Best subset selection via bi-objective mixed integer linear programming

Hadi Charkhgard[a,*], Ali Eshragh[b]

[a] *Department of Industrial and Management Systems Engineering, University of South Florida, Tampa, FL 33620, USA*

[b] *School of Mathematical and Physical Sceinces, University of Newcastle, NSW 2308, Australia*

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 either by combining the two objectives using some weights or by treating one of them as a constraint. We explain the main weaknesses of both of these approaches, and to overcome their drawbacks, we propose a bi-objective mixed integer linear programming approach with the property that it can handle additional constraints as well. We conduct a computational study and show that existing bi-objective optimization solvers are able to solve the problem in a reasonable time.

Keywords: 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* (BBSP), 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* (Tibshirani, 1996). The traditional approach of constructing

*Corresponding author

Email address: hcharkhgard@usf.edu (Hadi Charkhgard)

regression models is to minimize the sum of 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 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 (Tibshirani, 1996). 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 is no study on obtaining a desirable regression model using 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 a bi-objective optimization problem is NP-Hard[1] (Papadimitriou and Yannakakis, 2000), 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.

[1]It implies that no-one has found an efficient solution algorithm to solve it, yet.

2. Preliminaries

A *Bi-Objective Mixed Integer Linear Program* (BOMILP) can be stated as follows:

$$\min_{(\boldsymbol{x}_1,\boldsymbol{x}_2)\in\mathcal{X}} \{z_1(\boldsymbol{x}_1,\boldsymbol{x}_2), z_2(\boldsymbol{x}_1,\boldsymbol{x}_2)\}, \tag{1}$$

where $\mathcal{X} := \left\{(\boldsymbol{x}_1,\boldsymbol{x}_2) \in \mathbb{Z}^{n_1}_{\geq} \times \mathbb{R}^{n_2}_{\geq} : A_1\boldsymbol{x}_1 + A_2\boldsymbol{x}_2 \leq \boldsymbol{b}\right\}$ represents the *feasible set in the decision space*, $\mathbb{Z}^{n_1}_{\geq} := \{\boldsymbol{s} \in \mathbb{Z}^{n_1} : \boldsymbol{s} \geq \boldsymbol{0}\}$, $\mathbb{R}^{n_2}_{\geq} := \{\boldsymbol{s} \in \mathbb{R}^{n_2} : \boldsymbol{s} \geq \boldsymbol{0}\}$, $A_1 \in \mathbb{R}^{m\times n_1}$, $A_2 \in \mathbb{R}^{m\times n_2}$, and $\boldsymbol{b} \in \mathbb{R}^m$. It is assumed that $\mathcal{X}$ is *bounded* and $z_i(\boldsymbol{x}_1,\boldsymbol{x}_2) = \boldsymbol{c}^{\intercal}_{i,1}\boldsymbol{x}_1 + \boldsymbol{c}^{\intercal}_{i,2}\boldsymbol{x}_2$ 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 $\mathcal{Y}$ of $\mathcal{X}$ under vector-valued function $\boldsymbol{z} := (z_1, z_2)^{\intercal}$ represents the *feasible set in the objective/criterion space*, that is $\mathcal{Y} := \{\boldsymbol{o} \in \mathbb{R}^2 : \boldsymbol{o} = \boldsymbol{z}(\boldsymbol{x}_1,\boldsymbol{x}_2) \text{ for all } (\boldsymbol{x}_1,\boldsymbol{x}_2) \in \mathcal{X}\}$. 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 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 $\mathcal{Y}_N$ and referred to as the nondominated frontier.*

Definition 2. *If there exists a vector $(\lambda_1,\lambda_2)^{\intercal} \in \mathbb{R}^2_{>} := \{\boldsymbol{s} \in \mathbb{R}^2 : \boldsymbol{s} > \boldsymbol{0}\}$ 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 3. *Let $\mathcal{Y}^e$ be the set of extreme points of the convex hull of $\mathcal{Y}$, that is the smallest convex set containing the set $\mathcal{Y}$. A point $\boldsymbol{z}(\boldsymbol{x}_1,\boldsymbol{x}_2) \in \mathcal{Y}$ is called an extreme supported nondominated point, if $\boldsymbol{z}(\boldsymbol{x}_1,\boldsymbol{x}_2)$ is a supported nondominated point and $\boldsymbol{z}(\boldsymbol{x}_1,\boldsymbol{x}_2) \in \mathcal{Y}^e$.*

In summary, based on Definition 1, the elements of $\mathcal{Y}$ can be divided into two groups including dominated and nondominated points. Furthermore, based on Definitions 2 and 3, the nondominated points can be divided into unsupported nondominated points, non-extreme supported nondominated points and extreme supported nondominated points. Overall, bi-objective optimization

problems are concerned with finding all elements of $\mathcal{Y}_N$, that is all nondominated points, including supported and unsupported nondominated points. An illustration of the set $\mathcal{Y}$ and its corresponding categories are shown in Figure 1.

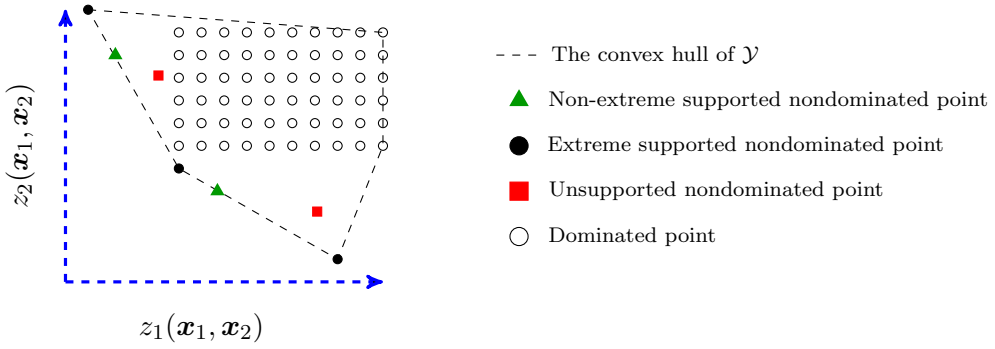
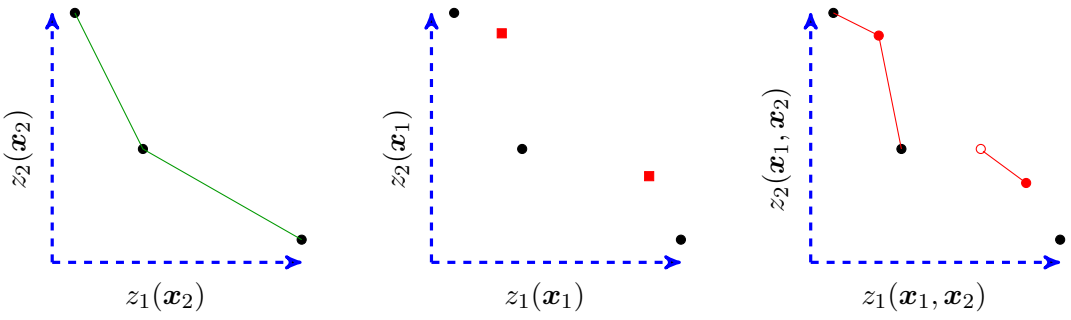


Figure 1: An illustration of different types of (feasible) points in the criterion space

It is well-known that in a BOLP, both the set of efficient solutions $\mathcal{X}_E$ and the set of nondominated points $\mathcal{Y}_N$ are supported and connected. 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 2a.

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 2b where the rectangles are unsupported nondominated points.



(a) BOLP (b) BOILP (c) BOMILP

Figure 2: An illustration of the nondominated frontier

Finding all nondominated points of a BOMILP is even more challenging. Nonetheless, if at most one of the objective functions of a BOMILP contains continuous decision variables, then the set of nondominated points is finite and BOILP solution approaches can be utilized to solve it (Stidsen et al., 2014). However, in all other cases that more than one objective function contain 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 anymore. A typical nondominated frontier of a BOMILP is illustrated in Figure 2c, where even half-open (or open) line segments may exist in the nondominated frontier. Interested readers are referred to Boland et al. (2015a,b) for further discussions on the properties of BOILPs and BOMILPs and algorithms to solve them.

3. Bi-objective vs. 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 vector $\hat{\boldsymbol{\beta}}$, $z_1(\hat{\boldsymbol{\beta}})$ 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

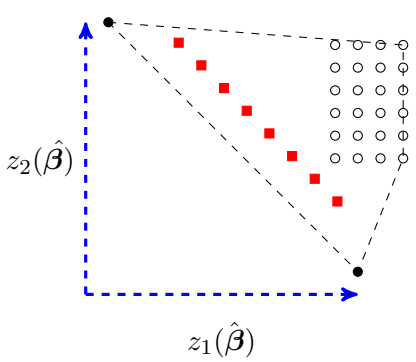
$$\min_{\hat{\boldsymbol{\beta}}\in\mathcal{F}} \; z_1(\hat{\boldsymbol{\beta}}) + \lambda z_2(\hat{\boldsymbol{\beta}}).$$

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

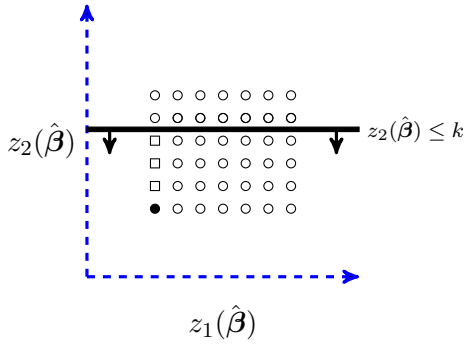
$$\min_{\hat{\boldsymbol{\beta}}\in\mathcal{F}:\; z_2(\hat{\boldsymbol{\beta}})\leq k} \; z_1(\hat{\boldsymbol{\beta}}).$$

For further details, interested readers are referred to Chen et al. (1998), Meinshausen and Bhlmann (2006), Zhang and Huang (2008), Bickel et al. (2009), Candés and Plan (2009), and Ren and Zhang (2010) for the weighted sum approach, and Miller (2002) and Bertsimas et al. (2016) for the goal

programming approach. Although, those two optimization programs (i) and (ii) can be solved significantly faster than a bi-objective optimization problem, their drawbacks are explained and illustrated here.



(a) The wighted 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

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 3a and 3b 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 3a, the region defined by the dashed lines is the convex hull of all feasible points. In this case, it is impossible that the wighted sum approach finds the filled rectangles for any arbitrary weight as all filled rectangles are unsupported nondominated points (i.e., 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 3b 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. However, to be able to exploit existing bi-objective mixed integer linear programming solvers, we use the sum of absolute residuals for $z_1(\hat{\boldsymbol{\beta}})$. We conclude this section by providing two remarks.

Remark 4. *If we incorporate additional linear constraints on the vector of parameter estimators of the regression model, $\hat{\boldsymbol{\beta}}$, it will be more likely that the goal programming approach fails to find a nondominated point.*

Remark 5. *Unlike the weighted sum and goal programming approaches that 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`:

$$\min \{\sum_{i=1}^{n} \gamma_i, \ \sum_{j=1}^{p} r_j\} \tag{2}$$

$$\text{such that: } r_j l_j \leq \hat{\beta}_j \leq r_j u_j \qquad \text{for } j = 1, \ldots, p \tag{3}$$

$$y_i - \sum_{j=1}^{p} x_{ij} \hat{\beta}_j \leq \gamma_i \qquad \text{for } i = 1, \ldots, n \tag{4}$$

$$\sum_{j=1}^{p} x_{ij} \hat{\beta}_j - y_i \leq \gamma_i \qquad \text{for } i = 1, \ldots, n \tag{5}$$

$$r_j \in \{0, 1\} \qquad \text{for } j = 1, \ldots, p \tag{6}$$

$$\gamma_i \geq 0 \qquad \text{for } i = 1, \ldots, n \tag{7}$$

$$\hat{\beta}_j \in \mathbb{R} \qquad \text{for } j = 1, \ldots, p, \tag{8}$$

where $l_j \in \mathbb{R}$ and $u_j \in \mathbb{R}$ are, respectively, a lower bound and an upper bound (known) for $\hat{\beta}_j$ for $j = 1, \ldots, p$, γ_i for $i = 1, \ldots, n$ is a non-negative continuous variable that captures the value of

$|y_i - \sum_{j=1}^{p} x_{ij}\hat{\beta}_j|$ in any efficient solution, and r_j for $j = 1, \ldots, p$ 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 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 (3) ensures that if $\hat{\beta}_j \neq 0$ then $r_j = 1$ for $j = 1, \ldots, p$. Constraints (4) and (5) guarantee that $|y_i - \sum_{j=1}^{p} x_{ij}\hat{\beta}_j| \leq \gamma_i$ for $i = 1, \ldots, n$. 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 $i = 1, \ldots, n$ in an efficient solution.

Remark 6. *The* `BSSP-BOMILP` *can handle additional linear constraints and variables. Furthermore, by choosing tight bounds in Constraint* (3)*, 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 7. *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 $p+1$ number of nondominated points as $\sum_{j=1}^{p} r_j \in \{0, 1, \ldots, p\}$. So we can use BOILP solvers such as the ϵ-constraint method or the balanced box method to solve* `BSSP-BOMILP` *(Chankong and Haimes, 1983; Boland et al., 2015a).*

Remark 8. *The solution $(\boldsymbol{\gamma}^B, \boldsymbol{r}^B, \hat{\boldsymbol{\beta}}^B) := (|\boldsymbol{y}|, \mathbf{0}, \mathbf{0})$ 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

In this section, we develop a data-driven approach to find bounds l_j and u_j for $j = 1, \ldots, p$ such that $l_j \leq \hat{\beta}_j \leq u_j$, in the lack of any additional information. To achieve this, we firstly present the following proposition.

Proposition 9. *Let m be the median of response observations $y_1, \ldots, y_n$. 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 consider the feasible solution $(\boldsymbol{\gamma}, \boldsymbol{r}, \hat{\boldsymbol{\beta}})$ where $r_1 = 1$, $\beta_1 = m$, $r_j = \beta_j = 0$ for $j = 2, \ldots, p$, $\gamma_i = |y_i - \sum_{j=1}^{p} x_{ij}\hat{\beta}_j|$ for $i = 1, \ldots, n$. So, we have $\gamma_i = |y_i - m|$ for $i = 1, \ldots, n$ because $x_{i1} = 1$ for $i = 1, \ldots, n$ in `BSSP-BOMILP`. Since by Remark 8, $\sum_{j=1}^{p} r_j^* \geq 1 = \sum_{j=1}^{p} r_j$, 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 $(\boldsymbol{\gamma}^*, \boldsymbol{r}^*, \hat{\boldsymbol{\beta}}^*)$ an efficient solution. □

Remark 10. *It is readily seen that if we replace m with any other real number, the inequality given in Proposition 9 still holds. However, as the minimum of $\sum_{i=1}^{n} |y_i - \hat{\beta}_1|$ is achieved at $\hat{\beta}_1 = m$ (Schwertman et al., 1990), Proposition 9 provides the best upper bound for $\sum_{i=1}^{n} \gamma_i^*$.*

Motivated from Proposition 9, we can solve the following optimization problem to find u_j for $j = 1, \ldots, p$:

$$u_j := \max\{\hat{\beta}_j : \sum_{i=1}^{n} |y_i - \sum_{j'=1}^{p} x_{ij'}\hat{\beta}_{j'}| \leq \sum_{i=1}^{n} |y_i - m|,\ \hat{\boldsymbol{\beta}} \in \mathbb{R}^p\}. \tag{9}$$

There are several ways to transform (9) to a linear program (e.g., see Dielman (2005)). Here, we propose the following linear programing model:

$$\begin{aligned} u_j := \max\{\hat{\beta}_j : & \sum_{i=1}^{n} \gamma_i \leq \sum_{i=1}^{n} |y_i - m|, \\ & y_i - \sum_{j'=1}^{p} x_{ij'}\hat{\beta}_{j'} \leq \gamma_i \quad \text{for } i = 1, \ldots, n, \\ & \sum_{j'=1}^{p} x_{ij'}\hat{\beta}_{j'} - y_i \leq \gamma_i \quad \text{for } i = 1, \ldots, n, \\ & \hat{\boldsymbol{\beta}} \in \mathbb{R}^p,\ \boldsymbol{\gamma} \in \mathbb{R}^n_{\geq}\}. \end{aligned} \tag{10}$$

It should be noted that (10) is a relaxation of (9) 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 (10).

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 (30MB), 128GB RAM, and the RedHat 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 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 third of its components are zeros, and the others are randomly drawn from the uniform distribution on 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 (10).

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**

Table 1 reports 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 in average, respectively.

To highlight the drawbacks of existing approaches including the weighted sum approach and the goal programming approach, the nondominated frontier of the Instance 1 from the Class $C(20, 40)$ is illustrated in Figure 4. The filled rectangles and circles are unsupported and supported nondominated points, respectively. As we discussed previously, it is impossible to find any of the unsupported nondominated points using the weighted sum approach. Also, observe that many of

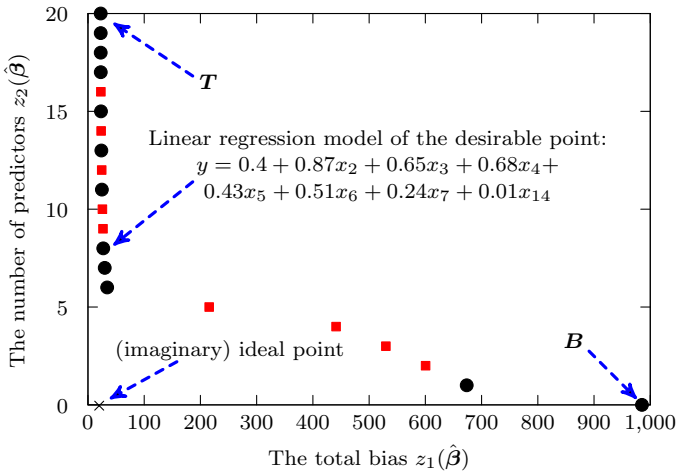


Figure 4: The nondominated frontier of the Instance 1 from the Class $C(20, 40)$

the nondominated points lies on an almost vertical line. This implies that all these points are almost optimal for the goal programming approach when $k = 7, \ldots, 20$.

We note that selecting a desirable nondominated point in the nondominated frontier depends on decision maker(s). 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 endpoints of the nondominated frontier. 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 4, 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 4 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.

6. Conclusion

We introduced a bi-objective mixed integer linear programing approach to solve BSSP and estimate a linear regression model. This new approach has two advantageous to the existing approaches: (i) it can compute all unsupported and supported nondominated points, (ii) it does not choose dominated points. 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.

References

Bertsimas, D., King, A., Mazumder, R., 2016. Best subset selection via a modern optimization lens. The Annals of Statistics 44 (2), 813–852.

Bickel, P. J., Ritov, Y., Tsybakov, A. B., 2009. Simultaneous analysis of lasso and dantzig selector. The Annals of Statistics 37 (4), 1705–1732.

Boland, N., Charkhgard, H., Savelsbergh, M., 2015a. A criterion space search algorithm for biobjective integer programming: The balanced box method. INFORMS Journal on Computing 27 (4), 735–754.

Boland, N., Charkhgard, H., Savelsbergh, M., 2015b. A criterion space search algorithm for biobjective mixed integer programming: The triangle splitting method. INFORMS Journal on Computing 27 (4), 597–618.

Candés, E. J., Plan, Y., 2009. Near-ideal model selection by l_1 minimization. The Annals of Statistics 37 (5A), 2145–2177.

Chankong, V., Haimes, Y. Y., 1983. Multiobjective Decision Making: Theory and Methodology. Elsevier Science, New York.

Chen, S. S., Donoho, D. L., Saunders, M. A., 1998. Atomic decomposition by basis pursuit. SIAM Journal on Scientific Computing 20 (1), 33–61.

Dielman, T. E., 2005. Least absolute value regression: recent contributions. Journal of Statistical Computation and Simulation 75 (4), 263–286.

Meinshausen, N., Bhlmann, P., 2006. High-dimensional graphs and variable selection with the lasso. The Annals of Statistics 34 (3), 1436–1462.

Miller, A., 2002. Subset Selection in Regression, 2nd Edition. Chapman & Hall/CRC Monographs on Statistics & Applied Probability.

Papadimitriou, C. H., Yannakakis, M., 2000. On the approximability of trade-offs and optimal access of web sources. In: 41st Annual Symposium on Foundations of Computer Science. Proceedings. pp. 86–92.

Ren, Y., Zhang, X., 2010. Subset selection for vector autoregressive processes via adaptive lasso. Statistics & Probability Letters 80 (23-24), 1705 – 1712.

Schwertman, N. C., Gilks, A. J., Cameron, J., 1990. A simple noncalculus proof that the median minimizes the sum of the absolute deviations. The American Statistician 44 (1), 38–39.

Stidsen, T., Andersen, K. A., Dammann, B., 2014. A branch and bound algorithm for a class of biobjective mixed integer programs. Management Science 60 (4), 1009–1032.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B 58, 267–288.

Zhang, C.-H., Huang, J., 2008. The sparsity and bias of the lasso selection in high-dimensional linear regression. The Annals of Statistics 36 (4), 1567–1594.