Improved Local Quantile Regression

Xi Liu [1], Keming Yu [1], and Xueqing Tang [2]

[1] Department of Mathematics, Brunel University London, London, UB8 3PH, UK

[2] School of Science, Jiangnan University, Wuxi, Jiangsu, China.

Address for correspondence: Keming Yu, Department of Mathematics, Brunel University London, London, UB8 3PH, UK.

E-mail: `Keming.Yu@brunel.ac.uk`.

Phone: (+440) 1895 266128.

Fax: (+440) 1895 266128.

Abstract: We investigate a new kernel-weighted likelihood smoothing quantile regression method. The likelihood is based on a normal scale-mixture representation of asymmetric Laplace distribution (ALD). This approach enjoys the same good design adaptation as the local quantile regression (Spokoiny et al., 2013), particularly for smoothing extreme quantile curves, and ensures non-crossing quantile curves for any given sample. The performance of the proposed method is evaluated via extensive Monte Carlo simulation studies and one real data analysis.

Key words: Bandwidth Selection; Nonparametric Quantile Regression; Quantile Crossing

1 Introduction

Parametric quantile regression (Koenker, 2005) has been used in a number of disciplines to explore the relationship between the response and covariates at both the center and extremes of the conditional distribution and obtain a more comprehensive analysis of the relationship between variables. While a parametric model is possibly misspecified, non-parametric models, on the other hand, require fewer assumptions about the data and offer a more flexible way of modelling a relationship than parametric models, consequently avoid model misspecification when a parametric model is not available, which is common in wide applications (Wand and Jones, 1995; Fan and Gijbels, 1996a; Takezawa, 2005). One of the popular nonparametric smoothing techniques is kernel smoothing. Nonparametric kernel smoothing quantile regression has attracted much attention in the literature (Chaudhuri, 1991; Hardle and Mammen, 1993; Fan and Gijbels, 1996a; Yu and Jones, 1998; Cai and Xu, 2008; Dette and Volgushev, 2008; Dabo-Niang and Laksaci, 2012; Schaumburg, 2012; Kong and Xia, 2015; among others).

However, the performance of kernel smoothing techniques, in spite of their advantages over parametric models in dealing with model misspecification, depends on smoothing parameter or bandwidth selection. While a global bandwidth such as the rule of thumb (Yu and Jones, 1998) is generally useful, a point-wise bandwidth, which depends on the values of covariate X or the design set should be considered for the complexity of the underlying regression functions. In particular, bandwidth selection in nonparametric smoothing quantile regression requires not only design adaptation but also quantile adaptation. Spokoiny, Wang and H$\ddot{a}$rdle (henceforth SWH) (Spokoiny et al., 2013) developed a kernel-weighted likelihood quantile regression with point-wise bandwidth selection and promising performance in practice.

But SWH's approach may not guarantee non-crossing quantile curves for any given sample (calculated for various percentile $\tau \in (0,1)$), which is a common problem in the estimation of conditional and structural quantile functions due to lack of monotonicity. Note that, monotonicity (for each x in the design set, it's a monotone function of percentile value τ) guarantees non-crossing quantile curves, but not vice versa. Such a phenomenon violates the basic principle of probability theory, that is, the associated distribution functions should be monotone increasing. Various methods were presented to address or avoid the quantile crossing in parametric quantile regression, but with few on nonparametric quantile regression. Recently, Jones and Yu (2007b) improved double kernel smoothing for quantile regression, Using spline-based constraints easily allows us to incorporate non-crossing conditions, as in Bondell et al. (2010) or Muggeo et al. (2013), for quantile estimation. Liu and Wu (2011) dealt with this issue via simultaneous multiple quantile smoothing, Qu and Yoon (2015) applied inequality constrains to ensure the monotonicity over quantiles.

In this paper, we explore a local quantile regression based on a normal scale-mixture representation of asymmetric Laplace distribution (ALD) and show that this method has the similar property of SWH's procedure but much better-adaptive for smoothing extreme quantile curves. Moreover, quantile function is monotone with respect to τ for all x, which is satisfied by the proposed method, but SWH's method, which may also be non-crossing practically but without theoretical justification. Therefore, the proposed method enjoys both design adaptation and non-crossing quantile curves simultaneously. This paper is organized as follows. We first review SWH's approach in Section 2, then propose a new local likelihood smoothing based on a normal scale-mixture representation of ALD and show that this approach satisfies the propagation condition (Spokoiny and Vial, 2009) in Section 3. In Section 4 we elaborate the proposed adaptive bandwidth selection rule and point out that the rule is able to avoid the problem of quantile curves crossing, especially for estimating extreme quantiles. Section 5 illustrates the numerical performance of the proposed method. Section 6 provides concluding remarks and discusses future work.

2 Kernel-Weighted Likelihood for Local Quantile Regression

Spokoiny et al. (2013) developed an interesting nonparametric quantile regression method: local quantile regression, which provides point-wise bandwidth selection and exhibits promising performance in practice. SWH claimed that their bandwidth selection rule is adaptive and novel, although the regression estimator named qMLE in their Eq.(8) is simply equivalent to a local polynomial quantile regression or a type of kernel-based weighting 'check function' approach, such as the local linear single-kernel approach of Yu and Jones (1998).

Let (X, Y) be the random variables, where Y is a continuous random variable and X is a univariate regressor $X \in \mathbb{R}^1$. Let $F_Y\left(Y|X\right)$ be the cumulative distribution function of Y given X. Let $Q_\tau\left(Y|X\right) = \inf\left\{Y : F_Y\left(Y|X\right) \geq \tau\right\}$ be the inverse function, which is also the value of a that minimizes the expected loss function:

$$Q_\tau\left(Y|X\right) = \operatorname*{argmin}_a E\rho_\tau\left(Y - a\right), \tag{2.1}$$

where, $\tau \in (0,1)$ and $\rho_\tau(\cdot)$ is an asymmetric loss function that satisfies $\rho_\tau(u) =$

$u\left(\tau - I(u<0)\right)$ with $I(\cdot)$ is an indicator function.

Under the quantile non-parametric model $Y = f(X) + \varepsilon$, given data in the form $\{X_i, Y_i\}_{i=1}^{n}$, where X_i and Y_i are independent scalar observations of X and Y, respectively. The τth conditional quantile of Y given X is estimated by

$$\hat{f}(x) = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\tau \left(Y_i - f(X_i)\right). \tag{2.2}$$

SWH took advantage of the link between the minimization of the sum of the loss function in Eq.(2.2) and the maximum likelihood theory given by the asymmetric Laplace distribution. For a random variable $Y \sim \text{ALD}(\mu, \sigma, \tau)$, its density function can be written as

$$f(y;\mu,\sigma,\tau) = \frac{\tau(1-\tau)}{\sigma} \exp\left\{ \frac{y-\mu}{\sigma}\left[\tau - \text{I}(y \le \mu)\right]\right\}, \qquad y \in (-\infty, +\infty) \tag{2.3}$$

where, $0 < \tau < 1$ is skew parameter, $\sigma > 0$ is scale parameter, and $-\infty < \mu < \infty$ is location parameter.

Based on an ALD log-likelihood, SWH considered

$$L_{SWH}(\boldsymbol{\theta}) \equiv \log\left\{\tau(1-\tau)\right\} \sum_{i=1}^{n} I - \sum_{i=1}^{n} \rho_\tau \left(Y_i - f_\theta(X_i)\right), \tag{2.4}$$

with $0 < \tau < 1$ is the level of the quantile. Then they fit $f(x)$ at point x by the local polynomial approach $Y_i = \boldsymbol{\psi}_i^T \boldsymbol{\theta} + \epsilon$, with basis $\boldsymbol{\psi}_i = \{1, (X_i - x), (X_i - x)^2/2!, \cdots, (X_i - x)^p/p!\}^T$ and $\boldsymbol{\theta} = (\theta_0, ..., \theta_p)^T$. Therefore, the local log-likelihood at x is given by

$$L_{SWH}(W, \boldsymbol{\theta}) \equiv \log \tau(1-\tau) \sum_{i=1}^{n} w_i - \sum_{i=1}^{n} \rho_\tau \left(Y_i - \boldsymbol{\psi}_i^T \boldsymbol{\theta}\right) w_i, \tag{2.5}$$

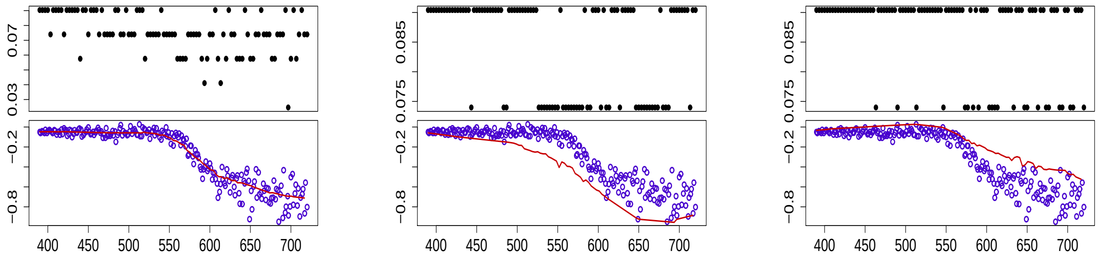
where the weights W is chosen via a kernel function $w_i = K\left(\frac{X_i - x}{h}\right)$, while h is a bandwidth controlling the degree of localization. Note that, Eq.(2.5) is similar to the global log-likelihood in Eq.(2.4), but each summand in $L_{SWH}(W, \boldsymbol{\theta})$ is multiplied with the weight w_i, so only the points from the local vicinity of x contribute to $L_{SWH}(W, \boldsymbol{\theta})$.

The corresponding local quantile MLE (they named it as qMLE) at x is then given via the maximization of $L_{SWH}(W, \boldsymbol{\theta})$ in Eq.(2.4)

$$\begin{aligned} \tilde{\boldsymbol{\theta}}_{SWH}(x) &\equiv \operatorname*{argmax}_{\theta \in \Theta} L_{SWH}(W, \boldsymbol{\theta}) \\ &= \operatorname*{argmin}_{\theta \in \Theta} \sum_{i=1}^{n} \rho_\tau \left(Y_i - \boldsymbol{\psi}_i^T \boldsymbol{\theta} \right) w_i. \end{aligned} \tag{2.6}$$

3 An Alternative Likelihood for Local Quantile Regression

Figure 1a displays the performance of SWH's approach, showing the bandwidth sequence (upper panel) and the smoothed 50% quantile curve (lower panel) based on the Lidar dataset (available in *R* package *'SemiPar'*), which adapts the data well. And this is also true for other moderate or central quantile curves. However, it can be seen from smoothing extreme quantile curves in Figure 1 here, the proposed bandwidth selection rule is lack of good adaptation and then results in the over-smoothing phenomenon. Figures 1b and 1c display the smoothed 1% and 99% quantile curves using SWH's method and shows that when the curves start to switch smoothness, the rule is not adaptive so that the estimated curves are too smoothing out of the data ranges. A possibly theoretical interpretation for this problem is: when $\tau \to 0$, the weighted 'check function' $\rho_\tau(Y_i - \boldsymbol{\psi}_i^T \boldsymbol{\theta}) w_i$ takes constant 0 if $Y_i > \boldsymbol{\psi}_i^T \boldsymbol{\theta}$ (also, when $\tau \to 1$ and if $Y_i < \boldsymbol{\psi}_i^T \boldsymbol{\theta}$). This may result in that the proposed significant test always picks constant bandwidth for smoothing extreme quantile curves although this is not a problem for the local quantile regression estimation equation. We want to point out that this over-smoothing problem will be solved by a new version of adaptive bandwidth selection rule.



(a) $\tau = 0.50$ (b) $\tau = 0.01$ (c) $\tau = 0.99$

Figure 1: The bandwidth sequences (upper panels) and smoothed quantile curves (lower panels) for the Lidar dataset using SWH's kernel-weighted likelihood.

Moreover, there is no guaranteed of this approach to avoid quantile crossing. Therefore we propose an alternative adaptive bandwidth selection rule based on a normal scale-mixture representation (henceforth NSM) of ALD and show that this alternative version has the similar property of SWH's procedure but much better-adaptive for smoothing extreme quantile curves.

Reed and Yu (2010) and Kozumi and Kobayashi (2011) noted that under the assumption of ALD-based 'working likelihood', the quantile regression model error $\epsilon \sim \text{ALD}(0, 1, \tau)$ can be represented as a scale mixture of normal variables, that is,

$$\epsilon = \mu z + \delta\sqrt{z}e, \tag{3.1}$$

where $\mu = \frac{1-2\tau}{\tau(1-\tau)}$, $\delta^2 = \frac{2}{\tau(1-\tau)}$, $z \sim Exp(1)$ and $e \sim N(0, 1)$, and z and e are independent. Hence, SWH's model (1) ($Y_i = f(X_i) + \epsilon_i$) could be re-written as

$$Y_i = f(X_i) + \mu z_i + \delta\sqrt{z_i}e_i. \tag{3.2}$$

That is, for given $\boldsymbol{z} = (z_1, z_2, ..., z_n)$,

$$Y_i \sim N\left(f(X_i) + \mu z_i,\ \delta^2 z_i\right), \tag{3.3}$$

i.e., the joint conditional density of $Y = (Y_1, Y_2, ..., Y_n)$ is given by

$$l\left(Y|\boldsymbol{z}, X\right) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\,\delta\sqrt{z_i}} \exp\left\{-\frac{\left(Y_i - f(X_i) - \mu z_i\right)^2}{2\delta^2 z_i}\right\}. \tag{3.4}$$

Clearly, if $\boldsymbol{z}$ is fixed in advance, then the local log-likelihood (SWH's Eq.(7)) can be replaced by a normal scale-mixture representation of ALD :

$$\begin{aligned} L_{NSM}(W, \boldsymbol{\theta}) &\equiv -\log(\sqrt{2\pi}\delta)\sum_{i=1}^{n} w_i - \frac{1}{2}\sum_{i=1}^{n}\log(z_i)\,w_i \\ &- \frac{1}{2\delta^2}\sum_{i=1}^{n}\frac{(Y_i - f(X_i) - \mu z_i)^2}{z_i}w_i - \sum_{i=1}^{n} z_i w_i, \end{aligned} \tag{3.5}$$

where the weights W is chosen via a kernel function $w_i = K\left(\frac{X_i - x}{h}\right)$, while h is a bandwidth controlling the degree of localization. Similar to Eq.(2.5), the local log-likelihood in Eq.(3.5) depends on the central point x via the structure of the basis vectors $\boldsymbol{\psi}_i$ and via the weights w_i.

Now, once a local pth-degree polynomial $\boldsymbol{\psi}_i^T\boldsymbol{\theta}$ is used to approximate $f(x)$ at $X = x$, the corresponding local qMLE at x could be defined via maximization of $L_{NSM}(W, \boldsymbol{\theta})$ above:

$$\begin{aligned} \tilde{\boldsymbol{\theta}}(x) &\equiv \left(\tilde{\theta}_0(x), \tilde{\theta}_1(x), ..., \tilde{\theta}_p(x)\right) \\ &= \operatorname*{argmax}_{\theta\in\Theta} L_{NSM}(W, \boldsymbol{\theta}) \\ &= \operatorname*{argmin}_{\theta\in\Theta} \sum_{i=1}^{n}\frac{(Y_i - \boldsymbol{\psi}_i^T\boldsymbol{\theta} - \mu z_i)^2}{\delta^2 z_i}w_i, \end{aligned} \tag{3.6}$$

where $\tilde{\theta}_0(x)$ estimates $f(x)$, and $\tilde{\theta}_m(x)$ estimates the m^{th} derivative of $f(x)$. Further, let $\boldsymbol{\psi} = (\boldsymbol{\psi}_1, .., \boldsymbol{\psi}_n)^T$ and $\boldsymbol{w}_k = diag\left(\frac{w_1^{(k)}}{\delta^2 z_1}, ..., \frac{w_n^{(k)}}{\delta^2 z_n}\right)$, we have

$$\tilde{\boldsymbol{\theta}}_k(x) = \left(\boldsymbol{\psi}\boldsymbol{w}_k\boldsymbol{\psi}^T\right)^{-1} \boldsymbol{\psi}\boldsymbol{w}_k \left(Y + \mu\boldsymbol{z} + \delta\boldsymbol{z}^{1/2}\boldsymbol{e}\right), \tag{3.7}$$

where the design matrix $\boldsymbol{\psi}$ consists of the columns $\boldsymbol{\psi}_i = \{1, (X_i - x), \cdots, (X_i - x)^p/p!\}^T$.

We note that the $L_{NSM}(W, \boldsymbol{\theta})$ involves in a specification of vector $\boldsymbol{z}$, and we point out that $\boldsymbol{z}$ could be fixed in advance via a sample from a data-driven inverse Gaussian distribution, and our extensive experiments in Section 5 show that the selection of the sample has no effect on the estimation. In fact, note that the joint likelihood function of $(Y, \boldsymbol{z})$ is given by

$$f(Y, \boldsymbol{z}|X) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\,\tau\sqrt{z_i}} \exp\left\{-\frac{(Y_i - f(X_i) - \mu z_i)^2}{2\tau^2 z_i}\right\} \prod_{i=1}^{n} \exp(-z_i).$$

Therefore, the conditional density of $f(\boldsymbol{z}|Y)$ is given by

$$\begin{aligned} f(\boldsymbol{z}|Y) &\propto f(Y, \boldsymbol{z}) \\ &\propto \prod_{i=1}^{n} \frac{1}{\sqrt{z_i}} \exp\left(-\frac{1}{2}\left[\frac{(Y_i - f(X_i))^2}{\delta^2} z_i^{-1} + \left(\frac{\mu^2}{\delta^2} + 2\right) z_i\right]\right). \end{aligned} \tag{3.8}$$

That is, $z_i, z_2, ..., z_n$ are i.i.d. with a generalized inverse Gaussian (GIG) distribution:

$$\begin{aligned} f(\boldsymbol{z}|Y) &\propto z_i^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\left[\frac{(Y_i - f(X_i))^2}{\delta^2} z_i^{-1} + \left(\frac{\mu^2}{\delta^2} + 2\right) z_i\right]\right) \\ &\sim GIG\left(\frac{1}{2}, \eta_i, \zeta_i\right), \end{aligned} \tag{3.9}$$

where $\eta_i^2 = \frac{(Y_i - f(X_i))^2}{\delta^2}$ and $\zeta_i^2 = \frac{\mu^2}{\delta^2} + 2$.

4 Performance of Adaptive Bandwidth Selection and Non-Crossing Estimation

4.1 Adaptive Bandwidth Selection

There are several methodologies for automatic smoothing parameter selection. One class of methods chooses the smoothing parameter value to minimize a criterion that incorporates both the tightness of the fit and model complexity. Such a criterion can usually be written as a function of the error mean square, and a penalty function designed to decrease with increasing smoothness of the fit. Examples of specific criteria are generalized cross-validation (Craven and Wahba, 1979) and the Akaike information criterion (AIC)(Akaike, 1973). These classical selectors have two undesirable properties when used with local polynomial and kernel estimators: they tend to under-smooth and tend to be non-robust in the sense that small variations in the input data can change the choice of smoothing parameter value significantly. Hurvich et al. (1998) obtained several bias-corrected AIC criteria that limit these unfavorable properties and perform comparably with the plug-in selectors (Ruppert et al., 1995).

The adaptive bandwidth selection rule in SWH's paper is different from the rule-of-thumb rule of Yu and Jones (1998) and AIC rule of Cai and Xu (2008). It does add a nice option to the bandwidth selection menu for practitioners. In this paper, we perform the local quantile curve estimation following the similar bandwidth selection procedures, but based on a normal scale-mixture representation of ALD.

First, we fix a finite ordered set of candidates of bandwidth $h_1 < h_2 < \cdots < h_K$, where h_1 is very small. According to SWH, the bandwidth sequence can be taken geometrically increasing of the form $h_k = ab^k$ with fixed $a > 0$, $b > 1$, and $n^{-1} < ab^k < 1$ for $k = 1, \cdots, K$. For each $k \leq K$, an ordered weighting scheme $W^{(k)} = \left(w_1^{(k)}, w_2^{(k)}, \cdots, w_n^{(k)}\right)$ is chosen via a kernel function $w_i^{(k)} = K\left(\frac{X_i - x}{h_k}\right)$ leading to the local quantile estimator at x, $\tilde{\boldsymbol{\theta}}_k(x)$, as:

$$\begin{aligned} \tilde{\boldsymbol{\theta}}_k(x) &= \operatorname*{argmax}_{\theta\in\Theta} L_{NSM}(W^{(k)}, \boldsymbol{\theta}) \\ &= \operatorname*{argmin}_{\theta\in\Theta} \sum_{i=1}^{n} \frac{(Y_i - \boldsymbol{\psi}_i^T \boldsymbol{\theta} - \mu z_i)^2}{\delta^2 z_i} w_i^{(k)}. \end{aligned} \tag{4.1}$$

Then, we start with the smallest bandwidth h_1. For any $k > 1$, compute the local qMLE $\tilde{\boldsymbol{\theta}}_k(x)$ and check whether it is consistent with all the previous estimators $\tilde{\boldsymbol{\theta}}_l(x)$ for $l < k$. We use a localized likelihood ratio test, i.e. the difference $L_{NSM}\left(W^{(l)}, \tilde{\boldsymbol{\theta}}_l(x)\right) - L_{NSM}\left(W^{(l)}, \tilde{\boldsymbol{\theta}}_k(x)\right)$ to reject $\tilde{\boldsymbol{\theta}}_k(x)$, where $\tilde{\boldsymbol{\theta}}_l(x)$ maximize the log-likelihood $L_{NSM}\left(W^{(l)}, \tilde{\boldsymbol{\theta}}_l(x)\right) = \sup_\theta L_{NSM}\left(W^{(l)}, \boldsymbol{\theta}\right)$ defined in Eq.(3.5) with bandwidth h_l and $L_{NSM}\left(W^{(l)}, \tilde{\boldsymbol{\theta}}_k(x)\right)$ is the other local likelihood under $\tilde{\boldsymbol{\theta}}_k(x)$ with bandwidth $h_k (l < k)$. The difference checks whether $\tilde{\boldsymbol{\theta}}_k(x)$ belongs to the confidence set $\varepsilon_l(\zeta)$ of $\tilde{\boldsymbol{\theta}}_l(x)$:

$$\varepsilon_l(\zeta) := \left\{\boldsymbol{\theta} \in \Theta : L_{NSM}\left(W^{(l)}, \tilde{\boldsymbol{\theta}}_l(x)\right) - L_{NSM}\left(W^{(l)}, \tilde{\boldsymbol{\theta}}_k(x)\right) \leq \zeta_l\right\},$$

where ζ_l refers the choice of critical values given in Theorem 1 below.

If the consistency check is negative, the procedure terminates and selects the latest accepted estimator.

The adaptation algorithm can be summarized as follows:

Algorithm 1

Step 1: Start with $\hat{\boldsymbol{\theta}}_1(x) = \tilde{\boldsymbol{\theta}}_1(x)$.

Step 2: For $k \geq 2$, $\tilde{\boldsymbol{\theta}}_k(x)$ is accepted and $\hat{\boldsymbol{\theta}}_k(x) = \tilde{\boldsymbol{\theta}}_k(x)$, if $\tilde{\boldsymbol{\theta}}_{k-1}(x)$ was accepted and

$$L_{NSM}\left(W^{(l)}, \tilde{\boldsymbol{\theta}}_l(x)\right) - L_{NSM}\left(W^{(l)}, \tilde{\boldsymbol{\theta}}_k(x)\right) \leq \zeta_l, \qquad l = 1, ..., k-1.$$

where the choice of critical values $\zeta_l, l = 1, ..., k-1$ are based on the propagation conditions (detailed in Theorem 1 below).

Step 3: Otherwise, $\hat{\boldsymbol{\theta}}_k(x) = \hat{\boldsymbol{\theta}}_{k-1}(x)$.

The adaptive estimator $\hat{\boldsymbol{\theta}}(x)$ is the latest accepted estimator after all K steps:

$$\hat{\boldsymbol{\theta}}(x) = \hat{\boldsymbol{\theta}}_K(x).$$

Moreover, all the estimators $\tilde{\boldsymbol{\theta}}_k(x)$ should be consistent to each other and the procedure should not terminate at any intermediate step $k < K$. This effect is called as 'propagation'. Hence, under the assumptions **(A1)-(A3)** in Appendix, and then according to Serdyukova (2012), the propagation conditions (PC) for this approach also satisfies:

Theorem 1. *(Theoretical choice of the critical values.) Assume **(A1)-(A3)**, given $\alpha \in (0,1]$ and $r > 0$, the critical values $\zeta_1, \cdots, \zeta_K$ satisfy*

$$\mathbb{E}\left|\left(\tilde{\boldsymbol{\theta}}_k(x) - \hat{\boldsymbol{\theta}}_k(x)\right)^T \left(\boldsymbol{\psi} w_k(x) \boldsymbol{\psi}^T\right) \left(\tilde{\boldsymbol{\theta}}_k(x) - \hat{\boldsymbol{\theta}}_k(x)\right)\right|^r \leq \alpha C(p,r), \tag{4.2}$$

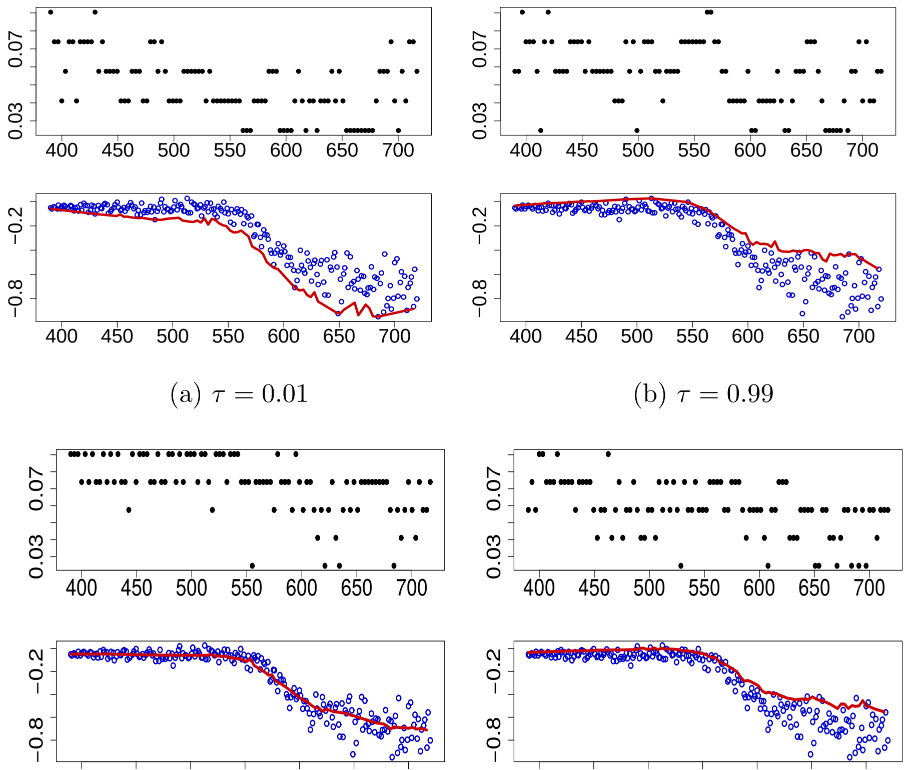
for all $k = 2, \cdots, K$, where $C(p,r) = 2^r \Gamma(r + p/2)/\Gamma(p/2)$, with the choice of the critical values of the form

$$\zeta_l = \frac{4}{\mu}\left\{r(K-l)\text{log}b + \text{log}\frac{K}{\alpha} - \frac{p}{4}\text{log}(1-4\mu) - \text{log}(1-b^{-r}) + \bar{C}(p,r)\right\}, l = 1, ..., k-1$$

where $\mu \in (0, 1/4)$ is an arbitrary constant, $b > 1$ and $\bar{C}(p,r) = \log\left\{\frac{2^{2r}[\Gamma(2r+p/2)\Gamma(p/2)]^{1/2}}{\Gamma(r+p/2)}\right\}$.

The critical values are selected to ensure the desired propagation condition which effectively means a 'no alarm' property, that is the selected adaptive estimator coincides in the most cases that the estimator $\tilde{\boldsymbol{\theta}}_k(x)$ corresponding to the largest bandwidth.

An advantage of the proposed alternative normal scale-mixture likelihood function over SWH's method is that the derived bandwidth has better adaptation when τ tends to 0 or 1. Figure 2 displays the bandwidth sequence (upper panel) and smoothed quantile curves for quantiles 1% (2a) and 99% (2b) based on the Lidar dataset, which provides much better fitting than those curves presented in Figure 1. The dependency structure changing on smoothness is more adaptive than the bandwidth sequence in Figure 1. This alternative normal scale-mixture likelihood method also works well for other moderate or central quantile curves. Figure 2 shows that the method gives quite similar estimates to SWH's method for $\tau = 0.5$ (2c) and 0.9 (2d) quantile curves.



(a) $\tau = 0.01$

(b) $\tau = 0.99$

(c) $\tau = 0.50$

(d) $\tau = 0.90$

Figure 2: The bandwidth sequences (upper panels) and smoothed quantile curves (lower panels) for the Lidar dataset using the alternative normal scale-mixture likelihood.

4.2 Non-crossing Quantile Curve Estimation

The proposed bandwidth selection rule in SWH's method seems to have no quantile crossing phenomenon when several smoothed quantile curves are provided together. This indicates the advantage of the local bandwidth selection rule. Whereas most of published articles on this topic, which include constrained smoothing spline (He, 1997; Bondell et al., 2010), double-kernel smoothing (Yu and Jones, 1998; Jones and Yu, 2007a) and monotone constraint on conditional distribution function (Hall et al., 1999; Dette and Volgushev, 2008), among others, focus on the development of new methods rather than adaptive bandwidth selection for avoiding quantile crossing. SWH showed, even working with 'local constant' kernel smoothing quantile regression via

$$\hat{q}_\tau(x) = \underset{a}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\tau \left(Y_i - a\right) K_h \left(x - X_i\right),$$

adaptive bandwidth selection rule may not have quantile crossing either. This may be true practically, but without a theoretical justification. Under our proposed approach, the justification of non-crossing quantiles could be outlined as below.

Recall the nonparametric quantile regression model $Y = f(X) + \epsilon$, where $Q_\tau(\epsilon) = 0$. Given data $\{X_i, Y_i\}_{i=1}^n$, and under the local polynomial approach, $\tilde{\theta}_0(x)$ estimates $f(x)$, with

$$\begin{aligned} \tilde{\boldsymbol{\theta}}_{NSM} &\equiv \left(\tilde{\theta}_0, \tilde{\theta}_1, \cdots, \tilde{\theta}_p\right) \\ &= \underset{\theta \in \Theta}{\operatorname{argmax}} \, L_{NSM}(W, \boldsymbol{\theta}), \end{aligned}$$

where the likelihood function $L_{NSM}(W, \boldsymbol{\theta})$ is expressed in Eq.(3.5) and $\tilde{\theta}_m(x)$ estimate the m^{th} derivative of $f(x)$.

That is, the derivative of $L_{NSM}(W, \boldsymbol{\theta})$ over $\tilde{\theta}_0(x)$ satisfies $\sum_{i=1}^n \frac{w_i}{z_i}(Y_i - \boldsymbol{\psi}_i^T \tilde{\boldsymbol{\theta}}_{NSM} - \mu z_i) = 0$. Therefore, $\tilde{\theta}_0(x)$ can be expressed as,

$$\tilde{\theta}_0(x) = \frac{\sum_{i=1}^n \frac{w_i}{z_i} \left(Y_i - \mu z_i - \sum_{j=1}^p \tilde{\theta}_j \frac{(X_i - x)^j}{j!}\right)}{\sum_{i=1}^n \frac{w_i}{z_i}}.$$

For each x, we aim to check the derivative of $\tilde{\theta}_0(x)$ over $\tau \in (0,1)$. If $\frac{d\tilde{\theta}_0(x)}{d\tau} > 0$, then $\tilde{\theta}_0(x)$ is an increasing function of τ.

Note that $\mu = \frac{1-2\tau}{\tau(1-\tau)}$, therefore, we have

$$\begin{aligned}
\frac{d\tilde{\theta}_0(x)}{d\tau} &= \frac{1}{\sum_{i=1}^{n}\frac{w_i}{z_i}}\sum_{i=1}^{n}\frac{-z_i w_i \frac{d\mu}{d\tau}}{z_i} \\
&= \frac{1}{\sum_{i=1}^{n}\frac{w_i}{z_i}}\sum_{i=1}^{n}\frac{-z_i w_i \frac{-2(\tau-1/2)^2-1/2}{\tau^2(1-\tau)^2}}{z_i} \\
&= \frac{1}{\sum_{i=1}^{n}\frac{w_i}{z_i}}\sum_{i=1}^{n} w_i \frac{2(\tau-1/2)^2+1/2}{\tau^2(1-\tau)^2} \\
&> 0. \qquad (4.3)
\end{aligned}$$

That is, $\hat{f}(x) \equiv \tilde{\theta}_0(x)$ is a strictly monotonic function of τ over x.

5 Numerical examples

In this section, we implement the proposed method via extensive Monte Carlo simulation studies and one real data analysis. All numerical experiments are carried out on one Inter Core i5-3470 CPU (3.20GMHz) processor and 8 GB RAM.

5.1 Simulation 1

In this simulation study, we aim to summarize our numerical results on choosing the critical values by the propagation condition as described in Section 4.1. We generate data of size 10^6 from an $ALD_\tau(0,1)$, which does coincide with the likelihood (ALD_τ) taken to simulate critical values. We mainly check the critical values at different quantile levels $\tau = 0.05, 0.25, 0.5, 0.75, 0.95$, and for different choices of α and r. We also study how bandwidth sequence affects the critical values.

Table 1 shows the critical values with several choices of α and r with $\tau = 0.2$ and $m =$ 5000 Monte Carlo samples, and a bandwidth sequence $(5, 7, 10, 13, 17, 21, 24, 28, 36, 45)/365$

scaled to the interval $[0, 1]$. Critical values decrease when α increases, and increase when r increases.

Table 1: Critical values with different α and r ($\tau = 0.2$).

α	r	Critical values					
0.25	0.5	16.971	11.539	8.133	3.584	0.044	0.000
0.25	0.75	20.218	13.743	9.336	3.131	0.000	0.000
0.25	1	24.676	16.270	9.308	4.214	1.561	0.000
0.5	0.5	12.823	9.619	7.205	3.703	0.949	0.000
0.75	0.5	11.249	7.222	4.244	0.181	0.000	0.000

Table 2 shows the critical values for different τs with $\alpha = 0.25, r = 0.5$ and $m = 5000$ Monte Carlo samples, and a bandwidth sequence $(5, 7, 10, 13, 17, 21, 24, 28, 36, 45)/365$ scaled to the interval $[0, 1]$. Critical values behave similarly for symmetric τ.

Table 2: Critical values with different τ ($\alpha = 0.25, r = 0.5$).

τ	Critical values					
0.05	10.357	7.605	4.888	1.248	0.000	0.000
0.25	15.782	11.332	8.440	4.354	0.908	0.000
0.50	21.714	15.427	10.351	3.594	0.000	0.000
0.75	15.283	10.932	8.396	3.949	0.840	0.000
0.95	10.789	7.686	4.943	1.208	0.000	0.000

Table 3 compares critical values for the following three bandwidth sequences, with $\alpha = 0.25, r = 0.5, \tau = 0.8$ and $m = 5000$ Monte Carlo samples.

$$\eta_1 = (5, 7, 10, 13, 17, 21, 24, 28, 36, 45)/365$$

$$\eta_2 = (10, 13, 17, 21, 24, 28, 36, 45, 49, 60)/365$$

$$\eta_3 = (2, 3, 5, 7, 10, 13, 17, 21, 24, 28)/365$$

Clearly, although the critical values differ for different bandwidth sequences, they indicate the same patterns (finite and decreasing). Moreover, the adaptation algorithm can be completed in maximum $K = 6$ steps, as all critical values decrease to zero in 6-step.

Table 3: Critical values with different bandwidth sequences ($\alpha = 0.25, r = 0.5, \tau = 0.8$).

η	Critical values					
η_1	11.002	6.508	3.089	0.000	0.000	0.000
η_2	23.187	13.810	7.775	3.690	0.000	0.000
η_3	6.871	4.737	2.046	0.389	0.000	0.000

5.2 Simulation 2

In this simulation study, we compare the performance of our proposed approach to SWH's method as well as two other bandwidth selection techniques. One proposal comes from Ng and Maechler (2007), in which they considered constrained quantile estimations using linear or quadratic splines (implemented with R function *cobs* in Package *cobs*), and the other is from Yu and Jones (1998), in which they considered a rule of thumb bandwidth (implemented with R function *lprq* in Package *quantreg*).

We generate one training data of size 2000 and 500 test data sets of size 500 from the model

$$Y = m(X) + \sigma(X)\varepsilon, \tag{5.1}$$

where the univariate input X follows a uniform distribution on $[4, 4]$ and $m(X)$ is a non-linear function of X

$$m(X) = (1 - X + 2X^2)e^{-0.5x^2},$$

and the scale factor $\sigma(X)$ is linearly increasing in X with the form

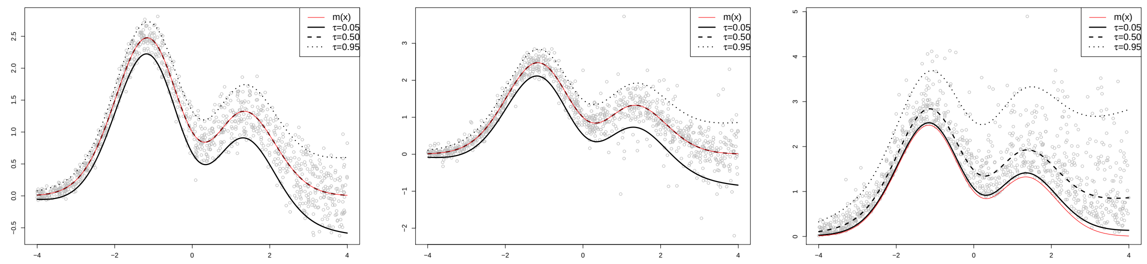
$$\sigma(X) = \frac{1}{5}(1 + 0.2x).$$

Therefore, Eq.(5.1) is a heteroskedastic model.

In this simulation, we consider three different types of random errors for ε: $N(0, 1)$, $t(3)$ and $\chi^2(3)$, respectively. Therefore, the true τ-th conditional quantile function of Y given $X = x$ can be expressed as

$$Q_Y(\tau|x) = m(x) + \sigma(x)F_\tau^{-1}(\varepsilon),$$

where $F_\tau^{-1}(\varepsilon)$ is the τ-th quantile of ε. Fig. 3 presents the training data generated under this scenario with their true τ-th conditional quantile functions $Q_Y(\tau|x), \tau \in c(0.05, 0.50, 0.95)$. Note that, the nonlinear function $m(X)$ in the right figure is not identical to the true conditional median function $Q_Y(0.50|x)$ as the random error $\chi^2(3)$ is an asymmetric distribution.



(a) $\varepsilon \sim N(0, 1)$ (b) $\varepsilon \sim t(3)$ (c) $\varepsilon \sim \chi^2(3)$

Figure 3: Simulated training data and true conditional quantile functions with $\tau \in c(0.05, 0.50, 0.95)$.

We aim to compare the prediction power of the above-mentioned four methods for the prediction of the conditional quantile function by 500 test data sets, in terms of three measurements, namely, the root mean square error (RMSE), the mean absolute errors (MAE), and the Theil-U statistic, which is a relative accuracy measure that compares the forecast results with the naïve forecast (Theil, 1966):

$$
\begin{aligned}
RMSE(\tau) &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Q_{Y_i}(\tau|x) - \hat{Q}_{Y_i}(\tau|x)\right)^2},\\
MAE(\tau) &= \frac{1}{n}\sum_{i=1}^{n}\left|Q_{Y_i}(\tau|x) - \hat{Q}_{Y_i}(\tau|x)\right|,\\
TheiU(\tau) &= \sqrt{\frac{\sum_{i=2}^{n}\left(\frac{\hat{Q}_{Y_i}(\tau|x)-Q_{Y_i}(\tau|x)}{Q_{Y_{i-1}}(\tau|x)}\right)^2}{\sum_{i=2}^{n}\left(\frac{Q_{Y_i}(\tau|x)-Q_{Y_{i-1}}(\tau|x)}{Q_{Y_{i-1}}(\tau|x)}\right)^2}},
\end{aligned}
$$

where $\hat{Q}_{Y_i}(\tau|x)$ is the prediction of the true conditional quantile $Q_{Y_i}(\tau|x)$. The smaller the measurement value is, the better the method is. The three measurements are implemented with R function *av.res* in package *AnalyzeTS*.

The superiority of the proposed normal-scale mixture approach is demonstrated in Table 4 which summarizes the results for three values of τs: 0.05, 0.50, and 0.95, based on the 500 replications. Note that, Simulation 2 is implemented with bandwidth sequence η= (5,7,10,13,17,21,24,28,36,45)/365, simulated from $ALD(0,1,\tau)$ (coincide with the likelihood) with α=0.25, r=0.5. The bold face values show that both SWH's method and the proposed normal scale-mixture approach are superior to LPQR and COBS, while the proposed approach performs slightly better than SWH. It is encouraging to see that the proposed approach approximates well under Gaussian error and also provides excellent results under the circumstance of heavy tail and asymmetric distributions, such as $t(3)$ and $\chi^2(3)$.

Table 4: Average value of the evaluation indices for 500 test data of size 500.

	$\varepsilon \sim N(0,1)$				$\varepsilon \sim t(3)$				$\varepsilon \sim \chi^2(3)$			
Indices	LPQR	COBS	SWH	NSM	LPQR	COBS	SWH	NSM	LPQR	COBS	SWH	NSM
$\tau = 0.05$												
RMSE	0.364	0.254	0.168	**0.157**	0.399	0.274	0.226	**0.213**	0.432	0.239	0.162	**0.154**
MAE	0.234	0.176	0.128	**0.121**	0.273	0.205	0.173	**0.163**	0.269	0.162	0.121	**0.116**
Thei U	17.773	12.414	8.196	**7.667**	19.293	13.264	10.896	**10.286**	20.974	11.640	7.863	**7.480**
$\tau = 0.5$												
RMSE	0.178	0.184	0.163	**0.140**	0.184	0.172	0.141	**0.139**	0.210	0.198	0.176	**0.170**
MAE	0.140	0.144	0.128	**0.114**	0.144	0.137	0.107	**0.103**	0.171	0.161	0.139	**0.132**
Thei U	8.524	8.865	7.839	**7.131**	8.942	8.403	6.875	**6.741**	10.246	9.695	8.587	**8.241**
$\tau = 0.95$												
RMSE	0.258	0.210	0.159	**0.157**	0.283	0.245	0.205	**0.195**	0.367	**0.324**	0.331	0.326
MAE	0.193	0.153	0.125	**0.123**	0.226	0.190	0.162	**0.153**	0.272	0.261	**0.250**	0.261
Thei U	12.507	10.176	7.735	**7.600**	8.983	9.553	**6.862**	7.570	16.743	**14.798**	15.159	14.852

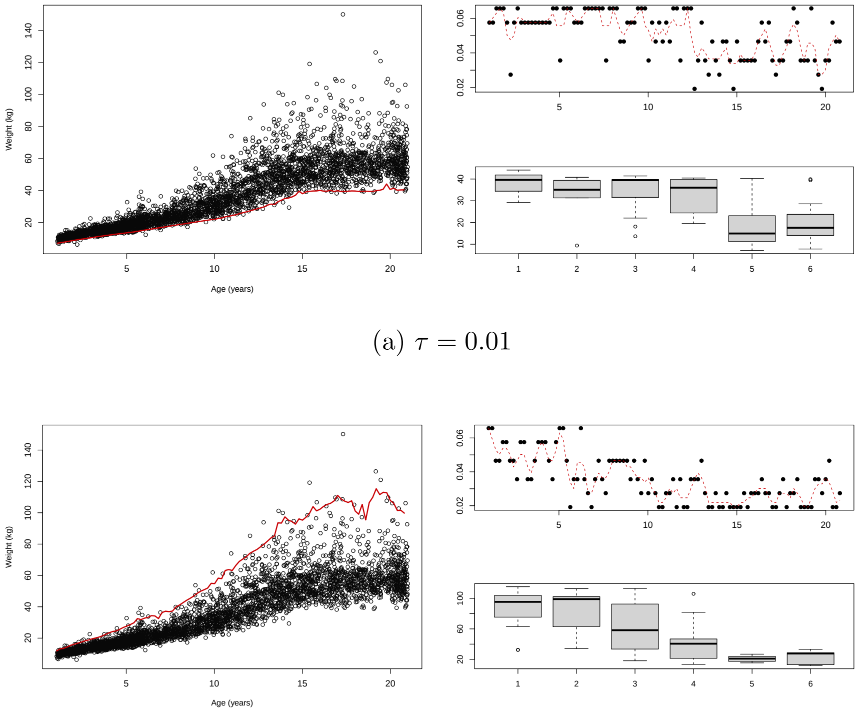
Note: The bandwidth h_τ at τ that controls the complexity of the LPQR model is selected by the rule of thumb in Fan and Gijbels (1996b).

5.3 Real-world data application

In this section we demonstrate the efficacy of our the proposed alternative approach with one benchmark example that comes from the second and third health examination surveys of the USA (National Center for US Health Examination Surveys, 1970; 1973). Taken together these provide data on the anthropometry of children between the ages of 6 years and under 18 years, with from 400 to 600 children of each sex seen in each year of age (Cole, 1988). Here, along with Yu and Jones (1998), the weights and ages of 4011 US girls were analysed.

The scatter plot in Figure 4a displays weight against age for a sample of 4011 US girls, where age is a univariate regressor $X \in R^1$ for simplicity. It is evident that the distribution is left-skewed and presents long tails, suggesting that focusing on the centre is not sufficient for a comprehensive description of a weight distribution. Such observation motivates the use of quantile regression, where a complete picture of weight distribution is captured by conditional quantiles.

We then continue by inspecting the relation between weight and age in the sample. In Figure 4, we display the bandwidth sequence (upper right panels), boxplot of adapted bandwidth (lower right panels) showing the relationship between the adapted estimator and the bandwidth index, and smoothed quantile curves for quantile 99% (4b) and 1% (4a) respectively by using the alternative normal scale-mixture likelihood function. Both adaptations show that the proposed bandwidth selection is well-adapted over the data distribution, which provides smooth fitting and better adaptation when τ tends to extreme quantiles. Furthermore, Figure 5 shows that the non-quantile crossing property holds for the rule in Section 4.2, which is based on the alternative normal scale-mixture likelihood function.



(a) $\tau = 0.01$

(b) $\tau = 0.99$

Figure 4: Smoothed quantile curves (in red) for US Health Examination Surveys with $\tau = 0.01$ and $\tau = 0.99$ via alternative normal scale-mixture likelihood (left panel). The bandwidth sequence (upper right); boxplot of adaptive bandwidth (lower right).

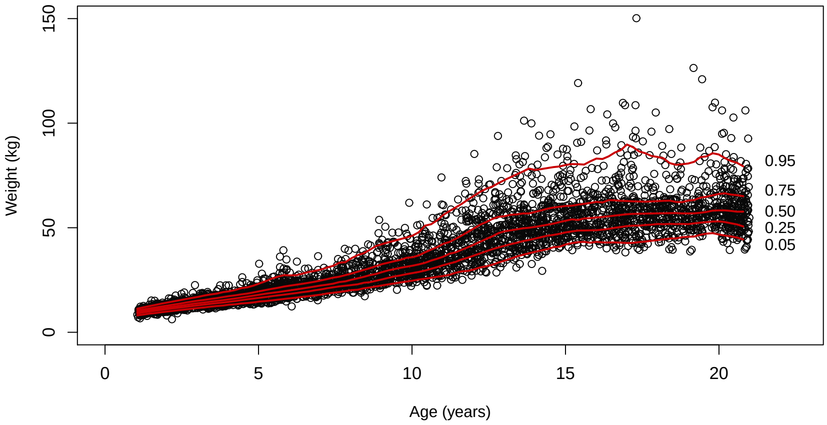


Figure 5: Smoothed quantile curves for US Health Examination Surveys with $\tau = c(0.05, 0.25, 0.5, 0.75, 0.95)$ via alternative normal scale-mixture likelihood function.

6 Discussions and Concluding Remarks

The kernel-weighted likelihood function Eq.(2.5) in SWH's paper is a local ALD-based likelihood function. The ALD-based inference has nowadays become a powerful tool for formulating different quantile regression techniques, particularly for the development of different Bayesian inference techniques for quantile regression. The ALD-based inference for non-Bayesian methods includes Taylor and Yu (2016) in financial risk analysis, Geraci and Bottai (2007) in longitudinal data analysis and among others. The local ALD-based likelihood approach in the paper uses an alternative ALD-type of likelihood. The resulting automatic bandwidth selection rule not only enjoys the propagation condition of SWH (which postulates that the risk is smaller than the upper bound for the risk of the estimator $\tilde{\boldsymbol{\theta}}_k(x)$) but also guarantees non-quantile curve crossing. Theoretical results also claim that the proposed adaptive procedure performs well, which would minimize the local estimation risk for the problem at hand. We illustrate the performance of the procedure by comparing the Lidar dataset with SWH's approach and analyzing an extended real data application. In particular, we show that the performance of the adaptive procedure is promising in practice, especially for smoothing extreme quantile curves.

Moreover, the proposed approach can also be extended to the d-dimensional case $X \in \mathbb{R}^d$ with $d > 1$, under the non-parametric additive modelling framework (Yu and Lu, 2004). That is, let Y be a real-valued dependent variable and $X = \left(X^{(1)}, \cdots, X^{(d)}\right) \in \mathbb{R}^d$ is a vector of explanatory variables. Let $f(x)$ be a d-dimensional τth quantile regression function of Y given $X = x$. Suppose that the τth quantile function $f(\boldsymbol{x})$ is modelled as an additive function of $\left(x^{(1)}, \cdots, x^{(d)}\right)$,

$$f(x) = \sum_{l=1}^{d} f^{(l)}\left(x^{(l)}\right), \tag{6.1}$$

where each $f^{(l)}(x^{(l)})$ can be fitted by the proposed approach in Section 3 and the whole $f(x)$ can be further derived via backfitting algorithm used in Yu and Lu (2004). For example, without of generality, consider a local linear regression with $p = 2$, for $l = 1, \cdots, d$,

$$(\hat{a}^{(l)}, \hat{b}^{(l)}) = \operatorname*{argmin}_{a,b} \sum_{i=1}^{n} \rho_\tau \left(Y_i - a - b\left(X_i^{(l)} - x^{(l)}\right)\right)\left(\frac{X_i^{(l)} - x^{(l)}}{h^{(l)}}\right),$$

where $K(\cdot)$ is a kernel function and $h^{(l)}(l = 1, \cdots, d)$ is the bandwidth for estimating $f^{(l)}(x^{(l)})$ in the setting above.

Acknowledgements

We thank the editor, the associate editor and two anonymous reviewers for their constructive comments, which helped us to improve the manuscript. The research was partially supported by Major Program of the National Natural Science Foundation of China (Grant No. 71490725), the BUL Research Leave funding, the National Science Foundation of China (No 11261048).

Appendix:

Recall: $\boldsymbol{w}_k = diag\left(\frac{w_1^{(k)}}{\delta^2 z_1}, ..., \frac{w_n^{(k)}}{\delta^2 z_n}\right)$.

Assumption 1. *Consider a finite sequence of scales* $w_k = diag\left(w_1^{(k)}, \cdots, w_n^{(k)}\right)$*, the* $p \times n$ *matrix* $\boldsymbol{\psi}^T w_1$*is of full row rank.*

Assumption 2. *For any fixed* x *and the method of localization with* $w_i^{(k)}(x) \geq 0$*, the following relation holds:*

$$w_1(x) \leq w_2(x) \leq \cdots \leq w_n(x).$$

Assumption 3. *Assume that the true regression model*

$$Y_i = f_0(X_i) + \mu_0 z_{0,i} + \delta_0^2 \sqrt{z_{0,i}} e_i,$$

considering the regression model (3.2), where $\boldsymbol{z}_0 = diag\left(\delta_0^2 z_{0,1}, \cdots, \delta_0^2 z_{0,n}\right)$ stands for the unknown true covariance matrix, with $z_{0,i}$ is the true value of Eq.(3.2), there exists $\eta \in [0, 1)$ such that

$$1 - \eta \leq \frac{\delta_0^2 z_{0,i}}{\delta^2 z_i} \leq 1 + \eta \quad for\ all\ i = 1, \cdots, n.$$

Assuming **(A3)**, the true covariance matrix $\boldsymbol{z}_0 \preceq \boldsymbol{z}(1 + \eta)$, and the conditional variance of the estimate $\tilde{\boldsymbol{\theta}}_k(x)$ is bounded with $\left(\boldsymbol{\psi}\boldsymbol{w}_k\boldsymbol{\psi}^T\right)^{-1}$: as follows :

$$\begin{aligned}
\operatorname{Var}\left(\tilde{\boldsymbol{\theta}}_k(x)\right) &= \left(\boldsymbol{\psi}\boldsymbol{w}_k\boldsymbol{\psi}^T\right)^{-1}\boldsymbol{\psi}\boldsymbol{w}_k\boldsymbol{z}_0\boldsymbol{w}_k\boldsymbol{\psi}^T\left(\boldsymbol{\psi}\boldsymbol{w}_k\boldsymbol{\psi}^T\right)^{-1} \\
&\preceq (1+\eta)\left(\boldsymbol{\psi}\boldsymbol{w}_k\boldsymbol{\psi}^T\right)^{-1}\boldsymbol{\psi}\boldsymbol{w}_k\boldsymbol{z}\boldsymbol{w}_k\boldsymbol{\psi}^T\left(\boldsymbol{\psi}\boldsymbol{w}_k\boldsymbol{\psi}^T\right)^{-1} \\
&= (1+\eta)\left(\boldsymbol{\psi}\boldsymbol{w}_k\boldsymbol{\psi}^T\right)^{-1}\boldsymbol{\psi}\boldsymbol{z}^{-1/2}w_k^2\boldsymbol{z}^{-1/2}\boldsymbol{\psi}^T\left(\boldsymbol{\psi}\boldsymbol{w}_k\boldsymbol{\psi}^T\right)^{-1} \\
&\preceq (1+\eta)\left(\boldsymbol{\psi}\boldsymbol{w}_k\boldsymbol{\psi}^T\right)^{-1}\boldsymbol{\psi}\boldsymbol{z}^{-1/2}w_k\boldsymbol{z}^{-1/2}\boldsymbol{\psi}^T\left(\boldsymbol{\psi}\boldsymbol{w}_k\boldsymbol{\psi}^T\right)^{-1} \\
&= (1+\eta)\left(\boldsymbol{\psi}\boldsymbol{w}_k\boldsymbol{\psi}^T\right)^{-1}\left(\boldsymbol{\psi}\boldsymbol{w}_k\boldsymbol{\psi}^T\right)\left(\boldsymbol{\psi}\boldsymbol{w}_k\boldsymbol{\psi}^T\right)^{-1} \\
&= (1+\eta)\left(\boldsymbol{\psi}\boldsymbol{w}_k\boldsymbol{\psi}^T\right)^{-1} \\
&= (1+\eta)\left(\sum_{i=1}^{n}\boldsymbol{\psi}_i\boldsymbol{\psi}_i^T\frac{w_i^{(k)}}{\delta^2 z_i}\right)^{-1}. \qquad (6.2)
\end{aligned}$$

According to the basic property of quadratic equation, consider a simple example $\left(\frac{1}{z_1}+\frac{1}{z_2}\right)^{-1}$ and there always holds $\left(\frac{1}{z_1}+\frac{1}{z_2}\right)^{-1}=\frac{z_1z_2}{z_1+z_2}\le z_1+z_2$, with $z_1,z_2>0$. The same procedure may be easily adapted to Eq.(6.2) as follows:

$$\begin{aligned}
\operatorname{Var}\left(\tilde{\boldsymbol{\theta}}_k(x)\right) &\preceq (1+\eta)\sum_{i=1}^{n}\boldsymbol{\psi}_i\boldsymbol{\psi}_i^T w_i^{(k)}\delta^2 z_i \\
&= (1+\eta)\delta^2\sum_{i=1}^{n}\boldsymbol{\psi}_i\boldsymbol{\psi}_i^T w_i^{(k)} z_i. \qquad (6.3)
\end{aligned}$$

Therefore, the unconditional variance of the estimate $\tilde{\boldsymbol{\theta}}_k(x)$ as follows is bounded with $\boldsymbol{\psi}w_k\boldsymbol{\psi}^T$

$$\begin{aligned}
\mathbf{V}_k(x) &\equiv E\left[\operatorname{Var}\tilde{\boldsymbol{\theta}}_k(x)\right] \\
&= E\left[(1+\eta)\delta^2\sum_{i=1}^{n}\boldsymbol{\psi}_i\boldsymbol{\psi}_i^T w_i^{(k)} z_i\right] \\
&= (1+\eta)\delta^2\sum_{i=1}^{n}\boldsymbol{\psi}_i\boldsymbol{\psi}_i^T w_i^{(k)} E\left[z_i\right] \\
&= (1+\eta)\delta^2\sum_{i=1}^{n}\boldsymbol{\psi}_i\boldsymbol{\psi}_i^T w_i^{(k)} \\
&= (1+\eta)\delta^2\boldsymbol{\psi}w_k\boldsymbol{\psi}^T. \qquad (6.4)
\end{aligned}$$

Proof. of **Theorem 1**.

$$\begin{aligned}
&\mathbb{E}\left|\left(\tilde{\boldsymbol{\theta}}_k(x)-\hat{\boldsymbol{\theta}}_k(x)\right)^T\left(\boldsymbol{\psi} w_k(x)\boldsymbol{\psi}^T\right)\left(\tilde{\boldsymbol{\theta}}_k(x)-\hat{\boldsymbol{\theta}}_k(x)\right)\right|^r \\
=&\sum_{m=1}^{k-1}\mathbb{E}\left|\left(\tilde{\boldsymbol{\theta}}_k(x)-\tilde{\boldsymbol{\theta}}_m(x)\right)^T\left(\boldsymbol{\psi} w_k\boldsymbol{\psi}^T\right)\left(\tilde{\boldsymbol{\theta}}_k(x)-\tilde{\boldsymbol{\theta}}_m(x)\right)\right|^r I\left\{\hat{\boldsymbol{\theta}}_k(x)=\tilde{\boldsymbol{\theta}}_m(x)\right\}. \qquad (6.5)
\end{aligned}$$

The event $\left\{\hat{\boldsymbol{\theta}}_k(x)=\tilde{\boldsymbol{\theta}}_m(x)\right\}$ happens if for some $l=1,\cdots,m$, $T_{l,m+1}>\zeta_l$, Hence,

$$\left\{\hat{\boldsymbol{\theta}}_k(x)=\tilde{\boldsymbol{\theta}}_m(x)\right\}\subseteq\bigcup_{l=1}^{m}\{T_{l,m+1}>\zeta_l\}.$$

Further, combined with the Cauchy-Schwarz inequality, for any positive a:

$$\begin{aligned}
&\mathbb{E}\left|\left(\tilde{\boldsymbol{\theta}}_k(x)-\tilde{\boldsymbol{\theta}}_m(x)\right)^T\left(\boldsymbol{\psi} w_k\boldsymbol{\psi}^T\right)\left(\tilde{\boldsymbol{\theta}}_k(x)-\tilde{\boldsymbol{\theta}}_m(x)\right)\right|^r I\left\{\hat{\boldsymbol{\theta}}_k(x)=\tilde{\boldsymbol{\theta}}_m(x)\right\} \\
=\ &\mathbb{E}\left|2L_{NSM}\left(W^{(k)},\tilde{\boldsymbol{\theta}}_k(x),\tilde{\boldsymbol{\theta}}_m(x)\right)\right|^r I\left\{\hat{\boldsymbol{\theta}}_k(x)=\tilde{\boldsymbol{\theta}}_m(x)\right\} \\
\leq\ &\sum_{l=1}^{m}e^{-\frac{a}{4}\zeta_l}\left\{\mathbb{E}\left[\left|2L_{NSM}\left(W^{(k)},\tilde{\boldsymbol{\theta}}_k(x),\tilde{\boldsymbol{\theta}}_m(x)\right)\right|^{2r}\right]\right\}^{\frac{1}{2}}\left\{\mathbb{E}\left[\exp\left\{aL_{NSM}\left(W^{(k)},\tilde{\boldsymbol{\theta}}_l(x),\tilde{\boldsymbol{\theta}}_{m+1}(x)\right)\right\}\right]\right\}^{\frac{1}{2}} (6.6).
\end{aligned}$$

Among which,

$$\begin{aligned}
&E\left[\left|2L_{NSM}\left(W^{(k)},\tilde{\boldsymbol{\theta}}_k(x),\tilde{\boldsymbol{\theta}}_m(x)\right)\right|^{2r}\right] && (6.7)\\
=\quad &2r\int_0^{\infty}P\left\{2L_{NSM}\left(W^{(k)},\tilde{\boldsymbol{\theta}}_k(x),\tilde{\boldsymbol{\theta}}_m(x)\right)\geq\zeta\right\}\zeta^{2r-1}d\zeta \\
\leq\quad &2r\int_0^{\infty}P\left\{\gamma\geq\zeta\left[2\left(1+\eta\right)\left(1+b^{(k-m)}\right)\right]^{-1}\right\}\zeta^{2r-1}d\zeta \\
=\quad &2^{2r}\left(1+\eta\right)^{2r}\left(1+b^{(k-m)}\right)^{2r}E\left|\chi_p^2\right|^r \\
=\eta=0\quad &2^{2r}C(p,2r)\left(1+b^{(k-m)}\right)^{2r}, && (6.8)
\end{aligned}$$

and

$$\begin{aligned}
&E\left[\exp\left\{aL_{NSM}\left(W^{(k)},\tilde{\boldsymbol{\theta}}_l(x),\tilde{\boldsymbol{\theta}}_{m+1}(x)\right)\right\}\right] \\
=\quad &\prod_{j=1}^{p}\left[1-a\lambda_j\left(V_{l,m+1}^{-1/2}\left(\boldsymbol{\psi} w_m\boldsymbol{\psi}^T\right)V_{l,m+1}^{-1/2}\right)\right]^{-1/2} \\
\leq\quad &\left[1-a\lambda_{max}\left(V_{l,m+1}^{-1/2}\left(\boldsymbol{\psi} w_m\boldsymbol{\psi}^T\right)V_{l,m+1}^{-1/2}\right)\right]^{-p/2} \\
\leq\quad &\left[1-2a\left(1+\eta\right)\left(1+b^{-(m+1-l)}\right)\right]^{-p/2} \\
=\eta=0\quad &\left[1-2a\left(1+b^{-(m+1-l)}\right)\right]^{-p/2}. && (6.9)
\end{aligned}$$

Therefore, we obtain

$$\begin{aligned} & E\left|\left(\tilde{\boldsymbol{\theta}}_k(x)-\hat{\boldsymbol{\theta}}_k(x)\right)^T\left(\boldsymbol{\psi} w_k(x)\boldsymbol{\psi}^T\right)\left(\tilde{\boldsymbol{\theta}}_k(x)-\hat{\boldsymbol{\theta}}_k(x)\right)\right|^r \\ \leq\ & 2^r\sqrt{C(p,2r)}(1-4a)^{-p/4}\sum_{m=1}^{k-1}\sum_{l=1}^{m} e^{-\frac{\mu}{4}\zeta_l}\left(1+b^{k-m}\right)^r \\ \leq\ & 2^{2r}\sqrt{C(p,2r)}(1-4a)^{-p/4}(1-b^{-r})\sum_{l=1}^{k-1} e^{-\frac{\mu}{4}\zeta_l} b^{r(k-l)}. \qquad (6.10) \end{aligned}$$

For any $l<k<K$, with an arbitrary constant $\mu\in(0,1/4)$ the choice of the threshold of the form

$$\zeta_l=\frac{4}{\mu}\left\{r(K-l)\log b+\log\frac{K}{\alpha}-\frac{p}{4}\log(1-4\mu)-\log(1-b^{-r})+\bar{C}(p,r)\right\},$$

where $\bar{C}(p,r)=\log\left\{\frac{2^{2r}[\Gamma(2r+p/2)\Gamma(p/2)]^{1/2}}{\Gamma(r+p/2)}\right\}$ provides the required PC bounds.

$$E\left|\left(\tilde{\boldsymbol{\theta}}_k(x)-\hat{\boldsymbol{\theta}}_k(x)\right)^T\left(\boldsymbol{\psi} w_k(x)\boldsymbol{\psi}^T\right)\left(\tilde{\boldsymbol{\theta}}_k(x)-\hat{\boldsymbol{\theta}}_k(x)\right)\right|^r\leq\alpha C(p,r),\ \textit{for all}\ k=2,\cdots,K.$$

□

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *In: B. N. PETROV and F. CSAKI, eds. Second International Symposium on Information Theory. Budapest: Akademiai Kiado*, pages 267–281.

Bondell, H., Reich, B., and Wang, H. (2010). Noncrossing quantile regression curve estimation. *Biometrika*, **97**, 825–838.

Cai, Z. and Xu, X. (2008). Nonparametric quantile estimations for dynamic smooth coefficient models. *Journal of the American Statistical Association*, **103**, 1595–1608.

Chaudhuri, P. (1991). Nonparametric estimates of regression quantiles and their local bahadur representation. *The Annals of statistics*, **19**, 760–777.

Cole, T. (1988). Fitting smoothed centile curves to reference data. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, pages 385–418.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerical Mathematics*, **31**, 377–403.

Dabo-Niang, S. and Laksaci, A. (2012). Nonparametric quantile regression estimation for functional dependent data. *Numerical Mathematics*, **41**, 1254–1268.

Dette, H. and Volgushev, S. (2008). Non-crossing non-parametric estimates of quantile curves. *Journal of the Royal Statistical Society B*, **70**, 609–627.

Fan, J. and Gijbels, I. (1996a). *Local polynomial modelling and its applications.* Chapman and Hall.

Fan, J. and Gijbels, I. (1996b). *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*, volume 66. CRC Press.

Geraci, M. and Bottai, M. (2007). Quantile regression for longitudinal data using the asymmetric laplace distribution. *Biostatistics*, **8**(1), 140–154.

Hall, P., Wolff, R. C., and Yao, Q. (1999). Methods for estimating a conditional distribution function. *Journal of the American Statistical Association*, **94**, 154–163.

Hardle, W. and Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *The Annals of Statistics*, **21**, 1926–1947.

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

Hurvich, C., Simonoff, J. S., and Tsai, C. (1998). Smoothing parameter selection in nonparametric regression using an improved akaike information criterion. *Journal of the Royal Statistical Society: B*, **60**, 271–293.

Jones, M. and Yu, K. (2007a). Improve double kernel local linear quantile regression. *Statistical Modelling*, **7**, 377–389.

Jones, M. and Yu, K. (2007b). Improved double kernel local linear quantile regression. *Statistical Modelling*, **7**(4), 377–389.

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

Kong, E. and Xia, Y. (2015). Uniform bahadur representation for nonparametric censored quantile regression: A redistribution-of-mass approach. *Econometric Theory.*

Kozumi, H. and Kobayashi, G. (2011). Gibbs sampling methods for bayesian quantile regression. *Journal of Statistical Computation and Simulation*, **81**, 1565–1578.

Liu, Y. and Wu, Y. (2011). Simultaneous multiple non-crossing quantile regression estimation using kernel constraints. *Journal of nonparametric statistics*, **23**(2), 415–437.

Muggeo, V. M., Sciandra, M., Tomasello, A., and Calvo, S. (2013). Estimating growth charts via nonparametric quantile regression: a practical framework with application in ecology. *Environmental and ecological statistics*, **20**(4), 519–531.

Ng, P. and Maechler, M. (2007). A fast and efficient implementation of qualitatively constrained quantile smoothing splines. *Statistical Modelling*, **7**(4), 315–328.

Qu, Z. and Yoon, J. (2015). Nonparametric estimation and inference on conditional quantile processes. *Journal of Econometrics*, **185**(1), 1–19.

Reed, C. and Yu, K. (2010). Efficient gibbs sampling for bayesian quantile regression. Technical report, Brunel University Mathematics Technical Report.

Ruppert, D., Sheather, S., and Wand, M. (1995). An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association*, **90**, 1257–1270.

Schaumburg, J. (2012). Predicting extreme value at risk: Nonparametric quantile regression with refinements from extreme value theory. *Computational Statistics & Data Analysis*, **56**, 4081–4096.

Serdyukova, N. (2012). Spatial adaptation in heteroscedastic regression: Propagation approach. *Electron. J. Stat.*, **6**, 861–907.

Spokoiny, V. and Vial, C. (2009). Parameter tuning in pointwise adaptation using a propagation approach. *Ann. Statist.*, **37**, 2783–2807.

Spokoiny, V., Wang, W., and Härdle, W. K. (2013). Local quantile regression. *Journal of Statistical Planning and Inference*, **143**(7), 1109–1129.

Takezawa, K. (2005). *Introduction to nonparametric regression*, volume 606. John Wiley & Sons.

Taylor, J. W. and Yu, K. (2016). Using auto-regressive logit models to forecast the exceedance probability for financial risk management. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*.

Theil, H. (1966). Applied economic forecasting.

Wand, M. and Jones, M. (1995). Kernel smoothing. 1995. *Chapman&Hall, London.*

Yu, K. and Jones, M. C. (1998). Local linear quantile regression. *Journal of the American Statistical Association*, **93**, 228–237.

Yu, K. and Lu, Z. (2004). Local linear additive quantile regression. *Scandinavian Journal of Statistics*, **31**(3), 333–346.