The Korean Journal of Applied Statistics (2017) DOI: https://doi.org/10.5351/KJAS.2017.30.1.057
30(1), 57–68

Bayesian analysis of finite mixture model with cluster-specific random effects

Hyejin Lee[a] · Minjung Kyung[a,1]

[a]Department of Statistics, Duksung Women's University

(Received September 22, 2016; Revised November 28, 2016; Accepted December 20, 2016)

Abstract

Clustering algorithms attempt to find a partition of a finite set of objects in to a potentially predetermined number of nonempty subsets. Gibbs sampling of a normal mixture of linear mixed regressions with a Dirichlet prior distribution calculates posterior probabilities when the number of clusters was known. Our approach provides simultaneous partitioning and parameter estimation with the computation of classification probabilities. A Monte Carlo study of curve estimation results showed that the model was useful for function estimation. Examples are given to show how these models perform on real data.

Keywords: clustering analysis, finite mixture model, cluster-specific random effect, Gibbs sampling

1. Introduction

군집분석이란 주어진 데이터에서 비슷한 특성을 가진 객체끼리 동일한 집단으로 분류해내는 것을 일컫는다. 즉 다른 집단과 구분 할 수 있는 대표적인 특성을 찾아내어 분류하고 분석하는 방법을 말한다. 결국 군집분석은 대량의 데이터가 수집되었을 경우에 모든 객체의 특성을 하나하나 파악하지 않고도 각 군집의 대표적인 특성을 찾아냄으로써 데이터의 전반적인 특성 및 구조를 파악 할 수 있게 도와준다. 이러한 군집분석 방법으로 Dempster 등 (1977)에 의해 정의된 expectation-maximization(EM) 알고리즘은 가장 보편적으로 널리 사용되는 추정방법이며, 각 객체의 군집 id는 군집의 갯수가 결정 된 후 각 개체들의 사후 확률에 의해 정해진다 (Banfield와 Raftery, 1993; Dasgupta와 Raftery, 1998; Fraley와 Raftery, 2002; McLachlan과 Basford, 1988; McLachlan과 Peel, 2000). 결국 각각의 혼합된 요소들은 동일한 일반적인 특성을 지님과 동시에 다른 집단과는 구분되는 특성을 갖는 모수들에 의존하게 된다.

특히 모형기반 군집방법(model-based clustering; MBC)은 모형에 근거를 둔 군집화 방법으로 Scott과 Symons (1971)가 제안하고 Banfield와 Raftery (1993), Dasgupta와 Raftery (1998), Fraley와 Raftery (2002) 등이 발전시킨 방법이다. 이는 확률분포에 관한 정보가 있을 경우 확률분포를 가정하여 군집화를 방법으로 각 개체가 최대 확률로 해당 그룹에 할당될 때 최종 결과로 수렴되며, 모형선택 및 군집의 개수는 Bayesian information criterion(BIC)를 계산하여 BIC가 최소가 되는 군집 개수를 최종모형으로 선택한다.

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (NRF-2015R1C1A1A01051837).
[1]Corresponding author: Department of Statistics, Duksung Women's University, 33, Samyang-ro 144-gil, Dobong-gu, Seoul 01369, Korea. Email: mkyung@duksung.ac.kr

만일 반응변수들의 관측값들을 군집화할 때, 설명변수들과의 관계성에 기반하여 군집화한다면, 모형기반 군집방법 중 선형회귀모형에 의한 유한혼합물을 고려할 수 있다. 선형모형의 유한혼합물(finite mixture of linear model) 기법은 군집분석 방법 중 다양한 분야에서 많이 사용되는 방법으로, Quandt (1958)의 교체회귀(switching regression) 방법의 확장형으로 정규혼합물 모형에 기반하여 소개되었다. Quandt (1958)와 Quandt와 Ramsey (1978)는 혼합물 모형의 모수를 추정하는데 적률생성 함수를 사용하였다. De Veaux (1989)는 두 그룹의 정규선형회귀선을 추정하고 군집화를 위해 EM 방법을 적용하였다. 유한혼합물 모형에 대한 자세한 설명 및 다양한 추정법은 McLachlen과 Peel (2000), Fruhwirth-Schnatter (2005) 등에서 확인할 수 있다.

우리는 이 논문에서 일반적인 선형회귀모형의 유한혼합물 기법을 확장하여 변량효과를 포함한 선형혼합모형의 유한혼합물(finite mixture of linear mixed model) 분석방법을 연구한다. 이러한 선형혼합모형의 유한혼합물에서 변량 효과들은 숨겨진 데이터의 구조를 설명하고, 혼합(mixture) 모형들로 다봉 분포 및 비대칭 분포의 특성을 파악할 수 있다. 특히 변량효과로는 군집특정적(cluster-specific) 변량효과를 고려하였다. 군집특정적 변량효과는 군집 간의 독립을 가정하고 비슷한 특성을 동일한 군집 내의 구성 원소들이 공유함으로써 관찰되지 않는 군집의 특성들을 조정할 수 있게 된다 (Kyung, 2015). 그리고 선형모형을 바탕으로 군집화를 하는 과정에서 각 군집 별 다른 절편을 고려하여 일반적인 선형모형의 유한혼합물 모형보다 더 정확한 군집화를 기대할 수 있다.

적절한 사전분포를 사용할 경우, 사후분포가 적합하게 되므로 깁스 표본추출방법과 같은 MCMC 방법을 사용하여 정확한 근사값을 제공할 수 있다. 정규오차를 가정한 선형혼합모형의 유한혼합물기법의 추정방법으로는 깁스 표본추출 알고리즘(Gibbs sampling algorithm)을 사용한 베이지안 분석방법을 활용한다. 군집모수인 각 그룹에 대한 가중치 확률의 사전분포로는 공액사전분포(conjugate prior distribution)인 디리슈레 분포(Dirichlet distribution)을 사용하고, 각 그룹 안에서의 변량 효과는 평균이 0인 정규분포를 가정하며 선형모형의 기울기 모수에 대해서도 공액사전분포인 정규분포를 가정한다.

현재 MCMC 방법의 적용이 보편적이지만, 혼합모형에서 다루어야 할 베이지안 접근법들에 관련해선 어려움들이 존재한다. 첫 번째로 주된 문제점은 적합하지 않은 사전분포의 사용은 적합하지 않은, 부적절한 사후분포를 도출한다는 것이다. 이 문제에 대한 대안은 “부분적으로(partially) 적합한 사전분포”를 사용하는 것이다. 또 다른 문제점은 구성 성분들의 수가 알려져 있지 않을 때, 모수공간이 동시에 잘못 정의되고 공간이 무한으로 발산한다는 것이다. 이것은 전통적인 검정 과정과 사전분포들의 사용을 방해한다. 따라서 일반적인 접근법은 고정된 구성성분들에 대한 혼합 모형에 적합하며, 그리고 소위 정보기준(information criterion)에 의해 그룹 선택을 고려한다. 최근, Phillips와 Smith (1996) 그리고 Richardson와 Green (1997)에서 알려지지 않은 모수를 도출하기 위해 구성성분들과 함께 완전 베이지안 접근법을 보였다. 그들의 MCMC 방법들은 변수 공간 모수들을 구성할 수 있으며 결국 구체화되지 않은 구성요소들을 다룰 수 있게 돕는다 (Geoffrey와 David, 2000).

베이지안 군집방법은 Bernardo와 Giron (1988)이 군집에 대한 가중치 확률만 모를 경우 처음 적용하였다. 군집 내 모수들 및 가중치 확률에 대해 공액사전분포를 사용하여도, 사후분포는 다루기 쉬운 형태의 분포족이 아니므로 마르코프 체인 몬테 카를로(Markov chain Monte Carlo) 기법을 사용해야 하는 어려움이 있다. 그러나 유한혼합물 모형에서 군집의 지수(index)를 결측자료로 가정하여 구한 Dempster 등 (1977)의 근본적인 내용을 바탕으로, 베이지안 방법에서는 자료확대(data augmentation) 방법과 깁스 표본추출 알고리즘을 적용하여 사후분포로부터 직접적인 표집법이 가능하다 (Diebolt와 Robert, 1994; Escobar와 West, 1995; Mengersen와 Robert, 1996; Raftery, 1996; Smith와 Roberts, 1993; West, 1992).

MCMC 방법을 사용한 혼합 모형의 베이지안 분석방법에 대한 주요 초기 논문은 Diebolt와 Robert

(1990, 1994)와 Escobar와 West (1995)를 포함한다. 또한 혼합(mixture) 문제는 Smith와 Roberts (1993)의 MCMC 방법 리뷰에서 다루어진다. 더 발전된 문제는 Bensmail 등 (1997), Cao와 West (1996), Carlin과 Chib (1995), Mengersen과 Robert (1996), Phillips와 Smith (1996), Raftery (1996), Richardson과 Green (1997), Robert (1996), Robert와 Mengersen (1999), Roeder와 Wasserman (1997), West 등 (1994), 그리고 Yu와 Tanner (1999)에서 다루어진다. 사후 모의실험을 통한 베이지안 혼합문 추정의 추가적인 적용에 대해선 Gelman 등 (1995)에서 언급된다. 혼합 분포들의 베이지안 분석방법을 포함한 최근의 다른 연구들은 Dellaportas (1998)와 Vounatsou 등 (1998)에서 보이고 있다.

본 논문의 구성은, 2절에서는 군집특정적 변량효과를 포함한 선형 혼합 모형의 유한혼합물 모형에 대해서 설명하고, 3절에서는 깁스 표본 추출법을 적용하여 모형의 모수들과 군집 도출에 대해서 설명한다. 그리고 4절에서는 모의실험을 통해 제안된 모형을 적용하며 5절에서는 실제 데이터에 적용한다. 마지막 6절에서는 요약과 결론으로 끝을 맺는다.

2. The finite mixture model of linear mixed regressions

일반적인 정규오차 선형모형의 유한혼합물 분포는

$$Y_i \sim \sum_{k=1}^{K} w_k N\left(\boldsymbol{x}_i\boldsymbol{\beta}_k, \sigma_k^2\right)$$

이며, $w_1, w_2, \ldots, w_K$는 혼합비율(mixing proportion)로 다음과 같은 성질을 만족한다.

$$0 \le w_k \le 1, \quad (k = 1, \ldots, K), \ \sum_{k=1}^{K} w_k = 1.$$

이를 확장한 군집특정적 변량효과를 가진 혼합모형은 다음과 같이 나타낼 수 있다.

$$\begin{aligned} Y_i | i \in C_k &= \boldsymbol{x}_i\boldsymbol{\beta}_k + \eta_k + \epsilon_{ik}, \quad i = 1, \ldots, n, \ k = 1, \ldots, K \\ \epsilon_{ik} | i \in C_k &\sim N\left(0, \sigma_k^2\right), \\ \eta_k &\sim N\left(0, \tau^2\right), \end{aligned}$$

여기서 Y_i는 i번째 반응변수이며, $\boldsymbol{x} = (x_{i1}, \ldots, x_{ip})'$는 i번째 개체의 독립변수들의 벡터를 나타낸다. 확장된 형태의 모형으로 일반적인 선형모형 대신 일반화 선형모형을 사용하는 경우도 있다. 그러나 본 논문에서는 설명변수벡터와 연관된 계수의 내적($\boldsymbol{x}_i\boldsymbol{\beta}_k$)인 선형모형을 가정하였다. C_k는 k번째 군집에 속한 원소들의 id 벡터를 나타내며, $\boldsymbol{\beta}_k = (\beta_{ik}, \ldots, \beta_{pk})'$는 k번째 군집에 대한 회귀계수의 벡터이고 η_k는 k번째 군집특정적 변량효과를 나타내며 평균이 0이고 분산이 τ^2인 정규분포를 가정한다. 마지막으로 σ_k^2는 k번째 군집의 오차의 분산을 나타낸다.

쉬운 표기법 및 군집화 과정을 위해 분할 레이블 $\boldsymbol{Z} = (Z_1, \ldots, Z_n)$을 고려하여 모형을 다시 표기한다. 반응변수 Y_i가 k번째 군집에서 관찰되었다고 가정 할 때 모형을 다음과 같이 나타낼 수 있다.

$$\begin{aligned} Y_i | Z_i = k, \boldsymbol{\beta}_k, \sigma_k^2 &\sim N\left(\boldsymbol{x}_i\boldsymbol{\beta}_k + \eta_k, \sigma_k^2\right), \\ \eta_k | Z_i = k, \tau^2 &\sim N\left(0, \tau^2\right), \end{aligned}$$

여기서 $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$는 모수들의 벡터로서 각 $\theta_k = (\boldsymbol{\beta}_k, \sigma_k^2)$으로 나타낼 수 있다. $\boldsymbol{Z} = (Z_1, \ldots, Z_n)$는 확률 $P(Z_i = k) = w_k, \sum_{k=1}^{K} w_k = 1$을 가진 관찰되지 않는 군집 레이블을 나타내며 이 변수는

각 관찰값들이 어떤 군집에 포함되는지를 확인할 수 있게 해준다. $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_K)$는 군집특정적 변량효과들의 벡터이다. 그러므로 군집에 대한 정보가 주어졌을 때, 즉 군집 레이블이 주어졌을 때 모형의 모수들의 가능도함수는 다음과 같다.

$$L(\boldsymbol{\beta}, \boldsymbol{\sigma}^2, \boldsymbol{\eta}, \tau^2 | \boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{y}) = \prod_{i=1}^{n} \left[I(Z_i = k) \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{1}{2\tau^2} \eta_{Z_i}^2 \right) \frac{1}{\sqrt{2\pi\sigma_{Z_i}^2}} \exp\left\{ -\frac{1}{2\sigma_{Z_i}^2} \left(y_i - x_i \boldsymbol{\beta}_{Z_i} - \eta_{Z_i} \right) \right\}^2 \right],$$

여기서 $\boldsymbol{\beta}$는 $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K)$, $\boldsymbol{\sigma}^2$는 $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_K^2)$이며 $\boldsymbol{\eta}$는 $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_K)$이다.

군집특정적 변량효과를 갖는 혼합모형에서 군집특정적 절편과 군집특정적 변량효과는 식별성(identifiability)을 갖지 못하게 되는데 절편은 위치모수에 대응되고 변량효과는 퍼짐도 형태에 영향을 주는 것과는 별개이기 때문이다. 이러한 식별성은 실제 β값이 제대로 추정되는지 모의실험을 통해 확인해 볼 수 있을 것이다. 하지만 본 논문에서는 모형의 군집 분리에 대한 성능을 확인하는 것이 목표이며 가정된 모형에서 각 군집에 사용되는 절편은 각 군집의 위치모수에 포함되는 모수이고, 변량효과는 군집별로 다른 분산값을 설명하기 위해 모형에 포함된 변량효과이다. 베이지안 분석법을 이용한 군집분석에서는 우도함수를 최대화는 maximum likelihood estimation(MLE) 방법과는 다르게 절편을 회귀식에 포함하여 위치모수를 추정하는데 사용하고, 위치모수가 주어진 조건 하에서 임의효과를 추정하여 각 군집의 퍼짐도 및 형태를 결정하는데 사용한다.

3. Sampling schemes

이 절에서는 앞 절에서 설명한 군집특정적 변량효과를 포함한 선형혼합모형의 유한혼합물 모형에 대한 추정법으로 공액사전분포를 활용하여 각 모수의 완전 조건부 사후분포 도출에 대하여 서술한다. 활용하는 표본 추출방법은 깁스 표본추출방법으로 이 추출법을 통해 모형의 모수들인 $\boldsymbol{\beta}_k, \sigma_k^2$, 군집 특정 변량효과 η_k와 가중치 w_k는 물론 분할 레이블 $\boldsymbol{Z}$까지 도출한다.

모수들의 추정에 있어 활용하는 공액사전분포는 다음과 같다.

$$\begin{aligned}
\boldsymbol{\beta}_k | \sigma_k^2 &\sim N\left(\mathbf{0}, d\sigma_k^2 I\right), \\
\sigma_k^2 &\sim \mathrm{IG}(a, b), \\
\eta_k &\sim N\left(0, \tau^2\right), \\
\tau^2 &\sim \mathrm{IG}(a_1, b_1), \\
(w_1, \ldots, w_k) &\sim D(\alpha_1, \ldots, \alpha_k),
\end{aligned}$$

여기서 $d > 1$이고 $\mathrm{IG}(a, b)$는 모수 a와 b를 갖는 역감마분포이다. D는 모수 $(\alpha_1, \ldots, \alpha_k)$를 갖는 디리슈레 분포이며 이때 모수는 α_k이고 $\alpha_k > 0$이다.

각 모수의 완전 조건부 사후분포는 다음과 같이 도출된다.

1. $\boldsymbol{\beta}_k, \eta_k, w_k$가 주어졌을 때, 새로운 σ_k^2 값은 완전 조건부 사후분포

$$\sigma_k^2 | \boldsymbol{\beta}_k, \eta_k, w_k, \boldsymbol{X}, \boldsymbol{y}, Z = k \sim \mathrm{IG}(a^*, b^*)$$

으로부터 생성되며 이때

$$a^* = a + \frac{n_k + 1}{2}, \quad b^* = b + \frac{1}{2} \left\{ (\boldsymbol{y}_k - \boldsymbol{x}_k \boldsymbol{\beta}_k - \eta_k \mathbf{1})' (\boldsymbol{y}_k - \boldsymbol{x}_k \boldsymbol{\beta}_k - \eta_k \mathbf{1}) + \frac{1}{d} \boldsymbol{\beta}_k' \boldsymbol{\beta}_k \right\}$$

이다. K는 군집을 나타내며 $k = 1, \ldots, K$이고 n_k는 k번째 군집에 포함된 원소의 수이다.

2. $k = 1, \ldots, K$에서 σ_k^2, η_k, w_k가 주어졌을 때, 새로운 $\boldsymbol{\beta}_k$값은 완전 조건부 사후분포

$$\boldsymbol{\beta}_k | \sigma_k^2, \eta_k, w_k, \boldsymbol{X}, \boldsymbol{y}, Z = k \sim N\left(\boldsymbol{\beta}_k^*, \sigma_k^2 \left(\boldsymbol{x}_k' \boldsymbol{x}_k + \frac{1}{d} I\right)^{-1}\right)$$

으로부터 생성되며 이때

$$\boldsymbol{\beta}_k^* = \left(\boldsymbol{x}_k' \boldsymbol{x}_k + \frac{1}{d} I\right)^{-1} \left\{\boldsymbol{x}_k' (\boldsymbol{y}_k - \eta_k \mathbf{1})\right\}$$

이다.

3. $k = 1, \ldots, K$에서 $\boldsymbol{\beta}_k, \sigma_k^2, w_k, \tau^2$가 주어졌을 때, 새로운 군집 특정 변량효과 η_k값은 완전 조건부 사후분포

$$\eta_k | \boldsymbol{\beta}_k, \sigma_k^2, w_k, \tau^2, \boldsymbol{X}, \boldsymbol{y}, Z = k \sim N\left(\frac{n_k \tau^2}{n_k \tau^2 + \sigma_k^2} \mathbf{1}'(\boldsymbol{y}_k - \boldsymbol{x}_k \boldsymbol{\beta}_k), \left(\frac{n_k}{\sigma_k^2} + \frac{1}{\tau^2}\right)^{-1}\right)$$

로부터 생성된다.

4. $k = 1, \ldots, K$에서 $\boldsymbol{\beta}_k, \sigma_k^2, \eta_k$가 주어졌을 때, 새로운 τ^2은 완전 조건부 사후분포

$$\tau^2 | \boldsymbol{\beta}_k, \sigma_k^2, w_k, \eta_k, \boldsymbol{X}, \boldsymbol{y}, Z = k \sim \mathrm{IG}\left(a_1 + \frac{K}{2}, b_1 + \frac{1}{2} \sum_{k=1}^{K} \eta_k^2\right)$$

로부터 생성된다. 이때 η_k는 군집 특정 변량효과이다.

5. $k = 1, \ldots, K$에서 $\boldsymbol{\beta}_k, \sigma_k^2, \eta_k$가 주어졌을 때, 군집 가중치 w_k는 완전 조건부 사후분포

$$w_k | \boldsymbol{\beta}_k, \sigma_k^2, w_k, \eta_k, \boldsymbol{X}, \boldsymbol{y}, Z = k \sim \mathrm{Dirichlet}\,(n_k + \alpha_k)$$

으로부터 생성된다. α_k는 디리슈레 분포의 모수이며($\alpha > 0$), n_k는 k번째 군집에 포함된 원소의 수이다.

6. $k = 1, \ldots, K,\ i = 1, \ldots, n$에서 $\boldsymbol{\beta}_k, \sigma_k^2, \eta_k, \tau^2, w_k$가 주어졌을 때, 군집을 구분하는 군집 레이블 Z_i의 완전조건부 사후 확률 $P(Z_i = k)$는

$$P\left(Z_i = k | \boldsymbol{\beta}_k, \sigma_k^2, \eta_k, w_k, \tau^2, \boldsymbol{X}, \boldsymbol{y}\right) = \frac{w_k N\left(x_i \boldsymbol{\beta}_k + \eta_k, \sigma_k^2\right)}{\sum_{k=1}^{K} w_k N\left(x_i \boldsymbol{\beta}_k + \eta_k, \sigma_k^2\right)}$$

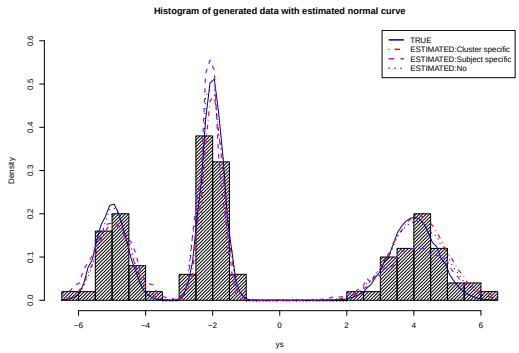
으로부터 생성되며, 관찰값은 이 확률값에 근거하여 가장 큰 확률 값을 갖는 군집으로 포함된다.

이러한 완전 조건부 사후분포에서 발생된 새로운 모수들이 수렴할 때까지 깁스 표본추출방법을 반복한다.

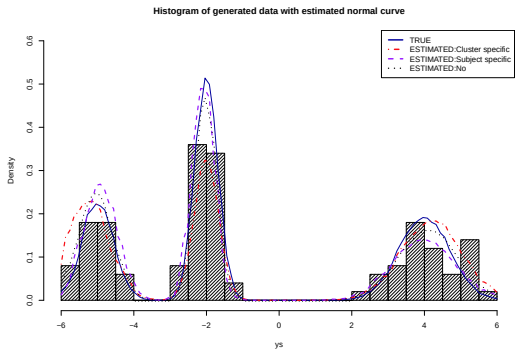
4. Simulation study

이 절에서는 실제 데이터에 적용하기 앞서 모의실험을 통해 군집 특정 변량효과에 대한 성능을 알아보고자 한다. 모의실험을 위해 다음과 같은 세 가지 경우의 모형에서 모의자료를 만들어낸다.

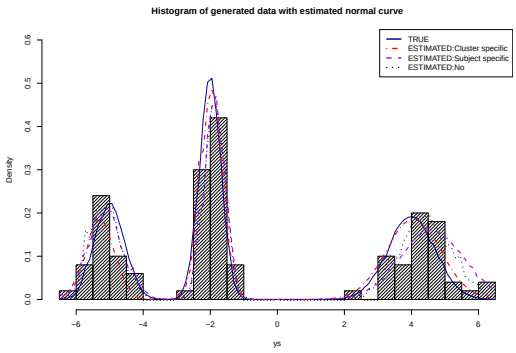
$$\begin{aligned}
&Y_i | z_i = k = x_i \boldsymbol{\beta}_k + \epsilon_{ik}, \\
&Y_i | z_i = k = x_i \boldsymbol{\beta}_k + \eta_i + \epsilon_{ik}, \\
&Y_i | z_i = k = x_i \boldsymbol{\beta}_k + \eta_k + \epsilon_{ik}, \quad \text{for } i = 1, \ldots, n,\ k = 1, \ldots, K,
\end{aligned}$$



(a) Case 1 (no random)



(b) Case 2 (subject-specific)



(c) Case 3 (cluster-specific)

Figure 4.1. Estimated curves of the Bayesian finite mixture models.

여기서 ϵ_{ik}는 $N(0, \sigma_k^2)$를 따르며 η_k는 군집 특정 변량효과, η_i는 개체 특정 변량효과이다. $K = 3$에서 $n = 100$개의 모의 자료를 생성하였다. X_1과 X_2를 각각 $N(-3, 0.01^2), N(2, 0.01^2)$ 생성하고 $\boldsymbol{X} = (1, X_1, X_2)$의 설계행렬을 고려하였다. 또한 각 군집 별로 참 $w_k, \boldsymbol{\beta}_k, \sigma_k^2$의 참값은

$$\begin{aligned}
&\text{Cluster 1} : w_1 = 0.3, \quad \boldsymbol{\beta}_1 = (0, 0, 2), \quad \sigma_1^2 = 0.5,\\
&\text{Cluster 2} : w_2 = 0.3, \quad \boldsymbol{\beta}_2 = (-1, 0, -2), \quad \sigma_2^2 = 0.2,\\
&\text{Cluster 3} : w_3 = 0.4, \quad \boldsymbol{\beta}_3 = (1, 1, 0), \quad \sigma_3^2 = 0.1
\end{aligned}$$

이고, 변량효과의 경우 $\eta_k \sim N(0, 0.01^2), \eta_i \sim N(0, 0.01^2)$에서 생성하였다. 분석을 위해 베이지안 추정값을 깁스 방법을 통해 50,000번의 반복 후 25,000개의 표본으로부터 계산 된 사후분포로부터의 추정값들을 사용하였다. $\boldsymbol{\beta_k|\sigma_k^2 \sim N(0, d\sigma_k^2 I)}$에서 d의 값은 10으로 고정하였다. 또한 군집 레이블은 3절에서 언급한 추정 단계를 반복 후 사후가능도함수를 계산하여 각 모형에 있어 사후가능도 함수 값 중 가장 큰 값을 갖는 경우를 선택하였다.

Figure 4.1은 앞서 설명한 세 가지 경우의 모의자료에서 각각 추정 모형에 변량효과가 없는 선형 모형, 개체 특정 변량효과가 있는 선형혼합모형, 군집 특정 변량효과가 있는 선형혼합모형에서 추정된 커브를 그린 것이다. Table 4.1은 각 모형들로부터 가장 큰 사후가능도함수 값에서 추정된 군집별 확률값이며 괄호의 값은 25,000개의 표본으로부터 계산된 각 군집별 확률값의 표준오차이다.

Table 4.1. Estimated component probabilities of the Bayesian finite mixture models

Cluster	Case1 (No RE)			Case2 (SS RE)			Case3 (CS RE)		
	1	2	3	1	2	3	1	2	3
True	0.3	0.3	0.4	0.3	0.3	0.4	0.3	0.3	0.4
No RE	0.37 (0.048)	0.25 (0.042)	0.38 (0.046)	0.30 (0.048)	0.34 (0.043)	0.36 (0.047)	0.27 (0.042)	0.36 (0.048)	0.37 (0.046)
Subject-specific RE	0.30 (0.047)	0.28 (0.043)	0.42 (0.048)	0.30 (0.047)	0.31 (0.043)	0.39 (0.049)	0.27 (0.043)	0.35 (0.048)	0.37 (0.046)
Cluster-specific RE	0.36 (0.048)	0.23 (0.043)	0.41 (0.047)	0.28 (0.048)	0.33 (0.043)	0.39 (0.046)	0.20 (0.042)	0.38 (0.047)	0.42 (0.048)

No RE = no random effect; SS RE = subject-specific random effect; CS RE = cluster-specific random effect.

Case 1의 경우, 실험 데이터는 변량효과가 없는 모형으로부터 생성하였다. Table 4.1에서 각 모형의 군집별 확률값과 표준오차 값을 살펴보면 표준오차 값은 세 모형 모두 비슷한 값이 확인된다. 각 군집별 확률값을 확인해 보면 두 개의 군집에 대해 변량효과가 없는 모형과 군집 특정 변량효과 모형의 경우 확률값이 실제 확률값과 다소 다른 값이 추정된 것을 확인할 수 있다. 그러나 Figure 4.1에서 첫 번째 커브를 포함한 히스토그램을 살펴보면 개체 특정 변량효과를 포함한 정규 혼합 선형 모형은 세 번째 그룹의 평균을 과소 추정한 것이 보여진다. 그러나 군집 특정 혼합모형과 변량효과가 없는 모형의 경우 실제 커브와 실제 평균에 가깝게 추정한 것이 관찰된다.

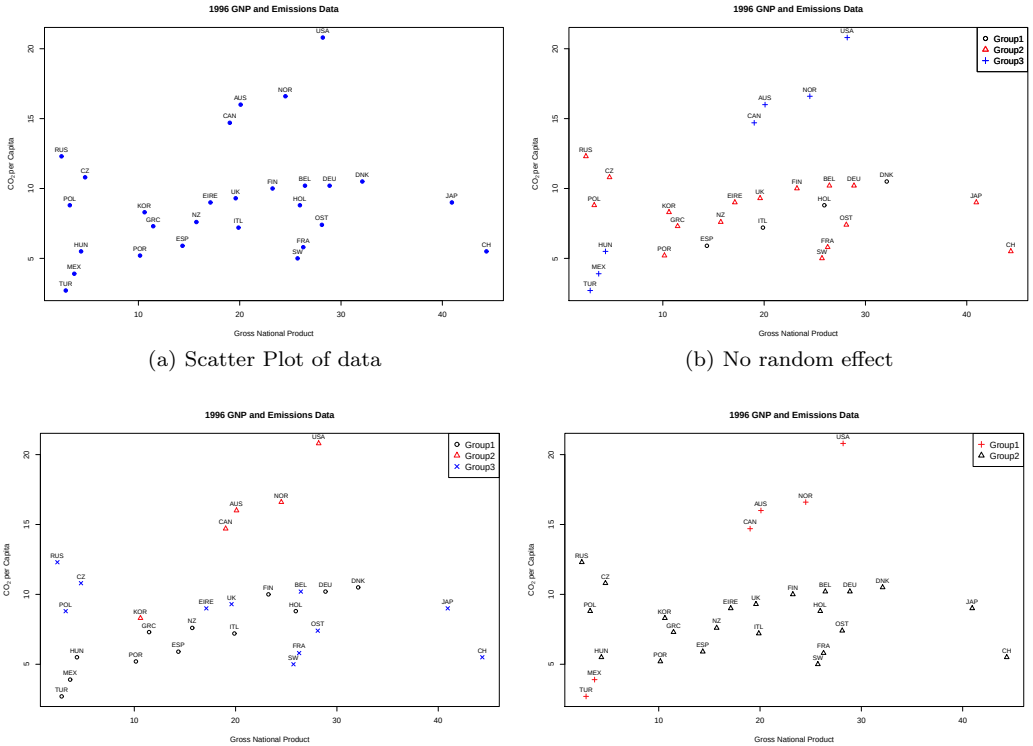
Case 2의 경우, 실험 데이터는 개체 특정 변량효과가 포함되어 있는 모형으로부터 생성하였다. Table 4.1에서 각 모형의 군집별 확률값과 표준오차 값을 살펴보면 표준오차 값은 세 모형 모두 비슷한 값이 확인된다. 각 군집별 확률값을 확인해 보면 실제 확률값에 개체 특정 변량효과 모형의 확률값이 가장 근접한 것이 확인된다. Figure 4.1에서 두 번째 그래프가 해당 데이터에 대한 히스토그램과 각 추정 모형에 대한 커브를 그린 것이다. 두 번째 군집의 경우 군집 특정 변량효과 모형이 실제 평균보다 과소 추정한 것이 확인된다. 또한 첫 번째와 세 번째 군집의 경우 군집 특정 변량효과 모형의 커브가 실제 커브에 비해 다소 분산이 큰 것처럼 관찰 된다.

Case 3의 실험 데이터는 군집 특정 변량효과가 포함되어 있는 모형으로부터 생성하였으며 Figure 4.1에서 세 번째 그래프에서 확인할 수 있다. Table 4.1에서 각 모형의 군집별 확률값과 표준오차 값을 살펴보면 표준오차 값은 세 모형 모두 비슷한 값이 확인된다. 각 군집별 확률값을 확인해 보면 실제 확률값에 각 모형의 추정된 확률값이 다소 다른 것이 보이나 그래프를 살펴보면 군집특정 변량효과가 포함된 모형이 가장 실제 평균과 가까운 것을 확인할 수 있다.

5. Data analysis

제안된 변량효과의 효용성을 판단하기 위해 실제 데이터에 모형을 적용하였다. Hurn 등 (2003)에서 사용된 CO2 데이터를 분석에 사용하였다. 해당 데이터는 1996년 28개의 나라에서 수집되었으며 국민총생산(gross national product; GNP)과 CO2 발생지수로 구성되어 있다. 본래 Hurn 등 (2003)에서는 해당 데이터를 활용하여 GNP가 낮은 국가의 발전을 돕기 위해 관련된 모형을 구성하고 활용에 필요한 그룹을 구성하고자 하였다.

Hurn 등 (2003)에서 CO2 데이터를 분석하기 위해 birth-and-death 과정을 기본으로 한 정규회귀(normal regression)식에서 군집 확률에 대한 사전분포로 포아송 분포, Poi($\lambda = 2$)를 사용하였다. 또한 분석에 활용한 모형은 본 연구에서 활용한 변량효과를 포함하지 않고 있으며 본 데이터에는 2개의



(a) Scatter Plot of data (b) No random effect

(c) Subject-specific random effect (d) Cluster-specific random effect

Figure 5.1. Representation of the allocation of the Bayesian finite mixture models to components.

군집이 가장 적합하다고 결론지었다. 이때 3개의 군집도 고려되었으나 세 번째 회귀선의 경우 그 가중치가 0.007로 매우 작아 무시될 정도였기 때문에 2개의 군집으로 결론지었다. 본 논문에서는 3개의 군집을 가정하고 분석을 시행하였다. 또한 Hurn 등 (2003)에서 분석에 활용한 모형은 본 연구에서 활용한 변량효과를 포함하지 않고 있다.

군집 특정 변량효과가 모수의 추정 및 군집 분류에 어떤 영향을 미치는지 확인하기 위해 개체 특정 변량효과 모형, 어떤 변량효과도 포함되지 않는 모형과 비교한다. 깁스 표본 추출방법을 50,000번 반복 후 후자 25,000개의 표본을 사후분포 추정에 사용한다. 25,000개의 표본 중 사후가능도함수의 값이 가장 큰 군집 구조를 선택한다.

Figure 5.1는 각 모형별로 추정한 것을 비교한 것이다. 첫 번째 그림은 변량효과가 없는 모형으로 관찰값들을 3개의 군집으로 분리하였다. USA, NOR, AUS, CAN, HUN, MEX, TUR을 한 군집으로, DNK, ITL, HOL, ESP를 군집으로 분리하고 나머지 국가를 또 다른 국가로 분리하였다. 개체특정 변량효과 모형의 경우 또한 3개의 군집으로 분리하였으나 분리된 군집에 속하는 국가의 차이가 나타나고 있음이 확인됐다. 그러나 군집특정 변량효과 모형의 경우 2군집으로만 관찰값들을 분리하였다. Table 5.1는 각 모형에 따른 군집과 BIC를 표기하였다.

모형을 비교하기 위하여 베이지안 정보 기준(BIC)를 고려하였다.

$$\mathrm{BIC} = -2L\left(\hat{\theta}\right) + 2k\ln\left(n\right),$$

Table 5.1. Estimated components of the cluster of the Bayesian finite mixture models

	Clust 1	Clust 2	Clust 3	BIC
No RE	DNK ITL HOL ESP	CAN MEX USA AUS HUN TUR NOR	NZ OST BEL CZ FIN FRA DEU GRC EIRE POL POR SW CH UK RUS KOR	137.28
Subject-specific RE	NZ DNK FIN DEU GRC MEX HUN ITL HOL POR ESP TUR	CAN USA KOR AUS NOR	JAP OST CZ FRA EIRE POL SW CH BEL UK RUS	209.97
Cluster-specific RE	CAN MEX USA AUS NOR TUR	JAP KOR NZ OST BEL CZ DNK FIN FRA DEU GRC HUN EIRE ITL HOL POL POR ESP SW CH UK RUS		16.76

여기서 $L(\hat{\theta})$는 로그사후가능도 함수값이고 k는 모수의 개수이다. 각 모델별 BIC를 비교 해 볼 때 변량효과가 없는 모형은 137.28, 개체특정 변량효과가 있는 모형은 209.97이다. 그리고 군집특정 변량효과가 있는 모형의 BIC 값은 16.76으로 세 모형 중 가장 작은 BIC값을 갖고 있으므로 BIC값을 기준으로 모형을 선택할 때 해당 데이터는 군집특정 변량효과를 포함한 모형을 선택하여 2개의 군집으로 분리하는 것이 가장 적합하다고 보여진다.

6. Discussion

이 논문에서는 군집 특정 변량효과를 포함한 정규 혼합 모형에 대해 깁스 표본법을 이용한 베이지안 분석에 대해 논술하였다. 군집을 구분 짓는데 활용되는 분류 가중치에 디리슈레 사전분포를 가정하여 섞는 정규혼합모형을 통해 모형의 모수들의 완전 조건 사후분포들을 도출하였다.

군집 특정 변량효과를 더한 베이지안 선형 혼합 모형은 데이터에 있어 겉으로 드러나지 않은 숨겨진 구조를 찾는데 도움을 준다. 군집 특정 변량효과는 각 군집 간의 독립을 가정하고 동시에 동일한 군집 안에서는 비슷한 특성을 공유한다는 것을 가정한다. 이러한 변량효과의 효용성을 판단하기 위해 객체 특정 변량효과가 포함된 모형의 경우와 어떤 변량효과도 가지고 있지 않는 모형을 비교하였다. 객체 특정 변량의효과는 각각의 개체가 다른, 이질적인 특성을 띄고 있다는 것을 가정한다.

모의실험에서는 세 가지 모형으로부터 데이터를 생성하여 변량효과가 없는 모형, 개체 특정 변량효과 모형, 군집 특정 변량효과 모형에서 모수들의 사후를 분포를 추정하여 관찰하였다. 각 모형들이 3개로 구성한 군집을 정확히 구분해 내었지만 각각의 커브를 살펴보면 평균값을 과대 혹은 과소 추정한 것이 관찰되었다. 이는 데이터구조가 갖는 오차에 의해 모형으로 추정할 수 있는 한계점으로 확인된다. 그러나 추정된 평균들의 전반적인 특징을 살펴보게 될 때 군집특정 변량효과 모형의 경우가 참모수 값에 가장 근접한 것이 확인된다. 결국 각각의 데이터가 가지고 있는 특성을 고려하여 필요한 변량효과를 추가하거나 모형의 변화를 통해 오차를 최소화하고 BIC 등의 기준값을 근거로 하여 모형을 선택하여야 함을

확인하였다.

마지막으로 제안한 모형을 실제 CO2 데이터에 적용하였다. 변량효과가 없는 모형과 개체특정 변량효과가 있는 모형의 경우 데이터를 3개의 군집으로 분리해 내었으나 군집 특정 변량효과 모형의 경우 2개의 군집으로 분리하였다. 또한 세 모형의 BIC를 비교했을 때 2개의 군집으로만 분리한 군집특정 변량효과를 포함한 모형이 가장 작은 값을 보여 해당 데이터는 2개의 군집으로 분리하는 것이 가장 적합한 것으로 판단되었다. 이는 Hurn 등 (2003)와 같은 결론을 보여주고 있으며 다른 모형들과의 차이는 변량효과에 의해 발생한 부분으로 확인할 수 있다.

이 논문에서는 군집의 수를 가정 즉, 군집의 수를 유한개(finite)로 한정해 놓고 모의실험 및 데이터 분석을 시행하였다. 하지만 일반적으로 분석을 위해 접하는 자료의 경우 이러한 군집의 수가 알려지지 않은 경우가 대부분으로 최대 군집의 수를 한정하는 것에 어려움이 따를 수밖에 없다. 따라서 이러한 군집수를 유한개로 한정하는 것이 아닌 무한개(infinite) 설정 후 추정해 나가는 방법으로의 확장이 추가적인 연구주제가 될 수 있을 것이다. 또한 2절에서 언급한바와 같이 본 방법론은 모형의 군집 분리 성능에 목표를 두고 있다. 따라서 β자체에 관심이 있거나 변량효과에 흥미가 있다면 추정된 사후확률 값으로 각 군집을 분리한 후 각 군집에 속하는 관찰값들을 이용하여 군집별 모형으로부터 β값을 추정할 수 있을 것이다. 하지만 이러한 사전분포에 대한 가정에 있어 공액사전분포가 아닌 다른 사전분포를 가정함으로써 또 다른 결론을 얻을 수 있을 것으로 생각되며 사전분포의 변형 역시 또 다른 연구주제가 될 수 있을 것이다.

References

Banfield, J. D. and Raftery A. E. (1993). Model-based Gaussian and non-Gaussian clustering, *Biometrics*, **49**, 803–821.

Bensmail, H., Celeux, G., Raftery, A. E., and Robert, C. P. (1997). Inference in modelbased cluster analysis, *Statistics and Computing*, **7**, 1–10.

Bernardo, J. M. and Giró n, F. J. (1988). A Bayesian analysis of simple mixture problems. In *Bayesian Statistics 3*, Bernardo, J. M., DeGroot, M. H., Lindley, D. V., and Smith, A. F. M. (Eds), Clarendon, New York, 67–78.

Cao, G. and West, M. (1996). Practical Bayesian inference using mixtures of mixtures, *Biometrics*, **52**, 1334–1341.

Carlin, B. and Chib, S. (1995). Bayesian model choice via Markov chain Monte Carlo methods, *Journal of the Royal Statistical Society Series B (Methodological)*, **57**, 473–484.

Dasgupta, A. and Raftery, A. E. (1998). Detecting features in spatial point processes with clutter via model-based clustering, *Journal of the American Statistical Association*, **93**, 294–302.

Dellaportas, P. (1998). Bayesian classification of neolithic tools, *Applied Statistics*, **47**, 279–297.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society Series B (Methodological)*, **39**, 1–38.

De Veaux, R. D. (1989). Mixtures of linear regressions, *Journal Computational Statistics & Data Analysis*, **8**, 227–245.

Diebolt, J. and Robert, C. (1990). Bayesian estimation of finite mixture distributions, part ii: sampling implementation, *Technical Report 111, LSTA*, Université Paris VI, Paris.

Diebolt, J. and Robert, C. (1994). Estimation of finite mixture distributions through Bayesian sampling, *Journal of the Royal Statistical Society Series B (Methodological)*, **56**, 363–375.

Escobar, M. and West, M. (1995). Bayesian density estimation and inference using mixtures, *Journal of the American Statistical Association*, **90**, 577–588.

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation, *Journal of the American Statistical Association*, **97**, 611–631.

Frühwirth-Schnatter, S. (2005). *Finite Mixture and Markov Switching Models*, Springer Science & Business

Media, New York.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian Data Analysis* (pp. 526), Chapman and Hall, Boca Raton, FL.

Geoffery, M. and David, P. (2000). *Finite Mixture Models*, John Wiley & Sons, New York.

Hurn, M., Justel, A., and Robert, C. P. (2003). Estimating mixtures of regressions, *Journal of Computational and Graphical Statistics*, **12**, 55–79.

Kyung, M. (2015). Dirichlet process mixtures of linear mixed regressions, *Communications for Statistical Applications and Methods*, **22**, 625–637.

McLachlan, G. J. and Basford, K. E. (1988). *Mixture Models: Inference and Applications to Clustering*, Marcel Dekker, New York.

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*, Wiley, New York.

Mengersen, K. and Robert, C. (1996). Testing for mixtures: a Bayesian entropic approach. In *Bayesian Statistics 5, Proceedings of the Fifth Valencia International Meeting*, Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M. (Eds), Oxford University Press, Oxford, 255–276.

Phillips, D. B. and Smith, A. F. M. (1996). Bayesian model comparison via jump diffusions. In *Markov Chain Monte Carlo in Practice*, Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (Eds), Chapman and Hall, London, 215–239.

Quandt, R. E. (1958). The estimation of the parameters of a linear regression system obeying two separate regimes, *Journal of the American Statistical Association*, **53**, 873–880.

Quandt, R. E. and Ramsey, J. B. (1978). Estimating mixtures of normal distributions and switching regressions, *Journal of the American Statistical Association*, **73**, 730–738.

Raftery, A. E. (1996). Approximate Bayes factors and accounting for model uncertainty in generalized linear models, *Biometrika*, **83**, 251–266.

Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion), *Journal of the Royal Statistical Society B (Statistical Methodology)*, **59**, 731–792.

Robert, C. P. (1996). Mixtures of distributions: inference and estimation. In *Markov Chain Monte Carlo in Practice*, Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (Eds), Chapman and Hall, London, 441–464.

Robert, C. P. and Mengersen, K. L. (1999). Reparameterization issues in mixture modelling and their bearings on MCMC algorithms, *Computational Statistics and Data Analysis*, **29**, 325–343.

Roeder, K. and Wasserman, L. (1997). Practical density estimation using mixtures of normals, *Journal of the American Statistical Association*, **92**, 894–902.

Scott, A. J. and Symons, M. J. (1971). Clustering methods based on likelihood ratio criteria, *Biometrics*, **27**, 387–389.

Smith, A. E. M. and Roberts, G. O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods, *Journal of the Royal Statistical Society Series B (Methodological)*, **55**, 3–23.

Vounatsou, P., Smith, T., and Smith, A. F. M. (1998). Bayesian analysis of two-component mixture distributions applied to estimating malaria attributable fractions, *Applied Statistics*, **47**, 575–587.

West, M. (1992). Modelling with mixtures. In *Bayesian Statistics 4, Proceedings of the Fourth Valencia International Meeting*, Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M. (Eds), Oxford University Press, Oxford, 503–524.

West, M., Müller, P., and Escobar, M. D. (1994). Hierarchical priors and mixture models with application in regression and density estimation. In *Aspects of Uncertainty: A tribute to D.V. Lindley, Smith, A. F. M. and Freeman, P. (Eds)*, Wiley, New York, 363–386.

Yu, J. Z. and Tanner, M. A. (1999). An analytical study of several Markov chain Monte Carlo estimators of the marginal likelihood, *Journal of Computational and Graphical Statistics*, **8**, 839–853.

군집 특정 변량효과를 포함한 유한 혼합 모형의 베이지안 분석

이혜진[a] · 경민정[a,1]

[a]덕성여자대학교 통계학과

(2016년 9월 22일 접수, 2016년 11월 28일 수정, 2016년 12월 20일 채택)

요 약

대량의 데이터에 있어 전반적인 특성 및 구조를 파악하는데 유용하기 때문에 다양한 분야에서 군집분석을 사용하고 있다. Dempster 등 (1977)에서 정의된 expectation-maximization(EM) 알고리즘은 가장 보편적으로 사용되는 군집분석 방법이다. 선형모형의 유한혼합물(finite mixture of linear model) 기법 또한 군집분석 방법 중 많이 사용되는 방법이며 베이지안 군집방법은 Bernardo와 Giron (1988)이 군집에 대한 가중치 확률만 모를 경우 처음 적용하였다. 우리는 이 연구에서 일반적인 선형모형의 유한혼합물이 아닌 군집특정(cluster-specific) 변량효과를 모형에 포함하여 베이지안 분석방법인 깁스표집법(Gibbs sampling)을 사용한다. 제안한 모형의 특성 및 표집법에 대하여 설명하였고 모의실험 및 실제 데이터 분석을 통하여 모형의 유용성을 파악하였다. Hurn 등 (2003)의 CO2 데이터에 모형을 적용하여 변량효과가 없는 모형, 개체특정(subject-specific) 변량효과 모형과 비교하였다.

주요용어: 군집분석, 유한혼합물모형, 군집특정변량효과, 깁스표본추출법

이 논문은 2016년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. NRF-2015R1C1A1A01051837).

[1]교신저자: (01369) 서울 도봉구 삼양로144길 33, 덕성여자대학교 통계학과. E-mail: mkyung@duksung.ac.kr