Journal of Applied Sciences 8 (6): 1118-1121, 2008
ISSN 1812-5654

Parameter Estimation for the Heavy Tailed Distributions with the Empirical Distribution

H. Jabbari Khamnei, H. Bevrani and A.A. Haydari
Department of Statistics, Faculty of Mathematical Science,
University of Tabriz and Research Institute for Fundamental Science, Tabriz, Iran

Abstract: One of the most challenging characteristics of heavy tailed distribution is their high or even infinite variance or burstiness. Heavy-tailed distributions are the distributions, the tails of which cannot be cut off. So, we cannot neglect the large-scale but rare events. The most evident and natural problem of statistical estimation connected with heavy tailed distributions are the problem of their parameter estimation. For estimating their parameter, known statistical procedures can be used, for example, estimation methods of Hill, de Haan and others. This research addresses a new method for estimating the parameters of heavy tailed distributions based on the empirical distribution and, using the list squares method. To demonstrate the ability of the proposed approach, it is applied to famous Parteo and Student distributions as two different examples. The related numerical results are reported.

Key words: Empirical distribution, Glivenko theorem, least squares method

INTRODUCTION

Let F(x) be the distribution function of random variable X. Then F(x) is said to be heavy tailed distributions class, if there exists positive constants C and $0 < \gamma \leq 2$ such that:

$$\lim_{x\to\infty} x^{\gamma}\bar{F}(x) = C \tag{1}$$

Where:
$\bar{F}(x) = 1 - F(x)$.

One of the most challenging characteristics of heavy tailed distribution is its high or even infinite variance or burstiness. This guarantees the condition that $0 < \gamma \leq 2$. Heavy tailed distributions have been observed in many natural phenomena including hydrology, geology, electrotechnology, informatics, physics, insurance, they describe the behavior of financial markets (Mandelbort, 1982; Hosking and Wallis, 1987; Brockwell and Davis, 1991; Janicki and Weron, 1994; Adler *et al*., 1998; Nikias and Shao, 1995). When analyzing statistical dependences we often neglect the probabilities of large-scale events on the tail of distribution.

Heavy tailed distributions are the distributions, the tail of which cannot be cut off. So, we cannot neglect the large-scale but rare events.

Further we will investigate Pareto and Student distributions, as heavy tailed distributions class representatives.

Pareto distribution: We know that Pareto density function is given by:

$$p_{\alpha,\beta}(x) = \frac{\alpha\beta^{\alpha}}{x^{\alpha+1}}, \quad x > \beta,\ \alpha > 0. \tag{2}$$

Its distribution function has the form:

$$F(x) = 1 - \left(\frac{\beta}{x}\right)^{\alpha} \tag{3}$$

In case $\beta = 1$ this distribution is considered to be the distribution with one parameter. Mean and variance of this distribution are:

$$E(X) = \frac{\alpha\beta}{\alpha-1},\ \alpha > 1;$$
$$Var(X) = \frac{\alpha\beta^2}{(\alpha-2)} - \frac{\alpha\beta}{(\alpha-1)^2},\ \alpha > 2.$$

The distribution has infinite mean and variance if $\alpha \leq 1$ and $\alpha \leq 2$, respectively. Especially in case $\alpha \leq 2$, when variance is very high, Pareto distribution is a heavy tailed distribution. Comparing Eq. 1 and 3 we realize that

Corresponding Author: H. Jabbari Khamnei, Department of Statistics, Faculty of Mathematical Science, University of Tabriz and Research Institute for Fundamental Science, Tabriz, Iran Tel: +98(0411)339 2844

heavy tailed distributions and Pareto distribution are similar. That is why we often use the first definition for heavy tailed distribution.

Student distribution: A random variable X has a Student distribution with probability mass function:

$$p_\gamma(x) = \frac{\Gamma((\gamma+1)/2)}{\sqrt{\pi\gamma}\,\Gamma(\gamma/2)}(1+x^2/\gamma)^{-\frac{\gamma+1}{2}}, \quad -\infty < x < \infty. \tag{4}$$

It's not complicated to realize that for the Student distribution with parameter γ moments of degrees $\delta \geq \gamma$ do not exist.

Moreover if Y_γ is random value with Student's distribution and $x \to \infty$ then:

$$P(|Y_\gamma| > x) \approx \frac{C_{(\gamma)}}{x^\gamma}, \tag{5}$$

Where:

$$C_{(\gamma)} = \frac{2\Gamma((\gamma+1)/2)\gamma^{\frac{\gamma}{2}-1}}{\sqrt{\pi}\,\Gamma(\gamma/2)}$$

Equation 5 satisfies the heavy tailed distributions definition. That is why Student distribution belongs to heavy tailed distributions class. The parameter γ is the index of tail gravity.

We pay attention that $x \to \infty$ only when $\gamma \leq 2$ exist. This condition exactly is equivalent to infinite variance.

The most evident and natural problem of statistical estimation connected with heavy tailed distributions is the problem of parameter γ estimation. For estimating the parameter known procedures can be used, for example, estimations of Hill (1975), de Haan (1970) and others (Hosking and Wallis, 1987). The mentioned estimation methods, as usual, are calculated easier than maximum likelihood or moments estimation methods. In this study we shall consider the estimation of using the least squares method and empirical distribution. Furthermore, an unbiased estimation is introduced for $\frac{\hat{1}}{\gamma_k}$.

ESTIMATION OF γ USING EMPIRICAL DISTRIBUTION

Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a heavy tailed distribution. Let $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ be the corresponding order statistics and $F_n(x)$ be the appropriate empirical distribution function. Than $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i < x), -\infty < x < \infty.$

According to Glivenko theorem for large n

$$p(X_i > x) \approx 1 - F_n(x) \tag{6}$$

It's obvious that:

$$F_n(X_{(i)}) = \frac{i-1}{n}, \quad i = 1, \ldots, n. \tag{7}$$

Comparing 4, 6 and 7, we note, that for sufficiently large i, the equations

$$\frac{C_{(\gamma)}}{X_{(i)}^\gamma} \approx \frac{n-i+1}{n}$$

or

$$\log(C_{(\gamma)}) - \gamma \log(X_{(i)}) \approx \log(\frac{n-i+1}{n}), \tag{8a}$$

or

$$\log(X_{(i)}) \approx \frac{\log(C_{(\gamma)})}{\gamma} - \frac{1}{\gamma}\log(\frac{n-i+1}{n}), \tag{8b}$$

are fulfilled.

It's natural to consider that Eq. 8a and b are fulfilled for all i, that are greater or equal than some k, so, for $i \geq k$. Using Eq. 8a and b for $i = k, \ldots, n$, we'll calculate estimation $\hat{\gamma}_k$ of γ using condition

$$\hat{\gamma}_k = \arg\min_\gamma \sum_{i=k}^{n} [\log C_{(\gamma)} - \gamma \log(X_{(i)}) - \log(\frac{n-i+1}{n})]^2 \tag{9a}$$

or

$$\hat{\gamma}_k = \arg\min_\gamma \sum_{i=k}^{n} [\log X_{(i)} - \frac{\log(C_{(\gamma)})}{\gamma} + \frac{1}{\gamma}\log(\frac{n-i+1}{n})]^2 \tag{9b}$$

Conditions Eq. 9a and b represent the two-parametered model of the least squares. The solution of Eq. 9a is the estimation:

$$\hat{\gamma}_k = -\frac{(n-k+1)\sum_{i=k}^{n} \log X_{(i)} \log(\frac{n-i+1}{n}) - \sum_{i=k}^{n} \log X_{(i)} \sum_{i=k}^{n} \log(\frac{n-i+1}{n})}{(n-k+1)\sum_{i=k}^{n} (\log X_{(i)})^2 - \left(\sum_{i=k}^{n} \log X_{(i)}\right)^2} \tag{10a}$$

and for (9b) the estimation is

$$\frac{\hat{1}}{\gamma_k} = \frac{(n-k+1)\sum_{i=k}^{n}\log X_{(i)}\log(\frac{n-i+1}{n}) - \sum_{i=k}^{n}\log X_{(i)}\sum_{i=k}^{n}\log(\frac{n-i+1}{n})}{(n-k+1)\sum_{i=k}^{n}(\log(\frac{n-i+1}{n}))^2 - \left(\sum_{i=k}^{n}\log(\frac{n-i+1}{n})\right)^2} \qquad (10b)$$

Using the second estimation Eq. 10b we can calculate the unbiased estimation for $\frac{1}{\gamma_k}$. According the feature of order statistics (Arnold *et al*., 1992):

$$E(\log X_{(i)}) = \sum_{r=1}^{i}\frac{E(\log X_i)}{n-r+1} = E(\log X_1)\sum_{r=1}^{i}\frac{1}{n-r+1}$$

but mean log X_i is:

$$E(\log X_{(i)}) = \frac{1+\log C}{\gamma}$$

That is why unbiased estimation for $\frac{1}{\gamma_k}$ is:

$$\frac{\hat{\hat{1}}}{\gamma_k} = \frac{\hat{1}}{\gamma_k}\frac{(n-k+1)\sum_{i=k}^{n}(\log(\frac{n-i+1}{n}))^2 - \left(\sum_{i=k}^{n}\log(\frac{n-i+1}{n})\right)^2}{(n-k+1)\sum_{i=k}^{n}\sum_{r=1}^{i}\frac{1}{n-r+1}(\log(\frac{n-i+1}{n}) - \sum_{i=k}^{n}\log(\frac{n-i+1}{n}))(1+\log\hat{C})} \qquad (10c)$$

When the values of the argument of function are moderate, the function is not a hyperbola. That is why insertion to Eq. 10a-c order statistics with small or moderate numbers can strongly distort the final result. But as two coefficients of the linear model Eq. 9a and b depend on unknown parameter γ, the condition Eq. 9a and b is redundant. This fact allows us to use the estimation of the second coefficient $C_{(\gamma)}$ which we calculate using conditions Eq. 9a and b, for optimal k value. This is the number of order statistics, used for calculating parameter γ estimation from conditions Eq. 8a and b. Exactly the value of $C_{(\gamma)}$ estimated by the least squares method using conditions Eq. 9a and b is:

$$\hat{C}_k = \exp\{\frac{1}{n-k+1}[\sum_{i=k}^{n}\log(\frac{n-i+1}{n}) + \hat{\gamma}_k\sum_{i=k}^{n}\log X_{(i)}]\} \qquad (11a)$$

and

$$\hat{C}_k = \exp\{\frac{\hat{\gamma}}{n-k+1}[\sum_{i=k}^{n}\log X_{(i)} + \frac{\hat{1}}{\gamma_k}\sum_{i=k}^{n}\log(\frac{n-i+1}{n})]\} \qquad (11b)$$

RESULTS ANALYSIS

In this research, we have estimated the parameter of heavy tailed distribution using empirical distribution. Results of these estimation are given in the Fig. 1a-c for Student distribution and the Fig. 2a-c for Pareto distributions.

We have made the most accurate estimations of parameter when γ ≤ 2. For clearness there are graphics of

$$\bar{\hat{\gamma}} \text{ and } \bar{\hat{\gamma}} \pm S_\gamma,$$

Where:
S_γ = Deviation.

For Student distribution when γ > 2 we get more accurate results.

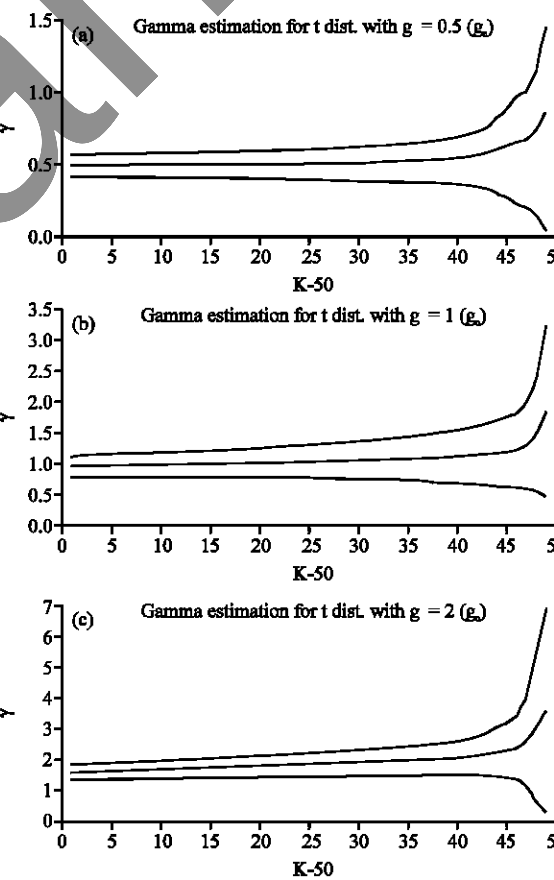

Fig. 1: $\bar{\hat{\gamma}}$ and $\bar{\hat{\gamma}}$ $\pm S_\gamma$ for student distribution, where (a) γ = 0.5, (b) γ = 1 and (c) γ = 2

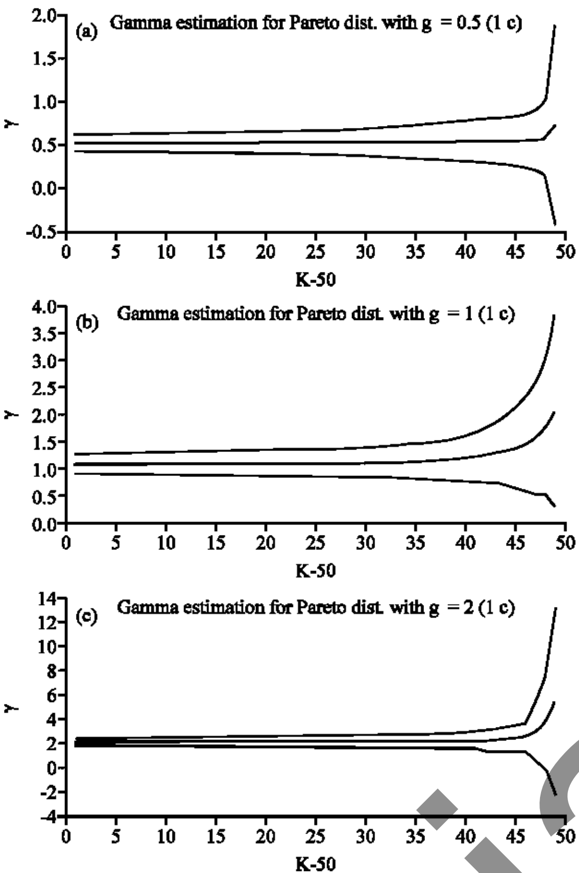

Fig. 2: $\bar{\gamma}$ and $\bar{\gamma}$ $\pm S_{\gamma}$ for Pareto distribution, where (a) $\gamma = 0.5$, (b) $\gamma = 1$ and (c) $\gamma = 2$

When γ is increasing, the accuracy of estimations abruptly falls. In contrast with Student distribution estimations made by mentioned methods for Pareto distribution will be accurate also for $\gamma > 2$. For Student distribution this is not valid. Because, in this case ($\gamma > 2$) the related distribution can not be considered as a heavy tailed distribution.

We can also use our method not only for estimating parameter of Student and Pareto distributions but also for other distributions, which belong to the heavy tailed distributions class.

ACKNOWLEGMENT

This research has been supported by the Research Institute for Fundamental Science, Tabriz, Iran. The authors would like to thank this support.

REFERENCES

Adler, R., R. Feldman and M.S. Taqqu, 1998. A Practical Guide to Heavy Tails: Statistical Techniques and Applications. Chap. No. 1. Birkhauser, Basel.

Arnold, C., N. Balakrishnan and H.N. Nagaraja, 1992. A First Course in Order Statistics. Chap. No. 1. Wiley, New York.

Brockwell, P. and R. Davis, 1991. Time Series: Theory and Methods. 2nd Edn. Springer, New York, pp: 535-552.

de Haan, L., 1970. On regular variation and its application to the weak convergence of sample extremes. Mathematical center tracts. Mathematisch Centrum. Amsterdam, 32: 104.

Hill, B., 1975. A simple general approach to inference about the tail of a distribution. Ann. Stat., 3: 1163-1174.

Hosking, J. and J. Wallis, 1987. Parameter and quantile estimation for the generalized Pareto distribution. Technometrics, 29: 339-349.

Janicki, A. and A. Weron, 1994. Can one see alpha stable variables and processes? Stat. Sci., 9: 109-126.

Mandelbort, B.B., 1982. The Fractal Geometry of Nature. Freeman, W.H. (Ed.). New York, pp: 150.

Nikias, C. and M. Shao, 1995. Signal Processing with Alpha Stable Distributions and Applications. Chap No. 1, pp: 4-10. Wiley, New York.