Smoothed High-Quantile Regression in Low and High Dimensions under Data Contamination

Abstract

We consider smoothed high-quantile regression in both low and high dimensions when the observed data are contaminated by arbitrary outliers. A convolution-type kernel smoothing is used to enable efficient first-order gradient-based algorithms for the estimation with and without a penalty term in the loss function. We provide analytical results to understand and quantify the effect of a high quantile level when it is chosen close to the boundary, and the effect of the percentage of outliers whose distribution can be unknown and arbitrary. In particular, the convergence rate in the current setting is shown to depend on an interplay among multiple factors including, for example, how extremal the quantile level is, the degree to which the data is contaminated, and the tail distribution of the representative data. We consider the problem in both low and high dimensions, where an additional penalty term is needed to handle the large number of variables in the high-dimensional case. Numerical experiments, including a Monte Carlo simulation study and a real data example, are provided to illustrate the developed results.

Keywords: contaminated data, high-dimensional regression, high-quantile regression, smoothed quantile regression

1 Introduction

Since the seminal work of Koenker and Bassett (1978), quantile regression has been celebrated as a powerful alternative to the traditional least squares for studying the relationship between the response variable and a set of explanatory variables. It has enjoyed tremendous developments, and an extensive amount of research has been carried out to study quantile regression estimators; see for example, Bai et al. (1992), Gutenbrunner and Jureckova (1992), He (1997), Wu (2007), Zhou and Shao (2013), and He et al. (2023), among many others. We also refer to the book by Koenker (2005) for a nice review and additional references. Besides its different conceptual focus on the quantile instead of the mean, quantile regression models are also frequently adopted as a robust approach for estimating the underlying regression coefficients in the presence of outliers. For example, Suhaila et al. (2021) considered the use of quantile to obtain robust estimators for linear regression models in the presence of multicollinearity and outliers. Ling et al. (2022) used quantile regression to remove microbiome batch effects in the form of systematic variation originating from differential handling and processing of specimens. Akrami et al. (2024) considered the use of quantile regression in a deep learning framework with an application to the task of medical imaging translation. The robustness of quantile regression estimators mainly

stems from the fact that outliers in the form of unusually large observations generally do not have a large effect on the data order around moderate quantiles that are under consideration.

In many applications, however, it is often desirable to consider quantile regression at quantile levels that are close to the boundary to focus on the study of tail phenomena. For example, Abrevaya (2001) considered the analysis on low percentiles of the birth weight distribution. Elsner et al. (2008) studied the increasing intensity of the strongest tropical cyclones from a quantile regression perspective. Tsai (2012) considered the relationship between stock price index and exchange rates when exchange rates are extremely high or low. When the quantile level is in the tail close to the boundary, the problem of high-quantile regression was studied by Chernozhukov (2005), Belloni and Chernozhukov (2011), Wang and Li (2013), He et al. (2016a,b), Wang and Wang (2016), Zhang (2018), and Zhang (2021), among others. Since high quantiles are more related to extremal observations, it is expected that the effect of outliers may become more noticeable in the high quantile setting. In addition, outliers in many data applications appear as samples from a different distribution and may exist over the whole data range; see for example Huber (1964), Punzo and McNicholas (2016), Kluin et al. (2018), Mei et al. (2018), Zhang et al. (2022), Zhang (2025), and references therein. This motivates us to consider the problem of high-quantile regression in the presence of outliers to quantify the interplay between a high quantile level and the degree of data contamination on the quality of quantile regression estimators in both low and high dimensions.

The technique of classical quantile regression, however, is fundamentally based on solving non-smooth optimization problems, since the check loss is non-differentiable. Such a non-Such a non-smoothness can lead to nontrivial theoretical and computational challenges, particularly in the high-dimensional quantile regression setting; see, for example, the discussions in He and Shao (1996, 2000). Horowitz (1998) proposed to smooth the indicator part of the check function by a survival function, which gains smoothness at the cost of convexity and thus may cause optimization-related issues. Tan et al. (2022) and He et al. (2023) considered a different approach that applies kernel smoothing to the gradient of the check function, which has been shown to preserve convexity and facilitate optimization. However, their analysis was mainly conducted in the conventional quantile regression setting, where the quantile level is treated as fixed and there is no data contamination. In the presence of data contamination, Zhang et al. (2022) studied how the percentage of outliers from an arbitrary distribution may affect certain statistical properties of M-estimators. Their results, however, were developed under the major assumption that the idealized noise has a symmetric distribution with mean zero, and therefore cannot be applied directly to quantile regression models.

In this paper, we consider smoothed high-quantile regression in low and high dimensions when the observed data are contaminated by some outliers whose distribution can be unknown and arbitrary. We consider the smoothed quantile regression approach proposed by Tan et al. (2022) and He et al. (2023), which uses kernel smoothing of the gradient of the check function to preserve convexity and facilitate computational efficiency via first-order gradient algorithms. We first develop theoretical results for the smoothed quantile regression model in the low-dimensional case, starting with a characterization of the effect of outliers on the population risk function. In particular, we show that the existence of outliers may alter the stationary point of the population

risk function, yet it remains confined within a neighborhood of the true underlying parameter when the degree of data contamination is controlled. In this case, the gradient descent algorithm is shown to converge to the optimal solution at an exponential rate, and we establish an explicit bound for the resulting high-quantile regression estimator with respect to the true underlying parameter. Our result reveals that the convergence rate in the high-quantile regression setting with data contamination may depend on an interplay among multiple factors, including, for example, how extremal the quantile level is, the amount of outliers in the data, and the tail distribution of the representative data. Unlike the conventional quantile regression setting, where the quantile level is treated as fixed, and the convergence rate is typically universal across different underlying distributions, the convergence rate in the current high-quantile regression setting can be affected by the tail behavior of the underlying distribution, with heavier tails generally leading to slower convergence. We then extend our results to the high-dimensional case, where the number of explanatory variables can be comparable to or even larger than the sample size. In this case, an additional penalty term is needed for regularization to handle the curse of dimensionality, and the established convergence rate may differ from that in the low-dimensional case. Our results are among the earliest efforts to provide an analytical quantification of the joint effect of a tail quantile level and data contamination on quantile regression estimators in both low and high dimensions.

The remainder of the paper is organized as follows. Section 2 introduces the setup for smoothed high-quantile regression under data contamination, and our main results are developed and presented in Section 3. Numerical experiments including a Monte Carlo simulation study and a real data illustration are provided in Sections 4 and 5, respectively. Section 6 concludes the paper. We shall here introduce some notations. Given vectors $\mu = (\mu_1, \ldots, \mu_p)^\top \in \mathbb{R}^p$ and $\nu = (\nu_1, \ldots, \nu_p)^\top \in \mathbb{R}^p$ with $^\top$ denoting the transpose operator, we write the Euclidean norm $\|\mu\|_2 = (\sum_{i=1}^p \mu_i^2)^{1/2}$ and their inner product $\langle \mu, \nu \rangle = \sum_{i=1}^p \mu_i \nu_i$. For any matrix $M \in \mathbb{R}^{m \times m}$, we use $\lambda_{\max}(M)$ and $\lambda_{\min}(M)$ to denote its largest eigenvalue and the smallest eigenvalue, respectively. The operator norm of M is denoted by $\|M\|_{op}$, which equals to $\max\{\lambda_{\max}(M), -\lambda_{\min}(M)\}$. We also denote I_m as the identity matrix in $\mathbb{R}^{m \times m}$. Let $B_q^p(a, r) = \{x \in \mathbb{R}^p : \|x - a\|_q \leq r\}$ be the ℓ_q ball in $\mathbb{R}^p$ with center a and radius r, and we write $B_q^p(r) = B_q^p(0, r)$.

2 Smoothed High-Quantile Regression with Data Contamination

Given random samples $(y_i, \mathbf{x}_i)$, $i = 1, \ldots, n$, with $y_i \in \mathbb{R}, \mathbf{x}_i \in \mathbb{R}^p$, we consider the quantile regression model (Koenker, 2005)

$$y_i = \langle \theta_0, \mathbf{x}_i \rangle + \epsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where θ_0 is the true unknown regression coefficient for the $(1-\alpha)$-th quantile, and ϵ_i, $i = 1, \ldots, n$, are the associated auxiliary variables. In the conventional quantile regression model, the quantile level is typically chosen to be moderate and away from the boundary, and the auxiliary variables are properly centered at the associated quantile with $\mathbf{P}(\epsilon \leq 0 \mid \mathbf{x}) = 1 - \alpha$; see for example, Koenker and Bassett (1978) and Koenker (2005). In the current paper, however, we consider

the situation when the quantile level is chosen close to one in the tail and when the data are contaminated by outliers with an arbitrary distribution. In particular, we assume that the auxiliary variable ϵ follows the Huber's gross error model as used in Huber (1964) whose conditional density is given by the mixture $(1-\delta)f_0 + \delta g$, where $f_0 = f_0(\epsilon \mid \mathbf{x})$ represents the conditional density of the auxiliary variable given the feature vector for idealized samples, g represents the counterpart for contaminated samples whose distribution can be arbitrary, and $\delta \in [0, 1/2]$, which is also known as the contamination ratio in robust statistics literature, indicates the proportion of data contamination. Given a random variable X with probability density function f, we denote the corresponding probability and expectation by $\boldsymbol{P}_f$ and $\boldsymbol{E}_f$ respectively. We will often omit the density function subscript f when it is clear from the context, and the expectation is taken over all variables. Under data contamination, it is assumed that $\mathbf{P}_{f_0}(\epsilon \leq 0 \mid \mathbf{x}) = 1 - \alpha$, namely the target regression coefficient θ_0 in (1) is related to the desired quantile of the idealized samples only, and as a result data contamination may affect the stationary point of the population risk function making it a nontrivial problem.

For model (1), the conventional quantile regression estimator can be obtained by

$$\underset{\theta}{\text{Minimize:}} \quad \frac{1}{n}\sum_{i=1}^{n} \rho_{1-\alpha}(y_i - \langle \theta, \mathbf{x}_i \rangle), \tag{2}$$

where $\rho_{1-\alpha}(t) = (1-\alpha)t^+ + \alpha(-t)^+$ is the check function with $t^+ = \max(t, 0)$. Let $I(\cdot)$ denote the indicator function, then the associated estimating equation can be derived by differentiating the objective function with respect to θ as

$$\sum_{i=1}^{n} \{I(y_i - \langle \theta, \mathbf{x}_i \rangle > 0) - \alpha\}\mathbf{x}_i = 0.$$

Note that the indicator function is not differentiable at the point $y_i = \langle \theta, \mathbf{x}_i \rangle$, to enable efficient first-order gradient-based algorithms, we consider smoothing the indicator function as in Tan et al. (2022) and He et al. (2023). In particular, let $K(\cdot)$ be a kernel function that is bounded, symmetric, and integrated to one, and we write $K_h(t) = K(t/h)/h$ for some positive bandwidth h, then we can approximate the indicator function by the smoothed counterpart $G_h(y_i - \langle \theta, \mathbf{x}_i \rangle)$ with $G_h(t) = \int_{-\infty}^{t} K_h(u)du$. Let $\phi_h(t) = \int_{-\infty}^{\infty} \rho_{1-\alpha}(u)K_h(u-t)du$ and $\psi_h(t) = G_h(t) - \alpha$, then $d\phi_h(t)/dt = \psi_h(t)$ and $\psi_h(t) \to I(t > 0) - \alpha$ as $h \to 0$ for any $t \in \mathbb{R}$. We consider the smoothed quantile regression estimator with a constraint on the ℓ_2 norm of θ, which yields the following constrained optimization problem:

$$\begin{aligned} &\underset{\theta}{\text{Minimize:}} \quad \hat{R}_n(\theta) := \frac{1}{n}\sum_{i=1}^{n} \phi_h(y_i - \langle \theta, \mathbf{x}_i \rangle) \\ &\text{subject to:} \quad \|\theta\|_2 \leq r \end{aligned} \tag{3}$$

for some search radius $r > 0$. Due to the smoothness of the convolution-type function $\phi_h(\cdot)$, the

minimization problem in (3) is equivalent to finding solutions of

$$\sum_{i=1}^{n} \psi_h(y_i - \langle \theta, \mathbf{x}_i \rangle)\mathbf{x}_i = 0$$
$$\text{subject to:} \quad \|\theta\|_2 \le r.$$

Before presenting our main theorems, we make the following mild assumptions on the kernel function $K(\cdots)$, the feature vectors $\mathbf{x}_i$, and the distribution of the idealized auxiliary variable f_0.

Assumption 1.

(a) *The kernel function $K(\cdot)$ is a symmetric and non-negative function that integrates to one with $|K(t)| \le L$ and $|K'(t)| \le L$ for some $L < \infty$ for all $t \in \mathbb{R}$. Moreover, $\mu_1 = \int_{-\infty}^{\infty} |t|K(t)dt < \infty$ and $\mu_2 = \int_0^{\infty} t^2 K(t)dt < \infty$.*

(b) *The feature vector $\mathbf{x}$ is τ^2-sub-Gaussain with zero mean, that is, $\boldsymbol{E}[e^{\langle \lambda, \mathbf{x} \rangle}] \le \exp(\tau^2 \|\lambda\|_2^2/2)$ for all $\lambda \in \mathbb{R}^p$.*

(c) *The feature vector $\mathbf{x}$ spans all possible directions in $\mathbb{R}^p$, that is, $\lambda_{\min}(\boldsymbol{E}[\mathbf{x}\mathbf{x}^\top]) \ge \gamma\tau^2$ for some $0 < \gamma \le 1$.*

(d) *There exist positive numbers m_0 and M_0 such that $f_0(0 \mid \mathbf{x}) \ge m_0$ for any $\mathbf{x} \in \mathbb{R}^p$ and $f_0(\epsilon \mid \mathbf{x}) \le M_0$ for all $\epsilon \in \mathbb{R}$ and $\mathbf{x} \in \mathbb{R}^p$.*

(e) *There exists a constant $\ell > 0$ such that $|f_0(\epsilon_1 \mid \mathbf{x}) - f_0(\epsilon_0 \mid \mathbf{x})| \le \ell|\epsilon_0 - \epsilon_1|$ for all $\epsilon_0, \epsilon_1 \in \mathbb{R}$ and $\mathbf{x} \in \mathbb{R}^p$.*

We shall here provide a brief discussion on Assumption 1. In particular, part (a) on the kernel function is very mild and is satisfied by many popular kernel choices such as the Gaussian kernel, the uniform kernel, the triangular kernel, and the Laplace kernel, among many others. Part (b) essentially assumes that the feature vector has sub-Gaussian tails, which has been widely used in the literature of high-dimensional data analysis; see, for example, (Mei et al., 2018; He et al., 2023). Part (c) assumes that the covariance matrix of the feature vector is positive semidefinite, which is very mild. Part (d) requires the conditional density of the idealized auxiliary variable to be bounded away from zero and infinity, which is mild and satisfied by many distributions. Part (e) concerns the Lipschitz continuity of the conditional density, which is also mild and satisfied by most continuous distributions. The aforementioned assumptions are standard in the quantile regression literature, whereas in the current context of high-quantile regression, the constant m_0 in part (d) may depend on the quantile level $1 - \alpha$.

3 Main Results

3.1 Smoothed High-Quantile Regression in Low Dimensions

We consider the low-dimensional case in which the feature vector $\mathbf{x}$ has much lower dimensionality than the sample size. In this case, we study the estimator obtained from (3), and we first

define the associated population risk under Huber's gross error model as

$$R(\theta) = \boldsymbol{E}\hat{R}_n(\theta) = \boldsymbol{E}\{\phi_h(Y - \langle\theta, \mathbf{x}\rangle)\} = \boldsymbol{E}\{\phi_h(\langle\theta_0 - \theta, \mathbf{x}\rangle + \epsilon)\}. \tag{4}$$

For our theoretical developments, we first analyze the population risk and then build a link between the population risk and the empirical risk to solve the original estimation problem. Theorem 3.1 summarizes the results for the population risk function $R(\theta)$ in (4).

Theorem 3.1. *Assume Assumption 1 and that the parameter θ_0 satisfies $\|\theta_0\|_2 \leq r/3$.*

(a) *There exist positive constants C_1 and C_2 that depend on μ_2, ℓ, r and γ but not on h, m_0, α, δ and the outlier distribution g such that*

$$\|\theta^* - \theta_0\|_2 \leq \eta_0$$

holds for any stationary point θ^* of $R(\theta)$ with

$$\eta_0 = m_0^{-2}\left(h^2 C_1 + \frac{\delta}{1-\delta}C_2\right).$$

(b) *There exist positive constants C_3, C_4 and C_5 that depend on μ_1, ℓ, r and γ but not on h, m_0, α, δ and the outlier distribution g such that*

$$\lambda_{\min}\{\nabla^2 R(\theta)\} \geq 4(1-\delta)\ell\tau^3\eta_1$$

holds for any $\|\theta - \theta_0\|_2 \leq \eta_1$ with

$$\eta_1 = C_3 m_0 - C_4 h - C_5\frac{\delta}{(1-\delta)h}.$$

(c) *There is a unique stationary point of $R(\theta)$ in the closed ball $B_2^p(r)$ as long as $\eta_0 < \eta_1$.*

We shall here provide some remarks to help better understand Theorem 3.1. First, under data contamination, the auxiliary variables are assumed to follow a gross error model with δ percentage of outliers with an arbitrary distribution. This, together with the use of a smoothed approximation $G_h(\cdot)$ to the indicator function, makes the desired unknown parameter θ_0 no longer necessarily the stationary point of the population risk function $R(\theta)$; see also the discussion in Section 2. By Theorem 3.1 (a), however, the stationary point of $R(\theta)$ will be bounded in a small neighborhood of the desired unknown parameter θ_0 if h and δ are small. In particular, it can be seen from Theorem 3.1 (a) that the associated bias between θ^* and θ_0 goes to zero if the bandwidth $h \to 0$ and the gross error ratio $\delta \to 0$. This indicates that the smoothed high-quantile regression is robust to a small amount of outliers with arbitrary distributions. On the other hand, the bias increases with the increase of bandwidth h and the gross error ratio δ. We also emphasize that the constant m_0 in the current high-quantile regression setting depends on the quantile level $1-\alpha$, and we will later work out the detailed expression of it with respect to α for some specific idealized auxiliary variable distributions. In particular, our results indicate that as $1-\alpha$ approaches 1 or α approaches 0, m_0 approaches 0, suggesting that the current

high-quantile regression setting can be more challenging than conventional quantile regression. On the other hand, by part (b) of Theorem 3.1, the population risk $R(\theta)$ is $4(1-\delta)\ell\tau^3\eta_1$-strongly convex in a neighborhood of θ_0, which serves as the foundation for a gradient-based search algorithm. The strongly convex region around θ_0, however, will decrease with the increase of δ and the decrease of m_0 from Theorem 3.1 (b). As a result, the optimization problem can become more challenging if we consider situations with a higher gross error ratio δ or a more extremal quantile level $1-\alpha$. Based on the aforementioned results, Theorem 3.1 (c) then provides the existence of a unique stationary point to the population risk $R(\theta)$. Note that $\eta_1 = C_3 m_0 - C_4 h$ and $\eta_0 = C_1 h^2/m_0^2$ when $\delta = 0$, one can choose the bandwidth h with $C_3 m_0 - C_4 h > C_1 h^2/m_0^2$ so that $\eta_1 > \eta_0$ to guarantee the uniqueness of the stationary point within the closed ball. Since η_0 and η_1 are both continuous in δ from their explicit expressions, there exists a positive $\delta^* > 0$ such that $\eta_0 < \eta_1$ holds for all $0 \le \delta \le \delta^*$. This indicates that the smoothed high-quantile regression problem in the population case has a nice tractability property when the proportion of outliers is small. However, we may lose the nice tractability property for the high-quantile regression problem when α is close to 0. In a nutshell, when data contains outliers, the estimation and the corresponding optimization problem become more difficult both in terms of solution quality and computational efficiency for smoothed high-quantile regression. While the former is well expected, we find the latter — that a higher quantile makes the estimation more difficult to numerically compute – an interesting and somewhat surprising discovery.

Our next step is to link the empirical risk function $\hat{R}_n(\theta)$ and the corresponding estimator to their population version $R(\theta)$. In particular, Lemma 3.2 provides the global uniform convergence theorem of the sample gradient and Hessian.

Lemma 3.2. *Assume Assumption 1 and $0 < h < 1$. Then for any $\pi > 0$ there exist positive constants C_0 and C_1 that depend only on L, r, π, τ and γ but not on p, n, h, α, δ and the outlier distribution g such that the followings hold.*

(a) *The sample gradient converges uniformly to the population gradient in Euclidean norm in the sense that*

$$\boldsymbol{P}\left(\sup_{\theta\in B_2^p(r)} \|\nabla\hat{R}_n(\theta) - \nabla R(\theta)\|_2 \le C_0\sqrt{\frac{p\log n + p\log(1/h)}{n}}\right) \ge 1-\pi$$

holds for any $p \le C_0^{-1} n/\log n$.

(b) *The sample Hessian converges uniformly to the population Hessian in operator norm in the sense that*

$$\boldsymbol{P}\left(\sup_{\theta\in B_2^p(r)} \|\nabla^2\hat{R}_n(\theta) - \nabla^2 R(\theta)\|_{op} \le \sqrt{\frac{C_1 p\log n}{nh^3}}\right) \ge 1-\pi$$

holds for any $p \le C_1^{-1} nh^3/\log n$.

From Lemma 3.2, it is interesting to see that the convergence rates of the sample gradient and the sample Hessian do not depend on the contamination ratio δ and the outlier distribution. The main reason is that due to the boundedness of the loss function ϕ_h and its derivative ψ_h, any outliers will have limited impacts on the risk functions. In addition, the convergence rate

of the gradient is less sensitive to the choice of the bandwidth h than the second-order Hessian. Intuitively, note that $\nabla R(\theta) = \boldsymbol{E}[-\mathbf{x}\psi_h(Y - \langle\theta, \mathbf{x}\rangle)]$ where $|\psi_h(t)| \leq 1$ for any $h \geq 0$, and thus the smoothing parameter h does not directly affect the scale of the gradient, making its convergence rate less sensitive to the choice of h. On the other hand, note that the second-order Hessian $\nabla^2 R(\theta) = \boldsymbol{E}[\mathbf{x}\mathbf{x}^\top K_h(Y - \langle\theta, \mathbf{x}\rangle)]$ where $K_h(t) = K(t/h)/h \leq L/h$, and thus the scale of the Hessian depends directly on the smoothing parameter h, making its convergence heavily affected by the choice of h. The detailed proof of Lemma 3.2 is provided in the Appendix. We are now ready to present our main results regarding the solution of the optimization problem in (3) by investigating the empirical risk function $\hat{R}_n(\theta)$.

Theorem 3.3. *Assume Assumption 1 and that the parameter θ_0 satisfies $\|\theta_0\|_2 \leq r/3$. Let η_0 and η_1 be defined as in Theorem 3.1. Then for any $\pi > 0$ there exists a constant C_π that depends on L, r, π and τ but not on p, n, h, α, δ and the outlier distribution g satisfying $\{C_\pi(p\log n)/n\}^{1/2} \leq \{(1-\delta)\ell h^2\mu_2 + \delta\}\tau$ and $\{C_\pi(p\log n)/(nh^3)\}^{1/2} \leq 2(1-\delta)\ell\tau^3\eta_1$ for all large n so that the following statements hold with probability at least $1-\pi$.*

(a) $\langle\theta - \theta_0, \nabla\hat{R}_n(\theta)\rangle > 0$ *for all* $\|\theta - \theta_0\|_2 > 2\eta_0$.

(b) $\lambda_{\min}\{\nabla^2\hat{R}_n(\theta)\} \geq 2(1-\delta)\ell\tau^3\eta_1 > 0$ *for all* $\|\theta - \theta_0\|_2 < \eta_1$.

Therefore, as long as $2\eta_0 < \eta_1$, there is a unique stationary point of $\hat{R}_n(\theta)$ that lies in the closed ball $B_2^p(\theta_0, \eta_0(1+h))$ which also serves as the unique global optimal solution of (3). We denote this unique stationary point by $\hat{\theta}_n$. Then we have the following:

(c) *There exists a constant C that depends on L, r, π, τ, γ and ℓ but not on p, n, h, m_0, δ and the outlier distribution g so that*

$$\|\hat{\theta}_n - \theta_0\|_2 \leq \eta_0 + \frac{1}{(1-\delta)\ell\tau^3\eta_1}\sqrt{\frac{Cp\log n}{n}}. \tag{5}$$

(d) *There exist constants C_1, C_2 and $\eta_{\max}$ that depend on π, γ, r, η, h, ℓ, μ_1, μ_2, m_0 and δ but not on n, p and η so that the gradient descent algorithm with fixed step size $\eta \leq \eta_{\max}$ converges exponentially fast to the global minimizer, namely for any initialization $\theta_n(0) \in B_2^p(0, r)$ we have*

$$\|\theta_n(k) - \hat{\theta}_n\|_2^2 \leq C_1(1 - C_2\eta)^k\|\theta_n(0) - \hat{\theta}_n\|_2^2, \tag{6}$$

where $0 < 1 - C_2\eta < 1$.

We shall here provide some discussions on the results in Theorem 3.3. First, by Theorem 3.1 the radius $\eta_0 = m_0^{-2}\{C_1h^2 + C_2\delta/(1-\delta)\}$, and thus part (a) of Theorem 3.3 implies that, when the sample size n is large and the bandwidth h and contamination ratio δ are small, with high probability all stationary points of the empirical risk function $\hat{R}_n(\theta)$ are close to the true parameter θ_0. Second, by Theorem 3.3 (b), the empirical risk function $\hat{R}_n(\theta)$ is $2(1-\delta)\ell\tau^3\eta_1$-strongly convex in the closed ball $B_2^p(\theta_0, \eta_1)$. These serve as the empirical counterparts of the population results studied in Theorem 3.1, and are obtained by leveraging the uniform convergence of the sample gradient and Hessian established in Lemma 3.2. Third, by part (c) of Theorem 3.3, the convergence rate of the smoothed high-quantile regression estimator under

data contamination can depend on an interplay among how extremal the quantile level is, the amount of outliers in the data, and the tail distribution of the representative data. To provide an accessible illustration for this, we start with the simplest case when $\delta = 0$ and m_0 is a constant. In this case, by Theorem 3.3 (c), the estimation error satisfies $\|\hat{\theta}_n - \theta_0\|_2 \lesssim h^2 + \{(p \log n)/n\}^{1/2}$, where $a_n \lesssim b_n$ means that $a_n \leq C b_n$ for some constant $0 < C < \infty$. Therefore, if we set the bandwidth to be of order $h \asymp \{(p \log n)/n\}^{1/4}$, then $\|\hat{\theta}_n - \theta_0\|_2 \lesssim \{(p \log n)/n\}^{1/2}$ which coincides with the optimal result developed in the literature without an outlier He et al. (2023). We then consider the more complicated situation when there are outliers from an arbitrary distribution under the gross error model with $\delta > 0$. In this case, if we set the oracle $h \asymp \delta^{1/2}$, then the two constraints $\{C_\pi (p \log n)/n\}^{1/2} \leq \{(1-\delta)\ell h^2 \mu_2 + \delta\}\tau$ and $\{C_\pi (p \log n)/(nh^3)\}^{1/2} \leq 2(1-\delta)\ell\tau^3\eta_1$ will be automatically satisfied for all large n and by Theorem 3.3 (c) we have $\|\hat{\theta}_n - \theta_0\|_2 \lesssim \delta + \{(p \log n)/n\}^{1/2}$. Therefore, our result in this case early achieves the minimax rate $O(\delta + (p/n)^{1/2})$ of M-estimators under Huber's gross error model (Chen et al., 2016). We shall now consider the high-quantile regression setting when α is chosen close to zero under data contamination, for which we consider situations with the Gaussian tail and the generalized Pareto tail in Corollaries 3.4 and 3.5, respectively.

Corollary 3.4. *(Gaussian Case) Suppose the idealized auxiliary variable has a conditional Gaussian distribution with $f_0(\epsilon \mid \mathbf{x}) = (2\pi)^{-1/2} \exp\{-(\epsilon + c_\alpha)^2/2\}$ where c_α represents the $(1-\alpha)$-th quantile of a standard normal so that $\mathbf{P}(\epsilon > 0 \mid \mathbf{x}) = \alpha$.As $\alpha \to 0$, for any $\pi > 0$, by choosing the smoothing parameter $h \asymp \sqrt{\delta}$, if $\delta \lesssim \alpha^3$, and the sample size n be large enough such that $\{(p \log n)/n\}^{1/2} \lesssim \min(\delta, \alpha\delta^{3/4})$, there exists a constant C depending on L, r, π, τ but is independent of p, n, α, δ, and the outlier distribution g such that as n is large enough, with probability at least $1 - \pi$, $\hat{R}_n(\theta)$ has a unique stationary point $\hat{\theta}_n$ satisfying*

$$\|\hat{\theta}_n - \theta_0\|_2 \leq C \left(\frac{\delta}{\alpha^2} + \frac{1}{\alpha} \sqrt{\frac{p \log n}{n}} \right). \tag{7}$$

Corollary 3.5. *(Generalized Pareto Case) Suppose the idealized auxiliary variable has a conditional distribution from the generalized Pareto family with*

$$f_0(\epsilon \mid \mathbf{x}) = \begin{cases} \{1 + \beta(c'_\alpha + \epsilon)\}^{-1/\beta - 1} I(\epsilon \geq -c'_\alpha), & \text{if } \beta > 0; \\ \exp(-\epsilon + \log \alpha) I(\epsilon \geq \log \alpha), & \text{if } \beta = 0; \\ \{1 + \beta(c'_\alpha + \epsilon)\}^{-1/\beta - 1} I(-c'_\alpha \leq \epsilon \leq -1/\beta - c'_\alpha), & \text{if } \beta < 0, \end{cases}$$

where $c'_\alpha = (\alpha^{-\beta} - 1)/\beta$ if $\beta > 0$ and $c'_\alpha = -\log \alpha$ if $\beta = 0$. As special cases, the generalized Pareto family includes the uniform distribution when $\beta = -1$, the exponential distribution when $\beta = 0$, and the heavy-tailed Pareto distribution when $\beta > 0$. As $\alpha \to 0$, for any $\pi > 0$, by choosing the smoothing parameter $h \asymp \sqrt{\delta}$, if $\delta \lesssim \alpha^{3(1+\beta)}$, and the sample size n be large enough such that $\{(p \log n)/n\}^{1/2} \lesssim \min(\delta, \alpha^{1+\beta}\delta^{3/4})$, there exists a constant C depending on L, r, π, τ but is independent of $p, n, \alpha, \delta, \beta$, and the outlier distribution g such that as n is large enough,

with probability at least $1-\pi$, $\hat{R}_n(\theta)$ *has a unique stationary point* $\hat{\theta}_n$ *satisfying*

$$\|\hat{\theta}_n - \theta_0\|_2 \leq \frac{C}{\alpha^{1+\beta}} \left(\frac{\delta}{\alpha^{1+\beta}} + \sqrt{\frac{p \log n}{n}} \right). \tag{8}$$

By Corollaries 3.4 and 3.5, the convergence rate of smoothed high-quantile regression estimators under data contamination may no longer be universal for different quantile levels as in traditional quantile regression models. In the current high quantile setting, the convergence rate may depend on, in addition to the contamination ratio, an interplay between how extremal the quantile level is and the tail behavior of idealized samples. In particular, if the auxiliary variable of idealized samples exhibits a heavy-tailed behavior as in the Pareto case of Corollary 3.5 with $\beta > 0$, then the convergence rate can be slowed down by a polynomial of α when compared with the rate in Corollary 3.4 under Gaussian tails. We then consider the following extension to the high-dimensional case, for which an additional penalty term is needed to handle the curse of dimensionality.

3.2 Smoothed High-Quantile Regression in High Dimensions

For the high-quantile regression model (1), when the dimension of the feature vector $\mathbf{x}$ is comparable to or larger than the sample size n, the estimator that solves (3) as studied in Section 3.1 may start to become vulnerable to the curse of dimensionality. We shall now consider its high-dimensional generalization, which solves the following optimization problem with an additional penalty.

$$\begin{aligned} \underset{\theta}{\text{Minimize:}} \quad & \hat{L}_n(\theta) := \frac{1}{n} \sum_{i=1}^{n} \phi_h(y_i - \langle \theta, \mathbf{x}_i \rangle) + \lambda_n \|\theta\|_1 \\ \text{subject to:} \quad & \|\theta\|_2 \leq r. \end{aligned} \tag{9}$$

To study the penalized smoothed high-quantile regression estimator (9) in the high-dimensional setting, we assume the following additional condition on the feature vector $\mathbf{x}$.

Assumption 2. *The feature vector* $\mathbf{x}$ *has a probability density function in* $\mathbb{R}^p$*, and there exists a constant* $M > 1$ *that does not depend on the dimension* p *so that* $\|\mathbf{x}\|_\infty \leq M\tau$ *holds almost surely.*

We remark that Assumption 2 is used mainly to simplify the statement of our theoretical results in the high-dimensional case, and it can be relaxed to allow unbounded sub-Gaussian feature vectors. To be more specific, if the feature vectors $\mathbf{x}_i$, $i = 1, \ldots, n$, are independent τ^2-sub-Gaussian, then one can show that, with high probability, $\max_{1 \leq i \leq n} \|\mathbf{x}_i\|_\infty \leq C\tau\sqrt{\log(np)}$. As a result, Theorem 3.7 can be supplemented by taking a truncation at $M = C\sqrt{\log(np)}$, and the conclusions will continue to hold with an additional $\log(np)$ term. Therefore, for simplicity of the statement of Theorem 3.7, we consider the case when Assumption 2 holds. Throughout this section, we assume that the target underlying coefficient θ_0 is sparse in the sense that $\|\theta_0\|_0 \leq s_0$ for some s_0 that is much smaller than the dimension p. Lemma 3.6 provides the results on the gradient and second-order Hessian of the empirical risk function and is useful in proving Theorem 3.7.

Lemma 3.6. *Suppose Assumptions 1 and 2 hold. Then, for any $\pi > 0$,*

a *If $h < 1$, the sample directional gradient converges uniformly to the population directional gradient along the direction $(\theta - \theta_0)$. Namely, we have*

$$\boldsymbol{P}\left(\sup_{\theta \in B_2^p(r)\setminus\{0\}} \frac{|\langle \nabla \hat{R}_n(\theta) - \nabla R(\theta), \theta - \theta_0\rangle|}{\|\theta - \theta_0\|_1} \leq \frac{C_0}{h}\sqrt{\frac{\log(np)}{n}}\right) \geq 1 - \pi,$$

where C_0 is a constant depending on M, L, τ, r but independent of $s_0, h, p, n, m_0, \delta$, and the outlier distribution g.

b *As $n \geq \frac{C_2 s \log(np)}{h^3}$, the sample restricted Hessian converges uniformly to the population restricted Hessian in the set $B_2^p(r) \cap B_2^p(s)$ for any $s \leq p$. Namely, we have*

$$\boldsymbol{P}\left(\sup_{\substack{\theta \in B_2^p(r) \cap B_2^p(s),\\ \nu \in B_2^p(1) \cap B_0^p(s)}} |\langle \nu, \left(\nabla^2 \hat{R}_n(\theta) - \nabla^2 R(\theta)\right)\nu\rangle| \leq C_1\sqrt{\frac{s\log(np)}{nh^3}}\right) \geq 1 - \pi,$$

where C_1, C_2 are constants depending on L, r, π, τ but are independent of s, p, n, h, m_0, δ, and the outlier distribution g.

To ease the statement of the following theorem, we assumed that $n \leq p^{100}$, which is a technical assumption so that we can have the bound $\log(pn) \leq 101\log(p)$. Since the current section focuses on the high-dimensional setting when p is comparable to or larger than the sample size n, it may not be very intriguing to discuss the case when $n > p^{100}$. Theorem 3.7 provides the convergence rate of penalized smoothed high-quantile regression estimators under data contamination in the high-dimensional case.

Theorem 3.7. *Assume Assumptions 1 and 2 and that the parameter θ_0 satisfies $\|\theta_0\|_2 \leq r/3$ and $\|\theta_0\|_0 \leq s_0$. Let η_0 and η_1 be defined as in Theorem 3.1. Then there exist constants C_0, C_1, C_2 and C_3 that depend on M, L, τ^2, r, γ, π and ℓ but not on s_0, n, p, m_0, h, δ and the outlier distribution g such that, for any $nh^2 \geq C_0 m_0^{-4} s_0 \log p$ and $\lambda_n \geq C_1[\delta + h^2 + h^{-3/2}\{(\log p)/n\}^{1/2}]$, the followings hold with probability as least $1 - \pi$.*

(a) *All stationary points of the optimization problem in (9) satisfy*

$$\|\hat{\theta}_n - \theta_0\|_2 \leq \eta_0 + 4C_2 m_0^{-2} s_0^{1/2}\lambda_n.$$

(b) *For sufficiently large n with $n \geq C_3 h^{-5} m_0^{-4} s_0 (\log p)^2$ and $\eta_0 + 4C_2 m_0^{-2} s_0^{1/2}\lambda_n < \eta_1$, the optimization problem in (9) has a unique local stationary point which also serves as the global minimizer.*

We shall here provide a discussion on the implications of Theorem 3.7. For this, we first consider the simple baseline case when m_0 is treated as a constant as in the conventional quantile regression setting. In this case, if we set the bandwidth $h \asymp \{(\log p)/n\}^{1/7}$ and choose $\lambda_n \asymp \delta + \{(\log p)/n\}^{2/7}$, then by Theorem 3.7 the estimation error $\|\hat{\theta}_n - \theta_0\|_2 \lesssim s_0^{1/2}\left(\delta + \{(\log p)/n\}^{2/7}\right)$ holds with a high probability. As a result, it allows situations when the dimension p can grow

much faster than the sample size n. We then consider the high-quantile regression setting when α is chosen close to zero under data contamination, for which we consider situations with the Gaussian tail and the generalized Pareto tail in Corollaries 3.8 and 3.9 respectively.

Corollary 3.8. *(Gaussian Case) Suppose the idealized auxiliary variable has a conditional Gaussian distribution with* $f_0(\epsilon \mid \mathbf{x}) = (2\pi)^{-1/2}\exp\{-(\epsilon + c_\alpha)^2/2\}$ *where* c_α *represents the* $(1-\alpha)$*-th quantile of a standard normal so that* $\mathbf{P}(\epsilon > 0 \mid \mathbf{x}) = \alpha$*. If* $\lambda_n \asymp \delta + \{(\log p)/n\}^{2/7}$*,* $h \asymp \max(\sqrt{\delta}, \{(\log p)/n\}^{1/7})$*, then as* $\alpha \to 0$ *for any* $\pi > 0$ *there exists a constant* C *that depends on* L*,* r*,* π *and* τ *but not on* p*,* n*,* α*,* δ *and the outlier distribution* g *such that any stationary point of* $\hat{L}_n(\theta)$ *satisfying*

$$\|\hat{\theta}_n - \theta_0\|_2 \le C\frac{s_0^{1/2}}{\alpha^2}\left\{\delta + \left(\frac{\log p}{n}\right)^{2/7}\right\}$$

with probability at least $1-\pi$ *for all large* n*.*

Corollary 3.9. *(Generalized Pareto Case) Suppose the idealized auxiliary variable has a conditional distribution from the generalized Pareto family with*

$$f_0(\epsilon \mid \mathbf{x}) = \begin{cases} \{1+\beta(c'_\alpha+\epsilon)\}^{-1/\beta-1} I(\epsilon \ge -c'_\alpha), & \text{if } \beta > 0; \\ \exp(-\epsilon + \log\alpha) I(\epsilon \ge \log\alpha), & \text{if } \beta = 0; \\ \{1+\lambda(c'_\alpha+\epsilon)\}^{-1/\beta-1} I(-c_\alpha \le \epsilon \le -1/\beta - c_\alpha), & \text{if } \beta < 0, \end{cases}$$

where $c'_\alpha = (\alpha^{-\beta}-1)/\beta$ *if* $\beta > 0$ *and* $c'_\alpha = -\log\alpha$ *if* $\beta = 0$*. If* $h \asymp \max(\sqrt{\delta}, \{(\log p)/n\}^{1/7})$ *and* $\delta \le \alpha^{2\beta+2}$*, then as* $\alpha \to 0$ *for any* $\pi > 0$ *there exists a constant* C *that depends on* L*,* r*,* π *and* τ *but not on* p*,* n*,* α*,* δ *and the outlier distribution* g *such that any stationary points of* $\hat{L}_n(\theta)$ *satisfying*

$$\|\hat{\theta}_n - \theta_0\|_2 \le C\frac{s_0^{1/2}}{\alpha^{2(1+\beta)}}\left\{\delta + \left(\frac{\log p}{n}\right)^{2/7}\right\}$$

with probability at least $1-\pi$ *for all large* n*.*

Similar to the low-dimensional case in Section 3.1, the convergence rate of smoothed high-quantile regression estimators under data contamination in the high-dimensional case will continue to be affected by an interplay between how extremal the quantile level is and what tail behavior the idealized samples have. For example, if the auxiliary variable of idealized samples exhibits heavy tails as in the Pareto distribution, then the convergence rate of the corresponding smoothed high-quantile regression estimators can be slowed down by a polynomial of α when compared with the case of Gaussian tails. Our results serve as an exploratory step toward rigorously understanding the effect of a high quantile level under data contamination in both low and high dimensions.

4 Numerical Experiments

4.1 A Monte Carlo Simulation Study

In this section, we provide a Monte Carlo simulation study to investigate the finite-sample performance of the smoothed high-quantile regression estimator under data contamination in both low and high dimensions. For this, we consider the quantile regression model

$$y_i = \langle \theta_0, \mathbf{x}_i \rangle + \epsilon_i, \quad i = 1, \ldots, n, \tag{10}$$

where the regressors $\mathbf{x}_i$, $i = 1, \ldots, n$, are generated as independent standard multivariate normal random vectors and we set the true regression coefficient as $\theta_0 = (1, 2, 3, 0, \ldots, 0)^\top \in \mathbb{R}^p$ for $p \in \{5, 10, 100, 200\}$ so that it has a sparse structure in the high dimensional case. For the auxiliary variables ϵ_i, $i = 1, \ldots, n$, we use the gross error model of Huber (1964) and generate them from a mixture of idealized and contaminated samples. In particular, the idealized samples are assumed to follow a shifted standard normal distribution centered at its $(1-\alpha)$-th quantile, while the contaminated samples are assumed to occur with probability δ from a different normal distribution that has mean zero and variance 10. Let the sample size $n \in \{500, 1000\}$ and smoothing parameter $h \in \{0.25, 0.5, 0.75\}$, we consider applying the smoothed high-quantile regression estimator in (3) and its penalized version in (9) to three upper high quantiles $1-\alpha \in \{0.9, 0.95, 0.99\}$ and three lower tail quantiles $1-\alpha \in \{0.1, 0.05, 0.01\}$ under data contamination with $\delta \in \{0.001, 0.01, 0.05\}$. As a comparison, we also report results for the median regression. For each realization, we compute the smoothed high-quantile regression estimate in (3) by iterations of the projected gradient descent (PGD) algorithm, which updates at each iteration by

$$\hat{\theta}_n^{(k+1)} = \mathcal{P}_{B_2^p(r)}[\hat{\theta}_n^{(k)} - \text{step size} \times \nabla \hat{R}_n\{\hat{\theta}_n^{(k)}\}],$$

where $\mathcal{P}_{B_2^p(r)}$ denotes the projection onto the Euclidean ball of radius $r = 10$ in $\mathbb{R}^p$ and $\nabla \hat{R}_n\{\hat{\theta}_n^{(k)}\}$ is the corresponding gradient of the smoothed high-quantile regression criterion function. The estimate is then obtained by iteratively performing the PGD algorithm with a fixed step size of 1 until the relative difference is within 0.1% or $K = 300$ iterations. For the penalized optimization problem (9), we obtain a solution using a proximal PGD algorithm, which includes an additional soft-thresholding step after each gradient descent step but before the projection step. For the tuning parameter that controls the degree of penalization, we follow Tan et al. (2022) and consider the rule of thumb choice of $\lambda = 0.5\{\log(p)/n\}^{1/2}$. We in Tables 1–4 report results based on 100 replications for configurations when $h = 0.5$ and $\delta \in \{0.01, 0.05\}$, and a complete summary of simulation results can be found in the supplementary material.

From Tables 1–4, we can observe the followings. First, a higher contamination ratio typically leads to a larger relative mean squared error (RMSE) for the unpenalized estimate, where tail quantiles seem to be the most affected. For example, if we focus on the setting when $h = 0.5$, $n = 500$ and $p = 200$, then the RMSE of the unpenalized estimate at the $1-\alpha = 0.99$ high quantile can increase from 0.156 to 0.238 as the contamination ratio δ increases from 0.01 as in Table 1 to 0.05 as in Table 3. The RMSE for the median in the same setting, however, only increases from 0.055 to 0.061 and is much less affected. Compared with the unpenalized estimate,

Table 1: Relative mean squared errors (RMSE), average misclassification rate of zero variables (AMR_0) and average misclassification rate of nonzero variables (AMR_1) of the smoothed high-quantile regression estimate (3) and its penalized version (9) for different sample sizes, regressor dimensions and quantile levels when $h = 0.5$, $\delta = 0.01$ and $n = 500$.

		Unpenalized			Penalized		
p	$1-\alpha$	RMSE	AMR_0	AMR_1	RMSE	AMR_0	AMR_1
5	0.01	0.005	1.000	0.000	0.045	0.005	0.000
	0.05	0.002	1.000	0.000	0.012	0.005	0.000
	0.1	0.001	1.000	0.000	0.007	0.010	0.000
	0.5	0.001	1.000	0.000	0.002	0.110	0.000
	0.9	0.001	1.000	0.000	0.007	0.020	0.000
	0.95	0.002	1.000	0.000	0.013	0.005	0.000
	0.99	0.005	1.000	0.000	0.043	0.000	0.000
10	0.01	0.010	1.000	0.000	0.058	0.001	0.000
	0.05	0.004	1.000	0.000	0.017	0.001	0.000
	0.1	0.003	1.000	0.000	0.009	0.001	0.000
	0.5	0.002	1.000	0.000	0.003	0.050	0.000
	0.9	0.003	1.000	0.000	0.009	0.001	0.000
	0.95	0.004	1.000	0.000	0.017	0.001	0.000
	0.99	0.009	1.000	0.000	0.055	0.000	0.000
100	0.01	0.085	1.000	0.000	0.079	0.000	0.000
	0.05	0.046	1.000	0.000	0.031	0.000	0.000
	0.1	0.033	1.000	0.000	0.017	0.000	0.000
	0.5	0.020	1.000	0.000	0.005	0.006	0.000
	0.9	0.033	1.000	0.000	0.017	0.000	0.000
	0.95	0.045	1.000	0.000	0.030	0.000	0.000
	0.99	0.086	1.000	0.000	0.083	0.000	0.000
200	0.01	0.150	1.000	0.000	0.090	0.000	0.000
	0.05	0.094	1.000	0.000	0.034	0.000	0.000
	0.1	0.083	1.000	0.000	0.018	0.000	0.000
	0.5	0.055	1.000	0.000	0.006	0.003	0.000
	0.9	0.081	1.000	0.000	0.018	0.000	0.000
	0.95	0.094	1.000	0.000	0.035	0.000	0.000
	0.99	0.156	1.000	0.000	0.089	0.000	0.000

the performance of the penalized estimate seems to be more robust to the contamination ratio, where the RMSE generally does not change too much as we increase the contamination ratio δ from 0.01 as in Table 1 to 0.05 as in Table 3. Second, the performance of both unpenalized and penalized smoothed high-quantile regression estimates seems to exhibit a decay as the quantile level gets more and more extremal, which is now reflected by the newly developed theoretical results in Section 3 that allow the study of the effect of tail quantiles. Unlike the comparison made when we increase the contamination ratio, the performance of the penalized estimate in this case seems to be more sensitive to a more extremal quantile level. For example, if we focus on the setting when $h = 0.5$, $\delta = 0.01$, $n = 500$ and $p = 200$, then the RMSE of the penalized estimate at the $1 - \alpha = 0.99$ high quantile is 0.089 from Table 1, which is almost fifteen times the RMSE 0.006 at the median level. As a comparison, the RMSE of the unpenalized estimate at the $1 - \alpha = 0.99$ high quantile approximately only triples that at the median level. Third, the performance generally becomes worse as we increase the dimension, for which the penalized estimate seems to exhibit more robustness with respect to the increasing dimension. In addition, the penalized estimate seems to deliver a reasonably good performance in shrinking coefficient estimates of irrelevant variables to zero, as the reported average misclassification rates of zero variables (AMR_0) are all close to zero for the penalized estimate. Fourth, the performance of both unpenalized and penalized estimates improve as we increase the sample size, which

Table 2: Relative mean squared errors (RMSE), average misclassification rate of zero variables (AMR_0) and average misclassification rate of nonzero variables (AMR_1) of the smoothed high-quantile regression estimate (3) and its penalized version (9) for different sample sizes, regressor dimensions and quantile levels when $h = 0.5$, $\delta = 0.01$ and $n = 1000$.

		Unpenalized			Penalized		
p	$1-\alpha$	RMSE	AMR_0	AMR_1	RMSE	AMR_0	AMR_1
5	0.01	0.003	1.000	0.000	0.028	0.000	0.000
	0.05	0.001	1.000	0.000	0.007	0.000	0.000
	0.1	0.001	1.000	0.000	0.003	0.015	0.000
	0.5	0.000	1.000	0.000	0.001	0.095	0.000
	0.9	0.001	1.000	0.000	0.003	0.010	0.000
	0.95	0.001	1.000	0.000	0.007	0.000	0.000
	0.99	0.003	1.000	0.000	0.028	0.000	0.000
10	0.01	0.005	1.000	0.000	0.034	0.000	0.000
	0.05	0.002	1.000	0.000	0.010	0.000	0.000
	0.1	0.001	1.000	0.000	0.004	0.006	0.000
	0.5	0.001	1.000	0.000	0.001	0.057	0.000
	0.9	0.001	1.000	0.000	0.004	0.001	0.000
	0.95	0.002	1.000	0.000	0.010	0.000	0.000
	0.99	0.005	1.000	0.000	0.034	0.000	0.000
100	0.01	0.046	1.000	0.000	0.053	0.000	0.000
	0.05	0.021	1.000	0.000	0.016	0.000	0.000
	0.1	0.014	1.000	0.000	0.008	0.000	0.000
	0.5	0.009	1.000	0.000	0.002	0.005	0.000
	0.9	0.015	1.000	0.000	0.008	0.000	0.000
	0.95	0.020	1.000	0.000	0.016	0.000	0.000
	0.99	0.045	1.000	0.000	0.052	0.000	0.000
200	0.01	0.086	1.000	0.000	0.057	0.000	0.000
	0.05	0.045	1.000	0.000	0.019	0.000	0.000
	0.1	0.034	1.000	0.000	0.009	0.000	0.000
	0.5	0.021	1.000	0.000	0.003	0.003	0.000
	0.9	0.032	1.000	0.000	0.009	0.000	0.000
	0.95	0.045	1.000	0.000	0.019	0.000	0.000
	0.99	0.087	1.000	0.000	0.057	0.000	0.000

corroborates with our theoretical results developed in Section 3. Additional simulation results are provided in the supplementary material, where the aforementioned observations continue to hold.

5 A Real Data Illustration

We in this section provide a further illustration of the smoothed high-quantile regression estimator under data contamination by using a wine quality data. The data is available from the UC Irvine Machine Learning Repository at `https://archive.ics.uci.edu/dataset/186/wine+quality`, which contains physicochemical and sensory variables for red and white variants of the Portuguese Vinho Verde wine; see also Cortez et al. (2009) for a more detailed description of the data. Since the white wine data with 4898 observations constitute a major proportion, we treat them as the idealized samples and obtain a contaminated data by adding a random sample from the red wine data with $\delta = 0.05$. We consider applying the smoothed high-quantile regression estimate (3) and its penalized version (9) to study the effect of physicochemical variables (fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol) on high quantiles of the sensory variable to focus on wines of the highest quality. In particular, we consider three upper high quantiles

Table 3: Relative mean squared errors (RMSE), average misclassification rate of zero variables (AMR_0) and average misclassification rate of nonzero variables (AMR_1) of the smoothed high-quantile regression estimate (3) and its penalized version (9) for different sample sizes, regressor dimensions and quantile levels when $h = 0.5$, $\delta = 0.05$ and $n = 500$.

		Unpenalized			Penalized		
p	$1-\alpha$	RMSE	AMR_0	AMR_1	RMSE	AMR_0	AMR_1
5	0.01	0.009	1.000	0.000	0.042	0.005	0.000
	0.05	0.002	1.000	0.000	0.014	0.010	0.000
	0.1	0.002	1.000	0.000	0.007	0.015	0.000
	0.5	0.001	1.000	0.000	0.002	0.115	0.000
	0.9	0.002	1.000	0.000	0.007	0.040	0.000
	0.95	0.003	1.000	0.000	0.014	0.020	0.000
	0.99	0.009	1.000	0.000	0.046	0.010	0.000
10	0.01	0.015	1.000	0.000	0.054	0.000	0.000
	0.05	0.005	1.000	0.000	0.017	0.001	0.000
	0.1	0.003	1.000	0.000	0.009	0.004	0.000
	0.5	0.002	1.000	0.000	0.003	0.057	0.000
	0.9	0.003	1.000	0.000	0.010	0.007	0.000
	0.95	0.005	1.000	0.000	0.017	0.006	0.000
	0.99	0.016	1.000	0.000	0.053	0.001	0.000
100	0.01	0.163	1.000	0.000	0.083	0.000	0.000
	0.05	0.081	1.000	0.000	0.031	0.000	0.000
	0.1	0.048	1.000	0.000	0.017	0.000	0.000
	0.5	0.023	1.000	0.000	0.005	0.008	0.000
	0.9	0.047	1.000	0.000	0.017	0.000	0.000
	0.95	0.076	1.000	0.000	0.030	0.000	0.000
	0.99	0.162	1.000	0.000	0.081	0.000	0.003
200	0.01	0.238	1.000	0.000	0.090	0.000	0.000
	0.05	0.185	1.000	0.000	0.035	0.000	0.000
	0.1	0.127	1.000	0.000	0.020	0.000	0.000
	0.5	0.061	1.000	0.000	0.006	0.003	0.000
	0.9	0.134	1.000	0.000	0.018	0.000	0.000
	0.95	0.165	1.000	0.000	0.036	0.000	0.000
	0.99	0.238	1.000	0.000	0.086	0.000	0.000

$1-\alpha \in \{0.9, 0.95, 0.99\}$ that relate to wines of top 10%, 5% and 1% quality respectively. As a comparison, we also report results for the median when $1-\alpha = 0.5$. Following the mean analysis as in Cortez et al. (2009), explanatory variables are first standardized to have a zero mean and unit standard deviation before entering the regression. To compute the smoothed high-quantile regression estimate, we use $h = 0.5$ and iterate the PGD algorithm with a radius of $r = 10$ and a fixed step size of 1 until the relative difference is within 0.1% or $K = 300$ iterations. For the penalized estimate, we follow the setting in Section 4.1 and use the rule of thumb tuning parameter $\lambda = 0.5\{\log(p)/n\}^{1/2}$ from Tan et al. (2022). The results are summarized in Figures 1 and 2 for the unpenalized and penalized estimates respectively.

We can observe the followings from Figures 1 and 2. First, the regression analysis in the upper tail for high quality wines seems to differ a lot from that in the median for middle tier wines. For example, the fixed acidity has a negative impact on the taste of middle tier wines as the coefficient estimate is negative in both Figures 1 and 2 indicated by the thick lines at quantile level 0.5. However, such a negative effect vanishes in the penalized estimate and becomes positive in the unpenalized estimate for high quality wines as can be seen from the thick lines at quantile levels 0.9, 0.95 and 0.99. Therefore, although a higher level of fixed acidity may negatively impact the taste of middle tier wines, it may not matter or be even desirable in high quality wines. On the other hand, free sulfur dioxide and sulphates seem to positively improve

Table 4: Relative mean squared errors (RMSE), average misclassification rate of zero variables (AMR_0) and average misclassification rate of nonzero variables (AMR_1) of the smoothed high-quantile regression estimate (3) and its penalized version (9) for different sample sizes, regressor dimensions and quantile levels when $h = 0.5$, $\delta = 0.05$ and $n = 1000$.

		Unpenalized			Penalized		
p	$1-\alpha$	RMSE	AMR_0	AMR_1	RMSE	AMR_0	AMR_1
5	0.01	0.005	1.000	0.000	0.027	0.000	0.000
	0.05	0.001	1.000	0.000	0.008	0.015	0.000
	0.1	0.001	1.000	0.000	0.004	0.010	0.000
	0.5	0.000	1.000	0.000	0.001	0.095	0.000
	0.9	0.001	1.000	0.000	0.004	0.035	0.000
	0.95	0.001	1.000	0.000	0.007	0.000	0.000
	0.99	0.005	1.000	0.000	0.028	0.000	0.000
10	0.01	0.009	1.000	0.000	0.035	0.000	0.000
	0.05	0.002	1.000	0.000	0.009	0.003	0.000
	0.1	0.002	1.000	0.000	0.005	0.006	0.000
	0.5	0.001	1.000	0.000	0.001	0.054	0.000
	0.9	0.002	1.000	0.000	0.005	0.007	0.000
	0.95	0.003	1.000	0.000	0.010	0.000	0.000
	0.99	0.008	1.000	0.000	0.034	0.000	0.000
100	0.01	0.079	1.000	0.000	0.049	0.000	0.000
	0.05	0.030	1.000	0.000	0.017	0.000	0.000
	0.1	0.019	1.000	0.000	0.008	0.000	0.000
	0.5	0.010	1.000	0.000	0.002	0.007	0.000
	0.9	0.019	1.000	0.000	0.008	0.000	0.000
	0.95	0.031	1.000	0.000	0.017	0.000	0.000
	0.99	0.082	1.000	0.000	0.051	0.000	0.000
200	0.01	0.166	1.000	0.000	0.056	0.000	0.000
	0.05	0.076	1.000	0.000	0.018	0.000	0.000
	0.1	0.047	1.000	0.000	0.010	0.000	0.000
	0.5	0.022	1.000	0.000	0.003	0.004	0.000
	0.9	0.046	1.000	0.000	0.010	0.000	0.000
	0.95	0.076	1.000	0.000	0.019	0.000	0.000
	0.99	0.171	1.000	0.000	0.055	0.000	0.000

the taste of middle tier wines, while their positive effects become less substantial in effect or even vanish for high quality wines. As a result, it seems desirable to increase the level of both free sulfur dioxide and sulphates in the production of middle tier wines, while these two variables may not be that important in the production of high quality wines. In addition, the inclusion of data contamination can lead to a bias in coefficient estimates, and the amount of bias in tail quantiles seems to be comparable to that in the median. In addition, the direction of the bias is not necessarily the same for the tail when compared with the median. For example, it can be seen from Figure 2 that the inclusion of data contamination raises the coefficient estimate of the volatile acidity in the median regression while lowering it in the 0.99 tail quantile regression. This suggests that the effect difference of volatile acidity on red and white variants can be different in high quality wines compared to that in the in the median for middle tier wines. Furthermore, the estimates at different quantile levels seem to exhibit different degrees of sensitivity to data contamination for certain sensory variables, as the box plots tend to possess different widths across different quantile levels. This indicates that the difference between red and white variants are inhomogeneous across different quantile levels beyond the center, and therefore the effect of data contamination can be different in high quality wines from that in middle tier wines in multiple aspects of the underlying distribution.

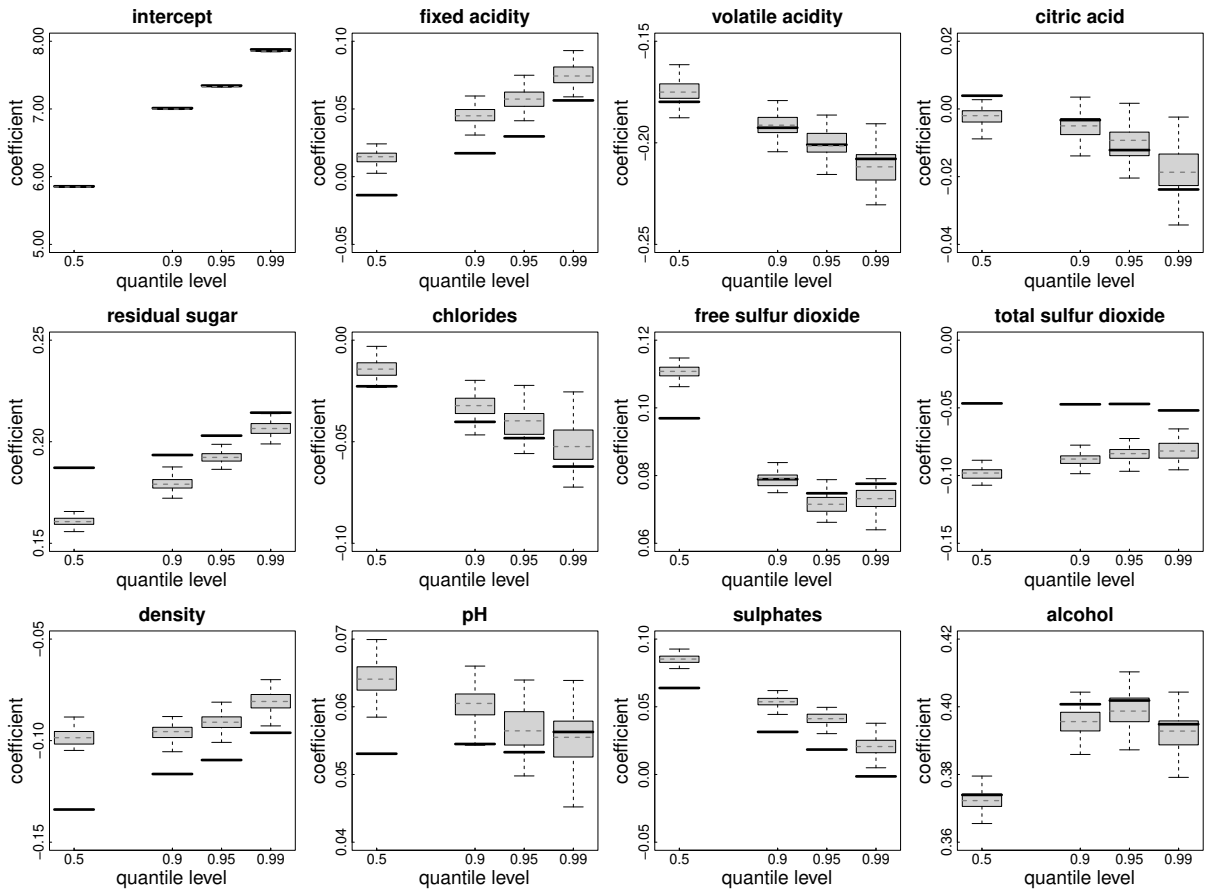

Figure 1: Smoothed high-quantile regression estimates at $1-\alpha \in \{0.9, 0.95, 0.99\}$ quantile levels for different physicochemical variables when the white wine data is contaminated by a random sample from the red wine. The results are summarized using box plots at each quantile level, where the thick lines represent the idealized situation as if no data contamination is present. For comparison, results for the median are also provided.

6 Conclusion

In this paper, we explore high-quantile regression in low and high dimensions, allowing the quantile level to be chosen close to the boundary to study tail phenomena, and when the data are contaminated with samples from an arbitrary, different distribution. To facilitate efficient computation via first-order gradient-based algorithms, we consider the use of a convolution-type kernel smoothing with and without a penalty function. Our analytical results developed in Sections 3.1 and 3.2 are among the first to understand and quantify the joint effect of the tail quantile level and the data contamination ratio on quantile regression estimators in both low and high dimensions. In particular, we establish uniform convergence results for the gradient and Hessian of the empirical risthose of the population risk in the current high-quantilentile setting with data contamination. Compared with Mei et al. (2018), which focused on linear regression models and relied critically on the assumption of symmetric error distributions, the current high-quantile regression setting requires novel techniques to handle auxiliary variables that are fundamentally asymmetric and may depend on the explanatory variables. We in addition examine the computational tractability of the smoothed quantile regression estimator under the gross error model. Unlike existing results for traditional linear regression models, the com-

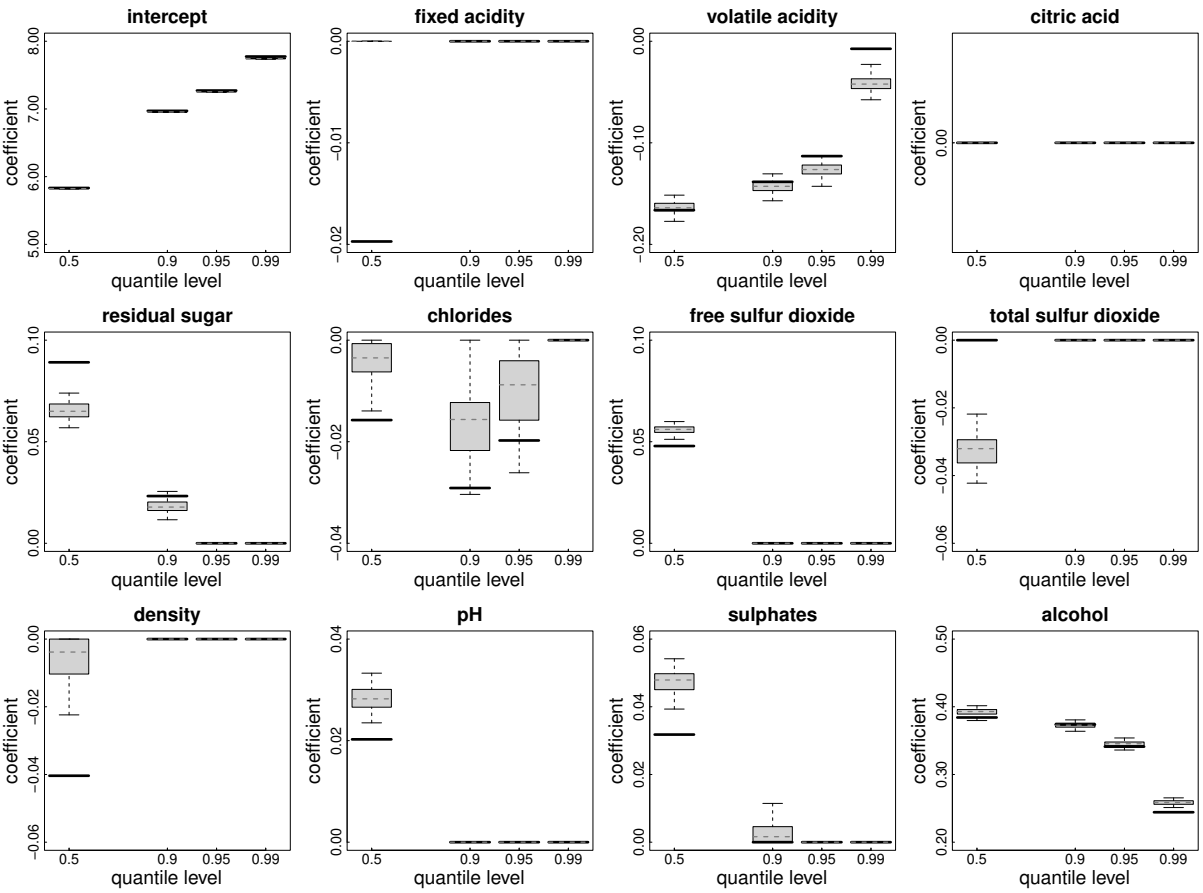

Figure 2: Penalized smoothed high-quantile regression estimates at $1-\alpha \in \{0.9, 0.95, 0.99\}$ quantile levels for different physicochemical variables when the white wine data is contaminated by a random sample from the red wine. The results are summarized using box plots at each quantile level, where the thick lines represent the idealized situation as if no data contamination is present. For comparison, results for the median are also provided.

putational tractability in the current high-quantile regression setting is shown to be determined by an interplay between the degree of data contamination and the extremeness of the target tail quantile. In particular, our results reveal that not only the quality but also the computational efficiency can be affected by a more extremal quantile under data contamination, which is an interesting and somewhat surprising discovery. Also, we develop nonasymptotic upper bounds of the estimation error in both low- and high-dimensional cases, which provide explicit characterizations on how the convergence rates are affected by the sample size, dimension of explanatory variables, the degree of data contamination, the extremeness of tail quantile, the tail distribution of idealized samples, and how they interplay with each other. We illustrate the results through both synesthetic data and a wine quality data, in which it is desirable to consider high quantiles to study the effect of different physicochemical variables on high qIt can be seen from our analysis that, compared with existing results on the mean or median that are more related to middle-tier wines, the way high-quality wines depend on physicochemical variables can be different in multiple aspects.tiple aspects. Therefore, our results can be potentially useful for the wine industry to improve the production procehigher-quality higher quality wines.

Supplementary Material

The online Supplementary Material contains technical proofs for our main results in Sections 3.1 and 3.2, detailed descriptions of the algorithms used to compute the smoothed high-quantile regression estimates in both low and high dimensions, and additional simulation results complementing those reported in Section 4.

References

Abrevaya, J. (2001). The effects of demographics and maternal behavior on the distribution of birth outcomes. *Empirical Economics*, **26**, 247–257.

Akrami, H., Zamzam, O., Joshi, A., Aydore, S. and Leahy, R. (2024). Beta quantile regression for robust estimation of uncertainty in the presence of outliers. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 7480–7484.

Bai, Z. D., Rao, C. R. and Wu, Y. (1992). M-estimation of multivariate linear regression parameters under a convex discrepancy function. *Statistica Sinica*, **2**, 237–254.

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

Chen, M., Gao, C. and Ren, Z. (2016). A general decision theory for huber's ϵ-contamination model. *Electronic Journal of Statistics*, **10**, 3752–3774.

Chernozhukov, V. (2005). Extremal quantile regression. *The Annals of Statistics*, **33**, 806–839.

Cortez, P., Cerdeira, A., Almeida, F., Matos, T. and Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, **47**, 547–553.

Elsner, J. B., Kossin, J. P. and Jagger, T. H. (2008). The increasing intensity of the strongest tropical cyclones. *Nature*, **455**, 92–95.

Gutenbrunner, C. and Jureckova, J. (1992). Regression rank scores and regression quantiles. *The Annals of Statistics*, **20**, 305–330.

He, F., Cheng, Y. and Tong, T. (2016a). Estimation of extreme conditional quantiles through an extrapolation of intermediate regression quantiles. *Statistics & Probability Letters*, **113**, 30–37.

He, F., Cheng, Y. and Tong, T. (2016b). Estimation of high conditional quantiles using the hill estimator of the tail index. *Journal of Statistical Planning and Inference*, **176**, 64–77.

He, X. (1997). Quantile curves without crossing. *The American Statistician*, **51**, 186–192.

He, X., Pan, X., Tan, K. M. and Zhou, W.-X. (2023). Smoothed quantile regression with large-scale inference. *Journal of Econometrics*, **232**, 367–388.

He, X. and Shao, Q.-M. (1996). A general bahadur representation of m-estimators and its application to linear regression with nonstochastic designs. *The Annals of Statistics*, **24**, 2608–2630.

He, X. and Shao, Q.-M. (2000). On parameters of increasing dimensions. *Journal of multivariate analysis*, **73**, 120–135.

Horowitz, J. L. (1998). Bootstrap methods for median regression models. *Econometrica*, **66**, 1327–1351.

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, **35**, 73–101.

Kluin, R. J. C., Kemper, K., Kuilman, T., de Ruiter, J. R., Iyer, V., Forment, J. V., Cornelissen-Steijger, P., de Rink, I., ter Brugge, P., Song, J.-Y., Klarenbeek, S., McDermott, U., Jonkers, J., Velds, A., Adams, D. J., Peeper, D. S. and Krijgsman, O. (2018). Xenofilter: computational deconvolution of mouse and human reads in tumor xenograft sequence data. *BMC Bioinformatics*, **19**.

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

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

Ling, W., Lu, J., Zhao, N., Lulla, A., Plantinga, A. M., Fu, W., Zhang, A., Liu, H., Song, H., Li, Z., Chen, J., Randolph, T. W., Koay, W. L. A., White, J. R., Launer, L. J., Fodor, A. A., Meyer, K. A. and Wu, M. C. (2022). Batch effects removal for microbiome data via conditional quantile regression. *Nature Communications*, **13**.

Mei, S., Bai, Y. and Montanari, A. (2018). The landscape of empirical risk for nonconvex losses. *The Annals of Statistics*, **46**, 2747–2774.

Punzo, A. and McNicholas, P. D. (2016). Parsimonious mixtures of multivariate contaminated normal distributions. *Biometrical Journal*, **58**, 1506–1537.

Suhaila, M., Chand, S. and Kibria, B. M. G. (2021). Quantile-based robust ridge m-estimator for linear regression model in presence of multicollinearity and outliers. *Communications in Statistics - Simulation and Computation*, **50**, 3194–3206.

Tan, K. M., Wang, L. and Zhou, W.-X. (2022). High-dimensional quantile regression: Convolution smoothing and concave regularization. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **84**, 205–233.

Tsai, I.-C. (2012). The relationship between stock price index and exchange rate in asian markets: A quantile regression approach. *Journal of International Financial Markets, Institutions and Money*, **22**, 609–621.

Wang, H. J. and Li, D. (2013). Estimation of extreme conditional quantiles through power transformation. *Journal of the American Statistical Association*, **108**, 1062–1074.

Wang, K. and Wang, H. J. (2016). Optimally combined estimation for tail quantile regression. *Statistica Sinica*, **26**, 295–311.

Wu, W. B. (2007). M-estimation of linear models with dependent errors. *The Annals of Statistics*, **35**, 495–521.

Zhang, R. (2025). Robust sequential change detection: The approach based on breakdown points and influence functions. *IEEE Transactions on Information Theory*, **71**, 5620–5632.

Zhang, R., Mei, Y., Shi, J. and Xu, H. (2022). Robustness and tractability for nonconvex M-estimators. *Statistica Sinica*, **32**, 1295–1316.

Zhang, T. (2021). High-quantile regression for tail-dependent time series. *Biometrika*, **108**, 113–126.

Zhang, Y. (2018). Extremal quantile treatment effects. *The Annals of Statistics*, **46**, 3707–3740.

Zhou, Z. and Shao, X. (2013). Inference for linear models with dependent errors. *Journal of*

the Royal Statistical Society: Series B (Statistical Methodology), **75**, 323–343.