Estimation of the Parameters of Extreme Value Distributions from Truncated Data Via the EM Algorithm

Tewfik Kernane, Zohrh A. Raizah

Estimation of the Parameters of Extreme Value Distributions from Truncated Data Via the EM Algorithm

Tewfik Kernane[1]*and Zohrh A. Raizah[2]

[1] Department of Probability and Statistics,

Faculty of Mathematics,

University of Sciences and Technology USTHB, Algiers, Algeria

tkernane@gmail.com

[2]Department of Mathematics, College of Education,

King Khalid University, Abha, Saudi Arabia

zo.hrh@hotmail.com

Abstract

EM algorithm is used to obtain the maximum likelihood estimates for the parameters of extreme value distributions when the data are truncated. The method is used for the parameters of the type I least extreme value distribution from right truncated data. Using transforms between the different types of extreme value distributions, the algorithm can be used to estimate the parameters of the Type I greatest extreme value distribution from left truncated data, for the two parameters Weibull distribution from right truncated data and for the Fréchet distribution from left truncated data. The algorithm is illustrated with simulated examples.

1 Introduction

Extreme value distributions are popular models in lifetime and reliability analysis where samples are often either truncated or censored. They are also useful in the analysis of environmental data such as rainfall, flood flow, earthquake among others. They approximate distributions of extremes (least or greatest) in large random samples and are more widely known as the Gumbel (type I), Fréchet (type II) and Weibull (type III) distributions. They are also particularly useful in the context of insurance or reinsurance, as they allow to identify extreme events, for the pricing of small or non-working insurance contributions or for constituting sufficient capital to ensure solvency of the company. For a thorough account of the theory of extreme value distributions we refer to the book by Coles [3].

*Corresponding author: tkernane@gmail.com

The data are said to be truncated when measuring devices fail to report observations below and/or above certain readings. For example, truncated data frequently arise in the statistical analysis of astronomical observations (see Efron and Petrosian [5]) and in medical data (see Klein and Zhang [7]), and if the truncation is ignored this can cause considerable bias in the estimation. The estimation of the distribution of losses in insurance field with left truncated data is also useful when it comes to coverage of excess-of-loss reinsurance where the data usually exceeds a certain underlying retention (see Klugman et *al.* [8]). There exists in the literature many approaches of estimation from "incomplete data" such as moment based estimators, maximum likelihood based approach of the EM algorithm (Dempster et *al* [4]) or nonparametric methods (see for example [5]). Samples to be considered in this paper include those that are singly right or singly left truncated. The EM algorithm is a powerful iterative procedure which by repetition fill in the missing data with estimated values to update the parameter estimates.

The type-I least extreme values distribution, which is also called the Gumbel minimum $\mathcal{G}_{\min}(\mu, \sigma)$, is defined by its distribution function as:

$$F(x) = 1 - \exp\left\{-\exp\left[\frac{1}{\sigma}(x-\mu)\right]\right\},$$

for $x \in \mathbb{R}$, and it has the probability density

$$f(x) = \frac{1}{\sigma}\exp\left[\frac{1}{\sigma}(x-\mu)\right]\exp\left\{-\exp\left[\frac{1}{\sigma}(x-\mu)\right]\right\},$$

where $\sigma > 0$ and $\mu \in \mathbb{R}$.

If X is a random variable from a type I greatest extreme values distribution or Gumbel maximum $\mathcal{G}_{\max}(\mu, \sigma)$ with location parameter μ and shape parameter σ then $-X$ follows a type I least extreme values distribution with location parameter $-\mu$ and shape parameter σ [2].

In most cases, an insurance company provides a database that does not contain all of its claims to the reinsurer. It provides information on claims that exceed a threshold of communication (see Chernobai et *al.* [1]). The sample will be truncated to the left and the error on the choice of the adjusted distribution for this sample will then be found throughout the pricing process. Let $w_1, ..., w_{n_u}$ be a left truncated sample from the type I greatest extreme value distribution, where n_u is the known number of untruncated observations. Consider C the cutoff point below which observations are discarded, that is we observe only the observations $w_i \geq C$. Transforming the data we obtain a sample $x_i = -w_i$, $i = 1, ..., n_u$ from a Gumbel minimum $\mathcal{G}_{\min}(\mu, \sigma)$ truncated on the right ($x_i \leq C$). Consider $\underline{y} = (y_1, ..., y_{n_t})'$ the discarded data where n_t is the unknown number of truncated observations which is a random variable. Hence, the missing data is the pair $\left\{\underline{y}, n_t\right\}$. We can say that the complete data, denoted by z, is $z = \left(\underline{x}, n_t, \underline{y}\right)$ such that $\underline{x} = \left\{z_i : z_i \geq C\right\}$, $\underline{y} = \left\{z_i : z_i < C\right\}$ and $n_t = \#\left\{z_i : z_i < C\right\}$.

We shall estimate the Gumbel (minimum) distribution parameters from right truncated data, then the likelihood function for the observed data is

$$L_{obs} = \prod_{i=1}^{n_u} \frac{f(x_i)}{F(C)} \tag{1}$$

where $F(C)$ is the value of the distribution function at the truncation point. The relation (1) will be

$$L_{obs} =_{i=1}^{n_u} \frac{1}{\sigma F(C)} \exp\left\{\frac{x_i-\mu}{\sigma}\right\} \exp\left\{-\exp\left\{\frac{x_i-\mu}{\sigma}\right\}\right\}$$

where $F(C) = 1 - \exp\left\{-\exp\left\{\frac{C-\mu}{\sigma}\right\}\right\}$.

The complete likelihood function is given by

$$L_c(\mu, \sigma; z) =_{i=1}^{n} f(z_i);\ \mu \in \mathbb{R}, \sigma > 0, \tag{2}$$

where $n = n_u + n_t$ the number of complete data. It should be noted that n is unknown since n_t is a random variable following a negative binomial distribution with parameters n_u and $F(C)$ (see McLachlan and Krishnan [10], pp 78-79). It is difficult to estimate the parameters from (2) since it lacks to know z_i completely, for this we will use the Expectation-Maximisation algorithm (EM) which is used generally for incomplete data, it globally recover the missed information from the expectation of the known one.

Our main focus is on the estimation of the parameters of extreme value distributions from truncated data by using the method of the EM algorithm. In his book on truncated and censored samples, Cohen [2] did not treat the case of truncated samples for extreme value distributions and from the best of our knowledge there is no reported work on this subject until now. In section 2, we provide an EM algorithm for the estimation of the parameters of the type I least extreme value distribution (or Gumbel minimum) from right truncated data. Using transforms between the different types of extreme value distributions, the algorithm is used to estimate the parameters of the Type I greatest extreme value distribution (Gumbel maximum) from left truncated data, for the two parameters Weibull distribution from right truncated data and for the Fréchet distribution from left truncated data. In section 3 we present simulated examples.

2 EM Algorithm for truncated data

The EM algorithm is one of the numerical algorithms which helps to compute the maximum likelihood estimations based on missing or latent data. The first reported work about this method was the paper of Dempster et *al.* [4] applied it for censored and truncated data from some distributions and also for mixture of distributions. For a detail review on this method, see the book by McLachlan and Krishnan [10]. For applying this algorithm, we need first to find the expectation of the likelihood function for complete data given by 2, and then compute the estimations by finding the maximum of the expectation of the likelihood.

2.1 The E-step

Consider $\theta = (\mu, \sigma)$ the set of parameters to estimate, then $\theta_{(k)} = \left(\mu_{(k)}, \sigma_{(k)}\right)$ for $k = 0, 1, ...$ the estimations corresponding to the kth step of the algorithm. In this case, the log-likelihood is given by

$$\ln L = -(n_u + n_t)\ln\sigma + \frac{1}{\sigma}\sum_{i=1}^{n_u}(x_i-\mu) - \sum_{i=1}^{n_u}\exp\left[\frac{1}{\sigma}(x_i-\mu)\right] + \frac{1}{\sigma}\sum_{i=1}^{n_t}(y_i-\mu) - \sum_{i=1}^{n_t}\exp\left[\frac{1}{\sigma}(y_i-\mu)\right] \tag{3}$$

Taking the expectation and using Wald's formula we obtain

$$\begin{aligned} E(\ln L) &= -n_u \ln\sigma + E(n_t)\ln\sigma + \frac{1}{\sigma}\sum_{i=1}^{n_u}(x_i-\mu) - \sum_{i=1}^{n_u}\exp\left[\frac{1}{\sigma}(x_i-\mu)\right] - \frac{\mu}{\sigma}E(n_t) \\ &\quad + \frac{1}{\sigma}E(n_t)E(Y_i) - E(n_t)\exp\left(-\frac{\mu}{\sigma}\right)E\left(\exp\frac{Y_i}{\sigma}\right) \end{aligned} \tag{4}$$

Then, using the fact that n_t follows a negative binomial distribution with parameters n_u and $F(C)$,we obtain

$$\begin{aligned} E(\ln L) &= -n_u \ln\sigma + \frac{n_u(1-F(C))}{F(C)}\ln\sigma + \frac{1}{\sigma}\sum_{i=1}^{n_u}(x_i-\mu) - \sum_{i=1}^{n_u}\exp\left[\frac{1}{\sigma}(x_i-\mu)\right] - \frac{\mu}{\sigma}\frac{n_u(1-F(C))}{F(C)} \\ &\quad + \frac{n_u(1-F(C))}{\sigma F(C)}E(Y_i) - \frac{n_u(1-F(C))}{F(C)}\exp\left(-\frac{\mu}{\sigma}\right)E\left(\exp\frac{Y_i}{\sigma}\right). \end{aligned} \tag{5}$$

To compute the quantities $E(Y_i)$ and $E\left(\exp\frac{Y_i}{\sigma}\right)$ we make use of moment generating function of the Gumbel distribution of least extreme values truncated on the right used by Ng et *al.* [9], given by the equations:

$$M_{\frac{Y_i-\mu}{\sigma}}(t) = \exp(\exp(\lambda))\Gamma\left(t+1, e^{\lambda}\right) = \Gamma(t+1)\left[e^{e^{\lambda}} - \sum_{p=0}^{\infty}\frac{e^{(t+p+1)\lambda}}{\Gamma(t+p+2)}\right],$$

where $f(y_i/y_i > C) = \frac{1}{\sigma}\exp(\exp(\lambda))\exp\left[\frac{y_i-\mu}{\sigma} - \exp\left(\frac{y_i-\mu}{\sigma}\right)\right]$; $C < y_i < \infty$, $\lambda = (C-\mu)/\sigma$.

$\Gamma\left(t+1, e^{\lambda}\right)$ is the incomplete gamma function and $\Gamma(t+1)$ the complete gamma function, hence we can deduce the conditional expectations which are the derivatives of the moment generating function at $t = 0$ and given by:

$$\begin{aligned} E(Y_i \mid \lambda, \mu, \sigma) &= E_{1,i}\sigma + \mu, \\ E\left(e^{Y_i/\sigma} \mid \lambda, \mu, \sigma\right) &= e^{\mu/\sigma}\left(e^{\lambda}+1\right), \\ E\left(Y_i e^{Y_i/\sigma} \mid \lambda, \mu, \sigma\right) &= e^{\mu/\sigma}\left[E_{2,i}\sigma + \mu\left(e^{\lambda}+1\right)\right], \end{aligned} \tag{6}$$

where

$$E_{1,i} = \psi(1)\exp(\exp(\lambda)) + \sum_{p=0}^{\infty}\frac{e^{(p+1)\lambda}\psi(p+2)}{\Gamma(p+2)} - [\lambda+\psi(1)]\sum_{p=0}^{\infty}\frac{e^{(p+1)\lambda}}{\Gamma(p+2)},$$

$$E_{2,i} = \psi(2)\exp(\exp(\lambda)) + \sum_{p=0}^{\infty}\frac{e^{(p+2)\lambda}\psi(p+3)}{\Gamma(p+3)} - [\lambda+\psi(2)]\sum_{p=0}^{\infty}\frac{e^{(p+2)\lambda}}{\Gamma(p+3)},$$

and $\psi(1)$ and $\psi(2)$ are the digamma functions obtained as the first and second derivatives of the gamma function at $t = 0$. precisely, $\psi(1) = \frac{d}{dt}\ln[\Gamma(t+1)]\mid_{t=0}$ and $\psi(2) = \frac{d^2}{dt^2}\ln[\Gamma(t+1)]\mid_{t=0}$.

2.2 The M-step

In this step we obtain the estimations from the formulas

$$\mu = \sigma \left[\ln \left(\sum_{i=1}^{n} \exp\left(\frac{z_i}{\sigma}\right) \right) - \ln(n) \right] \tag{7}$$

$$\sigma = \frac{\sum_{i=1}^{n} z_i \exp\left(\frac{z_i}{\sigma}\right)}{\sum_{i=1}^{n} \exp\left(\frac{z_i}{\sigma}\right)} - \frac{\sum_{i=1}^{n} z_i}{n}$$

from which we will obtain the $(k+1)$ iteration of the algorithm. The estimations will be obtained using the fixed point iteration used by Kernane and Raizah [6] as following:

$$\sigma_{(k+1)} = \frac{\sum_{i=1}^{n_u} x_i \exp\left(\frac{x_i}{\sigma_{(k+1)}}\right) + E(n_t)\, E\left(Y_i e^{Y_i/\sigma} \mid \lambda, \mu_{(k)}, \sigma_{(k)}\right)}{\sum_{i=1}^{n} \exp\left(\frac{z_i}{\sigma_{(k+1)}}\right) + E(n_t)\, E\left(e^{Y_i/\sigma} \mid \lambda, \mu_{(k)}, \sigma_{(k)}\right)} - \frac{\sum_{i=1}^{n_u} x_i + E(n_t)\, E\left(Y_i \mid \lambda, \mu_{(k)}, \sigma_{(k)}\right)}{n_u + E(n_t)}. \tag{8}$$

In fact $E(n_t) = E\left(n_t \mid \lambda, \mu_{(k)}, \sigma_{(k)}\right)$, and equation (7) becomes

$$\mu_{(k+1)} = \sigma_{(k+1)} \left[\ln \left(\sum_{i=1}^{n_u} \exp\left(\frac{x_i}{\sigma_{(k+1)}}\right) + E(n_t)\, E\left(e^{Y_i/\sigma} \mid \lambda, \mu_{(k)}, \sigma_{(k)}\right) \right) - \ln\left(n_u + E(n_t)\right) \right], \tag{9}$$

with $E(n_t) = \frac{n_u(1-F(C))}{F(C)}$ the expectation of the random variable n_t which follows a negative binomial distribution. Also $F(C) = 1 - \exp\left[-\exp\left(\frac{C-\mu_{(k)}}{\sigma_{(k)}}\right)\right]$ and $\lambda = \frac{C-\mu_{(k)}}{\sigma_{(k)}}$.

Remark 1 *The two parameters Weibull distribution $\mathcal{W}(\theta, \lambda)$ has the pdf given by*

$$f(x; \theta, \lambda) = \frac{\lambda}{\theta^{\lambda}} x^{\lambda - 1} \exp\left[-(x/\theta)^{\lambda}\right], \ x > 0, \ \theta, \lambda > 0.$$

If X follows a Weibull distribution $\mathcal{W}(\theta, \lambda)$ then $Y = \log X$ follows a Gumbel minimum $\mathcal{G}_{\min}(\mu, \sigma)$ with $\sigma = 1/\theta$ and $\mu = \ln \lambda$. If we have a data truncated on the right from a Weibull distribution $X \rightsquigarrow \mathcal{W}(\theta, \lambda)$ that is $X < C$ observed, then $Y = \ln X$ will follow a type-I least extreme value distribution treated above with data truncated on the right ($Y < \ln C$ observed) and we can use the EM algorithm above to estimate $\mu = \ln \lambda$ and $\sigma = 1/\theta$.

Remark 2 *The two-parameter Fréchet distribution $\mathcal{F}(\delta, \nu)$ or type II extreme value distribution is largely used as a model for extremes of flood and rainfall data. Its probability density function is*

$$f(x) = \frac{\lambda}{\delta} \left(\frac{\delta}{x}\right)^{\nu+1} \exp\left[-\left(\frac{\delta}{x}\right)^{\nu}\right],$$

where $x > 0$, $\delta, \nu > 0$. If a random variable X follows a Fréchet distribution with parameters δ and ν then $Y = \ln(X)$ follows a Gumbel maximum distribution with parameters $\mu = \ln(\delta)$ and $\sigma = 1/\nu$.For the Fréchet distribution $X \rightsquigarrow \mathcal{F}(\delta, \nu)$, if data are truncated on the left $X > C$ then $Y = -\ln X$ follow a type-I least extreme value distribution treated above with data truncated on the right ($Y < -\ln C$ observed) and we can use the EM algorithm above to estimate $\mu = -\ln \delta$ and $\sigma = 1/\nu$.

3 Simulation Example

Example 3 *We generated samples of different sizes* $n = 50, 100, 200, ..., 1000$ *from the extreme value distribution* $\mathcal{G}_{\min}(\mu, \sigma)$ *with* $\mu = 2$ *and* $\sigma = 1.2$. *We choose the threshold point* $C = 2.5$. *By using the EM algorithm explained in this paper we get the results summarized in the following table. For 1000 replications, the table gives the mean estimations* $\widehat{\mu}$ *and* $\widehat{\sigma}$, *the mean square errors of the estimations* $MSE(\widehat{\mu}) = \sum_{i=1}^{1000}(\widehat{\mu}_i - \mu)^2/1000$ *and* $MSE(\widehat{\sigma}) = \sum_{i=1}^{1000}(\widehat{\sigma}_i - \sigma)^2/1000$, *the mean of the untruncated sample sizes* $\overline{n_u}$, *and the variances of the estimations* $Var(\widehat{\mu})$ *and* $Var(\widehat{\sigma})$.

n	$\widehat{\mu}$	$\widehat{\sigma}$	$\overline{n_u}$	$MSE(\widehat{\mu})$	$MSE(\widehat{\sigma})$	$Var(\widehat{\mu})$	$Var(\widehat{\sigma})$
50	2.0428	1.1717	39.139	0.2823	0.0431	0.5299	0.2058
100	2.0843	1.1969	78.039	0.2497	0.0313	0.4928	0.177
200	2.0756	1.2031	156.392	0.1599	0.0165	0.3928	0.1284
300	2.0505	1.1999	234.509	0.108	0.0112	0.325	0.1061
400	2.0204	1.1972	312.343	0.0742	0.0085	0.2717	0.0922
500	2.0114	1.1976	389.92	0.051	0.0061	0.2258	0.0782
600	2.0207	1.1999	468.579	0.0467	0.0059	0.2153	0.0768
700	2.0177	1.2003	545.982	0.0389	0.0049	0.1964	0.0701
800	2.0186	1.2019	623.901	0.0306	0.0041	0.1739	0.0641
900	2.0031	1.1977	702.483	0.0271	0.0036	0.1646	0.0603
1000	2.0119	1.1999	781.331	0.0241	0.0034	0.1549	0.058

References

[1] Chernobai, A., Burnecki, K., Rachev, S., Trück, S. and Weron, R., Modelling catastrophe claims with left-truncated severity distributions, *Computational Statistics 21, pp. 537-555, 2006.*

[2] Cohen, A.C., *Truncated and Censored Samples: Theory and Applications.* Dekker, New York 1991.

[3] Coles, S., *An Introduction to Statistical Modeling of Extreme Values*, Springer-Verlag, 2001.

[4] Dempster, A.P, Laird, N.M. and Rubin, D.B., Maximum likelihood from incomplete data via the EM algorithm, *JRSS, B, 39 (1), pp. 1-38, 1977.*

[5] Efron, B. and Petrosian, V. (1999). Nonparametric methods for doubly truncated data, *J. Am. Stat. Assoc.* Vol. 94. No. 447. pp. 824-834.

[6] Kernane, T. and Raizah, Z. A., Fixed Point Iteration for Estimating the Parameters of Extreme Value Distributions, *Communications in Statistics: Simulation and Computation, 38 (10) pp. 2161-2170, 2009.*

[7] Klein, J. P. and Zhang, M. J. (1996). Statistical challenges in comparing chemotherapy and bone marrow transplantation as a treatment for leukemia, *Lifetime Data: Models in Reliability and Survival Analysis, N.P. Jewel, 175-185.*

[8] Klugman, S.A., Panjer, H.H. and Willmot G.E., *Loss models: from data to decisions,* Wiley, New York, 1998.

[9] Ng, H.K.T.; Chan, P.S.; Balakrishnan, N., Estimation of parameters from progressively censored data using EM algorithm, *Computational Statistics & Data Analysis* **39** (2002) 371 – 386.

[10] McLachlan and Krishnan, *The EM algorithm and Extensions*, Wiley, New York, 2008.