cemmap
centre for microdata methods and practice

Sparse Quantile Regression

Le-Yu Chen
Sokbae Lee

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP30/20

Economic and Social Research Council

Sparse Quantile Regression*

Le-Yu Chen†

Institute of Economics, Academia Sinica

Sokbae Lee‡

Department of Economics, Columbia University

Centre for Microdata Methods and Practice, Institute for Fiscal Studies

June 2020

Abstract

We consider both ℓ_0-penalized and ℓ_0-constrained quantile regression estimators. For the ℓ_0-penalized estimator, we derive an exponential inequality on the tail probability of excess quantile prediction risk and apply it to obtain non-asymptotic upper bounds on the mean-square parameter and regression function estimation errors. We also derive analogous results for the ℓ_0-constrained estimator. The resulting rates of convergence are minimax-optimal and the same as those for ℓ_1-penalized estimators. Further, we characterize expected Hamming loss for the ℓ_0-penalized estimator. We implement the proposed procedure via mixed integer linear programming and also a more scalable first-order approximation algorithm. We illustrate the finite-sample performance of our approach in Monte Carlo experiments and its usefulness in a real data application concerning conformal prediction of infant birth weights (with $n \approx 10^3$ and up to $p > 10^3$). In sum, our ℓ_0-based method produces a much sparser estimator than the ℓ_1-penalized approach without compromising precision.

Keywords: quantile regression, sparse estimation, mixed integer optimization, finite sample property, conformal prediction, Hamming distance

*We would like to thank Roger Koenker, Rahul Mazumder, Guillaume Pouliot and participants at "Optimization-Conscious Econometrics Conference," Chicago, November 2019 for helpful comments. We are also grateful to Rahul Mazumder for providing us his codes. This work was supported in part by the Ministry of Science and Technology, Taiwan (MOST107-2410-H-001-029-MY2), Academia Sinica (AS-CDA-106-H01), the European Research Council (ERC-2014-CoG-646917-ROMIA), and the UK Economic and Social Research Council (ESRC) through research grant (ES/P008909/1) to the CeMMAP.

†E-mail: lychen@econ.sinica.edu.tw

‡E-mail: sl3841@columbia.edu

1 Introduction

Quantile regression has been increasingly popular since the seminal work of Koenker and Bassett (1978). See Koenker (2005) for a classic and comprehensive text on quantile regression and Koenker (2017) for a review of recent developments. This paper is concerned with estimating a sparse high-dimensional quantile regression model:

$$Y = X^\top \theta_* + U, \tag{1.1}$$

where $Y \in \mathbb{R}$ is the outcome of interest, $X \in \mathbb{R}^p$ is a p-dimensional vector of covariates, θ_* is the vector of unknown parameters, and U is a regression error. Let $Q_\tau(U|X)$ denote the τ-th quantile of U conditional on X. Assume that $Q_\tau(U|X) = 0$ almost surely for a given $\tau \in (0, 1)$ and that the data consist of a random sample of n observations $(Y_i, X_i)_{i=1}^n$. As usual, p can be much larger than n; however, sparsity s, the number of nonzero elements of θ_*, is less than n.

To date, an ℓ_1-penalized approach to estimating (1.1) has been predominant in the literature mainly thanks to its computational advantages. See e.g., Belloni and Chernozhukov (2011), Wang, Wu, and Li (2012), Wang (2013), Belloni, Chernozhukov, and Kato (2014, 2019), Fan, Fan, and Barut (2014), Zheng, Peng, and He (2015), Lee, Liao, Seo, and Shin (2018), Lv, Lin, Lian, and Huang (2018), Wang, Van Keilegom, and Maidman (2018) and Wang (2019) among many others. The ℓ_1-penalized quantile regression (ℓ_1-PQR hereafter) is akin to the well known LASSO (Tibshirani, 1996). Recently, there is emerging interest in adopting an ℓ_0-based approach since the latter is regarded as a more direct solution to estimation problem under sparsity. For instance, Bertsimas, King, and Mazumder (2016) took an ℓ_0-constrained approach in order to solve the best subset selection problem in linear regression models. Huang, Jiao, Liu, and Lu (2018) proposed a scalable computational algorithm for ℓ_0-penalized least squares solutions. Chen and Lee (2018a,b) studied the ℓ_0-constrained and ℓ_0-penalized empirical risk minimization approaches to high dimensional binary classification problems.

In this paper, we pursue an ℓ_0-based approach to estimating sparse quantile regression. We are inspired by Bertsimas, King, and Mazumder (2016, Section 6), who provided a piece of numerical evidence—without theoretical analysis—that the ℓ_0-constrained least absolute deviation (LAD) estimator outperforms ℓ_1-penalized LAD estimator in terms of both sparsity and predictive accuracy. That is, Bertsimas, King, and Mazumder (2016) made a convincing case for adopting an ℓ_0-based approach in median regression. In convex optimization, a constrained approach is equivalent to a penalized method (see, e.g., Boyd and Vandenberghe, 2004). For non-convex problems, both are distinct and it is unclear which method is better. Therefore, in the paper, we consider both ℓ_0-constrained and ℓ_0-penalized quantile regression (ℓ_0-CQR and ℓ_0-PQR hereafter).

The main contributions of this paper are twofold. First, we derive an exponential inequality on the tail probability of the excess quantile predictive risk and apply it to obtain non-asymptotic upper bounds on a triplet of population quantities for the ℓ_0-PQR estimator: the mean excess predictive risk, the mean-square regression function estimation error, and the mean-square parameter estimation error. The resulting rates of convergence for the triplets are minimax optimal at the order of $s \ln p/n$ and the same as those of ℓ_1-PQR (see, e.g., Belloni and Chernozhukov, 2011; Wang, 2019). However, the optimal tuning parameter λ in ℓ_1-PQR is of order $\sqrt{\ln p/n}$, whereas it is of order $\ln p/n$ in ℓ_0-PQR. We also characterize expected Hamming loss for the ℓ_0-penalized estimator. In a nutshell, ℓ_0-PQR produces a sparser estimator than ℓ_1-PQR, while maintaining the same level of prediction and estimation errors. In addition, we establish analogous results for the ℓ_0-CQR estimator under the assumption that the imposed sparsity is at least as large as true sparsity. Our non-asymptotic results build on Bousquet (2002) and Massart and Nédélec (2006) and are applicable for ℓ_0-based, general M-estimation with a Lipschitz objective function that includes sparse logistic regression as a special case. Therefore, our theoretical results may be of independent interest beyond quantile regression.

The second contribution is computational. Both ℓ_0-CQR and ℓ_0-PQR estimation problems can be equivalently reformulated as mixed integer linear programming (MILP) problems. This reformulation enables us to employ efficient mixed integer optimization (MIO) solvers to compute exact solutions to the ℓ_0-based quantile regression problems. However, the method of MIO is concerned with optimization over integers, which could be computationally challenging for large scale problems. To scale up ℓ_0-based methods, Bertsimas, King, and Mazumder (2016) developed fast first-order approximation methods for both ℓ_0-constrained least squares and absolute deviation estimators. Huang, Jiao, Liu, and Lu (2018) also proposed a scalable computational algorithm for approximating the ℓ_0-penalized least squares solutions. Building on these papers, we propose a new first-order computational approach, which can deliver high-quality, approximate ℓ_0-PQR solutions and thus can be used as a warm-start strategy for boosting the computational performance of the MILP. As a standalone algorithm, our first-order approach renders the ℓ_0-PQR computationally as scalable as commonly used ℓ_1-PQR.

As an illustrative application, we consider conformal prediction of birth weights and have a horse race among ℓ_0-CQR, ℓ_0-PQR, and ℓ_1-PQR with $n \approx 1000$ and p ranging from $p \approx 20$ to $p \approx 1600$. Recently, Romano, Patterson, and Candes (2019) combined conformal prediction with quantile regression and proposed conformalized quantile regression that rigorously ensures a non-asymptotic, distribution-free coverage guarantee, independent of the underlying regression algorithm. When we implement conformal prediction using competing estimation methods, we find that both ℓ_0-CQR and ℓ_0-PQR are capable of delivering much sparser solutions than ℓ_1-PQR, while maintaining the same length and coverage of prediction confidence intervals. Furthermore, we obtain similar results in Monte Carlo experiments. Therefore, ℓ_0-CQR and ℓ_0-PQR are worthy competitors to ℓ_1-PQR—superior if a researcher prefers sparsity—as supported by non-asymptotic theory, a real-data application and Monte Carlo experiments.

The rest of this paper is organized as follows. In Section 2, we set up the sparse quantile regression model and present the ℓ_0-based approaches. In Section 3, we establish non-asymptotic statistical properties of the proposed ℓ_0-PQR and ℓ_0-CQR estimators. In Section 4, we provide both MILP- and first-order (FO)-based computational approaches for solving the ℓ_0-PQR problems. In Section 5, we perform a simulation study on the finite-sample performance of our proposed estimators. In Section 6, we illustrate our method in a real data application concerning conformal prediction of birth weights. We then conclude the paper in Section 7. Proofs of all theoretical results of the paper are collated in Appendix A.

2 ℓ_0-Based Approaches to Quantile Regression

Let $\|a\|_0$ be the ℓ_0 norm of a vector a, which is the number of nonzero components of a. The usual ℓ_1 and ℓ_2 norms are denoted by $\|\cdot\|_1$ and $\|\cdot\|_2$, respectively. For any t and u, let

$$\rho(t,u) \equiv (t-u)[\tau - 1(t \leq u)]. \tag{2.1}$$

Let Θ denote a parameter space, which is assumed to be a compact subspace of $\mathbb{R}^p$. Define

$$S_n(\theta) \equiv n^{-1} \sum_{i=1}^{n} \rho(Y_i, X_i^\top \theta). \tag{2.2}$$

We first define ℓ_0-CQR. For any given sparsity $q \geq 0$, let $\widetilde{\theta}$ denote an ℓ_0-constrained quantile regression (ℓ_0-CQR) estimator, which is defined as a solution to the following minimization problem:

$$\min_{\theta \in \mathbb{B}(q)} S_n(\theta), \text{ where } \mathbb{B}(q) \equiv \{\theta \in \Theta : \|\theta\|_0 \leq q\}. \tag{2.3}$$

In practice, choosing q is important: ℓ_0-CQR will result in selecting many more (or far fewer) covariates if the imposed sparsity is too large (or too small).

To mitigate the issue of unknown true sparsity s, we now focus on ℓ_0-PQR. Let $\widehat{\theta}$ denote an ℓ_0-PQR estimator, which is defined as a solution to the following minimization problem:

$$\min_{\theta \in \mathbb{B}(k_0)} S_n(\theta) + \lambda \|\theta\|_0, \tag{2.4}$$

where λ is a nonnegative tuning parameter and k_0 is a fixed upper bound for the true sparsity s. In ℓ_0-PQR, the main tuning parameter is λ. To adapt to an unknown s, we rely on ℓ_0-penalization that is controlled by λ. The sparsity bound k_0 is different from q in ℓ_0-CQR. The latter acts as a tuning parameter, which will be calibrated to maximize the predictive performance, whereas the former is predetermined and imposed mainly for technical regularity conditions, just like the compactness assumption on Θ. We will set k_0 with a large value in numerical exercises.

To make our proposed estimators operational, we follow the standard machine learning approach. That is, we first randomly split the dataset into three samples: training, validation and test samples. For each candidate value of the tuning parameter q or λ, we estimate the model using the training sample. Then, the tuning parameter is selected based on the quantile prediction risk using the validation sample. Finally, out-of-sample performance is evaluated using the test sample.

3 Theory for ℓ_0-Based Quantile Regression

3.1 Assumptions

We provide general regularity conditions that include quantile regression as a special case. Define $S(\theta) \equiv \mathbb{E}\left[\rho(Y, X^\top \theta)\right]$.

Assumption 1. $S(\theta) \geq S(\theta_*)$ *for any* $\theta \in \Theta$.

Note that for quantile regression,

$$S(\theta) - S(\theta_*) = \int \int_0^{x^\top(\theta-\theta_*)} \left[F_{U|X}(z|x) - F_{U|X}(0|x)\right] dz \, dF_X(x), \tag{3.1}$$

where $F_{U|X}(\cdot|x)$ is the cumulative distribution function of U conditional on $X = x$ and F_X is the cumulative distribution function of X. Thus, Assumption 1 is satisfied.

Assumption 2. *There exists a Lipschitz constant L such that*

$$|\rho(t, u_1) - \rho(t, u_2)| \leq L\, |u_1 - u_2| \tag{3.2}$$

for all $t, u_1, u_2 \in \mathbb{R}$.

Assumption 2 is satisfied for quantile regression with $L = 1$. For any two real numbers x and y, let $x \vee y \equiv \max\{x, y\}$ and $x \wedge y \equiv \min\{x, y\}$.

Assumption 3. *There exists a positive and finite constant B such that*

$$\max_{1 \leq j \leq p} \{\left|X^{(j)}\right| \vee \left|\theta^{(j)}\right|\} \leq B, \tag{3.3}$$

where $X^{(j)}$ and $\theta^{(j)}$ denote the j-th component of X and that of θ, respectively.

Assumption 3 requires that each component of X and that of θ be bounded by a universal constant. This condition could be restrictive yet is commonly adopted in the literature. For example, Zheng, Peng, and He (2018) assumed the uniform boundedness of each of the covariates, citing the literature that points out that "a global linear quantile regression model is most sensible when the covariates are confined to a compact set" to avoid the problem of quantile crossing.

Assumption 4 (Separability Condition)**.** *There exists a countable subset Θ' of Θ that satisfies the following conditions: (i) for any $\theta \in \Theta$, there exists a sequence (θ_j) of elements of Θ' such that,* *for every realization of* (Y, X), $\rho(Y, X^\top\theta_j)$ *converges to* $\rho(Y, X^\top\theta)$ *as* $j \to \infty$. *(ii) Furthermore, for any given* $\varepsilon_* > 0$, *there exists a point* $\theta'_* \in \Theta'$ *such that* $\|\theta'_*\|_0 = \|\theta_*\|_0$ *and* $S(\theta'_*) \leq S(\theta_*) + \varepsilon_*$.

Assumption 4 is very mild. A similar condition is assumed in Massart and Nédélec (2006) to avoid measurability issues and to use the concentration inequality by Bousquet (2002). By Assumption 2 and taking $\Theta' = \Theta \cap \mathbb{Q}^p$, Assumption 4 (i) holds by the denseness of the set of rational numbers and the continuity of the function ρ. Suppose that, for some non-negative random variable Z with $\mathbb{E}(Z) < \infty$, $\rho(Y, X^\top\theta) \leq Z$ holds with probability 1 for every $\theta \in \Theta$. Then using this condition together with Assumption 4 (i), we can also deduce from the dominated convergence theorem that $S(\theta_*) = \inf_{\theta\in\Theta'} S(\theta)$ and thus Assumption 4 (ii) also holds. In the quantile regression case, we can take the dominating variable Z to be $|Y| + pB^2$, which has finite mean provided that the mean of $|Y|$ is also finite. The requirement that $\mathbb{E}|Y| < \infty$ is not strictly necessary because we can redefine the quantile regression objective function by $\rho(Y, X^\top\theta) - \rho(Y, X^\top\theta_*)$, whose magnitude is uniformly bounded above by $2pB^2$.

For each θ, define $R(\theta) \equiv \mathbb{E}[|X^\top(\theta - \theta_*)|^2]$, which is the the expected squared difference of the true quantile regression function $X^\top\theta_*$ and a linear fit evaluated at a given parameter vector θ.

Assumption 5. *For some* $k \geq k_0$ *in* (2.4), *there exists a constant* $\kappa_0 > 0$ *such that*

$$S(\theta) - S(\theta_*) \geq \kappa_0^2 R(\theta) \text{ for all } \theta \in \mathbb{B}(k). \tag{3.4}$$

Assumption 5 relates $R(\theta)$ to the difference of their corresponding quantile predictive risks. Given Assumption 3, if the distribution $F_{U|X}(z|x)$ admits a Lebesgue density $f_{U|X}(z|x)$ that is bounded below by a positive constant c_u for all z in an open interval containing $[-2B^2(k_0 + s), 2B^2(k_0 + s)]$ and for all x in the support of X, then Assumption 5 holds with $\kappa_0 = \sqrt{c_u/2}$.

Assumption 6. *For some $k \geq k_0$ in* (2.4)*, there exists a constant $\kappa_1 > 0$ such that*

$$R(\theta) \geq \kappa_1^2 \|\theta - \theta_*\|_2^2 \text{ for all } \theta \in \mathbb{B}(k). \tag{3.5}$$

For any subset $J \subset \{1, ..., p\}$, let X_J denote the $|J|$-dimensional subvector of $X \equiv (X^{(1)}, \ldots, X^{(p)})^\top$ formed by keeping only those elements $X^{(j)}$ with $j \in J$. Suppose that, for any subset $J \subset \{1, ..., p\}$ such that $|J| \leq (k_0 + s)$, the smallest eigenvalue of $\mathbb{E}\left(X_J X_J^\top\right)$ is bounded below by a positive constant ω. Then Assumption 6 holds with $\kappa_1 = \sqrt{\omega}$. This assumption is related to the sparse eigenvalue condition used in the high dimensional regression literature (see, e.g. Raskutti, Wainwright, and Yu (2011)). For example, if X is a random vector with mean zero and the covariance matrix Σ whose (i, j) component is $\Sigma_{i,j} = r^{|i-j|}$ for some constant $r > 0$, then the smallest eigenvalue of Σ is bounded away from zero where the lower bound is independent of the dimension p (van de Geer and Bühlmann (2009, p. 1384)) and thus $\mathbb{E}\left(X_J X_J^\top\right)$ is bounded below by a universal positive constant for every $J \subset \{1, ..., p\}$.

3.2 Non-Asymptotic Bounds and Minimax Optimal Rates

The following theorem is the key step to the main results of this section for ℓ_0-PQR.

Theorem 1. *Let Assumptions 1–6 hold. Suppose $s \leq k_0$. Then, for any given positive scalar $\eta \leq 1$, there is a universal constant M, which depends only on η, such that, for every $y \geq 1$,*

$$\mathbb{P}\left[S(\widehat{\theta}) - S(\theta_*) \geq \frac{2\lambda s}{1-\eta} + 32C^2(s + k_0)\left(\frac{1 + \eta + M\eta y}{1-\eta}\right)\frac{\ln(2p)}{n}\right] \leq \exp(-y), \tag{3.6}$$

$$\mathbb{P}\left[\|\widehat{\theta} - \theta_*\|_0 \geq \frac{4-2\eta}{1-\eta}s + 32\lambda^{-1}C^2(s + k_0)\left(\frac{1 + \eta + M\eta y}{1-\eta}\right)\frac{\ln(2p)}{n}\right] \leq \exp(-y), \tag{3.7}$$

where

$$C \equiv 8LB\kappa_1^{-1}\kappa_0^{-1}, \tag{3.8}$$

provided that

$$\ln(2p) \geq \left(\frac{\kappa_1^2\kappa_0^2}{64L} \vee 1\right). \tag{3.9}$$

Results (3.6) and (3.7) of Theorem 1 are non-asymptotic and establish exponential inequalities on the tail probabilities of the excess quantile predictive risk $S(\widehat{\theta}) - S(\theta_*)$ as well as the ℓ_0-distance between the ℓ_0-PQR estimator and the true parameter value. Applying inequality (3.6), we can obtain non-asymptotic upper bounds on a triplet of population quantities: (i) the mean excess predictive risk $\mathbb{E}[S(\widehat{\theta}) - S(\theta_*)]$; (ii) the mean-square regression function estimation error $\mathbb{E}[R(\widehat{\theta})]$; (iii) the mean-square parameter estimation error $\mathbb{E}[\|\widehat{\theta} - \theta_*\|_2^2]$. The results concerning these bounds are given in the next theorem.

Theorem 2. *Let Assumptions 1–6 hold. Suppose $s \leq k_0$. Given condition (3.9) of Theorem 1, there is a universal constant K, which depends only on the constants L and B, such that the following bounds hold:*

$$\begin{aligned}
\mathbb{E}\left[S(\widehat{\theta}) - S(\theta_*)\right] &\leq 4\lambda s + \frac{K(s+k_0)}{\kappa_1^2\kappa_0^2}\frac{\ln(2p)}{n}, & (3.10)\\
\mathbb{E}\left[R(\widehat{\theta})\right] &\leq \kappa_0^{-2}\left(4\lambda s + \frac{K(s+k_0)}{\kappa_1^2\kappa_0^2}\frac{\ln(2p)}{n}\right), & (3.11)\\
\mathbb{E}\left[\left\|\widehat{\theta} - \theta_*\right\|_2^2\right] &\leq \kappa_1^{-2}\kappa_0^{-2}\left(4\lambda s + \frac{K(s+k_0)}{\kappa_1^2\kappa_0^2}\frac{\ln(2p)}{n}\right). & (3.12)
\end{aligned}$$

If k_0/s is bounded by a fixed constant, we can deduce from Theorem 2 that

$$\begin{aligned}
\mathbb{E}\left[S(\widehat{\theta}) - S(\theta_*)\right] &= O\left[\left(\lambda + n^{-1}\ln p\right)s\right],\\
\mathbb{E}\left[R(\widehat{\theta})\right] &= O\left[\left(\lambda + n^{-1}\ln p\right)s\right],\\
\mathbb{E}\left[\left\|\widehat{\theta} - \theta_*\right\|_2^2\right] &= O\left[\left(\lambda + n^{-1}\ln p\right)s\right],
\end{aligned}$$

which suggests that the optimal λ be of the following form:

$$\lambda = C_\lambda\frac{\ln p}{n}, \tag{3.13}$$

where C_λ is a constant that needs to be chosen by a researcher. Under (3.13) and the side condition that $k_0/s \leq C_k$ for some fixed constant C_k, we have that

$$\mathbb{E}\left[S(\widehat{\theta}) - S(\theta_*)\right] = O\left(\frac{s \ln p}{n}\right), \mathbb{E}\left[R(\widehat{\theta})\right] = O\left(\frac{s \ln p}{n}\right) \text{ and } \mathbb{E}\left[\left\|\widehat{\theta} - \theta_*\right\|_2^2\right] = O\left(\frac{s \ln p}{n}\right).$$

We now specialize Theorem 2 to quantile regression. The following corollary provides the main results for ℓ_0-PQR.

Corollary 1. *Assume that (i)* (3.3) *holds and $k_0 \in [s, C_k s]$ for a fixed constant $C_k \geq 1$, (ii) $\lambda = C_\lambda \ln p/n$ for a fixed constant C_λ, (iii) $\mathbb{E}|Y| < \infty$, (iv) $f_{U|X}(z|x)$ is bounded below by $c_u > 0$ for all z in an open interval containing $[-2B^2(k_0+s), 2B^2(k_0+s)]$ and for all x in the support of X, (v) for any subset $J \subset \{1, ..., p\}$ such that $|J| \leq (k_0+s)$, the smallest eigenvalue of $\mathbb{E}\left(X_J X_J^\top\right)$ is bounded below by a positive constant ω. Then, there is a universal constant $\bar{K}$, which depends only on the constants B, such that*

$$\mathbb{E}\left[S(\widehat{\theta}) - S(\theta_*)\right] \leq 4C_\lambda \frac{s \ln p}{n} + \frac{\bar{K}(C_k+1)s}{c_u \omega}\frac{\ln(2p)}{n}, \tag{3.14}$$

$$\mathbb{E}\left[R(\widehat{\theta})\right] \leq \frac{8C_\lambda}{c_u}\frac{s \ln p}{n} + \frac{2\bar{K}(C_k+1)s}{c_u^2 \omega}\frac{\ln(2p)}{n}, \tag{3.15}$$

$$\mathbb{E}\left[\left\|\widehat{\theta} - \theta_*\right\|_2^2\right] \leq \frac{8C_\lambda}{c_u \omega}\frac{s \ln p}{n} + \frac{2\bar{K}(C_k+1)s}{c_u^2 \omega^2}\frac{\ln(2p)}{n}, \tag{3.16}$$

provided that

$$\ln(2p) \geq \left(\frac{c_u \omega}{128} \vee 1\right). \tag{3.17}$$

Corollary 1 provides non-asymptotic bounds on the mean-square regression function and parameter estimation errors as well as the excess quantile prediction risk. The resulting rates of convergence are of order $s \ln p/n$, which is the same as those of ℓ_1-PQR (see, e.g., Belloni and Chernozhukov (2011) and Wang (2013) for earlier results and Wang (2019) for the latest results). These are minimax optimal rates of convergence (up to a logarithmic factor) because it is shown in Wang (2019, Theorem 4.1(i)) that the minimax lower bound for $\mathbb{E}[\|\widehat{\theta} - \theta_*\|_2^2]$ is of order $s\ln(p/s)/n$. The optimal tuning parameter λ in ℓ_1-PQR is of order $\sqrt{\ln p/n}$, whereas it is of order $\ln p/n$ in ℓ_0-PQR.

Remark 1. Instead of assuming condition (iv) in Corollary 1, one may assume the regularity conditions that are similar to those imposed in Belloni and Chernozhukov (2011): that is, $f_{U|X}(0|x)$ is bounded below by a positive constant for all x in the support of X, $\partial f_{U|X}(z|x)/\partial z$ exists and is bounded in absolute value by a constant uniformly in (z, x), and

$$\inf_{\theta \in \mathbb{B}(k_0):\theta \neq \theta_*} \frac{\left\{\mathbb{E}\left[\left|X^\top(\theta - \theta_*)\right|^2\right]\right\}^{3/2}}{\mathbb{E}\left[\left|X^\top(\theta - \theta_*)\right|^3\right]} > 0.$$

The last condition is called the restricted nonlinearity condition (Belloni and Chernozhukov, 2011). In a recent working paper, Wang (2019) established theoretical results for ℓ_1-PQR without relying on the restricted nonlinearity condition. In fact, Wang (2019) only assumed a uniform lower bound for $f_{U|X}(\cdot|x)$ in a neighborhood of zero, which is weaker than condition (iv) in Corollary 1. It is an open question whether we can verify Assumption 5 under a weaker condition imposed in Wang (2019).

Using the method for proving Corollary 1, we can obtain the following result for ℓ_0-CQR.

Corollary 2. *Assume that (i)* (3.3) *holds, (ii) $s \leq q$, (iii) $\mathbb{E}|Y| < \infty$, (iv) $f_{U|X}(z|x)$ is bounded below by $c_u > 0$ for all z in an open interval containing $[-2B^2(q+s), 2B^2(q+s)]$ and for all x in the support of X, (v) for any subset $J \subset \{1, ..., p\}$ such that $|J| \leq (q+s)$, the smallest eigenvalue of $\mathbb{E}\left(X_J X_J^\top\right)$ is bounded below by a positive constant ω. Then, there is a universal constant $\tilde{K}$, which depends only on the constant B, such that*

$$\mathbb{E}\left[S(\widetilde{\theta}) - S(\theta_*)\right] \leq \frac{\tilde{K}(s+q)}{c_u\omega}\frac{\ln(2p)}{n}, \quad (3.18)$$

$$\mathbb{E}\left[R(\widetilde{\theta})\right] \leq \frac{2\tilde{K}(s+q)}{c_u^2\omega}\frac{\ln(2p)}{n}, \quad (3.19)$$

$$\mathbb{E}\left[\left\|\widetilde{\theta} - \theta_*\right\|_2^2\right] \leq \frac{2\tilde{K}(s+q)}{c_u^2\omega^2}\frac{\ln(2p)}{n}, \quad (3.20)$$

provided that (3.17) *holds.*

Corollary 2 shows that the ℓ_0-CQR estimator is minimax optimal, provided that the imposed sparsity q is at least as large as the true sparsity s and that q/s is bounded by a fixed constant. Therefore, our theory predicts that ℓ_0-PQR and ℓ_0-CQR would perform similarly in applications.

3.3 Hamming Loss

Applying (3.7) of Theorem 1, we now derive a theoretical result regarding the ℓ_0-PQR in terms of expected Hamming loss. Specifically, the following theorem presents an upper bound on the expectation of the Hamming distance between $\widehat{\theta}$ and θ_*.

Theorem 3. *Let Assumptions 1–6 hold. Furthermore,* (3.9) *holds,* $k_0 \in [s, C_k s]$ *for a fixed constant* $C_k \geq 1$*, and* $\lambda = C_\lambda \ln(p)/n$*. For any given* $\nu > 0$*, there exists a sufficiently large constant* C_λ*, which does not depend on* (s, n, p)*, such that*

$$\mathbb{E}\left[\frac{\|\widehat{\theta} - \theta_*\|_0}{s}\right] \leq (4 + \nu).$$

Note that $D_H(\widehat{\theta}, \theta_*) \equiv s^{-1}\|\widehat{\theta} - \theta_*\|_0$ is the Hamming distance—normalized by diving it by s—between $\widehat{\theta}$ and θ_*, that is, s^{-1} times the number of elements of the ℓ_0-PQR estimator that are different from the corresponding elements of the true parameter vector. Theorem 3 shows that $\mathbb{E}[D_H(\widehat{\theta}, \theta_*)]$ can be bounded by a constant that is slightly larger than 4, provided that the tuning parameter λ is suitably chosen. Note that

$$\mathbb{P}\left(\widehat{\theta} \neq \theta_*\right) = \mathbb{P}\left(\|\widehat{\theta} - \theta_*\|_0 \geq 1\right) \leq \mathbb{E}\left[\|\widehat{\theta} - \theta_*\|_0\right].$$

Thus, we do not expect that $\mathbb{E}[\|\widehat{\theta} - \theta_*\|_0]$ can be small since it is impossible to make $\mathbb{P}(\widehat{\theta} \neq \theta_*)$ small. Instead, what we obtain in Theorem 3 is that $\mathbb{E}[\|\widehat{\theta} - \theta_*\|_0]$ is bounded by $(4 + \nu)s$, independent of p.

In view of Theorem 2, Theorem 3 suggests that the estimated sparsity and the selected set of covariates of ℓ_0-PQR cannot be too distinct from s and the true set of nonzero elements of θ_*. By Theorem 3 , the resulting sparsity of ℓ_0-PQR is likely to be substantially smaller than k_0 with a suitable choice of λ and k_0; therefore, we expect that the constraint $\theta \in \mathbb{B}(k_0)$ in (2.4) will not be binding in practice. Moreover, since the choice of C_λ in Theorem 3 is independent of (s, n, p), the minimax optimal rates are still intact.

Remark 2. Using a simple Gaussian mean model, Butucea, Ndaoud, Stepanova, and Tsybakov (2018) considered variable selection under expected Hamming loss. Among other things, they derive sufficient and necessary conditions under which the following term converges to zero (using our notation):

$$\mathbb{E}\left[\frac{1}{s}\sum_{j=1}^{p}\left|1(\widehat{\theta}_j \neq 0) - 1(\theta_{*,j} \neq 0)\right|\right], \tag{3.21}$$

where $1(\cdot)$ is the indicator function and $\widehat{\theta}_j$ and $\theta_{*,j}$, respectively, are the j-th elements of $\widehat{\theta}$ and θ_*. Their conditions involve the size of the smallest non-zero elements of a signal vector. It is an interesting future research topic to investigate the behavior of (3.21) in ℓ_0-PQR.

4 Implementation of ℓ_0-PQR

In Section 4.1, we propose a data-dependent rule-of-thumb choice of λ. In Section 4.2, we present a mixed integer optimization (MIO) algorithm for computing the ℓ_0-PQR estimator. In Section 4.3, we develop a first-order (FO) approximation algorithm that can be used as either a standalone solution algorithm or a warm-start strategy for our MIO algorithm.

4.1 Choice of λ

Note that the scale of the objective function $S_n(\theta)$ varies if that of Y changes. To relate the penalty term to the scale of Y, we propose the following simple rule:

$$\lambda = c\left(n^{-1}\sum\nolimits_{i=1}^{n}|Y_i|\right)\frac{\ln p}{n}, \tag{4.1}$$

which is proportional to the sample average of the absolute value of Y. We implement this rule-of-thumb choice and calibrate the scalar c in our numerical exercises.

4.2 Computation through Mixed Integer Optimization

The MIO approach is useful for solving variable selection problems with ℓ_0-norm constraints or penalties (see, e.g., Bertsimas, King, and Mazumder, 2016; Chen and Lee, 2018a,b). Assume that the parameter space Θ takes the form $\Theta = \prod_{j=1}^{p}\left[\underline{\theta}_j, \overline{\theta}_j\right]$, where $\underline{\theta}_j$ and $\overline{\theta}_j$ are lower and upper parameter bounds such that $-\infty < \underline{\theta}_j \leq \theta_j \leq \overline{\theta}_j < \infty$ for $j \in \{1, ..., p\}$. Our implementation builds on the method of mixed integer linear programming (MILP). Specifically, the ℓ_0-penalized minimization problem (2.4) can be equivalently reformulated as the following MILP problem:

$$\min_{\theta\in\boldsymbol{\Theta},(r_i,s_i)_{i=1}^{n},(d_j)_{j=1}^{p}} \frac{1}{n}\sum\nolimits_{i=1}^{n}\left[\tau r_i + (1-\tau)\, s_i\right] + \lambda\sum\nolimits_{j=1}^{p} d_j \tag{4.2}$$

subject to

$$r_i - s_i = Y_i - X_i^{\top}\theta,\ i \in \{1, ..., n\}, \tag{4.3}$$

$$d_j\underline{\theta}_j \leq \theta_j \leq d_j\overline{\theta}_j,\ j \in \{1, ..., p\}, \tag{4.4}$$

$$d_j \in \{0, 1\},\ j \in \{1, ..., p\}, \tag{4.5}$$

$$r_i \geq 0,\ s_i \geq 0,\ i \in \{1, ..., n\}, \tag{4.6}$$

$$\sum_{j=1}^{p} d_j \leq k_0. \tag{4.7}$$

We now explain the equivalence between (2.4) and (4.2). If we remove from the problem (2.4) the second term of the objective function as well as all the $(d_1, ..., d_p)$ control variables together with their constraints (4.4) and (4.5), the resulting minimization problem reduces to the linear programming reformulation of the standard linear quantile regression problem (Koenker, 2005, Section 6.2). In the presence of the penalty term and the $(d_1, ..., d_p)$ controls, the inequality and dichotomization constraints (4.4) and (4.5) ensure that, whenever $d_j = 0$, the value θ_j must also be zero and the sum $\sum_{j=1}^{p} d_j$ thus captures the number of non-zero components of the vector θ. The last constraint (4.7) imposes that the estimated sparsity is at most k_0. As a result, both minimization problems (2.4) and (4.2) are equivalent. This equivalence enables us to employ modern MIO solvers to solve ℓ_0-PQR problems.

4.3 Computation through First-Order Approximation

The MIO formulation (4.2) is concerned with optimization over integers, which could be computationally challenging for large scale problems. Bertsimas, King, and Mazumder (2016, Section 3) have developed discrete first-order algorithms enabling fast computation of near optimal solutions to ℓ_0-constrained least squares and least absolute deviation estimation problems. Huang, Jiao, Liu, and Lu (2018) have also proposed fast and scalable algorithms for computing approximate solutions to ℓ_0-penalized least squares estimation problems. These algorithms build on the necessary conditions for optimality in the ℓ_0-constrained or penalized optimization problems. Motivated from these papers, in this subsection, we present a first-order approximation algorithm that can be used as either a standalone solution algorithm or a warm-start strategy for enhancing the computational performance of our MIO approach to the ℓ_0-PQR problem.

For $\tau \in (0, 1)$, the quantile regression objective function (2.2) can be equivalently expressed as

$$S_n(\theta) = n^{-1} \max_{\tau-1 \le w_i \le \tau} \sum_{i=1}^{n} w_i(Y_i - X_i^\top \theta). \tag{4.8}$$

The function $S_n(\theta)$ is nonsmooth. Following Nesterov (2005), we can construct a smooth approximation of $S_n(\theta)$ by

$$S_n(\theta;\delta) \equiv n^{-1} \max_{\tau-1 \le w_i \le \tau} \left[\sum_{i=1}^{n} w_i(Y_i - X_i^\top \theta) - \frac{\delta}{2} \|w\|_2^2 \right] \tag{4.9}$$

where w denote the vector of controls $(w_1, ..., w_n)$ in the maximization problem (4.9).

Assume that the parameter space Θ is of an equilateral cube form $\Theta = [-B, B]^p$ for some $B > 0$. Let t be any given vector in $\mathbb{R}^p$. Let $\widehat{\beta}$ be a solution to the following ℓ_0-penalized minimization problem:

$$\min_{\beta \in \mathbb{B}(k_0)} \|\beta - t\|_2^2 + \lambda \|\beta\|_0, \tag{4.10}$$

where λ is a non-negative penalty tuning parameter. It is straightforward to see that the solution $\widehat{\beta}$ can be computed as follows. Let $\widetilde{\beta}$ be a p dimensional vector given by

$$\widetilde{\beta}_j = \begin{cases} B \times 1\left\{B^2 - 2t_j B + \lambda < 0\right\} \text{ if } t_j > B \\ t_j \times 1\{|t_j| > \sqrt{\lambda}\} \text{ if } -B \le t_j \le B \\ -B \times 1\left\{B^2 + 2t_j B + \lambda < 0\right\} \text{ if } t_j < -B \end{cases},$$

for $j \in \{1, ..., p\}$. Then the solution $\widehat{\beta} = \widetilde{\beta}$ if $\|\widetilde{\beta}\|_0 \le k_0$. Otherwise, letting $S(t)$ denote the set of k_0 indices that keep track of the largest k_0 components of t in absolute value, we set $\widehat{\beta}_j = \widetilde{\beta}_j$ for $j \in S(t)$ and $\widehat{\beta}_j = 0$ for $j \notin S(t)$. Therefore, the problem (4.10) admits a simple closed-form solution. We will exploit this fact and develop a first-order approximation algorithm.

Define

$$Q_n(\theta;\delta) \equiv S_n(\theta;\delta) + \lambda \|\theta\|_0. \tag{4.11}$$

For any vector $t \in \mathbb{R}^p$, suppose we can construct a quadratic envelope of $S_n(\theta;\delta)$ with respect to the vector t in the sense that

$$S_n(\theta;\delta) \leq \widetilde{S}_n(\theta;t,\delta,l) \equiv S_n(t;\delta) + \bigtriangledown_\theta S_n(t;\delta)^\top (\theta - t) + \frac{l}{2}\|\theta - t\|_2^2 \tag{4.12}$$

for some non-negative real scalar l, which does not depend on the parameter vector θ. Note that (4.12) holds whenever the gradient function $\bigtriangledown_\theta S_n(\cdot;\delta)$ is Lipschitz continuous such that

$$\| \bigtriangledown_\theta S_n(t;\delta) - \bigtriangledown_\theta S_n(t';\delta)\|_2 \leq h\|t - t'\|_2 \tag{4.13}$$

for some Lipschitz constant h, which does not depend on t and t'. By the envelope theorem,

$$\bigtriangledown_\theta S_n(t;\delta) = -\frac{1}{n}\sum_{i=1}^n X_i \widehat{w}_{i,\delta},$$

where $(\widehat{w}_{1,\delta}, ..., \widehat{w}_{n,\delta})$ is the solution to the minimization problem (4.9). Using Nesterov (2005, Theorem 1), we can deduce that (4.13) holds with

$$h = \frac{1}{n\delta}\text{ trace }\left(\sum_{i=1}^n X_i X_i'\right) \tag{4.14}$$

and hence (4.12) holds for every $l \geq h$.

Define

$$\widetilde{Q}_n(\theta;t,\delta,l) \equiv \widetilde{S}_n(\theta;t,\delta,l) + \lambda\|\theta\|_0.$$

Note that $\widetilde{Q}_n(\theta;t,\delta,l)$ is an upper envelope of $Q_n(\theta;\delta)$ around the vector t with the property that $\widetilde{Q}_n(t;t,\delta,l) = Q_n(t;\delta)$.

For $t \in \mathbb{R}^p$, define the mapping

$$H_{\delta,l}(t) \equiv \arg\min_{\theta\in\mathbb{B}(k_0)}\left\{\left\|\theta - \left(t - \frac{1}{l}\bigtriangledown_\theta S_n(t;\delta)\right)\right\|_2^2 + \lambda\|\theta\|_0\right\}. \tag{4.15}$$

Arranging the terms, we can easily deduce

$$H_{\delta,l}(t) = \arg\min_{\theta\in\mathbb{B}(k_0)} \widetilde{Q}_n(\theta; t, \delta, l). \tag{4.16}$$

We say that a point $t \in \mathbb{R}^p$ is a stationary point of the mapping $H_{\delta,l}$ if $t \in H_{\delta,l}(t)$. For each given value of δ, let $\widehat{\theta}_\delta$ denote a solution to the problem of minimizing $Q_n(\theta;\delta)$ over $\theta \in \mathbb{B}(k_0)$. We propose to approximate $\widehat{\theta}_\delta$ by solving for the stationary point of the mapping $H_{\delta,l}$. This can be justified by the following proposition.

Proposition 1. *The following statements hold:*

(a) If $\widehat{\theta}_\delta \in \arg\min_{\theta\in\mathbb{B}(k_0)} Q_n(\theta;\delta)$, then $\widehat{\theta}_\delta \in H_{\delta,l}(\widehat{\theta}_\delta)$.

(b) Let $l > h$ and t_m be a sequence such that $t_{m+1} \in H_{\delta,l}(t_m)$. Then, for some limits t^* and Q^*, we have that $t_m \longrightarrow t^*$, $Q_n(t_m;\delta) \downarrow Q^*$ as $m \longrightarrow \infty$. Moreover,

$$\min_{m=1,\ldots,N} \|t_{m+1} - t_m\|_2^2 \leq \frac{2\left(Q_n(t_1;\delta) - Q^*\right)}{N\left(l-h\right)}. \tag{4.17}$$

Proposition 1 implies that any solution to the minimization of $Q_n(\theta;\delta)$ over $\theta \in \mathbb{B}(k_0)$ is also a stationary point of the mapping $H_{\delta,l}$. Moreover we can solve for a stationarity point by iterating until convergence. Result (4.17) indicates that the convergence rate is $O(N^{-1})$, where N is the number of performed iterations. Note that we can use (4.10) to obtain a closed-form solution to the ℓ_0-penalized minimization problem (4.15) for every $t \in \mathbb{R}^p$ and therefore solving for a stationary point of $H_{\delta,l}$ would incur relatively little computational cost.

We now turn to the ℓ_0-PQR problem (2.4). Let $c_\tau \equiv \tau^2 \vee (1-\tau)^2$.

Proposition 2. *For $\delta \geq 0$, if $\widehat{\theta}_\delta \in \arg\min_{\theta\in\mathbb{B}(k_0)} Q_n(\theta;\delta)$, then*

$$S_n(\widehat{\theta}_\delta) + \lambda\|\widehat{\theta}_\delta\|_0 \leq \min\nolimits_{\theta\in\mathbb{B}(k_0)} \left\{S_n(\theta) + \lambda\|\theta\|_0\right\} + \frac{\delta c_\tau}{2}. \tag{4.18}$$

Given a tolerance level ϵ, Proposition 2 implies that, for any given $\delta \leq 2\epsilon c_\tau^{-1}$, if we solve for the minimization of $Q_n(\theta;\delta)$ over $\theta \in \mathbb{B}(k_0)$, the resulting solution $\widehat{\theta}_\delta$ is an ϵ-level approximate ℓ_0-PQR estimator in the sense that

$$S_n(\widehat{\theta}_\delta) + \lambda\|\widehat{\theta}_\delta\|_0 \leq \min_{\theta\in\mathbb{B}(k_0)} \{S_n(\theta) + \lambda\|\theta\|_0\} + \epsilon.$$

This thus yields the following algorithm for computing a near optimal solution to the ℓ_0-PQR problem (2.4).

Algorithm 1. *Given an initial guess $\widehat{\theta}_1$, set $\delta = 2\epsilon c_\tau^{-1}$ and perform the following iterative procedure starting with $k = 1$:*

Step 1. For $k \geq 1$, compute $\widehat{\theta}_{k+1} \in H_{\delta,l}(\widehat{\theta}_k)$.

Step 2. Repeat Step 1 until the objective function $Q_n(\cdot;\delta)$ converges.

5 Simulation Study

In this section, we perform Monte Carlo simulation experiments to evaluate the performance of our ℓ_0-PQR approach. We consider the following data generating setup. Let $Z = (Z_1, ..., Z_{p-1})$ be a $p-1$ dimensional multivariate normal random vector with mean zero and covariance matrix Σ with its element $\Sigma_{i,j} = (0.5)^{|i-j|}$. Let $X = (X_1, ..., X_p)$ be a p dimensional covariate vector with its components $X_1 = 1$ and $X_j = Z_{j-1}1\{|Z_{j-1}| \leq 6\}$ for $j \in \{2, .., p\}$. The outcome Y is generated according to the model:

$$Y = X^\top\theta_* + X_2\varepsilon,$$

where ε is a random disturbance which is independent of X and follows the univariate normal distribution with mean zero and standard deviation 0.25. We set the sparsity $s = 5$ and the true parameter value $\theta^*_j = 1$ for s equispaced values.

We compared the finite-sample performance among ℓ_0-PQR, ℓ_0-CQR, and ℓ_1-PQR estimators. We considered simulation configurations with $p \in \{10, 500\}$ to assess the performance in both the low and high dimensional settings. In each simulation repetition, we generated a training sample of $n = 100$ observations for estimating the parameter vector θ and another independent validation sample of 100 observations for calibrating the tuning parameters of these estimation approaches. Moreover, we also generated a test sample of 5000 observations for evaluating the out-of-sample predictive performance.

We focused our simulation study on median regression ($\tau = 0.5$). To implement the ℓ_1-PQR approach, we adopted the ℓ_1-penalized quantile regression estimator of Belloni and Chernozhukov (2011) with the penalty level given by

$$\lambda_{BC} \equiv c_{BC} \Lambda \left(1 - \alpha | X\right),$$

where $\Lambda \left(1 - \alpha | X\right)$ is the $(1 - \alpha)$ level quantile of the random variate Λ, which is defined in Belloni and Chernozhukov (2011, equation (2.6)), conditional on the covariate vector X. Following Belloni and Chernozhukov (2011), we set $\alpha = 0.1$. Moreover, we calibrated the optimal tuning value c_{BC} from the set $\mathcal{S} = \{0.1, 0.2, ..., 1.9, 2\}$ using the aforementioned validation sample in the setup with $p \geq 100$. For the low dimensional setup with $p < 100$, we performed this calibration over an expanded set $\mathcal{S} \cup \{0\}$, thereby allowing for an estimating model that did not penalize any parameter.

To implement the ℓ_0-CQR method, we solved over the training sample the ℓ_0-constrained estimation problem (2.3) for sparsity level q ranging from 1 up to $p \wedge 25$. To solve (2.3) with $\tau = 0.5$ for a given value of q, following Bertsimas, King, and Mazumder (2016, Section 6), we used the MIO-based, ℓ_0-constrained LAD approach with a warm-start strategy by supplying the MIO solver an initial guess computed via the discrete first-order approximation algorithms. We then calibrated the optimal sparsity level among this set of q values using the calibration sample. The resulting ℓ_0-CQR estimator was then constructed based on the model associated with the calibrated optimal sparsity level.

Finally, for our ℓ_0-PQR method, for a given value of c in (4.1), we solved the problem (2.4) with $k_0 = 100$ using our MIO computational approach of Section 4, where we warm-started the MIO solver by supplying as an initial guess the approximate solution obtained through the first-order method of Section 4.3. As in the ℓ_1-PQR case, we calibrated the optimal tuning scalar c over the set $\mathcal{S}$ using the calibration sample in the setup with $p \geq k_0$ and over the expanded set $\mathcal{S} \cup \{0\}$ in the setup with $p < k_0$.

We provided further details here on the implementation of our first-order approximation procedure in Algorithm 1. We set the tolerance level ϵ to be $2 \cdot 10^{-4}$ and parameter l of the quadratic envelope in (4.12) to be $2h$, where h is the Lipschitz constant given by (4.14). Note that Algorithm 1 also requires an initial guess. We therefore ran it for $T = 50$ times, each of which was performed with a different initial guess and used the output that delivered the best penalized objective function value in (2.4) as the resulting first-order approximate solution. We chose these T initial guesses sequentially where the first one was the ℓ_1-PQR solution of Belloni and Chernozhukov (2011) implemented with its tuning value c_{BC} set to be identical to the given value c in (4.1) whereas, for $t \in \{2, ..., T\}$, the t-th initial guess was subsequently constructed as the solution to the standard quantile regression of the outcome Y on those covariates selected in the output of Algorithm 1 which was initiated with the $(t-1)$-th initial guess. We found this implementation procedure worked very well in both our simulation study here and the empirical application of Section 6.

We specified the parameter space Θ to be $[-10, 10]^p$ for the MIO computation of both the ℓ_0-PQR and ℓ_0-CQR estimators. Throughout this paper, we used the MATLAB implementation of the Gurobi Optimizer (version 8.1.1) to solve all the MIO problems. Moreover, all numerical computations were done on a desktop PC (Windows 7) equipped with 128 GB RAM and a CPU processor (Intel i9-7980XE) of 2.6 GHz. To reduce computation cost in all MIO computations associated with the covariate configuration of $p = 500$, we set the MIO solver time limit to be 10 minutes beyond which we forced the solver to stop early and used the best discovered feasible solution to construct the resulting ℓ_0-PQR and ℓ_0-CQR estimators.

We reported performance results based on 100 simulation repetitions. We considered the following performance measures. Abusing the notation a bit, let $\widehat{\theta}$ denote the estimated parameters under a given quantile regression approach. To assess the predictive performance, we reported the relative risk, which is the ratio of the median predictive risk evaluated at the estimate $\widehat{\theta}$ over that evaluated at the true value θ_*. We approximated the out-of-sample predictive risk using the generated 5000-observation test sample. Let in_RR and out_RR respectively denote the average of in-sample and that of out-of-sample relative risks over the simulation repetitions.

We also reported the estimation performance in terms of both the average parameter estimation error defined as $\mathbb{E}[\|\widehat{\theta}-\theta_*\|_2]$ and the average regression function estimation error defined as $\mathbb{E}[|X^\top(\widehat{\theta}-\theta_*)|^2]$. Finally, we examined the variable selection performance. We say that a covariate X_j is effectively selected if and only if the magnitude of $\widehat{\theta}_j$ is larger than a small tolerance level (e.g., 10^{-5} as used in our numerical study) which is distinct from zero in numerical computation. Let $Avg\_sparsity$ denote the average number of effectively selected covariates. Let $Corr\_sel$ be the proportion of the truly relevant covariates being effectively selected. Let $Orac\_sel$ be the proportion of obtaining an oracle variable selection outcome where the set of effectively selected covariates coincides exactly with that of the truly relevant covariates. Finally, let Num_irrel denote the average number of effectively selected covariates whose true regression coefficients are zero.

5.1 Simulation Results

We now present in Tables 1 and 2 the simulation results for the performance comparison between the ℓ_0-PQR, ℓ_0-CQR, and ℓ_1-PQR approaches. For ℓ_0-PQR, we report performance measures for both the implementation based on the first-order (FO) approximation and that based on the MIO, which was warm-started by using the FO solutions as initial guesses. We find that, regarding the predictive performance, all the three competing approaches performed comparably well for both in-sample and out-of-sample relative risks under the low dimensional covariate design. By contrast, for the high dimensional design, both ℓ_0-PQR and ℓ_0-CQR considerably dominated ℓ_1-PQR in terms of out-of-sample predictive performance.

Table 1: Simulation comparison for $p = 10$

$p = 10$	ℓ_0-PQR		ℓ_0-CQR	ℓ_1-PQR
	MIO	FO		
$Corr\_sel$	1	1	1	1
$Orac\_sel$	0.78	0.77	0.6	0.05
Num_irrel	1.10	1.11	1.02	3.13
$Avg\_sparsity$	6.10	6.11	6.02	8.13
$\mathbb{E}\left[\left\|\widehat{\theta} - \theta_*\right\|_2\right]$	0.0355	0.0364	0.0382	0.0476
$\mathbb{E}[\|X^\top(\widehat{\theta} - \theta_*)\|^2]$	0.0014	0.0015	0.0016	0.0021
in_RR	0.9766	0.9761	0.9744	0.9692
out_RR	1.0292	1.0300	1.0309	1.0407

Table 2: Simulation comparison for $p = 500$

$p = 500$	ℓ_0-PQR		ℓ_0-CQR	ℓ_1-PQR
	MIO	FO		
$Corr\_sel$	1	1	1	1
$Orac\_sel$	1	1	0.97	0
Num_irrel	0	0	0.03	29.26
$Avg\_sparsity$	5	5	5.03	34.26
$\mathbb{E}\left[\left\|\widehat{\theta} - \theta_*\right\|_2\right]$	0.0285	0.0285	0.0289	0.1358
$\mathbb{E}[\|X^\top(\widehat{\theta} - \theta_*)\|^2]$	0.0010	0.0010	0.0010	0.0193
in_RR	0.9811	0.9811	0.9810	0.7930
out_RR	1.0246	1.0246	1.0253	1.2828

Turning to the variable selection results, we see that all the three estimation approaches had perfect $Corr\_sel$ rates and hence were effective for selecting the relevant covariates.

However, superb $Corr\_sel$ performance might just be a consequence of overfitting, which may result in excessive selection of irrelevant covariates and adversely impact on the out-of-sample predictive performance. From the results on the variable selection performance measures, we note that the number of irrelevant variables selected under ℓ_1-PQR was quite large relatively to those under both ℓ_0-PQR and ℓ_0-CQR in the high dimensional setup even though all these approaches exhibited the effect of reducing the covariate space dimension. This echoes with the finding in the setup of $p = 500$ that ℓ_1-PQR had far better in-sample fit in terms of in_RR yet worse out-of-sample fit in terms of out_RR relatively to the other two ℓ_0-norm-based approaches. Besides, while we could observe nonzero and high values of $Orac\_Sel$ for the ℓ_0-PQR and ℓ_0-CQR approaches in both the low and high dimensional setups, the ℓ_1-PQR approach could not induce any oracle variable selection outcome in these simulations. Regarding the performance in parameter and regression function estimation, we find that all the three approaches performed quite well for the estimation of the true regression function; however, the ℓ_1-PQR approach incurred much larger parameter estimation error relative to both the ℓ_0-PQR and ℓ_0-CQR approaches. Finally, for the ℓ_0-PQR approach, the FO-based algorithm as a standalone solution algorithm also performed very well. In fact, we find that in the high dimensional setup, all the MIO-based ℓ_0-PQR computations could not converge within the 10-minute computational time limit and the discovered solutions upon early stopping coincided with the FO-based solutions which were used to warm-start the MIO solver. This indicates that the first-order algorithm already located high-quality ℓ_0-PQR solutions upon which further improvements through the global optimization solver of MIO could not be obtained within the given computational time constraint. These simulation results thus shed lights on the usefulness of our first-order approximation approach for solving ℓ_0-PQR problems in the presence of computational resource constraints.

6 An Application to Conformal Prediction

In this section, we compare ℓ_0-PQR and ℓ_0-CQR with ℓ_1-PQR via a real data application to conformal prediction of birth weights. In particular, we employ conformalized quantile regression (Romano, Patterson, and Candes, 2019) to construct prediction intervals for birth weights.

We now describe the split conformalized quantile regression procedure (see Algorithm 1 of Romano, Patterson, and Candes, 2019). First, we split the data into a proper training set, indexed by $\mathcal{I}_1$, and a calibration set, indexed by $\mathcal{I}_2$. For each quantile regression algorithm, we use the proper training set $\mathcal{I}_1$ to obtain the estimates of two conditional quantile functions $\widehat{Q}_{\alpha/2}(Y|X = x)$ and $\widehat{Q}_{1-\alpha/2}(Y|X = x)$ for a given level $\alpha \in (0, 0.5)$. Then, the following scores are evaluated on the calibration set $\mathcal{I}_2$ as

$$E_i \equiv \max\{\widehat{Q}_{\alpha/2}(Y|X = X_i) - Y_i, Y_i - \widehat{Q}_{1-\alpha/2}(Y|X = X_i)\}$$

for each $i \in \mathcal{I}_2$. Finally, given new covariates X_{n+1}, construct the prediction interval for Y_{n+1} as

$$C(X_{n+1}) \equiv \Big[\widehat{Q}_{\alpha/2}(Y|X = X_{n+1}) - Q_{1-\alpha}(E, \mathcal{I}_2), \widehat{Q}_{1-\alpha/2}(Y|X = X_{n+1}) + Q_{1-\alpha}(E, \mathcal{I}_2)\Big] \tag{6.1}$$

where $Q_{1-\alpha}(E, \mathcal{I}_2)$ is the $(1-\alpha)(1+1/|\mathcal{I}_2|)$-th empirical quantile of $\{E_i : i \in \mathcal{I}_2\}$. Remarkably, Theorem 1 of Romano, Patterson, and Candes (2019) guarantees that the prediction interval (6.1) satisfies the marginal, distribution-free, finite-sample coverage in the sense that

$$\mathbb{P}\{Y_{n+1} \in C(X_{n+1})\} \geq 1 - \alpha,$$

provided that the data $\{(Y_i, X_i) : i = 1, \ldots, n+1\}$ are exchangeable.

We look at the dataset on birth weights originally analyzed by Almond, Chay, and Lee (2005). We use the excerpt from Cattaneo (2010) available at `http://www.stata-press.`

com/data/r13/cattaneo2.dta. The sample size is 4642 and the outcome of interest is infant birth weight measured in kilograms. The basic covariates include 20 variables concerning mother's age, mother's years of education, father's age, father's years of education, number of prenatal care visits, trimester of first prenatal care visit, birth order of an infant, months since last birth, an indicator variable whether a newborn died in previous births, mother's smoking behavior during pregnancy, mother's alcohol consumption during pregnancy, mother's marital status, mother's and father's hispanic status and race (being white or not), an indicator variable whether a mother was born abroad, and three dummy variables indicating seasons of the birth.

We conduct conformal prediction of infant birth weights with nominal level $\alpha = 0.1$. We split the sample randomly into four subsets of about equal size: $\mathcal{I}_1$, $\mathcal{I}_2$, $\mathcal{I}_3$ and $\mathcal{I}_4$. As described above, the set $\mathcal{I}_1$ is the training sample for the estimation of conditional quantile functions. We perform this estimation respectively using the ℓ_0-PQR, ℓ_0-CQR and ℓ_1-PQR approaches. We calibrate the tuning parameters in these three competing quantile regression approaches using the validation sample $\mathcal{I}_2$. Estimation and calibration are performed in the same fashion as described in the simulation study. With the calibrated optimal tuning parameter value, we use the set $\mathcal{I}_3$ to estimate the out-of-sample quantile prediction risk and conformalize quantile regression estimates by constructing $\{E_i : i \in \mathcal{I}_3\}$. We then evaluate the coverage performance of the prediction interval (6.1) over the test sample $\mathcal{I}_4$. We carry out a number of replications of such random splitting exercises and report the averages of estimated sparsity, out-of-sample prediction risk as well as length and coverage of the prediction interval across these replications. Specifically, we use 100 replications for the cases of ℓ_1-PQR and the first-order approximation (FO)-based implementation of ℓ_0-PQR. To mitigate the computational cost, we reduce the number of replications to 30 for the MIO-based ℓ_0-PQR and ℓ_0-CQR cases. Moreover, every MIO computation in this empirical study is conducted under a 10-minute computational time constraint.

We consider the following five different dictionary specifications. The first one is concerned with a covariate vector of $p = 21$ that includes a regression intercept together with the aforementioned 20 basic explanatory variables. The second specification modifies the first by discretizing both the maternal and paternal years of education into four categories indicating whether the schooling year is less than 12, exactly 12, between 12 and 16, or at least 16. In addition, we replace number of prenatal care visits, months since last birth and both parents' ages by cubic B-spline terms using 4 interior B-spline knots. These allow us to approximate smooth functions of these variables in the quantile regression analysis. We exclude the B-spline intercept terms so that the resulting covariate vector for the second specification has dimension $p = 49$. The third covariate specification consists of all variables in the second specification together with those obtained by interacting the B-spline expansion terms with the other explanatory variables. This then renders $p = 609$ in the third covariate specification scenario. Both the fourth and fifth specifications are constructed using the same procedure as for the third case except that we enlarge the covariate dimensions by using respectively 12 and 16 interior B-spline knots for these two high dimensional scenarios, each of which comprises 1281 and 1617 covariates respectively.

6.1 Empirical Results

We summarize in Figures 1 and 2 statistical performance results under the approaches of ℓ_1-PQR, ℓ_0-CQR and both the FO- and MIO-based implementations of ℓ_0-PQR. For each estimation method, we also report in Figure 3 its computational performance, which is measured by the employed CPU seconds that are averaged over the range of tuning parameter values and across the random splitting replications.

From Figure 1, we find that ℓ_1-PQR tends to induce a far denser estimating model than the ℓ_0-based approaches. At the 5% quantile level, the number of selected covariates under ℓ_1-PQR could reach around 30 when $p > 10^3$, whereas that quantity is only about 7, 12 and 2 under ℓ_0-CQR, FO- and MIO-based ℓ_0-PQR respectively. Similar sparsity pattern also emerges at the 95% quantile level. On the other hand, all these quantile regression

Figure 1: Results on Estimated Sparsities and Out-of-Sample Prediction Risks

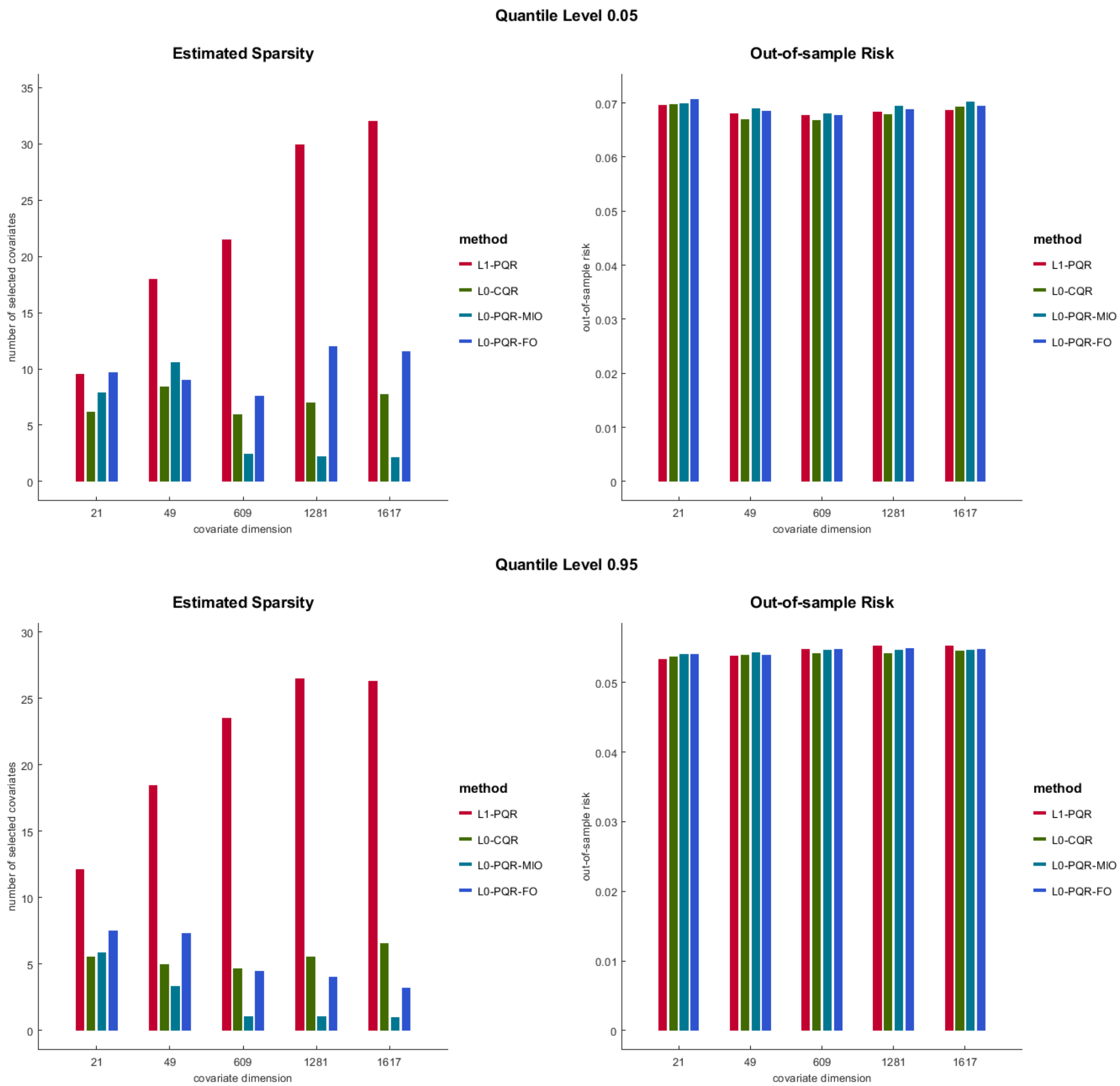
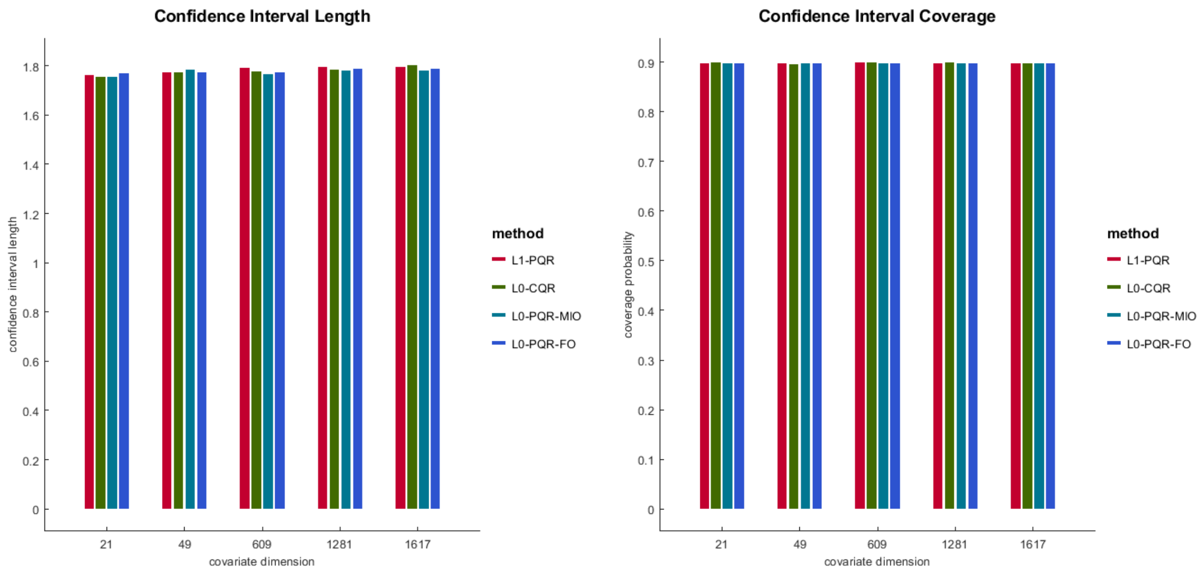


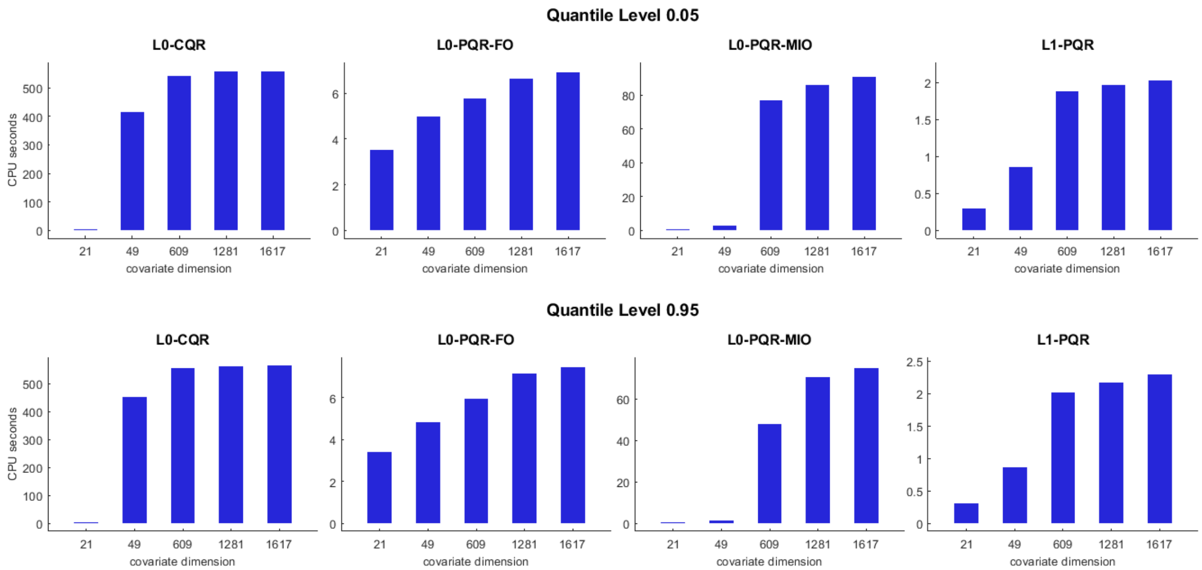
Figure 2: Lengths and Coverages of Conformalized Prediction Intervals



approaches exhibit comparable out-of-sample quantile prediction risks for all the estimation scenarios. Figure 2 also indicates that lengths and coverages of the conformalized prediction intervals are essentially identical across all the estimation approaches. Moreover, actual sizes of these prediction intervals are on average close to the nominal size. On the whole, these statistical performance results reveal that both ℓ_0-CQR and ℓ_0-PQR are capable of delivering sparser solutions than ℓ_1-PQR whilst maintaining commensurate prediction accuracy.

We now turn attention to computational performance of these competing approaches. From Figure 3, it is evident that ℓ_1-PQR enjoys the best computational performance with its average computation time being capped below 2.5 CPU seconds across all the estimation scenarios. FO-based ℓ_0-PQR also performs very well as its average computation time does not exceed 7.5 seconds in all scenarios. Relative to ℓ_1-PQR and FO-based ℓ_0-PQR, the other two methods are implemented through MIO and can be observed to be far more computationally intensive in this numerical study. For high dimensional scenarios, Figure 3 indicates a substantial computational performance difference between the MIO and non-MIO-based approaches. This can be anticipated because of the non-convexity in

Figure 3: Results on Computational Performance



the estimation problems for both ℓ_0-CQR and MIO-based ℓ_0-PQR. It is interesting to note that MIO-based ℓ_0-PQR appears to incur much milder computational cost than ℓ_0-CQR. In particular, the average computation time for MIO-based ℓ_0-PQR is still within 3 seconds in the case with $p = 49$ and does not go over 100 seconds across all estimation scenarios, whereas that for ℓ_0-CQR already reaches more than 400 seconds in the $p = 49$ case and is well above 500 seconds in most estimation configurations. Based on the numerical results, we recommend using FO-based ℓ_0-PQR for a large scale problem and MIO-based ℓ_0-PQR for a moderate size problem.

7 Conclusions

In this paper, we study estimation of a sparse high dimensional quantile regression model. The main contributions of this paper are twofold. First, we derive non-asymptotic expectation bounds on the excess quantile prediction risk as well as the mean-square parameter and regression function estimation errors of both the ℓ_0-PQR and ℓ_0-CQR estimators. These theoretical results imply the minimax optimal rates of convergence. Moreover, we characterize expected Hamming loss for the ℓ_0-penalized estimator. The second contribution is computational. We provide an exact computation approach for ℓ_0-PQR through the method of mixed integer optimization. We also develop a first-order approximation algorithm for solving large scale ℓ_0-PQR problems. Through Monte Carlo simulations and a real-data application, we find that both ℓ_0-PQR and ℓ_0-CQR perform fairly well and produce much sparser solutions than ℓ_1-PQR does. Our theoretical and numerical results suggest that ℓ_0-based approaches are worthy competitors to ℓ_1-based approaches in sparse quantile regression. Recently, Hazimeh and Mazumder (2020) developed fast computational methods for ℓ_0-penalized least squares with an additional ℓ_1- or ℓ_2-penalty term. It is an interesting future research topic to extend their approach to quantile regression and investigate its statistical properties.

A Proofs

A.1 Lemmas

For any $\theta, \theta_1, \theta_2 \in \Theta$, define $\overline{S}_n(\theta) \equiv S_n(\theta) - S(\theta)$, $\Delta(\theta_1, \theta_2) \equiv S(\theta_1) - S(\theta_2)$, and $\overline{\Delta}_n(\theta_1, \theta_2) \equiv \overline{S}_n(\theta_1) - \overline{S}_n(\theta_2)$. By Assumption 4, we have that, for some given scalar ε_*^2, which will be chosen later, we can find a point θ_*' in Θ' such that $\|\theta_*'\|_0 = \|\theta_*\|_0$ and $\Delta(\theta_*', \theta_*) \leq \varepsilon_*^2$. We start with the following basic inequality.

Lemma 1 (Basic inequality)**.** *For $k_0 \geq s$,*

$$\Delta(\widehat{\theta}, \theta_*') + \lambda \|\widehat{\theta} - \theta_*'\|_0 \leq \overline{\Delta}_n(\theta_*', \widehat{\theta}) + 2\lambda s.$$

Proof of Lemma 1. Using (2.4), we have that, for $k_0 \geq s$,

$$S_n(\widehat{\theta}) + \lambda \|\widehat{\theta}\|_0 \leq S_n(\theta_*') + \lambda \|\theta_*'\|_0. \tag{A.1}$$

Using (A.1), we can deduce that

$$\Delta(\widehat{\theta}, \theta'_*) + \lambda\|\widehat{\theta}\|_0 \leq \overline{\Delta}_n(\theta'_*, \widehat{\theta}) + \lambda s. \tag{A.2}$$

Then, the lemma follows from (A.2) together with an application of triangle inequality. ∎

Let

$$V_x \equiv \sup_{\theta \in \mathbb{B}(k_0)} \frac{\overline{\Delta}_n(\theta'_*, \theta)}{\Delta(\theta, \theta_*) + \varepsilon_*^2 + x^2}. \tag{A.3}$$

Lemma 2 (Preliminary Probability Bounds)**.** *Let Assumptions 1 and 4 hold. Suppose $k_0 \geq s$. Then for a constant $0 < \eta < 1$,*

$$\mathbb{P}\left[\Delta(\widehat{\theta}, \theta_*) \geq \frac{2}{1-\eta}\lambda s + \frac{1+\eta}{1-\eta}\varepsilon_*^2 + \frac{\eta}{1-\eta}x^2\right] \leq \mathbb{P}\left(V_x \geq \eta\right), \tag{A.4}$$

$$\mathbb{P}\left[\|\widehat{\theta} - \theta_*\|_0 \geq \frac{4-2\eta}{1-\eta}s + \lambda^{-1}\left\{\frac{1+\eta}{1-\eta}\varepsilon_*^2 + \frac{\eta}{1-\eta}x^2\right\}\right] \leq \mathbb{P}\left(V_x \geq \eta\right). \tag{A.5}$$

Proof of Lemma 2. First, since $\lambda\|\widehat{\theta} - \theta'_*\|_0$ is always non-negative, it follows from Lemma 1 that

$$\Delta(\widehat{\theta}, \theta_*) = \Delta(\widehat{\theta}, \theta'_*) + \Delta(\theta'_*, \theta_*) \leq \varepsilon_*^2 + 2\lambda s + \overline{\Delta}_n(\theta'_*, \widehat{\theta}). \tag{A.6}$$

Using (A.3) and (A.6), we have that, if $V_x < \eta$ for some positive constant $\eta < 1$, then

$$\begin{aligned} \Delta(\widehat{\theta}, \theta_*) &\leq 2\lambda s + \varepsilon_*^2 + V_x\left[\Delta(\widehat{\theta}, \theta_*) + \varepsilon_*^2 + x^2\right] \\ &< \frac{2}{1-\eta}\lambda s + \frac{1+\eta}{1-\eta}\varepsilon_*^2 + \frac{\eta}{1-\eta}x^2. \end{aligned} \tag{A.7}$$

Furthermore, note that

$$\Delta(\widehat{\theta}, \theta_*) + \lambda\|\widehat{\theta} - \theta'_*\|_0 \leq \varepsilon_*^2 + \Delta(\widehat{\theta}, \theta'_*) + \lambda\|\widehat{\theta} - \theta'_*\|_0.$$

By Assumption 1, $\Delta(\widehat{\theta}, \theta_*) \geq 0$. Therefore, using Lemma 1, it follows that $\|\widehat{\theta} - \theta'_*\|_0 \leq$

$2s + \lambda^{-1}[\overline{\Delta}_n(\theta'_*, \widehat{\theta}) + \varepsilon_*^2]$. Thus, if $V_x < \eta$, we have

$$\begin{aligned}\|\widehat{\theta} - \theta'_*\|_0 &\leq 2s + \lambda^{-1} V_x \left[\Delta(\widehat{\theta}, \theta_*) + \varepsilon_*^2 + x^2\right] + \lambda^{-1}\varepsilon_*^2 \\ &< 2s + \lambda^{-1}\eta \left[\frac{2}{1-\eta}\lambda s + \frac{1+\eta}{1-\eta}\varepsilon_*^2 + \frac{\eta}{1-\eta}x^2 + \varepsilon_*^2 + x^2\right] + \lambda^{-1}\varepsilon_*^2 \text{ by (A.7).}\end{aligned}$$

Arranging the terms, we therefore have that

$$\mathbb{P}\left[\|\widehat{\theta} - \theta'_*\|_0 \geq \frac{2}{1-\eta}s + \lambda^{-1}\left\{\frac{1+\eta}{1-\eta}\varepsilon_*^2 + \frac{\eta}{1-\eta}x^2\right\}\right] \leq \mathbb{P}\left(V_x \geq \eta\right),$$

which yields (A.5) by an application of triangle inequality. ∎

By Assumption 4,

$$V_x = \sup_{\theta \in \mathbb{B}'(k_0)} \frac{\overline{\Delta}_n(\theta'_*, \theta)}{\Delta(\theta, \theta_*) + \varepsilon_*^2 + x^2}.$$

In other words, it suffices to take the supremum over the countable subset Θ'.

To bound $\mathbb{P}(V_x \geq \eta)$ in (A.4) and (A.5), we will use Bousquet's inequality and a technical lemma from Massart and Nédélec (2006). For the sake of easy referencing, these results are reproduced below.

Lemma 3 (Bousquet's inequality)**.** *Suppose that $\mathcal{F}$ is a countable family of measurable functions such that for every $f \in \mathcal{F}, P(f^2) \leq v$ and $\|f\|_\infty \leq b$ for some positive constants v and b. Define $Z \equiv \sup_{f \in \mathcal{F}}(P_n - P)(f)$. Then for every positive y,*

$$\mathbb{P}\left[Z - \mathbb{E}\left[Z\right] \geq \sqrt{2\frac{(v + 4b\mathbb{E}[Z])\, y}{n}} + \frac{2by}{3n}\right] \leq e^{-y}. \tag{A.8}$$

Lemma 4 (Lemma A.5 of Massart and Nédélec (2006))**.** *Let $\mathcal{S}$ be a countable set, $u \in \mathcal{S}$ and $a : \mathcal{S} \to \mathbb{R}_+$ such that $a(u) = \inf_{t \in \mathcal{S}} a(t)$. Define $\mathcal{B}(\varepsilon) = \{t \in \mathcal{S} : a(t) \leq \varepsilon\}$. Let Z be a process indexed by $\mathcal{S}$ and assume that the nonnegative random variable $\sup_{t \in \mathcal{B}(\varepsilon)}[Z(u) - Z(t)]$ has finite expectation for any positive number ε. Let ψ be a nonnegative function on $\mathbb{R}_+$ such that $\psi(x)/x$ is* *nonincreasing on $\mathbb{R}_+$ and satisfies for some positive ε_* :*

$$\mathbb{E}\left[\sup_{t\in\mathcal{B}(\varepsilon)}[Z(u)-Z(t)]\right]\leq\psi(\varepsilon)\quad \textit{for any } \varepsilon\geq\varepsilon_*.$$

Then, one has, for any positive number $x \geq \varepsilon_*$,

$$\mathbb{E}\left[\sup_{t\in\mathcal{S}}\left(\frac{Z(u)-Z(t)}{a^2(t)+x^2}\right)\right]\leq 4x^{-2}\psi(x).$$

A.2 Proof of Theorem 1

Proof of Theorem 1. We prove Theorem 1 by adopting the ideas behind the proof of Theorem 2 in Massart and Nédélec (2006). For $q \geq 0$, let $\mathbb{B}'(q) \equiv \{\theta \in \Theta' : \|\theta\|_0 \leq q\}$. By Assumption 1, $\Delta(\theta, \theta_*) \geq 0$ for all $\theta \in \mathbb{B}'(k_0)$. Using this fact and Assumptions 2 and 3, we have that, for all $i = 1, \ldots, n$ and for all $\theta \in \mathbb{B}'(k_0)$,

$$\left|\frac{\rho(Y_i, X_i^\top\theta_*') - \rho(Y_i, X_i^\top\theta)}{\Delta(\theta,\theta_*)+\varepsilon_*^2+x^2}\right| \leq \frac{2LB^2\,(s+k_0)}{x^2} \equiv b_x.$$

Moreover, by Assumptions 2, 3, 5 and 6, for all $\theta \in \mathbb{B}'(k_0)$,

$$\begin{aligned}
\mathbb{E}\left(\left[\rho(Y_i, X_i^\top\theta_*') - \rho(Y_i, X_i^\top\theta)\right]^2\right) &\leq L^2B^2\,\|\theta-\theta_*'\|_1^2 \\
&\leq L^2B^2\|\theta-\theta_*'\|_0\left(\|\theta-\theta_*\|_2+\|\theta_*'-\theta_*\|_2\right)^2 \\
&\leq L^2B^2\,(s+k_0)\,\kappa_1^{-2}\kappa_0^{-2}\left(\sqrt{\Delta(\theta,\theta_*)}+\varepsilon_*\right)^2 \\
&\leq 4L^2B^2\,(s+k_0)\,\kappa_1^{-2}\kappa_0^{-2}\left(\Delta(\theta,\theta_*)+\varepsilon_*^2\right)
\end{aligned}$$

and therefore

$$\mathbb{E}\left(\left[\frac{\rho(Y_i, X_i^\top\theta_*') - \rho(Y_i, X_i^\top\theta)}{\Delta(\theta,\theta_*) + \varepsilon_*^2 + x^2}\right]^2\right) \leq \frac{4L^2B^2\,(s+k_0)\,\kappa_1^{-2}\kappa_0^{-2}\,(\Delta(\theta,\theta_*)+\varepsilon_*^2)}{[\Delta(\theta,\theta_*)+\varepsilon_*^2+x^2]^2}$$
$$\leq \frac{4L^2B^2\,(s+k_0)\,\kappa_1^{-2}\kappa_0^{-2}}{x^2}\sup_{\varepsilon\geq 0}\frac{\varepsilon}{\varepsilon+x^2}$$
$$\leq \frac{4L^2B^2\,(s+k_0)\,\kappa_1^{-2}\kappa_0^{-2}}{x^2} \equiv v_x.$$

Choose $b = b_x$ and $v = v_x$ in Lemma 3. By (A.8), we then have that for every positive y,

$$\mathbb{P}\left[V_x - \mathbb{E}[V_x] \geq \sqrt{2\frac{(v_x + 4b_x\mathbb{E}[V_x])y}{n}} + \frac{2b_x y}{3n}\right] \leq \exp(-y). \tag{A.9}$$

We now bound $\mathbb{E}[V_x]$ using Lemma 4. Let $a^2(\theta) \equiv \Delta(\theta_*', \theta_*) \vee \Delta(\theta,\theta_*)$ for any $\theta \in \Theta$. Then $\Delta(\theta,\theta_*) \leq a^2(\theta) \leq \Delta(\theta,\theta_*) + \varepsilon_*^2$. Therefore,

$$\mathbb{E}[V_x] \leq \mathbb{E}\left[\sup_{\theta\in\mathbb{B}'(k_0)}\frac{\overline{\Delta}_n(\theta_*',\theta)}{a^2(\theta)+x^2}\right]$$

and for every $\varepsilon \geq \varepsilon_*$,

$$\mathbb{E}\left[\sup_{\theta\in\mathbb{B}'(k_0):a(\theta)\leq\varepsilon}\overline{\Delta}_n(\theta_*',\theta)\right] \leq \mathbb{E}\left[\sup_{\theta\in\mathbb{B}'(k_0):\Delta(\theta,\theta_*)\leq\varepsilon^2}\overline{\Delta}_n(\theta_*',\theta)\right].$$

The next step is to find a function ψ such that

$$\mathbb{E}\left[\sup_{\theta\in\mathbb{B}'(k_0):\Delta(\theta,\theta_*)\leq\varepsilon^2}\overline{\Delta}_n(\theta_*',\theta)\right] \leq \psi(\varepsilon) \quad \text{for any } \varepsilon \geq \varepsilon_*.$$

Let $\epsilon_1, \ldots, \epsilon_n$ denote a Rademacher sequence that is independent of $\{(Y_i, X_i) : i = 1, \ldots, n\}$. By the symmetrization and contraction theorems (e.g., Theorems 14.3 and 14.4 of Bühlmann and Van De Geer (2011)),

$$
\begin{aligned}
&\mathbb{E}\left[\sup_{\theta\in\mathbb{B}'(k_0):\Delta(\theta,\theta_*)\le\varepsilon^2}\overline{\Delta}_n(\theta'_*,\theta)\right]\\
&\le \mathbb{E}\left[\sup_{\theta\in\mathbb{B}'(k_0):\Delta(\theta,\theta_*)\le\varepsilon^2}|\overline{\Delta}_n(\theta'_*,\theta)|\right]\\
&\le 2\mathbb{E}\left[\sup_{\theta\in\mathbb{B}'(k_0):\Delta(\theta,\theta_*)\le\varepsilon^2}\left|n^{-1}\sum_{i=1}^n\epsilon_i\left\{\rho(Y_i,X_i^\top\theta)-\rho(Y_i,X_i^\top\theta'_*)\right\}\right|\right]\\
&\le 4L\mathbb{E}\left[\sup_{\theta\in\mathbb{B}'(k_0):\Delta(\theta,\theta_*)\le\varepsilon^2}\left|n^{-1}\sum_{i=1}^n\epsilon_i X_i^\top(\theta-\theta'_*)\right|\right].
\end{aligned}
$$

By Hölder's inequality

$$
\left|n^{-1}\sum_{i=1}^n\epsilon_i X_i^\top(\theta-\theta'_*)\right|\le\|\theta-\theta'_*\|_1\max_{1\le j\le p}\left|n^{-1}\sum_{i=1}^n\epsilon_i X_i^{(j)}\right|,
$$

where $X_i^{(j)}$ denotes the j-th component of the covariate vector X_i.

For $\theta\in\mathbb{B}'(k_0)$ that satisfies $\Delta(\theta,\theta_*)\le\varepsilon^2$, we have that, by Assumptions 5 and 6,

$$
\begin{aligned}
\|\theta-\theta'_*\|_1 &\le \sqrt{\|\theta-\theta'_*\|_0}\left(\|\theta-\theta_*\|_2+\|\theta'_*-\theta_*\|_2\right)\\
&\le (s+k_0)^{1/2}\kappa_1^{-1}\kappa_0^{-1}\left(\sqrt{\Delta(\theta,\theta_*)}+\sqrt{\Delta(\theta'_*,\theta_*)}\right)\\
&\le (s+k_0)^{1/2}\kappa_1^{-1}\kappa_0^{-1}(\varepsilon+\varepsilon_*)
\end{aligned}
$$

Therefore, we have that for any $\varepsilon\ge\varepsilon_*$,

$$
\begin{aligned}
\mathbb{E}\left[\sup_{\theta\in\mathbb{B}'(k_0):\Delta(\theta,\theta_*)\le\varepsilon^2}\overline{\Delta}_n(\theta'_*,\theta)\right]&\le 8L(s+k_0)^{1/2}\kappa_1^{-1}\kappa_0^{-1}\varepsilon\mathbb{E}\left[\max_{1\le j\le p}\left|n^{-1}\sum_{i=1}^n\epsilon_i X_i^{(j)}\right|\right]\\
&\le C(s+k_0)^{1/2}\varepsilon\sqrt{\frac{2\ln(2p)}{n}},
\end{aligned}
$$

where the last inequality follows from Hoeffding's inequality (e.g., Lemma 14.14 of Bühlmann and Van De Geer (2011)) together with (3.8) and Assumption 3. Hence, we set $\psi(x)\equiv$

$C\,(s+k_0)^{1/2}\,x\sqrt{2\ln(2p)/n}$. Thus, by Lemma 4, for any $x \geq \varepsilon_*$,

$$\mathbb{E}[V_x] \leq \frac{4C}{x}(s+k_0)^{1/2}\sqrt{\frac{2\ln(2p)}{n}}.$$

For every $y \geq 1$, set $x = \sqrt{My}\varepsilon_*$ for some constant $M \geq 1$, which will be chosen below, and

$$\varepsilon_* = 4C(s+k_0)^{1/2}\sqrt{\frac{2\ln(2p)}{n}}. \tag{A.10}$$

By (3.8) and (3.9), we have that $C^2\ln(2p) \geq LB^2$ and $\ln(2p) \geq 1$. Therefore,

$$\begin{aligned}
\mathbb{E}[V_x] &\leq \frac{1}{\sqrt{My}} \leq \frac{1}{M^{1/2}}, \\
\frac{b_x y}{n} &= \frac{LB^2}{16MC^2\ln(2p)} \leq \frac{1}{16M}, \\
\frac{v_x y}{n} &= \frac{1}{512M\ln(2p)} \leq \frac{1}{64M},
\end{aligned}$$

which implies that

$$\begin{aligned}
&\mathbb{E}[V_x] + \sqrt{2\frac{(v_x + 4b_x\mathbb{E}[V_x])y}{n}} + \frac{2b_x y}{3n} \\
\leq\ & \frac{1}{M^{1/2}} + \sqrt{\frac{1}{32M} + \frac{1}{2M^{3/2}}} + \frac{1}{24M} \\
\leq\ & \frac{3}{M^{1/2}}.
\end{aligned}$$

It thus follows from (A.9) that

$$\mathbb{P}\left[V_x \geq 3M^{-1/2}\right] \leq \exp(-y) \tag{A.11}$$

for $y \geq 1$. Now choose a sufficiently large M that satisfies

$$\eta \geq 3M^{-1/2} \tag{A.12}$$

for any given positive constant $\eta < 1$. Putting together (3.8), (A.4), (A.5), (A.10), (A.11) and (A.12), we then have that

$$\mathbb{P}\left[\Delta(\widehat{\theta}, \theta_*) \geq \frac{2}{1-\eta}\lambda s + 32C^2(s+k_0)\left(\frac{1+\eta+M\eta y}{1-\eta}\right)\frac{\ln(2p)}{n}\right] \leq \exp(-y),$$
$$\mathbb{P}\left[\|\widehat{\theta} - \theta_*\|_0 \geq \frac{4-2\eta}{1-\eta}s + 32\lambda^{-1}C^2(s+k_0)\left(\frac{1+\eta+M\eta y}{1-\eta}\right)\frac{\ln(2p)}{n}\right] \leq \exp(-y)$$

for $y \geq 1$. ∎

A.3 Proof of Theorem 2

Proof of Theorem 2. By (3.6) of Theorem 1 with the choice of $\eta = 1/2$, we have that, for every $y \geq 1$,

$$\mathbb{P}\left[S(\widehat{\theta}) - S(\theta_*) \geq A + By\right] \leq \exp(-y), \tag{A.13}$$

where

$$A \equiv 4\lambda s + 96C^2(s+k_0)\frac{\ln(2p)}{n},$$
$$B \equiv 32MC^2(s+k_0)\frac{\ln(2p)}{n}.$$

and the constant C is given by (3.8). Since $S(\widehat{\theta}) \geq S(\theta_*)$, result (3.10) thus follows by noting that

$$\begin{aligned}
&\mathbb{E}\left[S(\widehat{\theta}) - S(\theta_*)\right] \\
&= \int_0^\infty \mathbb{P}\left[S(\widehat{\theta}) - S(\theta_*) \geq t\right] dt \\
&= B\int_{-A/B}^\infty \mathbb{P}\left[S(\widehat{\theta}) - S(\theta_*) \geq A + By\right] dy \\
&= B\int_{-A/B}^1 \mathbb{P}\left[S(\widehat{\theta}) - S(\theta_*) \geq A + By\right] dy + B\int_1^\infty \mathbb{P}\left[S(\widehat{\theta}) - S(\theta_*) \geq A + By\right] dy \\
&\leq A + 2B,
\end{aligned}$$

where the last inequality above follows from applying (A.13) for $y \in [1, \infty)$. Moreover, using (3.4) and (3.5), we can deduce that

$$\begin{aligned}\mathbb{E}\left[R(\widehat{\theta})\right] &\leq \kappa_0^{-2}\mathbb{E}\left[S(\widehat{\theta}) - S\left(\theta_*\right)\right], \\ \mathbb{E}\left[\left\|\widehat{\theta} - \theta_*\right\|_2^2\right] &\leq \kappa_1^{-2}\kappa_0^{-2}\mathbb{E}\left[S(\widehat{\theta}) - S\left(\theta_*\right)\right],\end{aligned}$$

which, given (3.10), therefore imply (3.11) and (3.12). ∎

A.4 Proof of Theorem 3

Proof of Theorem 3. By (3.9), $p \geq 2$ so that $\ln(2p) \leq 2\ln p$. Thus using (3.7) in Theorem 1, we have that

$$\mathbb{P}\left[\|\widehat{\theta} - \theta_*\|_0 \geq \frac{4-2\eta}{1-\eta}s + 64\lambda^{-1}C^2(s+k_0)\left(\frac{1+\eta+M\eta y}{1-\eta}\right)\frac{\ln(p)}{n}\right] \leq \exp(-y)$$

for $y \geq 1$. Choose the smallest M that satisfies (A.12), i.e., $\eta = 3M^{-1/2}$. Because it is assumed that $k_0 \leq C_k s$ for a fixed constant C_k and $\lambda = C_\lambda \ln p/n$, taking $C_\lambda = 64\zeta_\lambda C^2(C_k+1)$ for some constant $\zeta_\lambda \geq 1$, we have that for $y \geq 1$,

$$\mathbb{P}\left[\|\widehat{\theta} - \theta_*\|_0 \geq As + Bsy\right] \leq \exp(-y), \tag{A.14}$$

where

$$A = \frac{4-2\eta}{1-\eta} + \zeta_\lambda^{-1}\frac{1+\eta}{1-\eta} \quad \text{and} \quad B = \zeta_\lambda^{-1}\frac{9}{\eta(1-\eta)}.$$

Using the integrated tail probability expectation formula for nonnegative integer valued random variables (see e.g. Lo (2019)) and following similar steps in the Proof of Theorem 2, we have that

$$\mathbb{E}\left[\|\widehat{\theta} - \theta_*\|_0\right] \leq (A+2B)s.$$

The conclusion of the theorem follows by first choosing a sufficiently small η and then selecting a sufficiently large ζ_λ. ∎

A.5 Proof of Corollary 1

Proof of Corollary 1. As discussed in Section 3.1, Assumptions 1–6 are satisfied for quantile regression with the Lipschitz constant $L = 1$, Assumption 4 holds by (3.3) and the presumption on the finiteness of $\mathbb{E}|Y|$, and Assumption 6 holds with $\kappa_1 = \sqrt{\omega}$.

Note that, for any $\theta \in \mathbb{B}(k_0)$, we have that $\left|x^\top(\theta - \theta_*)\right| \leq 2B^2\,(k_0 + s)$ by (3.3). Using (3.1), it hence follows from assumption (iv) of this corollary that

$$S(\theta) - S(\theta_*) \geq \frac{c_u}{2}\mathbb{E}\left[\left|X^\top(\theta - \theta_*)\right|^2\right] \text{ for all } \theta \in \mathbb{B}(k_0).$$

Thus, Assumption 5 of Theorem 1 also holds with $\kappa_0 = \sqrt{c_u/2}$. As a result, the corollary follows from Theorem 2. ∎

A.6 Proof of Corollary 2

Proof of Corollary 2. Repeating arguments used in the proof of Theorem 1 with $\lambda = 0$, we have that

$$\mathbb{P}\left[\Delta(\tilde{\theta}, \theta_*) \geq 32C^2\,(s+q)\left(\frac{1+\eta+M\eta y}{1-\eta}\right)\frac{\ln(2p)}{n}\right] \leq \exp(-y)$$

for $y \geq 1$. Then we can proceed as in the proof of Corollary 1. ∎

A.7 Proof of Proposition 1

Proof of Proposition 1. (a) Let $\widehat{t}$ denote a point in $H_{\delta,l}(\widehat{\theta}_\delta)$. By (4.15) and (4.16),

$$\widetilde{Q}_n(\widehat{t}; \widehat{\theta}_\delta, \delta, l) \leq \widetilde{Q}_n(\widehat{\theta}_\delta; \widehat{\theta}_\delta, \delta, l). \tag{A.15}$$

Using (4.12), we have that

$$\widetilde{Q}_n(\widehat{\theta}_\delta;\widehat{\theta}_\delta,\delta,l) = Q_n(\widehat{\theta}_\delta;\delta) \le Q_n(\widehat{t};\delta) \le \widetilde{Q}_n(\widehat{t};\widehat{\theta}_\delta,\delta,l). \tag{A.16}$$

Putting (A.15) and (A.16) together, we have that

$$\widetilde{Q}_n(\widehat{t};\widehat{\theta}_\delta,\delta,l) = \widetilde{Q}_n(\widehat{\theta}_\delta;\widehat{\theta}_\delta,\delta,l)$$

so that $\widehat{\theta}_\delta$ is also a minimizer to the problem (4.15) with $t = \widehat{\theta}_\delta$. It thus follows that $\widehat{\theta}_\delta \in H_{\delta,l}(\widehat{\theta}_\delta)$.

(b) Proof of part (b) follows closely that of Theorem 3.1 of Bertsimas, King, and Mazumder (2016). Note that, for any $l \ge h$, if $t' \in H_{\delta,l}(t)$, then

$$\begin{aligned} Q_n(t;\delta) &= \widetilde{Q}_n(t;t,\delta,l) \\ &\ge \widetilde{Q}_n(t';t,\delta,l) \\ &= \frac{l-h}{2}\left\|t'-t\right\|_2^2 + \widetilde{Q}_n(t';t,\delta,h) \\ &\ge \frac{l-h}{2}\left\|t'-t\right\|_2^2 + Q_n(t';\delta) \end{aligned} \tag{A.17}$$

so that

$$\left\|t'-t\right\|_2^2 \le \frac{2\left(Q_n(t;\delta) - Q_n(t';\delta)\right)}{l-h}. \tag{A.18}$$

Now let $l > h$ and consider the sequence t_m satisfying $t_{m+1} \in H_{\delta,l}(t_m)$. Since the parameter space Θ is compact, inequality (A.17) implies that $Q_n(t_m;\delta)$ decreases with m and thus converges to a limit Q^* as $m \longrightarrow \infty$. Using this fact together with (A.18), it follows that t_m also converges to a limit t^*. Applying (A.18) with $t' = t_{m+1}$ and $t = t_m$, we can deduce (4.17) by noting that

$$\min_{m=1,\ldots,N}\left\|t_{m+1}-t_m\right\|_2^2 \le \frac{1}{N}\sum\nolimits_{m=1}^{N}\left\|t_{m+1}-t_m\right\|_2^2 \le \frac{2\left(Q_n(t_1;\delta) - Q^*\right)}{N\,(l-h)}.$$

∎

A.8 Proof of Proposition 2

Proof of Proposition 2. By (4.9) and (4.11), we have that $Q_n(\theta;0) = S_n(\theta) + \lambda\|\theta\|_0$.

For each $\theta \in \Theta$,

$$Q_n(\theta;\delta) \le Q_n(\theta;0) \le Q_n(\theta;\delta) + \frac{\delta c_\tau}{2}.$$

Therefore, we can deduce that

$$\begin{aligned} Q_n(\widehat{\theta}_\delta;\delta) &\le \min_{\theta \in \mathbb{B}(k_0)} Q_n(\theta;0) \\ &\le Q_n(\widehat{\theta}_\delta;0) \\ &\le Q_n(\widehat{\theta}_\delta;\delta) + \frac{\delta c_\tau}{2} \\ &\le \min_{\theta \in \mathbb{B}(k_0)} Q_n(\theta;0) + \frac{\delta c_\tau}{2}. \end{aligned}$$

∎

References

Almond, D., K. Y. Chay, and D. S. Lee (2005): "The costs of low birth weight," *Quarterly Journal of Economics*, 120(3), 1031–1083.

Belloni, A., and V. Chernozhukov (2011): "ℓ_1-penalized quantile regression in high-dimensional sparse models," *Annals of Statistics*, 39(1), 82–130.

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

——— (2019): "Valid Post-Selection Inference in High-Dimensional Approximately Sparse Quantile Regression Models," *Journal of the American Statistical Association*, 114(526), 749–758.

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

Bousquet, O. (2002): "A Bennett Concentration Inequality and its Application to Suprema of Empirical Processes," *C. R. Math. Acad. Sci. Paris*, 334, 495–500.

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

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

Butucea, C., M. Ndaoud, N. A. Stepanova, and A. B. Tsybakov (2018): "Variable selection with Hamming loss," *Annals of Statistics*, 46(5), 1837–1875.

Cattaneo, M. D. (2010): "Efficient semiparametric estimation of multi-valued treatment effects under ignorability," *Journal of Econometrics*, 155(2), 138–154.

Chen, L.-Y., and S. Lee (2018a): "Best subset binary prediction," *Journal of Econometrics*, 206(1), 39–56.

——— (2018b): "High Dimensional Classification through ℓ_0-Penalized Empirical Risk Minimization," *arXiv preprint arXiv:1811.09540*.

Fan, J., Y. Fan, and E. Barut (2014): "Adaptive Robust Variable Selection," *Annals of Statistics*, 42(1), 324–351.

Hazimeh, H., and R. Mazumder (2020): "Fast best subset selection: Coordinate descent and local combinatorial optimization algorithms," *Operations Research*, forthcoming.

Huang, J., Y. Jiao, Y. Liu, and X. Lu (2018): "A Constructive Approach to L_0 Penalized Regression," *Journal of Machine Learning Research*, 19(10), 1–37.

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

——— (2017): "Quantile Regression: 40 Years On," *Annual Review of Economics*, 9(1), 155–176.

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

Lee, S., Y. Liao, M. H. Seo, and Y. Shin (2018): "Oracle Estimation of a Change Point in High-Dimensional Quantile Regression," *Journal of the American Statistical Association*, 113(523), 1184–1194.

Lo, A. (2019): "Demystifying the integrated tail probability expectation formula," *The American Statistician*, 73(4), 367–374.

Lv, S., H. Lin, H. Lian, and J. Huang (2018): "Oracle Inequalities for Sparse Additive Quantile Regression in Reproducing Kernel Hilbert Space," *Annals of Statistics*, 46(2), 781–813.

Massart, P., and E. Nédélec (2006): "Risk bounds for statistical learning," *Annals of Statistics*, 34(5), 2326–2366.

Nesterov, Y. (2005): "Smooth minimization of non-smooth functions," *Mathematical programming*, 103(1), 127–152.

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

Romano, Y., E. Patterson, and E. Candes (2019): "Conformalized quantile regression," in *Advances in Neural Information Processing Systems*, pp. 3538–3548.

Tibshirani, R. (1996): "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267–288.

van de Geer, S. A., and P. Bühlmann (2009): "On the conditions used to prove oracle results for the Lasso," *Electronic Journal of Statistics*, 3, 1360–1392.

Wang, L. (2013): "The L_1 penalized LAD estimator for high dimensional linear regression," *Journal of Multivariate Analysis*, 120, 135–151.

Wang, L. (2019): "L_1-regularized Quantile Regression with Many Regressors under Lean Assumptions," Retrieved from the University of Minnesota Digital Conservancy, `http://hdl.handle.net/11299/202063`.

Wang, L., I. Van Keilegom, and A. Maidman (2018): "Wild residual bootstrap inference for penalized quantile regression with heteroscedastic errors," *Biometrika*, 105(4), 859–872.

Wang, L., Y. Wu, and R. Li (2012): "Quantile Regression for Analyzing Heterogeneity in Ultra-High Dimension," *Journal of the American Statistical Association*, 107(497), 214–222.

Zheng, Q., L. Peng, and X. He (2015): "Globally adaptive quantile regression with ultra-high dimensional data," *Annals of Statistics*, 43(5), 2225–2258.

——— (2018): "High dimensional censored quantile regression," *Annals of Statistics*, 46(1), 308–343.