The Korean Journal of Applied Statistics (2016) DOI: http://dx.doi.org/10.5351/KJAS.2016.29.3.513
29(3), 513–523

An approximate fitting for mixture of multivariate skew normal distribution via EM algorithm

Seung-Gu Kim[a,1]

[a]Department of Data and Information, Sangji University

(Received February 19, 2016; Revised March 5, 2016; Accepted March 7, 2016)

Abstract

Fitting a mixture of multivariate skew normal distribution (MSNMix) with multiple skewness parameter vectors via EM algorithm often requires a highly expensive computational cost to calculate the moments and probabilities of multivariate truncated normal distribution in E-step. Subsequently, it is common to fit an asymmetric data set with MSNMix with a simple skewness parameter vector since it allows us to compute them in E-step in an univariate manner that guarantees a cheap computational cost. However, the adaptation of a simple skewness parameter is unrealistic in many situations. This paper proposes an approximate estimation for the MSNMix with multiple skewness parameter vectors that also allows us to treat them in an univariate manner. We additionally provide some experiments to show its effectiveness.

Keywords: multivariate skew normal distribution, mixture model, EM algorithm, multivariate normal cdf

1. 서론

많은 응용분야에서 비대칭 자료를 반영하는 통계적 모형에 대한 요구가 증가하고 있다. 특히 Cytometry와 같은 의생학 분야에서는 구조적으로 발생할 수 밖에 없는 비대칭 자료에 대해 군집분석을 수행해야 한다. 이에 따라 최근 비대칭 다변량 자료들의 군집분석을 위해 다변량 치우친 분포의 혼합모형에 대한 연구가 활발하다. 다변량 치우친 분포 혼합모형에 대한 연구사를 간략히 요약하면 다음과 같다.

Azzalini (1985) 및 Azzalini와 Dalla-Valle (1996)에 의해 multivariate skew normal distribution(MSN)이 소개된 이후 Sahu 등 (2003)에 의해 multivariate skew t-distribution(MST)가 소개되었으며, Arellano-Valle와 Genton (2005)은 보다 일반적 형태의 치우침 모수(skewness parameter)를 가지는 다변량 canonical fundamental skew t-distribution(CFUST)를 소개 하였다. Lin (2010)은 Sahu 등 (2003)의 MST 혼합모형을 적합하기 위해 Monte Carlo EM 알고리즘(MC-EM) 개발하였는데, 이 방법은 E-step에서 다변량 절단 t-분포의 적률에 대해 모의연구를 수행하므로 막대한 처리시간을 요한다. 이 문제를 개선하기 위해 Lee와 McLachlan (2013, 2014a)은 다변량 절단분포의 적률을 명시적 수식으로 표현한 정확한 EM 알고리즘(Exact-EM)을 개발하였다. 또한 Lee와 McLachlan (2014b)은 다변량 CFUST 혼합모형을 위한 Exact-EM을 개발하였다. 그러나 Exact-EM 알고리즘 역시 다중 치우

This work was supported by the research grant of the Sangji University in 2014.
[1]Department of Data and information, Sangji University, 83 Sangjidae-gil, Wonju-si, Gangwon-do 26339, Korea. E-mail: sgukim@sangji.ac.kr

침 모수를 고려한 모형을 적용할 경우 E-step에서 다변량 절단분포의 확률과 적률을 계산하는데 매우 큰 시간을 요한다. 이러한 문제는 보다 일반적인 모형인 CFUST 혼합모형을 적용하는데 현실적인 어려움을 가지게 된다. 예를 들어 약 200개의 관측치를 가지는 자료에 5개 치우침 모수를 가진 모형을 적합하는데, 일반적인 PC로 EM 알고리즘의 1 반복 당 10분 이상이 소요된다. 따라서 100회 반복을 수행하는데 약 1.5일 이상이 소요될 것으로 예상된다. 이러한 계산적 문제 때문에 다중 치우침 모수 대신 단순 치우침 모수를 가지는 MSN 혹은 MST 혼합모형, 예를 들면 Pyne 등 (2009) 혹은 Cabral 등 (2012)의 모형을 사용할 수 밖에 없게 한다. 이와같은 제약된 모형은 매우 빠른 처리시간을 보장하기 때문이다. 그러나 현실에서는 반드시 다중 치우침 모수를 사용해야 하는 경우도 있는 만큼, 이에 대한 계산 속도의 개선이 반드시 필요한 시점이다. Kim (2014)는 보다 빠른 CFUST 혼합모형의 적합을 위한 방법을 제공하였으나 제안된 EM 알고리즘은 다중 치우침 모수를 가지는 상황에서 불안정한 상태를 보였다. 그 외에는 저자가 아는 한 계산시간 개선을 위한 어떠한 연구도 없는 것으로 보인다.

본 논문에서는 canonical fundamental skew normal distribution(CFUSN) 형식의 다중 치우침 모수를 가지는 MSN 혼합모형의 적합 시 처리시간을 개선하기 위한 근사적인 방법을 제공한다. 이 모형을 Exact-EM으로 적합할 때 여러번의 다변량 누적분포함수를 계산해야 하는데, 본 논문에서는 처리시간의 주된 요인은 바로 다변량 누적분포함수의 계산시간에 있다고 보고, 이에 대해 Olson과 Weissfeld (1991)의 근사 계산법을 적용하여 모형 적합의 계산시간을 얼마나 단축시킬 수 있는지 실험해 보고자 한다.

다음 절에서는 CFUSN 혼합모형과 Exact-EM 알고리즘을 간략히 요약하며, 3절에서는 Olson과 Weissfeld (1991)의 다변량 정규분포 누적분포함수의 계산방법을 설명하며 이를 바탕으로 근사적인 방법을 제안한다. 4절에서는 제안된 방법을 Exact-EM 방법과 비교하여 실험한다. 5절에서는 결론을 정리하고 추후 연구할 만한 과제를 제안한다.

2. MSN 혼합모형 및 Exact-EM

2.1. MSNMix의 정의

p-변량 관측치 $\boldsymbol{y}_j = (y_{1j}, \ldots, y_{pj})^T$, $(j = 1, \ldots, n)$에 대해 g-성분 혼합모형

$$f\left(\boldsymbol{y}_j; \boldsymbol{\Theta}\right) = \sum_{i=1}^{g} \pi_i f_i\left(\boldsymbol{y}_j; \boldsymbol{\theta}_i\right), \quad j = 1, \ldots, n \tag{2.1}$$

을 고려하자. 여기서 π_i는 i번째 성분의 혼합비율이며,

$$f_i\left(\boldsymbol{y}_j; \boldsymbol{\theta}_i\right) = 2^q \phi_p\left(\boldsymbol{y}_j; \boldsymbol{\mu}_i, \boldsymbol{\Omega}_i\right) \Phi_q\left(\tilde{\boldsymbol{x}}_{ij}; \boldsymbol{\Psi}_i\right), \quad i = 1, \ldots, g \tag{2.2}$$

은 q-중 치우침 모수(q-ple skewness parameter) $\boldsymbol{\Delta}_i = (\boldsymbol{\delta}_{i1}, \ldots, \boldsymbol{\delta}_{iq})$를 가지는 MSN 밀도로서, $\boldsymbol{\delta}_{ik} = (\delta_{ik1}, \ldots, \delta_{ikp})^T$는 $(p \times 1)$ 벡터이며, $\boldsymbol{\Omega}_i = \boldsymbol{\Sigma}_i + \boldsymbol{\Delta}_i \boldsymbol{\Delta}_i^T$, $\tilde{\boldsymbol{x}}_{ij} = \boldsymbol{\Delta}_i^T \boldsymbol{\Omega}_i^{-1} (\boldsymbol{y}_j - \boldsymbol{\mu}_i)$, $\boldsymbol{\Psi}_i = \boldsymbol{I}_q - \boldsymbol{\Delta}_i^T \boldsymbol{\Omega}_i^{-1} \boldsymbol{\Delta}_i$이며, $\phi_p(\cdot; \boldsymbol{m}, \boldsymbol{S})$는 평균 벡터 $\boldsymbol{m}$ 및 공분산 행렬 $\boldsymbol{S}$를 가지는 p-변량 정규분포밀도이며, $\Phi_q(\boldsymbol{x}; \boldsymbol{S})$는 평균 벡터 $\boldsymbol{0}$ 및 공분산 행렬 $\boldsymbol{S}$를 가지는 q-변량 정규분포 누적분포함수를 나타낸다. 그리고 $\boldsymbol{\theta}_i$는 i번째 성분밀도의 모수들을 원소로하는 벡터이며, $\boldsymbol{\Theta}$는 모형에 포함된 모든 모수를 포함하는 벡터이다. 이때 모형 (2.1)을 q-중 치우침 모수를 가지는 MSN의 혼합모형이라 정의하며 MSNMix_q로 표시하자.

2.2. Exact EM 알고리즘

우리의 목표는 관측치 $\boldsymbol{y} = (\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_n^T)^T$에 대한 로그-우도

$$L(\boldsymbol{\Theta}) = 2^q \sum_{j=1}^{n} \log \left\{ \sum_{i=1}^{g} \pi_i \phi_p(\boldsymbol{y}_j; \boldsymbol{\mu}_i, \boldsymbol{\Omega}_i) \Phi_q\left(\tilde{\boldsymbol{x}}_{ij}; \boldsymbol{\Psi}_i\right) \right\}$$

를 최대화하는 $\boldsymbol{\Theta}$를 얻는 것이다. 그러나 직접 최대화는 거의 불가능하므로 EM 알고리즘을 이용하는 것이 일반적이다.

모형 식 (2.1)의 자료 $\boldsymbol{y}_j = (y_{1j}, \ldots, y_{pj})^T$의 생성을 위한 위계적 구조는 다음과 같다. 즉, 관측치 $\boldsymbol{y}_j$가 i번째 성분으로부터 왔다면 $Z_{ij} = 1$ 그렇지 않으면 0을 나타내는 미관측 성분지시변수를 추가함으로써

$$\begin{aligned}
&\boldsymbol{Y}_j = \boldsymbol{y}_j | (z_{ij} = 1, \boldsymbol{X}_j = \boldsymbol{x}_j) \sim N_p(\boldsymbol{\mu}_i + \boldsymbol{\Delta}_i \boldsymbol{x}_j, \boldsymbol{\Sigma}_i), \\
&\boldsymbol{X}_j = \boldsymbol{x}_j | z_{ij} = 1 \sim N_q(\boldsymbol{0}, \boldsymbol{I}), \\
&\boldsymbol{Z}_j = (Z_{1j}, \ldots, Z_{gj})^T \sim \text{multinomial}(1, \pi_1, \ldots, \pi_g)
\end{aligned}$$

와 같이 나타낼 수 있다. 이로부터 완비자료(complete data) $(\boldsymbol{y}_j, \boldsymbol{x}_j, \boldsymbol{z}_j)(j = 1, \ldots, n)$의 로그-우도는 (상수항은 제외하고)

$$L_c(\boldsymbol{\Theta}) = \sum_{j=1}^{n} \sum_{i=1}^{g} z_{ij} \left[\log \pi_i - \frac{1}{2} \log |\boldsymbol{\Sigma}| - \frac{1}{2} (\boldsymbol{y}_j - \boldsymbol{\mu} - \boldsymbol{\Delta} \boldsymbol{x}_j)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{y}_j - \boldsymbol{\mu} - \boldsymbol{\Delta} \boldsymbol{x}_j) \right] \tag{2.3}$$

이며, EM 알고리즘은 $(t+1)$번째 반복에서 관측 불완비자료 $\boldsymbol{y}$에 대한 완비자료 로그-우도의 조건부 기대값 $Q(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(t)}) = E[L_c(\boldsymbol{\Theta})|\boldsymbol{y}, \boldsymbol{\Theta}^{(t)}]$를 최대화 한다. 여기서 $\boldsymbol{\Theta}^{(t)}$는 (t)번째 반복에서 얻은 추정치이다.

이 문제는 결국 E-step에서 미관측 확률변수 Z_{ij}와 $\boldsymbol{X}_j$에 대한 조건부 기대값

$$\begin{aligned}
\tau_{ij}^{(t+1)} &= E\left[Z_{ij} | \boldsymbol{y}_j, \boldsymbol{\Theta}^{(t)} \right] \\
&= \frac{\phi_p\left(\boldsymbol{y}_j; \boldsymbol{\mu}_i^{(t)}, \boldsymbol{\Omega}_i^{(t)}\right) \Phi_q\left(\tilde{\boldsymbol{x}}_{ij}^{(t)}; \boldsymbol{\Psi}_i^{(t)}\right)}{\sum_{h=1}^{g} \phi_p\left(\boldsymbol{y}_j; \boldsymbol{\mu}_h^{(t)}, \boldsymbol{\Omega}_h^{(t)}\right) \Phi_q\left(\tilde{\boldsymbol{x}}_{hj}^{(t)}; \boldsymbol{\Psi}_h^{(t)}\right)}, \qquad (2.4) \\
\boldsymbol{e}_{ij}^{(t+1)} &= E\left[\boldsymbol{X}_j | \boldsymbol{y}_j, Z_{ij} = 1, \boldsymbol{\Theta}^{(t)} \right], \qquad (2.5) \\
\boldsymbol{E}_{ij}^{(t+1)} &= E\left[\boldsymbol{X}_j \boldsymbol{X}_j^T | \boldsymbol{y}_j, Z_{ij} = 1, \boldsymbol{\Theta}^{(t)} \right], \quad i = 1, \ldots, g;\ j = 1, \ldots, n \qquad (2.6)
\end{aligned}$$

을 계산하는 것이다. 여기서 $\boldsymbol{X}_j$에 대한 1, 2차 적률인 $\boldsymbol{e}_{ij}^{(t+1)}$과 $\boldsymbol{E}_{ij}^{(t+1)}$는 $q = \dim(\boldsymbol{X}_j) > 1$이면 명시적인 수식으로 표현할 수 없다. 다만,

$$\boldsymbol{X}_j | (\boldsymbol{y}_j, Z_{ij} = 1) \sim TN_q\left(\tilde{\boldsymbol{x}}_{ij}^{(t+1)}, \boldsymbol{\Psi}_i^{(t+1)} | (0, \infty)^q\right)$$

즉 양의 영역에 지지역(support)을 가지는 평균벡터 $\tilde{\boldsymbol{x}}_{ij}^{(t)} = \boldsymbol{\Delta}_i^{(t)^T} \boldsymbol{\Omega}_i^{(t)^{-1}} (\boldsymbol{y}_j - \boldsymbol{\mu}_i^{(t)})$ 및 공분산 행렬 $\boldsymbol{\Psi}_i^{(t)} = \boldsymbol{I}_q - \boldsymbol{\Delta}_i^{(t)^T} \boldsymbol{\Omega}_i^{(t)^{-1}} \boldsymbol{\Delta}_i^{(t)}$인 q-변량 절단 정규분포를 따른다. 그래서 본 논문에서는 절단 정규분포 확률변수 $\boldsymbol{X}_j$의 두 적률을 Ho 등 (2012)의 루틴을 이용하여 계산하였다.

M-step에서는 이를 바탕으로 혼합비율 및 평균벡터 추정치

$$\pi_i^{(t+1)} = \frac{1}{n}\sum_{j=1}^{n}\tau_{ij}^{(t)}, \qquad \boldsymbol{\mu}_i^{(t+1)} = \frac{1}{n_i^{(t+1)}}\sum_{j=1}^{n}\left(\boldsymbol{y}_j - \boldsymbol{\Delta}_i^{(t)}\boldsymbol{e}_{ij}^{(t)}\right), \quad i = 1, \ldots, g$$

및 공분산 행렬 추정치

$$\begin{aligned}\boldsymbol{\Sigma}_i^{(t+1)} = \frac{1}{n_i^{(t+1)}}\sum_{j=1}^{n}\tau_{ij}^{(t+1)}\Big[&\tilde{\boldsymbol{y}}_{ij}^{(t+1)}\tilde{\boldsymbol{y}}_{ij}^{(t+1)^T} - \tilde{\boldsymbol{y}}_{ij}^{(t+1)}\boldsymbol{e}_{ij}^{(t+1)^T}\boldsymbol{\Delta}_i^{(t)^T} - \boldsymbol{\Delta}_i^{(t)}\boldsymbol{e}_{ij}^{(t+1)}\tilde{\boldsymbol{y}}_{ij}^{(t+1)^T} \\ &+\boldsymbol{\Delta}_i^{(t)}\boldsymbol{E}_{ij}^{(t+1)}\boldsymbol{\Delta}_i^{(t)^T}\Big], \quad i = 1, \ldots, g\end{aligned}$$

그리고 치우침 모수 추정치

$$\boldsymbol{\Delta}_i^{(t+1)} = \left[\sum_{j=1}^{n}\tau_{ij}^{(t+1)}\tilde{\boldsymbol{y}}_{ij}^{(t+1)}\boldsymbol{e}_{ij}^{(t+1)^T}\right]\left[\sum_{j=1}^{n}\tau_{ij}^{(t+1)}\boldsymbol{E}_{ij}^{(k+1)}\right]^{-1}, \quad i = 1, \ldots, g$$

를 갱신하는 것으로 귀결된다. 여기서 $n_i^{(t+1)} = \sum_{j=1}^{n}\tau_{ij}^{(t+1)}$ 그리고 $\tilde{\boldsymbol{y}}_{ij}^{(t+1)} = \boldsymbol{y}_j - \boldsymbol{\mu}_i^{(t+1)}$을 나타낸다. 이 EM 알고리즘의 M-step을 수행하는데 처리시간이 오래 걸리는 부분은 없다. 그러나 본 논문의 주된 관심사인 E-step은 q가 크면 비현실적인 처리시간이 소요된다. 저자가 추적한 바에 의하면 그 이유는 오직 q-변량 누적분포함수의 계산시간에 소요되는 처리시간 때문이다. 식 (2.4)의 사후확률 추정치 $\tau_{ij}^{(t+1)}$를 계산하는데 1개의 관측치 당 g번의 $\Phi_q(\tilde{\boldsymbol{x}}_{ij}^{(t)}; \boldsymbol{\Psi}_i^{(t)})$를 계산해야 한다. 그리고 두 적률 $\boldsymbol{e}_{ij}^{(t+1)}$과 $\boldsymbol{E}_{ij}^{(t+1)}$를 계산하기 위한 Ho 등 (2012)의 루틴에서는 1개의 관측치 당 g번의 q-변량 누적분포함수, $g(q-1)$번의 $(q-1)$-변량 누적분포함수 그리고 $gq(q+1)/2$번의 $(q-2)$-변량 누적분포함수의 계산이 필요하다. 이런 이유로 만약 성분의 개수 $g > 1$이며, $q > 2$이고 $n > 100$이면, MSNMix$_q$ 모형에 대한 비실용성 논란을 일으킬 만한 처리시간이 걸리게 된다. 결국 MSNMix$_q$ 모형을 적합하기 위해서는 다변량 정규분포 누적분포함수를 (일반적인 PC 수준에서) 빠르게 계산하는 방법을 찾아야 하는 원초적인 문제에 봉착한다.

만약 다변량 정규분포 누적분포함수의 결과가 근사적일지라도 MSNMix$_q$ 모형을 빠르게 적합할 수 있고, 그 적합 결과가 정확한 결과와 아주 큰 차이를 보이지 않는다면 현실적인 측면에서 받아들일 수 있을 것이다. 저자는 이 문제에 대한 관련 문헌을 찾던 중 비교적 오래된 그러나 이 분야에서 (저자가 아는 한) 전혀 인용된 바 없는 꽤 유익한 논문 Olson과 Weissfeld (1991)을 발견하였다. 다음 절에서는 이 논문에서 제공하는 방법을 보다 일반적 형태로 소개하며, 정규분포 누적분포함수에 특정한 수식을 제공할 것이며, 정확성을 높이기 위한 한 가지 트릭을 제안한다.

3. 근사적 정규분포 누적분포함수

Olson과 Weissfeld (1991)은 다중적분에 대한 한 가지 근사적인 방법을 소개하였다. 여기서는 그들의 방법을 따라 다변량 누적분포함수의 근사공식을 제공한다. 먼저 표기의 간편함을 위해 q-변량 정규분포 밀도 $\phi_q(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$와 누적분포함수 $\Phi_q(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$를 각각 $\phi_q(\boldsymbol{x})$와 누적분포함수 $\Phi_q(\boldsymbol{x})$로서 나타내자. 그리고 X_i의 단변량주변밀도와 단변량누적분포함수를 $\phi_1(x_i)$와 누적분포함수 $\Phi_1(x_i)$로 나타내자.

우선 본 연구에서는 확률변수 X의 어떤 (연속) 함수 $h(X)$에 대해

$$E(h(X)) \approx h(E(X)) \tag{3.1}$$

와 같은 근사를 고려한다. 식 (3.1)은 함수 $h(X)$의 기대값에 대한 $\eta = E(X)$에서의 1차 Taylor 근사 $E(h(X)) \approx h(\eta) + h^{[1]}(\eta)E(X-\eta) = h(\eta)$에 바탕을 두고 있다. 만약 좀 더 정확한 근사를 원한다면, Taylor 전개식의 추가적인 항을 고려할 수 있을 것이지만, 본 논문에서는 연구의 목적 상 1차 Taylor 근사만을 사용할 것이다. 실제 확률변수 X 밀도가 정규분포와 같이 두꺼운 꼬리를 가지지 않는다면 식 (3.1)은 꽤 좋은 근사를 제공한다는 점에 주목한다.

이제 식 (3.1)을 적용하여

$$\begin{aligned}
\Phi_q(\boldsymbol{x}) &= \int_{-\infty}^{x_1}\int_{-\infty}^{x_2}\cdots\int_{-\infty}^{x_q}\phi_q(t_1,t_2,\ldots,t_q)\,dt_q\cdots dt_2dt_1 \\
&= \int_{-\infty}^{x_1}\phi_1(t_1)\int_{-\infty}^{x_2}\cdots\int_{-\infty}^{x_q}\phi_{q-1}(t_2,\ldots,t_q|t_1)\,dt_q\cdots dt_2dt_1 \\
&= \Phi_1(x_1)\int_{-\infty}^{x_1}\frac{\phi_1(t_1)}{\Phi_1(x_1)}\int_{-\infty}^{x_2}\cdots\int_{-\infty}^{x_q}\phi_{q-1}(t_2,\ldots,t_q|t_1)\,dt_q\cdots dt_2dt_1 \\
&= \Phi_1(x_1)E\left[\int_{-\infty}^{x_2}\cdots\int_{-\infty}^{x_q}\phi_{q-1}(t_2,\ldots,t_q|X_1)\,dt_q\cdots dt_2\middle| X_1\le x_1\right] \\
&\approx \Phi_1(x_1)\int_{-\infty}^{x_2}\cdots\int_{-\infty}^{x_q}\phi_{q-1}(t_2,\ldots,t_q|\eta_1)\,dt_q\cdots dt_2 \qquad (3.2)
\end{aligned}$$

이 됨을 알 수 있다. 여기서 $\eta_1 = E(X_1|X_1\le x_1) = \mu_1 - \sigma_1\phi_1(w_1)/\Phi_1(w_1)$로서 하방절단 일변량 정규분포의 평균이며, $w_1 = (x_1-\mu_1)/\sigma_1$을 나타낸다.

그리고 식 (3.2) 우변의 다중적분항에 대해 비슷한 과정을 적용하면

$$\begin{aligned}
&\int_{-\infty}^{x_2}\int_{-\infty}^{x_3}\cdots\int_{-\infty}^{x_q}\phi_{q-1}(t_2,t_3,\ldots,t_q|\eta_1)\,dt_q\cdots dt_3dt_2 \\
&= \Phi_1(x_2|\eta_1)\int_{-\infty}^{x_2}\frac{\phi_1(t_2|\eta_1)}{\Phi_1(x_2|\eta_1)}\int_{-\infty}^{x_3}\cdots\int_{-\infty}^{x_q}\phi_{q-2}(t_3,\ldots,t_q|\eta_1,t_2)\,dt_q\cdots dt_3dt_2 \\
&= \Phi_1(x_2|\eta_1)E\left[\int_{-\infty}^{x_3}\cdots\int_{-\infty}^{x_q}\phi_{q-2}(t_3,\ldots,t_q|X_2,\eta_1)\,dt_q\cdots dt_3\middle| X_2\le x_2\right] \\
&\approx \Phi_1(x_2|\eta_1)\int_{-\infty}^{x_3}\cdots\int_{-\infty}^{x_q}\phi_{q-2}(t_3,\ldots,t_q|\eta_1,\eta_2)\,dt_q\cdots dt_3 \qquad (3.3)
\end{aligned}$$

이 된다. 여기서 $\eta_2 = E(X_2|X_1=\eta_1, X_2\le x_2) = \mu_{2|1} - \sigma_{2|1}\phi_1(w_2)/\Phi_1(w_2)$로서 하방절단 일변량 정규분포의 평균이며, $w_2 = (x_1-\mu_{2|1})/\sigma_{2|1}$ 그리고 $\mu_{2|1} = E(X_2|X_1=\eta_1)$ 및 $\sigma^2_{2|1} = \mathrm{var}(X_2|X_1=\eta_1)$을 나타낸다.

식 (3.3) 이후 비슷한 과정을 반복하면 결국

$$\Phi_q(\boldsymbol{x}) \approx \Phi_1(x_1)\Phi_1(x_2|\eta_1)\Phi_1(x_3|\eta_1,\eta_2)\times\cdots\times\Phi_1(x_q|\eta_1,\ldots,\eta_{q-1}) \stackrel{\text{let}}{=} \Phi^*_q(\boldsymbol{x}) \qquad (3.4)$$

와 같은 일반식을 얻는다. 여기서 $\Phi_1(x_i|\eta_1,\ldots,\eta_{i-1})$은 일변량 조건부 정규밀도 $\phi_1(x_i|\eta_1,\ldots,\eta_{i-1})$의 누적분포함수이며, $\eta_i = E(X_i|X_1=\eta_1,\ldots,X_{i-1}=\eta_{i-1},X_i\le x_i) = \mu_{i|\{1,\ldots,i-1\}} - \sigma_{i|\{1,\ldots,i-1\}}\phi_1(w_i)/\Phi_1(w_i)$이고, $w_i = (x_i-\mu_{i|\{1,\ldots,i-1\}})/\sigma_{i|\{1,\ldots,i-1\}}$ 그리고 조건부 평균 $\mu_{i|\{1,\ldots,i-1\}} = E(X_i|X_1=\eta_1,\ldots,X_{i-1}=\eta_{i-1})$ 및 $\sigma^2_{i|\{1,\ldots,i-1\}} = \mathrm{var}(X_i|X_1=\eta_1,\ldots,X_{i-1}=\eta_{i-1})$을 나타낸다.

식 (3.4)의 특징은 모든 성분이 처리속도가 매우 빠른 $2q-1$번의 단변량 누적분포함수 계산으로 이루어져 있다는 것이다. 우선 $q=1$이면 $\Phi^*_1(x) = \Phi_1(x)$이므로 처리속도는 동일하다. $q>1$에 대해, 임의

Table 3.1. Comparison between Exact CDF and Approximate CDF. Second entries in each case indicate the errors

(x_1, x_2, x_3)	$\sigma_{12} = \sigma_{13} = \sigma_{23}$											
	0.1				0.5				0.9			
	$\Phi_3(\boldsymbol{x})$	$\Phi_3^*(\boldsymbol{x})$	$\Phi_3^{*d}(\boldsymbol{x})$	$\Phi_3^{*a}(\boldsymbol{x})$	$\Phi_3(\boldsymbol{x})$	$\Phi_3^*(\boldsymbol{x})$	$\Phi_3^{*d}(\boldsymbol{x})$	$\Phi_3^{*a}(\boldsymbol{x})$	$\Phi_3(\boldsymbol{x})$	$\Phi_3^*(\boldsymbol{x})$	$\Phi_3^{*d}(\boldsymbol{x})$	$\Phi_3^{*a}(\boldsymbol{x})$
$(5,5,5)$	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
		0.0000	0.0000	0.0000		0.0000	0.0000	0.0000		0.0000	0.0000	0.0000
$(-5,-5,-5)$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
		0.0000	−0.0000	−0.0000		−0.0000	−0.0000	−0.0000		−0.0000	−0.0000	−0.0000
$(2,2,2)$	0.9343	0.9355	0.9355	0.9355	0.9425	0.9617	0.9617	0.9617	0.9617	0.9772	0.9772	0.9772
		0.0012	0.0012	0.0012		0.0192	0.0192	0.0192		0.0155	0.0155	0.0155
$(-2,-2,-2)$	0.0001	0.0000	0.0000	0.0000	0.0014	0.0013	0.0013	0.0013	0.0101	0.0107	0.0107	0.0107
		−0.0000	−0.0000	−0.0000	-	−0.0001	−0.0001	−0.0001	-	0.0005	0.0005	0.0005
$(1,1,1)$	0.6106	0.6119	0.6119	0.6119	0.6778	0.7084	0.7084	0.7084	0.7732	0.8393	0.8393	0.8393
		0.0012	0.0012	0.0012		0.0307	0.0307	0.0307		0.0661	0.0661	0.0661
$(0,0,0)$	0.1489	0.1487	0.1487	0.1487	0.2500	0.2542	0.2542	0.2542	0.3923	0.4624	0.4624	0.4624
		−0.0002	−0.0002	−0.0002		0.0042	0.0042	0.0042		0.0700	0.0700	0.0700
$(-1,-1,-1)$	0.0074	0.0073	0.0073	0.0073	0.0338	0.0325	0.0325	0.0325	0.0973	0.1117	0.1117	0.1117
		−0.0000	−0.0000	−0.0000		−0.0013	−0.0013	−0.0013		0.0144	0.0144	0.0144
$(1,2,3)$	0.8227	0.8230	0.8250	0.8230	0.8316	0.8357	0.8816	0.8357	0.8411	0.8413	0.9942	0.8413
		0.0003	0.0023	0.0003		0.0041	0.0499	0.0041		0.0002	0.1531	0.0002
$(3,2,1)$	0.8227	0.8250	0.8250	0.8230	0.8316	0.8816	0.8816	0.8357	0.8411	0.9942	0.9942	0.8413
		0.0023	0.0023	0.0003		0.0499	0.0499	0.0041		0.1531	0.1531	0.0002
$(-2,-1,0)$	0.0032	0.0032	0.0032	0.0032	0.0126	0.0127	0.0080	0.0127	0.0225	0.0226	0.0005	0.0226
		−0.0000	−0.0001	−0.0000		0.0001	−0.0046	0.0001		0.0001	−0.0220	0.0001
$(0,-1,-2)$	0.0032	0.0032	0.0032	0.0032	0.0126	0.0080	0.0080	0.0127	0.0225	0.0005	0.0005	0.0226
		−0.0001	−0.0001	−0.0000		−0.0046	−0.0046	0.0001		−0.0220	−0.0220	0.0001
$(2,1,3)$	0.8227	0.8238	0.8250	0.8230	0.8316	0.8622	0.8816	0.8357	0.8411	0.9694	0.9942	0.8413
		0.0011	0.0023	0.0003		0.0306	0.0499	0.0041		0.1283	0.1531	0.0002
$(3,1,2)$	0.8227	0.8243	0.8250	0.8230	0.8316	0.8703	0.8816	0.8357	0.8411	0.9880	0.9942	0.8413
		0.0017	0.0023	0.0003		0.0387	0.0499	0.0041		0.1469	0.1531	0.0002

적으로 100개의 관측치를 발생시켜 평균벡터와 공분산행렬을 얻은 후 반복 실험해 본 결과, 평균적으로 2-변량의 경우 $\Phi_2^*(\boldsymbol{x})$의 계산시간은 $\Phi_2(\boldsymbol{x})$ 보다 약 60배 정도, 3-변량의 경우 약 150배, 5-변량의 경우 약 2,000배 정도 빨랐다. 물론 MSNMix$_q$ 모형을 적합 시에도 이 정도로 향상된 처리속도가 유지된다고는 말할 수 없지만 매우 개선된 처리속도를 보장할 것은 분명하다.

문제는 근사 누적분포함수의 정확도일 것이다. Table 3.1에 $\boldsymbol{\mu} = (0,0,0)^T$이며 $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 1$이고 $\sigma_{12} = \sigma_{13} = \sigma_{23}$가 0.1, 0.5 및 0.9인 상황에서 3-변량 누적분포함수 $\Phi_3(\boldsymbol{x})$와 근사값 $\Phi_3^*(\boldsymbol{x})$의 계산결과를 비교하였다. 각 관측치에서 첫 번째 행은 누적분포함수값이며 두 번째 행은 오차 $\Phi_3^*(\boldsymbol{x}) - \Phi_3(\boldsymbol{x})$를 나타낸다.

대체적으로 $\Phi_3^*(\boldsymbol{x})$의 정확도는 만족스러워 보이나, 변수들 간 상관성이 높으며 (x_1, x_2, x_3)가 모두 아주 크지 않은 큰 양의 값을 가질 때 정확도는 낮아지는 것으로 보인다. 그리고 상관계수가 작거나 (x_1, x_2, x_3)들이 작은 값일 때 $\Phi_3^*(\boldsymbol{x})$의 정확도가 높아지는 것을 보여주고 있다. 실제 $\sigma_{12} = \sigma_{13} = \sigma_{23} = 0$이면 $\Phi_3(\boldsymbol{x}) = \Phi_3^*(\boldsymbol{x}) = \prod_{i=1}^3 \Phi_1(x_i)$이 되는데, 이것은 식 (3.4)로부터도 확인할 수 있다.

Table 3.1의 상단은 $x_1 = x_2 = x_3$인 변량값인 반면, 하단은 서로 다른 값으로 구성하였다. 그런데 주목할 것은 누적확률 $\Phi_3(x_1, x_2, x_3) = \Phi_3(x_3, x_2, x_1)$이지만 근사누적확률 $\Phi_3^*(x_1, x_2, x_3) \neq \Phi_3^*(x_3, x_2, x_1)$라는 사실이다. Table 3.1에서 $\Phi_3^{*a}(\boldsymbol{x})$는 x_1, x_2, x_3를 오름차순으로 정렬한 후 구한 근사값이며, 반대로 $\Phi_3^{*d}(\boldsymbol{x})$는 내림차순 정렬에 의한 근사값이다.

예를 들어 상관계수가 0.1일 때, $(x_1 = 2, x_2 = 1, x_3 = 3)$인 경우 $\Phi_3^*(2,1,3) = 0.8238$이며, $\Phi_3^{*d}(2,1,3) = \Phi_3^*(3,2,1) = 0.8250$인 반면 $\Phi_3^{*a}(2,1,3) = \Phi_3^*(1,2,3) = 0.8230$으로서 참값 $\Phi_3(2,1,3) = \Phi_3(1,2,3) = \Phi_3(3,2,1) = 0.8227$에 가장 가깝다. 다시 말해서 내림차순으로 되어있는 자료의 근사값이 정확도가 가장 나쁘며, 오름차순으로 되어 있는 자료의 정확도가 가장 높다. 이러한 사실은 상관계수가 커질수록 더 분명해진다. 본 논문에는 수록하지는 않았으나 다른 공분산 구조 하에서도 이러한 성질은 동일하게 나타난다. 따라서 근사치의 정확성을 향상시키기 위해서 관측치의 표준화 변량값 $\boldsymbol{x} = \text{diag}(\boldsymbol{\Sigma})^{-1/2}(\boldsymbol{y}_j - \boldsymbol{\mu})$들을 오름차순으로 정렬한 후 그 순서에 따라 근사누적확률을 구해야할 필요가 있다.

위의 성질은 직관적으로 다음과 같이 설명할 수 있을 것이다.

$$\Phi_3^*(\boldsymbol{x}) = \Phi_1(x_1)\Phi_1(x_2|\eta_1)\Phi_1(x_3|\eta_1,\eta_2)$$

는 식 (3.2)–(3.4)를 통한 유도과정에서 알 수 있듯이 오른쪽으로 진행할수록 오차는 증폭된다. 이때 가장 좁은 범위인 $(-\infty, x_1]$에서 정확한 값 $\Phi_1(x_1)$을 계산하고, 가장 큰 범위 $(-\infty, x_3]$에서 증폭된 오차가 상대적으로 작은 비중을 가지도록 하기 때문이다.

다음 절에서는 근사누적분포함수 $\Phi_q^*(\boldsymbol{x})$와 (오름차순) 정렬 후 근사누적분포함수 $\Phi_q^{*a}(\boldsymbol{x})$를 적용한 MSNMix_q의 결과를 제공할 것이다. 여기서 유의할 점은 $(x_1, \ldots, x_q)$의 정렬에 대응하도록 평균벡터의 원소의 위치와 공분산행렬의 행과 열의 위치를 바꾸어주어야 한다는 것이다. 모든 n개의 관측치에 대해 이 정렬과정을 수행하는데 따르는 적지 않은 처리시간이 소요된다. 이것은 근사값의 부가적인 정확성을 위해 어쩔 수 없이 지불하는 비용이라 해야할 것이다.

4. 실험

정확한 $\Phi_q(\boldsymbol{x})$을 사용하는 Exact-EM 알고리즘에 대응하여 근사누적분포함수 $\Phi_q^*(\boldsymbol{x})$와 정렬 후 근사누적분포함수 $\Phi_q^{*a}(\boldsymbol{x})$가 적용된 적합 알고리즘을 각각 A-EM 및 AS-EM 알고리즘이라 칭하겠다. 그리고 세 가지 적합 알고리즘에 대응하는 MSN 혼합모형을 각각 MSNMix_q, MSNMix_q^* 및 MSNMix_q^{*a}라 표시하겠다. 즉, MSNMix_q^*과 MSNMix_q^{*a} 모형은 식 (2.2)에서 $\Phi_q(\boldsymbol{x})$ 대신 각각 $\Phi_q^*(\boldsymbol{x})$과 $\Phi_q^{*a}(\boldsymbol{x})$를 바꾼 모형이다.

이 절에서는 Cook과 Weisberg (1994)에서 제공된 Australian Institution of Sport(AIS) 자료를 이용하여 A-EM 및 AS-EM 알고리즘의 결과가 Exact-EM알고리즘의 적합결과와 근사한지 확인하고 처리속도는 어떠한지 비교할 것이다. AIS 자료는 100명의 여성과 102명의 남성 육상선수의 11개 신체특성을 측정한 자료인데, 우리는 이들 특성 중에 4변량 body mass index(BMI; Y_1), percentage of body fat(Bfat; Y_2), lean body mass(LBM; Y_3), height(Ht; Y_4) 만을 사용하여 실험한다.

먼저 2변량 $\boldsymbol{Y} = (Y_1, Y_2)^T = (\text{BMI}, \text{Bfat})^T$의 자료에 대한 2-성분 혼합모형 MSNMix를 적용하되, 치우침 모수 $\boldsymbol{\Delta} = (\boldsymbol{\delta}_1, \boldsymbol{\delta}_2)$와 같이 $q = 2$로 하겠다. Figure 4.2에 그 적합 결과를 나타내었다. 그림에서 마커 ◯와 △은 각각 여자와 남자 관측치를 나타내며, 등고선은 적합된 혼합모형의 높이를 나타낸다. 가장 왼쪽이 Exact-EM 알고리즘에 의한 MSNMix_q를 적합한 것인데, 두 성분밀도가 여자와 남자 그룹의 분포을 비교적 잘 구분해 주고 있다. 그런데 A-EM 및 AS-EM 알고리즘에 의한 두 모형 MSNMix_q^* (중간 그림) 및 MSNMix_q^{*a} (오른쪽 그림)의 적합 결과가 MSNMix_q를 적합 결과와 구분이 어려울 정도로 비슷하다. 실제 Figure 4.1의 반복에 따른 로그-우도 플롯 그림 (왼쪽)을 보면 (동일한 초기 추정치를 사용한다는 전제 하에서) 세 알고리즘의 수렴 경로는 거의 일치하고 있다. 이것은 $q = 2$인 경우 $\Phi_q(\boldsymbol{x})$와 근사치 $\Phi_q^*(\boldsymbol{x})$ 및 정렬 후 근사치 $\Phi_q^{*a}(\boldsymbol{x})$의 차이는 크게 나타나지 않음을 말한다.

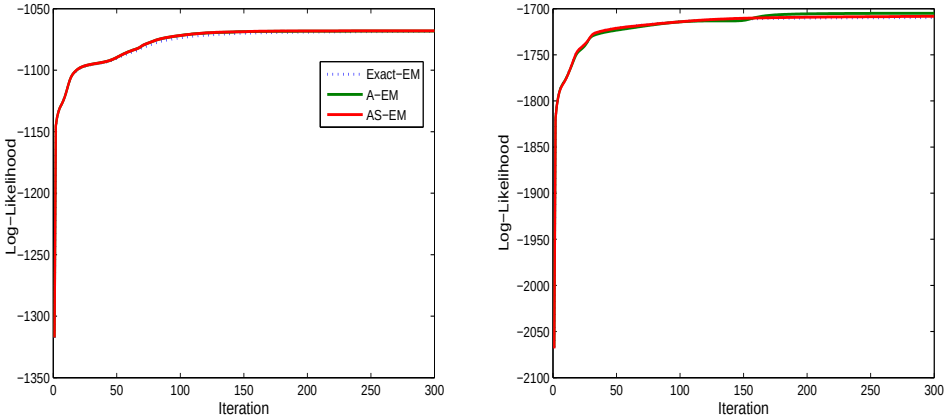

Figure 4.1. Log-Likelihood over the iterations for MSNMix_2 (Exact-EM), MSNMix_2^* (A-EM) and MSNMix_2^{*a} (AS-EM) (left: For two variables (BMI, Bfat), right: For three variables (BMI, Bfat, LBM)).

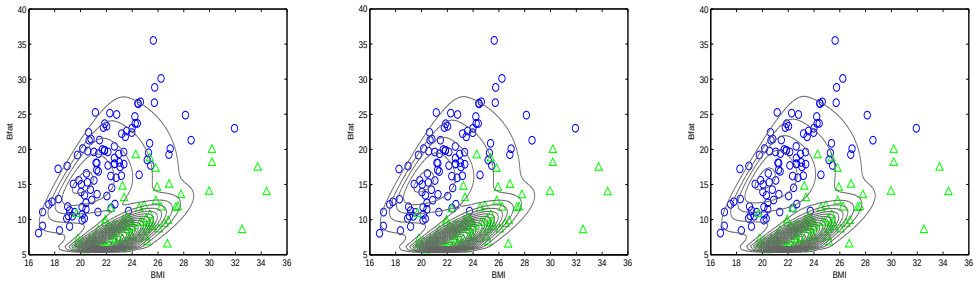

Figure 4.2. Results of fit for (BMI, Bfat) (left: MSNMix_2, middle: MSNMix_2^*, right: MSNMix_2^{*a}). The symbols ◯ and △ indicate female and male, respectively.

다음은 3 변량 $\boldsymbol{Y} = (Y_1, Y_2, Y_3)^T = (\text{BMI}, \text{Bfat}, \text{LMB})^T$ 데이터 셋에 대해 $\boldsymbol{\Delta} = (\boldsymbol{\delta}_1, \boldsymbol{\delta}_2, \boldsymbol{\delta}_2)$로 하여 적합을 시도하였다. Figure 4.1의 오른쪽 그림에 우도의 증가 상태를 나타내었는데, A-EM 알고리즘 보다는 AS-EM 알고리즘이 보다 Exact-EM 알고리즘과 유사한 수렴경로를 보이고 있다. 실제로 Figure 4.3에 나타낸 그들의 적합 결과 역시 MSNMix_q (첫 번째 열)와 MSNMix_q^{*a} (두 번째 열)의 결과는 유사한 반면 MSNMix_q^* (세 번째 열)는 다소 이질적이다.

Table 4.1에 세 알고리즘이 도달하는 로그-우도와 처리시간을 정리하였다. 이상의 결과로 볼 때, $q = 2$이면 A-EM 알고리즘을 사용하여 MSNMix_q^* 적합 모형을 얻는 것이 현명하며, $q > 2$일 때는 AS-EM 알고리즘을 이용하여 MSNMix_q^{*a}의 적합결과를 얻어 MSNMix_q을 추측해 볼 것을 추천한다.

본 실험에서 사용된 PC는 Pentium Dual-Core 2.5GHz이다.

5. 결론 및 토의

본 논문에서는 q개의 다중 치우침 모수를 가지는 다변량 치우친 혼합모형 MSNMix_q에 대한 근사적 EM 알고리즘을 제안하였다. Lee와 McLachlan (2013)의 Exact-EM 알고리즘의 사용은 너무 큰 처리시간을 요하기 때문에 좀 빠른 알고리즘이 필요한 실정이다. Exact-EM 알고리즘의 처리가 지연되는

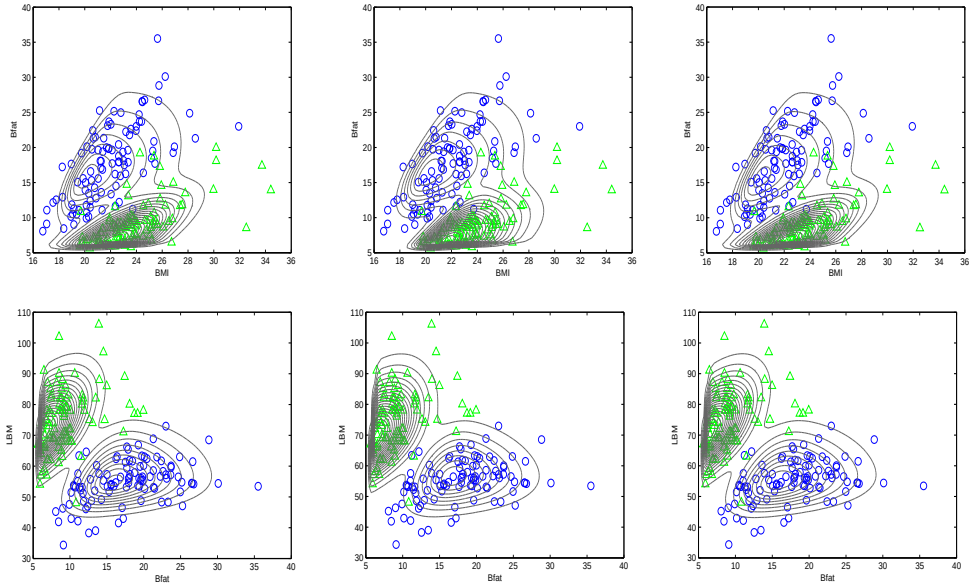

Figure 4.3. Results of fit for (BMI, Bfat, LBM) ((a) 1st column: MSNMix_3, (b) 2nd column: MSNMix_3^*, (c) 3rd column: MSNMix_3^{*a}). The symbols ◯ and △ indicate female and male, respectively.

Table 4.1. Log-Likelihoods and Execution time (PC: Pentium Dual-Core 2.5GHz)

	$q=2$			$q=3$			$q=4$		
	Exact-EM	A-EM	AS-EM	Exact-EM	A-EM	AS-EM	Exact-EM	A-EM	AS-EM
Log-Likelihood	−1068.5	−1068.2	−1067.9	−1709.7	−1705.0	−1708.3	-	-	-
# of iterations	300	300	300	300	300	300	10	10	10
Etime	561.6	183.2	233.2	1962.7	480.5	509.6	2434.1	30.6	34.7
Etime per iteration	1.9	0.6	0.8	6.5	1.6	1.7	243.4	3.1	3.5

주된 부분이 q-변량 정규분포 누적함수 $\Phi_q(\boldsymbol{x})$ 계산에 있다고 보고, Olson과 Weissfeld (1991)이 제안한 다중적분의 근사기법에 따라 $\Phi_q(\boldsymbol{x})$에 대한 근사값 $\Phi_q^*(\boldsymbol{x})$을 유도하고 이 보다 좀 더 정확도가 높은 $\Phi_q^{*a}(\boldsymbol{x})$를 적용하면 참 MSNMix_q 적합결과에 근접한 결과를 얻을 수 있음을 보였다.

본 연구는 정규분포 혼합모형에 한정하였다. 그러나 최근에는 이상치에 대한 로버스트 성질을 반영하기 위해 다변량 치우친 t-분포를 기반으로 한 MSTMix_q의 적용이 대세이다. 제안된 근사 방법이 과연 이 모형에도 잘 적용될 수 있을지 궁금하다. 그리고 제안된 A-EM과 AS-EM 알고리즘이 Exact-EM 보다 빠른 처리시간을 보장하지만 EM 알고리즘의 확률적 구조에 Olson과 Weissfeld (1991) 원리를 직접 이식하는 방법도 고려해 볼만하다고 사료된다.

References

Azzalini, A. (1985). A class of distribution which includes the normal ones, *Scandinavian Journal of Statistics*, **33**, 561–574.

Azzalini, A. and Dalla-Valle, A. (1996). The multivariate skew normal distribution, *Biometrika*, **83**, 715–726.

Arellano-Valle, R. B. and Genton, M. G. (2005). On fundamental skew distributions, *Journal of Multivariate Analysis*, **96**, 93–116.

Cabral, C. S., Lachos, V. H., and Prates, M. O. (2012). Multivariate mixture modeling using skew-normal independent distribution, *Computational Statistics and Data Analysis*, **56**, 126–142.

Cook, R. D. and Weisberg, S. (1994). *An Introduction to Regression Graphics*, Wiley, New York.

Ho, H. J., Lin, T. I., Chen, H.-Y., and Wang, W.-L. (2012). Some results on the truncated multivariate t distribution, *Journal of Statistical Planning & Inference*, **142**, 25–40.

Kim, S.-G. (2014). An alternating approach of maximum likelihood estimation for mixture of multivariate skew t-distribution, *The Korean Journal of Applied Statistics*, **27**, 819–831.

Lee, S. X. and McLachlan, G. J. (2013). On mixtures of skew normal and skew t-distributions, *Advances in Data Analysis and Classification*, **7**, 241–266.

Lee, S. X. and McLachlan, G. J. (2014a). Finite mixtures of multivariate skew t-distributions: some recent and new results, *Statistics and Computing*, **24**, 181–202.

Lee, S. X. and McLachlan, G. J. (2014b). Finite mixtures of canonical fundamental skew t-distributions, *arXiv: 1405.0685v1 [Stat. ME] 4 May 2014*.

Lin, T.-I. (2010). Robust mixture modeling using multivariate skew t-distributions, *Statistics and Computing*, **20**, 343–356.

Olson, J. M. and Weissfeld, L. A. (1991). Approximation of certain multivariate integrals, *Statistics & Probability Letters*, **11**, 309–317.

Pyne, S., Hu, X., Wang, K., Rossin, E., Lin, T. I., Maier, L., Baecher-Allan, C., McLachlan, G. J., Tamayo, P., Hafler, D. A., De Jager, P. L., and Mesirov, J. P. (2009). Automated high-dimensional flow cytometric data analysis, In *Proceedings of the National Academy of Sciences*, **106** , 8519–8524.

Sahu, S. K., Dey, D. K., and Branco, M. D. (2003). A new class of multivariate skew distribution with application to Bayesian regression model, *The Canadian Journal of Statistics*, **31**, 129–150.

EM 알고리즘에 의한 다변량 치우친 정규분포 혼합모형의 근사적 적합

김승구[a,1]

[a]상지대학교 컴퓨터데이터정보학과

(2016년 2월 19일 접수, 2016년 3월 5일 수정, 2016년 3월 7일 채택)

요 약

다중 치우침 모수벡터를 가진 다변량 치우친 정규분포 (MSNMix)를 EM 알고리즘으로 적합하려면 E-step에서 다변량 절단 정규분포의 적률과 확률을 계산해야 하는데 이것은 매우 큰 계산 시간을 요구한다. 그래서 비대칭 자료를 적합하는데 흔히 단순 치우침 모수를 가진 모형을 적용한다. 이 모형은 단변량 처리방식으로 적합하는 것이 가능하기 때문에 처리속도가 매우 빠르다. 그러나 단순 치우침 모수를 적용하는 것은 응용에서 비현실적인 경우가 많다. 본 논문에서는 다중 치우침 모수를 가지는 MSNMix의 근사적 추정법을 제안하는데, 이 방법은 단변량 처리방식이 적용되므로 향상된 처리속도를 보장한다. 그리고 제안된 방법의 실효성을 보이기 위해 몇 가지 실험 결과를 제공한다.

주요용어: 치우친 다변량 정규분포, 혼합모형, EM 알고리즘, 다변량 정규 cdf

본 연구는 상지대학교 2014 교내 연구비 지원에 의해 수행되었음.

[1](04310) 강원도 원주시 상지대길 83, 상지대학교 컴퓨터데이터정보학과. E-mail: sgukim@sangji.ac.kr