



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

Tewfik Kernane, Zohrh A. Raizah

► To cite this version:

Tewfik Kernane, Zohrh A. Raizah. Estimation of the Parameters of Extreme Value Distributions from Truncated Data Via the EM Algorithm. 2010. hal-00503252v1

HAL Id: hal-00503252

<https://hal.science/hal-00503252v1>

Preprint submitted on 18 Jul 2010 (v1), last revised 2 Dec 2014 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

TEWFIK KERNANE^{1*} and ZOHRH A. RAIZAH²

¹ *Laboratory of Research in Artificial Intelligence (LRIA)*

Faculty of Electronics and Informatics

University of Sciences and Technology USTHB

Algiers, Algeria

tkernane@gmail.com

² *Department of Mathematics, Girls College of Education*

King Khalid University

Abha, Saudi Arabia

zo.hrh@hotmail.com

July 17, 2010

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 (or Gumbel minimum) 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 (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. 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

*Corresponding author: tkernane@gmail.com

known as the Gumbel (type I), Fréchet (type II) and Weibull (type III) distributions. For a thorough account of the theory of extreme value distributions we refer to the book by Coles [2].

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 [4]) and in medical data (see Klein and Zhang [6]), and if the truncation is ignored this can cause considerable bias in the estimation. 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 [3]) or nonparametric methods (see for example [4]). Samples to be considered in this paper include those that are singly right or singly left truncated. The EM algorithm is a powerfull iterative procedure which by repetition fill in the missing data with estimated values and to update the parameter estimates.

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 [1] did not treat the case of truncated samples for extreme value distributions and from the best of our knowlegde 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 can be 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 two simulated examples.

2 Estimations of the parameters of Extreme value distribution from truncated data

The Type-I least extreme values distribution function is defined as:

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

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 [1].

For the two parameters Weibull distribution $\mathcal{W}(\theta, \lambda)$ the pdf is given by

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

If X follows $\mathcal{W}(\theta, \lambda)$ then $Y = \log X$ follows a Gumbel minimum $G_{\min}(\mu, \sigma)$ with $\sigma = 1/\theta$ and $\mu = \ln \lambda$.

The two-parameter Fréchet distribution or type II extreme value distribution is largely used as a model for extremes of flood and rainfall data. The probability density function of the two-parameter Fréchet distribution $\mathcal{F}(\delta, \nu)$ 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 $\sigma = 1/\nu$.

Let x_1, \dots, x_{n_u} a right truncated sample from the type I least extreme value distribution which is also called the Gumbel minimum distribution which we denote by $\mathcal{G}_{\min}(\mu, \sigma)$, where n_u is the known number of untruncated observations. Consider C the cutoff point above which observations are discarded, that is we observe only the observations $x_i < C$. Consider $\underline{y} = (y_1, \dots, y_{n_t})'$ 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 $\{\underline{y}, n_t\}$. We can say that the complete data, denoted by z , is $z = (\underline{x}, n_t, \underline{y})$ such that $\underline{x} = \{z_i : z_i \leq C\}$, $\underline{y} = \{z_i : z_i > C\}$ and $n_t = \#\{z_i : z_i > C\}$.

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

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

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

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

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

The complete likelihood function is given by

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

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 [8], 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.

2.1 The EM Algorithm

This method is one of the numerical algorithms which helps to compute the maximum likelihood estimations based on missing or latent data. This method was introduced in the paper

of Dempster et *al.* [3] and they applied it for censored and truncated data for some distributions and also for mixture of distributions. For a detail review on this method, see the book by McLachlan and Krishnan [8]. 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.

2.1.1 The E-step

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

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

Taking the expectation and using Wald's formula we obtain

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

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

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

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.* [7], given by the equations:

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

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

$\Gamma(t + 1, e^\lambda)$ 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:

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

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, $\psi(1) = \frac{d}{dt} \ln [\Gamma(t+1)]|_{t=0}$ and $\psi(2) = \frac{d^2}{dt^2} \ln [\Gamma(t+1)]|_{t=0}$.

2.1.2 The M-step

In this step we obtain the estimations from the formulas

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

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

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 [5] as following:

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

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

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

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 If X follows a Gumbel maximum distribution $\mathcal{G}_{\max}(\alpha, \beta)$ then if data arising from X are truncated on the left, that is observations below a threshold C ($X > C$) are discarded, then $Y = -X$ will be a type-I lest extreme value distribution treated above with data truncated on the right ($Y < -C$ observed) and we can use the EM algorithm above to estimate $\mu = -\alpha$ and $\sigma = \beta$.

Remark 2 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 lest 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 3 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 left 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 Examples

Example 4 Consider a simulated data of size 300 from the extreme value distribution by taking $\mu = -4$ and $\sigma = 2$. The data was truncated at $C = -3$. The untruncated sample size was $n_u = 246$. The procedure will follow the following steps. Firstly we take for example the starting values $\mu_{(0)} = 0$ and $\sigma_{(0)} = 1$ and compute the conditional expectations given in (6) then we use the fixed point iterations given in equations (8,9). repeat the procedure until the algorithm converges. By using this algorithm we obtained the estimations $\hat{\sigma} = 2.0543$ and $\hat{\mu} = -4.3249$.

Example 5 We generated a data of size 200 from the extreme value distribution with $\mu = 2$ and $\sigma = 1.2$. We choose the threshold point $C = 2.5$ and obtained the untruncated sample size $n_u = 155$. By using the EM algorithm explained in this paper we get the results $\hat{\sigma} = 1.1932$ and $\hat{\mu} = 1.8947$.

References

- [1] Cohen, A.C., *Truncated and Censored Samples: Theory and Applications*. Dekker, New York 1991.
- [2] Coles, S., *An Introduction to Statistical Modeling of Extreme Values*, Springer-Verlag, 2001.
- [3] 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.
- [4] Efron, B. and Petrosian, V. (1999). Nonparametric methods for doubly truncated data, *J. Am. Stat. Assoc.* Vol. 94. No. 447. pp. 824-834.
- [5] 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.
- [6] 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. Jewell, 175-185.
- [7] 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.
- [8] McLachlan and Krishnan, *The EM algorithm and Extensions*, Wiley, New York, 2008.