应用概率统计 第 32 卷
第 2 期 2016 年 4 月

Chinese Journal of Applied Probability and Statistics
Apr., 2016, Vol. 32, No. 2, pp. 121-131
doi: 10.3969/j.issn.1001-4268.2016.02.002

Estimation of Parameters based on Type-II Hybrid Censored Data by EM Algorithm *

WANG Yashi*

(*Department of Science and Technology, China University of Political Science and Law, Beijing, 100027, China*)

WANG Xuanhe

(*School of Finance, Dongbei University of Finance and Economics, Dalian, 116025, China*)

WANG Ziyue

(*Department of Mathematics, Shanghai Jiao Tong University, Shanghai, 200240, China*)

Abstract: In this paper, we estimate parameters of gamma life distribution and normal life distribution by EM algorithm based on Type-II hybrid censored data. The covariance matrices are derived as well. Some numerical examples are also presented for illustration.

Keywords: EM algorithm; Type-II hybrid censored data; parameters estimation

2010 Mathematics Subject Classification: 62F12; 62F99

§1. Introduction

For a life testing experiment, it is very time consuming that if we wait until the last component fails. Thus, several censoring schemes have been proposed in the literature. For example, Type-I, Type-II censoring, and hybrid censoring. Interested readers may refer to [1–3] and others. Recently, Childs et al. [4] propose a new hybrid censoring scheme known as Type-II hybrid censoring scheme: Suppose n components $X_1, X_2, \ldots, X_n$ are on test, where X_i are i.i.d. nonnegative random variables. The experiment are terminated at a random time $T^* = \max\{X_{r:n}, T\}$, where $1 \leqslant r \leqslant n$, $T > 0$ are fixed in advance and $X_{r:n}$ is the rth order statistics from X_i, $i = 1, 2, \ldots, n$. If we do not fix time T, this will reduce to Type-II censoring, and if without fixed r, this will be Type-I censoring. This scheme has the advantage of guaranteeing that at least r failures are observed. As described in

*The research was supported by the National Natural Science Foundation of China (71171193, 11501575), the Humanities and Social Science Project of China University of Political Science and Law (12ZFQ11001), and the Young Faculty Academic Innovation Team of China University of Political Science and Law (2014CXTD01).

*Corresponding author, E-mail: wangysh0727@163.com.

Received August 21, 2013. Revised February 2, 2015.

[4], this censoring scheme may arise when the experimenter prepares T units of time for testing and at least r failures must be observed. Hence, we may have the following types of observations:

- $\{x_{1:n} \leqslant x_{2:n} \leqslant \cdots \leqslant x_{n:n}\}$ for $x_{n:n} \leqslant T$.
- $\{x_{1:n} \leqslant x_{2:n} \leqslant \cdots \leqslant x_{r:n}\}$ for $x_{r:n} \geqslant T$.
- $\{x_{1:n} \leqslant x_{2:n} \leqslant \cdots \leqslant x_{r:n} \leqslant \cdots \leqslant x_{m:n}\}$ for $x_{m:n} \leqslant T < x_{m+1:n}$.

Due to censoring, it is complicated to determine the maximum likelihood estimations (MLE) of the parameters. Some numerical procedures are developed for the estimators, for example, Newton-Raphson method. If we consider the censored observations as the missing data, the EM algorithm could also be used, which is very powerful to deal with the incomplete data problem (see also [5]). Compared to New-Raphson method, EM algorithm is slow, however it is more reliable. For related topics, readers may refer to [6] and [7] for EM algorithm in progressively censored data. Recently, Banerjee and Kundu [8] derived the MLE estimators of Weibull parameters based on the Type-II hybrid censored data by the iterative procedure. In this paper, we will estimate parameters of two popular lifetime models, gamma and normal distributions by EM algorithm.

The rest of this paper is organized as follows. In Section 2, we give the details how to use the EM algorithm to estimate gamma parameters and we also give the covariance matrix of our estimation. In Section 3, we estimate the normal parameters and the covariance matrix is given. We give some numerical examples in Section 4 for illustration purpose.

§2. Gamma Lifetime Model

2.1 EM Algorithm

Let X_i, $i = 1, 2, \ldots, n$ be gamma random variables $G(\gamma, \lambda)$, with distribution function F, density function f and survival function $S = 1 - F$. Denote $\boldsymbol{\Theta} = (\gamma, \lambda)$, the parameters of the gamma distribution. The joint likelihood function for complete data is

$$L(\boldsymbol{\Theta}) = n! \prod_{i=1}^{n} \frac{\lambda}{\Gamma(\gamma)} (\lambda x_{i:n})^{\gamma-1} \exp\{-\lambda x_{i:n}\}, \qquad \lambda > 0,\ \gamma > 0.$$

Denote the observed data set $\boldsymbol{X} = (X_{1:n}, X_{2:n}, \ldots, X_{m:n})$ and the missing data set $\boldsymbol{Y} = (Y_{m+1}, Y_{m+2}, \ldots, Y_n)$. Now, $(\boldsymbol{X}, \boldsymbol{Y})$ forms the complete data set. The joint density function of $(\boldsymbol{X},\boldsymbol{Y})$ is, for $y_i > \max\{T, x_{m:n}\}$, $i=m+1,m+2,\ldots,n$,

$$f(\boldsymbol{y},\boldsymbol{x}\,|\,\boldsymbol{\Theta}) = \frac{n!}{(n-m)!}\prod_{i=1}^{m}\frac{\lambda}{\Gamma(\gamma)}(\lambda x_{i:n})^{\gamma-1}\exp\{-\lambda x_{i:n}\}\prod_{i=m+1}^{n}\frac{\lambda}{\Gamma(\gamma)}(\lambda y_i)^{\gamma-1}\exp\{-\lambda y_i\}.$$

The joint density function of $\boldsymbol{X}$ is

$$f(\boldsymbol{x}\,|\,\boldsymbol{\Theta}) = \frac{n!}{(n-m)!}\prod_{i=1}^{m}\frac{\lambda}{\Gamma(\gamma)}(\lambda x_{i:n})^{\gamma-1}\exp\{-\lambda x_{i:n}\}S^{n-m}(\max\{T, x_{m:n}\}).$$

Hence,

$$f(\boldsymbol{y}\,|\,\boldsymbol{x},\boldsymbol{\Theta}) = \prod_{i=m+1}^{n}\frac{\lambda}{\Gamma(\gamma)}\frac{(\lambda y_i)^{\gamma-1}\exp\{-\lambda y_i\}}{S(\max\{T, x_{m:n}\})},\qquad y_i > \max\{T, x_{m:n}\}.$$

The E-step of the the algorithm calculates the expected log-likelihood,

$$\mathsf{E}[\ln L(\boldsymbol{\Theta}\,|\,\boldsymbol{x},\boldsymbol{Y})\,|\,\boldsymbol{\Theta}^{(r)},\boldsymbol{x}].$$

The M-step finds the maximum,

$$\boldsymbol{\Theta}^{(r+1)} = \text{the value that maximaizes } \mathsf{E}[\ln L(\boldsymbol{\Theta}\,|\,\boldsymbol{x},\boldsymbol{Y})\,|\,\boldsymbol{\Theta}^{(r)},\boldsymbol{x}].$$

Now, we compute the log-likelihood function,

$$\ln L(\boldsymbol{\Theta}\,|\,\boldsymbol{x},\boldsymbol{y}) = h(\lambda,\gamma) + (\gamma-1)\sum_{i=1}^{m}\ln x_{i:n} - \lambda\sum_{i=1}^{m}x_{i:n} + (\gamma-1)\sum_{i=m+1}^{n}\ln y_i - \lambda\sum_{i=m+1}^{n}y_i,$$

where $h(\lambda,\gamma)=\ln(n!/(n-m)!)+n\ln(\lambda^{\gamma}/\Gamma(\gamma))$, for $\lambda,\gamma>0$. Hence, we have to compute the following moment of the truncated distribution:

$$\begin{aligned}\mathsf{E}[Y_i\,|\,\boldsymbol{\Theta},\boldsymbol{x}] &= \mathsf{E}[Y_i\,|\,Y_i>\max\{T,x_{m:n}\}]\\ &= \frac{1}{S(\max\{T,x_{m:n}\})}\int_{\max\{T,x_{m:n}\}}^{\infty} y\frac{\lambda}{\Gamma(\gamma)}(\lambda y)^{\gamma-1}\exp\{-\lambda y\}\mathrm{d}y\\ &= \frac{\Gamma(\gamma)}{\Gamma(\gamma,\xi)}\frac{1}{\lambda\Gamma(\gamma)}\Gamma(\gamma+1,\xi) = \frac{\Gamma(\gamma+1,\xi)}{\lambda\Gamma(\gamma,\xi)},\end{aligned}$$

where $\xi=\lambda\max\{T,x_{m:n}\}$, and

$$\Gamma(a,x) = \int_x^{\infty} t^{a-1}\exp\{-t\}\mathrm{d}t$$

is the upper incomplete gamma function. Also,

$$\begin{aligned}\mathsf{E}[\ln Y_i\,|\,\boldsymbol{\Theta},\boldsymbol{x}] &= \mathsf{E}[\ln Y_i\,|\,Y_i>\max\{T,x_{m:n}\}]\\ &= \frac{\Gamma(\gamma)}{\Gamma(\gamma,\xi)}\int_{\max\{T,x_{m:n}\}}^{\infty}\ln(y)\frac{\lambda}{\Gamma(\gamma)}(\lambda y)^{\gamma-1}\exp\{-\lambda y\}\mathrm{d}y\end{aligned}$$

$$= \frac{\lambda}{\Gamma(\gamma,\xi)}\int_{\max\{T,x_{m:n}\}}^{\infty}\ln(y)(\lambda y)^{\gamma-1}\exp\{-\lambda y\}\mathrm{d}y.$$

Note that,

$$\begin{aligned}
&\int_{\max\{T,x_{m:n}\}}^{\infty}\ln(y)(\lambda y)^{\gamma-1}\exp\{-\lambda y\}\mathrm{d}y\\
=&\int_{\max\{T,x_{m:n}\}}^{\infty}\ln(\lambda y)(\lambda y)^{\gamma-1}\exp\{-\lambda y\}\mathrm{d}y-\ln\lambda\int_{\max\{T,x_{m:n}\}}^{\infty}(\lambda y)^{\gamma-1}\exp\{-\lambda y\}\mathrm{d}y\\
=&\frac{1}{\lambda}\int_{\xi}^{\infty}\ln(t)t^{\gamma-1}\exp\{-t\}\mathrm{d}t-\frac{\ln\lambda}{\lambda}\int_{\xi}^{\infty}t^{\gamma-1}\exp\{-t\}\mathrm{d}t\\
=&\frac{1}{\lambda}\Gamma'_{\gamma}(\gamma,\xi)-\frac{\ln\lambda}{\lambda}\Gamma(\gamma,\xi).
\end{aligned}$$

Hence,

$$\mathsf{E}[\ln Y_i\,|\,\boldsymbol{\Theta},\boldsymbol{x}]=\frac{\Gamma'_{\gamma}(\gamma,\xi)}{\Gamma(\gamma,\xi)}-\ln\lambda.$$

Note that the first term is actually the first derivative of the logarithm of the incomplete gamma function with respect to γ.

Now, the MLE for the complete data is,

$$\begin{aligned}
\frac{n\widehat{\gamma}}{\widehat{\lambda}}-\sum_{i=1}^{m}x_{i:n}-\sum_{i=m+1}^{n}y_i&=0,\\
n\ln\widehat{\lambda}-\frac{n\Gamma'(\widehat{\gamma})}{\Gamma(\widehat{\gamma})}+\sum_{i=1}^{m}\ln x_{i:n}+\sum_{i=m+1}^{n}\ln y_i&=0.
\end{aligned}$$

Taking the expectations, we derive the MLE for the censored data as follows:

$$\begin{aligned}
n\ln\widehat{\lambda}^{(r+1)}&=\frac{n\Gamma'(\widehat{\gamma}^{(r+1)})}{\Gamma(\widehat{\gamma}^{(r+1)})}-\sum_{i=1}^{m}\ln x_{i:n}-(n-m)\mathsf{E}[\ln Y_n\,|\,\widehat{\gamma}^{(r)},\widehat{\lambda}^{(r)},\boldsymbol{x}],\\
\widehat{\gamma}^{(r+1)}&=\frac{\widehat{\lambda}^{(r+1)}}{n}\sum_{i=1}^{m}x_{i:m}+\frac{\widehat{\lambda}^{(r+1)}}{n}(n-m)\mathsf{E}[Y_n\,|\,\widehat{\gamma}^{(r)},\widehat{\lambda}^{(r)},\boldsymbol{x}].
\end{aligned}$$

2.2 Asymptotic Covariance Matrix

In order to get the asymptotic covariance matrix, we first derive the observed information matrix. Note that the observed information matrix is

$$\begin{pmatrix}
-\mathsf{E}\dfrac{\partial^2\ln f(\boldsymbol{x}\,|\,\boldsymbol{\Theta})}{\partial^2\gamma} & -\mathsf{E}\dfrac{\partial^2\ln f(\boldsymbol{x}\,|\,\boldsymbol{\Theta})}{\partial\gamma\partial\lambda}\\
-\mathsf{E}\dfrac{\partial^2\ln f(\boldsymbol{x}\,|\,\boldsymbol{\Theta})}{\partial\lambda\partial\gamma} & -\mathsf{E}\dfrac{\partial^2\ln f(\boldsymbol{x}\,|\,\boldsymbol{\Theta})}{\partial^2\lambda}
\end{pmatrix}.$$

However, when EM algorithm is used, Louis [9] proposed a procedure for finding the observed information matrix. That is (see also [10] and [6]),

$$I_X(\boldsymbol{\Theta})=I_{(X,Y)}(\boldsymbol{\Theta})-I_{(Y|X)}(\boldsymbol{\Theta}),$$

where I_X, $I_{(X,Y)}$, $I_{(Y|X)}$ denotes the observed information, complete information, and missing information, respectively. It is easy to get the information matrix based on gamma distribution

$$I_{(X,Y)}(\boldsymbol{\Theta}) = n\begin{pmatrix} \dfrac{\Gamma''(\gamma)\Gamma(\gamma)-\Gamma'^2(\gamma)}{\Gamma^2(\gamma)} & -\dfrac{1}{\lambda} \\ -\dfrac{1}{\lambda} & \dfrac{\gamma}{\lambda^2} \end{pmatrix}.$$

Note that,

$$f(\boldsymbol{y}\,|\,\boldsymbol{x},\boldsymbol{\Theta}) = \prod_{i=m+1}^{n} \frac{\lambda}{\Gamma(\gamma,\xi)}(\lambda y_i)^{\gamma-1}\exp\{-\lambda y_i\}.$$

Hence,

$$\ln(f(\boldsymbol{y}\,|\,\boldsymbol{x},\boldsymbol{\Theta})) = \sum_{i=m+1}^{n}[\gamma\ln\lambda - \ln\Gamma(\gamma,\xi) + (\gamma-1)\ln y_i - \lambda y_i].$$

Now,

$$\begin{aligned}
\frac{\partial^2\ln(f(\boldsymbol{y}\,|\,\boldsymbol{x},\boldsymbol{\Theta}))}{\partial^2\gamma} &= -(n-m)\frac{\Gamma''_{\gamma}(\gamma,\xi)\Gamma(\gamma,\xi)-\Gamma'^2_{\gamma}(\gamma,\xi)}{\Gamma^2(\gamma,\xi)},\\
\frac{\partial^2\ln(f(\boldsymbol{y}\,|\,\boldsymbol{x},\boldsymbol{\Theta}))}{\partial\gamma\partial\lambda} &= \frac{n-m}{\lambda} - (n-m)\frac{\Gamma''_{\gamma,\lambda}(\gamma,\xi)\Gamma(\gamma,\xi)-\Gamma'_{\gamma}(\gamma,\xi)\Gamma'_{\lambda}(\gamma,\xi)}{\Gamma^2(\gamma,\xi)}\max\{T,x_{m:n}\},\\
\frac{\partial^2\ln(f(\boldsymbol{y}\,|\,\boldsymbol{x},\boldsymbol{\Theta}))}{\partial^2\lambda} &= \frac{-(n-m)\gamma}{\lambda^2} - \frac{\Gamma''_{\lambda}(\gamma,\xi)\Gamma(\gamma,\xi)-\Gamma'^2_{\lambda}(\gamma,\xi)}{\Gamma^2(\gamma,\xi)}(\max\{T,x_{m:n}\})^2.
\end{aligned}$$

These values form the observed information matrix $I_{(Y|X)}(\boldsymbol{\Theta})$, hence $I_X(\boldsymbol{\Theta})$. Inverting $I_X(\boldsymbol{\Theta})$, we get the covariance matrix the MLE estimator $\widehat{\Theta}$.

§3. Normal Lifetime Model

3.1 EM Algorithm

Let X_i, $i=1,2,\ldots,n$ be normal random variables $\mathrm{N}(\mu,\sigma^2)$, with distribution function F, density function f and survival function $S=1-F$. Denote $\boldsymbol{\Theta}=(\mu,\sigma)$, the parameters of normal distribution. The joint likelihood function for complete data is

$$L(\boldsymbol{\Theta}) = n!\prod_{i=1}^{n} f(x_{i:n}) = n!\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}\exp\Big\{-\frac{(x_{i:n}-\mu)^2}{2\sigma^2}\Big\}.$$

Denote the observed data set $\boldsymbol{X}=(X_{1:n},X_{2:n},\ldots,X_{m:n})$ and the missing data set $\boldsymbol{Y}=(Y_{m+1},Y_{m+2},\ldots,Y_n)$. Note that the joint density function of $(\boldsymbol{Y},\boldsymbol{X})$ is

$$f(\boldsymbol{y},\boldsymbol{x}\,|\,\boldsymbol{\Theta}) = \frac{n!}{(n-m)!}\prod_{i=1}^{m}\frac{1}{\sqrt{2\pi}\sigma}\exp\Big\{-\frac{(x_{i:n}-\mu)^2}{2\sigma^2}\Big\}\prod_{i=m+1}^{n}\frac{1}{\sqrt{2\pi}\sigma}\exp\Big\{-\frac{(y_i-\mu)^2}{2\sigma^2}\Big\}.$$

The joint density function of $\boldsymbol{X}$ is

$$f(\boldsymbol{x}\,|\,\boldsymbol{\Theta}) = \frac{n!}{(n-m)!}\prod_{i=1}^{m}\frac{1}{\sqrt{2\pi}\sigma}\exp\Big\{-\frac{(x_{i:n}-\mu)^2}{2\sigma^2}\Big\}S^{n-m}(\max\{T, x_{m:n}\}).$$

Hence,

$$f(\boldsymbol{y}\,|\,\boldsymbol{x},\boldsymbol{\Theta}) = \prod_{i=m+1}^{n}\frac{1}{\sqrt{2\pi}\sigma}\frac{\exp\{-(y_i-\mu)^2/(2\sigma^2)\}}{S(\max\{T, x_{m:n}\})}, \qquad y_i > \max\{T, x_{m:n}\}. \tag{1}$$

Similarly, let us compute the log-likelihood function,

$$\ln L(\boldsymbol{\Theta}\,|\,\boldsymbol{x},\boldsymbol{y}) = C - n\ln\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{m}(x_{i:n}-\mu)^2 - \frac{1}{2\sigma^2}\sum_{i=m+1}^{n}(y_i-\mu)^2,$$

where C is the constant. Note that

$$\mathsf{E}\Big[\sum_{i=m+1}^{n}(Y_i-\mu)^2\,|\,\boldsymbol{\Theta},\boldsymbol{x}\Big] = \sum_{i=m+1}^{n}\mathsf{E}[Y_i^2\,|\,\boldsymbol{\Theta},\boldsymbol{x}] - 2\mu\sum_{i=m+1}^{n}\mathsf{E}[Y_i\,|\,\boldsymbol{\Theta},\boldsymbol{x}] + (n-m)\mu^2.$$

From Equation (1),

$$\begin{aligned}\mathsf{E}[Y_i^2\,|\,\boldsymbol{\Theta},\boldsymbol{x}] &= \int_{\max\{T,x_{m:n}\}}^{\infty} y^2\frac{1}{\sqrt{2\pi}\sigma}\frac{\exp\{-(y-\mu)^2/(2\sigma^2)\}}{S(\max\{T, x_{m:n}\})}\mathrm{d}y\\ &= \mathsf{E}[Y_i^2\,|\,Y_i > \max\{T, x_{m:n}\}]\\ &= \sigma^2(1+\xi Q) + 2\sigma\mu Q + \mu^2,\end{aligned}$$

where the last equality follows from [2] (see also [6]),

$$\xi = \frac{\max\{T, x_{m:n}\}-\mu}{\sigma} \qquad \text{and} \qquad Q = \frac{\phi(\xi)}{1-\Phi(\xi)},$$

is the hazard rate function of standard normal distribution.

Similarly,

$$\mathsf{E}[Y_i\,|\,\boldsymbol{\Theta},\boldsymbol{x}] = \mathsf{E}[Y_i\,|\,Y_i > \max\{T, x_{m:n}\}] = \sigma Q + \mu.$$

It is well-known that the explicit formulas for MLE of $\boldsymbol{\Theta}$,

$$\widehat{\mu} = \frac{1}{n}\Big[\sum_{i=1}^{m}x_{i:n} + \sum_{i=m+1}^{n}y_i\Big], \qquad \widehat{\sigma} = \Big[\frac{1}{n}\sum_{i=1}^{m}(x_{i:n}-\widehat{\mu})^2 + \frac{1}{n}\sum_{i=m+1}^{n}(y_{i:n}-\widehat{\mu})^2\Big]^{1/2}.$$

Taking the expectation, it follows that,

$$\widehat{\mu}^{(r+1)} = \frac{1}{n}\Big[\sum_{i=1}^{m}x_{i:n} + (n-m)\mathsf{E}[Y_{m+1}\,|\,\boldsymbol{\Theta}^{(r)},\boldsymbol{x}]\Big] = \frac{1}{n}\Big[\sum_{i=1}^{m}x_{i:n} + (n-m)(\sigma^{(r)}Q^{(r)}+\mu^{(r)})\Big],$$

$$\widehat{\sigma}^{(r+1)} = \Big\{\frac{1}{n}\sum_{i=1}^{m}x_{i:n}^2 + \frac{n-m}{n}\mathsf{E}[Y_{m+1}^2\,|\,Y_{m+1} > \max\{T, x_{m:n}\}, \mu^{(r)}, \sigma^{(r)}] - (\widehat{\mu}^{(r+1)})^2\Big\}^{1/2}.$$

3.2 Asymptotic Covariance Matrix

Similarly, we compute $I_{(X,Y)}$ and $I_{(Y|X)}$, the complete information, and the missing information, respectively. It is well-known that the information matrix based on normal distribution is

$$I_{(X,Y)}(\boldsymbol{\Theta}) = \frac{n}{\sigma^2}\begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}.$$

Note that,

$$\ln f(\boldsymbol{y}\,|\,\boldsymbol{x},\boldsymbol{\Theta}) = C - (n-m)\ln\sigma - \sum_{i=m+1}^{n}\frac{(y_i-\mu)^2}{2\sigma^2} - (n-m)\ln S(\max\{T, x_{m:n}\}).$$

Taking the derivative with respect to μ and σ,

$$\begin{aligned}
\frac{\partial f(\boldsymbol{y}\,|\,\boldsymbol{x},\boldsymbol{\Theta})}{\partial\mu} &= \sum_{i=m+1}^{n}\frac{(y_i-\mu)}{\sigma^2} - (n-m)\frac{f(\max\{T,x_{m:n}\})}{S(\max\{T,x_{m:n}\})} \\
&= \sum_{i=m+1}^{n}\frac{(y_i-\mu)}{\sigma^2} - (n-m)\frac{Q}{\sigma}, \\
\frac{\partial f(\boldsymbol{y}\,|\,\boldsymbol{x},\boldsymbol{\Theta})}{\partial\sigma} &= -\frac{n-m}{\sigma} + \sum_{i=m+1}^{n}\frac{(y_i-\mu)^2}{\sigma^3} - (n-m)\frac{Q\xi}{\sigma}.
\end{aligned}$$

From [6], it is known that,

$$\begin{aligned}
-\mathsf{E}\frac{\partial^2 \ln f(\boldsymbol{Y}\,|\,\boldsymbol{x},\boldsymbol{\Theta})}{\partial^2\mu} &= \frac{n-m}{\sigma^2}[1+\xi Q - Q^2], \\
-\mathsf{E}\frac{\partial^2 \ln f(\boldsymbol{Y}\,|\,\boldsymbol{x},\boldsymbol{\Theta})}{\partial^2\sigma} &= \frac{n-m}{\sigma^2}[2+\xi Q(1-\xi Q+\xi^2)], \\
-\mathsf{E}\frac{\partial^2 \ln f(\boldsymbol{Y}\,|\,\boldsymbol{x},\boldsymbol{\Theta})}{\partial\sigma\partial\mu} &= \frac{n-m}{\sigma^2}[Q+\xi Q(\xi - Q)].
\end{aligned}$$

These values form the observed information matrix $I_{(Y|X)}$, hence, $I_X(\boldsymbol{\Theta})$. Inverting $I_X(\boldsymbol{\Theta})$, we get the covariance matrix of $\boldsymbol{\Theta}$.

§4. Numerical Examples

4.1 Simulations

To illustrate our method, we generate standard Normal samples with size $n = 10$. The program is written in R to execute the EM algorithm (R code is available based on request). The generated samples are in the following table.

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
-1.2212	-1.1296	-0.7851	-0.2846	0.1464	0.1888	0.4441	0.6696	1.7121	2.1553

Assuming $r = 7$ and $T = 1$, due to the censoring, $\max\{T, x_{7:10}\} = 1$, we only have 8 observations. We choose the observed sample mean and sample standard deviation as the initial values $\mu^{(0)} = -0.2464$ and $\sigma^{(0)} = 0.7252$. Using the EM algorithm, we get our estimated values $\widehat{\mu} = 0.1044$, and $\widehat{\sigma} = 0.9474$. From our estimation, we calculate the observed information matrix

$$I_{(X,Y)}(\widehat{\boldsymbol{\Theta}}) = \begin{pmatrix} 11.1416 & 0 \\ 0 & 22.2833 \end{pmatrix}, \qquad I_{(Y|X)}(\widehat{\boldsymbol{\Theta}}) = \begin{pmatrix} 0.4582 & 1.6279 \\ 1.6279 & 5.9955 \end{pmatrix}.$$

Hence, the observed information matrix is

$$I_{(X)}(\widehat{\boldsymbol{\Theta}}) = \begin{pmatrix} 10.6348 & -1.6279 \\ -1.6279 & 16.2878 \end{pmatrix}.$$

Invert this matrix, we get the variance-covariance matrix

$$\mathsf{Var}\,(\widehat{\boldsymbol{\Theta}}) = \begin{pmatrix} 0.0951 & 0.0095 \\ 0.0095 & 0.0623 \end{pmatrix}.$$

4.2 Real Data Analysis

In this section, we analyze one real data for illustration in [11] (see also [12]). This data set is the strength measured in GPA for single carbon fibers of 10mm gauge length with sample size $n = 63$. For convenience, the data is presented as follows.

1.901,	2.132,	2.203,	2.228,	2.257,	2.350,	2.361,	2.396,	2.397,	2.445,	2.454,
2.474,	2.518,	2.522,	2.525,	2.532,	2.575,	2.614,	2.616,	2.618,	2.624,	2.659,
2.675,	2.738,	2.740,	2.856,	2.917,	2.928,	2.937,	2.937,	2.977,	2.996,	3.030,
3.125,	3.139,	3.145,	3.220,	3.223,	3.235,	3.243,	3.264,	3.272,	3.294,	3.332,
3.346,	3.377,	3.408,	3.435,	3.493,	3.501,	3.537,	3.554,	3.562,	3.628,	3.852,
3.871,	3.886,	3.971,	4.024,	4.027,	4.225,	4.395,	5.020.			

It is worth remarking that Kundu and Gupta [12] fitted the modified data to the Weibull distribution. It fits pretty good. However, if we make the log-transformation on the data set ($w = \ln(x)$), it fits the norm distribution very well. We present the QQ-plot for the transformed data. It looks good. The Shapiro-Wilk normality test statistics is 0.9888 with associated p-value 0.8378. Fit the transformed data to normal distribution gives $\widehat{\mu} = 1.0985$ and $\widehat{\sigma} = 0.1975$. We also present the empirical distribution and fitted normal distribution in the same graph. It could be seen that they fit very good.

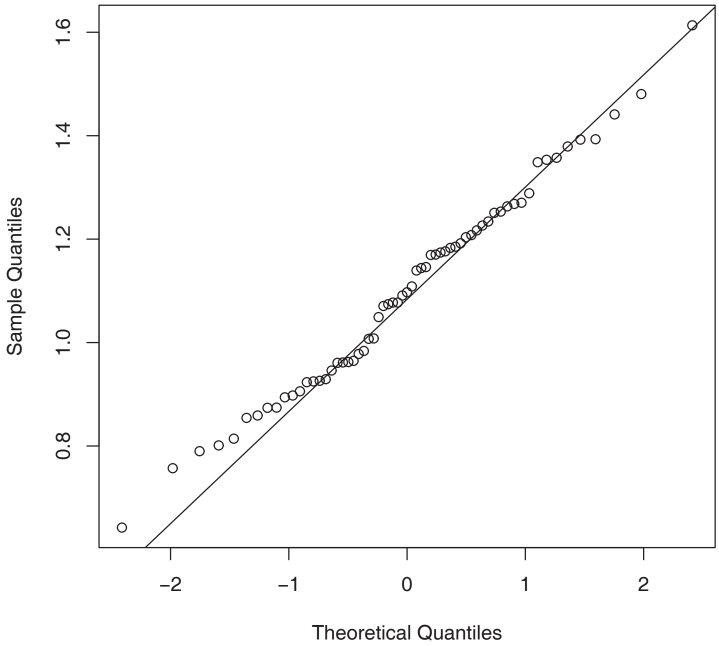

Figure 1 Normal Q-Q plot

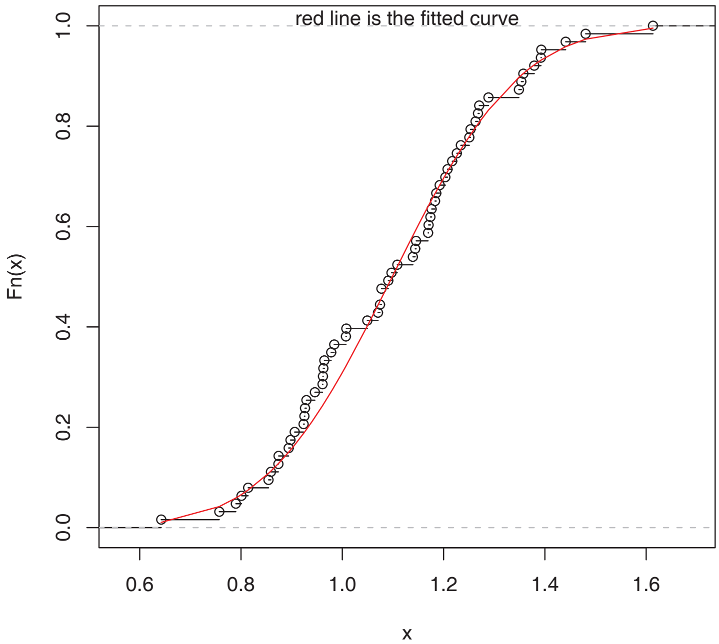

Figure 2 ecdf(w)

Now, let us first consider the following scheme 1: $r=40$, $T=1.2$. Then, $\max\{x_{m:n},T\}$ $=1.2$. Using the EM-algorithm, it follows that $\widehat{\mu}=1.0004$ and $\widehat{\sigma}=0.1996$. From our estimation, we calculate the observed information matrix

$$I_{(X,Y)}(\widehat{\boldsymbol{\Theta}}) = \begin{pmatrix} 1\,580.719 & 0 \\ 0 & 3\,161.438 \end{pmatrix}, \qquad I_{(Y|X)}(\widehat{\boldsymbol{\Theta}}) = \begin{pmatrix} 244.3802 & 384.2217 \\ 384.2217 & 810.3629 \end{pmatrix}.$$

Hence, the observed information matrix is

$$I_{(X)}(\widehat{\boldsymbol{\Theta}}) = \begin{pmatrix} 1\,336.34 & -384.222 \\ -384.222 & 2\,351.08 \end{pmatrix}.$$

Invert this matrix, we get the variance-covariance matrix

$$\mathsf{Var}\,(\widehat{\boldsymbol{\Theta}}) = \begin{pmatrix} 0.0008 & 0.0001 \\ 0.0001 & 0.0004 \end{pmatrix}.$$

Let us consider the scheme 2: $r=40$, $T=1$. Then, $\max\{x_{m:n},T\}=1.1764988$. Using the EM-algorithm, it follows that $\widehat{\mu}=1.1014$ and $\widehat{\sigma}=0.2007$. From our estimation, we calculate the observed information matrix

$$I_{(X,Y)}(\widehat{\boldsymbol{\Theta}}) = \begin{pmatrix} 1\,563.785 & 0 \\ 0 & 3\,127.570 \end{pmatrix}, \qquad I_{(Y|X)}(\widehat{\boldsymbol{\Theta}}) = \begin{pmatrix} 278.3359 & 436.9655 \\ 436.9655 & 921.1398 \end{pmatrix}.$$

Hence, the observed information matrix is

$$I_{(X)}(\widehat{\boldsymbol{\Theta}}) = \begin{pmatrix} 1\,285.45 & -436.966 \\ -436.966 & 2\,206.43 \end{pmatrix}.$$

Invert this matrix, we get the variance-covariance matrix

$$\mathsf{Var}\,(\widehat{\boldsymbol{\Theta}}) = \begin{pmatrix} 0.0008 & 0.0002 \\ 0.0002 & 0.0005 \end{pmatrix}.$$

Acknowledgment The authors thank anonymous referees for their insightful comments which have greatly improve the presentation of this manuscript.

References

[1] Epstein B. Truncated life tests in the exponential case [J]. *Ann. Math. Statist.*, 1954, **25(3)**: 555–564.

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

[3] Gupta R D, Kundu D. Hybrid censoring schemes with exponential failure distribution [J]. *Comm. Statist. Theory Methods*, 1998, **27(12)**: 3065–3083.

[4] Childs A, Chandrasekar B, Balakrishnan N, et al. Exact likelihood inference based on Type-I and Type-II hybrid censored samples from the exponential distribution [J]. *Ann. Inst. Statist. Math.*, 2003, **55(2)**: 319–330.

[5] Dempster A P, Laird N M, Rubin D B. Maximum likelihood from incomplete data via the EM algorithm [J]. *J. Roy. Statist. Soc. Ser. B*, 1977, **39(1)**: 1–38.

[6] Ng H K T, Chan P S, Balakrishnan N. Estimation of parameters from progressively censored data using EM algorithm [J]. *Comput. Statist. Data Anal.*, 2002, **39(4)**: 371–386.

[7] Balakrishnan N, Kannan N, Nagaraja H N. *Advances in Ranking and Selection, Multiple Comparisons, and Reliability* [M]. Boston: Birkhäuser, 2005.

[8] Banerjee A, Kundu D. Inference based on Type-II hybrid censored data from a Weibull distribution [J]. *IEEE T. Reliab.*, 2008, **57(2)**: 369–378.

[9] Louis T A. Finding the observed information matrix when using the EM algorithm [J]. *J. Roy. Statist. Soc. Ser. B*, 1982, **44(2)**: 226–233.

[10] Tanner M A. *Tools for Statistical Inference: Observed Data and Data Augmentation Methods* [M]. 2nd ed. New York: Springer, 1993.

[11] Bader M G, Priest A M. Statistical aspects of fiber and bundle strength in hybrid composites [C] // Hayashi T, Kawata K, Umekawa S. *Progress in Science and EngIneerIng Composites*. Tokyo: ICCM-IV, 1982: 1129–1136.

[12] Kundu D, Gupta R D. Estimation of $\mathsf{P}(Y < X)$ for Weibull distributions [J]. *IEEE T. Reliab.*, 2006, **55(2)**: 270–280.

基于EM算法的II型混合删失数据的参数估计

王雅实

(中国政法大学科学技术教学部, 北京, 100027)

王选鹤

(东北财经大学金融学院, 大连, 116025)

王子悦

(上海交通大学数学系, 上海, 200240)

摘　要： 本文基于EM算法, 对于II型混合删失数据, 分别在Gamma和正态寿命分布族下进行了参数估计, 并且推导了协防差矩阵. 最后给出了几个很好的数值例子, 对本文的结论进行了诠释.

关键词： EM算法; II型混合删失数据; 参数估计

中图分类号： O212.1