Subset Selection with Shrinkage: Sparse Linear Modeling When the SNR Is Low

Subset Selection with Shrinkage: Sparse Linear Modeling when the SNR is low

Rahul Mazumder*[1], Peter Radchenko [2], and Antoine Dedieu†[1]

[1]Massachusetts Institute of Technology

[2]University of Sydney

December, 2021 ‡

Abstract

We study a seemingly unexpected and relatively less understood overfitting aspect of a fundamental tool in sparse linear modeling – best subset selection, which minimizes the residual sum of squares subject to a constraint on the number of nonzero coefficients. While the best subset selection procedure is often perceived as the "gold standard" in sparse learning when the signal to noise ratio (SNR) is high, its predictive performance deteriorates when the SNR is low. In particular, it is outperformed by continuous shrinkage methods, such as ridge regression and the Lasso. We investigate the behavior of best subset selection in the high-noise regimes and propose an alternative approach based on a regularized version of the least-squares criterion. Our proposed estimators (a) mitigate, to a large extent, the poor predictive performance of best subset selection in the high-noise regimes; and (b) perform favorably, while generally delivering substantially sparser models, relative to the best predictive models available via ridge regression and the Lasso. We conduct an extensive theoretical analysis of the predictive properties of the proposed approach and provide justification for its superior predictive performance relative to best subset selection when the noise-level is high. Our estimators can be expressed as solutions to mixed integer second order conic optimization problems and, hence, are amenable to modern computational tools from mathematical optimization.

*Rahul Mazumder's research was partially supported by the Office of Naval Research (N000141512342, N000141812298 – Young Investigator Award) and the National Science Foundation (NSF-IIS-1718258).

†Now at Vicarious AI; performed a major part of his work while a graduate student at MIT.

‡This is a major revision of an earlier manuscript dated August 2017.

1 Introduction

We consider the usual linear regression framework, with response $\mathbf{y} \in \mathbb{R}^n$, model matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ and regression coefficients $\boldsymbol{\beta} \in \mathbb{R}^p$. We assume that columns of $\mathbf{X}$ have been standardized to have zero means and unit ℓ_2-norms. In many classical and modern statistical applications it is desirable to obtain a parsimonious model with good data-fidelity. Towards this end, a natural candidate is the well-known *best-subsets* estimator [42], given by the following combinatorial optimization problem:

$$\hat{\boldsymbol{\beta}}_{\ell_0} \quad \in \quad \arg\min \quad \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \qquad \text{s.t.} \qquad \|\boldsymbol{\beta}\|_0 \leq k. \tag{1}$$

Problem (1) has a simple interpretation: it seeks to obtain the best least squares fit with at most k nonzero regression coefficients. There is a rich body of theoretical work studying the statistical properties of this estimator – see, for example, [21, 22, 51, 61] and the references therein. The caveat, however, is that Problem (1) is often perceived as computationally *infeasible* [45] – the popular `R`-package *leaps*, for example, is unable to obtain solutions to (1) when $p > 30$. Inability to compute the best-subsets estimator has perhaps contributed towards an aura of mystery around its operational characteristics on problem-instances that arise in practice. Recently, [9] demonstrated that Problem (1) can be solved to certifiable global optimality via mixed integer optimization (MIO) techniques [46, 8], leveraging the impressive advances in MIO over the past ten or so years – see [9, 40, 27] and the references therein. From a practical viewpoint, this line of research has made it possible to use subset selection procedures on real and synthetic datasets and gather insights regarding their operating characteristics, previously unseen due to the perceived computational limits. This paper investigates one such insight.

Does best subset selection overfit? Suppose that the data are generated from a linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}$, where matrix $\mathbf{X}$ is deterministic and the elements of $\boldsymbol{\epsilon} \in \mathbb{R}^n$ are independent $N(0, \sigma^2)$. We focus on the case where $\boldsymbol{\beta}^*$ is sparse, with few nonzero elements. It is well known that if the noise-level, measured by σ, is small relative to the signal-level, measured by $\|\mathbf{X}\boldsymbol{\beta}^*\|_2$, for example, then the best-subsets estimator leads to models with excellent statistical properties [51, 61, 14] in terms of prediction, estimation and variable selection (minor additional assumptions are required for the latter two metrics). However, the situation is different when the noise level is high – this was observed in [13], which highlighted the instability of best subset selection. Deterioration of the predictive performance in high-noise regimes is a significant drawback of best-subsets that,

to our knowledge, has received limited attention in the literature thus far. It is important to note that SNR alone does not control the difficulty of the underlying statistical problem; model parameters p, n, k^*, $\boldsymbol{\beta}^*$, and $\mathbf{X}$, also affect the performance of the estimator. In our theoretical analysis, presented in Section 3, we use ratios $\|\boldsymbol{\beta}^*\|_1/\sigma$ and $\|\boldsymbol{\beta}^*\|_2/\sigma$ to characterize the relevant noise-level regimes.

The best-subsets estimator given by Problem (1) focuses on two goals: (a) searching for the best subset, $\mathcal{I}$, containing k features; and (b) estimating $\hat{\boldsymbol{\beta}}_{\ell_0}$ by implementing the unconstrained least-squares method on the selected features $\mathcal{I}$. Even if best-subsets selects $\mathcal{I}$ to be the support of $\boldsymbol{\beta}^*$, the un-regularized fit on features $\mathcal{I}$ can be improved by shrinking the coefficients when σ is large. For a simple illustration of this, consider the setting where $n > p$ and $k = p$. Here, estimator $\hat{\boldsymbol{\beta}}_{\ell_0}$ is the usual least-squares solution, which benefits from additional shrinkage [30] to achieve a better bias-variance trade-off in the presence of noise. Further problems arise when the SNR is low due to the variability associated with the choice of $\mathcal{I}$. See for example, the works of [59, 16, 20] discussing the impossibility of variable selection when the signal is weak.

The discussion above suggests that the best-subsets estimator is not the right approach when the noise-level is high. Figure 1 presents a concrete example illustrating this point. The data are generated from a linear model with $n = 40$, $p = 60$, five true coefficients equal to one, and the rest equal to zero. The rows of $\mathbf{X}$ are drawn from a multivariate Gaussian distribution with the mean equal to zero and all the pairwise correlations equal to ρ. The features are standardized to have unit ℓ_2-norm, and σ^2 is set to match specific values of SNR $= \|\mathbf{X}\boldsymbol{\beta}^*\|_2^2/\|\boldsymbol{\epsilon}\|_2^2$. Figure 1 illustrates the performance of the best-subsets estimator, computed using the framework of [9] for different values of k; the results are averaged over ten different replications of $(\mathbf{X}, \boldsymbol{\epsilon})$. As expected, the predictive accuracy of best-subsets deteriorates as the SNR decreases – it is outperformed by continuous shrinkage methods such as ridge regression [28] and the Lasso [56]. The overfitting behavior of best-subsets can be attributed to its aggressive search for the best feature subset $\mathcal{I}$ and not performing any shrinkage on the selected coefficients.

We contend that the classical best-subsets estimator (1) is not designed to be used in high-noise regimes. Our theoretical and empirical investigations in Sections 3 and 5 highlight the shortcomings of best-subsets when contrasted with shrinkage methods. A natural question to ask at this point is: how might we *fix* this problem? Addressing this question with an associated methodological development is the main focus of this paper. We rule out the ambitious goal of correct variable

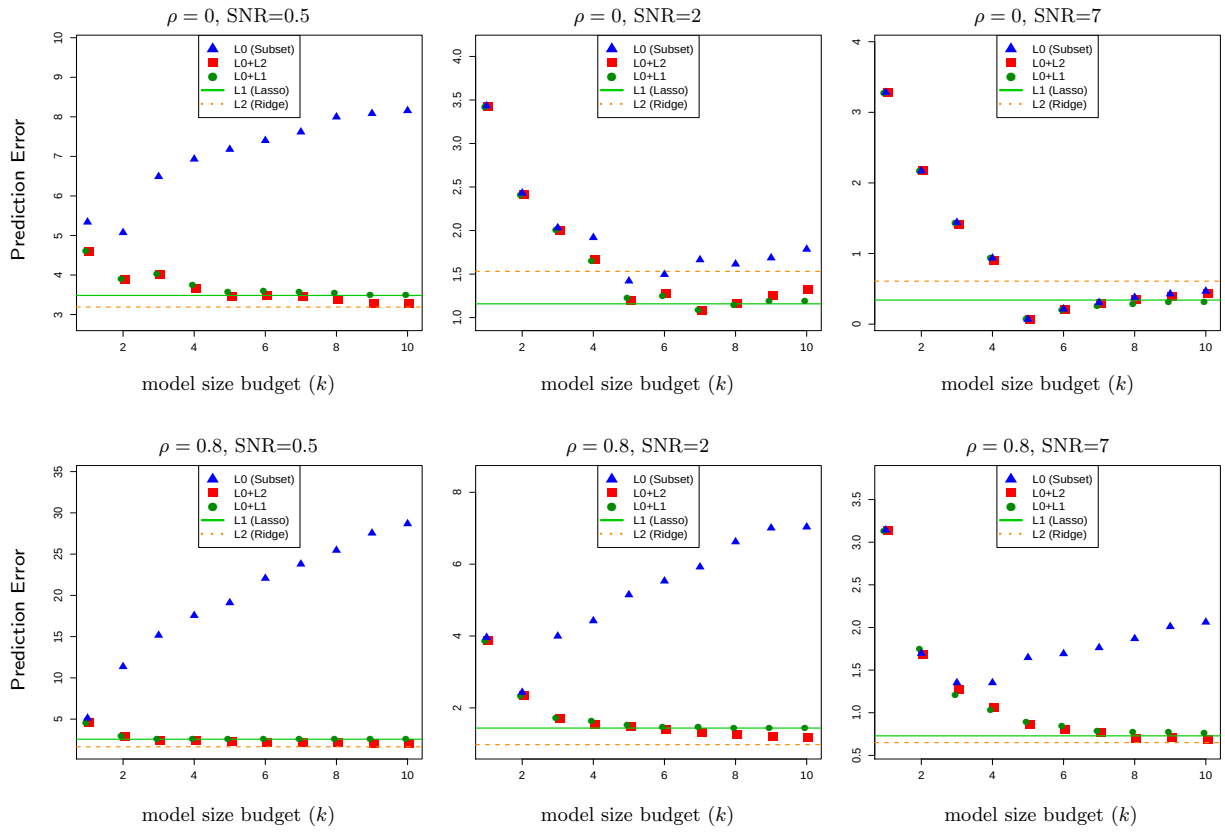


Figure 1: Prediction error $\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2$, averaged over the simulated datasets described in the text, for the Lasso (L1), ridge regression (L2), best-subsets (L0), and the estimators proposed in Problem (3): L0+L1 ($q = 1$) and L0+L2 ($q = 2$). Given the model size parameter k (irrelevant to L1 and L2), the average prediction error of best predictive model across λ (irrelevant to L0) is plotted for each method. The best L1 models have average sizes 11.4, 17.8, 18.4 [top panel] and 8.3, 12.0, 16.6 [bottom panel], while the L2 models are completely dense.

selection, as this may be not be statistically possible when the noise-level is high. Instead, we focus on improving the predictive performance of the best-subsets approach, with an explicit control of the model size – we also wish to devise an estimator that is based on a simple and easy-to-interpret optimization criterion.

In Section 2 we formulate the optimization problem for our proposed estimator and describe how to compute the corresponding solutions using modern computational tools from mathematical optimization. In Section 3 we study the theoretical properties of our proposed approach. First, we establish non-asymptotic error bounds for the new estimators. Second, we derive novel lower-bounds on the prediction error for the best-subsets estimator in settings where the noise-level is

high, and then contrast the predictive performance of best-subsets with that of our estimators. In Section 4 we discuss the connections between our proposal and existing work, and in Section 5 we evaluate the performance of the proposed estimators empirically. Theoretical proofs and some computational details are provided in the Supplementary Material.

2 Methodological Framework

Continuous shrinkage methods that solve optimization problems of the form

$$\underset{\boldsymbol{\beta}}{\text{minimize}} \qquad \frac{1}{2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2+\lambda\|\boldsymbol{\beta}\|_q^q, \tag{2}$$

such as ridge regression ($q=2$) and the Lasso ($q=1$), are generally recognized for producing estimators with excellent predictive performance, however, their estimated models are denser than those produced by best-subsets (see Figure 1). Similarly to the best-subsets approach, the Lasso *searches* for a subset of features, however, unlike best-subsets, it then regularizes the least-squares regression performed on the selected features. The superior predictive performance of the Lasso can be attributed in part to the shrinkage effect of the ℓ_1-penalty. Perhaps even more compelling is the example of ridge regression – there is no searching here per se, as all the estimated coefficients are generally nonzero. The excellent predictive performance of ridge regression can be attributed fully to the shrinkage induced by the ℓ_2-penalty.

2.1 The proposed estimator

The above discussion suggests the possibility of obtaining a *sparse* linear model with predictive performance better than best-subsets and comparable to, or even better than, ridge regression and the Lasso. In terms of sparsity, we desire an estimator with fewer nonzero coefficients than the Lasso, for example. We propose the following regularized best-subsets estimator[1]:

$$\underset{\boldsymbol{\beta}}{\text{minimize}} \qquad \frac{1}{2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2+\underbrace{\lambda\|\boldsymbol{\beta}\|_q}_{\text{Shrinkage}} \qquad \text{s.t.} \qquad \underbrace{\|\boldsymbol{\beta}\|_0\leq k}_{\text{Sparsity}}. \tag{3}$$

Above, the cardinality constraint on $\boldsymbol{\beta}$ directly controls the model size, and the ℓ_q-penalty[2] with $q\in\{1,2\}$ shrinks the regression coefficients towards zero using $\lambda>0$ as the shrinkage parameter.

[1]Estimator (3) is inspired by *regularized SVD* estimators (involving a nuclear norm penalty and a rank constraint) commonly used in collaborative filtering [32] and matrix completion [23].

[2]Note that Problem (3) uses the ℓ_q rather than the ℓ_q^q penalization, to be consistent with the theoretical results in Section 3. However, our computational framework can handle both versions of the problem.

Informally speaking[3], Problem (3) separates out the effects of shrinkage (via $\lambda\|\boldsymbol{\beta}\|_q$) and sparsity (via $\|\boldsymbol{\beta}\|_0 \leq k$) – this may be contrasted with the Lasso, where the penalty simultaneously controls both shrinkage and sparsity, and best subset selection, which only selects but does not shrink. The family of estimators (3) contains as special cases the best-subsets estimator given by Problem 1 ($\lambda = 0$), the Lasso family ($k = p$, $q = 1$) and the ridge regression[4] family ($k = p$, $q = 2$) of estimators. For other values of λ and k, Problem (3) combines the best of both worlds: best-subsets (Problem 1) and continuous shrinkage methods (Problem 2).

Figure 1 shows that when $k > \|\boldsymbol{\beta}^*\|_0$, continuous shrinkage regulates the overfitting behavior of best-subsets: as k increases, estimator (3) overfits more slowly when compared to best-subsets. This observation is also supported by our theory in Section 3. When the SNR is low, shrinkage imparted via ℓ_q-regularization becomes critical – estimator (3) prefers to choose a strictly positive value of λ to produce a good predictive model. The ℓ_1-penalty in estimator (3) with $q = 1$ can also act as an additional sparsification tool when k is large – this partially explains its (marginally) superior predictive accuracy over $q = 2$ for larger SNR values. Overall, Figure 1 illustrates that estimator (3) produces sparser models than the Lasso, while its predictive performance is consistently as good as or better than that of the continuous shrinkage methods.

Problem (3) is a nonconvex optimization problem. However, as we show in Section 2.2, it can be expressed as a mixed integer second order conic optimization (MISOCO) problem and solved (in practice) to certifiable optimality by leveraging advances in modern integer optimization techniques, using standard solvers like Cplex, Gurobi, Knitro, Mosek, Glpk, Scip [34, 58]. To obtain high-quality solutions to Problem (3) at low computational cost, we develop specialized discrete first order methods [48] in Section 2.3, by extending the framework in [9, 40]. When these algorithms are used with our proposed continuation schemes across (λ, k) and randomized local search heuristics [1, 44], a family of (near optimal) feasible solutions to Problem (3) can be computed within minutes. These algorithms, however, do not certify the quality of the solutions in terms of lower-bounds on the objective function. For this we need the power of MIO techniques. When our heuristic algorithms are used in conjunction with MISOCO solvers for Problem (3), they lead to improved computational performance – see, for example, [9, 40] for similar observations on related problems.

[3]When $q = 1$, the shrinkage penalty may induce further sparsity.

[4]The coefficient path for Problem (3) contains the ridge regression coefficient path.

2.2 Mixed Integer Optimization formulations

Here we present the MIO formulation for Problem (3). Denoting $\{1, \ldots, p\}$ by $[p]$ and assuming, without loss of generality,[5] that $\boldsymbol{\beta} \in [-\mathcal{M}, \mathcal{M}]^p$, we can rewrite (3) as follows:

$$\text{minimize} \;\; \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_q \;\; \text{s.t.} \;\; -\mathcal{M}z_j \leq \beta_j \leq \mathcal{M}z_j, j \in [p]; \; \mathbf{z} \in \{0,1\}^p; \; \sum_j z_j = k. \quad (4)$$

Here, $\boldsymbol{\beta}$ and $\mathbf{z}$ are the optimization variables and $\mathcal{M} < \infty$ is a BigM parameter [8, 9], which is sufficiently large, so that a solution to Problem (4) is also a solution to Problem (3). The binary variable z_j controls whether β_j is zero or not: $z_j = 1$ implies that β_j is *free* to vary in $[-\mathcal{M}, \mathcal{M}]$, and $z_j = 0$ implies $\beta_j = 0$. The constraint $\sum_j z_j = k$ allows at most k regression coefficients to be nonzero. The nonconvexity in (4) stems from the binary variables in $\mathbf{z}$. Problem (4) can be reformulated as a MISOCO, i.e., a second order conic optimization problem [12] where a subset of the variables is binary. Thanks to the impressive advances in MIO, these problems can be solved in practice using state-of-the-art MIO solvers [see, for example, the recent work of 58]. To this end, we note that (4) can be written as follows:

$$\text{minimize} \;\; u/2 + \lambda v \;\; \text{s.t.} \;\; \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \leq u, \; \|\boldsymbol{\beta}\|_q \leq v, \; (\boldsymbol{\beta}, \mathbf{z}) \in \mathcal{C}, \quad (5)$$

where the optimization variables are $(u, v, \boldsymbol{\beta}, \mathbf{z}) \subset \mathbb{R} \times \mathbb{R} \times \mathbb{R}^p \times \{0,1\}^p$, and $\mathcal{C}$ denotes the mixed integral polyhedral constraint in (4). The first term in the constraint can be expressed as a second order cone [12],

$$\left\{(\boldsymbol{\beta}, u) : \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \leq u, u \geq 0\right\} \quad \equiv \quad \left\{(\boldsymbol{\beta}, u) : \left\|\left([\mathbf{y} - \mathbf{X}\boldsymbol{\beta}]^\top, [u-1]/2\right)\right\|_2 \leq (u+1)/2, u \geq 0\right\}.$$

For $q = 1$, the term $\|\boldsymbol{\beta}\|_q \leq v$ in the constraint can be expressed via linear inequalities using auxiliary continuous variables $\{\bar{\beta}\}_1^p$:

$$\{(\boldsymbol{\beta}, v) : \|\boldsymbol{\beta}\|_1 \leq v, v \geq 0\} \quad \equiv \quad \{(\boldsymbol{\beta}, v) : \exists \, \bar{\boldsymbol{\beta}} \geq \mathbf{0} \;\; \text{s.t.} \;\; -\bar{\beta}_j \leq \beta_j \leq \bar{\beta}_j, \; \sum_j \bar{\beta}_j \leq v, v \geq 0\}, \quad (6)$$

thereby leading to a MISOCO formulation for (5) when $q = 1$. When $q = 2$, the epigraph version of $\|\boldsymbol{\beta}\|_q \leq v$ is already a second order cone, so (5) admits a MISOCO formulation.

Other Formulations. Computational performance of MISOCO solvers (Gurobi, for example) is found to improve by adding structural implied inequalities, or cuts, to the basic formulation (5) –

[5] Note that every solution to (3) is bounded when $\lambda > 0$, because the level sets of the objective function are bounded. The case for $\lambda = 0$ has been addressed in [9].

see Section A.2 of the Supplementary Material. Computation of problem-specific BigM parameters and other bounds is discussed in Section A.3.

Problem (4) with $q = 1$ can also be expressed as a mixed integer quadratic optimization (MIQO) problem. Note that if we replaced the ℓ_2-penalty in (5) with the squared-ℓ_2-penalty, then the resulting problem would be readily expressed as MIQO as well – both problems leading to the same family of solutions[6]. In what follows, we will focus on the MISOCO formulation presented above to be consistent with our theoretical results in Section 3.

2.3 Discrete First Order Algorithms

Inspired by proximal gradient methods [48, 47], popularly used in convex optimization, we present discrete first order (DFO) methods to obtain good upper bounds for (3). The DFO methods have a low iteration complexity and can nicely exploit warm-start information across the (λ, k)-space: Using a combination of neighborhood continuation schemes and local combinatorial search methods proposed here, they lead to near-optimal[7] solutions to (3). We note that the DFO methods are heuristics– they do not certify solution quality (i.e., global optimality) via dual-bounds. For the latter, we critically rely on MIO technology. The MIO solvers accept warm-starts available from the DFO algorithm, then subsequently improve the solution and certify optimality, at the cost of additional (but still reasonable) computation times.

We describe a DFO method for the following problem (in composite form [47]):

$$\text{minimize} \quad F(\boldsymbol{\beta}) := \quad f(\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_q \quad \text{s.t.} \quad \|\boldsymbol{\beta}\|_0 \leq k, \tag{7}$$

where $f(\boldsymbol{\beta})$ is a L_0-smooth convex function, i.e., it satisfies

$$\|\nabla f(\boldsymbol{\beta}) - \nabla f(\boldsymbol{\alpha})\|_2 \leq L_0 \|\boldsymbol{\beta} - \boldsymbol{\alpha}\|_2 \quad \forall\, \boldsymbol{\beta}, \boldsymbol{\alpha} \in \mathbb{R}^p. \tag{8}$$

For $f(\boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$, we can use $L_0 = \sigma_{\max}(\mathbf{X})^2$, where $\sigma_{\max}(\cdot)$ is the maximum singular value of $\mathbf{X}$. As a consequence of (8), for any $L \geq L_0$, we have the following bound [48] in place:

$$f(\boldsymbol{\beta}) \leq f(\boldsymbol{\alpha}) + \langle \nabla f(\boldsymbol{\alpha}), \boldsymbol{\beta} - \boldsymbol{\alpha} \rangle + \frac{L}{2}\|\boldsymbol{\beta} - \boldsymbol{\alpha}\|_2^2 := Q_L(\boldsymbol{\beta}; \boldsymbol{\alpha}), \quad \forall \boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^p. \tag{9}$$

[6] If we denote the solution to the modified problem by $\hat{\boldsymbol{\beta}}_{\ell_2^2}(\lambda', k)$, then, for every fixed k, the solution path $\{\hat{\boldsymbol{\beta}}_{\ell_2^2}(\lambda', k)\}_{\lambda' \geq 0}$ recovers the corresponding path for the original Problem (4) with $q = 2$.

[7] In our experiments, we observed that the solutions obtained by our elaborate heuristics are often close to the optimal solutions returned by the MIO solvers in the neighborhood of the optimal (λ, k) choice, made by minimizing the prediction error on a separate validation set.

Given a current solution $\boldsymbol{\alpha}$, our algorithm minimizes an upper bound to $F(\boldsymbol{\beta})$ around $\boldsymbol{\alpha}$:

$$\underset{\|\boldsymbol{\beta}\|_0 \leq k}{\text{minimize}} \;\; Q_L(\boldsymbol{\beta}; \boldsymbol{\alpha}) + \lambda \|\boldsymbol{\beta}\|_q \iff \underset{\|\boldsymbol{\beta}\|_0 \leq k}{\text{minimize}} \;\; \frac{L}{2} \left\| \boldsymbol{\beta} - \left(\boldsymbol{\alpha} - \frac{1}{L} \nabla f(\boldsymbol{\alpha}) \right) \right\|_2^2 + \lambda \|\boldsymbol{\beta}\|_q. \tag{10}$$

A key ingredient in solving the above is the thresholding operator,

$$\mathrm{S}(\mathbf{u}; k; \lambda \ell_q) := \underset{\boldsymbol{\beta}: \|\boldsymbol{\beta}\|_0 \leq k}{\arg\min} \;\; \frac{1}{2} \|\boldsymbol{\beta} - \mathbf{u}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_q, \tag{11}$$

where $\mathrm{S}(\mathbf{u}; k; \lambda \ell_q)$ denotes the set of optimal solutions to Problem (11). We note that $\mathrm{S}(\mathbf{u}; k; \lambda \ell_q)$ may be set-valued – the non-uniqueness of an optimal solution to Problem (11) arises from the fact that the ordering of $|u_j|$ for $j \in [p]$ may have ties.

Proposition 1. *Let $(1), \ldots, (p)$ be a permutation of the indices $1, \ldots, p$, such that the entries in $\mathbf{u}$ are sorted as: $|u_{(1)}| \geq |u_{(2)}| \geq \ldots \geq |u_{(p)}|$. Then, the thresholding operator* (11) *has the following form:*

(a) For the ℓ_1-regularizer (with $q = 1$) any $\hat{\boldsymbol{\beta}} \in \mathrm{S}(\mathbf{u}; k; \lambda \ell_q)$ is given by:

$$\hat{\beta}_i = \begin{cases} \mathrm{sgn}(u_i) \max\{|u_i| - \lambda, 0\} & i \in \{(1), (2), \ldots, (k)\} \\ 0 & \textit{otherwise.} \end{cases} \tag{12}$$

(a) For the ℓ_2-regularizer (with $q = 2$) any $\hat{\boldsymbol{\beta}} \in \mathrm{S}(\mathbf{u}; k; \lambda \ell_q)$ is given by:

$$\hat{\beta}_i = \begin{cases} \frac{u_i}{\tau_u} \max\{\tau_u - \lambda, 0\} & i \in \{(1), (2), \ldots, (k)\} \\ 0 & \textit{otherwise,} \end{cases} \tag{13}$$

where $\tau_u = \sqrt{\sum_{i=1}^k u_{(i)}^2}$ is the ℓ_2-norm of the k largest (in magnitude) entries of $\mathbf{u}$.

The DFO algorithm performs the following updates (for $m \geq 1$)

$$\boldsymbol{\beta}^{(m+1)} \in \mathrm{S}\left(\boldsymbol{\beta}^{(m)} - \tfrac{1}{L} \nabla f(\boldsymbol{\beta}^{(m)}); k; \tfrac{\lambda}{L} \ell_q \right), \tag{14}$$

till some convergence criterion is met. The algorithm is summarized below for convenience.

Discrete First Order Algorithm (DFO)

1. Fix $L \geq L_0$ and a convergence threshold $\tau > 0$. Initialize with $\boldsymbol{\beta}^{(1)}$ that is k-sparse. Repeat update (14) until $\|\boldsymbol{\beta}^{(m+1)} - \boldsymbol{\beta}^{(m)}\|_2^2 \leq \tau$.

2. Let $\mathtt{I}(\widetilde{\boldsymbol{\beta}})$ denote the support of the $\widetilde{\boldsymbol{\beta}}$ obtained from Step 1, i.e., $\mathtt{I}(\widetilde{\boldsymbol{\beta}}) = \{i : \widetilde{\beta}_j \neq 0, j \in [p]\}$. Solve the convex problem (7) restricted to the support $\mathtt{I}(\widetilde{\boldsymbol{\beta}})$: min $F(\boldsymbol{\beta})$ s.t. $\beta_j = 0, j \notin \mathtt{I}(\widetilde{\boldsymbol{\beta}})$.

For the sake of completeness, we establish convergence properties of the sequence $\{\boldsymbol{\beta}^{(m)}\}_{m\geq 1}$ in terms of reaching a first order stationary point. Our work adapts the framework proposed in [9] to the composite form. Towards this end, we need the following definition.

Definition 1. *We say that $\boldsymbol{\eta}$ is a first order stationary point of Problem (7) if $\boldsymbol{\eta} \in \mathrm{S}(\boldsymbol{\eta} - \frac{1}{L}\nabla f(\boldsymbol{\eta}); k; \frac{\lambda}{L}\ell_q)$. We say that $\boldsymbol{\eta}$ is an ϵ-accurate first order stationary point if $\|\boldsymbol{\eta}\|_0 \leq k$ and $\|\boldsymbol{\eta} - \mathrm{S}(\boldsymbol{\eta} - \frac{1}{L}\nabla g(\boldsymbol{\eta}); k; \frac{\lambda}{L}\ell_q)\|_2^2 \leq \epsilon$.*

The following result presents convergence properties of the sequence $\{\boldsymbol{\beta}^{(m)}\}_{m\geq 1}$ in terms of reaching a first order stationary point (see Section A.1, Supplementary Material for the proof).

Proposition 2. *Let $\{\boldsymbol{\beta}^{(m)}\}$ denote a sequence generated by the DFO algorithm. Then,*

(a) for $L \geq L_0$, the sequence $F(\boldsymbol{\beta}^{(m)})$ is decreasing, and it converges to some $F^* \geq 0$;

(b) for $L > L_0$, we have the following finite-time convergence rate:

$$\min_{1\leq m\leq M} \|\boldsymbol{\beta}^{(m+1)} - \boldsymbol{\beta}^{(m)}\|_2^2 \leq \frac{2(F(\boldsymbol{\beta}^{(1)}) - F^*)}{M(L - L_0)}.$$

Proposition 2 suggests that the DFO algorithm applied to Problem (7) leads to a decreasing sequence of objective values, which eventually converges. When $L > L_0$ the algorithm reaches an ϵ-accurate first order stationary point (Definition 1) in $O(\epsilon^{-1})$ iterations. We note that the proposition makes no assumption on the data at hand – improved convergence rates may be achievable by making further assumptions on the problem data (see, for example, [9] and the discussion therein). In practice however, the DFO algorithm converges much faster (especially when using warm-start continuation) than the sublinear rate suggested by Proposition 2.

2.4 Neighborhood continuation and local search heuristics

Due to the nonconvexity of Problem (3), the DFO algorithm is sensitive to the initialization $\boldsymbol{\beta}^{(1)}$. The effect of initialization becomes particularly pronounced when n is relatively small compared to p, the pairwise (sample) correlations among the features are high; and the SNR is low. These solutions can be improved, often substantially (in terms of the objective value), using continuation schemes and randomized local search-heuristics, as we discuss below. The continuation scheme, which makes use of the warm-starting capabilities of the DFO algorithm, is quite efficient. Note that these algorithms serve as stand-alone methods to obtain good feasible solutions for (3), for a family of tuning parameters (λ, k) – this makes them practically appealing. Furthermore, these

methods can be used to obtain a good estimate of an optimal tuning parameter (for example, based on validation set tuning) with relatively low computational cost.

Neighborhood Continuation. Let $\hat{\boldsymbol{\beta}}(\lambda, k)$ denote a solution delivered by the DFO algorithm for (3) (we drop the dependence on q for notational convenience). We let $F(\lambda, k)$ denote the corresponding objective value. We consider a 2D grid of tuning parameters in $\Lambda \times K = \{\lambda_1, \ldots, \lambda_N\} \times \{k_1, \ldots, k_r\}$ with $\lambda_i > \lambda_{i+1}$ and $k_i > k_{i+1}$ for all i. We set $k_1 = p, k_r = 1$. We set $\lambda_1 = \|\mathbf{X}^\top \mathbf{y}\|_{\bar{q}}$ with $\bar{q} = \infty$ if $q = 1$ and $\bar{q} = 2$ if $q = 2$ – the rationale being that if $\lambda = \lambda_1$, then an optimal solution to Problem (3) is zero.

Algorithm 1: Neighborhood Continuation

(i) Initialize $\hat{\boldsymbol{\beta}}(\lambda_i; k_j) \leftarrow \mathbf{0}$ for every $i, j \in [N] \times [r]$. Repeat Step (ii) until the array of objective values $\{F(\lambda_i; k_j)\}_{i,j}$ stops changing between successive sweeps across the 2D grid $\Lambda \times K$:

(ii) For $i \in [N], j \in [r]$ do the following:

  (a) Set $(\lambda, k) = (\lambda_i, k_j)$ and use the DFO algorithm with (at most) four different neighborhood initializations $\hat{\boldsymbol{\beta}}(\lambda_a; k_b)$, $(a, b) \in \mathcal{N}(i, j)$ where, $\mathcal{N}(i, j)$ are the neighbors of (i, j). For every (a, b) in the neighborhood $\mathcal{N}(i, j)$, let $\hat{\boldsymbol{\beta}}_{a,b}$ and $F_{a,b}$ denote the corresponding estimate and objective value, respectively.

  (b) Set $\hat{\boldsymbol{\beta}}(\lambda_i; k_j)$ equal to the estimate $\hat{\boldsymbol{\beta}}_{a,b}$ with the smallest objective value: $F(\lambda_i; k_j) = \min\{F_{a,b} : (a, b) \in \mathcal{N}(i, j)\}$.

We make a series of remarks pertaining to Algorithm 1:

- If we denote one execution of Step-(ii) (formed by looping across all $i, j \in [N] \times [r]$) as a sweep, then successive sweeps may lead to a strict improvement[8] in the objective values $\{F(\lambda_i, k_j)\}_{i,j}$ for several (i, j).
- During the first sweep of Algorithm 1 many neighbors $\hat{\boldsymbol{\beta}}(\lambda_a, r_b)$ of (i, j) are zero. After the first sweep, however, all entries (i, j) get populated.
- The neighborhood initializations $\hat{\boldsymbol{\beta}}(\lambda_a; k_b)$ for $(a, b) \in \mathcal{N}(i, j)$ serve as excellent warm-starts for (3) at (λ_i, r_j). This improves the overall runtime of the algorithm (as compared to independently computing the solutions on the 2D grid) and also results in a solution with good objective values.

A (randomized) local search heuristic. We present a local-search heuristic, which, loosely

[8] By construction, given $(i, j) \in [N] \times [r]$, the objective value $F(\lambda_i, k_j)$ cannot increase between successive sweeps.

speaking, is capable of navigating different parts of the model space by perturbing the support of a DFO solution. We draw inspiration from local search schemes commonly used in combinatorial optimization problems [1, 44]. Our local search scheme works as follows: for every nonzero initialization $\hat{\boldsymbol{\beta}}(\lambda_a, k_b)$, we randomly swap roughly 50% of the nonzero coefficients with an equal number of zero coefficients before passing the resulting estimate as an initialization to the DFO algorithm. This stochastic search scheme is performed as a part of the 2D continuation scheme (described above) – we register the estimate if it leads to an improvement in the objective value.

3 Statistical Theory

We study the performance of the proposed approach in the regression setting with deterministic design. In Sections 3.1-3.3 we establish non-asymptotic oracle error bounds for the corresponding estimators. In Section 3.4 we contrast the predictive performance of the new approach with that of best-subsets selection, by deriving novel lower-bounds on the prediction error of $\widehat{\boldsymbol{\beta}}_{\ell_0}$. The comparison between the estimators is done for each fixed value of the model size tuning parameter k. In Section 3.5 we analyze a BIC-type approach for selecting the optimal value of k. Our results provide new insights on the benefits of additional regularization in best subset selection.

3.1 Notation and preliminary results

We assume that the observed data follows the model

$$\mathbf{y} = \mathbf{f}^* + \boldsymbol{\epsilon}. \tag{15}$$

The components in the equation above are vectors in $\mathbb{R}^n$, vector $\mathbf{f}^*$ is an unknown deterministic mean, and the elements of $\boldsymbol{\epsilon}$ are independent $N(0, \sigma^2)$ with $\sigma > 0$. A special case of (15) is the linear model $\mathbf{f}^* = \mathbf{X}\boldsymbol{\beta}^*$. As before, we assume that the columns of $\mathbf{X}$ have unit ℓ_2-norm.

We use the following notation for the regularized best-subsets solutions to Problem (3):

$$\widehat{\boldsymbol{\beta}}_q = \arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_q \quad \text{s.t.} \quad \|\boldsymbol{\beta}\|_0 \leq k, \qquad \text{for } q = 1, 2. \tag{16}$$

The dependence of $\widehat{\boldsymbol{\beta}}_q$ on k and λ is understood implicitly. From here on, we drop the subscript in the notation $\|\cdot\|_2$, used for the Euclidean norm. To simplify the presentation, we refer to $\|\mathbf{f}^* - \widehat{\boldsymbol{\beta}}_q\|^2$ as the prediction error for $\widehat{\boldsymbol{\beta}}_q$, multiplying the usual prediction error by n. Given an

integer $s \in [p]$, we define $B_0(s) = \{\mathbf{u} \in \mathbb{R}^p : \|\mathbf{u}\|_0 \leq s\}$ and let γ_s denote the minimal s-sparse eigenvalue of $\mathbf{X}$:

$$\gamma_s = \min_{\mathbf{u} \neq \mathbf{0}, \mathbf{u} \in B_0(s)} \frac{\|\mathbf{X}\mathbf{u}\|}{\|\mathbf{u}\|}.$$

Given a vector $\mathbf{u} \in \mathbb{R}^p$, we write $u_1^\sharp, ..., u_p^\sharp$ for a non-increasing rearrangement of $|u_1|, ..., |u_p|$. We say that a constant is *universal* if it does not depend on other parameters, such as k, p or λ. We write $\gtrsim$ and $\lesssim$ to indicate that inequalities $\geq$ and $\leq$, respectively, hold up to positive universal multiplicative factors, and use $\asymp$ when the two inequalities hold simultaneously. We use the notation $a \vee b = \max\{a, b\}$, $a \wedge b = \min\{a, b\}$, and treat algebraic expressions of the form $0 \cdot \infty$ or $0/0$ as zero.

As is typical in high-dimensional regression settings, we establish the error bounds by conducting deterministic arguments on suitably chosen random events:

$$\begin{aligned}
\mathcal{E}_s &= \big\{\boldsymbol{\epsilon}^\top \mathbf{X}\mathbf{u} \leq [4+\sqrt{2}]\sigma\sqrt{s\log(2ep/s)}\|\mathbf{u}\|,\ \forall \mathbf{u} \in B_0(s)\big\} \\
\mathcal{F} &= \big\{\boldsymbol{\epsilon}^\top \mathbf{X}\mathbf{u} \leq [4+\sqrt{2}]\sigma \max\big(\textstyle\sum_{j=1}^p u_j^\sharp\sqrt{\log(2p/j)}\,,\ \sqrt{\log(1/\delta_0)}\|\mathbf{X}\mathbf{u}\|\big),\ \forall \mathbf{u} \in \mathbb{R}^p\big\} \\
\mathcal{G}_s &= \big\{\boldsymbol{\epsilon}^\top \mathbf{X}\mathbf{u} \leq \sigma\sqrt{5s\log(ep/s) + \log(1/\delta_0)}\|\mathbf{X}\mathbf{u}\|,\ \forall \mathbf{u} \in B_0(s)\big\} \\
\mathcal{H} &= \big\{\|\mathbf{X}^\top\boldsymbol{\epsilon}\|_\infty \leq \sigma\sqrt{2\log(2p)} + \sigma\sqrt{2\log(1/\delta_0)}\big\}.
\end{aligned}$$

When s/p and δ_0 are small, all four events hold with high probability.

Theorem 1. *Suppose that $s \in [p]$ and $\delta_0 \in (0, 1]$. Then,*

$$\mathbb{P}(\mathcal{E}_s) \geq 1 - s/(4ep), \quad \mathbb{P}(\mathcal{F}) \geq 1 - \delta_0/2, \quad \mathbb{P}(\mathcal{G}_s) \geq 1 - \delta_0 \quad \textit{and} \quad \mathbb{P}(\mathcal{H}) \geq 1 - \delta_0.$$

Some of the above probability bounds have appeared in the literature. In particular, the bound for $\mathcal{F}$, which is an important component of our analysis, was recently established in [5].

3.2 Results for the ℓ_2-regularized best-subsets estimator

We follow the common convention in the literature [17, for example] by referring to prediction error rates that involve terms of order λ^2 as *fast* and referring to prediction error rates that involve terms of order λ as *slow*. The slow rates are especially relevant to our study, because they tend to outperform the fast rates in the high-noise regimes. The following result focuses on $\widehat{\boldsymbol{\beta}}_2$ and provides both the slow and the fast rate prediction error bounds. We note that an important attractive feature of the last two error bounds in Theorem 2 is the independence of the uncertainty parameter δ_0 from the tuning parameters λ and k. This feature allows us to control the expected prediction error, as we demonstrate in Corollary 3 below.

Theorem 2. *(A) Slow rate. If* $\lambda \geq [8+2\sqrt{2}]\sigma\sqrt{2k\log(ep/k)}$*, then on the event* $\mathcal{E}_{2k}$*,*

$$\|\mathbf{f}^* - \mathbf{X}\widehat{\boldsymbol{\beta}}_2\|^2 \leq \inf_{\boldsymbol{\beta}\in B_0(k)} \Big[\|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|^2 + 2\lambda\|\boldsymbol{\beta}\|\Big];$$

and on the event $\mathcal{F}$*,*

$$\|\mathbf{f}^* - \mathbf{X}\widehat{\boldsymbol{\beta}}_2\|^2 \lesssim \inf_{\boldsymbol{\beta}\in B_0(k)} \Big[\|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|\Big] + \sigma^2\log(1/\delta_0).$$

(B) Fast rate. On the event $\mathcal{G}_{2k}$*,*

$$\|\mathbf{f}^* - \mathbf{X}\widehat{\boldsymbol{\beta}}_2\|^2 \lesssim \inf_{\boldsymbol{\beta}\in B_0(k)} \|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|^2 + \sigma^2 k\log(ep/k) + \gamma_{2k}^{-2}\lambda^2 + \sigma^2\log(1/\delta_0)$$

for every $\lambda \geq 0$*.*

The above result establishes oracle inequalities for the prediction error under potential model misspecification. The added generality allows us to avoid restrictions on the model size parameter k. This is relevant to the discussion in Section 3.4 on the relationship between decreasing k and the predictive performance of best-subsets. We note that our oracle inequalities are restricted to $B_0(k)$ for each fixed value of the model size tuning parameter k. In Section 3.5 we present a data-driven approach for selecting k and establish oracle inequalities in a more general form.

To illustrate the rates of convergence in Theorem 2 more clearly, we consider the linear case, $\mathbf{f}^* = \mathbf{X}\boldsymbol{\beta}^*$, and set δ_0 equal to some specific small values.

Corollary 1. *Let* $\mathbf{f}^* = \mathbf{X}\boldsymbol{\beta}^*$ *for some* $\boldsymbol{\beta}^* \in B_0(k)$*. If* $\lambda \geq [8+2\sqrt{2}]\sigma\sqrt{2k\log(ep/k)}$*, then*

$$\|\mathbf{X}\widehat{\boldsymbol{\beta}}_2 - \mathbf{X}\boldsymbol{\beta}^*\|^2 \leq 2\lambda\|\boldsymbol{\beta}^*\|$$

with probability at least $1 - k/(2ep)$*, and*

$$\|\mathbf{X}\widehat{\boldsymbol{\beta}}_2 - \mathbf{X}\boldsymbol{\beta}^*\|^2 \lesssim \lambda\|\boldsymbol{\beta}^*\| + \sigma^2\log(p)$$

with probability at least $1 - 1/p$*. Furthermore, with probability at least* $1 - (k/p)^k$*,*

$$\|\mathbf{X}\widehat{\boldsymbol{\beta}}_2 - \mathbf{X}\boldsymbol{\beta}^*\|^2 \lesssim \sigma^2 k\log(ep/k) + \gamma_{2k}^{-2}\lambda^2 \quad \textit{and} \quad \|\widehat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}^*\| \lesssim \gamma_{2k}^{-1}\sigma\sqrt{k\log(ep/k)} + \gamma_{2k}^{-2}\lambda$$

for every $\lambda \geq 0$*.*

We make the following observations regarding the established error bounds for $\widehat{\boldsymbol{\beta}}_2$.

Remark 1. *Letting* $k = k^*$*, we note that the fast prediction error rate,* $\sigma^2 k^*\log(ep/k^*)$*, matches the minimax rate over* $\boldsymbol{\beta}^* \in B_0(k^*)$ *[52, 36, 51].*

Remark 2. *When $\lambda = 0$, the fast rate part of Corollary 1 yields the prediction and estimation error bounds for the best-subsets estimator, $\widehat{\boldsymbol{\beta}}_{\ell_0}$.*

Remark 3. *The slow rate for $\widehat{\boldsymbol{\beta}}_2$ is $\sigma\sqrt{k^* \log(ep/k^*)}\|\boldsymbol{\beta}^*\|$, which improves on the prediction error bound for $\widehat{\boldsymbol{\beta}}_{\ell_0}$ when $\|\boldsymbol{\beta}^*\|/\sigma \lesssim \sqrt{k^* \log(ep/k^*)}$ with a sufficiently small universal constant.*

Remark 4. *The lower-bound on λ needed for the slow rate results contains the unknown parameter σ, i.e., the standard deviation of the noise in the model. The noise variance can be estimated by employing a preliminary regression estimator [see, for example, the discussion in 6] that is unrestricted in terms of the model size. In practice, parameter λ can be tuned based on a separate validation set (or by cross-validation), leading to the best k-sparse model with respect to the validation error.*

The error rates presented above can also apply to approximate solutions, obtained after an early termination of the MIO solver. Upon termination, the solver provides the upper and lower bounds on the value of the objective in (16). We denote these bounds by UB and LB, respectively, and write $\tau = (UB - LB)/UB$ for the corresponding optimality gap. The next result focuses on an approximate ℓ_2-regularized best-subsets solution $\widetilde{\boldsymbol{\beta}}_2$ in the linear setting of Corollary 1.

Corollary 2. *Let $\mathbf{f}^* = \mathbf{X}\boldsymbol{\beta}^*$ for some $\boldsymbol{\beta}^* \in B_0(k)$ and suppose that $\tau \leq 1 - c$ for some positive universal constant c. Then, with probability at least $1 - 1/p$,*

$$\|\mathbf{X}\widetilde{\boldsymbol{\beta}}_2 - \mathbf{X}\boldsymbol{\beta}^*\|^2 \lesssim \lambda\|\boldsymbol{\beta}^*\| + \sigma^2\big[\log(p) + \tau n\big] \quad \text{for } \lambda \geq [8 + 2\sqrt{2}]\sigma\sqrt{2k\log(ep/k)}.$$

In addition, with probability at least $1 - (k/p)^k$,

$$\|\mathbf{X}\widetilde{\boldsymbol{\beta}}_2 - \mathbf{X}\boldsymbol{\beta}^*\|^2 \lesssim \sigma^2\big[k\log(ep/k) + \tau n\big] + \gamma_{2k}^{-2}\lambda^2 \quad \text{for every } \lambda \geq 0.$$

We note that $\widetilde{\boldsymbol{\beta}}_2$ achieves the second slow error rate in Corollary 1 when $\tau \lesssim \log(p)/n$, and it achieves the corresponding fast error rate when $\tau \lesssim k\log(ep/k)/n$. Furthermore, we show in the proof of Corollary 2 that the multiplicative increase in the slow rate error bound relative to the case $\tau = 0$ is at most $1 + \frac{\tau}{1-\tau}\big\{1 \vee \frac{n}{58\log(p)}\big\}$; the corresponding multiplicative increase in the fast rate error bound is at most $1 + \frac{\tau}{1-\tau}\big\{1 \vee \frac{n}{43k\log(ep/[2k])}\big\}$. These expressions illustrate the trade-off between the optimality gap and the quality of the prediction error bounds.

The next result bounds the expected prediction error of $\widehat{\boldsymbol{\beta}}_2$.

Corollary 3. *If* $\lambda \geq [8+2\sqrt{2}]\sigma\sqrt{2k\log(ep/k)}$, *then*

$$\mathbb{E}\|\mathbf{f}^* - \mathbf{X}\widehat{\boldsymbol{\beta}}_2\|^2 \lesssim \inf_{\boldsymbol{\beta}\in B_0(k)} \Big[\|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|\Big] + \sigma^2.$$

Furthermore, for every $\lambda \geq 0$,

$$\mathbb{E}\|\mathbf{f}^* - \mathbf{X}\widehat{\boldsymbol{\beta}}_2\|^2 \lesssim \inf_{\boldsymbol{\beta}\in B_0(k)} \|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|^2 + \sigma^2 k\log(ep/k) + \gamma_{2k}^{-2}\lambda^2.$$

Comparing the slow rate bounds in Theorem 2 and Corollary 3, we note that the additional σ^2 term in the corollary matches the expected prediction error rate for the oracle least-squares estimator, achieved in the setting where $\|\boldsymbol{\beta}^*\|_0$ is bounded above by a universal constant.

3.3 Results for the ℓ_1-regularized best-subsets estimator

There exists extensive literature [for example, 11, 31, 4, 55, 7, 17] on the prediction error bounds for the Lasso, which is an ℓ_1-regularized least-squares estimator. The following theorem focuses on the ℓ_1-regularized estimator with an additional ℓ_0 constraint. It establishes both the slow and the fast rate prediction error bounds for $\widehat{\boldsymbol{\beta}}_1$. Like the estimator $\widehat{\boldsymbol{\beta}}_1$, the presented oracle inequalities are restricted to $B_0(k)$ for each fixed value of the model size tuning parameter k. In this respect, they are not as strong as the bounds in the literature that are stated without such a restriction. In Section 3.5 we analyze a data-driven approach for selecting the optimal value of k and establish oracle inequalities in a more general form.

Theorem 3. *(A) Slow rate. If* $\lambda = 2\sigma\sqrt{2\log(2p)} + 2\sigma\sqrt{2\log(1/\delta_0)}$, *then on the event* $\mathcal{H}$,

$$\|\mathbf{f}^* - \mathbf{X}\widehat{\boldsymbol{\beta}}_1\|^2 \leq \inf_{\boldsymbol{\beta}\in B_0(k)} \Big[\|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|^2 + 2\lambda\|\boldsymbol{\beta}\|_1\Big].$$

If $\lambda \geq [8+2\sqrt{2}]\sigma\sqrt{\log(2p)}$, *then on the event* $\mathcal{F}$,

$$\|\mathbf{f}^* - \mathbf{X}\widehat{\boldsymbol{\beta}}_1\|^2 \lesssim \inf_{\boldsymbol{\beta}\in B_0(k)} \Big[\|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_1\Big] + \sigma^2\log(1/\delta_0).$$

(B) Fast rate. On the event $\mathcal{G}_{2k}$,

$$\|\mathbf{f}^* - \mathbf{X}\widehat{\boldsymbol{\beta}}_1\|^2 \lesssim \inf_{\boldsymbol{\beta}\in B_0(k)} \|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|^2 + \sigma^2 k\log(ep/k) + \gamma_{2k}^{-2}\lambda^2 k + \sigma^2\log(1/\delta_0)$$

for every $\lambda \geq 0$.

Focusing on the linear case, $\mathbf{f}^* = \mathbf{X}\boldsymbol{\beta}^*$, we make the following observations.

Remark 5. *The slow rate prediction error bound for $\widehat{\boldsymbol{\beta}}_1$ is $\sigma\sqrt{\log(ep)}\|\boldsymbol{\beta}^*\|_1$, which is better than the $\sigma^2 k^* \log(ep/k^*)$ bound for best-subsets when $\|\boldsymbol{\beta}^*\|_1/\sigma \lesssim k^* \log(ep/k^*)/\sqrt{\log(ep)}$ with a sufficiently small universal constant.*

Remark 6. *As is the case with $\widehat{\boldsymbol{\beta}}_2$, the fast prediction error rate for $\widehat{\boldsymbol{\beta}}_1$ matches the minimax rate over ℓ_0-balls. The slow rate for $\widehat{\boldsymbol{\beta}}_1$ matches the corresponding rate for the Lasso and the minimax lower bound over ℓ_1-balls derived in [51]. This rate is slightly worse than the corresponding minimax rate established in [52]. However, we note that the latter rate can be derived for $\widehat{\boldsymbol{\beta}}_1$ with an appropriate tuning of the parameter k, using the arguments in the proof of Corollary 4.1 in [52], which bounds the prediction error of a modified BIC estimator.*

Remark 7. *Similarly to the ℓ_2 case (Corollary 2), the established error rates can also apply to solutions obtained after an early termination of the MIO solver. More specifically, if the optimality gap, τ, is bounded away from one, then the approximate solution achieves the second slow error rate in Theorem 3 when $\tau \lesssim \log(1/\delta_0)/n$, and it achieves the corresponding fast error rate when $\tau \lesssim k\log(ep/k)/n$.*

Remark 8. *Similarly to Corollary 1, the fast rate part of Theorem 3 implies an estimation error bound: $\|\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}^*\| \lesssim \gamma_{2k}^{-1}\sigma\sqrt{k\log(ep/k)} + \gamma_{2k}^{-2}\lambda\sqrt{k}$.*

Remark 9. *The first slow rate bound in Theorem 3 can be potentially improved [see, for example, the discussion in 7] by replacing the approximation $\sqrt{2\log(2p)}+\sqrt{2\log(1/\delta_0)}$, used in the definition of λ, directly with the $(1-\delta_0)$ quantile of $\|\mathbf{X}^\top\boldsymbol{\epsilon}/\sigma\|_\infty$. As before, σ can be estimated by employing a preliminary regression estimator, unrestricted in terms of the model size. In practice, λ can be tuned based on a separate validation set or by cross-validation.*

The next result uses Theorems 1 and 3 to bound the expected prediction error for $\widehat{\boldsymbol{\beta}}_1$.

Corollary 4. *If $\lambda \geq [8+2\sqrt{2}]\sigma\sqrt{\log(2p)}$, then*

$$\mathbb{E}\|\mathbf{f}^* - \mathbf{X}\widehat{\boldsymbol{\beta}}_1\|^2 \lesssim \inf_{\boldsymbol{\beta}\in B_0(k)} \Big[\|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_1\Big] + \sigma^2.$$

Furthermore, for every $\lambda \geq 0$,

$$\mathbb{E}\|\mathbf{f}^* - \mathbf{X}\widehat{\boldsymbol{\beta}}_1\|^2 \lesssim \inf_{\boldsymbol{\beta}\in B_0(k)} \|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|^2 + \sigma^2 k\log(ep/k) + \gamma_{2k}^{-2}\lambda^2 k.$$

We now compare the slow rate prediction error bounds for the two proposed estimators: $\widehat{\boldsymbol{\beta}}_1$ and $\widehat{\boldsymbol{\beta}}_2$. In the case where all the non-zero coefficients of $\boldsymbol{\beta}^*$ are of the same order of magnitude, the pre-

diction error rate for $\widehat{\boldsymbol{\beta}}_2$ is superior to the one for $\widehat{\boldsymbol{\beta}}_1$, because the former replaces the $\log(ep)$ term with $\log(ep/k^*)$. Alternatively, the slow rate for $\widehat{\boldsymbol{\beta}}_1$ is better when the ratio $\|\boldsymbol{\beta}^*\|_1/\|\boldsymbol{\beta}^*\|$ is sufficiently small. The following result formalizes the last observation in the asymptotic setting.

Corollary 5. *Denote the slow prediction error rates for $\widehat{\boldsymbol{\beta}}_1$ and $\widehat{\boldsymbol{\beta}}_2$ by SR_1 and SR_2, respectively. Suppose that $k = k^*$, $\mathbf{f}^* = \mathbf{X}\boldsymbol{\beta}^*$ and*

$$\|\boldsymbol{\beta}^*\|_1/\big(\sqrt{k^*}\|\boldsymbol{\beta}^*\|\big) = o\big(\sqrt{\log(p/k^*)/\log(p)}\big)$$

as $p \to \infty$. Then, $SR_1/SR_2 \to 0$.

In the next section we complement the slow rate prediction error bounds for $\widehat{\boldsymbol{\beta}}_2$ and $\widehat{\boldsymbol{\beta}}_1$ with a corresponding lower-bound for $\widehat{\boldsymbol{\beta}}_{\ell_0}$.

3.4 Lower bounds for the best-subsets estimator

Focusing on the linear setting and comparing the slow rate prediction error bound for $\widehat{\boldsymbol{\beta}}_2$ in Corollary 1 with the one provided for $\widehat{\boldsymbol{\beta}}_{\ell_0}$ by the fast rate part of the same result, we note that the former bound is superior when $\|\boldsymbol{\beta}^*\|/\sigma \lesssim \sqrt{k\log(ep/k)}$ with a sufficiently small constant. The following novel result demonstrates that in this regime of low $\|\boldsymbol{\beta}^*\|/\sigma$ the above comparison is meaningful, because the error bound for $\widehat{\boldsymbol{\beta}}_{\ell_0}$ is tight.

Theorem 4. *Suppose that $k \in [p]$ and $\|\boldsymbol{\beta}^*\|/\sigma \lesssim \gamma_k\sqrt{k\log(ep/k)}$ with a sufficiently small universal constant. Then, there exists a positive universal constant c, such that*

$$\|\mathbf{X}\hat{\boldsymbol{\beta}}_{\ell_0} - \mathbf{X}\boldsymbol{\beta}^*\|^2 \gtrsim \sigma^2\gamma_k^2 k\log(ep/k)$$

with (high) probability of at least $1 - 2(ep/k)^{-c\gamma_k^2 k} - (ep/k)^{-k}$.

Suppose that γ_k is bounded away from zero by a positive universal constant. Note that this holds under the sparse eigenvalue condition, which is standard in the literature (see the discussion in Section 8 of [5], for example). In particular, this condition holds with high probability for a wide class of random matrices $\mathbf{X}$ with i.i.d. rows, provided $k\log(ep/k) \lesssim n$ with an appropriate universal constant [33]. Under this setting, we make the following key observations.

Remark 10. *Combining the upper-bound for $\widehat{\boldsymbol{\beta}}_{\ell_0}$ from Corollary 1 with the lower-bound from Theorem 4 yields $\|\mathbf{X}\hat{\boldsymbol{\beta}}_{\ell_0} - \mathbf{X}\boldsymbol{\beta}^*\|^2 \asymp \sigma^2 k\log(ep/k)$. Comparing this prediction error to the slow rate*

prediction error bound for $\widehat{\boldsymbol{\beta}}_2$, we conclude that

$$\|\mathbf{X}\boldsymbol{\beta}^* - \mathbf{X}\widehat{\boldsymbol{\beta}}_{\ell_0}\|^2 / \|\mathbf{X}\boldsymbol{\beta}^* - \mathbf{X}\widehat{\boldsymbol{\beta}}_2\|^2 \gtrsim \big(\sigma/\|\boldsymbol{\beta}^*\|\big)\sqrt{k\log(ep/k)} \tag{17}$$

with high probability.

Remark 11. *In the regime of interest, where $\|\boldsymbol{\beta}^*\|/\sigma \lesssim \sqrt{k\log(ep/k)}$, the ratio of prediction errors in (17) can be made arbitrarily large by decreasing $\|\boldsymbol{\beta}^*\|/\sigma$ or increasing k. Similarly, Theorem 3 implies that the prediction error for $\widehat{\boldsymbol{\beta}}_1$ is smaller than the one for $\widehat{\boldsymbol{\beta}}_{\ell_0}$ in the regime of low $\|\boldsymbol{\beta}^*\|_1/\sigma$. These observations are supported empirically, as illustrated by the left column in Figure 1, where the predictive performance of $\widehat{\boldsymbol{\beta}}_{\ell_0}$ steadily deteriorates relative to that of $\widehat{\boldsymbol{\beta}}_2$ and $\widehat{\boldsymbol{\beta}}_1$ as k increases.*

The lower-bound in Theorem 4, together with the companion upper-bound implied by Corollary 1, suggests that in the setting where $\|\boldsymbol{\beta}^*\|/\sigma$ is low, the prediction error for $\widehat{\boldsymbol{\beta}}_{\ell_0}$ could be reduced by decreasing k below k^*. Thus, decreasing the model size parameter k may have a regularizing effect on the best-subsets estimator. However, if we tune k in order to improve the predictive performance, then we lose the attractive feature of subset selection that allows the user to select the model size based on external considerations. In contrast, estimator $\widehat{\boldsymbol{\beta}}_2$ is regularized via the tuning parameter λ, for each given model size k. Moreover, the next example illustrates that, even with optimal data-dependent choice of k, best subset selection does not achieve the $\sigma\sqrt{k^*\log(p/k^*)}\|\boldsymbol{\beta}^*\|$ prediction error rate available for $\widehat{\boldsymbol{\beta}}_2$.

Example. Suppose that all pairwise correlations among the predictors are equal to a fixed universal constant $\rho \in (0,1)$. Recall the notation $k^* = \|\boldsymbol{\beta}^*\|_0$, let $k^* > 0$ and assume that each nonzero element of $\boldsymbol{\beta}^*$ is equal to $b\sigma\sqrt{\log(ep)}/k^*$ for some positive b.

Proposition 3. *Let $\delta \in (0,1]$ be a fixed universal constant. Under the setting of the Example, there exist positive universal constants b_0 and a, such that if $b \in [\delta b_0, b_0]$, then*

$$\min_{k\in\{0,1,\ldots,p\}} \|\mathbf{X}\boldsymbol{\beta}^* - \mathbf{X}\widehat{\boldsymbol{\beta}}_{\ell_0}\|^2 \gtrsim \sigma\sqrt{k^*\log(ep)}\|\boldsymbol{\beta}^*\|$$

with probability at least $1 - 2(ep)^{-a}$. Moreover, the result holds uniformly over $\boldsymbol{\beta}^*$.

We note that, under the setting of the Example and up to universal multiplicative constants, the above lower-bound matches the minimax rate on the intersection of ℓ_0 and ℓ_1 balls [52, Section 5.2]. Comparing this lower-bound with the $\sigma\sqrt{k^*\log(ep/k^*)}\|\boldsymbol{\beta}^*\|$ upper-bound in the slow rate part of Corollary 1, we conclude that in general the best-subsets estimator is not able to achieve the slow

rate of ℓ_2-regularized best-subsets estimator. We emphasize that the lower-bound in Proposition 3 holds with high probability, and is uniform over k and $\boldsymbol{\beta}^*$. In particular, even if best-subsets were able to choose an optimal k for each given sample, the prediction error rate for the resulting "oracle" estimator would still be worse than the one for $\widehat{\boldsymbol{\beta}}_2$.

The next result shows that for larger k the difference between the prediction errors for $\widehat{\boldsymbol{\beta}}_{\ell_0}$ and $\widehat{\boldsymbol{\beta}}_2$ is substantially greater than the one suggested by the uniform lower-bound in Proposition 3.

Proposition 4. *Suppose that $k \in [p]$. Under the setting of the Example, there exist positive universal constants b_0, k_0 and a, such that if either $b \leq b_0$ or $\max\{k^*, k\} \geq k_0$, then*

$$\|\mathbf{X}\hat{\boldsymbol{\beta}}_{\ell_0} - \mathbf{X}\boldsymbol{\beta}^*\|^2 \gtrsim \sigma^2 k \log(ep/k)$$

with probability at least $1 - 3(ep/k)^{-ak}$.

We now compare the prediction errors for $\widehat{\boldsymbol{\beta}}_{\ell_0}$ and $\widehat{\boldsymbol{\beta}}_2$ in the concrete case where $k = k^*$. Proposition 4 and the slow rate part of Corollary 1 imply that

$$\|\mathbf{X}\boldsymbol{\beta}^* - \mathbf{X}\widehat{\boldsymbol{\beta}}_{\ell_0}\|^2 / \|\mathbf{X}\boldsymbol{\beta}^* - \mathbf{X}\widehat{\boldsymbol{\beta}}_2\|^2 \gtrsim k^* [\log(ep/k^*)/\log(ep)]^{1/2}$$

with high probability. In particular, if we let $k^* = O(p^{1-c})$ for some positive c, then the lower-bound in the above display grows linearly in k^*.

3.5 Data-driven choice of k

In this section we study a BIC-type approach for selecting the model size k. We define

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_2^{\mathrm{B}} &= \arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\| + \mu_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_0 \\ \widehat{\boldsymbol{\beta}}_1^{\mathrm{B}} &= \arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|_1 + \mu_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_0, \end{aligned}$$

where $\lambda_{\boldsymbol{\beta}} = a\sqrt{\|\boldsymbol{\beta}\|_0 \log(ep/\|\boldsymbol{\beta}\|_0)}$ and $\mu_{\boldsymbol{\beta}} = b \log(ep/\|\boldsymbol{\beta}\|_0)$ for some nonnegative a and b. The above optimization problems are equivalent to first solving the corresponding constrained problems (16), for each k, and then identifying the optimal model size k via BIC-type penalization. The value of λ in the corresponding constrained formulation for $\widehat{\boldsymbol{\beta}}_2^{\mathrm{B}}$ is $a\sqrt{k \log(ep/k)}$.

The following result establishes general oracle inequalities for $\widehat{\boldsymbol{\beta}}_2^{\mathrm{B}}$ and $\widehat{\boldsymbol{\beta}}_1^{\mathrm{B}}$. To simplify the presentation, we focus on the expected prediction error.

Theorem 5. *There exist universal constants a_0, b_0 and c_0, such that if $a \geq a_0\sigma$ or $b \geq b_0\sigma^2$, then*

$$\mathbb{E}\|\mathbf{f}^* - \mathbf{X}\widehat{\boldsymbol{\beta}}_2^{\mathrm{B}}\|^2 \lesssim \inf_{\boldsymbol{\beta}\in\mathbb{R}^p} \Big[\|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\beta}\| + \mu_{\boldsymbol{\beta}}\|\boldsymbol{\beta}\|_0\Big] + \sigma^2;$$

and if $\lambda \geq c_0\sigma\sqrt{\log(ep)}$ or $b \geq b_0\sigma^2$, then

$$\mathbb{E}\|\mathbf{f}^* - \mathbf{X}\widehat{\boldsymbol{\beta}}_1^{\mathrm{B}}\|^2 \lesssim \inf_{\boldsymbol{\beta}\in\mathbb{R}^p} \Big[\|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_1 + \mu_{\boldsymbol{\beta}}\|\boldsymbol{\beta}\|_0\Big] + \sigma^2.$$

The next result, which focuses on the linear case for concreteness, shows that the new estimators achieve the error rates in Corollaries 3 and 4 while producing model sizes of the same order as the true model size k^*.

Corollary 6. *Let $\mathbf{f}^* = \mathbf{X}\boldsymbol{\beta}^*$ and consider the universal constants that appear in the statement of Theorem 5. If $a_0\sigma \leq a \lesssim \sigma$ and $b \asymp \big(\lambda_{\boldsymbol{\beta}^*}\|\boldsymbol{\beta}^*\| + \sigma^2\big)/\big\{[k^* \vee 1]\log(ep/[k^* \vee 1])\big\}$, then*

$$\mathbb{E}\|\mathbf{f}^* - \mathbf{X}\widehat{\boldsymbol{\beta}}_2^{\mathrm{B}}\|^2 \lesssim \sigma\sqrt{k^*\log(ep/k^*)}\|\boldsymbol{\beta}^*\| + \sigma^2 \quad \textit{and} \quad \mathbb{E}\|\widehat{\boldsymbol{\beta}}_2^{\mathrm{B}}\|_0 \lesssim k^* \vee 1.$$

If $a \lesssim \sigma^2\sqrt{k^*\log(ep/k^*)}/\|\boldsymbol{\beta}^*\|$ and $b_0\sigma^2 \leq b \lesssim \sigma^2$, then

$$\mathbb{E}\|\mathbf{f}^* - \mathbf{X}\widehat{\boldsymbol{\beta}}_2^{\mathrm{B}}\|^2 \lesssim \sigma^2 k^*\log(ep/k^*) + \sigma^2 \quad \textit{and} \quad \mathbb{E}\|\widehat{\boldsymbol{\beta}}_2^{\mathrm{B}}\|_0 \lesssim k^* \vee 1.$$

If $c_0\sigma\sqrt{\log(ep)} \leq \lambda \lesssim \sigma\sqrt{\log(ep)}$ and $b \asymp \big(\lambda\|\boldsymbol{\beta}^*\|_1 + \sigma^2\big)/\big\{[k^* \vee 1]\log(ep/[k^* \vee 1])\big\}$, then

$$\mathbb{E}\|\mathbf{f}^* - \mathbf{X}\widehat{\boldsymbol{\beta}}_1^{\mathrm{B}}\|^2 \lesssim \sigma\sqrt{\log(ep)}\|\boldsymbol{\beta}^*\|_1 + \sigma^2 \quad \textit{and} \quad \mathbb{E}\|\widehat{\boldsymbol{\beta}}_1^{\mathrm{B}}\|_0 \lesssim k^* \vee 1. \tag{18}$$

If $\lambda \lesssim \sigma^2 k^*\log(ep/k^*)/\|\boldsymbol{\beta}^*\|_1$ and $b_0\sigma^2 \leq b \lesssim \sigma^2$, then

$$\mathbb{E}\|\mathbf{f}^* - \mathbf{X}\widehat{\boldsymbol{\beta}}_1^{\mathrm{B}}\|^2 \lesssim \sigma^2 k^*\log(ep/k^*) + \sigma^2 \quad \textit{and} \quad \mathbb{E}\|\widehat{\boldsymbol{\beta}}_1^{\mathrm{B}}\|_0 \lesssim k^* \vee 1. \tag{19}$$

It is useful to compare $\widehat{\boldsymbol{\beta}}_1^{\mathrm{B}}$ to the related Lasso estimator. We first note that the slow rate in (18) also holds for the Lasso, as a consequence of the second bound in Theorem 5 when $b = 0$. The fast rate in (19) holds for the Lasso as well [5, Corollary 4.4], however, under a "strong restricted eigen value condition". In contrast, all the error bounds in this section hold without imposing any assumptions on the design beyond the usual normalization of the columns of $\mathbf{X}$. This can be viewed as a non-trivial advantage of ℓ_0-based approaches over Lasso-type methods: [62] gives examples of design matrixes for which the Lasso[9] prediction error is lower-bounded by a constant multiple of $\sqrt{n}$, which is generally much larger than the fast rate error bound in (19). Similarly, the sparsity bounds for the Lasso estimator [6] require sparse eigen value conditions, while the corresponding bounds the proposed approach hold without any additional assumptions on the design.

[9]The lower-bound holds for a wide range of coordinate-separable M-estimators, including popular nonconvex regularizers such as SCAD and MCP.

4 Related work and connections to existing estimators

The literature on penalized estimation in high-dimensional regression is extensive. Here we discuss a subset of this work that is closely related to the topic of our paper.

When $q = 2$, estimator (3) is related[10] to the elastic net estimator [63]. Similarly, when $q = 1$, a relaxation of (4) leads to the Lasso problem. However, as we demonstrate in Section 5, the operating characteristics of estimator (3) are quite different from these relaxations.

Estimator (3) bears similarities with the nonconvex approaches in [19, 29, 64, 35, 18], however, the particular form of (3) is not considered in these works. Despite apparent similarities, our work is different in terms of motivation, context and computational methods. More specifically, our primary motivation is to *regularize* the overfitting behavior of best subsets selection and obtain sparse models with good predictive power. From a computational standpoint, our MIO framework delivers a *global* solution for the corresponding optimization problem.

[64, 29] propose improvements over the elastic net by replacing the ℓ_1-penalty with more aggressive penalties (for example, adaptive Lasso and MCP). They consider the penalized formulation, different from the cardinality constrained version (3). While these works focus on improved estimation accuracy in low-noise regimes, the resulting estimators may also perform well in the high-noise settings. [18] impose both a concave penalty and the ℓ_1-penalty on $\boldsymbol{\beta}$, demonstrating theoretically that their estimator combines the predictive strength of the ℓ_1 regularization with the variable selection strength of the nonconvex regularization. [35] impose a convex combination of the ℓ_0 and the ℓ_1 penalties on $\boldsymbol{\beta}$, and study statistical properties of their estimator in the low-dimensional setting. There are differences in the computational approaches as well: [35] propose using a piecewise linear approximation to the ℓ_0-penalty for computational purposes; their numerical experiments are mostly limited to the case $p \leq 15$. [29] and [18] rely on local approximations to nonconvex optimization problems, which may potentially lead to sub-optimal local solutions.

Our approach has interesting connections with Bayesian procedures that use sparsity-inducing prior distributions for the regression coefficients – for example, the spike-and-slab priors [43, 50, 54]. In the Bernoulli-Gaussian mixture model [54], each coefficient follows a mixture distribution involving a point mass at zero and a zero mean Gaussian distribution: $\beta_j|\theta, \sigma_\beta \sim (1-\theta)\delta_0 + \theta N(0, \sigma_\beta^2)$.

[10] A convex relaxation of (4) with $q = 2$, obtained by relaxing $z_j \in \{0, 1\}$ to $z_j \in [0, 1]$, leads to a slight modification of the elastic net optimization problem, where the squared-ℓ_2-penalty is replaced by the ℓ_2-penalty.

One may represent β_j as a product of two independent random variables: $\beta_j = \gamma_j \alpha_j$, where $\gamma_j|\theta \sim \text{Bernoulli}(\theta)$, $\alpha_j|\sigma_\beta \sim N(0, \sigma_\beta^2)$. The corresponding MAP estimator then minimizes

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda_1 \|\boldsymbol{\alpha}\|^2 + \lambda_2 \|\boldsymbol{\gamma}\|_0$$

with respect to variables $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\alpha})$, for a suitable choice of parameters λ_1, λ_2. The above problem is an ℓ_0-penalized version of Problem (2) with $q = 2$, in which the squared ℓ_2-penalty replaces the ℓ_2-penalty. Such problems are known to pose computational challenges in large-scale settings. [50] study a special case of this problem with $\lambda_1 \approx 0$, which corresponds to the high-SNR regime, and consider a number of approximate algorithms (for example, proximal gradient [9] and single-best-replacement [54]) for the Lagrangian version of Problem (1). Another possibility is to use the Bernoulli-Laplace prior for the regression coefficients [2, 50] – the corresponding MAP formulation leads to an ℓ_0-penalized form of Problem (2) with $q = 1$. Our proposed algorithms may potentially be used to obtain (near-optimal or optimal) solutions for both of these problems.

Another popular approach is to employ continuous spike-and-slab priors, such as a mixture of two Laplace distributions. When $q = 1$, the penalized modification of estimator (3) corresponds to the limiting case in which the spike distribution is a point mass. Importantly, our estimator (3) is constrained rather than penalized, providing a direct control over the sparsity level. When the mixture weight in the aforementioned Laplace mixture follows its own prior distribution, the resulting approach is the powerful spike-and-slab Lasso procedure of [53]. Some other state-of-the-art Bayesian shrinkage methods include the horseshoe regression [15] and the empirical Bayes method of [38]. These methods are known to improve on the predictive performance of the global shrinkage approaches such as ridge regression [10, 37]. In particular, [37] propose a Monte-Carlo scheme to approximate the predictive density, allowing for uncertainty quantification. From an algorithmic standpoint, the main difference between our approach and the related Bayesian methods for computing MAP estimators is our use of mixed integer programming. Furthermore, our theoretical analysis focuses on the low-SNR regime. To the best of our knowledge, the earlier works discussed above do not consider the low-SNR regime in their theoretical development.

The topic of this paper is closely related to the interesting recent work of [24], where the authors also observe that in the low-SNR regimes the Lasso leads to better predictive models than best subset selection, while the reverse is true in the high-SNR regimes. As a compromise between the

two approaches, [24] propose a variant[11] of relaxed Lasso [41]. Interestingly, the original form of the relaxed Lasso estimator can be interpreted as a feasible solution to Problem (3), with $q = 1$, for a suitable choice of tuning parameters k and λ. The key advantages of our approach are as follows. Unlike relaxed Lasso, estimator (3) is given by a transparent optimization formulation with an explicit control on the support size. We conduct an extensive theoretical analysis of the predictive properties of estimator (3), including its superior performance relative to best-subsets in high-noise regimes. To our knowledge, similar results are not available for the relaxed Lasso estimator.

After an earlier version of this paper became publicly available, some interesting follow-up work has been conducted with the focus on the computational aspects of the regularized best-subset estimators [for example, 25, 3, 26].

5 Experiments

We explore the properties of our estimator empirically on synthetic datasets with varying values of n, p, SNR and correlations among the predictors, as well as on several real datasets. An implementation of the algorithms we propose in this paper is available on github[12].

5.1 Synthetic Datasets

We generate the rows of the model matrix $\mathbf{X}$ as n independent realizations from a p-dimensional multivariate Gaussian distribution with mean zero and covariance matrix $\mathbf{\Sigma} = (\sigma_{jk})$. We standardize the columns of $\mathbf{X}$ to have zero mean and unit ℓ_2-norm, and generate $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}$ with $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$ and $\boldsymbol{\beta}^* \in \mathbb{R}^p$. Recall that we define $\text{SNR} = \|\mathbf{X}\boldsymbol{\beta}^*\|_2^2 / \|\boldsymbol{\epsilon}\|_2^2$ and let $k^* = \|\boldsymbol{\beta}^*\|_0$ denote the true number of nonzeros. We consider the following examples:

Example 1. $\sigma_{jk} = \rho^{|j-k|}$ (with the convention $0^0 = 1$), $\beta_j^* = 1$ for $k^* = 7$ equispaced values in $[p]$ and $\beta_j^* = 0$ otherwise.

Example 2. $\sigma_{jk} = \rho + (1-\rho)I\{j = k\}$, $\beta_j^* = 1$ for $j \leq k^* = 7$ and $\beta_j^* = 0$ otherwise.

In the above examples all the nonzero coefficients in $\boldsymbol{\beta}^*$ have the same magnitude. We focus on this setting to get a clear understanding of how our proposed estimator regulates the overfitting

[11] This is given by a convex combination of the Lasso estimator and its polished version (obtained by performing a least squares fit on the Lasso support).

[12] Link to repository: `https://github.com/antoine-dedieu/subset_selection_with_shrinkage`

behavior of best-subsets and compares with estimators such as ridge regression and the Lasso, as the SNR is varied. In our simulations, we also vary the values of ρ, n and p.

We conduct a comparison across the following methods:

(L1+L0) Estimator (3) with $q = 1$. The 2D grid of tuning parameters has λ taking values in a geometrically spaced sequence $\{\lambda_i\}_1^{100}$, with $\lambda_1 = \|\mathbf{X}^\top \mathbf{y}\|_\infty$ and $\lambda_{100} \sim 10^{-4}\lambda_1$, while k takes values in $\{0, \ldots, 15\}$.

(L2+L0) Estimator (3) with $q = 2$. The 2D grid was similar to the above, with $\lambda_1 = \|\mathbf{X}^\top \mathbf{y}\|_2$, which ensures a zero solution.

(L0) Best-subsets estimator (1) with $k \in \{0, \ldots, 15\}$.

(L1) The Lasso estimator given by Problem (2) with $q = 1$ on a grid of 100 values of λ.

(L1P) Polished version of the Lasso estimator, computed as the least-squares estimator on the support of every L1 solution.

(L2) Ridge regression estimator given by Problem (2) with $q = 2$ on a grid of 100 values of λ.

(L1+L2) Elastic net estimator [63]. For each value of parameter λ, we consider a sequence of 20 values $\alpha \in [0.05, 0.95]$ for weighting the ℓ_1 and ℓ_2^2 penalties.

The estimators in (3) are computed via 3 rounds of Algorithm 1 (Neighborhood Continuation) with stochastic local search, as described in Section 2.4. Let $\{\hat{\boldsymbol{\beta}}(\lambda, k)\}$ denote the corresponding 2-dimensional family of solutions. The discrete first order algorithm (DFO) is run until reaching the convergence threshold of $\tau = 10^{-3}$ or a maximum of 1000 iterations, whichever is earlier. Once the family $\{\hat{\boldsymbol{\beta}}(\lambda, k)\}$ is obtained, the best pair $(\hat{\lambda}, \hat{k})$ is chosen on a held-out validation set as discussed below. For this choice of $(\hat{\lambda}, \hat{k})$, we solve the MIO formulation (4) with a time-limit of 30 minutes[13] – the resultant solutions are referred to as L1+L0 or L2+L0. We obtain the L0 solution in a similar fashion, using $\hat{\boldsymbol{\beta}}(\lambda_N, k)$ from Problem (3) with $q = 1$ to warm-start the DFO. Methods L1, L1P, L2 and L1+L2 are computed using Python's `scikit-learn` suite of algorithms.

Selecting the tuning parameters. For each of the above methods, we pick the estimator that minimizes the least squares criterion on a validation set simulated as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}$, with the fixed $\mathbf{X}$ and an independent realization of $\boldsymbol{\epsilon}$, with the same SNR. For each selected estimator we

[13] We use a Python interface to the Gurobi solver for our experiments.

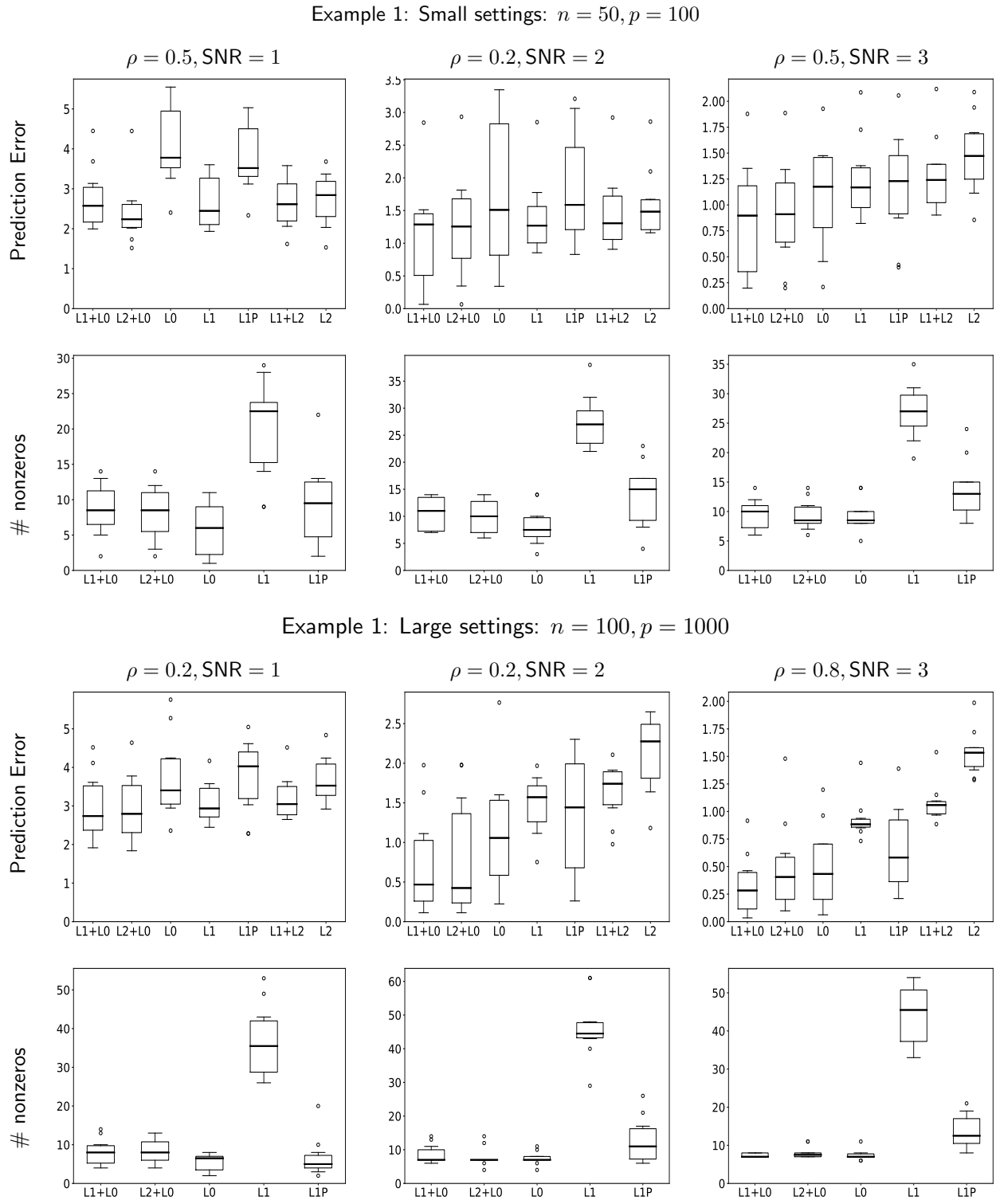


Figure 2: Example 1 simulations for different values of n, p, ρ, and SNR. Prediction error refers to the best predictive models obtained after tuning on a separate validation set. # nonzeros refers to the corresponding number of nonzero coefficients. For low SNR values, L0 led to poor predictive models and was outperformed by L1 and L2. Overall, the best predictive models were produced by L1+L0/L2+L0 – in some instances they were comparable to the best L1/L2 models, but much sparser.

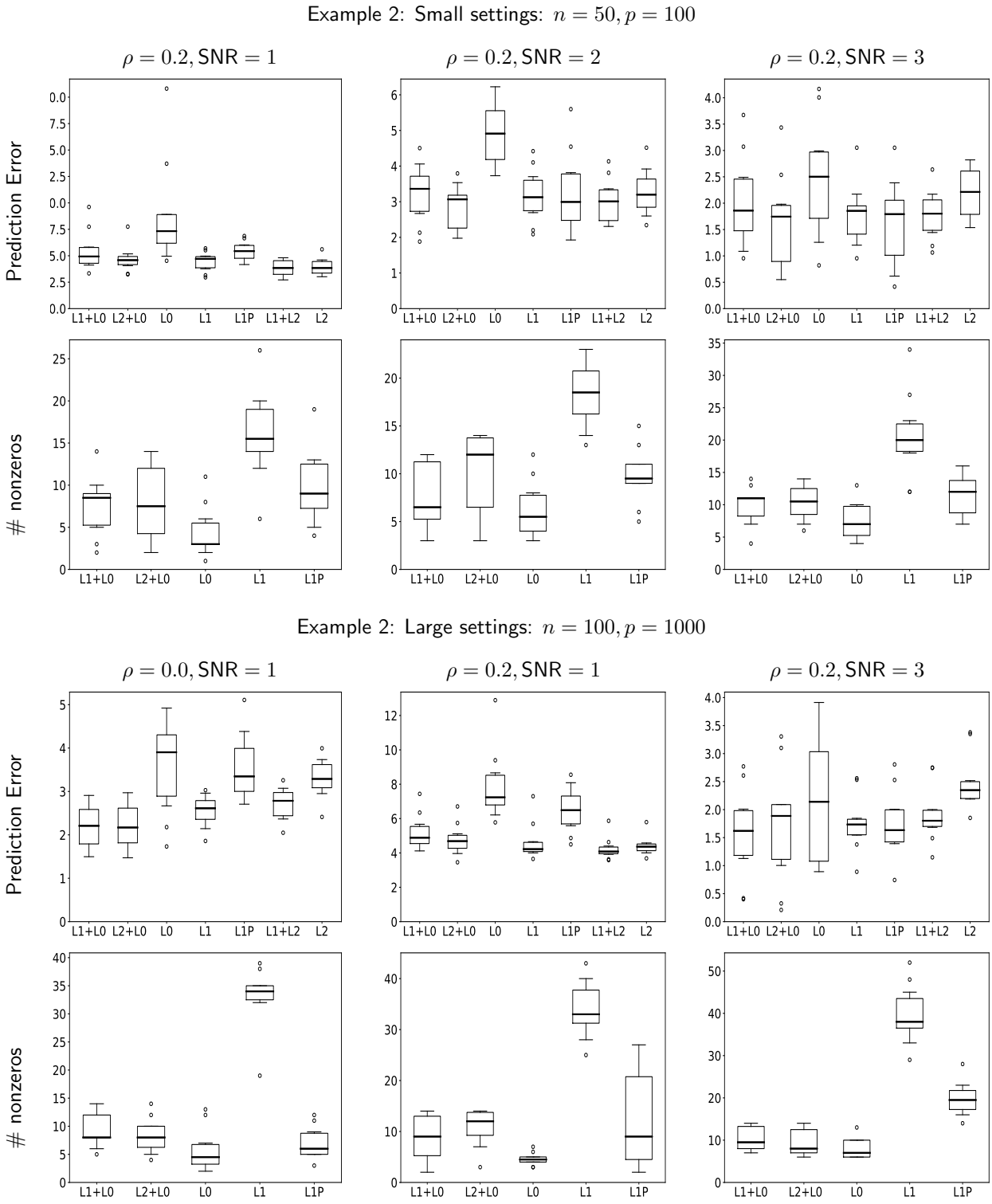


Figure 3: Experimental results for Example 2. The results are qualitatively similar to Figure 2 – however, this example is "harder" than Example 1 due to the increased correlation among the features – a larger nominal value of SNR is required before L0 matches the performance of L1+L0/L2+L0. The L1+L0/L2+L0 methods performed the best in terms of obtaining a good predictive model that is also sparse – the model sizes were larger than k^* but smaller than those available from the best L1 models.

compute the prediction error, $\|\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}^*\|_2^2/n$, and the associated number of nonzero regression coefficients. Figures 2, 3 and 4 summarize the results via box plots, in which the boxes extend from the lower to the upper quartile of the data with a line at the median, to aggregate the results over the ten independent simulations. We do not display the sparsity levels of L1+L2 and L2, as these methods are considerably denser than L1, which, in turn, produces the densest solutions among the remaining methods in the examples we consider.

Summary of observations. We summarize our general observations below:

- When the noise level is high (SNR=1), L0 performs poorly in terms of prediction accuracy. To mitigate its overfitting behavior, L0 attempts to regularize by selecting very sparse models – the best predictive model for L0 has fewer nonzeros than $\boldsymbol{\beta}^*$. In this setting, methods L1 and L2 work better than L0 in terms of the prediction accuracy. However, the estimated models are rather dense. The polished version of the Lasso, L1P, selects a model that is sparser than the Lasso but suffers in prediction accuracy.

  The two new methods, L1+L0 and L2+L0, display the best prediction accuracy overall. They *fix* the overfitting behavior of L0 via the additional shrinkage. This observation agrees with the theoretical results and the discussion in Sections 3.2-3.4. The best predictive models available from L1+L0/L2+L0 are similar in performance to the best predictive models available via L1 and L2, however, the new methods lead to estimators that are significantly sparser. The L0 models are sparser than those for L1+L0 and L2+L0, however, L0 suffers in terms of the prediction accuracy. In summary, the new L1+L0/L2+L0 methods significantly improve upon the predictive performance of L0 at the cost of marginally decreasing the model sparsity.

- As SNR increases, L1+L0 and L2+L0 become more similar to L0, in terms of both sparsity and the prediction accuracy. Additional shrinkage marginally helps the prediction accuracy, and the model sparsity becomes comparable to that of L0, with the model size concentrating around $\|\boldsymbol{\beta}^*\|_0$. This observation is consistent with the results in the fast rate parts of Theorems 2 and 3. L1 performs better than both L1+L2 and L2; it also benefits from polishing – L1P gets closer to L0 in terms of the prediction accuracy but selects a denser model.

In the Supplementary Material, we discuss additional experiments corresponding to the challenging ultra-high dimensional setting [57] with $k^* \log(p/k^*) > n/2$. These experiments provide further support for the observations listed above.

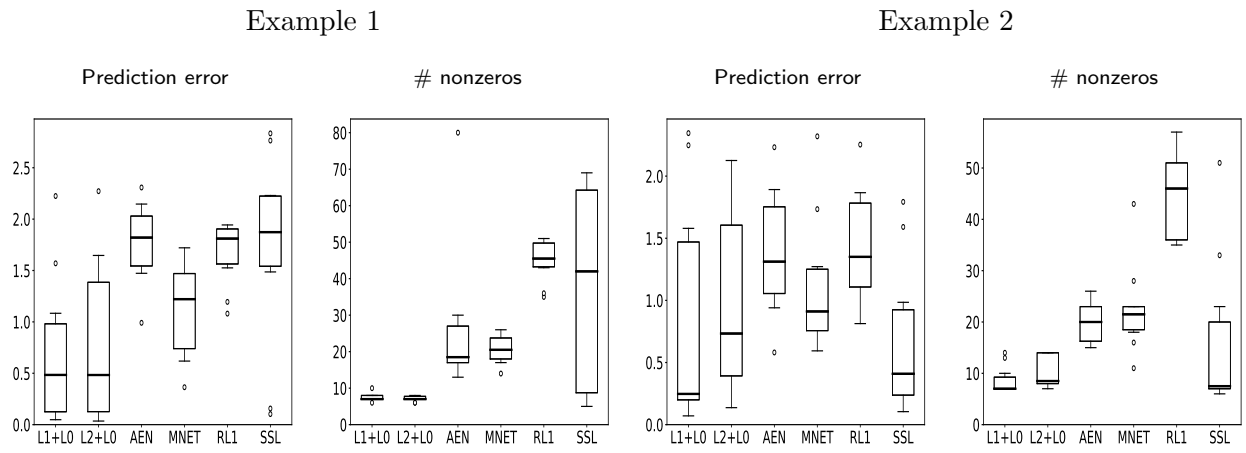


Figure 4: Experimental results for the proposed methods, L1+L0 and L2+L0, as well as adaptive elastic net (AEN), Mnet, relaxed Lasso (RL1), and spike-and-slab Lasso (SSL) methods (as described in the text). Here, $\rho = 0.2$, SNR $= 2$ for Example 1 and $\rho = 0.1$, SNR $= 3$ for Example 2; $n = 100$, $p = 1000$ in both settings. Overall, our proposed approach performed favorably in terms of both the model sparsity and the prediction accuracy.

Comparisons with adaptive elastic net (AEN), Mnet, relaxed Lasso and spike-and-slab Lasso (SSL). We present simulation results that compare our proposal with methods Mnet [29], AEN [64], relaxed Lasso [24], and SSL [53] . Mnet and AEN reduce the estimation error of elastic net, and encourage greater sparsity, by using a nonconvex penalty on $\boldsymbol{\beta}$ instead of the usual ℓ_1-norm. The proposed estimator with $q = 2$ is a natural alternative to Mnet and AEN in the regimes where these methods are found to be useful – however, our motivation for estimator (3) is different. Empirically, we observe important differences in the statistical performance of Mnet, AEN and our approach. These differences are likely a consequence of (a) the optimization algorithms[14] and (b) the exact forms of the estimators, including the choice of the penalty function.

Figure 4 compares the methods on the data generated as per Examples 1 and 2, with $n = 100$ and $p = 1000$. For AEN, we used R package `gcdnet` with weights chosen based on Example 1 in [64]. For Mnet, we used R package `ncvreg`, with the MCP penalty and ridge regularization. For the relaxed Lasso, we implemented the code in [24]; and for SSL, we used R package `SSLASSO`. For AEN, Mnet, and relaxed Lasso, we used the same number of tuning parameters as for our proposed

[14] [29] use a coordinate descent method directly on the ℓ_2^2+MCP penalized problem; [64] work with the ℓ_2^2 + adaptive Lasso regularized least squares, which is a convex problem.

methods[15]. As before, the tuning parameters were selected based on a held-out validation set. For SSL, we used the default settings of R package `SSLASSO` (with the exception of the variance parameter, set to be unknown). In summary, estimator (3) produced models with significantly fewer nonzeros and overall better predictive performance.

In the Supplementary Material, we compare estimator (3) to two additional state-of-the-art Bayesian shrinkage methods – the horseshoe regression [15] and the empirical Bayes method of [38], which were outperformed in our experiments by the spike-and-slab Lasso approach considered in Figure 4.

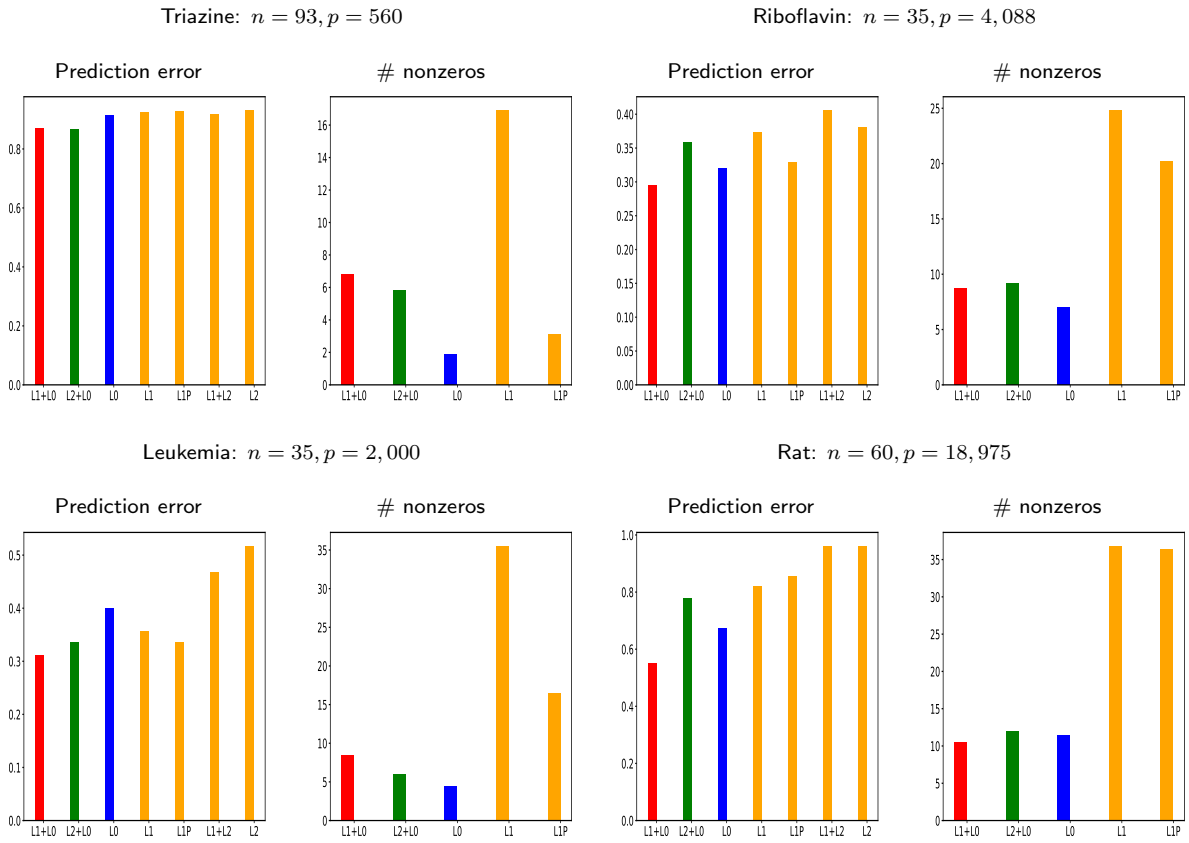


Figure 5: Performance of the methods on four real datasets. We observe that pure L0 tended to underfit by selecting models that are overly sparse. L1+L0/L2+L0 worked well both in terms of prediction and in terms of sparsity, when compared to the best available L1 models. L1 led to models with good predictive accuracy, but at the cost of a significant increase in density.

[15]For Mnet, we used 15 values for the tuning parameter that combines the ridge and MCP penalties, and 100 values for the MCP penalty weight. We made a similar choice for AEN. For the relaxed Lasso, we used 15 values for the weight in the convex combination, and 100 tuning parameter values for the Lasso.

5.2 Real Datasets

We now compare the performance of the methods on real datasets, as described below.

Triazine dataset is taken from the `libsmv` website (`https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/regression/triazines`). It contains 186 observations and 60 features, to which we added 500 features generated as Gaussian noise.

Riboflavin dataset, taken from R package `hdi`, pertains to riboflavin production for $n = 71$ observations of Bacillus subtilis. Each observation contains $p = 4088$ gene expression features.

Leukemia dataset, available at `http://cilab.ujn.edu.cn/datasets.htm`, is a classification dataset, with 72 observations and 7129 features. We keep the top 2000 features based on correlation screening and create a semi-synthetic response using $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}$ with $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$, where we set SNR $= 4$ and let $\beta_j^* \in \{0, 1\}$ with 10 randomly chosen coefficients set to 1.

Rat dataset. Using the same processing steps as [60], we analyze the RNA from the eyes of 120 twelve-week old male rats by considering $18,975$ probes expressed in the eye tissue. We thank Dr. Haolei Weng for providing the microarray dataset and the preprocessing code.

For each example, we standardize the features and the response. We randomly split each dataset into new training and test sets, compute all the estimators and, for each method, keep the estimator with the best test accuracy. Figure 5 displays the results averaged over 10 random splits.

Acknowledgements

We thank the anonymous referees for their constructive comments that helped us improve the paper.

References

[1] E. H. Aarts and J. K. Lenstra. *Local search in combinatorial optimization*. Princeton University Press, 1997.

[2] A. Amini, U. S. Kamilov, and M. Unser. The analog formulation of sparsity implies infinite divisibility and rules out bernoulli-gaussian priors. In *2012 IEEE Information Theory Workshop*, pages 682–686. Ieee, 2012.

[3] A. Atamturk and A. Gomez. Rank-one convexification for sparse regression. *arXiv preprint arXiv:1901.10334*, 2019.

[4] P. L. Bartlett, S. Mendelson, and J. Neeman. L1-regularized linear regression: persistence and oracle inequalities. *Probability theory and related fields*, 154(1-2):193–224, 2012.

[5] P. C. Bellec, G. Lecué, and A. B. Tsybakov. Slope meets lasso: improved oracle bounds and optimality. *The Annals of Statistics*, 46(6B):3603–3642, 2018.

[6] A. Belloni and V. Chernozhukov. Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547, 2013.

[7] A. Belloni, V. Chernozhukov, and L. Wang. Pivotal estimation via square-root lasso in nonparametric regression. *The Annals of Statistics*, 42(2):757–788, 2014.

[8] D. Bertsimas and R. Weismantel. *Optimization over integers*. Dynamic Ideas Belmont, 2005.

[9] D. Bertsimas, A. King, and R. Mazumder. Best subset selection via a modern optimization lens. *Annals of Statistics*, 44(2):813–852, 2016.

[10] A. Bhadra, J. Datta, Y. Li, N. G. Polson, and B. Willard. Prediction risk for the horseshoe regression. *The Journal of Machine Learning Research*, 20(1):2882–2920, 2019.

[11] P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37:1705–1732, 2009.

[12] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, 2004.

[13] L. Breiman. Heuristics of instability and stabilization in model selection. *The annals of statistics*, 24 (6):2350–2383, 1996.

[14] P. Bühlmann and S. van-de-Geer. *Statistics for high-dimensional data*. Springer, 2011.

[15] C. M. Carvalho, N. G. Polson, and J. G. Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480, 2010.

[16] L. Comminges, A. S. Dalalyan, et al. Tight conditions for consistency of variable selection in the context of high dimensionality. *The Annals of Statistics*, 40(5):2667–2696, 2012.

[17] A. S. Dalalyan, M. Hebiri, and J. Lederer. On the prediction performance of the lasso. *Bernoulli*, 23 (1):552–581, 2017.

[18] Y. Fan and J. Lv. Asymptotic properties for combined L1 and concave regularization. *Biometrika*, 101 (1):57–70, 2013.

[19] I. Frank and J. Friedman. A statistical view of some chemometrics regression tools (with discussion). *Technometrics*, 35(2):109–148, 1993.

[20] D. Gamarnik and I. Zadik. High dimensional regression with binary coefficients. estimating squared error and a phase transtition. In *Conference on Learning Theory*, pages 948–953, 2017.

[21] E. Greenshtein and Y. Ritov. Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli*, 10:971–988, 2004.

[22] E. Greenshtein. Best subset selection, persistence in high-dimensional statistical learning and optimization under ℓ_1 constraint. *The Annals of Statistics*, 34(5):2367–2386, 2006.

[23] T. Hastie, R. Mazumder, J. D. Lee, and R. Zadeh. Matrix completion and low-rank svd via fast alternating least squares. *Journal of Machine Learning Research*, 16:3367–3402, 2015.

[24] T. Hastie, R. Tibshirani, and R. Tibshirani. Best subset, forward stepwise or lasso? Analysis and recommendations based on extensive comparisons. *Statistical Science*, 35(4):579–592, 2020.

[25] H. Hazimeh and R. Mazumder. Fast best subset selection: Coordinate descent and local combinatorial optimization algorithms. *arXiv preprint arXiv:1803.01454*, 2018.

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

[27] H. Hazimeh, R. Mazumder, and P. Radchenko. Grouped variable selection with discrete optimization: Computational and statistical perspectives. *arXiv preprint arXiv:2104.07084*, 2021.

[28] A. E. Hoerl and R. Kennard. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67, 1970.

[29] J. Huang, P. Breheny, S. Lee, S. Ma, and C. Zhang. The mnet method for variable selection. *Statistica Sinica*, 26:903?923, 2016.

[30] W. James and C. Stein. Estimation with quadratic loss. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 361–379, 1961.

[31] V. Koltchinskii, K. Lounici, and A. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329, 2011.

[32] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *IEEE Computer*, 42(8), 2009.

[33] G. Lecué and S. Mendelson. Sparse recovery under weak moment assumptions. *Journal of the European Mathematical Society*, 19(3):881–904, 2017.

[34] J. T. Linderoth and A. Lodi. MILP software. *Wiley encyclopedia of operations research and management science*, 2010.

[35] Y. Liu and Y. Wu. Variable selection via a combination of the l0 and l1 penalties. *Journal of Computational and Graphical Statistics*, 16(4):782–798, 2007.

[36] K. Lounici, M. Pontil, A. Tsybakov, and S. Geer. Oracle inequalities and optimal inference under group sparsity. *The Annals of Statistics*, 39(4):2164–2204, 2011.

[37] R. Martin and Y. Tang. Empirical priors for prediction in sparse high-dimensional linear regression. *Journal of Machine Learning Research*, 21(144):1–30, 2020.

[38] R. Martin, R. Mess, and S. G. Walker. Empirical bayes posterior concentration in sparse high-dimensional linear models. *Bernoulli*, 23(3):1822–1847, 2017.

[39] P. Massart. *Concentration inequalities and model selection*, volume 6. Springer, 2007.

[40] R. Mazumder and P. Radchenko. The Discrete Dantzig Selector: Estimating sparse linear models via mixed integer linear optimization. *IEEE Transactions on Information Theory*, 63 (5):3053 – 3075, 2017.

[41] N. Meinshausen. Relaxed lasso. *Computational Statistics & Data Analysis*, 52(1):374–393, 2007.

[42] A. Miller. *Subset selection in regression*. CRC Press Washington, 2002.

[43] T. J. Mitchell and J. J. Beauchamp. Bayesian variable selection in linear regression. *Journal of the american statistical association*, 83(404):1023–1032, 1988.

[44] N. Mladenović and P. Hansen. Variable neighborhood search. *Computers & operations research*, 24(11): 1097–1100, 1997.

[45] B. Natarajan. Sparse approximate solutions to linear systems. *SIAM journal on computing*, 24(2): 227–234, 1995.

[46] G. L. Nemhauser and L. A. Wolsey. Integer programming and combinatorial optimization. *Wiley, Chichester. GL Nemhauser, MWP Savelsbergh, GS Sigismondi (1992). Constraint Classification for Mixed Integer Programming Formulations. COAL Bulletin*, 20:8–12, 1988.

[47] Y. Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140 (1):125–161, 2013.

[48] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Kluwer, Norwell, 2004.

[49] G. Pisier. Remarques sur un résultat non publié de b. maurey. *Séminaire Analyse fonctionnelle (dit" Maurey-Schwartz")*, pages 1–12, 1980.

[50] N. G. Polson and L. Sun. Bayesian ℓ_0-regularized least squares. *Applied Stochastic Models in Business and Industry*, 35(3):717–731, 2019.

[51] G. Raskutti, M. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over-balls. *Information Theory, IEEE Transactions on*, 57(10):6976–6994, 2011.

[52] P. Rigollet and A. Tsybakov. Exponential screening and optimal rates of sparse estimation. *The Annals of Statistics*, 39(2):731–771, 2011.

[53] V. Rocková and E. I. George. The spike-and-slab lasso. *Journal of the American Statistical Association*, 113(521):431–444, 2018.

[54] C. Soussen, J. Idier, D. Brie, and J. Duan. From bernoulli–gaussian deconvolution to sparse signal restoration. *IEEE Transactions on Signal Processing*, 59(10):4572–4584, 2011.

[55] T. Sun and C.-H. Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 2012.

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

[57] N. Verzelen. Minimax risks for sparse regressions: Ultra-high dimensional phenomenons. *Electronic Journal of Statistics*, 6:38–90, 2012.

[58] J. P. Vielma, I. Dunning, J. Huchette, and M. Lubin. Extended formulations in mixed integer conic quadratic programming. *Mathematical Programming Computation*, pages 1–50, 2016.

[59] M. J. Wainwright. Sharp thresholds for high-dimensional and noisy recovery of sparsity using l1-constrained quadratic programming. *IEEE Transactions on Information Theory*, 2009.

[60] H. Weng, Y. Feng, and X. Qiao. Regularization after retention in ultrahigh dimensional linear regression models. *arXiv preprint arXiv:1311.5625*, 2013.

[61] C.-H. Zhang and T. Zhang. A general theory of concave regularization for high-dimensional sparse estimation problems. *Statistical Science*, 27(4):576–593, 2012.

[62] Y. Zhang, M. J. Wainwright, and M. I. Jordan. Optimal prediction for sparse linear models? Lower bounds for coordinate-separable M-estimators. *Electronic Journal of Statistics*, 11(1):752–799, 2017.

[63] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B.*, 67(2):301–320, 2005.

[64] H. Zou and H. H. Zhang. On the adaptive elastic-net with a diverging number of parameters. *Annals of statistics*, 37(4):1733, 2009.

Supplementary Material for "Subset Selection with Shrinkage: Sparse Linear Modeling when the SNR is Low"

A Computational details

A.1 Proof of Proposition 2

(a) It follows from (9) that for any $\boldsymbol{\beta}$ satisfying $\|\boldsymbol{\beta}\|_0 \leq k$:

$$\begin{aligned}
F(\boldsymbol{\beta}) &= Q_L(\boldsymbol{\beta}, \boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_q \\
&\geq \inf_{\|\boldsymbol{\eta}\|_0 \leq k} \; (Q_L(\boldsymbol{\eta}, \boldsymbol{\beta}) + \lambda \|\boldsymbol{\eta}\|_q) \\
&= \inf_{\|\boldsymbol{\eta}\|_0 \leq k} \; \left(\frac{L}{2} \|\boldsymbol{\eta} - \boldsymbol{\beta}\|_2^2 + \langle \nabla f(\boldsymbol{\beta}), \boldsymbol{\eta} - \boldsymbol{\beta} \rangle + f(\boldsymbol{\beta}) + \lambda \|\boldsymbol{\eta}\|_q \right) \\
&= \inf_{\|\boldsymbol{\eta}\|_0 \leq k} \; \left(\frac{L}{2} \left\| \boldsymbol{\eta} - \left(\boldsymbol{\beta} - \frac{1}{L} \nabla f(\boldsymbol{\beta}) \right) \right\|_2^2 - \frac{1}{2L} \|\nabla f(\boldsymbol{\beta})\|_2^2 + f(\boldsymbol{\beta}) + \lambda \|\boldsymbol{\eta}\|_q \right) \qquad (20) \\
&= \left(\frac{L}{2} \left\| \widehat{\boldsymbol{\eta}} - \left(\boldsymbol{\beta} - \frac{1}{L} \nabla f(\boldsymbol{\beta}) \right) \right\|_2^2 - \frac{1}{2L} \|\nabla f(\boldsymbol{\beta})\|_2^2 + f(\boldsymbol{\beta}) \right) + \lambda \|\hat{\boldsymbol{\eta}}\|_q. \qquad (21)
\end{aligned}$$

Note that in (21) above we use the notation $\hat{\boldsymbol{\eta}}$ to denote a minimizer of (20). We now follow the proof in Proposition 6 in [9] to arrive at:

$$F(\boldsymbol{\beta}) \geq \frac{L - L_0}{2} \|\widehat{\boldsymbol{\eta}} - \boldsymbol{\beta}\|_2^2 + F(\widehat{\boldsymbol{\eta}}). \qquad (22)$$

In particular, using $\hat{\boldsymbol{\eta}} = \boldsymbol{\beta}^{(m+1)}$, $\boldsymbol{\beta} = \boldsymbol{\beta}^{(m)}$ and $L \geq L_0$, we see that the sequence $F(\boldsymbol{\beta}^{(m)})$ is decreasing. Because $F(\boldsymbol{\beta}) \geq 0$, we observe that the sequence $F(\boldsymbol{\beta}^{(m)})$ converges to some $F^* \geq 0$.

(b) Summing inequalities (22) for $1 \leq m \leq M$, we obtain

$$\sum_{m=1}^{M} \left(F(\boldsymbol{\beta}^{(m)}) - F(\boldsymbol{\beta}^{(m+1)}) \right) \geq \frac{L - L_0}{2} \sum_{m=1}^{M} \|\boldsymbol{\beta}^{(m+1)} - \boldsymbol{\beta}^{(m)}\|_2^2, \qquad (23)$$

leading to

$$F(\boldsymbol{\beta}^{(1)}) - F(\boldsymbol{\beta}^{(M+1)}) \geq \frac{M(L - L_0)}{2} \min_{m=1,\ldots,M} \|\boldsymbol{\beta}^{(m+1)} - \boldsymbol{\beta}^{(m)}\|_2^2.$$

Because the decreasing sequence $F(\boldsymbol{\beta}^{(m)})$ converges to $F(\boldsymbol{\beta}^*) = F^*$, say, we arrive at the conclusion in Part (b).

A.2 Stronger formulations: adding implied inequalities

We use the following notation for the model matrix: $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_p]$. We consider a structured version of Problem (5) with additional implied inequalities (cuts) for improved lower bounds:

$$
\begin{aligned}
\text{minimize} \quad & \frac{u}{2} + \lambda v \\
\text{s.t.} \quad & \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \leq u && \text{(24a)} \\
& \|\boldsymbol{\beta}\|_q \leq v && \text{(24b)} \\
& -\mathcal{M}_j z_j \leq \beta_j \leq \mathcal{M}_j z_j, j \in [p] \\
& z_j \in \{0,1\}, j \in [p] \\
& \sum_j z_j = k \\
& -\mathcal{M}_i \leq \beta_i \leq \mathcal{M}_i, i \in [p] && \text{(24c)} \\
& -\bar{\mathcal{M}}_i^- \leq \langle \mathbf{x}_i, \boldsymbol{\beta} \rangle \leq \bar{\mathcal{M}}_i^+, i \in [n] && \text{(24d)} \\
& \|\boldsymbol{\beta}\|_1 \leq \mathcal{M}_{\ell_1}, && \text{(24e)}
\end{aligned}
$$

where (a) $\mathcal{M}_i, i \in [p]$ denote bounds on β_i's via constraint (24c); (b) $-\bar{\mathcal{M}}_i^-, \bar{\mathcal{M}}_i^+$ denote bounds on the predicted values $\langle \mathbf{x}_i, \boldsymbol{\beta} \rangle$ for $i \in [n]$ via constraint (24d); (c) $\mathcal{M}_{\ell_1}$, in constraint (24e), denotes an upper bound on the ℓ_1-norm of the regression coefficients $\|\boldsymbol{\beta}\|_1$.

The additional cuts in Problem (24) help the progress of the MIO solver – the implied inequalities rule out several fractional solutions, thereby helping in obtaining superior lower bounds within a fixed computational budget. The caveat, however, is that the resulting formulation has additional variables – hence more work needs to be done within every node of the branch-and-bound tree. Section A.3 presents ways to compute these bounds – Section A.3.1 describes ways to compute them via convex optimization – these are bounds implied by an optimal solution to Problem (3). Section A.3.2 describes ways to compute these bounds based on good heuristic solutions.

A.3 Computing problem specific parameters

A.3.1 Computing parameters via convex optimization

Formulation (4) involves a BigM value $\mathcal{M}$ – tighter formulations can be obtained by using variable dependent BigM values for the β_i:

$$
-\mathcal{M}_i z_i \leq \beta_i \leq \mathcal{M}_i z_i, \quad i \in [p].
$$

In addition, implied constraints (or bounds) on $\langle \mathbf{x}_i, \boldsymbol{\beta} \rangle$'s can also be added:

$$-\bar{\mathcal{M}}_i \leq \langle \mathbf{x}_i, \boldsymbol{\beta} \rangle \leq \bar{\mathcal{M}}_i, \quad i \in [n].$$

We discuss how to compute these from data using convex optimization. Note that, because $\boldsymbol{\beta}$ is k-sparse, we have $|\langle \mathbf{x}_i, \boldsymbol{\beta} \rangle| \leq \mathcal{M} \|\mathbf{x}_i\|_{k,1}$, where for a vector $\mathbf{a} \in \mathbb{R}^p$ the quantity $\|\mathbf{a}\|_{k,1}$ denotes the ℓ_1-norm of the k-largest (in absolute value) entries of $\mathbf{a}$. We can set $\bar{\mathcal{M}}_i \leq \mathcal{M} \|\mathbf{x}_i\|_{k,1}$. Note also that $\|\boldsymbol{\beta}\|_1 \leq \mathcal{M}k := \mathcal{M}_{\ell_1}$. We now upper bound each coefficient β_i by solving the quadratic optimization problems:

$$\begin{array}{lll} \mathcal{M}_i^+ = \max & \beta_i & \\ \text{s.t.} & \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_q \leq \text{UB} & \\ & \|\boldsymbol{\beta}\|_\infty \leq \mathcal{M} & \\ & \|\boldsymbol{\beta}\|_1 \leq \mathcal{M}_{\ell_1} & \\ & -\bar{\mathcal{M}}_i^- \leq \langle \mathbf{x}_i, \boldsymbol{\beta} \rangle \leq \bar{\mathcal{M}}_i^+, i \in [n] & \end{array} \qquad \begin{array}{lll} \mathcal{M}_i^- = \max & -\beta_i & \\ \text{s.t.} & \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_q \leq \text{UB} & \\ & \|\boldsymbol{\beta}\|_\infty \leq \mathcal{M} & \\ & \|\boldsymbol{\beta}\|_1 \leq \mathcal{M}_{\ell_1} & \\ & -\bar{\mathcal{M}}_i^- \leq \langle \mathbf{x}_i, \boldsymbol{\beta} \rangle \leq \bar{\mathcal{M}}_i^+, i \in [n] & \end{array} \tag{25}$$

where UB is an upper bound to Problem (3) obtained via Algorithm 1, for example. Upon solving Problem (25), we set $\mathcal{M}_i = \max\{\mathcal{M}_i^+, \mathcal{M}_i^-\}$ for all $i \in [p]$. Consequently, we can update the bounds $\mathcal{M} = \|\mathcal{M}_i\|_\infty$, $\bar{\mathcal{M}}_i$ and $\mathcal{M}_{\ell_1}$ – such bound tightening methods have been proposed in [40] in the context of the Discrete Dantzig Selector problem.

Similarly, we can also obtain bounds on $\langle \mathbf{x}_j, \boldsymbol{\beta} \rangle$ by solving the following pair of optimization problems for all $j \in [n]$.

$$\begin{array}{lll} \bar{\mathcal{M}}_j^+ = \max & \langle \mathbf{x}_j, \boldsymbol{\beta} \rangle & \\ \text{s.t.} & \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_q \leq \text{UB} & \\ & -\mathcal{M}_i^- \leq \beta_i \leq \mathcal{M}_i^+, i \in [p] & \\ & \|\boldsymbol{\beta}\|_1 \leq \mathcal{M}_{\ell_1} & \\ & -\bar{\mathcal{M}}_i^- \leq \langle \mathbf{x}_i, \boldsymbol{\beta} \rangle \leq \bar{\mathcal{M}}_i^+, i \in [n] & \end{array} \qquad \begin{array}{lll} \bar{\mathcal{M}}_j^- = \max & -\langle \mathbf{x}_j, \boldsymbol{\beta} \rangle & \\ \text{s.t.} & \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_q \leq \text{UB} & \\ & -\mathcal{M}_i^- \leq \beta_i \leq \mathcal{M}_i^+, i \in [p] & \\ & \|\boldsymbol{\beta}\|_1 \leq \mathcal{M}_{\ell_1} & \\ & -\bar{\mathcal{M}}_i^- \leq \langle \mathbf{x}_i, \boldsymbol{\beta} \rangle \leq \bar{\mathcal{M}}_i^+, i \in [n]. & \end{array} \tag{26}$$

Upon solving Problem (26), we can set $\bar{\mathcal{M}}_i = \max\{|\bar{\mathcal{M}}_j^+|, |\bar{\mathcal{M}}_j^-|\}$. The bounds thus obtained can be used to tighten the bounds used in Problems (25) and (26). New bounds on $\{\mathcal{M}_i\}$ and $\{\bar{\mathcal{M}}_i\}$ can be obtained by solving the new problems with the updated bounds.

Remark 12. *Problems* (25), (26) *drop the cardinality constraint on* $\boldsymbol{\beta}$ *– hence the derived bounds need not be tight, i.e.,* $\mathcal{M}_i > |\hat{\beta}_i(\lambda; k)|$*, where* $\hat{\boldsymbol{\beta}}(\lambda; k)$ *denotes an optimal solution to Problem* (3).

A.3.2 Computing parameters via Algorithm 1

We note that the BigM values $\mathcal{M}_i, i \in [p]$ can also be based on the solutions obtained from the heuristic algorithms. For example, we can set $\mathcal{M}_i = \tau \|\hat{\boldsymbol{\beta}}(\lambda; k)\|_\infty$ for all $i \in [p]$ for some multiplier $\tau \in \{1.5, 2\}$, for example. Similarly, the bounds $\bar{\mathcal{M}}_i$ can be set to $\tau |\langle \mathbf{x}_i, \hat{\boldsymbol{\beta}}(\lambda; k)\rangle|$ for all $i \in [n]$. Such bounds are usually tighter and are obtained as a simple by-product of Algorithm 1.

B Proofs of the results in Section 3

B.1 Proof of Theorem 1

We first note that the probability of event $\mathcal{F}$ is at least $1 - \delta_0/2$ by Theorem 4.1 in [5]. Next, we establish the probability bound for $\mathcal{E}_s$.

Because the columns of $\mathbf{X}$ have unit Euclidean norm, we can write $\|\mathbf{X}\mathbf{u}\| \leq \|\mathbf{u}\|_1 \leq \sqrt{s}\|\mathbf{u}\|$ for every $\mathbf{u} \in B_0(s)$. Hence, taking $\delta_0 = s/(2ep)$, we derive

$$\sqrt{\log(1/\delta_0)}\|\mathbf{X}\mathbf{u}\| \leq \sqrt{s \log(2ep/s)}\|\mathbf{u}\|. \tag{27}$$

It follows from Stirling's formula that $\log(s!) \geq s \log(s/e)$, and hence

$$\sum_{j=1}^{s} \log(2p/j) = s \log(2p) - \log(s!) \leq s \log(2ep/s).$$

Thus, using the Cauchy-Schwarz inequality and taking into account $\|\mathbf{u}\|_0 \leq s$, we arrive at

$$\sum_{j=1}^{p} u_j^\sharp \sqrt{\log(2p/j)} \leq \|\mathbf{u}\| \sqrt{\sum_{j=1}^{s} \log(2p/j)} \leq \sqrt{s \log(2ep/s)}\|\mathbf{u}\|. \tag{28}$$

Inequalities (28) and (27) yield

$$[4 + \sqrt{2}]\sigma \max\Big(\sum_{j=1}^{p} u_j^\sharp \sqrt{\log(2p/j)}, \sqrt{\log(1/\delta_0)}\|\mathbf{X}\mathbf{u}\| \Big) \leq [4 + \sqrt{2}]\sigma \sqrt{s \log(2ep/s)}\|\mathbf{u}\|.$$

Consequently, when $\delta_0 = s/(2ep)$, we have $\mathcal{F} \subseteq \mathcal{E}_s$. Because the probability of event $\mathcal{F}$ is at least $1 - s/(4ep)$, we have established the stated probability bound for $\mathcal{E}_s$.

The result for $\mathcal{H}$ follows from the standard tail probability bounds for maxima of Gaussian random variables (for example, those in [14]). The result for $\mathcal{G}_s$ follows from the argument in the

proof of Lemma 8 in [51], with appropriate modifications in order to incorporate the uncertainty parameter δ_0.

B.2 Proof of Theorem 2

We consider an arbitrary $\boldsymbol{\beta} \in B_0(k)$ and note that

$$\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_2\|^2 + \lambda\|\hat{\boldsymbol{\beta}}_2\| \leq \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|,$$

which implies

$$\|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}_2\|^2 + \lambda\|\hat{\boldsymbol{\beta}}_2\| \leq \|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|^2 + 2\boldsymbol{\epsilon}^\top\mathbf{X}(\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|. \tag{29}$$

We will derive prediction error bounds for $\hat{\boldsymbol{\beta}}_2$ by controlling the term $\boldsymbol{\epsilon}^\top\mathbf{X}(\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta})$.

We first focus on establishing the slow rate. On the event $\mathcal{E}_{2k}$ we have

$$\boldsymbol{\epsilon}^\top\mathbf{X}(\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}) \leq [4 + \sqrt{2}]\sigma\sqrt{2k\log(ep/k)}\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_2\|. \tag{30}$$

Combining this inequality with (29) and using the lower bound imposed on λ, we derive

$$\|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}_2\|^2 \leq \|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|^2 + 2\lambda\|\boldsymbol{\beta}\|. \tag{31}$$

Thus, we have established the first slow rate prediction error bound.

Repeating the arguments in the proof of Theorem 1, we see that on the event $\mathcal{F}$ we have either (a) inequality (30), which implies (31), or (b) the following inequality:

$$\boldsymbol{\epsilon}^\top\mathbf{X}(\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}) \leq [4 + \sqrt{2}]\sigma\sqrt{\log(1/\delta_0)}\|\mathbf{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_2)\|, \tag{32}$$

which implies

$$\|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}_2\|^2 \leq \|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\| + 2[4 + \sqrt{2}]\sigma\sqrt{\log(1/\delta_0)}\big(\|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}_2\| + \|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|\big). \tag{33}$$

We bound the last term in the above display by two applications of the inequality

$$2ab \leq \alpha a^2 + \alpha^{-1}b^2, \tag{34}$$

which holds for every $\alpha > 0$ and $a, b \in \mathbb{R}$. Setting $\alpha = 2$, we derive inequalities

$$2[4 + \sqrt{2}]\sigma\sqrt{\log(1/\delta_0)}\|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}_2\| \leq 2[4 + \sqrt{2}]^2\sigma^2\log(1/\delta_0) + \|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}_2\|^2/2$$

and

$$\sigma\sqrt{\log(1/\delta_0)}\|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\| \lesssim \sigma^2\log(1/\delta_0) + \|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|^2.$$

Taking into account inequality (33), we then arrive at the second slow rate prediction error bound:

$$\|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}_2\|^2 \lesssim \|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\| + \sigma^2\log(1/\delta_0).$$

We now establish the fast rate. Starting with inequality (29) and restricting our attention to event $\mathcal{G}_{2k}$, we derive

$$\|\mathbf{f}^*-\mathbf{X}\hat{\boldsymbol{\beta}}_2\|^2 \leq \|\mathbf{f}^*-\mathbf{X}\boldsymbol{\beta}\|^2+2\sigma\big[10k\log(ep/[2k])+\log(1/\delta_0)\big]^{1/2}\big(\|\mathbf{f}^*-\mathbf{X}\hat{\boldsymbol{\beta}}_2\|+\|\mathbf{f}^*-\mathbf{X}\boldsymbol{\beta}\|\big)+\lambda\|\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_2\|.$$

We bound the second term on the right-hand side by two applications of inequality (34), in which we set $\alpha = 4$ in order to have $\|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}_2\|^2$ appear with the multiplier $1/4$. We bound the last term on the right-hand side using

$$\lambda\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_2\| \leq \gamma_{2k}^{-1}\lambda\|\mathbf{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_2)\| \leq \gamma_{2k}^{-1}\lambda\big(\|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}_2\| + \|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|\big),$$

and then apply (34) with $\alpha = 2$ again to derive

$$\gamma_{2k}^{-1}\lambda\|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}_2\| \leq \gamma_{2k}^{-2}\lambda^2 + \|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}_2\|^2/4$$

and

$$\gamma_{2k}^{-1}\lambda\|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\| \lesssim \gamma_{2k}^{-2}\lambda^2 + \|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|^2.$$

Rearranging the resulting terms we arrive at the fast rate prediction error bound:

$$\|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}_2\|^2 \lesssim \|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|^2 + \sigma^2 k\log(ep/[2k]) + \gamma_{2k}^{-2}\lambda^2 + \sigma^2\log(1/\delta_0).$$

B.3 Proof of Corollary 1

The first prediction error bound is a direct consequence of Theorems 1 and 2. The last two prediction error bounds are derived from Theorems 1 and the corresponding bounds in Theorem 2 by setting $\delta_0 = 1/p$ and $\delta_0 = (k/p)^k$, respectively. The estimation error bound follows from the inequality $\gamma_{2k}^2\|\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}^*\|^2 \leq \|\mathbf{X}(\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}^*)\|^2$.

B.4 Proof of Corollary 2

We let $Q(\boldsymbol{\beta})$ denote the the objective function in (16) when $q = 2$. Because $UB = Q(\widetilde{\boldsymbol{\beta}}_2)$, $LB \leq Q(\boldsymbol{\beta}^*)$, and $UB = LB/(1-\tau)$, we derive

$$Q(\widetilde{\boldsymbol{\beta}}_2) \leq Q(\boldsymbol{\beta}^*)/(1-\tau).$$

Because $Q(\widetilde{\boldsymbol{\beta}}_2) = \|\mathbf{y} - \mathbf{X}\widetilde{\boldsymbol{\beta}}_2\|^2 + \lambda\|\widetilde{\boldsymbol{\beta}}\|$ and $Q(\boldsymbol{\beta}^*) = \|\boldsymbol{\epsilon}\|^2 + \lambda\|\boldsymbol{\beta}^*\|$, we then have

$$\|\mathbf{y} - \mathbf{X}\widetilde{\boldsymbol{\beta}}_2\|^2 + \lambda\|\widetilde{\boldsymbol{\beta}} \leq \|\boldsymbol{\epsilon}\|^2/(1-\tau) + \lambda\|\boldsymbol{\beta}^*\|/(1-\tau).$$

Repeating the arguments in the proof of the second slow rate in Theorem 2 while incorporating the optimality gap, we derive that

$$\|\mathbf{f}^* - \mathbf{X}\widetilde{\boldsymbol{\beta}}_2\|^2 \leq 2\lambda\|\boldsymbol{\beta}^*\|/(1-\tau) + 4[4+\sqrt{2}]^2\sigma^2\log(1/\delta_0) + 2\|\boldsymbol{\epsilon}\|^2\tau/(1-\tau) \tag{35}$$

on the event $\mathcal{F}$. Standard chi-square tail bounds imply that, with an appropriate multiplicative constant, inequality $\|\boldsymbol{\epsilon}\|^2 \lesssim \sigma^2[n \vee \log(p)]$ holds with probability at least $1 - 1/(2p)$. Letting $\delta_0 = 1/(2p)$, noting $\tau \leq 1$, and recalling that $1/(1-\tau)$ is upper-bounded by a universal constant, we then conclude that inequality

$$\|\mathbf{f}^* - \mathbf{X}\widetilde{\boldsymbol{\beta}}_2\|^2 \lesssim \lambda\|\boldsymbol{\beta}^*\| + \sigma^2[\log(p) + \tau n]$$

holds with probability at least $1 - 1/p$, establishing the first error bound in Corollary 2.

Revisiting inequality (35) with $\delta_0 = 1/p$, we note that (as $n \to \infty$) the right-hand side is of the order

$$2\lambda\|\boldsymbol{\beta}^*\|\Big(1 + \frac{\tau}{1-\tau}\Big) + 4[4+\sqrt{2}]^2\sigma^2\log(p)\left\{1 + \frac{\tau n}{2[4+\sqrt{2}]^2(1-\tau)\log(p)}\right\}.$$

Thus, the multiplicative increase in the error bound relative to the case $\tau = 0$ is at most

$$1 + \frac{\tau}{1-\tau}\left\{1 \vee \frac{n}{58\log(p)}\right\}.$$

We now focus on the second error bound in Corollary 2. Repeating the arguments in the proof of the fast rate in Theorem 2, incorporating the optimality gap, and keeping track of the constants, we arrive at the following error bound:

$$\|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}_2\|^2 \leq 8\sigma^2\big[10k\log(ep/[2k]) + \log(1/\delta_0)\big] + 2\gamma_{2k}^{-2}\lambda^2\Big(1 + \frac{\tau}{1-\tau}\Big) + 2\|\boldsymbol{\epsilon}\|^2\tau/(1-\tau), \tag{36}$$

which holds on the event $\mathcal{G}_{2k}$. Letting $\delta_0 = (k/p)^k/2$, and again using the chi-square tail bounds to control $\|\boldsymbol{\epsilon}\|^2$, we then conclude that inequality

$$\|\mathbf{f}^* - \mathbf{X}\widetilde{\boldsymbol{\beta}}_2\|^2 \lesssim \sigma^2[k\log(ep/k) + \tau n] + \gamma_{2k}^{-2}\lambda^2$$

holds with probability at least $1-(k/p)^k$, establishing the second error bound in Corollary 2.

Revisiting inequality (36) with $\delta_0 = (k/p)^k$, we note that (as $n \to \infty$) the right-hand side is of the order

$$88\sigma^2\log(ep/[2k])\left\{1 + \frac{\tau n}{44(1-\tau)\log(ep/[2k])}\right\} + 2\gamma_{2k}^{-2}\lambda^2\Big(1 + \frac{\tau}{1-\tau}\Big).$$

Thus, the multiplicative increase in the error bound relative to the case $\tau = 0$ is at most

$$1 + \frac{\tau}{1-\tau}\left\{1 \vee \frac{n}{43\log(ep/[2k])}\right\}.$$

B.5 Proof of Corollary 3

Let c_0 be the universal constant from the second slow rate error bound in Theorem 2. Take an arbitrary $\boldsymbol{\beta} \in B_0(k)$ and define

$$W = \|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}_2\|^2 - c_0\|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|^2 - c_0\lambda\|\boldsymbol{\beta}\|.$$

By Theorems 1 and 2 we have $W \le c_0\sigma^2\log(1/\delta_0)$ with probability at least $1-\delta_0/2$. Thus,

$$2\mathbb{P}\big(W > w\big) \le e^{-w/[c_0\sigma^2]},$$

for every non-negative w. Consequently,

$$\mathbb{E}W \le \int_0^\infty \mathbb{P}\big(W > w\big)dw \le \tfrac{1}{2}\int_0^\infty e^{-w/[c_0\sigma^2]}dw \le \frac{c_0\sigma^2}{2},$$

and the first stated bound follows from the definition of W.

The second stated bound follows by an analogous argument, together with an additional observation that $k\log(ep/[2k])$ is bounded away from zero by a positive universal constant.

B.6 Proof of Theorem 3

We consider an arbitrary $\boldsymbol{\beta} \in B_0(k)$. In the ℓ_1 setting, inequality (29) becomes

$$\|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}_1\|^2 + \lambda\|\hat{\boldsymbol{\beta}}_1\|_1 \le \|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|^2 + 2\boldsymbol{\epsilon}^\top\mathbf{X}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1. \tag{37}$$

On the event $\mathcal{H}$, we then have

$$2\boldsymbol{\epsilon}^\top \mathbf{X}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}) \leq 2\|\mathbf{X}^T\boldsymbol{\epsilon}\|_\infty \Big[\|\boldsymbol{\beta}\|_1 + \|\hat{\boldsymbol{\beta}}_1\|_1\Big] \leq \lambda\Big[\|\hat{\boldsymbol{\beta}}_1\|_1 + \|\boldsymbol{\beta}\|_1\Big].$$

Consequently,

$$\|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}_1\|^2 \leq \|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|^2 + 2\lambda\|\boldsymbol{\beta}\|_1,$$

which completes the proof of the first slow rate error bound.

We now restrict our attention to the event $\mathcal{F}$. Note that, because

$$\sum_{j=1}^{p} u_j^\sharp \sqrt{\log(2p/j)} \leq \sqrt{\log(2p)}\|\mathbf{u}\|_1,$$

we must have either (a) inequality

$$\boldsymbol{\epsilon}^\top \mathbf{X}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}) \leq [4+\sqrt{2}]\sigma\sqrt{\log(2p)}\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_1\|_1,$$

which implies

$$\|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}_1\|^2 \leq \|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|^2 + 2\lambda\|\boldsymbol{\beta}\|_1;$$

or (b) the following inequality:

$$\boldsymbol{\epsilon}^\top \mathbf{X}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}) \leq [4+\sqrt{2}]\sigma\sqrt{\log(1/\delta_0)}\|\mathbf{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_2)\|,$$

which implies

$$\|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}_1\|^2 \leq \|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_1 + [8+2\sqrt{2}]\sigma\sqrt{\log(1/\delta_0)}\Big(\|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}_1\| + \|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|\Big).$$

Bounding the last term in the above display by two applications of (34) with $\alpha = 2$ yields

$$\|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}_1\|^2 \lesssim \|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_1 + \sigma^2\log(1/\delta_0), \tag{38}$$

which establishes the second slow rate error bound.

We now move to the fast rate. Starting with (37), using inequalities

$$\lambda\|\boldsymbol{\beta}\|_1 - \lambda\|\hat{\boldsymbol{\beta}}_1\|_1 \leq \lambda\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_1\|_1 \leq \lambda\sqrt{2k}\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_1\|,$$

and restricting our attention to the event $\mathcal{G}_{2k}$, we derive

$$\|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}_1\|^2 \leq \|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|^2 + \sigma\Big[10k\log(ep/[2k]) + \log(1/\delta_0)\Big]^{1/2}\|\mathbf{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_1)\| + \lambda\sqrt{k}\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_1\|.$$

Repeating the argument used to establish the fast rate part of Theorem 2, we arrive at

$$\|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}_1\|^2 \lesssim \|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|^2 + \sigma^2 k\log(ep/[2k]) + \gamma_{2k}^{-2}\lambda^2 k + \sigma^2\log(1/\delta_0).$$

B.7 Proof of Corollary 4

This result follows by an argument analogous to the one used in the proof of Corollary 3.

B.8 Proof of Corollary 5

This result follows directly from the slow rate parts of Theorems 2 and 3.

B.9 Proof of Theorem 4

The following result will allow us to lower-bound the magnitude of the cross-product term in the sum of squares function.

Lemma 1. *Let $S \subset \{1, ..., p\}$ have cardinality q, and let s be an integer in $[1, q]$. There exists a positive universal constant $\tilde{c}$, such that*

$$\max_{supp(\mathbf{v}) \subset S,\, \|\mathbf{v}\|_0 \leq s,\, \|\mathbf{X}\mathbf{v}\| = 1} |\boldsymbol{\epsilon}^\top \mathbf{X}\mathbf{v}| \gtrsim \sigma \gamma_{2s} \sqrt{s \log(eq/s)}$$

with probability at least $1 - 2(eq/s)^{-\tilde{c}\gamma_{2s}^2 s}$.

Lemma 1 is proved in the next subsection.

Using Maurey's argument [49], we can bound the error in approximating $\mathbf{X}\boldsymbol{\beta}^*$ with $\mathbf{X}\boldsymbol{\beta}$, when $\boldsymbol{\beta}$ is restricted to an ℓ_0 ball. More specifically, by Lemma A.1 in [52], there exists a vector $\tilde{\boldsymbol{\beta}}^* \in B_0(k/2)$ such that $\|\mathbf{X}\boldsymbol{\beta}^* - \mathbf{X}\tilde{\boldsymbol{\beta}}^*\|^2 \leq 2\|\boldsymbol{\beta}^*\|_1^2/k$. For convenience, we define $\boldsymbol{\Delta}^* = \mathbf{X}\boldsymbol{\beta}^* - \mathbf{X}\tilde{\boldsymbol{\beta}}^*$. Minimizing the sum of squares is equivalent to minimizing the function

$$G(\boldsymbol{\beta}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 - \|\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}^*\|^2 = \|\mathbf{X}\boldsymbol{\beta} - \mathbf{X}\tilde{\boldsymbol{\beta}}^*\|^2 + 2(\boldsymbol{\Delta}^* + \boldsymbol{\epsilon})^\top (\mathbf{X}\tilde{\boldsymbol{\beta}}^* - \mathbf{X}\boldsymbol{\beta}).$$

Given a vector $\mathbf{u} \in \mathbb{R}^p$, we define

$$H(\mathbf{u}) = \|\mathbf{X}\mathbf{u}\|^2 - 2(\boldsymbol{\Delta}^* + \boldsymbol{\epsilon})^\top \mathbf{X}\mathbf{u}.$$

Given an index set $\mathcal{I}$ and a vector $\boldsymbol{\beta}$, we will write $\boldsymbol{\beta}_{\mathcal{I}}$ for the vector that (a) matches $\boldsymbol{\beta}$ element by element on the index set $\mathcal{I}$; and (b) has its support contained in $\mathcal{I}$. Let $\tilde{S}$ denote the support of $\tilde{\boldsymbol{\beta}}^*$. Note that if $\boldsymbol{\beta}_{\tilde{S}} = \tilde{\boldsymbol{\beta}}^*$ and $\|\boldsymbol{\beta}\|_0 \leq k$, then

$$G(\boldsymbol{\beta}) = \|\mathbf{X}\boldsymbol{\beta}_{\tilde{S}^c}\|^2 - 2(\boldsymbol{\Delta}^* + \boldsymbol{\epsilon})^\top \mathbf{X}\boldsymbol{\beta}_{\tilde{S}^c} = H(\boldsymbol{\beta}_{\tilde{S}^c}).$$

Note that $|\tilde{S}| \leq k/2$, and hence

$$\min_{\|\boldsymbol{\beta}\|_0 \leq k} G(\boldsymbol{\beta}) \leq \min_{\boldsymbol{\beta}_{\tilde{S}} = \tilde{\boldsymbol{\beta}}^*,\, \|\boldsymbol{\beta}\|_0 \leq k} G(\boldsymbol{\beta}) \leq \min_{\boldsymbol{\beta}_{\tilde{S}} = \tilde{\boldsymbol{\beta}}^*,\, \|\boldsymbol{\beta}\|_0 \leq k} H(\boldsymbol{\beta}_{\tilde{S}^c}) \leq \min_{\mathrm{supp}(\mathbf{u}) \subseteq \tilde{S}^c,\, \|\mathbf{u}\|_0 \leq k/2} H(\mathbf{u}). \tag{39}$$

To simplify the notation, we define $\mathcal{V}_k = \{\mathbf{v} \in \mathbb{R}^p, \text{ s.t. } \text{supp}(\mathbf{v}) \subseteq \tilde{S}^c, \|\mathbf{v}\|_0 \leq k/2, \|\mathbf{X}\mathbf{v}\| = 1\}$ and $c_{\mathbf{v}} = (\boldsymbol{\Delta}^* + \boldsymbol{\epsilon})^\top \mathbf{X}\mathbf{v}$. In addition to the inequalities in (39) we also have

$$\min_{\text{supp}(\mathbf{u}) \subseteq \tilde{S}^c, \|\mathbf{u}\|_0 \leq k/2} H(\mathbf{u}) \leq \min_{\mathcal{V}_k} H(c_{\mathbf{v}}\mathbf{v}) = \min_{\mathcal{V}_k} \big[- c_{\mathbf{v}}^2 \big] = - \max_{\mathcal{V}_k} |(\boldsymbol{\Delta}^* + \boldsymbol{\epsilon})^\top \mathbf{X}\mathbf{v}|^2.$$

Consequently,

$$\min_{\|\boldsymbol{\beta}\|_0 \leq k} G(\boldsymbol{\beta}) \leq - \max_{\mathcal{V}_k} |(\boldsymbol{\Delta}^* + \boldsymbol{\epsilon})^\top \mathbf{X}\mathbf{v}|^2. \tag{40}$$

Note that if $\|\mathbf{X}\mathbf{v}\| = 1$, then $|(\boldsymbol{\Delta}^* + \boldsymbol{\epsilon})^\top \mathbf{X}\mathbf{v}| \geq |\boldsymbol{\epsilon}^\top \mathbf{X}\mathbf{v}| - \|\boldsymbol{\Delta}^*\|$. Also note that

$$\|\boldsymbol{\Delta}^*\| \leq \|\boldsymbol{\beta}^*\|_1 / \sqrt{k/2} \lesssim \sigma \gamma_k \sqrt{k \log(ep/k)}, \tag{41}$$

with a sufficiently small multiplicative constant due to the assumption on $\|\boldsymbol{\beta}^*\|_1$. Note that the cardinality of $\tilde{S}^c$ is at least $p/2$. Thus, applying Lemma 1, with $s = k/2$ and $q = p/2$, to lower bound $\max_{\mathcal{V}_k} |\boldsymbol{\epsilon}^\top \mathbf{X}\mathbf{v}|$, we derive that

$$\max_{\mathcal{V}_k} |(\boldsymbol{\Delta}^* + \boldsymbol{\epsilon})^\top \mathbf{X}\mathbf{v}| \gtrsim \sigma \gamma_k \sqrt{k \log(ep/k)},$$

with probability at least $1 - 2(ep/k)^{-\tilde{c}\gamma_k^2 k/2}$. Thus, inequality (40) and the definition of $\hat{\boldsymbol{\beta}}_{\ell_0}$ yield

$$G(\hat{\boldsymbol{\beta}}_{\ell_0}) \leq \min_{\boldsymbol{\beta}:\ \|\boldsymbol{\beta}\|_0 \leq k} G(\boldsymbol{\beta}) \lesssim -\sigma^2 \gamma_k^2 k \log(ep/k).$$

Because $2(\boldsymbol{\Delta}^* + \boldsymbol{\epsilon})^\top (\mathbf{X}\tilde{\boldsymbol{\beta}}^* - \mathbf{X}\boldsymbol{\beta}) \leq G(\boldsymbol{\beta})$ for each $\boldsymbol{\beta}$, we derive

$$|(\boldsymbol{\Delta}^* + \boldsymbol{\epsilon})^\top (\mathbf{X}\hat{\boldsymbol{\beta}}_{\ell_0} - \mathbf{X}\tilde{\boldsymbol{\beta}}^*)| \gtrsim \sigma^2 \gamma_k^2 k \log(ep/k). \tag{42}$$

Taking into account (41), which holds with a sufficiently small multiplicative constant, we derive

$$|{\boldsymbol{\Delta}^*}^\top (\mathbf{X}\hat{\boldsymbol{\beta}}_{\ell_0} - \mathbf{X}\tilde{\boldsymbol{\beta}}^*)| \leq \|\boldsymbol{\Delta}^*\| \|\mathbf{X}\hat{\boldsymbol{\beta}}_{\ell_0} - \mathbf{X}\tilde{\boldsymbol{\beta}}^*\| \lesssim \sigma \gamma_k \sqrt{k \log(ep/k)} \|\mathbf{X}\hat{\boldsymbol{\beta}}_{\ell_0} - \mathbf{X}\tilde{\boldsymbol{\beta}}^*\|. \tag{43}$$

Furthermore, on the event $\mathcal{G}_{2k}$ with $\delta_0 = (ep/k)^{-k}$, which holds with probability at least $1 - \delta_0$ by Theorem 1, we have

$$|\boldsymbol{\epsilon}^\top (\mathbf{X}\hat{\boldsymbol{\beta}}_{\ell_0} - \mathbf{X}\tilde{\boldsymbol{\beta}}^*)| \lesssim \sigma \sqrt{k \log(ep/k)} \|\mathbf{X}\hat{\boldsymbol{\beta}}_{\ell_0} - \mathbf{X}\tilde{\boldsymbol{\beta}}^*\|. \tag{44}$$

Combining inequalities (42), (43) and (44), we arrive at

$$\|\mathbf{X}\hat{\boldsymbol{\beta}}_{\ell_0} - \mathbf{X}\boldsymbol{\beta}^*\| + \|\boldsymbol{\Delta}^*\| \geq \|\mathbf{X}\hat{\boldsymbol{\beta}}_{\ell_0} - \mathbf{X}\tilde{\boldsymbol{\beta}}^*\| \gtrsim \sigma \gamma_k \sqrt{k \log(ep/k)}.$$

Note that $\|\mathbf{X}\hat{\boldsymbol{\beta}}_{\ell_0} - \mathbf{X}\tilde{\boldsymbol{\beta}}^*\| \leq \|\mathbf{X}\hat{\boldsymbol{\beta}}_{\ell_0} - \mathbf{X}\boldsymbol{\beta}^*\| + \|\boldsymbol{\Delta}^*\|$, by the triangle inequality. Let $\psi = \tilde{c}/2$. Applying (41), which holds with a sufficiently small multiplicative constant, we conclude that

$$\|\mathbf{X}\hat{\boldsymbol{\beta}}_{\ell_0} - \mathbf{X}\boldsymbol{\beta}^*\| \gtrsim \sigma\gamma_k\sqrt{k\log(ep/k)},$$

with probability at least $1 - 2(ep/k)^{-c\gamma_k^2 k} - (ep/k)^{-k}$.

B.10 Proof of Lemma 1

Note that if $s > q/2$, then we can establish the bound for $s = \lfloor q/2 \rfloor$ and use

$$\max_{\|\mathbf{v}\|_0 \leq s,\, \|\mathbf{X}\mathbf{v}\|=1} |\boldsymbol{\epsilon}^\top \mathbf{X}\mathbf{v}| \geq \max_{\|\mathbf{v}\|_0 \leq \lfloor q/2 \rfloor,\, \|\mathbf{X}\mathbf{v}\|=1} |\boldsymbol{\epsilon}^\top \mathbf{X}\mathbf{v}|.$$

Hence, we will focus on the case $s \leq q/2$.

We write $|\cdot|$ for the cardinality of a set. Applying Lemma F.1 in [5], which is closely related to the results in [57], we deduce that there exists a subset $\mathcal{H}$ of the set $\{-1, 0, 1\}^p$, with

$$\log\left(|\mathcal{H}|\right) \gtrsim s\log(eq/s),$$

such that $\mathrm{supp}(\mathbf{v}) \subset S$, $\|\mathbf{v}\|_0 \leq s$, $\|\mathbf{X}\mathbf{v}\|^2 \leq s$ and $\|\mathbf{v}_1 - \mathbf{v}_2\|^2 \geq s/4$, for all $\mathbf{v}, \mathbf{v}_1, \mathbf{v}_2 \in \mathcal{H}$. Note that the last inequality implies

$$\|\mathbf{X}\mathbf{v}_1 - \mathbf{X}\mathbf{v}_2\|^2 \geq \gamma_{2s}^2 s/4.$$

Consequently, by Sudakov's minoration [for example, Proposition 3.15 in 39],

$$E\max_{\mathbf{v}\in\mathcal{H}} \boldsymbol{\epsilon}^\top \mathbf{X}\mathbf{v} \gtrsim \sigma\gamma_{2s}\sqrt{s\log(|\mathcal{H}|)} \gtrsim \sigma\gamma_{2s}s\sqrt{\log(eq/s)}.$$

Define $W = \max_{\mathbf{v}\in\mathcal{H}} \boldsymbol{\epsilon}^\top \mathbf{X}\mathbf{v}$ and $v = \max_{\mathbf{v}\in\mathcal{H}} SD(\boldsymbol{\epsilon}^\top \mathbf{X}\mathbf{v})$ by v. By the concentration inequality for the supremum of a Gaussian process [for example, Theorem 3.12 in 39], we have, for all $t \geq 0$,

$$P\left(W \leq EW - vt\right) \leq 2\exp(-t^2/2).$$

Note that

$$v \leq \sigma \max_{\mathbf{v}\in\mathcal{H}} \|\mathbf{X}\mathbf{v}\| \leq \sigma\sqrt{s}.$$

Consequently, if $t \leq \gamma_{2s}\sqrt{\tilde{c}s\log(eq/s)}$ with a sufficiently small positive universal constant $\tilde{c}$, then $EW - vt \gtrsim \sigma\gamma_{2s}s\sqrt{\log(eq/s)}$, and hence

$$\max_{\mathbf{v}\in\mathcal{H}} \boldsymbol{\epsilon}^\top \mathbf{X}\mathbf{v} \gtrsim \sigma\gamma_{2s}s\sqrt{\log(eq/s)},$$

with probability at least $1 - 2\exp(-\tilde{c}\gamma_{2s}^2 s\log(eq/s))$. We complete the proof by noting that

$$\max_{\|\mathbf{v}\|_0 \leq s,\, \|\mathbf{X}\mathbf{v}\|=1} \boldsymbol{\epsilon}^\top \mathbf{X}\mathbf{v} \geq s^{-1/2} \max_{\mathbf{v}\in\mathcal{H}} \boldsymbol{\epsilon}^\top \mathbf{X}\mathbf{v}.$$

B.11 Proof of Proposition 3

Note that the assumptions imposed on b imply

$$b \gtrsim 1 \qquad \text{and} \qquad b \lesssim 1, \tag{45}$$

where the universal constant in the second bound can be chosen to be sufficiently small. Also note that

$$\|\boldsymbol{\beta}^*\| = b\sigma\sqrt{[\log(ep)]/k^*} \qquad \text{and} \qquad \|\boldsymbol{\beta}^*\|_1 = b\sigma\sqrt{\log(ep)}. \tag{46}$$

Thus, to establish the result of Proposition 3, we only need to demonstrate that

$$\min_{k\in[0,p]} \|\mathbf{X}\boldsymbol{\beta}^* - \mathbf{X}\widehat{\boldsymbol{\beta}}_{\ell_0}\|^2 \gtrsim \sigma^2 \log(ep), \tag{47}$$

with high probability.

Let $\mathbf{1}$ denote a p-dimensional vector of ones, and note that

$$\|\mathbf{X}\boldsymbol{\beta}^*\|^2 \geq \rho_l \boldsymbol{\beta}^{*\top}\mathbf{1}\mathbf{1}^\top\boldsymbol{\beta}^* = \rho_l\|\boldsymbol{\beta}^*\|_1^2 = \rho_l b^2\sigma^2\log(ep) \gtrsim \sigma^2\log(ep).$$

We conclude that for $k = 0$ bound (47) holds with probability one. For the remainder of the proof we focus on the case of $k \in [p]$.

Minimizing the sum of squares is equivalent to minimizing the function

$$L(\boldsymbol{\beta}) = \|\mathbf{X}\boldsymbol{\beta}\|^2 - 2\mathbf{y}^\top\mathbf{X}\boldsymbol{\beta}.$$

Define $c_j = \mathbf{y}^\top\mathbf{X}_j$ and let $\mathbf{e}_j$ denote the j-th coordinate vector in $\mathbb{R}^p$. Because $\mathbf{X}\mathbf{e}_j = \mathbf{X}_j$ and $\|\mathbf{X}_j\| = 1$, we have

$$\min_{\|\boldsymbol{\beta}\|_0=1} L(\boldsymbol{\beta}) \leq \min_j L(c_j\mathbf{e}_j) = \min_j -c_j^2 = -\max_j |\mathbf{y}^\top\mathbf{X}_j|^2.$$

We also have

$$\begin{aligned}
\max_j |\mathbf{y}^\top\mathbf{X}_j|^2 &= \max_j \Big(|\boldsymbol{\epsilon}^\top\mathbf{X}_j|^2 + 2(\boldsymbol{\epsilon}^\top\mathbf{X}_j)(\mathbf{X}_j^\top\mathbf{X}\boldsymbol{\beta}^*) + |\mathbf{X}_j^\top\mathbf{X}\boldsymbol{\beta}^*|^2\Big) \\
&\geq \max_j \Big(|\boldsymbol{\epsilon}^\top\mathbf{X}_j|^2 - 2|\boldsymbol{\epsilon}^\top\mathbf{X}_j|\|\mathbf{X}\boldsymbol{\beta}^*\|\Big) \\
&\geq \max_j \Big(|\boldsymbol{\epsilon}^\top\mathbf{X}_j|^2/2 - 2\|\mathbf{X}\boldsymbol{\beta}^*\|^2\Big),
\end{aligned}$$

where we used bound (34) with $a = |\boldsymbol{\epsilon}^\top\mathbf{X}_j|$, $b = \|\mathbf{X}\boldsymbol{\beta}^*\|$ and $\alpha = 1/2$ to get the last inequality. Applying Lemma 1, we derive that

$$\max_j |\boldsymbol{\epsilon}^\top\mathbf{X}_j| \gtrsim (1-\rho_u)\sigma^2\log(ep),$$

with probability at least $1 - 2(ep)^{-\tilde{c}(1-\rho_u)}$, for some positive universal constant $\tilde{c}$.

Inequalities (46) and (45), together with the fact that columns of $\mathbf{X}$ have unit norm, yield

$$\|\mathbf{X}\boldsymbol{\beta}^*\|^2 \leq \|\boldsymbol{\beta}^*\|_1^2 \leq b^2\sigma^2 \log(ep) \lesssim \sigma^2 \log(ep), \tag{48}$$

with a sufficiently small universal constant. Consequently,

$$\min_{\|\boldsymbol{\beta}\|_0 \leq k} L(\boldsymbol{\beta}) \leq \min_{\|\boldsymbol{\beta}\|_0 = 1} L(\boldsymbol{\beta}) \lesssim -\sigma^2 \log(ep), \tag{49}$$

uniformly over $k \in [p]$ and with probability at least $1 - 2(ep)^{-a}$, for some positive universal constant a.

We conduct the rest of the argument on the high-probability event where (49) holds. On this event we have the bound

$$L(\hat{\boldsymbol{\beta}}_{\ell_0}) = \|\mathbf{X}\hat{\boldsymbol{\beta}}_{\ell_0}\|^2 - 2\mathbf{y}^\top \mathbf{X}\hat{\boldsymbol{\beta}}_{\ell_0} \lesssim -\sigma^2 \log(ep), \tag{50}$$

in which the universal constant does not depend on k. Given a set $S \subseteq \{1, ..., p\}$, we define $\hat{\boldsymbol{\beta}}_S = \arg\min_{\text{supp}(\boldsymbol{\beta}) \subseteq S} L(\boldsymbol{\beta})$ and note that $\|\mathbf{X}\hat{\boldsymbol{\beta}}_S\|^2 = \mathbf{y}^\top \mathbf{X}\hat{\boldsymbol{\beta}}_S$. Consequently, $\|\mathbf{X}\hat{\boldsymbol{\beta}}_{\ell_0}\|^2 = \mathbf{y}^\top \mathbf{X}\hat{\boldsymbol{\beta}}_{\ell_0}$, and hence bound (50) implies

$$\|\mathbf{X}\hat{\boldsymbol{\beta}}_{\ell_0}\|^2 \gtrsim \sigma^2 \log(ep).$$

Bound (47) then follows from the inequality $\|\mathbf{X}\boldsymbol{\beta}^* - \mathbf{X}\hat{\boldsymbol{\beta}}_{\ell_0}\| \geq \|\mathbf{X}\hat{\boldsymbol{\beta}}_{\ell_0}\| - \|\mathbf{X}\boldsymbol{\beta}^*\|$ and bound (48), applied with a sufficiently small universal constant.

B.12 Proof of Proposition 4

Let $\mathbf{1}$ denote a p-dimensional vector of ones, and note that

$$\mathbf{X}^\top \mathbf{X} = (1 - \rho)\mathbf{I} + \rho \mathbf{1}\mathbf{1}^\top.$$

Hence, for every $\mathbf{u} \in \mathbb{R}^p$,

$$\|\mathbf{X}\mathbf{u}\|^2 = (1 - \rho)\|u\|^2 + \rho(\mathbf{1}^\top \mathbf{u})^2 \geq (1 - \rho)\|u\|^2,$$

which implies $\gamma_k^2 \geq 1 - \rho$. Also note that, by (46), $\|\boldsymbol{\beta}^*\| = b\sigma\sqrt{[\log(ep)]/k^*}$, which can be made smaller than any given multiple of $\sigma\sqrt{k \log(ep/k)}$ under the assumptions imposed on b, k and k^* in Proposition 4. Under this scenario, we can apply Theorem 4, which leads to

$$\|\mathbf{X}\hat{\boldsymbol{\beta}}_{\ell_0} - \mathbf{X}\boldsymbol{\beta}^*\|^2 \gtrsim \sigma^2 k \log(ep/k).$$

B.13 Proof of Theorem 5

We first establish the bound for $\hat{\boldsymbol{\beta}}_2^B$, and then establish the one for $\hat{\boldsymbol{\beta}}_1^B$. We note that throughout the proof the positive multiplicative factors in inequalities $\lesssim$ and $\gtrsim$ are universal constants, which are independent from all other parameters such as n, p, σ, $\boldsymbol{\beta}^*$, $\boldsymbol{\beta}$ and s.

Expected prediction error bound for $\hat{\boldsymbol{\beta}}_2^B$.
To simplify the expressions, we drop the subscript and the superscript in $\hat{\boldsymbol{\beta}}_2^B$ and simply write $\hat{\boldsymbol{\beta}}$.

Taking an arbitrary $\boldsymbol{\beta} \in \mathbb{R}^p$, we note that

$$\|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 + \lambda_{\hat{\boldsymbol{\beta}}}\|\hat{\boldsymbol{\beta}}\| + \mu_{\hat{\boldsymbol{\beta}}}|||\hat{\boldsymbol{\beta}}\|_0 \leq \|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|^2 + 2\boldsymbol{\epsilon}^\top \mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\beta}\| + \mu_{\boldsymbol{\beta}}\|\boldsymbol{\beta}\|_0. \tag{51}$$

In the setting where $a \gtrsim \sigma$, with a sufficiently large multiplicative constant, we will bound the term $2\boldsymbol{\epsilon}^\top \mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \lambda_{\hat{\boldsymbol{\beta}}}\|\hat{\boldsymbol{\beta}}\|$. Similarly, in the case $b \gtrsim \sigma^2$ we will bound $2\boldsymbol{\epsilon}^\top \mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \mu_{\hat{\boldsymbol{\beta}}}|||\hat{\boldsymbol{\beta}}\|_0$.

We first consider the case $a \gtrsim \sigma$, which implies $\lambda_{\boldsymbol{\beta}} \gtrsim \sigma\sqrt{\|\boldsymbol{\beta}\|_0 \log(ep/\|\boldsymbol{\beta}\|_0)}$. We let $\hat{\mathbf{u}} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ and restrict our attention to event $\mathcal{F}$, defined in Section 3.1, which holds with probability at least $1 - \delta_0/2$. On event $\mathcal{F}$, we have either $\boldsymbol{\epsilon}^\top \mathbf{X}\hat{\mathbf{u}} \lesssim \sigma \sum_{j=1}^p \hat{u}_j^\sharp \sqrt{\log(2p/j)}$ or $\boldsymbol{\epsilon}^\top \mathbf{X}\hat{\mathbf{u}} \lesssim \sigma\sqrt{\log(1/\delta_0)}\|\mathbf{X}\hat{\mathbf{u}}\|$. In the latter scenario, repeating the argument after inequality (32) in the proof of Theorem 2 yields

$$\|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 + \mu_{\hat{\boldsymbol{\beta}}_2}|||\hat{\boldsymbol{\beta}}\|_0 \lesssim \|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\beta}\| + \sigma^2 \log(1/\delta_0) + \mu_{\boldsymbol{\beta}}\|\boldsymbol{\beta}\|_0. \tag{52}$$

We now focus on the event $\boldsymbol{\epsilon}^\top \mathbf{X}\hat{\mathbf{u}} \lesssim \sigma \sum_{j=1}^p \hat{u}_j^\sharp \sqrt{\log(2p/j)}$. In view of (28), we have

$$\begin{aligned}
\sigma \sum_{j=1}^p \hat{u}_j^\sharp \sqrt{\log(2p/j)} &\leq \sigma \sum_{j=1}^p \hat{\beta}_j^\sharp \sqrt{\log(2p/j)} + \sigma \sum_{j=1}^p \beta_j{}^\sharp \sqrt{\log(2p/j)} \\
&\leq \sigma\sqrt{\hat{k}\log(2ep/\hat{k})}\|\hat{\boldsymbol{\beta}}\| + \sigma\sqrt{\|\boldsymbol{\beta}\|_0 \log(2ep/\|\boldsymbol{\beta}\|_0)}\|\boldsymbol{\beta}\|.
\end{aligned}$$

Thus, if $a \gtrsim \sigma$ with a sufficiently large universal constant, then $2\boldsymbol{\epsilon}^\top \mathbf{X}\hat{\mathbf{u}} \leq \lambda_{\hat{\boldsymbol{\beta}}}\|\hat{\boldsymbol{\beta}}\| + \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\beta}\|$. Combining this bound with inequality (51), we derive

$$\|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 + \mu_{\hat{\boldsymbol{\beta}}_2}|||\hat{\boldsymbol{\beta}}\|_0 \lesssim \|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|^2 + 2\lambda_{\boldsymbol{\beta}}\|\boldsymbol{\beta}\| + \mu_{\boldsymbol{\beta}}\|\boldsymbol{\beta}\|_0. \tag{53}$$

Bounds (52) and (53) imply that, for each $\delta_0 \in (0, 1)$,

$$\|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 + \mu_{\hat{\boldsymbol{\beta}}_2}|||\hat{\boldsymbol{\beta}}\|_0 \lesssim \|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\beta}\| + \sigma^2 \log(1/\delta_0) + \mu_{\boldsymbol{\beta}}\|\boldsymbol{\beta}\|_0 \tag{54}$$

with probability at least $1 - \delta_0/2$.

We now focus on the case $b \gtrsim \sigma^2$, which implies $\mu_{\boldsymbol{\beta}} \gtrsim \sigma^2 \log(ep/\|\boldsymbol{\beta}\|_0)$. Given an $s \in [p]$, we consider the event $\mathcal{G}_s$, defined in Section 3.1, where we take $\delta_0 = (s/[ep])^s \epsilon_0$. Here, $\epsilon_0 \in (0,1)$ is an arbitrary value that does not depend on s. On the event $\mathcal{G} = \cap_{s=1}^p \mathcal{G}_s$ we have

$$\begin{aligned}\boldsymbol{\epsilon}^\top \mathbf{X}\hat{\mathbf{u}} &\lesssim \sigma\sqrt{\|\hat{\mathbf{u}}\|_0 \log(ep/\|\hat{\mathbf{u}}\|_0) + \log(1/\epsilon_0)}\|\mathbf{X}\hat{\mathbf{u}}\| \\ &\lesssim \sigma\sqrt{\|\hat{\boldsymbol{\beta}}\|_0 \log(ep/\|\hat{\boldsymbol{\beta}}\|_0) + \|\boldsymbol{\beta}\|_0 \log(ep/\|\boldsymbol{\beta}\|_0) + \log(1/\epsilon_0)}\|\mathbf{X}\hat{\mathbf{u}}\|.\end{aligned}$$

Noting that $\|\mathbf{X}\hat{\mathbf{u}}\| \le \|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}\| + \|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|$ and applying inequality (34) twice, we derive

$$2\boldsymbol{\epsilon}^\top \mathbf{X}\hat{\mathbf{u}} \le c\Big[\sigma^2\|\hat{\boldsymbol{\beta}}\|_0 \log(ep/\|\hat{\boldsymbol{\beta}}\|_0) + \sigma^2\|\boldsymbol{\beta}\|_0 \log(ep/\|\boldsymbol{\beta}\|_0) + \sigma^2 \log(1/\epsilon_0) + \|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|^2\Big] + \|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2/2,$$

for some universal constant c. Taking into account inequality (51), we deduce that

$$\|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 + 2[b - c\sigma^2]\|\hat{\boldsymbol{\beta}}\|_0 \log(ep/\|\hat{\boldsymbol{\beta}}\|_0) \lesssim \mu_{\boldsymbol{\beta}}\|\boldsymbol{\beta}\|_0 + \sigma^2 \log(1/\epsilon_0) + \|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\beta}\|$$

on the event $\mathcal{G}$. Note that

$$\mathbb{P}(\mathcal{G}^c) \le \sum_{s=1}^p \mathcal{P}(\mathcal{G}_s^c) \le \sum_{s=1}^p (s/[ep])^s \epsilon_0 \le \sum_{s=1}^p e^{-s}\epsilon_0 \le \epsilon_0.$$

Consequently, if we let $b \ge 2c\sigma^2$, then

$$\|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 + \mu_{\hat{\boldsymbol{\beta}}}|\|\hat{\boldsymbol{\beta}}\|_0 \lesssim \|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|^2 + \sigma^2 \log(1/\epsilon_0) + \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\beta}\| + \mu_{\boldsymbol{\beta}}\|\boldsymbol{\beta}\|_0 \tag{55}$$

with probability at least $1 - \epsilon_0$.

Combining bounds (54) and (55), we conclude that, for each $\delta_0 \in (0,1)$,

$$\|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 + \mu_{\hat{\boldsymbol{\beta}}}|\|\hat{\boldsymbol{\beta}}\|_0 \lesssim \inf_{\boldsymbol{\beta}\in\mathbb{R}^p}\Big[\|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\beta}\| + \mu_{\boldsymbol{\beta}}\|\boldsymbol{\beta}\|_0\Big] + \sigma^2 \log(1/\delta_0) \tag{56}$$

with probability at least $1-\delta_0$. Repeating the argument in the proof of Corollary 3, we derive

$$\mathbb{E}\|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 \lesssim \inf_{\boldsymbol{\beta}\in\mathbb{R}^p}\Big[\|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\beta}\| + \mu_{\boldsymbol{\beta}}\|\boldsymbol{\beta}\|_0\Big] + \sigma^2, \tag{57}$$

which establishes the first bound in the statement of Theorem 5.

Expected prediction error bound for $\hat{\boldsymbol{\beta}}_1^B$.

To simplify the expressions, we drop the subscript and the superscript in $\hat{\boldsymbol{\beta}}_1^B$ and simply write $\hat{\boldsymbol{\beta}}$.

In the case $b \gtrsim \sigma^2$ we repeat the argument in the corresponding part of the proof for $\hat{\boldsymbol{\beta}}_2^B$ to derive a counterpart of inequality (56). We deduce that, for each $\delta_0 \in (0,1)$,

$$\|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 + \mu_{\hat{\boldsymbol{\beta}}}|\|\hat{\boldsymbol{\beta}}\|_0 \lesssim \inf_{\boldsymbol{\beta}\in\mathbb{R}^p}\Big[\|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_1 + \mu_{\boldsymbol{\beta}}\|\boldsymbol{\beta}\|_0\Big] + \sigma^2 \log(1/\delta_0) \tag{58}$$

with probability at least $1 - \delta_0$.

In the case $\lambda \geq c_0 \sigma \sqrt{\log(ep)}$, we repeat the argument in the second slow rate part of the proof of Theorem 3 to derive a slight modification of inequality (38), containing the additional ℓ_0 penalty terms. Thus, we again deduce that inequality (58) holds for each $\delta_0 \in (0,1)$ with probability at least $1 - \delta_0$.

As before, starting with probability bound (58) and repeating the argument in the proof of Corollary 3 we conclude that

$$\mathbb{E}\|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 \lesssim \inf_{\boldsymbol{\beta} \in \mathbb{R}^p} \Big[\|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|_1 + \mu_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_0 \Big] + \sigma^2, \tag{59}$$

which establishes the second bound in the statement of Theorem 5.

B.14 Proof of Corollary 6

The error rates in Corollary 6 follow directly from inequalities (57) and (59).

To control the sparsity of $\hat{\boldsymbol{\beta}}_2^B$, we first establish that inequality $\|\hat{\boldsymbol{\beta}}_2^B\|_0 \lesssim (k^* \vee 1)\{1 + [\log(1/\delta_0)]^2\}$ holds with probability at least $1 - \delta_0$, and then use this fact to bound $\mathbb{E}\|\hat{\boldsymbol{\beta}}_2^B\|_0$.

We define $g(x) = \log(ep/x)x$ for $x \in [0,p]$, with $g(0) = 0$. We note that $g(x) \geq x$, and $g(x)$ is monotone increasing and continuous. We also note that if $C > 0$, $x_1 \in [0,p]$ and $x_2 \in [0,p]$, then

$$g(x_1) \leq C g(x_2) \quad \Rightarrow \quad x_1 \leq 2C[1 \vee \log(2C)]x_2. \tag{60}$$

To establish the above relationship we note that

$$g(x_1) \leq C g(x_2) \leq g(2Cx_2)/2 + Cx_2 \log(2C) \leq \max\{g(2Cx_2), 2Cx_2 \log(2C)\}$$

and consider two cases. If $g(x_1) \leq g(2Cx_2)$, then $x_1 \leq 2Cx_2$. Alternatively, if $g(x_1) \leq 2Cx_2 \log(2C)$, then $x_1 \leq 2C \log(2C) x_2$.

We write $\hat{k} = \|\hat{\boldsymbol{\beta}}_2^B\|_0$ and note that under each corresponding set of assumptions on the tuning parameters in Corollary 6, inequality (56) yields $(\mu_{\hat{\boldsymbol{\beta}}_2^B})\hat{k} \lesssim b \log(ep/[k^* \vee 1])[k^* \vee 1]\{1 + \log(1/\delta_0)\}$. Taking into account $(\mu_{\hat{\boldsymbol{\beta}}_2^B})\hat{k} = b g(\hat{k})$, we rewrite the last inequality as $g(\hat{k}) \lesssim g(k^* \vee 1)\{1 + \log(1/\delta_0)\}$. Hence, by property 60, we have $\hat{k} \lesssim \{1 + \log(1/\delta_0)\}[1 \vee \log(1 + 2\log(1/\delta_0))](k^* \vee 1)$. Consequently, $\hat{k}/[k^* \vee 1] \lesssim 1 + [\log(1/\delta_0)]^2$ with probability at least $1 - \delta_0$. Finally, we bound $\mathbb{E}\hat{k}/[k^* \vee 1]$ using an argument analogous to the one in the proof of Corollary 3: $\mathbb{E}\hat{k}/[k^* \vee 1] \lesssim 1 + \int_0^\infty e^{-w^{1/2}} dw \lesssim 1$.

To establish the sparsity bound for $\hat{\boldsymbol{\beta}}_1^B$, we note that for all of corresponding tuning parameter settings in Corollary 6, inequality (58) yields $(\mu_{\hat{\boldsymbol{\beta}}_1^B})\|\hat{\boldsymbol{\beta}}_1^B\|_0 \lesssim b\log(ep/[k^* \vee 1])[k^* \vee 1]\{1+\log(1/\delta_0)\}$, with probability at least $1-\delta_0$. The sparsity bound then follows by repeating the argument in the last paragraph of the proof for $\hat{\boldsymbol{\beta}}_2^B$.

C Additional experiments

C.1 Ultra-high dimensional examples

Figures 6 and 7 summarize the results of additional experiments corresponding to the challenging ultra-high dimensional setting [57] with $k^* \log(p/k^*) > n/2$. These experiments complement the ones that are reported in Figure 2. Qualitatively, the results are overall similar to those in Figure 2, especially with respect to the effect of adding ℓ_1 or ℓ_2 regularization to best subset selection. However, the predictive performance of all the methods in the challenging ultra-high dimensional setting is significantly worse than before, while the corresponding relative standing of dense models such as Ridge, Elastic net, and Lasso is improved.

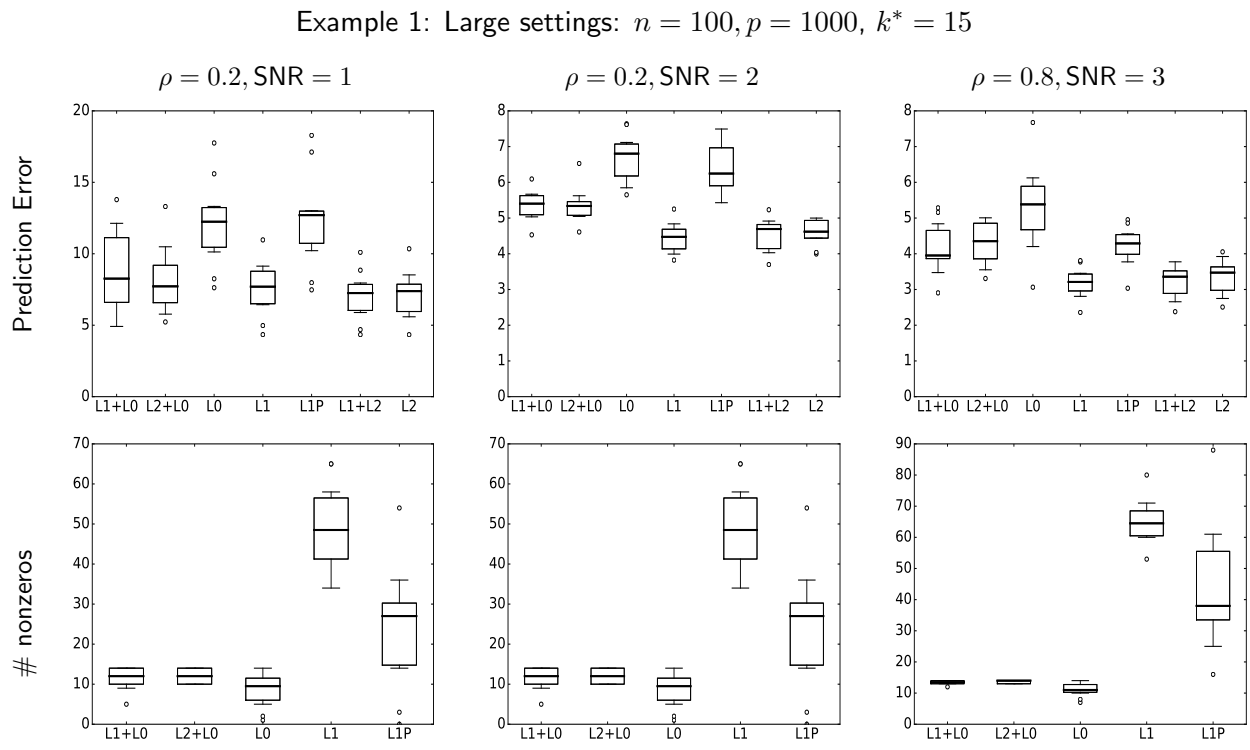


Figure 6: Example 1 simulations for different values of n, p, ρ, and SNR.

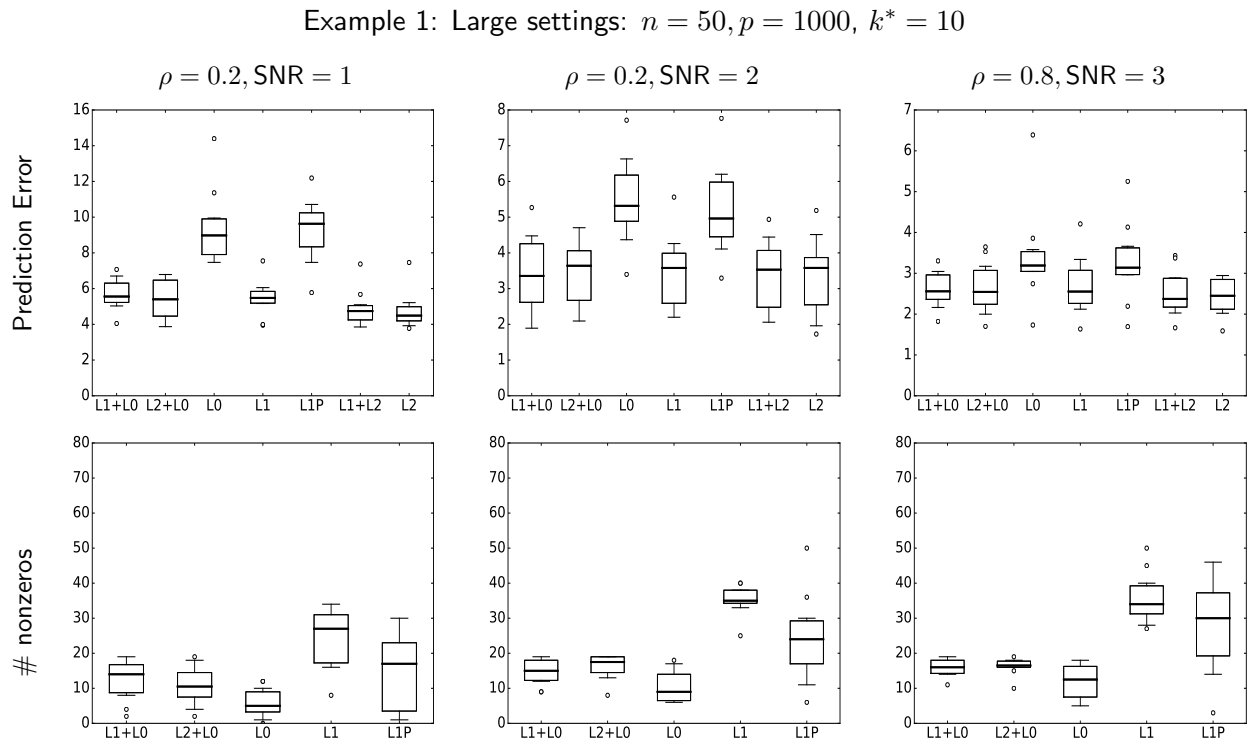
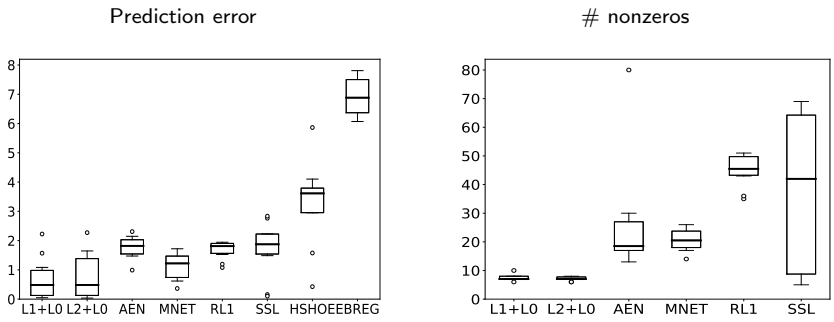


Figure 7: Example 1 simulations for different values of n, p, ρ, and SNR.

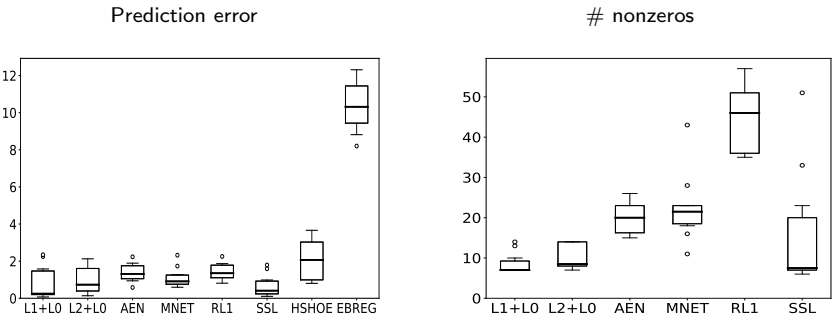
C.2 Comparisons with Bayesian methods

Figure 8 summarizes the results of additional experiments that include three state-of-the-art Bayesian approaches: the spike-and-slab Lasso method [53], implemented using R package `SSLASSO`; the empirical Bayes method of [38], implemented using R package `ebreg`; and the horseshoe regression [15], implemented using R package `horseshoe`. In the experiments that we consider, and with the default settings for the tuning parameters, the predictive performance of the last two methods is not quite as good as that of the competitors. The predictive performance of spike-and-slab Lasso is on par with the best performing methods (but somewhat worse overall than that of the proposed approach); however, their models are denser than those of the proposed approach. Overall, the proposed approach performed favorably in terms of both the model sparsity and the prediction accuracy. We note that the experiments in the top two panels of Figure 8 are the same as those in Figure 4; however, Figure 8 also includes the results for `horseshoe` and `ebreg`.

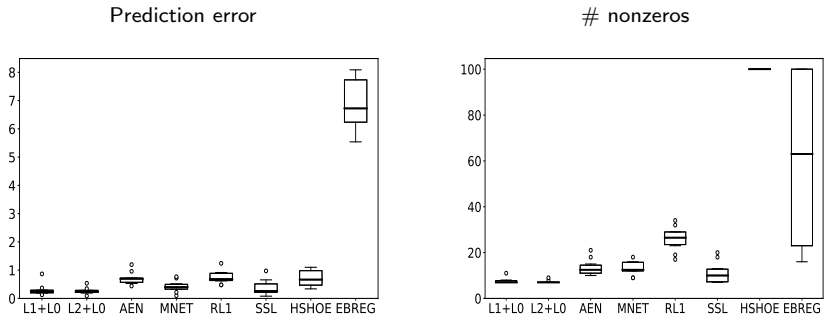
Example 1: $n = 100, p = 1000,\ \rho = 0.2$, SNR=2.



Example 2: $n = 100, p = 1000,\ \rho = 0.1$, SNR=3.



Example 1: $n = 100, p = 100,\ \rho = 0.2$, SNR=2.



Example 2: $n = 100, p = 100,\ \rho = 0.1$, SNR=3.

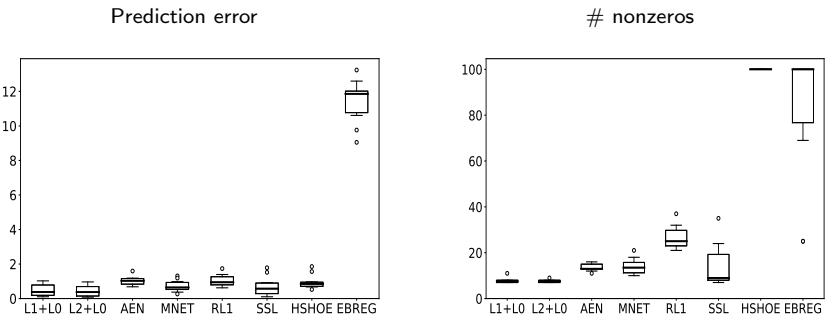


Figure 8: Experimental results for the proposed methods, L0+L1 and L0+L2, as well as adaptive elastic net (AEN), Mnet, relaxed Lasso (RL1), `SSLASSO` (SSL), `horseshoe` (HSHOE), and `ebreg` (EBREG) methods. Due to the density of the corresponding solutions, we do not report the sparsity for HSHOE and EBREG in the top two panels.