⊠연구논문

몬테칼로 깁스방법을 적용한 소프트웨어 신뢰도 성장모형에 대한 베이지안 추론과 모형선택에 관한 연구

김희철
동국대학교 통계학과
이승주
청주대학교 응용통계학과

Bayesian Inference and Model Selection for Software Growth Reliability Models using Gibbs Sampler

Heecheul Kim
Dept. of statistics, Dongguk University, Seoul
Seungjoo Lee
Dept. of Applied Statistics, Chongju University, Chongju

Abstract

Bayesian inference and model selection method for software reliability growth models are studied. Software reliability growth models are used in testing stages of software development to model the error content and time intervals between software failures. In this paper, we could avoid the multiple integration by the use of Gibbs sampling, which is a kind of Markov Chain Monte Carlo method to compute the posterior distribution. Bayesian inference and model selection method for Jelinski-Moranda and Goel-Okumoto and Schick-Wolverton models in software reliability with Poisson prior information are studied. For model selection, we explored the relative error.

1. 서론

우리들의 주변에는 복잡한 소프트웨어 시스템(System)들로 둘러 쌓여 있으며 이러한 시스템의 혜택을 받는 일이 커짐에 따라 소프트웨어 신뢰성의 역할은 점차 커지게 되었다. 그러나 시스템이 고장이 나면 고장이 난 원인을 찾아 필요할 경우 새로운 디자인을 개발하거나 새로운 기술을 도입하게 된다. 따라서 시간이 지남에 따라 신뢰도의 증가가 기대되어 진다. 이런 모형을 신뢰도 성장모형(Reliability growth model)이라고 한다. 신뢰도(Reliability)는 시스템, 제품 또는 부품이 어떤 정하여진 조건 아래서 의도하는 기간 동안에 요구된 기능을 발휘하는 확률이라고 정의할 수 있다. 따라서, 본 논문은 신뢰도 측면을 고찰하기 위해 신뢰도 이론을 사용하고자 한다. 신뢰도 이론은 장비나 무기의 일부분이나 소프트웨어 시스템 전체가 규정된 환경조건하에서 의도하는 기간 동안에 요구된 기능을 만족스럽게 수행할 수 있는 확률을 예측하고 신뢰도를 향상시키기 위한 실제적인 도구가 되며, 시스템, 제품 또는 부품이 의도된 기간동안 요구된 기능을 발휘할 수 있는 확률을 보다 정확한 방법으로 추정하는 연구는 이 분야의 관심사항이 된다. 베이즈 추정법에서 사전 확률분포인 수명분포가 복잡하면 적분이 불가능하므로 사후정보의 추출이 불가능해 진다.

본 연구에서는 깁스 샘플링(Gibbs sampling)을 이용하여 적분이 풀리지 않는 경우에 근사적인 깁스 추정량을 유도하여 최우추정량과 비교하고자 한다. 데이터 증대(Data augmentation)를 위한 MCMC(Markov chain monte carlo)기법의 하나인 깁스 추정법은 사후분포의 특징을 계산하기 위해 제시되었다. 이러한 데이터증대 접근방법은 마코브체인을 사용하여 추이측도(Transition measure)의 계산을 용이하게 한다. 몬테칼로 적분 추정기법은 종래의 수치해석적 적분을 대신할 수 있는 강력한 도구로 각광받고 있다. 적당한 차수까지 연속인 도함수가 존재하는 경우에는 종래의 수치해석적 적분 알고리즘이 여전히 유용하고 정밀한 답을 제시한다. 그러나 정상적인 함수의 적분문제에서도 몬테칼로 적분추정기법이 수치해석적 방법에 비하여 효율이 떨어지지 않을 뿐만 아니라 연속성이 깨지는 상황이거나 다차원 적분문제에서는 몬테칼로 적분추정이 아주 유용하게 이용된다.

소프트웨어 테스팅 단계에서 소프트웨어 오류수와 고장간격시간에 의해 소프트웨어 고장현상을 수리적으로 모형화 하면 소프트웨어에 대한 평가를 쉽게 할 수 있으며, 신뢰도 성장모형에 의해 소프트웨어 오류수, 소프트웨어 고장발생간격시간, 소프트웨어 신뢰도 및 고장률등의 신뢰성 평가 측도들이 추정되어 예측할 수 있다.

Jelinski-Moranda모형은 소프트웨어 신뢰도 모형에서 가장 기본적인 모형이다. 그러나 이 모형은 모든 오류들이 같은 크기를 가진다는 가정 하에서 순수하게 결정적이고 중요한 비율의 열만 취급한다는 비판이 있었다. 이러한 비판을 개선하기 위해 Goel-Okumoto(1978)가 제시한 개선된 Jelinski-Moranda 모형과 Schick-Wolverton 모형이 제시되었다.

본 연구는 Jelinski-Moranda 모형과 Goel-Okumoto 모형 그리고 Schick-Wolverton

모형에 전통적인 최우추정법을 적용하여 모수를 추정을 하고, 위험함수(hazard function)의 형태와 초기오류수의 형태의 사전분포가 포아송분포인 경우를 선택하여 깁스추출법을 이용한 베이지안 추론과 상대오차의 합을 이용하여 모형 선택을 시도하였다.

2. 신뢰성 모형

2.1 Jelinski-Moranda 모형

신뢰도 모형 중에서 대표적인 모형은 Jelinski-Moranda(JM)(1972)모형이다. 이 모형은 초기 고장 수에 의존하는 모형이며 소프트웨어 공학 분야에서 일반화된 소프트웨어 신뢰도 성장 모형으로 알려져 왔고, 소프트웨어 고장 데이터를 기술하는 가장 간단한 모형이다. N을 테스팅 스테이지 초기의 소프트웨어 총오류의 수이고, β는 한 오류마다의 오류 발견률을 나타내는 상수라 하자. 그러면 JM모형의 가정은 다음과 같다.

1. N은 소프트웨어에 존재하는 초기 총오류의 수이고, 미지의 고정된 상수이다.
2. 고장의 원인은 한 개의 오류에 의해 발생하고, 추가적인 오류는 발생하지 않으며 디버깅 절차는 완벽하다.
3. 고장이 발견되면 즉시 수정되며 수정에 걸리는 시간은 거의 없고 고장시간을 x_i 라고 하면 고장발생 간격시간 $t_i = x_i - x_{i-1}$는 각각 독립적으로 평균이 $1/\beta(N-i+1)$인 지수분포를 따른다. 즉, T_i에 대한 위험함수는 모든 t에 대하여 $h_{T_i}(t) = \beta(N-i+1)$ 이 되고 t_i에 대한 확률밀도함수는 다음과 같다.

$$f(t_i) = (N-i+1)\beta e^{-\beta(N-i+1)t_i}.$$

2.2 Goel-Okumoto 모형

Goel-Okumoto(1978)는 개선된 JM모형(GO모형)을 제시하였다. 이 모형은 JM모형과 비슷하지만 하나의 버그가 발견되었을 때 그것을 고치는 데에 확률 $w(0 \le w \le 1)$가 있는 불완전한 디버깅 모형(Imperfect debugging model)을 제안했다. Goel-Okumoto 모형에서의 가정은 JM 모형에서의 가정 1과 2는 같고 i번째의 고장 간격시간에 대한 위험함수는 모든 t에 대하여 $h_{T_i}(t) = \beta(N-w(i-1))$ 이 된다. 즉 t_i에 대한 확률밀도함수는 다음과 같다.

$$f(t_i) = (N-w(i-1))\beta e^{-\beta(N-w(i-1))t_i}.$$

2.3 Schick-Wolverton 모형

Schock-Wolverton(1978)은 JM모형을 수정한 위험함수 $h_{T_i}(t)=\beta(N-i+1)(t-x_{i-1})$ 이고 t는 $x_{i-1}<t\leq x_i$을 만족하는 톱니 연결 위험함수(Sawtooth concatenated hazard function)를 제안하였다. 이 SW모형의 경우에는 n번째 고장이 발생되었을 때 (단, 이때 시간을 x_i 라 가정함), $D_{t_n}=\{t_1, t_2, \ldots, t_n\}$를 관측된 연속적인 고장간격시간 (t_i)이라고 하면 고장간격시간 $t_1, \cdots, t_n$과 N, β가 주어졌을 때 t_i는 위험함수 $h_{T_i}(t) = \beta(N-i+1)t_i$을 가지는 레일리(Rayleigh)분포를 따르고 조건적으로 독립이라고 가정한다. 즉 t_i에 대한 확률밀도함수는 다음과 같다.

$$f(t_i) = (N-i+1)\beta t_i e^{-\frac{\beta}{2}(N-i+1)t_i^2}.$$

3. 신뢰도 모형에 대한 깁스 샘플링

사전분포가 포아송인 경우의 신뢰도 모형을 본 절에서 논의하고자 한다. 추가정보인 사전 분포를 정의함에 있어서, $N \perp \beta$는 N와 β가 독립이라는 것을 의미하며, $P(\theta)$는 평균이 θ인 포아송분포를, 그리고 $\Gamma(a,b)$는 평균이 a/b인 감마분포(Gamma distribution)를 나타낸다.

3.1 Jelinski-Moranda 모형

위험함수가 $h_{T_i}(t) = \beta(N-i+1)$ 인 JM모형의 경우에는 n번째 고장이 발생되었을 때 (단, 이때 시간을 x_i 라 가정함), $D_{t_n}=\{t_1, t_2, \ldots, t_n\}$를 관측된 연속적인 고장간격시간(t_i)은 $t_i = x_i - x_{i-1}$이 되고 소프트웨어에 존재하는 미지의 총 N개의 오류로 인한 총 고장시간은 $T = \sum_{i=1}^{n}(N-i+1)t_i$ 임을 고려하면 n번째 까지 관찰된 JM모형에 대한 우도함수는 다음과 같다.

$$\begin{aligned} L(N,\beta;D_{t_n}) &= \prod_{i=1}^{n}(N-i+1)\beta e^{-(N-i+1)\beta t_i} \\ &= \{\prod_{i=1}^{n}(N-i+1)\}\beta^n e^{-\beta T} \end{aligned} \tag{3.1}$$

식(3.1)에서 모수 N과 β를 추정하기 위하여 최우추정법과 깁스 추출법을 사용하고자 한다.

(A) 최우추정법(MLE)에 의한 접근

Schick와 Wolverton(1978)은 식(3.1)에서 최대우도를 사용하여 모수 N과 β는 다음과 같은 두 식을 만족하는 최우추정치 $\hat{N}_{MLE}$, $\hat{\beta}_{MLE}$를 유도하였다.

$$\sum_{i=1}^{n}\frac{1}{\hat{N}_{MLE}-(i-1)}=\frac{n\sum_{i=1}^{n}t_i}{\hat{N}_{MLE}\cdot\sum_{i=1}^{n}t_i-\sum_{i=1}^{n}(i-1)t_i} \tag{3.2}$$

$$\hat{\beta}_{MLE}=\frac{n}{\hat{N}_{MLE}\cdot\sum_{i=1}^{n}t_i-\sum_{i=1}^{n}(i-1)t_i} \tag{3.3}$$

(3.3)식을 이용하여 근사값을 계산할때 수치해석적인 이분법(Bisection method)이나 뉴톤-렙슨법(Newton-Raphson method)을 이용하여 계산할 수 있다. 그러나 참값을 알고 있는 경우는 초기값의 범위를 잘 설정할 수 있기 때문에 기본적인 방법인 이분법이 효율적인 방법이 됨이 알려져 있다.

(B) 베이지안 접근

깁스알고리즘을 이용한 베이지안 접근은 사전분포 $N\sim P(\theta)$; $\Gamma(a,b)$; $N\perp\beta$을 사용하면 N과 β의 사후 결합밀도함수는 다음과 같다.

$$\begin{aligned}P(N,\beta\mid D_{t_n}) &\propto L(N,\beta\mid D_{t_n})\cdot P(N,\beta)\\ &\propto L(N,\beta\mid D_{t_n})\cdot P(N)\cdot P(\beta)\\ &\propto \{\prod_{i=1}^{n}(N-i+1)\}\beta^n e^{-\beta T}\cdot\frac{e^{-\theta}\theta^N}{N!}\cdot\frac{b^a\beta^{a-1}e^{-b\beta}}{\Gamma(a)}\end{aligned} \tag{3.4}$$

베이즈 정리와 장애(Nusiance) 모수의 개념을 이용하여 N의 사후 조건부 밀도함수를 구하기 난해하므로 고유변수 $N'=N-n$를 이용하면 다음과 같다.

$$
\begin{aligned}
P(N' \mid \beta, D_{t_n}) &\propto L(N,\beta \mid D_{t_n}) \cdot P(N) \\
&\propto \{\prod_{i=1}^{n}(N-i+1)\}\beta^n e^{-\beta T} \cdot \frac{e^{-\theta}\theta^N}{N!} \\
&\propto \{\prod_{i=1}^{n}(N-i+1)\}\beta^n e^{-\beta[\sum_{i=1}^{n}(N-i+1)t_i]} \cdot \frac{e^{-\theta}\theta^N}{N!} \\
&\propto \frac{N!}{(N-n)!}\beta^n \exp[-\beta[\sum_{i=1}^{n}(N-i+1)t_i]] \cdot \frac{e^{-\theta}\theta^N}{N!} \\
&\propto \frac{(\theta e^{-\beta\sum_{i=1}^{n}t_i})^{N-n} \cdot e^{-\theta\exp[-\beta\sum_{i=1}^{n}t_i]}}{(N-n)!} \\
&\propto \frac{\theta^{(N-n)}}{(N-n)!} \cdot \exp[-\beta\sum_{i=1}^{n}t_i]^{N-n+1} \cdot e^{-\theta}
\end{aligned}
$$

따라서 $N' = N-n$에 대한 사후분포는 평균 $\theta\exp[-\beta\sum_{i=1}^{n}t_i]$을 가지는 포아송분포가 된다.

$$
P(N' \mid \beta, D_{t_n}) \sim P\left(\theta\exp[-\beta\sum_{i=1}^{n}t_i]\right) \tag{3.5}
$$

그리고 β 에 대한 사후 조건부분포는 (3.4)를 이용하면 다음과 같다.

$$
\begin{aligned}
P(\phi \mid N, D_{t_n}) &\propto \{\prod_{i=1}^{n}(N-i+1)\}\beta^n e^{[-\beta\sum_{i=1}^{n}(N-i+1)t_i]} \cdot \frac{b^a\beta^{a-1}e^{-b\beta}}{\Gamma(a)} \\
&\propto \frac{N!}{(N-n)!}\exp[-\beta(\sum_{i=1}^{n}(N-i+1)t_i+b)] \cdot \beta^{n+a-1} \cdot \frac{b^a}{\Gamma(a)} \\
&\propto \frac{(\sum_{i=1}^{n}(N-i+1)t_i+b)^{a+n}}{\Gamma(a+n)} \cdot \beta^{n+a-1}\exp[-\beta(\sum_{i=1}^{n}(N-i+1)t_i+b)] \\
&\propto \frac{(b+N'x_n+\sum_{i=1}^{n}x_i)^{a+n}}{\Gamma(a+n)}\beta^{n+a-1}\exp[-\beta(b+N'x_n+\sum_{i=1}^{n}x_i)]
\end{aligned}
$$

그러므로 β 에 대한 사후분포는 다음과 같은 감마분포가 된다.

$$
P(\beta \mid N, D_{t_n}) \sim \Gamma(a+n,\ b+N'x_n+\sum_{i=1}^{n}x_i) \tag{3.6}
$$

3.2 Goel-Okumoto 모형

위험함수가 $h_{T_i}(t)=\beta(N-w(i-1))$인 GO모형(단, $0\le w\le 1$)인 경우는 앞의 경우와 유사하게 $T=\sum_{i=1}^{n}(N-w(i-1))t_i$이 되고, n번 까지 관찰된 우도함수를 구해보면 다음과 같다.

$$\begin{aligned} L(N,\beta \mid D_{t_n}) &= \prod_{i=1}^{n}(N-w(i-1))\beta e^{-(N-w(i-1))\beta t_i} \\ &= \{\prod_{i=1}^{n}(N-w(i-1))\}\beta^n e^{-\sum_{i=1}^{n}(N-w(i-1))\beta t_i} \end{aligned} \tag{3.7}$$

식(3.7)에서 모수 N,β를 추정하기 위하여 최우추정법과 깁스 추출법을 사용하고자 한다.

(A) 최우추정법에 의한 접근

최우추정치 $\hat{N}_{MLE}$, $\hat{\beta}_{MLE}$는 앞 절의 Jelinski-Moranda 모형과 유사하게 다음과 같은 식을 유도할 수 있다.

$$\sum_{i=1}^{n}\frac{1}{\hat{N}_{MLE}-w\cdot(i-1)}=\frac{n\sum_{i=1}^{n}t_i}{\hat{N}_{MLE}\cdot\sum_{i=1}^{n}t_i-\sum_{i=1}^{n}w\cdot(i-1)t_i} \tag{3.8}$$

$$\hat{\beta}_{MLE}=\frac{n}{\hat{N}_{MLE}\cdot\sum_{i=1}^{n}t_i-\sum_{i=1}^{n}w\cdot(i-1)t_i} \tag{3.9}$$

식(3.9)는 비선형형태를 가지고 있으므로 수치해석적 방법을 이용하여 근을 계산할 수 있다.

(B) 베이지안 접근

깁스 추출 알고리즘에 의한 사전정보인 동시에 초모수(Hyperparameter)인 N과 β의 사전분포는 다음과 같다.

$$N\sim P(\theta);\ \beta\sim\Gamma(a,b);\ N\perp\beta \tag{3.10}$$

따라서 N과 β의 사후 결합밀도함수는 식(3.7)의 우도함수와 식(3.10)의 사전분포를

가지고 베이즈정리에 의해서 다음과 같이 유도된다.

$$P(N,\beta \mid D_{t_n}) \propto \{\prod_{i=1}^{n}(N-w(i-1))\}\beta^n e^{-\beta\sum_{i=1}^{n}(N-w(i-1))t_i} \cdot \frac{e^{-\theta}\theta^N}{N!} \cdot \frac{b^a\beta^{a-1}e^{-b\phi}}{\Gamma(a)} \quad (3.11)$$

베이즈 정리와 장애모수의 개념을 이용하여 N의 사후 조건부밀도를 구하기 난해하므로 고유변수 $N'' = N-nw$를 이용하여 JM모형과 유사한 방법으로 계산하면 다음과 같다.

$$P(N' \mid \beta, D_{t_n}) \propto L(N,\beta \mid D_{t_n}) \cdot P(N)$$

$$\propto \frac{\left[\theta e^{-\beta\sum_{i=1}^{n}t_i}\right]^{N-nw} e^{-\theta e^{-\beta\sum_{i=1}^{n}t_i}}}{(N-nw)!}$$

그러므로 $N'' = N-nw$에 대한 사후 분포는 평균 $\theta\exp[-\beta\sum_{i=1}^{n}t_i]$을 가지는 포아송분포가 된다.

$$P(N'' \mid \beta, D_{t_n}) \sim P\left(\theta\exp[-\beta\sum_{i=1}^{n}t_i]\right) \quad (3.12)$$

β 에 대한 사후분포는 (3.11) 식으로부터

$$P(\beta \mid N, D_{t_n}) \propto \frac{\left(\sum_{i=1}^{n}(N-w(i-1))t_i+b\right)^{a+n}}{\Gamma(a+n)} \cdot \beta^{n+a-1}\exp\left(-\beta(\sum_{i=1}^{n}(N-w(i-1))t_i+b)\right)$$

그러므로 β에 대한 사후분포는 다음과 같은 감마분포가 된다.

$$\begin{aligned} P(\beta \mid N, D_{t_n}) &\sim \Gamma\left(a+n,\ \sum_{i=1}^{n}(N-w(i-1))t_i+b\right) \\ &\sim \Gamma\left(a+n,\ b+N'x_n+w\sum_{i=1}^{n}x_i\right) \end{aligned} \quad (3.13)$$

3.3 Schick-Wolverton 모형

위험함수가 $h_{T_i}(t) = \beta(N-i+1)t_i$인 SW모형의 경우에는 n번째 고장이 발생되었을 때 (단, 이때 시간을 x_i 라 가정함), $D_{t_n} = \{t_1, t_2, \ldots, t_n\}$를 관측된 연속적인

고장간격시간 (t_i)이라고 하면 소프트웨어에 존재하는 미지의 총 N개의 오류로 인한 총 고장시간은 $T=\sum_{i=1}^{n}(N-i+1)t_i$ 임을 고려하면 n번 까지 관찰된 SW모형에 대한 우도함수는 다음과 같다.

$$\begin{aligned} L(N,\beta \mid D_{t_n}) &= \prod_{i=1}^{n}[(N-i+1)\beta t_i \exp\{-\frac{\beta}{2}(N-i+1)\ t_i^2\}] \\ &= \{\prod_{i=1}^{n}(N-i+1)\}\beta^n\{\prod_{i=1}^{n}t_i\}\exp\{-\frac{\beta}{2}\sum_{i=1}^{n}(N-i+1)\ t_i^2\} \end{aligned} \tag{3.14}$$

(A) 최우추정법에 의한 접근

Schick와 Wolverton(1978)은 식(4.15)에서 최대우도를 사용하여 모수 N과 β는 다음과 같은 최우추정치 $\hat{N}_{MLE}$, $\hat{\beta}_{MLE}$를 유도하였다.

$$\hat{N}_{MLE}=\left[\frac{2n}{\hat{\beta}_{MLE}}+\sum_{i=1}^{n}(i-1)t_i^2\right]\cdot\frac{1}{\sum_{i=1}^{n}t_i^2} \tag{3.15}$$

$$\hat{\beta}_{MLE}=\left[\sum_{i=1}^{n}\frac{2}{\hat{N}_{MLE}-(i-1)}\right]\cdot\frac{1}{\sum_{i=1}^{n}t_i^2} \tag{3.16}$$

식(3.15)와 식(3.16)은 각각 비선형이기 때문에 수치해석적 방법을 이용하여 근을 계산할 수 있다.

(B) 베이지안 접근

사전정보이고 초모수(Hyperparameter)인 N과 β의 사전분포는 다음과 같다.

$$N\sim P(\theta);\ \beta\sim\Gamma(a,b);\ N\perp\beta \tag{3.17}$$

따라서 N과 β의 사후 결합밀도함수는 식(3.14)의 우도함수와 식(3.17)의 사전분포를 가지고 베이즈 정리에 의해서 다음과 같이 유도된다.

$$\begin{aligned} &P(N,\beta \mid D_{t_n}) \\ &\propto\{\prod_{i=1}^{n}(N-i+1)\}\beta^n\{\prod_{i=1}^{n}t_i\}\exp\{-\frac{\beta}{2}\sum_{i=1}^{n}(N-i+1)\ t_i^2\}\frac{e^{-\theta}\theta^N}{N!}\cdot\frac{b^a\beta^{a-1}e^{-b\beta}}{\Gamma(a)} \end{aligned} \tag{3.18}$$

베이즈 정리와 장애(Nusiance) 모수의 개념을 이용하여 N의 사후 조건부밀도를 구하기 난해하므로 고유변수 $N'=N-n$를 이용하여 JM모형과 유사한 방법으로 계산하면 다음과 같다.

$$P(N' \mid \beta, D_{t_n}) \propto \frac{(\theta e^{-\frac{\beta}{2}\sum_{i=1}^{n} t_i^2})^{N-n} \cdot e^{-\theta \exp[-\frac{\beta}{2}\sum_{i=1}^{n} t_i^2]}}{(N-n)!}$$

그러므로 $N'=N-n$ 에 대한 사후분포는 평균 $\theta \exp[-\frac{\beta}{2}\sum_{i=1}^{n} t_i^2]$ 을 가지는 포아송분포가 된다.

$$P(N' \mid \beta, D_{t_n}) \sim P\left(\theta \exp[-\frac{\beta}{2}\sum_{i=1}^{n} t_i^2]\right) \tag{3.19}$$

β에 대한 사후분포는 (3.18)식으로부터

$$P(\beta \mid N, D_{t_n}) \propto \frac{\left(\frac{1}{2}[(N-n)\sum_{i=1}^{n} t_i^2 + n\sum_{i=1}^{n} t_i^2 - \sum_{i=1}^{n}(i-1)\ t_i^2 + 2b]\right)^{a+n}}{\Gamma(a+n)}$$

$$\times\ \beta^{n+a-1} \exp\left[-\frac{\beta}{2}\left((N-n)\sum_{i=1}^{n} t_i^2 + n\sum_{i=1}^{n} t_i^2 - \sum_{i=1}^{n}(i-1)\ t_i^2 + 2b\right)\right]$$

그러므로 β 에 대한 사후분포는 다음과 같은 감마분포가 된다.

$$P(\beta \mid N, D_{t_n}) \sim \Gamma\left[a+n,\ \frac{1}{2}\left((N-n)\sum_{i=1}^{n} t_i^2 + n\sum_{i=1}^{n} t_i^2 - \sum_{i=1}^{n}(i-1)\ t_i^2 + 2b\right)\right] \tag{3.20}$$

3.4 깁스 알고리즘 시행단계

JM모형의 조건부 밀도를 이용한 깁스 알고리즘 (3.5)식과 (3.6)식을 이용하여 다음과 같은 단계를 이용하여 시행한다. 본 논문에서는 사전분포를 $N \sim P(30)$ 로 초기치를 주었으며, β에 대한 사전정보는 확산(diffuse) 사전분포을 이용하여 넓은 범위에서 표본발생이 이루어지도록 비교적 분산이 큰 감마분포 $\Gamma(1, 0.0001)$를 선택하였다.

(0단계)

P(30)와 $\Gamma(1, 0.0001)$의 분포에서 데이터를 랜덤추출하여 초기값을 각각 $N^{(0)}, \beta^{(0)}$을 정한다.

(1-1단계)

$\beta=\beta^{(0)}$로 고정시켰을 경우 (3.5)에 대입하여 생성된 랜덤표본 하나를 $N^{(1)}$이라 한다. 즉,

$$N^{(1)}-n \sim P\left(30\exp[-\beta^{(0)}\sum_{i=1}^{n}t_i]\right).$$

(1-2단계)

(1-1)단계의 $N=N^{(1)}$로 고정시켰을 경우 (3.6)에 대입하여 생성된 랜덤표본 하나를 $\phi^{(1)}$이라 한다. 즉,

$$\beta^{(1)} \sim \Gamma\left(1+n,\ 0.0001+(N^{(1)}-n)\,x_n+\sum_{i=1}^{n}x_i\right).$$

(2단계)

(1-1단계), (1-2단계)로부터 고정시킨 N, β의 값을 가장 최근에 생성된 랜덤표본의 값으로 대체하면서 (1-1단계), (1-2단계)를 충분히 큰 수 k 만큼 반복 수행한다. 이렇게 하여 얻은 최종표본을 ($N^{(k)}$, $\beta^{(k)}$) 이라 한다.

(3단계)

(1-1단계), (1-2단계)를 다시 (m-1)번 충분히 반복 적용하면 총 m개의랜덤 표본 ($N_1^{(k)}$, $\beta_1^{(k)}$), ($N_2^{(k)}$, $\beta_2^{(k)}$), …, ($N_m^{(k)}$, $\beta_m^{(k)}$)이 얻어진다.

(4단계)

최종적인 결과에 의해 N, β의 추정은 다음과 같다.

$$\hat{N}_{Gibbs}=\frac{1}{m}\sum_{l=1}^{m}N_l^{(k)},\quad \hat{\beta}_{Gibbs}=\frac{1}{m}\sum_{l=1}^{m}\beta_l^{(k)}$$

GO모형이나 SW모형도 유사한 방법을 적용하여 사후정보를 계산 할 수 있다.

4. 모형선택

본 연구에서는 모형선택에 있어서 상대오차의 합(the sum of relative errors)으로 비교하고자 한다. 상대오차의 합은 다음과 같이 정의된다.

$$RE(l) = \sum_{i=1}^{n} \left| \frac{\text{참값모형} - \text{추정값모형}}{\text{참값모형}} \right|$$

단, 참값모형은 각각의 모형에 관찰된 값 t_i와 N, ϕ값을 대입했을 때의 값이고 추정값모형은 각각의 모형에 관찰된 값 t_i와 MLE에 의한 모수추정치 $\hat{N}_{MLE}$, $\hat{\beta}_{MLE}$을 대입하거나 깁스 알고리즘에 의해 구해진 $\hat{N}_{GIBBS}$과 $\hat{\beta}_{GIBBS}$를 대입했을 때의 값이고, l은 인덱스(index)된 모형을 나타내고 $RE(l)$의 값이 작으면 보다 좋은 모형이라고 할 수 있을 것이다.

5. 수치적인 예

t_i에 대한 자료는 $N=35$, $\beta=0.00045$인 레일리분포에서 생성된 난수를 이용하였다. <표 1>에 대한 정보를 계산하면 다음과 같다.

$$\sum_{i=1}^{30} t_i = 419, \quad \sum_{i=1}^{30} i\,t_i = 7190, \quad \sum_{i=1}^{30} t_i^2 = 8021, \quad \sum_{i=1}^{30} i\,t_i^2 = 153572.$$

<표 1> 모의실험된 자료 t_i

고장번호 i	t_i	고장번호	t_i	고장번호	t_i
1	14	11	11	21	11
2	17	22	10	22	38
3	20	13	5	23	14
4	4	14	16	24	2
5	7	15	8	25	6
6	8	16	18	26	20
7	14	17	25	27	18
8	5	18	1	28	14
9	13	19	13	29	35
10	11	20	24	30	17

최우추정치의 계산은 비선형식이 되기 때문에 수치해석적인 이분법을 사용하여 C-언어를 이용하여 근사값을 계산하였다. N에 대한 참값이 35이므로 초기값은 30과 45를 각각 주었고 β는 참값이 0.00045이므로 초기값을 0.0001과 0.0006를 주었고 반복은 100번을 하였다.

깁스 샘플링을 하는데 있어서의 N에 대한 사전분포를 $N \sim P(30)$로 초기치를 주었

으며, β에 대한 사전분포는 확산(Diffuse) 사전분포를 이용하기 위해 분산이 비교적 큰 $\Gamma(1,0.0001)$을 택하였다. 깁스 샘플링에서 수렴성을 고려하기 위해 반복(k)은 500번, 반복에 대한 적용(m)은 각각 1000, 2000, 3000번씩 하였다. <표 2>는 사전분포가 포아송분포일때 JM모형에서 추정된 N과 β값이다. 추정된 값은 N이 36.9, β이 0.0036으로 간주할 수 있음을 알 수 있다.

<표 2> JM모형

적 용	반 복	$\hat{N}_{GIBBS}$	$\hat{N}_{MLE}$	$\hat{\beta}_{GIBBS}$	$\hat{\beta}_{MLE}$
1000	500	36.91317032	42.000028	0.0035924382	0.0030033
2000	500	36.88713961		0.0035931321	
3000	500	36.95629959		0.0035903111	

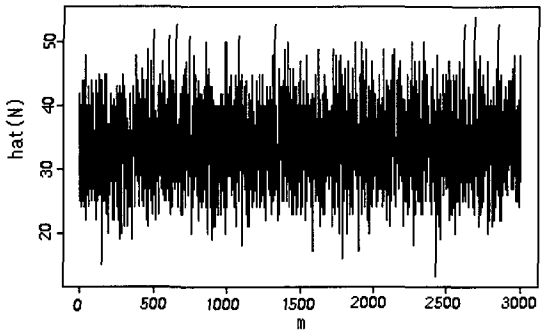
<표 3>은 사전분포가 포아송분포일때 GO모형에서 추정된 값이다. 여기서 모형에 대한 가중치 $w = 0.5$ 에서 0.9까지 고려하여 계산되었다. 위 표에서 w가 커짐에 따라 N추정값과 β추정값이 커지고 있음을 알 수 있다. <표 4>는 사전분포가 포아송일때 SW모형에서 추정된 값이다. 이 표에서 N의 추정값은 35, β의 추정값은 0.00046으로 간주할 수 있고 <그림 1-1>과 <그림 1-2>에서 500번 반복 이후 3000번 적용한 표본의 분포도는 각각 35와 0.00046의 근방에서 분포됨을 알 수 있고 <그림 2-1>과 <그림 2-2>에서는 깁스 알고리즘에서 얻어진 데이터(본 논문에서는 3000개)들이 어떤 특정한 분포로 수렴되어 있음을 진단하는 역활을 하는 Q-Q그림을 사용한 결과 N과 β에 대한 그림이 거의 직선의 형태를 보이고 있으므로 특정한 안정분포로 수렴되고 있음을 보여주고 있다.

<표 3> GO모형

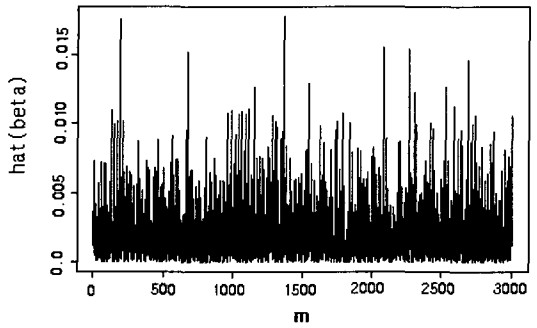
w	적 용	반 복	$\hat{N}_{GIBBS}$	$\hat{N}_{MLE}$	$\hat{\beta}_{GIBBS}$	$\hat{\beta}_{MLE}$
w=0.5	1000	500	27.121052749	39.38386107	0.0021459272	0.0022431
	2000	500	27.181253578		0.0022009502	
	3000	500	27.180085483		0.0021827217	
w=0.6	1000	500	29.507054190	41.19985153	0.0023291275	0.0023627
	2000	500	29.460410486		0.0023331315	
	3000	500	29.452477182		0.0023392818	
w=0.7	1000	500	31.524685779	41.43427165	0.0025565968	0.0024968
	2000	500	31.545646316		0.0025445056	
	3000	500	31.628092876		0.0025403081	
w=0.8	1000	500	33.462432787	41.59955293	0.0028242289	0.0026448
	2000	500	33.546544904		0.0027890575	
	3000	500	33.489501816		0.0028145645	
w=0.9	1000	500	35.322909582	41.8004473	0.0031410629	0.0028126
	2000	500	35.310287388		0.0031403270	
	3000	500	35.354818170		0.0031309736	

<표 4> SW모형

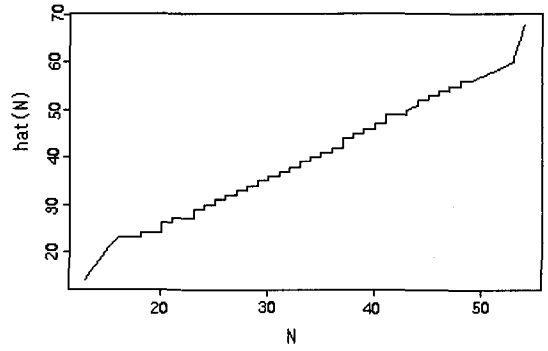
적 용	반 복	$\hat{N}_{GIBBS}$	$\hat{N}_{MLE}$	$\hat{\beta}_{GIBBS}$	$\hat{\beta}_{MLE}$
1000	500	35.095903780	35.2003281	0.0004593424	0.000537179
2000	500	34.975803379		0.0004650525	
3000	500	35.040001268		0.0004636040	



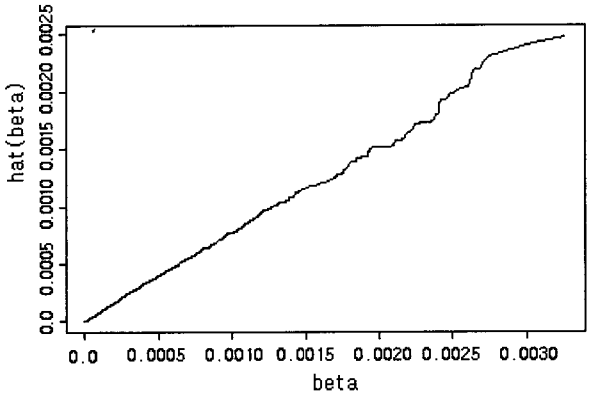
<그 림 1-1> $\hat{N}_{GIBBS}$의 3000(m)번 적용한 표본 분포도



<그 림 1-2> $\hat{\beta}_{GIBBS}$의 3000(m)번 적용한 표본 분포도



<그 림 2-1> $\hat{N}_{GIBBS}$에 대한3000번 적용한 표본에 대한 Q-Q 그림



<그림 2-2> $\hat{\beta}_{GIBBS}$의 3000번 적용한 표본에 대한 Q-Q 그림

<표 5>는 각 모형에 대한 N과 β의 추정값으로 부터 계산된 상대오차의 합을 나타낸 것이다. 이 상대오차의 합은 깁스 추출에서 각 1000, 2000, 3000번씩 적용한 것 중 차이가 거의 없으므로 3000번 적용한 값을 가지고 구한 값이다.

<표 5> 상대오차값의 합

상대측도 / 추정방법	n	상대오차의 합 GIBBS	상대오차의 합 MLE
JM	30	0.528001	0.661235
GO(W=0.5)	30	0.263799	0.378341
GO(W=0.6)	30	0.315252	0.456396
GO(W=0.7)	30	0.369177	0.467297
GO(W=0.8)	30	0.423044	0.528715
GO(W=0.9)	30	0.475485	0.567823
SW	30	0.005848	0.191254

6. 결론

일반적인 추론문제에서 적분 계산이 복잡하여 많은 시간과 노력을 필요로 하는 경우를 직면하게 된다. 이러한 점 때문에 복잡한 계산을 보다 쉽게 해결하려는 연구가 다양하게 이루어져 왔다. 이러한 문제는 Gelman과 Rubin(1992)에 의해 구체화된 깁스 알고리즘이 소개되었고 이 기법은 적분대신 적절한 조건부분포로부터 반복표본을 이용한 몬테칼로 적분으로 쉽게 추정할 수 있게 되었다. 본 연구에서는 소프트웨어 고장현상을 수리적으로 모형화를 하기 위한 소프트웨어 신뢰도 성장모형에 최우추정법과 깁스 기법을 적용하여 모수추정과 모형선택에 사용되었다. 레일리분포에서 발생된

자료를 이용한 모의실험 결과를 보면 상대오차의 합이 JM모형 보다는 GO모형이 상대오차의 합이 작게 나왔고, GO모형보다는 SW모형이 상대오차의 합이 작게 나왔다. 따라서 모형선택에 있어서는 우리가 기대했던 것처럼 JM모형보다는 GO모형이, GO모형보다는 SW모형이 소프트웨어 신뢰도 모형에서 더 적합하다고 결론 지을 수 있다.

참고문헌

[1] Aitkin, M.(1991) "Posterior Bayes Factors," *Journal of the Royal Statistical Society B*, pp. 111-142.

[2] Box, G.(1991) "Sampling and Bayes' Inference in Scientific Modeling and Robustness (with discussion)," *Journal of the Royal Statistical Society, Ser. A*, 143, pp. 382-430.

[3] Casella, G. and George, E. I.(1992) "Explaining the Gibbs Sampler," *The American Statistician*, 46, pp. 167-174.

[4] Cox, D. R. and Lewis, P. A.(1966) "*Statistical Analysis of Series of Events*," *Londen: Methuen.*

[5] Gelfand, A. E. and Smith, A. F. M.(1990) ,"Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, pp. 398-409.

[6] Gelman, A. E., and Rubin D.(1992) "Inference from Iterative Simulation Using Multiple Sequences," *Statistical Science*, 7, pp. 457-472.

[7] Geman, S and Geman, D.(1984) "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Distributions, and the Bayesian Restoration of Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, pp. 721-741.

[8] Goel, A. L. and Okumoto, K.(1978) "An analysis of recurrent software failures on a real-timecontrol systems," *Proceedings of the ACM Annual Technical Conference, ACM: Washing D. C.* pp. 496-500.

[9] Jelinski, Z., and Moranda, P. B.(1972) "*Software Reliability Research,* in Statistical Computer Performance Evaluation," ed. W. Freiberger, New York: Academic Press, pp. 465-497.

[10] Kuo, L., and Yang, T. Y.(1995) "Bayesian Computation of Software Reliability," *Journal of Computational and Graphical Statistics*, pp. 65-82.

[11] Langberg, N., and Singpurwalla, N. D.(1985) A Unification of Some Software Reliability Models, *SIAM Journal on Scientific and Statistical Computing*, 6, pp. 781-790.

[12] Lawless, J. F.(1982) "*Statistical Models and Methods for lifetime Data*," New York: Jonn Wiley & Sons.

[13] Musa,J. D. and. Iannino, A., and Okumoto, K.(1987) "*Software Reliability: Measurement, Prediction, Application*," New York: McGraw Hill.

[14] Musa, J. D., and Okumoto, K., A.(1984) "Logarithmic Poisson Execution Time Model for Software Reliability Measurement," *in Proceedings Seventh International Conference on Software Engineering Orlando*, pp. 230-238.

[15] Parzen, E.(1962) "*Stochastic Process*," San Francisco: Holden-Day.

[16] Pesnick, S. I.(1987) "Exterme Values, Regular Variation, and Point*Process*," *Berlin:*Springer-Verlag

[17] Schick, G. J and Wolverton, R. W.(1978) "An Analysis of Competing Software reliability Models," *IEEE Transactions on Software Engineering*, SE-4, 2, pp. 104-120.

[18] Tanner, M. and Wong, W.(1987) "The Calculation of Posterior Distributions by Data Augmentation" (with discussion), *Journal of the American Statistical Association*, 81, pp. 82-86.

[19] "*USER'S MANUAL*(1987) *STAT/LIBRARY FORTRAN Subroutines for statistical analysis*," IMSL, Volume 3, pp. 1050-1054.