A Stochastic EM Algorithm for Progressively Censored Data Analysis

Mimi ZHANG, Zhi-Sheng YE and Min XIE

Department of Systems Engineering & Engineering Management, City University of Hong Kong

Abstract

Progressive censoring technique is useful in lifetime data analysis. Simple approaches to progressive data analysis are crucial for its widespread adoption by reliability engineers. This study develops an efficient yet easy-to-implement framework for analyzing progressively censored data by making use of the stochastic EM algorithm. Based on this framework, we develop specific stochastic EM procedures for several popular lifetime models. These procedures are shown to be very simple. We then demonstrate the applicability and efficiency of the stochastic EM algorithm by a fatigue life dataset with proper modification and by a progressively censored dataset from a life test on hard disk drives.

Key words: stochastic EM algorithm; progressively censored data; lifetime distribution

1 Introduction

Censoring is a common phenomenon in most life-testing experiments due to time constraints. Among the existing censoring schemes, the progressive censoring schemes have become very popular in the last few decades because of its flexibility in removing un-failed units from the

tests. Generally speaking, a progressively censored sample of size n consists of m failures and $n-m$ progressively censored observations. Censoring may happen right after a failure occurrence, i.e., remove R_i functioning units upon the ith failure, which is called progressive type II censoring. The progressive Type II censoring scheme has been applied to Burr-XII distributions (Wang and Cheng[1]), Weibull distributions (Pareek et al.[2]), Gaussian distributions (Balakrishnan et al.[3]), exponential distributions (Lee and Pan[4]) and log-logistic distributions (Balakrishnan and Saleh[5]), etc. Censoring can also happen in a random manner, which is common in medical studies, e.g., Davis and Feldstein[6]. Another example of random censoring is a life test conducted on a batch of raw hard disk drives (HDDs) in Seagate, a leading HDD company. Detailed description of the test will be introduced in Section 4. However, a problem faced by the HDD engineers is a lack of simple tools to analyze the data.

As noted by Ng et al.[7], the complicated calculation of the likelihood function of progressively censored data when deriving the maximum likelihood estimates (MLEs) has greatly restricted the wide adoption of this scheme by reliability engineers. What engineers need is an efficient yet simple tool to compute the MLE. Employment of the standard or modified Newton–Raphson algorithm requires the Hessian matrix of the likelihood function, which differs from distributions to distributions and whose expression is most often quite complicated. An intuitive yet brute-force approach to maximizing the likelihood function of the progressively censored data is to directly apply some derivative-free algorithms, e.g., the ones reviewed by Kolda et al.[8]. However, as is well-known in the optimization literature, an algorithm has to visit every point in the feasible region in order to guarantee the global optimality, which is almost impossible when the parameter space is continuous. Moreover, most of these derivative-free algorithms will eventually converge to some local optimal points far away from the global optimum, if an educated starting point is not available.

Another popular means to handle the progressively censored data is to treat them as a missing data problem, and thus the EM algorithm can be invoked, e.g., see Ng et al.[7] for the Weibull and log-normal distribution, Lin et al.[9] for log-gamma distributions and Pradhan and Kundu[10] for generalized exponential distribution, among others. Because the EM algorithm relies on complete-data computations, it is generally simple to implement. A typical EM algorithm iteratively applies two steps, i.e. the expectation step (E-Step) and the maximization step (M-Step). The E-step involves taking expectation over complete-data conditional distribution, while the M-Step only involves maximum likelihood estimation

based on complete data, which often has a simple closed form. However, with the increasing complexity of the progressively censored data and lifetime model, one of the biggest shortcomings of EM is that it is only a local optimization procedure and can easily get stuck in a saddle point. Moreover, when the E-step involves intricate or even infeasible computation, the EM paradigm is no longer directly applicable. A possible solution to overcoming the computational inefficiencies, i.e., the intractable E-step and the saddle point problem, is to invoke stochastic EM implementations such as the Monte Carlo EM algorithm (Wei and Tanner[11]). The Monte Carlo EM algorithm approximates the expectation in the E-Step by the Monte Carlo average. Therefore, the maximization of the averaged log-likelihood may be very complicated and thus more time-consuming, e.g., see Wang and Cheng[1] for an application of the Monte Carlo EM algorithm to the Burr-XII distribution.

The stochastic EM (SEM) algorithm proposed by Celeux and Diebolt[12] is also a stochastic version of the EM implementations as a way for executing the E-step by simulation. A very attractive merit of this algorithm is that it replaces the E-Step with an S-Step, which is very easy to implement whatever the underlying distribution and the missing data are. Compared with the Monte Carlo EM algorithm, the SEM algorithm completes the observed sample by replacing each missing datum by a value randomly drawn from the distribution conditional on results from the previous step. The M-step is thus a complete-data maximum likelihood estimation, which is often very easy to solve. The SEM algorithm has been shown to be computationally less burdensome and more appropriate than the EM algorithm in a lot of problems (Celeux and Diebolt[12], Tregouet et al.[13], Delignon et al.[14], Cariou and Chehdi[15]). It is shown by Nielsen[16] that the SEM algorithm always converges to some local optimum. Some applications of the SEM algorithm suggest that this algorithm tends to converge to the global optimum or a non-significant local optimum (Diebolt and Celeux[17], Cariou and Chehdi[15], Svensson and Sjöstedt-de Luna[18]).

Motivated by the need of the HDD engineers and inspired by the numerous attractive properties of the SEM algorithm and the current computational challenges faced by the analysis of progressively censored data, the broad objective of this paper is to promote the SEM algorithm to the analysis of progressively censored data. More specifically, we develop a generic framework for the parametric maximum likelihood inference of progressively censored data. Under this framework, the S-Step imputes a single value for the censored data from the original distribution truncated at the left by making use of the parameters estimated

from the previous SEM cycle, while the M-Step maximizes a complete sample likelihood function, which can be easily accomplished by most statistical software, e.g., Matlab, R, JMP, Minitab, SAS, SPSS, and S–PLUS. The procedure is thus very easy to implement yet efficient, which meets the needs of engineers. This framework is then applied to several common distributions, including the Weibull, lognormal, inverse Gaussian, and Birnbaum-Saunders distributions, etc. We only focus on point estimation. Confidence intervals for the parameters can be constructed based on the Hessian matrixes derived in the literature for each distribution, (Wang and Cheng[1], Pareek et al.[2], Balakrishnan et al.[3], Balakrishnan and Saleh[5]).

The remainder of this paper is organized as follows. Section 2 develops the framework of progressively censored data analysis using the SEM algorithm. Section 3 elaborates on how this framework can be applied to a number of common distributions. In Section 3.5, a dataset from Birnbaum and Saunders[19] and Ng et al.[20], after proper modification, is fitted by the Birnbaum-Saunders distribution to demonstrate the simplicity of our method. We also apply the methodology to analyze a real dataset from a HDD test. Section 5 concludes the paper and points out possible topics for future research.

2 Analysis of Progressively Censored Data via SEM: A General Framework

To simplify the notation, we consider data from a progressive type-II right censoring scheme, because of its popularity and its standards in notations. But we shall underscore that the framework applies to general progressively censored data, as will be demonstrated in Section 4.2. Under this scheme, $n \in \mathbb{N}$ identical units are placed on a life-test. Their lifetimes are described by independent and identically distributed random variables $T_1, \dots, T_n$, each with probability density function (PDF) $f(t; \Theta)$ and cumulative distribution function (CDF) $F(t; \Theta)$, where Θ denotes the vector of model parameters. At the time of the first failure t_1, R_1 of the $n-1$ surviving units are randomly withdrawn from the experiment. R_2 of the $n-2-R_1$ surviving units are withdrawn at the time of the second failure t_2 and so on. Finally, at the time of the mth failure t_m, all the remaining $R_m = n - m - R_1 - \cdots - R_{m-1}$ surviving units are withdrawn. The scheme $(R_1, R_2, \dots, R_m)$ is referred to as progressive

Type-II right censoring scheme with $R_j > 0$ and $\sum_{j=1}^{m} R_j + m = n$. The likelihood function based on the progressively censored data is

$$L(\Theta) = c \prod_{j=1}^{m} f(t_j;\ \Theta)[1 - F(t_j;\ \Theta)]^{R_j} \tag{1}$$

where $c = n(n - R_1 - 1) \ldots (n - R_1 - R_2 - \cdots - R_{m-1} - m + 1)$. Equation (1) is generally difficult to optimize directly.

Denote the observed (ordered) failure data by $\boldsymbol{T} = (T_{1:m:n}, T_{2:m:n}, \ldots, T_{m:m:n})$ and the unobserved censored data by $\boldsymbol{Z} = (\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_m)$. $\boldsymbol{Z}_j$ $(j = 1, \ldots, m)$ is a $1 \times R_j$ random vector with $\boldsymbol{Z}_j = \left(Z_{j1}, Z_{j2}, \ldots, Z_{jR_j}\right)$ which represents the survival times of R_j withdrawn units. The conditional distribution of the random variable Z_{jl} $(l = 1, \ldots, R_j)$ given $T_{j:m:n} = t_j$ $(j = 1, \ldots, m)$ is given by (Ng et al.[7])

$$G_j\left(z;\ \Theta \middle| T_{j:m:n} = t_j\right) = \frac{F(z;\ \Theta) - F\left(t_j;\ \Theta\right)}{1 - F\left(t_j;\ \Theta\right)} \qquad z > t_j. \tag{2}$$

The unobserved $\boldsymbol{Z}$ can be regarded as missing data. Given the complete data $\boldsymbol{T}$ and $\boldsymbol{Z}$, the log-likelihood function of the complete sample can be specified as

$$Q(\Theta) = \sum_{j=1}^{m} \log f(t_j;\ \Theta) + \sum_{j=1}^{m} \sum_{l=1}^{R_j} \log f(z_{jl};\ \Theta), \tag{3}$$

The EM algorithm is an iterative algorithm. The E-step of the regular EM algorithm obtains the *Q*-function by taking the expectation of (3) with respect to the $\boldsymbol{Z}$ conditional on observed data $\boldsymbol{Z}$ and $\Theta^{(k)}$, the current value of Θ after *k* cycles of the EM algorithm. In view of the difficulty that the expectation is often very difficult, or even intractable to compute, the main idea of the SEM algorithm is to replace the E-step by a stochastic step where the missing data $\boldsymbol{Z}$ are imputed with a single draw from the distribution of the missing data conditional on the observed $\boldsymbol{T}$. The imputed $\boldsymbol{Z}$ are then substituted to (3) to form the pseudo *Q*-function, which is then optimized in the M-step to obtain $\Theta^{(k+1)}$ for the next cycle. More formally, given the parameter estimate $\Theta^{(k)}$ at the kth SEM cycle, the $(k + 1)$st cycle of the SEM algorithm evolves as follows:

S-Step. Given the current $\Theta^{(k)}$, simulate R_j independent values from the conditional distribution $G_j\left(z;\Theta^{(k)}|T_{j:m:n}=t_j\right)$ respectively for $j=1,\ldots,m$ to form a realization of $\boldsymbol{Z}$.

M-Step. Maximize the pseudo *Q*-function given $(\boldsymbol{T},\boldsymbol{Z})$ to obtain $\Theta^{(k+1)}$.

The S-Step completes the data set in a very simple way, while the maximization in the M-Step deals with a complete sample. Hence the M-Step is easy to solve either explicitly or iteratively for most commonly-used distributions. Under mild regularity conditions, the sequence of $\{\Theta^{(k)}\}$ starting from a specific $\Theta^{(0)}$ converges to a stationary distribution whose mean is a consistent and asymptotically efficient estimator of parameters for the progressively censored data, i.e., the large sample variance-covariance matrix of the mean is exactly the Fisher information matrix (Diebolt and Celeux[17], Diebolt and Ip[21], Nielsen[16], Svensson and Sjöstedt-de Luna[18]). It is easy to show that the regularity condition on smoothness of the underlying model is met for distributions discussed in Section 3. Therefore, we can run the SEM algorithm for a specific number of iterations, discard a few initial iterations for burn-in purpose, and average over $\Theta^{(k)}$ of the remaining iterations to obtain the MLEs. According to our experience, a burn-in period of 100 cycles is usually long enough under moderate censoring. A trace plot of the $\{\Theta^{(k)}\}$ sequence versus the cycles is also helpful for checking sufficiency of the burn-in period or for determining a more appropriate burn-in duration.

3 SEM Algorithms for Some Commonly Used Lifetime Distributions

To show the wide applicability of the framework developed in Section 2 and to elaborate on the flexibility of the SEM algorithm in handling progressively censored data, this section applied this framework to some common distributions used in lifetime data analysis in order.

3.1 Birnbaum-Saunders Lifetime Data

The Birnbaum-Saunders distribution proposed by Birnbaum and Saunders[19] has been successfully applied to model failures due to crack. In addition, it is often used to approximate the distribution of failures caused by independent accumulation of damage. The

respective CDF and PDF of a two-parameter Birnbaum-Saunders random variable with a shape parameter μ and a scale parameter λ can be written as

$$F(t;\,\mu,\lambda)=\Phi\left\{\frac{1}{\mu}\left[\left(\frac{t}{\lambda}\right)^{\frac{1}{2}}-\left(\frac{\lambda}{t}\right)^{\frac{1}{2}}\right]\right\},\quad t>0,$$

$$f(t;\,\mu,\lambda)=\frac{1}{2\sqrt{2\pi}\mu\lambda}\left[\left(\frac{\lambda}{t}\right)^{\frac{1}{2}}+\left(\frac{\lambda}{t}\right)^{\frac{3}{2}}\right]\times\exp\left[-\frac{1}{2\mu^2}\left(\frac{t}{\lambda}+\frac{\lambda}{t}-2\right)\right],\quad t>0.$$

In the literature, Ng et al.[20] have studied MLE of this distribution under type-II censoring by means of direct optimization of the likelihood function and the Monte Carlo EM algorithm. Their procedures tend to be very complicated due to the complex form of the likelihood function. On the contrary, the SEM algorithm is a handy gadget to handle this kind of censored data. Consider the progressively censored data $(\boldsymbol{T},\boldsymbol{Z})$ generated from this distribution. Denoting $\Theta^{(k)}=\left(\mu^{(k)},\lambda^{(k)}\right)$ the value of Θ at the kth SEM cycle, the $(k+1)$st cycle proceeds as follows.

The S-Step

The conditional distribution function of the unobserved Z_{jl} $(l=1,\ldots,R_j)$ given $T_{j:m:n}=t_j$ $(j=1,\ldots,m)$ is

$$G_j\left(z;\mu,\lambda\middle|T_{j:m:n}=t_j\right)=\frac{F(z;\,\mu,\lambda)-F\left(t_j;\,\mu,\lambda\right)}{1-F\left(t_j;\,\mu,\lambda\right)}\qquad t>t_j. \tag{4}$$

In other words, given $T_{j:m:n}=t_j$, Z_{jl} is a Birnbaum-Saunders variable left-truncated at t_j. Given (4), a random realization of $\boldsymbol{Z}$ is readily generated from $G_j\left(z;\,\mu^{(k)},\lambda^{(k)}|T_{j:m:n}=t_j\right)$.

The M-Step

Given the observed data $\boldsymbol{T}$ and the imputed $\boldsymbol{Z}$ from the S-Step, the pseudo *Q*-function can be expressed as

$$Q(\mu,\lambda) = -\mathrm{n}\log 2\sqrt{2\pi} - \mathrm{n}\log\mu - \mathrm{n}\log\lambda + \sum_{i=1}^{m}\log\left[\left(\frac{\lambda}{t_i}\right)^{\frac{1}{2}} + \left(\frac{\lambda}{t_i}\right)^{\frac{3}{2}}\right]$$
$$- \sum_{i=1}^{m}\frac{1}{2\mu^2}\left(\frac{t_i}{\lambda} + \frac{\lambda}{t_i} - 2\right) + \sum_{i=1}^{m}\sum_{j=1}^{R_j}\log\left[\left(\frac{\lambda}{z_{ij}}\right)^{\frac{1}{2}} + \left(\frac{\lambda}{z_{ij}}\right)^{\frac{3}{2}}\right]$$
$$- \sum_{i=1}^{m}\sum_{j=1}^{R_j}\frac{1}{2\mu^2}\left(\frac{z_{ij}}{\lambda} + \frac{\lambda}{z_{ij}} - 2\right).$$

To maximize this *Q*-function, we define the following function

$$K(x) = \left[\frac{1}{n}\left(\sum_{i=1}^{m}(x+t_i)^{-1} + \sum_{i=1}^{m}\sum_{j=1}^{R_j}(x+z_{ij})^{-1}\right)\right]^{-1} \qquad x \geq 0,$$

where

$$s = \frac{1}{n}\left(\sum_{i=1}^{m} t_i + \sum_{i=1}^{m}\sum_{j=1}^{R_j} z_{ij}\right), \qquad r = \left[\frac{1}{n}\left(\sum_{i=1}^{m} t_i{}^{-1} + \sum_{i=1}^{m}\sum_{j=1}^{R_j} z_{ij}{}^{-1}\right)\right]^{-1}.$$

are the sample arithmetic and the sample harmonic means, respectively. From the results of complete sample MLE discussed by Birnbaum and Saunders[19], we can see that the value of $\lambda^{(k+1)}$ is the unique positive root of the equation

$$\lambda^2 - \lambda[2r + K(\lambda)] + r[s + K(\lambda)] = 0.$$

Once $\lambda^{(k+1)}$ is obtained, $\mu^{(k+1)}$ can be derived explicitly as

$$\mu^{(k+1)} = \left[\frac{s}{\lambda^{(k+1)}} + \frac{\lambda^{(k+1)}}{r} - 2\right]^{\frac{1}{2}}.$$

Again, most statistical software provides packages that are able to do the complete sample estimation for the Birnbaum-Saunders distribution. Therefore, the M-step is indeed very easy to implement.

3.2 Gamma Lifetime Data

The PDF of the gamma distribution is expressed in terms of the gamma function

$$f(t;\mu,\lambda)=\frac{1}{\Gamma(\mu)}\frac{t^{\mu-1}}{\lambda^{\mu}}\exp\left\{-\frac{t}{\lambda}\right\},\qquad t>0,$$

where $\mu>0$ is a shape parameter and $\lambda>0$ is a scale parameter. The CDF is

$$F(t;\mu,\lambda)=\frac{1}{\Gamma(\mu)}\int_0^{t/\lambda}v^{\mu-1}e^{-v}dv=\frac{\gamma(\mu,t/\lambda)}{\Gamma(\mu)},\qquad t>0,$$

where $\gamma(a,b)$ is the lower incomplete gamma function given by

$$\gamma(a,b)=\int_0^b v^{a-1}e^{-v}dv.$$

Consider the progressively censored data $(\boldsymbol{T},\boldsymbol{Z})$ generated from the above gamma distribution. Given the parameter values $\Theta^{(k)}=\left(\mu^{(k)},\lambda^{(k)}\right)$ obtained from the kth SEM cycle, The missing datum Z_{jl} $(l=1,\dots,R_j)$ can be imputed from the following CDF

$$G_j\left(z;\mu^{(k)},\lambda^{(k)}\middle|T_{j:m:n}=t_j\right)=\frac{\gamma\left(\mu^{(k)},z/\lambda^{(k)}\right)-\gamma\left(\mu^{(k)},t_j/\lambda^{(k)}\right)}{\Gamma(\mu^{(k)})-\gamma\left(\mu^{(k)},t_j/\lambda^{(k)}\right)},\qquad t>t_j.$$

The pseudo *Q*-function based on data the imputed data is

$$Q(\mu,\lambda)=-n\ln\Gamma(\mu)-n\mu\ln\lambda+(\mu-1)\sum_{j=1}^{m}\ln t_j-\sum_{j=1}^{m}\frac{t_j}{\lambda}+(\mu-1)\sum_{j=1}^{m}\sum_{l=1}^{R_j}\ln z_{jl}$$
$$-\sum_{j=1}^{m}\sum_{l=1}^{R_j}\frac{z_{jl}}{\lambda}.$$

The standard procedure to derive MLEs from this complete sample gamma *Q*-function is to first get $\mu^{(k+1)}$ by solving

$$\sum_{j=1}^{m}\ln t_j+\sum_{j=1}^{m}\sum_{l=1}^{R_j}\ln z_{jl}+n\ln n+n\ln\mu^{(k+1)}-n\Psi\left(\mu^{(k+1)}\right)$$
$$=n\ln\left(\sum_{j=1}^{m}t_j+\sum_{j=1}^{m}\sum_{l=1}^{R_j}z_{jl}\right)$$

where $\Psi(\mu)=d\ln\Gamma(\mu)/d\mu$ is the digamma function, after which $\lambda^{(k+1)}$ can thus be obtained as

$$\lambda^{(k+1)} = \frac{\sum_{j=1}^{m} t_j + \sum_{j=1}^{m} \sum_{l=1}^{R_j} z_{jl}}{n\mu^{(k+1)}}.$$

Remark: The above SEM procedure is also applicable to some variants of the gamma variable. For example, the inverse gamma distribution as a lifetime model has been suggested by Glen[24]. An inverse gamma random variable can be converted to a gamma variable by taking the reciprocal. Therefore, we can first take the reciprocal of the progressively censored data from an inverse gamma distribution, and then obtain MLEs of the parameters using the above procedure. Similarly, progressively censored data from the log-gamma distribution can be analyzed using the above procedure after an exponential transformation.

3.3 Inverse-Gaussian Lifetime Data

The inverse Gaussian distribution as a lifetime model has been promoted by many researchers (e.g., Chhikara and Folks[23]) due to its physical interpretation, i.e., the first passage time distribution of a Wiener process with a drift. This distribution has numerous good properties. Its PDF accommodates a variety of shapes, from highly skewed to almost normal; its failure rate function is unimodal with a limiting value greater than 0, providing a suitable choice for a lifetime model. See Chhikara and Folks[23] for more details of this distribution. The respective PDF and CDF of an inverse Gaussian variable with the shape parameter μ and the mean λ are

$$f(t;\, \mu, \lambda) = \sqrt{\frac{\mu}{2\pi t^3}} \exp\left\{\frac{-\mu(t-\lambda)^2}{2\lambda^2 t}\right\}, \qquad t > 0;.$$

$$F(t;\, \mu, \lambda) = \Phi\left[\sqrt{\frac{\mu}{t}}\left(\frac{t}{\lambda} - 1\right)\right] + \exp\left(\frac{2\mu}{\lambda}\right) \Phi\left[-\sqrt{\frac{\mu}{t}}\left(\frac{t}{\lambda} + 1\right)\right], \qquad t > 0;$$

where $\Phi(\cdot)$ is the CDF of the standard normal distribution. Consider the progressively censored data $(\boldsymbol{T}, \boldsymbol{Z})$ generated from this distribution. If the missing data $\boldsymbol{Z}$ were known, MLE of the parameters based on $(\boldsymbol{T}, \boldsymbol{Z})$ can be specified as

$$\hat{\lambda} = \frac{\sum_{i=1}^{m} t_i + \sum_{i=1}^{m} \sum_{j=1}^{R_j} z_{ij}}{n}, \quad \hat{\mu}^{-1} = \frac{\sum_{i=1}^{m} (t_i^{-1} - \hat{\lambda}^{-1}) + \sum_{i=1}^{m} \sum_{j=1}^{R_j} (z_{ij}^{-1} - \hat{\lambda}^{-1})}{n}. \tag{5}$$

However, $\boldsymbol{Z}$ are indeed unknown. Under the SEM paradigm, they have to be imputed from the conditional distribution

$$G_j\left(z;\mu,\lambda\middle|T_{j:m:n}=t_j\right)=\frac{F(z;\ \mu,\lambda)-F\left(t_j;\ \mu,\lambda\right)}{1-F\left(t_j;\ \mu,\lambda\right)}\qquad t>t_j$$

in the S-Step, after which the M-Step can be readily updated in terms of (5).

3.4 Lognormal Lifetime Data

The lognormal distribution is another commonly used lifetime model. Its PDF is given by

$$f(t;\ \mu,\lambda)=\frac{1}{t\sqrt{2\pi\lambda^2}}\exp\left\{-\frac{(ln\,t-\mu)^2}{2\lambda^2}\right\} \tag{6}$$

In the presence of censoring data, the log-likelihood function does not have close form expression because of the survival function of the lognormal distribution involved (Bennett[22]). The EM algorithm for this distribution is thus preferable, and it has been developed by Ng et al.[7]. We will show that such data can also be easily handled by the SEM algorithm. Denoting $\Theta^{(k)}=\left(\mu^{(k)},\lambda^{(k)}\right)$ the value of Θ at the kth SEM cycle, the $(k+1)$st step evolves as follows.

S-Step

To impute the missing data, we need the conditional distribution of Z_{jl} $(l=1,\dots,R_j)$, which is given by

$$G_j\left(z;\mu,\lambda\middle|T_{j:m:n}=t_j\right)=\frac{\Phi\left(\frac{\ln z-\mu}{\lambda}\right)-\Phi\left(\frac{t_j-\mu}{\lambda}\right)}{1-\Phi\left(\frac{t_j-\mu}{\lambda}\right)}\qquad z>t_j.$$

A realization of the missing data $\boldsymbol{Z}$ can then be easily generated by imputing R_j *i.i.d.* samples from $G_j\left(z;\ \mu^{(k)},\lambda^{(k)}|T_{j:m:n}=t_j\right),\ j=1,\dots,m.$

M-Step

The realization of $\boldsymbol{Z}$ in conjunction with the observed data $\boldsymbol{T}$ makes up a complete sample for the lognormal distribution, which has close-form expressions for the MLEs. Therefore, the value of $\Theta^{(k+1)}$ are given by

$$\mu^{(k+1)} = \frac{\sum_{i=1}^{m} \ln t_i + \sum_{i=1}^{m} \sum_{j=1}^{R_j} \ln z_{ij}}{n},$$

$$\lambda^{(k+1)} = \sqrt{\frac{\sum_{i=1}^{m} [\ln t_i - \mu^{(k+1)}]^2 + \sum_{i=1}^{m} \sum_{j=1}^{R_j} [\ln z_{ij} - \mu^{(k+1)}]^2}{n}}.$$

3.5 Weibull Lifetime Data

The Weibull distribution is one of the most widely used models in reliability and survival analysis. The respective PDF and CDF of this two-parameter model are

$$f(t;\, \mu, \lambda) = \frac{\mu}{\lambda} \times \left(\frac{t}{\lambda}\right)^{\mu-1} \times \exp\left\{-\left(\frac{t}{\lambda}\right)^{\mu}\right\}, \qquad t > 0, \tag{7}$$

$$F(t;\, \mu, \lambda) = 1 - \exp\left\{-\left(\frac{t}{\lambda}\right)^{\mu}\right\}, \qquad t > 0,$$

where $\mu > 0$ is a shape parameter, and $\lambda > 0$ is a scale parameter. Ng et al.[7] has shown how to estimate the model parameters via the EM algorithm. Their method has to work on the log-lifetime that follows an extreme value distribution, and then transform the MLEs of the parameters in the extreme value distribution to the MLEs of the Weibull parameters. On the contrary, the SEM algorithm is able to avoid such transformation and work directly on the Weibull distribution, making the estimation more straightforward and simpler.

Consider the progressively censored data $(\boldsymbol{T}, \boldsymbol{Z})$ generated from the Weibull distribution (7). $\boldsymbol{T}$ are the observed data while $\boldsymbol{Z}$ are considered as missing data. Given the complete data $(\boldsymbol{T}, \boldsymbol{Z})$, The log-likelihood function can be expressed as

$$Q(\mu, \lambda) = n(\ln \mu - \ln \lambda) + (\mu - 1) \sum_{j=1}^{m} (\ln t_j - \ln \lambda) - \sum_{j=1}^{m} \left(\frac{t_j}{\lambda}\right)^{\mu}$$

$$+ (\mu - 1) \sum_{j=1}^{m} \sum_{l=1}^{R_j} (\ln z_{jl} - \ln \lambda) - \sum_{j=1}^{m} \sum_{l=1}^{R_j} \left(\frac{z_{jl}}{\lambda}\right)^{\mu}. \tag{8}$$

This complete sample likelihood function is easy to maximize. Therefore, the main idea of the SEM algorithm is to impute a realization of $\boldsymbol{Z}$ in the S-Step in order to facilitate the M-Step. Denoting $\Theta^{(k)} = (\mu^{(k)}, \lambda^{(k)})$ the value of Θ at the kth SEM cycle, the procedure of the $(k+1)$st step is detailed as follows.

The S-Step

To implement the S-step, the conditional CDF of Z_{jl} $(l = 1, \dots, R_j)$ given $T_{j:m:n} = t_j$ $(j = 1, \dots, m)$ is needed, which is readily obtained as

$$G_j\left(z;\ \mu, \lambda \middle| T_{j:m:n} = t_j\right) = 1 - \exp\left(\left(\frac{t_j}{\lambda}\right)^{\mu} - \left(\frac{z}{\lambda}\right)^{\mu}\right) \qquad z > t_j.$$

Based on this conditional CDF, we can readily impute $\boldsymbol{Z}_j$, $j = 1, \dots, m$, by generating R_j *i.i.d.* samples from $G_j\left(z;\ \mu^{(k)}, \lambda^{(k)} | T_{j:m:n} = t_j\right)$. The imputed $\boldsymbol{Z}_j$ makes up a random realization of $\boldsymbol{Z}$. This random realization of $\boldsymbol{Z}$ in conjunction with the observed data is substituted into (8) to get the pseudo *Q*-function for the M-Step.

The M-Step

After attaining the pseudo *Q*-function (8) by making use of the imputed $\boldsymbol{Z}$ from the S-Step, the M-Step aims to obtaining $\Theta^{(k+1)}$ by maximizing this *Q*-function. The standard procedure to maximize this complete sample Weibull likelihood function is to first derive $\mu^{(k+1)}$ by solving

$$\frac{\sum_{j=1}^{m} t_j{}^{\mu} \times \ln t_j + \sum_{j=1}^{m} \sum_{l=1}^{R_j} z_{jl}{}^{\mu} \times \ln z_{jl}}{\sum_{j=1}^{m} t_j{}^{\mu} + \sum_{j=1}^{m} \sum_{l=1}^{R_j} z_{jl}{}^{\mu}} - \frac{1}{\mu} - \frac{1}{n}\left(\sum_{j=1}^{m} \ln t_j + \sum_{j=1}^{m} \sum_{l=1}^{R_j} \ln z_{jl}\right) = 0,$$

and then obtain $\lambda^{(k+1)}$ by making use of $\mu^{(k+1)}$ as

$$\lambda^{(k+1)} = \left[\frac{\sum_{j=1}^{m} t_j^{\mu^{(k+1)}} + \sum_{j=1}^{m} \sum_{l=1}^{R_j} z_{jl}^{\mu^{(k+1)}}}{n}\right]^{1/\mu^{(k+1)}}.$$

Most statistical software is able to do the complete sample estimation for the Weibull distribution. Uses made of these standard software packages further simplify the M-Step, and thus facilitating the application of the SEM algorithm. After an initial burn-in period, the sequence of $\left\{\Theta^{(k)}\right\}$ is averaged to get an approximation of the MLEs.

4 Two Illustrative Examples

4.1 The Birnbaum-Saunders Distribution for the Fatigue Life Data

Consider a dataset from Birnbaum and Saunders[19] on the fatigue life of 6061-T6 aluminum coupons cut parallel to the direction of rolling and oscillated at 18 cycles per second, with a maximum stress per cycle at 31,000 psi. The original data were presented in Table 2 in Birnbaum and Saunders[19]. MLE of the Birnbaum-Saunders parameters based on this complete sample is

$$\hat{\mu} = 0.170,\ \hat{\lambda} = 131.8.$$

Ng et al.[20] have used this dataset to demonstrate their inference procedure under type-II censoring. In this study, we use this dataset to demonstrate the applicability of the SEM algorithm for progressively censored data from the Birnbaum-Saunders distribution. Based on the original data, we randomly generate a progressively censored dataset with $m = 50$ and a progressively censored scheme $R_i = 1$ for $i < 50$ and $R_{50} = 2$. The generated data are displayed in Table 1.

We apply the SEM procedure developed in Section 3.1 to analyze this progressively censored dataset. The initial value of μ and λ for the algorithm are set as $\mu^{(0)} = 1$ and $\lambda^{(0)} = 100$, which is far away from the MLE. The number SEM cycles is set as 1100. The first 100 cycles are used as burn-in period, while the additional 1000 cycles are averaged to estimate model parameters.

Figure 1 shows the trace plots of these two parameters versus the SEM cycles. The values of the parameters oscillate with the SEM cycles around the bold horizontal lines in Figure 1, but do not show any uptrend or downtrend. This suggests that the Markov Chain $\{\Theta^{(k)}\}$ has converged to a stationary distribution. The average of the sequence $\{\Theta^{(k)}\}$ would be sufficient to approximate the MLE, which is

$$\hat{\mu} = 0.1735, \hat{\lambda} = 131.2.$$

To demonstrate the validity of the SEM algorithm, we use these estimated values from the SEM algorithm as a starting point and directly maximize the likelihood function using some

derivative-free algorithms reviewed in Kolda et al.[8]. The brute-force maximization yields an estimate of

$$\hat{\mu} = 0.1739, \hat{\lambda} = 131.2,$$

which is very close to the results given by the SEM algorithm. The difference may be due to some simulation variations. This similarity implies the efficacy of the SEM algorithm in dealing with the censored data.

Given the complete sample with sample size $n = 101$, an extensive simulation study is conducted to investigate the impact of the sampling scheme, i.e. the failure number m and the combination of withdrawn numbers $\{R_1, R_2, \dots, R_m\}$, on the performance of the SEM algorithm. The progressively Type-II censored dataset is generated randomly, that is, the withdrawn number R_i at the ith failure is randomly simulated under the constraint $\sum_{i=1}^{m} R_i = n - m$. For each value of m, we randomly generate a progressively Type-II censored dataset and, using the SEM algorithm, we obtain an estimate vector of (μ, λ), denoted as $(\hat{\mu}, \hat{\lambda})$. The Kolmogorov-Smirnov test (K-S test) is served as a goodness-of-fit test (compare the fitted distribution to the original complete data). By ranging m from 21 to 50, the results are given in Table 2. In Table 2, the notation p represents the p-value of the Kolmogorov-Smirnov test.

Form the p-value we can see that the increase of the failure number m does not necessarily guarantee the improvement in the performance of the SEM algorithm. This result may seem counterintuitive in the sense that one would typically expect, with the failure number increasing, the p-value of the K-S test should increase. However, the fluctuating p-values can be explained as follows: for predetermined sample size n and failure number m, the combination $\{R_1, R_2, \dots, R_m\}$ has a considerable influence on the SEM algorithm. We repeat the preceding procedure for 20 times and plot the p values in Figure 2. The increasing trend in Figure 2 indicates that, with the failure number increasing, the SEM algorithm will more fit the data.

4.2 The Lognormal Distribution for the HDD Failure Data

In this section we apply the SEM algorithm to a progressively Type-I censored dataset, i.e. the HDD failure data, to show the competence of the proposed technique. Analysis of HDD

failures reveals that most failures attribute to particles accumulated in the disks. Therefore, the life test was conducted by injecting particles into 36 raw disks for a certain duration, during which the cumulative particle counts to failure for the failed disks were recorded. After the test, some units were still working, and their cumulative particle counts to failure data were censored. Due to the unstable injection rate for each unit, the censored cumulative particle counts for these units were different. Therefore, the data can be regarded as progressively censored. The data are shown in Table 3.

HDD engineers demand a convenient method that can effectively analyze the data. The SEM algorithm is thus favorable. We try all the distributions presented in Section 3 on this data set, and the results are listed in Table 4. The MLEs, which are obtained based on the exactly observed failure data, are set as the initial values of μ and λ in the SEM algorithm. As with the previous example, the number of SEM cycles is set as 1100, while the first 100 cycles are used for burn-in. Obviously, based on the value of the log-likelihood, the lognormal distribution is well-fit to the data. For each distribution, we also record the time the SEM algorithm consumes. All computations were coded in MATLAB (MathWorks, R2011b) on an Intel Core 2 6420 (Intel), 2.13 GHz PC with 2 GB RAM. As can be seen, the Birnbaum-Saunders distribution and the Inverse-Gaussion distribution take much more time than the other three distributions. This is rooted in the way we impute the missing data: To generate a sample from $F_j(Z_{jl}\,;\,\Theta)$ in (2), we generate samples from $F(t;\,\Theta)$ until we get the first one greater than t_j; see the appendix. If we use built-in random number generator, as in the other three cases, the algorithm is indeed very fast.

Therefore, the Lognormal distribution can be served as the underlying distribution, and the estimates for μ and λ are, respectively $\hat{\mu} = 12.9327$ and $\hat{\lambda} = 1.1091$. The evolution paths of the parameters in the algorithm are depicted in Figure 3. As can be seen from this figure, no obvious trend is detected for the $\{\Theta^{(k)}\}$ sequence. By using the estimates from the SEM algorithm as a starting point, we numerically maximize the likelihood function by some optimization algorithms. The MLEs are

$$\hat{\mu} = 12.920, \hat{\lambda} = 1.084,$$

which are very close to the results from the SEM algorithm. This implies the effectiveness of the SEM algorithm.

5 Conclusions

This study has developed a generic framework for analyzing the progressively censored data by using the stochastic EM algorithm. The algorithm iteratively implements the S-Step by drawing a sample from the conditional distribution of the missing data based on the parameter values from the previous step, while the M-Step is a complete sample likelihood maximization. Both steps can be easily implemented by making use of distribution packages provided by most statistical software. In addition, the stochastic nature enables the SEM algorithm to avoid getting stuck at a saddle point of the likelihood function, which is a headache faced by the traditional EM algorithm, and the Newton–Raphson method. In view of the fact that progressively censored data are common in real applications and that engineers prefer handy tools for the analysis, our framework is potentially very useful. We then presented a real dataset from a life test on 36 HDDs. The data were progressively censored, and the SEM algorithm was shown to be capable to effectively estimate the parameters with a short runtime.

In Section 4 we investigate the impact of the sampling scheme on the performance of the SEM algorithm. The results show that, with the failure number m increasing, the p-value of the K-S test increases, i.e., the performance improves. However, for fixed sample size n and failure number m, the combination of withdrawn numbers $\{R_1, R_2, \dots, R_m\}$ is of great influence on the SEM algorithm. Future research can be done in determining the optimal sampling scheme. Also, future research can be done in studying whether the number of burn-in iterations depends on specific distributions and/or on initial parameter settings. This will help the practitioners to choose a favorable sampling scheme and appropriate number of cycles in their domain problems.

Appendix

This appendix shows how to draw a random sample from Equation (2).

The standard method is to use the fact that $F_j(Z_{jl}\,;\,\Theta)$ follows a standard uniform distribution, i.e., $\mathcal{U} = F_j(Z_{jl}\,;\,\Theta) \sim U(0,1)$. By using (2), we can see that

$$\mathcal{U}[1 - F(t_j;\,\Theta)] + F(t_j;\,\Theta) = F(Z_{jl}\,;\,\Theta)$$

Therefore, we can first generate a random realization of $\mathcal{U}$, say, u, and then obtain realization of Z_{jl} as

$$z_{jl} = F^{-1}(u + (1-u)F(t_j;\,\Theta)\,;\,\Theta), \tag{9}$$

where $F^{-1}(\cdot)$ is the inverse function of $F(\cdot)$. For example if $F(\cdot)$ if the Weibull CDF, (9) can be explicitly written as

$$z_{jl} = \lambda[(t_j/\lambda)^{\mu} - \ln(1-u)]^{1/\mu},$$

However, close-form expressions of (9) for some distributions do not exist, e.g., the lognormal distribution and the gamma distribution. Luckily, most statistical software provides packages that can directly compute $F^{-1}(\cdot)$ for most distributions. This greatly simplifies the problem.

Another way to generate a sample from $F_j(Z_{jl}\,;\,\Theta)$ in (2) is to generate samples from $F(t;\,\Theta)$ until we get the first one greater than t_j. This method is somewhat brute-force. However, it is much easier to implement. In addition, almost all statistical software has random number generators for most common distributions, e.g., the ones discussed in Section 3.

References

1. Wang FK, Cheng YF. EM algorithm for estimating the Burr XII parameters with multiple censored data. *Quality and Reliability Engineering International* 2010; **26:** 615-630.
2. Pareek B, Kundu D, Kumar S. On progressively censored competing risks data for Weibull distributions. *Computational Statistics & Data Analysis* 2009; **53:** 4083-4094.

3. Balakrishnan N, Kannan N, Lin CT, Ng HKT. Point and interval estimation for Gaussian distribution, based on progressively Type-II censored samples. *IEEE Transactions on Reliability* 2003; **52:** 90-95.
4. Lee J, Pan R. Bayesian analysis of step-stress accelerated life test with exponential distribution. *Quality and Reliability Engineering International* 2012; **28:** 353-361.
5. Balakrishnan N, Saleh HM. Relations for moments of progressively type-II censored order statistics from log-logistic distribution with applications to inference. *Communications in Statistics-Theory and Methods* 2012; **41:** 880-906.
6. Davis HT, Feldstein ML. The generalized Pareto law as a model for progressively censored survival data. *Biometrika* 1979; **66:** 299-306.
7. Ng HKT, Chan PS, Balakrishnan N. Estimation of parameters from progressively censored data using EM algorithm. *Computational Statistics & Data Analysis* 2002; **39:** 371-386.
8. Kolda TG, Lewis RM, Torczon V. Optimization by direct search: New perspectives on some classical and modern methods. *SIAM Review* 2003; **45:** 385-482.
9. Lin CT, Wu SJS, Balakrishnan N. Inference for log-gamma distribution based on progressively type-II censored data. *Communications in Statistics-Theory and Methods* 2006; **35:** 1271-1292.
10. Pradhan B, Kundu D. On progressively censored generalized exponential distribution. *Test* 2009; **18:** 497-515.
11. Wei GCG, Tanner MA. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association* 1990; **85:** 699-704.
12. Celeux G, Diebolt J. The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational Statistics Quarterly* 1985; **2:** 73-82.
13. Tregouet DA, Escolano S, Tiret L, Mallet A, Golmard JL. A new algorithm for haplotype-based association analysis: the Stochastic-EM algorithm. *Annals of human genetics* 2004; **68:** 165-177.
14. Delignon Y, Marzouki A, Pieczynski W. Estimation of generalized mixtures and its application in image segmentation. *IEEE Transactions on Image Processing* 1997; **6:** 1364-1375.

15. Cariou C, Chehdi K. Unsupervised texture segmentation/classification using 2-D autoregressive modeling and the stochastic expectation-maximization algorithm. *Pattern Recognition Letters* 2008; **29:** 905-917.
16. Nielsen FS. The stochastic EM algorithm: estimation and asymptotic results. *Bernoulli* 2000; **6:** 457-489.
17. Diebolt J, Celeux G. Asymptotic properties of a stochastic EM algorithm for estimating mixing proportions. *Stochastic Models* 1993; **9:** 599-613.
18. Svensson I, Sjöstedt-de Luna S. Asymptotic properties of a stochastic EM algorithm for mixtures with censored data. *Journal of Statistical Planning and Inference* 2010; **140:** 111-127.
19. Birnbaum ZW, Saunders SC. Estimation for a family of life distributions with applications to fatigue. *Journal of Applied Probability* 1969; **6:** 328-347.
20. Ng HKT, Kundu D, Balakrishnan N. Point and interval estimation for the two-parameter Birnbaum–Saunders distribution based on Type-II censored samples. *Computational Statistics & Data Analysis* 2006; **50:** 3222-3242.
21. Diebolt J, Ip EHS. Stochastic EM: Method and application, in *Markov chain Monte Carlo in practice*, Chapman & Hall/CRC, 1996; 259-273.
22. Bennett S. Log-logistic regression models for survival data. *Applied Statistics* 1983; **32:** 165-171.
23. Chhikara RS, Folks JL. The inverse Gaussian distribution as a lifetime model. *Technometrics* 1977; **19:** 461-468.
24. Glen AG. On the inverse gamma as a survival distribution. *Journal of Quality Technology* 2011; **43:** 158-166.

Table 1. A progressively censored dataset generated from the original fatigue life data presented in Birnbaum and Saunders[19] and Ng et al.[20].

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
$t_{i:m:n}$	70	90	96	97	99	100	103	104	104	105	107	108	108	108	109	109	112
R_i	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
i	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
$t_{i:m:n}$	112	113	114	114	116	119	120	120	121	124	124	124	128	128	129	139	130
R_i	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
i	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	
$t_{i:m:n}$	130	131	131	132	132	132	134	134	136	136	138	141	142	144	146	148	
R_i	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	

Table 2. Estimated parameter values and p-value with changing m.

m	21	22	23	24	25	26	27	28	29	30
$\hat{\mu}$	0.1936	0.1921	0.1911	0.1869	0.1918	0.1871	0.1873	0.1864	0.1882	0.1827
$\hat{\lambda}$	133.98	133.52	134.48	133.52	133.25	133.57	132.21	133.05	131.99	132.68
p	0.3044	0.4027	0.2394	0.4537	0.4646	0.4389	0.4377	0.5726	0.3815	0.5835
m	31	32	33	34	35	36	37	38	39	40
$\hat{\mu}$	0.1973	0.1921	0.1861	0.1761	0.1931	0.1939	0.1794	0.1765	0.1824	0.1801
$\hat{\lambda}$	134.61	133.85	134.57	132.01	134.41	134.23	132.04	131.75	132.31	131.30
p	0.1911	0.3392	0.2558	0.4496	0.2386	0.2612	0.4385	0.3850	0.4885	0.2763
m	41	42	43	44	45	46	47	48	49	50
$\hat{\mu}$	0.1747	0.1887	0.1873	0.1728	0.1816	0.1785	0.1856	0.1874	0.1785	0.1605
$\hat{\lambda}$	131.61	134.45	133.15	132.12	132.81	131.83	133.46	134.66	132.21	130.66
p	0.3628	0.2583	0.5363	0.4994	0.6279	0.3946	0.4821	0.2334	0.4865	0.2241

Table 3. Progressively censored data of 36 HDDs based on cumulative particle counts: Censoring is indicated by +.

38248	44592	63184	70888	168536	170408	186944	193084	217956
217992	242064+	242064+	242064+	253744	266240	272468	290220+	290220+
292380	295780	299797+	301564+	301920+	304563+	307636+	313936+	316224
345568	517636	797168+	797196	812716+	822136	905952	1341996	1345544+

Table 4. Estimated parameter values and the likelihood result.

	Birnbaum-Saunders	Gamma	Inverse-Gaussion	Lognormal	Weubull
$\hat{\mu}$	1.2210	1.3716	3.2278e+5	12.9327	1.2051
$\hat{\lambda}$	3.9194e+5	4.3325e+5	8.6029e+5	1.1091	6.1591e+5
log-likelihood	-315.1581	-315.1807	-315.8342	-314.8521	-315.3574
time elapsed	62.649984	1.838062	5.878271	0.885176	1.409312