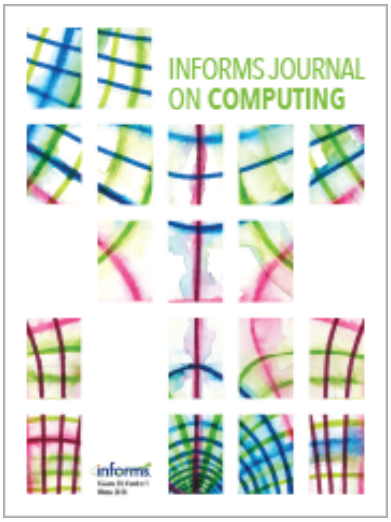


INFORMS Journal on Computing

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

A Mixed-Integer Fractional Optimization Approach to Best Subset Selection

Andrés Gómez, Oleg A. Prokopyev

To cite this article:
Andrés Gómez, Oleg A. Prokopyev (2021) A Mixed-Integer Fractional Optimization Approach to Best Subset Selection. INFORMS Journal on Computing 33(2):551-565. https://doi.org/10.1287/ijoc.2020.1031

http://pubsonline.informs.org/journal/ijoc

INFORMS JOURNAL ON COMPUTING
Vol. 33, No. 2, Spring 2021, pp. 551–565
ISSN 1091-9856 (print), ISSN 1526-5528 (online)

A Mixed-Integer Fractional Optimization Approach to Best Subset Selection

Andrés Gómez,[a] **Oleg A. Prokopyev**[b]

[a] Department of Industrial and Systems Engineering, Viterbi School of Engineering, University of Southern California, Los Angeles, California 90089; [b] Department of Industrial Engineering, Swanson School of Engineering, University of Pittsburgh, Pittsburgh, Pennsylvania 15261

Contact: gomezand@usc.edu, https://orcid.org/0000-0003-3668-0653 (AG); droleg@pitt.edu, https://orcid.org/0000-0003-2888-8630 (OAP)

Received: July 1, 2019
Revised: March 30, 2020; August 9, 2020; August 12, 2020; September 3, 2020
Accepted: September 18, 2020
Published Online in Articles in Advance: March 12, 2021

https://doi.org/10.1287/ijoc.2020.1031

Abstract. We consider the best subset selection problem in linear regression—that is, finding a parsimonious subset of the regression variables that provides the best fit to the data according to some predefined criterion. We are primarily concerned with alternatives to cross-validation methods that do not require data partitioning and involve a range of information criteria extensively studied in the statistical literature. We show that the problem of interest can be modeled using fractional mixed-integer optimization, which can be tackled by leveraging recent advances in modern optimization solvers. The proposed algorithms involve solving a sequence of mixed-integer quadratic optimization problems (or their convexifications) and can be implemented with off-the-shelf solvers. We report encouraging results in our computational experiments, with respect to both the optimization and statistical performance.

Summary of Contribution: This paper considers feature selection problems with information criteria. We show that by adopting a fractional optimization perspective (a well-known field in nonlinear optimization and operations research), it is possible to leverage recent advances in mixed-integer quadratic optimization technology to tackle traditional statistical problems long considered intractable. We present extensive computational experiments, with both synthetic and real data, illustrating that the new fractional optimization approach is orders of magnitude faster than existing approaches in the literature.

History: Accepted by Ram Ramesh, Area Editor for Knowledge Management and Machine Learning.
Funding: This paper is based on work supported by the Division of Mathematical Sciences of the National Science Foundation [Grant 1818700].
Supplemental Material: The online supplement and code are available at https://doi.org/10.1287/ijoc.2020.1031.

Keywords: fractional optimization • conic optimization • sparse regression • information criteria

1. Introduction

We consider the linear regression model (Seber and Lee 2003, Weisberg 2005) in which, given a *design matrix* $\boldsymbol{X} = [\boldsymbol{x_1}, \ldots, \boldsymbol{x_p}] \in \mathbb{R}^{n\times p}$ of *explanatory* (*independent*) *variables* and a vector $\boldsymbol{y} \in \mathbb{R}^n$ of *response* (*dependent*) *variables*, the relationship between them is

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ is a vector *regression coefficients* and $\boldsymbol{\epsilon} \in \mathbb{R}^n$ are the *error* terms; throughout the paper, we assume $n > p$. The linear regression approach involves finding appropriate values for parameters $\boldsymbol{\beta}$ such that the data fitting error is minimized according to some predefined criteria. The ordinary least squares estimate, found by minimizing the residual squared error, is easy to compute but suffers from poor *prediction accuracy* and *interpretability*. Model *overfitting* is one of the key challenges, which naturally leads to the problem of finding a parsimonious *best subset* of explanatory variables. By removing unnecessary or noise variables and keeping only the most important and critical ones, we obtain more interpretable and robust regression models. This *subset selection* problem has attracted significant attention in the statistical, machine learning, and optimization literatures. A classical model for the subset selection problem (Miller 2002) is

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\boldsymbol{y} - \boldsymbol{X\beta}\|_2^2 \text{ subject to} \|\boldsymbol{\beta}\|_0 \le k, \tag{2}$$

where k is some predefined sparsity parameter and $\|\cdot\|_0$ is the ℓ_0-norm (i.e., $\|\boldsymbol{\beta}\|_0 = \sum_{i=1}^p \mathbb{1}_{\{\beta_i\neq 0\}}$, with $\mathbb{1}_{(\cdot)}$ denoting the indicator function). Problem (2) is strongly *NP*-hard (Chen et al. 2019), and several approaches to tackle it approximately or exactly have been proposed in the literature.

Perhaps the most widely known approximation approach is Lasso (Tibshirani 1996), where the ℓ_0-norm is replaced by the convex ℓ_1-norm. The resulting convex problem can be solved very efficiently (Efron et al. 2004). Lasso has some desirable theoretical

properties under appropriate conditions on data (Zhang and Huang 2008, Wainwright 2009, Bühlmann and van de Geer 2011, Tibshirani 2011) and is widely used for finding sparse models in practice. However, Lasso is only a surrogate and may potentially lead to low-quality solutions; we refer the reader to the detailed discussion in Bertsimas et al. (2016) and the references therein.

Alternatively, globally optimal solutions for (2) can be sought. Earlier approaches, including exhaustive enumerations of all subsets (Garside 1965, 1971a, b) and the leaps and bounds procedure (Furnival and Wilson 1974), do not scale well for large instances. Nevertheless, recent approaches based on mixed-integer optimization (MIO) have proven more effective at solving problem (2), see Bertsimas and Shioda (2009), Bertsimas and King (2015), Bertsimas et al. (2016), Bertsimas and Van Parys (2020), Cozad et al. (2015), Miyashiro and Takano (2015b), and Wilson and Sahinidis (2017). Specifically, by introducing binary variables $\boldsymbol{z} \in \{0,1\}^p$ such that $z_i = 1$ if $\beta_i \neq 0$, problem (2) can be formulated as

$$\min \|\boldsymbol{y} - \boldsymbol{X\beta}\|_2^2 \tag{3a}$$

$$\text{s.t. } \mathbf{1}'\boldsymbol{z} \leq k, \tag{3b}$$

$$-M\boldsymbol{z} \leq \boldsymbol{\beta} \leq M\boldsymbol{z}, \tag{3c}$$

$$\boldsymbol{z} \in \{0,1\}^p,\ \boldsymbol{\beta} \in \mathbb{R}^p, \tag{3d}$$

where $\mathbf{1}$ denotes a p-dimensional vector of all ones, and big-M constraints (3c) are used to link the indicator and regression coefficient variables (Glover 1975). Problem (3) is a mixed-integer quadratic optimization (MIQO) problem, which can be solved directly with off-the-shelf solvers for convex MIO.

Note that (3) requires specifying a priori the desired sparsity k at the right-hand side of (3b). The standard technique for determining k is based on using cross validation, which considers (3) for multiple values of k and then selects the one that performs best in a held-out validation set. A naive approach would be to simply solve (3) for all possible values of k. Clearly, it is prohibitively expensive in many settings. Thus, various ideas have been explored in the literature to avoid such enumeration. For example, Kenney et al. (2018) propose warm-starting and novel bisection schemes to reduce the burden of solving multiple MIO problems; other warm-start-like and related ideas are explored by Bertsimas et al. (2016, 2019a). Nevertheless, the approach based on (3) and cross validation may remain relatively expensive. Hence, the primary goal of this study is to explore alternatives to cross-validation methods that can be performed effectively and do not require partitioning the data.

1.1. Criteria

Several criteria have been proposed in the statistics literature to evaluate the quality of a given regression model. The measures involve a trade-off between the residual squared error $\|\boldsymbol{y} - \boldsymbol{X\beta}\|_2^2$ and the size of the model $\|\boldsymbol{\beta}\|_0$. We present a brief description of the information criteria used in this paper and refer the reader to Konishi and Kitagawa (2008) for an in-depth treatment of the statistical merits of the information criteria used (and others). To simplify the discussion, we assume in this section that the noise $\boldsymbol{\epsilon}$ is independent and identically distributed (i.i.d.) Gaussian with unknown variance σ^2, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\boldsymbol{I})$.

1.1.1. Mean Squared Error. The mean squared error (MSE; Wherry 1931) of a regression model is given by

$$\frac{\|\boldsymbol{y} - \boldsymbol{X\beta}\|_2^2}{n - \|\boldsymbol{\beta}\|_0}. \tag{4}$$

Minimizing the MSE is equivalent to maximizing the adjusted R^2 and is one of the most widely used criteria to compare regression models because of its simplicity. Note that $\frac{\|\boldsymbol{y}-\boldsymbol{X\beta}\|_2^2}{n-p}$ is precisely the least squares estimator of the variance σ^2 for (1) with p regressors; thus, intuitively, optimization with respect to the MSE criterion selects the model that "promises" the lowest noise variance.

1.1.2. Akaike Information Criterion and Corrected Akaike Information Criterion. The Akaike information criterion (AIC; Akaike 1974) is

$$n \ln\left(\frac{\|\boldsymbol{y} - \boldsymbol{X\beta}\|_2^2}{n}\right) + 2\|\boldsymbol{\beta}\|_0 + K, \tag{5}$$

where K is a constant that does not depend on the model. Note that the term inside the logarithm in (5) is the maximum likelihood estimator of σ^2, but this estimator is biased. Akaike (1974) corrects this bias using the term $2\|\boldsymbol{\beta}\|_0$ and shows that if the data are indeed generator according to (1) for some sparse vector $\boldsymbol{\beta}$, then minimizing the AIC criterion yields estimates with minimal Kullback–Leibler divergence with respect to a true distribution. Sugiura (1978) note that AIC bias needs to be further corrected when n is close to p and propose the corrected AIC (or AICc):

$$n \ln\left(\frac{\|\boldsymbol{y} - \boldsymbol{X\beta}\|_2^2}{n}\right) + 2\|\boldsymbol{\beta}\|_0 + \frac{2\|\boldsymbol{\beta}\|_0^2 + 2\|\boldsymbol{\beta}\|_0}{n - \|\boldsymbol{\beta}\|_0 - 1}; \tag{6}$$

see also Hurvich and Tsai (1989).

1.1.3. Bayesian Information Criterion. The Bayesian information criterion (BIC; Schwarz 1978) is

$$n \ln\left(\frac{\|\boldsymbol{y} - \boldsymbol{X\beta}\|_2^2}{n}\right) + \ln(n)\|\boldsymbol{\beta}\|_0 + K. \tag{7}$$

Whereas AIC seeks to minimize the Kullback–Leibler divergence between the true and estimated models, BIC is obtained by maximizing the model that is a posteriori most probable, under the prior that all subsets of $\{1, \ldots, p\}$ are equally likely to be the model generating the data.

The criteria just outlined are widely used to compare linear regression models. Furthermore, they are also used as stopping rules for heuristics (2) such as forward selection or backward elimination (Miller 2002). However, currently few approaches exist to find the best model according to one of these criteria.

In particular, Park and Klabjan (2017) propose a mixed-integer quadratically constrained programming approach for optimization with respect to MSE. Kimura and Waki (2018) proposed a tailored branch-and-bound algorithm for minimization of the AIC criterion. Wilson and Sahinidis (2017) exploit the fact that if the variance of the error terms $\boldsymbol{\epsilon}$ is known, problems with AIC and BIC can be simplified to MIQO problems. Cozad et al. (2014) tackle subset selection problems with information criteria by solving problem (3) for different values of k and choosing the best one:

$$\min_{k \in \{0,\ldots,p\}} \{\min\{F(\boldsymbol{\beta}, k) : (3b) - (3d)\}\}, \tag{8}$$

where $F(\boldsymbol{\beta}, k)$ corresponds to one of the above-mentioned criteria given by (4)–(6).

Observe that for a fixed $k = \|\boldsymbol{\beta}\|_0$, finding the best model with respect to any criterion in (4)–(7) can be done by minimizing $\|\boldsymbol{y} - \boldsymbol{X\beta}\|_2^2$. Thus, approach (8) requires solving $p + 1$ different MIO problems and is, to the best of our knowledge, the most efficient method to date. Note that this approach can be improved by warm-starting each MIO problem with the solution found from the previous one, as pointed out by Bertsimas et al. (2016).

Miyashiro and Takano (2015a) propose using mixed-integer second-order conic optimization (MISOCO) for the best subset selection problem with information criteria. The best model can be found by solving a single MIO, but it requires the addition of $p + 1$ additional binary variables. The authors report that the MISOCO formulations perform worse than (8) by an order of magnitude. Finally, Takano and Miyashiro (2020) also propose the MIO approach for the best subset selection using the cross-validation criterion.

1.2. Contributions and Outline

In this paper we propose new MIO formulations and techniques for the best subset selection problem with information criteria. In particular, the problems considered are modeled as convex mixed-integer fractional optimization problems (MIFO). The formulations are stronger than the existing alternatives proposed in the literature; the proposed approach is faster than (8) by at least an order of magnitude in large instances and several orders of magnitude faster than previous MISOCO approaches. The algorithms proposed can be easily implemented using off-the-shelf mathematical optimization software, resulting in several advantages over customized methods: additional constraints can easily be incorporated into the formulations (e.g., see Bertsimas and King (2015) and Cozad et al. (2015)), and the proposed algorithms benefit from the continuous improvements to commercial software.

The remainder of the paper is organized as follows. In Section 2 we describe our MIFO approach and compare it against the existing modeling alternatives. In Section 3 we discuss how to solve the resulting MIFO by (partially) solving a sequence of MIQO problems. In Section 4 we provide computational experiments on synthetic and real data sets, and in Section 5 we conclude the paper and highlight directions for future research. Finally, we note that all proofs as well as some modeling and algorithmic details are relegated to the online supplement.

2. Formulations

In this section we give MIFO formulations for the subset selection problems with the information criteria discussed in Section 1. In particular, one of the main challenges for solving best subset selection (with respect to criteria other than the MSE) is handling the (nonconvex) logarithmic term in the objective function; see (5)–(7). In order to do so, we first show in Section 2.1 how to model the best subset selection problems as the (possibly nonconvex) MIFO problem:

$$\min_{\boldsymbol{z} \in \{0,1\}^p, \boldsymbol{\beta} \in \mathbb{R}^p} \frac{\|\boldsymbol{y} - \boldsymbol{X\beta}\|_2^2}{g(\mathbf{1}'\boldsymbol{z})} \quad \text{subject to } -M\boldsymbol{z} \le \boldsymbol{\beta} \le M\boldsymbol{z}, \tag{9}$$

where $g : \mathbb{R}_+ \to \mathbb{R}_+$ is a one-dimensional nonincreasing convex function that depends on the criterion used. Then in Section 2.2 we discuss how to obtain mixed-integer convex formulations of (9) by exploiting submodularity. Finally, in Section 2.3, we show that the resulting formulations are at least as strong as the alternative formulations proposed in the literature.

2.1. Fractional Formulations

We now discuss a MIFO framework that is able to handle most feature selection problems with information criteria.

2.1.1. MSE Criterion. Observe that optimization with respect to the MSE criterion can be directly formulated as

$$\min_{z\in\{0,1\}^p,\boldsymbol{\beta}\in\mathbb{R}^p} \frac{\|\boldsymbol{y}-\boldsymbol{X\beta}\|_2^2}{n-\mathbf{1}'\boldsymbol{z}} \quad \text{subject to } -M\boldsymbol{z}\le\boldsymbol{\beta}\le M\boldsymbol{z}.$$

That is, function $g(x)=n-x$ is affine.

2.1.2. AIC and BIC Criteria. Consider optimization with respect to either AIC or BIC, given by (5) or (7), respectively. The best model with respect to such criteria can be found by solving

$$\begin{aligned}\min_{z\in\{0,1\}^p,\boldsymbol{\beta}\in\mathbb{R}^p}\ & \ln\left(\frac{\|\boldsymbol{y}-\boldsymbol{X\beta}\|_2^2}{n}\right)+\alpha\mathbf{1}'\boldsymbol{z}\\ \text{subject to}\ & -M\boldsymbol{z}\le\boldsymbol{\beta}\le M\boldsymbol{z},\end{aligned}\tag{10}$$

where the constant terms in the definition of the criterion is dropped, and α is a constant that may depend on n; that is, $\alpha=2/n$ for AIC and $\alpha=\ln(n)/n$ for BIC. Because the exponential function is nondecreasing and monotone, we can take the exponential of the objective function and find that (10) is equivalent to

$$\begin{aligned}&\frac{1}{n}\cdot\min_{z\in\{0,1\}^p,\boldsymbol{\beta}\in\mathbb{R}^p}\ \|\boldsymbol{y}-\boldsymbol{X\beta}\|_2^2\cdot e^{\alpha\mathbf{1}'\boldsymbol{z}}\\ &\quad\text{subject to}\ -M\boldsymbol{z}\le\boldsymbol{\beta}\le M\boldsymbol{z}\\ &\quad=\frac{1}{n}\cdot\min_{z\in\{0,1\}^p,\boldsymbol{\beta}\in\mathbb{R}^p}\ \frac{\|\boldsymbol{y}-\boldsymbol{X\beta}\|_2^2}{e^{-\alpha\mathbf{1}'\boldsymbol{z}}}\\ &\qquad\text{subject to}\ -M\boldsymbol{z}\le\boldsymbol{\beta}\le M\boldsymbol{z}.\end{aligned}$$

From these derivations, we see that (10) is a special case of (9), where $g(x)=e^{-\alpha x}$.

2.1.3. AICc Criterion. A similar approach can be used for optimization with respect to AICc given by (6), resulting in

$$\min\ n\ln\left(\frac{\|\boldsymbol{y}-\boldsymbol{X\beta}\|_2^2}{n}\right)+2(\mathbf{1}'\boldsymbol{z})+\frac{2(\mathbf{1}'\boldsymbol{z})^2+2(\mathbf{1}'\boldsymbol{z})}{n-\mathbf{1}'\boldsymbol{z}-1}\tag{11a}$$

$$\text{subject to } -M\boldsymbol{z}\le\boldsymbol{\beta}\le M\boldsymbol{z},\ \boldsymbol{z}\in\{0,1\}^p,\ \boldsymbol{\beta}\in\mathbb{R}^p.\tag{11b}$$

After dividing by n, taking the exponential of the objective function and some algebraic manipulations (please see details in the online supplement), problem (11) can be equivalently written as

$$\frac{1}{ne^2}\min_{z\in\{0,1\}^p,\boldsymbol{\beta}\in\mathbb{R}^p}\ \frac{\|\boldsymbol{y}-\boldsymbol{X\beta}\|_2^2}{e^{-2\frac{n-1}{n-1'z-1}}}\quad\text{subject to } -M\boldsymbol{z}\le\boldsymbol{\beta}\le M\boldsymbol{z}.\tag{12}$$

Therefore, we see that (11) is a special case of (9), where $g(x)=e^{-2\frac{n-1}{n-x-1}}$.

2.2. Convexification

Consider the mixed-integer set

$$\mathcal{F}=\left\{\boldsymbol{z}\in\{0,1\}^p,\ s\in\mathbb{R}_+ : s\le g(\mathbf{1}'\boldsymbol{z})\right\}.\tag{13}$$

Because g is convex, the function $g(\mathbf{1}'\boldsymbol{z})$ is supermodular. Define $\pi_i=g(i)-g(i-1)$, $i=1,\dots,p$, and given a permutation $((1),(2),\dots,(p))$ of $[p]$, consider the inequality

$$s\le g(0)+\sum_{i=1}^{p}\pi_i z_{(i)}.\tag{14}$$

The coefficients $-\boldsymbol{\pi}$ in (14) correspond to an extreme point of the *extended polymatroid* associated with the submodular function $-g$, and (14) is referred to as an *extended polymatroid inequality* (Atamtürk and Narayanan 2008). Additionally, extended polymatroid inequalities and bound constraints are sufficient to describe the convex hull of $\mathcal{F}$ (Lovász 1983); that is,

$$\begin{aligned}\operatorname{conv}(\mathcal{F})=\Bigg\{&(\boldsymbol{z},s)\in[0,1]^p\times\mathbb{R}_+ : s\le g(0)\\ &+\sum_{i=1}^{p}\pi_i z_{(i)},\ \text{for all permutations of } [p]\Bigg\}.\end{aligned}$$

Thus, we can formulate (9) as the convex MIFO problem

$$\min\ \frac{\|\boldsymbol{y}-\boldsymbol{X\beta}\|_2^2}{s}\tag{15a}$$

$$\text{s.t. } s\le g(0)+\sum_{i=1}^{p}\pi_i z_{(i)},\quad\text{for all permutations of } [p],\tag{15b}$$

$$-M\boldsymbol{z}\le\boldsymbol{\beta}\le M\boldsymbol{z},\tag{15c}$$

$$\boldsymbol{z}\in\{0,1\}^p,\ \boldsymbol{\beta}\in\mathbb{R}^p,\ s\ge 0.\tag{15d}$$

Note that there is a factorial number of constraints (15b). Therefore, to implement formulations (15) in practice, a lazy constraint generation scheme for (15b) should be used, which is a standard feature of modern off-the-shelf solvers. In particular, finding which

inequality (15b) to add at a particular point $(\bar{z}, \bar{\beta}, \bar{s})$ (if any) can be done using a greedy algorithm (Edmonds 1970), as formalized in Proposition 1.

Proposition 1 (Edmonds 1970)**.** *A most violated inequality* (15b) *at* $(\bar{z}, \bar{\beta}, \bar{s})$ *is precisely* $s \le g(0) + \sum_{i=1}^{p} \pi_{(i)} z_{(i)}$ *for the permutation where variables are ordered in nonincreasing order,* $\bar{z}_{(1)} \ge \bar{z}_{(2)} \ge \cdots \ge \bar{z}_{(p)}$.

Remark 1 (MSE Criterion)**.** If $g(x) = n - x$, corresponding to the MSE criterion, then each inequality (14) reduces to $s \le n - \mathbf{1}'z$. This inequality can be changed to an equality constraint without loss of generality. Hence, in such case, (15) reduces simply to the convex MIFO:

$$\min_{z \in \{0,1\}^p, \beta \in \mathbb{R}^p} \frac{\|y - X\beta\|_2^2}{n - \mathbf{1}'z} \text{ subject to } -Mz \le \beta \le Mz. \quad (16)$$

2.3. Comparison with Existing Results

In this section we compare formulation (15) with other MIO formulations for optimization with respect to information criteria.

2.3.1. Linearization for MSE Criterion. Park and Klabjan (2017) propose a MIO formulation for optimization with respect to the MSE criterion. They formulate problem (16) as

$$\min_{z,\beta,t} \ t \quad (17a)$$

$$\text{s.t. } \|y - X\beta\|_2^2 \le t\left(n - \sum_{i=1}^{p} z_i\right), \quad (17b)$$

$$-Mz \le \beta \le Mz, \quad (17c)$$

$$z \in \{0,1\}^p, \beta \in \mathbb{R}^p, t \in \mathbb{R}_+. \quad (17d)$$

Then, in order to model the nonlinear constraint (17b), the authors linearize the bilinear terms. Specifically, by introducing additional variables v_i, they replace (17b) with the system

$$\|y - X\beta\|_2^2 \le tn - \sum_{i=1}^{p} v_i, \quad (18a)$$

$$0 \le v_i \le t, \ t - M(1 - z_i) \le v_i \le Mz_i, \quad \forall i = 1, \ldots, p, \quad (18b)$$

where M is sufficiently large. Because each bilinear term tz_i is replaced by its convex envelope, the system (18a) and (18b) is weaker than (17b). Also, for the MSE criterion, (15) is equivalent to (17) in terms of its continuous relaxation strength. Thus, (15) is stronger than the formulations induced by (18a) and (18b), and it avoids the inclusion of additional big-M constraints.

2.3.2. MISOCO Formulations. Miyashiro and Takano (2015a) propose to tackle subset selection problems with information criteria using `MISOCO` formulations, discussed next. As we show in this section, the relaxations induced by our approach are stronger than the existing MISOCO formulations for criteria other than MSE (and is equivalent for MSE).

MSE Criterion. Constraint (17b) is a rotated cone constraint, and problem (17) can be directly formulated as a MISOCO. Thus, the strength of the convex relaxation of (16) is the same as that of the MISOCO formulation (17) used in Miyashiro and Takano (2015a).

General Criteria. For tackling (9), Miyashiro and Takano (2015a) propose to use

$$\min_{z,\beta,w,s,t} \ t \quad (19a)$$

$$\text{s.t. } \|y - X\beta\|_2^2 \le ts, \quad (19b)$$

$$s \le \sum_{i=0}^{p} g(i) w_i, \quad (19c)$$

$$\sum_{i=0}^{p} i w_i = \mathbf{1}'z, \quad (19d)$$

$$\mathbf{1}'w = 1, \quad (19e)$$

$$-Mz \le \beta \le Mz, \quad (19f)$$

$$z \in \{0,1\}^p, w \in \{0,1\}^{p+1}, \beta \in \mathbb{R}^p, s \ge 0, t \ge 0 \quad (19g)$$

(i.e., using special ordered sets of type 1 (SOS 1) with the introduction of additional variables w). We have the following theoretical observation.

Proposition 2. *Formulation* (15) *has a stronger convex relaxation than* (19).

Our result implies that the formulations proposed in this paper, which do not require the introduction of additional binary variables, result in a stronger convex relaxation than the MISOCO formulation.

Finally, we want to point out that current technology for solving MISOCO is lagging far behind MIQO technology. Specifically, convexifications for MIQO sparse regression problems have been extensively studied in the literature (Günlük and Linderoth 2010, Jeon et al. 2017, Atamtürk and Gómez 2018; Atamtürk et al. 2018; Wei et al. 2020a, b; Xie and Deng 2020), whereas there are relatively few results concerning the corresponding MISOCO structures (Atamtürk and Jeon 2019, Gómez 2020). Using parametric approaches for fractional optimization, discussed in Section 3, our method fully leverages the advanced technology for MIQO problems and far outperforms the MISOCO formulations even in the case of the MSE criterion.

3. Parametric MIQO Approaches

Formulations (15) and (19) can be tackled with convex MIO solvers such as Bonmin (Bonami et al. 2008) and FilMINT (Abhishek et al. 2010); see also the work of Mahajan et al. (2017) and the references therein for further details. However, a MIQO problem such as (3) admits specialized and better solution approaches. Specifically, the convex subproblems arising in MIQO can be solved with the simplex method, which is amenable to warm starts and is a better choice for branch-and-bound algorithms. As a consequence, current codes for MIQO are more efficient than the corresponding codes for convex MIO. To leverage the superior performance of solvers for MIQO, recent works have proposed to tackle MISOCO with a polyhedral feasible region by solving a sequence of MIQO problems (Atamtürk and Gómez 2019b, Atamtürk et al. 2020), and they report significant speedups in solution times. By exploiting the fractional structure of problem (15), similar approaches can be used in our context.

Consider the MIQO problems parameterized by t:

$$(\texttt{MIQO}_t) \quad d(t) = \min \ \|\boldsymbol{y}-\boldsymbol{X\beta}\|_2^2 - ts \tag{20a}$$

$$\text{s.t. (15b)–(15d)}, \tag{20b}$$

and recall that $s = g(\mathbf{1}'\boldsymbol{z})$ in any optimal solution. A classical result from the fractional optimization literature (see, e.g., Radzik (1998)) is that if $d(t^*) = 0$, then t^* is the optimal objective function value of (9). Hence, problem (9) reduces to finding a root for the function $d(t)$, for example, via bisection or Newton-like methods (Dinkelbach 1967, Megiddo 1979, Radzik 1998, Borrero et al. 2017).

Given a parameter $\xi > 0$, let $\texttt{solve}_\xi$ be a routine that either returns a feasible solution $(\hat{\boldsymbol{\beta}}(t), \hat{\boldsymbol{z}}(t), \hat{d}(t))$ of $\texttt{MIQO}_t$ with corresponding objective function value $\hat{d}(t)$ less than $-\xi$,

$$\hat{d}(t) = \|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}(t)\|_2^2 - tg(\mathbf{1}'\hat{\boldsymbol{z}}(t)) < -\xi,$$

or proves that $d(t) \geq -\xi$. For example, $\texttt{solve}_\xi$ can be naturally implemented using branch-and-bound solvers for MIQO by either solving (20) to optimality and checking whether $d(t) < -\xi$, or stopping the algorithm when an incumbent solution with a value less than $-\xi$ is found or when a tight lower bound is proven.

Define the function $h : \{0,1\}^p \times \mathbb{R}^p \to \mathbb{R}_+$ as

$$h(\bar{\boldsymbol{\beta}}, \bar{\boldsymbol{z}}) = \frac{\|\boldsymbol{y} - \boldsymbol{X}\bar{\boldsymbol{\beta}}\|_2^2}{g(\mathbf{1}'\bar{\boldsymbol{z}})}.$$

Furthermore, let $(\boldsymbol{\beta}^*, \boldsymbol{z}^*)$ be an optimal solution for (9) and define for any feasible solution $(\bar{\boldsymbol{\beta}}, \bar{\boldsymbol{z}})$ the relative optimality gap as

$$\texttt{gap} = \frac{h(\bar{\boldsymbol{\beta}}, \bar{\boldsymbol{z}}) - h(\boldsymbol{\beta}^*, \boldsymbol{z}^*)}{h(\bar{\boldsymbol{\beta}}, \bar{\boldsymbol{z}})}. \tag{21}$$

Next, we consider the Newton method approach given in Algorithm 1.

Algorithm 1 (Newton method for (9))

Input: $\boldsymbol{y}$ response vector; $\boldsymbol{X}$ model matrix; ϵ precision parameter.

Output: $\boldsymbol{\beta}$, regression coefficients; $\boldsymbol{z}$, selected features.

1: Compute initial bounds
2: $(\bar{\boldsymbol{\beta}}, \bar{\boldsymbol{z}}) \leftarrow$ any feasible solution ▷ for example, $\bar{\boldsymbol{\beta}} = \bar{\boldsymbol{z}} = \mathbf{0}$
3: $t \leftarrow h(\bar{\boldsymbol{\beta}}, \bar{\boldsymbol{z}})$
4: **while** time limit not exceeded **do**
5: $\quad \xi \leftarrow \epsilon t g(p)$ ▷ Precision for subproblem
6: $\quad (\hat{\boldsymbol{\beta}}(t), \hat{\boldsymbol{z}}(t), \hat{d}(t)) \leftarrow \texttt{solve}_\xi$
7: $\quad$ **If** $\hat{d}(t) < -\xi$ **then**
8: $\quad\quad (\bar{\boldsymbol{\beta}}, \bar{\boldsymbol{z}}) \leftarrow (\hat{\boldsymbol{\beta}}(t), \hat{\boldsymbol{z}}(t))$
9: $\quad\quad t \leftarrow h(\bar{\boldsymbol{\beta}}, \bar{\boldsymbol{z}})$
10: $\quad$ **else if** $\hat{d}(t) \geq -\xi$ **then**
11: $\quad\quad$ **return** $(\bar{\boldsymbol{\beta}}, \bar{\boldsymbol{z}})$ ▷ Optimal solution found
12: $\quad$ **end if**
13: **end while**
14: **return** $(\bar{\boldsymbol{\beta}}, \bar{\boldsymbol{z}})$ ▷ Best solution found within the time limit

Proposition 3. *If the time limit is not reached, then Algorithm 1 terminates with a feasible solution with gap $\leq \epsilon$.*

The result of Proposition 3 holds independently of the quality of the feasible solutions found in line 6 of the algorithm. However, as Proposition 4 shows, high-quality solutions may lead to substantially fewer iterations.

Proposition 4. *If all problems $MIQO_t$ in line 6 are solved to optimality, then Algorithm 1 finds an optimal solution in at most $p + 1$ iterations.*

The proof of Proposition 4 follows standard arguments in fractional combinatorial optimization literature, see similar results in Radzik (1998). More important, Proposition 4 provides some intuition on why Algorithm 1 performs better than using (8): in the worst case, both approaches involve solving $p + 1$ MIQO, but in practice, Algorithm 1 requires significantly fewer iterations. Furthermore, in our computations discussed next, we found out that stopping the optimization of $\texttt{MIQO}_t$ whenever a feasible solution with objective value less than $-\xi$ is found, results in a better performance. Indeed, it is well known that algorithms for MIO find high quality and even optimal solutions in a fraction of the time required to prove optimality. Thus, if problems (20) are solved partially, then in practice all iterations except the last one or two are solved in seconds or milliseconds with few branch-and-bound nodes. Even if such an approach requires more iterations (in our computations the number of iterations is still bounded by $p + 1$), the overall solution times are reduced significantly.

3.1. Exploiting Conic Relaxations for Sparse Regression

There has been a recent research thrust toward designing strong convex relaxations of problem (3), and either using them as stand-alone methods to obtain estimators for sparse regression (Dong et al. 2015, Pilanci et al. 2015, Atamtürk and Gómez 2019a) or embedding them into branch-and-bound methods (Bertsimas et al. 2019b, Bertsimas and Van Parys 2020). It is possible to modify Algorithm 1 to solve at each iteration any such relaxation (instead of a MIO), thus solving a strong relaxation of the fractional problem (9)—the resulting estimator can then be used directly as a proxy of the optimal estimator with respect to a given information criteria.

In this paper we implemented this approach using the convex relaxation by Atamtürk and Gómez (2019), which is the strongest of the three mentioned and the only one that does not require an additional ridge regularization term $\|\boldsymbol{\beta}\|_2^2$. Note that solving to optimality each problem in Algorithm 1 requires solving an SDP with lazy constraints. As SDPs are solved in current off-the-shelf solvers via interior point methods and lack warm-start capabilities, a naïve implementation of this lazy constraint method may be tantamount to solving several SDPs from scratch, and it may be prohibitively expensive. To address this issue, we modify Algorithm 1 to integrate the cut generation and Newton method, reducing the number of SDPs to be solved. The details of the convex relaxation used and the modified Newton method are given in the online supplement.

4. Computations

In this section we report computational experiments performed on synthetic and real data sets to test the proposed approaches for the best subset selection problems with respect to MSE, BIC, and AICc criteria. We set the following specifications:

- Computations were performed using CPLEX 12.7.1 (for MIO) and MOSEK 8.1.0 (for conic relaxations) on a computer with a 3.50 GHz Intel® Xeon® E5-1620 v4 CPU and 16 GB main memory and with a single thread.
- All solver parameters were set to their default values.
- The code used in the implementations is available in the online supplement of the paper.

4.1. Instances

We now describe the instances used in our experiments.

4.1.1. Synthetic Instances. We generate synthetic data sets as done in Bertsimas et al. (2016) and Hastie et al. (2017). Given dimensions n and p, a sparsity parameter $k_0 \in \mathbb{Z}_+$, an autocorrelation parameter ρ, and a signal-to-noise parameter ν, the instances are generated as follows:

(i) The "true" regression coefficients $\boldsymbol{\beta}^0$ have their first k_0 components equal to 1 and the remaining equal to 0.

(ii) Each row of the design matrix $\boldsymbol{X}$ is generated i.i.d. from a multivariate normal distribution $\mathcal{N}_p(\mathbf{0}, \Sigma)$, where $\Sigma \in \mathbb{R}^{p\times p}$ satisfies $\Sigma_{ij} = \rho^{|i-j|}$.

(iii) The response variable $\boldsymbol{y}$ is generated from a normal distribution $\mathcal{N}_n(\boldsymbol{X\beta^0}, \sigma^2\boldsymbol{I})$, where $\sigma^2 = (\boldsymbol{\beta^0}\Sigma\boldsymbol{\beta^0})/\nu$ is defined to meet the desired SNR level.

4.1.2. Real Instances. We test the proposed methods on the Diabetes data set used in Efron et al. (2004) and later in Bertsimas et al. (2016). We also use the data sets used in Miyashiro and Takano (2015a): the data sets Housing, AutoMPG, SolarFlare, BreastCancer, and Crime, as well as the Insurance data set; their sizes (n, p) are reported in the left column in Table 1. All data sets except for Diabetes are available from the UCI Machine Learning Repository (Dheeru and Karra Taniskidou 2017).

4.2. Optimization Performance

We first focus on the performance of the methods from an optimization perspective (i.e., their solution times and end gaps). We point out that although there exist techniques for sparse regression that can solve *to optimality* problems with thousands of variables (Bertsimas et al. 2019b, Atamtürk and Gómez 2020, Bertsimas and Van Parys 2020), those methods involve and exploit additional regularization terms. By contrast, the regression problems with respect to information criteria call for solving the "core" best subset selection problem with no additional regularization, where such techniques cannot be applied or do not perform well.

4.2.1. Methods. We compare the following methods for tackling the feature selection problems with information criteria:

`Misoco`: The MISOCO formulation (as in Miyashiro and Takano (2015a))

$$\min\ t \tag{22a}$$

$$\text{s.t. } \boldsymbol{\gamma} = \boldsymbol{y} - \boldsymbol{X\beta},\ \boldsymbol{\gamma}'\boldsymbol{\gamma} \le ts,\ (19c)\text{–}(19g),\ \boldsymbol{\gamma} \in \mathbb{R}^n \tag{22b}$$

`Fractional`: The fractional optimization approach with Algorithm 1

`Cardinality`: The approach described in (8), where the MIQO (3) is solved for all values of $k = 1, \ldots, p$; solutions obtained from solving the method with cardinality k are used to warm-start the solvers when solving the problem with cardinality $k + 1$

Table 1. Performance for MIO Methods in Real Data Sets

Instance	Method	MSE			BIC			AICc		
		Time	Gap (%)	Nodes	Time	Gap (%)	Nodes	Time	Gap (%)	Nodes
Housing	Misoco	†	†	†	13.4	—	416	4.1	—	151
$n = 506$, $p = 13$	Fractional	0.2	—	31	0.3	—	160	0.2	—	34
	Cardinality	1.2	—	304	1.2	—	304	1.2	—	304
AutoMPG	Misoco	†	†	†	612.3	—	39,387	43.9	—	6,336
$n = 392$, $p = 25$	Fractional	1.3	—	4,999	5.1	—	27,708	1.9	—	8,121
	Cardinality	10.3	—	50,562	10.3	—	50,562	10.3	—	50,562
SolarFlareC	Misoco	155.0	—	4,502	177.2	—	15,476	1,712	—	359,532
$n = 1{,}066$, $p = 26$	Fractional	0.6	—	824	2.7	—	9,704	1.3	—	3,576
	Cardinality	11.2	—	21,396	5.1	—	21,396	5.1	—	21,396
SolarFlareM	Misoco	†	†	†	65.8	—	11,270	366.6	—	148,084
$n = 1{,}066$, $p = 26$	Fractional	0.3	—	54	2.2	—	8,265	1.6	—	5,239
	Cardinality	23.1	—	108,409	23.1	—	108,409	23.1	—	108,405
SolarFlareX	Misoco	†	†	†	2.2	—	178	19.5	—	10,730
$n = 1{,}066$, $p = 26$	Fractional	0.2	—	20	0.5	—	554	0.4	—	599
	Cardinality	9.6	—	22,472	9.6	—	22,472	9.6	—	22,472
BreastCancer	Misoco	Limit	5.3	2.9×10^6	Limit	6.0	489,650	Limit	7.3	785,634
$n = 196$, $p = 37$	Fractional	119.9	—	648,348	825.0	—	3.6×10^6	860.4	—	4.7×10^6
	Cardinality	515.9	—	3.0×10^6	515.9	—	3.0×10^6	515.9	—	3.0×10^6
Diabetes	Misoco	Limit	22.2	52,651	Limit	41.0	117,190	Limit	8.9	155,747
$n = 442$, $p = 64$	Fractional	Limit	4.3	1.4×10^7	Limit	16.1	1.3×10^7	Limit	7.9	$1{,}3\times10^7$
	Cardinality	Limit	6.0	8.9×10^6	Limit	6.0	8.9×10^6	Limit	6.0	8.9×10^6
Crime	Misoco	Limit	100.0	4,201	Limit	41.3	22,283	Limit	7.2	24,544
$n = 1{,}993$, $p = 100$	Fractional	Limit	3.1	6.3×10^6	Limit	13.4	6.0×10^6	Limit	5.1	5.7×10^6
	Cardinality	Limit	11.8	4.8×10^6	Limit	11.8	4.8×10^6	Limit	11.8	4.8×10^6
Insurance	Misoco	†	†	†	Limit	100.0	2,871	Limit	100.0	1,100
$n = 5{,}822$, $p = 151$	Fractional	Limit	2.1	3.1×10^6	Limit	4.2	2.6×10^6	Limit	2.5	3.2×10^6
	Cardinality	Limit	3.1	2.7×10^6	Limit	3.1	2.7×10^6	Limit	3.1	2.7×10^6

Note. In instances not solved to optimality, the best solution found by `Cardinality` with respect to a given criterion matches the solution found by `Fractional`.

†Numerical errors occurred during branch and bound.

In addition, we also test methods $\texttt{Fractional}_{\texttt{SDP}}$, corresponding to the conic relaxations described in the online supplement, and $\texttt{Cardinality}_{\texttt{SDP}}$, which solves the conic relaxation propose in Atamtürk and Gómez (2019a) for all cardinalities.

Finally, for MIO formulations, we use the logical constraints $z_i = 0 \Rightarrow \beta_i = 0$ in CPLEX to impose constraints (3c), which essentially delegates to the solver the task of computing adequate big-M values—note that the conic relaxations $\texttt{Fractional}_{\texttt{SDP}}$ and $\texttt{Cardinality}_{\texttt{SDP}}$ do not use big-M values.

4.2.2. Time Limits. When solving the conic relaxations $\texttt{Fractional}_{\texttt{SDP}}$ and $\texttt{Cardinality}_{\texttt{SDP}}$, each problem is solved to optimality, and there is no time limit. For the MIO-based `Fractional` and `Cardinality` methods, we set a time limit of one hour. Note that for the `Cardinality` method, this is a time limit to solve *all* problems:[1] we initially allocate a time limit of (1/p) hours to each problem to each problem, and if a problem is solved before the time limit, then we allocate the unused time evenly among remaining problems.

4.2.3. Results. Table 1 reports for each instance, MIO method, and criterion the solution time (in seconds) required to solve problem (9) to optimality or the optimality gap proven when a time limit of one hour is reached, as well as the number of branch and bound nodes.

The optimality gaps are computed as follows: for methods `Fractional`, the optimality gap is given by Equation (27) in the online appendix; for `Misoco`, the optimality gap just corresponds to the gap reported by the solver; and for `Cardinality`, we report the worst optimality gap among all problems (8). Note that although the gaps of `Misoco` and `Fractional` correspond to the gap with respect to the optimal solution of problem (9), the gap of `Cardinality` has a different interpretation as it corresponds to the gap with respect to the optimal solution of a cardinality constrained problem (3), and thus it is not directly comparable with the other optimality gaps.

We see from Table 1 that the performance of `Misoco` is very poor, struggling in almost all instances; note that, by default, CPLEX uses linear outer approximations

to tackle MISOCO optimization problems, and poor quality of such approximations may be the cause of this bad performance. By contrast, the other MIO formulations, `Fractional` and `Cardinality`, perform well in the smaller data sets with $p \leq 40$, solving the problems to optimality in seconds or minutes. In addition, in all instances that are solved to optimality (except BreastCancer), `Fractional` is consistently an order of magnitude faster than `Cardinality` (in BreastCancer, the formulations are approximately equal, depending on the criterion used). However, in instances with $p \geq 64$, all MIO formulations are unable to prove optimality, and end gaps are in some cases above 10%. Solving these instances to optimality (by either method) would require substantially larger time limits.

In addition, Table 2 reports the time required to solve the problems for relaxations $\texttt{Fractional}_{\texttt{SDP}}$ and $\texttt{Cardinality}_{\texttt{SDP}}$, as well as the quality of the resulting relaxations. Specifically, the relaxation quality is the gap between the objective value of the best solution found via the MIO method `Fractional` (val_{MIO}) and the corresponding lower bound proven by relaxation (val_{SDP}):

$$\textbf{Relax} = \frac{\text{val}_{\text{MIO}} - \text{val}_{\text{SDP}}}{\text{val}_{\text{MIO}}}.$$

We observe from Table 2 that $\texttt{Fractional}_{\texttt{SDP}}$ is between 20 and 50 times faster than $\texttt{Cardinality}_{\texttt{SDP}}$. Also, larger speedups correspond to instances with larger values of p. In addition, by comparing the MIO formulations (Table 1) and the conic relaxations (Table 2), we make the following observations. First, conic relaxations can be substantially faster than MIOs, which is not surprising as the problems tackled are simpler. Nonetheless, as shown by the small optimality gaps in Table 2 and as will be argued further in Section 4.3, the solutions from the conic relaxation can be excellent estimators. Second, in the instances not solved to optimality by the MIO methods, the gaps reported by the conic relaxations are smaller, indicating that the lower bound obtained by solving this relaxation is better than the lower bound obtained after one hour of branch and bound. These results also indicate that the feasible solution found by MIO is better than what the optimality gap from branch and bound indicates and may, in fact, be optimal in many cases (indeed, the solution with respect to MSE in "Crime" was, in fact, proven optimal after solving the conic relaxation).

We conclude from our experiments that by adopting a fractional optimization perspective, feature selection problems with information criteria can be solved substantially faster than by using the existing approaches. These benefits are further compounded when the fractional optimization methods are combined with novel approaches for tackling MIQO problems.

4.3. Statistical Performance

In this section, we replicate the simulation setup used by Hastie et al. (2017) to compare the statistical performance of feature selection methods with different information criteria and test the performance of solving the cardinality constrained problem (3) while using hold-out validation to select the right parameter k. In our computations, we use $n = 1,000$, $p = 100$, $\rho = 0.35$, $\nu \in \{0.05, 0.09, 0.14, 0.25, 0.42, 0.71, 1.22, 2.07, 3.52, 6.00\}$, and $k_0 = \{5, 10, 25\}$.

Table 2. Performance for Conic Relaxations in Real Data Sets

		MSE		BIC		AICc	
Instance	Method	Time	Relax (%)	Time	Relax (%)	Time	Relax (%)
Housing	$\texttt{Fractional}_{\texttt{SDP}}$	0.1	0.0	0.1	0.0	0.1	0.0
	$\texttt{Cardinality}_{\texttt{SDP}}$	0.7	—	0.7	—	0.7	—
AutoMPG	$\texttt{Fractional}_{\texttt{SDP}}$	0.7	1.0	0.9	4.3	0.8	1.8
	$\texttt{Cardinality}_{\texttt{SDP}}$	6.8	—	6.8	—	6.8	—
SolarFlareC	$\texttt{Fractional}_{\texttt{SDP}}$	0.5	0.3	0.6	1.5	0.4	0.4
	$\texttt{Cardinality}_{\texttt{SDP}}$	8.7	—	8.7	—	8.7	—
SolarFlareM	$\texttt{Fractional}_{\texttt{SDP}}$	0.5	0.2	0.5	0.6	0.5	0.3
	$\texttt{Cardinality}_{\texttt{SDP}}$	8.7	—	8.7	—	8.7	—
SolarFlareX	$\texttt{Fractional}_{\texttt{SDP}}$	0.5	0.1	0.4	0.0	0.5	0.1
	$\texttt{Cardinality}_{\texttt{SDP}}$	9.0	—	9.0	—	9.0	—
BreastCancer	$\texttt{Fractional}_{\texttt{SDP}}$	1.4	1.4	6.6	5.4	1.6	3.6
	$\texttt{Cardinality}_{\texttt{SDP}}$	28.1	—	28.1	—	28.1	—
Diabetes	$\texttt{Fractional}_{\texttt{SDP}}$	17.6	2.9	36.1	6.7	17.0	4.2
	$\texttt{Cardinality}_{\texttt{SDP}}$	503.1	—	503.1	—	503.1	—
Crime	$\texttt{Fractional}_{\texttt{SDP}}$	125.2	0.0	158.6	2.4	125.7	0.9
	$\texttt{Cardinality}_{\texttt{SDP}}$	4,004.0	—	4,004.0	—	4,004.0	—
Insurance	$\texttt{Fractional}_{\texttt{SDP}}$	1,305.7	1.1	1,140.0	1.1	1,113.3	1.2
	$\texttt{Cardinality}_{\texttt{SDP}}$	16hrs	—	16hrs	—	16hrs	—

4.3.1. Methods. We compare the performance of the following methods.

Hold-out validation: The data are partitioned into a training set and a validation set, each of size $n/2$. The best subset selection problem (3) is solved on the training set for all values of $k = 0, 1, \ldots, 2k_0$, and the estimator that results in the smallest prediction error on the validation set is used. This method corresponds to the best subset selection method used in Hastie et al. (2017).

MSE: The estimator that minimizes the MSE.

BIC: The estimator that minimizes the BIC.

AICc: The estimator that minimizes the AICc.

For both the hold-out validation and information criteria methods, we tested the MIO-based and conic relaxation methods. For MIO methods, we set a time limit of 3 minutes to compute the estimators; note that for hold-out validation, this amounts to an average of $3/(2k_0)$ minutes per problem (although some problems may be allocated more time, as discussed in Section 4.2.2). For conic relaxations, we do not set a time limit and solve all convex problems to optimality.

4.3.2. Metrics. To evaluate the performance of each method, we consider the following metrics:

(i) The relative test error given by

$$\frac{\mathbb{E}\left(y_0 - \boldsymbol{x}_0'\hat{\boldsymbol{\beta}}\right)^2}{\sigma^2} = \frac{\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{\mathbf{0}}\right)'\Sigma\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{\mathbf{0}}\right) + \sigma^2}{\sigma^2},$$

where $\boldsymbol{x}_0 \in \mathbb{R}^p$ denotes a test predictor drawn from $\mathcal{N}_p(\mathbf{0}, \Sigma)$, y_0 is its associated response drawn from $\mathcal{N}_p(\boldsymbol{x}_0'\boldsymbol{\beta}^{\mathbf{0}}, \sigma^2)$, and $\hat{\boldsymbol{\beta}}$ is an estimator obtained from a given regression procedure

(ii) The *support recovery* (i.e., the number of correctly/incorrectly identified predictor variables)

(iii) The total time required to compute the estimator

Observe that the relative test error was also used as a metric in Hastie et al. (2017).

4.3.3. Results. We generated, for each combination of parameters ν and k_0, 10 instances with identical parameters, and we report the averages across all replications. Specifically, Table 3 reports for each value of k_0 the average total time required to compute the estimators (the averages are also taken across all SNRs). Consistent with the results reported in Section 4.2, we observe that the conic relaxations with respect to information criteria are substantially faster to compute and that the conic problems are solved to optimality in less than three minutes, the time limit given for MIO problems. We also point out that, on average, the conic relaxation of each cardinality-constrained problem solved in hold-out validation requires 62 seconds; thus we see that conic problems with respect to information criteria are solved in the time required to solve two cardinality-constrained problems (or less).

Table 3. Average Computational Time (in Seconds) of Conic Relaxations in Synthetic Instances with $n = 1,000$, $p = 100$, and $\rho = 0.35$

Setting	MSE	BIC	AICc	Hold-out validation
$k_0 = 5$	133	69	132	575
$k_0 = 10$	131	75	129	1,153
$k_0 = 25$	57	33	52	3,075

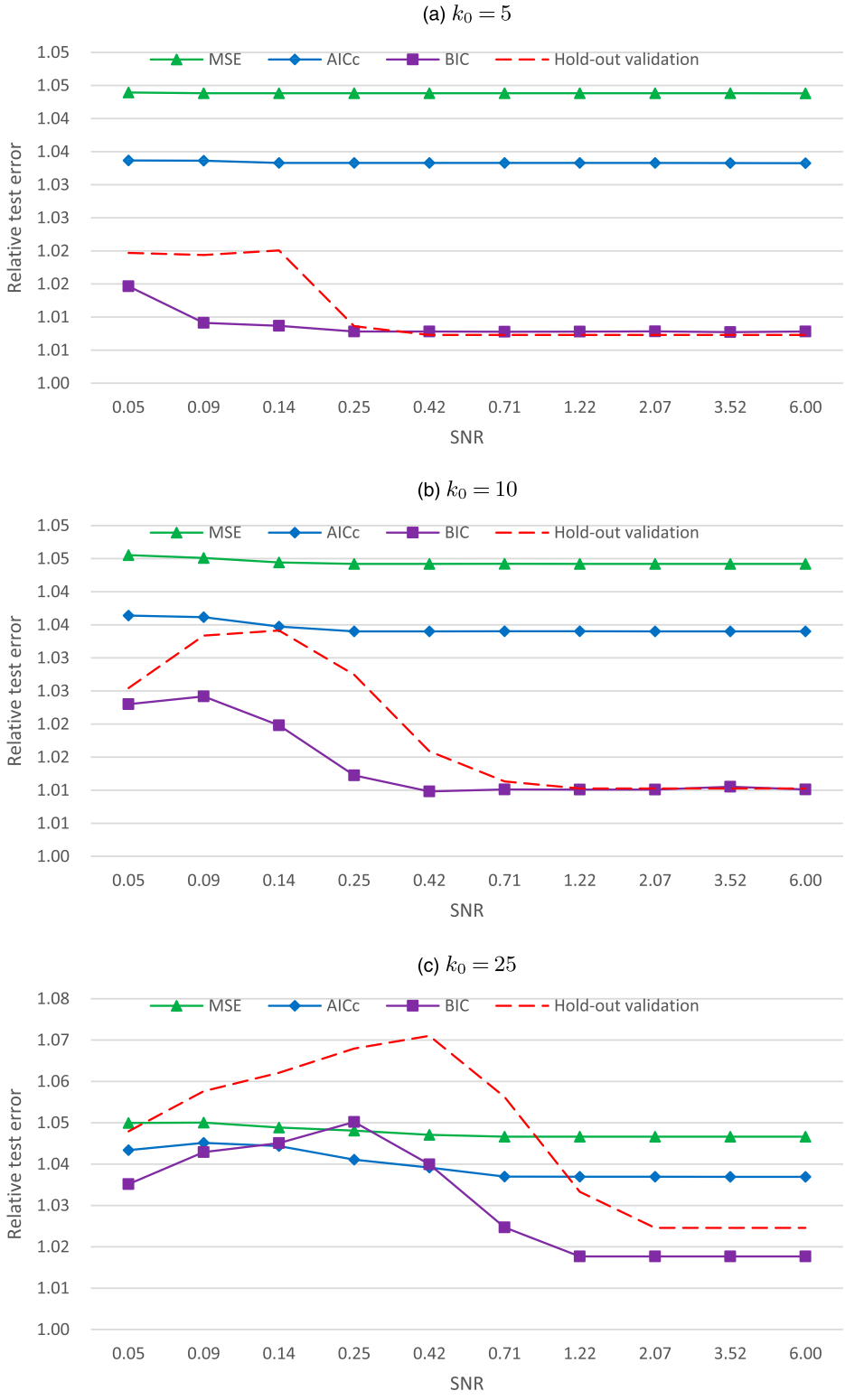
Note. All MIO methods hit the time limit of three minutes without proving optimality.

In terms of statistical performance, we observe that MIO formulations and conic relaxations deliver almost identical solutions: in SNR regimes with $\nu \leq 0.25$, the estimators from the conic relaxations have a slightly lower error than their MIO counterparts (by an average of 0.2%), whereas in regimes with $\nu > 0.25$, the two errors are, on average, the same. This similarity indicates on the one hand that conic relaxations indeed produce very-high-quality estimators and on the other hand that MIO methods find optimal or near-optimal solutions in a short time limit, even if proving optimality would require a long time. Because both methods result in very similar performance, but the conic relaxations are slightly superior in low SNR regimes, we report only those results.

Figures 1 and 2 depict for different values of parameters k_0 and signal-noise ratios ν the test error and support recovery, respectively. We observe from Figure 1 that AICc dominates MSE and that BIC dominates hold-out validation in terms of prediction accuracy. Moreover, although the quality of the predictions of MSE and AICc are fairly insensitive to the SNR and true sparsity parameter k_0, the performance of BIC and hold-out validation depends on those parameters. In particular, both BIC and hold-out validation perform (comparatively) better when the true model is very sparse (i.e., low values of k_0) and in very low and very high SNRs. By contrast, AICc performs better for denser models and for medium SNR values. We see that the performance of hold-out validation is especially poor for $k_0 = 25$, being outperformed by all other methods for several SNR values (i.e., for $0.09 \leq \nu \leq 0.71$). We attribute, in part, the superior performance of information criteria approaches such as BIC to the lack of hold-out validation, which requires holding out a portion of the data for validation purposes.

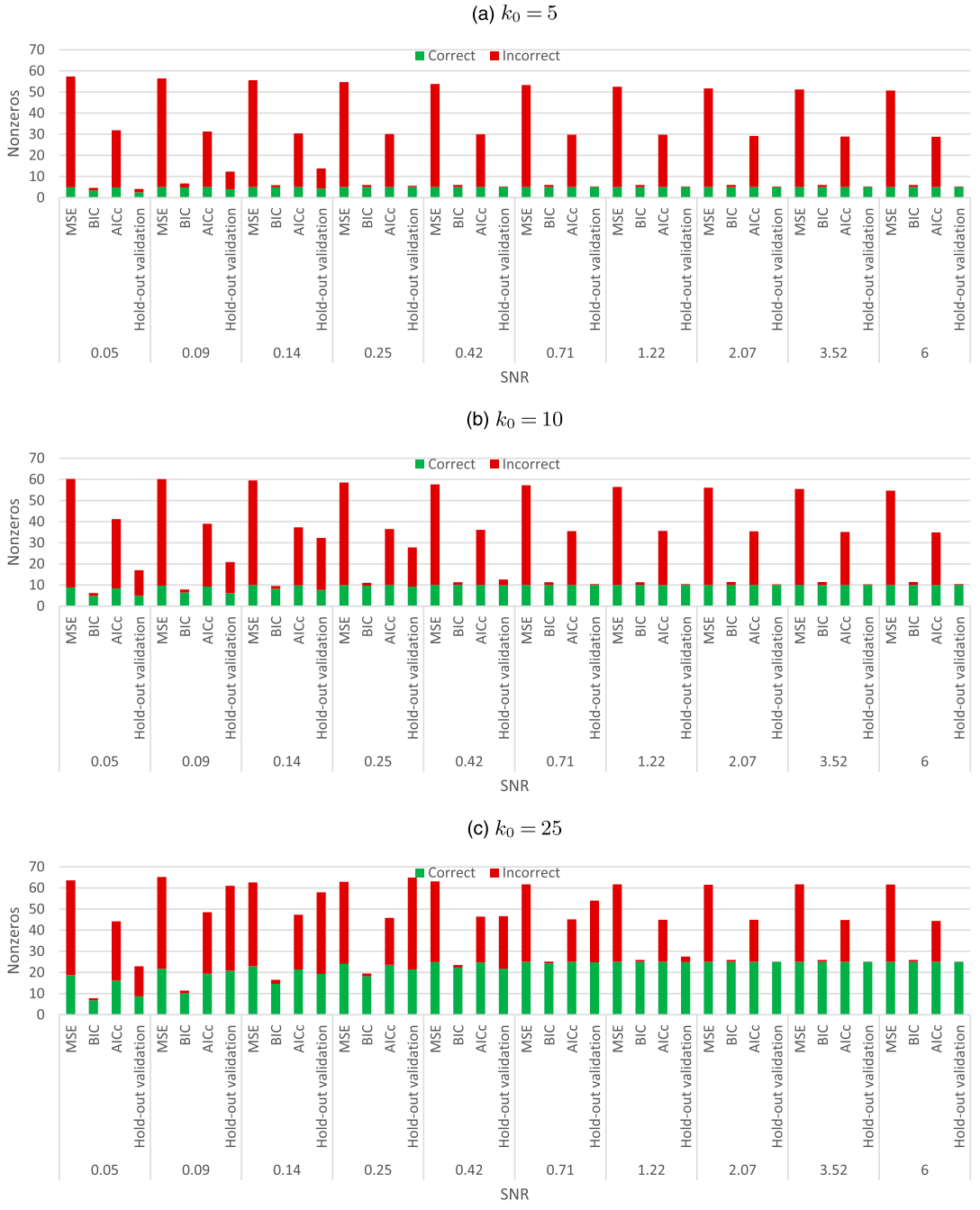
From Figure 2 we see that BIC achieves its good prediction performance in low SNRs by selecting a small number of predictor variables, but most of those match the support of the "true" regression

Figure 1. (Color online) Relative Test Error of Conic Estimators as a Function of the SNR in Synthetic Instances with $n = 1{,}000$, $p = 100$, and $\rho = 0.35$



coefficients β^0. As the SNR increases, the number of predictor variables chosen by BIC gradually increases until achieving an almost exact recovery of the true support of β^0. We see that hold-out validation is also able to recover the true support in large SNR regimes but may choose a relatively large number of incorrect regression predictors in low SNR regimes when compared with BIC. By contrast, MSE and AICc

Figure 2. (Color online) Support Recovery of Conic Estimators as a Function of the SNR in Synthetic Instances with $n = 1,000$, $p = 100$, and $\rho = 0.35$



fail to recover the true support for $\nu \leq 6$; in general, MSE selects a larger number of "incorrect" predictors, which explains why its prediction performance is worse than that of AICc. Nonetheless, for medium values of the SNR value, AICc chooses a larger number of true predictors than does BIC or hold-out validation with a modest amount of incorrect ones, leading to better prediction performance.

We conclude this section by summarizing our main computational findings:

- The parametric method described in Section 3 is able to solve to optimality problems an order of magnitude faster than previous approaches for information criteria. The speedup is more pronounced when paired with recent convexification methods.

• Optimizing with respect to information criteria via the new conic relaxations can be substantially "cheaper" than performing simple hold-out validation (and would be much faster than k-fold cross validation) while delivering comparable or even superior statistical performance (depending on the regime and criterion used).

• In terms of the performance of information criteria, we report the following findings:

—The BIC criterion delivers sparser solutions than other criteria and is the best at identifying the true sparsity pattern (validating the theoretical derivation of the criterion). It also delivers excellent prediction capabilities (although it is highly dependent on the SNR) and consistently outperforms hold-out validation.

—The AICc, although unable to identify the correct sparsity pattern, delivers good predictions and is fairly insensitive to the SNR in terms of the relative test error. It outperforms other methods when the underlying model is relatively dense (25% of nonzeros).

—The MSE criterion, corresponding to the simple and popular "adjusted" R^2 metric (but without the theoretical justifications of other criteria), is dominated by the AICc criterion in terms of both prediction and support recovery. It also performs worse than BIC and hold-out validation in most (but not of all) of the scenarios considered.

4.4. Additional Discussion

One of the main advantages of the parametric approach given in Section 3 is that it reduces optimization with respect to highly nonlinear criteria such as (6) and (7) to solving a sequence of MIQO optimization problems, for which specialized methods, well beyond simply using off-the-shelf solvers, exist. In this paper, we illustrate one such approach by using the conic relaxations for MIQO derived in Atamtürk and Gómez (2019a). We now give pointers to alternatives and briefly discuss their integration with the parametric method.

• Hazimeh and Mazumder (2018) propose a coordinate-descent method for the ℓ_0-regularized best subset selection that delivers locally optimal solutions and scales to problems with $p \sim 10^5$. This method can be used to solve each subproblem in the Newton method described in Section 3, resulting in a method that quickly finds high-quality solutions to problems with respect to information criteria. The integration with respect to the MSE criterion is straightforward, whereas other criteria require minor modifications to account for the submodular regularization $-tg(\mathbf{1}'z)$.

In addition, if an additional regularization term is added in the numerator of (4) and inside the logarithm in (5)–(7)—in which case the resulting problem can be interpreted as a robustification of the original problem—the following methods could be used as well:

• Bertsimas and Van Parys (2020) propose a linear outer approximation algorithm to tackle the sparse regression problems, and they show that computational times can be reduced in cross validation by reusing the approximation constructed in earlier problems. The same algorithm can be used to solve subproblems arising in Algorithm 1, and the linear outer approximations constructed can be reused in subsequent iterations.

• Pilanci et al. (2015), Dong et al. (2015), and Xie and Deng (2020) propose to solve conic relaxations that, although weaker than the one used in this paper, are simpler and scale to larger instances while preserving good statistical properties. These relaxations could be used instead of the relaxation proposed by Atamturk and Gomez (2019a).

• Atamtürk and Gómez (2020) propose safe screening rules to quickly fix discrete variables to 0 or 1 while preserving optimality guarantees. By using the Lovász extension of the submodular function $-tg(\mathbf{1}'z)$, similar rules could be derived for problems with information criteria.

As the methods for tackling problem (3) keep improving rapidly, the fractional optimization approach presented in this paper allows the direct incorporation of those methods to tackle problems with information criteria.

5. Conclusion

We present an MIFO framework to support the best subset selection in linear regression under a variety of criteria proposed in the literature. We use an underlying submodular function that arises with most of the criteria considered to strengthen the formulations, and we propose to tackle the resulting optimization problems by solving a sequence of MIQO problems (or their relaxations). We report encouraging results in our computational experiments, with respect to both the optimization and statistical performance. Because of the ubiquity of the information criteria in subset selection and other more general feature selection problems, the proposed methodologies may be potentially applicable in contexts other than linear regression.

Acknowledgments

The authors thank the associate editor and three anonymous reviewers for their constructive and helpful comments.

Endnote

[1] We also tested a parallel implementation of this method by allocating one hour to each problem and not using warm starts. We found that the sequential implementation with warm starts produces solutions close or equal to the parallel implementation, although gaps

can be much larger in the more difficult instances. For the sake of brevity, we omit this approach from our computations.

References

Abhishek K, Leyffer S, Linderoth J (2010) FilMINT: An outer approximation-based solver for convex mixed-integer nonlinear programs. *INFORMS J. Comput.* 22(4):555–567.

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

Atamtürk A, Gómez A (2018) Strong formulations for quadratic optimization with M-matrices and indicator variables. *Math. Programming* 170(1):141–176.

Atamtürk A, Gómez A (2019a) Rank-one convexification for sparse regression. Preprint, submitted January 29, https://arxiv.org/abs/1901.10334.

Atamtürk A, Gómez A (2019b) Simplex QP-based methods for minimizing a conic quadratic objective over polyhedra. *Math. Programming Comput.* 11(2):311–340.

Atamtürk A, Gómez A (2020) Safe screening rules for ℓ_0-regression. Daume H III, Singh A, eds. *Proc. Internat. Conf. Machine Learn.* 119:421–430.

Atamtürk A, Jeon H (2019) Lifted polymatroid inequalities for mean-risk optimization with indicator variables. *J. Global Optim.* 73(4): 677–699.

Atamtürk A, Narayanan V (2008) Polymatroids and mean-risk minimization in discrete optimization. *Oper. Res. Lett.* 36(5):618–622.

Atamtürk A, Deck C, Jeon H (2020) Successive quadratic upper-bounding for discrete mean-risk minimization and network interdiction. *INFORMS J. Comput.* 32(2):346–355.

Atamtürk A, Gómez A, Han S (2018) Sparse and smooth signal estimation: Convexification of L0 formulations. Preprint, submitted November 6, https://arxiv.org/abs/1811.02655.

Bertsimas D, King A (2015) An algorithmic approach to linear regression. *Oper. Res.* 64(1):2–16.

Bertsimas D, Shioda R (2009) Algorithm for cardinality-constrained quadratic optimization. *Comput. Optim. Appl.* 43(1):1–22.

Bertsimas D, Van Parys B (2020) Sparse high-dimensional regression: Exact scalable algorithms and phase transitions. *Ann. Statist.* 48(1):300–323.

Bertsimas D, Cory-Wright R, Pauphilet J (2019a) A unified approach to mixed-integer optimization: Nonlinear formulations and scalable algorithms. Preprint, submitted July 3, https://arxiv.org/abs/1907.02109.

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

Bertsimas D, Pauphilet J, Van Parys B (2019b) Sparse regression: Scalable algorithms and empirical performance. Preprint, submitted February 18, https://arxiv.org/abs/1902.06547.

Bonami P, Biegler LT, Conn AR, Cornuéjols G, Grossmann IE, Laird CD, Lee J, Lodi A, Margot F, Sawaya N (2008) An algorithmic framework for convex mixed integer nonlinear programs. *Discrete Optim.* 5(2):186–204.

Borrero JS, Gillen C, Prokopyev OA (2017) Fractional 0–1 programming: Applications and algorithms. *J. Global Optim.* 69(1):255–282.

Bühlmann P, van de Geer S (2011) *Statistics for High-Dimensional Data: Methods, Theory and Applications* (Springer Science & Business Media, Berlin).

Chen Y, Ye Y, Wang M (2019) Approximation hardness for a class of sparse optimization problems. *J. Machine Learn. Res.* 20(38):1–27.

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

Cozad A, Sahinidis NV, Miller DC (2015) A combined first-principles and data-driven approach to model building. *Comput. Chemical Engrg.* 73(February):116–127.

Dheeru D, Karra Taniskidou E (2017) UCI Machine Learning Repository. University of California, Irvine, Irvine. http://archive.ics.uci.edu/ml.

Dinkelbach W (1967) On nonlinear fractional programming. *Management Sci.* 13(7):492–498.

Dong H, Chen K, Linderoth J (2015) Regularization vs. relaxation: A conic optimization perspective of statistical variable selection. Preprint, submitted October 20, https://arxiv.org/abs/1510.06083.

Edmonds J (1970) Submodular functions, matroids, and certain polyhedra. Guy RK, ed. *Combin. Structures Their Appl.: Proc. Calgary Internat. Conf. Combin. Structures Their Appl.* (Gordon and Breach, New York), 69–87.

Efron B, Hastie T, Johnstone I, Tibshirani R (2004) Least angle regression. *Ann. Statist.* 32(2):407–499.

Furnival GM, Wilson RW Jr (1974) Regressions by leaps and bounds. *Technometrics* 16(4):499–511.

Garside M (1965) The best subset in multiple regression analysis. *J. Roy. Statist. Soc. Ser. C Appl. Statist.* 14(2-3):196–200.

Garside M (1971a) Algorithm AS 38: Best subset search. *J. Roy. Statist. Soc. Ser. C Appl. Statist.* 20(1):112–115.

Garside M (1971b) Some computational procedures for the best subset problem. *J. Roy. Statist. Soc. Ser. C Appl. Statist.* 20(1):8–15.

Glover F (1975) Improved linear integer programming formulations of nonlinear integer problems. *Management Sci.* 22(4): 455–460.

Gómez A (2020) Strong formulations for conic quadratic optimization with indicator variables. *Math. Programming*, ePub ahead of print April 30, https://doi.org/10.1007/s10107-020-01508-y.

Günlük O, Linderoth JT (2010) Perspective reformulations of mixed integer nonlinear programs with indicator variables. *Math. Programming* 124(1–2):183–205.

Hastie T, Tibshirani R, Tibshirani RJ (2017) Extended comparisons of best subset selection, forward stepwise selection, and the Lasso. Preprint, submitted July 27, https://arxiv.org/abs/1707.08692.

Hazimeh H, Mazumder R (2018) Fast best subset selection: Coordinate descent and local combinatorial optimization algorithms. Preprint, submitted March 5, https://arxiv.org/abs/1803.01454.

Hurvich CM, Tsai C-L (1989) Regression and time series model selection in small samples. *Biometrika* 76(2):297–307.

Jeon H, Linderoth JT, Miller A (2017) Quadratic cone cutting surfaces for quadratic programs with on–off constraints. *Discrete Optim.* 24(May):32–50.

Kenney A, Chiaromonte F, Felici G (2018) Efficient and effective ℓ_0 feature selection. Preprint, submitted August 7, https://arxiv.org/abs/1808.02526.

Kimura K, Waki H (2018) Minimization of Akaike's information criterion in linear regression analysis via mixed integer nonlinear program. *Optim. Methods Software* 33(3):633–649.

Konishi S, Kitagawa G (2008) *Information Criteria and Statistical Modeling* (Springer Science & Business Media, New York).

Lovász L (1983) Submodular functions and convexity. Bachem A, Grötschel M, Korte B, eds. *Mathematical Programming The State of the Art* (Springer, Berlin), 235–257.

Mahajan A, Leyffer S, Linderoth J, Luedtke, J, Munson T (2017) Minotaur: A mixed-integer nonlinear optimization toolkit. Preprint, submitted October 16, http://www.optimization-online.org/DB_HTML/2017/10/6275.html.

Megiddo N (1979) Combinatorial optimization with rational objective functions. *Math. Oper. Res.* 4(4):414–424.

Miller A (2002) *Subset Selection in Regression* (CRC Press, Boca Raton, FL).

Miyashiro R, Takano Y (2015a) Mixed integer second-order cone programming formulations for variable selection in linear regression. *Eur. J. Oper. Res.* 247(3):721–731.

Miyashiro R, Takano Y (2015b) Subset selection by Mallows' Cp: A mixed integer programming approach. *Expert Systems Appl.* 42(1):325–331.

Park YW, Klabjan D (2017) Subset selection for multiple linear regression via optimization. Preprint, submitted January 27, https://arxiv.org/abs/1701.07920.

Pilanci P, Wainwright MJ, El Ghaoui L (2015) Sparse learning via Boolean relaxations. *Math. Programming* 151(1):63–87.

Radzik T (1998) Fractional combinatorial optimization. Du D-Z, Pardalos PM, eds. *Handbook of Combinatorial Optimization* (Springer, New York), 429–478.

Schwarz G (1978) Estimating the dimension of a model. *Ann. Statist.* 6(2):461–464.

Seber GA, Lee AJ (2003) *Linear Regression Analysis*, 2nd ed. (John Wiley & Sons, Hoboken, NJ).

Sugiura N (1978) Further analysts of the data by Akaike's information criterion and the finite corrections. *Comm. Statist. Theory Methods* 7(1):13–26.

Takano Y, Miyashiro R (2020) Best subset selection via cross-validation criterion. *TOP* 28(2):475–488.

Tibshirani R (1996) Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc. Ser. B.* 58(1):267–288.

Tibshirani R (2011) Regression shrinkage and selection via the Lasso: A retrospective. *J. Roy. Statist. Soc. Ser. B Statist. Methodol.* 73(3): 273–282.

Wainwright MJ (2009) Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Inform. Theory* 55(5): 2183–2202.

Wei L, Gómez A, Küçükyavuz S (2020a) Ideal formulations for constrained convex optimization problems with indicator variables. Preprint, submitted June 30, https://arxiv.org/abs/2007.00107.

Wei L, Gómez A, Küçükyavuz S (2020b) On the convexification of constrained quadratic optimization problems with indicator variables. Bienstock D, Zambelli G, eds. *Internat. Conf. Integer Programming Combin. Optim.* (Springer, Cham, Switzerland), 433–447.

Weisberg S (2005) *Applied Linear Regression* (John Wiley & Sons, Hoboken, NJ).

Wherry R (1931) A new formula for predicting the shrinkage of the coefficient of multiple correlation. *Ann. Math. Statist.* 2(4): 440–457.

Wilson ZT, Sahinidis NV (2017) The ALAMO approach to machine learning. *Comput. Chemical Engrg.* 106(November):785–795.

Xie W, Deng X (2020) Scalable algorithms for sparse ridge regression. Preprint, submitted June 11, https://arxiv.org/abs/1806.03756.

Zhang C-H, Huang J (2008) The sparsity and bias of the Lasso selection in high-dimensional linear regression. *Ann. Statist.* 36(4): 1567–1594.