The Korean Journal of Applied Statistics (2014) DOI: http://dx.doi.org/10.5351/KJAS.2014.27.6.891
27(6), 891–907

Beta Processes and Survival Analysis

Yongdai Kim[a,1] · Minwoo Chae[a]

[a]Department of Statistics, Seoul National University

(Received October 21, 2014; Revised November 14, 2014; Accepted November 19, 2014)

Abstract

This article is concerned with one of the most important prior distributions for Bayesian analysis of survival and event history data, called *Beta processes*, proposed in Hjort (1990). We review the current state of the art of beta processes and their application to survival analysis. Relevant methodological and practical areas of research that we touch on relate to constructions, posterior distributions, large-sample properties, Bayesian computations, and mixtures of Beta processes.

Keywords: Beta processes, event history analysis, hazard rates, survival analysis.

1. 서론

본 논문은 생존자료 또는 사건사 자료를 모형화하고 분석하는 비모수 베이지안 방법론에 관한 것이다. 이러한 방법들은 전통적으로 생존자료분석에서 사용되는 Kaplan-Meier 추정량, Nelson-Aalen 추정량, 공변량이 있는 경우에는 Cox 비례위험모형과 Aalen의 가법위험모형, 그리고 시간이질(time inhomogeneous) 마코프 과정에 적용할 수 있는 Aalen-Johansen의 방법 등에 대응되는 베이지안의 생존자료분석 기법들이다. 지금 언급한 빈도론 관점에서의 생존자료분석과 관련된 방대한 이론 및 방법론은 Anderson 등 (1993)이 저술한 책에 자세히 기술되어 있다.

베이지안 생존분석에서 가장 근본적인 문제는 위험률, 누적위험률 또는 누적강도함수(cumulative intensity function)의 사전분포를 부여하는 것으로 공변량이 있는 회귀모형의 경우 회귀계수에 대한 사전분포 또한 함께 고려해야 한다. 이러한 사전 분포로는 여러 종류가 있지만, 그 중에서도 가장 중요한 분포족은 단연 Hjort (1985, 1990)가 제안한 *베타과정*이다. 본 논문에서는 베타과정을 기반으로 하는 베이지안 생존분석 또는 사건사자료(event history data)의 분석 대한 최신 이론과 방법론을 다루며, 나아가 향후 연구되어야 할 문제들에 대해서 소개한다.

Ferguson (1973)은 미지의 분포에 대한 최초의 비모수 사전분포족인 *디리끌레 과정*을 제안하였다. 이 이후로 많은 사람들이 중도절단된 자료가 주어졌을 때 이에 대한 사후분포를 구하려고 노력하였다. 이러한 결실로 Susarla와 Van Ryzin (1976)은 베이즈 추정량을 구하였고, Ferguson과 Phadia (1979)는 Doksum (1974)이 제안한 NtR 과정(neutral to the right process)이 중도절단된 자료에 대한 켤레사전분포가 된다는 사실을 증명해냈다. 하지만 NtR 과정 사전분포족은 그 범위가 상당히 넓기 때문에 모든 NtR 과정이 유용하다고 보기 어려울 뿐만아니라 수학적으로도 다루기가 쉽지 않다. 디리끌레 과정은

[1]Corresponding author: Department of Statistics, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 151-747, Korea. E-mail: ydkim0903@gmail.com

NtR 과정의 한 종류이기는 하지만 중도절단자료에 대한 켤레사전분포가 아니기 때문에 이 또한 훌륭한 대안이 될 수는 없다. 중도절단자료에 대한 사전분포에 대하여 1980년대에는 큰 진전이 없다가 1990년에 Hjort가 누적위험함수에 대한 사전분포로 베타과정을 제안하였고 이것이 중도절단자료에 대한 켤레 사전분포가 된다는 사실을 증명하였다. 또한, Hjort는 베타과정이 시간이질 마코프 과정의 누적강도함수에 대한 사전분포로 활용되어 다중사건자료, illness-death 모형, recovery 모형 (Anderson 등, 1993) 등에 적용할 수 있다는 사실을 발견하였다. 이는 베타과정의 발견으로 인해 사건사 자료에 대한 베이지안 분석 방법이 현저하게 발전햇다는 것을 의미한다.

Hjort가 처음 베타과정을 발견한 이래로 많은 사람들이 다양한 목적으로 이를 연구하기 시작하였다. Lo (1993)은 감마과정을 사용하여 베타과정을 유도해냈고 이를 *beta-neutral 과정*이라고 명명하였다. Kim (1999)는 Aalen의 일반적인 승법 셈과정 모형에서 누적강도함수에 대한 사후분포를 유도해 내었다. Walker와 Muliere (1997)은 소위 beta-Stacy 과정이라는 사전분포를 연구하였는데, 이는 베타과정과 거의 유사하지만 누적위험함수가 아니라 누적분포함수 관점에서 기술한 것이다. Kim과 Lee (2003)은 Cox의 비례위험모형에서 기저누적함수에 대한 사전분포로 베타과정을 사용하였을 때 회귀계수에 대한 사후분포를 구하였다. (이에 대한 결과의 일부는 Hjort의 1990년 첫 논문에도 소개되어 있다.) Kim과 Lee (2001, 2004) 및 Kim (2006)에서는 베타과정 사전분포에 대한 사후분포의 대표본 이론을 다루고 있으며 소위 말하는 좋은 성질인 점근적 일치성, 정규성 등이 성립한다는 것을 증명하였다. De Blasi와 Hjort (2007)은 로지스틱 링크를 사용한 비례위험모형에서 사후분포의 대표본 이론을 밝혀냈다. Damian 외 2인 (1996), Wolpert와 Ickstadt (1998), Lee와 Kim (2004)은 베타과정 표본을 생성하는 계산 알고리즘을 각각 개발하였으며, Laud 등 (1998)은 베타과정을 사전분포로 하는 Cox 비례위험모형에서 마코프 연쇄 몬테 까를로 (MCMC) 알고리즘을 개발하였다. Kim (2001)은 모수모형 근처에서 정의되는 비모수 사전분포인 혼합베타과정을 제안하였고 De Blasi 등 (2009)는 여러 베타과정을 중첩하는 비모수 사전분포를 개발하였다. Kim 등 (2012)은 베타과정을 다변수로 확장하는 격인 베타-디리끌레 과정을 개발하였고 이것이 마코프 과정의 누적강도함수에 대한 켤레 사전분포가 된다는 사실을 증명하였다.

본 논문의 2장부터 7장까지는 베타과정과 관련된 여섯 가지 분야 (베타과정의 생성, 사후분포, 대표본 이론, 베이지안 계산법, 혼합 베타과정, 다변수로의 확장)에 대한 연구 결과를 다루었다. 마지막 장인 8장에서는 몇 가지 논의 사항과 함께 향후 연구 방향을 모색하기로 한다.

2. 베타과정의 생성

음이 아닌 실수 집합 $[0,\infty)$에서 정의된 분포함수 F가 주어졌을 때 이에 대응하는 누적위험함수 A는

$$A(t) = \int_0^t \frac{dF(s)}{1-F(s)} = \int_0^t \frac{dF(s)}{F[s,\infty)} \tag{2.1}$$

로 정의된다. 역으로 주어진 누적위험함수 A에 대하여 이에 대응하는 분포함수는

$$F(t) = 1 - \prod_{s\in[0,t]} \{1 - dA(s)\}$$

로 구할 수 있는데 여기서 $\prod$는 곱적분 (Gill과 Johansen, 1990)을 의미한다. 그러므로 F를 추정하는 문제는 A를 추정하는 문제로 변환할 수 있고 그 반대 또한 마찬가지이다. 누적위험함수의 극소적 의미는

$$dA(s) = \Pr\{\text{transition in } [s, s+ds] | \text{survival up to time } s\} \tag{2.2}$$

이기 때문에 F보다는 A를 추정하는 문제가 위험률 (또는 강도함수, 전이율)이라는 개념에 보다 가깝다고 볼 수 있다. 특히 경쟁위험모형이나, 이질 시간 마코프 과정 등을 분석하는 경우 누적위험함수 A를 다루는 것이 F를 고려하는 것보다 개념적으로 훨씬 자연스럽다.

베타과정은 누적위험함수들의 공간인 $\mathcal{A}$ 상의 사전분포 및 사후분포를 구하기 위해 고안되었다. 이번 장에서는 베타과정을 생성하는 네 가지 방법에 대한 내용을 다룬다. 누적함수 공간 $\mathcal{A}$는 $[0,\infty)$ 상에서 정의된 함수 A 중에서 단조증가, 우연속, $A(0)=0$, 그리고 모든 $t\in[0,\infty)$에 대하여 $\Delta A(t)\le 1$을 만족함과 동시에 $\lim_{t\to\infty}A(t)=\infty$ 또는 어떤 $t>0$에 대하여 $\Delta A(t)=1$를 만족하는 함수들을 모아놓은 집합이다. 여기서 $\Delta A(t)=A(t)-A(t-)=A\{t\}$는 시간 t에서 함수 A의 점프 크기를 의미한다.

2.1. 시간 0이산 과정의 극한 (Hjort, 1990)

A_0를 $\mathcal{A}$ 원소 중 연속인 함수라 하고 $c(\cdot)$을 조각별 연속인 음이아닌 함수라고 하자. 주어진 자연수 m과 $i=1,2,\ldots$에 대하여 시간 공간 $[0,\infty)$을 서로 겹치지 않는 구간들 $((i-1)/m, i/m]$로 나눈 후 $\beta\beta(a_{m,i},b_{m,i})$를 따르는 서로 독립인 확률변수 $X_{m,i}$를 정의하자. 여기서 상수 $a_{m,i}$과 $b_{m,i}$는 각각

$$a_{m,i}=c_{m,i}A_0\left(\frac{i-1}{m},\frac{i}{m}\right],\quad b_{m,i}=c_{m,i}\left(1-A_0\left(\frac{i-1}{m},\frac{i}{m}\right]\right)$$

로 정의되며 $c_{m,i}=c((i-1/2)/m)$이다. 그리고 확률과정 $A_m(\cdot)$을 $A_m(0)=0$,

$$A_m(t)=\sum_{i/m\le t}X_{m,i},\quad \forall t\ge 0$$

으로 정의하자. Hjort (1990)은 독립 증분을 갖는 어떤 확률과정 A가 존재하여 확률 1로 표본 경로가 $\mathcal{A}$에 포함되며 모든 $t>0$에 대하여 $D[0,t]$ 상에서 $A_m(\cdot)\to_d A(\cdot)$가 성립한다는 사실을 증명하였다. 여기서 $D[0,t]$는 $[0,t]$ 상에서 정의된 우연속이며 좌극한을 갖는 함수들을 모은 집합으로 Skorohod 위상을 갖는 거리공간이다. 또한, 동일 논문에서 Hjort는 A의 Laplace 변환이

$$E\exp\{-\theta A(t)\}=\exp\left\{-\int_0^1(1-e^{-\theta s})\,dL_t(s)\right\},\quad \forall\theta\ge 0 \tag{2.3}$$

로 된다는 사실을 증명하였다. 여기서, $t\ge 0$에 대하여 L_t는 A의 Lévy 측도로,

$$dL_t(s)=\left\{\int_0^t c(z)s^{-1}(1-s)^{c(z)-1}\,dA_0(z)\right\}1_{0<s<1}ds$$

로 정의되는 측도이며 수식 (2.3)을 A의 Lévy 표현법이라고 부른다.

이러한 이론을 기반으로 Hjort (1990)은 베타과정을 다음과 같이 정의하였다. 우선 A_0를 $\mathcal{A}$의 원소라 하고 $t_1,t_2,\ldots$에서 점프를 갖는다고 하자. 또한 $c(\cdot)$을 $[0,\infty)$ 상의 조각별 연속인 음이 아닌 함수라고 하자. 이제 $(c(\cdot),A_0(\cdot))$을 모수로 하는 베타과정 A는 Lévy 과정으로써 Lévy 표현법이

$$E\exp\{-\theta A(t)\}=\left\{\prod_{j:\,t_j\le t}E\exp(-\theta S_j)\right\}\exp\left\{-\int_0^1(1-e^{-\theta s})\,dL_t(s)\right\}$$

와 같이 되는 확률과정으로 정의한다. 여기서

$$S_j=\Delta A(t_j)\sim\text{Beta}\big\{c(t_j)\Delta A_0(t_j),c(t_j)(1-\Delta A_0(t_j))\big\},$$
$$dL_t(s)=\int_0^t c(z)s^{-1}(1-s)^{c(z)-1}\,dA_{0,\text{cont}}(z)\,ds$$

이며 $A_{0,\mathrm{cont}}(t) = A_0(t) - \sum_{t_j \le t} \Delta A_0(t_j)$는 A_0에서 점프 부분을 제거하고 남은 연속함수이다. 확률과정 A가 베타과정을 따를 때 $A \sim \mathrm{Beta}\{c(\cdot), A_0(\cdot)\}$라고 표기하기로 한다. 또한, Hjort (1990)는 베타과정의 평균과 분산에 대하여 $EA(t) = A_0(t)$와

$$\mathrm{Var}A(t) = \int_0^t \frac{dA_0(s)\{1 - dA_0(s)\}}{c(s)+1}$$

이 성립한다는 사실을 증명하였다. 약간의 계산을 더 하면

$$E\{A(t) - A_0(t)\}^3 = \int_0^t \frac{2\,dA_0(s)\,\{1 - dA_0(s)\}\{1 - 2\,dA_0(s)\}}{\{c(s)+1\}\{c(s)+2\}}$$

를 보일 수 있다.

그러므로 베타과정에서 A_0는 A에 대한 사전 추측치 생각할 수 있고, c는 이에 대한 믿음의 정도를 조절하는 모수라고 생각할 수 있다. 또한 F가 A에 대응하는 랜덤분포함수라면 $F(t) = 1 - \prod_{s\in[0,t]}\{1 - dA(s)\}$가 성립하기 때문에 $EF(t) = 1 - \prod_{s\in[0,t]}\{1 - dA_0(s)\}$ 또한 만족된다는 것을 알 수 있다. 그러므로 A_0는 역시나 F에 대응되는 누적위험함수에 대한 사전 추측치로 여길 수 있다.

A가 베타과정이면 그에 대응되는 랜덤분포함수 F는 NtR 과정 (Doksum, 1974)이 된다. 다시 말해, 단조증가, 음이아닌 함수이면서 독립 증분을 갖는 확률과정 Y가 존재하여 $1 - F(t) = \exp\{-Y(t)\}$를 만족한다. 또한 A와는 관계가 $1 - dA(s) = \exp\{-dY(s)\}$로 된다.

베타과정은 디리끌레 과정과도 흥미로운 관계가 있는데 F가 $[0,\infty)$ 상의 디리끌레 과정이고 유한측도 aF_0를 기저측도로 갖는다고 하자. 그러면 F에 대응되는 누적위험함수 A는 A_0와 c를 모수로 하는 베타과정이 되는데 (Hjort, 1990), 여기서 A_0는 $F_0(\cdot) = EF(\cdot)$의 누적위험함수이고 $c(t) = aF_0[t,\infty)$이다. 즉, 베타과정은 디리끌레 과정의 자연스러운 확장이 된다는 것이다.

2.2. 감마과정의 위험률 (Lo, 1993)

Lo (1993)은 다음과 같이 두 개의 독립인 감마과정을 이용하여 베타과정을 생성하는 방법을 발견하였다. 먼저, 주어진 $[0,\infty)$ 상의 유한측도 α가 있을 때 독립증분과정 γ의 유한차원 분포 $\gamma(t) - \gamma(s) = \gamma((s,t])$가 평균, 분산이 모두 $\alpha((s,t])$인 감마분포를 따를 때, 확률과정 γ는 α를 모수로 하는 감마과정이라고 정의한다.

이제 γ_α와 γ_β를 각각 α와 β를 모수로 하는 독립인 감마과정이라고 하자. 다음으로 A라는 확률과정을

$$A(t) = \int_0^t \frac{\gamma_\alpha(ds)}{\gamma_\alpha[s,\infty) + \gamma_\beta[s,\infty)}$$

로 정의하자. 그러면 A가 $c(t) = \alpha[t,\infty) + \beta[t,\infty)$와

$$A_0(t) = \int_0^t \frac{\alpha(ds)}{\alpha[s,\infty) + \beta[s,\infty)}$$

를 모수로 하는 베타과정이 된다는 사실이 Lo (1993)에 의해 증명되었다. 이러한 생성법은 유용하기는 하지만 $c(t)$가 반드시 단조감소함수여야 한다는 제약이 있기 때문에 완전히 일반적이지는 않다.

2.3. 포아송 측도를 통한 접근 (Kim, 1999)

Kim (1999)는 포아송 랜덤 측도의 개념을 통해 베타과정을 설명하였다. 이는 준마팅게일을 방식의 접근법으로 $\mathcal{A}$ 상의 독립증분 과정 A의 Lévy 측도를 다음과 같이 정의하는 것이다. 주어진 $\mathcal{A}$ 상의 독립

증분 과정 A가 있을 때 $[0,\infty)\times[0,1]$ 상의 랜덤측도를 $\mu(dt,dx)=I\{A[t,t+dt]\in[x,x+dx]\}$로 정의하자. 그러면 μ가 포아송 랜덤측도 (Jacod과 Shiryaev, 1987)가 된다는 사실을 쉽게 증명할 수 있는데 포아송 랜덤측도는 평균 측도인 $\nu(dt,dx)=E(\mu(dt,dx))$에 의해 그 성질이 규명된다. 역으로, $[0,\infty)\times[0,1]$ 상의 σ-유한 측도인 ν가 주어지면 평균측도가 ν인 포아송 랜덤측도 μ가 유일하게 존재하는데 이 때 μ의 subordinator를

$$A(t)=\int_0^t\int_0^1 x\,\mu(ds,dx)$$

로 정의할 수 있다. 따라서 $\mathcal{A}$ 상의 독립증분 확률과정은 $[0,\infty)\times[0,1]$ 상의 σ-유한 측도 하나를 선택하기만 하면 완전히 정해진다.

A를 $\mathcal{A}$ 상의 확률적으로 연속(stochastically continuous)이고 Lévy 표현법이 (2.3)으로 주어지는 독립증분과정이라고 가정하자. 그러면 모든 $t>0$와 $[0,1]$ 상의 보렐집합 B에 대하여

$$\nu([0,t]\times B)=\int_B dL_t(x)$$

가 성립한다는 사실을 증명할 수 있다 (Jacod과 Shiryaev (1987)의 Theorem II.4.8). 그러므로 ν는 Lévy 측도 L_t의 또다른 표현법이라고 할 수 있다.

$\mathcal{A}$ 상에서 평균이 A_0인 동시에 확률적으로 연속인 독립증분과정 A를 생성하는 문제를 생각해보자. 이 때 Lévy 측도는

$$\nu(dt,dx)=f_t(x)\,dx\,dA_0(t)$$

로 주어져 있다고 하자. $EA(t)=A_0(t)$가 성립하기 때문에

$$EA(t)=\int_0^t\int_0^1 x f_s(x)\,dx\,dA_0(s)=A_0(t). \tag{2.4}$$

또한 성립해야 한다. 식 (2.4)를 만족하는 $f_t(x)$를 선택하는 하나의 방법으로 $xf_t(x)$가 모든 $t>0$에 대하여 $[0,1]$ 상의 확률밀도함수가 되도록 하는 것이 있다. 따라서 $xf_t(x)$에 대한 자연스러운 선택은 $\alpha(t)$와 $\beta(t)$를 모수로 하는 베타분포의 확률밀도함수가 되도록 하는 것이다. 이러한 맥락 하에서 Kim과 Lee (2001)은 $(\alpha(t),\beta(t),A_0(t))$를 모수로 하는 확장된 베타과정을 제안하였는데 이는 독립증분을 가지는 동시에

$$\nu(dt,dx)=\frac{1}{x}\frac{\Gamma(\alpha(t)+\beta(t))}{\Gamma(\alpha(t))\Gamma(\beta(t))}x^{\alpha(t)-1}(1-x)^{\beta(t)-1}\,dA_0(t). \tag{2.5}$$

를 Lévy 측도로 갖는 확률과정을 말한다. 참고로 $(c(\cdot),A_0(\cdot))$를 모수로 하는 베타과정은 수식 (2.5)에서 $\alpha(t)=1$이고 $\beta(t)=c(t)$인 특별한 경우이다.

3. 사후 분포

베타과정을 처음 소개한 Hjort는 1990년 논문에서 누적위험함수뿐만 아니라 시간이질 마코프 과정의 누적강도함수에 대한 사전분포로도 베타과정을 사용하였다. Kim (1999)는 Hjort의 결과들을 셈과정에 대한 Aalen의 일반적인 승법강도모형 (이러한 모형은 Andersen 등 (1993)에 잘 설명되어 있다)으로 확장하였다. 이번 장에서는 우선 누적위험함수의 사전분포로 베타과정을 사용했을 때의 사후분포를 유도하고 그 다음 시간 이질 마코프 과정과 Aalen의 일반적 승법 셈과정 모형을 다루기로 한다. 또한 Cox의 비례위험모형을 베이지안 관점에서 다루는 방법을 소개하기로 한다.

3.1. 누적위험함수의 사후분포

$X_1, \ldots, X_n$를 누적분포함수가 F인 *i.i.d.* 생존시간이라 하고 $C_1, \ldots, C_n$을 X_i와 독립인 중도절단시간이라 하자. 관측치가 중도절단되었기 때문에 실제 관측치는 $(T_1, \delta_1), \ldots, (T_n, \delta_n)$로 표현할 수 있다. 여기서 $T_i = \min(C_i, X_i)$이고 $\delta_i = I\{X_i \le C_i\}$이다. A를 F의 누적위험함수라고 하자.

A에 대한 사전분포로 $(c(\cdot), A_0(\cdot))$를 모수로 하는 베타과정을 고려하자. 사후 분포에 관하여 Hjort (1990)에 나와있는 가장 중요한 내용은 사후분포 또한 $(c_0^p(\cdot), A_0^p(\cdot))$를 모수로 하는 베타과정이 된다는 것이다. 여기서

$$A_0^p(t) = \int_0^t \frac{c(s)\, dA_0(s) + dN(s)}{c(s) + Y(s)} \quad \text{and} \quad c_0^p(t) = c(t) + Y(t) \tag{3.1}$$

이며 $N(t) = \sum_{i=1}^n I\{T_i \le t, \delta_i = 1\}$는 셈과정, $Y(t) = \sum_{i=1}^n I\{T_i \ge t\}$는 위험에 노출되어 있는 개체수이다.

사후분포에 관해서는 몇 가지 짚고 넘어가야 할 것들이 있다. 첫째로 베타과정 분포족은 우측으로 중도절단된 자료에 대하여 켤레 사전분포가 된다는 것이다. 참고로 디리끌레 과정과 감마 과정은 모두 우측중도절단 자료에 대한 켤레 사전분포가 아니다. 둘째는 제곱손실함수에 대한 베이즈 추정량이 사후 분포의 평균, 즉 $A_0^p(t)$이 된다는 것이다. 이 형태를 자세히 들여다 보면 사전 추측치와 빈도론의 Nelson-Aalen 추정량 $\hat{A}(t) = \int_0^t dN(s)/Y(s)$의 가중평균으로 이루어져 있어 매우 흥미롭다. 또한, $c(s)$는 시간 s에서 위험에 노출되어 있는 가상의 사전 표본 수로 해석할 수가 있는데 여기서 가상의 표본은 누적위험함수가 A_0인 분포를 따른다고 생각하면 된다. 셋째로 $c(\cdot)$이 0으로 수렴할 때 베이즈 추정량은 빈도론의 추정량으로 수렴한다. 따라서 비모수 최대우도추정량이라 할 수 있는 Nelson-Aalen 추정량은 사전 정보가 전혀 없을 때의 베이즈 추정량이라고 생각할 수 있다. 마지막으로 Hjort (1990)에서 언급되었듯이 사전분포의 정확도에 해당하는 함수 $c(t)$는 용도에 따라 유동적으로 정해질 수 있다. 예를 들어, $t > t_0$인 $A(t)$을 추론하고자 할 때 $[0, t_0]$까지의 관측치들을 바탕으로 사전분포를 정할 수도 있다는 것인데 이는 사전분포에 필요한 두 모수 $c(t)$와 $A_0(t)$가 실험 전부터 완전히 결정될 필요가 없다는 것을 의미한다.

3.2. 누적강도함수의 사후분포

$X_1, \ldots, X_n$를 상태공간 $\{1, \ldots, k\}$ 상에서 정의된 독립인 시간이질 마코프 과정이라고 하고 $j \neq h$일 때의 누적강도함수를 $A_{h,j}$라고 하자. 누적강도함수는 직관적으로 다음과 같이 생각할 수 있다 (*cf.* 식 (2.1)과 식 (2.2)).

$$dA_{h,j}(s) = \Pr\{\text{transition occurs } h \to j, \text{inside } [s, s+ds] \mid X_i(s) = h\}. \tag{3.2}$$

전통적인 생존분석이라 함은 상태가 단 두 종류(살아 있거나 죽은)이며 가능한 전이 방향 또한 한 가지 밖에 없는 경우이다. 그러므로 생존분석에서 마코프 과정을 고려하게 되면 그 응용 범위가 훨씬 넓어진다.

사전분포로 $A_{h,j}$가 확률적으로 연속이고 독립이며 평균 및 정확도 모수가 $\tilde{A}_{h,j}$, $c_{h,j}$인 베타과정을 따른다고 하자. Hjort (1990)과 Kim (1999)는 각각 $A_{h,j}$의 사후분포가 마찬가지로 독립인 베타과정이며 평균, 정확도 모수가 가 $\tilde{A}_{h,j}^p$, $c_{h,j}^p$로 주어진다는 사실을 증명하였다. 여기서

$$\tilde{A}_{h,j}^p(t) = \int_0^t \frac{c_{h,j}(s)\, d\tilde{A}_{h,j} + dN_{h,j}(s)}{c_{h,j}(s) + Y_h(s)}$$

이며 $c^p_{h,j}(t) = c_{h,j}(t) + Y_h(t)$이다. 또한,

$$N_{h,j}(t) = \sum_{i=1}^{n} \sum_{s \le t} I\{X_i(s) = j, X_i(s-) = h\}, \quad Y_h(t) = \sum_{i=1}^{n} I\{X_i(t-) = h\}$$

는 각각 $h \to j$로 전이되는 셈과정과 현재 h에 있는 지의 여부를 나타내는 확률과정이다.

Kim (1999)은 시간이질 마코프 과정를 통하지 않고도, Aalen의 일반적인 승법강도모형에서 비슷한 결과를 도출해냈다. 우선 주어진 셈과정 N이 Aalen의 승법강도모형을 따른다는 것은 어떤 예측가능과정(predictable process) $Y(t)$와 단조증가이면서 음이아닌 함수 $A(t)$가 존재하여 $N(t) - \int_0^t Y(s)dA(s)$가 마팅게일이 된다는 것이다. 이 때 A를 N의 누적강도함수라고 부른다. 하나의 예를 들자면, $Y(t) \equiv 1$인 경우 $N(t)$는 평균이 $A(t)$인 포아송 과정이 된다. 또 하나의 중요한 예는 중도절단된 포아송 과정으로 N이 포아송 과정이고 Y가 0-1의 값을 가지는 조각별 상수 과정일 때 $Y(t) = 1$인 경우에만 셈과정 $N(t)$를 관측하고 $Y(t) = 0$인 경우에는 관측할 수 없는 경우이다. 이 때 관측되는 셈과정 $N^c(t) = \int_0^t Y(s)\, dN(s)$ 또한 Aalen의 승법강도모형을 따른다는 사실이 잘 알려져 있다. Kim (1999)는 A에 대한 사전분포로 베타과정을 사용했을 때 사후분포 또한 식 (3.1)을 모수로 하는 베타과정이 된다는 것을 증명하였다.

3.3. Cox의 비례위험모형

비례위험모형은 다음과 같이 정의된다. 먼저 $i = 1, \ldots, n$에 대하여 $X_1, \ldots, X_n$은 생존시간이라고 하고 $Z_1, \ldots, Z_n$를 R^p 상의 공변량이라고 하자. 공변량 Z_i가 주어졌을 때 X_i의 분포 F_i를

$$1 - F_i(t) = \{1 - F(t)\}^{\exp(\beta^t Z_i)} \tag{3.3}$$

라고 가정한다. 여기서 $\beta \in R^p$는 미지의 회귀계수이고 F 또한 미지의 분포함수로 공변량이 0벡터일 때의 생존시간에 대한 분포함수이다. 대부분의 응용 분야에서 생존시간이 우측 중도절단되었다는 사실을 고려하면 C_i를 X_i와 Z_i에 독립인 중도절단 시간, $T_i = \min(C_i, T_i)$, $\delta_i = I\{X_i \le C_i\}$라고 했을 때 관측치는 $\mathcal{D}_\backslash = \{(\mathcal{T}_\infty, \delta_\infty, \mathcal{Z}_\infty), \ldots, (\mathcal{T}_\backslash, \delta_\backslash, \mathcal{Z}_\backslash)\}$가 된다.

수식 (3.3)은 위험률이 아니라 로그 생존함수에 대한 비례모형처럼 보이지만 동치인 수식 (*cf.* 식 (2.1))

$$1 - dA_i(s) = \{1 - dA(s)\}^{\exp(\beta^t Z_i)}, \tag{3.4}$$

로 바꾸어 놓고 보면 그 이름의 유래를 짐작할 수 있다. 이러한 표현은 시간이 이산인 경우를 자연스럽게 확장한 것으로 수학적으로 다루기 편리하다.

비례위험모형 식 (3.4)에는 두 개의 모수(회귀계수 β와 기저누적위험함수 A)가 있다. 준모수 베이지안 분석을 위한 자연스러운 방법은 기저누적위험함수 A에는 베타과정을, 그리고 회귀계수 β에는 적절한 밀도함수 $\pi(\beta)$를 사전분포로 부여하는 것이다. 이는 Hjort (1990)의 6장에서 소개된 베이지안 Cox 회귀모형의 원형으로 볼 수 있다. 이 논문에서 Hjort는 모든 관측시간이 서로 다를 경우 (A, β)에 대한 사후분포를 유도해냈다. 동일한 관측치가 있는 경우에 대한 확장은 Kim과 Lee (2003)의 논문에서 다루었는데 여기서는 이러한 일반적인 경우를 다루기로 한다.

먼저 $i = 1, \ldots, n$에 대하여 셈과정 $N_i(t) = I\{T_i \le t, \delta_i = 1\}$를 정의하고 $Y_i(t) = I\{T_i \ge t\}$, 합과정은 $N(t) = \sum_{i=1}^n N_i(t)$, 차분을 $\Delta N(t) = N(t) - N(t-)$, 그리고 $Y(t) = \sum_{i=1}^n Y_i(t)$라고 하자. 다음으로 q_n을 중도절단되지 않은 서로 다른 관측 시간의 수라 하고 $t_1 < t_2 < \cdots < t_{q_n}$을 그 값으로 표기하자.

이제

$$D_n(t) = \{i \le n \colon T_i = t, \delta_i = 1\}, \quad R_n(t) = \{i \le n \colon T_i \ge t\}$$

라 하고 $R_n^+(t) = R_n(t) - D_n(t)$라 하자. 그러면 사후분포에 대하여 다음이 성립한다.

(i) β와 $\mathcal{D}_\setminus$이 주어졌을 때, A의 사후분포는 Lévy 측도가

$$\nu(dt, dx \mid \beta, \mathcal{D}_\setminus) = \frac{1}{x}(1-x)^{\sum_{j\in R_n(t)} \exp\left(\beta^t Z_j\right)+c(t)-1}\, dx\, dt + \sum_{i=1}^{q_n} dH_{n,i}(x \mid \beta)\delta_{t_i}(dt),$$

로 주어지는 독립증분과정이다. 여기서 δ_a 는 a에서 전체 질량 1을 갖는 디락 측도이고 $H_{n,i}(\cdot \mid \beta)$는 $[0,1]$ 상에서 정의되는 확률 측도로

$$h_{n,i}(x \mid \beta) = \frac{1}{x}\left[\prod_{j\in D_n(t_i)} \left\{1-(1-x)^{\exp\left(\beta^t Z_j\right)}\right\}\right](1-x)^{\sum_{j\in R_n^+(t_i)} \exp\left(\beta^t Z_j\right)+c(t_i)+1}. \tag{3.5}$$

에 비례하는 밀도함수를 가진다.

(ii) β의 주변사후분포는

$$\pi(\beta \mid \mathcal{D}_n) \propto \pi(\beta)\exp\{-\rho_n(\beta)\}\prod_{i=1}^{q_n}\int_0^1 h_{n,i}(x \mid \beta)\, dx$$

로 주어지며, 여기서

$$\rho_n(\beta) = \sum_{i=1}^{n}\int_0^{T_i}\int_0^1 \frac{1}{x}\left\{1-(1-x)^{\exp\left(\beta^t Z_i\right)}\right\}(1-x)^{\sum_{j=i+1}^{n} \exp\left(\beta^t Z_j\right)+c(t)-1}\, dx\, dt$$

이다.

참고로 β가 주어졌을 때 A는 베타과정과 거의 비슷하지만 $\delta_i = 1$인 T_i, 즉, 관측된 생존시간에서 랜덤 점프의 분포가 베타분포는 아니다. Hjort (1990)의 6장이나 Kim과 Lee (2003)의 논문에 A의 사후분포에 대한 평균과 분산이 자세히 기술되어 있다.

4. 대표본 이론

베이지안 통계에서는 사후분포의 여러가지 대표본 이론을 다루기도 한다. 첫째가 사후분포의 일치성에 관한 것인데 이는 자료의 수가 커짐에 따라 사후분포의 확률 대부분이 자료를 생성하는 참분포 주변으로 몰려드는 현상을 말한다. 두 번째 성질은 소위 Bernstein-von Mises(BvM) 정리라고 불리는 것인데 사후분포의 점근 분포가 최대우도추정량의 표본분포와 점근적으로 같아지는 성질이다. 사후분포의 이러한 성질들은 사후분포의 평균 추정량, 즉 베이즈 추정량의 극한분포를 결정한다. 본 장에서는 베타과정 사전분포에 대한 이러한 대표본 이론을 살펴보기로 한다.

모수공간이 클 때 Diaconis와 Freedman (1986)은 사후분포가 불일치성을 가질 수 있다는 것을 보였는데 이 이후로 사후분포의 대표본 이론에 관한 것이 많이 연구되기 시작하였다 (*e.g.* Hjort (1986), Barron (1988), Barron 등 (1999), Ghosal 등 (1999), Ghosal 등 (2000), Shen과 Wasserman (2001), Walker와 Hjort (2001), Ghosh와 Ramamoorthi (2003), Walker (2003, 2004)). 하지만, 이러한 이론

과 방법론은 보통 모형에 속하는 모든 분포함수를 지배하는 σ-유한 측도의 존재성 하에서 전개된다. 사후분포를 구하기 위해 베이즈 정리를 사용하는 위의 일반적 이론에서 이 가정은 필수적이다. 하지만 베타과정에서 나오는 표본 경로는 확률 1로 이산분포인데 이로 인해 모든 이산확률분포뿐만 아니라 모든 연속확률분포를 고려해야 한다. 이를 모두 지배하는 σ-유한 측도는 존재하지 않을뿐만 아니라 사후분포를 구하기 위해 베이즈 정리를 직접적으로 사용하기도 어렵다. 따라서 위에서 언급한 논문과 이론들을 생존모형에 직접적으로 적용할 수는 없다.

사후분포에 비해 베이즈 추정량의 점근적 성질을 구하는 것은 상대적으로 쉽기 때문에 먼저 이를 살펴보기로 하자. Nelson-Aalen 추정량 $\hat{A}(t) = \int_0^t dN(s)/Y(s)$ 은 점근적으로 일치하며 정규분포로 수렴할 뿐만아니라 효율적이라는 사실이 잘 알려져 있다 (Andersen 등, 1993). 수식 (3.1)의 베이즈추정량 $A_0^p(t)$와 Nelson-Aalen 추정량을 비교해보면 $\sup_{t\in[0,\tau]} c(t)$가 유계이며 $X_1, \ldots, X_n$이 연속인 누적위험함수 $A^*(t)$로부터 나온 랜덤 표본일 경우

$$\sup_{t\in[0,\tau]} \left| A_0^p(t) - \hat{A}(t) \right| = O_p\left(\frac{1}{n}\right)$$

가 성립한다는 것을 증명할 수 있다. 따라서 베이즈 추정량은 Nelson-Aalen 추정량과 똑같이 좋은 점근적 성질을 가진다는 것을 알 수 있다. 이러한 결과는 Hjort (1990)에 의해 처음 밝혀졌다.

베타과정의 사후분포에 대한 점근이론은 Kim과 Lee (2003)에 의해 처음 밝혀졌다. 이들은 우측 중도절단된 자료에서 누적위험함수에 대한 사전분포로 베타과정을 사용하였을 때 그 사후분포가 일치성을 갖는다는 사실을 증명하였다. 즉, $X_1, \ldots, X_n$이 연속인 누적위험함수 $A^*(t)$로부터 나온 랜덤 표본일 경우 적당한 조건 하에서 확률 1로 어떠한 $\epsilon > 0$에 대해서도

$$\pi^p\left(\sup_{t\in[0,\tau]} |A(t) - A^*(t)| < \epsilon | \mathcal{D}_n \right) \to 1$$

이 성립한다는 것을 보인 것이다. 여기서 $\pi^p(\cdot|\text{data})$는 A의 사후분포, $\mathcal{D}_n = \{(T_1, \delta_1), \ldots, (T_n, \delta_n)\}$이다. 또한, 식 (2.5)에서 정의된 확장된 베타과정 중 원래의 베타과정 ($\alpha(t) \equiv 1$)만이 점근적으로 일치한다는 사실을 증명하였다. 이는 독립증분과정 중 불일치하는 사전분포가 있다는 것이기 때문에 매우 놀라운 사실이다. 이전까지만 해도 사람들은 독립증분과정의 성질이 매우 좋기 때문에 모든 독립증분과정에 대한 사후분포가 일치할 것이라고 생각하였다. Kim과 Lee (2001)은 이러한 믿음이 사실이 아니라는 것을 증명한 셈이다.

나아가 Kim과 Lee (2003)는 어떤 독립증분과정을 사전분포로 사용하면 사후분포는 일치하지만 BvM 정리가 성립하지 않는다는 사실을 증명하였다. 다행히도 이들은 같은 논문에서 베타과정을 사용하면 BvM 정리가 성립한다는 것을 증명하였다. 즉, 확률 1로

$$\pi^p\left(\sqrt{n}\left(A(\,\cdot\,) - \hat{A}(\,\cdot\,) \right) \in \cdot\, | \mathcal{D}_n \right) - \mathcal{L}\left(\sqrt{n}\left(\hat{A}(\,\cdot\,) - A^*(\,\cdot\,) \right) \in \cdot\, \right) \to 0$$

이 성립한다는 것이다. 여기서 $\hat{A}$는 Nelson-Aalen 추정량이고 $\mathcal{L}(\,\cdot\,)$는 $\hat{A}$의 표본분포이다.

비례위험모형 하에서는 Kim (2006)의 결과에 따르면 기저위험함수뿐만 아니라 회귀계수에 대해서도 BvM 정리가 성립한다.

5. 베이지안 계산법

베타과정에 대한 이론이 발달함에 따라 베타과정과 관련된 계산 방법에 대한 연구 또한 중요해졌다. 가장 중요한 것은 베타과정의 표본 경로를 생성하는 알고리즘에 관한 것이다. 이러한 알고리즘은 Hjort

(1990)에 나오는 시간 이산화 방법 (2.1장)에 더하여 Damien 등 (1996), Wolpert와 Ickstadt (1998), Lee와 Kim (2004) 등에 의해서 개발되었다. 본 장에서는 베타과정의 표본경로를 생성하는 세 가지 방법을 다루고 이러한 방법들이 비례위험모형 하의 MCMC에서 어떻게 쓰이는 지에 대해 설명하기로 한다.

5.1. 베타과정의 표본경로 생성법

우선, A가 $(c(\cdot), A_0(\cdot))$를 모수로 하는 확률적으로 연속인 베타과정이라고 하고 주어진 구간 $[0,\tau]$에서 A의 표본 경로를 생성하는 다음 세 가지 방법을 살펴보자.

시간 이산화 알고리즘 (Hjort, 1990)*:*

(1) 충분히 큰 m을 잡고, $i \le m\tau$에 대하여 2.1장에 나온 대로 독립인 확률변수 $X_{m,i} \sim \beta(a_{m,i}, b_{m,i})$를 생성한다.

(2) $A(t) = \sum_{i/m \le t} X_{m,i}$라고 놓는다.

가중 포아송 알고리즘 (Damien 등, 1996)*:*

(1) 충분히 큰 m에 대하여 $dA_0(t)/A_0(\tau)$로부터 *i.i.d.* 확률변수 $T_1, \ldots, T_n$을 생성한다.

(2) $i = 1, \ldots, m$에 대하여 $X_i \sim \text{Beta}(1, c(T_i))$를 생성한다.

(3) $i = 1, \ldots, m$에 대하여, $\lambda_i = A_0(\tau)/(nX_i)$라 놓고 $Z_i \sim \text{Pois}(\lambda_i)$를 생성한다.

(4) $A(t) = \sum_{i \le m} X_i Z_i I\{T_i \le t\}$라 놓는다.

ε-근사 알고리즘 (Lee와 Kim, 2004)*:*

(1) 충분히 작은 $\varepsilon > 0$에 대하여, 점프의 총 수 M을 $\text{Pois}(\lambda)$에서 생성한다. 여기서 $\lambda = \varepsilon^{-1} \int_0^\tau c(s)\, dA_0(s)$이다.

(2) 점프 시간 $(s_1, \ldots, s_M)$을 다음과 같이 생성한다: $i = 1, \ldots, M$에 대하여 $[0,\tau]$ 상에서 확률밀도함수가 $c(s)\, dA_0(s)$에 비례하도록 i.i.d. 확률변수 $r_1, \ldots, r_M$를 생성한 후 $s_i = r_{(i)}$라고 한다. 여기서, $r_{(i)}$는 i번째 순서통계량이다.

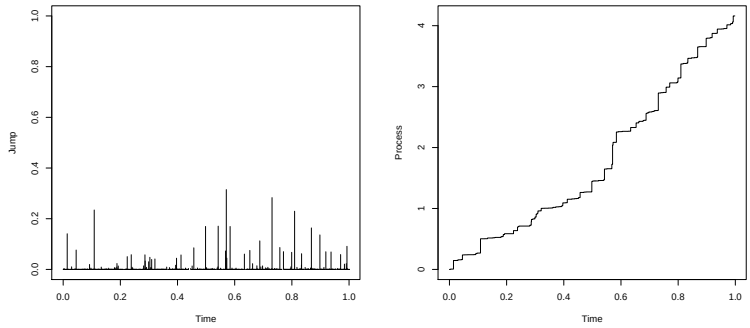
(3) 점프 크기 $(x_1, \ldots, x_M)$를 $x_i \mid s_i \sim \text{Beta}(\varepsilon, c(s_i))$로부터 생성한다.

(4) $A(t) = \sum_{i \le M} x_i I\{s_i \le t\}$로 놓는다.

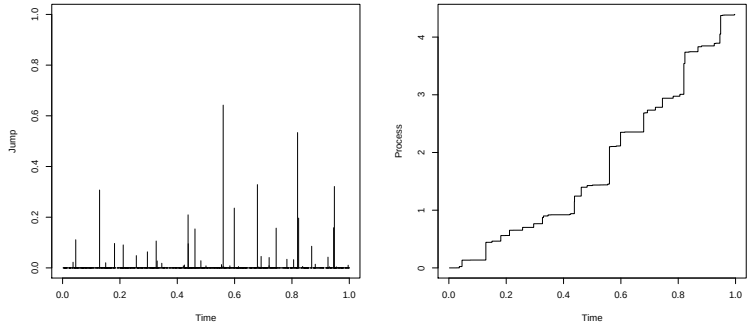
(1) 위의 세 알고리즘은 모두 주어진 베타과정을 근사하는 알고리즘이다. 예를 들어, 포아송 가중 알고리즘에서는 m이 ∞로 갈 때, ϵ-근사 알고리즘에서는 ϵ이 0으로 갈 때 생성된 표본경로가 베타과정으로 수렴한다.

(2) Damien 등 (1996)은 원래 베타과정의 증분을 계산하기 위해 포아송 가중 알고리즘을 개발하였고 앞에서 소개한 알고리즘은 이를 약간 변형한 것이다. 포아송 가중 알고리즘의 문제점 중 하나는 생성된 표본 경로의 점프 크기가 1보다 클 수 있다는 것인데 이는 사후분포의 질량이 음의값을 가질 수도 있게 한다. 다른 알고리즘들은 이러한 문제점이 없다.

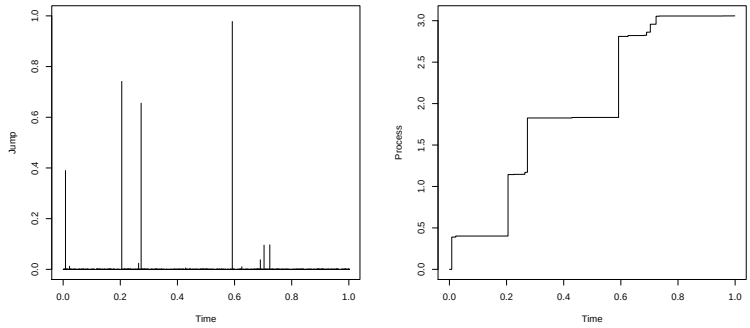
Figure 5.1은 가중 포아송 알고리즘을 통해 $A_0(t) = 2t$와 여러 상수값 c에 대한 베타과정의 점프와 표본 경로를 그린 것이다. 그림을 보면 c 값이 작을 때는 소수의 큰 점프가 전체 경로의 대부분을 차지하고 반대의 경우 작은 점프가 많이 모여 전체 경로를 결정한다는 것을 알 수 있다.



(a)



(b)



(c)

Figure 5.1. Beta process relaization with $A_0(t) = 3t$ and (a) $c(t) \equiv 10$, (b) $c(t) \equiv 3$ and (c) $c(t) \equiv 1$. The left panels and right panels draw jumps and sample paths, respectively.

5.2. 비례위험모형에서의 MCMC 알고리즘

본 장에서는 베이지안 생존분석에서 베타과정의 표본경로 생성 알고리즘이 어떻게 사용되는지를 설명한

다. 이를 위해 비례위험모형에서 베이지안 방법에 사용되는 MCMC 알고리즘을 소개한다.

비례위험모형에 관한 기본적인 사항들은 3.3장에 소개했었다. A에 대한 사전분포로는 $\mathrm{Beta}(c(\cdot), A_0(\cdot))$, β에 대한 사전분포는 $\pi(\beta)$를 사용하기로 한다. β와 관측치가 주어졌을 때 A의 사후분포는 3.3장에 주어져 있다. A의 사후분포는 연속인 부분 A_{cont}와 이산부분 A_{disc} 둘로 분해하여 $A = A_{\mathrm{cont}} + A_{\mathrm{disc}}$로 쓸 수 있는데 여기서

$$A_{\mathrm{cont}} \sim \mathrm{Beta}\left(R_n(t,\beta) + c(t),\ \frac{c(t)}{R_n(t,\beta)+c(t)} A_0(t)\right)$$

이고 $R_n(t,\beta) = \sum_{j \in R_n(t)} \exp(\beta^t Z_j)$이다. 또한 A_{disc}는 중도절단되지 않은 관측 시간인 $t_1 < \cdots < t_{q_n}$에서의 고정 점프만으로 이루어져 있으며 점프의 분포는 수식 (3.5)를 따른다. 앞서 말했듯이 점프 크기의 분포가 베타분포가 아니기 때문에 β가 주어졌을 때 A의 사후분포는 베타과정이 아니다. 하지만 연속인 부분 A_{cont}는 베타과정이므로 A_{cont}의 표본경로는 6.1장에서 소개한 알고리즘을 통해 생성할 수 있다. Laud명 등 (1998)은 두 개의 보조변수를 사용하여 A_{disc}를 생성하는 멋진 알고리즘을 개발하였다.

일단 A를 생성하면 β는 다음과 같이 생성하면 된다. 먼저 $u_1, \ldots, u_{q_n}$을 $t_1, \ldots, t_{q_n}$에서 A_d의 점프 크기라 하고 $\{(s_1, x_1), \ldots, (s_M, x_M)\}$를 A_c의 점프 시간과 크기라고 하자. 참고로 $t \in [0, \tau]$에 대하여

$$A(t) = A_c(t) + A_d(t) = \sum_{i=1}^{M} x_i I\{0 \le s_i \le t\} + \sum_{i=1}^{q_n} u_i I\{0 \le t_i \le t\}$$

로 쓸 수 있다. 그러면 A가 주어졌을 때 회귀계수 β의 사후분포 밀도함수는

$$\pi(\beta) \prod_{i=1}^{q_n} \left\{ \prod_{j=1}^{k_i} \left(1 - (1-u_i)^{\exp\left(\beta^t z_{i(j)}\right)}\right) (1-u_i)^{R_n^+(t_i,\beta)} \right\} \times \prod_{j=1}^{M} (1-x_j)^{R_n(s_j,\beta)} \tag{5.1}$$

에 비례한다. 여기서, $R_n^+(t,\beta) = \sum_{j \in R_n^+(t)} \exp(\beta^t Z_j)$이고 k_i는 t_i에서의 중도절단되지 않은 관측 시간의 수, $z_{i(j)}$는 $D_n(t_i)$에 속하는 j번째 관측치의 공변량 벡터이다. β를 생성하는 것은 임의보행 Matropolis–Hastings 알고리즘을 사용하면 되는데 저자들의 경험에 의하면 이 방법은 매우 유용하다. 대안으로 β 생성 시 $\log \pi(\beta)$가 위로 볼록한 함수이면 Laud 등 (1998)처럼 Gilks와 Wild (1992)의 알고리즘을 사용할 수도 있다.

전체 MCMC 알고리즘은 위에서 설명한 방법대로 A와 β가 수렴할 때까지 반복적으로 시행하면 된다.

6. 혼합 베타과정

이번 장에서는 베타과정을 응용하는 한 방법을 다루기로 한다. 많은 문제에서 모수 분포족을 사용할 수 있기는 하지만 보통의 경우 자료가 완전히 모수 모형을 따른다고 보기는 힘들다. 이런 경우 베이지안은 모수 분포족 주변에 대부분의 질량을 갖는 비모수 사전분포를 사용할 수 있다. Doss (1994)는 혼합 디리끌레 과정을 사용하였고 그 후에 Kim (2001)은 다음과 같은 혼합 베타과정을 고안하였다. 먼저 누적위험함수에 대한 모수 분포족 A_θ를 고른다. 다음으로 주어진 모수족 A_θ와 조각별 연속이면서 음이아닌 함수 $c_\theta(t)$ $(\theta \in \Theta \subset R^p)$에 대하여 확률 측도 ν를 따르는 θ를 선택한다. 여기서 ν는 Θ 상에 정의된 사전분포이다. 이제 A가 $\beta\{c_\theta(\cdot), A_\theta(\cdot)\}$로부터 생성하면 이 A가 혼합 베타과정이 된다. 혼합 베타과정을 따르는 A를

$$A \sim \int \mathrm{Beta}\,\{c_\theta(\cdot), A_\theta(\cdot)\}\ \nu(d\theta). \tag{6.1}$$

로 표기하도록 한다.

이제 A를 수식 (6.1)의 혼합베타과정이라고 하고 주어진 관측치 $(T_1, \delta_1), \ldots, (T_n, \delta_n)$가 있을 때 $T_{(1)} \leq \cdots \leq T_{(n)}$를 $T_1, \ldots, T_n$의 순서통계량이라고 하자. 그리고 q_n은 $T_1, \ldots, T_n$ 중 중도절단되지 않은 서로 다른 값의 수라 하고 $v_1, \ldots, v_{q_n}$ 을 그 관측값이라고 하자. Kim (2001)은 $(T_1, \delta_1), \ldots, (T_n, \delta_n)$가 주어졌을 때 A의 사후분포 또한 혼합 베타과정

$$\int \text{Beta}\left\{c_\theta^p(\,\cdot\,), A_\theta^p(\,\cdot\,)\right\} \nu^p\left(d\theta \mid T^n, \delta^n\right),$$

이 된다는 것을 보였다. 여기서

$$\nu^p\left(d\theta \mid T^n, \delta^n\right) \propto \exp\left\{-\sum_{i=1}^{n} \int_0^{T_{(i)}} \frac{c_\theta(s)}{c_\theta(s) + n - i} dA_\theta(s)\right\} \times \prod_{k=1}^{q_n} c_\theta(v_k)\lambda_\theta(v_k) \left[\prod_{j=1}^{d_n(v_k)} \{c_\theta(v_k) + Y_n(v_k) - j\}\right]^{-1} \nu(d\theta),$$

이고

$$dA_\theta^p(s) = \frac{c_\theta(s)\, dA_\theta(s) + dN(s)}{c_\theta(s) + Y(s)} \quad \text{and} \quad c_\theta^p(s) = c_\theta(s) + Y(s)$$

이며 N과 Y는 3.1장에서 정의된 것과 같다.

Kim (2001)은 θ의 사후분포가 c_θ의 선택에 크게 의존한다는 사실을 보였다. 특히 $c_\theta(t) \equiv c > 0$인 경우 θ의 사후분포가 이상한 경향을 보였는데 (사적인 대화를 통해) Hjort 또한 비슷한 현상을 발견하였다. Hjort는 $A_\theta(t) = \theta t$일 때 베이즈 추정량이 $O(n/\log n)$의 속도로 수렴한다는 사실을 증명하였다. 이는 θ의 사후분포가 A의 사후분포에 막대한 영향을 미친다는 것이며 그 결과 A의 사후분포에 대한 대표본 성질이 일반적인 이론과 크게 다를 수 있다는 것을 의미한다. 이와는 대조적으로, Kim (2003)에서 $c_\theta(t) = c\exp\{-A_\theta(t)\}$이고 $c > 0$인 경우 A와 θ의 사후분포가 모두 정상적으로 행동한다는 것을 보였다. 게다가 Kim과 Hjort (2012)에서는 A가 적당한 조건 하에 있고

$$0 < \inf_{\theta\in\Theta, t\in[0,\tau]} c_\theta(t)\exp(A_\theta(t)) \leq \sup_{\theta\in\Theta, t\in[0,\tau]} c_\theta(t)\exp(A_\theta(t)) < \infty \tag{6.2}$$

이 만족되면 A에 대한 BvM 정리 또한 성립한다는 사실을 증명하였다. 현재 본 저자들은 혼합 베타과정에 대하여 BvM 정리가 성립하기 위해서는 수식 (6.2)가 반드시 필요한 조건일 거라 믿고 있다.

7. 베타-디리끌레 과정: 다변수로의 확장

베타-디리끌레 과정을 쉽게 설명하기 위해 먼저 K 종류의 사건이 있는 경쟁위험모형을 생각해보자. $X(t)$를 t 시점에서 개체의 상태, 즉 시간 t까지 아무런 사건도 발생하지 않았다면 $X(t) = 0$, t 시점 이전에 k번째 종류의 사건이 발생했다면 $X(t) = k$로 하기로 한다. $A(t) = (A_1(t), \ldots, A_K(t))$는 마코프 과정 $X(t)$의 누적강도함수, 다시 말해 $k = 1, \ldots, K$에 대하여 $\Pr(X(t) = 0 | X(t-) = 0) = 1 - \sum_{k=1}^{K} \Delta A_k(t)$, 그리고 $\Pr(X(t) = k | X(t-) = 0) = \Delta A_k(t)$ 라고 하자.

먼저 A를 $\mathcal{T} = \{t_1, \ldots, t_m\}$에서만 점프를 하는 이산확률과정이라 하자. 그러면 A에 대한 자연스러운 사전분포는

$$\pi(A) \propto \prod_{j=1}^{m}\prod_{k=1}^{K} \left(\Delta A_k\left(t_j\right)\right)^{\alpha_k(t_j)} \left(1 - \sum_{k=1}^{K} \Delta A_k(t_j)\right)^{\alpha_{K+1}(t_j)}$$

와 같이 서로 독립인 디리끌레 분포를 부여하는 것이다. 하지만 Hjort (1990)는 이러한 독립 디리끌레 분포가 $\max_j |t_j - t_{j-1}| \to 0$일 때 $A^* = (A_1^*, \ldots, A_K^*)$로 수렴한다는 사실을 보였는데 여기서 $A_1^*, \ldots, A_K^*$는 서로 독립인 연속 베타과정이다. 참고로 연속 시간 상의 베타과정을 누적강도함수에 대한 사전분포로 사용하면 켤레가 되지 않을 수도 있다 (Kim 등, 2012).

A_k들이 서로 독립이 아닌 켤레 사전분포가 되게 하기 위해 Kim 등 (2012)은 다음과 같은 베타-디리끌레 과정을 개발하였다. 먼저 A가 이산과정으로 $\mathcal{T} = \{t_1, \ldots, t_m\}$에서만 점프를 하는 경우부터 살펴보자. $\Delta A(t) = (\Delta A_1(t), \ldots, \Delta A_K(t))$라고 표기하고 $j = 1, \ldots, m$에 대하여 $\Delta A(t_j)$를 서로 독립이라 하자. $\Delta A.(t_j) = \sum_{k=1}^K \Delta A_k(t_j)$가 $(\alpha(t_j), \beta(t_j))$를 모수로 하는 베타분포를 따르면서 $\Delta A.(t_j)$가 주어졌을 때 $(\Delta A_1(t_j)/\Delta A.(t_j), \ldots, \Delta A_K(t_j)/\Delta A.(t_j))$의 분포가 $(\gamma_1(t_j), \ldots, \gamma_K(t_j))$를 모수로 하는 디리끌레 분포라고 하자. 그러면 이에 대응되는 $\Delta A(t_j)$의 밀도함수는

$$(\Delta A.\,(t_j))^{\alpha(t_j)-\sum_{k=1}^K \gamma_k(t_j)}\,(1-\Delta A.\,(t_j))^{\beta(t_j)-1}\prod_{k=1}^K \Delta A_k\,(t_j)^{\gamma_k(t_j)} \tag{7.1}$$

에 비례하는데 이 분포를 베타-디리끌레 분포라고 한다 (Kim 등, 2012). 여기서 $\Delta A_k(t_j) \geq 0$이고 $0 \leq \Delta A.(t_j) \leq 1$이다. 또한 이런 식으로 형성된 확률과정 A를 이산 베타-디리끌레 과정이라고 한다. 동일 논문에서 저자들은 시간 간격이 0으로 갈 때 시간 이산 베타-디리끌레 과정이 독립이 아닌 다변수 Lévy 과정으로 수렴한다는 것을 보였는데 이것이 바로 일반적인 베타-디리끌레 과정이다. 정확한 정의는 다음과 같다. $A.$이 (A_0, c)를 모수로 하는 베타과정이고 $(V_1(s), \ldots, V_K(s))$이 $(\gamma_1(s), \ldots, \gamma_K(s))$를 모수로 하는 디리끌레 분포를 따를 때 $k = 1, \ldots, K$에 대하여

$$A_k(t) = \sum_{s \leq t} V_k(s)\Delta A.(s)$$

로 정의하면 $(A_1, \ldots, A_K)$는 $(A_0(\,\cdot\,), c(\,\cdot\,), \gamma_1(\cdot), \ldots, \gamma_K(\cdot))$를 모수로 하는 베타-디리끌레 과정이라고 한다.

주어진 A에 대하여 $X_1, \ldots, X_n$는 A를 누적강도함수로 하는 경쟁위험모형 나온 독립 사건사 자료라고 하자. A에 대한 사전분포로 $(A_0, c, \gamma_1, \ldots, \gamma_K)$를 모수로 하는 베타-디리끌레 과정을 부여하면 Kim 등 (2012)은 그 사후분포가 다시 $(A_0^p, c^P, \gamma_1^p, \ldots, \gamma_K^p)$를 모수로 하는 베타-디리끌레 과정이 된다는 사실을 증명하였다. 여기서

$$dA_0^p(t) = \frac{c(t)}{c(t)+n}dA_0(t) + \frac{1}{c(t)+n}\sum_{i=1}^n dN_i(t),$$
$$c^p(t) = c(t) + n$$

이고 $k = 1, \ldots, K$에 대하여 $\gamma_k^p(t) = \gamma_k(t) + \sum_{i=1}^n I(X_i(t-) = 0, X_i(t) = k)$이다.

Kim 등 (2012)는 베타-디리끌레 과정이 경쟁위험모형이나 illness-death 모형 등 유한 상태공간을 취하는 일반 마코프 과정의 누적강도함수에 대한 켤레 사전분포가 된다는 사실을 증명하였다. 또한, 이러한 마코프 과정 자료에 적용할 수 있는 베이지안 준모수 회귀모형을 제안하였고 베타-디리끌레 과정을 이 모형에서 기저 누적강도함수로 사용하였다.

8. 결론 및 향후 과제

지금까지 베타과정 사전분포에 대한 좋은 성질들을 살펴보았다 그 중에서도 가장 중요한 성질은 베타과정이 우측 중도절단자료에 대한 켤레 사전분포가 된다는 사실이고 베이즈 추정량이 사전 추측치와 빈도

추정량의 가중평균으로 표현된다는 것이다. 둘째로 두 개의 모수인 A_0와 c가 각각 사전 추측치와 그에 대한 신뢰도라는 의미로 해석될 수 있기 때문에 사전분포를 정하는 것이 매우 용이하다는 것 또한 큰 장점이다. 세 번째 장점은 점근적 일치성이나 BvM 정리 등 사후분포에 대한 대표본 이론이 잘 성립되어 있다는 것이다. 넷째로 다양한 셈과정 모형의 누적강도함수에 대한 사전분포로 활용될 수 있다는 점을 들 수 있고, 다섯 번째 중요한 성질로 디리끌레 과정을 특수한 형태로 포함하고 있다는 것이다. 마지막으로 표본 경로를 쉽게 근사시킬 수 있을 뿐만 아니라 실용적으로 계산법 또한 다양한데 이를 요약하자면 베이지안 생존자료 분석에 있어 베타과정이 매우 편리하고 유용한 도구가 된다는 것이다.

베타과정과 베이지안 생존분석 관련 향후 연구과제로는 다음과 같은 것들을 고려할 수 있다. 먼저, 여러 개의 누적강도함수에 의존성을 줄 수 있는 사전분포의 개발이 필요하다. 현재는 각 누적위험함수가 독립이라는 가정하에서 베이지안분석방법이 개발되었다. Aalen의 가법 모형에 대한 베이지안 분석방법의 개발 또한 이론적/기술적 어려움으로 인하여 아직 연구가 되지 않고 있는 분야이다. 나아가 대부분의 회귀모형을 포함하는 일반변형모형(general trnasformation model)에 대한 베이지안 분석방법의 개발도 연구를 기다리고 있다.

References

Andersen, P. K., Borgan, Ø., Gill, R. D. and Keiding, N. (1993). *Statistical Models Based on Counting Processes*, Springer, New York.

Barron, A. R. (1988). The exponential convergence of posterior probabilities with implications for Bayes estimators of density functions, Technical report, University of Illinois.

Barron, A., Schervish, M. J. and Wasserman, L. (1999). The consistency of posterior distributions in non-parametric problems, *Annals of Statistics*, **27**, 536–561.

Damien, P., Laud, P. W. and Smith, A. F. M. (1996). Implementation of Bayesian non-parametric inference based on Beta processes, *Scandinavian Journal of Statistics*, **23**, 27–36.

De Blasi, P., Favaro, S. and Muliere, P. (2009). A class of neutral to the right priors induced by superposition of beta processes, Working paper of Collegio Carlo Alberto.

De Blasi, P., Hjort, N. L. (2007). Bayesian survival analysis in proportional hazard models with logistic relative risks, *Scandinavian Journal of Statistics*, **34**, 229–257.

Diaconis, P. and Freedman, D. A. (1986). On the consistency of Bayes estimates, *Annals of Statistics*, **14**, 1–26.

Doksum, K. A. (1974). Tailfree and neutral random probabilities and their posterior distributions, *Annals of Probability*, **2**, 183–201.

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

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

Ferguson. T. S. and Phadia, E. G. (1979). Bayesian nonparametric estimation based on censored data, *Annals of Statistics*, **7**, 163–186.

Ghosal, S., Ghosh, J. K. and Ramamoorthi, R. V. (1999). Posterior consistency of Dirichlet mixtures in density estimation, *Annals of Statistics*, **27**, 143–158.

Ghosal, S., Ghosh, J. K. and van der Vaart, A. D. (2000). Convergence rates of posterior distributions, *Annals of Statistics*, **28**, 500–531.

Ghosh, J. K. and Ramamoorthi, R. V. (2003). *Bayesian Nonparametircs*, Springer.

Gilks, W. R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling, *Applied Statistics*, **41**, 337–348.

Gill, R. D. and Johansen, S. (1990). A survey of product-integration with a view toward application in survival analysis, *Annals of Statistics*, **18**, 1501–1555.

Hjort, N. L. (1985). Contribution to the discussion of Andersen and Borgan's 'Counting process models for

life history data: a review', *Scandinavian Journal of Statistics*, **12**, 141–150.

Hjort, N. L. (1986). Contribution to the discussion of Diaconis and Freedman's 'On the consistency of Bayes estimates', *Annals of Statistics*, **14**, 49–55.

Hjort, N. L. (1990). Nonparametric Bayes estimators based on Beta processes in models for life history data, *Annals of Statistics*, **18**, 1259–1294.

Jacod, J. and Shiryaev, A. N. (1987). *Limit Theorems for Stochastic Processes*, Springer, New York.

Kim, Y. (1999). Nonparametric Bayesian estimators for counting processes, *Annals of Statistics*, **27**, 562–588.

Kim, Y. (2001). Mixture of Beta processes for right censored data, *Journal of the Korean Statistical Society*, **30**, 127–138.

Kim, Y. (2003). On posterior consistency of mixtures of Dirichlet processes with censored observations, *Scandinavian Journal of Statistics*, **30**, 535–547.

Kim, Y. and Hjort, N. (2012). Beta process envelopes around parametric models for survival analysis, Unpublished manuscript.

Kim, Y., James, L. and Weisbach, R. (2012). Bayesian analysis for multi-state event history data: The Beta-Dirichlet process prior, *Biometrika*.

Kim, Y. and Lee, J. (2001). On posterior consistency of survival models, *Annals of Statistics*, **29**, 666–686.

Kim, Y. and Lee, J. (2003). Bayesian analysis of proportional hazard models, *Annals of Statistics*, **31**, 493–511.

Kim, Y. and Lee, J. (2004). The Bernstein-von Mises theorem for survival models, *Annals of Statistics*, **32**, 1492–1512.

Kim, Y. (2006). The Bernstein-von Mises theorem of semiparametric Bayesian models for survival data, *Annals of Statistics*, **34**, 1678–1700.

Laud, P. W., Damien, P. and Smith, A. F. M. (1998). Bayesian nonparametric and covariate analysis of failure time data, In *Practical Nonparametric and Semiparametric Bayesian Statistics*, (eds: Dey, D., Muller, P. and Sinha, D.).

Lee, J. and Kim, Y. (2004). A new algorithm to generate Beta processes, *Computational Statistics and Data Analysis*, **47**, 441–453.

Lo, A. Y. (1993). A Bayesian bootstrap for censored data, *Annals of Statistics*, **21**, 100–123.

Shen, X. and Wasserman, L. (2001). Rates of convergence of posterior distributions, *Annals of Statistics*, **29**, 687–714.

Susarla, V. and Van Ryzin, J. (1976). Nonparametric Bayesian estimation of survival curves from incomplete observations, *Journal of the American Statistical Association*, **71**, 897–902.

Walker, S. G. (2003). On sufficient conditions for Bayesian consistency, *Biometrika*, **90**, 482–488.

Walker, S. G. (2004). A new approach to Bayesian consistency, *Annals of Statistics*, **32**, 2028–2043.

Walker, S. and Hjort, N. L. (2001). On Bayesian consistency, *Journal of the Royal Statistical Society*, **63**, 811–821.

Walker, S. and Muliere, P. (1997). Beta-Stacy processes and a generalization of the Polya-urn scheme, *Annals of Statistics*, **25**, 1762–1780.

Wolpert, R.L. and Ickstadt K. (1998). Simulation of Lévy random fields, In *Practical Nonparametric and Semiparametric Bayesian Statistics*, (eds. Dey, D., Muller, P. and Sinha, D.), 227–242.

베타과정과 베이지안 생존분석

김용대[a,1] · 최민우[a]

[a]서울대학교 통계학과

(2014년 10월 21일 접수, 2014년 11월 14일 수정, 2014년 11월 19일 채택)

요 약

Hjort (1990)가 제안한 베타과정은 베이지안 생존분석 또는 사건사 분석에서 널리 쓰이는 사전분포이다. 본 논문은 베타과정에 대한 최신 이론과 이를 기반으로 하는 베이지안 생존자료분석 방법을 주로 다룬다. 구체적으로는 베타과정의 생성법, 사후 분포, 대표본 이론, 베이지안 계산법, 혼합베타과정 등을 소개하기로 한다.

주요용어: 베타 과정, 사건사 분석, 위험률, 생존분석.

[1]교신저자: (151-747) 서울특별시 관악구 관악로 1, 서울대학교 통계학과. E-mail: ydkim@gmail.com