Journal of the Korean
Data & Information Science Society
1999, Vol. 10, No. 1, pp. 119 ~ 133

소프트웨어 신뢰모형에 대한 베이지안 접근 [1]

최기헌 [2]

요 약

마코브체인 몬테칼로 방법을 소프트웨어 신뢰모형에 이용하였다. 베이지안 추론에서 조건부 분포를 가지고 사후분포를 결정하는데 있어서의 계산 문제를 고찰하였다. 특히 레코드값을 통계량을 갖고서 혼합과정과 중첩과정에 대하여 깁스샘플링 알고리즘과 메트로폴리스 알고리즘을 활용하여 베이지안 계산과 모형선택을 제시하고 모의실험자료를 이용하여 수치적인 계산을 시행하고 그 결과를 비교하였다.

주제어: 레코드값 통계량, 깁스 샘플링, 메트로폴리스 알고리즘, 모형선택, 비동질적 포아송과정

1. 서론

우리들의 주변에는 복잡한 소프트웨어 시스템들로 둘러 쌓여 있으며, 이러한 시스템의 혜택을 받는 일이 커짐에 따라 소프트웨어 신뢰성의 역할은 점차 커지게 되었다. 소프트웨어 고장들로 인한 컴퓨터 시스템의 고장은 우리 사회에 엄청난 손실을 초래할 수도 있다. 시스템이 고장이 나면 고장 난 원인을 찾아 필요할 경우 새로운 디자인을 개발하거나 새로운 기술을 도입하게 된다. 따라서, 시간이 지남에 따라 신뢰도의 증가가 기대되어진다. 이런 모형을 신뢰도 성장모형(reliability growth model)이라고 한다. 소프트웨어 신뢰성의 예측 문제에 대한 많은 경험기법들은 그들의 정도에 있어서 많은 차이를 보이고 있다. 확실한 것은 모든 환경 하에서 신뢰성 있는 결과를 만족시켜 주는 기법은 아주 어렵다. 소프트웨어 테스팅 단계에서 소프트웨어 오류 수와 고장간격시간에 의해 소프트웨어 고장현상을 수리 모형화를 하면 소프트웨어에 대한 평가를 쉽게 할 수 있으며, 신뢰도 성장모형에 의해 소프트웨어 오류수, 소프트웨어 고장발생간격시간, 소프트웨어 신뢰도 및 고장률등의 신뢰성 평가측도들이 추정되어 예측할 수 있다. 베이지안적 측면에서의 베이즈 이론은 알려져 있는 사실에 대한 주관적 의견을 경험이나 지식을 바탕으로 하여 사전정보를 만든 다

[1]이 논문은 1998년 한국학술진흥재단의 대학교수 해외파견 연구지원에 의하여 연구되었음.
[2](132-714) 서울시 도봉구 쌍문동 419, 덕성여자대학교 통계학과 부교수

음 실험이나 관측을 통하여 얻어진 자료와 결합시켜 사후정보를 추출하는 과정이다. 베이즈 이론에 의한 베이즈 추정량이 갖는 좋은 성질들 때문에 이 방법의 이용으로 많은 연구결과를 낳고 있는 추세이지만 베이즈 추정법에서 사전 확률분포인 수명분포가 복잡하면 적분이 불가능하므로 사후정보의 추출이 불가능해 진다.

본 논문에서는 깁스샘플링(Gibbs sampling)을 이용하여 적분이 풀리지 않는 경우에 근사적인 깁스 추정량을 유도하고 추정량이 갖는 위험도 추이를 분석하고자 한다. 여기서 데이터 증대(data augmentation)를 위한 마코브 체인 몬테칼로(Markov Chain Monte Carlo, MCMC)기법은 사후분포의 특징을 계산하기 위해 제시되었다. 이러한 데이터 증대 접근방법은 마코브 체인을 사용하여 전이측도(transition measure)의 구체화를 용이하게 한다. 몬테칼로 적분 추정기법은 종래의 수치 해석적 적분을 대신할 수 있는 강력한 도구로 각광받고 있다. 적당한 차수까지 연속인 도함수가 존재하는 경우에는 종래의 수치 해석적 적분 알고리즘이 여전히 유용하고 정밀한 답을 제시한다. 그러나 정상적인 함수의 적분문제에서도 몬테칼로 적분추정기법이 수치 해석적 방법에 비하여 효율이 떨어지지 않을 뿐만 아니라 연속성이 성립하지 않는 상황이거나 다차원 적분문제에서는 몬테칼로 적분추정이 아주 유용하게 이용된다.

지금까지 연구들은 강도함수가 하나인 Musa-Okumoto 모형이나 Weibull 모형 같은 단순모형을 개선하는 수정모형이 연구되어 왔다. 그러나 소프트웨어 시스템이 복잡해지면 고장의 원인이 하나의 강도함수에 의해서만 일어나지 않고 여러 원인이 중첩(superposition)되어 발생할 수도 있고, 혼합(mixture)하여 발생할 수도 있다. 이러한 복잡한 시스템에 의한 우도함수를 적분하기가 힘들어지므로 반복표본을 이용하는 몬테칼로 방법인 깁스 알고리즘이 제안되었다. 본 논문에서는 비동질적 포아송 과정(nonhomogeneous Poisson process, NHPP)모형의 중첩과 혼합에 대한 잠재변수들을 이용하여 깁스 알고리즘을 적용하였고, 단순과정, 중첩과정, 혼합과정의 비교를 위해 사후 베이즈요인을 이용하여 모형선택을 하였다. 또한 본 논문에서 는 레코드값 통계량(record value statistics, RVS) 속성을 가진 감마족일 경우의 중첩과정과 혼합과정을 비교하였다. 레코드값에 대한 것은 Arnold, Balakrishnan 그리고 Nagaraja (1998)가 자세히 논하였다. 단순 강도함수가 감마족일 경우에 기본적인 Musa-Okumoto 모형에 Erlang 모형의 중첩과 혼합과정을 비교하고자 한다.

따라서 2절에서는 제안된 혼합과정을 유도하는데 관련된 NHPP와 RVS 그리고 Musa-Okumoto 모형과 연관된 여러 모형의 관계를 설명하였고, 3절에서는 혼합과정에 대한 깁스 샘플링 방법을 설명하였으며, RVS 속성을 가진 Musa-Okumoto 모형과 Erlang 모형의 혼합과정을 제시하였다. 4절에서는 모수에 대한 베이지안 추론과 모형선택에 논하고, 5절에서는 수치적인 예를 들어 중첩과정(Choi and Kim, 1998)과 혼합과정을 비교하였다.

2. 레코드값 통계량 모형

소프트웨어 신뢰도에서 발견된 고장의 수를 모형화하는 것은 비동질적 포아송과정 (non-

homogeneous Poisson process, NHPP)으로 널리 사용하여 왔다 (Musa, Iannino and Okumoto, 1990).

$M(t)$를 시간 $(0,t]$사이에 발견된 고장의 수라 정의하면 $M(t)$는 평균값 함수(mean value function) $m(t) = EM(t)$인 NHPP에 의해 다음과 같이 모형화 될 수 있다.

$$P(M(t) = n) = \frac{m(t)^n}{n!} e^{-m(t)} \tag{1}$$

여기서 $n = 0, 1, 2, \cdots$이다. $m(t)$가 t에 대한 비감소 함수(nondecreasing function) 추세를 가진 미분가능함수이면 강도함수(intensity function), 혹은 고장 발생율(rate of occurrence of failures, ROCOF)은 $\lambda(t) = m'(t)$가 됨이 알려져 있다. 예를 들어 $\lambda(t)$가 상수, 즉 $m(t)$가 선형(linear) 추세이면 동질적 포아송 과정(homogeneous Poisson process)이고, t에 대한 함수형태이면 NHPP가 된다. 즉, 동질적 포아송 과정, Musa-Okumoto 과정, Weibull 과정, 그리고 Cox-Lewis 과정은 각각 강도함수가 상수, 부분(fraction)함수, 멱(power)함수, 대수선형(log-linear)함수이다.

시간 t까지 조사하기 위한 시간 절단(time truncated)모형은 n번 자료를 $x_1, x_2, \cdots, x_n$ 이라고 하면 데이터 집합 D_t는 $\{n, x_1, x_2, \cdots, x_n; t\}$와 같이 구성된다. n번째까지 고장시점이 관찰된 고장 절단 모형일 경우에 데이터 집합 D_{x_n}은 $\{x_1, x_2, \cdots, x_n\}$으로 구성된다. 시간 절단 모형에서의 우도함수는 다음과 같다.

$$L_{NHPP}(\beta | D_t) = \prod_{i=1}^{n} \lambda(x_i) exp(-m(t)) \tag{2}$$

이 우도함수는 Lawless(1982)에 의해서 제시되었다. 고장절단모형은 식(2)의 t을 x_n으로 대치하면 유사한 표현이 된다.

한편, 고장의 시점은 특정밀도 $f(S|\beta)$에 의존하는 서로 독립이고 동일한 확률변량 $S_1, S_2, \cdots$이라 가정하면 레코드값의 열 $\{X_n\}_{n \geq 1}$이 다음과 같이 정의된 모형을 레코드값 통계량 모형이라고 한다(Kuo and Yang, 1996).

$$\begin{aligned} R_{k+1} &= min\{i : S_i > S_{R_k}\}, \quad (k = 1, 2, \cdots), \\ R_1 &= 1, \\ X_n &= S_{R_n}, \quad (n \geq 1). \end{aligned} \tag{3}$$

따라서 레코드값 $X_1, X_2, \cdots$에 따라서 관찰된 고장시점 $x_1, x_2, \cdots$에 의해서 모형화될 수 있다. 이러한 레코드값의 열은 무한히 이루어짐이 Glick(1978)에 의해서 증명되었다. 그러므로 $t \to \infty$이고 고장의 수가 한정되지 않을 때(unbounded)일 때 고장시간에 대한 모형화가 가능하다. 따라서 RVS모형을 NHPP와 관련하여 다음과 같은 관계를 제시하였다.

누적함수 $F(t)$가 실수영역 R^+에서 존재하면 $(0,t]$사이에 구성되는 레코드값은 다음과 같은 평균값을 가진 NHPP의 고장시점이 되는 것을 증명하였다.

$$m(t) = -ln(1 - F(t)) \tag{4}$$

NHPP에서는 $m(t)$가 't에 대한 비감소함수 추세를 가진 미분가능함수이면 강도함수는 $\lambda(t) = m'(t)$가 됨이 알려져 있으므로 강도함수는 위험함수가 되며 다음과 같다.

$$\lambda(t) = m'(t) = \frac{f(t)}{1 - F(t)} \tag{5}$$

따라서 식(2)에 식(4)과 식(5)를 대입하면 다음과 같은 우도함수가 된다.

$$L_{NHPP}(\beta|D_t) = (\prod_{i=1}^{n} f(x_i|\beta)/(1 - F(x_i|\beta))(1 - F(t|\beta)) \tag{6}$$

그러므로 고장절단모형은 식(6)의 t을 x_n으로 대치하면 유사한 우도함수의 표현을 얻을 수 있다.

따라서, RVS 모형에서 식(6)과 관련하여 강도함수 $\lambda(t) = \alpha/(t + \beta)$인 경우를 Musa-Okumoto 모형이라 하고 $\lambda(t) = \alpha\beta t^{\alpha-1}$와 $\lambda(t) = exp(\alpha + \beta t)$인 경우를 각각 Duane, Cox-Lewis 모형이라고 한다. Musal-Okumoto 모형과 연관하여 NHPP 중에서 $\lambda(t)$가 이중지수, 곰퍼르츠(Gompertz), 랄리(Rayleigh), 감마, 그리고 와이블(Weibull)일 때 이중지수 NHPP, 곰퍼르츠 NHPP, 랄리 NHPP, 감마 NHPP, 그리고 와이블 NHPP 모형으로 제시할 수 있다.

3. 혼합과정에 대한 깁스추출법

깁스샘플링은 적분이 복잡하거나 난해한 경우에 반복표본을 이용하여 정보를 얻는 수치 해석적 방법이다. 이 기법은 MCMC의 기법중 하나이며 마코브 연쇄의 전이분포는 여러 개의 조건부밀도로서 이루어진다. 이러한 조건부분포의 반복표본추출을 이용하여 깁스샘플링을 시행한다. 결국 마코브 연쇄의 정상분포가 원하는 사후분포가 된다. 마코브 연쇄의 다중열(multiple sequences)을 얻기위하여 독립 초기점을 가진 연쇄를 반복한다. 이러한 기법은 Tanner와 Wong(1987), Gelfand와 Smith(1990), Casella와 George(1992) 등 많은 학자들에 의해서 제시되었다.

한편, 소프트웨어 시스템이 복잡해지면 고장의 원인이 하나의 강도함수에 의해서만 일어 나지 않고 여러 원인이 중첩되어 발생할 수 있다. 이런 경우에 강도함수와 평균값 함수를 알고 있으면 모형화가 가능하다.

$M_j(t)$을 시간 $(0,t]$사이에서 강도함수가 $\lambda_j(t|\beta_j)$을 가진 j번째 요소로부터 고장이 발생되는 NHPP라고 표현하고 $\lambda_j(t|\beta_j)$은 미지 모수 β_j을 가진 알고 있는 값이고 j번째 요소에 의해 발생된 고장의 수 $M_j(t)(j = 1, 2, \cdots, J)$는 독립이라고 가정하면 $M(t) = \sum_{j=1}^{J} M_j(t)$이다. 즉, 중첩과정에서 $(0,t]$의 구간에서 발생된 고장의 총수는 강도함수와 평균값함수가 각각 다음과 같은 NHPP가 된다(Cinlar, 1975). 단, $\underline{\beta} = \{\beta_1, \beta_2, \cdots, \beta_J\}$ 이다.

$$\begin{aligned} \lambda(x_i|\underline{\beta}) &= \lambda_1(x_i|\beta_1) + \lambda_2(x_i|\beta_2) + \cdots + \lambda_J(x_i|\beta_J) \\ m(t|\underline{\beta}) &= m_1(t|\beta_1) + m_2(t|\beta_2) + \cdots + m_J(t|\beta_J) \end{aligned} \tag{7}$$

따라서 식(2)와 관련하여 시간절단모형의 RVS 속성을 가진 중첩과정에 대한 우도함수는 다음과 같다.

$$L_{NHPP}(\underline{\beta}|D_t) = \left(\prod_{i=1}^{n}(\lambda(x_i|\underline{\beta})\right)\exp(-m(t|\underline{\beta})) \tag{8}$$

중첩과정에서 $\underline{\beta}$의 사후결합분포는 식(3.2)의 우도함수와 사전분포를 베이즈정리에 의해 다음과 같이 나타낼 수 있다.

$$\begin{aligned} f(\underline{\beta}|D_t) \quad \propto \quad & \left(\prod_{i=1}^{n}[\lambda_1(x_i|\beta_1) + \lambda_2(x_i|\beta_2) + \cdots + \lambda_J(x_i|\beta_J)]\right) \\ & \cdot \exp(-[m_1(t|\beta_1) + \cdots + m_J(t|\beta_J)]) \times \prod_{j=1}^{J}\pi_j(\beta_j) \end{aligned} \tag{9}$$

단, $\underline{\beta}$는 모수벡터이고, π_j는 β_j에 대한 사전분포이고, π_j와 β_j은 독립이라고 가정하자. 식(9)에 있는

$$\prod_{i=1}^{n}[\lambda_1(x_i|\beta_1) + \lambda_2(x_i|\beta_2) + \cdots + \lambda_J(x_i|\beta_J)] \tag{10}$$

의 표현때문에 D_t의 조건하에서 $\underline{\beta}$의 조건부분포의 형태를 구체화되기 어렵다. 이러한 어려움을 해소하기 위하여 잠재변수를 사용한다. I를 잠재변수들의 모임이라고 표시하고 이 I를 사용하여 $\underline{\beta}$의 사후분포를 계산한다. 각 $i(i = 1, 2, \cdots, n)$에 대해 잠재변수 $I_i = (I_{i1}, \cdots, I_{iJ})$을 설정하자. j번째 요소에 의해서 i번째 고장의 원인이 발생했을 때를 $I_{ij} = 1$이라고 하자. 그 외에는 $I_{ij} = 0$이라 하면, $i = 1, 2, \cdots, n$, $\sum_{j=1}^{J} I_{ij} = 1$의 조건을 만족한다.

$\underline{\beta}$와 D_t의 조건에서의 I_i의 조건부밀도는 다항분포(multinomial distribution, MN)가 되는데 이 분포는 모수가 1이고, 셀확률 $(p_{i1}, \cdots, p_{iJ})$을 가진다. 단,

$$p_{ij} = \lambda_j(x_i|\beta_j)/[\lambda_1(x_i|\beta_1) + \lambda_2(x_i|\beta_2) + \cdots + \lambda_J(x_i|\beta_J)]. \tag{11}$$

이다. $I = (I_1, \cdots, I_n)^T$라고 표시하면 i번째 고장발생의 요인은 $n \times J$ 행렬이 된다. 여기서 시뮬레이트된 $I_1, \cdots, I_n$은 독립이고 다음과 같다.

$$I_i \sim MN(1, (p_{i1}, \cdots, p_{iJ})), \quad i = 1, \cdots, n. \tag{12}$$

따라서 I와 D_t의 조건하에서 $\underline{\beta}$의 사후결합밀도는 다음과 같이 표현할 수 있다.

$$\begin{aligned} f(\underline{\beta}|I,D_t) &\propto L(\underline{\beta}|D_t)\times\prod_{i=1}^{n}P(I_i|\underline{\beta},D_t)\times\prod_{j=1}^{J}\pi_j(\beta_j) \\ &\propto \prod_{j=1}^{J}\prod_{i:I_{ij=1}}^{n}\lambda_j(x_i|\beta_j)\times\prod_{j=1}^{J}\exp(-m_j(t))\times\prod_{j=1}^{J}\pi_j(\beta_j) \end{aligned} \tag{13}$$

조건부밀도는 I와 D_t의 조건하에서 $\underline{\beta}$의 독립 사후결합밀도들로 구성된 결합밀도가 된다. 그러므로 깁스샘플링은 반복을 통해서 $P(I|\underline{\beta},D_t)$로부터 I을 추출하고 $f(\underline{\beta}|I,D_t)$로부터 $\underline{\beta}$를 추출하는 반복형식을 통해 독립적으로 최종 표본을 얻을 수 있다.

그러므로 혼합과정을 $m(t)=p_1m_1(t)+p_2m_2(t)$라고 제시하고 NHPP는 $m(t)$가 t에 대한 비감소함수 추세를 가진 미분가능함수이라면 강도함수는 $\lambda(t)=m'(t)$가 됨이 알려져 있으므로

$$\lambda(t)=p_1\lambda_1(t)+p_2\lambda_2(t) \tag{14}$$

라고 정의할 수 있다. 따라서 우도함수는 식(2)와 관련하여 다음과 같이 유도할 수 있다.

$$L_{NHPP}(\underline{\beta}|D_t)=\Big[\prod_{i=1}^{n}p_1\lambda_1(x_i)+p_2\lambda_2(x_i)]\cdot exp[-p_1m_1(t)-p_2m_2(t)\Big] \tag{15}$$

여기서, $p_1+p_2=1$이고, p_1은 0 혹은 1를 가진다.

혼합과정에서 $\underline{\beta}$의 사후결합분포는 식(15)의 우도함수와 사전분포를 베이즈정리에 의해 다음과 같이 나타낼 수 있다.

$$\begin{aligned} f(\underline{\beta}|D_t) \propto &\Big(\prod_{i=1}^{n}\Big[p_1\lambda_1(x_i|\beta_1)+(1-p_1)\lambda_2(x_i|\beta_2)\Big]\Big) \\ &\cdot\exp\left(-[p_1m_1(t|\beta_1)+(1-p_1)m_2(t|\beta_2)]\right)\prod_{j=1}^{2}\pi_j(\beta_j)\cdot\pi_3(p_1) \end{aligned} \tag{16}$$

여기서, $\underline{\beta}=\{\beta_1,\beta_2,p_1\}$는 모수벡터이고, π_j는 $\beta_j(j=1,2)$에 대한 사전분포이고, π_3는 p_1에 대한 사전분포이고, $\underline{\beta}$와 $\pi_j(j=1,2,3)$은 독립이라고 가정하자.

식(16)에 있는 $\prod_{i=1}^{n}[p_1\lambda_1(x_i|\beta_1)+(1-p_1)\lambda_2(x_i|\beta_2)]$ 의 표현때문에 D_t의 조건하에서 $\underline{\beta}$의 조건부분포의 형태를 구체화되기 어렵다. 이러한 어려움을 해소하기 위하여 잠재변수를 사용한다. Z를 잠재변수들의 집합이라고 표시하고 이 Z을 사용하여 $\underline{\beta}$의 사후분포를 계산한다. 각 $i(i=1,\cdots,n)$에 대해 잠재변수 $Z_i=(Z_{i1},Z_{i2})$을 설정하자. $j(=1,2)$번째 요소에 의해서 i번째 고장의 원인이 발생했을 때를 $Z_{ij}=1$이라고 하자. 그 외에는 $Z_{ij}=0$이라 하면, $i=1,\cdots,n$, $\sum_{j=1}^{2}Z_{ij}=1$의 조건을 만족한다.

$\underline{\beta}$와 D_t의 조건에서의 Z_i의 조건부밀도함수는 베르누이 시행이 되는데 이 분포는 모수가 1이고, 셀확률 p_{i1}을 가진다. 여기서 $p_{i1} = p_1\lambda_j(x_i)/[p_1\lambda_1(x_i) + (1-p_1)\lambda_2(x_i)]$ 이다. 따라서 Z와 D_t의 조건하에서 $\underline{\beta}$의 사후결합밀도는 다음과 같이 표현할 수 있다.

$$\begin{aligned} & f(\underline{\beta}|Z, D_t) \\ & \propto L(\underline{\beta}|D_t) \times \prod_{i=1}^{n} P(Z_i|\underline{\beta}, D_t) \times \prod_{j=1}^{2} \pi_j(\beta_j) \cdot \pi_3(p_1) \\ & \propto \Big[\prod_{j=1}^{n}(p_1\lambda_1(x_i|\beta_1)Z_i + (1-p_1)\lambda_2(x_i|\beta_2)(1-Z_i))\Big] \\ & \quad \times \prod_{j=1}^{2} \exp(-m_j(t)) \times \prod_{j=1}^{2} \pi_j(\beta_j) \cdot \pi_3(p_1) \end{aligned} \tag{17}$$

조건부밀도는 Z와 D_t의 조건하에서 $\underline{\beta}$의 독립 사후결합밀도함수들로 구성된 결합밀도함수가 된다. 그러므로 깁스샘플링은 반복을 통해서 $P(Z|\underline{\beta}, D_t)$로부터 Z을 추출하고 $f(\underline{\beta}|Z, D_t)$로부터 $\underline{\beta}$를 추출하는 반복형식을 통해 독립적으로 최종표본을 얻을 수 있다.

본 절에서는 RVS속성을 가진 감마족 $F(t)$일 경우의 혼합과정을 비교하고자 한다. 단순강도함수가 감마족일 경우에 기본적인 Musa–Okumoto 모형과 Erlang(2) 모형의 혼합과정을 유도하기 위한 강도함수는 다음과 같다.

$$\lambda_1(t) = \frac{\alpha_1}{\alpha_2 + t}, \quad \lambda_2(t) = \frac{\alpha_3^2 t}{\beta_3 t + 1} \tag{18}$$

여기서, $\alpha_i > 0 (i = 1, 2, 3)$이고, 사후결합분포를 유도하기 위한 혼합과정 우도함수는 식(15)을 이용하면 다음과 같다.

$$\begin{aligned} L(\underline{\beta}|D_t) = \quad & \prod_{i=1}^{n}\left(p_1 \cdot \frac{\alpha_1}{\alpha_2 + x_i} + p_2 \cdot \frac{\alpha_3^2 x_i}{\alpha_3 x_i + 1}\right) \\ & \cdot \exp\left(-p_1\alpha_1 \ln(1 + \frac{t}{\alpha_2}) - p_2\alpha_3 t + p_2 \ln(1 + \alpha_3 t)\right) \end{aligned} \tag{19}$$

여기서, $\underline{\beta} = \{\alpha_1, \alpha_2, \alpha_3, p_1\}$이다. 그리고 $p_1 + p_2 = 1$이고, p_1은 0 혹은 1을 가진다. $\Gamma(a, b)$은 평균이 a/b인 감마분포를 표시하고 독립인 사전분포는 다음과 같다.

$$\alpha_1 \sim \Gamma(a_1, b_1); \quad \alpha_2 \sim \pi_1(\alpha_2); \quad \alpha_3 \sim \pi_2(\alpha_3); \quad p_1 \sim Beta(a_2, b_2) \tag{20}$$

여기서, $\pi_1(\alpha_2)$와 $\pi_2(\alpha_3)$는 각각 $\alpha_2 > 0$, $\alpha_3 > 0$에 대한 임의의 사전분포이고, $\underline{\beta}$와 D_t의 조건에서의 Z_i의 조건부밀도함수는 $i (i = 1, 2, \cdots, n)$에 대한 p_{i1}의 모수를 가진 베르누이분 포로부터 발생된다. 여기서

$$p_{i1} = \frac{p_1\alpha_1}{\alpha_2 + x_i} \Big/ \left(\frac{p_1\alpha_1}{\alpha_2 + x_i} + \frac{p_2\alpha_3^2 x_i}{\alpha_3 x_i + 1} \right) \tag{21}$$

이고 $Z_{i2} = 1 - Z_{i1}$이 된다. 그리고 $p_1 + p_2 = 1$이며 p_1은 0 혹은 1를 가진다. Z와 D_t의 조건하에서 $\underline{\beta}$의 사후결합밀도는 식(19)에 의해서 다음과 같다.

$$\begin{aligned} f(\underline{\beta}|Z, D_t) \quad \propto \quad & \prod_{j=1}^{n} \left[\frac{p_1\alpha_1}{\alpha_2 x_i} Z_i + \frac{p_2\alpha_3^2 x_i}{\alpha_3 x_i + 1}(1 - Z_i) \right] \\ & \cdot \exp[-p_1\alpha_1 ln(1 + \frac{t}{\alpha_2}) Z_i - p_2\alpha_3 t(1 - Z_i) \\ & + p_2 \ln(1 + \alpha_3 t)(1 - Z_i)] \\ & \cdot \frac{b_1^{a_1}\alpha_1^{a_1} e^{-b_1\alpha_1}}{\Gamma(a_1)} \cdot \pi_1(\alpha_2) \cdot \pi_2(\alpha_3) \cdot p_1^{a_1 - 1} p_2^{b_2 - 1} \end{aligned} \tag{22}$$

식(22)을 베이즈정리와 장애모수의 개념을 사용하여 깁스샘플링 알고리즘을 이용하기 위한 모든 조건부 밀도함수들은 다음과 같은 4개의 조건식이 필요하다.

$$(1) \quad p(p_1|\alpha_1, \alpha_2, \alpha_3, Z, D_t) \sim \Gamma(\sum z_i + a_2 - 1,\ \alpha_1 \ln(1 + \frac{1}{\alpha_2}) \sum z_i); \tag{23}$$

$$(2) \quad p(\alpha_1|p_1, \alpha_2, \alpha_3, Z, D_t) \sim \Gamma(a_1 + \sum z_i,\ p_1 \ln(1 + \frac{t}{\alpha_2}) \sum z_i + b_1); \tag{24}$$

$$\begin{aligned} (3) \quad & p(\alpha_2|p_1, \alpha_1, \alpha_3, Z, D_t) \\ & \propto \frac{1}{\prod_{i=1}^{n} z_i(x_i + \alpha_2)} \cdot \exp[-p_1\alpha_1 \ln(1 + \frac{t}{\alpha_2}) \sum z_i] \cdot \pi_1(\alpha_2); \end{aligned} \tag{25}$$

$$\begin{aligned} (4) \quad & p(\alpha_3|p_1, \alpha_1, \alpha_2, Z, D_t) \propto \left(\prod_{i=1}^{n} (1 - z_i) \frac{\alpha_3^2 x_i}{\alpha_3 x_i + 1} \right) \\ & \cdot \exp[-(1 - p_1)\alpha_3 t \sum (1 - z_i) + (1 - p_1) \ln(1 + \alpha_3 t) \sum (1 - z_i)] \cdot \pi_2(\alpha_3); \end{aligned} \tag{26}$$

깁스샘플링 알고리즘은 각 모수의 사전분포의 초기값을 대입하여 위 식의 $p(p_1|\alpha_1, \alpha_2, \alpha_3, Z, D_t)$의 분포로부터 p_1을 추출하고 갱신된(추출된) p_1과 나머지 초기값을 대입하여 $p(\alpha_1|p_1, \alpha_2, \alpha_3, Z, D_t)$의 분포로부터 α_1를 추출한다. 그리고 갱신된 p_1와 α_1, 그리고 나머지 초기값을 대입하여 $p(\alpha_2|p_1, \alpha_1, \alpha_3, Z, D_t)$ 로부터 α_2를 추출하고 갱신된 p_1와 α_1, α_2, 그리고 나머지 초기값을 대입하여 $p(\alpha_3|p_1, \alpha_1, \alpha_2, Z, D_t)$로부터 α_3를 추출하는 반복형식을 통해 독립적으로 최종표본을 얻을 수 있다. 반복과정에서 p_1, α_1는 감마분포에서 표본이 발생하여 사용할 수 있지만 α_2와 α_3는 일정한 분포를 알지 못하므로 메트로폴리스 알고리즘에 의해서 발생된다. 즉 깁스단계에서 메트로폴리스와 깁스샘플링을 혼합하여 사용하는 방법은 다음과 같다. 예를 들어 깁스단계에서 α_2를 추출 하는 메트로폴리스 알고리즘은 다음과 같은 기법을 사용하여 이루어진다.

식(25)의 오른쪽 식의 목적분포를 간편하게 $f(\alpha_2)$로 표기하고, 여기서 정규화 상수는 필요하지 않다. p_1과 α_1, α_3는 앞 단계에서 추출된 값이고, $\pi_1(\alpha_2)$의 분포는 $1/\alpha_2(>0)$이라 가정되었고 추이커널은 분산을 보다 크게 한 거의 대칭을 이루는 감마분포 $\Gamma(1, 10^{-4})$에서 β^{ω}을 발생하고 균등분포 $(0,1]$에서 확률변량을 ω라 하고, $\log f(\omega) \leq \log f(\beta^{\omega}) - \log f(\alpha_2)$을 만족하면 β^{ω}을 α_2으로 간주되고 만족하지 않으면 α_2을 β^{ω}으로 대치되면서 만족할 때까지 계속 반복된다. 다시 갱신된 최종 추정치를 새로운 초기값으로 대치되면서 계속 반복 적용된다.

4. 베이지안 추론과 모형선택

베이지안 예측은 고장절단상황에서 고려하는 것이 일반적이기 때문에 n번째 고장까지 조사하여 t를 x_n으로 대치함으로서 해결할 수 있다(Kuo와 Yang, 1996). 만일 시점 x_i로부터 x만큼 떨어진 예측 패턴은 미래생존함수(future survival function)의 통해서 해결할 수 있고 다음과 같은 방법을 통하여 추론이 이루어진다(Cinlar 1975, p.97).

$$\begin{aligned} E(S(x)|D_{x_n}) &= E[E(P(X_{n+1} > x_n + x)|\underline{\beta}, D_{x_n})|D_{x_n}] \\ &= E[\exp(-m(x_n + x|\underline{\beta}) + m(x_n\underline{\beta}))|D_{x_n}] \\ &= \int \cdots \int \exp\{-m(x_n + x|\underline{\beta})\} f(\beta|D_{x_n}) d\underline{\beta} \end{aligned} \tag{27}$$

식(27)은 Gelman & Rubin(1992)이 제시한 깁스추출방법을 사용하여 다음과 같이 적용을 하였다.

$$\hat{S}(x|D_{x_n}) = \frac{2}{RI} \sum_{r=1}^{R} \sum_{i=\frac{I}{2}+1}^{I} \exp\left(-m(x_n + x|\underline{\beta}^{(i,r)}) + m(x_n|\underline{\beta}^{(i,r)})\right) \tag{28}$$

식(28)에서 $\underline{\beta}^{(i,r)}$은 깁스 반복알고리즘을 사용하여 i번 반복 후 r번 적용을 통해 $\underline{\beta}$를 발생시킨 깁스 표본추출을 의미하고, $i = I/2+1, \cdots, I, r = 1, \cdots, R$이며 충분히 큰 R과 짝수인 I를 대입한 깁스 표본추출을 사용하면 식(28)의 추정치를 유도할 수 있다. 예를 들어 앞 절에서 제시한 Musa-Okumoto와 Erlang(2) 모형의 혼합과정의 경우에는 다음과 같이 계산할 수 있다.

$$\begin{aligned} \hat{S}(x|D_{x_n}) = \frac{2}{IR} \sum_{r=1}^{R} \sum_{i=\frac{I}{2}+1}^{I} & \exp[-p_1^{(i,r)}\alpha_1^{(i,r)} ln\left(1 + \frac{x_n + x}{\alpha_2^{(i,r)}}\right) - (1 - p_1^{(i,r)})\alpha_3^{(i,r)}(x + x_n) \\ & + (1 - p_1^{(i,r)}) \ln(1 + \alpha_3^{(i,r)}(x + x_n)) + p_1^{(i,r)}\alpha_1^{(i,r)} \ln\left(1 + \frac{x_n}{\alpha_2^{(i,r)}}\right) \\ & + (1 - p_1^{(i,r)})\alpha_3^{(i,r)} x_n - (1 - p_1^{(i,r)}) ln(1 + \alpha_3^{(i,r)} x_n)]. \end{aligned} \tag{29}$$

NHPP는 여러 가지 모형이 존재할 수 있으며 어떤 모형이 적당한가는 모형의 정도와 모형선택으로 평가할 수있다. NHPP는 여러 과정들이 존재할 수 있으므로 최적과정 선택은 Dawid(1984)가 제안한 PCPO(Prequencial Conditional Probability Ordinate)를 사용하여 해결할 수 있다. 따라서 미래시점 x_{i+1}에 대한 PCPO는 $(x_1, \cdots, x_n)$이 주어진 미래의 관찰시점 X_{i+1}의 조건부 밀도함수 $c_{i+1} = p(x_{i+1}|D_{x_i}), i \geq 1$에 의해 정의된다. 이 PCPO는 과거의 데이터가 주어진 상태에서 X_{i+1}의 값을 예측하기 때문에 과정선택을 하는데도 적절한 도구가 된다. 계열 $\{X_i\}_{i\geq 1}$이 주어지면 PCPO는 다음과 같은 식을 통하여 계산된다.

$$\begin{aligned} p(X_{i+1}|D_{x_i}) &= \int \cdots \int p(X_{i+1}|\underline{\beta}, D_{x_i}) p(\underline{\beta}|D_{x_i}) d\underline{\beta} \\ &= \int \cdots \int \lambda(X_{i+1}) \exp(-m(X_{i+1}) + m(x_i)) \times p(\underline{\beta}|D_{x_i}) d\underline{\beta} \end{aligned} \tag{30}$$

식(30)을 이용하여 앞 절에서 제시한 Musa-Okumoto와 Erlang(2)의 혼합과정의 경우에는 다음과 같이 계산 할 수 있고

$$\begin{aligned} p(x_{i+1}|x_i) = \frac{2}{IR} \sum_{r=1}^{R} \sum_{i=\frac{I}{2}+1}^{I} [& \left(\frac{p_1^{(i,r)} \alpha_1^{(i,r)}}{\alpha_2^{(i,r)} + x_{i+1}} + \frac{(1-p_1^{(i,r)}) \alpha_3^{(i,r)} x_{i+1}}{\alpha_3^{(i,r)} x_{i+1} + 1} \right) \\ & \cdot \exp(-p_1^{(i,r)} \alpha_1^{(i,r)} \ln(1 + \frac{x_{i+1}}{\alpha_2^{(i,r)}}) - (1-p_1^{(i,r)}) \alpha_3^{(i,r)} x_{i+1} \\ & + (1-p_1^{(i,r)}) \ln(1 + \alpha_3^{(i,r)} x_{i+1}) + \alpha_1^{(i,r)} \ln(1 + \frac{x_i}{\alpha_2^{(i,r)}}) \\ & + (1-p_1^{(i,r)}) \alpha_3^{(i,r)} x_i - (1-p_1^{(i,r)}) \ln(1 + \alpha_3^{(i,r)} x_i))] \end{aligned} \tag{31}$$

모형선택을 위하여 사전자료를 사용하여 c_{i+1}을 각 시점에 대하여 계산하여 곱하면 베이지안 예측우도기준은 예측밀도가 되고 다음과 같이 표현할 수 있다(Shina와 Dey, 1997).

$$Pd = \prod_{i=1}^{n} c_{i+1} = p(x_1, x_2, \cdots, x_n) \tag{32}$$

이 예측밀도를 최대화하는 모형이 더 좋은 모형으로 간주된다(Dawid, 1984).

5. 수치적인 예

모수추정과 모형선택을 위해 다음과 같은 RVS의 강도함수를 가진 NHPP를 가진 과정을 제시하고 비교하고자 한다. 즉, 부분함수 $\lambda_1(t) = \alpha_1/(\alpha_2 + t), \lambda_2(t) = \alpha_3^2 t/(\alpha_3 t + 1), \lambda_1(t) + \lambda_2(t)$, 그리고 혼합과정 $p_1\lambda_1(t) + p_2\lambda_2(t)$. 여기서 $p_1 + p_2 = 1$이고 $\alpha_i > 0 (i = 1, 2, 3)$이다. λ_1은 Musal-Okumoto모형의 강도함수이고, λ_2는 랄리모형의 특수한 강도함수를 의미한

다. 데이터들은 일반적으로 NHPP에 대한 자료생성에 이용되는 Lewis와 Shedler(1979)에 의해 제시된 thinning 알고리즘을 이용하여 IMSL RANPP($t = 40$) 루틴에 의하여 모수 $p = 0.5, \alpha_1 = 0.05, \alpha_2 = 50, \alpha_3 = 0.05$를 가정하여 시뮬레이트되었고 이 루틴에 의한 자료들은 다음과 같다(IMSL 메뉴얼 p.1051, 1987).

$$0.24, 5.36, 7.05, 9.34, 10.3, 10.7, 10.8, 11.8, 13.1, 15.0,$$

$$15.2, 17.1, 17.4, 18.1, 18.2, 20.4, 21.8, 22.7, 23.8, 24.5,$$

$$25.7, 30.1, 31.2, 32.6, 32.7, 33.4, 33.8, 34.8, 37.5, 37.6$$

깁스샘플링을 시행하기 위하여 각 모수들에 대한 사전분포는 상대적으로 확산분포(diffuse priors)를 제시하여 경험적 결과를 유도하였다. 사전분포는 $p = beta(3,1), \alpha_1 = \Gamma(1, 0.0001)$을 선택 이용하였고, α_2, α_3에 대한 비정보 사전밀도(noninformative prior density)는 각각 $\pi_1(\alpha_2) = 1/\alpha_2, \alpha > 0, \pi_2(\alpha_3) = 1/\alpha_3, \alpha > 0$을 가정하였다. 이 과정들에 대한 확산 사전분포는 표 5.1에 요약되었다. 반복이 얼마나 필요한지를 결정하기 위하여 Gelman과 Rubin(1992)이 제시한 방법을 사용하여 깁스샘플링의 수렴성을 고려하여 제안된 사전분포를 가지고 FORTRAN IMSL언어를 이용하여 충분한 1000번의 반복을 시행하였다. 모든 경험적 결과들의 깁스추출은 1000번의 반복 중 50개의 분리된 깁스체인을 발생하였다.

표 5.2는 모수에 대한 사후평균을 요약하였고, 혼합과정에서 $\{\hat{p}, \hat{\alpha}_1, \hat{\alpha}_2, \hat{\alpha}_3\}$의 베이지안 추정치는 다른 모형들과 비교되었다. 또 이 과정들에 대한 PCPO의 값(Pd)을 요약하였다. 그 결과 예측된 것처럼 혼합과정 Pd의 값이 비교적 크므로 혼합과정이 중첩과정과 단순과정보다 효율적인 과정으로 간주할 수 있다. 그림 5.1은 시간의 흐름에 따라 비증가함수의 형태를 가진 미래생존함수 $\hat{S}(x|D_{x_n})$을 나타낸 것이고, 그림 5.2는 시간에 따라 혼합과정에 대한 진(true)의 강도함수와 베이즈 추정치에 의한 강도함수를 비교하여 본 결과 거의 밀접한 추세를 보이고 있으므로 수렴하고 있다고 간주할 수 있다.

표 5.1 사전분포

$\lambda(t)$	λ_1	λ_2	$\lambda_1 + \lambda_2$	$p_1\lambda_1 + p_2\lambda_2$
사전분포	$\alpha_1 = \Gamma(1, 10^{-4})$		$\alpha_1 = \Gamma(1, 10^{-4})$	$p_1 = beta(3,1)$ $\alpha_1 = \Gamma(1, 10^{-4})$
	$\alpha_2 = \Gamma(1, 10^{-2})$		$\alpha_2 = \Gamma(1, 10^{-2})$	$\alpha_2 = \Gamma(1, 10^{-2})$
	$\alpha_3 = \Gamma(1, 10^{-2})$	$\alpha_3 = \Gamma(1, 10^{-2})$	$\alpha_3 = \Gamma(1, 10^{-2})$	$\alpha_3 = \Gamma(1, 10^{-2})$

표 5.2 사후평균의 추정값과 PCPO의 값

추세	λ_1	λ_2	$\lambda_1 + \lambda_2$	$p_1\lambda_1 + p_2\lambda_2$
사후평균의 추정값	$\hat{\alpha}_1 = 0.04014$ $\hat{\alpha}_2 = 48.14271$	$\hat{\alpha}_3 = 0.03176$	$\hat{\alpha}_1 = 0.039542$ $\hat{\alpha}_2 = 46.76281$ $\hat{\alpha}_3 = 0.01021$	$\hat{p} = 0.42138$ $\hat{\alpha}_1 = 0.042262$ $\hat{\alpha}_2 = 45.26719$ $\hat{\alpha}_3 = 0.02182$
$log(Pd)$	-41.67535	-40.65786	-37.23981	-35.25671

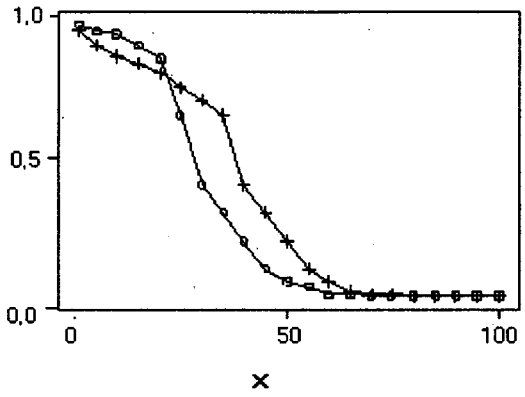

그림 1 시간(x)에 대한 중첩과정(+)과 혼합과정(○)에 대한 예측생존함수 $\hat{S}(x|D_{x_n})$

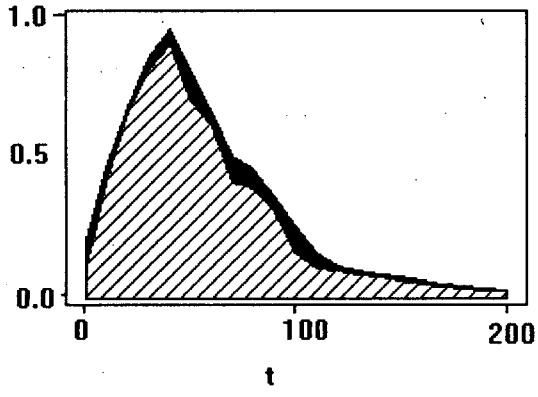

그림 2 시간(t)에 대한 진(true)의 강도함수(■)와
베이즈 추정치에 의한 강도함수(///)의 비교

참 고 문 헌

1. Arnold, B. C., Balakrishnan, N. and Nagaraja, H. (1998). *Records*, New York: Jonn Wiley and Sons.

2. Casella, G. and George, E. I. (1992). Explaining the Gibbs Sampler, *The American Statistician*, vol.46, 167-174.

3. Cinlar, E.(1975). *Introduction To Stochastic Process*, New Jersey : Prentice-Hall.

4. Choi, K. and Kim, H.(1998). Bayesian Computation for Superposition of MUSA–OKUMOTO and ERLANG(2) Processes, *The Korean Journal of Applied Statistics*, vol.11, No. 2, 377-387.

5. Dawid, A. P.(1984). Statistical Theory : The Prequential Approach, *Journal of the Royal Statistical Society*, Ser. A, vol.147, 278-292.

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

7. Glick, N.(1978). Breaking Records and Breaking Boards, *American Mathematical Monthly*, vol. 85, 2-26.

8. Kuo, L. and Yang, T. Y.(1995). Bayesian Computation of Software Reliability, *Journal of Computational and Grphical Statistics*, 65-82. 781-790.

9. Kuo, L. and Yang, T. (1996). Bayesian computation of Software Reliability, *Journal of the American Statistical Association*, Vol.91, pp.763-773.

10. Lewis, P. A. W. and Shedler, G. S. (1979). Simulation of Nonhomogeneous Poisson Process by Thinning, *Naval Research Logistic Quarterly*, 26, 403-413.

11. Lawless, J. F.(1982). *Statistical Models and Methods for lifetime Data*, New York: Jonn Wiley and Sons.

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

13. Shiha, D. and Dey, D. K. (1987). Semiparametric Bayesian Analysis of Survival Data, *Journal of the American Statistical Association*, 81, 82-86.

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

15. *USER'S MANUAL STAT/LIBRARY FORTRAN Subroutines for statistical analysis*, IMSL, Volume 3, 1050-1054.

Bayesian Approach for Software Reliability Models

Kiheon Choi [3]

Abstract

A Markov Chain Monte Carlo method is developed to compute the software reliability model. We consider computation problem for determining of posterior distibution in Bayseian inference. Metropolis algorithms along with Gibbs sampling are proposed to preform the Bayesian inference of the Mixed model with record value statistics. For model determiniation, we explored the prequential conditional predictive ordinate criterion that selects the best model with the largest posterior likelihood among models using all possible subsets of the component intensity functions. To relax the monotonic intensity function assumptions. A numerical example with simulated data set is given.

Key Words and Phrases: record value statistics, gibbs sampling, Metropolis algorithm, model selection, nonhomogeneous Poisson process.

[3]Department of Statistics, Duksung Women's University, 132-714, Seoul, Korea