A Nonparametric Method for Extremes

Mei Ling Huang[a]*, Lucas Thorpe[a] and Percy Brill[b]†

[a]Department of Mathematics, Brock University, Canada
[b]Department of Mathematics and Statistics, University of Windsor, Canada

August 20, 2013

Abstract

The Pareto distribution is a popular distribution with many applications in real world extreme events. Many existing estimation methods for the Pareto distribution are based on estimating the tail index. There are bias and infinite moments issues. In this paper, we first use a truncation method to overcome these difficulties, then propose a nonparametric kernel distribution estimation method. The paper also studies an example of real world applications on extreme values by using the proposed method. We compare the results on goodness-of-fit tests by applying the proposed methods and also existing parametric methods with real-world data.

Keywords: Extreme value, heavy tailed distributions; goodness of fit test; order statistics; Pareto distribution.

1. Introduction

Natural disasters such as earthquakes, hurricanes, wild fires, air pollution or floods can cause huge damage. Scientists study these extreme events. In the literature, extreme value problems are 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. Several authors suggest using a truncated Pareto distribution (TPD), which 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 in a complicated manner. For example, losses from hurricane damage may not be fitted to a single Pareto distribution with an estimated tail index. In this paper we propose a nonparametric method to study an example of 49 most damaging Atlantic hurricanes occurring between years 1900 and 2005 (U.S. National Hurricane Center, 2008). Figure 1.1 gives the costs standardized to 2005 USD. Figure 1.2 is a histogram of Atlantic hurricane losses.

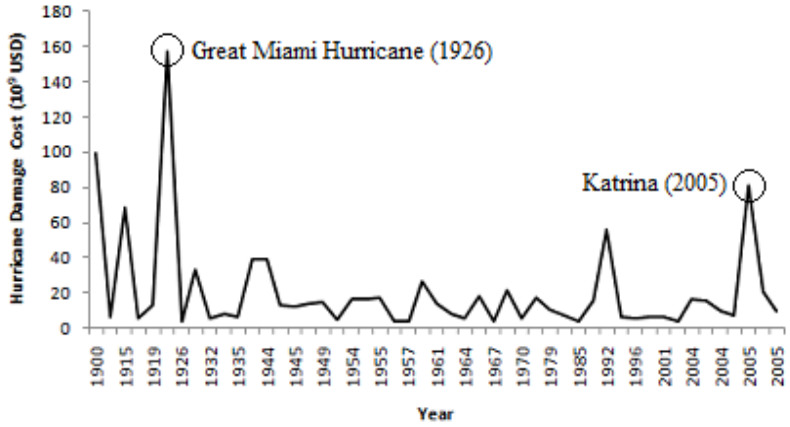

Figure 1.1. The 49 costliest Atlantic hurricanes between the years 1900-2005.

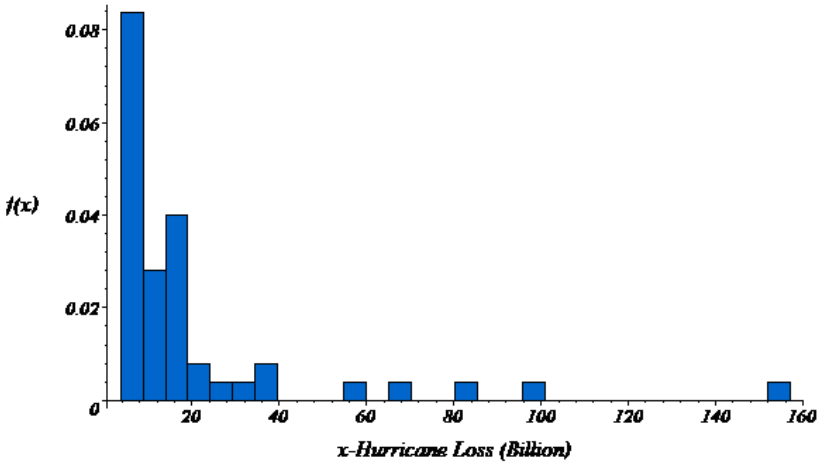

Figure 1.2. Histogram of hurricane loss data

Huang and Zhao (2013) used the 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 1.3. Huang and Zhao (2013) 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 in Figure 1.3; 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 may not fit a single parameter distribution. This motivates us to try a nonparametric estimation method to find a suitable distribution to fit this empirical data better.

In this paper, we review the Pareto distribution and truncated Pareto distribution with their estimation methods in Section 2. We propose a nonparametric kernel distribution estimator (KDE) 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 KDE and three other existing estimation methods (see Figure 4.2). The results show that the proposed nonparametric method is superior to other existing estimation methods.

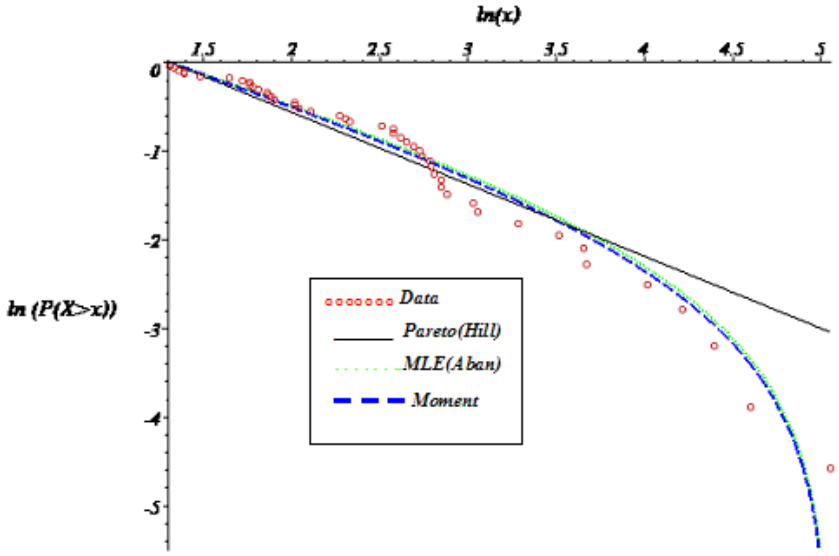

Figure 1.3. Log-log plot of hurricane example with estimated distribution curves. The red circles are the data; the black straight line is the original Pareto distribution; the green dot line is the MLE estimated truncated Pareto distribution; the blue dash line is the moment estimated truncated Pareto distribution.

2. Pareto Distribution and 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*

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

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

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 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

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

The mean, second moment and variance of X are

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

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

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

Let $X_{1,n}, X_{2,n}, ..., X_{n,n}$ be order statistics of a random sample $X_1, X_2, ..., X_n$ from the TPD in *(2.3)*. Three estimation methods for the shape parameter α are given by

1. **Hill Estimator** (original Pareto distribution in *(2.1)*):
The Hill (1975) *MLE* for α is defined as

$$\widehat{\alpha}_{Hill} = \left[r^{-1} \sum_{i=1}^{r} \{ \ln X_{n-i+1,n} - \ln X_{n-r,n} \} \right]^{-1}, \tag{2.9}$$

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

2. **Moment Estimator** (truncated Pareto distribution in *(2.3)*):
A moment estimator $\widehat{\alpha}_M$ for α can be obtained by solving the following equation:

$$\frac{1}{n} \sum_{i=1}^{n} X_i = \frac{\widehat{\alpha}_M \gamma^{\widehat{\alpha}_M} (\gamma^{1-\widehat{\alpha}_M} - \nu^{1-\widehat{\alpha}_M})}{(\widehat{\alpha}_M - 1)(1 - (\frac{\gamma}{\nu})^{\widehat{\alpha}_M}),} \tag{2.10}$$

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

3. **MLE method** (truncated Pareto distribution in *(2.3)*)
The *MLE* for α (Aban et al, 2006) is obtained by solving the following equation:

$$\frac{n}{\widehat{\alpha}_{Aban}} + \frac{n(\frac{\gamma}{\nu}) \widehat{\alpha}_{Aban} \ln(\frac{\gamma}{\nu})}{1 - (\frac{\gamma}{\nu})^{\widehat{\alpha}_{Aban}}} - \sum_{i=1}^{n} [\ln X_{n-i+1,n} - \ln \gamma] = 0, \tag{2.11}$$

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

We use $\widehat{\gamma} = \min(X_1, X_2, ..., X_n)$, $\widehat{\nu} = \max(X_1, X_2, ..., X_n)$ in *(2.10)* and *(2.11)*.

3. A Nonparametric Kernel Distribution Estimator (KDE)

We apply the kernel density estimation method (Silverman, 1986) which is given by
Definition 3.1. *The kernel density estimator for a true density function (d.f) f from a random sample $X_1, X_2, ..., X_n$ is defined by*

$$\widehat{f_n}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h} \right), \ x \in \Re, \tag{3.1}$$

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

Definition 3.2. *The kernel distribution estimator (KDE) for a true distribution function (c.d.f) F from a random sample $X_1, X_2, ..., X_n$ is defined by*

$$\widehat{F_n}(x) = \int_{-\infty}^{\infty} \widehat{f_n}(x) dx = \int_{-\infty}^{\infty} \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h} \right) dx. \tag{3.2}$$

We will compare the *KDE* estimator and other estimators in the next section.

4. Applications

4.1. Kernel Estimation Method

Now we apply the four estimation methods to the hurricane example.

1) Pareto distribution in *(2.1)* with Hill estimator $\widehat{\alpha}_{Hill}$ in *(2.9)*;

2) Truncated Pareto distribution in *(2.3)* with Aban MLE estimator $\widehat{\alpha}_{Aban}$ in *(2.10);*

3) Truncated Pareto distribution in *(2.3)* with Moment estimator $\widehat{\alpha}_{Moment}$ in *(2.11);*

4) Kernel density estimator and kernel distribution estimator in *(3.1)* and *(3.2)*, where we use a standard normal kernel

$$K(t) = \frac{1}{\sqrt{2\pi}} e^{\left(-x^2/2\right)}, \quad -\infty < t < \infty; \tag{4.1}$$

and optimal bandwidth (Silverman, 1986, p.40)

$$h_{opt} = \left\{ \int t^2 K(t) dt \right\}^{-2/5} \left\{ \int \left(K(t)\right)^2 dt \right\}^{1/5} \left\{ \int \left(f''(x)\right)^2 dx \right\}^{-1/5} n^{-1/5} = 1.23899, \tag{4.2}$$

where $K(t)$ and $f(x)$ are given in *(4.1)* and *(2.1)* respectively.

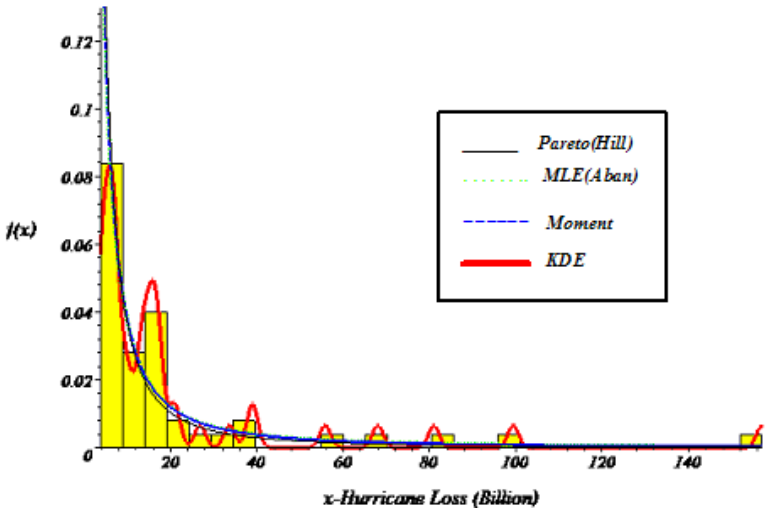

Figure 4.1. Histogram with four estimated density curves of hurricane loss data. The red circles are the data; the black straight line is the original Pareto distribution; the green dot line is the MLE estimated truncated Pareto distribution; the blue dash line is the moment estimated truncated Pareto distribution; the thick red line is the kernel estimated distribution.

Figure 4.1 displays the histogram of hurricane losses with four estimated densities. We note that the KDE density fits the histogram well, including the multiple modes. The other three parametric estimated Pareto and truncated density curves miss some modes.

Table 4.1 gives $\widehat{\alpha}$, $\widehat{\mu}$, Median, 5% Value-at-Risk (VaR) and 1% VaR of each of the four estimation methods. We note that the KDE method gives the smallest mean, largest median, smallest 5% VaR and 2nd largest 1% VaR.

Table 4.1 Four Estimation Methods for Hurricane Example

Method	$\widehat{\alpha}$	$\widehat{\mu}$	Median	5% VaR	1% 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
KDE	-	20.48 billion	11.83 billion	81.16 billion	157.04 billion

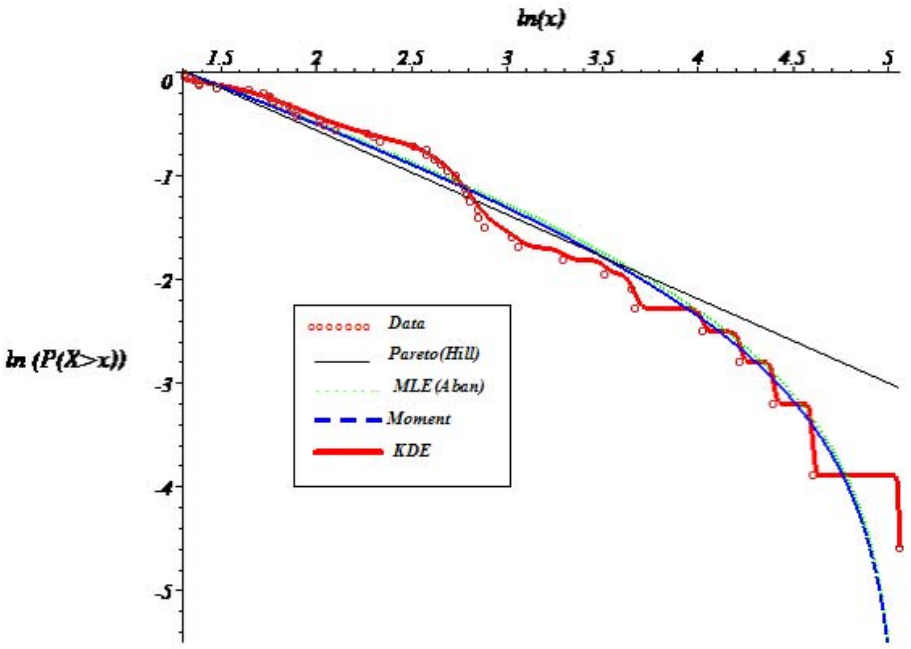

Figure 4.2. Log-log plot of hurricane example with estimated distribution curves. The red circles are the data; the black straight line is the original Pareto distribution; the green dot line is the MLE estimated truncated Pareto distribution; the blue dash line is the moment estimated truncated Pareto distribution; the thick red line is the kernel estimated distribution.

Figure 4.2 exhibits data and four estimated log-log 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 the data well in the right tail, but not so well in the smaller or middle values data. The kernel distribution estimator overcomes this problem, and has the best fitting to the data over the whole range. Figure 4.2 suggests a single distribution may not totally represent how natural data is distributed, and the estimated distribution by the kernel nonparametric method realistically represents the data.

The result in Figure 4.2 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 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 *(2.1)* or truncated Pareto distribution in *(2.3),* or the KDE method in *(3.2).*

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 by using:

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) Kernel Distribution Estimator (KDE) in *(3.2).*

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,

$$T = \sup_x |F^*(x) - S_n(x)|, \quad -\infty < x < \infty, \tag{4.3}$$

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

(2) **Anderson and Darling (A-D)** (1952). This test 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,

$$W_n^2 = n \int_{-\infty}^{\infty} [S_n(x) - F^*(x)]^2 \, \psi\left(F^*(x)\right) dF, \quad \text{where} \quad \psi(u) = \frac{1}{u(1-u)}. \tag{4.4}$$

(3) **Cramer-von Mises (C-v-M)** (Anderson and Darling, 1952). This test proposed using $\psi(u) = 1$ in *(4.2)*. Thus under H_0 the test statistic is given by

$$n\omega^2 = \frac{1}{12n} + \sum_{j=1}^{n} \left(u_j - \frac{2j-1}{2n} \right)^2 . \qquad (4.5)$$

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$_{(Hill)}$	0.1340	0.2900	2.7141	0.0383	0.2057	0.2568
MLE$_{(Aban)}$	0.0948	0.6282	2.3126	0.0622	0.0964	0.6030
Moment	0.1053	0.5308	2.3672	0.0582	0.1095	0.5402
KDE	**0.0689**	**0.8408**	**0.2009**	**0.9902**	**0.0177**	**0.9987**

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 KDE method has the smallest test statistics (i.e., smallest errors) and the largest p-values in each of the three tests respectively (we highlighted the values as bold in the table). This implies 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$_{(Hill)}$	0.1340	0.0584	0.0584	0.0032	0.0027	0.0027
MLE$_{(Aban)}$	0.0948	0.0839	0.0832	0.0020	0.0018	0.0016
Moment	0.1053	0.0738	0.0737	0.0019	0.0015	0.0013
KDE	**0.0689**	**0.0326**	**0.0172**	**0.0005**	**0.0003**	**0.0003**

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

$$AE = \sup_x |F^*(x) - S_n(x)|, \quad -\infty < x < \infty, \qquad (4.6)$$

$$IE = \frac{1}{(X_{n:n} - X_{n-r+1:n})} \left[\int_{X_{n-r+1,n}}^{X_{n,n}} (S_n(x) - F^*(x))^2 dx \right]^{1/2} . \qquad (4.7)$$

Table 4.4 gives absolute errors and integrated errors of the four estimation methods in $r = 49, 18, 10$ cases. We note that the KDE method has the smallest errors in all 6 cases (we highlighted the values as bold in the table). This implies the nonparametric kernel distribution estimator (KDE) 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, 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, 1163-1174.*

[6] Huang, M. L. and Zhao, K. 2013. A Weighted Estimation for risk models, *ISRN Probability and Statistics, in press.*

[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] Silverman, B. W. 1986. *Density Estimation for Statistics and Data Analysis,* Chapman and Hall, New York.

[10] U.S. National Hurricane Center, 2008. *Hurricane History*, on the Website: http://www.nhc.noaa.gov/HAW2/english/history.shtml.