Noname manuscript No.
(will be inserted by the editor)

A mixed-integer optimization approach to an exhaustive cross-validated model selection for regression

Dennis Kreber

Received: date / Accepted: date

Acknowledgements This research was supported within the research training group 2126 Algorithmic Optimization (ALOP) founded by the German Research Foundation DFG.

Abstract We consider a linear regression model for which we assume that many of the observed regressors are irrelevant for the prediction. To avoid overfitting, we conduct a variable selection and only include the true predictors for the least square fitting. The best subset selection gained much interest in recent years for addressing this objective. For this method, a mixed-integer optimization problem is solved, which finds the optimal subset not larger than a given natural number k concerning the in-sample error. In practice, a best subset selection is computed for each k, and the ideal k is then chosen via a validation.

We argue that the notion of the best subset selection might be misaligned with the statistical intention. Instead, we propose a subset selection formulation based on the cross-validation loss function. We present a discrete optimization formulation which fits coefficients to training data and decides to in- or exclude variables to minimize the cross-validation error. Hence, we do not require a fixed sparsity bound and do not have to solve successive discrete optimization problems. Moreover, we present bounds for the regression coefficients, which allows us to construct a tighter mixed-integer formulation. Finally, we conduct a simulation study and provide evidence that the

After the author (Dennis Kreber) handed in his dissertation (Kreber 2019) on February 8th 2019 he was made aware of a preprint by Takano and Miyashiro (2019). They propose a very similar approach to variable selection, which was also developed in the dissertation. In addition to these results, a major part of the work presented here is finding novel bounds to strengthen the formulation so that the resulting program can be solved faster.

D. Kreber
Trier University
Department of Mathematics
Universitätsring 15, 54296 Trier
Germany
Tel.: +49-651-201-3452
E-mail: kreberd@uni-trier.de

novel mixed-integer formulation provides excellent predictions surpassing the results of competing state-of-the-art approaches.

Keywords Best Subset Selection · Sparse Regression · Cross-Validation · Mixed-Integer Quadratic Programming · Bilevel Optimization

1 Introduction

We are considering a linear regression model $\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta}^0 + \boldsymbol{\epsilon}$ with $\boldsymbol{y} \in \mathbb{R}^n$ being a response, $\mathbf{X} = [\, \mathbf{x}_1 \cdots \mathbf{x}_p \,] \in \mathbb{R}^{n \times p}$ being a design matrix and $\boldsymbol{\epsilon} \in \mathbb{R}^n$ being some unknown error. In many cases only some of the observed variables $\mathbf{X}_1, \ldots, \mathbf{X}_p$ are "true" predictors, that is, only a portion of the entries of $\boldsymbol{\beta}^0$ are non-zero. Applying an ordinary least squares fitting in this setting would result in overfitting and would consequently lead to bad predictions. Many applications arise from such sparse regression problems. Bühlmann (2013) gives two examples where finding the "true" variables is essential:

- *Time to flowering*: Which genes in the plant *Arabidopsis thaliana* must be modified so that the time to flowering is reduced? Here, we are interested in quantifying the effect of certain genes on the response.
- Causal effects between genes of yeast: The objective is to quantify the effect of a gene intervention by determining the causal effects between all genes.

Friedman et al. (2001) feature the problem of prostate cancer prediction, where the objective is to predict the number of antigens based on measured medical characteristics and Schoofs et al. (1997) argue that the sparse regression problem arises in engineering optimization where function approximation is a regular occurrence. As such, model selection constitutes a major part in statistical learning and is an important element of predictive statistics.

Many approaches proposed over the years build upon the idea to induce sparsity of the fitted coefficients, i.e., the number of nonzero entries of $\boldsymbol{\beta}$ ought to be limited. One of the most prominent method intended to achieve this effect is the Lasso approach (Tibshirani 1996). Here, in addition to the usual least squares regression a ℓ_1-regularization term is added:

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \; \|\mathbf{X}\boldsymbol{\beta} - \boldsymbol{y}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \tag{LASSO}$$

where $\lambda > 0$ controls the magnitude of the penalty term. The ℓ_1-regularization intends to force coefficients to zero and hence causes $\boldsymbol{\beta}$ to be sparse. While theoretical results exists attesting sparsity inducing effects and the correct choice of predictors (Bühlmann and van de Geer 2011), they require strict assumptions on the design matrix $\mathbf{X}$, which are in practice not guaranteed to hold and are computationally difficult to verify (Tillmann and Pfetsch 2014). If those assumptions do not hold, neither sparsity nor the selection of correct regressors are assured.

In recent years mixed-integer optimization has taken a bigger role in data science and machine learning. As such the best subset selection (Konno and Yamamoto 2009;

Dong et al. 2015; Bertsimas et al. 2016; Miller 1990) has produced significant research interest. The problem is formulated as the discrete optimization problem

$$\begin{aligned} \min \quad & \|\mathbf{X}\boldsymbol{\beta} - \boldsymbol{y}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2 \\ \text{s. t.} \quad & \|\boldsymbol{\beta}\|_0 \leq k, \end{aligned} \tag{BS$_{k,\lambda}$}$$

where $\|\boldsymbol{\beta}\|_0 := |\text{supp}(\boldsymbol{\beta})| := |\{i \in \{1, \ldots, p\} : \boldsymbol{\beta}_i \neq 0\}|$ and $\lambda \geq 0$. Originally, the problem does not include a regularization term as written out here. However, Bertsimas and Copenhaver (2018) argue that regularizations in context of a linear regression problem can be understood as a robustification, that is, coefficients are fitted such that they are protected against new unknown settings. Moreover, Mazumder et al. (2017) present empirical evidence that the inclusion of a regularization term has positive effects on the predictive quality. Therefore, we extend the original best subset selection by the ℓ_2-regularization but leave the option open to set λ to 0.

Unlike (LASSO) the formulation (BS$_{k,\lambda}$) guarantees that there are no more than k many nonzero entries in $\boldsymbol{\beta}$, and hence, sparsity is ensured. Bertsimas et al. (2016) and Mazumder et al. (2011) present empirical evidence that solving (BS$_{k,\lambda}$) yields predictions which are superior to those produced by Lasso. One of the earliest mixed-integer formulations for the best subset selection problem (BS$_{k,\lambda}$) was given by Konno and Yamamoto (2009). Since then, interest in solving (BS$_{k,\lambda}$) via modern discrete optimization methods grew rapidly. Dong et al. (2015) apply the perspective reformulation to the best subset selection enabling a much stronger relaxation and consequently allowing for the mixed-integer program to be solved faster. Bertsimas et al. (2016) present a mixed-integer quadratic formulation, a first-order warm start approach and an extensive study on the statistical quality of the best subset selection. They argue that discrete optimization methods can play an important role in statistics. They provide evidence that the critique of discrete optimization being computationally impractical in fields like statistics is outdated and that proper mixed-integer optimization can be highly valuable and worthwhile.

1.1 Model selection via (BS$_{k,\lambda}$)

The sparsity bound k is a fixed parameter of (BS$_{k,\lambda}$) and therefore choosing k requires additional considerations. Let us identify subsets $S \subseteq [p]$ with a regression model, that is, a subset encodes a variable selection used to predict the response $\boldsymbol{y}$. Ideally, we would like to minimize the *prediction error* PE(S) with respect to the choice of variables. In light of this, an ideal subset $\hat{S} \in [p]$ is defined as the subset which minimizes the expected ℓ_2-loss

$$\text{PE}(S) := \mathbb{E}\left(\|\mathbf{x}_S^0 \hat{\boldsymbol{\beta}}(S) - y^0\|_2^2\right), \tag{1}$$

where $\boldsymbol{x}^0$ is the random vector which encodes a new observation, y^0 is the respective response and $\hat{\boldsymbol{\beta}}(S)$ are the coefficients fitted to training data which only contains the variables indexed by S. Unfortunately, the prediction error of a subset does not correlate wit the objective function of (BS$_{k,\lambda}$). The least squares loss decreases when k increases and therefore the sparsity level $k = p$ yields the smallest ℓ_2-loss. Since

$\boldsymbol{\beta_0}$ is sparse by assumption, the ℓ_2-loss is not an appropriate metric to select the best variables. Bertsimas et al. (2016) choose the best model by solving ($\mathrm{BS}_{k,\lambda}$) for every possible $k \in \{1, \dots, p\}$ and then validating the computed variable selection on a different validation data set. The best performing model is then picked and tested on a separate test data set (see Algorithm 1). More precisely, the sampled data $\mathbf{X}$ and $\boldsymbol{y}$ is divided into three parts $\mathbf{X}^{(1)} \in \mathbb{R}^{n_1 \times p}$, $\mathbf{X}^{(2)} \in \mathbb{R}^{n_2 \times p}$, $\mathbf{X}^{(3)} \in \mathbb{R}^{n_3 \times p}$ and $\boldsymbol{y}^{(1)} \in \mathbb{R}^{n_1}, \boldsymbol{y}^{(2)} \in \mathbb{R}^{n_2}, \boldsymbol{y}^{(3)} \in \mathbb{R}^{n_3}$. For each $k \in [p]$ problem ($\mathrm{BS}_{k,\lambda}$) is solved with $\mathbf{X}^{(1)}$ and $\boldsymbol{y}^{(1)}$ yielding the optimal solution $\hat{\boldsymbol{\beta}}^k$. Then, the validation error

$$\|\mathbf{X}^{(2)}\hat{\boldsymbol{\beta}}^k - \boldsymbol{y}^{(2)}\|_2^2 \tag{2}$$

is determined. At the end $\hat{k} = \operatorname{argmin}\{k \in [p] : \|\mathbf{X}^{(2)}\hat{\boldsymbol{\beta}}^k - \boldsymbol{y}^{(2)}\|_2^2\}$ is selected and the test error

$$\|\mathbf{X}^{(3)}\hat{\boldsymbol{\beta}}^{\hat{k}} - \boldsymbol{y}^{(3)}\|_2^2 \tag{3}$$

is calculated to assess the end result. Here, the objective value of ($\mathrm{BS}_{k,\lambda}$) is called the *training error* while the squared prediction error of a selected set of coefficients applied to the validation data is called the *validation error*, i.e., the value (2). Usually, when the final model is determined the predictive quality is then certified on an independent test data set. The resulting prediction error is then called *test error*, i.e., the result of (3). See also the book by Friedman et al. (2001) for a detailed and concise explanation of the model selection process. In the described process the ridge parameter is ignored. However, usually it is also part of the training and part of the parameters to validate. Hence, we are looking at a two-dimensional grid of validation points $G := \{(k, \lambda) : k \in [p], \lambda \in \boldsymbol{\Lambda}\}$ where $\boldsymbol{\Lambda}$ is the finite set of different values of λ.

```
Input: $\boldsymbol{y} \in \mathbb{R}^n$, $\mathbf{X} \in \mathbb{R}^{n \times p}$
Output: A model $S$.
for $(k, \lambda) \in G$ do
    Solve ($\mathrm{BS}_{k,\lambda}$) on training data and obtain model $S$;
    Validate $S$ and update the best model $\hat{S}$;
return $\hat{S}$;
```

Algorithm 1: Schematic process used to select a model in combination with ($\mathrm{BS}_{k,\lambda}$). The optimization problem ($\mathrm{BS}_{k,\lambda}$) is solved iteratively for each k. Each model is then tested on a validation data set and then the model with the best predictive quality in respect to the validation process is returned.

Validating on separate data $\mathbf{X}^{(2)}$ gives us a specific estimate of (1). We can, however, also apply other estimates of the prediction error (1) like for example a *cross-validation* (Friedman et al. 2001), adjusted R^2 (Draper and Smith 2014), BIC (Schwarz 1978) or AIC (Akaike 1974). Such an estimate of PE is denoted by $\widehat{\mathrm{PE}}$.

Commonly, when doing variable selection, a finite collection of possible models $S_1, \dots, S_d \subseteq [p]$ is chosen and then

$$\hat{S} = \operatorname*{argmin}_{S \in \{S_1, \dots, S_d\}} \widehat{\mathrm{PE}}(S)$$

is picked as the model of choice. However, this approach requires a careful selection of possible subsets from the beginning. In case of the best subset selection we define the set $\mathcal{S} := \{\text{supp}(\hat{\boldsymbol{\beta}}) : \hat{\boldsymbol{\beta}}$ is an optimal solution for $(\text{BS}_{k,\lambda}),\ k \in [p],\ \lambda \in \boldsymbol{\Lambda}\}$ and choose the best model according to the rule

$$\hat{S} = \underset{S \in \mathcal{S}}{\operatorname{argmin}}\ \widehat{\text{PE}}(S). \tag{MS}$$

In other words, the optimal solutions of all possible configurations of $(\text{BS}_{k,\lambda})$ form the set of models to validate.

1.2 Critique of the best subset selection

Given a collection $\mathcal{S}$ of models, for each set $S \in \mathcal{S}$ the coefficients should be fitted to the data indexed by S in the training stage. However, in case of problem $(\text{BS}_{k,\lambda})$ a subset S is selected with respect to the best training error and *not* a priori. In fact, only the sparsity is controlled and selected in the validation process. Hence, the model selection approach used for the best subset selection does not follow the model selection procedure described beforehand. The methodology of looking for the best model solely by training error can be problematic (see for instance Friedman et al. 2001, pp. 193 - 196) since training error and validation error do not necessarily correlate with each other. In the context of this work, we would rather like to have the model selection process happening at the validation stage and the model fitting process happening at the training stage. The process explained in Algorithm 1 puts part of the model selection into the training stage, that is, only the sparsity is selected in the validation stage.

Bertsimas and King (2016) address these problems by adding additional constraints to the mixed-integer program based on specific statistical insights. Those constraints are meant to exclude solutions, which are considered statistically insignificant with respect to external selection criteria. However, those metrics are not evaluated in the mixed-integer model itself but applied and enforced a posteriori, which requires significant more solver invocations. In comparison, Miyashiro and Takano (2015) use various information criteria like adjusted R^2 (Draper and Smith 2014), BIC (Schwarz 1978) and AIC (Akaike 1974) to select variables. They present a mixed-integer second-order model, which produces the intended outcome.

1.3 Contribution

Our contributions surrounds a novel formulation for the best subset selection. Instead of utilizing the training error to choose variables we use the cross-validation as our objective function. In part, we reflect results from the dissertation by Kreber (2019). In detail, we present the following results:

- We present a MIQP formulation which conducts an in-model cross-validation. Our program is constrained to only fit coefficients to training data but it can control which variables to include or exclude. With that, the objective is to minimize the validation error determined by a cross-validation. Using the cross-validation as an estimate for the prediction error enables us to require no further assumptions on the underlying data. This formulation allows us then to solve the optimization problem

$$\hat{S} = \operatorname*{argmin}_{S \subseteq [p]} \widehat{\mathrm{PE}}(S),$$

which greatly expands the search space for the best variable selection in comparison to (MS).

- We present an initial formulation using logical constraints. On one hand such constraints are easy to understand but on the other hand they cause CPLEX to have trouble finding a solution. Note that, Takano and Miyashiro (2019) independently and simultaneously developed a similiar formulation, which uses logical constraints. Our experience show that in general tight algebraic formulations can be solved more efficiently by the commercial solvers. Hence, we refrain from using only logical constraints and develop an algebraic formulation, which requires additional bounds on several terms. We proceed to present such bounds.
- We conduct a simulation study where we assess the cross-validation subset selection against the best subset selection ($\mathrm{BS}_{k,\lambda}$) and Lasso under several noise and multicollinearity settings.

1.4 Structure

In Section 2 we consider the validation process behind variable selection. In light of this, we formulate a mixed-integer quadratic program, which is used to minimize the validation error over all variable subsets using the cross-validation as an estimate for the prediction error. For this, we utilize logical constraints in order to construct a readily understandable program. Afterwards, we proceed to reformulate the program to only utilize algebraic constraints.

As we require "Big-M" constraints for the formulation we develop the necessary bounds in Section 3. We utilize bounds for the entries of inverse matrices to arrive at the desired result.

Subsequently, we conduct a simulation study comparing the best subset selection, the cross-validation subset selection and Lasso in Section 4. Here, we generate synthetic data with specific noise and multicollinearity setups. Since we know the true predictors we can assert the variable selection and predictive quality of all approaches.

2 A MINLP formulation for a cross-validation model selection

In this section we are presenting our formulation, which combines variable selection and model validation in one optimization model. For the model validation we are using a m-fold cross-validation. In other words we propose a novel MIQP formulation which is used to solve the problem

$$\min_{S \subseteq [p]} \widehat{\mathrm{PE}}(S). \tag{MINCV}$$

Here, we do not require a predetermined collection $\mathcal{S}$ of models, which we wish to validate. Instead, we select the best cross-validated subset out of all possible subsets. Assuming that $\widehat{\mathrm{PE}}$ is a good estimate of the predictive error, we consider a larger model space than with the best subset selection, and hence we can expect to find solutions which provide better predictions. We call problem (MINCV) the *cross-validation subset selection*.

We first define some notations we are using. The writing $\mathbf{X}_i$ denotes the i-th column of $\mathbf{X}$. The notation X_{ij} will be used to describe the entry of $\mathbf{X}$ which is located in the i-th row and j-th column. For a subset $S = \{i_1, \dots i_q\} \subseteq [p]$ with $q \in \mathbb{N}$ we will define

$$\mathbf{X}_S := \left[\mathbf{X}_{i_1}\ \mathbf{X}_{i_2}\ \cdots\ \mathbf{X}_{i_q}\right],$$

that is, $\mathbf{X}_S$ is the sub matrix with columns only indexed by S. For a vector $u \in \mathbb{R}^n$ we use the notation $u_S := (u_{i_1}, \dots, u_{i_q})$ to get subview of u indexed by $S \subseteq [n]$. We write $\mathbf{X}_{T,*}$ to select the rows of $\mathbf{X}$ indexed by $T \subseteq [n]$. For some generic set $M \subseteq [n]$ we will write $\bar{M}$ to denote the complement of M with respect to $[n]$. In order to apply a cross-validation we partition the index set of observations into m subsets $T_1, \dots, T_m$. Let $\bar{\mathbf{X}}^{(l)}$ be the matrix $\mathbf{X}_{\bar{T}_l,*}$ and let $\mathbf{X}^{(l)}$ be the matrix $\mathbf{X}_{T_l,*}$. The same notation is applied to the response $\boldsymbol{y}$, i.e., $\boldsymbol{y}_{T_l}$ is denoted by $\boldsymbol{y}^{(l)}$ and $\boldsymbol{y}_{\bar{T}_l}$ is denoted by $\bar{\boldsymbol{y}}^{(l)}$.

For some fixed subset $S \subseteq [p]$ the coefficients

$$\hat{\boldsymbol{\beta}}^{(l)}(S) = \underset{\boldsymbol{\beta} \in \mathbb{R}^{|S|}}{\operatorname{argmin}}\ \|\bar{\mathbf{X}}_S^{(l)}\boldsymbol{\beta} - \bar{\boldsymbol{y}}^{(l)}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2 \tag{4}$$

are computed for each $l \in [m]$. The cross-validation estimate of the prediction error is then given by

$$\widehat{\mathrm{PE}}(S) = \frac{1}{m}\sum_{l=1}^{m}\|\mathbf{X}_S^{(l)}\hat{\boldsymbol{\beta}}^{(l)}(S) - \boldsymbol{y}^{(l)}\|_2^2.$$

The issue we face with the aforementioned concept of a coherent MIQP is that we have to ensure strict separation between training and validation, i.e., the coefficients $\hat{\boldsymbol{\beta}}^{(l)}$ should strictly be fitted to the training data and must not be able to optimize the validation error. Therefore, we calibrate the coefficients using the normal equation

$$\mathrm{NE}_\lambda^l(\boldsymbol{\beta}, S) := (\bar{\mathbf{X}}_S^{(l)})^\mathsf{T}\bar{\mathbf{X}}_S^{(l)}\boldsymbol{\beta}_S + \lambda\boldsymbol{\beta}_S - (\bar{\mathbf{X}}_S^{(l)})^\mathsf{T}\bar{\boldsymbol{y}}^{(l)} = \mathbf{0}.$$

In other words $\mathrm{NE}_\lambda^l(\boldsymbol{\beta}, S) = \mathbf{0}$ holds if and only if $\boldsymbol{\beta}$ is an optimal solution of (4). Note, that it is important that for every $S \subseteq [p]$ the equation system $\mathrm{NE}_\lambda^l(\boldsymbol{\beta}, S) = \mathbf{0}$ has a unique solution $\boldsymbol{\beta}$. Furthermore, rewriting (4) as an algebraic formulation is important to keep the separation between training and validation intact. For instance, replacing the ℓ_2 regularization with ℓ_1 results in the inability to formulate the fitting process as an algebraic equation, making ℓ_1 an impractical regularization choice, as long as we do not rely on complementary programming techniques.

We can now formulate (MINCV) as following optimization problem.

$$\begin{aligned} \min \quad & \frac{1}{m}\sum_{l=1}^{m}\|\mathbf{X}_S^{(l)}\boldsymbol{\beta}_S^{l} - \boldsymbol{y}^{(l)}\|_2^2 \\ \text{s. t.} \quad & \mathrm{NE}_\lambda^l(\boldsymbol{\beta}^{(l)}, S) = \mathbf{0} \qquad \forall l \in [m] \\ & \boldsymbol{\beta}_{\bar{S}}^{(l)} = \mathbf{0} \qquad \forall l \in [m] \\ & |S| \leq k \\ & \boldsymbol{\beta}^{(1)}, \dots, \boldsymbol{\beta}^{(m)} \in \mathbb{R}^p, S \subseteq [p] \end{aligned} \tag{P}$$

Note that $|S| \leq k$ for some $k \in \mathbb{N}$ is an extension to (MINCV), in the sense that the modeler might choose $k \in \mathbb{N}$ in accordance to some anticipation or assumption about the sparsity the program should return. In that case, the sparsity can be restricted. However, in comparison to the best subset selection ($\mathrm{BS}_{k,\lambda}$) the sparsity constraint is not required and could be removed with little influence on the effectiveness of the program.

Although, (P) is a convenient illustration of the general idea we propose, its formulation cannot be entered into any of the commonly used MIQP solvers because of the set-valued decision variable S. Consequently, we present the following MIQP formulation.

$$\begin{aligned} \min \quad & \frac{1}{m}\sum_{l=1}^{m}\|\mathbf{X}^{(l)}\boldsymbol{\beta} - \boldsymbol{y}^{(l)}\|_2^2 \\ \text{s. t.} \quad & z_i = 1 \;\Rightarrow\; (\bar{\mathbf{X}}_i^{(l)})^\mathsf{T}\bar{\mathbf{X}}^{(l)}\boldsymbol{\beta}^{(l)} + \lambda\beta_i^{(l)} = (\bar{\mathbf{X}}_i^{(l)})^\mathsf{T}\bar{\boldsymbol{y}}^{(l)} \qquad \forall i \in [p], l \in [m] \\ & z_i = 0 \;\Rightarrow\; \beta_i^{(l)} = 0 \qquad \forall i \in [p], l \in [m] \\ & \mathbf{1}^\mathsf{T}\boldsymbol{z} \leq k \\ & \boldsymbol{\beta}^{(1)}, \dots, \boldsymbol{\beta}^{(m)} \in \mathbb{R}^p, \boldsymbol{z} \in \{0,1\}^p \end{aligned} \tag{$\mathrm{P}_{\mathrm{Ind}}$}$$

The formulation uses logical constraints, which most solvers can translate to algebraic constraints. We will later do this translation by ourselves and provide the necessary bounds required for this. Note that logical constraints can be replaced with SOS1 constraints. At least with CPLEX, our experience shows that SOS1 constraints are handled better, and hence they are preferable. However, for simpler exposition we use logical constraints to better illustrate the idea behind the formulation. First, we show that both formulations are indeed equivalent.

Theorem 2.1 *The formulation* (P) *is equivalent to* ($\mathrm{P}_{\mathrm{Ind}}$)*, that is, when considering $\boldsymbol{z}$ as an indicator vector, i.e., $S = \{i : z_i = 1\}$, both optimization problems yield the same optimal solution.*

Proof Let $(\hat{\boldsymbol{\beta}}^{(1)}, \dots, \hat{\boldsymbol{\beta}}^{(m)}, \hat{S})$ be an optimal solution of (P) and let $\hat{\boldsymbol{z}}$ be the indicator vector of $\hat{S}$, i.e., $\hat{S} = \{i : \hat{z}_i = 1\}$ holds. We first show that $(\hat{\boldsymbol{\beta}}^{(1)}, \dots, \hat{\boldsymbol{\beta}}^{(m)}, \hat{\boldsymbol{z}})$ is feasible for ($\mathrm{P}_{\mathrm{Ind}}$). For any $l \in [m]$ we have

$$\mathbf{X}_{\hat{S}}\hat{\boldsymbol{\beta}}_{\hat{S}}^{(l)} = \sum_{i \in S}\mathbf{X}_i\hat{\beta}_i^{(l)} = \sum_{i=1}^{p}\mathbf{X}_i\hat{\beta}_i^{(l)} = \mathbf{X}\hat{\boldsymbol{\beta}}^{(l)} \tag{5}$$

because $\hat{\beta}_i^{(l)} = 0$ holds for every $i \notin \hat{S}$. Since $\mathrm{NE}_\lambda^l(\hat{\boldsymbol{\beta}}^{(l)}, \hat{S}) = \mathbf{0}$ implies

$$(\bar{\mathbf{X}}_i^{(l)})^\mathsf{T} \bar{\mathbf{X}}_{\hat{S}}^{(l)} \hat{\boldsymbol{\beta}}_{\hat{S}}^{(l)} + \lambda \hat{\beta}_i^{(l)} - (\bar{\mathbf{X}}_i^{(l)})^\mathsf{T} \bar{\boldsymbol{y}}^{(l)} = 0$$

for every $i \in \hat{S}$, by (5) we get

$$(\bar{\mathbf{X}}_i^{(l)})^\mathsf{T} \bar{\mathbf{X}}^{(l)} \hat{\boldsymbol{\beta}}^{(l)} + \lambda \hat{\beta}_i^{(l)} - (\bar{\mathbf{X}}_i^{(l)})^\mathsf{T} \bar{\boldsymbol{y}}^{(l)} = 0$$

for every $i \in \hat{S}$. Therefore, for every $i \in [p]$ and every $l \in [m]$ the logical constraint $\hat{z}_i = 1 \Rightarrow (\bar{\mathbf{X}}_i^{(l)})^\mathsf{T} \bar{\mathbf{X}}^{(l)} \hat{\boldsymbol{\beta}}^{(l)} + \lambda \hat{\beta}_i^{(l)} = (\bar{\mathbf{X}}_i^{(l)})^\mathsf{T} \bar{\boldsymbol{y}}^{(l)}$ is satisfied. Furthermore, it is easy to see that the last two constraints $\hat{z}_i = 0 \Rightarrow \hat{\beta}_i^l = 0$ and $\mathbf{1}^\mathsf{T} \hat{\boldsymbol{z}} \leq k$ are satisfied since $\hat{\beta}_i^{(l)} = 0$ for every $i \notin \hat{S}$ and $|\hat{S}| \leq k$ hold. Additionally, by (5) both objective values are equal as well. Analogously, for every optimal solution $(\tilde{\boldsymbol{\beta}}^1, \dots, \tilde{\boldsymbol{\beta}}^m, \tilde{z})$ of $(\mathrm{P}_{\mathrm{Ind}})$ it follows that $(\tilde{\boldsymbol{\beta}}^{(1)}, \dots, \tilde{\boldsymbol{\beta}}^{(m)}, \tilde{S})$ is feasible for (P) and that both solutions provide the same objective value. Hence, the optimization problems (P) and $(\mathrm{P}_{\mathrm{Ind}})$ are equivalent.
□

Since mixed-integer programs utilizing logical constraint are more difficult to solve than programs with deliberately constructed algebraic constraints, we present the following Big-M formulation of $(\mathrm{P}_{\mathrm{Ind}})$.

$$\begin{aligned}
\min \quad & \frac{1}{m} \sum_{l=1}^{m} \left((\mathbf{X}^{(l)} \boldsymbol{\beta}^{(l)})^\mathsf{T} \mathbf{X}^{(l)} \boldsymbol{\beta}^{(l)} - 2 (\boldsymbol{y}^{(l)})^\mathsf{T} \mathbf{X}^{(l)} \boldsymbol{\beta}^{(l)} \right) \\
\text{s.t.} \quad & \qquad -L_i^{(l)} z_i \leq \beta_i^{(l)} \leq L_i^{(l)} z_i \\
& (\bar{\mathbf{X}}_i^{(l)})^\mathsf{T} \bar{\mathbf{X}}^{(l)} \boldsymbol{\beta}^{(l)} + \lambda \beta_i^{(l)} \leq M_i^{(l)} (1 - z_i) + (\bar{\mathbf{X}}_i^{(l)})^\mathsf{T} \bar{\boldsymbol{y}}^{(l)} \\
& (\bar{\mathbf{X}}_i^{(l)})^\mathsf{T} \bar{\mathbf{X}}^{(l)} \boldsymbol{\beta}^{(l)} + \lambda \beta_i^{(l)} \geq -m_i^{(l)} (1 - z_i) + (\bar{\mathbf{X}}_i^{(l)})^\mathsf{T} \bar{\boldsymbol{y}}^{(l)} \\
& \qquad\qquad\qquad\qquad\qquad\qquad (\forall i \in [p], l \in [m]) \\
& \qquad\quad \mathbf{1}^\mathsf{T} \boldsymbol{z} \leq k \\
& \boldsymbol{\beta}^{(1)}, \dots, \boldsymbol{\beta}^{(m)} \in \mathbb{R}^p, \boldsymbol{z} \in \{0,1\}^p
\end{aligned} \tag{$\mathrm{P}_{\mathrm{BigM}}$}$$

For sufficiently large constants $L_i^{(l)}, m_i^{(l)}, M_i^{(l)}$ the proposed program is equivalent to $(\mathrm{P}_{\mathrm{Ind}})$. The tighter we can choose the model constants, the stronger our formulation becomes. Therefore, we will propose appropriate bounds in Section 3. In order for the formulation to be statistically meaningful we have to make some key assumptions on the data $\mathbf{X}$ before considering technical details about the program $(\mathrm{P}_{\mathrm{BigM}})$.

2.1 Data preprocessing

Since data preprocessing for $(\mathrm{P}_{\mathrm{BigM}})$ is much more technical than it is for the best subset selection, we explain the data preparation in detail in this section. In order for the ridge penalization to produce consistent results, we have to standardize all variables, and hence adapt the validation data as well. Otherwise, the regularization terms $\lambda\|\boldsymbol{\beta}^{(l)}\|_2^2$ would influence variables with various magnitudes, which would lead to undesirable results. The transformations are applied *before* partitioning the data. Usually data is standardized in the following way:

- Normalization: each variable is scaled such that the variance is equal to 1.
- Centering: each variable is shifted such that the mean is 0.
- Intercept: a **1** column is added to account for an affine displacement in the data.

Normalization requires no further considerations, and we simply scale all variables such that the columns have ℓ_2-norm of 1. However, centering the covariates requires a more deliberate approach. Assume we have some ridge regression model with added intercept and centered variables

$$\min_{\beta_0\in\mathbb{R},\boldsymbol{\beta}\in\mathbb{R}^p} \|\mathbf{1}\beta_0 + \sum_{i=1}^{p}(\mathbf{X}_i - \mu_i\mathbf{1})\beta_i - \boldsymbol{y}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2 \tag{6}$$

with $\mu_i = \frac{1}{n}\sum_{j=1}^{n} X_{ij}$ being the mean of $\mathbf{X}_i$. By this formulation it is easy to see that centering a variable already accumulates the intercept. That is, if $(\hat{\beta}_0,\hat{\beta}_1,\dots,\hat{\beta}_p)$ is an optimal solution for (6), then the vector $(\hat{\beta}_0 - \sum_{i=1}^{p}\mu_i\hat{\beta}_i,\hat{\beta}_1,\dots,\hat{\beta}_p)$ is an optimal solution for

$$\min_{\beta_0\in\mathbb{R},\boldsymbol{\beta}\in\mathbb{R}^p} \|\mathbf{1}\beta_0 + \sum_{i=1}^{p}\mathbf{X}_i\beta_i - \boldsymbol{y}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2.$$

and vice versa. Therefore, adding an intercept already accounts for centering the variables. However, centering a variable generates implicit intercept, even though the true model might not have any constant shift. In light of our presented model, the intercept is handled exactly like the covariates, that is, it can be ex- or included by the solver. Hence, the former requires a conscious decision by the user while the later gives the freedom of choice to the algorithm. Since we are interested in presenting an automated approach to model selection, we add an intercept without centering the variables.

However, when adding a column of 1's we do not want to penalize this intercept by the ridge regression. Thus, it must be excluded from the regularization. This issue would normally pose no problem, and in fact it doesn't produce any issues for the optimization problem ($\mathrm{P}_{\mathrm{Ind}}$). With an additional intercept column we would simply reformulate the problem to

$$\begin{array}{lll}
\min & \frac{1}{m}\sum_{l=1}^{m}\left(\left(\left[\mathbf{1}\ \mathbf{X}^{(l)}\right]\boldsymbol{\beta}^{(l)}\right)^{\mathsf{T}}\left[\mathbf{1}\ \mathbf{X}^{(l)}\right]\boldsymbol{\beta}^{(l)} - 2(\boldsymbol{y}^{(l)})^{\mathsf{T}}\left[\mathbf{1}\ \mathbf{X}^{(l)}\right]\boldsymbol{\beta}^{(l)}\right) & \\
\text{s. t.} & z_i = 0 \ \Rightarrow\ \beta_i^{(l)} = 0 & \forall i \in [p],\ l \in [m] \\
& z_1 = 1 \ \Rightarrow\ \mathbf{1}^{\mathsf{T}}\left[\mathbf{1}\ \bar{\mathbf{X}}^{(l)}\right]\boldsymbol{\beta}^{(l)} = \mathbf{1}^{\mathsf{T}}\bar{\boldsymbol{y}}^{(l)} & \forall l \in [m] \\
& z_i = 1 \ \Rightarrow\ (\bar{\mathbf{X}}_{i-1}^{(l)})^{\mathsf{T}}\left[\mathbf{1}\ \bar{\mathbf{X}}^{(l)}\right]\boldsymbol{\beta}^{(l)} + \lambda\beta_i^{(l)} = (\bar{\mathbf{X}}_{i-1}^{(l)})^{\mathsf{T}}\bar{\boldsymbol{y}}^{(l)} & \forall i \in [p]\setminus\{1\},\ l \in [m] \\
& \quad \mathbf{1}^{\mathsf{T}}\boldsymbol{z} \le k & \\
& \boldsymbol{\beta}^{(1)},\dots,\boldsymbol{\beta}^{(m)} \in \mathbb{R}^p, \boldsymbol{z} \in \{0,1\}^p &
\end{array}$$

However, it causes some complications when finding bounds for $(\mathrm{P}_{\mathrm{BigM}})$. To better deal with different ridge penalizations we generalize problem $(\mathrm{P}_{\mathrm{BigM}})$ to

$$
\begin{aligned}
\min \quad & \frac{1}{m}\sum_{l=1}^{m}\left((\mathbf{X}^{(l)}\boldsymbol{\beta}^{(l)})^{\mathsf{T}}\mathbf{X}^{(l)}\boldsymbol{\beta}^{(l)} - 2(\boldsymbol{y}^{(l)})^{\mathsf{T}}\mathbf{X}^{(l)}\boldsymbol{\beta}^{(l)}\right) \\
\text{s. t.} \quad & \qquad -L_i^{(l)} z_i \le \beta_i^{(l)} \le L_i^{(l)} z_i \\
& (\bar{\mathbf{X}}_i^{(l)})^{\mathsf{T}}\bar{\mathbf{X}}^{(l)}\boldsymbol{\beta}^{(l)} + \gamma_i\beta_i^{(l)} \le M_i^{(l)}(1-z_i) + (\bar{\mathbf{X}}_i^{(l)})^{\mathsf{T}}\bar{\boldsymbol{y}}^{(l)} \\
& (\bar{\mathbf{X}}_i^{(l)})^{\mathsf{T}}\bar{\mathbf{X}}^{(l)}\boldsymbol{\beta}^{(l)} + \gamma_i\beta_i^{(l)} \ge -m_i^{(l)}(1-z_i) + (\bar{\mathbf{X}}_i^{(l)})^{\mathsf{T}}\bar{\boldsymbol{y}}^{(l)} \\
& \qquad\qquad\qquad\qquad\qquad\qquad (\forall i \in [p], l \in [m]) \\
& \qquad \mathbf{1}^{\mathsf{T}}\boldsymbol{z} \le k \\
& \boldsymbol{\beta}^{(1)},\dots,\boldsymbol{\beta}^{(m)} \in \mathbb{R}^p, \boldsymbol{z} \in \{0,1\}^p
\end{aligned}
\tag{$\mathrm{Q}_{\mathrm{BigM}}$}
$$

with parameters $\gamma_1,\dots,\gamma_p \ge 0$. Hence, an added intercept would be a special case of $(\mathrm{Q}_{\mathrm{BigM}})$. Furthermore, we denote the diagonal matrix $\mathrm{diag}(\gamma_1,\dots,\gamma_p)$ by $\boldsymbol{\Gamma} \in \mathbb{R}^{p\times p}$.

In conclusion, we assume that all covariates are normalized and that the validation data is modified in accordance to the normalization. Furthermore, we assume that the model accounts for an intercept, i.e., that $\mathbf{X}_1^{(l)} = \mathbf{1}$ for all $l \in \{1,\dots,m\}$ with parameter $\gamma_1 = 0$.

2.2 Data assumptions

After pre-processing the design matrix $\mathbf{X}$, we have to ensure that the fitted coefficients $\boldsymbol{\beta}^{(l)}$ are unique for every combination of selected variables. Otherwise, the solver would choose each $\boldsymbol{\beta}^{(l)}$ such that the validation error is minimized. However, this would lead to a dependence between validation and training, which would most likely result in overfitting. Thus, from now on we assume that $(\bar{\mathbf{X}}^{(l)})^{\mathsf{T}}\bar{\mathbf{X}}^{(l)} + \boldsymbol{\Gamma}$ is positive definite. We can weaken this assumption by requiring that $(\bar{\mathbf{X}}_S^l)^{\mathsf{T}}\bar{\mathbf{X}}_S^l + \sqrt{\boldsymbol{\Gamma}_S^{\mathsf{T}}\boldsymbol{\Gamma}_S}$ is positive definite for all $S \subseteq [p]$ with $|S| \le k$. However, then the bounds presented in the next section would cease to work, and we would have to fall back to the formulation $(\mathrm{P}_{\mathrm{Ind}})$. Therefore, we assume the former.

2.3 Implementation remarks

We noticed that often there are many feasible solutions having objective values very close to each other. In this case, since the cross-validation is an estimation of the prediction error, it can happen that a solution is determined to be the best optimal solution of $(\mathrm{Q}_{\mathrm{BigM}})$ even though it provides a worse prediction error than other approximately equal-valued solutions. Hence, experience showed that it is much better to select the sparsest solution from a pool of nearly identically valued solutions at the end of the optimization. Intuitively, that means we select a solution from a candidate pool, which requires the least information and assumptions to produce the prediction.

With CPLEX this is possible by allowing the solver to store the best solutions with regard to a defined range in a solution pool, which can be iterated at the end of the optimization. For the presented problem a tolerance of 0.05 is chosen, i.e., solutions which are at most 5% worse (by means of the relative gap) than the best integer solution are kept in the solution pool.

3 Bounds for the model constants

Most modern MINLP solvers support use of logical constraints like in ($\mathrm{P}_{\mathrm{Ind}}$) and thus we could put the formulation directly into a solver and search for a global optimum. However, the performance difference between MINLPs with logical constraints and MINLPs with only algebraic constraints can be quite stark, if the "Big-M" constants are chosen sufficiently tight. Hence, we are enticed to find strong bounds on the solutions to derive tight model constants.

Finding bounds on the absolute values of the coefficients seem to be central to this task. Here, the difficulty lies in finding valid bounds for all possible subsets of variables. Bertsimas et al. (2016) develop analytic bounds for the coefficients for the subset selection problem. They do not regard a regularization and neither assume full column rank on the whole data set. Hence, their assumption to derive bounds requires a variation of diagonal dominance. However, this assumptions is quite restrictive and is not satisfied in most practical applications. Nevertheless, they also present a data driven, algorithmic approach to derive bounds. Since we assume to have a unique solution for each training set, we can derive tighter analytic bounds without any further requirements.

The section is structured as follows. First, we present a bound for the norm of the regularized predicted values over all possible subsets. That means, for some design matrix $\mathbf{X}$ with full column rank let us define

$$\hat{\boldsymbol{\beta}}(S) := \underset{\boldsymbol{\beta}\in\mathbb{R}^{|S|}}{\operatorname{argmin}} \; \|\mathbf{X}_S\boldsymbol{\beta} - \boldsymbol{y}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2$$

and the full dimensional vector

$$\begin{aligned} \tilde{\boldsymbol{\beta}}(S) := \underset{\boldsymbol{\beta}\in\mathbb{R}^{p}}{\operatorname{argmin}} \quad & \|\mathbf{X}\boldsymbol{\beta} - \boldsymbol{y}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2 \\ \text{s.t.} \quad & \boldsymbol{\beta}_{\bar{S}} = \mathbf{0}. \end{aligned}$$

We find a bound $c(\mathbf{X}, \boldsymbol{y}, k)$ such that $\|\mathbf{X}_S\hat{\boldsymbol{\beta}}(S)\|_2^2 + \lambda\|\hat{\boldsymbol{\beta}}(S)\|_2^2 \leq c(\mathbf{X}, \boldsymbol{y}, k)$ for all $S \subseteq [p]$ and $|S| \leq k$. This result helps us in the second part of the section to find tighter bounds for $|\tilde{\boldsymbol{\beta}}(S)_i|$ over all possible subsets S. Finally, we derive the model constants necessary for the "Big-M" approach in the last part of the section. In the two first parts we are mostly concerned with one single design matrix $\mathbf{X}$ and a consistent ridge scalar and later apply the results to the various training matrices in the presented formulation. Hence, most results are also applicable to the ridge regularized best subset selection.

3.1 Bound on the norm of the regularized predicted value

Let $S \subseteq [p]$ be a valid subset, i.e., its cardinality is less or equal than k and let λ be positive. We present an upper bound for $\|\mathbf{X}_S\hat{\boldsymbol{\beta}}(S)\|_2^2 + \lambda\|\hat{\boldsymbol{\beta}}(S)\|_2^2$ only dependent on $\mathbf{X}$ and the required sparsity k. We first prove the following

Lemma 3.1 *The identity*

$$\|\boldsymbol{y}\|_2^2 - \|\mathbf{X}_S\hat{\boldsymbol{\beta}}(S)\|_2^2 - \lambda\|\hat{\boldsymbol{\beta}}(S)\|_2^2 = \|\mathbf{X}_S\hat{\boldsymbol{\beta}}(S) - \boldsymbol{y}\|_2^2 + \lambda\|\hat{\boldsymbol{\beta}}(S)\|_2^2. \tag{7}$$

holds.

Proof By the KKT conditions, $\hat{\boldsymbol{\beta}}(S)$ is equivalently an optimal solution of

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^{|S|}} \left\| \begin{bmatrix} \mathbf{X}_S \\ \sqrt{\lambda}\mathbf{I} \end{bmatrix} \boldsymbol{\beta} - \begin{pmatrix} \boldsymbol{y} \\ \mathbf{0} \end{pmatrix} \right\|_2^2.$$

Hence, by well-known results (see for example Seber (1977, p. 43)) on least squares problems the following orthogonality holds

$$\left(\begin{bmatrix} \mathbf{X}_S \\ \sqrt{\lambda}\mathbf{I} \end{bmatrix} \hat{\boldsymbol{\beta}}(S) - \begin{pmatrix} \boldsymbol{y} \\ \mathbf{0} \end{pmatrix} \right)^\top \left(\begin{bmatrix} \mathbf{X}_S \\ \sqrt{\lambda}\mathbf{I} \end{bmatrix} \hat{\boldsymbol{\beta}}(S) \right) = 0.$$

Thus, by using

$$\|\boldsymbol{y}\|_2^2 - \|\mathbf{X}_S\hat{\boldsymbol{\beta}}(S)\|_2^2 - \lambda\|\hat{\boldsymbol{\beta}}(S)\|_2^2 = \left\| \begin{pmatrix} \boldsymbol{y} \\ \mathbf{0} \end{pmatrix} \right\|_2^2 - \left\| \begin{bmatrix} \mathbf{X}_S \\ \sqrt{\lambda}\mathbf{I} \end{bmatrix} \hat{\boldsymbol{\beta}}(S) \right\|_2^2$$

and

$$\|\mathbf{X}_S\hat{\boldsymbol{\beta}}(S) - \boldsymbol{y}\|_2^2 + \lambda\|\hat{\boldsymbol{\beta}}(S)\|_2^2 = \left\| \begin{bmatrix} \mathbf{X}_S \\ \sqrt{\lambda}\mathbf{I} \end{bmatrix} \hat{\boldsymbol{\beta}}(S) - \begin{pmatrix} \boldsymbol{y} \\ \mathbf{0} \end{pmatrix} \right\|_2^2$$

the equality (7) follows easily. □

From this we conclude that a lower bound on $\|\mathbf{X}_S\hat{\boldsymbol{\beta}}(S) - \boldsymbol{y}\|_2^2 + \lambda\|\hat{\boldsymbol{\beta}}(S)\|_2^2$ can be transformed to an upper bound for $\|\mathbf{X}_S\hat{\boldsymbol{\beta}}(S)\|_2^2 + \lambda\|\hat{\boldsymbol{\beta}}(S)\|_2^2$.

Immediately, a trivial bound could be derived. Since $\|\mathbf{X}_S\hat{\boldsymbol{\beta}}(S) - \boldsymbol{y}\|_2^2 + \lambda\|\hat{\boldsymbol{\beta}}(S)\|_2^2 \geq 0$ holds, we have $\|\mathbf{X}_S\hat{\boldsymbol{\beta}}(S)\|_2^2 + \lambda\|\hat{\boldsymbol{\beta}}(S)\|_2^2 \leq \|\boldsymbol{y}\|_2^2$. Even though, this bound is easy to compute we can do better and tighten this further with the following result. Note that the $\max_{[k]}$ operator is defined as the sum over the k largest elements.

Lemma 3.2 *Let S be a subset of $[p]$ with $|S| \leq k$ and let $\lambda > 0$. Then, the ridge regression value is bounded from below by*

$$\frac{\|\boldsymbol{y}\|_2^4}{\|\boldsymbol{y}\|_2^2 + \frac{1}{\lambda}\max_{[k]}\left\{\boldsymbol{y}^\top\mathbf{X}_i\mathbf{X}_i^\top\boldsymbol{y} : i \in [p]\right\}} \leq \|\mathbf{X}\hat{\boldsymbol{\beta}}(S) - \boldsymbol{y}\|_2^2 + \lambda\|\hat{\boldsymbol{\beta}}(S)\|_2^2.$$

Proof Since $\hat{\boldsymbol{\beta}}(S)$ satisfies $\mathbf{X}_S^\top \mathbf{X}_S \hat{\boldsymbol{\beta}}(S) + \lambda \hat{\boldsymbol{\beta}}(S) = \mathbf{X}_S^\top \boldsymbol{y}$, we have $\hat{\boldsymbol{\beta}}(S) = (\mathbf{X}_S^\top \mathbf{X}_S + \lambda \mathbf{I})^{-1} \mathbf{X}_S^\top \boldsymbol{y}$. Replacing $\hat{\boldsymbol{\beta}}(S)$ with this term yields

$$\|\mathbf{X}_S \hat{\boldsymbol{\beta}}(S) - \boldsymbol{y}\|_2^2 + \lambda \|\hat{\boldsymbol{\beta}}(S)\|_2^2 = \boldsymbol{y}^\top \boldsymbol{y} - \boldsymbol{y}^\top \mathbf{X}_S \left(\mathbf{X}_S^\top \mathbf{X}_S + \lambda \mathbf{I}\right)^{-1} \mathbf{X}_S^\top \boldsymbol{y}$$

Using the Sherman-Morrison-Woodbury matrix identity (Meyer 2000) and the decomposition $\mathbf{X}_S \mathbf{X}_S^\top = \sum_{i \in S} \mathbf{X}_i \mathbf{X}_i^\top$ gives us

$$\begin{aligned} \boldsymbol{y}^\top \boldsymbol{y} - \boldsymbol{y}^\top \mathbf{X}_S \left(\mathbf{X}_S^\top \mathbf{X}_S + \lambda \mathbf{I}\right)^{-1} \mathbf{X}_S^\top \boldsymbol{y} &= \boldsymbol{y}^\top \left(\mathbf{I} + \frac{1}{\lambda} \mathbf{X}_S \mathbf{X}_S^\top\right)^{-1} \boldsymbol{y} \\ &= \boldsymbol{y}^\top \left(\mathbf{I} + \frac{1}{\lambda} \sum_{i \in S} \mathbf{X}_i \mathbf{X}_i^\top\right)^{-1} \boldsymbol{y} \end{aligned}$$

Since $\mathbf{I} + \frac{1}{\lambda} \sum_{i \in S} \mathbf{X}_i \mathbf{X}_i^\top$ is symmetric and positive definite we can take the square root of the matrix in $\mathbb{R}^{p \times p}$. With the Cauchy-Schwarz inequality it follows that

$$\begin{aligned} (\boldsymbol{y}^\top \boldsymbol{y})^2 &= \left(\boldsymbol{y}^\top \left(\mathbf{I} + \frac{1}{\lambda} \sum_{i \in S} \mathbf{X}_i \mathbf{X}_i^\top\right)^{-\frac{1}{2}} \left(\mathbf{I} + \frac{1}{\lambda} \sum_{i \in S} \mathbf{X}_i \mathbf{X}_i^\top\right)^{\frac{1}{2}} \boldsymbol{y}\right)^2 \\ &\leq \boldsymbol{y}^\top \left(\mathbf{I} + \frac{1}{\lambda} \sum_{i \in S} \mathbf{X}_i \mathbf{X}_i^\top\right)^{-1} \boldsymbol{y} \cdot \boldsymbol{y}^\top \left(\mathbf{I} + \frac{1}{\lambda} \sum_{i \in S} \mathbf{X}_i \mathbf{X}_i^\top\right) \boldsymbol{y}. \end{aligned}$$

Thus, by reordering the terms we get

$$\begin{aligned} \frac{\|\boldsymbol{y}\|_2^4}{\boldsymbol{y}^\top \boldsymbol{y} + \frac{1}{\lambda} \max_{[k]} \left\{\boldsymbol{y}^\top \mathbf{X}_i \mathbf{X}_i^\top \boldsymbol{y} : i \in [p]\right\}} &\leq \frac{\|\boldsymbol{y}\|_2^4}{\boldsymbol{y}^\top \left(\mathbf{I} + \frac{1}{\lambda} \sum_{i \in S} \mathbf{X}_i \mathbf{X}_i^\top\right) \boldsymbol{y}} \\ &\leq \boldsymbol{y}^\top \left(\mathbf{I} + \frac{1}{\lambda} \sum_{i \in S} \mathbf{X}_i \mathbf{X}_i^\top\right)^{-1} \boldsymbol{y} \\ &= \|\mathbf{X}_S \hat{\boldsymbol{\beta}}(S) - \boldsymbol{y}\|_2^2 + \lambda \|\hat{\boldsymbol{\beta}}(S)\|_2^2 \end{aligned}$$

We can then apply this lemma to the regularized, squared predicted values by using Equation (7).

Corollary 3.3 *Let be $\lambda > 0$ and let be $S \in [p]$, $|S| \leq k$. Then, the inequality*

$$\|\mathbf{X}_S \hat{\boldsymbol{\beta}}(S)\|_2^2 + \lambda \|\hat{\boldsymbol{\beta}}(S)\|_2^2 \leq \|\boldsymbol{y}\|_2^2 - \frac{\|\boldsymbol{y}\|_2^4}{\|\boldsymbol{y}\|_2^2 + \frac{1}{\lambda} \max_{[k]} \left\{\boldsymbol{y}^\top \mathbf{X}_i \mathbf{X}_i^\top \boldsymbol{y} : i \in [p]\right\}} =: c(\mathbf{X}, \boldsymbol{y}, k)$$

holds.

In the proof of Lemma 3.2 it is possible to derive more computational intense, non-analytic bounds. Instead of finding a lower estimate by using the Cauchy-Schwartz inequality one could as well solve the relaxation

$$\begin{aligned} \min \quad & \boldsymbol{y}^\top \left(\mathbf{I} + \frac{1}{\lambda} \sum_{i=1}^{p} \mathbf{X}_i \mathbf{X}_i^\top z_i \right)^{-1} \boldsymbol{y} \\ \text{s. t.} \quad & \sum_{i=1}^{p} z_i \le t \\ & 0 \le z_i \le 1 \end{aligned}$$

Bertsimas and Van Parys (2017) show that the relaxation can be efficiently solved as a second-order cone program. With that, paying an additional computational cost would enable $c(\mathbf{X}, \boldsymbol{y}, k)$ to be tightened even more.

3.2 Bound on the absolute regression coefficients

In the previous section, we focused on computing an upper bound on the regularized, squared predicted values and now extend this work to derive a bound for the absolute values of the coefficients of a best subset selection ridge regression, i.e., we find an upper estimate to $|\tilde{\boldsymbol{\beta}}(S)_i|$. To our knowledge most approaches estimating the coefficients of a subset selection come across the necessity to calculate an upper bound of $(\lambda_{\min}(\mathbf{X}_S^\top \mathbf{X}_S))^{-1}$, which should be independent of the subset S. In general the Cauchy Interlacing theorem (see for example Horn and Johnson 2013, pp. 242) can be used to estimate $\lambda_{\min}(\mathbf{X}_S^\top \mathbf{X}_S)$ against $\lambda_{\min}(\mathbf{X}^\top \mathbf{X})$. However, if $\mathbf{X}$ does not have full column rank, this estimation yields the trivial bound 0, which is impractical in this setting since it cannot be inverted. Bertsimas et al. (2016) circumvent this issue by requiring more restrictive assumptions, which are akin to the diagonal dominance property. Our approach assumes that $\mathbf{X}^\top \mathbf{X} + \boldsymbol{\Gamma}$ is positive definite, which we require to derive a lower estimate of the minimal eigenvalue.

In order to derive bounds on the individual entries of $\hat{\boldsymbol{\beta}}(S)$, we utilize the following result, which estimates the diagonal entries of the inverse of some positive definite matrix $\mathbf{A}$.

Proposition 3.4 (Robinson and Wathen (1992)) *For a positive definite matrix $A \in \mathbb{R}^{m \times m}$ assume $\rho, \tau \in \mathbb{R}$ to be chosen such that they satisfy $\lambda_{\max}(\mathbf{A}) \le \rho$ and $0 < \tau \le \lambda_{\min}(\mathbf{A})$. Then, for $i \in [m]$ the following bounds hold*

i) $(\mathbf{A}^{-1})_{ii} \le \frac{1}{4}\left(\frac{\rho}{\tau} + \frac{\tau}{\rho} + 2\right) \cdot (\mathbf{A}_{ii})^{-1} =: g_i^1(\mathbf{A}, \tau, \rho)$

ii) $(\mathbf{A}^{-1})_{ii} \le \frac{1}{\tau} - (\mathbf{A}_{ii} - \tau)^2 \cdot \left(\tau \left(\sum_{k=1}^{m} \mathbf{A}_{ik}^2 - \tau \mathbf{A}_{ii}\right)\right)^{-1} =: g_i^2(\mathbf{A}, \tau)$

Using the proposition we can prove the following bounds for the absolute values of the coefficient entries.

Theorem 3.5 *Let be* $S \subseteq [p]$ *with* $|S| \leq k$ *and* $\lambda > 0$. *Then, the two inequalities*

$$|\tilde{\boldsymbol{\beta}}(S)_i| \leq \sqrt{c(\mathbf{X}, \boldsymbol{y}, k) \cdot g_i^1(\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda\mathbf{I},\ \lambda_{\min}(\mathbf{X}^\mathsf{T}\mathbf{X}) + \lambda,\ \lambda_{\max}(\mathbf{X}^\mathsf{T}\mathbf{X}) + \lambda)}$$

and

$$|\tilde{\boldsymbol{\beta}}(S)_i| \leq \sqrt{c(\mathbf{X}, \boldsymbol{y}, k) \cdot g_i^2(\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda\mathbf{I},\ \lambda_{\min}(\mathbf{X}^\mathsf{T}\mathbf{X}) + \lambda)}$$

hold.

Proof We first define

$$\mathbf{W} := \begin{pmatrix} \mathbf{X} \\ \sqrt{\lambda} \cdot \mathbf{I} \end{pmatrix}.$$

Clearly, $\mathbf{W}$ has full column rank. Denoting the unit vector with entry 1 at position i by $\boldsymbol{e}_i$ and the pseudoinverse of $\mathbf{W}$ by $\mathbf{W}^+$, we first note that

$$\begin{aligned}
|\tilde{\boldsymbol{\beta}}(S)_i| &= |\boldsymbol{e}_i{}^\mathsf{T}\tilde{\boldsymbol{\beta}}(S)| \\
&= |\boldsymbol{e}_i{}^\mathsf{T}\mathbf{W}^+\mathbf{W}\tilde{\boldsymbol{\beta}}(S)| \\
&= |((\mathbf{W}^+)^\mathsf{T}\boldsymbol{e}_i)^\mathsf{T}\mathbf{W}\tilde{\boldsymbol{\beta}}(S)| \\
&\leq \|(\mathbf{W}^+)^\mathsf{T}\boldsymbol{e}_i\|_2 \|\mathbf{W}_S\hat{\boldsymbol{\beta}}(S)\|_2 \\
&= \|(\mathbf{W}^+)^\mathsf{T}\boldsymbol{e}_i\|_2 \sqrt{\|\mathbf{X}_S\hat{\boldsymbol{\beta}}(S)\|_2^2 + \lambda\|\hat{\boldsymbol{\beta}}(S)\|_2^2} \\
&\leq \sqrt{((\mathbf{W}^+)^\mathsf{T}\boldsymbol{e}_i)^\mathsf{T}(\mathbf{W}^+)^\mathsf{T}\boldsymbol{e}_i \cdot c(\mathbf{X}, \boldsymbol{y}, k)} \\
&= \sqrt{\boldsymbol{e}_i{}^\mathsf{T}\mathbf{W}^+(\mathbf{W}^+)^\mathsf{T}\boldsymbol{e}_i \cdot c(\mathbf{X}, \boldsymbol{y}, k)} \\
&= \sqrt{\boldsymbol{e}_i{}^\mathsf{T}(\mathbf{W}^\mathsf{T}\mathbf{W})^+\boldsymbol{e}_i \cdot c(\mathbf{X}, \boldsymbol{y}, k)} \\
&= \sqrt{((\mathbf{W}^\mathsf{T}\mathbf{W})^{-1})_{ii} \cdot c(\mathbf{X}, \boldsymbol{y}, k)}.
\end{aligned}$$

We are going to use the bounds presented in Proposition 3.4 to prove our claim. Therefore, we determine an appropriate ρ and τ independent of S.

It holds that $\lambda_{\min}(\mathbf{W}^\mathsf{T}\mathbf{W}) = \lambda_{\min}(\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda\mathbf{I}) = \lambda_{\min}(\mathbf{X}^\mathsf{T}\mathbf{X}) + \lambda$ and $\lambda_{\max}(\mathbf{W}^\mathsf{T}\mathbf{W}) = \lambda_{\max}(\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda\mathbf{I}) = \lambda_{\max}(\mathbf{X}^\mathsf{T}\mathbf{X}) + \lambda$. Thus, we have $\tau := \lambda_{\min}(\mathbf{X}^\mathsf{T}\mathbf{X}) + \lambda$ and $\rho := \lambda_{\max}(\mathbf{X}^\mathsf{T}\mathbf{X}) + \lambda$ as feasible choices for the bounds presented in Proposition 3.4.

Therefore, we get

$$\sqrt{((\mathbf{W}^\mathsf{T}\mathbf{W})^{-1})_{ii} \cdot c(\mathbf{X}, \boldsymbol{y}, k)} \leq \sqrt{g_i^1\left(\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda\mathbf{I},\ \tau,\ \rho\right) \cdot c(\mathbf{X}, \boldsymbol{y}, k)}$$

and

$$\sqrt{((\mathbf{W}^\mathsf{T}\mathbf{W})^{-1})_{ii} \cdot c(\mathbf{X}, \boldsymbol{y}, k)} \leq \sqrt{g_i^2\left(\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda\mathbf{I},\ \tau\right) \cdot c(\mathbf{X}, \boldsymbol{y}, k)}$$

which proves the original statement. □

Let us denote the minimum of the two bounds by

$$g_i(\mathbf{A}, \tau, \rho) := \min\left\{g_i^1(\mathbf{A}, \tau, \rho),\ g_i^2(\mathbf{A}, \tau)\right\}.$$

We use the results to derive the necessary Big-M constants for $(\mathrm{Q}_{\mathrm{BigM}})$.

3.3 Model constants

The results presented until now are valid for the best subset selection problem with a ridge regularization term. However, we would like to apply the results to a more general setting as presented in (Q_{BigM}). To begin with we show the connection between the ordinary best subset selection and the ridge regularized best subset selection. This will be helpful to reformulate

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^{|S|}} \|\mathbf{X}_S\boldsymbol{\beta} - \boldsymbol{y}\|_2^2 + \|(\sqrt{\boldsymbol{\Gamma}})_S\boldsymbol{\beta}\|_2^2$$

with possible zero-diagonal-entries in $\boldsymbol{\Gamma}$ as an ordinary ridge regression problem in the form

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^{|S|}} \|\mathbf{X}_S\boldsymbol{\beta} - \boldsymbol{y}\|_2^2 + r\|\boldsymbol{\beta}\|_2^2$$

with $r > 0$ and then utilize the bounds proven previously.

Lemma 3.6 *Let* $\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}$ *be the singular value decomposition of* $\mathbf{X}$. *Assume that either* $\mathbf{X}$ *has full column rank or that* $\lambda > 0$. *For* $r \in (-\infty, \lambda_{\min}(\mathbf{X}^{\mathsf{T}}\mathbf{X}) + \lambda)$ *denote* $\widetilde{\boldsymbol{\Sigma}} := \sqrt{\boldsymbol{\Sigma}^{\mathsf{T}}\boldsymbol{\Sigma} + \lambda\mathbf{I} - r\mathbf{I}}$, $\widetilde{\mathbf{X}} := \widetilde{\boldsymbol{\Sigma}}\mathbf{V}$ *and* $\tilde{\boldsymbol{y}} := \widetilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}^{\mathsf{T}}\mathbf{U}^{\mathsf{T}}\boldsymbol{y}$. *Then, the equations*

$$\operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^{|S|}} \|\mathbf{X}_S\boldsymbol{\beta} - \boldsymbol{y}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2 = \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^{|S|}} \|\widetilde{\mathbf{X}}_S\boldsymbol{\beta} - \tilde{\boldsymbol{y}}\|_2^2 + r\|\boldsymbol{\beta}\|_2^2$$

and

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^{|S|}} \|\mathbf{X}_S\boldsymbol{\beta} - \boldsymbol{y}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2 = \min_{\boldsymbol{\beta}\in\mathbb{R}^{|S|}} \|\widetilde{\mathbf{X}}_S\boldsymbol{\beta} - \tilde{\boldsymbol{y}}\|_2^2 + r\|\boldsymbol{\beta}\|_2^2 + \|\boldsymbol{y}\|_2^2 - \|\tilde{\boldsymbol{y}}\|_2^2$$

hold for every subset $S \subseteq [p]$.

Proof First note, that $\widetilde{\boldsymbol{\Sigma}}$ is well-defined and non-singular. Since

$$\boldsymbol{\Sigma}^{\mathsf{T}}\boldsymbol{\Sigma} = \operatorname{diag}(\lambda_p(\mathbf{X}^{\mathsf{T}}\mathbf{X}), \ldots, \lambda_1(\mathbf{X}^{\mathsf{T}}\mathbf{X}))$$

holds for the eigenvalues $\lambda_p(\mathbf{X}^{\mathsf{T}}\mathbf{X}) \geq \cdots \geq \lambda_1(\mathbf{X}^{\mathsf{T}}\mathbf{X})$, the diagonal matrix $\boldsymbol{\Sigma}^{\mathsf{T}}\boldsymbol{\Sigma}+\lambda\mathbf{I}-r\mathbf{I}$ has positive diagonal entries and therefore the square root of $\boldsymbol{\Sigma}^{\mathsf{T}}\boldsymbol{\Sigma} + \lambda\mathbf{I} - r\mathbf{I}$ is well-defined in $\mathbb{R}^{p\times p}$. Additionally, since all diagonal entries are positive, $\widetilde{\boldsymbol{\Sigma}}$ is non-singular as well. Since by assumption $\mathbf{X}^{\mathsf{T}}\mathbf{X}+\lambda\mathbf{I}$ is positive definite, both optimization problems

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^{|S|}} \|\mathbf{X}_S\boldsymbol{\beta} - \boldsymbol{y}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2, \qquad \min_{\boldsymbol{\beta}\in\mathbb{R}^{|S|}} \|\widetilde{\mathbf{X}}_S\boldsymbol{\beta} - \tilde{\boldsymbol{y}}\|_2^2 + r\|\boldsymbol{\beta}\|_2^2$$

are strictly convex and thus have unique solutions, which we denote by $\hat{\boldsymbol{\beta}}_1 = (\mathbf{X}_S^{\mathsf{T}}\mathbf{X}_S + \lambda\mathbf{I})^{-1}\mathbf{X}_S^{\mathsf{T}}\boldsymbol{y}$ and $\hat{\boldsymbol{\beta}}_2 = (\widetilde{\mathbf{X}}_S^{\mathsf{T}}\widetilde{\mathbf{X}}_S + r\mathbf{I})^{-1}\widetilde{\mathbf{X}}_S^{\mathsf{T}}\tilde{\boldsymbol{y}}$. Furthermore, it holds that

$$\mathbf{X}_S^{\mathsf{T}}\mathbf{X}_S + \lambda\mathbf{I} = \mathbf{V}_S^{\mathsf{T}}\boldsymbol{\Sigma}^{\mathsf{T}}\boldsymbol{\Sigma}\mathbf{V}_S + \lambda\mathbf{I} = V_S^{\mathsf{T}}(\widetilde{\boldsymbol{\Sigma}}^2 - \lambda\mathbf{I} + r\mathbf{I})\mathbf{V}_S + \lambda\mathbf{I} = \widetilde{\mathbf{X}}_S^{\mathsf{T}}\widetilde{\mathbf{X}}_S + r\mathbf{I} \tag{8}$$

and

$$\mathbf{X}_S^{\mathsf{T}}\boldsymbol{y} = \mathbf{V}_S^{\mathsf{T}}\boldsymbol{\Sigma}^{\mathsf{T}}\mathbf{U}^{\mathsf{T}}\boldsymbol{y} = \mathbf{V}_S^{\mathsf{T}}\widetilde{\boldsymbol{\Sigma}}\widetilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}^{\mathsf{T}}\mathbf{U}^{\mathsf{T}}\boldsymbol{y} = \widetilde{\mathbf{X}}_S^{\mathsf{T}}\tilde{\boldsymbol{y}}.$$

Thus, we have $\hat{\boldsymbol{\beta}}_1 = (\mathbf{X}_S^\top \mathbf{X}_S + \lambda \mathbf{I})^{-1} \mathbf{X}_S^\top \boldsymbol{y} = (\widetilde{\mathbf{X}}_S^\top \widetilde{\mathbf{X}}_S + rI)^{-1} \widetilde{\mathbf{X}}_S^\top \tilde{\boldsymbol{y}} = \hat{\boldsymbol{\beta}}_2$. Furthermore, the objective values of both optimization problems are equal modulo an additive constant:

$$\begin{aligned}
\|\mathbf{X}\hat{\boldsymbol{\beta}}_1 - \boldsymbol{y}\|_2^2 + \lambda \|\hat{\boldsymbol{\beta}}_1\|_2^2 &= \hat{\boldsymbol{\beta}}_1^\top \mathbf{X}^\top \mathbf{X} \hat{\boldsymbol{\beta}}_1 + \lambda \hat{\boldsymbol{\beta}}_1^\top \hat{\boldsymbol{\beta}}_1 - 2\boldsymbol{y}^\top \mathbf{X} \hat{\boldsymbol{\beta}}_1 + \boldsymbol{y}^\top \boldsymbol{y} \\
&= \hat{\boldsymbol{\beta}}_1^\top \widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}} \hat{\boldsymbol{\beta}}_1 + r \hat{\boldsymbol{\beta}}_1^\top \hat{\boldsymbol{\beta}}_1 - 2\tilde{\boldsymbol{y}}^\top \widetilde{\mathbf{X}} \hat{\boldsymbol{\beta}}_1 + \boldsymbol{y}^\top \boldsymbol{y} \\
&= \|\widetilde{\mathbf{X}} \hat{\boldsymbol{\beta}}_1 - \tilde{\boldsymbol{y}}\|_2^2 + r \|\hat{\boldsymbol{\beta}}_1\|_2^2 + \boldsymbol{y}^\top \boldsymbol{y} - \tilde{\boldsymbol{y}}^\top \tilde{\boldsymbol{y}},
\end{aligned}$$

which proves the lemma. □

Theorem 3.7 *Let be $l \in [m]$, let $(\boldsymbol{\beta}^{(1)}, \ldots, \boldsymbol{\beta}^{(m)}, z)$ be a feasible solution of* $(\mathrm{Q}_{\mathrm{BigM}})$ *and assume*

$$\begin{bmatrix} \bar{\mathbf{X}}^{(l)} \\ \sqrt{\boldsymbol{\Gamma}} \end{bmatrix} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}$$

to be a singular value decomposition as in Lemma 3.6. Define $\rho^{(l)} := \lambda_{\min}((\mathbf{X}^{(l)})^\top \mathbf{X}^{(l)} + \boldsymbol{\Gamma})$ and $\tau^{(l)} := \lambda_{\max}((\mathbf{X}^{(l)})^\top \mathbf{X}^{(l)} + \boldsymbol{\Gamma})$. Furthermore, denote

$$G_i^{(l)} := g_i((\bar{\mathbf{X}}^{(l)})^\top \bar{\mathbf{X}}^{(l)} + \boldsymbol{\Gamma},\ \rho^{(l)},\ \tau^{(l)})$$

and

$$C^{(l)} := c\left(\left(\boldsymbol{\Sigma}^\top \boldsymbol{\Sigma} - rI\right)^{\frac{1}{2}} V, \quad \left(\boldsymbol{\Sigma}^\top \boldsymbol{\Sigma} - rI\right)^{-\frac{1}{2}} \boldsymbol{\Sigma}^\top U^\top \bar{\boldsymbol{y}}^{(l)}, \quad k\right)$$

for $0 < r < \rho^l$. Then,

$$|\boldsymbol{\beta}_i^{(l)}| \le \sqrt{G_i^{(l)} \cdot C^{(l)}} =: L_i^{(l)}$$

is a valid inequality for $(\mathrm{Q}_{\mathrm{BigM}})$.

Proof Let be $S := \{i : z_i = 1\}$ and define

$$\mathbf{W} := \begin{bmatrix} \bar{\mathbf{X}}^{(l)} \\ \sqrt{\boldsymbol{\Gamma}} \end{bmatrix}, \qquad \boldsymbol{w} := \begin{pmatrix} \bar{\boldsymbol{y}}^{(l)} \\ \mathbf{0} \end{pmatrix}.$$

Since $\boldsymbol{\beta}^{(l)}$ is feasible, it satisfies

$$(\bar{\mathbf{X}}_S^{(l)})^\top \bar{\mathbf{X}}^{(l)} \boldsymbol{\beta}^{(l)} + \boldsymbol{\Gamma}\boldsymbol{\beta}^{(l)} = (\bar{\mathbf{X}}_S^{(l)})^\top \bar{\boldsymbol{y}}^{(l)}$$

and $\boldsymbol{\beta}_{\bar{S}} = \mathbf{0}$. Hence, $\boldsymbol{\beta}_S^{(l)}$ is an optimal solution of

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^{|S|}} \|\mathbf{W}_S \boldsymbol{\beta} - \boldsymbol{w}\|_2^2$$

and because $\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}$ is a singular value decomposition of $\mathbf{W}$, Lemma 3.6 yields that $\boldsymbol{\beta}_S^{(l)}$ is also an optimal solution of

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^{|S|}} \left\| \left(\boldsymbol{\Sigma}^\top \boldsymbol{\Sigma} - r\mathbf{I}\right)^{\frac{1}{2}} \mathbf{V}_S \boldsymbol{\beta} - \left(\boldsymbol{\Sigma}^\top \boldsymbol{\Sigma} - r\mathbf{I}\right)^{-\frac{1}{2}} \boldsymbol{\Sigma}^\top \mathbf{U}^\top \bar{\boldsymbol{y}}^{(l)} \right\|_2^2 + r \|\boldsymbol{\beta}\|_2^2. \tag{9}$$

Additionally, by (8) it holds that the Gramian matrix $\mathbf{H} := \mathbf{V}_S^\mathsf{T}\left(\mathbf{\Sigma}^\mathsf{T}\mathbf{\Sigma} - r\mathbf{I}\right)\mathbf{V}_S + r\mathbf{I}$, which comes from the design matrix of (9), is equal to $\mathbf{W}_S^\mathsf{T}\mathbf{W}_S = \mathbf{X}_S^\mathsf{T}\mathbf{X}_S + \mathbf{\Gamma}$. Thus,

$$g_i(\mathbf{H}, \lambda_{\min}(\mathbf{H}), \lambda_{\max}(\mathbf{H})) = G_i^{(l)}$$

holds. From this and by Proposition 3.5 it follows that the bound

$$|\boldsymbol{\beta}_i^{(l)}| \leq \sqrt{G_i^{(l)} \cdot C^{(l)}}$$

is valid. □

Consequently, we can set the model constants $L_i^{(l)}$ to $\sqrt{G_i^{(l)} \cdot C^{(l)}}$ without altering the solution set of ($\mathrm{Q}_{\mathrm{BigM}}$). Now, we consider the constants M_i^l and m_i^l. They are parts of the inequalities

$$(\bar{\mathbf{X}}_i^{(l)})^\mathsf{T}\bar{\mathbf{X}}^{(l)}\boldsymbol{\beta}^{(l)} + \lambda_i\beta_i^{(l)} \leq M_i^{(l)}(1 - z_i) + (\bar{\mathbf{X}}_i^{(l)})^\mathsf{T}\bar{\boldsymbol{y}}^{(l)} \qquad \forall i \in [p], l \in [m] \tag{10}$$

$$(\bar{\mathbf{X}}_i^{(l)})^\mathsf{T}\bar{\mathbf{X}}^{(l)}\boldsymbol{\beta}^{(l)} + \lambda_i\beta_i^{(l)} \geq -m_i^{(l)}(1 - z_i) + (\bar{\mathbf{X}}_i^{(l)})^\mathsf{T}\bar{\boldsymbol{y}}^{(l)} \qquad \forall i \in [p], l \in [m] \tag{11}$$

of problem ($\mathrm{Q}_{\mathrm{BigM}}$).

Theorem 3.8 *For each $l \in [m]$ let $C^{(l)}$ be defined as in Proposition 3.7. Then,*

$$M_i^{(l)} = \sqrt{\left(\|\bar{\mathbf{X}}_i^{(l)}\|_2^2 + \gamma_i\right)C^{(l)}} - (\bar{\mathbf{X}}^{(l)})^\mathsf{T}\bar{\boldsymbol{y}}^{(l)}$$

and

$$m_i^{(l)} = \sqrt{\left(\|\bar{\mathbf{X}}_i^{(l)}\|_2^2 + \gamma_i\right)C^{(l)}} + (\bar{\mathbf{X}}^{(l)})^\mathsf{T}\bar{\boldsymbol{y}}^{(l)}$$

are valid constants for ($\mathrm{Q}_{\mathrm{BigM}}$).

Proof Let $(\boldsymbol{\beta}^{(1)}, \ldots, \boldsymbol{\beta}^{(k)}, \boldsymbol{z})$ be a feasible solution of ($\mathrm{Q}_{\mathrm{BigM}}$). We find an upper estimate for $|(\bar{\mathbf{X}}_i^{(l)})^\mathsf{T}\bar{\mathbf{X}}^{(l)}\boldsymbol{\beta}^{(l)} + \lambda_i\beta_i^{(l)}|$ and consequently derive $M_i^{(l)}$ and $m_i^{(l)}$. By the Cauchy-Schwarz inequality we have that

$$\begin{aligned}
|(\bar{\mathbf{X}}_i^{(l)})^\mathsf{T}\bar{\mathbf{X}}^{(l)}\boldsymbol{\beta}^{(l)} + \lambda_i\beta_i^{(l)}| &= \left|\begin{pmatrix}\bar{\mathbf{X}}_i^{(l)} \\ \sqrt{\gamma_i}\boldsymbol{e}_i\end{pmatrix}^\mathsf{T}\begin{bmatrix}\bar{\mathbf{X}}^{(l)} \\ \sqrt{\mathbf{\Gamma}}\end{bmatrix}\boldsymbol{\beta}^{(l)}\right| \\
&\leq \left\|\begin{pmatrix}\bar{\mathbf{X}}_i^{(l)} \\ \sqrt{\gamma_i}\boldsymbol{e}_i\end{pmatrix}\right\|_2 \left\|\begin{bmatrix}\bar{\mathbf{X}}^{(l)} \\ \sqrt{\mathbf{\Gamma}}\end{bmatrix}\boldsymbol{\beta}^{(l)}\right\|_2 \\
&\leq \sqrt{\left(\|\bar{\mathbf{X}}_i^{(l)}\|_2^2 + \gamma_i\right)C^{(l)}}
\end{aligned}$$

Accounting for $(\bar{\mathbf{X}}^{(l)})^\mathsf{T}\bar{\boldsymbol{y}}^{(l)}$ in the inequalities (10) and (11) yields the result. □

4 Simulation study

Our simulation study is setup in the following way: we synthetically generate the design matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$, sparse coefficients $\boldsymbol{\beta}^0 \in \mathbb{R}^p$ and noise $\boldsymbol{\epsilon} \in \mathbb{R}^n$. Then, the response $\boldsymbol{y} \in \mathbb{R}^n$ is computed by $\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta}^0 + \boldsymbol{\epsilon}$. In this way, we know the true coefficients, can try to recover them with various algorithms and compare the results. Two setup parts emerge from the experiment description. In the first part we are concerned with how to generate the data and in the second part we determine what algorithms to use and how to set the corresponding parameters.

4.1 Data generation

We first consider the design of $\mathbf{X}$. In light of this, we analyze the following problem size:

$$n = 2000, \quad p = 20, \quad \|\boldsymbol{\beta}^0\|_0 = 5$$

We draw each row of $\mathbf{X}$ i.i.d. from $N_p(0, \boldsymbol{\Sigma})$ with

- **multicoll-none**: $\boldsymbol{\Sigma} = \mathbf{I}$.
- **multicoll-1**: $\Sigma_{i,j} = 0.5^{|i-j|}$ for all $i, j \in [p]$.
- **multicoll-2**: $\Sigma_{i,j} = \begin{cases} 1, & \text{if } i = j, \\ 0.9, & \text{if } i \neq j. \end{cases}$

We then select coefficients $\boldsymbol{\beta}^0 \in \mathbb{R}^p$ subject to the sparsity condition for the respective dimension setting. Non-zero entries of $\boldsymbol{\beta}^0$ are uniformly drawn from the interval $[1, 10]$. The placement of the entries are drawn from the uniform distribution. In other words, we uniformly draw $v_1, \ldots, v_5$ from the interval $[1, 10]$. We then draw a subset $\{s_1, \ldots, s_5\} \subseteq [p]$ and create the coefficients $\boldsymbol{\beta}^0$ according to the rule

$$\boldsymbol{\beta}_i^0 := \begin{cases} v_j, & \text{if } i = s_j \text{ for some } j \in \{1, \ldots, 5\}, \\ 0, & \text{otherwise.} \end{cases}$$

After creating the coefficients we generate the noise $\boldsymbol{\epsilon}$ added to $\mathbf{X}\boldsymbol{\beta}^0$, which is drawn i.i.d from $N_n(\mathbf{0}, \sigma^2\mathbf{I})$, with σ^2 detailed below. The noise setting is the standard requirement usually assumed for least squares regression. In order to measure the severance of the noise we consider the signal-to-noise ratio (SNR). The SNR describes the proportion of the signal in comparison to the noise. A high SNR means there is very little noise compared to the signal whereas a low SNR describes the effect of a significant noise interference. The ratio is defined as the quotient of the variance of the predicted response and the variance of the noise, i.e.,

$$\text{SNR} := \frac{\text{Var}(\boldsymbol{x}^0\boldsymbol{\beta}^0)}{\text{Var}(\epsilon)} = \frac{(\boldsymbol{\beta}^0)^\top \boldsymbol{\Sigma} \boldsymbol{\beta}^0}{\sigma^2}.$$

In our experiment we consider the values

$$\text{SNR} = 0.3, \ 1, \ 3, \ 6.$$

Accordingly, we choose the noise variance σ^2 to fit the desired SNR value. In this sense, low SNR leads to a high noise variance whereas a high SNR leads to a low noise variance.

4.2 Algorithm setting

We consider different algorithms for the simulation all of which are described in this work. We are disabling intercept for all methods.

- **BS**: The best subset selection ($\mathrm{BS}_{k,\lambda}$) with ridge parameter $\lambda \geq 0$. The sparsity level k and the ridge parameter is selected via a cross-validation. The grid for the ridge parameters is detailed below.
- **CVBS**: The cross-validation subset selection as proposed in this article with ridge parameter $\lambda \geq 0$ and 10 folds. The optimal ridge parameter is selected on a grid, which is detailed below, via a k-fold cross-validation also consisting of 10 folds. We provide a warm start computed by SparseNet (Mazumder et al. 2011).
- **LASSO**: The Lasso method due to Tibshirani (1996). We are using the `R` implementation found in the package `glmnet` (Friedman et al. 2010). The k-fold cross-validation used by `glmnet` utilizes 10 folds and the mean squared error as the loss function.

The methods BS and CVBS are developed in `C++` and called in `R`. The MIPs are solved via CPLEX. For the algorithms BS and CVBS we are cross-validating each ridge parameter on a predefined grid. That is, let G denote the number of evaluation points on the grid, $\bar{\lambda}$ the upper regularization parameter limit and $\underline{\lambda}$ the lower regularization parameter limit. Then, the ridge parameter grid is constructed by

$$\text{grid} = \{e^{(i-1)\cdot\frac{\log(\bar{\lambda}-\underline{\lambda}+1)}{G-1}} - 1 + \underline{\lambda} : i \in [G]\}.$$

In the experiments we choose $\bar{\lambda} = 10$ and $\underline{\lambda} = 0$. Table 1 shows the algorithm setup.

Algorithm	#Repetitions	Time limit	Reg. grid size
BS	99	720 per (k, μ)	20
CVBS	99	720 per μ	20
LASSO	99	NA	100

Table 1 Algorithmic settings.

4.3 Evaluation setup

Let $\hat{\boldsymbol{\beta}}$ be the estimated coefficients and x^0 a new observation drawn from $N_p(\mathbf{0}, \boldsymbol{\Sigma})$ with response $y^0 = \boldsymbol{x}^0\boldsymbol{\beta}^0 + \epsilon^0$.

- **Model difference**: Even if a method produces the correct sparsity, the non-zero entries could still be placed in deviation from the true model. The metric tells us how many coefficients are out of place. It is defined by

$$\text{MDIFF} = \|\mathbb{I}_{\text{supp}(\boldsymbol{\beta}^0)} - \mathbb{I}_{\text{supp}(\hat{\boldsymbol{\beta}})}\|_0$$

  where $\mathbb{I}_S$ denotes the indicator vector representing a set S.
- **ℓ_2-difference from true coefficients**: This metric evaluates how far the estimated coefficients $\hat{\boldsymbol{\beta}}$ deviate from the true coefficients $\boldsymbol{\beta}^0$ with respect to the ℓ_2-norm.

$$\text{L2DIFF} = \|\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\|_2^2$$

- **Relative test error**: The metric measures the test error divided by the noise variance.

$$\begin{aligned} \text{RTE} &= \frac{\mathbb{E}(y^0 - \boldsymbol{x}^0\hat{\boldsymbol{\beta}})^2}{\text{Var}(\epsilon^0)} \\ &= \frac{\mathbb{E}(\epsilon - \boldsymbol{x}^0(\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}))^2}{\text{Var}(\epsilon^0)} \\ &= \frac{(\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}})\boldsymbol{\Sigma}(\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}) + \sigma^2}{\sigma^2} \end{aligned}$$

  The perfect score is 1 whereas the null score is SNR + 1. The metric is also used by Hastie et al. (2017) and in some deviation by Bertsimas et al. (2016).

4.4 Hardware

We conducted the experiments on a machine with two Intel Xeon CPU E5-2699 v4 @ 2.20GHz (2 × 44 threads) and a random access memory capacity of 756 GB.

4.5 Simulation results

Predictive quality We first examine the relative test error over all scenarios. According to Figure 1 we identify the following key observations:

- CVBS produces the lowest prediction errors in most cases.
- LASSO fails to produce competitive results in all cases.
- CVBS and BS produce more consistent results than LASSO, i.e., the prediction errors produced by them have significant less variance.

The study shows evidence that CVBS produces considerably better predictions than the competing approaches. Only in the case when noise and multicollinearity are high, BS yields better predictions than CVBS. Figure 2 shows a detailed comparison between CVBS and BS. Although both approaches yield good predictions, over all scenarios the distance between the median of the relative test error of BS and the perfect score 1 is about 1.5 times larger than the same statistic for CVBS. That is, we have

$$\frac{\text{median}(\text{RTE}(\text{BS})) - 1}{\text{median}(\text{RTE}(\text{CVBS})) - 1} = 1.515418.$$

Exchanging the median for the mean gives the value of 1.315038. Hence, the relative difference in prediction quality is still considerable.

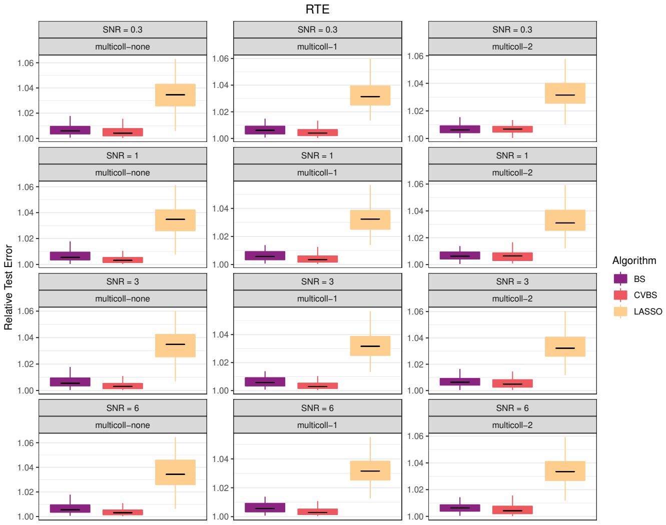

Fig. 1 Comparison of the relative test error between BS (in purple), CVBS (in orange) and LASSO (in yellow) for all SNR and multicollinearity settings.

It should also be emphasized that the cross-validation subset selection does not require the solution of p many mixed-integer optimization problems. Instead, we only have to solve one discrete optimization problem (for each ridge parameter). Admittedly, the mixed-integer program associated with the cross-validation subset selection is about m times larger than the program associated with BS, i.e., for every fold used for the cross-validation we need a new set of variables. Nevertheless, m is usually a constant (5 or 10) whereas p is part of the input size.

Yet oftentimes BS runs faster due to various improvements and advancements in recent years (Dong et al. 2015; Bertsimas et al. 2016; Bertsimas and Van Parys 2017; Atamtürk and Gómez 2018), whereas the proposed method certainly still provides plenty of untapped potential in terms of efficiency.

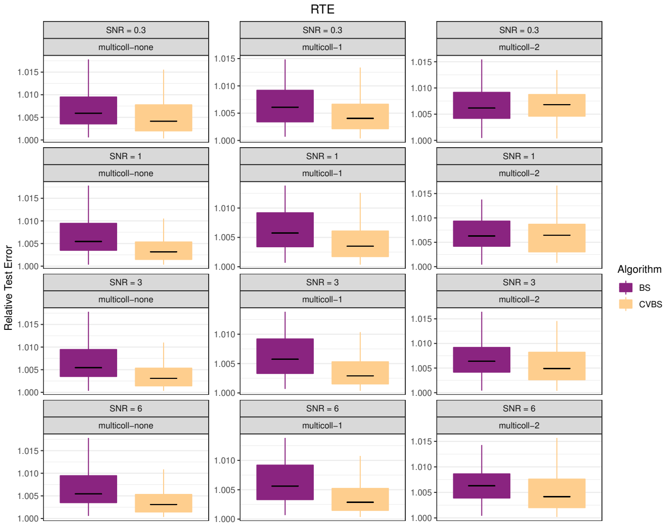

Fig. 2 Detailed comparison of the relative test error between BS (in purple) and CVBS (in yellow) for all SNR values and multicollinearity settings.

Considering LASSO, we observe that it produces results which have significantly higher test errors. While LASSO can be solved considerably quicker and more efficient than the subset selection methods, the simulation study shows that a loss of predictive quality is the price for such a computational advantage.

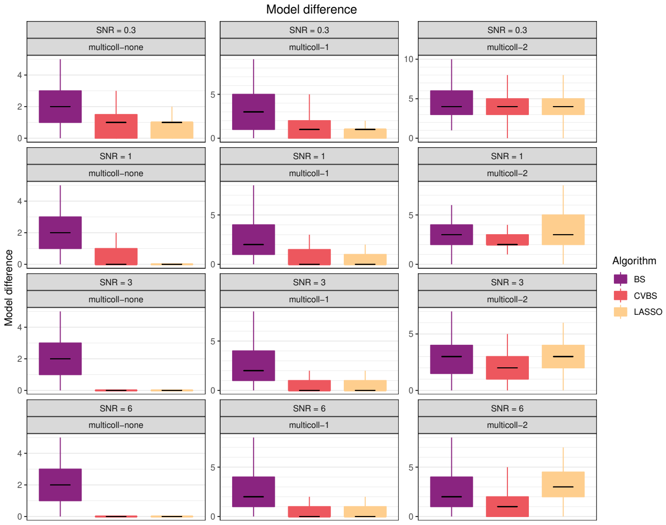

Fig. 3 Comparison of the variable selection accuracy between BS (in purple), CVBS (in orange) and LASSO (in yellow) for all SNR values and multicollinearity settings.

Selection accuracy We now want to consider the model difference and the accuracy of the coefficient estimation. From Figure 3 we can deduce the following observations. CVBS and LASSO select the variables most accurately. LASSO produces the best results when there is no multicollinearity whereas CVBS shows a larger error variance in these scenarios. When multicollinearity is high, LASSO produces significantly worse selections while CVBS yields the most accurate variable selections in this setting. It can be observed that in all scenarios the median of the model differences produced by CVBS is less or equal than the median corresponding to LASSO. In summary, CVBS yields very accurate variable selections, however in few cases it outputs less consistent selections than LASSO.

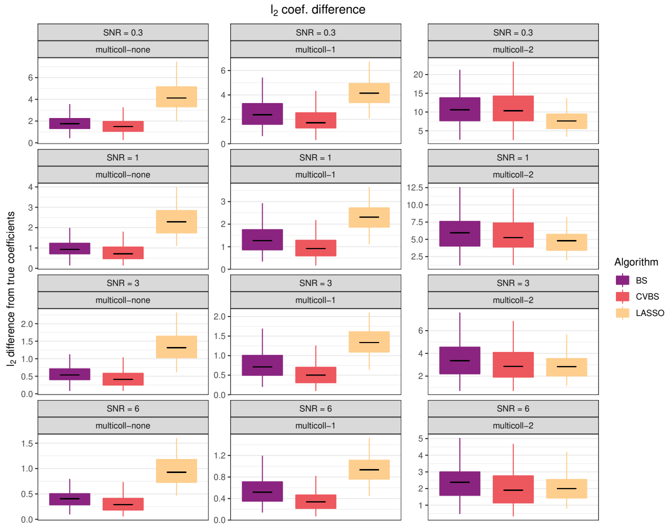

Fig. 4 Comparison of the ℓ_2 coefficient difference between BS (in purple), CVBS (in orange) and LASSO (in yellow) for all SNR values and multicollinearity settings.

Quality of coefficient estimation Surprisingly, although BS output good predictions it generates the least accurate variable selections. On the other hand, LASSO yields good selections but fails to generate low prediction errors. Hence, we also examine the ℓ_2-difference between the true coefficients and the estimated coefficients to get a better understanding of the previous observations. Figure 4 shows that for multicol-none and multicol-1 LASSO has difficulties estimating the coefficients, i.e, the ℓ_2-difference is significantly higher compared to the subset selection approaches. The cause for this effect is that the method underestimates the coefficients by penalizing them with the ℓ_1 regularization, which at the same time is required for the variable selection. Therefore, to have a better variable selection LASSO has to sacrifice coefficient quality. CVBS and BS do not have this disadvantage as the coefficient estimation and variable selection are decoupled. For multicol-2 LASSO produces better estimates but as we have seen before it also outputs a worse variable selection in this scenario.

Discussion In summary, CVBS generates the best results in most categories followed by BS. LASSO, on the other hand, outputs less accurate predictions caused by inaccurate coefficient estimations. Considering the selection accuracy, LASSO fares better if no multicollinearity is present, however it generates worse selections in the case of higher multicollinearity. Here, CVBS produces much more consistent results, i.e., the selection accuracy is more robust under multicollinearity. The study supports the initial reasoning that BS is misaligned with the actual statistical intention as it produces results with higher prediction errors.

5 Conclusion

We considered the problem of variable selection in regression and reviewed two prominent approaches – the Lasso method and the best subset selection. The best subset selection has drawn increasing interest in recent years and has become a popular research topic.

We argued that the notion of the best subset selection has some flaws and leaves room for improvement. Hence, we proposed a cross-validation subset selection method. For this approach an estimation of the prediction error is used as the objective function, making a cardinality constraint obsolete. Moreover, with the cross-validation subset selection the whole space of subsets is validated, i.e., checked against the estimate of the prediction error. In comparison, only the best cardinality-constrained subsets with respect to the in-sample error are validated with the best subset selection. Hence, from a theoretical point of view the cross-validation subset selection requires fewer assumptions about the data and has a concise objective without intermediate steps.

The first program we proposed was based on logical constraints. However, implementing a program in this form can lead to slow solver performance. Therefore, we have reformulated the program to only have algebraic constrains and determined the necessary Big-M constants. For deriving the bounds we relied on the assumption that the normal equations yield a unique solution due to $\mathbf{X}^\mathsf{T}\mathbf{X} + \boldsymbol{\Gamma}$ being positive definite.

Finally, we asserted the cross-validation subset selection in a simulation study. We observed that the cross-validation subset selection yields excellent predictions while requiring a single mixed-integer program to determine the necessary sparsity. The findings showed that the cross-validation subset selection method produces considerably better predictions than the Lasso method and the best subset selection. Furthermore, we observed that the proposed method produces the most consistent variable selections.

It is evident that discrete optimization plays a major role in statistics and data science and provides novel opportunities for statistics. The proposed cross-validation subset selection regression shows that the application of mixed-integer optimization can lead to superior results compared to the state-of-the-art approach Lasso and has advantages over the best subset selection.

References

Akaike H (1974) A new look at the statistical model identification. IEEE Transactions on Automatic Control 19(6):716–723, DOI 10.1109/TAC.1974.1100705

Atamtürk A, Gómez A (2018) Strong formulations for quadratic optimization with M-matrices and indicator variables. Mathematical Programming 170(1):141–176, DOI 10.1007/s10107-018-1301-5

Bertsimas D, Copenhaver MS (2018) Characterization of the equivalence of robustification and regularization in linear and matrix regression. European Journal of Operational Research 270(3):931–942, DOI 10.1016/J.EJOR.2017.03.051

Bertsimas D, King A (2016) OR Forum - An algorithmic approach to linear regression. Operations Research 64(1):2–16, DOI 10.1287/opre.2015.1436

Bertsimas D, Van Parys B (2017) Sparse high-dimensional regression: Exact scalable algorithms and phase transitions. ArXiv e-prints pp 1–22, 1709.10029

Bertsimas D, King A, Mazumder R (2016) Best subset selection via a modern optimization lens. The Annals of Statistics 44(2):813–852, DOI 10.1214/15-AOS1388

Bühlmann P (2013) Causal statistical inference in high dimensions. Mathematical Methods of Operations Research 77(3):357–370, DOI 10.1007/s00186-012-0404-7

Bühlmann P, van de Geer S (2011) Statistics for High-Dimensional Data: Methods, Theory and Applications. Springer, DOI 10.1080/02664763.2012.694258

Dong H, Chen K, Linderoth J (2015) Regularization vs. relaxation: A conic optimization perspective of statistical variable selection. ArXiv e-prints 1510.06083

Draper NR, Smith H (2014) Applied Regression Analysis. John Wiley & Sons

Friedman J, Hastie T, Tibshirani R (2001) The Elements of Statistical Learning. Springer Series in Statistics, New York

Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software 33(1):1–22, DOI 10.18637/jss.v033.i01

Hastie T, Tibshirani R, Tibshirani RJ (2017) Extended comparisons of best subset selection, forward stepwise selection, and the Lasso. ArXiv e-prints 1707.08692

Horn RA, Johnson CR (2013) Matrix Analysis, 2nd edn. Cambridge University Press

Konno H, Yamamoto R (2009) Choosing the best set of variables in regression analysis using integer programming. Journal of Global Optimization 44:273–282, DOI 10.1007/s10898-008-9323-9

Kreber D (2019) Cardinality-Constrained Discrete Optimization for Regression. PhD thesis, Trier University

Mazumder R, Friedman JH, Hastie T (2011) SparseNet: Coordinate descent with nonconvex penalties. Journal of the American Statistical Association 106(495):1125–1138, DOI 10.1198/jasa.2011.tm09738

Mazumder R, Radchenko P, Dedieu A (2017) Subset selection with shrinkage: Sparse linear modeling when the SNR is low. ArXiv e-prints 1708.03288v1

Meyer CD (2000) Matrix analysis and applied linear algebra. SIAM

Miller A (1990) Subset Selection in Regression. Chapman and Hall, Melbourne

Miyashiro R, Takano Y (2015) Mixed integer second-order cone programming formulations for variable selection in linear regression. European Journal of Operational Research 247(3):721–731, DOI 10.1016/J.EJOR.2015.06.081

Robinson PD, Wathen AJ (1992) Variational bounds on the entries of the inverse of a matrix. IMA Journal of Numerical Analysis 12(4):463–486, DOI 10.1093/imanum/12.4.463

Schoofs AJG, van Houten MH, Etman LFP, van Campen DH (1997) Global and mid-range function approximation for engineering optimization. Mathematical Methods of Operations Research 46(3):335–359, DOI 10.1007/BF01194860

Schwarz G (1978) Estimating the dimension of a model. The Annals of Statistics 6(2):461–464, DOI 10.1214/aos/1176344136

Seber GAF (1977) Linear Regression Analysis. John Wiley and Sons

Takano Y, Miyashiro R (2019) Best subset selection via cross-validation criterion. Optimization Online preprint URL http://www.optimization-online.org/DB_FILE/2019/01/7028.pdf

Tibshirani R (1996) Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society Series B (Methodological) 58(1):267–288

Tillmann AM, Pfetsch ME (2014) The computational complexity of the restricted isometry property, the nullspace property, and related concepts in compressed sensing. IEEE Transactions on Information Theory 60(2):1248–1259, DOI 10.1109/TIT.2013.2290112