Department of Industrial Engineering and Management

Technical Report

No. 2014-1

Subset Selection by Mallows' Cp: A Mixed Integer Programming Approach

Ryuhei Miyashiro and Yuichi Takano

January, 2014

Tokyo Institute of Technology

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

Subset Selection by Mallows' C_p: A Mixed Integer Programming Approach

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

Yuichi Takano takano.y.ad@m.titech.ac.jp
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 a method of selecting the best subset of explanatory variables for a linear regression model. Employing Mallows' C_p as a goodness-of-fit measure, we formulate the subset selection problem as a mixed integer quadratic programming problem. Computational results demonstrate that our method provides the best subset of variables in a few seconds when the number of candidate explanatory variables is less than 30. Furthermore, when handling datasets consisting of a large number of samples, it finds better-quality solutions faster than stepwise regression methods do.

Keywords: Subset selection, Mixed integer programming, Mallows' C_p, Linear regression model

1 Introduction

Subset selection, or variable/feature/attribute selection, is of essential importance in statistics [7, 15, 23], and it has recently received considerable attention in data mining and machine learning as a result of the increased size of the datasets dealt with in these fields [21]. Previous research has proposed a number of computational methods for subset selection (see, e.g., [5, 12, 18, 21]), and many of them are categorized as heuristic algorithms. Among them is the well-known stepwise regression method [9], which repeats forward selection (adding one significant variable) and backward elimination (eliminating one redundant variable) until a stopping condition is satisfied. Although these heuristic algorithms often yield good-quality solutions to large-scale problems, they do not necessarily find an optimal solution.

This paper focuses on determining the best subset of explanatory variables for a linear regression model by means of the mixed integer programming (MIP) methodology. It should be understood that finding an optimal solution under some goodness-of-fit (GOF) measure does not correspond to identifying a "true" model. It is expected, nevertheless, that optimal solutions lead to better models than heuristic solutions do. Hence, they are quite useful for identifying a model that captures the essence of a system. Additionally, optimal solutions can be used to evaluate the performance of heuristic algorithms and the quality of their solutions. For instance, if the optimal GOF value is available, it is possible to know how far a heuristic solution is from it. Even if an optimal solution can be found only for datasets with a small number of samples, this solution can be used as a starting point in heuristic algorithms for datasets with a larger number

of samples. Exact algorithms can also be employed as a subprocedure of heuristic algorithms. This sort of hybrid algorithm has been drawing interest from researchers in recent years (see, e.g., [29]).

When the number of explanatory variables to be selected is given a priori, subset selection is usually carried out so that the residual sum of squares (RSS) of the subset regression model is minimized. It is known that this subset selection problem can be framed as a mixed integer quadratic programming (MIQP) problem [2, 4, 20]. Bertsimas and Shioda [4] developed a tailored branch-and-bound procedure to solve such MIQP problems efficiently. Meanwhile, the studies [2, 19, 20] employed the mean absolute deviation instead of RSS and reformulated the subset selection problem as a mixed integer linear programming (MILP) problem. Moreover, Konno and Takaya [19] developed a multi-step method, which iteratively reduces the number of explanatory variables by using integer programming. This method can be regarded as one of the hybrid heuristic algorithms mentioned above. These MIQP/MILP approaches, however, need to decide how many explanatory variables to select before solving the problems. This is disadvantageous in many cases; accordingly, it is more effective to select a subset of explanatory variables according to an appropriate GOF measure.

There are several GOF measures for evaluating a subset regression model. One of the most widely used is the Akaike information criterion (AIC) [1]. Although Emet [10] and Skrifvars et al. [26] studied MIP formulations of minimizing AIC, this problem is computationally intractable due to its integrality, nonlinearity and nonconvexity. On the other hand, the authors of the present paper proposed in [24] mixed integer second-order cone programming (MISOCP) formulations for subset selection. This sort of formulation enables one to find the best subset of explanatory variables in terms of various GOF measures such as AIC, adjusted R^2 [31], Bayesian information criterion [25], corrected AIC [28] and Hannan-Quinn information criterion [13]. The proposed MISOCP formulations, whose continuous relaxation problems are nonlinear but convex, can be handled by recent mathematical programming solvers using a branch-and-bound procedure; however, the experiments in [24] showed that substantial time is required for solving them.

This paper makes use of Mallows' C_p [22] as a GOF measure and derives an MIQP formulation for subset selection. It is known that minimizing AIC is approximately equivalent to minimizing C_p for a linear regression model (see e.g., [6, 23]). Although similar MIQP formulations were presented in [10, 26] as a variant of the AIC minimization problem, the authors of those studies made no mention of Mallows' C_p. Additionally, detailed computational results were not provided in [10, 26]. By contrast, we conducted careful computational experiments to assess the effectiveness of the MIQP formulation on various datasets from the UCI Machine Learning Repository [3].

The contributions of the present paper are summarized as follows:

- We propose the MIQP formulation for minimizing Mallows' C_p. It is noteworthy that solving this problem is much more efficient than solving the MISOCP problems in [24]. The computational results demonstrate that our method can provide a subset of variables

with optimality guarantees in a few seconds when the number of candidate explanatory variables is less than 30.

- We verify that our MIQP formulation, together with a state-of-the-art mathematical programming solver, has a computational advantage over stepwise regression methods, the commonly-used heuristics. It takes a fair amount of time to solve an MIQP problem exactly for large-sized instances. Nevertheless, we show that our MIQP formulation can find better-quality solutions faster than stepwise regression methods do.

2 Linear Regression Model and Mallows' C_p

2.1 Linear regression model

Let us focus on the linear regression model:

$$y = a_0 + a_1 x_1 + a_2 x_2 + \cdots + a_p x_p + \varepsilon, \tag{1}$$

where y is an explained variable (or dependent variable), x_j for $j = 1, 2, \ldots, p$ are explanatory variables (or independent variables), a_j for $j = 0, 1, \ldots, p$ are unknown parameters to be estimated, and ε is a prediction residual. Since the model (1) incorporates all candidate explanatory variables, we refer to (1) as a full model.

Suppose that we have data consisting of n samples $(y_i; x_{i1}, x_{i2}, \ldots, x_{ip})$ for $i = 1, 2, \ldots, n$. Then, the full model (1) can be rewritten as

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

where $\boldsymbol{y} := (y_1\ y_2\ \cdots\ y_n)^\top$, $\boldsymbol{a} := (a_0\ a_1\ \cdots\ a_p)^\top$, $\boldsymbol{\varepsilon} := (\varepsilon_1\ \varepsilon_2\ \cdots\ \varepsilon_n)^\top$, and

$$\boldsymbol{X} := \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}.$$

The ordinary least squares (OLS) method estimates the coefficients, $\boldsymbol{a}$, such that the residual sum of squares (RSS):

$$\boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon} = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{a})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{a}) = \boldsymbol{y}^\top \boldsymbol{y} - 2\boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{a} + \boldsymbol{a}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{a} \tag{2}$$

is minimized. After partial differentiation, the OLS estimator for the full model (1) becomes

$$\hat{\boldsymbol{a}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}.$$

2.2 Mallows' C_p

Mallows' C_p [22] is a goodness-of-fit (GOF) measure that is frequently used for evaluating the linear regression model. It is assumed that the residuals ε_i for $i = 1, 2, \ldots, n$ are independent random variables with zero mean and unknown variance σ^2. Accordingly, C_p for the full model (1) is defined as

$$C_p^{\text{Full}} = \frac{(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{a}})^\top (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{a}})}{\hat{\sigma}^2} + 2(p+1) - n,$$

where $\hat{\sigma}^2$ is an estimator of the residual variance, σ^2. This estimator is usually set to the unbiased estimator of the full model (1) (see e.g., [22, 23]),

$$\hat{\sigma}^2 = \frac{(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{a}})^\top (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{a}})}{n - p - 1}. \tag{3}$$

The first term of C_p^{Full} is the minimum RSS divided by $\hat{\sigma}^2$. This shows how the model fits the given samples $(y_i; x_{i1}, x_{i2}, \ldots, x_{ip})$ for $i = 1, 2, \ldots, n$ in the least squares sense. On the other hand, the second term, $2(p+1)$, of C_p^{Full} represents the model complexity to be decreased. The principle of parsimony helps to avoid overfitting and computational error, and accordingly improves the generalization capability of the predictive model (see e.g., [14]). Note that C_p^{Full} can be converted into

$$C_p^{\text{Full}} = \min_{\boldsymbol{a}} \frac{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{a})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{a})}{\hat{\sigma}^2} + 2(p+1) - n,$$

because the OLS estimator, $\hat{\boldsymbol{a}}$, minimizes RSS (2).

In what follows, we will consider how to select the best subset of explanatory variables. Specifically, we address the subset regression model with explanatory variables x_j for $j \in S$, where $S \subseteq \{1, 2, \ldots, p\}$ is the index set of selected variables.

Eliminating the explanatory variable x_j is equivalent to fixing its coefficient a_j to zero. Accordingly, C_p for the subset model is

$$C_p(S) = \min_{\boldsymbol{a}} \left\{ \frac{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{a})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{a})}{\hat{\sigma}^2} \;\middle|\; a_j = 0 \ (j \notin S) \right\} + 2(|S| + 1) - n, \tag{4}$$

where $|S|$ is the number of elements of the set S, i.e., the number of selected variables. Substituting (3) into C_p^{Full}, we see that $C_p^{\text{Full}} = p + 1$. Hence, if $C_p(S)$ is minimized with respect to $S \subseteq \{1, 2, \ldots, p\}$, it will not be more than $p + 1$.

3 Mixed Integer Programming Formulations

3.1 Selecting the best k explanatory variables

Let us consider the case in which the number $k = |S|$ is given, i.e., k explanatory variables are to be selected from p candidate ones. In this case, by omitting constant terms, the minimization of $C_p(S)$ reduces to the following RSS minimization problem:

$$\min_{\boldsymbol{a}, S} \left\{ (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{a})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{a}) \;\middle|\; a_j = 0 \ (j \notin S), \ |S| = k, \ S \subseteq \{1, 2, \ldots, p\} \right\}. \tag{5}$$

This subset selection problem can be posed as a mixed integer quadratic programming (MIQP) problem.

Let us introduce 0-1 decision variables z_j for $j = 1, 2, \ldots, p$ to determine whether the j-th candidate explanatory variable is selected or not; $z_j = 1$ if the j-th variable is selected; $z_j = 0$, otherwise. By using the big-M formulation, the subset selection problem with the fixed k can be expressed as an MIQP problem (see [2, 4, 20]):

$$\begin{aligned}
\underset{\boldsymbol{a},\boldsymbol{z}}{\text{minimize}} \quad & \sum_{i=1}^{n}\left(y_i - \left(a_0 + \sum_{j=1}^{p} a_j x_{ij}\right)\right)^2 && (6)\\
\text{subject to} \quad & -Mz_j \le a_j \le Mz_j \quad (j = 1, 2, \ldots, p), && (7)\\
& \sum_{j=1}^{p} z_j = k, && (8)\\
& z_j \in \{0, 1\} \quad (j = 1, 2, \ldots, p), && (9)
\end{aligned}$$

where M is a sufficiently-large positive constant. Constraint (7) is called a big-M constraint. If $z_j = 0$, the j-th candidate explanatory variable is eliminated from the regression model, because its coefficient a_j has to be 0 from Constraint (7). If the interval $[-M, M]$ is sufficiently large, $z_j = 1$ implies that a_j can take an arbitrary value. Constraint (8) forces the number of selected explanatory variables to be k. Consequently, Problem (6)–(9) is equivalent to Problem (5).

3.2 Subset selection by Mallows' C_p

Problem (6)–(9) enables one to find k explanatory variables that minimize RSS. It is often the case, however, that one would like to determine an appropriate number $k = |S|$ simultaneously. To accomplish this, we shall use Mallows' C_p as a GOF measure.

Considering the representation (4) of $C_p(S)$, the subset selection problem of minimizing Mallows' C_p can be formulated as an MIQP problem:

$$\begin{aligned}
\underset{\boldsymbol{a},\boldsymbol{z}}{\text{minimize}} \quad & \frac{\sum_{i=1}^{n}\left(y_i - \left(a_0 + \sum_{j=1}^{p} a_j x_{ij}\right)\right)^2}{\hat{\sigma}^2} + 2\left(\sum_{j=1}^{p} z_j + 1\right) - n && (10)\\
\text{subject to} \quad & -Mz_j \le a_j \le Mz_j \quad (j = 1, 2, \ldots, p), && (11)\\
& z_j \in \{0, 1\} \quad (j = 1, 2, \ldots, p). && (12)
\end{aligned}$$

Here, the number of selected explanatory variables, $\sum_{j=1}^{p} z_j$, is not pre-specified, whereas it is a given constant, k, in Problem (6)–(9). We can select the best subset of explanatory variables according to C_p by solving Problem (10)–(12).

In Problem (10)–(12), the positive constant M needs to be sufficiently large. If M is not sufficiently large, Problem (10)–(12) cannot guarantee the optimality of the selected explanatory variables. On the other hand, it is known that a large M can cause numerical instabilities in computations (see, e.g., [32]). Mixed logical programming [8, 16] is a remedy for this problem,

and it is supported by several mathematical programming solvers. Specifically, we replace the big-M constraint (11) with its logical implication:

$$\underset{\boldsymbol{a},\boldsymbol{z}}{\text{minimize}} \quad \frac{\sum_{i=1}^{n}\left(y_i-\left(a_0+\sum_{j=1}^{p}a_jx_{ij}\right)\right)^2}{\hat{\sigma}^2}+2\left(\sum_{j=1}^{p}z_j+1\right)-n \tag{13}$$

$$\text{subject to} \quad z_j=0 \;\Rightarrow\; a_j=0 \quad (j=1,2,\ldots,p), \tag{14}$$

$$z_j\in\{0,1\} \quad (j=1,2,\ldots,p). \tag{15}$$

The logical implications (14) mean that if $z_j = 0$, the j-th candidate explanatory variable is eliminated from the regression model. This sort of logical implication can be efficiently handled in a branch-and-bound procedure for MIP problems (see, e.g., [16] for the details).

4 Computational Experiments

The computational results reported in this section compare the efficiency of our MIQP formulation with those of well-known stepwise regression methods.

We downloaded ten datasets for the regression analysis from the UCI Machine Learning Repository [3]. The "Solar Flare" dataset has three variables (i.e., three classes of flares production) to be predicted, and accordingly, twelve instances were prepared. Tables 1 and 2 list these instances, where n and p are the number of samples and number of candidate explanatory variables, respectively.

Table 1: List of small-sized (ten) instances

abbreviation	n	p	original dataset [3]
`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`	1993	100	Communities and Crime

For the `ForestFires` instance, we created interaction terms from the variables of the x-axis and y-axis spatial coordinates. In the `Crime` instance, variables having missing values for most of the samples were removed. For all datasets, each categorical variable was transformed into as many dummy variables as its distinct values. In addition, samples including a missing value and redundant variables having a constant value were all eliminated.

Table 2: List of large-sized instances

abbreviation	n	p	original dataset [3]
`YearMSD`	51630	90	YearPredictionMSD (test set)
`CTSlices`	53500	384	Relative location of CT slices on axial axis

The numerical experiments tested the computational performance of the following methods:

- SWconst: stepwise regression starting with no explanatory variables (i.e., $S = \emptyset$),
- SWall: stepwise regression starting with all candidate explanatory variables (i.e., $S = \{1, 2, \ldots, p\}$),
- MIQP: MIQP formulation (13)–(15).

Both stepwise regression methods iteratively add or eliminate an explanatory variable such that Mallows' C_p is minimized. The estimator, $\hat{\sigma}^2$, of the residual variance was set as in (3).

All computations were performed on a Dell Precision T5500 PC[1]. For each instance, a single thread and 16 GB of memory were allocated to the branch-and-bound procedure. We used CPLEX 12.5 [17] as the mathematical programming solver for MIQP problem (13)–(15). In addition, `indicator`, a function implemented in CPLEX was used to impose logical implications (14). We also performed stepwise regression methods with the `step` function implemented in R 3.0.2 [27].

Table 3 shows the computational results for the small-sized instances listed in Table 1. The columns labeled "C_p", "AIC" and "adj. R^2" are respectively the values of Mallows' C_p [22], Akaike information criterion [1], and adjusted R^2 [31] of the subset regression models built by each method. Note that the best C_p/AIC/adj. R^2 values for each instance are bold-faced. The column labeled "$|S|$" is the number of selected variables, and the column labeled "time (s)" is computation time in seconds. The MIQP computation was terminated if it did not finish by itself after 1000 seconds.

From Table 3, we can see that the two stepwise regression methods, SWconst and SWall, provided different subsets of explanatory variables for all instances except `Housing`. The number of explanatory variables selected by them also varied considerably in several cases; for instance, SWconst and SWall respectively selected 12 and 21 variables for `ForesetFires`. For `BreastCancer`, `Automobile` and `Crime`, the C_p values obtained by the stepwise regression methods were inferior to those of MIQP, which implies that the stepwise regression methods failed to find the best subset of variables. For the other seven instances, at least one of the C_p values obtained by SWconst and SWall was equal to that of MIQP. SWconst and SWall finished their computations in a very short time. More specifically, they required less than 100 seconds for all instances in Table 3.

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

Table 3: Results of the small-sized instances

instance	n	p	method	C_p	AIC	adj. R^2	$\|S\|$	time (s)
`Housing`	506	13	SWconst	**10.11**	**3023.7**	**0.7348**	11	0.40
			SWall	**10.11**	**3023.7**	**0.7348**	11	0.09
			MIQP	**10.11**	**3023.7**	**0.7348**	11	0.30
`Servo`	167	19	SWconst	**6.64**	**408.8**	**0.7405**	9	0.36
			SWall	8.47	410.6	0.7391	10	0.28
			MIQP	**6.64**	**408.8**	**0.7405**	9	0.62
`AutoMPG`	392	25	SWconst	**11.50**	**1945.8**	**0.8685**	15	0.79
			SWall	16.63	1950.9	0.8677	18	0.36
			MIQP	**11.50**	**1945.8**	**0.8685**	15	1.30
`SolarFlareC`	1066	26	SWconst	**5.09**	**2435.8**	**0.1856**	9	0.61
			SWall	10.04	2440.7	0.1849	13	1.26
			MIQP	**5.09**	**2435.8**	**0.1856**	9	6.54
`SolarFlareM`	1066	26	SWconst	**−0.78**	**381.9**	**0.0949**	7	0.40
			SWall	3.23	385.9	0.0931	9	1.29
			MIQP	**−0.78**	**381.9**	**0.0949**	7	4.37
`SolarFlareX`	1066	26	SWconst	**−6.50**	**−2333.8**	**0.1283**	3	0.18
			SWall	2.31	−2325.1	0.1260	9	1.31
			MIQP	**−6.50**	**−2333.8**	**0.1283**	3	0.64
`BreastCancer`	194	32	SWconst	2.83	1885.7	0.2265	8	0.33
			SWall	4.16	1886.3	**0.2389**	12	0.95
			MIQP	**2.19**	**1884.6**	0.2383	10	317.95
`ForestFires`	517	63	SWconst	**−6.79**	**1778.4**	**0.0928**	12	1.00
			SWall	6.25	1791.1	0.0858	21	8.53
			MIQP	**−6.79**	**1778.4**	**0.0928**	12	1000.00
`Automobile`	159	65	SWconst	44.49	2733.4	0.9569	21	1.42
			SWall	28.74	2708.8	0.9658	37	3.37
			MIQP	**16.13**	**2698.3**	**0.9667**	28	1000.00
`Crime`	1993	100	SWconst	37.13	−2379.9	0.6804	41	26.51
			SWall	24.56	−2393.5	**0.6840**	50	95.83
			MIQP	**24.21**	**−2393.8**	0.6839	49	1000.00

By contrast, we quit MIQP computation after 1000 seconds in three instances, `ForestFires`, `Automobile` and `Crime`. In these cases, the subsets of explanatory variables obtained within 1000 seconds were not necessarily the best. Nevertheless, for all instances, MIQP successfully found a subset that was as good as or better than those found by the stepwise regression methods. In particular, for `Automobile`, the obtained C_p values differed substantially among SWconst, SWall and MIQP (44.49, 28.74 and 16.13, respectively). MIQP also attained good AIC and adj. R^2 values relative to those of the stepwise regression methods. In view of these observations, we

Table 4: Mallows' C_p for the `YearMSD` and `CTSlices` instances

instance	n	p	set	SWconst	SWall	MIQP1000s	MIQP10000s
`YearMSD`	1000	90	A	**18.21**	20.77	**18.21**	**18.21**
			B	20.42	19.74	**19.06**	**19.06**
			C	16.67	17.64	**16.51**	**16.51**
			D	36.01	28.43	**28.34**	**28.34**
			E	23.54	**23.15**	**23.15**	**23.15**
	10000	90	A	44.79	**42.53**	**42.53**	**42.53**
			B	51.26	**48.33**	**48.33**	**48.33**
			C	50.97	**49.21**	**49.21**	**49.21**
			D	48.78	**46.08**	**46.08**	**46.08**
			E	37.77	38.67	**37.45**	**37.45**
`CTSlices`	1000	384	A	121.09	**114.87**	120.61	115.21
			B	115.19	102.66	99.89	**99.33**
			C	107.88	111.41	107.58	**107.28**
			D	117.48	**113.12**	120.93	116.34
			E	128.71	132.80	131.64	**123.17**
	10000	384	A	228.54	224.20	225.77	**222.58**
			B	227.90	228.88	232.11	**227.72**
			C	231.99	229.90	229.69	**228.58**
			D	259.98	**243.31**	245.31	243.40
			E	231.05	230.18	230.60	**229.18**

expect that MIQP will be able to find a good-quality solution even if its computation has to be terminated before optimality is proven.

Next, we evaluated the usefulness of MIQP as a heuristic method with a time limit. For this purpose, in addition to SWconst and SWall, we tested the following methods:

- MIQP1000s: MIQP formulation (13)–(15); the computation was terminated in 1000 seconds,
- MIQP10000s: MIQP formulation (13)–(15); the computation was terminated in 10000 seconds,

on the two large-sized instances, `YearMSD` and `CTSlices`, listed in Table 2. Here, five sample sets (A, B, C, D, E) of $n = 1000$ and 10000 were created by drawing samples from each instance. Table 4 shows the Mallows' C_p obtained by each method. Here, the best C_p values for each sample set are bold-faced. Table 5 shows computation time taken by each method to find the best solution.

We begin by evaluating the results of `YearMSD`. Table 4 shows that among the four methods, MIQP1000s and MIQP10000s attained the best C_p values for all ten sample sets. This means that

Table 5: Computation time in seconds for finding the best solution of each method for the `YearMSD` and `CTSlices` instances

instance	n	p	set	SWconst	SWall	MIQP1000s	MIQP10000s
`YearMSD`	1000	90	A	11.9	38.9	126.9	126.9
			B	7.9	47.1	67.4	67.4
			C	9.4	39.2	90.0	90.0
			D	12.0	40.3	892.2	892.2
			E	12.8	37.9	195.4	195.4
	10000	90	A	216.1	328.1	9.2	9.2
			B	189.1	282.1	0.4	0.4
			C	161.2	281.9	109.0	109.0
			D	132.6	294.5	8.7	8.7
			E	145.7	300.5	5.4	5.4
`CTSlices`	1000	384	A	2085.5	6702.2	907.1	4902.2
			B	1098.7	6486.1	940.6	1976.9
			C	1521.4	6355.0	910.6	1655.1
			D	1564.7	6336.1	735.5	9867.2
			E	1931.3	6138.1	887.2	3205.5
	10000	384	A	36077.6	62672.3	862.2	7732.0
			B	38648.2	60506.6	904.8	9334.4
			C	33538.5	68777.5	804.7	8043.8
			D	36055.3	63243.6	849.6	9338.4
			E	38157.2	67799.0	942.6	1790.7

MIQP needs only 1000 seconds to find a good-quality solution. On the other hand, SWconst and SWall always arrived at different solutions. Moreover, we can see that these stepwise regression methods failed to find the best subset of variables for sample sets B, C, D of $n = 1000$ and E of $n = 10000$ because their solutions were inferior to the MIQP ones. Table 5 shows that when $n = 1000$, SWconst found its best solution faster than the other methods found theirs. Conversely, when $n = 10000$, MIQP1000s and MIQP10000s found their best solution much faster than the stepwise regression methods did. It seems strange, however, that MIQP found its best solution faster on average for $n = 10000$ than for $n = 1000$. This is probably because a large number of samples changes the nature of the problem such that good solutions have similar structures. This property is known as the proximate optimality principle [11], and it enhances the performance of heuristic procedures in a mathematical programming solver.

Turning to the results of `CTSlices`, Table 4 reveals that MIQP10000s always attained C_p values lower than those of MIQP1000s. For this reason, we can see that MIQP requires more than 1000 seconds to find an optimal solution. For seven out of ten sample sets, MIQP10000s attained a better C_p value than those of SWconst and SWall. Moreover, Table 5 shows that when $n = 10000$, SWconst and SWall were much more time-consuming than MIQP10000s. The

most prominent example of this is sample set E of $n = 10000$; MIQP10000s found its best solution in 1790.7 seconds, whereas SWconst and SWall respectively took 38157.2 and 67799.0 seconds to arrive at solutions inferior to that of MIQP10000s. Accordingly, we can conclude that our MIQP formulation has clear advantages over stepwise regression methods especially when it comes to handling datasets with a large number of samples.

5 Conclusion

This paper dealt with the problem of subset selection for a linear regression model. To select the best subset of explanatory variables according to Mallows' C_p, we developed a computational method based on mixed integer programming (MIP). The computational results confirmed the effectiveness of our subset selection method. Given sufficient time, the method can find an optimal solution, and even if it must quit before proving optimality, the resulting solution is of good quality.

The contributions of this research are twofold. First, we proposed a mixed integer quadratic programming (MIQP) formulation for minimizing Mallows' C_p without pre-specifying the number of explanatory variables to be selected. It is known that recent mathematical programming solvers can solve MIQP problems efficiently. Indeed, our method found the best subset of explanatory variables in a few seconds when the number of candidate explanatory variables is less than 30. Second, we demonstrated that our MIQP formulation with a state-of-the-art mathematical programming solver outperforms stepwise regression methods as a heuristic method. Even when it had to be terminated before proving optimality, the obtained solution, though not necessarily optimal, was in most cases better in terms of C_p than those of the stepwise regression methods. Moreover, when the number of samples was large (e.g., $n = 10000$), our MIQP formulation frequently found better-quality solutions faster than the stepwise regression methods did.

This study illustrates the fact that the MIP approach is a powerful tool for statistical data analysis. Previously, one had no choice but to depend on heuristic algorithms like the stepwise regression method in order to select a statistical model from a huge number of candidate models. However, such heuristic algorithms do not necessarily find an optimal model. Indeed, we showed that the stepwise regression methods arrived at solutions that were markedly inferior to the optimal ones; see, for instance, the results for `Automobile` in Table 1. The performance of mathematical programming solvers has rapidly improved in recent years. As shown in this paper, the MIP approach has great potential for selecting a good statistical model efficiently.

A future direction of study will be to speed up the MIQP computation by exploiting its problem structure. For instance, Bertsimas and Shioda [4] developed a tailored branch-and-bound algorithm to efficiently solve particular MIQP problems. Moreover, Konno and Takaya [19] developed a multi-step method, which is a hybrid heuristic algorithm using the MIP formulation as a subprocedure. Similar approaches suited to the problem at hand will lead to a fast algorithm.

Acknowledgments

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

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] T.S. Arthanari and Y. Dodge, *Mathematical Programming in Statistics* (John Wiley & Sons, 1981).

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

[4] D. Bertsimas and R. Shioda, "Algorithm for Cardinality-Constrained Quadratic Optimization," *Computational Optimization and Applications*, Vol.43, No.1, pp.1–22 (2009).

[5] 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).

[6] A. Boisbunon, S. Canu, D. Fourdrinier, W. Strawderman, and M.T. Wells, "AIC and C_p as Estimators of Loss for Spherically Symmetric Distributions," Eprint arXiv:1308.2766 (2013).

[7] K.P. Burnham and D.R. Anderson, *Model Selection and Multimodel Inference: A Practical Information Theoretic Approach, 2nd Edition* (Springer, 2002).

[8] R.A Carbonneau, G. Caporossi, and P. Hansen, "Globally Optimal Clusterwise Regression by Mixed Logical-Quadratic Programming," *European Journal of Operational Research*, Vol.212, No.1, pp.213–222 (2011).

[9] M.A. Efroymson, "Multiple Regression Analysis," In A. Ralston and H.S. Wilf (Eds.), *Mathematical Methods for Digital Computers*, pp.191–203 (Wiley, 1960).

[10] S. Emet, "A Model Identification Approach Using MINLP Techniques," *Proceedings of the 9th WSEAS International Conference on Applied Mathematics*, pp.347–350 (2006).

[11] F. Glover and M. Laguna, *Tabu Search* (Kluwer, 1998).

[12] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," *Journal of Machine Learning Research*, Vol.3 (March), pp.1157–1182 (2003).

[13] 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).

[14] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning, 2nd Edition* (Springer, 2009).

[15] R.R. Hocking, "The Analysis and Selection of Variables in Linear Regression," *Biometrics*, Vol.32, No.1, pp.1–49 (1976).

[16] J.N. Hooker and M.A. Osorio, "Mixed Logical-Linear Programming," *Discrete Applied Mathematics*, Vol.96–97, pp.395–442 (1999).

[17] IBM ILOG, IBM ILOG CPLEX 12.5 (2012).

[18] R. Kohavi and G.H. John, "Wrappers for Feature Subset Selection," *Artificial Intelligence*, Vol.97, No.1–2, pp.273–324 (1997).

[19] 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).

[20] 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).

[21] H. Liu and H. Motoda, *Computational Methods of Feature Selection* (Chapman and Hall/CRC, 2007).

[22] C.L. Mallows, "Some Comments on C_p," *Technometrics*, Vol.15, No.4, pp.661–675 (1973).

[23] A. Miller, *Subset Selection in Regression, 2nd Edition* (Chapman and Hall/CRC, 2002).

[24] R. Miyashiro and Y. Takano, "Mixed Integer Second-Order Cone Programming Formulations for Variable Selection," Technical Report, Department of Industrial Engineering and Management, Tokyo Institute of Technology (2013).

[25] G. Schwarz, "Estimating the Dimension of a Model," *Annals of Statistics*, Vol.6, No.2, pp.461–464 (1978).

[26] 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).

[27] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, URL `http://www.R-project.org` (2013).

[28] 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, pp.13–26 (1978).

[29] E.G. Talbi, "A Taxonomy of Hybrid Metaheuristics," *Journal of Heuristics*, Vol.8, No.5, pp.541–564 (2002).

[30] H. Theil, *Economic Forecasts and Policy* (North-Holland Publishing Company, 1961).

[31] R.J. Wherry, "A New Formula for Predicting the Shrinkage of the Coefficient of Multiple Correlation," *The Annals of Mathematical Statistics*, Vol.2, No.4, pp.440–457 (1931).

[32] H.P. Williams, *Model Building in Mathematical Programming, 5th edition* (Wiley, 2013).