

⊗ 연구논문

비동질적 포아송과정을 사용한 소프트웨어 베이지안 신뢰성
분석에 관한 연구

Bayesian Analysis for Nonhomogeneous Poisson Process
Software Reliability

김 회 철¹⁾
Kim, Hee Cheul
이 동 철²⁾
Lee, Dong Cheol

Abstract

Bayesian approach using nonhomogeneous Poisson process is considered for modelling software reliability problem. The usefulness of the iterative sampling-based method increases greatly as the dimension of a problem increases. Maximum likelihood estimator and Gibbs estimator are derived. Model selection based on a predictive likelihood is studied. A numerical example is given.

1. 서론

신뢰도 이론은 소프트웨어 시스템 전체가 규정된 환경조건하에서 의도하는 기간동안에 요구된 기능을 만족스럽게 수행할 수 있는 확률을 예측하고 증대시키기 위한 실제적인 도구가 된다. 제품이 의도된 기간동안 고장 없이 제 기능을 발휘할 수 있는 확률을 보다 정확한 방법으로 추정하는 연구는 이 분야의 관심사항이 된다.

신뢰도의 정량적인 값을 얻는데 있어 기존의 방법은 최우추정법을 많이 사용하였다. 그러나 새로운 자료가 얻어지면 그 자료를 이제까지 얻었던 자료와 결합시켜 새로운 결론에 도달하려는 이론이 베이즈 추정법이다. 즉, 베이즈 이론은 알려져 있는 사실에 대한 주관적 의견을 경험이나 지식을 바탕으로 하여 사전정보를 만든 다음 실험을 통하여 얻어진 자료와 결합시켜 사후정보를 추출하는 과정이다. 그러나 베이즈 추정법에서 사전확률분포인 수명분포가 복잡하면 적분이 불가능해지기 때문에 사후정보의 추출이 불가능해진다.

본 연구에서는 최우추정법과 적분이 난해한 경우에 깁스 샘플링(Gibbs sampling)을 이용하여 근사적 깁스추정량 유도하여 비교하고 그 특징을 분석하고자 한다.

우리가 가능한 소프트웨어 시스템에서 $(0, t]$ 사이에 발견된 고장의 수 $M(t)$ 을 모형화 하는데 비동질적 포아송과정(nonhomogeneous poisson process, NHPP)이 널리 사용되어 왔다(Musa, Iannino & Okumoto, 1987). 강도함수(intensity function) 혹은 고장발생률(rate of occurrence of failure, ROCOF) $\lambda(t) = dE[M(t)]/dt$ 은 t 에 대한 단조(monotonic)함수로 흔히 가정한다. 예를 들어 동질적(homogeneous) 포아송과정, Musa-Okumoto(1984)과정, Weibull 과정 그리고, Cox-Lewis 과정은 각각 ROCOF를 상수, 시간에 관한 부분(fraction)함수, 멱(power)함수, 대수선형(log-linear)함수를 가정하였다. 시스템에서 고장시간은 Raftery(1987)가 제안한 일반순서통계량 모형(general order statistics model, GOS)를 사용하고자 한다. GOS 모형에서 소프트웨어 검정시점에서 미지의 N 개의 결합 수가 있다고 가정하고, 관찰된 고장시점은 N 으로부터 추출된 n 개의 순서통계량은 R^+ 에서 만들어진(supported) 밀도 f 을 가진 $i.i.d$ 관찰값이 된다. f 가 지수분포라면

1) 동국대학교 통계학과

2) 제주관광대학 관광정보처리과

Jelinski-Moranda 모형이 된다. 그러면 이 f 을 다양하게 할 수 있다. 예를 들면 랄리, 파레토, 와이블, 지수, 로그정규 그리고, 절단 극값 분포를 따른다고 생각 할수 있다. 이러한 모형을 랄리, 파레토, 와이블, 지수, 로그 정규 그리고 절단 극값 순서 통계량 모형으로 제안 할 수 있다.

고장시점들을 모델링 함에 있어서 $(0, t]$ 사이에 발견된 고장의 수를 $M(t)$ 라고 하고, t 에 대한 감소함수(nondecreasing function)이라고 가정하면 평균함수 $m(t) = EM(t)$ 가 된다. 따라서 $M(t)$ 를 비동질적 포아송과정으로 접근할 수 있다. 결국 GOS모형에서 N 은 평균 θ 를 가진 포아송 분포를 따른다면 $M(t)$ 는 $m(t) = \theta F(t)$ 을 가진 NHPP가 된다. 단, F 는 밀도 f 을 따르는 누적분포 함수이다. $\lim_{t \rightarrow \infty} m(t)$ 가 유한할 때의 과정을 NHPP(f)라고 표현하자. $F(t) = (1 - e^{-\beta t})$ 일 때 Jelinski-Moranda 모형의 $M(t)$ 는 평균함수(mean function) $m(t) = \theta (1 - e^{-\beta t})$ 를 가진 NHPP(f)가 되고 이것을 Goel-Okumoto(1979) 과정이라고 한다.

Jelinski-Moranda 모형과 Goel-Okumoto(1979) 과정의 관계는 Langberg와 Singpurwalla (1985)에 의해서 지적되었다. Goel-Okumoto(1979) 과정과 연관하여, F 가 이중지수, 로그정규, Gompertz, 랄리, 감마, 그리고 갬블이 될 때 이중지수 NHPP(F), 로그정규 NHPP(f), Gompertz NHPP(f), 랄리 NHPP(f), 감마 NHPP(f), 갬블 NHPP(f) 등을 제시할 수 있다.

시간 t 까지 조사하기 위한 시간 절단(time truncated)모형은 n 번의 고장시점의 순서화된 자료를 x_1, x_2, \dots, x_n 이라고 하면 데이터 집합 D_t 는 $(n, x_1, x_2, \dots, x_n; t)$ 와 같이 구성된다. n 번째까지 고장시점이 관찰된 고장 절단 모형일 경우에 데이터 집합 D_{x_n} 은 (x_1, x_2, \dots, x_n) 으로 구성된다. 시간 절단 모형에서의 우도함수는 다음과 같다.

$$L_{NHPP}(\theta, \beta | D_t) = [\prod_{i=1}^n \lambda(x_i)] \exp(-m(t)) \tag{1.1}$$

이 우도함수는 Basawa & Prakasa, Rao(1980)와 Crowder et al (1991)에 의해서 제시되었다. 고장 절단 모형은 (1.1)의 t 을 x_n 으로 대치하면 유사한 형태가 된다.

2. 일반 순서 통계량(GOS)과 NHPP

NHPP(f)에서 $\lim_{t \rightarrow \infty} m(t) = \theta < \infty$ 이고, N 은 조사시점에서의 결함의 수라고 가정하자.

GOS모형에서 Rafty(1987)와 Joe(1989)는 θ, β, N 에 대한 우도함수를 사용하였다. (1.1)의 우도함수 N 이 포아송분포를 따른다는 가정에 의해 주변우도를 만들 수 있다. NHPP(f)에서 시간 절단 모형에 대한 데이터 D_t 가 주어졌을 때 우도함수는 다음과 같다.

$$L_{NHPP(1)}(\theta, \beta, N | D_t) = \frac{N!}{(N-n)!} [\prod_{i=1}^n \frac{\lambda(x_i)}{\theta}] \left(1 - \frac{m(t)}{\theta}\right)^{N-n} \tag{2.1}$$

여기서, $p(N|\theta)$ 을 평균이 θ 인 포아송 확률질량함수(probability mass function)라고 한다면 다음과 같이 표현할 수 있다.

$$\sum_{N=n}^{\infty} L_{NHPP(1)}(\theta, \beta, N | D_t) \times p(N|\theta) = L_{NHPP}(\theta, \beta | D_t) \tag{2.2}$$

그러므로, (2.1)에 있는 우도함수 N 이 포아송분포를 따를 때 주변 우도함수가 된다.

GOS 모형에서의 고장시점은 분포함수가 $F(X|\beta)$ 을 가진 밀도함수 $f(X|\beta)$ 에 따라 분포되는 N 개의 *i.i.d* 관찰값으로부터 추출된 n 개의 순서통계량이 된다. 그러므로, 고장 절단 모형에서의 우도함수는 다음과 같다.

$$L_{GOS}(\theta, \beta | D_{x_n}) = \frac{N!}{(N-n)!} [\prod_{i=1}^n f(x_i | \beta)] [1 - F(x_n | \beta)]^{N-n} \quad (2.3)$$

따라서 $m(t) = \theta F(t)$ ($\lambda(t) = \theta f(t)$)라고 하면 $L_{GOS}(\theta, \beta | D_i)$ 와 (2.1)의 조건부 우도함수는 동일하다.

결과적으로 $m(t) = \theta F(t)$ 와 검정시점에서 결합의 총수를 N 이라고 가정함으로써 NHPP(f)로부터 GOS모형을 유도할 수 있다(Parzen 1962, p.143). 역으로 생각하면 고장시점이 GOS모형과 평균이 θ 인 포아송분포를 따르는 N 을 가정하면 $M(t)$ 는 평균함수 $m(t) = \theta F(t)$ 인 NHPP(f)가 됨을 알 수 있다. (Ross, 1980 참조).

3. GOS 소프트웨어 신뢰성모형에 대한 깃스추정량과 최우추정량

(1.1)의 우도함수에서 $m(t) = \exp[-\theta F(t)]$ 의 형태 때문에 깃스 표본추출을 적용하는데 필요한 조건부분포의 형태를 구체화하기 어렵다. 그러므로 모수 $\theta[1 - F(t)]$ 을 가지는 포아송 분포를 따르는 잠재변수(latent variable) $N' = N - n$ 을 이용한다. 이 경우 사후분포 $p(\theta, \beta | D_i)$ 는 결합밀도 $p(\theta, \beta, N' | D_i)$ 을 주변밀도화 함으로서 얻을 수 있다. 이 밀도는 다음과 같은 구체화가 가능한 조건부밀도의 반복을 이용하여 깃스 표본추출을 통하여 근사밀도를 구할 수 있다.

즉, $p(N' | \theta, \beta, D_i)$, $p(\theta | N', \beta, D_i)$, $p(\beta | N', \theta, D_i)$.

본 논문에서는 $X \perp Y$ 는 X 와 Y 가 독립, $N' = N - n$ 을 의미하고, $\Gamma(a, b)$ 는 평균이 a/b 인 감마 분포를, $P(\theta)$ 는 평균이 θ 인 포아송분포를 나타내고, β 은 미지의 모수를 의미한다.

본 절에서는 다음과 같은 몇 개의 분포를 적용하여 깃스추정량과 기존의 최우추정량을 구하고자 한다.

3.1. 지수(Exponential) 순서 통계량 모형

지수모형에서는 $m(t) = \theta F(t) = \exp[-\theta(1 - e^{-\beta t})]$, $\lambda(t) = \theta f(t) = \theta \beta e^{-\beta x_i}$ 을 (1.1)의 우도함수에 대입하면 다음과 같이 나타낼 수 있다.

$$L_{NHPP(f)}(\theta, \beta | D_i) = \left(\prod_{i=1}^n \theta \beta e^{-\beta x_i} \right) \exp[-\theta(1 - e^{-\beta t})]. \quad (3.1)$$

이 우도함수에 로그를 취하여 각 모수에 대한 편미분을 시행하면 다음과 같이 최우추정치를 계산할 수 있다.

$$\frac{\partial \ln L_{NHPP(f)}(\theta, \beta | D_i)}{\partial \theta} = \frac{n}{\theta} - 1 + e^{-\beta t} = 0 \quad (3.2)$$

$$\frac{\partial \ln L_{NHPP(f)}(\theta, \beta | D_i)}{\partial \beta} = \frac{n}{\beta} - \sum_{i=1}^n x_i - \theta t e^{-\beta t} = 0 \quad (3.3)$$

그러나 (3.2)와 (3.3)의 관계식을 이용한 계산을 수치해석적인 반복법의 하나인 뉴우튼-랩슨(Newton-Raphson)법을 이용하여 모수의 최우추정치 $\hat{\theta}_{mle}$ 와 $\hat{\beta}_{mle}$ 를 계산 할 수 있다. 베이즈 추정법을 시행하기 위하여 사전분포는 $\theta \sim \Gamma(a, b)$, $a, b > 0$, $\beta \sim \Gamma(c, d)$, $c, d > 0$, $\theta \perp \beta$ 이라고 하면 사후결합밀도는 사전분포와 우도함수 (3.1)를 이용하면 다음과 같은 식이 된다(베이즈 정리 이용).

$$\begin{aligned}
 p(\theta, \beta, N' | D_t) &\propto L_{NHPP(t)}(\theta, \beta | D_t) \cdot p(\theta) \cdot p(\beta) \cdot p(N') \\
 &= \frac{(\theta e^{-\beta t})^{N'}}{N'!} \cdot e^{-\theta e^{-\beta t}} \cdot \theta^{n+a-1} \cdot \beta^{c+n-1} \cdot e^{-b\theta} \cdot e^{-(d+\sum x_i)\beta} \exp[-\theta(1-e^{-\beta t})] \\
 &= \frac{1}{N'!} \theta^{N'+n+a-1} \cdot \beta^{c+n-1} \cdot e^{-(b+1)\theta} e^{-(d+tN'+\sum_{i=1}^n x_i)\beta}
 \end{aligned}$$

이식을 장애(nuisance)모수 개념을 이용하여 깃스 샘플링 알고리즘에 사용되는 각 모수에 대한 사후조건부밀도(full posterior conditional densities)는 다음과 같다.

$$N' | \theta, \beta, D_t \sim P(\theta e^{-\beta t}) \tag{3.4}$$

$$\theta | N', \beta, D_t \sim \Gamma(a+n+N', b+1) \tag{3.5}$$

$$\beta | N', \theta, D_t \sim \Gamma(c+n, d+tN'+\sum_{i=1}^n x_i) \tag{3.6}$$

위 조건부 밀도를 이용한 깃스 알고리즘은 다음과 같은 단계를 이용하여 시행한다.

(0단계)

$\Gamma(a, b)$ 와 $\Gamma(c, d)$ 의 분포에서 데이터를 랜덤추출하여 초기값을 각각 $\theta^{(0)}, \beta^{(0)}$ 를 정한다. 단, a, b, c, d 는 0보다 큰 임의의 알려진 상수를 의미한다.

(1-1단계)

$\theta = \theta^{(0)}, \beta = \beta^{(0)}$ 로 고정시켰을 경우 (3.4)에 대입하여 생성된 랜덤포본 하나를 $N^{(1)}$ 이라 한다. 즉

$$N^{(1)} \sim P(\theta^{(0)} e^{-\beta^{(0)} t})$$

위와 동일한 방법으로(1-1단계)를 이용하여 다음 (1-2단계), (1-3단계)를 얻는다.

(1-2단계)

$$\theta^{(1)} \sim \Gamma(a+n+N^{(1)}, b+1)$$

(1-3단계)

$$\beta^{(1)} \sim \Gamma(c+n, d+tN^{(1)}+\sum_{i=1}^n x_i)$$

(2단계)

(1-1단계), (1-2단계), (1-3단계)로부터 고정시킨 θ, β, N' 의 값을 가장 최근에 생성된 랜덤포본의 값으로 대체하면서 (1-1단계), (1-2단계), (1-3단계)를 충분히 큰 수 (본 논문에서는 70번) 만큼 반복 수행한다. 이렇게 하여 얻은 최종표본을 $(\theta_1^{(70)}, \beta_1^{(70)}, N_1^{(70)})$ 이라 한다.

(3단계)

(1-1단계), (1-2단계), (1-3단계)를 다시 (2000-1)번 반복 수행하면 총 2000개의 랜덤 표본 $(\theta_1^{(70)}, \beta_1^{(70)}, N_1^{(70)}), (\theta_2^{(70)}, \beta_2^{(70)}, N_2^{(70)}), \dots, (\theta_{2000}^{(70)}, \beta_{2000}^{(70)}, N_{2000}^{(70)})$ 이 얻어진다.

(4단계)

최종적인 결과에 의해 θ, β, N' 의 추정치는 다음과 같다.

$$\hat{\theta}_{Gibbs} = \frac{1}{2000} \sum_{k=1}^{2000} \theta_k^{(70)}, \quad \hat{\beta}_{Gibbs} = \frac{1}{2000} \sum_{k=1}^{2000} \beta_k^{(70)}, \quad \hat{N}'_{Gibbs} = \frac{1}{2000} \sum_{k=1}^{2000} N'_k^{(70)} \quad (3.7)$$

3.2 감마(Gamma) 순서 통계량 모형

형상모수 (shape parameter)가 2인 경우를 고려하여 보면 우도함수는 다음과 같이 표현된다.

$$L_{NHPP(1)}(\theta, \beta | D_t) = \left(\prod_{i=1}^n \theta \beta^2 x_i \exp^{-\beta x_i} \right) \exp[-\theta \{1 - e^{-\beta t} (1 + \beta t)\}]$$

사전분포는 $\theta \sim \Gamma(a, b)$, $\beta \sim \pi_1(\beta)$. 단, π_1 은 $\beta > 0$ 을 만족하는 임의의 사전분포이고 $\theta \perp \beta$ 을 만족한다. 이 우도함수에 로그를 취하여 각 모수에 대한 편미분을 시행하면 다음과 같이 최우추정치를 계산할수 있다.

$$\frac{\partial \ln L_{NHPP(1)}(\theta, \beta | D_t)}{\partial \theta} = \frac{n}{\theta} - 1 + e^{-\beta t} (1 + \beta t) = 0 \quad (3.8)$$

$$\frac{\partial \ln L_{NHPP(1)}(\theta, \beta | D_t)}{\partial \beta} = \frac{2n}{\beta} - \sum_{i=1}^n x_i - \theta e^{-\beta t} (\beta t^2) = 0 \quad (3.9)$$

그러나 (3.8)와 (3.9)의 관계식을 이용한 계산을 수치해석적인 반복법의 하나인 뉴우튼-랩슨 (Newton-Raphson)법을 이용하여 모수의 최우추정치 $\hat{\theta}_{mle}$ 와 $\hat{\beta}_{mle}$ 를 계산 할 수 있다.

깁스 표본 추출 알고리즘은 지수모형과 유사한 방법으로 다음과 같은 조건부 밀도를 이용하여 시행한다.

$$N' | \theta, \beta, D_t \sim P[\theta e^{-\beta t} (1 + \beta t)] \quad (3.10)$$

$$\theta | N', \beta, D_t \sim \Gamma(a + n + N', b + 1) \quad (3.11)$$

$$\beta | N', \theta, D_t \propto \beta^{2n} e^{-\beta \sum x_i} [e^{-\beta t} (1 + \beta t)]^{N'} \cdot \pi_1(\beta) \quad (3.12)$$

앞의 지수모형 알고리즘과 유사하게 적용되지만 식(3.12)은 따르는 분포가 구체화 되지 않기 때문에 앞 절의 (1-3단계)에서 메트로폴리스 기법을 적용해야 한다.

즉,

$$\beta | N', \theta, D_t \propto \beta^{2n} e^{-\beta \sum x_i} [e^{-\beta t} (1 + \beta t)]^{N'} \times \pi_1(\beta)$$

위 식에서 오른쪽 식 즉, 목적분포를 간편하게 $f(\beta)$ 라고 표기하고, N' 는 앞 단계에서 깁스추출법으로 추정된 값이고 $\pi_1(\beta) \propto \text{상수} (>0)$ 라고 가정하였고 추이커널(transitional kernel)은 거의 대칭을 이루는 $\Gamma(a, b)$ 에서 β' 을 발생하고 균등분포 (0, 1]에서 확률변량을 w 라고 하면 $\log w \leq \log f(\beta') - \log f(\beta)$ 을 만족하면 β' 를 β 로 간주되고 만족되지 않으면 β 를 β' 로 대치되면서 충분한 반복을 시행한다. 따라서 β 는 Metropolis 알고리즘에 의하여 발생되어진다.

3.3 랄리(Rayleigh) 순서 통계량 모형

랄리분포는 형상모수를 2로 한 와이불분포의 특수한 경우이므로 우도함수는 다음과 같이 표현된다.

$$L_{NHPP(1)}(\theta, \beta | D_t) = \left(\prod_{i=1}^n 2\theta \beta x_i \exp(-\beta x_i^2) \right) \cdot \exp[-\theta (1 - e^{-\beta t^2})].$$

최우추정법과 깃스추정량은 앞 절의 경우와 유사한 방법으로 시행하여 그 결과를 얻을 수 있다. 사전분포를 $\theta \sim \Gamma(a, b)$, $\beta \sim \Gamma(c, d)$, $\theta \perp \beta$ 라고 하면 깃스추출법에 사용되는 조건부 밀도는 다음과 같다.

$$\begin{aligned} N' | \theta, \beta, D_t &\sim P(\theta e^{-\beta t^2}); \\ \theta | N', \beta, D_t &\sim \Gamma(a + n + N', b + 1); \\ \beta | N', \theta, D_t &\sim \Gamma(n + c, N'^2 + d + \sum_{i=1}^n x_i^2); \end{aligned}$$

4. NHPP에 대한 베이지안 추론 및 모형선택

미지의 모수를 추론하기 위해서는 조건부 예측 순서량(ordinates)과 미래 신뢰성 함수(future reliability function)를 이용하여 고장 시점 사이의 평균 시간을 추정해야 한다.

N, θ 그리고 β 와 같은 모수를 추정하기 위해서는 깃스 추출법의 경험적 측도(empirical measure)를 사용할 수 있다. 예측은 고장 절단 상황에서 고려하는 것이 보다 쉽기 때문에 n 번째 고장까지 조사하고 모든 사후분포는 t 를 x_n 으로 대치함으로써 가능하게 된다.

x_i 로부터 거리 x 만큼 떨어진 시점에서의 미래 생존 함수(future survival function)의 추론은 다음과 같은 식을 통해 얻을 수 있다.(Cinlar 1975, p.97 참조)

$$\begin{aligned} E(S(x) | D_{x_i}) &= E[E(P(X_{i+1} > x_i + x) | \beta, D_{x_i}) | D_{x_i}] \\ &= E[\exp(-m(x_i + x) + m(x_i)) | D_{x_i}] \end{aligned} \tag{4.1}$$

미래 시점 x_{i+1} 에 대한 PCPO(Prequential Conditional Predictive Ordinate)는 (x_1, \dots, x_i) 가 주어진 미래의 관찰 시점 x_{i+1} 에서 나타나는 X_{i+1} 의 조건부 밀도 $c_{i+1} = p(x_{i+1} | D_{x_i})$ 으로 정의할 수 있다. 이 PCPO는 과거의 데이터가 주어진 상태에서 X_{i+1} 의 값을 예측하기 때문에 모델 선택을 하는데도 적절한 도구가 된다. 계열 $\{X_i\}_{i \geq 1}$ 이 주어지면 $c_{i+1} = p(x_{i+1} | x_i)$ 으로 표기할 수 있다. PCPO는 다음과 같은 식을 통해 계산된다.

$$\begin{aligned} p(x_{i+1} | x_i) &= \int p(x_{i+1} | \beta, D_{x_i}) p(\beta | D_{x_i}) d\beta \\ &= \int \lambda(x_{i+1}) \exp(-m(x_{i+1}) + m(x_i)) p(\beta | D_{x_i}) d\beta \end{aligned} \tag{4.2}$$

(4.1)과 (4.2)는 깃스 추출법을 이용하거나 Rao-Blackwell정리를 사용하여 평가할 수 있다.

이것을 Gelman & Rubin(1992)이 제시한 마코브체인 몬테칼로(MCMC)방법을 적용하고자 한다. $\theta^{(r,s)}$ 은 MCMC의 반복 알고리즘을 사용하여 r 번 반복 후 s 번 적용을 통해 θ 를 발생시킨 깃스 표본추출을 의미하고 유사한 방법으로 $\beta^{(r,s)}$ 을 발생시킨다. S 가 충분히 큰 짝수를 대입한 깃스 표본추출을 사용하면 (4.2)의 추정치를 유도할 수 있다.

예를 들어 지수분포를 고려하면 다음과 같다.

$$\hat{p}(x_{i+1} | x_i) = \frac{2}{RS} \sum_{r=1}^R \sum_{s=\frac{S}{2}+1}^S \theta^{(r,s)} \beta^{(r,s)} e^{-x_i \beta^{(r,s)}} \exp(\theta^{(r,s)} e^{-\beta^{(r,s)} x_{i+1}} - \theta^{(r,s)} e^{-\beta^{(r,s)} x_i})$$

NHPP는 여러 모형들이 존재할 수 있고 최적 모형의 선택은 베이지안 PCPO를 사용하여 해결할 수 있다. x_{i+1} 에 대한 PCPO는 $c_{i+1} = p(x_{i+1} | x_1, \dots, x_i), i \geq 1$ 에 의해 정의된다. 데이터에 대한 사전분포를 사용하여 각 모형에서 i 에 대한 c_i 를 플롯(plot)시킨다. 평균적으로 가장 큰 c

를 가진 모형이 더 좋은 모형이 된다. 그러나, 플롯에서 c 의 값이 아주 작을 때는 이상치(outlier)가 존재할 수도 있다. 따라서 베이저안 예측 우도 기준은 예측밀도

$$C(l) = \prod_{i=1}^n c_i(l) = p(x_1, x_2, \dots, x_n)$$

를 최대화시키는 모형이 선택되어진다. 단, l 은 모형들에 대한 지수를 의미한다. 이러한 계산은 (4.2)로부터 직접 이루어진다. 모형선택의 대안으로서 상대오차(relative errors)의 합으로 비교할 수 있으며 다음과 같이 정의된다.

$$RE(l) = \sum_{i=1}^{n_i} (n_i - \hat{m}(x_i))^2 / \hat{m}(x_i)$$

단, n_i 는 $(0, x_i]$ 사이에 고장 발생횟수를 의미하고 $\hat{m}(x_i)$ 는 평균함수의 베이저안 추정량을 의미한다. 따라서 $RE(l)$ 가 작은값을 가지는 모형이 좋은 모형이라고 간주 할 수 있다.

5. 수치적인 예

<표1>에 있는 자료는 파레토 평균함수 $\theta[1 - \beta^t / (t + \beta)^\alpha]$ 에서 $\alpha=0.003, \beta=0.98, t=150$ 를 사용하여 FORTRAN IMSL RNGAM함수를 이용하여 난수 발생된 모의 자료이다.

<표3-1>, <표3-2>, <표3-3>은 적당한 사전분포를 사용하여 2000번 적용을 하고 70번 반복하여 나타난 깃스 추정치들이고 이 표에서 수렴성을 확인하기 위해서 500, 2000번 적용에 50, 70번의 결과를 나타내었다. 사전분포는 분산이 큰 $\beta \sim \Gamma(1, 0.0001)$, $\theta \sim \Gamma(1, 0.0001)$ 을 선택 이용하였고, 메트로폴리스 기법의 적용이 용이하도록 하기 위해 사전 분포 $\pi_1(\beta) = 1$ (상수), $\beta > 0$ 을 가정하였고, 최우추정치는 뉴우튼-랩슨법에서 초기치가 0.1로 하고 100번 반복을 사용하였다. 그 결과 두 추정치는 θ 의 추정치는 거의 유사하지만 β 의 추정치는 깃스 추정치가 자료의 모수 $\beta=0.98$ 에 보다 가까운 추정치가 추정되므로 효율적 방법임을 알 수 있다. <그림 1>은 생존함수를 이용하여 미래의 신뢰도를 예측한 결과 비증가함수의 추세로 나타나고 있으며 <표 2>는 밀도함수와 평균함수를 요약하였고, <표4>는 PCPO를 이용하여 계산된 사후베이지스요인($\ln C'$)과 상대오차의 합($RE(l)$)을 이용하여 모델 선택을 하였다. 그 결과로 파악되는 것은 본 모의 자료는 탈리분포가 지수, 감마보다 효율적 모형이라고 결론을 내릴 수 있다. 상대오차의 합도 같은 결과를 나타내고 있다. <그림 2>는 각 시점에서의 PCPO의 그림을 나타내었다.

6. 결 론

기존의 최우추정법을 이용하거나 사전분포를 추가한 사후분포를 구체화 하는데 있어서 수식의 복잡성 때문에 깃스 추출법을 통하여 얻은 표본의 추정값을 계산 할 수 있어서 MCMC 방법을 도구로 하여 NHPP모형에 대한 정보를 계산할 수 있었다. 모형선택도 사후베이지스요인과 상대오차의 합을 이용하여 결론을 내릴 수 있었다. 그러나 이 기법의 단점은 수렴성의 증명에 있다. 몇 가지 탐색적 방법(예 Q-Q 플롯)이 있지만 매 시행마다 조사하려면 번거롭기 때문에 이 방면에 대한 연구에 관심을 가지고 연구가 확대되어야 한다.

<표 1> 소프트웨어의 모의고장시간

고장번호	고장간격시간	고장번호	고장간격시간
1	4	16	3
2	3	17	3
3	4	18	5
4	6	19	3
5	4	20	3
6	3	21	3
7	4	22	14
8	5	23	5
9	6	24	14
10	5	25	5
11	6	26	6
12	5	27	6
13	1	28	1
14	2	29	6
15	5	30	10

<표 2> 밀도함수와 평균 함수

	밀도	GOS/NHPP(1) 평균함수
지수	$\beta \exp(-\beta t)$	$\theta[(1 - e^{-\beta t})]$
감마	$\beta^2 t \exp(-\beta t)$	$\theta[1 - e^{-\beta t}(1 + \beta t)]$
랄리	$2\beta t \exp(-\beta t^2)$	$\theta[1 - \exp(-\beta t^2)]$

<표3-1>지수분포모형

S	R	$\hat{\theta}_{Gibbs}$	$\hat{\theta}_{MLE}$	$\hat{\beta}_{Gibbs}$	$\hat{\beta}_{MLE}$
500	50	31.33334	31.0088	0.008824	0.00207
500	70	31.33434		0.008922	
2000	50	31.35833		0.008904	
2000	70	31.39831		0.008701	

<표3-2> 감마분포(2) 모형

S	R	$\hat{\theta}_{Gibbs}$	$\hat{\theta}_{MLE}$	$\hat{\beta}_{Gibbs}$	$\hat{\beta}_{MLE}$
500	50	32.48748	31.7150	0.001023	0.000991
500	70	32.18952		0.001106	
2000	50	32.52896		0.000989	
2000	70	32.50198		0.001006	

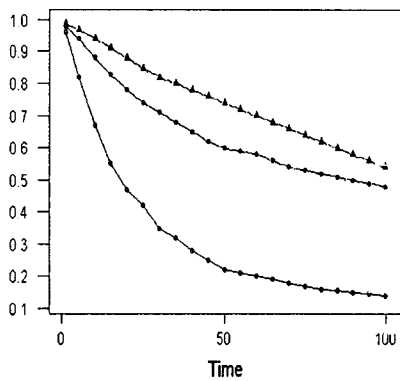
<표3-3> 랄리분포모형

S	R	$\hat{\theta}_{Gibbs}$	$\hat{\theta}_{MLE}$	$\hat{\beta}_{Gibbs}$	$\hat{\beta}_{MLE}$
500	50	32.37362	31.4723	0.000491	0.000391
500	70	32.16149		0.000428	
2000	50	32.84469		0.000437	
2000	70	32.01269		0.000429	

<표 4> 모형 선택

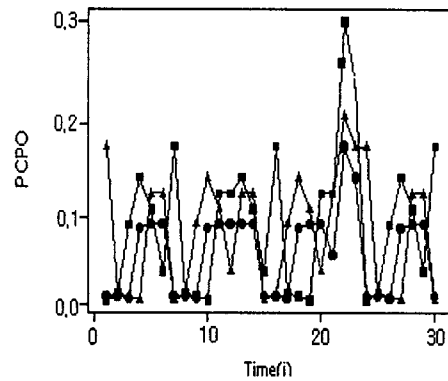
사후베이지스요인과 ($\ln C'$) 과 상대오차의 합($RE(l)$)

모형	지수모형	감마 모형	랄리 모형
$\ln C'$	-68.20	-72.23	-65.15
$RE(l)$	23.3929	10.6438	7.19660



단, ●은 지수모형 ○은 랄리모형 ▲은 감마 모형

<그림 1> NHPP $P(\text{Time} | D_i)$



단, ●은 지수모형 ■은 랄리모형 ▲은 감마 모형

<그림 2> 각 시점에 대한 PCPO

참 고 문 헌

- [1] Aitkin, M.(1991), "Posterior Bayes Factors," *Journal of the Royal Statistical Society, B*, 111-142.
- [2] Basawa, I. V., and Prakasa and Rao, B.L.S (1980), *Statistical Inference for statistic Process*, London; Academic Press.
- [3] Blumenthal, S., and Marcus, R.(1975), "Estimating Population Size with Exponential Failure," *Journal of the American Statistical Association*, 70, 913-922.
- [4] Casella, G., and George, E. I. (1992), "Explaining the Gibbs Sampler," *The American Statistician*, 46, 167-174.
- [5] Cinlar, E.(1975), *Introduction To Stochastic Process*, New Jersey : Prentice-Hall.
- [6] Crowder, et al (1991), *Statistical Analysis of Reliability*, London; Champman and Hall
- [7] Dawid, A. P.(1984), "Statistical Theory : The Prequential Approach," *Journal of the Royal Statistical Society, Ser. A*, 147, 278-292.
- [8] Joe, H., and Reid, N. (1989), "Statistical Inference for General-Order-Statistics and Nonhomogeneous-Poisson-Press Software Reliability Models," *IEEE Transaction on Software Engineering*, 15, 1485-1490
- [9] Gelman, A. E., and Rubin D.(1992), "Inference from Iterative Simulation Using Multiple Sequences," *Statistical Science*, 7, 457-472.
- [10] Geol, A. L., and Okumoto, K. (1979), " Time Dependent Error Detection Rate Model for Software Reliability and Other Performance Measures," *IEEE Transaction on Reliability*,28, 206-211.
- [11] Kuo, L., and Yang, T. Y.(1995), "Bayesian Computation of Software Reliability", *Journal of Computational and Graphical Statistics*, 65-82.
- [12] Langberg, N., and Singpurwalla, N. D. (1985), "A Unification of Software Reliability Models." *SIA Journal on Scientific and Statistical Computing*" 6, 781-790.
- [13] Musa, Iannino and Okumoto (1987), "*Software Reliability : Measurement, Prediction, Application*", New York:McGraw Hill
- [14] Parzen, E. (1962), *Stochastic Process*, San Francisco ; Hold-Day
- [15] Raftery, A. E.(1987), "Inference and Prediction for a General Order Statistic Model with Unknown Population Size," *Journal of the American Statistical Association*, 82, 1163-1168.
- [16] Ross, S. M.(1980), *Stochastic Process*.
- [17] Stone, M.(1974), "Cross-Validatory Choice and Assessment of Statistical Predictions," *Journal of the Royal Statistical Society, Ser. B*, 111-147.
- [18] User Manual *STAT/LIBRARY(1987) Fortran Subroutines for statiatical analysis*, IMSL vol 3.