Journal of the Korean
Data & Information Science Society
2006, Vol. 17, No. 4, pp. 1209~1217

Estimation in Mixture of Shifted Poisson Distributions

Changhyuck Oh[1)]

Abstract

For the mixture of shifted Poisson distributions, a method of parameter estimation is proposed. The range of the shifted parameters are estimated first and for each shifted parameter set EM algorithm is applied to estimate the other parameters of the distribution. Among the estimated parameter sets, one with minimum likelihood for given data is to be set as the final estimate. In simulation experiments, the suggested estimation method shows to have a good performance.

Keywords : 우도, 혼합 이동 포아송분포, EM 알고리즘

1. 머리말

혼합분포에서의 모수의 추정은 Dempster 외(1977)에 의해 제시된 EM 알고리즘을 이용하여 추정하는 방법이 Liu 외(2006)에 의하여 소개되었다. EM 알고리즘을 이용한, 여러 가지 종류의 혼합분포에서의 모수 추정에 관한 연구가 이루어져 왔다. 이에 관하여는 McLachlan과 Peel(2001)과 Copsey와 Webb(2003)를 참조. 그 중 혼합포아송분포에 관하여도 많은 연구가 이루어져 왔으며 혼합포아송분포는 인터넷트래픽, 문서분류, 영상복원검색 등의 분야에서 응용되고 있다. Lee와 Oh(2006) 참조. 혼합분포에서의 EM 알고리즘의 적용은 관측된 자료를 불완전한 것으로 간주하는 것에서부터 시작한다. 즉, 자료에는 관측치를 생성한 혼합 성분에 관한 정보가 결측되어 있다고 가정하는 것이다.

한편 이동포아송분포는 포아송분포에서 이동모수를 추가된 모형이며, 이동모수의 값 보다 적은 값은 관측되지 않는 모형이다. 이러한 이동모수의 특성으로 인하여 이동포아송분포를 혼합한 경우에는 EM 알고리즘을 단순 적용하여 모수를 추정하는 데에는 어려움이 따른다. 이러한 어려움 때문에 Lee와 Oh(2006)은 이동모수가 고정된 경우에 EM 알고리즘을 적용하여 혼합이동포아송분포의 모수를 추정하는 방법을 제시하였다. 본 연구에서는 이를 바탕으로 이동모수가 미지인 경우에 혼합포아송분포의

1) 경북 경산시 대동 214-1 영남대학교 통계학과 교수
E-mail : choh@yu.ac.kr

모수를 추정하는 법을 살펴 본다. 제시된 방법의 기본적인 생각은 주어진 자료에 의해 추정 가능한 모든 이동모수의 각 쌍에 대하여 Lee와 Oh(2006)의 방법으로 다른 모수들을 추정한 후에, 이들 중에서 주어진 자료에 대한 우도를 최대로 하는 것을 선택하여 최종 추정치로 하는 것이다. 제 2절에서는 혼합 이동 포아송분포에서 이동모수를 비롯한 모든 모수가 미지인 경우에 모수추정 절차를 소개한다. 그리고, 제 3절에서는 제시된 모수 추정 절차의 효율성을 살펴 보기 위하여 시뮬레이션 실험을 한다. 마지막으로 제 4절에서는 토의와 결론을 다룬다.

2. 혼합 이동 포아송분포의 모수 추정 절차

혼합 이동 포아송분포 모형

$$f(x;\Phi)=\sum_{i=1}^{g}\pi_i f_i(x;\theta_i,k_i) \tag{1}$$

를 생각하자. 여기서 g는 혼합분포의 성분의 개수이고, 혼합가중치 $\pi_i(i=1,\dots,g)$는 $0<\pi_i<1$, $\pi_1+\cdots+\pi_g=1$를 만족하며, 각 성분확률함수 f_i는 이동 포아송분포의 확률함수로 다음과 같이 주어진다.

$$f_i(x;\theta_i,k_i)=\frac{\theta_i^{(x-k_i)}e^{-\theta_i}}{(x-k_i)!},\quad x-k_i=0,1,2,\dots \tag{2}$$

여기서 $\theta_i>0$는 위치모수, k_i는 정수인 이동모수이다. 한편 전체 모수 벡터를 $\Phi=(\pi_1,\dots,\pi_g;\theta_1,\dots,\theta_g;k_1,\dots,k_g)$로 나타내자.

혼합 이동 포아송분포 모형 (1)로 부터의 n개의 관측치 $\boldsymbol{x}=(x_1,\dots,x_n)$이 주어졌다고 하자. 이 자료에는 각각의 관측치 x_j, $j=1,\dots,n$, 가 어느 성분으로부터 관측되었는지에 관한 정보가 없으므로, 성분정보가 결측된 관측치로 가정되며, 관측치 x_j에 대응되는 확률함수 $f(x;\Phi)$를 불완전 자료에 대한 확률함수라고도 부른다. 주어진 관측치 $\boldsymbol{x}$에 대하여 모수 Φ의 로그우도함수는

$$L_{\boldsymbol{x}}(\Phi)=\sum_{j=1}^{n}\log f(x_j;\Phi) \tag{3}$$

로 주어진다. 최우추정치 $\hat{\Phi}_{\boldsymbol{x}}$는 로그우도함수 $L_{\boldsymbol{x}}(\Phi)$를 최대로 하는 Φ를 찾음으로 구할 수 있다. 그러나 혼합 이동 포아송분포 모형에 대하여 로그우도함수 $L_{\boldsymbol{x}}(\Phi)$를 직접 미분하는 등의 방법으로 $\hat{\Phi}_{\boldsymbol{x}}$를 구하는 것은 $L_{\boldsymbol{x}}(\Phi)$의 구조적 형태 때문에 매우 어려운 문제로 알려져 있다. 따라서 여기에서는 관측치 $\boldsymbol{x}$를 완전자료의 일부분으로 간주하고 반복적 방법으로 해를 구하는 EM 알고리즘을 적용하여 추정치를 구하는 방법을 제시한다. 관측치 x_j, $j=1,\dots,n$, 에는 관측치를 발생시킨 성분 정보가 결측되었다고 가정한다. 각 관측치 x_j를 발생시킨 성분확률함수의 정보를 나타내기 위하여 성분 벡터

$z_j = (z_{j1}, z_{j2}, \ldots, z_{jg})$를 생각한다. 벡터 z_j의 원소 $z_{j1}, z_{j2}, \ldots, z_{jg}$은 관측값을 발생시킨 원소에 대하여는 1, 나머지는 모두 0의 값을 취한다.

각 관측치에 대하여 성분정보도 함께 관측한 경우에 얻어지는 자료 $(x_1, z_1), \ldots, (x_n, z_n)$를 완전자료라고 부르며, 이에 대한 로그우도함수는

$$L_{\boldsymbol{x},\boldsymbol{z}}(\Phi) = \sum_{i=1}^{g}\sum_{j=1}^{n} z_{ji}\{\log\pi_i + \log f_i(x_j;\theta_i)\} \tag{4}$$

로 주어진다. 완전자료에 대한 모수의 최우추정치는 식 (4)로부터 식 (5)와 같이 주어짐을 쉽게 보일 수 있다.

$$\begin{aligned} \hat{k}_i &= \min\{x_j : z_{ji} > 0,\ j = 1, \ldots, n\}, \\ \hat{\pi}_i &= \frac{\sum_{j=1}^{n} z_{ji}}{n}, \\ \hat{\theta}_i &= \frac{\sum_{j=1}^{n} z_{ji} x_j}{\sum_{j=1}^{n} z_{ji}} - \hat{k}_i, \quad i = 1, 2, \ldots, g. \end{aligned} \tag{5}$$

여기서는 각 성분으로부터 최소한 한 개 이상의 관측치가 얻어졌다고 가정한다. 한편, 고려하고 있는 상황은 성분정보가 관측되지 않는 경우이므로, 식 (5)의 추정치를 사용할 수 없으므로, 먼저 성분정보를 추정한다. $p+1$번째 반복에서 EM 알고리즘의 E 단계는 관측치 $(x_1, \ldots, x_n)$가 주어진 조건 하에서 성분벡터 z_{ji}의 조건부기대치 $\hat{z}_{ji}$를 구한 후 추정된 완전자료 $(x_1, z_1^{(p)}), \ldots, (x_n, z_n^{(p)})$대하여 식 (5)의 $\hat{\pi}_i$와 $\hat{\theta}_i$를 이용하여 π_i와 θ_i를 추정한다.

성분벡터 $z_j,\ j = 1, \ldots, n,$ 의 추정을 위하여 먼저 이동모수의 집합

$$\mathbb{K} = \{K = (k_1, \ldots, k_g) : x_{(1)} = k_1 < \cdots < k_g = x_{(n)}\} \tag{6}$$

를 생각하자. 단, $x_{(1)} = \min(x_1, \ldots, x_n)$과 $x_{(n)} = \max(x_1, \ldots, x_n)$이다. 집합 $\mathbb{K}$의 원소의 개수를 V이라고 하고 $\mathbb{K} = \{K^{(1)}, \ldots, K^{(V)}\}$로 나타내기로 하자. 하나의 이동모수 $K^{(v)} = (k_1^{(v)}, \ldots, k_g^{(v)})$이 주어져 있다고 하자. 고정된 이동모수 값 $K^{(v)}$에 대하여 Lee와 Oh(2006)은 모수 $\Phi_S = (\pi_1, \ldots, \pi_g;\ \theta_1, \ldots, \theta_g)$에 대한 추정 방법을 제시한다. 즉, 고정된 이동모수 $K^{(v)}$에 대하여 초기값 $\Phi_S^{(p,v)} = (\pi_1^{(p,v)}, \ldots, \pi_g^{(p,v)};\ \theta_1^{(p,v)}, \ldots, \theta_g^{(p,v)})$가 주어져 있다고 하자. 만일 $k_g^{(v)} < x_j$이면,

$$z_{ji}^{(p+1,v)} = \frac{\pi_i^{(p,v)} f_i(x_j; \theta_i^{(p,v)}, k_i^{(v)})}{\sum_{h=1}^{g} \pi_h^{(p,v)} f_h(x_j; \theta_h^{(p,v)}, k_i^{(v)})},\quad i = 1, 2, \ldots, g. \tag{7}$$

만일 $k_1^{(v)} < \cdots < k_{l-1}^{(v)} < x_j < k_l^{(v)} < \cdots < k_g^{(v)}$이면

$$z_{ji}^{(p+1,v)}=\begin{cases}\dfrac{\pi_i^{(p,v)}f_i(x_j;\theta_i^{(p,v)},k_i^{(v)})}{\sum_{h=1}^{l-1}\pi_h^{(p,v)}f_h(x_j,;\theta_h^{(p,v)},k_i^{(v)})}, & i=1,2,\ldots,l-1,\\ 0, & i=l,l+1,\ldots,g.\end{cases} \tag{8}$$

그리고 추정된 완전자료 $(x_1,z_1^{(p+1,v)}),\ldots,(x_n,z_n^{(p+1,v)})$에 대하여 $p+1$ 번째의 반복에서 추정된 모수는 다음과 같이 나타내자.

$$\pi_i^{(p+1,v)}=\frac{\sum_{j=1}^{n}z_{ji}^{(p+1,v)}}{n},\ \ \theta_i^{(p+1,v)}=\frac{\sum_{j=1}^{n}z_{ji}^{(p+1,v)}x_j}{\sum_{j=1}^{n}z_{ji}^{(p+1,v)}}-k_i^{(v)},\ \ i=1,\ldots,g. \tag{9}$$

$\pi_i^{(p,v)}\leftarrow\pi_i^{(p+1,v)}$와 $\theta_i^{(p,v)}\leftarrow\theta_i^{(p+1,v)}$로 두고 (7), (8), (9)를 반복하는 과정을 거쳐 수렴되는 경우 $\Phi_S^{(v)}$라 두고 주어진 이동모수 $K^{(v)}=(k_1^{(v)},\ldots,k_g^{(v)})$에 대한 Φ_S의 추정치라고 하며, $K^{(v)}$에 대응되는 전체 모수 Φ의 추정치를 $\Phi^{(v)}=(\pi_1^{(v)},\ldots,\pi_g^{(v)};\ \theta_1^{(v)},\ldots,\theta_g^{(v)};\ k_1^{(v)},\ldots,k_g^{(v)})$로 나타내기로 하자. 이동모수집합 $\mathbb{K}$에 속한 모든 $K^{(v)}$에 대하여 얻은 $\Phi^{(v)}$에 대하여, 모수 $\Phi=(\pi_1,\ldots,\pi;\theta_1,\ldots,\theta_g;\ k_1,\ldots,k_g)$의 추정치는 다음과 같이 주어진다.

$$\hat{\Phi}=\mathrm{argmax}\left\{L_{\boldsymbol{x}}(\Phi^{(v)})\,|\,v=1,2,\ldots,V\right\}. \tag{10}$$

이동모수를 추정할 때 주어진 자료에 대하여 가능한 모든 이동모수의 조합에 대하여 Φ_S를 추정하는 방법을 사용한다. 식 (5)의 이동모수에 대한 추정식을 사용하지 않는 이유는 다음과 같다. 반복 p번째에서 K의 값이 주어지는 경우 $p+1$번째 반복에서 성분 i에 대한 이동모수 k_i 보다 적은 관측값 x_j 에 대한 성분 추정값이 $z_{ji}^{(p+1,v)}=0$이 되며, 그리고 k_i 보다 크거나 같은 관측값 x_j에 대하여는 $z_{ji}^{(p+1,v)}>0$이 되어 값이 K의 추정값이 변하지 않기 때문이다.

관측값 $\boldsymbol{x}$가 주어져 있을 때, 혼합 이동 포아송분포에 대한 모수 추정절차를 다음과 같이 정리한다.

모수 추정절차 1

초기치: $v\leftarrow 1$로 둔다.

단계 1: 이동위치모수 $K^{(v)}=(k_1^{(v)},\ldots,k_g^{(v)})\in\mathbb{K}$에 대하여

하위단계 1-1: $p\leftarrow 1$로 둔다.

모수 Φ_S의 초기값 $\Phi_S^{(p,v)}=(\pi_1^{(p,v)},\ldots,\pi_g^{(p,v)};\theta_1^{(p,v)},\ldots,\theta_g^{(p,v)})$을 지정한다.

하위단계 1-2: 주어진 모수 $\Phi_S^{(p,v)}$에 대하여 식 (7) 또는 (8)의 방법으로 성분정보의 추정값 $z_1^{(p+1,v)},\ldots,z_n^{(p+1,v)}$을 구한다. 추정된 완전자료 $(x_1, z_1^{(p+1,v)}),\ \ldots,(x_n,z_n^{(p+1,v)})$와 식 (9)의 방법을 이용하여 추정치 $\Phi_S^{(p+1,v)}$ 를

얻는다.
하위단계 1-3: 수렴 조건을 만족하지 않으면, $p \leftarrow p+1$ 그리고 $\Phi_S^{(p,v)} \leftarrow \Phi_S^{(p+1,v)}$로 두고 하위단계 1-2로 간다. 수렴 조건을 만족하면 모수의 추정치를 $\hat{\Phi}^{(v)}$라고 표시한다. 만약 $v < V$이면 $v \leftarrow v+1$로 두고 단계 1로 가고, 아니면 단계 2로 간다.
단계 2: 모수의 추정치는 식 (10)으로 주어진다.

모수 추정절차 1의 하위단계 1-3에서 수렴 조건은 $\left|L_{\boldsymbol{x}}(\Phi^{(p+1)})-L_{\boldsymbol{x}}(\Phi^{(p)}\right|<\delta$ 인 경우, 즉, 식 (3)의 로그우도값의 증가치가 일정 범위 내에 드는 것으로 정하였다. 단, $\delta > 0$는 수렴한계이다. 모수의 초기값을 정하는 하나의 방법은 Lee와 Oh(2006)에 제시되어 있다.

3. 모의실험

모수 추정절차 1의 추정법의 효율성을 알아보기 위해 몬테카를로 모의실험을 실시하였다. 고려한 혼합이동포아송 모형은 성분의 개수가 g=2와 3인 경우이며, 각 시뮬레이션에서 표본의 크기는 $n = g \times 50$, 반복회수는 3000번으로 하였다. 각 표의 값은 3000번 반복에 대한 추정치의 표본평균과 표준오차이다. 표준오차는 괄호 안에 표시하였다. 각 표에서 i 행은 불완전 자료에 대한 추정치, 행 c는 완전자료에 대한 최우추정치에 관한 것이다. 매 반복에서 모수에 대한 n개의 혼합이동포아송 표본을 생성한다. 이 때 성분정보를 자료에 포함시킨다. 이 표본에 대하여 불완전 자료에 대한 모수 추정 절차 1의 방법과 완전 자료에 대한 최우추정법으로 추정치를 구한다.

표 1은 g=2, $(\theta_1, \theta_2) = (1,1)$일 때, (k_1, k_2)=(0, 3), (0, 4), (0, 5), 그리고 (π_1, π_2)=(.7, .3), (.6, .4), (.5, .5)인 경우에 대한 모의실험 결과이다. 모든 모수 조합의 경우에서 이동모수 k_1은 참값 0으로 추정되었으며 표준오차는 0이다. 이는 모든 경우에서 관측된 자료의 최소값이 0이고 따라서 k_1이 정확하게 추정되었음을 의미한다. 한편 각 (π_1, π_2)의 값에 대하여 k_2의 참값이 3, 4, 5로 커짐에 따라서, k_2, π_1, π_2, θ_1, θ_2의 추정값의 평균이 참값에 가까와지며 표준오차가 작아지는 모습을 보이고 있다. 이동모수의 차이가 큰 것은 두 성분 함수의 구분이 더 잘 되는 경우이므로 이와 같은 시뮬레이션 결과는 당연한 것이라고 할 수 있다. 그리고 각 (k_1, k_2)의 값에 대하여 π_1의 참값이 .7, .6, .5로 1/2에 가까워 짐에 따라 k_1, k_2, π_1, π_2, θ_1, θ_2의 추정값의 평균이 참값에 가까와지며 표준오차가 작아지는 모습을 보이고 있다. 예컨대 (k_1, k_2)=(0, 3)인 경우에 π_1 = .7, .6, .5의 순으로 π_1의 추정값의 평균은 각각 .682, .588, .494로 참값에 가까우며, 표준오차는 .085, .076, .066으로 작아진다. 이는 두 성분 비율이 같은 경우에는 각 성분의 자료의 수가 비슷하게 관측되므로 다른 경우에 비해 더 좋은 결과를 가져 올 것이라는 직관과도 일치한다. 또한 (π_1, π_2)의 각 경우에서 이동모수의 차이가 커질수록 불완전자료에 의한 평균과 표준오차는 완전자료에 의한 평균값과 표준오차에 가까워져 가는 것을 볼 수 있다. 한편 전체적으로 k_2, π_1, θ_1은 저추정되며, π_2와 θ_2는 과추정되는 경향을 나타내고 있는 데 이는 k_2의 저추정에 기인하는 것으로 짐작된다. 이러한 경향은 g=2, $(\theta_1, \theta_2) = (1,2)$인 표 2의 경우에서도 마찬가지로 나타남을 관측할 수 있다. 차이점으로는 표 2에서 θ_2=2이며 표 1의 θ_2=1

보다 큰 관계로 표준오차가 표 1의 결과와 비해서 더 큰 모습을 나타내고 있다.

<표 1> 모의실험 결과. g=2, (θ_1, θ_2)=(1, 1)인 경우의 추정치의 표본평균(표준오차). 자료형의 열에서 행 i와 행 c는 각각 불완전자료와 완전자료에 대한 것임.

(π_1, π_2)	(k_1, k_2)	자료형	$\hat{k}_1$	$\hat{k}_2$	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\theta}_1$	$\hat{\theta}_2$
.7, .3	0, 3	i	0.000	2.863	0.682	0.318	0.955	1.072
			(0.000)	(0.609)	(0.085)	(0.085)	(0.225)	(0.403)
		c	0.000	3.000	0.700	0.300	0.997	1.001
			(0.000)	(0.000)	(0.045)	(0.045)	(0.120)	(0.182)
	0, 4	i	0.000	3.898	0.694	0.306	0.987	1.074
			(0.000)	(0.379)	(0.051)	(0.051)	(0.147)	(0.330)
		c	0.000	4.000	0.699	0.301	1.004	1.005
			(0.000)	(0.000)	(0.046)	(0.046)	(0.121)	(0.186)
	0, 5	i	0.000	4.951	0.700	0.300	0.997	1.040
			(0.000)	(0.249)	(0.046)	(0.046)	(0.125)	(0.281)
		c	0.000	5.000	0.701	0.299	1.003	1.001
			(0.000)	(0.000)	(0.046)	(0.046)	(0.118)	(0.182)
.6, .4	0, 3	i	0.000	2.902	0.588	0.412	0.971	1.062
			(0.000)	(0.420)	(0.076)	(0.076)	(0.223)	(0.314)
		c	0.000	3.000	0.599	0.401	1.000	1.002
			(0.000)	(0.000)	(0.050)	(0.050)	(0.131)	(0.157)
	0, 4	i	0.000	3.942	0.597	0.403	0.993	1.046
			(0.000)	(0.261)	(0.051)	(0.051)	(0.154)	(0.266)
		c	0.000	4.000	0.600	0.400	1.002	1.002
			(0.000)	(0.000)	(0.048)	(0.048)	(0.129)	(0.162)
	0, 5	i	0.000	4.981	0.600	0.400	0.999	1.014
			(0.000)	(0.140)	(0.050)	(0.050)	(0.136)	(0.204)
		c	0.000	5.000	0.600	0.400	1.001	0.998
			(0.000)	(0.000)	(0.050)	(0.050)	(0.130)	(0.156)
			(0.000)	(0.000)	(0.049)	(0.049)	(0.143)	(0.141)
.5, .5	0, 3	i	0.000	2.940	0.494	0.506	0.981	1.041
			(0.000)	(0.276)	(0.066)	(0.066)	(0.228)	(0.247)
		c	0.000	3.000	0.499	0.501	0.999	0.999
			(0.000)	(0.000)	(0.049)	(0.049)	(0.141)	(0.143)
	0, 4	i	0.000	3.977	0.499	0.501	0.998	1.017
			(0.000)	(0.164)	(0.052)	(0.052)	(0.165)	(0.200)
		c	0.000	4.000	0.499	0.501	1.000	0.998
			(0.000)	(0.000)	(0.050)	(0.050)	(0.140)	(0.143)
	0, 5	i	0.000	4.995	0.499	0.501	1.001	1.004
			(0.000)	(0.075)	(0.050)	(0.050)	(0.150)	(0.157)
		c	0.000	5.000	0.499	0.501	1.002	1.000
			(0.000)	(0.000)	(0.050)	(0.050)	(0.144)	(0.142)

<표 2> 모의실험 결과. g=2, (θ_1, θ_2)=(1, 2)인 경우의 표본평균(표준오차). 자료형의 열에서 행 i와 행 c는 각각 불완전자료와 완전자료에 대한 것임.

(π_1, π_2)	(k_1, k_2)	자료형	$\hat{k}_1$	$\hat{k}_2$	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\theta}_1$	$\hat{\theta}_2$
.7, .3	0, 3	i	0.000	2.866	0.685	0.315	0.972	2.053
			(0.000)	(1.024)	(0.074)	(0.074)	(0.200)	(0.768)
		c	0.000	3.012	0.698	0.302	1.000	1.980
			(0.000)	(0.110)	(0.046)	(0.046)	(0.120)	(0.275)
	0, 4	i	0.000	3.778	0.693	0.307	0.983	2.158
			(0.000)	(0.861)	(0.053)	(0.053)	(0.151)	(0.737)
		c	0.000	4.013	0.700	0.300	1.004	1.985
			(0.000)	(0.113)	(0.046)	(0.046)	(0.119)	(0.277)
	0, 5	i	0.000	4.843	0.697	0.303	0.984	2.123
			(0.000)	(0.580)	(0.048)	(0.048)	(0.126)	(0.576)
		c	0.000	5.014	0.700	0.300	0.995	1.980
			(0.000)	(0.116)	(0.046)	(0.046)	(0.117)	(0.273)
.6, .4	0, 3	i	0.000	2.853	0.587	0.413	0.972	2.090
			(0.000)	(0.886)	(0.075)	(0.075)	(0.227)	(0.674)
		c	0.000	3.004	0.598	0.402	1.000	1.994
			(0.000)	(0.060)	(0.049)	(0.049)	(0.130)	(0.227)
	0, 4	i	0.000	3.827	0.596	0.404	0.979	2.134
			(0.000)	(0.660)	(0.053)	(0.053)	(0.162)	(0.591)
		c	0.000	4.005	0.601	0.399	0.998	1.995
			(0.000)	(0.068)	(0.048)	(0.048)	(0.131)	(0.234)
	0, 5	i	0.000	4.894	0.598	0.402	0.994	2.089
			(0.000)	(0.411)	(0.049)	(0.049)	(0.136)	(0.436)
		c	0.000	5.005	0.600	0.400	1.003	1.993
			(0.000)	(0.068)	(0.048)	(0.048)	(0.127)	(0.235)
.5, .5	0, 3	i	0.000	2.841	0.488	0.512	0.971	2.115
			(0.000)	(0.761)	(0.073)	(0.073)	(0.253)	(0.595)
		c	0.000	3.001	0.498	0.502	1.001	2.002
			(0.000)	(0.032)	(0.048)	(0.048)	(0.141)	(0.202)
	0, 4	i	0.000	3.854	0.495	0.505	0.980	2.116
			(0.000)	(0.499)	(0.053)	(0.053)	(0.176)	(0.470)
		c	0.000	4.001	0.500	0.500	1.002	1.997
			(0.000)	(0.032)	(0.050)	(0.050)	(0.144)	(0.199)
	0, 5	i	0.000	4.917	0.499	0.501	0.995	2.073
			(0.000)	(0.317)	(0.049)	(0.049)	(0.151)	(0.360)
		c	0.000	5.001	0.500	0.500	1.003	1.998
			(0.000)	(0.026)	(0.048)	(0.048)	(0.143)	(0.202)

<표 3> 모의실험 결과. g=3, (π_1, π_2, π_3)=(.33, .33, .34), $(\theta_1, \theta_2, \theta_3)$=(1, 1, 1)인 경우 의 추정치의 표본평균(표준오차). 자료형의 열에서 행 i와 행 c는 각 각 불완전자료와 완전자료에 대한 것임.

(k_1, k_2, k_3)		0	3	6	.33	.33	.34	1.00	1.00	1.00
0, 3, 6	i	.00	2.86	5.96	.32	.34	.34	.96	1.12	1.02
		(.00)	(.43)	(.32)	(.05)	(.07)	(.06)	(.24)	(.52)	(.26)
	c	.00	3.00	6.00	.33	.33	.34	1.00	1.00	1.00
		(.00)	(.00)	(.00)	(.04)	(.04)	(.04)	(.14)	(.14)	(.14)
		0	4	8	.33	.33	.34	1.00	1.00	1.00
0, 4 ,8	i	.00	3.96	7.98	.33	.33	.34	1.00	1.04	1.02
		(.00)	(.23)	(.16)	(.04)	(.04)	(.04)	(.17)	(.30)	(.20)
	c	.00	4.00	8.00	.33	.33	.34	1.00	1.00	1.00
		(.00)	(.00)	(.00)	(.04)	(.04)	(.04)	(.14)	(.15)	(.14)
		0	5	10	.33	.33	.34	1.00	1.00	1.00
0, 5, 10	i	.00	4.99	10.0	.33	.33	.34	1.00	1.00	1.00
		(.00)	(.08)	(.07)	(.04)	(.04)	(.04)	(.15)	(.17)	(.16)
	c	.00	5.00	1.00	.33	.33	.34	1.00	1.00	1.00
		(.00)	(.00)	(.00)	(.04)	(.04)	(.04)	(.14)	(.14)	(.14)

마지막으로 표 3은 성분의 개수 g=3, 성분 비율 모수 (π_1, π_2, π_3)=(.33, .33, .34), 성분의 분산 모수 $(\Theta_1, \Theta_2, \Theta_3)$=(1, 1, 1)인 경우에 대한 시뮬레이션 결과이다. 모의실험에서는 이동모수 간의 거리가 3배의 표준편차 즉, $3\sqrt{\Theta_1}$ 이상인 경우를 고려하였다. 표 3에서 불완전자료에 대한 추정치의 평균과 표준오차는 완전자료에 대한 그것과 근접한 모습을 나타내고 있다.

4. 토의

Liu 외(2006)은 혼합포아송분포의 모수 추정에서 EM 알고리즘을 적용하는 것을 소개하였다. 그러나, 이동포아송분포의 혼합분포에서 관측된 자료로부터 모수를 추정하는 문제는 이동모수의 특성으로 일반적인 EM 알고리즘의 단순한 적용을 할 수 없다. 따라서 본 연구에서는 주어진 자료의 범위 내에서 모든 가능한 이동모수의 조합을 생각하고 이들 각 이동모수의 값에 대하여 Lee와 Oh(2006)의 방법으로 다른 모수들을 추정한 후, 이들 이동모수와 추정된 모수들 중에서 불완전자료에 대한 우도함수를 최저로 하는 것을 추정치로 선택하는 방법을 제시하였다. 시뮬레이션 실험으로 제시된 방법의 성질을 조사와 완전자료에 대한 추정치와의 비교를 하였다. 실험 결과는 제시된 방법이 '합리적'임을 나타내 보이고 있다. 더욱이 이동모수의 차이가 큰 경우 불완전 자료에 대한 추정치가 완전자료에 대한 최우추정치와 근접함을 보인다. 한편, 제시된 방법에 대한 대표본 성질은 추후 연구되어야 할 과제이다. 혼합이동포아송분포에서 이동모수가 모두 0인 경우가 혼합포아송분포이므로 여기서 제시된 방법은 혼합포아송분포에서의 Liu 외(2006)의 모수추정방법을 포함하는 방법이다.

참고문헌

1. Copsey, K. and Webb, A. (2003). Bayesian gamma mixture model approach to radar target recognition. IEEE Transactions on Aerospace and Electronic Systems, Vol. 39, 1201-1217.
2. Dempster, A. P., Laird, N .M. and Rubin, D. R. (1977). Maximum likelihood from incomplete data. Journal of the Royal Statistical Society, Series B, Vol. 39, 1-38.
3. Lee, H. J. and Oh, C. (2006). Estimation in mixture of shifted Poisson distributions with Known shift parameters. Journal of the Korean Data & Information Science Society, Vol. 17, No. 3, 785-794.
4. Liu, Z., Almhana, J., Choulakian, V. and McGorman, R. (2006). Online EM algorithm for mixture with application to internet traffic modeling. Computational Statistics & Data Analysis, Vol. 50, 1052-1071.
5. McLachlan, G. J. and Peel, D. (2001). Finite Mixture Models. John Wiley & Sons, Inc.

[2006년 10월 접수, 2006년 11월 채택]