Journal of Modern Applied Statistical Methods
2025, Vol.24, No.1,

*Copyright © 2024 JMASM*
*ISSN 1538 – 9472*

# Estimating Extreme Quantiles of Unknown Distributions using the Peak Over Thresholds Method[1]

**Mahfuza KHATUN[a*], Sikandar SIDDIQUI[b*]**

[a)] Jahangirnagar University, Savar, 1342 Dhaka, Bangladesh.
[b)] Deloitte Audit Analytics GmbH, Europa-Allee 91, 60486 Frankfurt, Germany.

The purpose of this paper is to present an analytically easy to use procedure for estimating of extreme quantiles of continuous random variables using the Peak Over Threshold approach, and a statistically sound approach to the problem of threshold selection that needs to be resolved in this context. A web link included in the text points to a ready-to-use implementation of the proposed method in the popular programming language Python.

**Keywords**: Extreme Value Theory, Peak over Threshold approach, L-moments

**JEL Classification**: C46, C58

[1] Corresponding author, **Sikandar Siddiqui** – siddiqui@web.de

## 1. Introduction

Estimating the probability of very rare events is an important task in many scientific fields, including hydrology, climate science, engineering, insurance mathematics, and finance. The origins of the corresponding sub-field of statistics, named Extreme Value Analysis (EVA) go back more than 100 years, and a large variety of related methods exist and have been discussed in several specialist books and articles, including (but by no means limited to) the works by Dixon (1950), Gumbel (1935), Pickands (1975), Castillo (1988), Smith (1989), Hosking (1990), Coles (2001) Beirlant et al (2004), Castillo et al. (2005), McNeil et al. (2005), and Reiss and Thomas (2007). A key advantage of EVA is that, in general, the original parent distribution of the random variable under investigation does not have to be known, because the distribution of the extremes approaches a known distribution as the sample size goes to Infinity.

EVA is often performed by expressing the conditional distribution of the data points located beyond a given threshold as a mathematical function of a limited number of unknown parameters, which then need to be estimated numerically, either by applying the method of Maximum Likelihood or by using a Weighted Generalized Method of Moments approach. This is particularly true when considering that the choice of the threshold value beyond which an observation is considered "extreme", must also be made with computational means. Against this background, the purpose of this paper is to present an analytically tractable, statistically well-founded solution approach to the problem of threshold determination and the issue of parameter estimation. The proposed solution is based on a set of statistics termed L-moments, each of which constitutes a specific linear combination of order statistics of the underlying

*The project underlying this publication was funded by the German Federal Ministry of Economic Affairs and Climate Action under project funding reference number 01MK21002G.

variable. The proposed procedure is based on earlier work by Hosking (1990), Hamdan (2009) and Simkova (2017), who have shown that such statistics can be used to estimate extreme quantiles of a univariate, continuous, real-valued random variable *X* based on a sample of identically and independently distributed realizations $x_1, \ldots, x_N$, in an analytically convenient and tractable manner. The remainder of this paper is organized as follows: Section 2 describes a simple estimator of the cumulative distribution function of a continuous, univariate random variable, which is based solely on the available sample of actual observations and does not require the introduction of any further, potentially restrictive, assumptions. Section 3 summarizes some basic properties of the Peak Over Thresholds approach, a popular method for estimating the probability of extreme events on which the following considerations are based. In Section 4, details of the parameter estimation and threshold selection procedures are presented. The feasibility of the proposed approach is demonstrated in Section 5. Section 6 concludes.

## 1. Estimating the cumulative distribution function of X

Let $x_i^{(s)}$, $i = 1, \ldots N$, denote the *i*-th order statistic (i.e. the *i*-th smallest realisation of *X*) in the sample. Then, a consistent nonparametric estimate of the unknown cumulative distribution function $\hat{F}\left(x_i^{(s)}\right)$ at point $x_i^{(s)}$ can be calculated as

$$\hat{F}\left(x_i^{(s)}\right) = \frac{\sum_{j=1}^{N} I\left(x_j^{(s)} \leq x_i^{(s)}\right)}{N+1}$$

(1)

(see, e.g., Maakonen, 2005), where $I(\cdot)$ is an indicator function that is set to 1 if the condition in brackets is fulfilled and to zero otherwise. The denominator in (1) has been set to $(N + 1)$ rather than *N* in order to keep the estimates for all observed data points in the interior of the interval [0, 1].

Moreover, estimates of the unknown cumulative distribution for arbitrary values $x$ of *X* can be calculated by means of log-linear interpolation between adjacent observations as follows:

$$\hat{F}_X(x) = \begin{cases} \hat{F}\left(x_i^{(s)}\right) \cdot \exp\left(\omega_1\left(x - x_1^{(s)}\right)\right) & if\ x < x_1^{(s)} \\ \hat{F}\left(x_i^{(s)}\right) \cdot \exp\left(\omega_i\left(x - x_i^{(s)}\right)\right) & if\ x_i^{(s)} \leq x < x_{i+1}^{(s)} \\ 1 - \left[1 - \hat{F}\left(x_N^{(s)}\right)\right] \cdot \exp\left(\omega_N\left(x - x_N^{(s)}\right)\right) & if\ x \geq x_N^{(s)} \end{cases} \tag{2}$$

$$\text{with} \quad \omega_i := \begin{cases} \dfrac{\ln \hat{F}\left(x_{i+1}^{(s)}\right) - ln\hat{F}\left(x_i^{(s)}\right)}{x_{i+1}^{(s)} - x_i^{(s)}} & \text{if i} = 1, \ldots, \text{N} - 1 \\ \dfrac{\ln\left[1 - \hat{F}\left(x_N^{(s)}\right)\right] - ln\left[1 - \hat{F}\left(x_{N-1}^{(s)}\right)\right]}{x_N^{(s)} - x_{N-1}^{(s)}} & \text{if } i = \text{N} \end{cases}$$

A difficulty associated with the estimation procedure sketched above is that whenever values near the upper and lower ends of the range of *X* are observed only rarely, substantial uncertainty prevails as to the true profile of the cumulative distribution in these regions. This is of high relevance whenever it is precisely the likelihood of extreme events, for which few or no data on historical precedents may be available, that is to be estimated.

# Estimating Extreme Quantiles of Unknown Distributions using the Peak Over Thresholds Method[1]

.

## 2. Estimating the Likelihood of Extreme Events

Reliably estimating extreme quantiles of an unknown distribution based on a finite sample is by no means trivial, given that the particular regions of the range of the related variable where such extreme values tend to be located are often only thinly populated with data points, and that the possibility of observations that lie beyond the observed sample extrema often needs to be taken into consideration.

Among practitioners, the Peaks Over Threshold (POT) method has become the most popular solution approach to this problem. The POT method models the distribution of the excesses over a given size threshold by the Generalized Pareto Distribution (GPD); see Pickands (1975).

The theoretical basis for the POT method is the Pickands-Balkema-de Haan theorem (see Balkema and de Haan, 1974, and Pickands, 1975). It deals with the distribution of the excess ($Y := X - \theta$) of a random variable $X$ over a threshold $\theta$ conditional on the $X$ exceeding $\theta$. This conditional excess distribution can be expressed as

$$\begin{aligned} F_Y(y) &= \Pr(X - \theta \leq y \mid X > \theta) \\ &= \frac{F_X(y+\theta) - F_X(\theta)}{1 - F_X(\theta)} \end{aligned} \tag{3}$$

The authors show that for a large variety of continuous cumulative distribution functions $F_X$, given a sufficiently high value of the threshold parameter $\theta$ and a sufficiently large sample size, the above distribution is well approximated by a Generalized Pareto Distribution (GPD). The GPD is a continuous probability distribution characterized by three parameters: location (μ), scale (σ), and shape (ξ); its cumulative distribution function reads

$$\Psi\ (y;\ \mu, \sigma, \zeta) = \begin{cases} 1 - \left[1 + \zeta \cdot \frac{y-\mu}{\sigma}\right]^{-\frac{1}{\zeta}}\ \ if\ \zeta \neq 0 \\ 1 - exp\left(-\frac{y-\mu}{\sigma}\right)\ \ if\ \zeta = 0 \end{cases}, \tag{4}$$

while the corresponding probability density function is

$$\psi\ (y;\ \mu, \sigma, \zeta) = \begin{cases} \frac{1}{\sigma}\left[1 + \zeta\left(\frac{y-\mu}{\sigma}\right)\right]^{-\left(\frac{1}{\zeta}+1\right)}\ \ if\ \zeta \neq 0 \\ \quad \frac{1}{\sigma} exp\left(-\frac{y-\mu}{\sigma}\right)\ \ if\ \zeta = 0 \end{cases}$$

(5)

and the related quantile function equals

$$\Psi^{-1}\ (u;\ \mu, \sigma, \zeta) = \begin{cases} \mu + \frac{\sigma}{\zeta}(u^{-\zeta} - 1)\ \ if\ \zeta \neq 0 \\ \mu - \sigma ln(u)\ \ if\ \zeta = 0 \end{cases} \tag{6}$$

## 3. Parameter Estimation for the Generalized Pareto Distribution

### 4.1 Preliminaries: L-Moments

*Description*

Traditionally, central moments such as the mean, the variance, skewness and kurtosis, have been used to describe univariate distributions. An alternative to this way of proceeding was introduced by Hosking (1990). This alternative approach is based on L-moments, which are certain linear combinations of order statistics. According to Šimková (2017), their main advantage, as compared to "conventional" central moments, is that if the mean of the underlying distribution is finite, their existence of L-moments all orders can be taken for granted.

*Population L-Moments*

For a random variable *X*, the Hosking (1990) defines the r-th population L-moment as

$$\lambda_r = \frac{1}{r} \cdot \sum_{k=0}^{r-1} (-1)^k \binom{r-1}{k} E(Y_{r-k:k}),$$

$$\text{with } \binom{a}{b} := \frac{a!}{b!\,(a-b)!}, \quad m! := \begin{cases} 1 \; if \; m = 0 \\ 1 \cdot 2 \cdot \ldots m \; if \; m > 0 \end{cases} \tag{7}$$

and $Y_{m:n} :=$ the m-th smallest value in an independent sample of size n from the distribution of *Y*.

*Sample L-Moments*

Sample L-Moments can probably be computed most efficiently by drawing on the concept of probability-weighted moments introduced by Greenwood et al. (1979). Sample probability weighted moments, computed from order statistics $y_i^{(s)}$, *i* = 1,…*N*, are given by

$$b_0^{\square} := N^{-1} \sum_{i=1}^{N} y_i^{(s)} \tag{8}$$

and

$$b_r^{\square} := N^{-1} \sum_{j=r+1}^{N} \frac{(j-1)\cdot(j-2)\cdot\ldots\cdot(j-r)}{(N-1)\cdot(N-2)\cdot\ldots\cdot(N-r)} y_j^{(s)} \tag{9}$$

The sample L-moments are linear combinations of the sample probability weighted moments and can be calculated as follows (see Hamdan, 2009, p. 80):

- The first-order sample L-moment equals $\hat{\lambda}_1 = b_0^{\square}$.
- The higher-order sample L-moments are given by $\hat{\lambda}_{r+1} = \sum_{k=0}^{r} p_{r,k}^{*}\, b_k$, with $p_{r,k}^{*} := (-1)^{r-k} \binom{r}{k} \binom{r+k}{k}$

By dividing the higher-order L-moments by the dispersion measure $\hat{\lambda}_2$, we obtain the L-moment ratios,

$$\hat{\tau}_r := \hat{\lambda}_r / \hat{\lambda}_2, \tag{10}$$

which are dimensionless, i.e. independent of the units in which the data have been measured, with $\tau_3$ being a measure of skewness and $\tau_4$ is a measure of kurtosis.

### 4.2 Deriving GPD parameter estimates from sample L-moments

In line with Hosking (1990), the sample L-moments obtained as above can be used to calculate estimates of the GPD distribution as follows:

$$\hat{\zeta} := \frac{3\hat{\tau}_3 - 1}{1 + \hat{\tau}_3} \tag{11}$$

$$\hat{\sigma} \;\coloneqq\; \left(1-\hat{\zeta}\right)\cdot\left(2-\hat{\zeta}\right)\cdot\hat{\lambda}_2 \tag{12}$$

$$\hat{\mu} \;\coloneqq\; \hat{\lambda}_1-\left(2-\hat{\zeta}\right)\cdot\hat{\lambda}_2 \tag{13}$$

The resulting estimate of the cumulative distribution function GPD distribution then reads

$$\hat{\Psi}\,(y;\,.) = \begin{cases} 1-\left[1+\hat{\zeta}\cdot\frac{y-\hat{\mu}}{\hat{\sigma}}\right]^{-\frac{1}{\hat{\zeta}}} & if\ \hat{\zeta}\neq 0 \\ 1-exp\left(-\frac{y-\hat{\mu}}{\hat{\sigma}}\right) & if\ \hat{\zeta}=0 \end{cases}, \tag{14}$$

### 4.3 Setting the Threshold Parameter

*Motivation*

The problem of choosing the threshold parameter $\theta$ involves a tradeoff between the objectives of (a) setting the threshold sufficiently high for the theoretical preconditions for the GPD method to apply, and (b) setting it low enough to obtain a number of threshold exceedances that is so large that it allows for reliable estimates.

*Decision Rule*

A decision rule by which this tradeoff can be resolved can be set by first specifying a confidence level $\alpha$, which choices determines the range of values in which the GPD-based estimate of the cumulative distribution function must lie for each point above the threshold for the overall estimate (14) to be acceptable.

- A comparably high value of $\alpha$ will allow for relatively substantial deviations between the nonparametric estimates and its GPD-based counterparts to be accepted, thus resulting in a relatively large risk of a Type 2 error (failure to reject the null hypothesis “GPD estimate = true” if it is actually false).
- On the other hand, a relatively low value of $\alpha$ will only allow comparably small deviations between nonparametric and GPD-based estimates and estimates and therefore imply a relatively large risk of a Type 1 error (rejection of the null hypothesis “GPD estimate = true” that is actually true).

The proposed decision rule for choosing the actual parameter value used for estimation can be summarized as follows:

(i) Specify a dense grid of trial values $\tilde{\theta}_1, \ldots, \tilde{\theta}_J$ for the threshold parameters.

(ii) Set $j$ to 1.

(iii) Set the candidate value of the threshold parameter to $\tilde{\theta}_j$.

(iv) Test, separately for at each data point $x$ above $\tilde{\theta}_j$, whether the GPD-based estimate of the cumulative distribution function of $X$, which reads $\hat{\Psi}\left(x-\tilde{\theta}_j;\,.\right)\cdot\left[1-\hat{F}_X\left(\tilde{\theta}_j\right)\right]+\hat{F}_X\left(\tilde{\theta}_j\right)$, lies inside a two-sided a × 100% confidence interval around its empirical counterpart $\hat{F}_X(x)$

(v) If the condition stated in (iv) has been met for each data point above $\tilde{\theta}_j$, set the actual trial parameter used for estimation, $\theta$, to $\tilde{\theta}_j$, and terminate the threshold selection procedure.

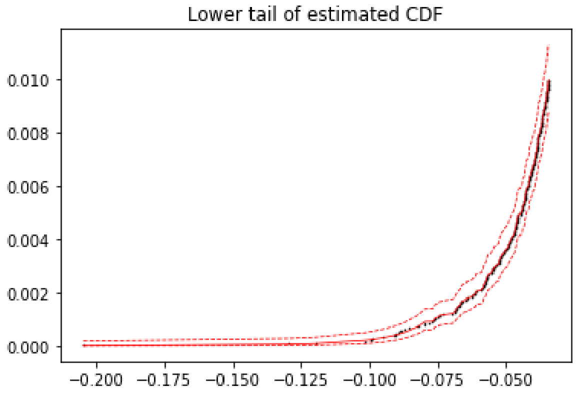
(vi) If not, increase $j$ by 1 and repeat steps (iii) to (v).

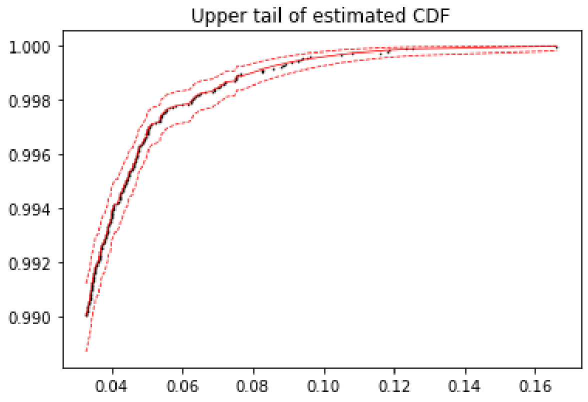
*Confidence intervals*

Pointwise confidence intervals for $\hat{F}_X(x)$ are estimated using Wilson's (1927) Score Interval.

## 4. Application

The example given below shows that the proposed procedure can be successfully applied to the problem of estimating extreme quantiles of the return distribution of the S&P 500 stock market index, using historical data ranging from January 1928 to November 2022 (source: Yahoo! Finance). The p-value below which the null hypothesis that the GPD-based estimates is compatible with the empirical cumulative distribution function is rejected is set to a rather high value (0.20) in order to ensure a rather high degree of agreement between the estimated and empirically observed cumulative distribution in the tails. The threshold selection procedure described in Section 4 led to a lower threshold of -0.07715 and an upper threshold of 0.07543, respectively. A graphical summary of the results obtained is given below:





A csv file with the data in use, together with a commented version of the code for implementing the proposed procedure, prepared in the language *Python,* has been provided on the web at https://drive.google.com/drive/folders/1jfl1w00r0-5pGnJFlyB8MzVkEfUcXwxA?usp=sharing (under the file names: S&P500History.csv and peakOverThreshold.ipnb , respectively).

## 5. Concluding Remarks

The method presented here provides an analytically easy to use procedure for estimating of extreme quantiles of continuous random variables using the Peak Over Threshold approach, along with a statistically sound approach to threshold selection. A weblink included in the text points to a ready-to-use implementation of the proposed method in the popular programming language *Python*.

The weakness of the current approach is that it assumes the data points in use to be identically and independently distributed over time. In many applications to economic, social, and financial data series, this premise is unlikely to be fulfilled. Future research efforts could therefore be directed at integrating extreme value models into a framework capable of capturing complex, possibly nonlinear dependence patterns between successive realisations of a given random variable, and/or allowing for changes in the nature of the data-generating process during the passage of time.

**References:**

Balkema, A., and L. de Haan, 1974. Residual life time at great age. *Annals of Probability* 2, pp. 792–804.

Beirlant, J., Goegebeur, Y., Segers, J., and Teugels, J., 2004. *Statistics of Extremes.* Chichester: Wiley.

Castillo, E., Hadi, A.S., Balakrishnan, N., and Sarabia, J.M., 2005. *Extreme Value and Related Models with Applications in Engineering and Science*. Hoboken, NJ: Wiley.

Coles, S., 2001. *An Introduction to Statistical Modelling of Extreme Values.* Berlin (Springer).

Dixon, W.J., 1950. Analysis of extreme values. *Annals of Mathematical Statistics* 21, pp. 488-506.

Ferreira, A., and de Haan, L., 2015. On the block maxima method in extreme value theory. *The Annals of Statistics* 43 (1), pp. 276–298.

Fisher, R.A., and Tippett, L.H.C., 1928. Limiting forms of the frequency distribution of the largest and smallest member of a sample. *Proceedings of the Cambridge Philosophical Society* 24 (2), pp. 180–190.

Gnedenko, B.V., 1943. Sur la distribution limite du terme maximum d'une serie aleatoire. *Annals of Mathematics* 44 (3), pp. 423–453

Greenwood. J.A., Landwehr, J.M., Matalas, N.C., and Wallis, J.R., 1979. Probability weighted moments: Definition and relation to parameters of several distributions expressable in inverse form. *Water Resources Research* 15 (5), pp. 1049-1054.

Gumbel, E.J., 1935. Les valeurs extrêmes des distributions statistiques. *Annales de l'Institut Henri Poincaré*, 5 (2), pp. 115–158.

Hamdan, M. S., 2009. *The Properties of L-moments Compared to Conventional Moments.* MSc thesis, Department of Mathematics, The Islamic University of Gaza.

Hosking, J. R. M., 1990. L-moments: Analysis and Estimation of Distributions Using Linear Combinations of Order Statistics. *Journal of Royal Statistical Society* (Series B) 52, pp. 105–124.

Makkonen, L., 2005. Plotting Positions in Extreme Value Analysis. *Journal of Applied Meteorology and Climatology 45*, pp. 334-340.

McNeil, A. J., Frey, R., and Embrechts, P., 2005. *Quantitative Risk Management: Concepts, Techniques, and Tools.* Princeton (University Press).

Pickands, J. III, 1975. Statistical Inference using Extreme Order Statistics. *Annals of Statistics* 3, pp. 119–131.

Reiss, R.D., and Thomas, M., 2007. *Statistical Analysis of Extreme Values with Applications to Insurance, Finance Hydrology and Other Fields*. Boston (Birkhäuser).

Šimková, T., 2017. Statistical Inference Based on L-Moments. *Statistika: Statistics and Economy Journal* 97(1), pp. 44-58.

Smith, R.L., 1989. Extreme value analysis of environmental time series: an application in trend detection of ground-level ozone. *Statistical Science* 4, pp. 367-393

Wilson, E. B. 1927. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association* 22 (158), pp. 209–212.