Statistica Sinica Preprint No: SS-2020-0051	
Title	Extreme Quantile Estimation Based on the Tail Single-index Model
Manuscript ID	SS-2020-0051
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202020.0051
Complete List of Authors	Wen Xu, Huixia Judy Wang and Deyuan Li
Corresponding Author	Deyuan Li
E-mail	deyuanli@fudan.edu.cn

Statistica Sinica

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0051

Extreme Quantile Estimation Based on the Tail Single-index Model

Wen Xu[1], Huixia Judy Wang[2] and Deyuan Li[1]

[1] *Fudan University,* [2] *The George Washington University*

Abstract: It is important to quantify and predict rare events that have significant societal effects. Existing works on analyzing such events rely mainly on either inflexible parametric models or nonparametric models that are subject to "the curse of dimensionality." We propose a new semiparametric approach based on the tail single-index model to obtain a better balance between model flexibility and parsimony. The procedure involves three steps. First, we obtain a $\sqrt{n}$-estimator of the index parameter. Next, we apply the local polynomial regression to estimate the intermediate conditional quantiles. Lastly, these quantiles are extrapolated to the tails to estimate the extreme conditional quantiles. We establish the asymptotic properties of the proposed estimators. Furthermore, we demonstrate using a simulation and an analysis of Los Angeles mortality and air pollution data that the proposed method is easy to compute and leads to more stable and accurate estimations than those of alternative methods.

Key words and phrases: extreme quantile, local linear regression, semi-parametric, single-index, tail.

1. Introduction

An important problem in fields such as econometrics, finance, hydrol-

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0051

ogy, and climate science is to model and predict events that are rare, but that have significant consequences. Examples include a large financial loss, heavy snowfall, extreme temperatures, high medical costs, and a low birth weight, among others. For such data, modeling and estimating the tail quantiles are of more interest than doing so for the mean. Numerous works have examined the estimation of extreme quantiles for univariate data; see Embrechts et al. (2013) and De Haan and Ferreira (2006), and the references therein.

To predict rare events, it would be helpful to quantify the tail quantiles of the response by accounting for information provided by relevant predictors (covariates). Studies on conditional tail quantiles can be roughly divided into two classes. The first models extreme conditional quantiles by fitting either parametric distributions, such as a generalized extreme value distribution or a generalized Pareto distribution (GPD), or linear quantile regression models; see Davison et al. (1990), Beirlant and Goegebeur (2003), Beirlant and Goegebeur (2004), Chavez et al. (2005), Chernozhukov (2005), Wang and Tsai (2009), Wang et al. (2012), Wang and Li (2013), and Li and Wang (2019). These methods assume that the conditional quantiles are parametric functions of the covariates and, thus, are not flexible in some applications. The second class estimates extreme quantiles by fitting non-

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0051

parametric models; see Gardes et al. (2010), Gardes et al. (2012), Daouia et al. (2011), and Daouia et al. (2013). These methods are based on local estimations using observations in a small neighbourhood. Thus, the finite-sample behavior depends heavily on the richness of the data in the neighborhood. However, owing to the "curse of dimensionality," these methods generally do not work well when the number of covariates increases.

To overcome the "curse of dimensionality," while still allowing for model flexibility, we propose a new extreme quantile estimation method based on a tail single-index model. The single-index model is a semiparametric regression model that captures the nonlinear relationship between the response and the covariates using an unspecified univariate link function and the index, an unknown linear combination of covariates. Therefore, the model provides a convenient tool to overcome the "curse of dimensionality" encountered in nonparametric regressions with multivariate covariates; see Powell et al. (1989) and Haedle et al. (1993). Some works have integrated the single-index model and quantile regression; see Wu et al. (2010), Zhu et al. (2012), Kong and Xia (2012), and Zhong et al. (2016), among others. To the best of our knowledge, only one work (Gardes, 2018) discusses the estimation of extreme conditional quantiles for single-index and multi-index models. Gardes (2018) proposed a new dimension-reduction approach and

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0051

a conditional extremal quantile estimator by considering the tail dimension-reduction subspace. However, this method is computationally complex, and the authors do not formally establish the theoretical properties of the estimator when the index parameters are unknown and have to be estimated from the data.

In this study, we consider a new tail single-index model that assumes there exists a single-index structure at the tail and, thus, is less restrictive than the global single-index models assumed in Zhu et al. (2012) and Zhong et al. (2016). The estimation of the extreme conditional quantiles involves estimating three unknown quantities, namely, the index parameter, link function, and extreme value index that characterizes the heaviness of the tail distribution. We propose a convenient three-step estimator for the extreme conditional quantiles based on the tail single-index model. In the first step, we construct a $\sqrt{n}$-estimator of the unknown index parameter under a misspecified linear quantile regression model at a central quantile level close to the tail. In the second step, we apply a local polynomial regression technique (Fan and Gijbels, 1996) to estimate the intermediate conditional quantiles. These estimates are then extrapolated in the third step to extreme tails by adapting the univariate extreme value theory to the regression setup. Our method provides a convenient and flexible tool

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0051

to analyze rare events by considering the effects of multiple covariates with possibly large dimensions.

Our proposed method differs from existing works in the following ways. First, to the best of our knowledge, this is the first work to systematically examine the extreme quantile estimation using single-index models, and to provide theoretical guarantees for cases with unknown index parameters. Second, the proposed tail single-index model not only provides more flexibility than parametric models, but also leads to a simple approach for estimating the index parameters with a $\sqrt{n}$-convergence rate. As a result, this estimation does not affect the asymptotic properties of the final extreme quantile estimation. In contrast, the index estimation method in Gardes (2018) is more complicated and numerically less stable, and its theoretical properties and effects on the extreme quantile estimation have not been studied formally. Third, instead of indirectly estimating the conditional quantiles by inverting the conditional cumulative distribution function, as in Gardes (2018), our procedure is based on a direct estimation of the conditional quantiles in all three steps. We show that this coherence helps reduce errors from different layers of the modeling and ameliorates the tuning parameter selection, leading to numerically more accurate estimations. Furthermore, the direct estimation helps quantify the effect of the covari-

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0051

ates on the extreme tails of the response in a more straightforward and interpretable way.

The rest of this paper is organized as follows. In Section 2, we present the proposed method and investigate its theoretical properties. In Section 3, we assess the finite-sample performance of the proposed method using a simulation study and an analysis of Los Angeles mortality and air pollution data. Section 4 concludes the paper. All technical details are given in the online Supplementary Material.

2. Methodology

2.1 Notation and the tail single-index model

Let Y be the response variable of interest, and $F_Y(\cdot|\mathbf{X})$ be the cumulative distribution function (CDF) of Y conditional on the covariate $\mathbf{X} = (X_1, X_2, ..., X_p)^T$. Denote $Q_\tau(Y|\mathbf{X})$ as the τth conditional quantile of Y given $\mathbf{X}$, namely, $Q_\tau(Y|\mathbf{X}) = \inf\{y : F_Y(y|\mathbf{X}) \leq \tau\}$. Suppose that we observe a random sample $\{(\mathbf{X}_i, Y_i), i = 1, 2, ..., n\}$ from $(\mathbf{X}, Y)$. Our main objective is to estimate the extreme conditional high quantile $Q_{\tau_n^*}(Y|\mathbf{X})$. Here, τ_n^* may approach one at any rate, including special cases, such as the intermediate quantiles with $n(1-\tau_n^*) \to \infty$ and the extreme quantiles with $n(1-\tau_n^*) \to c \geq 0$. For simplicity, we denote $\tau_n^* = \tau^*$.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0051

Throughout, we assume that the conditional distribution of $Y|\mathbf{X}$ for the given $\mathbf{X}$ belongs to the maximum domain of attraction of some extreme value distribution $H_{\gamma(\mathbf{X})}$ with the extreme value index (EVI) $\gamma(\mathbf{X})$, denoted by $Y|\mathbf{X} \in D(H_{\gamma(\mathbf{X})})$. That means, for independent and identically distributed (i.i.d.) sample $\{U_i : i = 1, 2, ..., n\}$ from the conditional distribution of $Y|\mathbf{X}$, there exist $a_n > 0$ and $b_n \in R$, such that

$$P\Big(\frac{\max_{i=1,\dots,n} U_i - b_n}{a_n} \leq u\Big) \to H_{\gamma(\mathbf{X})}(u) := \exp\{-(1+\gamma(\mathbf{X})u)^{-1/\gamma(\mathbf{X})}\},$$

as $n \to \infty$, for all u with $1+\gamma(\mathbf{X})u > 0$. We assume $\gamma(\mathbf{X}) > 0$, which means that $Y|\mathbf{X}$ has a heavy-tailed distribution. Such distributions are common in applications such as financial returns and insurance claims, and the heavy tails often make the estimation of extreme quantiles more challenging.

Here, we consider a new tail single-index model, which assumes that there exists $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ and the unknown function $G_\tau(\cdot)$, such that

$$Q_\tau(Y|\mathbf{X}) = G_\tau(\mathbf{X}^T\boldsymbol{\beta}_0) \quad \text{for} \quad \tau \in (\tau_c, 1), \tag{2.1}$$

where τ_c is a fixed quantile level close to one. For model identifiability, we assume throughout that $||\boldsymbol{\beta}_0|| = 1$, where $||\cdot||$ denotes the L_2 norm. Model (2.1) requires that the single-index structure holds only in the right tail, which is a weaker assumption than the global single-index quantile regression model considered by Zhu et al. (2012) and Zhong et al. (2016).

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0051

2.2 Three-step estimation

We propose a three-step estimation procedure. The first step estimates the index parameter $\boldsymbol{\beta}_0$. The second step estimates the unknown link function G_τ and the conditional quantile at intermediate quantile levels. In the third step, we use extrapolation and extreme value theory to estimate $Q_{\tau^*}(Y|\mathbf{X})$.

We first discuss the estimation of the index parameter $\boldsymbol{\beta}_0$. Zhu et al. (2012) and Zhong et al. (2016) showed that under the global single-index quantile regression model and some conditions on $\mathbf{X}$, the direction of $\boldsymbol{\beta}_0$ can be estimated consistently using the slope estimation obtained by fitting a misspecified linear quantile regression model. We show in Proposition 2.1 that this result still holds under the tail single-index model (2.1) and a relaxed assumption on $\mathbf{X}$.

Let $\rho_\tau(r) = \tau r - r\mathbb{I}(r < 0)$ be the quantile check loss function (Koenker et al., 2005), and $\mathcal{L}_\tau(u, \boldsymbol{\beta}) = E\{\rho_\tau(Y - u - \mathbf{X}^T\boldsymbol{\beta}) - \rho_\tau(Y)\}$. Define

$$(u_\tau, \boldsymbol{\beta}_\tau) = \operatorname*{argmin}_{u,\boldsymbol{\beta}} \mathcal{L}_\tau(u, \boldsymbol{\beta}), \tag{2.2}$$

which are the population parameters resulting from fitting the misspecified linear quantile regression model.

Proposition 2.1. Let $\tau \in (\tau_c, 1)$ be a given quantile level. Under model

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0051

(2.1), if the covariate vector $\mathbf{X}$ satisfies

$$E(\mathbf{X}|\boldsymbol{\beta}_0^T\mathbf{X}) = \mathbf{C}\boldsymbol{\beta}_0^T\mathbf{X}, \tag{2.3}$$

where $\mathbf{C}$ is a p-dimensional constant vector, then $\boldsymbol{\beta}_\tau = k\boldsymbol{\beta}_0$, for some constant k.

When $\mathbf{X}$ follows an elliptically symmetric distribution (e.g., the normal distribution), the linearity assumption (2.3) is satisfied. Li (1991) and Hall et al. (1993) showed that the linearity condition (2.3) is typically regarded as mild, particularly when p is fairly large.

Proposition 2.1 implies that the direction of $\boldsymbol{\beta}_\tau$, defined in (2.2) for $\tau \in (\tau_c, 1)$, is the same as that of $\boldsymbol{\beta}_0$. Obviously, because $||\boldsymbol{\beta}_0|| = 1$, k is the L_2-norm of $\boldsymbol{\beta}_\tau$. Hence, the conditional distribution of $Y|(\mathbf{X}^T\boldsymbol{\beta}_0)$ is equivalent to that of $Y|(\mathbf{X}^T\boldsymbol{\beta}_\tau)$. Based on the observed data, we obtain the sample version of $(u_\tau, \boldsymbol{\beta}_\tau)$ as $(\hat{u}_\tau, \hat{\boldsymbol{\beta}}_\tau) = \underset{u,\boldsymbol{\beta}}{\operatorname{argmin}}\, \mathcal{L}_{\tau n}(u, \boldsymbol{\beta})$, where $\mathcal{L}_{\tau n}(u, \boldsymbol{\beta}) = n^{-1}\sum_{i=1}^n \rho_\tau(Y_i - u - \mathbf{X}_i^T\boldsymbol{\beta})$.

We propose estimating the index parameter $\boldsymbol{\beta}_0$ using $\hat{\boldsymbol{\beta}}_{\tau_0}$ at $\tau_0 \in (\tau_c, 1)$. Theoretically, τ_0 can be any value in $(\tau_c, 1)$, and this results in a $\sqrt{n}$-consistent estimator of $\boldsymbol{\beta}_0$. The following proposition presents the asymptotic normality of $(\hat{u}_{\tau_0}, \hat{\boldsymbol{\beta}}_{\tau_0})^T$.

Proposition 2.2. Let $\epsilon = Y - \mathbf{X}^T\boldsymbol{\beta}_{\tau_0}$, and denote $F_\epsilon(t|\mathbf{X})$ and $f_\epsilon(\cdot|\mathbf{X})$ as

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0051

the conditional CDF and conditional density function of ϵ given $\mathbf{X}$, respectively. Then,

$$\begin{pmatrix} n^{1/2}(\hat{u}_{\tau_0} - u_{\tau_0}) \\ n^{1/2}(\hat{\boldsymbol{\beta}}_{\tau_0} - \boldsymbol{\beta}_{\tau_0}) \end{pmatrix} \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\Sigma} = \mathbf{B}^{-1}\mathbf{V}\mathbf{B}^{-1}$, with

$$\mathbf{B} = \begin{pmatrix} E\{f_\epsilon(u_{\tau_0}|\mathbf{X})\} & E\{\mathbf{X}^T f_\epsilon(u_{\tau_0}|\mathbf{X})\} \\ E\{\mathbf{X} f_\epsilon(u_{\tau_0}|\mathbf{X})\} & E\{\mathbf{X}\mathbf{X}^T f_\epsilon(u_{\tau_0}|\mathbf{X})\} \end{pmatrix}, \quad \mathbf{V} = Var\begin{pmatrix} F_\epsilon(u_{\tau_0}|\mathbf{X}) - \tau_0 \\ \mathbf{X}\{F_\epsilon(u_{\tau_0}|\mathbf{X}) - \tau_0\} \end{pmatrix}.$$

Remark 1. We propose estimating the index parameter using the linear quantile slope estimator $\hat{\boldsymbol{\beta}}_{\tau_0}$ at a central quantile level τ_0, because this estimator is $\sqrt{n}$-consistent to $\boldsymbol{\beta}_{\tau_0}$. We can also use the estimator $\hat{\boldsymbol{\beta}}_{\tau_0}$ at an intermediate extreme quantile level $\tau_0 \to 1$ and $n(1-\tau_0) \to \infty$. For this case, we can follow similar arguments to those in Chernozhukov (2005) and Angrist (2006) to establish the asymptotic normality of $\hat{\boldsymbol{\beta}}_{\tau_0}$, but this estimator has a lower convergence rate of $\sqrt{n} f_Y\{G_{\tau_0}(\mathbf{x})|\mathbf{x}\}/\sqrt{1-\tau_0}$, where the $f_Y\{G_{\tau_0}(\mathbf{x})|\mathbf{x}\}$ is the conditional density function of Y evaluated at the τ_0th conditional quantile given $\mathbf{X} = \mathbf{x}$.

In the second step, we estimate the intermediate conditional quantiles of Y by applying the local linear quantile regression, and then use the results to estimate the EVI. For ease of presentation, let $z = \mathbf{X}_0^T \boldsymbol{\beta}_{\tau_0}$, $\hat{z} = \mathbf{X}_0^T \hat{\boldsymbol{\beta}}_{\tau_0}$, $Z_i = \mathbf{X}_i^T \boldsymbol{\beta}_{\tau_0}$, and $\hat{Z}_i = \mathbf{X}_i^T \hat{\boldsymbol{\beta}}_{\tau_0}$. Note that by Model (2.1) and Proposition

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0051

2.1, we have $Q_\tau(Y|\mathbf{X}) = Q_\tau(Y|\mathbf{X}^T\boldsymbol{\beta}_0) = Q_\tau(Y|\mathbf{X}^T\boldsymbol{\beta}_{\tau_0})$. Using the pseudo sample data $\{(\mathbf{X}_i^T\hat{\boldsymbol{\beta}}_{\tau_0}, Y_i) : i = 1, 2, ..., n\}$, we can estimate $G_\tau(\mathbf{X}_0^T\boldsymbol{\beta}_{\tau_0})$ for a given new $\mathbf{X}_0$ using a local linear regression. For Z in the neighborhood of z, $G_\tau(Z)$ can be approximated as $G_\tau(Z) \approx G_\tau(z) + G'_\tau(z)(Z - z)$. Define $(\hat{a}, \hat{b}) = \underset{a,b}{\operatorname{argmin}}\, n^{-1}\sum_{i=1}^n \rho_\tau\{Y_i - a - b(\hat{Z}_i - \hat{z})\}K\left(\frac{\hat{Z}_i - \hat{z}}{h}\right)$. Let $\hat{G}_\tau(\hat{z}) = \hat{a}$ and $\hat{G}'_\tau(\hat{z}) = \hat{b}$. We can estimate $G_\tau(z)$ as $\hat{G}_\tau(\hat{z})$ at a sequence of quantile levels $\tau_j = 1 - j/n$, with $j = \lceil n^\eta \rceil, ..., k$, for $0 < \eta < 1$, where $\lceil a \rceil$ denotes the ceiling function that returns the smallest integer greater than or equal to a, k satisfies $k = k(n) \to \infty$, $k/n \to 0$, and $\lceil n^\eta \rceil = o(k^{1/2})$.

We can then estimate the EVI $\gamma(\mathbf{x}) = \gamma(z)$ based on the estimated intermediate quantiles $\{\hat{G}_{\tau_j}(\hat{z}) : j = \lceil n^\eta \rceil, \lceil n^\eta \rceil + 1, ..., k\}$. For heavy-tailed distributions, a commonly used estimator for the extreme value index is Hill's estimator. We propose estimating $\gamma(z)$ using the following Hill-type estimator:

$$\hat{\gamma}(\hat{z}) = \frac{1}{k}\sum_{j=\lceil n^\eta \rceil}^{k}\Big[\log\{\hat{G}_{\tau_j}(\hat{z})\} - \log\{\hat{G}_{\tau_k}(\hat{z})\}\Big].$$

In the third step, we adapt the univariate extreme value theory and extrapolate from the intermediate quantile level to the extreme tail to estimate the extreme conditional quantile $G_{\tau^*}(z)$ for $\tau^* \to 1$. Specifically, by adapting Weissman's estimator to the conditional case (Weissman, 1978),

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0051

we obtain the extreme conditional quantile estimator,

$$\hat{G}_{\tau^*}(\hat{z}) = \left(\frac{1-\tau_k}{1-\tau^*}\right)^{\hat{\gamma}(\hat{z})} \hat{G}_{\tau_k}(\hat{z}),$$

where $\tau_k = 1-k/n$. In addition to Hill-type estimators, we can consider alternative methods for estimating the EVI, such as the moment estimator of Li and Wang (2019), Pickands estimator of Daouia et al. (2013), and peaks over random threshold (PORT) estimator of Santos et al. (2006). Our numerical study (in Section S3 of the Supplementary Material) suggests that the proposed extreme conditional quantile estimator is stable for different EVI estimators.

3. Theoretical properties

In order to derive the asymptotic properties of $\hat{\gamma}(\hat{z})$ and $\hat{G}_{\tau^*}(\hat{z})$, we need to assume some second-order condition. A positive function h is called regularly varying at infinity with index $\alpha \in \mathbb{R}$, denoted by $h \in RV(\alpha)$, if $\lim_{t\to\infty} h(tx)/h(t) = x^{\alpha}$, for $x > 0$. Let $U(t; z) = G_{1-1/t}(z)$. We assume the following second-order condition:

C_1 There exists a function $A(t; z) \in RV(\varrho(z))$, for some $\varrho(z) \leq 0$ and $A(t; z) \to 0$, as $t \to \infty$, such that

$$\frac{\frac{U(tx;z)}{U(t;z)} - x^{\gamma(z)}}{A(t;z)} \to x^{\gamma(z)} \frac{x^{\varrho(z)}-1}{\varrho(z)}, \quad x > 0. \tag{3.1}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0051

Most families of continuous distributions satisfy condition (3.1), for instance, the t distribution and the Pareto distribution. We also need the following regularity conditions:

C_2 The quantile function $G_\tau(Z)$ has a continuous and bounded second derivative $G''_\tau(Z)$ with respect to Z.

C_3 The density function of $\mathbf{X}^T\boldsymbol{\beta}$ is positive and uniformly continuous for $\boldsymbol{\beta}$ in a neighborhood of $\boldsymbol{\beta}_0$. Furthermore, the density function of $Z = \mathbf{X}^T\boldsymbol{\beta}_0$ is continuous and bounded away from zero and infinity on its support.

C_4 The conditional density of Y given $\mathbf{x}^T\boldsymbol{\beta}_0$, $f_Y(y|\mathbf{x}^T\boldsymbol{\beta}_0)$, is continuous in $\mathbf{x}^T\boldsymbol{\beta}_0$, for each $y \in \mathbb{R}$. Moreover, there exist positive constants ε and δ and a positive function $\bar{f}(y|\mathbf{x}^T\boldsymbol{\beta}_0)$, such that $\sup_{||\mathbf{x}^T\boldsymbol{\beta}-\mathbf{x}^T\boldsymbol{\beta}_0||\le\varepsilon} f_Y(y|\mathbf{x}^T\boldsymbol{\beta}) \le \bar{f}(y|\mathbf{x}^T\boldsymbol{\beta}_0)$. For any fixed value of $\mathbf{x}^T\boldsymbol{\beta}_0$, $\int \bar{f}(y|\mathbf{x}^T\boldsymbol{\beta}_0)dy < \infty$; furthermore, as $t \to 0$, $\int\{\rho_\tau(y-t) - \rho_\tau(y) - \dot{\rho}_\tau(y)t\}^2 \bar{f}(y|\mathbf{x}^T\boldsymbol{\beta}_0)dy = o(t^2)$, where $\dot{\rho}_\tau(u) = \{\text{sgn}(u) + (2\tau - 1)\}/2$, for $u \le 0$ and $\dot{\rho}_\tau(0) = 0$.

C_5 The kernel function $K(\cdot)$ is symmetric with a compact support $[-1, 1]$, and satisfies the first-order Lipschitz condition.

C_6 $U(t; z) = G_{1-1/t}(z)$ has the first-order derivative $U'(t; z)$ with respective to t, and satisfies $\lim_{t\to\infty} tU'(t; z)/U(t; z) = \gamma(z)$ uniformly for z

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0051

in a compact support $\mathcal{Z}$.

Condition C_2 is a common assumption in semiparametric regression for the true link function. Condition C_3 presents assumptions on the density of the single index. Condition C_4 is a mild condition that is weaker than the Lipschitz condition on the function $\dot{\rho}_\tau(\cdot)$. Condition C_5 requires the kernel function to be a proper density function with a compact support. Condition C_6 includes some classic assumptions on the extreme value index and the distribution function in extreme value theory.

Theorems 1-3 present the asymptotic properties of the conditional quantile estimator at the intermediate order, extreme value index estimator, and extrapolation estimator of the extreme conditional quantile, respectively. Throughout this paper, we denote $\mu_2 = \int_{-1}^{1} u^2 K(u)du$ and $\nu_0 = \int_{-1}^{1} K^2(u)du$.

Theorem 1. *Suppose that model (2.1) and conditions C_2-C_6 hold. Define $\mathcal{T} = \{\tau_m < \cdots < \tau_k\}$, with $m = \lceil n^\eta \rceil$ for $0 < \eta < 1, \tau_j = 1 - j/n$ for $j = \lceil n^\eta \rceil, ..., k$, where k satisfies $k = k(n) \to \infty$, $k/n \to 0$, and $\lceil n^\eta \rceil = o(k^{1/2})$. If $h \to 0$ and $nh \to \infty$ as $n \to \infty$, we have*

$$\frac{\{nh(1-\tau)\}^{1/2}}{\gamma(z)G_\tau(z)}\big\{\hat{G}_\tau(\hat{z}) - G_\tau(z) - \frac{1}{2}h^2 G''_\tau(z)\mu_2\big\} = W_n(\tau)\{1 + o_p(1)\}$$

uniformly for $\tau \in \mathcal{T}$, where $W_n(\tau) = \{nh(1-\tau)\}^{-1/2} f_Z^{-1}(z) \sum_{i=1}^{n}[\tau -$

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0051

$I\{Y_i \leq G_\tau(Z_i)\}]K_i$, which converges to a Gaussian process with mean zero and covariance $\mathbf{\Sigma}(\tau_t, \tau_s) = \nu_0\{\min(\tau_t, \tau_s) - \tau_t\tau_s\} f_Z^{-1}(z)/\sqrt{(1-\tau_t)(1-\tau_s)}$, where $K_i = K\{(Z_i - z)/h\}$ and $f_Z(z)$ is the density function of $Z = \mathbf{X}^T\boldsymbol{\beta}_{\tau_0}$.

Theorem 2. *Suppose that the conditions in Theorem 1 and the second-order condition (3.1) hold with $\gamma(z) > 0$ and $\varrho(z) < 0$, and k and h satisfy $kh \to \infty$, $(kh)^{1/2}h^2 \log(n/k) \to \lambda_1 \in \mathbb{R}$, and $(kh)^{1/2}A(n/k; z) \to \lambda_2 \in \mathbb{R}$. Then, there exist a sequence of Brownian motions $\{\tilde{W}_n(t) : t \in [0, 1]\}$, such that*

$$(kh)^{1/2}\Big\{\hat{\gamma}(\hat{z}) - \gamma(z) - \frac{A(n/k; z)}{1-\varrho(z)} - \tilde{I}_{3n}(z)\Big\}$$
$$= \gamma(z)\sqrt{\frac{\nu_0}{f_Z(z)}}\int_0^1 \{x^{-1}\tilde{W}_n(x) - \tilde{W}_n(1)\}dx + o_p(1),$$

where

$$\tilde{I}_{3n}(z) = \begin{cases} h^2\mu_2 \log(n/k)(\gamma'(z))^2, & \gamma'(z) \neq 0, \\ \frac{1}{2}h^2\mu_2 \frac{d(z)(\varrho'(z))^2}{c(z)}(\frac{n}{k})^{\varrho(z)}\{\log(n/k)\}^2 \frac{\varrho(z)}{1-\varrho(z)}, & \gamma'(z) = 0, \varrho'(z) \neq 0, \\ -\frac{1}{2}h^2\mu_2\Big\{\frac{c''(z)d(z)}{c^2(z)} - \frac{d''(z)}{c(z)}\Big\}\big(\frac{n}{k}\big)^{\varrho(z)}\frac{\varrho(z)}{1-\varrho(z)}, & \gamma'(z) = \varrho'(z) = 0. \end{cases}$$

Remark 2. By Theorem 2, the asymptotic bias of $\hat{\gamma}(\hat{z})$ consists of two parts, $\tilde{I}_{3n}(z)$ and $A(n/k; z)/\{1 - \varrho(z)\}$. The first $\tilde{I}_{3n}(z)$ is from the kernel estimation, and the second $A(n/k; z)/\{1 - \varrho(z)\}$ is the result of the second-order approximation to the conditional distribution $F_Y(y|\mathbf{X})$. The

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0051

convergence rate of $\hat{\gamma}(\hat{z})$ is $(kh)^{1/2}$, which is slower than $k^{1/2}$ for the ordinary Hill estimator in the univariate extreme analysis without a kernel estimation.

Theorem 3. *Assume that the conditions in Theorem 2 hold. Then, we have*

$$\frac{(kh)^{1/2}}{\log\{k/(np_n)\}}\Big\{\frac{\hat{G}_{\tau^*}(\hat{z})}{G_{\tau^*}(z)}-1-\frac{1}{2}h^2G_{\tau_k}^{-1}(z)G''_{\tau_k}(z)\mu_2+A(\frac{n}{k};z)\frac{(\frac{k}{np_n})^{\varrho(z)}-1}{\varrho(z)}\Big\}$$
$$=\gamma(z)\sqrt{\frac{\nu_0}{f_Z(z)}}\int_0^1\big\{x^{-1}\tilde{W}_n(x)-\tilde{W}_n(1)\big\}dx(1+o_p(1)),$$

where $p_n=1-\tau^*$, $\tau^*\to 1$, $k/(np_n)\to\infty$, *and* $(kh)^{-1/2}\log\{k/(np_n)\}\to 0$.

Remark 3. Similarly to $\hat{\gamma}(\hat{z})$, the asymptotic bias of $\hat{G}_{\tau^*}(\hat{z})$ consists of two parts. The first, $(1/2)h^2G_{\tau_k}^{-1}(z)G''_{\tau_k}(z)\mu_2$, is from the kernel estimation, and the second, $-A(n/k;z)[\{k/(np_n)\}^{\varrho(z)}-1]/\varrho(z)$, is from the second-order approximation of the conditional distribution of Y. The convergence rate of the extreme conditional quantile estimator obtained under the single-index model is $(kh)^{1/2}[\log\{k/(np_n)\}]^{-1}$, which slower than the rate of $k^{1/2}[\log\{k/(np_n)\}]^{-1}$ under the parametric regression models. In addition, the condition $k/(np_n)\to\infty$ implies τ^* approaches one at a faster rate than τ_k does, which makes the extrapolation feasible.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0051

4. Tuning parameters selection

4.1 Bandwidth selection

The bandwidth h balances the bias and the variance: a smaller h leads to a smaller modeling bias, but a larger variance. We can choose h by minimizing the mean squared error (MSE) of the nonparametric conditional quantile estimator at an intermediate quantile level τ. By Theorem 1, at an intermediate quantile level τ, where $\tau \to 1$ and $n(1-\tau) \to \infty$, we have

$$MSE\{\hat{G}_\tau(\hat{z})\} = \frac{1}{4}h^4 G_\tau''^2(z)\mu_2^2 + \frac{\gamma^2(z)G_\tau^2(z)\nu_0\tau f_Z^{-1}(z)}{nh(1-\tau)}.$$

Minimizing $MSE\{\hat{G}_\tau(\hat{z})\}$ gives

$$\begin{aligned} h^{opt}(\hat{z}) &= \Big[\frac{\gamma^2(z)G_\tau^2(z)\nu_0\tau f_Z^{-1}(z)}{(1-\tau)\{G_\tau''(z)\}^2\mu_2^2}\Big]^{1/5} n^{-1/5} \\ &\approx \Big[\frac{\nu_0\tau(1-\tau)f_Z^{-1}(z)}{f_Y^2\{G_\tau(z)|z\}\{G_\tau''(z)\}^2\mu_2^2}\Big]^{1/5} n^{-1/5}. \end{aligned} \quad (4.1)$$

The approximation in (4.1) is from $f_Y\{G_\tau(z)|z\} \approx (1-\tau)\{\gamma(z)G_\tau(z)\}^{-1}$, by Condition C_6.

Fan and Gijbels (1996) showed that in local linear mean regression, the optimal bandwidth is

$$h_m^{opt}(\hat{z}) = \Big[\frac{\nu_0\sigma^2(z)}{\mu_2^2 f_Z(z)\{G''(z)\}^2}\Big]^{1/5} n^{-1/5}, \quad (4.2)$$

where $G(z)$ and $\sigma^2(z)$ are the conditional mean and the variance of Y given the covariate z, respectively.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0051

Combining (4.1) and (4.2), we have

$$h^{opt}(\hat{z}) \approx h_m^{opt}(\hat{z}) \Big[\frac{\tau(1-\tau)\{G''(z)\}^2}{\sigma^2(z) f_Y^2\{G_\tau(z)|z\}\{G_\tau''(z)\}^2} \Big]^{1/5}. \tag{4.3}$$

The optimal bandwidth in (4.3) depends on the unknown conditional density function $f_Y(\cdot|z)$ and $G''(z)$. For simple calculation, we take the following approximations: (1) assume that the curvatures of the quantile function $G_\tau''(z)$ and the conditional mean function $G''(z)$ are similar; (2) take $\sigma^2(z) f_Y^2\{G_\tau(z)|z\} = \phi^2\{\Phi^{-1}(\tau)\}$ under the normal distribution, where $\phi(\cdot)$ and $\Phi(\cdot)$ are the standard normal density and distribution functions, respectively.

Finally, we choose the bandwidth using the following rule of thumb:

$$\hat{h}^{opt}(\hat{z}) = \hat{h}_m^{opt}(\hat{z}) \Big[\frac{\tau(1-\tau)}{\phi^2\{\Phi^{-1}(\tau)\}} \Big]^{1/5},$$

where $\hat{h}_m^{opt}(\hat{z})$ can be attained using the plug-in method, using the "lpbwselect" function in the R package *nprobust*.

4.2 Selection of τ_0

The quantile level τ_0 is involved in estimating the index parameter $\boldsymbol{\beta}_0$. As discussed in Remark 1, we can choose a fixed $\tau_0 \in (\tau_c, 1)$, and this results in a $\sqrt{n}$-consistent estimation of $\boldsymbol{\beta}_0$. Alternatively, we can choose τ_0 at an intermediate quantile level, such that $\tau_0 \to 1$, and $n(1-\tau_0) \to \infty$. Correspond-

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0051

ingly, the convergence rate of $\hat{\boldsymbol{\beta}}_{\tau_0}$ is $\sqrt{n}f_Y\{G_{\tau_0}(\mathbf{x})|\mathbf{x}\}/\sqrt{1-\tau_0}$, which is slower than $\sqrt{n}$ from a fixed quantile level. If τ_0 also satisfies $h(1-\tau)/(1-\tau_0) \to 0$ uniformly for $\tau \in \mathcal{T}$, the conclusion of Theorem 1 still holds. The main reason is that by Condition C_6, we have $f_Y\{G_\tau(z)|z\} \approx (1-\tau)\{\gamma(z)G_\tau(z)\}^{-1}$ and, consequently, $\sqrt{nh(1-\tau)}/\{\gamma(z)G_\tau(z)\}[\sqrt{n}f_Y\{G_{\tau_0}(\mathbf{x})|\mathbf{x}\}/\sqrt{1-\tau_0}]^{-1} \to 0$. Therefore, the estimation error involved in $\boldsymbol{\beta}_0$ does not affect the asymptotic properties of the estimators of the EVI and extreme conditional quantile in Theorems 2 and 3, respectively. For instance, we can choose $\tau = 1 - n^\eta/n$ and $h = h^{opt}(\hat{z})$. Thus, the condition $h^{opt}(\hat{z})(1-\tau)/(1-\tau_0) \to 0$ is equivalent to $n^{-(7-6\eta)/5}/(1-\tau_0) \to 0$; that is, τ_0 approaches one at a slower rate than $n^{-(7-6\eta)/5}$. Because $0 < \eta < 1$, we suggest the following rule of thumb: $\tau_0 = 1 - cn^{-1/5}$, where c is a constant. Our numerical study in Section 5.1 suggests that this rule of thumb leads to a stable estimation for $c \in (0.1, 0.4)$.

5. Numerical studies

In this section, we investigate the finite-sample performance of our proposed method, referred to as the single-index model extreme quantile (SIMEXQ) method, using a simulation study and an analysis of the National Morbidity, Mortality, and Air Pollution Study (NMMAPS) data set of Los Angeles

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0051

(LA).

5.1 Simulation

We consider the following three models to generate the simulation data:

- Case 1 (univarite x): Conditional on $X = x$, Y is distributed from $F(y|x) = \exp\{-y^{-1/\gamma(x)}\}, y > 0$, where the extreme value index $\gamma(x) = \frac{1}{2}\left\{\frac{1}{10} + \sin(\pi x)\right\}\left[\frac{11}{10} - \frac{1}{2}\exp\{-64(x-1/2)^2\}\right]$. Therefore, the true extreme conditional quantile function is $Q_\tau(Y|x) = (-\log\tau)^{-\gamma(x)}$. The covariate X is generated from the standard uniform distribution $U(0,1)$. This model was also considered in Daouia et al. (2011).

- Cases 2 (single-index model): Conditional on $X = \mathbf{x}$, the response variable is generated from $Y = \sin\{2(\mathbf{x}^T\boldsymbol{\beta}_0)\} + 2\exp\{-16(\mathbf{x}^T\boldsymbol{\beta}_0)^2\} + (\mathbf{x}^T\boldsymbol{\beta}_0)\varepsilon$, where $\boldsymbol{\beta}_0 = (2,-2,-1,1,0,...,0)^T/\sqrt{10}$ is a $p\times 1$ vector, the covariate vector $\mathbf{X} = (X_1,\ldots,X_p)^T$ is multivariate normal with mean zero and covariance matrix $Cov(\mathbf{X}) = (\sigma_{ij})_{p\times p}$, with $\sigma_{ij} = 0.5^{|i-j|}$, and $\varepsilon \sim t(3)$ is the random noise. Therefore, the extreme value index is $\gamma(\mathbf{x}) = 1/3$. This model was also considered in Zhu et al. (2012). We consider $p = 4, 50$, and 100.

- Case 3 (tail dimension-reduction subspace): Conditional on $X = \mathbf{x}$,

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0051

the τth conditional quantile of Y for $\tau \in (0,1)$ is defined as $Q_\tau(Y|\mathbf{x}) = \{\ln(1/\tau)\}^{-g_0(\mathbf{B}_0^\top \mathbf{x})}\left[1 + g_1\left(\mathbf{B}_1^\top \mathbf{x}\right)\exp\{-(1-\tau)^{-1}\}\right]^{-1}$, where $\mathbf{x}^T = (x_1, ..., x_4)$, $\mathbf{B}_0^\top = (2,1,0,0)/\sqrt{5}$, $\mathbf{B}_1^\top = (0,0,1,1)$, $g_0(z) = \tilde{g}(z; 1/3, 8/3)$, $\tilde{g}(z; a, b) = a\mathbb{I}_{(-\infty,0)}(z) + \left(a + b\frac{\exp(2z)-1}{\exp(6/\sqrt{5})-1}\right)\mathbb{I}_{[0,3/\sqrt{5})}(z) + (a+b)\mathbb{I}_{[3/\sqrt{5},\infty)}(z)$, and $g_1(z) = \mathbb{I}_{(-\infty,0)}(z) + \exp(5z)\mathbb{I}_{[0,2)}(z) + \exp(10)\mathbb{I}_{[2,\infty)}(z)$. The covariates x_j, for $j = 1, ..., 4$ are generated as independent normal variables with mean 1/2 and variance 1/9. Gardes (2018) also considered this case, showing that the extreme value index $\gamma(\mathbf{x}) = g_0\left(\mathbf{B}_0^\top \mathbf{x}\right)$ in this model.

The EVI varies with the covariates in Cases 1 and 3, but is constant in Case 2. In Case 1, the conditional quantiles of Y depend on the univariate x. In Case 2, the tail single-index model assumption in (2.1) is satisfied. Case 3 is a multi-index model that depends on two indices and satisfies the TDR space assumption in Gardes (2018). As $\tau \to 1$, the quantile of Y depends on $\mathbf{x}$ approximately through the single index $\mathbf{B}_0^\top \mathbf{x}$. The sample size is set to $n = 1000$. For each scenario, the simulation is repeated 500 times. As suggested in Wang et al. (2012), we choose $k = \lceil 4.5n^{1/3} \rceil$ and $\eta = 0.1$ when estimating the EVI. We choose $\tau_0 = 1 - 0.2n^{-1/5}$, resulting in $\tau_0 = 0.95$ for $n = 1000$.

We include the following four methods for comparison: (i) the method of Beirlant and Goegebeur (2004), denoted by BG, which is based on the

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0051

local polynomial maximum likelihood estimation and the generalized Pareto distribution, fitted locally to excedances over a high specified threshold; (ii) the inverse CDF method of Daouia et al. (2011), denoted by ICDF, which first gets the estimator of the conditional kernel survival function, inverses it to get conditional quantile estimates, and then extrapolates these to estimate the extreme quantiles; (iii) the tail dimension-reduction method of Gardes (2018), denoted by TDR, which first estimates the unknown index to reduce the dimension of the covariate, and then uses a kernel-based method to estimate the conditional extreme quantiles; and (iv) the local linear estimator of Zhu et al. (2012), denoted by SIMQ, which is developed for the single-index quantile regression model at central quantiles. The tuning parameters u_x and h in BG are chosen as the minimizers of the asymptotic MSE of $\hat{\gamma}(x)$. The bandwidth parameter in ICDF is chosen using the cross-validation method proposed in Daouia et al. (2011). The parameter and kernel function of TDR are chosen in the same way as in Gardes (2018). The TDR method is for general multiple-index models and we apply this method with $p = 1$ when estimating the single index. The bandwidth h in SIMQ is chosen to be the same as in SIMEXQ.

Estimation of extreme conditional quantiles. We first compare the performance of the five methods when estimating the extreme conditional

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0051

quantiles $Q_\tau(Y|\mathbf{x})$ at $\tau = 0.99$, 0.995, and 0.999. For each simulation, we calculate the integrated squared error (ISE), defined as

$$\text{ISE} = \frac{1}{L}\sum_{l=1}^{L}\left\{\frac{\hat{Q}_\tau(Y|\mathbf{x}_l^*)}{Q_\tau(Y|\mathbf{x}_l^*)} - 1\right\}^2, \tag{5.1}$$

where $\mathbf{x}_1^*, \ldots, \mathbf{x}_L^*$ are evaluation points of the covariates, and we define the mean integrated squared error (MISE) as the average ISE across 500 simulations. In our simulation, we set $L = 50$. We choose fixed evaluation points $x_l^* = l/(1+L)$, for $l = 1, 2, ..., L$, in Case 1, and let $\mathbf{x}_l^*$ be random replicates of $\mathbf{X}$ in Case 2 with $p = 4$ and in Case 3. Table 1 summarizes the MISE for different estimators of the extreme conditional quantiles at $\tau = 0.99, 0.995$, and 0.999. The values in parentheses are the standard errors of the MISE.

In general, ICDF gives the least accurate estimators at high quantiles, while the proposed SIMEXQ method performs best in most cases. The larger MISE of ICDF is mainly due to the overestimation of the conditional quantiles. Case 1 can be regarded as a special case of the single-index model with $\beta_0 = 1$, the TDR methods, for which SIMQ and SIMEXQ do not involve an index estimation error. In Case 1, the BG method performs reasonably well and better than TDR, but its performance deteriorates quickly when the number of covariates increases. In all the scenarios considered, SIMEXQ is more efficient than SIMQ, and the advantage of SIMEXQ is more visibile at higher quantile levels. Compared to SIMEXQ, TDR per-

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0051

forms competitively when estimating the single index, but is less stable and gives a larger bias when estimating the extreme conditional quantiles.

Estimation of extreme value index. Because the estimation of the EVI is very important in extremal analysis, we also compare the performance of BG, ICDF, TDR, and the proposed SIMEXQ methods for estimating $\gamma(\mathbf{x})$. For each method, we calculate the MISE as the mean of the ISE across 500 simulations, where

$$\text{ISE} = \frac{1}{L}\sum_{l=1}^{L}\left\{\frac{\hat{\gamma}(\mathbf{x}_l^*)}{\gamma(\mathbf{x}_l^*)} - 1\right\}^2,$$

where $\mathbf{x}_1^*, \ldots, \mathbf{x}_L^*$ are set in (5.1).

Table 2: The mean integrated squared error (standard errors) of different estimators of $\gamma(\mathbf{x})$.

Case	BG	ICDF	TDR	SIMEXQ
Case 1, $p = 1$	0.02 (0.10)	0.02 (0.02)	0.03 (0.06)	0.01 (0.03)
Case 2, $p = 4$	0.25 (0.09)	0.17 (0.07)	0.34 (0.12)	0.11 (0.07)
Case 3, $p = 4$	0.98 (0.09)	0.52 (0.05)	0.24 (0.04)	0.30 (0.09)

BG: the estimator proposed of Beirlant and Goegebeur (2004); ICDF: the inverse CDF estimator; TDR: the tail dimension-reduction estimator; SIMEXQ: the proposed extreme quantile estimator.

Table 2 summarizes the MISE of different estimators of $\gamma(\mathbf{x})$ in Cases

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0051

1-3. The four methods perform similarly in Case 1. However, in Cases 2 and 3, BG and ICDF are clearly worse than SIMEXQ, with BG being the worst. The TDR method suffers from its complex estimation procedure, and leads to a more unstable estimation than that of SIMEXQ in Case 2. In Case 3, the quantile function depends on two indices, except when $\tau \to 1$, and the TDR method is based on estimating both indices, while SIMEXQ estimates only the single index. The more accurate index estimation in the TDR method leads to smaller MISEs in the EVI estimation in Case 3.

To better understand the performance of the different methods, we plot in Figures 1-3 the true and estimated conditional quantiles and the corresponding EVI estimators for BG, TDR, and SIMEXQ at $\tau = 0.995$ from one typical example in each case. For Case 1, the conditional quantile of Y is a *sine* function of x, and the true quantile curve has two peaks. We can see in Figure 1 that the proposed SIMEXQ method performs best, especially around the two sides of the conditional quantile curve. The BG method captures the two-peak structure, but overestimates them; hence, its MISE is large. The overestimation also occurs in BG's EVI estimation. For Case 2, the data are generated from a single-index model, so the x-axis is the single index $z = \mathbf{x}^T\boldsymbol{\beta}_0$. The conditional quantile curve is smooth, but not symmetric. The BG estimators are conditioned on $\mathbf{x}$, whereas the TDR

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0051

and SIMEXQ estimators are conditioned on their own index estimators $\hat{z}$. That is why their conditional quantile estimation curves are not smooth against the real index z. The EVI in Case 2 is a constant, so we present a box plot of $\hat{\gamma}(\hat{z})$ in Figure 2, which shows clearly that BG overestimates the EVI and has outliers, while TDR has the biggest range. For Case 3, the data are generated from tail single-index models, so the x-axis is the single index $z = \mathbf{x}^T\boldsymbol{\beta}_0$. In Figure 3, we can see that BG also overestimates the extreme conditional quantile, while TDR has more underestimation.

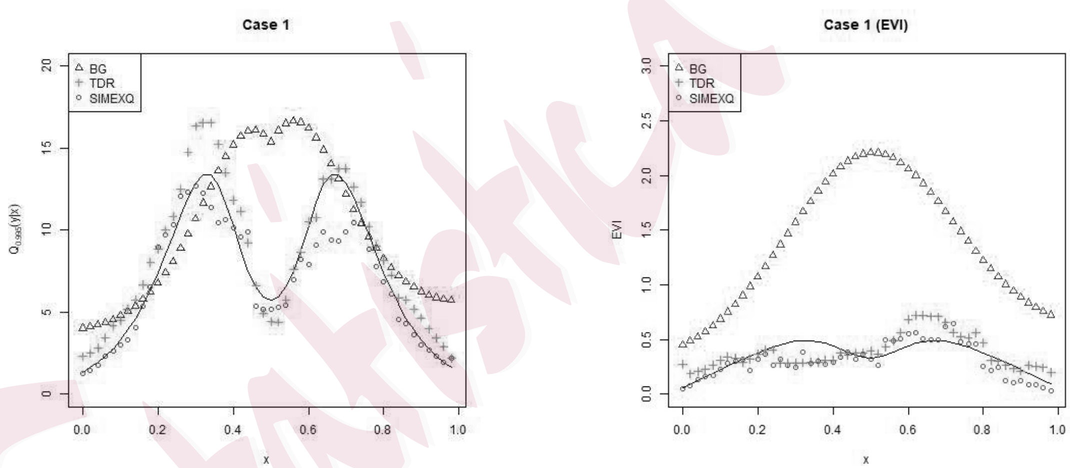

Figure 1: The truth (solid) and the estimations from BG (triangle), TDR (cross), and SIMEXQ (circle) for the conditional quantiles at $\tau^* = 0.995$ (left), and the EVI $\gamma(\mathbf{x})$ (right) for one example in Case 1.

Performance in high dimensions. We also investigate the performance

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0051

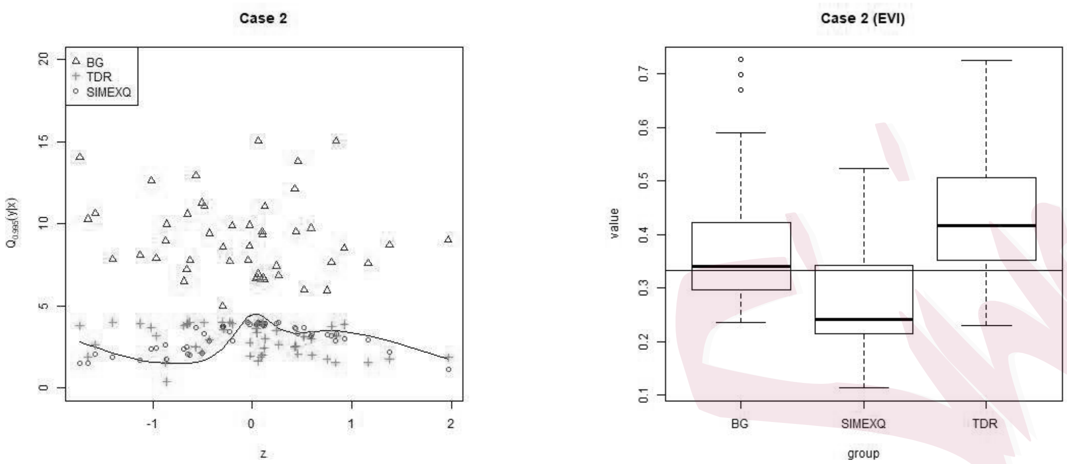

Figure 2: The truth (solid) and the estimations from BG (triangle), TDR (cross), and SIMEXQ (circle) for the conditional quantiles at $\tau^* = 0.995$ (left), and the EVI $\gamma(\mathbf{x})$ (right) for one example in Case 2 with $p = 4$.

of our proposed method when p is relatively large. Table 3 reports the MISE when $p = 50$ and 100, together with the previously considered $p = 4$, for different estimators in Case 2 with $\varepsilon \sim t(3)$. The results show that as p increases, the MISEs of TDR, SIMQ, and SIMEXQ increase much more slowly than those of ICDF and BG, manifesting the advantage of the dimension-reduction procedure in the former methods. The SIMEXQ performs similarly to TDR for $p = 4$, but the former is consistently more efficient for $p = 50$ and 100. In addition, SIMEXQ performs better than SIMQ across all of the scenarios and quantile levels considered.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0051

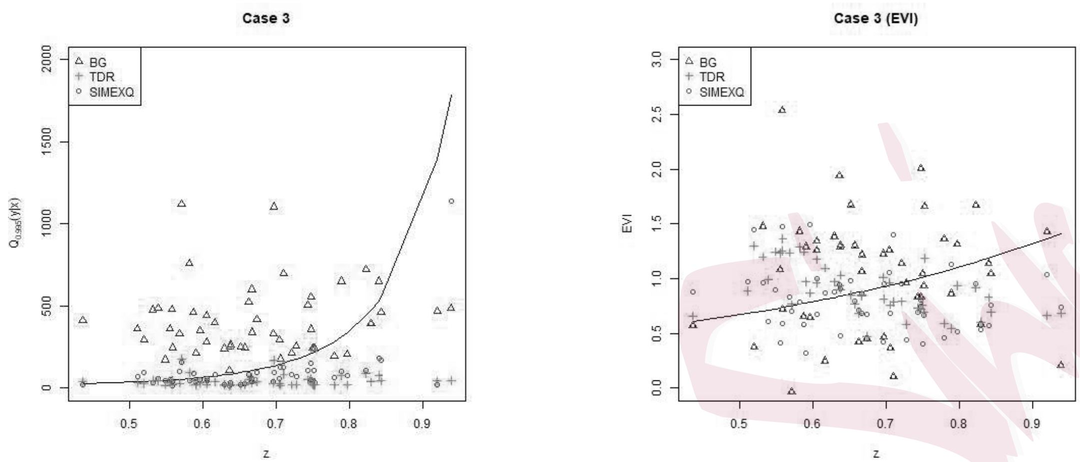

Figure 3: The truth (solid) and the estimations from BG (triangle), TDR (cross), and SIMEXQ (circle) for the conditional quantiles at $\tau^* = 0.995$ (left) and the EVI $\gamma(\mathbf{x})$ (right) for one example in Case 3.

5.2 Mortality data analysis

For centuries, the impact of weather and air pollution on people has been a public health concern. In this section, we analyse a subset of the National Morbidity, Mortality, and Air Pollution Study (NMMAPS) data to study the influence of weather and air pollution on the high quantile of mortality. The NMMAPS database consists of daily data on mortality, weather, and air pollution (e.g., pm10) for 109 US cities for the period 1987-2000. We use data from the city of Los Angeles (LA). The data set consists of daily mortality counts (all causes, CVD, respiratory), weather (temperature, dew

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0051

point temperature, relative humidity), and pollution factors (O_3, NO_2, SO_2, CO). We are interested in how the mortality count $\tilde{Y}$ in Los Angeles is affected by the following six variables: temperature, relative humidity, O_3, NO_2, SO_2, and CO, denoted by $X_1, X_2, ..., X_6$, respectively. After deleting observations with missing values, we have 4017 observations; that is, $n = 4017$. We scale all covariates to have a zero sample mean and a unit sample variance. Peng and Dominici (2006) also analyze mortality data from the NMMAPS database by fitting a Poisson regression to assess the effect of pollution on the mean of mortality. In contrast, we estimate the extreme high quantiles of mortality and examine how they depend on air pollution and weather.

Because the mortality $\tilde{Y}$ is count data, which are discrete, we perform the *jittering* process of Machado and Silva (2005). Specifically, add an independent random variable U from standard uniform distribution on $\tilde{Y}$; that is, $Y = \tilde{Y} + U$. Then, we consider the single-index model $Q_\tau(Y|\mathbf{X}) = G_\tau(\mathbf{X}^T\boldsymbol{\beta}_0)$, where $\mathbf{X} = (X_1, X_2, ..., X_6)^T$. After estimating $Q_\tau(Y|\mathbf{X})$, denoted by $\hat{Q}_\tau(Y|\mathbf{X})$, we obtain the estimation of the conditional quantile of $\tilde{Y}$ using $\hat{Q}_\tau(\tilde{Y}|\mathbf{X}) = \lceil \hat{Q}_\tau(Y|\mathbf{X}) - 1 \rceil$.

To reduce the variability of the estimates, we repeat the *jittering* processes 20 times, and take the average as our final estimator. That is, for the

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0051

lth time, we estimate the extreme conditional quantiles based on the typo sample $\{(Y_i^{(l)}, \mathbf{X}_i) : i = 1, 2, ..., n\}$, where $Y_i^{(l)} = \tilde{Y}_i + U_i^{(l)}$, $U_i^{(l)}$ are drawn independently from $U[0, 1]$, and $l = 1, \ldots, 20$. As suggested in Section 4.2, we take $\tau_0 = 1 - 0.2n^{-1/5} = 0.96$. We estimate the index parameter as $\hat{\boldsymbol{\beta}}_{\tau_0} = (-0.61, -0.29, -0.23, -0.13, 0.11, 0.66)^T$. Because all the covariates are scaled, the absolute values of the estimators imply that the pollutant CO (X_6) is likely to have the largest impact on the mortality variable (Y), and hence on the mortality ($\tilde{Y}$), followed by the temperature variable (X_1).

To better understand the performance of our proposed SIMEXQ method, we compare it with that of SIMQ. We set $\tau_0 = 0.96$ for both methods, and choose $k = \lceil 4.5n^{1/3} \rceil$ and $\eta = 0.1$ when estimating the EVI. Figure 4 plots the estimation of extreme conditional quantiles at $\tau^* = 0.995$ and 0.999 against $\hat{z} = \mathbf{x}^T \hat{\boldsymbol{\beta}}_{0.96}$. The plots shows that the SIMEXQ method gives a much smoother estimation than that of SIMQ. When the quantile level increases, SIMQ cannot capture the extreme behavior well, owing to the data sparsity. In addition, SIMQ also has a quantile crossing issue.

To compare the performance of the methods for predicting the extreme conditional quantiles of mortality counts, we carry out a cross-validation study. We randomly divide the data set into a training set (20% of the data set, 827 observations) and a testing set (the remaining 80% of the data set,

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0051

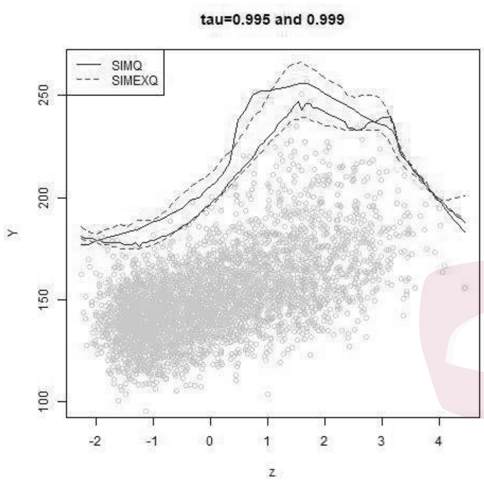

Figure 4: Estimation of the conditional quantile of mortality counts at $\tau^* = 0.995$ and 0.999 against $\hat{z} = \mathbf{x}^T \hat{\boldsymbol{\beta}}_{0.96}$ by SIMEXQ (dashed) and SIMQ (lined).

3190 observations). We apply BG, ICDF, TDR, SIMQ, and SIMEXQ to analyze the training set, and predict the extreme quantiles of the mortality counts conditional on the covariates in the testing set. Let $\hat{Q}_\tau(\tilde{Y}|\mathbf{X}_i)$ and $Q_\tau(\tilde{Y}|\mathbf{X}_i)$, with $i = 1, 2, ..., m = 3190$, denote the estimated and the true conditional quantiles of the mortality counts for subject i in the testing set, respectively. Conditional on $\mathbf{X}_i$, $I\{\tilde{Y}_i < Q_\tau(\tilde{Y}|\mathbf{X}_i)\}$ has mean τ and variance $\tau(1-\tau)$. We consider the following prediction error (PE) measurement, PE $= \{m\tau(1-\tau)\}^{-1/2} \sum_{i=1}^m [I\{\tilde{Y}_i < \hat{Q}_\tau(\tilde{Y}|\mathbf{X}_i)\} - \tau]$. We repeat the cross-validation 500 times. Table 4 summarizes the mean absolute PE of different methods at $\tau = 0.99, 0.995$, and 0.999. The values in

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0051

the parentheses are the corresponding standard errors. The results suggest that SIMEXQ has the highest prediction accuracy for extreme conditional quantile estimation, even for $\tau = 0.99$.

Table 4: Mean absolute prediction error (standard errors) of different methods at $\tau = 0.99, 0.995$, and 0.999 for predicting the extremal conditional quantiles of mortality counts.

Method	$\tau = 0.99$	$\tau = 0.995$	$\tau = 0.999$
BG	3.79 (0.10)	5.02 (0.21)	9.73 (0.43)
ICDF	6.67 (0.07)	13.53 (0.12)	25.34 (0.18)
TDR	4.21 (0.08)	5.72 (0.09)	10.26 (0.14)
SIMQ	3.58 (0.12)	6.04 (0.23)	12.16 (0.31)
SIMEXQ	3.32 (0.03)	5.22 (0.07)	8.67 (0.13)

BG: the estimator proposed of Beirlant and Goegebeur (2004); ICDF: the inverse CDF estimator; TDR: the tail dimension-reduction estimator; SIMQ: the single-index model estimator of Zhu et al. (2012) for central quantiles; SIMEXQ: the proposed extreme quantile estimator.

6. Discussion

We have proposed a the new tail single-index model to estimate the extreme quantile conditional on multi-dimensional covariates. We propose an

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0051

efficient three-step procedure for estimating extreme conditional quantiles. We establish the asymptotic properties of our new estimators for the extreme value index and extreme conditional quantiles. The results of our numerical and empirical studies imply that the proposed SIMEXQ method performs more effectively and is more stable than competing methods.

Although we assume heavy-tailed distributions, the proposed method can be extended to general cases with $\gamma(\mathbf{x}) \in \mathbb{R}$ by considering other types of estimators for the extreme value index, such as the moment estimators of De Haan and Ferreira (2006) and Li and Wang (2019). For single-index models with high dimensional covariates, variable selection is important, and research in this direction under the extreme quantile setting deserves further investigation.

Supplementary Material

The online Supplementary Material contains some remarks, additional simulation results, and all technical details.

Acknowledgments

This work was partially supported by the China Scholarship Council grant 201906100118, the National Natural Science Foundations of China grants 11571081, 11971115, 11690012, and 71531006, the Key Laboratory for Applied Statistics of MOE, North Mormal University, the IR/D program

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0051

from the U.S. National Science Foundation (NSF), and grant DMS-1712760. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors, and do not necessarily reflect the views of the NSF.

References

Angrist, J., Chernozhukov, V. and Fernández-Val, I. (2006). Quantile regression under misspecification with an application to the U.S. wage structure. *Econometrica* **74**, 539-563.

Beirlant, J. and Goegebeur, Y. (2003). Regression with response distributions of Pareto-type. *Comput. Statist. Data Anal.* **42**, 595-619.

Beirlant, J. and Goegebeur, Y. (2004). Local polynomial maximum likelihood estimation for Pareto-type distributions. *J. Multiariate Statist.* **89**, 97-118.

Chavez-Demoulin, V. and Davison, A. (2005). Generalized additive modelling of sample extremes. *J. Roy. Statist. Soc. Ser. C)* **54**, 207-222.

Chernozhukov, V. (2005). Extremal quantile regression. *Ann. Statist.* **33**, 806-839.

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

Daouia, A., Gardes, L. and Girard, S. (2013). On kernel smoothing for extremal quantile regression. *Bernoulli* **19**, 2557-2589.

Davison, A. and Smith, R. (1990). Models for exceedances over high thresholds. *J. Roy. Statist.*

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0051

Soc. Ser. B **52**, 393-425.

De Haan, L. and Ferreira, A. (2006). *Extreme value theory: an introduction.* Springer Science & Business Media.

Embrechts, P., Klüppelberg, C. and Mikosch, T. (2013). *Modelling extremal events: for insurance and finance.* Springer Science & Business Media.

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

Gannoun, A., Girard, S., Guinot, C. and Saracco, J. (2004). Sliced inverse regression in reference curves estimation. *Comput. Statist. Data Anal.* **46**, 103-122.

Gardes, L. (2018). Tail dimension reduction for extreme quantile estimation. *Extremes* **21**, 57-95.

Gardes, L. and Girard, S. (2010). Conditional extremes from heavy-tailed distributions: An application to the estimation of extreme rainfall return levels. *Extremes* **13**, 177-204.

Gardes, L., Girard, S. and Lekina, A. (2010). Functional nonparametric estimation of conditional extreme quantiles. *J. Multiariate Statist.* **101**, 419-433.

Gardes, L., Guillou, A. and Schorgen, A. (2012). Estimating the conditional tail index by integrating a kernel conditional quantile estimator. *J. Statist. Plann. Inference* **142**, 1586-1598.

Hall, P. and Li, K. (1993). On almost linearity of low dimensional projections from high

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0051

dimensional data. *Ann. Statist.* **93**, 867-889.

Hardle, W., Hall, P. and Ichimura, H. (1993). Optimal smoothing in single-index models. *Ann. Statist.* **21**, 157-178.

Kato, K. (2009). Asymptotics for argmin processes: Convexity arguments. *J. Multiariate Statist.* **100**, 1816-1829.

Kong, E. and Xia, Y. (2012). A single-index quantile regression model and its estimation. *Econom. Theory* **28**, 730-768.

Koenker, R., Chesher, A. and Jackson, M. (2005). Quantile Regression. *Cambridge University Press.*

Li, K. (1991). Sliced inverse regression for dimension reduction. *J. Amer. Statist. Assoc.* **86**, 316-327.

Li, D. and Wang, H. (2019). Extreme quantile estimation for autoregressive models. *J. Business Econom. Statist.* **37**, 661-670.

Li, B. (2018). Sufficient dimension reduction: Methods and applications with R. *CRC Press.*

Machado, J. and Silva, J. (2005). Quantiles for counts. *J. Amer. Statist. Assoc.* **100**, 1226-1237.

Peng, R., Dominici, F. and Louis, T. (2006). Model choice in time series studies of air pollution and mortality *J. Roy. Statist. Soc. Ser. A* **169**, 179-203.

Powell, J., Stock, J. and Stoker, T. (1989). Semiparametric estimation of index coefficients. *Econometrica* **57**, 1403-1430.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0051

Santos, P. A., Alves, M. I. and Gomes, M. I. (2006). Peaks over random threshold methodology for tail index and high quantile estimation. *REVSTAT* **4**, 227-247.

Shorack, G. R. (1979). The weighted empirical process of row independent random variables with arbitrary distribution functions. *Statist. Neerlandica* **33**, 169-189.

Wang, H. and Tsai, C. (2009). Tail index regression. *J. Amer. Statist. Assoc.* **104**, 1233-1240.

Wang, H. and Li, D. (2013). Estimation of extreme conditional quantiles through power transformation. *J. Amer. Statist. Assoc.* **503**, 1062-1074.

Wang, H., Li, D. and He, X. (2012). Estimation of high conditional quantiles for heavy-tailed distributions. *J. Amer. Statist. Assoc.* **107**, 1453-1464.

Weissman, I. (1978). Estimation of parameters and large quantiles based on the k largest observations. *J. Amer. Statist. Assoc.* **73**, 812-815.

Wu, T., Yu, K. and Yu, Y. (2010). Single-index quantile regression. *J. Multivariate Anal.* **101**, 1607-1621.

Zhong, W., Zhu, L., Li, R. and Cui, H. (2016). Regularized quantile regression and robust feature screening for single index models. *Statist. Sinica* **26**, 69-95.

Zhu, L., Huang, M. and Li, R. (2012). Semiparametric quantile regression with high-dimensional covariates. *Statist. Sinica* **22**, 1379-1424.

Department of Statistics, Fudan University, Shanghai, China

E-mail: (18110690007@fudan.edu.cn)

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0051

Department of Statistics, George Washington University, Washington DC, USA

E-mail: (judywang@email.gwu.edu)

Department of Statistics, Fudan University, Shanghai, China

E-mail: (deyuanli@fudan.edu.cn)

Statistica Sinica

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0051

Table 1: The mean integrated squared error (standard errors) for different estimators of the extreme conditional quantiles at $\tau = 0.99, 0.995$, and 0.999.

Case	Method	$\tau = 0.99$	$\tau = 0.995$	$\tau = 0.999$
Case 1, $p = 1$	BG	0.64 (0.12)	0.72 (0.37)	0.89 (0.23)
	ICDF	1.51 (0.09)	2.94 (0.21)	12.23 (0.27)
	TDR	0.87 (0.08)	1.35 (0.18)	2.67 (0.29)
	SIMQ	0.16 (0.15)	0.24 (0.22)	0.37 (0.27)
	SIMEXQ	0.15 (0.01)	0.18 (0.04)	0.27 (0.07)
Case 2, $p = 4$	BG	12.42 (0.04)	8.37 (0.07)	5.04 (0.11)
	ICDF	6.22 (0.05)	11.23 (0.08)	57.38 (0.12)
	TDR	0.18 (0.04)	0.67 (0.07)	0.89 (0.09)
	SIMQ	0.06 (0.09)	0.13 (0.12)	0.24 (0.15)
	SIMEXQ	0.04 (0.02)	0.05 (0.03)	0.07 (0.07)
Case 3, $p = 4$	BG	18.76 (0.14)	26.53 (0.27)	37.27 (0.91)
	ICDF	12.43 (0.12)	31.26 (0.24)	52.31 (0.67)
	TDR	0.64 (0.08)	0.86 (0.35)	1.51 (0.51)
	SIMQ	0.41 (0.11)	0.98 (0.42)	1.67 (0.87)
	SIMEXQ	0.16 (0.09)	0.41 (0.13)	0.99 (0.24)

BG: the estimator proposed of Beirlant and Goegebeur (2004); ICDF: the inverse CDF estimator; TDR: the tail-dimension reduction estimator; SIMQ: the single-index model estimator of Zhu et al. (2012) for central quantiles; SIMEXQ: the proposed extreme quantile estimator.

Statistica Sinica

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0051

Table 3: The mean integrated squared error (standard errors) for different estimators of the conditional quantiles with $\tau = 0.99, 0.995$, and 0.999 in Case 2.

p	Method	$\tau = 0.99$	$\tau = 0.995$	$\tau = 0.999$
4	BG	12.42 (0.04)	8.37 (0.07)	5.04 (0.11)
	ICDF	6.22 (0.05)	11.23 (0.08)	57.38 (0.12)
	TDR	0.05 (0.02)	0.06 (0.03)	0.07 (0.07)
	SIMQ	0.13 (0.09)	0.13 (0.12)	0.24 (0.15)
	SIMEXQ	0.04 (0.02)	0.05 (0.03)	0.07 (0.07)
50	BG	43.79 (0.21)	80.12 (0.25)	123.30 (0.38)
	ICDF	12.57 (0.28)	36.21 (0.36)	64.78 (0.42)
	TDR	0.32 (0.07)	0.41 (0.13)	0.57 (0.17)
	SIMQ	0.35 (0.12)	0.59 (0.14)	0.98 (0.19)
	SIMEXQ	0.27 (0.08)	0.39 (0.12)	0.52 (0.15)
100	BG	52.72 (0.27)	81.34 (0.28)	133.20 (0.41)
	ICDF	11.84 (0.23)	42.67 (0.32)	69.83 (0.39)
	TDR	0.36 (0.09)	0.53 (0.17)	0.68 (0.21)
	SIMQ	0.52 (0.13)	0.65 (0.18)	1.02 (0.20)
	SIMEXQ	0.31 (0.08)	0.45 (0.13)	0.59 (0.16)

BG: the estimator proposed of Beirlant and Goegebeur (2004); ICDF: the inverse CDF estimator; TDR: the tail dimension-reduction estimator; SIMQ: the single-index model estimator of Zhu et al. (2012) for central quantiles; SIMEXQ: the proposed extreme quantile estimator.