A Monte Carlo Expectation Maximization Algorithm for Statistical Inference of Weibull Process with Left Censored Data

Jun-Ming Hu\textsuperscript{1,2,3} and Hong-Zhong Huang\textsuperscript{1,3}

\textsuperscript{1} School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu, Sichuan 611731, P.R. China
\textsuperscript{2} School of Transportation and Automotive Engineering, Xihua University, Chengdu, Sichuan 610039, P.R. China
\textsuperscript{3} Center for System Reliability and Safety, University of Electronic Science and Technology of China, Chengdu, Sichuan 611731, P.R. China

E-mail: hujunming@std.uestc.edu

Abstract. The Weibull process plays an important role in the failure analysis of repairable systems. In practice, there exists a situation that the data collected are incomplete. Some of the failure data are missing due to various reasons. Statistical inferences of a Weibull process with incomplete data using Monte Carlo expectation maximization algorithm is proposed. The estimation procedures are derived. A case study is performed to illustrate and compare the performances of this algorithm. It is observed that this method is effective and can simplify the estimation.

1. Introduction

Maintenance is a crucial factor for repairable systems to maintain their performances. Trend analysis provides a powerful method to help maintenance policy making and reliability assessment. It aims to investigate whether the failures collected exhibit a decreasing or increasing trend by statistical modeling the failure intensity of repairable systems. Multiple methods are proposed for modeling failure intensity. One typical form of them is the Weibull process that has a time dependent failure intensity same as the failure rate of Weibull distribution. Due to ease of use and simplicity, the Weibull process is widely used for trend analysis of repairable systems. It is not only adopted to model the reliability growth [1], but also utilized for minimal repairs [2]. Relevant problems on the Weibull process attract much attention, but a great deal of research rests on the assumption that the failure data are reported exactly [3-8].

However, there exists incomplete data because of human error or malfunction of sensors in practice. Incomplete data may result in a distorted estimation [9]. Thus, analyzing incomplete failure data is critical to inform decision-making. Taghipour and Banjevic [10] estimated a power law intensity with interval-censored data using expectation maximization algorithm. Peng et al. [11] proposed a method to estimate and predict the reliability of repairable systems subject to interval censoring by the Monte Carlo expectation maximization algorithm. Yu et al. [12] studied statistical analysis for the Weibull process with left censored data via the classical method. Na and Chang [13] proposed a model to estimate the failure intensity given left censored data from multiple machines.
Chumnaul and Sepehrifar [14] studied the generalized confidence interval for the scale parameter of the Weibull process via generalized pivotal quantity. However, the fact that the successive failure times of a power law process are dependent results in a complicated derivation. The dependent variables can be transferred into independent variables from the order statistic model. In this paper, a Monte Carlo expectation maximization algorithm is utilized to simplify the estimation from the perspective of the order statistic.

The rest of the paper is structured as follows. Section 2 gives a brief introduction to the Weibull process. Section 3 gives the Monte Carlo expectation maximization for the Weibull process with left censored data. A simulation case is used to illustrate this proposed method in Section 4. Section 5 concludes this work.

2. Weibull Process
The Weibull process is a nonhomogeneous Poisson process with a time-variant failure intensity as

$$\lambda(t) = \beta t^{\beta-1} \alpha^{-\beta}, \alpha > 0, \beta > 0,$$

where $\beta$ is the shape parameter, and $\alpha$ is the scale parameter. It has been widely used to model the failure of repairable systems because of its simplicity and flexibility. The shape parameter $\beta$ dominates the failure intensity curve. When $\beta > 1$, the failure intensity is increasing which implies failures occur more frequently. If $\beta = 1$, the failure intensity is a constant and the process is reduced to a homogeneous Poisson process. For $\beta < 1$, there exists a decreasing failure intensity which represents the system reliability is improving.

2.1. Mean value function
Let $N(t)$ denote the number of failures occurred up to time $t$ and $m(t)$ stand for the expected number of failures through time $t$, i.e., the mean value function for the process. $N(t)$ is a random variable which follows a Poisson distribution with parameter $m(t)$. $m(t)$ is defined by:

$$m(t) = \int_0^t \lambda(t)dt = \frac{(t)^\beta}{\alpha},$$

Thus, the probability mass function of $N(t)$ is given by:

$$P[N(t) = k] = \frac{(m(t))^k e^{-m(t)}}{k!},$$

where $k$ is a non-negative integer, $k!$ is the factorial of $k$.

2.2. Likelihood for complete data
Let $T = (T_1, T_2, ..., T_n)$ denote the successive failure data of a Weibull process where $T_i$ is the arrival time of the $i$ th failure. It has been proved that the first failure time $T_1$ follows a Weibull distribution with the same shape parameter and scale parameters as that of the failure intensity of the Weibull process and $T_i, i \geq 2$ follows a left-truncated Weibull distribution with the truncation point $t_{i-1}$. Taking account of the independent increments of the power law process, it is easy to derive the joint likelihood of a realization of $T$.

In practice, there exist two types of failure data, time truncated data and failure truncated data. If the failures are observed until a prespecified time $t$, the data is termed as time truncated data. If the failures are recorded until the predetermined failure occurs (e.g., the $n$ th failure), the data is regarded as failure truncated data. The different terminal time results in a little difference between the inference procedures for the time truncated data and failure truncated data. Given the time truncated data
\( T = (t_1, t_2, \ldots, t_n) \) and the terminal time \( t_c \), the joint likelihood function of \( 0 < t_1 < t_2 < \cdots < t_n < t_c \) is defined by:

\[
f(t_1, t_2, \cdots, t_n) = \frac{\beta^n}{\alpha^n} \left( \prod_{i=1}^{n} t_i \right)^{\beta-1} e^{-\left( \frac{t_c}{\alpha} \right)^\beta},
\]

For a failure truncated case, the joint likelihood function of \( 0 < t_1 < t_2 < \cdots < t_n \) is defined by:

\[
f(t_1, t_2, \cdots, t_n) = \frac{\beta^n}{\alpha^n} \left( \prod_{i=1}^{n} t_i \right)^{\beta-1} e^{-\left( \frac{t_c}{\alpha} \right)^\beta}.
\]

2.3. The conditional probability distribution
Combining Eq. (1) and (3), it can be easily derived that the conditional distribution of \( T_1 < T_2 < \cdots < T_n \) given \( N(t) = n \) is that of \( n \) order statistics whose cumulative distribution function is

\[
F(T_i) = \frac{m(t_i)}{m(t_c)} = \left( \frac{t_i}{t_c} \right)^\beta, \quad 0 \leq t_i < t_c.
\]

The corresponding probability density function is

\[
f(t_i) = \beta t_i^{\beta-1} t_c^{-\beta}, \quad 0 \leq t_i < t_c.
\]

The conditional distribution of \( T_1 < T_2 < \cdots < T_n \) given \( N(t) = n \) is that of \( n-1 \) order statistics whose cumulative distribution function is

\[
F(t_i) = \frac{m(t_i)}{m(t_n)} = \left( \frac{t_i}{t_n} \alpha \right)^\beta = \left( \frac{t_i}{t_n} \right)^\beta, \quad i = 1, 2, \ldots, n-1.
\]

3. Monte Carlo expectation maximization algorithm
The expectation maximization algorithm proposed by Dempster et al. [15] is an effective method to handle the incomplete data. For a left censored data, the failures that occurred at the early stage of the testing are missing just as shown in Figure 1. Let us denote the missing data and the observed data by \( Z = (Z_1, Z_2, \ldots, Z_m) \) and \( X = (X_1, X_2, \ldots, X_n) \) separately. Here for a given \( m, Z_1, Z_2, \ldots, Z_m \) are not observable, thus they are thought of as missing data. The combination of \( W = (Z, X) \) comprises the complete data.

\[
\begin{array}{cccccccc}
& 1^\text{st} & 2^\text{nd} & \cdots & m^\text{th} & (m+1)^\text{th} & \cdots & (m+n)^\text{th} \\
\text{failure} & \text{failure} & \cdots & \text{failure} & \text{failure} & \cdots & \text{failure} & \\
\end{array}
\]

\[0 \quad z_1 \quad z_2 \quad \cdots \quad z_m \quad x_1 \quad \cdots \quad x_n \quad t \]

Figure 1. Left censored data for a Weibull process.

The log-likelihood function based on the complete data \( W \) is

\[
I(\alpha, \beta; W) = (m + n) \log \beta - (m + n) \beta \log \alpha + (\beta - 1) \sum_{i=1}^{m+n} \log w_i - \left( \frac{t}{\alpha} \right)^\beta \]

\[
= (m + n) \log \beta - (m + n) \beta \log \alpha + (\beta - 1) \left( \sum_{i=1}^{m} \log z_i + \sum_{i=1}^{n} \log x_i \right) - \left( \frac{t}{\alpha} \right)^\beta.
\]
For the time truncated case, \( t \) is equal to \( t_c \), and for the failure truncated case, \( t \) is equal to \( t_r \).

### 3.1. Expectation maximization algorithm

The expectation maximization algorithm estimates the parameters by filling in the missing data with estimated values and updating the estimates iteratively. Each iteration of this algorithm consists of two steps: E-step and M-step. The E-step is to compute the expected complete data log-likelihood function given the observed data at the current parameter estimate \( E[l(\alpha, \beta; W) | X] \). Therefore,

\[
E[l(\alpha, \beta; W) | X] = (m + n) \log \beta - (m + n) \beta \log \alpha + (\beta - 1) \sum_{i=1}^{n} \log x_i - \left( \frac{t}{\alpha} \right)^\beta
\]

\[
+ (\beta - 1) E \left[ \sum_{i=1}^{n} \log Z_i \left| Z_i < X_1 \right. \right]
\]

(8)

The M-step is to maximize the pseudo log-likelihood function \( E[l(\alpha, \beta; W) | X] \) and update the parameter estimates. If the estimate of \((\alpha, \beta)\) at the \( k \) th stage is \((\alpha^k, \beta^k)\), then \((\alpha^{k+1}, \beta^{k+1})\) can be obtained by maximizing Eq. (8).

### 3.2. Monte Carlo expectation maximization

Because the variables \( Z_1, Z_2, \ldots, Z_m \) are dependent, the integral term \( E \left[ \sum_{i=1}^{n} \log Z_i \left| Z_i < X_1 \right. \right] \) is complicated and no close form solution of this integral exists. Trying to estimate the conditional expectation by simulation provides an alternative to overcome this difficulty [16].

Taking samples \( (z_i, z_2, \ldots, z_m) \) from the Weibull process given \((\alpha, \beta)\) is critical to the approximation. According to Eq. (5) and (6), order statistic can simplify the sampling procedures. The sampling procedures are summarized in Table 1.

**Table 1.** Sampling procedures for \((z_i, z_2, \ldots, z_m)\).

| Step 1. Setting the cumulative number of missing data \( m \), the terminal time \( t \), and the \( k \) th stage estimates \((\alpha^k, \beta^k)\); |
| Step 2. Drawing \( m \) independent identical distribution random samples \( u_i, i = 1, 2, \ldots, m \) from a uniform distribution over \((0, (x_i)/m(t))\); |
| Step 3. Based on the Eq. (5) or (6), transforming \( m \) samples \( u_i \) into samples of missing failure times \( z_i \), i.e., \( z_i = \{t(u_i)\}^{\frac{1}{\beta}} \); |
| Step 4. Sorting the above samples \( z_i \) in ascending order to form the missing data. |

Repeating the sample procedures \( s \) times, \((z_1, z_2, \ldots, z_m), (z_1, z_2, \ldots, z_m), \ldots, (z_1, z_2, \ldots, z_m)\) are generated. Thus, given \((\alpha^k, \beta^k)\) the conditional expectation can be approximated by

\[
E \left[ \sum_{i=1}^{m} \log Z_i \left| Z_i < X_1, \alpha^k, \beta^k \right. \right] = \frac{1}{s} \sum_{j=1}^{s} \sum_{i=1}^{m} \log z_i^{j}
\]

(9)

In the M-step, we update the parameter estimates by maximizing Eq. (8). Taking derivatives to \( \alpha \) and \( \beta \) of Eq. (8), we obtain

\[
\frac{\partial E[l(\alpha, \beta; W) | X]}{\partial \alpha} = -\frac{(m + n) \beta}{\alpha} \left( \frac{t}{\alpha} \right)^\beta
\]

\[
\frac{\partial E[l(\alpha, \beta; W) | X]}{\partial \beta} = \frac{(m + n) \alpha}{\beta} \left( \frac{t}{\alpha} \right)^\beta
\]

(10)
\[
\frac{\partial E[l(\alpha, \beta; W) \mid X]}{\partial \beta} = \frac{(m + n)}{\beta} \left[ (m + n) \log \alpha + \sum_{i=1}^{n} \log x_i - \left( \frac{t}{\alpha} \right) \log \left( \frac{t}{\alpha} \right) \right] + E[\sum_{i=1}^{n} \log z_i \mid Z_i < X_i, \alpha^k, \beta^k] 
\]

(11)

Setting Eq. (10) and (11) equal to zero, we can find \( \alpha^{k+1} \) and \( \beta^{k+1} \) as

\[
\alpha^{k+1} = \frac{t}{(m + n)^{1/\alpha}} 
\]

(12)

\[
\beta^{k+1} = \frac{m + n}{(m + n) \log t - \sum_{i=1}^{n} \log x_i - E[\sum_{i=1}^{n} \log z_i \mid Z_i < X_i, \alpha^k, \beta^k]} 
\]

(13)

The procedures of Monte Carlo expectation maximization for parameter estimation are summarized in Table 2.

### Table 2. Procedures for Monte Carlo expectation maximization.

1. **Step 1.** Setting the parameter \( (\alpha^0, \beta^0) \);
2. **Step 2.** Drawing samples \((z_1, z_2, \ldots, z_m)^t, (z_1, z_2, \ldots, z_m)^t, \ldots, (z_1, z_2, \ldots, z_m)^t\) from the Weibull process based on the algorithm in Table 1;
3. **Step 3.** Calculating the conditional expectation \( E[\sum_{i=1}^{m} \log z_i \mid Z_i < X_i, \alpha^k, \beta^k] \), according to the Equation (9);
4. **Step 4.** Computing \( (\alpha^{k+1}, \beta^{k+1}) \) according to the Equation (12) and (13), and update \( (\alpha^k, \beta^k) \) by \( (\alpha^{k+1}, \beta^{k+1}) \);
5. **Step 5:** output the parameter estimation \( (\alpha, \beta) \).

### 4. Case study

In this section, an engine failure data from Chumnaul and Sepehrifar [14] is used to demonstrate the proposed methodologies. The total failure times during the system development testing are listed in Table 3, where * denotes the time is missing.

### Table 3. Failure data.

|   |   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|---|
| * | * | * | 171 | 234 | 274 | 377 | 530 |
| 1074 | 1188 | 1248 | 2298 | 2347 | 2347 | 2381 | 2456 |
| 2913 | 3022 | 3038 | 3728 | 3873 | 4724 | 5147 | 5179 |
| 6824 | 6983 | 7106 | 7106 | 7568 | 7568 | 7593 | 7642 |
| 7928 | 8063 | 171 | 2347 | 2347 | 2381 | 2456 | 2500 |
| 1074 | 1188 | 1248 | 2298 | 2347 | 2347 | 2381 | 2456 |
| 2913 | 3022 | 3038 | 3728 | 3873 | 4724 | 5147 | 5179 |
| 6824 | 6983 | 7106 | 7106 | 7568 | 7568 | 7593 | 7642 |
| 7928 | 8063 | 171 | 2347 | 2347 | 2381 | 2456 | 2500 |

According to Table 3, \( m = 3, n = 37, x_i = 171 \), and \( t = 8063 \). Three different initial parameters \( (\alpha^0, \beta^0) \) are used and 1000 missing data samples are drawn. Using the Monte Carlo expectation maximization, we attain the parameter estimate converged. Figure 2 shows the dynamic traces of iteration under three different initial values of \( \beta \).
Figure 2. Dynamic traces of $\beta$.

The solid line in Figure 2 denotes the trace of $\beta^0=0.3$; the dashed line is the trace of $\beta^0=0.6$; and the dotted line represents the trace of $\beta^0=2$. Although different initial values of $\beta$ lead to different iteration numbers, all the three settings converge finally. The detail estimates of $\beta$ and $\alpha$ are listed in Table 4.

Table 4. Parameter estimate under different initial parameters.

| $(\alpha^0, \beta^0)$ | Parameter estimate converged |
|------------------------|----------------------------|
| (0.0368, 0.3)          | 34.4477 0.6762             |
| (17.2344, 0.6)         | 34.4230 0.6761             |
| (1274.8722, 2)         | 34.4635 0.6762             |

Table 4 shows that all the different initial parameters are updated to converged values which are close mutually. The estimates proposed by Chumnaul and Sepehrifar [14] are $\hat{\beta}=0.6761$ and $\hat{\alpha}=34.4219$. There is no significant difference among them. It demonstrates that the method proposed is effective.

5. Conclusions

A Monte Carlo expectation maximization algorithm is proposed to estimate unknown parameters of the Weibull process with left censored data. The case study shows that this method is effective. Order statistics provides an alternative to draw samples to approximate the missing data. This framework is promising and could be extended to interval censored data. Other consideration regarding this method is the effect of the sample numbers specified on the convergence which is also worth studying further.
Acknowledgments
Thanks to the National Natural Science Foundation of China for financial support under the contract number 51775090.

References
[1] Wang W, Xu Y and Hou L 2019 Optimal allocation of test times for reliability growth testing with interval-valued model parameters P. I. Mech. Eng. O-J. Ris. 233 791-802
[2] Asfaw Z G and Lindqvist B H 2015 Extending minimal repair models for repairable systems: A comparison of dynamic and heterogeneous extensions of a nonhomogeneous Poisson process Reliab. Eng. Syst. Saf. 140 53-8
[3] Wang Y-P and Lu Z-Z 2013 Bayesian inference and prediction analysis of the power law process based on a natural conjugate prior J. Stat. Comput. Simul. 85 881-98
[4] Wang Y-P and Lu Z-Z 2011 Bayesian inference and prediction analysis of the power law process based on a gamma prior distribution Commun. Stat. Simul. Comput. 40 1383-401
[5] Wang Y, Lu Z and Xiao S 2018 Parametric bootstrap confidence interval method for the power law process with applications to multiple repairable systems IEEE Access 6 49157-69
[6] Tang Y, Fu J, Liu W and Xu A 2017 Bayesian analysis of repairable systems with modulated power law process Appl. Math. Modell. 44 357-73
[7] Somboonsavatdee A and Sen A 2015 Statistical inference for power-law process with competing risks Technometrics 57 112-22
[8] Yu J W, Tian G L and Tang M L 2007 Predictive analyses for nonhomogeneous poisson processes with power law using bayesian approach Comput. Stat. Data Anal. 51 4254-68
[9] Si W, Yang Q, Monplaisir L and Chen Y 2018 Reliability analysis of repairable systems with incomplete failure time data IEEE Trans. Reliab. 67 1043-59
[10] Taghipour S and Banjevic D 2011 Trend analysis of the power law process using expectation–maximization algorithm for data censored by inspection intervals Reliab. Eng. Syst. Saf. 96 1340-48
[11] Peng Y, Wang Y, Zi Y, Tsui K-L and Zhang C 2017 Dynamic reliability assessment and prediction for repairable systems with interval-censored data Reliab. Eng. Syst. Saf. 159 301-09
[12] Yu J W, Tian G L, and Tang M L 2008 Statistical inference and prediction for the Weibull process with incomplete observation Comput. Stat. Data Anal. 52 1587-603
[13] Na I Y and Chang W 2017 Multi-system reliability trend analysis model using incomplete data with application to tank maintenance Qual. Reliab. Eng. Int. 33 2385-95
[14] Chumnaul J and Sepehrifar M 2018 Generalized confidence interval for the scale parameter of the power-law process with incomplete failure data Comput. Stat. Data Anal. 128 17-33
[15] Dempster A P, Laird N M and Rubin D B 1977 Maximum likelihood from incomplete data via the EM algorithm J. Roy. Statist. Soc. Ser. B 39 1-38
[16] Celeux G and Diebolt J 1985 The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. Comput. Stat. Q. 2 73-82