碩 士 學 位 論 文

# 소량자료를 위한 베이지안 다중 변환점 모형


高麗大學校 大學院

經濟統計學科

兪 文 星

2012年 06月

全 秀 榮 敎 授 指 導

碩 士 學 位 論 文

# 소량자료를 위한 베이지안 다중 변환점 모형

이 論文을 經濟學 碩士學位 論文으로 提出함.

2012年 06月

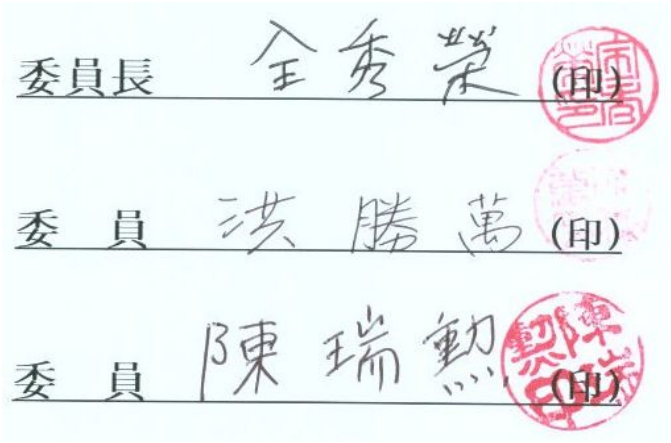
高 麗 大 學 校 大 學 院

經 濟 統 計 學 科

兪 文 星 (印)

俞 文 星의 經濟學 碩士學位論文
審査를 完了함.

2012年 06月



# 요 약 문

일반적으로 자료는 대량자료와 소량자료로 나눌 수 있다. 대량자료에 대한 연구는 아주 많지만 소량자료에 관한 연구는 많지 않다. 소량자료(small data)의 다중 변환점(change-point) 추정은 베이지안 통계학에서 중요한 부분이다. 본 논문은 소량자료에 적합이 잘 되는 베이지안 t 분포를 제안한다. 이를 해결하기 위하여 본 논문에서는 메트로폴리스-해스팅스를 포함한 깁스 샘플링(Metropolis Hastings-Within-Gibbs sampling) 알고리즘을 이용하여 해결하고자 한다. 이 방법을 이용하여 소량자료의 변환점의 개수 및 위치를 추정할 수 있다.

본 논문에서 t 분포와 메트로폴리스-해스팅스를 포함한 깁스 샘플링 알고리즘의 개념을 살펴보고 새로 만든 모형을 이용하여 모의실험자료 및 실제자료 즉 태풍 발생 수의 자료와 한우의 지방함량자료를 분석했다.

# 목 차

# 표 목 차

# 그림 목차

# 제 1장 서 론

온난화 현상을 일으키는 원인 중 온실효과를 일으키는 온실기체가 가장 중요한 원인으로 꼽힌다. 온실기체로는 이산화탄소가 가장 대표적이며 인류의 산업화와 함께 그 양은 계속 증가하고 있다. 이외에도 메텐, 수증기가 대표적인 온실기체이다. 또한, 자연적 요인인 극심한 가뭄과 장기간에 걸친 건조화 현상 그리고 인위적 요인인 과도한 경작 및 관개, 산림벌채, 환경오염으로 인한 기후변화 등 원인으로 지구의 온난화가 빠른 속도로 진행되고 있다. 이와 같은 지구 온난화의 상황이 지속되면서 이상 기상현상들이 지구 상 각지에서 빈발함에 따라 이러한 현상을 보다 정확히 파악하기 위해 지구 온난화와 태풍과의 관계에 대하여 Sugi et al.(2002)이 수치실험을 통해 태풍발생의 변환점을 찾는 연구를 진행하여 왔다.

한우의 지방함량에 따른 부드러움(JUICY) 및 육즙(TENDER)은 주로 한우의 맛을 평가한다. 지방함량이 어느 특정한 비율일 때 맛은 가일층 변화한다. 대한민국 한우는 이제 더 이상 우리만의 것이 아니다. 세계가 주목하고 해외시장에서 당당하게 대접받는 한우, 글로벌 한우로 태어나기 위하여 과학자의 노력은 계속되고 있다. 그러므로 한우 지방함량 비율에 대한 변환점을 찾는 것 또한 중요한 과제이다.

앞에서 언급한 자료는 주로 대량자료보다 소량자료가 더 많다. 따라서 정규분포를 가정한 통상적인 통계적 방법으로 정확한 추론을 할 수 없다. 그러므로 소량자료에 대한 분석이 필요하나 현재까지 이에 대한 연구가 미진한 것으로 보여 본 논문에서 제안한 베이지안 t 분포 모형을 이용하여 이를 해결하고자 한다. 기상학, 금융학 등 여러 분야에서 변환점에 대한 관심이 증가하고 있다. 이러한 분야에서 변환점의 개수를 찾고 그 위치를 정확히 찾는 것은 새로운 관심사로 될 수 있다. 변환점을 찾을 때 일반적으로 마코브 연쇄 몬테카를로(Markov chain Monte Carlo, MCMC) 및 깁스 샘플링(Gibbs sampling) 방법을 이용하여 문제를 해결한다. 하나의 변환점인 경우, Hinkley(1970)는 최대우도추정치(maximum likelihood estimate)를 이용하여 변환점을 찾는 연구를 하였고, Smith(1975)는 이항분포와 정규분포에서의 변환점 추론을 위해 베이지안 방법을 제안하였다. Carlin et al. (1992)은 MCMC 알고리즘(Metropolis et al., 1953; Hastings, 1970)을 이용하여 Smith(1975)의 방법을 확정하였다. 다중 변환점 모형인 경우에, Venter and Steel(1996)은 가설검정을 통하여 순차적으로 배열된 관측치들에 대하여 다중 변환점을 발견하였고, Barry and Hartigan(1993)은 곱분할 모형(product partition model)에서 모수들의 평균 변화를 통해 변환점을 찾았으며, Chib(1998)는 전이확률에 관한 마코프 과정을 통하여 잠재(latent)이산 상태변수에 관한 다중 변환점 모형에 대해 베이지안 방법을 이용한 연구를 진행하였다. 또한 최근에 Kim and Cheon(2010)과 Cheon and Kim(2010)은 일변량과 이변량의 다양한 분포에서 베이지안 다중변환점 모형을 제안하였다. 지금까지 제안된 대부분의 베이지안 다중 변환점 모형은 대량자료 분

석을 통해 MCMC와 깁스 샘플링으로  해결하였다. 하지만 소량자료 분석을 위한 베이지안 다중 변환점 모형에 대한 연구는 미진하다.

일반적으로 소량자료는 정규분포를 따르기보다 비중심(noncentral) t 분포를 따르는 경향이 크다. 본 논문에서는 베이지안 다변량 비중심 t 분포 모형을 제안하고 메트로폴리스-해스팅스를 포함한 깁스 샘플링 알고리즘을 이용하여 소량자료에서 변환점 분석을 하고자 한다.

본 논문에서 다루는 실증 분석 자료는 지구 온난화로 인해 발생하는 1951년부터 2008년까지 총 58년간의 태풍 발생수의 일변량 연도별 자료와 2006년도 10군데 서로 다른 부위에서 잘라낸 한우의 이변량 자료이다.

본 논문의 2절에서는 베이지안 다중 변환점 모형 및 베이지안 모형 선택에 대하여 제안하고, 3절에서는 일변량 베이지안 다중 변환점 모형을 제안하여 모의실험 및 실증 분석을 하였고, 4절에서는 이변량 베이지안 다중 변환점 모형을 제안하고 모의실험 및 실증 분석을 하였다. 5절에서는 본 논문의 결론을 정리하였다.

# 제 2장 베이지안 다중 변환점 분석

## 2.1 베이지안 다중 변환점 모형

$X$가 비중심 t 분포를 따를 때 $X=(\overrightarrow{x_1},\overrightarrow{x_2},...,\overrightarrow{x_n})'$를 순서에 따라 독립적으로 관측된 자료라고 한다. 변환점이 없을 때 $X$의 확률분포(probability density function)는 다음과 같다.

$$f(X|p,\overrightarrow{u},\Sigma)\propto|\Sigma|^{-1/2}\left[1+\frac{(X-\overrightarrow{u})'\Sigma^{-1}(X-\overrightarrow{u})}{p}\right]^{-(p+d)/2}, \qquad (2.1)$$

여기서 $p$는 자유도, $\overrightarrow{u}$는 비중심 모수, $\Sigma$는 분산-공분산 행렬이고 $d$는 차원수이다.

전체영역 S={1,2,...,$n$}'은 순서에 따라 서로 다른 분포를 가지는 여러 부분영역으로 나누어 지고 각각의 부분영역들은 변환점에 의해 분할된다고 하자. $T=(t_1,t_2,...,t_n)$는 $t_{c_1}=t_{c_2}=...=t_{c_k}=1$ 또는 0인 이항 벡터라고 하자. $0=c_0<c_1<c_2<...<c_{k+1}=n$이다. 변환점이 $k$개 있을 때 다중 변환점 모형을 다음과 같이 정의한다.

$$f(\overrightarrow{x})=\prod_{i=1}^{n}f_r(\overrightarrow{x_i}), \quad where \quad \overrightarrow{x_i}\sim f_r(.|\phi_r), \quad c_{r-1}<i<c_r, \qquad (2.2)$$

이때 $r=1,2,...,k+1$이고 $f_r$은 모수 $\phi_r\in\Phi$에 의존한다. 각 자료들은 $c_1+1,c_2+1,...,c_{k+1}$에서 변화하므로 $c_1,c_2,...,c_k$를 $k+1$개의 부분공간으로 분할하는 변환점

(change-points)이라고 부른다.

만약 $f_r$함수가 $d$차원 다변량 비중심 t 분포라고 하고 모수 $\phi_r$는 $d$차원 자유도 $p$이고 위치벡터 $\vec{u}$와 분산행렬 $\Sigma$로 나뉜다. $T^{(k)}$를 $X$의 $k$차 변환점 형태라고 할 때 전체 모수 $\eta^{(k)} = (T^{(k)}, p_1, \overrightarrow{u_1}, \Sigma_1, ..., p_{k+1}, \overrightarrow{u_{k+1}}, \Sigma_{k+1})$이다.

본 논문에서는 식 (2.1)에서 일변량과 이변량 비중심 t 분포에 관하여 분석하였다. 즉, $(d = 1,2)$.

### 2.2 베이지안 모형 선택

결합분포에서 일부 자료를 직접 얻기 힘든 경우가 발생했을 때 일반적으로 채택-기각(acceptance-rejection) 알고리즘과 같은 맞춤형(customized) 알고리즘을 이용하여 해결하려고 노력한다. Müller(1991,1993)가 제안한 절충(compromised) 깁스 알고리즘 즉 메트로폴리스-해스팅스를 포함한 깁스 샘플링 알고리즘을 이용한다. 메트로폴리스-해스팅스를 포함한 깁스 샘플링 알고리즘의 자료생성 절차는 다음과 같다.

메트로폴리스-해스팅스를 포함한 깁스 샘플링:
$i = 1,...,M$이고 현 상태는 t번째 반복 후에 $z = \left(z_1^{(t+1)}, ..., z_{i-1}^{(t+1)}, z_i^{(t)}, ..., z_M^{(t)}\right)$가 주어졌다고 하자.

1. $z_i^* \sim q_i\left(z_i | z_1^{(t+1)}, ..., z_{i-1}^{(t+1)}, z_i^{(t)}, ..., z_M^{(t)}\right)$를 생성한다.
2. $r = \dfrac{f_i\left(z_i^* | z_1^{(t+1)}, ..., z_{i-1}^{(t+1)}, z_{i+1}^{(t)}, ..., z_M^{(t)}\right)}{f_i\left(z_i^{(t)} | z_1^{(t+1)}, ..., z_{i-1}^{(t+1)}, z_{i+1}^{(t)}, ..., z_M^{(t)}\right)} \times \dfrac{q_i\left(z_i^{(t)} | z_1^{(t+1)}, ..., z_{i-1}^{(t+1)}, z_i^*, z_{i+1}^{(t)}, ..., z_M^{(t)}\right)}{q_i\left(z_i^* | z_1^{(t+1)}, ..., z_{i-1}^{(t+1)}, z_i^{(t)}, z_{i+1}^{(t)}, ..., z_M^{(t)}\right)}$를 계산한다.
3. $\min(1,r)$의 확률을 가지고 채택이면 $z_i^{(t+1)} = z_i^*$이고, 기각이면 $z_i^{(t+1)} = z_i^{(t)}$이다.

이 알고리즘에서 메트로폴리스-해스팅스 절차는 각각의 반복에서 오직 한 번만 시행한다. 만약 메트로폴리스-해스팅스 절차를 각각의 반복에서 여러 번 시행한다면 $f_i(\bullet)$는 더 정확한 근사치가 나올 것이다. 하지만 Chen과 Schmeiser(1998)에 의해 오직 한 번의 절차만 시행해도 좋은 결과를 얻을 수 있다.

## 제 3장 일변량 베이지안 다중 변환점 모형

### 3.1 베이지안 일변량 비중심(noncentral) t 분포 다중 변환점 모형

$N$을 평균이 0이고 분산이 1인 정규분포의 확률변수라고 하고, $V$는 $N$과 독립이고

자유도가 $p$인 카이제곱분포의 확률변수라고 하자. 이때 $X=\frac{N-u}{\sqrt{V/p}}$는 자유도가 $p$이고 비중심 모수가 $u$인 비중심 t 분포의 확률변수이다. 즉 $\vec{\phi}=(p,u)'$라 하면 $X \sim t(\vec{\phi})$이다. 비중심 t 분포의 확률분포(probability density function)는 아래와 같다.

$$f(x|p,u)=\frac{p^{p/2}\exp(-\frac{pu^2}{2(x^2+p)})}{\sqrt{\pi}\,\Gamma(p/2)2^{(p-1)/2}(x^2+p)^{(p+1)/2}}\int_0^\infty y^p\exp(-\frac{1}{2}(y-\frac{ux}{\sqrt{x^2+p}})^2)dy. \quad (3.1)$$

식 (2.2)에서 $f_r$은 모수가 $\vec{\phi}=\overrightarrow{\phi_r}=(p_r,u_r)$인 비중심 t 분포를 따르는 함수라 하고, $T^{(k)}$는 $k$차 변환점의 형태라 하며 $\eta^{(k)}=(T^{(k)},p_1,u_1,...,p_{k+1},u_{k+1})$라고 하자.

베이지안 추론을 위해 $u_i(i=1,...,k+1)$의 사전분포(prior distribution)로 균일분포(uniform)를 사용하였다. $p_i$는 자유도이므로 $p_i=n_i-1$ 즉 $p_i=c_i-c_{i-1}-1$이다. 따라서, 자유도 $p_i$ 대신 변환점 위치 $c_i$를 이용하면 $c_i$의 사전확률분포로 이산균일분포(discrete uniform distribution)를 선택하여 자유도 $p_i$를 구할 수 있다.

변환점이 $k$개로 주어졌을 때, 변환점 위치는 $c_1,c_2,...,c_k$등 $k$개이고, 비중심 모수는 $u_1,u_2,...u_{k+1}$등 $k+1$개이다. $X$의 우도함수(likelihood function)는

$$L(\eta^{(k)}|X)=\prod_{j=c_0+1}^{c_1} f_1(x_j|p_1,u_1)\times...\times\prod_{j=c_k+1}^{c_{k+1}} f_{k+1}(x_j|p_{k+1},u_{k+1}). \quad (3.2)$$

$\eta^{(k)}$의 사전확률분포는

$$\pi_1(\eta^{(k)})=\frac{1}{(b-a)^{k+1}}\times\frac{1}{(n-5)^k}\times I_{(a,b)}(u_1,u_2,...,u_{k+1})\times I_{[3,..,n-3]}(c_1,c_2,...,c_k). \quad (3.3)$$

여기서 $a<b$는 비중심 모수 $u_i$의 사전확률분포 구간이다.

$X$의 사후확률분포(posterior distribution)는

$$\pi_1(\eta^{(k)}|X)=\frac{L(\eta^{(k)}|X)\times\pi_1(\eta^{(k)})}{\pi_1(X)}\propto L(\eta^{(k)}|X)\times\pi_1(\eta^{(k)}). \quad (3.4)$$

예를 들면 변환점이 2개로 주어졌을 때, 변환점 위치는 $c_1,c_2$ 2개이고, 비중심 모수는 $u_1,u_2,u_3$ 3개이다. $X$의 우도함수는

$$L(\eta^{(2)}|X)=\prod_{j=1}^{c_1} f_1(x_j|p_1,u_1)\times\prod_{j=c_1+1}^{c_2} f_2(x_j|p_2,u_2)\times\prod_{j=c_2+1}^{n} f_3(x_j|p_3,u_3) \quad (3.5)$$

이고, $\eta^{(2)}$의 사전확률분포는 다음과 같다.

$$\pi_1(\eta^{(2)}) = \frac{1}{(b-a)^3} \times \frac{1}{(n-5)^2} \times I_{(a,b)}(u_1,u_2,u_3) \times I_{[3,\ldots,n-3]}(c_1,c_2). \tag{3.6}$$

(3.5)와 (3.6)를 이용하여 $\eta^{(2)}$의 사후분포인 베이지안 t 분포 모형을 구하면 다음과 같다.

$$\pi_1(\eta^{(2)}|X) \propto L(\eta^{(2)}|X) \times \pi(\eta^{(2)}). \tag{3.7}$$

본 논문은 $X$의 사후확률분포 $\pi_1(\eta^{(k)}|X)$를 메트로폴리스-해스팅스를 포함한 깁스 샘플링을 이용하여 사후분포확률과 BIC(Bayesian Information Criterion)를 계산하여 변환점을 찾고자 한다.

$$BIC = -2(\log(\text{최대우도값})) + (\log(\text{자료의 수}))(\text{모수의수}). \tag{3.8}$$

### 3.2 일변량 베이지안 모형 선택

일변량 베이지안 모형인 경우, 비중심 모수는 $u_1,u_2,\ldots,u_{k+1}$이고 변환점 위치는 $c_1,c_2,\ldots,c_k$ 인 k개의 변환점이 주어진 경우를 고려해 보자. 이때 두 변수들을 통합하여 새로운 변수 $z=(z_1,\ldots,z_{k+1},z_{k+2},\ldots,z_{2k+1})=(u_1,\ldots,u_{k+1},c_1,\ldots,c_k)$를 설정한다. 이때 2.2절의 베이지안 모형 분석에 메트로폴리스-해스팅스를 포함한 깁스 샘플링을 이용한다. 일변량 베이지안 t 분포 모형을 이용하여 아래 모의실험과 실증 분석을 한다.

메트로폴리스-해스팅스를 포함한 깁스 샘플링:
$i=1,\ldots,2k+1(=M)$이고 현상태는 t번째 반복후에 $z=\left(z_1^{(t+1)},\ldots,z_{i-1}^{(t+1)},z_i^{(t)},\ldots,z_{2k+1}^{(t)}\right)$가 주어졌다고 하자.

1. $z_i^* \sim q\left(z_i|z_1^{(t+1)},\ldots,z_{i-1}^{(t+1)},z_i^{(t)},\ldots,z_{2k+1}^{(t)}\right)$를 생성한다.

2. $$r=\frac{\pi_1\left(z_i^*|z_1^{(t+1)},\ldots,z_{i-1}^{(t+1)},z_{i+1}^{(t)},\ldots,z_{2k+1}^{(t)}\right)}{\pi_1\left(z_i^{(t)}|z_1^{(t+1)},\ldots,z_{i-1}^{(t+1)},z_{i+1}^{(t)},\ldots,z_{2k+1}^{(t)}\right)} \times \frac{q\left(z_i^{(t)}|z_1^{(t+1)},\ldots,z_{i-1}^{(t+1)},z_i^*,z_{i+1}^{(t)},\ldots,z_{2k+1}^{(t)}\right)}{q\left(z_i^*|z_1^{(t+1)},\ldots,z_{i-1}^{(t+1)},z_i^{(t)},z_{i+1}^{(t)},\ldots,z_{2k+1}^{(t)}\right)}$$
를 계산한다.

3. $\min(1,r)$의 확률을 가지고 채택이면 $z_i^{(t+1)}=z_i^*$이고, 기각이면 $z_i^{(t+1)}=z_i^{(t)}$이다.

### 3.3 모의실험

3.3에서는 앞에서 설명한 알고리즘들을 좀 더 쉽게 비교 분석하기 위해 각 알고리즘

들을 이용하여 모의실험을 진행하였다. 모의실험을 위한 통계 패키지로는 R-software 2-14-2 버전을 사용하였다.

### 3.3.1 모의실험 자료

본 모의실험으로 60개 독립적 자료를 각각 20개씩 3개 부분으로 나누어 차례로 생성한다. (그림 3.1) ; 즉, $x_1,...,x_{20} \sim T(p_1 = 19, u_1 = 4)$ ; $x_{21},...,x_{40} \sim T(p_2 = 19, u_2 = 7)$ ; $x_{41},...,x_{60} \sim T(p_3 = 19, u_3 = 2)$.

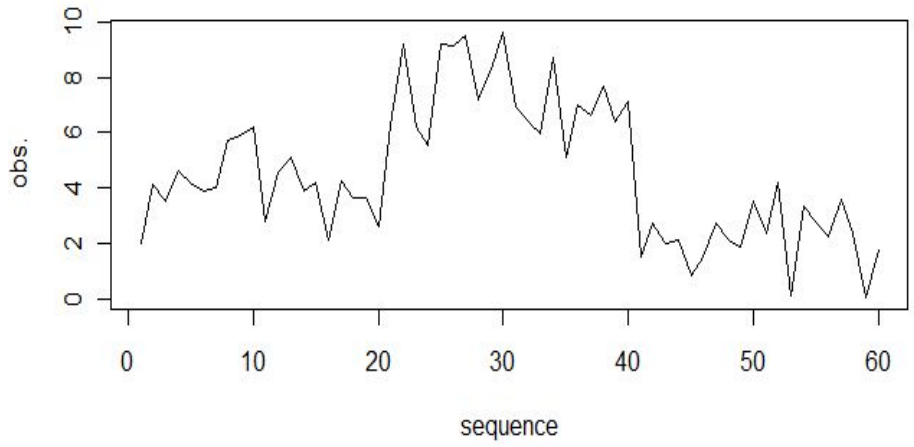



그림 3.1: 모의실험 자료의 그래프

### 3.3.2 모의실험 결과

본 모의실험은 변환점이 각각 1개, 2개, 3개인 경우만을 고려하였다. <표 3.1>은 메트로폴리스-해스팅스를 포함한 깁스 샘플링 알고리즘을 이용하여 변환점이 1개, 2개, 3개인 경우에 대해 각각 10000번씩 시행하여 로그 사후확률이 가장 큰 10개 값으로부터 얻은 결과이다. <표 3.1>로부터 2개의 변환점에서의 사후확률이 다른 변환점에서의 사후확률보다 크며, 변환점이 2개인 경우중 비중심 모수 값에 따라 로그 사후확률이 달라지는데, 비중심 모수$(u_1, u_2, u_3)$가 (3.9406, 7.1919, 2.1164)일 때 로그 사후확률값이 -105.6310으로 가장 크고 BIC값도 223.5450으로 가장 작게 얻어졌다. 따라서, 사후확률과 BIC에 의해 최적의 모형은 변환점 위치$(c_1, c_2)$는 (20, 40)이고 비중심 모수 (3.9406, 7.1919, 2.1164)일 때 얻어진다.

표 3.1: 모의실험 결과

| No | CP=1 | | | |
|---|---|---|---|---|
| | Change patterns | noncentral parameters | Log−posterior | BIC |
| 1 | (40) | (5.3725,2.1198) | −140.7706 | 289.7300 |
| 2 | (40) | (5.3605,2.1096) | −140.7715 | 289.7318 |
| 3 | (40) | (5.3546,2.1184) | −140.7723 | 289.7333 |
| 4 | (40) | (5.3626,2.1041) | −140.7724 | 289.7335 |
| 5 | (40) | (5.3769,2.1066) | −140.7730 | 289.7347 |
| 6 | (40) | (5.3546,2.1093) | −140.7732 | 289.7351 |
| 7 | (40) | (5.3812,2.1153) | −140.7733 | 289.7353 |
| 8 | (40) | (5.3769,2.1331) | −140.7734 | 289.7355 |
| 9 | (40) | (5.3826,2.1099) | −140.7745 | 289.7377 |
| 10 | (40) | (5.3745,2.0986) | −140.7747 | 289.7381 |
| No | CP=2 | | | |
| | Change patterns | noncentral parameters | Log−posterior | BIC |
| 1 | (20,40) | (3.9406,7.1919,2.1164) | −105.6310 | 223.5450 |
| 2 | (20,40) | (3.9333,7.1344,2.0980) | −105.6450 | 223.5729 |
| 3 | (20,40) | (3.9717,7.1591,2.1402) | −105.6510 | 223.5850 |
| 4 | (20,40) | (3.9024,7.2333,2.0868) | −105.6522 | 223.5874 |
| 5 | (20,40) | (3.9815,7.1633,2.1164) | −105.6533 | 223.5896 |
| 6 | (20,40) | (3.9394,7.1707,2.1695) | −105.6539 | 223.5908 |
| 7 | (20,40) | (3.9815,7.1633,2.1088) | −105.6541 | 223.5912 |
| 8 | (20,40) | (3.8658,7.1971,2.1111) | −105.6564 | 223.5958 |
| 9 | (20,40) | (3.8658,7.2399,2.1111) | −105.6684 | 223.6198 |
| 10 | (20,40) | (3.9220,7.1160,2.1664) | −105.6701 | 223.6232 |
| No | CP=3 | | | |
| | Change patterns | noncentral parameters | Log−posterior | BIC |
| 1 | (3,20,40) | (3.0559,4.0387,7.2331,2.0125) | −108.0118 | 232.4010 |
| 2 | (3,20,40) | (2.8027,4.0387,7.2331,2.0125) | −108.0381 | 232.4536 |
| 3 | (3,20,40) | (3.1860,4.0523,7.0899,2.2244) | −108.0475 | 232.4723 |
| 4 | (3,20,40) | (2.8015,4.0523,7.0899,2.2244) | −108.0621 | 232.5017 |
| 5 | (3,20,40) | (3.2764,4.0523,7.0899,2.2244) | −108.0656 | 232.5085 |
| 6 | (3,20,40) | (3.3869,4.0387,7.2331,2.0125) | −108.0747 | 232.5268 |
| 7 | (3,20,40) | (2.7049,4.0523,7.0899,2.2244) | −108.0901 | 232.5576 |
| 8 | (3,20,40) | (3.4416,4.0387,7.2331,2.0125) | −108.0954 | 232.5682 |
| 9 | (3,20,40) | (3.4370,4.0523,7.0899,2.2244) | −108.1173 | 232.6120 |
| 10 | (3,20,40) | (2.7949,3.9785,7.1432,1.9926) | −108.1200 | 232.6174 |

변환점 위치의 상대빈도 히스토그램으로부터 최적의 모형이 (20, 40)임을 알 수 있고, 또한 최적의 모형이 정확히 세 개의 서로 다른 이질적인 영역으로 분리하고 있음을 알 수 있다(그림 3.2). 따라서 제안된 베이지안 모형에 메트로폴리스-해스팅스를 포함한

깁스 샘플링 알고리즘을 이용하면 정확히 변환점 개수 및 위치를 찾아 낼 수 있음을 알 수 있다.

<그림 3.2(a)>에서는 변환점이 2개인 경우 변환점 위치의 상대빈도 히스토그램을 보여주며, <그림 3.2(b)>에서는 로그 사후확률이 가장 크고 BIC가 가장 작을때의 최적의 모형인 변환점위치 (20, 40)을 보여주고 있다.

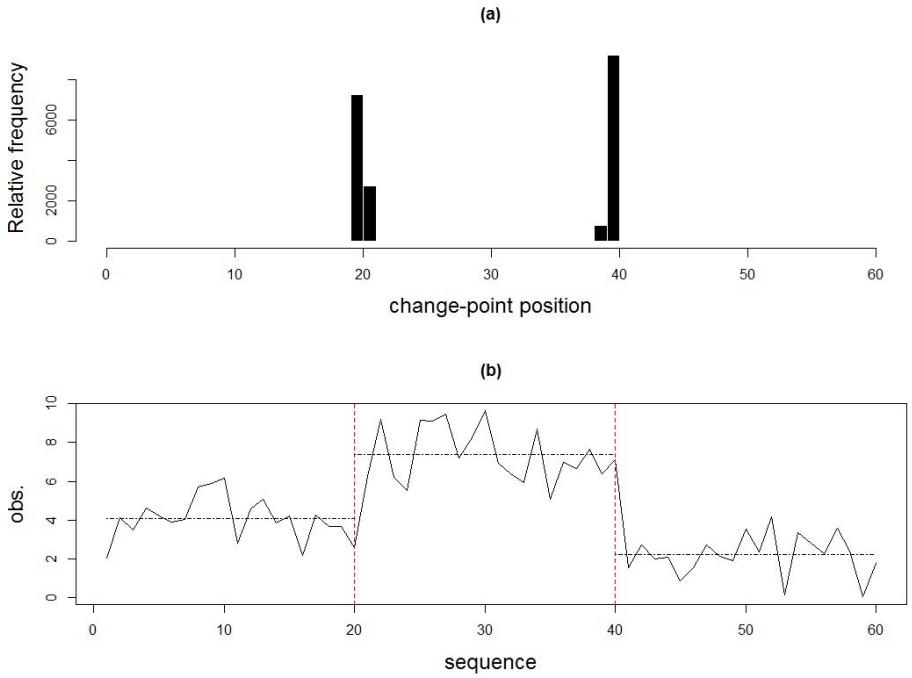



그림 3.2: (a):변환점위치 상대빈도; (b):최대 로그 사후확률값의 변환점 위치 및 평균값

## 3.4 실증 분석

### 3.4.1 실증 분석 자료

변환점을 검출하는 방법은 기상학에서 체제적 변화를 찾는데서 아주 유용하게 쓰인다. 특히, 강수량, 기압, 습도, 온도 등의 방면에서 자주 쓰인다. 기상학의 체제적 변화는 측정하는 과정이나 계절요인을 이유로 하여 생긴다. 이러한 특징을 알아보기 위해 본 논문에서 변환점을 추정하고자 한다. 기상학 자료는 1951년부터 최근인 2008년까지 총 58년간 세계에서 매년 일어나는 태풍 발생 수에 대한 자료이다. 이 자료를 이용하여 태풍횟수에 대한 변화 경향을 분석하고자 한다. 본 연구를 위하여 사용한 자료는 한국기상청에서 제공하는 기상연보에서의 자료이다.

(그림 3.3)은 1951-2008년의 58년간의 자료에 대하여 태풍의 연별 발생수를 시간에 따라 그래프로 나타낸 것이다. 태풍의 연별 발생 수는 시간의 흐름에 따라 서서히 감소하고 있는 경향이 보인다. 그 변화시점이 대략 1969년(19시점)과 1997년(47시점)인 것으로 추정된다. 이는 변환점을 위에서 제안한 모형을 이용하여 정확히 찾으려 한다.

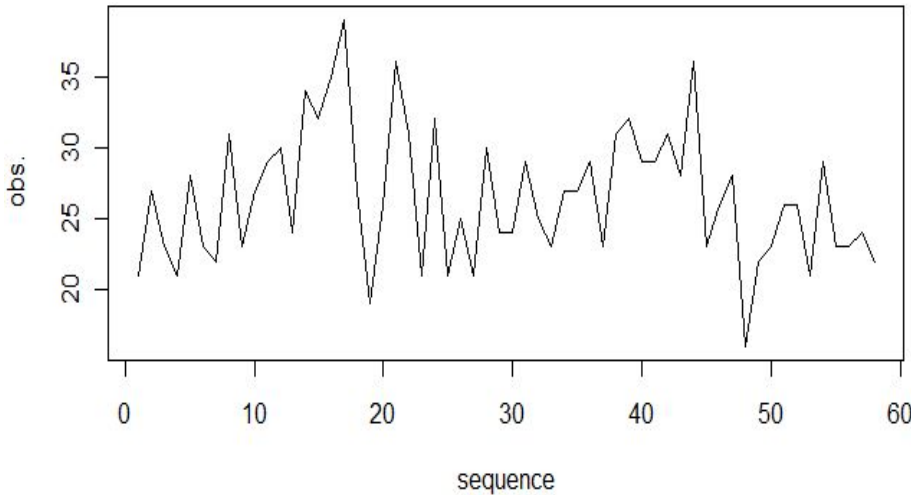



그림 3.3: 1951-2008년 태풍 발생 수의 시계열 자료의 그래프

### 3.4.2 실증 분석 결과

<표 3.2>는 모의실험과 마찬가지로 메트로폴리스-해스팅스를 포함한 깁스 샘플링 알고리즘을 이용하여 변환점이 1개, 2개, 3개인 경우에 각각 10000번씩 시행하여 로그 사후확률이 가장 큰 10개 값을 찾아서 그 값에 해당되는 변환점 위치 및 로그 사후확률(log-posterior) 및 BIC를 구한 결과를 보여 준다 (그림 3.4). 로그 사후확률과 BIC값을 보면 변환점이 2개이며 비중심 모수 $(u_1, u_2, u_3)$가 (25.5302, 26.3424, 22.6523)일 때, 로그 사후확률값이 -179.4510으로 가장 크게 얻어지고 BIC값도 371.0833으로 가장 작게 얻어짐을 알 수 있다. 따라서 변환점 위치$(c_1, c_2)$가 (19, 47)인 변환점 2개의 모형이 최적의 모형으로 선택되었다.

<그림 3.4(a)>에서 변환점이 2개인 경우 변환점 위치의 상대빈도 히스토그램을 보여주며, 또한 <그림 3.4(b)>에서 로그 사후확률이 가장 크고 BIC가 가장 작을때의 최적의 모형인 변환점위치 (19, 47)을 보여주고 있다.

표 3.2: 실증 분석 결과

| No | CP=1 | | | |
|---|---|---|---|---|
| | Change patterns | noncentral parameters | Log-posterior | BIC |
| 1 | (24) | (26.1189,24.8816) | −182.1526 | 372.4262 |
| 2 | (24) | (26.0884,24.9193) | −182.1530 | 372.4270 |
| 3 | (24) | (26.0795,24.8924) | −182.1539 | 372.4287 |
| 4 | (24) | (26.2014,24.9157) | −182.1551 | 372.4311 |
| 5 | (24) | (26.0640,24.9591) | −182.1592 | 372.4393 |
| 6 | (24) | (26.0640,24.8533) | −182.1598 | 372.4405 |
| 7 | (24) | (26.0935,24.9862) | −182.1626 | 372.4461 |
| 8 | (24) | (26.0433,24.9724) | −182.1642 | 372.4493 |
| 9 | (24) | (26.0243,24.9724) | −182.1671 | 372.4551 |
| 10 | (24) | (26.3064,24.9416) | −182.1762 | 372.4733 |
| No | CP=2 | | | |
| | Change patterns | noncentral parameters | Log-posterior | BIC |
| 1 | **(19,47)** | **(25.5302,26.3424,22.6523)** | **−179.4510** | **371.0833** |
| 2 | (19,47) | (26.0828,26.3424,22.6523) | −179.4872 | 371.1557 |
| 3 | (19,47) | (25.4457,26.7655,22.1363) | −179.4872 | 371.1558 |
| 4 | (19,47) | (25.4910,26.7655,22.9298) | −179.4930 | 371.1674 |
| 5 | (19,47) | (26.0667,26.7655,22.1363) | −179.4972 | 371.1758 |
| 6 | (19,47) | (25.4658,26.7655,22.9298) | −179.4995 | 371.1803 |
| 7 | (19,47) | (25.9233,26.2090,22.3891) | −179.5167 | 371.2147 |
| 8 | (19,47) | (25.4910,26.7655,21.9116) | −179.5184 | 371.2181 |
| 9 | (19,47) | (25.4658,26.8395,22.9298) | −179.5356 | 371.2525 |
| 10 | (19,47) | (26.0667,26.7655,21.9116) | −179.5406 | 371.2625 |
| No | CP=3 | | | |
| | Change patterns | noncentral parameters | Log-posterior | BIC |
| 1 | (13,24,47) | (24.3419,27.6084,26.5967,22.3339) | −180.4592 | 377.1602 |
| 2 | (13,24,47) | (24.3419,28.6207,26.5967,22.3339) | −180.4629 | 377.1676 |
| 3 | (13,24,47) | (23.9560,28.1323,26.2934,22.5597) | −180.4732 | 377.1882 |
| 4 | (13,24,47) | (24.6189,27.6926,26.2109,23.2338) | −180.4832 | 377.2082 |
| 5 | (13,24,47) | (24.8298,27.8013,26.8417,22.9895) | −180.5980 | 377.4378 |
| 6 | (13,25,47) | (24.6827,27.0277,26.4670,22.0042) | −180.6084 | 377.4586 |
| 7 | (13,24,47) | (25.1673,27.4283,25.8481,22.9008) | −180.6337 | 377.5092 |
| 8 | (13,24,47) | (25.1673,27.4283,25.8481,21.9808) | −180.6488 | 377.5394 |
| 9 | (13,25,47) | (25.3218,27.6263,26.4961,22.2856) | −180.6765 | 377.5948 |
| 10 | (13,24,47) | (25.6534,27.6926,26.2109,23.2338) | −180.7030 | 377.6478 |

변환점의 위치가 (19, 47)인 시점은 각각 1969년과 1997년이다. 설동일(2010)은 연구 기간의 전기 20년간(1951년-1970년)의 연 평균 태풍 발생이 27.2건인데 비하여, 후기

20년간(1989년-2008년)의 연 평균 발생은 25.9건으로 1.3건의 감소를 보인다. 본 논문에서 제안한 베이지안 모형을 이용하면 첫 번째 변환점 위치는 정확히 일치함을 알았지만 두 번째 변환점 위치는 약간의 차이를 보인다. 이는 아마도 1951년부터 2008년 전체 자료를 사용한 것이 아니고 앞부분과 뒷부분 각각 20년 자료만 사용한 것으로 판단된다. 만약 전체자료를 사용하면 <그림 3.4>에서와 같이 두 번째 변환점 위치는 1989년보다 1997년이 더욱 적합하다고 본다.

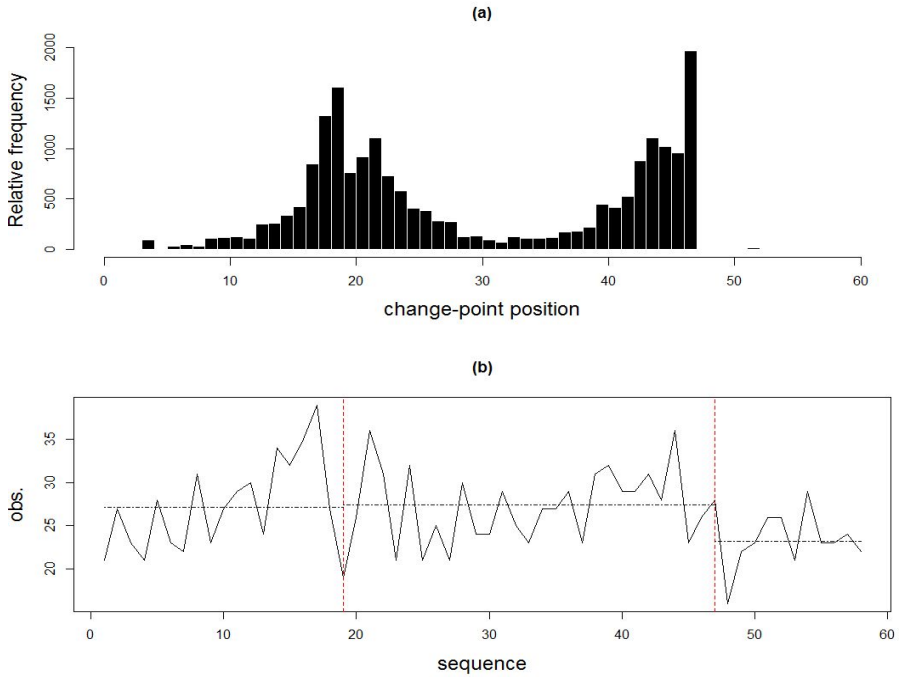



그림 3.4: (a):변환점위치 상대빈도; (b):최대 로그 사후확률값의 변환점 위치 및 평균값

조사 자료에 의하면 사하라 사막의 주변에는 매년 적어도 25만 에이커의 토지가 사막으로 변하고 있고 남미 칠레 북부의 사막에서는 최근 10년간 사막이 100km 남쪽으로 확장되었으며 미국의 애리조나 주 뉴멕시코의 사막에서도 이와 같은 규모의 사막화 현상이 일어나고 있다. 1969년에는 사막화가 점차 증가되어 사하라 사막에서 모래바람으로 잃어버린 토지의 양이 최고 6천만 톤에 달하고 있다. 이와 같이 지구 온난화 현상은 점점 심각해지고 있으며 태풍의 발생수도 1970년부터 상승하는 추세를 보이고 있다. 이러한 현상을 줄이기 위해 온실효과 방지를 위한 국제간의 공동노력의 일환으로 "기후변화에 관한 교토의정서"가 1997년에 채택되었다. 이후 지구 온난화 현상이 조금씩 완화되어 태풍의 발생수가 1998년부터 하강하는 추세를 보이고 있다. 이러한 결과는 변

환점의 개수 및 그 시점을 볼 때 본 논문에서 제안한 모형에 적합하여 얻은 결과와 거의 비슷하게 보여주고 있음을 알 수 있다.

# 제 4장 이변량 베이지안 다중 변환점 모형

## 4.1 베이지안 이변량 t 분포 다중 변환점 모형

$\vec{\phi}=(p,\vec{u},\Sigma)$라 하면 $Y(=\vec{y_i})$는 비중심 모수가 $\vec{u}$인 이변량 비중심 t 분포의 확률변수이다. 즉 $\vec{y_i}\sim BT(\vec{\phi})$이다. 변환점이 k개 있을 때, 이변량 t 분포의 확률분포는 아래와 같다. $\vec{u}$는 $2\times1$인 벡터이고 $\Sigma$는 $2\times2$인 행렬이다.

$$f(\vec{y})=\prod_{i=1}^{n} f_r(\vec{y_i}|p_r,\vec{u_r},\Sigma_r)\,,\qquad where$$

$$f_r(\vec{y_i}|p_r,\vec{u_r},\Sigma_r)\propto|\Sigma_r|^{-1/2}\left[1+\frac{(\vec{y_i}-\vec{u_r})'\Sigma_r^{-1}(\vec{y_i}-\vec{u_r})}{p_r}\right]^{-(p_r+2)/2},\quad r=1,2,...k+1 \qquad (4.1)$$

식 (2.2)에서 $f_r$은 모수가 $\vec{\phi}=\vec{\phi_r}=(p_r,\vec{\mu_r},\Sigma_r)$인 이변량 t 분포를 따르는 함수라고 하고, $T^{(k)}$는 $k$차 변환점의 형태라고 하며 $\eta^{(k)}=(T^{(k)},p_1,\vec{u_1},\Sigma_1...,p_{k+1},\overrightarrow{u_{k+1}},\Sigma_{k+1})$라고 하자. 이때 $\eta^{(k)}$의 우도함수는 다음과 같이 구할 수 있다.

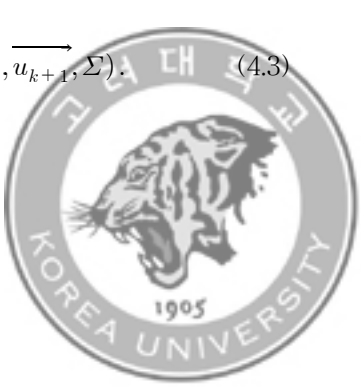
$$L(\eta^{(k)}\mid Y)=\prod_{j=1}^{c_1} f_1(y_j|p_1,\vec{u_1},\Sigma_1)\times...\times\prod_{j=c_k+1}^{n} f_{k+1}(y_j|p_{k+1},\overrightarrow{u_{k+1}},\Sigma_{k+1})\,. \qquad (4.2)$$

베이지안 추론을 위해 $\vec{u_i}(i=1,...,k+1)$의 사전분포로 균일분포를 사용하였고 $\Sigma_i(i=1,...,k+1)$은 똑같은 분산 $\Sigma$로 하였다. $p_i$는 자유도이므로 $p_i=n_i-1$ 즉 $p_i=c_i-c_{i-1}-1$이다. 따라서, 자유도 $p_i$ 대신 변환점 위치 $c_i$를 이용하면 $c_i$의 사전확률분포로 이산균일분포를 선택하여 자유도 $p_i$를 구할 수 있다.

변환점이 $k$개로 주어졌을 때, 변환점 위치는 $c_1,c_2,...,c_k$ $k$개이고, 비중심 모수 벡터는 $\vec{u_1},\vec{u_2},...,\overrightarrow{u_{k+1}}$등 $k+1$개이다. $Y$의 우도함수는

$$L(\eta^{(k)}\mid Y)=\prod_{j=1}^{c_1} f_1(y_j|p_1,\vec{u_1},\Sigma)\times...\times\prod_{j=c_k+1}^{n} f_{k+1}(y_j|p_{k+1},\overrightarrow{u_{k+1}},\Sigma)\,. \qquad (4.3)$$

$\eta^{(k)}$의 사전확률분포는



$$\pi_2(\eta^{(k)}) = |\Lambda_0|^{v_0/2}|\Sigma|^{-(v_0+3)/2} \times \exp(-\frac{1}{2}tr(\Lambda_0\Sigma^{-1})) \tag{4.4}$$
$$\times \frac{1}{(b-a)^{2(k+1)}} \times \frac{1}{(n-5)^k} \times I_{(a,b)}(\overrightarrow{u_1},...,\overrightarrow{u_{k+1}}) \times I_{[3,...,n-3]}(c_1,...,c_k)$$

이고, 분산 $\Sigma$의 사전확률분포는 Inverse-Wishart분포 $IW(v_0,\Lambda_0^{-1})$를 따른다고 가정한다.

$Y$의 사후확률분포은

$$\pi_2(\eta^{(k)}\mid Y) = \frac{L(\eta^{(k)}\mid Y)\times\pi_2(\eta^{(k)})}{\pi_2(Y)} \propto L(\eta^{(k)}\mid Y)\times\pi_2(\eta^{(k)}). \tag{4.5}$$

예를 들면 변환점이 2개로 주어졌을 때, 변환점 위치는 $c_1, c_2$ 2개이고, 비중심 모수 벡터는 $\overrightarrow{u_1},\overrightarrow{u_2},\overrightarrow{u_3}$ 3개이다. $Y$의 우도함수는

$$L(\eta^{(2)}\mid Y) = \prod_{j=1}^{c_1} f_1(y_j|p_1,\overrightarrow{u_1},\Sigma) \times \prod_{j=c_1+1}^{c_2} f_2(y_j|p_2,\overrightarrow{u_2},\Sigma) \times \prod_{j=c_2+1}^{n} f_3(y_j|p_3,\overrightarrow{u_3},\Sigma), \tag{4.6}$$

이고, $\eta^{(2)}$의 사전확률분포는 다음과 같다.

$$\pi_2(\eta^{(2)}) = |\Lambda_0|^{v_0/2}|\Sigma|^{-(v_0+3)/2} \times \exp(-\frac{1}{2}tr(\Lambda_0\Sigma^{-1})) \tag{4.7}$$
$$\times \frac{1}{(b-a)^6} \times \frac{1}{(n-5)^2} \times I_{(a,b)}(\overrightarrow{u_1},\overrightarrow{u_2},\overrightarrow{u_3}) \times I_{[3,...,n-3]}(c_1,c_2).$$

(4.6)와 (4.7)를 이용하여 $\eta^{(2)}$의 사후분포인 베이지안 t 분포 모형을 구하면 다음과 같다.

$$\pi_2(\eta^{(2)}\mid Y) \propto L(\eta^{(2)}\mid Y)\times\pi_2(\eta^{(2)}). \tag{4.8}$$

본 논문에서는 $Y$의 사후확률분포 $\pi_2(\eta^{(k)}\mid Y)$를 메트로폴리스-해스팅스를 포함한 깁스 샘플링을 이용하여 사후분포확률과 BIC를 계산하여 변환점을 찾고자 한다.

$$BIC = -2(\log(\text{최대 우도값})) + (\log(\text{자료의 수}))(\text{모수의 수}) \tag{4.9}$$

### 4.2 이변량 베이지안 모형 선택

이변량 베이지안 모형인 경우, 비중심 모수는 $\overrightarrow{u_1},\overrightarrow{u_2},...,\overrightarrow{u_{k+1}}$이고 변환점 위치는 $c_1,c_2,...,c_k$이고 분산 모수는 $\Sigma$인 k개의 변환점이 주어진 경우를 고려해 보자. 변수들을 합하여 새로운 변수 $z=(z_1,...,z_{k+1},z_{k+2},...,z_{k+1},z_{2k+2})=(u_1,...,u_{k+1},c_1,...,c_k,\Sigma)$를 설정한다. 이때 3장의 베이지안 모형 메트로폴리스-해스팅스를 포함한 깁스 샘플링을

이용한다. 이변량 베이지안 t 분포 모형을 이용하여 아래 모의실험과 실증 분석을 한다.

메트로폴리스-해스팅스를 포함한 깁스 샘플링:
$i=1,...,2k+2(=M)$이고 현 상태는 t번째 반복 후 $z=\left(z_1^{(t+1)},...,z_{i-1}^{(t+1)},z_i^{(t)},...,z_{2k+2}^{(t)}\right)$가 주어졌다고 하자.
1. $z_i^* \sim q\left(z_i | z_1^{(t+1)},...,z_{i-1}^{(t+1)},z_i^{(t)},...,z_{2k+2}^{(t)}\right)$를 생성한다.
2.
$r=\frac{\pi_2\left(z_i^* | z_1^{(t+1)},...,z_{i-1}^{(t+1)},z_{i+1}^{(t)},...,z_{2k+2}^{(t)}\right)}{\pi_2\left(z_i^{(t)} | z_1^{(t+1)},...,z_{i-1}^{(t+1)},z_{i+1}^{(t)},...,z_{2k+2}^{(t)}\right)} \times \frac{q\left(z_i^{(t)} | z_1^{(t+1)},...,z_{i-1}^{(t+1)},z_i^*,z_{i+1}^{(t)},...,z_{2k+2}^{(t)}\right)}{q\left(z_i^* | z_1^{(t+1)},...,z_{i-1}^{(t+1)},z_i^{(t)},z_{i+1}^{(t)},...,z_{2k+2}^{(t)}\right)}$를 계산한다.
3. $\min(1,r)$의 확률을 가지고 채택이면 $z_i^{(t+1)}=z_i^*$이고, 기각이면 $z_i^{(t+1)}=z_i^{(t)}$이다.

## 4.3 모의실험

### 4.3.1 모의실험 자료

본 모의실험으로 60쌍의 이변량 자료를 각각 20쌍씩 3개 부분으로 나누어 차례로 생성한다 (그림 4.1).
즉, $\overrightarrow{y_1},...,\overrightarrow{y_{20}} \sim BT(p_1=19, \overrightarrow{u_1}=(3.31,2.85))$; $\overrightarrow{y_{21}},...,\overrightarrow{y_{40}} \sim BT(p_2=19, \overrightarrow{u_2}=(6.82,6.26))$; $\overrightarrow{y_{41}},...,\overrightarrow{y_{60}} \sim BT(p_3=19, \overrightarrow{u_3}=(11.56,10.65))$.

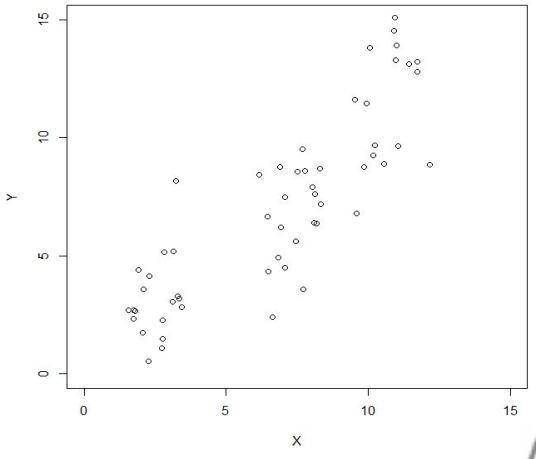


그림 4.1: 모의실험 자료의 그래프

4.3.2 모의실험 결과

본 모의실험은 변환점이 각각 1개, 2개, 3개인 경우만을 고려하였다. <표 4.1>은 메트

표 4.1: 모의실험 결과

| No | CP=1 | | |
|---|---|---|---|
| | Change patterns | Log-posterior | BIC |
| 1 | (20) | -153.9919 | 328.4555 |
| 2 | (20) | -154.1950 | 328.8617 |
| 3 | (20) | -154.2049 | 328.8815 |
| 4 | (20) | -154.3283 | 329.1283 |
| 5 | (20) | -154.3403 | 329.1523 |
| 6 | (20) | -154.5539 | 329.5795 |
| 7 | (20) | -154.7713 | 330.0143 |
| 8 | (20) | -154.7799 | 330.0315 |
| 9 | (20) | -154.8825 | 330.2367 |
| 10 | (20) | -155.0688 | 330.6093 |
| No | CP=2 | | |
| | Change patterns | Log-posterior | BIC |
| 1 | **(20,40)** | **-112.2521** | **261.3533** |
| 2 | (20,40) | -112.2719 | 261.3929 |
| 3 | (20,40) | -112.5306 | 261.9103 |
| 4 | (20,40) | -112.6247 | 262.0985 |
| 5 | (20,40) | -112.8906 | 262.6303 |
| 6 | (20,40) | -113.2024 | 263.2539 |
| 7 | (20,40) | -113.3555 | 263.5601 |
| 8 | (20,40) | -113.3606 | 263.5703 |
| 9 | (20,40) | -113.3764 | 263.6019 |
| 10 | (20,40) | -113.4420 | 263.7331 |
| No | CP=3 | | |
| | Change patterns | Log-posterior | BIC |
| 1 | (20,29,40) | -120.4417 | 290.0155 |
| 2 | (20,29,40) | -120.4860 | 290.1041 |
| 3 | (20,29,40) | -120.5067 | 290.1455 |
| 4 | (20,35,40) | -120.6468 | 290.4257 |
| 5 | (20,35,40) | -120.7028 | 290.5377 |
| 6 | (20,23,40) | -120.7173 | 290.5667 |
| 7 | (20,36,40) | -120.7321 | 290.5963 |
| 8 | (20,35,40) | -120.8092 | 290.7505 |
| 9 | (20,29,40) | -120.8862 | 290.9045 |
| 10 | (20,36,40) | -120.8874 | 290.9069 |

로폴리스-해스팅스를 포함한 깁스 샘플링 알고리즘을 이용하여 변환점이 1개, 2개, 3개인 경우에 대해 각각 10000번씩 시행하여 로그 사후확률이 가장 큰 10개 값으로부터 얻은 결과이다. <표 4.1>로부터 2개의 변환점에서 로그 사후확률이 다른 변환점에서 로그 사후확률보다 크고 BIC값도 가장 작은 것을 보아낼 수 있다. 즉, 변환점 위치 $(c_1, c_2)$가 (20,40)일 때 로그 사후확률값은 -112.2521으로 가장 크고 BIC값도 261.3533으로 가장 작게 얻어졌다. 따라서, 사후확률과 BIC에 의해 최적의 모형은 변환점 위치 (20, 40)일 때이다.

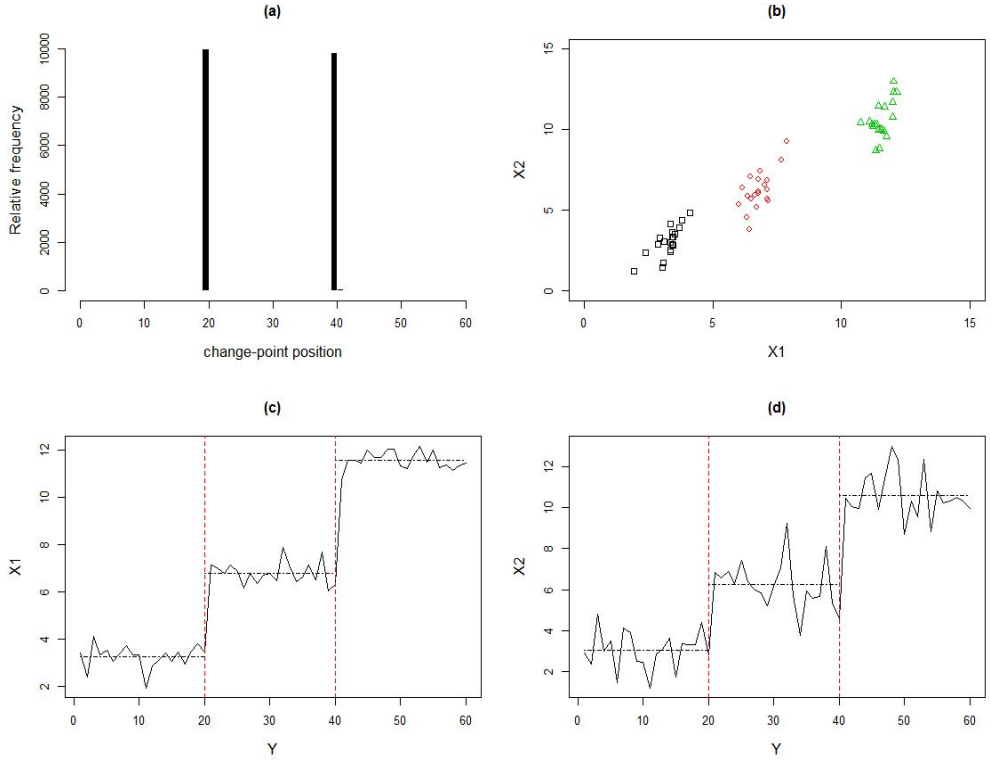



그림 4.2: (a): 변환점위치 상대빈도; (b): 두 변수의 해당 위치 그래프; (c): 변수1의 변환점 위치 및 부분영역의 평균 값; (d): 변수2의 변환점 위치 및 부분영역의 평균 값

변환점 위치의 상대빈도 히스토그램으로부터 최적의 모형이 (20, 40)임을 알 수 있고, 또한 최적의 모형이 정확히 세 개의 서로 다른 이질적인 영역으로 분리하고 있음을 알 수 있다(그림 4.2). 따라서 제안된 베이지안 모형에 메트로폴리스-해스팅스를 포함한 깁스 샘플링 알고리즘을 이용하면 정확히 변환점 개수 및 위치를 찾아 낼 수 있음을 알 수 있다.

## 4.4 실증 분석

### 4.4.1 실증 분석 자료

한우의 맛은 일반적으로 부드러움(JUICY), 육즙(TENDER)으로 맛을 평가한다(Forrest, 1975). 맛을 결정하는 다양한 변수들 중 근육 내 지방 함량은 한우의 맛, 즉 부드러움과 육즙과 직접 관련된다(Wheeler et al, 1994). 이러한 맛은 2006년도 10군데 서로 다른 부위에서 잘라낸 한우로 평가한다. 한국 요리에서 가장 인기 있는 한우조리법은 한우구이인데 각 부위에서의 지방함량의 비율은 한우의 맛과 부드러움, 육즙의 관계를 분석하기 위하여 측정된다(Jennings et al., 1978; Indurain et al., 2009). 한우 맛의 평가 점수는 0부터 100까지이다. 그림 4.3은 육즙과 부드러움 두 변수가 각각 지방 함량이 높아짐에 따라 얻은 그래프이다. 또한 그림 4.4(b)에서는 육즙과 부드러움의 관계를 나타내는 그래프이다. 이는 어느 정도의 지방함량의 비율이 한우의 맛 평가를 변화시키는가를 결정하는데 관심이 있다. 본 논문에서 제안한 베이지안 다중 변환점 모형을 이용하여 이변량 자료의 변환점을 찾으려 한다. 이 자료는 Kim and Cheon (2010)의 한우의 지방함량 자료이다.

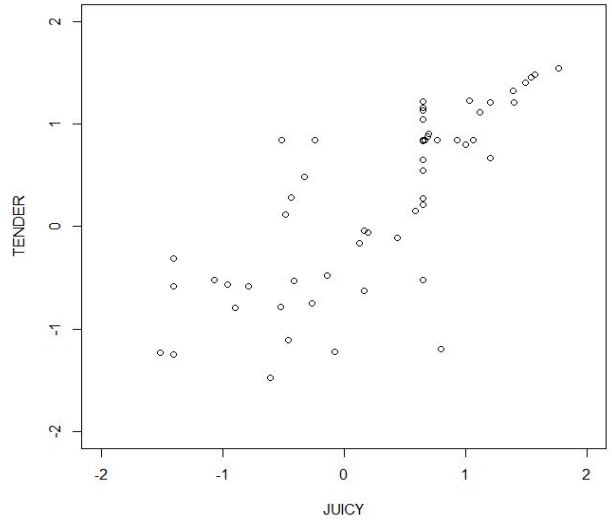



그림 4.3: 한우의 지방함량자료의 그래프

### 4.4.2 실증 분석 결과

<표 4.2>는 모의실험과 마찬가지로 메트로폴리스-해스팅스를 포함한 깁스 샘플링 알고리즘을 이용하여 변환점이 1개, 2개, 3개인 경우에 각각 10000번씩 시행하여 로그

표 4.2: 실증 분석 결과

| No | CP=1 | | |
|---|---|---|---|
| | Change patterns | Log-posterior | BIC |
| 1 | (27) | -14.7105 | **53.9871** |
| 2 | (27) | -14.7327 | 54.0315 |
| 3 | (27) | -14.9964 | 54.5589 |
| 4 | (27) | -15.0501 | 54.6662 |
| 5 | (27) | -15.0657 | 54.6974 |
| 6 | (22) | -15.2512 | 55.0684 |
| 7 | (22) | -15.2623 | 55.0907 |
| 8 | (22) | -15.3023 | 55.1707 |
| 9 | (27) | -15.4431 | 55.4522 |
| 10 | (43) | -15.4556 | 55.4774 |
| No | CP=2 | | |
| | Change patterns | Log-posterior | BIC |
| 1 | **(43,55)** | **-14.4847** | 65.8185 |
| 2 | (43,55) | -14.5022 | 65.8534 |
| 3 | (43,55) | -14.9717 | 66.7925 |
| 4 | (43,55) | -15.2261 | 67.3012 |
| 5 | (43,55) | -15.2482 | 67.3454 |
| 6 | (43,55) | -15.3916 | 67.6323 |
| 7 | (43,55) | -15.4110 | 67.6711 |
| 8 | (42,55) | -15.4647 | 67.7785 |
| 9 | (42,55) | -15.5515 | 67.9521 |
| 10 | (43,55) | -15.5825 | 68.0142 |
| No | CP=3 | | |
| | Change patterns | Log-posterior | BIC |
| 1 | (27,43,55) | -21.3942 | 91.9204 |
| 2 | (27,43,55) | -21.4704 | 92.0729 |
| 3 | (21,43,55) | -21.4708 | 92.0736 |
| 4 | (22,43,55) | -21.5082 | 92.1485 |
| 5 | (27,43,55) | -21.5674 | 92.2669 |
| 6 | (19,22,55) | -21.7648 | 92.6617 |
| 7 | (28,43,55) | -21.8883 | 92.9088 |
| 8 | (24,43,55) | -21.9040 | 92.9401 |
| 9 | (10,43,55) | -21.9071 | 92.9463 |
| 10 | (20,43,55) | -21.9218 | 92.9758 |

사후확률이 가장 큰 10개 값을 찾아서 그 값에 해당되는 변환점 위치와 로그 사후확률(log-posterior) 및 BIC를 구한 결과를 보여준다. 변환점이 2개이며 변환점 위치 $(c_1, c_2)$

가 (43. 55)일 때 비록 BIC 값은 비록 제일 작지 않지만 여기서 로그 사후확률값이 -14.4847로 가장 크기 때문에 변환점 위치 (43, 55)인 변환점 2개의 모형이 최적의 모형으로 선택되었다.

<그림 4(a)>에서 변환점이 2개인 경우 변환점 위치의 상대빈도 히스토그램을 보여주며, 또한 <그림 4(b)>에서 로그 사후확률이 가장 크고 BIC가 가장 작을 때 최적의 모형인 변환점 위치 (43, 55)를 보여주고 있다.

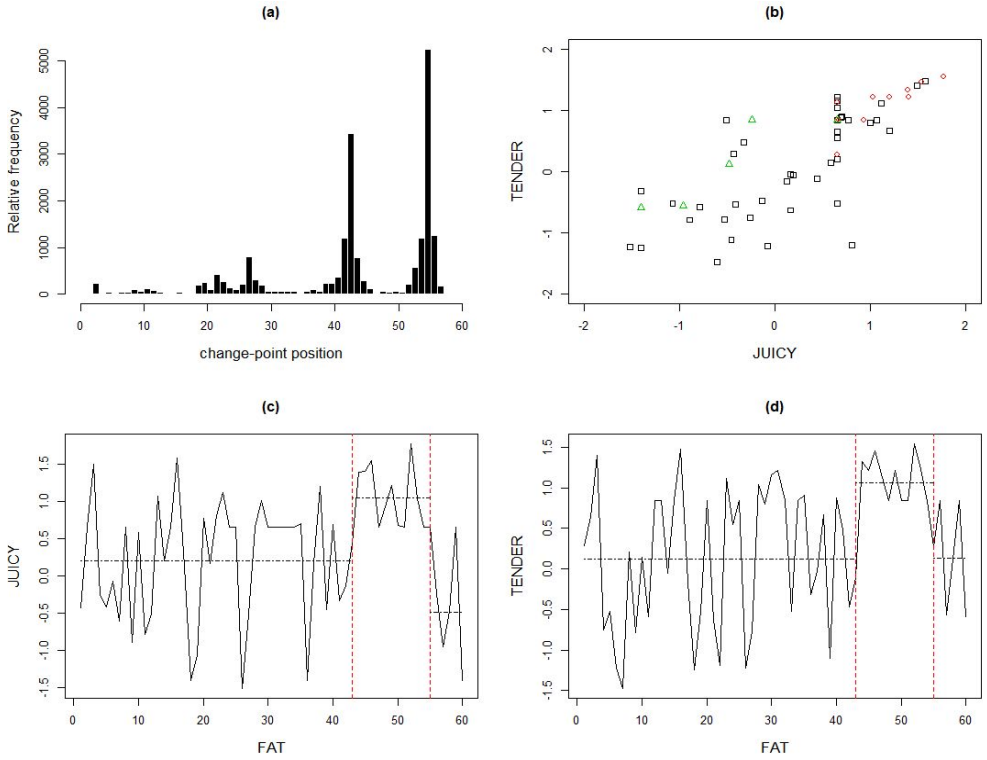



그림 4.4: (a): 변환점위치 상대빈도; (b): 두 변수의 해당 위치 그래프; (c): 육즙의 변환점위치 및 부분영역의 평균 값; (d): 부드러움의 변환점위치 및 부분영역의 평균 값

이 결과는 Kim and Cheon (2010)의 결과와 비슷하게 나온다. Kim and Cheon (2010)은 한우의 지방함량 자료로 272개 자료를 선택하였다. 하지만 본 논문은 변화가 보다 심한 소량자료에 관심이 있어서 272개 자료 중 60개 자료(시점,191-250)만 사용하여 분석한 결과 변환점 위치(43, 55)를 찾았고 이 위치는 Kim and Cheon에서 찾아낸 (233, 245)시점을 정확하게 일치함을 알 수 있었다. Kim and Cheon (2010)은 베이지안 다변량 정규분포를 이용하였고 본 논문에서는 소량자료를 위해 다변량 비중심 t 분포

를 사용하였다.

## 제 5장 결 론

본 논문에서 소량자료의 변환점 개수 및 변환점의 위치를 찾고자 할 때 베이지안 다변량 비중심 t 분포 모형을 제안하였고 Müller(1991, 1993)가 제안한 메트로폴리스-해스팅스를 포함한 깁스 샘플링 알고리즘을 이용하여 분석을 하였다. 제안된 모형을 모의실험 및 태풍횟수자료와 한우 지방함량자료 등의 실증 분석에 이용한 결과 변환점 개수와 그 위치를 정확히 찾아주고 있다. 결론적으로 소량자료에 관한 변환점 및 위치를 찾고자 할 때 본 논문에서 제안한 베이지안 다변량 비중심 t 분포 모형을 이용하는 것이 좋다고 본다. 또한 결합분포에서 표본을 추출하기 힘들 때 메트로폴리스-해스팅스를 포함한 깁스 샘플링 알고리즘을 제안하고자 한다.

# 참고문헌


[1] 설동일 (2010). 지구 온난화와 태풍의 변화 경향. *한국항해항만학회 학술발표*, 2010(1), 238-239.

[2] Barry, D. and Hartigan, J.A. (1993). A Bayesian analysis for change-point problems. *Journal of the American Statistical Association*, 88, 309-319.

[3] Carlin B.P., Gelfand A.E. and Smith A.F.M. (1992). Hierarchical Bayesian analysis of change point problem. *Applied statistics*, 41(2), 389-405.

[4] Chen, M.H. and Schmeiser, B.W. (1998). Towards black-box sampling. *Journal of computational and Graphical Satistics*, 7(1), 1-22.

[5] Chib, S. (1998). Estimation and comparison of multiple change-point models. *Journal of Econometrics*, 86, 221-241.

[6] Forrest, R.J. (1975). Effects of castration, sire and hormone treatments on the quality of rib roasts from Holstein-Friesian males. *Canadian Journal of Animal Science*, 55-87.

[7] Geman, S. and Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and Bayesian Restoration of Images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 6, 721-741.

[8] Hastings, W.K. (1970). Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, 57, 97-109.

[9] Hinkley, D.V. (1970). Inference about the change-point in a sequence of random variables. *Biometrika*, 57, 1-17.

[10] Indurain, G., Carr, T.R., Gonim M.V., Insausti, K. and Beriain, M.J. (2009). The relationship of carcass measurements to carcass composition and intramuscular fat in Spanish beef. *Meat Science*, 82, 155-161.

[11] Jennings, T.G., Berry, B.W. and Joseph, A.L. (1978). Influence of fat thickness, marbling and length of aging on beef palatability and shelf-life characteristics. *Journal of Animal Science*, 46(3), 658-665.

[12] Kim, J. and Cheon, S. (2010). Bayesian multiple change-point estimation with annealing stochastic approximation Monte Carlo. *Computational statistics*, 25(2), 215-239.

[13] Kim, J. and Cheon, S. (2010). Multiple change-point detection of multivariate mean vectors with the Bayesian approach. *Computational Statistics and Data*

*Analysis*, 54, 406-415.

[14] Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E. (1953). Equations of State Calculations by Fast Computing Machines. *Journal of Chemical Physics,* 21, 1087-1091.

[15] Müller, P. (1991). A generic approach to posterior integration and Gibbs sampling. Technical Report, Purdue University, West Lafayette IN.

[16] Müller, P. (1993). Alternatives to the Gibbs sampling scheme,Technical, Report. *Institute of Statistics and Decision Sciences*, Duke University.

[17] Smith, A.F.M. (1975). A Bayesian approach to inference about a change-point in a sequence of random variables. *Biometrika*, 62(2), 407-416.

[18] SUGI, M., NODA, A. and SATO, N. (2002). Influence of the Global Warning on Tropical Cyclone Climatology: An Experiment with the JMA Global Model. *Journal of the Meteorological Society of Japan*, 80(2), 249-272.

[19] Venter, J.H. and Steel, S.J. (1996). Finding multiple abrupt change points. *Computational Statistics and Data Analysis*, 22, 481-50 4.

[20] Wheeler, T.L., Cundiff, L. V. and Koch, R.M. (1994). Effect of marbling degree on beef palatability in Bos Taurus and Bos Inbdicus cattle. *Journal of Animal Science*, 72, 3145-3151.