Journal of Global Optimization (2023) 85:315–349
https://doi.org/10.1007/s10898-022-01220-5

Sparse estimation via lower-order penalty optimization methods in high-dimensional linear regression

Xin Li[1] · Yaohua Hu[2] · Chong Li[3] · Xiaoqi Yang[4] · Tianzi Jiang[5]

Received: 2 September 2020 / Accepted: 28 July 2022 / Published online: 6 September 2022

Abstract
The lower-order penalty optimization methods, including the ℓ_q minimization method and the ℓ_q regularization method ($0 < q \leq 1$), have been widely applied to find sparse solutions of linear regression problems and gained successful applications in various mathematics and applied science fields. In this paper, we aim to investigate statistical properties of the ℓ_q penalty optimization methods with randomly noisy observations and a deterministic/random design. For this purpose, we introduce a general q-Restricted Eigenvalue Condition (REC) and provide its sufficient conditions in terms of several widely-used regularity conditions such as sparse eigenvalue condition, restricted isometry property, and mutual incoherence property. By virtue of the q-REC, we exhibit the ℓ_2 recovery bounds of order $O(\epsilon^2)$ and $O(\lambda^{\frac{2}{2-q}} s)$ for the ℓ_q minimization method and the ℓ_q regularization method, respectively, with high probability for either deterministic or random designs. The results in this paper are nonasymptotic and only assume the weak q-REC. The preliminary numerical results

✉ Yaohua Hu
mayhhu@szu.edu.cn

Xin Li
lixin@nwu.edu.cn

Chong Li
cli@zju.edu.cn

Xiaoqi Yang
mayangxq@polyu.edu.hk

Tianzi Jiang
jiangtz@nlpr.ia.ac.cn

1 School of Mathematics, Northwest University, Xi'an 710069, People's Republic of China

2 College of Mathematics and Statistics, Shenzhen Key Laboratory of Advanced Machine Learning and Applications, Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen University, Shenzhen 518060, P. R. China

3 School of Mathematical Sciences, Zhejiang University, Hangzhou 310027, People's Republic of China

4 Department of Applied Mathematics, The Hong Kong Polytechnic University, Kowloon, Hong Kong

5 Brainnetome Center, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, People's Republic of China

verify the established statistical properties and demonstrate the advantages of the ℓ_q penalty optimization methods over existing sparse optimization methods.

Keywords Sparse optimization · Lower-order penalty methods · Restricted eigenvalue condition · Recovery bound

1 Introduction

In various areas of applied sciences and engineering, a fundamental problem is to estimate an unknown parameter $\beta^* \in \mathbb{R}^n$ of a linear regression model

$$y = X\beta^* + e, \tag{1}$$

where $X \in \mathbb{R}^{m\times n}$ is a design matrix and $y \in \mathbb{R}^m$ is an observation vector with random noise $e \in \mathbb{R}^m$. In this paper, we assume that the random noise is Gaussian noise, i.e., $e \sim \mathcal{N}(0, \sigma^2\mathbb{I}_m)$. According to the context of practical applications, the design matrix could be either deterministic or random.

The curse of dimensionality always occurs in many application problems. For example, in magnetic resonance imaging [1], portfolio selection [2], systems biology [3], one is only able to collect limited samples of experimental data due to physical or economical constraints, i.e., $m \ll n$. In the high-dimensional scenario, it is a vital challenge to estimate the true underlying parameter of problem (1) because the corresponding linear system is seriously ill-conditioned and has infinitely many solutions.

1.1 ℓ_1 penalty optimization methods

Fortunately, in practical applications, a wide class of problems usually have certain special structures, which could help to eliminate the nonidentifiability and enhance the predictability. One of the most popular structures is the sparsity structure. One common way to measure the sparsity is the ℓ_q norm, which is defined by

$$\|\beta\|_q := \left(\sum_{i=1}^{n} |\beta_i|^q\right)^{1/q} \quad \text{for } 0 < q \le 1,$$

and $\|\beta\|_0$ is defined by the number of nonzero entries of β.

We first discuss the case when the design matrix X is deterministic. In the presence of a bounded noise (i.e., $\|e\|_2 \le \epsilon$), in order to find the sparest solution, [4] proposed the following (constrained) ℓ_0 minimization problem:

$$(\mathrm{CP}_{0,\epsilon}) \quad \min \|\beta\|_0 \quad \text{s.t.} \quad \|y - X\beta\|_2 \le \epsilon.$$

Unfortunately, it is NP-hard to compute its global solution due to the nonconvex and combinational natures [5].

To overcome this obstacle, a common technique is to use the (convex) ℓ_1 norm to approach the ℓ_0 norm:

$$(\mathrm{CP}_{1,\epsilon}) \quad \min \|\beta\|_1 \quad \text{s.t.} \quad \|y - X\beta\|_2 \le \epsilon,$$

which can be efficiently solved by convex optimization algorithms; see [6, 7] and references therein. In the practical applications, the amplitude of noise is usually difficult to estimate. In

such situations, the regularization method is a practical technique to avoid the noise estimation and has been widely used in optimization and statistics. Specifically, the ℓ_1 regularization problem is

$$(\mathrm{RP}_{1,\lambda}) \quad \min \frac{1}{2m}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1,$$

where $\lambda > 0$ is the regularization parameter, providing a tradeoff between data fidelity and sparsity. The ℓ_1 regularization problem, also named the least absolute shrinkage and selection operator (Lasso) in statistics [8], has attracted a great deal of attention in variable selection and gained wide applications in the high-dimensional scenario; see [9, 10] and references therein. In order to investigate the statistical properties of $(\mathrm{CP}_{1,\epsilon})$ and $(\mathrm{RP}_{1,\lambda})$, researchers have proposed several types of regularity conditions as follows.

Definition 1 (Restricted Isometry Constants (RICs), [11])

(i) The s-restricted isometry constant of X is denoted by $\eta_s(X)$ and defined to be the smallest quantity such that, for any $\beta \in \mathbb{R}^n$ and $J \subseteq \{1, \ldots, n\}$ with $|J| \leq s$,

$$(1 - \eta_s(X))\|\beta_J\|_2^2 \leq \|X\beta_J\|_2^2 \leq (1 + \eta_s(X))\|\beta_J\|_2^2. \tag{2}$$

(ii) The (s, t)-restricted orthogonality constant of X is denoted by $\theta_{s,t}(X)$ and defined to be the smallest quantity such that, for any $\beta \in \mathbb{R}^n$ and $J, T \subseteq \{1, \ldots, n\}$ with $|J| \leq s$, $|T| \leq t$ and $J \cap T = \emptyset$,

$$|\langle X\beta_J, X\beta_T\rangle| \leq \theta_{s,t}(X)\|\beta_J\|_2\|\beta_T\|_2. \tag{3}$$

Definition 2 (Mutual Incoherence Constant (MIC), [12]) The mutual incoherence constant of X is denoted by $M(X)$ and defined as

$$M(X) = \sup\{|X_{\cdot j}^\top X_{\cdot i}| : \forall 1 \leq i, j \leq n\}.$$

Definition 3 (Restricted Eigenvalue Condition (REC), [13]) X is said to satisfy the restricted eigenvalue condition if

$$\min_{\delta \in \mathbb{R}^n : \delta \neq 0} \left\{ \frac{\|X\delta\|_2}{\sqrt{m}\|\delta_{J \cup J(\delta;t)}\|_2} : |J| \leq s, \|\delta_{J^c}\|_1 \leq a\|\delta_J\|_1 \right\} > 0,$$

where $a > 0$ and (s, t) are a pair of integers such that $1 \leq s \leq t \leq n$, $s + t \leq n$, and $J(\delta; t)$ refers to the index set corresponding to the first t largest coordinates in absolute value of δ in J^c.

Definition 4 (Sparse Eigenvalue Condition (SEC), [14]) The s-sparse minimal eigenvalue and s-sparse maximal eigenvalue of X are respectively defined by

$$\sigma_{\min}(s, X) := \min_{\beta \in \mathbb{R}^n : 1 \leq \|\beta\|_0 \leq s} \frac{\beta^\top X^\top X\beta}{\beta^\top \beta}, \quad \sigma_{\max}(s, X) := \max_{\beta \in \mathbb{R}^n : 1 \leq \|\beta\|_0 \leq s} \frac{\beta^\top X^\top X\beta}{\beta^\top \beta}. \tag{4}$$

Conditions that the RIC or MIC is less than some positive constant are usually referred to the Restricted Isometry Property (RIP) or the Mutual Incoherence Property (MIP), respectively. It was claimed in [15] that the RIP can be implied by the MIP, while the RIC is more difficult to be calculated than the MIC. It was also reported in [13] that the REC can be implied by the RIP, and in [16] that a broad class of correlated Gaussian design matrices satisfy the REC but violate the RIP with high probability.

An important statistical property is the ℓ_2 recovery bound property, which aims to estimate an upper bound of the error between the global solution of the optimization problem and the

true underlying parameter in terms of the noise level ϵ. More specifically, let $s \ll n$ and β^* be an s-sparse solution of the linear regression problem (1). The ℓ_2 recovery bound of $(\mathrm{CP}_{1,\epsilon})$ was provided in [1] and [4] under the RIP and MIP, respectively; that is

$$\|\bar{\beta}_{1,\epsilon} - \beta^*\|_2^2 = O(\epsilon^2),$$

where $\bar{\beta}_{1,\epsilon}$ stands for the global solution of $(\mathrm{CP}_{1,\epsilon})$. In the noise-free case, the ℓ_2 recovery bound of $(\mathrm{RP}_{1,\lambda})$ was provided in [17] under the RIP or the REC:

$$\|\hat{\beta}_{1,\lambda} - \beta^*\|_2^2 = O(\lambda^2 s),$$

where $\hat{\beta}_{1,\lambda}$ denotes the global solution of $(\mathrm{RP}_{1,\lambda})$. Furthermore, when the noise is normally distributed as $e \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_m)$, it was established in [13, 18, 19] that the following ℓ_2 recovery bound holds with high probability under the RIP, REC or other regularity conditions:

$$\|\hat{\beta}_{1,\lambda} - \beta^*\|_2^2 = O\left(\sigma^2 s \frac{\log n}{m}\right),$$

when the regularization parameter is chosen as $\lambda = \sigma\sqrt{\frac{\log n}{m}}$. However, the ℓ_1 penalty optimization methods, including the ℓ_1 minimization method and the ℓ_1 regularization method, suffer from several dissatisfactions in both theoretical properties and practical applications. In detail, it was reported by extensive theoretical and empirical studies that the ℓ_1 penalty optimization methods suffer from significant estimation bias when parameters have large absolute values; the induced solution is much less sparse than the true parameter leading to sub-optimal sparsity in practice; they cannot recover a sparse signal with the least samples when applied to compressed sensing; see, e.g., [20–25]. Therefore, there is a great demand for developing an alternative sparse estimation technique that enjoys nice statistical theory and successful applications.

1.2 ℓ_q penalty optimization methods

To address the bias and the sub-optimal issues induced by the ℓ_1 penalty optimization methods, several nonconvex regularizers have been proposed to improve the sparsity-promoting capability. One of the most important nonconvex approaches is the ℓ_q $(0 < q < 1)$ penalty optimization methods:

$$(\mathrm{CP}_{q,\epsilon}) \quad \min \|\beta\|_q \quad \text{s.t.} \quad \|y - X\beta\|_2 \le \epsilon,$$

and

$$(\mathrm{RP}_{q,\lambda}) \quad \min \frac{1}{2m}\|y - X\beta\|_2^2 + \lambda\|\beta\|_q^q.$$

The numerical results in [20] and [25] showed that the ℓ_q minimization and the $\ell_{\frac{1}{2}}$ regularization methods admit a significantly stronger sparsity-promoting capability than the ℓ_1 minimization and the ℓ_1 regularization methods, respectively, in the sense that they allow to obtain a more sparse solution from a smaller amount of samples. [3] revealed that the $\ell_{\frac{1}{2}}$ regularization method achieved a more reliable biological solution than the ℓ_1 regularization method in gene regulatory network inference. In view of the lower-order penalty methods as investigated in [26–29], a main advantage of the lower-order penalty functions over the classical ℓ_1 penalty functions is that they require weaker conditions to guarantee an exact penalization property and that their least exact penalty parameter is smaller. Nowadays, the

ℓ_q penalty optimization methods have widely applied and gained successful applications in a wide range of fields.

The advantage of the ℓ_q penalty optimization methods has also been shown in theory that they require a weaker regularity condition to guarantee the stable statistical property than the classical ℓ_1 penalty optimization methods. In detail, let $\bar{\beta}_{q,\epsilon}$ and $\hat{\beta}_{q,\lambda}$ denote the global solution of the ℓ_q constrained minimization problem ($\mathrm{CP}_{q,\epsilon}$) and the ℓ_q regularization problem ($\mathrm{RP}_{q,\lambda}$), respectively. The ℓ_2 recovery bound of ($\mathrm{CP}_{q,\epsilon}$) was established in [30] and [31] under MIP (cf. Definition 2) and RIP (cf. Definition 1), respectively, that

$$\|\bar{\beta}_{q,\epsilon} - \beta^*\|_2^2 = O(\epsilon^2), \tag{5}$$

where the MIP or RIP is weaker than the one for ($\mathrm{CP}_{1,\epsilon}$). [32] established an ℓ_2 recovery bound of ($\mathrm{RP}_{q,\lambda}$) in the noise-free case that

$$\|\hat{\beta}_{q,\lambda} - \beta^*\|_2^2 = O(\lambda^{\frac{2}{2-q}} s) \tag{6}$$

under an introduced q-REC, which is weaker than the classical REC (cf. Definition 3). However, the theoretical study for the ℓ_q penalty optimization methods is still limited; particularly, there is still no paper devoted to establishing the recovery bound property of the ℓ_q minimization method when the noise is randomly distributed, and that of the ℓ_q regularization method in the noise-aware case.

Besides the ℓ_q penalty optimization methods, there are several popular nonconvex regularizers, including the Smoothly Clipped Absolute Deviation (SCAD) [21], Minimax Concave Penalty (MCP) [24], Folded Concave Penalty (FCP) [33], and capped ℓ_1 norm [34]. Specifically, fixing $a > 2$ and $b > 0$, the SCAD regularizer is defined as $\mathcal{R}_{\mathrm{SCAD},\lambda}(\beta) := \sum_{j=1}^n f_{\mathrm{SCAD},\lambda}(\beta_j)$ with

$$f_{\mathrm{SCAD},\lambda}(t) := \begin{cases} \lambda|t|, & \text{if } |t| \le \lambda, \\ -\frac{t^2 - 2a\lambda|t| + \lambda^2}{2(a-1)}, & \text{if } \lambda < |t| \le a\lambda, \\ \frac{(a+1)\lambda^2}{2}, & \text{if } |t| > a\lambda, \end{cases} \tag{7}$$

and the MCP regularizer is defined as $\mathcal{R}_{\mathrm{MCP},\lambda}(\beta) := \sum_{j=1}^n f_{\mathrm{MCP},\lambda}(\beta_j)$ with

$$f_{\mathrm{MCP},\lambda}(t) := \begin{cases} \lambda|t| - \frac{t^2}{2b}, & \text{if } |t| \le b\lambda, \\ \frac{b\lambda^2}{2}, & \text{if } |t| > b\lambda, \end{cases} \tag{8}$$

In particular, SCAD and MCP fall into the category of FCP defined as $\mathcal{R}_{\mathrm{FCP},\lambda}(\beta) = \sum_{j=1}^n f_{\mathrm{FCP},\lambda}(|\beta_j|)$, where $f_{\mathrm{FCP},\lambda}(\cdot)$ is defined on $[0, \infty)$ and satisfies the following assumption.

Assumption 1 (i) $f_{\mathrm{FCP},\lambda}(t)$ is increasing and concave in $t \in [0, \infty)$ with $f_{\mathrm{FCP},\lambda}(0) = 0$;
(ii) $f_{\mathrm{FCP},\lambda}(t)$ is differentiable in $t \in (0, \infty)$ with $f'_{\mathrm{FCP},\lambda}(0+) \ge a_1\lambda$;
(iii) $f'_{\mathrm{FCP},\lambda}(t) \ge a_1\lambda$ for $t \in (0, a_2\lambda]$;
(iv) $f'_{\mathrm{FCP},\lambda}(t) = 0$ for $t \in [a\lambda, \infty)$ with a pre-specified constant $a > a_2$.

It was studied in [35] that the global solution of the FCP linear regression enjoys the oracle property under the SEC (cf. Definition 4). Nevertheless, the ℓ_q penalty optimization methods are beyond the category of the FCP considered in [35]; consequently, this work is not applicable to provide the oracle property for the ℓ_q penalty optimization methods.

1.3 Contributions of this paper

The main contribution of this paper is to establish the statistical properties of the ℓ_q penalty optimization methods, including the ℓ_q constrained minimization problem ($\mathrm{CP}_{q,\epsilon}$) and the ℓ_q regularization problem ($\mathrm{RP}_{q,\lambda}$), in the noise-aware case; specifically, in the case when the linear regression model (1) involves a Gaussian noise as $e \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_m)$. For this purpose, we first extend the q-REC [32] to a more general one, which is one of the weakest regularity conditions to ensure the ℓ_2 recovery bounds of sparse regression models, and provide some sufficient conditions for guaranteeing the general q-REC in terms of REC, RIP, and MIP (with a less restrictive constant); see Propositions 1 and 2. Under the general q-REC, we show that the ℓ_2 recovery bounds (5) and

$$\|\hat{\beta}_{q,\lambda} - \beta^*\|_2^2 = O\left(\left(\sigma^2 \frac{\log n}{m}\right)^{\frac{1}{2-q}} s\right)$$

hold with high probability for ($\mathrm{CP}_{q,\epsilon}$) and ($\mathrm{RP}_{q,\lambda}$), respectively; see Theorems 1 and 2. These results provide a unified framework for the statistical properties of the ℓ_q penalty optimization methods, and improve the ones of the ℓ_q minimization method [30, 31] and the ℓ_1 regularization method [13, 18, 19] under a weak q-REC; see Remark 5. They are not only of independent interest in establishing statistical properties of the lower-order penalty optimization methods with randomly noisy data, but also provide a useful tool for the study of the random design case.

Another contribution of the present paper is to explore the ℓ_2 recovery bounds of the ℓ_q penalty optimization methods with a random design matrix X and random noise e. This case is more realistic in real-world applications; e.g., compressed sensing [36], signal processing [1], statistical learning [37]. Motivated by real-world applications, we consider the common case when X is a Gaussian random design with i.i.d. $\mathcal{N}(0, \Sigma)$ rows. For this case, we explore a sufficient condition for ensuring the q-REC of X with high probability by virtue of Σ, and apply the preceding results to establish the ℓ_2 recovery bounds (5) and (6) for ($\mathrm{CP}_{q,\epsilon}$) and ($\mathrm{RP}_{q,\lambda}$), respectively; see Theorems 3 and 4. These results provide a unified framework for the statistical properties of the ℓ_q penalty optimization methods with a Gaussian random design under the q-REC, which cover the ones of the ℓ_1 penalty optimization methods (see [38, Theorem 3.1]) as special cases; see Corollaries 3 and 4. To the best of our knowledge, most results presented here are new, either for the deterministic or random design.

We also carry out numerical experiments on standard simulated data. The preliminary numerical results verify the established statistical properties and show that the ℓ_q penalty optimization methods possess better recovery performance than the ℓ_1 penalty optimization methods, SCAD (7) and MCP (8), which coincides with existing numerical studies [25, 32] on the ℓ_q regularization method.

The remainder of this paper is organized as follows. In Sect. 2, we introduce the lower-order REC and discuss its sufficient conditions. The ℓ_2 recovery bounds for the ℓ_q penalty optimization methods with the deterministic and random designs are presented in Sects. 3 and 4, respectively. The preliminary numerical results are illustrated in Sect. 5. Preliminary results and technical proofs of are deferred to Appendix.

We end this section by presenting the notations adopted in this paper. We use Greek lowercase letters α, β, δ to denote the vectors, capital letters J, T to denote the index sets, and script capital letters $\mathscr{A}, \mathscr{B}, \mathscr{C}$ to denote the random events. For $\beta \in \mathbb{R}^n$ and $J \subseteq \{1, 2, \ldots, n\}$, we use β_J to denote the vector in $\mathbb{R}^n$ with $(\beta_J)_i = \beta_i$ for $i \in J$ and zero elsewhere, $|J|$ to denote the cardinality of J, J^c to denote the complement of J, and $\mathrm{supp}(\beta)$ to denote

the support of β, i.e., the index set of nonzero entries of β. For a matrix $X \in \mathbb{R}^{m\times n}$, let X_{ij} $(i = 1,\ldots,m, j = 1,2,\cdots,n)$ denote its (i,j)-th entry, $X_{i\cdot}$ $(i = 1,\ldots,m)$ denote its i-th row, $X_{\cdot j}$ $(j = 1,2,\cdots,n)$ denote its j-th column. As usual, $\mathbb{I}_m$ stands for the identity matrix in $\mathbb{R}^{m\times m}$, and $\mathbb{P}(\mathscr{A})$ and $\mathbb{P}(\mathscr{A}|\mathscr{B})$ denote the probability of event $\mathscr{A}$ and the conditional probability event $\mathscr{A}$ provided event $\mathscr{B}$, respectively. Throughout the whole paper, we always assume that $0 < q \le 1$ unless otherwise specified.

2 Restricted eigenvalue conditions

This section aims to discuss some regularity conditions imposed on the design matrix that are required to guarantee the stable statistical properties of the ℓ_q constrained minimization problem (CP$_{q,\epsilon}$) and the ℓ_q regularization problem (RP$_{q,\lambda}$).

The ordinary least squares (OLS) is a classical technique to estimate the unknown parameters in a linear regression (1) and has favourable properties if some regularity conditions are satisfied; see, e.g., [39]. For example, the OLS always requires the positive definiteness of the Gram matrix $X^\top X$, that is,

$$\min_{\beta\in\mathbb{R}^n:\beta\neq 0} \frac{(\beta^\top X\beta)^{1/2}}{\|\beta\|_2} = \min_{\beta\in\mathbb{R}^n:\beta\neq 0} \frac{\|X\beta\|_2}{\|\beta\|_2} > 0. \tag{9}$$

However, the OLS does not work well in the high-dimensional scenario because the associated Gram matrix is always seriously degenerate, i.e.,

$$\min_{\beta\in\mathbb{R}^n:\beta\neq 0} \frac{\|X\beta\|_2}{\|\beta\|_2} = 0.$$

To deal with the challenges caused by the high-dimensional data, the ℓ_1 regularization method (also named Lasso) was introduced by [8] and has gained a great success in sparse representation of high-dimensional data; see, e.g., [13, 19, 40] and references therein. It was pointed out that the ℓ_1 regularization method requires a weak REC (cf. Definition 3) [13] to ensure nice statistical properties; see, e.g., [17, 41, 42]. In the definition of REC, the minimum in (9) is replaced by a minimum over a restricted set of vectors measured by an ℓ_1 norm inequality, and the norm $\|\beta\|_2$ in the denominator is replaced by the ℓ_2 norm of only a part of β. The notion of REC was extended to the group-wised lower-order REC in [32], which was used there to explore the oracle property and ℓ_2 recovery bound of the $\ell_{p,q}$ regularization problem in the noise-free case.

Inspired by the ideas in [13] and [32], we here introduce a lower-order REC for the ℓ_q optimization problems, similar to but more general than the one in [32], where the minimum is taken over a restricted set of vectors measured by an ℓ_q norm inequality. To proceed, we shall introduce some notations used in the lower-order REC. In the remainder of this paper, let $a > 0$ and (s,t) be a pair of integers such that

$$1 \le s \le t \le n \quad \text{and} \quad s + t \le n. \tag{10}$$

For $\delta \in \mathbb{R}^n$ and $J \subseteq \{1,2,\ldots,n\}$, we define by $J(\delta;t)$ the index set corresponding to the first t largest coordinates in absolute value of δ in J^c. For $X \in \mathbb{R}^{m\times n}$, its q-restricted eigenvalue modulus relative to (s,t,a) is defined by

$$\phi_q(s,t,a,X) := \min_{\delta\in\mathbb{R}^n:\delta\neq 0} \left\{ \frac{\|X\delta\|_2}{\sqrt{m}\|\delta_{J\cup J(\delta;t)}\|_2} : |J| \le s, \|\delta_{J^c}\|_q^q \le a\|\delta_J\|_q^q \right\}. \tag{11}$$

Definition 5 (q-Restricted Eigenvalue Condition (q-REC)) Let $0 \le q \le 1$ and $X \in \mathbb{R}^{m\times n}$. X is said to satisfy the q-restricted eigenvalue condition relative to (s, t, a) (q-REC(s, t, a) for short) if

$$\phi_q(s, t, a, X) > 0.$$

Remark 1 Clearly, the q-REC(s, t, a) in Definition 5 provides a unified framework for the REC-type conditions, e.g., it includes the classical REC in [13] (when $q = 1$) and the q-REC(s, t) in [32] (when $a = 1$) as special cases. Particularly, the q-REC(s, t, a) extends the classical REC (that is $q = 1$) in [13] to the lower-order case, where the latter one is weaker (see Proposition 1 below) and applicable to establish nice statistical properties of the ℓ_q sparse optimization problems. Moreover, the q-REC(s, t, a) in Definition 5 is an extension of the q-REC (that is $a = 1$) in [32] to the general $a > 0$, where the latter one with $a > 1$ is stronger but required for the establishment of nice statistical properties for the ℓ_q regularization problem of noisy linear regression. (The q-REC in [32] was applied there to explore the oracle property and recovery bound for the ℓ_q regularization in the noiseless case.) This is because, in the noisy case, the dominant property (24) in Proposition 4 is satisfied only when $a > 1$.

It is natural to study the relationships between the q-RECs and other types of regularity conditions. First, by extending [32, Proposition 5] to the general q-REC, we validate the relationship between the q-RECs in the following proposition: the lower the q, the weaker the q-REC. However, the inverse of this implication is not true; see [32, Example 1] for a counter example.

Proposition 1 *Suppose that* $0 < q_1 \le q_2 \le 1$ *and that* X *satisfies the* q_2-REC(s, t, a). *Then* X *satisfies the* q_1-REC(s, t, a).

It is revealed from Proposition 1 that the classical REC is a sufficient condition for the lower-order REC. In the sequel, we will further discuss some other types of regularity conditions: SEC (cf. Definition 4), RIP (cf. Definition 1), and MIP (cf. Definition 2), to ensure the lower-order REC.

The SEC is a popular regularity condition to guarantee nice properties of sparse representation; see [13, 14, 35] and references therein. The SEC was first introduced in [14] to show that the global solution of ($\mathrm{CP}_{1,\epsilon}$) approximates that of ($\mathrm{CP}_{0,\epsilon}$) whenever $\sigma_{\min}(2s, X) > 0$. The RIP is a well-known regularity condition in the scenario of sparse learning, which was introduced by [11] and has been widely used in the study of oracle property and recovery bound for the high-dimensional regression model; see [1, 13, 43] and references therein. The MIP is another well-known regularity condition in the scenario of sparse learning, which was introduced by [12] and has been used in [4, 13–15] and references therein. In the case when each diagonal element of X is 1, $\theta_{1,1}(X)$ coincides with the MIC; see [12].

The following proposition provides the sufficient conditions for the q-REC in terms of SEC, RIP and MIP; see terms (a), (b) and (c) below respectively.

Proposition 2 *Let* $X \in \mathbb{R}^{m\times n}$, $0 < q \le 1$, $a > 0$, *and* (s, t) *be a pair of integers satisfying (10). Then* X *satisfies the* q-REC(s, t, a) *provided one of the following conditions:*

(a) $\sigma_{\min}(s + t, X) > a\left(\frac{as}{t}\right)^{\frac{2}{q}-1}\sigma_{\max}(t, X)$.

(b) $\eta_t(X) + \theta_{s,t}(X) + a^{\frac{1}{2}}\left(\frac{as}{t}\right)^{\frac{1}{q}-\frac{1}{2}}\theta_{t,s+t}(X) < 1$.

(c) each diagonal element of $\frac{X^\top X}{m}$ *is* 1 *and* $\theta_{1,1}(X) < \frac{1}{s+t}\left(1 + 2a\left(\frac{as}{t}\right)^{\frac{1}{q}-1}\right)^{-1}$.

Remark 2 It was established in [13, Lemma 4.1(ii)], [17, Corollary 7.1 and 3.1] and [13, Assumption 5] that X satisfies the classical *REC* under one of the following conditions:

(a′) $\sigma_{\min}(s+t, X) > \frac{s}{t}a^2\sigma_{\max}(t, X)$.

(b′) $\eta_t(X) + \theta_{s,t}(X) + \left(\frac{s}{t}\right)^{\frac{1}{2}} a\theta_{t,s+t}(X) < 1$.

(c′) each diagonal element of $\frac{X^\top X}{m}$ is 1 and $\theta_{1,1}(X) < \frac{1}{(s+t)(1+2a)}$.

Proposition 2 extends these results to the general case when $0 < q \leq 1$ and partially improves them; in particular, each of conditions (a)–(c) in Proposition 2 required for the q-REC is less restrictive than the corresponding one of conditions (a')–(c') required for the classical REC in the situation when $a < t/s$, which usually occurs in the high-dimensional scenario (see, e.g., [1, 13, 38]). Moreover, by Propositions 1 and 2, we achieve that the q-REC(s, t, a) is satisfied provided that one of the following conditions:

(a°) $\sigma_{\min}(s+t, X) > \min\left\{1, \left(\frac{as}{t}\right)^{\frac{2}{q}-2}\right\} \frac{s}{t}a^2\sigma_{\max}(t, X)$.

(b°) $\eta_t(X) + \theta_{s,t}(X) + \min\left\{1, \left(\frac{as}{t}\right)^{\frac{1}{q}-1}\right\} \left(\frac{s}{t}\right)^{\frac{1}{2}} a\theta_{t,s+t}(X) < 1$.

(c°) each diagonal element of $\frac{X^\top X}{m}$ is 1 and $\theta_{1,1}(X) < \frac{1}{s+t}\left(1 + 2a\min\left\{1, \left(\frac{as}{t}\right)^{\frac{1}{q}-1}\right\}\right)^{-1}$.

3 Recovery bounds for deterministic design

This section is devoted to establishing the ℓ_2 recovery bounds for the ℓ_q constrained minimization problem (CP$_{q,\epsilon}$) and the ℓ_q regularization problem (RP$_{q,\lambda}$) in the case when X is deterministic. Recall that $e \sim \mathcal{N}(0, \sigma^2\mathbb{I}_m)$, and adopt the following notations:

$$\text{let } \beta^* \text{ be a solution of (1), } J := \text{supp}(\beta^*), s := |J|, \text{ and let } t \in \mathbb{N} \text{ satisfy (10).}$$

The ℓ_2 recovery bound of the ℓ_1 regularization problem was established in [13] under the classical REC (cf. Definition 3). The deduction is based on an important property of the global solution. More precisely, let $\bar{\beta}_{1,\epsilon}$ and $\hat{\beta}_{1,\lambda}$ be the global solutions of the ℓ_1 minimization problem and the ℓ_1 regularization problem, respectively. It was reported in [1, Eq. (2.2)] and [13, Corollary B.2] that the corresponding residuals satisfy the following dominant properties with high probability,

$$\|(\bar{\beta}_{1,\epsilon} - \beta^*)_{J^c}\|_1 \leq \|(\bar{\beta}_{1,\epsilon} - \beta^*)_J\|_1$$

and

$$\|(\hat{\beta}_{1,\lambda} - \beta^*)_{J^c}\|_1 \leq 3\|(\hat{\beta}_{1,\lambda} - \beta^*)_J\|_1$$

for the ℓ_1 minimization problem and the ℓ_1 regularization problem, respectively. Here the term "high probability" means that the probability tends to 1 as long as the sample size m and/or the problem dimension n tend to infinity.

In the study of the ℓ_q penalty optimization methods, a natural question arises whether the residuals of global solutions of the ℓ_q constrained minimization problem (CP$_{q,\epsilon}$) or the ℓ_q regularization problem (RP$_{q,\lambda}$) satisfy such a dominant property on the support of the true underlying parameter of linear regression (1) with high probability. Below, we provide a positive answer for this question in Propositions 3 and 4.

To this end, some useful notations are provided. In the remainder of this paper, let

$$a > 1, \quad 0 \le \theta < 1, \quad b \ge 0, \tag{12}$$

unless otherwise specified, and let $r > 0$ be such that

$$r \ge \|\beta^*\|_q. \tag{13}$$

Let

$$\epsilon := \sigma\sqrt{5m} \quad \text{and} \quad \rho := \left(\frac{5\sigma^2}{2\lambda} + r^q\right)^{1/q}, \tag{14}$$

and select the regularization parameter in $(\mathrm{RP}_{q,\lambda})$ as

$$\lambda = \max\left\{\frac{a+1}{a-1}\sigma(1+\theta)2^{1-q}(1+r^q)^{\frac{1-q}{q}}\sqrt{\frac{2(1+b)\log n}{m}},\ \frac{5}{2}\sigma^2\right\}. \tag{15}$$

Define the following two random events relative to linear regression model (1) by

$$\mathscr{A} := \{e : \|e\|_2 \le \epsilon\} \tag{16}$$

and

$$\mathscr{B} := \left\{e : \frac{a+1}{(a-1)m}(2\rho)^{1-q}\|X^\top e\|_\infty \le \lambda\right\}. \tag{17}$$

The following lemma estimates the probabilities of events $\mathscr{A}$ and $\mathscr{B}$. From (18) to (21), one can see that the events $\mathscr{A}$ and $\mathscr{A} \cap \mathscr{B}$ happen with probability achieving 1 as long as the sample size m and/or the problem dimension n tend to infinity.

Lemma 1 *The probability of event $\mathscr{A}$ satisfies*

$$\mathbb{P}(\mathscr{A}) \ge 1 - \exp(-m). \tag{18}$$

Moreover, suppose that X satisfies

$$\max_{1\le j\le n}\|X_{\cdot j}\|_2 \le (1+\theta)\sqrt{m}. \tag{19}$$

Then

$$\mathbb{P}(\mathscr{B}) \ge 1 - \left(n^b\sqrt{\pi\log n}\right)^{-1}, \tag{20}$$

$$\mathbb{P}(\mathscr{A}\cap\mathscr{B}) \ge 1 - \exp(-m) - \left(n^b\sqrt{\pi\log n}\right)^{-1}. \tag{21}$$

We show in the following two propositions that the global solutions of the ℓ_q constrained minimization problem $(\mathrm{CP}_{q,\epsilon})$ and the ℓ_q regularization problem $(\mathrm{RP}_{q,\lambda})$ satisfy the dominant property on the support of the true underlying parameter with high probability:

$$\|(\hat\beta - \beta^*)_{J^c}\|_q^q \le c\|(\hat\beta - \beta^*)_J\|_q^q$$

with $c = 1$ or $c = a$, respectively.

Proposition 3 *Let $\bar\beta_{q,\epsilon}$ be a global solution of $(\mathrm{CP}_{q,\epsilon})$ with ϵ given by (14). Then it holds under the event $\mathscr{A}$ that*

$$\|(\bar\beta_{q,\epsilon} - \beta^*)_{J^c}\|_q \le \|(\bar\beta_{q,\epsilon} - \beta^*)_J\|_q. \tag{22}$$

Proposition 4 *Let $\hat{\beta}_{q,\lambda}$ be a global solution of* $(\mathrm{RP}_{q,\lambda})$ *with* λ *given by (15). Suppose that (19) is satisfied. Then*

$$\|\hat{\beta}_{q,\lambda} - \beta^*\|_1 \leq (2\rho)^{1-q}\|\hat{\beta}_{q,\lambda} - \beta^*\|_q^q \tag{23}$$

under the event $\mathscr{A}$, and

$$\|(\hat{\beta}_{q,\lambda} - \beta^*)_{J^c}\|_q^q \leq a\|(\hat{\beta}_{q,\lambda} - \beta^*)_J\|_q^q \tag{24}$$

under the event $\mathscr{A} \cap \mathscr{B}$.

Remark 3 By Lemma 1, Propositions 3 and 4 show that (22) holds with probability at least $1 - \exp(-m)$, and (24) holds with probability at least $1 - \exp(-m) - \left(n^b\sqrt{\pi \log n}\right)^{-1}$ if (19) is satisfied, respectively.

By virtue of Lemma 1 and Proposition 3, one of the main theorems of this section is as follows, in which we establish the ℓ_2 recovery bound for the ℓ_q constrained minimization problem $(\mathrm{CP}_{q,\epsilon})$ under the weak q-REC (cf. Definition 5). This theorem shows that one can stably recover the underlying parameter with high probability via solving the ℓ_q constrained minimization problem when the design matrix satisfies the weak q-REC.

Theorem 1 *Let $\bar{\beta}_{q,\epsilon}$ be a global solution of* $(\mathrm{CP}_{q,\epsilon})$ *with ϵ given by (14). Suppose that X satisfies the q-REC(s, t, a) with $a \geq 1$. Then, with probability at least $1 - \exp(-m)$, we have that*

$$\|\bar{\beta}_{q,\epsilon} - \beta^*\|_2^2 \leq \frac{4\left(1 + \left(\frac{s}{t}\right)^{\frac{2}{q}-1}\right)}{m\phi_q^2(s, t, a, X)}\epsilon^2. \tag{25}$$

Remark 4 (i) As shown in Proposition 3, the global solution of the ℓ_q constrained minimization problem $(\mathrm{CP}_{q,\epsilon})$ satisfies the dominant property on the support of the true underlying parameter with high probability, in which the coefficient in the right hand side of (22) is 1. Thus the q-REC(s, t, a) with $a = 1$ is enough to guarantee the recovery bound results. Particularly, in the special case when $a = 1$, the q-REC$(s, t, 1)$ is a weaker condition and (25) provides a tighter bound than the ones as $a > 1$. This phenomenon also occurs for Theorem 3 and Corollaries 1 and 3.

(ii) In the special case when the underlying data is noise-free, Theorem 1 shows that $(\mathrm{CP}_{q,\epsilon})$ can exactly predict the parameter of the linear regression with high probability under the lower-order REC. For the realistic scenario where the measurements are noisy, Theorem 1 illustrates the stable recovery capability of $(\mathrm{CP}_{q,\epsilon})$ in the sense that its global solution approaches to the true sparse parameter within a tolerance proportional to the noise level with high probability. Moreover, Theorem 1 establishes the ℓ_2 recovery bound of order $O(\epsilon^2)$ under the q-REC, which is weaker than the RIP-type or MIP-type condition used in [30, 31] to obtain the same ℓ_2 recovery bound of $(\mathrm{CP}_{q,\epsilon})$, respectively.

As a special case of Theorem 1 when $q = 1$, the following corollary presents the ℓ_2 recovery bound of the ℓ_1 minimization problem $(\mathrm{CP}_{1,\epsilon})$ as

$$\|\bar{\beta}_{1,\epsilon} - \beta^*\|_2^2 = O(\epsilon^2) \tag{26}$$

under the classical REC. This result improves the ones in [1, 15] under a weaker assumption, in which the ℓ_2 recovery bound (26) was obtained under the RIP-type conditions.

Corollary 1 *Let $\bar{\beta}_{1,\epsilon}$ be a global solution of* ($\mathrm{CP}_{1,\epsilon}$) *with ϵ given by (14). Suppose that X satisfies the* 1-REC(s,t,a) *with $a \geq 1$. Then, with probability at least* $1-\exp(-m)$*, we have that*

$$\|\bar{\beta}_{1,\epsilon} - \beta^*\|_2^2 \leq \frac{1+\frac{s}{t}}{m\phi_1^2(s,t,a,X)} 4\epsilon^2.$$

The other main theorem of this section is as follows, in which we exploit the statistical properties of the ℓ_q regularization problem ($\mathrm{RP}_{q,\lambda}$) under the q-REC. The results include the estimation of prediction loss and recovery bound of parameter approximation, and also the oracle property, which provides an upper bound on the prediction loss plus the violation of false variable selection.

Theorem 2 *Let $\hat{\beta}_{q,\lambda}$ be a global solution of* ($\mathrm{RP}_{q,\lambda}$) *with λ given by (15). Suppose that X satisfies the q-*REC(s,t,a) *with $a > 1$ and that (19) is satisfied. Then, with probability at least* $1-\exp(-m)-\left(n^b\sqrt{\pi\log n}\right)^{-1}$*, we have that*

$$\frac{1}{m}\|X\hat{\beta}_{q,\lambda} - X\beta^*\|_2^2 \leq \left(\frac{2a\lambda}{\phi_q^q(s,t,a,X)}\right)^{\frac{2}{2-q}} s, \tag{27}$$

$$\frac{1}{2m}\|X\hat{\beta}_{q,\lambda} - X\beta^*\|_2^2 + \lambda\|(\hat{\beta}_{q,\lambda})_{J^c}\|_q^q \leq \left(\frac{2^{\frac{q}{2}}a\lambda}{\phi_q^q(s,t,a,X)}\right)^{\frac{2}{2-q}} s, \tag{28}$$

$$\|\hat{\beta}_{q,\lambda} - \beta^*\|_2^2 \leq \left(1 + a^{\frac{2}{q}}\left(\frac{s}{t}\right)^{\frac{2}{q}-1}\right)\left(\frac{2a\lambda}{\phi_q^2(s,t,a,X)}\right)^{\frac{2}{2-q}} s. \tag{29}$$

Remark 5 (i) Theorem 2 provides a unified framework for the statistical properties of the ℓ_q regularization problem under the weak q-REC that is one of the weakest regularity conditions in the literature, in which each of the obtained estimations depends on the noise amplitude and sample size. In particular, for the regularization parameter scaling as $\lambda \asymp \max\left(\sigma\sqrt{\frac{\log n}{m}}, \sigma^2\right)$[1] (cf. (15)), Theorem 2 indicates the prediction loss and the ℓ_2 recovery bound of ($\mathrm{RP}_{q,\lambda}$) scale as

$$\frac{1}{m}\|X\hat{\beta}_{q,\lambda} - X\beta^*\|_2^2 = O\left(\left(\sigma^2\frac{\log n}{m}\right)^{\frac{1}{2-q}} s\right),$$

and

$$\|\hat{\beta}_{q,\lambda} - \beta^*\|_2^2 = O\left(\left(\sigma^2\frac{\log n}{m}\right)^{\frac{1}{2-q}} s\right). \tag{30}$$

Though the rate (30) in the case $q < 1$ is not as good as that of the ℓ_1 regularization method, the required regularity condition is substantially weaker. Specifically, for problems where the q-REC is satisfied but not classical REC, the recovery bound of the ℓ_1 regularization method may violate and lead to a false estimation while the ℓ_q regularization method still works and produces a comprehensive estimation; see Example C.1 in Appendix.

(ii) Theorem 2 obtains the recovery bound of the ℓ_q regularization method under the assumption of general q-REC(s,t,a), which covers (and improves) the one for the ℓ_1 regularization method in [13] as a special case when $q = 1$ and $a = 3$. From (11) and (27)–(29),

[1] For two functions f and g, we use $f \asymp g$ to denote that $f = cg$ for a universal positive constant c.

we can see that the larger the value of a, the smaller the value of $\phi_q(s,t,a,X)$, and thus the stronger condition the q-REC(s,t,a) is and the looser bounds obtained in Theorem 2. Hence Theorem 2 can be understood that the bounds of (27)–(29) holds as $a \to 1$ whenever the q-REC(s,t,a) is satisfied for some $a > 1$. In fact, when a stronger q-REC(s,t,a) is satisfied with a larger value of a, the estimated result of the regularization problem may be better; see Example C.2 in Appendix.

(iii) It was shown in [35] that the global solution of the FCP (cf. Assumption 1) sparse linear regression, including the SCAD (cf. (7)) and MCP (cf. (8)) as special cases, has an ℓ_2 recovery bound $O(\lambda^2 s)$ under the SEC. Though the recovery bounds are slightly better than (30), the condition required is substantially stronger than the q-REC. In [35], the authors also established the oracle property of the ℓ_0 regularization method under the SEC; while its ℓ_2 recovery bound cannot be guaranteed in their work. We shall see in section 5 that the ℓ_q regularization method performs better in parameter estimation than either the SCAD/MCP or the ℓ_0 regularization method via several numerical experiments.

(iv) [44, 45] considered the following ℓ_0 optimization problems

$$\min \ \|\beta\|_0, \quad \text{s.t.} \quad \left\| \frac{1}{m} X^\top (y - X\beta) \right\|_\infty \le \epsilon, \tag{31}$$

and

$$\min \ \frac{1}{2m}\|y - X\beta\|_2^2 + \lambda\|\beta\|_p, \quad \text{s.t.} \quad \|\beta\|_0 \le s \ (p = 1 \text{ or } 2). \tag{32}$$

[44] provided the ℓ_2 recovery bound of order $O(\epsilon^2)$ for problem (31) under the SEC-type condition, which is stronger than the q-REC; see Proposition 2. [45] established the prediction loss of order $O(\sigma\sqrt{\log n}\|\beta^*\|_1)$ and $O(\sigma\sqrt{s\log n}\|\beta^*\|_2)$ for problem (32) when $p = 1$ and $p = 2$, respectively. However, the ℓ_2 recovery bound was not obtained yet therein.

Remark 6 Recently, some works concerned the statistical property of the local minimum of some nonconvex regularization methods; see [33, 34].

(i) [34] studied the ℓ_2 recovery bound of the local minimum of a general regularization method:

$$\min \ \mathcal{L}_m(\beta; X) + \sum_{j=1}^{n} \rho_\lambda(\beta_j), \tag{33}$$

where $\mathcal{L}_m : \mathbb{R}^n \times \mathbb{R}^{m\times n} \to \mathbb{R}$ is the loss function, and $\rho_\lambda : \mathbb{R} \to \mathbb{R}$ is the (possibly nonconvex) penalty function. In [34], the penalty function ρ_λ is assumed to satisfy the following assumptions:

(a) $\rho_\lambda(0) = 0$ and is symmetric around zero;
(b) ρ_λ is nondecreasing on $\mathbb{R}_+$;
(c) For $t > 0$, the function $t \mapsto \frac{\rho_\lambda(t)}{t}$ is nonincreasing in t;
(d) ρ_λ is differentiable for each $t \neq 0$ and subdifferentiable at $t = 0$, with $\lim_{t\to 0^+} \rho_\lambda'(t) = \lambda L$;
(e) There exists $\mu > 0$ such that $\rho_{\lambda,\mu}(t) := \rho_\lambda(t) + \frac{\mu}{2}t^2$ is convex.

It was established in [34, Theorem 1] the ℓ_2 recovery bound of the critical point satisfying the first-order necessary condition of (33) under the restricted strong convex condition (RSC), which is a variant of the classical REC.

The ℓ_q norm can be reformulated as the penalty function $\rho_\lambda(\beta_j) := \lambda|\beta_j|^q$, however, it does not satisfy assumptions (d) or (e); in particular, assumption (e) plays a key role in the

establishment of the oracle property and ℓ_2 recovery bound of the local minimum. Therefore, the result in [34] cannot be directly applied to the ℓ_q regularization method, and the oracle property of the general local minimum of the ℓ_q regularization method is still an open question at this moment.

(ii) [33] studied the statistical property of the FCP sparse linear regression and presented the oracle property and ℓ_2 recovery bound of the certain local minimum, which satisfies a subspace second-order necessary condition and lies in the level set of the FCP regularized function at the true solution, under the SEC. Although the ℓ_q regularizer is beyond the FCP, our established Theorem 2 provides a theoretical result similar to [33] in the sense that the oracle property and ℓ_2 recovery bound are shown for the points within the level set of the ℓ_q regularized function at the true solution.

As an application of Theorem 2 to the case when $q = 1$, the following corollary presents statistical properties of the ℓ_1 regularization problem under the classical REC, which covers [13, Theorem 7.2] as a special case when $a = 3$, $\theta = 0$ and $b = 0$. The same ℓ_2 recovery bound rate $O(\sigma^2 s \log n/m)$ was reported in [46] under the sparse Riesz condition, which is comparable with the classical REC; while the same oracle inequality rate $O(\sigma^2 s \log n/m)$ was established in [17] under the compatibility condition, which is slightly weaker than the classical REC but cannot guarantee the ℓ_2 recovery bound.

Corollary 2 *Let $\hat{\beta}_{1,\lambda}$ be a global solution of* ($\mathrm{RP}_{1,\lambda}$) *with*

$$\lambda = 2\sigma(1+\theta)\sqrt{\frac{2(1+b)\log n}{m}}.$$

Suppose that X satisfies the 1-REC(s,t,a) *with $a > 1$ and that (19) is satisfied. Then, with probability at least $1 - \left(n^b\sqrt{\pi \log n}\right)^{-1}$, we have that*

$$\frac{1}{m}\|X\hat{\beta}_{1,\lambda} - X\beta^*\|_2^2 \le \frac{32a^2(1+b)(1+\theta)^2}{\phi_1^2(s,t,a,X)}\sigma^2\frac{\log n}{m}s,$$

$$\frac{1}{2m}\|X\hat{\beta}_{1,\lambda} - X\beta^*\|_2^2 + \lambda\|(\hat{\beta}_{1,\lambda})_{J^c}\|_1 \le \frac{16a^2(1+b)(1+\theta)^2}{\phi_1^2(s,t,a,X)}\sigma^2\frac{\log n}{m}s,$$

$$\|\hat{\beta}_{1,\lambda} - \beta^*\|_2^2 \le \frac{32a^2(1+b)(1+\theta)^2\left(1+9\frac{s}{t}\right)}{\phi_1^4(s,t,a,X)}\sigma^2\frac{\log n}{m}s.$$

4 Recovery bounds for random design

In practical applications, it is more realistic that the design matrix X is random. In this section, we consider this situation and present the ℓ_2 recovery bounds for the ℓ_q constrained minimization problem ($\mathrm{CP}_{q,\epsilon}$) and the ℓ_q regularization problem ($\mathrm{RP}_{q,\lambda}$) by virtue of the results obtained in the preceding section. In particular, throughout this section, we shall assume that

$$X \text{ is a Gaussian random design with i.i.d. } \mathcal{N}(0,\Sigma) \text{ rows},$$

that is, $X_{1\cdot}, \ldots, X_{m\cdot}$ are i.i.d. random vectors with each $X_{i\cdot} \sim \mathcal{N}(0,\Sigma)$. Recall that a, θ, and b are given by (12), and let (s,t) be a pair of integers satisfying (10).

To study the statistical properties of ($\mathrm{CP}_{q,\epsilon}$) and ($\mathrm{RP}_{q,\lambda}$) with a random design X, we first provide a sufficient condition for the q-REC of X by virtue of the population covariance

matrix Σ. For this purpose, we use $\Sigma^{\frac{1}{2}}$ to denote the square root of Σ and $\zeta(\Sigma) := \max_{1\le j\le n} \Sigma_{jj}$ to denote the maximal variance. Let $a > 1$ and assume the pair (s, t) satisfies (10). For $\delta \in \mathbb{R}^n$ and $J \subseteq \{1, 2, \ldots, n\}$, recall that $J(\delta; t)$ refers to the index set corresponding to the first t largest coordinates in absolute value of δ in J^c. The population covariance matrix $\Sigma \in \mathbb{R}^{n\times n}$ is assumed to satisfy the following condition

$$\Phi_q(s, t, a, \Sigma) := \min_{\delta\in\mathbb{R}^n:\delta\neq 0} \left\{ \frac{\|\Sigma^{1/2}\delta\|_2}{\|\delta_{J\cup J(\delta;t)}\|_2} : |J| \le s, \|\delta_{J^c}\|_q^q \le a\|\delta_J\|_q^q \right\} > 0. \tag{34}$$

Note that this condition is a quasi-REC for the square root of the population covariance matrix $\Sigma^{\frac{1}{2}}$ with the factor $\sqrt{n}$ moved, to make it consistent for the REC of the random design matrix X. Then two random events related to the linear regression model (1) with X being a Gaussian random design are

$$\mathscr{C}_a := \left\{ \phi_q(s, t, a, X) > \frac{1}{2}\Phi_q(s, t, a, \Sigma) \right\}, \tag{35}$$

and

$$\mathscr{D} := \left\{ \max_{1\le j\le n} \|X_{\cdot j}\|_2 \le (1+\theta)\sqrt{m} \right\}. \tag{36}$$

The following lemma calculates the probabilities of events $\mathscr{C}_a$ and $\mathscr{D}$, which is crucial for establishing the ℓ_2 recovery bounds of $(\mathrm{CP}_{q,\epsilon})$ and $(\mathrm{RP}_{q,\lambda})$ with a random design X. In particular, part (i) of this lemma shows that the Gaussian random design X satisfies the q-REC with high probability as long as the sample size m is sufficiently large and the population covariance matrix Σ satisfies (34); part (ii) of this lemma presents that each column of the Gaussian random design X has an Euclidean norm scaling as $\sqrt{m}$ with overwhelming probability.

Lemma 2 (i) *Suppose that Σ satisfies (34). Then, there exist universal positive constants (c_1, c_2) (independent of m, n, q, s, t, a, Σ) such that, if*

$$m > \frac{c_1\zeta(\Sigma)}{\Phi_q^2(s, t, a, \Sigma)} \left(\sqrt{s+t} + a\sqrt{s}\left(\frac{as}{t}\right)^{\frac{1}{q}-1} \right)^2 \log n, \tag{37}$$

then

$$\mathbb{P}(\mathscr{C}_a) \ge 1 - \exp(-c_2 m). \tag{38}$$

(ii) *Suppose that $\Sigma_{jj} = 1$ for all $j = 1, \ldots, n$. Then, there exist universal positive constants (c_3, c_4) and $\tau \ge 1$ (independent of m, n, θ, Σ) such that, if*

$$m > \frac{c_3\tau^4}{\theta^2} \log n, \tag{39}$$

then

$$\mathbb{P}(\mathscr{D}) \ge 1 - 2\exp(-c_4\theta^2 m/\tau^4). \tag{40}$$

Remark 7 As a direct application of Lemma 2(i), the classical REC is satisfied by X with high probability if Σ satisfies (34) with $q = 1$ and

$$m > \frac{c_1\zeta(\Sigma)}{\Phi_1^2(s, t, a, \Sigma)} \left(\sqrt{s+t} + a\sqrt{s}\right)^2 \log n, \tag{41}$$

which covers [16, Corollary 1] as a special case when $t = 0$.

Below, we consider the dominant property in the situation when X is a Gaussian random design. For the ℓ_q constrained minimization problem ($\mathrm{CP}_{q,\epsilon}$), Proposition 3 is still applicable in the case when X is Gaussian random since it does not rely on the assumption of X, and thus, (22) holds with the same probability for the random scenario; see Remark 3. In the following proposition, we show the dominant property (24) for the ℓ_q regularization problem ($\mathrm{RP}_{q,\lambda}$) with a random design by virtue of Proposition 4. Recall that ϵ, λ, ρ and the events $\mathscr{A}$ and $\mathscr{B}$ are given in the preceding section; see (14)-(17) for details.

Proposition 5 *Let $\hat{\beta}_{q,\lambda}$ be a global solution of ($\mathrm{RP}_{q,\lambda}$) with λ given by (15). Suppose that $\Sigma_{jj} = 1$ for all $j = 1, \ldots, n$. Then, there exist universal positive constants (c_1, c_2) and $\tau \geq 1$ (independent of $m, n, q, a, \theta, b, \epsilon, r, \lambda, \Sigma$) such that, if*

$$m > \frac{c_1 \tau^4}{\theta^2} \log n, \tag{42}$$

then (24) holds with probability at least $(1 - \left(n^b \sqrt{\pi \log n}\right)^{-1})(1 - 2\exp(-c_2\theta^2 m/\tau^4)) - \exp(-m)$.

Now we are ready to present the main theorems of this section, in which we establish the ℓ_2 recovery bounds for the ℓ_q constrained minimization problem ($\mathrm{CP}_{q,\epsilon}$) and the ℓ_q regularization problem ($\mathrm{RP}_{q,\lambda}$) when X is a Gaussian random design. The first theorem illustrates the stable recovery capability of the ℓ_q minimization method ($\mathrm{CP}_{q,\epsilon}$) (within a tolerance proportional to the noise) with high probability with a random design as long as β^* is sufficiently sparse and the sample size m is sufficiently large.

Theorem 3 *Let $\bar{\beta}_{q,\epsilon}$ be a global solution of ($\mathrm{CP}_{q,\epsilon}$) with ϵ given by (14). Suppose that Σ satisfies (34) with $a \geq 1$. Then, there exist universal positive constants (c_1, c_2) (independent of $m, n, q, s, t, \epsilon, \Sigma$) such that, if (37) is satisfied, then it holds with probability at least $(1 - \exp(-m))(1 - \exp(-c_2 m))$ that*

$$\|\bar{\beta}_{q,\epsilon} - \beta^*\|_2^2 \leq \frac{16\left(1 + \left(\frac{s}{t}\right)^{\frac{2}{q}-1}\right)}{m\Phi_q^2(s, t, a, \Sigma)}\epsilon^2. \tag{43}$$

As a direct application of Theorem 3 to the special case when $q = 1$, the following corollary presents the ℓ_2 recovery bound of the ℓ_1 minimization method ($\mathrm{CP}_{1,\epsilon}$) with a Gaussian random design as

$$\|\bar{\beta}_{1,\epsilon} - \beta^*\|_2 = O(\epsilon)$$

under the classical REC.

Corollary 3 *Let $\bar{\beta}_{1,\epsilon}$ be a global solution of ($\mathrm{CP}_{1,\epsilon}$) with ϵ given by (14). Suppose that Σ satisfies (34) with $q = 1$ and $a \geq 1$. Then, there exist universal positive constants (c_1, c_2) (independent of $m, n, q, s, t, \epsilon, \Sigma$) such that, if (41) is satisfied, then it holds with probability at least $(1 - \exp(-m))(1 - \exp(-c_2 m))$ that*

$$\|\bar{\beta}_{1,\epsilon} - \beta^*\|_2^2 \leq \frac{16(1 + \frac{s}{t})}{m\Phi_1^2(s, t, a, \Sigma)}\epsilon^2.$$

The other main theorem of this section is as follows, in which we exploit the estimation of prediction loss, the oracle property and the ℓ_2 recovery bound of parameter approximation of the ℓ_q regularization method ($\mathrm{RP}_{q,\lambda}$) with a Gaussian random design, by virtue of the q-REC of the square root of its population covariance matrix.

Theorem 4 *Let $\hat{\beta}_{q,\lambda}$ be a global solution of* (RP$_{q,\lambda}$) *with λ given by (15). Suppose that $\Sigma_{jj} = 1$ for all $j = 1, \ldots, n$ and Σ satisfies (34) with $a > 1$. Then, there exist universal positive constants (c_1, c_2, c_3, c_4) and $\tau \geq 1$ (independent of $m, n, q, s, t, a, \theta, b, \epsilon, r, \lambda, \Sigma$) such that, if*

$$m > \max\left\{ \frac{c_1(\sqrt{s+t} + a^{\frac{1}{q}}\sqrt{s}\left(\frac{s}{t}\right)^{\frac{1}{q}-1})^2}{\Phi_q^2(s,t,a,\Sigma)} \log n,\ \frac{c_3\tau^4}{\theta^2}\log n \right\}, \tag{44}$$

then it holds with probability at least

$$\left(1 - \exp(-m) - \left(n^b\sqrt{\pi \log n}\right)^{-1}\right)\left(1 - \exp(-c_2 m) - 2\exp(-c_4\theta^2 m/\tau^4)\right)$$

that

$$\frac{1}{m}\|X\hat{\beta}_{q,\lambda} - X\beta^*\|_2^2 \leq \left(\frac{2^{q+1}a\lambda}{\Phi_q^q(s,t,a,\Sigma)}\right)^{\frac{2}{2-q}} s, \tag{45}$$

$$\frac{1}{2m}\|X\hat{\beta}_{q,\lambda} - X\beta^*\|_2^2 + \lambda\|(\hat{\beta}_{q,\lambda})_{J^c}\|_q^q \leq \left(\frac{8^{\frac{q}{2}}a\lambda}{\Phi_q^q(s,t,a,\Sigma)}\right)^{\frac{2}{2-q}} s, \tag{46}$$

$$\|\hat{\beta}_{q,\lambda} - \beta^*\|_2^2 \leq \left(1 + a^{\frac{2}{q}}\left(\frac{s}{t}\right)^{\frac{2}{q}-1}\right)\left(\frac{8a\lambda}{\Phi_q^2(s,t,a,\Sigma)}\right)^{\frac{2}{2-q}} s. \tag{47}$$

As an application of Theorem 4 to the special case when $q = 1$, the following corollary presents the statistical properties of the ℓ_1 regularization method with a Gaussian random design under the classical REC. A similar ℓ_2 recovery bound was shown in [38, Theorem 3.1] by using a different analytic technique.

Corollary 4 *Let $\hat{\beta}_{1,\lambda}$ be a global solution of* (RP$_{1,\lambda}$) *with*

$$\lambda = 2\sigma(1+\theta)\sqrt{\frac{2(1+b)\log n}{m}}.$$

Suppose that $\Sigma_{jj} = 1$ for all $j = 1, \ldots, n$ and Σ satisfies (34) with $q = 1$ and $a > 1$. Then, there exist universal positive constants (c_1, c_2, c_3, c_4) and $\tau \geq 1$ (independent of $m, n, s, t, \theta, b, \Sigma$) such that, if

$$m > \max\left\{ \frac{c_1(\sqrt{s+t} + a\sqrt{s})^2}{\Phi_1^2(s,t,a,\Sigma)} \log n,\ \frac{c_3\tau^4}{\theta^2}\log n \right\},$$

then it holds with probability at least

$$(1 - \exp(-m) - \left(n^b\sqrt{\pi \log n}\right)^{-1})(1 - \exp(-c_2 m) - 2\exp(-c_4\theta^2 m/\tau^4))$$

that

$$\frac{1}{m}\|X\hat{\beta}_{1,\lambda} - X\beta^*\|_2^2 \leq \frac{128a^2(1+b)(1+\theta)^2}{\Phi_1^2(s,t,a,\Sigma)}\sigma^2\frac{\log n}{m}s,$$

$$\frac{1}{2m}\|X\hat{\beta}_{1,\lambda} - X\beta^*\|_2^2 + \lambda\|(\hat{\beta}_{1,\lambda})_{J^c}\|_1 \leq \frac{64a^2(1+b)(1+\theta)^2}{\Phi_1^2(s,t,a,\Sigma)}\sigma^2\frac{\log n}{m}s,$$

$$\|\hat{\beta}_{1,\lambda} - \beta^*\|_2^2 \leq \frac{512a^2(1+b)(1+\theta)^2(1+9\frac{s}{t})}{\Phi_1^4(s,t,a,\Sigma)}\sigma^2\frac{\log n}{m}s.$$

5 Numerical experiments

The purpose of this section is to carry out numerical experiments to illustrate the stability of the ℓ_q penalty optimization methods, verify the established theory of the ℓ_2 recovery bounds and compare the numerical performance of the ℓ_q regularization methods with another two widely-used nonconvex regularization methods, namely the SCAD (cf. (7)) and MCP (cf. (7)). In particular, we are concerned with the cases when $q = 0, 1/2, 2/3$ and 1. To solve the ℓ_q constrained minimization problems, we will apply the iterative reweighted algorithm [47, 48]. To solve the ℓ_q regularization problems, we will apply the iterative hard thresholding algorithm [49] for $q = 0$, the proximal gradient algorithm [32] for $q = 1/2$ and $2/3$, and FISTA [9] for $q = 1$, respectively. The proximal gradient algorithm proposed in [34] will be used to solve the SCAD and MCP. All numerical experiments are performed in MATLAB R2014b and executed on a personal desktop (Intel Core i7-4790, 3.60 GHz, 8.00 GB of RAM).

The simulated data are generated via a standard process; see, e.g., [32, 37]. Specifically, the design matrix are generated via two different ways – one is for the deterministic case and the other one is for the random case. For the deterministic vase, we randomly generate an i.i.d. Gaussian ensemble $X \in \mathbb{R}^{m\times n}$. For the random case, we first randomly generated a covariance matrix $\Sigma \in \mathbb{R}^{n\times n}$, and then generate the design matrix X with i.i.d. $\mathcal{N}(0, \Sigma)$ rows. The sparse vector $\beta^* \in \mathbb{R}^n$ with the sparsity being equal to s. The observation y is then generated by the MATLAB script

$$y = X * \beta^* + \sigma * randn(m, 1),$$

where σ is the noise level, i.e., the standard deviation of Gaussian noise. In the numerical experiments, the dimension of variables and the noise level are set as $n = 1024$ and $\sigma = 0.01$, respectively.

For each sparsity level s/n, we randomly generate the data X, β^*, y 100 times and run the algorithms mentioned above to solve the ℓ_q optimization problems for $q = 0, 1/2, 2/3$ and 1 as well as the SCAD and MCP. To simplify the notations, the solution of different problems will all be denoted as $\hat{\beta}$. The parameter ϵ in the ℓ_q constrained minimization problems ($\mathrm{CP}_{q,\epsilon}$) is set as $\epsilon = \sigma * \sqrt{m + 2\sqrt{2m}}$ in order to guarantee that $\|e\|_2^2$ is no more than ϵ^2 with overwhelming probability [1, 47]. In numerical experiments, the regularization parameter λ for each solver is chosen as the best one among $[0.01, 5]$ in terms of estimate error $\|\hat{\beta} - \beta^*\|_2$. In order to reveal the dependence of ℓ_2 recovery bounds on the sample size, we report numerical results for a range of sample sizes of the form $m = \Omega(s \log n)$, inspired by the established theorems (e.g., (44)).

The following first, second and fifth experiments are carried out with a deterministic design matrix, while the third and forth experiments are performed with a random design matrix.

The first experiment is conducted to show the performance on parameter estimation of the ℓ_q minimization methods (cf. (25)). Fig. 1 plots the logarithmic estimated error $\log(\|\hat{\beta} - \beta^*\|_2)$ along with different sample size m. From Fig. 1, we can see that the estimated error of each minimization method decreases consistently as the sample size increases. In addition, we find that the lower the q, the better the corresponding minimization method to achieve a more accurate solution.

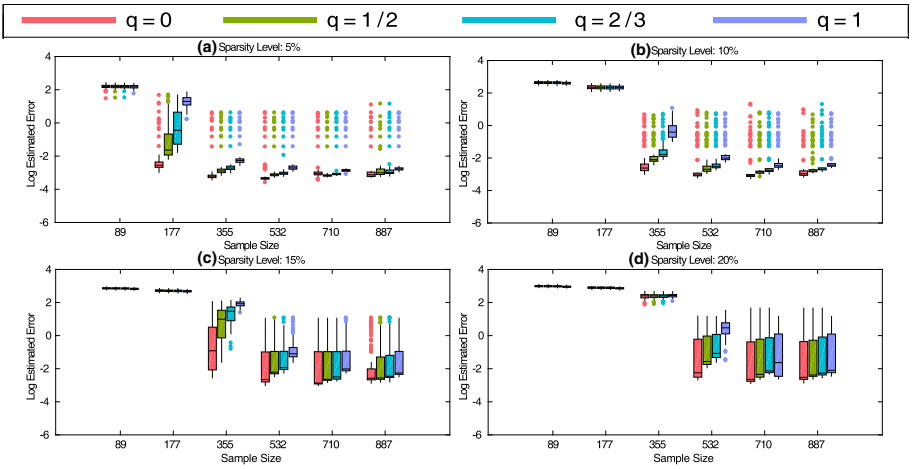


Fig. 1 Boxplots of the estimated error versus the sample size for different ℓ_q minimization methods with a deterministic design

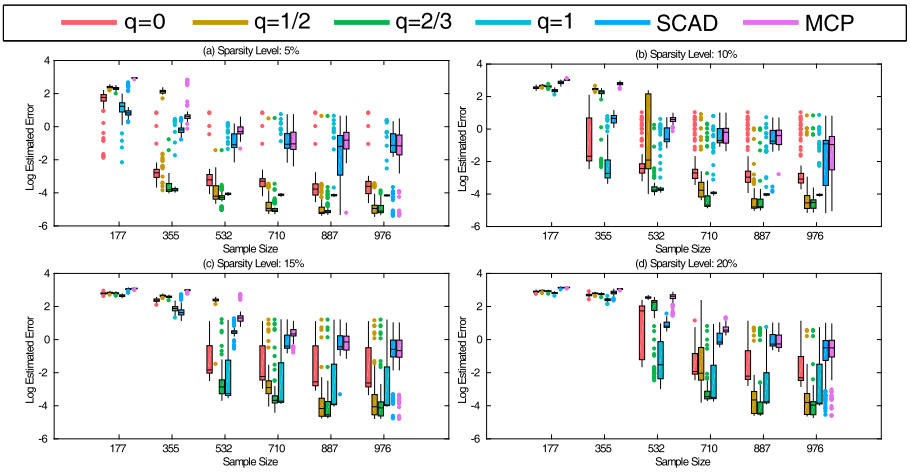


Fig. 2 Boxplots of the estimated error versus the sample size for different regularization methods with a deterministic design

The second experiment is carried out to show the performance on parameter estimation of the ℓ_q regularization methods (cf. (29)) and compare the performance with the SCAD and MCP. The corresponding result is displayed in Fig. 2, which plots the logarithmic estimated error $\log(\|\hat{\beta} - \beta^*\|_2)$ along with the sample size m. As shown by Fig. 2, the estimated error of each regularization method decreases consistently as the sample size increases. We also find that the lower-order regularization method (e.g., when $q = 1/2, 2/3$) outperforms the ℓ_0/ℓ_1 regularization method, in the sense that its estimated error decreases faster as the sample size increases and achieves a more accurate solution than the ℓ_0/ℓ_1 regularization method. This is due to the fact that the q-REC is satisfied when the sample size is larger than a certain level (see Lemma 2(i)) and that the lower-order regularization method only requires a weaker q-REC to guarantee its nice statistical property (see Theorems 2 and 4). This result is consistent

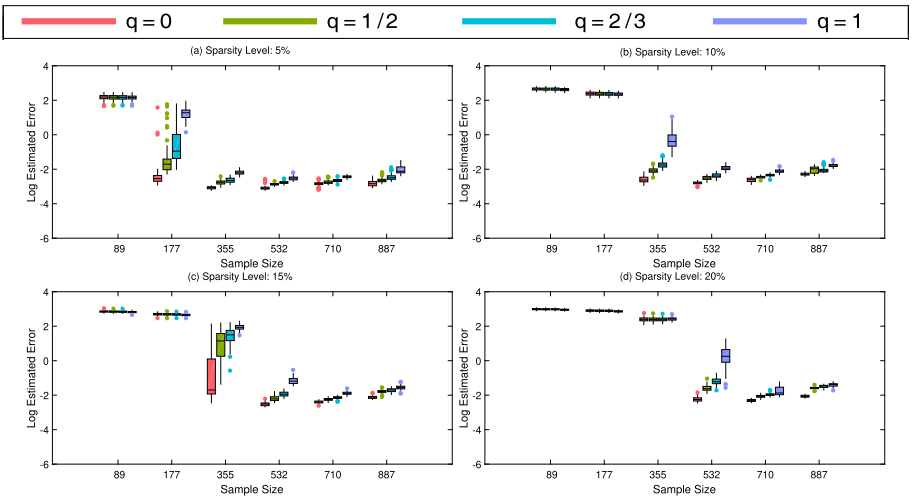


Fig. 3 Boxplots of the estimated error versus the sample size for different ℓ_q minimization methods with a random design

with existing empirical studies on the ℓ_q regularization methods as in [25, 32]. In addition, it is obvious that the lower-order regularization methods perform much better than the SCAD and MCP to achieve an accurate solution no matter whether the sparsity level is high or low. Moreover, we also note in Fig. 2b and d that there is a sudden increase of variation when the sample size is 532. This may be due to that when the sample size is not enough (e.g., 177, 352), all the methods exhibit poor recovery performance. When the sample size grows to $m = 532$, the recovery performance becomes better but the success probability is still not so high (see Theorem 2) because of the relatively limited sample size, thus leading to a sudden increase of variation.

The third and forth experiments are similar to the first and second experiments, respectively, with the only difference that the design matrix X is random. Corresponding theoretical conclusions are Eqs. (43) and (47), respectively. The results are respectively displayed in Figs. 3 and 4. The corresponding conclusions are similar to those in the first and second experiments, respectively, except that the variance of the estimate error in the random case is much smaller than that in the deterministic case. This phenomenon may be due to the high probability to guarantee the q-REC in the random case. Moreover, we also note in Figs. 3c and 4 that the lower-order minimization/regularization methods have larger variation than the ℓ_1 minimization/regularization methods. This may be due to the limited sample size for the exact recovery guarantee (see Theorems 3 and 4), and the variation turns smaller as long as the sample size becomes larger.

The fifth experiment is implemented to study the performance on variable selection of the ℓ_q regularization methods as well as the SCAD and MCP. We use following two criteria to characterize the capability of variable selection:

$$\text{sensitivity} = \frac{\text{true positive}}{\text{true positive+false negative}} \quad \text{and} \quad \text{specificity} = \frac{\text{true negative}}{\text{true negative+false positive}},$$

which respectively measures the proportion of positives and negatives that are correctly identified. The larger values of both sensitivity and specificity mean the higher capability of variable selection. The results are illustrated by averaging over 100 random trials. Tables 1

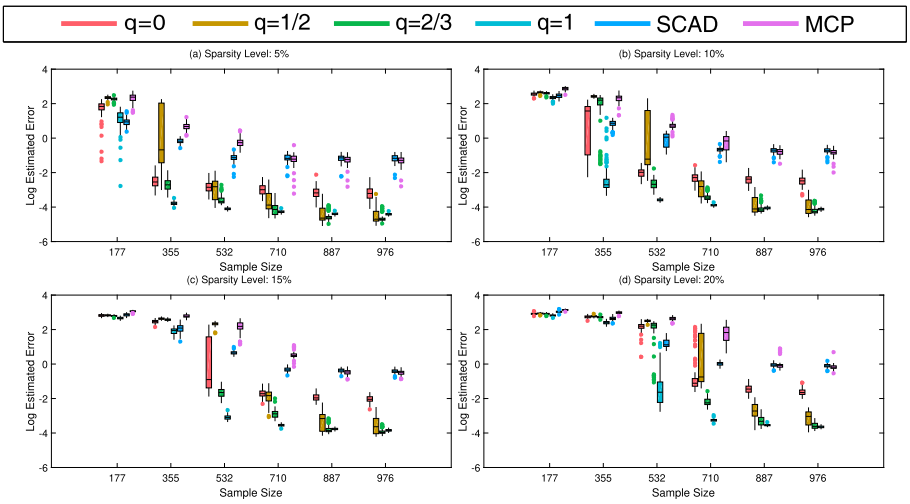


Fig. 4 Boxplots of the estimated error versus the sample size for different regularization methods with a random design

Table 1 Sensitivity of different regularization methods

Method	Sample size					
	177	355	532	710	887	976
$q = 0$	0.3029	0.8931	0.9824	0.9873	0.9902	0.9912
$q = 1/2$	0.2902	0.5108	0.9412	0.9873	0.9892	0.9941
$q = 2/3$	0.3108	0.9333	0.9922	0.9931	0.9941	0.9971
$q = 1$	0.5088	0.9980	1.0000	1.0000	1.0000	1.0000
SCAD	0.2882	0.8471	0.9157	0.9363	0.9324	0.9422
MCP	0.1373	0.4539	0.8461	0.9088	0.9353	0.9382

Table 2 Specificity of different regularization methods

Method	Sample size					
	177	355	532	710	887	976
$q = 0$	0.9229	0.9882	0.9980	0.9986	0.9989	0.9990
$q = 1/2$	0.8810	0.8119	1.0000	1.0000	1.0000	1.0000
$q = 2/3$	0.8782	0.9999	1.0000	1.0000	1.0000	1.0000
$q = 1$	0.8088	0.7454	0.7680	0.7473	0.7357	0.6120
SCAD	0.9653	0.9900	0.9906	0.9908	0.9919	0.9925
MCP	0.9466	0.9757	0.9909	0.9919	0.9925	0.9925

and 2 respectively chart the sensitivity and specificity of these methods at a sparsity level 10% corresponding to Fig. 2b. It is illustrated that the sensitivity and specificity of all these methods increase as the sample size grows, except for the specificity of the ℓ_1 regularization method, which is resulted from the fact that there are many small nonzero coefficients estimated by the ℓ_1 regularization method. We also note that the lower-order regularization method (e.g.,

when $q = 1/2, 2/3$) outperforms the other regularization methods in the sense that it can almost completely select the true model when the size of samples is getting large.

Finally, it is worth mentioning that the existing ℓ_q optimization algorithms (see, e.g., [25, 32, 47, 48]) are only proved to converge to a critical point, while their convergence to a global optimum is still an open question. Nevertheless, it is demonstrated by the numerical results above, as well as the ones in the literature, that the limiting point of these algorithms performs well in estimating the true underlying parameter.

Acknowledgements The authors are grateful to the editor and the anonymous reviewers for their valuable comments and suggestions toward the improvement of this paper. Xin Li's work was supported in part by the Natural Science Foundation of Shaanxi Province of China (2022JQ-045). Yaohua Hu's work was supported in part by the National Natural Science Foundation of China (12071306, 32170655, 11871347), Natural Science Foundation of Guangdong Province of China (2019A1515011917, 2020B1515310008), Project of Educational Commission of Guangdong Province of China (2021KTSCX103, 2019KZDZX1007), and Natural Science Foundation of Shenzhen (JCYJ20190808173603590). Chong Li's work was supported in part by the National Natural Science Foundation of China (11971429) and Zhejiang Provincial Natural Science Foundation of China (LY18A010004). Xiaoqi Yang's work was supported in part by the Research Grants Council of Hong Kong (PolyU 15212817). Tianzi Jiang's work was supported in part by the Science and Technology Innovation 2030 - Brain Science and Brain-Inspired Intelligence Project of China (2021ZD0200200).

Appendix

A Preliminary lemmas

We first recall some basic properties of the ℓ_q norm in the following lemmas; particularly, the first inequality in (A.1) is from [32, Eq. (7)] and the second inequality in (A.1) is from [50, Eq. (104)] and (A.2) is from [32, Lemma 2] by taking $p = 1$ and a simple variable substitution.

Lemma A.1 *Let $\alpha, \beta \in \mathbb{R}^n$. Then the following relations are true:*

$$\|\beta\|_{q_2} \le \|\beta\|_{q_1} \le n^{\frac{1}{q_1}-\frac{1}{q_2}} \|\beta\|_{q_2} \quad \text{for any } 0 < q_1 \le q_2 < +\infty, \tag{A.1}$$

$$\|\alpha\|_q^q - \|\beta\|_q^q \le \|\alpha+\beta\|_q^q \le \|\alpha\|_q^q + \|\beta\|_q^q \quad \text{for any } 0 < q \le 1. \tag{A.2}$$

Lemma A.2 *Let $p \ge 1$, $n_1, n_2 \in \mathbb{N}$, $\alpha \in \mathbb{R}_+^{n_1}$, $\beta \in \mathbb{R}_+^{n_2}$ and $c > 0$ be such that*

$$\max_{1\le i\le n_1} \alpha_i \le \min_{1\le j\le n_2} \beta_j \quad \text{and} \quad \sum_{i=1}^{n_1} \alpha_i \le c \sum_{j=1}^{n_2} \beta_j. \tag{A.3}$$

Then

$$\sum_{i=1}^{n_1} \alpha_i^p \le c \sum_{j=1}^{n_2} \beta_j^p. \tag{A.4}$$

Proof Let $\alpha_{\max} := \max_{1\le i\le n_1} \alpha_i$ and $\beta_{\min} := \min_{1\le j\le n_2} \beta_j$. Then it holds that

$$\alpha_{\max} \sum_{i=1}^{n_1} \alpha_i^p \le \alpha_{\max}^p \sum_{i=1}^{n_1} \alpha_i \quad \text{and} \quad \beta_{\min}^p \sum_{j=1}^{n_2} \beta_j \le \beta_{\min} \sum_{j=1}^{n_2} \beta_j^p. \tag{A.5}$$

Without loss of generality, we assume that $\alpha_{\max} > 0$; otherwise, (A.4) holds automatically. Thus, by the first inequality of (A.3) and noting $p \geq 1$, we have that

$$0 < \alpha_{\max}^p \beta_{\min} \leq \alpha_{\max} \beta_{\min}^p. \tag{A.6}$$

Multiplying the inequalities in (A.5) by $\beta_{\min} \sum_{j=1}^{n_2} \beta_j$ and $\alpha_{\max} \sum_{i=1}^{n_1} \alpha_i$ respectively, we obtain that

$$\begin{aligned}
\alpha_{\max}\beta_{\min} \sum_{i=1}^{n_1} \alpha_i^p \sum_{j=1}^{n_2} \beta_j &\leq \alpha_{\max}^p \beta_{\min} \sum_{i=1}^{n_1} \alpha_i \sum_{j=1}^{n_2} \beta_j \\
&\leq \alpha_{\max} \beta_{\min}^p \sum_{i=1}^{n_1} \alpha_i \sum_{j=1}^{n_2} \beta_j \\
&\leq \alpha_{\max}\beta_{\min} \sum_{i=1}^{n_1} \alpha_i \sum_{j=1}^{n_2} \beta_j^p,
\end{aligned}$$

where the second inequality follows from (A.6). This, together with the second inequality of (A.3), yields (A.4). The proof is complete. □

The following lemmas are useful for establishing the relationship between the q-REC and other types of regularity conditions; in particular, Lemmas A.3 and A.4 are taken from [11, Lemma 1.1] and [17, Lemma 3.1], respectively.

Lemma A.3 *Let $X \in \mathbb{R}^{m\times n}$ and $s, t \in \mathbb{N}$ be such that $s + t \leq n$. Then*

$$\theta_{s,t}(X) \leq \eta_{s+t}(X) \leq \theta_{s,t}(X) + \max\{\eta_s(X), \eta_t(X)\}.$$

Lemma A.4 *Let $\alpha, \beta \in \mathbb{R}^n$ and $0 < \tau < 1$ be such that $-\langle \alpha, \beta\rangle \leq \tau\|\alpha\|_2^2$. Then $(1-\tau)\|\alpha\|_2 \leq \|\alpha+\beta\|_2$.*

For the sake of simplicity, a partition structure and some notations are presented. For a vector $\delta \in \mathbb{R}^n$ and an index set $J \subseteq \{1, 2, \ldots, n\}$, we use $\mathrm{rank}(\delta_i; J^c)$ to denote the rank of the absolute value of δ_i in J^c (in a decreasing order) and $J_k(\delta; t)$ to denote the index set of the k-th batch of the first t largest coordinates in absolute value of δ in J^c. That is,

$$J_k(\delta; t) := \left\{ i \in J^c : \mathrm{rank}(\delta_i; J^c) \in \{kt+1, \ldots, (k+1)t\} \right\} \quad \text{for each } k \in \mathbb{N}. \tag{A.7}$$

Furthermore, we let $r := \lceil \frac{n-s}{t} \rceil$ (where $\lceil u \rceil$ denotes the largest integer not greater than u), $J_k := J_k(\delta; t)$ (defined by (A.7)) for each $k \in \mathbb{N}$ and $J_* := J \cup J_0$. With these notations, the Lemma A.5 is taken from [32, Lemma 7] when the group structure is degenerated.

Lemma A.5 *Let $\delta \in \mathbb{R}^n$, $0 < q \leq 1$ and $\tau \geq 1$. Then the following inequalities hold*

$$\|\delta_{J_*^c}\|_\tau \leq \sum_{k=1}^{r} \|\delta_{J_k}\|_\tau \leq t^{\frac{1}{\tau}-\frac{1}{q}} \|\delta_{J^c}\|_q.$$

Lemma A.6 *Let $X \in \mathbb{R}^{m\times n}$, $0 < q \leq 1$, $a > 0$, and (s, t) be a pair of integers satisfying (10). Then the following relations are true:*

$$\phi_q(s, t, a, X) \geq \frac{1}{\sqrt{m}} \left(\sqrt{\sigma_{\min}(s+t, X)} - a^{\frac{1}{q}} \left(\frac{s}{t}\right)^{\frac{1}{q}-\frac{1}{2}} \sqrt{\sigma_{\max}(t, X)} \right), \tag{A.8}$$

$$\phi_q(s, t, a, X) \leq \frac{1}{\sqrt{m}} \left(\sqrt{\sigma_{\max}(s+t, X)} + a^{\frac{1}{q}} \left(\frac{s}{t}\right)^{\frac{1}{q}-\frac{1}{2}} \sqrt{\sigma_{\max}(t, X)} \right). \tag{A.9}$$

Proof Fix $\delta \in C_q(s,a)$, as defined by (B.1). Then there exists $J \subseteq \{1,2,\ldots,n\}$ such that

$$|J| \le s \quad \text{and} \quad \|\delta_{J^c}\|_q^q \le a\|\delta_J\|_q^q. \tag{A.10}$$

Write $r := \lceil \frac{n-s}{t} \rceil$, $J_k := J_k(\delta;t)$ (defined by (A.7)) for each $k \in \mathbb{N}$ and $J_* := J \cup J_0$. Then it follows from Lemma A.5 and (A.10) that

$$\sum_{k=1}^{r} \|\delta_{J_k}\|_2 \le t^{\frac{1}{2}-\frac{1}{q}} \|\delta_{J^c}\|_q \le a^{\frac{1}{q}} t^{\frac{1}{2}-\frac{1}{q}} \|\delta_J\|_q \le a^{\frac{1}{q}} \left(\frac{s}{t}\right)^{\frac{1}{q}-\frac{1}{2}} \|\delta_J\|_2 \tag{A.11}$$

(due to (A.1)). Noting by (A.7) and (A.10) that $|J_*| \le s+t$ and $|J_k| \le t$ for each $k \in \mathbb{N}$, one has by (4) that

$$\sqrt{\sigma_{\min}(s+t,X)}\|\delta_{J_*}\|_2 \le \|X\delta_{J_*}\|_2 \le \sqrt{\sigma_{\max}(s+t,X)}\|\delta_{J_*}\|_2,$$
$$\|X\delta_{J_k}\|_2 \le \sqrt{\sigma_{\max}(t,X)}\|\delta_{J_k}\|_2 \quad \text{for each } k \in \mathbb{N}.$$

These, together with (A.11), imply that

$$\begin{aligned}\|X\delta\|_2 &\ge \|X\delta_{J_*}\|_2 - \sum_{k=1}^{r} \|X\delta_{J_k}\|_2 \\ &\ge \left(\sqrt{\sigma_{\min}(s+t,X)} - a^{\frac{1}{q}} \left(\frac{s}{t}\right)^{\frac{1}{q}-\frac{1}{2}} \sqrt{\sigma_{\max}(t,X)}\right) \|\delta_{J_*}\|_2.\end{aligned}$$

Since δ and J satisfying (A.10) are arbitrary, (A.8) is shown to hold by (11) and the fact that $J_* = J \cup J(\delta;t)$. One can prove (A.9) in a similar way, and thus, the details are omitted. □

The next two lemmas provide some preliminary lemmas to measure the probabilities of random events related to the linear regression model (1), in which Lemma A.7 is taken from [38, Lemma C.1].

Lemma A.7 *Let $0 \le \theta < 1$ and $b \ge 0$. Suppose that*

$$\max_{1\le j\le n} \|X_{\cdot j}\|_2 \le (1+\theta)\sqrt{m}. \tag{A.12}$$

Then

$$\mathbb{P}\left(\frac{\|X^\top e\|_\infty}{m} \ge \sigma(1+\theta)\sqrt{\frac{2(1+b)\log n}{m}}\right) \le \left(n^b\sqrt{\pi \log n}\right)^{-1}.$$

Lemma A.8 *Let $d \ge 5$. Then*

$$\mathbb{P}\left(\|e\|_2^2 \ge dm\sigma^2\right) \le \exp\left(-\frac{d-1}{4}m\right).$$

Proof Recall that $e = (e_1,\ldots,e_m)^\top \sim \mathcal{N}(0,\sigma^2\mathbb{I}_m)$. Let $u_i := \frac{1}{\sigma}e_i$ for $i = 1,\ldots,m$. Then one has that $u_1,\ldots,u_m$ are i.i.d. Gaussian variables with $u_i \sim \mathcal{N}(0,1)$ for $i = 1,\ldots,m$. Let $u := (u_1,\ldots,u_m)^\top$. Clearly, $\|u\|_2^2 = \frac{1}{\sigma^2}\|e\|_2^2$ is a chi-square random variable with m degrees of freedom (see, e.g., [51, Section 5.6]). Then it follows from standard tail bounds of chi-square random variable (see, e.g., [52, Appendix I]) that

$$\mathbb{P}\left(\frac{\|u\|_2^2 - m}{m} \ge d-1\right) \le \exp\left(-\frac{d-1}{4}m\right)$$

(as $d \geq 5$). Consequently, we obtain that

$$\mathbb{P}\left(\|e\|_2^2 \geq dm\sigma^2\right) = \mathbb{P}\left(\|u\|_2^2 \geq dm\right) \leq \exp\left(-\frac{d-1}{4}m\right).$$

The proof is complete. □

Recall that β^* satisfies the linear regression model (1). The following lemma is beneficial in proving Theorem 2.

Lemma A.9 *Let $\hat{\beta}_{q,\lambda}$ be an optimal solution of* (RP$_{q,\lambda}$). *Then*

$$\frac{1}{2m}\|X\beta^* - X\hat{\beta}_{q,\lambda}\|_2^2 \leq \lambda\|\beta^*\|_q^q - \lambda\|\hat{\beta}_{q,\lambda}\|_q^q + \frac{1}{m}\|\hat{\beta}_{q,\lambda} - \beta^*\|_1\|X^\top e\|_\infty.$$

Proof Since $\hat{\beta}_{q,\lambda}$ is an optimal solution of (RP$_{q,\lambda}$), it follows that

$$\frac{1}{2m}\|y - X\hat{\beta}_{q,\lambda}\|_2^2 + \lambda\|\hat{\beta}_{q,\lambda}\|_q^q \leq \frac{1}{2m}\|y - X\beta^*\|_2^2 + \lambda\|\beta^*\|_q^q.$$

This, together with (1), yields that

$$\begin{aligned}\lambda\|\hat{\beta}_{q,\lambda}\|_q^q - \lambda\|\beta^*\|_q^q &\leq \frac{1}{2m}\|y - X\beta^*\|_2^2 - \frac{1}{2m}\|y - X\hat{\beta}_{q,\lambda}\|_2^2\\ &= \frac{1}{m}\left\langle X(\hat{\beta}_{q,\lambda} - \beta^*), e\right\rangle - \frac{1}{2m}\|X\beta^* - X\hat{\beta}_{q,\lambda}\|_2^2\\ &\leq \frac{1}{m}\|\hat{\beta}_{q,\lambda} - \beta^*\|_1\|X^\top e\|_\infty - \frac{1}{2m}\|X\beta^* - X\hat{\beta}_{q,\lambda}\|_2^2.\end{aligned}$$

The proof is complete. □

Assume $X_{1\cdot}, \ldots, X_{m\cdot}$ are i.i.d. random vectors with each $X_{i\cdot} \sim \mathcal{N}(0, \Sigma)$. Then the following lemma is taken from [37, Supplementary, Lemma 6], which is useful for providing a sufficient condition for the q-REC of the random design X.

Lemma A.10 *There exist universal positive constants (c_1, c_2) (independent of m, n, Σ) such that it holds with probability at least $1 - \exp(-c_2 m)$ that, for each $\delta \in \mathbb{R}^n$*

$$\frac{\|X\delta\|_2^2}{m} \geq \frac{1}{2}\|\Sigma^{\frac{1}{2}}\delta\|_2^2 - c_1\zeta(\Sigma)\frac{\log n}{m}\|\delta\|_1^2. \tag{A.13}$$

B Technical proofs

Proof of Proposition 1 Associated with the q-REC(s, t, a), we define the feasible set

$$C_q(s, a) := \{\delta \in \mathbb{R}^n : \|\delta_{J^c}\|_q^q \leq a\|\delta_J\|_q^q \text{ for some } |J| \leq s\}. \tag{B.1}$$

By Definition 5, it remains to show that $C_{q_1}(s, a) \subseteq C_{q_2}(s, a)$. To this end, let $\delta \in C_{q_1}(s, a)$, and let J_0 denote the index set of the first s largest coordinates in absolute value of δ. By the assumption that $\delta \in C_{q_1}(s, a)$ and by the construction of J_0, one has $\|\delta_{J_0^c}\|_{q_1}^{q_1} \leq a\|\delta_{J_0}\|_{q_1}^{q_1}$. Then we obtain by Lemma A.2 (with q_2/q_1 in place of p) that $\|\delta_{J_0^c}\|_{q_2}^{q_2} \leq a\|\delta_{J_0}\|_{q_2}^{q_2}$; consequently, $\delta \in C_{q_2}(s, a)$. Hence, it follows that $C_{q_1}(s, a) \subseteq C_{q_2}(s, a)$, and the proof is complete. □

Proof of Proposition 2 It directly follows from Lemma A.6 (cf. (A.8)) that X satisfies the q-REC(s,t,a) provided that condition (a) holds. Fix $\delta \in C_q(s,a)$, and let J, r, J_k (for each $k \in \mathbb{N}$) and J_* be defined, respectively, as in the beginning of the proof of Lemma A.6. Then (A.11) follows directly and it follows from Lemma A.5 and (17) that

$$\|\delta_{J_*^c}\|_1 = \sum_{k=1}^{r} \|\delta_{J_k}\|_1 \le t^{1-\frac{1}{q}} \|\delta_{J^c}\|_q \le a^{\frac{1}{q}} t^{1-\frac{1}{q}} \|\delta_J\|_q \le a^{\frac{1}{q}} \left(\frac{s}{t}\right)^{\frac{1}{q}-1} \|\delta_J\|_1. \tag{B.2}$$

Suppose that condition (b) is satisfied. By Definition 2 (cf. (3)), one has that

$$|\langle X\delta_{J_*}, X\delta_{J_*^c}\rangle| \le \sum_{k=1}^{r} |\langle X\delta_{J_*}, X\delta_{J_k}\rangle| \le \theta_{t,s+t}(X)\|\delta_{J_*}\|_2 \sum_{k=1}^{r} \|\delta_{J_k}\|_2.$$

Then it follows from (A.11) that

$$\begin{aligned} |\langle X\delta_{J_*}, X\delta_{J_*^c}\rangle| &\le a^{\frac{1}{q}} \left(\frac{s}{t}\right)^{\frac{1}{q}-\frac{1}{2}} \theta_{t,s+t}(X)\|\delta_{J_*}\|_2\|\delta_J\|_2 \\ &\le \frac{a^{\frac{1}{q}} \left(\frac{s}{t}\right)^{\frac{1}{q}-\frac{1}{2}} \theta_{t,s+t}(X)}{1-\eta_{s+t}(X)} \|X\delta_{J_*}\|_2^2 \end{aligned} \tag{B.3}$$

(by (2)). Since $s \le t$ (by (10)), one has by Definition 1(i) that $\eta_s(X) \le \eta_t(X)$, and then by Lemma A.3 that $\eta_{s+t}(X) \le \theta_{s,t}(X) + \eta_t(X)$. Then it follows from (b) that

$$0 < \frac{a^{\frac{1}{q}} \left(\frac{s}{t}\right)^{\frac{1}{q}-\frac{1}{2}} \theta_{t,s+t}(X)}{1-\eta_{s+t}(X)} \le \frac{a^{\frac{1}{q}} \left(\frac{s}{t}\right)^{\frac{1}{q}-\frac{1}{2}} \theta_{t,s+t}(X)}{1-(\eta_t(X)+\theta_{s,t}(X))} < 1. \tag{B.4}$$

This, together with (B.3), shows that Lemma A.4 is applicable (with $X\delta_{J_*}$, $X\delta_{J_*^c}$, $\frac{a^{\frac{1}{q}} \left(\frac{s}{t}\right)^{\frac{1}{q}-\frac{1}{2}} \theta_{t,s+t}(X)}{1-\eta_{s+t}(X)}$ in place of α, β, τ) to concluding that

$$\begin{aligned} \|X\delta\|_2^2 &\ge \left(1 - \frac{a^{\frac{1}{q}} \left(\frac{s}{t}\right)^{\frac{1}{q}-\frac{1}{2}} \theta_{t,s+t}(X)}{1-\eta_{s+t}(X)}\right)^2 \|X\delta_{J_*}\|_2^2 \\ &\ge (1-\eta_{s+t}(X)) \left(1 - \frac{a^{\frac{1}{q}} \left(\frac{s}{t}\right)^{\frac{1}{q}-\frac{1}{2}} \theta_{t,s+t}(X)}{1-\eta_{s+t}(X)}\right)^2 \|\delta_{J_*}\|_2^2 \end{aligned}$$

(due to (2)). Since δ and J satisfying (A.10) are arbitrary, we derive by (11) and (B.4) that

$$\phi_q(s,t,a,X) \ge \frac{1}{\sqrt{m}} \left(\sqrt{1-\eta_{s+t}(X)} \left(1 - \frac{a^{\frac{1}{q}} \left(\frac{s}{t}\right)^{\frac{1}{q}-\frac{1}{2}} \theta_{t,s+t}(X)}{1-\eta_{s+t}(X)}\right)\right) > 0;$$

consequently, X satisfies the q-REC(s,t,a).

Suppose that (c) is satisfied. Then we have by (B.2) and Definition 2 (cf. (3)) that

$$\begin{aligned} \|X\delta\|_2^2 &= \|X\delta_{J_*}\|_2^2 + 2\langle X\delta_{J_*}, X\delta_{J^c}\rangle + \|X\delta_{J_*^c}\|_2^2 \\ &\ge \|X\delta_{J_*}\|_2^2 - 2|\langle X\delta_{J_*}, X\delta_{J_*^c}\rangle| \\ &\ge \|X\delta_{J_*}\|_2^2 - 2\theta_{1,1}(X)\|\delta_{J_*}\|_1\|\delta_{J_*^c}\|_1 \\ &\ge \|X\delta_{J_*}\|_2^2 - 2a^{\frac{1}{q}} \left(\frac{s}{t}\right)^{\frac{1}{q}-1} \theta_{1,1}(X)\|\delta_{J_*}\|_1^2. \end{aligned} \tag{B.5}$$

Separating the diagonal and off-diagonal terms of the quadratic form $\delta_{J_*}^T X^T X \delta_{J_*}$, one has by (A.1) and (c) that

$$\begin{aligned}\|X\delta_{J_*}\|_2^2 &= \sum_{i=1}^n (X^TX)_{i,i}(\delta_{J_*})_i(\delta_{J_*})_i + \sum_{j\neq k}^n (X^TX)_{j,k}(\delta_{J_*})_j(\delta_{J_*})_k \\ &= \|\delta_{J_*}\|_2^2 + \sum_{j\neq k}^n \langle X_{\cdot j}(\delta_{J_*})_j, X_{\cdot k}(\delta_{J_*})_k\rangle \\ &\geq \|\delta_{J_*}\|_2^2 - \theta_{1,1}(X)\|\delta_{J_*}\|_1^2 \\ &\geq (1-(s+t)\theta_{1,1}(X))\|\delta_{J_*}\|_2^2.\end{aligned}$$

Combining this inequality with (B.5), we get that

$$\|X\delta\|_2^2 \geq \left(1-\left(1+2a^{\frac{1}{q}}\left(\frac{s}{t}\right)^{\frac{1}{q}-1}\right)(s+t)\theta_{1,1}(X)\right)\|\delta_{J_*}\|_2^2.$$

Since δ and J satisfying (A.10) are arbitrary, we derive by (11) and (c) that

$$\phi_q(s,t,a,X) \geq \frac{1}{\sqrt{m}}\left(1-\left(1+2a^{\frac{1}{q}}\left(\frac{s}{t}\right)^{\frac{1}{q}-1}\right)(s+t)\theta_{1,1}(X)\right) > 0;$$

consequently, X satisfies the q-REC(s,t,a). The proof is complete. □

Proof of Lemma 1 By (14) and (16), Lemma A.8 is applicable (with $d=5$) to showing that $\mathbb{P}(\mathscr{A}^c) \leq \exp(-m)$, that is, (18) is proved. Then it remains to show (20) and (21). For this purpose, we have by (15) that $\lambda \geq \frac{5}{2}\sigma^2$, and noting that $0 < q \leq 1$,

$$\begin{aligned}\lambda &\geq \frac{a+1}{a-1}\sigma(1+\theta)2^{1-q}\left(\frac{5\sigma^2}{2\lambda}+r^q\right)^{\frac{1-q}{q}}\sqrt{\frac{2(1+b)\log n}{m}} \\ &= \frac{a+1}{a-1}\sigma(1+\theta)(2\rho)^{1-q}\sqrt{\frac{2(1+b)\log n}{m}}\end{aligned}$$

(due to (14)). Then one has by (17) that

$$\begin{aligned}\mathbb{P}(\mathscr{B}^c) &\leq \mathbb{P}\left(\frac{a+1}{(a-1)m}(2\rho)^{1-q}\|X^\top e\|_\infty \geq \frac{a+1}{a-1}\sigma(1+\theta)(2\rho)^{1-q}\sqrt{\frac{2(1+b)\log n}{m}}\right) \\ &= \mathbb{P}\left(\frac{\|X^\top e\|_\infty}{m} \geq \sigma(1+\theta)\sqrt{\frac{2(1+b)\log n}{m}}\right).\end{aligned}$$

Hence, by assumption (A.12), Lemma A.7 is applicable to ensuring (20). Moreover, it follows from the elementary probability theory that

$$\mathbb{P}(\mathscr{A}\cap\mathscr{B}) \geq \mathbb{P}(\mathscr{A}) - \mathbb{P}(\mathscr{B}^c) \geq 1-\exp(-m) - \left(n^b\sqrt{\pi\log n}\right)^{-1}.$$

The proof is complete. □

Proof of Proposition 3 Let $e \in \mathscr{A}$. Recall that β^* satisfies the linear regression model (1), one has that $\|y - X\beta^*\|_2 = \|e\|_2 \leq \epsilon$ (under the event $\mathscr{A}$), and so, β^* is a feasible vector of (CP$_{q,\epsilon}$). Consequently, by the optimality of $\bar{\beta}_{q,\epsilon}$ for (CP$_{q,\epsilon}$), it follows that $\|\bar{\beta}_{q,\epsilon}\|_q \leq \|\beta^*\|_q$. Write $\delta := \bar{\beta}_{q,\epsilon} - \beta^*$. Then we obtain that

$$\|\beta^*\|_q^q \geq \|\beta^*+\delta\|_q^q = \|\beta^*+\delta_J+\delta_{J^c}\|_q^q = \|\beta^*+\delta_J\|_q^q + \|\delta_{J^c}\|_q^q, \tag{B.6}$$

where the last equality holds because $\beta^*_{J^c} = 0$. On the other hand, one has by (A.2) that $\|\beta^* + \delta_J\|_q^q \geq \|\beta^*\|_q^q - \|\delta_J\|_q^q$. This, together with (B.6), implies (22). The proof is complete. □

Proof of Proposition 4 Let $e \in \mathscr{A}$. Since $\hat{\beta}_{q,\lambda}$ is an optimal solution of $(\mathrm{RP}_{q,\lambda})$, one has that

$$\frac{1}{2m}\|y - X\hat{\beta}_{q,\lambda}\|_2^2 + \lambda\|\hat{\beta}_{q,\lambda}\|_q^q \leq \frac{1}{2m}\|y - X\beta^*\|_2^2 + \lambda\|\beta^*\|_q^q.$$

Then, by (1) and (13), it follows that

$$\|\hat{\beta}_{q,\lambda}\|_q^q \leq \frac{1}{2m\lambda}\|y - X\beta^*\|_2^2 + \|\beta^*\|_q^q \leq \frac{1}{2m\lambda}\|e\|_2^2 + r^q \leq \rho^q$$

(due to (14) and (16)). Write $\delta := \hat{\beta}_{q,\lambda} - \beta^*$. Then, we obtain by (A.1) and (13) that

$$\|\delta\|_1 \leq \|\hat{\beta}_{q,\lambda}\|_1 + \|\beta^*\|_1 \leq \|\hat{\beta}_{q,\lambda}\|_q + \|\beta^*\|_q \leq \rho + r < 2\rho.$$

Consequently, noting that $0 < q \leq 1$, one sees that $\frac{\|\delta\|_1}{2\rho} \leq \left(\frac{\|\delta\|_1}{2\rho}\right)^q$, and then, by (A.1) that

$$\|\delta\|_1 \leq (2\rho)^{1-q}\|\delta\|_1^q \leq (2\rho)^{1-q}\|\delta\|_q^q. \tag{B.7}$$

This shows that (23) is proved. Then it remains to claim (24). To this end, noting that $\beta^*_{J^c} = 0$, we derive by Lemma A.9 that

$$\begin{aligned} -\frac{1}{m}\|\delta\|_1\|X^\top e\|_\infty &\leq \lambda\|\beta^*\|_q^q - \lambda\|\beta^* + \delta\|_q^q \\ &= \lambda\|\beta^*_J\|_q^q - \lambda\|\beta^*_J + \delta_J\|_q^q - \lambda\|\delta_{J^c}\|_q^q \\ &\leq \lambda\left(\|\delta_J\|_q^q - \|\delta_{J^c}\|_q^q\right) \end{aligned}$$

(by (A.2)). This, together with (B.7), yields that

$$\lambda\left(\|\delta_J\|_q^q - \|\delta_{J^c}\|_q^q\right) \geq -\frac{1}{m}(2\rho)^{1-q}\|\delta\|_q^q\|X^\top e\|_\infty.$$

Then, under the event $\mathscr{A} \cap \mathscr{B}$, we obtain by (17) that

$$(a+1)\left(\|\delta_J\|_q^q - \|\delta_{J^c}\|_q^q\right) \geq -(a-1)\|\delta\|_q^q = -(a-1)(\|\delta_J\|_q^q + \|\delta_{J^c}\|_q^q),$$

which yields (24). The proof is complete. □

Proof of Theorem 1 Write $\delta := \bar{\beta}_{q,\epsilon} - \beta^*$, and let $J_* := J \cup J_0(\delta; t)$ (defined by (A.7)). Fix $e \in \mathscr{A}$. Then it follows from Lemma A.5 and Proposition 3 that

$$\|\delta_{J_*^c}\|_2^2 \leq t^{1-\frac{2}{q}}\|\delta_{J^c}\|_q^2 \leq t^{1-\frac{2}{q}}\|\delta_J\|_q^2 \leq \left(\frac{s}{t}\right)^{\frac{2}{q}-1}\|\delta_J\|_2^2 \leq \left(\frac{s}{t}\right)^{\frac{2}{q}-1}\|\delta_{J_*}\|_2^2$$

(by (A.1)), and so

$$\|\delta\|_2^2 = \|\delta_{J_*}\|_2^2 + \|\delta_{J_*^c}\|_2^2 \leq \left(1 + \left(\frac{s}{t}\right)^{\frac{2}{q}-1}\right)\|\delta_{J_*}\|_2^2. \tag{B.8}$$

Recalling that β^* satisfies the linear regression model (1), we have that $\|y - X\beta^*\|_2 = \|e\|_2 \leq \epsilon$ (by (16)), and then

$$\|X\delta\|_2 = \|X\bar{\beta}_{q,\epsilon} - X\beta^*\|_2 \leq \|X\bar{\beta}_{q,\epsilon} - y\|_2 + \|X\beta^* - y\|_2 \leq 2\epsilon. \tag{B.9}$$

On the other hand, Proposition 3 is applicable to concluding that (22) holds, which shows $\delta \in C_q(s,1) \subseteq C_q(s,a)$ due to $a \geq 1$ (cf. (B.1)). Consequently, we obtain by the assumption of the q-REC(s,t,a) that

$$\|\delta_{J_*}\|_2 \leq \frac{\|X\delta\|_2}{\sqrt{m}\phi_q(s,t,a,X)}.$$

This, together with (B.8) and (B.9), implies that (25) holds under the event $\mathscr{A}$. Noting from Lemma 1 that $\mathbb{P}(\mathscr{A}) \geq 1-\exp(-m)$, we obtain the conclusion. The proof is complete. □

Proof of Theorem 2 Write $\delta := \hat{\beta}_{q,\lambda} - \beta^*$ and fix $e \in \mathscr{A} \cap \mathscr{B}$. Note by (23) and (17) that

$$\frac{1}{m}\|\delta\|_1 \|X^\top e\|_\infty \leq \frac{a-1}{a+1}\lambda\|\delta\|_q^q.$$

This, together with Lemma A.9, implies that

$$\begin{aligned}\frac{1}{2m}\|X\hat{\beta}_{q,\lambda} - X\beta^*\|_2^2 &\leq \lambda\|\beta^*\|_q^q - \lambda\|\hat{\beta}_{q,\lambda}\|_q^q + \frac{a-1}{a+1}\lambda\|\delta\|_q^q \\ &\leq \lambda\|\delta_J\|_q^q - \lambda\|(\hat{\beta}_{q,\lambda})_{J^c}\|_q^q + \frac{a-1}{a+1}\lambda\|\delta\|_q^q\end{aligned} \tag{B.10}$$

(noting that $\beta^*_{J^c} = 0$ and by (A.2)). Let $J_* := J \cup J_0(\delta;t)$. One has by (24) and (A.1) that

$$\lambda\|\delta_J\|_q^q + \frac{a-1}{a+1}\lambda\|\delta\|_q^q \leq a\lambda\|\delta_J\|_q^q \leq a\lambda s^{1-\frac{q}{2}}\|\delta_J\|_2^q,$$

and by the assumption of the q-REC(s,t,a) that

$$\|\delta_J\|_2 \leq \|\delta_{J^*}\|_2 \leq \frac{\|X\delta\|_2}{\sqrt{m}\phi_q(s,t,a,X)}.$$

These two inequalities, together with (B.10), imply that

$$\frac{1}{2m}\|X\hat{\beta}_{q,\lambda} - X\beta^*\|_2^2 + \lambda\|(\hat{\beta}_{q,\lambda})_{J^c}\|_q^q \leq \frac{a\lambda s^{1-\frac{q}{2}}}{m^{\frac{q}{2}}\phi_q^q(s,t,a,X)}\|X\hat{\beta}_{q,\lambda} - X\beta^*\|_2^q.$$

This yields that

$$\text{(27) and (28) hold under the event } \mathscr{A} \cap \mathscr{B}. \tag{B.11}$$

Furthermore, it follows from Lemma A.5 that

$$\|\delta_{J_*^c}\|_2^2 \leq t^{1-\frac{2}{q}}\|\delta_{J^c}\|_q^2 \leq a^{\frac{2}{q}}t^{1-\frac{2}{q}}\|\delta_J\|_q^2 \leq a^{\frac{2}{q}}\left(\frac{s}{t}\right)^{\frac{2}{q}-1}\|\delta_J\|_2^2.$$

(by (24) and (A.1)). By the assumption of the q-REC(s,t,a), one has by (27) that

$$\|\delta_{J_*}\|_2^2 \leq \frac{\|X\delta\|_2^2}{m\phi_q^2(s,t,a,X)} \leq \left(\frac{2a\lambda}{(\phi_q^2(s,t,a,X)}\right)^{\frac{2}{2-q}} s.$$

Hence we obtain that

$$\begin{aligned}\|\hat{\beta}_{q,\lambda} - \beta^*\|_2^2 = \|\delta_{J_*}\|_2^2 + \|\delta_{J_*^c}\|_2^2 &\leq \left(1 + a^{\frac{2}{q}}\left(\frac{s}{t}\right)^{\frac{2}{q}-1}\right)\|\delta_{J_*}\|_2^2 \\ &\leq \left(1 + a^{\frac{2}{q}}\left(\frac{s}{t}\right)^{\frac{2}{q}-1}\right)\left(\frac{2a\lambda}{(\phi_q^2(s,t,a,X))}\right)^{\frac{2}{2-q}} s.\end{aligned}$$

This shows that

$$\text{(29) holds under the event } \mathscr{A} \cap \mathscr{B}. \tag{B.12}$$

By assumption (A.12), Lemma 1 is applicable to concluding that

$$\mathbb{P}(\mathscr{A} \cap \mathscr{B}) \geq 1 - \exp(-m) - \left(n^b \sqrt{\pi \log n}\right)^{-1}.$$

This, together with (B.11) and (B.12), yields that (27)-(29) hold with probability at least $1 - \exp(-m) - \left(n^b \sqrt{\pi \log n}\right)^{-1}$. The proof is complete. □

Proof of Lemma 2 (i) We first claim that

$$\phi_q(s, t, a, X) > \frac{1}{2}\Phi_q(s, t, a, \Sigma), \tag{B.13}$$

whenever (A.13) holds for each $\delta \in \mathbb{R}^n$. To this end, we suppose that (A.13) is satisfied for each $\delta \in \mathbb{R}^n$. Fix $\delta \in C_q(s, a)$, and let J, r, J_k (for each $k \in \mathbb{N}$) and J_* be defined, respectively, as in the beginning of the proof of Lemma A.6. Then (B.2) follows directly, and one has that

$$\begin{aligned} \|\delta\|_1 &= \|\delta_{J_*}\|_1 + \|\delta_{J_*^c}\|_1 \\ &\leq \sqrt{s+t}\|\delta_{J_*}\|_2 + a\sqrt{s}\left(\frac{as}{t}\right)^{\frac{1}{q}-1}\|\delta_J\|_2 \\ &\leq \left(\sqrt{s+t} + a\sqrt{s}\left(\frac{as}{t}\right)^{\frac{1}{q}-1}\right)\|\delta_{J_*}\|_2. \end{aligned} \tag{B.14}$$

By the assumption that Σ satisfies (34), it follows that

$$\|\Sigma^{\frac{1}{2}}\delta\|_2^2 \geq \Phi_q^2(s, t, a, \Sigma)\|\delta_{J_*}\|_2^2.$$

Substituting this inequality and (B.14) into (A.13) yields

$$\frac{\|X\delta\|_2^2}{m} \geq \left(\frac{1}{2}\Phi_q^2(s, t, a, \Sigma) - c_1\zeta(\Sigma)\left(\sqrt{s+t} + a\sqrt{s}\left(\frac{as}{t}\right)^{\frac{1}{q}-1}\right)^2\frac{\log n}{m}\right)\|\delta_{J_*}\|_2^2.$$

This, together with (37), shows that

$$\frac{\|X\delta\|_2^2}{m} \geq \frac{1}{4}\Phi_q^2(s, t, a, \Sigma)\|\delta_{J_*}\|_2^2.$$

Since δ and J satisfying (A.10) are arbitrary, we derive by (11) that (B.13) holds, as desired. Then, Lemma A.10 is applicable to concluding (38).

(ii) Noting by the assumption that $\Sigma_{jj} = 1$ for all $j = 1, \ldots, n$, [38, Theorem 1.6] is applicable to showing that there exist universal positive constants (c_1, c_2) and $\tau \geq 1$ such that

$$\mathbb{P}\left(\cap_{j=1}^n \left\{(1-\theta)\sqrt{m} \leq \|X_{\cdot j}\|_2 \leq (1+\theta)\sqrt{m}\right\}\right) \geq 1 - 2\exp(-c_2\theta^2 m/\tau^4),$$

whenever m satisfies (39). Then it immediately follows from (36) that

$$\begin{aligned} \mathbb{P}(\mathscr{D}) &= \mathbb{P}(\cap_{j=1}^n \{\|X_{\cdot j}\|_2 \leq (1+\theta)\sqrt{m}\}) \\ &\geq 1 - 2\exp(-c_2\theta^2 m/\tau^4), \end{aligned}$$

that is, (40) is proved. □

Proof of Proposition 5 By (36), one sees by Proposition 4 that (24) holds under the event $\mathscr{A} \cap \mathscr{B} \cap \mathscr{D}$. Then it remains to estimate $\mathbb{P}(\mathscr{A} \cap \mathscr{B} \cap \mathscr{D})$. By Lemma 2(ii), there exist universal positive constants (c_1, c_2) and $\tau \geq 1$ such that

$$\mathbb{P}(\mathscr{D}) \geq 1 - 2\exp(-c_2\theta^2 m/\tau^4),$$

whenever m satisfies (42). From Lemma 1 (cf. (20)), we have also by (36) that

$$\mathbb{P}(\mathscr{B}|\mathscr{D}) \geq 1 - (n^b\sqrt{\pi \log n})^{-1}.$$

Then, it follows that

$$\begin{aligned}\mathbb{P}(\mathscr{B} \cap \mathscr{D}) &= \mathbb{P}(\mathscr{B}|\mathscr{D})\mathbb{P}(\mathscr{D}) \\ &\geq (1 - (n^b\sqrt{\pi \log n})^{-1})(1 - 2\exp(-c_2\theta^2 m/\tau^4)),\end{aligned}$$

and then by the elementary probability theory and (18) that,

$$\begin{aligned}\mathbb{P}(\mathscr{A} \cap \mathscr{B} \cap \mathscr{D}) &= \mathbb{P}(\mathscr{B} \cap \mathscr{D}) - \mathbb{P}(\mathscr{B} \cap \mathscr{D} \cap \mathscr{A}^c) \\ &\geq \mathbb{P}(\mathscr{B} \cap \mathscr{D}) + \mathbb{P}(\mathscr{A}) - 1 \\ &\geq \left(1 - \left(n^b\sqrt{\pi \log n}\right)^{-1}\right)(1 - 2\exp(-c_2\theta^2 m/\tau^4)) - \exp(-m),\end{aligned}$$

whenever m satisfies (42). The proof is complete. □

Proof of Theorem 3 To simplify the proof, corresponding to inequalities (25) and (43), we define the following two events

$$\mathscr{E}_1 := \left\{ \|\bar{\beta}_{q,\epsilon} - \beta^*\|_2^2 \leq \frac{4(1 + \left(\frac{s}{t}\right)^{\frac{2}{q}-1})}{\phi_q^2(s,t,a,X)}\epsilon^2 \right\},$$

$$\mathscr{E}_2 := \left\{ \|\bar{\beta}_{q,\epsilon} - \beta^*\|_2^2 \leq \frac{16(1 + \left(\frac{s}{t}\right)^{\frac{2}{q}-1})}{m\Phi_q^2(s,t,a,\Sigma)}\epsilon^2 \right\}.$$

Then, by the definition of $\mathscr{C}_1$ (35), we have that $\mathscr{C}_1 \cap \mathscr{E}_1 \subseteq \mathscr{E}_2$ and thus

$$\mathbb{P}(\mathscr{E}_2) \geq \mathbb{P}(\mathscr{E}_1 \cap \mathscr{C}_1) = \mathbb{P}(\mathscr{E}_1|\mathscr{C}_1)\mathbb{P}(\mathscr{C}_1). \tag{B.15}$$

Note by Theorem 1 that

$$\mathbb{P}(\mathscr{E}_1|\mathscr{C}_1) \geq 1 - \exp(-m). \tag{B.16}$$

By Lemma 2(i) (with $a = 1$), there exist universal positive constants (c_1, c_2) such that (37) ensures (38). Then we obtain by (B.15) and (B.16) that

$$\mathbb{P}(\mathscr{E}_2) \geq (1 - \exp(-m))(1 - \exp(-c_2 m)),$$

whenever m satisfies (37). The proof is complete. □

Proof of Theorem 4 To simplify the proof, we define the following six events

$$\begin{aligned}&\mathscr{F}_1 = \{(31) \text{ happens}\}, \quad \mathscr{F}_2 = \{(32) \text{ happens}\}, \quad \mathscr{F}_3 = \{(33) \text{ happens}\}, \\ &\mathscr{G}_1 = \{(45) \text{ happens}\}, \quad \mathscr{G}_2 = \{(46) \text{ happens}\}, \quad \mathscr{G}_3 = \{(47) \text{ happens}\}.\end{aligned}$$

Fix $i \in \{1, 2, 3\}$. Then, we have by (35) that $\mathscr{C}_a \cap \mathscr{F}_i \subseteq \mathscr{G}_i$ and thus

$$\mathbb{P}(\mathscr{G}_i) \geq \mathbb{P}(\mathscr{C}_a \cap \mathscr{F}_i). \tag{B.17}$$

By Lemma 2, there exist universal positive constants (c_1, c_2, c_3, c_4) and $\tau \geq 1$ such that, (44) ensures (38) and (40). Then it follows from (38) and (40) that

$$\mathbb{P}(\mathscr{C}_a \cap \mathscr{D}) \geq \mathbb{P}(\mathscr{C}_a) + P(\mathscr{D}) - 1 \geq 1 - \exp(-c_2 m) - 2\exp(-c_4\theta^2 m/\tau^4), \quad \text{(B.18)}$$

whenever m satisfies (44). Recall from Theorem 2 that

$$\mathbb{P}(\mathscr{F}_i|\mathscr{C}_a \cap \mathscr{D}) \geq 1 - \exp(-m) - \left(n^b\sqrt{\pi \log n}\right)^{-1}.$$

This, together with (B.18), implies that

$$\begin{aligned}\mathbb{P}(\mathscr{C}_a \cap \mathscr{F}_i) &\geq \mathbb{P}(\mathscr{F}_i|\mathscr{C}_a \cap \mathscr{D})\,\mathbb{P}(\mathscr{C}_a \cap \mathscr{D}) \\ &\geq \left(1 - \exp(-m) - \left(n^b\sqrt{\pi \log n}\right)^{-1}\right)(1 - \exp(-c_2 m) - 2\exp(-c_4\theta^2 m/\tau^4)).\end{aligned}$$

Then, one has by (B.17) that

$$\mathbb{P}(\mathscr{G}_i) \geq \left(1 - \exp(-m) - \left(n^b\sqrt{\pi \log n}\right)^{-1}\right)(1 - \exp(-c_2 m) - 2\exp(-c_4\theta^2 m/\tau^4)),$$

whenever m satisfies (44). The proof is complete. □

C Example to illustrate the recovery bound

The following example shows the performance of the $\ell_{1/2}$ regularization method and the ℓ_1 regularization method in the case where 1/2-REC(1, 1, 1) is satisfied but not the classical REC(1, 1, 1).

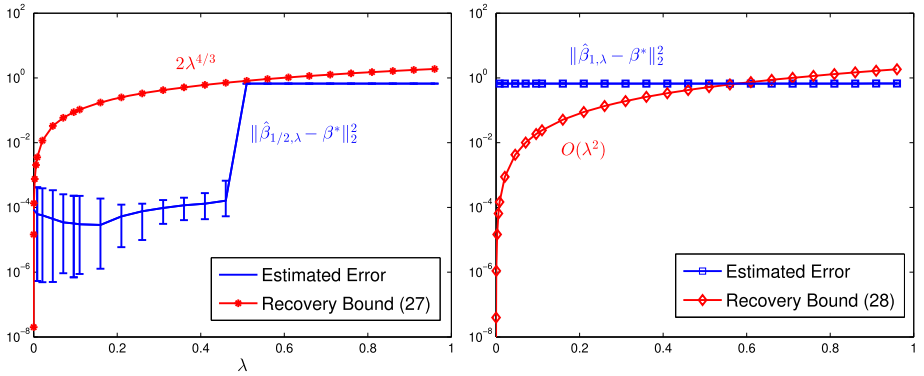
Example C.1 Consider the linear regression model (1), where

$$X := \begin{pmatrix} 2 & 3 & 1 \\ 2 & 1 & 3 \end{pmatrix}, \quad \beta^* := (1, 0, 0)^\top, \quad e \sim \mathcal{N}(0, 0.01).$$

It was validated in [32, Example 1] that the matrix X satisfies 1/2-REC(1, 1, 1) but not the classical REC(1, 1, 1); hence the recovery bound of the $\ell_{1/2}$ regularization method is satisfied but may not for the ℓ_1 regularization method.

To show the performance of the $\ell_{1/2}$ regularization method and the ℓ_1 regularization method in this case, for each regularization parameter λ varying from 10^{-8} to 1, we randomly generate the Gaussian noise 500 times and calculate the estimated errors $\|\hat{\beta}_{q,\lambda} - \beta^*\|_2^2$ for the $\ell_{1/2}$ regularization method and the ℓ_1 regularization method, respectively. We employ FISTA [9] and the filled function method [53] to find the global solution of the ℓ_1 regularization problem and the $\ell_{1/2}$ regularization problem, respectively. The results are illustrated in Fig. 5, in which the error bars represent the 95% confidence intervals and the curves of recovery bounds stand for the terms in the right-hand side of (29) (cf. [32, Example 2]) and (30), respectively. It is observed from Fig. 5a that the recovery bound (29) is satisfied with high probability for most of λ's and tight when $\lambda \approx \frac{1}{2}$ for the $\ell_{1/2}$ regularization method. Fig. 5b shows that the estimated error (30) for the ℓ_1 regularization method is not satisfied when λ is small because the classical REC violates. Moreover, the solutions of the ℓ_1 regularization problem are always equal-contributed among 3 components that leads to the failure approach to a sparse solution.

The next example is to illustrate the influence of the parameter a as in the REC on estimated errors obtained by the ℓ_1 regularization method.



(a) The $\ell_{1/2}$ regularization method. **(b)** The ℓ_1 regularization method.

Fig. 5 The illustration of recovery bounds and estimated errors

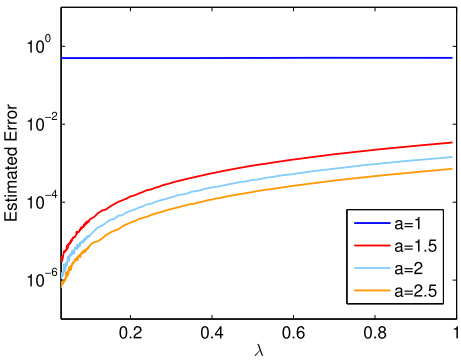


Fig. 6 The illustration of influence of parameter a as in the REC on estimated errors

Example C.2 Consider the linear regression model (1), where

$$X := \begin{pmatrix} 1 & 1 & -a \\ 1 & 2 & -(a+1) \end{pmatrix}, \quad \beta^* := (0, 0, 1)^\top, \quad e \sim \mathcal{N}(0, 0.01). \tag{C.1}$$

Due to the geometric interpretation of the REC, one sees that the REC$(1, 1, a)$ holds if and only if the null space of X does not intersect the feasible set $C_1(1, a)$ (B.1); one can also refer to [32, Fig. 1]. Hence one can check by the construction of X that it satisfies the REC$(1, 1, c)$ for each $c < \min\{a, 2\}$ and fails otherwise. That is, if X is given by (C.1) with a larger parameter $a \in [1, 2]$, then X satisfies a stronger REC$(1, 1, a)$.

To show the influence of the parameter a on the estimated error, we select $a := 1, 1.5, 2, 2.5$ as an instance to construct the matrix X by (C.1). Then for each regularization parameter λ varying from 10^{-8} to 1, we randomly generate the Gaussian noise 500 times and calculate the estimated errors $\|\hat{\beta}_{1,\lambda} - \beta^*\|_2^2$ using FISTA [9] to find the global solution of the ℓ_1 regularization problem. The averaged result is displayed in Fig. 6. One can see that (i) the recovery bound in (29) is satisfied when $a > 1$ but fails when $a = 1$ that is consistent with Theorem 2; and (ii) as parameter a becomes larger, the estimated error becomes smaller that shows a better result than Theorem 2.

References

1. Candès, E.J., Romberg, J.K., Tao, T.: Stable signal recovery from incomplete and inaccurate measurements. Commun. Pure Appl. Math. **59**(8), 410–412 (2006)
2. Pun, C.S., Wong, H.Y.: A linear programming model for selection of sparse high-dimensional multiperiod portfolios. Eur. J. Oper. Res. **273**(2), 754–771 (2019)
3. Qin, J., Hu, Y.H., Xu, F., Yalamanchili, H.K., Wang, J.W.: Inferring gene regulatory networks by integrating ChIP-seq/chip and transcriptome data via LASSO-type regularization methods. Methods **67**(3), 294–303 (2014)
4. Donoho, D.L., Elad, M., Temlyakov, V.N.: Stable recovery of sparse overcomplete representations in the presence of noise. IEEE Trans. Inf. Theory **52**(1), 6–18 (2006)
5. Natarajan, B.K.: Sparse approximate solutions to linear systems. SIAM J. Comput. **24**(2), 227–234 (1995)
6. Chen, S.S., Donoho, D.L., Saunders, M.A.: Atomic decomposition by basis pursuit. SIAM Rev. **43**(1), 129–159 (2001)
7. Figueiredo, M.A.T., Nowak, R.D., Wright, S.J.: Gradient projection for sparse reconstruction: application to compressed sensing and other inverse problems. IEEE J. Sel. Top. Signal Process. **1**(4), 586–597 (2007)
8. Tibshirani, R.: Regression shrinkage and selection via the Lasso. J. R. Stat. Soc. Ser. B **58**(1), 267–288 (1996)
9. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci. **2**(1), 183–202 (2009)
10. Daubechies, I., Devore, R., Fornasier, M.: Iteratively reweighted least squares minimization for sparse recovery. Commun. Pure Appl. Math. **63**(1), 1–38 (2010)
11. Candès, E.J., Tao, T.: Decoding by linear programming. IEEE Trans. Inf. Theory **51**(12), 4203–4215 (2005)
12. Donoho, D.L., Huo, X.M.: Uncertainty principles and ideal atomic decomposition. IEEE Trans. Inf. Theory **47**(7), 2845–2862 (2001)
13. Bickel, P.J., Ritov, Y., Tsybakov, A.B.: Simultaneous analysis of Lasso and Dantzig selector. Annal. Stat. **37**(4), 1705–1732 (2009)
14. Donoho, D.L.: For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution. Commun. Pure Appl. Math. **59**(6), 797–829 (2006)
15. Cai, T.T., Xu, G.W., Zhang, J.: On recovery of sparse signals via ℓ_1 minimization. IEEE Trans. Inf. Theory **55**(7), 3388–3397 (2009)
16. Raskutti, G., Wainwright, M.J., Yu, B.: Restricted eigenvalue properties for correlated gaussian designs. J. Mach. Learn. Res. **11**(2), 2241–2259 (2010)
17. van de Geer, S.A., Bühlmann, P.: On the conditions used to prove oracle results for the Lasso. Electr. J. Stat. **3**, 2009 (2009)
18. Bunea, F., Tsybakov, A., Wegkamp, M.: Sparsity oracle inequalities for the Lasso. Electr. J. Stat. **64**(3), 330–2 (2007)
19. Zhang, T.: Some sharp performance bounds for least squares regression with ℓ_1 regularization. Annal. Stat. **37**, 2109–2144 (2009)
20. Chartrand, R.: Exact reconstruction of sparse signals via nonconvex minimization. IEEE Signal Process. Lett. **14**(10), 707–710 (2007)
21. Fan, J.Q., Li, R.Z.: Variable selection via nonconcave penalized likelihood and its oracle properties. J. Am. Stat. Assoc. **96**(456), 1348–1360 (2001)
22. Giuzio, M., Ferrari, D., Paterlini, S.: Sparse and robust normal and t-portfolios by penalized lq-likelihood minimization. Eur. J. Oper. Res. **250**(1), 251–261 (2016)
23. Le Thi, H.A., Dinh, T.P., Le, H.M., Vo, X.T.: DC approximation approaches for sparse optimization. Eur. J. Oper. Res. **244**(1), 26–46 (2015)
24. Zhang, C.H.: Nearly unbiased variable selection under minimax concave penalty. Annal. Stat. **38**(2), 894–942 (2010)
25. Xu, Z.B., Chang, X.Y., Xu, F.M., Zhang, H.: $L_{1/2}$ regularization: a thresholding representation theory and a fast solver. IEEE Trans. Neural Netw. Learn. Syst. **23**(7), 1013–1027 (2012)
26. Burachik, R.S., Rubinov, A.: Abstract convexity and augmented Lagrangians. SIAM J. Optim. **18**(2), 413–436 (2007). https://doi.org/10.1137/050647621
27. Huang, X., Yang, X.: A unified augmented Lagrangian approach to duality and exact penalization. Math. Oper. Res. **28**(3), 533–552 (2003). https://doi.org/10.1287/moor.28.3.533.16395
28. Luo, Z., Pang, J., Ralph, D.: Mathematical Programs with Equilibrium Constraints. Cambridge University Press, Cambridge (1996)
29. Yang, X., Huang, X.: A nonlinear Lagrangian approach to constrained optimization problems. SIAM J. Optim. **11**(4), 1119–1144 (2001)

30. Dong, Z.L., Yang, X.Q., Dai, Y.H.: A unified recovery bound estimation for noise-aware ℓ_q optimization model in compressed sensing. arXiv preprint arXiv:1609.01531 (2016)
31. Song, C.B., Xia, S.T.: Sparse signal recovery by ℓ_q minimization under restricted isometry property. IEEE Signal Process. Lett. **21**(9), 1154–1158 (2014)
32. Hu, Y.H., Li, C., Meng, K.W., Qin, J., Yang, X.Q.: Group sparse optimization via $\ell_{p,q}$ regularization. J. Mach. Learn. Res. **18**(30), 1–52 (2017)
33. Liu, H.C., Yao, T., Li, R.Z., Ye, Y.Y.: Folded concave penalized sparse linear regression: sparsity, statistical performance, and algorithmic theory for local solutions. Math. Progr. **166**(1–2), 207–240 (2017)
34. Loh, P.L., Wainwright, M.J.: Regularized M-estimators with nonconvexity: statistical and algorithmic theory for local optima. J. Mach. Learn. Res. **16**(1), 559–616 (2015)
35. Zhang, C.H., Zhang, T.: A general theory of concave regularization for high-dimensional sparse estimation problems. Stat. Sci. **27**(4), 576–593 (2012)
36. Candès, E.J., Romberg, J.K., Tao, T.: Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. IEEE Trans. Inf. Theory **52**(2), 489–509 (2006)
37. Agarwal, A., Negahban, S., Wainwright, M.J.: Fast global convergence of gradient methods for high-dimensional statistical recovery. Ann. Stat. **40**(5), 2452–2482 (2012)
38. Zhou, S.H.: Restricted eigenvalue conditions on subgaussian random matrices. arXiv preprint arXiv:0912.4045 (2009)
39. Rao, C.R., Statistiker, M.: Linear Statistical Inference and Its Applications. Wiley, New York, New York (1973)
40. van de Geer, S.A.: High-dimensional generalized linear models and the Lasso. Annal. Stat. **36**(2), 614–645 (2008)
41. Loh, P.L., Wainwright, M.J.: High-dimensional regression with noisy and missing data: provable guarantees with nonconvexity. Annal. Stat. **40**(3), 1637–1664 (2012)
42. Negahban, S., Ravikumar, P., Wainwright, M.J., Yu, B.: A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. Stat. Sci. **27**(4), 538–557 (2012)
43. Recht, B., Fazel, M., Parrilo, P.A.: Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. SIAM Rev. **52**(3), 471–501 (2010)
44. Mazumder, R., Radchenko, P.: The discrete dantzig selector: estimating sparse linear models via mixed integer linear optimization. IEEE Trans. Inf. Theory **63**(5), 3053–3075 (2017)
45. Mazumder, R., Radchenko, P., Dedieu, A.: Subset selection with shrinkage: sparse linear modeling when the SNR is low. arXiv preprint arXiv:1708.03288 (2017)
46. Zhang, C.H., Huang, J.: The sparsity and bias of the Lasso selection in high-dimensional linear regression. Annal. Stat. **36**(4), 1567–1594 (2008)
47. Candès, E.J., Wakin, M.B., Boyd, S.P.: Enhancing sparsity by reweighted l_1 minimization. J. Fourier Anal. Appl. **14**(5–6), 877–905 (2008)
48. Chartrand, R., Yin, W.T.: Iteratively reweighted algorithms for compressive sensing. In: IEEE International Conference on Acoustics, pp. 3869–3872 (2008)
49. Blumensath, T., Davies, M.E.: Iterative thresholding for sparse approximations. J. Fourier Anal. Appl. **14**(5–6), 629–654 (2008)
50. Herman, J., Kucera, R., Simsa, J.: Equations and Inequalities: Elementary Problems and Theorems in Algebra and Number Theory. Springer, Berlin (2000)
51. Ross, S.: A First Course in Probability. Pearson, London (2009)
52. Raskutti, G., Wainwright, M.J., Yu, B.: Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. IEEE Trans. Inf. Theory **57**(10), 6976–6994 (2011)
53. Ge, R.: A filled function method for finding a global minimizer of a function of several variables. Math. Progr. **46**(1–3), 191–204 (1990). https://doi.org/10.1007/BF01585737