ANZIAM J. 61(2019), 64–75
doi:10.1017/S1446181118000275

A NEW APPROACH TO SELECT THE BEST SUBSET OF PREDICTORS IN LINEAR REGRESSION MODELLING: BI-OBJECTIVE MIXED INTEGER LINEAR PROGRAMMING

HADI CHARKHGARD ✉[1] and ALI ESHRAGH[2]

(Received 8 January, 2018; accepted 10 October, 2018; first published online 11 January 2019)

Abstract

We study the problem of choosing the best subset of p features in linear regression, given n observations. This problem naturally contains two objective functions including minimizing the amount of bias and minimizing the number of predictors. The existing approaches transform the problem into a single-objective optimization problem. We explain the main weaknesses of existing approaches and, to overcome their drawbacks, we propose a bi-objective mixed integer linear programming approach. A computational study shows the efficacy of the proposed approach.

2010 *Mathematics subject classification*: primary 62J05; secondary 90B50, 90C11.

Keywords and phrases: linear regression, best subset selection, bi-objective mixed integer linear programming.

1. Introduction

The availability of cheap computing power and significant algorithmic advances in optimization have caused a resurgence of interest in solving classical problems in different fields of study using modern optimization techniques. The focus of this study is on one of the classical problems in statistics, the so-called *best subset selection problem* (BSSP), that is, finding the best subset of p predictors in linear regression, given n observations.

Linear regression models should have two important characteristics in practice including *prediction accuracy* and *interpretability* [20]. The traditional approach of constructing regression models is to minimize the sum of the squared residuals. It is evident that models obtained in this approach have low biases. However, their prediction accuracy can be low due to their large variances. Furthermore, models

[1] Department of Industrial and Management Systems Engineering, University of South Florida, Tampa, FL 33620, USA; e-mail: hcharkhgard@usf.edu.

[2] School of Mathematical and Physical Sciences, University of Newcastle, New South Wales 2308, Australia; e-mail: ali.eshragh@newcastle.edu.au.

constructed by this approach may contain a large number of predictors and so data analysts struggle in interpreting them.

In general, reducing the number of predictors in a regression model can improve not only the interpretability but also, sometimes, the prediction accuracy by reducing the variance [20]. Hence, there is often a trade-off between the amount of bias and the practical characteristics of a regression model. In other words, finding a desirable regression model is naturally a bi-objective optimization problem that minimizes the amount of bias and the number of predictors simultaneously.

To the best of our knowledge, there has been no study on obtaining a desirable regression model using a bi-objective optimization approach. This may be due to the fact that bi-objective optimization problems are usually computationally intensive, much more than single-objective optimization problems. However, recent algorithmic and theoretical advances in bi-objective optimization (in particular, bi-objective mixed integer linear programming) have now made these problems computationally *tractable* in practice. More precisely, although bi-objective optimization problems are NP-hard [15], under some mild conditions, we are now able to solve them reasonably fast in practice. We believe that this is the first work to construct a regression model utilizing a bi-objective optimization approach.

The structure of the paper is organized as follows. In Section 2, the main concepts in bi-objective mixed integer linear programming are explained. In Section 3, the drawbacks of existing (single-objective) optimization techniques for BSSP are presented. In Section 4, the proposed bi-objective mixed integer linear programming formulation is introduced. In Section 5, the computational results are reported. Finally, in Section 6, some concluding remarks are provided.

2. Preliminaries

A *bi-objective mixed integer linear program* (BOMILP) can be stated as follows:

$$\min_{(\boldsymbol{x}_1,\boldsymbol{x}_2)\in\mathcal{X}} \{z_1(\boldsymbol{x}_1,\boldsymbol{x}_2), z_2(\boldsymbol{x}_1,\boldsymbol{x}_2)\},$$

where $\mathcal{X} = \{(\boldsymbol{x}_1,\boldsymbol{x}_2) \in \mathbb{Z}^{n_1}_{\geq} \times \mathbb{R}^{n_2}_{\geq} \mid A_1\boldsymbol{x}_1 + A_2\boldsymbol{x}_2 \leq \boldsymbol{b}\}$ represents the *feasible set in the decision space*, $\mathbb{Z}^{n_1}_{\geq} = \{\boldsymbol{s} \in \mathbb{Z}^{n_1} \mid \boldsymbol{s} \geq \boldsymbol{0}\}$, $\mathbb{R}^{n_2}_{\geq} = \{\boldsymbol{s} \in \mathbb{R}^{n_2} \mid \boldsymbol{s} \geq \boldsymbol{0}\}$, $A_1 \in \mathbb{R}^{m\times n_1}$, $A_2 \in \mathbb{R}^{m\times n_2}$, and $\boldsymbol{b} \in \mathbb{R}^m$. It is assumed that $\mathcal{X}$ is *bounded* and $z_i(\boldsymbol{x}_1,\boldsymbol{x}_2) = \boldsymbol{c}_{i,1}^{\intercal}\boldsymbol{x}_1 + \boldsymbol{c}_{i,2}^{\intercal}\boldsymbol{x}_2$, where $\boldsymbol{c}_{i,1} \in \mathbb{R}^{n_1}$ and $\boldsymbol{c}_{i,2} \in \mathbb{R}^{n_2}$, for $i = 1, 2$, represents a linear objective function. The image $\mathcal{Z}$ of $\mathcal{X}$ under the vector-valued function $\boldsymbol{z} = (z_1, z_2)^{\intercal}$ represents the *feasible set in the objective/criterion space*, that is, $\mathcal{Z} = \{\boldsymbol{o} \in \mathbb{R}^2 \mid \boldsymbol{o} = \boldsymbol{z}(\boldsymbol{x}_1,\boldsymbol{x}_2)$ for all $(\boldsymbol{x}_1,\boldsymbol{x}_2) \in \mathcal{X}\}$. Note that BOMILP is called *bi-objective linear program* (BOLP) and *bi-objective integer linear program* (BOILP) for the special cases of $n_1 = 0$ and $n_2 = 0$, respectively.

Definition 2.1. A feasible solution $(\boldsymbol{x}_1,\boldsymbol{x}_2) \in \mathcal{X}$ is called *efficient* or *Pareto optimal* if there is no other $(\boldsymbol{x}'_1,\boldsymbol{x}'_2) \in \mathcal{X}$ such that $z_1(\boldsymbol{x}'_1,\boldsymbol{x}'_2) \leq z_1(\boldsymbol{x}_1,\boldsymbol{x}_2)$ and $z_2(\boldsymbol{x}'_1,\boldsymbol{x}'_2) < z_2(\boldsymbol{x}_1,\boldsymbol{x}_2)$, or $z_1(\boldsymbol{x}'_1,\boldsymbol{x}'_2) < z_1(\boldsymbol{x}_1,\boldsymbol{x}_2)$ and $z_2(\boldsymbol{x}'_1,\boldsymbol{x}'_2) \leq z_2(\boldsymbol{x}_1,\boldsymbol{x}_2)$. If $(\boldsymbol{x}_1,\boldsymbol{x}_2)$ is efficient, then $\boldsymbol{z}(\boldsymbol{x}_1,\boldsymbol{x}_2)$ is called a *nondominated point*. The set of all efficient solutions is denoted by $\mathcal{X}_E$. The set of all nondominated points $\boldsymbol{z}(\boldsymbol{x}_1,\boldsymbol{x}_2)$ for $(\boldsymbol{x}_1,\boldsymbol{x}_2) \in \mathcal{X}_E$ is denoted by $\mathcal{Z}_N$ and referred to as the *nondominated frontier*.

https://doi.org/10.1017/S1446181118000275 Published online by Cambridge University Press

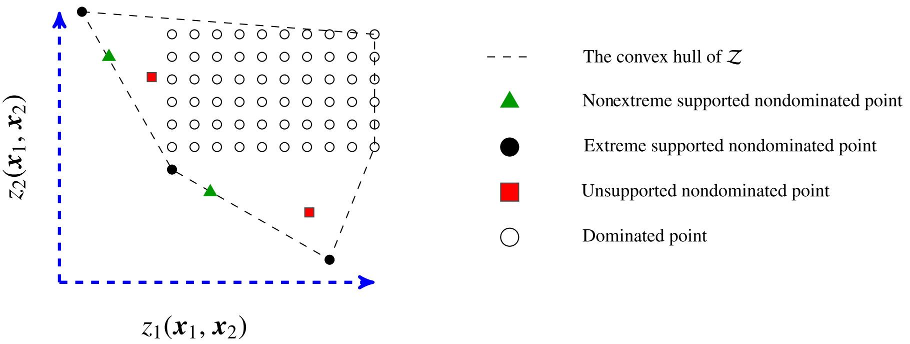

$z_2(\boldsymbol{x}_1, \boldsymbol{x}_2)$
$z_1(\boldsymbol{x}_1, \boldsymbol{x}_2)$
The convex hull of $\mathcal{Z}$
Nonextreme supported nondominated point
Extreme supported nondominated point
Unsupported nondominated point
Dominated point

FIGURE 1. An illustration of different types of (feasible) points in the criterion space.

DEFINITION 2.2. If there exists a vector $(\lambda_1, \lambda_2)^\intercal \in \mathbb{R}^2_> = \{\boldsymbol{s} \in \mathbb{R}^2 \mid \boldsymbol{s} > \boldsymbol{0}\}$ such that

$$(\boldsymbol{x}_1^*, \boldsymbol{x}_2^*) \in \arg \min_{(\boldsymbol{x}_1, \boldsymbol{x}_2) \in \mathcal{X}} \lambda_1 z_1(\boldsymbol{x}_1, \boldsymbol{x}_2) + \lambda_2 z_2(\boldsymbol{x}_1, \boldsymbol{x}_2),$$

then $(\boldsymbol{x}_1^*, \boldsymbol{x}_2^*)$ is called a *supported efficient solution* and $\boldsymbol{z}(\boldsymbol{x}_1^*, \boldsymbol{x}_2^*)$ is called a *supported nondominated point*.

DEFINITION 2.3. Let $\mathcal{Z}^e$ be the set of extreme points of the convex hull of $\mathcal{Z}$, that is, the smallest convex set containing the set $\mathcal{Z}$. A point $\boldsymbol{z}(\boldsymbol{x}_1, \boldsymbol{x}_2) \in \mathcal{Z}$ is called an *extreme supported nondominated point* if $\boldsymbol{z}(\boldsymbol{x}_1, \boldsymbol{x}_2)$ is a supported nondominated point and $\boldsymbol{z}(\boldsymbol{x}_1, \boldsymbol{x}_2) \in \mathcal{Z}^e$.

In summary, based on Definition 2.1, the elements of $\mathcal{Z}$ can be partitioned into dominated and nondominated points. Furthermore, based on Definitions 2.2 and 2.3, the points can be partitioned into unsupported, nonextreme supported, and extreme supported nondominated points. Overall, bi-objective optimization problems are concerned with finding all elements of $\mathcal{Z}_N$, that is, all nondominated points, including supported and unsupported nondominated points. An illustration of the set $\mathcal{Z}$ and its corresponding categories is shown in Figure 1.

It is well known that in a BOLP, both the set of efficient solutions $\mathcal{X}_E$ and the set of nondominated points $\mathcal{Z}_N$ are supported and connected [17]. Consequently, to describe all nondominated points in a BOLP, it suffices to find all extreme supported nondominated points. A typical illustration of the nondominated frontier of a BOLP is displayed in Figure 2(a), where (solid) circles are extreme supported nondominated points.

Since we assume that $\mathcal{X}$ is bounded, the set of nondominated points of a BOILP is finite. However, due to the existence of unsupported nondominated points in a BOILP, finding all nondominated points is more challenging than in a BOLP. A typical nondominated frontier of a BOILP is shown in Figure 2(b), where the rectangles are unsupported nondominated points.

Finding all nondominated points of a BOMILP is even more challenging. Nonetheless, if at most one of the objective functions of a BOMILP contains

https://doi.org/10.1017/S1446181118000275 Published online by Cambridge University Press

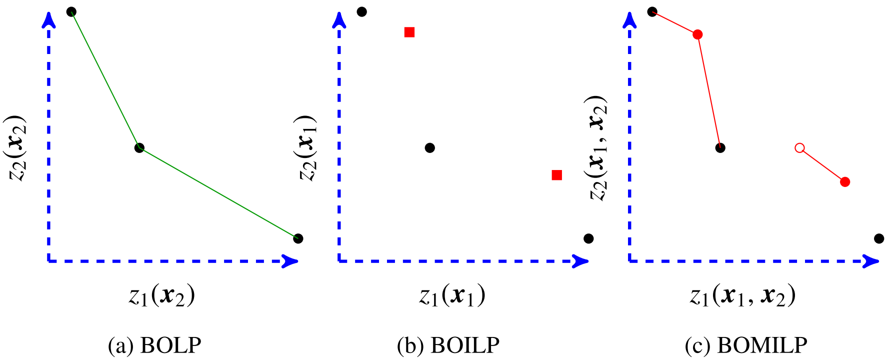


FIGURE 2. An illustration of the nondominated frontier.

continuous decision variables, then the set of nondominated points is finite and BOILP solution approaches can be utilized to solve it [19]. However, in all other cases in which more than one objective function contains continuous decision variables, the nondominated frontier of a BOMILP may contain connected parts as well as supported and unsupported nondominated points. Therefore, in these cases, the set of nondominated points may not be finite and BOILP algorithms cannot be applied to solve them any more. A typical nondominated frontier of a BOMILP is illustrated in Figure 2(c), where even half-open (or open) line segments may exist in the nondominated frontier. Interested readers are referred to the literature [3, 4, 10, 11] for further discussions on the properties of BOILPs and BOMILPs and algorithms to solve them.

3. Bi-objective versus single objective optimization models for BSSP

As discussed in Section 1, BSSP is naturally a bi-objective optimization problem (BOOP), which can be stated as $\min_{\hat{\boldsymbol{\beta}}\in\mathcal{F}} \{z_1(\hat{\boldsymbol{\beta}}), z_2(\hat{\boldsymbol{\beta}})\}$, where $\mathcal{F}$ is the feasible set of parameter estimator vectors $\hat{\boldsymbol{\beta}}$, $z_1(\hat{\boldsymbol{\beta}})$ is the total bias, and $z_2(\hat{\boldsymbol{\beta}})$ is the number of predictors. Since there is no bi-objective optimization technique in the literature of BSSP, the following two approaches have widely been used to convert BOOP to a single-objective optimization problem.

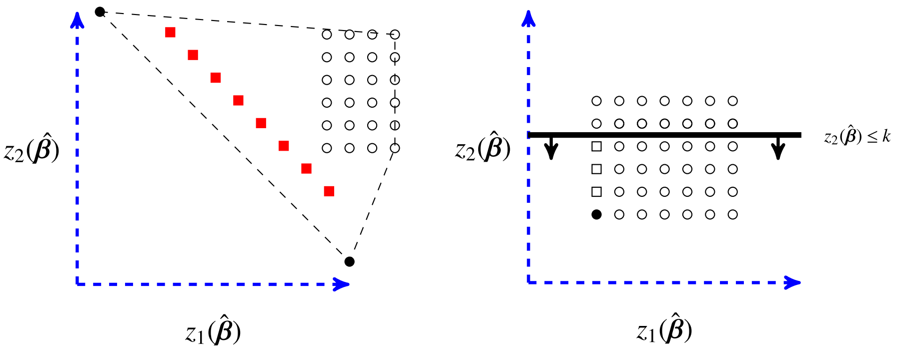
(i) *The weighted sum approach:* Given some $\lambda > 0$, BOOP has been reformulated as

$$\min_{\hat{\boldsymbol{\beta}}\in\mathcal{F}} z_1(\hat{\boldsymbol{\beta}}) + \lambda z_2(\hat{\boldsymbol{\beta}}).$$

(ii) *The goal programming approach:* Given some $k \in \mathbb{Z}_{\geq}$, BOOP has been reformulated as

$$\min_{\hat{\boldsymbol{\beta}}\in\mathcal{F}:z_2(\hat{\boldsymbol{\beta}})\leq k} z_1(\hat{\boldsymbol{\beta}}).$$

https://doi.org/10.1017/S1446181118000275 Published online by Cambridge University Press



(a) The weighted sum approach fails to find rectangles

(b) The goal programming approach may incorrectly find rectangles

FIGURE 3. The set of feasible points in the criterion space.

For further details, interested readers are referred to [2, 5, 7, 12, 16, 22] for the weighted sum approach and [1, 13, 14] for the goal programming approach. Although those two optimization problems (i) and (ii) can be solved significantly faster than a bi-objective optimization problem, their drawbacks are explained and illustrated here.

Suppose that for each $\hat{\boldsymbol{\beta}} \in \mathcal{F}$, the corresponding point $(z_1(\hat{\boldsymbol{\beta}}), z_2(\hat{\boldsymbol{\beta}}))^\intercal$ is plotted into the criterion space. Figures 3(a) and 3(b) show two typical plots of such pairs for all $\hat{\boldsymbol{\beta}} \in \mathcal{F}$. In these two figures, all filled circles and rectangles are nondominated points of the problem and unfilled rectangles and circles are dominated points. In Figure 3(a), the region defined by the dashed lines is the convex hull of all feasible points. In this case, it is impossible that the weighted sum approach finds the filled rectangles for any arbitrary weight, as all filled rectangles are unsupported nondominated points (that is, they are interior points of the convex hull). So, this illustrates that there may exist many nondominated points, but the weighted sum approach can fail to find most of them for any arbitrary weight. Figure 3(b) is helpful for understanding the main drawback of the goal programming approach. It is obvious that depending on the value of k, the goal programming approach may find one of the unfilled rectangles which are dominated points. So, the main drawback of the goal programming approach is that it may even fail to find a nondominated point.

The main contribution of our research presented here is to overcome both of these disadvantages by utilizing bi-objective optimization techniques. We note that in the literature of BSSP, $z_1(\hat{\boldsymbol{\beta}})$ is mainly defined as the sum of squared residuals. The reason lies in the fact that the sum of squared residuals is a smooth (convex) function. Hence, it is easy to minimize such a function (over a convex feasible set), which yields a unique solution. However, to be able to exploit existing bi-objective mixed integer linear programming solvers, we use the sum of absolute residuals for $z_1(\hat{\boldsymbol{\beta}})$. Such solvers transform a bi-objective optimization problem into a sequence of single-objective integer linear programs in which each one can often be solved efficiently,

https://doi.org/10.1017/S1446181118000275 Published online by Cambridge University Press

in practice, by using commercial solvers such as CPLEX or Gurobi. Note that if we used the sum of squared residuals, then the bi-objective optimization problem would become nonlinear. So, a sequence of single-objective integer nonlinear programs would be solved. However, commercial integer nonlinear programming solvers are not as mature and fast as integer linear programming solvers. So, it is the main reason that the focus of this study is on the sum of absolute residuals. Another advantage of using the latter is that it is a superior model in the presence of outliers [8]. More precisely, while the sum of squared residuals gives more weights to large residuals, the sum of absolute residuals gives equal weights to all residuals, which results in a robust estimation. We conclude this section with the following two remarks.

Remark 3.1. If we incorporate additional linear constraints on the vector of parameter estimators of the regression model $\hat{\boldsymbol{\beta}}$, it is more likely that the goal programming approach fails to find a nondominated point.

Remark 3.2. Unlike the weighted sum and goal programming approaches, where new parameters λ and k, respectively, should be employed and tuned by the user, the bi-objective optimization approach does not need any extra parameter.

4. A bi-objective mixed integer linear programming formulation

Let $X = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p] \in \mathbb{R}^{n\times p}$ be the model matrix (it is assumed that $\boldsymbol{x}_1 = \mathbf{1}$), $\boldsymbol{\beta} \in \mathbb{R}^{p\times 1}$ be the vector of regression coefficients, and $\boldsymbol{y} \in \mathbb{R}^{n\times 1}$ be the response vector. It is assumed that $\boldsymbol{\beta}$ is unknown and should be estimated. Let $\hat{\boldsymbol{\beta}} \in \mathbb{R}^{p\times 1}$ denote an estimate for $\boldsymbol{\beta}$. To solve BSSP for this set of data, we construct the following BOMILP and denote it by BSSP-BOMILP:

$$
\begin{aligned}
&\min \Big\{\sum_{i=1}^{n} \gamma_i, \sum_{j=1}^{p} r_j\Big\} && \\
\text{such that}\quad & r_j l_j \le \hat{\beta}_j \le r_j u_j && \text{for } j = 1, \ldots, p, && (4.1)\\
& y_i - \sum_{j=1}^{p} x_{ij}\hat{\beta}_j \le \gamma_i && \text{for } i = 1, \ldots, n, && (4.2)\\
& \sum_{j=1}^{p} x_{ij}\hat{\beta}_j - y_i \le \gamma_i && \text{for } i = 1, \ldots, n, && (4.3)\\
& r_j \in \{0, 1\}, \quad \hat{\beta}_j \in \mathbb{R} && \text{for } j = 1, \ldots, p, &&\\
& \gamma_i \ge 0 && \text{for } i = 1, \ldots, n, &&
\end{aligned}
$$

where $l_j \in \mathbb{R}$ and $u_j \in \mathbb{R}$ are, respectively, a lower bound and an upper bound (known) for $\hat{\beta}_j$, γ_i is a nonnegative continuous variable that takes the value of $|y_i - \sum_{j=1}^{p} x_{ij}\hat{\beta}_j|$ in any efficient solution, and r_j is a binary decision variable that takes the value of one if $\hat{\beta}_j \neq 0$, implying that the predictor j is active. By these definitions, for any efficient solution, the first objective function, $z_1(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{n} \gamma_i$, takes the value of the

https://doi.org/10.1017/S1446181118000275 Published online by Cambridge University Press

sum of absolute residuals and the second objective function, $z_2(\hat{\boldsymbol{\beta}}) = \sum_{j=1}^{p} r_j$, computes the number of predictors. Constraint (4.1) ensures that if $\hat{\beta}_j \neq 0$, then $r_j = 1$ for $j = 1, \dots, p$. Constraints (4.2) and (4.3) guarantee that

$$\left| y_i - \sum_{j=1}^{p} x_{ij}\hat{\beta}_j \right| \leq \gamma_i \quad \text{for } i = 1, \dots, n.$$

Note that since we minimize the first objective function, we have $|y_i - \sum_{j=1}^{p} x_{ij}\hat{\beta}_j| = \gamma_i$ for $i = 1, \dots, n$ in an efficient solution.

Remark 4.1. The BSSP-BOMILP can handle additional linear constraints and variables. Furthermore, by choosing tight bounds in Constraint (4.1), we can speed up the solution time of BSSP-BOMILP. Hence, we should try to choose l_j/u_j as large/small as possible.

Remark 4.2. Since only one of the objective functions in BSSP-BOMILP contains continuous variables, based on our discussion in Section 2, the set of nondominated points of BSSP-BOMILP is finite. More precisely, the nondominated frontier of BSSP-BOMILP can have at most $p + 1$ nondominated points, as $\sum_{j=1}^{p} r_j \in \{0, 1, \dots, p\}$. So, we can use BOILP solvers, such as the ϵ-constraint method or the balanced box method, to solve BSSP-BOMILP [3, 6].

Remark 4.3. The solution $(\boldsymbol{\gamma}^B, \boldsymbol{r}^B, \hat{\boldsymbol{\beta}}^B) = (|\boldsymbol{y}|, \boldsymbol{0}, \boldsymbol{0})$ is a trivial efficient solution of BSSP-BOMILP, which attains the minimum possible value for the second objective function. Accordingly, the point $(\sum_{i=1}^{n} \gamma_i^B, \sum_{j=1}^{p} r_j^B) = (\sum_{i=1}^{n} |y_i|, 0)$ is a trivial nondominated point of BSSP-BOMILP, where there is no parameter selected in the estimated regression model. Hence, we exclude this trivial nondominated point by adding the constraint $\sum_{j=1}^{p} r_j \geq 1$ to BSSP-BOMILP.

4.1. Bounds for the regression coefficients Establishing tight bounds on parameter estimators $\hat{\beta}_j$ is important in practice, as it can result in generating a strong formulation [21]. If the bounds are not tight enough, the solution time may increase significantly. Furthermore, some numerical issues may arise. For example, some nondominated points may not be found or even some infeasible points may be reported as nondominated points incorrectly. Consequently, in this section, we develop a data-driven approach to find good bounds l_j and u_j for $j = 1, \dots, p$ such that $l_j \leq \hat{\beta}_j \leq u_j$, in the lack of any additional information. For this purpose, we first present a proposition.

Proposition 4.4. *Let m be the median of response observations $y_1, \dots, y_n$. If $(\boldsymbol{\gamma}^*, \boldsymbol{r}^*, \hat{\boldsymbol{\beta}}^*)$ is an efficient solution of* BSSP-BOMILP, *then $\sum_{i=1}^{n} \gamma_i^* \leq \sum_{i=1}^{n} |y_i - m|$.*

Proof. Let us consider the feasible solution $(\boldsymbol{\gamma}, \boldsymbol{r}, \hat{\boldsymbol{\beta}})$, where $r_1 = 1$, $\beta_1 = m$, $r_j = \beta_j = 0$ for $j = 2, \dots, p$, and $\gamma_i = |y_i - \sum_{j=1}^{p} x_{ij}\hat{\beta}_j|$ for $i = 1, \dots, n$. So, we have $\gamma_i = |y_i - m|$ for $i = 1, \dots, n$, because $x_{i1} = 1$ for $i = 1, \dots, n$ in BSSP-BOMILP. Since, by Remark 4.3, $\sum_{j=1}^{p} r_j^* \geq 1 = \sum_{j=1}^{p} r_j$, we must have $\sum_{i=1}^{n} \gamma_i^* \leq \sum_{i=1}^{n} |y_i - m| = \sum_{i=1}^{n} \gamma_i$, to keep $(\boldsymbol{\gamma}^*, \boldsymbol{r}^*, \hat{\boldsymbol{\beta}}^*)$ as an efficient solution. □

https://doi.org/10.1017/S1446181118000275 Published online by Cambridge University Press

REMARK 4.5. It is readily seen that if we replace m with any other real number, the inequality given in Proposition 4.4 still holds. However, as the minimum of $\sum_{i=1}^{n} |y_i - \hat{\beta}_1|$ is achieved at $\hat{\beta}_1 = m$ [18], Proposition 4.4 provides the best upper bound for $\sum_{i=1}^{n} \gamma_i^*$.

Motivated by Proposition 4.4, we solve the following optimization problem to find u_j for $j = 1, \ldots, p$:

$$u_j = \max\Big\{\hat{\beta}_j \;\Big|\; \sum_{i=1}^{n}\Big|y_i - \sum_{j'=1}^{p} x_{ij'}\hat{\beta}_{j'}\Big| \le \sum_{i=1}^{n} |y_i - m|,\ \hat{\boldsymbol{\beta}} \in \mathbb{R}^p\Big\}. \tag{4.4}$$

There are several ways to transform (4.4) to a linear program (for example, see the article by Dielman [9]). Here, we propose the linear programming model

$$u_j = \max\Big\{\hat{\beta}_j \;\Big|\; \sum_{i=1}^{n} \gamma_i \le \sum_{i=1}^{n} |y_i - m|, y_i - \sum_{j'=1}^{p} x_{ij'}\hat{\beta}_{j'} \le \gamma_i, \sum_{j'=1}^{p} x_{ij'}\hat{\beta}_{j'} - y_i \le \gamma_i$$
$$\text{for } i = 1, \ldots, n \text{ and } \hat{\boldsymbol{\beta}} \in \mathbb{R}^p, \boldsymbol{\gamma} \in \mathbb{R}^n_{\ge}\Big\}. \tag{4.5}$$

Note that model (4.5) is a relaxation of model (4.4), since γ_i over-calculates $|y_i - \sum_{j'=1}^{p} x_{ij'}\hat{\beta}_{j'}|$ for $i = 1, \ldots, n$. Analogously, l_j for $j = 1, \ldots, p$ can be computed by changing "max" into "min" in (4.5).

5. Computational results

We conduct a computational study to show the performance of the ϵ-constraint method on BSSP-BOMILP numerically. We use C++ to code the ϵ-constraint method. In this computational study, the algorithm uses CPLEX 12.7 as the single-objective integer programming solver. All computational experiments are carried out on a Dell PowerEdge R630 with two Intel Xeon E5-2650 2.2 GHz 12-core processors (30 MB), 128 GB RAM, and the Red Hat Enterprise Linux 6.8 operating system. We allow CPLEX to employ at most 10 threads at the same time.

We design six classes of instances, each denoted by $C(p, n)$, where $p \in \{20, 40\}$ and $n \in \{2p, 3p, 4p\}$. Based on this construction, we generate three instances for each class as follows.

- We set all $x_{i1} = 1$, and all x_{ij} with $j > 1$ are randomly drawn from the discrete uniform distribution on the interval $[-50, 50]$.
- To construct y_i for $i = 1, \ldots, n$, two steps are taken: (1) a vector $\boldsymbol{\beta}$ is generated such that two-thirds of its components are zeros and the rest of the components are randomly drawn from the uniform distribution on the interval $(0, 1)$; (2) we set $y_i = \varepsilon_i + \sum_{j=1}^{p} x_{ij}\beta_j$ (with at most one decimal place), where ε_i is randomly generated from the standard normal distribution.
- Optimal values of l_j and u_j for $j = 1, \ldots, p$ are computed by solving model (4.5).

https://doi.org/10.1017/S1446181118000275 Published online by Cambridge University Press

TABLE 1. Numerical results obtained by running the ϵ-constraint method.

Class	Instance 1		Instance 2		Instance 3		Average	
	Time (sec.)	#NDPs	Time (sec.)	#NDPs	Time (sec.)	#NDPs	Time (sec.)	#NDPs
C(20,40)	4.1	21	3.7	21	3.8	21	3.8	21.0
C(20,60)	4.6	21	5.4	21	4.3	21	4.8	21.0
C(20,80)	5.2	21	6.0	21	6.1	21	5.8	21.0
C(40,80)	264.4	41	385.5	41	290.6	41	313.5	41.0
C(40,120)	313.0	41	921.5	41	247.0	41	493.8	41.0
C(40,160)	275.6	41	327.9	41	591.0	41	398.2	41.0

In Table 1 we report the numerical results for all 18 instances. For each instance, there are two columns "Time (sec.)" and "#NDPs" showing the solution time in seconds and the number of nondominated points, respectively. All nondominated points can be found, for instances with $p = 20$ and $p = 40$, in about 5 seconds and 7 minutes on average, respectively. Note that the numbers in columns "Time (sec.)" include the computational times to find the lower and upper bounds for $\hat{\beta}_j$ through solving the linear programming model given in (4.5). Of course, CPLEX can solve each of those linear programs efficiently in a fraction of a second for the instances used in this study.

In order to demonstrate the important role of good bounds for $\hat{\beta}_j$ derived through solving the linear programming model (4.5), we solve all 18 instances by setting a wide range for the parameter estimators, say $l_j = -1000$ and $u_j = 1000$ for $j = 1, \ldots, p$. As stated in Section 4.1 and illustrated in Figure 4, due to numerical issues, a significant ratio of nondominated points (that is, around 50% to 70%) fail to be found by the model. Note that, in practice, most of the missing points belong to the part of the nondominated frontier in which the total bias is almost the same for them (see, for example, the vertical segment in Figure 5). So, one may argue that finding and listing all such points are not very appealing to the final decision. However, the main point is that employing good bounds prescribed by the data-driven approach developed in Section 4.1 helps the model to list all $p + 1$ nondominated points.

REMARK 5.1. Table 1 displays that all $p + 1$ nondominated points have been found for each instance. This implies that for each of these generated instances, the goal programming approach should also return a nondominated point for $k = 0, \ldots, p$. Of course, this is not a surprising result, since by adding more nonzero regression coefficients to a model, the optimal value of the sum of absolute residuals often decreases unless the following hold.

(i) Decision makers impose additional constraints, as stated in Remark 3.1. For instance, one may impose that if $r_1 = 1$, then we should have $r_2 + r_3 = 0$, that is, $r_2 + r_3 \leq 2(1 - r_1)$.

https://doi.org/10.1017/S1446181118000275 Published online by Cambridge University Press

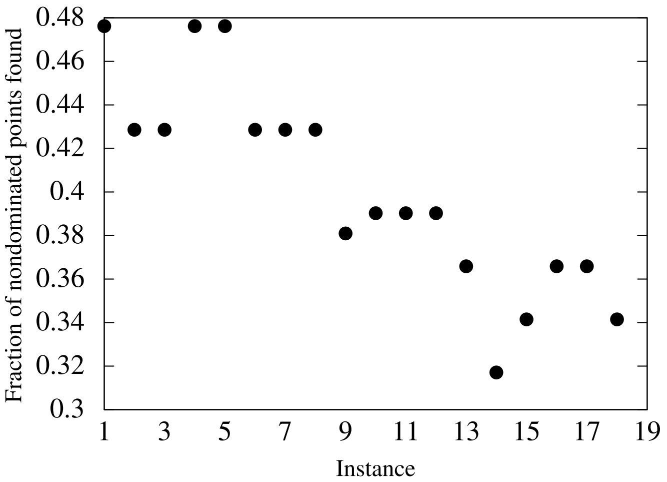


FIGURE 4. The impact of setting a wide bound for $\hat{\beta}_j$ with $l_j = -1000$ and $u_j = 1000$ for $j = 1, \ldots, p$.

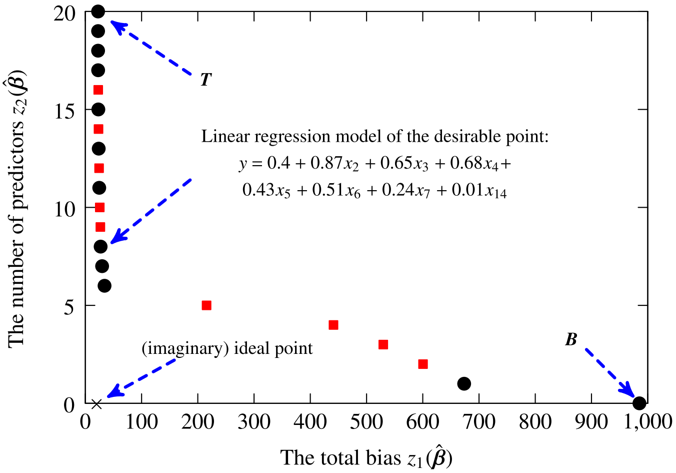


FIGURE 5. The nondominated frontier of Instance 1 from Class $C(20, 40)$.

(ii) Some columns of the model matrix X are linearly dependent.

(iii) There is an ideal regression model with the sum of absolute residuals equal to zero.

(iv) Numerical issues arise.

To highlight the drawbacks of existing approaches including the weighted sum approach and the goal programming approach, the nondominated frontier of Instance 1 from Class $C(20, 40)$ is illustrated in Figure 5. The filled rectangles and circles are unsupported and supported nondominated points, respectively. As we discussed

https://doi.org/10.1017/S1446181118000275 Published online by Cambridge University Press

previously, it is impossible to find any of the unsupported nondominated points using the weighted sum approach. Also, observe that many of the nondominated points lie on an almost vertical line. This implies that all these points are almost optimal for the goal programming approach when $k = 7, \dots, 20$.

We note that selecting a desirable nondominated point in the nondominated frontier depends on decision makers. Here, we introduce a heuristic algorithm to do so. Let $\boldsymbol{T} = (T_1, T_2)^\intercal \in \mathbb{R}^2$ and $\boldsymbol{B} = (B_1, B_2)^\intercal \in \mathbb{R}^2$ be the top and bottom end points of the nondominated frontier, respectively. One may simply choose the point that has the minimum Euclidean distance from the (imaginary) *ideal* point, that is, $(T_1, B_2)^\intercal$. Based on this algorithm, in Figure 5, the (imaginary) ideal point is $(22.6, 0)^\intercal$ and the closest nondominated point to it is $(27.2, 8)^\intercal$. The generated instance that we discuss in Figure 5 is $y = 0.42 + 0.86x_2 + 0.63x_3 + 0.68x_4 + 0.42x_5 + 0.50x_6 + 0.25x_7$ and the estimated linear regression model corresponding to the selected nondominated point is $y = 0.4 + 0.87x_2 + 0.65x_3 + 0.68x_4 + 0.43x_5 + 0.51x_6 + 0.24x_7 + 0.01x_{14}$, which are very close together.

Remark 5.2. From Figure 5, we observe that the selected nondominated point which has the minimum Euclidean distance from the (imaginary) ideal point is a “supported” nondominated point. However, this is not always the case and, in many other instances, such a point is an unsupported nondominated point. This implies that those points cannot be found by the weighted sum approach.

6. Conclusion

Minimizing the amount of bias and the number of predictors is a natural objective that is typically considered for choosing the best subset of p features in linear regression, given n observations. The existing approaches transform this problem into a single-objective optimization problem by using a weighted summation of the objectives or treating one of the objectives as a constraint. We explained the main weaknesses of the existing approaches and, to overcome their drawbacks, we proposed to directly use the bi-objective optimization techniques for solving the problem. We hope that the simplicity, versatility, and performance of our approach encourage practitioners to consider using exact bi-objective optimization methods for constructing linear regression models.

Acknowledgement

The authors are grateful to an anonymous reviewer for his/her valuable comments that helped improve the previous version of this paper.

References

[1] D. Bertsimas, A. King and R. Mazumder, “Best subset selection via a modern optimization lens”, *Ann. Statist.* **44** (2016) 813–852; doi:10.1214/15-AOS1388.

[2] P. J. Bickel, Y. Ritov and A. B. Tsybakov, “Simultaneous analysis of Lasso and Dantzig selector”, *Ann. Statist.* **37** (2009) 1705–1732; doi:10.1214/08-AOS620.

https://doi.org/10.1017/S1446181118000275 Published online by Cambridge University Press

[3] N. Boland, H. Charkhgard and M. Savelsbergh, “A criterion space search algorithm for biobjective integer programming: the balanced box method”, *INFORMS J. Comput.* **27** (2015) 735–754; doi:10.1287/ijoc.2015.0657.

[4] N. Boland, H. Charkhgard and M. Savelsbergh, “A criterion space search algorithm for biobjective mixed integer programming: the triangle splitting method”, *INFORMS J. Comput.* **27** (2015) 597–618; doi:10.1287/ijoc.2015.0646.

[5] E. J. Candés and Y. Plan, “Near-ideal model selection by l_1 minimization”, *Ann. Statist.* **37** (2009) 2145–2177; doi:10.1214/08-AOS653.

[6] V. Chankong and Y. Y. Haimes, *Multiobjective decision making: theory and methodology* (Elsevier Science, New York, 1983).

[7] S. S. Chen, D. L. Donoho and M. A. Saunders, “Atomic decomposition by basis pursuit”, *SIAM J. Sci. Comput.* **20** (1998) 33–61; doi:10.1137/S1064827596304010.

[8] T. E. Dielman, “A comparison of forecasts from least absolute value and least squares regression”, *J. Forecast.* **5** (1986) 189–195; doi:10.1080/0094965042000223680.

[9] T. E. Dielman, “Least absolute value regression: recent contributions”, *J. Stat. Comput. Simul.* **75** (2005) 263–286; doi:10.1002/for.3980050305.

[10] D. Ghosh and D. Chakraborty, “A new Pareto set generating method for multi-criteria optimization problems”, *Oper. Res. Lett.* **42** (2014) 514–521; doi:10.1016/j.orl.2014.08.011.

[11] H. W. Hamacher, C. R. Pedersen and S. Ruzika, “Finding representative systems for discrete bicriterion optimization problems”, *Oper. Res. Lett.* **35** (2007) 336–344; doi:10.1016/j.orl.2006.03.019.

[12] N. Meinshausen and P. Bühlmann, “High-dimensional graphs and variable selection with the Lasso”, *Ann. Statist.* **34** (2006) 1436–1462; doi:10.1214/009053606000000281.

[13] A. Miller, *Subset selection in regression*, 2nd edn, *Monogr. Statistics and Applied Probability* (Chapman and Hall/CRC Press, Boca Raton, FL, 2002).

[14] R. Miyashiroa and Y. Takanon, “Mixed integer second-order cone programming formulations for variable selection in linear regression”, *European J. Oper. Res.* **247** (2015) 721–731; doi:10.1214/009053606000000281.

[15] C. H. Papadimitriou and M. Yannakakis, “On the approximability of trade-offs and optimal access of web sources”, in: *Proceedings 41st Annual Symposium on Foundations of Computer Science* (IEEE, Redondo Beach, CA, 2000) 86–92; doi:10.1109/SFCS.2000.892068.

[16] Y. Ren and X. Zhang, “Subset selection for vector autoregressive processes via adaptive Lasso”, *Statist. Probab. Lett.* **80** (2010) 1705–1712; doi:10.1016/j.spl.2010.07.013.

[17] S. Sayın, “An algorithm based on facial decomposition for finding the efficient set in multiple objective linear programming”, *Oper. Res. Lett.* **19** (1996) 87–94; doi:10.1016/0167-6377(95)00046-1.

[18] N. C. Schwertman, A. J. Gilks and J. Cameron, “A simple noncalculus proof that the median minimizes the sum of the absolute deviations”, *Amer. Statist.* **44** (1990) 38–39; doi:10.1080/00031305.1990.10475690.

[19] T. Stidsen, K. A. Andersen and B. Dammann, “An algorithm based on facial decomposition for finding the efficient set in multiple objective linear programming”, *Manag. Sci.* **60** (2014) 1009–1032; doi:10.1287/mnsc.2013.1802.

[20] R. Tibshirani, “Regression shrinkage and selection via the Lasso”, *J. R. Stat. Soc. Ser.* B **58** (1996) 267–288; https://www.jstor.org/stable/i316036.

[21] L. A. Wolsey, *Integer programming*, 2nd edn (Wiley-Interscience, New York, 1998).

[22] C. Zhang and J. Huang, “The sparsity and bias of the Lasso selection in high-dimensional linear regression”, *Ann. Statist.* **36** (2008) 1567–1594; doi:10.1214/07-AOS520.

https://doi.org/10.1017/S1446181118000275 Published online by Cambridge University Press