분포 혼합비율의 모수추정을 위한 효율적인 알고리즘에 관한 연구

*황강진, **박경탁, ***유희경
*강릉영동대학 인터넷사무정보전공, **동국대학교 통계학과
***삼척대학교 컴퓨터공학과

A Study for Efficient EM Algorithms for Estimation of the Proportion of a Mixed Distribution

*Kangjin Hwang, **Gyungtag Park, ***Heekyung Yoo
*Internet Office Information Course, Kangnung Yeongdong college
**Dept. of Statistics, Dongguk University
***Dept. of Computer Engineering, Samchok National University

Key Words : 동적 EM 알고리즘, Titterington 알고리즘

Abstract

EM algorithm has good convergence rate for numerical procedures which converges on very small step. In the case of proportion estimation in a mixed distribution which has very big incomplete data or of update of new data continuously, however, *EM* algorithm highly depends on a initial value with slow convergence ratio. There have been many studies to improve the convergence rate of *EM* algorithm in estimating the proportion parameter of a mixed data.

Among them, dynamic *EM* algorithm by Murray Jorgensen and *Titterington* algorithm by D. M. Titterington are proven to have better convergence rate than the standard *EM* algorithm, when a new data is continuously updated.

In this paper we suggest dynamic *EM* algorithm and *Titterington* algorithm for the estimation of a mixed Poisson distribution and compare them in terms of convergence rate by using a simulation method.

1. 서 론

혼합모형(Mixture model)은 생물통계 뿐만 아니라 여러 통계적 응용분야에서 비동질적 모집단을 설명하는데 널리 사용된다. 비동질성은 생물통계 모집단에서 더욱 일반적인 현상이므로 혼합모형은 이러한 모집단을

설명하는데 유용하다(Dimitris Karlis & *Evdokia Xekalaki, 1999*). 특히 최근에는 컴퓨터의 연산능력이 급격히 발달하면서 그러한 혼합모형의 추정을 쉽게 하는 효율적인 알고리즘 개발이 가능하게 되었다.

EM 알고리즘은 결측자료(missing data), 절단분포(truncated distributi on), 중도절단자료(censored data) 등 불완전한 자료(incomplete data)가 주어졌을 때 이 자료가 가지고 있는 정보를 이용해서 완전한 자료를 도출한 후 최우추정량(MLE) 등과 같은 함수의 최대값을 찾고자 하는 반복 알고리즘 방법이다(Dempster, Laird, Ruvin, 1977). 이러한 알고리즘은 여러 회에 걸쳐 반복하는 수치해석적 모수추정 과정이 필요하다.

EM 알고리즘은 매우 작은 단계로 수렴하는 수치해석적 방법에 대하여 좋은 수렴 정도를 갖는 것으로 알려져 있다. 그러나, 매우 큰 불완전 자료를 갖는 혼합비율 분포의 모수를 추정하는 경우 *EM* 알고리즘과 같은 방법은 모수를 추정하는데 있어 느린 수렴정도를 갖고 초기값에 높이 의존한다(D. M. Titterington, 1984).

이와 같이 혼합비율(mixture proportion)의 모수추정에서 느린 수렴정도를 갖는 *EM* 알고리즘을 어떻게 개선하여 실용화 할 것인가는 여러 학자들에 의해 연구되었다. Bohning et al.(1994)은 반복의 회수를 줄이면서 추정값을 찾는 쉬운 방법을 제안하였고, Fruman & Lindsay(1994)는 *EM* 알고리즘의 초기값으로 적률추정값을 사용하는 효율적인 초기값의 사용을 제안하였다. 또 다른 방법은 McLachland & Krishnan(1997, Chapter 4)이 조사한 빠른 수렴방법을 찾는 것이다.

본 연구에서는 새로운 관측치가 사용됨으로서 추정이 연속적으로 새롭게 이루어지는 보조적인 접근법을 논의할 것이다. 또한 한 번의 모수추정 과정에서 연속적으로 순환과정이 이루어짐으로써 수렴의 정도를 높일 수 있는 *Titterington*알고리즘 방법을 논의할 것이다. 이러한 과정은 다음과 같은 식으로 표현 가능하다.

$$\theta^{(n+1)} = G_n(\theta^{(n)}, y_{n+1}),$$

$$n = 0, 1, \cdots,$$

여기서 θ는 모수를, $\theta^{(n)}$은 n개 관측자료에서의 추정치를, y_{n+1}은 $(n+1)$번째의 관측치를 나타낸다. 이것은 각각의 연속된 관측치가 알려진 후 반복은 단지 적은 회수로 이루어지는 연속된 추정과정을 제안한다.

또한 *Titterington*알고리즘 방법을 논의하는데 있어 두 개의 알려진 분포를 포함하는 혼합분포모형에서 한 구성분포의 비율에 관한 추정 문제와 같이 단순한 하나의 모수 문제를 상세하게 다룰 것이다.

위에서 언급한 혼합비율분포는 포아송 분포(Poisson distribution)를 모델로 하였고, 이러한 모델을 통한 실험을 하기 위해 1장에서 표준 *EM*알고리즘을, 2장에서 동적 *EM*알고리즘(Dynamic *EM* algorithm)에 대하여, 3장에서는 *Titterington* 알고리즘의 개념을 설명한다. 4장에서는 혼합분포의 혼합비율 추정을 설명하고, 5장에서는 포아송 분포를 이용한 동적 *EM*알고리즘, *Titterington* 알고리즘의 모의실험을 통해 모수추정 결과에 대해 종합적인 비교를 하고자 한다.

2. 본 론

2.1 *EM* 알고리즘

A. P. Dempster, N. M. Laird, 그리고 D. B. Rubin으로부터 1977년에 소개된 *EM* 알고리즘은 다양한 불완전한 자료(incomplete data)로부터 최우추정치(MLE)를 반복적인 기법(iterative method)을 통해 구할 수 있는 방법으로 *E*(Expectation)-단계와 *M*(Maximization)-단계로 구성되어 모수를 추정한다.

확률변수 $\{y_1, \cdots, y_n\}$는 모수 θ에 의한 확률분포함수 $f(y \mid \theta)$를 갖고, y는 관측된 불완전한 자료라고 가정한다. 불완전자료는 자료의 손실 때문만이 아니라 자료를 모으는 과정에서 관측할 수 없기 때문에 발생할 수도 있다. $\{x_1, \cdots, x_n\}$는 확률분포 $f(x \mid \theta)$를 갖는 완전자료이고, z를 잠재자료(latent data)라 하면, 완전자료 형태의 확률변수 $X=(Y, Z)$로 표현 가능하다. 보통 x_i는 다차원이고, y_i는 관측되어진 x_i로 구성된다.

완전자료의 로그우도함수(Log Likelihood Function)는

$$l_c(\theta) = \sum_{i=1}^{n} \log f(x_i \mid \theta)$$

이다. 우리는

$$l(\theta) = \sum_{i=1}^{n} \log f_{obs}(y_i \mid \theta)$$

를 최대로 하는 θ를 찾고자 한다. 이때 $f_{obs}(y_i)$는 x_i의 관측되지 않은 부분을 제외한 관측된 자료를 적분하거나 합함으로써 함수 $f(x_i)$로부터 얻어진다. 독립된 관측치에 대한 *EM* 알고리즘은 반복법 $\theta^{(n+1)} = g(\theta^{(n)})$을 요구한다. 이때 함수 $g(\cdot)$는 $g(\theta')$이 함수 $Q(\theta, \theta')$를 최대화하는 θ의 값이 되도록 정의되고, 실제 모수벡터는 θ'임을 가정한다. 또한 $Q(\theta, \theta')$는 관측된 자료 $\{y_1, \cdots, y_n\}$이 주어진 완전자료 우도함수 l_c의 조건부 기대값이다. 우리는 g를 자료 $\{y_1, \cdots, y_n\}$을 근거로 하는 *EM*-map이라 한다.

이를 정리하면 다음과 같다.

Step-1 : *E*-단계(Expectation step)

초기추정치를 θ'라 하고 아래의 수식에 따라 기대값을 구한다

$$\begin{aligned} Q(\theta, \theta') &= E_{\theta'}[l_c(\theta) \mid y_1, \cdots, y_n] \\ &= E_{\theta'}[\sum \log f(x_i \mid \theta) \mid y_1, \cdots, y_n] \\ &= \int \sum \log f(x_i \mid \theta) f(z \mid y_i, \theta') dz \end{aligned}$$

잠재변수 Z가 이산확률변수인 경우 함수 $Q(\theta, \theta')$는 다음과 같이 표현된다. 실제적으로 응용하는 경우에는 이 형태가 주로 사용된다.

$$\begin{aligned} Q(\theta, \theta') &= \sum_{i=1}^{n} \sum_{z \in Z} \log f(x_i \mid \theta) f(z \mid y_i, \theta') \\ &= \sum_{i=1}^{n} \sum_{z \in Z} \log f(x_i \mid \theta) \\ &\quad \times f(y_i, z \mid \theta') / f(y_i \mid \theta') \end{aligned}$$

Step-2 : *M*-단계(Maximization step)

$Q(\theta, \theta')$을 최대화하는 $\theta^{(1)}$은 다음을 이

용하여 구한다.

$$\partial Q(\theta^{(1)}, \theta')/\partial\theta = 0$$

Step-3 : 반복과정

$\theta' = \theta^{(1)}$라 하여 E-단계를 계산하고 다시 최대화 단계에서 $Q(\theta, \theta^{(1)})$을 최대화하는 $\theta^{(2)}$를 구하는 순환과정 $\theta^{(k+1)} = g(\theta^{(k)})$을 반복한다. 이러한 순환을 통해 *EM* 알고리즘은 θ^*값에 수렴하게 된다(Martin A. Tanner, 1993).

2.2 동적 *EM*알고리즘

EM 알고리즘이 실행되는 그 순간 새로운 관측치가 관측되는 상황을 고려하자. g_n은 관측 가능한 자료 $\{y_1, \cdots, y_n\}$을 근거로 하는 *EM*-map이라 한다. 자료 $\{y_1, \cdots, y_m\}$과 이에 대응하는 m개의 자료에서 추정된 추정치 $\theta^{(m)}$을 가지고 시작한다. 이때, 우리는 수렴값을 찾기 위해 반복 g_m을 시행하고, 자료 $\{y_1, \cdots, y_m\}$으로부터 θ의 최우추정량이 되는 $\theta^{(m)}$을 얻을 수 있다. 또한 이러한 방법을 통해 계산상의 효율성을 얻는다. 자료의 초기집합 크기 m은 $\theta^{(m)}$의 신뢰도를 위해 충분히 커야 한다. 그러면, 각각의 새로운 표본 관측치가 생성되므로 다음 식 (1)에 의하여 새로운 모수추정이 이루어진다.

$$\theta^{(n+1)} = g_{n+1}(\theta^{(n)}) \qquad (1)$$

달리 표현하자면, *EM*-map은 모수 벡터의 다음 반복을 실행하기 이전에 새로이 관측된 자료를 포함한 모든 자료의 모수를 추정하기 위해 기존 자료에서 추정된 모수를 초기추정치로 이용하여 새로운 *EM* 알고리즘의 과정을 실행하게 된다. 우리는 이러한 과정을 동적 *EM*알고리즘(Dynamic *EM* algorithm)이라 하고, 식 (1)에서 추정된 동적 *EM*알고리즘의 $(n+1)$단계 최우추정식은 다음과 같다.

$$Q(\theta, \theta^{(n)}) = \sum_{i=1}^{n+1} E_{\theta^{(n)}}[\log f(x_i \mid \theta) \mid y_i] \quad (2)$$

이에 대한 설명은 혼합비율 추정의 예를 통해 더 상세히 설명된다.

2.3. *Titterington* 알고리즘

Titterington 알고리즘을 요약하기에 앞서 순환과정(Recusive procedure)을 살펴보기로 하자.

독립적으로 관측된 자료 $\{y_1, \cdots, y_n\}$은 확률분포 $f(y \mid \theta)$를 갖는다. 이때 임의의 s에 대하여 $\theta \in \Theta \subset R^s$ 이다. $S(y, \theta)$를 다음과 같이 정의한다.

$$S_j(y, \theta) = \frac{\partial}{\partial\theta_j} \log f(y \mid \theta)$$
$$, j = 1, \cdots, s$$

$D^2(y, \theta)$는 $\log f(y \mid \theta)$의 2차 미분된 행렬, $I(\theta)$는 관측치에 대응하는 Fisher의 정보행렬이다. 모든 모수에 대해 미분 가능(조건 ①)하고 기대값이 존재(조건 ②)하면서 다음 식이 성립(조건 ③)함을 가정하자.

$$E_\theta[\, S(y,\theta)] = \int S(y,\theta)g(y\mid\theta)dy = 0$$

$$I(\theta) = E_\theta[S(y,\theta)S'(y,\theta)]$$

$$= -E\,[D^2(y,\theta)]$$

그러면 통계적 근사 과정인 순환과정은 다음과 같다.

$$\theta^{(n+1)} = \theta^{(n)} + \{nI(\theta^{(n)})\}^{-1} \times S(y_{n+1}, \theta^{(n)}) \tag{3}$$

여기서 θ는 모수를, $\theta^{(n)}$은 n개 관측자료에서의 추정치를, y_{n+1}은 $(n+1)$번째의 관측치를 나타낸다. 위에서 언급된 가정들과 일반조건(regularity condition) 하에서 순환과정은 식 (4)와 같다.

$$\theta^{(n+1)} = \theta^{(n)} + \{(n+1)I(\theta^{(n)})\}^{-1} \times S(y_{n+1}, \theta^{(n)}) \tag{4}$$

Titterington(1984, p. 259)은 불완전 자료의 문제에 있어 식 (3)과 (4)는 각각 식 (5), (6)로 대체될 것을 제안하고 있다.

$$\theta^{(n+1)} = \theta^{(n)} + \{nI_c(\theta^{(n)})\}^{-1} S(y_{n+1}, \theta^{(n)}) \tag{5}$$

$$\theta^{(n+1)} = \theta^{(n)} + \{(n+1)I_c(\theta^{(n)})\}^{-1} S(y_{n+1}, \theta^{(n)}) \tag{6}$$

여기서 $I_c(\theta)$는 완전자료에 대응하는 Fisher의 정보행렬이다.

Titterington(1984, p. 264)은 다음과 같이 정의된 *EM* 알고리즘의 반복적 방법(귀납적 방법, 순환적 방법; recursive version)을 소개하였다. 최근 모수추정량을 $\theta^{(n)}$과 추정된 로그우도함수 $l_n(\theta)$를 가지고 새로이 관측된 자료를 포함한 $(n+1)$단계를 고려한다면, $(n+1)$에서 모수추정량과 추정된 로그우도함수는 각각 $\theta^{(n+1)}$과 $l_{n+1}(\theta)$로 표현되고 로그우도함수는 식 (7)과 같이 정의한다.

$$l_{n+1}(\theta) = E_{\theta^{(n)}}[\log f(x_{n+1}\mid\theta) \mid y_{n-1}] + l_n(\theta) \tag{7}$$

여기서 모수추정량 $\theta^{(n+1)}$은 $l_{n+1}(\theta)$를 최대화하는 θ의 값이다.

EM 알고리즘과 순환적 방법인 *Titterington* 알고리즘은 사후분포를 최대화하는 모수값을 찾기 위한 베이지안 통계분석에 응용된다. 식 (7)에서 초기값으로 $L_0(\theta) = \log P(\theta)$를 사용하고, 이때 $P(\cdot)$는 최대값으로 θ_0를 갖는 θ에 대한 사전확률 분포이다.

Theorem.(Titterington)

근사적으로 적합한 조건 하에서, 식 (7)은 다음과 같이 순환과정 식으로 표현된다.

$$\hat{\theta}^{(n+1)} = \hat{\theta}^{(n)} + \{(n+1)I_c(\hat{\theta}^{(n)})\}^{-1} \times S(y_{n+1}, \hat{\theta}^{(n)})$$

동적 *EM* 알고리즘과 *Titterington* 알고리즘은 연속된 상황에서 *EM* 알고리즘보다 계

산상의 절약을 보여준다(Dimitris Karlis & Evdokia Xekalaki, 1999).

*Titteringtong*알고리즘은 우도에서 단지 한 항의 추정을 요구한다. 이에 반하여 동적 *EM* 알고리즘은 각 단계에서 재 추정이 이루어지기 위하여 모든 항을 필요로 한다. 그러나 Murray Jorgensen(1999)은 정보손실 비율이 큰 경우 동적 *EM* 알고리즘이 *Titterington* 알고리즘보다 더 효율적임을 보여주었다.

2.4. 혼합분포의 비율 추정

동적 *EM* 알고리즘의 사용은 예제로 자세히 설명된다.

자료 { y 1 , *cdots*, y n}는

분포 $\pi f_1(y)+(1-\pi)f_2(y)$에서 추출한 이산형 확률변수(random variable)이다. 이때 분포 f_1과 f_2는 실수에 대해 양의 값을 갖는 알려진 분포임을 가정한다. *EM* 알고리즘에서 완전자료는 불완전 관측자료 y_i와, 함수 f_1에 대해 지시변수인 z_i의 쌍($x_i=(y_i, z_i)$)으로 구성되어 있다. 완전자료 $\{x_1, \cdots, x_n\}$를 근거로 한 우도함수는 식 (8)과 같고, 로그우도함수는 (9)와 같다.

$$L_c=\prod_{i=1}^{n}[\pi f_1(y_i)]^{z_i}[(1-\pi)f_2(y_i)]^{1-z_i} \quad (8)$$

$$\begin{aligned} l_c &= \sum_{i=1}^{n} z_i[\log\pi+\log f_1(y_i)] \\ &+ \sum_{i=1}^{n}(1-z_i)[\log(1-\pi)+\log f_2(y_i)] \end{aligned} \quad (9)$$

먼저 추정해야할 모수의 초기값은 π'라고 하자. 그러면 초기 모수가 주어졌을 때 완전한 로그우도함수의 기대값을 $Q(\pi, \pi')$라고 하면 다음과 같이 나타낼 수 있다.

$$Q(\pi, \pi')=E_{\pi'}[l_c \mid y_i] \quad (10)$$

그런데 l_c는 관측되지 않은 지시변수 z_i의 선형함수이므로 관측값 y_i와 초기 모수들이 주어지는 경우인 식 (10)의 우변은 간단히 z_i의 조건부 기대값으로 바꾸어 구할 수 있다. 그리고 이때의 $E_{\pi'}[z_i \mid y_i]$는 확률이 되면서 0과 1사이의 값을 갖는다. 이를 전개해보면 식 (11)과 같고, *EM* 알고리즘의 반복식 $g_n(\pi')=\frac{1}{n}\sum_{i=1}^{n}q(\pi', y_i)$을 얻을 수 있다.

$$\begin{aligned} &E[z_i \mid \{y_1, \cdots, y_n\}, \pi'] \\ &= E[z_i \mid y_i, \pi'] = P[z_i=1 \mid y_i, \pi'] \\ &= \pi' f_1(y_i)/\{\pi' f_1(y_i)+(1-\pi')f_2(y_i)\} \\ &= q(\pi', y_i) \end{aligned} \quad (11)$$

고정된 자료 $\{y_1, \cdots, y_n\}$에 대하여 *EM*알고리즘은 반복 $\pi^{(n+1)}=g_n(\pi^{(n)})$을 요구한다. 이때 $n=0, 1, \cdots$이고, $\pi'=\pi^{(0)}=0.5$이다.

우리가 자료 $\{y_1, \cdots, y_n\}$과 대응추정량 $\pi^{(n)}$을 갖고 있다하자. 그리고 새로운 자료 y_{n+1}을 관측했다면 $\pi^{(n)}$의 갱신된 동적 *EM* 알고리즘은

$$\begin{aligned} \pi^{(n+1)} &= g_{n+1}(\pi^{(n)}) \\ &= \{1/(n+1)\}\sum_{i=1}^{n+1}q(\pi^{(n)}, y_i) \end{aligned} \quad (12)$$

이다. 이와 달리 *Titterington* 알고리즘은

성립조건 ①, ②, ③을 만족하면 $\pi^{(n)}$을 다음 식으로 대체할 수 있다.

$$\pi^{(n+1)} = \pi^{(n)} + \{(n+1)I_c(\pi^{(n)})\}^{-1} \times (y_{n+1}, \pi^{(n)})$$

여기서, $S(y, \pi)$는 관측치 y에 대응하는 점수인 $\frac{\partial}{\partial \pi} f_{obs}(y, \pi)$이고, $I_c(\pi)$는 완전 자료 x 에 대한 Fisher의 정보량이다. 우리는 불완전 자료 y 에 대한 Fisher의 정보량으로 $I(\pi)$를 사용할 것이다. 현재 모형에 대하여 *Titterington* 알고리즘은 식 (13)과 같이 계산된다.

$$\begin{aligned}\pi^{(n+1)} &= \pi^{(n)} + [(n+1)\times\{1/(\pi^{(n)}(1-\pi^{(n)}))\}]^{-1} \\ &\times\{q(\pi^{(n)}, y_{n+1}) - \pi^{(n)}\}/\{\pi^{(n)}(1-\pi^{(n)})\} \\ &= \pi^{(n)} + \{q(\pi^{(n)}, y_{n+1}) - \pi^{(n)}\}/(n+1)\end{aligned} \tag{13}$$

2.5. 모의실험

2.5.1. 모형의 소개

실제로 두 알고리즘이 어떻게 사용되는가는 모의 실험을 통한 결과로 얻어질 수 있다.

자료 $\{y_1, \cdots, y_n\}$는 혼합분포 $\pi f_1(y) + (1-\pi) f_2(y)$에서 추출된 자료로 $f_1(y) = e^{-1}/y!$, $(y=0,1,\cdots)$, $f_2(y) = (e^{-k}k^y)/y!$, $(y=0,1,\cdots; k>0)$인 포아송 분포이다. 이때 혼합비율 추정값은 $\pi=0.3, \pi=0.6$ 으로 가정하고 $k=1.3$, $1.5, \cdots, 3.1$ 으로 가정하고, 에 대해 각 100개의 자료이다. 두 알고리즘은 초기 추정치 $\pi^{(0)}=0.5$로 이루어지고, 마지막 값 $\pi^{(100)}$은 저장된다. 이는 k의 각 값에 대하여 50회씩 반복된다.

포아송 분포를 이용한 $q(\pi', y_i)$함수는 식 (11)를 통해 다음과 같다.

$$q(\pi', y_i) = (\pi' e^{-1}) / (\pi' e^{-1} + (1-\pi')k^{y_i}e^{-k})$$

$\pi^{(n)}$은 *EM* 알고리즘의 반복식에 의해 얻어진 최우추정치라고 가정하면 동적 *EM* 알고리즘의 반복식은 식 (12)에 의해 다음과 같이 얻어진다.

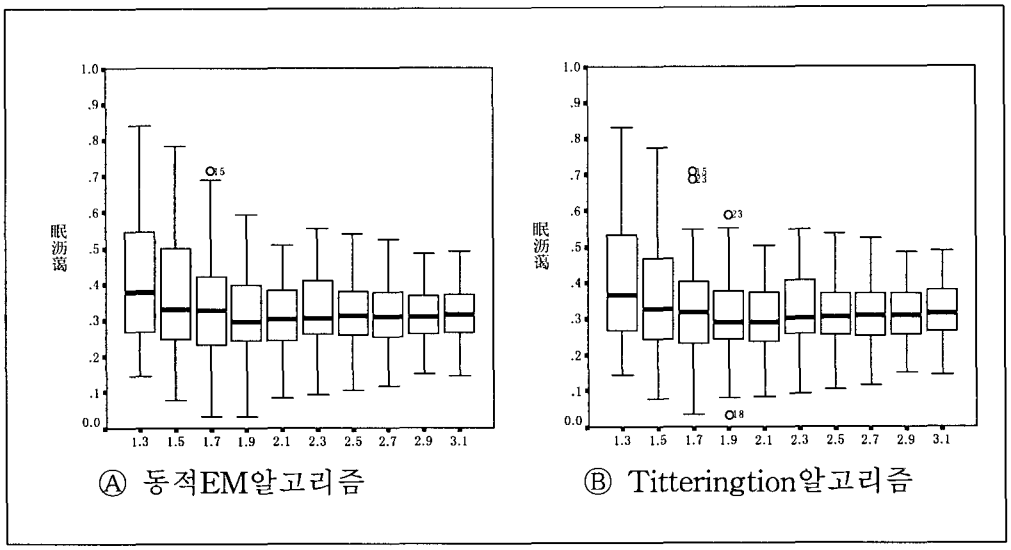
$$\begin{aligned}\pi^{(n+1)} &= g_{n+1}(\pi^{(n)}) \\ &= \{1/(n+1)\}\sum_{i=1}^{n+1} q(\pi', y_i) \\ &= \{1/(n+1)\} \\ &\times \sum_{i=1}^{n+1}(\pi' e^{-1})/\{\pi' e^{-1} + (1-\pi')k^{y_i}e^{-k}\}\end{aligned}$$

Titterington 알고리즘의 반복식은 식 (13)을 이용하여 다음과 같이 얻어진다.

$$\pi^{(n+1)} = \left(\frac{n}{n+1}\right)\pi^{(n)} + \pi^{(n)}e^{-1} / \{(n+1)(\pi^{(n)}e^{-1} + (1-\pi^{(n)})k^{y_{n+1}}e^{-k})\}$$

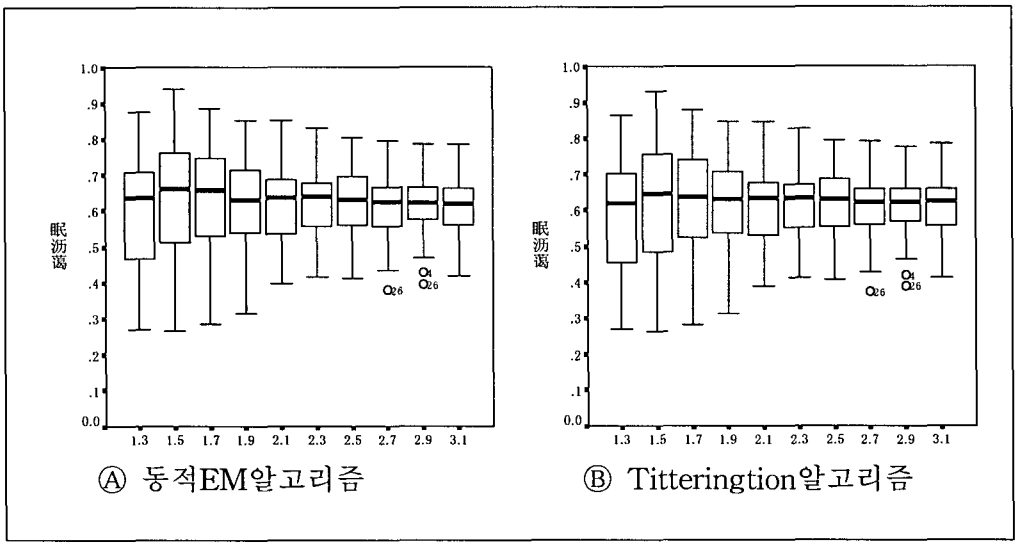
2.5.2. 결과

[그림 1]과 [그림 2]는 각각 π가 0.3, 0.6 일 때 두 알고리즘에 의한 혼합비율 추정값을 나타낸 것이다. [그림 1]과 [그림 2]의 Ⓐ는 100개 자료에서 *EM* 알고리즘을 통해 추정된 추정값을 초기추정치로 이용한 동적



Ⓐ 동적EM알고리즘 Ⓑ Titteringtion알고리즘

[그림 1] 혼합비율추정값($\pi = 0.3$)



Ⓐ 동적EM알고리즘 Ⓑ Titteringtion알고리즘

[그림 2] 혼합비율추정값($\pi = 0.6$)

EM 알고리즘의 추정값을 나타낸다. 또한 각 그림의 Ⓑ는 같은 추정값을 초기추정치로 이용한 *Titterington* 알고리즘의 추정값이다.

[그림 1], [그림 2]에서 알 수 있듯이 관측된 자료에 하나의 자료가 추가로 측정되었을 때, 기존 관측자료에서 추정된 추정값을 이용한 두 알고리즘의 혼합비율 추정값에는 차이가 없는 것으로 나타났다. 그러나 [표 1]에

[표 1] 평균 수렴횟수

구분 \ π		1.3	1.5	1.7	1.9	2.1	2.3	2.5	2.7	2.9	3.1
$k=0.3$	동적 *EM*	80.01	69.26	59.81	47.36	38.05	30.38	24.19	20.81	16.85	15.33
	Titterintton	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
$k=0.6$	동적 *EM*	31.08	20.72	14.03	8.60	3.81	1.40	1.00	1.00	1.00	1.00
	Titterintton	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

서 보여주듯이 평균 수렴횟수는 동적 *EM* 알고리즘이 각 단계에서 재 추정이 이루어지므로 *Titterington* 알고리즘보다 많은 반복이 이루어지고 있다.

동적 *EM* 알고리즘과 *Titterington* 알고리즘은 [그림 1], [그림 2]에 나타나듯이 같은 결과를 보여주고 있다. 그러나 [표 1]의 결과와 같이 평균 수렴횟수에서는 *Titterington* 알고리즘이 작은 단일반복으로 동일한 수렴값을 갖는 것으로 나타났다.

이는 자료집합의 수정 후 혼합비율 추정은 기존 자료에서 추정된 결과값을 이용하여 단일반복을 통해 추정하는 *Titterington* 알고리즘의 결과가 적합함을 나타낸다. 뿐만 아니라 수정된 자료집합을 이용한 반복식의 계산과정인 동적 *EM* 알고리즘에 비해 기존 자료의 추정값을 이용한 단일반복이 계산과정과 시간에서 단순화됨을 보여준다.

3. 결 론

EM 알고리즘은 불완전 표본에서 모수를 추정해야 하는 경우 기대와 극대의 반복과정을 통하여 모수의 최우추정치를 구하는 통계적 방법으로 혼합분포의 경우에도 활용 가능하다. 그러나, 혼합분포 추정의 경우에 있어서 *EM* 알고리즘은 느리게 수렴하기 때문에 새로운 자료가 연속적으로 생성되는 경우 이를 포함하는 새로운 모수 추정값(updated estimator)들의 계산에는 적합하지 않은 것으로 알려져 있다.(Dimitris Karlis & Evdokia Xekalaki, 1999).

본 논문은 혼합분포에서 새로운 자료가 연속적으로 생성되는 경우 *EM* 알고리즘이 혼합분포 혼합비율을 새로이 추정하는데 필요하게 되는 과다한 수렴 반복 횟수를 줄이는 방법에 대해 연구하였다. 그 방법으로는 기존자료에서 추정된 추정치를 이용한 단일반복 혼합비율 모수추정 방법인 Murray Jorgensen의 동적 *EM* 알고리즘과 D. M. Titterington의 *Titterintton* 알고리즘이 있으며, 모의실험을 통해 두 알고리즘의 수렴정도를 비교해 보았다.

모의실험 결과([그림 1], [그림 2]) 동적 *EM* 알고리즘과 *Titterintton* 알고리즘의 수렴값에는 큰 차이가 없었으나, 두 방법의 평균 수렴 횟수에는 많은 차이를 보였다. 그 이유는 *Titterington* 알고리즘은 우도에서 단지 한 항의 추정을 요구하는데 반하여 동적

EM 알고리즘은 각 단계에서 재 추정이 이루어지기 위하여 모든 항을 필요로 하기 때문이다. 그러나 Murray Jorgensen(1999)은 정보손실 비율이 큰 경우 동적 *EM* 알고리즘이 *Titterington* 알고리즘보다 더 효율적이라는 것을 보여주었다.

참고문헌

[1] Bohning, D., Dietz, EK., Schaub, R., Schlatman, P. and Lindsay, B.(1994) The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. *Annals of the Institute of Statistical Mathematics*, 46, pp. 373-388.

[2] Dempster, A. P., Laird, N. M. and Rubin, D. B (1977) Maximum likelihood from incomplete data via the EM algorithm(with discussion). *J. R. Statist.* Soc. B, 39, pp. 1-38

[3] Dimitris Karlis and Evdokia Xekalaki (1999) Improving the EM Algorithm for Mixtures. *Statistics and Computing*. 9, pp. 303-307.

[4] Everitt, B. S. & D. J. Hand (1981) Finite Mixture Distributions, Chapman & Hall : London.

[5] Fruman, W. D. & Lindsay, B. (1994) Measuring the relative effectiveness of moment estimators as starting values in maximizing mixture likelihoods. *Computational Statistics and Data Analysis*, 17, pp. 493-507.

[6] Martin A. Tanner (1993) *Tools for Statistical Inference*. Springer -Verlag : New York.

[7] McLachlan, G. J. and Krishnan, T. (1997) *The EM Algorithm and Extensions*. Wiley: New York.

[8] Murray Jorgensen (1999) A Dynamic EM Algorithm for Estimating Mixture Proportions. Statistics and Computing. 9, pp. 299-302.

[9] Titterington, D. M. (1984) Recursive parameter estimation using incomplete data. *J. R. Statist.* Soc. B, 46, pp. 257-67.