BCOL RESEARCH REPORT 19.01

Industrial Engineering & Operations Research
University of California, Berkeley, CA 94720–1777

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 ℓ_0 constraint restricting the support of the estimators is a challenging ($\mathcal{NP}$-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 ℓ_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; October 2020; October 2022

A. Atamtürk: Department of Industrial Engineering & Operations Research, University of California, Berkeley, CA 94720. atamturk@berkeley.edu
A. Gómez: Department of Industrial & Systems Engineering, Viterbi School of Engineering, University of Southern California, CA 90089. gomezand@usc.edu .

1. Introduction

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

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2 + \mu \|\boldsymbol{\beta}\|_1 \text{ s.t. } \|\boldsymbol{\beta}\|_0 \leq k, \tag{1}$$

where $\|\boldsymbol{\beta}\|_0$ denotes cardinality of the support of $\boldsymbol{\beta}$. Problem (1) encompasses a broad range of the regression models. It includes as special cases: `ridge` regression [32], when $\lambda > 0$, $\mu = 0$ and $k \geq p$; `lasso` [52], when $\lambda = 0$, $\mu \geq 0$ and $k \geq p$; `elastic net` [65] when $\lambda, \mu > 0$ and $k \geq p$; `best subset selection` [44], when $\lambda = \mu = 0$ 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. [43] 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 $\lambda \|\boldsymbol{A}\boldsymbol{\beta}\|_2^2 + \mu \|\boldsymbol{C}\boldsymbol{\beta}\|_1$, resulting in sparse variants of the `fused lasso` [48, 53], `generalized lasso` [38, 54] and `smooth lasso` [31], among others.

Regularization techniques. The motivation and benefits of the regularization and sparsity are well-documented in the literature. In particular, two key principals drive sparsity in machine learning models: generalization and interpretability. Generalization refers to the ability of a model to perform well out-of-sample. The principle of parsimony postulates that observed phenomena often admit simple explanations, and thus sparse models are preferable as they are more likely to capture such explanations; in fact, Hastie et al. [25] 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. Interpretability, on the other hand, is becoming increasingly important due to the deployment of machine learning models in high-stakes situations [50], where complex models can result in undesirable, unfair or even discriminatory behavior. Moreover, interpretable learning models are also preferable when the output of the model is meant to serve as input to a downstream decision-making process [11, 40, 41].

`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 $\boldsymbol{X}$ is not orthogonal, but it does not result in sparsity. On the other hand, `lasso`, the ℓ_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 [44, 60, 61, 62, 63, 64]. Moreover, achieving a target sparsity level k with `lasso` requires significant experimentation with the penalty parameter μ [10]. When $k \geq p$, the cardinality constraint on ℓ_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 $\mathcal{N}P$-hard [46], thus finding an optimal solution may require excessive computational effort and methods to solve it approximately are used instead [33, 47]. Due to the perceived difficulties of tackling the non-convex ℓ_0 constraint in (1), `lasso`-type simpler approaches are still preferred for inference problems with sparsity [27].

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

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

where $\rho_i:\mathbb{R}\to\mathbb{R}$ are non-convex regularization functions. Examples of such regularization functions include ℓ_q penalties with $0<q<1$ [18] 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 [34, 42, 66], and suboptimal solutions may not benefit from the improved statistical properties. To address such difficulties, Zhang et al. [59] propose the `minimax concave penalty` ($\texttt{MC}_+$), a class of sparsity-inducing penalty functions where the non-convexity of ρ is offset by the convexity of $\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2$ for sufficiently sparse solutions, so that (2) remains convex – Zhang et al. [59] refer to this property as sparse convexity. Thus, in the ideal scenario (and with proper tuning of the parameter controlling the concavity of ρ), the $\texttt{MC}_+$ 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 $\texttt{MC}_+$ greatly depends on the diagonal dominance of the matrix $\boldsymbol{X}^\top\boldsymbol{X}$ (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 $\boldsymbol{X}^\top\boldsymbol{X}$ 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 [21, 45], as well as the `diabetes` dataset with all second interactions used in [7, 14]. The diagonal dominance of a positive semidefinite matrix $\boldsymbol{A}$ is computed as

$$\texttt{dd}(\boldsymbol{A}):=(1/\mathrm{tr}(\boldsymbol{A}))\max_{\boldsymbol{d}\in\mathbb{R}^p_+}\boldsymbol{e}^\top\boldsymbol{d}\ \text{s.t.}\ \boldsymbol{A}-\mathrm{diag}(\boldsymbol{d})\succeq 0,$$

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

TABLE 1. Diagonal dominance of $\boldsymbol{X}^\top \boldsymbol{X}$ for benchmark datasets.

dataset	p	n	dd$\times$100%
housing	13	506	26.7%
servo	19	167	0.0%
auto MPG	25	392	1.5%
solar flare	26	1,066	8.8%
breast cancer	37	196	3.6%
diabetes	64	442	0.0%
crime	100	1993	13.5 %

$\boldsymbol{z} \in \{0,1\}^p$, where $z_i = \mathbb{1}_{\beta_i \neq 0}$, problem (1) can be reformulated as

$$\boldsymbol{y}^\top \boldsymbol{y} + \min_{\boldsymbol{\beta},\boldsymbol{z},\boldsymbol{u}} \; -2\boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\top \left(\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}\right)\boldsymbol{\beta} + \mu \sum_{i=1}^p u_i \tag{3a}$$

$$\text{s.t.} \sum_{i=1}^p z_i \leq k \tag{3b}$$

$$\beta_i \leq u_i, \; -\beta_i \leq u_i \quad i = 1, \dots, p \tag{3c}$$

$$\beta_i(1 - z_i) = 0 \quad i = 1, \dots, p \tag{3d}$$

$$\boldsymbol{\beta} \in \mathbb{R}^p, \; \boldsymbol{z} \in \{0,1\}^p, \; \boldsymbol{u} \in \mathbb{R}^p_+. \tag{3e}$$

The non-convexity of (1) is captured by the complementary constraints (3d) and the integrality constraints $\boldsymbol{z} \in \{0,1\}^p$. 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

$$-Mz_i \leq \beta_i \leq Mz_i \tag{4}$$

for a sufficiently large number M to bound the variables β_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 $M = \tau \|\hat{\boldsymbol{\beta}}\|_\infty$ where $\tau \in \mathbb{R}_+$ and $\hat{\boldsymbol{\beta}}$ is a feasible solution of (1) found via a heuristic[1]. While using such heuristic values yield reasonable performance for small enough values of τ, 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 [4]. 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 than the natural convex relaxation

[1]This method with $\tau = 2$ was used in the computations in [7].

of (3) with $\lambda = \mu = 0$ and big-M constraints is precisely `lasso`, see [13] for example. Therefore, sparsity-inducing techniques that more accurately capture the properties of the non-convex constraint $\|\boldsymbol{\beta}\|_0 \leq k$ can be found by deriving tighter convex relaxations of (1).

Pilanci et al. [49] 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.

The perspective relaxation. Problem (3) is a mixed-integer convex quadratic optimization problem with indicator variables, a class of problems which has received a fair amount of attention in the optimization literature. In particular, the `perspective relaxation` [1, 16, 22] is, by now, a standard technique that can be 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,

$$Q_1 = \left\{ z \in \{0,1\}, \beta \in \mathbb{R}, t \in \mathbb{R}_+ : \beta_i^2 \leq t,\ \beta_i(1 - z_i) = 0 \right\} \cdot$$

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 β_i^2, expressed as a rotated cone constraint:

$$\text{cl conv}(Q_1) = \left\{ z \in [0,1], \beta \in \mathbb{R}, t \in \mathbb{R}_+ : \frac{\beta_i^2}{z_i} \leq t \right\} \cdot$$

Xie and Deng [58] apply the `perspective relaxation` to the separable quadratic regularization term $\lambda\|\boldsymbol{\beta}\|_2^2$, i.e., reformulate (3) as

$$\boldsymbol{y}^\top \boldsymbol{y} + \min_{\boldsymbol{\beta}, \boldsymbol{z}, \boldsymbol{u}} \; -2\boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\top \left(\boldsymbol{X}^\top \boldsymbol{X}\right) \boldsymbol{\beta} + \lambda \sum_{i=1}^p \frac{\beta_i^2}{z_i} + \mu \sum_{i=1}^p u_i \tag{5a}$$

$$\text{s.t.} \; \sum_{i=1}^p z_i \leq k \tag{5b}$$

$$\beta_i \leq u_i,\ -\beta_i \leq u_i \quad i = 1, \ldots, p \tag{5c}$$

$$\boldsymbol{\beta} \in \mathbb{R}^p,\ \boldsymbol{z} \in \{0,1\}^p,\ \boldsymbol{u} \in \mathbb{R}_+^p. \tag{5d}$$

Moreover, they show that the continuous relaxation of (5) is equivalent to the continuous relaxation of the formulation used by Bertsimas and Van Parys [8]. Dong et al. [13] also study the `perspective relaxation` in the context of regression: first, they show that using the `reverse Huber penalty` [49] is, in fact, equivalent to just solving the convex relaxation of (5) — thus the relaxations of [8, 49, 58] all coincide; second, they propose to use an *optimal* `perspective relaxation`, i.e., by applying the perspective relaxation to a separable quadratic function $\boldsymbol{\beta}^\top \boldsymbol{D}\boldsymbol{\beta}$, where $\boldsymbol{D}$ is a nonnegative diagonal matrix such that $\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I} - \boldsymbol{D} \succeq 0$; finally, they show that solving this stronger convex relaxation of the optimal `perspective relaxation` is, in fact, equivalent to using the $\texttt{MC}_+$ penalty [59]. However, the authors also point out that if $\lambda = 0$ and a suitable matrix D cannot be found, then the optimal perspective relaxation reduces to `lasso`. For example, from Table 1,

[2]We use the convention that $\frac{\beta_i^2}{z_i} = 0$ when $\beta_i = z_i = 0$ and $\frac{\beta_i^2}{z_i} = \infty$ if $z_i = 0$ and $\beta_i \neq 0$.

we see that the optimal perspective relaxation would reduce to `lasso` in the `servo` and `diabetes` datasets.

The perspective relaxation is now a state-of-the-art method to convexify problems with separable terms and indicator variables. However, there are relatively few convexification techniques for problems without separable terms [17, 20, 23, 35]. In fact, among the previously discussed methods for sparse regression, the optimal `perspective relaxation` of Dong et al. [13] is the only one that does not explicitly require the use of the Tikhonov regularization $\lambda\|\boldsymbol{\beta}\|_2^2$. Nonetheless, as the authors point out, if $\lambda = 0$ then the method is effective only when the matrix $\boldsymbol{X}^\top \boldsymbol{X}$ 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, λ 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 λ and, critically, are able to achieve 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 $\rho_{\texttt{R1}}(\boldsymbol{\beta})$ which: *(i)* imposes larger penalties than the separable minimax concave penalty [59] $\rho_{\texttt{MC}_+}(\boldsymbol{\beta})$ 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 $\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2$, 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:

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

where $\boldsymbol{X}^\top \boldsymbol{X} = \begin{pmatrix} 1+\delta & 1 \\ 1 & 1+\delta \end{pmatrix}$ 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 $\texttt{MC}_+$penalty for different values of δ and the proposed rank-one `R1` regularization. The graphs of $\texttt{MC}_+$ and `R1` are obtained by setting $\lambda = \mu = 0$ and $\kappa = 1$, and using the appropriate convex strengthening, see §3 for details. Observe that the `R1` regularization results in larger penalties than $\texttt{MC}_+$for all values of δ, and the improvement increases as $\delta \to 0$. In addition, Figure 2 shows the effect of using the `lasso` constraint $\|\boldsymbol{\beta}\|_1 \leq k$, the $\texttt{MC}_+$ constraint $\rho_{\texttt{MC}_+}(\boldsymbol{\beta}) \leq k$, and the rank-one constraint $\rho_{\texttt{R1}}(\boldsymbol{\beta}) \leq k$ in a two-dimensional problem to achieve sparse solutions satisfying $\|\boldsymbol{\beta}\|_0 \leq 1$. Specifically, let

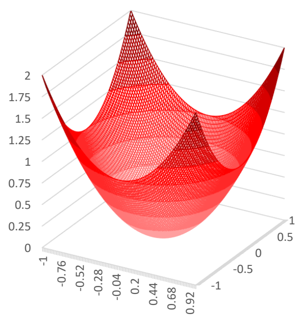
$$\varepsilon^* = \min_{\|\boldsymbol{\beta}\|_0\leq 1} \|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2$$

be the minimum residual error of a sparse solution of the least squares problem. Figure 2 shows in gray the (possibly dense) points satisfying $\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 \leq \varepsilon^*$, and it shows in color the set of feasible points satisfying $\rho(\boldsymbol{\beta}) \leq k$, where ρ 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 $\texttt{MC}_+$ is able to exactly recover an optimal sparse solution for any diagonal dominance parameter δ, despite significant shrinkage ($k < 1$). In contrast, the rank-one constraint $\rho_{\texttt{R1}}(\boldsymbol{\beta}) \leq k$ adapts to the curvature of the error function $\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2$ to induce higher sparsity: in particular, the "natural" constraint $\rho_{\texttt{R1}}(\boldsymbol{\beta}) \leq 1$, with the target sparsity $k = 1$, results in exact recovery without shrinkage in all cases.

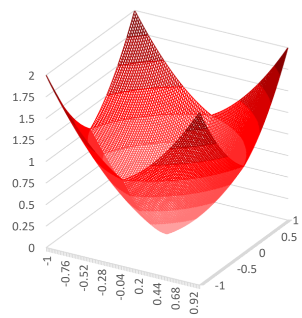
Finally, Figure 3 shows the strength of relaxations of (1) discussed in this paper. The "big-M" relaxation is the natural convex relaxation of (3) obtained by replacing $z \in \{0,1\}^p$ by $z \in [0,1]^p$, used in [7, 11]. The perspective relaxation is the natural convex relaxation of (5), which is the basis of recent methods [8, 30, 49, 58] – note that this formulation may only be used if $\lambda > 0$. The "optimal perspective" relaxation, also referred to as $\texttt{sdp}_\texttt{1}$ in this paper, was explicitly given in [13]. Interestingly, it was recently shown [24] that $\texttt{sdp}_\texttt{1}$ is equivalent to the standard Shor's SDP relaxation [51] for problem (3)– a convex relaxation that has been proven to be very effective at approximating discrete optimization problems [19]. This paper proposes new relaxations $\texttt{sdp}_\texttt{r}$, discussed in §2, which dominate all existing relaxations in terms of strength. It also proposes the new formulation $\texttt{sdp}_\texttt{LB}$, discussed in §4, which is easier to solve than $\texttt{sdp}_\texttt{r}$ but still compares favorably with the "big-M" and perspective formulations.

Outline. The rest of the paper is organized as follows. In §2 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 §3 we derive an explicit form of the regularization penalty for the two-dimensional case. In §4 we discuss the implementation of the proposed relaxation in a conic quadratic framework. In §5 we present computational experiments with synthetic as well as benchmark datasets, demonstrating that *(i)* the proposed formulation delivers near-optimal solutions (with provable 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.

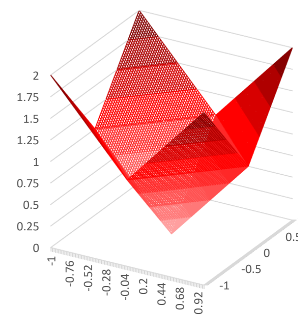
Notation. Define $P = \{1, \ldots, p\}$ and $\boldsymbol{e} \in \mathbb{R}^p$ be the vector of ones. Given $T \subseteq P$ and a vector $\boldsymbol{a} \in \mathbb{R}^p$, define $\boldsymbol{a_T}$ as the subvector of $\boldsymbol{a}$ induced by T, $a_i = \boldsymbol{a_{\{i\}}}$ as the i-th element of $\boldsymbol{a}$, and define $a(T) = \sum_{i \in T} a_i$. Given a symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{p \times p}$, let $\boldsymbol{A_T}$ be the submatrix of $\boldsymbol{A}$ induced by $T \subseteq P$, and let $\mathcal{S}_+^T$ be the set of $|T| \times |T|$ symmetric positive semidefinite matrices, i.e., $\boldsymbol{A_T} \succeq 0 \Leftrightarrow \boldsymbol{A_T} \in \mathcal{S}_+^T$. We use $\boldsymbol{a_T}$ or $\boldsymbol{A_T}$ to make explicit that a given vector or matrix is indexed by the elements of T or $T \times T$, respectively. Given matrices $\boldsymbol{A}$, $\boldsymbol{B}$ of the same dimension, $\boldsymbol{A} \circ \boldsymbol{B}$ denotes the Hadamard product of $\boldsymbol{A}$ and $\boldsymbol{B}$, and $\langle \boldsymbol{A}, \boldsymbol{B} \rangle$ denotes their inner product. Given a vector $\boldsymbol{a} \in \mathbb{R}^n$, let $\text{diag}(\boldsymbol{a})$ be the $n \times n$ diagonal matrix $\boldsymbol{A}$ with $A_{ii} = a_i$. For a set $X \subseteq \mathbb{R}^p$, $\text{cl conv}(X)$ denotes the closure of the convex hull



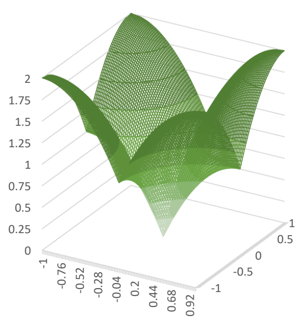
(A) `ridge`



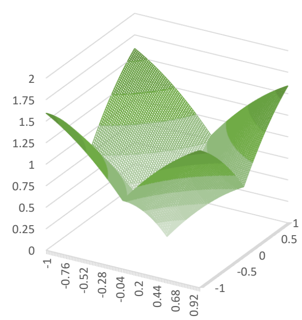
(B) `elastic net`



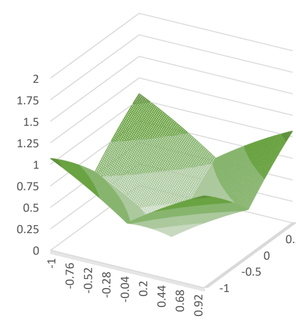
(C) `lasso`



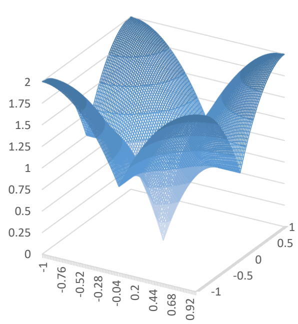
(D) $\mathtt{MC}_+$, $\delta = 1.0$



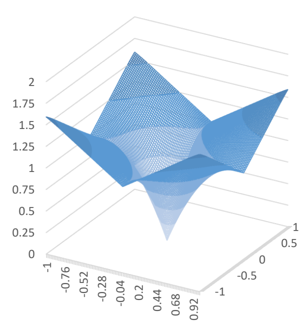
(E) $\mathtt{MC}_+$, $\delta = 0.3$



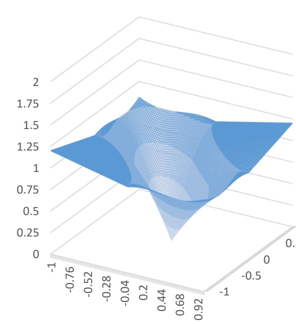
(F) $\mathtt{MC}_+$, $\delta = 0.1$



(G) `R1`, $\delta = 1.0$

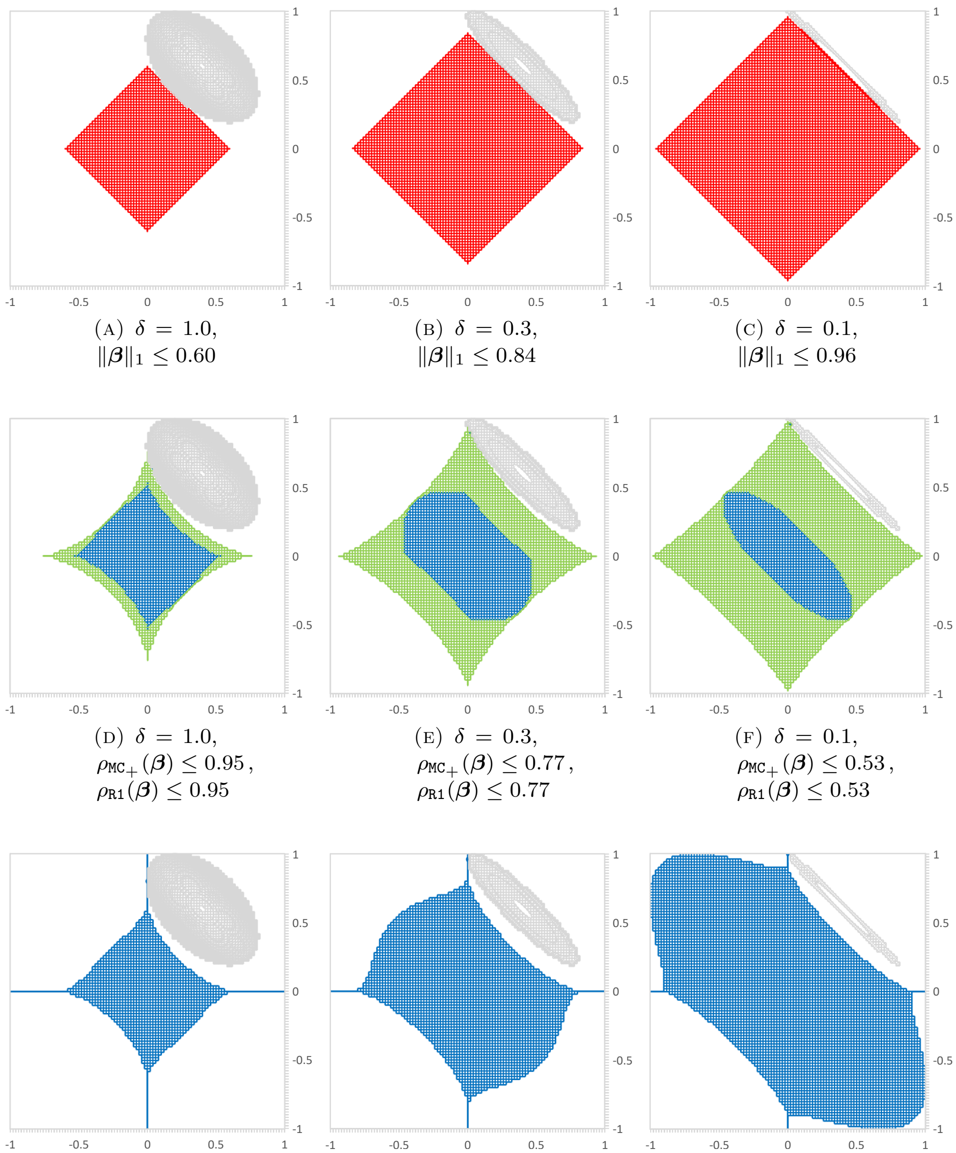


(H) `R1`, $\delta = 0.3$



(I) `R1`, $\delta = 0.1$

FIGURE 1. Graphs of regularization penalties with $p = 2$. The horizontal axes correspond to values of β_1 and β_2, and the vertical axis corresponds to the regularization penalty. The `ridge`, `elastic net`, and `lasso` (top row) regularizations do not depend on the diagonal dominance, but induce substantial bias. The $\mathtt{MC}_+$ regularization (second row) does not induce as much bias, but it depends on the diagonal dominance (δ). The new non-separable, non-convex `R1` regularization (bottom row) induces larger penalties than $\mathtt{MC}_+$ for all diagonal dominance values and is a closer approximation for the exact ℓ_0 penalty.



(A) $\delta = 1.0$, $\|\boldsymbol{\beta}\|_1 \leq 0.60$

(B) $\delta = 0.3$, $\|\boldsymbol{\beta}\|_1 \leq 0.84$

(C) $\delta = 0.1$, $\|\boldsymbol{\beta}\|_1 \leq 0.96$

(D) $\delta = 1.0$, $\rho_{\mathtt{MC}_+}(\boldsymbol{\beta}) \leq 0.95$, $\rho_{\mathtt{R1}}(\boldsymbol{\beta}) \leq 0.95$

(E) $\delta = 0.3$, $\rho_{\mathtt{MC}_+}(\boldsymbol{\beta}) \leq 0.77$, $\rho_{\mathtt{R1}}(\boldsymbol{\beta}) \leq 0.77$

(F) $\delta = 0.1$, $\rho_{\mathtt{MC}_+}(\boldsymbol{\beta}) \leq 0.53$, $\rho_{\mathtt{R1}}(\boldsymbol{\beta}) \leq 0.53$

(G) $\delta = 1.0$, $\rho_{\mathtt{R1}}(\boldsymbol{\beta}) \leq 1.00$

(H) $\delta = 0.3$, $\rho_{\mathtt{R1}}(\boldsymbol{\beta}) \leq 1.00$

(I) $\delta = 0.1$, $\rho_{\mathtt{R1}}(\boldsymbol{\beta}) \leq 1.00$

FIGURE 2. The axes correspond to the sparse solutions satisfying $\|\boldsymbol{\beta}\|_0 \leq 1$. In gray: level sets given by $\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 \leq \varepsilon^*$; in red: feasible region for $\|\boldsymbol{\beta}\|_1 \leq k$; in green: feasible region for $\rho_{\mathtt{MC}_+}(\boldsymbol{\beta}) \leq k$; in blue: feasible region for $\rho_{\mathtt{R1}}(\boldsymbol{\beta}) \leq k$. All `lasso` and $\mathtt{MC}_+$ 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 δ.

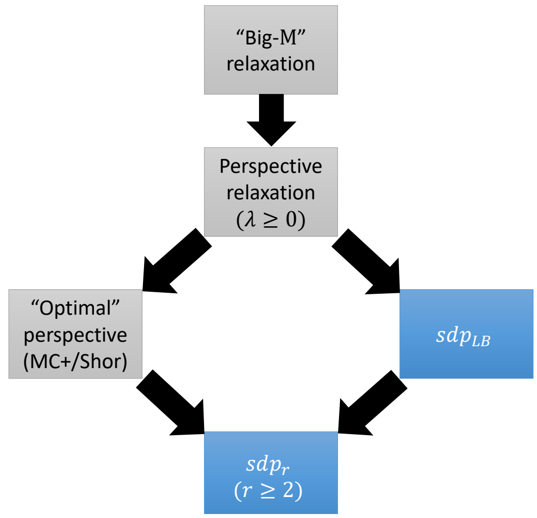


FIGURE 3. Strength of relaxations discussed in the paper. "$A \Rightarrow B$" indicates that B is a stronger relaxation than A, i.e., is a better approximation for the non-convex problem (1). Blue boxes correspond to the new formulations proposed in this paper.

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. In §2.4 we present an explicit SDP representation of the proposed relaxations in an extended space. Finally, in §2.5, we comment on the strength of the proposed relaxations.

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 $A_T \in \mathcal{S}_+^T$, consider the set

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

Proposition 1. *The inequality*

$$\frac{\boldsymbol{\beta}^\top \boldsymbol{A_T} \boldsymbol{\beta}}{z(T)} \leq t \tag{7}$$

is valid for Q_T.

Proof. Let $(\boldsymbol{z}, \boldsymbol{\beta}, t) \in Q_T$, and we verify that inequality (7) is satisfied. First observe that if $\boldsymbol{z} = \boldsymbol{0}$, then $\boldsymbol{\beta} = \boldsymbol{0}$ 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 $\frac{\boldsymbol{\beta}^\top \boldsymbol{A_T}\boldsymbol{\beta}}{z(T)} \leq \boldsymbol{\beta}^\top \boldsymbol{A_T}\boldsymbol{\beta} \leq t$, and inequality (7) is satisfied again. $\square$

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 $\boldsymbol{A_T}$ is rank-one, i.e., $\boldsymbol{A_T} = a_T a_T^\top$ with $a_T = (a_i\ a_j)^\top$ and $\boldsymbol{\beta}^\top \boldsymbol{A_T}\boldsymbol{\beta} = |A_{ij}|\left(a\beta_i^2 \pm 2\beta_i\beta_j + (1/a)\beta_j^2\right)$ for $A_{ij}= a_i a_j$ and $a = a_i/a_j$, then (7) reduces to

$$|A_{ij}|\left(a\beta_i^2 \pm 2\beta_i\beta_j + (1/a)\beta_j^2\right) \leq t(z_i + z_j), \tag{8}$$

one of the inequalities proposed in [35] 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 [2] for additional discussion); and inequality (7) is, in general, weak for arbitrary matrices $\boldsymbol{A_T} \in \mathcal{S}_+^T$. Nonetheless, as we show next, inequality (7) is sufficient to describe the *ideal* (convex hull) description for Q_T if $\boldsymbol{A_T}= \boldsymbol{a_T}\boldsymbol{a_T}^\top$ is a rank-one matrix. Consider the special case of Q_T defined with a rank-one matrix:

$$Q_T^{r1} = \left\{(\boldsymbol{z}, \boldsymbol{\beta}, t) \in \{0,1\}^{|T|} \times \mathbb{R}^{|T|} \times \mathbb{R}_+ : (\boldsymbol{a_T}^\top \boldsymbol{\beta})^2 \leq t,\ \beta_i(1 - z_i) = 0, \forall i \in T\right\}.$$

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

$$cl\ conv(Q_T^{r1}) = \left\{(\boldsymbol{z}, \boldsymbol{\beta}, t) \in [0,1]^{|T|} \times \mathbb{R}^{|T|} \times \mathbb{R}_+ : (\boldsymbol{a_T}^\top \boldsymbol{\beta})^2 \leq t,\ \frac{(\boldsymbol{a_T}^\top \boldsymbol{\beta})^2}{z(T)} \leq t\right\}.$$

Proof. Consider the optimization of an arbitrary linear function over Q_T^{r1} and $\bar{Q}_T := \left\{(\boldsymbol{z}, \boldsymbol{\beta}, t) \in [0,1]^{|T|} \times \mathbb{R}^{|T|} \times \mathbb{R}_+ : (\boldsymbol{a_T}^\top \boldsymbol{\beta})^2 \leq t,\ \frac{(\boldsymbol{a_T}^\top \boldsymbol{\beta})^2}{z(T)} \leq t\right\}$:

$$\min_{(\boldsymbol{z},\boldsymbol{\beta},t)\in Q_T^{r1}} \boldsymbol{u_T}^\top \boldsymbol{z} + \boldsymbol{v_T}^\top \boldsymbol{\beta} + \kappa t, \tag{9}$$

$$\min_{(\boldsymbol{z},\boldsymbol{\beta},t)\in \bar{Q}_T} \boldsymbol{u_T}^\top \boldsymbol{z} + \boldsymbol{v_T}^\top \boldsymbol{\beta} + \kappa t, \tag{10}$$

where $\boldsymbol{u_T}, \boldsymbol{v_T} \in \mathbb{R}^{|T|}$ 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^{r1}, or that (9) and (10) are both unbounded.

Observe that if $\kappa < 0$, then letting $\boldsymbol{z} = \boldsymbol{\beta} = \boldsymbol{0}$ and $t \to \infty$ we see that both problems are unbounded. If $\kappa = 0$ and $\boldsymbol{v_T} = \boldsymbol{0}$, then (10) reduces to $\min_{\boldsymbol{z}\in[0,1]^{|T|}} \boldsymbol{u_T}^\top \boldsymbol{z}$, which has an optimal integral solution $\boldsymbol{z}^*$, and $(\boldsymbol{z}^*, \boldsymbol{0}, 0)$ 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 $\boldsymbol{a_T}$ has no zero entry, we may assume, without loss of generality, that $\boldsymbol{a_T} = \boldsymbol{e_T}$, since otherwise $\boldsymbol{\beta}$ and $\boldsymbol{v_T}$ can be scaled by letting $\bar{\beta}_i = a_i\beta_i$ and $\bar{v}_i = v_i/a_i$ to arrive at an equivalent problem. Moreover, a necessary condition for (9)–(10) to be bounded is that

$$-\infty < \min_{\boldsymbol{\beta}\in\mathbb{R}^{|T|}} \boldsymbol{v_T}^\top \boldsymbol{\beta} \text{ s.t. } \beta(T) = \zeta \tag{11}$$

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

$\boldsymbol{v}_{\boldsymbol{T}}^\top \boldsymbol{\beta} = v_0\beta(T)$ for some scalar v_0. Performing the above simplifications, we find that (10) reduces to

$$\min_{\boldsymbol{z}\in[0,1]^{|T|},\boldsymbol{\beta}\in\mathbb{R}^{|T|},t\in\mathbb{R}} \boldsymbol{u}_{\boldsymbol{T}}^\top \boldsymbol{z} + v_0\beta(T) + t \text{ s.t. } \beta(T)^2 \leq t,\ \beta(T)^2 \leq tz(T). \tag{12}$$

Since the one-dimensional optimization $\min_{\beta\in\mathbb{R}}\left\{v_0\beta+\beta^2\right\}$ 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 $\boldsymbol{z}$ and satisfies $\boldsymbol{\beta}\circ(\boldsymbol{e}-\boldsymbol{z})=0$.

Let $(\boldsymbol{z}^*,\boldsymbol{\beta}^*,t^*)$ be an optimal solution of (12). First note that if $0<z^*(T)<1$, then $(\gamma\boldsymbol{z}^*,\gamma\boldsymbol{\beta}^*,\gamma t^*)$ is feasible for (10) for γ sufficiently close to 1, with objective value $\gamma\left(\boldsymbol{u}_{\boldsymbol{T}}^\top\boldsymbol{z}^* + v_0\beta^*(T)+t^*\right)$. If $\boldsymbol{u}_{\boldsymbol{T}}^\top\boldsymbol{z}^* + v_0\beta^*(T)+t^* \geq 0$, then for $\gamma=0$, $(\gamma\boldsymbol{z}^*,\gamma\boldsymbol{\beta}^*,\gamma t^*)$ has an objective value equal or lower. Otherwise, for $\gamma = 1/z^*(T)$, $(\gamma\boldsymbol{z}^*,\gamma\boldsymbol{\beta}^*,\gamma t^*)$ is feasible and has a lower objective value. Thus, we find that either $\boldsymbol{0}$ 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 $(\bar{\boldsymbol{z}},\boldsymbol{\beta}^*,t^*)$ with $\bar{\boldsymbol{z}} \in \arg\min\{\boldsymbol{u}_{\boldsymbol{T}}^\top\boldsymbol{z} : z^*(T)\geq 1, z\in[0,1]^{|T|}\}$ is also optimal for (12), an in particular there exists an optimal solution with $\bar{\boldsymbol{z}}$ 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 $(\bar{\boldsymbol{z}},\bar{\boldsymbol{\beta}},t^*)$ for (12) that satisfies the complementary constraints, and thus is feasible and optimal for (9). □

Remark 1. Observe that describing cl conv(Q_T^{r1}) requires two nonlinear inequalities in the original space of variables. More compactly, we can specify cl conv(Q_T^{r1}) using a single convex inequality, as

$$\text{cl conv}(Q_T^{r1}) = \left\{(\boldsymbol{z},\boldsymbol{\beta},t)\in[0,1]^{|T|}\times\mathbb{R}^{|T|}\times\mathbb{R}_+ : \frac{(\boldsymbol{a}_{\boldsymbol{T}}^\top\boldsymbol{\beta})^2}{\min\{1,z(T)\}}\leq t\right\}.$$

Finally, we point out that cl conv(Q_T^{r1}) is conic quadratic representable, as $(\boldsymbol{z},\boldsymbol{\beta},t)\in$ cl conv(Q_T^{r1}) if and only if there exists w such that the system

$$\boldsymbol{z}\in[0,1]^{|T|},\ \boldsymbol{\beta}\in\mathbb{R}^{|T|},\ t\in\mathbb{R}_+,\ w\in\mathbb{R}_+,\ w\leq 1,\ w\leq z(T),\ (\boldsymbol{a}_{\boldsymbol{T}}^\top\boldsymbol{\beta})^2\leq tw$$

is feasible, where the last constraint is a rotated conic quadratic constraint and all other constraints are linear. □

2.2. General case.

Now consider again the mixed-integer optimization (3)

$$\boldsymbol{y}^\top\boldsymbol{y} + \min_{\boldsymbol{\beta},\boldsymbol{z},\boldsymbol{u}}\ -2\boldsymbol{y}^\top\boldsymbol{X}\boldsymbol{\beta} + \mu\left(\boldsymbol{e}^\top\boldsymbol{u}\right) + t \tag{13a}$$

$$\text{s.t. } \boldsymbol{\beta}^\top\left(\boldsymbol{X}^\top\boldsymbol{X}+\lambda\boldsymbol{I}\right)\boldsymbol{\beta}\leq t \tag{13b}$$

$$\boldsymbol{e}^\top\boldsymbol{z}\leq k \tag{13c}$$

$$\boldsymbol{\beta}\leq\boldsymbol{u},\ -\boldsymbol{\beta}\leq\boldsymbol{u} \tag{13d}$$

$$\boldsymbol{\beta}\circ(\boldsymbol{e}-\boldsymbol{z})=\boldsymbol{0} \tag{13e}$$

$$\boldsymbol{\beta}\in\mathbb{R}^p,\ \boldsymbol{z}\in\{0,1\}^p,\ \boldsymbol{u}\in\mathbb{R}_+^p,\ t\in\mathbb{R} \tag{13f}$$

where the nonlinear terms of the objective is moved to constraint (13b). A direct application of (7) yields the inequality $\boldsymbol{\beta}^\top\left(\boldsymbol{X}^\top\boldsymbol{X}+\lambda\boldsymbol{I}\right)\boldsymbol{\beta}\leq tz(P)$, which is weak and has no effect when $z(P)\geq 1$. Instead, a more effective approach is to decompose the matrix $\boldsymbol{X}^\top\boldsymbol{X}+\lambda\boldsymbol{I}$ 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 $\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I} = \begin{pmatrix} 25 & 15 & -5 \\ 15 & 18 & 0 \\ -5 & 0 & 11 \end{pmatrix}$.

Then, it follows that

$$\boldsymbol{\beta}^\top \left(\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}\right) \boldsymbol{\beta} = (5\beta_1 + 3\beta_2 - \beta_3)^2 + (3\beta_2 + \beta_3)^2 + 9\beta_3^2$$

and we have the corresponding valid inequality

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

$\square$

The decomposition of $\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}$ illustrated in Example 1 is not unique. Since one does not obtain a strengthening when the denominator is one, it is important to have decomposition both rank-one and sparse. This motivates the question on how to find a decomposition that results in the best convex relaxation, i.e., that maximizes the left hand side of (14). Specifically, let $\mathcal{P} \subseteq 2^P$ be a subset of the power set of P, i.e.,

$$\mathcal{P} = \{T_1, \ldots, T_m\}$$

with $T_h \subseteq P$, $h = 1, \ldots, m$. For each h, define a matrix variable $\boldsymbol{A_h}$ whose nonzero elements correspond to the submatrix induced by T_h, and consider the valid inequality $\phi_{\mathcal{P}}(\boldsymbol{z}, \boldsymbol{\beta}) \leq t$, where $\phi_{\mathcal{P}} : [0,1]^p \times \mathbb{R}^p \to \mathbb{R}$ is defined as

$$\begin{aligned}
\phi_{\mathcal{P}}(\boldsymbol{z}, \boldsymbol{\beta}) := \max_{\boldsymbol{A_h}, \boldsymbol{R}}\ & \boldsymbol{\beta}^\top \boldsymbol{R} \boldsymbol{\beta} + \sum_{h=1}^m \frac{\boldsymbol{\beta}^\top \boldsymbol{A_h} \boldsymbol{\beta}}{\min\{1, z(T_h)\}} && \text{(15a)} \\
\text{s.t.}\ & \sum_{h=1}^m \boldsymbol{A_h} + \boldsymbol{R} = \boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I} && \text{(15b)} \\
& (A_h)_{ij} = 0 \qquad \forall h = 1, \ldots, m,\ i \notin T_h \text{ or } j \notin T_j && \text{(15c)} \\
& \boldsymbol{A_h} \in \mathcal{S}_+^P \qquad \forall h = 1, \ldots, m && \text{(15d)} \\
& \boldsymbol{R} \in \mathcal{S}_+^P, && \text{(15e)}
\end{aligned}$$

where strengthening (7) is applied to each low-dimensional quadratic term $\boldsymbol{\beta}^\top \boldsymbol{A_h} \boldsymbol{\beta}$. For a fixed value of $(\boldsymbol{z}, \boldsymbol{\beta})$, problem (15) finds the best decomposition of the matrix $\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}$ as a sum of positive semidefinite matrices $\boldsymbol{A_h}$, $h = 1, \ldots, m$, and a remainder positive semidefinite matrix $\boldsymbol{R}$ to maximize the strengthening.

For a given decomposition, the objective (15a) is convex in $(\boldsymbol{z}, \boldsymbol{\beta})$, thus $\phi_{\mathcal{P}}$ 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 $\phi_{\mathcal{P}}$, and we assume for simplicity that $\emptyset \in \mathcal{P}$.

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

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

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

$$
\begin{aligned}
\frac{\boldsymbol{\beta_T}^\top \boldsymbol{A_T} \boldsymbol{\beta_T}}{\min\{1, z(T)\}} &= \frac{\boldsymbol{\beta_T}^\top \left(\sum_{i \neq j} (\boldsymbol{L_i} \boldsymbol{L_i}^\top)\right) \boldsymbol{\beta_T}}{\min\{1, z(T)\}} + \frac{\boldsymbol{\beta_T}^\top (\boldsymbol{L_j} \boldsymbol{L_j}^\top) \boldsymbol{\beta_T}}{\min\{1, z(T)\}} \\
&\leq \frac{\boldsymbol{\beta_T}^\top \left(\sum_{i \neq j} (\boldsymbol{L_i} \boldsymbol{L_i}^\top)\right) \boldsymbol{\beta_T}}{\min\{1, z(T)\}} + \frac{\boldsymbol{\beta_T}^\top (\boldsymbol{L_j} \boldsymbol{L_j}^\top) \boldsymbol{\beta_T}}{\min\{1, z(\bar{T}_j)\}}. \qquad (16)
\end{aligned}
$$

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

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

$$
\begin{aligned}
\boldsymbol{y}^\top \boldsymbol{y} + \min_{\boldsymbol{\beta}, \boldsymbol{z}, \boldsymbol{u}} \; & -2\boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{\beta} + \mu \left(\boldsymbol{e}^\top \boldsymbol{u}\right) + \phi_{\mathcal{P}}(\boldsymbol{z}, \boldsymbol{\beta}) && (17a)\\
& \boldsymbol{e}^\top \boldsymbol{z} \leq k && (17b)\\
& \boldsymbol{\beta} \leq \boldsymbol{u}, \; -\boldsymbol{\beta} \leq \boldsymbol{u} && (17c)\\
& \boldsymbol{\beta} \in \mathbb{R}^p, \; \boldsymbol{z} \in [0,1]^p, \; \boldsymbol{u} \in \mathbb{R}^p_+ && (17d)
\end{aligned}
$$

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:

$$
\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2 + \mu \|\boldsymbol{\beta}\|_1 + \rho_{\mathtt{R1}}(\boldsymbol{\beta}; k)
$$

where

$$
\rho_{\mathtt{R1}}(\boldsymbol{\beta}; k) := \min_{\boldsymbol{z} \in [0,1]^p} \phi_{\mathcal{P}}(\boldsymbol{z}, \boldsymbol{\beta}) - \boldsymbol{\beta}^\top (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})\boldsymbol{\beta} \text{ s.t. } \boldsymbol{e}^\top \boldsymbol{z} \leq k. \qquad (18)
$$

is the (non-convex) *rank-one regularization penalty*. Observe that $\rho_{\mathtt{R1}}(\boldsymbol{\beta}; k)$ is the difference of two convex functions: the quadratic function $\boldsymbol{\beta}^\top (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})\boldsymbol{\beta}$ arising from the fitness term and the Tikhonov regularization; and the projection of its convexification $\phi_{\mathcal{P}}(\boldsymbol{z}, \boldsymbol{\beta})$ in the original space of the regression variables $\boldsymbol{\beta}$. As we now show, unlike the usual ℓ_1 penalty, the *rank-one regularization penalty* does not induce a bias when $\boldsymbol{\beta}$ is sparse.

Theorem 2. *If $\|\boldsymbol{\beta}\|_0 \leq k$, then $\rho_{R1}(\boldsymbol{\beta}; k) = 0$.*

Proof. Let $(\boldsymbol{\beta}, \boldsymbol{z}) \in \mathbb{R}^p \times [0,1]^p$, and let $\boldsymbol{R}$ and $\boldsymbol{A_h}$, $h = 1, \ldots, m$, correspond to an optimal solution of (15). Since

$$\begin{aligned}\boldsymbol{\beta}^\top(\boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{I})\boldsymbol{\beta} &= \boldsymbol{\beta}^\top\boldsymbol{R}\boldsymbol{\beta} + \sum_{h=1}^m \boldsymbol{\beta}^\top\boldsymbol{A_h}\boldsymbol{\beta}\\ &\le \boldsymbol{\beta}^\top\boldsymbol{R}\boldsymbol{\beta} + \sum_{h=1}^m \frac{\boldsymbol{\beta}^\top\boldsymbol{A_h}\boldsymbol{\beta}}{\min\{1, z(T_h)\}} = \phi_{\mathcal{P}}(\boldsymbol{z}, \boldsymbol{\beta}),\end{aligned}$$

it follows that $\rho_{\mathtt{R1}}(\boldsymbol{\beta}; k) \ge 0$ for any $\boldsymbol{\beta} \in \mathbb{R}^p$. Now let $\hat{\boldsymbol{\beta}}$ satisfy $\|\hat{\boldsymbol{\beta}}\|_0 \le k$, let $\hat{T} = \left\{i \in P : \hat{\beta}_i \ne 0\right\}$ be the support of $\hat{\boldsymbol{\beta}}$ and let $\hat{\boldsymbol{z}}$ such that $\hat{z}_i = \mathbb{1}_{i \in \hat{T}}$ be the indicator vector of $\hat{T}$. By construction, $\boldsymbol{e}^\top\hat{\boldsymbol{z}} \le k$ and $\hat{\boldsymbol{z}}$ is feasible for problem (18). Moreover

$$\begin{aligned}\rho_{\mathtt{R1}}(\hat{\boldsymbol{\beta}}; k) &\le \phi_{\mathcal{P}}(\hat{\boldsymbol{z}}, \hat{\boldsymbol{\beta}}) - \hat{\boldsymbol{\beta}}^\top(\boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{I})\hat{\boldsymbol{\beta}}\\ &= \sum_{\substack{1 \le h \le m\\ T_h \cap \hat{T} \ne \emptyset}} \left(\frac{\hat{\boldsymbol{\beta}}^\top\boldsymbol{A_h}\hat{\boldsymbol{\beta}}}{\min\{1, \hat{z}(T_h)\}} - \hat{\boldsymbol{\beta}}^\top\boldsymbol{A_h}\hat{\boldsymbol{\beta}}\right) = 0.\end{aligned}$$

Thus, $\rho_{\mathtt{R1}}(\hat{\boldsymbol{\beta}}; k) = 0$. □

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

In §3 we derive an explicit form of $\rho_{\mathtt{R1}}(\boldsymbol{\beta}; k)$ when $p = 2$; Figure 1 plots the graphs of the usual regularization penalties and $\rho_{\mathtt{R1}}$ for the two-dimensional case, and Figure 2 illustrates the better sparsity inducing properties of regularization $\rho_{\mathtt{R1}}$. Deriving explicit forms of $\rho_{\mathtt{R1}}$ is cumbersome for $p \ge 3$. 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 $\boldsymbol{z} \in [0,1]^p$ and $\boldsymbol{\beta} \in \mathbb{R}^p$, we introduce variables $\boldsymbol{w} \in [0,1]^m$ corresponding to terms $w_h := \min\{1, z(T_h)\}$ and $\boldsymbol{B} \in \mathbb{R}^{p \times p}$ corresponding to terms $B_{ij} = \beta_i\beta_j$.

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

$$\begin{aligned}
\boldsymbol{y}^\top\boldsymbol{y} + \min\ & -2\boldsymbol{y}^\top\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}^\top\boldsymbol{u} + \langle \boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{I}, \boldsymbol{B}\rangle && \text{(19a)}\\
\text{s.t. } & \boldsymbol{e}^\top\boldsymbol{z} \le k && \text{(19b)}\\
& \boldsymbol{\beta} \le \boldsymbol{u},\ -\boldsymbol{\beta} \le \boldsymbol{u} && \text{(19c)}\\
& w_h \le \boldsymbol{e}_{\boldsymbol{T_h}}^\top\boldsymbol{z}_{\boldsymbol{T_h}} \qquad \forall h = 1, \ldots, m && \text{(19d)}\\
& w_h\boldsymbol{B}_{\boldsymbol{T_h}} - \boldsymbol{\beta}_{\boldsymbol{T_h}}\boldsymbol{\beta}_{\boldsymbol{T_h}}^\top \in \mathcal{S}_+^{T_h} \qquad \forall h = 1, \ldots, m && \text{(19e)}\\
& \boldsymbol{B} - \boldsymbol{\beta}\boldsymbol{\beta}^\top \in \mathcal{S}_+^P && \text{(19f)}\\
& \boldsymbol{\beta} \in \mathbb{R}^p,\ \boldsymbol{z} \in [0,1]^p,\ \boldsymbol{u} \in \mathbb{R}_+^p,\ \boldsymbol{w} \in [0,1]^m,\ \boldsymbol{B} \in \mathbb{R}^{p \times p}. && \text{(19g)}
\end{aligned}$$

Observe that (19) is indeed an SDP, as

$$w_h \boldsymbol{B_{T_h}} - \boldsymbol{\beta_{T_h}}\boldsymbol{\beta_{T_h}}^\top \in \mathcal{S}_+^{T_h} \Leftrightarrow \begin{pmatrix} w_h & \boldsymbol{\beta_{T_h}}^\top \\ \boldsymbol{\beta_{T_h}} & \boldsymbol{B_{T_h}} \end{pmatrix} \succeq 0. \tag{20}$$

Indeed, from Schur's complement, the right hand side of (20) is equivalent to $w_h \geq 0$ (automatically satisfied) and $\boldsymbol{B_{T_h}} - \frac{1}{w_h}\boldsymbol{\beta_{T_h}}\boldsymbol{\beta_{T_h}}^\top \succeq 0$. Similarly, constraint (19f) can be modeled as $\begin{pmatrix} 1 & \boldsymbol{\beta}^\top \\ \boldsymbol{\beta} & \boldsymbol{B} \end{pmatrix} \succeq 0$. 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 $\boldsymbol{\beta} = 0$, $\boldsymbol{z} = \boldsymbol{e}$, $\boldsymbol{w} > \boldsymbol{0}$ and $\boldsymbol{B} = \boldsymbol{I}$). Adding surplus variables $\boldsymbol{\Gamma}$, $\boldsymbol{\Gamma_h}$ for $h = 1, \ldots, m$, write (19) as

$$\begin{aligned}
\boldsymbol{y}^\top\boldsymbol{y} + \min_{(\boldsymbol{\beta},\boldsymbol{z},\boldsymbol{u},\boldsymbol{w})\in C} \Big\{ -2\boldsymbol{y}^\top\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}^\top\boldsymbol{u} + \min_{\boldsymbol{B},\boldsymbol{\Gamma_h},\boldsymbol{\Gamma}} \langle \boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{I}, \boldsymbol{B}\rangle \Big\} \\
\text{s.t. } & w_h\boldsymbol{B_{T_h}} - \boldsymbol{\Gamma_h} = \boldsymbol{\beta_{T_h}}\boldsymbol{\beta_{T_h}}^\top \quad \forall h \qquad (\boldsymbol{A_h}) \\
& \boldsymbol{B} - \boldsymbol{\Gamma} = \boldsymbol{\beta}\boldsymbol{\beta}^\top \qquad (\boldsymbol{R}) \\
& \boldsymbol{\Gamma_h} \in \mathcal{S}_+^{T_h} \qquad \forall h \\
& \boldsymbol{\Gamma} \in \mathcal{S}_+^P \\
& \boldsymbol{B} \in \mathbb{R}^{p\times p},
\end{aligned}$$

where $C = \big\{\boldsymbol{\beta} \in \mathbb{R}^p, \boldsymbol{z} \in [0,1]^p, \boldsymbol{u} \in \mathbb{R}_+^p, \boldsymbol{w} \in [0,1]^m : (19\text{b}), (19\text{c}), (19\text{d})\big\}$. Using conic duality for the inner minimization problem, we find the dual

$$\begin{aligned}
\boldsymbol{y}^\top\boldsymbol{y} + \min_{(\boldsymbol{\beta},\boldsymbol{z},\boldsymbol{u},\boldsymbol{w})\in C} \Big\{ -2\boldsymbol{y}^\top\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}^\top\boldsymbol{u} + \max_{\boldsymbol{A_h},\boldsymbol{R}} \langle \boldsymbol{\beta}\boldsymbol{\beta}^\top, \boldsymbol{R}\rangle + \sum_{h=1}^m \langle \boldsymbol{\beta}\boldsymbol{\beta}^\top, \boldsymbol{A_h}\rangle \Big\} \\
\text{s.t. } & \sum_{j=1}^m w_h\boldsymbol{A_h} + \boldsymbol{R} = \boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{I} \\
& (A_h)_{ij} = 0 \qquad \text{for } i \notin T_h \text{ or } j \notin T_h \\
& \boldsymbol{A_h} \in \mathcal{S}_+^P \quad \forall T \in \mathcal{P} \\
& \boldsymbol{R} \in \mathcal{S}_+^P.
\end{aligned}$$

After substituting $\bar{\boldsymbol{A}}_{\boldsymbol{h}} = w_h\boldsymbol{A_h}$ and noting that there exists an optimal solution with $w_h = \min\{1, z(T_h)\}$, 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 ℓ_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^p semidefinite constraints and is impractical. Proposition 2 suggests using down-monotone sets $\mathcal{P}$ with limited size

$$
(\mathtt{sdp_r}) \quad
\begin{aligned}
\boldsymbol{y}^\top \boldsymbol{y} + \min\ & -2\boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}^\top \boldsymbol{u} + \langle \boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}, \boldsymbol{B} \rangle && && \text{(21a)}\\
\text{s.t. } & \boldsymbol{e}^\top \boldsymbol{z} \le k && && \text{(21b)}\\
& \boldsymbol{\beta} \le \boldsymbol{u},\ -\boldsymbol{\beta} \le \boldsymbol{u} && && \text{(21c)}\\
& 0 \le w_T \le \min\{1, \boldsymbol{e}_{\boldsymbol{T}}^\top \boldsymbol{z}_{\boldsymbol{T}}\} && \forall T : |T| \le r && \text{(21d)}\\
& w_T \boldsymbol{B}_{\boldsymbol{T}} - \boldsymbol{\beta}_{\boldsymbol{T}} \boldsymbol{\beta}_{\boldsymbol{T}}^\top \in \mathcal{S}_+^T && \forall T : |T| \le r && \text{(21e)}\\
& \boldsymbol{B} - \boldsymbol{\beta}\boldsymbol{\beta}^\top \in \mathcal{S}_+^P && && \text{(21f)}\\
& \boldsymbol{\beta} \in \mathbb{R}^p,\ \boldsymbol{z} \in [0,1]^p,\ \boldsymbol{u} \in \mathbb{R}_+^p,\ \boldsymbol{B} \in \mathbb{R}^{p\times p}, \boldsymbol{w} \in \mathbb{R}^m && && \text{(21g)}
\end{aligned}
$$

for some $r \in \mathbb{Z}_+$, where $m = |\{T \subseteq P : |T| \le r\}|$. Note that in the above formulation, w_T is a scalar corresponding to the T-th coordinate of the m-dimensional vector $\boldsymbol{w}$. If $r = 1$, then $\mathtt{sdp_1}$ reduces to the formulation of the optimal `perspective relaxation` proposed in [13], which is equivalent to using $\mathtt{MC}_+$ regularization. Our computations experiments show that whereas $\mathtt{sdp_1}$ may be a weak convex relaxation for problems with low diagonal dominance, $\mathtt{sdp_2}$ achieves excellent relaxation bounds even for the case of low diagonal-dominance within reasonable compute times. For clarity, we give the explicit form of the case $\mathtt{sdp_2}$:

$$
(\mathtt{sdp_2}) \quad
\begin{aligned}
\boldsymbol{y}^\top \boldsymbol{y} + \min\ & -2\boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}^\top \boldsymbol{u} + \langle \boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}, \boldsymbol{B} \rangle && && \text{(22a)}\\
\text{s.t. } & \boldsymbol{e}^\top \boldsymbol{z} \le k && && \text{(22b)}\\
& \boldsymbol{\beta} \le \boldsymbol{u},\ -\boldsymbol{\beta} \le \boldsymbol{u} && && \text{(22c)}\\
& \begin{pmatrix} z_i & \beta_i \\ \beta_i & B_{ii} \end{pmatrix} \succeq 0 && \forall i = 1, \ldots, p && \text{(22d)}\\
& 0 \le w_{ij} \le \min\{1, z_i + z_j\} && \forall i < j && \text{(22e)}\\
& \begin{pmatrix} w_{ij} & \beta_i & \beta_j \\ \beta_i & B_{ii} & B_{ij} \\ \beta_j & B_{ij} & B_{jj} \end{pmatrix} \succeq 0 && \forall i < j && \text{(22f)}\\
& \begin{pmatrix} 1 & \boldsymbol{\beta}^\top \\ \boldsymbol{\beta} & \boldsymbol{B} \end{pmatrix} \succeq 0 && && \text{(22g)}\\
& \boldsymbol{\beta} \in \mathbb{R}^p,\ \boldsymbol{z} \in [0,1]^p,\ \boldsymbol{u} \in \mathbb{R}_+^p,\ \boldsymbol{B} \in \mathbb{R}^{p\times p}, \boldsymbol{w} \in \mathbb{R}^{p(p-1)/2}. && && \text{(22h)}
\end{aligned}
$$

Constraints (22e) and (22f) correspond exactly to constraints (21d)-(21e) for the cases with $|T| = 2$. Moreover, for cases where T is a singleton, that is, $T = \{i\}$ for some $i \in \{1, \ldots, p\}$, constraints $0 \le w_i \le z_i$ can be omitted since it can be easily verified that $w_i = z_i$ in any optimal solution; thus constraints (21e) reduce, after substitution of w_i, to constraints (22d). Finally, constraints (22b), (22c) and (22g) correspond to constraints (21b), (21c) and (21f), respectively.

2.5. **Additional notes on the strength of the relaxations.** Just like `lasso` can be interpreted as the best possible convex relaxation obtained from the set with bounded continuous variables $\{z \in \{0,1\}, \beta \in [0,1] : \beta(1-z) = 0\}$ and the `perspective relaxation` is the best possible relaxation obtained from set Q_1, the relaxations proposed in this section are the best possible convex relaxations obtained from study of Q_T^{r1}. Since Q_T^{r1} generalizes these two simpler sets, it follows that the proposed formulations are a better approximation of (3) than `lasso` and the `perspective relaxation`. Nonetheless, Q_T^{r1} is still considerably simpler than the feasible region of (3) – which involves constraints on the binary variables and general convex quadratic functions as described in set Q_T. We now briefly discuss two recent results that shed additional light into the strength of the formulations.

Wei et al. [56] consider a generalization of Q_T^{r1} in which the binary variables are subject to additional constraints. The authors find that with the k-sparsity constraint $\sum_{i=1}^p z_i \leq k$, the rank-one relaxation described in Theorem 1 is still the best possible formulation. Moreover, the authors show that the convex hull for other classes of constraints (e.g., hierarchical constraints [29]) has a similar structure to cl conv(Q_T^{r1}), and formulation $\texttt{sdp}_\texttt{r}$ can be extended to those cases.

Wei et al. [55] study set Q_T, and show that cl conv(Q_T) can be described in an extended formulation with $O(p^2)$ additional variables as the intersection of a single conic constraint and a polyhedron Ψ. The proof of this result is not constructive: an explicit description of Ψ is not given, as it requires an exponential number of linear inequalities. Nonetheless, the authors show that formulations $\texttt{sdp}_\texttt{r}$ can be interpreted as relaxations in this extended space obtained by adding linear inequalities that are *guaranteed* to define high-dimensional faces of Ψ – the dimension of the face is larger for small values of r, providing additional theoretical justification that the nonlinear inequalities (22d) and (22f) are in fact strong approximations of cl conv(Q_T).

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 κ:

$$\begin{aligned} \boldsymbol{y}^\top \boldsymbol{y} + \min\ & -2\boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}^\top \boldsymbol{u} + \phi_{\mathcal{P}}(\boldsymbol{z},\boldsymbol{\beta}) + \kappa \boldsymbol{e}^\top \boldsymbol{z} && \text{(23a)} \\ & \boldsymbol{\beta} \leq \boldsymbol{u},\ -\boldsymbol{\beta} \leq \boldsymbol{u} && \text{(23b)} \\ & \boldsymbol{\beta} \in \mathbb{R}^P,\ \boldsymbol{z} \in [0,1]^P,\ \boldsymbol{u} \in \mathbb{R}_+^P, && \text{(23c)} \end{aligned}$$

where $p = 2$, and

$$\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I} = \begin{pmatrix} 1+\delta_1 & 1 \\ 1 & 1+\delta_2 \end{pmatrix}. \tag{24}$$

Observe that assumption (24) is without loss of generality, provided that $\boldsymbol{X}^\top \boldsymbol{X}$ 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 $\bar{\beta}_1 = \alpha\beta_1$ and $\bar{\beta}_2 = (a_{12}/\alpha)\beta_2$ with $|a_{12}|/a_2 \leq \alpha \leq a_1$ yields a quadratic form satisfying (24). Also note that we are using the Lagrangean form instead of the cardinality constrained form given in (18) for simplicity; however, since $\phi_{\mathcal{P}}(\boldsymbol{z},\boldsymbol{\beta})$ is convex in $\boldsymbol{z}$, there exists a value of κ such that both forms are equivalent, i.e., result in the same optimal solutions $\hat{\boldsymbol{\beta}}$ for the regression problem, and the objective values differ by the constant $\kappa \cdot k$.

If $\mathcal{P} = \{\emptyset, \{1\}, \{2\}\}$, then (23) reduces to a perspective strengthening of the form

$$\boldsymbol{y}'\boldsymbol{y} + \min_{\boldsymbol{z}\in[0,1]^2, \boldsymbol{\beta}\in\mathbb{R}^2,} -2\boldsymbol{y}'\boldsymbol{X}\boldsymbol{\beta} + (\beta_1+\beta_2)^2 + \delta_1\frac{\beta_1^2}{z_1} + \delta_2\frac{\beta_2^2}{z_2} + \mu\|\boldsymbol{\beta}\|_1 + \kappa\|\boldsymbol{z}\|_1. \tag{25}$$

The links between (25) and regularization were studied[3] in [13].

Proposition 3 (Dong et al. [13])**.** *Problem* (25) *is equivalent to the regularization problem*

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

where

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

Regularization $\rho_{\mathtt{MC}_+}$ 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 [59]; and, if $\lambda=\delta_1=\delta_2$, then the regularization given in Proposition 3 reduces to the reverse Huber penalty derived in [49]. Observe that the regularization function $\rho_{\mathtt{MC}_+}$ is highly dependent on the diagonal dominance $\boldsymbol{\delta}$: specifically, in the low diagonal dominance setting with $\boldsymbol{\delta}=\boldsymbol{0}$, we find that $\rho_{\mathtt{MC}_+}(\boldsymbol{\beta};\kappa,\boldsymbol{0})=0$.

We now consider conic formulation (23) for the case $\mathcal{P}=\{\emptyset,\{1\},\{2\},\{1,2\}\}$, corresponding to the full rank-one strengthening:

$$\boldsymbol{y}^\top\boldsymbol{y}+\min_{\boldsymbol{z}\in[0,1]^2,\boldsymbol{\beta}\in\mathbb{R}^2,}-2\boldsymbol{y}^\top\boldsymbol{X}\boldsymbol{\beta}+\frac{(\beta_1+\beta_2)^2}{\min\{1,z_1+z_2\}}+\delta_1\frac{\beta_1^2}{z_1}+\delta_2\frac{\beta_2^2}{z_2}+\mu\|\boldsymbol{\beta}\|_1+\kappa\|\boldsymbol{z}\|_1. \tag{26}$$

Proposition 4. *Problem* (26) *is equivalent to the regularization problem*

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^2} \|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda\|\boldsymbol{\beta}\|_2^2+\mu\|\boldsymbol{\beta}\|_1+\rho_{R1}(\boldsymbol{\beta};\kappa,\boldsymbol{\delta})$$

where

$$\rho_{R1}(\boldsymbol{\beta};\kappa,\boldsymbol{\delta})=\begin{cases}2\sqrt{\kappa}\sqrt{\boldsymbol{\beta}'(\boldsymbol{X}^\top\boldsymbol{X}+\lambda\boldsymbol{I})\boldsymbol{\beta}+2\sqrt{\delta_1\delta_2}|\beta_1\beta_2|}-\boldsymbol{\beta}'(\boldsymbol{X}^\top\boldsymbol{X}+\lambda\boldsymbol{I})\boldsymbol{\beta} & \\ \qquad \text{if } \boldsymbol{\beta}'(\boldsymbol{X}^\top\boldsymbol{X}+\lambda\boldsymbol{I})\boldsymbol{\beta}+2\sqrt{\delta_1\delta_2}|\beta_1\beta_2|<\kappa & \\ \kappa+2\sqrt{\delta_1\delta_2}|\beta_1\beta_2| & \\ \qquad \text{if } \left(\sqrt{\delta_1}|\beta_1|+\sqrt{\delta_2}|\beta_2|\right)^2\le\kappa\le\boldsymbol{\beta}'(\boldsymbol{X}^\top\boldsymbol{X}+\lambda\boldsymbol{I})\boldsymbol{\beta}+2\sqrt{\delta_1\delta_2}|\beta_1\beta_2| & \\ \sum_{i=1}^2\left(2\sqrt{\kappa\delta_i}|\beta_i|-\delta_i\beta_i^2\right) & \\ \qquad \text{if } \left(\sqrt{\delta_1}|\beta_1|+\sqrt{\delta_2}|\beta_2|\right)^2>\kappa \ \&\ \delta_i\beta_i^2\le\kappa,\ i=1,2 & \\ \kappa+\sqrt{\kappa\delta_i}|\beta_i|-\delta_i\beta_i^2 \qquad \text{if } \delta_i\beta_i^2\le\kappa \ \&\ \delta_j\beta_j^2>\kappa & \\ 2\kappa \qquad \text{if } \delta_i\beta_i^2>\kappa,\ i=1,2. & \end{cases}$$

Observe that, unlike $\rho_{\mathtt{MC}_+}$, the function $\rho_{\mathtt{R1}}$ is not separable in β_1 and β_2 and does not vanish when $\boldsymbol{\delta}=0$: indeed, for $\boldsymbol{\delta}=0$ we find that

$$\rho_{\mathtt{R1}}(\boldsymbol{\beta};\kappa,\boldsymbol{0})=\begin{cases}2\sqrt{\kappa}\sqrt{\boldsymbol{\beta}'(\boldsymbol{X}^\top\boldsymbol{X}+\lambda\boldsymbol{I})\boldsymbol{\beta}}-\boldsymbol{\beta}'(\boldsymbol{X}^\top\boldsymbol{X}+\lambda\boldsymbol{I})\boldsymbol{\beta} & \text{if } \boldsymbol{\beta}'(\boldsymbol{X}^\top\boldsymbol{X}+\lambda\boldsymbol{I})\boldsymbol{\beta}<\kappa\\ \kappa & \text{if } 0\le\kappa\le\boldsymbol{\beta}'(\boldsymbol{X}^\top\boldsymbol{X}+\lambda\boldsymbol{I})\boldsymbol{\beta}.\end{cases}$$

Proof of Proposition 4. We prove the result by projecting out the z variables in (26), i.e., giving closed form solutions for them. There are three cases to consider, depending on the optimal value for z_1+z_2.

[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.

• Case 1: $z_1+z_2<1$. In this case, we find by setting the derivatives of the objective in (26) with respect to z_1 and z_2 that

$$\left.\begin{aligned}\kappa-\delta_1\frac{\beta_1^2}{z_1^2}-\frac{(\beta_1+\beta_2)^2}{(z_1+z_2)^2}=0\\ \kappa-\delta_2\frac{\beta_2^2}{z_2^2}-\frac{(\beta_1+\beta_2)^2}{(z_1+z_2)^2}=0\end{aligned}\right\}\implies z_2=\sqrt{\frac{\delta_2}{\delta_1}}\frac{|\beta_2|}{|\beta_1|}z_1.$$

Define $\bar z:=\frac{z_1}{\sqrt{\delta_1}|\beta_1|}$, so $z_2=\sqrt{\delta_2}|\beta_2|\bar z$, and $z_1+z_2=\left(\sqrt{\delta_1}|\beta_1|+\sqrt{\delta_2}|\beta_2|\right)\bar z$. Moreover, we find that (26) reduces to

$$\begin{aligned}\boldsymbol{y}^\top\boldsymbol{y}+\min_{\bar z>0,\boldsymbol{\beta}\in\mathbb{R}^2}&-2\boldsymbol{y}^\top\boldsymbol{X}\boldsymbol{\beta}+\mu\|\boldsymbol{\beta}\|_1\\&+\frac{(\beta_1+\beta_2)^2+\left(\sqrt{\delta_1}|\beta_1|+\sqrt{\delta_2}|\beta_2|\right)^2}{\left(\sqrt{\delta_1}|\beta_1|+\sqrt{\delta_2}|\beta_2|\right)\bar z}+\kappa\left(\sqrt{\delta_1}|\beta_1|+\sqrt{\delta_2}|\beta_2|\right)\bar z.\end{aligned}\tag{27}$$

An optimal solution of (27) is attained at

$$\bar z^*=\sqrt{\frac{\frac{(\beta_1+\beta_2)^2+\left(\sqrt{\delta_1}|\beta_1|+\sqrt{\delta_2}|\beta_2|\right)^2}{\left(\sqrt{\delta_1}|\beta_1|+\sqrt{\delta_2}|\beta_2|\right)}}{\kappa\left(\sqrt{\delta_1}|\beta_1|+\sqrt{\delta_2}|\beta_2|\right)}}=\frac{\sqrt{(\beta_1+\beta_2)^2+\left(\sqrt{\delta_1}|\beta_1|+\sqrt{\delta_2}|\beta_2|\right)^2}}{\sqrt{\kappa}\left(\sqrt{\delta_1}|\beta_1|+\sqrt{\delta_2}|\beta_2|\right)}$$

with objective value

$$\begin{aligned}&\boldsymbol{y}^\top\boldsymbol{y}+\min_{\boldsymbol{\beta}\in\mathbb{R}^2}-2\boldsymbol{y}^\top\boldsymbol{X}\boldsymbol{\beta}+\mu\|\boldsymbol{\beta}\|_1+2\sqrt{\kappa}\sqrt{(\beta_1+\beta_2)^2+\left(\sqrt{\delta_1}|\beta_1|+\sqrt{\delta_2}|\beta_2|\right)^2}\\&=\min_{\boldsymbol{\beta}\in\mathbb{R}^2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda\|\boldsymbol{\beta}\|_2^2+\mu\|\boldsymbol{\beta}\|_1\\&\quad+\left(2\sqrt{\kappa}\sqrt{(\beta_1+\beta_2)^2+\left(\sqrt{\delta_1}|\beta_1|+\sqrt{\delta_2}|\beta_2|\right)^2}-(\beta_1+\beta_2)^2-\delta_1\beta_1^2-\delta_2\beta_2^2\right).\end{aligned}$$

Finally, this case happens when $z_1+z_2<1\Leftrightarrow(\beta_1+\beta_2)^2+(\sqrt{\delta_1}|\beta_1|+\sqrt{\delta_2}|\beta_2)^2<\kappa$.

• Case 2: $z_1+z_2>1$. In this case, we find by setting the derivatives of the objective in (26) with respect to z_1 and z_2 that $\bar z_i=\sqrt{\frac{\delta_i}{\kappa}}|\beta_i|$ for $i=1,2$. Thus, in this case, for an optimal solution $\boldsymbol{z}^*$ of (26), we have $z_i^*=\min\{\bar z_i,1\}$, and problem (26) reduces to

$$\begin{aligned}\boldsymbol{y}^\top\boldsymbol{y}+\min_{\boldsymbol{\beta}\in\mathbb{R}^2,}&-2\boldsymbol{y}^\top\boldsymbol{X}\boldsymbol{\beta}+(\beta_1+\beta_2)^2+\sum_{i=1}^2\max\left\{\delta_i\beta_i^2,\sqrt{\kappa\delta_i}|\beta_i|\right\}+\mu\|\boldsymbol{\beta}\|_1\\&+\sum_{i=1}^2\min\left\{\sqrt{\kappa\delta_i}|\beta_i|,\kappa\right\}\end{aligned}$$

$$=\min_{\boldsymbol{\beta}\in\mathbb{R}^2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda\|\boldsymbol{\beta}\|_2^2+\mu\|\boldsymbol{\beta}\|_1+\sum_{i=1}^2\left(\max\left\{\delta_i\beta_i^2,\sqrt{\kappa\delta_i}|\beta_i|\right\}+\min\left\{\sqrt{\kappa\delta_i}|\beta_i|,\kappa\right\}-\delta_i\beta_i^2\right)$$

$$=\min_{\boldsymbol{\beta}\in\mathbb{R}^2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda\|\boldsymbol{\beta}\|_2^2+\mu\|\boldsymbol{\beta}\|_1+\begin{cases}\sum_{i=1}^2\left(2\sqrt{\kappa\delta_i}|\beta_i|-\delta_i\beta_i^2\right)&\text{if }\delta_i\beta_i^2\le\kappa,\ i=1,2\\\sqrt{\kappa\delta_i}|\beta_i|-\delta_i\beta_i^2+\kappa&\text{if }\delta_i\beta_i^2\le\kappa\ \&\ \delta_j\beta_j^2>\kappa\\2\kappa&\text{if }\delta_i\beta_i^2>\kappa,\ i=1,2.\end{cases}$$

Finally, this case happens when $z_1+z_2>1 \Leftrightarrow \left(\sqrt{\delta_1}|\beta_1|+\sqrt{\delta_2}|\beta_2|\right)^2>\kappa$. 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 (26) reduces to

$$\boldsymbol{y}^\top\boldsymbol{y}+\min_{0\le z_1\le 1,\boldsymbol{\beta}\in\mathbb{R}^2,} -2\boldsymbol{y}^\top\boldsymbol{X}\boldsymbol{\beta}+(\beta_1+\beta_2)^2+\delta_1\frac{\beta_1^2}{z_1}+\delta_2\frac{\beta_2^2}{1-z_1}+\mu\|\boldsymbol{\beta}\|_1+\kappa. \tag{28}$$

Setting derivative with respect to z_1 in (28) to 0, we have

$$\begin{aligned}0&=\delta_1\beta_1^2(1-z_1)^2-\delta_2\beta_2^2z_1^2\\&=\delta_1\beta_1^2-2\delta_1\beta_1^2z_1+(\delta_1\beta_1^2-\delta_2\beta_2^2)z_1^2.\end{aligned}$$

Thus, we find that

$$\begin{aligned}z_1&=\frac{2\delta_1\beta_1^2\pm\sqrt{4\delta_1^2\beta_1^4-4\delta_1\beta_1^2(\delta_1\beta_1^2-\delta_2\beta_2^2)}}{2\left(\delta_1\beta_1^2-\delta_2\beta_2^2\right)}\\&=\frac{\delta_1\beta_1^2\pm\sqrt{\delta_1\delta_2}|\beta_1\beta_2|}{\delta_1\beta_1^2-\delta_2\beta_2^2}=\frac{\sqrt{\delta_1}|\beta_1|(\sqrt{\delta_1}|\beta_1|\pm\sqrt{\delta_2}|\beta_2|)}{(\sqrt{\delta_1}|\beta_1|+\sqrt{\delta_2}|\beta_2|)(\sqrt{\delta_1|\beta_1|}-\sqrt{\delta_2}|\beta_2|)}.\end{aligned}$$

Moreover, since $0\le z_1\le 1$, 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 (28), we find the equivalent form

$$\begin{aligned}&\boldsymbol{y}^\top\boldsymbol{y}+\min_{\boldsymbol{\beta}\in\mathbb{R}^2,} -2\boldsymbol{y}^\top\boldsymbol{X}\boldsymbol{\beta}+(\beta_1+\beta_2)^2+\left(\sqrt{\delta_1}|\beta_1|+\sqrt{\delta_2}|\beta_2|\right)^2+\mu\|\boldsymbol{\beta}\|_1+\kappa\\=&\min_{\boldsymbol{\beta}\in\mathbb{R}^2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda\|\boldsymbol{\beta}\|_2^2+\mu\|\boldsymbol{\beta}\|_1+\kappa+2\sqrt{\delta_1\delta_2}|\beta_1\beta_2|.\end{aligned}$$

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

The plots of $\rho_{\texttt{MC}_+}$ and $\rho_{\texttt{R1}}$ 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 $\texttt{sdp}_\texttt{r}$, 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, conic quadratic optimization problems are considerably easier to solve than semidefinite optimization problems, thus scaling to larger dimensions. Moreover there exist off-the-shelf mixed-integer conic quadratic optimization solvers that are actively maintained and improved by numerous software vendors. In this section we show how the proposed conic relaxations, and specifically $\texttt{sdp}_2$, can be implemented in a conic quadratic framework. The resulting convex formulations can then be directly used as a fast approximation to the SDP formulations presented in §2, and pave the way towards an integration with branch-and-bound solvers[4].

[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.

4.1. **Two-dimensional PSD constraints.** Constraint (22d), $\beta_i^2 \le z_i B_{ii}$, is a rotated cone constraint as $z_i \ge 0$ and $B_{ii} \ge 0$ in any feasible solution of (21), and thus conic quadratic representable.

4.2. **Three-dimensional PSD constraints.** As we now show, constraints (22f) can be accurately approximated using conic quadratic constraints.

Proposition 5. *Problem $\texttt{sdp}_2$ is equivalent to the optimization problem*

$$
\begin{aligned}
\boldsymbol{y}^\top \boldsymbol{y} + \min \ & -2\boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}^\top \boldsymbol{u} + \langle \boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}, \boldsymbol{B}\rangle && \text{(29a)}\\
\text{s.t. } & \boldsymbol{e}^\top \boldsymbol{z} \le k && \text{(29b)}\\
& \boldsymbol{\beta} \le \boldsymbol{u},\ -\boldsymbol{\beta} \le \boldsymbol{u} && \text{(29c)}\\
& z_i B_{ii} \ge \beta_i^2 \qquad \forall i \in P && \text{(29d)}\\
& 0 \le w_{ij} \le 1,\ w_{ij} \le z_i + z_j \qquad \forall i \ne j && \text{(29e)}\\
& 0 \ge \max_{\alpha \ge 0} \left\{ \frac{\alpha\beta_i^2 + 2\beta_i\beta_j + \beta_j^2/\alpha}{w_{ij}} - 2B_{ij} - \alpha B_{ii} - B_{jj}/\alpha \right\} \forall i \ne j && \text{(29f)}\\
& 0 \ge \max_{\alpha \ge 0} \left\{ \frac{\alpha\beta_i^2 - 2\beta_i\beta_j + \beta_j^2/\alpha}{w_{ij}} + 2B_{ij} - \alpha B_{ii} - B_{jj}/\alpha \right\} \forall i \ne j && \text{(29g)}\\
& \boldsymbol{B} - \boldsymbol{\beta}\boldsymbol{\beta}' \in \mathcal{S}_+^P && \text{(29h)}\\
& \boldsymbol{\beta} \in \mathbb{R}^p,\ \boldsymbol{z} \in [0,1]^p,\ \boldsymbol{u} \in \mathbb{R}_+^p,\ \boldsymbol{B} \in \mathbb{R}^{p \times p}. && \text{(29i)}
\end{aligned}
$$

Proof. It suffices to compute the optimal value of α in (29f)–(29g). Observe that the rhs of (29f) can be written as

$$
v = \frac{2\beta_i\beta_j}{w_{ij}} - 2B_{ij} - \min_{\alpha \ge 0} \left\{ \alpha \left(B_{ii} - \frac{\beta_i^2}{w_{ij}} \right) + \frac{1}{\alpha} \left(B_{jj} - \frac{\beta_j^2}{w_{ij}} \right) \right\}. \tag{30}
$$

Moreover, in an optimal solution of (29), we have that $w_{ij} = \min\{1, z_i + z_j\}$. Thus, due to constraints (29d), we find that $B_{ii} - \beta_i^2/w_{ij} \ge 0$ in optimal solutions of (29), 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 (30) is $v = 2\beta_i\beta_j/\min\{1, z_i+z_j\} - 2B_{ij}$, by setting $\alpha \to \infty$ or $\alpha = 0$, respectively. Otherwise, the optimal α equals

$$
\alpha = \sqrt{\frac{B_{jj}w_{ij} - \beta_j^2}{B_{ii}w_{ij} - \beta_i^2}}, \tag{31}
$$

with the objective value

$$
v = \frac{2\beta_i\beta_j}{w_{ij}} - 2B_{ij} - 2\sqrt{\left(B_{ii} - \frac{\beta_i^2}{w_{ij}} \right)\left(B_{jj} - \frac{\beta_j^2}{w_{ij}} \right)}.
$$

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 (29f) reduces to

$$
0 \ge \beta_i\beta_j - B_{ij}w_{ij} - \sqrt{(B_{ii}w_{ij} - \beta_i^2)(B_{jj}w_{ij} - \beta_j^2)}. \tag{32}
$$

Similarly, it can be shown that constraint (29g) reduces to

$$
0 \ge -\beta_i\beta_j + B_{ij}w_{ij} - \sqrt{(B_{ii}w_{ij} - \beta_i^2)(B_{jj}w_{ij} - \beta_j^2)}. \tag{33}
$$

More compactly, constraints (32)–(33) are equivalent to

$$\left(w_{ij}B_{ii}-\beta_i^2\right)\left(w_{ij}B_{jj}-\beta_j^2\right)\geq\left(w_{ij}B_{ij}-\beta_i\beta_j\right)^2. \tag{34}$$

Moreover, note that constraints (21e) with $T=\{i,j\}$ are equivalent to

$$\begin{aligned}&\begin{pmatrix}w_{ij}B_{ii}-\beta_i^2 & w_{ij}B_{ij}-\beta_i\beta_j\\ w_{ij}B_{ij}-\beta_i\beta_j & w_{ij}B_{jj}-\beta_j^2\end{pmatrix}\in\mathcal{S}_+^2\\ \Leftrightarrow\ & w_{ij}B_{ii}-\beta_i^2\geq 0,\ w_{ij}B_{jj}-\beta_j^2\geq 0,\ \text{and (34).}\end{aligned}$$

Since the first two constraints are implied by (29d) and $w_{ij}=\min\{1,z_i+z_j\}$ in optimal solutions, the proof is complete. □

Observe that, for any fixed value of α, constraints (29f)–(29g) are conic quadratic representable. Thus, we can obtain relaxations of (29) of the form

$$\begin{aligned}\boldsymbol{y}^\top\boldsymbol{y}+\min\ & -2\boldsymbol{y}^\top\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{e}^\top\boldsymbol{u}+\langle\boldsymbol{X}^\top\boldsymbol{X}+\lambda\boldsymbol{I},\boldsymbol{B}\rangle && \text{(35a)}\\ \text{s.t.}\ & \text{(29b), (29c), (29d), (29e), (29h), (29i)} && \text{(35b)}\\ & 0\geq\frac{\alpha\beta_i^2+2\beta_i\beta_j+\beta_j^2/\alpha}{\min\{1,z_i+z_j\}}-2B_{ij}-\alpha B_{ii}-B_{jj}/\alpha,\forall i\neq j,\alpha\in V_{ij}^+ && \text{(35c)}\\ & 0\geq\frac{\alpha\beta_i^2-2\beta_i\beta_j+\beta_j^2/\alpha}{\min\{1,z_i+z_j\}}+2B_{ij}-\alpha B_{ii}-B_{jj}/\alpha,\forall i\neq j,\alpha\in V_{ij}^-, && \text{(35d)}\end{aligned}$$

where V_{ij}^+ and V_{ij}^- are any finite subsets of $\mathbb{R}_+$. Relaxation (35) can be refined dynamically: given an optimal solution of (35), new values of α generated according to (31) (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 (35), 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 $\boldsymbol{\Lambda}\in\mathcal{S}_+^P$ be diagonally dominant matrix. Observe that for any $(\boldsymbol{z},\boldsymbol{\beta})\in\{0,1\}^p\times\mathbb{R}^p$ such that $\boldsymbol{\beta}\circ(\boldsymbol{e}-\boldsymbol{z})=\boldsymbol{0}$,

$$\begin{aligned}t\geq\boldsymbol{\beta}^\top\boldsymbol{\Lambda}\boldsymbol{\beta}\Leftrightarrow t&\geq\sum_{i=1}^p\left(\Lambda_{ii}-\sum_{j\neq i}|\Lambda_{ij}|\right)\beta_i^2+\sum_{i=1}^p\sum_{j=i+1}^p|\Lambda_{ij}|\left(\beta_i+\operatorname{sign}(\Lambda_{ij})\beta_j\right)^2\\ \Leftrightarrow t&\geq\sum_{i=1}^p\left(\Lambda_{ii}-\sum_{j\neq i}|\Lambda_{ij}|\right)\frac{\beta_i^2}{z_i}+\sum_{i=1}^p\sum_{j=i+1}^p|\Lambda_{ij}|\frac{\left(\beta_i+\operatorname{sign}(\Lambda_{ij})\beta_j\right)^2}{\min\{1,z_i+z_j\}},\end{aligned} \tag{36}$$

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 [3] for a similar strengthening for signal estimation based on nonnegative pairwise quadratic terms.

We now consider using decompositions of the form $\boldsymbol{\Lambda}+\boldsymbol{R}=\boldsymbol{X}^\top\boldsymbol{X}+\lambda\boldsymbol{I}$, where $\boldsymbol{\Lambda}$ is a diagonally dominant matrix and $\boldsymbol{R}\in\mathcal{S}_+^P$. Given such a decomposition, inequalities (36) can be used to strengthen the formulations. Specifically, we consider relaxations of (3) of the form

$$\boldsymbol{y}^\top \boldsymbol{y} + \min \; -2\boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}^\top \boldsymbol{u} + \hat{\phi}(\boldsymbol{z},\boldsymbol{\beta}) \tag{37a}$$

$$\text{(17b)},\ \text{(17c)},\ \text{(17d)}, \tag{37b}$$

where

$$\hat{\phi}(\boldsymbol{z},\boldsymbol{\beta}) := \max_{\boldsymbol{\Lambda},\boldsymbol{R}} \boldsymbol{\beta}^\top \boldsymbol{R}\boldsymbol{\beta} + \sum_{i=1}^p \Big(\Lambda_{ii} - \sum_{j\neq i} |\Lambda_{ij}|\Big)\frac{\beta_i^2}{z_i} + \sum_{i=1}^p \sum_{j=i+1}^p |\Lambda_{ij}| \frac{\left(\beta_i + \text{sign}(\Lambda_{ij})\beta_j\right)^2}{\min\{1, z_i + z_j\}} \tag{38a}$$

$$\text{s.t. } \boldsymbol{\Lambda} + \boldsymbol{R} = \boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I} \tag{38b}$$

$$\Lambda_{ii} \geq \sum_{j<i} |\Lambda_{ji}| + \sum_{j>i} |\Lambda_{ij}| \qquad \forall i \in P \tag{38c}$$

$$\boldsymbol{R} \in \mathcal{S}_+^P. \tag{38d}$$

Proposition 6. *Problem* (37) *is equivalent to*

$$
(\texttt{sdp}_{\texttt{dd}}) \qquad
\begin{aligned}
& \boldsymbol{y}^\top \boldsymbol{y} + \min \; -2\boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}^\top \boldsymbol{u} + \langle \boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}, \boldsymbol{B}\rangle && && \text{(39a)}\\
& \quad \textit{s.t. } \boldsymbol{e}^\top \boldsymbol{z} \leq k && && \text{(39b)}\\
& \quad \boldsymbol{\beta} \leq \boldsymbol{u},\ -\boldsymbol{\beta} \leq \boldsymbol{u} && && \text{(39c)}\\
& \quad z_i B_{ii} \geq \beta_i^2 && \forall i \in P && \text{(39d)}\\
& \quad 0 \leq w_{ij} \leq 1,\ w_{ij} \leq z_i + z_j && \forall i \neq j && \text{(39e)}\\
& \quad 0 \geq \frac{\beta_i^2 + 2\beta_i\beta_j + \beta_j^2}{w_{ij}} - 2B_{ij} - B_{ii} - B_{jj} && \forall i \neq j && \text{(39f)}\\
& \quad 0 \geq \frac{\beta_i^2 - 2\beta_i\beta_j + \beta_j^2}{w_{ij}} + 2B_{ij} - B_{ii} - B_{jj} && \forall i \neq j && \text{(39g)}\\
& \quad \boldsymbol{B} - \boldsymbol{\beta}\boldsymbol{\beta}' \in \mathcal{S}_+^P && && \text{(39h)}\\
& \quad \boldsymbol{\beta} \in \mathbb{R}^p,\ \boldsymbol{z} \in [0,1]^p,\ \boldsymbol{u} \in \mathbb{R}_+^p,\ \boldsymbol{B} \in \mathbb{R}^{p\times p}. && && \text{(39i)}
\end{aligned}
$$

Proof. Let $\boldsymbol{\Gamma}, \boldsymbol{\Gamma}^+, \boldsymbol{\Gamma}^-$ be nonnegative $p \times p$ matrices such that: $\Gamma_{ii} = \Lambda_{ii}$ and $\Gamma_{ij} = 0$ for $i \neq j$; $\Gamma_{ii}^+ = \Gamma_{ii}^- = 0$ and $\Gamma_{ij}^+ - \Gamma_{ij}^- = \Lambda_{ij}$ for $i \neq j$. Problem (38) can be written as

$$\hat{\phi}(\boldsymbol{z},\boldsymbol{\beta}) := \max_{\boldsymbol{\Gamma},\boldsymbol{\Gamma}^+,\boldsymbol{\Gamma}^- \boldsymbol{R}} \boldsymbol{\beta}^\top \boldsymbol{R}\boldsymbol{\beta} + \sum_{i=1}^p \Big(\Gamma_{ii} - \sum_{j\neq i}(\Gamma_{ij}^+ + \Gamma_{ij}^-)\Big)\frac{\beta_i^2}{z_i} \tag{40a}$$

$$+ \sum_{i=1}^p \sum_{j=i+1}^p \left(\Gamma_{ij}^+ \frac{(\beta_i+\beta_j)^2}{\min\{1, z_i+z_j\}} + \Gamma_{ij}^- \frac{(\beta_i-\beta_j)^2}{\min\{1, z_i+z_j\}}\right) \tag{40b}$$

$$\text{s.t. } \boldsymbol{\Gamma} + \boldsymbol{\Gamma}^+ + \boldsymbol{\Gamma}^- + \boldsymbol{R} = \boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I} \tag{40c}$$

$$\Gamma_{ii} \geq \sum_{j<i}(\Gamma_{ji}^+ + \Gamma_{ji}^-) + \sum_{j>i}(\Gamma_{ij}^+ + \Gamma_{ij}^-) \qquad \forall i \in P \tag{40d}$$

$$\boldsymbol{R} \in \mathcal{S}_+^P. \tag{40e}$$

Then, similarly to the proof of Theorem 3, it is easy to show that the dual of (40) is precisely (39). □

4.4. **Relaxing the $(p+1)$-dimensional PSD constraint.** We now discuss a relaxation of the p-dimensional semidefinite constraint $\boldsymbol{B}-\boldsymbol{\beta}\boldsymbol{\beta}^\top \in \mathcal{S}_+^P$, present in all formulations. Let $\boldsymbol{V}$ be a matrix whose j-th column $\boldsymbol{V_j}$ is an eigenvector of $\boldsymbol{X}^\top\boldsymbol{X}$. Consider the optimization problem

$$\underline{\phi_{\mathcal{P}}}(\boldsymbol{z},\boldsymbol{\beta}) := \max_{\boldsymbol{A_T},\boldsymbol{R},\boldsymbol{\pi}} \boldsymbol{\beta}^\top\boldsymbol{R}\boldsymbol{\beta} + \sum_{T\in\mathcal{P}} \frac{\boldsymbol{\beta_T}^\top\boldsymbol{A_T}\boldsymbol{\beta_T}}{\min\{1, z(T)\}} \tag{41a}$$
$$\text{s.t.} \sum_{T\in\mathcal{P}} \boldsymbol{A_T} + \boldsymbol{R} = \boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{I} \tag{41b}$$
$$\boldsymbol{A_T} \in \mathcal{S}_+^T \qquad \forall T\in\mathcal{P} \tag{41c}$$
$$\boldsymbol{R} = \boldsymbol{V}\operatorname{diag}(\boldsymbol{\pi})\boldsymbol{V}^\top \tag{41d}$$
$$\boldsymbol{\pi} \in \mathbb{R}_+^n. \tag{41e}$$

Observe that the objective and constraints (41a)–(41c) are identical to (15). However, instead of (15e), we have $\boldsymbol{R} = \sum_{j=1}^{\min\{p,n\}} \pi_j \boldsymbol{V_j}\boldsymbol{V_j}^\top$. Moreover, since $\boldsymbol{\pi}\geq\boldsymbol{0}$, $\boldsymbol{R}\in\mathcal{S}_+^P$ in any feasible solution of (41), thus (15) is a relaxation of (41), and, hence, $\underline{\phi_{\mathcal{P}}}$ is indeed a lower bound on $\phi_{\mathcal{P}}$. Finally, (41) is feasible if $\lambda=0$ or $\mathcal{P}$ contains all singletons, as it is possible to set $\boldsymbol{A_{\{i\}}}=\lambda$, $\boldsymbol{A_T}=0$ for $|T|>1$, and set $\boldsymbol{\pi}$ equal to the eigenvalues of $\boldsymbol{X}^\top\boldsymbol{X}$. Therefore, instead of (17), one may use the simpler convex relaxation

$$\boldsymbol{y}^\top\boldsymbol{y} + \min\ -2\boldsymbol{y}^\top\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}^\top\boldsymbol{u} + \underline{\phi_{\mathcal{P}}}(\boldsymbol{z},\boldsymbol{\beta}) \tag{42a}$$
$$\boldsymbol{e}^\top\boldsymbol{z} \leq k \tag{42b}$$
$$\boldsymbol{\beta}\leq\boldsymbol{u},\ -\boldsymbol{\beta}\leq\boldsymbol{u} \tag{42c}$$
$$\boldsymbol{\beta}\in\mathbb{R}^p,\ \boldsymbol{z}\in[0,1]^p,\ \boldsymbol{u}\in\mathbb{R}_+^p \tag{42d}$$

for (1).

Proposition 7. *If $\mathcal{P}=\{T\subseteq P : |T|\leq 2\}$, then problem (42) is equivalent to*

$$\boldsymbol{y}^\top\boldsymbol{y} + \min\ -2\boldsymbol{y}^\top\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}^\top\boldsymbol{u} + \langle \boldsymbol{X}^\top\boldsymbol{X}+\lambda\boldsymbol{I}, \boldsymbol{B}\rangle \tag{43a}$$
$$\textit{s.t. } \boldsymbol{e}^\top\boldsymbol{z}\leq k \tag{43b}$$
$$\boldsymbol{\beta}\leq\boldsymbol{u},\ -\boldsymbol{\beta}\leq\boldsymbol{u} \tag{43c}$$
$$\begin{pmatrix} z_i & \beta_i \\ \beta_i & B_{ii}\end{pmatrix} \succeq 0 \qquad \forall i=1,\dots,p \tag{43d}$$
$$(\boldsymbol{sdp_{LB}}) \qquad 0\leq w_{ij}\leq\min\{1, z_i+z_j\} \qquad \forall i<j \tag{43e}$$
$$\begin{pmatrix} w_{ij} & \beta_i & \beta_j \\ \beta_i & B_{ii} & B_{ij} \\ \beta_j & B_{ij} & B_{jj}\end{pmatrix} \succeq 0 \qquad \forall i<j \tag{43f}$$
$$\boldsymbol{V_j}^\top\left(\boldsymbol{B}-\boldsymbol{\beta}\boldsymbol{\beta}^\top\right)\boldsymbol{V_j}\geq 0 \qquad \forall j=1,\dots,\min\{n,p\} \tag{43g}$$
$$\boldsymbol{\beta}\in\mathbb{R}^p,\ \boldsymbol{z}\in[0,1]^p,\ \boldsymbol{u}\in\mathbb{R}_+^p,\ \boldsymbol{B}\in\mathbb{R}^{p\times p}. \tag{43h}$$

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

Observe that in formulation (43), the $(p+1)$-dimensional semidefinite constraint (19f) is replaced with $\min\{p,n\}$ rank-one quadratic constraints (43g). We denote by $\mathtt{sdp_{LB}}$ the relaxation of $\mathtt{sdp_2}$ obtained by replacing (21f) with (43g). In general, $\mathtt{sdp_{LB}}$ is still an SDP due to constraints (43f); however, note that $\mathtt{sdp_{LB}}$ can be implemented in a conic quadratic framework by using cuts, as described in §4.2. Moreover, constraints (43g) could also be dynamically refined to better approximate the SDP constraint, or formulation (43) could be improved with ongoing research on approximating SDP via mixed-integer conic quadratic optimization, e.g., see [36, 37].

Remark 2. We observe that formulation (43) is solved substantially faster than $\mathtt{sdp_2}$ (with Mosek) with constraints (43f) formulated as semi-definite constraints. Indeed, the $\mathcal{O}(p^2)$ low-dimensional constraints (22f) can actually be handled efficiently, but the major computational bottleneck towards solving $\mathtt{sdp_2}$ is handling the single large-dimensional positive semi-definite constraint (22g).

5. Computations

In this section, we report computational experiments with the proposed conic relaxations on synthetic as well as benchmark datasets. Semidefinite optimization problems are solved with MOSEK 8.1 solver, and conic quadratic optimization problems (continuous and mixed-integer) are solved with CPLEX 12.8 solver. All computations are performed 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 §5.2, we focus on the relaxation quality of $\mathtt{sdp_r}$ and its ability to approximate the exact ℓ_0-problem (1); then, in §5.3, we adopt the same experimental framework used in [7, 26] to generate synthetic instances and evaluate the proposed conic formulations from an inference perspective. In both cases, our results compare favorably with existing approaches in the literature. Finally, in §5.4, we summarize our findings and discuss possible extensions.

5.1. **Datasets.** We use the benchmark datasets in Table 1. The first five were first used in [45] in the context of MIO algorithms for best subset selection, and later used in [21]. 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`.

In addition, we also use synthetic datasets generated similarly to [7, 26]. Here we present a summary of the simulation setup and refer the readers to [26] for an extended description. . For given dimensions n, p, sparsity s, predictor autocorrelation ρ, and signal-to-noise ratio SNR, the instances are generated as follows:

(1) The (true) coefficients $\boldsymbol{\beta}_0$ have the first s components equal to one, and the rest equal to zero.
(2) The rows of the predictor matrix $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ are drawn from i.i.d. distributions $\mathcal{N}_p(\boldsymbol{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$ has entry (i, j) equal to $\rho^{|i-j|}$.
(3) The response vector $\boldsymbol{y} \in \mathbb{R}^n$ is drawn from $\mathcal{N}_p(\boldsymbol{X\beta_0}, \sigma^2\boldsymbol{I})$, where $\sigma^2 = \boldsymbol{\beta_0}^\top \boldsymbol{X\beta_0}/\mathrm{SNR}$.

Similar data generation has been used in the literature [7, 26].

[5] In our experiments, the datasets were standardized first.

5.2. **Relaxation quality.** In this section we test the ability of $\texttt{sdp}_2$, given in (22), and of $\texttt{sdp}_{\texttt{LB}}$, given in (43), to provide near-optimal solutions to problem (1), and compare its performance with MIO approaches. In §5.2.1, we focus on the pure best subset selection problem with $\lambda = 0$, which has received relatively little attention in the literature [7]; in §5.2.2 we consider problems with ℓ_0-ℓ_2 regularization, which has received more attention in the literature [8, 28, 30, 58]; in §5.2.3 we study the impact of model complexity parameter r on the relaxation quality, and in §5.2.4 we study the scalability of the proposed methods.

Computing optimality gaps for $\texttt{sdp}_\texttt{r}$. The optimal objective value ν_ℓ^* of $\texttt{sdp}_\texttt{r}$ provides a lower bound on the optimal objective value of (1). To obtain an upper bound, we use a simple greedy heuristic to retrieve a feasible solution for (1): given an optimal solution vector $\bar{\boldsymbol{\beta}}^*$ for $\texttt{sdp}_\texttt{r}$, let $\bar{\beta}^*_{(k)}$ denote the k-th largest absolute value. For $T = \left\{i \in P : |\bar{\beta}^*_i| \geq \bar{\beta}^*_{(k)}\right\}$, let $\hat{\boldsymbol{\beta}}_{\boldsymbol{T}}$ be the k-dimensional `ols/ridge` estimator using only predictors in T, i.e.,

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{T}} = (\boldsymbol{X}_{\boldsymbol{T}}^\top \boldsymbol{X}_{\boldsymbol{T}} + \lambda \boldsymbol{I}_{\boldsymbol{T}})^{-1} \boldsymbol{X}_{\boldsymbol{T}}^\top \boldsymbol{y},$$

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

$$\texttt{gap} = \frac{\nu_u - \nu_\ell^*}{\nu_\ell^*} \times 100. \tag{44}$$

While stronger relaxations result in improved lower bounds ν_ℓ^*, the corresponding heuristic upper bounds ν_u are not necessarily better; thus, the optimality gaps are not guaranteed to improve with stronger relaxations. Nevertheless, as shown next, stronger relaxations in general yield much smaller gaps in practice.

We point out that the main focus of the strong relaxations is to obtain improved lower bounds ν_ℓ^*. Randomized rounding methods [49, 58], more sophisticated rounding heuristics [13], or alternative heuristic methods [28] can be used to obtain improved upper bounds. Nevertheless, the quality of the upper bounds obtained from the greedy rounding method can be used to estimate how well the solutions from the relaxations match the sparsity pattern of the optimal solution.

5.2.1. $\lambda = 0$ *case.* For each dataset with $\lambda = \mu = 0$, we solve the conic relaxations of (1) $\texttt{sdp}_1$ and $\texttt{sdp}_2$ as well as $\texttt{sdp}_{\texttt{LB}}$ and the mixed-integer formulation `big-M` given by (3)–(4). In our experiments, we set $M = 3\|\boldsymbol{\beta}_{\text{ols}}\|_\infty$, where $\boldsymbol{\beta}_{\text{ols}}$ is the ordinary least square estimator[6] and set a time limit of 10 minutes. For data with $p \leq 40$ we solve problems with cardinalities $k \in \{3, \ldots, 10\}$, and for `diabetes` and `crime` we solve problems with $k \in \{3, \ldots, 30\}$. Table 2 shows, for each dataset and method, the average lower bound (`LB`) and upper bound (`UB`) found by each method, the `gap` (44), and the `time` required to solve the problems (in seconds) – the average is taken across all k values. In all cases, lower and upper bounds are scaled so that the best upper bound for any given instance has value $\nu_u^* = 100$.

The `big-M` method is highly inconsistent and prone to numerical difficulties, due to the use of big-M constraints. First, for three datasets (`servo`, `auto MPG` and

[6] Bertsimas et al. [7] set $M = 2\|\hat{\boldsymbol{\beta}}\|_\infty$ for some heuristic solution $\hat{\boldsymbol{\beta}}$

TABLE 2. Results with $\lambda = 0$ on real instances. Lower and upper bounds are scaled so that the best upper bound for a given instance has value 100. Mean $\pm$ stdev are reported.

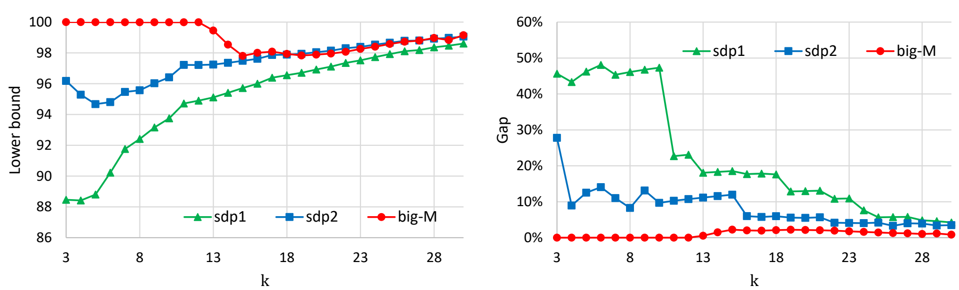
dataset	method	LB	UB	gap(%)	time
housing	$\mathtt{sdp}_1$	99.4±0.6	100.1±0.1	0.7±0.0	0.03±0.02
	$\mathtt{sdp}_2$	99.6±0.6	100.1±0.1	0.5±0.6	0.07±0.03
	$\mathtt{sdp}_{\mathtt{LB}}$	98.8±0.6	100.4±0.1	1.6±0.8	0.06±0.03
	big-M	100.0±0.0	100.0±0.0	0.0±0.0	0.01±0.01
servo	$\mathtt{sdp}_1$	86.8±5.5	109.5±10.3	27.3±20.6	0.02±0.01
	$\mathtt{sdp}_2$	94.9±2.9	106.2±16.5	12.2±19.8	0.10±0.01
	$\mathtt{sdp}_{\mathtt{LB}}$	89.5±2.7	109.2±15.5	21.8±14.9	0.17±0.03
	big-M†	†	†	†	†
auto MPG	$\mathtt{sdp}_1$	75.3±10.3	115.3±6.0	55.8±23.7	0.07±0.04
	$\mathtt{sdp}_2$	96.7±3.3	100.5±0.8	4.0±4.2	0.24±0.02
	$\mathtt{sdp}_{\mathtt{LB}}$	78.8±7.7	101.6±2.7	30.0±14.0	0.40±0.09
	big-M†	†	†	†	†
solar flare	$\mathtt{sdp}_1$	97.5±1.5	103.3±1.1	6.0±2.0	0.07±0.03
	$\mathtt{sdp}_2$	99.2±0.8	100.0±0.0	1.0±0.6	0.28±0.06
	$\mathtt{sdp}_{\mathtt{LB}}$	97.8±1.6	102.3±1.9	4.6±2.7	0.13±0.02
	big-M††	98.1±1.7	98.1±1.7	-	0.01±0.01
breast cancer	$\mathtt{sdp}_1$	88.9±3.1	101.5±1.7	14.4±5.6	0.15±0.02
	$\mathtt{sdp}_2$	98.0±0.6	100.4±0.8	2.4±1.1	0.77±0.07
	$\mathtt{sdp}_{\mathtt{LB}}$	94.8±0.5	100.5±0.7	6.0±0.5	0.40±0.03
	big-M†	†	†	†	†
diabetes	$\mathtt{sdp}_1$	95.2±3.2	115.2±11.8	22.2±16.3	3.58±0.77
	$\mathtt{sdp}_2$	97.4±1.3	105.4±4.2	8.2±5.2	9.28±1.12
	$\mathtt{sdp}_{\mathtt{LB}}$†	†	†	†	†
	big-M	99.0±0.9	100.0±0.0	1.0±0.9	416.17±260.57
crime	$\mathtt{sdp}_1$	97.8±1.3	103.2±2.4	5.6±3.6	17.82±0.98
	$\mathtt{sdp}_2$	99.0±0.8	101.6±2.0	2.7±2.7	45.29±4.06
	$\mathtt{sdp}_{\mathtt{LB}}$	94.6±2.0	109.7±2.8	16.0±4.9	5.87±0.43
	big-M	96.4±1.7	100.0±0.0	3.7±1.8	527.03±185.64

† Error in solving problem.
†† Infeasible solution is reported as optimal.

`breast cancer`) the method fails due to numerical issues ("failure to solve MIP subproblem"). In addition, for `solar flare` the solver reports very fast solution times but the solutions are in fact infeasible for problem (1): by default in CPLEX, if $z_i \leq 10^{-5}$ in a solution then z_i is deemed to satisfy the integrality constraint $z_i \in \{0, 1\}$. Thus, if the big-M constant is large enough, then constraint (4) may in fact allow nonzero values for β_i even when "$z_i = 0$". In particular, in `solar flare` we found that the solution $\boldsymbol{\beta}_{\text{mio}}$ reported by the MIO solver satisfies[7] $\|\boldsymbol{\beta}_{\text{mio}}\|_0 = 20$, regardless of the value of k used, violating the sparsity constraint. We also point out that $\texttt{sdp}_{\texttt{LB}}$ struggles with numerical difficulties in `diabetes`: the problems are incorrectly found to be unbounded. In contrast, $\texttt{sdp}_{\texttt{r}}$ methods are solved without numerical difficulties.

In terms of the relaxation quality, we find that $\texttt{sdp}_2$ is the best as expected. It consistently delivers better lower and upper bounds compared to the other conic relaxations, and even outperforming `big-M` in terms of lower bounds and gaps in the largest dataset (`crime`). The strength of the relaxation comes at the expense 2–4-fold larger computation time than $\texttt{sdp}_1$, but on the other hand $\texttt{sdp}_2$ is substantially faster than `big-M` on large datasets. We see that neither $\texttt{sdp}_1$ nor $\texttt{sdp}_{\texttt{LB}}$ dominates each other in terms of relaxation quality. While $\texttt{sdp}_1$ is faster on the smaller datasets, $\texttt{sdp}_{\texttt{LB}}$ is faster on `crime`, indicating that $\texttt{sdp}_{\texttt{LB}}$ may scale better (we corroborate this statement in §5.2.4). Finally `big-M`, in datasets where numerical issues do not occur, is able to find high quality solutions consistently, but struggles to find matching lower bound in larger instances, despite significantly higher computation time spent.

Figures 4 and 5 present detailed results on lower bounds and gaps as a function of the sparsity parameter k for the `diabetes` and `crime` datasets. For small values of k, `big-M` is arguably the best method, solving the problems to optimality. However, as k increases, the quality of the lower bounds and gaps deteriorate: for `diabetes`, $\texttt{sdp}_2$ finds better solutions than `big-M` for $k \geq 18$; for `crime`, $\texttt{sdp}_1$ and $\texttt{sdp}_2$ find better lower bounds for $k \geq 8$ (and, in the case of $\texttt{sdp}_2$, better gaps as well), and $\texttt{sdp}_{\texttt{LB}}$ matches the lower bound found by `big-M` for $k \geq 14$, despite requiring only five seconds (instead of 10 minutes) to find such lower bounds. Observe that the number of possible supports $\binom{p}{k} = \mathcal{O}(p^k)$ for problem (1) scales exponentially with k, thus enumerative methods such as branch-and-bound may struggle as k grows.



(A) Lower bounds

(B) Gaps

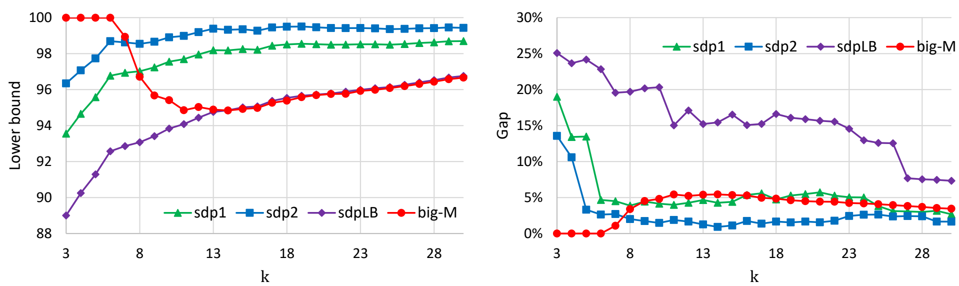
FIGURE 4. Detailed results on the `diabetes` dataset with $\lambda = 0$.

5.2.2. $\lambda > 0$ *case.* For each dataset with[8] $\lambda = 0.05$ and $\mu = 0$, we solve the conic relaxations of (1) $\texttt{sdp}_1$, $\texttt{sdp}_2$ and $\texttt{sdp}_{\texttt{LB}}$ and the "big-M free" mixed-integer formulation (5) with a time limit of 10 minutes (`persp`). This MIO formulation is

---

[7]We consider $\beta_i \neq 0$ whenever $\|\beta_i\| > 10^{-4}$.

[8]Since data is standardized so that each column has unit norm, a value of $\lambda = 0.05$ corresponds to an increase of 5% in the diagonal elements of the matrix $\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}$.



(A) Lower bounds (B) Gaps

FIGURE 5. Detailed results on the `crime` dataset with $\lambda = 0$.

possible since $\lambda > 0$, and has been shown to be competitive [58, 30] with the tailored algorithm proposed in [8]. For datasets with $p \leq 40$ we solve the problems with cardinalities $k \in \{3, \ldots, 10\}$, and for `diabetes` and `crime` we solve the problems with $k \in \{3, \ldots, 30\}$. Table 3 shows, for each dataset and method, the average lower bound (`LB`) and upper bound (`UB`) found by each method, the `gap` (44), and the `time` required to solve the problems (in seconds) – the average is taken across all k values. In all cases, lower and upper bounds are scaled so that the best upper bound for any given instance has value $\nu_u^* = 100$.

We observe that instances with $\lambda = 0.05$ are much easier to solve than those with $\lambda = 0$: no numerical issues occur for $\texttt{sdp}_{\texttt{LB}}$ or `persp`, and lower and upper bounds are much better for all methods. The mixed integer formulation `persp` comfortably solves the small instances with $p \leq 40$ to optimality, but $\texttt{sdp}_2$ yields better lower bounds and gaps for the larger instances `diabetes` and `crime` in a fraction of the time used by `persp`.

Figures 6 and 7 present lower bounds and gaps as a function of the regularization parameter λ, for `diabetes` and `crime` datasets (with $k = 15$). We observe that for low value of λ, `persp` struggles to find good lower bounds, e.g., it is outperformed by all conic relaxations in `crime` for $\lambda \leq 0.02$, and is worse than $\texttt{sdp}_2$ for $\lambda \leq 0.1$ in terms of lower bounds and gaps in both datasets. As λ increases, all methods deliver better bounds, and `persp` is eventually able to solve all problems to optimality.

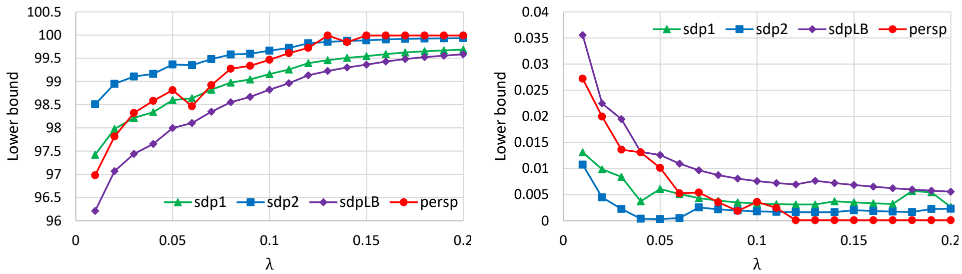
As expected, the performance of `persp` improves as λ increases. The perspective relaxation discussed in §1 exploits the separable terms introduced by the ℓ_2-regularization: as λ increases, this separable terms have a larger weight in the objective, and the strength of the relaxation improves as a consequence. Note that the conic relaxations also improve with larger λ: they are based on decompositions of the matrix $\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}$ into one- and two-variable terms, and the addition of the separable terms allows for a much richer set of decompositions. For large values of λ, $\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}$ becomes highly diagonal dominant, and the perspective relaxation alone provides a substantial strengthening. In this case, the advanced conic relaxations have a marginal impact and MIO methods with perspective strengthening performs better overall. In contrast, for low values of λ, the conic relaxations result in substantial strengthening over the perspective relaxation, and $\texttt{sdp}_{\texttt{r}}$ outperforms `persp` as a consequence.

TABLE 3. Results with $\lambda = 0.05$ on real instances. Lower and upper bounds are scaled so that the best upper bound for a given instance has value 100. Mean $\pm$ stdev are reported.

dataset	method	LB	UB	gap(%)	time
housing	$\texttt{sdp}_1$	99.7±0.4	100.2±0.3	0.5±0.6	0.03±0.02
	$\texttt{sdp}_2$	99.8±0.3	100.1±0.2	0.3±0.5	0.06±0.02
	$\texttt{sdp}_{\texttt{LB}}$	99.5±0.4	100.3±0.3	0.8±0.6	0.06±0.02
	persp	100.0±0.0	100.0±0.0	0.0±0.0	0.11±0.03
servo	$\texttt{sdp}_1$	95.9±3.0	102.2±6.7	6.7±4.1	0.03±0.01
	$\texttt{sdp}_2$	99.5±0.5	100.6±1.1	1.1±1.6	0.11±0.01
	$\texttt{sdp}_{\texttt{LB}}$	97.6±1.4	102.0±2.1	4.6±3.3	0.16±0.02
	persp	100.0±0.0	100.0±0.0	0.0±0.0	0.28±0.13
auto MPG	$\texttt{sdp}_1$	89.1±6.1	101.4±1.2	14.4±8.5	0.05±0.01
	$\texttt{sdp}_2$	99.8±0.2	100.0±0.1	0.2±0.3	0.25±0.04
	$\texttt{sdp}_{\texttt{LB}}$	92.7±3.1	101.1±1.5	9.2±4.0	0.35±0.02
	persp	100.0±0.0	100.0±0.0	0.0±0.0	1.29±0.60
solar flare	$\texttt{sdp}_1$	99.3±0.5	100.1±0.1	0.8±0.5	0.07±0.01
	$\texttt{sdp}_2$	99.9±0.1	100.1±0.1	0.2±0.1	0.28±0.03
	$\texttt{sdp}_{\texttt{LB}}$	99.2±0.7	100.4±1.2	1.2±1.0	0.16±0.03
	persp	100.0±0.0	100.0±0.0	0.0±0.0	1.75±1.07
breast cancer	$\texttt{sdp}_1$	94.9±1.8	100.8±0.4	6.3±2.4	0.18±0.04
	$\texttt{sdp}_2$	99.6±0.2	100.1±0.2	0.5±0.3	0.72±0.06
	$\texttt{sdp}_{\texttt{LB}}$	97.5±0.6	100.5±0.4	2.9±0.9	0.36±0.05
	persp	100.0±0.0	100.0±0.0	0.0±0.0	56.12±44.34
diabetes	$\texttt{sdp}_1$	98.9±0.6	100.2±0.2	1.2±0.7	2.13±0.24
	$\texttt{sdp}_2$	99.6±0.2	100.1±0.1	0.5±0.3	5.83±0.79
	$\texttt{sdp}_{\texttt{LB}}$	98.2±1.3	100.3±0.3	2.2±1.4	1.48±0.18
	persp	99.4±0.5	100.0±0.0	0.6±0.5	441.90±258.29
crime	$\texttt{sdp}_1$	99.3±0.9	100.3±0.9	1.1±1.7	19.15±1.30
	$\texttt{sdp}_2$	99.7±0.4	100.2±0.8	0.5±1.0	43.86±2.38
	$\texttt{sdp}_{\texttt{LB}}$	98.7±1.0	100.7±1.3	2.0±2.3	5.30±0.35
	persp	99.5±0.4	100.1±0.1	0.6±0.4	518.03±175.65

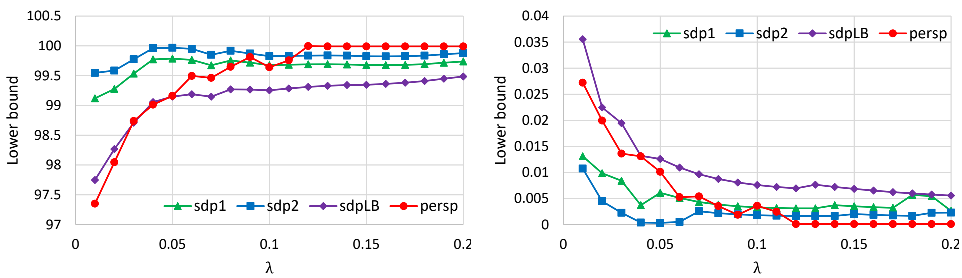
5.2.3. *The effect of model complexity r.* In §5.2.1–5.2.2 we reported computations with $\texttt{sdp}_\texttt{r}$ with $r \leq 2$. In experiments with those datasets, $\texttt{sdp}_3$ yields almost the same strengthening as $\texttt{sdp}_2$, but with much larger computational cost. Since $\texttt{sdp}_2$ already achieves gaps close to 0 in those instances, there is little room for improvement with higher values of r.

If the matrix $\boldsymbol{X}^\top \boldsymbol{X}+\lambda \boldsymbol{I}$ has high diagonal dominance, which happens if $n > p$ or if λ is large, then there are many ways to decompose it into low-dimensional rank-one terms, and $\texttt{sdp}_\texttt{r}$ with r small achieves good relaxations. In contrast, if the



(A) Lower bounds (B) Gaps

FIGURE 6. Detailed results on the `diabetes` dataset with $k = 15$.



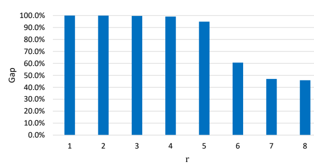
(A) Lower bounds (B) Gaps

FIGURE 7. Detailed results on the `crime` dataset with $k = 15$.

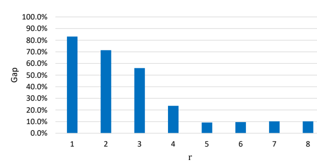
matrix $\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}$ has low diagonal dominance, it may be difficult to extract low-dimensional rank-one terms. In the extreme case of a rank-one case matrix, while $\texttt{sdp}_\texttt{p}$ results in the convex description, $\texttt{sdp}_\texttt{r}$ with $r < p$ achieves no improvement. In this section we illustrate this phenomenon.

First, we test on small synthetic instances with $p = 15$ and $n = 10$, setting the true sparsity to $s = 5$, autocorrelation $\rho = 0.35$, signal-noise-ration SNR $\in \{1, 5\}$, sparsity $k \in \{3, 4, 5, 6, 7, 8\}$, and for each combination of parameters we generate five instances. We report in Figure 8 the gaps obtained by $\texttt{sdp}_\texttt{r}$ for different values of r and λ – averaging across instances and different values of SNR and k. In addition, Figure 9 depicts the distribution of computational times required to solve the problems. We observe that for $\lambda = 0$, $\texttt{sdp}_\texttt{r}$ with $r \leq 4$ results in no strengthening and gaps of 100%; $\texttt{sdp}_5$ results in a small improvement (note that $5 = p - n$), while $\texttt{sdp}_\texttt{r}$ with $r \geq 6$ results in larger improvements. These results suggest that, with $\lambda = 0$, stronger formulations require rank-one strengthening with at least $p - n$ variables. We also observe that, as λ increases, the gaps reported by all methods decrease substantially, and the incremental strengthening obtained from larger values of r decreases: for $\lambda \geq 0.05$ $\texttt{sdp}_4$ performs almost identical to $\texttt{sdp}_8$, and for $\lambda = 0.15$ $\texttt{sdp}_3$ is similar to $\texttt{sdp}_8$ and $\texttt{sdp}_2$ already results in low optimality gaps. The computational time required to solve $\texttt{sdp}_\texttt{r}$ scales exponentially with r since the number of constraints increases exponentially as well. We conclude that

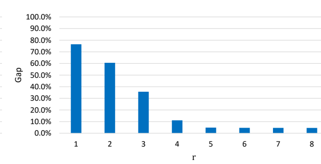
$\mathtt{sdp_2}$ is well suited for the $n > p$ case or for medium values of λ (for larger values $\mathtt{sdp_1}$ or even the simple perspective relaxation may be preferable), while $\mathtt{sdp_r}$ with $r \geq 3$ achieves a good improvement in relaxation quality for low values of λ, at the expense of larger computational times.



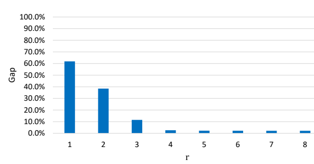
(A) $\lambda = 0$



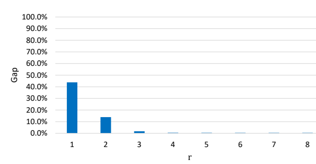
(B) $\lambda = 0.01$



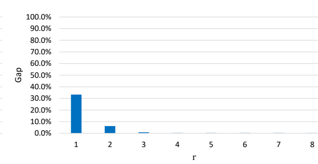
(C) $\lambda = 0.02$



(D) $\lambda = 0.05$



(E) $\lambda = 0.10$



(F) $\lambda = 0.15$

FIGURE 8. Optimality gaps of $\mathtt{sdp_r}$ with synthetic data, $1 \leq r \leq 8$.

Next, we test $\mathtt{sdp_r}$ on the `housing` dataset with $\lambda = 0$. Since this dataset has high diagonal dominance (see Table 1), even $\mathtt{sdp_1}$ results in small optimality gaps (see the first row of Table 2). Thus, instead of using all $n = 506$ datapoints, we randomly select only $n_0 < n$ datapoints, resulting in small diagonal dominance – for $n_0 = 10$, matrix $\boldsymbol{X}^\top \boldsymbol{X}$ is rank-deficient. Figure 10 shows the optimality gaps as a function of r and n_0. The relaxation quality improves with increasing n_0 and increasing r. For $n_0 = 20$, setting $r \geq 5$ produces optimality gaps close to 0% while $\mathtt{sdp_1}$ results in a gap of 13%, $\mathtt{sdp_2}$ a gap of 8%, and $\mathtt{sdp_3}$ yields a gap of 3%. For $n_0 = 50$, $\mathtt{sdp_2}$ results in a gap of 2%, and $\mathtt{sdp_3}$ yields a gap almost equal to 0%. Overall, the results are consistent with the experiments on synthetic data.

5.2.4. *On scalability.* As discussed in §5.2.3, formulation $\mathtt{sdp_r}$ for large values of r can be expensive to solve. Moreover, even $\mathtt{sdp_1}$ and $\mathtt{sdp_2}$ are semidefinite programs,

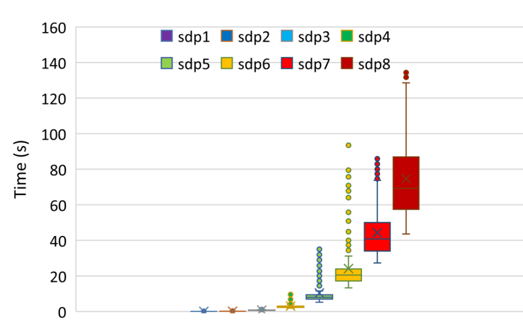
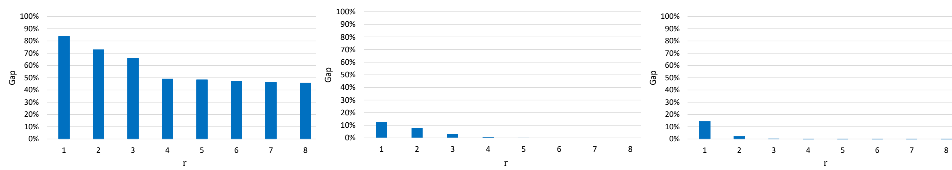


FIGURE 9. Time required to solve $\mathtt{sdp_r}$, $1 \leq r \leq 8$.



(A) $n_0 = 10$ (B) $n_0 = 20$ (C) $n_0 = 50$

FIGURE 10. Optimality gaps of $\texttt{sdp}_\texttt{r}$ with the altered `housing` dataset(using only n_0 datapoints), $1 \leq r \leq 8$.

which may not scale well for large values of p. In this section we present computations illustrating that while this is indeed the case, formulation $\texttt{sdp}_\text{LB}$ –which replaces the semidefinite constraint $\boldsymbol{B} - \boldsymbol{\beta\beta'} \in S_+^P$ with the quadratic constraints (43g)– scales much better and in fact can significantly outperform `persp` in terms of relaxation quality.

We generate synthetic instances with $p \in \{100, 150, \ldots, 500\}$, $n = 500$, true sparsity parameter $s = 30$, autocorrelation $\rho = 0.35$, signal-noise-ration SNR $\in \{1, 5\}$, sparsity $k = 30$; for each combination of parameters we generate five instances, and solve them for $\lambda \in \{0.01, 0.02, 0.05, 0.15\}$ and $\mu = 0$. Table 4 reports, for $\texttt{sdp}_1$, $\texttt{sdp}_2$, $\texttt{sdp}_\text{LB}$ and `persp` –using formulation (5) with a time limit of 600 seconds–, the time required to solve the problems and the optimality gap proven.

TABLE 4. Computational times and gaps on synthetic instances as a function of p. TL= Time Limit. †= Unable to solve (either due to very large computational times or memory issues). Numbers after "±" are the sample standard deviation.

p	$\texttt{sdp}_1$		$\texttt{sdp}_2$		$\texttt{sdp}_\text{LB}$		`persp`	
	`time(s)`	`gap(%)`	`time(s)`	`gap(%)`	`time(s)`	`gap(%)`	`time(s)`	`gap(%)`
100	19±3	0.5±0.6	44±9	0.1±0.1	5±1	1.0±1.1	TL	3.3±4.2
150	153±19	1.1±1.5	356±61	0.2±0.4	20±2	1.9±2.2	TL	6.2±7.2
200	673±64	2.7±3.0	1,691±165	0.6±0.9	42±3	4.3±4.0	TL	12.5±10.5
250	†	†	†	†	79±4	7.1±5.6	TL	17.2±13.2
300	†	†	†	†	147±7	12.2±7.6	TL	21.7±14.6
350	†	†	†	†	248±14	17.6±11.0	TL	25.9±17.0
400	†	†	†	†	391±36	24.0±14.9	TL	29.1±18.9
450	†	†	†	†	394±46	32.2±18.3	TL	34.3±21.0
500	†	†	†	†	462±43	39.3±21.9	TL	38.8±22.8

We observe that `persp` is unable to solve the problems within the 10 minute time limit and results in larger gaps than all other approaches, despite using substantially more time in most cases. We also observe that $\texttt{sdp}_\texttt{r}$ formulations struggle in instances with $p \geq 200$. Interestingly, $\texttt{sdp}_2$ requires consistently 2-4 times more than $\texttt{sdp}_1$ regardless of the dimension p. A similar factor was observed in Tables 2 and 3 with real data, suggesting that computational times with $\texttt{sdp}_2$ are within the same order-of-magnitude as $\texttt{sdp}_1$. Finally, $\texttt{sdp}_\text{LB}$ is substantially faster than both $\texttt{sdp}_1$ and $\texttt{sdp}_2$. While it results in larger gaps than $\texttt{sdp}_2$ as expected, since the high-dimensional constraint (22g) is relaxed, it still yields better optimality gaps than `persp`.

5.3. **Inference study on synthetic instances.** We now present inference results on synthetic data using the same simulation setup as in [7, 26], see [26] for an extended description. Specifically, we generate synthetic data as described in §5.1, and use the evaluation metrics used in [26], described next.

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

Relative risk:

$$\mathrm{RR}(\hat{\boldsymbol{\beta}}) = \frac{\mathbb{E}\left(\boldsymbol{x_0}^\top \hat{\boldsymbol{\beta}} - \boldsymbol{x_0}^\top \boldsymbol{\beta_0}\right)^2}{\mathbb{E}\left(\boldsymbol{x_0}^\top \boldsymbol{\beta_0}\right)^2}$$

with a perfect score 0 and null score of 1.

Relative test error:

$$\mathrm{RTE}(\hat{\boldsymbol{\beta}}) = \frac{\mathbb{E}\left(\boldsymbol{x_0}^\top \hat{\boldsymbol{\beta}} - y_0\right)^2}{\sigma^2}$$

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

Proportion of variance explained:

$$1 - \frac{\mathbb{E}\left(\boldsymbol{x_0}^\top \hat{\boldsymbol{\beta}} - y_0\right)^2}{\mathrm{Var}(y_0)}$$

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

Sparsity: We record the number of nonzeros[9], $\|\hat{\boldsymbol{\beta}}\|_0$, as done in [26]. Additionally, we also report the number of variables correctly identified, given by $\sum_{i=1}^p \mathbb{1}\{\hat{\beta}_i \neq 0 \text{ and } (\beta_0)_i \neq 0\}$.

5.3.2. *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 $\hat{\boldsymbol{\beta}}$.

elastic net: We solve the elastic net procedure using the parametrization

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\left(\alpha\|\boldsymbol{\beta}\|_1 + (1-\alpha)\|\boldsymbol{\beta}\|_2^2\right)$$

where $\alpha, \lambda \geq 0$ are the regularization parameters. We let $\alpha = 0.1\ell$ for integer $0 \leq \ell \leq 10$, we generated 50 values of λ ranging from $\lambda_{max} = \|\boldsymbol{X}^\top \boldsymbol{y}\|_\infty$ to $\lambda_{max}/200$ on a log scale, and using the pair (λ, μ) that results in the best prediction error on the validation set. A total of 500 (α, λ) pairs are tested.

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

[9]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.

The `elastic net` procedure approximately corresponds to the `lasso` procedure with 100 tuning parameters used in [26]. Similarly, $\texttt{sdp}_2$ with cross-validation approximately corresponds to the `best subset` procedure with 51 tuning parameters[10] used in [26]; nonetheless, the estimators from [26] are obtained by running a MIO solver for 3 minutes, while ours are obtained from solving to optimality a strong convex relaxation.

5.3.3. *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 5 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*[11] for (1). For a comparison, Hastie et al. [26] 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. [26] 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 5 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 5. Optimality gap and computation time (in seconds) of $\texttt{sdp}_2$ with $n = 500$, $p = 100$, $s = k = 5$, $\lambda = \mu = 0$.

SNR		0.05	0.09	0.14	0.25	0.42	0.71	1.22	2.07	3.52	6.00	**avg**
$\rho = 0.00$	`gap`	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**0.0**
	`time`	45.2	38.8	38.6	29.5	29.3	28.4	27.4	26.3	26.4	25.9	**31.6**
$\rho = 0.35$	`gap`	0.3	0.2	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0	**0.1**
	`time`	48.0	47.6	49.4	44.1	39.3	30.7	29.0	29.1	27.3	28.0	**37.3**

5.3.4. *Results: accuracy metrics.* Figure 11 plots the relative risk, relative test error, proportion of variance explained and sparsity results as a function of the SNR for instances with $n = 500$, $p = 100$, $s = 5$ and $\rho = 0$. Figure 12 plots the same results for instances with $\rho = 0.35$. The setting with $\rho = 0.35$ was also presented in [26].

We see that `elastic net` outperforms $\texttt{sdp}_2$ in low SNR settings, i.e., in SNR= 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 $\boldsymbol{\beta_0}$ for sufficiently large SNR, while `elastic net` is unable to do so. We also see that $\texttt{sdp}_2$ performs comparatively better than `elastic net` in instances with $\rho = 0$. Indeed, for large autocorrelations ρ, features where $(\beta_0)_i = 0$ still have

[10] Hastie et al. [26] use values of $k = 0, \dots, 50$. 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.

[11] A solution is deemed optimal if $\texttt{gap}< 10^{-4}$, which is the default parameter in MIO solvers.

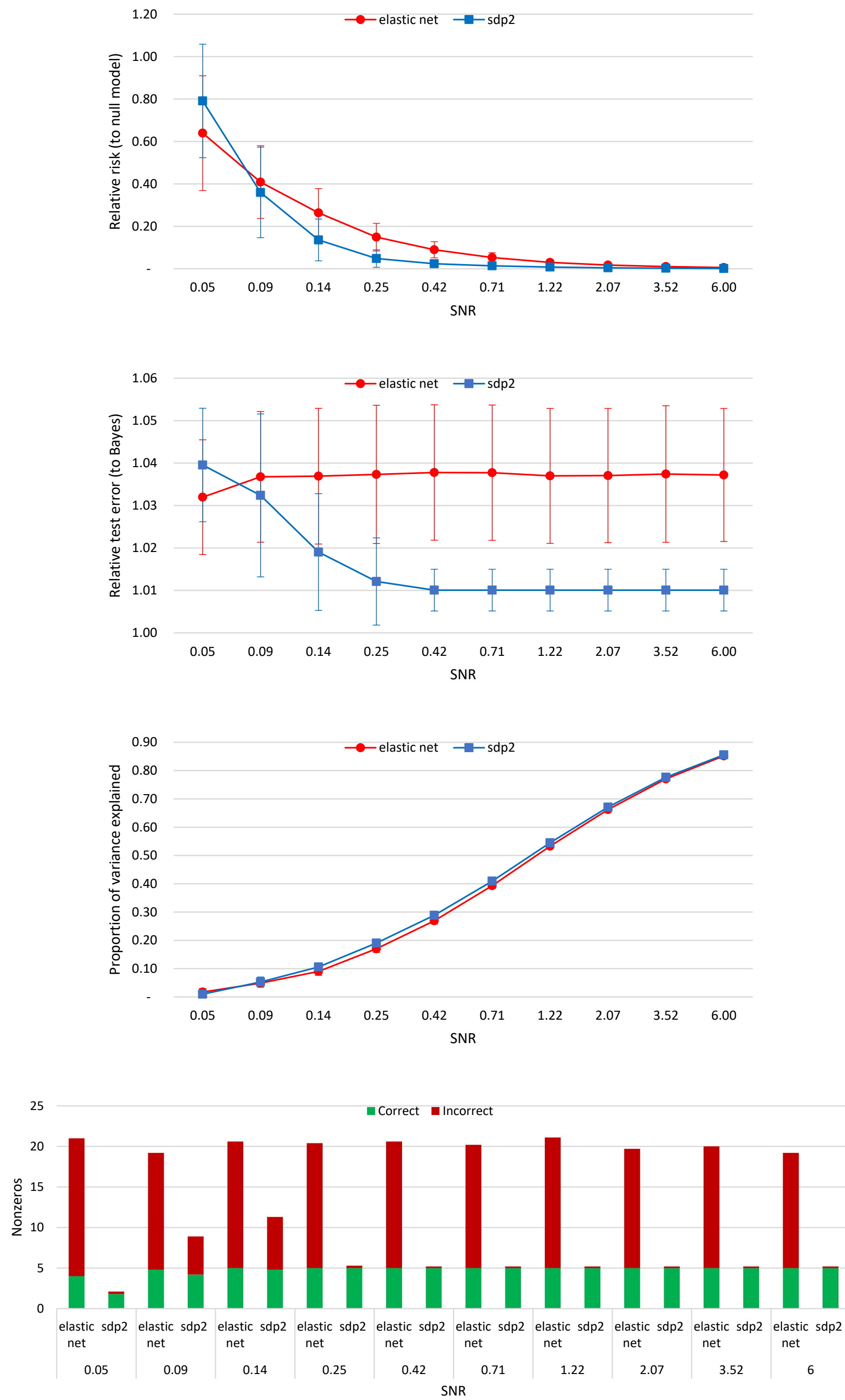


FIGURE 11. Relative risk, relative test error, proportion of variance explained and sparsity as a function of SNR, with $n = 500$, $p = 100$, $s = 5$ and $\rho = 0.00$.

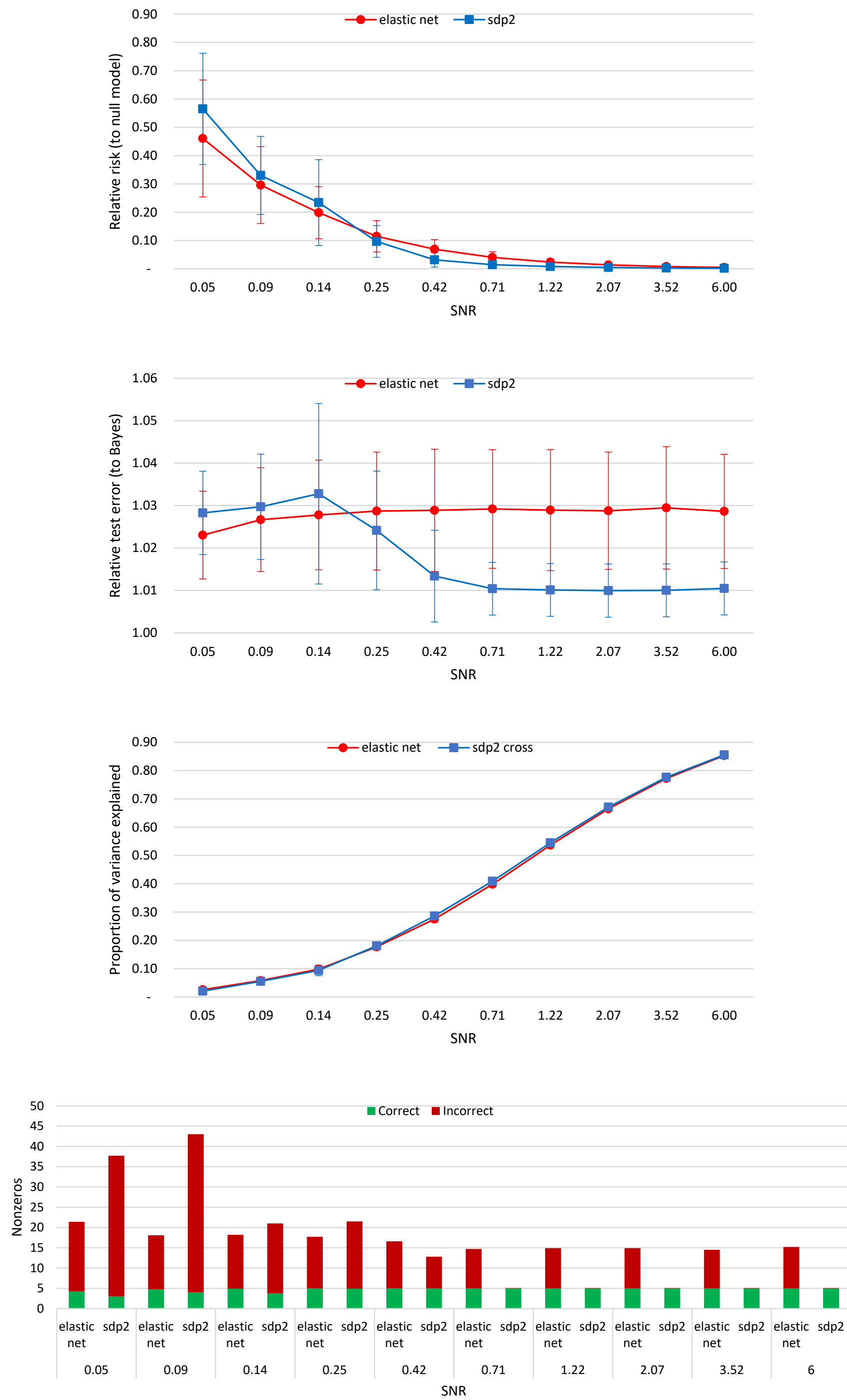


FIGURE 12. Relative risk, relative test error, proportion of variance explained and sparsity as a function of SNR, with $n = 500$, $p = 100$, $s = 5$ and $\rho = 0.35$.

predictive value, thus the dense estimator obtained by `elastic net` retains a relatively good predictive performance (however, such dense solutions are undesirable from an *interpretability* perspective). In contrast, when $\rho = 0$, such features are simply noise and `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 $\texttt{sdp}_2$ selects model corresponding to sparsities $k < s$ in low SNRs, while it consistently selects models with $k \approx s$ in high SNRs. We point out that, as suggested in [43], the results for low SNR could potentially be improved by fitting models with $\mu > 0$.

5.4. **Summary and extensions.** The conic formulations $\texttt{sdp}_2$ and $\texttt{sdp}_{\texttt{LB}}$ are able to deliver near-optimal solutions to problem (1) with p in the hundreds, and results in substantially better performance than MIO methods (computational times and numerical stability) in instances with low or no ℓ_2 regularization. Such performance makes the approaches directly applicable in several high-stakes domains where interpretability is a major consideration (e.g., $p \approx 50$ in the highly-publicized COMPAS recidivism case [57]; $p \approx 20$ in the setting considered in [11], where the output of the regression is used to optimize a carbon capture adsorber). Due to the computational challenges with solving conic optimization problems via second order methods, the solution approach presented in this paper (using an off-the-shelf solver) does not scale to larger instances. Nonetheless, as we now point out, the formulations presented here may serve as a basis for methods that scale to larger values of p, or tackle statistical problems other than sparse regression.

First, tailored implementations that do not rely on off-the-shelf solvers are possible. For example, Liu et al. [39] consider sparse inference problems with graphical models, which can be interpreted as a special case of (1) where matrix $\boldsymbol{X}^\top \boldsymbol{X}$ is sparse. They use $\texttt{sdp}_2$ as a base relaxation of their method and develop a tailored primal-dual method to solve it: the resulting approach scales comfortably to problems with p in the thousands.

Second, $\texttt{sdp}_\texttt{r}$ can be used as a subroutine of a more sophisticated method. For example, Hazimeh and Mazumder [28] propose a method that approximately solves problem (1), scales to problems with $p \approx 10^5$ and produces "combinatorially local" solutions. A key idea of their approach is to solve MIO problems involving only a small subset of the variables to escape local minima. In a similar vein, it would be possible to use relaxations $\texttt{sdp}_2$ with a small subset of the variables to identify descent directions.

Finally, we point out that both the quadratic loss function and sparsity are used pervasively in machine learning; thus the theory and methods developed in this paper may be generally applicable. For example, Bertsimas et al. [5] incorporate some the ideas discussed in §4 to develop algorithms that solve matrix completion problems to certifiable optimality.

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. 2112533 from the National Science Foundation and Grant No. 12951270 from the Office of Naval Research. A. Gómez is supported, in part, by Grants No. 1818700 and 2006762 from the National Science Foundation.

References

[1] Aktürk, M. S., Atamtürk, A., and Gürel, S. (2009). A strong conic quadratic reformulation for machine-job assignment with controllable processing times. *Operations Research Letters*, 37:187–191.

[2] Atamtürk, A. and Gómez, A. (2018). Strong formulations for quadratic optimization with M-matrices and indicator variables. *Mathematical Programming*, 170:141–176.

[3] Atamtürk, A., Gómez, A., and Han, S. (2021). Sparse and smooth signal estimation: Convexification of L0-formulations. *Journal of Machine Learning Research*, 22(52):1–43.

[4] Atamtürk, A. and Narayanan, V. (2007). Cuts for conic mixed integer programming. In Fischetti, M. and Williamson, D. P., editors, *Proceedings of the 12th International IPCO Conference*, pages 16–29.

[5] Bertsimas, D., Cory-Wright, R., and Pauphilet, J. (2021). Mixed-projection conic optimization: A new paradigm for modeling rank constraints. *Operations Research*.

[6] Bertsimas, D. and King, A. (2015). OR forum – an algorithmic approach to linear regression. *Operations Research*, 64:2–16.

[7] Bertsimas, D., King, A., Mazumder, R., et al. (2016). Best subset selection via a modern optimization lens. *The Annals of Statistics*, 44:813–852.

[8] Bertsimas, D. and Van Parys, B. (2017). Sparse high-dimensional regression: Exact scalable algorithms and phase transitions. *arXiv preprint arXiv:1709.10029*.

[9] Bixby, R. E. (2012). A brief history of linear and mixed-integer programming computation. *Documenta Mathematica*, pages 107–121.

[10] Chichignoud, M., Lederer, J., and Wainwright, M. J. (2016). A practical scheme and fast algorithm to tune the lasso with optimality guarantees. *The Journal of Machine Learning Research*, 17:8162–8181.

[11] Cozad, A., Sahinidis, N. V., and Miller, D. C. (2014). Learning surrogate models for simulation-based optimization. *AIChE Journal*, 60:2211–2227.

[12] Dheeru, D. and Karra Taniskidou, E. (2017). UCI machine learning repository.

[13] Dong, H., Chen, K., and Linderoth, J. (2015). Regularization vs. relaxation: A conic optimization perspective of statistical variable selection. *arXiv preprint arXiv:1510.06083*.

[14] Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al. (2004). Least angle regression. *The Annals of Statistics*, 32:407–499.

[15] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:1348–1360.

[16] Frangioni, A. and Gentile, C. (2006). Perspective cuts for a class of convex 0–1 mixed integer programs. *Mathematical Programming*, 106:225–236.

[17] Frangioni, A., Gentile, C., and Hungerford, J. (2020). Decompositions of semidefinite matrices and the perspective reformulation of nonseparable quadratic programs. *Mathematics of Operations Research*, 45(1):15–33.

[18] Frank, L. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35:109–135.

[19] Goemans, M. X. and Williamson, D. P. (1995). Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM (JACM)*, 42(6):1115–1145.

[20] Gómez, A. (2021). Strong formulations for conic quadratic optimization with indicator variables. *Mathematical Programming*, 188(1):193–226.

[21] Gómez, A. and Prokopyev, O. A. (2021). A mixed-integer fractional optimization approach to best subset selection. *INFORMS Journal on Computing*, 33(2):551–565.

[22] Günlük, O. and Linderoth, J. (2010). Perspective reformulations of mixed integer nonlinear programs with indicator variables. *Mathematical Programming*, 124:183–205.

[23] Han, S., Gómez, A., and Atamtürk, A. (2020). 2x2 convexifications for convex quadratic optimization with indicator variables. *arXiv preprint arXiv:2004.07448*.

[24] Han, S., Gómez, A., and Atamtürk, A. (2022). The equivalence of optimal perspective formulation and Shor's SDP for quadratic programs with indicator variables. *Operations Research Letters*, 50(2):195–198.

[25] Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The elements of statistical learning: Data mining, inference, and prediction*, volume 1. Springer series in statistics New York, NY, USA.

[26] Hastie, T., Tibshirani, R., and Tibshirani, R. J. (2017). Extended comparisons of best subset selection, forward stepwise selection, and the lasso. *arXiv preprint arXiv:1707.08692*.

[27] Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC press.

[28] Hazimeh, H. and Mazumder, R. (2020a). Fast best subset selection: Coordinate descent and local combinatorial optimization algorithms. *Operations Research*, 68(5):1517–1537.

[29] Hazimeh, H. and Mazumder, R. (2020b). Learning hierarchical interactions at scale: A convex optimization approach. In *International Conference on Artificial Intelligence and Statistics*, pages 1833–1843. PMLR.

[30] Hazimeh, H., Mazumder, R., and Saab, A. (2020). Sparse regression at scale: Branch-and-bound rooted in first-order optimization. *arXiv preprint arXiv:2004.06152*.

[31] Hebiri, M., Van De Geer, S., et al. (2011). The smooth-lasso and other $\ell_1+\ell_2$-penalized methods. *Electronic Journal of Statistics*, 5:1184–1226.

[32] Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67.

[33] Huang, J., Jiao, Y., Liu, Y., and Lu, X. (2018). A constructive approach to L0 penalized regression. *The Journal of Machine Learning Research*, 19:403–439.

[34] Hunter, D. R. and Li, R. (2005). Variable selection using MM algorithms. *Annals of Statistics*, 33:1617.

[35] Jeon, H., Linderoth, J., and Miller, A. (2017). Quadratic cone cutting surfaces for quadratic programs with on–off constraints. *Discrete Optimization*, 24:32–50.

[36] Kocuk, B., Dey, S. S., and Sun, X. A. (2016). Strong socp relaxations for the optimal power flow problem. *Operations Research*, 64:1177–1196.

[37] Kocuk, B., Dey, S. S., and Sun, X. A. (2018). Matrix minor reformulation and SOCP-based spatial branch-and-cut method for the AC optimal power flow problem. *Mathematical Programming Computation*, 10:557–596.

[38] Lin, X., Pham, M., and Ruszczyński, A. (2014). Alternating linearization for structured regularization problems. *The Journal of Machine Learning Research*, 15:3447–3481.

[39] Liu, P., Fattahi, S., Gómez, A., and Küçükyavuz, S. (2021). A graph-based decomposition method for convex quadratic optimization with indicators. *arXiv preprint arXiv:2110.12547 (Forthcoming in Mathematical Programming)*.

[40] Lombardi, M., Milano, M., and Bartolini, A. (2017). Empirical decision model learning. *Artificial Intelligence*, 244:343–367.

[41] Maragno, D., Wiberg, H., Bertsimas, D., Birbil, S. I., Hertog, D. d., and Fajemisin, A. (2021). Mixed-integer optimization with constraint learning. *arXiv preprint arXiv:2111.04469*.

[42] Mazumder, R., Friedman, J. H., and Hastie, T. (2011). Sparsenet: Coordinate descent with nonconvex penalties. *Journal of the American Statistical Association*, 106:1125–1138.

[43] Mazumder, R., Radchenko, P., and Dedieu, A. (2022). Subset selection with shrinkage: Sparse linear modeling when the SNR is low. *Operations Research*.

[44] Miller, A. (2002). *Subset Selection in Regression*. CRC Press.

[45] Miyashiro, R. and Takano, Y. (2015). Mixed integer second-order cone programming formulations for variable selection in linear regression. *European Journal of Operational Research*, 247:721–731.

[46] Natarajan, B. K. (1995). Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24:227–234.

[47] Nevo, D. and Ritov, Y. (2017). Identifying a minimal class of models for high-dimensional data. *The Journal of Machine Learning Research*, 18:797–825.

[48] Padilla, O. H. M., Sharpnack, J., Scott, J. G., and Tibshirani, R. J. (2017). The dfs fused lasso: Linear-time denoising over general graphs. *The Journal of Machine Learning Research*, 18:176–1.

[49] Pilanci, P., Wainwright, M. J., and El Ghaoui, L. (2015). Sparse learning via boolean relaxations. *Mathematical Programming*, 151:63–87.

[50] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.

[51] Shor, N. Z. (1987). Quadratic optimization problems. *Soviet Journal of Computer and Systems Sciences*, 25:1–11.

[52] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

[53] Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused Lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67:91–108.

[54] Tibshirani, R. J. and Taylor, J. (2011). The solution path of the generalized Lasso. *The Annals of Statistics*, 39(3):1335–1371.

[55] Wei, L., Atamtürk, A., Gómez, A., and Küçükyavuz, S. (2022a). On the convex hull of convex quadratic optimization problems with indicators. *arXiv preprint arXiv:2201.00387*.

[56] Wei, L., Gómez, A., and Küçükyavuz, S. (2022b). Ideal formulations for constrained convex optimization problems with indicator variables. *Mathematical Programming*, 192(1):57–88.

[57] Wexler, R. (2017). When a computer program keeps you in jail. *New York Times*.

[58] Xie, W. and Deng, X. (2020). Scalable algorithms for the sparse ridge regression. *SIAM Journal on Optimization*, 30(4):3359–3386.

[59] Zhang, C.-H. et al. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38:894–942.

[60] Zhang, C.-H., Huang, J., et al. (2008). The sparsity and bias of the Lasso selection in high-dimensional linear regression. *The Annals of Statistics*, 36:1567–1594.

[61] Zhang, C.-H., Zhang, T., et al. (2012). A general theory of concave regularization for high-dimensional sparse estimation problems. *Statistical Science*, 27:576–593.

[62] Zhang, Y., Wainwright, M. J., and Jordan, M. I. (2014). Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. In *Conference on Learning Theory*, pages 921–948.

[63] Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *The Journal of Machine Learning Research*, 7(Nov):2541–2563.

[64] Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429.

[65] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67:301–320.

[66] Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics*, 36:1509–1533.