The Korean Journal of Applied Statistics (2015) DOI: http://dx.doi.org/10.5351/KJAS.2015.28.5.1035
28(5), 1035–1046

Effect of Experimental Layout on Model Selection under Variance Components Models: A Simulation Study

Yonghee Lee[a,1]

[a]Department of Statistics, University of Seoul

(Received September 14, 2015; Revised October 12, 2015; Accepted October 14, 2015)

Abstract

Variance components models incorporate various random factors in the form of linear models. There are two experimental Layouts for the classification of factors under variance components models: nested classification and crossed classification. We consider two-way variance components models and investigate the effect of experimental Layout on the performance of model selection criteria AIC and BIC. The effect of experimental Layout is studied through a simulation study with various combinations of parameters in a systematic fashion. The simulation study shows differences in performance of model selection methods between the two classification. There is a particular tendency to prefer the smaller model than the true model when the variance component of a nested factor becomes relatively larger than a nesting factor that is persistent even when the sample size is not small.

Keywords: variance components models, linear mixed models, nested classification, crossed classification, model selection, AIC, BIC, experimental layout

1. 서론

분산성분모형(Variance components models)은 여러 가지의 임의 요인들(random factors)이 관측값에 미치는 구조를 선형식으로 나타내는 모형이다. 요인의 효과가 고정효과(fixed effects)인 경우 일반선형모형(General linear models)을 사용하여 자료를 분석하지만 요인들 중에 임의효과(random effect)가 있는 경우 고정효과와 임의효과를 동시에 고려하는 선형혼합모형(Linear mixed effects models)을 사용하며 분산성분모형은 선형혼합모형의 특별한 형태이다. 분산성분모형에서 모수를 추정하는 방법으로 전통적인 분산분석법(analysis of variance; ANOVA)이 오랜 기간 널리 사용되었고 근래에 들어서는 컴퓨터를 이용한 계산법의 발달로 인하여 복잡한 모형에서도 모수의 추정이 가능한 최대가능도 추정법(Maximum likelihood estimation) 또는 분산성분을 편이없이 추정하는 제한적 최대가능도 추정법(Restricted maximum likelihood estimation)이 널리 사용된다 (Searle 등, 1992).

일반선형모형에서 유의한 요인 또는 변수를 선택하는 방법은 변수선택법, 분산분석표를 이용한 F-검정법 또는 가능도비 검정법(likelihood ratio test)이 있으며 개선된 검정방법들이 다양하게 제시되어 있다. 또한 정보기준(Information criteria)에 근거한 AIC(Akaike Information Criteria) 또는

[1]Department of Statistics, University of Seoul, 163 Seoulsiripdae-ro, Dongdaemun-gu, Seoul 130-743, Korea. E-mail: ylee@uos.ac.kr

BIC(Bayesian Information Criteria) 등을 사용하여 모형선택이 가능하다. 하지만 임의효과에 대해서는 고정효과를 선택하는 변수선택법(예를 들어 forward selection, stepwise selection)을 그대로 적용하기 힘들다. 예를 들어 다변량 분산성분인 경우 하나의 분산성분이 영이면 해당하는 공분산도 영이 되는데 이러한 구조적 문제를 변수선택법에 구현하기 쉽지 않다. 또한 임의효과에 대한 통계적 가설검정은 일부 균형 배치(balanced design)인 경우에만 제곱합을 이용하는 단순한 F-검정법이 존재하고 불균형 배치인 경우는 F-검정법이 존재하지 않는 경우가 많다. 수정된 F-검정법을 사용하더라도 검정 통계량의 분포가 카이제곱분포의 혼합분포(mixture distributions) 형태로 나타나고 모형의 종류에 따라 혼합의 형태가 틀려지기 때문에 실제 분석에서 이용하는 것이 매우 어렵다. 가능도비 검정(likelihood ratio test)도 마찬가지로 분산성분이 영을 가지는 귀무가설 하에서 검정통계량의 분포가 카이제곱분포를 따르지 않고 그 분포의 형태가 문제마다 달라지기 때문에 일반 카이제곱검정을 이용하면 검정력(power)이 떨어질 수 있다는 것이 알려져 있다 (Stram과 Lee, 1994; Crainiceanu와 Ruppert, 2004; Pinheiro와 Bates, 2006). 더 나아가 F-검정법과 가능도비 검정은 포함 관계가 없는 모형들(non-nested models)을 고려하는 경우 적용할 수 없다. 정보기준에 의한 모형선택법은 자료의 균형성이나 요인의 배치에 관계없이 가능도 함수와 모수의 갯수만 알면 쉽게 적용할 수 있다는 장점이 있어 근래에는 선형혼합모형에서 정보기준을 이용한 모형의 선택법을 이용하려는 경향이 강하며 다양한 정보기준 모형선택법들이 제안되고 있다 (Müller 등, 2013). 다만 정보기준에 의한 모형선택법은 주어진 모형에 대해 잘 설계되어진 특정한 검정법보다 모형선택의 효율이 떨어질 수 있다

선형혼합모형에서는 임의효과가 가지는 구조와 통계적 분포의 특성으로 인하여 모형선택 방법을 설계하는 것이 쉽지 않다. 최근에 AIC를 변형한 방법들이 많이 연구되었으며 대표적인 성과로 임의효과의 유효 자유도(effective degrees of freedom)를 고려한 조건부 AIC(conditional AIC)가 제안되었지만 (Vaida와 Blanchard, 2005; Greven과 Kneib, 2010; Müller 등, 2013) 조건부 AIC는 여러 모의실험 연구를 통하여 자료를 생성하는 참모형(data generating model; true model)보다 과대모형(over-parameterized model)을 선택하는 경향이 심하다는 것이 알려져 있다 (Dimova 등, 2011; Lee, 2015). 베이지안 방법을 이용한 임의효과의 선택법에서 분산성분에 대한 사전분포(prior distribution)을 선택하는 경우 분산성분이 경계값(boundary value)인 영의 값을 가질 수 있도록 사전확률을 주는 여러 가지 방법들도 제안되었지만 (Pauler 등, 1999; Chen과 Dunson, 2003; Saville와 Herring, 2009) 복잡한 형태의 사전확률분포와 사후확률을 계산하는 수치적인 어려움때문에 실제 문제에 쉽게 적용할 수 없는 것이 현실이다. 앞에서 언급한 이유로 인하여 가장 단순한 형태의 AIC와 BIC가 많은 연구에서 선형혼합모형의 모형선택법으로 사용되고 있다.

선형혼합모형에서 정보기준에 근거한 모형선택법, 특히 AIC와 BIC의 효율은 많은 연구들에서 이론적인 성질보다 모의실험에 근거한 경험적 증거로 제시되었지만 대부분의 연구들이 고려한 모형은 임의효과가 하나인 가장 단순한 일원배치모형(one-way classification models)이다. 많은 연구에서 사용한 모형이 반복측정자료(repeated measures)이며 이는 비록 다차원의 임의벡터를 고려하더라도 그룹(group)을 나누는 요인이 하나인 일원배치모형이다. 이렇게 가장 단순한 모형만을 고려한 모의 실험의 한계는 Müller 등 (2013)에 의해 제기되었다. 최근까지 제시된 연구들 중 가장 광범위한 모형들과 선택법들을 모의실험에 의해 비교한 연구는 Dimova 등 (2011)이 수행하였다. 이 연구에서 가장 좋은 성능을 보인 선택법은 BIC와 수정항이 표본의 개수의 제곱근($\sqrt{n}$)에 비례하는 GIC(General Information Criteria)로 나타났다. 하지만 Dimova 등 (2011)이 수행한 모의실험에서도 일원배치모형만을 사용하였다. 최근 조건부 AIC의 성능을 비교한 Lee (2015)는 이원배치모형(two-way classification models)의 형태를 고려하였으며 이 때 모형선택법의 효율이 요인의 배치구조와 분산성분의 크기에 따라 다르게 나타나는 현상을 발견하였다.

본 연구은 이원배치 분산성분모형에서 대표적인 설계구조인 교차배치(crossed classification)과 지분배치(nested classification) 모형을 고려한다. 이러한 모형 하에서 분산 성분들을 선택하는 AIC와 BIC의 성능이 모수의 변화와 요인의 배치 방법에 따라 어떻게 다른지 체계적인 모의실험을 수행하여 그 결과를 제시하고자 한다.

본 논문에서 2장은 두 가지 이원배치모형과 모형선택법인 AIC와 BIC를 소개하고 3장에서는 모의실험을 통하여 요인의 배치 방법과 모수의 변화에 따른 모형선택법의 성능 차이를 보여주고자 한다. 4장에서는 결론과 향후 연구 방향을 제시한다.

2. 이원배치 분산성분 모형과 모형선택법

2.1. 이원배치 분산성분 모형

요인이 두 개인 이원배치 분산성분모형은 크게 교차배치와 지분배치로 나누어 진다. 본 논문에서는 간결한 표현을 위하여 균형배치(balanced classification)만을 고려하였다. 불균형이 심하지 않은 자료에 대한 모형선택법의 성능은 균형배치와 유사하며 불균형의 정도가 심각한 자료는 모수의 추정에서부터 많은 문제가 먼저 발생하므로 선택법의 효율을 비교하는데 적절하지 않다.

첫 번째로 이원 교차배치 모형(two-way crossed classification model)의 구조는 다음과 같다.

$$y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk}, \quad i = 1, 2, \ldots, a,\ j = 1, 2, \ldots, b,\ k = 1, 2, \ldots, n. \tag{2.1}$$

교차배치 모형에서 첫 번째 그룹을 나누는 임의효과 α_i는 두 번째 그룹에 대한 임의효과 β_j와 교차되어있다(crossed). 첫 번째 요인에 대한 임의효과 α_i는 독립적으로 평균이 0이고 분산이 σ_a^2인 정규분포 $N(0, \sigma_a^2)$를 따른다. 두 번째 요인의 임의효과 β_j는 정규분포 $N(0, \sigma_b^2)$을 따른다. 마지막으로 오차항 e_{ijk}는 각 그룹의 조합에서 반복에 대한 오차이며 독립적으로 $N(0, \sigma_e^2)$을 따르며 모든 효과들은 서로 독립이다. 이원교차배치 모형의 예로 반응값이 곡물 생산량이라 하면 첫번째 그룹은 구역(plot), 두 번째 그룹은 사용한 비료의 종류로 생각할 수 있다.

이원 지분배치 모형(two-way nested classification model)의 구조는 다음과 같다.

$$y_{ijk} = \mu + \alpha_i + \beta_{ij} + e_{ijk}, \quad i = 1, 2, \ldots, a,\ j = 1, 2, \ldots, b,\ k = 1, 2, \ldots, n. \tag{2.2}$$

반응값 y_{ijk}는 첫 번째 그룹 i와 내포된 두 번째 그룹 j에서 k번째 관측값을 의미한다. 예를 들어 반응값이 학생의 성적이라 하면 첫 번째 그룹은 학교, 두 번째 그룹은 학교에 속한 학급으로 생각할 수 있다. 각 임의효과의 분포는 모형 (2.1)과 동일하다.

식 (2.1)과 (2.2)에서 분산성분 $\theta = (\sigma_a^2, \sigma_b^2)$의 모수 공간(parameter space)는 0을 포함하는 비음수 실수 공간이다. 즉 분산성분의 추정량은 0이 될 수 있다. 하지만 오차항의 분산 σ_e^2의 모수공간을 0은 제외한 양의 실수 공간이다. 모든 성분이 정규분포를 따르기 때문에 모형 (2.1)과 (2.2)에 대한 가능도 함수(likeihood function) L은 추정에 관계없는 상수를 제외하면 다음과 나타낼 수 있다.

$$\log L\left(\mu, \theta, \sigma_e^2 | y\right) \simeq -\frac{1}{2}\det(V) - \frac{1}{2}(y - \mu 1_N)^t V^{-1}(y - \mu 1_N). \tag{2.3}$$

위의 식에서 1_N은 모든 원소가 1이고 길이가 $N = abn$인 벡터이다. 행렬 V는 관측값 벡터 $y = (y_{111}, y_{112}, \ldots, y_{abn})^t$의 공분산 행렬이며 다음과 같이 나타난다.

$$V = \sigma_e^2 I_N + \sigma_a^2 Z_1^t Z_1 + \sigma_b^2 Z_2^t Z_2.$$

위의 식에서 I_N은 N-차원 항등행렬이고 Z_1과 Z_2는 요인 α와 β의 설계행렬이며 이원지분배치 모형에서는 $Z_1 = I_a \otimes 1_b \otimes 1_n$, $Z_2 = I_a \otimes I_b \otimes 1_n$으로 나타나고 이원교차배치 모형에서는 $Z_1 = I_a \otimes 1_b \otimes 1_n$, $Z_2 = 1_a \otimes I_b \otimes 1_n$로 나타난다. 여기서 $\otimes$는 크로네커곱(Kronecker product)이다. 가능도함수 (2.3)에 기반한 최대가능도추정량은 다양한 제곱합들로 이루어진 가능도 방정식으로 구해지며 분산성분에 대한 불편추정량을 구할 수 있는 제한가능도추정법도 사용할 수 있다. 자세한 추정과정과 결과는 Searle 등 (1992, pp. 146–163)에서 찾을 수 있다.

2.2. AIC와 BIC

AIC(Akaike Information Criterion)는 Akaike (1973)에 의해 제안된 모형선택의 기준으로서 자료를 생성하는 참모형과 자료를 분석하기 위해 고려된 모형의 거리를 나타내는 Kullback-Leibler divergence(K-L divergence)를 반영하는 모형선택의 기준이다. 자료를 생성하는 참모형을 $f_t(y)$라고 하고 분석에서 사용된 모형의 집합을 $\{f_\psi(y)|\psi \in \Psi\}$라고 하면 K-L divergence는 다음과 같이 정의된다

$$\mathrm{KL}(f_t, f_\theta) = E_t(\log f_t) - E_t(\log f_\psi).$$

K-L divergence에서 $E_t(\log f_t)$는 고려된 모형과 상관이 없는 상수이기 때문에 $-E_t(\log f_\psi)$의 값이 모형의 선택에서 추정하고자 하는 양이며 그 값이 작을수록 실제 모형과 가깝다. 따라서 주어진 모형이 실제 모형에 가까운 정도를 나타내는 양으로서 Akaike Information을 다음과 같이 정의한다.

$$\mathrm{AI} = -2E_t(\log f_\psi).$$

모수 ψ는 보통 최대가능도 추정량 $\hat{\psi} = \hat{\psi}(y)$로 추정되며 Akaike Information에 대한 추정값이 AIC이며 이는 가능도함수와 모형의 복잡성이 결합한 형태이다. 선형혼합모형에 대한 AIC 다음과 같이 주어진다.

$$\mathrm{AIC} = -2\log L + 2K = -2\log L\left(\hat{\mu}, \hat{\theta}, \hat{\sigma}_e^2|y\right) + 2K, \tag{2.4}$$

여기서 L은 식 (2.3)에 주어진 가능도함수이며 모형의 복잡성을 나타내는 측도인 K는 주어진 모형에 모수의 수(number pf parameters)와 같다. 본 논문에서 고려하는 분산성분모형에서 모형의 복잡성을 나타내는 측도인 K는 식 (2.3)에 주어진 가능도함수에 대한 모수의 개수이며 다음과 같이 주어진다.

$$K = \dim(\mu) + \dim(\theta) + 1. \tag{2.5}$$

BIC(Bayesian Information Criteria 또는 Schwarz Information Criteria)는 베이지안 추론에서 베이즈 인자(Bayes factor)를 이용하여 모형을 선택하는 경우에 사용하는 기준이다. r개의 모형 $\{M_1, M_2, \ldots, M_r\}$을 후보모형으로 고려하는 경우 i번째 모형에 대한 확률모형이 $f_i(y|\psi_i)$로 주어지고 사전분포가 $\pi_i(\psi_i)$로 주어졌다고 하자. BIC는 모형의 확률이 모두 같은 경우, $P(M_1) = P(M_2) = \cdots = P(M_r)$, 사후확률 $P(M_i|y)$가 y의 주변분포 $f_i(y)$에 비례하므로 이를 라플라스 근사(Laplace approximation)을 사용하여 근사한 기준이다.

$$\mathrm{BIC} = -2\log L + \log(N)K \approx -2\log f(y) = -2\log \int f(y|\psi)\pi(\psi)d\psi. \tag{2.6}$$

BIC의 K는 AIC를 정의할 때 사용한 모수의 개수 (2.5)와 같다.

정보기준 모형선택법의 대표적인 방법인 AIC와 BIC는 가능도함수를 기반으로 하는 것이 공통이지만 모수의 개수를 포함하는 벌칙항(penalty term)에 AIC는 2를, BIC는 연관된 표본 수의 로그값 $\log(N)$을 곱해주는 차이가 있으며 참모형을 포함한 모형들을 고려하면 일반적으로 BIC가 AIC보다 모수의 개수가 적은 모형을 선호하는 경향이 있다는 것이 잘 알려져 있다. 본 논문에서 사용하는 AIC와 BIC는 분산성분이나 선형혼합모형을 가정하고 이에 맞추어 유도된 기준은 아니지만 위에서 정의된 기본적인 AIC와 BIC가 선형혼합모형하에서 개선된 형태의 방법과 경험적 성질이 크게 다르지 않으므로 (Dimova 등, 2011) 본 논문에서는 식 (2.4)과 (2.6)에서 정의된 AIC와 BIC를 사용하여 모의실험을 수행한다. 또한 가능도함수는 불편추정량을 제시하는 제한적 최대가능도함수를 사용한다.

3. 모의실험

3.1. 실험설계

이원지분배치 모형과 이원교차배치 모형에서 모형선택법의 경험적 성질을 알아보기 위하여 모수의 형태를 체계적으로 설계하였다. 먼저 오차항의 분산은 언제나 1.0으로 고정하고($\sigma_e = 1.0$), 두 분산성분(σ_a^2, σ_b^2)에 대하여 다음과 같은 조합들을 고려하였다.

$$\sigma_a = r_i, \qquad \sigma_b = r_i \times r_j, \quad i, j = 1, 2, \ldots, 9, \tag{3.1}$$

여기서 r은 다음과 같은 9개의 숫자로 주어진다.

$$r = (0.0, 0.25, 0.50, 0.75, 1.00, 1.25, 1.50, 1.75, 2.00).$$

예를 들어 $\sigma_a = 0.5$으로 고정된 경우 σ_b의 값은 0.0을 포함하여 $(0.5)(0.25) = 0.125$부터 $(0.5)(2.0) = 1.0$까지 9단계로 점차 증가하는 경우를 고려하여 각각의 경우에 모의실험을 실시한다. 하나의 분산성분이 0인 경우도 고려하며 분산성분이 0이 되면 이에 해당하는 효과가 모형에 포함되지 않은다.

본 모의실험에서 고려하는 모형은 세가지 모형을 고려한다. 분산성분이 한 개만 있는 부분모형과 두 개가 있는 완전모형이다. 이원 지분배치 모형인 경우 다음과 같은 3개의 모형들이 후보모형이다.

$$[\mathbf{M_1}]\ y_{ijk} = \mu + \alpha_i + e_{ijk}, \quad [\mathbf{M_2}]\ y_{ijk} = \mu + \beta_{ij} + e_{ijk}, \quad [\mathbf{M_3}]\ y_{ijk} = \mu + \alpha_i + \beta_{ij} + e_{ijk}.$$

이원 교차배치 모형인 경우 다음과 같은 3개의 모형들이 후보모형이다.

$$[\mathbf{M_1}]\ y_{ijk} = \mu + \alpha_i + e_{ijk}, \quad [\mathbf{M_2}]\ y_{ijk} = \mu + \beta_j + e_{ijk}, \quad [\mathbf{M_3}]\ y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk}.$$

모의실험에서 식 (3.1)에서 정의된 각 조합에 대하여 이원지분배치 모형과 교차배치 모형에서 표본의 크기를 $a = b = n = 5$ ($N = 75$)인 경우와 $a = b = 8, n = 5$ ($N = 325$)인 두 가지 경우에 대하여 고려하고 각 모수의 조합에서 1000개의 모의자료를 생성하여 세 개의 후보모형들 중 AIC와 BIC 기준으로 선택된 모형의 상대적인 도수를 비교하였다. 자료를 생상하는 참모형은 분산성분이 0인 경우를 제외하면 식 (2.1)와 (2.2)에 주어진 모형 $\mathbf{M_3}$이다.

3.2. 실험의 결과

Figure 3.1은 이원교차배치 모형 (2.1)을 고려하고 표본의 크기가 75인 경우($a = b = n = 5$) 분산성분들의 상대적인 크기의 조합에 따라 세 개의 후보모형들이 각각 BIC와 AIC에 의하여 선택된 상대도수를 나타낸 그림들이다. 모든 분산성분의 조합에서 오차항의 분산이 1.0으로 고정되어 있다($\sigma_e = 1.0$).

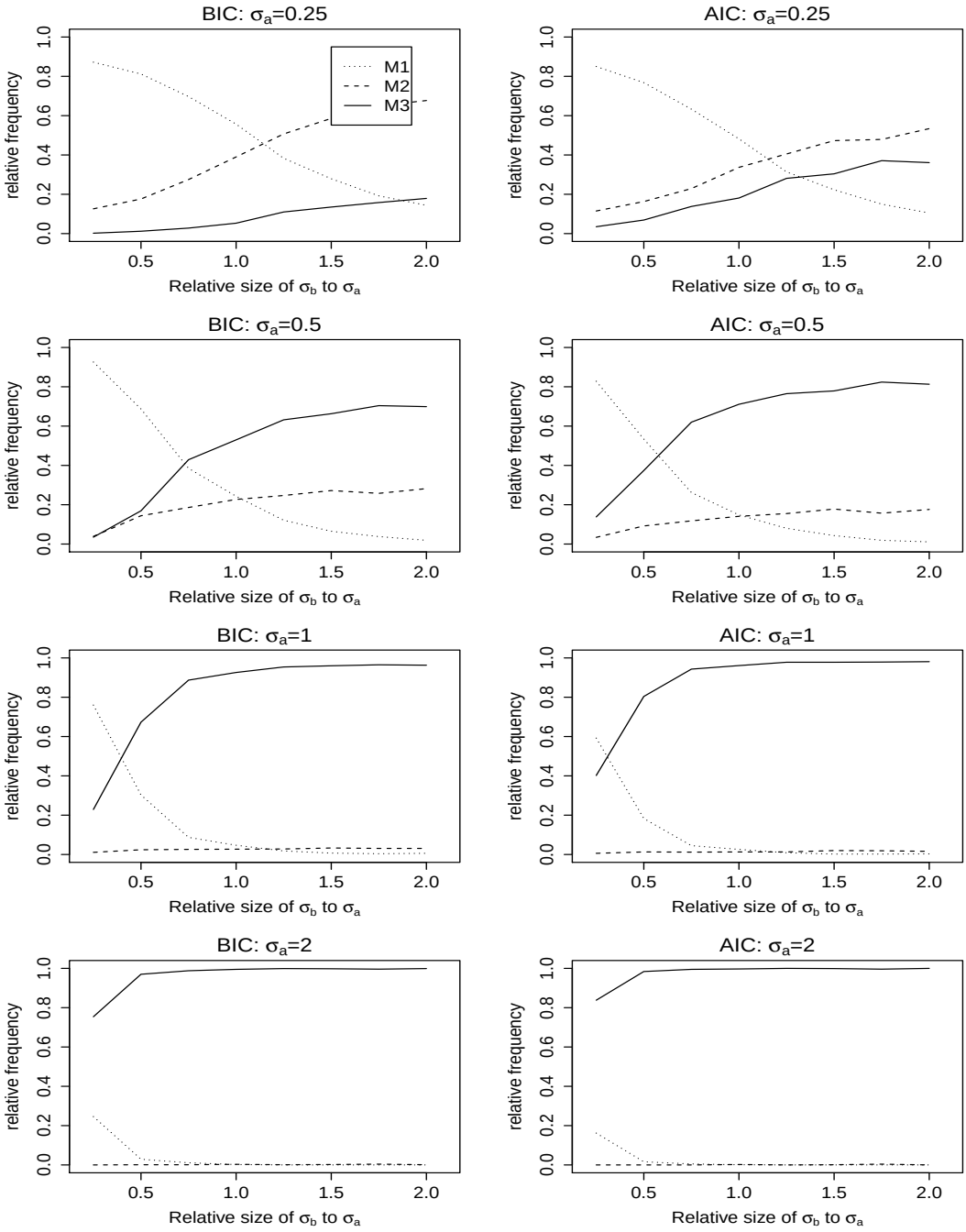

Figure 3.1. Relative frequency of selected candidate models by BIC and AIC under two-way crossed classification models; $\sigma_e = 1.0$ and sample size is $N = 75$ ($a = b = n = 5$); Dotted line indicates relative frequency of the model M_1 selected, dashed line indicates the model M_2, and solid line indicates the model M_3.

두 분산성분의 크기가 오차항의 분산의 크기보다 상대적으로 매우 작은 경우 두 분산성분의 상대적인 크기의 차이에 비례하여 모형 M_1 또는 모형 M_2가 더 많이 선택된다. 예를 들어 $\sigma_a = 0.25$인 경우 σ_b가 σ_a보다 작은 경우에는 모형 M_1이 주로 선택되고 큰 경우에는 모형 M_2가 주로 선택된다 (Figure 3.1에서 상단 그림). 오차항의 분산에 대한 두 분산성분의 상대적 크기가 커짐에 따라 자료를 생성하는 참모형인 M_3를 선택하는 상대도수가 빠르게 증가하는 것을 알 수 있다. 예를 들어 $\sigma_a = 1.0$인 경우 σ_b의 상대적인 크기가 0.75보다 크면($\sigma_b > 0.75\sigma_a$) 90% 이상 참모형인 M_3를 선택한다. Figure 3.1에서 볼 수 있듯이 AIC가 참모형을 선택하는 경향의 속도가 BIC보다 빠르다는 것을 알 수 있다. 표본의 크기가 $n = 325$인 경우 이원교차배치에서 모형선택의 패턴은 $n = 75$인 경우와 유사하며 또한 모든 조합에서 참모형인 M_3를 선택하는 상대도수가 빠르게 증가하고 BIC와 AIC의 상대적 성질이 유사하므로 생략하였다.

Figure 3.2는 이원지분배치 모형 (2.2)을 고려하고 표본의 크기가 75인 경우와 325인 경우에 분산성분들의 상대적인 크기의 조합에 따라 세 개의 후보모형들이 BIC에 의하여 선택된 상대도수를 나타낸 그림들이다. 모든 분산성분의 조합에서 오차항의 분산이 1.0으로 고정되어 있다($\sigma_e = 1.0$). 두 분산성분의 크기가 오차항의 분산의 크기보다 상대적으로 매우 작은 경우 Figure 3.1에서 나타난 이원교차배치 모형의 경우와 유사한 패턴을 보여준다(Figure 3.2의 상단 그림). 오차항의 분산에 대한 두 분산성분의 상대적 크기가 커짐에 따라 자료를 생성하는 참모형인 M_3를 선택하는 상대도수가 증가하지만 σ_b의 상대적인 크기가 σ_a보다 커지면 모형 M_2를 선호하는 경향이 강해진다. 이러한 성질은 이원교차배치 모형에서 나타나지 않은 지분배치 모형의 특별한 패턴이다. 예를 들어 $\sigma_a = 1.0$인 경우 σ_b의 상대적인 크기가 1.0보다 크면($\sigma_b > \sigma_a$) 50% 이상 참모형보다 작은 모형인 M_2를 더 많이 선택한다. 표본의 크기가 증가하면 참모형을 선택하는 상대적인 비율이 증가하지만 σ_b의 상대적인 크기가 σ_a보다 큰 경우 작은 모형인 M_2를 선택하는 경향이 증가하는 현상은 계속 일어난다.

Figure 3.3은 이원지분배치 모형에서 표본의 크기가 75인 경우와 325인 경우에 분산성분들의 상대적인 크기에 따라 세 개의 후보모형들이 AIC에 의하여 선택된 상대도수를 나타낸 그림들이다. 두 분산성분의 크기가 오차항의 분산의 크기보다 상대적으로 매우 작은 경우 BIC와 유사한 패턴을 보여준다(Figure 3.3의 상단 그림). 오차항의 분산에 대한 두 분산성분의 상대적 크기가 커짐에 따라 자료를 생성하는 참모형인 M_3를 선택하는 상대도수가 증가하지만 σ_b의 상대적인 크기가 σ_a보다 커지면 모형 M_2를 선호하는 경향이 강해진다. 이러한 성질은 Figure 3.2에서 나타난 BIC와 유사하지만 Figure 3.2와 Figure 3.3에서 볼 수 있듯이 AIC가 BIC보다 참모형을 선택하는 상대도수가 약간 크다.

Table 3.1은 지분배치모형에서 특정한 모수들의 조합을 고려한 경우 BIC와 AIC가 세 개의 모형을 선택하는 상대도수를 비교한 결과를 보여준다. 두 개의 분산성분를 고정하고($\sigma_a = 1$, $\sigma_e = 1$) σ_b의 크기와 표본의 개수의 변화에 따라 모형을 선택하는 경향을 비교하였다. BIC의 경우 표본의 수가 작고($N = 75$) $\sigma_b = 0.25$인 경우 참모형인 이원배치모형을 선택하는 비율은 5.6%이다. $\sigma_b = 0.75$로 그 값이 증가하면서 참모형을 선택하는 비율은 63.4%로 크게 증가하지만 $\sigma_b = 1.0$인 경우에는 참모형을 선택하는 비율이 줄어들기 시작하면서 $\sigma_b = 2.0$인 경우에는 참모형을 선택하는 비율이 13.3%까지 낮아지며 모형 M_2를 선택하는 비율이 빨리 증가한다. 표본의 수가 클 때는($N = 325$) 참모형을 선택하는 패턴은 비슷하게 나타나지만 참모형을 선택하는 비율은 상대적으로 증가한다. 하지만 $\sigma_b = 2.0$인 경우 참모형을 선택하는 비율이 36.8%로 모형 M_2를 선택하는 비율 63.2% 보다 낮다. AIC의 경우 표본의 수가 작고($N = 75$) $\sigma_b = 0.25$인 경우 참모형인 이원배치모형을 선택하는 비율은 22.3%이다. $\sigma_b = 0.75$로 그 값이 증가하면서 참모형을 선택하는 비율은 82.4%로 크게 증가하지만 $\sigma_b = 1.0$인 경우에는 참모형을 선택하는 비율이 줄어들기 시작하면서 $\sigma_b = 2.0$인 경우에는 참모형을 선택하는 비율이 31.5%까지 낮아지며 모형 M_2를 선택하는 비율이 빨리 증가한다. 표본의 수가 클 때는($N = 325$) 참모

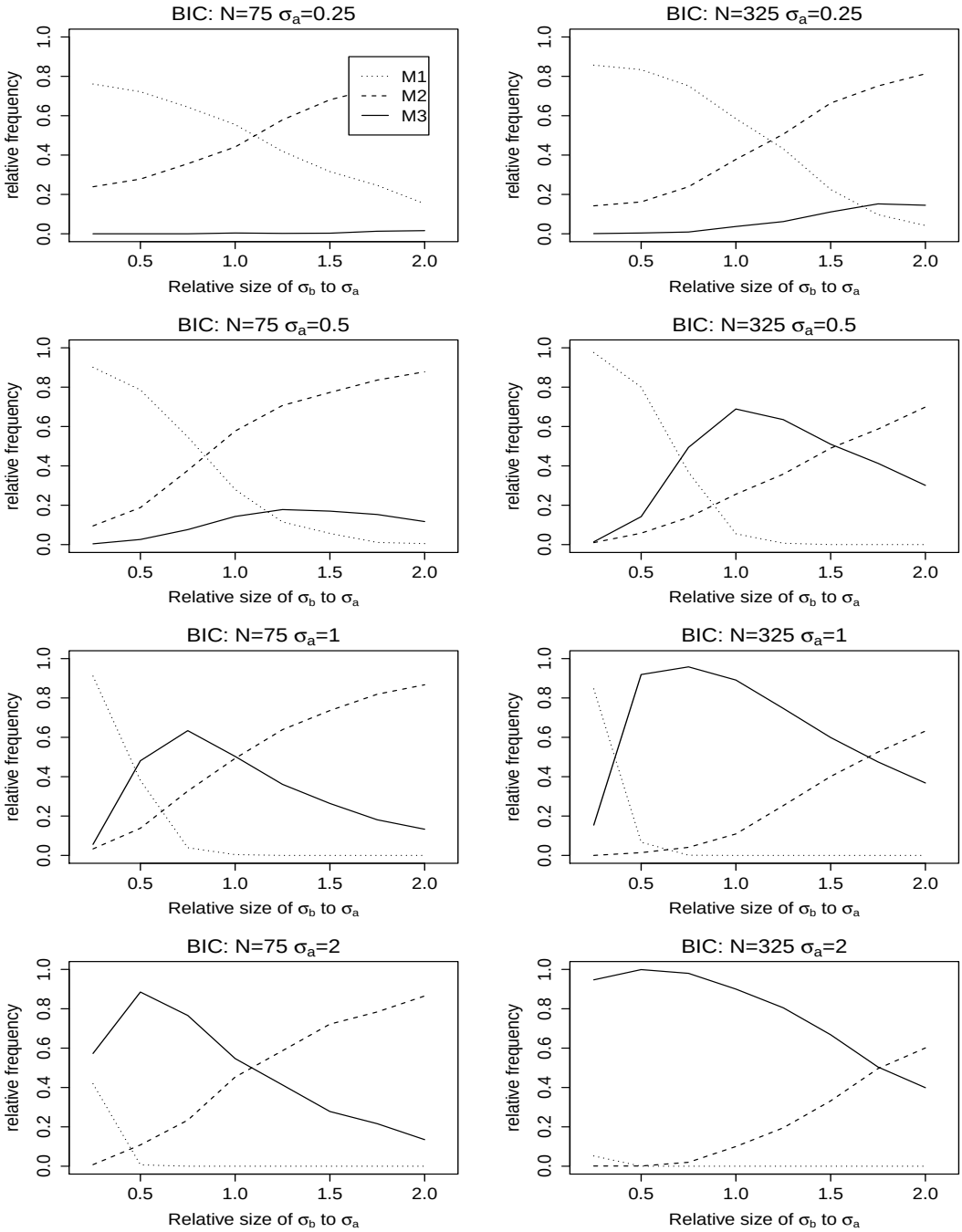

Figure 3.2. Relative frequency of selected candidate models by BIC under two-way nested classification models; $\sigma_e = 1.0$ and sample size is $N = 75$ ($a = b = n = 5$) or $N = 325$ ($a = b = 8, n = 5$); Dotted line indicates relative frequency of the model M_1 selected, dashed line indicates the model M_2, and solid line indicates the model M_3.

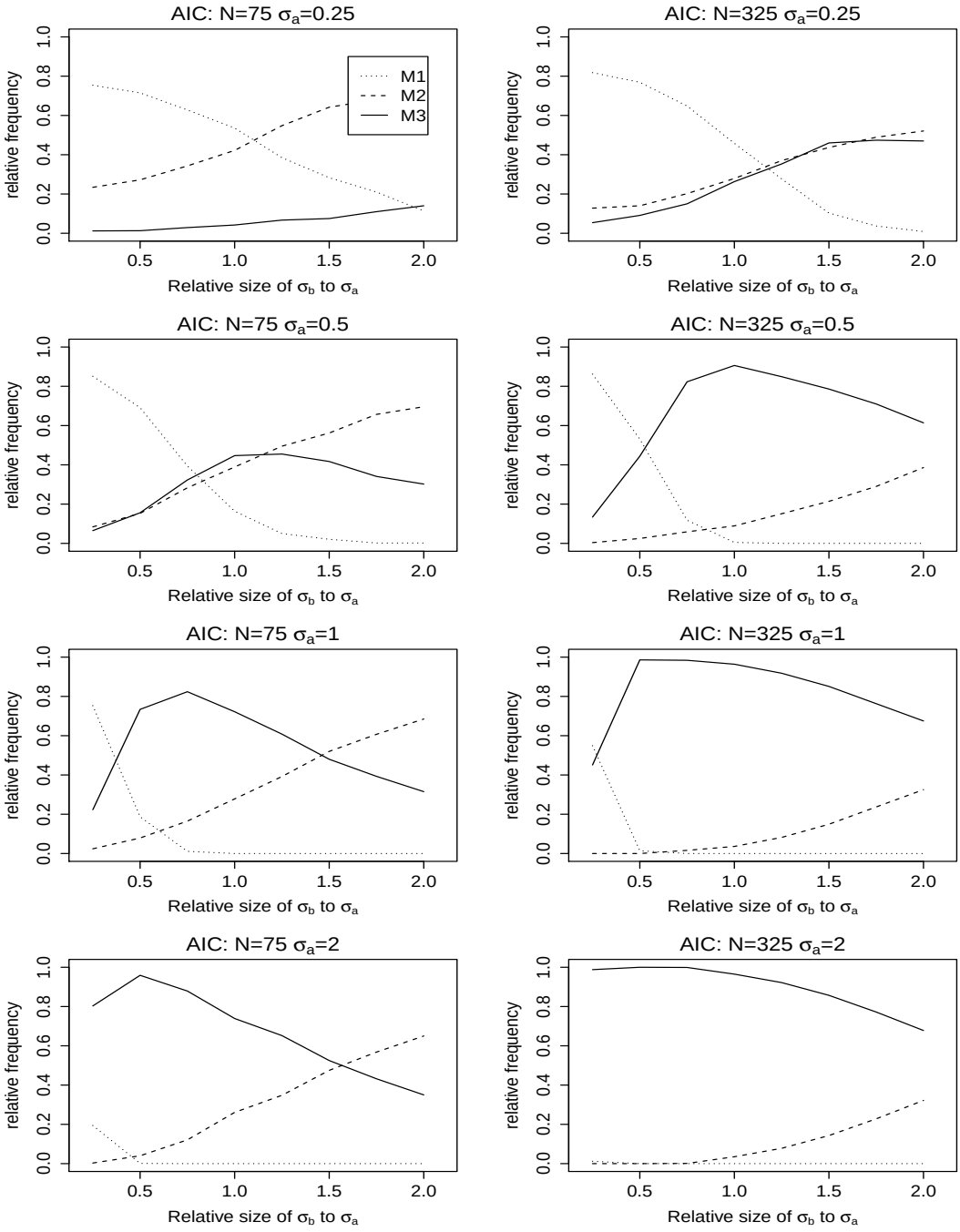

Figure 3.3. Relative frequency of selected candidate models by AIC under two-way nested classification models; $\sigma_e = 1.0$ and sample size is $N = 75$ ($a = b = n = 5$) or $N = 325$ ($a = b = 8, n = 5$); Dotted line indicates relative frequency of the model M_1 selected, dashed line indicates the model M_2, and solid line indicates the model M_3.

Table 3.1. Relative frequency(percentage) of selected candidate models by BIC and AIC under two-way nested classification models ($\sigma_e = 1.0$)

N	σ_a	σ_b	BIC			AIC		
			M_1	M_2	M_3	M_1	M_2	M_3
75	1.0	0.25	91.1	3.3	5.6	75.3	2.4	22.3
	1.0	0.75	3.9	32.7	63.4	1.1	16.5	82.4
	1.0	1.0	0.4	49.3	50.3	0.0	27.8	72.2
	1.0	2.0	0.0	86.7	13.3	0.0	68.5	31.5
325	1.0	0.25	84.6	0.0	15.4	54.9	0.0	45.1
	1.0	0.75	0.1	4.1	95.8	0.0	1.6	98.4
	1.0	1.0	0.0	10.9	89.1	0.0	3.6	96.4
	1.0	2.0	0.0	63.2	36.8	0.0	32.5	67.5

형을 선택하는 패턴은 표본의 수가 작은 경우와 비슷하게 나타나지만 참모형을 선택하는 비율은 상대적으로 증가한다. $\sigma_b = 2.0$인 경우에 참모형을 선택하는 비율이 67.5%로 나타난다.

Table 3.1의 결과를 보면 지분배치모형에서는 σ_b/σ_a가 커지는 경우 AIC와 BIC 모두 참모형인 이원배치모형을 선택하는 비율이 감소하지만 표본의 개수에 따라 AIC와 BIC의 효율이 매우 달라지는 것을 알 수 있다. 특히 $\sigma_a = 1$, $\sigma_b = 2$, $\sigma_e = 1$인 경우는 분산성분의 상대적 크기가 매우 큰 차이가 아님에도 불구하고 표본의 크기가 큰 경우에도 BIC의 경우 참모형을 선택하는 비율이 약 40%밖에 되지 않는다.

4. 결론

분산성분모형에서 요인의 배치구조와 분산성분의 크기에 따라 모형선택법의 경험적인 성질이 다르게 나타나는 현상을 체계적인 모의실험을 통하여 제시하였다. 이원배치 분산성분모형에서 요인 배치 또는 설계의 구조에 따라 교차배치와 지분배치를 고려하고 오차항의 분산에 대한 두 분산성분의 상대적인 크기에 따라 정보기준에 근거한 모형선택법 AIC와 BIC가 참모형을 선택하는 경험적 성질을 비교하였다.

이원 교차배치 모형에서는 두 분산성분의 크기가 커짐에 따라 참모형을 선택하는 경향이 강해지며 표본의 개수가 증가하면 그 효율이 빠르게 높아진다. 하지만 이원 지분배치 모형에서는 두 분산성분의 크기가 커짐에 따라 참모형을 선택하는 경향이 증가하다가 내포된 그룹에 대한 분산성분의 상대적 크기가 증가하면 참모형보다 작은 내포된 그룹에 대한 임의효과만 포함한 작은 모형을 선호하는 경향이 AIC와 BIC 모두 증가한다. 하지만 AIC와 BIC는 모형선택을 하는 경향이 서로 다르다. 실제 자료를 분석하고 모형을 세울 경우 이원 지분배치 모형에서는 표본의 개수가 매우 크더라도 내포된 분산성분의 크기가 상대적으로 커지면 참모형보다 작은 모형을 선택하는 경향이 크다는 것을 참고하여 모형의 선택에 신중을 기할 필요가 있다. 이러한 성질은 요인의 개수가 많아지면 분산 성분간에 더 복잡한 관계를 가지고 나타날 수 있다는 것도 예상할 수 있다.

AIC와 BIC를 이용한 모형선택은 서로 다른 두개의 모형 M_i과 M_j에 대한 정보기준 값의 차이로 결정된다, 즉 AIC인 경우 정보기준 값의 차이는

$$\mathrm{AIC}(M_i) - \mathrm{AIC}(M_j) = -2\log\frac{L(M_i)}{L(M_j)} + 2[K(M_i) - K(M_j)].$$

또한 BIC의 경우 아래와 같다.

$$\mathrm{BIC}(M_i) - \mathrm{BIC}(M_j) = -2\log\frac{L(M_i)}{L(M_j)} + (\log N)[K(M_i) - K(M_j)].$$

AIC와 BIC는 모두 가능도비 $L(M_i)/L(M_j)$에 의존하고 있으나 그 차이는 벌칙항에서 BIC는 표본의 개수에 영향을 받으며 AIC는 영향을 받지 않는다. Crainiceanu와 Ruppert (2004)는 일원배치모형에서 가능도비 $L(M_i)/L(M_j)$의 통계적 분포가 카이제곱분포의 선형결합으로 나타남을 보였다. 따라서 본 논문에서 보여진 결과를 설명하기 위해서는 이원배치 모형에서 가능도비 $L(M_i)/L(M_j)$가 분산성분의 비율에 따라, 또한 벌칙항이 표본 개수를 포함하느냐에 따라 정보기준의 차이에 어떤 영향을 미치는가에 대한 통계적인 성질을 규명해야 한다. 또한 지분배치모형에서 분산성분의 추정량은 내포된 요인들의 제곱합의 차이로 구해지기 때문에 추정의 정도(예를 들어 추정량의 평균제곱오차)가 모형선택에 미치는 영향도 중요하다고 예상된다. 모의실험에서 확인된 경험적인 성질을 구체적이고 이론적으로 규명하는 것이 향후 연구과제이다.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, *Second International Symposium on Information Theory*, 267–281, Akademiai Kiado, Budapest.

Chen, Z. and Dunson, D. B. (2003). Random effects selection in linear mixed models, *Biometrics*, **59**, 762–769.

Crainiceanu, C. M. and Ruppert, D. (2004). Likelihood ratio tests in linear mixed models with one variance component, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **66**, 165–185.

Dimova, R. B., Markatou, M. and Talal, A. H. (2011). Information methods for model selection in linear mixed effects models with application to HCV data, *Computational Statistics and Data Analysis*, **55**, 2677–2697.

Greven, S. and Kneib, T. (2010). On the behaviour of marginal and conditional AIC in linear mixed models, *Biometrika*, **97**, 773–789.

Lee, Y. (2015). A note on performance of conditional Akaike Information Criteria in linear mixed models, *Communications of the Korean Statistical Society*, Accepted.

Müller, S., Scealy, J. L. and Welsh, A. H. (2013). Model selection in linear mixed models, *Statistical Science*, **28**, 135–167.

Pauler, D. K., Wakefield, J. C. and Kass, R. E. (1999). Bayes factors and approximations for variance component models, *Journal of the American Statistical Association*, **94**, 1242–1253.

Pinheiro, J. and Bates, D. (2006). Mixed-effects models in S and S-PLUS, *Springer Science and Business Media*.

Saville, B. R. and Herring, A. H. (2009). Testing random effects in the linear mixed model using approximate Bayes factors, *Biometrics*, **65**, 369–376.

Searle, S. R., Casella, G. and McCulloch, C. E. (1992). *Variance Components*, **391**, John Wiley & Sons.

Stram, D. O. and Lee, J. W. (1994). Variance components testing in the longitudinal mixed effects model, *Biometrics*, **50**, 1171–1177.

Vaida, F. and Blanchard, S. (2005). Conditional Akaike information for mixed-effects models, *Biometrika*, **92**, 351–370.

분산성분모형에서 요인의 배치구조가 모형선택법에 미치는 영향에 대한 실험연구

이용희[a,1]

[a]서울시립대학교 통계학과

(2015년 9월 14일 접수, 2015년 10월 12일 수정, 2015년 10월 14일 채택)

요 약

분산성분모형은 다양한 임의 요인들이 반응변수에 미치는 영향을 선형식의 형태로 나타내는 매우 유용하고 널리 사용되는 통계적 모형이다. 분산성분모형은 요인의 배치나 관측 자료의 구조에 따라 크게 교차배치와 지분배치로 나누어진다. 본 논문은 분산성분모형에서 요인의 배치구조와 분산성분의 크기에 따라 모형선택법의 경험적인 성질이 다르게 나타나는 현상을 체계적인 모의실험을 통하여 제시하고자 한다. 이원배치 분산성분모형에서 정보기준에 근거한 모형선택법, 즉 BIC 또는 AIC를 사용하는 경우 요인의 배치구조와 분산성분의 크기에 따라 모형선택법의 경험적인 성질이 다르게 나타나는 현상을 소규모 모의실험을 통하여 보여준다. 모의실험 결과에서 모형선택법의 경험적 성질이 요인의 배치 설계에 따라 다르게 나타난다는 사실을 확인하였으며 특히 요인의 배치구조가 지분 설계구조일 때 내포된 요인의 분산성분의 상대적인 크기가 커짐에 따라 자료를 생성하는 모형보다 작은 모형을 선택하는 경향이 있다는 것이 모의실험으로 확인되었다.

주요용어: 분산성분모형, 선형혼합모형, 교차배치, 지분배치, 모형선택, BIC, AIC, 요인의 배치

[1](130-743) 서울특별시 동대문구 서울시립대로 163, 서울시립대학교 통계학과. E-mail: ylee@uos.ac.kr