Research Article

A Cluster Truncated Pareto Distribution and Its Applications

Mei Ling Huang,¹ Vincenzo Coia,² and Percy Brill³

¹ Department of Mathematics, Brock University, St. Catharines, ON, Canada L2S 3A1

² Department of Statistics, University of British Columbia, Vancouver, BC, Canada V6T 1Z4

³ Department of Mathematics & Statistics, University of Windsor, Windsor, ON, Canada N9B 3P4

Correspondence should be addressed to Mei Ling Huang; mhuang@brocku.ca

Received 20 June 2013; Accepted 25 July 2013

Academic Editors: E. Omev and S. Sagitov

Copyright © 2013 Mei Ling Huang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The Pareto distribution is a heavy-tailed distribution with many applications in the real world. The tail of the distribution is important, but the threshold of the distribution is difficult to determine in some situations. In this paper we consider two real-world examples with heavy-tailed observations, which leads us to propose a mixture truncated Pareto distribution (MTPD) and study its properties. We construct a cluster truncated Pareto distribution (CTPD) by using a two-point slope technique to estimate the MTPD from a random sample. We apply the MTPD and CTPD to the two examples and compare the proposed method with existing estimation methods. The results of log-log plots and goodness-of-fit tests show that the MTPD and the cluster estimation method produce very good fitting distributions with real-world data.

1. Introduction

There are many real-world problems modelled as heavy-tailed distributions, especially the Pareto distribution [1, 2]. However, there are some difficulties in estimation of Pareto distributions. First, the Pareto distribution has infinite moments in some heavy-tailed cases. Therefore the moment estimation method for the shape parameter cannot be used in these situations. It is a loss for the estimation process since the moment estimator is a robust estimator. Several authors suggest using a truncated Pareto distribution which always has finite moments (e.g., [3–5]).

In some situations, data will behave differently within different thresholds. For example, losses from hurricane damage can be classified into small, medium, and large hurricane groups. The data in these classes may have different distributions, or by grouping, data with self-similarity may have the same kind of distribution but with different parameters. A cluster method for data is needed to determine these groups when dealing with real data sets. In this paper we study an example of 49 most damaging Atlantic hurricanes occurring between years 1900 and 2005 [6]. The costs are standardized to 2005 USD; see Figure 1.

Coia and Huang [7] applied Pareto and truncated Pareto models to fit the hurricane data set. The maximum

likelihood estimator (MLE) and the moment estimator for the shape parameter were used. The results are shown in a log-log plot in Figure 2. Coia and Huang [7] also used Kolmogorov-Smirnov, Anderson-Darling, and Cramer-von-Mises goodness-of-fit tests. We note that the two estimated (by MLE and moment method) truncated Pareto curves fit the data set quite well; they fit much better in the tail than the original Pareto distribution (which is in a straight line). But the truncated Pareto curves do not fit the data uniformly well, especially for the middle value data. We observed that the pattern of data can be classified into three groups. The data in these three groups may still be Pareto distributed but with different shape parameters. In the literature, researchers study similar data sets by using cluster methods; for example, Coia and Huang [7] proposed a sieve model.

In this paper, we propose a more generalized method—mixture truncated Pareto distribution (MTPD)—in Section 2. We study the properties of the MTPD in Section 3. In Section 4, we propose a cluster method by using a two-point slope technique to estimate the MTPD from data which utilizes a cluster truncated Pareto distribution (CTPD). In Section 5, we review the nonparametric kernel density estimation method. In Section 6, we analyze the hurricane data and a second example regarding sizes of fish caught off the Atlantic shore of Massachusetts, USA, by using

FIGURE 1: The 49 costliest Atlantic hurricanes between the years 1900 and 2005.

the CTPD, nonparametric kernel estimation method and three other existing semiparametric estimation methods in log-log plots (see Figure 3 in Section 6). We also perform Kolmogorov-Smirnov, Anderson Darling, and Cramer-von Mises goodness-of-fit tests on these two data sets. The results show that the proposed cluster method and nonparametric method are superior to other existing estimation methods, in both examples.

2. Mixture Truncated Pareto Distribution

Definition 1. The probability density function (p.d.f.) and the cumulative distribution function (c.d.f.) of a random variable Y having the Pareto distribution are given by

$$f_P(y; \gamma, \alpha) = \frac{\alpha \gamma^\alpha}{y^{(\alpha+1)}}, \quad 0 < \gamma \leq y < \infty, \alpha > 0, \quad (1)$$

$$F_P(y; \gamma, \alpha) = 1 - \left(\frac{\gamma}{y}\right)^\alpha, \quad 0 < \gamma \leq y < \infty, \alpha > 0, \quad (2)$$

where α is the shape parameter.

When $0 < \alpha \leq 1$, which is a heavy-tailed case, the mean and variance of Y are infinite, and the distribution is heavier in the right tail as α decreases.

The truncated Pareto distribution (TPD) was originally used to describe the distribution of oil fields by size. It has a lower limit γ , an upper limit ν , and a shape parameter α . In fact, it has been shown that the truncated Pareto distribution fits better than the nontruncated Pareto distribution for some positively skewed populations [3].

Definition 2. The p.d.f. and c.d.f. of a random variable X having the truncated Pareto distribution are given by

$$f(x; \gamma, \nu, \alpha) = \frac{\alpha \gamma^\alpha x^{-\alpha-1}}{1 - (\gamma/\nu)^\alpha}, \quad 0 < \gamma \leq x \leq \nu < \infty, \alpha > 0, \quad (3)$$

$$F(x; \gamma, \nu, \alpha) = 1 - \frac{\gamma^\alpha (x^{-\alpha} - \nu^{-\alpha})}{1 - (\gamma/\nu)^\alpha}, \quad (4)$$

$$0 < \gamma \leq x \leq \nu < \infty, \alpha > 0,$$

where γ and ν are the left and right truncation points.

FIGURE 2: Log-log plot of hurricane example with estimated distribution curves. The red circles are the data; the black straight line is the original Pareto distribution; the green dot line is the MLE estimated truncated Pareto distribution; the brown dash line is the moment estimated truncated Pareto distribution.

The quantile function of the truncated Pareto distribution is

$$F^{-1}(u) = \left(\frac{1-u}{\gamma^\alpha} + \frac{u}{\nu^\alpha} \right)^{-1/\alpha}, \quad 0 \leq u \leq 1, \alpha > 0. \quad (5)$$

The mean, second moment, and variance of X are, respectively,

$$\mu = \frac{\alpha \gamma^\alpha (\gamma^{1-\alpha} - \nu^{1-\alpha})}{(\alpha-1)(1 - (\gamma/\nu)^\alpha)}, \quad 0 < \gamma < \nu < \infty, \alpha > 1, \quad (6)$$

$$\mu_{(2)} = \frac{\alpha \gamma^\alpha (\gamma^{2-\alpha} - \nu^{2-\alpha})}{(2-\alpha)(1 - (\gamma/\nu)^\alpha)^2}, \quad 0 < \gamma < \nu < \infty, \alpha > 2, \quad (7)$$

$$\sigma^2 = \frac{\alpha \gamma^\alpha (\gamma^{2-\alpha} - \nu^{2-\alpha})}{(2-\alpha)(1 - (\gamma/\nu)^\alpha)^2} - \frac{\alpha^2 \gamma^{2\alpha} (\gamma^{1-\alpha} - \nu^{1-\alpha})^2}{(1-\alpha)^2 (1 - (\gamma/\nu)^\alpha)^2}, \quad (8)$$

$$0 < \gamma < \nu < \infty, \alpha > 2.$$

We consider a vector of thresholds

$$\mathbf{T} = (t_0, t_1, \dots, t_k)^T,$$

$$\text{where } 0 < a = t_0 < t_1 \cdots < t_k = b < \infty, \quad (9)$$

$$a, b \in \mathfrak{R}, k = 1, 2, \dots$$

Consider a vector $\mathbf{\Lambda} = (\alpha_1, \alpha_2, \dots, \alpha_k)^T$, $\alpha_i > 0$, $i = 1, \dots, k$. We define a mixture truncated Pareto distribution as follows.

Definition 3. The c.d.f. of a random variable X having a mixture truncated Pareto distribution (MTPD) is given by

$$F_{\text{MTP}}(x; \mathbf{T}, \mathbf{A}; \mathbf{W}) = \sum_{i=1}^k w_i F_i(x; t_{i-1}, t_i, \alpha_i), \quad (10)$$

$$0 < a \leq x \leq b < \infty, \quad a = t_0, \quad b = t_k,$$

where $F_i(x; t_{i-1}, t_i, \alpha_i)$ is the c.d.f. of the truncated Pareto distribution in (4), the truncation points t_{i-1}, t_i , are related to thresholds $\mathbf{T} = (t_0, t_1, \dots, t_k)^T$, and $\mathbf{W}$ is a vector of weights

$$\mathbf{W} = (w_1, w_2, \dots, w_k)^T, \quad 0 \leq w_i \leq 1, \quad \sum_{i=1}^k w_i = 1. \quad (11)$$

The p.d.f. of a mixture truncated Pareto distribution is given by

$$f_{\text{MTP}}(x; \mathbf{T}, \mathbf{A}; \mathbf{W}) = \sum_{i=1}^k w_i f_i(x; t_{i-1}, t_i, \alpha_i), \quad (12)$$

$$0 < a \leq x \leq b < \infty, \quad a = t_0, \quad b = t_k,$$

where $f_i(x; t_{i-1}, t_i, \alpha_i)$ is the p.d.f. of the truncated Pareto distribution in (3).

3. Properties of Mixture Truncated Pareto Distribution

Proposition 4. The mean μ_{MTP} and variance Var_{MTP} of a mixture truncated Pareto distribution are given by

$$\mu_{\text{MTP}} = \sum_{i=1}^k w_i \mu_i, \quad (13)$$

$$\text{Var}_{\text{MTP}} = \sum_{i=1}^k w_i \mu_{(2)i} - \left(\sum_{i=1}^k w_i \mu_i \right)^2,$$

where μ_i and $\mu_{(2)i}$ are the mean and the second moment of the $f_i(x; t_{i-1}, t_i, \alpha_i)$ in (6) and (7), $i = 1, 2, \dots, n$.

Proof.

$$\begin{aligned} \mu_{\text{MTP}} &= \int_{\gamma}^{\nu} x \sum_{i=1}^k w_i f_i(x; t_{i-1}, t_i, \alpha_i) dx \\ &= \sum_{i=1}^k w_i \int_{t_{i-1}}^{t_i} x f_i(x; t_{i-1}, t_i, \alpha_i) dx \\ &= \sum_{i=1}^k w_i \mu_i, \\ \mu_{(2)\text{MTP}} &= \int_{\gamma}^{\nu} x^2 \sum_{i=1}^k w_i f_i(x; t_{i-1}, t_i, \alpha_i) dx \\ &= \sum_{i=1}^k w_i \int_{t_{i-1}}^{t_i} x^2 f_i(x; t_{i-1}, t_i, \alpha_i) dx \\ &= \sum_{i=1}^k w_i \mu_{(2)i}. \end{aligned} \quad (14)$$

□

Proposition 5. For given p , $0 < p < 1$, the quantile function of a mixture truncated Pareto distribution in (10) is given by solving for x in the following equation:

$$\sum_{i=1}^k w_i F_i(x; t_{i-1}, t_i, \alpha_i) = p, \quad 0 < a \leq x \leq b < \infty. \quad (15)$$

4. A Cluster Truncated Pareto Distribution Estimator

Consider a random sample $X_1, X_2, \dots, X_n$ from the MTPD in (10), and let $X_{1,n} \leq X_{2,n} \leq \dots \leq X_{n,n}$ denote its order statistics. We divide data into k clusters by the domains (t_i, t_{i+1}) , $i = 0, 1, \dots, k-1$, $\mathbf{T} = (t_0, t_1, \dots, t_k)^T$, where $0 < a = t_0 < t_1 < \dots < t_k = b < \infty$, $a, b \in \mathfrak{R}$. We define a cluster truncated Pareto distribution (CTPD) as an estimator of the MTPD.

Definition 6. The c.d.f. of a random variable X having the cluster truncated Pareto distribution (CTPD) is given by

$$F_C(x; \mathbf{T}, \mathbf{A}; \mathbf{W}) = \sum_{i=1}^k \left(\frac{n_i}{n} \right) F_i(x; t_{i-1}, t_i, \alpha_i), \quad (16)$$

$$0 < a \leq x \leq b < \infty,$$

where $F_C(x; \mathbf{T}, \mathbf{A}; \mathbf{W})$ is a c.d.f. of the MTPD in (10) and n_i is the sample size in the i th cluster in the i th domain: (t_{i-1}, t_i) :

$$\mathbf{W} = \left(\frac{n_1}{n}, \frac{n_2}{n}, \dots, \frac{n_k}{n} \right)^T, \quad w_i = \frac{n_i}{n}, \quad i = 1, 2, \dots, k, \quad (17)$$

where n_i 's depend on the vector $\mathbf{C} = (c_0, c_1, \dots, c_k)^T$, where $0 = c_0 < c_1 < \dots < c_k = n$, c_i is the number of data less than or equal to the threshold t_i . Note that c_i is a function of t_i and the random sample $(X_1, X_2, \dots, X_n)$. Thus

$$c_i(t_i; X_1, X_2, \dots, X_n) = \sum_{j=1}^n I_{(-\infty, t_i]}(X_j), \quad i = 1, 2, \dots, k-1,$$

$$n_i = c_i - c_{i-1}, \quad i = 1, 2, \dots, k, \quad (18)$$

where I_A is the indicator function of set A .

We also note that c_i and n_i all depend on t_i , $i = 1, \dots, k$ in (16). The key point of applying the CTPD in (16) is to determine $\mathbf{T} = (t_0, t_1, \dots, t_k)^T$ from the random sample. In this paper, we propose a two-point slope technique in the log-log plot to estimate thresholds $\mathbf{T} = (t_0, t_1, \dots, t_k)^T$.

Definition 7. A two-point slope is defined as

$$S_i(X_1, X_2, \dots, X_n) = \begin{cases} \frac{\log(1 - (i+1)/n) - \log(1 - i/n)}{\log(X_{i+1,n}) - \log(X_{i,n})}, \\ \log(X_{i+1,n}) - \log(X_{i,n}) \neq 0, \quad i = 1, \dots, n-1; \\ 0, \quad \log(X_{i+1,n}) - \log(X_{i,n}) = 0, \quad i = 1, \dots, n-1. \end{cases} \quad (19)$$

Box 1: Construction of cluster truncated Pareto distribution from the data.

Then we construct $(n-1)$ order statistics $S_{1,n-1} \leq S_{2,n-1} \leq \dots \leq S_{n-1,n-1}$ from the absolute values of the two-point slopes $|S_i(X_1, X_2, \dots, X_n)|$, $i = 1, 2, \dots, n-1$. The cluster threshold points can be estimated by $\hat{t}_1(X_1, X_2, \dots, X_n), \dots, \hat{t}_{k-1}(X_1, X_2, \dots, X_n)$ which are determined by the $(k-1)$ largest absolute values of the two-point slopes

$$S_{n-k+1,n-1} \leq S_{n-k+2,n-1} \leq \dots \leq S_{n-1,n-1}, \quad (20)$$

where k depends upon empirical observations of differences between successive $S_{i,n-1}$'s. This usually occurs when $|S_{n-k+2,n-1} - S_{n-k+1,n-1}|$ is large compared with previous differences.

We propose seven steps to construct a cluster truncated Pareto distribution in (16).

Step 1. Compute $(n-1)$ two-point slopes $S_i(X_1, X_2, \dots, X_n)$ in (19), $i = 1, \dots, n-1$.

Step 2. Determine k by using (20); there are two main factors:

- (1) Determining k depends upon empirical observations of differences between successive $S_{i,n-1}$'s, when $|S_{n-k+2,n-1} - S_{n-k+1,n-1}|$ is much larger than the previous difference $|S_{n-k+1,n-1} - S_{n-k,n-1}|$. (This technique is used on the two examples in Section 6.)
- (2) We also ensure that the sample size n_i within each group is sufficiently large (usually $n_i \geq 5$).

Step 3. Find the $(k-1)$ estimated threshold points $\hat{t}_1, \dots, \hat{t}_{k-1}$ by using the values of the $(k-1)$ largest absolute slopes of the order statistics of $|S_i(X_1, X_2, \dots, X_n)|$ in (20), $i = 1, \dots, n-1$, corresponding to the $(k-1)$ values $\{X_1^*, X_2^*, \dots, X_{k-1}^*\}$ of the original sample, which now have been ordered as new order statistics

$$X_{1,k-1}^* \leq X_{2,k-1}^* \leq \dots \leq X_{k-1,k-1}^*. \quad (21)$$

Then we let

$$\begin{aligned} \hat{t}_i(X_1, X_2, \dots, X_n) &= X_{i,k-1}^*, \quad i = 1, \dots, k-1, \\ \hat{t}_0 &= X_{1,n} = a, \quad \hat{t}_k = X_{n,n} = b. \end{aligned} \quad (22)$$

Step 4. Determine $\mathbf{C} = (c_0, c_1, \dots, c_k)^T$, where $0 = c_0 < c_1 < \dots < c_k = n$, $c_i(t_i; X_1, X_2, \dots, X_n) = \sum_{j=1}^n I_{(-\infty, t_i]}(X_j)$. Thus

$$\begin{aligned} n_i &= c_i - c_{i-1}, \quad i = 1, 2, \dots, k, \\ \hat{t}_i &= X_{c_i,n}, \quad i = 1, \dots, k-1, \end{aligned} \quad (23)$$

(note that we replace $X_{c_i,n} = X_{i,k-1}^*$ in (22)),

$$\hat{t}_0 = X_{1,n} = a, \quad \hat{t}_k = X_{n,n} = b.$$

Then we have k clusters:

$$\{a = \hat{t}_0, \dots, X_{c_1,n}\}, \quad (24)$$

$$\{X_{c_2,n}, \dots, X_{c_3,n}\}, \dots, \{X_{c_{k-1},n}, \dots, \hat{t}_k = X_{n,n} = b\}.$$

This construction is shown in Box 1.

Step 5. Construct $\widetilde{F}_C(x; \widehat{\mathbf{T}}, \widehat{\Lambda}; \widehat{\mathbf{W}}) = \sum_{i=1}^k (n_i/n) F_i(x; \hat{t}_{i-1}, \hat{t}_i, \alpha_i)$, in (16).

Step 6. Estimate $\hat{\alpha}_{i,\text{Moment}}$. We suggest using the moment estimator $\hat{\alpha}_M$ in (26) since it has robust properties, but there are other estimators available in (25) and (27).

Step 7. Construct an estimator $\widehat{F}_C(x; \widehat{\mathbf{T}}, \widehat{\Lambda}; \widehat{\mathbf{W}}) = \sum_{i=1}^k (n_i/n) F_i(x; \hat{t}_{i-1}, \hat{t}_i, \hat{\alpha}_i)$, for (16).

Remark 8. There are three estimation methods for the shape parameters α_i for all sub-samples, given by the following.

(1) *Hill Estimator (original Pareto distribution in (1)).* The Hill [8] MLE $\hat{\alpha}_{\text{Hill}}$ for α is defined as

$$\hat{\alpha}_{\text{Hill}} = \left[r^{-1} \sum_{i=1}^r \{\ln X_{n-i+1,n} - \ln X_{n-r,n}\} \right]^{-1}, \quad (25)$$

where $X_{i,n}$ is the i th smallest order statistic and r is the cut-off point.

(2) *Moment Estimator (truncated Pareto distribution in (3)).* A moment estimator $\hat{\alpha}_M$ for α can be obtained by solving the following equation:

$$\frac{1}{n} \sum_{i=1}^n X_i = \frac{\hat{\alpha}_M \gamma^{\hat{\alpha}_M} (\gamma^{1-\hat{\alpha}_M} - \gamma^{1-\hat{\alpha}_M})}{(\hat{\alpha}_M - 1) (1 - (\gamma/\gamma)^{\hat{\alpha}_M})}, \quad (26)$$

where $0 < \gamma \leq X_i \leq \nu < \infty$, $\hat{\alpha}_M > 0$.

(3) *MLE Method (truncated Pareto distribution in (3))*. The Aban MLE $\hat{\alpha}_{\text{Aban}}$ [4] for α is obtained by solving the following equation:

$$\frac{n}{\hat{\alpha}_{\text{Aban}}} + \frac{n(\gamma/\nu)\hat{\alpha}_{\text{Aban}} \ln(\gamma/\nu)}{1 - (\gamma/\nu)^{\hat{\alpha}_{\text{Aban}}}} - \sum_{i=1}^n [\ln X_{n-i+1,n} - \ln \gamma] = 0, \quad (27)$$

where $X_{i,n}$ is the i th smallest order statistic, $\gamma \leq X_{i,n} \leq \nu$, $i = 1, 2, \dots, n$.

Note that, before using (26) and (27), the parameters γ and ν may be estimated by $\hat{\gamma} = X_{1,n}$, $\hat{\nu} = X_{n,n}$, which are used in this paper. There are other nonparametric methods in the literature. Cooke [9] proposed estimators

$$\begin{aligned} \hat{\gamma} &= 2X_{1,n} - \sum_{i=0}^{n-1} \left[\left(1 - \frac{i}{n}\right)^n - \left(1 - \frac{i+1}{n}\right)^n \right] X_{i+1,n}, \\ \hat{\nu} &= 2X_{n,n} - \sum_{i=0}^{n-1} \left[\left(1 - \frac{i}{n}\right)^n - \left(1 - \frac{i+1}{n}\right)^n \right] X_{n-i,n}. \end{aligned} \quad (28)$$

5. Nonparametric Kernel Density Estimation

We apply the kernel density estimation method (KDE) [10]. It is a smoothing technique for estimating distributions. It is well known that the classical kernel density estimator from a random sample $X_1, X_2, \dots, X_n$ for true probability density function (p.d.f.) f is

$$\hat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right), \quad x \in \mathfrak{R}, \quad (29)$$

where $K(\bullet)$ is a symmetric density function and $h > 0$ is a bandwidth.

We will compare the KDE estimator and the CTPD estimator in the next section.

6. Applications

Now we apply the cluster truncated Pareto distribution to the hurricane example in Section 1.

6.1. Cluster Method. By using 48 two-point slopes in (19) and the seven steps in Section 4, we construct $k = 3$ clusters. We select the $k - 1 = 2$ largest absolute values of the two-point slopes $S_{n-2,n-1} \leq S_{n-1,n-1}$ in (20) as

$$S_{48,48} = 14.2027, \quad S_{47,48} = 5.6574, \quad (30)$$

since the third largest two-point slope $S_{46,48} = 5.2276$ does not have a large change compared with $S_{47,48} = 5.6574$.

Then we determine $\hat{t}_i$'s and $k = 3$ groups as

$$\begin{aligned} \{a = \hat{t}_0 = X_{1,n}, \dots, X_{c_1,n}\}, \quad \{X_{c_1+1,n}, \dots, X_{c_2,n}\}, \\ \{X_{c_2+1,n}, \dots, \hat{t}_3 = X_{n,n} = b\}, \end{aligned} \quad (31)$$

where $\hat{t}_0 = X_{1,n} = 3.7 = a$, $\hat{t}_1 = X_{c_1,n} = 13.7$, $\hat{t}_2 = X_{c_2,n} = 21.29$, and $\hat{t}_3 = X_{c_3,n} = 157 = b$; $c_0 = 0$, $c_1 = 27$, $c_2 = 40$,

and $c_3 = 49$; $n_1 = 27$, $n_2 = 13$, and $n_3 = 9$; $n_1 + n_2 + n_3 = n = 49$; then we have an estimated CTPD in (16) as

$$\widehat{F}_C(x; \widehat{\mathbf{T}}, \widehat{\mathbf{A}}, \widehat{\mathbf{W}}) = \sum_{i=1}^3 \left(\frac{n_i}{n}\right) F_i(x; \widehat{t}_{i-1}, \widehat{t}_i, \widehat{\alpha}_i). \quad (32)$$

This construction is shown in Box 2.

6.2. Kernel and Other Estimation Methods in Log-Log Plot. We apply the kernel estimator in (29) which is normalized to the hurricane data. Here, we use a standard normal kernel and optimal bandwidth [10, p. 45]

$$h_{\text{opt}} = \left(\frac{4}{3n}\right)^{1/5} \approx 1.05922n^{-1/5}. \quad (33)$$

We ensure that the bandwidth is sufficiently large such that the estimated tail distribution is smooth enough.

Table 1 gives $\hat{\alpha}$, $\hat{\mu}$, Median, 5% Value-at-Risk (VaR), and 1% VaR of each of four estimation methods. We note that the cluster method and kernel method give the largest medians. The cluster method gives the smallest VaRs.

Figure 3 is a log-log plot which exhibits data and five estimated distribution curves. We note that the original Pareto distribution does not fit data well in the right tail. The moment and Aban estimated truncated Pareto distribution (TPD) fit data well in the right tail, but not so well in the smaller or middle values data. The cluster truncated Pareto distribution overcomes this problem and has the best fit to the data over the whole range. Figure 3 suggests a single distribution may not totally represent how natural data is distributed. We may consider grouping data by using the cluster method. Figure 3 also shows that the nonparametric kernel estimated distribution fits the data well.

Figure 3 provides a visual observation. It is necessary to run goodness-of-fit tests mathematically to confirm which estimated distribution best fits the hurricane data.

6.3. Goodness-of-Fit Tests. In this section we conduct three goodness-of-fit tests, Kolmogorov-Smirnov, Anderson Darling, and Cramer-von Mises. All three tests are based on the distance between the empirical distribution function and the proposed distribution function: original Pareto distribution in (1) or truncated Pareto distribution in (3) or mixture truncated Pareto distribution in (10).

Each test considers the same null and alternative hypothesis:

$$H_0 : F(x) = F^*(x) \quad \text{versus} \quad H_1 : F(x) \neq F^*(x), \quad (34)$$

where $F(x)$ is the unknown true distribution of the sample data and $F^*(x)$ is one of our proposed four estimated distributions:

- (1) Pareto distribution in (1) with Hill estimator $\hat{\alpha}_{\text{Hill}}$ in (25);
- (2) truncated Pareto distribution (TPD) in (3) with Aban estimator $\hat{\alpha}_{\text{Aban}}$ in (27);

TABLE 1: Five estimation methods for hurricane example.

Method	$\hat{\alpha}$	$\hat{\mu}$	Median	5% VaR	1% VaR
Pareto _(Hill)	0.8126	∞	8.68 billion	147.68 billion	1070.30 billion
TPD _(Aban)	0.6206	21.10 billion	9.73 billion	85.15 billion	136.17 billion
TPD _(moment)	0.6476	20.48 billion	9.47 billion	82.55 billion	134.90 billion
Kernel	—	19.88 billion	12.88 billion	80.58 billion	156.14 billion
Cluster	$\hat{\alpha}_1 = 0.6476$	27.65 billion	11.19 billion	77.24 billion	132.41 billion
	$\hat{\alpha}_2 = 5.6498$				
	$\hat{\alpha}_3 = 0.8416$				

$c_0 = 0$	$c_1 = 27$	$c_2 = 40$	$c_3 = n = 49$
----- $n_1 = 27$ -----	----- $n_2 = 13$ -----	----- $n_3 = 9$ -----	-----
$\hat{t}_0 = 3.7$	$\hat{t}_1 = 13.7$	$\hat{t}_2 = 21.29$	$\hat{t}_3 = 157$
$= X_{1,n}$	$= X_{c_1,n}$	$= X_{c_2,n}$	$= X_{c_3,n}$
$= a$			$= b$

Box 2: Construction of cluster truncated Pareto distribution from the data.

FIGURE 3: Log-log plot of hurricane example with estimated distribution curves. The red circles are the data; the black straight line is the original Pareto distribution; the green dot line is the MLE estimated truncated Pareto distribution; the brown dash line is the moment estimated truncated Pareto distribution; the thick green dash line is the kernel estimated distribution; the thick blue line is the cluster truncated Pareto distribution.

(3) truncated Pareto distribution in (TPD) (3) with moment estimator $\hat{\alpha}_M$ in (26);

(4) kernel estimated distribution in (29);

(5) cluster truncated Pareto distribution in (16) with moment estimator $\hat{\alpha}_{M(i)}$ in (26).

We ran a test for each estimated distribution as $F^*(x)$.

(1) The Kolmogorov-Smirnov (K-S) Test [11]. The test statistic is given by

$$T = \sup_x |F^*(x) - S_n(x)|, \quad -\infty < x < \infty, \quad (35)$$

where $S_n(x) = (1/n) \sum_{i=1}^n I_{(-\infty, x]}(X_i)$ is the empirical distribution function. Under H_0 the two-tailed P value for the K-S test is as follows:

P value

$$= 1 - \left[1 - T \sum_{j=0}^{[n(1-T)]} \frac{n!}{j!(n-j)!} \left(1 - T - \frac{j}{n}\right)^{n-j} \left(T + \frac{j}{n}\right)^{j-1} \right]^2, \quad (36)$$

where $[n(1-T)]$ is the integer part of $n(1-T)$.

(2) Anderson and Darling Test (A-D Test). Anderson and Darling [12] introduced a measure of “distance” between the empirical distribution $S_n(x)$ and the proposed c.d.f. $F^*(x)$ by using a metric function space,

$$W_n^2 = n \int_{-\infty}^{\infty} [S_n(x) - F^*(x)]^2 \psi(F^*(x)) dF^*, \quad (37)$$

where $\psi(u) = 1/(u(1-u))$, with $u = F^*(x)$. Let $u_i = F^*(X_{i,n})$; assume under H_0 the test statistic and P value are given by

$$W_n^2 = -n - \frac{1}{n} \sum_{j=1}^n (2j-1) (\log(u_j) + \log(1 - u_{n-j+1})),$$

P value

$$= 1 - \frac{\sqrt{2}}{z} \sum_{j=0}^{\infty} \left\{ \frac{(-1)^j \Gamma(j+1/2) (4j+1)}{j!} e^{-(4j+1)^2 \pi^2 / 8z} \cdot \int_0^{\infty} e^{(z/8/(w^2+1) - (4j+1)^2 \pi^2 w^2 / 8z)} dw \right\}, \quad (38)$$

where z is the observed value of W_n^2 and $\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt$, $x > 0$, is the Gamma function.

TABLE 2: Goodness-of-fit tests $n = 49$ for hurricane example.

Method	K-S test		Goodness-of-fit tests A-D test		C-v-M test	
	Test statistic	P value	Test statistic	P value	Test statistic	P value
Pareto _(Hill)	0.1340	0.2900	2.7141	0.0383	0.2057	0.2568
TPD _(Aban)	0.0948	0.6282	2.3126	0.0622	0.0964	0.6030
TPD _(moment)	0.1053	0.5308	2.3672	0.0582	0.1095	0.5402
Kernel	0.0890 ⁽²⁾	0.6805 ⁽²⁾	2.2344 ⁽¹⁾	0.0685 ⁽¹⁾	0.0427 ⁽¹⁾	0.9187 ⁽¹⁾
Cluster	0.0741 ⁽¹⁾	0.8035 ⁽¹⁾	2.2584 ⁽²⁾	0.0665 ⁽²⁾	0.0467 ⁽²⁾	0.8958 ⁽²⁾

Note. In this paper, we use (1) to denote the best values and (2) to denote the second-best values in the tables.

TABLE 3: Errors of goodness-of-fit tests $n = 49$ for hurricane example.

Method	Goodness-of-fit tests Absolute error (AE)			Integrated error (IE)		
	$r = 49$	$r = 18$	$r = 10$	$r = 49$	$r = 18$	$r = 10$
Pareto _(Hill)	0.1340	0.0584	0.0584	0.0032	0.0027	0.0027
TPD _(Aban)	0.0948	0.0839	0.0832	0.0020	0.0018	0.0016
TPD _(moment)	0.1053	0.0738	0.0737	0.0019	0.0015	0.0013
Kernel	0.0890 ⁽²⁾	0.0271 ⁽¹⁾	0.0138 ⁽¹⁾	0.0009 ⁽¹⁾	0.0007 ⁽¹⁾	0.0007 ⁽¹⁾
Cluster	0.0741 ⁽¹⁾	0.0510 ⁽²⁾	0.0278 ⁽²⁾	0.0011 ⁽²⁾	0.0010 ⁽²⁾	0.0009 ⁽²⁾

(3) *Cramer-von Mises Test (C-v-M Test)*. Anderson and Darling [12] proposed this test by using $\psi(u) = 1$ in (37). Thus under H_0 the test statistic and P value are given by

$$n\omega^2 = \frac{1}{12n} + \sum_{j=1}^n \left(u_j - \frac{2j-1}{2n} \right)^2,$$

$$P \text{ value} = 1 - \frac{1}{\pi \sqrt{z}} \sum_{j=0}^{\infty} \left\{ \frac{(-1)^{2j} \Gamma(j+1/2) (4j+1)^{1/2}}{j! \sqrt{\pi}} e^{-(4j+1)^2/16z} \cdot K_{1/4} \left(\frac{(4j+1)^2}{16z} \right) \right\}, \quad (39)$$

where $K_{1/4}(x)$ is the modified Bessel function of the second kind,

$$K_{\alpha}(x) = \frac{\pi}{2} \frac{I_{-\alpha}(x) - I_{\alpha}(x)}{\sin(\alpha\pi)},$$

$$I_{\alpha}(x) = \sum_{m=0}^{\infty} \frac{1}{m! \Gamma(m+\alpha+1)} \left(\frac{x}{2} \right)^{2m+\alpha}. \quad (40)$$

Table 2 gives the values of the test statistics and P values of three goodness-of-fit tests. Note that the cluster truncated Pareto distribution has the smallest test statistics in the K-S test (i.e., the smallest errors) and the largest P values. The kernel estimated distribution gives the smallest test statistics in the A-D test and C-v-M test, respectively. This means the cluster truncated Pareto distribution and kernel estimated distribution have the best fit to the hurricane data.

In Table 3, we took the r largest data in the sample. The absolute error and integrated error are defined by

$$AE = \sup_x |F^*(x) - S_n(x)|, \quad -\infty < x < \infty,$$

$$IE = \frac{1}{(X_{n,n} - X_{n-r+1,n})} \left[\int_{X_{n-r+1,n}}^{X_{n,n}} (S_n(x) - F^*(x))^2 dx \right]^{1/2}. \quad (41)$$

Table 3 gives absolute errors and integrated errors of the five estimation methods in $r = 49, 18, 10$ cases. We note that the cluster truncated Pareto distribution has the smallest errors in all 6 cases. This means the cluster method is superior in fitting the hurricane data compared with the other existing methods.

6.4. Fishing Example. Another example is determining the chance of catching a fish of record length (size). Overfishing is a serious issue that has been known to collapse many fish populations. To help control the population, limits are set upon anglers to determine the size of fish which can be kept.

The data is from a fishing trip, May 29–June 3, 2011, to Buzzard's Bay of the Cape Cod area in Massachusetts, USA, by Coia [13]. The largest black sea bass lengths were measured out of a total of 326 black sea bass caught on the trip, using 43.0 cm as a lower-limit threshold. The threshold of 43.0 cm was chosen to be conservative. When sampling, locations producing smaller fish were avoided and the largest fish were targeted. A time-series plot in Figure 4 shows the lengths of these largest 72 fish in the order of which they were caught.

By using 71 two-point slopes in (19) and the seven steps in Section 4, we construct $k = 4$ clusters. We select

$c_0 = 0$	$c_1 = 46$	$c_2 = 59$	$c_3 = 65$	$c_4 = n = 72$
$n_1 = 46$	$n_2 = 13$	$n_3 = 6$	$n_4 = 7$	
$\hat{t}_0 = 43.0$	$\hat{t}_1 = 46.0$	$\hat{t}_2 = 46.6$	$\hat{t}_3 = 47.9$	$\hat{t}_4 = 56.8$
$= X_{1,n}$	$= X_{c_1,n}$	$= X_{c_2,n}$	$= X_{c_3,n}$	$= X_{n,n}$
$= a$				$= b$

Box 3: Construction of cluster truncated Pareto distribution from the data.

FIGURE 4: Lengths of 72 black sea bass (*Centropristis striata*) caught at least 43 cm long, fishing in Buzzard's Bay, May 29–June 3, 2011.

the $(k-1) = 3$ largest absolute values of the two-point slopes $S_{n-3,n-1} \leq S_{n-2,n-1} \leq S_{n-1,n-1}$ in (20) as

$$S_{71,71} = 88.6994, \quad S_{70,71} = 63.8948, \quad S_{69,71} = 36.3351, \quad (42)$$

since the fourth largest two-point slope $S_{68,71} = 35.3847$ does not have a large change compared with $S_{69,71} = 36.3351$.

Then we determine $\hat{t}_i$'s and $k = 4$ groups as

$$\begin{aligned} \{a = \hat{t}_0, \dots, X_{c_1,n}\}, \quad \{X_{c_1+1,n}, \dots, X_{c_2,n}\}, \\ \{X_{c_2+1,n}, \dots, X_{c_3,n}\}, \quad \{X_{c_3+1,n}, \dots, \hat{t}_n = b\}, \end{aligned} \quad (43)$$

where $\hat{t}_0 = 3.7 = a$, $\hat{t}_1 = X_{c_1,n} = 13.7$, $\hat{t}_2 = X_{c_2,n} = 21.29$, $\hat{t}_3 = X_{c_3,n} = 157$, and $\hat{t}_4 = b$; $c_0 = 0$, $c_1 = 46$, $c_2 = 59$, $c_3 = 65$, and $c_4 = 72$; $n_1 = 46$, $n_2 = 13$, $n_3 = 6$, and $n_4 = 7$; $n_1 + n_2 + n_3 + n_4 = n = 72$; then we have an estimated CTPD in (16) as

$$\hat{F}_c(x; \hat{\mathbf{T}}, \hat{\mathbf{A}}, \hat{\mathbf{W}}) = \sum_{i=1}^4 \left(\frac{n_i}{n} \right) F_i(x; \hat{t}_{i-1}, \hat{t}_i, \hat{\alpha}_i). \quad (44)$$

This construction is shown in Box 3.

We also apply the kernel estimator in (29) and bandwidth in (33) to the fish data. Figure 5 is a log-log plot which exhibits data and five estimated distribution curves. We note that the original Pareto distribution does not fit the data well in the right tail. The moment and Aban estimated truncated Pareto distribution (TPD) fit the data well in the right tail, but not so well in the smaller or middle values data. The cluster truncated Pareto distribution overcomes this problem and has the best fit to the data over the whole range. Figure 5 suggests a single distribution may not totally represent how natural data

FIGURE 5: Log-log plot of fishing example with estimated distribution curves. The red circles are the data; the black straight line is the original Pareto distribution; the green dot line is the MLE estimated truncated Pareto distribution; the brown dash line is the moment estimated truncated Pareto distribution; the thick green dash line is the kernel estimated distribution; the thick blue line is the cluster Pareto distribution.

is distributed. We group the data by using the cluster method. Figure 5 also shows that the nonparametric kernel estimated distribution fits the data well.

Table 4 gave $\hat{\alpha}$, $\hat{\mu}$, Median, 5% VaR, and 1% VaR of each of four estimation methods. We note that the cluster method and kernel method give the largest median and the smallest VaRs.

We also ran goodness-of-fit tests mathematically to confirm which estimated distribution best fits the fish data.

Table 5 gives the values of the test statistics and P value of each of three goodness-of-fit tests. We note that the cluster truncated Pareto distribution has the smallest test statistics (i.e., the smallest errors) and the largest P values in the K-S test. The kernel estimated distribution has the smallest test statistics in the A-D test and C-v-M test, respectively. Thus the cluster truncated Pareto distribution and kernel estimated distribution fit best to the fish data.

Table 6 gives absolute errors and integrated errors of five estimation methods in $r = 72, 50, 10$ cases. We note that the cluster truncated Pareto distribution and the kernel estimated

TABLE 4: Five estimation methods for fishing example.

Method	$\hat{\alpha}$	$\hat{\mu}$	Median	5% VaR	1% VaR
Pareto _(Hill)	16.3582	45.6433	44.8612	51.6419	56.9812
TPD _(Aban)	15.6493	45.7437	44.9108	51.3511	54.7664
TPD _(moment)	15.7782	45.7250	44.8961	51.2993	54.7273
Kernel	—	45.7484	45.4767	49.3010	54.5943
Cluster	$\hat{\alpha}_1 = 2.7781$	43.4179	45.2801	50.3432	54.0029
	$\hat{\alpha}_2 = 162.6216$				
	$\hat{\alpha}_3 = 64.5429$				
	$\hat{\alpha}_4 = 17.0537$				

TABLE 5: Goodness-of-fit tests $n = 72$ for fishing example.

Method	Goodness-of-fit tests					
	K-S test		A-D test		C-v-M test	
	Test statistic	P value	Test statistic	P value	Test statistic	P value
Pareto _(Hill)	0.1496	0.0702	2.8568	0.0324	0.4956	0.0409
TPD _(Aban)	0.1407	0.1021	2.8413	0.0330	0.4442	0.0554
TPD _(moment)	0.1434	0.0916	2.8672	0.0320	0.4560	0.0516
Kernel	0.0697 ⁽²⁾	0.7240 ⁽²⁾	0.7034 ⁽¹⁾	0.5560 ⁽¹⁾	0.0379 ⁽¹⁾	0.9438 ⁽¹⁾
Cluster	0.0687 ⁽¹⁾	0.7348 ⁽¹⁾	0.8528 ⁽²⁾	0.4445 ⁽²⁾	0.0590 ⁽²⁾	0.8202 ⁽²⁾

TABLE 6: Errors of goodness-of-fit tests $n = 72$ for fishing example.

Method	Goodness-of-fit tests					
	Absolute error (AE)			Integrated error (IE)		
	$r = 72$	$r = 50$	$r = 10$	$r = 72$	$r = 50$	$r = 10$
Pareto _(Hill)	0.1496	0.1496	0.0739	0.0154	0.0148	0.0109
TPD _(Aban)	0.1407	0.1407	0.0770	0.0149	0.0144	0.0108
TPD _(moment)	0.1434	0.1434	0.0747	0.0149	0.0143	0.0105
Kernel	0.0697 ⁽²⁾	0.0561 ⁽¹⁾	0.0179 ⁽¹⁾	0.0040 ⁽¹⁾	0.0035 ⁽¹⁾	0.0022 ⁽¹⁾
Cluster	0.0687 ⁽¹⁾	0.0687 ⁽²⁾	0.0332 ⁽²⁾	0.0055 ⁽²⁾	0.0055 ⁽²⁾	0.0046 ⁽²⁾

distribution have the smallest errors in all 6 cases. The cluster method and the kernel estimation method are superior in fitting the fish data compared with the other existing semi-parametric estimation methods.

The k very much depends on the empirical observations. We displayed this technique in the two examples in Section 6. We plan to do further studies on more complex data sets in the future.

7. Conclusions

Overall, after the studies in this paper, we may conclude the following.

- (1) Truncated Pareto models are useful for analyzing real-world data.
- (2) Cluster truncated Pareto models are useful for grouped data.
- (3) The two-point slope technique is innovative and useful for determining the threshold points in clustering.
- (4) A nonparametric kernel estimated distribution fits heavy-tailed data very well.
- (5) It is a difficult problem to determine k —the number of groups which is based on the two-point slope in (20) in Step 2 of the construction of the CTPD in (16).

Acknowledgment

The authors thank the referees and the editor for their comments which helped to improve the paper. This research is supported by the Natural Sciences and Engineering Research Council of Canada.

References

- [1] C. Kleiber and S. Kotz, *Statistical Size Distributions in Economics and Actuarial Sciences*, Wiley Series in Probability and Statistics, Wiley-Interscience, Hoboken, NJ, USA, 2003.
- [2] R. Perline, "Strong, weak and false inverse power laws," *Statistical Science*, vol. 20, no. 1, pp. 68–88, 2005.
- [3] M. A. Beg, "Estimation of the tail probability of the truncated Pareto distribution," *Journal of Information & Optimization Sciences*, vol. 2, no. 2, pp. 192–198, 1981.

- [4] I. B. Aban, M. M. Meerschaert, and A. K. Panorska, "Parameter estimation for the truncated Pareto distribution," *Journal of the American Statistical Association*, vol. 101, no. 473, pp. 270–277, 2006.
- [5] M. L. Huang and K. Zhao, "On estimation of the truncated Pareto distribution," *Advances and Applications in Statistics*, vol. 16, no. 1, pp. 83–102, 2010.
- [6] R. A. Pielke, J. Gratz, C. W. Landsea, D. Collins, M. A. Saunders, and R. Musulin, "Normalized hurricane damage in the United States: 1900–2005," *Natural Hazards Review*, vol. 9, no. 1, pp. 29–42, 2008.
- [7] V. Coia and M. L. Huang, "A sieve model for extreme values," *Journal of Statistical Computation and Simulation*, vol. 83, 2013.
- [8] B. M. Hill, "A simple general approach to inference about the tail of a distribution," *The Annals of Statistics*, vol. 3, no. 5, pp. 1163–1174, 1975.
- [9] P. Cooke, "Statistical inference for bounds of random variables," *Biometrika*, vol. 66, no. 2, pp. 367–374, 1979.
- [10] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*, Monographs on Statistics and Applied Probability, Chapman & Hall, London, UK, 1986.
- [11] A. N. Kolmogorov, "Sulla determinazione empirica di una legge di distribuzione," *Giornale dell'Istituto Italiano degli Attuari*, vol. 4, pp. 83–91, 1933.
- [12] T. W. Anderson and D. A. Darling, "Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes," *Annals of Mathematical Statistics*, vol. 23, pp. 193–212, 1952.
- [13] V. Coia, "On estimation of extreme value distributions," Brock Report in Mathematics and Statistics 120809-01, Brock University, 2012.