Journal of Econometrics xxx (xxxx) xxx

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Smoothed quantile regression with large-scale inference

Xuming He [a], Xiaoou Pan [b], Kean Ming Tan [a,*], Wen-Xin Zhou [b]

[a] Department of Statistics, University of Michigan, Ann Arbor, MI, 48109, USA
[b] Department of Mathematics, University of California, San Diego, La Jolla, CA, 92093, USA

ARTICLE INFO

Article history:
Received 1 November 2020
Received in revised form 27 July 2021
Accepted 30 July 2021
Available online xxxx

Keywords:
Bahadur–Kiefer representation
Convolution
Quantile regression
Multiplier bootstrap
Non-asymptotic statistics

ABSTRACT

Quantile regression is a powerful tool for learning the relationship between a response variable and a multivariate predictor while exploring heterogeneous effects. This paper focuses on statistical inference for quantile regression in the "increasing dimension" regime. We provide a comprehensive analysis of a convolution smoothed approach that achieves adequate approximation to computation and inference for quantile regression. This method, which we refer to as *conquer*, turns the non-differentiable check function into a twice-differentiable, convex and locally strongly convex surrogate, which admits fast and scalable gradient-based algorithms to perform optimization, and multiplier bootstrap for statistical inference. Theoretically, we establish explicit non-asymptotic bounds on estimation and Bahadur–Kiefer linearization errors, from which we show that the asymptotic normality of the conquer estimator holds under a weaker requirement on dimensionality than needed for conventional quantile regression. The validity of multiplier bootstrap is also provided. Numerical studies confirm conquer as a practical and reliable approach to large-scale inference for quantile regression. Software implementing the methodology is available in the R package `conquer`.

1. Introduction

Quantile regression (QR) is a useful statistical tool for modeling and inferring the relationship between a scalar response y and a p-dimensional predictor $\boldsymbol{x}$ (Koenker and Bassett, 1978). Compared to the least squares regression that focuses on modeling the conditional mean of y given $\boldsymbol{x}$, QR allows modeling of the entire conditional distribution of y given $\boldsymbol{x}$, and thus provides valuable insights into heterogeneity in the relationship between $\boldsymbol{x}$ and y. Moreover, quantile regression is robust against outliers and can be performed for skewed or heavy-tailed response distributions without correct specification of the likelihood. These advantages make QR an appealing method to explore data features that are invisible to the least squares regression. We refer to Koenker (2005) and Koenker et al. (2017) for an extensive overview of QR from methods, theory, computation, to various extensions under complex data structures.

Quantile regression involves a convex optimization problem with a piecewise linear loss function, also known as the *check function*. One can reformulate the QR problem as a linear program (LP), solvable by the Frisch–Newton algorithm with an average-case computational complexity that grows as a cubic function of p, i.e., $\mathcal{O}_{\mathbb{P}}(n^{1+\alpha}p^3 \log n)$ for some constant $\alpha \in (0, 1/2)$ (Portnoy and Koenker, 1997), where n is the sample size and p is the parametric dimension. However, when applied to large-scale problems—both n and p are large, QR computation via LP reformulation tends to be slow or too memory-intensive. To better appreciate such a challenge, we take the empirical study of U.S. equities from Gu et al. (2020) as an example. The dataset consists of monthly total individual equity returns, which begins in March 1957 and

* Corresponding author.
E-mail addresses: xmhe@umich.edu (X. He), xip024@ucsd.edu (X. Pan), keanming@umich.edu (K.M. Tan), wez243@ucsd.edu (W.-X. Zhou).

https://doi.org/10.1016/j.jeconom.2021.07.010
0304-4076/

ends in December 2016, from CRSP for all firms listed in the NYSE, AMEX, and NASDAQ. Within this span of 60 years, the average number of stocks considered is around 6200 per month. After processing, the number of observations (over 60 years) in the entire panel exceeds 4 million, and the number of stock-level covariates is 920. Using half of the data (1957–1986) as the training sample, the training sample size is still as large as 2 million. Even with preprocessing, the interior point QR solver in R (Koenker, 2019) may either run out of memory or take too much time on a personal computer. This shortcoming arguably makes QR less attractive compared to various machine learning tools. We refer to Chapter 5 of Koenker et al. (2017) for an overview of the prevailing computational methods for quantile regression, such as simplex-based algorithms (Barrodale and Roberts, 1974; Koenker and d'Orey, 1987), interior point methods (Portnoy and Koenker, 1997), and alternating direction method of multipliers, among other first-order proximal methods (Parikh and Boyd, 2014).

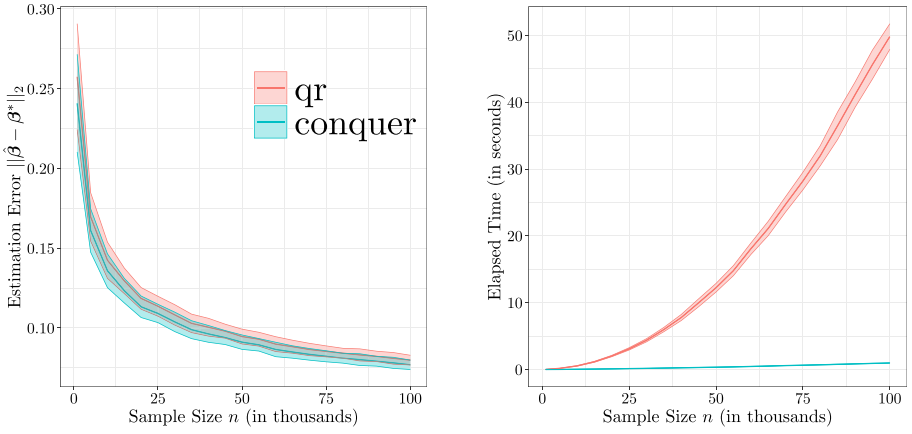
We consider conducting large-scale inference for quantile regression under the "increasing dimension" regime, namely, the dimension $p = p_n$ is subject to the growth condition $p \asymp n^a$ for some $a \in (0, 1)$. Two general principles have been widely used to suit this purpose. The first uses a nonparametric estimate of the asymptotic variance (Gutenbrunner and Jurečková, 1992) that involves the conditional density of the response given the covariates, yet such an estimate can be fairly unstable. Even if the asymptotic variance is well estimated, its approximation accuracy to the finite-sample variance depends on the design matrix and the quantile level. Resampling methods, on the other hand, provide a more reliable approach to inference for QR under a wide variety of settings (Parzen et al., 1994; He and Hu, 2002; Kocherginsky et al., 2005; Feng et al., 2011). Inevitably, the resampling approach requires repeatedly computing QR estimates up to thousands of times, and therefore is unduly expensive for large-scale data.

Theoretically, valid statistical inference is often justified by asymptotic normal approximations to QR estimators. The Bahadur–Kiefer representation of the nonlinear QR estimators are essential to this end, as shown in Arcones (1996) and He and Shao (1996). In large-p (non)asymptotic settings in which the parametric dimension p may tend to infinity with the sample size, we refer to Welsh (1989), He and Shao (2000), Belloni et al. (2019), and Pan and Zhou (2020) for normal approximation results of the QR estimators under fixed and random designs. The question of how large p can be relative to n to ensure asymptotic normality has been addressed by those authors. It is now recognized that we may have to pay a price here as compared to M-estimators with smooth loss functions that are at least twice continuously differentiable.

To circumvent the non-differentiability of the QR loss function, Horowitz (1998) proposed to smooth the indicator part of the check function via the survival function of a kernel. This smoothing method, which we refer to as *Horowitz's smoothing* throughout, has been widely used for various QR-related problems with complex data (Wang et al., 2012a; Wu et al., 2015; Galvao and Kato, 2016; de Castro et al., 2019; Chen et al., 2019). However, Horowitz's smoothing gains smoothness at the cost of convexity, which inevitably raises optimization-related issues. In general, computing a global minimum of a non-convex function is intractable: finding an ϵ-suboptimal point for a k-times continuously differentiable function $f : \mathbb{R}^p \to \mathbb{R}$ requires at least as many as $(1/\epsilon)^{p/k}$ evaluations of the function and its first k derivatives (Nemirovski and Yudin, 1983). As we shall see from the numerical studies in Section 5, the convergence of gradient-based algorithms can be relatively slow for high and low quantile levels. To address the aforementioned issue, Fernandes et al. (2021) proposed a convolution-type smoothing method that yields a convex and twice differentiable loss function, and studied the asymptotic properties of the smoothed estimator when p is fixed. To distinguish this approach from Horowitz's smoothing, we adopt the term *conquer* for convolution-type smoothed quantile regression.

In this paper, we first provide an in-depth statistical analysis of conquer under various nonstandard asymptotics settings in which p increases with n. Our results reveal a key feature of the smoothing parameter, often referred to as the bandwidth: the bandwidth adapts to both the sample size n and dimensionality p, so as to achieve a tradeoff between statistical accuracy and computational stability. Since the convolution smoothed loss function is globally convex and locally strongly convex, we propose an efficient gradient descent algorithm with the Barzilai–Borwein stepsize and a Huber-type initialization. The proposed algorithm is implemented via `RcppArmadillo` (Eddelbuettel and Sanderson, 2014) in the R package `conquer`. We next focus on large-scale statistical inference (hypothesis testing and confidence estimation) with large p and larger n. We propose a bootstrapped conquer method that has reduced computational complexity when the conquer estimator is used as initialization. Under appropriate restrictions on dimension, we establish the consistency (or concentration), Bahadur representation, asymptotic normality of the conquer estimator as well as the validity of the bootstrap approximation. In the following, we provide more details on the computational and statistical contributions of this paper.

Theoretically, by allowing p to grow with n, the 'complexity' of the function classes that we come across in the analysis also increases with n. Conventional asymptotic tools for proving the bootstrap validity are based on weak convergence arguments (van der Vaart and Wellner, 1996), which are not directly applicable in the increasing dimension setting, especially with a non-differentiable loss. In this paper we turn to a more refined and self-contained analysis, and prove a new local restricted strong convexity (RSC) property for the empirical smoothed quantile loss. This validates the key merit of convolution-type smoothing, i.e., local strong convexity. The smoothing method involves a bandwidth, denoted by h. Theoretically, we show that with sub-exponential random covariates (relaxing the bounded covariates assumption in Fernandes et al., 2021), conquer exhibits an ℓ_2-error of order $\sqrt{(p+t)/n} + h^2$ with probability at least $1 - 2e^{-t}$. When h is of order $\{(p + \log n)/n\}^\gamma$ for any $\gamma \in [1/4, 1/2]$, the conquer estimation is first-order equivalent to QR. Under slightly more stringent sub-Gaussian condition on the covariates, we show that the Bahadur–Kiefer linearization error of conquer is of order $(p + t)/(nh^{1/2}) + h^{3/2}\sqrt{(p + t)/n} + h^4$ with probability at least $1 - 3e^{-t}$. Based on such a representation, we establish a Berry–Esseen bound for linear functionals of conquer, which lays the theoretical foundation for testing general



(a) Estimation error (b) Runtime

Fig. 1. A numerical comparison between conquer and QR. The latter is implemented by the R package `quantreg` using the "pfn" method. Panels (a) and (b) display, respectively, the "estimation error and its standard deviation versus sample size" and "elapsed time and its standard deviation versus sample size" as the size of the problem increases.

linear hypotheses, encompassing covariate-effect analysis, analysis of variance, and model comparisons, to name a few. It is worth noting that with a properly chosen h, the linear functional of conquer is asymptotically normal as long as $p^{8/3}/n \to 0$, which improves the best known growth condition on p for standard QR (Welsh, 1989; He and Shao, 2000; Pan and Zhou, 2020). We attribute this gain to the effect of smoothing. Under similar conditions, we further establish upper bounds on both estimation and Bahadur–Kiefer linearization errors for the bootstrapped conquer estimator.

To better appreciate the computational feasibility of conquer for large-scale problems, we compare it with standard QR on large synthetic datasets, where the latter is implemented by the R package `quantreg` (Koenker, 2019) using the Frisch–Newton approach after preprocessing "pfn". We generate independent data vectors $\{y_i, \boldsymbol{x}_i\}_{i=1}^n$ from a linear model $y_i = \beta_0^* + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^* + \varepsilon_i$, where $(\beta_0^*, \boldsymbol{\beta}^{*\mathrm{T}})^{\mathrm{T}} = (1, \ldots, 1)^{\mathrm{T}} \in \mathbb{R}^{p+1}$, $\boldsymbol{x}_i \sim \mathcal{N}_p(0, \mathbf{I})$ and $\varepsilon_i \sim t_2$. We report the estimation error and elapsed time for increasing sample sizes $n \in \{1000, 5000, 10000, \ldots, 100000\}$ and dimension $p = \lfloor n^{1/2} \rfloor$, the largest integer that is less than or equal to $n^{1/2}$. Fig. 1 displays the average estimation error, average elapsed time and their standard deviations based on 100 Monte Carlo samples. This experiment shows promise of conquer as a practically useful tool for large-scale quantile regression analysis. More empirical evidence will be given in the latter section.

The rest of the paper is organized as follows. We start with a brief review of linear quantile regression and the convolution-type smoothing method in Section 2. Explicit forms of the smoothed check functions are provided for several representative kernel functions in nonparametric statistics. We introduce the multiplier bootstrap for statistical inference in Section 2.3. In Section 4, we provide a comprehensive theoretical study of conquer from a non-asymptotic viewpoint, which directly leads to array asymptotic results. Specifically, the bias incurred by smoothing the quantile loss is characterized in Section 4.1. In Section 4.2, we establish the rate of convergence, Bahadur–Kiefer representation, and Berry–Esseen bound for conquer in a large-p and larger-n regime. Results for its bootstrap counterpart are provided in Section 4.3. A Barzilai–Borwein gradient-based algorithm with a Huber-type warm start is detailed in Section 3. We conclude the paper with an extensive numerical study in Section 5 to illustrate the finite-sample performance of conquer in large-scale quantile regression analysis. We defer the proofs of all theoretical results as well as the full details of the one-step conquer to online supplementary materials.

Notation. For every integer $k \geq 1$, we use $\mathbb{R}^k$ to denote the k-dimensional Euclidean space. The inner product of any two vectors $\boldsymbol{u} = (u_1, \ldots, u_k)^{\mathrm{T}}$, $\boldsymbol{v} = (v_1, \ldots, v_k)^{\mathrm{T}} \in \mathbb{R}^k$ is defined by $\boldsymbol{u}^{\mathrm{T}}\boldsymbol{v} = \langle \boldsymbol{u}, \boldsymbol{v} \rangle = \sum_{i=1}^k u_i v_i$. We use $\|\cdot\|_p$ $(1 \leq p \leq \infty)$ to denote the ℓ_p-norm in $\mathbb{R}^k$: $\|\boldsymbol{u}\|_p = (\sum_{i=1}^k |u_i|^p)^{1/p}$ and $\|\boldsymbol{u}\|_\infty = \max_{1 \leq i \leq k} |u_i|$. Throughout this paper, we use bold capital letters to represent matrices. For $k \geq 2$, $\mathbf{I}_k$ represents the identity matrix of size k. For any $k \times k$ symmetric matrix $\mathbf{A} \in \mathbb{R}^{k \times k}$, $\|\mathbf{A}\|_2$ denotes the operator norm of $\mathbf{A}$. If $\mathbf{A}$ is positive semidefinite, we use $\|\cdot\|_{\mathbf{A}}$ to denote the vector norm linked to $\mathbf{A}$ given by $\|\boldsymbol{u}\|_{\mathbf{A}} = \|\mathbf{A}^{1/2}\boldsymbol{u}\|_2$, $\boldsymbol{u} \in \mathbb{R}^k$. For $r \geq 0$, define the Euclidean ball and unit sphere in $\mathbb{R}^k$ as $\mathbb{B}^k(r) = \{\boldsymbol{u} \in \mathbb{R}^k : \|\boldsymbol{u}\|_2 \leq r\}$ and $\mathbb{S}^{k-1} = \partial\mathbb{B}^k(1) = \{\boldsymbol{u} \in \mathbb{R}^k : \|\boldsymbol{u}\|_2 = 1\}$, respectively. For two sequences of non-negative numbers $\{a_n\}_{n \geq 1}$ and $\{b_n\}_{n \geq 1}$, $a_n \lesssim b_n$ indicates that there exists a constant $C > 0$ independent of n such that $a_n \leq C b_n$; $a_n \gtrsim b_n$ is equivalent to $b_n \lesssim a_n$; $a_n \asymp b_n$ is equivalent to $a_n \lesssim b_n$ and $b_n \lesssim a_n$.

2. Smoothed quantile regression

2.1. The linear quantile regression model

Given a univariate response variable $y \in \mathbb{R}$ and a p-dimensional covariate vector $\boldsymbol{x} = (x_1, \ldots, x_p)^{\mathrm{T}} \in \mathbb{R}^p$ with $x_1 \equiv 1$, the primary goal here is to learn the effect of $\boldsymbol{x}$ on the distribution of y. Let $F_{y|\boldsymbol{x}}(\cdot)$ be the conditional distribution function of y given $\boldsymbol{x}$. The dependence between y and $\boldsymbol{x}$ is then fully characterized by the conditional quantile functions of y given $\boldsymbol{x}$, denoted as $F^{-1}_{y|\boldsymbol{x}}(\tau)$, for $0 < \tau < 1$. We consider a linear quantile regression model at a given $\tau \in (0, 1)$, that is, the τth conditional quantile function is

$$F^{-1}_{y|\boldsymbol{x}}(\tau) = \langle \boldsymbol{x}, \boldsymbol{\beta}^*(\tau) \rangle, \tag{2.1}$$

where $\boldsymbol{\beta}^*(\tau) = (\beta^*_1(\tau), \ldots, \beta^*_p(\tau))^{\mathrm{T}} \in \mathbb{R}^p$ is the true quantile regression coefficient.

Let $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$ be a random sample from $(y, \boldsymbol{x})$. The standard quantile regression estimator (Koenker and Bassett, 1978) is then given as

$$\widehat{\boldsymbol{\beta}}(\tau) \in \min_{\boldsymbol{\beta} \in \mathbb{R}^p} \widehat{Q}(\boldsymbol{\beta}) = \min_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n \rho_\tau(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta} \rangle), \tag{2.2}$$

where $\rho_\tau(u) = u\{\tau - \mathbb{1}(u < 0)\}$ is the τ-quantile loss function, also referred to as the check function. Statistical properties of $\widehat{\boldsymbol{\beta}}(\tau)$ have been extensively studied. We refer the reader to Koenker (2005) and Koenker et al. (2017) for more details.

2.2. Smoothed estimation equation and convolution-type smoothing

Let $Q(\boldsymbol{\beta}) = \mathbb{E}\{\widehat{Q}(\boldsymbol{\beta})\}$ be the population quantile loss function. Under mild conditions, $Q(\cdot)$ is twice differentiable and strongly convex in a neighborhood of $\boldsymbol{\beta}^*$ with Hessian matrix $\mathbf{J} := \nabla^2 Q(\boldsymbol{\beta}^*) = \mathbb{E}\{f_{\varepsilon|\boldsymbol{x}}(0)\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\}$, where $\varepsilon = y - \langle \boldsymbol{x}, \boldsymbol{\beta}^*(\tau) \rangle$ is the random noise and $f_{\varepsilon|\boldsymbol{x}}(\cdot)$ is the conditional density of ε given $\boldsymbol{x}$. In contrast, the empirical quantile loss $\widehat{Q}(\cdot)$ is not differentiable at $\boldsymbol{\beta}^*$, and its "curvature energy" is concentrated at a single point. This is substantially different from other widely used loss functions that are at least locally strongly convex, such as the squared or logistic loss. The non-smoothness property not only brings challenge to theoretical analysis, but more importantly, also prevents gradient-based optimization methods from being efficient. In his seminal work, Horowitz (1998) proposed to directly smooth the check function $\rho_\tau(\cdot)$ to obtain

$$\ell_h^{\mathrm{Horo}}(u) = u\big\{\tau - \mathcal{G}(-u/h)\big\}, \tag{2.3}$$

where $\mathcal{G}(\cdot)$ is a smooth function that takes values between 0 and 1, and $h > 0$ is a smoothing parameter/bandwidth. However, Horowitz's smoothing gains smoothness at the cost of convexity, which inevitably raises optimization issues especially when p is large. On the other hand, by the first-order condition, the population parameter $\boldsymbol{\beta}^*$ satisfies the moment condition

$$\nabla Q(\boldsymbol{\beta}^*) = \mathbb{E}\left[\big\{\mathbb{1}(y < \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}) - \tau\big\}\boldsymbol{x}\right]\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} = \mathbf{0}.$$

This property motivates a smoothed estimating equation (SEE) estimator (Whang, 2006; Kaplan and Sun, 2017), defined as the solution to the smoothed moment condition

$$\frac{1}{n} \sum_{i=1}^n \Big[\mathcal{G}\big\{(\langle \boldsymbol{x}_i, \boldsymbol{\beta} \rangle - y_i)/h\big\} - \tau\Big]\boldsymbol{x}_i = \mathbf{0}. \tag{2.4}$$

Let $K(\cdot)$ be a kernel function that integrates to one, and $h > 0$ be a bandwidth. Throughout the paper, we write

$$K_h(u) = h^{-1}K(u/h), \quad \mathcal{K}_h(u) = \mathcal{K}(u/h) \ \text{ and } \ \mathcal{K}(u) = \int_{-\infty}^u K(v)\,\mathrm{d}v, \quad u \in \mathbb{R}. \tag{2.5}$$

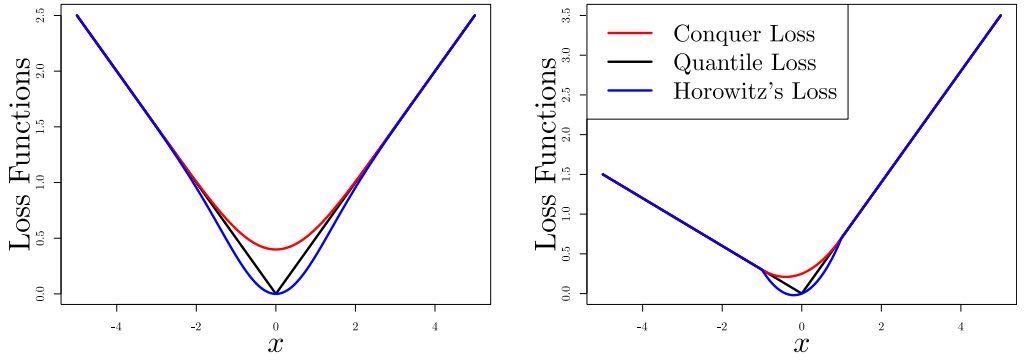
From an M-estimation viewpoint, the aforementioned SEE estimator can be equivalently defined as a minimizer of the empirical smoothed loss function

$$\widehat{Q}_h(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^n \ell_h(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta} \rangle) \quad \text{with} \quad \ell_h(u) = (\rho_\tau * K_h)(u) = \int_{-\infty}^{\infty} \rho_\tau(v) K_h(v - u)\,\mathrm{d}v, \tag{2.6}$$

where $*$ denotes the convolution operator. Therefore, as stated in the Introduction, we refer to the aforementioned smoothing method as *conquer*. The ensuing conquer estimator is given by

$$\widehat{\boldsymbol{\beta}}_h = \widehat{\boldsymbol{\beta}}_h(\tau) \in \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} \widehat{Q}_h(\boldsymbol{\beta}). \tag{2.7}$$

The key difference between the conquer loss (2.6) and Horowitz's loss (2.3) is that the former is globally convex, while Horowitz's loss is not. This is illustrated in Fig. 2.



(a) Guassian kernel under $\tau = 0.5$. (b) Uniform kernel under $\tau = 0.7$.

Fig. 2. Visualization of the quantile loss in (2.2), conquer loss (2.6), and Horowitz's smoothed loss (2.3) with Gaussian and uniform kernels, respectively.

As we shall see later, the ideal choice of bandwidth should adapt to the sample size n and dimension p, since the quantile level τ is prespecified and fixed. Thus, the dependence of $\widehat{\boldsymbol{\beta}}_h$ and $\widehat{Q}_h(\cdot)$ on τ will be assumed without display. Commonly used kernel functions include: (a) uniform kernel $K(u) = (1/2)\mathbb{1}(|u| \le 1)$, (b) Gaussian kernel $K(u) = \phi(u) := (2\pi)^{-1/2}e^{-u^2/2}$, (c) logistic kernel $K(u) = e^{-u}/(1+e^{-u})^2$, (d) Epanechnikov kernel $K(u) = (3/4)(1-u^2)\mathbb{1}(|u| \le 1)$, and (e) triangular kernel $K(u) = (1-|u|)\mathbb{1}(|u| \le 1)$. Explicit expressions of the corresponding smoothed loss function $\rho_\tau * K_h$ will be given in Section 3.

The convolution-type kernel smoothing yields an objective function $\boldsymbol{\beta} \mapsto \widehat{Q}_h(\boldsymbol{\beta})$ that is twice continuously differentiable with gradient and hessian matrix

$$\nabla\widehat{Q}_h(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\big\{\mathcal{K}_h\big(\langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle - y_i\big) - \tau\big\}\boldsymbol{x}_i \quad \text{and} \quad \nabla^2\widehat{Q}_h(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} K_h(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle)\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}, \tag{2.8}$$

respectively, where $\mathcal{K}_h(\cdot) = \mathcal{K}(\cdot/h)$ is defined in (2.5). Provided that K is non-negative, $\widehat{Q}_h(\cdot)$ is a convex function for any $h > 0$, and $\widehat{\boldsymbol{\beta}}_h = \widehat{\boldsymbol{\beta}}_h(\tau)$ satisfies the first-order condition $\nabla\widehat{Q}_h(\widehat{\boldsymbol{\beta}}_h) = \mathbf{0}$. This reveals the connection between the SEE and the conquer methods. Together, the smoothness and convexity of $\widehat{Q}_h(\cdot)$ warrant the superior computation efficiency of first-order gradient based algorithms for solving large-scale smoothed quantile regressions. The computational aspect of conquer will be discussed in Section 3.

When the dimension p is fixed, asymptotic properties of the SEE or conquer estimator have been studied by Kaplan and Sun (2017) and Fernandes et al. (2021). The former used a higher-order kernel to deal with the instrumental variables QR problem (see Section 2.4 for further discussions), and the latter showed that the conquer estimator has a lower asymptotic mean squared error than Horowitz's smoothed estimator, and also has a smaller Bahadur linearization error than the standard QR in the almost sure sense. The optimal order of the bandwidth based on the asymptotic mean squared error is unveiled as a function of n. In Section 4, we will establish exponential concentration inequalities and non-asymptotic Bahadur representation for the conquer estimator, while allowing the dimension p to grow with the sample size n. Our results reveal a key feature of the smoothing parameter: the bandwidth should adapt to both the sample size n and dimensionality p, so as to achieve a tradeoff between statistical accuracy and computational stability.

Remark 2.1. As discussed in Fernandes et al. (2021), another advantage of convolution smoothing is that it facilitates conditional density estimation for the quantile regression process. Assume $Q_y(\tau|\boldsymbol{x}) = F^{-1}_{y|\boldsymbol{x}}(\tau) = \langle \boldsymbol{x}, \boldsymbol{\beta}^*(\tau)\rangle$ for all $\tau \in [\tau_L, \tau_U] \subseteq (0,1)$. Under mild regularity conditions, $q_y(\tau|\boldsymbol{x}) := \partial Q_y(\tau|\boldsymbol{x})/\partial\tau = 1/f_{y|\boldsymbol{x}}(\langle \boldsymbol{x}, \boldsymbol{\beta}^*(\tau)\rangle)$ exists. The inverse conditional density function plays an important role in, for example, the study of quantile treatment effects through modeling inverse propensity scores (Firpo, 2007; Chen et al., 2008). By the linear conditional quantile model assumption, $\partial Q_y(\tau|\boldsymbol{x})/\partial\tau = \langle \boldsymbol{x}, \partial\boldsymbol{\beta}^*(\tau)/\partial\tau\rangle$ for $\tau \in (\tau_L, \tau_U)$. Recall that the conquer estimator $\widehat{\boldsymbol{\beta}}_h = \widehat{\boldsymbol{\beta}}_h(\tau)$ satisfies the first-order condition $\nabla\widehat{Q}_h(\widehat{\boldsymbol{\beta}}_h(\tau)) = 0$. Taking the partial derivative with respect to τ on both sides, it follows from (2.8) and the chain rule that

$$\frac{\partial\widehat{\boldsymbol{\beta}}_h(\tau)}{\partial\tau} = \big\{\nabla^2\widehat{Q}_h\big(\widehat{\boldsymbol{\beta}}_h(\tau)\big)\big\}^{-1}\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i = \left\{\frac{1}{n}\sum_{i=1}^{n}K_h\big(y_i - \boldsymbol{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_h(\tau)\big)\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\right\}^{-1}\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i.$$

Consequently, the inverse densities $1/f_{y_i|\boldsymbol{x}_i}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*(\tau))$ can be directly estimated by $\boldsymbol{x}_i^{\mathrm{T}}\frac{\partial\widehat{\boldsymbol{\beta}}_h(\tau)}{\partial\tau}$. This bypasses the use of any nonparametric method for density estimation with fitted residuals.

2.3. Multiplier bootstrap inference

In this section, we consider a multiplier bootstrap procedure to construct confidence intervals for conquer. Independent of the observed sample $\mathcal{X}_n = \{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$, let $\{w_i\}_{i=1}^n$ be independent and identically distributed random variables with $\mathbb{E}(w_i) = 1$ and $\mathrm{var}(w_i) = 1$. Recall that $\widehat{\boldsymbol{\beta}}_h = \widehat{\boldsymbol{\beta}}_h(\tau) = \min_{\boldsymbol{\beta}\in\mathbb{R}^p} \widehat{Q}_h(\boldsymbol{\beta})$ is the conquer estimator. If the minimizer is not unique, we take any of the minima as $\widehat{\boldsymbol{\beta}}_h = (\widehat{\beta}_{h,1}, \ldots, \widehat{\beta}_{h,p})^{\mathrm{T}}$.

The proposed bootstrap method, which dates back to Dudewicz (1992) and Barbe and Bertail (1995), is based on reweighting the summands of $\widehat{Q}_h(\cdot)$ with random weights w_i. More specifically, define the weighted quantile loss $\widehat{Q}_h^{\flat} : \mathbb{R}^p \to \mathbb{R}$ as

$$\widehat{Q}_h^{\flat}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n w_i \ell_h(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle), \tag{2.9}$$

where $\ell_h(u) = (\rho_\tau * K_h)(u)$ is as in (2.6). The ensuing multiplier bootstrap statistic is then given by

$$\widehat{\boldsymbol{\beta}}_h^{\flat} = \widehat{\boldsymbol{\beta}}_h^{\flat}(\tau) \in \operatorname*{argmin}_{\boldsymbol{\beta}\in\Theta} \widehat{Q}_h^{\flat}(\boldsymbol{\beta}), \tag{2.10}$$

where Θ is a predetermined subset of $\mathbb{R}^p$. If the random weights are allowed to take negative values, the weighted quantile loss may be non-convex. We therefore take Θ as a compact subset, such as $\Theta = \{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_h\|_2 \le R\}$ for some $R \ge 1$, in order to guarantee the existence of local/global optima.

Let $\mathbb{E}^*$ and $\mathbb{P}^*$ be the conditional expectation and probability given the observed data $\mathcal{X}_n$, respectively. Observe that $\mathbb{E}^*\{\widehat{Q}_h^{\flat}(\boldsymbol{\beta})\} = \widehat{Q}_h(\boldsymbol{\beta})$ for any $\boldsymbol{\beta} \in \mathbb{R}^p$. Consequently, we have

$$\operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \mathbb{E}^*\{\widehat{Q}_h^{\flat}(\boldsymbol{\beta})\} = \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \widehat{Q}_h(\boldsymbol{\beta}) = \widehat{\boldsymbol{\beta}}_h.$$

Intuitively, this means that $\widehat{\boldsymbol{\beta}}_h^{\flat}$ is an M-estimator of $\widehat{\boldsymbol{\beta}}_h$ in the bootstrap world. Since $\widehat{\boldsymbol{\beta}}_h$ is also an (approximate) M-estimator of $\boldsymbol{\beta}^*$, we expect that the distribution of $\widehat{\boldsymbol{\beta}}_h - \boldsymbol{\beta}^*$ can be well approximated with high probability by the conditional distribution of $\widehat{\boldsymbol{\beta}}_h^{\flat} - \widehat{\boldsymbol{\beta}}_h$. We will establish the validity of this approximation with explicit rates in Section 4.3. We refer to Chatterjee and Bose (2005) for a general asymptotic theory for weighted bootstrap for estimating equations, where a class of bootstrap weights is considered. Extensions to semiparametric M-estimation can be found in Ma and Kosorok (2005) and Cheng and Huang (2010).

To retain convexity of the loss function, non-negative random weights are preferred, such as $w_i \sim \mathrm{Exp}(1)$, i.e., exponential distribution with rate 1, and $w_i = 1 + e_i$, where e_i are independent Rademacher random variables. To compute the bootstrap estimate $\widehat{\boldsymbol{\beta}}_h^{\flat}$, we use the same gradient-based algorithm described in Section 3, which will have a faster convergence rate thanks to the (provably) good initialization $\widehat{\boldsymbol{\beta}}_h$ and the strong convexity of the smoothed loss in a neighborhood of $\widehat{\boldsymbol{\beta}}_h$.

We can construct confidence intervals based on the bootstrap estimates using one of the three classical methods, the percentile method, the pivotal method, and the normal-based method. To be specific, for each $q \in (0, 1)$ and $1 \le j \le p$, define the (conditional) q-quantile of $\widehat{\beta}_{h,j}^{\flat}$ – the jth coordinate of $\widehat{\boldsymbol{\beta}}_h^{\flat} \in \mathbb{R}^p$ – given the observed data as $\mathrm{c}_j^{\flat}(q) = \inf\{t \in \mathbb{R} : \mathbb{P}^*(\widehat{\beta}_{h,j}^{\flat} \le t) \ge q\}$. Then, for a prespecified nominal level $\alpha \in (0, 1)$, the corresponding $1-\alpha$ bootstrap percentile and pivotal confidence intervals (CIs) for β_j^* $(j = 1, \ldots, p)$ are, respectively,

$$\left[\mathrm{c}_j^{\flat}(\alpha/2),\ \mathrm{c}_j^{\flat}(1-\alpha/2)\right] \quad \text{and} \quad \left[2\widehat{\beta}_{h,j} - \mathrm{c}_j^{\flat}(1-\alpha/2),\ 2\widehat{\beta}_{h,j} - \mathrm{c}_j^{\flat}(\alpha/2)\right].$$

Numerically, $\mathrm{c}_j^{\flat}(q)$ $(q \in \{\alpha, 1-\alpha/2\})$ can be calculated with any specified precision by the simulation. In the R package `conquer`, the default number of bootstrap replications is set as $B = 1000$.

In the next section, we will present a finite-sample theoretical framework for convolution-type smoothed quantile regression, including the concentration inequality and non-asymptotic Bahadur representation for both the conquer estimator (2.7) and its bootstrap counterpart (2.10) using Rademacher multipliers. As a by-product, a Berry–Esseen-type inequality (see Theorem 4.3) states that, under certain constraints on the (growing) dimensionality and bandwidth, the distribution of any linear projection of $\widehat{\boldsymbol{\beta}}_h$ converges to a normal distribution as the sample size increases to infinity. Informally, for any given deterministic vector $\boldsymbol{a} \in \mathbb{R}^p$, the scaled statistic $n^{1/2}\langle \boldsymbol{a}, \widehat{\boldsymbol{\beta}}_h - \boldsymbol{\beta}^*\rangle$ is asymptotically normally distributed with asymptotic variance $\sigma_0^2(\boldsymbol{a}) := \tau(1-\tau)\boldsymbol{a}^{\mathrm{T}}\mathbf{J}^{-1}\boldsymbol{\Sigma}\mathbf{J}^{-1}\boldsymbol{a}$, where $\boldsymbol{\Sigma}$ is the population covariance matrix of the covariates $\boldsymbol{x}$. Another interesting implication from our theoretical analysis is that the unit variance requirement $\mathrm{var}(w_i) = 1$ for the random weight is not necessary to ensure (asymptotically) valid bootstrap inference after a proper variance adjustment. See Remark 4.7 for more details.

To make inference based on such asymptotic results, we need to consistently estimate the asymptotic variance. Fernandes et al. (2021) suggested the following estimators

$$\widehat{\mathbf{J}}_h := \nabla^2 \widehat{\mathcal{Q}}_h(\widehat{\boldsymbol{\beta}}_h) = \frac{1}{nh}\sum_{i=1}^n K(\widehat{\varepsilon}_i/h)\cdot \boldsymbol{x}\boldsymbol{x}_i^{\mathrm{T}} \quad \text{and} \quad \widehat{\mathbf{V}}_h := \frac{1}{n}\sum_{i=1}^n \left\{\mathcal{K}_h(-\widehat{\varepsilon}_i) - \tau\right\}^2 \boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}} \tag{2.11}$$

of $\mathbf{J}$ and $\tau(1-\tau)\boldsymbol{\Sigma}$, respectively, where $\widehat{\varepsilon}_i = y_i - \langle \boldsymbol{x}_i, \widehat{\boldsymbol{\beta}}_h \rangle$ are fitted residuals. The ensuing $1-\alpha$ normal-based CIs are given by $\widehat{\beta}_{h,j} \pm \Phi^{-1}(1-\alpha/2) \cdot n^{-1/2} (\widehat{\mathbf{J}}_h^{-1} \widehat{\mathbf{V}}_h \widehat{\mathbf{J}}_h^{-1})_{jj}^{1/2}$, $j = 1, \ldots, p$. The normal approximations to the CI may suffer from the sensitivity to the smoothing needed to estimate the conditional densities, namely, the matrix $\mathbf{J} = \mathbb{E}\{f_{\varepsilon|\boldsymbol{x}}(0)\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\}$. When p is large, inverting the estimated density matrix $\widehat{\mathbf{J}}_h$ may be numerically unstable. This is typically true when τ is in the upper or lower tail. See Section 5.3 for a numerical comparison between normal approximation and bootstrap calibration for confidence construction at high and low quantile levels. As we shall see, the normal-based CIs can be exceedingly wide and thus inaccurate under this scenario.

2.4. Connections to instrumental variable quantile regression

This work focuses on large-scale estimation and inference for linear quantile regression with many exogenous covariates. However, in many economic applications, some regressors of interest (e.g., education, prices) are endogenous, making conventional quantile regression inconsistent for estimating causal quantile effects. To address this problem, Chernozhukov and Hansen (2005) proposed an instrumental variable quantile regression (IVQR) model, which has become a popular tool for estimating quantile effects with endogenous covariates. Due to the non-convex and non-smooth nature of the problem, there is a burgeoning literature on estimation and inference of IVQR models and the related computational issues, dating back to Chen et al. (2003) and Chernozhukov and Hansen (2006). We refer to Chernozhukov et al. (2020) – Chapter 9 of Koenker et al. (2017) – for an overview of IVQR modeling, from identification conditions to estimation and inference. More specifically, see Horowitz and Lee (2007) and Chen and Pouzo (2009) for non- and semi-parametric IVQR estimation; Kaplan and Sun (2017) and de Castro et al. (2019) for smoothed methods; and Chen and Lee (2018) and Zhu (2018) for methods based on reformulation as mixed integer optimization (MIO), Machado and Santos Silva (2018) for moment-based estimators, and Kaido and Wüthrich (2021) for a decentralization approach which decomposes the IVQR estimation problem into a set of conventional QR sub-problems.

The convolution smoothing method studied in this paper can be directly linked to the SEE approach in Kaplan and Sun (2017). The latter addressed the more challenging IVQR problem, and derived both asymptotic mean squared error and normality for the SEE estimator when the dimension is fixed. Our study complements that of Kaplan and Sun (2017) in two ways. First, we provide a systematic analysis for smoothed (conventional) QR from an M-estimation viewpoint under the growing dimension setting. Our results provide explicit finite-sample bounds for the estimation error, Bahadur linearization error as long as their (multiplier) bootstrap counterparts. Asymptotic validity of the multiplier bootstrap is also rigorously established. Secondly, we propose tailored computational methods for smoothed QR computation, which rely on the use of non-negative kernels and the resulting local strong convexity. Compared with generic optimization toolboxes for solving linear programs, the computational efficiency of the gradient-based algorithm for conquer is considerably improved, especially for large-scale problems with many (exogenous) regressors and massive sample size. A potential application is empirical asset pricing via quantile regression, extending the existing machine learning tools for average return forecasting (Gu et al., 2020).

In the presence of both exogenous and endogenous covariates, the advantage of smoothing is diluted because the non-convexity issue prevails. The MIO-based IVQR estimation procedure can be implemented by the Gurobi commercial MIO solver, which is free for academic use. The MIO solver converges fast when the number of endogenous covariates varies in the range of 5 and 20 (Zhu, 2018). The MIO solver in moderate dimensions typically takes much longer to complete the optimization: optimal solutions may be found in a few seconds, but it can take much longer to certify optimality via the lower bounds (Bertsimas et al., 2016).[1]

Recently, Kaido and Wüthrich (2021) proposed a "decentralized" approach for IVQR estimation. The idea is to decompose the non-convex program into $p_d + 1$ conventional (weighted) quantile regression sub-problems. The IVQR estimator is then characterized as a fixed point of such sub-problems. Since p_d – the number of endogenous variables – is typically small, the overall computational complexity depends primarily on the QR fitting step. When the number of exogenous variables, p_x, is large in the range of hundreds to thousands, the proposed framework in this paper, along with the accompanying software `conquer`, provides a viable option to further reduce the computational cost of the above IVQR estimation method. We leave a rigorous theoretical investigation (when p_d is fixed, $p_x = p_x(n) \to \infty$ and $p_x/n \to 0$ as $n \to \infty$) as well as empirical applications with many (exogenous) regressors to future work.

3. Computational methods for conquer

To solve optimization problems (2.7) and (2.10) with non-negative weights, arguably the simplest algorithm is a vanilla gradient descent algorithm (GD). For a prespecified $\tau \in (0,1)$ and bandwidth $h > 0$, recall that $\widehat{Q}_h(\boldsymbol{\beta}) = (1/n)\sum_{i=1}^n \ell_h(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta} \rangle)$. Starting with an initial value $\boldsymbol{\beta}^0 \in \mathbb{R}^p$, at iteration $t = 0, 1, 2, \ldots$, GD computes

$$\boldsymbol{\beta}^{t+1} = \boldsymbol{\beta}^t - \eta_t \cdot \nabla \widehat{Q}_h(\boldsymbol{\beta}^t) = \boldsymbol{\beta}^t - \frac{\eta_t}{n} \sum_{i=1}^n \big\{ \mathcal{K}_h(\langle \boldsymbol{x}_i, \boldsymbol{\beta}^t \rangle - y_i) - \tau \big\} \boldsymbol{x}_i, \tag{3.1}$$

[1] MIO solvers provide both feasible solutions and lower bounds to the optimal value. As the MIO solver progresses toward the optimal solution, the lower bounds improve and provide an increasingly better guarantee of suboptimality. It is the lower bounds that take so long to converge.

where $\eta_t > 0$ is the stepsize. In the classical GD method, the stepsize is usually obtained by employing line search techniques. However, line search is computationally intensive for large-scale settings. One of the most important issues in GD is to determine a proper update step η_t decay schedule. A common practice in the literature is to use a diminishing stepsize or a best-tuned fixed stepsize. Neither of these two approaches can be efficient, at least compared to the Newton–Frisch algorithm with preprocessing (Portnoy and Koenker, 1997). Recall that the smoothed loss $\widehat{Q}_h(\cdot)$ is twice differentiable with Hessian $\nabla^2\widehat{Q}_h(\boldsymbol{\beta}) = (1/n)\sum_{i=1}^n K_h(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta}\rangle)\boldsymbol{x}_i\boldsymbol{x}_i^\top$. It is therefore natural to employ the Newton–Raphson method, which at iteration t would read

$$\boldsymbol{\beta}^{t+1} = \boldsymbol{\beta}^t + \boldsymbol{d}^t \quad \text{with} \quad \boldsymbol{d}^t := -\big\{\nabla^2\widehat{Q}_h(\boldsymbol{\beta}^t)\big\}^{-1}\nabla\widehat{Q}_h(\boldsymbol{\beta}^t). \tag{3.2}$$

In practice, the Newton method is often paired with Armoji stepsize: choose a stepsize $\lambda^t = \max\{1, 1/2, 1/4, \ldots\}$ such that $\widehat{Q}_h(\boldsymbol{\beta}^t) - \widehat{Q}_h(\boldsymbol{\beta}^t + \lambda^t\boldsymbol{d}^t) \geq -c\lambda^t\nabla\widehat{Q}_h(\boldsymbol{\beta}^t)\boldsymbol{d}^t$, where $c \in (0, 1/2)$. Then redefine the current iterate as $\boldsymbol{\beta}^{t+1} = \boldsymbol{\beta}^t + \lambda^t\boldsymbol{d}^t$. Since such a backtracking line search requires evaluations of the loss function itself, in the following remark we present the explicit expressions of the convolution smoothed check function for several commonly used kernels.

Remark 3.1. Recall that the check function can be written as $\rho_\tau(u) = |u|/2 + (\tau - 1/2)u$, which, after convolution smoothing, becomes $\ell_h(u) = (1/2)\int_{-\infty}^{\infty}|u + hv|K(v)\,\mathrm{d}v + (\tau - 1/2)u$.

- (Gaussian kernel $K(u) = (2\pi)^{-1/2}e^{-u^2/2}$): $\ell_h(u) = (h/2)\ell^{\mathrm{G}}(u/h) + (\tau - 1/2)u$, where $\ell^{\mathrm{G}}(u) := (2/\pi)^{1/2}e^{-u^2/2} + u\{1 - 2\Phi(-u)\}$.
- (Logistic kernel[2] $K(u) = e^{-u}/(1 + e^{-u})^2$): $\ell_h(u) = (h/2)\ell^{\mathrm{L}}(u/h) + (\tau - 1/2)u$, where $\ell^{\mathrm{L}}(u) := u + 2\log(1 + e^{-u})$.
- (Uniform kernel $K(u) = (1/2)\mathbb{1}(|u| \leq 1)$): $\ell_h(u) = (h/2)\ell^{\mathrm{U}}(u/h) + (\tau - 1/2)u$, where $\ell^{\mathrm{U}}(u) := (u^2/2 + 1/2)\mathbb{1}(|u| \leq 1) + |u|\mathbb{1}(|u| > 1)$ is a shifted Huber loss (Huber, 1973).
- (Epanechnikov kernel $K(u) = (3/4)(1 - u^2)\mathbb{1}(|u| \leq 1)$): $\ell_h(u) = (h/2)\ell^{\mathrm{E}}(u/h) + (\tau - 1/2)u$, where $\ell^{\mathrm{E}}(u) := (3u^2/4 - u^4/8 + 3/8)\mathbb{1}(|u| \leq 1) + |u|\mathbb{1}(|u| > 1)$.
- (Triangular kernel $K(u) = (1 - |u|)\mathbb{1}(|u| \leq 1)$): $\ell_h(u) = (h/2)\ell^{\mathrm{T}}(u/h) + (\tau - 1/2)u$, where $\ell^{\mathrm{T}}(u) := (u^2 - |u|^3/3 + 1/3)\mathbb{1}(|u| \leq 1) + |u|\mathbb{1}(|u| > 1)$.

3.1. The Barzilai–Borwein stepsize

In this section, we propose to solve conquer by means of the gradient descent with a Barzilai–Borwein update step (Barzilai and Borwein, 1988), which we refer to as the GD-BB algorithm. Motivated by quasi-Newton methods, the BB method has been proven to be very successful in solving nonlinear optimization problems.

Computing the inverse of the Hessian when p is large is an expensive operation at each Newton step (3.2). Moreover, in circumstances where h is small or τ is very close to 0 or 1, $\nabla^2\widehat{Q}_h(\cdot)$ may have a large condition number, thus leading to slow convergence. For this reason, many quasi-Newton methods seek a simple approximation of the inverse Hessian matrix, say $(\mathbf{J}^t)^{-1}$, satisfying the secant equation $\mathbf{J}^t\boldsymbol{\delta}^t = \boldsymbol{g}^t$, where

$$\boldsymbol{\delta}^t = \boldsymbol{\beta}^t - \boldsymbol{\beta}^{t-1} \quad \text{and} \quad \boldsymbol{g}^t = \nabla\widehat{Q}_h(\boldsymbol{\beta}^t) - \nabla\widehat{Q}_h(\boldsymbol{\beta}^{t-1}), \quad t = 1, 2, \ldots. \tag{3.3}$$

To mitigate the computational cost of inverting a large matrix, the BB method chooses η so that $\eta\nabla\widehat{Q}_h(\boldsymbol{\beta}^t) = (\eta^{-1}\mathbf{I}_p)^{-1}\nabla\widehat{Q}_h(\boldsymbol{\beta}^t)$ "approximates" $(\mathbf{J}^t)^{-1}\nabla\widehat{Q}_h(\boldsymbol{\beta}^t)$. Since $\mathbf{J}^t$ satisfies $\mathbf{J}^t\boldsymbol{\delta}^t = \boldsymbol{g}^t$, it is more practical to choose η such that $(1/\eta)\boldsymbol{\delta}^t \approx \boldsymbol{g}^t$ or $\boldsymbol{\delta}^t \approx \eta\boldsymbol{g}^t$. Via least squares approximations, one may use $\eta_{1,t}^{-1} = \operatorname{argmin}_\alpha \|\alpha\boldsymbol{\delta}^t - \boldsymbol{g}^t\|_2^2$ or $\eta_{2,t} = \operatorname{argmin}_\eta \|\boldsymbol{\delta}^t - \eta\boldsymbol{g}^t\|_2^2$. The BB step sizes are then defined as

$$\eta_{1,t} = \frac{\langle\boldsymbol{\delta}^t, \boldsymbol{\delta}^t\rangle}{\langle\boldsymbol{\delta}^t, \boldsymbol{g}^t\rangle} \quad \text{and} \quad \eta_{2,t} = \frac{\langle\boldsymbol{\delta}^t, \boldsymbol{g}^t\rangle}{\langle\boldsymbol{g}^t, \boldsymbol{g}^t\rangle}. \tag{3.4}$$

Consequently, the BB iteration takes the form

$$\boldsymbol{\beta}^{t+1} = \boldsymbol{\beta}^t - \eta_{\ell,t}\nabla\widehat{Q}_h(\boldsymbol{\beta}^t), \quad \ell = 1 \text{ or } 2. \tag{3.5}$$

Note that the BB step starts at iteration 1, while at iteration 0, we compute $\boldsymbol{\beta}^1$ using standard gradient descent with an initial estimate $\boldsymbol{\beta}^0$. The procedure is summarized in Algorithm 1. Based on extensive numerical studies, we find that at a fixed τ, the number of iterations is insensitive to varying (n, p) combinations. Moreover, as h increases, the number of iterations declines because the loss function is "more convex" for larger h. In Algorithm 1, the quantity $\delta > 0$ is a prespecified tolerance level, ensuring that the final iterate $\boldsymbol{\beta}^T$ satisfies $\|\nabla\widehat{Q}_h(\boldsymbol{\beta}^T)\|_2 \leq \delta$. Provided that $\delta \lesssim \sqrt{p/n}$, the statistical theory developed in Section 4 prevails. In our R package `conquer`, we set $\delta = 10^{-4}$ as the default value; this value can also be specified by the user.

As τ approaches 0 or 1, the Hessian matrix becomes more ill-conditioned. As a result, the step sizes computed in GD-BB may sometimes fluctuate drastically, causing instability of the algorithm. Therefore, in practice, we set an upper bound

[2] Logistic kernel smoothed approximation of the check function dates back to Amemiya, 1982, which is used as a technical device to simplify the analysis of the asymptotic behavior of a two-stage median regression estimator.

Algorithm 1 Gradient descent with Barzilai–Borwein stepsize (GD-BB) for solving conquer.

Input: data vectors $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$, $\tau \in (0,1)$, bandwidth $h \in (0,1)$, initialization $\boldsymbol{\beta}^0$, and gradient tolerance δ.

Compute $\boldsymbol{\beta}^1 \leftarrow \boldsymbol{\beta}^0 - \nabla \widehat{Q}_h(\boldsymbol{\beta}^0)$
for $t = 1, 2 \ldots$ **do**
  $\boldsymbol{\delta}^t \leftarrow \boldsymbol{\beta}^t - \boldsymbol{\beta}^{t-1}$, $\boldsymbol{g}^t \leftarrow \nabla \widehat{Q}_h(\boldsymbol{\beta}^t) - \nabla \widehat{Q}_h(\boldsymbol{\beta}^{t-1})$
  $\eta_{1,t} \leftarrow \langle \boldsymbol{\delta}^t, \boldsymbol{\delta}^t \rangle / \langle \boldsymbol{\delta}^t, \boldsymbol{g}^t \rangle$, $\eta_{2,t} \leftarrow \langle \boldsymbol{\delta}^t, \boldsymbol{g}^t \rangle / \langle \boldsymbol{g}^t, \boldsymbol{g}^t \rangle$
  $\eta_t \leftarrow \min\{\eta_{1,t}, \eta_{2,t}, 100\}$ if $\eta_{1,t} > 0$ and $\eta_t \leftarrow 1$ otherwise
  $\boldsymbol{\beta}^{t+1} \leftarrow \boldsymbol{\beta}^t - \eta_t \nabla \widehat{Q}_h(\boldsymbol{\beta}^t)$
end for when $\|\nabla \widehat{Q}_h(\boldsymbol{\beta}^t)\|_2 \le \delta$

for the step sizes by taking $\eta_t = \min\{\eta_{1,t}, \eta_{2,t}, 100\}$, for $t = 1, 2, \ldots$. Another case of an ill-conditioned Hessian arises when we have covariates with very different scales. In this case, the stepsize should be different for each covariate, and a constant stepsize will be either too small or too large for one or more covariates, which leads to slow convergence. To address this issue, we scale the covariate inputs to have zero mean and unit variance before applying gradient descent.

3.2. Warm start via asymmetric Huber regression

A good initialization helps reduce the number of iterations for GD, and hence facilitates fast convergence. Recall from Remark 3.1 that with a uniform kernel, the smoothed check function is proximal to a Huber loss (Huber, 1973). Motivated by this subtle proximity, we propose using the asymmetric Huber M-estimator as an initial estimate, and then proceed by iteratively applying gradient descent with BB update step.

Let $H_{\tau,\gamma}(u) = |\tau - \mathbb{1}(u < 0)| \cdot \{(u^2/2)\mathbb{1}(|u| \le \gamma) + \gamma(|u| - \gamma/2)\mathbb{1}(|u| > \gamma)\}$ be the asymmetric Huber loss parametrized by $\gamma > 0$. The asymmetric Huber M-estimator is then defined as

$$\widetilde{\boldsymbol{\beta}}_\gamma \in \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} \widehat{\mathcal{L}}_\gamma(\boldsymbol{\beta}), \quad \text{where} \quad \widehat{\mathcal{L}}_\gamma(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^n H_{\tau,\gamma}(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta} \rangle). \tag{3.6}$$

The quantity γ is a shape parameter that controls the amount of robustness. The main reason for choosing a fixed (neither diminishing nor diverging) tuning parameter γ in Huber (1981) is to guarantee robustness toward arbitrary contamination in a neighborhood of the model. This is at the core of the robust statistics idiosyncrasy. In particular, Huber (1981) proposed $\gamma = 1.35\sigma$ to gain as much robustness as possible while retaining 95% asymptotic efficiency for normally distributed data, where $\sigma > 0$ is the standard deviation of the random noise. We estimate σ using the median absolute deviation of the residuals at each iteration, i.e., $\mathrm{MAD}(\{r_i^t\}_{i=1}^n) = \mathrm{median}(|r_i^t - \mathrm{median}(r_i^t)|)$.

Noting that the asymmetric Huber loss is twice continuously differentiable, convex, and locally strongly convex, we use the GD-BB method described in the previous section to solve the optimization problem (3.6). Starting at iteration 0 with $\boldsymbol{\beta}^{0,0} = \mathbf{0}$, at iteration $t = 0, 1, 2, \ldots$, we compute

$$\boldsymbol{\beta}^{0,t+1} = \boldsymbol{\beta}^{0,t} - \eta_t \nabla \widehat{\mathcal{L}}_\gamma(\boldsymbol{\beta}^{0,t}) = \boldsymbol{\beta}^{0,t} + \frac{\eta_t}{n} \sum_{i=1}^n \psi_{\tau,\gamma}(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta}^{0,t} \rangle) \boldsymbol{x}_i \tag{3.7}$$

with $\eta_t > 0$ automatically obtained by the BB method, where $\psi_{\tau,\gamma}(u) = |\tau - \mathbb{1}(u < 0)| \cdot H'_{\tau,\gamma}(u) = |\tau - \mathbb{1}(u < 0)| \cdot \min\{\max(-\gamma, u), \gamma\}$. The final iterate $\boldsymbol{\beta}^{0,T'}$ for some $T' > 1$ will be used as the initial value in Section 3.1. We summarize the details in Algorithm 2.

Remark 3.2. The asymmetric Huber loss $H_{\tau,\gamma}(\cdot)$ approximates the check function $\rho_\tau(\cdot)$ as $\gamma \to 0$. Therefore, an alternative method for QR computing is to solve asymmetric Huber regression via gradient descent with a shrinking gamma. To evaluate its performance, we implement the above idea by setting $\gamma^t = c \cdot \gamma^{t-1}$ for some $c \in (0,1)$ at the tth iteration. We found that the aforementioned idea is not numerically stable across several simulated datasets, unless one controls the minimal magnitude of γ very carefully. Furthermore, the obtained solution has a higher estimation error than that of conventional QR and conquer.

4. Statistical analysis

Under the linear quantile regression model in (2.1), we write, for convenience, the generic data vector $(y, \boldsymbol{x})$ in a linear model form: given a quantile level $\tau \in (0,1)$ of interest,

$$y = \langle \boldsymbol{x}, \boldsymbol{\beta}^*(\tau) \rangle + \varepsilon(\tau), \tag{4.1}$$

Algorithm 2 GD-BB method for solving (3.6).

Input: $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$ and convergence criterion δ.

Initialize $\boldsymbol{\beta}^{0,0} = \mathbf{0}$
Compute $\gamma^0 = 1.35 \cdot \mathrm{MAD}(\{r_i^0\}_{i=1}^n)$, where $r_i^0 \leftarrow y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta}^{0,0} \rangle$, $i = 1, \ldots, n$, where $\mathrm{MAD}(\cdot)$ is the median absolute deviation
$\boldsymbol{\beta}^{0,1} \leftarrow \boldsymbol{\beta}^{0,0} - \nabla \widehat{\mathcal{L}}_{\gamma^0}(\boldsymbol{\beta}^{0,0})$
for $t = 1, 2 \ldots$ **do**
  $\gamma^t = 1.35 \cdot \mathrm{MAD}(\{r_i^t\}_{i=1}^n)$, where $r_i^t \leftarrow y_i - \langle \boldsymbol{x}_i, \boldsymbol{\beta}^{0,t} \rangle$, $i = 1, \ldots, n$
  $\boldsymbol{\delta}^t \leftarrow \boldsymbol{\beta}^{0,t} - \boldsymbol{\beta}^{0,t-1}$, $\boldsymbol{g}^t \leftarrow \nabla \widehat{\mathcal{L}}_{\gamma^t}(\boldsymbol{\beta}^{0,t}) - \nabla \widehat{\mathcal{L}}_{\gamma^t}(\boldsymbol{\beta}^{0,t-1})$
  $\eta_{1,t} \leftarrow \langle \boldsymbol{\delta}^t, \boldsymbol{\delta}^t \rangle / \langle \boldsymbol{\delta}^t, \boldsymbol{g}^t \rangle$, $\eta_{2,t} \leftarrow \langle \boldsymbol{\delta}^t, \boldsymbol{g}^t \rangle / \langle \boldsymbol{g}^t, \boldsymbol{g}^t \rangle$.
  $\eta_t \leftarrow \min\{\eta_{1,t}, \eta_{2,t}, 100\}$ if $\eta_{1,t} > 0$ and $\eta_t \leftarrow 1$ otherwise
  $\boldsymbol{\beta}^{0,t+1} \leftarrow \boldsymbol{\beta}^{0,t} - \eta_t \nabla \widehat{\mathcal{L}}_{\gamma^t}(\boldsymbol{\beta}^{0,t})$
end for when $\|\nabla \widehat{\mathcal{L}}_{\gamma^t}(\boldsymbol{\beta}^{0,t})\|_2 \le \delta$

where the random variable $\varepsilon(\tau)$ satisfies $\mathbb{P}\{\varepsilon(\tau) \le 0 | \boldsymbol{x}\} = \tau$. Let $f_{\varepsilon|\boldsymbol{x}}(\cdot)$ be the conditional density function of the regression error $\varepsilon = \varepsilon(\tau)$ given $\boldsymbol{x} = (x_1, \ldots, x_p)^{\mathrm{T}}$ ($p \ge 2$). We first derive upper bounds for the smoothing bias under mild regularity conditions on the conditional density $f_{\varepsilon|\boldsymbol{x}}$ and the kernel function. For any vector $\boldsymbol{u} \in \mathbb{R}^p$, we write $\boldsymbol{u}_- \in \mathbb{R}^{p-1}$ as the sub-vector of $\boldsymbol{u}$ with its first component removed. Recall that $x_1 \equiv 1$, and $\boldsymbol{x}_- = (x_2, \ldots, x_p)^{\mathrm{T}} \in \mathbb{R}^{p-1}$ is assumed to be random. Without loss of generality, we assume $\boldsymbol{\mu}_- := \mathbb{E}(\boldsymbol{x}_-) = \mathbf{0}$ throughout this section; otherwise, set $\widetilde{\boldsymbol{x}} = (1, (\boldsymbol{x}_- - \boldsymbol{\mu}_-)^{\mathrm{T}})^{\mathrm{T}}$, so that model (4.1) can be written as $y = \langle \widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{\beta}}^* \rangle + \varepsilon$, where $\widetilde{\boldsymbol{\beta}}^* = (\widetilde{\beta}_1^*, \beta_2^*, \ldots, \beta_p^*)^{\mathrm{T}}$ with $\widetilde{\beta}_1^* = \beta_1^* + \langle \boldsymbol{\mu}_-, \boldsymbol{\beta}_-^* \rangle$. The analysis then applies to $\{(y_i, \widetilde{\boldsymbol{x}}_i)\}_{i=1}^n$, and the probabilistic bounds for $\widetilde{\boldsymbol{\beta}}^*$ naturally lead to those for $\boldsymbol{\beta}^*$.

4.1. Smoothing bias

Condition 4.1 (*Kernel Function*)**.** Let $K(\cdot)$ be a symmetric and non-negative function that integrates to one, that is, $K(u) = K(-u)$, $K(u) \ge 0$ for all $u \in \mathbb{R}$ and $\int_{-\infty}^{\infty} K(u)\,\mathrm{d}u = 1$. Moreover, $K(\cdot)$ is bounded with $\kappa_u := \sup_{u \in \mathbb{R}} K(u) < \infty$.

We will use the notation $\kappa_k = \int_{-\infty}^{\infty} |u|^k K(u)\,\mathrm{d}u$ for $k \ge 1$. Furthermore, we define the population smoothed loss function $Q_h(\boldsymbol{\beta}) = \mathbb{E}\{\widehat{Q}_h(\boldsymbol{\beta})\}$, $\boldsymbol{\beta} \in \mathbb{R}^p$ and the pseudo parameter

$$\boldsymbol{\beta}_h^*(\tau) \in \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} Q_h(\boldsymbol{\beta}), \tag{4.2}$$

which is the population minimizer under the smoothed quantile loss. For simplicity, we write $\boldsymbol{\beta}^* = \boldsymbol{\beta}^*(\tau)$ and $\boldsymbol{\beta}_h^* = \boldsymbol{\beta}_h^*(\tau)$ hereinafter. In general, $\boldsymbol{\beta}_h^*$ differs from $\boldsymbol{\beta}^*$, and we refer to $\|\boldsymbol{\beta}_h^* - \boldsymbol{\beta}^*\|_2$ as the approximation error or smoothing bias.

Condition 4.2 (*Conditional Density*)**.** There exists $\underline{f} > 0$ such that $f_{\varepsilon|\boldsymbol{x}}(0) \ge \underline{f}$ almost surely (for all $\boldsymbol{x}$). Moreover, there exists a constant $l_0 > 0$ such that $|f_{\varepsilon|\boldsymbol{x}}(u) - f_{\varepsilon|\boldsymbol{x}}(v)| \le l_0 |u - v|$ for all $u, v \in \mathbb{R}$ almost surely (over $\boldsymbol{x}$).

Condition 4.3 (*Random Design: Moments*)**.** The (random) vector $\boldsymbol{x} \in \mathbb{R}^p$ of covariates satisfies $m_3 := \sup_{\boldsymbol{u} \in \mathbb{S}^{p-1}} \mathbb{E}(|\langle \boldsymbol{u}, \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x} \rangle|^3) < \infty$, where $\boldsymbol{\Sigma} = \mathbb{E}(\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}})$ is positive definite.

Condition 4.3 requires that all the one-dimensional marginals of $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}$ have bounded third absolute moments. When $\boldsymbol{x}_-$ follows a multivariate normal distribution, Condition 4.3 holds trivially. We refer to Remarks 4.1 and 4.2 below for more examples. The following result characterizes the smoothing bias from a non-asymptotic viewpoint.

Proposition 4.1. *Assume Conditions 4.1–4.3 hold, and let the bandwidth satisfy* $0 < h < \frac{1}{l_0\{\kappa_1 + (m_3\kappa_2)^{1/2}\}}\underline{f}$. *Then,* $\boldsymbol{\beta}_h^*$ *is the unique minimizer of* $\boldsymbol{\beta} \mapsto Q_h(\boldsymbol{\beta})$ *and satisfies*

$$\delta_h := \|\boldsymbol{\beta}_h^* - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma}} < \frac{l_0\kappa_2 h^2}{\underline{f} - l_0\kappa_1 h}. \tag{4.3}$$

In addition, assume $f_{\varepsilon|\boldsymbol{x}}(\cdot)$ *is continuously differentiable and satisfies almost surely (over* $\boldsymbol{x}$*) that* $|f'_{\varepsilon|\boldsymbol{x}}(u) - f'_{\varepsilon|\boldsymbol{x}}(0)| \le l_1|u|$ *for some constant* $l_1 > 0$. *Then*

$$\left\| \boldsymbol{\Sigma}^{-1/2}\mathbf{J}(\boldsymbol{\beta}_h^* - \boldsymbol{\beta}^*) + \frac{1}{2}\kappa_2 h^2 \cdot \boldsymbol{\Sigma}^{-1/2}\mathbb{E}\{f'_{\varepsilon|\boldsymbol{x}}(0)\boldsymbol{x}\} \right\|_2 \le \frac{1}{6}l_1\kappa_3 h^3 + \frac{1}{2}l_0 m_3 \delta_h^2 + l_0\kappa_1 h\delta_h, \tag{4.4}$$

where $\mathbf{J} = \mathbb{E}\{f_{\varepsilon|\boldsymbol{x}}(0)\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\}$.

To better understand the bounds (4.3) and (4.4), note that $\|\boldsymbol{\beta}_h^* - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma}}^2 = \mathbb{E}\langle \boldsymbol{x}, \boldsymbol{\beta}_h^* - \boldsymbol{\beta}^*\rangle^2$ is the average prediction smoothing error. Interestingly, the upper bound on the right-hand side is dimension-free given h as long as the uniform third moment m_3 in Condition 4.3 is dimension-free. See Remarks 4.1 and 4.2 for examples of multivariate distributions on $\mathbb{R}^p$ that have dimension-free uniform moment parameter. Another interesting implication is that, when both $f_{\varepsilon|\boldsymbol{x}}(0)$ and $f'_{\varepsilon|\boldsymbol{x}}(0)$ are independent of $\boldsymbol{x}$, i.e., $f_{\varepsilon|\boldsymbol{x}}(0) = f_\varepsilon(0)$ and $f'_{\varepsilon|\boldsymbol{x}}(0) = f'_\varepsilon(0)$, the leading term in the bias simplifies to

$$\frac{1}{2}\kappa_2 h^2 \cdot \mathbf{J}^{-1}\mathbb{E}\{f'_{\varepsilon|\boldsymbol{x}}(0)\boldsymbol{x}\} = \frac{f'_\varepsilon(0)}{2f_\varepsilon(0)}\kappa_2 h^2 \cdot \boldsymbol{\Sigma}^{-1}\mathbb{E}(\boldsymbol{x}) = \frac{f'_\varepsilon(0)}{2f_\varepsilon(0)}\kappa_2 h^2 \cdot \begin{bmatrix} 1 \\ \mathbf{0}_{p-1} \end{bmatrix}.$$

In other words, the smoothing bias is concentrated primarily on the intercept. In the asymptotic setting where p is fixed, and $h = o(1)$ as $n \to \infty$, we refer to Theorem 1 in Fernandes et al. (2021) for the expression of asymptotic bias.

4.2. Finite sample theory

In this section, we provide two non-asymptotic results, the concentration inequality and the Bahadur–Kiefer representation, for the conquer estimator under random design.

Condition 4.4 (*Random Design: Sub-Exponential Case*)**.** The predictor $\boldsymbol{x} = (x_1, \ldots, x_p)^{\mathrm{T}} \in \mathbb{R}^p$ is *sub-exponential* with $x_1 \equiv 1$ and $\mathbb{E}(x_j) = 0$ for $j = 2, \ldots, p$. That is, there exists $\upsilon_0 > 0$ such that $\mathbb{P}\{|\langle \boldsymbol{u}, \boldsymbol{w}\rangle| \geq \upsilon_0 t\} \leq e^{-t}$ for all $\boldsymbol{u} \in \mathbb{S}^{p-1}$ and $t \geq 0$, where $\boldsymbol{w} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}$ with $\boldsymbol{\Sigma} = \mathbb{E}(\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}})$ being positive definite.

Condition 4.4 asserts that the distribution of the covariates is sub-exponential, which encompasses the bounded case considered by Fernandes et al. (2021). For the standardized predictor $\boldsymbol{w} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}$, we define the uniform moment parameters (including m_3 that first occurred in Condition 4.3)

$$m_k = \sup_{\boldsymbol{u} \in \mathbb{S}^{p-1}} \mathbb{E}|\langle \boldsymbol{u}, \boldsymbol{w}\rangle|^k, \quad k = 1, 2, \ldots, \tag{4.5}$$

with $m_2 = 1$. In particular, m_4 can be viewed as the uniform kurtosis parameter. Under Condition 4.4, a straightforward calculation shows that $m_k \leq \upsilon_0^k k!$, valid for all $k \geq 1$.

Remark 4.1. The parameter υ_0 is often referred to as the sub-exponential parameter, which along with the sub-Gaussian parameter υ_1 defined in Condition 4.5 below, plays an important role in non-asymptotic analysis of statistical models with growing dimensions (Vershynin, 2018; Wainwright, 2019). For many "nice" distributions on $\mathbb{R}^p$, both υ_0 and υ_1 are absolute constants and thus are dimension-free. In the following, we list several p-dimensional multivariate distributions, all of which have a dimension-free sub-exponential parameter υ_0.

(i) (Multivariate normal). Let $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ for some positive definite matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$.
(ii) (Multivariate symmetric Bernoulli). Let $\boldsymbol{x} = (x_1, \ldots, x_p)^{\mathrm{T}} \sim \mathrm{Unif}(\{-1, 1\}^n)$. That is, $x_1, \ldots, x_p$ are independent, and satisfy $\mathbb{P}(x_j = 1) = \mathbb{P}(x_j = -1) = 1/2$.
(iii) (Uniform distribution on the sphere). Let $\boldsymbol{x} = (x_1, \ldots, x_p)^{\mathrm{T}}$ be a random vector uniformly distributed on the Euclidean sphere in $\mathbb{R}^p$ with center at the origin and radius $p^{1/2}$.
(iv) (Uniform distribution on the Euclidean ball). Let $\boldsymbol{x} = (x_1, \ldots, x_p)^{\mathrm{T}}$ be a random vector uniformly distributed on the Euclidean ball $\mathbb{B}^p(p^{1/2})$ in $\mathbb{R}^p$.
(v) (Uniform distribution on the unit cube). Let $\boldsymbol{x} = (x_1, \ldots, x_p)^{\mathrm{T}}$ be a random vector uniformly distributed on the unit cube $[-1, 1]^p$. That is, $x_1, \ldots, x_p$ are independent from Unif $[-1, 1]$.
(vi) (Uniform distribution on the ℓ_1-ball). Let $\boldsymbol{x} = (x_1, \ldots, x_p)^{\mathrm{T}}$ be a random vector uniformly distributed on the ℓ_1-ball $\{\boldsymbol{u} \in \mathbb{R}^p : \|\boldsymbol{u}\|_1 \leq r\}$ for $r \asymp p$.

The multivariate distributions from examples (i)–(v) are all sub-Gaussian with a constant parameter, and hence are also sub-exponential. The distribution from the last example is not sub-Gaussian, but is sub-exponential with a constant parameter. We refer to Section 3.4 in Vershynin (2018) for a detailed introduction of sub-exponential and sub-Gaussian random vectors, including examples for which the sub-Gaussian parameter does depend on and grow with p.

Remark 4.2. Another important multivariate distribution is the *elliptical distribution* (Fang et al., 1990). We say a random vector $\boldsymbol{x} \in \mathbb{R}^p$ follows an elliptical distribution, denoted by $\boldsymbol{x} \sim \mathrm{ED}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \xi)$, if it has a stochastic representation $\boldsymbol{x} \overset{d}{=} \boldsymbol{\mu} + \xi\mathbf{A}\mathbf{U}$, where ξ is a random scalar, $\mathbf{U}$ is a random vector uniformly distributed on the unit sphere $\mathbb{S}^{p-1}$ and is independent of ξ, and $\mathbf{A}$ is a deterministic matrix satisfying $\boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}^{\mathrm{T}}$. Assume $\boldsymbol{x} \sim \mathrm{ED}(\mathbf{0}, \boldsymbol{\Sigma}, \xi)$ for some $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$ and a random variable $\xi \in \mathbb{R}$. With slight abuse of notation, write $\boldsymbol{x} = \xi\mathbf{A}\mathbf{U}$. Then, for any unit vector $\boldsymbol{u} \in \mathbb{S}^{p-1}$, $|\langle \boldsymbol{u}, \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}\rangle| = |\langle \mathbf{A}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{u}, \mathbf{U}\rangle| \cdot |\xi| \leq \|\mathbf{A}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{u}\|_2 \cdot |\xi| \leq |\xi|$. This implies that (i) if $\mathbb{E}|\xi|^3 < \infty$, Condition 4.3 holds, (ii) if ξ is sub-exponential, Condition 4.4 holds with a dimension-free υ_0, and (iii) if ξ is sub-Gaussian, Condition 4.5 is satisfied with a dimension-free υ_1.

Theorem 4.1. *Assume Conditions 4.1, 4.2 and 4.4 hold. For any $t > 0$, the smoothed quantile regression estimator $\widehat{\boldsymbol{\beta}}_h$ with $\underline{f}^{-1}m_3^{1/2}\upsilon_0\sqrt{(p+t)/n} \lesssim h \lesssim \underline{f}m_3^{-1/2}$ satisfies the bound*

$$\|\widehat{\boldsymbol{\beta}}_h - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma}} \leq \frac{C}{\underline{f}}\left\{\upsilon_0\sqrt{\frac{\log_2(1/h)+p+t}{n}} + l_0\kappa_2 h^2\right\}, \tag{4.6}$$

with probability at least $1-2e^{-t}$, where $C>0$ is an absolute constant, and $\log_2(x) := \log\log(x \vee 1)$.

With high probability, the estimation error in (4.6) is upper bounded by two terms, $\underline{f}^{-1}l_0\kappa_2 h^2$ and $\underline{f}^{-1}\upsilon_0\sqrt{(p+t)/n}$, which can be interpreted as the bias and statistical rate of convergence, respectively. The parameter $t \geq 0$ controls the confidence level through $1-2e^{-t}$. The additional factor $\log_2(1/h)$ in the upper bound is a consequence of the peeling argument, which can be removed via a more refined analysis yet under slightly stronger technical conditions; see Section C.3 in the supplement for details. Adjusting the proof by changing high probability bounds to $\mathcal{O}_{\mathbb{P}}$ statements, it can be shown that $\|\widehat{\boldsymbol{\beta}}_h - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma}} = \mathcal{O}_{\mathbb{P}}(\sqrt{p/n})$ under the condition $h = \mathcal{O}((p/n)^{1/4})$ and $\sqrt{p/n} = \mathcal{O}(h)$. Next we explain the bandwidth constraint $\sqrt{p/n} \lesssim h \lesssim 1$ required in Theorem 4.1 and all the other results below. On the one side, the smoothing parameter should be sufficiently small, typically $h = h_n \to 0$, so that the smoothing bias is negligible and does not change the target parameter to be estimated. On the other side, the bandwidth cannot be too small in the sense that we need $h \gtrsim \sqrt{p/n}$. Intuitively, this is because the main motivation for smoothed QR is to seek a tradeoff between statistical rate of convergence and computational precision (unless the data is noiseless). The standard QR estimator $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\tau)$ has a convergence rate $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 = \mathcal{O}_{\mathbb{P}}(\sqrt{p/n})$ under the growth condition $p\sup_{\boldsymbol{x}\in\mathcal{X}}\|\boldsymbol{x}\|_2^2 \cdot (\log n)^2 = o(n)$; see Theorem 1 in Belloni et al. (2019). Here $\mathcal{X} \subseteq \mathbb{R}^p$ is the support of the covariate vector $\boldsymbol{x} \in \mathbb{R}^p$. Therefore, smoothing will become redundant if the bandwidth is set at a level below the best possible statistical convergence radius.

Our results provide non-asymptotic bounds via high probability statements, which complement the classical Big OP ($\mathcal{O}_{\mathbb{P}}$) and little op ($o_{\mathbb{P}}$) statements frequently used in statistics and econometrics. Probabilistic bounds of this kind can also be extended to analyze high-dimensional models (Belloni and Chernozhukov, 2011; Wang et al., 2012b) or nonparametric methods (Belloni et al., 2019).

Next, we establish a Bahadur representation for the conquer estimator, which lays the theoretical foundation for the ensuing statistical inference. To this end, we impose a slightly more stringent sub-Gaussian condition on the covariates.

Condition 4.5 (*Random Design: Sub-Gaussian Case*)**.** The predictor $\boldsymbol{x} = (x_1, \ldots, x_p)^{\mathrm{T}} \in \mathbb{R}^p$ is *sub-Gaussian* with $x_1 \equiv 1$ and $\mathbb{E}(x_j) = 0$ for $j = 2, \ldots, p$. That is, there exists $\upsilon_1 > 0$ such that $\mathbb{P}\{|\langle \boldsymbol{u}, \boldsymbol{w}\rangle| \geq \upsilon_1 t\} \leq 2e^{-t^2/2}$ for all $\boldsymbol{u} \in \mathbb{S}^{p-1}$ and $t \geq 0$, where $\boldsymbol{w} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}$.

We are primarily concerned with the cases where υ_1 is a dimension-free constant; see Remarks 4.1 and 4.2.

Theorem 4.2. *In addition to Conditions 4.1, 4.2 and 4.5, assume $\sup_{u\in\mathbb{R}} f_{\varepsilon|\boldsymbol{x}}(u) \leq \bar{f}$ almost surely. Let $t > 0$, and suppose the sample size n and bandwidth h satisfy $\underline{f}^{-1}m_3^{1/2}\upsilon_1\sqrt{(p+t)/n} \lesssim h \lesssim \underline{f}m_3^{-1/2}$. Then, with probability at least $1-3e^{-t}$,*

$$\left\|\boldsymbol{\Sigma}^{-1/2}\mathbf{J}_h(\widehat{\boldsymbol{\beta}}_h - \boldsymbol{\beta}^*) - \frac{1}{n}\sum_{i=1}^n\big\{\tau - \mathcal{K}_h(-\varepsilon_i)\big\}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}_i\right\|_2 \leq C\left(\frac{p+t}{nh^{1/2}} + h^{3/2}\sqrt{\frac{p+t}{n}} + h^4\right), \tag{4.7}$$

where $\mathbf{J}_h = \nabla^2 Q_h(\boldsymbol{\beta}^*) = \mathbb{E}\big\{K_h(\varepsilon)\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\big\}$, $\mathcal{K}_h(u) = \int_{-\infty}^{u/h} K(v)\,\mathrm{d}v$, and $C>0$ is a constant depending only on $(\upsilon_1, \kappa_2, \kappa_u, l_0, \bar{f}, \underline{f})$. When $\mathbf{J}_h$ on the left-hand side of (4.7) is replaced by $\mathbf{J} = \mathbb{E}\{f_{\varepsilon|\boldsymbol{x}}(0)\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\}$, the upper bound is of order $(p+t)/(nh^{1/2}) + h\sqrt{(p+t)/n} + h^3$.

With growing dimensions (many regressors), Theorem 4.2 is directly comparable to and complements Theorem 2 in Belloni et al. (2019), although the latter concerns the linear approximation of the quantile regression process. To see the connection, we write $n^{1/2}(\widehat{\boldsymbol{\beta}}_h - \boldsymbol{\beta}^*) = \mathbf{J}_h^{-1}\boldsymbol{U}_h + \boldsymbol{r}_h$, where $\boldsymbol{U}_h = n^{-1/2}\sum_{i=1}^n(1-\mathbb{E})\{\tau - \mathcal{K}_h(-\varepsilon_i)\}\boldsymbol{x}_i$ is a zero-mean random vector, and the remainder $\boldsymbol{r}_h$ is such that $\|\boldsymbol{r}_h\|_2 \lesssim (p+t)/(nh)^{1/2} + n^{1/2}h^2$ with high probability. Minimizing the right-hand side over h in terms of order leads to a convergence rate $p^{4/5}/n^{3/10}$. For standard QR with fixed design, Theorem 2 in Belloni et al. (2019) implies $n^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) = \mathbf{J}^{-1}\boldsymbol{U} + \boldsymbol{r}$, where $\boldsymbol{U} = n^{-1/2}\sum_{i=1}^n\{\tau - \mathbb{1}(\varepsilon_i \leq 0)\}\boldsymbol{x}_i$, and $\|\boldsymbol{r}\|_2 = \mathcal{O}_{\mathbb{P}}(p^{3/4}\zeta_p(\log n)^{1/2}n^{-1/4})$, where $\zeta_p = \sup_{\boldsymbol{x}\in\mathcal{X}}\|\boldsymbol{x}\|_2$. From an asymptotic perspective, the QR estimator has the advantage of being (conditionally) pivotal asymptotically. However, possibly due to the non-smoothness of the check function, the linear approximation error has a slower rate of convergence $(p^5/n)^{1/4}$ even for bounded design, i.e., $\zeta_p \leq Bp^{1/2}$ for some constant $B > 0$. For the conquer estimator, although the linear term $\boldsymbol{U}_h$ is not pivotal, we will show that Rademacher multiplier bootstrap provides accurate approximations both theoretically and numerically.

The Bahadur representation can be used to establish the limiting distribution of the estimator or its functionals. Here we consider a fundamental statistical inference problem for testing the linear hypothesis $H_0 : \langle \boldsymbol{a}, \boldsymbol{\beta}^*\rangle = 0$, where $\boldsymbol{a} \in \mathbb{R}^p$ is a deterministic vector that defines a linear functional of interest. It is then natural to consider a test statistic that depends on $n^{1/2}\langle \boldsymbol{a}, \widehat{\boldsymbol{\beta}}_h\rangle$. Based on the non-asymptotic result in Theorem 4.2, we establish a Berry–Esseen bound for the linear projection of the conquer estimator.

Theorem 4.3. *Assume that the conditions in Theorem 4.2 hold, and* $\sqrt{(p+\log n)/n} \lesssim h \lesssim 1$. *Then,*

$$\Delta_{n,p}(h) := \sup_{x\in\mathbb{R},\, \boldsymbol{a}\in\mathbb{R}^p} \left|\mathbb{P}\big(n^{1/2}\sigma_h^{-1}\langle \boldsymbol{a}, \widehat{\boldsymbol{\beta}}_h - \boldsymbol{\beta}^*\rangle \le x\big) - \Phi(x)\right| \lesssim \frac{p+\log n}{(nh)^{1/2}} + n^{1/2}h^2, \tag{4.8}$$

where $\sigma_h^2 = \sigma_h^2(\boldsymbol{a}) = \boldsymbol{a}^{\mathrm{T}}\mathbf{J}_h^{-1}\mathbb{E}\big[\{\mathcal{K}_h(-\varepsilon)-\tau\}^2\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\big]\mathbf{J}_h^{-1}\boldsymbol{a}$, *where* $\Phi(\cdot)$ *denotes the standard normal distribution function. Moreover,*

$$\sup_{\boldsymbol{a}\in\mathbb{R}^p}\left|\frac{\sigma_h^2(\boldsymbol{a})}{\boldsymbol{a}^{\mathrm{T}}\mathbf{J}_h^{-1}\boldsymbol{\Sigma}\mathbf{J}_h^{-1}\boldsymbol{a}} - \tau(1-\tau)\right| = O(h) \quad \text{as } h\to 0. \tag{4.9}$$

If, in addition, that $f_{\varepsilon|\boldsymbol{x}}(\cdot)$ *is twice continuously differentiable and satisfies* $|f''_{\varepsilon|\boldsymbol{x}}(u) - f''_{\varepsilon|\boldsymbol{x}}(v)| \le l_2(\boldsymbol{x})|u-v|$ *for all* $u, v\in\mathbb{R}$ *and* $\boldsymbol{x}\in\mathbb{R}^p$, *and* $l_2:\mathbb{R}^p\to\mathbb{R}^+$ *is such that* $\mathbb{E}\{l_2^2(\boldsymbol{x})\}\le C$ *for some* $C>0$. *Then,*

$$\sup_{x\in\mathbb{R},\, \boldsymbol{a}\in\mathbb{R}^p} \left|\mathbb{P}\big(n^{1/2}\sigma_h^{-1}\langle \boldsymbol{a}, \widehat{\boldsymbol{\beta}}_h - \boldsymbol{\beta}^* + 0.5\kappa_2 h^2\mathbf{J}_h^{-1}\mathbb{E}\{f'_{\varepsilon|\boldsymbol{x}}(0)\boldsymbol{x}\}\rangle \le x\big) - \Phi(x)\right| \lesssim \frac{p+\log n}{(nh)^{1/2}} + (p+\log n)^{1/2}h^{3/2} + n^{1/2}h^4. \tag{4.10}$$

Theorem 4.3 shows that for certain choice of bandwidth $h = h_n \to 0$, all the linear functionals of $\widehat{\boldsymbol{\beta}}_h$, after properly standardization, are asymptotically normal as $n, p \to \infty$ subject to some conditions. For example, if h satisfies $h = o(n^{-1/4})$, then the smoothing bias does not affect the asymptotic distribution. The Berry–Esseen bound (4.8) immediately yields a large-p asymptotic result. Taking $h = h_n = \{(p+\log n)/n\}^{2/5}$ therein, the Gaussian approximation error $\Delta_{n,p}(h)$ is of order $(p+\log n)^{4/5}n^{-3/10}$. Consequently, $n^{1/2}\langle \boldsymbol{a}, \widehat{\boldsymbol{\beta}}_h - \boldsymbol{\beta}^*\rangle$, for any given (deterministic) vector $\boldsymbol{a}\in\mathbb{R}^p$, is asymptotically normally distributed as long as $p^{8/3}/n \to 0$, which improves the best known growth condition on p for quantile regression (Welsh, 1989).

Remark 4.3 (*Optimal Bandwidth Under AMSE*)**.** From the refined Berry–Esseen bound (4.9) under an additional smoothness condition on $f_{\varepsilon|\boldsymbol{x}}(\cdot)$, we can characterize more precisely the asymptotic bias and variance–covariance matrix of $n^{1/2}(\widehat{\boldsymbol{\beta}}_h - \boldsymbol{\beta}^*)$, thus leading to the asymptotic mean squared error (AMSE). In fact, the asymptotic distribution of $n^{1/2}(\widehat{\boldsymbol{\beta}}_h - \boldsymbol{\beta}^*)$ can be approximated by that of

$$\boldsymbol{g}_h \overset{\mathrm{d}}{=} \mathcal{N}\left(\frac{1}{2}\kappa_2 n^{1/2}h^2\mathbf{J}_h^{-1}\mathbb{E}\{f'_{\varepsilon|\boldsymbol{x}}(0)\boldsymbol{x}\},\ \mathbf{J}_h^{-1}\mathbb{E}\big[\{\mathcal{K}_h(-\varepsilon)-\tau\}^2\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\big]\mathbf{J}_h^{-1}\right).$$

After some calculations (see Section A in the Appendix), we show that the AMSE of $n^{1/2}(\widehat{\boldsymbol{\beta}}_h - \boldsymbol{\beta}^*)$ is determined by

$$\mathbf{J}_h^{-1}\boldsymbol{\Sigma}^{1/2}\left[\tau(1-\tau)\mathbf{I}_p - 2\bar{\kappa}_1 h\mathbf{H} + \frac{\kappa_2^2}{4}nh^4\mathbb{E}(\boldsymbol{b})\mathbb{E}(\boldsymbol{b}^{\mathrm{T}})\right]\boldsymbol{\Sigma}^{1/2}\mathbf{J}_h^{-1}, \tag{4.11}$$

where $\boldsymbol{b} = f'_{\varepsilon|\boldsymbol{x}}(0)\boldsymbol{w} = f'_{\varepsilon|\boldsymbol{x}}(0)\boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}$, $\mathbf{H} = \mathbb{E}\{f_{\varepsilon|\boldsymbol{x}}(0)\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}\}$, and $\bar{\kappa}_1 = \int_{-\infty}^{\infty} uK(u)\mathcal{K}(u)\mathrm{d}u > 0$. Kaplan and Sun (2017) proposed to choose the bandwidth h by minimizing the trace of AMSE of the smoothed estimator. In this case, it is

$$h^* = \operatorname*{argmin}_{h>0} \operatorname{tr}\big\{0.25\kappa_2^2 nh^4\mathbb{E}(\boldsymbol{b})\mathbb{E}(\boldsymbol{b}^{\mathrm{T}}) - 2\bar{\kappa}_1 h\mathbf{H}\big\} = \left\{\frac{2\bar{\kappa}_1\mathbb{E}\{f_{\varepsilon|\boldsymbol{x}}(0)\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}\}}{\kappa_2^2 n\cdot\mathbb{E}(\boldsymbol{b}^{\mathrm{T}})\mathbb{E}(\boldsymbol{b})}\right\}^{1/3}. \tag{4.12}$$

Since $\boldsymbol{x} = (1, \boldsymbol{x}_-^{\mathrm{T}})^{\mathrm{T}}$ has 1 as its first component, we have $\boldsymbol{\Sigma}^{-1}\mathbb{E}(\boldsymbol{x}) = (1, \mathbf{0}_{p-1}^{\mathrm{T}})^{\mathrm{T}}$. As a result, in the special case where ε_i and $\boldsymbol{x}_i$ are independent, the above AMSE-optimal h^* can be reduced to

$$h^\star = \left[\frac{2\bar{\kappa}_1 f_{\varepsilon|\boldsymbol{x}}(0)}{\{\kappa_2 f'_{\varepsilon|\boldsymbol{x}}(0)\}^2}\frac{p}{n}\right]^{1/3}. \tag{4.13}$$

When $K(\cdot)$ is the Gaussian kernel, $\kappa_2 = \mathbb{E}(g^2) = 1$ and $\bar{\kappa}_1 = \mathbb{E}\{g\Phi(g)\}$, where $g\sim\mathcal{N}(0,1)$ and $\Phi(\cdot)$ is the standard normal CDF. The Monte Carlo method computes $\bar{\kappa}_1 \approx 0.282$. When $K(\cdot)$ is the logistic kernel (see Remark 3.1), we have $\kappa_2 = \pi^2/3$ and $\bar{\kappa}_1 = 1/2$. Although both h^* and $h^\star$ depend on $f_{\varepsilon|\boldsymbol{x}}(0)$ and $f'_{\varepsilon|\boldsymbol{x}}(0)$, which can only be estimated by nonparametric estimators with a rule-of-thumb bandwidth, they provide a benchmark bandwidth choice in simulation studies.

Remark 4.4. As previously noted, normal approximation result in the form of (4.8) is useful for testing linear hypothesis $H_0: \langle \boldsymbol{a}, \boldsymbol{\beta}^*\rangle = 0$, where $\boldsymbol{a}\in\mathbb{R}^p$ is a predetermined vector. For testing the global hypothesis $H_0: \boldsymbol{\beta}^* = \boldsymbol{b}_0$ with a given vector $\boldsymbol{b}_0$, Kaplan and Sun (2017) proposed a chi-square test based on a quadratic form of $\nabla\widehat{Q}_h(\boldsymbol{b}_0)$. Using Edgeworth expansion, they showed that the bandwidth that minimizes the approximate type I error of the chi-square test coincides with $h^\star$ in (4.12); see Theorem 3 therein for a chi-square approximation result via Edgeworth expansion. Instead, the choice $h \asymp \{(p+\log n)/n\}^{2/5}$ is obtained by minimizing the Berry–Esseen bound (4.8), which consists of two terms, Bahadur

Table 1
Summary of scaling conditions required for normal approximation under various loss functions.

Loss function	Design	Scaling condition
Huber loss (Huber, 1973)	Fixed design	$p^3 = o(n)$
Four times differentiable loss (Portnoy, 1985)	Fixed design (with symmetric error)	$(p \log n)^{3/2} = o(n)$
Four times differentiable loss (Mammen, 1989)	Fixed design	$p^{3/2} \log n = o(n)$
Huber loss (He and Shao, 2000)	Fixed design	$p^2 \log p = o(n)$
Huber loss (Chen and Zhou, 2020)	Sub-Gaussian	$p^2 = o(n)$
Quantile loss (Welsh, 1989; He and Shao, 2000)	Fixed design	$p^3(\log n)^2 = o(n)$
Quantile loss (Pan and Zhou, 2020)	Sub-Gaussian	$p^3(\log n)^2 = o(n)$
Convolution smoothed quantile loss	Sub-Gaussian	$p^{8/3} = o(n)$

linearization error and smoothing bias. To better understand this discrepancy, one may need to incorporate Edgeworth expansion techniques in the growing-p regime so as to characterize higher-order normal and chi-square approximation errors that cannot be seen from Berry–Esseen-type bounds. This poses additional technical challenges and warrants further investigation.

Remark 4.5 (*Large-p Asymptotics*)**.** A broader view of classical asymptotics recognizes that the parametric dimension of appropriate model sequences may tend to infinity with the sample size; that is $p = p_n \to \infty$ as $n \to \infty$. Results with increasing p are available in Welsh (1989), He and Shao (2000) and Belloni et al. (2019) when $p = o(n)$, and in Belloni et al. (2019), Wang et al. (2012b) and Koenker et al. (2017) for regularized quantile regression when $p \gg n$. In the large-p and larger-n setting—"$p \to \infty$ and $p/n \to 0$", Welsh (1989) shows that $p^3(\log n)^2/n \to 0$ suffices for a normal approximation. This growth condition remains the best known one although under weaker assumptions on the (fixed) design (He and Shao, 2000; Belloni et al., 2019). To our knowledge, the weakest fixed design assumption is $\max_{1\le i\le n} \|\boldsymbol{x}_i\|_2^2 = \mathcal{O}(p)$.

For smooth robust regression estimators, asymptotic normality can be proven under less restrictive conditions on p. Huber (1973) showed that if the loss is twice differentiable, the asymptotic normality for $\langle \boldsymbol{a}, \widehat{\boldsymbol{\beta}} \rangle$, where $\boldsymbol{a} \in \mathbb{R}^p$, holds if $p^3/n \to 0$ as n increases. Portnoy (1985) and Mammen (1989) weakened this condition to $(p \log n)^{3/2}/n \to 0$ and $p^{3/2} \log(n)/n \to 0$, respectively, when the loss function is four times differentiable. For Huber loss that has a Lipschitz continuous derivative, He and Shao (2000) obtained the scaling $p^2 \log p = o(n)$ that ensures the asymptotic normality of arbitrary linear combinations of $\widehat{\boldsymbol{\beta}}$. Table 1 summarizes our discussion here and shows that the smoothing for conquer helps ensure asymptotic normality of the estimator under weaker conditions on p than what we need for the usual quantile regression estimator.

Remark 4.6. In this paper, we show that the accuracy of conquer-based inference via the Bahadur representation (and normal approximations) has an error of rate faster than $n^{-1/4}$ yet slower than $n^{-1/2}$; see Theorems 4.2 and 4.3. For standard regression quantiles, Portnoy (2012) proposed an alternative expansion for the quantile process using the "Hungarian" construction of Komlós, Major and Tusnády. This stochastic approximation yields an error of order $n^{-1/2}$ (up to a factor of $\log n$), and hence provides a theoretical justification for accurate approximations for inference in regression quantile models.

4.3. Theoretical guarantees for inference

We next investigate the statistical properties of the bootstrapped estimator defined in (2.10), with a particular focus on the Rademacher multiplier bootstrap (RMB). To be specific, we use, in this section and the rest of the paper, the random weights $w_i = 1 + e_i$ for $i = 1, \ldots, n$, where $e_1, \ldots, e_n$ are independent Rademacher random variables, that is, symmetric sign variables with $\mathbb{P}(e_i = 1) = \mathbb{P}(e_i = -1) = 1/2$. As before, we consider array (non)asymptotics, and the obtained bootstrap approximation errors depend explicitly on (n, p) and h.

Theorem 4.4. *Assume Conditions 4.1, 4.2 and 4.5 hold. For any given $t \ge 0$, let the sample size and bandwidth satisfy $\underline{f}^{-1} m_3^{1/2} \upsilon_1 \sqrt{(p+t)/n} \lesssim h \lesssim \underline{f} m_3^{-1/2}$. Then, there exists some "good" event $\mathcal{E}(t)$ with $\mathbb{P}\{\mathcal{E}(t)\} \ge 1 - 3e^{-t}$ such that, with $\mathbb{P}^*$-probability at least $1 - 2e^{-t}$ conditioned on $\mathcal{E}(t)$,*

$$\|\widehat{\boldsymbol{\beta}}_h^\flat - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma}} \le \frac{C^\flat}{\underline{f}} \left\{ \upsilon_1 \sqrt{\frac{\log_2(1/h) + p + t}{n}} + l_0 \kappa_2 h^2 \right\}, \tag{4.13}$$

where $C^\flat > 0$ is a absolute constant.

Analogously to Theorem 4.2, we further provide a Bahadur representation result for the bootstrap estimator $\widehat{\boldsymbol{\beta}}_h^\flat$, which paves the way for validating the conquer-RMB method.

Theorem 4.5. *In addition to Conditions 4.1, 4.2 and 4.5, assume* $\sup_{u\in\mathbb{R}} f_{\varepsilon|\boldsymbol{x}}(u) \le \bar{f}$ *almost surely (in* $\boldsymbol{x}$*) and* $K(\cdot)$ *is* l_K*-Lipschitz continuous. Suppose the sample size satisfies* $n \gtrsim q := p + \log n$*, and set the bandwidth as* $h \asymp (q/n)^{2/5}$*. Then, there exists a sequence of events* $\{\mathcal{F}_n\}$ *with* $\mathbb{P}(\mathcal{F}_n) \ge 1 - 6n^{-1}$ *such that, with* $\mathbb{P}^*$*-probability at least* $1 - 3n^{-1}$ *conditioned on* $\mathcal{F}_n$*,*

$$\left\| \boldsymbol{\Sigma}^{-1/2}\mathbf{J}_h(\widehat{\boldsymbol{\beta}}_h^\flat - \widehat{\boldsymbol{\beta}}_h) - \frac{1}{n}\sum_{i=1}^n e_i\big\{\tau - \mathcal{K}_h(-\varepsilon_i)\big\}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}_i \right\|_2 \lesssim \left(\frac{q}{n}\right)^{4/5} \bigvee \left(\frac{q}{n}\right)^{3/5} \left(\frac{p\log n}{n}\right)^{1/4} \bigvee \left(\frac{q}{n}\right)^{3/5} \frac{p\log n}{n^{1/2}}. \tag{4.14}$$

As suggested by Theorem 4.3 and the discussion below, if we set the order of the bandwidth h as $\{(p+\log n)/n\}^{2/5}$, the normal approximation to the conquer estimator is asymptotically accurate provided that $p^{8/3} = o(n)$ as $n \to \infty$. For the same h, the right-hand side of (4.14) is of order $o(n^{-1/2})$ provided that $p^{8/3}(\log n)^{5/3} = o(n)$. Putting these two parts together, we have the following asymptotic bootstrap approximation result.

Corollary 4.1. *Assume the same conditions of Theorem 4.5, and let the bandwidth be of order* $h \asymp \{(p+\log n)/n\}^{2/5}$*. If the dimension* $p = p_n$ *is subject to* $p(\log n)^{5/8} = o(n^{3/8})$*, then for any deterministic vector* $\boldsymbol{a} \in \mathbb{R}^p$*,*

$$\sup_{x\in\mathbb{R}} \left|\mathbb{P}\big(n^{1/2}\langle \boldsymbol{a}, \widehat{\boldsymbol{\beta}}_h - \boldsymbol{\beta}^*\rangle \le x\big) - \mathbb{P}^*\big(n^{1/2}\langle \boldsymbol{a}, \widehat{\boldsymbol{\beta}}_h^\flat - \widehat{\boldsymbol{\beta}}_h\rangle \le x\big)\right| \xrightarrow{\mathbb{P}} 0 \text{ as } n \to \infty. \tag{4.15}$$

The proof of (4.15) follows the same argument as that in the proof of Theorem 4.3, and therefore is omitted. The additional logarithmic factor in the scaling may be an artifact of the proof technique. For standard quantile regression, Feng et al. (2011) established a fixed-p asymptotic bootstrap approximation result for wild bootstrap under fixed design.

Remark 4.7 (*Multiplier Bootstrap with More General Weighting Schemes*)**.** By examining the proof of Theorem 4.5, we see that the assumption $\mathbb{E}(e_i^2) = 1$ is not necessarily required for the bound (4.14) on bootstrap Bahadur linearization error. To retain the convexity of the bootstrap loss $\widehat{Q}_h^\flat(\boldsymbol{\beta}) = (1/n)\sum_{i=1}^n (1+e_i)\ell_h(y_i - \boldsymbol{x}_i^\mathrm{T}\boldsymbol{\beta})$, we restrict our attention to non-negative multipliers $1 + e_i \ge 0$. More generally, assume that $e_1, \ldots, e_n$ are i.i.d. satisfying

$$\mathbb{E}(e_i) = 0, \quad e_i \ge -1 \quad \text{and} \quad \log \mathbb{E}e^{\lambda e_i} \le \lambda^2\nu/2 \quad \text{for all } \lambda \ge 0 \text{ and some } \nu > 0. \tag{4.16}$$

This means that e_i has sub-Gaussian right tails. Typical examples satisfying (4.16) include: (i) uniform distribution on $[-1, 1]$, (ii) symmetric triangular distribution on $[-1, 1]$, (iii) shifted folded normal distribution $(\pi/2)^{1/2}|g| - 1$ where $g \sim \mathcal{N}(0, 1)$. The proof of the bound (4.14) under such a general scheme requires more involved argument; see, for example, the proof of Theorem 2.3 in Chen and Zhou (2020) (the unit variance assumption therein can also be relaxed). When $\kappa^2 := \mathbb{E}(e_i^2) \ne 1$, although the bootstrap approximation result (4.15) will no longer hold, by a simple variance adjustment it can be shown that

$$\sup_{x\in\mathbb{R}} \left|\mathbb{P}\big(n^{1/2}\langle \boldsymbol{a}, \widehat{\boldsymbol{\beta}}_h - \boldsymbol{\beta}^*\rangle \le x\big) - \mathbb{P}^*\big\{(n/\kappa)^{1/2}\langle \boldsymbol{a}, \widehat{\boldsymbol{\beta}}_h^\flat - \widehat{\boldsymbol{\beta}}_h\rangle \le x\big\}\right| \xrightarrow{\mathbb{P}} 0 \text{ as } n \to \infty.$$

The pivotal bootstrap confidence intervals can thus be constructed by slightly adapting the method described in Section 2.3.

The numerical performance of the Rademacher multiplier bootstrap inference for conquer will be examined in Section 5.2. The main advantage of the multiplier bootstrap method is that it does not require estimating the variance–covariance matrices in (2.11), which can be quite unstable and thus causes outliers when τ is close to 0 or 1; see Fig. 8 for a numerical comparison of the (multiplier) bootstrap percentile method and the normal-based method. When τ is reasonably bounded away from 0 and 1, say between 0.25 and 0.75, normal calibration also performs well empirically, and is computationally attractive because the estimator is only computed once.

The construction of normal-based confidence intervals is based on the estimated variances $\widehat{\sigma}_h^2(\boldsymbol{a}) = \boldsymbol{a}^\mathrm{T}\widehat{\mathbf{J}}_h^{-1}\widehat{\mathbf{V}}_h\widehat{\mathbf{J}}_h^{-1}\boldsymbol{a}$ for $\boldsymbol{a} \in \mathbb{R}^p$, where $\widehat{\mathbf{J}}_h$ and $\widehat{\mathbf{V}}_h$ are given in (2.11). In view of Theorem 4.3, the validity of normal calibration relies on the consistency of $\widehat{\mathbf{J}}_h$ and $\widehat{\mathbf{V}}_h$. In the following, we provide the consistency of $\widehat{\mathbf{J}}_h$ and $\widehat{\mathbf{V}}_h$ under the operator norm, again in the regime "$p/n \to 0$ as $p, n \to \infty$".

Note that both $\widehat{\mathbf{J}}_h$ and $\widehat{\mathbf{V}}_h$ depend on the conquer estimator, whose rate of convergence is already established in Theorem 4.1. For $\boldsymbol{\delta} \in \mathbb{R}^p$, define matrix-valued functions

$$\widehat{\mathbf{J}}_h(\boldsymbol{\delta}) = \frac{1}{n}\sum_{i=1}^n K_h(\varepsilon_i - \langle \boldsymbol{x}_i, \boldsymbol{\delta}\rangle)\boldsymbol{x}_i\boldsymbol{x}_i^\mathrm{T} \quad \text{and} \quad \widehat{\mathbf{V}}_h(\boldsymbol{\delta}) = \frac{1}{n}\sum_{i=1}^n \{\mathcal{K}_h(\langle \boldsymbol{x}_i, \boldsymbol{\delta}\rangle - \varepsilon) - \tau\}^2\boldsymbol{x}_i\boldsymbol{x}_i^\mathrm{T}, \tag{4.17}$$

so that $\widehat{\mathbf{J}}_h = \widehat{\mathbf{J}}_h(\widehat{\boldsymbol{\delta}})$ and $\widehat{\mathbf{V}}_h = \widehat{\mathbf{V}}_h(\widehat{\boldsymbol{\delta}})$ with $\widehat{\boldsymbol{\delta}} = \widehat{\boldsymbol{\beta}}_h - \boldsymbol{\beta}^*$. Conditioned on the event $\{\|\widehat{\boldsymbol{\delta}}\|_{\boldsymbol{\Sigma}} \le r\}$ for some prespecified $r > 0$ which determines the convergence rate of $\widehat{\boldsymbol{\beta}}_h$, we have

$$\|\widehat{\mathbf{J}}_h - \mathbf{J}_h\|_2 \le \sup_{\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}}\le r} \|\widehat{\mathbf{J}}_h(\boldsymbol{\delta}) - \mathbf{J}_h\|_2 \quad \text{and} \quad \|\widehat{\mathbf{V}}_h - \mathbf{V}_h\|_2 \le \sup_{\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}}\le r} \|\widehat{\mathbf{V}}_h(\boldsymbol{\delta}) - \mathbf{V}_h\|_2,$$

where $\mathbf{V}_h := \mathbb{E}\big[\{\mathcal{K}_h(-\varepsilon) - \tau\}^2 \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\big]$. The problem is thus reduced to controlling the above suprema over a local neighborhood.

Proposition 4.2. *In addition to the conditions in Theorem 4.2, assume that the kernel $K(\cdot)$ is l_K-Lipschitz continuous. For any given $r \geq 0$,*

$$\sup_{\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}} \leq r} \|\boldsymbol{\Sigma}^{-1/2}\{\widehat{\mathbf{J}}_h(\boldsymbol{\delta}) - \mathbf{J}_h\}\boldsymbol{\Sigma}^{-1/2}\|_2 \lesssim \sqrt{\frac{p\log n + t}{nh}} + r \tag{4.18}$$

holds with probability at least $1 - e^{-t}$, provided that $\max\{\sqrt{(p+t)/n}, p\log(n)/n\} \lesssim h \lesssim 1$. The same probabilistic bound also applies to $\sup_{\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}} \leq r} \|\boldsymbol{\Sigma}^{-1/2}\{\widehat{\mathbf{V}}_h(\boldsymbol{\delta}) - \mathbf{V}_h\}\boldsymbol{\Sigma}^{-1/2}\|_2$.

Following the discussions below Theorem 4.3, if we set the bandwidth as $h \asymp \{(p + \log n)/n\}^{2/5}$, $\|\widehat{\boldsymbol{\beta}}_h - \boldsymbol{\beta}^*\|_{\boldsymbol{\Sigma}} = \mathcal{O}_{\mathbb{P}}(\sqrt{(p+\log n)/n})$ and $n^{1/2}\boldsymbol{a}^{\mathrm{T}}(\widehat{\boldsymbol{\beta}}_h - \boldsymbol{\beta}^*)/\sigma_h(\boldsymbol{a}) \to \mathcal{N}(0,1)$ in distribution uniformly over $\boldsymbol{a} \in \mathbb{R}^p$ as $n \to \infty$ under the constraint $p^{8/3} = o(n)$. With the same bandwidth, it follows from Proposition 4.2 that

$$\max\big(\|\widehat{\mathbf{J}}_h - \mathbf{J}_h\|_2, \|\widehat{\mathbf{V}}_h - \mathbf{V}_h\|_2\big) = \mathcal{O}_{\mathbb{P}}\Big[\big\{(\log n)^{1/2}p^{3/10} + (\log n)^{3/10}p^{1/2}\big\}n^{-3/10}\Big] = o_{\mathbb{P}}(1).$$

This ensures the consistency of variance estimators, that is, $|\widehat{\sigma}_h^2(\boldsymbol{a})/\sigma_h^2(\boldsymbol{a}) - 1| \xrightarrow{\mathbb{P}} 0$.

5. Numerical studies

In this section, we assess the finite-sample performance of conquer via extensive numerical studies. We compare conquer to standard QR (Koenker and Bassett, 1978) and Horowitz's smoothed QR (Horowitz, 1998). Both the convolution-type and Horowitz's smoothed methods involve a smoothing parameter h. In view of Theorem 4.3, we take $h = \{(p + \log n)/n\}^{2/5}$ in all of the numerical experiments. As we will see from Fig. 5, the proposed method is insensitive to the choice of h. Therefore, we leave the fine tuning of h as an optional rather than imperative choice. In all the numerical experiments, the convergence criterion in Algorithms 1 and 2 is taken as $\delta = 10^{-4}$.

We first generate the covariates $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,p})^{\mathrm{T}}$ from a multivariate uniform distribution on the cube $3^{1/2}\cdot[-1,1]^p$ with covariance matrix $\boldsymbol{\Sigma} = (0.7^{|j-k|})_{1\leq j,k\leq p}$ using the `R` package `MultiRNG` (Falk, 1999). The random noise ε_i is generated from two different distributions: (i) Gaussian distribution, $\mathcal{N}(0,4)$; and (ii) t distribution with two degrees of freedom, t_2. Let $\boldsymbol{\beta}^* = (1, \ldots, 1)_p^{\mathrm{T}}$, and $\beta_0^* = 1$. Given $\tau \in (0,1)$, we then generate the response y_i from the following homogeneous and heterogeneous models, all of which satisfy the Assumption (2.1):

1. Homogeneous model:

$$y_i = \beta_0^* + \langle \boldsymbol{x}_i, \boldsymbol{\beta}^* \rangle + \{\varepsilon_i - F_{\varepsilon_i}^{-1}(\tau)\}, \quad i = 1, \ldots, n; \tag{5.1}$$

2. Linear heterogeneous model:

$$y_i = \beta_0^* + \langle \boldsymbol{x}_i, \boldsymbol{\beta}^* \rangle + (0.5x_{i,p} + 1)\{\varepsilon_i - F_{\varepsilon_i}^{-1}(\tau)\}, \quad i = 1, \ldots, n; \tag{5.2}$$

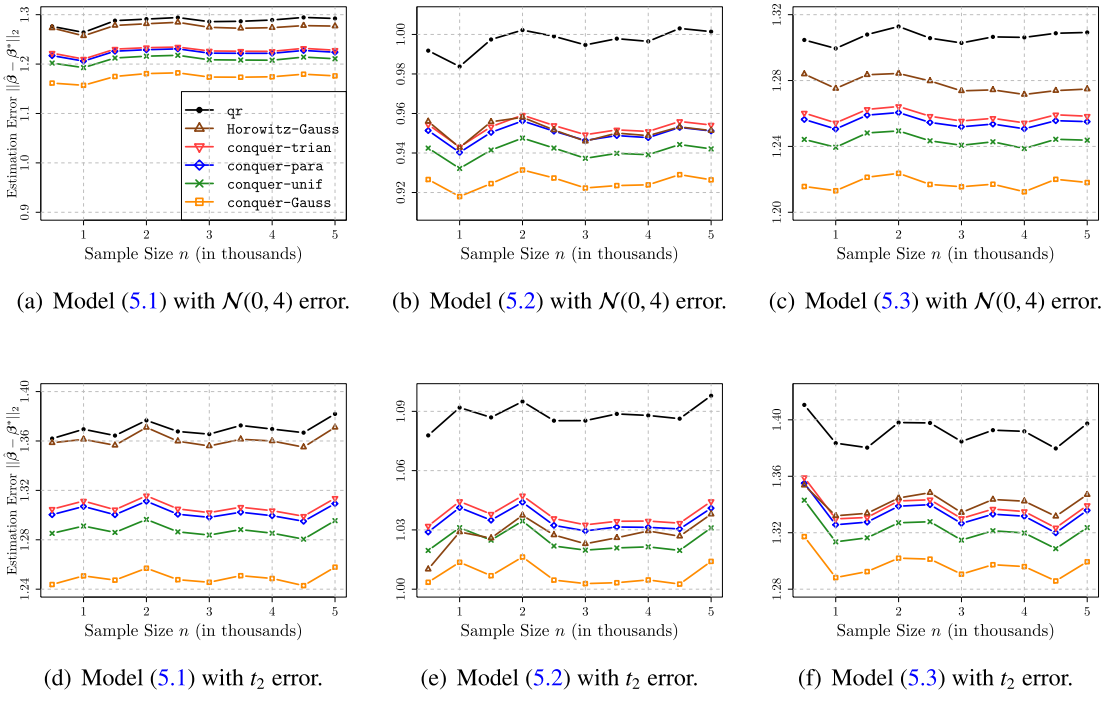
3. Quadratic heterogeneous model:

$$y_i = \beta_0^* + \langle \boldsymbol{x}_i, \boldsymbol{\beta}^* \rangle + 0.5\{1 + (x_{i,p} - 1)^2\}\{\varepsilon_i - F_{\varepsilon_i}^{-1}(\tau)\}, \quad i = 1, \ldots, n. \tag{5.3}$$

To evaluate the performance of different methods, we calculate the estimation error under the ℓ_2-norm, i.e., $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$, and record the elapsed time. The details are in Section 5.1. In Section 5.2, we examine the finite-sample performance of the multiplier bootstrap method for constructing confidence intervals in terms of coverage probability, width of the interval, and computing time.

5.1. Estimation

For all the numerical studies in this section, we consider a wide range of the sample size n, with the size-dimension ratio fixed at $n/p = 20$. That is, we allow the dimension p to increase as a function of n. We implement conquer with four different kernel functions as described in Remark 3.1: (i) Gaussian; (ii) uniform; (iii) Epanechnikov; and (iv) triangular. The classical quantile regression is implemented via a modified version of the Barrodale and Roberts algorithm (Koenker and d'Orey, 1987, 1994) by setting `method=` "`br`" in the `R` package `quantreg`, which is recommended for problems with up to several thousands of observations in Koenker (2019). For very large problems, the Frisch–Newton approach after preprocessing "pfn" is preferred. Since the same size taken to be at most 5000 throughout this section, the two methods, "br" and "pfn", have nearly identical runtime behaviors. In some applications where there are a lot of discrete covariates, it is advantageous to use method "sfn", a sparse version of Frisch–Newton algorithm that exploits sparse algebra to compute iterates (Koenker and Ng, 2003). Moreover, we implement Horowitz's smoothed quantile regression using the Gaussian kernel, and solve the resulting non-convex optimization via gradient descent with random initialization and stepsize



(a) Model (5.1) with $\mathcal{N}(0,4)$ error. (b) Model (5.2) with $\mathcal{N}(0,4)$ error. (c) Model (5.3) with $\mathcal{N}(0,4)$ error.

(d) Model (5.1) with t_2 error. (e) Model (5.2) with t_2 error. (f) Model (5.3) with t_2 error.

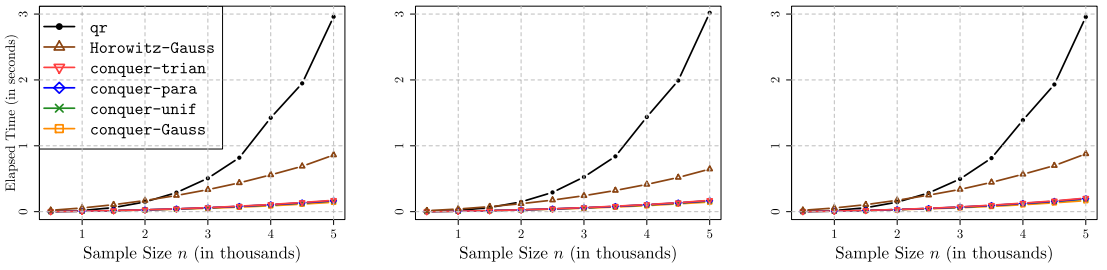
Fig. 3. Estimation error under models (5.1)–(5.3) in Section 5 with $\mathcal{N}(0,4)$ and t_2 errors, $\tau = 0.9$, averaged over 500 datasets for three different methods: (i) quantile regression `qr`, (ii) Horowitz's method with Gaussian kernel `Horowitz--Gauss`, and (iii) the conquer method with four different kernel functions `conquer-trian`, `conquer-para`, `conquer-unif`, and `conquer-Gauss`.

calibrated by backtracking line search (Section 9.3 of Boyd and Vandenberghe, 2004). The results, averaged over 500 replications, are reported.

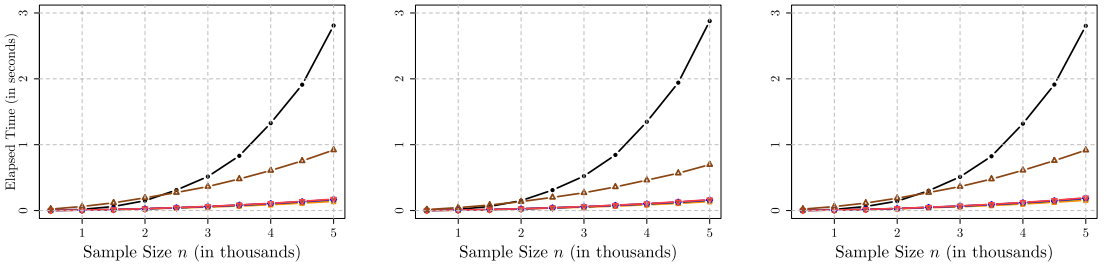
Fig. 3 depicts estimation error of the different methods under the simulation settings described in Section 5 with $\tau = 0.9$. We see that conquer has a lower estimation error than the classical QR across all scenarios, indicating that smoothing can improve estimation accuracy under the finite-sample setting. Moreover, compared to Horowitz's smoothing, conquer has a lower estimation error in most settings. Estimation error under various quantile levels $\tau \in \{0.1, 0.3, 0.5, 0.7\}$ with the $\mathcal{N}(0, 4)$ and t_2 random noise are also examined. The results are reported in Figs. E.1 and E.3 in Appendix E, from which we observe evident advantages of conquer, especially at low and high quantile levels.

To assess the computational efficiency, we compute the elapsed time for fitting the different methods. Figs. 4, E.2 and E.4 in Appendix E report the runtime for the different methods with growing sample size and dimension under the same settings as in Figs. 3 E.1 and E.3, respectively. We observe that conquer is computationally efficient and stable across all scenarios, and the runtime is insensitive to the choice of kernel functions. In contrast, the runtime for classical quantile regression grows rapidly as the sample size and dimension increase. Figs. 4, E.2 and E.4 in Appendix E show that the runtime of Horowitz's smoothing method increases significantly at extreme quantile levels $\tau \in \{0.1, 0.9\}$, possibly due to the combination of its non-convex nature and flatter gradient. As suggested by a referee, another set of simulations with covariates $\boldsymbol{x}_i$'s following multivariate normal distribution $\mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$ are also conducted, where $\boldsymbol{\Sigma} = (0.7^{|j-k|})_{1\le j,k\le p}$. When the covariates are unbounded, the assumption (2.1) is violated under the model (5.2), thus we exclude this setting from our experiments. The corresponding results of estimation and running time under t_2 noise are presented in Figs. E.5 and E.6, averaged over 500 datasets, and similar performance can be observed. In summary, we conclude that conquer significantly improves computational efficiency while retaining high statistical accuracy for fitting large-scale linear quantile regression models.

Next, we conduct a sensitivity analysis for conquer regarding the smoothing bandwidth h. We first set $(n, p) = (2000, 100)$ and consider the simulation settings (5.1)–(5.3) with $\mathcal{N}(0, 4)$ and t_2 noise. We perform conquer with h taken from a wide range, including the default value $h_{\text{de}} = \{(p+\log n)/n\}^{2/5} = 0.3107$ guided by Theorem 4.3, and the bandwidth h_{AMSE} determined by AMSE in (4.12), and compare the estimation error with that of QR in Fig. 5. For h_{AMSE}, we directly substitute in the oracle values of $f_{\varepsilon|\boldsymbol{x}}(0)$ and $f'_{\varepsilon|\boldsymbol{x}}(0)$ for illustration purpose, and in practice, estimating these two quantities usually involves non-parametric techniques. We see that the estimation error of conquer is uniformly lower than that of QR over a range of h, including our default choice, suggesting that conquer is insensitive to the choice of bandwidth h. We then consider a growing-scale scenario under model (5.2) with t_2 error, and compare conquer with QR in Fig. E.7. This can be regarded as an extension of panel (e) in Fig. 5. We observe that the estimation error of conquer using our default bandwidth is uniformly lower that of QR in various settings, and is comparable to that using AMSE-based bandwidth.

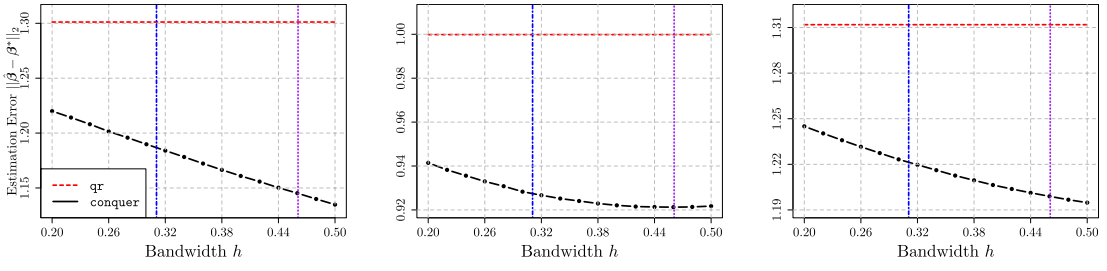


(a) Model (5.1) with $\mathcal{N}(0,4)$ error. (b) Model (5.2) with $\mathcal{N}(0,4)$ error. (c) Model (5.3) with $\mathcal{N}(0,4)$ error.

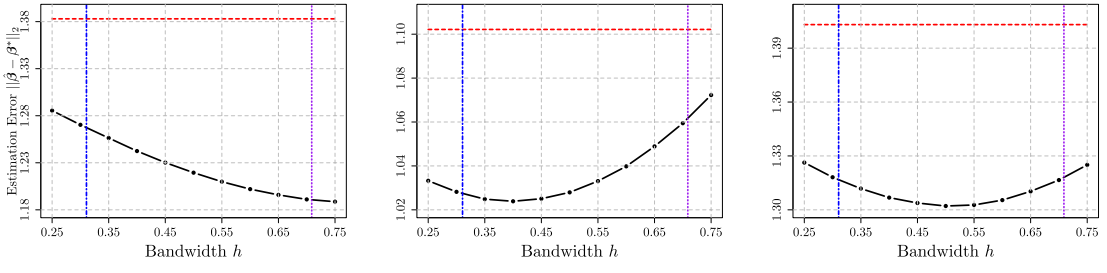


(d) Model (5.1) with t_2 error. (e) Model (5.2) with t_2 error. (f) Model (5.3) with t_2 error.

Fig. 4. Elapsed time of standard QR, Horowitz's smoothing, and conquer when $\tau = 0.9$. The model settings are the same as those in Fig. 3.



(a) Model (5.1) with $\mathcal{N}(0,4)$ error. (b) Model (5.2) with $\mathcal{N}(0,4)$ error. (c) Model (5.3) with $\mathcal{N}(0,4)$ error.



(d) Model (5.1) with t_2 error. (e) Model (5.2) with t_2 error. (f) Model (5.3) with t_2 error.

Fig. 5. Sensitivity analysis of conquer with a range of bandwidth parameter h. Results for $(n, p) = (2000, 100)$, averaged over 500 datasets, with conquer implemented using a Gaussian kernel. The blue vertical dash line represents our default choice $h_{\text{de}} = \{(p + \log n)/n\}^{2/5}$, the purple vertical dash line refers to the choice h_{AMSE} from (4.12) with some oracle knowledge, and the red horizontal dash line represents the estimation error of standard QR. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

5.2. Inference

In this section, we assess the performance of the multiplier bootstrap procedure for constructing confidence interval for each of the regression coefficients obtained from conquer. We implement conquer using the Gaussian kernel, and construct three types of confidence intervals: (i) the percentile `mb-per`; (ii) pivotal `mb-piv`; (iii) and regular `mb-norm` confidence intervals, as described in Section 2.3. We also refer to the proposed multiplier bootstrap procedure as `mb-conquer` for simplicity. We compare the proposed method to several widely used inference methods for QR. In particular, we consider confidence intervals by inverting a rank score test, `rank` (Gutenbrunner and Jurečková, 1992; Section 3.5 of Koenker, 2005); a bootstrap variant based on pivotal estimating functions, `pwy` (Parzen et al., 1994); and wild bootstrap with Rademacher weights, `wild` (Feng et al., 2011). The three methods `rank`, `pwy`, and `wild` are implemented using the `R` package `quantreg`. Note that `rank` is a non-resampling based procedure that relies on prior knowledge on the random noise, i.e., a user needs to specify whether the random noise are independent and identically distributed. In our simulation studies, we provide `rank` an unfair advantage by specifying the correct random noise structure.

We set $(n, p) = (800, 20)$, $\tau \in \{0.5, 0.9\}$, and significance level $\alpha = 0.05$. All of the resampling methods are implemented with $B = 500$ bootstrap samples. To measure the reliability, accuracy, and computational efficiency of different methods for constructing confidence intervals, we calculate the average empirical coverage probability, average width of confidence interval, and the average runtime. The average is taken over all regression coefficients without the intercept. Results based on 500 replications are reported in Fig. 6, and Figs. E.8–E.10 in Appendix E.

In Fig. 6, and Figs. E.8–E.10 in Appendix E, we use the rank-inversion method, `rank`, as a benchmark since we implement `rank` using information about the true underlying random noise, which is practically infeasible. In the case of $\tau = 0.9$, `pwy` is the most conservative as it produces the widest confidence intervals with slightly inflated coverage probability, and `wild` gives the narrowest confidence intervals but at the cost of coverage probability. The proposed methods `mb-per`, `mb-piv`, and `mb-norm` achieve a good balance between reliability (high coverage probability) and accuracy (narrow CI width), and moreover, has the lowest runtime.

To further highlight the computational gain of the proposed method, we now perform numerical studies with larger n and p. In this case, the rank inversion method `rank` is computationally infeasible. For example, when $(n, p) = (5000, 250)$, rank inversion takes approximately 80 min while conquer with multiplier bootstrap takes 41 s for constructing confidence intervals. We therefore omit `rank` from the following comparison. We consider the quadratic heterogeneous model (5.3) with $(n, p) = (4000, 100)$ and t_2 noise. The results are reported in Fig. 7. We see that `pwy` and `wild` take up to 200 s while `mb-conquer` takes less than 10 s. In summary, `mb-conquer` leads to a huge computational gain without sacrificing statistical efficiency.

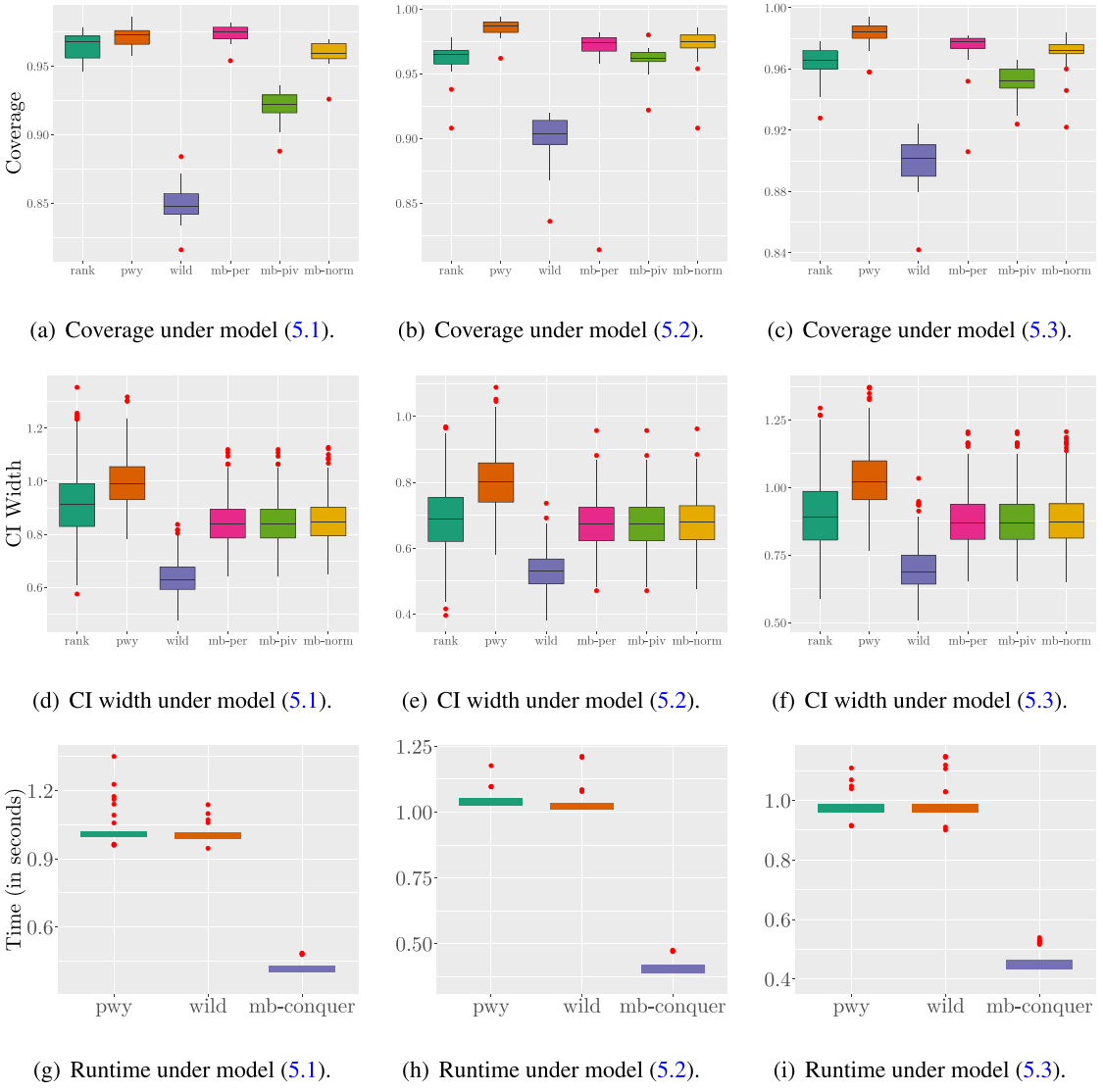
5.3. Comparison between normal approximation and bootstrap calibration

Finally, we complement the above studies with a comparison between the normal approximation and bootstrap calibration methods for confidence estimation. We consider model (5.1) with $(n, p) = (2000, 10)$. We use the same $\boldsymbol{\beta}^* \in \mathbb{R}^p$ and β_0^* as before, and generate random covariates and noise from a multivariate uniform distribution and $t_{1.5}$-distribution, respectively. For each of the p regression coefficients, we apply the proposed bootstrap percentile method and the normal-based method (Fernandes et al., 2021) to construct pointwise confidence intervals at quantile indices close to 0 and 1, that is, $\tau \in \{0.05, 0.1, 0.9, 0.95\}$. Boxplots of the empirical coverage and CI width for the two methods are reported in Fig. 8. Considering that extreme quantile regressions are notoriously hard to estimate, the bootstrap method can produce much more reliable (high coverage) and accurate (narrow width) confidence intervals than the normal-based counterpart. Therefore, for applications in which extreme quantiles are of particular interest, such as the problem of forecasting the conditional value-at-risk of a financial institution (Chernozhukov and Umantsev, 2001), the bootstrap provides a more reliable approach for quantifying the uncertainty of the estimates.

6. Discussion

In this paper, we provide a comprehensive study on the statistical properties of *conquer*, namely, convolution-type smoothed quantile regression, under the non-asymptotic setting in which p is allowed to increase as a function of n while p/n being small. When a non-negative kernel is used, the smoothed objective function is convex, twice continuously differentiable, and locally strongly convex in a neighborhood of $\boldsymbol{\beta}^*$ (with high probability). An efficient gradient-based algorithm is proposed to compute the conquer estimator, which is scalable to very large-scale problems. For traditional QR computation with linear programming, interior point algorithms are typically used to get solutions with high precision (low duality gap) (Portnoy and Koenker, 1997). When applied to large-scale datasets, this may be inefficient for two reasons: (i) it takes a lot more time to reach a duality gap of the order of machine precision, and (ii) such a generic algorithm, which is less tailored to problem structure, tends to be very slow or even run out of memory (unless with a high performance computing cluster). In this regard, convolution smoothing offers a balanced tradeoff between statistical accuracy and computational complexity.

In the context of nonparametric density or regression estimation, it is known that when higher-order kernels are used (and if the density or regression function has enough derivatives), the bias is proportional to h^ν for some $\nu \geq 4$ which

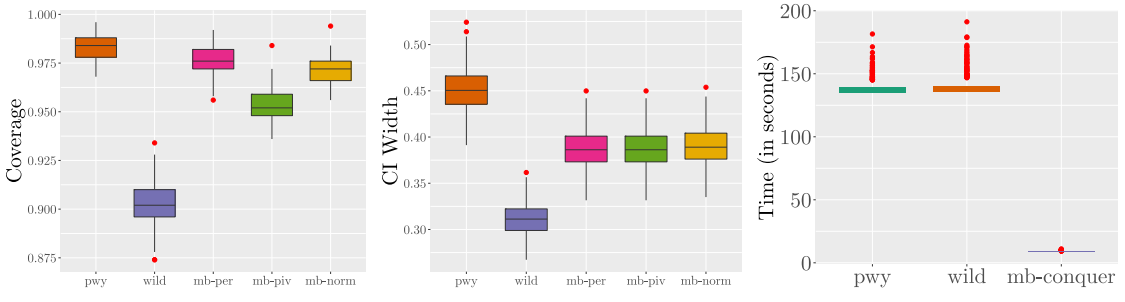


(a) Coverage under model (5.1). (b) Coverage under model (5.2). (c) Coverage under model (5.3).

(d) CI width under model (5.1). (e) CI width under model (5.2). (f) CI width under model (5.3).

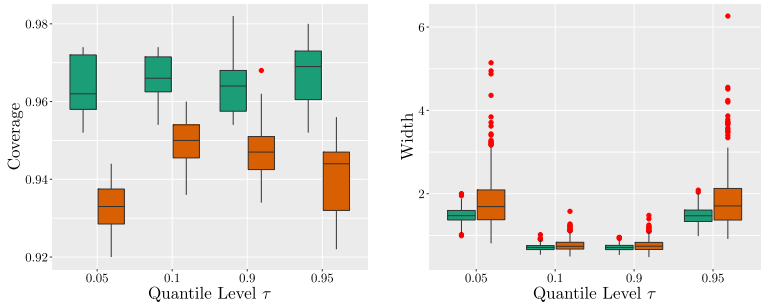
(g) Runtime under model (5.1). (h) Runtime under model (5.2). (i) Runtime under model (5.3).

Fig. 6. Empirical coverage, confidence interval width, and elapsed time of *six* methods: `rank`, `pwy`, `wild` and three types of mb-conquer: `mb-per`, `mb-piv`, and `mb-norm` under models (5.1)–(5.3) with t_2 errors. For the running time, `rank` is not included since it is not a resampling-based method. The quantile level τ is fixed to be 0.9, and the results are averaged over 500 datasets.



(a) Coverage (b) CI width (c) Runtime

Fig. 7. Empirical coverage, confidence interval width and elapsed time of `pwy`, `wild` and 3 types of mb-conquer: `mb-per`, `mb-piv`, and `mb-norm` under quadratic heterogeneous model (5.3) with t_2 errors. This figure extends the rightmost column of Fig. 6 to larger scale: $(n, p) = (4000, 100)$.



(a) Boxplots of empirical coverage (b) Boxplots of CI width

Fig. 8. Boxplots of the empirical coverage and CI width for the bootstrap percentile method and normal-based method under model (5.1) with $t_{1.5}$ error.

is of better order than h^2. Since a higher-order kernel has negative parts, the resulting smoothed loss is non-convex and thus brings the computational issue once again. Motivated by the two-stage procedure proposed by Bickel (1975) whose original idea is to improve an initial estimator that is already consistent but not efficient, we further propose a one-step conquer estimator using higher-order kernels but without the need for solving a large-scale non-convex optimization. With increasing degrees of smoothness, the one-step conquer is asymptotically normal under a milder dimension constraint of roughly $p^2/n \to 0$. Due to space limitations, the details of this method are relegated to Section B in the supplementary material.

In high-dimensional settings in which $p \gg n$, various authors have studied the regularized quantile regression under the sparsity assumption that most of the regression coefficients are zero (Belloni and Chernozhukov, 2011; Wang et al., 2012b; Zheng et al., 2015). Due to the vast literature in regularized quantile regression, we refer the reader to Chapters 15–16 of Koenker et al. (2017) for an overview. The computation of ℓ_1-penalized QR is based on either reformulation as linear programs or alternating direction method of multiplier algorithms (Gu et al., 2018). Since the conquer loss is convex and twice differentiable, we expect that gradient-based algorithms, such as coordinate gradient descent or composite gradient descent, will enjoy superior computational efficiency for solving regularized conquer without sacrificing statistical accuracy.

Acknowledgment

We sincerely thank the three anonymous referees, the Associate Editor, and Co-Editor (Xiaohong Chen) for many valuable suggestions that help improve the overall quality of the paper. We thank Roger Koenker, Stephen Portnoy, Yixiao Sun, and Hanghui Zhang for their helpful comments during early phase of the manuscript preparation. Kean Ming Tan was supported by NSF DMS-1949730, NSF DMS-2113346, and NIH RF1-MH122833. Wen-Xin Zhou was supported by NSF DMS-1811376 and NSF DMS-2113409. Xuming He was supported by NSF DMS-1951980 and NSF DMS-1914496.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jeconom.2021.07.010.

References

Amemiya, T., 1982. Two stage least absolute deviations estimators. Econometrica 50, 689–711.
Arcones, M.A., 1996. The Bahadur-Kiefer representation of L_p regression estimators. Econom. Theory 12, 257–283.
Barbe, P., Bertail, P., 1995. The Weighted Bootstrap. In: Lecture Notes in Statistics, vol. 98, Springer, New York.
Barrodale, I., Roberts, F., 1974. Solution of an overdetermined system of equations in the ℓ_1 norm. Commun. ACM 17, 319–320.
Barzilai, J., Borwein, J.M., 1988. Two-point step size gradient methods. IMA J. Numer. Anal. 8, 141–148.
Belloni, A., Chernozhukov, V., 2011. ℓ_1-Regularized quantile regression in high-dimensional sparse models. Ann. Statist. 39, 82–130.
Belloni, A., Chernozhukov, V., Chetverikov, D., Fernandez-Val, I., 2019. Conditional quantile processes based on series or many regressors. J. Econometrics 213, 4–29.
Bertsimas, D., King, A., Mazumder, R., 2016. Best subset selection via a modern optimization lens. Ann. Statist. 44, 813–852.
Bickel, P.J., 1975. One-step Huber estimates in the linear model. J. Amer. Statist. Assoc. 70, 428–434.
Boyd, S., Vandenberghe, L., 2004. Convex Optimization. Cambridge University Press, Cambridge.
de Castro, L., Galvao, A.F., Kaplan, D.M., Liu, X., 2019. Smoothed GMM for quantile models. J. Econometrics 213, 121–144.
Chatterjee, S., Bose, A., 2005. Generalized bootstrap for estimating equations. Ann. Statist. 33, 414–436.
Chen, X., Hong, H., Tarozzi, A., 2008. Semiparametric efficiency in GMM models with auxiliary data. Ann. Statist. 36, 808–843.
Chen, L.-Y., Lee, S., 2018. Exact computation of GMM estimators for instrumental variable quantile regression models. J. Appl. Econometrics 33, 553–567.

Chen, X., Linton, O., Van Keilegom, I., 2003. Estimation of semiparametric models when the criterion function is not smooth. Econometrica 71, 1591–1608.
Chen, X., Liu, W., Zhang, Y., 2019. Quantile regression under memory constraint. Ann. Statist. 47, 3244–3273.
Chen, X., Pouzo, D., 2009. Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals. J. Econometrics 152, 46–60.
Chen, X., Zhou, W.-X., 2020. Robust inference via multiplier bootstrap. Ann. Statist. 48, 1665–1691.
Cheng, G., Huang, J.Z., 2010. Bootstrap consistency for general semiparametric M-estimation. Ann. Statist. 38, 2884–2915.
Chernozhukov, V., Hansen, C., 2005. An IV model of quantile treatment effects. Econometrica 73, 245–261.
Chernozhukov, V., Hansen, C., 2006. Instrumental quantile regression inference for structural and treatment effects models. J. Econometrics 132, 491–525.
Chernozhukov, V., Hansen, C., Wüthrich, K., 2020. Instrumental variable quantile regression. arXiv:2009.00436.
Chernozhukov, V., Umantsev, L., 2001. Conditional value-at-risk: Aspects of modeling and estimation. Empir. Econ. 26, 271–293.
Dudewicz, E.J., 1992. The generalized bootstrap. In: Bootstrapping and Related Techniques. In: Lecture Notes in Economics and Mathematical Systems, vol. 376, Springer-Verlag, Berlin, pp. 31–37.
Eddelbuettel, D., Sanderson, C., 2014. RcppArmadillo: Accelerating R with high-performance C++ linear algebra. Comput. Statist. Data Anal. 71, 1054–1063.
Falk, M., 1999. A simple approach to the generation of uniformly distributed random variables with prescribed correlations. Comm. Statist. Simulation Comput. 28, 785–791.
Fang, K.T., Kotz, S., Ng, K.W., 1990. Symmetric Multivariate and Related Distributions. Chapman & Hall, London.
Feng, X., He, X., Hu, J., 2011. Wild bootstrap for quantile regression. Biometrika 98, 995–999.
Fernandes, M., Guerre, E., Horta, E., 2021. Smoothing quantile regressions. J. Bus. Econom. Statist. 39, 338–357.
Firpo, S., 2007. Efficient semiparametric estimation of quantile treatment effects. Econometrica 75, 259–276.
Galvao, A.F., Kato, K., 2016. Smoothed quantile regression for panel data. J. Econometrics 193, 92–112.
Gu, Y., Fan, J., Kong, L., Ma, S., Zou, H., 2018. ADMM for high-dimensional sparse regularized quantile regression. Technometrics 60, 319–331.
Gu, S., Kelly, B., Xiu, D., 2020. Empirical asset pricing via machine learning. Rev. Financial Stud. 33, 2223–2273.
Gutenbrunner, C., Jurečková, J., 1992. Regression rank scores and regression quantiles. Ann. Statist. 20, 305–330.
He, X., Hu, F., 2002. Markov chain marginal bootstrap. J. Amer. Statist. Assoc. 97, 783–795.
He, X., Shao, Q.-M., 1996. A general Bahadur representation of M-estimators and its application to linear regression with nonstochastic designs. Ann. Statist. 24, 2608–2630.
He, X., Shao, Q.-M., 2000. On parameters of increasing dimensions. J. Multivariate Anal. 73, 120–135.
Horowitz, J.L., 1998. Bootstrap methods for median regression models. Econometrica 66, 1327–1351.
Horowitz, J.L., Lee, S., 2007. Nonparametric instrumental variables estimation of a quantile regression model. Econometrica 75, 1191–1208.
Huber, P.J., 1973. Robust estimation: Asymptotics, conjectures and Monte Carlo. Ann. Statist. 1, 799–821.
Huber, P.J., 1981. Robust Statistics. John Wiley & Sons, New York.
Kaido, H., Wüthrich, K., 2021. Decentralization estimators for instrumental variable quantile regression models. Quant. Econ. in press.
Kaplan, D.M., Sun, Y., 2017. Smoothed estimating equations for instrumental variables quantile regression. Econom. Theory 33, 105–157.
Kocherginsky, M., He, X., Hu, F., 2005. Practical confidence intervals for regression quantiles. J. Comput. Graph. Statist. 14, 41–55.
Koenker, R., 2005. Quantile Regression. Cambridge University Press, Cambridge.
Koenker, R., 2019. Package `quantreg`, version 5.54. Reference manual: https://cran.r-project.org/web/packages/quantreg/quantreg.pdf.
Koenker, R., Bassett, G., 1978. Regression quantiles. Econometrica 46, 33–50.
Koenker, R., Chernozhukov, V., He, X., Peng, L., 2017. Handbook of Quantile Regression. CRC Press, New York.
Koenker, R., d'Orey, V., 1987. Computing quantile regressions. J. R. Stat. Soc. Ser. C. Appl. Stat. 36, 383–393.
Koenker, R., d'Orey, V., 1994. A remark on Algorithm AS 229: Computing dual regression quantiles and regression rank scores. J. R. Stat. Soc. Ser. C. Appl. Stat. 43, 410–414.
Koenker, R., Ng, P., 2003. SparseM: A sparse matrix package for R. J. Statist. Softw. 8, 1–9.
Ma, S., Kosorok, M.R., 2005. Robust semiparametric M-estimation and the weighted bootstrap. J. Multivariate Anal. 96, 190–217.
Machado, J.A.F., Santos Silva, J.M.C., 2018. Quantile via moments. J. Econometrics 213, 145–173.
Mammen, E., 1989. Asymptotics with increasing dimension for robust regression with applications to the bootstrap. Ann. Statist. 17, 382–400.
Nemirovski, A., Yudin, D., 1983. Problem Complexity and Method Efficiency in Optimization. Wiley.
Pan, X., Zhou, W.-X., 2020. Multiplier bootstrap for quantile regression: Non-asymptotic theory under random design. Inf. Inference in press.
Parikh, N., Boyd, S., 2014. Proximal algorithms. Found. Trends Optim. 1, 127–239.
Parzen, M.I., Wei, L.J., Ying, Z., 1994. A resampling method based on pivotal estimating functions. Biometrika 81, 341–350.
Portnoy, S., 1985. Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large; II. Normal approximation. Ann. Statist. 13, 1403–1417.
Portnoy, S., 2012. Nearly root-n approximation for regression quantile processes. Ann. Statist. 40, 1714–1736.
Portnoy, S., Koenker, R., 1997. The Gaussian hare and the Laplacian tortoise: Computability of squared-error versus absolute-error estimators. Statist. Sci. 12, 279–300.
van der Vaart, A.W., Wellner, J.A., 1996. Weak Convergence and Empirical Processes: With Applications to Statistics. Springer, New York.
Vershynin, R., 2018. High-Dimensional Probability: An Introduction with Applications in Data Science. Cambridge University Press, Cambridge.
Wainwright, M.J., 2019. High-Dimensional Statistics: A Non-Asymptotic Viewpoint. Cambridge University Press, Cambridge.
Wang, H.J., Stefanski, L.A., Zhu, Z., 2012a. Corrected-loss estimation for quantile regression with covariate measurement errors. Biometrika 99, 405–421.
Wang, L., Wu, Y., Li, R., 2012b. Quantile regression for analyzing heterogeneity in ultra-high dimension. J. Amer. Statist. Assoc. 107, 214–222.
Welsh, A.H., 1989. On M-processes and M-estimation. Ann. Statist. 15, 337–361.
Whang, Y.-J., 2006. Smoothed empirical likelihood methods for quantile regression models. Econom. Theory 22, 173–205.
Wu, Y., Ma, Y., Yin, G., 2015. Smoothed and corrected score approach to censored quantile regression with measurement errors. J. Amer. Statist. Assoc. 110, 1670–1683.
Zheng, Q., Peng, L., He, X., 2015. Globally adaptive quantile regression with ultra-high dimensional data. Ann. Statist. 43, 2225–2258.
Zhu, Y., 2018. K-step correction for mixed integer linear programming: A new approach for instrumental variable quantile regressions and related problems. arXiv:1805.06855.