A Mixture Truncated Pareto Distribution

Mei Ling Huang[a]*, Vincenzo Coia[a] and Percy Brill[b]†

[a]Department of Mathematics, Brock University,
St. Catharines, Ontario, Canada
[b]Department of Mathematics & Statistics, University of Winsdor
Windsor Ontario, Canada

August 23, 2012

Abstract

Heavy tailed distributions have many applications in the real world. The Pareto distribution is very popular. The tail of the distribution is important but the threshold of the distribution is difficult to determine. We propose a mixture truncated Pareto distribution (MTPD). This study leads us to construct a cluster Pareto distribution (CPD) estimation method. The paper studies a real world example which utilizes the MTPD. The results of goodness of fit tests show that the MTPD and the cluster estimation method produce very good fitting distributions with real world data.

Keywords: Cluster, heavy tailed distributions; goodness of fit tests; order statistics; truncated Pareto distribution.

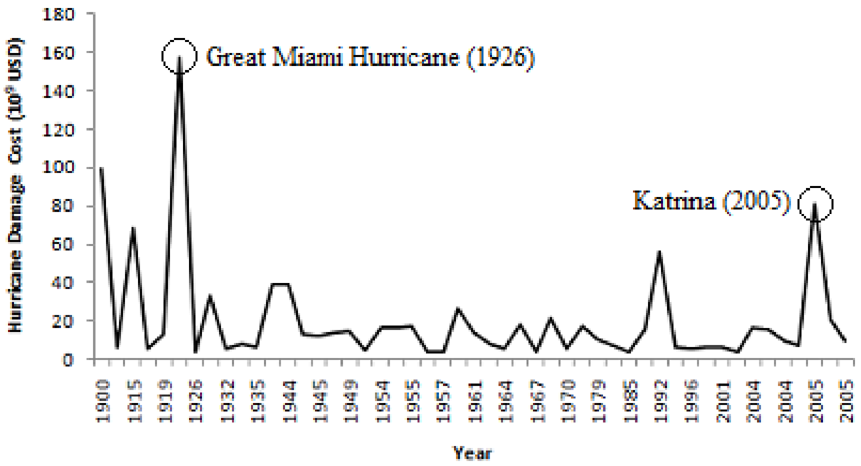
1. Introduction

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 can not be used in these situations. Several authors suggest using a truncated Pareto distribution (TPD) which always has finite moments (e.g., Beg, 1981; Aban, et al, 2006).

*Corresponding author, e-mail: mhuang@brocku.ca.

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

In some situations, data behaves differently within different thresholds. For example, losses from hurricane damages can be classified into small, medium and large hurricane groups. The data in these classes may have different distributions, or grouped data may have the same kind of distribution but with different parameters. A cluster method is needed 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 (U.S. National Hurricane Center, 2008). The costs are standardized to 2005 USD; see figure below.



The 49 Costliest Atlantic hurricanes between the years 1900-2005.

Huang and Zhao (2011) used Pareto and truncated Pareto models to fit the 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 1. Huang and Zhao (2011) 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 perfectly, 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 a cluster method. In this paper, we propose a mixture truncated Pareto distribution (MTPD) in Section 2. We propose a cluster Parato distribution (CPD) to estimate the MTPD in Section 3. In Section 4, we perform Kolmogorov-Smirnov, Anderson Darling, and Cramer-von Mises goodness of fit tests to analyze the hurricane data by using the CPD and three other existing estimation methods (see Figures 2). The results show that the proposed cluster method is superior to other existing estimation methods.

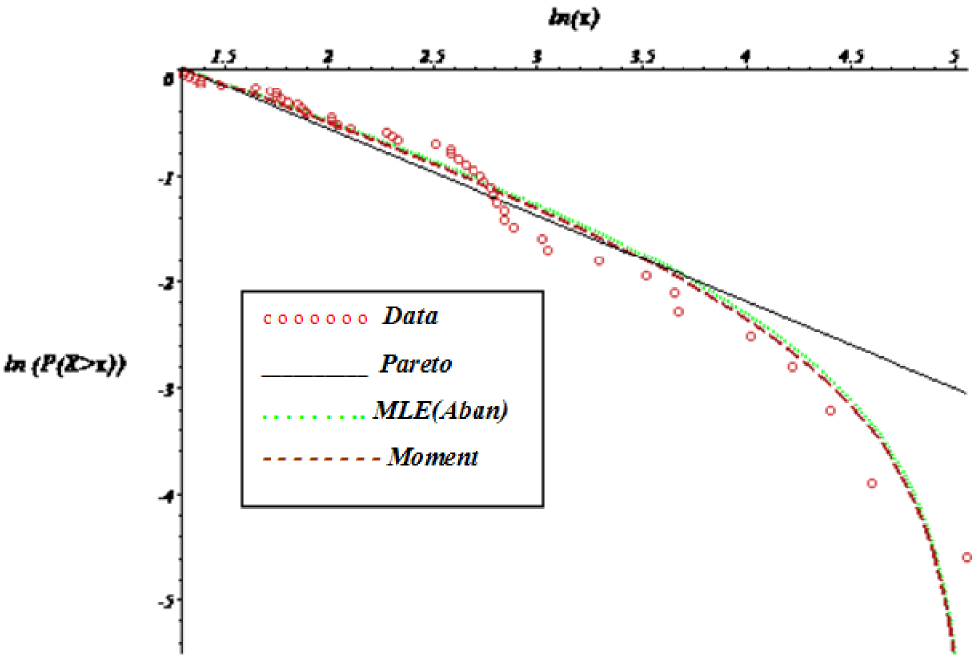


Figure 1: Log-log plot of hurricane example with estimated distribution curves. The red circles are the data; the black straight line is the original Pareto distribution; the green dot line is the MLE estimated truncated Pareto distribution; the brown dash line is the moment estimated truncatted Pareto distribution.

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 by*

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

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

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 is heavier in the right tail as α decreases.

The truncated Pareto distribution 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 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 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

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

The mean, second moment and variance of X are

$$\mu=\frac{\alpha\gamma^{\alpha}(\gamma^{1-\alpha}-\nu^{1-\alpha})}{(\alpha-1)(1-(\frac{\gamma}{\nu})^{\alpha})},\quad 0<\gamma<\nu<\infty,\ \alpha>1; \tag{2.6}$$

$$\mu_{(2)}=\frac{\alpha\gamma^{\alpha}(\nu^{2-\alpha}-\gamma^{2-\alpha})}{(2-\alpha)(1-(\frac{\gamma}{\nu})^{\alpha})^{2}},\quad 0<\gamma<\nu<\infty,\ \alpha>2; \tag{2.7}$$

$$\sigma^{2}=\frac{\alpha\gamma^{\alpha}(\nu^{2-\alpha}-\gamma^{2-\alpha})}{(2-\alpha)(1-(\frac{\gamma}{\nu})^{\alpha})^{2}}-\frac{\alpha^{2}\gamma^{2\alpha}(\nu^{1-\alpha}-\gamma^{1-\alpha})^{2}}{(1-\alpha)^{2}(1-(\frac{\gamma}{\nu})^{\alpha})^{2}},\ 0<\gamma<\nu<\infty,\ \alpha>2. \tag{2.8}$$

We consider thresholds as a vector $\mathbf{T}=(t_0,t_1,...,t_k)^T,$ where $0<\gamma=t_k<t_{k-1}...<t_0=\nu<\infty$. Also consider vector $\mathbf{\Lambda}=(\alpha_1,\alpha_2,...,\alpha_k)^T,\ \alpha_i>0,\ i=1,...,k$. We define a mixture truncated Pareto distribution as

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

$$F_c(x;\mathbf{T},\mathbf{\Lambda};\mathbf{W})=\sum_{i=1}^{k}w_iF_i(x;\gamma_i,\nu_i,\alpha_i),\ 0<\gamma\le x\le\nu<\infty,\ \gamma=\min_{i=1}^{k}\gamma_i,\ \nu=\max_{i=1}^{k}\nu_i, \tag{2.9}$$

where $F_i(x;\gamma_i,\nu_i,\alpha_i)$ is the c.d.f. of the truncated Pareto distribution in (2.3), and the truncation points γ_i, ν_i, are related to thresholds $\mathbf{T}=(t_0,t_1,...,t_k)^T$,

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

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

$$f_c(x;\mathbf{T},\mathbf{\Lambda};\mathbf{W})=\sum_{i=1}^{k} w_i f_i(x;\gamma_i,\nu_i,\alpha_i),\ \ 0<\gamma\le x\le\nu<\infty,\ \gamma=\min_{i=1}^{k}\gamma_i,\ \nu=\max_{i=1}^{k}\nu_i, \tag{2.10}$$

where $f_i(x;\gamma_i,\nu_i,\alpha_i)$ is the c.d.f. of the truncated Pareto distribution in (2.4).

3. A Cluster Distribution Estimator

Consider a random sample $X_{1,}X_{2,}..., X_n$ from the MTPD in *(2.9)*, and let $X_{1,n}\ge X_{2,n}\ge ...\ge X_{n,n}$ denote its order statistics. We divide data into k clusters by the domains $(t_{i+1},t_i),\ i=0,1,...,k-1$. $\mathbf{T}=(t_0,t_1,...,t_k)^T,$ where $0<\gamma=t_k<t_{k-1}...<t_0=\nu<\infty$. We define a cluster Pareto distribution as

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

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

where $FC(x;\mathbf{T},\mathbf{\Lambda};\mathbf{W})$ 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).

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

where 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 greater than or equal to the threshold t_i. c_i is a function of t_i and the random sample $(X_1,X_2,...X_n)$, thus

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

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

where I_A is an indicator function of set A.

In this paper, we propose a two-points slope method in the log-log plot to determine thresholds $\mathbf{T}=(t_0,t_1,...,t_k)^T$.

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

$$S_i(X_1,X_2,...X_n)=\begin{cases}\frac{\log(1-\frac{i+1}{n})-\log(1-\frac{i}{n})}{\log(X_{i+1,n})-\log(X_{i,n})}, & \log(X_{i+1,n})-\log(X_{i,n})\ne0,\ i=1,...n-1;\\ 0, & \log(X_{i+1,n})-\log(X_{i,n})=0,\ i=1,...n-1.\end{cases} \tag{3.2}$$

Then we make order statistics $S_{1,n} \leq S_{2,n} \leq ... \leq S_{n-1,n}$ of the absolute slope $|S_i(X_1, X_2, ...X_n)|\,,\ i = 1, 2, ...n-1$. 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 top $(k-1)$ slopes

$$S_{n-k.n} \leq S_{n-k+1,n} \leq ... \leq S_{n-1,n}. \qquad (3.3)$$

We propose six steps to construct a cluster Pareto distribution:

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

Step 2: Find the $(k-1)$ estimated threshold points $\widehat{t}_1, ..., \widehat{t}_{k-1}$ by using 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

$$X_{1,n}^* \geq X_{2,n}^* \geq ... \geq X_{k-1,n}^*;$$

then we let

$$\widehat{t}_i(X_1, X_2, ...X_n) = X_{i,n}^*,\ i = 1, ...k-1,\ and\, \widehat{t}_0 = X_{1,n} = \nu,\ \widehat{t}_k = X_{n,n} = \gamma.$$

Step 3: Determine $\mathbf{C} = (c_0, c_1, ..., c_k)^T$, where $0 = c_0 < c_1 ... < c_k = n,\ c_i(t_i; X_1, X_2, ...X_n) = \sum_{j=1}^{n} I_{X_j \geq t_i}(X_j)$. Thus

$$\begin{aligned} n_i &= c_i - c_{i-1},\ i = 1, 2, ..., k; \\ \widehat{t}_i &= X_{c_i,n},\ i = 1, ...k-1,\ and\ \widehat{t}_0 = X_{1,n} = \nu,\ \widehat{t}_k = X_{n,n} = \gamma. \end{aligned}$$

Then we have k clusters:

$$\{\gamma = \widehat{t}_k = X_{c_k,n}, ..., X_{c_{k-1},n}\}, \{X_{c_{k-1},n}, ..., X_{c_{k-2},n}\}, ..., \{X_{c_1,n}, ..., \widehat{t}_0 = X_{1,n} = \nu\}.$$

Table 3.1 Construction of Cluster Pareto Distribution

$c_k = n$	c_{k-1}	c_2	c_1	$c_0 = 0$
______ n_k ____	______...... __	____ n_2 __	______ n_1 ____	
$\widehat{t}_k$ $= X_{c_k,n}$ $= X_{n,n}$ $= \gamma$	$\widehat{t}_{k-1}$ $= X_{c_{k-1},n}$	$\widehat{t}_2$	$\widehat{t}_1$ $= X_{c_1,n}$	$\widehat{t}_0$ $= X_{1,n}$ $= \nu$

Step 4: Construct $\widetilde{FC}(x; \widehat{\mathbf{T}}, \boldsymbol{\Lambda}; \widehat{\mathbf{W}}) = \sum_{i=1}^{k} \left(\frac{n_i}{n}\right) \widetilde{F}_i(x; \widehat{t}_i, \widehat{t}_{i-1}, \alpha_i),$ *in (3.1).*

Step 5: Estimate $\widehat{\alpha}_{i,Moment}$. (We suggest using the moment estimator in *(3.4)* since it has robust properties, but there are other estimators available in *(3.4), (3.5), (3.6)*)

Step 6: Construct an estimator $\widehat{FC}(x;\widehat{\mathbf{T}},\widehat{\mathbf{\Lambda}};\widehat{\mathbf{W}}) = \sum_{i=1}^{k} \left(\frac{n_i}{n}\right) \widehat{F}_i(x;t_i,t_{i-1},\widehat{\alpha}_i),\ for\ (3.1).$

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

1. Hill Estimator (original Pareto distribution):

The Hill (1975) *MLE* for α is defined as

$$\widehat{\alpha}_{Hill} = \left[r^{-1} \sum_{i=1}^{r} \{\ln X_{i,n} - \ln X_{r+1,n}\} \right]^{-1}, \qquad (3.4)$$

where $X_{i,n}$ is the ith largest order statistic, and r is the cut off point, i.e., $X_{r+1,n} \leq \nu$.

2. Moment Estimator (truncated Pareto distribution):

A moment estimator $\widehat{\alpha}_M$ for α can be obtained by solving the following equation:

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \frac{\widehat{\alpha}_M \gamma^{\widehat{\alpha}_M}(\gamma^{1-\widehat{\alpha}_M} - \nu^{1-\widehat{\alpha}_M})}{(\widehat{\alpha}_M - 1)(1 - (\frac{\gamma}{\nu})^{\widehat{\alpha}_M}),} \qquad (3.5)$$

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

3. MLE method (truncated Pareto distribution)

The Aban *MLE* for α (Aban et al, 2006) is obtained by solving the following equation:

$$\frac{n}{\widehat{\alpha}_{Aban}} + \frac{n(\frac{\gamma}{\nu})\widehat{\alpha}_{Aban} \ln(\frac{\gamma}{\nu})}{1 - (\frac{\gamma}{\nu})^{\widehat{\alpha}_{Aban}}} - \sum_{i=1}^{n} [\ln X_{i,n} - \ln \gamma] = 0, \qquad (3.6)$$

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

4. Applications

Now we apply the cluster Pareto distribution to the hurricane example.

4.1. Cluster Method

By using *48* two-point slopes in *(3.2)* and the six steps in Section 4, we construct $k = 3$ clusters,

$$\{\gamma = \widehat{t}_3 = X_{c_3,n}, ..., X_{c_2,n}\},\ \{X_{c_2,n}, ..., X_{c_1,n}\}, \{X_{c_1,n}, ..., \widehat{t}_0 = X_{1,n} = \nu\},$$

$$\begin{aligned} \widehat{t}_0 &= X_{1,n} = 157 = \nu,\ \widehat{t}_1 = X_{c_1,n} = 26.8,\ \widehat{t}_2 = X_{c_2,n} = 13.7,\ \widehat{t}_3 = X_{c_3,n} = 3.7 = \gamma; \\ c_0 &= 0,\ c_1 = 9,\ c_2 = 22,\ c_3 = 49; \\ n_1 &= 9,\ n_2 = 13,\ n_3 = 27;\ n_1 + n_2 + n_3 = n = 49; \end{aligned}$$

Table 4.1 Construction of Cluster Pareto Distribution

$c_3 = n = 49$ $c_2 = 22$ $c_1 = 9$ $c_0 = 0$

|______ $n_3 = 27$____|____ $n_2 = 13$__|_____$n_1 = 9$____|

$\widehat{t}_3 = 3.7$ $\widehat{t}_2 = 13.7$ $\widehat{t}_1 = 26.8$ $\widehat{t}_0 = 157$

$= X_{c_3,n}$ $= X_{c_2,n}$ $= X_{c_1,n}$ $= X_{1,n}$

$= X_{n,n}$ $= \nu$

$= \gamma$

Table 4.2 Four Estimation Methods for Hurricane Example

Method	$\widehat{\alpha}$	$\widehat{\mu}$	Median	5% VaR	5% VaR
Pareto$_{(Hill)}$	0.8126	∞	8.68 billion	147.68 billion	1070.30 billion
MLE$_{(Aban)}$	0.6206	21.10 billion	9.73 billion	85.15 billion	136.17 billion
Moment	0.6476	20.48 billion	9.47 billion	82.55 billion	134.90 billion
Cluster	$\widehat{\alpha}_1$=0.6476 $\widehat{\alpha}_2$=5.6498 $\widehat{\alpha}_3$=0.8416	25.06 billion	11.19 billion	77.24 billion	132.41 billion

Table 4.2 gave $\widehat{\alpha}$, $\widehat{\mu}$, Median, 5% Value-at-Risk (VaR) and 1% Value-at-Risk (VaR) of each of four estimation methods. We note that the Cluster method gives the largest mean and median and the smallest VaRs

Figure 2 exhibits data and four 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 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 fitting to the data over the whole range. Figure 2 suggests a single distribution may not totally represent how natural data is distributed. We may consider grouping data by using the cluster method.

The result in Figure 2 is a visual observation. It is necessary to run goodness of fit tests to confirm which estimated distribution best fits the hurricane data.

4.2. Goodness of Fit Test

In this section we will 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 *(2.1)* or truncated Pareto distribution in *(2.3),* or mixture truncated Pareto distribution in *(2.9).*

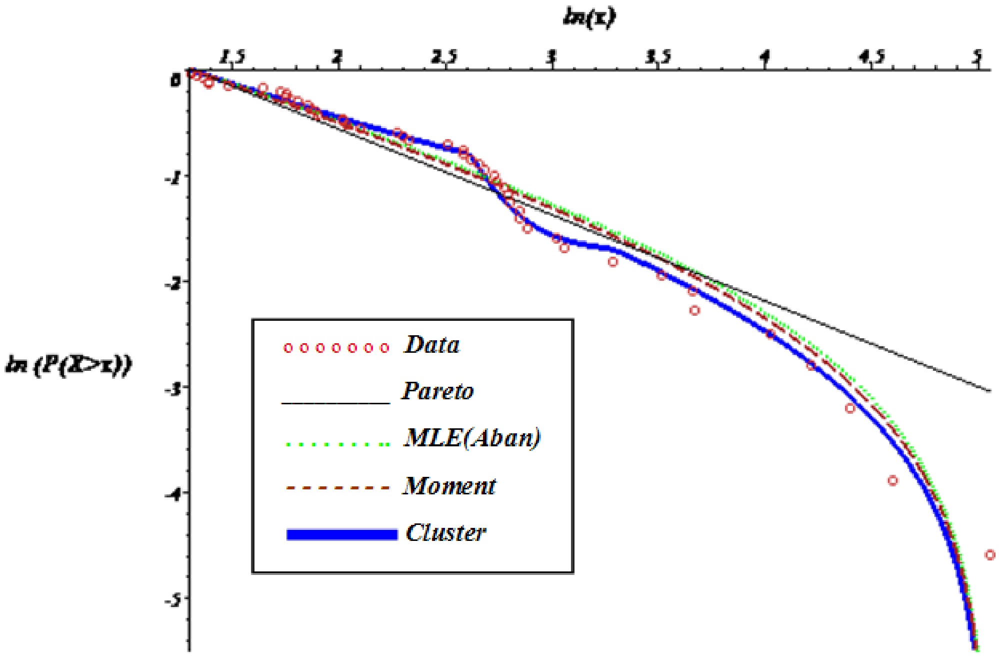


Figure 2: Log-log plot of hurricane example with estimated distribution curves. The red circles are the data; the black straight line is the original Pareto distribution; the green dot line is the MLE estimated truncated Pareto distribution; the brown dash line is the moment estimated truncatted Pareto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

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 $\widehat{\alpha}_{Hill}$;
2) Truncated Pareto distribution in *(2.3)* with Aban MLE estimator $\widehat{\alpha}_{Aban;}$
3) Truncated Pareto distribution in *(2.3)* with Moment estimator $\widehat{\alpha}_{Moment;}$
4) Cluster Pareto distribution in *(3.1)* with estimator $\widehat{\alpha}_{Moment(i);}$

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

(1) The Kolmogorov-Smirnov (K-S) test (Kolmogorov, 1933), the test statistic is given by,

$$T = \sup_x |F^*(x) - S_n(x)|, \quad -\infty < x < \infty, \qquad (4.1)$$

where $S_n(x)$ is the empirical distribution function.

(2) Anderson and Darling (1952) 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,

$$W_n^2 = n \int_{-\infty}^{\infty} \left[S_n(x) - F^*(x)\right]^2 \psi\left(F^*(x)\right) dF, \quad \text{where} \quad \psi(u) = \frac{1}{u(1-u)}. \qquad (4.2)$$

(3) Cramer-von Mises (Anderson and Darling, 1952) proposed using $\psi(u) = 1$ in *(4.2)*, thus under H_0 the test statistic and p-value are given by

$$n\omega^2 = \frac{1}{12n} + \sum_{j=1}^{n} \left(u_j - \frac{2j-1}{2n}\right)^2 . \qquad (4.3)$$

Table 4.3 Goodness of Fit Tests $n = 49$ for Hurricane 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$_{(Hill)}$	0.1130	0.4608	2.7141	0.0383	0.2057	0.2568
MLE$_{(Aban)}$	0.0839	0.7251	2.3126	0.0622	0.0964	0.6030
Moment	0.0828	0.7341	2.3672	0.0582	0.1095	0.5402
Cluster	**0.0700**	**0.8328**	**2.1258**	**0.0784**	**0.0429**	**0.9177**

Table 4.3 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 (means smallest errors) and the largest p-values in each of all three tests respectively (we highlighted the values as bold in the table). This means the cluster truncated Pareto distribution has the best fitting to the hurricane data.

Table 4.4 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$_{(Hill)}$	0.1130	0.0584	0.0584	0.4844	0.3818	0.3723
MLE$_{(Aban)}$	0.0839	0.0839	0.0832	0.3114	0.2565	0.2161
Moment	0.0828	0.0738	0.0738	0.2985	0.2171	0.1825
Cluster	**0.0700**	**0.0399**	**0.0228**	**0.1650**	**0.1325**	**0.1218**

In Table 4.4, we took the r largest data in the sample. The absolute error and integrated error are defined by

$$AE = \sup_x |F^*(x) - S_n(x)|, \quad -\infty < x < \infty, \tag{4.4}$$

$$IE = \left[\int_{X_{n-r+1,n}}^{X_{n,n}} (S_n(x) - F^*(x))^2 dx \right]^{1/2}. \tag{4.5}$$

Table 4.4 gives absolute errors and integrated errors of four estimation methods in $r = 49, 18, 10$ cases. We note that the cluster truncated Pareto distribution has the smallest errors in all 6 cases (we highlighted the values as bold in the table). This means the cluster method is superior in fitting the hurricane data compared with the other existing methods.

References

[1] Anderson, T. W. and Darling, D. A. 1952. Asymptotic theory of certain goodness of fit criteria based on stochastic processes, *The Annals of Mathematical Statistics, 23, 193-212.*

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

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

[4] David. H. A. 2005. *Order Statistics*, third edition. Wiley, New York.

[5] Hill, M. 1975. A Simple general approach to inference about the tail of a distribution, *The Annals of Statistics, Vol.3, No.5, pp1163-1174.*

[6] Huang, M. L. and Zhao, K. 2011. Weighted efficient estimator for risk models, *Working Paper*, *Brock University.*

[7] Kleiber, C. K. and Kotz, S. 2003. *Statistical Size Distribution in Economics and Actuarial Sciences.* John Wiley & Sons, New York.

[8] Kolmogorov, A. N. 1933. Sulla determinazione empirica di una legge di distribuzione, *Giornale dell' Istituto Italiano degli Attuari,* 4, 83-91 (6.1).

[9] U.S. National Hurricane Center, (2008). *Hurricane History*, http://www.nhc.noaa.gov/HAW2/english/history.shtml.