A Two-Stage Nonparametric Quantile Regression

Rachel Clemens and Mei Ling Huang*†

Department of Mathematics & Statistics, Brock University,
St. Catharines, Ontario, Canada L2S 3A1

August 31, 2020

Abstract

Estimating extreme conditional quantiles is an important problem. Many studies on this problem use a quantile regression (QR) method. The regular QR method often sets a linear model, which estimates the coefficients in the model to obtain the estimated conditional quantile. The real-world applications may be restricted by this approach's model setting. This project proposes a two-stage direct nonparametric extrapolation quantile regression method to overcome this restriction. Monte Carlo simulations show good efficiency for the proposed direct nonparametric QR extrapolation estimator relative to the linear QR extrapolation estimator. This project also investigates an example of rainfall in Toronto, Canada using the proposed method with comparisons to the linear methods.

Keywords: *Conditional quantile, extrapolation, extreme value distribution, Fréchet distribution, goodness of fit, multivariate kernel density estimation.*

AMS 2010 Subject Classifications: primary: 62G32; secondary: 62J05

1. Introduction

Extreme value events occur in many fields, such as financial markets, weather, industrial engineering, actuarial science, survival analysis, queueing networks, and other stochastic models. When statisticians are interested in estimating high quantiles of heavy-tailed distributions of extreme events, they often face theoretical difficulties. It is important to estimate extreme conditional quantiles of a random variable y given a variable vector $\mathbf{x} = (x_1, x_2, ..., x_d)^T \in R^d$. We recall the mean regression and original linear quantile regression models.

The mean linear regression model assumes

$$\mu_{y|\mathbf{x}} = E(y|x_1, x_2, ..., x_d) = \mathbf{x}_p^T \boldsymbol{\beta} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_d x_d. \quad (1)$$

*Corresponding author. E-mail: mhuang@brocku.ca.

†This research is supported by *the Natural Science and Engineering Research Council of Canada (NSERC) grant MLH, RGPIN-2019-04206.*

We estimate $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_d)^T \in R^p$ from a random sample $\{(y_i, \mathbf{x}_{pi}), i = 1, ..., n\} \in R \times R^p$, where $p = d + 1$, $\mathbf{x}_{pi} = (1, x_{i1}, x_{i2}, ..., x_{id})^T$ is the p-dimensional design vector and y_i is the univariate response variable from a continuous distribution with c.d.f. $F(y)$. The least squares (LS) estimator $\widehat{\boldsymbol{\beta}}_{LS}$ is a solution to the following equation

$$\widehat{\boldsymbol{\beta}}_{LS} = \arg\min_{\boldsymbol{\beta} \in R^p} \sum_{i=1}^{n} (y_i - \mathbf{x}_{pi}^T \boldsymbol{\beta})^2, \tag{2}$$

that is, $\widehat{\boldsymbol{\beta}}_{LS}$ is obtained by minimizing the L_2-distance.

The mean linear regression provides the mean relationship between a response variable and explanatory variables. However, there are limitations present in the conditional mean models. To address this concern, quantile regression estimates the conditional quantiles of y given $\mathbf{x}$.

$$Q_Y(\tau|\mathbf{x}) = \inf(t : F_Y(t|\mathbf{x}) \geq \tau\} = F_Y^{-1}(\tau|\mathbf{x}), \quad 0 < \tau < 1. \tag{3}$$

Koenker and Bassett (1978) proposed a linear quantile regression model for estimating the true τth conditional quantile $Q_Y(\tau|\mathbf{x})$ in *(3)*, which is defined as

$$Q_L(\tau|\mathbf{x}) = \mathbf{x}_p^T \boldsymbol{\beta}(\tau) = \beta_0(\tau) + \beta_1(\tau)x_1 + \cdots + \beta_d(\tau)x_d, \quad 0 < \tau < 1, \tag{4}$$

where $\boldsymbol{\beta}(\tau) = (\beta_0(\tau), \beta_1(\tau), \beta_2(\tau), ..., \beta_d(\tau))^T$.

In model *(4)*, we estimate the coefficient $\boldsymbol{\beta}(\tau) = (\beta_0(\tau), \beta_1(\tau), \beta_2(\tau), ..., \beta_d(\tau))^T \in R^p$ from a random sample $\{(y_i, \mathbf{x}_i),\ i = 1, ..., n\}$ by using an L_1-weighted loss function ρ_τ to obtain estimator $\widehat{\boldsymbol{\beta}}(\tau)$,

$$\widehat{\boldsymbol{\beta}}(\tau) = \arg\min_{\boldsymbol{\beta}(\tau) \in R^p} \sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_{pi}^T \boldsymbol{\beta}(\tau)),\ 0 < \tau < 1, \tag{5}$$

where ρ_τ is a loss function, namely

$$\rho_\tau(u) = u(\tau - I(u < 0)) = \begin{cases} u(\tau - 1),\ u < 0; \\ u\tau,\ u \geq 0. \end{cases}$$

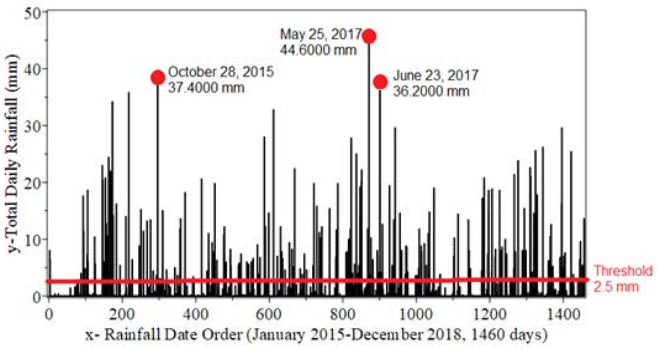
The linear quantile regression model in *(4)* needs the estimator in *(5)* to find the conditional quantile curves. The estimation of extremely high or low conditional quantile curves may be restricted by this model setting. We are motivated by rainfall in Toronto example in this section.

Example: Rainfall in Toronto, Canada (2015-2018)

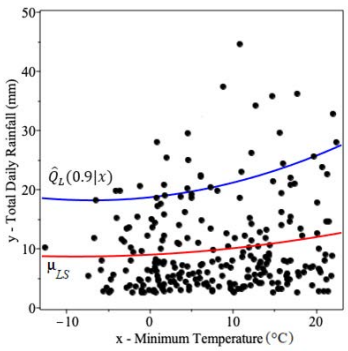
Large amounts of rainfall can be detrimental. Most obviously, torrential rainfall is associated with flooding and mudslides. Both flooding and mudslides can cause injury and death to humans and can also cause costly damage to houses and infrastructure. As a result, it is important to study torrential rain. In this paper, we will explore rainfall amounts in Toronto, Ontario.

A data set from January 2015 to December 2018 was taken from the Climate Change Canada (2018) (http://climate.weather. gc.ca/climate_data). Figure 1(a) shows the total daily rainfalls for the original 1460 days. The x-axis represents the rainfall in the order of occurrence and the y-axis represents the total daily rainfall in millimeters (mm). According to Climate

Change Canada (2019), rainfall less than 2.5 millimeters per hour (mm/hr) is classified as light rain, whereas rainfall above 2.6 mm/hr is considered moderate and rainfall above 7.6 mm/hr is considered heavy. We apply a threshold of 2.5 mm reducing the sample size to $n = 244$. The three highest daily rainfalls are highlighted in Figure 1(a). We set random variable y as total daily rainfall measured in millimeters, which is related to the minimum temperatures x.



(a) Rainfall in Toronto 2015-2018.



(b) μ_{LS} and $\widehat{Q}_L(0.9|x)$.

Figure 1. (a) Total daily rainfall in Toronto from January 2015 to December 2018, $n^* = 1460$ days, and a threshold of 2.5 mm (red); (b) Scatter diagram of total daily rainfall vs. minimum temperature with least squares mean regression curve μ_{LS} in *(2)* (red) and the linear quantile regression curve $\widehat{Q}_L(\tau|\mathbf{x})$ in *(4)* at $\tau = 0.90$ (blue), $n = 244$ daily total rainfalls over 2.5 mm.

Let y represent total daily rainfall in millimeters and x represent minimum daily temperature in degrees Celsius *(^oC)*, based on the $n = 244$ data, a mean regression estimate based on *(2)* is

$$\mu_{LS} = \widehat{\mu}_{y|\mathbf{x}} = \widehat{E}(y|x) = 8.9968 + 0.0705x + 0.0039x^2,$$

The $\tau = 0.90$ linear quantile regression estimate from *(4)* is

$$\widehat{Q}_L(0.90|x) = 18.7049 + 0.1452x + 0.0104x^2.$$

Both mean regression and linear quantile regression models have limitations for estimating extreme conditional quantiles. In Section 5 of this paper, we will apply a new proposed quantile regression method to this example.

When analyzing extreme value events, the response variable y has a heavy-tailed distribution. In recognition of this complication, many studies have focused on looking for improvements of estimator *(5)*. Some studies have used nonparametric quantile regression and two-stage quantile regression, for example, Chernozhukov and Du (2008); Daouia *et al.* (2011); Gardes and Girard (2011); Wang and Li (2013); Huang and Nguyen (2018).

To overcome the limitation of the model setting in *(1)* and *(4)*, we propose a two-stage direct nonparametric quantile regression method. The method uses the ideas of kernel density estimation and nonparametric kernel regression under the assumption that the response variable Y is heavy-tailed distributed. The proposed method is not only different from most of the existing nonparametric quantile regression methods, it also overcomes the crossing problem of estimating quantile curves. We also show the improvements of the proposed method relative to the two-stage linear quantile regression.

2. Proposed Extrapolation for Extreme Conditional Quantiles

We assume Y has a heavy-tailed distribution and we ignore the idea of the linear model *(4)* to obtain a nonparametric kernel estimator for the true conditional quantile in *(3):*

$$\widehat{Q}_Y(\tau|\mathbf{x}) = \inf(t : \widehat{F}_Y(t|\mathbf{x}) \geq \tau\} = \widehat{F}_Y^{-1}(\tau|\mathbf{x}),\ 0 < \tau < 1 \tag{6}$$

by using a given random sample, $\{(Y_i, \mathbf{X}_i), i = 1, ..., n\}$, for $\mathbf{X}_i = (x_{i1}, x_{i2}, ..., x_{id})^T$.

2.1. Extreme Value Distribution

Let $X_1, X_2, ..., X_n$ be independent and identically distributed (i.i.d.) continuous random variables with c.d.f. $F(x)$. It is important to study the limiting behavior of the $\max(X_1, X_2, ..., X_n)$ and $\min(X_1, X_2, ..., X_n)$. The main interest of extreme value theory (EVT) is in finding possible limiting distributions of the sample maxima of i.i.d. random variables. Any non-degenerate distribution that can be derived as such a limit is called an *extreme value distribution* (de Haan and Ferreira, 2006).

Definition 1. *(Fisher and Tippett, 1928, Gnedenko, 1943) The c.d.f. of any extreme value distribution is of the form $G_\gamma(ax + b)$ for some constants $a > 0$, $b \in R$, where*

$$G_\gamma(x) = \begin{cases} 1 - \exp(-(1+\gamma x)^{-1/\gamma}), & 1 + \gamma x > 0 \text{ and } \gamma \neq 0; \\ 1 - \exp(-e^{-x}), & \gamma = 0, \end{cases} \tag{7}$$

where the parameter γ is called the extreme value index (EVI).

Note that when $\gamma > 0$, the corresponding densities for both $G_\gamma(x)$ and $G_\gamma(ax+b)$ have heavier tails than the exponential distribution, which are referred to as *heavy-tailed distributions*. In many applications, it is important to include observations that take extremely high or low values in the statistical analysis. To estimate the index γ, we have

Definition 2. *(Hill Estimator, Hill, 1975) Consider a random sample $X_1, X_2, ...X_n$ with sample size n from the distribution in (7), and let $X_{1,n} \leq X_{2,n} \leq ... \leq X_{n,n}$ denote its order statistics. Hill (1975) introduced a maximum likelihood estimators (MLE) for the index γ as*

$$\widehat{\gamma}_{Hill} = \frac{1}{k}\sum_{i=1}^{k}(\log\{X_{n-i,n}\} - \log\{X_{n-k,n}\}) = \frac{1}{k}\sum_{i=0}^{k-1}\log\{X_{n-i,n}\} - \ln X_{n-k,n}, \tag{8}$$

where $X_{i,n}$ represents the ith order statistics, $i = 1, ..., n$, and $\widehat{\gamma}_{Hill}$ uses the k largest order statistics.

A conditional extreme value distribution for exceeding a threshold is the *generalized Pareto distribution* (GPD).

Definition 3. *(Pickands, 1975) The c.d.f. $H_\gamma(x)$ and its corresponding probability density function (p.d.f.) $f(x)$ of the two-parameter GPD(γ, σ) with shape parameter $\gamma > 0$ and scale parameter σ of a random variable X are given by*

$$H\gamma(x) = 1 - \left(1 + \gamma\frac{x}{\sigma}\right)^{1/\gamma}, \quad \gamma, \sigma > 0, \quad x > 0. \tag{9}$$

One natural choice for modeling such data is to consider heavy-tailed distributions that can provide a good fit for these extreme value events.

2.2. 6-Step Algorithm in Two-Stages of a Direct Nonparametric Extrapolation Method

In this paper, we propose the following 6-step algorithm divided in two stages to construct a direct nonparametric quantile regression estimator.

Stage 1: Direct Nonparametric Quantile Regression (5-Step)

Step 1: Estimate the conditional density of y for given $\mathbf{x} = (x_1,x_2,...,x_d)$ using a kernel density estimation method (Silverman, 1986; Scott, 2015):

$$\widehat{f}(y|\mathbf{x}) = \frac{\widehat{f}(y,\mathbf{x})}{\widehat{g}(\mathbf{x})},$$

where $\widehat{f}(y,\mathbf{x})$ is an estimator of the joint density of y and $\mathbf{x}$, and $\widehat{g}(\mathbf{x})$ is an estimator of the marginal density of $\mathbf{x}$.

A d-dimensional kernel density estimator from a random sample $\mathbf{X}_i = (X_{1i},X_{2i},...,X_{di})$, $i = 1,2,...,n$, from a population $\mathbf{x} = (x_1,x_2,...,x_d)$ for joint density $g(\mathbf{x})$, is given by

$$\widehat{g}(\mathbf{x}) = \frac{1}{nh^d}\sum_{i=1}^{n} K\left\{\frac{\mathbf{x}-\mathbf{X}_i}{h}\right\},$$

where $h > 0$ is the bandwidth and the kernel function $K(\mathbf{x})$ is a function defined for d-dimensional $\mathbf{x} = (\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_d)$ which satisfies $\int_{R^d} K(\mathbf{x})d\mathbf{x} = 1$.

If a multivariate normal kernel is used for smoothing the data, we use

$$h_{opt} = \left\{\frac{4}{d+2}\right\}^{1/(d+4)} n^{-1/(d+4)}.$$

Step 2: Estimate the conditional c.d.f. of y given $\mathbf{x}$:

$$\widehat{F}_Y(y|\mathbf{x}) = \int_{-\infty}^{y} \widehat{f}(y|\mathbf{x})dy.$$

Step 3: Normalization: Normalize the estimated conditional c.d.f. $\widehat{F}_Y(y|\mathbf{x}) \in [0,1]$.
Step 4: Localization: Estimate the local conditional quantile function $\xi(\tau|\mathbf{x})$ of y given $\mathbf{x}$ by inverting an estimated conditional c.d.f. $\widehat{F}_Y(y|\mathbf{x})$.

$$\widehat{\xi}(\tau|\mathbf{x}) = \widehat{Q}_Y(\tau|\mathbf{x}) = \inf\{y : \widehat{F}_Y(y|\mathbf{x}) \geq \tau\} = \widehat{F}_Y^{-1}(\tau|\mathbf{x}). \quad (10)$$

To avoid the computational difficulties of $\widehat{\xi}(\tau|\mathbf{x})$, we estimate the local conditional quantile function $\xi_i(\tau|\mathbf{x}_i)$ of y given $\mathbf{x}_i$ by inverting c.d.f. $\widehat{F}_Y(y|\mathbf{x}_i)$ at the ith data point $(y_i,\mathbf{x}_i)$:

$$\widehat{\xi}_i(\tau|\mathbf{x}_i) = \widehat{Q}_Y(\tau|\mathbf{x}_i) = \inf\{y : \widehat{F}_Y(y|\mathbf{x}_i) \geq \tau\} = \widehat{F}_Y^{-1}(\tau|\mathbf{x}_i), \quad i = 1,2,...,n.$$

Step 5: We propose a direct nonparametric quantile regression estimator for the τth conditional quantile curve of $\mathbf{x}$ by using Nadaraya-Watson (NW) nonparametric regression estimator on $\left(\mathbf{x}_i, \widehat{\xi}_i(\tau|\mathbf{x}_i)\right),\ i = 1, 2, ..., n$:

$$Q_D(\tau|\mathbf{x}) = \widehat{\xi}(\tau|\mathbf{x}) = \frac{\sum_{i=1}^{n} K\left\{\frac{\mathbf{x}-\mathbf{X}_i}{\mathbf{h}}\right\} \widehat{\xi}_i(\tau|\mathbf{x}_i)}{\sum_{j=1}^{n} K\left\{\frac{\mathbf{x}-\mathbf{X}_j}{\mathbf{h}}\right\}} = \sum_{i=1}^{n} W_{h_{\mathbf{x}}}(\mathbf{x}, \mathbf{X}_i)\widehat{\xi}_i(\tau|\mathbf{x}_i), \quad 0 < \tau < 1, \qquad (11)$$

where $W_{h_x}(\mathbf{x}, \mathbf{X}_i)$ is called an equivalent kernel, and $\mathbf{h} = (h_1, ..., h_d)$,

$$W_{h_{\mathbf{x}}}(\mathbf{x}, \mathbf{X}_i) = \frac{K\left\{\frac{\mathbf{x}-\mathbf{X}_i}{\mathbf{h}}\right\}}{\sum_{j=1}^{n} K\left\{\frac{\mathbf{x}-\mathbf{X}_j}{\mathbf{h}}\right\}}, \quad i = 1, 2, ..., n,$$

where

$$K\left\{\frac{\mathbf{x}-\mathbf{X}_i}{\mathbf{h}}\right\} = \frac{1}{nh_1...h_d} \prod_{j=1}^{d} K\left(\frac{x - x_{ij}}{h_j}\right), \quad i = 1, ..., n,$$

where K is the kernel function, and $h_j > 0$ is the bandwidth for the jth dimension.

Stage 2: Extrapolation (1-Step)

Step 6: Extrapolation for extreme conditional quantile,

$$\widehat{Q}_{DE}(\tau_n|\mathbf{x}) = \widehat{Q}_D(\alpha_n|\mathbf{x}) \left(\frac{1-\alpha_n}{1-\tau_n}\right)^{\widehat{\gamma}_D^H(\mathbf{x})}, \qquad (12)$$

where α_n is an intermediate quantile level, and $\widehat{Q}_D(\alpha_n|\mathbf{x})$ is intermediate quantiles defined in *(11)* for extreme order of quantiles, as $\tau_n \to 1$ at a rate faster than $1/n$.

$$n(1-\tau_n) > (\log(n))^p, \quad 0 \le \tau_n \le 1, \quad \left(\frac{1-\tau_n}{1-\alpha_n}\right) \to 0, \quad as \quad n \to \infty,$$

which means $(1-\tau_n) \to 0$ faster than $(1-\alpha_n) \to 0$ as $n \to \infty$, and $\widehat{\gamma}_D^H(\mathbf{x})$ is a kernel version of Hill estimator in *(8)* for conditional EVI $\gamma(\mathbf{x})$.

$$\widehat{\gamma}_D^H(\mathbf{x}) = \frac{\sum_{j=1}^{J} \left[\log\left\{\widehat{Q}_D(1 - w_j(1-\alpha_n^*)|\mathbf{x})\right\} - \log\left\{\widehat{Q}_D(\alpha_n^*)|\mathbf{x})\right\}\right]}{\sum_{j=1}^{J} \log(1/w_j)}, \quad 0 < w_j < 1, \quad j = 1, ..., J, \qquad (13)$$

where $w_1 > w_2 > ... > w_J > 0$ is a decreasing sequence of weights and J is a positive integer. $\widehat{Q}_D(\alpha_n^*|\mathbf{x})$ in *(11)* is the intermediate quantile curve starting at α_n^* level, then the intermediate

estimated quantile curves $\widehat{Q}_D(1-w_j(1-\alpha_n^*)|\mathbf{x}), \quad j=1,...,J$, are order statistics. Consequently, the quantile curves are non-crossing.

Remark: The extrapolation condition is that Y given a variable vector $\mathbf{x}$ must have conditional c.d.f. $F_Y(\bullet|\mathbf{x})$ that belongs to the Fréchet maximum domain of attraction with conditional extreme value index (EVI) $\gamma(\mathbf{x})$ for any $(\mathbf{x},y) \in R^p \times R$. The τth conditional quantile $Q_Y(\tau|\mathbf{x})$ is defined by *(3)*. The kernel estimator is not feasible as it cannot extrapolate beyond the maximum observation in the ball centered at $\mathbf{x}$ with radius h. (Daouia et al., 2011). This extrapolation allows the estimation of extreme conditional quantiles with $\tau_n \to 1$ arbitrarily fast.

In this paper, we compare the $\widehat{Q}_{DE}(\tau_n|\mathbf{x})$ in *(12)* with Chernozhukov and Du (2008) proposed method. They outlined an assumption that allows the quantile slope coefficient $\boldsymbol{\beta}(\tau)$ in model *(4)* to cross τ. Assume model *(4)* applies after being transformed by some auxiliary regression line, then the response variable Y has regularly varying tails with EVI $\gamma > 0$. Suppose that there exists an auxiliary slope $\boldsymbol{\beta}(\tau)$ such that the following tail-equivalence relationship holds as $\tau_n \to 1$,

$$\widehat{Q}_{LE}(\tau_n|\mathbf{x}) = \widehat{Q}_L(\alpha_n|\mathbf{x}) \left(\frac{1-\alpha_n}{1-\tau_n}\right)^{\widehat{\gamma}_L^H(\mathbf{x})}, \tag{14}$$

where α_n is an intermediate quantile level, and $\widehat{Q}_L(\alpha_n|\mathbf{x})$ is an estimated intermediate quantile given by *(5)*, and $\widehat{\gamma}_L^H(\mathbf{x})$ is

$$\widehat{\gamma}_L^H(\mathbf{x}) = \frac{1}{n(1-\alpha_n)} \sum_{i=1}^{n} \log\left(\frac{y_i}{\mathbf{X}_{pi}^T \widehat{\beta}(\alpha_n)}\right). \tag{15}$$

3. Monte Carlo Simulation

In this section, we run Monte Carlo simulations to investigate the efficiency of the proposed direct nonparametric extrapolation estimator $\widehat{Q}_{DE}(\tau|\mathbf{x})$ in *(12)* relative to the linear extrapolation estimator $\widehat{Q}_{LE}(\tau|\mathbf{x})$ in *(14)*. We generate m random samples with size n each from one-dimensional random variables X $(p=1)$ uniformly distributed on $E=[0,1]$. Suppose Y given $X=x$ is Fréchet distributed (de Haan and Ferreira, 2006). The conditional c.d.f. is

$$F(y|x) = e^{(-y^{-1/\gamma(x)})}, \quad 0 \le x \le 1, \quad \gamma(x) > 0, \tag{16}$$

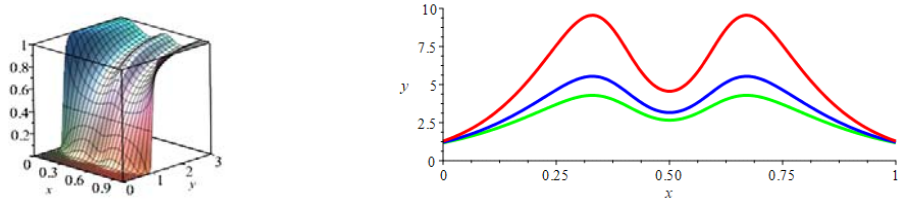
with the conditional tail index

$$\gamma(x) = -\frac{1}{2}\left(\frac{1}{10} + \sin(\pi x)\right)\left(\frac{11}{10} - \frac{1}{2}e^{(-64(x-1/2)^2}\right), \quad 0 \le x \le 1.$$

The true conditional quantile of *(16)* is

$$Q_Y(\tau|x) = (-\log(\tau))^{-\gamma(x)}, \quad 0 \le x \le 1, \quad 0 < \tau < 1. \tag{17}$$

We apply the two estimators $\widehat{Q}_{DE}(\tau|x)$ in *(12)* and $\widehat{Q}_{LE}(\tau|x)$ in *(14)* to estimate the true conditional quantile $Q_Y(\tau|x)$ in *(17)*. For each method, we generate size $n=200$, $m=100$ samples. $\widehat{Q}_{DE,i}(\tau|x)$ and $\widehat{Q}_{LE,i}(\tau|x)$, $i=1,2,...,m$, are estimated in the ith sample.



(a) Joint Distribution. (b) True conditional quantiles.

Figure 2. (a) The joint c.d.f. of Fréchet distribution of y and uniform distribution of x. (b) The true conditional quantiles at $\tau = 0.99$ (red), $\tau = 0.97$ (blue), $\tau = 0.95$ (green).

The simulation mean squared errors (SMSE) of the estimators *(12)* and *(14)* are:

$$SMSE\left(\widehat{Q}_{DE}(\tau)\right) = \frac{1}{m}\sum_{i=1}^{m}\int_0^1 \left(\widehat{Q}_{DE,i}(\tau|x) - Q_Y(\tau|x)\right)^2 dx; \quad (18)$$

$$SMSE\left(\widehat{Q}_{LE}(\tau)\right) = \frac{1}{m}\sum_{i=1}^{m}\int_0^1 \left(\widehat{Q}_{LE,i}(\tau|x) - Q_Y(\tau|x)\right)^2 dx, \quad (19)$$

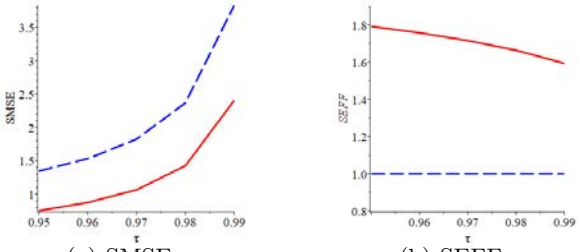
where the true τth conditional quantile $Q_Y(\tau|x)$ is in *(17)*. The simulation efficiencies (SEFF) are given by

$$SEFF\left(\widehat{Q}_{DE}(\tau)\right) = \frac{SMSE\left(\widehat{Q}_{LE}(\tau|x)\right)}{SMSE\left(\widehat{Q}_{DE}(\tau|x)\right)},$$

where $SMSE(\widehat{Q}_{DE}(\tau))$ and $SMSE(\widehat{Q}_{LE}(\tau))$ are defined in *(18)* and *(19)*, respectively.

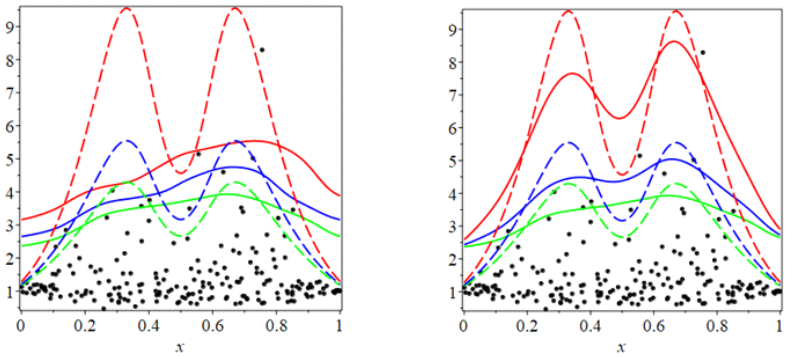
Table 1. Simulation mean squared errors (SMSE) and efficiencies (SEFF) of $\widehat{Q}_{DE}(\tau|\mathbf{x})$ relative to $\widehat{Q}_{LE}(\tau|\mathbf{x})$ estimating $Q_Y(\tau|x), m = 100$, $n = 200$.

τ	0.95	0.96	0.97	0.98	0.99
$SMSE(\widehat{Q}_{LE}(\tau))$	1.3439	1.5308	1.8242	2.3678	3.8316
$SMSE(\widehat{Q}_{DE}(\tau))$	0.7501	0.8712	1.0633	1.4235	2.4051
$SEFF(\widehat{Q}_{DE}(\tau))$	**1.7916**	**1.7571**	**1.7155**	**1.6633**	**1.5931**



(a) SMSE. (b) SEFF.

Figure 3. $m = 100, n = 200$, (a) Simulation MSE of $\widehat{Q}_{DE}(\tau|x)$ (red) and $\widehat{Q}_{LE}(\tau|x)$ (blue-dash); (b) Simulation Efficiency (SEFF) of $\widehat{Q}_{DE}(\tau|x)$ (red) relative to $\widehat{Q}_{LE}(\tau|x)$ (blue dash).



(a) Stage 1: $\widehat{Q}_D(\tau|x)$. (b) Stage 2: $\widehat{Q}_{DE}(\tau|x)$.

Figure 4. The true quantile curves (dash), $m = 1, n = 200$, data points are black dots, (a) Stage 1: The $\widehat{Q}_D(\tau|\mathbf{x})$ quantile estimates; (b) Stage 2: The $\widehat{Q}_{DE}(\tau|\mathbf{x})$ quantile estimates for $\tau = 0.99$ (red), $\tau = 0.97$ (blue) and $\tau = 0.95$ (green).

From the simulation results, we conclude that

(1) Table 1 and Figure 3 show that all of the $SEFF\left(\widehat{Q}_{DE}(\tau)\right) > 1$ when $\tau = 0.95$,..., 0.99. We can conclude that using the proposed direct nonparametric extrapolation estimator $\widehat{Q}_{DE}(\tau|x)$ in *(12)* is more efficient relative to the extrapolation based on linear quantile regression estimator $\widehat{Q}_{LE}(\tau|x)$ in *(14)*.

(2) Figure 4 shows that the Stage 2 $\widehat{Q}_{DE}(\tau|x)$ estimate curves are much closer to the true conditional quantile curves than the Stage 1 $\widehat{Q}_D(\tau|x)$ curves. Thus, the Stage 2 extrapolations contribute to a more accurate estimator.

4. Comparison of Goodness-of Fit on Quantile Regression Models

To compare the estimator $\widehat{Q}_{DE}(\tau|x)$ in *(12)* and the estimator $\widehat{Q}_{LE}(\tau|x)$ in *(14)*, we extend the idea of measuring goodness-of-fit by Koenker and Machado (1999). We suggest using a Relative $R_{DE}(\tau)$ of $\widehat{Q}_{DE}(\tau|x)$ to $\widehat{Q}_{LE}(\tau|x)$, $0 < \tau < 1$, which is defined as

$$Relative\ R_{DE}(\tau) = 1 - \frac{V_{DE}(\tau)}{V_{LE}(\tau)}, \quad -1 \le R_{DE}(\tau) \le 1, \quad \text{where} \tag{20}$$

$$V_{DE}(\tau) = \sum_{y_i \ge \widehat{Q}_{DE}(\tau|\mathbf{x}_i)} \frac{\tau}{n} \left| y_i - \widehat{Q}_{DE}(\tau|\mathbf{x}_i) \right| + \sum_{y_i < \widehat{Q}_{DE}(\tau|\mathbf{x}_i)} \frac{(1-\tau)}{n} \left| y_i - \widehat{Q}_{DE}(\tau|\mathbf{x}_i) \right|,$$

where $\widehat{Q}_{DE}(\tau|\mathbf{x}_i)$ is obtained by *(12)*, and

$$V_{LE}(\tau) = \sum_{y_i \ge \widehat{Q}_{LE}(\tau|\mathbf{x}_i)} \frac{\tau}{n} \left| y_i - \widehat{Q}_{LE}(\tau|\mathbf{x}_i) \right| + \sum_{y_i < \widehat{Q}_{LE}(\tau|\mathbf{x}_i)} \frac{(1-\tau)}{n} \left| y_i - \widehat{Q}_{LE}(\tau|\mathbf{x}_i) \right|,$$

where $\widehat{Q}_{LE}(\tau|\mathbf{x}_i)$ is given by *(14)*.

5. Rainfall in Toronto, Canada (2015-2018)

We recall the rainfall in Toronto example from Section 1. The mean regression and linear quantile regression models in *(1)* and *(4)* are

$$\begin{aligned} E(y|x) &= \beta_0 + \beta_1 x + \beta_2 x^2, \\ Q_Y(\tau|x) &= \beta_0(\tau) + \beta_1(\tau)x + \beta_2(\tau)x^2, \quad 0 < \tau < 1. \end{aligned}$$

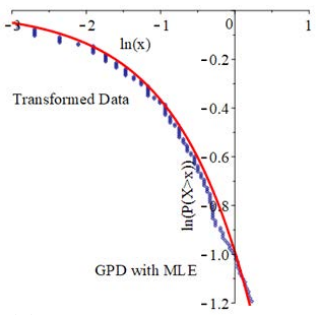
where y is the total daily rainfall in millimeters and x is the minimum daily temperature in ^oC.

To overcome the limitations and deficiency of the above models for estimating the extreme quantiles, in this section, we use extrapolation starting at $\alpha_n^* = 0.95$ level and $J = 9$. We will compare the proposed $\widehat{Q}_{DE}(\tau|x)$ with the $\widehat{Q}_{LE}(\tau|x)$:

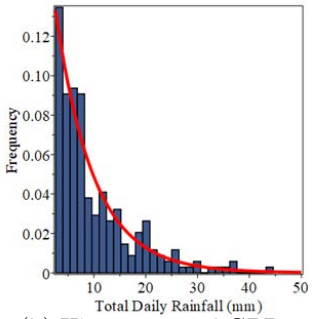
1. The linear extrapolation estimator in $\widehat{Q}_{LE}(\tau|x)$ in *(14)* uses intermediate quantile $\widehat{Q}_L(\alpha|x)$ given by *(5)* obtained by linear programming and $\widehat{\gamma}_L^H(\mathbf{x})$ in *(15)*.

2. The direct nonparametric extrapolation estimator $\widehat{Q}_{DE}(\tau|x)$ in *(12)* uses the intermediate quantile $\widehat{Q}_D(\alpha|x)$ in *(11)* obtained by nonparametric 6-step algorithm and and $\widehat{\gamma}_D^H(\mathbf{x})$ in *(13)* in Section 2.

We must confirm that y is heavy-tailed distributed to be able to use the proposed extrapolation estimators $\widehat{Q}_{DE}(\tau|x)$ and $\widehat{Q}_{LE}(\tau|x)$. Figure 5(a) and (b) show the log-log plot and histogram of the rainfall in Toronto data along with the GPD curve in *(9)*. The estimated GPD curve follows the shape of the rainfall data very well.



(a) Log-log Plot and GPD.

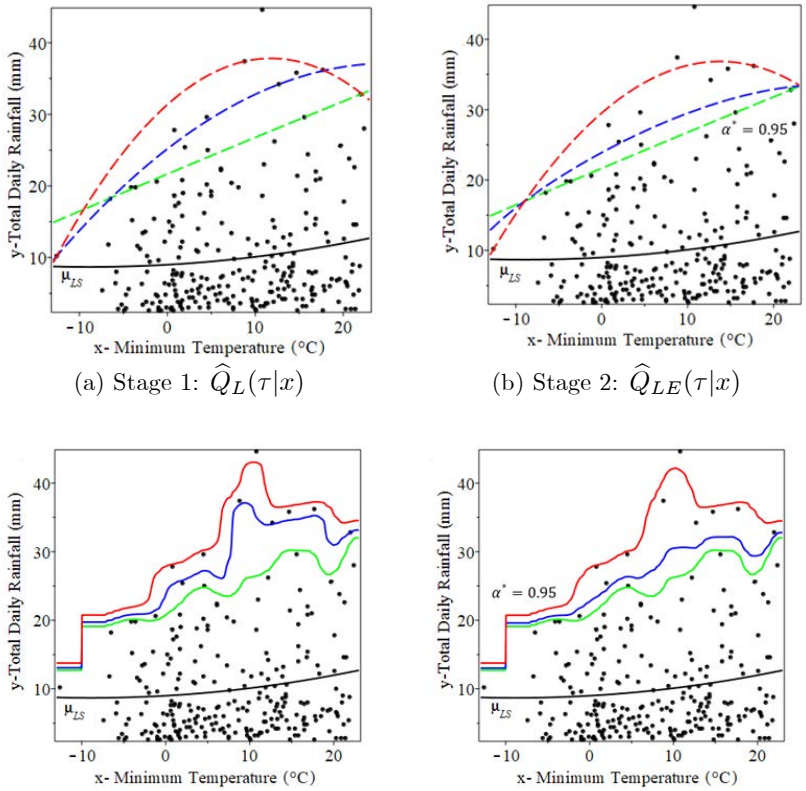


(b) Histogram and GPD.

Figure 5. (a) Log-Log plot of total daily rainfalls (blue dots) and GPD curve (red line); (b) Histogram of the total daily rainfalls with the GPD curve (red line), $n = 244$.

Figure 6 shows the fitted $\widehat{Q}_L(\tau|x)$, $\widehat{Q}_D(\tau|x)$ estimate curves from Stage 1 and fitted $\widehat{Q}_{LE}(\tau|x)$, $\widehat{Q}_{DE}(\tau|x)$ estimate curves from Stage 2. We observe that the proposed $\widehat{Q}_{DE}(\tau|x)$ curves fit the data much better than the $\widehat{Q}_{LE}(\tau|x)$ curves. It also displays that the proposed Stage 2 estimators $\widehat{Q}_{LE}(\tau|x)$ and $\widehat{Q}_{DE}(\tau|x)$ using extrapolations are not over fitting the extreme data as much as the Stage 1 estimators $\widehat{Q}_D(\tau|x)$ and $\widehat{Q}_L(\tau|x)$.

Table 2 shows the values of the Relative $R_{DE}(\tau)$ in *(20)* for given $\tau = 0.95, 0.96, 0.97, 0.98, 0.99, 0.995, 0.999$. We note that $R_{DE}(\tau) > 0$, which means that $V_{DE}(\tau) < V_{LE}(\tau)$. Thus, the $\widehat{Q}_{DE}(\tau|x)$ is a better fit to the data than the $\widehat{Q}_{LE}(\tau|x)$.



(a) Stage 1: $\widehat{Q}_L(\tau|x)$ (b) Stage 2: $\widehat{Q}_{LE}(\tau|x)$

(c) Stage 1: $\widehat{Q}_D(\tau|x)$ (d) Stage 2: $\widehat{Q}_{DE}(\tau|x)$

Figure 6. For rainfall in Toronto data, scatter plot (black dots). (a) Stage 1: $\widehat{Q}_L(\tau|x)$; (b) Stage 2: $\widehat{Q}_{LE}(\tau|x)$; (c) Stage 1: $\widehat{Q}_D(\tau|x)$; (d) Stage 2: $\widehat{Q}_{DE}(\tau|x)$ at $\tau = 0.95$ (green); $\tau = 0.97$ (blue), $\tau = 0.99$ (red), with LS mean regression curves (black), $n = 244$.

Table 2. Relative $R_{DE}(\tau)$ of $\widehat{Q}_{DE}(\tau|x)$ to $\widehat{Q}_{LE}(\tau|x)$ for the Rainfall in Toronto example.

	$\tau = 0.95$	$\tau = 0.96$	$\tau = 0.97$	$\tau = 0.98$	$\tau = 0.99$	$\tau = 0.995$	$\tau = 0.999$
$R_{DE}(\tau)$	0.0146	0.0118	0.0175	0.0313	0.1207	0.1357	0.3123

Study of the rainfall in Toronto example demonstrates that the proposed two-stage estimator $\widehat{Q}_{DE}(\tau|\mathbf{x})$ predicts that for moderate temperatures the rainfall amounts are more substantial in Toronto, Ontario. For higher temperatures, over 15 ^oC, it is less likely that we see torrential rainfall. We can conclude that the problematic torrential rains are related to transitional weather. We are more likely to see heavy rainfall in the fall or spring. More importantly, we have seen that our proposed method is a useful model for predicting heavy rainfall.

6. Conclusions

After the studies above, we can conclude:

1. In heavy-tailed population cases, the proposed two-stage direct nonparametric extrapolation method for predicting extreme conditional quantiles has advantages over other existing methods.

2. The Monte Carlo simulation shows that the proposed two-stage direct nonparametric extrapolation method is more efficient than the two-stage linear quantile regression method.

3. The proposed two-stage direct nonparametric extrapolation quantile regression method has useful applications in extreme events. This method proved useful in predicting torrential rainfall in the Toronto Rainfall example.

References

[1] Chernozhukov, V. and Du, S. (2008). Extreme quantiles and value-at risk, *The New Palgrave Dictionary of Economics, eds.* Durlauf, S. N. and Blume, L. E., Basingstoke: Palgrave Macmillian.

[2] Climate Change Canada. (2018). Station Results - Historical Data. Retrieved from http://climate.weather.gc.ca/climate_data

[3] Climate Change Canada. (2019). Glossary. Retrieved from http://climate.weather.gc.ca/glossary_e.html#r

[4] Daouia, A., Gardes, L., Girard, S. and Lekina, A. (2011). Kernel estimators of extreme level curves, *Test*, 20, 311-333.

[5] de Haan, L. and Ferreira, A. (2006). *Extreme Value Theory: An Introduction.* Springer, New York.

[6] Fréchet, M. (1927). Sur loi de probabilté de l'écart maximun, *Ann, Soc. Math. Polon,* 6, 93-116.

[7] Gardes, L. and Girard, S. (2011). Functional kernel estimators of conditional extreme quantiles, *Recent Advances in Functional Data Analysis and Related Topics Contributions to Statistics,* Springer, 135-140.

[8] Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution, *The Annals of Statistics, Vol.3, No.5, 1163-1174.*

[9] Huang, M. L. and Nguyen, C. (2018). A nonparametric approach for quantile regression, *Journal of Statistical Distributions and Applications,* Springer, Vol. 5(1), 1-14.

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

[11] Scott, D. W. (2015). *Multivariate Density Estimation, Theory, Practice and Visualization.* 2nd edition, John Wiley & Sons, New York.

[12] Silverman, B. W. (1986). *Density estimation for Statistics and Data Analysis.* Chapman & Hall, London, UK.

[13] Wang, H. J. and Li, D. (2013). Estimation of extreme conditional quantile through power transformation, *Journal of the American Statistical Association,* 108(503), 1062-1074.