Journal of the Society of Korea Industrial and Systems Engineering
Vol, 28, No. 3, pp.26- , September 2005.

베이지안 확률 모형을 이용한 위험률 함수의 추론*

김현묵 · 선웅**

한양대학교 산업공학과

Hazard Rate Estimation from Bayesian Approach

Hyunmook Kim · neung Ahn**

Department of Industrial Engineering, Hanyang University, Ansan, Kyeonggi-do

This paper is intended to compare the hazard rate estimations from Bayesian approach and maximum likelihood estimate(MLE) method. Hazard rate frequently involves unknown parameters and it is common that those parameters are estimated from observed data by using MLE method. Such estimated parameters are appropriate as long as there are sufficient data. Due to various reasons, however, we frequently cannot obtain sufficient data so that the result of MLE method may be unreliable. In order to resolve such a problem we need to rely on the judgement about the unknown parameters. We do this by adopting the Bayesian approach. The first one is to use a predictive distribution and the second one is a method called Bayesian estimate. In addition, in the Bayesian approach, the prior distribution has a critical effect on the result of analysis, so we introduce the method using computerized-simulation to elicit an effective prior distribution. For the simplicity, we use exponential and gamma distributions as a likelihood distribution and its natural conjugate prior distribution, respectively. Finally, numerical examples are given to illustrate the potential benefits of the Bayesian approach.

Keywords : ayesian Approach, Hazard Rate, Maximum Likelihood Estimate, Reliability

1. 서 론[3)]

일반적으로 공학시스템(engineering system)의 고장을 수학적으로 설명하는 것은 어렵다. 이러한 어려움은 시스템에 내재되어 있는 물리적 측면을 고려함으로써 해결할 수 있는데 그러한 방법으로 위험률함수(hazard rate)를 이용할 수 있다. 위험률함수는 t시간 이전에 고장이 발생하지 않았다는 가정하에 매우 짧은 시간 구간 $[t, t+dt]$동안의 고장발생율을 의미한다[12]. 위험률함수는 주어진 시간 동안의 고장발생비율, 시간에 따른 고장확률의 변화, 주어진 시간에 대한 조건부 고장확률 등 여러 정보를 제공하기 때문에[19] 위험률함수를 정확하게 추정할 수 있다면 공학시스템의 유지보수에 있어 매우 유용할 것이다.

만약 관측된 데이터가 있고 그 데이터가 위험률함수 추정에 영향을 미친다면, 위험률함수는 관측된 데이터에 의해 새롭게 갱신(update)되어야 할 것이다. 일반적으로 위험률함수는 미지의 파라미터(unknown parameter)에 의해 표현되고, 미지의 파라미터는 최대우도추정량(maximum likelihood estimate; MLE) 방법에 의해 추정되어왔다. 관측된 데이터에만 의존하는 최대우도추정량 방법은 충분한 데이터 존재 시 매우 효율적이지만 만약 데이터가 충분치 않을 경우 최대우도추정량 방법에 의한 결과는 신뢰할 수 없을 것이다. 따라서 충분한 데이터가 존재하

* 본 연구는 한국과학재단 지역대학우수과학자사업의 지원으로 수행되었음(R05-2003-000-10290-0).
** 연락저자 : : : 82-31-409-2423,
E-mail : nahn@hanyang.ac.kr

지 않다면 미지의 파라미터를 확률적으로 표현해야 할 것이고, 이러한 요구사항을 수렴하는 방법론으로 베이지안 확률 모형을 들 수 있다.

베이지안 확률 모형에서는 정확한 결과를 위해 적절한 사전분포(prior distribution)가 필요하다. 따라서 본 논문에서는 시뮬레이션 방법을 이용하여 가상데이터를 발생시켜 해당 시스템 전문가의 주관적 견해를 사전분포로 추론할 수 있는 방법을 소개한다. 또한 추론된 사전분포를 기반으로 베이지안 확률 모형에서는 위험률함수를 크게 두 가지 방법으로 추정할 수 있다. 첫째로, 우도함수와 사전분포로부터 예측분포(predictive distribution)를 계산하고, 이 예측분포로부터 위험률함수의 정의에 따라 위험률함수를 추정하는 것이다. 둘째로, 베이지안 추정치(Bayesian estimate)를 이용하는 것이다. 베이지안 추정치는 우도함수의 위험률함수를 확률분포를 갖는 우도함수의 파라미터에 대해 기대값(expectation)을 취하는 것을 말한다. 또한 베이지안 추정치는 제곱오차 손실함수(squared error loss function)에 대한 최적의 추정치이며, 동시에 주어진 데이터에 의해 얻은 사후분포(posterior distribution)의 기대값이다[14].

본 논문에서는 계산의 용이성을 위해 지수분포(exponential distribution)를 우도함수로, 감마분포(gamma distribution)를 사전분포로 이용한다. 두 분포를 이용하여 데이터 발생시 사전분포를 사후분포로 갱신하고, 갱신된 사후분포를 이용하여 위험률함수 또한 갱신할 것이다.

위의 서술된 연구를 수행하기 위해 본 논문은 다음과 같이 구성된다. 2장에서는 사전분포 추론 방법을 설명한다. 3장에서는 베이지안 확률 모형을 이용한 위험률함수 추정에 대해 다루고, 4장에서는 베이지안 모형과 최대우도추정량 방법을 비교한다. 5장에서는 베이지안 확률 모형의 장점을 예를 들어 설명하고, 마지막으로 6장에서 결론을 다룬다.

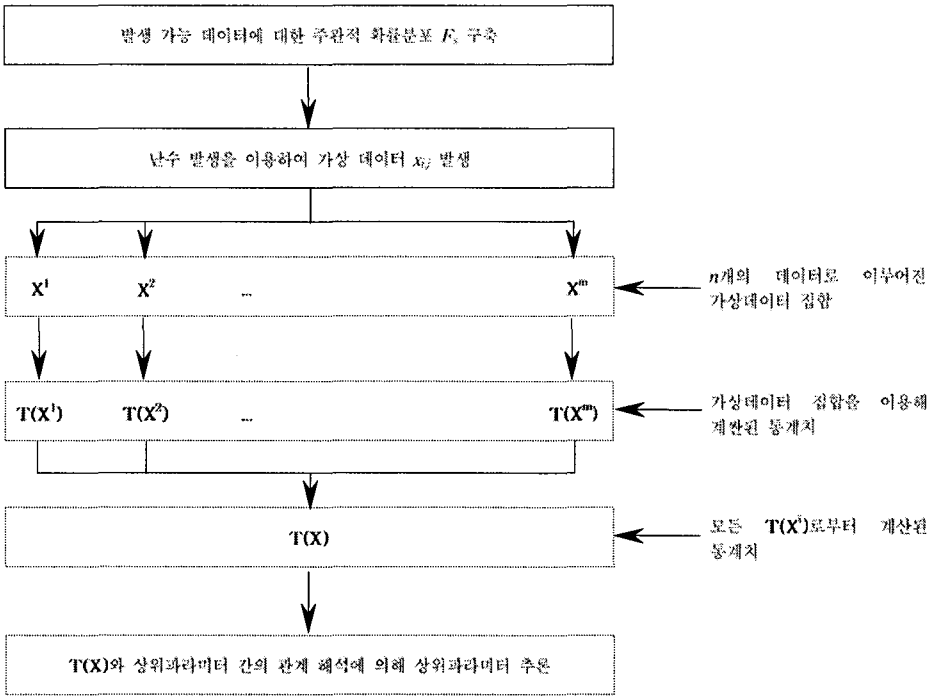
2. 사전분포 추론 방법

사전분포 추론을 위한 정형화된 방법론의 필요성이 증대되고 있지만 아직까지 표준적인 방법론이 존재하지 않는 것이 사실이다[4, 13]. 사전분포를 추론하기 위해서는 사전분포로 어떤 확률분포를 사용할 것인가(사전분포의 함수형태), 사전분포의 파라미터인 상위파라미터(hyperparameter)를 어떻게 결정할 것인가라는 두 가지 문제를 해결해야 한다. 우선 베이지안 확률 모형에서는 보통 자연공액사전분포(natural conjugate prior distribution)가 많이 사용되는데, 이는 자연공액사전분포의 함수형태가 우도함수의 형태와 같아 결국 사후분포 역시 같은 함수의 형태로 표현되기 때문이다[10]. 따라서 만약 우도함수가 자연공액사전분포를 갖는다면 자연공액사전분포를 이용함으로써 사전분포의 함수형태에 관한 문제는 해결될 것이다. 자연공액사전분포를 갖는 우도함수로는 지수분포, 정규분포, 포아송분포, 이항분포 등을 예로 들 수 있다[15]. 하지만 자연공액사전분포를 이용함으로써 사전분포 함수의 형태를 결정하더라도 여전히 상위파라미터를 결정하는 것은 해결되지 않는다. 따라서 본 장에서는 우도함수가 자연공액사전분포를 갖는 경우에 대해 시뮬레이션 방법을 이용하여 상위파라미터를 추론하는 방법을 소개한다.

만약 상위파라미터 추론을 위한 관측 데이터 혹은 정보가 존재한다면 큰 도움이 될 것이다. 하지만 때때로 공학 시스템에서는 데이터 및 정보의 부족이 빈번히 발생하고, 이는 통계적 분석을 어렵게 하는 이유가 된다. 따라서 본 연구에서는 상위파라미터 추론을 위한 관측 데이터 혹은 정보가 존재하지 않는 경우에 초점을 둔다.

상위파라미터 추론을 위한 관측 데이터 혹은 정보가 존재하지 않는 경우에는 해당 시스템 전문가의 주관적 견해(subjective knowledge)가 중요하게 활용될 수 있을 것이다. 또한 일반적으로 모수에 대한 관측은 쉽지 않기 때문에, 상위파라미터 추론을 위해 전문가의 주관적 견해가 필요할 때는 모수에 대한 주관적 견해를 직접적으로 요구하기 보다는 실제로 발생할 수 있는 데이터에 대한 주관적 견해를 요구하는 것이 더욱 적절할 것이다. 그 후 실제로 발생할 수 있는 데이터가 우도함수와 관련이 있고, 우도함수의 특정 통계치는 상위파라미터와 관계가 있다는 점을 이용하여 상위파라미터를 추론할 수 있다.

상위파라미터 추론을 위해, 시스템 혹은 부품의 고장과 관련하여, 전문가는 주관적 견해를 특정 시간기간동안에서의 고장확률, $\{p_j; (t_j, t_{j+1})\}$과 같이 표현할 수 있음을 가정하자. 이와 같은 전문가의 주관적 확률분포를 F_s로 표기하자. 실제로 전문가의 주관적 견해를 수량화하는 표준적인 방법이 존재하지 않아 F_s를 구축하는 것 역시 어렵지만, 이와 관련된 여러 연구들은[1, 2, 4, 5, 13, 15, 22, 23] 많은 유용한 방법들을 제공해주고 있다. 구축된 F_s를 이용하여 우도함수의 특정 통계치를 얻기 위한 가상데이터(artficial data) x_{ij}를 발생시키기 위해 두 개의 난수(random number), $r(i, j, 1)$과 $r(i, j, 2)$를 이용한다. i는 가상데이터 집합의 순번을, j는 i번째 집합내의 가상데이터의 순번을 나타낸다. 예를 들어, $r(i, j, 1)$에 의해 k번째 시간구간이 선택되었다면 가상데이터 x_{ij}는 $t_k+r(i, j, 2)\times(t_{k+1}-t_k)$과 같이 구할 수 있다.



<그림 1> 제시하는 상위파라미터 설정 방법의 절차

위와 같은 계산을 가상데이터 x_{im}을 얻을 때까지 반복 수행하면 i번째 가상데이터 집합 $\mathbf{X}^i$를 얻을 수 있고 마찬가지로 반복적인 절차를 통해 가상데이터 집합 $\mathbf{X}^1$, $\mathbf{X}^2$, ..., $\mathbf{X}^m$을 얻을 수 있다. 또한 각각의 가상데이터 집합을 이용하여 상위파라미터와 관련이 있는 우도함수의 특정 통계치 $T(\mathbf{X}^1)$, $T(\mathbf{X}^2)$, ..., $T(\mathbf{X}^m)$을 계산하고, $T(\mathbf{X}^1)$, $T(\mathbf{X}^2)$, ..., $T(\mathbf{X}^m)$을 종합적으로 고려하여 상위파라미터 추론을 위한 $T(\mathbf{X})$를 구한다. 마지막으로 상위파라미터와 $T(\mathbf{X})$간의 수학적 관계를 통해 상위파라미터를 추론할 수 있다. 이와 같은 일련의 과정이 <그림 1>에 요약되어 있다.

제시하는 방법론을 수행하기 위해서는 먼저 여러 가지 변수들을 설정해야 한다. 예를 들어, 가상데이터 집합의 크기 n, 가상데이터 집합의 개수 m, 통계치 $T(\mathbf{X}^i)$와 $T(\mathbf{X})$ 등이 있다. 먼저 n과 m은 상위파라미터 추론을 위해 사용될 가상데이터의 개수와 관련이 있는데, n과 m이 클수록 전문가의 주관적 확률분포 F_s에 근접한 사전분포를 표현해주는 상위파라미터를 얻을 수 있을 것이다. 그리고 $T(\mathbf{X}^i)$와 $T(\mathbf{X})$는 우도함수와 사전분포에 따라 달라질 수 있는데, 보통 $T(\mathbf{X}^i)$는 우도함수의 기대값, $T(\mathbf{X})$는 $T(\mathbf{X}^i)$의 표본평균(sample mean)과 표본분산(sample variance)으로 설정하는 것을 추천한다.

예제

본 절에서는 앞에서 서술한 방법의 이해를 돕기 위해 간단한 예제를 선보인다. 계산의 용이성을 위해 지수분포와 감마분포를 각각 우도함수와 자연공액사전분포로 사용한다.

파라미터가 $\theta \in \Theta$인 지수분포로부터 데이터 t가 발생했을 때, 이에 대한 우도함수는 아래와 같다.

$$L(\theta;t) = p(t|\theta) = \theta e^{-\theta t}, \quad \cdots\cdots (1)$$

이때 $t \geq$ 이고, t는 일반적으로 비율(rate)을 나타내는 파라미터이다. 지수분포의 파라미터 t에 대한 자연공액사전분포는 아래의 감마분포이다[14].

$$\pi(\theta|\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta}, \quad \cdots\cdots (2)$$

이때 $\alpha>0$, $\beta>0$는 상위파라미터이다.

제시하는 방법을 실행하기 위해서는 먼저 여러 변수들을 설정해야 한다. 문제를 간단히 하기 위해서 가상데이터 집합의 크기 n=10, 가상데이터 집합의 개수 m=100이라고 가정하자. 또한 지수분포의 파라미터 t는 지수

분포 기대값의 역수이므로 상위파라미터를 추론하기 위한 통계치로서 $T(X^i)$는 다음과 같이 정의하자.

$$T(X^i)= n/\sum_{j=1}^{n} x_{ij}. \quad \cdots\cdots (3)$$

또한 $T(X)$는 $T(X^1)$, $T(X^2)$, ..., $T(X^m)$의 표본평균과 표본분산이라고 하자.

해당 시스템 전문가의 주관적 확률분포 F_s는 <표 1>과 같이 구축되었다고 하자. <표 1>의 F_s를 이용하여 가상데이터 집합 X^1, X^2, ..., X^m를 발생시켰으며, <표 2>는 첫 번째 가상데이터 집합 X^1의 예를 보여주고 있다.

가상데이터 집합 X^1을 이용하여 우도함수의 기대값의 역수인 $T(X^1)$를 계산하면, $T(X^1)$=10/686이다. 이와 같은 과정을 X^{100}과 $T(X^{100})$을 얻을 때까지 반복한다. 또한 $T(X^1)$, $T(X^2)$, ..., $T(X^{100})$에 대한 표본평균과 표본분산, 즉 $T(X)$를 계산한 결과 각각 1.1654×10^{-2}과 1.9177×10^{-5}이었다. 지수분포의 파라미터 λ를 확률적으로 표현하는 감마분포의 경우 기대값이 α/β, 분산이 α/β^2이므로 이를 가상데이터를 이용하여 얻은 $T(X)$와 대응시키면 $\alpha/\beta=1.1654\times10^{-2}$와 $\alpha/\beta^2=1.9177\times10^{-5}$, 두 식을 얻을 수 있다. 이를 계산하면 사전분포로 $gamma(\theta|\alpha=7, \beta=608)$를 얻을 수 있다. 이는 비록 가상데이터를 이용한 것이지만 충분한 양의 가상데이터를 이용하여 θ에 대해 신뢰할만한 통계치를 계산한 것이므로 이를 상위파라미터에 대응시키는 것이 큰 무리가 없다라는 가정에 근거하고 있다.

상위파라미터 추론을 위한 관측 데이터 혹은 정보가 존재하지 않아 해당 시스템 전문가의 주관적 견해를 바탕으로 상위파라미터를 추론하였기 때문에 제시하는 방법론의 적합성을 검증하기 위해서는 추론된 사전분포가 얼마나 주관적 확률분포 F_s를 잘 반영하는가를 확인해야 한다. 하지만 사전분포는 우도함수의 파라미터를 표현하고 있고, F_s는 우도함수에서 실제로 발생 가능한 데이터에 대한것이기 때문에 둘 사이에는 직접적인 비교가 가능할 수 없다. 따라서 추론된 사전분포로부터 F_s와 직접적인 비교가 가능한 예측분포를 계산하여 이를 F_s와 비교함으로써 추론된 사전분포 즉, 상위파라미터의 적합성을 검증할 수 있을 것이다.

사전분포로 $gamma(\theta|\alpha, \beta)$를 사용할 때 우도함수인 지수분포의 예측분포는 아래와 같이 구할 수 있다.

$$f(y) = \int_{\Theta} f(y|\theta)\pi(\theta|\alpha,\beta)d\theta = \frac{\Gamma(\alpha+1)}{\Gamma(\alpha)}\frac{\beta^{\alpha}}{(\beta+y)^{\alpha+1}}, \quad \cdots\cdots (4)$$

이때 $y\geq$ 이고, $\Gamma(\cdot$: 감마함수이다. 위에서 구한 α=7, β=608를 식(4)에 대입하면 $f(y)=7\times608^7/(608+y)^8$를 구할 수 있고, <그림 2>는 <표 1>에서 제시한 F_s와 식(4)를 통해 구한 예측분포의 누적확률함수(cumulative distribution fuction)를 비교하고 있다. <그림 2>를 통해 제시한 방법에 의해 추론된 상위파라미터를 이용하여 얻은 예측분포가 전문가의 주관적 견해 즉, F_s와 매우 비슷함을 알 수 있다. 실제로 [0, 300]구간에서 F_s의 누적분포와 예측분포의 누적분포함수의 차이는 0.105보다 적다. 이는 제시하는 방법론을 통해 추론된 상위파라미터가 전문가의 주관적 견해를 효과적으로 표현하고 있음을 뒷받침하고 있다.

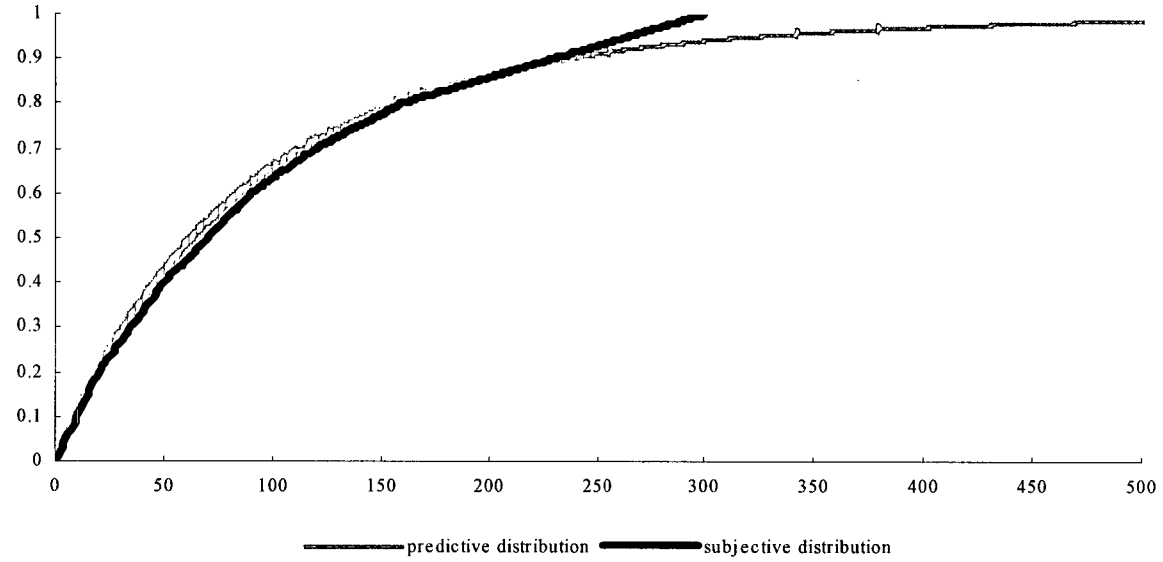
상위파라미터 추론을 위해 제시한 방법은 여러 가지 장점을 가지고 있다. 첫째로, 제시하는 방법의 기본 개념은 매우 간단하다. 둘째로, 제시하는 방법을 컴퓨터를 이용하여 프로그램화하여 실행하는 것 또한 매우 쉽다. 이와는 반대로 수학적 모형은 매우 복잡하여 때때로 원하는 결과를 얻기가 힘든 경우가 있다. 마지막으로, 제시하는 방법은 가상데이터 집합의 크기 n과 가상데이터 집합의 개수 m이 매우 큰 경우 좋은 결과 즉, F_s와 매우 근접한 사전분포를 표현하는 상위파라미터를 얻을 수

<표 1> 주관적 확률분포 Fs

시간구간	0-10	10-20	20-35	35-50	50-70	70-90	90-120	120-160	160-230	230-300
고장확률	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1

<표 2> 가상데이터 집합 X^1의 예

r(1, j, 1)	0.1465	0.559	0.4957	0.2339	0.6258	0.5919	0.6137	0.2143	0.4982	0.6442
선택된 시간구간	10-20	70-90	50-70	20-35	90-120	70-90	90-120	20-35	50-70	90-120
r(1, j, 2)	0.3784	0.8656	0.5106	0.5633	0.6030	0.9987	0.4570	0.4863	0.8945	0.3349
가상데이터 xij	14	87	60	28	108	90	104	27	68	100



<그림 2> 주관적 확률분포 Fs와 예측분포의 누적분포함수

있다.

마지막으로 본 장을 정리하면서 제시한 방법론의 기본 개념을 요약하면 다음과 같다.

- 당 시스템의 전문가는 주관적 견해를 표현할 때, 일반적으로 잘 관측되지 않는 파라미터에 대한 것보다 상대적으로 관측이 쉬운 실제 발생 데이터에 대한 주관적 견해를 보다 잘 표현할 수 있다.
- 회할 만한 우도함수의 특정 통계치를 가지고 있다면 우도함수의 특정 통계치와 상위파라미터간의 수학적 관계를 이용하여 상위파라미터를 추론할 수 있다.
- 측 데이터가 존재하지 않는 경우라도, 전문가의 주관적 견해와 시뮬레이션 방법을 사용하여 우도함수의 특정 통계치를 쉽게 계산할 수 있다.

3. 베이지안 확률 모형을 이용한 위험률함수의 추론

신뢰성공학(reliability engineering)에서 중요한 역할을 하는 위험률함수 $h(t)$는 다음과 같이 정의된다.

$$h(t|\theta) = \frac{f(t|\theta)}{S(t|\theta)}, \quad \cdots\cdots (5)$$

이때 $S(t)$는 생존함수(survival function)이고, θ는 미지의 파라미터이다. 본장에서는 베이지안 확률 모형을 이용하여 위험률함수를 추정하는 방법에 대해 다룬다. 앞에서 서술한 바와 같이 계산의 용이성을 위해 지수분포를 우도함수로, 감마분포를 사전분포로 사용한다.

베이지안 확률 모형에서는 위험률함수를 크게 두 가지 방법으로 추정할 수 있다. 첫째로, 우도함수와 사전분포의 혼합(mixture)을 통해 얻을 수 있는 예측분포를 위험률함수의 정의에 적용하여 위험률함수를 추정하는 것이다. 둘째로, 우도함수의 위험률함수를 확률분포를 갖는 우도함수의 파라미터에 대해 기대값을 취하는 방법이다. 이 방법은 베이지안 추정치라고 불리며, 오차제곱 손실함수에 대해 최적의 추정치를 제공하고, 주어진 데이터에 의해 얻은 사후분포의 기대값이다.

때때로 공학시스템에서 여러 가지 이유로 인해(예를 들어, 유지보수활동, 관측기계의 능력부족, 관측자의 부주의 등) 생존시간(lifetime) 데이터를 정확하게 관측하지 못할 때가 있다. 그러한 경우에 얻을 수 있는 관측 정보 중 하나는 생존시간이 특정 시간 T보다 크다는 것이다. 이러한 데이터 유형이 바로 중도절단자료(censored data)이다[7]. 따라서 중도절단자료를 반영한 우도함수는 아래와 같이 표현되어야 한다.

$$L(\theta;t) = \begin{cases} f(t|\theta) = \theta^{-\theta t} & , completely\ observed \\ f(t \geqq T|\theta) = e^{-\theta T} & , censored \end{cases} \quad \cdots\cdots (6)$$

만약 $k(\leq n)$개의 중도절단데이터와 $n-k$개의 완전관측(completely observed)데이터로 구성된 n개의 데이터 집합 $D=\{t_1, t_2, ..., t_n\}$를 관측했다면, 이에 따른 우도함수는 아래와 같다.

$$\begin{aligned} L(\theta;D) &= \Pi_i L(\theta;t_i) \\ &= \Pi_{j \in D_{CO}} L(\theta;t_j)(e^{-\theta T})^k \\ &= \theta^{n-k} e^{-\theta(\Sigma t_j + kT)}, \end{aligned} \quad \cdots\cdots (7)$$

이때 D_{CO}는 완전관측데이터 집합이고, Σt_j는 D_{CO}에 속하는 완전관측데이터들의 합이다.

식(7)에서 얻은 우도함수와 감마분포를 베이즈 공식(Bayes' formula)에 적용하면 아래와 같은 사후분포를 얻는다[10].

$$\begin{aligned} \pi(\theta|D) = &\frac{(\beta + \Sigma t_j + kT)^{\alpha+n-k}}{\Gamma(\alpha+n-k)} \\ &\times \theta^{\alpha+n-k-1} e^{-(\beta+\Sigma t_j+kT)\theta}. \end{aligned} \quad \cdots\cdots (8)$$

이제 갱신된 사후분포 식(8)을 이용하여 위험률함수를 갱신할 수 있다. 우선 예측분포를 이용하여 위험률함수를 추정하자. 위험률함수를 추정하기 위한 예측분포는 아래와 같이 구할 수 있다.

$$f(y|D) = \int_\Theta f(y|D,\theta)\pi(\theta|D)d\theta. \quad \cdots\cdots (9)$$

파라미터 θ가 주어졌을 때 y와 D가 서로 독립이기 때문에 $f(y|D, \theta)$는 식(1)을 따를 것이고, 식(9)에 식(8)을 대입하면,

$$\begin{aligned} f(y|D) = &(\alpha+n-k)\left(\frac{\beta+\Sigma t_j+kT}{\beta+\Sigma t_j+kT+y}\right)^{\alpha+n-k} \\ &\times\left(\frac{1}{\beta+\Sigma t_j+kT}\right) \end{aligned} \quad \cdots\cdots (10)$$

를 예측분포로 얻게 된다. 또한 예측분포 식(10)의 생존함수 $S(y|D)$는 다음과 같다.

$$S(y|D) = \left(\frac{\beta+\Sigma t_j+kT}{\beta+\Sigma t_j+kT+y}\right)^{\alpha+n-k} \quad \cdots\cdots (11)$$

따라서 위험률함수의 정의 식(4)에 따라 예측분포를 이용한 위험률함수의 추정치는 아래와 같다.

$$h(y|D) = \frac{\alpha+n-k}{\beta+\Sigma t_j+kT+y}. \quad \cdots\cdots (12)$$

이제 베이지안 추정치에 의한 위험률함수 추정치를 구하자. 베이지안 추정치는 오차제곱 손실함수에 대한 최적의 추정치를 제공하는데, 우도함수의 위험률함수를 확률분포를 갖는 우도함수의 파라미터에 대해 기대값을 취하면 얻을 수 있다. 따라서 $E[h(y|D, \theta)]$를 계산해야 하는데, 지수분포의 위험률함수는 파라미터 θ라는 사실[14]과 식(8)을 이용하여 위험률함수에 대한 베이지안 추정치를 계산하면 아래와 같다.

$$\begin{aligned} &E[h(y|D,\theta)] \\ &= E[E[h(y|D,\theta)]|D,y] \\ &= E[\theta|D,y] \\ &= \int_\Theta \theta \frac{(\beta+\Sigma t_j+kT+y)^{\alpha+n-k}}{\Gamma(\alpha+n-k+1)}\theta^{\alpha+n-k} \\ &\qquad \times e^{-(\beta+\Sigma t_j+kT+y)\theta}d\theta \\ &= \frac{\alpha+n-k+1}{\beta+\Sigma t_j+kT+y}. \end{aligned} \quad \cdots\cdots (13)$$

예측분포와 베이지안 추정치를 이용하여 두 개의 위험률함수 추정치, 식(12)와 식(13)을 계산하였다. 계산된 위험률함수 추정치, 식(12)와 식(13)은 모두 아직 관측되지 않은 미래의 데이터 y에 대한 함수이고, 또한 y에 대한 감소함수임을 알 수 있다. 이는 혼합에 의해 계산되는 지수분포의 위험률함수는 감소위험률함수(decreasing hazard rate)라는 사실과 일치하는 것이다[3].

4. 베이지안 확률 모형과 최대우도추정량 방법의 비교

기존의 전통적 통계(classical statistics)에서 미지의 파라미터를 추정함에 있어 가장 널리 쓰이는 최대우도추정량 θ_{MLE}는 다음의 조건을 만족시키는 파라미터 값이다.

$$L(\theta_{MLE};D) \geq L(\theta;D) \text{ for all } \theta \neq \theta_{MLE} \quad \cdots\cdots (14)$$

지수분포의 위험률함수에 대한 최대우도추정량은 앞에서 유도한 우도함수인 식(7)을 최대화시키는 파라미터 θ이며 아래와 같다.

$$h(y|D) = \theta_{MLE} = \frac{n-k+1}{\Sigma t_j + kT + y} \quad \cdots\cdots (15)$$

베이지안 확률 모형에 의한 위험률함수 추정치와 마찬가지로 최대우도추정량 방법에 의한 위험률함수 추정치도 역시 y에 대한 감소위험률함수임을 식(15)를 통해 알 수 있다.

베이지안 확률 모형과 최대우도추정량 방법을 비교하여 베이지안 확률 모형의 여러 가지 장점을 아래와 같이 제시할 수 있다.

- 약 충분한 양의 데이터가 존재하면 관측된 데이터만을 사용하는 최대우도추정량의 추정치는 신뢰적이지만, 여러 공학 시스템에서는 통계적 추론(statistical inference)을 위한 충분한 양의 데이터가 존재하지 않는 경우가 자주 발생한다. 따라서 이러한 경우 관측 데이터만을 사용하는 최대우도추정량의 결과는 신뢰적이지 못할 것이다. 하지만 베이지안 모형은 관측 데이터뿐만 아니라 기타 다른 정보들을 사전분포를 이용하여 표현할 수 있으므로 통계적 추론을 위한 데이터가 충분치 못한 경우라도 정확한 결과를 유도할 수 있다.
- 로운 관측 데이터가 발생하면 최대우도추정량 방법은 과거 데이터와 새로운 데이터의 결합확률분포를 이용하여 새로운 추정치를 계산한다. 이와는 다르게 베이지안 확률 모형은 새로운 데이터 발생시, 기존의 사전분포(혹은 사후분포)에 새로운 데이터를 반영하여 갱신된 사후분포를 계산하고 이를 이용하여 새로운 추정치를 계산한다. 따라서 최대우도추정량 방법과 같이 결합확률분포가 필요한 것이 아니다. 이러한 베이지안 확률 모형의 특징을 학습과정(learning process)이라 한다[15].
- 반적으로 기존의 전통적 통계방식에서는 적합도 검정(goodness of fit test)을 통해 발생한 데이터를 수학적으로 가장 잘 표현하는 분포를 선정하고 이를 분석에 활용하였다. 이러한 방법은 해당문제에 대해 선정된 분포와 그 분포의 파라미터가 물리적 의미를 가지고 있지 않은 경우가 대부분이다. 하지만 베이지안 확률 모형에서는 해당문제에 대한 물리적 의미와 관측 가능한 파라미터를 우도함수와 사전분포에 적용함으로써 보다 운용적인(operational) 분석이 가능하다. 예를 들어, 본 논문에서 다루는 문제에 대해 설명하면, θ는 단위시간당 고장의 개수를 의미하는 우도함수의 파라미터이고, 이를 표현하는 사전분포(감마분포)의 상위파라미터 α와 β에 각각 관측된(혹은 주관적인) 고장의 개수와 그때까지의 시간을 부여할 수 있다. 아마도 운용적인 분석을 위해서는 해당 문제에 내재되어 있는 물리적 측면과 여러 상황들의 이해가 선행되어야 할 것이다. 예를 들어, 고장물리, 확률적 상관관계 등을 들 수 있다.
- 반적인 확률 이론에 따르면 베이지안 확률 모형은 우도함수, 사전분포, 상위파라미터 등의 선정이 타당하다면 최적의 의사결정을 지원한다. 이와는 반대로 전통적 통계방법은 최적의 의사결정에 근접한 결과만을 유도한다는 것이 알려져 있다[18].

베이지안 확률 모형을 이용하여 얻은 위험률함수의 추정치 식(12)와 식(13)을 비교하면 식(12)는 수학적으로 위험률함수 추정치의 참값임을 알 수 있다. 이와는 다르게 식(13)은 식(12)에 대한 추정치로서 아직 관측되지 않은 y가 주어졌다는 가정하에 얻은 사후분포의 기대값이다. 베이지안 확률 모형에 의해 계산된 두 개의 추정치는 아직 관측되지 않은 미래의 데이터 y와 관측된 과거 데이터들에 대한 함수이기 때문에 여러 신뢰성 관련 문제에 적용가능 할 것이다. 또한 일반적으로 수학적으로 참값인 식(12)에 의한 방법이 보다 정확한 추정을 가능하게 할 것이라고 판단된다. 그러나 여러 통계적 문제에서 그러하듯이 식(12)와 식(13) 두 가지 방법 중 무엇이 더 정확할 것인가에 대한 문제는 상황에 따라 다를 것으로 생각된다.

마지막으로 지수분포의 위험률함수는 실제로 상수 형태(constant hazard rate)이지만 식(12), 식(13), 식(15)는 모두 감소함수 형태이다. 만약 상수 형태의 추정치가 필요하다면 식(12), 식(13), 식(15)의 분모에서 y를, 분자에서 1을 제외하여 상수 형태의 추정치를 얻을 수 있다. 이는 상수 형태의 추정치를 위해서 관측된 데이터와 사전에 결정된 상위파라미터만을 사용하겠다는 의미이다.

5. 실험예제

본 장에서는 무작위로 발생시킨 데이터를 이용하여 앞에서 설명한 방법을 이용하여 위험률함수를 추정하고 이들을 정확성(precision)과 변동성(fluctuation)의 관점에서 비교한다. 실험예제를 위해 아래와 같은 사항을 가정한다.

- 장은 파라미터가 θ=0.01인 지수분포를 따르고 독립적으로 발생한다.
- 약 생존시간이 70보다 크면 중도절단데이터가 발생한다(즉, T=70).
- 도함수인 지수분포에 대한 사전분포로 감마분포를 사용하고, 2장에서 설명한 방법에 의해 계산된 α=5, β=549를 상위파라미터로 사용한다.

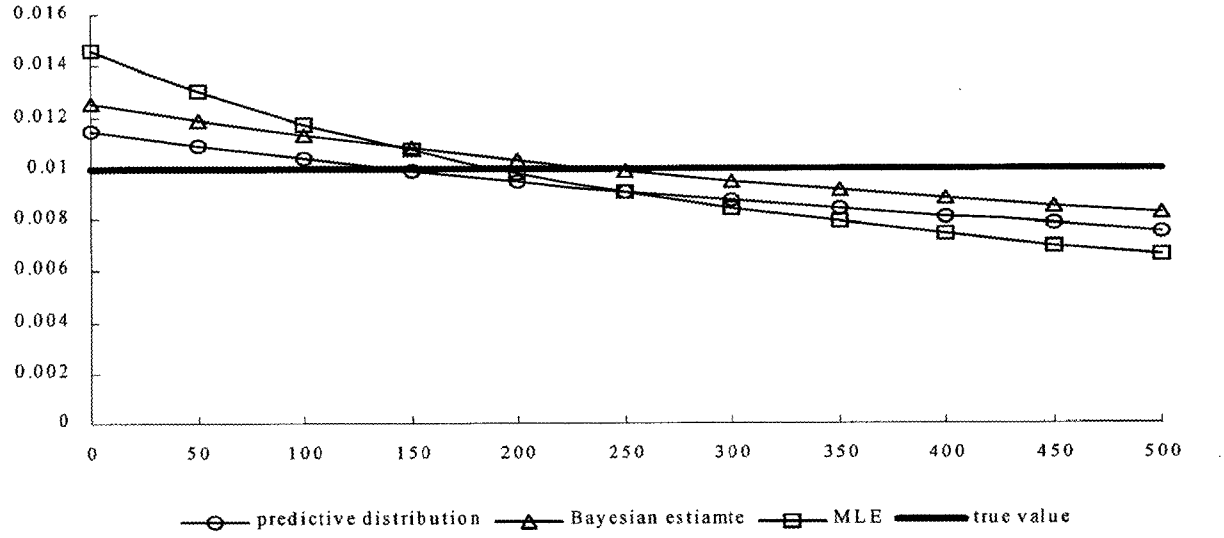
난수발생을 이용하여 4개의 중도절단데이터와 6개의 완전관측데이터로 이루어진 10개의 데이터를 아래와 같이 얻었다. 또한 별표시가 되어있는 데이터는 중도절단 데이터를 의미한다.

$$70^*,\ 70^*,\ 6,\ 11,\ 40,\ 70^*,\ 70^*,\ 18,\ 34,\ 22.$$

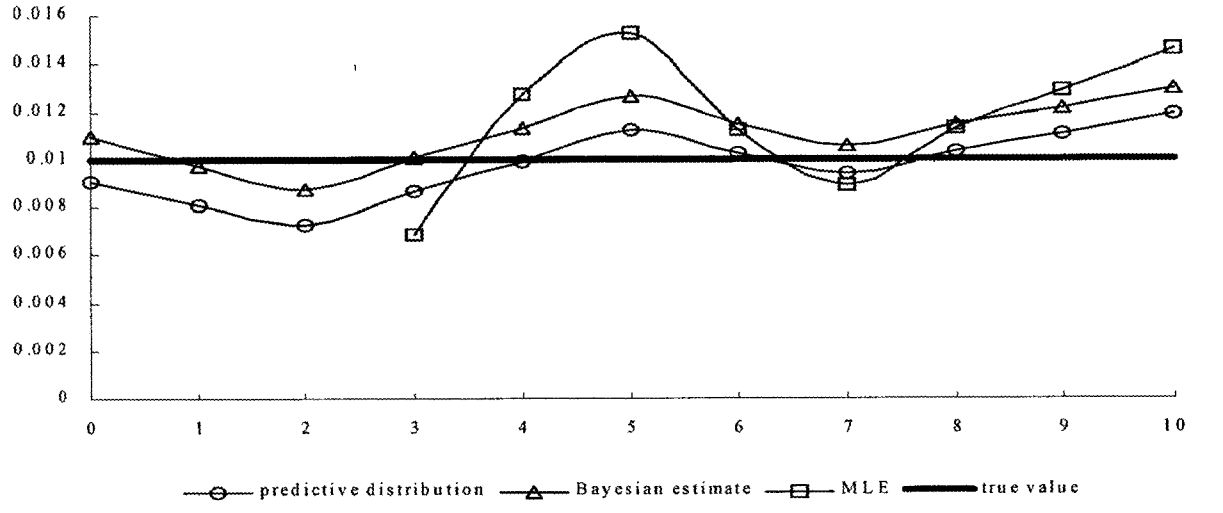
주어진 데이터와 사전분포를 식(8)에 대입하면 갱신된 사후분포 $gamma(\alpha=11,\ \beta=960)$을 얻을 수 있다. 또한 갱신된 사후분포를 식(12)와 식(13)에 대입하면, 위험률 함수의 추정치로 예측분포를 사용한 경우 $11/(960+y)$, 베이지안 추정치를 사용한 경우 $12/(960+y)$를 얻게 된다. 최대우도추정량 방법의 경우 식(15)를 이용하여 위험률 함수의 추정치로 $6/(411+y)$를 얻게 된다.

<그림 3>은 세 개의 위험률함수 추정치를 서로 비교하고 있다. <그림 3>을 통해 베이지안 확률 모형을 이용하여 얻은 위험률함수 추정치가 최대우도추정량 방법에 의한 위험률함수 추정치보다 참값 0.01에 가깝게, 즉 보다 정확한 추정치를 제공하고 있음을 알 수 있다.

또한 <그림 4>는 세 가지 방법에 의해 계산된 상수형태의 위험률함수 추정치를 보여주고 있다. <그림 4>에서 가로축은 데이터의 개수를 의미한다. <그림 4>를 보



<그림 3> 위험률함수 추정치



<그림 4> 상수 형태의 위험률함수 추정치

면 데이터가 없을 때(가로축의 0) 최대우도추정량 방법은 위험률함수 추정치를 얻을 수 없지만 베이지안 확률모형은 사전분포를 이용함으로써 데이터가 없을 때에도 위험률함수 추정치를 얻을 수 있다는 점을 알 수 있고, 두 번째 데이터를 얻을 때까지 최대우도추정량에 의한 위험률함수 추정치가 존재하지 않는 이유는 첫 번째, 두 번째 데이터가 중도절단데이터이기 때문이다. <그림 3>과 마찬가지로 참값이 0.01이기 때문에 상수 형태의 위험률함수의 경우 역시 베이지안 확률 모형이 최대우도 추정량보다 더 정확함을 알 수 있다.

또한 최대우도추정량 θ_{MLE}는, n'가 완전관측데이터의 개수일 때, 기대값이 $n'\theta/(n'-1)$이고, 분산이 $(n'\theta)^2/\{(n'-1)^2(n'-2)\}$인 역 감마분포(inverse gamma distribution)를 따른다[17]. 따라서 주어진 데이터에 대한 최대우도추정량 $\theta_{MLE} \approx 46\times10^{-2}$이고 분산은 $\mathrm{Var}(\theta_{MLE}) \approx 57\times10^{-5}$이다. 베이지안 확률 모형의 경우, 파라미터 θ는 감마분포를 따르기 때문에 θ에 대한 분산은 $\mathrm{Var}(\theta) \approx 19\times10^{-5}$이다. 분산에 대한 분석을 통해 최대우도추정량에 의한 결과가 베이지안 확률 모형에 의한 결과보다 더욱 변동이 큼을 알 수 있다. 이는 아마도 통계적 분석을 위한 데이터의 양이 적기 때문일 것이다. 또한 베이지안 확률 모형은 사전분포를 이용할 수 있기 때문에 데이터의 양이 적은 경우 최대우도추정량 방법보다 효과적임을 알 수 있다.

6. 결 론

공학시스템에서 발생하는 데이터는 여러 통계적·률적 문제를 해결하기 위해서 매우 중요하게 활용되어야 하지만, 만약 데이터가 매우 적은 경우 데이터만을 사용하는 접근법은 신뢰할 수 없을 것이다. 이러한 문제를 해결하기 위해 본 논문은 베이지안 확률 모형을 소개하였다. 또한 베이지안 확률모형의 여러 장점들은 타당한 사전분포를 이용해야 보장됨을 주의해야 할 것이다.

본 논문은 베이지안 확률 모형과 최대우도추정량 방법을 이용하여 위험률함수를 추정하고 두 방법을 비교하는데 중점을 두었다. 계산의 용이성을 위해 지수분포를 우도함수로, 감마분포를 지수분포의 자연공액사전분포로 사용하였다. 먼저 본 논문에서는 우도함수가 자연공액사전분포를 갖는 경우 시뮬레이션 방법을 이용하여 전문가의 주관적 견해를 사전분포로 표현하는, 즉 상위파라미터 추론 방법에 대해 다루었다. 제시한 상위파라미터 추론 방법은 다음의 세 가지 기본 개념을 가지고 있다. 첫째, 전문가는 주관적 견해를 표현할 때, 일반적으로 잘 관측되지 않는 파라미터에 대한 것보다 상대적으로 관측이 쉬운 실제 발생 데이터에 대한 주관적 견해를 보다 잘 표현할 수 있다. 둘째, 신뢰할 만한 우도함수의 특정 통계치를 가지고 있다면 우도함수의 특정 통계치와 상위파라미터 간의 수학적 관계를 이용하여 상위파라미터를 추론할 수 있다. 셋째, 실측 데이터가 존재하지 않는 경우라도, 전문가의 주관적 견해와 시뮬레이션 방법을 사용하여 우도함수의 특정 통계치를 쉽게 계산할 수 있다.

또한 베이지안 확률 모형과 최대우도추정량 방법을 이용하여 지수분포에 대한 위험률함수를 추정하였고, 두 가지 방법을 비교하였다. 그리고 실험예제를 통해 베이지안 확률 모형이 최대우도추정량 방법보다 효율적임을 보였다. 일반적으로 베이지안 확률 모형은 최대우도추정량 방법에 비해 다음과 같은 장점을 가지고 있다고 할 수 있다. 첫째, 베이지안 확률 모형은 통계적 분석을 위한 데이터가 충분히 존재하지 않는 경우, 사전분포를 활용함으로써 최대우도추정량 방법보다 신뢰할만한 결과를 제공할 수 있다. 둘째, 베이지안 확률 모형은 데이터로부터 얻은 정보를 체계적으로 갱신하는 즉, 학습과정을 포함한다. 셋째, 주어진 상황과 문제에 대한 물리적 측면 등을 고려하여 우도함수와 사전분포의 파라미터에 의미를 부여함으로써 운용적인 분석이 가능하다. 넷째, 일반적인 확률이론에 따르면 우도함수와 사전분포가 적절히 선택되었다면 베이지안 모형에 의한 결과는 최적해를 보장한다.

참고문헌

[1] Ang, A. H-S. and Tang, W. H. *Probability Concepts in Engineering Planning and Design*, John Wiley & Sons, Inc., New York, 1975.

[2] Arnold, B. C. "Back to Bayesics," *Journal of Statistical Planning and Inference*, Vol. 109, pp. 179-187, 2003.

[3] Barlow, R. E. and Proschan, F. *Statistical Theory of Reliability and Life Testing*, MD Silver Spring, 1981.

[4] Bosworth, K., Gingiss, P. M., Potthoff, S. and Roberts-Gray, C. "A Bayesian model to predict the success of the implementation of health and education innovations in school-centered programs," *Evaluation and Program Planning*, Vol. 22, pp. 1-11, 1999.

[5] Cagno, E., Caron, F., Mancini, M. and Ruggeri, F. "Using AHP in determining the prior distribution on gas pipeline failures in a robust Bayesian approach," *Reliability Engineering and System Safety*, Vol. 67, pp. 275-284, 2000.

[6] Chauhan, R. K. "Bayesian analysis of reliability and hazard

rate function of a mixture model," *Microelectronics Reliability*, Vol. 37, No. 6, pp. 953-941, 1997.

[7] Coolen, F. P. A. "On Bayesian reliability analysis with informative priors and censoring," *Reliability Engineering and System Safety*, Vol. 53, pp. 97-98, 1996.

[8] Davison, A. C. and Hinkley, D. V. *Bootstrap Methods and their Application*, CAMBRIDGE UNIVERSITY PRESS, Cambridge, 1997.

[9] Efron, B. and Tibshirani, R. G. *An Introduction to the Bootstrap*. CHAPMAN & HALL/CRC, Boca Raton, 1993.

[10] Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. *Bayesian Data Analysis*, Champman & Hall, New York, 2000.

[11] Hall, P. *The Bootstrap and Edgeworth Expansion*, Springer, New York, 1992.

[12] Leemis, L. M. *RELIABILITY Probabilistic Models and Statistical Methods*, Prentice-Hall, New Jersey, 1995.

[13] Lesley, W. and John, Q. "Building prior distributions to support Bayesian reliability growth modeling using expert judgement," *Reliability Engineering and System Safety*, Vol. 74, pp. 117-128, 2001.

[14] Martz, H. F. and Waller, R. A. *Bayesian Reliability Analysis*. John Wieley & Sons, Inc., New York, 1982.

[15] Migon, H. S. and Gamerman, D. *Statistical Inference : an Integrated Approach*, Arnold, LONDON, 1999.

[16] Nelson, W. "Hazard plotting of left truncated life data," *Journal of Quality Technology*, Vol. 22, pp. 230-238, 1990.

[17] Percy, D. F. "Bayesian enhanced starategic decision making for reliability," *European Journal of Operational Research*, Vol. 139, pp. 133-145, 2002.

[18] Percy, D. F., Kobbacy, K. A. H. and Fawzi, B. B. "Setting preventive maintenance schedules when data are sparse," *International Journal of Production Economics*, Vol. 51, pp. 223-234, 1997.

[19] Rai, B. and Shngh, N. "Hazard rate estimation from incomplete and unclean warranty data," *Reliability Engineering and Safety*, Vol. 81, pp. 79-92, 2003.

[20] Rosqvist, T. "Bayesian aggregation of experts' judgements on failure intensity," *Reliability Engineering and System Safety*, Vol. 70, pp. 283-289, 2000.

[21] Sharma, K. K., Krishna, H. and Singh, B. "Bayes estimation of the mixture of hazard rate model," *Reliability Engineering and System Safety*, Vol. 55, pp. 9-13, 1997.

[22] Siu, N. O. and Kelly, D. L. "Bayesian parameter estimation in probabilistic risk assessment," *Reliability Engineering and System Safety*, Vol. 62, pp. 89-116, 1998.

[23] Wilson, G. "Tides of change : Bayesianism the new paradigm in statistics?," *Journal of Statistical Planning and Inference*, Vol. 113, pp. 371-374, 2003.