2002年日本オペレーションズ・リサーチ学会
秋季研究発表会

Parameter Estimation of Additive NHPP-Based Software Reliability Models via EM Algorithm

H. Okamura† (01013754), Y. Watanabe†, T. Dohi† (01307065)
† Department of Information Engineering, Graduate School of Engineering, Hiroshima University

1. Introduction

Software reliability models (SRMs) based on the non-homogeneous Poisson process (NHPP) have been proposed in recent three decades. The NHPP-based SRMs are usually tractable in practical use and are intuitively reasonable. The most traditional SRMs are Goel and Okumoto model [1] and Yamada and Osaki model [2].

This paper describes a parameter estimation algorithm on the NHPP-based SRMs. Especially, we focus on the expectation-maximization (EM) principle, and develop the iteration algorithm to calculate the maximum likelihood estimates on the NHPP-based SRMs.

First, we introduce the NHPP-based SRMs and the underlying mathematical modeling framework. Based on the modeling framework, the EM algorithm for calculating the maximum likelihood estimates can be developed as an iterative scheme. Next, we extensively apply the EM algorithm to estimating the model parameters for additive NHPP-based SRMs.

2. NHPP-Based SRMs

The NHPP-based SRMs are usually tractable in practical use and are intuitively reasonable. In this section, we introduce the mathematical framework of the NHPP-based SRMs, and develop the EM algorithm for estimating the model parameters.

As the modeling frameworks of NHPP-based SRMs, some approaches are known to construct them from the stochastic behavior of software faults. Langberg and Singpurwalla [3] propose a modeling framework based on the generalized order statistics (GOSs). The GOS-based modeling framework is made under the following assumptions:

Assumption A: Software failures caused by software faults occur at independent and identically distributed (i.i.d.) random times.

Assumption B: The initial number of software faults is finite.

Let $F(t)$ and $f(t) = dF(t)/dt$ denote the probability distribution function for the software fault detection times and its probability density function, respectively. If the initial number of software faults is known as a constant N (> 0), the probability mass function of the number of faults detected before time t is given by

$$\Pr\{N(t) = n\} = \binom{N}{n} F(t)^n \overline{F}(t)^{N-n}, \tag{1}$$

where $\overline{F}(\cdot) = 1 - F(\cdot)$. Assuming that the initial number of faults follows the Poisson distribution with parameter ω (> 0), we have the number of faults detected before time t,

$$\Pr\{N(t) = k\} = \frac{\{\omega F(t)\}^k}{k!} \exp\{-\omega F(t)\}. \tag{2}$$

Equation (2) is equivalent to the probability mass function of the NHPP having the mean value function $\omega F(t)$. In this modeling framework, substituting typical probability distributions into $F(t)$ in Eq. (2) yields existing SRMs.

Let us consider the parameter estimation for the NHPP-based SRMs. In particular, this paper focuses on the maximum likelihood estimation.

Let $X_1, X_2, \ldots, X_N$ and $X_{[1]} < X_{[2]} < \cdots < X_{[N]}$ be fault detection times and their order statistics, respectively, where N is the initial number of faults and is the Poisson distributed random variable with parameter ω (> 0). If one can observe all the fault detection times $\mathcal{D}_\infty = (x_1, \ldots, x_n)$, which is the complete data, the logarithmic likelihood function is given by

$$\log L(\omega, \boldsymbol{\theta} | \mathcal{D}_\infty) = n \log \omega - \omega + \sum_{k=1}^{n} \log f(x_k; \boldsymbol{\theta}). \tag{3}$$

Under the complete data, the maximum likelihood estimates are provided as follows.

$$\hat{\omega} = n \tag{4}$$

and

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left\{ \sum_{k=1}^{n} \log f(x_k; \boldsymbol{\theta}) \right\}. \tag{5}$$

The EM algorithm is an iterative method for an estimation problem with incomplete data. Given the observed experiment, each step in the EM algorithm consists of calculating the expected logarithmic likelihood function under the incomplete data and of finding the estimates which maximizing it.

Given the fault detection time data at time t, $\mathcal{D}_t = (x_1, \ldots, x_n)$, $n < N$, we can see that $\mathcal{D}_t$ is the incomplete

data instead of the complete data $\mathcal{D}_\infty$. The EM algorithm is then developed as below [4]:

At the $(n+1)$-st step, the estimates of the model parameters are calculated as

$$\hat{\omega}^{(n+1)} = \mathrm{E}[N|\mathcal{D}_t; \hat{\omega}^{(n)}, \hat{\boldsymbol{\theta}}^{(n)}] \tag{6}$$

$$\hat{\boldsymbol{\theta}}^{(n+1)} = \underset{\boldsymbol{\theta}}{\mathrm{argmax}} \left\{ \mathrm{E}\left[\sum_{k=1}^{N} \log f(X_k;\boldsymbol{\theta}) \middle| \mathcal{D}_t; \hat{\omega}^{(n)}, \hat{\boldsymbol{\theta}}^{(n)} \right] \right\}, \tag{7}$$

where $\boldsymbol{\theta}^{(n)}$ is the estimated parameter set at the n-th step in the EM algorithm and $\mathrm{E}[\cdot\,;\boldsymbol{\theta}]$ denotes the mathematical expectation operator, provided that the probability density f has the parameter set $\boldsymbol{\theta}$.

In Eqs. (6) and (7), we use the following formula on the expectation operator: For any measurable function h,

$$\begin{aligned} &\mathrm{E}\left[\sum_{k=1}^{N} h(X_k) \middle| \mathcal{D}_t; \hat{\omega}^{(n)}, \hat{\boldsymbol{\theta}}^{(n)} \right] \\ &\quad = \sum_{k=1}^{n} h(x_k) + \hat{\omega}^{(n)} \int_t^\infty h(u) f(u; \hat{\boldsymbol{\theta}}^{(n)}) du. \end{aligned} \tag{8}$$

3. Additive NHPP-Based SRMs

The additive NHPP-based SRM consists of m software components. The m components are independent from each other. The software faults on the component i are detected according to the NHPP with the mean value function, $\Lambda_i(t;\boldsymbol{\theta}_i)$, $i = 1,\ldots,m$. Since each component is independent, the total number of detected faults is given by

$$\Pr\{N(t)=n\} = \frac{\sum_{i=1}^m \Lambda_i(t;\boldsymbol{\theta}_i)^n}{n!} \exp\left\{ -\sum_{i=1}^m \Lambda_i(t;\boldsymbol{\theta}_i) \right\}. \tag{9}$$

From Eq. (9), we can find that the additive NHPP-based SRMs are comprised by the NHPP-based SRMs. In particular, if the fault detection on each component can be described by the GOS-based modeling framework, the probability mass function for the total number of faults is given by

$$\begin{aligned} \Pr\{N(t)=n\} &= \frac{\left(\omega \sum_{i=1}^m p_i F_i(t;\boldsymbol{\theta}_i)\right)^n}{n!} \\ &\quad \times \exp\left\{ -\left(\omega \sum_{i=1}^m p_i F_i(t;\boldsymbol{\theta}_i) \right) \right\}, \end{aligned} \tag{10}$$

where p_i is the ratio of the software faults corresponding to the component i and $F_i(\cdot)$ denotes the fault detection probability for an inherent software fault in the component i.

Let us consider the parameter estimation problem. We first define two data sets, fault detection time data and their corresponding component data,

$$\mathcal{D}_t = (x_1,\ldots,x_n), \quad \mathcal{M}_t = (m_1,\ldots,m_n). \tag{11}$$

Then the logarithmic likelihood function is given by

$$\begin{aligned} &\log L(\boldsymbol{p},\boldsymbol{\theta}|\mathcal{D}_t,\mathcal{M}_t) \\ &\quad = \sum_{i=1}^m n_i \log p_i + \sum_{k=1}^n \log\left(\sum_{i=1}^m f_i(x_k;\boldsymbol{\theta}_i)\chi\{m_k=i\} \right), \end{aligned} \tag{12}$$

where $\boldsymbol{p} = (p_1,\ldots,p_m)$, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_m)$, $\chi\{\cdot\}$ is the indicator function and $n_i = \sum_{k=1}^n \chi\{m_k = i\}$. Hence, we have the following maximum likelihood estimates: For $i = 1,\ldots,m$,

$$\hat{p}_i = n_i/n, \tag{13}$$

$$\hat{\boldsymbol{\theta}}_i = \underset{\boldsymbol{\theta}_i}{\mathrm{argmax}} \left\{ \sum_{k=1}^n \log f_i(x_k;\boldsymbol{\theta}_i)\chi\{m_k=i\} \right\}. \tag{14}$$

In fact, we may not observe the component data $\mathcal{M}_t$. Thus, the EM algorithm can be applied to calculating the maximum likelihood estimates as follows:

E-step:

$$\begin{aligned} &\mathrm{E}\left[\sum_{k=1}^n h(X_k)\chi\{M_k=z\} \middle| \mathcal{D}_t; \hat{\boldsymbol{p}}^{(n)}, \hat{\boldsymbol{\theta}}^{(n)} \right] \\ &\quad = \sum_{k=1}^n \frac{\hat{p}_z^{(n)} h(x_k) f_z(x_k;\hat{\boldsymbol{\theta}}_z^{(n)})}{\sum_{i=1}^m \hat{p}_i^{(n)} f_i(x_k;\hat{\boldsymbol{\theta}}_i^{(n)})} \end{aligned} \tag{15}$$

M-step:

$$\hat{p}_i^{(n+1)} = \frac{\mathrm{E}\left[\sum_{k=1}^n \chi\{M_k=i\} \middle| \mathcal{D}_t; \hat{\boldsymbol{p}}^{(n)}, \hat{\boldsymbol{\theta}}^{(n)} \right]}{n}, \tag{16}$$

$$\begin{aligned} \hat{\boldsymbol{\theta}}_i^{(n+1)} &= \underset{\boldsymbol{\theta}_i}{\mathrm{argmax}} \left\{ \mathrm{E}\left[\sum_{k=1}^n \log f_i(X_k;\boldsymbol{\theta}_i)\chi\{M_k=i\} \right. \right. \\ &\qquad \left. \left. \middle| \mathcal{D}_t; \hat{\boldsymbol{p}}^{(n)}, \hat{\boldsymbol{\theta}}^{(n)} \right] \right\}. \end{aligned} \tag{17}$$

Combining Eqs. (15)–(17) with Eqs. (6)–(8), we develop the EM algorithm for additive NHPP-based SRMs.

References

[1] A. Goel and K. Okumoto, Time-dependent error-detection rate model for software reliability and other performance measures, IEEE Trans. Reliab., R-28, 206–211, 1979.

[2] S. Yamada and S. Osaki, Software reliability growth modeling: models and applications, IEEE Trans. Software Eng., SE-11, 1431–1437, 1985.

[3] N. Langberg and N. D. Singpurwalla, Unification of some software reliability models, SIAM J. Sci. Comput., 6, 781–790, 1985.

[4] H. Okamura, Y. Watanabe and T. Dohi, An estimation of software reliability models based on EM algorithm (in Japanese), Trans. of IEICE(A), J85-A, 442–450, 2002.