The Korean Journal of Applied Statistics (2012) DOI: http://dx.doi.org/10.5351/KJAS.2012.25.6.999
25(6), 999–1008

Bayesian Multiple Change-Point Estimation of Multivariate Mean Vectors for Small Data

Sooyoung Cheon[1] · Wenxing Yu[2]

[1]Department of Informational Statistics, Korea University
[2]Department of Economics and Statistics, Korea University

(Received July 31, 2012; Revised August 27, 2012; Accepted October 23, 2012)

Abstract

A Bayesian multiple change-point model for small data is proposed for multivariate means and is an extension of the univariate case of Cheon and Yu (2012). The proposed model requires data from a multivariate noncentral t-distribution and conjugate priors for the distributional parameters. We apply the Metropolis-Hastings-within-Gibbs Sampling algorithm to the proposed model to detecte multiple change-points. The performance of our proposed algorithm has been investigated on simulated and real dataset, Hanwoo fat content bivariate data.

Keywords: Small data, change-point, noncentral t-distribution, Metropolis-Hastings-Within-Gibbs sampling, Hanwoo fat content.

1. 서론

경제학, 금융학, 기상학, 금융학 등 여러 분야에서 변환점에 대한 관심이 증가하고 있다. 하지만, 일반적으로 변환점의 수와 위치는 알려져 있지 않기때문에, 변환점을 찾는 문제가 통계학에서 도전해 볼만한 분야이다. 본 연구에서는 축산학에서의 변환점분석을 위해 한우자료를 다룬다. 일반적으로 한우의 지방함량따라 한우의 맛이 달라지는데, 이러한 한우의 맛을 평가하는데 있어 지방함량에 따른 부드러움(Tenderness)과 육즙(Juiciness)이 주로 사용된다. 한우는 이제 더 이상 우리만의 것이 아니라 세계가 주목하고 있는 글로벌 한우가 되었다. 한우가 해외시장에서 당당하게 대접받기 위해서는 한우의 맛을 개선시킬 필요가 있다. 따라서 한우의 맛에 중요한 역할을 하는 지방함량 비율에 대한 변환점을 찾는 것이 중요한 과제이다.

일반적으로 자료가 정규분포를 따른다는 가정하에, 하나의 변환점인 경우, Chernoff와 Zacks (1964)와 Smith (1975)가 변환점 추론을 위해 베이지안 방법을 제안하였고, Carlin 등 (1992)은 마코브 연쇄 몬테카를로(Markov chain Monte Carlo; MCMC; Metropolis 등, 1953; Hastings, 1970) 알고리즘을 이용하여 Smith (1975)의 방법을 확장하였다. 다중 변환점 모형인 경우에, Barry와 Hartigan (1993)은 정규분포의 곱분할 모형(product partition model)에서 모수들의 평균 변화를 통해 변환점을 찾았으며, 최근에 Kim과 Cheon (2010)과 Cheon과 Kim (2010)은 stochastic approximation Monte Carlo 알

[1]Corresponding author: Assistant Professor, Department of Informational Statistics, Korea University, 2511 Sejong-ro, Sejong-city 339-700, South Korea. E-mail: scheon@korea.ac.kr

고리즘을 이용하여 일변량과 이변량의 정규분포에서 베이지안 다중변환점 모형을 제안하였다. 이와 같이 지금까지 제안된 대부분의 모형은 대량자료 분석을 위해 정규 분포 자료의 베이지안 다중 변환점 모형에 MCMC를 이용하여 해결하였다. 하지만 소량자료 분석을 위한 베이지안 다중 변환점 모형에 대한 연구는 아직까지 미진하다. 일반적으로 자료가 소량일때, 정규분포를 가정한 통상적인 통계적 방법으로는 정확한 추론을 할 수 없다. 보다 정확한 추론을 위해 소량자료에 대한 연구에 있어, 최근에 Cheon과 Yu (2012)가 단변량 자료의 베이지안 비중심 t 분포 다중 변환점 모형을 제안하였다. 이에 대한 확장으로 본 논문에서는 다변량 자료의 베이지안 비중심 t 분포 모형을 제안하여 보다 복잡한 모형에서의 변환점 분석을 하고자 한다.

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(Metropolis-Hastings-within-Gibbs Sampling; MHWGS) 알고리즘을 이용하여 소량자료에서 변환점 분석을 하고자 한다. 본 논문에서 다루는 실증 분석 자료는 2006년도 10군데 서로 다른 부위에서 잘라낸 한우의 지방함량에 따른 부드러움(Tenderness)과 육즙(Juiciness)의 이변량 자료이다.

본 논문의 2장에서 베이지안 다변량 비중심 t 분포 다중 변환점 모형 및 MHWGS를 이용한 베이지안 모형 선택에 대하여 소개하고, 3장에서는 모의실험 자료의 분석 및 그 결과를 보여준다. 4장에서는 한우의 이변량 자료의 실증 분석을 하며, 5장에서는 본 논문의 결론을 정리한다.

2. 베이지안 다중 변환점 분석

2.1. 베이지안 다변량 비중심 t 분포 다중 변환점 모형

본 논문의 베이지안 다중변환점 분석은 Cheon과 Yu (2012)에서 이용한 모형과 방법의 확장을 통해 분석한다. $d \times 1$ 확률 벡터 $\vec{X}$가 다변량 비중심 t 분포를 따른다고 할때, 변환점이 없는 다변량 $\vec{X}$의 확률분포(probability density function)는 다음과 같다.

$$f\left(\vec{x}|p,\vec{u},\Sigma\right) \propto |\Sigma|^{-\frac{1}{2}}\left[1+\frac{(\vec{x}-\vec{u})'\Sigma^{-1}(\vec{x}-\vec{u})}{p}\right]^{-\frac{p+d}{2}}. \tag{2.1}$$

여기서 p는 자유도, d는 차원수, $\vec{u}$는 $d \times 1$ 비중심 모수벡터이고 Σ는 $d \times d$ 분산-공분산 행렬이다.

전체영역 $S = (1, 2, \ldots, n)$가 서로 다른 분포를 가지는 여러 부분영역으로 나누어 지고, 각각의 부분영역들은 변환점에 의해 분할된다고 하자. 임의의 l $(\in [1, n])$에 대해, $T = (t_1, t_2, \ldots, t_n)'$는 $t_{c_1} = t_{c_2} = \cdots = t_{c_l} = 1$, 그 외는 0인 이항 벡터라고 하고, $0 = c_0 < c_1 < c_2 < \cdots < c_{l+1} = n$이라 하자. 또한, $i = 1, 2, \ldots, l+1$에 대해 $C_i = \{m : c_{i-1} < m \leq c_i\}$이라 하자. 만약 변환점이 k개라면 다중 변환점 모형을 다음과 같이 정의한다.

$$f(\vec{x}) = \prod_{j=1}^{n} f(\vec{x}_j) = \prod_{j \in C_1} f_1(\vec{x}_j) \times \cdots \times \prod_{j \in C_{k+1}} f_{k+1}(\vec{x}_j). \tag{2.2}$$

여기서 $c_{i-1} < j \leq c_i$ $(i = 1, 2, \ldots, k+1)$에 대해 $\vec{x}_j \sim f_i(\cdot|\phi_i)$이다. 이때 f_i는 모수 $\phi_i = (\vec{u}_i, \Sigma)$ $(\in \Phi = (\vec{u}, \Sigma))$에 의존한다. 각 자료들은 $c_1 + 1, c_2 + 1, \ldots, c_k + 1$에서 변화하므로 전체 공간을 $k+1$개의 부분공간으로 분할하는 $c_1, c_2, \ldots, c_k$를 변환점(change-points)이라고 부른다.

만약 함수 f_i가 d차원 다변량 비중심 t 분포라고 하면, 모수 ϕ_i는 d차원 자유도 p_i, 위치벡터 $\vec{u}_i$와 분산행렬 Σ로 나뉜다. 따라서 $T^{(k)}$를 k개의 변환점을 가지는 T라고 할때 전체 모수는 $\eta^{(k)} = (T^{(k)}, p_1, \vec{u}_1, \ldots, p_{k+1}, \vec{u}_{k+1}, \Sigma)$로 표현된다. $\vec{\phi} = (p, \vec{u}, \Sigma)$라 하면, 확률 벡터 $\vec{X}$는 비중심 모수가

$\vec{u}$이고 독립인 다변량 t 분포를 따른다고 한다. k개의 변환점이 있을 때, 자유도 p, $d \times 1$인 벡터 $\vec{u}$와 $d \times d$인 분산공분산 행렬 Σ를 가지는 다변량 $\vec{X}$의 확률분포(probability density function)는 다음과 같다.

$$f(\vec{x}|p,\vec{u},\Sigma) = \prod_{j\in C_1} f_1\left(\vec{x}_j|p_1,\vec{u}_1,\Sigma\right) \times \cdots \times \prod_{j\in C_{k+1}} f_{k+1}\left(\vec{x}_j|p_{k+1},\vec{u}_{k+1},\Sigma\right),$$

$$\text{where} \qquad f_i\left(\vec{x}_j|p_i,\vec{u}_i,\Sigma\right) \propto |\Sigma|^{-\frac{1}{2}}\left[1+\frac{(\vec{x}_j-\vec{u}_i)'\Sigma^{-1}(\vec{x}_j-\vec{u}_i)}{p_i}\right]^{-\frac{p_i+d}{2}}, \quad i=1,2,\ldots,k+1. \tag{2.3}$$

베이지안 추론을 위해 $\vec{u}_i\,(i=1,2,\ldots,k+1)$의 사전분포는 균일분포를, Σ의 사전분포는 Inverse-Wishart분포 IW(v_0,Λ_0^{-1})를 따른다고 가정한다. 각 분포의 자유도는 $p_i = n_i - 1 = c_i - c_{i-1} - 1$이다. 따라서, 자유도 p_i대신 변환점 위치 c_i를 이용하면 자유도 p_i를 구할 수 있다. 본 연구는 c_i의 사전확률분포로 이산균일분포를 선택하였다.

변환점이 k개 있을때, k개의 변환점 위치 $(c_1,c_2,\ldots,c_k)$와 $k+1$개의 비중심 모수 벡터 $(\vec{u}_1,\vec{u}_2,\ldots,\vec{u}_{k+1})$가 주어진다. $\eta^{(k)} = (T^{(k)},p_1,\vec{u}_1,\ldots,p_{k+1},\vec{u}_{k+1},\Sigma)$의 우도함수는 다음과 같다.

$$L\left(\eta^{(k)}|X\right) = \prod_{i=1}^{k+1}\prod_{j=c_{i-1}+1}^{c_i} f_i\left(\vec{x}_j|p_i,\vec{u}_i,\Sigma\right). \tag{2.4}$$

$\eta^{(k)}$의 사전확률분포로 다음과 같이 가정한다.

$$\begin{aligned}\pi\left(\eta^{(k)}\right) &\propto \pi(\Sigma)\times\pi\left(\vec{u}_1,\ldots,\vec{u}_{k+1}\right)\times\pi\left(\vec{c}_1,\ldots,\vec{c}_k\right)\\ &= |\Lambda_0|^{\frac{v_0}{2}}\times|\Sigma|^{-\frac{v_0+d+1}{2}}\times\exp\left[-\frac{1}{2}\mathrm{tr}\left(\Lambda_0\Sigma^{-1}\right)\right]\\ &\quad\times\frac{1}{(b-a)^{2(k+1)}}\times I_{(a,b)}(\vec{u}_1,\ldots,\vec{u}_{k+1})\times\frac{1}{(n-5)^k}\times I_{[3,\ldots,n-3]}(c_1,\ldots,c_k),\end{aligned} \tag{2.5}$$

여기서 $I(\cdot)$는 지시함수이다. 따라서 $\eta^{(k)}$의 로그사후(log-posterior) 확률분포인 베이지안 다변량 t 분포 다중 변환점모형은 다음과 같다.

$$\begin{aligned}\log\left(\pi\left(\eta^{(k)}|X\right)\right) &\propto \log\left(L\left(\eta^{(k)}|X\right)\pi\left(\eta^{(k)}\right)\right)\\ &= \sum_{i=1}^{k}\sum_{j=c_{i-1}+1}^{c_i}\left\{-\frac{1}{2}\log|\Sigma| - \frac{p_i+d}{2}\log\left[1+\frac{(\vec{x}_j-\vec{u}_i)'\Sigma^{-1}(\vec{x}_j-\vec{u}_i)}{p_i}\right]\right\}\\ &\quad+\frac{v_0}{2}\log|\Lambda_0| - \frac{v_0+d+1}{2}\log|\Sigma| - \frac{1}{2}\mathrm{tr}(\Lambda_0\Sigma^{-1}) - 2(k+1)\log(b-a)\\ &\quad+\log\left(I_{(a,b)}\left(\vec{u}_1,\ldots,\vec{u}_{k+1}\right)\right) - k\log(n-5) + \log\left(I_{[3,\ldots,n-3]}(c_1,\ldots,c_k)\right).\end{aligned} \tag{2.6}$$

본 논문에서는 $d=2$인 경우인 이변량 비중심 t 분포에 관하여 분석하였다. 따라서, $\eta^{(k)}$의 로그사후확률분포인 베이지안 이변량 t 분포 다중 변환점모형은 다음과 같이 구할 수 있다.

$$\begin{aligned}\log\left(\pi\left(\eta^{(k)}|X\right)\right) &\propto \sum_{i=1}^{k}\sum_{j=c_{i-1}+1}^{c_i}\left\{-\frac{1}{2}\log|\Sigma| - \frac{p_i+2}{2}\log\left[1+\frac{(\vec{x}_j-\vec{u}_i)'\Sigma^{-1}(\vec{x}_j-\vec{u}_i)}{p_i}\right]\right\}\\ &\quad+\frac{v_0}{2}\log|\Lambda_0| - \frac{v_0+3}{2}\log|\Sigma| - \frac{1}{2}\mathrm{tr}\left(\Lambda_0\Sigma^{-1}\right) - 2(k+1)\log(b-a)\\ &\quad+\log\left(I_{(a,b)}\left(\vec{u}_1,\ldots,\vec{u}_{k+1}\right)\right) - k\log(n-5) + \log\left(I_{[3,\ldots,n-3]}(c_1,\ldots,c_k)\right).\end{aligned} \tag{2.7}$$

2.2. 메트로폴리스-해스팅스를 포함한 깁스 샘플링과 베이지안 모형 선택

결합분포에서 자료를 직접 얻기 힘든 경우가 발생했을 때, 일반적으로 채택-기각(acceptance-rejection) 알고리즘과 같은 맞춤형(customized) 알고리즘을 이용하여 해결하려고 노력한다. 본 논문은 Müller (1991, 1993)가 제안한 절충(compromised) 깁스 알고리즘, 즉 메트로폴리스-해스팅스를 포함한 깁스 샘플링(Metropolis-Hastings-within-Gibbs Sampling; MHWGS) 알고리즘을 이용한다. MHWGS 알고리즘에서 $f_k(x_k|x_i, i \neq k)$로 부터 표본을 추출하는데 어려움이 있는 깁스 샘플러(Gibbs sampler; Geman과 Geman, 1984)의 샘플링 단계가 Metropolis Hastings(MH; Hastings, 1970) 단계로 대체된다.

이변량 베이지안 모형에서, 비중심 모수는 $(\vec{u}_1, \ldots, \vec{u}_{k+1})$, 변환점 위치는 $(c_1, \ldots, c_k)$, 분산 모수는 Σ인 k개의 변환점이 주어진 경우를 고려해 보자. MHWGS 알고리즘의 자료생성 절차는 다음과 같다. 먼저 모든 변수들을 고려한 변수 $\boldsymbol{z} = (z_1, \ldots, z_{k+1}, z_{k+2}, \ \ldots, z_{2k+1}, z_{2k+2}) = (\vec{u}_1, \ldots, \vec{u}_{k+1}, c_1, \ldots, c_k, \Sigma)$를 설정한다.

메트로폴리스-해스팅스를 포함한 깁스 샘플링:

$i = 1, 2, \ldots, 2k+2$이고 현재 표본이 t번째 반복 후인 $\boldsymbol{z} = \big(z_1^{(t+1)}, \ldots, z_{i-1}^{(t+1)}, z_i^{(t)}, \ldots, z_{2k+2}^{(t)}\big)$라고 하자.

1. $q_i\big(z_i|z_1^{(t+1)}, \ldots, z_{i-1}^{(t+1)}, z_i^{(t)}, \ldots, z_{2k+2}^{(t)}\big)$로 부터 z_i^*를 생성한다.
2. $r = \dfrac{\pi_i\big(z_i^*|z_1^{(t+1)}, \ldots, z_{i-1}^{(t+1)}, z_{i+1}^{(t)}, \ldots, z_{2k+2}^{(t)}\big)}{\pi_i\big(z_i^{(t)}|z_1^{(t+1)}, \ldots, z_{i-1}^{(t+1)}, z_{i+1}^{(t)}, \ldots, z_{2k+2}^{(t)}\big)} \times \dfrac{q_i\big(z_i^{(t)}|z_1^{(t+1)}, \ldots, z_{i-1}^{(t+1)}, z_i^*, z_{i+1}^{(t)}, \ldots, z_{2k+2}^{(t)}\big)}{q_i\big(z_i^*|z_1^{(t+1)}, \ldots, z_{i-1}^{(t+1)}, z_i^{(t)}, z_{i+1}^{(t)}, \ldots, z_{2k+2}^{(t)}\big)}$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.

여기서 MH 절차는 각각의 반복에서 오직 한 번만 시행한다. 만약 MH 절차를 각각의 반복에서 여러 번 시행했을 경우 $\pi_i(\cdot)$의 더 정확한 근사치를 얻을 수 있다. 하지만 오직 한 번의 절차만 시행해도 비교적 좋은 결과를 얻을 수 있기 때문에 효율성을 위해 본 논문에서는 오직 한번의 MH 절차만 시행한다 (Chen과 Schmeiser, 1998).

본 논문에서는 MHWGS 알고리즘을 이용하여 얻어진 표본으로 부터 계산된 사후 확률과 Bayesian Information Criterion(BIC)를 동시에 고려하여 최적의 모형인 최대 사후 확률(maximum a posteriori; MAP) 모형을 선택한다 (Cheon과 Yu, 2012).

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

3. 모의실험

3.1. 모의실험 자료

본 연구에서 제안한 모형의 모의실험을 위해 각각 20개씩 3개 부분으로 나뉜 60개의 이변량 자료들로 이루어진 서로 다른 4개 세트를 차례로 생성시켰다 (Figure 3.1). 즉, $\vec{x}_1, \ldots, \vec{x}_{20} \sim \text{BVT}(p_1 = 19, \vec{u}_1 = (\vec{u}_{11}, \vec{u}_{12}, \vec{u}_{13}, \vec{u}_{14}) = ((3.53, 2.71), (3.53, 10.62), (4.93, 3.60), (3.72, 8.03))$; $\vec{x}_{21}, \ldots, \vec{x}_{40} \sim \text{BVT}(p_2 = 19, \vec{u}_2 = (\vec{u}_{21}, \vec{u}_{22}, \vec{u}_{23}, \vec{u}_{24}) = ((6.29, 6.78), (6.29, 6.78), (7.63, 8.22), (7.52, 4.80))$; $\vec{x}_{41}, \ldots, \vec{x}_{60} \sim \text{BVT}(p_3 = 19, \vec{u}_3 = (\vec{u}_{31}, \vec{u}_{32}, \vec{u}_{33}, \vec{u}_{34}) = ((8.54, 10.62), (8.54, 2.71), (9.99, 6.31), (9.76, 7.56))$. 모의실험을 위한 통계 패키지로는 R-software(version 2.12.1)를 사용하였다.

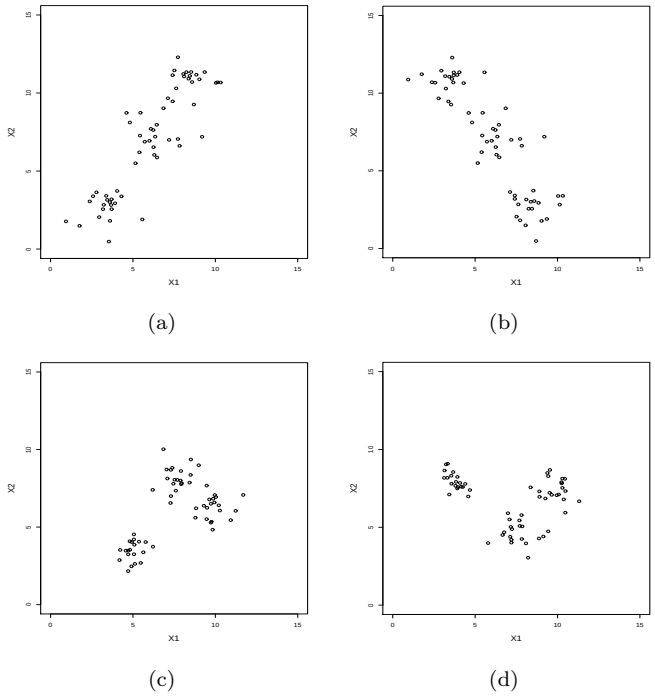

Figure 3.1. The plots of four simulated data from the bivariate t distribution.

3.2. 모의실험 결과

본 모의실험은 변환점이 각각 1개, 2개, 3개인 경우만을 고려하였다. Table 3.1은 MHWGS 알고리즘을 이용하여 변환점이 1개, 2개, 3개인 경우와 4개의 모의실험 자료에 대해 각각 10000번씩 시행하여 로그사후확률이 가장 큰 5개 값으로부터 얻은 결과이다. Table 3.1로부터 모든 경우에 있어 변환점 2개에서의 로그사후확률이 다른 변환점에서의 로그사후확률보다 크고 BIC값도 가장 작은 것을 볼 수 있다. 즉, 변환점 위치가 $(20, 40)$일 때 로그사후확률값은 각각 $(-69.8243, -71.0516, -31.6007, -30.7672)$로 가장 크고 BIC값도 각각 $(176.4977, 178.9524, 100.0506, 98.3835)$으로 가장 작게 얻어졌다. 따라서, 사후확률과 BIC에 의해 MAP 모형은 변환점 위치 $(20, 40)$일 때이다.

Figure 3.2로부터 각 모형의 (a)는 변환점 위치의 상대빈도 히스토그램으로 4개의 모의실험 자료의 MAP 모형이 모두 (20, 40)임을 알 수 있고, 또한 (b)에 의해 MAP 추정치에서 정확히 세 개의 서로 다른 이질적인 영역으로 분리되고 있음을 알 수 있다. (c)와 (d)는 각 변수에서 변환점의 MAP 추정치인 (20, 40)에 의해 분리된 부분영역들을 보여준다. 따라서 제안된 베이지안 모형에 MHWGS 알고리즘을 이용하면 정확히 변환점 개수 및 위치를 찾아 낼 수 있음을 알 수 있다.

4. 실증분석

4.1. 실증분석 자료

한우의 맛은 일반적으로 부드러움(Tenderness)과 육즙(Juiciness)으로 맛을 평가한다 (Forrest, 1975). 맛을 결정하는 다양한 변수들 중 근육 내 지방 함량이 한우의 맛을 내는 부드러움과 육즙과 직접 관련되

Table 3.1. Three five models with the largest log-posterior values in four simulated data set according to the number of change points.

CP	No	(a) Change patterns	(a) Log posterior	(a) BIC	(b) Change patterns	(b) Log posterior	(b) BIC
1	1	(20)	−113.7180	252.0021	(20)	−117.8768	260.3197
	2	(20)	−113.7204	252.0069	(20)	−118.1020	260.7701
	3	(20)	−113.7917	252.1495	(20)	−118.3465	261.2591
	4	(20)	−113.8655	252.2971	(20)	−118.3475	261.2611
	5	(20)	−113.9012	252.3685	(20)	−118.4261	261.4183
2	1	(20, 40)	−69.8243	176.4977	(20, 40)	−71.0516	178.9524
	2	(20, 40)	−69.9469	176.7428	(20, 40)	−71.0703	178.9897
	3	(20, 40)	−69.9916	176.8324	(20, 40)	−71.0861	179.0213
	4	(20, 40)	−69.9977	176.8444	(20, 40)	−71.2407	179.3306
	5	(20, 40)	−70.4274	177.7039	(20, 40)	−71.3068	179.4626
3	1	(7, 20, 40)	−79.0371	207.2064	(20, 35, 40)	−74.9210	198.9742
	2	(17, 20, 40)	−79.4618	208.0558	(20, 35, 40)	−75.1090	199.3501
	3	(11, 20, 40)	−79.4962	208.1245	(20, 35, 40)	−75.1779	199.4879
	4	(12, 20, 40)	−79.5388	208.2098	(20, 35, 40)	−75.2199	199.5719
	5	(7, 20, 40)	−79.6206	208.3733	(20, 35, 40)	−75.2578	199.6477

CP	No	(c) Change patterns	(c) Log posterior	(c) BIC	(d) Change patterns	(d) Log posterior	(d) BIC
1	1	(20)	−74.9694	174.5049	(20)	−86.4568	197.4796
	2	(20)	−75.0331	174.6323	(20)	−86.4952	197.5565
	3	(20)	−75.4836	175.5333	(20)	−86.7231	198.0123
	4	(20)	−75.4953	175.5567	(20)	−86.8292	198.2244
	5	(20)	−75.5601	175.6862	(20)	−86.8919	198.3498
2	1	(20, 40)	−31.6007	100.0506	(20, 40)	−30.7672	98.3835
	2	(20, 40)	−31.7505	100.3501	(20, 40)	−30.7991	98.4473
	3	(20, 40)	−31.7703	100.3897	(20, 40)	−30.8469	98.5428
	4	(20, 40)	−31.8245	100.4981	(20, 40)	−31.0136	98.8762
	5	(20, 40)	−31.9387	100.7265	(20, 40)	−31.1008	99.0507
3	1	(16, 20, 40)	−42.0882	133.3085	(11, 20, 40)	−38.7646	126.6614
	2	(6, 20, 40)	−42.1872	133.5066	(15, 20, 40)	−39.2003	127.5327
	3	(6, 20, 40)	−42.2566	133.6454	(11, 20, 40)	−39.2577	127.6475
	4	(5, 20, 40)	−42.3444	133.8208	(16, 20, 40)	−39.3930	127.9181
	5	(16, 20, 40)	−42.4502	134.0325	(17, 20, 40)	−39.4890	128.1101

어 있다고 알려져 있다 (Wheeler 등, 1994). 본 논문은 Cheon과 Kim (2010)에서 사용한 한우자료중 일부를 이용하였으며, 이 자료는 한우의 맛에 영향을 미치는 부드러움과 육즙을 평가하기 위해 각 부위에서의 지방함량 비율이 측정된 자료이다 (Jennings 등, 1978; Indurain 등, 2009). 한우 맛의 평가 점수는 0부터 100까지이다. Figure 4.1은 육즙과 부드러움 두 변수가 각각 지방 함량이 높아짐에 따라 얻어진 그림이다. 본 실증 분석에서는 어느 정도의 지방함량 비율이 한우의 맛의 평가를 변화시키는가를 알기 위해, 본 논문에서 제안한 베이지안 다중 변환점 모형을 이용하여 이변량 자료의 변환점을 찾으려 한다.

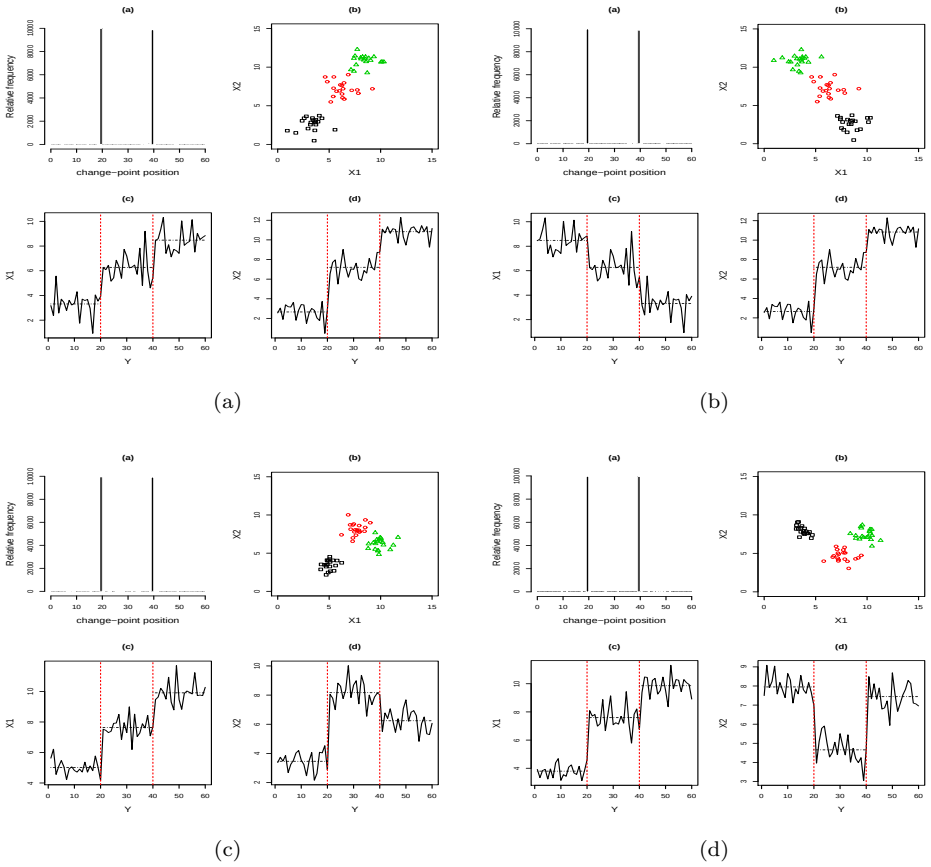

Figure 3.2. The simulation results of four multiple change-point examples: (a) The histogram for log posterior probabilities of change-point positions from the posterior samples; (b) A bivariate plot with the maximum log-posterior estimate of the change-point positions; (c) A comparison of the MAP estimates of the change-point positions and the true change-point positions in observations of the first variable. The vertical (dotted) lines indicated the change-point positions identified by the MAP model, and the horizontal (dashed) lines indicate the mean value of observations separated by change-point positions of the true model; (d) same with (c) except for the second variable)

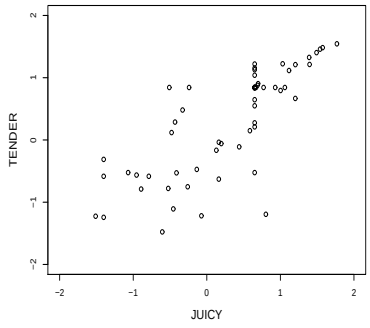

Figure 4.1. A plot of juiciness and tenderness

Table 4.1. Three ten models with the largest log-posterior values in fat data according to the number of change points.

CP	No	Change patterns	Log-posterior	BIC
1	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
2	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
3	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

4.2. 실증분석 결과

Table 4.1은 MHWGS 알고리즘을 이용하여 변환점이 1개, 2개, 3개인 경우에 각각 10000번씩 시행하여 로그사후확률이 가장 큰 10개 값을 찾아서 그 값에 해당되는 변환점 위치와 로그사후확률및 BIC를 구한 결과를 보여주고 있다. Figure 4.2(a)는 변환점 위치의 상대빈도 히스토그램으로 변환점 27보다 변환점 43과 55에서 훨씬 많은 빈도수를 보여 주고 있다. 따라서, 변환점이 2개이며 변환점 위치 (43, 55)일 때, 비록 BIC 값은 비록 제일 작지 않지만 빈도수가 가장 많고 또한 최대 로그사후확률값이 −14.4847로 가장 크기 때문에 변환점 위치 (43, 55)인 변환점 2개의 모형이 MAP 모형으로 선택되었다. 또한, Figure 4.2(b)로 부터 최적의 모형인 변환점 위치 (43, 55)에서 어느 정도 세 개의 서로 다른 영역으로 분리하고 있음을 알 수 있다. Figure 4.2(c)와 (d)는 지방함량이 높아짐에 따라 각각 육즙과 부드러움의 변화를 보여 주고 있으며, 지방함량 변환점의 MAP 추정치인 (43, 55)에 의해 두 변수의 변화를 잘 알 수 있다.

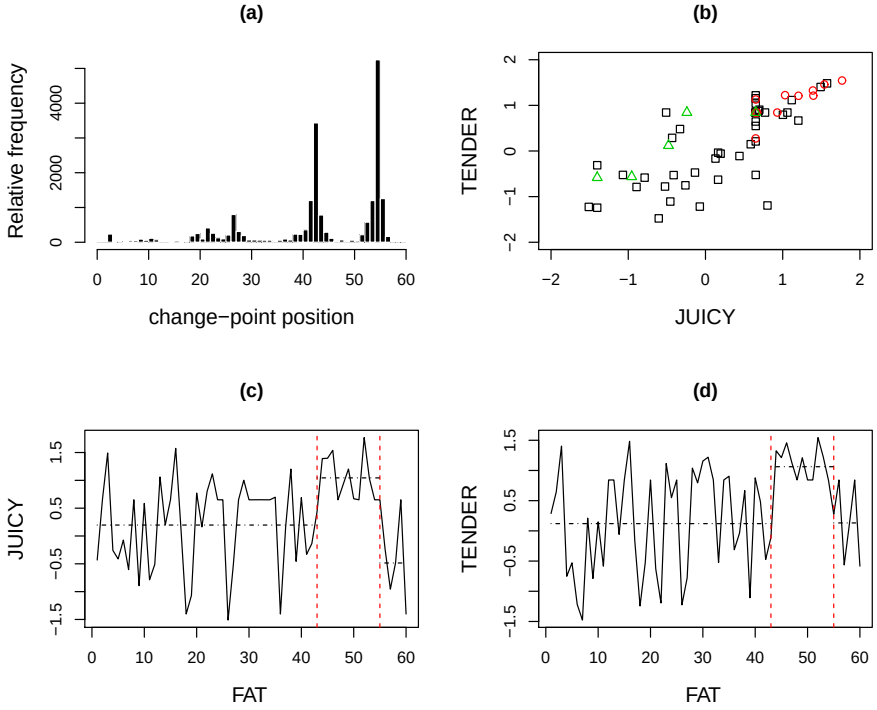

Figure 4.2. The simulation results of multiple change points estimation: (a) The histogram for log posterior probabilities of change points position; (b) A plot of juiciness and tenderness, with the maximum log-posterior estimate of the change-point positions; (c) A plot of the maximum log-posteriori estimated change positions with the change-point positions (the vertical lines) and means (horizontal lines) of juiciness in each subregion; (d) same with (c) except for tenderness)

이 결과는 Cheon과 Kim (2010)에서 제시한 결과와 비슷한 결과를 보여 준다. Cheon과 Kim (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 분석하였고, 그 결과 MAP 모형의 변환점 위치 (43, 55)를 찾았고 이 위치는 Cheon과 Kim (2010)에서 찾아낸 (233, 245) 시점과 정확하게 일치함을 알 수 있었다. 두 결과의 차이점으로, Cheon과 Kim (2010)은 베이지안 다변량 정규분포를 이용하였고 본 논문에서는 소량자료를 위해 다변량 비중심 t 분포를 사용하였다.

5. 결론

본 논문에서는 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 한우의 지방함량자료의 실증 분석에 적용한 결과, Cheon과 Kim (2010)의 결과와 비교해 볼때 변환점 개수와 그 위치를 적절하게 잘 찾아주고 있음을 알 수 있었다. 결론적으로 소량자료에 관한 변환점 및 위치를 찾고자 할 때 본 논문의 베이지안 다변량 비중심 t 분포 모형을 제안하며, 또한 결합분포에서 표본을 추출하기 힘들 때 메트로폴리스-해스팅스를 포함한 깁스 샘플링 알고리즘을 제안하고자 한다.

References

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

Carlin, B. P., Gelfand, A. E. and Smith, A. F. M. (1992). Hierarchical Bayesian analysis of change point problem, *Applied Statistics*, **41**, 389–405.

Chen, M. H. and Schmeiser, B. W. (1998). Towards black-box sampling, *Journal of Computational and Graphical Statistics*, **8**, 1–22.

Cheon, S. and Kim, J. (2010). Multiple change-point detection of multivariate mean vectors with the Bayesian approach, *Computational Statistics and Data Analysis*, **54**, 406–415.

Cheon, S. and Yu, W. (2012). Bayesian Multiple change-point for small data, *Communications of the Korean Statistical Society*, **19**, 237–246.

Chernoff, H. and Zacks, S. (1964). Estimating the current mean of a normal distribution which is subject to changes in time, *Annals of Mathematical Statistics*, **35**, 999–1018.

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**, 287–290.

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.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov Chains and their applications, *Biometrika*, **57**, 97–109.

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.

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**, 658–665.

Kim, J. and Cheon, S. (2010). Bayesian multiple change-point estimation with annealing stochastic approximation Monte Carlo, *Computational Statistics*, **25**, 215–239.

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.

Müller, P. (1991). A Bayesian analysis for change-point problems, A generic approach to posterior integration and Gibbs sampling. Technical Report, Purdue University, West Lafayette IN.

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

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

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.