A MIXED-INTEGER FRACTIONAL OPTIMIZATION APPROACH TO
BEST SUBSET SELECTION

ANDRÉS GÓMEZ AND OLEG A. PROKOPYEV

Abstract. We consider the best subset selection problem in linear regression, i.e., finding
a parsimonious subset of the regression variables that provides the best fit to the data
according to some predefined criterion. We show that, for a broad range of criteria used in
the statistics literature, the best subset selection problem can be modeled as a mixed-integer
fractional optimization problem. Then we show how to exploit underlying submodularity in
the problem to strengthen the formulations, and propose to tackle the problem by solving
a sequence of mixed-integer quadratic optimization problems. The proposed approach can
be implemented with off-the-shelf solvers and, in our computations, it outperforms existing
alternatives by orders of magnitude.

June 2019

1. Introduction

Given design matrix $X = [x_1, \ldots, x_p] \in \mathbb{R}^{n \times p}$ of explanatory (independent) variables and
vector $y \in \mathbb{R}^n$ of response (dependent) variables, the linear regression approach assumes that
the relationship between them is given by

$$y = X\beta + \epsilon,$$

(1)

where $\beta \in \mathbb{R}^p$ is a vector of parameters known as the regression coefficients and $\epsilon \in \mathbb{R}^n$
are the error terms. Vectors $x_1, \ldots, x_p$ are often referred to as regressors, while $y$ is also
known as regressand. We assume throughout the paper that $n > p$. Linear regression is the
most widely used and studied model in the regression analysis literature, see Seber and Lee
(2003), Weisberg (2005). It finds numerous applications for data analysis in a wide variety
of important real-life contexts.
Simply speaking, the goal of the linear regression approach is to find appropriate values for parameters $\beta$ such that the data fitting error is minimized according to some predefined criterion or criteria. The ordinary least squares estimates (OLS), found by minimizing the residual squared error, are easy to compute but suffer from poor prediction accuracy and interpretability. In particular, 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, also known as feature selection in the data mining area, has attracted significant attention in the statistical, machine learning and optimization literature.

A classical model for the subset selection problem (Miller 2002) is

$$\begin{align*}
\min_{\beta \in \mathbb{R}^p} & \quad \| y - X\beta \|^2_2 \\
\text{subject to} & \quad \| \beta \|_0 \leq k,
\end{align*}$$

where $k$ is some predefined sparsity parameter and $\| \cdot \|_0$ is the $\ell_0$-norm, i.e., $\| \beta \|_0 = \sum_{i=1}^p 1_{\{\beta_i \neq 0\}}$ with $1_{\cdot}$ denoting the indicator function. Problem (2) is NP-hard (Natarajan 1995), and several approaches to tackle it approximately or exactly have been proposed in the literature.

Perhaps the most widely known and used approximation approach is Lasso (Tibshirani 1996), where the $\ell_0$-norm is replaced by the convex $\ell_1$-norm. The resulting convex problems can be solved very efficiently, e.g., see Efron et al. (2004), and Lasso has indeed proven useful in finding sparse models in practice. Lasso has also been found to enjoy desirable theoretical properties under appropriate conditions on the data of the problem (Bühlmann and Van De Geer 2011, Tibshirani 2011, Wainwright 2009, Zhang et al. 2008).

However, since Lasso is only a surrogate, it is not without shortcomings. It may actually result in dense solutions, or solutions with poor statistical performance, e.g., see Miller (2002), Shen et al. (2013), Zheng et al. (2014). Researchers have also considered a variety of non-convex regularization terms (Frank and Friedman 1993, Mazumder et al. 2011) that
have theoretical advantages over Lasso. Typical algorithms for such problems return only local minima and thus, may potentially lead to low quality solutions.

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 (2017), Cozad et al. (2015), Miyashiro and Takano (2015b). Specifically, by introducing binary variables $z \in \{0, 1\}^p$ such that $z_i = 1$ iff $\beta_i \neq 0$, problem (2) can be formulated as

$$\min \| y - X\beta \|^2_2$$

s.t. $1'z \leq k$ \hspace{1cm} (3a)

$\ell \circ z \leq \beta \leq u \circ z$ \hspace{1cm} (3b)

$z \in \{0, 1\}^p$, $\beta \in \mathbb{R}^p$, \hspace{1cm} (3c)

where $1$ denotes a $p$-dimensional vector of all ones, $\ell$ and $u$ are $p$-dimensional vectors of sufficiently small and large numbers, respectively, and $\ell \circ z$ is the Hadamard product of $\ell$ and $z$. Problem (3) is a mixed-integer quadratic optimization (MIQO) problem, which can be solved directly with off-the-shelf solvers for convex MIO. Cozad et al. (2014) also propose a mixed-integer linear optimization (MILO) surrogate of (3).

Note that problem (2) requires specifying a priori the desired sparsity $k$ at the right-hand side of (3b). The standard technique for determining $k$ is using cross-validation, which solves (3) for multiple values of $k$ and selects the one that performs best in a held-out validation set. However, cross-validation has two main drawbacks: first, it requires solving the hard MIO problem (3) several times, and may be prohibitively expensive in many settings; second, it requires partitioning the dataset into training and validation sets, which reduces the number of datapoints used for training the model and effectively reduces the Signal-to-Noise Ratio.
(SNR). The second drawback is compounded by the fact that estimators produced from (3) have been found to be poor in low SNR settings when compared with simpler techniques such as Lasso (Hastie et al. 2017).

This paper is primarily concerned with alternatives to cross-validation, which can be performed efficiently and do not require partitioning the data.

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 $\|y - X\beta\|_2^2$ and the size of the model $\|\beta\|_0$.

**MSE:** The Mean Squared Error (Wherry 1931) of a regression model is given by

$$\frac{\|y - X\beta\|_2^2}{n} - \|\beta\|_0.$$  \hspace{1cm} (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 due to its simplicity. The MSE is an unbiased estimator of the variance of the errors $\epsilon$ under appropriate assumptions.

**AIC:** The Akaike Information Criterion (Akaike 1974) is

$$n \ln \left( \frac{\|y - X\beta\|_2^2}{n} \right) + 2\|\beta\|_0 + K,$$ \hspace{1cm} (5)

where $K$ is a constant that does not depend on the model. The AIC is derived from a maximum likelihood analysis and, under appropriate assumptions, minimization of AIC yields the model with minimal Kullback-Leibler divergence with respect to a true distribution.

**BIC:** The Bayesian Information Criterion (Schwarz et al. 1978) is

$$n \ln \left( \frac{\|y - X\beta\|_2^2}{n} \right) + \ln(n)\|\beta\|_0 + K.$$ \hspace{1cm} (6)

BIC is also derived from a maximum likelihood analysis and, under appropriate assumptions, minimization of BIC yields the model that is a posteriori most probable.
HQIC: The Hannan-Quinn Information Criterion (Hannan and Quinn 1979) is

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

which is derived as a consistent estimator of the order of an autoregression.

AICc: Hurvich and Tsai (1989) note that AIC is biased when $n$ is not much larger than $p$, and propose the corrected AIC

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

(8)

to correct the bias.

The criteria outlined above 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 $\epsilon$ is known, problems with AIC, BIC and HQIC can be simplified to MIQO. 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, i.e.,

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

(9)

where $F(\beta, k)$ corresponds to one of the above criteria given by (4)-(8). Note that for a fixed $k = \|\beta\|_0$, finding the best model with respect to any criterion in (4)-(8) can be done by minimizing $\|y - X\beta\|_2^2$. Thus, approach (9) requires solving $p + 1$ different MIO problems, and is, to the best of our knowledge, the most efficient method to date.

Finally, Miyashiro and Takano (2015a) propose to use 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 requires the addition of $p + 1$ additional
binary variables. The authors report that the MISOCO formulations perform worse than (9) by an order of magnitude.

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 (9) 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, 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 review the MISOCO formulations with \( p + 1 \) additional variables proposed by Miyashiro and Takano (2015a), and exploit submodularity to derive stronger convex MIO formulations in the original space of variables. In Section 3 we discuss how to solve the resulting MIO by (partially) solving a sequence of MIQO problems. In Section 4 we provide computational experiments on synthetic and real datasets, and in Section 5 we conclude the paper.

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 (non-convex) logarithmic term in the objective function, see (5)-(8). In order to do so, we first show in Section 2.1 how to model the best subset selection problems as the (possibly non-convex) MIFO problem

\[
\min_{z \in \{0,1\}^p, \beta \in \mathbb{R}^p} \frac{\|y - X\beta\|_2^2}{g(1'z)} \quad \text{subject to } \ell \circ z \leq \beta \leq u \circ z,
\]  

(10)
where \( g : \mathbb{R}_+ \rightarrow \mathbb{R}_+ \) is a one-dimensional non-increasing convex function that depends on the criterion used. Then in Section 2.2 we discuss how to obtain mixed-integer convex formulations of (10) by exploiting submodularity. Finally, in Section 2.3, we show that the resulting formulations are at least as strong as 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.

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

\[
\min_{z \in \{0,1\}^p, \beta \in \mathbb{R}^p} \frac{\|y - X\beta\|_2^2}{n - 1'z} \quad \text{subject to } \ell \circ z \leq \beta \leq u \circ z,
\]

i.e., function \( g(x) = 1 - x \) is affine.

**AIC and related information criteria.** Consider optimization with respect to either AIC, BIC or HQIC given by (5), (6) and (7), respectively. The best model with respect to such criteria can be found by solving the MIO

\[
\min_{z \in \{0,1\}^p, \beta \in \mathbb{R}^p} \ln \left( \frac{\|y - X\beta\|_2^2}{n} \right) + \alpha 1'z \quad \text{subject to } \ell \circ z \leq \beta \leq u \circ z,
\]

where the constant terms in the definition of the criterion is dropped, and \( \alpha \) is a constant that may depend on \( n \) that is, \( \alpha = 2/n \) for AIC, \( \alpha = \ln(n)/n \) for BIC and \( \alpha = \ln \ln(n)/n \) for HQIC. Since the exponential function is non-decreasing and monotone, we can take the exponential of the objective function and find that (11) is equivalent to

\[
\frac{1}{n} \cdot \min_{z \in \{0,1\}^p, \beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 \cdot e^{\alpha 1'z} \quad \text{subject to } \ell \circ z \leq \beta \leq u \circ z
\]

\[
= \frac{1}{n} \cdot \min_{z \in \{0,1\}^p, \beta \in \mathbb{R}^p} \frac{\|y - X\beta\|_2^2}{e^{-\alpha 1'z}} \quad \text{subject to } \ell \circ z \leq \beta \leq u \circ z.
\]

Based on the above derivations, we see that (11) is a special case of (10), where \( g(x) = e^{-\alpha x} \).
AICc criterion. A similar approach can be used for optimization with respect to AICc given by (8), resulting in

$$
\begin{align*}
\min n \ln \left( \frac{\|y - X\beta\|^2}{n} \right) + 2 (1'z) + \frac{2(1'z)^2 + 2(1'z)}{n - 1'z - 1} \\
\text{subject to } \ell \circ z \leq \beta \leq u \circ z, \ z \in \{0, 1\}^p, \ \beta \in \mathbb{R}^p.
\end{align*}
$$

(12a)

After dividing by $n$, taking exponential of the objective function and some algebraic manipulations, problem (12) can be equivalently written as

$$
\begin{align*}
\frac{1}{ne^2} \min_{z \in \{0, 1\}^p, \beta \in \mathbb{R}^p} \frac{\|y - X\beta\|^2}{e^{-\frac{n-1}{n-1}}} \text{subject to } \ell \circ z \leq \beta \leq u \circ z.
\end{align*}
$$

Therefore, we see that (12) is a special case of (10), where $g(x) = e^{-\frac{n-1}{n-1}}$.

Other MIFO formulations in feature selection. Fractional formulations arise naturally when modeling trade-offs between competing objectives. Since best subset selection involves a trade-off between the fit and the size of the model, it is not surprising that formulation (10) includes most criteria used as its special cases. For example, if $g(x) = 1_{\{x \leq k\}}$, then (10) reduces to the cardinality constrained problem (3) – in this case $g$ is monotone non-increasing, but is non-convex.

Finally, fractional formulations for feature selection problems in contexts other than linear regression have also been proposed in the data mining literature, e.g., for consistent biclustering (Busygin et al. 2005) and feature selection with correlation measures (Peng et al. 2005, Nguyen et al. 2009). Note that the formulations in these papers involve only 0–1 variables and linear numerators/denominators and, unlike the formulations presented in this paper, can be handled using standard linearization techniques, e.g., see Borrero et al. (2017).

2.2. Convexification. Consider the mixed-integer set

$$
\mathcal{F} = \{z \in \{0, 1\}^p, \ s \in \mathbb{R}_+ : s \leq g(1'z)\}.
$$

(13)
Since $g$ is convex, the set function $g(1'z)$ is supermodular. Define $\pi_i = g(i) - g(i - 1)$, $i = 1, \ldots, p$, and given a permutation $((1), (2), \ldots, (p))$ of $[p]$, consider the inequality

$$s \leq g(0) + \sum_{i=1}^{p} \pi_i z_i.$$  

(14)

The coefficients $-\pi$ in (14) correspond to an extreme point of the extended polymatroid associated with the submodular function $-g$, and inequality (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), i.e.,

$$\text{conv}(\mathcal{F}) = \left\{ (z, s) \in [0, 1]^p \times \mathbb{R}_+ : s \leq g(0) + \sum_{i=1}^{p} \pi_i z_i, \text{ for all permutations of } [p] \right\} .$$

Thus, we can formulate (10) as the convex MIFO problem

$$\min \frac{\|y - X\beta\|_2^2}{s}$$

(15a)

s.t. $s \leq g(0) + \sum_{i=1}^{p} \pi_i z_i$, \ for all permutations of $[p]$  

(15b)

$$\ell \circ z \leq \beta \leq u \circ z$$

(15c)

$$z \in \{0, 1\}^p, \ \beta \in \mathbb{R}^p, \ s \geq 0,$$

(15d)

Note that there is a factorial number of linear constraints (15b). Therefore, to implement formulations (15) in practice, a lazy constraint generation scheme for inequalities (15b) should be used. 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): if variables $z$ are indexed such that $\bar{z}_1 \geq \bar{z}_2 \geq \ldots \geq \bar{z}_p$, then a most violated inequality (15b) at $(\bar{z}, \bar{\beta}, \bar{s})$ is precisely $s \leq g(0) + \sum_{i=1}^{p} \pi_i \bar{z}_i$.

Remark 1 (MSE criterion). If $g(x) = n - x$, corresponding to the MSE criterion, then each inequality (14) reduces to $s \leq n - 1'z$ – and this inequality can be changed to an equality constraint without loss of generality. Thus, in such case, formulation (15) reduces simply to
the convex MIFO

\[
\min_{z \in \{0,1\}^p, \beta \in \mathbb{R}^p} \frac{\|y - X\beta\|^2}{n - 1'z} \quad \text{subject to } \ell \circ z \leq \beta \leq u \circ z.
\]  

(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

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

s.t. \(\|y - X\beta\|^2 \leq t(n - \sum_{i=1}^{p} z_i)\)  

(17b)

\(\ell \circ z \leq \beta \leq u \circ z\)  

(17c)

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

(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

\[
\|y - X\beta\|^2 \leq tn - \sum_{i=1}^{p} v_i \quad (18a)
\]

\[0 \leq v_i \leq t, \quad t - M(1 - z_i) \leq v_i \leq Mz_i, \quad \forall i = 1, \ldots, p,\]

(18b)

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

2.3.2. MISOCO formulations. Miyashiro and Takano (2015a) proposed to tackle subset selection problems with information criteria using MISOCO formulations, discussed next.
MSE criterion. Constraint (17b) is a rotated cone constraint, thus problem (17) can directly be formulated as a MISOCO. Thus, the strength of the convex relaxation of (16) is the same as the MISOCO formulation (17) used in Miyashiro and Takano (2015a). However, in this paper, we tackle (16) as a fractional optimization problem (see Section 3), which leads to substantially better computational performance than tackling it as a MISOCO.

General criteria. For tackling problem (10), Miyashiro and Takano (2015a) propose to use the formulation

\[
\begin{align*}
\min_{z, \beta, w, s, t} & \quad t \\
\text{s.t.} & \quad \|y - X\beta\|_2^2 \leq ts \\
& \quad s \leq \sum_{i=0}^{p} g(i)w_i \\
& \quad \sum_{i=0}^{p} iw_i = 1'z \\
& \quad 1'w = 1 \\
& \quad \ell \circ z \leq \beta \leq u \circ z \\
& \quad z \in \{0, 1\}^p, w \in \{0, 1\}^{p+1}, \beta \in \mathbb{R}^p, s \geq 0, t \geq 0,
\end{align*}
\]  

i.e., using Special Ordered Sets of type 1 (SOS 1) with the introduction of additional variables \( w \). As we now show, 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 (19).

**Proposition 1.** The convex relaxation of (19) is equivalent to the convex optimization problem

\[
\begin{align*}
\min_{z, \beta, w, s, t} & \quad \frac{\|y - X\beta\|_2^2}{s} \\
\text{s.t.} & \quad s \leq g(0) + \frac{g(p) - g(0)}{p} \sum_{i=1}^{p} z_i
\end{align*}
\]
\[ \ell \circ z \leq \beta \leq u \circ z \quad (20c) \]
\[ z \in [0, 1]^p, \beta \in \mathbb{R}^p, s \geq 0. \quad (20d) \]

**Proof.** First, note that \( t = \| y - X\beta \|_2^2 / s \) in any optimal solution of (19) or of its convex relaxation. Consider the optimization problem

\[
\max_w \sum_{i=0}^{p} g(i)w_i \quad (21a)
\]
\[
\text{s.t. } \sum_{i=0}^{p} iw_i = 1'z, \ 1'w = 1, \ w \geq 0, \quad (21b)
\]

and denote by \( \gamma^* \) its optimal objective function value. Observe that a point \((z, \beta, s) \in [0, 1]^p \times \mathbb{R}^p \times \mathbb{R}_+\) satisfying \( \ell \circ z \leq \beta \leq u \circ z \) is feasible for the convex relaxation of (19) if and only if \( s \leq \gamma^* \).

We now claim that there exists an optimal solution of (21) where \( w_i = 0 \) whenever \( 0 < i < p \). Indeed, suppose that an optimal solution \( w \) of (21) satisfies \( w_j > 0 \) with \( 0 < j < p \). Since function \( g \) is convex, we have that \( g(j) \leq \left( 1 - \frac{j}{p} \right) g(0) + \frac{j}{p} g(p) \). Therefore the solution \( \tilde{w} \) given by

\[
\tilde{w}_i = \begin{cases} 
  w_0 + (1 - j/p)w_j & \text{if } i = 0 \\
  w_p + (j/p)w_j & \text{if } i = p \\
  0 & \text{if } i = j \\
  w_i & \text{otherwise}
\end{cases}
\]

satisfies all constraints of (21) and results in an equal or better objective function value.

Therefore, we can assume without loss of generality that in an optimal solution of (21) only \( w_0 \) and \( w_p \) are nonzero, i.e., \( w_p = 1'z/p \) and \( w_0 = 1 - 1'z/p \). Thus, we find that \((z, \beta, s) \in [0, 1]^p \times \mathbb{R}^p \times \mathbb{R}_+\) is feasible for the convex relaxation of (19) if and only if \( s \leq g(0) + \frac{g(p)-g(0)}{p} 1'z \).

\( \Box \)

**Corollary 1.** Formulation (15) has a stronger convex relaxation than (19).
Proof. The natural convex relaxations of (15) and (19) differ only in the use of either constraint (15b) or (20b). Since (20b) is only valid for \( \text{conv}(\mathcal{F}) \) while (15b) in fact describes \( \text{conv}(\mathcal{F}) \), the result follows. \( \square \)

3. Parametric MIQO approaches

Formulations (15) and (19) can be tackled with convex MIO solvers. However, MIQO 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 2017, Atamturk et al. 2017), and 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 \( t \)

\[
\text{(MIQO}_t\text{)} \quad \min d(t) = \min \|y - X\beta\|_2^2 - ts \quad (22a)
\]

\[
\text{s.t. } (15b) - (15d), \quad (22b)
\]

and recall that \( s = g(1'z) \) 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 (10). Hence, problem (10) reduces to finding a root for the function \( d(t) \), e.g., via bisection or Newton-like methods (Dinkelbach 1967, Megiddo 1979, Radzik 1998, Borrero et al. 2017).

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

\[
\hat{d}(t) = \|y - X\hat{\beta}(t)\|_2^2 - tg(1'\hat{z}(t)) < -\xi,
\]
or proves that $d(t) \geq -\xi$. For example, solve can be naturally implemented using branch-and-bound solvers for MIQO by: solving (22) to optimality and checking whether $d(t) < -\xi$; or stopping the algorithm when an incumbent solution with 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

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

Furthermore, let $(\beta^*, z^*)$ be an optimal solution for (10), and define for any feasible solution $(\bar{\beta}, \bar{z})$ the relative optimality gap as

$$\text{gap} = \frac{h(\bar{\beta}, \bar{z}) - h(\beta^*, z^*)}{h(\bar{\beta}, \bar{z})}.$$  

(23)

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

**Algorithm 1** Newton method for (10).

**Input:** $y$, response vector; $X$, model matrix; $\epsilon$, precision parameter.

**Output:** $\beta$, regression coefficients; $z$, selected features.

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

**Proposition 2.** If the time limit is not reached, then Algorithm 1 terminates with a feasible solution with $\text{gap} \leq \epsilon$. 


Proof. Observe that \((\beta^*, z^*)\) has the objective function value less than or equal to 0 for any subproblem (22) in line 6 – otherwise, \(t = h(\beta, z) \leq h(\beta^*, z^*)\), contradicting the optimality of \((\beta^*, z^*)\). Now suppose that at any iteration of Algorithm 1, a lower bound \(d_{LB}(t)\) on the optimal objective function value \(d(t)\) has been proven. We find

\[
d_{LB}(t) \leq \|y - X\beta^*\|_2^2 - tg(1'z^*) \leq 0
\]

\[
\Leftrightarrow \frac{d_{LB}(t)}{g(1'z^*)} + t \leq h(\beta^*, z^*) \leq t \quad (24)
\]

\[
\Rightarrow \frac{d_{LB}(t)}{g(p)} + t \leq h(\beta^*, z^*) \quad (25)
\]

where the implication holds since \(d_{LB}(t) \leq 0\) and \(g\) is non-increasing. From (25) we conclude

\[
gap \leq - \frac{d_{LB}(t)}{tg(p)}. \quad (26)
\]

Thus, if \(d_{LB}(t) \geq -\epsilon tg(p)\), see line 10, then \(\text{gap} \leq \epsilon\), and the relative optimality gap of the solution returned by the algorithm in line 11 is at most \(\epsilon\).

The result of Proposition 2 holds independently of the quality of the feasible solutions found in line 6 of the algorithm. However, as Proposition 3 below shows, high quality solutions may lead to substantially fewer iterations.

**Proposition 3.** If all problems \(\text{MIQO}_k\) in line 6 are solved to optimality, then Algorithm 1 finds an optimal solution in at most \(p + 1\) iterations.

Proof. Let \((\beta^*, z^*(t))\) be an optimal solution of (22), and let \(t_k\) and \(t_{k+1}\) be two successive \(t\) values generated by the algorithm. Since \(t_k \geq t_{k+1}\) (a direct consequence of equation (24)), we find that \(g(1'z^*(t_k)) \geq g(1'z^*(t_{k+1}))\). Moreover, if \(g(1'z^*(t_k)) = g(1'z^*(t_{k+1}))\), then necessarily \(\|y - X\beta^*(t_k)\|_2^2 = \|y - X\beta^*(t_{k+1})\|_2^2\),

\[
d(t_{k+1}) = \|y - X\beta^*(t_{k+1})\|_2^2 - t_{k+1}g(1'z^*(t_{k+1}))
\]

\[
= \|y - X\beta^*(t_k)\|_2^2 - \frac{\|y - X\beta^*(t_k)\|_2^2}{g(1'z^*(t_k))} g(1'z^*(t_k)) = 0,
\]
and the algorithm terminates. Finally, since \( g(1'z) \in \{g(0), g(1), \ldots, g(p)\} \) can take at most \( p + 1 \) values, we find that \( g(1'z^*(t_k)) > g(1'z^*(t_{k+1})) \) in at most \( p \) iterations. Thus, if all subproblems are solved to optimality, then the algorithm finds an optimal solution in at most \( p + 1 \) iterations.

The above proof follows standard arguments in fractional combinatorial optimization literature, see similar results in Radzik (1998). More importantly, Proposition 3 provides some intuition on why Algorithm 1 performs better than using (9): 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 MIQO, 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 (22) 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.

4. Computations

In this section we report computational experiments performed on synthetic and real datasets to test the proposed approaches in best subset selection problems with respect to MSE, BIC, AICc criteria. We also compare with the approach that solves for all cardinalities as in (9).

Specifications. Computations were performed using CPLEX 12.7.1 on a computer with a 3.50GHz Intel®Xeon®E5-1620 v4 CPU and 16 GB main memory and with a single thread. All CPLEX parameters were set to their default values.

4.1. Optimization performance. We first present computations with the real datasets. In this section we focus in the performance of the methods from an optimization perspective, i.e., solution times and end gaps.
Instances. We test the proposed methods on the “Diabetes” dataset with 64 predictors and 442 observations used in Efron et al. (2004) and later in Bertsimas et al. (2016). We also use the datasets used in Miyashiro and Takano (2015a), i.e., the datasets “Housing” \((n = 506, p = 13)\), “Servo” \((n = 167, p = 19)\), “AutoMPG” \((n = 392, p = 25)\), “SolarFlare” \((n = 1066, p = 26)\), “BreastCancer” \((n = 196, p = 37)\) and “Crime” \((n = 1,993, p = 100)\) from the UCI Machine Learning Repository (Dheeru and Karra Taniskidou 2017).

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

Misoco_e: The MISOCO formulation

\[
\begin{align*}
\min & \quad t \\
\text{s.t.} & \quad \|y - X\beta\|_2^2 \leq ts, \quad (19c) - (19g),
\end{align*}
\]  

(27a)  

(27b)

with the following modification: additional variables \(\gamma \in \mathbb{R}^n\) are introduced, and the rotated cone constraint \(\|y - X\beta\|_2^2 \leq ts\) is replaced with the equivalent system

\[
\gamma = y - X\beta, \quad \gamma'\gamma \leq ts.
\]

(28)

Indeed, the original formulation (27) was not recognized as convex by the solver, while the extended formulation with the \(\gamma\) variables is.

Fractional: The fractional optimization approach with Algorithm 1.

Fractional_e: The fractional optimization approach with Algorithm 1 and the substitution (28). This substitution requires adding additional variables and performs worse than Fractional in all cases, but was implemented to ensure a fair comparison with Misoco_e.

Cardinality: The approach described in (9), where the MIQO (3) is solved for all values of \(k = 1, \ldots, p\).

Results. Table 1 reports for each instance, method and criterion, the solution time (in seconds) required to solve problem (10) to optimality, or the optimality gap proven when a
time limit of one hour is reached. For the method \textbf{Cardinality}, a time limit of one hour was allocated for solving each problem (3), thus the total time reported might exceed one hour and can be as large as $p$ hours.

The optimality gaps are computed as follows: for methods \textbf{Fractional} and \textbf{Fractional.e}, the optimality gap is given by (26); for \textbf{Misoco.e}, the optimality gap just corresponds to the gap reported by the solver; and for \textbf{Cardinality}, we report the worst optimality gap among all problems (9). Note that while the gaps of \textbf{Misoco.e}, \textbf{Fractional} and \textbf{Fractional.e} all correspond to the gap with respect to the optimal solution of problem (10), the gap of \textbf{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 is not directly comparable with the other optimality gaps.

We see that the performance of \textbf{Misoco.e} is in general poor, resulting in many cases in numerical errors during branch-and-bound or very large optimality gaps of 100%; 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. In contrast, \textbf{Fractional.e} and \textbf{Fractional}, which are tackled by (partially) solving a sequence of MIQO optimization problems, perform better and are unaffected by numerical difficulties. In particular \textbf{Fractional} has the best performance among all the approaches tested, resulting in smaller computational times and end gaps.

We see that that \textbf{Cardinality} performs worse than \textbf{Fractional} in all cases, and its performance degrades as the number of predictors increases. Surprisingly, in the “Diabetes” dataset, the computational times required to solve all cardinality constrained problems (3) is much more than $p$ times the time used by Algorithm 1. As a consequence, even if the \textbf{Cardinality} method were fully parallelized, method \textbf{Fractional} on a single thread would result in better computational times.

In particular, we note that the cardinality constrained problem (3) is solved quickly for very small or very large values of $k$, perhaps due to a small number of feasible solutions ($p$); however, problem (3) is challenging when $k \approx p/2$. Thus, when using the \textbf{Cardinality}
method, these challenging problems with $k \approx p/2$ still need to be solved, resulting in large computational times. In contrast, Algorithm 1 can be interpreted as a method that quickly
probes different values of $k$ and quickly narrows it down to a small range containing the optimal value of $k$; if this optimal value is far from $p/2$, Algorithm 1 is able to prove optimality without allocating computational effort to explore solutions with support close to $p/2$.

In summary, we conclude from our experiments that, by adopting a fractional optimization perspective and using Algorithm 1, feature selection problems with information criteria can be solved substantially faster than existing approaches.

4.2. Statistical performance. In this section we compare the statistical performance of feature selection methods with different information criteria; we also test the performance of solving the cardinality constrained problem (3) and using cross-validation to select the right parameter $k$, as suggested in Bertsimas et al. (2016).

Instances. We generate synthetic datasets 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 $\rho$ and a Signal-to-Noise parameter $\nu$, the instances are generated as follows:

(i) The “true” regression coefficients $\beta^0$ have their first $k_0$ components equal to 1, and the remaining equal to 0.

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

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

In our computations, we use $n = 1,000$, $p = 100$, $\rho = 0.35$ and $\nu \in \{0.05, 0.09, 0.14, 0.25, 0.42, 0.71, 1.22, 2.07\}$; the same parameters are used in Hastie et al. (2017).

Methods. We compare the performance of the following methods.

Cross-validation: The data is 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.

We use Algorithm 1 to compute the estimators corresponding MSE, BIC and AICc. Similar to Hastie et al. (2017), we set a time limit of three minutes for solving each cardinality constrained problem (3); we also allocate a time limit of three minutes to methods with respect to information criteria. Thus, using these settings, the cross-validation approach is up to $2k_0$ times *more expensive* than the others, either in terms of time or computational resources (in practice the parameter $k_0$ is rarely known, and cross-validation might require solving $p$ problems). If there were unlimited computational resources for parallelization, then all methods would require the same amount of time.

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

(i) the *relative test error* given by

$$
\frac{\mathbb{E} \left( y_0 - x_0' \hat{\beta} \right)^2}{\sigma^2} = \frac{(\hat{\beta} - \beta^0)' \Sigma (\hat{\beta} - \beta^0) + \sigma^2}{\sigma^2},
$$

where $x_0 \in \mathbb{R}^p$ denotes a test predictor drawn from $\mathcal{N}_p(\mathbf{0}, \Sigma)$, $y_0$ its associated response drawn from $\mathcal{N}_p(x_0' \beta^0, \sigma^2)$, and $\hat{\beta}$ is an estimator obtained from a given regression procedure;

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

(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).

**Results.** We generated, for each combination of $\nu \in \{0.05, 0.09, 0.14, 0.25, 0.42, 0.71, 1.22, 2.07, 3.52, 6.00\}$ and $k_0 \in \{5, 10, 25\}$, 10 instances with identical parameters and report the averages across all replications. Specifically, Table 2 reports for each value of $k_0$ the average total time required to compute the estimators (the averages are also taken across all SNRs); and Figures 1 and 2 show, for each value of $k_0$ and $\nu$ and each method, the statistical performance in terms of relative test error and support recovery, respectively.
Table 2. Average computational time of each method in synthetic instances with $n = 1,000$, $p = 100$ and $\rho = 0.35$.

<table>
<thead>
<tr>
<th>Setting</th>
<th>MSE</th>
<th>BIC</th>
<th>AICc</th>
<th>Cross validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k_0 = 5$</td>
<td>3.0 min</td>
<td>3.0 min</td>
<td>3.0 min</td>
<td>8.2 min</td>
</tr>
<tr>
<td>$k_0 = 10$</td>
<td>3.0 min</td>
<td>3.0 min</td>
<td>3.0 min</td>
<td>33.0 min</td>
</tr>
<tr>
<td>$k_0 = 25$</td>
<td>3.0 min</td>
<td>2.8 min</td>
<td>3.0 min</td>
<td>128.9 min</td>
</tr>
</tbody>
</table>

From Table 2, we observe that the time limit of three minutes is reached for MSE, BIC and AICc methods in most of the instances. When using cross-validation, instances with low values of $k$ are often solved to optimality quickly (note that complete enumeration is feasible for $k$ small), but more difficult instances with larger values of $k$ are not solved to optimality within the time limit. In all cases, cross-validation uses more time than the methods based on information criteria, and the difference is significant for large values of $k_0$.

Moreover, the computational burden of cross-validation would further increase if the search was not limited to values of $k \leq 2k_0$ (which requires some prior knowledge of the value of $k_0$) but all values $k \leq p$ had to be considered. Therefore, we see that feature selection with information criteria can be a cheap alternative to cross-validation, especially for casual users with limited access to computational resources for parallelization.

We observe from Figure 1 that AICc dominates MSE and that BIC dominates cross-validation in terms of prediction accuracy. Moreover, while 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 cross-validation depends on those parameters. In particular, both BIC and cross-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. In contrast, AICc performs better for denser models and for medium SNR values. We see that the performance of cross-validation is especially poor for $k_0 = 25$, being outperformed by all other methods for most SNR values, i.e., for $0.14 \leq \nu \leq 2.07$. We attribute, in part, the superior performance of information criteria approaches such as BIC to the lack of cross-validation, which requires holding out a portion of the data for validation purposes.
Figure 1. Relative test error as a function of the SNR in synthetic instances with \( n = 1,000, p = 100 \) and \( \rho = 0.35 \).
From Figure 2 we see that BIC and cross-validation achieve their 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 coefficients $\beta^0$. As the SNR increases, the number
of predictor variables chosen by BIC and cross-validation gradually increases until achieving an almost exact recovery of the true support of $\beta^0$. In general, BIC tends to select a slightly larger number of correct features and a smaller number of incorrect ones, which explains its better than cross-validation prediction performance across all SNRs. In contrast, MSE and AICc fail to recover the true support for $\nu \leq 6$; in general, MSE selects a larger number of “incorrect” predictors, which explains its worse than AICc prediction performance. Nonetheless, for medium values of the SNR value, AICc chooses a larger number of true predictors than BIC or cross-validation with a modest amount of incorrect ones, leading to better prediction performance.

In summary, we see that using a fractional optimization approach to solve problems with respect to information criteria can match and even consistently outperform (e.g., BIC) traditional cross-validation approaches in terms of statistical performance, despite requiring only a small fraction of the computational effort.

5. Conclusion

We present an MIFO framework to 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 propose to tackle the resulting optimization problems by solving a sequence of MIQO problems. The proposed approach is orders of magnitude faster than existing approaches in the literature. Moreover, our computations indicate that optimizing with respect to information criteria often results in better linear regression models and furthermore, is much more efficient than traditional best subset selection approaches which rely on cross-validation. Due to the ubiquity of the information criteria in feature selection problems, the proposed methodologies may be broadly applicable in contexts other than linear regression.

Acknowledgements

This paper is based upon work supported by the National Science Foundation under Grant No. 1818700.
REFERENCES


