The Korean Journal of Applied Statistics (2014) DOI: http://dx.doi.org/10.5351/KJAS.2014.27.5.819
27(5), 819–831

An Alternating Approach of Maximum Likelihood Estimation for Mixture of Multivariate Skew *t*-Distribution

Seung-Gu Kim[a,1]

[a]Department of Data and Information, Sangji University

(Received September 1, 2014; Revised September 22, 2014; Accepted September 22, 2014)

Abstract

The Exact-EM algorithm can conventionally fit a mixture of multivariate skew distribution. However, it suffers from highly expensive computational costs to calculate the moments of multivariate truncated *t*-distribution in E-step. This paper proposes a new SPU-EM method that adopts the AECM algorithm principle proposed by Meng and van Dyk (1997)'s to circumvent the multi-dimensionality of the moments. This method offers a shorter execution time than a conventional Exact-EM algorithm. Some experments are provided to show its effectiveness.

Keywords: Multivariate skew *t*-distribution, mixture model, EM algorithm, AECM algorithm.

1. 서론

최근 다변량 치우친 *t*-분포 (multivariate skew *t*-distribution; MST)의 혼합모형에 대한 연구가 활발하다. 그동안 군집분석을 수행하는 많은 응용문제에서 자료를 대칭적이라 가정하고 정규혼합모형(mixture of normal distribution)이나 혹은 좀 더 이상치들에 로버스트한 *t*-분포 혼합모형을 이용하여 왔다. 그러나 보건 의료 분야에서 나타나는 자료들은 비대칭적인 경우가 많아서 최근들어 비대칭적이면서 이상치를 가지는 자료를 적합하기 위하여 MST에 대한 연구가 활발하다.

Azzalini (1985) 및 Azzalini와 Dalla-Valle (1996)에 의해 MSN(multivariate skew normal distribution)이 소개된 이후 최근에는 MSN을 특수한 경우로서 포함하는 MST 분포의 개발에 집중되고 있다. Pyne 등 (2009)은 다변량 치우침 함수(skewing function)를 가지는 Sahu 등 (2003)의 MST 분포에서 치우침 함수를 단변량으로 제약한 버전을 소개하고 획기적으로 빠른 계산 속도로 EM(expectation-maximization)알고리즘에 의해 추정할 수 있음을 보였다. 그러나 단변량 치우침 함수의 한계로 인해 낮은 적합도를 가지는 단점이 있다. 한편 Lin (2010)은 Sahu 등 (2003)의 MST를 적합하기 위해 Monte Carlo EM 알고리즘(MC-EM) 개발하였는데, 이 방법은 E-step에서 요구되는 다변량 절단 *t*-분포의 적률에 대해 모의연구를 수행한다. 그로인해 이 추정법은 비현실적으로 큰 계산 시간을 요하여 실용성의 문제가 대두되었다.

This research was supported by Sangji University Research Fund, 2013.
[1]Department of Data and information, Sangji University, WooSan-Dong, Wonju 220-702, Korea.
E-mail: sgukim@sangji.ac.kr

MC-EM의 계산시간을 개선하기 위해 최근 Lee와 McLachlan (2013, 2014a)은 E-step에서 다변량 절단분포의 적률을 명시적 수식으로 표현한 정확한 EM 알고리즘(Exact-EM)을 개발하였다. 또한 극히 최근 Lee와 McLachlan (2014b)은 Arellano-Valle와 Genton (2005)의 다변량 CFUST(canonical fundamental skew t-distribution)에 대하여 Exact-EM 알고리즘을 적용하였다. 그러나 Exact-EM이 MC-EM보다는 개선되기는 하였으나 여전히 매우 큰 계산 시간을 요한다.

본 연구에서는 이 문제를 완화시키기 위해 E-step에서 요구되는 다변량 절단분포의 적률에 대한 새로운 근사를 통해 Exact-EM 만큼의 적합도(로그-우도)를 가지면서 계산시간을 크게 단축시킬 수 있는 기법을 제공하고자 한다.

다음 절에서는 다변량 CFUST 분포를 바탕으로 MST를 정의하고, Exact-EM 알고리즘을 소개한 후 현존하는 문제점과 본 연구의 동기를 설명한다. 그리고 3절에서는 제안된 방법을 소개하고, 4절에서는 MST 혼합모형으로 확장하며, 5절에서는 몇가지 실험을 통해 제안된 방법의 실효성을 보인다. 마지막으로 6절에서는 결론과 토의 그리고 추후연구과제를 정리한다.

2. MST 및 Exact-EM

2.1. MST의 정의

p-변량 관측치 $\boldsymbol{y}_j$ $(j = 1, \ldots, n)$에 대해 q차 치우침 모수(skewing parameter) $\boldsymbol{\Delta}_{p\times q} = (\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_q)$를 가지는 MST는 다음과 같이 정의한다.

$$f\left(\boldsymbol{y}_j; \boldsymbol{\Theta}\right) = 2^q t_p\left(\boldsymbol{y}_j; \boldsymbol{\mu}, \boldsymbol{\Omega}, \nu\right) T_q\left(\sqrt{\frac{\nu+p}{\nu+D_j}}\tilde{\boldsymbol{x}}_j; \boldsymbol{\Psi}, \nu+p\right), \tag{2.1}$$

여기서 $\boldsymbol{\Omega} = \boldsymbol{\Sigma} + \boldsymbol{\Delta}\boldsymbol{\Delta}^T$, $\tilde{\boldsymbol{x}}_j = \boldsymbol{\Delta}^T\boldsymbol{\Omega}^{-1}(\boldsymbol{y}_j - \boldsymbol{\mu})$, $\boldsymbol{\Psi} = \boldsymbol{I}_q - \boldsymbol{\Delta}^T\boldsymbol{\Omega}^{-1}\boldsymbol{\Delta}$, $D_j = (\boldsymbol{y}_j - \boldsymbol{\mu})^T\boldsymbol{\Omega}^{-1}(\boldsymbol{y}_j - \boldsymbol{\mu})$이며, $t_p(\cdot; \boldsymbol{m}, \boldsymbol{S}, \nu)$는 위치모수, 척도모수 및 자유도가 각각 $\boldsymbol{m}$, $\boldsymbol{S}$, ν인 p-변량 t-분포의 확률밀도이며 $T_q(\cdot; \boldsymbol{S}, \nu)$는 위치모수, 척도모수 및 자유도가 각각 $\boldsymbol{0}$, $\boldsymbol{S}$, ν인 q-변량 t 누적분포를 나타낸다. 그리고 $\boldsymbol{\Theta}$는 밀도에 포함된 모든 모수를 포함하는 벡터이다.

우리의 목표는 로그-우도

$$L(\boldsymbol{\Theta}) = \sum_{j=1}^{n} \log\left\{2^q t_p\left(\boldsymbol{y}_j; \boldsymbol{\mu}, \boldsymbol{\Omega}, \nu\right) T_q\left(\sqrt{\frac{\nu+p}{\nu+D_j}}\tilde{\boldsymbol{x}}_j; \boldsymbol{\Psi}, \nu+p\right)\right\}$$

를 최대화하는 $\boldsymbol{\Theta}$를 얻는 것이다.

MST (2.1)의 확률적 표기는 다음과 같다. 즉,

$$\begin{aligned}\boldsymbol{Y}_j &= \boldsymbol{\mu} + U_j^{-\frac{1}{2}}\boldsymbol{\Delta}\boldsymbol{X}_j + U_j^{-\frac{1}{2}}\boldsymbol{W}_j \\ &= \boldsymbol{\mu} + U_j^{-\frac{1}{2}}\left(\boldsymbol{\delta}_1 X_{1j} + \cdots + \boldsymbol{\delta}_q X_{qj}\right) + U_j^{-\frac{1}{2}}\boldsymbol{W}_j,\end{aligned}$$

여기서 $U_j \sim \text{gamma}(\nu/2, \nu/2)$, $\boldsymbol{X}_j = (X_{1j}, \ldots, X_{qj})^T \sim HN_q(\boldsymbol{0}, \boldsymbol{I}_q)$이며 $\boldsymbol{W}_j \sim N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$, 그리고 $\boldsymbol{X}_j$와 $\boldsymbol{W}_j$는 서로 독립이다.

$q = 1$일 때 즉 $\boldsymbol{\Delta}_{p\times 1} = \boldsymbol{\delta}$ 때 식 (2.1)은 Lee와 McLachlan (2013)이 'restricted' MST라고 부른 Pyne 등 (2009)의 MST 분포가 되며, $q = p$일 때 다변량 CFUST 분포가 된다. 그리고 $\boldsymbol{\Delta} = \boldsymbol{0}$이면 $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$로 축소된다.

2.2. Exact EM 알고리즘

MST (2.1)의 관측치 $\boldsymbol{y} = (\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_n^T)^T$ 단 $\boldsymbol{y}_j = (y_{1j}, \ldots, y_{pj})^T$의 생성을 위한 위계적 구조는 다음과 같다. 즉,

$$\begin{aligned}\boldsymbol{Y}_j = \boldsymbol{y}_j | (\boldsymbol{X}_j = \boldsymbol{x}_j, U_j = u_j) &\sim N_p(\boldsymbol{\mu} + \boldsymbol{\Delta}\boldsymbol{x}_j, \boldsymbol{\Sigma}/u_j),\\ \boldsymbol{X}_j = \boldsymbol{x}_j | (U_j = u_j) &\sim N_q(\boldsymbol{0}, \boldsymbol{I}/u_j),\\ U_j &= \text{gamma}(\nu/2, \nu/2).\end{aligned}$$

이때 완비자료(complete data) $(\boldsymbol{y}_j, u_j, \boldsymbol{x}_j)$의 로그-우도는

$$L_c(\boldsymbol{\Theta}) = L_{c1}(\boldsymbol{\theta}) + L_{c2}(\nu)$$

인데, 여기서

$$L_{c1}(\boldsymbol{\theta}) = \sum_{j=1}^{n}\left[-\frac{1}{2}\log u_j|\boldsymbol{\Sigma}| - \frac{u_j}{2}\left(\boldsymbol{y}_j - \boldsymbol{\mu} - \boldsymbol{\Delta}\boldsymbol{x}_j\right)^T\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{y}_j - \boldsymbol{\mu} - \boldsymbol{\Delta}\boldsymbol{x}_j\right)\right], \tag{2.2}$$

$$L_{c2}(\nu) = \sum_{j=1}^{n}\left[\frac{\nu}{2}\log\left(\frac{\nu}{2}\right) - \log\Gamma\left(\frac{\nu}{2}\right) + \left(\frac{\nu}{2}\right)(\log u_j - u_j)\right] \tag{2.3}$$

을 나타낸다. 그리고 $\boldsymbol{\theta}$는 $\boldsymbol{\Theta}$에서 ν를 제외한 벡터를 나타낸다.

이제 EM 알고리즘의 k 번째 반복에서 E-step은 $Q(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(k)}) = E_{\boldsymbol{\Theta}^{(k)}}[L_c(\boldsymbol{\Theta})|\boldsymbol{y}] = E_{\boldsymbol{\Theta}^{(k)}}[L_{c1}(\boldsymbol{\theta})|\boldsymbol{y}] + E_{\boldsymbol{\Theta}^{(k)}}[L_{c2}(\nu)|\boldsymbol{y}] = Q_1(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) + Q_2(\nu|\nu^{(k)})$의 계산을 필요로 하는데, 이것은 다음과 같은 조건부 기대값을 구하는 것으로 귀결된다. 즉, $j = 1, \ldots, n$에 대해

$$\begin{aligned}e_{1,j}^{(k+1)} &= E_{\boldsymbol{\Theta}^{(k)}}\left(U_j|\boldsymbol{y}_j\right)\\ &= \frac{\nu^{(k)} + p}{\nu^{(k)} + D_j^{(k)}}\frac{T_q\left(\sqrt{\frac{\nu^{(k)}+p+2}{\nu^{(k)}+D_j^{(k)}}}\tilde{\boldsymbol{x}}_j^{(k)}; \boldsymbol{\Psi}^{(k)}, \nu^{(k)} + p + 2\right)}{T_q\left(\sqrt{\frac{\nu^{(k)}+p}{\nu^{(k)}+D_j^{(k)}}}\tilde{\boldsymbol{x}}_j^{(k)}; \boldsymbol{\Psi}^{(k)}, \nu^{(k)} + p\right)},\\ \boldsymbol{e}_{2,j}^{(k+1)} &= E_{\boldsymbol{\Theta}^{(k)}}\left(U_j\boldsymbol{X}_j|\boldsymbol{y}_j\right) = e_{1,j}^{(k+1)}\boldsymbol{m}_j^{(k+1)}, \qquad (2.4)\\ \boldsymbol{E}_{3,j}^{(k+1)} &= E_{\boldsymbol{\Theta}^{(k)}}\left(U_j\boldsymbol{X}_j\boldsymbol{X}_j^T|\boldsymbol{y}_j\right) = e_{1,j}^{(k+1)}\boldsymbol{M}_j^{(k+1)}, \qquad (2.5)\\ e_{4,j}^{(k+1)} &= E_{\boldsymbol{\Theta}^{(k)}}\left(\log U_j|\boldsymbol{y}_j\right)\end{aligned}$$

의 계산이 필요하다. 여기서 $\boldsymbol{m}_j^{(k+1)} = E_{\boldsymbol{\Theta}^{(k)}}(\boldsymbol{X}_j|\boldsymbol{y}_j)$ 그리고 $\boldsymbol{M}_j^{(k+1)} = E_{\boldsymbol{\Theta}^{(k)}}(\boldsymbol{X}_j\boldsymbol{X}_j^T|\boldsymbol{y}_j)$로서

$$\boldsymbol{X}_j|\boldsymbol{y}_j \sim Tt_p\left(\tilde{\boldsymbol{x}}_j^{(k)}, \frac{\nu^{(k)} + D_j^{(k)}}{\nu^{(k)} + p + 2}\boldsymbol{\Psi}^{(k)}, \nu^{(k)} + p + 2\ |(0,\infty)^p\right)$$

와 같이 양의 영역에 지지도(support)를 가지는 q-변량 절단 t-분포를 따른다. $e_{4,j}^{(k+1)}$의 계산방법은 여러가지 있을 수 있으나 본 논문에서는 Kim (2012)의 방법을 사용하였다. 이것은 본 논문의 주된 관심사가 아니므로 자세한 내용은 생략하기로 한다.

한편 M-step에서는 $Q(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(k)})$를 $\boldsymbol{\Theta}$에 관해 최대화하는데, 이것은 다음을 계산하는 것으로 귀결된다. 즉,

$$\begin{aligned}
\boldsymbol{\mu}^{(k+1)} &= \frac{1}{\sum_{j=1}^{n} e_{1,j}^{(k+1)}} \left(\sum_{j=1}^{n} e_{1,j}^{(k+1)} \boldsymbol{y}_j - \boldsymbol{\Delta}^{(k)} \sum_{j=1}^{n} \boldsymbol{e}_{2,j}^{(k+1)} \right), \\
\boldsymbol{\Delta}^{(k+1)} &= \left[\sum_{j=1}^{n} \tilde{\boldsymbol{y}}_j^{(k+1)} \boldsymbol{e}_{2,j}^{(k+1)^T} \right] \left[\sum_{j=1}^{n} \boldsymbol{E}_{3,j}^{(k+1)} \right]^{-1}, \qquad (2.6) \\
\boldsymbol{\Sigma}^{(k+1)} &= \frac{1}{n} \sum_{j=1}^{n} \Big[e_{1,j}^{(k+1)} \tilde{\boldsymbol{y}}_j^{(k+1)} \tilde{\boldsymbol{y}}_j^{(k+1)^T} - \tilde{\boldsymbol{y}}_j^{(k+1)} \boldsymbol{e}_{2,j}^{(k+1)^T} \boldsymbol{\Delta}^{(k+1)^T} \\
&\quad - \boldsymbol{\Delta}^{(k+1)} \boldsymbol{e}_{2,j}^{(k+1)} \tilde{\boldsymbol{y}}_j^{(k+1)^T} + \boldsymbol{\Delta}^{(k+1)} \boldsymbol{E}_{3,j}^{(k+1)} \boldsymbol{\Delta}^{(k+1)^T} \Big] \qquad (2.7)
\end{aligned}$$

와 같이 갱신한다. 단, $\tilde{\boldsymbol{y}}_j^{(k)} = \boldsymbol{y}_j - \boldsymbol{\mu}^{(k)}$을 나타낸다. 그리고 자유도 ν에 대해서는

$$\nu^{(k+1)} = \underset{\nu}{\operatorname{argmax}} \sum_{j=1}^{n} \left[\frac{\nu}{2} \log\left(\frac{\nu}{2}\right) - \log\Gamma\left(\frac{\nu}{2}\right) + \left(\frac{\nu}{2}\right) \left(e_{4,j}^{(k+1)} - e_{1,j}^{(k+1)} \right) \right]$$

를 만족하도록 구한다.

한편 식 (2.7) 대신에

$$\boldsymbol{\Sigma}^{(k+1)} = \frac{1}{n} \sum_{j=1}^{n} \left[e_{1,j}^{(k+1)} \tilde{\boldsymbol{y}}_j^{(k+1)} \tilde{\boldsymbol{y}}_j^{(k+1)^T} - \frac{1}{2} \left(\tilde{\boldsymbol{y}}_j^{(k+1)} \boldsymbol{e}_{2,j}^{(k+1)^T} \boldsymbol{\Delta}^{(k+1)^T} + \boldsymbol{\Delta}^{(k+1)} \boldsymbol{e}_{2,j}^{(k+1)} \tilde{\boldsymbol{y}}_j^{(k+1)^T} \right) \right]$$

를 사용할 수 있는데, 그 이유는 식 (2.6)–(2.7)의 계산 순서를 지키면

$$2 \sum_{j=1}^{n} \boldsymbol{\Delta}^{(k+1)} \boldsymbol{E}_{3,j}^{(k+1)} \boldsymbol{\Delta}^{(k+1)^T} = \sum_{j=1}^{n} \left[\tilde{\boldsymbol{y}}_j^{(k+1)} \boldsymbol{e}_{2,j}^{(k+1)^T} \boldsymbol{\Delta}^{(k+1)^T} + \boldsymbol{\Delta}^{(k+1)} \boldsymbol{e}_{2,j}^{(k+1)} \tilde{\boldsymbol{y}}_j^{(k+1)^T} \right]$$

를 만족하기 때문이다. 이로써 $\boldsymbol{\Sigma}^{(k+1)}$을 구하는데 조금이나마 계산 시간을 줄일 수 있다.

그러나 본 논문에서 보다 주목하는 주된 문제는, E-step에서 식 (2.4)과 식 (2.5)의 두 조건부 기대값 $\boldsymbol{e}_{2,j}^{(k+1)}$과 $\boldsymbol{E}_{3,j}^{(k+1)}$를 구하기 위해서는 q-변량 절단 t-분포의 1, 2차 적률 $\boldsymbol{m}_j^{(k+1)}$과 $\boldsymbol{M}_j^{(k+1)}$를 계산해야 한다는 점이다. 두 적률에 대한 계산은 최근 Ho 등 (2012)에 의해 보다 정확한 계산 공식이 제공되었다. 그러나 잘 알려져 있는 바와 같이 다변량 절단 분포의 적률은 큰 계산 시간을 요한다. 특히 관측치의 개수 n과 차원 q가 큰 경우 MST의 적합은 비실용적일 수 밖에 없다.

본 논문의 연구 동기는 다음과 같다.

$\boldsymbol{e}_{2,j}^{(k+1)} = e_{1,j}^{(k+1)} \boldsymbol{m}_j^{(k+1)}$ 및 $\boldsymbol{E}_{3,j}^{(k+1)} = e_{1,j}^{(k+1)} \boldsymbol{M}_j^{(k+1)}$으로 계산된 $Q_1(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$은 반복이 진행함에 따라 증가하며 결국 로그-우도를 증가시킨다. 여기서 $\boldsymbol{e}_{2,j}^{(k+1)}$과 $\boldsymbol{E}_{3,j}^{(k+1)}$ 대신 어떤 $q \times 1$ 벡터 $\dot{\boldsymbol{e}}_{2,j}^{(k+1)} = e_{1,j}^{(k+1)} \dot{\boldsymbol{m}}_j$ 및 어떤 $q \times q$(양정치) 대각행렬 $\dot{\boldsymbol{E}}_{3,j}^{(k+1)} = e_{1,j}^{(k+1)} \dot{\boldsymbol{M}}_j$이 포함된 Q-함수를 $\dot{Q}_1(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$라 하자. 이때 $\dot{Q}_1$의 증가가 Q_1의 증가를 의미하는 $\dot{\boldsymbol{m}}_j$와 $\dot{\boldsymbol{M}}_j$를 찾는다면,

$$\begin{aligned}
\boldsymbol{\Delta}^{(k+1)} &= \left(\boldsymbol{\delta}_1^{(k+1)}, \ldots, \boldsymbol{\delta}_q^{(k+1)} \right) \\
&= \left[\sum_{j=1}^{n} \tilde{\boldsymbol{y}}_j^{(k+1)} \dot{\boldsymbol{e}}_{2,j}^{(k+1)^T} \right] \left[\sum_{j=1}^{n} \dot{\boldsymbol{E}}_{3,j}^{(k+1)} \right]^{-1} \\
&= \left\{ \frac{\sum_{j=1}^{n} \tilde{\boldsymbol{y}}_j^{(k+1)} \dot{e}_{2,hj}^{(k+1)}}{\sum_{j=1}^{n} \dot{e}_{3,h,j}^{(k+1)}} \right\}_{h=1}^{q} = \left\{ \frac{\sum_{j=1}^{n} \tilde{\boldsymbol{y}}_j^{(k+1)} e_{1,j}^{(k+1)} \dot{m}_{hj}^{(k+1)}}{\sum_{j=1}^{n} e_{1,j}^{(k+1)} \dot{M}_{hj}^{(k+1)}} \right\}_{h=1}^{q}
\end{aligned}$$

와 같이 $\boldsymbol{\Delta}^{(k+1)}$은 $\boldsymbol{\delta}_h^{(k+1)}$ 별로 분리해서 얻을 수 있다. 즉, 계산시간이 상대적으로 짧은 단변량 계산 방식으로 q번의 수행으로 처리할 수 있을 것이다. 단, $\dot{\boldsymbol{m}}_j = (\dot{m}_{1j}, \ldots, \dot{m}_{qj})^T$ 및 $\dot{\boldsymbol{M}}_j = \mathrm{diag}(\dot{M}_{1j}, \ldots, \dot{M}_{qj})$을 나타낸다.

이에 대한 구체적인 내용은 다음 절에서 다룬다.

3. 치우침 모수 $\boldsymbol{\Delta}$에 대한 편추정

3.1. 제안된 방법

우선 $\boldsymbol{\Delta}\boldsymbol{x}_j = \boldsymbol{\delta}_1 x_{1j} + \cdots + \boldsymbol{\delta}_q x_{qj}$이므로 로그-우도 (2.2)는

$$L_{c1}(\boldsymbol{\theta}) \propto \sum_{j=1}^{n} \left[-\frac{u_j}{2} \left(\boldsymbol{y}_j - \boldsymbol{\mu} - \sum_{h=1}^{q} \boldsymbol{\delta}_h x_{hj} \right)^T \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{y}_j - \boldsymbol{\mu} - \sum_{h=1}^{q} \boldsymbol{\delta}_h x_{hj} \right) \right]$$

와 같이 나타낼 수 있다. 이제 $(k+1)$번째 반복에서 $Q_1(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$를 최대화 하기 위해 본 논문에서 적용하는 방법은 다음과 같다. Meng과 van Dyk (1997)의 AECM(Alternating ECM) 알고리즘의 원리를 이용하여,

1 단계: 먼저 $\boldsymbol{x} = \dot{\boldsymbol{m}}_j^{(k)}$ 및 $\boldsymbol{\Delta} = (\boldsymbol{\delta}_1^{(k)}, \ldots, \boldsymbol{\delta}_q^{(k)})$로서 주어졌다는 가정 하에서

$$Q\left(\boldsymbol{\mu}, \boldsymbol{\Sigma} | \boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)}\right) = E_{\boldsymbol{\Delta}^{(k)}}[L_{c1}(\boldsymbol{y}, \boldsymbol{u}; \boldsymbol{\theta}) | \boldsymbol{y}, \boldsymbol{x}]$$

을 최대화 한다. 이 단계의 결과로 E-step에서 $u_j^{(k+1)} = E(U_j|\boldsymbol{y}_j, \boldsymbol{x}_j) = e_{1,j}^{(k+1)}$과 M-step에서 $\boldsymbol{\mu}^{(k+1)}$과 $\boldsymbol{\Sigma}^{(k+1)}$을 얻는다.

2 단계: 그리고 $U_j = e_{1,j}^{(k+1)}$과 $\boldsymbol{\mu} = \boldsymbol{\mu}^{(k+1)}$ 및 $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{(k+1)}$로서 주어졌다는 가정 하에서

$$Q\left(\boldsymbol{\Delta} | \boldsymbol{\Delta}^{(k)}\right) = E_{\boldsymbol{\theta}^{(k)}}[L_{c1}(\boldsymbol{y}, \boldsymbol{x}; \boldsymbol{\Delta}) | \boldsymbol{y}, \boldsymbol{u}]$$

를 최대화 한다. 그러나 이것을 최대화 하는 대신 이 단계에서 다시 AECM 원리를 사용하여

Partial Update (PU) 단계: $\boldsymbol{X}_{(h)} = (x_{1j}^{(k+1)}, \ldots, x_{h-1,j}^{(k+1)}, x_{h+1,j}^{(k)}, \ldots, x_{qj}^{(k)})^T$ 및 $\boldsymbol{\Delta}_{(h)} = (\boldsymbol{\delta}_1^{(k+1)}, \ldots, \boldsymbol{\delta}_{h-1}^{(k+1)}, \boldsymbol{\delta}_{h+1}^{(k)}, \ldots, \boldsymbol{\delta}_q^{(k)})$가 주어졌다는 조건 하에서

$$\dot{Q}\left(\boldsymbol{\delta}_h | \boldsymbol{\delta}_h^{(k)}\right) = E_{\boldsymbol{\delta}^{(k)}}\left[L_{c1}(\boldsymbol{y}, \boldsymbol{x}; \boldsymbol{\delta}_h) | \boldsymbol{y}, \boldsymbol{x}_{(h)}^{(k)}\right], \quad h = 1, \ldots, q$$

를 최대화 한다. 이 단계의 결과로 E-step에서 $E(X_{hj}|\boldsymbol{y}_j, u_j^{(k+1)}) = \dot{m}_{hj}^{(k+1)}$과 $E(X_{hj}^2|\boldsymbol{y}_j, u_j^{(k+1)}) = \dot{M}_{hj}^{(k+1)}$ 그리고 M-step에서 $\boldsymbol{\delta}_h^{(k+1)}$을 얻는다. 가능하면 PU 단계를 몇 번 반복한다.

1–2 단계를 반복하면 $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$를 증가시킴으로써 결국 로그-우도를 증가키게 된다. 좀 더 구체적으로 표현하자면 다음과 같다.

1 단계에서는 E-step에서 $e_{1,j}^{(k+1)}$을 식 (2.6)과 동일한 수식으로 얻으며, M-step에서는 $\boldsymbol{\mu}$와 $\boldsymbol{\Sigma}$의 추정치로서

$$\boldsymbol{\mu}^{(k+1)} = \frac{1}{\sum_{j=1}^{n} e_{1,j}^{(k+1)}} \left(\sum_{j=1}^{n} e_{1,j}^{(k+1)} \boldsymbol{y}_j - \boldsymbol{\Delta}^{(k)} \sum_{j=1}^{n} \dot{\boldsymbol{e}}_{2,j}^{(k+1)} \right),$$

$$\boldsymbol{\Sigma}^{(k+1)} = \frac{1}{n} \sum_{j=1}^{n} \left[e_{1,j}^{(k+1)} \tilde{\boldsymbol{y}}_j^{(k+1)} \tilde{\boldsymbol{y}}_j^{(k+1)^T} - \frac{1}{2} \left(\tilde{\boldsymbol{y}}_j^{(k+1)} \dot{\boldsymbol{e}}_{2,j}^{(k+1)^T} \boldsymbol{\Delta}^{(k)^T} + \boldsymbol{\Delta}^{(k)} \dot{\boldsymbol{e}}_{2,j}^{(k+1)} \tilde{\boldsymbol{y}}_j^{(k+1)^T} \right) \right]$$

를 얻는다. 여기서 $\dot{\boldsymbol{e}}_{2,j}^{(k+1)} = (\dot{e}_{2,1j}^{(k+1)}, \ldots, \dot{e}_{2,qj}^{(k+1)})^T = e_{1j}^{(k+1)}(\dot{m}_{1j}^{(k+1)}, \ldots, \dot{m}_{qj}^{(k+1)})^T$를 나타낸다.
그리고 2 단계에서는

$$\begin{aligned}\dot{Q}\left(\boldsymbol{\delta}_h|\boldsymbol{\delta}_h^{(k)}\right) &= \sum_{j=1}^{n} E_{\boldsymbol{\Theta}^{(k)}}\Big[\log L_{c1}\left(\boldsymbol{Y}_j, U_j, \boldsymbol{X}_j; \boldsymbol{\delta}_h\right)\big|\boldsymbol{y}_j, U_j = e_{1,j}^{(k+1)},\\ &\qquad \Big\{X_{lj} = m_{lj}^{(k+1)}\ (l<h)\Big\}, \Big\{X_{lj} = m_{lj}^{(k)}\ (l>h)\Big\}\Big]\\ &= \sum_{j=1}^{n} E_{\boldsymbol{\Theta}^{(k)}}\left[-\frac{e_{1,j}^{(k+1)}}{2}\left(\tilde{\boldsymbol{r}}_{hj}^{(k)} - \boldsymbol{\delta}_h X_{hj}\right)^T \boldsymbol{\Sigma}^{(k)^{-1}}\left(\tilde{\boldsymbol{r}}_{hj}^{(k)} - \boldsymbol{\delta}_h X_{hj}\right)\Big|\boldsymbol{y}_j\right]\end{aligned}$$

을 최대화 하는데, 여기서

$$\begin{aligned}\tilde{\boldsymbol{r}}_{hj}^{(k)} &= \tilde{\boldsymbol{y}}^{(k)} - \sum_{l<h}^{q}\boldsymbol{\delta}_l^{(k)}\dot{m}_{lj}^{(k+1)} - \sum_{l>h}^{q}\boldsymbol{\delta}_l^{(k)}\dot{m}_{lj}^{(k)}\\ &= \tilde{\boldsymbol{y}}^{(k)} - \frac{\sum_{l<h}^{q}\boldsymbol{\delta}_l^{(k+1)}\dot{e}_{2,lj}^{(k)} + \sum_{l>h}^{q}\boldsymbol{\delta}_l^{(k)}\dot{e}_{2,lj}^{(k)}}{e_{1,j}^{(k)}},\end{aligned}$$

그리고 $\tilde{\boldsymbol{y}}_j^{(k)} = \boldsymbol{y}_j - \boldsymbol{\mu}^{(k)}$ 및 $\dot{m}_{lj}^{(k)} = \dot{e}_{2,lj}^{(k)}/e_{1,j}^{(k)}$을 나타낸다.
이때 $\dot{Q}(\boldsymbol{\delta}_h|\boldsymbol{\delta}_h^{(k)})$를 $\boldsymbol{\delta}_h$ $(h = 1, \ldots, q)$에 관하여 최대화하면

$$\boldsymbol{\delta}_h^{(k+1)} = \frac{\sum_{j=1}^{n} e_{1,j}^{(k+1)}\dot{m}_{hj}^{(k+1)}\tilde{\boldsymbol{r}}_{hj}^{(k+1)}}{\sum_{j=1}^{n} e_{1,j}^{(k+1)}\dot{M}_{hj}^{(k+1)}} = \frac{\sum_{j=1}^{n}\dot{e}_{2,hj}^{(k+1)}\tilde{\boldsymbol{r}}_{hj}^{(k)}}{\sum_{j=1}^{n}\dot{e}_{3,hj}^{(k+1)}}, \quad h = 1, \ldots, q \tag{3.1}$$

을 얻게 된다. 이제 문제는 E-step에서 단변량 값인 두 조건부 기대값 $\dot{e}_{2,hj}^{(k+1)}$과 $\dot{e}_{3,hj}^{(k+1)}$을 구하는 것인데, 다소 긴 유도과정은 생략하고 여기서는 그 결과만 제공하기로 하겠다. 즉,

$$\begin{aligned}\dot{e}_{2,hj}^{(k+1)} &\overset{\text{let}}{=} E_{\boldsymbol{\Theta}^{(k)}}\Big[U_j X_{hj}|\boldsymbol{y}_j, U_j = e_{1,j}^{(k+1)}, \Big\{X_{lj} = \dot{m}_{lj}^{(k+1)}, l<h\Big\}, \Big\{X_{lj} = \dot{m}_{lj}^{(k)}, l<h\Big\}\Big]\\ &= \tilde{x}_{hj}^{(k)} e_{1,j}^{(k+1)} + \psi_h^{(k)} B_{hj}^{(k)}, \quad h = 1, \ldots, q,\\ \dot{e}_{3,hj}^{(k+1)} &\overset{\text{let}}{=} E_{\boldsymbol{\Theta}^{(k)}}\Big[U_j X_{hj}^2|\boldsymbol{y}_j, U_j = e_{1,j}^{(k+1)}, \Big\{X_{lj} = \dot{m}_{lj}^{(k+1)}, l<h\Big\}, \Big\{X_{lj} = \dot{m}_{lj}^{(k)}, l<h\Big\}\Big]\\ &= \tilde{x}_{hj}^{(k)}\dot{e}_{2,j}^{(k+1)} + {\psi_h^{(k)}}^2, \quad h = 1, \ldots, q\end{aligned}$$

이다. 단,

$$B_{hj}^{(k+1)} = \sqrt{\frac{\nu^{(k)}+p}{\nu^{(k)}+D_{hj}^{(k)}}} \times \frac{t_1\left(\frac{\tilde{x}_{hj}^{(k)}}{\psi_h^{(k)}}\sqrt{\frac{\nu^{(k)}+p}{\nu^{(k)}+D_{hj}^{(k)}}}; \nu^{(k)}+p\right)}{T_1\left(\frac{\tilde{x}_{hj}^{(k;s)}}{\psi_h^{(k)}}\sqrt{\frac{\nu^{(k)}+p}{\nu^{(k)}+D_{hj}^{(k)}}}; \nu^{(k)}+p\right)}, \quad h = 1, \ldots, q$$

을 나타내며, 그리고 각각 $\boldsymbol{\Omega}_h^{(k)} = \boldsymbol{\Sigma}^{(k)} + \boldsymbol{\delta}_h^{(k)}\boldsymbol{\delta}_h^{(k)^T}$, $\tilde{x}_{hj}^{(k)} = \boldsymbol{\delta}_h^{(k)^T}\boldsymbol{\Omega}_h^{(k)^{-1}}\tilde{\boldsymbol{r}}_{hj}^{(k)}$, ${\psi_h^{(k)}}^2 = 1 - \boldsymbol{\delta}_h^{(k)^T}\boldsymbol{\Omega}_h^{(k)^{-1}}\boldsymbol{\delta}_h^{(k)}$, $D_{hj}^{(k)} = \tilde{\boldsymbol{r}}_{hj}^{(k)^T}\boldsymbol{\Omega}_h^{(k)^{-1}}\tilde{\boldsymbol{r}}_{hj}^{(k)}$ 을 나타낸다.
앞으로 이 알고리즘을 "SPU(sequentially partial update)"-EM 알고리즘이라 부르도록 하겠다. 주목하고자 하는 것은 SPU-EM 알고리즘은 q-변량 절단 t-분포의 두 적률을 계산하는데 계산시간이 월등히 짧은 단변량 t-pdf와 t-cdf를 사용한다는 것이다. 따라서 Exact-EM 알고리즘에 비해 처리시간이 훨씬 줄어들 것으로 기대한다.

3.2. 정칙화 버전

여러 실험에 따른 경험에 의하면 SPU-EM 알고리즘에 의해 적합된 분포는 소위 "에지영역(edge area)"를 두드러지게 잘 표현한다. 에지영역이란 자료의 비대칭의 원인으로서 은닉변수의 절단 $\boldsymbol{X}_j > \mathbf{0}$의 경계영역에 대응하는 관측 자료 $\boldsymbol{y}_j \in R^p$ 공간 상의 영역이다. 그러나 종종 절단면이 두드러지게 나타나 있는 자료에 대해서 SPU-EM은 최대 우도를 가지기 위해 몇몇 $\boldsymbol{\delta}_h^{(k)^T}\boldsymbol{\Omega}^{(k)^{-1}}\boldsymbol{\delta}_h^{(k)}$이 거의 1이 될 때까지 알고리즘을 진행 시키는 경향이 있다. 이것은 결국 행렬 $\boldsymbol{\Psi}^{(k)} = \boldsymbol{I}_q - \boldsymbol{\delta}_h^{(k)^T}\boldsymbol{\Omega}^{(k)^{-1}}\boldsymbol{\delta}_h^{(k)}$을 0-정치 행렬이 되도록 만들어 MST가 정의되지 않게 한다.

본 논문에서는 이 문제를 제어하기 위해 평활상수 $\alpha \geq 0$을 도입하여 식 (3.1) 대신

$$\boldsymbol{\delta}_h^{(k+1)} = \frac{\sum_{j=1}^n \dot{e}_{2,hj}^{(k+1)} \tilde{\boldsymbol{r}}_{hj}^{(k)}}{\left\{\alpha + \sum_{j=1}^n \dot{e}_{3,hj}^{(k+1)}\right\}}, \quad h = 1, \ldots, q$$

를 사용할 것이다.

평활상수 α를 크게 줄수록 치우침 모수의 추정치 $\boldsymbol{\delta}_h^{(k)}$는 좀 더 $\mathbf{0}$에 가까워짐으로써 $\boldsymbol{\Psi}^{(k)}$의 대각원소가 0이 되는 상황을 통제할 수 있다. 5절의 실험에서는 공히 평활상수를 극히 작은 값인 $\alpha = 10^{-5}$으로 설정하여 0-정치 발생을 막을 수 있었다.

4. 혼합모형으로의 확장

SPU-EM 알고리즘을 혼합모형으로 확장하는 방법은 그렇게 어렵지 않다. 본 논문에서는 자세한 전개는 지양하고 그 결과만 가능한 한 간략히 제공한다.

g개의 성분을 가지는 MST 분포의 혼합모형은 다음과 같다.

$$\begin{aligned} f\left(\boldsymbol{y}_j; \boldsymbol{\Theta}\right) &= \sum_{i=1}^g \pi_i f_i\left(\boldsymbol{y}_j; \boldsymbol{\theta}_i\right), \quad j = 1, \ldots, g \\ &= 2^q \sum_{i=1}^g \pi_i t_p\left(\boldsymbol{y}_j; \boldsymbol{\mu}_i, \boldsymbol{\Omega}_i, \nu_i\right) T_q\left(\sqrt{\frac{\nu_i + p}{\nu + D_{ij}}}\tilde{\boldsymbol{x}}_{ij}; \boldsymbol{\Psi}_i, \nu_i + p\right), \end{aligned}$$

여기서 π_i는 i번째 성분의 혼합비율(mixing proportion)로서 $\sum_{i=1}^g \pi_i = 1$을 만족한다. 그리고 $\boldsymbol{\Omega}_i = \boldsymbol{\Sigma}_i + \boldsymbol{\Delta}_i\boldsymbol{\Delta}_i^T$, $\tilde{\boldsymbol{x}}_{ij} = \boldsymbol{\Delta}_i^T\boldsymbol{\Omega}_i^{-1}(\boldsymbol{y}_j - \boldsymbol{\mu}_i)$, $\boldsymbol{\Psi}_i = \boldsymbol{I}_q - \boldsymbol{\Delta}_i^T\boldsymbol{\Omega}_i^{-1}\boldsymbol{\Delta}_i$, $D_{ij} = (\boldsymbol{y}_j - \boldsymbol{\mu}_i)^T\boldsymbol{\Omega}_i^{-1}(\boldsymbol{y}_j - \boldsymbol{\mu}_i)$를 나타낸다.

한편 $\boldsymbol{y}_j = (y_{1j}, \ldots, y_{pj})^T$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

$$\begin{aligned} \boldsymbol{Y}_j = \boldsymbol{y}_j | \left(z_{ij} = 1, \boldsymbol{X}_j = \boldsymbol{x}_j, U_j = u_j\right) &\sim N_p\left(\boldsymbol{\mu}_i + \boldsymbol{\Delta}_i\boldsymbol{x}_j, \frac{\boldsymbol{\Sigma}_i}{u_j}\right), \\ \boldsymbol{X}_j = \boldsymbol{x}_j | \left(z_{ij} = 1, U_j = u_j\right) &\sim N_q\left(\mathbf{0}, \frac{\boldsymbol{I}}{u_j}\right), \\ U_j | z_{ij} = 1 &\sim \text{gamma}\left(\frac{\nu_i}{2}, \frac{\nu_i}{2}\right), \\ \boldsymbol{Z}_j = \left(Z_{1j}, \ldots, Z_{gj}\right)^T &\sim \text{multinomial}\left(1, \pi_1, \ldots, \pi_g\right) \end{aligned}$$

와 같이 고려할 수 있다. 이때 자유도 ν_i들을 제외한 모수벡터 $\boldsymbol{\theta}$에 대응하는 완비자료의 로그-우도는 다음과 같다. 즉,

$$L_{c1}(\boldsymbol{\theta}) \propto \sum_{j=1}^{n}\sum_{i=1}^{g}\left[-\frac{z_{ij}}{2}\log u_j\,|\boldsymbol{\Sigma}_i| - \frac{1}{2}z_{ij}u_j\left(\boldsymbol{y}_j - \boldsymbol{\mu}_i - \sum_{h=1}^{q}\boldsymbol{\delta}_{hi}x_{hj}\right)^T\boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{y}_j - \boldsymbol{\mu}_i - \sum_{h=1}^{q}\boldsymbol{\delta}_{hi}x_{hj}\right)\right].$$

이제 SPU-EM 알고리즘은 다음과 같이 구성하게 된다.

우선, 사후확률 추정치

$$\tau_{ij}^{(k+1)} = E_{\boldsymbol{\Theta}^{(k)}}\left(Z_{ij}=1|\boldsymbol{y}_j\right) = \frac{\pi_i^{(k)} f_i\left(\boldsymbol{y}_j;\boldsymbol{\theta}_i^{(k)}\right)}{\sum_{i'=1}^{g}\pi_{i'}^{(k)} f_{i'}\left(\boldsymbol{y}_j;\boldsymbol{\theta}_{i'}^{(k)}\right)}, \quad i=1,\ldots,g$$

를 구한다. 그리고 $\pi_i^{(k+1)} = n_i^{(k+1)}/n$ $(i=1,\ldots,g)$을 계산한다.

그리고 다음을 수행한다.

1 단계의 E-step에서는

$$E_{\boldsymbol{\Theta}^{(k)}}\left(Z_{ij}U_j|z_{ij}=1,\boldsymbol{y}_j\right) = \tau_{ij}^{(k+1)}e_{1,ij}^{(k+1)}$$

를 구한다. 여기서 $e_{1,ij}^{(k+1)}$는

$$\begin{aligned} e_{1,ij}^{(k+1)} &= E_{\boldsymbol{\Theta}^{(k)}}\left(U_j|z_{ij}=1,\boldsymbol{y}_j\right) \\ &= \frac{\nu_i^{(k)}+p}{\nu_i^{(k)}+D_{ij}^{(k)}}\frac{T_q\left(\sqrt{\frac{\nu_i^{(k)}+p+2}{\nu_i^{(k)}+D_{ij}^{(k)}}}\tilde{\boldsymbol{x}}_{ij}^{(k)};\boldsymbol{\Psi}_i^{(k)},\nu_i^{(k)}+p+2\right)}{T_q\left(\sqrt{\frac{\nu_i^{(k)}+p}{\nu_i^{(k)}+D_{ij}^{(k)}}}\tilde{\boldsymbol{x}}_{ij}^{(k)};\boldsymbol{\Psi}_i^{(k)},\nu_i^{(k)}+p\right)} \end{aligned}$$

와 같이 얻는다. 그리고 M-step에서는

$$\begin{aligned} \boldsymbol{\mu}_i^{(k+1)} &= \frac{\sum_{j=1}^{n}\tau_{ij}^{(k+1)}\left(e_{1,ij}^{(k+1)}\boldsymbol{y}_j - \boldsymbol{\Delta}_i^{(k+1)}\dot{\boldsymbol{e}}_{2,ij}^{(k+1)}\right)}{\sum_{j=1}^{n}\tau_{ij}^{(k+1)}e_{1,ij}^{(k+1)}}, \quad i=1,\ldots,g, \\ \boldsymbol{\Sigma}_i^{(k+1)} &= \frac{1}{n_i^{(k+1)}}\sum_{j=1}^{n}\tau_{ij}^{(k+1)}\Big[e_{1,ij}^{(k+1)}\tilde{\boldsymbol{y}}_{ij}^{(k)}\tilde{\boldsymbol{y}}_{ij}^{(k)^T} \\ &\quad - \frac{1}{2}\left(\tilde{\boldsymbol{y}}_{ij}^{(k)}\dot{\boldsymbol{e}}_{2,ij}^{(k+1)^T}\boldsymbol{\Delta}_i^{(k+1)^T} + \boldsymbol{\Delta}_i^{(k+1)}\dot{\boldsymbol{e}}_{2,ij}^{(k+1)}\tilde{\boldsymbol{y}}_{ij}^{(k)^T}\right)\Big], \quad i=1,\ldots,g \end{aligned}$$

을 갱신한다. 여기서 $n_i^{(k+1)} = \sum_{j=1}^{n}\tau_{ij}^{(k+1)}$ 및 $\tilde{\boldsymbol{y}}_{ij}^{(k)} = \boldsymbol{y}_j - \boldsymbol{\mu}_i^{(k)}$을 나타낸다.

2 단계인 SPU 과정을 수행한다. E-step에서는 $\dot{e}_{2,hij}^{(k+1)}$ 및 $\dot{e}_{3,hij}^{(k+1)}$를 각 성분 $i=1,\ldots,g$ 별로 앞 절에서 소개한 방법과 같은 방법으로 계산하고, M-step에서

$$\boldsymbol{\delta}_{hi}^{(k+1)} = \frac{\sum_{j=1}^{n}\tau_{ij}^{(k+1)}\dot{e}_{2,hij}^{(k+1)}\tilde{\boldsymbol{r}}_{hij}^{(k+1)}}{\sum_{j=1}^{n}\tau_{ij}^{(k+1)}\dot{e}_{3,hij}^{(k+1)}}$$

를 갱신한다. 단,

$$\tilde{\boldsymbol{r}}_{hij}^{(k+1)} = \tilde{\boldsymbol{y}}_{ij}^{(k)} - \frac{\sum_{l=1}^{h-1}\boldsymbol{\delta}_{li}^{(k+1)}\dot{e}_{2,lij}^{(k+1)} + \sum_{l=h+1}^{q}\boldsymbol{\delta}_{li}^{(k)}\dot{e}_{2,lij}^{(k)}}{e_{1,ij}^{(k+1)}},$$

Table 5.1. Execution Times and Goodness-of-Fits

q	Iterations	Method	Etime (sec.)	log-likelihood	AIC	BIC	Misallocation
1	400	Pyne-EM	5.52	−1984.5	4022.9	4112.2	8
2	400	Exact-EM	980.40	−1977.8	4021.7	4130.8	9
		SPU-EM	**323.26**	**−1975.9**	**4017.9**	**4127.0**	**8**
3	200	Exact-EM	2977.50	−1976.1	4030.2	4159.2	8
		SPU-EM	**905.70**	−1979.4	4036.7	4165.8	**4**

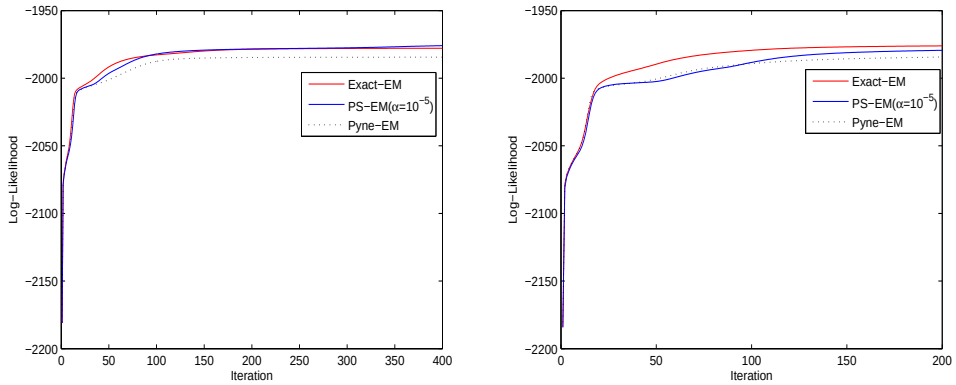

Figure 5.1. Increments of log-likelihood over the iterations. (a)Left: $q = 2$, (b)Right: $q = 3$. Dot line: Pyne-EM, red line: Exact-EM, blue line: SPU-EM.

$\tilde{\boldsymbol{y}}_{ij}^{(k)} = \boldsymbol{y}_j - \boldsymbol{\mu}_i^{(k)}$를 나타낸다.

마지막으로 1, 2 단계를 마친 후

$$\nu_i^{(k+1)} = \underset{\nu}{\operatorname{argmax}} \sum_{j=1}^{n} \left[\frac{\nu}{2} \log\left(\frac{\nu}{2}\right) - \log\Gamma\left(\frac{\nu}{2}\right) + \left(\frac{\nu}{2}\right)\left(e_{4,ij}^{(k+1)} - e_{1,ij}^{(k+1)}\right) \right], \quad i = 1, \ldots, g$$

와 같이 자유도 추정치를 갱신한다.

5. 실험

이 절에서는 Cook과 Weisberg (1994)에서 제공된 AIS(Australian Institution of Sport) 자료를 이용하여 Exact-EM 알고리즘과 비교하면서 제안된 SPU-EM 알고리즘의 성능을 보인다. AIS 자료는 100명의 여성과 102명의 남성 육상선수의 11개 특성을 측정한 자료인데, 우리는 이들 특성 중에 3 변량 Ht(height: Y_1), Bfat(percentage of body fat: Y_2), LBM(lean body mass: Y_3)만을 사용하여 실험한다.

본 실험에서는 $q = 1, 2, 3$에 대해 SPU-EM이 Exact-EM의 결과와 얼마나 비슷한지 알아보고자 하는 것이다. $q = 1$일 때, 즉 치우침 모수 행렬이 $\boldsymbol{\Delta} = \boldsymbol{\delta}$와 같이 1 벡터를 가지는 경우로서 이때는 Exact-EM이나 SPU-EM이 모두 동일한 결과를 제공하게 되는데 결국 Pyne 등 (2009)의 MST를 추정하는 것이므로 Pyne-EM이라 호칭하였다. 그리고 두 알고리즘의 공정한 비교를 위해 동일한 초기치를 사용하였다.

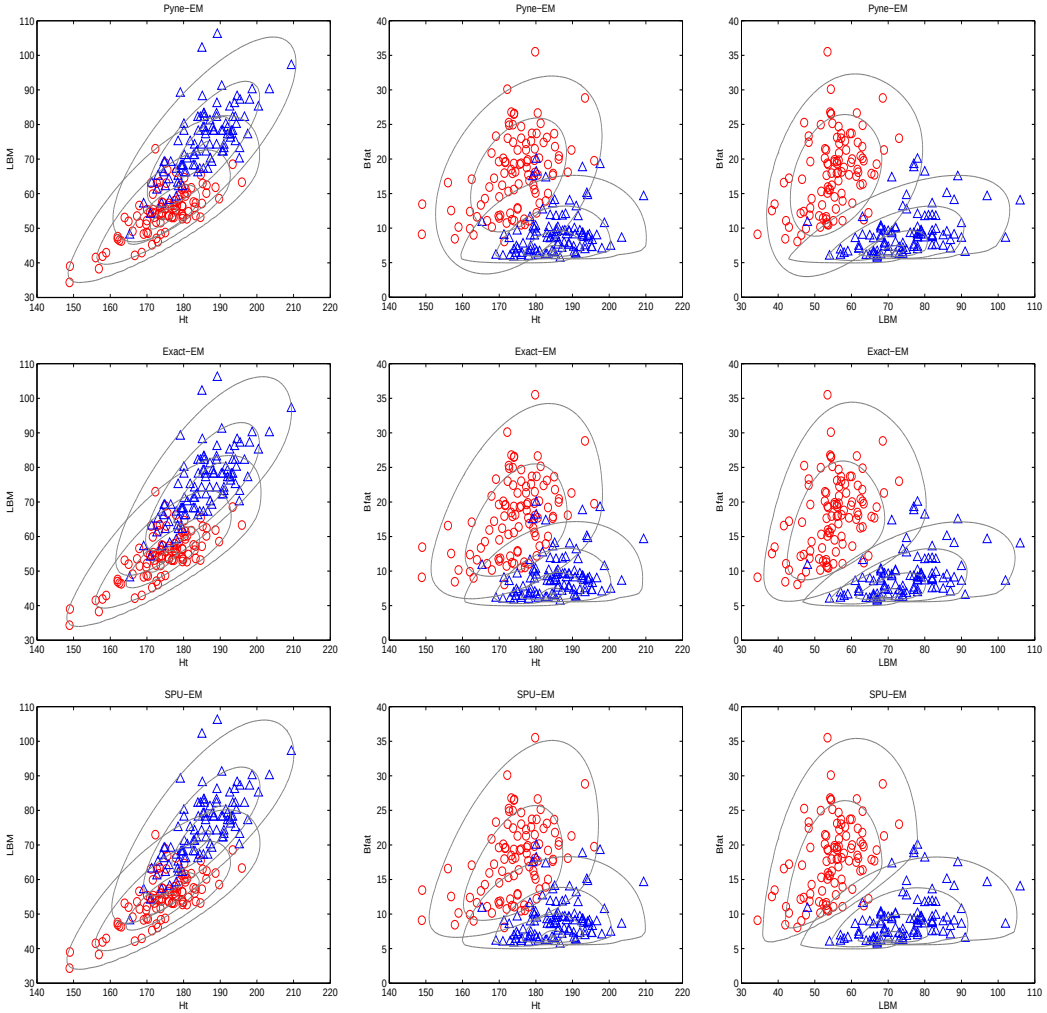

Figure 5.2. Contour plots of fitted model with $q = 2$. ◯: female, △: male. 1^{st} row: Pyne-EM, 2^{nd} row: Exact-EM, 3^{rd} row: SPU-EM.

$q = 2$인 경우 즉 $\boldsymbol{\Delta} = (\boldsymbol{\delta}_1, \boldsymbol{\delta}_2)$일 때 두 알고리즘을 400회 충분히 반복하여 처리 시간과 적합도를 Table 5.1에 나타내었다. 이 경우 Exact-EM이 약 980초(약 16분)이 걸린 반면 SPU-EM의 경우는 약 323초(약 5.3분)이 소요되어 제안된 방법이 3배 이상 빠른 처리 속도를 보였다. 그럼에도 로그-우도는 Exact-EM보다 약간 더 나은 결과를 나타내었다. 한편 Figure 5.1.(a) (왼쪽)에서 두 알고리즘의 반복에 따른 로그-우도의 증가 형태가 약간 다름을 보이고 있는데 이는 SPU-EM이 Exact-EM과는 다른 루트를 경유해 (국소) 최대우도에 도달함을 암시한다 하겠다. 그리고 이들의 적합 결과를 Figure 5.2에 나타내었는데 (남녀 자료 별로 적합한 것이 아니라 혼합모형 적합후 추정된 성분 별로 등고선을 그린 것임.), 첫 행은 Pyne-EM의 결과로서 남자 자료(△)에 비해 여자 자료(◯)의 비대칭이 잘 적합되어 있지 않다. 반면, Exact-EM과 SPU-EM의 결과는 비슷한데, SPU-EM의 결과가 소위 "에지영역"을 좀 더 두드러지게 잘 표현하고 있다.

$q = 3$인 경우 즉 $\boldsymbol{\Delta} = (\boldsymbol{\delta}_1, \boldsymbol{\delta}_2, \boldsymbol{\delta}_3)$일 때는 반복횟수가 200회로 물론 충분한 반복을 진행하지 않았지만 SPU-EM의 로그-우도가 Exact-EM보다 다소 못미치는 결과로 나타났다. 그러나 SPU-EM의 수행시간이 약 905초(약 15분)인 반면 Exact-EM은 약 2977초(약 50분)이나 소요되었다. 이것은 MST 적합을 위한 Exact-EM의 비실용성을 잘 보여주는 것이라 하겠다.

한편, 적합 후 얻은 사후확률 추정치를 이용하여

$$\text{cluster}_j = \underset{i}{\text{argmax}}\, \hat{\tau}_{ij}$$

를 통해 관측치 j를 최대 사후확률에 대응하는 성분 i로 할당하여 군집하여 보았다. Table 5.1의 마지막 열에 나타낸 군집 후 오할당 결과는 SPU-EM이 나머지 두 방법보다 같거나 좋았다. 그 이유는 적합된 밀도가 에지영역을 좀 더 예리하게 표현함으로써 군집 사이를 보다 확실하게 구분하려는 특성 때문인 것으로 보인다.

6. 결론 및 토의

본 연구에서 저자는 다변량 치우친 t-분포(MST) 혼합모형 적합을 위한 Lee와 McLachlan (2013, 2014a, 2014b)의 Exact-EM을 AECM 알고리즘의 원리를 도입하여 처리시간의 비실용성을 완화시키고자 SPU-EM 알고리즘을 제안하였다. 그 결과 약 3배 정도 빠른 처리시간으로 Exact-EM과 유사한 결과를 얻을 수 있음을 보였다. 그리고 자료에 따라서는 제안된 SPU-EM이 Exact-EM보다 더 좋은 결과를 얻을 수도 있다는 가능성을 확인하였다.

그럼에도 불구하고 SPU-EM 알고리즘도 "실용적"이라고 말할 수 있을 정도의 처리시간을 보인 것은 아니다. 사실 혼합모형 사용의 유용성은 다양한 성분의 개수 g의 보장에서 비롯한다. 그러나 보통 크기의 표본의 개수에서 조차 막대한 처리시간 때문에 성분의 개수를 3개 이상으로 시도해 보기 조차 어려웠다. 따라서 좀 더 획기적인 수준의 개량이 있어야 비대칭 자료에 대한 혼합모형의 사용을 현실적으로 보장할 수 있을 것으로 보인다.

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.

Cook, R. D. and Weisberg, S. (1994). *An Introduction to Regression Graphics*, **56**, 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. (2012). Diagnosis of observations after fit of multivariate skew t-distribution: Identification of outliers and edge observations from asymmetric Data, *The Korean Journal of Applied Statistics*, **25**, 1019–1026.

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.

Meng. X. L, and van Dyk, D. A. (1997). The EM-algorithm and old folk song sung to a fast new tune, *Journal of the Royal Statistical Society B*, **59**, 511–567.

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, *Proc. Natl. Acad. Sci. USA*, **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.

치우친 다변량 t-분포 혼합모형에 대한 최우추정

김승구[a,1]

[a]상지대학교 데이터정보학과

(2014년 9월 1일 접수, 2014년 9월 22일 수정, 2014년 9월 22일 채택)

요 약

치우친 다변량 t-분포 혼합을 적합하기 위해 Exact-EM 알고리즘이 사용된다. 그러나 이 방법은 E-step에서 매우 긴 처리시간을 요하는 다변량 절단 t-분포의 적률을 계산해야 한다. 본 논문에서는 이러한 문제점을 완화하기 위해 SPU-EM이라 명명한 알고리즘을 제안하는데, 이것은 Meng과 van Dyk (1997)의 AECM 알고리즘의 원리를 이용하여 다차원 적률의 계산상의 어려움을 해결한다. 결과적으로 제안된 방법은 Exact-EM 알고리즘 보다 빠른 처리시간으로 보장한다. 이를 입증하기 위해 실험을 통해 제안된 방법의 유효성을 보인다.

주요용어: 치우친 다변량 t-분포, 혼합모형, EM 알고리즘, AECM 알고리즘.

연구는 상지대학교 2013 교내 연구비 지원에 의해 수행되었음.

[1](220-702) 강원도 원주시 우산동, 상지대학교 데이터정보학과. E-mail: sgukim@sangji.ac.kr