TELKOMNIKA Indonesian Journal of Electrical Engineering
Vol.12, No.6, June 2014, pp. 4717 ~ 4723
DOI: 10.11591/telkomnika.v12i6.5491

Impact of Missing Data on EM Algorithm under Rayleigh Distribution

Zhendong Li[1], Mengmeng Li[2]
[1]School of Information Engineering, n, Lanzhou, 730020, China
[2]School of Statistics, Lanzhou University of Finance and Economics, Lanzhou, 730020, China
Corresponding author, email: lizd@lzcc.edu.com[1], limm2012@yeah.net[2]

Abstract

Is EM algorithm parameter estimation under Rayleigh distribution sensitive to missing data and if it is, what extent is it? By designing computer simulation methods, contrast and analyze the results of maximum likelihood estimation with complete data and EM algorithm estimation under different missing rate in small sample. It shows that the results were almost identical when the missing rate is below 0.30, but the efficiency of EM parameter estimation gradually deteriorates as the missing rate increases. Meanwhile the results also show that the EM algorithm is sensitive to sample size and the selection of initial value.

Keywords: *missing data, Rayleigh algorithm, EM, parameter estimation*

1. Introduction

In the study of reliability, Exponential distribution, Weibull distribution, Rayleigh distributions and so on, are important life distribution. Therefore, researchers explored good parameter estimation from various point of view, whether in the complete sample, or in the case of censored samples. Based on the Type-II censoring life test, Wei *et al* [1] discusses exponential distribution; Liu *et al* [2] discusses experimental Bayes parameter estimation of Weibull distribution. However, how missing data affects parameter estimation under common condition of missing data is still lack of in-depth study.

EM (Expectation-Maximization) algorithm plays an important role in parameter estimation, especially performs well in small sample with missing data. EM algorithm is an iterative algorithm with numerical stability, small storage capacity. It could ensure that in the parameter estimation process the likelihood function of observation data is nondecreasing in each iterative and the accuracy is relatively high. But under different missing rate, how EM algorithm affects the accuracy of parameter estimation need a further study. This paper is to work on the accuracy of EM parameter estimation of Rayleigh distribution under different missing rate based on foregoing case by the numerical example. It analyzes and evaluates the impact of different missing rate on the accuracy of parameter estimation.

2. Research Method

Let the density function of Rayleigh distribution is:

$$f(x;\theta) = (x/\theta^2)\exp(-x^2/2\theta^2),\ x>0\ \ (\theta>0) \tag{1}$$

The distribution function is:

$$F(x) = 1 - e^{-\frac{x^2}{2\theta^2}},\quad x>0 \tag{2}$$

$x_1, x_2, \mathrm{L}\ , x_n$ is the sample, then its likelihood function is:

Received December 29, 2013; Revised March 8, 2014; Accepted March 21, 2014

$$L(\theta)=\prod_{i=1}^{n}\frac{x_i}{\theta^2}\exp(-\frac{x_i}{2\theta^2})=\left(\prod_{i=1}^{n}x_i\right)\theta^{-2n}\exp\left(-\frac{\sum_{i=1}^{n}x_i^2}{2\theta^2}\right)\quad x_i>0 \tag{3}$$

Logarithmic:

$$\ln L(\theta)=\ln\left(\prod_{i=1}^{n}x_i\right)-2n\ln\theta-\frac{1}{2\theta^2}\sum_{i=1}^{n}x_i^2 \tag{4}$$

The maximum likelihood estimator of parameter is:

$$\hat{\theta}=\left(\frac{1}{2n}\sum_{i=1}^{n}x_i^2\right)^{\frac{1}{2}} \tag{5}$$

Each iteration of EM iterative algorithm of Rayleigh distribution parameter estimation consists of two steps: E(Expectation) step and M(Maximization) step.

Let $X=(x_1,x_2,\mathrm{L}\ ,x_n)$ be the observation data, but we only obtain $Z=(z_1,\mathrm{L}\ ,z_k,z_{k+1}^{+},\mathrm{L}\ ,z_n^{+})$ because of conditional limits, $z_{k+1}^{+},\mathrm{L}\ ,z_n^{+}$ represents the data which don't be observed, and X and Z have the following relations:

$$\begin{cases}x_j=z_j & j=1,2,\mathrm{L}\ ,k\\ x_j>z_j & j=k+1,k+2,\mathrm{L}\ ,n\end{cases} \tag{6}$$

Solve $\hat{\theta}$ which is the estimated value of θ by EM algorithm, starting from the initial value $\theta^{(i)}$ ($i=0,1,2,\mathrm{L}$), the two step of the $i+1$ iteration is:

1) E step：Calculate the conditional expectation.

$$Q(\theta|\theta^{(i)})\equiv E[\ln L(x,\theta)|Z,\theta^{(i)}]=\sum_{j=1}^{n}E(\ln x_j|Z,\theta^{(i)})-2n\ln\theta-\frac{1}{2\theta^2}\sum_{j=1}^{n}E(x_j^2|Z,\theta^{(i)}) \tag{7}$$

2) M step: Maximize $Q(\theta|\theta^{(i)})$ to get the updated $\theta^{(i+1)}$, that is:

$$Q(\theta^{(i+1)}|\theta^{(i)})=\max\ Q(\theta|\theta^{(i)}) \tag{8}$$

Let $\frac{dQ}{d\theta}=-\frac{2n}{\theta}+\frac{1}{\theta^3}\sum_{j=1}^{n}E(x_j^2|Z,\theta^{(i)})=0$ (9)

Including it:

$$\sum_{j=1}^{n}E(x_j^2|Z,\theta^{(i)})=\sum_{j=1}^{k}E(x_j^2|Z,\theta^{(i)})+\sum_{j=k+1}^{n}E(x_j^2|Z,\theta^{(i)})$$

$$=\sum_{j=1}^{k}z_j^2+\sum_{j=k+1}^{n}\frac{E(x_j^2\ge z_j,\theta^{(i)})}{p(x_j^2\ge z_j,\theta^{(i)})}=\sum_{j=1}^{k}z_j^2+\sum_{j=k+1}^{n}\frac{\int_{z_j}^{+\infty}\frac{x_j}{2\theta^{(i)2}}e^{-\frac{x_j}{2\theta^{(i)2}}}dx_j}{\int_{z_j}^{+\infty}\frac{1}{2\theta^{(i)2}}e^{-\frac{x_j}{2\theta^{(i)2}}}dx_j}$$

$$= \sum_{j=1}^{k} z_j^2 + \sum_{j=k+1}^{n} \frac{(2\theta^{(i)2} + z_j)e^{-\frac{z_j}{2\theta^{(i)2}}}}{e^{-\frac{z_j}{2\theta^{(i)2}}}}$$

$$= \sum_{j=1}^{k} z_j^2 + \sum_{j=k+1}^{n} (2\theta^{(i)2} + z_j) = \sum_{j=1}^{k} z_j^2 + (n-k)\theta^{(i)2} + \sum_{j=k+1}^{n} z_j = \sum_{j=1}^{k} z_j^2 + (n-k)\theta^{(i)2} \quad (10)$$

Because,

$$Q(\theta^{(i+1)} \mid \theta^{(i)}) = \max Q(\theta \mid \theta^{(i)}) \quad (11)$$

By (9),

$$-\frac{2n}{\theta^{(i+1)}} + \frac{1}{\theta^{(i+1)3}}(\sum_{j=1}^{k} z_j^2 + (n-k)\theta^{(i)2}) = 0 \quad (12)$$

Get the iterative formula is:

$$\theta^{(i+1)} = \left(\frac{\sum_{j=1}^{k} z_j^2 + (n-k)\theta^{(i)2}}{2n} \right)^{\frac{1}{2}} \quad (13)$$

Stop the iterations until $|\theta^{(i+1)} - \theta^{(i)}|$ is sufficiently small.

3. Analysis of Computer Simulation

3.1. Computer Simulation Designed [3]

Step 1: Generate observations of a random number $Y = (y_1, y_2, \mathrm{L}\ , y_n)$ with Parameter θ and restrictive random number $X = (x_1, x_2, \mathrm{L}\ , x_n)$ under Rayleigh distribution by the computer;

Step 2: Generate the observed data by computer which random delete some data from original data under different missing rate p (such as: 0.05 ,0.10 ,0.15 , etc.):

$$z_j = y_j I(x_j \geq y_j) + x_j I(x_j < y_j) \quad (j = 1, 2, \mathrm{L}\ , n) \quad (14)$$

Substitute it into the iterative formula (13);

Step 3: Take the initial value $\theta^{(0)}$ for a given $\varepsilon > 0$, test whether $|\theta^{(i+1)} - \theta^{(i)}| < \varepsilon$ or not, if it meet the above conditions, then $\tilde{\theta} = \theta^{(i+1)}$, or using (13) to continue calculating $\theta^{(i+1)}$;

Step 4: Calculate $\hat{\theta} = \left(\frac{1}{2n} \sum_{i=1}^{n} y_i^2 \right)^{\frac{1}{2}}$ which is the maximum likelihood estimated value of θ by the observed data $Y = (y_1, y_2, \mathrm{L}\ , y_n)$;

Step 5: Repeat Step1 to step4 to get the observed sequence $\hat{\theta}_1, \hat{\theta}_2, \mathrm{L}\ , \hat{\theta}_n$ and $\tilde{\theta}_1, \tilde{\theta}_2, \mathrm{L}\ , \tilde{\theta}_n$. Respectively calculate sequence $a_i = |\hat{\theta}_i - \theta|$ and $b_i = |\tilde{\theta}_i - \theta|$ under different

missing rate p and their MSE, analyze the differences between the likelihood estimates and EM estimates with incomplete data.

3.2. Comparison and Analysis of Simulation Results

1) The impact of missing rate on parameter estimation

When θ =0.5 and the initial value of EM algorithm $\theta^{(0)}=1$, the likelihood estimates and EM estimates under different missing rate and their MSE shows in Table 1 and Table 2, and its analysis table shows in Figure 1.

Table 1. Likelihood Estimates and EM Estimates ($\theta=0.5$)

Missing rate (*p*)	Likelihood estimate	EM estimate
0.05	0.4978	0.5316
0.10	0.4955	0.5520
0.15	0.4984	0.5250
0.20	0.4986	0.5392
0.25	0.5005	0.4797
0.30	0.4969	0.4389
0.35	0.4946	0.4553
0.40	0.4974	0.4255
0.45	0.5012	0.4206

Table 2. MSE of Likelihood Estimation and EM Estimation ($\theta=0.5$)

Missing rate (*p*)	Likelihood estimation	EM estimaton
0.05	0.0004	0.0004
0.10	0.0005	0.0007
0.15	0.0005	0.0012
0.20	0.0004	0.0013
0.25	0.0004	0.0017
0.30	0.0004	0.0014
0.35	0.0005	0.0022
0.40	0.0005	0.0017
0.45	0.0005	0.0015

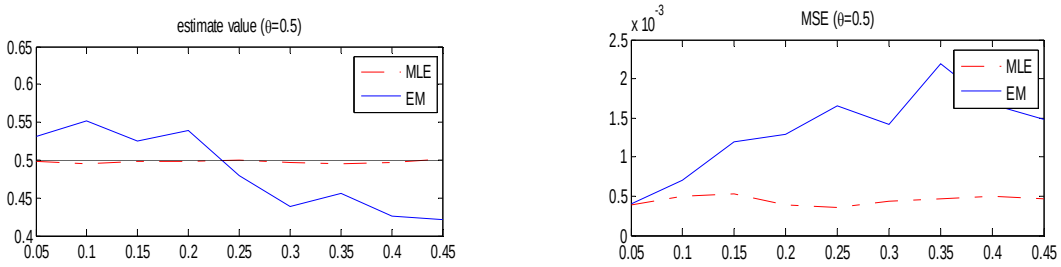

Figure 1. Likelihood Estimates and EM Estimates and their MSE ($\theta=1$)

It can be seen from Figure 1, when the missing rate $p > 0.30$, the difference between likelihood estimation and EM estimation is significant, meanwhile the MSE also increasingly increases.

When θ =1 and the initial value of EM algorithm $\theta^{(0)}=0.5$, the likelihood estimate and EM estimate under different missing rate and their mean square error shows in Table 3 and Table 4, and its analysis table is shown in Figure 2.

Table 3. Likelihood Estimates and EM Estimates ($\theta = 1$)

Missing rate (*p*)	Likelihood estimate	EM estimate
0.05	0.9942	0.8874
0.10	1.0004	0.8857
0.15	0.9902	0.8982
0.20	1.0077	0.8372
0.25	0.9819	0.8078
0.30	0.9924	0.7597
0.35	1.0013	0.7420
0.40	0.9954	0.5785
0.45	0.9983	0.5889

Table 4. MSE of Likelihood Estimation and EM Estimation ($\theta = 1$)

Missing rate (*p*)	Likelihood estimation	EM estimation
0.05	0.0020	0.0025
0.10	0.0017	0.0025
0.15	0.0014	0.0039
0.20	0.0019	0.0051
0.25	0.0015	0.0054
0.30	0.0022	0.0059
0.35	0.0016	0.0070
0.40	0.0026	0.0090
0.45	0.0015	0.0060

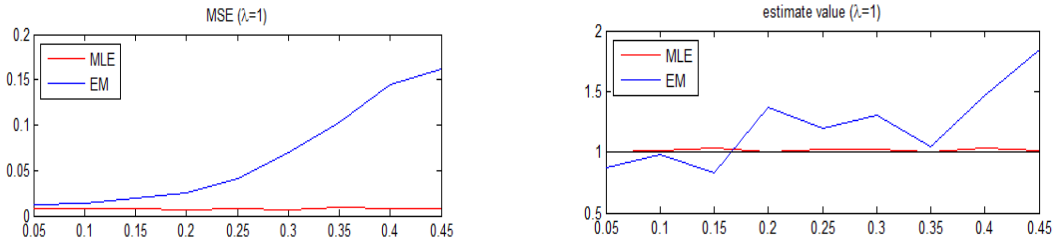

Figure 2. Likelihood Estimates and EM Estimates and their MSE ($\theta = 1$)

It can be seen from Figure 1, Figure 2, though θ and the initial value of EM algorithm $\theta^{(0)}$ takes different values ,when the missing rate is greater than 0.30, the differences in the MSE also increases.That is, when the missing rate is greater than 0.30,the results of EM algorithm parameter estimation are significantly different to the results of maximum likelihood estimation and the difference increases significantly as the missing rate increases.

2) The impact of sample size on parameter estimation

EM algorithm performs well in small sample with missing data. When θ =1, the initial value of EM algorithm $\theta^{(0)} = 0.5$, sampling size $n = 50,30,20$,the MSE of the likelihood estimation and EM estimation with missing data shows in Figure 3 to Figure 5.

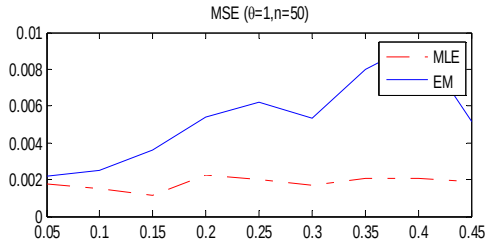

Figure 3. MSE of Different Missing Rate ($n = 50$)

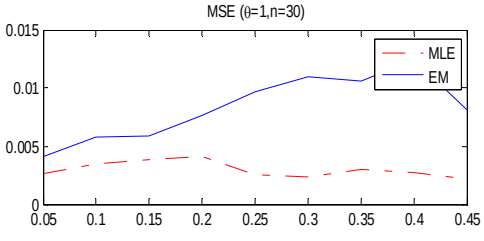

Figure 4. MSE of Different Missing Rate ($n = 30$)

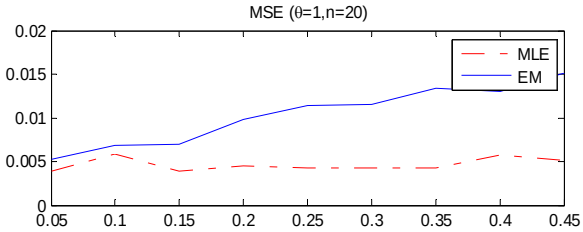

Figure 5. MSE of Different Missing Rate $(n = 20)$

It shows in Figure 3, when sample size $n = 50$,missing rate $p > 0.30$, the difference between likelihood estimation and EM estimation becomes clear. It shows in Figure 4 and Figure 5, when $n = 30,20$, $p > 0.15$,the difference is clear. It indicates that sample size has a big imfact on the accuracy of EM estimation.

3) The impact of initial value on parameter estimation

It shows in Figure 6, when θ =1, the effect of EM algorithm was not good with the initial value $\theta^{(0)} = 0.5$,when the missing rate is fairly low, the effect will be greatly improved if selecting 0.8 as its initial value. It indicates that the EM algorithm is sensitive to the selection of initial value, to choose a reasonable initial value $\theta^{(0)}$ will increase the accuracy of EM algorithm.

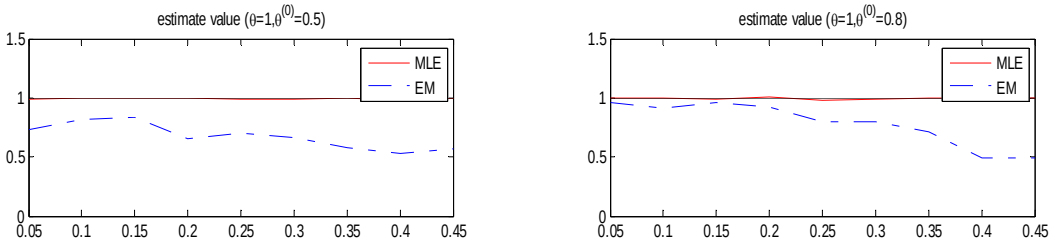

Figure 6. Likelihood Estimates and EM Estimates and their MSE ($\theta = 1$, $\theta^{(0)} = 0.5,0.8$)

4. Conclusion

EM algorithm performs well in parameter estimation with missing data. Through computer simulated calculation, it can be seen that when the missing rate is less than 0.30, EM algorithm is almost identical to likelihood estimation , when the missing rate is greater than 0.30, the difference between the two parameter estimation methods is increasing. Simulated calculation indicates that sample size the selection of initial value make an effect on the accuracy of EM algorithm, and it needs further study to choose a reasonable initial value of EM algorithm according to different problems.

Acknowledgements

The research is supported by Science and Technology Support Project of Gansu Province (Project No. 1204GKCA010).

References

[1] Ling Wei, Jianjun Qi, Yimin Shi. The EB Estimation of Scale-parameter for the Two-Parameter Exponential Distribution Under the Type-II Censoring Life Test. *Mathematica App licata*. 2001; 14(4): 66-701.

[2] Yushuang Liu, Lixin Song. The EB Estimation of Scale-parameter Under Type-II Censoring Life Test for Two-parameter Weibull Distribution. *Journal of Jilin Normal university*. 2004; 5(2): 16~18.

[3] Zhendong Li, Yingshu Liao. *Impact of Missing Data on Parameter Estimation of EM Algorithm Under Exponential Distribution*. The International Conference on Automatic Control and Artificial Intelligence (ACAI 2012). Xiamen. 2012; 3597-3599.
[4] Ng HKT, Chan PS, Balakrishnan N. Estimation of Parameters From Progressively Censored Data Using EM Algorithm. *Computational Statistics & Data Analysis*. 2002; 39:371- 386.
[5] Wu CFJ. On the Convergence Properties of the EM Algorithm. *The Annals of Statistics*. 1983; 11:95-103.