응용통계연구 제 15권 1호, 2002년, pp. 57-71

지수 생존 모형에서의 베이지안 모형 선택

정 윤 식 [1] 김 미 숙 [2]

요 약

본 논문에서는 지수생존 모형의 형태들로써 단순 지수모형, 변환 점 지수모형과 유한 혼합 지수모형 등 세 가지 모형을 소개한다. 이러한 모형들 중에서, 최적의 모형을 찾기 위하여 Gelfand와 Ghosh(1998)의 방법을 이용한 모형 선택 방법을 제안한다. 이때, 계산상의 어려움을 피하기 위하여 자료 확장 기법(Tanner와 Wong, 1987)과 깁스 샘플러(Gelfand와 Smith, 1990)를 사용하였다. 제안된 베이지안 방법을 설명하기 위하여 모의 실험 자료와 Stangl의 항 우울제 자료에 적용한다. 모형 선택 방법은 사전 분포와 모형 선택 기준의 가중치에 민감하지 않다는 것을 제한된 우리의 실험으로 알 수 있었다.

주요용어: 깁스 샘플러; 변환점 모형; 베이지안 모형선택; 균형손실 함수; 유한 혼합모형; 잠재 변수.

1. 서론

임상 실험이나 임상 연구는 환자나 관심이 되는 대상에게 투약되는 새로운 약같이 하나 이상의 의학적 치료의 안전성과 효능을 평가하기 위하여 설계된 실험을 의미한다. 질병, 사망 혹은 다른 심각한 사건이 관심의 대상이 된다. 종종, 임상 실험에서는 질병에 대한 두 가지 다른 치료법의 효능을 비교하는 것이 주된 목표이다. 대개 연구자들은 기존의 처방으로 치료를 받은 환자와 실험되는 처방으로 치료를 받은 환자 각각을 비교하게 된다. 또한 연구자들은 기존의 약과 새로이 개발된 약의 효능을 비교한다. 때때로, 몇 몇의 대상들은 흥미가 있는 사건을 경험하지 않을 수도 있다. 이 경우 이러한 대상들을 우리는 중도절단 되었다고 한다. 중도절단은 실험의 대상이 통계적 분석이 행하여 질 수 있는 임상 연구가 끝날 때까지도 관심이 되는 사건을 경험하지 못했을 경우, 임상 연구 중간에 누락이 되는 경우, 그리고 다른 곳으로 연구 대상이 이주하였거나 연구를 더 이상 진행할 수 없을 경우 일어나게 된다. 그래서, 임상 자료를 분석할 때는 중도 절단된 자료를 어떻게 처리해야하는지를 결정해야만 한다.

역사적으로, 생존 분석은 추정, 신뢰구간 그리고 가설 검정과 같이 고전적인 개념을 적용해왔다. Nelson(1982)이 공학적 응용을 연구하였고 반면에 Cox와 Oakes(1984)가 임상 자료에 관심을 보였다. Kalbfleisch와 Prentice(1980)가 비례 위험 방법에 특별히 관심을 보이며 연구를 확장했다.

1) (609-735) 부산시 금정구 장전동, 부산대학교 자연과학대학 통계학과 부교수
yschung@hyowon.pusan.ac.kr
2) (100-666) 서울시 중구 을지로1가 87번지, 삼성카드 주식회사 신용관리 팀

지난 과거 수십년 동안, 생존 분석에 베이지안 방법을 적용시키는데 많은 관심이 증가되어왔다. 예를 들면, Martz와 Waller(1982)는 베이지안 접근에 초점을 두었다. 90년대에 들어와 깁스 샘플러(Gelfand와 Smith, 1990)를 이용한 복잡한 베이지안 통계 계산이 빠른 속도로 발달하여짐에 따라 베이지안 방법이 생존 분석과 같이 복잡한 모형에 잘 적용되어 오고 있다. 최근에, Qian(1994)은 베이지안 와이블 생존 모형을 연구했다. 특히, Stangl과 Greenhouse (1998)는 다 기관 임상 실험에서 임상 기관들 사이에서의 치료 효과에 대한 이질성을 모형화 하는데 베이지안 위계지수, 혼합 그리고 변환 점 모형을 사용하여 연구했다.

본 논문에서는 베이지안 지수 모형들을 사용할 것이다. 각각의 치료에서 환자의 생존 시간은 지수 분포를 따른다고 가정한다. 즉, 만약 확률변수 T가 모수 θ를 가지고 지수 분포를 따른다고 하면, 확률 밀도 함수는

$$f(t|\theta) = \theta e^{-\theta t}, \ \ t > 0,$$

이고 간단히 다음과 같이 쓰기로 한다.

$$T \sim Exp(t|\theta).$$

지금부터 세 가지 형태의 지수 모형 즉, 지수 모형, 변환 점 모형 그리고 유한 혼합 모형 등을 다루어 보고자 한다. 이러한 지수 모형들 중에서 모형 선택을 위하여 Gelfand와 Ghosh(1998)가 제안한 모형 선택에 기인한 모형 선택 방법을 제안한다. 사후 평균, 사후 분산, 생존 분포, 위험률 그리고 사후 주변 밀도 함수와 같은 관심 있는 사후 특성들을 깁스샘플러(Gibbs sampler)를 이용하여 연구할 것이다.

본 논문의 나머지 부분은 다음과 같이 구성되어진다. 2장에서는 지수 생존 모형의 세 가지 형태가 소개된다. 각각의 모형에서 우도함수가 제시된다. 3장에서는 베이지안 계산의 어려움을 피하기 위하여 깁스 샘플러와 Tanner와 Wong(1987)에 의해서 소개된 자료 확장 알고리즘이 사용되어진다. 4장에서는 모형을 선택하는 방법을 제안한다. 이 방법은 Gelfand와 Ghosh(1998)에 의해서 제안 방법을 본 모형에 적용한 후 제시된 항들의 의미와 성질을 설명한다. 5장에서는 제안된 모형 선택 방법을 적용하기 위하여 하나의 모의 실험 자료와 Stangl(1991)의 실제 자료를 소개한다. 또한 모수의 값의 변화에 따라 우리가 얻은 결과가 얼마나 민감한지를 살펴본다.

2. 지수 생존 모형들

이 장에서 우리는 세 가지 형태의 지수 생존 모형, 즉 지수 모형, 변환 점 모형 그리고 유한 혼합 모형등을 소개하고자 한다. 사건이 일어날 때까지의 시간 T를 모형화 하기 위하여, 몇 가지 함수에 대한 정의가 필요하다. 생존 함수는

$$S(t|\theta) = P(T > t|\theta),$$

이고 시간 t까지 생존한 환자의 갑작스러운 사망률을 의미하는 위험 함수는 다음과 같다.

$$h(t|\theta) = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t | T \ge t)}{\Delta t} = \frac{f(t|\theta)}{S(t|\theta)}$$

여기에서 $f(t|\theta)$는 $S(t|\theta)$에 대한 확률 밀도 함수이고 누적 위험 함수는

$$H(t|\theta) = \int_0^t h(u|\theta)du$$

이다. 우도 함수 $L(t|\theta)$를 구하기 위하여 지표함수 δ_j를 정의하자:

$$\delta_j = \begin{cases} 1 & j\text{번째 대상이 중도 절단 되지 않은 경우} \\ 0 & j\text{번째 대상이 중도 절단된 경우.} \end{cases}$$

지금부터 처음 d명은 중도 절단 되지 않았고 나머지 $(n-d)$명은 절단된 형태로 재배치 하였다고 가정한다. 따라서, 자료에 대한 우도 함수는 사건이 일어나는 시간에 대한 밀도 함수와 생존 분포 함수의 곱에 비례하는 다음과 같은 형태로 표현할 수 있다.

$$L(\mathbf{t}|\theta) \propto \prod_{j=1}^{n} f(t_j|\theta)^{\delta_j} S(t_j|\theta)^{1-\delta_j} \propto \prod_{j=1}^{d} f(t_j|\theta) \prod_{j=d+1}^{n} S(t_j|\theta). \tag{2.1}$$

2.1. 지수 모형

만약 생존 시간 T가 지수 분포 $Exp(t|\theta)$를 가진다면 위험 함수는 상수이다. 즉, $h(t|\theta) = \theta$이고, 그것의 생존 분포는 $S(t|\theta) = \exp(-\theta t)$로 여기에서 θ는 척도 모수이며 양수이다. 누적 분포는 $F(t|\theta) = 1 - \exp(\theta t)$이고 식(2.1)에 주어진 우도 함수는 다음과 같은 형태로 쓰여질 수 있다:

$$L(\mathbf{t}|\theta) \propto \prod_{j=1}^{n} [\theta \exp(-\theta t_j)]^{\delta_j} [\exp(-\theta t_j)]^{1-\delta_j}. \tag{2.2}$$

(2.2)식에 소개된 모형을 앞으로 소개될 모형들과 구별하고자 단순 지수모형이라 하자.

2.2. 변환 점 모형

환자의 생존 시간은 모두 독립이지만 생존 시간 T의 분포는 시간이 변함에 따라 하나 이상의 점에서 변한다고 가정하자. 즉, 처음 실패는 하나의 실패율로 일어나고 두 번째 실패는 또 다른 실패율로 일어나는 등과 같은 형태이다. 전통적인 견지로는 Nguyen, Rogers와 Walker(1997)가 상수 위험률 모형에서의 변환 점이 하나인 경우를 연구하였다. 최근에는 Ebrahimi, Gelfand, Ghosh와 Ghosh(1997)가 변환 점 문제에 있어서 베이지안 접근을 시도하였다. 더욱이 Chung, Jeong와 Han(1999)은 Ebrahimi외 (1997)가 제시한 방법들을 확장하여 다중 변환 점 모형들중에서 최적의 모형을 찾기 위하여 베이즈 인자를 사용하였다. 여기서는 모형에 k개의 변환 점이 있다고 가정한다. 그러면 위험 함수는 다음과 같이 표현되며

$$\begin{aligned} h(t) &= h_1(t)I(0 \le t \le \tau_1) + h_2(t)I(\tau_1 < t \le \tau_2) \\ &\quad + \cdots + h_k(t)I(\tau_{k-1} < t \le \tau_k) + h_{k+1}(t)I(t > \tau_k), \end{aligned} \tag{2.3}$$

여기서 $\tau = (\tau_1, \cdots, \tau_k)'$는 $\tau \in (R^+)^k$인 변환 점 모수 벡터이고,

$$I(A) = \begin{cases} 1 & x \in A \text{ 인 경우} \\ 0 & \text{그 외 경우} \end{cases}$$

이다. $\tau_l \le t < \tau_{l+1}$인 경우 누적 위험 함수는 다음과 같이 정의되며

$$\begin{aligned} H(t) &= \int_0^t h(u)du \\ &= H_1[\min(t, \tau_1)] + \sum_{i=1}^{l} [H_i(\tau_i) - H_i(\tau_{i-1})]^+ + [H_{l+1}(t) - H_{l+1}(\tau_l)]^+, \end{aligned}$$

여기서 $H_i(t)$는 $\tau_{i-1} < t \le \tau_i$에서의 누적 위험 함수이고 $H_1(\tau_1) - H_1(\tau_0) = 0$으로 두기로 하며

$$[a]^+ = \begin{cases} a & a > 0 \text{ 인 경우} \\ 0 & \text{그 외 경우} \end{cases}$$

라 정의하자. 총 대상들 $T_1, T_2, \cdots, T_n$에 대해서 식(2.1)의 우도 함수는 다음과 같이 표현되고

$$\begin{aligned} & L(\theta_1, \cdots, \theta_{k+1}, \tau_1, \cdots, \tau_k; \mathbf{t}) = \prod_{j=1}^{n} [h(t_j)\exp(-H(t_j))]^{\delta_j} [\exp(-H(t_j))]^{1-\delta_j} \\ = & \prod_{i=1}^{d_1-1} h_1(t_{(i)}; \theta_1) \cdots \prod_{i=d_{k-1}}^{d_k-1} h_k(t_{(i)}; \theta_k) \times \prod_{j=d_k}^{d} h_{k+1}(t_{(i)}; \theta_{k+1}) \\ \times & \exp[-\sum_{j=1}^{n} \{H_1[\min(t, \tau_1)] + \sum_{i=1}^{l} [H_i(\tau_i) - H_i(\tau_{i-1})]^+ + [H_{l+1}(t) - H_{l+1}(\tau_l)]^+\}], \end{aligned} \tag{2.4}$$

여기서 d_i는 모든 j에 대해서 $\tau_i \ge t_j$ 조건을 만족하는 중도 절단되지 않은 대상의 수이고 d는 총 중도 절단되지 않은 대상의 수이다. $d_0 = 1$로 $d_{k+1} = d + 1$로 두기로 하자. 그러면, 식(2.4)로부터 (2.1)에 주어진 우도 함수는 다음과 같은 형태로 쓰여질 수 있으며

$$\begin{aligned} & L(\theta_1, \cdots, \theta_{k+1}, \tau_1, \cdots, \tau_k; \mathbf{t}) \\ = & \theta_1^{d_1-d_0} \cdots \theta_{k+1}^{d_{k+1}-d_k} \exp\{-[\sum_{j=1}^{n_1-n_0} \theta_1 t_{(j)} + \cdots + \sum_{j=1}^{n_{k+1}-n_k} \theta_{k+1} t_{(j)})] \\ & - \sum_{l=1}^{k} (n_{k+1} - n_l)[\theta_l \tau_l - \theta_{l+1} \tau_l]\}, \end{aligned} \tag{2.5}$$

여기서 n_i는 모든 j에 대해서 $\tau_i \ge t_j$ 조건을 만족하는 대상의 수이고 n은 연구에 동원된 모든 대상의 수를 의미한다. $n_0 = 1$로 $n_{k+1} = n + 1$로 두기로 한다.

2.3. 유한 혼합 모형

유한 혼합 모형은 많은 분야에서 모형의 동질성 연구에 있어서 광범위하게 사용되어져 왔다. 특별히, 유한 혼합 모형은 위험률이나 실패율을 비교하는 경우에 실패하는 데까지 걸리는 시간을 모형화 할 수 있게 방법을 제시해준다. 생존 시간 T의 확률 밀도 함수 $f(t|\theta)$가 유한 혼합 형태를 가진다고 가정하면 다음과 같은 형태로 표현되고

$$f(\mathbf{t}|\theta) = \sum_{i=1}^{k} \pi_i f_i(\mathbf{t}|\theta_i), \tag{2.6}$$

여기서 $f_1(t|\theta_1), \cdots, f_k(t|\theta_k)$ 는 각각 $\pi_1, \cdots, \pi_k$의 확률을 가지고 발생되는 k개의 확률 밀도함수이고 $0 \leq \pi_i \leq 1$, $i = 1, \cdots, k$ 이고 $\sum_{i=1}^{k} \pi_i = 1$이다. 식(2.6)에 주어진 생존 분포 $S(t|\theta)$는 다음과 같은 혼합 형태를 가지고

$$S(\mathbf{t}|\theta) = \sum_{i=1}^{k} \pi_i S_i(\mathbf{t}|\theta_i),$$

여기에서 $S_i(t|\theta_i) = \int_t^{\infty} f_i(x|\theta_i)dx$이고 $i = 1, \cdots, k$이다. 실패 시간 T의 확률 밀도 함수가 식(2.6)의 혼합 형태를 가진다면 $f_i(t|\theta)$는 실패 형태 i의 경우에 있어서 T의 조건부 확률 밀도 함수로 π_i는 실패 형태 i의 사전 확률을 의미한다. 그러면 식(2.1)의 우도 함수는 다음과 같이 쓸 수 있다:

$$L(\mathbf{t}|\theta, \pi) = \prod_{j=1}^{d} [\sum_{l=1}^{k} \theta_l \pi_l \exp(-\theta_l t_j)] \times \prod_{j=d+1}^{n} [\sum_{l=1}^{k} \pi_l \exp(-\theta_l t_j)]. \tag{2.7}$$

3. 계산 문제

이 장에서는 각각의 모형에 대한 사후 밀도를 계산한다. 사후 밀도는 우도 함수와 사전 밀도의 곱에 비례하는 형태로 표현된다. 계산상의 어려움을 완화시키기 위하여 Tanner와 Wong(1987)에 의해서 제안된 잠재 벡터 변수를 이용한 깁스 샘플러를 사용할 것이다. $t = (t_1, \cdots, t_n)$ 라 두자.

3.1. 단순 지수 모형

단계 I: $(t_j|\theta) \sim Exp(t_j|\theta)$, $j = 1, 2, \cdots, n$
단계 II: $\theta \sim G(u_\theta\ ,\ v_\theta)$ 여기에서 $G(a, b)$는 위치 모수 a와 척도 모수 b를 가지는 감마 분포를 나타내며 그것의 밀도는 $b^a x^{a-1} e^{-bx}/\Gamma(a)$이다. 식(2.1)을 이용하여 다음과 같은 식을 구할 수 있으며

$$p(\theta|\mathbf{t}) \equiv G(d + u_\theta\ ,\ v_\theta + \sum_{j=1}^{n} t_j),$$

d는 중도 절단되지 않은 대상의 수이고 $p(\theta|t)$는 자료t가 주어진 경우 θ에 대한 사후 밀도를 의미한다.

3.2. 변환 점 모형

모형이 k개의 변환 점을 가진다고 가정하자. Chung외 (1999)이 Ebrahimi외 (1997)의 결과를 $k(\geq 2)$개의 변환 점을 가지는 모형으로 다음과 같이 확장했다:

단계 I: $(t_j|\theta,\tau) \sim f(\mathbf{t}|\theta,\tau)$, $j=1,\cdots,n$, 여기서 $f(\mathbf{t}|\theta,\tau)$ 는 식(2.3)에서 정의된 위험 함수 $h(t)$를 갖는 분포 함수이다.

단계 II: $\tau \sim \pi(\tau)$, $\theta_i \sim G(u_{\theta_i}\ ,\ v_{\theta_i})$ $i=1,\cdots,k+1$.

여기서 $\pi(\tau)$는 $(\tau_1,\cdots,\tau_k)$에 대한 사전밀도이고 $\tau_0=0$라 두기로 한다. 단계 II에 주어진 사전 분포를 이용한 θ와 τ에 대한 결합 사후 밀도는 아래와 같다:

$$\begin{aligned} & p(\theta_1,\cdots,\theta_{k+1},\tau_1,\cdots,\tau_k|\mathbf{t}) \\ \propto\ & h(\mathbf{t};d_1,\cdots,d_k)\exp\{-H(\mathbf{t};n_1,\cdots,n_k)-\sum_{l=1}^{k}(n_{k+1}-n_l)[\theta_l\tau_l-\theta_{l+1}\tau_l]\} \\ & \times\prod_{i=1}^{k+1}\theta_i^{u_{\theta_i}-1}\exp(-v_{\theta_i}\theta_i)\times\pi(\tau). \end{aligned}$$

$t_{(n_i-1)}<\tau_i<t_{(n_i)}, n_i=n_{i-1}+1,\cdots,n-k+i$를 만족하는 구간에서, 깁스 샘플러를 사용하기 위하여 아래와 같은 완전 조건부 분포가 필요하다;

$$[\theta_i|\theta_{(-i)},\tau_1,\cdots,\tau_k,\mathbf{t}]\equiv G(n_i-n_{i-1}+u_{\theta_i}\ ,\ \beta_i),$$

여기서 $\beta_i=(n_i-n_{i-1})\{(n_{k+1}-n_i)\tau_i-(n_{k+1}-n_{i-1})\tau_{i-1}\}+v_{\theta_i}+\sum_{j=n_{i-1}}^{n_i}t_{(j)}$이고 $\theta_{(-i)}=(\theta_1,\cdots,\theta_{i-1},\theta_{i+1},\cdots,\theta_{k+1})$이다.

$$\begin{aligned} [\tau_i|\tau_{(-i)},\theta_1,\cdots,\theta_{k+1},\ t]\ \propto\ & K_{n_i}\exp\{-[(n_{k+1}-n_i)(H_i(\tau_i;\theta_i)-H_{i+1}(\tau_i;\theta_{i+1})]\} \\ & \times\pi(\tau), \end{aligned} \tag{3.1}$$

여기서

$$\begin{aligned} K_{n_i}\ =\ & h(t;d_1,\cdots,d_k)\exp\{-H(t;n_1,\cdots,n_k)-\sum_{l\neq i}^{k}(n_{k+1}-n_l)[H_l(\tau_l;\theta_l) \\ & -H_{l+1}(\tau_l;\theta_{l+1})]\} \end{aligned}$$

이다.

$$C_{n_i}=K_{n_i}\int_{t_{(n_i-1)}}^{t_{(n_i)}}\pi(\tau)\exp\{-(n_{k+1}-n_l)(\theta_l\tau_l-\theta_{l+1}\tau_l)\}d\tau_i,$$

이라 두고, $C=\sum_{n_i=i+1}^{n-k+i}C_{n_i}$, $p_{n_i}=C_{n_i}/C$로 정의하자. τ_i를 샘플링하기 위하여 처음으로 구간을 랜덤하게 결정해야 한다. 즉, 확률 p_{n_i}를 가지고 구간 $t_{(n_i-1)}<\tau_i<t_{(n_i)}$을 선택한다. τ_i가 있을 구간이 선택되면 밀도는 식(3.1)을 C로 나누면 구할 수가 있다. 결국 구간 $t_{(n_i-1)}<\tau_i<t_{(n_i)}$의 누적 분포 함수에 의해서 τ의 값은 쉽게 결정할 수 있다.

3.3. 유한 혼합 모형

단계 I: $(t_j|\theta_1,\pi_1,\cdots,\theta_{k-1},\pi_{k-1},\theta_k) \sim f(t_j|\theta)\quad j=1,\cdots,n$. 여기서 $f(t_j|\theta)$는 식(2.6)에서 정의되었다.

단계 II: $\theta_l \sim G(u_{\theta_l}\ ,\ v_{\theta_l}), \pi_l \sim Dir(u_{\pi_1},u_{\pi_2},\cdots,u_{\pi_k}),\ l=1,\cdots,k$

여기서 $Dir(a_1,a_2,\cdots,a_k)$는 모수 a_i들을 가지는 디리슈레 분포를 나타낸다. 그러면 자료 t가 주어진 θ와 π의 결합 사후 밀도는 다음과 같이 표현된다.

$$\begin{aligned} p(\theta,\pi|\mathbf{t}) \quad \propto \quad & \prod_{j=1}^{d}[\sum_{l=1}^{k}\theta_l\pi_l\exp(-\theta_l t_j)] \times \prod_{j=d+1}^{n}[\sum_{l=1}^{k}\pi_l\exp(-\theta_l t_j)] \\ & \times \prod_{l=1}^{k-1}\pi_l^{u_{\pi_l}-1}(1-\sum_{l=1}^{k-1}\pi_l)^{u_{\pi_k}-1}\,[\prod_{l=1}^{k}\theta_l^{u_{\theta_l}-1}\exp(-v_{\theta_l}\theta_l)]. \end{aligned} \tag{3.2}$$

식(3.2)에 있는 밀도 함수의 계산이 매우 복잡하기 때문에 깁스 샘플러를 바로 적용시키기는 힘들다. 그래서 우리는 Tanner와 Wong(1987)에 의해서 제안된 잠재 벡터 변수를 사용할 것이다. 첫 번째 잠재 벡터 변수 $I_j=(I_{jl})$는 아래와 같이 정의되며

$$I_{jl} \sim M(1,p_{j1},p_{j2},\cdots,p_{jk-1}),\ \ j=1,2,\cdots,d,\ \ l=1,2,\cdots,k-1,$$

여기서 $M(1,a_1,a_2,\cdots,a_{k-1})$ 는 모수 $a_1,a_2,\cdots,a_{k-1}$를 가지는 다항분포를 나타내고 $l=1,\cdots,k-1$, $j=1,2,\cdots,d$에 대해서 $p_{jl}=\frac{\pi_l\theta_l\exp[-\theta_l t_j]}{\sum_{i=1}^{k}\theta_i\pi_i\exp[-\theta_i t_j]}$이다. 두 번째 잠재 벡터 변수 $\lambda_j=(\lambda_{jl})$는 다음과 같이 정의한다:

$$\lambda_{jl} \sim M(1,q_{j1},q_{j2},\cdots,q_{j,k-1}),\ \ j=d+1,\cdots,n,\ \ l=1,2,\cdots,k-1,$$

$l=1,\cdots,k-1$, $j=d+1,\cdots,n$에 대해서 $q_{jl}=\frac{\pi_l\exp[-\theta_l t_j]}{\sum_{i=1}^{k}\pi_i\exp[-\theta_i t_j]}$이다. 잠재 변수들을 사용함으로써, 자료 t가 주어질 경우 θ,π,I와 λ의 결합 밀도를 구할 수 있다:

$$\begin{aligned} p(\theta,\pi,\mathbf{I},\lambda|\mathbf{t}) \quad \propto \quad & \prod_{j=1}^{d}[\sum_{l=1}^{k}\theta_l\pi_l\exp(-\theta_l t_j)] \\ & \times \prod_{j=d+1}^{n}[\sum_{l=1}^{k}\pi_l\exp(-\theta_l t_j)]\prod_{l=1}^{k-1}\pi_l^{u_{\pi_l}-1}(1-\sum_{l=1}^{k-1}\pi_l)^{u_{\pi_k}-1} \\ & \times[\prod_{l=1}^{k}\theta_l^{u_{\theta_l}-1}\exp(-v_{\theta_l}\theta_l)] \times \prod_{j=1}^{d}\prod_{l=1}^{k}(p_{jl})^{I_{jl}}\prod_{j=d+1}^{n}\prod_{l=1}^{k}(q_{jl})^{\lambda_{jl}}, \end{aligned}$$

여기서 $p_{jk}=1-\sum_{l=1}^{k-1}p_{jl}$, $q_{jk}=1-\sum_{l=1}^{k-1}q_{jl}$, $I_{jk}=1-\sum_{l=1}^{k-1}I_{jl}$이고 $\lambda_{jk}=1-\sum_{l=1}^{k-1}\lambda_{jl}$이다. 깁스 샘플러를 적용하기 위하여 아래와 같은 완전 조건부 분포의 계산이 필요하다:

$$\begin{aligned} [\theta_l|\theta_{(-l)},\pi,\lambda,\mathbf{I}] \quad \propto \quad & \theta_l^{u_{\theta_l}-1}\exp(-v_{\theta_l}\theta_l)\theta_l^{\sum_{j=1}^{d}I_{jl}} \\ \times \quad & \exp(-\sum_{j=1}^{d}I_{jl}\theta_l t_j)\exp(-\sum_{j=d+1}^{n}\lambda_{jl}\theta_l t_j) \\ \equiv \quad & G(a\ ,\ b) \end{aligned}$$

여기서 $a = u_{\theta_l} + \sum_{j=1}^{d} I_{jl}$이고 $b = v_{\theta_l} + \sum_{j=1}^{d} I_{jl}t_j + \sum_{j=d+1}^{n} \lambda_{jl}t_j$이다.

$$\begin{aligned}[\pi|\theta,\lambda,\mathbf{I}] &\propto \prod_{l=1}^{k-1} \pi_l^{u_{\pi_l}-1}(1-\sum_{l=1}^{k-1}\pi_l)^{u_{\pi_k}-1} \\ &\quad \times \prod_{l=1}^{k-1} \pi_l^{\sum_{j=1}^{d} I_{jl}+\sum_{j=d+1}^{n}\lambda_{jl}+u_{\pi_l}-1}(1-\sum_{l=1}^{k-1}\pi_k)^{\sum_{j=1}^{d} I_{jl}+\sum_{j=d+1}^{n}\lambda_{jl}+u_{\pi_k}-1} \\ &\equiv Dir(c_1, c_2, \cdots, c_k),\end{aligned}$$

여기서 $c_l = u_{\pi_l} + \sum_{j=1}^{d} I_{jl} + \sum_{j=d+1}^{n} \lambda_{jl} - 1$, $l = 1, 2, \cdots, k$이고 Dir 은 디리슈레 분포를 나타낸다. 또한 다음과 같은 식도 정의할 수 가 있는데

$$[I_{jl}|\mathbf{I}_{(-\mathrm{jl})}, \theta, \lambda, \pi] \sim M(1, p_{j1}, p_{j2}, \cdots, p_{jk-1})$$

여기서 $p_{jl} = \frac{\pi_l\theta_l \exp[-\theta_l t_j]}{\sum_{i=1}^{k}\theta_i\pi_i \exp[-\theta_i t_j]}$, $l = 1, 2, \cdots, k-1$이며 $j = 1, 2, \cdots, d$이다.

$$[\lambda_{jl}|\lambda_{(-jl)}, \theta, \mathbf{I}, \pi] \sim M(1, q_{j1}, q_{j2}, \cdots, q_{jk-1})$$

여기서 $q_{jl} = \frac{\pi_l \exp[-\theta_l t_j]}{\sum_{i=1}^{k}\pi_i \exp[-\theta_i t_j]}$, $l = 1, 2, \cdots, k-1$ 이며 $j = d+1, d+2, \cdots, n$이다.

4. 베이지안 모형 선택

본 장에서는 Gelfand와 Ghosh(1998)의 방법에 기인한 베이지안 모형 선택 방법을 소개하고자한다. 적절한 편차 함수를 사용함으로써 지수 분포에서의 적합도 항과 벌점 항을 포함하는 측도를 정의 할 수 있다. y_{obs}의 l번째 성분과 행위 a에 대해서, 일변량 손실 함수 $L(y, a)$를 가지고, Gelfand와 Ghosh(1998)는 다음과 같은 손실 함수를 정의했다:

$$L(y_{l,rep}, a_l; y_{l,obs}) = L(y_{l,rep}, a_l) + kL(y_{l,obs}, a_l), \quad (k \geq 0) \tag{4.1}$$

여기에서 $y_{l,rep}$는 $y_{l,obs}$의 독립적인 반복을 의미하고, a_l은 $y_{l,rep}$에 대한 '추측'을 의미한다. 이러한 것은 Zellner(1994)에 의해서 균형 손실 함수라 정의되었다. k는 가중치로서 $y_{l,rep}$로부터 떨어져있는 것에 대해 $y_{l,obs}$로 부터 떨어져있는 것에 대한 상대적인 정도를 나타낸다. 모든 구성 요소에 대해서, 모형 선택 기준은 다음과 같다

$$D_k(m) = \sum_{l=1}^{n} \min_{a_l}\{E_{y_{l,rep}|y_{l,obs},m} L(y_{l,rep}; a_l) + kL(y_{l,obs}, a_l)\}. \tag{4.2}$$

여기서, m은 관심의 대상이 되는 모형을 표시한다. 이러한 기준은 베이지안 편차 측도 부분과 모형의 복잡성에 의한 벌점 부분으로 구성된다. 식(4.1)로부터, 적절한 손실 함수를 선택하는 것이 중요하다. 지수 분포의 경우에는, McCullah와 Nelder(1989)가 편차 함수를 다음과 같이 제안했다:

$$D(y, a) = -2\left[\log\left(\frac{y}{a}\right) - \left(\frac{y-a}{a}\right)\right] = -2\log\frac{y}{a} + 2\frac{y-a}{a}. \tag{4.3}$$

계산을 간단히 하기 위하여 손실함수의 경우 $D(y,a)$의 절반, 즉 $L(y,a) = -\log\frac{y}{a} + \frac{y-a}{a}$, 만을 사용하기로 한다. a_l을 고정시키고 결정된 손실 함수를 사용해서 식(4.2)에 정의된 $D_k(m)$의 l번째 항을 구하면 아래와 같고

$$-\mu_l^{(m^2)} + \frac{\mu_l^{(m)}}{a_l} + k\left[-\log\frac{y_{l,obs}}{a_l} + \frac{y_{l,obs}-a_l}{a_l}\right] + \log a_l - 1, \tag{4.4}$$

여기서 $\mu_l^{(m)} = E_{y_{l,rep}|y_{l,obs,m}}\{y_{l,rep}\}$이고 $\mu_l^{(m^2)} = E_{y_{l,rep}|y_{l,obs,m}}\{\log y_{l,rep}\}$이다. 그러면 식(4.4)에 정의된 $D_k(m)$의 l번째 항을 최소화 시켜주는 $\hat{a}_l = (k+1)^{-1}(\mu_l^{(m)} + ky_{l,obs})$이고, 최소 인자 $\hat{a}_l$을 식(4.2)에 대입하자. 그러면 제안된 기준은 다음과 같이 표현된다

$$\begin{aligned} D_k(m) &= \sum_{l=1}^{n}\{-\mu_l^{(m^2)} - k\log y_{l,obs} + (k+1)\log\hat{a}_l\} \\ &= (k+1)\sum_{l=1}^{n}\left[\log\hat{a}_l - \frac{\log\mu_l^{(m)} + k\log y_{l,obs}}{k+1}\right] + \sum_{l=1}^{n}[\log\mu_l^{(m)} - \mu_l^{(m^2)}]. \end{aligned} \tag{4.5}$$

식(4.5)에서

$$G_k(m) = (k+1)\sum_{l=1}^{n}\left[\log\hat{a}_l - \frac{\log\mu_l^{(m)} + k\log y_{l,obs}}{k+1}\right] \tag{4.6}$$

이라 두고

$$P_k(m) = \sum_{l=1}^{n}\log\mu_l^{(m)} - \mu_l^{(m^2)}. \tag{4.7}$$

라 두자. 식(4.5)에 있는 첫 번째 항, $G_k(m)$, 은 적합도 항으로 보여질 수 있다. 만약 반복 $y_{l,rep}$가 $y_{l,obs}$에 잘 예측되었다면, $G_k(m)$은 0가 될 것이다. 식(4.5)에 있는 두 번째 항, $P_k(m)$, 은 벌점 항으로 보여질 수 있다. $\log y_{l,rep}$를 $E_{y_{l,rep}|y_{l,obs,m}}(y_{l,rep}) = \mu_l^{(m)}$에 대해서 두 번째 항까지 테일러 급수 전개하면, 아래와 같은 근사 식을 구할 수 있다

$$\log y_{l,rep} \simeq \log\mu_l^{(m)} + \frac{1}{\mu_l^{(m)}}(y_{l,rep} - \mu_l^{(m)}) - \frac{1}{2(\mu_l^{(m)})^2}(y_{l,rep} - \mu_l^{(m)})^2. \tag{4.8}$$

식(4.8)의 양변에 기대값을 취하면 다음과 같다

$$E_{y_{l,rep}|y_{l,obs,m}}(\log y_{l,rep}) \simeq \log\mu_l^{(m)} - \frac{1}{2(\mu_l^{(m)})^2}E_{y_{l,rep}|y_{l,obs,m}}(y_{l,rep} - \mu_l^{(m)})^2.$$

결론적으로 아래와 같은 근사 식을 얻을 수 있으며

$$\log\mu_l^{(m)} - \mu_l^{(m^2)} \simeq \frac{1}{2(\mu_l^{(m)})^2}\sigma_l^{2(m)},$$

여기서 $\sigma_l^{2(m)} = E_{y_{l,rep}|y_{l,obs},m}(y_{l,rep} - \mu_l^{(m)})^2 = Var(y_{l,rep}|y_{l,obs}, m)$이다. 잘 적합되지 못한 모형에 대해서는 예측 분산이 굉장히 클 것이며, 이에 따라 $P_k(m)$ 역시 그럴 것이다. 특별한 경우로 만약 l번째 자료가 중도 절단된다면, 실제 값 y_l은 우리가 알수 없을 것이다. 다만 y_l이 집합 $A_{l,obs} = [s_l, \infty]$에 속한다는 것만 알 뿐이다. 이러한 중도 절단된 자료에 대해서는, Gelfand와 Ghosh(1998)는 다음과 같이 정의했다

$$L(y_{l,rep}, a_l; A_{obs}) = L(y_{l,rep}; a_l) + k \inf_{y_l \in A_{l,obs}} L(y_l, a_l).$$

지수 함수의 경우 적합한 식(4.3)에 있는 손실 함수를 이용하여 식(4.5)에 있는 기준의 l번째 항은 다음과 같다:

$$\hat{Q} = -\mu_l^{(m^2)} + (k+1)\log \hat{a}_l - k \sup_{y_l \in A_{l,obs}} \{\log y_l\}.$$

따라서 모형 선택 기준은 아래와 같이 표현될 수 있고

$$D_k(m) = \sum_{l=1}^{n} [-\mu_l^{(m^2)} - k \log v_l^{(m)} + (k+1)\log \hat{a}_{v_l}],$$

여기서 $\hat{a}_{v_l^{(m)}} = \frac{kv_l + \mu_l^{(m)}}{k+1}$ 이고 $v_l^{(m)} = \max(\mu_l^{(m)}, s_l)$이다.

5. 실례

5.1. 모의 실험 자료

이 절에서는 모의 실험 자료에서의 최적의 모형을 찾고자 한다. 본 논문에서는 난수를 생성할 때나 깁스 출력치를 생성할 때 Digital Fortran 언어를 사용하였다. 표본의 크기가 20이고 변환 점이 하나인 모의 실험 모형을 가정하자. 따라서 위험률 θ_1, θ_2와 변환점 τ를 아래와 같은 방법으로 정한다.

i) 상수 위험률 모형에 대한 척도 모수의 값들로 $\theta_1 = 1.0$, $\theta_2 = 0.01$로 둔다.

ii)변환 점 τ의 위치를 $P(T \leq \tau) = 0.3$을 만족하도록 고정시키자.

생존 시간 T의 분포 형태를 이미 알고 있으므로 그 것의 누적분포 함수의 역함수를 이용하면 변환점 τ의 값을 구할 수 있다. 이때 주어진 $theta_1, \theta_2$ 값들에 의하면 τ의 실제 값은 0.3568이다. 이를 이용한 모의 실험 자료는 표 5.1에 주어져 있다.

표 5.1. 모의 실험 자료

303.5241	38.878040	0.3250003	31.36641	205.8886	30.502870	0.0847665
284.3023	0.2943283	0.2880020	117.4680	63.00681	0.1633024	0.2720628
3.255565	0.2959414	21.218900	17.51542	0.119714	0.3226901	

지금부터 2장에서 소개된 지수분포에 대한 세 가지 모형을 아래와 같이 고려해 본다.:
M_1(단순 지수 모형), M_2(변환 점이 하나인 변환 점 모형) 그리고 M_3(두 개의 부분군을 가

진 혼합 모형) 이다. 이러한 모형들에 대해서, θ에 대한 사전 분포 형태를 세 가지 경우를 고려해보자. 만약 표 5.1에 있는 모의 실험 자료가 지수 모형이라는 가정을 한다면 θ에 대한 최대 우도 추정치는 0.02이다. 따라서 사전 평균은 0.02가 되도록 하고 θ에 대한 사전 분산을 변화시키면서 민감성을 살펴보기로 한다. θ에 대한 세 가지 형태의 사전 분포는 아래와 같다.
사전 분포 I: $\theta \sim G(2 \times 10^{-2}, 1.0)$, 사전 분포 II: $\theta \sim G(2 \times 10^{-4}, 10^{-2})$,
사전 분포 III: $\theta \sim G(2 \times 10^{-6}, 10^{-4})$.

표 5.2. 생존 모형들에 대한 $G_k(m), P_k(m)$ 그리고 $D_k(m)$

		$G_1(m)$	$P_1(m)$	$D_1(m)$	$D_3(m)$	$D_9(m)$
사전 분포 I	M_1	42.32923	12.07562	54.40485	121.6942	274.6113
	M_2	24.39680	22.07313	46.46992	80.57495	144.6363
	M_3	41.74116	48.52312	90.26425	156.8096	308.4895
사전 분포 II	M_1	42.32082	12.07563	54.39645	121.6678	274.5355
	M_2	25.41024	21.00597	46.41620	64.63758	131.2165
	M_3	43.20689	56.93528	100.1422	168.9208	326.2477
사전 분포 III	M_1	42.32082	12.07563	54.39645	121.6675	274.5348
	M_2	25.44113	21.03331	46.47444	82.76027	152.4940
	M_3	42.45724	55.08880	97.54603	165.1985	319.3089

변환 점 모형(M_2)의 경우 τ에 대한 사전 분포는 최소 관측치와 최대 관측치 사이에 정의된 균일 분포를 가정한다. 또한 혼합 모형의 경우 π의 사전 분포는 균일분포, 즉 Be(1,1)을 사용한다.

표 5.3. 변환 점이 하나인 모형 (M_2)에 대한 추정치

		사후 평균	사후 분산
사전 분포 I	θ_1	0.7031870	0.1710739
	θ_2	0.0142196	0.0000243
	τ	0.4009477	0.4446456
사전 분포 II	θ_1	0.8446819	0.2442662
	θ_2	0.0145143	0.0000241
	τ	0.3691046	0.3650596
사전 분포 III	θ_1	0.8470752	0.2453275
	θ_2	0.0145144	0.0000241
	τ	0.3686798	0.3638609

깁스 샘플러의 수렴성을 확인하기 위해서 CODA를 사용한 Geweke(1992)의 방법을 이용하여 수렴성을 확인하였다. 이때 깁스 샘플러를 30,000번 반복 후 수렴성을 만족하였다. 따라서 마지막 100개를 깁스 출력(Gibbs output)으로 이용하였다.

또한, 각 각의 모형에 대해서, 4장에 소개된 베이지안 모형 선택을 위해서 y_{rep}에 대한 100개의 추측치를 반복 생성하였다. 식(4.5), (4.6) 그리고 (4.7)에 정의된 $G_k(m), P_k(m)$ 그리고 $D_k(m)$을 각각 구할 수 있었고, θ에 대한 사전 분포의 세 가지 형태에 따라 가중치 k를 1,3과 9로 변화시켜 가면서 그 값을 표 5.2에 실었다. 각 경우마다 모형 M_2의 $P_k(m)$의 값이 모형 M_1의 값보다 상대적으로 컸다. 그렇지만 $G_k(m)$은 가장 작았다. 예상한 것 처럼, $G_k(m)$은 일반적으로 모형이 복잡해짐에 따라 증가하는 경향을 보였다. $k = 1, 3, 9$와 사전 분포 I, II, III의 경우 모형 M_2에 대한 $D_k(m)$의 값이 가장 작았기 때문에 변환 점이 하나인 변환 점 모형이 최적의 모형으로 선택되는 것이 이상적이라 할 수 있다. 모형 선택은 k의 값이나 사전 분포의 형태에 민감하지 않았다. $k = 3, 9$의 경우 $G_k(m)$과 $P_k(m)$의 경향이 $G_1(m)$과 $P_1(m)$의 경향과 비슷하므로, $k = 3, 9$의 경우 $G_k(m)$과 $P_k(m)$의 값은 표 5.2에서 생략하였다. 각각의 사전 분포에 대해서, 표 5.3은 모형 M_2의 θ_1, θ_2와 τ의 사후 평균과 분산을 나타낸다. 그림 1, 2은 모형 M_2아래서 위험률 θ_1, θ_2에 대한 사전, 사후 분포함수들을 각각 나타낸다. 여기서 굵은 선(—)은 사전분포 I, 즉 $G(2 \times 10^{-2}, 1)$, 점선$(\cdots)$은 100개의 깁스출력을 이용한 사후분포 함수를 나타낸다.

5.2. 실제 자료

표 5.4에 주어진 자료는 Stangl(1991)자료의 일부분이다. 원 자료는 기관 1, 2, 3, 4 그리고 5의 다섯 개의 임상 기관의 두 가지 치료 군(항 우울제 사용 과 항 우울제 미사용)으로 구성되어져 있다. 이러한 치료 하에서, 환자가 우울증이 재발될 때까지의 시간이 기록되어진다. 여기서, 우리는 단지 표본의 수가 $n = 15$인 기관 1의 항 우울제 사용 군만을 살펴보기로 한다. 표 5.4에서, 각각의 관측치 (t_i, δ_i)에서 t_i는 재발될 때까지의 시간이고 δ_i는 중도 절단 여부를 가리키는 지시함수이다. 즉, $\delta_i = 1$은 i번째 환자가 중도 절단되지 않음을 의미한다. 재발될 때까지 걸리는 시간은 우울 증세가 처음 재발되는 시간까지 월로 표현된다. 5.1.절과 같이 세 가지 모형을 고려하도록 한다: M_1(지수 모형), M_2(변환 점이 하나인 변환 점 모형), M_3(두 개의 부분군을 가진 혼합·모형)이다. 만약 우리가 이 모형이 지수 모형이라고 가정한다면 최대 우도 추정치는 대략 0.003임을 알 수 있다. 따라서, 사전 평균이 0.003이 되게하고 θ에 대한 사전 분산을 변화시키면서 사분 분포에 대한 민감성을 알아보고자 사전 분포 형태를 다음과 같이 두 가지 경우를 고려하기로 하자. 즉,
사전 분포 I: $\theta \sim G(3 \times 10^{-3}, 1.0)$, 사전 분포 II: $\theta \sim G(3 \times 10^{-6}, 10^{-3})$.

표 5.4. 항 우울제 자료

(105.143 , 0)	(74.571 , 0)	(102.143 , 0)	(8.429 , 1)	(108.857 , 0)
(106.429 , 0)	(13.429 , 1)	(105.143 , 0)	(83.000 , 0)	(27.286 , 1)
(104.000 , 0)	(83.000 , 0)	(98.000 , 0)	(98.000 , 0)	(88.000 , 0)

또한 τ와 π에 대한 사전 분포는 앞에서와 같이 균일 분포라 정의한다. 여기서도, 깁스 샘플러가 적용되며, 이때, 총 30,000번의 깁스 출력을 행하여 Geweke 통계량을 이용한 수렴성을 조사하였다. 100번의 반복 생성(y_{rep})을 한 후, 식(4.5), (4.6) 그리고 (4.7)에 정의된 $G_k(m), P_k(m)$과 $D_k(m)$을 각각 계산하여, 그 결과를 표 5.5에 실었다. 이 경우 자료가 많

이 중도 절단되었기 때문에 직관적으로 생존 절단 모형이 최적의 모형이라 생각될 수 있다. 그러나, 표 5.5에서 알 수 있듯이 M_1의 $D_k(m)$이 두 가지 사전 분포의 경우와 가중치 $k=1,3,9$에 대해 가장 작은 값을 가지므로 지수 모형 M_1이 최적이 된다.

표 5.5. 생존 모형들에 대한 $G_k(m), P_k(m)$ 그리고 $D_k(m)$

		$G_1(m)$	$P_1(m)$	$D_1(m)$	$D_3(m)$	$D_9(m)$
사전 분포 I	M_1	6.841372	11.0176	17.85934	27.87381	46.34124
	M_2	11.46738	57.25982	68.72720	87.42741	133.2450
	M_3	130.0802	44.82274	174.9030	430.9051	1183.904
사전 분포 II	M_1	6.842966	11.02188	17.86485	27.88225	46.35691
	M_2	10.41218	53.67030	64.08249	80.63152	120.0999
	M_3	113.3995	72.81187	186.2114	408.8531	1061.777

모의실험 자료로부터, 우리의 제안된 방법이 주어진 지수 생존 모형들 중에서 최적의 모형을 찾는데 이상적임을 보였다. 제안된 기법은 Stangl(1991)의 실제 자료에도 적용하였다. 이때, 모든 사전 분포와 가중치 k에 대해서 지수 모형이 가장 적합함을 알 수 있었다. 또한, 제한된 실험의 결과로 미루어 보아 제시한 모형 선택 방법은 사전 분포나 가중치 k에 민감하지 않다는 것을 알 수 있다.

참고문헌

[1] Chung, Y., Jeong, K. and Han, M. (1999) "Bayesian Analysis for Multiple Change-point Hazard Rate Models", *The Korean Communications in Statistics*, vol.6, 801-811

[2] Cox, D.R. and Oakes, D. (1984) *Analysis of Survival Data*, Chapman and Hall: London.

[3] Ebrahimi, N., Gelfand, A., Ghosh, M. and Ghosh, S.K. (1997) "Bayesian Analysis of Change-point Hazard Rate Models", Technical Report 9707, Department of Statistics, Conneeticut University.

[4] Gelfand, A. and Ghosh, S.K. (1998) "Model Choice: A Minimum Posterior Predictive Loss Approach", *Biometrika*, vol.85,1, 327-335.

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

[6] Geweke, J. (1992) "Evaluating the Accuracy of Sampling-Based Approaches to Calculating Posterior Moments", *In Bayesian Statistics 4*, ed. J.M. Bernado, J.O. Berger, A.P.Dawid and A.F.M. Smith, Oxford, UK; Oxford University Press, 169-193.

[7] Kalbfleisch, J.D. and Prentice, R.L. (1980) *The Statistical Analysis of Failure Time Data*, John Wiley and Sons: New York.

[8] Martz, H.F. and Waller, R.A. (1982) *Bayesian Reliability Analysis*, Wiley: New York.

[9] McCullagh, P. and Nelder, J.A. (1989) *Generalized Linear Models*, Chapman and Hall: London.

[10] Nelson, W.B. (1982) *Applied Life Data Analysis*, Wiley: New York.

[11] Nguyen, H.T., Rogers, G.S. and Walker, E.A. (1984) "Estimation in Change Point Hazard Rate Models", *Biometrika,* vol.71, 229-304.

[12] Qian, J. (1994) "A Bayesian Weibull Survival Model." Ph.D. Thesis, Duke University.

[13] Stangl, D.K. (1991) "Modeling Heterogeneity in Multi-center Clinical Trials Using

[14] Bayesian Hierarchical Survival Models." Ph.D.Thesis, Carnegie-Mellon University.

[15] Stangl, D.K. and Greenhouse, J.B. (1998)"Assesing Placebo Response Using Bayesian Hierarchical Survival Models", *Lifetime Data Analysis*, vol.4, 5-28.

[16] Tanner, M.A. and Wong, W.H. (1987) "The Calculation of Posterior Distributions by Data Augmentation", *Journal of the American Statistical Association*, vol.82, 528-540.

[17] Zellner, A. (1994) "Bayesian and Non-Bayesian Estimation Using Balanced Loss Functions", *In Statistical Decision theory and Related Topics V, Ed. S.S. Gupta and J.O. Berger*, 377-390 , Springer Verlag : New York.

[2001년 5월 접수, 2001년 11월 채택]

Bayesian model selection in exponential survival models

Younshik Chung [1] Misook Kim [2]

ABSTRACT

We introduce three types of exponential survival models, such as simple model, change-point model and finite mixture model in this paper. Among these models, in order to choose the best model, the model choice method is proposed using Gelfand and Ghosh(1998)'s idea. Then to avoid the computational difficulties, data augmentation method (Tanner and Wong, 1987) and Gibbs sampler (Gelfand and Smith, 1990) are employed. Our methodology is applied to both simulated data and Stangl (1991)'s On-impramint Hydrochloride data.

Keywords: Balanced loss function; Bayesian model selection; Change-point model; Finite mixture model; Gibbs sampler; Latent variable.

1) Associate professor, Department of Statistics, Pusan National University, Pusan 609-735 Korea
E-mail:yschung@hyowon.pusan.ac.kr

2) Statistical consultant, Credit Management team, Samsung Card Co. LTD, Seoul 100-666 Korea
E-mail:coglass@samsung.co.kr