

Department of Industrial Engineering and Management

Technical Report

No. 2013-7

Mixed Integer Second-Order Cone Programming Formulations for Variable Selection

Ryuhei Miyashiro and Yuichi Takano



June, 2013

Tokyo Institute of Technology

2-12-1 Ookayama, Meguro-ku, Tokyo 152-8552, JAPAN
<http://www.me.titech.ac.jp/index-e.html>

Mixed Integer Second-Order Cone Programming Formulations for Variable Selection

Ryuhei Miyashiro^{a,*}, Yuichi Takano^b

^a*Department of Computer and Information Sciences,
Institute of Engineering,*

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

^b*Department of Industrial Engineering and Management,
Graduate School of Decision Science and Technology,*

*Tokyo Institute of Technology,
2-12-1-W9-77 Ookayama, Meguro-ku, Tokyo 152-8552, Japan*

Abstract

This paper concerns the method of selecting the best subset of explanatory variables in a multiple linear regression model. To evaluate a subset regression model, some goodness-of-fit measures, e.g., adjusted R^2 , AIC and BIC, are generally employed. Although variable selection is usually handled via a stepwise regression method, the method does not always provide the best subset of explanatory variables according to adjusted R^2 , AIC and BIC. In this paper, we propose mixed integer second-order cone programming formulations for selecting the best subset of variables. Computational experiments show that, in terms of the goodness-of-fit measures, the proposed formulations yield solutions having a clear advantage over common stepwise regression methods.

Keywords: Integer programming, Variable selection, Multiple linear regression, Information criterion, Second-order cone programming, Statistics

*Corresponding author.

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

1. Introduction

Variable selection in statistics, also known as feature selection or attribute selection in machine learning, is the method of choosing a set of significant variables from many candidate variables for model construction. Potential benefits of variable selection are as follows (see e.g., [14, 33]): (i) improving predictive performance of a statistical model by preventing overfitting, (ii) identifying a model that captures the essence of a system, and (iii) providing a computationally-efficient set of explanatory variables. From these benefits, studies on variable selection are of supreme importance in multiple regression analysis [9, 16, 26].

There are several goodness-of-fit (GOF) measures, such as adjusted R^2 [32], Akaike information criterion (AIC) [1] and Bayesian information criterion (BIC) [28], to evaluate a subset regression model. A straightforward way to search for the best-subset regression model is evaluating all possible subset models. Though some procedures have been described, e.g., [12, 13, 18], this task is practically infeasible unless the number of candidate variables is small. Accordingly, existing studies have focused on a wide range of search strategies for approximately solving the problem (see e.g., [8, 14, 20, 23]). Among them is a well-known stepwise regression method [10], which repeats forward selection (adding one significant variable) and backward elimination (eliminating one redundant variable) until a stopping condition is satisfied. Ridge regression [17], Lasso [31] and metaheuristics (e.g., [27]) are also used for variable selection.

Although these heuristic optimization algorithms can handle large-scale variable selection problems, they do not necessarily select the best set of variables. In addition, other shortcomings of stepwise regression have been pointed out, e.g., Whittingham et al. [33].

In contrast to heuristic optimization algorithms, integer programming methodology has the potential to determine the best subset of explanatory variables in a multiple linear regression model. When the number of variables to be selected is given, the variable selection problems can be formulated as mixed integer quadratic programming (MIQP) problems [3, 7, 22]. In particular, Bertsimas and Shioda [7] utilized a tailored branch-and-bound procedure to solve the problem of minimizing the sum of squared deviation. They reported that their algorithm had significant computational advantages over a commercial mixed integer programming (MIP) solver. It is also known that, by employing the mean absolute deviation as a deviation measure,

the variable selection problem can be formulated as a mixed integer linear programming (MILP) problems (e.g., [3]). Konno and Yamamoto [22] solved similar MILP problems by using a MIP solver, and Konno and Takaya [21] developed a multi-step method to obtain a nearly optimal solution to large-scale problems.

To adopt these MIQP or MILP approaches, however, the number of selected variables has to be fixed in advance. This is disadvantageous because the optimal number of variables in terms of a GOF measure is unknown before solving the corresponding variable selection problem. On the assumption that the residual variance of the best-subset regression model is given, the variable selection problem of minimizing AIC can be formulated as MIP problems with a nonlinear convex objective function [11, 29]; however, this assumption clearly does not coincide with reality. A straightforward formulation for minimizing AIC or BIC generally leads a MIP problem with a nonconvex objective function, which is intractable even if integrality of variables is relaxed.

The purpose of this paper is to develop an exact and practical method for selecting the best subset of explanatory variables in a multiple linear regression model. To this end, we propose mixed integer second-order cone programming (MISOCP) formulations to build the best-subset regression model in terms of the adjusted \bar{R}^2 , AIC and BIC, without prespecifying the number of variables to be selected. The continuous relaxation of an MISOCP problem is a second-order cone programming (SOCP) problem, which can be solved in polynomial time; thus, an MISOCP problem can be handled by recent MIP solvers using a branch-and-bound procedure.

Using data sets from UCI Machine Learning Repository [4], we conduct computational experiments to assess the effectiveness of the proposed MISOCP formulations. Computational results show that the proposed method is able to provide the best subset of variables for small-sized instances in minutes. Furthermore, for medium-sized instances, the method often generates a better subset of variables than stepwise regression methods do.

2. Variable selection and goodness-of-fit measures

This section makes a brief review of variable selection and GOF measures, and mentions previous researches on variable selection using integer programming.

2.1. Multiple linear regression analysis and variable selection

Given n data points, $(y_i; x_{i1}, x_{i2}, \dots, x_{ik})$ for $i = 1, 2, \dots, n$, y_i is referred to as an explained variable (or dependent variable), and x_{ij} ($j = 1, 2, \dots, k$) to as explanatory variables (or independent variables). In multiple linear regression analysis, the following linear model is constructed for predicting the value of y_i :

$$y_i = b + a_1x_{i1} + a_2x_{i2} + \dots + a_kx_{ik} + \varepsilon_i, \quad (1)$$

where ε_i is a prediction residual corresponding to the i -th data point. The ordinary least squares method estimates the value of the intercept b and coefficient vector \mathbf{a} such that the sum of squared residuals $\sum_{i=1}^n \varepsilon_i^2$ is minimized.

This paper considers the variable selection problem, i.e., selecting the best subset of variables from the set of candidate explanatory variables. To evaluate a subset regression model, in the following some GOF measures are explained. Throughout this paper, it is assumed that the number of all candidate variables is p , and that the number n of data points is much larger than p .

2.2. Adjusted R^2

The adjusted R^2 [32], hereafter \bar{R}^2 , for the regression model (1) is defined as follows:

$$\bar{R}^2 = 1 - \frac{\sum_{i=1}^n \varepsilon_i^2 / (n - k - 1)}{\sum_{i=1}^n (y_i - \bar{y})^2 / (n - 1)},$$

where $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$. Note that maximizing \bar{R}^2 is equivalent to minimizing

$$\sum_{i=1}^n \varepsilon_i^2 / (n - k - 1), \quad (2)$$

because other terms are all constants. Accordingly, if the sum of squared residuals $\sum_{i=1}^n \varepsilon_i^2$ are the same in two models, the model with smaller k is better. Conversely, if k is fixed in advance, minimizing $\sum_{i=1}^n \varepsilon_i^2$ leads the best model.

2.3. Information criteria: AIC and BIC

Assume that the prediction residuals, i.e., $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$, are independent and all normally distributed with zero mean and the variance σ^2 . Then, the log likelihood function of the regression model (1) can be written as follows:

$$\ell(\mathbf{a}, b, \sigma^2) = -\frac{n}{2} \log 2\pi - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n \varepsilon_i^2, \quad (3)$$

where $\mathbf{a} = (a_1, a_2, \dots, a_p)$. By partial differentiation, the maximum likelihood estimator of σ^2 becomes

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n \varepsilon_i^2. \quad (4)$$

By substituting (4) into (3), the maximal value of the log likelihood function is expressed as:

$$\max_{\mathbf{a}, b, \sigma^2} \ell(\mathbf{a}, b, \sigma^2) = \max_{\mathbf{a}, b} \left(-\frac{n}{2} \log 2\pi - \frac{n}{2} \log \left(\frac{1}{n} \sum_{i=1}^n \varepsilon_i^2 \right) - \frac{n}{2} \right).$$

The Akaike information criterion (AIC) [1] of the regression model (1) is defined as follows:

$$-2 \max_{\mathbf{a}, b, \sigma^2} \ell(\mathbf{a}, b, \sigma^2) + 2(k+2) \quad (5)$$

$$= \min_{\mathbf{a}, b} \left(n \log 2\pi + n \log \left(\frac{1}{n} \sum_{i=1}^n \varepsilon_i^2 \right) + n \right) + 2(k+2), \quad (6)$$

where $k+2$ is the number of parameters (i.e., \mathbf{a}, b , and σ^2) in the model. By omitting constant terms from (6), the variable selection problem with respect to AIC is reduced to minimization of

$$n \log \left(\frac{1}{n} \sum_{i=1}^n \varepsilon_i^2 \right) + 2k. \quad (7)$$

The Bayesian information criterion (BIC) [28] is another information criterion as popular as AIC. The BIC of the regression model (1) is defined as follows:

$$-2 \max_{\mathbf{a}, b, \sigma^2} \ell(\mathbf{a}, b, \sigma^2) + (k+2) \log n.$$

As in the case for AIC, the following function is minimized to solve the variable selection problem in terms of BIC:

$$n \log \left(\frac{1}{n} \sum_{i=1}^n \varepsilon_i^2 \right) + k \log n. \quad (8)$$

2.4. Previous researches on variable selection via integer programming

In this subsection, previous researches on variable selection are described from the viewpoint of integer programming.

From (2), (7) and (8), if the number k of selected variables is predetermined, only minimizing the sum of squared residuals $\sum_{i=1}^n \varepsilon_i^2$ is necessary to select the best k variables by means of \bar{R}^2 , AIC and BIC. This minimization problem is known to be formulated as a MIQP problem [3, 7, 22], explained as follows.

Let z_j ($j = 1, 2, \dots, p$) be a 0-1 variable such that $z_j = 1$ if the j -th candidate variable is selected, otherwise $z_j = 0$. The variable selection problem with specified k is formulated as the following MIQP problem¹:

$$\underset{\mathbf{a}, \mathbf{b}, \boldsymbol{\varepsilon}, \mathbf{z}}{\text{minimize}} \quad \sum_{i=1}^n \varepsilon_i^2 \quad (9)$$

$$\text{subject to} \quad \varepsilon_i = y_i - \left(b + \sum_{j=1}^p a_j x_{ij} \right) \quad (i = 1, 2, \dots, n), \quad (10)$$

$$- Mz_j \leq a_j \leq Mz_j \quad (j = 1, 2, \dots, p), \quad (11)$$

$$\sum_{j=1}^p z_j = k, \quad (12)$$

$$z_j \in \{0, 1\} \quad (j = 1, 2, \dots, p), \quad (13)$$

where M is a sufficiently large positive constant. If $z_j = 0$, the j -th candidate variable is eliminated from a regression model, because its coefficient a_j has to be 0 from Constraint (11); if $z_j = 1$, Constraint (11) is invalidated. Thus, the number of variables chosen becomes k due to Constraint (12). Hence, the problem (9)–(13) is a correct formulation for minimizing the sum of squared

¹When solving the MIQP (9)–(13), substituting Constraint (10) into the objective function (9) leads a simpler formulation.

residuals under the condition that the number of selected explanatory variables is k .

The formulation (9)–(13) is a problem with a convex quadratic objective function subject to linear and integrality constraints, i.e., MIQP. A branch-and-bound procedure handles an integer programming problem by relaxing the integrality constraints and then by solving the relaxation problem repeatedly; hence, a key factor for a branch-and-bound procedure is whether the continuous relaxation problem is computationally tractable. In this regard, the continuous relaxation of an MIQP problem is a quadratic programming (QP) problem, which is solvable in polynomial time. Therefore a branch-and-bound procedure works well for solving the problem (9)–(13).

Other than minimizing the sum of squared residuals, some researches considered minimizing the mean absolute deviations, i.e., $\frac{1}{n} \sum_{i=1}^n |\varepsilon_i|$ (see [3, 21, 22]). In this case, the corresponding variable selection problem with fixed k can be formulated as an MILP problem, which is easier than an MIQP one.

Nevertheless, in these MIQP and MILP approaches, the value of k needs to be fixed before solving the problems; this restriction is impractical. To find the best subset of variables by the approaches, it is necessary to solve all problems for $k = 0, 1, \dots, p$.

3. Mixed integer second-order cone programming formulations for variable selection

In this section, we propose MISOCP formulations for maximizing \bar{R}^2 , minimizing AIC and BIC, to select the best set of variables based on these GOF measures. Note that the proposed formulations treat k as a variable, and the continuous relaxation problems of the formulations belong to a computationally tractable class, SOCP (see Appendix A for SOCP.)

3.1. Mixed Integer SOCP formulation for maximizing \bar{R}^2

In view of (2), the variable selection problem of maximizing \bar{R}^2 can be formulated as follows:

$$\begin{aligned} & \underset{\mathbf{a}, \mathbf{b}, \varepsilon, k, \mathbf{z}}{\text{minimize}} && \sum_{i=1}^n \varepsilon_i^2 / (n - k - 1) \\ & \text{subject to} && \text{Constraints (10), (11), (12) and (13).} \end{aligned}$$

Except for the integrality constraint (13), the above problem has the same structure as the problem (A.1)–(A.2), which can be transformed into the SOCP problem (A.3)–(A.8) (see Appendix A.) Hence, importing the integrality constraint, we easily have an MISOCP formulation for maximizing \bar{R}^2 as follows:

$$\underset{\substack{\mathbf{a}, b, \boldsymbol{\varepsilon}, f, \\ g, k, \mathbf{z}}}{\text{minimize}} \quad f \quad (14)$$

$$\text{subject to} \quad \varepsilon_i = y_i - \left(b + \sum_{j=1}^p a_j x_{ij} \right) \quad (i = 1, 2, \dots, n), \quad (15)$$

$$\sum_{i=1}^n \varepsilon_i^2 \leq f \cdot g, \quad (16)$$

$$g = n - k - 1, \quad (17)$$

$$-Mz_j \leq a_j \leq Mz_j \quad (j = 1, 2, \dots, p), \quad (18)$$

$$\sum_{j=1}^p z_j = k, \quad (19)$$

$$z_j \in \{0, 1\} \quad (j = 1, 2, \dots, p). \quad (20)$$

3.2. Mixed integer SOCP formulations for minimizing AIC and BIC

In view of (7), the variable selection problem of minimizing AIC can be formulated as follows:

$$\underset{\mathbf{a}, b, \boldsymbol{\varepsilon}, k, \mathbf{z}}{\text{minimize}} \quad n \log \left(\frac{1}{n} \sum_{i=1}^n \varepsilon_i^2 \right) + 2k \quad (21)$$

subject to Constraints (10), (11), (12) and (13).

However, this straightforward formulation forces us to solve a MIP problem with the nonconvex objective function (21). Hence, even its continuous relaxation problem is computationally intractable. This is an undesirable outcome.

In the following, we make the above problem computationally tractable.

The objective function (21) is converted as follows:

$$\begin{aligned}
& \underset{\mathbf{a}, \mathbf{b}, \boldsymbol{\varepsilon}, k, \mathbf{z}}{\text{minimize}} && n \log \left(\frac{1}{n} \sum_{i=1}^n \varepsilon_i^2 \right) + 2k \\
\iff & \underset{\mathbf{a}, \mathbf{b}, \boldsymbol{\varepsilon}, k, \mathbf{z}}{\text{minimize}} && \log \left(\frac{1}{n} \sum_{i=1}^n \varepsilon_i^2 \right) + \frac{2k}{n} \\
\iff & \underset{\mathbf{a}, \mathbf{b}, \boldsymbol{\varepsilon}, k, \mathbf{z}}{\text{minimize}} && \exp \left(\log \left(\frac{1}{n} \sum_{i=1}^n \varepsilon_i^2 \right) + \frac{2k}{n} \right) \\
\iff & \underset{\mathbf{a}, \mathbf{b}, \boldsymbol{\varepsilon}, k, \mathbf{z}}{\text{minimize}} && \left(\frac{1}{n} \sum_{i=1}^n \varepsilon_i^2 \right) \cdot \exp \left(\frac{2k}{n} \right) \\
\iff & \underset{\mathbf{a}, \mathbf{b}, \boldsymbol{\varepsilon}, k, \mathbf{z}}{\text{minimize}} && \left(\sum_{i=1}^n \varepsilon_i^2 \right) \cdot \exp \left(\frac{2k}{n} \right).
\end{aligned}$$

Introducing a continuous variable f that represents an upper bound of $(\sum_{i=1}^n \varepsilon_i^2) \cdot \exp(2k/n)$, we have an intermediate formulation as follows:

$$\begin{aligned}
& \underset{\mathbf{a}, \mathbf{b}, \boldsymbol{\varepsilon}, f, k, \mathbf{z}}{\text{minimize}} && f \\
\text{subject to} & && \sum_{i=1}^n \varepsilon_i^2 \leq f \cdot \exp \left(-\frac{2k}{n} \right), \tag{22}
\end{aligned}$$

Constraints (10), (11), (12) and (13).

Now we resolve the nonlinearity in Constraint (22). Note that k is always integer-valued due to Constraints (12) and (13). Let w_j ($j = 0, 1, \dots, p$) be a 0-1 variable such that $w_j = 1$ if and only if $j = k$; this is achieved by the following constraints: $\sum_{j=0}^p (j \cdot w_j) = k$ and $\sum_{j=0}^p w_j = 1$. With

0-1 variables w_j , Constraint (22) is equivalently changed as follows:

$$\begin{aligned}
& \sum_{i=1}^n \varepsilon_i^2 \leq f \cdot \exp\left(-\frac{2k}{n}\right) \\
\iff & \sum_{i=1}^n \varepsilon_i^2 \leq f \cdot g, \quad g = \exp\left(-\frac{2k}{n}\right) \\
\iff & \begin{cases} \sum_{i=1}^n \varepsilon_i^2 \leq f \cdot g, \quad g = \sum_{j=0}^p \left(w_j \cdot \exp\left(-\frac{2j}{n}\right)\right), \\ \sum_{j=0}^p (j \cdot w_j) = k, \quad \sum_{j=0}^p w_j = 1, \quad w_j \in \{0, 1\} \quad (j = 0, 1, \dots, p), \end{cases}
\end{aligned}$$

where g is another continuous variable. Note that the constraint $g = \sum_{j=0}^p (w_j \cdot \exp(-2j/n))$ is a linear function with respect to variables w_j . This linearization technique of a nonlinear function using 0-1 variables is called “special ordered set type 1” [5, 6], which is well-known in the area of integer programming.

Although the constraint $\sum_{i=1}^n \varepsilon_i^2 \leq f \cdot g$ is still nonlinear, it is a hyperbolic constraint and thus representable as a second-order cone constraint (see Appendix A.) Consequently, we obtain an MISOCP formulation for minimizing

AIC as follows:

$$\begin{array}{ll} \text{minimize} & f \\ & \mathbf{a}, \mathbf{b}, \boldsymbol{\varepsilon}, f \\ & g, k, \mathbf{w}, \mathbf{z} \end{array} \quad (23)$$

$$\text{subject to} \quad \varepsilon_i = y_i - \left(b + \sum_{j=1}^p a_j x_{ij} \right) \quad (i = 1, 2, \dots, n), \quad (24)$$

$$\sum_{i=1}^n \varepsilon_i^2 \leq f \cdot g, \quad (25)$$

$$g = \sum_{j=0}^p \left(w_j \cdot \exp \left(-\frac{2j}{n} \right) \right), \quad (26)$$

$$\sum_{j=0}^p (j \cdot w_j) = k, \quad (27)$$

$$\sum_{j=0}^p w_j = 1, \quad (28)$$

$$\sum_{j=1}^p z_j = k, \quad (29)$$

$$-Mz_j \leq a_j \leq Mz_j \quad (j = 1, 2, \dots, p), \quad (30)$$

$$w_j \in \{0, 1\} \quad (j = 0, 1, \dots, p), \quad (31)$$

$$z_j \in \{0, 1\} \quad (j = 1, 2, \dots, p). \quad (32)$$

Next, we propose an MISOCP formulation for minimizing BIC. The difference between minimizing AIC and BIC lies only the second terms of the objective functions (7) and (8), respectively. Hence, replacing $2j$ in Constraint (26) with $j \log n$ yields the following constraint:

$$g = \sum_{j=0}^p \left(w_j \cdot \exp \left(-\frac{j \log n}{n} \right) \right) = \sum_{j=0}^p (w_j \cdot n^{-j/n}). \quad (33)$$

Consequently, as an MISOCP formulation for minimizing BIC, we obtain the problem consisting of (23)–(25), (27)–(32) and (33).

4. Computational experiments and discussion

In this section, we report computational results to evaluate the proposed MISOCP formulations, which are compared to well-known stepwise regres-

sion methods, and have discussion on the results.

For computational experiments, we downloaded eight data sets from UCI Machine Learning Repository [4] for regression analysis. The data set **SolarFlare** has three variables (i.e., each class of flares production) to be predicted, and accordingly 10 instances of variable selection problems were prepared. The list of the 10 instances are shown in Table 1, where n and p are the number of data points and that of candidate variables, respectively.

For the data set **ForestFires**, we created interaction terms from the variables of x-axis and y-axis spatial coordinates. In the data set **Crime**, variables having missing values for most of the samples (i.e., data points) were removed. For all data sets, each categorical variable was transformed into as many dummy variables as its distinct values. To avoid numerical instability, each integer and real variable was standardized so that its mean becomes zero and its standard deviation becomes one. In addition, samples including a missing value and redundant variables having a constant value were all eliminated.

Table 1: List of the instances.

abbreviation	n	p	original dataset [4]
Housing	506	13	Housing
Servo	167	19	Servo
AutoMPG	392	25	Auto MPG
SolarFlareC	1066	26	Solar Flare (C-class flares production)
SolarFlareM	1066	26	Solar Flare (M-class flares production)
SolarFlareX	1066	26	Solar Flare (X-class flares production)
BreastCancer	194	32	Breast Cancer Wisconsin
ForestFires	517	63	Forest Fires
Automobile	159	65	Automobile
Crime	1933	100	Communities and Crime

We solved \bar{R}^2 maximization, AIC and BIC minimization problems via the proposed MISOCP formulations² using CPLEX 12.5 [19] as a mathematical programming solver. All computations for solving MISOCP problems were

²In MISOCP formulations, we implemented the big- M method via `indicator`, a function implemented in CPLEX. Using this functions allows us not to manually determine the value of M in Constraints (18) and (30).

performed on a Dell Precision T5500 PC³. For each instances, 16 GB memory, eight threads and up to 10,000 seconds were assigned for a branch-and-bound procedure. For comparison, we also solved the same instances via stepwise regression methods using `LinearModel.stepwise` function implemented in the statistics toolbox of MATLAB R2012b [25] on a NEC Mate J PC⁴.

The results for maximizing \bar{R}^2 , minimizing AIC and BIC are shown in Tables 2, 3 and 4, respectively. The “method” column shows

- $\text{SWR}_{\text{const}}$: stepwise regression starting with no explanatory variables,
- SWR_{all} : stepwise regression starting with all candidate variables,
- MISOCP: the proposed MISOCP formulations, i.e., (14)–(20) for maximizing \bar{R}^2 , (23)–(32) for minimizing AIC, and (23)–(25), (27)–(32) and (33) for minimizing BIC,

where both stepwise regression methods iteratively add or eliminate a variable to improve the corresponding GOF measure. For each instance in the tables, the best \bar{R}^2 /AIC/BIC value(s) among by $\text{SWR}_{\text{const}}$, SWR_{all} and MISOCP is bold-faced. The column “ k ” is the number of selected variables and the column “time (s)” is computational time in seconds. In MISOCP, each computation was terminated if computational time took more than 10,000 seconds; in such cases, the obtained subset of variables are not necessarily optimal, otherwise the subset is the best.

First, from the results, we found out that the difference between the results by $\text{SWR}_{\text{const}}$ and those by SWR_{all} are large. From Tables 2, 3 and 4, we observed that $\text{SWR}_{\text{const}}$ and SWR_{all} selected quite different sets of variables in many cases. For example, in the **Automobile** instance, the differences in the obtained AIC values between $\text{SWR}_{\text{const}}$ and SWR_{all} are more than 20, which is hard to ignore. Additionally, in many cases, the number k of selected variables also greatly differs between $\text{SWR}_{\text{const}}$ and SWR_{all} .

Next, the results show that MISOCP finished selecting the subsets of variables in minutes for small-sized instances involving less than 30 candidate variables. Note that these obtained subsets of variables are proved to be the

³CPU: Intel Xeon W5590 3.33 GHz×2; RAM: 24 GB; OS: 64bit Windows 7 Ultimate SP1; chipset: Intel 5520.

⁴CPU: Intel Core i7-2600S 2.80 GHz; RAM: 8 GB; OS: 64bit Windows 7 Professional SP1; chipset: Intel Q67 Express.

Table 2: Results for maximizing \bar{R}^2 .

instance	n	p	method	\bar{R}^2	k	time (s)
Housing	506	13	SWR _{const}	0.7348	11	1.12
			SWR _{all}	0.7338	13	0.18
			MISOCP	0.7348	11	9.03
Servo	167	19	SWR _{const}	0.7419	10	1.75
			SWR _{all}	0.7348	15	0.52
			MISOCP	0.7419	10	4.43
AutoMPG	392	25	SWR _{const}	0.8683	17	3.39
			SWR _{all}	0.8669	22	0.71
			MISOCP	0.8686	16	29.83
SolarFlareC	1066	26	SWR _{const}	0.1869	11	3.28
			SWR _{all}	0.1818	20	1.34
			MISOCP	0.1869	11	184.97
SolarFlareM	1066	26	SWR _{const}	0.0955	9	2.80
			SWR _{all}	0.0873	20	1.24
			MISOCP	0.0955	9	95.61
SolarFlareX	1066	26	SWR _{const}	0.1295 [†]	6	1.90
			SWR _{all}	0.1195	20	1.30
			MISOCP	0.1295[‡]	6	19.03
BreastCancer	194	32	SWR _{const}	0.2305	11	3.41
			SWR _{all}	0.1999	32	0.51
			MISOCP	0.2494	16	3211.08
ForestFires	517	63	SWR _{const}	0.1006	22	15.65
			SWR _{all}	0.0558	60	3.72
			MISOCP	0.1024	26	> 10000
Automobile	159	65	SWR _{const}	0.9656	43	24.02
			SWR _{all}	0.9630	55	5.71
			MISOCP	0.9674	35	> 10000
Crime	1933	100	SWR _{const}	0.6839	65	104.63
			SWR _{all}	0.6796	100	8.89
			MISOCP	0.6841	53	> 10000

[†]0.129511, [‡]0.129512

best through integer programming methodology, and this is in clear contrast

Table 3: Results for minimizing AIC.

instance	n	p	method	AIC	k	time (s)
Housing	506	13	SWR _{const}	776.36	11	1.31
			SWR _{all}	776.36	11	0.51
			MISOCP	776.36	11	10.62
Servo	167	19	SWR _{const}	258.66	9	1.85
			SWR _{all}	266.36	14	0.75
			MISOCP	258.66	9	8.41
AutoMPG	392	25	SWR _{const}	333.22	15	3.96
			SWR _{all}	339.44	19	1.59
			MISOCP	333.22	15	51.23
SolarFlareC	1066	26	SWR _{const}	2816.34	9	3.09
			SWR _{all}	2819.73	13	4.02
			MISOCP	2816.34	9	227.25
SolarFlareM	1066	26	SWR _{const}	2926.93	7	2.38
			SWR _{all}	2926.93	7	5.99
			MISOCP	2926.93	7	92.18
SolarFlareX	1066	26	SWR _{const}	2882.81	3	1.20
			SWR _{all}	2882.81	3	7.59
			MISOCP	2882.81	3	10.73
BreastCancer	194	32	SWR _{const}	509.72	8	3.07
			SWR _{all}	510.58	14	8.13
			MISOCP	508.73	10	> 10000
ForestFires	517	63	SWR _{const}	1429.81	12	9.56
			SWR _{all}	1429.81	12	71.19
			MISOCP	1430.25	13	> 10000
Automobile	159	65	SWR _{const}	-26.87	21	14.20
			SWR _{all}	-47.50	38	24.75
			MISOCP	-58.49	32	> 10000
Crime	1933	100	SWR _{const}	3424.26	41	84.94
			SWR _{all}	3410.92	50	312.82
			MISOCP	3419.65	51	> 10000

to heuristic approaches. Although the MISOCP problems for `ForestFires`, `Automobile` and `Crime` instances were not solved within 10,000 seconds, the obtained subsets of variables are comparable to those obtained by SWR_{const}

Table 4: Results for minimizing BIC.

instance	n	p	method	BIC	k	time (s)
Housing	506	13	SWR _{const}	834.88	8	1.04
			SWR _{all}	827.07	11	0.47
			MISOCP	827.07	11	13.26
Servo	167	19	SWR _{const}	288.93	8	1.60
			SWR _{all}	303.87	11	1.50
			MISOCP	288.93	8	10.51
AutoMPG	392	25	SWR _{const}	390.96	11	3.20
			SWR _{all}	405.71	14	2.92
			MISOCP	390.96	11	59.92
SolarFlareC	1066	26	SWR _{const}	2855.93	6	2.04
			SWR _{all}	2855.89	6	6.51
			MISOCP	2855.89	6	73.73
SolarFlareM	1066	26	SWR _{const}	2956.02	4	1.52
			SWR _{all}	2954.42	4	6.96
			MISOCP	2954.42	4	20.59
SolarFlareX	1066	26	SWR _{const}	2900.12	2	0.93
			SWR _{all}	2900.12	2	7.83
			MISOCP	2900.12	2	5.52
BreastCancer	194	32	SWR _{const}	529.28	3	1.34
			SWR _{all}	528.90	3	11.70
			MISOCP	527.86	3	1198.73
ForestFires	517	63	SWR _{const}	1463.81	3	2.82
			SWR _{all}	1463.81	3	71.63
			MISOCP	1463.81	3	> 10000
Automobile	159	65	SWR _{const}	31.28	15	10.56
			SWR _{all}	42.59	27	35.19
			MISOCP	20.81	23	> 10000
Crime	1933	100	SWR _{const}	3574.68	13	24.92
			SWR _{all}	3594.84	22	390.06
			MISOCP	3591.94	16	> 10000

and SWR_{all}. In addition, for all instances in maximizing \bar{R}^2 , MISOCP had the maximal values among the three methods. This result suggests that the stepwise regression methods have difficulty in selecting the best subset of

Table 5: The number of the best \bar{R}^2 /AIC/BIC values obtained out of 10 instances.

method	\bar{R}^2	AIC	BIC
SWR _{const}	4	7	5
SWR _{all}	5	5	5
MISOCP	10	8	9

Table 6: The number of the best \bar{R}^2 /AIC/BIC values obtained for instances that were not solved within 10,000 seconds via MISOCP.

method	\bar{R}^2	AIC	BIC
SWR _{const}	0	1	2
SWR _{all}	0	2	1
MISOCP	3	2	2

variables based on \bar{R}^2 .

Table 5 shows the number of the best \bar{R}^2 /AIC/BIC values obtained by each method out of 10 instances; Table 6 shows the number of the best \bar{R}^2 /AIC/BIC values only for instances that were not solved within 10,000 seconds via MISOCP. These tables clearly prove the superiority of MISOCP, even for large instances. Stepwise regression methods are greedy-type heuristic methods and naturally do not always provide the best subset of explanatory variables; however, it is also observed that the results by SWR_{const} and SWR_{all} are less robust than we expected.

About computational time, the MISOCP approaches took much longer than the stepwise regression methods did. This is the difference between the exact method that pursues the proof of optimality by a branch-and-bound procedure, and the heuristic nature of stepwise regression methods.

In fact, for instances that were solved within 10,000 seconds via the MISOCP approach, solving $p+1$ MIQP problems (see Section 2.4) was faster than solving an MISOCP problem; whereas the MIQP approach did not solve `ForestFires`, `Automobile` and `Crime` instances within 10,000 seconds, as neither the MISOCP approach did. This phenomenon is explained as follows. In each node of a branch-and-bound procedure for an MIQP problem, a dual-simplex method implemented in a MIP solver handles a QP problem. At a child node in a branch-and-bound tree, a QP problem to be solved is al-

most the same as its parent node. In that case, a dual-simplex method needs a few iterations to solve the problem, i.e., “warm-start” works well for an MIQP problem. In contrast, for an MISOCP problem, a branch-and-bound procedure solves a continuous SOCP problem, which needs an interior point method. Developing warm-start algorithms for an interior point method is still in progress in the area of mathematical programming, and such algorithms are not yet implemented in current commercial solvers that can handle MISOCP.

Although the MISOCP approach needs a longer computational time than the MIQP approach at this time, we emphasize that the proposed formulation technique allows us to transform the variable selection problem into a single MISOCP problem, not a collection of problems. Numerical techniques for solving SOCP/MISOCP problems are areas of active research, and thus the proposed MISOCP formulations are expected to be more valuable in the near future.

Finally, we discuss minimizing other information criteria. Other than AIC and BIC, there are several information criteria proposed so far, e.g., corrected AIC [30] and Hannan-Quinn information criterion [15]. Using the proposed transformation technique, we can also formulate a problem of minimizing such an information criterion as an MISOCP problem.

5. Conclusion

This paper considered selecting the best subset of variables through the use of several GOF measures in a multiple linear regression model. We proposed formulations for maximizing \bar{R}^2 , minimizing AIC and BIC without prespecifying the number k of selected variables, whereas previous researches using integer programming need to specify k . The proposed formulations are MISOCP problems, whose continuous relaxation belong to the class SOCP, and thus solvable using a branch-and-bound procedure.

Through computational experiments, we compared the performance of the proposed MISOCP formulations with stepwise regression methods, well-known variable selection algorithms. We observed that the MISOCP formulations successfully selected the best subset of variables in minutes when the number of candidate variables is less than 30. Moreover, even when the number of candidate variables is more than 60, in many cases the MISOCP formulations found a better subset of variables than that generated by the stepwise regression methods.

In contrast to the stepwise methods, the proposed method proves the optimality of the selected subset of explanatory variables when the associated MISOCP problem is successfully solved. Since proper variable selection is essential for obtaining a correct result of data analysis, this study has a great advantage over heuristic methods.

Acknowledgment

This work was partially supported by Grants-in-Aid for Scientific Research, by the Ministry of Education, Culture, Sports, Science and Technology of Japan.

Appendix A. Second-order cone programming

This appendix adds a supplementary explanation for second-order cone programming (SOCP). For more detail, see [24].

A general form of an SOCP problem is given as follows:

$$\begin{aligned} & \underset{\mathbf{x}}{\text{minimize}} && \mathbf{c}_0^\top \mathbf{x} \\ & \text{subject to} && \|A_i \mathbf{x} + \mathbf{b}_i\| \leq \mathbf{c}_i^\top \mathbf{x} + d_i \quad (i = 1, 2, \dots, m), \end{aligned}$$

where $\|\mathbf{u}\| = (\mathbf{u}^\top \mathbf{u})^{1/2}$. The constraint $\|A_i \mathbf{x} + \mathbf{b}_i\| \leq \mathbf{c}_i^\top \mathbf{x} + d_i$ is called a second-order cone constraint. Since the constraint becomes linear when A_i is a null matrix and \mathbf{b}_i is a zero vector, the class SOCP includes linear programming as a special case. As well as a linear programming problem, an SOCP problem is solvable in polynomial time by using an interior point method [2, 24]; several mathematical programming solvers can handle an SOCP problem.

A hyperbolic constraint takes the form of $x^2 \leq f \cdot g$, $f \geq 0$, $g \geq 0$ for scalar variables x , f and g . A linear programming problem with a hyperbolic constraint is representable as an SOCP problem, because a hyperbolic constraint can be represented as a second-order cone constraint as below:

$$x^2 \leq f \cdot g, \quad f \geq 0, \quad g \geq 0 \quad \iff \quad \left\| \begin{pmatrix} 2x \\ f - g \end{pmatrix} \right\| \leq f + g.$$

In addition, when the left-hand-side of a hyperbolic constraint is a product of a variable vector \mathbf{x} , the constraint is also expressed as a second-order cone

constraint, because

$$\mathbf{x}^\top \mathbf{x} \leq f \cdot g, \quad f \geq 0, \quad g \geq 0 \quad \iff \quad \left\| \begin{pmatrix} 2\mathbf{x} \\ f - g \end{pmatrix} \right\| \leq f + g.$$

On the assumption that $\mathbf{c}^\top \mathbf{x} + d$ is always positive, a problem of the following form can also be casted as an SOCP problem:

$$\underset{\mathbf{x}}{\text{minimize}} \quad \frac{\mathbf{x}^\top \mathbf{x}}{\mathbf{c}^\top \mathbf{x} + d} \tag{A.1}$$

$$\text{subject to} \quad \mathbf{A}\mathbf{x} \leq \mathbf{b}; \tag{A.2}$$

because it is equivalent to the following SOCP problem:

$$\underset{\mathbf{x}, f, g}{\text{minimize}} \quad f \tag{A.3}$$

$$\text{subject to} \quad \mathbf{x}^\top \mathbf{x} \leq f \cdot g, \tag{A.4}$$

$$\mathbf{A}\mathbf{x} \leq \mathbf{b}, \tag{A.5}$$

$$\mathbf{c}^\top \mathbf{x} + d = g, \tag{A.6}$$

$$f \geq 0, \tag{A.7}$$

$$g \geq 0. \tag{A.8}$$

A mixed integer SOCP (MISOCP) problem is an SOCP problem with integrality constraints over a part of variables. Making good use of the fact that the continuous relaxation problem of MISOCP is an SOCP problem, which is efficiently solvable, some recent integer programming solvers can handle an MISOCP problem by using a branch-and-bound procedure.

References

- [1] H. Akaike, “A New Look at the Statistical Model Identification,” *IEEE Transactions on Automatic Control*, Vol.19, No.6, pp.716–723 (1974).
- [2] F. Alizadeh and D. Goldfarb, “Second-Order Cone Programming,” *Mathematical Programming*, Vol.95, No.1, pp.3–51 (2003).
- [3] T.S. Arthanari and Y. Dodge, *Mathematical Programming in Statistics* (John Wiley & Sons, 1981).

- [4] K. Bache and M. Lichman: UCI Machine Learning Repository. University of California, School of Information and Computer Science, Irvine, 2013. <http://archive.ics.uci.edu/ml>
- [5] E.M.L. Beale, “Two Transportation Problems,” *Proceedings of the 3rd International Conference on Operational Research*, pp.780–788 (1963).
- [6] E.M.L. Beale and J.A. Tomlin, “Special Facilities in a General Mathematical Programming System for Non-Convex Problems Using Ordered Sets of Variables,” *Proceedings of the 5th International Conference on Operational Research*, pp.447–454 (1970).
- [7] D. Bertsimas and R. Shioda, “Algorithm for Cardinality-Constrained Quadratic Optimization,” *Computational Optimization and Applications*, Vol.43, No.1, pp.1–22 (2009).
- [8] A.L. Blum and P. Langley, “Selection of Relevant Features and Examples in Machine Learning,” *Artificial Intelligence*, Vol.97, No.1–2, pp.245–271 (1997).
- [9] K.P. Burnham and D.R. Anderson, *Model Selection and Multimodel Inference: A Practical Information Theoretic Approach, 2nd Edition* (Springer, 2002).
- [10] M.A. Efron, “Multiple Regression Analysis,” In A. Ralston, and H.S. Wilf (Eds.), *Mathematical Methods for Digital Computers*, pp. 191–203 (Wiley, 1960).
- [11] S. Emet, “A Model Identification Approach Using MINLP Techniques,” *Proceedings of the 9th WSEAS International Conference on Applied Mathematics*, pp.347–350 (2006).
- [12] G.M. Furnival and R.W. Wilson Jr., “Regressions by Leaps and Bounds,” *Technometrics*, Vol.16, No.4, pp.499–511 (1974).
- [13] C. Gatu and E.J. Kontoghiorghes, “Branch-and-Bound Algorithms for Computing the Best-Subset Regression Models,” *Journal of Computational and Graphical Statistics*, Vol.15, No.1, pp.139–156 (2006).
- [14] I. Guyon and A. Elisseeff, “An Introduction to Variable and Feature Selection,” *Journal of Machine Learning Research*, Vol.3 (March), pp.1157–1182 (2003).

- [15] E.J. Hannan and B.G. Quinn, “The Determination of the Order of an Autoregression,” *Journal of the Royal Statistical Society, Series B*, Vol.41, No.2, pp.190–195 (1979).
- [16] R.R. Hocking, “The Analysis and Selection of Variables in Linear Regression,” *Biometrics*, Vol.32, No.1, pp.1–49 (1976).
- [17] A.E. Hoerl and R.W. Kennard, “Ridge Regression: Biased Estimation for Non-Orthogonal Problems,” *Technometrics*, Vol.20, No.1, pp.55–67 (1970).
- [18] M. Hofmann, C. Gatu, and E.J. Kontoghiorghes, “Efficient Algorithms for Computing the Best Subset Regression Models for Large-Scale Problems,” *Computational Statistics & Data Analysis*, Vol.52, No.1, pp.16–29 (2007).
- [19] IBM ILOG, IBM ILOG CPLEX 12.5, 2012.
- [20] R. Kohavi and G.H. John, “Wrappers for Feature Subset Selection,” *Artificial Intelligence*, Vol.97, No.1–2, pp.273–324 (1997).
- [21] H. Konno and Y. Takaya, “Multi-Step Methods for Choosing the Best Set of Variables in Regression Analysis,” *Computational Optimization and Applications*, Vol.46, No.3, pp.417–426 (2010).
- [22] H. Konno and R. Yamamoto, “Choosing the Best Set of Variables in Regression Analysis Using Integer Programming,” *Journal of Global Optimization*, Vol.44, No.2, pp.272–282 (2009).
- [23] H. Liu and H. Motoda, *Computational Methods of Feature Selection* (Chapman and Hall/CRC, 2007).
- [24] M.S. Lobo, L. Vandenberghe, S. Boyd, and H. Lebert, “Applications of Second-Order Cone Programming,” *Linear Algebra and its Applications*, Vol.284, No.1, pp.193–228 (1998).
- [25] The MathWorks Inc., MATLAB R2012b, 2012.
- [26] A. Miller, *Subset Selection in Regression, 2nd Edition* (Chapman and Hall/CRC, 2002).

- [27] R. Meiri and J. Zahavi, “Using Simulated Annealing to Optimize the Feature Selection Problem in Marketing Applications,” *European Journal of Operational Research*, Vol.171, No.3, pp.842–858 (2006).
- [28] G. Schwarz, “Estimating the Dimension of a Model,” *Annals of Statistics*, Vol.6, No.2, pp.461–464 (1978).
- [29] H. Skrifvars, S. Leyffer, and T. Westerlund, “Comparison of Certain MINLP Algorithms When Applied to a Model Structure Determination and Parameter Estimation Problem,” *Computers & Chemical Engineering*, Vol.22, No.12, pp.1829–1835 (1998).
- [30] N. Sugiura, “Further Analysts of the Data by Akaike’s Information Criterion and the Finite Corrections,” *Communications in Statistics — Theory and Methods*, Vol.7, No.1 (1978), pp. 13–26.
- [31] R. Tibshirani, “Regression Shrinkage and Selection via the Lasso,” *Journal of the Royal Statistical Society, Series B*, Vol.58, No.1, pp.267–288 (1996).
- [32] H. Theil, *Economic Forecasts and Policy* (North-Holland Publishing Company, 1961).
- [33] M.J. Whittingham, P.A. Stephens, R.B. Bradbury, and R.P. Freckleton, “Why Do We Still Use Stepwise Modelling in Ecology and Behaviour?” *Journal of Animal Ecology*, Vol.75, No.5, pp.1182–1189 (2006).