Mixed Integer Quadratic Optimization Formulations for Eliminating Multicollinearity Based on Variance Inflation Factor

Ryuta Tamura[a], Ken Kobayashi[b], Yuichi Takano[c], Ryuhei Miyashiro[d,*], Kazuhide Nakata[e], Tomomi Matsui[e]

[a] *Graduate School of Engineering, Tokyo University of Agriculture and Technology, 2-24-16 Naka-cho, Koganei-shi, Tokyo 184-8588, Japan*
[b] *Knowledge Information Processing Laboratory, Fujitsu Laboratories Ltd., 4-1-1 Kamikodanaka, Nakahara-ku, Kawasaki-shi, Kanagawa 211-8588, Japan*
[c] *School of Network and Information, Senshu University, 2-1-1 Higashimita, Tama-ku, Kawasaki-shi, Kanagawa 214-8580, Japan*
[d] *Institute of Engineering, Tokyo University of Agriculture and Technology, 2-24-16 Naka-cho, Koganei-shi, Tokyo 184-8588, Japan*
[e] *Department of Industrial Engineering and Economics, School of Engineering, Tokyo Institute of Technology, 2-12-1 Ookayama, Meguro-ku, Tokyo 152-8552, Japan*

Abstract

The variance inflation factor, VIF, is the most frequently used indicator for detecting multicollinearity in multiple linear regression models. This paper proposes two mixed integer quadratic optimization formulations for selecting the best subset of explanatory variables under upper-bound constraints on VIF of selected variables. Computational results illustrate the effectiveness of our optimization formulations based on comparisons with conventional local search algorithms.

Keywords: Integer programming, Subset selection, Multicollinearity, Variance inflation factor, Multiple Linear Regression, Statistics

*Corresponding author.
Email address: r-miya@cc.tuat.ac.jp (Ryuhei Miyashiro)

1. Introduction

Multiple regression analysis is a statistical process for estimating the relationship between explanatory variables and an explained variable. The reliability of the analysis is decreased when some of explanatory variables are highly correlated because of the low quality of the resultant estimates. This problem is known as *multicollinearity* [1, 4, 5].

Several methods have been proposed to avoid the deleterious effects of multicollinearity [1, 4, 5]. Even among them, this paper focuses on the subset selection, which is a common and effective method for eliminating multicollinearity from a regression model. Conventionally in this method, explanatory variables are removed one by one on the basis of indicators for detecting multicollinearity, such as a condition number and variance inflation factor (VIF). On the other hand, the potential disadvantage of this iterative procedure is that the best subset of variables, e.g., in the least-squares sense, is not necessarily found.

Recently, mixed integer optimization (MIO) approaches to subset selection have received much attention because they have the potential to provide the best subset of variables with respect to several goodness-of-fit measures, such as the ordinary least squares [2, 3], least absolute deviation [3, 8, 9], Mallows' C_p [11], adjusted R^2 [12] and some information criteria [7, 12, 14, 15].

As an MIO approach for avoiding multicollinearity, Bertsimas and King [2] suggest the use of cutting plane algorithm, which iteratively adds valid inequalities for cutting off sets of collinear variables. These valid inequalities can be strengthened by means of a local search algorithm [16]; however, this algorithm is not computationally efficient because it must repeatedly solve a large number of MIO problems, each of which is NP-hard.

Meanwhile, the authors of this paper devised a mixed integer semidefinite optimization (MISDO) formulation for subset selection to eliminate multicollinearity [16]. In contrast to the cutting plane algorithm, this approach needs to solve only 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 the most common indicator for detecting multicollinearity, to the best of our knowledge, none of the existing studies have developed a single MIO formulation for eliminating multicollinearity based on VIF.

In light of these circumstances, this paper proposes mixed integer quadratic optimization (MIQO) formulations for subset selection to eliminate multi-

collinearity based on VIF. Our two MIQO formulations are respectively derived based on the two equivalent definitions of VIF. Computational results demonstrate that our MIQO formulations provided solutions of better quality than those of local search algorithms within a time limit of 10000 s.

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 an explained 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 Sections 2 and 3 we assume that all explanatory and explained 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

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

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

In what follows, we consider selecting a subset $S \subseteq P$ of explanatory variables in order to reduce negative influence of multicollinearity on regression estimates. For this purpose, we focus on the correlation matrix of selected variables. On account of assumption (1), it 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 sub-matrix 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 $\boldsymbol{R}_S$, i.e.,

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

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

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

where α is a user-defined parameter larger than one. When the VIF value is greater than 10, the set S often has collinear problems [4].

On the other hand, VIF has another easily interpretable definition. To describe this, we consider a linear regression model of explaining 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 defined 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 [1, 5]:

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

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 VIF. 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:

$$\begin{aligned}
\text{minimize} \quad & \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{a}\|_2^2 && (8)\\
\text{subject to} \quad & z_j = 0 \;\Rightarrow\; a_j = 0 \quad (j \in P), && (9)\\
& z_\ell = 1 \;\Rightarrow\; \mathrm{VIF}(\ell, S(\boldsymbol{z})) \le \alpha \quad (\ell \in P), && (10)\\
& \boldsymbol{a} \in \mathbb{R}^p,\ \boldsymbol{z} \in \{0,1\}^p. && (11)
\end{aligned}$$

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 (SOS1) constraint.

3.1. Inverse-matrix-based formulation

We first propose a mixed integer quadratic optimization (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 $\boldsymbol{R}_{S(\boldsymbol{z})}$, we make use of the following constraints:

$$\begin{aligned}
& \boldsymbol{Q}\boldsymbol{R}_P + \boldsymbol{U} = \boldsymbol{I}_p, && (12)\\
& z_j = 1 \;\Rightarrow\; u_{\ell j} = 0 \quad (\ell \in P,\ j \in P), && (13)\\
& z_j = 0 \;\Rightarrow\; q_{\ell j} = q_{j\ell} = 0 \quad (\ell \in P,\ j \in P), && (14)
\end{aligned}$$

where $\boldsymbol{I}_p$ is the identity matrix of size p.

Theorem 1. *If* $(\boldsymbol{Q}, \boldsymbol{U}, \boldsymbol{z}) \in \mathbb{R}^{p\times p} \times \mathbb{R}^{p\times p} \times \{0,1\}^p$ *satisfies constraints* (12)–(14)*, the equality* $q_{\ell\ell} = [\boldsymbol{R}_{S(\boldsymbol{z})}^{-1}]_{\ell\ell}$ *holds for each* $\ell \in S(\boldsymbol{z})$.

PROOF. Let s be the number of nonzero elements of $\boldsymbol{z}$. Without loss of generality, we may assume that $S(\boldsymbol{z}) = \{1, 2, \ldots, s\}$. We partition $\boldsymbol{Q}, \boldsymbol{R}_P, \boldsymbol{U}$ and $\boldsymbol{I}_p$ according to $S(\boldsymbol{z})$ and rewrite constraints (12) 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 (13) and (14). As a result, the above constraints are reduced to

$$\begin{pmatrix} \boldsymbol{Q}_1\boldsymbol{R}_1 \\ \boldsymbol{O}^\top \end{pmatrix} = \begin{pmatrix} \boldsymbol{I}_s \\ \boldsymbol{O}^\top \end{pmatrix}, \tag{15}$$

while other constraints are satisfied through free decision variables $\boldsymbol{U}_2$ 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. □

Since the correlation matrix is symmetric, the same holds for its inverse, and thus $\boldsymbol{Q} = \boldsymbol{Q}^\top$. According to the definition (2), the subset selection problem (8)–(11) is reformulated as an MIQO problem, which we call the inverse-matrix-based formulation:

$$\begin{aligned}
\text{minimize} \quad & \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{a}\|_2^2 && (16)\\
\text{subject to} \quad & z_j = 0 \;\Rightarrow\; a_j = 0 \quad (j \in P), && (17)\\
& q_{\ell\ell} \le \alpha \quad (\ell \in P), && (18)\\
& \boldsymbol{Q}\boldsymbol{R}_P + \boldsymbol{U} = \boldsymbol{I}_p, \quad \boldsymbol{Q} = \boldsymbol{Q}^\top, && (19)\\
& z_j = 1 \;\Rightarrow\; u_{\ell j} = 0 \quad (\ell \in P,\ j \in P), && (20)\\
& z_j = 0 \;\Rightarrow\; q_{\ell j} = 0 \quad (\ell \in P,\ j \in P), && (21)\\
& \boldsymbol{a} \in \mathbb{R}^p,\ \boldsymbol{Q} \in \mathbb{R}^{p\times p},\ \boldsymbol{U} \in \mathbb{R}^{p\times p},\ \boldsymbol{z} \in \{0,1\}^p. && (22)
\end{aligned}$$

3.2. Normal-equation-based formulation

We next propose another MIQO formulation based on the definition (7) of VIF. Let us introduce a vector of decision variables,

$$\boldsymbol{a}^{(\ell)} := (a_1^{(\ell)}, a_2^{(\ell)}, \ldots, a_{\ell-1}^{(\ell)}, a_{\ell+1}^{(\ell)}, a_{\ell+2}^{(\ell)}, \ldots, a_p^{(\ell)})^\top \in \mathbb{R}^{p-1}$$

for each $\ell \in P$. To convert the VIF constraints (10) into a set of linear constraints, we exploit the normal-equation-based constraints proposed in Tamura et al. [16]. Specifically, we make use of the following constraints:

$$z_j = 1 \quad \Rightarrow \quad \boldsymbol{x}_j^\top \boldsymbol{X}_{P\setminus\{\ell\}} \boldsymbol{a}^{(\ell)} = \boldsymbol{x}_j^\top \boldsymbol{x}_\ell \quad (j \in P \setminus \{\ell\}), \tag{23}$$

$$z_j = 0 \quad \Rightarrow \quad a_j^{(\ell)} = 0 \quad (j \in P \setminus \{\ell\}). \tag{24}$$

Theorem 2. *Suppose that* $(\boldsymbol{a}^{(\ell)}, \boldsymbol{z}) \in \mathbb{R}^{p-1} \times \{0,1\}^p$ *satisfies constraints* (23)–(24)*, and* $\ell \in S(\boldsymbol{z})$*. Then,* $a_j^{(\ell)}$ $(j \in S(\boldsymbol{z}) \setminus \{\ell\})$ *solves the normal equation* (6) *for* $S = S(\boldsymbol{z})$*, and* $a_j^{(\ell)} = 0$ $(j \notin S(\boldsymbol{z}))$*.*

Proof. Similarly to Theorem 1, we may assume without loss of generality 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 (24). Therefore constraints (23) correspond to the following normal equation:

$$\boldsymbol{X}_{S(\boldsymbol{z})\setminus\{\ell\}}^\top \boldsymbol{X}_{S(\boldsymbol{z})\setminus\{\ell\}} \boldsymbol{a}_1^{(\ell)} = \boldsymbol{X}_{S(\boldsymbol{z})\setminus\{\ell\}}^\top \boldsymbol{x}_\ell, \tag{25}$$

which completes the proof. □

Since the OLS estimator provides the minimum value of RSS (5), Theorem 2 implies that

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

under constraints (23) and (24). Thus, the VIF constraints (10) can be rewritten by its definition (7) as follows:

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

However, this is a reverse convex constraint, which is intractable to handle in an MIO approach.

To resolve this difficulty, we recall that $\boldsymbol{X}_{P\setminus\{\ell\}}\boldsymbol{a}^{(\ell)} = \boldsymbol{X}_{S(\boldsymbol{z})\setminus\{\ell\}}\boldsymbol{a}_1^{(\ell)}$ because $\boldsymbol{a}_2^{(\ell)} = \boldsymbol{0}$ in the proof of Theorem 2. We also exploit the normal equation (25) to show that

$$\begin{aligned} & \|\boldsymbol{x}_\ell - \boldsymbol{X}_{P\setminus\{\ell\}}\boldsymbol{a}^{(\ell)}\|_2^2 \\ = \; & \boldsymbol{x}_\ell^\top\boldsymbol{x}_\ell - 2\boldsymbol{x}_\ell^\top\boldsymbol{X}_{P\setminus\{\ell\}}\boldsymbol{a}^{(\ell)} + (\boldsymbol{a}_1^{(\ell)})^\top\boldsymbol{X}_{S(\boldsymbol{z})\setminus\{\ell\}}^\top\boldsymbol{X}_{S(\boldsymbol{z})\setminus\{\ell\}}\boldsymbol{a}_1^{(\ell)} \\ = \; & \boldsymbol{x}_\ell^\top\boldsymbol{x}_\ell - 2\boldsymbol{x}_\ell^\top\boldsymbol{X}_{P\setminus\{\ell\}}\boldsymbol{a}^{(\ell)} + (\boldsymbol{a}_1^{(\ell)})^\top\boldsymbol{X}_{S(\boldsymbol{z})\setminus\{\ell\}}^\top\boldsymbol{x}_\ell \\ = \; & \boldsymbol{x}_\ell^\top\boldsymbol{x}_\ell - \boldsymbol{x}_\ell^\top\boldsymbol{X}_{P\setminus\{\ell\}}\boldsymbol{a}^{(\ell)}. \end{aligned} \tag{26}$$

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

$$\text{minimize} \quad \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{a}\|_2^2 \tag{27}$$

$$\text{subject to} \quad z_j = 0 \;\; \Rightarrow \;\; a_j = 0 \quad (j \in P), \tag{28}$$

$$z_\ell \le \alpha(\boldsymbol{x}_\ell^\top\boldsymbol{x}_\ell - \boldsymbol{x}_\ell^\top\boldsymbol{X}_{P\setminus\{\ell\}}\boldsymbol{a}^{(\ell)}) \quad (\ell \in P), \tag{29}$$

$$z_j = 1 \;\; \Rightarrow \;\; \boldsymbol{x}_j^\top\boldsymbol{X}_{P\setminus\{\ell\}}\boldsymbol{a}^{(\ell)} = \boldsymbol{x}_j^\top\boldsymbol{x}_\ell \quad (\ell \in P,\; j \in P\setminus\{\ell\}), \tag{30}$$

$$z_j = 0 \;\; \Rightarrow \;\; a_j^{(\ell)} = 0 \quad (\ell \in P,\; j \in P\setminus\{\ell\}), \tag{31}$$

$$\boldsymbol{a} \in \mathbb{R}^p,\; \boldsymbol{z} \in \{0,1\}^p, \quad \boldsymbol{a}^{(\ell)} \in \mathbb{R}^{p-1} \quad (\ell \in P). \tag{32}$$

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 efficiency of preprocessing (i) and (ii) in the next section.

Preprocessing (i): 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, where

$$P_0 := \{\ell \in P \mid \mathrm{VIF}(\ell, P) \le \alpha\}. \tag{33}$$

Preprocessing (ii): Cutting-plane-based constraints. First, find sets $S_k \subseteq P$ $(k \in K)$ of collinear variables such that $\mathrm{VIF}(\ell, S_k) > \alpha$ for some $\ell \in P$. Next, cut them off by means of the following cutting-plane-based constraints [2, 16]:

$$\sum_{j\in S_k} z_j \le |S_k| - 1 \quad (k \in K). \tag{34}$$

Preprocessing (iii): Tightening constraints. Constraints (29) 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 (26) 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 based on VIF.

We downloaded five datasets for regression analysis from the UCI Machine Learning Repository [10]. Table 1 lists the instances used for computational experiments, where n and p are the number of samples and number of candidate explanatory variables, respectively. In the `SolarFlareC` instance, C-class flares production was employed as an explained variable. Each categorical variable was transformed into one or more dummy variables. Samples containing missing values and redundant variables having the same value in all samples were removed.

Table 1: List of instances

Abbreviation	n	p	Original dataset [10]
`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

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

FwS Forward selection method: Starts with $S = \emptyset$ and iteratively adds the variable j (i.e., $S \leftarrow S \cup \{j\}$) with a greedy manner in terms of decrement of RSS; this operation is repeated while the VIF constraints (3) are satisfied.

BwE Backward elimination method: Starts with $S = \{1, 2, \ldots, p\}$ and iteratively eliminates the variable j (i.e., $S \leftarrow S \setminus \{j\}$) with a stingy method in terms of increment of RSS; this operation is repeated until the VIF constraints (3) hold.

IMF Inverse-matrix-based MIQO formulation (16)–(22).

IMF+ Inverse-matrix-based MIQO formulation (16)–(22) with the preprocessing (i) and (ii).

NEF Normal-equation-based MIQO formulation (27)–(32).

NEF+ Normal-equation-based MIQO formulation (27)–(32) 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 [13], and MIQO problems were solved by using IBM ILOG CPLEX 12.6.3.0 [6] with eight threads. Here the `indicator` function implemented in CPLEX was used to impose the logical implications (17), (20), (21), (28), (30) and (31). The upper bound on VIF was set as $\alpha = 10$ in accordance with Chatterjee and Hadi [4].

Note that IMF+ and NEF+ involved the preprocessing (i) and (ii) as explained in Section 3.3. Precisely, the VIF constraints for the set (33) were deleted in advance, and the cutting-plane-based constraints (34) were included. Here, the sets 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, where $|P_0|$ and $|K|$ are the number of redundant VIF constraints and that of 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.

Table 3 shows the computational results of the subset selection algorithms. 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 for solving the MIQO

Table 2: Results of preprocessing

			Preprocessing (i)		Preprocessing (ii)	
Instance	n	p	$\|P_0\|$	Time (s)	$\|K\|$	Time (s)
`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

problem was terminated if it did not finish by itself within 10000 s. In this case, the best feasible solution obtained within 10000 s was taken as the result.

Our MIQO problems for the `AutoMPG` and `SolarFlareC` instances were all solved completely within a few tens of seconds. In this case, their R^2 values were always the largest because the obtained solutions were proved to be optimal. We can also see that our preprocessing significantly reduced the computation time of solving the MIQO problems. For instance, in the case of `BreastCancer` instance, the computation of IMF+ finished more than 20 times faster than that of IMF did.

The MIQO computations for the `Automobile` and `Crime` instances were terminated due to the time limit of 10000 s; nevertheless, they successfully found quality solutions within 10000 s. Indeed, IMF+ provided a solution of the largest R^2 value to each of `Automobile` and `Crime` instances. These results support the effectiveness of our MIQO approaches to subset selection for eliminating multicollinearity. On the other hand, IMF failed to deliver a solution of good quality to the `Crime` instance; precisely, it found only a feasible solution $S = \emptyset$ within 10000 s.

Acknowledgments

This work was partially supported by JSPS KAKENHI Grant Number JP26560165.

References

[1] Belsley, D.A., Kuh, E., and Welsch, R.E. (2005). Regression diagnostics: Identifying influential data and sources of collinearity. John Wiley &

Table 3: Results of subset selection algorithms

Instance	n	p	Method	R^2	$\text{VIF}_{\max}$	$\lvert S\rvert$	Time (s)
`AutoMPG`	392	25	FwS	**0.87334**	9.549	20	0.92
			BwE	0.87149	5.899	16	0.75
			IMF	**0.87334**	8.523	20	21.09
			IMF+	**0.87334**	8.523	20	1.72
			NEF	**0.87334**	8.523	20	13.12
			NEF+	**0.87334**	8.523	20	1.83
`SolarFlareC`	1066	26	FwS	0.19713	3.083	19	1.28
			BwE	0.18232	7.661	9	1.50
			IMF	**0.19715**	4.348	19	12.93
			IMF+	**0.19715**	9.102	19	2.00
			NEF	**0.19715**	4.348	19	9.86
			NEF+	**0.19715**	4.348	19	1.93
`BreastCancer`	194	32	FwS	0.27039	9.981	16	1.24
			BwE	0.25424	9.973	8	1.81
			IMF	**0.29158**	9.765	16	>10000.00
			IMF+	**0.29158**	9.765	16	499.92
			NEF	**0.29158**	9.765	16	>10000.00
			NEF+	**0.29158**	9.765	16	1523.34
`Automobile`	159	65	FwS	0.96605	9.996	31	5.25
			BwE	0.91367	1.596	10	13.01
			IMF	0.96626	9.923	41	>10000.00
			IMF+	**0.96970**	9.937	43	>10000.00
			NEF	0.96281	9.411	32	>10000.00
			NEF+	0.96568	8.679	29	>10000.00
`Crime`	1993	100	FwS	0.66906	9.999	24	13.12
			BwE	0.64953	5.773	7	115.52
			IMF	0	—	0	>10000.00
			IMF+	**0.67660**	9.988	45	>10000.00
			NEF	0.67444	9.989	40	>10000.00
			NEF+	0.67587	9.934	49	>10000.00

Sons.

[2] Bertsimas, D. and King, A. (2016). OR forum—An algorithmic approach to linear regression. Operations Research, 64(1), 2–16.

[3] Bertsimas, D., King, A., and Mazumder, R. (2016). Best subset selection via a modern optimization lens. The Annals of Statistics, 44(2), 813–852.

[4] Chatterjee, S. and Hadi, A.S. (2012). Regression analysis by example. John Wiley & Sons.

[5] 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., and Lautenbach, S. (2013). Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. Ecography, 36(1), 27–46.

[6] IBM. (2015). IBM ILOG CPLEX Optimization Studio. [`https://www-01.ibm.com/software/commerce/optimization/cplex-optimizer/`]

[7] Kimura, K. and Waki, H. (2016). Minimization of Akaike's information criterion in linear regression analysis via mixed integer nonlinear program. arXiv preprint, arXiv:1606.05030.

[8] Konno, H. and Takaya, Y. (2010). Multi-step methods for choosing the best set of variables in regression analysis. Computational Optimization and Applications, 46(3), 417–426.

[9] Konno, H. and Yamamoto, R. (2009). Choosing the best set of variables in regression analysis using integer programming. Journal of Global Optimization, 44(2), 273–282.

[10] Lichman, M. (2013). UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science. [`http://archive.ics.uci.edu/ml`]

[11] Miyashiro, R. and Takano, Y. (2015). Subset selection by Mallows' C_p: A mixed integer programming approach. Expert Systems with Applications, 42(1), 325–331.

[12] Miyashiro, R. and 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.

[13] R Core Team. (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing. [https://www.R-project.org]

[14] Sato, T., Takano, Y., Miyashiro, R., and Yoshise, A. (2016). Feature subset selection for logistic regression via mixed integer optimization. Computational Optimization and Applications, 64(3), 865–880.

[15] Sato, T., Takano, Y., and Miyashiro, R. (in press). Piecewise-linear approximation for feature subset selection in a sequential logit model. Journal of the Operations Research Society of Japan.

[16] Tamura, R., Kobayashi, K., Takano, Y., Miyashiro, R., Nakata, K., and Matsui, T. (2016). Best subset selection for eliminating multicollinearity. Optimization Online. [http://www.optimization-online.org/DB_HTML/2016/07/5559.html]