RANK-ONE CONVEXIFICATION FOR SPARSE REGRESSION

ALPER ATAMTÜRK AND ANDRÉS GÓMEZ

Abstract. Sparse regression models are increasingly prevalent due to their ease of interpretability and superior out-of-sample performance. However, the exact model of sparse regression with an $\ell_0$ constraint restricting the support of the estimators is a challenging ($NP$-hard) non-convex optimization problem. In this paper, we derive new strong convex relaxations for sparse regression. These relaxations are based on the ideal (convex-hull) formulations for rank-one quadratic terms with indicator variables. The new relaxations can be formulated as semidefinite optimization problems in an extended space and are stronger and more general than the state-of-the-art formulations, including the perspective reformulation and formulations with the reverse Huber penalty and the minimax concave penalty functions. Furthermore, the proposed rank-one strengthening can be interpreted as a non-separable, non-convex, unbiased sparsity-inducing regularizer, which dynamically adjusts its penalty according to the shape of the error function without inducing bias for the sparse solutions. In our computational experiments with benchmark datasets, the proposed conic formulations are solved within seconds and result in near-optimal solutions (with 0.4% optimality gap) for non-convex $\ell_0$-problems. Moreover, the resulting estimators also outperform alternative convex approaches from a statistical perspective, achieving high prediction accuracy and good interpretability.

Keywords Sparse regression, best subset selection, lasso, elastic net, conic formulations, non-convex regularization

January 2019
1. Introduction

Given a model matrix \( X = [x_1, \ldots, x_p] \in \mathbb{R}^{n \times p} \) of explanatory variables, a vector \( y \in \mathbb{R}^n \) of response variables, regularization parameters \( \lambda, \mu \geq 0 \) and a desired sparsity \( k \in \mathbb{Z}^+ \), we consider the least squares regression problem

\[
\min_{\beta \in \mathbb{R}^p} \| y - X \beta \|_2^2 + \lambda \| \beta \|_2^2 + \mu \| \beta \|_1 \quad \text{s.t.} \quad \| \beta \|_0 \leq k,
\]

where \( \| \beta \|_0 \) denotes cardinality of the support of \( \beta \). Problem (1) encompasses a broad range of the regression models. It includes as special cases: ridge regression [24], when \( \lambda > 0, \mu = 0 \) and \( k \geq p \); lasso [40], when \( \lambda = 0, \mu \geq 0 \) and \( k \geq p \); elastic net [51] when \( \lambda, \mu > 0 \) and \( k \geq p \); best subset selection [33], when \( \lambda = 0, \mu > 0 \) and \( k < p \). Additionally, Bertsimas and Van Parys [8] propose to solve (1) with \( \lambda > 0, \mu = 0 \) and \( k < p \) for high-dimensional regression problems, while Mazumder et al. [32] study (1) with \( \lambda = 0, \mu > 0 \) and \( k < p \) for problems with low Signal-to-Noise Ratios (SNR). The results in this paper cover all versions of (1) with \( k < p \); moreover, they can be extended to problems with non-separable regularizations of the form \( \lambda \| A \beta \|_2^2 + \mu \| C \beta \|_1 \), resulting in sparse variants of the fused lasso [37, 41], generalized lasso [30, 42] and smooth lasso [23], among others.

Regularization techniques. The motivation and benefits of the regularization are well-documented in the literature. Hastie et al. [20] coined the bet on sparsity principle, i.e., using an inference procedure that performs well in sparse problems since no procedure can do well in dense problems. Best subset selection with \( k < p \) and \( \lambda = \mu = 0 \) is the direct approach to enforce sparsity without incurring bias. In contrast, ridge regression with \( \lambda > 0 \) (Tikhonov regularization) is known to induce shrinkage and bias, which can be desirable, for example, when \( X \) is not orthogonal, but it does not result in sparsity. On the other hand, lasso, the \( \ell_1 \) regularization with \( \mu > 0 \) simultaneously causes shrinkage and induces sparsity, but the inability to separately control for shrinkage and sparsity may result in subpar performance in some cases [33, 46, 47, 48, 49, 50]. Moreover, achieving a target sparsity level \( k \) with lasso requires significant experimentation with the penalty parameter \( \mu \) [10]. When \( k \geq p \), the cardinality constraint on \( \ell_0 \) is redundant and (1) reduces to a convex optimization problem and can be solved easily. On the other hand, when \( k < p \), problem (1) is non-convex and \( NP \)-hard [35], thus finding an optimal solution may require excessive computational effort and methods to solve it approximately are used instead [25, 36]. Due to the perceived difficulties of tackling the non-convex \( \ell_0 \) constraint in (1), lasso-type simpler approaches are still preferred for inference problems with sparsity [22].

Nonetheless, there has been a substantial effort to develop sparsity-inducing methodologies that do not incur as much shrinkage and bias as lasso does. The resulting techniques often result in optimization problems of the form

\[
\min_{\beta \in \mathbb{R}^p} \| y - X \beta \|_2^2 + \sum_{i=1}^{p} \rho_i(\beta_i)
\]

where \( \rho_i : \mathbb{R} \to \mathbb{R} \) are non-convex regularization functions. Examples of such regularization functions include \( \ell_q \) penalties with \( 0 < q < 1 \) [17] and SCAD [15]. Although optimal solutions of (2) with non-convex regularizations may substantially improve upon the estimators obtained by lasso, solving (2) to optimality
is still a difficult task [26, 31, 52], and suboptimal solutions may not benefit from the improved statistical properties. To address such difficulties, Zhang et al. [45] propose the minimax concave penalty ($\text{MC}_+$), a class of sparsity-inducing penalty functions where the non-convexity of $\rho$ is offset by the convexity of $\|y - X\beta\|_2^2$ for sufficiently sparse solutions, so that (2) remains convex – Zhang et al. [45] refer to this property as sparse convexity. Thus, in the ideal scenario (and with proper tuning of the parameter controlling the concavity of $\rho$), the $\text{MC}_+$ penalty is able to retain the sparsity and unbiasedness of best subset selection while preserving convexity, resulting in the best of both worlds. However, due to the separable form of the regularization term, the effectiveness of $\text{MC}_+$ greatly depends on the diagonal dominance of the matrix $X^\top X$ (this statement will be made more precise in §3), and may result in poor performance when the diagonal dominance is low.

Unfortunately, in many practical applications, the matrix $X^\top X$ has low eigenvalues and is not diagonally dominant at all. To illustrate, Table 1 presents the diagonal dominance of five datasets from the UCI Machine Learning Repository [12] used in [18, 34], as well as the diabetes dataset with all second interactions used in [7, 14]. The diagonal dominance of a positive semidefinite matrix $A$ is computed as

$$\text{dd}(A) := \left(\frac{1}{\text{tr}(A)}\right) \max_{d \in \mathbb{R}_+^p} e^\top d \text{ s.t. } A - \text{diag}(d) \succeq 0,$$

where $e$ is the $p$-dimensional vector of ones, $\text{diag}(d)$ is the diagonal matrix such that $\text{diag}(d)_i = d_i$ and $\text{tr}(A)$ denotes the trace of $A$. Accordingly, the diagonal dominance is the trace of the largest diagonal matrix that can be extracted from $A$ without violating positive semidefiniteness, divided by the trace of $A$. Observe in Table 1 that the diagonal dominance of $X^\top X$ is very low or even 0%, and $\text{MC}_+$ struggles for these datasets as we demonstrate in §5.

**Mixed-integer optimization formulations.** An alternative to utilizing non-convex regularizations is to leverage the recent advances in mixed-integer optimization (MIO) to tackle (1) exactly [6, 7, 11]. By introducing indicator variables
\( z \in \{0,1\}^p \), where \( z_i = 1_{\beta_i \neq 0} \), problem (1) can be reformulated as

\[
\begin{align*}
\min & \quad y^\top y + (-2y^\top X \beta + \beta^\top (X^\top X + \lambda I) \beta + \mu \sum_{i=1}^p u_i) \\
\text{s.t.} & \quad \sum_{i=1}^p z_i \leq k \quad (3a) \\
& \quad \beta_i \leq u_i, -\beta_i \leq u_i \quad i = 1, \ldots, p \quad (3b) \\
& \quad \beta_i (1 - z_i) = 0 \quad i = 1, \ldots, p \quad (3c) \\
& \quad \beta \in \mathbb{R}^p, z \in \{0,1\}^p, u \in \mathbb{R}^p_+ . \quad (3e)
\end{align*}
\]

The non-convexity of (1) is captured by the complementary constraints (3d) and the integrality constraints \( z \in \{0,1\}^p \). In fact, one of the main challenges for solving (3) is handling constraints (3d). A standard approach in the MIO literature is to use the so-called big-\( M \) constraints and replace (3d) with

\[
-M z_i \leq \beta_i \leq M z_i \quad (4)
\]

for a sufficiently large number \( M \) to bound the variables \( \beta_i \). However, these so-called big-\( M \) constraints (4) are poor approximations of constraints (3d), especially in the case of regression problems where no natural big-\( M \) value is available. Bertsimas et al. [7] propose approaches to compute provable big-\( M \) values, but such values often result in prohibitively large computational times even in problems with a few dozens variables (or, even worse, may lead to numerical instabilities and cause convex solvers to crash). Alternatively, heuristic values for the big-\( M \) values can be estimated, e.g., setting \( M = \tau \|\hat{\beta}\|_\infty \) where \( \tau \in \mathbb{R}_+ \) and \( \hat{\beta} \) is a feasible solution of (1) found via a heuristic.\(^1\) While using such heuristic values yield reasonable performance for small enough values of \( \tau \), it may eliminate optimal solutions.

Branch-and-bound algorithms for MIO leverage strong convex relaxations of problems to prune the search space and reduce the number of sub-problems to be enumerated (and, in some cases, eliminate the need for enumeration altogether). Thus, a critical step to speed-up the solution times for (3) is to derive convex relaxations that approximate the non-convex problem well [5]. Such strong relaxations can also be used directly to find good estimators for the inference problems (without branch-and-bound): in fact, it is well-known that the natural convex relaxation of (3) with \( \lambda = \mu = 0 \) and big-\( M \) constraints is precisely \texttt{lasso}, see [13] for example. Therefore, sparsity-inducing techniques that more accurately capture the properties of the non-convex constraint \( \|\beta\|_0 \leq k \) can be found by deriving tighter convex relaxations of (1). Pilanci et al. [38] exploit the Tikhonov regularization term and convex analysis to construct an improved convex relaxation using the reverse Huber penalty. In a similar vein, Bertsimas and Van Parys [8] leverage the Tikhonov regularization and duality to propose an efficient algorithm for high-dimensional sparse regression.

\(^1\)This method with \( \tau = 2 \) was used in the computations in [7].
used to substantially strengthen the convex relaxations by exploiting separable quadratic terms. Specifically, consider the mixed-integer epigraph of a one-dimensional quadratic function with an indicator constraint,
\[ Q_1 = \{ z \in \{0, 1\}, \beta \in \mathbb{R}, t \in \mathbb{R}_+ : \beta_i^2 \leq t, \beta_i (1 - z_i) = 0 \} \cdot \]
The convex hull of \( Q_1 \) is obtained by relaxing the integrality constraint to bound constraints and using the closure of the perspective function\(^2\) of \( \beta_i^2 \), expressed as a rotated cone constraint:
\[ \text{conv}(Q_1) = \{ z \in [0, 1], \beta \in \mathbb{R}, t \in \mathbb{R}_+ : \frac{\beta_i^2}{z_i} \leq t \} \cdot \]

Xie and Deng \[44\] apply the perspective relaxation to the separable quadratic regularization term \( \lambda \| \beta \|_2^2 \), i.e., reformulate (3) as
\[
\begin{align*}
y^\top y + \min & -2y^\top X\beta + \beta^\top (X^\top X) \beta + \lambda \sum_{i=1}^p \frac{\beta_i^2}{z_i} + \mu \sum_{i=1}^p u_i \quad (5a) \\
\text{s.t.} & \sum_{i=1}^p z_i \leq k \quad (5b) \\
& \beta_i \leq u_i, -\beta_i \leq u_i \quad i = 1, \ldots, p \quad (5c) \\
& \beta \in \mathbb{R}^p, z \in \{0, 1\}^p, u \in \mathbb{R}_+^p. \quad (5d)
\end{align*}
\]
Moreover, they show that the continuous relaxation of (5) is equivalent to the continuous regularization term \( \lambda \| \beta \|_2^2 \), i.e., reformulate (3) as
\[
\begin{align*}
y^\top y + \min & -2y^\top X\beta + \beta^\top (X^\top X) \beta + \lambda \sum_{i=1}^p \frac{\beta_i^2}{z_i} + \mu \sum_{i=1}^p u_i \quad (5a) \\
\text{s.t.} & \sum_{i=1}^p z_i \leq k \quad (5b) \\
& \beta_i \leq u_i, -\beta_i \leq u_i \quad i = 1, \ldots, p \quad (5c) \\
& \beta \in \mathbb{R}^p, z \in \{0, 1\}^p, u \in \mathbb{R}_+^p \cdot \quad (5d)
\end{align*}
\]

Among these approaches the optimal perspective relaxation of Dong et al. \[13\] is the only one that does not explicitly require the use of the Tikhonov regularization \( \lambda \| \beta \|_2^2 \). Nonetheless, as the authors point out, if \( \lambda = 0 \) then the method is effective only when the matrix \( X^\top X \) is sufficiently diagonally dominant, which, as illustrated in Table 1, is not necessarily the case in practice. As a consequence, perspective relaxation techniques may be insufficient to tackle problems when large shrinkage is undesirable and, hence, \( \lambda \) is small.

**Our contributions.** In this paper we derive stronger convex relaxations of (3) than the optimal perspective relaxation. These relaxations are obtained from the study of ideal (convex-hull) formulations of the mixed-integer epigraphs of non-separable rank-one quadratic functions with indicators. Since the perspective relaxation corresponds to the ideal formulation of a one-dimensional rank-one quadratic function, the proposed relaxations generalize and strengthen the existing results. In particular, they dominate perspective relaxation approaches for all values of the regularization parameter \( \lambda \) and, critically, are able to achieve

\[ \text{We use the convention that } \frac{\beta_i^2}{z_i} = 0 \text{ when } \beta_i = z_i = 0 \text{ and } \frac{\beta_i^2}{z_i} = \infty \text{ if } z_i = 0 \text{ and } \beta_i \neq 0. \]
high-quality approximations of (1) even in low diagonal dominance settings with \( \lambda = 0 \). Alternatively, our results can also be interpreted as a new non-separable, non-convex, unbiased regularization penalty \( \rho_{\text{MC}}(\beta) \) which: (i) imposes larger penalties than the separable minimax concave penalty \([45]\) \( \rho_{\text{MC}+}(\beta) \) to dense estimators, thus achieving better sparsity-inducing properties; and (ii) the nonconvexity of the penalty function is offset by the convexity of the term \( \|y - X\beta\|_2^2 \), and the resulting continuous problem can be solved to global optimality using convex optimization tools. In fact, they can be formulated as semidefinite optimization and, in certain special cases, as conic quadratic optimization.

To illustrate the regularization point of view for the proposed relaxations, consider a two-predictor regression problem in Lagrangean form:

\[
\min_{\beta \in \mathbb{R}^2} \|y - X\beta\|_2^2 + \lambda \|\beta\|_2^2 + \mu \|\beta\|_1 + \kappa \|\beta\|_0, \tag{6}
\]

where \( X^T X = \begin{pmatrix} 1 + \delta & 1 + \delta \\ 1 & 1 \end{pmatrix} \) and \( \delta \geq 0 \) is a parameter controlling the diagonal dominance. Figure 1 depicts the graphs of well-known regularizations including lasso (\( \lambda = \kappa = 0, \mu = 1 \)), ridge (\( \mu = \kappa = 0, \lambda = 1 \)), elastic net (\( \kappa = 0, \lambda = \mu = 0.5 \)), the MC\(_+\) penalty for different values of \( \delta \) and the proposed rank-one R1 regularization. The graphs of MC\(_+\) and R1 are obtained by setting \( \lambda = \mu = 0 \) and \( \kappa = 1 \), and using the appropriate convex strengthening, see \S3 for details. Observe that the R1 regularization results in larger penalties than MC\(_+\) for all values of \( \delta \), and the improvement increases as \( \delta \to 0 \). In addition, Figure 2 shows the effect of using the lasso constraint \( \|\beta\|_1 \leq k \), the MC\(_+\) constraint \( \rho_{\text{MC}}(\beta) \leq k \), and the rank-one constraint \( \rho_{\text{R1}}(\beta) \leq k \) in a two-dimensional problem to achieve sparse solutions satisfying \( \|\beta\|_0 \leq 1 \). Specifically, let

\[
\epsilon^* = \min_{\|\beta\|_0 \leq 1} \|y - X\beta\|_2^2
\]

be the minimum residual error of a sparse solution of the least squares problem. Figure 2 shows in gray the (possibly dense) points satisfying \( \|y - X\beta\|_2^2 \leq \epsilon^* \), and it shows in color the set of feasible points satisfying \( \rho(\beta) \leq k \), where \( \rho \) is a given regularization and \( k \) is chosen so that the feasible region (color) intersects the level sets (gray). We see that neither lasso nor MC\(_+\) is able to exactly recover an optimal sparse solution for any diagonal dominance parameter \( \delta \), despite significant shrinkage \( (k < 1) \). In contrast, the rank-one constraint \( \rho_{\text{R1}}(\beta) \leq k \) adapts to the curvature of the error function \( \|y - X\beta\|_2^2 \) to induce higher sparsity: in particular, the “natural” constraint \( \rho_{\text{R1}}(\beta) \leq 1 \), with the target sparsity \( k = 1 \), results in exact recovery without shrinkage in all cases.

Outline. The rest of the paper is organized as follows. In \S2 we derive the proposed convex relaxations based on ideal formulations for rank-one quadratic terms with indicator variables. We also give an interpretation of the convex relaxations as unbiased regularization penalties, and we give an explicit semidefinite optimization (SDP) formulation in an extended space, which can be implemented with off-the-shelf conic optimization solvers. In \S3 we derive an explicit form of the regularization penalty for the two-dimensional case. In \S4 we discuss the implementation of the proposed relaxation in a conic quadratic framework. In \S5 we present computational experiments with synthetic as well as benchmark datasets, demonstrating that (i) the proposed formulation delivers near-optimal solutions (with provable
Figure 1. Graphs of regularization penalties with $p = 2$. The ridge, elastic net, and lasso (top row) regularizations do not depend on the diagonal dominance but induce substantial bias. The $\mathcal{MC}_+$ regularization (second row) does not induce as much bias, but it depends on the diagonal dominance ($\delta$). The new non-separable, non-convex $\mathcal{R}_1$ regularization (bottom row) induces larger penalties than $\mathcal{MC}_+$ for all diagonal dominance values and is a closer approximation for the exact $\ell_0$ penalty.
Figure 2. The axes correspond to the sparse solutions satisfying $\|\beta\|_0 \leq 1$. In gray: level sets given by $\|y - X\beta\|_2^2 \leq \epsilon^*$; in red: feasible region for $\|\beta\|_1 \leq k$; in green: feasible region for $\mu_{nc}(\beta) \leq k$; in blue: feasible region for $\rho_{mc}(\beta) \leq k$. All lasso and MC solutions above are dense even with significant shrinkage ($k < 1$). Rank-one constraint attains sparse solutions on the axes with no shrinkage ($k = 1$) for all diagonal dominance values $\delta$. 
optimality gaps) of (1) in most cases, (ii) using the proposed convex relaxation results in superior statistical performance when compared with usual estimators obtained from convex optimization approaches. In §6 we conclude the paper with a few final remarks.

**Notation.** Define $P = \{1, \ldots, p\}$ and $e \in \mathbb{R}^p$ be the vector of ones. Given $T \subseteq P$ and a vector $a \in \mathbb{R}^p$, define $a_T$ as the subvector of $a$ induced by $T$, $a_i = a_{\{i\}}$ as the $i$-th element of $a$, and define $a(T) = \sum_{i \in T} a_i$. Given a symmetric matrix $A \in \mathbb{R}^{P \times P}$, let $A_T$ be the submatrix of $A$ induced by $T \subseteq P$, and let $S^T_+$ be the set of $T \times T$ symmetric positive semidefinite matrices, i.e., $A_T \succeq 0 \iff A_T \in S^T_+$. We use $a_T$ or $A_T$ to make explicit that a given vector or matrix belongs to $\mathbb{R}^T$ or $\mathbb{R}^{T \times T}$, respectively. Given matrices $A$, $B$ of the same dimension, $A \circ B$ denotes the Hadamard product of $A$ and $B$, and $\langle A, B \rangle$ denotes their inner product. Given a vector $a \in \mathbb{R}^n$, let $\text{diag}(a)$ be the $n \times n$ diagonal matrix $A$ with $A_{ii} = a_i$. For a set $X \subseteq \mathbb{R}^p$, $\text{conv}(X)$ denotes the closure of the convex hull of $X$. Throughout the paper, we adopt the following convention for division by 0: given a scalar $s \geq 0$, $s/0 = \infty$ if $s > 0$ and $s/0$ if $s = 0$. For a scalar $a \in \mathbb{R}$, let $\text{sign}(a) = a/|a|$.

2. Convexification

In this section we introduce the proposed relaxations of problem (1). First, in §2.1, we describe the ideal relaxations for the mixed-integer epigraph of a rank-one quadratic term. Then, in §2.2, we use the relaxations derived in §2.1 to give strong relaxations of (1). Next, in §2.3, we give an interpretation of the proposed relaxations as unbiased sparsity-inducing regularizations. Finally, in §2.4 we present an explicit SDP representation of the proposed relaxations in an extended space.

2.1. Rank-one case. We first give a valid inequality for the mixed-integer epigraph of a convex quadratic function defined over the subsets of $P$. Given $A_T \in S^T_+$, consider the set

$$Q_T = \left\{ (z, \beta, t) \in \{0, 1\}^T \times \mathbb{R}^T \times \mathbb{R}_+ : \beta^T A_T \beta \leq t, \beta_i (1 - z_i) = 0, \forall i \in T \right\}.$$

**Proposition 1.** The inequality

$$\frac{\beta^T A_T \beta}{z(T)} \leq t$$

is valid for $Q_T$.

**Proof.** Let $(z, \beta, t) \in Q_T$, and we verify that inequality (7) is satisfied. First observe that if $z = 0$, then $\beta = 0$ and inequality (7) reduces to $0 \leq t$, which is satisfied. Otherwise, if $z_i = 1$ for some $i \in T$, then $z(T) \geq 1$ and we find that $\frac{\beta^T A_T \beta}{z(T)} \leq \beta^T A_T \beta \leq t$, and inequality (7) is satisfied again. 

Observe that if $T$ is a singleton, i.e., $T = \{i\}$, then (7) reduces to the well-known perspective inequality $A_{ii} \beta_i^2 \leq tz_i$. Moreover, if $T = \{i, j\}$ and $A_T$ is rank-one, i.e., $\beta^T A_T \beta = |A_{ij}| \left( a \beta_i^2 + 2 \beta_i \beta_j + (1/a) \beta_j^2 \right)$ for $A_{ij} \neq 0$ and scalar $a > 0$, then (7) reduces to

$$|A_{ij}| \left( a \beta_i^2 + 2 \beta_i \beta_j + (1/a) \beta_j^2 \right) \leq t(z_i + z_j),$$

one of the inequalities proposed in [27] in the context of quadratic optimization with indicators and bounded continuous variables. Note that inequality (8) is, in general, weak for bounded continuous variables (as non-negativity or other bounds
can be used to strengthen the inequalities, see [3] for additional discussion; and inequality (7) is, in general, weak for arbitrary matrices \( A_T \in \mathcal{S}_T^+ \). Nonetheless, as we show next, inequality (7) is sufficient to describe the ideal (convex hull) description for \( Q_T \) if \( A_T \) is a rank-one matrix. Consider the special case of \( Q_T \) defined with a rank-one matrix:

\[
Q_T^1 = \{(z, \beta, t) \in \{0,1\}^T \times \mathbb{R}^T \times \mathbb{R}_+: (a_T^T \beta)^2 \leq t, \beta_i(1-z_i) = 0, \forall i \in T\}.
\]

**Theorem 1.** If \( a_i \neq 0 \) for all \( i \in T \), then

\[
\text{conv}(Q_T^1) = \left\{(z, \beta, t) \in [0,1]^T \times \mathbb{R}^T \times \mathbb{R}_+: (a_T^T \beta)^2 \leq t, \frac{(a_T^T \beta)^2}{z(T)} \leq t\right\}.
\]

**Proof.** Consider the optimization of an arbitrary linear function over \( Q_T^1 \) and \( \bar{Q}_T := \{(z, \beta, t) \in [0,1]^T \times \mathbb{R}^T \times \mathbb{R}_+: (a_T^T \beta)^2 \leq t, \frac{(a_T^T \beta)^2}{z(T)} \leq t\} :\)

\[
\begin{align*}
\min_{(z, \beta, t) \in Q_T^1} & \quad u_T^T z + v_T^T \beta + \kappa t, \\
\min_{(z, \beta, t) \in \bar{Q}_T} & \quad u_T^T z + v_T^T \beta + \kappa t,
\end{align*}
\]

where \( u_T, v_T \in \mathbb{R}^T \) and \( \kappa \in \mathbb{R} \). We now show that either there exists an optimal solution of (10) that is feasible for (9), hence also optimal for (9) as \( \bar{Q}_T \) is a relaxation of \( Q_T^1 \), or that (9) and (10) are both unbounded.

Observe that if \( \kappa < 0 \), then letting \( z = \beta = 0 \) and \( t \to \infty \) we see that both problems are unbounded. If \( \kappa = 0 \) and \( v_T = 0 \), then (10) reduces to \( \min_{z \in [0,1]^T} u_T^T z \), which has an optimal integral solution \( z^* \), and \( (z^*, 0, 0) \) is optimal for (9) and (10). If \( \kappa = 0 \) and \( v_i \neq 0 \) for some \( i \in T \), then letting \( \beta_i \to \pm \infty \), \( z_i = 1 \), and \( \beta_j = z_j = t = 0 \) for \( j \neq i \), we find that both problems are unbounded. Thus, we may assume, without loss of generality that \( \kappa > 0 \), and, by scaling, \( \kappa = 1 \).

Additionally, as \( a_T \) has no zero entry, we may assume, without loss of generality, that \( a_T \) can be scaled by letting \( \tilde{\beta}_i = a_i \beta_i \) and \( \tilde{v}_i = v_i/a_i \) to arrive at an equivalent problem. Moreover, a necessary condition for (9)–(10) to be bounded is that

\[
-\infty < \min_{\beta \in \mathbb{R}^T} v_T^T \beta \text{ s.t. } \beta(T) = \zeta
\]

for any fixed \( \zeta \in \mathbb{R} \). It is easily seen that (11) has an optimal solution if and only if \( v_i = v_j \) for all \( i \neq j \). Thus, we may also assume without loss of generality that \( v_T^T \beta = v_0 \beta(T) \) for some scalar \( v_0 \). Performing the above simplifications, we find that (10) reduces to

\[
\min_{z \in [0,1]^T, \beta \in \mathbb{R}^T, t \in \mathbb{R}} u_T^T z + v_0 \beta(T) + t \text{ s.t. } \beta(T)^2 \leq t, \beta(T)^2 \leq tz(T).
\]

Since the one-dimensional optimization \( \min_{\beta \in \mathbb{R}} \{v_0 \beta + \beta^2\} \) has an optimal solution, it follows that (12) is bounded and has an optimal solution. We now prove that (12) has an optimal solution that is integral in \( z \) and satisfies \( \beta \circ (e-z) = 0 \).

Let \( (z^*, \beta^*, t^*) \) be an optimal solution of (12). First note that if \( 0 < z^*(T) < 1 \), then \( (\gamma z^*, \gamma \beta^*, \gamma t^*) \) is feasible for (10) for \( \gamma \) sufficiently close to 1, with objective value \( \gamma (u_T^T z^* + v_0 \beta^*(T) + t^*) \). If \( u_T^T z^* + v_0 \beta^*(T) + t^* \geq 0 \), then for \( \gamma = 0 \), \( (\gamma z^*, \gamma \beta^*, \gamma t^*) \) has an objective value equal or lower. Otherwise, for \( \gamma = 1/z^*(T) \), \( (\gamma z^*, \gamma \beta^*, \gamma t^*) \) is feasible and has a lower objective value. Thus, we find that either \( 0 \) is optimal for (12) (and the proof is complete), or there exists an optimal
solution with \( z^*(T) \geq 1 \). In the later case, observe that any \((\bar{z}, \bar{\beta}', t^*)\) with \( \bar{z} \in \arg \min \{ u_T^T z : z^*(T) \geq 1, z \in [0, 1]^n \}\) is also optimal for \((12)\), an in particular there exists an optimal solution with \( \bar{z} \) integral.

Finally, let \( i \in T \) be any index with \( \bar{z}_i = 1 \). Setting \( \bar{\beta}_i = \beta^*(T) \) and \( \bar{\beta}_j = 0 \) for \( i \neq j \), we find another optimal solution \((\bar{z}, \bar{\beta}, t^*)\) for \((12)\) that satisfies the complementary constraints, and thus is feasible and optimal for \((9)\). \( \square \)

**Remark 1.** Observe that describing \( \text{conv}(Q_T^{1}) \) requires two nonlinear inequalities in the original space of variables. More compactly, we can specify \( \text{conv}(Q_T^{1}) \) using a single convex inequality, as

\[
\text{conv}(Q_T^{1}) = \left\{ (z, \beta, t) \in [0, 1]^T \times \mathbb{R}^T \times \mathbb{R}_+ : \frac{(a_T^T \beta)^2}{\min\{1, z(T)\}} \leq t \right\}.
\]

Finally, we point out that \( \text{conv}(Q_T^{1}) \) is conic quadratic representable, as \((z, \beta, t) \in \text{conv}(Q_T^{1})\) if and only if there exists \( w \) such that the system

\[
z \in [0, 1]^T, \quad \beta \in \mathbb{R}^P, \quad t \in \mathbb{R}_+, \quad w \in \mathbb{R}_+, \quad w \leq 1, \quad w \leq z(T), \quad (a_T^T \beta)^2 \leq tw
\]

is feasible, where the last constraint is a rotated conic quadratic constraint and all other constraints are linear. \( \square \)

2.2. **General case.** Now consider again the mixed-integer optimization (3)

\[
\begin{align*}
y^T y + \min & -2y^T X \beta + \mu (e^T u) + t \\
\text{s.t.} & \beta^T \left( X^T X + \lambda I \right) \beta \leq t \quad (13a) \\
& e^T z \leq k \quad (13c) \\
& \beta \leq u, \quad -\beta \leq u \quad (13d) \\
& \beta \circ (e - z) = 0 \quad (13e) \\
& \beta \in \mathbb{R}^P, \quad z \in \{0, 1\}^P, \quad u \in \mathbb{R}_+^P, \quad t \in \mathbb{R} \quad (13f)
\end{align*}
\]

where the nonlinear terms of the objective is moved to constraint \((13b)\). A direct application of \((7)\) yields the inequality \( \beta^T \left( X^T X + \lambda I \right) \beta \leq tz(P) \), which is weak and has no effect when \( z(P) \geq 1 \). Instead, a more effective approach is to decompose the matrix \( X^T X + \lambda I \) into a sum of low-dimensional rank-one matrices, and use inequality \((7)\) to strengthen each quadratic term in the decomposition separately, as illustrated in Example 1 bellow.

**Example 1.** Consider the example with \( p = 3 \) and \( X^T X + \lambda I = \begin{pmatrix} 25 & 15 & -5 \\ 15 & 18 & 0 \\ -5 & 0 & 11 \end{pmatrix} \).

Then, it follows that

\[
\beta^T \left( X^T X + \lambda I \right) \beta = (5\beta_1 + 3\beta_2 - \beta_3)^2 + (3\beta_2 + \beta_3)^2 + 9\beta_3^2
\]

and we have the corresponding valid inequality

\[
\frac{(5\beta_1 + 3\beta_2 - \beta_3)^2}{\min\{1, z_1 + z_2 + z_3\}} + \frac{(3\beta_2 + \beta_3)^2}{\min\{1, z_2 + z_3\}} + 9\frac{\beta_3^2}{z_3} \leq t. \quad (14)
\]

\( \square \)
The decomposition of $X^TX + \lambda I$ illustrated in Example 1 is not unique; therefore, we want to use the decomposition that results in the best convex relaxation, i.e., that maximizes the left hand side of (14). Specifically, let $P \subseteq 2^P$ be a subset of the power set of $P$, and for $A_T \in \mathbb{R}^{T \times T}$ define $\tilde{A}_T$ as the $P \times P$ matrix obtained by filling the missing entries by zeros. Consider the valid inequality $\phi_P(z, \beta) \leq t$, where $\phi_P : [0, 1]^P \times \mathbb{R}^P \to \mathbb{R}$ is defined as

$$\phi_P(z, \beta) := \max_{A_T \in P} \beta^T R \beta + \sum_{T \in P} \frac{\beta^T A_T \beta_T}{\min \{1, z(T)\}}$$

s.t. $\sum_{T \in P} \tilde{A}_T + R = X^TX + \lambda I$ \hspace{1cm} (15a)

$A_T \in S^T_+$ \hspace{1cm} (15b)

$R \in S^P_+$ \hspace{1cm} (15c)

where strengthening (7) is applied to each low-dimensional quadratic term $\beta^T A_T \beta_T$. For a fixed value of $(z, \beta)$, problem (15) finds the best decomposition of the matrix $X^TX + \lambda I$ as a sum of positive semidefinite matrices $\tilde{A}_T$, $T \in P$, and a remainder positive semidefinite matrix $R$ to maximize the strengthening.

For a given decomposition, the objective (15a) is convex in $(z, \beta)$, thus $\phi_P$ is a supremum of convex functions and is convex on its domain. Observe that the inclusion or omission of the empty set does not affect function $\phi_P$, and we assume for simplicity that $\emptyset \in P$.

Since inequalities (7) are ideal for rank-one matrices, inequality $\phi_P(z, \beta) \leq t$ is particularly strong if matrices $A_T$ are rank-one in optimal solutions of (15). As we now show, this is indeed the case if $P$ is downward closed.

**Proposition 2.** If $P$ is downward closed, i.e., $V \in P \implies U \in P$ for all $U \subseteq V$, then there exists an optimal solution to (15) where all matrices $A_T$ are rank-one.

**Proof.** Let $T \in P$, suppose $A_T$ is not rank-one in an optimal solution to (15), also suppose for simplicity that $T = \{1, \ldots, p_0\}$ for some $p_0 \leq p$, and let $\tilde{T}_i = \{i, \ldots, p_0\}$ for $i = 1, \ldots, p_0$. Since $A_T$ is positive semidefinite, there exists a Cholesky decomposition $A_T = LL^T$ where $L$ is a lower triangular matrix (possibly with zeros on the diagonal if $A_T$ is not positive definite). Let $L_i$ denote the $i$-the column of $L$. Since $A_T$ is not a rank-one matrix, there exist at least two non-zero columns of $L$. Let $L_j$ with $j > 1$ be the second non-zero column. Then

$$\frac{\beta^T A_T \beta_T}{\min \{1, z(T)\}} = \beta^T \left( \frac{\sum_{i \neq j} (L_i L_i^T)}{\min \{1, z(T)\}} \right) \beta_T + \frac{\beta^T (L_j L_j^T) \beta_T}{\min \{1, z(T)\}} \leq \beta^T \left( \frac{\sum_{i \neq j} (L_i L_i^T)}{\min \{1, z(T)\}} \right) \beta_T + \frac{\beta^T (L_j L_j^T) \beta_T}{\min \{1, z(T)\}}. \hspace{1cm} (16)$$

Finally, since $\tilde{T}_j \in P$, the (better) decomposition (16) is feasible for (15), and the proposition is proven. □

By dropping the complementary constraints (13e), replacing the integrality constraints $z \in \{0, 1\}^P$ with bound constraints $z \in [0, 1]^P$, and utilizing the convex
function $\phi_P$ to reformulate (13b), we obtain the convex relaxation of (1)

\[
\begin{align*}
\mathbf{y}^\top \mathbf{y} + \min_{\mathbf{y} \in \mathbb{R}^P} & -2\mathbf{y}^\top \mathbf{X}\beta + \mu (\mathbf{e}^\top \mathbf{u}) + \phi_P(z, \beta) & (17a) \\
\mathbf{e}^\top \mathbf{z} & \leq k & (17b) \\
\beta & \leq \mathbf{u}, -\beta \leq \mathbf{u} & (17c) \\
\beta & \in \mathbb{R}^P, \ z \in [0,1]^P, \ \mathbf{u} \in \mathbb{R}_+^P & (17d)
\end{align*}
\]

for a given $\mathcal{P} \subseteq 2^P$. In the next section, we give an interpretation of formulation (17) as a sparsity-inducing regularization penalty.

### 2.3. Interpretation as regularization

Note that the relaxation (17) can be rewritten as:

\[
\min_{\beta \in \mathbb{R}^P} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_2^2 + \mu \|\beta\|_1 + \rho_{\mathcal{R}_1}(\beta; k)
\]

where

\[
\rho_{\mathcal{R}_1}(\beta; k) := \min_{z \in [0,1]^P} \phi_P(z, \beta) - \beta^\top (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})\beta \text{ s.t. } \mathbf{e}^\top \mathbf{z} \leq k.
\]

is the (non-convex) rank-one regularization penalty. Observe that $\rho_{\mathcal{R}_1}(\beta; k)$ is the difference of two convex functions: the quadratic function $\beta^\top (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})\beta$ arising from the fitness term and the Tikhonov regularization; and the projection of its convexification $\phi_P(z, \beta)$ in the original space of the regression variables $\beta$. As we now show, unlike the usual $\ell_1$ penalty, the rank-one regularization penalty does not induce a bias when $\beta$ is sparse.

**Theorem 2.** If $\|\beta\|_0 \leq k$, then $\rho_{\mathcal{R}_1}(\beta; k) = 0$.

**Proof.** Let $(\beta, z) \in \mathbb{R}^P \times [0,1]^P$, and let $\mathbf{R}$ and $\mathbf{A}_T$, $T \in \mathcal{P}$, correspond to an optimal solution of (15). Since

\[
\beta^\top (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})\beta = \beta^\top \mathbf{R}\beta + \sum_{T \in \mathcal{P}} \beta_T^\top \mathbf{A}_T \beta_T \\
\leq \beta^\top \mathbf{R}\beta + \sum_{T \in \mathcal{P}} \beta_T^\top \mathbf{A}_T \beta_T \min\{1, z(T)\} = \phi_P(z, \beta),
\]

it follows that $\rho_{\mathcal{R}_1}(\beta; k) \geq 0$ for any $\beta \in \mathbb{R}^P$. Now let $\hat{\beta}$ satisfy $\|\hat{\beta}\|_0 \leq k$, let $\hat{T} = \{i \in \mathcal{P} : \hat{\beta}_i \neq 0\}$ be the support of $\hat{\beta}$ and let $\hat{z}$ such that $\hat{z}_i = 1_{i \in \hat{T}}$ be the indicator vector of $\hat{T}$. By construction, $\mathbf{e}^\top \hat{z} \leq k$ and $\hat{z}$ is feasible for problem (18). Moreover

\[
\rho_{\mathcal{R}_1}(\hat{\beta}; k) \leq \phi_P(\hat{z}, \hat{\beta}) - \hat{\beta}^\top (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})\hat{\beta} \\
= \sum_{T \in \mathcal{P}} \sum_{T \cap \hat{T} \neq \emptyset} \left( \hat{\beta}_T^\top \mathbf{A}_T \hat{\beta}_T \min\{1, z(T)\} - \hat{\beta}_T^\top \mathbf{A}_T \hat{\beta}_T \right) = 0.
\]

Thus, $\rho_{\mathcal{R}_1}(\hat{\beta}; k) = 0$. \qed

The rank-one regularization penalty $\rho_{\mathcal{R}_1}$ can also be interpreted from an optimization perspective: note that problem (15) is the separation problem that, given
\((\beta, z) \in \mathbb{R}^P \times [0, 1]^P\), finds a decomposition that results in a most violated inequality after applying the rank-one strengthening. Thus, the regularization penalty \(\rho_{R1}(\beta; k)\) is precisely the violation of this inequality when \(z\) is chosen optimally.

In §3 we derive an explicit form of \(\rho_{R1}(\beta; k)\) when \(p = 2\); Figure 1 plots the graphs of the usual regularization penalties and \(\rho_{R1}\) for the two-dimensional case, and Figure 2 illustrates the better sparsity inducing properties of regularization \(\rho_{R1}\). Deriving explicit forms of \(\rho_{R1}\) is cumbersome for \(p \geq 3\). Fortunately, problem (17) can be explicitly reformulated in an extended space as an SDP and tackled using off-the-shelf conic optimization solvers.

### 2.4. Extended SDP formulation

To state the extended SDP formulation, in addition to variables \(z \in [0, 1]^P\) and \(\beta \in \mathbb{R}^P\), we introduce variables \(w \in [0, 1]^P\) corresponding to terms \(w^T := \min\{1, z(T)\}\) and \(B \in \mathbb{R}^{P \times P}\) corresponding to terms \(B_{ij} = \beta_i \beta_j\). Observe that for \(T \in \mathcal{P}\), \(\beta_T\) and \(B_T\) represent the subvector of \(\beta\) and submatrix of \(B\) induced by \(T\), whereas \(w_T\) is a scalar corresponding to the \(T\)-th coordinate of the \(|\mathcal{P}|\)-dimensional vector \(w\).

**Theorem 3.** Problem (17) is equivalent to the SDP

\[
\begin{align*}
\min_{y} & \quad y^T y + \min_{\beta, z} -2y^T X \beta + e^T u + (X^T X + \lambda I, B) \\
\text{s.t.} & \quad e^T z \leq k \\
& \quad \beta \leq u, -\beta \leq u \\
& \quad w_T \leq e^T_T z_T \quad \forall T \in \mathcal{P} \\
& \quad w_T B_T - \beta_T \beta_T^T \in S_+^T \quad \forall T \in \mathcal{P} \\
& \quad B - \beta \beta^T \in S_+^P \\
& \quad \beta \in \mathbb{R}^P, z \in [0, 1]^P, u \in \mathbb{R}_+^P, w \in [0, 1]^P, B \in \mathbb{R}^{P \times P}.
\end{align*}
\]  

Observe that (19) is indeed an SDP, as

\[
w_T B_T - \beta_T \beta_T^T \in S_+^T \iff \begin{pmatrix} w_T & \beta_T^T \\ \beta_T & B_T \end{pmatrix} \succeq 0;
\]

thus constraints (19e) and (19f) are indeed SDP-representable and the remaining constraints and objective are linear.

**Proof of Theorem 3.** It is easy to check that (19) is strictly feasible (set \(\beta = 0, z = e, w > 0\) and \(B = I\)). Adding surplus variables \(\Gamma, \Gamma_T\) write (19) as

\[
\begin{align*}
\min_{(\beta, z, u, w)} & \quad \min_{\beta, z} \left\{ -2y^T X \beta + e^T u + \min_{B, \Gamma_T, \Gamma} (X^T X + \lambda I, B) \right\} \\
\text{s.t.} & \quad w_T B_T - \Gamma_T = \beta_T \beta_T^T \quad \forall T \in \mathcal{P} \quad (A_T) \\
& \quad B - \Gamma = \beta \beta^T \quad (R) \\
& \quad \Gamma_T \in S_+^T \quad \forall T \in \mathcal{P} \\
& \quad \Gamma \in S_+^P \\
& \quad B \in \mathbb{R}^{P \times P}.
\end{align*}
\]
where \( C = \{ \beta \in \mathbb{R}^P, z \in [0, 1]^P, u \in \mathbb{R}^P, w \in [0, 1]^P : (19b), (19c), (19d) \} \). Using conic duality for the inner minimization problem, we find the dual

\[
y^\top y + \min_{(\beta, z, u, w) \in C} \left\{ -2y^\top X \beta + e^\top u + \max_{A_T, R} \langle \beta \beta^\top, R \rangle + \sum_{T \in \mathcal{P}} \langle \beta_T \beta_T^\top, A_T \rangle \right\}
\]

\[
\text{s.t.} \quad \sum_{T \in \mathcal{P}} w_T \hat{A}_T + R = X^\top X + \lambda I
\]

\[
\hat{A}_T \in S^P_+ \quad \forall T \in \mathcal{P}
\]

\[
R \in S^P_+.
\]

After substituting \( \hat{A}_T = w_T A_T \) and noting that there exists an optimal solution with \( w_T = \min \{1, z(T)\} \), we obtain formulation (15).

Note that if \( \mathcal{P} = \{\emptyset\} \), there is no strengthening and (19) is equivalent to elastic net \((\lambda, \mu > 0)\), lasso \((\lambda = 0, \mu > 0)\), ridge regression \((\lambda > 0, \mu = 0)\) or ordinary least squares \((\lambda = \mu = 0)\). As \(|\mathcal{P}|\) increases, the quality of the conic relaxation (19) for the non-convex \( \ell_0 \)-problem (1) improves, but the computational burden required to solve the resulting SDP also increases. In particular, the full rank-one strengthening with \( \mathcal{P} = 2^P \) requires \( 2^{|\mathcal{P}|} \) semidefinite constraints and is impractical. Proposition 2 suggests using down-monotone sets \( \mathcal{P} \) with limited size

\[
y^\top y + \min -2y^\top X \beta + e^\top u + \langle X^\top X + \lambda I, B \rangle
\]

\[
\text{s.t.} \quad e^\top z \leq k
\]

\[
\beta \leq u, \quad -\beta \leq u
\]

\[
(\text{sdp}_r)
\]

\[
0 \leq w_T \leq \min \{1, e^\top T z_T \} \quad \forall T : |T| \leq r
\]

\[
w_T B_T - \beta_T \beta_T^\top \in S^P_+ \quad \forall T : |T| \leq r
\]

\[
B - \beta \beta^\top \in S^P_+
\]

\[
\beta \in \mathbb{R}^P, \quad z \in [0, 1]^P, \quad u \in \mathbb{R}^P, \quad B \in \mathbb{R}^{P \times P}
\]

for some \( r \in \mathbb{Z}_+ \). In fact, if \( r = 1 \), then \( \text{sdp}_1 \) reduces to the formulation of the optimal perspective relaxation proposed in [13], which is equivalent to using \( \mathcal{MC}_+ \)-regularization. Our computations experiments show that whereas \( \text{sdp}_1 \) may be a weak convex relaxation for problems with low diagonal dominance, \( \text{sdp}_2 \) and \( \text{sdp}_3 \) achieve excellent relaxation bounds even for the case of low diagonal-dominance within reasonable compute times.

### 3. Regularization for the Two-Dimensional Case

To better understand the properties of the proposed conic relaxations, in this section, we study them from a regularization perspective. Consider formulation (17b) in Lagrangean form with multiplier \( \kappa \):

\[
y^\top y + \min -2y^\top X \beta + e^\top u + \phi_T(z, \beta) + \kappa e^\top z \tag{21a}
\]

\[
\beta \leq u, \quad -\beta \leq u \tag{21b}
\]

\[
\beta \in \mathbb{R}^P, \quad z \in [0, 1]^P, \quad u \in \mathbb{R}^P, \tag{21c}
\]

where \( p = 2 \), and

\[
X^\top X + \lambda I = \begin{pmatrix} 1 + \delta_1 & 1 \\ 1 & 1 + \delta_2 \end{pmatrix} \tag{22}
\]
Observe that assumption (22) is without loss of generality, provided that \(X^\top X\) is not diagonal: given a two-dimensional convex quadratic function \(a_1\beta_1^2 + 2a_12\beta_1\beta_2 + a_2\beta_2^2\) (with \(a_12 \neq 0\)), the substitution \(\beta_1 = \alpha\beta_1\) and \(\beta_2 = (a_12/\alpha)\beta_2\) with \(|a_12|/a_2 \leq \alpha \leq a_1\) yields a quadratic form satisfying (22). Also note that we are using the Lagrangean form instead of the cardinality constrained form given in (18) for simplicity; however, since \(\phi_P(z, \beta)\) is convex in \(z\), there exists a value of \(\kappa\) such that both forms are equivalent, i.e., result in the same optimal solutions \(\hat{\beta}\) for the regression problem, and the objective values differ by the constant \(\kappa \cdot k\).

If \(P = \{\emptyset, \{1\}, \{2\}\}\), then (21) reduces to a perspective strengthening of the form

\[
y'y + \min_{z \in [0, 1]^2, \beta \in \mathbb{R}^2} -2y'X\beta + (\beta_1 + \beta_2)^2 + \delta_1 \frac{\beta_1^2}{z_1} + \delta_2 \frac{\beta_2^2}{z_2} + \mu \|\beta\|_1 + \kappa \|z\|_1. 
\]

The links between (23) and regularization were studied\(^3\) in [13].

**Proposition 3** (Dong et al. [13]). Problem (23) is equivalent to the regularization problem

\[
\min_{\beta \in \mathbb{R}^2} \|y - X\beta\|^2_2 + \lambda \|\beta\|^2_2 + \mu \|\beta\|_1 + \rho_{MC}(\beta; \kappa, \delta)
\]

where

\[
\rho_{MC}(\beta; \kappa, \delta) = \begin{cases}
\sum_{i=1}^2 \left(2\sqrt{\kappa \delta_i} |\beta_i| - \delta_i \beta_i^2 \right) & \text{if } \delta_i \beta_i^2 \leq \kappa, \ i = 1, 2 \\
\kappa + 2\sqrt{\kappa \delta_i} |\beta_i| - \delta_i \beta_i^2 & \text{if } \delta_i \beta_i^2 \leq \kappa \text{ and } \delta_j \beta_j^2 > \kappa \\
2\kappa & \text{if } \delta_i \beta_i^2 > \kappa, \ i = 1, 2.
\end{cases}
\]

Regularization \(\rho_{MC}\) is non-convex and separable. Moreover, as pointed out in [13], the regularization given in Proposition 3 is the same as the Minimax Concave Penalty given in [45]; and, if \(\lambda = \delta_1 = \delta_2\), then the regularization given in Proposition 3 reduces to the reverse Huber penalty derived in [38]. Observe that the regularization function \(\rho_{MC}\) is highly dependent on the diagonal dominance \(\delta\): specifically, in the low diagonal dominance setting with \(\delta = 0\), we find that \(\rho_{MC}(\beta; \kappa, 0) = 0\).

We now consider conic formulation (21) for the case \(P = \{\emptyset, \{1\}, \{2\}, \{1, 2\}\}\), corresponding to the full rank-one strengthening:

\[
y'y + \min_{z \in [0, 1]^2, \beta \in \mathbb{R}^2} -2y'X\beta + \frac{(\beta_1 + \beta_2)^2}{\min\{1, z_1 + z_2\}} + \delta_1 \frac{\beta_1^2}{z_1} + \delta_2 \frac{\beta_2^2}{z_2} + \mu \|\beta\|_1 + \kappa \|z\|_1. 
\]

**Proposition 4.** Problem (24) is equivalent to the regularization problem

\[
\min_{\beta \in \mathbb{R}^2} \|y - X\beta\|^2_2 + \lambda \|\beta\|^2_2 + \mu \|\beta\|_1 + \rho_{RI}(\beta; \kappa, \delta)
\]

\(^3\)The case with \(\mu = 0\) is explicitly considered in Dong et al. [13], but the results extend straightforwardly to the case with \(\mu > 0\). The results presented here differ slightly from those in [13] to account for a different scaling in the objective function.
where
\[
\rho_{bl}(\beta;\kappa,\delta) = \begin{cases} 
2\sqrt{\kappa}\sqrt{\beta'(X^\top X + \lambda I)\beta + 2\sqrt{\delta_1}\delta_2|\beta_1\beta_2| - \beta'(X^\top X + \lambda I)\beta} & \text{if } \beta'(X^\top X + \lambda I)\beta + 2\sqrt{\delta_1}\delta_2|\beta_1\beta_2| < \kappa \\
\kappa + 2\sqrt{\delta_1}\delta_2|\beta_1\beta_2| & \text{if } (\sqrt{\delta_1}|\beta_1| + \sqrt{\delta_2}|\beta_2|)^2 \leq \kappa \leq (\beta'(X^\top X + \lambda I)\beta + 2\sqrt{\delta_1}\delta_2|\beta_1\beta_2|)^2 \\
\sum_{i=1}^2 (2\sqrt{\kappa}\delta_i|\beta_i| - \delta_i\beta_i^2) & \text{if } (\sqrt{\delta_1}|\beta_1| + \sqrt{\delta_2}|\beta_2|)^2 > \kappa \\
2\kappa & \text{if } \delta = 0.
\end{cases}
\]

Observe that, unlike \(\rho_{bc_+}\), the function \(\rho_{bn}\) is not separable in \(\beta_1\) and \(\beta_2\) and does not vanish when \(\delta = 0\): indeed, for \(\delta = 0\) we find that
\[
\rho_{bn}(\beta;\kappa,0) = \begin{cases} 
2\sqrt{\kappa}\sqrt{\beta'(X^\top X + \lambda I)\beta - \beta'(X^\top X + \lambda I)\beta} & \text{if } \beta'(X^\top X + \lambda I)\beta < \kappa \\
\kappa & \text{if } 0 \leq \kappa \leq (\beta'(X^\top X + \lambda I)\beta)^2.
\end{cases}
\]

**Proof of Proposition 4.** We prove the result by projecting out the \(z\) variables in (24), i.e., giving closed form solutions for them. There are three cases to consider, depending on the optimal value for \(z_1 + z_2\).

- Case 1: \(z_1 + z_2 < 1\). In this case, we find by setting the derivatives of the objective in (24) with respect to \(z_1\) and \(z_2\) that
\[
\begin{align*}
\kappa - \delta_1 \beta_1^2 + \frac{(\beta_1 + \beta_2)^2}{z_1} - \frac{(\beta_1 + \beta_2)^2}{(z_1 + z_2)^2} &= 0 \\
\kappa - \delta_2 \beta_2^2 + \frac{(\beta_1 + \beta_2)^2}{z_2} - \frac{(\beta_1 + \beta_2)^2}{(z_1 + z_2)^2} &= 0
\end{align*}
\]

Define \(\bar{z} := \sqrt{\delta_2} \beta_2^2 \bar{z}_1\), so \(z_2 = \sqrt{\delta_2} |\beta_2| \bar{z}_1\), and \(z_1 + z_2 = (\sqrt{\delta_1}|\beta_1| + \sqrt{\delta_2}|\beta_2|) \bar{z}_1\). Moreover, we find that (24) reduces to
\[
y^\top y + \min_{\bar{z}>0,\beta\in\mathbb{R}^2} -2y^\top X\beta + \mu\|\beta\|_1 \\
+ \frac{(\beta_1 + \beta_2)^2}{(\sqrt{\delta_1}|\beta_1| + \sqrt{\delta_2}|\beta_2|) \bar{z}_1} + \kappa \left(\sqrt{\delta_1}|\beta_1| + \sqrt{\delta_2}|\beta_2|\right)^2 \bar{z}_1.
\]

An optimal solution of (25) is attained at
\[
z^* = \sqrt{\frac{(\beta_1 + \beta_2)^2 + (\sqrt{\delta_1}|\beta_1| + \sqrt{\delta_2}|\beta_2|)^2}{\kappa \left(\sqrt{\delta_1}|\beta_1| + \sqrt{\delta_2}|\beta_2|\right)^2}}
\]
with objective value
\[
y^\top y + \min_{\beta\in\mathbb{R}^2} -2y^\top X\beta + \mu\|\beta\|_1 + 2\sqrt{\kappa} \sqrt{(\beta_1 + \beta_2)^2 + (\sqrt{\delta_1}|\beta_1| + \sqrt{\delta_2}|\beta_2|)^2} \\
= \min_{\beta\in\mathbb{R}^2} \|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2 + \mu\|\beta\|_1 \\
+ \left(2\sqrt{\kappa} \sqrt{(\beta_1 + \beta_2)^2 + (\sqrt{\delta_1}|\beta_1| + \sqrt{\delta_2}|\beta_2|)^2} - (\beta_1 + \beta_2)^2 - \delta_1 \beta_1^2 - \delta_2 \beta_2^2\right).
\]
Finally, this case happens when $z_1 + z_2 < 1 \iff (\beta_1 + \beta_2)^2 + (\sqrt{\delta_1}|\beta_1| + \sqrt{\delta_2}|\beta_2|)^2 < \kappa$.

- Case 2: $z_1 + z_2 > 1$. In this case, we find by setting the derivatives of the objective in (24) with respect to $z_1$ and $z_2$ that $\bar{z}_i = \frac{2z_i}{\kappa}$ for $i = 1, 2$. Thus, in this case, for an optimal solution $z^*$ of (24), we have $z^*_i = \min\{\bar{z}_i, 1\}$, and problem (24) reduces to

$$y^T y + \min_{\beta \in \mathbb{R}^2} -2y^T X\beta + (\beta_1 + \beta_2)^2 + \sum_{i=1}^2 \max\{\delta_i\beta_i^2, \sqrt{\delta_i}|\beta_i|\} + \mu\|\beta\|_1 + \frac{\sum_{i=1}^2 (\sqrt{\delta_i}|\beta_i| - \delta_i\beta_i^2)}{\lambda}$$

Finally, this case happens when $z_1 + z_2 > 1 \iff (\sqrt{\delta_1}|\beta_1| + \sqrt{\delta_2}|\beta_2|)^2 > \kappa$. Observe that, in this case, the penalty function is precisely the one given in Proposition 3.

- Case 3: $z_1 + z_2 = 1$. In this case, problem (24) reduces to

$$y^T y + \min_{0 \leq z_1 \leq 1, \beta \in \mathbb{R}^2} -2y^T X\beta + (\beta_1 + \beta_2)^2 + \frac{\delta_1\beta_1^2}{z_1} + \frac{\delta_2\beta_2^2}{1 - z_1} + \mu\|\beta\|_1 + \kappa. \quad (26)$$

Setting derivative with respect to $z_1$ in (26) to 0, we have

$$0 = \frac{\delta_1\beta_1^2}{z_1^2} - \frac{\delta_2\beta_2^2}{1 - z_1}$$

Thus, we find that

$$z_1 = \frac{2\delta_1\beta_1^2 \pm \sqrt{4\delta_1^2\beta_1^4 - 4\delta_1\beta_1^2(\delta_1\beta_1^2 - \delta_2\beta_2^2)}}{2(\delta_1\beta_1^2 - \delta_2\beta_2^2)} = \frac{\delta_1\beta_1^2 \pm \sqrt{\delta_1\beta_1^2(\delta_1\beta_1^2 + \delta_2\beta_2^2)}}{\delta_1\beta_1^2 - \delta_2\beta_2^2} = \frac{\sqrt{\delta_1}\beta_1}{\sqrt{\delta_1}\beta_1 + \sqrt{\delta_2}\beta_2}$$

Moreover, since $0 \leq z_1 \leq 1$, we have $z_1 = \frac{\sqrt{\delta_1}\beta_1}{\sqrt{\delta_1}\beta_1 + \sqrt{\delta_2}\beta_2}$ and $1 - z_1 = \frac{\sqrt{\delta_2}\beta_2}{\sqrt{\delta_1}\beta_1 + \sqrt{\delta_2}\beta_2}$.

Substituting in (26), we find the equivalent form

$$y^T y + \min_{\beta \in \mathbb{R}^2} -2y^T X\beta + (\beta_1 + \beta_2)^2 + (\sqrt{\delta_1}|\beta_1| + \sqrt{\delta_2}|\beta_2|)^2 + \mu\|\beta\|_1 + \kappa$$

This final case occurs when neither case 1 or 2 does, i.e., when $(\sqrt{\delta_1}|\beta_1| + \sqrt{\delta_2}|\beta_2|)^2 \leq \kappa \leq (\beta_1 + \beta_2)^2 + (\sqrt{\delta_1}|\beta_1| + \sqrt{\delta_2}|\beta_2|)^2$. \hfill \Box

The plots of $\rho_{MC}$ and $\rho_{K1}$ shown in Figures 1 and 2 correspond to setting the natural value $\kappa = 1$. 


4. Conic Quadratic Relaxations

As mentioned in §1, strong convex relaxations of problem (1), such as sdp, can either be directly used to obtain good estimators via conic optimization, which is the approach we use in our computations, or can be embedded in a branch-and-bound algorithm to solve (1) to optimality. However, using SDP formulations such as (19) in branch-and-bound may be daunting since, to date, efficient branch-and-bound algorithms with SDP relaxations are not available. In contrast, off-the-shelf conic quadratic mixed-integer optimization solvers are successful in solving a broad class of practical problems, and are actively maintained and improved by numerous software vendors. In this section, to facilitate the integration with branch-and-bound solvers, we show how the proposed conic relaxations, and specifically sdp, can be implemented in a conic quadratic framework.

4.1. Constraints with $S^T$ and $|T| = 1$. If $T = \{i\}$, then constraint (20e), $\beta_i^2 \leq w_i B_{ii}$, is a rotated cone constraint as $w_i \geq 0$ and $B_{ii} \geq 0$ in any feasible solution of (20), and thus conic quadratic representable. Moreover, observe that the variable $w_i$ can be dropped from the formulation, resulting in the constraint $\beta_i^2 \leq z_i B_{ii}$.

4.2. Constraints with $S^T$ and $|T| = 2$. As we now show, constraints (20e) with $|T| = 2$ can be accurately approximated using conic quadratic constraints.

**Proposition 5.** Problem sdp is equivalent to the optimization problem

$$\begin{align*}
\min & \quad y^T y + \min -2y^T X \beta + e^T u + \langle X^T X + \lambda I, B \rangle \\
\text{s.t.} & \quad e^T z \leq k \tag{27a}
\beta \leq u, -\beta \leq u \tag{27b} \\
& \quad z_i B_{ii} \geq \beta_i^2 \quad \forall i \in P \tag{27c} \\
& \quad 0 \leq w_{ij} \leq 1, w_{ij} \leq z_i + z_j \quad \forall i \neq j \tag{27d} \\
& \quad 0 \geq \max_{\alpha \geq 0} \left\{ \frac{\alpha \beta_i^2 + 2 \beta_i \beta_j + \beta_j^2 / \alpha}{w_{ij}} - 2B_{ij} - \alpha B_{ii} - B_{jj} / \alpha \right\} \quad \forall i \neq j \tag{27f} \\
& \quad 0 \geq \max_{\alpha \geq 0} \left\{ \frac{\alpha \beta_i^2 - 2 \beta_i \beta_j + \beta_j^2 / \alpha}{w_{ij}} + 2B_{ij} - \alpha B_{ii} - B_{jj} / \alpha \right\} \quad \forall i \neq j \tag{27g} \\
& \quad B - \beta \beta' \in S^P_+ \tag{27h} \\
& \quad \beta \in \mathbb{R}^P, z \in [0, 1]^P, u \in \mathbb{R}_+^P, B \in \mathbb{R}^{P \times P}. \tag{27i}
\end{align*}$$

**Proof.** It suffices to compute the optimal value of $\alpha$ in (27f)–(27g). Observe that the rhs of (27f) can be written as

$$v = \frac{2 \beta_i \beta_j}{w_{ij}} - 2B_{ij} - \min_{\alpha \geq 0} \left\{ \alpha \left( B_{ii} - \frac{\beta_i^2}{w_{ij}} \right) + \frac{1}{\alpha} \left( B_{jj} - \frac{\beta_j^2}{w_{ij}} \right) \right\}. \tag{28}$$

Moreover, in an optimal solution of (27), we have that $w_{ij} = \min\{1, z_i + z_j\}$. Thus, due to constraints (27d), we find that $B_{ii} - \beta_i^2 / w_{ij} \geq 0$ in optimal solutions of (27).

---

4 An effective implementation would require careful constraint management strategies and integration with the different aspects of branch-and-bound solvers, e.g., branching strategies and heuristics. Such an implementation is beyond the scope of the paper.
Similarly, it can be shown that constraint (27g) reduces to
\[ \alpha = \sqrt{\frac{B_{ij}w_{ij} - \beta_i^2}{B_{ii}w_{ij} - \beta_i^2}}, \]
with the objective value
\[ v = \frac{2\beta_i\beta_j}{w_{ij}} - 2B_{ij} - 2\sqrt{(B_{ii}w_{ij} - \beta_i^2)(B_{jj}w_{ij} - \beta_j^2)}. \]
Observe that this expression is also correct when \( B_{ii} = \beta_i^2/\min(1, z_i + z_j) \) or \( B_{jj} = \beta_j^2/\min(1, z_i + z_j) \). Thus, constraint (27f) reduces to
\[ 0 \geq \beta_i\beta_j - B_{ij}w_{ij} - \sqrt{(B_{ii}w_{ij} - \beta_i^2)(B_{jj}w_{ij} - \beta_j^2)}. \]
and equality only occurs if either \( z_i = 1 \) or \( z_j = 0 \). If either \( B_{ii} = \beta_i^2/\min(1, z_i + z_j) \) or \( B_{jj} = \beta_j^2/\min(1, z_i + z_j) \), then the optimal value of (28) is \( v = \frac{2\beta_i\beta_j}{\min(1, z_i + z_j)} - 2B_{ij} \), by setting \( \alpha \to \infty \) or \( \alpha = 0 \), respectively. Otherwise, the optimal \( \alpha \) equals
\[ \alpha = \sqrt{\frac{B_{ij}w_{ij} - \beta_i^2}{B_{ii}w_{ij} - \beta_i^2}}, \]
resulting in tighter relaxations. Note that the use of cuts (as described here) to improve the continuous relaxations of mixed-integer optimization problems is one of the main reasons of the dramatic improvements of MIO software [9].

Moreover, note that constraints (20e) with \( T = \{i, j\} \) are equivalent to
\[ (w_{ij}B_{ii} - \beta_i^2)(w_{ij}B_{jj} - \beta_j^2) \geq (w_{ij}B_{ij} - \beta_i\beta_j)^2. \]
Since the first two constraints are implied by (27d) and \( w_{ij} = \min\{1, z_i + z_j\} \) in optimal solutions, the proof is complete.

Observe that, for any fixed value of \( \alpha \), constraints (27f)–(27g) are conic quadratic representable. Thus, we can obtain relaxations of (27) of the form
\[
y^\top y + \min - 2y^\top X\beta + e^\top u + \langle X^\top X + \lambda I, B \rangle
\]
s.t. (27b), (27c), (27d), (27e), (27f), (27i)
\[
0 \geq \alpha \beta_i^2 + 2\beta_i\beta_j + \beta_j^2/\alpha \min\{1, z_i + z_j\} - 2B_{ij} - \alpha B_{ii} - B_{jj}/\alpha, \forall i \neq j, \alpha \in V_{ij}^+ \]
\[
0 \geq \alpha \beta_i^2 + 2\beta_i\beta_j + \beta_j^2/\alpha \min\{1, z_i + z_j\} + 2B_{ij} - \alpha B_{ii} - B_{jj}/\alpha, \forall i \neq j, \alpha \in V_{ij}^-. \]
where \( V_{ij}^+ \) and \( V_{ij}^- \) are any finite subsets of \( \mathbb{R}_+ \). Relaxation (33) can be refined dynamically: given an optimal solution of (33), new values of \( \alpha \) generated according to (29) (resulting in most violated constraints) can be added to sets \( V_{ij}^+ \) and \( V_{ij}^- \), resulting in tighter relaxations. Note that the use of cuts (as described here) to improve the continuous relaxations of mixed-integer optimization problems is one of the main reasons of the dramatic improvements of MIO software [9].

In relaxation (33), \( V_{ij}^+ \) and \( V_{ij}^- \) can be initialized with any (possibly empty) subsets of \( \mathbb{R}_+ \). However, setting \( V_{ij}^+ = V_{ij}^- = \{1\} \) yields a relaxation with a simple interpretation, discussed next.
4.3. Diagonally dominant matrix relaxation. Let \( \Lambda \in \mathcal{S}_+^P \) be diagonally dominant matrix. Observe that for any \((z, \beta) \in \{0,1\}^P \times \mathbb{R}^P\) such that \(\beta^\top (e - z) = 0\),

\[
t \geq \beta^\top \Lambda \beta \iff t \geq \sum_{i=1}^{P} \left( \Lambda_{ii} - \sum_{j \neq i} |\Lambda_{ij}| \right) \beta_i^2 + \sum_{i=1}^{P} \sum_{j=i+1}^{P} |\Lambda_{ij}| \left( \beta_i + \text{sign}(\Lambda_{ij}) \beta_j \right)^2
\]

\[
\iff t \geq \sum_{i=1}^{P} \left( \Lambda_{ii} - \sum_{j \neq i} |\Lambda_{ij}| \right) \beta_i^2 + \frac{\sum_{i=1}^{P} \sum_{j=i+1}^{P} |\Lambda_{ij}| \left( \beta_i + \text{sign}(\Lambda_{ij}) \beta_j \right)^2}{\min\{1, z_i + z_j\}}.
\]

where the last line follows from using perspective strengthening for the separable quadratic terms, and using (7) for the non-separable, rank-one terms. See [4] for a similar strengthening for signal estimation based on nonnegative pairwise quadratic terms.

We now consider using decompositions of the form \(\Lambda + R = X^\top X + \lambda I\), where \(\Lambda\) is a diagonally dominant matrix and \(R \in \mathcal{S}_+^P\). Given such a decomposition, inequalities (34) can be used to strengthen the formulations. Specifically, we consider relaxations of (3) of the form

\[
y^\top y + \min \{-2y^\top X \beta + e^\top u + \hat{\phi}(z, \beta)\} \quad (35a)
\]

\[
(17b), (17c), (17d), \quad (35b)
\]

where

\[
\hat{\phi}(z, \beta) := \max_{\Lambda, R} \beta^\top R \beta + \sum_{i=1}^{P} \left( \Lambda_{ii} - \sum_{j \neq i} |\Lambda_{ij}| \right) \beta_i^2 + \frac{\sum_{i=1}^{P} \sum_{j=i+1}^{P} |\Lambda_{ij}| \left( \beta_i + \text{sign}(\Lambda_{ij}) \beta_j \right)^2}{\min\{1, z_i + z_j\}}.
\]

\[
s.t. \ \Lambda + R = X^\top X + \lambda I \quad (36a)
\]
\[
\Lambda_{ii} \geq \sum_{j < i} |\Lambda_{ij}| + \sum_{j > i} |\Lambda_{ij}| \quad \forall i \in P \quad (36b)
\]

\[
R \in \mathcal{S}_+^P. \quad (36c)
\]

**Proposition 6.** Problem (35) is equivalent to

\[
y^\top y + \min \{-2y^\top X \beta + e^\top u + \langle X^\top X + \lambda I, B \rangle\} \quad (37a)
\]

\[
s.t. \ e^\top z \leq k \quad (37b)
\]

\[
\beta \leq u, \ -\beta \leq u \quad (37c)
\]

\[
z_i B_{ii} \geq \beta_i^2 \quad \forall i \in P \quad (37d)
\]

\[
(\text{sdp}_2) \quad 0 \leq w_{ij} \leq 1, \ w_{ij} \leq z_i + z_j \quad \forall i \neq j \quad (37e)
\]

\[
0 \geq \frac{\beta_i^2 + 2\beta_i \beta_j + \beta_j^2}{w_{ij}} - 2B_{ij} - B_{ii} - B_{jj} \quad \forall i \neq j \quad (37f)
\]

\[
0 \geq \frac{\beta_i^2 - 2\beta_i \beta_j + \beta_j^2}{w_{ij}} + 2B_{ij} - B_{ii} - B_{jj} \quad \forall i \neq j \quad (37g)
\]

\[
B - \beta B' \in \mathcal{S}_+^P \quad (37h)
\]

\[
\beta \in \mathbb{R}^P, \ z \in [0,1]^P, \ u \in \mathbb{R}_+^P, \ B \in \mathbb{R}^{P \times P}. \quad (37i)
\]
Proof. Let $\Gamma, \Gamma^+, \Gamma^-$ be nonnegative $p \times p$ matrices such that: $\Gamma_{ii} = \Lambda_{ii}$ and $\Gamma_{ij} = 0$ for $i \neq j$; $\Gamma_{ii}^+ = \Gamma_{ii}^-$ and $\Gamma_{ij}^+ - \Gamma_{ij}^- = \Lambda_{ij}$ for $i \neq j$. Problem (36) can be written as

$$
\hat{\phi}(z, \beta) := \max_{\Gamma, \Gamma^+, \Gamma^- R} \beta^T R \beta + \sum_{i=1}^p \left( \Gamma_{ii} - \sum_{j \neq i} \left( \Gamma_{ij}^+ + \Gamma_{ij}^- \right) \right) \frac{\beta_i^2}{z_i} + \sum_{i=1}^p \sum_{j=i+1}^p \left( \Gamma_{ij}^+ \min\{1, z_i + z_j\} + \Gamma_{ij}^- \min\{1, z_i + z_j\} \right) - \frac{(\beta_i - \beta_j)^2}{z_i},
$$

s.t. $\Gamma + \Gamma^+ + \Gamma^- + R = X^T X + \lambda I$ \hspace{1cm} (38b)

$$
\Gamma_{ii} \geq \sum_{j<i} \left( \Gamma_{ji}^+ + \Gamma_{ji}^- \right) + \sum_{j>i} \left( \Gamma_{ij}^+ + \Gamma_{ij}^- \right) \quad \forall i \in P \hspace{1cm} (38d)
$$

$$
R \in S^P_+ \hspace{1cm} (38e).
$$

Then, similarly to the proof of Theorem 3, it is easy to show that the dual of (38) is precisely (37). \hfill \Box

4.4. Relaxing the constraint with $S^P_+$. We now discuss a relaxation of the $p$-dimensional semidefinite constraint $B - \beta \beta^T \in S^P_+$, present in all formulations.

Consider the optimization problem

$$
\bar{\phi}_P(z, \beta) := \max_{A_T, R, \pi} \beta^T R \beta + \sum_{T \in P} \frac{\beta_T^T A_T \beta_T}{\min\{1, z(T)\}} \hspace{1cm} (39a)
$$

s.t. $\sum_{T \in P} A_T + R = X^T X + \lambda I$ \hspace{1cm} (39b)

$$
A_T \in S^T_+ \hspace{1cm} \forall T \in P \hspace{1cm} (39c)
$$

$$
R = X^T \text{diag}(\pi) X \hspace{1cm} (39d)
$$

$$
\pi \in \mathbb{R}^n_+ \hspace{1cm} (39e).
$$

Observe that the objective and constraints (39a)-(39c) are identical to (15). However, instead of (15d), we have $R = \sum_{j=1}^n \pi_j X_j X_j^T$, where $X_j$ is the $j$-th row of $X$ (as a column vector). Moreover, since $\pi \geq 0$, $R \in S^P_+$ in any feasible solution of (39), thus (15) is a relaxation of (39), and, hence, $\bar{\phi}_P$ is indeed a lower bound on $\phi_P$. Therefore, instead of (17), one may use the simpler convex relaxation

$$
y^T y + \min \left[ -2y^T X \beta + e^T u + \hat{\phi}_P(z, \beta) \right] \hspace{1cm} (40a)
$$

$$
e^T z \leq k \hspace{1cm} (40b)
$$

$$
\beta \leq u, -\beta \leq u \hspace{1cm} (40c)
$$

$$
\beta \in \mathbb{R}^P, z \in [0, 1]^P, u \in \mathbb{R}^P_+ \hspace{1cm} (40d)
$$

for (1).
Proposition 7. Problem (40) is equivalent to the SDP

\[ \begin{align*}
    y^\top y + \min & -2y^\top X \beta + e^\top u + \langle X^\top X + \lambda I, B \rangle \\
    \text{s.t.} & \quad e^\top z \leq k \quad (41a) \\
    & \quad \beta \leq u, \quad -\beta \leq u \quad (41b) \\
    & \quad w^\top B_T - \beta_T \beta_T^\top \in S_+^T \quad \forall T \in \mathcal{P} \quad (41e) \\
    & \quad X_j^\top (B - \beta \beta^\top) X_j \geq 0 \quad \forall j = 1, \ldots, n \quad (41f) \\
    & \quad \beta \in \mathbb{R}^p, \quad z \in [0,1]^p, \quad u \in \mathbb{R}_+^p, \quad w \in [0,1]^p, \quad B \in \mathbb{R}^{p \times p}.
\end{align*} \]

Proof. The proof is based on conic duality similar to the proof of Theorem 3. □

Observe that in formulation (41), the \((p+1)\)-dimensional semidefinite constraint

\[(19f)\]

is replaced with \(n\) rank-one quadratic constraints (41f). We denote by \(\text{cqp}_r\) the relaxation of \(\text{sdp}_r\) obtained by replacing (20f) with (41f). In general, \(\text{cqp}_r\) is still an SDP due to constraints (41e); however, note that special cases \(\text{cqp}_1\) and \(\text{cqp}_2\) are conic quadratic formulations, and \(\text{cqp}_2\) can be implemented in a conic quadratic framework by using cuts, as described in §4.2. Moreover, constraints (41f) could also be dynamically refined to better approximate the SDP constraint, or formulation (41) could be improved with ongoing research on approximating SDP via mixed-integer conic quadratic optimization, e.g., see [28, 29].

Remark 2 (Further rank-one strengthening). Since constraints (41f) are rank-one quadratic constraints, additional strengthening can be achieved using the inequalities given in §2.1. Specifically, let \(T_j = \{i \in P : X_{ji} \neq 0\}\), and inequalities (41f) may be replaced with stronger versions

\[ \langle B, X_j X_j^\top \rangle - \frac{(X_j^\top \beta)^2}{\min\{1, z(T_j)\}} \geq 0 \quad \forall j \in N. \]

This strengthening is particularly effective when \(X\) is sparse. □

5. Computations

In this section we report computational experiments with the proposed conic relaxations on synthetic as well as benchmark datasets. The convex optimization problems are solved with MOSEK 8.1 solver on a laptop with a 1.80GHz Intel® Core™ i7-8550U CPU and 16 GB main memory. All solver parameters were set to their default values. We divide our discussion in two parts: first, in §5.1, we test \(\text{sdp}_r\) on benchmark instances previously used in the literature, and focus on the relaxation quality and ability to approximate the exact \(\ell_0\)-problem (1); then, in §5.2, we adopt the same experimental framework used in [7, 21] to generate synthetic instances and evaluate the proposed conic formulations from an inference perspective. In both cases, our results compare favorably with the results reported in previous works using the same instances.

5.1. Approximation study on benchmark instances. In this section we focus on the ability of \(\text{sdp}_r\) to provide near-optimal solutions to problem (1).
5.1.1. Computing optimality gaps. Observe that the optimal objective value $\nu_\star^r$ of $\text{sdp}_r$ provides a lower bound on the optimal objective value of (1). To obtain an upper bound, we use a simple heuristic to retrieve a feasible solution of (1): given an optimal solution vector $\hat{\beta}^r$ for $\text{sdp}_r$, let $\hat{\beta}^r_{(k)}$ denote the $k$-th largest absolute value, let $T = \{i \in P : |\hat{\beta}^r_i| \geq \hat{\beta}^r_{(k)}\}$, let $\hat{\beta}_T$ be the $T$-dimensional ols/ridge estimator using only predictors in $T$, i.e.,

$$\hat{\beta}_T = (X_T^T X_T + \lambda I_T)^{-1} X_T^T y,$$

where $X_T$ denotes the $n \times |T|$ matrix obtained by removing the columns with indexes not in $T$, and let $\hat{\beta}$ be the $P$-dimensional vector obtained by filling the missing entries in $\hat{\beta}_T$ with zeros. Since $\|\hat{\beta}\|_0 \leq k$ by construction, $\hat{\beta}$ is feasible for (1), and its objective value $\nu_u$ is an upper bound on the optimal objective value of (1). Moreover, the optimality gap can be computed as

$$\text{gap} = \frac{\nu_u - \nu_\star^r}{\nu_\star^r} \times 100. \quad (42)$$

Note that while stronger relaxations always result in improved lower bounds $\nu_\star^r$, the corresponding heuristic upper bounds $\nu_u$ are not necessarily better, thus the optimality gaps are not guaranteed to improve with stronger relaxations. Nevertheless, as shown next, stronger relaxation do in general yield much better gaps in practice.

5.1.2. Datasets. For these experiments, we use the benchmark datasets in Table 1. The first five were first used in [34] in the context of MIO algorithms for best subset selection, and later used in [18]. The diabetes dataset with all second interactions was introduced in [14] in the context of lasso, and later used in [7]. A few datasets require some manipulation to eliminate missing values and handle categorical variables. The processed datasets before standardization\(^5\) can be downloaded from http://atamturk.ieor.berkeley.edu/data/sparse.regression.

5.1.3. Results. For each dataset, $\lambda \in \{0, 0.05, 0.1\}^6$ and $\mu = 0$, we solve the conic relaxations of (1) for $k = 3, 4, \ldots, \lceil p/3 \rceil$. Specifically, we solve $\text{sdp}_r$ for $r = 1, 2, 3$, and two orders-of-magnitude less than ols/ridge. Specifically, for $\lambda = 0.05$, the average gaps (across all datasets) are $0.4\%$ for $\text{sdp}_2$ and $0.3\%$ for $\text{sdp}_3$; for $\lambda = 0.1$, the average gap is $0.2\%$ for both $\text{sdp}_2$ and $\text{sdp}_3$. In contrast, achieving small optimality gaps in the low diagonal dominance setting $\lambda = 0$ is much more difficult: formulation $\text{sdp}_1$.

\(^5\)In our experiments, the datasets were standardized first.

\(^6\)Since the data is standardized, $\lambda = 0.05$ and $\lambda = 0.1$ correspond to increasing the magnitude of the diagonal elements of $X^T X$ by $5\%$ and $10\%$, respectively.
Table 2. Average computation times (in seconds).

<table>
<thead>
<tr>
<th>dataset</th>
<th>p</th>
<th>n</th>
<th>ols/ridge</th>
<th>cqp</th>
<th>sdp</th>
<th>sdp</th>
<th>sdp</th>
</tr>
</thead>
<tbody>
<tr>
<td>housing</td>
<td>13</td>
<td>506</td>
<td>&lt;0.1</td>
<td>0.1</td>
<td>&lt;0.1</td>
<td>0.1</td>
<td>0.3</td>
</tr>
<tr>
<td>servo</td>
<td>19</td>
<td>167</td>
<td>&lt;0.1</td>
<td>0.2</td>
<td>&lt;0.1</td>
<td>0.1</td>
<td>1.3</td>
</tr>
<tr>
<td>auto MPG</td>
<td>25</td>
<td>392</td>
<td>&lt;0.1</td>
<td>0.5</td>
<td>0.1</td>
<td>0.2</td>
<td>3.4</td>
</tr>
<tr>
<td>solar flare</td>
<td>26</td>
<td>1,066</td>
<td>&lt;0.1</td>
<td>1.0</td>
<td>0.1</td>
<td>0.3</td>
<td>3.3</td>
</tr>
<tr>
<td>breast cancer</td>
<td>37</td>
<td>196</td>
<td>&lt;0.1</td>
<td>0.6</td>
<td>0.2</td>
<td>0.7</td>
<td>11.3</td>
</tr>
<tr>
<td>diabetes</td>
<td>64</td>
<td>442</td>
<td>&lt;0.1</td>
<td>3.8</td>
<td>2.8</td>
<td>7.2</td>
<td>113.5</td>
</tr>
</tbody>
</table>

proposed in [13] and corresponding to MC+, is barely able to improve over ols, and other existing perspective-based relaxations [8, 38, 44] cannot be used in this setting. Formulation cqp is equally ineffective in this setting. Nonetheless, formulations sdp and sdp are able to reduce the optimality gaps by an order-of-magnitude, and result in optimality gaps of 4% or less in all instances except for diabetes: the average optimality gaps with λ = 0 (excluding the housing dataset) is 3.9% for sdp and 3.5% for sdp, which are much lower than 27.6% for sdp and 37.3% for ols.

In summary, cqp is competitive with sdp, with neither formulation consistently outperforming the other, sdp results in considerable improvements over sdp and cqp, and sdp results in a small improvement over sdp.

To illustrate the effect of the sparsity parameter k in the optimality gaps, we show in Figure 4 the optimality gaps of ols, sdp, sdp, and sdp as a function of k for the breast cancer and diabetes datasets. Observe that, in general, optimality gaps decrease as the cardinality increases (note that if k = p, the optimality gaps of all relaxations are 0%). Formulations ols and sdp are particularly poor for low values of k, while formulations sdp and sdp perform consistently well across all values of k. The results suggest that the stronger relaxations sdp and sdp are especially beneficial for low values of the sparsity parameter k.

5.1.4. Computation times. Table 2 presents, for each dataset and method, the average time required to solve the relaxations. While the new relaxations are certainly more expensive to solve than the simple relaxations such as ols/ridge, with the exception of sdp, they can still be solved efficiently, under ten seconds in all cases. Formulation sdp, which has \( \binom{p}{2} \) additional 4-dimensional SDP constraints, requires substantially more time, although the computation time under two minutes is still quite reasonable for many practical applications. Formulation sdp seems to achieve the best balance between quality and efficiency, resulting in excellent relaxation quality without incurring in excessive computational costs.

We now compare these solution times with the computation times required to solve to optimality the MIO problems, as reported in other papers. Gómez and Prokopyev [18] use CPLEX solver with provable big-M constraints on the same datasets: while the first four instances can be solved to optimality in seconds, problems with the breast cancer dataset require around 80 seconds to solve to optimality, and problems with the diabetes dataset cannot be solved to optimality within one hour of branch-and-bound. Bertsimas et al. [7] use Gurobi solver coupled with tailored warm-start methods and heuristic big-M constraints on the diabetes dataset: in general, the optimization problems are not solved within one hour of branch-and-bound due to the difficulty of improving the lower bounds of the
Figure 3. Average optimality gaps for the convex formulations.
algorithm, although the solver reports gaps\(^7\) of the order of 0.3% after an hour of branch-and-bound. In conclusion, we see that sdp\(_2\) runs much faster than branch-and-bound methods, although it does not attempt to solve (1) to optimality, and is able to report guaranteed optimality gaps without resorting to big-M constraints.

5.2. **Inference study on synthetic instances.** We now present inference results on synthetic data, using the same simulation setup as in [7, 21]. Here we present a summary of the simulation setup and refer the readers to [21] for an extended description.

5.2.1. **Instance generation.** For given dimensions \(n, p\), sparsity \(s\), predictor autocorrelation \(\rho\), and signal-to-noise ratio SNR, the instances are generated as follows:

1. The (true) coefficients \(\beta_0\) have the first \(s\) components equal to one, and the rest equal to zero.
2. The rows of the predictor matrix \(X \in \mathbb{R}^{n \times p}\) are drawn from i.i.d. distributions \(\mathcal{N}_p(0, \Sigma)\), where \(\Sigma \in \mathbb{R}^{p \times p}\) has entry \((i, j)\) equal to \(\rho^{|i-j|}\).
3. The response vector \(y \in \mathbb{R}^n\) is drawn from \(\mathcal{N}_p(X\beta_0, \sigma^2 I)\), where \(\sigma^2 = \beta_0^\top X\beta_0 / \text{SNR}\).

In each experiment ten instances are generated with the same parameters and the averages are reported. We use the same SNR values as [21], i.e., SNR = 0.05, 0.09, 0.14, 0.25, 0.42, 0.71, 1.22, 2.07, 3.52, 6.0. We use \(n = 500\), \(p = 100\) and \(s = 5\), also corresponding to experiments presented in [21].

5.2.2. **Evaluation metrics.** Let \(x_0\) denote the test predictor drawn from \(\mathcal{N}_p(0, \Sigma)\) and let \(y_0\) denote its associated response value drawn from \(\mathcal{N}(x_0^\top \beta_0, \sigma^2)\). Given an estimator \(\hat{\beta}\) of \(\beta_0\), the following metrics are reported:

- **Relative risk:**
  \[
  \text{RR}(\hat{\beta}) = \frac{\mathbb{E} \left( x_0^\top \hat{\beta} - x_0^\top \beta_0 \right)^2 }{\mathbb{E} \left( x_0^\top \beta_0 \right)^2 }
  \]
  with a perfect score 0 and null score of 1.

\(^7\)Since Bertsimas et al. [7] use heuristic big-M constraints, the convex formulations used in their method are not guaranteed to be relaxations of (1), thus the reported gaps do not have the same interpretation as (42).
Relative test error:

\[
\text{RTE}(\hat{\beta}) = \frac{\mathbb{E}(x_0^T \hat{\beta} - y_0)^2}{\sigma^2}
\]

with a perfect score of 1 and null score of SNR+1.

Proportion of variance explained:

\[
1 - \frac{\mathbb{E}(x_0^T \hat{\beta} - y_0)^2}{\text{Var}(y_0)}
\]

with perfect score of SNR/(1+SNR) and null score of 0.

Sparsity: We record the number of nonzeros

\[
\|\hat{\beta}\|_0,
\]

as done in [21]. Additionally, we also report the number of variables correctly identified, given by

\[
\sum_{i=1}^{p} \mathbb{1}\{\hat{\beta}_i \neq 0 \text{ and } (\beta_0)_i \neq 0\}.
\]

5.2.3. Procedures. In addition to the training set of size \(n\), a validation set of size \(n\) is generated with the same parameters, matching the precision of leave-one-out cross-validation. We use the following procedures to obtain estimators \(\hat{\beta}\).

- **elastic net**: The elastic net procedure was tuned by generating 10 values of \(\lambda\) ranging from \(\lambda_{\text{max}} = \|X^Ty\|_{\infty}\) to \(\lambda_{\text{max}}/200\) on a log scale, generating 10 values of \(\mu\) on the same interval, and using the pair \((\lambda, \mu)\) that results in the best prediction error on the validation set. A total of 100 tuning parameters values are used.

- **sdp\_2 cross**: The estimator obtained from solving sdp\_2 (\(\lambda = \mu = 0\)) for all values of \(k = 0, \ldots, 7\) and choosing the one that results in the best prediction error on the validation set.

- **sdp\_2 k = s**: The estimator obtained from solving sdp\_2 with \(k = s\) and \(\lambda = \mu = 0\) on the training set. Corresponds to a method with perfect insight on the true sparsity \(s\).

- **sdp\_2 + k = s**: The estimator obtained from solving sdp\_2 with \(k = s\) and \(\lambda = \mu = 0\) on the dataset obtained from merging the training and validation sets. The intuition is that, since cross-validation is not used at all, there is no reason to discard half of the data.

The elastic net procedure approximately corresponds to the lasso procedure with 100 tuning parameters used in [21]. Similarly, sdp\_2 with cross-validation approximately corresponds to the best subset procedure with 51 tuning parameters used in [21]; nonetheless, the estimators from [21] are obtained by running a MIO solver for 3 minutes, while ours are obtained from solving to optimality a strong convex relaxation. The last two procedures are included to illustrate what is, in our opinion, one of the main advantages of solving (3) to optimality or using a strong convex relaxation: the interpretability of parameter \(k\), and the ability to enforce a sparsity prior without a need for cross-validation. In these experiments we use

---

8 An entry \(\hat{\beta}_i\) is deemed to be non-zero if \(|\hat{\beta}_i| > 10^{-5}\). This is the default integrality precision in commercial MIO solvers.

9 Hastie et al. [21] use values of \(k = 0, \ldots, 50\). Nonetheless, in our computations with the same tuning parameters, we found that values of \(k \geq 8\) are never selected after cross-validation. Thus our procedure with 8 tuning parameters results in the same results as the one with 51 parameters from a statistical viewpoint, but requires only a fraction of the computational effort.
since, as pointed out in §5.1, it achieves the best balance between relaxation accuracy and efficiency.

5.2.4. Optimality gaps and computation times. Before describing the statistical results, we briefly comment on the relaxation quality and computation time of \texttt{sdp}_2. Table 3 shows, for instances with \( n = 500, \ p = 100, \) and \( s = 5, \) the optimality gap and relaxation quality of \texttt{sdp}_2 — each column represents the average over ten instances generated with the same parameters. In all cases, \texttt{sdp}_2 produces optimal or near-optimal estimators, with optimality gap at most 0.3%. In fact, with \texttt{sdp}_2, we find that 97% of the estimators for \( \rho = 0.00 \) and 68% of the estimators with \( \rho = 0.35 \) are provably optimal\(^{10} \) for (1). For a comparison, Hastie et al. [21] report that, in their experiments, the MIO solver (with a time limit of three minutes) is able to prove optimality for only 35% of the instances generated with similar parameters. Although Hastie et al. [21] do not report optimality gaps for the instances where optimality is not proven, we conjecture that such gaps are significantly larger than those reported in Table 3 due to weak relaxations with big-\( M \) formulations.

In summary, for this class of instances, \texttt{sdp}_2 is able produce optimal or practically optimal estimators of (1) in about 30 seconds.

<table>
<thead>
<tr>
<th>SNR</th>
<th>0.05</th>
<th>0.09</th>
<th>0.14</th>
<th>0.25</th>
<th>0.42</th>
<th>0.71</th>
<th>1.22</th>
<th>2.07</th>
<th>3.52</th>
<th>6.00</th>
<th>avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \rho = 0.00 )</td>
<td>gap</td>
<td>0.1</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>time</td>
<td>45.2</td>
<td>38.8</td>
<td>38.6</td>
<td>29.5</td>
<td>29.3</td>
<td>28.4</td>
<td>27.4</td>
<td>26.3</td>
<td>26.4</td>
<td>25.9</td>
</tr>
<tr>
<td>( \rho = 0.35 )</td>
<td>gap</td>
<td>0.3</td>
<td>0.2</td>
<td>0.3</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>time</td>
<td>48.0</td>
<td>47.6</td>
<td>49.4</td>
<td>44.1</td>
<td>39.3</td>
<td>30.7</td>
<td>29.0</td>
<td>29.1</td>
<td>27.3</td>
<td>28.0</td>
</tr>
</tbody>
</table>

5.2.5. Results: accuracy metrics. Figure 5 plots the relative risk, relative test error, proportion of variance explained and sparsity results as a function of the SNR for instances with \( n = 500, \ p = 100, \) and \( \rho = 0. \) Figure 6 plots the same results for instances with \( \rho = 0.35. \) The setting with \( \rho = 0.35 \) was also presented in [21]. We see that \texttt{elastic net} outperforms \texttt{sdp}_2 with cross-validation in low SNR settings, i.e., in SNR\( \leq 0.05 \) for \( \rho = 0 \) and SNR\( \leq 0.14 \) for \( \rho = 0.35, \) but results in worse predictive performance for all other SNRs. Moreover, \texttt{sdp}_2 is able to recover the true sparsity pattern of \( \beta_0 \) for sufficiently large SNR, while \texttt{elastic net} is unable to do so. We also see that \texttt{sdp}_2 performs comparatively better than \texttt{elastic net} in instances with \( \rho = 0. \) Indeed, for large autocorrelations \( \rho, \) features where \((\beta_0)_i = 0\) still have predictive value, thus the dense estimator obtained by \texttt{elastic net} retains a relatively good predictive performance (however, such dense solutions are undesirable from an \textit{interpretability} perspective). In contrast, when \( \rho = 0, \) such features are simply noise and \texttt{elastic net} results in overfitting, while methods that deliver sparse solution such as \texttt{sdp}_2 perform much better in comparison. We also note that \texttt{sdp}_2 with cross-validations selects model corresponding to sparsities \( k < s \) in low SNRs, while it consistently selects models with \( k \approx s \) in high SNRs.

\(^{10}\) A solution is deemed optimal if \textit{gap} \( < 10^{-4}, \) which is the default parameter in MIO solvers.
Figure 5. Relative risk, relative test error, proportion of variance explained and sparsity as a function of SNR, with \( n = 500, p = 100, s = 5 \) and \( \rho = 0.00 \).
Figure 6. Relative risk, relative test error, proportion of variance explained and sparsity as a function of SNR, with $n = 500$, $p = 100$, $s = 5$ and $\rho = 0.35$. 
We point out that, as suggested in [32], the results for low SNR could potentially be improved by fitting models with $\mu > 0$.

We now compare the performance of $\text{sdp}_2$ with cross-validation with the methods that use the prior $k = s$. When the validation set is discarded by the method that enforces the prior, we see that using cross-validation yields much better results in low SNRs, as values with $k < s$ result in better performance; in contrast, simply enforcing the prior results in (slightly) better results for large SNRs. However, when the prior is enforced and the validation set is also used for training the model, the resulting method $\text{sdp}_2^+$ outperforms all other methods for all SNRs. Note that, in general, it is not possible for $\ell_1$-based methods to obtain a model with a desired sparsity without cross-validation, as the sparsity-inducing properties of parameter $\mu$ cannot be interpreted naturally.

Note that the “true” parameter $s$ is rarely known in practice (although in some applications it may be possible to narrow down $s$ to a small interval using prior knowledge). However, independently of the “true” value of $s$ (or whether the data-generating process is sparse at all), it may be desirable to restrict the search to parsimonious models of size at most $k$ for interpretability, especially when the estimators $\hat{\beta}$ are meant to be analyzed by human experts. In such situations, methods like $\text{sdp}_2$ may be preferred to $\ell_1$ methods that require extensive cross-validation to achieve a desired sparsity level. Finally, even if $s$ is unknown and interpretability is not a concern, the ability to accurately approximate (1) through relaxations $\text{sdp}_r$ allows the decision-maker to use information criteria such as AIC [1] and BIC [39], which do not require cross-validation, to select $k$, see also [18, 43].

6. Conclusions

In this paper we derive strong convex relaxations for sparse regression. The relaxations are based on the ideal formulations for rank-one quadratic terms with indicator variables. The new relaxations are formulated as semidefinite optimization problems in an extended space and are stronger and more general than the state-of-the-art formulations. In our computational experiments, the proposed conic formulations outperform the existing approaches, both in terms of accurately approximating the best subset selection problems and of achieving desirable estimation properties in statistical inference problems with sparsity.

Acknowledgments

A. Atamtürk is supported, in part, by Grant No. 1807260 from the National Science Foundation. A. Gómez is supported, in part, by Grant No. 1818700 from the National Science Foundation.

References


---


