The Korean Communications in Statistics **Vol. 14** No. 1, 2007, pp. 111–119

Comparison of Nonparametric Maximum Likelihood and Bayes Estimators of the Survival Function Based on Current Status Data*

Heejeong Kim[1] Yongdai Kim[2] and Young Sook Son[3]

Abstract

In this paper, we develop a nonparametric Bayesian methodology of estimating an unknown distribution function F at the given survival time with current status data under the assumption of Dirichlet process prior on F. We compare our algorithm with the nonparametric maximum likelihood estimator through application to simulated data and real data.

Keywords: Current status data; Dirichlet process prior; MCMC algorithm; Bayesian estimation; nonparametric maximum likelihood estimation.

1. 서론

현재상태자료 (current status data)는 신뢰성 (reliability), 품질공학 (quality engineering), 역학 (epidemiology), 생의약학 (biomedicine), 인구학 (demography) 등의 많은 연구 분야에서 발생되며, 구간절단자료 (interval censored data)의 특별한 형태이다. 현재상태자료는 어느 지정된 시각에 관심 있는 사건의 발생 유무만이 관측되기 때문에 사건의 정확한 발생시간은 알 수 없는 자료이다. 예를 들면, 어느 통조림 생산 공장에서는 통조림의 부패시간을 추정하기 위하여 어느 조사시점 T에서 통조림의 표본을 추출한 후, 개봉하여 통조림의 부패 여부를 관측한다. 만약 조사시점 T에서 부패가 발생된

* This work was supported in part by the Program for the Training of Graduate Students in Regional Innovation which was conducted by the Ministry of Commerce Industry and Energy of the Korean Government and in part by Grant R01-2004-000-10284-0 from the Basic Research Program of the Korea Science and Engineering Foundation.

1) Graduate Student, Department of Statistics, Chonnam National University, 300 Yongbong-Dong, Buk-Gu, Gwangju 500-757, Korea.
E-mail : 0909hehe@hanmail.net

2) Professor, Department of Statistics, Seoul National University, San 56-1 Sillim-Dong, Gwanak-Gu, Seoul 151-742, Korea.
E-mail : ydkim@stats.snu.ac.kr

3) Professor, Department of Statistics, Chonnam National University, 300 Yongbong-Dong, Buk-Gu, Gwangju 500-757, Korea.
Correspondence : ysson@chonnam.ac.kr

상태로 관측이 되면 부패의 정확한 발생시간 X는 구간 $(0,T]$의 어느 한 시점이 될 것이고, 조사시점 T에서 부패가 일어나지 않은 상태로 관측이 되면 부패의 정확한 발생시간 X는 구간 (T,∞)의 어느 한 시점이 될 것이라고만 알 수 있다.

생존분석에서는 생존시간 X의 분포함수 F에 대해서 지수 (exponential), 와이블 (Weibull), 감마 (gamma), 로그 정규 (lognormal), 로그 로지스틱 (loglogistic) 분포등과 같은 확률분포를 가정하는 모수적 접근방법이 있고, F에 대해 분포의 가정을 주지 않는 비모수적 접근방법이 있다.

Ferguson (1973)은 F에 대해 Dirichlet process prior를 가정하는 비모수적 베이지안 방법론을 제안하였다. 이때, F의 사후분포도 역시 Dirichlet process가 된다는 것을 증명하였다. Doss (1994)는 구간절단자료와 같은 불완전자료 (incomplete data)에 대하여 F의 사전분포로서 Dirichlet process의 혼합형 (mixture of Dirichlet process) 분포의 가정 하에서 깁스샘플링 (Gibbs sampling)을 사용하여 F의 사후분포를 얻는 비모수적 베이지안 방법을 제안하였다.

본 논문에서는 현재상태자료가 주어졌을 때, 생존시간 X의 알려지지 않은 분포함수 F에 대한 Dirichlet process prior의 가정 하에 시점 t에서 $F(t)$의 베이지안 추정방법을 논의하고, 얻어지는 베이즈 추정치 (Bayes estimate: BE)를 비모수적 최우추정치 (nonparametric maximum likelihood estimate: NPMLE)와 모의실험을 통하여 비교해 본다.

논문의 구성은 다음과 같다. 2절에서는 Dirichlet process prior와 결과로서 얻어지는 F의 사후분포를 소개한다. 3절에서는 현재상태자료가 주어졌을 때 주어진 시점 t에서 $F(t)$를 추정하는 MCMC 알고리즘을 제안한다. 4절에서는 금속 터빈 휠 (turbine wheel)에 금 (crack)이 발생하기까지의 시간을 측정한 현재상태자료에 제안된 알고리즘을 적용해보고, 또한 총합평균제곱오차 (integrated mean squared error: IMSE) 기준 하에서 NPMLE와의 우수성 비교를 위한 모의실험을 수행한다.

2. 분포함수 F에 대한 Dirichlet process prior와 사후분포

$X_1,\ldots,X_n$을 알려지지 않은 분포함수 F로부터의 확률표본이라고 하자. F에 대한 사전분포 (prior)로 가장 널리 사용되는 분포는 다음과 같이 정의되는 Dirichlet process prior이다 (Ferguson, 1973).

정의 2.1 $\mathcal{F}$를 R (또는 생존분석에서는 $[0,\infty)$)위에서의 모든 분포함수들의 집합이라 하자. R위에서 유한측도 (finite measure) α가 주어질 때, $\mathcal{F}$ 위의 확률측도 (probability measure) F가 임의의 유한개의 분할 (partition) $A_1,\ldots A_m$ (즉, $i\neq j$에 대해서 $A_i\bigcap A_j=\emptyset$이고, $\bigcup_{i=1}^m A_i=R$)에 대하여 $(F(A_1),\ldots,F(A_m))\sim Dir(\alpha(A_1),\ldots,\alpha(A_m))$을 만족할 때 F를 기저측도 (base measure) α를 갖는 Dirichlet process라고 한다. 여기서, $F(A_i)=\int_{A_i} dF(x)$ 이고, Dir은 Dirichlet 분포를 의미하고, $F\sim Dir(\alpha)$라고 나타낸다.

Dirichlet process의 존재, 표본의 성질 등, 몇몇 이론적 특징들은 Ghosh와 Ramamoorthi (2003)에서 확인할 수 있다.

만약 $F \sim Dir(\alpha)$이면, 모든 $t \in R$에 대하여,

$$E(F(t)) = \frac{\alpha(-\infty, t]}{\alpha(R)}$$

이고,

$$\mathrm{Var}(F(t)) = \frac{\alpha(-\infty, t]\alpha(t, \infty]}{\{\alpha(-\infty, t] + \alpha(t, \infty]\}^2\{\alpha(-\infty, t] + \alpha(t, \infty) + 1\}}$$

이다. 이것은 주어진 t에 대하여, $F(t) \sim Beta(\alpha(-\infty, t], \alpha(t, \infty))$이기 때문이다. 여기서, α를 $c = \alpha(R)$과 $H(t) = \alpha(-\infty, t]/\alpha(R)$로 다시 표현하면, $H(t)$는 $F(t)$의 기대값이 되고, $F(t)$의 분산은 c에 반비례하게 된다. 특히, $c \to \infty$인 경우는 $\mathrm{Var}(F(t)) \to 0$가 되어 $F(t)$는 $H(t)$에 확률적으로 수렴한다. 이런 의미에서, 모수 c는 정도모수 (precision parameter)라고 불려진다. 앞으로 $Dir(\alpha)$와 $Dir(c, H)$를 같은 의미로 사용한다.

이제 $F \sim Dir(c, H)$를 가정하자. Ferguson (1973)은 $X_1, \ldots, X_n$이 주어졌을 때 F의 사후분포가 다시 기저 측도 α^p,

$$\alpha^p(\cdot) = \alpha(\cdot) + \sum_{i=1}^{n} \delta_{X_i}(\cdot) \tag{2.1}$$

를 갖는 Dirichlet process가 된다는 것을 보였다. 따라서 사후분포의 평균 H^p는

$$H^p(t) = \frac{c}{c+n} H(t) + \frac{n}{c+n} F_n(t) \tag{2.2}$$

로 주어지게 되고, 정도모수는 $c^p = c + n$이 된다. 여기서 $F_n(t) = \sum_{i=1}^{n} I(X_i \le t)/n$은 경험적 분포함수 (empirical distribution)이다. 이때, c는 사전 샘플 크기 (prior sample size)로서 고려할 수 있으므로 베이즈 추정치는 사전분포의 평균 (prior mean)과 경험적 분포함수의 가중평균임을 주목한다. 따라서 $n \to \infty$이면, 베이즈 추정치는 F_n의 극한분포인 실제 분포함수에 수렴하게 된다.

F의 사후분포와 함께 흔히 관심 있는 F의 함수들, 즉, 중위수 (median), 분위수 (quantile), 또는 분산 (variance) 등의 사후분포는 일반적으로 명백한 산술식으로 존재하지 않는다. 그러나, 우리는 간단한 몬테 카를로 (Monte Carlo) 방법을 이용하여 이들의 사후분포를 얻을 수 있다. Z:$\mathcal{F} \to R$를 주어진 함수라고 하자. 그러면, 먼저 사후 분포로부터 F를 생성하고, $Z(F)$를 계산하면 된다. 이 과정을 여러 번 반복하여 생성되는 $Z(F)$의 샘플들로부터 원하는 함수의 사후분포를 계산할 수 있다.

3. 현재상태자료에 대한 분포함수 F의 베이즈 추정

현재상태자료의 경우에 생존시간 X_i는 정확히 관측되지 못하고 대신 X_i가 관측시점 T_i보다 더 큰지 혹은 더 작은지 만을 관측할 수 있다. 알려지지 않은 생존시간들을 $\boldsymbol{X} = (X_1, \ldots, X_n)$, 실제 관측치들을 $D = \{(T_1, \delta_1), \ldots, (T_n, \delta_n)\}$라 놓자. 여기서 T_i들은 관측시간이고 $\delta_i = I(X_i \le T_i)$은 절단지시변수 (censoring indicator variable)이다. 만약

$\delta_i = 1$이면, X_i는 관측시점 T_i보다 작거나 같고, $\delta_i = 0$이면, X_i는 절단시간 (censoring time) T_i보다 더 크다는 것을 알 수 있다. 이제 관측자료 D가 주어졌을 때 F를 추정 (베이지안통계에서는 F의 사후분포를 유도하는 것)하는 것이다. 관심 있는 함수들의 사후분포 식을 구하기는 매우 어렵지만, Doss (1994)의 MCMC 알고리즘을 이용하여 F의 사후분포로부터 F의 표본들을 생성할 수 있다. Doss (1994)의 알고리즘은 F의 사후분포 $\mathcal{L}(F|D)$대신 F와 $\boldsymbol{X}$의 결합 분포인 $\mathcal{L}(F, \boldsymbol{X}|D)$로부터 깁스샘플링에 의해 F의 사후분포를 얻는다. 이때, 완전 조건부 분포 (full conditional distribution)인 $\mathcal{L}(F|\boldsymbol{X}, D)$로부터 Sethuraman (1994)의 알고리즘을 사용하여 F의 sample path를 생성한다.

본 논문에서는 주어진 시점 t에서 $F(t)$의 추정을 위해 보다 간단한 MCMC 알고리즘으로 다음의 혼합알고리즘 (composition algorithm)을 사용한다.

$$\mathcal{L}(F|D) = \int_{R^n} \mathcal{L}(F|\boldsymbol{X}, D)\mathcal{L}(d\boldsymbol{X}|D).$$

이 알고리즘에서는 먼저 $\mathcal{L}(\boldsymbol{X}|D)$로부터 $\boldsymbol{X}$를 생성하고, 식 (2.1)의 기저측도 α^p를 갖는 Dirichlet process인 $\mathcal{L}(F|\boldsymbol{X}, D)$로부터 F를 생성하여 관심있는 F의 함수들을 계산할 수 있다. 한편, $\mathcal{L}(\boldsymbol{X}|D)$로부터 $\boldsymbol{X}$의 생성은 깁스샘플링을 이용할 수 있다. 즉, $\boldsymbol{X}_{-k}$를 $\boldsymbol{X}$의 k번째 확률변수를 제외하고 얻어진 $(n-1)$차원 부분벡터라고 놓고, 완전 조건부 분포 $\mathcal{L}(X_k|\boldsymbol{X}_{-k}, D)(k = 1, \ldots, n)$로부터 마코브체인 (Markov chain)이 안정적인 분포로 수렴할 때까지 반복하여 $\boldsymbol{X}$를 생성한다.

이제 분포 $\mathcal{L}(X_k|\boldsymbol{X}_{-k}, D)$를 구해 보자. 먼저 다음과 같은 식은 쉽게 정의된다.

$$\mathcal{L}(X_k \le t|F, D) = \begin{cases} \dfrac{F(t)}{F(T_k)}, & t \in [0, T_k], \quad \text{if } \delta_k = 1, \\ \dfrac{F(t) - F(T_k)}{1 - F(T_k)}, & t \in (T_k, \infty), \text{ if } \delta_k = 0. \end{cases} \tag{3.1}$$

만약, $F \sim Dir(\alpha)$이면, 임의의 $0 < s < t$에 대하여, $(F(s), F(t) - F(s), 1 - F(t)) \sim Dir(\alpha(0, s], \alpha(s, t], \alpha(t, \infty))$이므로 다음의 결과를 얻을 수 있다.

$$\begin{cases} \dfrac{F(s)}{F(t)} & \sim Beta(\alpha(0, s], \alpha(s, t]) \\ \dfrac{F(t) - F(s)}{1 - F(s)} & \sim Beta(\alpha(s, t], \alpha(t, \infty)). \end{cases} \tag{3.2}$$

한편, $\mathcal{L}(X_k|\boldsymbol{X}_{-k}, D)$는 다음과 같이 쓸 수 있다.

$$\mathcal{L}(X_k|\boldsymbol{X}_{-k}, D) = \int_{F \in \mathcal{F}} \mathcal{L}(X_k|F, D)\mathcal{L}(dF|\boldsymbol{X}_{-k}). \tag{3.3}$$

식 (3.1), (3.2), 그리고 $\mathcal{L}(F|\boldsymbol{X}_{-k})$가 기저측도 $\alpha(\cdot) + \sum_{i \ne k}^{n} \delta_{X_i}(\cdot)$를 가지는 Dirichlet process라는 사실을 이용하면 식 (3.3)은 다음과 같은 식 (3.4)의 결론에 이른다.

$$\mathcal{L}(X_k \le t|\boldsymbol{X}_{-k}, D) = \begin{cases} \dfrac{\alpha(0,t] + \sum_{i\neq k}\delta_{X_i}(0,t]}{\alpha(0,T_k] + \sum_{i\neq k}\delta_{X_i}(0,T_k]}, & t \in (0,T_k], \quad \text{if } \delta_k = 1, \\ 1 - \dfrac{\alpha(t,\infty) + \sum_{i\neq k}\delta_{X_i}(t,\infty)}{\alpha(T_k,\infty) + \sum_{i\neq k}\delta_{X_i}(T_k,\infty)}, & t \in (T_k,\infty), \text{ if } \delta_k = 0. \end{cases} \tag{3.4}$$

이제 다음에 주목하자.

$$E(F(t)|D) = E(E(F(t)|\boldsymbol{X}, D)) \quad \text{그리고} \quad E(F(t)|\boldsymbol{X}, D) = H^p(t).$$

결국 깁스샘플링에 의해 $\mathcal{L}(\boldsymbol{X}|D)$로부터 생성된 $\boldsymbol{X}$를 사용하여 반복해서 얻어지는 $H^p(t)$의 평균이 시점 t에서 분포함수 값 $F(t)$의 베이즈 추정량이다. 이 알고리즘은 Dirichlet process의 표본을 생성할 필요가 없다는 점에서 효율적인 알고리즘이라고 할 수 있다.

4. 수치분석

4.1. 모의실험

이제 본 논문에서 제안된 알고리즘을 평가하기 위하여 모의실험을 수행해 본다. 실제 생존시간 $\{X_i\}$는 형태모수 $a = 0.5, 1, 2$와 척도모수 $b = 2, 5, 10$을 갖는 와이블 분포로부터 생성하였다. 와이블 생존분포는 형태모수 $a = 1$ (이 경우의 와이블분포는 지수분포에 해당한다.)을 기준으로 위험함수 (hazard function)의 형태가 달라진다. 즉 $a = 0.5, 1, 2$일 때의 위험함수는 각각 감소함수, 상수, 그리고 증가함수의 형태를 갖는다. 관측시점 $\{T_i\}$은 각 모형에서 평균에 해당하는 모수값을 가지는 지수분포로부터 크기가 각각 $n = 10, 30, 50, 100$인 표본을 $M(= 100)$회 반복하여 생성하였다. 그림 4.1은 실험에 사용된 모수별 분포의 형태를 보여주고 있다. 또한, 생성된 각 생존시간 X_i가 관측시점 T_i보다 작거나 같으면 $\delta_i = 1$로, 그렇지 않으면, $\delta_i = 0$로 지정하였다.

베이지안추정에서 $H(t)$는 모수 θ를 갖는 지수분포로 가정하고 (θ는 지수분포 평균의 역수), 관측 자료에 대한 지수분포의 적합에서 최우 추정된 θ값을 모수로 갖는 절단 지수분포 (truncated exponential distribution)로부터 생성된 생존시간들의 평균의 역수를 θ의 값으로, 생존 시간들의 평균을 $\alpha(R)$의 값으로 사용하는 경험적 (empirical) 베이즈 추정방법을 사용하였다. 초모수 (hyperparameter)의 결정을 사전정보가 아닌 자료로부터 결정하는 경험적 베이지안 방법의 적용에 대하여 흔히 베이지안들이 비판적이긴 하지만 본 논문에서는 베이즈 추정량을 계산하는 하나의 방법으로서 제안한다. 참고로 Doss (1994)는 초모수 θ가 감마분포를 한다는 계층적 (hierarchial) 사전분포를 가정하였다.

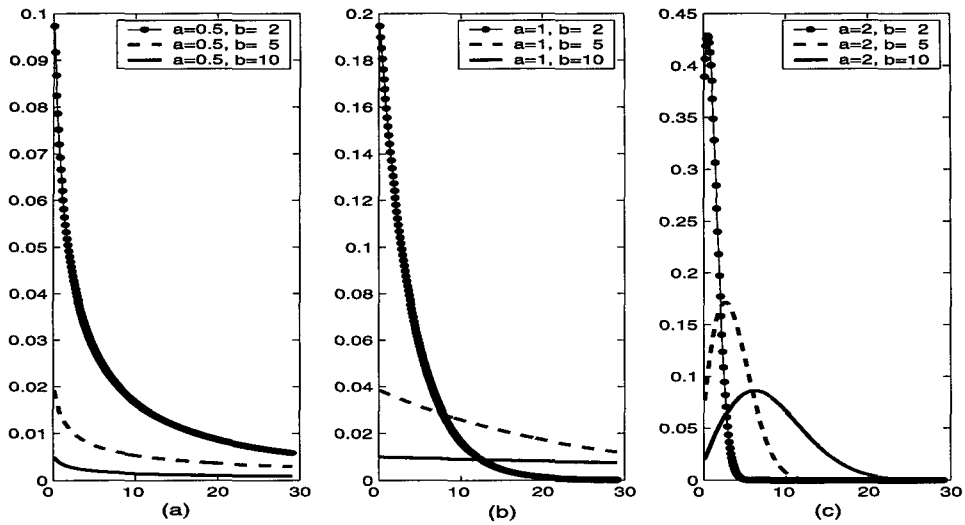

그림 4.1: 와이블분포 $(a) = 0.5$ $(b)a = 1$ $(c)a = 2$

표 4.1: $F(t)$에 대한 베이즈 추정치와 NPMLE의 IMSE 비교

X의 분포		NPMLE			Bayes estimate		
	n	$b=2$	$b=5$	$b=10$	$b=2$	$b=5$	$b=10$
	10	0.03417	0.02018	0.00143	0.01369	0.01603	0.00087
weibull	30	0.04472	0.02018	0.00145	0.02529	0.01373	0.00023
$(a=0.5)$	50	0.01878	0.01651	0.00143	0.01363	0.01496	0.00035
	100	0.01212	0.02017	0.00144	0.01446	0.01514	0.00038
	n	$b=2$	$b=5$	$b=10$	$b=2$	$b=5$	$b=10$
	10	0.00555	0.01797	0.02278	0.00123	0.00389	0.00670
weibull	30	0.00487	0.01099	0.01627	0.00118	0.00374	0.00079
$(a=1)$	50	0.00848	0.00894	0.01326	0.00275	0.00354	0.00227
	100	0.00234	0.00832	0.01220	0.00027	0.00045	0.00209
	n	$b=2$	$b=5$	$b=10$	$b=2$	$b=5$	$b=10$
	10	0.00724	0.03250	0.04756	0.00432	0.00770	0.00670
weibull	30	0.00936	0.00628	0.02941	0.00666	0.00454	0.00425
$(a=2)$	50	0.00327	0.00858	0.01304	0.00236	0.00383	0.00449
	100	0.00236	0.00351	0.00498	0.00204	0.00305	0.00327

비교를 위한 비모수적 최우추정량 NPMLE는 명백한 수식으로 다음과 같이 주어진다 (Lawless, 2003, p126): $s_1, s_2, \ldots, s_{k-1}$를 관측시점 $\{T_i, T_2, \ldots, T_n\}$중 서로 다른 시점들로서 $0 = s_0 < s_1 < s_2 < \cdots < s_{k-1} < s_k = \infty$을 만족한다고 놓을 때,

$$\hat{F}(s_j) = \max_{u \le j} \min_{v \ge j} \Big(\sum_{l=u}^{v} d_l / \sum_{l=u}^{v} n_l \Big), \quad j = 1, 2, \ldots, k-1,$$

여기서 $d_l = \sum_{i=1}^{n} I(X_i \le T_i, T_i = s_l)$, $n_l = \sum_{i=1}^{n} I(T_i = s_l)$이다. 비교를 위한 측도로서 다음과 같은 총합평균제곱오차 (integrated mean squared error: IMSE)를 정의한다.

$$IMSE = \frac{1}{M} \sum_{k=1}^{M} \frac{1}{N} \sum_{i=1}^{N} \big(\hat{F}^{(k)}(t_i) - F(t_i) \big)^2.$$

여기서 $F(t_i)$는 반복표본을 생성케 했던 이론적 분포의 시점 t_i에서의 분포함수의 값을 의미하고, $\hat{F}^{(k)}(t_i)$는 k번째 반복표본에 대해 시점 t_i에서 $F(t_i)$의 추정된 값을 의미한다. $t_1, t_2, \ldots, t_N$는 M =100개의 모든 반복표본에 대하여 구간 $[0, 30]$을 1 단위로 나눈 고정된 값들을 사용하였다.

한편, 베이지안추정을 위한 깁스샘플링에서는 처음 100번까지는 생성된 표본을 버린 후 반복회수를 100회부터 시작하여 100회씩 증가시킨 IMSE의 결과들을 살펴보는 것으로 깁스샘플링의 수렴성을 검토하였으며 최종적으로 1,000개의 반복으로 얻어지는 표본을 추정에 사용하였다. 표 4.1에는 자료가 생성된 각 모형별로 제안된 베이즈 추정치와 NPMLE의 IMSE가 계산되어 있다. $n = 100$, $a = 0.5$, $b = 2$의 경우를 제외하고 대부분의 경우에서 베이즈 추정치가 NPMLE보다 IMSE기준에서 더 우수하다고 보여진다.

4.2. 실제 자료분석

이제 본 논문에서 논의된 알고리즘을 실제 자료 (real data)에 적용해보기로 한다. 사용된 자료는 Nelson (1982)과 Meeker와 Escobar (1998)가 수행한 금속 터빈 휠에서 금 (crack)이 발생하기까지의 시간에 대한 연구에서 사용된 자료로서 표 4.2에 제시된 현재상태자료이다.

이제 F에 대하여 Dirichlet Process Prior의 가정 하에 본 논문에서 제안된 베이지안 방법을 적용하여 주어진 자료를 분석해 보기로 하자. 여기서 $H(t)$는 와이블 분포로 가정하고, 분포에 포함된 형상모수와 척도모수는 다음과 같은 경험적 베이즈 추정방법에 의해 결정되었다. 즉, 관측 자료에 대한 와이블 분포의 적합에서 최우 추정된 모수들을 갖는 절단 와이블 분포 (truncated Weibulll distribution)로 부터 생성된 생존시간들에 대해 와이블 분포의 적합에서 최우추정을 통하여 얻어지는 추정치들을 사용하였다. 또한, $\alpha(R) = 1,50$인 경우에 대해 분석하였다. $\boldsymbol{X}$를 생성하기 위하여 100개의 독립적인 각 깁스샘플러에서 처음 100번까지는 생성된 표본을 버린 후 반복회수를 10회부터 시작하여 10회씩 증가시킨 후의 $\hat{F}(t)$의 결과들을 살펴보면서 깁스샘플링의 수렴성을 검토하여 최종적으로 100회의 반복으로 얻어지는 표본을 추정에 사용하였다. 그리고 편의상 관측시점 $t = 4, 10, 14, 18, 22, 26, 30, 34, 38, 42, 46, 60$에서 $F(t)$를 추정하였다. 표 4.2 및 그림 4.2에는 $F(t)$의 NPMLE, 베이즈 추정치 (BE), 그리고 모수적 방법으로서 와이블 분포의 가정 하에서의 MLE가 제시되어 있다. 와이블 분포의 추정된 MLE는 형태모수는 2.176, 척도모수는 46.78 이다 (Lawless, 2003, p.177).

표 4.2: 현재상태자료와 $F(t)$에 대한 베이즈 추정치 (BE), NPMLE, MLE의 비교

현재상태자료			$\hat{F}(t)(s.d(\hat{F}(t)))$			
t	금이 간 휠의 수	금이 가지 않은 휠의 수	NPMLE	$BE_{\alpha(R)=1}$	$BE_{\alpha(R)=50}$	MLE
4	0	39	0.000(0.000)	0.005(0.015)	0.031(0.012)	0.005
10	4	49	0.070(0.027)	0.088(0.028)	0.083(0.015)	0.034
14	2	31	0.070(0.027)	0.093(0.027)	0.113(0.018)	0.070
18	7	66	0.096(0.034)	0.107(0.028)	0.148(0.018)	0.118
22	5	25	0.167(0.068)	0.155(0.052)	0.190(0.024)	0.176
26	9	30	0.222(0.046)	0.197(0.049)	0.243(0.033)	0.243
30	9	33	0.222(0.046)	0.218(0.055)	0.302(0.043)	0.316
34	6	7	0.462(0.138)	0.443(0.148)	0.408(0.047)	0.393
38	22	12	0.581(0.057)	0.566(0.043)	0.479(0.035)	0.471
42	21	19	0.581(0.057)	0.569(0.044)	0.510(0.035)	0.547
46	21	15	0.583(0.082)	0.579(0.043)	0.535(0.033)	0.619
60			1.000	0.643(0.113)	0.613(0.042)	0.821

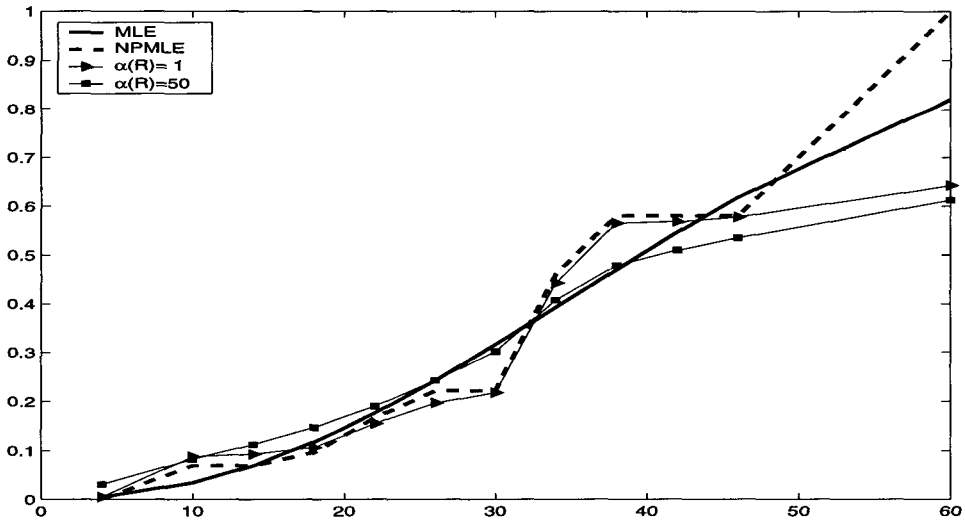

그림 4.2: $F(t)$에 대한 베이즈 추정치 (BE), NPMLE, MLE의 비교

표 4.2에서 보듯이 NPMLE는 관측자료가 존재하는 시점 ($t = 14, 30, 42$)에서도 확률의 증가가 없는 반면, 베이지안 결과는 관측자료의 정보가 반영되어 매 구간마다 추정량이 조금씩 증가하고 있음을 확인 할 수 있다. 또한, NPMLE의 결과에서는 시점 4에서 추정 확률이 0이어서 사건이 발생하지 않는 것으로 결과가 얻어졌다. 그러나, 자료에서는 시점 4를 포함한 시점 10 이전에 사건이 4회 발생한 것으로 관측되었으므로 $F(4)$에서

값을 갖는 것이 더 타당한 것으로 생각된다. 위의 자료는 비교를 위해 NPMLE와 동일한 시점에서 계산한 것이나, 베이지안 방법은 자료가 관측된 시점뿐만 아니라, 관측시점이 아닌 임의의 시점 t에서도 $F(t)$를 추정할 수 있다는 장점을 갖는다. 실제로, $t = 60$을 기준으로 $F(60)$, 즉 시점 60까지 사건이 발생할 확률이 추정되었다. 그러나 NPMLE는 관측시점끼리 단순히 이어주는 형태 (혹은 관측시점에서만 점프를 갖는 계단함수)를 가지며 마지막 관측시점 이후에는 일괄적으로 1의 추정값을 갖는다. 이러한 사실들은 앞선 모의실험에서 베이즈 추정치가 NPMLE보다 IMSE기준하에서 더 우수한 근거가 된다고 생각된다.

그림 4.2에서 베이즈 추정치는 $\alpha(R)$을 1의 작은 값을 주었을 때는 자료가 관측된 시점을 기준으로 NPMLE의 결과와 유사한 반면, $\alpha(R)$을 50의 큰 값을 주었을 때는 와이블 분포를 가정한 모수적 결과와 보다 유사한 것을 확인 할 수 있다.

5. 맺음말

본 논문에서는 현재상태자료가 주어졌을 때, F에 대한 Dirichlet process prior의 가정 하에 생존시간 $\boldsymbol{X}$의 알려지지 않은 분포함수 F의 베이즈 추정량을 구하는 방법을 제안하고, 모의실험 및 자료분석을 통하여 베이즈 추정량을 비모수적 최우추정량과 비교해 보았다. 본 논문에서 제안된 추정방법은 F의 사후분포인 Dirichlet process로부터 표본을 생성할 필요가 없다는 점에서 효율적이라고 할 수 있다. 또한, 비모수적 최우추정치와 비교하는 모의실험에서 총합평균제곱오차 기준으로 볼 때 베이즈 추정치가 더 우수함을 볼 수 있었다. 보다 포괄적인 자료의 형태인 구간절단자료에 대한 적용을 향후 연구에서 기대해 본다.

참고문헌

Doss, H. (1994). Bayesian nonparametric estimation for incomplete data via successive substitution sampling. *The Annals of Statistics*, **22**, 1763–1786.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, **1**, 209–230.

Ghosh, J. K. and Ramamoorthi. R. V. (2003). *Bayesian Nonparametrics.* Springer, New York.

Lawless, J. F. (2003). *Statistical Models and Methods for Lifetime Data.* 2nd ed, John Wiley & Sons, New York.

Meeker, W. Q. and Escobar, L. A. (1998). *Statistical Methods for Reliability Data.* John Wiley & Sons, New York.

Nelson, W. B. (1982). *Applied Life Data Analysis.* John Wiley & Sons, New York.

Sethuraman, J. (1994). A constructive definition of dirichlet priors. *Statistica Sinica*, **4**, 639–650.

[Received October 2006, Accepted December 2006]