On a Mixture Pareto Distribution

Mei Ling Huang[1]*, Justyne Mottola[1] and Percy Brill[2]†

[1]Department of Mathematics and Statistics, Brock University, Canada
[2]Department of Mathematics and Statistics, University of Windsor, Canada

September 1, 2015

Abstract

The Pareto distribution is a heavy tailed distribution with many applications. In this paper we consider a real world example with heavy tailed observations, which leads us to propose a mixture truncated Pareto distribution (MTPD) and study its properties. There are difficulties in the estimation of thresholds of the MTPD. We construct a cluster truncated Pareto distribution (CTPD) by using a two-point slope technique to estimate the MTPD from a random sample. The results of Monte Carlo simulations show that the two-point slope technique is useful for estimating thresholds. Finally, we apply the MTPD and CTPD to the example which we observed in the beginning 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 a good fitting distribution with real world data.

Keywords: Extreme value distribution, goodness-of-fit test, Hill estimator, finite mixture distribution, two-point slope, truncated Pareto distribution.

1. Introduction

There are many real world problems modelled as heavy tailed distributions, especially the Pareto distribution. 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., Beg, 1981; Aban, et al., 2006; Coia and Huang, 2014).

The aim of this paper is to explore better modelling methods for complicated heavy tailed observed real examples, like dangerous storms such as hurricanes, and floods. We would like to provide better-fit estimated distributions to these extreme weather data sets in order to prepare for the unknown future.

*Corresponding author, Address to: Mei Ling Huang, Department of Mathematics and Statistics, Brock University, St. Catharines, Ontario, Canada, L2S 3A1. E-mail: mhuang@brocku.ca.

†This research is supported by the Natural Sciences and Engineering Research Councile of Canada.

Example. Flood Damage in Canada

Floods are events that can damage homes, businesses, and crops. One never knows the extent to which the flood will damage a city or area, and when it happens the money to fix the damage needs to be available. The data set for flooding in Canada is from Environment Canada (2010), http://www.ec.gc.ca/eau-water/default.asp?lang=En&n=02A7110-1, where it reports the cost that the federal government paid to respective provinces and territories after major floods. The floods range from 1970 to 1998 in all of Canada. The data includes the top 34 flood damage costs that have been fully paid by the government ($n = 34$). The top 10 flood losses in Canada are shown in Table 1. Figure 1 shows the histogram and estimated Pareto and truncated Pareto distributions by using the maximum likelihood estimate (MLE) of the shape parameter from the flood damage costs data. Figure 2 is a log-log plot of the data and estimated Pareto distribution and truncated Pareto distribution.

Table 1. The top 10 flood damage costs in Canada, 1970-1998

Province or Territory	Year	Damage Cost ($)
Quebec	1983	17,346,772
Manitoba	1979	14,670,604
Manitoba	1974	11,464,005
Manitoba (Winnipeg River)	1993	11,291,186
Quebec	1974	8,670,477
Alberta	1990	8,229,503
Alberta	1988	7,787,911
Quebec	1976	7,582,330
British Columbia	1990	7,343,629
Alberta	1986	6,809,368

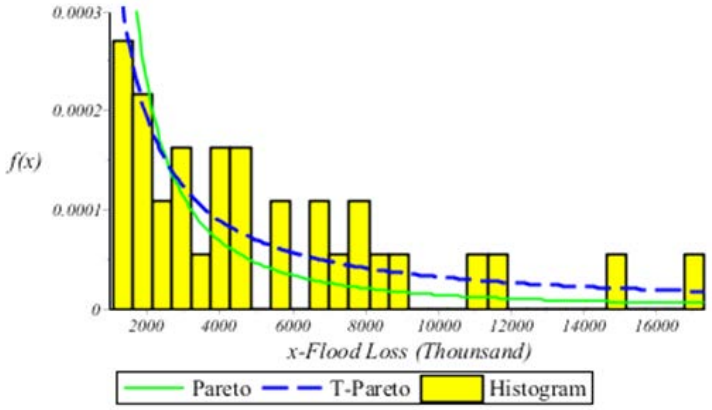

Figure 1. Histogram and fitting Pareto and truncated Pareto distributions of top 34 flood damage costs in Canada, 1970-1998. The green line is the MLE estimated original Pareto distribution; the blue dash line is the MLE estimated truncated Pareto distribution.

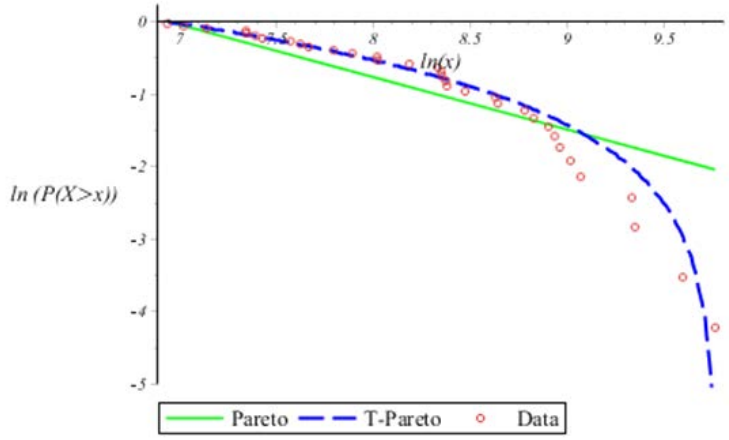

Figure 2. Log-log plots and fitting Pareto and truncated Pareto distributions of top 34 flood damage costs in Canada, 1970-1998. The red circles are the data; the green straight line is the MLE estimated original Pareto distribution; the blue dash line is the MLE estimated truncated Pareto distribution.

We apply Pareto and truncated Pareto models to fit the flood damage costs 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; at first, we note that the estimated truncated Pareto curve (blue dash line) fits the data set quite well and fits much better in the tail than the estimated original Pareto distribution (which is a straight line). But the truncated Pareto curve does not fit the data uniformly well, especially for the middle and tail data. We observed that the pattern of data can be classified into three groups, small, medium and large flood 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. 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 (2013) proposed a sieve model.

In this paper, we propose that a new estimation method of a mixture Pareto distribution will fit our data sets better since the data set seems to be separated into clusters of data points with different slopes. Through this study we hope to improve on the already known truncated Pareto distribution to be ready for costly damaged events. We will discuss the flood damage example mentioned above by the new cluster method in Section 5.

In this paper, we propose a more generalized method, mixture truncated Pareto distribution (MTPD), in Section 2. In Section 3, 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 4, the results of Monte Carlo simulations confirmed the precision of estimating group thresholds. In Section 5, we analyze the flood damage costs data by using the CTPD and two other existing semi-parametric estimation methods in log-log plots (see Figure 4 in Section 5). We also perform Kolmogorov-Smirnov, Anderson Darling, and Cramer-von Mises goodness-of-fit tests on this data set. The results show that the proposed cluster method is superior to other existing estimation methods, in this example.

2. Mixture Truncated Pareto Distribution

Definition 2.1. 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 respectively by*

$$f_P(y;\gamma,\alpha) = \frac{\alpha\gamma^\alpha}{y^{\alpha+1}}, \quad 0<\gamma\le y<\infty, \quad \alpha>0, \tag{2.1}$$

$$F_P(y;\gamma,\alpha) = 1-\left(\frac{\gamma}{y}\right)^\alpha, \quad 0<\gamma\le y<\infty, \quad \alpha>0, \tag{2.2}$$

where α is the shape parameter.

When $0<\alpha\le 1$, which is a heavy tailed case, the mean and variance of Y are infinite, and the distribution gets 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 non-truncated Pareto distribution for some positively skewed populations (Beg, 1981).

Definition 2.2. *The p.d.f. and c.d.f. of a random variable X having the truncated Pareto distribution are given respectively by*

$$f(x;\gamma,\nu,\alpha) = \frac{\alpha\gamma^\alpha x^{-\alpha-1}}{1-(\frac{\gamma}{\nu})^\alpha}, \quad 0<\gamma\le x\le\nu<\infty, \quad \alpha>0, \tag{2.3}$$

$$F(x;\gamma,\nu,\alpha) = 1-\frac{\gamma^\alpha(x^{-\alpha}-\nu^{-\alpha}),}{1-(\frac{\gamma}{\nu})^\alpha}, \quad 0<\gamma\le x\le\nu<\infty,\ \alpha>0, \tag{2.4}$$

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

The quantile function of the truncated Pareto distribution is

$$F^{-1}(u) = \left(\frac{1-u}{\gamma^\alpha}+\frac{u}{\nu^\alpha}\right)^{-\frac{1}{\alpha}}, \ \ 0\le u\le 1,\ \alpha>0. \tag{2.5}$$

The mean, second moment and variance of X are respectively, for $0<\gamma<\nu<\infty$,

$$\mu = E(X) = \begin{cases} \frac{\alpha\gamma^\alpha(\gamma^{1-\alpha}-\nu^{1-\alpha})}{(\alpha-1)(1-(\gamma/\nu)^\alpha)}, & \alpha\ne 1,\ \alpha>0; \\ \frac{\gamma\ln(\nu/\gamma)}{1-(\gamma/\nu)}, & \alpha=1, \end{cases} \tag{2.6}$$

$$\sigma^2 = Var(X) = \begin{cases} \frac{\alpha\gamma^\alpha(\gamma^{2-\alpha}-\nu^{2-\alpha})}{(\alpha-2)(1-(\gamma/\nu)^\alpha)} - \frac{\alpha^2\gamma^{2\alpha}(\gamma^{1-\alpha}-\nu^{1-\alpha})^2}{(\alpha-1)^2(1-(\gamma/\nu)^\alpha)^2}, & \alpha\ne 1,\ \alpha\ne 2,\ \alpha>0; \\ \frac{\gamma(\nu-\gamma)}{1-(\gamma/\nu)} - \frac{\gamma^2[\ln(\nu/\gamma)]^2}{[1-(\gamma/\nu)]^2}, & \alpha=1; \\ \frac{2\gamma^2\ln(\nu/\gamma)}{1-(\gamma/\nu)^2} - \frac{4\gamma^4(\gamma^{-1}-\nu^{-1})^2}{(1-(\gamma/\nu)^2)^2}, & \alpha=2. \end{cases} \tag{2.7}$$

The mth moments are, for $m=1,2,...$

$$\mu_{(m)} = E\left[X^m\right] = \begin{cases} \frac{\alpha\gamma^\alpha(\gamma^{m-\alpha}-\nu^{m-\alpha})}{(\alpha-m)(1-(\gamma/\nu)^\alpha)}, & \alpha\ne m,\ \alpha>0; \\ \frac{m\gamma^m\ln(\nu/\gamma)}{1-(\gamma/\nu)^m}, & \alpha=m. \end{cases} \tag{2.8}$$

Finite mixture distributions have been studied for complicated data (Frühwirth-Schnatter, 2006; Everitt, et al., 2011). We consider a vector of group thresholds

$$\mathbf{T} = (t_0, t_1, ..., t_k)^T, \quad \text{where} \quad 0 < a = t_0 < t_1 ... < t_k = b < \infty,\ a, b \in \Re,\ k = 1, 2, ...$$

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

Definition 2.3. *The c.d.f. of a random variable X having a mixture truncated Pareto distribution (MTPD) is given by*

$$F_{MTP}(x; \mathbf{T}, \mathbf{\Lambda}; \mathbf{W}) = \sum_{i=1}^{k} w_i F_i(x; t_{i-1}, t_i, \alpha_i),\ 0 < a \le x \le b < \infty,\ a = t_0,\ b = t_k, \qquad (2.9)$$

where $F_i(x; t_{i-1}, t_i, \alpha_i)$ is the c.d.f. of the truncated Pareto distribution in (2.4), and the truncation points t_{i-1}, t_i, are related to thresholds $\mathbf{T} = (t_0, t_1, ..., t_k)^T$, and $\mathbf{W}$ is a vector of weights

$$\mathbf{W} = (w_1, w_2, ..., w_k)^T,\ 0 < w_i \le 1,\ \sum_{i=1}^{k} w_i = 1.$$

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

$$f_{MTP}(x; \mathbf{T}, \mathbf{\Lambda}; \mathbf{W}) = \sum_{i=1}^{k} w_i f_i(x; t_{i-1}, t_i, \alpha_i),\ 0 < a \le x \le b < \infty,\ a = t_0,\ b = t_k, \qquad (2.10)$$

where $f_i(x; t_{i-1}, t_i, \alpha_i)$ is the p.d.f. of the truncated Pareto distribution in (2.3).

3. A Cluster Truncated Pareto Distribution Estimator

Consider a random sample $X_1, X_2, ..., X_n$ from the c.d.f of MTPD in *(2.9),* to estimate parameter vectors $\mathbf{T}$, $\mathbf{\Lambda}$, $\mathbf{W}$; we let $X_{1,n} \le X_{2,n} \le ... \le X_{n,n}$ denote its order statistics. We divide data into k clusters by the domains $(t_i, t_{i+1}),\ i = 0, 1, ..., k-1,\ \mathbf{T} = (t_0, t_1, ..., t_k)^T$, where $0 < a = t_0 < t_1 < ... < t_k = b < \infty,\ a, b \in \Re$. We define a cluster truncated Pareto distribution (CTPD) as an estimator of the MTPD.

Definition 3.1. *The c.d.f. of a random variable X having the cluster truncated Pareto distribution (CTPD) is given by*

$$F_C(x; \mathbf{T}, \mathbf{\Lambda}; \mathbf{W}) = \sum_{i=1}^{k} \left(\frac{n_i}{n}\right) F_i(x; t_{i-1}, t_i, \alpha_i),\ 0 < a \le x \le b < \infty, \qquad (3.1)$$

where $F_C(x; \mathbf{T}, \mathbf{\Lambda}; \mathbf{W})$ is a c.d.f. of the MTPD in (2.9), and n_i is the sample size in the ith cluster in the ith domain (t_{i-1}, t_i).

$$\mathbf{W} = \left(\frac{n_1}{n}, \frac{n_2}{n}, ..., \frac{n_k}{n}\right)^T, \quad w_i = \frac{n_i}{n}, \quad 0 < n_i \le n, \quad i = 1, 2, ..., k,$$

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

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

$$n_i = c_i - c_{i-1},\ i = 1, 2, ..., k,$$

where I_A is the indicator function of set A.

Also c_i and n_i all depend on $t_i,\ i = 1, ..., k$ in *(3.1)*. The key point of applying the CTPD in *(3.1)* is to determine $\mathbf{T} = (t_0, t_1, ..., t_k)^T$ from the random sample. Here we propose a two-point slope technique in the log-log plot to estimate thresholds $\mathbf{T} = (t_0, t_1, ..., t_k)^T$.

Definition 3.2. *A two-point slope is defined as*

$$S_{i,n-1}(X_1, X_2, ..., X_n) = \begin{cases} \frac{\log(1-\frac{i+1}{n}) - \log(1-\frac{i}{n})}{\log(X_{i+1,n}) - \log(X_{i,n})}, & \log(X_{i+1,n}) - \log(X_{i,n}) \neq 0,\ i = 1, ..., n-1; \\ 0, & \log(X_{i+1,n}) - \log(X_{i,n}) = 0,\ i = 1, ..., n-1. \end{cases} \tag{3.2}$$

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

$$S_{n-k+1,n-1} \leq S_{n-k+2,n-1} \leq ... \leq S_{n-1,n-1}, \tag{3.3}$$

where 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 large compared with previous differences.

We propose seven steps to construct a cluster truncated Pareto distribution as in *(3.1)*:

Step 1: Compute $n-1$ two-point slopes $S_{i,n-1}(X_1, X_2, ..., X_n)$ in *(3.2)*, $i = 1, ..., n-1$.

Step 2: Determine k by using *(3.3)*; 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 example in Section 5.)
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 $\widehat{t}_1, ..., \widehat{t}_{k-1}$ by using the values of the $k-1$ largest absolute slopes of the order statistics of $|S_i(X_1, X_2, ..., X_n)|$ in *(3.3)*, $i = 1, ..., n-1$, corresponding to the $k-1$ values $\left\{X_1^*, X_2^*, ..., X_{k-1}^*\right\}$ of the original sample, which now have been ordered as new order statistics

$$X_{1,k-1}^* \leq X_{2,k-1}^* \leq ... \leq X_{k-1,k-1}^*, \quad \text{then we let}$$

$$\widehat{t}_i(X_1, X_2, ..., X_n) = X^*_{i,k-1},\ i = 1, ..., k-1, \text{ and } \widehat{t}_0 = X_{1,n} = a,\ \widehat{t}_k = X_{n,n} = b. \qquad (3.4)$$

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

$$\begin{aligned} n_i &= c_i - c_{i-1},\ i = 1, 2, ..., k; \\ \widehat{t}_i &= X_{c_i,n},\ i = 1, ..., k-1, \text{ (Note: we replace } X_{c_i,n} = X^*_{i,k-1} \text{ in } (3.4)) \\ \text{and } \widehat{t}_0 &= X_{1,n} = a,\ \widehat{t}_k = X_{n,n} = b. \end{aligned}$$

Then we have k clusters:

$$\{a = \widehat{t}_0, ..., X_{c_1,n}\}, \{X_{c_2,n}, ..., X_{c_3,n}\}, ..., \{X_{c_{k-1},n}, ..., \widehat{t}_k = X_{n,n} = b\}.$$

Table 2. Construction of a cluster truncated Pareto distribution from data

$c_0 = 0$	c_1	c_{k-2}	c_{k-1}	$c_k = n$
|_____ n_1 ____|__ •••••• __|____			n_{k-1} __|_____ n_k ____|	
$\widehat{t}_0$	$\widehat{t}_1$	$\widehat{t}_{k-2}$	$\widehat{t}_{k-1}$	$\widehat{t}_k$
$= X_{1,n}$	$= X_{c_1,n}$		$= X_{c_{k-1},n}$	$= X_{n,n}$
$= a$				$= b$

Step 5: Construct $\widetilde{F_C}(x; \widehat{\mathbf{T}}, \mathbf{\Lambda}; \widehat{\mathbf{W}}) = \sum_{i=1}^{k} \left(\frac{n_i}{n}\right) \widetilde{F}_i(x; \widehat{t}_{i-1}, \widehat{t}_i, \alpha_i)$, in *(3.1)*.

Step 6: Estimate α_i. We suggest using the estimator $\widehat{\alpha}$ in *(3.5), (3.6)* and *(3.7)* in Remark 1.

Step 7: Construct an estimator $\widehat{F_C}(x; \widehat{\mathbf{T}}, \widehat{\mathbf{\Lambda}}; \widehat{\mathbf{W}}) = \sum_{i=1}^{k} \left(\frac{n_i}{n}\right) \widehat{F}_i(x; \widehat{t}_{i-1}, \widehat{t}_i, \widehat{\alpha}_i)$, for *(3.1)*.

Remark 1. There are three estimation methods for the shape parameters α_i given by

1. **Hill Estimator** : The Hill (1975) *MLE* $\widehat{\alpha}_{Hill}$ is defined as

$$\widehat{\alpha}_{Hill} = \left[r^{-1} \sum_{i=1}^{r} \{\ln X_{n-i+1,n} - \ln X_{n-r,n}\} \right]^{-1}, \qquad (3.5)$$

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

2. **Moment Estimator** : A moment estimator $\widehat{\alpha}_M$ can be obtained by solving

$$\frac{1}{n} \sum_{i=1}^{n} X_i = \frac{\widehat{\alpha}_M \gamma^{\widehat{\alpha}_M} (\gamma^{1-\widehat{\alpha}_M} - \nu^{1-\widehat{\alpha}_M})}{(\widehat{\alpha}_M - 1)(1 - (\frac{\gamma}{\nu})^{\widehat{\alpha}_M}),} \qquad (3.6)$$

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

3. **MLE method** : The Aban *MLE* $\widehat{\alpha}_{Aban}$ (Aban et al, 2006) for α is obtained by solving

$$\frac{n}{\widehat{\alpha}_{Aban}} + \frac{n(\frac{\gamma}{\nu})\widehat{\alpha}_{Aban} \ln(\frac{\gamma}{\nu})}{1 - (\frac{\gamma}{\nu})^{\widehat{\alpha}_{Aban}}} - \sum_{i=1}^{n} [\ln X_{n-i+1,n} - \ln \gamma] = 0, \qquad (3.7)$$

4. Simulations

One of the most difficult parts of estimating a mixture truncated Pareto distribution is dividing the data set into appropriate groups. In Section 3, we propose a two-point slope method to determine group thresholds of the data. Now we would like to examine the accuracy of this technique. We construct a cluster truncated Pareto distribution function in *(3.1)* for $1 \leq x \leq 10$ and three groups ($k = 3$). The t values were set at $t_1 = 1$, $t_2 = 4$, $t_3 = 7$, and $t_4 = 10$; and $\alpha_1 = 2$, $\alpha_2 = 5$, $\alpha_3 = 0.5$; and $w_1 = 1/6$, $w_2 = 1/3$, $w_3 = 1/2$. Then the c.d.f. of the cluster truncated Pareto distribution is

$$F_C(x;\mathbf{T},\mathbf{\Lambda};\mathbf{W}) = \begin{cases} 0, & x < 1, \\ \frac{0.1778(x^2-1)}{x^2}, & 1 \leq x < 4, \\ 0.5163 - \frac{363.4790}{x^5}, & 4 \leq x < 7, \\ 3.5611 - \frac{8.0989}{x^{0.5}}, & 7 \leq x < 10, \\ 1, & x \geq 10. \end{cases} \tag{4.1}$$

We generated $m = 1000$ random samples from *(4.1)* of size $n = 100$, and found $n - 1$ two-point slopes for each 1000 random samples. The corresponding vectors, $\mathbf{C}$, $\mathbf{W}$, and $\mathbf{T}$, were found accordingly by using a two-point slope technique. Figure 3 shows the boxplot of the estimated threshold limits. In this figure we can see that the two-point slope estimator seems to be very accurate with minimal variance. The only threshold limit with variance is $t_3 = 7$; the other three threshold limits have little variance. Table 3 shows the means and RMSE (root of the mean square error) of the estimated four threshold limits.

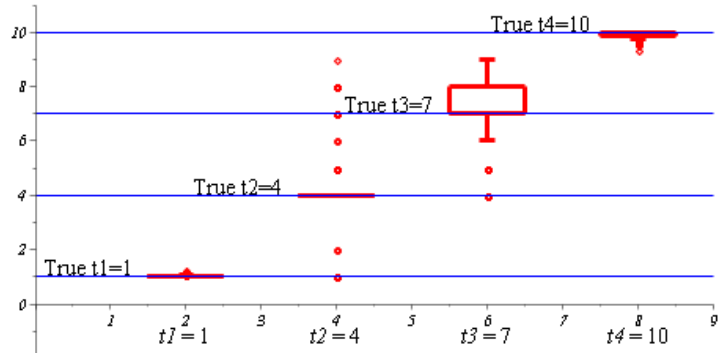

Figure 3. Boxplots of the estimated threshold limits using two-point slopes with $m = 1000$ simulated datasets of size $n = 100$ from a cluster truncated Pareto distribution.

Table 3. The means and RMSEs of the estimated threshold limits in the simulations

Threshold	t_1	t_2	t_3	t_4
True t *value*	1	4	7	10
Mean	1.0311	4.4250	7.2940	9.9215
RMSE	0.0445	1.6072	1.2806	1.1112

5. Applications

Now we apply the proposed cluster method to the flood damage example in Section 1.

5.1. Cluster Method

By using 33 two-point slopes defined in *(3.2)* and the seven steps in Section 3, we construct $k = 3$ clusters. We select $k - 1 = 2$ of the ten largest absolute values of the two-point slopes in *(3.2)* to ensure $\alpha > 0$ and appropriate n_i values as

$$S_{31,33} = 12.4657; \quad S_{25,33} = 4.1745.$$

Then we determine $\widehat{t}_i$'s, $i = 0, 1, 2, 3$, and $k = 3$ groups as

$$\left\{a = \widehat{t}_0 = X_{(1)}, ..., X_{(c_1)}\right\}, \left\{X_{(c_1+1)}, ..., X_{(c_2)}\right\}, \left\{X_{(c_2+1)}, ..., \widehat{t}_3 = X_{(n)} = b\right\},$$

$$\begin{aligned} \text{where } \widehat{t}_0 &= X_{(1)} = 1030,\ \widehat{t}_1 = X_{(c_1)} = 4151.3,\ \widehat{t}_2 = X_{(c_2)} = 7343.6,\ \widehat{t}_3 = X_{(n)} = 17346.8; \\ c_0 &= 0,\ c_1 = 16,\ c_2 = 26,\ c_3 = 34; \\ n_1 &= 16,\ n_2 = 10,\ n_3 = 8;\ n_1 + n_2 + n_3 = n = 34; \end{aligned}$$

then we have an estimated CTPD in *(3.1)* as

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

Table 4 shows the construction of the CTPD from the data.

Table 4. Construction of a CTPD from the flood damage data.

$c_0 = 0$	$c_1 = 16$	$c_2 = 26$	$c_3 = 34$
$\lvert\_\_\_ n_1 = 16 \_\_\_\rvert\_\_\_ n_2 = 10 \_\_\_\rvert\_\_\_ n_3 = 8 \_\_\_\rvert$			
$\widehat{t}_0$	$\widehat{t}_1$	$\widehat{t}_2$	$\widehat{t}_3$
$= 1030$	$= 4151.3$	$= 7343.6$	$= 17346.8$

Once the thresholds were decided based on the two-point slope method, four truncated Pareto distribution functions were created.

Table 5 provides the comparison between the estimation methods of the Canadian flood loss dataset: Pareto distribution using Hill's estimator, the truncated Pareto distributions (TPD) using both Aban's estimator and the Moment estimator, and the new MTPD method. The table compares the estimation methods through $\widehat{\alpha}$, $\widehat{\mu}$, median, 5% Value-at-Risk, and 1% Value-at-Risk.

Table 5. Comparisons of the estimation methods on the flood damage example.

Estimation Method	$\widehat{\alpha}$	$\widehat{\mu}$	Median	5% Value-at Risk	1% Value-at Risk
Pareto$_{(Hill)}$	0.7244	∞	2681.76	64410.33	594167.01
TPD$_{(Aban)}$	0.1085	5413.77	3795.23	14726.64	16783.97
TPD$_{(moment)}$	0.1838	5168.64	3526.63	14460.35	16718.54
Cluster	$\widehat{\alpha}_1$=0.1680	5122.54	4309.75	13092.99	16250.19
	$\widehat{\alpha}_2$=1.7282				
	$\widehat{\alpha}_3$=1.3281				

Figure 4 shows the log-log plot for the dataset of Canadian flood damage costs ($n = 34$). The four estimation methods were used to construct distribution functions that were plotted on the same log-log scale. Visually, the same trend occurs as in Figure 2. The original Pareto distribution function does not fit the data well as it does not curve to take the most extreme values into account. The truncated Pareto distribution with the Aban estimator seems to have a better fit than the original Pareto distribution, but still does not follow the data well in the tail of the data. The new mixture truncated Pareto distribution seems to fit the data the best.

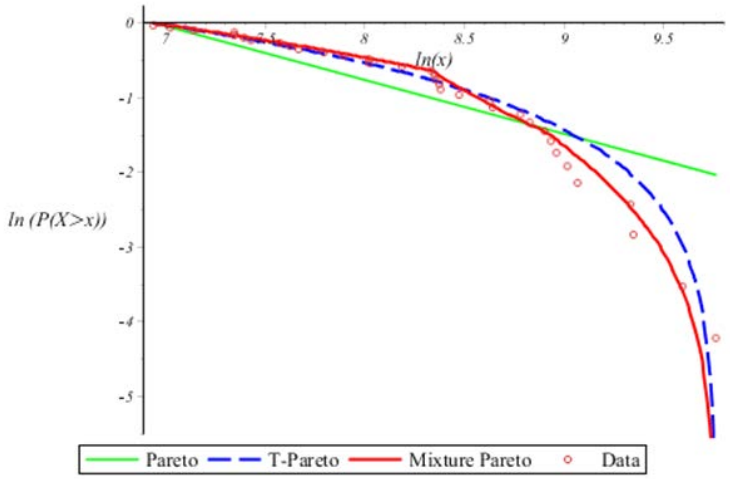

Figure 4. Flood damage data log-log plot and the estimated distribution functions. The original Pareto distribution with Hill's estimator is the green straight line, the truncated Pareto distribution with Aban's estimator is the blue dash line and the MTPD is the red line.

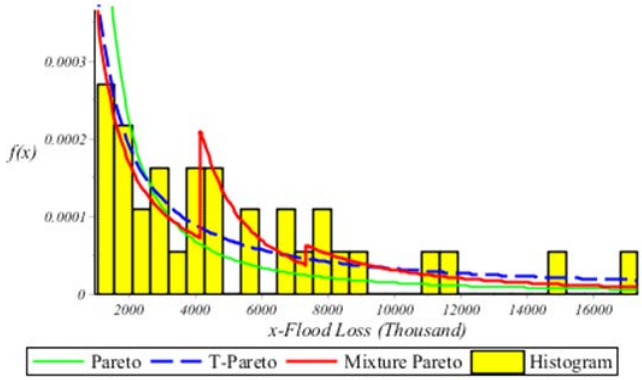

Figure 5. Histogram of the flood damage data with the estimated original Pareto density function is the green line, the Aban estimated truncated Pareto density function is the blue dash line, and the estimated mixture Pareto density function is the red line.

Figure 5 shows the histogram of the Flood damage data with three estimated probability density functions. We see that the mixture Pareto distribution models the data better as it

has peaks where the data increases whereas the estimated original Pareto and the estimated truncated Pareto distribution do not follow those peaks.

5.2. Goodness of Fit Tests

In this section we conduct three goodness of fit tests, Kolmogorov-Smirnov, Anderson Darling, and Cramer-von Mises tsets. All three tests are based on the distance between the empirical distribution function and the proposed distribution function: original Pareto distribution in *(2.1)* or truncated Pareto distribution in *(2.3)* or mixture truncated Pareto distribution in *(2.10).*

Each test considers the same null and alternative hypothesis:

$$H_0 : F(x) = F^*(x) \quad vs \quad H_1 : F(x) \neq F^*(x),$$

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 *(2.1)* with Hill estimator $\widehat{\alpha}_{Hill}$ in *(3.5)*;
2) Truncated Pareto distribution (TPD) in *(2.3)* with Aban estimator $\widehat{\alpha}_{Aban}$ in *(3.7);*
3) Truncated Pareto distribution in (TPD) *(2.3)* with moment estimator $\widehat{\alpha}_M$ in *(3.6);*
4) Cluster truncated Pareto distribution in *(3.1)* with moment estimator $\widehat{\alpha}_{M(i)}$ in *(3.6).*

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

Table 6. Goodness of fit tests $n = 34$ for the flood damage 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
$\text{Pareto}_{(Hill)}$	0.1945	0.1290	2.7048	0.0388	0.4151	0.0659
$\text{TPD}_{(Aban)}$	0.1003	0.7224	1.4119	0.1990	0.0516	0.8663
$\text{TPD}_{(moment)}$	0.1170	0.5979	1.4110	0.1992	0.0721	0.7390
Cluster	0.0803^{best}	0.8500^{best}	1.2075^{best}	0.2647^{best}	0.0196^{best}	0.9973^{best}

Note: In this paper, we use "^{best}" to denote the best values in the tables.

Table 6 gives the values of the test statistics and p-values of three goodness-of-fit tests. The cluster truncated Pareto distribution has the smallest test statistics (i.e., the smallest errors) and the largest p-values. This means the cluster truncated Pareto distribution has the best fit to the Canadian flood damage costs data.

Table 7. Errors of goodness-of-fit tests $n = 34$ for flood damage example

Method	Goodness-of-Fit Tests					
	Absolute Error (AE)			Integrated Error (IE)		
	$r = 34$	$r = 20$	$r = 10$	$r = 34$	$r = 20$	$r = 10$
$\text{Pareto}_{(Hill)}$	0.1945	0.1945	0.1664	0.1007	0.0911	0.0894
$\text{TPD}_{(Aban)}$	0.1003	0.0907	0.0750	0.0412	0.0423	0.0445
$\text{TPD}_{(moment)}$	0.1170	0.1170	0.0893	0.0399	0.0366	0.0338
Cluster	0.0803^{best}	0.0803^{best}	0.0603^{best}	0.0216^{best}	0.0223^{best}	0.0199^{best}

In Table 7, we took the r largest data in the sample. The absolute error and integrated error are defined by

$$AE = \sup_x |F^*(x) - S_n(x)|, \quad -\infty < x < \infty, \tag{5.1}$$

$$IE = \frac{1}{(X_{n,n} - X_{n-r+1,n})} \left[\int_{X_{n-r+1,n}}^{X_{n,n}} (S_n(x) - F^*(x))^2 dx \right]^{1/2}. \tag{5.2}$$

Table 7 gives absolute errors and integrated errors of the five estimation methods in $r = 34, 20, 10$ cases. The cluster truncated Pareto distribution has the smallest errors in all 6 cases. This means the cluster method is superior in fitting the flood damage costs data compared with the other existing methods.

6. Conclusions

In this paper, we found that the estimated mixture Pareto distribution has better fitting than only one single estimated Pareto or truncated Pareto distribution, for a complicated data set with heavy tailed and cluster properties. The new method based on the two-point slope technique breaks the data into different groups.

Summary of some useful results in this paper are as follows:

1. Truncated Pareto models are useful for analyzing real world data.

2. The results of the goodness-of-fit tests show that the cluster truncated Pareto distribution is a better model for fitting data than just using a single Pareto distribution model.

3. The results of simulations show that the two-point slope technique is innovate and useful, and seems to be an accurate method to determine the thresholds.

4. This method has the best fit in the flood damage costs data set of examples compared to the existing methods as seen by the goodness-of-fit tests.

References

[1] Aban, I. B., Meerschaert, M. M. and Panorska, A. K. 2006. Parametric estimation for truncated Pareto distribution, *Journal of the American Statistical Association, Vol. 101, No. 473, 270-277.*

[2] Beg, M. A. 1981. Estimation of the tail probability of the truncated Pareto distribution, *Journal of Information & Optimization Sciences, 2, 192-198.*

[3] Coia, V. and Huang, M. L. 2014. A sieve model for extreme values, *Journal of Statistical Computation and Simulation., Vol. 84, No. 8, pp. 1692-1710.*

[4] Everitt, B. S., Landau, S. Leese, M. and Stahl, D. 2011. *Cluster Analysis,* 5th ed., Wiley, New York.

[5] Frühwirth-Schnatter, S. 2006, *Finite Mixture and Markov Switching Models*, Springer, New York.

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