L_1-regularized Quantile Regression with Many Regressors under Lean Assumptions

Lan Wang

Abstract

L_1-regularized quantile regression (l_1-QR) provides a fundamental technique for analyzing high-dimensional economic data that are heterogeneous with potentially heavy-tailed random errors. This paper investigates conditional quantile estimation when the number of regressors is larger than the sample size. It establishes that l_1-QR possesses properties resembling those of L_1-regularized least squares regression (or LS-Lasso) under generally weaker conditions and enjoys near-optimal performance for a much richer class of error distributions, including some error distributions for which the performance of LS-Lasso is sub-optimal. Our results build upon and substantially generalize the earlier interesting work of Belloni and Chernozhukov (2011) and Wang (2013). We obtain interesting properties for l_1-QR in several novel directions under both the popular hard sparsity assumption and a more relaxed soft sparsity condition that allows many regressors to have small effects. These new theoretical guarantees fill important gaps in the literature and render strong support for the applicability of quantile regression in the high-dimensional setting.

KEY WORDS: Heterogeneous data, L_1 penalty, minimax rate, regularized quantile regression, soft sparsity, strong sparsity, ultra-high dimension.

[1]Lan Wang is Professor, School of Statistics, University of Minnesota. Emails: wangx346@umn.edu. Wang's research was supported by NSF DMS-1712706.

1 Introduction

The semiparametric technique of quantile regression (QR) provides a useful alternative to least-squares regression and has been widely applied to analyze data arising in economics and finance, since its introduction in the seminal paper of Koenker and Bassett (1978). For example, Buchinsky et al. (1994), Chamberlain (1994), Buchinsky (1998), Abadie et al. (2002), Horowitz and Spokoiny (2002), Angrist et al. (2006), Firpo et al. (2009), Galvao et al. (2013), Arellano and Bonhomme (2017), Graham et al. (2018), among others, employed quantile regression to study the wage distribution. See also Fitzenberger et al. (2013) for other interesting applications of quantile regression in economics. Quantile regression helps characterize the entire conditional distribution and often leads to discoveries of interesting features of the data that would otherwise be imperceptible. It also has the advantage of being robust to outlier contamination.

In the traditional framework where the number of regressors is small, the properties of QR have been extensively investigated. We refer to Koenker (2005) and Koenker (2017) for a comprehensive review and related references. In the classical setting, it is well known that the properties of QR estimator resemble those of the least squares regression estimator (LSE) under comparable regularity conditions. In the important special case where the random errors are independent and identically distributed, the asymptotic normality of the QR estimator requires essentially the same conditions on the regressor matrix as LSE does but much weaker conditions on the random error distribution. When the error distribution is non-normal, the QR estimator may be more efficient than LSE. Such solid theoretical grounding contributes crucially to the popularity of classical quantile regression.

The main focus of this paper is on L_1-regularized quantile regression (l_1-QR), which has gradually emerged as a powerful technique for analyzing high-dimensional data that are heterogeneous with potentially heavy-tailed random errors. A notable advantage of quantile regression over least-squares regression in high dimension is that the sparsity pattern is allowed to be heterogeneous. It could happen that the subset of variables relevant for modeling the higher conditional quantile is completely different from that relevant for modeling the lower conditional quantile. In modern economics, it has become common for researchers to collect data on a large number of variables, see for instance the examples in Fan et al. (2011), Horowitz (2015), Belloni et al. (2017), Cattaneo et al. (2018), Ando and Bai (2018). In the last decade, L_1-regularized least squares regression (or LS-Lasso, Tibshirani (1996)) and its variants

have been extensively studied and proven useful for analyzing high-dimensional data, see the monographs of Bühlmann and van de Geer (2011) and Hastie et al. (2015) and the references therein.

We are interested in the properties of l_1-QR in the setting where the number of regressors is greater than the sample size. The nonsmoothness of quantile loss functions brings substantial challenges in developing the theory for l_1-QR in high dimension. The novel work of Belloni and Chernozhukov (2011) was the first to establish the estimation error bounds for l_1-QR when the number of regressors p is large but the number of active regressors is relatively small. See also Wang (2013). Kato (2011) obtained such an estimation error bound for quantile regression with group L_1 penalty in high dimension. Lee et al. (2018) recently studied high-dimensional l_1-QR with a change point. Harding and Lamarche (2017) and Harding and Lamarche (2018) investigated L_1-regularized quantile regression for panel data.

In contrast to the traditional setting, existing work on l_1-QR entail more stringent conditions than those needed for LS-Lasso in the high-dimensional setting. Moreover, it is unknown whether the error bound rates in these work are optimal; whether similar error bounds still hold when many regressors exhibit small effects rather than a few regressors exhibit large effects; and whether there exist situations where l_1-QR possesses more favorable performance than LS-Lasso. Hence, there still exist important gaps in fully understanding the properties of l_1-QR, which seriously hinder its scope of application.

Building on the aforementioned earlier work, this paper made several novel contributions to the fundamental properties of L_1-regularized quantile regression, which would render strong support for the applicability of quantile regression in the high-dimensional setting. We carefully study the properties of l_1-QR under both the popular hard sparisty assumption (as assumed in Belloni and Chernozhukov (2011) and Wang (2013)) and a more relaxed soft sparsity assumption, the latter of which permits many regressors to have small effects. In particular, this paper addresses the following critical questions.

- Does l_1-QR enjoy properties resembling LS-Lasso in high dimension under similar or even weaker conditions?

- Is the estimation error rate derived in the earlier work for l_1-QR optimal in some sense or can some estimator improve upon this rate?

- Is there any scenario for which l_1-QR can outperform LS-Lasso?

The *first main contribution* of this paper is to derive the L_2-error bound for l_1-QR under a set of relaxed conditions in the ultrahigh dimensional setting, where p can grow at an exponential rate of n. The error bound shares the same rate as in Belloni and Chernozhukov (2011) and Wang (2013) but requires much weaker conditions. It is worthy emphasizing that our conditions are generally weaker than those needed for establishing similar error bounds for LS-Lasso.

The *second main contribution* of this paper is to prove that the above error bound is near minimax optimal. To the best of our knowledge, the minimax error rate for l_1-QR has not been investigated in the literature for the high-dimensional scenario. An important discovery is that this minimax lower bound for l_1-QR holds for a much richer class of error distributions. In contrast, similar minimax results were derived under Gaussian random error assumption for LS-Lasso. Further, we provide both theoretical and numerical evidence for scenarios where l_1-QR can outperform LS-Lasso for a class of heavy-tailed error distributions. As the *third main contribution*, this paper also studies the prediction error for the conditional quantile under a set of relaxed conditions.

The outline of the paper is as follows. Section 2 introduces the background and sets up the problem. Section 3 presents the estimation error bound for l_1-QR in ultra-high dimension. Section 4 derives the minimax error bounds; while Section 5 studies the quantile prediction error bound. Section 6 reports results from a Monte Carlo study. Lastly, we conclude with some discussions in Section 7. All the technical details are deferred to the appendices.

2 Identification of quantile regression parameter in high dimension

2.1 Background

Let Y be a random variable and $X \in \mathbb{R}^p$ be a p-dimensional vector of regressors. For a given $0 < \tau < 1$, the τth conditional quantile of Y given X is defined as

$$Q_{Y|X}(\tau) = \inf\{t : F_{Y|X}(t) \geq \tau\},$$

where $F_{Y|X}(\cdot)$ is the conditional distribution function of Y given X.

The popular linear quantile regression assumes $Q_{Y|X}(\tau) = X'\beta^*$, where $\beta^* = (\beta_1^*, \beta_2^*, ..., \beta_p^*)'$ depends on the quantile level of interest τ but we ignore such dependence in notation for simplicity. Equivalently, we may write

$$Y = X'\beta^* + \epsilon, \qquad P(\epsilon \leq 0|X) = \tau, \tag{1}$$

where the random error $\epsilon = Y - Q_{Y|X}(\tau)$ is allowed to have a heteroscedastic distribution.

We study estimating β^* in the scenario where the number of regressors p is much larger than the sample size n. Our theory allows p to grow at an exponential rate of n, that is $p = \exp(o(n^{d_0}))$ for some positive constant $d_0 > 0$, which is known as the ultrahigh dimensional scheme.

In the classical asymptotic framework where p is smaller than n, it is usually assumed that $\mathbb{X} = (\mathbb{X}_1, \ldots, \mathbb{X}_p)$, the $n \times p$ matrix of regressors, has a full column rank. This allows to uniquely identify β^*. In contrast, in the high-dimensional scenario where $p \gg n$, β^* is generally not identifiable in the absence of additional structural assumption as $\mathbb{X}$ has at most column rank n.

2.2 Hard Sparsity and Weak Sparsity in High Dimension

This subsection introduces two different structural assumptions for β^*, strong sparsity and soft sparsity, to help identify β^* in high dimension. As is customary for a regression model in the classical setting, the intercept term is always included in the model. The sparsity constraints are thus imposed on the slope components of β^*. Note that this is a subtle difference from high-dimensional least squares regression where the intercept is generally taken as zero, which can be done for mean regression without loss of generality by centering both the response variable and the regressors but is not feasible for quantile regression.

Under the hard sparsity assumption,

$$\beta^* \in \mathbb{B}_0(s) = \Big\{\beta \in \mathbb{R}^p : \sum_{i=2}^{p} \mathrm{I}(\beta_i \neq 0) \leq s - 1\Big\}, \tag{2}$$

for some positive constant $2 \leq s \ll n$, where $\mathrm{I}(\cdot)$ denotes the indicator function. Hence, $||\beta^*||_0 \leq s$ and most of the components in β^* are exactly zero. This is the most

popular structural assumption for high-dimensional linear regression and is adopted for most of the existing theory on L_1-regularized least squares regression.

The hard sparsity constraint may be overly restrictive for some applications where many weak signals, rather than just a few strong signals, are likely to be present. This paper also considers a more relaxed sparsity constraint, which allows β^* to have many smallish nonzero coefficients. More specifically, the soft sparsity constraint assumes

$$\beta^* \in \mathbb{B}_1(R) = \big\{\beta \in \mathbb{R}^p : \sum_{i=2}^{p} |\beta_i| \leq R\big\} \tag{3}$$

for some positive number R, which may depend on the sample size. In (3), instead of using L_1 norm, we may also use a L_q norm for some $0 < q < 1$. The results of the paper would still hold under minor modification. It is worth noting that both $\mathbb{B}_0(s)$ and $\mathbb{B}_1(R)$ depend on the quantile level of interest.

2.3 Identification of Parameters

In high dimension, β^* in general is not uniquely defined. Suppose model (1) is satisfied by $\beta^* = \beta_0^*$. Consider the affine space $\{\beta^* \in \mathbb{R}^p : \mathbb{X}\beta^* = \mathbb{X}\beta_0^*\}$. *We emphasize that* the error bounds derived in this paper apply to any β^* from this affine space and does not require the unique identification.

Let $\text{Ker}(\mathbb{X}) = \{\beta \in \mathbb{R}^p : \mathbb{X}\beta = 0\}$ be the null space of $\mathbb{X}$. If β^* satisfies (1), then $\beta^* + \beta$ also satisfies (1), $\forall\ \beta \in \text{Ker}(\mathbb{X})$. The extent of identifiability can be measured by the diameter of the set $N_0(\mathbb{X}) = \text{Ker}(\mathbb{X}) \cap \mathbb{B}$, defined as $\max_{\beta \in N_0(\mathbb{X})} ||\beta||_2$, where $\mathbb{B} = \mathbb{B}_0(s)$ under the hard sparsity assumption, while $\mathbb{B} = \mathbb{B}_1(R)$ under the soft sparsity assumption; The following lemma characterizes the properties of the diameter of $N_0(\mathbb{X})$.

Lemma 2.1 *Assume the vector $X_{-i} = (X_{i2}, \ldots, X_{ip})'$ is a mean-zero sub-Gaussian random vector, i.e., Assumption 2 in Section 3.3 is satisfied.*
(i) (hard sparsity case) Assume $\eta_{min}(s) = \inf_{v:||v||_2=1, ||v||_0 \leq s} v'\Sigma v > 0$, where $\eta_{min}(s)$ is the smallest s-sparse eigenvalue of $\Sigma = E(X_i X_i')$. Then

$$P\big(\max_{\beta \in N_0(\mathbb{X})} ||\beta||_2 = 0\big) \geq 1 - \alpha_1^* \exp(-\alpha_2^* \log p),$$

where α_1^* and α_2^* are positive universal constants.

(ii) (soft sparsity case),

$$P\big(\max_{\beta\in N_0(\mathbb{X})}||\beta||_2 \le \alpha'\xi_{min}^{-1/2}R\sqrt{\log p/n}\big) \ge 1-\alpha_1^*\exp(-\alpha_2^*\log p),$$

where α_1^*, α_2^* *and* α' *are positive universal constants, and* ξ_{min} *is the smallest eigenvalue of* Σ.

The above lemma can be considered as a generalization of Lemma 1 of Raskutti et al. (2011) to the random regressor case. From this lemma, it can be seen that under some general conditions, $\mathrm{Ker}(\mathbb{X}) = \{0\}$ with high probability for the hard sparsity case, i.e., the sparse β^* satisfying (1) is unique. While $\mathrm{Ker}(\mathbb{X})$ is a shrinking neighborhood around 0 with high probability for the soft sparsity case if $\xi_{min}^{-1/2}R\sqrt{\log p/n} \to 0$.

3 Estimation error bounds for L_1-regularized Quantile Regression with Many Regressors

3.1 L_1-regularized Quantile Regression

We strat with a brief review of the methodology of l_1-QR. Consider a random sample $\{Y_i, X_i\}_{i=1}^n$, where $X_i = (X_{i1}, \ldots, X_{ip})'$. To explicitly incorporate the intercept term, we write $X = (X_1, \ldots, X_p)' = (1, X_-')'$ where $X_- = (X_2, \ldots, X_p)'$; and correspondingly $\beta^* = (\beta_1^*, \beta_-^{*'})'$ where $\beta_-^* = (\beta_2^*, \ldots, \beta_p^*)'$. To avoid overfitting, we estimate β^* by the L_1-regularized quantile regression estimator

$$\widehat{\beta} = \arg\min_{\beta\in\mathbb{R}^p}\Big\{n^{-1}\sum_{i=1}^n \rho_\tau(Y_i - X_i'\beta) + \lambda||\beta_-||_1\Big\},$$

where $\rho_\tau(u) = u\{\tau - \mathrm{I}(u<0)\}$ is the quantile loss function (or the so-called check loss function) and $\beta = (\beta_1, \beta_-')' \in \mathbb{R}^p$, $||\cdot||_1$ denotes the L_1 norm of a vector, and $\lambda > 0$ is a penalization parameter that controls the complexity of the solution. A larger value of λ encourages sparser solutions.

As the regularized quantile loss function is convex and piece-wise linear, the regularized estimator can be conveniently computed by linear programming. Hence, the computation of l_1-QR is feasible even if p is very large. For this setting, best subset selection is known to be NP hard and computationally infeasible.

Let $\widehat{\gamma} = \widehat{\beta} - \beta^*$, then

$$\widehat{\gamma} = \underset{\gamma \in \mathbb{R}^p}{\arg\min} \Big\{ Q_n(\gamma) + \lambda ||\beta^*_- + \gamma_-||_1 \Big\},$$

where $\gamma = (\gamma_1, \gamma'_-)'$, $Q_n(\gamma) = n^{-1} \sum_{i=1}^n \rho_\tau(\epsilon_i - X_i'\gamma)$ with ϵ_i being independent random errors satisfying the constraint in (1). Let

$$S_n(0) = n^{-1} \sum_{i=1}^n X_i \xi_i, \tag{4}$$

where $\xi_i = \mathrm{I}(\epsilon_i < 0) - \tau$, then $S_n(0)$ is a subgradient of $Q_n(\gamma)$ at $\gamma = 0$. By the Karush-Kuhn-Tucker (KKT) condition of convex optimization (Boyd and Vandenberghe (2004)), a necessary and sufficient condition for $\widehat{\gamma}$ to be the minimizer is that zero belongs to the subdifferential of the regularized quantile loss function evaluated at $\widehat{\gamma}$. Motivated by the general principal of tuning parameter selection in penalized regression (Bickel et al. (2009)), one would take λ such that the event

$$\Lambda_n = \big\{ \lambda \geq c ||S_n(0)||_\infty \big\} \tag{5}$$

happens with a large probability, where $c > 1$ is a positive constant and $||\cdot||_\infty$ denotes the L_∞ norm of a vector.

3.2 Geometric Structure of the Solution

An essential phenomenon regarding l_1-QR is that despite $\widehat{\beta}$ being a high-dimensional vector, it lies in a restricted set with certain geometric structure with high probability. Under the hard sparsity assumption, this restricted set has a cone shape; while under the soft sparsity assumption, this restricted set has a star shape.

More specifically, under the hard sparsity condition, let $S_- = \{j : \beta^*_j \neq 0,\ 2 \leq j \leq p\}$ and $S = S_- \cup \{1\}$. The cardinality $||S||_0 = s$ is called the sparsity size of β^* in this setting. Define the cone set

$$\Gamma_H = \{\gamma \in \mathbb{R}^p : ||\gamma_{S^c}||_1 \leq \bar{c} ||\gamma_S||_1\}, \tag{6}$$

where $\bar{c} = \frac{c+1}{c-1}$ and c is the positive constant in (5).

Under the hard sparsity assumption, the sparsity size s is assumed to be of smaller

order of the sample size n, i.e., most of the regression coefficients are exactly zero. While under the soft sparsity assumption, s can be much larger than n, potentially of the same order of p. Let $S_{-a} = \{j : |\beta_j^*| > a, 2 \leq j \leq p\}$, where a is a small positive thresholding parameter to be discussed later. Let $S_a = S_{-a} \cup \{1\}$. Define

$$\Gamma_W = \big\{\gamma \in \mathbb{R}^p : ||\gamma_{S_a^c}||_1 \leq \bar{c}||\gamma_{S_a}||_1 + \frac{2c}{c-1}||\beta^*_{S_a^c}||_1\big\}, \tag{7}$$

where $c > 1$ is the positive constant in (5), $\bar{c}$ is the same as that in (6), γ_{S_a} denotes the subvector of γ containing the elements whose indice are in S_a, $\gamma_{S_a^c}$ and $\beta^*_{S_a^c}$ are defined similarly. It is observed that Γ_W is start-shaped, that is, if $\gamma \in \Gamma_W$, then the whole line segment $\{t\gamma | t \in (0,1)\}$ is contained in Γ_W.

Lemma 3.1 *On the event Λ_n defined in (5), $\widehat{\gamma} \in \Gamma_H$ under the hard sparsity assumption; and $\widehat{\gamma} \in \Gamma_W$ under the soft sparsity assumption, where a in the definition of Γ_W can be an arbitrary positive constant.*

Remark 1. The fact stated in the above lemma is a straightforward result of the definition of $\widehat{\gamma}$ and the convexity property of the regularized quantile loss function. The first part of the result under hard sparsity was observed in Belloni and Chernozhukov (2011); while the result under soft sparsity is a generalization of Negahban et al. (2012). Furthermore, it is known to be practically possible to choose an appropriate λ such that the event Λ_n is guaranteed to have a high probability. An upper bound of such a λ is given in Lemma **??** in the appendix. Alternatively, such a λ can be simulated, as observed in Belloni and Chernozhukov (2011).

3.3 Three Basic Assumptions

We now introduce a set of three basic assumptions needed to establish the statistical properties of the L_1-regularized quantile regression estimator.

Assumption 1. The random error term ϵ_i satisfies $P(\epsilon_i \leq 0 | X_i) = \tau$, $i = 1, \ldots, n$. There exist positive constants m_0 and b_0 such that $\inf_{1 \leq i \leq n} f_i(t) \geq m_0 > 0$, for all $|t| \leq b_0$, where $f_i(t)$ is the conditional probability density function of ϵ_i given X_i, $i = 1, \ldots, n$.

Assumption 2. The vector $X_{-i} = (X_{i2}, \ldots, X_{ip})'$ is a sub-Gaussian random vector

with mean zero and parameter (or variance proxy) bounded by a positive constant σ_x, that is, for any $(p-1)$-dimensional unit vector v, $\mathrm{E}\{\exp(tX_-^T v)\} \leq \exp(t^2\sigma_x^2/2)$, for any $t \in \mathbb{R}$.

Assumption 3. Let $\Sigma = \mathrm{E}(X_i X_i')$. It is assumed that

$$v'\Sigma v \geq m_1 ||v||_2^2, \quad \text{for all nonzero vector } v \in \Gamma, \tag{8}$$

where $\Gamma = \Gamma_H$ under the hard sparsity assumption, and $\Gamma = \Gamma_W$ under the soft sparsity assumption, and $||\cdot||_2$ denotes the L_2 norm of a vector.

Remarks 2-4 below compare the above three assumptions with those required for LS-Lasso. In short words, the three basic assumptions are no stronger than the conditions usually required to establish similar properties for LS-Lasso in the high-dimensional setting, see Meinshausen and Bühlmann (2006), Zhao and Yu (2006), Bunea et al. (2007), Van de Geer et al. (2008), Zhang et al. (2008), Bickel et al. (2009), Wainwright (2009), among others. Moreover, these assumptions are substantially weaker than the conditions in the existing literature for high-dimensional l_1-QR, which will be discussed in details in Remark 5 of Section 3.4.

Remark 2. We first note that the conditions imposed on the random error ϵ_i in Assumption 1 are weaker than those usually required for LS-Lasso in high dimension, which rely on the sub-Gaussian tail condition for ϵ_i. The b_0 in Assumption 1 can be an arbitrarily small positive constant or even be a sequence of positive constants that converges to zero. Hence our assumptions on ϵ_i are essentially the same as those for classical quantile regression, which assume $f_i(t)$ is continuous and uniformly lower bounded by a positive constant at 0. This permits a much richer class of error distributions, including heavy-tailed distributions such as Cauchy distribution.

Remark 3. The sub-Gaussian distribution assumption for X_- in Assumption 2 is also a popular assumption for studying the theory of high-dimensional LS-Lasso. The class of sub-Gaussian distributions encompasses many commonly used discrete and continuous distributions. Alternatively, one may replace the random regressors assumptions with fixed regressors assumptions, for which case the restricted eigenvalue conditions hold for the sample covariance matrix with high probability under the conditions stated

here.

Remark 4. Assumptions 3 is similar to the the restricted eigenvalue assumption (Bickel et al. 2009) for LS-Lasso. This is an identifiability condition in high dimension. It essentially requires that the expected quantile loss function has sufficient curvature, that is, not too flat, when being confined to the restricted subset Γ_H or Γ_W. As indicated by Lemma 3.1, with the appropriate choice of the regularization parameter λ, $\widehat{\gamma} = \widehat{\beta} - \beta^*$ belongs to the restricted set Γ_H or Γ_W, under the hard sparsity or the soft sparsity setting, respectively.

3.4 Estimation Error Bound

This subsection derives the L_2 estimation error bound for $\widehat{\gamma} = \widehat{\beta} - \beta^*$ under both the hard sparsity assumption and the soft sparsity assumption. It is worth emphasizing that the results in this and other sections of this paper are nonasymptotic in the sense that the bounds hold for each n greater than a universal constant. The theory allows the number of regressors p to grow at an exponential rate of the sample size n, i.e., the ultra-high dimensional setting.

Theorem 3.2 below shows that under the hard sparsity assumption, the L_2 estimation error for l_1-QR enjoys a near-oracle performance under a set of relatively weak assumptions. The L_2 estimation error is of order $\sqrt{s \log p/n}$. In the oracle case when one knows which variables are relevant, the estimation error is of order $\sqrt{s/n}$. Hence, the rate for the high-dimensional setting is near-oracle, up to an order $\sqrt{\log p}$, the price needed to pay for the fact that there are $p \gg n$ variables and it is not known in advance which ones are relevant.

Theorem 3.2 *(hard sparsity case) Suppose Assumptions 1-3 are satisfied and β^* satisfies the hard sparsity assumption (2). Let $\lambda = k_0\sqrt{\log p/n}$, where $k_0 \geq c \max\{2\sigma_x, 1\}$. Assume there exist positive constants a_i, $i = 0, 1$ and 2, such that the triple (n, s, p) satisfies*

$$1 \leq s \leq a_0 n^{\eta_0}, \quad a_1 n^{\eta_1} \leq p \leq a_2 \exp(n^{\eta_2})$$

for some constants $0 < \eta_0 < 1$, $\eta_1 > 0$ and $0 < \eta_2 < 1 - \eta_0$. Then there exist universal positive constants N_0, Δ_0 and c^*, all independent of the triple (n, s, p), such that for

any $n > N_0$, $\forall\ \Delta > \Delta_0$,

$$P\big(||\widehat{\beta} - \beta^*||_2 \leq \Delta\sqrt{s\log p/n}\big) \geq 1 - c^* \exp(-\log p).$$

Remark 5. Theorem 3.2 builds on the earlier work of Belloni and Chernozhukov (2011) and Wang (2013), who were the first to study the L_2 estimation error bound for l_1-QR with the focus on the hard-sparsity case. Our work establishes that the same convergence rate can be achieved under substantially weaker conditions. Comparing with the conditions in Belloni and Chernozhukov (2011), we relaxed the conditions on both Σ and ϵ_i.

- We have dropped their *restricted nonlinearity condition* on Σ (their condition D.4), which would require $q := \inf_{\delta\in A,\delta\neq 0} \frac{\{E(|X_i'\delta|^2)\}^{3/2}}{E(|X_i'\delta|^3)} > 0$ for some restricted set A. Such a similar condition is not needed for the parallel theory of LS-Lasso. Furthermore, if the non-linear impact coefficient q converges to zero at a sufficiently fast rate, this may have a negative impact on the feasible range of n and p through the *growth condition* $\sqrt{s\log(p\vee n)} \leq O(q\sqrt{n})$, which was assumed in the main theorem (Theorem 2) of Belloni and Chernozhukov (2011).

- Unlike Belloni and Chernozhukov (2011), we do not require the conditional random error density $f_i(t)$ to be continuously differential nor the derivative to be uniformly bounded everywhere. We only need a uniform lower bound for $f_i(t)$ in a small local region. This essentially reduces the random error assumption to that in the classical quantile regression setting.

Our assumptions are also significantly weaker than those in Wang (2013), which required independent and identically random errors and a restricted isometry type condition in additional to the restricted eigenvalue condition.

Remark 6. The regularization parameter λ is taken to be of the order $\sqrt{\log p/n}$, the universal penalty level introduced in Donoho and Johnstone (1994). In practice, an appealing approach (Belloni and Chernozhukov (2011)) is to directly simulate λ as the $(1-\alpha)$-quantile of the distribution of $c||S_n(0)||_\infty$, for some small $\alpha > 0$. This is feasible by observing that the distribution of $||S_n(0)||_\infty$ is pivotal and completely known. With this simulated choice of λ, the same estimation error bound would hold with probability $1 - \alpha - c^* \exp\big(-\log p\big)$ instead of $1 - c^* \exp\big(-\log p\big)$.

Theorem 3.3 below provides the estimation error bound under the soft sparsity assumption. Note that the soft sparsity assumption does not require the existence of an approximately L_0-sparse (or hard-sparse) model.

Theorem 3.3 *(soft sparsity case) Suppose Assumptions 1-3 are satisfied and β^* satisfies the soft sparsity assumption (3). Let $\lambda = k_0\sqrt{\log p/n}$, where $k_0 \geq c\max\{2\sigma_x, 1\}$. Assume there exist positive constants a_i, $i = 0, 1, 2$ and 3, such that the triple (n, R, p) satisfies*

$$a_0 n^{-\eta_0} \leq R \leq a_1 n^{\eta_1}, \quad a_2 n^{\eta_2} \leq p \leq a_3 \exp(n^{\eta_3})$$

for some constants $0 < \eta_0, \eta_1 < (1-\eta_3)/2$, $\eta_2 > 0$ and $0 < \eta_3 < 1$. Then there exist universal positive constants N_0', Δ_0' and c_* (all independent of the triple (n, R, p)) such that for any $n > N_0'$, $\forall\ \Delta > \Delta_0'$,

$$P\big(||\widehat{\beta} - \beta^*||_2 \leq \Delta\sqrt{R}(\log p/n)^{1/4}\big) \geq 1 - c_* \exp\big(-\log p\big).$$

Remark 7. Comparing with the hard-sparsity case, the soft sparsity case has been relatively little discussed in the high-dimensional regression literature and has not been investigated by Belloni and Chernozhukov (2011) and Wang (2013). Our result generalized Negahban et al. (2012) for LS-Lasso. The radius R is allowed to shrink or diverge with the sample size n.

4 Minimax lower bounds for estimation

4.1 Minimax Lower Bounds

We now address the following important questions: Are the rates of the estimation error bound derived in Section 3 for l_1-QR optimal in certain sense? And if so, for which class of distributions?

These problems have not been investigated in the literature but are of critical importance for understanding the properties of l_1-QR in high dimension. Only when one fully understands these properties, it is possible to have a fair comparison of l_1-QR with LS-Lasso.

For LS-Lasso, minimax lower bounds for estimation error were derived in the concurrent work of Raskutti et al. (2011) and Ye and Zhang (2010). Our results show

that the same minimax rates of estimation error hold for l_1-QR while allowing for a broader class of distributions. Formally, let $\{\mathbb{P}(\beta) : \beta \in \mathbb{B}\}$ denote the class of joint distributions of $\{(X_i, Y_i) : i = 1, \ldots, n\}$ indexed by β such that:

i. the τth conditional quantile function of Y_i given X_i has the coefficient vector $\beta \in \mathbb{B}$, where $\mathbb{B} = \mathbb{B}_0(s)$ under the hard sparsity assumption, while $\mathbb{B} = \mathbb{B}_1(R)$ under the soft sparsity assumption;

ii. (X_i, Y_i) satisfies Assumptions 1-3 in Section 3.3.

Note that for each $\beta \in \mathbb{B}$, $\mathbb{P}(\beta)$ encompasses a large semiparametric class of distributions. Our goal in this section is to provide a lower bound for the following minimax estimation error risk

$$\min_{\widehat{\beta}} \sup_{\mathbb{P}(\beta^*), \beta^* \in \mathbb{B}} \mathrm{E}_{\mathbb{P}(\beta)} \big\{ ||\widehat{\beta} - \beta^*||^2 \big\}, \tag{9}$$

where $\widehat{\beta}$ denotes an arbitrary estimator.

Before we present the main results, we first point out that an intrinsic lower bound exists due to the nature of the problem. Recall in Section 2.3, we defined $N_0(\mathbb{X}) = \mathrm{Ker}(\mathbb{X}) \cap \mathbb{B}$, where $\mathrm{Ker}(\mathbb{X}) = \{\beta \in \mathbb{R}^p : \mathbb{X}\beta = 0\}$ is the null space of $\mathbb{X}$. It is straightforward to see that for any estimator $\widehat{\beta}$, there exists a $\beta \in N_0(\mathbb{X})$ such that $||\widehat{\beta} - \beta|| \geq \frac{1}{2} \max_{\beta \in N_0(\mathbb{X})} ||\beta||_2$, where $\max_{\beta \in N_0(\mathbb{X})} ||\beta||_2$ is defined as the diameter of the $N_0(\mathbb{X})$. Lemma 2.1 characterized the properties of the diameter of $N_0(\mathbb{X})$ under the hard sparsity condition and the soft sparsity condition, respectively. This helps provide an intrinsic lower bound. Any lower bound of smaller order than this would not be of interest.

Theorem 4.1 below states the lower bounds for the minimax estimation error under both the hard sparsity assumption and the soft sparsity assumption. Let $\eta_{max}(k) = \sup_{v: ||v||_2 = 1, ||v||_0 \leq k} v'\Sigma v$ be the largest k-sparse eigenvalue of Σ, where k is a positive integer.

Theorem 4.1 *(i) (hard sparsity case) Assume the conditions of Theorem 3.2 are satisfied. There exist positive universal constants c and N_1 such that $\forall\ n > N_1$,*

$$\min_{\widehat{\beta}} \sup_{\mathbb{P}(\beta^*), \beta^* \in \mathbb{B}_0(s)} E_{\mathbb{P}(\beta^*)} \big\{ ||\widehat{\beta} - \beta^*||^2 \big\} \geq c \eta_{\max}^{-1}(2s) n^{-1} s \log\big(s^{-1} p\big).$$

(ii) (soft sparsity case) Assume the conditions of Theorem 3.3 are satisfied. There exist positive universal constants c' and N_1' such that $\forall\ n > N_1'$,

$$\min_{\widehat{\beta}} \sup_{\mathbb{P}(\beta^*),\beta^* \in \mathbb{B}_1(R)} E_{\mathbb{P}(\beta^*)}\big\{||\widehat{\beta} - \beta^*||^2\big\} \geq c' \eta_{\max}^{-1}(2\tilde{s}) \frac{R}{\sqrt{n}} \log\big(\frac{p}{R\sqrt{n}}\big).$$

where $\tilde{s} = \lceil R\sqrt{n} \rceil$ with $\lceil \cdot \rceil$ being the ceiling function.

Remark 8. The above minimax lower bounds for l_1-QR in high dimension are new and match the rates for LS-Lasso. Both Ye and Zhang (2010) and Raskutti et al. (2011) assume independent and identical normal distribution for ϵ_i in studying LS-Lasso. In contrast, we allow ϵ_i to have a heteroscedastic and potentially heavy-tailed distribution. These verified that the estimation error bounds derived in Section 3 for l_1-QR are near-minimax for a large class of error distributions.

4.2 Could L_1-regularized Quantile Regression be Superior?

A curious and important question is whether there exist situations where l_1-QR would be superior to LS-Lasso? This subsection provides a positive answer to this question.

The LS-Lasso estimator $\widehat{\beta}^{LS}$ is defined as

$$\widehat{\beta}^{LS} = \arg\min_{\beta \in \mathbb{R}^p} \Big\{ (2n)^{-1} ||\mathbb{Y} - \mathbb{X}\beta||_2^2 + \lambda ||\beta_-||_1 \Big\},$$

where $\mathbb{Y} = (Y_1, \ldots, Y_n)'$. A thought-provoking example regarding the sub-optimality of Lasso was given in Loh (2017) when the random error ϵ_i belongs to the class of α-stable distributions. Formally, ϵ_i has an α-stable distribution with scale parameter ξ if its characteristic function

$$E\{\exp(it\epsilon_i)\} = \exp(-\xi^\alpha |t|^\alpha), \quad \forall\ t \in \mathcal{R},$$

and $\alpha \in (0, 2]$. For example, the standard Cauchy distribution is an α-stable distribution with $\alpha = \xi = 1$, see Nolan (2003) for a general introduction to stable distributions.

Following the line of arguments in the proof of Lemma 2 of Loh (2017), if $p = o(\exp(n^{\frac{2-\alpha}{\alpha}}))$, then for any $b_n > 0$ such that $n^{(1-\alpha^{-1})} b_n \sqrt{\log p/n} \to 0$,

$$P\big(n^{-1}||\mathbb{X}'\epsilon||_\infty \geq b_n \sqrt{\log p/n}\big) \geq c_\alpha > 0, \tag{10}$$

where $\epsilon = (\epsilon_1, \ldots, \epsilon_n)'$ and $c_\alpha \leq 1$ is a constant that depends only on the sub-Gaussian parameter of the rows of $\mathbb{X}$ and does not scale with the problem dimension. Consider the hard sparsity case. Following the analysis in Bickel et al. (2009), for $\lambda > \xi_1 ||\mathbb{X}'\epsilon||_\infty$ where $\xi_1 > 1$ is some constant, $\widehat{\beta}^{LS}$ satisfies

$$||\widehat{\beta}^{LS} - \beta^*||_2 \leq \xi_2 \sqrt{s} \max\left\{\lambda, n^{-1}||\mathbb{X}'\epsilon||_\infty\right\}, \tag{11}$$

with high probability, where ξ_2 is a universal positive constant. As (10) suggests that for $0 < \alpha < 2$, there is a positive probability $||\mathbb{X}'\epsilon||_\infty \gg \sqrt{\log p/n}$. The error bound in (10) for $\widehat{\beta}^{LS}$ can be unfavorably large with a positive probability for the heavy-tailed error case, comparing with what would be obtained under the normal error distribution .

As the above analysis is based on the standard approach of Bickel et al. (2009), one may wonder whether a tighter upper bound can be derived using some alternative proof technique. We can strengthen the result of Lemma 2 of Loh (2017) to show that this would not happen by exploring the Karush-Kuhn-Tucker (KKT) condition. Again, we consider the hard sparsity case where $\mathbb{X}$ is a sub-Gaussian random matrix, ϵ_i has an α-stable distribution with $0 < \alpha < 2$, and $\lambda > \xi_1 ||\mathbb{X}'\epsilon||_\infty$ for some $\xi_1 > 1$, a standard choice due to the KKT condition. Consider the L_1-regularized least squares regression $\widehat{\beta}^{LS}$. Assume $\widehat{\beta}^{LS}$ is a non-degenerate solution in the sense that it has at least one non-zero component, say $\widehat{\beta}_j^{LS} \neq 0$ for some $1 \leq j \leq p$. By the KKT condition, $\widehat{\beta}^{LS}$ must satisfy

$$e_j'(n^{-1}\mathbb{X}'\mathbb{X})(\beta^* - \widehat{\beta}^{LS}) + e_j' n^{-1}\mathbb{X}'\epsilon + \lambda \text{sign}(\widehat{\beta}_j^{LS}) = 0, \tag{12}$$

where e_j is a p-dimensional unit vector with the j entry being one and all the other entries being zero. Assume the contrary is true, that is, $||\beta^* - \widehat{\beta}^{LS}||_2 \leq \xi_3 \sqrt{s \log p/n}$ for some positive constant ξ_3. Then $\left|e_j'(n^{-1}\mathbb{X}'\mathbb{X})(\beta^* - \widehat{\beta}^{LS})\right| \leq \xi_4 \sqrt{s \log p/n}$ with high probability, where the positive constant ξ_4 depends on the largest eigenvalue of Σ; on the other hand $\left|n^{-1}\mathbb{X}'\epsilon + \lambda \text{sign}(\widehat{\beta}_j^{LS})\right| \geq (\xi_3 - 1)n^{-1}||\mathbb{X}'\epsilon||_\infty$. It follows from (10) that

$$P\big(n^{-1}||\mathbb{X}'\epsilon||_\infty \geq b_n \sqrt{s \log p/n}\big) \geq c_\alpha > 0,$$

as long as $n^{(1-\alpha^{-1})} b_n \sqrt{s \log p/n} \to 0$. For example if ϵ_i has a standard Cauchy distribution, this means $b_n \sqrt{s \log p/n} \to 0$. As we assume $s \log p/n = o(1)$. This allows

$b_n \to \infty$ as $n \to \infty$. Hence, the KKT condition in (12) breaks down with a positive probability.

In summary, l_1-QR achieves the near-minimax estimation error rate for a larger class of error distributions, including those heavy-tailed error distribution for which the estimation error rate for LS-Lasso would be sub-optimal.

5 Quantile prediction error bounds

In this section, we develop some theoretical guarantee for the prediction error for l_1-QR. Recall the notation $\widehat{\gamma} = \widehat{\beta} - \beta^* = (\widehat{\gamma}_1, \widehat{\gamma}'_-)'$ and $Q_n(\gamma) = n^{-1} \sum_{i=1}^n \rho_\tau(\epsilon_i - X_i'\gamma)$. The prediction error for quantile regression can be evaluated using the check loss function. Specifically, we will consider the empirical prediction error defined as $R_n(\widehat{\gamma})$, where

$$R_n(\gamma) = Q_n(\gamma) - Q_n(0).$$

Our results extend the prediction error bound for L_1-regularized least-squares regression, see Greenshtein et al. (2004), Bunea et al. (2007), Bickel et al. (2009) and Raskutti et al. (2011). Different from quantile regression, the precition error bound for LS-Lasso is usually studied based on the the least squares loss function $n^{-1} \sum_{i=1}^n (\epsilon_i - X_i'\gamma)^2$.

The main messages from this section are the following: (i) Unlike estimation, prediction consistency can be obtained without restricted eigenvalue type conditions and no sparseness condition on the regression coefficient vector. This result is interesting despite the convergence rate is not optimal. (ii) Under additional conditions required by Theorem 3.2 and Theorem 3.3, faster convergence rates for the quantile prediction error can be achieved.

Theorem 4 below summarizes the results for the quantile prediction error. These results are new for l_1-QR in high dimension.

Theorem 5.1 *(i) (prediction error bound without sparsity assumption) For any tuning parameter λ, for any n, we have*

$$R_n(\widehat{\gamma}) \leq 2\lambda ||\beta^*||_1.$$

(ii) (faster rate, hard sparsity case) Assume the conditions of Theorem 3.2 are satisfied.

Then for any $n > N_0$, $\forall\ \Delta > \Delta_0$,

$$R_n(\widehat{\gamma}) \leq \frac{(1+\bar{c})}{k_0}\Delta s\lambda^2,$$

with probability at least $1 - c_1^* \exp\big(-c_2^* s \log p\big)$, where c_1^* and c_2^* are positive universal constants.

(iii) (faster rate, soft sparsity case) Assume the conditions of Theorem 3.3 are satisfied. Then for any $n > N_0'$, $\forall\ \Delta > \Delta_0'$,

$$R_n(\widehat{\gamma}) \leq \frac{2c}{c-1}\big\{k_0^{-1/2}\sqrt{||S_a||_0}\Delta\sqrt{R}\lambda^{3/2} + ||\beta^*_{S_a^c}||_1\lambda\big\},$$

with probability at least $1 - c_3^* \exp\big(-c_4^* s \log p\big)$, where c_3^* and c_4^* are positive universal constants and the set S_a is defined in Section 3 with a being an arbitrary positive constant.

Remark 9. The rates given in (i) and (ii) are the same as those in the literature for LS-Lasso. The rate in (i) is usually referred to as the "slow rate", while the rate in (ii) is referred to as as "faster rate". Intuitively, in the classical oracle case, the prediction error is of order n^{-1}; while in such a setting with the universal tuning parameter $\lambda = k_0\sqrt{\log p/n}$, the bound in (i) is roughly of order $n^{-1/2}$, while the bound in (ii) is roughly of order n^{-1}. However, (i) does imply that quantile prediction consistency can be achieved if $||\beta^*||_1 = o(\sqrt{n/\log p})$ without requiring restricted eigenvalue conditions on Σ. In the soft sparsity case, the upper bound is slightly different from the rate in Raskutti et al. (2011) for LS-Lasso, which is of order $R\lambda$ and coincides with our "slow rate" in (i). In contrast, our upper bound in (iii) permits faster rate when the true value β^* has certain desirable structural property: particularly when the number of relatively large signals in β^* is of relatively small order, while the number of relatively small signals are much smaller comparing with R. Finally, we note that for all scenarios, we require overall milder conditions on the random errors comparing with those in the literature for LS-Lasso.

6 A Monte Carlo experiment

We first generate $(X_1, X_2, \cdots, X_p)$ from the multivariate normal distribution $N_p(0, \Sigma)$ with $\Sigma = (\sigma_{jk})_{p\times p}$ and $\sigma_{jk} = 0.5^{|j-k|}$, $1 \leq j, k \leq p$. For the regression parameter β^*,

we consider two different models.

- Model 1 (sparser model): $\beta^* = (2, 1, 1.5, 1.75, 0'_{p-4})'$,
- Model 2 (denser model): $\beta^* = \frac{3}{n}(1_n, 0'_{p-n})'$,

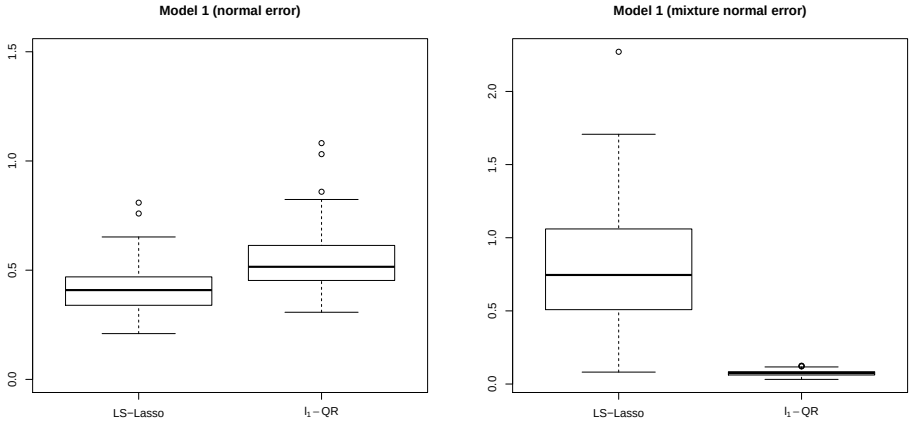
where 0_k denotes a k-dimensional vector of zeros, while 1_k denotes a k-dimensional vector of ones. For each model, we consider two different random error distributions for ϵ_i: the $N(0,1)$ dsitribution and the mixture normal distribution $aN(0,1) + (1-a)N(0,10^2)$, where $a \sim$ Bernoulli(0.95).

The boxplots for LS-Lasso and l_1-QR (with $\tau = 0.5$) for the four scenarios are depicted based on 100 simulation runs for $n = 100$ and $p = 500$. For both estimators, the tuning parameter was selected using a 5-fold cross-validation. The top panel of Figure 1 summarizes the results for Model 1. We observe that for the normal random error case LS-Lasso is slightly more efficient than l_1-QR but its performance deteriorates substantially for the mixture normal random error case. The bottom panel of Figure 1 summarizes the results for Model 2. For this denser model, l_1-QR behaves competitively for the normal random error case and has much smaller error for the mixture normal random error case.

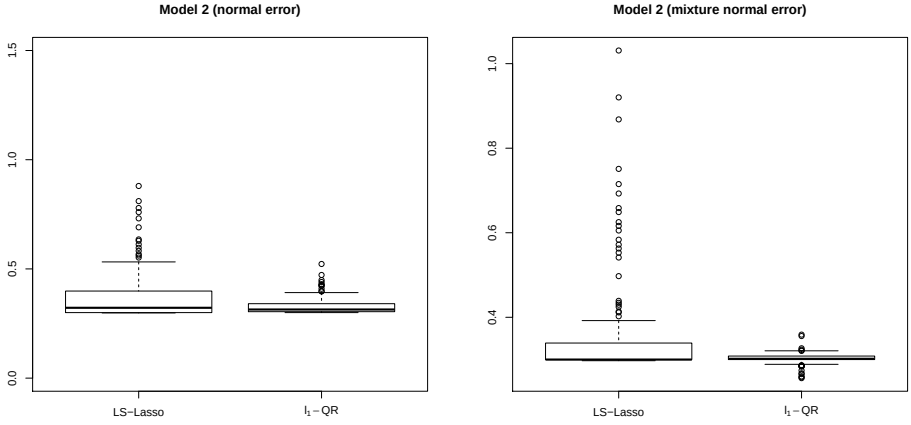
7 Concluding remarks

This paper presents several novel results on the fundamental properties of L_1-regularized quantile regression in high dimension, where the number of regressors can grow at an exponential rate of the sample size. We demonstrate that l_1-QR possesses properties resembling those of LS-Lasso under generally weaker conditions and enjoys near-optimal performance for a much richer class of error distributions, including some error distributions for which the performance of LS-Lasso could be sub-optimal. We expect these findings will render strong support for the applicability of quantile regression in the high-dimensional setting. Indeed, quantile regression has several unique advantages for analyzing high-dimensional heterogeneous data, especially for those with heavy-tailed errors. It has the flexibility to allow the sparsity pattern to depend on the quantile level of interest. That is, one variable can be active (or influential) at one quantile level but not at another quantile level.

This paper focuses on estimating the conditional quantile function at a given quantile level τ. Belloni and Chernozhukov (2011) studied the error bound for l_1-QR uni-



(a) (b)



(c) (d)

Figure 1: Box plots for L_1-regularized quantile regression estimator

formly for τ over an interval. Our results can also be extended to l_1-QR processes. This will be an interesting direction of future work. Finally, our results are also of importance for those procedures that use l_1-QR as the initial value, as those procedures tend to automatically assume the conditions in the literature are met for the first-step l_1-QR to have desirable performance. These in particular include the post-Lasso refitted quantile regression (Belloni and Chernozhukov (2011)) which can help improve finite-sample performance; related inference procedures (Belloni et al. (2014), Zhao et al. (2014)) which is based on a debiased version of l_1-QR; and adaptively weighted L_1 penalized or nonconvex penalized quantile regression (Wang et al. (2012), Fan et al. (2014), Zheng et al. (2015)) which are consistent for variable selection.

References

Abadie, A., Angrist, J., and Imbens, G. (2002). Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. *Econometrica*, 70(1):91–117.

Ando, T. and Bai, J. (2018). Quantile co-movement in financial markets: A panel quantile model with unobserved heterogeneity. *Journal of the American Statistical Association*, (just-accepted):1–35.

Angrist, J., Chernozhukov, V., and Fernández-Val, I. (2006). Quantile regression under misspecification, with an application to the us wage structure. *Econometrica*, 74(2):539–563.

Arellano, M. and Bonhomme, S. (2017). Quantile selection models with an application to understanding changes in wage inequality. *Econometrica*, 85(1):1–28.

Belloni, A. and Chernozhukov, V. (2011). L1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics*, 39:82–130.

Belloni, A., Chernozhukov, V., Fernández-Val, I., and Hansen, C. (2017). Program evaluation and causal inference with high-dimensional data. *Econometrica*, 85(1):233–298.

Belloni, A., Chernozhukov, V., and Kato, K. (2014). Uniform post-selection inference for least absolute deviation regression and other z-estimation problems. *Biometrika*, 102(1):77–94.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge: Cambridge University Press.

Buchinsky, M. (1998). The dynamics of changes in the female wage distribution in the usa: a quantile regression approach. *Journal of applied econometrics*, 13(1):1–30.

Buchinsky, M. et al. (1994). Changes in the us wage structure 1963-1987: Application of quantile regression. *ECONOMETRICA-EVANSTON ILL-*, 62:405–405.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

Bunea, F., Tsybakov, A., Wegkamp, M., et al. (2007). Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics*, 1:169–194.

Cattaneo, M. D., Jansson, M., and Newey, W. K. (2018). Inference in linear regression models with many covariates and heteroscedasticity. *Journal of the American Statistical Association*, 113(523):1350–1361.

Chamberlain, G. (1994). Quantile regression, censoring, and the structure of wages. In *Advances in econometrics: sixth world congress*, volume 2, pages 171–209.

Donoho, D. L. and Johnstone, I. M. (1994). Minimax risk overl p-balls forl p-error. *Probability Theory and Related Fields*, 99(2):277–303.

Fan, J., Fan, Y., and Barut, E. (2014). Adaptive robust variable selection. *Annals of statistics*, 42(1):324.

Fan, J., Lv, J., and Qi, L. (2011). Sparse high-dimensional models in economics. *Annu. Rev. Econ.*, 3(1):291–317.

Firpo, S., Fortin, N. M., and Lemieux, T. (2009). Unconditional quantile regressions. *Econometrica*, 77(3):953–973.

Fitzenberger, B., Koenker, R., and Machado, J. A. (2013). *Economic applications of quantile regression*. Springer Science & Business Media.

Galvao, A. F., Lamarche, C., and Lima, L. R. (2013). Estimation of censored quantile regression for panel data with fixed effects. *Journal of the American Statistical Association*, 108(503):1075–1089.

Graham, B. S., Hahn, J., Poirier, A., and Powell, J. L. (2018). A quantile correlated random coefficients panel data model. *Journal of Econometrics*, 206(2):305–335.

Greenshtein, E., Ritov, Y., et al. (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli*, 10(6):971–988.

Harding, M. and Lamarche, C. (2017). Penalized quantile regression with semiparametric correlated effects: An application with heterogeneous preferences. *Journal of Applied Econometrics*, 32(2):342–358.

Harding, M. and Lamarche, C. (2018). A panel quantile approach to attrition bias in big data: Evidence from a randomized experiment. *Journal of Econometrics*.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical learning with sparsity: the lasso and generalizations*. CRC press.

Horowitz, J. L. (2015). Variable selection and estimation in high-dimensional models. *Canadian Journal of Economics/Revue canadienne d'économique*, 48(2):389–407.

Horowitz, J. L. and Spokoiny, V. G. (2002). An adaptive, rate-optimal test of linearity for median regression models. *Journal of the American Statistical Association*, 97(459):822–835.

Kato, K. (2011). Group lasso for high dimensional sparse quantile regression models. *arXiv preprint arXiv:1103.1458*.

Koenker, R. (2005). *Quantile Regression*. Cambridge: Cambridge University Press.

Koenker, R. (2017). Quantile regression: 40 years on. *Annual Review of Economics*, 9:155–176.

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica*, 46:33–50.

Ledoux, M. and Talagrand, M. (2013). *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media.

Lee, S., Liao, Y., Seo, M. H., and Shin, Y. (2018). Oracle estimation of a change point in high dimensional quantile regression. *Journal of the American Statistical Association*, 43:1184–1194.

Loh, P.-L. (2017). Statistical consistency and asymptotic normality for high-dimensional robust m-estimators. *The Annals of Statistics*, 45(2):866–896.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, pages 1436–1462.

Negahban, S. N., Ravikumar, P., Wainwright, M. J., Yu, B., et al. (2012). A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557.

Nolan, J. (2003). *Stable distributions: models for heavy-tailed data*. Birkhauser New York.

Raskutti, G., Wainwright, M. J., and Yu, B. (2011). Minimax rates of estimation for high-dimensional linear regression over l_q-balls. *IEEE transactions on information theory*, 57(10):6976–6994.

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

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

Van de Geer, S. (2016). *Estimation and testing under sparsity*. Springer.

Van de Geer, S. A. et al. (2008). High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 36(2):614–645.

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using l_1-constrained quadratic programming (lasso). *IEEE transactions on information theory*, 55(5):2183–2202.

Wang, L. (2013). The l1 penalized lad estimator for high dimensional linear regression. *Journal of Multivariate Analysis*, 120:135–151.

Wang, L., Wu, Y., and Li, R. (2012). Quantile regression for analyzing heterogeneity in ultra-high dimension. *Journal of the American Statistical Association*, 107(497):214–222.

Ye, F. and Zhang, C.-H. (2010). Rate minimaxity of the lasso and dantzig selector for the lq loss in lr balls. *Journal of Machine Learning Research*, 11(Dec):3519–3540.

Zhang, C.-H., Huang, J., et al. (2008). The sparsity and bias of the lasso selection in high-dimensional linear regression. *The Annals of Statistics*, 36(4):1567–1594.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2563.

Zhao, T., Kolar, M., and Liu, H. (2014). A general framework for robust testing and confidence regions in high-dimensional quantile regression. *arXiv preprint arXiv:1412.8724*.

Zheng, Q., Peng, L., and He, X. (2015). Globally adaptive quantile regression with ultra-high dimensional data. *Annals of statistics*, 43(5):2225.

A Proofs of the main theorems

This section provides the proofs of the main theorems. Additional technical results are available in the supplemental material.

The following notation will be used throughout the proof.

$$\begin{aligned} Q_n(\gamma) &= n^{-1}\sum_{i=1}^{n}\rho_\tau(\epsilon_i - X_i'\gamma), \\ L_n(\gamma) &= Q_n(\gamma) + \lambda||\beta_-^* + \gamma_-||_1, \\ Q(\gamma) &= \mathrm{E}(Q_n(\gamma)), \\ D_n(\gamma) &= Q_n(\gamma) - Q_n(0) - Q(\gamma) + Q(0). \end{aligned}$$

The centered L_1 regularized estimator $\widehat{\gamma} = \widehat{\beta} - \beta^*$ satisfies $\widehat{\gamma} = \arg\min_{\gamma \in \mathbb{R}^p} L_n(\gamma)$. Recall the notation $h_n = \sqrt{s \log p/n}$ and $h_n^* = \sqrt{R}(\log p/n)^{1/4}$, where R is the positive constant in (3). For a given positive constant Δ,

$$\begin{aligned} \Gamma_n &= \big\{\gamma \in \mathbb{R}^p : \gamma \in \Gamma_H, ||\gamma||_2 = \Delta h_n\big\}, \quad (13) \\ \Gamma_n^* &= \big\{\gamma \in \mathbb{R}^p : \gamma \in \Gamma_W, ||\gamma||_2 = \Delta h_n^*\big\}, \quad (14) \end{aligned}$$

where Γ_H and Γ_W are defined in (6) and (7), respectively. Throughout the proofs, a universal constant means a constant that is independent of (n, s, p) (for the hard sparsity case) or (n, R, p) (for the soft sparsity case).

Proof of Theorem 3.2.
For any given constant $\Delta > 0$,

$$P\big(||\widehat{\gamma}||_2 > \Delta h_n\big) \leq P\big(||\widehat{\gamma}||_2 > \Delta h_n \big| \widehat{\gamma} \in \Gamma_H\big) + P\big(\widehat{\gamma} \notin \Gamma_H\big). \quad (15)$$

We observe

$$P\big(||\widehat{\gamma}||_2 \leq \Delta h_n \big| \widehat{\gamma} \in \Gamma_H\big) \geq P\big(\inf_{\gamma \in \Gamma_n} L_n(\gamma) > L_n(0)\big), \quad (16)$$

where the inequality follows from the convexity of $L_n(\gamma)$ and the observation that Γ_H is a cone. Combining (15) and (16), we have

$$\begin{aligned} P\big(||\widehat{\gamma}||_2 \leq \Delta h_n\big) &\geq P\big(\inf_{\gamma \in \Gamma_n} L_n(\gamma) > L_n(0)\big) - P\big(\widehat{\gamma} \notin \Gamma_H\big) \\ &\geq P\big(\inf_{\gamma \in \Gamma_n} L_n(\gamma) > L_n(0)\big) - 4\exp(-\log p), \quad (17) \end{aligned}$$

where the last inequality is a result of Lemma B.2 in the supplement. So the crux of proving the theorem is to derive an appropriate lower bound for $P\big(\inf_{\gamma \in \Gamma_n} L_n(\gamma) > L_n(0)\big)$. Observe that

$$\begin{aligned} & \inf_{\gamma \in \Gamma_n} \big\{L_n(\gamma) - L_n(0)\big\} \\ \geq\ & \inf_{\gamma \in \Gamma_n} \big\{Q(\gamma) - Q(0)\big\} - \sup_{\gamma \in \Gamma_n} \lambda\big|||\beta_-^* + \gamma_-||_1 - ||\beta_-^*||_1\big| \\ & - \sup_{\gamma \in \Gamma_n} \big|Q_n(\gamma) - Q_n(0) - Q(\gamma) + Q(0)\big| \\ =\ & I_1 - I_2 - I_3, \quad (18) \end{aligned}$$

where the definition of I_i $(i = 1, 2, 3)$ is clear from the context. First, by Lemma B.3 in the supplement.

$$\begin{aligned} I_1 &\geq \inf_{\gamma\in\Gamma_n}\Big\{\frac{m_0}{2}\mathrm{E}\big\{|X_i'\gamma|^2\big\} - 2(b_1||\gamma||_2^2 + b_2||\gamma||_2)\exp\Big\{-\frac{b_0^2}{16\sigma_x^2||\gamma||_2^2}\Big\}\Big\} \\ &\geq \frac{m_0 m_1}{3}\inf_{\gamma\in\Gamma_n}||\gamma||_2^2 = \frac{m_0 m_1}{3}\Delta^2 h_n^2, \end{aligned}$$

for all $n > N_1$, where N_1 is a universal positive constant. This can be seeing by observing that under the conditions of the theorem there exists such an N_1 such that $\forall\ n > N_1$, $\forall\ \gamma \in \Gamma_n$, $\exp\Big\{-\frac{b_0^2}{16\sigma_x^2||\gamma||_2^2}\Big\} \leq \min\big\{\frac{1}{24b_1}, \frac{1}{24b_2}||\gamma||_2\big\}$.

Next, we evaluate I_2. We note that

$$\begin{aligned} I_2 &\leq \lambda\sup_{\gamma\in\Gamma_n}||\gamma||_1 \leq \lambda(1+\bar{c})\sup_{\gamma\in\Gamma_n}||\gamma_S||_1 \\ &\leq \lambda(1+\bar{c})\sqrt{s}\sup_{\gamma\in\Gamma_n}||\gamma_S||_2 \ \leq\ \lambda(1+\bar{c})\sqrt{s}\Delta h_n. \end{aligned}$$

Finally, we need to establish an upper bound for $I_3 = \sup_{\gamma\in\Gamma_n}|D_n(\gamma)|$, where $D_n(\gamma) = n^{-1}\sum_{i=1}^n\{Z_i(\gamma) - \mathrm{E}Z_i(\gamma)\}$ and $Z_i(\gamma) = \rho_\tau(\epsilon_i - X_i^T\gamma) - \rho_\tau(\epsilon_i)$. To handle the potentially unbounded regressors, we next apply a symmetrization technique. Define $D_n^\xi(\gamma) = n^{-1}\sum_{i=1}^n\xi_i Z_i(\gamma)$, where $\{\xi_i, i = 1, \ldots, n\}$ is a sequence of independent Rademacher variables, that is, $p(\xi_i = 1) = p(\xi_i = -1) = 1/2$. The decomposition in the proof of Lemma B.3 implies that $|Z_i(\gamma)| \leq 2|X_i^T\gamma|$. We observe that

$$\sup_{\gamma\in\Gamma_n} n^{-1}\sum_{i=1}^n \mathrm{E}(Z_i^2(\gamma)) \leq 4\sup_{\gamma\in\Gamma_n} n^{-1}\sum_{i=1}^n \mathrm{E}\big((X_i^T\gamma)^2\big) \leq 4\xi_{max}\Delta^2 h_n^2.$$

Hence the condition of Lemma 16.1 in Van de Geer (2016) is satisfied. In that lemma, taking $R = 2\xi_{max}^{1/2}\Delta h_n$, $t = \frac{9}{2}b_1^2(1+\bar{c})^2\xi_{max}^{-1}s\log p$ where b_1 is a universal positive constant specified in Lemma B.5 and noting that $t \geq 4$ for all $n > N_2$, where N_2 is a universal positive constant. We have $\forall\ n > N_2$,

$$P\Big(I_3 > 24b_1(1+\bar{c})\Delta h_n^2\Big) \leq 4P\Big(\sup_{\gamma\in\Gamma_n}|D_n^\xi(\gamma)| > 6b_1(1+\bar{c})\Delta h_n^2\Big). \tag{19}$$

To evaluate the probability at the right side of (19), we note that

$$\begin{aligned} & P\Big(\sup_{\gamma\in\Gamma_n} |D_n^{\xi}(\gamma)| > 6b_1(1+\bar{c})\Delta h_n^2 \Big) \\ = \; & \mathrm{E}_{(X,\boldsymbol{\epsilon})} P_{\xi|(X,\boldsymbol{\epsilon})}\Big(\sup_{\gamma\in\Gamma_n} |D_n^{\xi}(\gamma)| > 6b_1(1+\bar{c})\Delta h_n^2 \Big). \end{aligned} \tag{20}$$

In the following, we will derive an upper bound for the conditional probability at the right side of (20). To do so, we define two following events:

$$A_{n1} = \Big\{ n^{-1}\sum_{i=1}^{n} (X_i^T\gamma)^2 \le 2\xi_{max}\Delta^2 h_n^2, \; \forall\, \gamma\in\Gamma_n \Big\}, \tag{21}$$

$$A_{n2} = \Big\{ \mathrm{E}_{\xi|X} || n^{-1}\sum_{i=1}^{n} \xi_i X_i ||_{\infty} \le b_1\sqrt{\log p/n} \Big\}, \tag{22}$$

where $\mathrm{E}_{\xi|X}$ denotes the conditional expectation with respect to the distribution of $(\xi_1,\ldots,\xi_n)$ given the regressors, ξ_{max} denotes the largest eigenvalue of Σ and b_1 is a positive universal constant. Now define $B_n = A_{n1}\cap A_{n2}$ and $T_n = \sup_{\gamma\in\Gamma_n} \big| \frac{1}{2\xi_{max}^{1/2}\Delta h_n} D_n^{\xi}(\gamma) \big|$. We have

$$\begin{aligned} & P_{\xi|(X,\boldsymbol{\epsilon})}\Big(\sup_{\gamma\in\Gamma_n} |D_n^{\xi}(\gamma)| > 6b_1(1+\bar{c})\Delta h_n^2 \Big) \\ \le \; & P_{\xi|(X,\boldsymbol{\epsilon})}\Big(T_n > 3b_1(1+\bar{c})\xi_{max}^{-1/2} h_n \big| B_n \Big) + P(B_n^c). \end{aligned} \tag{23}$$

By Lemma B.4, $P\big(A_{n1}\big) \ge 1 - c_1\exp(-c_2 n)$ for any $n > N_1'$, where c_1, c_2 and N_1' are positive universal constants (without loss of generality, it is assumed that $N_1' > N_2$); while by Lemma B.5, $P\big(A_{n2}\big) \ge 1 - \exp(\log p - n/2)$. Therefore, $P(B_n^c) \le c_1\exp(-c_2 n) + \exp(\log p - n/2)$, $\forall\, n > N_1'$. On the event B_n, the condition of Massart's concentration inequality (e.g., Theorem 14.2, Bühlmann and van de Geer (2011)) is satisfied by T_n. Hence, $\forall\, t > 0$,

$$P_{\xi|(X,\boldsymbol{\epsilon})}\Big(T_n \ge \mathrm{E}_{\xi|(X,\boldsymbol{\epsilon})}(T_n|B_n) + t \big| B_n \Big) \le \exp\big(-nt^2/8\big). \tag{24}$$

On the other hand,

$$\begin{aligned}
& \mathrm{E}_{\xi|(X,\boldsymbol{\epsilon})}(T_n\big|B_n) \\
= \;& \mathrm{E}_{\xi|(X,\boldsymbol{\epsilon})}\Big\{ \sup_{\gamma\in\Gamma_n} n^{-1}(2\xi_{max}^{1/2}\Delta h_n)^{-1}\big|\sum_{i=1}^{n}\xi_i(\rho_\tau(\epsilon_i - X_i\gamma) - \rho_\tau(\epsilon_i))\big|\Big|B_n\Big\} \\
\leq \;& 4\mathrm{E}_{\xi|(X,\boldsymbol{\epsilon})}\Big\{ \sup_{\gamma\in\Gamma_n} n^{-1}(2\xi_{max}^{1/2}\Delta h_n)^{-1}\big|\sum_{i=1}^{n}\xi_i X_i\gamma\big|\Big|B_n\Big\} \\
\leq \;& 2(\xi_{max}^{1/2}\Delta h_n)^{-1}\sup_{\gamma\in\Gamma_n}||\gamma||_1\mathrm{E}_{\xi|(X,\boldsymbol{\epsilon})}\Big\{||n^{-1}\sum_{i=1}^{n}\xi_i X_i||_\infty\Big|B_n\Big\} \\
\leq \;& 2b_1(\xi_{max}^{1/2}\Delta h_n)^{-1}(1+\bar{c})\sqrt{s}\Delta h_n\sqrt{\log p/n} \;=\; 2b_1(1+\bar{c})\xi_{\max}^{-1/2}h_n,
\end{aligned}$$

where the first inequality applies the contraction inequality (Ledoux and Talagrand (2013)). Taking $t = b_1(1+\bar{c})\xi_{\max}^{-1/2}h_n$ in (24), then $P_{\xi|(X,\boldsymbol{\epsilon})}\Big(T_n \geq 3b_1(1+\bar{c})\xi_{\max}^{-1/2}h_n|B_n\Big) \leq \exp\big(-b_2 s\log p\big)$, for some positive universal constant b_2. This implies

$$P_{\xi|(X,\boldsymbol{\epsilon})}\Big(\sup_{\gamma\in\Gamma_n}|D_n^{\xi}(\gamma)| > 6b_1(1+\bar{c})\Delta h_n^2\big|B_n\Big) \leq \exp\big(-b_2 s\log p\big).$$

This result combined with (19) and (20) implies, $\forall\; n > \max\{N_2, N_1'\}$,

$$\begin{aligned}
& P\Big(I_3 \leq 24b_1(1+\bar{c})\Delta h_n^2\Big) \\
\geq \;& 1 - 4\big\{\exp\big(-b_2 s\log p\big) + c_1\exp(-c_2 n) + \exp(\log p - n/32)\big\}.
\end{aligned}$$

Under the conditions of the theorem, there exist positive universal constants b_3, b_4 and N_3 such that $4\big\{\exp\big(-b_2 s\log p\big) + c_1\exp(-c_2 n) + \exp(\log p - n/32) \leq b_3\exp\big(-b_4 s\log p\big)$ for any $n > N_3$.

Define $N_4 = \max\{N_1, N_2, N_3, N_1'\}$ and define $\Delta_0 = \frac{3(k_0+24b_1)(1+\bar{c})}{m_0 m_1}$. By (18) and the above analysis of I_i $(i = 1, 2, 3)$, $\forall\; n > N_4$, $\forall\; \Delta > \Delta_0$, with probability at least $1 - b_3\exp\big(-b_4 s\log p\big)$,

$$\begin{aligned}
& \inf_{\gamma\in\Gamma_n}\big\{L_n(\gamma) - L_n(0)\big\} \\
\geq \;& m_0 m_1\Delta^2 h_n^2/3 - \lambda(1+\bar{c})\sqrt{s}\Delta h_n - 24c'(1+\bar{c})\Delta h_n^2 \\
= \;& \Delta h_n^2\big\{m_0 m_1\Delta/3 - (k_0 + 24c')(1+\bar{c})\big\} > 0.
\end{aligned}$$

By (17), we conclude that $\forall\ n > N_4$, $\forall\ \Delta > \Delta_0$,

$$P\big(||\widehat{\gamma}||_2 < \Delta\sqrt{s\log p/n}\big) \geq 1 - 4\exp(-\log p) - b_3\exp\big(-b_4 s\log p\big).$$

There exist positive universal constants c^* and N_5 such that $4\exp(-\log p) - b_3\exp\big(-b_4 s\log p\big) \leq c^*\exp(-\log p)$, for any $n > N_5$. The conclusion of the theorem follows by setting $N_0 = \max\{N_4, N_5\}$. $\square$

Proof of Theorem 3.3.
Let Γ_n^* be defined as in (14) with $a = \sqrt{\log p/n}$ and $h_n^* = \sqrt{R}(\log p/n)^{1/4}$. For any given constant $\Delta > 0$,

$$P\big(||\widehat{\gamma}||_2 > \Delta h_n^*\big) \quad \leq \quad P\big(||\widehat{\gamma}||_2 > \Delta h_n^* \big| \widehat{\gamma} \in \Gamma_W\big) + P\big(\widehat{\gamma} \notin \Gamma_W\big). \tag{25}$$

We can show

$$P\big(||\widehat{\gamma}||_2 \leq \Delta h_n^* \big| \widehat{\gamma} \in \Gamma_W\big) \geq P\big(\inf_{\gamma\in\Gamma_n^*} L_n(\gamma) > L_n(0)\big), \tag{26}$$

using similar argument as that for Lemma 4 of Negahban et al. (2012)(arivx version). Specifically, we will show conditional on the event $\{\widehat{\gamma} \in \Gamma_W\}$, the event $\{\inf_{\gamma\in\Gamma_n^*} \tilde{L}_n(\gamma) > 0\}$ implies $\{||\widehat{\gamma}||_2 \leq \Delta h_n^*\}$, where $\tilde{L}_n(\gamma) = L_n(\gamma) - L_n(0)$. Assume the contrary is true, i.e., $||\widehat{\gamma}||_2 > \Delta h_n^*$. Note that Γ_W is a star-shapred set, i.e., if $\gamma \in \Gamma_W$, then $t\gamma \in \Gamma_W$, $\forall\ t \in (0,1)$. On the event $\{\widehat{\gamma} \in \Gamma_W\}$, there exists some $t^* \in (0,1)$ such that the line connecting 0 and $\widehat{\gamma}$ intersects Γ_n^* at some point $t^*\widehat{\gamma}$. By the convexity of $\tilde{L}_n(\gamma)$,

$$\tilde{L}_n(t^*\widehat{\gamma}) \leq t^*\tilde{L}_n(\widehat{\gamma}) + (1-t^*)\tilde{L}_n(0) = t^*\tilde{L}_n(\widehat{\gamma}).$$

By definition $\widehat{\gamma}$, $\tilde{L}_n(\widehat{\gamma}) \leq 0$. On the other hand, $\tilde{L}_n(t^*\widehat{\gamma}) \geq \inf_{\gamma\in\Gamma_n^*} \tilde{L}_n(\gamma) > 0$. This leads to a contradiction. We conclude that (26) holds.

Combining (25) and (26), we have, for any given constant $\Delta > 0$,

$$P\big(||\widehat{\gamma}||_2 \leq \Delta h_n^*\big) \geq P\big(\inf_{\gamma\in\Gamma_n^*} L_n(\gamma) > L_n(0)\big) - 4\exp(-\log p). \tag{27}$$

We have

$$\inf_{\gamma\in\Gamma_n^*}\big\{L_n(\gamma)-L_n(0)\big\}\geq I_1^*-I_2^*-I_3^*,$$

where $I_1^*=\inf_{\gamma\in\Gamma_n^*}\big\{Q(\gamma)-Q(0)\big\}$, $I_2^*=\sup_{\gamma\in\Gamma_n^*}\lambda\big|||\beta_-^*+\gamma_-||_1-||\beta_-^*||_1\big|$, and $I_3^*=\sup_{\gamma\in\Gamma_n^*}\big|D_n(\gamma)\big|$. By Lemma B.3,

$$\begin{aligned}I_1^* &\geq \inf_{\gamma\in\Gamma_n^*}\Big\{\frac{m_0}{2}\mathrm{E}\big\{|X_i'\gamma|^2\big\}-2(b_1||\gamma||_2^2+b_2||\gamma||_2)\exp\Big\{-\frac{b_0^2}{16\sigma_x^2||\gamma||_2^2}\Big\}\Big\}\\ &\geq \frac{m_0m_1}{3}\inf_{\gamma\in\Gamma_n^*}||\gamma||_2^2 \;=\; m_0m_1\Delta^2h_n^{*2}/3=\frac{m_0m_1}{3}\Delta^2R\sqrt{\log p/n},\end{aligned}$$

$\forall\ n>N_1^*$, where N_1^* is a positive universal constant. Furthermore,

$$I_2^*\leq\lambda\sup_{\gamma\in\Gamma_n^*}||\gamma||_1\leq\tilde{c}\Delta\lambda R,$$

for any $n>\tilde{N}$, where $\tilde{c}$ and $\tilde{N}$ are positive universal constants, by Lemma B.6.

Finally, we establish an upper bound for $I_3^*=\sup_{\gamma\in\Gamma_n^*}|D_n(\gamma)|$. Let $D_n^\xi(\gamma)$ be the symmetrized version of $D_n(\gamma)$, as defined as in the proof of Theorem 3.2. Let

$$A_{n1}' \;=\; \Big\{n^{-1}\sum_{i=1}^n(X_i^T\gamma)^2\leq 2\xi_{max}\Delta^2h_n^{*2},\ \forall\ \gamma\in\Gamma_n^*\Big\},\tag{28}$$

and let A_{n2} be defined as in (22). Define $B_n'=A_{n1}'\cap A_{n2}$. By Lemma B.4 and Lemma B.5, $P(B_n'^c)\leq c_1\exp(-c_2n)+\exp(\log p-n/2)$, for any $n>N_1'$, where c_1, c_2 and N_1' are positive universal constants. Similarly as in the proof for Theorem 3.2, Applying the inequality on symmetrization for probability (e.g., Lemma 16.1 in Van de Geer (2016) with its t set to be $\frac{9}{2}b_1^2\tilde{c}^2\xi_{max}^{-1}nh_n^{*2}$). There exists a universal positive constant N_2^* such that $\frac{9}{2}b_1^2\tilde{c}^2\xi_{max}^{-1}nh_n^{*2}\geq 4$, hence the condition of that lemma is satisfied, $\forall\ n>N_2^*$. Without loss of generality, we assume $N_2^*>\max\{\tilde{N},N_1'\}$. Then $\forall\ n>N_2^*$, we have

$$\begin{aligned}&P\Big(I_3^*>24b_1\tilde{c}\Delta h_n^{*2}\Big)\\ \leq\;& 4P\Big(\sup_{\gamma\in\Gamma_n^*}|D_n^\xi(\gamma)|>6b_1\tilde{c}\Delta h_n^{*2}\Big)\\ \leq\;& 4\mathrm{E}_{(X,\boldsymbol{\epsilon})}P_{\xi|(X,\boldsymbol{\epsilon})}\Big(T_n^*>3b_1\tilde{c}\xi_{max}^{-1/2}h_n^*\big|B_n'\Big)\\ &+4\big(c_1\exp(-c_2n)+\exp(\log p-n/2)\big),\end{aligned}\tag{29}$$

where $T_n^* = \sup_{\gamma\in\Gamma_n^*} \big|\frac{1}{2\xi_{max}^{1/2}\Delta h_n^*} D_n^\xi(\gamma)\big|$. It follows from Massart's concentration inequality that $\forall\ t > 0$,

$$P_{\xi|(X,\boldsymbol{\epsilon})}\Big(T_n^* \geq \mathrm{E}_{\xi|(X,\boldsymbol{\epsilon})}(T_n^*|B_n') + t\big|B_n'\Big) \leq \exp\big(-nt^2/8\big). \tag{30}$$

Similarly as the argument in the proof of Theorem 3.2, on the event B_n', we have

$$\begin{aligned}
\mathrm{E}_{\xi|(X,\boldsymbol{\epsilon})}(T_n^*) &\leq 2(\xi_{max}^{1/2}\Delta h_n^*)^{-1}\big(\sup_{\gamma\in\Gamma_n^*}||\gamma||_1\big)\mathrm{E}_{\xi|(X,\boldsymbol{\epsilon})}\Big\{||n^{-1}\sum_{i=1}^n \xi_i X_i||_\infty\Big\} \\
&\leq 2b_1(\xi_{max}^{1/2}\Delta h_n^*)^{-1}\tilde{c}\Delta R\sqrt{\log p/n} \;=\; 2b_1\tilde{c}\xi_{max}^{-1/2}h_n^*,
\end{aligned}$$

Take $t = b_1\tilde{c}\xi_{max}^{-1/2}h_n^*$ in (30), then $\forall\ n > N_2^*$,

$$\begin{aligned}
&P_{\xi|(X,\boldsymbol{\epsilon})}\Big(T_n^* \geq 3b_1\tilde{c}\xi_{max}^{-1/2}h_n^*|B_n'\Big) \\
\leq\ & \exp\big(-b_2 n h_n^{*2}\big) \leq \exp\big(-b_2 R\sqrt{n\log p}\big),
\end{aligned}$$

for some positive constant b_2. Hence, by (29), $\forall\ n > N_2^*$,

$$\begin{aligned}
&P\Big(I_3^* > 24b_1\tilde{c}\Delta h_n^{*2}\Big) \\
\leq\ & 4\big[\exp\big(-b_2R\sqrt{n\log p}\big) + c_1\exp(-c_2 n) + \exp(\log p - n/2)\big].
\end{aligned}$$

There exist positive universal constants c_3, c_4 and N_3^* (without loss of generality, we assume $N_3^* > N_2^*$) such that $\exp\big(-b_2R\sqrt{n\log p}\big) + c_1\exp(-c_2 n) + \exp(\log p - n/2) \leq c_3\exp\big(-c_4\log p\big)$, for any $n > N_3^*$. Hence, with probability at least $1 - c_3\exp\big(-c_4\log p\big)$, we have $I_3^* \leq 24c'\tilde{c}\Delta h_n^{*2}$, $\forall\ n > N_3^*$.

Let $N_4^* = \max\{N_1^*, N_3^*\}$. Putting together the above analysis of I_i^* $(i = 1, 2, 3)$, we have with probability at least $1 - c_3\exp\big(-c_4 s\log p\big)$, $\forall\ n > N_4^*$,

$$\begin{aligned}
\inf_{\gamma\in\Gamma_n^*}\big\{L_n(\gamma) - L_n(0)\big\} &\geq \frac{m_0 m_1}{3}\Delta^2 R\sqrt{\log p/n} - \tilde{c}\Delta\lambda R - 24c'\tilde{c}\Delta h_n^{*2} \\
&= \Delta h_n^{*2}\big[m_0 m_1\Delta/3 - (k_0 + 24b_1)\tilde{c}\big].
\end{aligned}$$

Take $\Delta_0' = \max\{1, \frac{3(k_0+24b_1)\tilde{c}}{m_0 m_1}\}$, then by 27, $\forall\ \Delta > \Delta_0'$, $\forall\ n > N_4^*$

$$P(||\widehat{\gamma}||_2 < \Delta h_n^*) \geq 1 - c_3\exp\big(-c_4 R\sqrt{n\log p}\big) - 4\exp(-\log p).$$

Note that under the conditions of the theorem there exist positive universal con-

stants c_* and N_5^* such that such that $\forall\ n > N_4^*$, we have $c_3 \exp\big(-c_4 R\sqrt{n \log p}\big) - 4\exp(-\log p) \leq c_* \exp(-\log p)$. The conclusion of the theorem following by setting $N_0' = \max\{N_4^*, N_5^*\}$. $\square$

Proof of Theorem 4.1. Our proof generalizes that of Raskutti et al. (2011). It is based on the information-theoretical approach and applies the generalized Fano method. To construct a packing set, we make use of Lemma A.3 of Rigollet et al. (2011), which provides a sparse extension of the Varshamov-Gilbert lemma: for any p and s satisfying $p \geq 2$ and $2 \leq s \leq (p+1)/2$, there exist $u_1, \ldots, u_M \in \{0,1\}^{p-1}$ such that

$$\rho_H(u_i, u_j) \geq s/4,\ \forall\ i \neq j; \quad ||u_i||_0 = s-1,\ \forall\ i;$$

and $\log(M) \geq C_1(s-1)\log\big(1 + \frac{e(p-1)}{s-1}\big)$, where C_1 is a positive universal constant, e is the natural constant, $\rho_H(u_i, u_j) = \sum_{k=1}^{p-1} \mathrm{I}(u_{ik} \neq u_{jk})$ denotes the Hamming distance between the two vectors $u_i = (u_{i1}, \ldots, u_{i(p-1)})'$ and $u_j = (u_{j1}, \ldots, u_{j(p-1)})'$.

(i) (hard sparsity case) Define $\widetilde{u}_i = (1, u_i^T)^T$, $i = 1, \ldots, M$. Then $\rho_H(\widetilde{u}_i, \widetilde{u}_j) \geq s/4$, $\forall\ i \neq j$ and $||\widetilde{u}_i||_0 = s$, $\forall\ i$. Denote $\beta^j = \delta_n \sqrt{4/s}\widetilde{u}_j$, $j = 1, \ldots, M$, where δ_n is a positive constant to be determined later. Note that $\beta^i \in \mathbb{B}_0(s)$, $i = 1, \ldots, M$. Furthermore, $\forall\ i \neq j$,

$$\delta_n^2 \leq \delta_n^2 \frac{4}{s} \rho_H(\beta^i, \beta^j) \leq ||\beta^i - \beta^j||_2^2 \leq (2s)(\delta_n^2 4/s) = 8\delta_n^2.$$

Hence, $\{\beta^1, \ldots, \beta^M\}$ form a δ_n-packing of $\mathbb{B}_0(s)$. Let V be a random variable that is uniformly distributed over the index set $\{1, \ldots, M\}$. For a lower bound of the minimax estimation error, it suffices to consider particular distributions. Conditional on the choice $V = v \in \{1, \ldots, M\}$, we generate a random sample

$$Y_i = X_i' \beta^v + \epsilon_i, \quad i = 1, \ldots, n, \tag{31}$$

where X_i has a mean-zero multivariate normal distribution with covariance matrix Σ, $\epsilon_i \sim N(0,1)$ is independent of X_i. Note that the conditional quantile $Q_{Y_i|X_i}(\tau) = X_i'\beta^v + \Phi^{-1}(\tau)$, where $\Phi^{-1}(\tau)$ is the τth quantile of the standard normal distribution. The coefficient vector of $Q_{Y_i|X_i}(\tau)$ belongs to $\mathbb{B}_0(s)$. Given the observed sample $Z^n = \{X_i, Y_i\}_{i=1}^n$, we consider the problem of testing if $V = v$. Let $\Psi : Z^n \to \{1, \ldots, M\}$ be an arbitrary test function. Then it follows from the standard argument in minimax

theory that

$$\min_{\widehat{\beta}} \sup_{\mathbb{P}(\beta^*),\beta^* \in \mathbb{B}_0(s)} \mathrm{E}_{\mathbb{P}(\beta^*)}\big\{||\widehat{\beta} - \beta^*||^2\big\} \geq \frac{\delta_n^2}{4} \inf_{\Psi} \mathbb{P}(\Psi \neq V), \tag{32}$$

where $\mathbb{P}$ denotes the joint distribution of V and Z^n. Note that Z^n can be thought of a random sample from the mixture distribution $M^{-1}\sum_{i=1}^{M}\mathbb{P}_v$, where $\mathbb{P}_v$ denotes the joint distribution of (X_i, Y_i) induced by β^v, according to (31). By Fano's inequality,

$$\mathbb{P}(\Psi \neq V) \geq 1 - \frac{I(V; Z^n) + \log 2}{\log M},$$

where $I(V; Z^n)$ is the mutual information between V and Z^n. By the convexity of the mutual information, $I(V; Z^n) \leq M^{-2}\sum_{1\leq v,v'\leq M} D_{KL}\big(\mathbb{P}(Z^n|v)||\mathbb{P}(Z^n|v')\big)$, where $D_{KL}\big(\mathbb{P}(Z^n|v)||\mathbb{P}(Z^n|v')\big)$ denotes the KL-divergence of the joint distributions of Z_n, induced by v and v', respectively. Direct calculation of the KL-divergence yields,

$$\begin{aligned}
& D_{KL}\big(\mathbb{P}(Z^n|v)||\mathbb{P}(Z^n|v')\big) \\
= \;& \mathrm{E}_{\{X_i\}_{i=1}^n}\Big\{D_{KL}\big(\prod_{i=1}^{n}\mathbb{P}(Y_i|X_i, v)||\prod_{i=1}^{n}\mathbb{P}(Y_i|X_i, v')\big)\Big\} \\
= \;& E_{X_i}\big\{\frac{n}{2}(\beta^v - \beta^{v'})'(X_i X_i^T)(\beta^v - \beta^{v'})\big\} \\
= \;& \frac{n}{2}(\beta^v - \beta^{v'})'\Sigma(\beta^v - \beta^{v'}) \\
\leq \;& 4\eta_{max}(2s)n\delta_n^2,
\end{aligned}$$

where $\eta_{max}(2s)$ denotes the largest $(2s)$-sparse eigenvalue of Σ. Hence, $I(V; Z^n) \leq 4\eta_{max}(2s)n\delta_n^2$. We have

$$\inf_{\Psi}\mathbb{P}(\Psi \neq V) \;\; \geq \;\; 1 - \frac{4\eta_{max}(2s)n\delta_n^2 + \log 2}{C_1(s-1)\log\big(1 + \frac{e(p-1)}{s-1}\big)}.$$

Note that there exists a positive universal constant N_1 such that $\forall\ n > N_1$, $\log\big(1 + \frac{e(p-1)}{s-1}\big) \leq 2\log(p/s)$ and $2\log 2 \leq C_1 s\log(p/s))$. Hence, $\forall\ n > N_1$,

$$\inf_{\Psi}\mathbb{P}(\Psi \neq V) \geq 1 - \frac{4\eta_{max}(2s)n\delta_n^2 + C_1 s\log(p/s)/2}{2C_1 s\log(p/s))}.$$

Taking $\delta_n^2 = \frac{C_1 s\log(p/s)}{8\eta_{max}(2s)n}$, then $\inf_{\Psi}\mathbb{P}(\Psi \neq V) \geq 1/2$. It follows from (32) that there

exists some positive constant c such that

$$\min_{\widehat{\beta}} \sup_{\mathbb{P}(\beta),\beta\in\mathbb{B}_0(s)} \mathrm{E}_{\mathbb{P}(\beta)}\big\{||\widehat{\beta}-\beta||^2\big\} \geq c\eta_{max}^{-1}(2s)sn^{-1}\log(s^{-1}p).$$

(ii) (soft sparsity case) Let $\tilde{s} = \lceil R\sqrt{n}\rceil$, where $\lceil\cdot\rceil$ is the ceiling function. With this choice of $\tilde{s}$, we generate $\widetilde{u}_i$, $i = 1,\ldots,M$ the same way as in (i). Denote $\beta^j = \delta_n\widetilde{u}_j$, $j = 1,\ldots,M$, where $0 < \delta_n \leq R/\tilde{s}$. Then $||\beta^j||_1 \leq R$. Hence, $\beta^j \in \mathbb{B}_1(R)$, $j = 1,\ldots,M$. Let $\kappa_n = \tilde{s}\delta_n^2/4$. Similarly as in (i), $\forall\ i \neq j$,

$$\kappa_n^2 \leq ||\beta^i - \beta^j||_2^2 \leq 8\kappa_n^2.$$

Hence, $\{\beta^1,\ldots,\beta^M\}$ form a κ_n-packing of $\mathbb{B}_1(R)$. Let V be defined the same as in the proof for (i). Given $V = v$, we generate a random sample according to (31) with the β^v specified above. Then the coefficient vector of $Q_{Y_i|X_i}(\tau)$ for this generated random sample belongs to $\mathbb{B}_1(R)$. Furthermore, we have

$$\min_{\widehat{\beta}} \sup_{\mathbb{P}(\beta^*),\beta^*\in\mathbb{B}_1(R)} \mathrm{E}_{\mathbb{P}(\beta^*)}\big\{||\widehat{\beta}-\beta^*||^2\big\} \geq \frac{\kappa_n^2}{4}\inf_{\Psi}\mathbb{P}(\Psi \neq V).$$

To evaluate the lower bound, we apply the same argument based on Fano's inequality and KL-divergence as in the proof of (i). Similarly as in the proof for (i), there exists a positive universal constant N_1' such that $\forall\ n > N_1'$,

$$\inf_{\Psi}\mathbb{P}(\Psi \neq V) \ \geq\ 1 - \frac{4\eta_{max}(2\tilde{s})n\kappa_n^2 + C_1\tilde{s}\log(p/\tilde{s})/2}{2C_1\tilde{s}\log(p/\tilde{s}))}.$$

Taking $\kappa_n^2 = \frac{C_1 s\log(p/\tilde{s})}{8\eta_{max}(2\tilde{s})n}$, then $\inf_\Psi \mathbb{P}(\Psi \neq V) \geq 1/2$. there exist a positive universal constant c' such that $\forall\ n > N_1'$,

$$\begin{aligned}\min_{\widehat{\beta}} \sup_{\mathbb{P}(\beta^*),\beta^*\in\mathbb{B}_1(R)} \mathrm{E}_{\mathbb{P}(\beta)}\big\{||\widehat{\beta}-\beta^*||^2\big\} \ &\geq\ \frac{C_1}{32\eta_{max}(2\tilde{s})}sn^{-1}\log(p/\tilde{s}) \\ &\geq\ c'\eta_{max}^{-1}(2\tilde{s})\frac{R}{\sqrt{n}}\log\big(\frac{p}{R\sqrt{n}}\big)\end{aligned}$$

due to the choice $\tilde{s}$. □

Proof of Theorem 5.1. (i) (prediction error bound without sparsity) Recall the notation $Q_n(\gamma) = n^{-1}\sum_{i=1}^n \rho_\tau(\epsilon_i - X_i'\gamma)$, $R_n(\gamma) = Q_n(\gamma) - Q_n(0)$ and $\widehat{\gamma} = \widehat{\beta} - \beta^* =$

$(\widehat{\gamma}_1, \widehat{\gamma}'_-)'$. By the definition of $\widehat{\gamma}$, we have

$$Q_n(\widehat{\gamma}) - Q_n(0) \leq \lambda\Big(||\beta^*_-||_1 - ||\beta^*_- + \widehat{\gamma}_-||_1\Big). \tag{33}$$

Note that as $||\beta^*_- + \widehat{\gamma}_-||_1 \geq ||\beta^*_-||_1 - ||\widehat{\gamma}_-||_1$, the right-hand side of (33) immediately implies that

$$R_n(\gamma) \leq \lambda\Big(2||\beta^*_-||_1 - ||\widehat{\gamma}_-||_1\Big) \leq 2\lambda||\beta^*_-||_1 \leq 2\lambda||\beta^*||_1.$$

(ii) (hard sparsity case). Under the conditions of Theorem 3.2, $P(\widehat{\gamma} \in \Gamma_H) \geq 1 - 4\exp(-\log p)$, a result of Lemma B.2. It follows from (33) that on the event $\{\widehat{\gamma} \in \Gamma_H\}$,

$$R_n(\widehat{\gamma}) = Q_n(\widehat{\gamma}) - Q_n(0) \leq \lambda||\widehat{\gamma}||_1 \leq (1+\bar{c})\lambda\sqrt{s}||\widehat{\gamma}||_2, \tag{34}$$

where the last inequality is due to the property of Γ_H. Then there exist universal positive constants N_0, Δ_0 and c^* (all independent of the triple (n, s, p)) such that for any $n > N_0$, $\forall\ \Delta > \Delta_0$,

$$P\big(||\widehat{\gamma}||_2 \leq \Delta\sqrt{s\log p/n}\big) \geq 1 - c^*\exp(-\log p).$$

Therefore, we have for any $n > N_0$, $\forall\ \Delta > \Delta_0$,

$$R_n(\widehat{\gamma}) \leq (1+\bar{c})k_0\Delta\frac{s\log p}{n} = \frac{(1+\bar{c})}{k_0}\Delta s\lambda^2,$$

with probability at least $1 - c_1^*\exp\big(-c_2^* s\log p\big)$, where c_1^* and c_2^* are positive universal constants.

(iii) (soft sparsity case) Under the conditions of Theorem 3.3, by Lemma B.2, $P(\widehat{\gamma} \in \Gamma_W) \geq 1 - 4\exp(-\log p)$, where the a in the definition of Γ_W can be taken as an arbitrary positive constant. It follows from (33) that on the event $\{\widehat{\gamma} \in \Gamma_W\}$,

$$\begin{aligned} R_n(\widehat{\gamma}) &= Q_n(\widehat{\gamma}) - Q_n(0) \leq \lambda||\widehat{\gamma}||_1 \\ &\leq \frac{2c}{c-1}\lambda\big\{||\widehat{\gamma}_{S_a}||_1 + ||\beta^*_{S_a^c}||_1\big\} \\ &\leq \frac{2c}{c-1}\lambda\big\{\sqrt{||S_a||_0}||\widehat{\gamma}_{S_a}||_2 + ||\beta^*_{S_a^c}||_1\big\}, \end{aligned} \tag{35}$$

where the second inequality is due to the property of Γ_W, the set S_a is defined in

Section 3. By Theorem 3.3, there exist universal positive constants N_0', Δ_0' and c_* (all independent of the triple (n, R, p)) such that for any $n > N_0'$, $\forall\ \Delta > \Delta_0'$,

$$P\big(||\widehat{\gamma}||_2 \leq \Delta\sqrt{R}(\log p/n)^{1/4}\big) \geq 1 - c_* \exp\big(-\log p\big).$$

Therefore for any $n > N_0'$, $\forall\ \Delta > \Delta_0'$,

$$\begin{aligned} R_n(\widehat{\gamma}) &\leq \frac{2c}{c-1} k_0 \big\{ \sqrt{||S_a||_0}\Delta\sqrt{R}(\log p/n)^{3/4} + ||\beta^*_{S_a^c}||_1 (\log p/n)^{1/2} \big\} \\ &\leq \frac{2c}{c-1} \big\{ k_0^{-1/2} \sqrt{||S_a||_0}\Delta\sqrt{R}\lambda^{3/2} + ||\beta^*_{S_a^c}||_1 \lambda \big\}, \end{aligned}$$

with probability at least $1 - c_3^* \exp\big(-c_4^* s \log p\big)$, where c_3^* and c_4^* are positive universal constants. $\square$