Article

A Method for Confidence Intervals of High Quantiles

Mei Ling Huang [1,*] and **Xiang Raney-Yan** [2]

1 Department of Mathematics, Brock University, St. Catharines, ON L2S 3A1, Canada
2 Department of Mathematcs, Niaraga College, Welland, ON L3C 7L3, Canada; xyan@niagaracollege.ca
* Correspondence: mhuang@brocku.ca

Abstract: The high quantile estimation of heavy tailed distributions has many important applications. There are theoretical difficulties in studying heavy tailed distributions since they often have infinite moments. There are also bias issues with the existing methods of confidence intervals (CIs) of high quantiles. This paper proposes a new estimator for high quantiles based on the geometric mean. The new estimator has good asymptotic properties as well as it provides a computational algorithm for estimating confidence intervals of high quantiles. The new estimator avoids difficulties, improves efficiency and reduces bias. Comparisons of efficiencies and biases of the new estimator relative to existing estimators are studied. The theoretical are confirmed through Monte Carlo simulations. Finally, the applications on two real-world examples are provided.

Keywords: efficiency; extreme value distributions; generalized Pareto distribution; Hill estimator; mean square errors; order statistics; tail index; Weissman estimator

1. Introduction

Extreme value analysis (EVA) was first introduced by Leonard Tippett (Fisher and Tippett, 1928 [1]). Tippett was working on how to make cotton thread stronger, he realized that the strength of the weakest threads were the only factor that matters when it comes to deciding the strength of the cotton thread. Nowadays, extreme value analysis is widely used in almost all fields, from engineering, social science, economics, traffic predictions to insurance and so on. People are interested in extreme events in these fields such as, the shortest life span of a new engine, the maximum appreciation of the stock market, the longest driving time on a highway at rush hour, or the biggest medical claim to an insurance company. The distributions of these extreme events are usually unknown. In general, EVA involves the extrapolation of an unknown distribution and its high quantiles. Estimating high quantile based on observation is very important in EVA, since it gives the corresponding value x for a very small exceeding possibility p.

There are certain risks, ones that are not decided by us or can barely be predicted until right before they are about to happen. This can include things such as an earthquake, terrorist attacks, a virus breakout, and so forth. For these events, we will need risk management which is in place to minimize, monitor, and control the impact of unfortunate events, or to maximize the realization of opportunities. Estimating the confidence interval of high quantiles plays an important role in risk management. Since a high quantile is located at the tail area, it heavily depends on the behaviour of the tail distribution, or from the statistical point of view, it depends on the k largest order statistics. This leads to the challenges of the instability in the choice of k, and the bias issues. There are many research on the mathematical models and theoretical studies in the literature for estimating confidence intervals of high quantiles, we review them in Section 2.

This paper proposes a new method to estimate high quantile of a heavy-tailed distribution. The new method has interesting improvements compared with other existing methods. This paper makes three main contributions to methodology.

(1) This paper proposes a new estimation method based on a geometric mean with good asymptotic properties. It is consistent and stable relative to the existing methods. The

paper provides a computational algorithm which overcomes the mathematical difficulties and bias problems of the estimation of confidence intervals of high quantiles of a heavy tailed distribution.

(2) The Monte Carlo simulation studies on three heavy tailed distribution models: Fréchet (0.25), GPD (0.5) and GPD(2) (GPD: generalized Pareto distribution). The simulation results confirm that the proposed method is more efficient relative to the existing quantile estimators.

(3) This paper uses the proposed estimation method to predict extreme values in the flu in Canada, and gamma ray from solar flare examples. It is interesting to see that these data sets fit the GPD model very well. We apply the proposed method to estimate the confidence intervals of high quantiles. The numerical results show that the proposed method gives more efficient results compared with other existing methods.

In this paper, we review several existing high quantile estimators with their behavior in Section 2. We propose a new estimator for the confidence interval of high quantiles based on the geometric mean and explore its asymptotic properties in Section 3. To compare the new estimator with the existing estimators, Section 4 presents Monte Carlo simulation results and the improvement of the proposed quantile estimator relative to existing methods. In Section 5 we apply the proposed new method to construct confidence intervals of high quantiles on flu in Canada and gamma ray examples. Finally, conclusions and discussions are given in Section 6.

2. Existing Estimator for High Quantiles

Heavy-tailed distributions (de Haan and Ferreira, 2006 [2]) is important to extreme value events.

Definition 1. *A random variable X is said to have a heavy tail distribution if its distribution function F(x) satisfies*

$$1 - F(x) = L(x)x^{-1/\gamma}, x \in (-\infty, \infty), \text{ as } x \to \infty, \gamma > 0,$$

where $L(t)$ is a slowly varying function with $\lim\limits_{t \to \infty} \frac{L(tx)}{L(t)} = 1$, *for all $x > 0$. γ is the tail index.*

Notice that we can have $L(x) = (\ln(x))^b$, $b \in \mathbb{R}$ (de Hann and Ferreira, 2006, p. 362 [2]). Since $L(t)$ behaves approximately as a constant c, for simplicity, we assume that a heavy tailed distribution satisfies

$$1 - F(x, \gamma) \to cx^{-1/\gamma}, x \in (-\infty, \infty), \text{ as } x \to \infty, c > 0, \gamma > 0. \tag{1}$$

Since the heavy tailed distributions decay slower than the exponential distributions and have longer tails. A tail function is defined as

Definition 2. *A tail function $U(t)$ of any distribution function F(x) is defined as*

$$U(t) = \left(\frac{1}{1 - F}\right)^{\leftarrow}, \quad \textit{where "} \leftarrow \textit{" denotes the inverse function.}$$

For the heavy tailed distribution in (1), we can rewrite the tail function as

$$U(t) = \left(\frac{1}{ct^{-1/\gamma}}\right)^{\leftarrow} = c^{\gamma}t^{\gamma} = Ct^{\gamma}, \text{ as } t \to \infty, \text{ where } c^{\gamma} = (L(t))^{\gamma}, \text{ let } C = c^{\gamma}. \tag{2}$$

Definition 3. *The quantile function $Q(1 - p, \gamma)$ of a heavy tailed distribution $F(x, \gamma)$ in (1) for a given probability $1 - p$ is defined by*

$$x_{1-p} = Q(1 - p, \gamma) = \inf\{x : F(x, \gamma) \geq 1 - p\}, x \in (-\infty, \infty), 0 < p < 1$$

where $Q(1-p,\gamma)$ is the generalized inverse function of F, we call $Q(1-p,\gamma)$ the $(1-p)th$ quantile function of $F(x)$.

Value at Risk (*VaR*) is widely used in risk management. When p is very small, x_{1-p} becomes a high quantile as the pth value at risk, we define

$$VaR_{p,\gamma} = x_{1-p} = Q(1-p,\gamma), 0 < p < 1, p \text{ is very small.} \tag{3}$$

Also we can use the tail function in (2) to write $VaR_{p,\gamma}$ as

$$VaR_{p,\gamma} = U\left(\frac{1}{p},\gamma\right) = Ct^{\gamma}, t = \frac{1}{p}, \ p = p_n \to 0, np_n \to 0, \text{ as } n \to \infty.$$

The heavy-tailed models have a compulsory infinite right endpoint. In the case of negative observations in the model, the sample size should be exclusively the number of positive observations, n^+, although a deterministic shift in the data is preferred by some authors, to work only with positive values. In this paper, we use the real line $(0,\infty)$.

To estimate $VaR_{p,\gamma}$, let $X_{1:n} \leq X_{2:n} \leq ... \leq X_{n:n}$ be the order statistics from a random sample $X_1, X_2, ...X_n$. We review the four high quantile estimation methods in the literature.

2.1. Quantile Function-Tail Index Method

For estimating high quantiles, we use the ln function, and estimate the tail index first

$$\ln Q_{\widehat{\gamma}}^{(p)} = \ln VaR_{p,\widehat{\gamma}} = \ln x_{1-p,\widehat{\gamma}}, \ 0 < p < 1, \ p \text{ is very small.} \tag{4}$$

To estimate high quantile function, we estimate the tail index first (Dekkers and de Haan 1989 [3]). Hill (1975) [4] estimator is a well known consistent estimator for tail index γ.

Definition 4. *Consider the order statistics $X_{n-k:n}$, and k as an intermediate sequence of integers, Hill estimator is defined as*

$$\widehat{\gamma}_H = H(k) = \frac{1}{k}\sum_{i=1}^{k} U_i, \ U_i = i\left(\ln \frac{X_{n-i+1:n}}{X_{n-i:n}}\right), 1 \leq i \leq k. \tag{5}$$

where $k = k_n \to \infty, k \in [1,n), k = o(n)$ as $n \to \infty$.

The Hill estimator $\widehat{\gamma}_H = H(k)$ in (5) used largest k order statistics of a random sample. Substitute $\widehat{\gamma}_H = H(k)$ defined in (5) into (4), then we obtain $\ln (1-p)$th high quantile as

$$\ln q_{H,p}(k) = \ln Q_H^{(p)}(k) = \ln x_{1-p,H}(k), 0 < p < 1, p \text{ is very small.} \tag{6}$$

This estimator depends on k, small values of k provide high volatility whereas large values of k induce considerable bias. Hence, semi-parametric extensions may be considered for increasing the degree of freedom in the trade-off between variance and bias. Note that the tail index γ is a parameter of a given distribution, and a quantile of a distribution is a function of γ.

2.2. Weissman Method

Weissman (1978) [5] proposed the following semiparametric estimator of a high quantile

$$Q_{\widehat{\gamma}}^{(p)}(k) = VaR_{p,\widehat{\gamma}} = x_{1-p,\widehat{\gamma}} = X_{n-k:n}\left(\frac{k}{np}\right)^{\widehat{\gamma}}, 0 < p < 1, 1 \leq k \leq n-1, \quad \text{and}$$

$$\ln Q_{\widehat{\gamma}}^{(p)}(k) = \ln X_{n-k:n} + \widehat{\gamma}\ln\left(\frac{k}{np}\right).$$

We substitute $\widehat{\gamma}_H = H(k)$ in (5) into the function above, then we have,

$$\ln \widehat{Q}_H^{(p)}(k) = \ln X_{n-k:n} + H(k) \ln \left(\frac{k}{np} \right), \quad 1 \le k \le n-1. \tag{7}$$

Without any prior indication on *k*, the Weissman estimator shows a large volatility as it depends on the fraction sample *k*. Although the minimization of the bias and MSE can be considered as a criterion to select *k*, it is impractical as they are unknown. Other methods for the selection of sample fraction *k* can be found in Beirlant et al. (1996) [6]; Dreea and Kaufmann (1998) [7]; Guillou and Hall (2001) [8]; Gomes and Oliveira (2001) [9].

The optimal *k* value through the tail index Hill estimator *H*, k_0, is given by formula (15) in Section 2.4 Optimal *k* Values.

2.3. Reduced-Bias Method

Hall and Welsh (1985) [10] proposed a second-order expansion on the tail function *U* in (2)

$$U(t) = Ct^{\gamma} \left(1 + \frac{A(t)}{\rho} + o(t^{\rho}) \right), \quad A(t) = \gamma \beta t^{\rho}, \ as \ t \to \infty, \tag{8}$$

with $C, \gamma > 0, \rho < 0$, and $\beta \neq 0$. Where β is the scale second-order parameter and ρ is the shape second-order parameter.

To further reduce the bias of quantile estimators which requires us to observe the behavior of the estimation of the second-order parameters β and ρ. Second-order reduced-bias was discussed by Peng (1998) [11], Beirlant, Dierckx, Goegebeur and Mattys (1999) [12], Freueverger and Hall (1999) [13], Gomes, Martins and Neves (2000) [14], Caeiro and Gomes (2002) [15], Gomes, Figueiredo and Mendonea (2004) [16], among others. Comes and Pestana (2007) [17] considered the estimators $\left(\widehat{\rho}_{\tau}(k), \widehat{\beta}_{\widehat{\rho}}(k) \right)$ for the second-order parameters (ρ, β).

Careiro et al. (2005, p. 122) [18] advises the the use of turning parameter τ in the estimation of ρ. It provides higher stability as functions of *k*, the number of the top order statistics used, for a wide range of large *k* value, by means of any stability criterion.

Definition 5. *Caeiro et al. (2005) [18] defined the bias-corrected Hill estimator*

$$\overline{H}(k) \equiv \overline{H}_{\widehat{\beta},\widehat{\rho}}(k) = H(k) \left(1 - \frac{\widehat{\beta}}{1-\widehat{\rho}} \left(\frac{n}{k} \right)^{\widehat{\rho}} \right), \tag{9}$$

where $H(k)$ is defined in (5). For a tuning real parameter $\tau \in \mathbb{R}$,

$$\widehat{\rho}_{\tau}(k) \equiv \widehat{\rho}_n^{(\tau)}(k) = \min \left(0, \frac{3 \left(T_n^{(\tau)}(k) - 1 \right)}{T_n^{(\tau)}(k) - 3} \right) \tag{10}$$

$$T_n^{(\tau)}(k) = \begin{cases} \dfrac{\left(M_n^{(1)}(k) \right)^{\tau} - \left(\frac{M_n^{(2)}(k)}{2} \right)^{\frac{\tau}{2}}}{\left(\frac{M_n^{(2)}(k)}{2} \right)^{\frac{\tau}{2}} - \left(\frac{M_n^{(3)}(k)}{6} \right)^{\frac{\tau}{3}}} & \text{if } \tau \neq 0; \\[3em] \dfrac{\ln \left(M_n^{(1)}(k) \right) - \frac{1}{2} \ln \left(\frac{M_n^{(2)}(k)}{2} \right)}{\frac{1}{2} \ln \left(\frac{M_n^{(2)}(k)}{2} \right) - \frac{1}{3} \ln \left(\frac{M_n^{(3)}(k)}{6} \right)} & \text{if } \tau = 0, \end{cases}$$

$$M_n^{(j)}(k) = \frac{1}{k} \sum_{i=1}^{k} (\log X_{n-i+1:n} - \log X_{n-k:n})^j, \ j = 1,2,3.$$

$$\widehat{\beta}_{\widehat{\rho}}(k) = \left(\frac{k}{n} \right)^{\widehat{\rho}} \frac{d_{\widehat{\rho}}(k) D_0(k) - D_{\widehat{\rho}}(k)}{d_{\widehat{\rho}}(k) D_{\widehat{\rho}}(k) - D_{\widehat{2\rho}}(k)}, \tag{11}$$

where for any $\theta \leq 0$, $d_\theta(k) = \frac{1}{k} \sum_{i=1}^{k} \left(\frac{i}{k}\right)^{-\theta}$ and $D_\theta(k) = \frac{1}{k} \sum_{i=1}^{k} \left(\frac{i}{k}\right)^{-\theta} U_i$,

with U_i as defined in (5) that $1 \leq i \leq k$, and (10) achieves consistency if $\sqrt{k}A(n/k) \to \infty$ as $n \to \infty$ and $\widehat{\rho} - \rho = o_\rho(1/\ln n)$.

The corresponding ln-quantile estimator with the tail index estimator $\overline{H}$ in (9) is

$$\ln \widehat{Q}_{\overline{H}}^{(p)}(k) = \ln X_{n-k+1:n} + \overline{H}(k) \ln\left(\frac{k}{np}\right), \quad 1 \leq k \leq n - 1. \tag{12}$$

A similar estimator to the estimator in (12) is considered in Lekina et al. (2014) [19] and Lekina (2010) [20].

Gomes and Pestana (2007) [17] considered the ln-Var estimator

$$\ln \overline{Q}_{\widehat{\gamma}}^{(p)}(k) = \ln X_{n-k+1:n} + \widehat{\gamma}\left(\ln\left(\frac{k}{np}\right) + C_p\left(k; \widehat{\beta}, \widehat{\rho}\right)\right), \quad C_p(k; \widehat{\beta}, \widehat{\rho}) = \widehat{\beta}\left(\frac{n}{k}\right)^{\widehat{\rho}} \frac{\left(\frac{k}{np}\right)^{\widehat{\rho}} - 1}{\widehat{\rho}}. \tag{13}$$

Substitute the estimator $\overline{H}$ in (9) into (13), we have another estimator for high quantile as

$$\ln \overline{Q}_{\overline{H}}^{(p)}(k) = \ln X_{n-k+1:n} + \overline{H}(k)\left(\ln\left(\frac{k}{np}\right) + C_p\left(k; \widehat{\beta}, \widehat{\rho}\right)\right). \tag{14}$$

2.4. Optimal k Values

As discussed previously, we have problem that the estimation varies as the k varies, and it become very unreliable when k is large. Gomes and Pestana (2007) [17] suggested to use the numerically estimated optimal k values.

The optimal k for the tail index estimator through Hill estimator $H(k)$ in (5) is k_0,

$$k_0 = \left(\frac{(1-\rho)n^{-\rho}}{\beta\sqrt{-2\rho}}\right)^{\frac{2}{1-2\rho}}. \tag{15}$$

The optimal k for the semiparametric quantile estimator $\ln Q_H(k)$ in (7), is $k_0^{Q_H}$,

$$k_0^{Q_H} = \arg\min_k \left\{\ln^2\left(\frac{k}{np}\right)\left(\frac{1}{k} + \frac{\beta^2(n/k)^{2\rho}}{(1-\rho)^2}\right)\right\}. \tag{16}$$

The optimal k for the second-order reduced-bias quantile estimator $\ln Q_{\overline{H}}^{(p)}(k)$ in (12) and $\ln \overline{Q}_{\overline{H}}^{(p)}(k)$ in (14) should be larger than k_0, is k_{01},

$$k_{01} = \left(\frac{1.96(1-\rho)n^{-\rho}}{|\beta|}\right)^{\frac{2}{1-2\rho}}. \tag{17}$$

By using these optimal k values, all the quantile estimators provide better results. However, with an unknown distribution, and estimated second-order parameters, these numerically estimated k values are not always accurate. Since all the quantile estimators are so sensitive to the k value, in this paper, we propose a new quantile estimator which does not depends on k.

3. New Estimator for High Quantile
3.1. New Estimator

Our goal is to improve the quantile estimators in Section 2. There are bias issues and difficult in determining k with the existing estimating methods. In order to overcome these

problems, Huang (2011) [21] proposed a new quantile estimator which is the geometric mean of the reduced-bias quantile estimator in (14).

Definition 6.

$$\widehat{Q}_{New,\widehat{\gamma}}^{(p)} = \left[\prod_{k=1}^{n-1}\left(X_{n-k+1:n}\left(\frac{k}{np}\right)^{\widehat{\gamma}}\right)\right]^{\frac{1}{n-1}}, 0 < p < 1, 1 \le k \le n-1. \tag{18}$$

where $X_{n-k:n}$ is the $(k+1)$th top order statistic, $\widehat{\gamma}$ is any consistent estimator for γ, and Q stands for quantile function.

Based on (16), (20) can be written as

$$\ln \widehat{Q}_{New,\widehat{\gamma}}^{(p)} = \frac{1}{n-1}\sum_{k=1}^{n-1}\left[\ln X_{n-k+1:n} + \widehat{\gamma}\left(\ln\left(\frac{k}{np}\right) + \alpha C_p(k;\widehat{\beta},\widehat{\rho})\right)\right], \tag{19}$$

where $0 < p < 1$ and α is a constant that $\alpha \in \mathbb{R}$.

$\alpha C_p(k;\widehat{\beta},\widehat{\rho})$ is the adjustment term, where $C_p(k;\widehat{\beta},\widehat{\rho})$ is defined in (13) that reduces bias using the second-order parameters. α is a key value depends only on n to furthermore reduce the bias by observing the behavior of the second-order parameters. We will discuss the choice of α in Section 4.

Sections 3–5 will show that the new estimator $\ln \widehat{Q}_{New,\widehat{\gamma}}^{(p)}$ has good properties, and

1. The new quantile estimator $\ln \widehat{Q}_{New,\widehat{\gamma}}^{(p)}$ has the least bias, the smallest *MSE* and the highest efficiency.

2. The new quantile estimator $\ln \widehat{Q}_{New,\widehat{\gamma}}^{(p)}$ is consistent and does not depend on k as the existing quantile estimators does.

3. The confidence interval based on the new quantile estimator $\ln \widehat{Q}_{New,\widehat{\gamma}}^{(p)}$ is the most efficient compared to the existing methods, where it not only has the shortest length of the interval, but also has the highest probability coverage of the true value in most cases.

3.2. Asymptotic Properties of the New Estimator $\ln\widehat{Q}_{New,\overline{H}}^{(p)}$

Using the Hall-Welsh class of model in (8), we derive that the new estimator $\ln \widehat{Q}_{New,\widehat{\gamma}}^{(p)}$ in (19) has the asymptotic properties under following conditions, when $\widehat{\gamma} = \overline{H}$ in (5).

Condition 1 (C1). *For intermediate k, $k = k_n \to \infty$. $k \in [1, n)$, $k = o(n)$, as $n \to \infty$.*

Condition 2 (C2). $\ln(np_n) = o\left(\sqrt{k}\right)$, $\lim_{n\to\infty}\sqrt{k}A\left(\frac{n}{k}\right) = \lambda \in \mathbb{R}$, where A is in (8).

Theorem 1. *Under (C1) and (C2), if we use $\widehat{\gamma} = \overline{H}$ in (5), then $\ln\widehat{Q}_{New,\overline{H}}^{(p)}$ has a asymptotic normal distribution*

$$\left(\ln \widehat{Q}_{New,\overline{H}}^{(p)} - \ln VaR_p\right) \xrightarrow[n\to\infty]{d} Normal\left(0, \frac{\gamma^2}{(n-1)^2}\left[\sum_{k=1}^{n-1}\frac{[\ln\frac{k}{np}]^2}{k} + w\sum_{i<j}^{n-1}\sum^{n-1}\left[\frac{\ln\left(\frac{i}{np}\right)\ln\left(\frac{j}{np}\right)}{\sqrt{i \bullet j}}\right]\right]\right). \tag{20}$$

The asymptotic mean, variance and efficiency of $\ln \widehat{Q}_{New,\overline{H}}^{(p)}(k)$ in (19) relative to $\ln \overline{Q}_{\overline{H}}^{(p)}(k)$ in (14) are given by

$$E\left(\ln \widehat{Q}_{New,\overline{H}}^{(p)}\right) \underset{n\to\infty}{\approx} \ln VaR_p;$$

$$Var\left(\ln \widehat{Q}^{(p)}_{New,\overline{H}}\right) \underset{n\to\infty}{\leq} \frac{\gamma^2}{(n-1)^2}\left[\sum_{k=1}^{n-1}\frac{[\ln \frac{k}{np}]^2}{k} + w\sum_{i<j}^{n-1}\sum^{n-1}\left[\frac{\ln\left(\frac{i}{np}\right)\ln\left(\frac{j}{np}\right)}{\sqrt{i\bullet j}}\right]\right]; \tag{21}$$

$$EFF\left(\ln \widehat{Q}^{(p)}_{New,\overline{H}}\right)_{\ln \widehat{Q}^{(p)}_{\overline{H}}(k)} \underset{n\to\infty}{\geq} \frac{(n-1)^2\frac{\left(\ln\left(\frac{k}{np}\right)\right)^2}{k}}{\sum_{k=1}^{n-1}\frac{[\ln\frac{k}{np}]^2}{k} + w\sum_{i<j}^{n-1}\sum^{n-1}\left[\frac{\ln\left(\frac{i}{np}\right)\ln\left(\frac{j}{np}\right)}{\sqrt{i\bullet j}}\right]} > 1, \tag{22}$$

$$for\ k = 1, ..., n-1,$$

where w *is the weight,* $w = \max_{i\neq j}\rho^+_{ij}$, $0 \leq w \leq 1$, $\rho^+_{ij} = \left|\rho_{ij}\right|$, $0 \leq \rho^+_{ij} \leq 1$; ρ_{ij} *is correlation coefficient of* $\ln \overline{Q}^{(p)}_{\overline{H}}(i)$ *and* $\ln \overline{Q}^{(p)}_{\overline{H}}(j)$

$$\rho_{ij} = \frac{Cov\left[\ln \overline{Q}^{(p)}_{\overline{H}}(i), \ln \overline{Q}^{(p)}_{\overline{H}}(j)\right]}{\sqrt{\left(Var\left(\ln \overline{Q}^{(p)}_{\overline{H}}(i)\right)Var\left(\ln \overline{Q}^{(p)}_{\overline{H}}(j)\right)\right)}}, \quad i \neq j, i, j = 1, ..., n-1, \ -1 \leq \rho_{ij} \leq 1, \ and$$

$$EFF\left(\ln \widehat{Q}^{(p)}_{New,\overline{H}}\right)_{\ln \widehat{Q}^{(p)}_{\overline{H}}(k)} = \frac{Var\left(\ln \overline{Q}^{(p)}_{\overline{H}}(k)\right)}{Var\left(\ln \widehat{Q}^{(p)}_{New,\overline{H}}\right)}, \quad where\ Var\left(\ln \overline{Q}^{(p)}_{\overline{H}}(k)\right) \underset{n\to\infty}{\approx} \gamma^2\frac{\left(\ln\left(\frac{k}{np}\right)\right)^2}{k}.$$

See Appendix A for the proof of Theorem 1.

3.3. The C.I. for The New Estimator $ln\widehat{Q}^{(p)}_{New,\overline{H}}$

Theorem 2. *Under conditions (C1) and (C2), a* $(1-\alpha)100\%$ *confidence interval for* $\ln VaR_p$ *by using* $\ln \overline{Q}^{(p)}_{New,\overline{H}}$ *in (19) is given by*

$$\left(LCL_{\ln \widehat{Q}^{(p)}_{New,\overline{H}}}(k), \ UCL_{\ln \widehat{Q}^{(p)}_{New,\overline{H}}}(k)\right) = \left(\ln \widehat{Q}^{(p)}_{New,\overline{H}} - UCL_{\overline{H}}(k)b3, \ \ln \widehat{Q}^{(p)}_{New,\overline{H}} + UCL_{\overline{H}}(k)b3\right) \tag{23}$$

where $z_{1-\alpha/2}$ *is the* $(1-\alpha/2)th$ *quantile of standard normal distribution, and*

$$b3 = \frac{z_{1-\alpha/2}}{n-1}\sqrt{\sum_{k=1}^{n-1}\frac{[\ln\left(\frac{k}{np}\right)]^2}{k} + w\sum_{i\neq j}^{n-1}\sum^{n-1}\left[\frac{\ln\left(\frac{i}{np}\right)\ln\left(\frac{j}{np}\right)}{\sqrt{i\bullet j}}\right]}, \quad UCL_{\overline{H}}(k) = \frac{\overline{H}}{1 - \frac{z_{1-\alpha/2}}{\sqrt{k}}}.$$

See Appendix A for the proof of Theorem 2.

Remark 1. *Note that in the CI in (23), the main term* $\ln \widehat{Q}^{(p)}_{new,\overline{H}}$ *does not depend on* k, *only the error terms* $UCL_{\overline{H}}(k)b_3$ *depends on* k.

Remark 2. *In Section 4 Simulations and Section 5 Applications, we use the maximum weight* $w = 1$ *in Formula (23), thus, we use maximum CI length for new proposed estimator* $ln\widehat{Q}^{(p)}_{New,\overline{H}}$ *comparing with existing methods. Even with maximum CI length. Sections 4 and 5 show that the new estimator obtained confidence interval in (23) is still shorter than existing estimators obtained confidence intervals for most of* k *values.*

4. Simulations

4.1. Computer Simulations of Quantile Estimators

To verify that the new estimator $\ln \widehat{Q}_{new,\overline{H}}^{(p)}$ has good properties, we use simulations and compare the new estimator to the existing estimators using the following statistics

1.   The expected value $E[\cdot]$.
2.   The root of mean squared errors $RMSE[\cdot]$.
3.   The relative efficiencies $REFF[\cdot]$

$$REFF_{\widetilde{Q}_{H\,or\,\overline{H}}} = \sqrt{\frac{MSE\left[\ln q_H^{(p)}(k_0)\right]}{MSE\left[\ln \widetilde{Q}_{H\,or\,\overline{H}}^{(p)}(k_0)\right]}} \text{ for } \widetilde{Q} = Q \text{ or } \overline{Q}, \ p = 1/(2n), \ k_0 \text{ is defined in (15).} \tag{24}$$

In this Section, we choose models of Fréchet (0.25), GPD (0.5), GPD (2) to compare with the simulation results of Gomes and Pestana (2007) [17]. We use four quantile estimators in Table 1 to run simulations. When $|\rho| \leq 1$, estimators $\widehat{\beta}$ and $\widehat{\rho}$ in $\overline{H}$ use the tuning parameter $\tau = 0$, otherwise, use $\tau = 1$

Table 1. The four ln-quantile estimators we use in simulations.

Quantile Estimators	Defined in	Tail Index Estimator
$\ln Q_{\widehat{\gamma}=H} = \ln q_H$	(6)	H in (5)
$\ln Q_H$	(7)	H in (5)
$\ln \widetilde{Q}_{\overline{H}}$	$\ln Q_{\overline{H}}$ when $\rho \neq -1$ in (12) $\ln \overline{Q}_{\overline{H}}$ when $\rho = -1$ in (14)	$\overline{H}$ in (9)
$ln\widehat{Q}_{new,\overline{H}}$	(19)	$\overline{H}$ in (9)

(1) The Fréchet distribution (Fréchet, 1927) [22] has the c.d.f.

$$F(x) = \exp\left(-x^{-\frac{1}{\gamma}}\right), x > 0, \ \gamma > 0. \tag{25}$$

An estimator of the *p*th ln-high quantile function is

$$\ln Q_{\widehat{\gamma}}^{(p)} = \ln x_{1-p,\widehat{\gamma}} = -\widehat{\gamma} \ln\left(\ln\left(\frac{1}{1-p}\right)\right), \ 0 < p < 1, \ p \text{ is very small.}$$

(2) The generalized Pareto distribution (GPD) (de Zea Bermudeza and Kotz, 2010) [23] has the c.d.f.

$$F(x; \gamma) = 1 - (1 + \gamma x)^{-\frac{1}{\gamma}}, \ x \geq 0, \ \gamma \neq 0, \tag{26}$$

for $\gamma > 0$, an estimator of the *p*th ln-high quantile function is

$$\ln Q_{\widehat{\gamma}}^{(p)} = \ln VaR_{\widehat{\gamma}} = \ln x_{1-p,\widehat{\gamma}} = \ln\left(\frac{p^{-\widehat{\gamma}} - 1}{\widehat{\gamma}}\right), \ 0 < p < 1, \ p \text{ is very small.}$$

4.2. The Choice of α

As mentioned in Section 3, α is a key value to reduce the bias of the $\ln \widehat{Q}_{New,\widehat{\gamma}}^{(p)}$ defined in (19). We developed an algorithm to estimate α based on the results of $m-$ simulation runs:

Step 1: For a fixed sample size n, the $\alpha_i(n)$ in ith iteration, $i = 1, ..., m$, $m = 500$, is the true solution of equation

$$\ln \widehat{Q}_{New,\overline{H}_i}^{(p)} = \frac{1}{n-1} \sum_{k=1}^{n-1}\left[\ln X_{i,n-k+1:n} + \overline{H}_i\left(\ln\left(\frac{k}{np}\right) + \alpha C_p(k; \widehat{\beta}, \widehat{\rho})\right)\right] = \ln VaR_p, i = 1..., m,$$

then $\alpha(n) = \frac{1}{m}\sum_{i=1}^{m}\alpha_i(n)$. Note that $\alpha(n)$ depends on n. $\ln VaR_p$ is the true lnVaR value.

Step 2: Obtain estimator $\widehat{\alpha}(n)$ based on the linear regression (LR) models where α is related to n. We collect data set (α_j, n_j), $j = 1, ..., l$, with the sample size l.

$$\widehat{\alpha}(n) = \begin{cases} \widehat{\rho}(n), & |\rho(n)| < 1, \text{ for GPD(0.5);} \\ 1.7488 - 0.0002n + 2.9693X_1 + 2.6604X_2, & |\rho(n)| \geq 1, \\ & \text{for Fréchet (0.25), } X_1 = \widehat{\beta}(n); X_2 = \widehat{\rho}(n); \\ & \text{for GPD(2), } X_1 = \widehat{\rho}(n); X_2 = \widehat{\beta}(n). \end{cases} \tag{27}$$

Note that the estimate $\widehat{\alpha}(n)$ in (27) depends the parameters of the models and LR relationship with sample size n.

Remark 3. *If we assume α_j in (α_j, n_j), $j = 1, ..., l$, is normally distributed, based on (Bickel and Doksum, 2015, pp. 286–388) [24], then $\widehat{\alpha}(n)$ is a maximum likelihood estimator (MLE) and has an asymptotic normal distribution. Since the estimator $\widehat{\alpha}(n)$ only depands to n not related to the order statistics, it will not affect the asymptotic proprties of the proposed estimator $\ln \widehat{Q}_{New,\widehat{\gamma}}^{(p)}$ in (19).*

4.3. Simulation of Fréchet (0.25). GPD (0.5) and GPD (2)

Tables 2–4 list the results of simulations under the Fréchet (0.25), GPD (0.5) and GPD (2), where $N = 500$ iterations for sample size $n = 500, 1000, 2000, 5000$ and $p = 1/(2n)$. With $\widehat{\alpha}(n)$ in (27), we compare mean values, mean squared errors (MSE) and REFF of the four $\ln VaR$ estimators in Table 1, at optimal level $k = k_0$ based on (15) Note that the new estimator $\ln \widehat{Q}_{New,\widehat{\gamma}}^{(p)}$ has the highest REFF values among the four estimators which are in bold in all three models. The simulation MSE of $\ln \widehat{Q}_{New,\widehat{\gamma}}^{(p)}$ is defined as

$$MSE\left(\ln \widehat{Q}_{New,\widehat{\gamma}}^{(p)}\right) = \frac{1}{N}\sum_{i=1}^{N}\left(\ln \widehat{Q}_{New,\widehat{\gamma},i}^{(p)} - \ln VaR_p\right)^2,$$

where $\ln \widehat{Q}_{New,\widehat{\gamma},i}^{(p)}$ is the $\ln \widehat{Q}_{New,\widehat{\gamma}}^{(p)}$ in the ith iteration, $i = 1, ..., N$. So do for other ln-quatile estimators.

Figures 1–3 are based on Tables 2–4 results, Figure 1 is for Fréchet (0.25), we use $N = 500$ iterations, sample size $n = 1000$, $\gamma = 0.25$, $\rho = -1$, $\beta = 0.5$, $p = 1/2n$. The new estimator $\ln \widehat{Q}_{New,\widehat{\gamma}}^{(p)}$ has the best performance with the least bias and *RMSE*. It does not change as k varies. Figures 2 and 3 are for GPD(0.5) and GPD(2), $N = 500$ iterations, sample size $n = 1000$, $\gamma = 0.5$ and 2, $\rho = -\gamma$, $\beta = 1$, $p = 1/2n$. We note that the new estimator $\ln \widehat{Q}_{new,\overline{H}}$ is the best estimator as well, with the least bias, consistency as k varies, and the smallest *RMSE*. Note that $\ln \widehat{Q}_{New,\widehat{\gamma}}^{(p)}$ values are very close to the true lnVaR$_p$ values.

Table 2. Fréchet (0.25), $N = 500$, $\beta = 0.5$, $\rho = -1$. Mean, MSE, REFF of the $\ln VaR$ Estimators. The highest REFF values are in bold.

n		500	1000	2000	5000
$\ln VaR_p$, $p = 1/(2n)$		1.7268	1.9002	2.0735	2.3026
k_0		126	200	318	585
$\widehat{\alpha}(n) = LR$		1.1218	1.1400	0.4991	-0.0357
$\ln q_H$	Mean (MSE)	1.8526 (0.0429)	2.0038 (0.0300)	2.1657 (0.0228)	2.3755 (0.0147)
	REFF	1	1	1	1
$\ln Q_H$	Mean (MSE)	1.7906 (0.0219)	1.9540 (0.0154)	2.1239 (0.0115)	2.3431 (0.0074)
	REFF	1.4004	1.3933	1.4104	1.4125
$\ln \overline{Q}_{\overline{H}}$	Mean (MSE)	1.7092 (0.0206)	1.8849 (0.0141)	2.0764 (0.0111)	2.3073 (0.0072)
	REFF	1.4419	1.4576	1.4347	1.4257
$\ln \widehat{Q}_{new,\overline{H}}$	Mean (MSE)	1.7185 (0.0095)	1.8791 (0.0065)	2.0716 (0.0051)	2.2798 (0.0044)
	REFF	**2.1252**	**2.1399**	**2.1139**	**1.8231**

Table 3. GPD (0.5), $N = 500$, $\beta = 1$, $\rho = -0.5$. Mean, MSE, REFF of the $\ln VaR$ estimators. The highest REFF values are in bold.

n		500	1000	2000	5000
$\ln VaR_p$, $p = 1/(2n)$		4.1149	4.4710	4.8242	5.2883
k_0		34	48	68	107
$\widehat{\alpha}(n) = \widehat{\rho}$		-0.7512	-0.7482	-0.7427	-0.7244
$\ln q_H$	Mean (MSE)	4.7019 (0.6349)	4.9773 (0.4863)	5.3065 (0.4554)	5.7209 (0.3427)
	REFF	1	1	1	1
$\ln Q_H$	Mean (MSE)	4.2913 (0.2172)	4.6258 (0.1628)	4.9904 (0.1491)	5.4485 (0.1074)
	REFF	1.7159	1.7282	1.7478	1.7865
$\ln Q_{\overline{H}}$	Mean (MSE)	4.1140 (0.1654)	4.4801 (0.1267)	4.8656 (0.1166)	5.3434 (0.0825)
	REFF	1.9663	1.9591	1.9763	2.0379
$\ln \widehat{Q}_{new,\overline{H}}$	Mean (MSE)	3.9076 (0.0779)	4.4239 (0.0233)	4.8674 (0.0241)	5.4359 (0.0382)
	REFF	**2.8657**	**4.5666**	**4.3428**	**2.9954**

Table 4. GPD(2), $N = 500$, $\beta = 1$, $\rho = -2$. Mean, MSE, REFF of the $\ln VaR$ estimators. The highest REFF values are in bold.

| n		500	1000	2000	5000
$\ln VaR_p$, $p = 1/(2n)$		13.1224	14.5087	15.8949	17.7275
k_0		170	269	515	1071
$\widehat{\alpha}(n) = LR$		-2.7893	-2.8417	-2.8687	-3.1684
$\ln q_H$	Mean (MSE)	13.6276 (1.3232)	14.9415 (0.9745)	16.2733 (0.6548)	18.0099 (0.3833)
	REFF	1	1	1	1
$\ln Q_H$	Mean (MSE)	13.4502 (0.9965)	14.8004 (0.7283)	16.1618 (0.4779)	17.9283 (0.2804)
	REFF	1.1523	1.1567	1.2412	1.1693
$\ln Q_{\overline{H}}$	Mean (MSE)	13.1933 (0.8926)	14.5960 (0.6477)	15.9719 (0.4491)	17.7717 (0.2751)
	REFF	1.2175	1.2267	1.2075	1.1804
$\ln \widehat{Q}_{new,\overline{H}}$	Mean (MSE)	13.0009 (0.6007)	14.4907 (0.3680)	15.8926 (0.3070)	17.6429 (0.2127)
	REFF	**1.4841**	**1.6274**	**1.4606**	**1.3426**
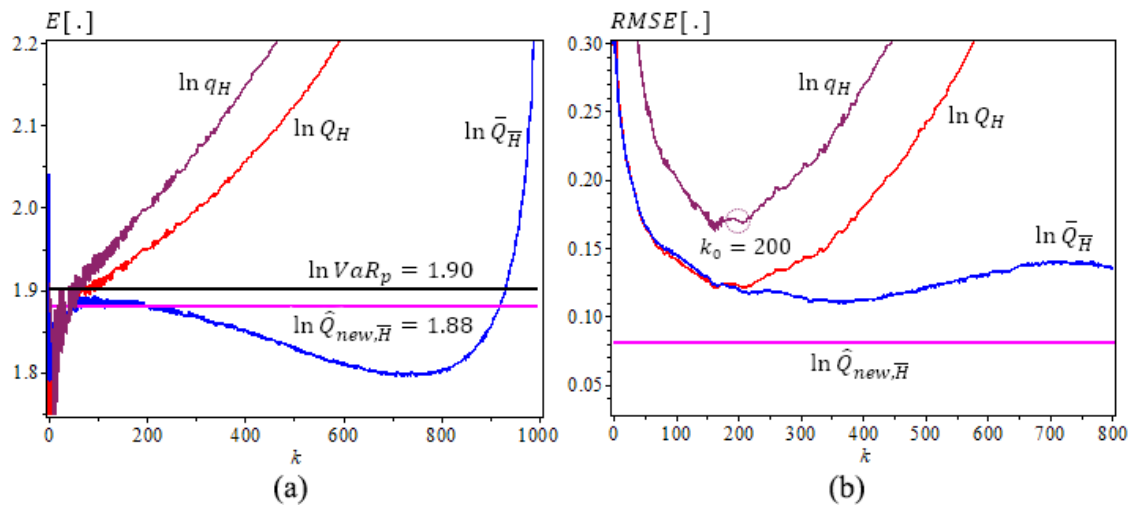
Figure 1. Underlying *Fréchet* (0.25), $\rho = -1$, $\beta = 0.5$. $N = 500$, $n = 1000$. (**a**) The means of ln-quantile estimators with the true $\ln VaR_{0.0005} \approx 1.9$ ($\ln \widehat{Q}_{new,\overline{H}}^{0.0005} \approx 1.88$). (**b**) The RMSE of Ln-quantile estimation, $p = 0.0005$, $\widehat{\alpha} = 1.14$.
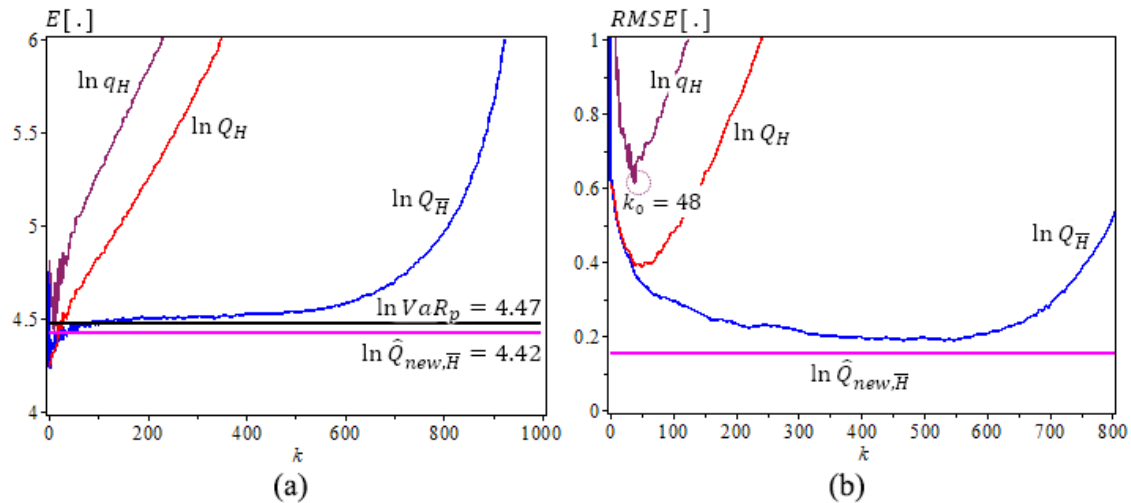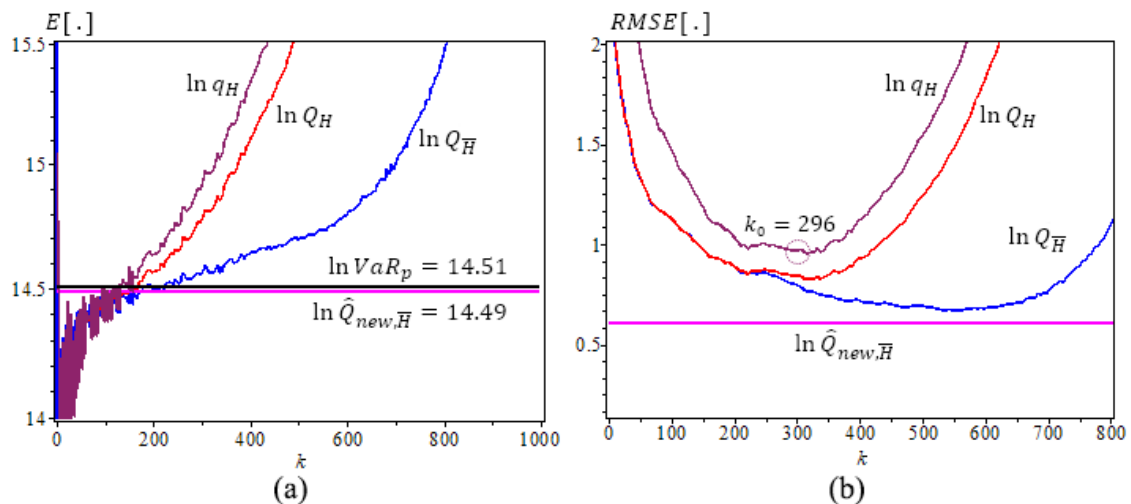

Figure 2. Underlying *GPD* (0.5), $\rho = -0.5$, $\beta = 1$, $N = 500$, $n = 1000$. (**a**) The means of ln-quantile estimators with the true $\ln VaR_{0.0005} \approx 4.47$ ($\ln \widehat{Q}_{new,\overline{H}}^{0.0005} \approx 4.42$). (**b**) The RMSE of Ln-quantile estimation, $p = 0.0005$, $\widehat{\alpha} = -0.7482$.

Figure 3. Underlying *GPD* (2), $\rho = -2$, $\beta = 1$, $N = 500$, $n = 1000$. (**a**) The means of ln-quantile estimators with the true $\ln VaR_{0.0005} \approx 14.51$ ($\ln \widehat{Q}_{new,\overline{H}}^{0.0005} \approx 14.49$). (**b**) The RMSE of ln-quantile estimators, $p = 0.0005$, $\widehat{\alpha} = -2.8417$.

4.4. Simulations of Confidence Intervals

By Gomes and Pestana (2007) [17], the 95% confidence interval of the true tail index using H is

$$(LCL_H(k), UCL_H(k)) = \left(\frac{H(k)}{1 + \frac{\beta(n/k)^\rho}{1-\rho} + \frac{1.96}{\sqrt{k}}}, \frac{H(k)}{1 + \frac{\beta(n/k)^\rho}{1-\rho} - \frac{1.96}{\sqrt{k}}} \right) \tag{28}$$

and the 95% confidence interval of the true tail index using $\overline{H}$ is

$$(LCL_{\overline{H}}(k), UCL_{\overline{H}}(k)) = \left(\frac{\overline{H}(k)}{1 + \frac{1.96}{\sqrt{k}}}, \frac{\overline{H}(k)}{1 - \frac{1.96}{\sqrt{k}}} \right) \tag{29}$$

Next, we compute the confidence intervals for the true ln-quantile by using the quantile estimators. We only use three out of four quantile estimators in Table 1, except $\ln q_H$ which has the worst result. Therefore, we compare CIs only using $\ln Q_H$, $\ln \widetilde{Q}_{\overline{H}}$ and $\ln \widehat{Q}_{new,\overline{H}}$ in (30), (31) and (23). Thus

(1) The 95% confidence interval for the true $\ln VaR_p$ using $\ln Q_H$ is

$$LCL_{\ln Q_H}(k) = \min \ln Q_H(k) - LCL_H(k) \ln\left(\frac{k}{np}\right) b_2, \ln Q_H(k) - UCL_H(k) \ln\left(\frac{k}{np}\right) b_2;$$

$$UCL_{\ln Q_H}(k) = \max \ln Q_H(k) + LCL_H(k) \ln\left(\frac{k}{np}\right) b_1, \ln Q_H(k) + UCL_H(k) \ln\left(\frac{k}{np}\right) b_1.$$

$$\tag{30}$$

where $LCL_H(k)$, $UCL_H(k)$ is given in (28), and

$$b_1 = \frac{1.96}{\sqrt{k}} - \frac{\beta(n/k)^\rho}{1-\rho}, \; b_2 = \frac{1.96}{\sqrt{k}} + \frac{\beta(n/k)^\rho}{1-\rho}.$$

(2) The 95% confidence interval for the true $\ln VaR_p$ using $\ln \widetilde{Q}_{\overline{H}}$ is

$$LCL_{\ln \widetilde{Q}_{\overline{H}}}(k) = \ln \widetilde{Q}_{\overline{H}} - UCL_{\overline{H}}(k) \ln\left(\frac{k}{np}\right) \frac{1.96}{\sqrt{k}},$$

$$UCL_{\ln \widetilde{Q}_{\overline{H}}}(k) = \ln \widetilde{Q}_{\overline{H}} + UCL_{\overline{H}}(k) \ln\left(\frac{k}{np}\right) \frac{1.96}{\sqrt{k}}; \tag{31}$$

where $LCL_{\overline{H}}(k)$, $UCL_{\overline{H}}(k)$ is given in (29), and

(3) The 95% confidence interval for the true $\ln VaR_p$ using $\ln \widehat{Q}_{new,\overline{H}}$ is given in (24).

To compare new proposed CI in (23) to CIs in (30) and *(31)*, we use evaluate the length and probability coverage of the CIs.

The length of CI is given as

$$length \text{ of CI} = UCL_{\text{quantile estimator}} - LCL_{\text{quantile estimator}}.$$

and the efficiency of the length of 95% *CI* is given as

$$EFF_{length} = \frac{C.I. \; length \; of \; \ln Q_H \; at \; k_0^{Q_H}}{C.I. \; length \; of \; \ln \overline{Q}_{\overline{H}} \; or \; \ln \widehat{Q}_{new,\overline{H}} \; at \; k_{01}}. \tag{32}$$
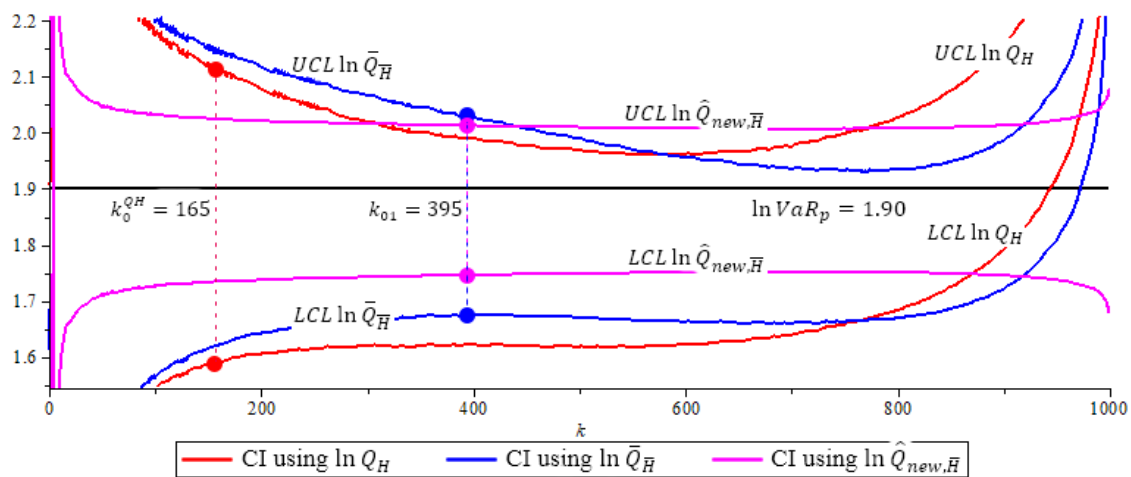
Figure 4. *Fréchet* (0.25) model, 95% confidence interval of quantile estimators, $N = 500$, $n = 1000$, $p = 0.0005$, $\beta = 0.5$, $\rho = -1$, $\alpha(1000) = LR = 1.14$, $k_0^{Q_H} = 165$, $k_{01} = 395$. Note that $\ln \widehat{Q}_{new,\overline{H}}$ (purple) has shortest CI with length 0.2668. (The solid circles "•" in the plot are the values of the quantile estimators at their optimal k level).
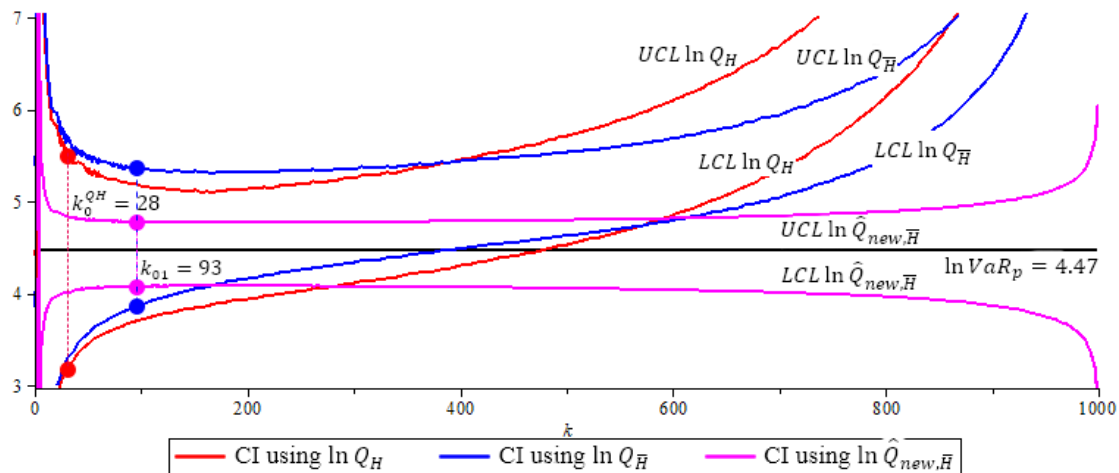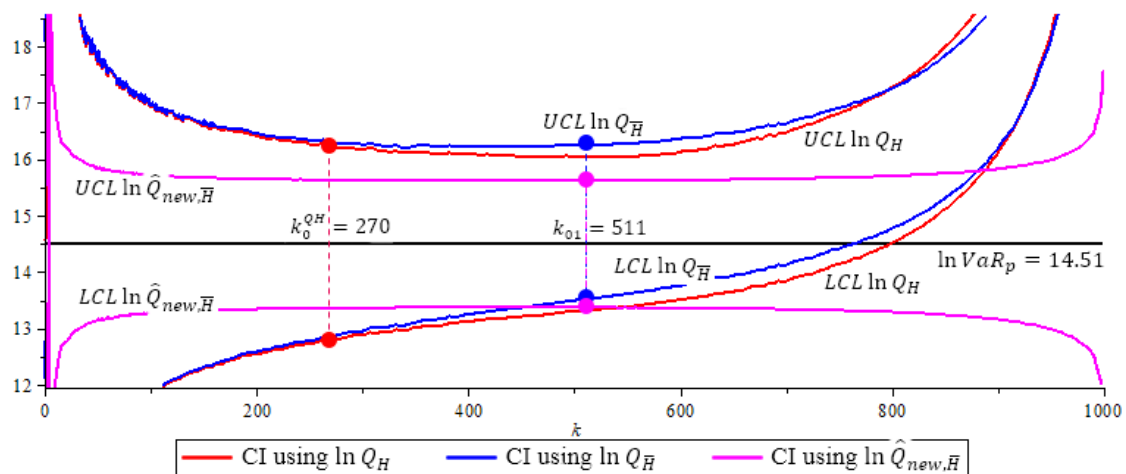

Figure 5. The *GPD* (0.5) model, 95% confidence interval of quantile estimators, $N = 500$, $n = 1000$, $p = 0.0005$, $\beta = 1$, $\rho = -0.5$, $\widehat{\alpha}(1000) = \widehat{\rho} = -0.7482$, $k_0^{Q_H} = 28$, $k_{01} = 93$. Note that $\ln \widehat{Q}_{new,\overline{H}}$ (purple) has shortest CI with length 0.7094. (The solid circles "•" in the plot are the values of the quantile estimators at their optimal k level).

Figure 6. The *GPD* (2) model, 95% confidence interval of quantile estimators $N = 500$, $n = 1000$, $p = 0.0005$, $\beta = 1$, $\rho = -2$, $\widehat{\alpha}(1000) = LR = -2.8417$, $\widehat{k}_0 = 75$, $\widehat{k}_0^{Q_H} = 80$, $\widehat{k}_{01} = 70$. Note that $\ln \widehat{Q}_{new,\overline{H}}$ (purple) has shortest CI with length 2.2511. (The solid circles "•" in the plot are the values of the quantile estimators at their optimal k level).

Also, the confidence interval is more efficient when it has a higher coverage of the true value under the simulations, where the probability coverage of 95% CI is defined as

$$P.C. = \frac{\text{number of 95\% } CI\text{'s contains the true value}}{\text{number of 95\% } CI\text{'s simulated in total}} * 100\%.$$

and the efficiency of the probability coverage of 95% CI is given as

$$EFF_{P.C.} = \frac{|P.C._{\cdot \ln Q_H} - 95\%|}{\left| \left(P.C._{\cdot \ln \overline{Q_H}} \text{ or } P.C._{\cdot \ln \widehat{Q}_{new,\overline{H}}} \right) - 95\% \right|}. \tag{33}$$

when $EFF_{P.C.}$ is bigger means it is more efficient.

Figures 4–6 show the 95% confidence interval of the three ln-quantile estimators under Fréchet (0.25), GPD (0.5 and 2) with $p = 0.0005$. We compare the size of each confidence interval at their optimal k level, and the probability coverage of each confidence interval at their optimal k level. Recall, the optimal k level for $\ln Q_H$ is at $k_0^{Q_H}$ based in (16), the optimal k level for $\ln \widetilde{Q_H}$ and $\ln \widehat{Q}_{new,\overline{H}}$ is at k_{01} based in (15).

Table 5 compare the efficiencies of 95% CI of the three quantile estimators under *Fréchet* (0.25), *GPD* (0.5 and 2). The efficiency of 95% CI can be compared by the length of CI and the probability coverage of CI, denoted by EFF_{length} and $EFF_{P.C.}$.

Table 5. $N = 500, n = 1000$, efficiencies of 95% CI for $\ln VaR_{0.01}$.

	CI of	at Optimal k	Length	EFF_{length}	Probabolity Coverage	$EFF_{P.C.}$
	$\ln Q_H$	$k_0^{Q_H} = 165$	0.5142	1	94.2%	1
Fréchet (0.25)	$\ln \overline{Q_H}$	$k_{01} = 395$	0.3564	1.4517	96.7%	0.4706
	$\ln \widehat{Q}_{new,\overline{H}}$	$k_{01} = 395$	0.2668	1.9275	99.6%	0.1739
	$\ln Q_H$	$k_0^{Q_H} = 28$	2.4922	1	47.4%	1
GPD(0.5)	$\ln Q_{\overline{H}}$	$k_{01} = 93$	1.5204	1.6392	79.0%	2.9750
	$\ln \widehat{Q}_{new,\overline{H}}$	$k_{01} = 93$	0.7094	3.5130	99.6%	10.3478
	$\ln Q_H$	$k_0^{Q_H} = 270$	3.4410	1	79.7%	1
GPD(2)	$\ln Q_{\overline{H}}$	$k_{01} = 511$	2.7291	1.2609	83.2%	1.2966
	$\ln \widehat{Q}_{new,\overline{H}}$	$k_{01} = 511$	2.2511	1.5286	99.6%	3.3261

In this section, we compared the new quantile estimator $\ln \widehat{Q}^{(p)}_{new,\overline{H}}$ in (19) with the existing methods. $\ln \widehat{Q}^{(p)}_{new,\overline{H}}$ has the least bias, the smallest $RMSE$, and not depends on k too much. It also has the smallest length and the highest probability coverage in 95% confidence interval in most cases. The simulation results verify that $\ln \widehat{Q}_{new,\overline{H}}$ is the best quantile estimator among all three methods. Next section, we apply the new estimator $\ln \widehat{Q}^{(p)}_{new,\overline{H}}$ to real world examples.

5. Applications

We will study two real-world examples in this Section. We are interested in the population that is above the threshold for each example. The goal is to estimate the $(1-p)th$ high quantiles of the example, where $0 < p < 1$ is a very small. We use the four quantile estimators in table 1 $\ln q_H$, $\ln Q_H$, $\ln Q_{\overline{H}}$ and $\ln \widehat{Q}_{new,\widehat{\gamma}}$ in (21), and compare their performances.

A. Procedure:

   Step 1: Choose and collect data of examples of real life extreme events.

   Step 2: Run Goodness-of-Fit tests to check if data is heavy distributed.

   Step 3: Estimate the high quantiles and construct the confidence intervals by using the new method and the existing methods.

   Step 4: Compare the results of the new method to the existing methods and make Prediction

B. Estimators

   1. Two tail index estimators $H(k)$ in (5) and $\overline{H}(k)$ in (9).

   2. Four quantile estimators (6), (7), (12) and (19) are in Table 1.

   3. We use $\widehat{\alpha}(n)$ in (27) for the new estimator $\ln \widehat{Q}^{(p)}_{new,\overline{H}}$ in (19) for the GPD model.

Remark 4. *In applications, the GPD is used as a tail approximation to the population distribution from which a sample of excesses $x - \mu$ above some suitably high threshold μ are observed. The GPD is parameterized by location, scale and shape parameters μ, $\lambda > 0$ and γ, and can equivalently be specified in terms of threshold excesses $x - \mu$ or, as here, exceedances $x > \mu$, as three parameters (γ, μ, λ) GPD in (34) (de Zea Bermudeza and Kotz, 2010) [23],*

$$H_\gamma(x) = 1 - \left(1 + \gamma \frac{x - \mu}{\lambda}\right)^{-\frac{1}{\gamma}}, 0 < \mu < x < (0 \vee (-\gamma))^{-1}, \lambda > 0, \tag{34}$$

Traditionally, the threshold was chosen before fitting, giving the so-called fixed threshold approach (Pickands, 1975 [25], Balkema and de Haan, 1974 [26]). It is common for practitioners to assume a constant quantile level, determined by some assessment of fit across all or a subset of the datasets (Scarrott and McDonald, 2012, p.36 [27]). In our application, the threshold is pre-determined by physical considerations, that is, number of type A flu viruses detected weekly in Canada above the average in flu season, and the counts of gamma ray released from significant solar flares (M and X rated) during the Sun's active years. Although it is possible to make some arbitrary definition of the choice of the threshold, it is preferable not to become involved with such delicate question. The application of the proposed method is presented in both examples for illustrative purpose.

5.1. Flu in Canada Example

   According to the WHO (World Health Organization, 2020 [28]), seasonal influenza is a common infection of the airways and lungs that can spread easily among humans. There are 37 million people in Canada, and flu season usually runs from November to April. Most people recover from the flu in about a week. However, influenza may be associated with serious complications such as pneumonia, especially in infants, the elderly and those with underlying medical conditions like diabetes, anemia, cancer, and immune suppression. On average, the flu and its complications send about 12,200 Canadians to the hospital every year, and around 3500 Canadians die. There are 3 types of flu viruses, A, B and C. Type A flu virus is the most harmful, and it is constantly changing and is generally responsible for the large flu epidemics. The 1918 Spanish Flu, 1957 Asian Flu, 1968 Hong Kong Flu, 2009 Swine flu, and the most recent 2014 H5N1 Bird Flu are all type A flu. In this paper, we study type A viruses in Canada.

   We collected the number of the type A flu viruses detected weekly in Canada, from 1 January 1997 to 31 December 2019, resulting in a sample size of $n^* = 994$ weeks. According to the WHO, the average number of type A flu viruses detested per week in the flu season, November to April, is 953, for the past 10 years. We set 953 viruses/week as the threshold, which reduced our sample size to $n = 111$ weeks. Full data-set is available at http://apps.who.int/influenza/gisrs_laboratory/flunet/en.

   Figure 7a shows a Flu chart in $n^* = 994$ weeks of type A flu viruses detected in Canada, and $n = 111$ weeks remaining after the threshold, of average 953 flu viruses. For each flu incubation period, a flu virus can last from one up to few weeks, that is why some arches are narrow and some arches are more bell shaped in this figure. The top three weeks are

circled in the plot. Figure 7b shows a histogram of $n^* = 994$ weeks data. We are interested in the 99% quantile, $x_{0.99}$, such that 99% chance that the viruses detected in a given week would be less than this value, or equivalently, with a 1% possibility, the number of flu viruses detested in a given week would be in excess of this value. This information is useful for monitoring and studying the virus, also is helpful for medical organizations that deal with disease control and prevention, pharmaceutical availability, and hospital resource readiness, especially during a serious flu outbreak. $\hat{x}_{0.99}$ is approximately located in the plot. In this paper, we propose a new estimate high quantiles method, and compare it with existing methods.
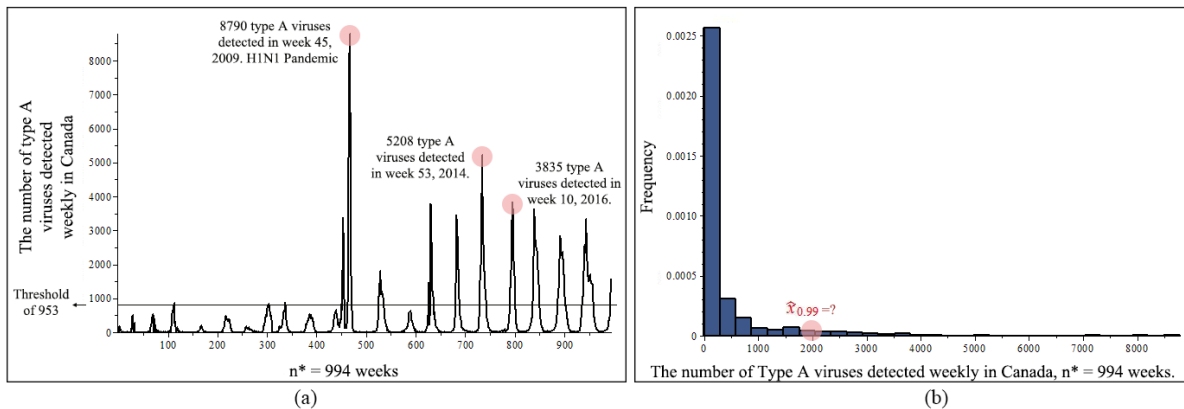

Figure 7. Flu original data from 1 January 1997 to December 31 2019, $n^* = 994$ weeks, (**a**) Flu chart of type A flu viruses detected in Canada, and $n = 111$ weeks remaining after the threshold, of average 953 flu viruses. (**b**) Histogram of the number of type A flu viruses detected in Canada.

Our interest is to find the 5%VaR and 1% VaR of the number of type A flu viruses detested in a week, and their 95% confidence intervals.

5.1.1. Goodness-of-Fit Test

Through data transformation $Y_i = \frac{X_i - \mu}{\lambda}$, $i = 1, ..., n$, $n = 111$. Take $\mu = 953$ as the threshold, the maximum likelihood estimators (MLE) are $\hat{\lambda}_{MLE} = 1275.97287$ and $\hat{\gamma}_{MLE} = 0.01345$. Figure 8a is the log-log plot of GPD curve with the horizontal axis $\ln(x)$ against the vertical axis $\ln(P\{x < X\})$. Visually the transformed data fit the one parameter GPD in (26) the best using $\hat{\gamma}_{MLE}$ (red curve). Figure 8b shows the GPD density curve (red curve) fits the histogram very well.
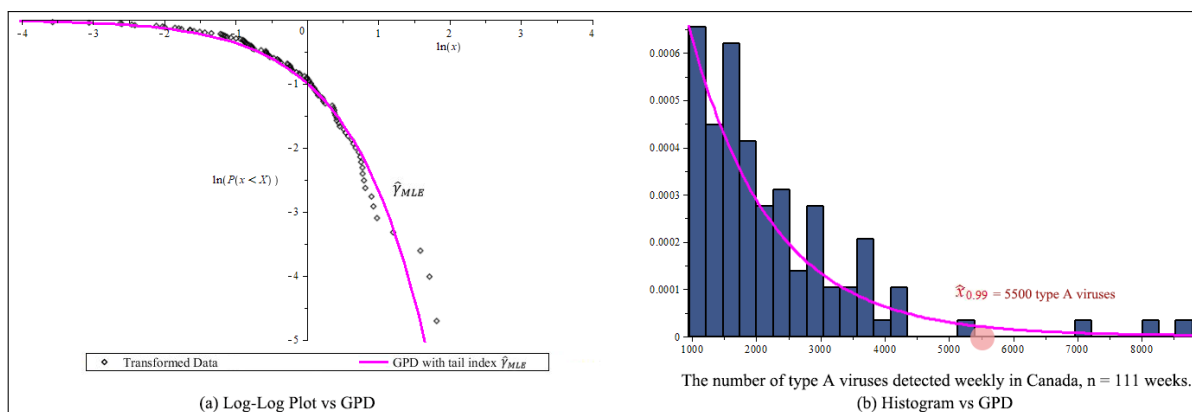

(a) Log-Log Plot vs GPD

(b) Histogram vs GPD

Figure 8. After threshold 953 flu viruses, Flu transformation data, $n = 111$, (**a**) Log-log plot of flu in Canada example. (**b**) Estimate GPD curve and the 99% high quantile and histogram of the distribution of type A flu viruses detested weekly.

Beside visual view of Figure 8, we also carry on the three goodness-of-fit tests: the Kolmogorov-Smirnov (*K-S*) test (Kolmogorov, 1933 [29]), Anderson-Darling (*A-D*) test, and Cramér von Mises (*C-v-M*) test (Anderson-Darling, 1952 [30]). All three tests are based on the maximum vertical distance between the empirical distribution function and the observations, and the parent distribution function is the *GPD*.

The Hypothesis for all three tests is

$$
\begin{aligned}
H_0 &: \quad F(x) = F^*(x), \text{ for all values of } x \\
H_1 &: \quad F(x) \neq F^*(x), \text{ for at least one value of } x
\end{aligned}
$$

$F(x)$ is the true but unknown distribution of the sample. $F^*(x)$ is the theoretical distribution, in our project, the parent distribution, *GPD*. $S_n(x)$ is the empirical distribution and step function of the sample. It is defined as

$$
S_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty, x]}(X_i), \quad \text{where} \qquad I_A = \left\{ \begin{array}{ll} 1, & \text{if } x \in A; \\ 0 & \text{if } x \notin A. \end{array} \right.
$$

where $-\infty < x < \infty$, $0 \leq S_n(x) \leq 1$.

The test statistics under H_0 of $K - S$ test is

$$
T = \sup_x |F^*(x) - S_n(x)|. \tag{35}
$$

Based on Table 6 goodness of fit tests' results, we set the GPD model for the flu in Canada data. We define the absolute errors (*AE*) in (34) and integrated errors (*IE*) in (35) as

$$
IE = \frac{1}{X_{n:n} - X_{n-r+1:n}} \left(\int_{X_{n-r+1:n}}^{X_{n:n}} [S_n(x) - F^*(x)]^2 dx \right)^{1/2}. \tag{36}
$$

Table 6. The goodness-of-fit tests under the *GPD* model for the flu in Canada data.

	Goodness-of-Fit Tests					
	K-S Test		A-D Test		C-v-M Test	
	Test Statistics	*p*-Value	Test Statistics	*p*-Value	Test Statistics	*p*-Value
$\widehat{\gamma}_{MLE}$	0.0628	0.6406	0.4475	0.8007	0.0621	0.8006

For both *AE* and *IE*, we use 3 different r values by letting $r = \frac{n}{10}th$, $r = \frac{n}{2}th$, and $r = nth$ top statistics. Table 7 lists the *AE* and *IE* errors which are very small.

Table 7. *AE* and *IE* under the *GPD* model for the flu in Canada data by using $\widehat{\gamma}_{MLE}$.

	Absolute Errors (*AE*)			Integrated Errors (*IE*)		
	*r*th Highest Amount of Type A Viruses			*r*th Highest Amount of Type A Viruses		
	$r = 12$	$r = 56$	$r = 111$	$r = 12$	$r = 56$	$r = 111$
$\widehat{\gamma}_{MLE}$	0.0450	0.0450	0.0628	0.0085	0.0071	0.0074

Next, we estimate the high quantiles and their confidence interval for this example.

5.1.2. Compare Four Estimation Methods

We use the four estimators in Table 1: $\ln q_H$, $\ln Q_H$, $\ln Q_{\overline{H}}$, and the new estimator $\ln \widehat{Q}_{new,\overline{H}}$.

We use $\widehat{\rho}_\tau(k)$ in (10), and $\widehat{\beta}_{\widehat{\rho}_0}(k)$ in (11). To decide if the tuning parameter $\tau = 0$ or 1, consider $\{\widehat{\rho}_\tau(k)\}_{k\in\Bbbk}$, for $k \in \Bbbk = \left(\left[n^{0.995}\right], \left[n^{0.999}\right]\right)$, and compute their median x_τ, then

$$\tau = \arg\min_k \sum_{k\in\Bbbk}(\widehat{\rho}_\tau(k) - x_\tau)^2.$$

With $n = 111$, we get $k \in \Bbbk = (108, 110)$ and $x_\tau = 109$, then $\sum_{k\in\Bbbk}(\widehat{\rho}_0(k) - x_\tau)^2 \approx 36116 < \sum_{k\in\Bbbk}(\widehat{\rho}_1(k) - x_\tau)^2 \approx 37033$, conclude that $\tau = 0$, thus we have $\widehat{\rho}_0(k_1) = -0.7101$ and $\widehat{\beta}_{\widehat{\rho}_0}(k_1) = 1.026571$, where k_1 is the optimal k value. Figure 9 shows the results.
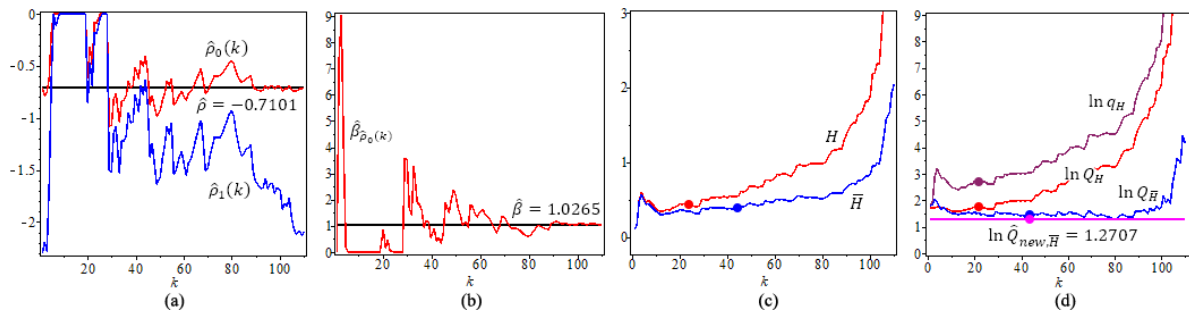

Figure 9. For flu in the Canadian data, $n = 111$, (**a**) Estimates of the second-order parameter $\widehat{\rho}$ and $\widehat{\rho}_\tau(k)$, $\tau = 0$; (**b**) Estimates $\widehat{\beta}$ and $\widehat{\beta}_{\widehat{\rho}_0}(k)$. (**c**) Tail index estimators, H, $\overline{H}$; (**d**) ln-quantile estimators, $p = 0.01$. The solid circles "•" in the plot are the values of the quantile estimators at their optimal k level.

Figure 9a shows estimates of the second-order parameters ρ through $\widehat{\rho}$ and $\widehat{\rho}_\tau(k)$, $\tau = 0$; Figure 9b shows Estimates $\widehat{\beta}$ and $\widehat{\beta}_{\widehat{\rho}_0}(k)$. Figure 9c shows the two estimated tail index, H, $\overline{H}$, $H = 0.4379$ at its optimal level using $\widehat{k}_0 = 21$ based on (15) and $\overline{H} = 0.3736$ at its optimal level using $\widehat{k}_{01} = 42$ based on (17). Figure 9d shows four quantile estimators of flu in Canada example, with $p = 0.01$. The full circles "•" in the plot are the values of the quantile estimators at their optimal k level. We note that $\ln \widehat{Q}_{new,\overline{H}}^{(p)}$ has a constant value, which does not depend on k.

Figure 10 compares the confidence intervals of three quantile estimators in (7), (12) and (19). This figure shows that the new quantile estimator $\ln \widehat{Q}_{new,\overline{H}}^{(p)}$ has the smallest confidence interval with length 0.7966, where we use $\widehat{\alpha} = \widehat{\rho} = -0.7101$. (The solid circles "•" in the plot are the values of the quantile estimators at their optimal k level).
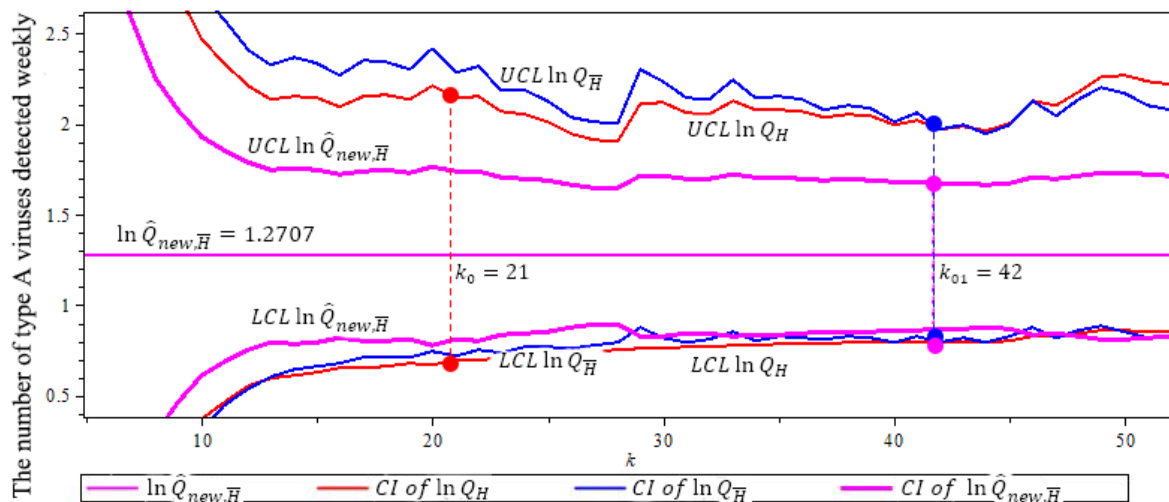

Figure 10. 95% confidence interval of three ln-quantile estimators after the threshold 953 for the flu in Canada example. $n = 111$, $p = 0.01$. Note that $\ln \widehat{Q}_{new,\overline{H}}$ (purple) has shortest CI with length 0.7966. (The solid circles "•" in the plot are the values of the quantile estimators at their optimal k level).

In Table 8, we compare the four ln-quantile estimators and their mean, median, $VaR_{0.05}$ and $VaR_{0.01}$. Table 9 compares the size of confidence intervals at $\ln VaR_{0.01}$ and $VaR_{0.01}$ of the three quantile estimators.

Table 8. Estimated $VaR_{0.05}$ and $VaR_{0.01}$ for the flu in Canada data. (Unit: Type A flu viruses).

Estimation	$\widehat{\alpha}$	$\widehat{\gamma}$	Mean	Median	$VaR_{0.05}$	$VaR_{0.01}$
$\ln Q_H$	N/A	$H = 0.4370$	3219.29	2257.03	4519.70	8159.10
$\ln Q_{\overline{H}}$	N/A	$\overline{H} = 0.3736$	2989.93	2130.78	3736.80	6031.79
$\ln \widehat{Q}_{new,\overline{H}}$	$\widehat{\rho} = -0.7101$	$\overline{H} = 0.3736$	2989.93	1690.07	2924.80	5499.85

Table 9. The 95% confidence interval for $\ln VaR_{0.01}$ and $VaR_{0.01}$.

Estimation Method	k	LCL	$\ln VaR_{0.01}$ $(VaR_{0.01})$	UCL	Length	EFF
$\ln Q_H$	$\widehat{k}_0 = 21$	0.6920	1.7312	2.1452	1.4531	1
(Q_H)		(3502.14)	(8159.10)	(11854.31)	(8352.17)	(1)
$\ln Q_{\overline{H}}$	$\widehat{k}_{01} = 42$	0.7929	1.3814	1.9698	1.1770	1.2346
$(Q_{\overline{H}})$		(3772.58)	(6031.78)	(10101.19)	(6328.21)	(1.3197)
$\ln Q_{new,\overline{H}}$	$\widehat{k}_{01} = 42$	0.8724	1.2707	1.6690	0.7966	1.8242
$(Q_{new,\overline{H}})$		(4006.07)	(5499.85)	(7724.49)	(3718.42)	(2.2462)

In Table 9, we compared Q_H, $Q_{\overline{H}}$ and $Q_{new,\overline{H}}$, the $Q_{new,\overline{H}}$ has the shortest confidence interval with the highest efficiency of 2.2462.

5.1.3. Summary

Based on Figure 10 and Table 9, we conclude that the new estimator $\ln \widehat{Q}_{new,\overline{H}}$ in (19) is the best estimator for Flu in Canada example. We can predict that at $VaR_{0.01}$, we expect 5500 type A flu viruses during a flu outbreak after threshold 953/week. This is shown in Figure 8b.

5.2. *Gamma Ray of Solar Flare Example*

Gamma ray has the most penetrating power among all the radiations. The burst of gamma rays are thought to be, due to the collapse of stars called hypernovas, the most powerful events so far discovered in the cosmos. The measurement of gamma rays are in counts, and it is the number of atoms in a given quantity of radioactive material that are detected by an instrument to have decayed. We have collected gamma ray data from solar flares, from November 2008 to September 2020, from NASA (National Aeronautics and Space Administration, 2020 [31]). Full data-set is available at http://hesperia.gsfc.nasa.gov/fermi/gbm/qlook/fermi_gbm_flare_list.txt.

The solar flare travels hundreds of miles per second, and can reach the Earth within hours. It can disrupt communication navigational equipment, damage satellites, and even cause blackouts by damaging power plants. In 1989, a strong solar storm knocked out the power grid in Québec, Canada, causing 6 million people to lose power for more than 9 hours, and it cost millions of dollars to repair. It can bring additional radiation around the north and south poles, a risk that forces airlines to reroute flights. The Fermi Gamma-ray Space Telescope was launched in late 2008 to explore high-energy phenomena in the Universe. It is worth noting that more than one trigger may have occurred during the flare, the one nearest the peak of the flare is listed, resulting in a sample size of 5128.

Solar flares are classified as A, B, C, M or X according to the peak flux (in watts per square meter, W/m^2) of 1 to 8 angstrom (The angstrom is a unit of length equal to 1/10,000,000,000 (one ten-billionth) of a meter.) X-rays near the Earth, as measured on the GOES spacecraft. Gamma ray activity is correlated with the X ray activity, as shown in Figure 11 (NOAA, 2020 [32]. When the amount of gamma ray released is over 5 million counts, it usually corresponds to an X rated flare or significant M rated flares.
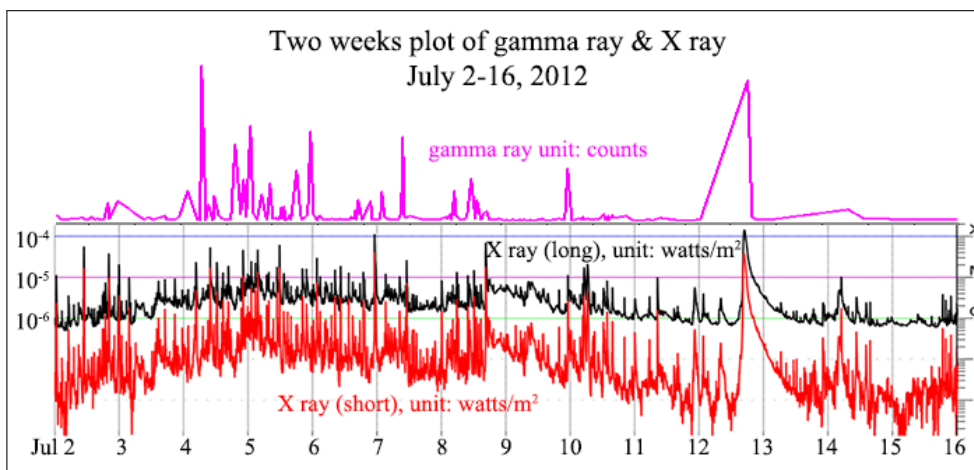

Figure 11. Two weeks plot of gamma ray & X ray from July 2 to 16, 2012.

Figure 12a shows a Gamma ray chart of $n^* = 5128$ flares, and $n = 104$ flares remaining after the threshold of 86 million counts. The most powerful gamma ray was released in March 7, 2012 with nearly 1.5 billion counts, the sun was brightened by 1000 times, and became the brightest object in the gamma ray sky. The top three events are circled in the chart. Figure 12b shows a histogram of $n^* = 5128$ flares. We are interested in the 99% quantile, $x_{0.99}$, such that 99% gamma ray released from solar flares are under this value, or equivalently, with a 1% possibility, the amount of gamma ray a solar flare releases would be in excess of this value. During the spring and fall, the satellites that are used to detect solar flares experience eclipses, in which the Earth or the Moon blocks between the satellites and the Sun for a short period every day. Eclipse season lasts for about 45 to 60 days and ranges from minutes to just over an hour. The quantile estimation would provide useful predictions for these times. $\widehat{x}_{0.99}$ is approximately located in the plot since we do not know this value yet.
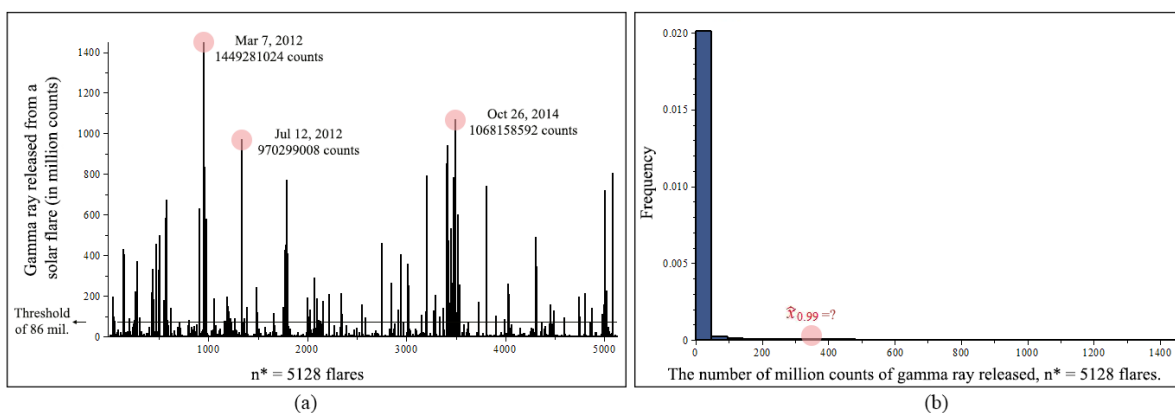

Figure 12. Gamma ray original data from November 2008 to April 2017, $n^* = 5182$, (**a**) Gamma ray released V.S solar flare occurred. After the threshold of 86 million counts, $n = 104$ flares remaining. (**b**) Histogram of gamma ray released from solar flares.

We chose the threshold as the mean of the data from the peak period. The solar cycle is every 11.6 years, and the sun's activity peaked from 2011 to 2014. In Figure 12a we can see that the top 3 flares, in fact, almost 90% of the top 100 flares, are from the 2011 to 2014 time period. Taking the average of all the X rated and significant M rated flares from this peak period, we obtained a mean of 86 million counts, resulting in a remaining sample size of $n = 104$.

For the Gamma ray of solar flare example, our goal is to find out the high quantiles, specifically, the 5% *VaR* and 1% *VaR* of the amount of gamma ray a solar flare would release, and their 95% confidence intervals.

5.2.1. Gooness-of-Fit Tests

Similar as Flu in Canada Example, we set $\mu = 86 \ million$, and obtain $\widehat{\lambda}_{MLE} = 171.0708592$, and $\widehat{\gamma}_{MLE} = 0.2580384847$. Figure 13a is a log-log plot of gamma ray data under *GPD* model, with the horizontal axis $\ln(x)$ against the vertical axis $\ln(P\{x < X\})$. Figure 13b shows the histogram fits the GPD model.
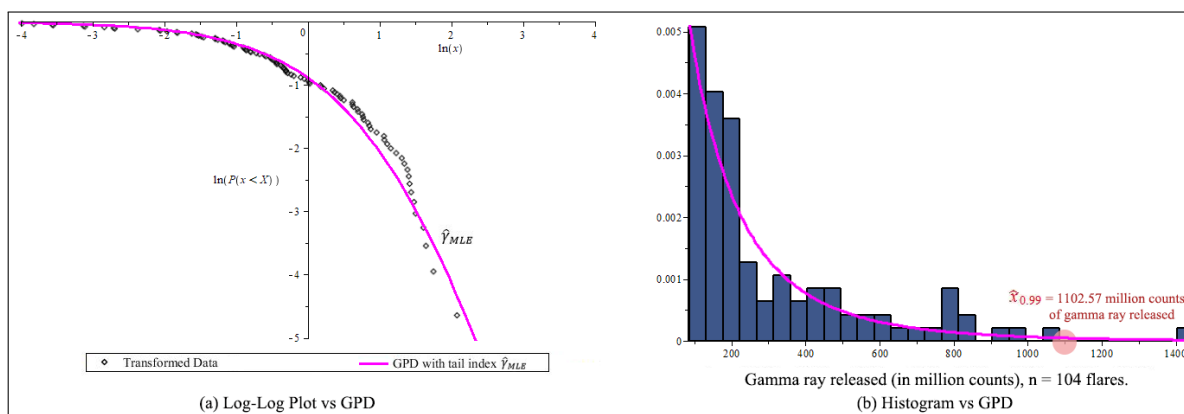

Figure 13. After threshold 86 millions count, transformation data, $n = 104$, (**a**) Log-log plot of gamma ray from solar flare example. (**b**) The Estimate GPD and the 99% high quantile of the distribution of gamma ray released by solar flare.

Next, we will perform three goodness-of-fit tests: Kolmogorov-Smirnov test, the Anderson-Darling test and the Cramér-von-Mises test. The results listed in Table 10, the data fits the *GPD* with $\widehat{\gamma}_{MLE}$ the best, nearly 59%.

Table 10. Compare the goodness-of-fit tests under the *GPD* model for the gamma ray data.

	Goodness-of-Fit Tests					
	K-S Test		*A-D* Test		*C-v-M* Test	
	Test Statistics	*p*-Value	Test Statistics	*p*-Value	Test Statistics	*p*-Value
$\widehat{\gamma}_{MLE}$	0.0697	0.5750	0.7276	0.5362	0.0991	0.5893

In Table 11, all the errors are less than 0.07 for *AE*, and less than 0.01 for *IE*.

Table 11. *AE* and *IE* under the *GPD* model for the gamma ray data using $\widehat{\gamma}_{MLE}$.

	Absolute Errors (*AE*)			Integrated Errors (*IE*)		
	*r*th Highest Gamma Ray Released			*r*th Highest Gamma Ray Released		
	r = 11	*r* = 53	*r* = 104	*r* = 11	*r* = 53	*r* = 104
$\widehat{\gamma}_{MLE}$	0.0359	0.0697	0.0697	0.0062	0.0092	0.0089

Next, we can compare the four high quantile estimators and their confidence intervals of this example.

5.2.2. Compare Four Estimation Methods

Similar as Example 1, we use the four quantile estimators in Table 1: $\ln q_H$, $\ln Q_H$, $\ln Q_{\overline{H}}$, and the $\ln \widehat{Q}_{new,\overline{H}}$.

We use $\widehat{\rho}_\tau(k)$ and $\widehat{\beta}_{\widehat{\rho}_0}(k)$, and $\tau = 0$, thus we have $\widehat{\rho}_0(k1) = -0.7269$ and $\widehat{\beta}_{\widehat{\rho}_0}(k_1) = 1.0257$, where k_1 is the optimal k value for the second-order parameters. The results are in Figure 14.
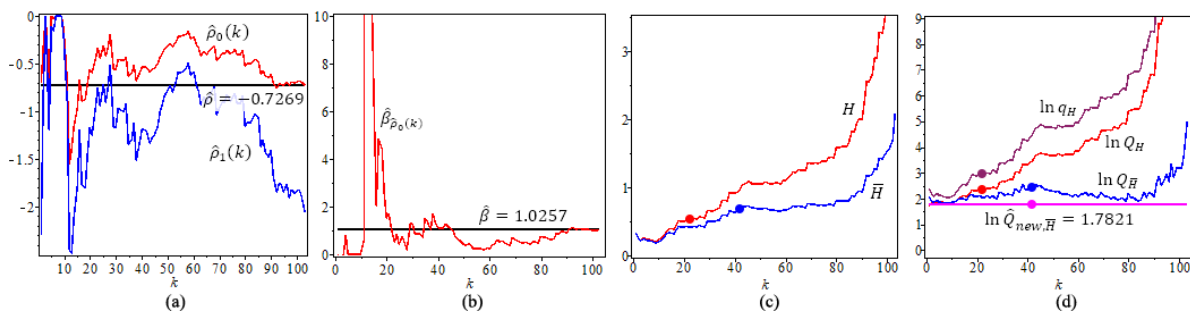

Figure 14. For gamma ray of solar flare example, $n = 104$, (**a**) Estimates of the second-order parameters $\widehat{\rho}$ and $\widehat{\rho}_\tau(k)$, $\tau = 0$, (**b**) Estimates $\widehat{\beta}$ and $\widehat{\beta}_{\widehat{\rho}_0}(k)$. (**c**) Tail index estimators, H, $\overline{H}$. (**d**) ln-quantile estimators, $p = 0.01$. The solid circles "•" in the plot are the values of the quantile estimators at their optimal k level.

Figure 14a shows the estimates of the second-order parameters $\widehat{\rho}$ and $\widehat{\rho}_\tau(k)$, $\tau = 0$. Figure 14b shows $\widehat{\beta}$ and $\widehat{\beta}_{\widehat{\rho}_0}(k)$. Figure 14c shows the two different tail index estimators, H, $\overline{H}$. We have $H = 0.5324$ at its optimal level with $\widehat{k}_0 = 21$, $\overline{H} = 0.6517$ at its optimal level with $\widehat{k}_{01} = 41$. Figure 14d shows all four quantile estimators of gamma ray example, with $p = 0.01$. We note that $\ln \widehat{Q}_{new,\overline{H}}$ has a constant value which does not depend on k.

Figure 15 compares the confidence intervals of our ln-quantile estimators in (7), (12) and (19). This figure shows that the new quantile estimator $\ln \widehat{Q}_{new,,\overline{H}}$ has the smallest confidence interval with length 1.4451, where we use $\widehat{\alpha} = \widehat{\rho} = -0.7269$. The solid circles "•" in the plot are the values of the quantile estimators at their optimal k level.
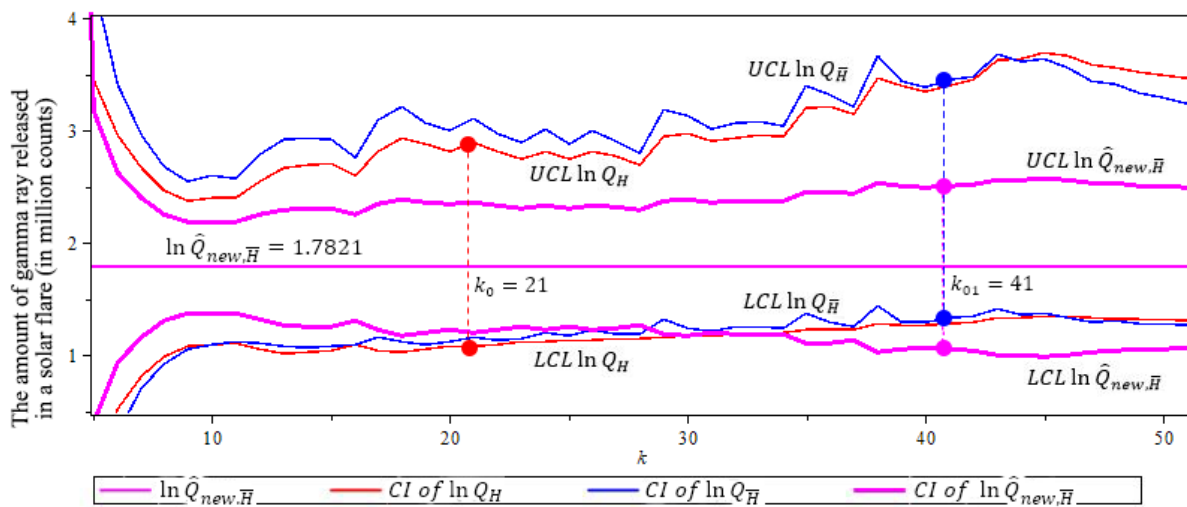

Figure 15. 95% confidence interval of three ln-quantile estimators after threshold of 86 million counts for the gamma ray example. $n = 104$, $p = 0.01$. Note that $\ln \widehat{Q}_{new,\overline{H}}$ (purple) has shortest CI with length 1.4451. (The solid circles "•" in the plot are the values of the quantile estimators at their optimal k level).

In Table 12, we compare all four quantile estimators under $VaR_{0.05}$ and $VaR_{0.01}$. Table 13 compares the size of confidence intervals of $\ln VaR_{0.01}$ and $VaR_{0.01}$ by three quantile estimators.

Table 12. Estimated $VaR_{0.05}$ and $VaR_{0.01}$ in the gamma ray example. (Unit: million counts).

Estimation Method	$\widehat{\alpha}$	$\widehat{\gamma}$	Mean	Median	$VaR_{0.05}$	$VaR_{0.01}$
$\ln Q_H$	N/A	$H = 0.5324$	451.82	315.27	867.12	1926.04
$\ln Q_{\overline{H}}$	N/A	$\overline{H} = 0.6517$	577.22	232.28	742.01	1958.67
$\ln \widehat{Q}_{new,\overline{H}}$	$\widehat{\rho} = -0.7269$	$\overline{H} = 0.6517$	577.22	189.35	441.60	1102.57

Table 13. The 95% confidence interval of $\ln VaR_{0.01}t$.and $VaR_{0.01}$.

Estimation Method	k	LCL	$\ln VaR_{0.01}$ ($VaR_{0.01}$)	UCL	Length	EFF
$\ln Q_H$	$\widehat{k}_0 = 21$	1.0807	2.3755	2.8864	1.8057	1
(Q_H)		(590.12)	(1926.04)	(3153.18)	(2563.06)	(1)
$\ln Q_{\overline{H}}$	$\widehat{k}_{01} = 41$	1.3367	2.3930	3.4494	2.1128	0.8547
$(Q_{\overline{H}})$		(737.14)	(1958.67)	(5471.71)	(4734.56)	(0.5414)
$\ln Q_{new,\overline{H}}$	$\widehat{k}_{01} = 41$	1.0595	1.7821	2.5047	1.4451	1.2495
$(Q_{new,\overline{H}})$		(579.55)	(1102.57)	(2179.85)	(1600.30)	(1.6016)

Table 13 shows that the new estimator has the shortest confidence interval, compared to $\ln Q_H$, and $\ln Q_{\overline{H}}$, with the highest efficiency of 1.6016.

5.2.3. Summary

Based on Figure 15 and Table 13, we conclude that the new estimator $\ln \widehat{Q}_{new,\overline{H}}$ in (19) is the best estimator for Gamma Ray example. We predict that $VaR_{0.01}$ is a gamma ray release of 1102.57 million counts, this is most likely an X rated solar flare. This is shown in Figure 13b.

6. Conclusions

Based on the studies in this paper, we conclude that:

1. High quantile and its CI estimation provides important information for risk management and for extreme event predictions.

2. Based on the theoretical and simulation results, the proposed new method for estimating confidence interval of high quantiles has advantages properties comparing with other existing methods. The estimation is consistent and stable with less error. The proposed method provides a useful computational algorithm to the readers.

3. The confidence interval of high quantile obtained by the new proposed method also has the highest efficiency compared to the existing methods, in terms of having the smallest size of confidence interval, and the highest probability coverage of the true quantile values in most cases.

4. Based on the analysis of the two real-world examples, flu in Canada and gamma ray from the solar flare, we can see that the new proposed method can be applied to many more fields, including other extreme events such as insurance claims, natural disasters, stock market predictions and pandemic disease monitoring.

Author Contributions: The authors M.L.H. and X.R.-Y. carried this work and drafted the manuscript together. All authors have read and agreed to the published version of the manuscript.

Appendix A. Proofs of Theorems 1 and 2

Lemma 1. *The sum of* $Cov\left[\ln \overline{Q}_H^{(p)}(i), \ln \overline{Q}_H^{(p)}(j)\right], i \neq j, i, j = 1, ..., n-1,$ *satisfy the inequality*

$$\sum_{i \neq j}^{n-1} \sum^{n-1} Cov\left[\ln \overline{Q}_H^{(p)}(i), \ln \overline{Q}_H^{(p)}(j)\right] \leq w \sum_{i \neq j}^{n-1} \sum^{n-1} \sqrt{\left(Var\left(\ln \overline{Q}_H^{(p)}(i)\right) Var\left(\ln \overline{Q}_H^{(p)}(j)\right)\right)}, \tag{A1}$$

where w is a weight

$$w = \max_{i \neq j} \rho_{ij}^+, \quad 0 \leq w \leq 1, \quad where \ \rho_{ij}^+ = \left|\rho_{ij}\right|, \quad 0 \leq \rho_{ij}^+ \leq 1;$$

ρ_{ij} *is correlation coefficient of* $\ln \overline{Q}_H^{(p)}(i)$ *and* $\ln \overline{Q}_H^{(p)}(j)$

$$\rho_{ij} = \frac{Cov\left[\ln \overline{Q}_H^{(p)}(i), \ln \overline{Q}_H^{(p)}(j)\right]}{\sqrt{\left(Var\left(\ln \overline{Q}_H^{(p)}(i)\right) Var\left(\ln \overline{Q}_H^{(p)}(j)\right)\right)}}, \quad i \neq j, \ i, j = 1, ..., n-1, \quad -1 \leq \rho_{ij} \leq 1.$$

Proof of Lemma 1. For each pair $(i, j), i \neq j, i, j = 1, ..., n-1$

$$Cov\left[\ln \overline{Q}_H^{(p)}(i), \ln \overline{Q}_H^{(p)}(j)\right] = \rho_{ij} \sqrt{\left(Var\left(\ln \overline{Q}_H^{(p)}(i)\right) Var\left(\ln \overline{Q}_H^{(p)}(j)\right)\right)}$$

then we have a boundary if we only add the positive ρ_{ij}^+ *terms in the following summation*

$$\begin{aligned}
\sum_{i \neq j}^{n-1} \sum^{n-1} Cov\left[\ln \overline{Q}_H^{(p)}(i), \ln \overline{Q}_H^{(p)}(j)\right] &= \sum_{i \neq j}^{n-1} \sum^{n-1} \rho_{ij} \sqrt{\left(Var\left(\ln \overline{Q}_H^{(p)}(i)\right) Var\left(\ln \overline{Q}_H^{(p)}(j)\right)\right)} \\
&\leq \sum_{i \neq j}^{n-1} \sum^{n-1} \rho_{ij}^+ \sqrt{\left(Var\left(\ln \overline{Q}_H^{(p)}(i)\right) Var\left(\ln \overline{Q}_H^{(p)}(j)\right)\right)} \\
&\leq w \sum_{i \neq j}^{n-1} \sum^{n-1} \sqrt{\left(Var\left(\ln \overline{Q}_H^{(p)}(i)\right) Var\left(\ln \overline{Q}_H^{(p)}(j)\right)\right)}. \quad \square
\end{aligned}$$

Lemma 2. *Under conditions (C1) and (C2), for* $\ln \overline{Q}_H^{(p)}(k)$ *in (14) by use Theorem 5.1, formula (5.2) in Gomes and Pestana (2007, p.285 [17]), as* $n \to \infty$,

$$\frac{\sqrt{k}}{\ln\left(\frac{k}{np}\right)} \left(\ln \overline{Q}_H^{(p)}(k) - \ln VaR_p\right) \xrightarrow[n \to \infty]{d} Normal\left(0, \gamma^2\right),$$

then the asymptotic expected value and variance are

$$E\left(\ln \overline{Q}_{\overline{H}}^{(p)}(k) - \ln VaR_p\right) \approx 0, \quad as\ n \to \infty;$$

$$Var\left(\ln \overline{Q}_{\overline{H}}^{(p)}(k)\right) \approx \gamma^2 \frac{\left(\ln\left(\frac{k}{np}\right)\right)^2}{k}, \quad as\ n \to \infty. \tag{A2}$$

Proof of Theorem 1. Under conditions (C1) and (C2), in the Hall-Welsh class of models in (6), where $\overline{H}$ is in (8) with conditions

$$\sqrt{k}\left(\overline{H}_{\beta,\rho}(k) - \gamma\right) \xrightarrow[n\to\infty]{d} Normal(0,\gamma),$$

and

$$\overline{H}_{\widehat{\beta},\widehat{\rho}}(k) \xrightarrow[n\to\infty]{d} = \gamma + \frac{\gamma}{\sqrt{k}}V_k + o_p\left(A\left(\frac{n}{k}\right)\right),$$

where V_k is an asymptotic standard normal random variable.

By Schwartz inequality and Lemma 1 formula (A1), sinece α is a contant in (19), based on asympototic properties of $C_p(k;\widehat{\beta},\widehat{\rho})$ in (13) (Gomes and Pestana, 2007, p.286 [17]), we have that

$$
\begin{aligned}
& Var\left(\ln \widehat{Q}_{New,\overline{H}}^{(p)}\right) \\
=\ & Var\left[\left(\frac{1}{n-1}\sum_{k=1}^{n-1}\ln \overline{Q}_{\overline{H}}^{(p)}(k)\right)\right] \\
=\ & \frac{1}{(n-1)^2}\left\{\sum_{k=1}^{n-1}Var\left[\left(\ln \overline{Q}_{\overline{H}}^{(p)}(k)\right)\right] + \sum_{i\neq j}^{n-1}\sum_{}^{n-1}Cov\left[\ln \overline{Q}_{\overline{H}}^{(p)}(i), \ln \overline{Q}_{\overline{H}}^{(p)}(j)\right]\right\} \\
=\ & \frac{1}{(n-1)^2}\left\{\sum_{k=1}^{n-1}Var\left[\left(\ln \overline{Q}_{\overline{H}}^{(p)}(k)\right)\right] + \sum_{i\neq j}^{n-1}\sum_{}^{n-1}\rho_{ij}\sqrt{\left(Var\left(\ln \overline{Q}_{\overline{H}}^{(p)}(i)\right)Var\left(\ln \overline{Q}_{\overline{H}}^{(p)}(j)\right)\right)}\right\} \\
\leq\ & \frac{1}{(n-1)^2}\left\{\sum_{k=1}^{n-1}Var\left[\left(\ln \overline{Q}_{\overline{H}}^{(p)}(k)\right)\right] + w\sum_{i\neq j}^{n-1}\sum_{}^{n-1}\sqrt{\left(Var\left(\ln \overline{Q}_{\overline{H}}^{(p)}(i)\right)Var\left(\ln \overline{Q}_{\overline{H}}^{(p)}(j)\right)\right)}\right\}.
\end{aligned}
$$

Therefore when n is large enough, use Lemma 2, formula (A2),

$$Var\left(\ln \overline{Q}_{\overline{H}}^{(p)}(k)\right) \approx \gamma^2 \frac{\left(\ln\left(\frac{k}{np}\right)\right)^2}{k}, \quad as\ n \to \infty;$$

and

$$E\left(\ln \widehat{Q}_{New,\overline{H}}^{(p)} - \ln VaR_p\right) \approx 0, \quad as\ n \to \infty.$$

we can have the following approximate relation

$$
\begin{aligned}
Var\left(\ln \widehat{Q}_{New,\overline{H}}^{(p)}\right) \leq\ & \frac{1}{(n-1)^2}\left[\sum_{k=1}^{n-1}\gamma^2\frac{\left[\ln\left(\frac{k}{np}\right)\right]^2}{k} + w\sum_{i\neq j}^{n-1}\sum_{}^{n-1}\gamma^2\sqrt{\frac{\left[\ln\left(\frac{i}{np}\right)\right]^2}{i}}\sqrt{\frac{\left(\ln\left(\frac{j}{np}\right)\right)^2}{j}}\right] \\
=\ & \frac{\gamma^2}{(n-1)^2}\left[\sum_{k=1}^{n-1}\frac{\left[\ln\left(\frac{k}{np}\right)\right]^2}{k} + w\sum_{i\neq j}^{n-1}\sum_{}^{n-1}\left[\frac{\ln\left(\frac{i}{np}\right)\ln\left(\frac{j}{np}\right)}{\sqrt{i \bullet j}}\right]\right], \quad n \to \infty.
\end{aligned}
$$

this proved (21). Therefore, we have the asymptotic normal distribution in (20)

$$\left(\ln \widehat{Q}_{New,\overline{H}}^{(p)} - \ln VaR_p\right) \xrightarrow[n\to\infty]{d} Normal\left(0, \frac{\gamma^2}{(n-1)^2}\left[\sum_{k=1}^{n-1}\frac{\left[\ln\frac{k}{np}\right]^2}{k} + w\sum_{i<j}^{n-1}\sum_{}^{n-1}\left[\frac{\ln\left(\frac{i}{np}\right)\ln\left(\frac{j}{np}\right)}{\sqrt{i \bullet j}}\right]\right]\right). \tag{A3}$$

Furthermore, we use (21) and (A2), we obtain (22) as

$$
\begin{aligned}
EFF\left(\ln \widehat{Q}_{New,\overline{H}}^{(p)}\right)_{\ln \widehat{Q}_{\overline{H}}^{(p)}(k)} &= \frac{Var\left(\ln \overline{Q}_{\overline{H}}^{(p)}(k)\right)}{Var\left(\ln \widehat{Q}_{New,\overline{H}}^{(p)}\right)} \\
&\geq \frac{(n-1)^2 \frac{\left(\ln\left(\frac{k}{np}\right)\right)^2}{k}}{\sum_{k=1}^{n-1}\frac{[\ln\left(\frac{k}{np}\right)]^2}{k} + w\sum^{n-1}\sum^{n-1}\left[\frac{\ln\left(\frac{i}{np}\right)\ln\left(\frac{j}{np}\right)}{\sqrt{i\bullet j}}\right]}, \quad n \to \infty. \quad \square
\end{aligned}
$$

Proof of Theorem 2. Under conditions (C1) and (C2), Use $Var\left(\ln \widehat{Q}_{New,\overline{H}}^{(p)}\right)$ in Theorem 1, Formula (22), and $z_{\alpha/2} = -z_{1-\alpha/2}$, with

$$
b3 = \frac{z_{1-\alpha/2}}{n-1}\sqrt{\sum_{k=1}^{n-1}\frac{[\ln\left(\frac{k}{np}\right)]^2}{k} + w\sum_{i\neq j}^{n-1}\sum^{n-1}\left[\frac{\ln\left(\frac{i}{np}\right)\ln\left(\frac{j}{np}\right)}{\sqrt{i\bullet j}}\right]}, \quad UCL_{\overline{H}}(k) = \frac{\overline{H}}{1 - \frac{z_{1-\alpha/2}}{\sqrt{k}}}.
$$

we have

$$
Z = \frac{\sqrt{n-1}}{\gamma\sqrt{\sum_{k=1}^{n-1}\frac{[\ln\left(\frac{k}{np}\right)]^2}{k} + w\sum_{i\neq j}^{n-1}\sum^{n-1}\left[\frac{\ln\left(\frac{i}{np}\right)\ln\left(\frac{j}{np}\right)}{\sqrt{i\bullet j}}\right]}}\left(\ln \widehat{Q}_{New,\overline{H}}^{(p)} - \ln VaR_p\right) \xrightarrow[n\to\infty]{d} Normal(0,1),
$$

therefore approximately

$$
P\left\{-z_{1-\alpha/2} \leq \frac{n-1}{\gamma\sqrt{\sum_{k=1}^{n-1}\frac{[\ln\left(\frac{k}{np}\right)]^2}{k} + w\sum_{i\neq j}^{n-1}\sum^{n-1}\left[\frac{\ln\left(\frac{i}{np}\right)\ln\left(\frac{j}{np}\right)}{\sqrt{i\bullet j}}\right]}}\left(\ln \widehat{Q}_{New,\overline{H}}^{(p)} - \ln VaR_p\right) \leq z_{1-\alpha/2}\right\} = 1 - \alpha,
$$

then

$$
P\left\{-\gamma b3 \leq \left(\ln \widehat{Q}_{New,\overline{H}}^{(p)} - \ln VaR_p\right) \leq \gamma b3\right\} = 1 - \alpha
$$

$$
P\left\{-\ln \widehat{Q}_{New,\overline{H}}^{(p)} - \gamma b3 \leq -\ln VaR_p \leq -\ln \widehat{Q}_{New,\overline{H}}^{(p)} + \gamma b3\right\} = 1 - \alpha,
$$

and using $\widehat{\gamma} = \overline{H}(k) \leq UCL_{\overline{H}}(k)$, to guarantee a coverage probability of at least $(1 - \alpha)100\%$, we have

$$
P\left\{\ln \widehat{Q}_{New,\overline{H}}^{(p)} - UCL_{\overline{H}}(k)b3 \leq \ln VaR_p \leq \ln \widehat{Q}_{New,\overline{H}}^{(p)} + UCL_{\overline{H}}(k)b3\right\} \geq 1 - \alpha. \quad \square
$$

References

1. Fisher, R.A.; Tippett, L.H.C. Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Procs. Camb. Philos. Soc.* **1928**, *24*, 180–190. [CrossRef]
2. de Haan, L.D.; Ferreira, A. *Extreme Value Theory*; Springer: New York, NY, USA, 2006.
3. Dekkers, A.L.M.; de Haan, L. On the estimation of the extreme value index and large quantile estimation. *Ann. Stat.* **1989**, *17*, 1795–1832. [CrossRef]
4. Hill, B.M. A simple general approach to inference about the tail of a distribution. *Ann. Statist.* **1975**, *3*, 1163–1174. [CrossRef]
5. Weissman, I. Estimation of parameters and large quantiles based on the k largest observations. *J. Am. Stat. Assoc.* **1978**, *73*, 812–815.

6. Beirlant, J.; Vynckier, P.; Teugels, J.L. Tail index estimation, Pareto quantile plots and regression diagnostics. *J. Am. Statist. Assoc.* **1996**, *91*, 1659–1667.
7. Dreea, H.; Kaufmann, E. Selecting the optimal sample fraction in univaraiate extreme values estimation. *Stoch. Proc. Appl.* **1998**, *75*, 149–172.
8. Guillou, A.; Hall, P.G. A diagnostic for selecting the threshold in extreme value analysis. *J. R. Stat. Soc. B* **2001**, *63*, 293–305. [CrossRef]
9. Gomes, M.I.; Oliveira, O. The bootstrap methodology in statistics of extremes: choice odf the optimal sample fraction. *Extremes* **2001**, *4*, 331–358. [CrossRef]
10. Hall, P.; Welsh, A.H. Adaptive estimates of parameters of regular variation. *Ann. Stat.* **1985**, *13*, 331–341. [CrossRef]
11. Peng, L. Asymptotic unbiased estimator for extreme-value index. *Stat. Probab. Lett.* **1998**, *38*, 107–115. [CrossRef]
12. Beirlant, J.; Dierckx, G.; Goegebeur, Y.; Matthys, G. Tail index estimation and an exponential regression model. *Extremes* **1999**, *2*, 177–200. [CrossRef]
13. Feuerverger, A.; Hall, P.G. Estimating a tail exponent by modelling departure from a Pareto. *Ann. Stat.* **1999**, *27*, 760–781.
14. Gomes, M.I.; Martins, M.J.; Neves, M. Alternatives to a semi-parametric estimator of parametric of rare events-the jackknife methodology. *Extremes* **2000**, *3*, 207–229. [CrossRef]
15. Caeiro, F.; Gomes, M.I. A class of asymptotically unbiased semi-parametric estimations of the tail index. *Test* **2002**, *11*, 345–364. [CrossRef]
16. Gomes, M.I.; Figueiredo, F.; Mendonca, S. Asymptotically best linear unbiased tail estimators under second order regular variation. *J. Stat. Plan. Inference* **2004**, *134*, 409–433. [CrossRef]
17. Gomes, M.I.; Pestana, D. A study reduced-bias extreme quantile (VaR) estimator. *J. Am. Assoc.* **2007**, *102*, 280–292. [CrossRef]
18. Caeiro, F.; Gomes, M.I.; Pestana, D. Direct reduction of bias of the classical Hill estimator. *Rev. Stat.* **2005**, *3*, 113–136.
19. Lekina, A.; Chebana, F.; Ouarda, T.B.M.J. Weighted estimate of extreme quantile: An application to the estimation of high flood return periods. *Stoch. Environ. Res. Risk Assess.* **2014**, *28*, 147–165. [CrossRef]
20. Lekina, A. Estimation Non-Paramétrique des Quantiles Extrêmes Conditionnels. Ph.D. Thesis, Université de Grenoble, Saint-Martin-d'Hères, France, 2010.
21. Huang, M.L. *A New High Quantile Estimator for Heavy Tailed Distributions;* (Working Paper); Department of Mathematics, Brock University: St. Catharines, ON, Canada, 2011.
22. Fréchet, M. Sur la loi de probabilite de l'écart maximum. *Ann. Soc. Pol. Math.* **1927**, *6*, 93–116.
23. de Zea, Bermudeza, P.; Kotz, S. Parameter estimation of the generalized Pareto distribution. *J. Stat. Inference* **2010**, *140*, 1374–1388. [CrossRef]
24. Bickel, P.J.; Docksum, K.A. *Mathematical Statistics, Basic Ideas and Selected Topic*, 2nd ed.; CRC Press, Taylor & Frances Group: Boca Raton, FL, USA, 2015; Volume 1.
25. Pickands, J. Statistical inference using extreme order statistics. *Ann. Stat.* **1975**, *3*, 119–131.
26. Balkema, A.A.; de Haan, L. Residual life time at great age. *Ann. Prob.* **1974**, *2*, 792–804. [CrossRef]
27. Scarrott, G.; McDonald, A. A Review of extreme value threshold estimation and uncertainty quantification. *Revstat* **2012**, *10*, 33–60.
28. World Health Organization (WHO). 2020. Available online: https://www.who.int/influenza/gisrs_laboratory/flunet/en (accessed on 31 December 2020).
29. Kolmogorov, A.N. Sulla determinazione empirica di una legge di distribuzione. *G. Dell. Istituto Ital. Degli Attuari* **1933**, *4*, 83–91.
30. Anderson, T.W.; Darling, D.A. Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *Ann. Math. Stat.* **1952**, *23*, 193–212. [CrossRef]
31. National Aeronautics and Space Administration (NASA). Gamma Ray. 2020. Available online: http://hesperia.gsfc.nasa.gov/fermi/gbm/qlook/fermi_gbm_flare_list.txt (accessed on 31 December 2020).
32. National Weather Service (NOAA). Space Weather Prediction Center. 2020. Available online: https://satdat.ngdc.noaa.gov/sem/goes/data/plots (accessed on 31 December 2020).