_3, \boldsymbol{R}_3, \boldsymbol{U}_3 \in \mathbb{R}^{(p-s)\times s}$, $\boldsymbol{Q}_4, \boldsymbol{R}_4, \boldsymbol{U}_4 \in \mathbb{R}^{(p-s)\times(p-s)}$, and $\boldsymbol{O}$ is the zero matrix of size $s \times (p-s)$. Note here that $\boldsymbol{Q}_2$, $\boldsymbol{Q}_3$, $\boldsymbol{Q}_4$, $\boldsymbol{U}_1$ and $\boldsymbol{U}_3$ become zero matrices due to constraints (23)–(24). As a result, the above constraints are reduced to

$$\begin{pmatrix} \boldsymbol{Q}_1\boldsymbol{R}_1 \\ \boldsymbol{O}^\top \end{pmatrix} = \begin{pmatrix} \boldsymbol{I}_s \\ \boldsymbol{O}^\top \end{pmatrix}, \tag{25}$$

while other constraints are satisfied through free decision variables in $\boldsymbol{U}_2$ and $\boldsymbol{U}_4$. Since $\boldsymbol{R}_1 = \boldsymbol{R}_{S(\boldsymbol{z})}$, it follows that $\boldsymbol{Q}_1 = \boldsymbol{R}_{S(\boldsymbol{z})}^{-1}$, which completes the proof. □

Using the definition (2), the subset selection problem (8)–(11) is reformulated as an MIQO problem, which we call the inverse-matrix-based formulation:

$$\begin{aligned}
\text{minimize} \quad & \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{a}\|_2^2 && (26)\\
\text{subject to} \quad & z_j = 0 \;\Rightarrow\; a_j = 0 \quad (j \in P), && (27)\\
& q_{\ell\ell} \le \alpha \quad (\ell \in P), && (28)\\
& \boldsymbol{Q}\boldsymbol{R}_P + \boldsymbol{U} = \boldsymbol{I}_p, && (29)\\
& \boldsymbol{Q} = \boldsymbol{Q}^\top, && (30)\\
& z_j = 1 \;\Rightarrow\; u_{\ell j} = 0 \quad (\ell \in P,\ j \in P), && (31)\\
& z_j = 0 \;\Rightarrow\; q_{\ell j} = 0 \quad (\ell \in P,\ j \in P), && (32)\\
& \boldsymbol{a} \in \mathbb{R}^p,\ \boldsymbol{Q} \in \mathbb{R}^{p\times p},\ \boldsymbol{U} \in \mathbb{R}^{p\times p},\ \boldsymbol{z} \in \{0,1\}^p. && (33)
\end{aligned}$$

Note that the correlation matrix is symmetric, so is its inverse; thus, constraint (30) is redundant. However, we explicitly add the constraint because it improves computational speed.

3.3 Preprocessing for faster computation

In this subsection, we propose some ideas for speeding up the MIQO computation. In our preliminary experiments, however, preprocessing (iii) and (iv) did not shorten the computation time; hence, we will evaluate the efficiency of preprocessing steps (i) and (ii) in the next section.

Preprocessing (i): Deleting redundant VIF constraints The definition (7) of VIF implies that $\mathrm{VIF}(\ell, S) \le \mathrm{VIF}(\ell, P)$ for all $S \subseteq P$. Therefore, the VIF constraints for $\ell \in P_0$ can be deleted, with

$$P_0 := \{\ell \in P \mid \mathrm{VIF}(\ell, P) \le \alpha\}. \tag{34}$$

Table 1 List of instances

Abbreviation	n	p	Original dataset [25]
`Servo`	167	19	Servo
`AutoMPG`	392	25	Auto MPG
`SolarFlareC`	1066	26	Solar flare (C-class flares production)
`BreastCancer`	194	32	Breast cancer Wisconsin
`Automobile`	159	65	Automobile
`Crime`	1993	100	Communities and crime

Preprocessing (ii): Adding cutting-plane-based constraints This step is the following. First, find subsets $S_k \subseteq P$ $(k \in K)$ of collinear variables such that $\mathrm{VIF}(\ell, S_k) > \alpha$ for some $\ell \in S_k$. Next, cut them off by means of the following cutting-plane-based constraints [7,36]:

$$\sum_{j \in S_k} z_j \leq |S_k| - 1 \quad (k \in K). \tag{35}$$

Preprocessing (iii): Tightening constraints Constraints (18) can be tightened by using the minimum RSS (5) of $S = P$ as follows:

$$z_\ell + (1 - z_\ell)\alpha \|\boldsymbol{x}_\ell - \boldsymbol{X}_{P\setminus\{\ell\}} \hat{\boldsymbol{a}}^{(\ell,P)}\|_2^2 \leq \alpha(\boldsymbol{x}_\ell^\top \boldsymbol{x}_\ell - \boldsymbol{x}_\ell^\top \boldsymbol{X}_{P\setminus\{\ell\}} \boldsymbol{a}^{(\ell)}) \quad (\ell \in P).$$

Preprocessing (iv): Linearization of the objective function The objective function can be linearized by applying the transformation (15) to $\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{a}\|_2^2$, which changes the proposed MIQO formulations into mixed integer linear optimization formulations.

4 Computational results

This section evaluates the computational performance of our MIQO formulations for subset selection to eliminate multicollinearity as characterized by VIF.

We downloaded six datasets for regression analysis from the UCI Machine Learning Repository [25]. Table 1 lists the instances used for computational experiments, where n and p are the numbers of samples and candidate explanatory variables, respectively. In the `SolarFlareC` instance, C-class flares production was employed as a response variable. Categorical variables were encoded into sets of dummy variables. Samples containing missing values were removed, and then redundant variables (i.e., those with the same value in all samples) were also removed.

We compare the computational performance of the following subset selection algorithms.

- FwS: Forward selection method: Starts with $S = \emptyset$ and iteratively adds the jth variable (i.e., $S \leftarrow S \cup \{j\}$) that decreases RSS the most; this operation is repeated while the VIF constraints (3) are satisfied.
- BwE: Backward elimination method: Starts with $S = P$ and iteratively eliminates the jth variable (i.e., $S \leftarrow S \setminus \{j\}$) that increases RSS the most; this operation is repeated until the VIF constraints (3) are satisfied.
- CPA: Cutting plane algorithm [7,36]: Solves the MIQO problem (8)–(9) and (11), and adds a valid inequality (cf. (35)) to the problem to cut off subsets of collinear variables; this operation is repeated until the VIF constraints (3) hold.

Table 2 Results of preprocessing ($\alpha = 5$)

Instance	n	p	Preprocessing (i)		Preprocessing (ii)	
			$\lvert P_0 \rvert$	Time (s)	$\lvert K \rvert$	Time (s)
`Servo`	167	19	0	0.06	7	1.80
`AutoMPG`	392	25	1	0.09	19	9.03
`SolarFlareC`	1066	26	4	0.14	18	8.37
`BreastCancer`	194	32	2	0.12	22	7.12
`Automobile`	159	65	1	0.47	56	79.90
`Crime`	1993	100	22	3.46	61	1354.53

CPA*: Cutting plane algorithm in which valid inequalities are strengthened by using a backward elimination method [36].
NEF: Normal-equation-based MIQO formulation (16)–(21).
NEF+: Normal-equation-based MIQO formulation (16)–(21) with the preprocessing (i) and (ii).
IMF: Inverse-matrix-based MIQO formulation (26)–(33).
IMF+: Inverse-matrix-based MIQO formulation (26)–(33) with the preprocessing (i) and (ii).

These computations were performed on a Windows 7 PC with an Intel Core i7-4770 CPU (3.40 GHz) and 8 GB memory. The algorithms FwS and BwE were implemented with R 3.1.1 [33]; the algorithms CPA and CPA* were implemented in Python 2.7 with Gurobi Optimizer 7.5.2 [13], where the logical implications (9) were incorporated in the form of SOS 1 constraint as in Tamura et al. [36]. The MIQO problems (i.e., NEF, NEF+, IMF, and IMF+) were solved using IBM ILOG CPLEX 12.6.3.0 [19] with eight threads. Here the `indicator` function implemented in CPLEX was used to impose the logical implications (17), (19), (20), (27), (31), and (32).

A subset of explanatory variables often has collinearity problems when its VIF value is greater than 5 or 10 (see [20], p. 101). Thus, we tested two values ($\alpha = 5$ and 10) of the upper bound on VIF to evaluate the effect on computational results.

The algorithms NEF+ and IMF+ involved preprocessing steps (i) and (ii), as explained in Sect. 3.3. Specifically, the VIF constraints for the set (34) were deleted in advance, and the cutting-plane-based constraints (35) were included. Here, the subsets S_k ($k \in K$) of collinear variables were found by applying the algorithm FwS to the regression model (4) for each $\ell \in P$.

Results of the preprocessing are summarized in Table 2 ($\alpha = 5$) and Table 3 ($\alpha = 10$), where $|P_0|$ and $|K|$ are the numbers of redundant VIF constraints and cutting-plane-based constraints, respectively. The column labeled "Time (s)" shows the computation time in seconds. We can see that preprocessing (i) required only a few seconds, but the computation time of preprocessing (ii) increased greatly with the number of candidate explanatory variables.

We evaluate the results of the subset selection algorithms on the basis of the following evaluation criteria.

– VIF: must be smaller than the specified upper bound.
  Note that if VIF is minimized as the objective function, then the optimal solution becomes a trivial and meaningless one, that is, a subset consisting of only one explanatory variable. Therefore, VIF should not be treated as the objective function but should be kept small by means of the upper-bound constraints.

Table 3 Results of preprocessing ($\alpha = 10$)

Instance	n	p	Preprocessing (i)		Preprocessing (ii)	
			$\lvert P_0 \rvert$	Time (s)	$\lvert K \rvert$	Time (s)
`Servo`	167	19	0	0.05	9	2.49
`AutoMPG`	392	25	1	0.10	20	9.33
`SolarFlareC`	1066	26	4	0.16	17	8.65
`BreastCancer`	194	32	8	0.12	22	15.73
`Automobile`	159	65	5	0.49	58	126.23
`Crime`	1993	100	35	3.57	54	2099.16

- Computation time: lower is better. Several days of computation are rarely acceptable for subset selection because it needs to be performed repeatedly (e.g., for cross-validation) in many practical situations. In addition, we are faced with a growing number of samples and explanatory variables because of recent advances in information technology. As a result, a faster algorithm is required for subset selection.
- R^2 value: higher is better. The goodness-of-fit of regression model is quantified by the coefficient of determination, R^2. Accordingly, an algorithm that produces a larger R^2 value is desirable.
- Implementation: simpler is better. If two algorithms offer a similar performance, an easily implementable one is highly preferred.

Tables 4 and 5 show the computational results of the subset selection algorithms for $\alpha = 5$ and 10, respectively. The column labeled "R^2" shows the value of the coefficient of determination of a subset regression model; the largest R^2 values for each instance are indicated in bold. The column labeled "$\mathrm{VIF}_{\max}$" shows the value of $\max\{\mathrm{VIF}(\ell, S) \mid \ell \in S\}$, and the column labeled "$|S|$" shows the number of selected explanatory variables. The computation of each subset selection algorithm was terminated if it did not finish by itself within 10,000 s. In these cases, the best feasible solution obtained within 10,000 s was taken as the result.

First, we can confirm that all solutions satisfy the constraint $\mathrm{VIF}_{\max} \leq \alpha$ unless no feasible solution was found within 10,000 s. Note that if a solution satisfies the VIF constraint, the value itself of $\mathrm{VIF}_{\max}$ does not matter.

Next, we discuss the computation time and R^2 values of each algorithm. Note that FwS and BwE are local search algorithms, and thus they complete the search process quickly without certificates of optimality of the obtained solutions. Meanwhile, other MIO approaches require a sufficient amount of computation time to verify optimality of the obtained solutions. Nevertheless, most of the MIQO problems for the `Servo`, `AutoMPG`, and `SolarFlareC` instances were solved to optimality within a few tens of seconds. In this case, R^2 values of our MIQO formulations were always the largest because the obtained solutions were verified to be optimal. We can also see that our preprocessing significantly reduced the time used in solving the MIQO problems. For instance, in Table 4, IMF required a relatively long computation time for the `AutoMPG` and `SolarFlareC` instances, but the computation of IMF+ finished much earlier for the same instances. In the case of the `BreastCancer` instance, the computation was completed about 20 times faster by IMF+ than by IMF for both $\alpha = 5$ and 10.

The MIQO computations for the `Automobile` and `Crime` instances were terminated due to the time limit of 10,000 s; nevertheless, they successfully found solutions of good

Table 4 Results of subset selection algorithms ($\alpha = 5$)

Instance	n	p	Method	R^2	$\text{VIF}_{\max}$	$\lvert S\rvert$	Time (s)
Servo	167	19	FwS	**0.75600**	3.604	13	0.47
			BwE	0.75482	3.194	13	0.41
			CPA	**0.75600**	3.604	13	589.61
			CPA*	**0.75600**	3.604	13	13.99
			NEF	**0.75600**	3.604	13	40.61
			NEF+	**0.75600**	3.604	13	12.81
			IMF	**0.75600**	3.604	13	29.06
			IMF+	**0.75600**	3.604	13	6.93
AutoMPG	392	25	FwS	0.86606	3.720	19	0.95
			BwE	0.86521	1.751	14	0.81
			CPA	–	–	–	> 10000.00
			CPA*	**0.87082**	2.373	19	9.42
			NEF	**0.87082**	2.224	19	43.65
			NEF+	**0.87082**	2.508	19	11.67
			IMF	**0.87082**	1.998	19	> 10000.00
			IMF+	**0.87082**	2.554	19	12.45
SolarFlareC	1066	26	FwS	0.19713	3.083	19	1.23
			BwE	0.17538	2.100	8	1.54
			CPA	**0.19715**	2.874	19	4802.29
			CPA*	**0.19715**	4.348	19	118.13
			NEF	**0.19715**	4.348	19	9.64
			NEF+	**0.19715**	4.348	19	1.72
			IMF	**0.19715**	4.348	19	2185.79
			IMF+	**0.19715**	3.505	19	2.34
BreastCancer	194	32	FwS	0.26848	4.984	14	1.02
			BwE	0.24493	3.062	7	1.84
			CPA	–	–	–	> 10000.00
			CPA*	**0.28192**	4.993	14	809.54
			NEF	**0.28192**	4.994	14	> 10000.00
			NEF+	**0.28192**	4.994	14	2246.02
			IMF	**0.28192**	4.994	14	> 10000.00
			IMF+	**0.28192**	4.994	14	536.07
Automobile	159	65	FwS	0.93659	4.990	16	2.59
			BwE	0.91369	1.596	10	12.95
			CPA	–	–	–	> 10000.00
			CPA*	0.96098	4.889	19	> 10000.00
			NEF	0.96110	4.994	33	> 10000.00
			NEF+	0.95745	4.591	31	> 10000.00
			IMF	0.90508	4.939	19	> 10000.00
			IMF+	**0.96626**	4.766	34	> 10000.00
Crime	1993	100	FwS	0.66248	4.996	21	11.44
			BwE	0.64681	4.989	6	115.08

Table 4 continued

Instance	n	p	Method	R^2	$\text{VIF}_{\max}$	$\lvert S\rvert$	Time (s)
			CPA	–	–	–	> 10000.00
			CPA*	0.65872	4.762	13	> 10000.00
			NEF	0.66070	4.995	26	> 10000.00
			NEF+	**0.66456**	4.880	30	> 10000.00
			IMF	0	–	0	> 10000.00
			IMF+	0.65161	4.726	20	> 10000.00

Table 5 Results of subset selection algorithms ($\alpha = 10$)

Instance	n	p	Method	R^2	$\text{VIF}_{\max}$	$\lvert S\rvert$	Time (s)
`Servo`	167	19	FwS	0.75862	8.677	14	0.42
			BwE	0.75862	8.677	13	0.28
			CPA	**0.75877**	8.741	15	3.56
			CPA*	**0.75877**	8.503	15	0.56
			NEF	**0.75877**	8.741	15	36.18
			NEF+	**0.75877**	8.741	15	30.41
			IMF	**0.76877**	8.741	15	29.67
			IMF+	**0.75877**	8.741	15	21.54
`AutoMPG`	392	25	FwS	**0.87334**	9.549	20	0.92
			BwE	0.87149	5.899	16	0.75
			CPA	**0.87334**	8.523	20	875.00
			CPA*	**0.87334**	8.523	20	16.18
			NEF	**0.87334**	8.523	20	13.12
			NEF+	**0.87334**	8.523	20	1.83
			IMF	**0.87334**	8.523	20	21.09
			IMF+	**0.87334**	8.523	20	1.72
`SolarFlareC`	1066	26	FwS	0.19713	3.083	19	1.28
			BwE	0.18232	7.661	9	1.50
			CPA	**0.19715**	2.874	19	4830.02
			CPA*	**0.19715**	5.989	19	214.53
			NEF	**0.19715**	4.348	19	9.86
			NEF+	**0.19715**	4.348	19	1.93
			IMF	**0.19715**	4.348	19	12.93
			IMF+	**0.19715**	9.102	19	2.00
`BreastCancer`	194	32	FwS	0.27039	9.981	16	1.24
			BwE	0.25424	9.973	8	1.81
			CPA	–	–	–	>10000.00
			CPA*	**0.29158**	9.765	16	1197.84
			NEF	**0.29158**	9.765	16	>10000.00
			NEF+	**0.29158**	9.765	16	1523.34
			IMF	**0.29158**	9.765	16	>10000.00
			IMF+	**0.29158**	9.765	16	499.92

Table 5 continued

Instance	n	p	Method	R^2	$\text{VIF}_{\max}$	$\lvert S\rvert$	Time (s)
`Automobile`	159	65	FwS	0.96605	9.996	31	5.25
			BwE	0.91367	1.596	10	13.01
			CPA	–	–	–	>10000.00
			CPA*	**0.96984**	9.912	31	>10000.00
			NEF	0.96281	9.411	32	>10000.00
			NEF+	0.96568	8.679	29	>10000.00
			IMF	0.96626	9.923	41	>10000.00
			IMF+	0.96970	9.937	43	>10000.00
`Crime`	1993	100	FwS	0.66906	9.999	24	13.12
			BwE	0.64953	5.773	7	115.52
			CPA	–	–	–	>10000.00
			CPA*	0.66946	9.766	21	>10000.00
			NEF	0.67444	9.989	40	>10000.00
			NEF+	0.67587	9.934	49	>10000.00
			IMF	0	–	0	>10000.00
			IMF+	**0.67660**	9.988	45	>10000.00

Table 6 Number of times the best solutions were generated

α	FwS	BwE	CPA	CPA*	NEF	NEF+	IMF	IMF+
5	1	0	2	4	4	5	4	5
10	1	0	3	5	4	4	4	5
Total	2	0	5	9	8	9	8	10

quality within 10,000 s. On the other hand, IMF failed to deliver a solution of good quality to the `Crime` instance; specifically, it found only the feasible solution $S = \emptyset$ within 10,000 s for $\alpha = 5$ and 10. This issue was observed due to numerical instability in IMF, whereas it did not happen in IMF+.

The algorithm CPA took a long time to solve even small instances (e.g., `AutoMPG` and `SolarFlareC`), and it failed to find a feasible solution to the `BreastCancer`, `Automobile`, and `Crime` instances within the time limit of 10,000 s. The algorithm CPA* is much faster than CPA, but it seems slower than IMF+, which is the fastest of four MIQO formulations (NEF, NEF+, IMF, and IMF+).

Table 6 shows the number of times the best solution was generated by each algorithm. We can see that CPA* was competitive with our MIQO formulations in terms of solution quality. Here, we also take into account the simplicity of implementation of these algorithms. The implementation of CPA* is relatively complicated; MIQO problems combined with the backward elimination method must be solved repeatedly. On the other hand, our MIQO formulations enable us to pose the subset selection problem as a single MIQO problem, which can be handled using standard MIO software. From the point of view of implementation, our MIQO formulations have a crucial advantage over the cutting plane algorithm.

We conclude this section by comparing our MIQO formulations with the MISDO formulation devised by Tamura et al. [36] for subset selection with the condition number constraint.

The MISDO formulation spent 5563.34 s ($\kappa = 100$) and 336.14 s ($\kappa = 225$) for the `AutoMPG` instance, where κ is the upper bound on the condition number. These results suggest that our VIF-based MIQO formulations had a clear computational advantage over the condition-number-based MISDO formulation. In addition, Tamura et al. [36] could not solve many of the MISDO problems for larger instances because of numerical instability.

5 Conclusions

This paper dealt with the problem of selecting the best subset of explanatory variables under upper-bound constraints on VIFs of selected variables. For this problem, we presented two MIQO formulations (a normal-equation-based formulation and an inverse-matrix-based formulation), based on two equivalent definitions of VIF.

The research contribution in this paper is computationally tractable MIQO formulations for eliminating multicollinearity by using VIF. Previously, no tractable formulation of the VIF constraint was known, and we have successfully written it as a set of linear constraints. As a result, we reduced the subset selection to a single MIO problem, which can be handled using standard MIO software.

Our MIQO formulations must spend a certain amount of time to verify optimality of the selected subset of variables. Nevertheless, it was demonstrated that when the number of candidate explanatory variables was less than 30, most of the MIQO problems were solved to optimality within a few tens of seconds. Our MIQO formulations also have the advantage of being able to find solutions of good quality in the early stage of the computation. Indeed, even when the number of candidate explanatory variables was more than 30, our MIQO formulations with preprocessing provided solutions of better quality within the time limit than those obtained using local search algorithms. These results reveal that our method is particularly effective for small-to-medium-sized problems (e.g., $p \leq 100$). We believe that our method has a potential value in reality because the number of candidate variables for regression analysis is less than one hundred in many cases.

A conventional way of avoiding multicollinearity is to remove explanatory variables one at a time according to the VIF value of each variable. Our MIQO formulations, however, have a clear advantage in terms of solution quality over such a heuristic algorithm. A solution of better quality leads to more reliable results for description purposes; hence, our MIQO formulations will be helpful in enhancing reliability of the regression analysis.

A future direction of study will be to extend our formulations to other regression/classification models. Multicollinearity generates a harmful effect on most statistical models, so our formulations are expected to be useful in various regression/classification models.

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

References

1. Arthanari, T.S., Dodge, Y.: Mathematical Programming in Statistics. Wiley, New York (1981)
2. Beale, E.M.L.: Two transportation problems. In: Kreweras, G., Morlat, G. (eds.) Proceedings of the Third International Conference on Operational Research, pp. 780–788 (1963)

3. Beale, E.M.L., Tomlin, J.A.: Special facilities in a general mathematical programming system for non-convex problems using ordered sets of variables. In: Lawrence, J. (ed.) Proceedings of the Fifth International Conference on Operational Research, pp. 447–454 (1970)
4. Belsley, D.A., Kuh, E., Welsch, R.E.: Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. Wiley, Hoboken (2005)
5. Benati, S., García, S.: A mixed integer linear model for clustering with variable selection. Comput. Oper. Res. **43**, 280–285 (2014)
6. Bertsimas, D., Dunn, J.: Optimal classification trees. Mach. Learn. **136**, 1039–1082 (2017)
7. Bertsimas, D., King, A.: OR forum: an algorithmic approach to linear regression. Oper. Res. **64**, 2–16 (2016)
8. Bertsimas, D., King, A.: Logistic regression: from art to science. Stat. Sci. **32**, 367–384 (2017)
9. Bertsimas, D., King, A., Mazumder, R.: Best subset selection via a modern optimization lens. Ann. Stat. **44**, 813–852 (2016)
10. Blum, A.L., Langley, P.: Selection of relevant features and examples in machine learning. Artif. Intell. **97**, 245–271 (1997)
11. Chatterjee, S., Hadi, A.S.: Regression Analysis by Example. Wiley, Hoboken (2012)
12. Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., García Marquéz, J.R., Gruber, B., Lafoourcade, B., Leitão, P.J., Münkemüller, T., McClean, C., Osborne, P.E., Reineking, B., Schröder, B., Skidmore, A.K., Zurell, D., Lautenbach, S.: Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography **36**, 27–46 (2013)
13. Gurobi Optimization, Inc.: Gurobi Optimizer Reference Manual. http://www.gurobi.com (2016). Accessed 6 Oct 2017
14. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. J. Mach. Learn. Res. **3**, 1157–1182 (2003)
15. Hastie, T., Tibshirani, R., Tibshirani, R.J.: Extended comparisons of best subset selection, forward stepwise selection, and the lasso. arXiv preprint arXiv:1707.08692 (2017)
16. Hocking, R.R.: The analysis and selection of variables in linear regression. Biometrics **32**, 1–49 (1976)
17. Hoerl, A.E., Kennard, R.W.: Ridge regression: biased estimation for nonorthogonal problems. Technometrics **12**, 55–67 (1970)
18. Huberty, C.J.: Issues in the use and interpretation of discriminant analysis. Psychol. Bull. **95**, 156–171 (1984)
19. IBM: IBM ILOG CPLEX Optimization Studio. https://www-01.ibm.com/software/commerce/optimization/cplex-optimizer/ (2015). Accessed 6 Oct 2017
20. James, G., Witten, D., Hastie, T., Tibshirani, R.: An Introduction to Statistical Learning. Springer, New York (2013)
21. Jolliffe, I.T.: A note on the use of principal components in regression. Appl. Stat. **31**, 300–303 (1982)
22. Kimura, K., Waki, H.: Minimization of Akaike's information criterion in linear regression analysis via mixed integer nonlinear program. Optim. Methods Softw. **33**, 633–649 (2018)
23. Kohavi, R., John, G.H.: Wrappers for feature subset selection. Artif. Intell. **97**, 273–324 (1997)
24. Konno, H., Yamamoto, R.: Choosing the best set of variables in regression analysis using integer programming. J. Glob. Optim. **44**, 273–282 (2009)
25. Lichman, M.: UCI Machine Learning Repository. School of Information and Computer Science, University of California, Irvine. http://archive.ics.uci.edu/ml (2013)
26. Liu, H., Motoda, H.: Computational Methods of Feature Selection. CRC Press, Boca Raton (2007)
27. Maldonado, S., Pérez, J., Weber, R., Labbé, M.: Feature selection for support vector machines via mixed integer linear programming. Inf. Sci. **279**, 163–175 (2014)
28. Massy, W.F.: Principal components regression in exploratory statistical research. J. Am. Stat. Assoc. **60**, 234–256 (1965)
29. Mazumder, R., Radchenko, P.: The discrete Dantzig selector: estimating sparse linear models via mixed integer linear optimization. IEEE Trans. Inf. Theory **63**, 3053–3075 (2017)
30. Miller, A.: Subset Selection in Regression. CRC Press, Boca Raton (2002)
31. Miyashiro, R., Takano, Y.: Subset selection by Mallows' C_p: a mixed integer programming approach. Expert. Syst. Appl. **42**, 325–331 (2015)
32. Miyashiro, R., Takano, Y.: Mixed integer second-order cone programming formulations for variable selection in linear regression. Eur. J. Oper. Res. **247**, 721–731 (2015)
33. R Core Team: R: a language and environment for statistical computing. R Foundation for Statistical Computing. https://www.R-project.org (2014). Accessed 6 Oct 2017
34. Sato, T., Takano, Y., Miyashiro, R., Yoshise, A.: Feature subset selection for logistic regression via mixed integer optimization. Comput. Optim. Appl. **64**, 865–880 (2016)

35. Sato, T., Takano, Y., Miyashiro, R.: Piecewise-linear approximation for feature subset selection in a sequential logit model. J. Oper. Res. Soc. Jpn. **60**, 1–14 (2017)
36. Tamura, R., Kobayashi, K., Takano, Y., Miyashiro, R., Nakata, K., Matsui, T.: Best subset selection for eliminating multicollinearity. J. Oper. Res. Soc. Jpn. **60**, 321–336 (2017)
37. Ustun, B., Rudin, C.: Supersparse linear integer models for optimized medical scoring systems. Mach. Learn. **102**, 349–391 (2016)
38. Wilson, Z.T., Sahinidis, N.V.: The ALAMO approach to machine learning. Comput. Chem. Eng. **106**, 785–795 (2017)
39. Wold, S., Ruhe, A., Wold, H., Dunn III, W.J.: The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. SIAM J. Sci. Stat. Comput. **5**, 735–743 (1984)
40. Wold, S., Sjöström, M., Eriksson, L.: PLS-regression: a basic tool of chemometrics. Chemom. Intell. Lab. Syst. **58**, 109–130 (2001)