응용통계연구 제 13권 1호, 2000년, pp. 115-131

변량모형 자료에서의 베이지안 이상점검출

정윤식[1] 이상진[2]

요 약

이 논문에서는 평균-이동모형(mean-shift model)을 이상점을 위한 대립모형으로 사용하여 변량모형(random effect model)에서의 이상점 검출을 위한 베이즈인자(Bayes factor)를 제시한다. 그러나 가능한 사전 정보가 없어서 무정보사전분포(noninformative prior distribution)가 사 용되어야만 할 때, 대부분의 무정보사전분포는 부적절분포(improper distribution)이기 때문에 베이즈인자에는 사전분포로 부터 나온 미지의 상수가 포함되어있다. 이 문제를 해결하기 위해 이 논문에서는 Berger 와 Pericchi (1996)가 제시한 내재베이즈인자 (the intrinsic Bayes factor; IBF)를 사용한다. 또한 이 베이즈인자를 계산상 어려움을 해결하기 위해 Verdinelli 와 Wasserman(1995)의 일반화 세비지디키 밀도비를 이용하여 수정하고 이것을 이용하여 이상점을 검출하는 방법을 제시한다. 마지막으로 인위적으로 이상점을 포함하고 있는 데이터를 만들고 제시된 방법으로 가상실험을 하고 또한 실제 데이터에서 제시한 방법으로 이상점을 찾아보았다.

주요용어: 내재베이즈인자, 메트로폴리스 표본추출법, 변량모형, 이상점, 평균이동모형.

1. 서론

통계적 자료분석을 할 때 자료에 이상점이 존재하는 것은 심각한 문제이 다. 베이지안 자료분석도 마찬가지이여서, 많은 베이지안 통계학자들은 이상 점검출에 많은 관심을 가져왔다. 베이지안 이상점검출 방법은 이상점을 위한 대립모형 을 사용하느냐 아니냐에 따라 크게 두 가지로 나누어진다.

대립모형을 사용하지 않은 방법으로는 Geisser (1985) 와 Pettit과 Smith (1985) 등의 예측분포(predictive distribution)를 이용하는 검출법과 Johnson과 Geisser (1983), Chaloner와 Brant (1988) 그리고 Guttman과 Pena (1993) 등의 사후확률분포(posterior distribution)를 사용하는 검출법이 있다.

이상점을 위한 대립모형으로는 평균이동모형(mean-shift model)과 분산팽창모형 (variance - inflation model)이 주로 사용된다. 평균이 μ이고 분산이 σ^2인 정규모집단으로 부터 자료 $\mathbf{y}$를 추출하였다 하자. 이때, 평균이동모형은 이상점 y_i가 $N(\mu+m_i,\sigma^2)$분포를 따른다고 가정하는 것이고, 분산팽창모형은 이상점이 $N(\mu,b_i\sigma^2)$분포로 부터 추출되었다고 생각하는 것이다, 이때, $m_i \neq 0$ 이고 $b_i >> 1$ 이다. Guttman (1973)은 평균이동모형을 선형모

1) (609-735) 부산시 금정구 장전동 산 30 부산대학교 통계학과, 부교수
E-mail: yschung@hyowon.pusan.ac.kr
2) (682-090) 울산시 남구 우거2동 산 29, 울산과학대학 컴퓨터정보학부, 전임강사

형에 적용했고, Sharples (1990)은 분산팽창모형이 일반계층적모형에 얼마나 쉽게 적용이 가능한가를 보였다.

이 논문에서는, 평균이동모형을 변량모형에 적용하여 이상점검출하는 방법을 제시하고자 한다. $\mathbf{Y} = (y_{ij})_{I\times J}$를 변량모형,

$$y_{ij} = \mu + e_i + \epsilon_{ij}, \;\; i = 1, \ldots, I, \;\; j = 1, \ldots, J, \tag{1.1}$$

으로부터 나온 자료행렬이라 하자. 여기서, μ는 y_{ij}의 평균이고 e_i와 ϵ_{ij}는 각각 평균이 0이고 분산이 σ_e^2와 σ^2인 독립정규확률변수 이다. 여기서 한 관측치, y_{ks}가 평균이동모형,

$$y_{ks} = \mu + m + e_k + \epsilon_{ks}, \;\; m \neq 0, \tag{1.2}$$

에서 나온 이상점으로 의심이 된다고 가정하자. 이때, m 는 관측치 y_{ks}의 평균이동모수 이다. 만약 $m = 0$ 이면 관측치 y_{ks}는 이상점이 아니고, 반대로 $m \neq 0$ 이면 y_{ks}는 이상점이 된다.

베이지안 검정은 주로 베이즈인자(Bayes factor)를 사용한다. 그러나 불완전사전분포가 사용하여 베이즈인자를 계산할 때, 불완전사전분포가 포함하고 있는 미지의 상수가 계산된 베이즈인자에 남아있는 문제점이 있다. 이 문제점을 극복하기위해, Berger 와 Pericchi (1996)가 제시한 내재베이즈인자 (the intrinsic Bayes factor; IBF)를 사용할 것이다. 원래의 데이터 $\mathbf{Y}$를 l번째 최소훈련표본 (the minimal training sample; MTS) $\mathbf{Y}(l)$와 나머지 $\mathbf{Y}(-l)$로 나눌 수 있다. 이러한 MTS는 L가지가 존재한다고 하자. 또한 $f_i(\mathbf{Y}|\boldsymbol{\theta}_i)$와 $\pi_i^N(\boldsymbol{\theta}_i)$를 각각 가설 H_i하에서의 우도함수와 주어진 부적절사전분포함수라 하자. 이때, 귀무가설 H_0를 선호하는 산술IBF (the arithemetic IBF; AIBF)와 기하IBF (geometric IBF; GIBF)는 각각

$$B_{01}^{AI} = \frac{1}{L}\sum_{l=1}^{L} B_{01}^{*}(l) \;\text{ 와 }\; B_{01}^{GI} = \{\prod_{l=1}^{L} B_{01}^{*}(l)\}^{1/L} \tag{1.3}$$

로 표현된다. 여기서,

$$B_{01}^{*}(l) = \frac{\int_{\boldsymbol{\Theta}_0} f_0(\mathbf{Y}(-l)|\boldsymbol{\theta}_0, \mathbf{Y}(l))\pi_0^N(\boldsymbol{\theta}_0|\mathbf{Y}(l))d\boldsymbol{\theta}_0}{\int_{\boldsymbol{\Theta}_1} f_1(\mathbf{Y}(-l)|\boldsymbol{\theta}_1, \mathbf{Y}(l))\pi_1^N(\boldsymbol{\theta}_1|\mathbf{Y}(l))d\boldsymbol{\theta}_1} \tag{1.4}$$

이다.

하지만 아직도 이 베이즈인자들은 계산적 어려움은 해결되지 않았다. 그래서 이 논문에서는 이 AIBF와 GIBF를 다음의 보조정리 1.1의 일반화 Savage-Dickey 밀도비를 이용하여 계산량을 줄여서 이상점검출을 위한 검정에 적용하고자 한다. Dickey (1971, 1976)는 단순가설 $H_0 : m = m_0$와 $H_1 : m \neq m_0$의 검정에서 Dickey의 조건, 즉 $\pi_1^N(\xi|m) = \pi_0^N(\xi)$, 이 만족되면 베이즈인자는 $B_{01} = \pi_1^N(m_0|\mathbf{Y})/\pi_1^N(m_0)$ 로 표현됨을 보였고 이것을 *Savage-Dickey* 밀도비라 불렀다. 여기서, ξ는 장애모수벡터이다.

Verdinelli와 Wasserman (1995)는 이 밀도비를 Dickey의 조건이 만족되지 않는 상황에서도 적용할 수 있도록 다음 보조정리 1.1와 같이 일반화 시켰다.

보조정리 1.1 (Verdinelli와 Wasserman, 1995) 만약 ξ에 대해서 $0 < \pi_1^N(m_0|\mathbf{Y}), \pi_1^N(m_0, \xi) < \infty$라면, $H_0 : m = m_0$를 선호하는 베이즈인자는

$$B_{01} = \frac{\pi_1^N(m_0|\mathbf{Y})}{\pi_1^N(m_0)} \cdot E^{\pi_1^N(\xi|m_0,\mathbf{Y})}[\frac{\pi_0^N(\xi)}{\pi_1^N(\xi|m_0)}] \tag{1.5}$$

와 같이 표현된다. 여기서, $E^{\pi_1^N(\xi|m_0,\mathbf{Y})}$ 는 확률분포 $\pi_1^N(\xi|m_0, \mathbf{Y})$에 대한 기대값 을 의미한다. 이것을 일반화 Savage-Dickey 밀도비라 말한다. 만약 Dickey의 조건이 만족되면 식 (1.5)에 있는 기대값 부분은 없어진다.

2절에서는 이상점검출에 대한 변량모형에서의 베이지안적 접근을 다루고, 3절에서는 이상점검출을 위한 IBF를 일반화 Savage-Dickey 밀도비를 이용하여 계산량을 줄이기 위한 수정을 할 것이다. 그리고 4절에서는 몇가지 자료에 제시한 방법을 적용하여 계산하고 그 성능에대해 논 할 것이다. 지금까지 제시하였던 한개의 이상점 검출에 대한 방법을 다중이상점을 검출할 수 있는 방법으로의 확장을 5절에서 할 것이다. 6절에서는 이에대한 예제를 다룬다.

2. 변량모형의 베이지안 접근

$\mathbf{Y} = \{y_{ij}, i = 1, \ldots, I, j = 1, \ldots, J\}$을 모형 (1.1) 과 (1.2)로 부터 나온 관측치행렬이라 하자. 한 관측치 y_{ks}가 이상점인가 아닌가를 판단하기 위해서 귀무가설 H_0 : "$\mathbf{Y}$에는 이상점이 없다" 와 대립가설 H_1 : "y_{ks}가 이상점이다" 가 검정된다. 이 검정은 주어진 k와 s에 대해

$$H_0 : m = 0 \text{와 } H_1 : m \neq 0 \tag{2.1}$$

를 비교하는 것과 같다.

편리를 위해, 분산비 $\phi = J\sigma_c^2/\sigma^2$를 정의하고 모수벡터를 $\boldsymbol{\theta} = (\mu, \sigma^2, \phi)$로 한다. H_0하에서의 우도함수는

$$L_0(\mu, \sigma^2, \phi) \propto \sigma^{-IJ}(1+\phi)^{-I/2} \exp\{-\frac{1}{2\sigma^2}(\frac{S_1^2 + IJ(\bar{y}_{..} - \mu)^2}{1+\phi} + S_2^2)\} \tag{2.2}$$

로 주어진다. 단, $\bar{y}_{i.} = \sum_j y_{ij}/J$, $\bar{y}_{..} = \sum_i \sum_j y_{ij}/IJ$ $S_1^2 = J\sum_i(\bar{y}_{i.} - \bar{y}_{..})^2$ 그리고 $S_2^2 = \sum_i \sum_j (y_{ij} - \bar{y}_{i.})^2$이다.

다음으로, H_1하에서의 우도함수를 구해보자. H_1하에서, 자료 $\mathbf{Y}$에 y_{ks} 대신 $y_{ks} - m$를 넣은 자료

$$\{y_{ij}, (i,j) \neq (k,s),\ y_{ks} - m\} \tag{2.3}$$

는 이상점이 포함되지 않은 자료로 볼 수 있다. $\bar{y}_{m..}$, $\bar{y}_{mk.}$, S_{m1}^2 그리고 S_{m2}^2를 각각 $\mathbf{Y}$ 대신에 식(2.3)에 있는 자료로 계산된 $\bar{y}_{..}$, $\bar{y}_{k.}$, S_1^2 그리고 S_2^2라 하자. 그러면 $\bar{y}_{m..} = \bar{y}_{..} - m/IJ$,

$\bar{y}_{mk.} = \bar{y}_{k.} - m/J$, $S_{m1}^2 = S_1^2 - 2(\bar{y}_{k.} - \bar{y}_{..})m + \frac{I-1}{IJ}m^2$, 그리고 $S_{m2}^2 = S_2^2 - 2(y_{ks} - \bar{y}_{k.})m + \frac{J-1}{J}m^2$ 로 표현된다. 그래서 H_1하에서의 우도함수는

$$\begin{aligned} L_1(\mu, m, \sigma^2, \phi) &\propto \sigma^{-IJ}(1+\phi)^{-I/2}\exp[-\frac{1}{2\sigma^2}\{\frac{S_{m1}^2 + IJ(\bar{y}_{m..} - \mu)^2}{1+\phi} + S_{m2}^2\}] \\ &\propto L_0(\mu, \sigma^2, \phi) \cdot \exp[-\frac{1}{2\sigma^2}\{(\frac{IJ - I + 1}{IJ(1+\phi)} - 2\frac{I-1}{IJ})m^2 \\ &\quad -2(\frac{1}{1+\phi}(y_{ks} - \mu) + \frac{\phi}{1+\phi}(\bar{y}_{k.} - \bar{y}_{..}))m\}] \end{aligned} \tag{2.4}$$

로 쓸 수 있다.

모든 모수에 대한 사전정보가 전혀 없다고 가정하자. 이때, 무정보적 사전분포가 귀무가설과 대립가설 모두에 대해 사용되어지며, Tiao와 Tan (1966) 그리고 Box와 Tiao (1973)의 무정보적 사전분포

$$\pi_0^N(\mu, \sigma^2, \phi) \propto \sigma^{-2}(1+\phi)^{-1} \tag{2.5}$$

를 귀무가설하에서의 사전분포로 사용할 수 있다. 평균이동모수 m은 위치모수 이고 $\{\mu, \sigma^2, \phi\}$와 독립이라 가정하면, H_1하에서의 사전분포는

$$\pi_1^N(\mu, m, \sigma^2, \phi) \propto \sigma^{-2}(1+\phi)^{-1}. \tag{2.6}$$

로 표현할 수 있다.

관측치 y_{ks}가 이상점 인가 아닌가를 검정한다고 하자. $\mathbf{Y}(l)$과 (I_M, J_M)는 각각 l번째 MTS와 그것의 크기라 하자, $l = 1, \ldots, L$. 여기서 L은 가능한 MTS의 수이다. 이때, 식 (1.3)에 있는 귀무가설 H_0을 선호하는 AIBF와 GIBF는 각각

$$B_{01}^{AI}(\mathbf{Y}) = \frac{1}{L}\sum_{l=1}^{L} B_{01}^{*N}(l) \text{와 } B_{01}^{GI}(\mathbf{Y}) = \prod_{l=1}^{L}\{B_{01}^{*N}(l)\}^{\frac{1}{L}} \tag{2.7}$$

로 계산된다. 여기서,

$$B_{01}^{*}(l) = B_{01}^{N} \cdot B_{10}^{N}(l) \tag{2.8}$$

이고, 또한

$$B_{01}^N = m_0(\mathbf{Y})/m_1(\mathbf{Y}) \text{ 그리고 } B_{10}^N(l) = m_1(\mathbf{Y}(l))/m_0(\mathbf{Y}(l)),$$

$$m_0^N(\mathbf{X}) = \int\int\int \pi_0^N(\mu, \sigma^2, \phi) L_0(\mu, \sigma^2, \phi; \mathbf{X}) d\mu d\sigma^2 d\phi,$$

$$m_1^N(\mathbf{X}) = \int\int\int\int \pi_1^N(\mu, m, \sigma^2, \phi) L_1(\mu, m, \sigma^2, \phi; \mathbf{X}) d\mu dm d\sigma^2 d\phi, \tag{2.9}$$

이다, $l = 1, \ldots, L$.

부적절사전분포로 부터 생긴 베이즈인자에 포함되어있는 미지의 상수 들은 식 (2.8)에서 상쇄된다. 그리고 귀무가설 H_0하에서의 주변분포함수 $m_0^N(\mathbf{X})$는 구할 수 있으나, 대립가설 H_1하에서의 주변분포함수 $m_1^N(\mathbf{X})$는 m에 대한 직접적분이 불가능 하므로 정확한 계산이 힘들다. 그래서 다음 절에서는, 이 AIBF와 GIBF에 일반화 Savage-Dickey 밀도비를 적응시켜 계산량을 줄이고 중요표본계산법 (importance sampling method)을 이용하여 적분하는 방법을 제시할 것이다.

3. 이상점검출을 위한 IBF의 계산

모수벡터 $\boldsymbol{\theta} = (m, \mu, \sigma^2, \phi)$를 가지고 있는 한 통계모형을 고려해보자. 여기서 우리는 모수 m에 관심을 가지고 있으며 (μ, σ^2, ϕ)는 장애모수 (nuisance parameter) 벡터이고 이를 ξ라 표기하자. 식 (2.1)에 있는 단순가설을 베이지안 검정하기위해 식 (2.5)의 $\pi_0^N(\mu, \sigma^2, \phi)$와 식 (2.6)의 $\pi_1^N(m, \mu, \sigma^2, \phi)$를 각각 식 (2.1)의 H_0와 H_1하에서의 사전확률분포로 사용한다. 계산량의 부담을 줄이기 위해, 식 (2.7)에서 구한 AIBF와 GIBF를 보조정리 1.1의 일반화 Savage-Dickey 밀도비 개념을 이용하면 다음과 같이 표현할 수 있다.

정리 3.1 귀무가설 $H_0 : m = m_0$를 선호하는 AIBF와 GIBF는 각각

$$B_{01}^{AI} = \frac{1}{L}\sum_{l=1}^{L} B_{01}^*(l) \text{ 와 } B_{01}^{GI} = \{\prod_{l=1}^{L} B_{01}^*(l)\}^{\frac{1}{L}} \tag{3.1}$$

로 표현할 수 있다. 여기서, $l = 1, \ldots, L$에 대하여

$$B_{01}^*(l) = \frac{\pi_1^N(m_0|\mathbf{Y}) E^{\pi_1^N(\boldsymbol{\xi}|m_0,\mathbf{Y})}\{\frac{\pi_0^N(\boldsymbol{\xi})}{\pi_1^N(\boldsymbol{\xi}|m_0)}\}}{\pi_1^N(m_0|\mathbf{Y}(l)) E^{\pi_1^N(\boldsymbol{\xi}|m_0,\mathbf{Y}(l))}\{\frac{\pi_0^N(\boldsymbol{\xi})}{\pi_1^N(\boldsymbol{\xi}|m_0)}\}} \tag{3.2}$$

이고 $E^{\pi_1^N(\boldsymbol{\xi}|m_0,\mathbf{X})}$는 밀도함수 $\pi_1^N(\boldsymbol{\xi}|m_0, \mathbf{X})$에 대한 기대값을 말한다.

증명: B_{01}^N와 $B_{01}^N(l)$는 각각 불완전사전분포와 모든 데이터와 l번째 MTS를 사용해서 계산한 베이즈인자 이다, $l = 1, \ldots, L$. 그래서 B_{01}^N와 $1/B_{10}^N(l)$는 모두 단순귀무가설에 대한 베이즈인자 이므로 보조정리 1.1을 사용하여 다음과 같이 계산할 수 있다,

$$B_{01}^N = \frac{\pi_1^N(m_0|\mathbf{Y})}{\pi_1^N(m_0)} E^{\pi_1^N(\boldsymbol{\xi}|m_0,\mathbf{Y})}\{\frac{\pi_0^N(\boldsymbol{\xi})}{\pi_1^N(\boldsymbol{\xi}|m_0)}\}$$

그리고

$$1/B_{10}^N(l) = B_{01}^N(l) = \frac{\pi_1^N(m_0|\mathbf{Y}(l))}{\pi_1^N(m_0)} E^{\pi_1^N(\boldsymbol{\xi}|m_0,\mathbf{Y}(l))}\{\frac{\pi_0^N(\boldsymbol{\xi})}{\pi_1^N(\boldsymbol{\xi}|m_0)}\}.$$

그래서 식(2.8)을 이용하여 식 (3.2)는

$$\begin{aligned} B_{01}^{*}(l) &= B_{01}^{N} \cdot B_{10}^{N}(l) = \frac{B_{01}^{N}}{B_{01}^{N}(l)} \\ &= \frac{\pi_1^N(m_0|\mathbf{Y})}{\pi_1^N(m_0|\mathbf{Y}(l))} \frac{E^{\pi_1^N(\boldsymbol{\xi}|m_0,\mathbf{Y})}\{\frac{\pi_0^N(\boldsymbol{\xi})}{\pi_1^N(\boldsymbol{\xi}|m_0)}\}}{E^{\pi_1^N(\boldsymbol{\xi}|m_0,\mathbf{Y}(l))}\{\frac{\pi_0^N(\boldsymbol{\xi})}{\pi_1^N(\boldsymbol{\xi}|m_0)}\}} \end{aligned}$$

로 쉽게 유도할 수 있다. □

식(3.1)의 AIBF B_{01}^{AI}와 GIBF B_{01}^{GI}를 각각 세비지디키 *(Savage-Dickey) AIBF*와 세비지디키 *(Savage-Dickey) GIBF*라 부르기로 하고 줄여서는 SDAIBF와 SDGIBF라 표기하기로 하자.

관심모수 m의 주변밀도함수 $\pi_1^N(m)$는 조건부밀도함수 $\pi_1^N(\boldsymbol{\xi}|m,\mathbf{Y})$와 $\pi_1^N(\boldsymbol{\xi}|m,\mathbf{Y}(l))$에 독립이므로, 식 (3.2) $B_{01}^{*}(l)$의 분모와 분자에 $\pi_1^N(m_0)$을 각각 나누어 주어도 그 값에는 변화가 없다. 그러므로 식 (3.2)의 우변에 있는 분모와 분자의 기대값 속의 분모 부분인 $\pi_1^N(\boldsymbol{\xi}|m_0)$ 대신에 $\pi_1^N(\boldsymbol{\xi},m_0)$을 사용해도 된다. 그래서 계산적인 편리성을 위해서 이 바꾸어진 식을 사용하기로 한다. 또한, Dickey의 조건, $\pi_1^N(\boldsymbol{\xi}|m_0)=\pi_0^N(\boldsymbol{\xi})$이 만족되는 상황이면 식 (3.2)에서 기대값 부분이 사라진다는 것을 쉽게 유도할 수 있다.

이제, 정리 3.1를 이용하여 변량모형에서 이상점검출을 위한 SDAIBF와 SDGIBF를 계산해 보자. 먼저 이상점으로 의심받는 한 관측치 y_{ks}에 대한 평균이동모수 m의 대립가설 H_1하에서의 사후주변확률밀도함수의 계산이 필요하다.

보조정리 3.1 식 (2.6)에 있는 확률함수를 가설 H_1하의 사전확률함수로 사용하고 식 (2.4)에 있는 $L_1(\mu,m,\sigma^2,\phi)$를 우도함수로 사용한 m의 사후주변확률함수는

$$\pi_1(m|\mathbf{Y}) = C\beta_{p,q}(\frac{W_m}{W_m+1})(S_{m2}^2)^{-(p+q)}W_m^{-p} \tag{3.3}$$

로 계산된다. 여기서,

$$C^{-1} = \int_{-\infty}^{\infty} \beta_{p,q}(\frac{W_m}{W_m+1})W_m^{-p}S_{m2}^{2(-p-q)}dm \tag{3.4}$$

이고 S_{m1}^2와 S_{m2}^2는 식 (2.4)에 정의되어있고 $W_m = S_{m1}^2/S_{m2}^2$, $p=(I-1)/2$, $q=I(J-1)/2$ 그리고 $\beta_{i,j}(x)=\int_0^x t^{i-1}(1-t)^{j-1}dt$는 불완전베타함수 이다.

증명: (m,μ,σ^2,ϕ)의 결합사후확률함수는

$$\begin{aligned} \pi_1(m,\mu,\sigma^2,\phi|y_{..},s_1^2,s_2^2) \propto\ & \sigma^{-(IJ+2)}(1+\phi)^{-(I+2)/2} \\ & \cdot\exp[-\frac{1}{2\sigma^2}\{\frac{S_{m1}^2+IJ(y_{m..}-\mu)^2}{1+\phi}+S_{m2}^2\}] \end{aligned}$$

이다. 그래서 m의 사후주변확률함수는 다음과 같이 적분함으로써 구할 수 있다.

$$\begin{aligned}\pi_1(m|\mathbf{Y}) &\propto \int_0^\infty\int_0^\infty\int_{-\infty}^\infty \pi_1(\mu,m,\sigma^2,\phi|\mathbf{Y})d\mu d\sigma^2 d\phi \\ &\propto \int_0^\infty\int_0^\infty \sigma^{-(IJ+2)}(1+\phi)^{-(I+2)/2}\exp\{-\frac{1}{2\sigma^2}(S_{m2}^2+\frac{S_{m1}^2}{1+\phi})\} \\ &\quad\cdot\{\frac{2\pi}{IJ}\sigma^2(1+\phi)\}^{1/2}d\sigma^2 d\phi \\ &\propto \int_0^\infty (1+\phi)^{-(I+1)/2}(S_{m2}^2+\frac{S_{m1}^2}{1+\phi})^{-(IJ-1)/2}d\phi.\end{aligned}$$

여기서 $Z_m = W_m/(W_m+1+\phi)$라 하자. 그러면,

$$\begin{aligned}\pi_1(m|\mathbf{Y}) &\propto (S_{m2}^2)^{-(p+q)}W_m^{-p}\int_0^{\frac{W_m}{W_m+1}} Z_m^{p-1}(1-Z_m)^{q-1}dZ_m \\ &= (S_{m2}^2)^{-(p+q)}W_m^{-p}\beta_{p,q}(\frac{W_m}{W_m+1}).\end{aligned}$$

□

변량모형에서의 MTS의 크기를 (I_M, J_M)라 하자. 또한 $p_M = (I_M-1)/2$라 하고 $q_M = I_M(J_M-1)/2$라 하자. 그리고 $S_1^2(l)$와 $S_2^2(l)$, $W(l)$는 각각 l번째 MTS로 계산한 S_1^2와 S_2^2, W라 표기하자. 이 표기들을 사용하여 이상점 검출을 위한 SDAIBF와 SDGIBF는 다음 정리 3.2에서 유도된다.

정리 3.2 식 (2.1)에 있는 귀무가설 H_0를 선호하는 SDAIBF와 SDGIBF는 각각

$$B_{01}^{AI}(\mathbf{Y}) = \frac{1}{L}\sum_{l=1}^{L} B_{01}^*(l) \text{ 와 } B_{01}^{GI}(\mathbf{Y}) = \prod_{l=1}^{L}\{B_{01}^*(l)\}^{\frac{1}{L}} \tag{3.5}$$

이다. 여기서 $l=1,\ldots,L$에 대해서

$$B_{01}^*(l) = \frac{C_1\beta_{p,q}(\frac{W}{W+1})(S_2^2)^{-(p+q)}W^{-p}}{C_1(l)\beta_{p_M,q_M}(\frac{W(l)}{W(l)+1})(S_2(l)^2)^{-(p_M+q_M)}W(l)^{-p_M}} \tag{3.6}$$

이다. C_1^{-1}은 보조정리 3.1에서 구해진 것이고 $C_1(l)^{-1}$은 $\mathbf{Y}$와 p, q 대신에 l번째 MTS $\mathbf{Y}(l)$과 p_M, q_M을 사용하여 보조정리 3.1과 같은 방법으로 구해진 C_1^{-1} 이다. 또한, $S_{m1}(l)^2$ 와 $S_{m2}(l)^2$, $W_m(l)$ 는 각각 전체 자료 대신에 l번째 MTS $\mathbf{Y}(l)$를 사용하여 계산된 S_{m1}^2 와 S_{m2}^2, W_m 이다, 여기서 $l=1,\ldots,L$ 이다.

증명: 사용된 사전분포가 디키의 조건을 만족하므로 전체자료 $\mathbf{Y}$와 l번째 MTS $\mathbf{Y}(l)$을 각각 사용한 m의 사후주변확률함수의 비를 계산하면 된다. $\pi_1(m|\mathbf{Y})$는 보조정리 3.1에서 구했고 $\pi_1(m|\mathbf{Y}(l))$도 같은 방법으로

$$\pi_1(m|\mathbf{Y}(l)) = C_1(l)\{S_{m2}^2(l)\}^{-(p_M+q_M)}W_m(l)^{-p_M}\beta_{p_M,q_M}(\frac{W_m(l)}{W_m(l)+1}), \tag{3.7}$$

$l = 1, \ldots, L$, 와 같이 구할 수 있다. 만약 $m = 0$이면, $y_{ks} - m$는 y_{ks}와 같아지므로 S^2_{m2} 와 W_m, $S_{m2}(l)^2$, $W_m(l)$ 는 각각 S_2^2 와 W, $S_2(l)^2$, $W_l(l)$ 와 같아진다, $l = 1, \ldots, L$. 그러므로 $B^*_{01}(l)$를 계산하여 정리 3.2는 쉽게 증명된다. □

구해진 SDAIBF와 SDGIBF에 있는 상수 C_1와 $C_1(l), l = 1, \ldots, L$는 해석학 적으로 구하기는 불가능하다. 그러나 중요표본계산법 (importance sampling method)과 같은 표본추출계산법 (sampling based computation)을 사용한 수치적 방법으로 이들을 추정할 수 있다. 즉,

$$g(m) = \beta_{p,q}(\frac{W_m}{W_m + 1})W_m^{-p}(S^2_{m2})^{-(p+q)}$$

라 하면,

$$C^{-1} = \int_{-\infty}^{\infty} g(m)dm = \int_{-\infty}^{\infty} \frac{g(m)}{I(m)} I(m)dm = E\{\frac{g(m)}{I(m)}\}$$

와 같이 둘 수 있다. 단, $I(m)$는 $g_r(m)$와 같은 정의구역을 가지는 확률밀도함수이다. 그래서 이 함수로 부터 표본 $\{m^{(1)}, \ldots, m^{(G)}\}$을 생성시킨 후, 다음과 같은 Monte Carlo 방법으로 C^{-1}의 값을 추정할 수 있다;

$$\hat{C}^{-1} = \frac{1}{G}\sum_{g=1}^{G} \frac{g(m^{(g)})}{I(m^{(g)})}. \tag{3.8}$$

이 계산법에서는 중요함수 $I(m)$의 선택이 매우 중요하다. 여기서는 메트로폴리스 표본추출법 (Metropolis 외4명, 1953) 으로 $g(m)$으로부터 표본을 추출한 다음 이 표본의 표본평균 $\bar{m}$와 표본분산 s^2_m을 평균과 분산으로 가지는 정규밀도함수를 중요함수 (importance function) $I(m)$로 사용하는 방법을 제시한다.

4. 한 이상점 검출의 예제들

4.1. 생성자료

이 절에서는 먼저 앞 절에서 제시한 SDAIBF와 SDGIBF를 이용한 이상점검출 방법을 이상점이 포함된 생성된 자료로 모의실험을 수행한다. 표4.1에 있는 자료는 평균이 $\mu = 5$이고 분산이 각각 $\sigma_e^2 = 6$과 $\sigma^2 = 8$ 이고 $I = 6, J = 5$를 사용하여 균형변량모형 (1.1)로 부터 생성된 자료이다. 이 자료의 한 관측치 y_{52}는 위와 같은 분산들을 가지며 평균이 $\mu = 0$인 균형변량모형 (1.1)로 부터 생성된 관측치 이다. 즉, $m = -5$인 모형 (1.2)로 부터 생성되었다. 생성된 자료는 표4.1에 있다. 모든 관측치가 이상점인가 아닌가에 대해 각각 수행해 보기로 한다.

우선, MTS의 크기가 결정되어야 한다. MTS는 귀무가설 H_0와 대립가설 H_1하에서의 주변확률함수

$$m_0(\mathbf{Y}) \propto \beta_{p,q}(\frac{W}{W+1})\frac{1}{(S_2^2)^{p+q}W^p} \tag{4.1}$$

표 4.1: 이상점이 포함된 생성자료

Batch	1	2	3	4	5	6
obs. 1	7.8925	-0.0030	10.1009	13.6895	0.5623	5.3777
obs. 2	12.6125	7.0934	5.0114	10.6080	-8.4583	10.1637
obs. 3	4.3213	12.4114	7.9833	9.6563	0.7844	3.4680
obs. 4	13.1566	7.8590	11.1319	11.2744	5.6431	5.4790
obs. 5	12.8839	9.0184	6.7217	3.8906	5.1731	7.5221

표 4.2: 생성된 자료에서의 이상점 검출을 위한 SDAIBF와 SDGIBF

s	k	1	2	3	4	5	6
1	AIBF	1.0693	0.7329	0.9458	0.8759	0.9446	1.0567
	GIBF	1.0657	0.7305	0.9384	0.8712	0.9386	1.0459
2	AIBF	0.8952	1.0048	1.1211	0.9617	0.6034	0.9040
	GIBF	0.8889	0.9991	1.1159	0.9522	0.6021	0.9004
3	AIBF	0.8093	0.8487	1.0224	0.9818	0.9534	0.9151
	GIBF	0.8051	0.8453	1.0128	0.9758	0.9447	0.9052
4	AIBF	0.8824	0.9909	0.9228	0.9489	0.8939	1.0511
	GIBF	0.8798	0.9830	0.9157	0.9421	0.8925	1.0425
5	AIBF	0.8898	0.9639	1.0576	0.8048	0.9121	0.9956
	GIBF	0.8852	0.9560	1.0498	0.8016	0.9067	0.9890

와

$$m_1(\mathbf{Y}) \propto \int \beta_{p,q}(\frac{W_m}{W_m+1}) \frac{1}{(S_{m2}^2)^{p+q} W_m^p} dm \tag{4.2}$$

모두가 유한성을 만족하는 가장 작은 크기이다. 식 (4.1)과 식 (4.2)가 모두 유한성을 만족하기 위해서는 각각의 함수에 포함되어 있는 불완전베타함수 $\beta_{p,q}(\frac{W}{W+1})$와 $\beta_{p,q}(\frac{W_m}{W_m+1})$가 다 유한해야 한다. 그러기 위해서는 p와 q가 0보다는 커야 한다. 이를 만족하는 MTS의 크기는 $I_M = 2$와 $J_M = 2$ 이다. 상수 C_1와 $C_1(l)$는 3절에서 제시한 중요표본방법으로 계산하였다.

각 관측치에 대한 귀무가설, H_0 : 이상점이 없다, 를 선호하는 SDAIBF와 SDGIBF의 값들이 표4.2에 나열하였다. 이 표를 보면, 관측치 y_{52}에 대한 베이즈 인자 값들이 1 보다 현저하게 작은 값이다. 그러므로 제시한 SDAIBF와 SDGIBF의 방법에 따라 관측치 y_{52}를 이상점이라 할 수 있다.

그러나 관측치 y_{21}에 대한 SDAIBF와 SDGIBF도 1 보다 작은 값을 가진다고 볼 수 있다. 이는 자료를 생성할 때 이상점으로 만들려고 하지는 않았으나 e_2와 ϵ_{21}가 중심점 0보다

상당히 작은 값으로 생성되어서 y_{21}가 만들어졌기 때문이다. 변량모형의 특성상 각 관측치에 영향을 주는 인자가 여러개 존재 하므로 이들이 한 쪽 방향으로 값이 생성되는 경우 이 상점에 가까운 관측치가 나올 수 있다. 즉, 이러한 관측치는 중심에서 상당히 떨어진 값을 가진다. 그래서 이 관측치에 대한 이상점 검출을 위한 베이즈 인자가 1 보다 작은 값을 가진다.

4.2. 실제자료

이번에는 실제 자료에 제시된 방법을 적용해보자. 실제 자료로 사용한 것은 Dyestuff 자료로써 여섯가지 종류의 생산에 각각 크기 5의 표본을 추출하여 표준색감의 생산량을 그람으로 나타낸 것이다. 이 자료를 표4.3에 있다.

표 4.3: Dyestuff 자료

Batch	1	2	3	4	5	6
obs. 1	1545	1540	1595	1445	1595	1520
obs. 2	1440	1555	1550	1440	1630	1455
obs. 3	1440	1490	1605	1595	1515	1450
obs. 4	1520	1560	1510	1465	1635	1480
obs. 5	1580	1495	1560	1545	1625	1445

표 4.4: Dyestuff 자료에서 이상점 검출을 위한 SDAIBF 와 SDGIBF

s	k	1	2	3	4	5	6
1	AIBF	0.9918	0.9967	0.9907	1.0152	0.9963	0.9921
	GIBF	0.9918	0.9966	0.9907	1.0150	0.9962	0.9921
2	AIBF	1.0172	0.9938	1.0005	1.0149	0.9891	1.0078
	GIBF	1.0172	0.9937	1.0004	1.0148	0.9890	1.0077
3	AIBF	1.0170	1.0080	0.9898	0.9777	1.0170	1.0076
	GIBF	1.0170	1.0079	0.9898	0.9777	1.0170	1.0074
4	AIBF	0.9971	0.9940	1.0106	1.0114	0.9896	1.0017
	GIBF	0.9970	0.9940	1.0106	1.0114	0.9894	1.0017
5	AIBF	0.9835	1.0084	0.9995	0.9900	0.9892	1.0072
	GIBF	0.9834	1.0083	0.9995	0.9900	0.9891	1.0071

생성된 자료와 같은 MTS의 크기를 사용하였다. 모든 자료 각각이 이상점인가를 알아보기 위해, 귀무가설을 선호하는 SDAIBF와 SDGIBF를 계산하여 표4.4에 나타내었다.

표 4.4의 결과를 보면 SDAIBF와 SDGIBF의 값들이 1보다 확실히 작다고 할 수 있는 것이 없다. 그러므로 Dyestuff 자료는 이상점이 없는 자료라 할 수 있다. 그리고 첫번째 열의 값들이 1과는 비슷하지만 다른 값들에 비해 비교적 약간 작은 성향을 볼 수 있으므로 첫번째 확률모형인자 e_1가 다른 인자들과 조금 특이한 성향을 가진다는 것을 판단할 수도 있다.

5. 다중이상점 검출

데이터 행렬 $\mathbf{Y}$는 (1.1)의 모형을 따른다고 할 때, 관측치 집합 $\{y_{k_1 s_1}, \ldots, y_{k_a s_a}\}$이 다중이상점 집합 인가 아닌가를 알려고 한다. 여기서, a는 고려하고 있는 다중이상점군의 크기이고 (k_i, s_i)들은 모두 다르다. 이것은 가설

$$H_0 : \mathbf{m} = 0 \ \text{와} \ H_1 : \mathbf{m} \neq 0 \tag{5.1}$$

을 검정함으로써 할 수 있다. 여기서, $\mathbf{m} = \{m_{k_1 s_1}, \ldots, m_{k_a s_a}\}$ 이고 0은 a-차원 영벡터 이다. 표현의 편리성을 위해, 지금부터 $m_i = m_{k_i s_i}$라 하자. 단, $i = 1, \ldots, a$.

귀무가설 H_0 하에서는 모든 m_i들이 0이므로, 우도함수는 식 (3.2)의 $L_0(\mu, \sigma^2, \phi)$과 같다. 대립가설 H_1 하에서의 우도함수를 구하기 위하여 다음의 상상자료집합을 생각해 보자,

$$\mathbf{Y}_m^a = \{y_{ij}, (i,j) \notin \{(k_1, s_1), \ldots, (k_a, s_a)\}, \ y_{k_v s_v} - m_v, v = 1, \ldots a, \}. \tag{5.2}$$

이 상상자료는 대립가설 H_1 하에서 이상점이 전혀 없는 자료가 된다. 그러므로 이 자료를 이용하여 3절의 방법과 비슷한 방식으로 대립가설 H_1 하에서의 우도함수 $L_1(\mu, \mathbf{m}, \sigma^2, \phi)$를 구할 수 있다. 구하여진 우도함수는 (2.3)의 자료 대신에 $\mathbf{Y}_m^a$을 사용한 식 (2.4)와 같게 된다.

모수 벡터 $\mathbf{m}$은 위치모수이고 서로 독립이므로 사전분포함수를 $\pi_1(\mathbf{m}) = 1$로 사용하여 대립가설 하에서의 사전분포함수를 구하면 식 (2.6)과 같게 된다.

가설 (5.1)을 검정하기 위한 SDAIBF와 SDGIBF를 계산하기위해서 우선 대립가설 H_1 하에서 $\mathbf{m}$의 주변사후분포함수를 구해보자. 앞에서 구한 사전분포함수와 우도함수 $L_1(\mu, \mathbf{m}, \sigma^2, \phi)$를 사용하여 보조정리 3.1과 같은 방법으로 $\mathbf{m}$의 주변분포함수를

$$\pi_1(\mathbf{m}|\mathbf{Y}) = C_a \beta_{p,q}\left(\frac{W_m}{W_m + 1}\right)(S_{m2}^2)^{-(p+q)} W_m^{-p} \tag{5.3}$$

와 같이 구할 수 있다. 여기서,

$$1/C_a = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \beta_{p,q}\left(\frac{W_m}{W_m + 1}\right) W_m^{-p} S_{m2}^{2(-p-q)} dm_1 \cdots dm_a, \tag{5.4}$$

이고 S_{m1}^2과 S_{m2}^2, W_m는 식 (2.3)의 데이터 대신에 $\mathbf{Y}_m^a$을 사용하여 구한 통계량들이다. 평균이동모수 $\mathbf{m}$이 귀무가설 하에서는 $\mathbf{0}$이므로 데이터 $\mathbf{Y}_m^a$는 원래의 데이터 $\mathbf{Y}$가 된다. 그러므로 귀무가설 H_0를 선호하는 SDAIBF와 SDGIBF는 각각 다음과 같이 구해진다;

$$B_{01}^{AI}(\mathbf{Y}) = \frac{1}{L}\sum_{l=1}^{L} B_{01}^{*}(l) \ \text{ and } \ B_{01}^{GI}(\mathbf{Y}) = \prod_{l=1}^{L} \{B_{01}^{*}(l)\}^{\frac{1}{L}}. \tag{5.5}$$

여기서,

$$B^*_{01}(l) = \frac{C^a_1 \beta_{p,q}(\frac{W}{W+1})(S^2_2)^{-(p+q)}W^{-p}}{C^a_1(l)\beta_{p_M,q_M}(\frac{W(l)}{W(l)+1})(S_2(l)^2)^{-(p_M+q_M)}W(l)^{-p_M}} \tag{5.6}$$

이고, $l = 1, \ldots, L$, $1/C_a(l)$는 전체자료 대신에 l번째 최소훈련표본을 사용하여 $1/C_a$와 같은 방법으로 정의되는 값이다, $l = 1, \ldots, L$.

식 (3.6)에 있는 상수들과 마찬가지로, $1/C_a$와 $1/C_a(l)$의 계산은 다변량 중요표분추출법을 사용하여 구할 수 있다. 여기서는 독립정규밀도함수를 중요함수로 사용하여 계산한다.

6. 다중이상점 검출의 예제들

이 절에서는 앞 절에서 제시한 다중이상점 검출법을 이상점이 두 개 있는 모의자료를 이용하여 모의실험을 하고 실존 자료에 적용하여 베이즈 인자 값들을 계산한다. 앞 절에서 여러개의 이상점을 검출하는 방법들은 비슷한 방법들 이므로 여기에서는 두 개의 이상점을 찾는 베이즈인자만을 계산한다.

모의자료는 모형 (1.1)에서 평균 $\mu = 8$, 분산 $\sigma^2_e = 3^2$와 $\sigma^2 = 4^2$를 사용하여 생성하였고 관측치 y_{21}와 y_{32}를 $m_{21} = -5$와 $m_{32} = 5$를 각각 사용하여 이상점으로 생성하였다. 이 생성된 자료는 표6.1에 있다.

표 6.1: 두 개의 이상점을 가진 생성 자료

Batch	1	2	3	4	5	6
obs. 1	7.0598	-2.3653	1.9783	7.5280	6.3272	6.8342
obs. 2	7.8652	8.4604	11.2092	5.9631	7.2833	9.3246
obs. 3	7.1009	7.3435	0.7544	9.9128	2.8004	9.3951
obs. 4	7.6271	5.0220	4.4633	8.7149	3.9802	9.5916
obs. 5	11.1953	6.7601	2.1682	10.7456	5.0674	9.0282

베이즈인자에 들어 있는 상수들의 계산을 위하여 독립 이변량 정규분포 밀도함수를 중요함수로 사용하였다. 생성자료의 모든 관측치를 두 개씩 짝을 지어서 이상점군 인가 아닌가를 위한 베이즈인자의 값들을 계산하여 표 6.2에 정리 했다. 두 개씩의 짝들이 너무 많아서 표에는 베이즈인자의 값이 작은 것부터 순서대로 축약하여 정리를 하였다. 이상점군으로 생성된 관측치군 외에도 이상점으로 판단되어지는 군이 상당히 존재함을 알 수 있다. 이는 두 개의 이상점 중 한 개가 포함된 군들과 두 이상점외에도 중심으로 부터 멀리 떨어져 있는 관측치군이 존재함을 알 수 있다.

다음으로 4절에 있는 실제 자료인 Dyestuff 자료에 적용해 보았다. 계산된 베이즈인자 값들은 표 6.3에 축약하여 정리 하였다. 분명한 이상점군으로 판단하기에는 다소 무리가 있으나 중심으로 부터 어느 정도 떨어져 있는 관측치군이 있슴을 결과표로 부터 알 수 있다.

이 논문에서는 2개로 이루어진 이상점군에 대한 결관만 실었다. 3개로 이루어지거나 더 큰 이상점군에 대해서도 평균이동모수 $\mathbf{m}$을 3차원이나 그이상 크기의 벡터로 하여 같은 방법으로 다중 이상점군을 찾을 수 있다.

표 6.2: 생성자료에서 크기 2의 이상점군 검출을 위한 SDAIBF와 SDGIBF

Observ.'s	SDAIBF	SDGIBF
y_{21}, y_{32}	.5414	.5413
y_{15}, y_{21}	.5797	.5771
y_{12}, y_{21}	.6051	.6017
y_{14}, y_{21}	.6165	.6093
y_{13}, y_{21}	.6192	.6180
y_{11}, y_{21}	.6208	.6197
y_{32}, y_{33}	.6699	.6308
y_{32}, y_{35}	.7102	.7038
Observ.'s	SDAIBF	SDGIBF
y_{32}, y_{53}	.7401	.7162
y_{32}, y_{61}	.8037	.7919
y_{13}, y_{32}	.8086	.7971
...	...	...
y_{21}, y_{22}	.8688	.8536
y_{15}, y_{35}	.8697	.8585
y_{32}, y_{65}	.8727	.8544
y_{32}, y_{62}	.8731	.8607
y_{24}, y_{32}	.8761	.8488
y_{32}, y_{63}	.8795	.8655
y_{32}, y_{55}	.8808	.8464
y_{32}, y_{41}	.8903	.8473
y_{14}, y_{32}	.8927	.8826
y_{12}, y_{32}	.8938	.8892
y_{32}, y_{64}	.8963	.8827
y_{21}, y_{45}	.8987	.8640
y_{22}, y_{33}	.9139	.8287
...	...	...
y_{13}, y_{22}	.9997	.9754
y_{12}, y_{35}	1.0002	.9868
...	...	...

표 6.3: Dyestuff 자료에서 크기 2의 이상점군 검출을 위한 SDAIBF와 SDGIBF

Observ.'s	SDAIBF	SDGIBF
y_{13}, y_{53}	.8403	.8340
y_{12}, y_{53}	.8403	.8340
y_{12}, y_{13}	.8403	.8371
y_{13}, y_{42}	.8558	.8475
y_{12}, y_{42}	.8558	.8475
y_{12}, y_{41}	.8581	.8485
y_{13}, y_{41}	.8581	.8485
y_{42}, y_{13}	.8664	.8581
y_{42}, y_{53}	.8688	.8615
y_{41}, y_{13}	.8834	.8665
y_{41}, y_{12}	.8834	.8665
y_{41}, y_{53}	.8864	.8700
Observ.'s	SDAIBF	SDGIBF
...	...	...
y_{12}, y_{44}	.8826	.8743
y_{13}, y_{44}	.8826	.8743
y_{13}, y_{34}	.8846	.8772
y_{12}, y_{34}	.8846	.8772
y_{53}, y_{44}	.8869	.8829
y_{41}, y_{42}	.8975	.8834
...	...	...
y_{13}, y_{62}	.9083	.8975
y_{12}, y_{62}	.9083	.8975
y_{42}, y_{44}	.9031	.8993
y_{53}, y_{25}	.9067	.9017
y_{41}, y_{64}	.9991	.9780
y_{53}, y_{14}	.9863	.9817
y_{42}, y_{35}	.9963	.9848
y_{41}, y_{32}	1.0144	.9871
y_{44}, y_{25}	.9982	.9881
y_{44}, y_{63}	1.0083	.9906
...	...	...

참고문헌

[1] Berger, J.O. and Pericchi, L.R. (1996). The Intrinsic Bayes Factor for Model Selection and Prediction. *Journal of the American Statistical Association,* **96,** No. 433, 109-122.

[2] Box, G.E.P. and Tiao, G.C. (1968). A Bayesian Approach to Some Outlier Problems. *Biometrika,* **55,** 119-129.

[3] Box, G.E.P. and Tiao, G.C. (1973). *Bayesian Inference in Statistical Analysis,* Addison-Wesley Publishing Co., U.S.A.

[4] Chaloner, K. and Brant, R. (1988). A Bayesian Approach to Outlier Detection and Residual Analysis. *Biometrika,* **75,** 651-659.

[5] Dickey, J. (1971). The Weighted Likelihood Ratio Linear Hypotheses on Normal Location Parameters. *the Annals of Mathematical Statistics,* **42,** 204-223.

[6] Dickey, J. (1976). Approximate Posterior Distributions. *Journal of the American Statistical Association,* **71,** 680-689.

[7] Geisser, S. (1985). On the Predicting of Observables: a Selective Update. *Bayesian Statistics 2,* Ed. Bernardo, J.M., DeGroot, M.H., Lindley, D.V. and Smith, A.F.M., 203-230, Amsterdam: North Holland.

[8] Guttman, I. (1973). Care and Handling of Univariate or Multivariate Outliers in Detecting Spuriosity - A Bayesian Approach. *Technometrics,* **15,** 4, 723-738.

[9] Guttman, I. and Pena, D. (1993). A Bayesian Look at Diagnostics in the Univariate Linear Model. *Statistical Sinica,* **3**, 367-390.

[10] Johnson, W. and Geisser, S. (1983). A Predictive View of the Detection and Characterization of Influential Observations in Regression Analysis. *Journal of the American Statistical Association,* **78,** 137-144.

[11] 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.

[12] O'Hagan, A. (1995). Fractional Bayes Factors for Model Comparison. *Journal of the Royal Statistical Society,* **B, 57,** 99-138.

[13] Pettit, L.I. and Smith, A.F.M. (1985). Outliers and Influential Observations in Linear Models. *Bayesian Statistics 2,* Ed. Bernardo, J.M., DeGroot, M.H., Lindley, D.V. and Smith, A.F.M., 473-494, Amsterdam: Elsevier.

[14] Sharples, L.D. (1990) Identification and Accommodation of Outliers in General Hierarchical Models. *Biometrika,* **77,** 3, 445-453.

[15] Tiao, G.C. and Tan, W.Y. (1966). Bayesian Analysis of Random Effect Models in the Analysis of Variance. II. Effect of Autocorrelated Errors *Biometrika,* **53,** 477.

[16] Verdinelli, I. and Wasserman, L. (1995). Computing Bayes Factors Using a Generalization of Savage-Dickey Density Ratio. *Journal of the American Statistical Association,* **90,** 614-618.

[1999년 8월 접수, 2000년 1월 채택]

A Bayesian Outlier Detection in Random Effects Model

Younshik Chung [1] Sangjeen Lee [2]

ABSTRACT

When no information is available and hence improper noninformative priors should be used, Bayes factor for model selection includes the unspecified constants and can not be calibrated. To solve this problems of Bayes factor, we use the intrinsic Bayes factor (IBF; Berger and Pericchi, 1996) and modify it with the generalized Savage-Dickey density ratio (Verdinelli and Wasserman, 1995) to reduce the computational burden. These modified IBF's are applied to detecting outlier in random effects model with a mean-shift structure. We propose the detecting procedures for multiple outlier as well as the single outlier. Our proposed methods are examplified by a simulation experiment with a hypothetical data set including an outlier and also analysis of a real data set.

Keywords: Intrinsic Bayes factor; Mean-shift model; Metropolis algorithm; Outliers; Random effects model.

1) Associate professor, Department of Statistics, Pusan National University.
E-mail: yschung@hyowon.pusan.ac.kr
2) Fulltime Lecturer, Division of Computer and Information, Ulsan College.