The Korean Communications
in Statistics Vol. 7, No. 1, 2000
pp. 141-150

Outlier Detection in Random Effects Model Using Fractional Bayes Factor

Younshik Chung[1] and Sangjeen Lee[2]

Abstract

In this paper, we propose a method of computing Bayes factor to detect an outlier in a random effects model. When no information is available and hence improper noninformative priors should be used, Bayes factor includes the unspecified constants and has complicated computational burden. To solve this problem, we use the fractional Bayes factor (FBF) of O'Hagan (1995) and the generalized Savage-Dickey density ratio of Verdinelli and Wasserman (1995). The proposed method is applied to outlier detection problem. We perform a simulation of the proposed approach with a simulated data set including an outlier and also analyze a real data set.

1. 서론

베이지안 이상점 검출 방법은 이상점을 위한 대립모형을 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 평균-이동모형(mean-shift model)과 분산팽창모형 (variance-inflation model)이 주로 사용된다. 평균이 μ이고 분산이 σ^2인 정규모집단으로부터 자료 $\boldsymbol{y}$를 추출하였다 하자. 이때, 평균이동모형은 이상점 y_i가 $N(\mu+m_i, \sigma^2)$분포를 따른다고 가정하는 것이고, 분산팽창모형은 이상점이 $N(\mu, b_i\sigma^2)$분포로부터 추출되었다고 생각하는 것이다, 이때, $m_i \neq 0$이고 $b_i \gg 1$이다. Guttman(1973)은 평균이동모형을 선형모형에 적용했고, Sharples (1990)은 분산팽창모형이 일반계층적모형에 얼마나 쉽게 적용이 가능한가를 보였다.

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

$$y_{ij}=\mu+e_i+\varepsilon_{ij},\ i=1,\cdots,I,\ j=1,\cdots,J, \tag{1.1}$$

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,

1) Associate Professor, Department of Statistics, Pusan National University, Pusan, 609-735, Korea.
2) Full-time Lecturer, Division of Computer and Information, Ulsan College, Ulsan, 682-090, Korea.

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

에서 나온 이상점으로 의심이 된다고 가정하자. 이때, m은 관측치 y_{ks}의 평균이동모수이다. 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 사용되어 베이즈인자를 계산할 때, 불완전사전분포가 포함하고 있는 미지의 상수가 계산된 베이즈인자에 남아있는 문제점이 있다. 이를 극복하기위해, O'Hagan(1995)이 제시한 부분베이즈인자(the fractional Bayes factor; FBF)를 사용할 것이다. $f_i(\boldsymbol{Y}|\boldsymbol{\theta}_i)$와 $\pi_i^N(\boldsymbol{\theta}_i)$를 각각 가설 H_i하에서의 우도함수와 주어진 불완전사전분포라 하면, $i = 0, 1$, H_0를 선호하는 FBF는

$$B_{01}^r(\boldsymbol{Y}) = \frac{q_0(r, \boldsymbol{Y})}{q_1(r, \boldsymbol{Y})}, \tag{1.3}$$

로 표현된다. 여기서, $i = 0, 1$에 대하여 $q_i(r, \boldsymbol{Y}) = \dfrac{\int \pi_i(\boldsymbol{\theta}_i) f_i(\boldsymbol{Y}|\boldsymbol{\theta}_i)\, d\boldsymbol{\theta}_i}{\int \pi_i(\boldsymbol{\theta}_i) f_i^r(\boldsymbol{Y}|\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i}$이고 r은 부분근사계수(fractional approximation coefficient)이다. 하지만 아직도 이 베이즈인자들의 계산적 어려움은 해결되지 않았다. 그러므로 본 이 논문에서는 이 FBF를 다음의 보조정리 1.1의 일반화 Savage-Dickey 밀도비를 이용하여 계산량을 줄이고 이상점 검출을 위한 검정에 적용하고자 한다. Dickey(1971)는 단순가설 $H_0: m = m_0$와 $H_1: m \neq m_0$의 검정에서 Dickey의 조건,

$$\pi_1^N(\xi|m) = \pi_0^N(\xi)$$

이 만족되면 베이즈인자는 $B_{01} = \pi_1^N(m_0|\boldsymbol{Y}) / \pi_1^N(m_0)$로 표현됨을 보였고 이것을 Savage-Dickey 밀도비라 불렀다, 여기서 ξ는 장애모수벡터이다.

Verdinelli and Wasserman (1995)는 이 밀도비를 Dickey의 조건이 만족되지 않는 상황에서도 적용할 수 있도록 다음 보조정리와 같이 일반화 시켰다.

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

$$B_{01} = \frac{\pi_1^N(m_0|\boldsymbol{Y})}{\pi_1^N(m_0)} \cdot E^{\pi_1^N(\xi|m_0,\ \boldsymbol{Y})}\left[\frac{\pi_0^N(\xi)}{\pi_1^N(\xi|m_0)}\right] \tag{1.4}$$

와 같이 표현된다. 여기서, $E^{\pi_1^N(\xi|m_0,\ \boldsymbol{Y})}$는 확률분포 $\pi_1^N(\xi|m_0, \boldsymbol{Y})$에 대한 기대값을 의미한다. 이것을 일반화 Savage-Dickey 밀도비라 말한다. 만약 Dickey의 조건이 만족되면 식 (1.4)에 있는 기대값부분은 없어진다.

2절에서는 이상점 검출에 대한 변량모형에서의 베이지안적 접근을 다루고 3절에서는 이상점 검출을 위한 FBF를 일반화 Savage-Dickey 밀도비를 이용하여 계산량을 줄이는 수정을 할 것이다. 4절에서는 몇 가지 자료에 제시한 방법을 적용하여 계산하고 그 성능에 대해 논할 것이다.

2. 변량모형의 베이지안 접근

$\boldsymbol{Y}=\{y_{ij},\ i=1,\cdots,I,\ j=1,\cdots,J\}$를 모형 (1.1)과 (1.2)로부터 나온 관측치행렬이라 하자. 한 관측치 y_{ks}가 이상점인가 아닌가를 판단하기 위해서 귀무가설 H_0: "$\boldsymbol{Y}$에는 이상점이 없다."와 대립가설 H_1: "y_{ks}가 이상점이다." 를 검정해야 된다. 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 $\phi=J\sigma_e^2/\sigma^2$를 정의하고 모수벡터를 $\boldsymbol{\theta}=(\mu,\ \sigma^2,\ \phi)$로 한다. H_0하에서의 우도함수는

$$L_0(\mu,\ \sigma^2,\ \phi)\propto\sigma^{-IJ}(1+\phi)^{-I/2}\exp\left\{-\frac{1}{2\sigma^2}\left(\frac{S_1^2+IJ(\overline{y_{..}}-\mu)^2}{1+\phi}+S_2^2\right)\right\} \tag{2.2}$$

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

다음으로, H_1하에서의 우도함수를 구해보자. H_1하에서, 자료 $\boldsymbol{Y}$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. $\overline{y_{m..}}$, $\overline{y_{mk.}}$, S_{m1}^2 그리고 S_{m2}^2를 각각 $\boldsymbol{Y}$ 대신에 식(2.3)에 있는 자료로 계산된 $\overline{y_{..}}$, $\overline{y_{k.}}$, S_1^2 그리고 S_2^2라 하자. 그러면 $\overline{y_{m..}}=\overline{y_{..}}-m/IJ$, $\overline{y_{mk.}}=\overline{y_{k.}}-m/J$, $S_{m1}^2=S_1^2-2(\overline{y_{k.}}-\overline{y_{..}})m+\frac{I-1}{IJ}m^2$이고 $S_{m2}^2=S_2^2-2(y_{ks}-\overline{y_{k.}})m+\frac{I-1}{J}m^2$로 표현된다. 그러므로 H_1하에서의 우도함수는

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

로 쓸 수 있다.

모든 모수에 대한 사전정보가 전혀 없다고 가정하자. 이때, 무정보사전분포가 귀무가설과 대립가설 모두에 대해 사용되어지며, Tiao와 Tan(1966) 그리고 Box와 Tiao(1992)의 무정보사전분포

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

를 귀무가설 하에서의 사전분포로 사용할 수 있다. 평균이동모수 m은 위치모수이고 $\{\mu,\ \sigma^2,\ \phi\}$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.

식 (2.1)의 귀무가설 H_0와 대립가설 H_1하에서 무정보사전분포 (2.5)와 (2.6)을 각각 사용하여 H_0를 선호하는 FBF를 구하면

$$B_{01}^{r}(\boldsymbol{Y}) = \frac{q_0(r, \boldsymbol{Y})}{q_1(r, \boldsymbol{Y})} \tag{2.7}$$

로 계산된다. 여기서

$$q_0(r, \boldsymbol{Y}) = \frac{\int\int\int \pi_0^N(\mu, \sigma^2, \phi) L_0(\mu, \sigma^2, \phi; \boldsymbol{Y}) d\mu d\sigma^2 d\phi}{\int\int\int \pi_0^N(\mu, \sigma^2, \phi) L_0^r(\mu, \sigma^2, \phi; \boldsymbol{Y}) d\mu d\sigma^2 d\phi},$$

$$q_1(r, \boldsymbol{Y}) = \frac{\int\int\int\int \pi_1^N(\mu, m, \sigma^2, \phi) L_1(\mu, m, \sigma^2, \phi; \boldsymbol{Y}) d\mu dm d\sigma^2 d\phi}{\int\int\int\int \pi_1^N(\mu, m, \sigma^2, \phi) L_1^r(\mu, m, \sigma^2, \phi; \boldsymbol{Y}) d\mu dm d\sigma^2 d\phi},$$

이다.

부적절사전분포로 부터 생긴 베이즈인자에 포함되어있는 미지의 상수들은 식 (2.7)에서 상쇄된다. 그러나 FBF에 들어있는 m에 대한 적분이 매우 어렵고 또한 계산량의 부담도 매우 크다고 생각된다. 그러므로 우리는 이 FBF에 일반화 Savage-Dickey 밀도비를 적용시켜 계산량을 줄이고 중요표본계산법 (importance sampling method)을 이용한 적분을 다음 장에서 할 것이다.

3. 이상점 검출을 위한 FBF의 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)벡터이며 이를 ξ라 하자. 식 (2.1)에 있는 단순가설을 베이지안 검정하기 위해 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_1하의 사전확률분포로 사용한다. 계산량의 부담을 줄이기 위해, 식 (2.7)에서 구한 O'Hagan(1995)의 FBF를 보조정리 1.1의 일반화 Savage-Dickey 밀도비 개념을 이용하여 다음과 같이 표현할 수 있다.

보조정리 3.1 적당한 근사계수 r에 대하여, FBF는 두 베이즈인자의 곱,

$$B_{01}^{r} = B_{01}^{N} \cdot B_{r10}^{N} \tag{3.1}$$

으로 표현될 수 있다. 여기서, B_{01}^{N}와 B_{r10}^{N}는 우도함수로 각각 $f(\boldsymbol{Y}|m, \xi)$와 $f^r(\boldsymbol{Y}|m, \xi)$를 사용하고 주어진 무정보사전분포를 이용하여 얻어진 베이즈인자들이다.

증명. $m_i(\boldsymbol{Y})$와 $m_{i,r}(\boldsymbol{Y})$를 각각 H_i하에서 우도함수 $f(\boldsymbol{Y}|m, \xi)$와 $f^r(\boldsymbol{Y}|m, \xi)$를 사용하여 구한 자료 $\boldsymbol{Y}$의 주변확률밀도함수라 하자, $i = 0, 1$. 그러면

$$\begin{aligned} B_{01}^{r} &= \frac{q_0(r, \boldsymbol{Y})}{q_1(r, \boldsymbol{Y})} \\ &= \frac{\int \pi_0^N(\xi) f(\boldsymbol{Y}|m_0, \xi) d\xi}{\int \pi_0^N(\xi) f^r(\boldsymbol{Y}|m_0, \xi) d\xi} \Big/ \frac{\int\int \pi_1^N(m, \xi) f(\boldsymbol{Y}|m, \xi) dm d\xi}{\int\int \pi_1^N(m, \xi) f^r(\boldsymbol{Y}|m, \xi) dm d\xi} \end{aligned}$$

$$= \frac{\int \pi_0^N(\xi) f(\boldsymbol{Y}|m_0, \xi)d\xi}{\int\int \pi_1^N(m, \xi) f(\boldsymbol{Y}|m, \xi)dmd\xi} \cdot \frac{\int\int \pi_1^N(m, \xi) f^r(\boldsymbol{Y}|m, \xi)dmd\xi}{\int \pi_0^N(\xi) f^r(\boldsymbol{Y}|m_0, \xi)d\xi}$$
$$= \frac{m_0(\boldsymbol{Y})}{m_1(\boldsymbol{Y})} \cdot \frac{m_{1,r}(\boldsymbol{Y})}{m_{0,r}(\boldsymbol{Y})} = B_{01}^N \cdot B_{r10}^N.$$

여기서 사용된 기호들을 설명해 보자. $m_i(\boldsymbol{Y})$와 $m_{i,r}(\boldsymbol{Y})$는 H_i하에서 우도함수로 각각 $f(\boldsymbol{Y}|m, \xi)$와 $f^r(\boldsymbol{Y}|m, \xi)$를 사용한 $\boldsymbol{Y}$의 주변밀도함수들이다, $i=0, 1$. 만약 $r=1$이면, $m_{i,r}(\boldsymbol{Y}) = m_i(\boldsymbol{Y})$, $i=0, 1$이므로 $B_{10}^N = B_{r10}^N$이 된다. 또한, $c(r, m, \xi) = \int f^r(\boldsymbol{Y}|m, \xi)d\boldsymbol{Y}$라 하고 $c(r, \xi) = c(r, m_0, \xi)$라 하면, H_0과 H_1하에서 자료 $\boldsymbol{Y}$의 주변밀도함수는 각각 $m_{0,r}(\boldsymbol{Y}) = \int \frac{\pi_0^N(\xi)}{c(r,\xi)} \cdot c(r,\xi) f^r(\boldsymbol{Y}|m_0, \xi)d\xi$ 와 $m_{1,r}(\boldsymbol{Y}) = \int\int \frac{\pi_1^N(m,\xi)}{c(r,m,\xi)} \cdot c(r,m,\xi) f^r(\boldsymbol{Y}|m,\xi)dmd\xi$로 표현된다. 이때, $c(r, \xi)\times f^r(\boldsymbol{Y}|m_0, \xi)$와 $c(r, m, \xi)\times f^r(\boldsymbol{Y}|m, \xi)$를 각각 H_0과 H_1하에서의 우도함수로 생각하고 $\frac{\pi_0^N(\xi)}{c(r,\xi)}$와 $\frac{\pi_1^N(m,\xi)}{c(r,m,\xi)}$는 각각 H_0과 H_1하에서의 ξ와 (m, ξ)의 사전밀도함수로 보면, B_{r10}^N도 하나의 베이즈인자로 취급할 수 있다. 또한 B_{01}^N도 하나의 단순검정을 위한 베이즈인자이므로, 우리는 Savage-Dickey 밀도비 개념을 적용할 수 있다. 그러나 실제로는 $c_1(r, m, \xi)$와 $c(r, \xi)$를 계산할 필요는 없다.

정리 3.1 적당한 근사계수 r에 대해, 가설 $H_0: m = m_0$를 선호하는 FBF는

$$B_{01}^r = \frac{\pi_1^N(m_0|\boldsymbol{Y})}{\pi_{1,r}^N(m_0|\boldsymbol{Y})} \cdot \frac{E^{\pi_1^N(\xi|m_0,\boldsymbol{Y})}[\pi_0^N(\xi)/\pi_1^N(\xi|m_0)]}{E^{\pi_{1,r}^N(\xi|m_0,\boldsymbol{Y})}[\pi_0^N(\xi)/\pi_1^N(\xi|m_0)]} \tag{3.2}$$

로 계산된다. 여기서, $\pi_{1,r}^N(\xi|m_0, \boldsymbol{Y}) = \pi_{1,r}^N(m_0, \xi|\boldsymbol{Y})/\pi_{1,r}^N(m_0|\boldsymbol{Y})$, $\pi_{1,r}^N(m_0|\boldsymbol{Y}) = \int \pi_{1,r}^N(m_0, \xi|\boldsymbol{Y})d\xi$, $\pi_{1,r}^N(m_0, \xi|\boldsymbol{Y}) = \pi_1^N(m_0, \xi) f^r(\boldsymbol{Y}|m_0, \xi)/m_{1,r}(\boldsymbol{Y})$이고 $E^{g(\xi)}(\cdot)$는 확률밀도함수 $g(\xi)$에 대한 기대값을 뜻한다.

증명. 보조정리 3.1로부터 FBF는 두 베이즈인자의 곱으로 표현될 수 있다. 또한, 이들은 각각 보조정리 1.1에 의하여

$$B_{01}^N = \frac{\pi_1^N(m_0|\boldsymbol{Y})}{\pi_1^N(m_0)} \cdot E\left[\frac{\pi_0^N(\xi)}{\pi_1^N(\xi|m_0)}\right]$$

와

$$B_{r01}^N = 1/B_{r10}^N = \frac{m_{0,r}(\boldsymbol{Y})}{m_{1,r}(\boldsymbol{Y})} = \frac{\int \pi_0^N(\xi) f^r(\boldsymbol{Y}|m_0, \xi)d\xi}{m_{1,r}(\boldsymbol{Y})}$$
$$= \pi_{1,r}^N(m_0|\boldsymbol{Y}) \cdot \int \frac{\pi_0^N(\xi) f^r(\boldsymbol{Y}|m_0, \xi)\pi_{1,r}^N(\xi|m_0, \boldsymbol{Y})}{m_{1,r}(\boldsymbol{Y})\pi_{1,r}^N(m_0|\boldsymbol{Y})\pi_{1,r}^N(\xi|m_0, \boldsymbol{Y})} d\xi$$

$$= \pi^N_{1,r}(m_0 \mid \boldsymbol{Y}) \cdot \int \frac{\pi^N_0(\xi)\pi^N_{1,r}(\xi \mid m_0, \boldsymbol{Y})}{\pi^N_1(m_0, \xi)} d\xi$$

$$= \frac{\pi^N_{1,r}(m_0 \mid \boldsymbol{Y})}{\pi^N_1(m_0)} \cdot E^{\pi^N_{1,r}(\xi \mid m_0,\ Y)}\left[\frac{\pi^N_0(\xi)}{\pi^N_1(\xi \mid m_0)}\right]$$

로 표현된다. 보조정리 3.1로부터 $B^r_{01} = B^N_{01}/B^N_{r01}$이므로 증명은 완성된다.

FBF의 이 형태를 Savage-Dickey FBF(SDFBF)라 부르자. 식 (3.2) 안에 있는 기대값의 분모와 분자에 이들과 독립인 $\pi^N_1(m_0)$을 곱해도 전체는 아무런 변화가 없으므로 식 (3.2)에서 $\pi^N_1(\xi \mid m_0)$ 대신에 $\pi^N_1(m_0, \xi)$를 사용할 수 있고 또 이 형태가 실제 계산에서 훨씬 편리하다.

따름정리 3.1 Dickey의 조건이 만족되면 식 (3.2)의 SDFBF는 기대값 항들이 없어진다.

이제, 정리 3.1을 이용하여 변량모형에서 이상점 검출을 위한 SDFBF를 계산해 보자. 우선, 한 관측치 y_{ks}의 평균이동모수 m의 사후주변확률밀도함수를 계산하자.

보조정리 3.2 식 (2.6)에 있는 확률함수를 가설 H_1하의 사전확률함수로 사용하고 식 (2.4)에 있는 $\{L_1(\mu, m, \sigma^2, \phi)\}^r$을 우도함수로 사용하여 구한 m의 사후주변확률함수는

$$\pi_{1,r}(m \mid \boldsymbol{Y}) = C_r \beta_{p_r, q_r}\left(\frac{W_m}{W_m+1}\right)(S^2_{m2})^{-(p_r+q_r)} W_m^{-p_r}, \quad 0 < r \le 1 \tag{3.3}$$

로 계산된다. 여기서,

$$C_r^{-1} = \int_{-\infty}^{\infty} \beta_{p_r, q_r}\left(\frac{W_m}{W_m+1}\right) W_m^{-p_r} S_{m2}^{2(-p_r-q_r)} dm, \quad 0 < r \le 1 \tag{3.4}$$

이고 S^2_{m1}와 S^2_{m2}는 식 (2.4)에 정의되어있고 $W_m = S^2_{m1}/S^2_{m2}$, $p_r = (rI-1)/2$, $q_r = rI(J-1)/2$이고 $\beta_{i,j}(x) = \int_0^x t^{i-1}(1-t)^{j-1} dt$, $0 < r \le 1$, 이다.

증명. (m, μ, σ^2, ϕ)의 결합사후확률함수는

$$\pi_{1,r}(m, \mu, \sigma^2, \phi \mid y_{\cdot\cdot}, s_1^2, s_2^2) \propto \sigma^{-(rIJ+2)}(1+\phi)^{-(rI+2)/2}$$

$$\cdot \exp\left[-\frac{r}{2\sigma^2}\left\{\frac{S^2_{m1} + IJ(y_{m\cdot\cdot} - \mu)^2}{1+\phi} + S^2_{m2}\right\}\right]$$

이다. 그래서 m의 주변확률함수는 다음과 같이 적분으로 구할 수 있다.

$$\pi_{1,r}(m \mid \boldsymbol{Y}) \propto \int_0^\infty \int_0^\infty \int_{-\infty}^\infty \pi_{1,r}(\mu, m, \sigma^2, \phi \mid \boldsymbol{Y}) d\mu d\sigma^2 d\phi$$

$$\propto \int_0^\infty \int_0^\infty \sigma^{-(rIJ+2)}(1+\phi)^{-(rI+2)/2}$$

$$\cdot \exp\left\{-\frac{r}{2\sigma^2}(S^2_{m2}+\frac{S^2_{m1}}{1+\phi})\right\}\cdot\{\frac{2\pi}{rIJ}\sigma^2(1+\phi)\}^{1/2}d\sigma^2 d\phi$$

$$\propto \int_0^\infty (1+\phi)^{-(rI+1)/2}\int(\sigma^2)^{-(rIJ+1)/2}\exp\{-\frac{r}{2\sigma^2}(S^2_{m2}+\frac{S^2_{m1}}{1+\phi})\}d\sigma^2 d\phi$$

$$\propto \int_0^\infty (1+\phi)^{-(rI+1)/2}(S^2_{m2}+\frac{S^2_{m1}}{1+\phi})^{-(rIJ-1)/2}d\phi\ .$$

여기서, $Z_m = W_m/(W_m+1+\phi)$라 하면

$$\pi_{1,r}(m\mid \boldsymbol{Y}) \propto \int_0^\infty (1+\phi)^{-(rI+1)/2}(1+\frac{W_m}{1+\phi})^{-(rIJ-1)/2}S_{m2}^{-(rIJ-1)}d\phi$$

$$\propto (S^2_{m2})^{-(p_r+q_r)}W_m^{-p_r}\int_0^{\frac{W_m}{W_m+1}} Z_m^{p_r-1}(1-Z_m)^{q_r-1}dZ_m$$

$$= (S^2_{m2})^{-(p_r+q_r)}W_m^{-p_r}\beta_{p_r,q_r}(\frac{W_m}{W_m+1}).$$

식 (3.3)의 S^2_{m1}과 S^2_{m2}가 양수이고 유한하므로 W_m도 양수이고 유한하다. 그러므로 불완전베타함수 $\beta_{p_r,q_r}(\frac{W_m}{W_m+1})$도 양수이고 유한하다. 그래서 C_r도 양수이고 유한하다. 이 성질들로부터 함수 $\pi_{1,r}(m\mid \boldsymbol{Y})$가 확률밀도함수의 공리를 만족한다는 것을 쉽게 알 수 있다. 구해진 m의 주변 사전분포함수를 사용하여 정리 3.1에 적용하면 변량모형의 이상점검출을 위한 SDFBF를 다음 정리와 같이 구할 수 있다.

정리 3.2 식 (2.1)의 H_0를 선호하는 SDFBF는

$$B^r_{01}(Y) = \frac{C_1}{C_r}\frac{\beta_{p,q}(\frac{W}{W+1})}{\beta_{p_r,q_r}(\frac{W}{W+1})}W^{-p+p_r}S_2^{2(p_r-p+q_r-q)} \tag{3.5}$$

로 계산된다. 여기서 C_r^{-1}와 p_r, q_r는 식 (3.4)와 그 아래에 정의되어 있고 $r=1$일 때의 값들을 각각 C_1^{-1}와 p, q로 표현했다. W_m와 S^2_{m1}, S^2_{m2}, $\beta_{i,j}(x)$도 식 (3.4)의 아래에 정의되어 있다.

증명. 식 (2.6)으로부터 $\pi_1^N(m)=1$ 이고 $\pi_1^N(\mu,\sigma^2,\phi\mid m)=\pi_1^N(\mu,m,\sigma^2,\phi)/\pi_1^N(m)$ 이므로 $\pi_1^N(\mu,\sigma^2,\phi\mid m)=\pi_0^N(\mu,\sigma^2,\phi)$ 이다. 즉, 사전분포함수 (2.5)와 (2.6)은 Dickey의 조건이 만족된다. 그러므로 따름정리 3.1에 의하여 $B^{r_{01}}=\frac{\pi_1^N(m_0\mid Y)}{\pi_{1,r}^N(m_0\mid Y)}$ 이 된다. 또한, 보조정리 3.2에서 m의 주변사후확률밀도함수 $\pi_{1,r}(m\mid \boldsymbol{Y})$를 구했고 $\pi_1(m\mid \boldsymbol{Y})$는 $\pi_{1,r}(m\mid \boldsymbol{Y})$이 $r=1$일 때이므로 식 (3.5)는 바로 구해진다.

식 (3.5)에 있는 상수 C_r은 해석학적으로 구하기는 불가능하다. 그러나 중요표본계산법(importance sampling method)과 같은 표본추출계산법(sampling based computation)을 사용한 수치적 방법으로 이들

을 추정할 수 있다.

$$g_r(m) = \beta_{p_r,q_r}\left(\frac{W_m}{W_m+1}\right) W_m^{-p_r}(S_{m2}^2)^{-(p_r+q_r)}$$

라 하면 상수 C_r은

$$C_r^{-1} = \int_{-\infty}^{\infty} g_r(m)\,dm = \int_{-\infty}^{\infty} \frac{g_r(m)}{I(m)} I(m)dm = E\left\{\frac{g_r(m)}{I(m)}\right\}$$

와 같이 계산할 수 있다. 여기서 $I(m)$는 $g_r(m)$와 같은 정의구역을 가지는 확률밀도함수이다. $I(m)$으로부터 표본 $\{m^{(1)}, \cdots, m^{(G)}\}$을 생성시킨 후, 다음과 같은 Monte Carlo 방법으로 C_r^{-1}의 값을 추정할 수 있다 ;

$$\hat{C}_r^{-1} = \frac{1}{G}\sum_{g=1}^{G} \frac{g_r(m^{(g)})}{I(m^{(g)})} . \tag{3.6}$$

이 계산법에서는 함수 $I(m)$의 선택이 매우 중요하다. 이 논문에서는 메트로폴리스 표본추출법 (Metropolis 외4명, 1953)의 방법으로 원래의 함수 $g_r(m)$으로부터 추출한 표본의 표본평균 $\overline{m}$와 표본분산 s_m^2을 평균과 분산으로 가지는 정규밀도함수를 중요함수 (importance function) $I(m)$으로 사용하였다.

4. 예제들

4.1 생성자료

이 절은 앞 절에서 제시한 SDFBF를 이용한 이상점 검출 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 변량모형 (1.1)로부터 생성된 자료이다. 이 자료의 한 관측치 y_{52}는 위와 같은 분산들을 가지며 평균이 $\mu=0$인 변량모형 (1.1)로부터 생성된 관측치이다. 즉, $m=-5$인 모형 (1.2)로부터 생성하였다.

모든 관측치 각각이 이상점인가 아닌가에 대한 SDFBF를 계산하였다. 근사계수 r을 구하기 위해 Berger와 Pericchi (1996)의 최소훈련표본(minimal training sample; MTS)을 구해보면 자료의 H_0와 H_1하의 주변분포함수 가 모두 유한하기위한 표본의 크기는 $I=2$ 그리고 $J=2$이다. 즉, MTS의 크기가 4가 된다. 그래서 O'Hagan(1995)의 한 방법 $r = Max\{4, \sqrt{30}\}/30 = \sqrt{30}/30$을 사용하여 SDFBF를 구한다. 계산된 각각의 결과 값을 표4.2에 나열하였다.

결과를 살펴보면 이상점인 y_{52}가 1보다 훨씬 작고 또 가장 작은 값을 가진다. 그래서 이상점 검출을 위한 식 (3.5)의 SDFBF는 하나의 좋은 방법이라는 것을 알 수 있다. 그러나 5번째 배치의 관측치들이 모두 비교적 작은 SDFBF값들을 가지는 것을 보면 이 관측치들 모두가 중심으로부터 약간 떨어져 있다고 생각되어진다. 즉, 5번째 배치 확률인자 e_5가 비교적 큰 값을 가지거나 아니면 작은 값을 가지는 것을 알 수 있다. 이것은 표4.1에 있는 실제 자료에서 5번째 관측치들이 다른 관측치들에 비해 모두 비교적 작은 값을 가지는 것과 일치하므로 제시한 방법의 분별력이 뛰어남을 알 수 있다.

표4.1 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 생성자료에서 이상점 검출을 위한 SDFBF 값들

s k	1	2	3	4	5	6
1	1.0673	0.6704	0.9292	0.8395	0.9109	1.0651
2	0.8869	1.0207	1.1297	0.9524	0.5180	0.8811
3	0.7738	0.8223	1.0103	0.9887	0.9223	0.8619
4	0.8672	0.9907	0.8910	0.9273	0.9014	1.0610
5	0.8770	0.9463	1.0602	0.7651	0.9191	0.9805

4.2 실제자료

이번에는 실제 자료에 제시된 방법을 적용해보자. 사용된 실제 자료는 Box와 Tiao (1973)의 Dyestuff 자료로써 여섯 가지 종류의 생산에 각각 크기 5의 표본을 추출하여 표준색감의 생산량을 그람으로 나타낸 것이다. 이 자료는 표4.3에 나타내었다.

표4.3 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

생성된 자료와 마찬가지로 $\sqrt{n}/n = \sqrt{30}/30$을 근사계수 r로 사용하여 모든 관측값이 이상점인가를 제시된 방법을 적용하여 각각 계산하였다. 즉, 모든 관측치 각각에 대해 식 (3.5)에 있는 SDFBF값들을 계산

표 4.4 Dyestuff 자료에서 이상점 검출을 위한 SDFBF 값들

s k	1	2	3	4	5	6
1	0.9915	0.9970	0.9900	1.0152	0.9966	0.9913
2	1.0178	0.9933	1.0012	1.0165	0.9879	1.0075
3	1.0178	1.0095	0.9875	0.9779	1.0166	1.0088
4	0.9978	0.9921	1.0112	1.0102	0.9867	1.0013
5	0.9829	0.1008	0.9987	0.9903	0.9892	1.0100

하여 표4.4에 나타내었다.

실제자료인 Dyestuff 자료의 결과를 보면 거의 모든 값이 1 근처로 나타난다. 그러므로 이 자료는 이상점이 없는 것으로 판단된다.

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.(1992). *Bayesian Inference in Statistical Analysis*, Wiley Classics Library, John Wiley and son

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

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

[5] 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.

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

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

[8] 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.

[9] 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.

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

[11] 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.

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

[13] 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.

[14] 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.