변곡 S-형 소프트웨어 신뢰도성장모형의 베이지안 모수추정

김희수* · 이종형** · 박동호***†

* 한국관세무역개발원
** 건양대학교 병원관리학과
*** 한림대학교 수리정보과학부

Bayesian Estimation for Inflection S-shaped Software Reliability Growth Model

Hee Soo Kim* · Chong Hyung Lee** · Dong Ho Park***†

* Customs and Trade Laboratory, Korea Customs and Trade Development Institute
** Department of Hospital Management, Konyang University
*** Department of Information and Statistics, Hallym University

Key Words : SRGM, failure detection rate, inflection rate, Metropolis-Hastings algorithm, prior and posterior distribution

Abstract

The inflection S-shaped software reliability growth model (SRGM) proposed by Ohba(1984) is one of the most commonly used models and has been discussed by many authors. The main purpose of this paper is to estimate the parameters of Ohba's SRGM within the Bayesian framework by applying the Markov chain Monte Carlo techniques. While the maximum likelihood estimates for these parameters are well known, the Bayesian method for the inflection S-shaped SRGM have not been discussed in the literature. The proposed methods can be quite flexible depending on the choice of prior distributions for the parameters of interests. We also compare the Bayesian methods with the maximum likelihood method numerically based on the real data.

1. Introduction

The software reliability growth model is one of the fundamental tools for assessing and predicting the reliability of the software system and thus, describing the fault debugging process as well.

Among those models discussed in the literature, the exponential reliability growth model suggested by Goel and Okumoto (1979) has been widely used mainly due to its simplicity.

The inflection S-shaped software reliability gro wth model (SRGM) has been proposed by Ohba (1984) and can be applied in many practical situations. The inflection S-shaped SRGM introduces the conc ept of inflection rate, assuming that some of the faults are dependent among them. Kim, Yamada and Park(2005) recently propose a new software reliability growth model which is the mixture of two exponential reliability growth models. The inflection S-shaped SRGM is character -ized by three parameters representing the failure detection rate, number of faults and inflection rate and they need

† Corresponding author :dhpark@hallym.ac.kr

to be estimated from the failure data obtained by executing the software system. Ohba(1984) assumes that the inflection rate is known and estimates the failure detection rate and the number of faults by applying the maximum likelihood techniques. Kuo, Lee, Choi and Yang (1997) propose two Gibbs sampling approaches to compute the Bayes estimates for the mean number of faults remaining and the current software reliability.

The main objective of this paper is to adapt the Bayesian approach to estimate three parameters of the inflection S-shaped SRGM. Based on the joint posterior distribution of these parameters for the Ohba's model, we may use the Gibbs sampling which is a special case of Markov chain Monte Carlo.

This paper consists of six sections. Section 2 int roduces the inflection S-shaped SRGM suggested by Ohba (1984). Section 3 presents the maximum likelihood method and the Bayesian approach to estimate the unknown parameters of the inflection S-shaped SRGM. In Section 4, we assume certain prior distributions for the unknown parameters of our interests and derive its corresponding conditional marginal posterior distribution for each parameter, which is necessary to apply the Bayesian estimation method. Section 5 uses real data to com pute the maximum likelihood estimates and the Bayesian estimates by applying the procedures discussed in Section 3. Section 6 gives concluding remarks.

2. Inflection S-shaped SRGM

Let $m(t)$ denote the mean value function of an NHPP (Nonhomogeneous Poisson process) describ ing the software fault detection process. Then, Ohba's(1984) inflection S-shaped SRGM is defined as

$$m(t) = N\left(\frac{1-e^{-\phi t}}{1+((1-\gamma)/\gamma)e^{-\phi t}}\right) \qquad (2.1)$$

under the following assumptions.

I) Some of the faults contained in a program are mutually dependent.

ii) The probability of failure detection at a given time is proportional to the current number of detectable faults remaining in a program.

iii) The proportionality is constant.

iv) The isolated faults can be entirely removed.

Here, $m(t)$ represent the expected cumulative number of software faults detected up to the testing time t starting from $t=0$, while N and ϕ denote the number of faults contained in the software system initially and the failure detection rate, respectively and $\gamma(0<\gamma<1))$ is the inflection rate which is the ratio of the number of detectable faults to the total number of faults latent in the system. When $\gamma=1$, then the model is reduced to the exponential software reliability growth model discussed by Goel and Okumoto (1979).

As for the interpretation of N, Nayak, Bose and Kundu(2008) asserts that N needs to be interpreted as "expected number of failures observed over -infinite testing time", instead of "initial number of faults in the system".

The intensity function for the inflection S-shaped SRGM is obtained as

$$\lambda(t) = \frac{dm(t)}{dt} = \frac{N\phi(1+((1-\gamma)/\gamma))e^{-\phi t}}{(1+((1-\gamma)/\gamma)\,e^{-\phi t})^2} \qquad (2.2)$$

Kimura, Toyota and Yamada(1999) derive the optimal software release time based on the special case of the intensity function given in (2.2) when $\gamma=1$.

3. Estimation of parameters

Assume that we have observed n pairs of data $(z_k,t_k), k=1,\cdots,n$, where z_k denotes the number of failures observed up to time $t_k$with $z_{k+1} \geq z_k$ for $t_{k+1} > t_k$.

Let $M(t)$ represent the cumulative number of fa ults detected until time t. If the number of failures is assumed to follow a Poisson distribution, the joint probability that the pairs of data$(z_k,t_k), k=1,\cdots,n$, are observed can be expressed as

$$\Pr\{M(0)=0, M(t_1)=z_1,\cdots,M(t_n)=z_n\}=$$

$$\prod_{i=1}^{n} \frac{m(t_i)-m(t_{i-1})^{(z_i-z_{i-1})}}{(z_i-z_{i-1})!} \times \exp\{-(m(t_i)-m(t_{i-1}))\} \quad (3.1)$$

Denote the observed n pairs of data by $D_t=\{(z_0,t_0),(z_1,t_1),\cdots,(z_n,t_n)\}$ Then, its likelihood function for the inflection S-shaped SRGM can be written as

$$L(N,\phi,\gamma \mid D_t) = N^{z_n}\prod_{i=1}^{n}\frac{\left\{\frac{1-e^{-\phi t_i}}{1+((1-\gamma)/\gamma)e^{-\phi t_i}}-\frac{1-e^{-\phi t_{i-1}}}{1+((1-\gamma)/\gamma)e^{-\phi t_{i-1}}}\right\}^{(z_i-z_{i-1})}}{(z_i-z_{i-1})!} \times \exp\left\{\frac{N(1-e^{-\phi_n})}{1+((1-\gamma)/\gamma)e^{-\phi_n}}\right\}, \quad (3.2)$$

where $(z_0,t_0)\equiv(0,0)$.

3.1 Maximum likelihood estimation

Taking the logarithm on both sides of the expression (3.2), we obtain the log likelihood function for three parameters N,ϕ and γ as

$$\ln L = \ln L(N,\phi,\gamma \mid D_t) = z_n \ln N + \sum_{i=1}^{n}(z_i-z_{i-1}) \times \ln\left\{\frac{1-e^{-\phi t_i}}{1+((1-\gamma)/\gamma)e^{-\phi t_i}}-\frac{1-e^{-\phi t_{i-1}}}{1+((1-\gamma)/\gamma)e^{-\phi t_{i-1}}}\right\} - \sum_{i=1}^{n}\ln(z_i-z_{i-1})! - \frac{N(1-e^{-\phi t_n})}{1+((1-\gamma)/\gamma)e^{-\phi t_n}} \quad (3.3)$$

By taking the partial derivatives of $\ln L$ with respect to each of three parameters N,ϕ and γ and setting them equal to zero, we obtain the following log-likelihood equations.

$$\frac{\partial \ln L}{\partial N} = \frac{\partial \ln L}{\partial \phi} = \frac{\partial \ln L}{\partial \gamma} = 0 \quad (3.4)$$

The maximum likelihood estimates for N,ϕ and γ can be calculated by solving these equations, given in (3.4), simultaneously. Although the explicit solutions for N,ϕ and γ are quite difficult to obtain analytically, they may be solved numerically.

3.2 Bayes estimation

In this subsection, we present the Bayes methodol ogy to estimate the unknown parameters N,ϕ and γ of the inflection S-shaped SRGM based on D_t.

Given the data D_t, the joint posterior distribution of N,ϕ and γ can be expressed as

$$\pi(N,\phi,\gamma|D_t) = \frac{f(N,\phi,\gamma)f(D_t|N,\phi,\gamma)}{f(D_t)} = \frac{f(N,\phi,\gamma)f(D_t|N,\phi,\gamma)}{\iiint f(N,\phi,\gamma)f(D_t|N,\phi,\gamma)dNd\phi d\gamma} \propto f(N,\phi,\gamma)f(D_t|N,\phi,\gamma). \quad (3.5)$$

Although the joint prior distributions and the likelihood function are easy to calculate, the marginal probability of the data denoted by $f(D_t)$, which is the normalizing constant, is quite difficult to obtain as an explicit expression in many situations.

Under the inflection S-shaped SRGM, the joint posterior distribution for N, ϕ and γ, given D_t can be expressed as

$$\pi(N,\phi,\gamma \mid D_t) \propto L(N,\phi,\gamma \mid D_t)\, g_1(N)\, g_2(\phi)\, g_3(\gamma) = N^{z_n}\prod_{i=1}^{n}\left\{\frac{1-e^{-\phi t_i}}{1+((1-\gamma)/\gamma)e^{-\phi t_i}}-\frac{1-e^{-\phi t_{i-1}}}{1+((1-\gamma)/\lambda)e^{-\phi t_{i-1}}}\right\}^{(z_i-z_{i-1})} \times \exp\left\{-\frac{1-e^{-\phi t_n}}{1+((1-\gamma)/\gamma)e^{-\phi t_n}}\right\} g_1(N)\, g_2(\phi)\, g_3(\gamma),.$$

where $g_1(N)$, $g_2(\phi)$, and $g_3(\gamma)$ denote the prior distributions of N,ϕ and γ, respectively and these three parameters are assumed to be independent .

To compute the Bayesian estimates for N,ϕ and γ, we adopt the Gibbs sampling which is a special case of the MCMC (Markov chain Monte Carlo) technique. Using such sampling scheme, we generate a Markov chain, whose states are the parameters N,ϕ and γ, and whose steady-state (stationary) distributions are $\pi(N|\phi,\gamma,D_t)$, $\pi(\phi|N,\gamma,D_t)$ and $\pi(\gamma|N,\phi,D_t)$, which are the posterior distributions of N,ϕ and γ, respectively. Using these three posterior densities, the random variates are generated sequentially for N,ϕ and γ by following the steps stated below.

i) Take an initial values of N,ϕ and γ, which

are denoted by $N^{(0)}, \phi^{(0)}$ and $\gamma^{(0)}$.

ii) Generate a random draw $N^{(1)}$ of N from $\pi(N|\phi^{(0)}, \gamma^{(0)}, D_t)$..

iii) Generate a random draw $\phi^{(1)}$ of ϕ from $\pi(\phi|N^{(1)}, \gamma^{(0)}, D_t)$.

iv) Generate a random draw $\gamma^{(1)}$of γ from $\pi(\gamma|N^{(1)}, \phi^{(1)}, D_t)$.

v) Repeat the steps ii) - iv) until I random draws of N, ϕ and γ are completed.

At the end of such repetition, we obtain a stationary distribution of this Markov chain for large I and the vector $(N^{(I)}, \phi^{(I)}, \gamma^{(I)})$ has a distribution that is approximately equal to $\pi(N, \phi, \gamma|D_t)$. The convergence of this Markov chain has been discussed in several articles, including Best, Cowles and Vines(1999), in which several convergence me asures were used to develop a checking program for convergence, called CODA (convergence diagnosis and output analysis software for Gibbs sampling output). By starting with independent initial choices, we can replicate the above iterations J times. Let $(N^{(i,j)}, \phi^{(i,j)}, \gamma^{(i,j)})$ denote the realization of Markov chain for the ith iteration and the jth replication. For the purpose of calculating the posterior mean from the sampler, we assign the weight $2/(IJ)$ to each $(N^{(i,j)}, \phi^{(i,j)}, \gamma^{(i,j)}), i=(I/2)+1, \cdots, I$ and $j=1,2,\cdots,J$ as suggested by Gel man and Rubin (1992). In case when the conditional posterior densities are not easily identified to carry out the steps (ii - (iv, we employ the Metropolis-Hastings algorithm within Gibbs sampling to generate the random draws of N, ϕ and γ. For detailed discussions on the Metropolis-Hasting algorithm, refer to Gelman and Rubin(1992).

For our estimation purpose, the states are represented by N, ϕ and γ and its target densities are $\pi(N|\phi, \gamma, D_t), \pi(\phi|N, \gamma, D_t)$ and $\pi(\gamma|N, \phi, D_t)$, respectively.

4. Prior and posterior distribution of N, ϕ and γ

As discussed in the previous section, we need to specify a prior distribution for an unknown parameter of our interests in order to apply the Gibbs sampling and the Metropolis-Hastings algorithm. In this section, we consider certain prior distributions for the parameters N, ϕ and γ to derive its corresponding conditional marginal posterior distributions and to obtain its Bayes estimates based on the Metropolis-Hastings algorithm.

4.1 Prior and posterior distribution of N

The Bayesian procedure is based on different degrees of prior information, ranging from the absence of any information up to a quite detailed kno wledge on the failure process.

Firstly, in absence of any information on the failure process, a commonly used prior is the non-informative prior. The non-informative prior is obtained by applying Jeffrey's rule, which suggests to take the prior density to be proportional to the square root of the determinant of the Fisher's information matrix. Let $g_1(N)$ denote the prior density for N. Then, under the hypothesis of prior independence, Jeffrey's prior for N as a non -infor mative prior is given by $g_1(N) \propto N^{r-1}$, where $\tau=0$ for failure truncated sampling $(T=t_n)$ and $\tau=1/2$ for time truncated sampling $(T>t_n)$. Thus, the conditional marginal posterior distribution for N can be expressed as

$$\pi(N|\phi, \gamma, D_t) \propto N^{z_n+\tau-1} \exp\left\{-\frac{N(1-e^{-\phi t_n})}{1+((1-\gamma)/\gamma)\, e^{-\phi t_n}}\right\}. \quad (4.1)$$

The right hand side of (4.1) can be used as the target density for N to apply the Metropolis -Hast ings algorithm. Since we consider a failure truncated sampling only in this paper, we take $\tau=0$ in equat ion (4.1) for the non-informative prior.

In case the only information available for$N$ is

that the values of a parameter are bounded on (a,b), the uniform prior for N can be used. Based on this prior, the conditional marginal posterior distribution for N can be written as

$$\pi(N|\phi,\gamma,D_t)$$
$$\propto N^{z_n}\exp\left\{-\frac{N(1-e^{-\phi t_n})}{1+((1-\gamma)/\gamma)e^{-\phi t_n}}\right\}$$

Next, we suppose that only a prior mean can be elicited on N, denote it by μ_N. Under this situation, an exponential prior can be used and the conditional marginal posterior distribution for N becomes

$$\pi(N|\phi,\gamma,D_t)$$
$$\propto N^{z_n}\exp\left\{-\left(\frac{1}{\mu_N}N+\frac{N(1-e^{-\phi t_n})}{1+((1-\gamma)/\gamma)e^{-\phi t_n}}\right)\right\}$$

When both prior mean and prior standard deviation, denote it by σ_N, are available, the prior information on the failure process is more detailed and so we apply the following gamma prior with shape and scale parameters of α and β.

$$g_1(N)=\frac{\beta^\alpha}{\Gamma(\alpha)}N^{\alpha-1}\exp(-\beta N),$$
where $\alpha=(\mu_N/\sigma_N)^2$, $\beta=\mu_N/\sigma_N^2>0$.

Under the gamma prior, the conditional marginal posterior distribution for N becomes

$$\pi(N|\phi,\gamma,D_t)\propto N^{z_n+\alpha-1}$$
$$\times\exp\left\{-\left(\beta N+\frac{N(1-e^{-\phi t_n})}{1+((1-\gamma)/\gamma)e^{-\phi t_n}}\right)\right\}$$

Another possible prior distribution for the parameter N is a Poisson distribution with mean λ, in which $g_1(N)=e^{-\lambda}\lambda^N/N!$, $N=0,1,\cdots$. In this case, the conditional marginal posterior distribution can be expressed as

$$\pi(N|\phi,\gamma,D_t)$$
$$\propto\frac{N^{z_n}\lambda^N}{N!}\exp\left\{-\frac{N(1-e^{-\phi t_n})}{1+((1-\gamma)/\gamma)e^{-\phi t_n}}\right\}$$

4.2 Prior and posterior distribution of ϕ and γ

The parameters ϕ and γ assume the values between 0 and 1 and so it is natural to consider a beta distribution as their priors. The prior distribution for γ with the hyper parameters α and β has the following probability density.

$$g_3(\gamma)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\gamma^{\alpha-1}(1-\gamma)^{\beta-1},\alpha>0,\beta>0.$$

In the similar manner as for N, the conditional marginal posterior distribution for γ can be expressed as

$$\pi(\gamma\mid N,\phi,D_t)$$
$$\propto\gamma^{\alpha-1}(1-\gamma)^{\beta-1}\exp\left\{-\frac{N(1-e^{-\phi t_n})}{1+((1-\gamma)/\gamma)e^{-\phi t_n}}\right\}\times$$
$$N^{z_n}\prod_{i=1}^{n}\left\{\frac{1-e^{-\phi t_i}}{1+((1-\gamma)/\gamma)e^{-\phi t_i}}-\frac{1-e^{-\phi t_{i-1}}}{1+((1-\gamma)/\lambda)e^{-\phi t_{i-1}}}\right\}^{(z_i-z_{i-1})}$$

The prior and conditional marginal posterior distributions for ϕ are obtained by exchanging γ and ϕ of $g_3(\gamma)$ and $\pi(\gamma|N,\phi,D_t)$ in the above expressions. By applying the Gibbs sampling and the Metropolis-Hastings algorithm, the Bayes estimates for ϕ and γ may be obtained numerically.

5. Numerical examples

Table 5.1 gives the actual test data for a PL/I application program, which is obtained during the testing period of software having approximately 1,317,000 lines of code. (cf. Ohba (1984))

Using the failure data of Table 5.1, the maximum likelihood estimates and the Bayes estimates for N,ϕ and γ are calculated. As the priors for these parameters, we use two particular sets of priors which would provide the best estimates among those discussed in Chapter 4. Table 5.2 gives the numerical results and it appears that there is not much differences between maximum

likelihood method and Bayes method in terms of SSE for these particular priors.

Table 5.1 Failure data for a PL/I application program

Time of observation (week)	Cumulative execution time	Cumulative number of failures
1	2.45	15
2	4.90	44
3	6.86	66
4	7.84	103
5	9.52	105
6	12.89	110
7	17.10	146
8	20.47	175
9	21.43	179
10	23.35	206
11	26.23	233
12	27.67	255
13	30.93	276
14	34.77	298
15	38.61	304
16	40.91	311
17	42.67	320
18	44.66	325
19	47.65	328

The SSE, which is often used to measure the goodness of fit of the estimate, is computed based on the equation (5.1) below.

$$SSE = \sum_{k=1}^{n}(z_k - \hat{z_k})^2 = \sum_{k=1}^{n}(z_k - \widehat{m}(t_k))^2$$

$$= \sum_{k=1}^{n}\left(z_k - \widehat{N}\left(\frac{1-e^{-\hat{\phi}t_k}}{1+((1-\hat{\gamma})/\hat{\gamma})e^{-\hat{\phi}t_k}}\right)\right)^2, \quad (5.1)$$

where z_k and $\widehat{z_k}$ denote the cumulative number of failures observed up to time t_k and its estimate, respectively.

Although the true values of N, ϕ and γ are unknown in this case, the Bayes estimates and the maximum likelihood estimates are fairly close, especially for the estimates of ϕ.

6. Concluding remarks

This paper proposes a Bayesian approach to estimate three parameters characterizing Ohba's (1984) SRGM by adopting the Markov chain Monte Carlo techniques and compare the proposed method with the well-known maximum likelihood method. When applying the maximum likelihood estimation method, there may be some cases for which the estimates for these parameters may not be found due to a wrong choice of initial value or failure to converge to certain values. On the other hand, the Bayesian method provides an efficient tool in estimating the parameters in such cases, being less

Table 5.2 Estimates of N, ϕ and γ based on the ML and Bayesian method using actual failure data

	Point estimates			
Parameter	N	ϕ	γ	SSE
Bayesian	Non-informative	U(0,1)	U(0,1)	
	358.967 356.125	0.079306 0.081034	0.30942 0.27469	3911.04066 2885.07362
	U(0, 1000)	Beta(1,2)	Beta(2,2)	
	361.735 358.862	0.077169 0.078642	0.32259 0.28985	3990.66828 2877.66597
ML	351.748	0.08400	0.25696	2929.53283

(*) the first entry is the posterior mean and the second entry is the posterior median.

prone to the convergence failure compared to the maximum likelihood method. The convergence of Markov chain in MCMC has not been proved analy tically. However, the MCMC technique is known to achieve the convergence in most of the cases with out much difficulties. Kho and Yae(2005) state "reg ardless of starting state, if we run this simulation long enough, we will end up sampling from the po sterior probability distribution". Numerical example based on real data indicates that the Bayesian met hod is quite competitive with the maximum likelihood method in most cases.

Acknowledgement

This research was supported by Hallym University Research Fund, 2009(HRF-2009-000).

References

[1] Best, N. G., Cowles, M. K. and Vines, S. K. (1996), "CODA : Convergence Diagnosis and Output Analysis Software for Gibbs Sampling Output, Version 0.40, *University of Nebraska Medica Center*, USA.

[2] Gelman, A. and Rubin, D. B.(1992), "Inference from iterative simulation using multiple sequences", *Statistical Science*, vol.7, pp.457-472.

[3] Goel, A. L. and Okumoto, K.(1979), "Time Dependent ErrorDetection Rate Model for Software Reliability and Other Performance Measures", *IEEE Transactions on Reliability*, vol.28, pp.206-211.

[4] Kim, H. S., Yamada, S.and Park, D. H. (2005), "Bayesian Approach to Optimal Release Policy of Software System", *IEICE TRANS on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E88-A No.12. pp.3618 -3626

[5] Kimura, M., Toyota, T. and Yamada, S. (1999), "Economic Analysis of Software Release Problems with Warranty Cost and Reliability Requirement", *Reliability Engineering & System Safety*, vol. 66, pp.49-55.

[6] Nayak, T. K., Bose, S. and Kundu, S. (2008), "On inconsistency of estimators of parameters of non-homogeneous Poisson process models for software reliability", *Statistics and Probability Letters*, vol. 78, pp.2217-2221.

[7] Kho, B. C. and Yae, S. M. (2005), "Bayesian Analysis of a Stochastic Beta Model in Korean Stock Markets", *The Korean Journal of Financial Management*, vol.22, No. 2, pp.43-69.

[8] Kuo, L., Lee, J.C., Choi, K. and Yang, T.Y. (1997), "Bayesian Inference for S-Shaped Software Reliability Growth Models", *IEEE Transactions on Reliability*, vol.46, pp.76-80.

[9] Ohba, M.(1984), "Software reliability analysis models", *IBM J. Research and Development*, vol.28, 428-443.