Journal of the Korean
Data & Information Science Society
2006, Vol. 17, No. 3, pp. 785~794

Estimation in Mixture of Shifted Poisson Distributions with Known Shift Parameters

Hyun Jung Lee[1] · Changhyuck Oh[2]

Abstract

Suggested is an EM algorithm for estimation in mixture of shifted Poisson distributions with known shift parameters. For this type of mixture distribution, we have to utilize values of shift parameters to determine whether each of data belongs to some component distribution. We propose a method of estimating values of component information and then follow typical EM methodology. Simulation results show that the algorithm provides reasonable performance for the distribution.

Keywords : 혼합 이동 포아송분포, EM 알고리즘

1. 머리말

결측치를 가지는 불완전한 자료에 대하여 최우추정치를 구하는 방법으로 제시된 EM 알고리즘은 반복적인 수치해법이다. Dempster 외(1977)에 의해 정리 및 소개되어진 이 알고리즘은, 기본적으로 자료와 모형으로부터 결측치를 추정하여 완전 자료를 만들고, 이 추정된 완전 자료에 관한 최우추정치를 구하는 일을 반복한다. EM 알고리즘은 결측치를 가지는 다변량자료의 분석, 혼합분포에서 얻은 자료로부터의 모수추정 등의 다양한 분야에 적용되고 있다. 혼합정규분포를 비롯한 여러 가지 혼합분포에 관한 EM 알고리즘에 대하여는 McLachlan과 Peel(2001)을 참조하기 바란다. 혼합 정규분포 이외에도 혼합 감마분포(Copsey와 Webb, 2003)나 혼합 포아송분포와 같은 다양한 분포에 대한 혼합분포에 대한 연구가 진행되고 있다.

Hasselbald(1969)는 2개의 성분을 가진 혼합 포아송분포를 사용하여 신문에 발표된 사망자 수의 분포를 분석하였다. Church와 Gale(1995)는 여러 개의 성분을 가지는 혼합포아송 분포를 이용하여 문서에 포함된 단어율에 관한 분석을 하였다. Liu 외(2006)은 혼합 포아송분포에 대해 EM 알고리즘과 Online EM 알고리즘을 소개하고 이를 이

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

용하여 인터넷 트래픽 자료를 모형화하였다. 김인영 외(2006)은 대한민국의 개인들의 기부 행태에 관한 조사 연구에서 혼합 포아송분포를 적용하였다. Schlatman(2005)는 혼합 포아송분포에서 성분의 개수를 구하기 위한 붓스트랩 방법을 제시하였다. 한편 Lia와 Zha(2006)은 이중 혼합 포아송분포를 이용하여 문서와 단어 군집분석을 하였다. 또한, Yavuz 외(1998)은 positron emission tomography 즉, PET 스캔자료로부터 영상복원을 하는 방법에서 이동 포아송분포를 적용하였다.

모수가 θ인 포아송분포는 평균이 θ, 표준편차가 $\sqrt{\theta}$이므로, 평균이 크면 표준편차도 큰 특성을 가지고 있다. 따라서 혼합 포아송분포로 계수 자료에 대한 히스토그램을 적합시킬 때, 0에서부터 멀리 떨어져 있는 봉우리는, 0에 가까이 있는 봉우리에 비해 펑퍼짐한 모양을 하고 있어야 한다. 이러한 모양의 히스토그램으로 간주되는 계수 자료로는 자동차 사고 자료, 기부에 관한 자료 등이 있다. 한편 Liu 외(2006)가 적용한 네트워킹 트래픽 데이터는 혼합 포아송분포에 대한 히스토그램의 모양 보다는 혼합 이동 포아송 분포가 더 적절해 보이는 형태를 취하고 있다. 따라서 본 논문에서는 혼합 이동 포아송분포에 대한 모수 추정의 방법을 연구한다. 제시되는 추정방법은 전통적인 EM 알고리즘 방식을 따르는 것이며 이동 포아송분포의 특성에 맞추어 알고리즘을 개발하였다. 제 2절에서는 혼합 이동 포아송분포에서의 모수 추정 절차를 소개한다. 본 논문에서는 방법론적 접근을 하므로, 3절에서는 제시된 모수 추정 절차에 대하여 모의실험으로 추정법의 효율성을 살펴본다. 마지막 4절에서는 토의와 결론을 내린다.

2. 혼합 이동 포아송분포의 모수 추정

양의 정수 g개의 성분을 가지는 혼합 이동 포아송분포 모형

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

를 생각하자. 단, π_i는 각 성분 함수에 대응하는 미지의 혼합 가중치이며 $0<\pi_i<1$, $(i=1,\ldots,g)$, $\pi_1+\cdots+\pi_g=1$이다. 그리고 성분확률함수 f_i는 일변량 이동 포아송분포의 확률함수이라고 가정한다. 즉,

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

여기서 $\theta_i>0$는 미지의 모수, 이동모수 k_i는 기지의 정수인 상수 값으로 가정한다. 또한 Φ는 전체 모수 벡터를 나타낸다. 즉, $\Phi=(\pi_1,\ldots,\pi_g;\theta_1,\ldots,\theta_g)$이다.

혼합 이동 포아송분포 (1)로 부터 관측 자료 $\boldsymbol{x}=(x_1,\ldots,x_n)$이 주어졌다고 하자. 자료에 결측치가 있다고 가정하는 상황에서는 $f(x;\Phi)$를 불완전자료의 확률함수라고 부르며, 주어진 불완전자료에 대한 모수 Φ의 로그우도함수는

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

로 주어진다. 최우추정법은 주어진 모수 공간에 대하여, $L_{\boldsymbol{x}}(\Phi)$를 최대로 하는 $\hat{\Phi}_{\boldsymbol{x}}$를 찾는다. 그러나 혼합모형에 대하여 이와 같은 $\hat{\Phi}_{\boldsymbol{x}}$를 구하는 것은 어려운 문제로 알려져 있다. 따라서 자료 $\boldsymbol{x}$를, 관측되지 못한 완전자료의 일부분인 불완전자료로 간주하여 문제를 해결하고자 한다. 여기서 관측된 자료는, 각 관측값 x_j, $j=1,\ldots,n$, 을 발생시킨 성분에 관한 정보가 결측된 불완전 자료로 간주할 수 있다. 관측값 x_j의 발생 성분 정보를 나타내기 위하여 벡터 $z_j=(z_{j1},z_{j2},\ldots,z_{jg})$를 생각한다. 벡터 z_j의 성분 $z_{j1},z_{j2},\ldots,z_{jg}$은 관측값을 발생시킨 성분에 대하여서만 1의 값을, 나머지는 모두 0의 값을 취한다. 성분정보를 포함한 자료 $(x_1,z_1),\ldots,(x_n,z_n)$을 완전자료라고 한다. 완전자료에 대한 모수의 최우추정치는 식 (4)와 같이 주어짐을 보일 수 있다.

$$\hat{\pi}_i=\frac{\sum_{j=1}^{n}z_{ji}}{n},\qquad \widehat{\Theta}_i=\frac{\sum_{j=1}^{n}z_{ji}x_j}{\sum_{j=1}^{n}z_{ji}}-k_i,\qquad i=1,2,\ldots,g \tag{4}$$

고려하는 경우에서는 성분 정보에 관한 값이 관측되지 않았으므로, EM 알고리즘의 E 단계에서는 불완전 자료 $x_1,\ldots,x_n$와 사전에 정해진 확률모형을 이용하여 성분정보 $z_{j1},z_{j2},\ldots,z_{jg}$를 추정하게 된다. 추정된 완전자료를 $(x_1,z_1^{(p)}),\ldots,(x_n,z_n^{(p)})$이라고 할 때, 혼합 이동 포아송분포에서의 모수 Φ는 추정된 완전자료를 식 (4)에 대입하여 추정되게 된다.

불완전 자료로부터 성분정보를 추정하기 위하여 EM 알고리즘에서는 Q 함수를 이용한다. Dempster 외(1977)를 참조하기 바란다. 함수 Q는 기본적으로 불완전자료가 주어졌다는 조건 하에서 관측되지 않은 값에 대한 기대값을 구함으로써 성분정보의 값을 추정하게 된다. 완전자료 $(x_1,z_1),\ldots,(x_n,z_n)$에 대한 로그우도함수 $LAPLACE(\Phi)$에 대하여 Q 함수는

$$\begin{aligned}Q(\Phi|\Phi^{(p)})&=E[LAPLACE(\Phi)|\,\boldsymbol{x},\Phi^{(p)}]\\&=\sum_{j=1}^{g}\sum_{j=1}^{n}E[z_{ji}|\,\boldsymbol{x},\Phi^{(p)}]\log\pi_i+\sum_{j=1}^{n}E[\log f_{z_j'q}(x_j;\Theta_{z_j'q})|\,\boldsymbol{x},\Phi^{(p)}]\end{aligned} \tag{5}$$

로 주어진다. 단 $q'=(1,2,\ldots,n)$이며, $z_j'q$는 주어진 관측치 x_j를 발생시킨 성분함수를 나타내는 지표값을 나타내게 된다. 일반성을 잃지 않고, 기지의 이동 모수 k_i에 대하여 $k_1<k_2<\cdots<k_g$라고 가정하고, 모수의 초기값 $\Phi^{(p)}=(\pi_1^{(p)},\ldots,\pi_g^{(p)};\ \Theta_1^{(p)},\ldots,\Theta_g^{(p)})$가 주어져 있다고 하자. 혼합 이동 포아송분포에서 성분 벡터 z_j, $j=1,\ldots,n$에 대하여 다음의 (6) 또는 (7)의 추정치를 함수 Q로부터 유도해 낼 수 있다. 만일 $k_g\le x_j$이면,

$$z_{ji}^{(p+1)}=\frac{\pi_i^{(p)} f_i(x_j;\theta_i^{(p)})}{\sum_{h=1}^{g}\pi_h^{(p)} f_h(x_j;\theta_h^{(p)})},\quad i=1,2,\ldots,g, \tag{6}$$

만일 $k_1<\cdots<k_{l-1}\le x_j<k_l<\cdots<k_g$이면

$$z_{ji}^{(p+1)}=\begin{cases}\dfrac{\pi_i^{(p)} f_i(x_j;\theta_i^{(p)})}{\sum_{h=1}^{l-1}\pi_h^{(p)} f_h(x_j;\theta_h^{(p)})}, & i=1,2,\ldots,l-1,\\ 0, & i=l,\ldots,g.\end{cases} \tag{7}$$

식 (7)에서 $k_1<\cdots<k_{l-1}\le x_j<k_l<\cdots<k_g$일 때, $z_{jl}^{(p+1)}=\cdots=z_{jg}^{(p+1)}=0$으로 추정되어야 하는 이유는 각 성분확률함수 $f_l,\ldots,f_g$에 대한 영역의 최소값이 관측값 x_j보다 크기 때문이다. 관측치 x_j에 대한 성분정보자료 $(z_{j1},\ldots,z_{jg})$의 각 성분은 0 또는 1로 주어지나, 추정된 성분자료 $(z_{j1}^{(p+1)},\ldots,z_{jg}^{(p+1)})$의 각 성분은 0과 1 사이의 실수값으로 주어진다. 즉, $0\le z_{ji}\le 1, i=1,\ldots,g$, 그리고 $z_{j1}^{(p+1)}+\cdots+z_{jg}^{(p+1)}=1$이다.

추정된 완전자료 $(x_1,z_1^{(p+1)}),\ldots,(x_n,z_n^{(p+1)})$에 대하여 모수 $\Phi=(\pi_1,\ldots,\pi_g;\theta_1,\ldots,\theta_g)$의 추정치를 다음과 같이 주어진다.

$$\pi_{i}^{(p+1)}=\frac{\sum_{j=1}^{n} z_{ji}^{(p+1)}}{n},\qquad \theta_i^{(p+1)}=\frac{\sum_{j=1}^{n} z_{ji}^{(p+1)} x_j}{\sum_{j=1}^{n} z_{ji}^{(p+1)}}-k_i,\quad i=1,2,\ldots,g \tag{8}$$

이동모수가 알려져 있는 혼합 이동 포아송분포에 대하여, 관측값 $x_1,\ldots,x_n$가 주어져 있을 때, 모수 추정절차를 다음과 같이 나타낸다.

모수 추정절차 1

단계 1: 모수 Φ의 초기값 $\Phi^{(p)}=(\pi_1^{(p)},\ldots,\pi_g^{(p)};\theta_1^{(p)},\ldots,\theta_g^{(p)})$을 지정한다.

단계 2: 식 (6) 또는 (7)의 방법으로 성분의 추정값 $z_1^{(p+1)},\ldots,z_n^{(p+1)}$을 구한다.

단계 3: 추정된 완전자료 $(x_1,z_1^{(p+1)}),\ldots,(x_n,z_n^{(p+1)})$를 이용하여 식 (8)의 방법으로 모수의 추정치 $\Phi^{(p+1)}$를 얻는다.

단계 4: 수렴 조건을 만족하지 않으면, $\Phi^{(p)}\leftarrow\Phi^{(p+1)}$로 두고 단계 2로 간다.

<표 1> 자료의 도수분포표

	값	0	1	2	3	4	5	6	7	8	9	10	합계
도수	그룹1	8	9	9	7	7	1						41
	그룹2					8	9	15	15	4	7	1	59
	합동	8	9	9	7	15	10	15	15	4	7	1	100

모수 추정절차 1의 단계 4에서 수렴 조건을 위하여 불완전자료에 대한 로그우도함수을 사용하였다. 즉, 추정된 두 모수 집합에 대하여, 식 (3)의 로그우도값의 증가치가 일정 범위 내에 드는 $|L_{\boldsymbol{x}}(\Phi^{(p+1)})-L_{\boldsymbol{x}}(\Phi^{(p)})|<\delta$ 인 경우를 수렴의 조건으로 정하였다. 단, $\delta>0$는 적당한 값의 수렴한계이다.

일반적으로 반복법에서 모수의 초기값을 정하는 일은 중요한 문제이다. 여기서는 모수의 초기값을 정할 때 주어진 자료에 근거하는 방법을 사용한다. 인접한 두 이동모수 k_i와 k_{i+1} 사이에 있는 관측값의 개수의 비율과 평균을 혼합가중치 π_i와 Θ_i의 초기치로 정한다. 각 $i=1,\ldots,g$에 대하여 $A_i=\{x_j: k_i \le x_j < k_{i+1}, j=1,\ldots,n\}$, 단, $k_{g+1}=\infty$, 그리고 c_i는 집합 A_i에 속하는 원소의 개수라고 하자. 모수 Φ의 초기치는 다음과 같이 설정한다.

$$\pi_i^{(0)}=c_i/n, \qquad \Theta_i^{(0)}=(1/c_i)\sum_{x_j\in A_i} x_j-k_i, \qquad i=1,2,\ldots,g \tag{9}$$

예제 1. 제시된 모수 추정절차 1과 초기값 설정에 대한 수치적 예를 든다. 표 1의 자료는 생성된 혼합 이동 포아송 분포수에 대한 도수분포이다. 이들 자료를 생성하는데 사용한 모수는 다음과 같다. k_1=0, k_2=4, π_1=0.4, π_2=0.6, $\Theta_1=\Theta_2$=2이며 모두 100개의 수를 생성하였다.

<표 2> 모수 추정절차 1에 대한 반복

반복	$\pi_1^{(p)}$	$\pi_2^{(p)}$	$\Theta_1^{(p)}$	$\Theta_2^{(p)}$	$-\log L(\Phi^{(p)})$
0	0.33017	0.66983	1.45234	2.11782	10^{99*}
1	0.35614	0.64386	1.65529	2.19774	233.7143832
2	0.37101	0.62899	1.76213	2.24215	233.3894853
3	0.38050	0.61950	1.82766	2.27050	233.2633684
4	0.38687	0.61313	1.87065	2.28954	233.2080266
5	0.39127	0.60873	1.89992	2.30270	233.1820657
6	0.39437	0.60563	1.92029	2.31194	233.1693855
7	0.39657	0.60343	1.93468	2.31852	233.1630289
…	…	…	…	…	…
31	0.40228	0.59772	1.97161	2.33554	233.1562762
32	0.40228	0.59772	1.97161	2.33554	233.1562762
cMLE**	0.41	0.59	1.97561	2.38981	

* 주어진 초기치, ** 완전 자료에 대한 최우추정치

표 1은 각 그룹과 두 그룹의 자료에서 발생성분을 없앤 경우에 대한 도수분포이다. 표 1의 완전자료를 이용한 그룹 1과 2에 대하여 혼합가중치의 추정값은 각각 $\hat{\pi}_1$=0.41, $\hat{\pi}_2$=0.59이며, 평균의 최우추정값은 각각 $\hat{\Theta}_1$=1.97561과 $\hat{\Theta}_2$=2.38981이다. 한편 표 2는 표 1에서 행 합동의 불완전 자료와 모수 추정절차 1과 초기값 설정 방법 (9)에 대한 반복과정이다. 반복 0에서의 모수 추정값은 식 (9)에 의해 설정된 초기값

이며 그때의 로그우도값은 초기치로 지정된 아주 적은 값으로 하였다. 그리고 반복 1 이후는 추정절차 1에 의한 모수의 추정값과 이에 대한 로그우도함수이다. 반복절차는 반복 32에서 종료되며 불완전자료에 대하여 혼합가중치의 추정치는 $\tilde{\pi}_1$=0.40228, $\tilde{\pi}_2$ =0.59772이며, 평균의 추정값은 각각 $\tilde{\Theta}_1$=1.97161과 $\tilde{\Theta}_2$=2.33554이다. 이들 값과 완전자료에 의한 추정값과의 오차는 $|\hat{\pi}_i-\tilde{\pi}_i|<0.01$, $|\hat{\Theta}_i-\tilde{\Theta}_i|<0.05$, $i=1,2$이다.

3. 모의실험

혼합 이동 포아송분포에 대하여 제시한 모수 추정절차 1의 추정법의 효율성을 알아보기 위해 몬테카를로 모의실험을 한다. 시뮬레이션은 성분의 개수가 g=2와 3인 경우에 대하여 시행하였다. 각 시뮬레이션에서 표본의 크기는 $n=g\times 50$, 반복회수는 3000번으로 하였다. 모수추정 절차 1의 반복을 종료하기 위한 수렴의 한계는 δ =0.000000001로 하였으며, 최대 가능 반복의 횟수는 100으로 하였다. 각 표의 값은 추정치의 표본평균이며 괄호 안의 값은 표준오차이다. 이동모수 k_i의 값은 알려져 있는 경우이고, 자료에 의해 추정하는 모수는 혼합가중치 π_i와 성분 포아송 분포의 모수 Θ_i이다. 각 표에서 cMLE 행은 완전자료에 대한 최우추정치의 표본평균과 표준오차이다.

표 3은 g=2 인 경우에 대한 모의실험 결과이다. 각 (π_1,π_2)의 값에 대하여 k_1=0으로 고정된 경우에, k_2의 값이 2, 3, 4, 5로 증가하면 π_i와 Θ_i에 대한 표본평균이 대응되는 모수의 참값에 가까워 짐을 볼 수 있다. 이는 이동모수의 값이 충분히 떨어져 있으면 각 자료에 대한 성분 정보가 없더라도 잘 구별될 수 있는 상황을 반영한다. 한편 k_2=5인 경우는, 성분에 관한 정보를 가지고 있는 완전자료에 대한 최우추정치에 대한 표본평균과 표준오차가 같아짐을 볼 수 있다. 이러한 상황은 혼합가중치가 (π_1,π_2)=(.6, .4)나 (.5, .5)인 경우에도 동일함을 볼 수 있다. 이동모수 k_1=0이고 Θ_1 =1인 포아송분포에서는 평균과 표준편차가 모두 1로 같으므로 두 번째 성분의 이동모수 k_2=3일 때, 즉 표준편자의 3배인 경우에 모수들이 잘 추정되고 있다고 볼 수 있다.

표 4는 g=2, (Θ_1,Θ_2)=(2, 1), (π_1,π_2)=(.7, .3), (.6, .4), (.5, .5)인 경우이다. 여기서는 첫 번째 성분의 포아송분포의 모수 Θ_1=2이므로 평균은 Θ_1=2, 표준편차는 $\sqrt{\Theta_1}$ $\fallingdotseq$1.4가 된다. 따라서 $k_2\fallingdotseq 2\times 1.4=2.8$인 경우에 추정치의 표본평균이 모수의 참값에 가깝게 됨을 볼 수 있다. 여기서도 $k_2=7=\sqrt{\Theta_1}\times 5$인 경우에는 완전자료에 대한 cMLE 행의 값과 거의 같음을 볼 수 있다.

한편 표 5는 g=3, (π_1,π_2,π_3)=(.33, .33, .34), $(\Theta_1,\Theta_2,\Theta_3)$=(1, 1, 1)인 경우이다. 고려한 경우는 이동모수간의 거리가 3배의 표준편차 $3\sqrt{\Theta_1}$ 이상인 경우이다. 이동모수 k_2=5인 경우는 완전자료에 대한 표본평균과 표준오차에 근접함을 볼 수 있다.

4. 결론

<표 3> 모의실험 결과. g=2, (Θ_1,Θ_2)=(1, 1)인 경우의 표본평균(표준오차)

(π_1,π_2)	(k_1,k_2)	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\Theta}_1$	$\hat{\Theta}_2$
(.7, .3)	(0,2)	0.714 (0.108)	0.286 (0.108)	1.018 (0.255)	1.048 (0.435)
	(0,3)	0.702 (0.057)	0.298 (0.057)	1.004 (0.167)	1.009 (0.211)
	(0,4)	0.701 (0.048)	0.299 (0.048)	1.006 (0.138)	1.000 (0.193)
	(0,5)	0.701 (0.045)	0.299 (0.045)	0.999 (0.125)	1.001 (0.184)
	cMLE*	0.700 (0.046)	0.300 (0.046)	1.001 (0.121)	1.002 (0.187)
(.6, .4)	(0,2)	0.616 (0.114)	0.384 (0.114)	1.022 (0.291)	1.002 (0.287)
	(0,3)	0.604 (0.059)	0.396 (0.059)	1.014 (0.185)	1.006 (0.174)
	(0,4)	0.599 (0.050)	0.401 (0.050)	1.004 (0.146)	1.004 (0.164)
	(0,5)	0.601 (0.048)	0.399 (0.048)	1.004 (0.134)	1.001 (0.160)
	cMLE	0.599 (0.048)	0.401 (0.048)	0.999 (0.129)	0.994 (0.155)
(.5, .5)	(0,2)	0.521 (0.114)	0.479 (0.114)	1.038 (0.323)	1.006 (0.238)
	(0,3)	0.505 (0.059)	0.495 (0.059)	1.013 (0.202)	1.009 (0.150)
	(0,4)	0.502 (0.052)	0.498 (0.052)	1.000 (0.160)	1.002 (0.143)
	(0,5)	0.501 (0.049)	0.499 (0.049)	1.001 (0.146)	0.999 (0.138)
	cMLE	0.501 (0.050)	0.499 (0.050)	0.999 (0.143)	1.001 (0.143)

* 완전자료에 대한 최우추정치

본 연구에서는 혼합 이동 포아송분포에 대하여 이동모수의 값이 알려진 경우에 EM 알고리즘을 이용한 모수 추정법을 다루었다. Liu 외(2006)에서 소개된 혼합 포아송분포에 대한 EM 알고리즘과의 차이점은 이동모수에 대한 자료의 특성에 있다. 혼합분포의 한 성분분포의 이동모수 보다 적은 관측 자료는 그 성분분포로 부터의 관측치가 될 수 없다는 성질을 이용하여 성분에 관한 값을 추정하는 식을 유도하였다. 추정된

<표 4> 모의실험 결과. g=2, (Θ_1,Θ_2)=(2, 1)인 경우의 표본평균(표준오차)

(π_1,π_2)	(k_1,k_2)	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\Theta}_1$	$\hat{\Theta}_2$
(.7, .3)	(0,3)	0.709 (0.111)	0.291 (0.111)	1.995 (0.329)	0.965 (0.364)
	(0,4)	0.706 (0.069)	0.294 (0.069)	2.015 (0.262)	1.000 (0.239)
	(0,5)	0.702 (0.052)	0.298 (0.052)	2.009 (0.214)	1.000 (0.197)
	(0,6)	0.702 (0.046)	0.298 (0.046)	2.009 (0.189)	0.996 (0.189)
	(0,7)	0.700 (0.046)	0.300 (0.046)	2.000 (0.180)	1.000 (0.186)
	cMLE*	0.698 (0.046)	0.302 (0.046)	2.002 (0.167)	1.002 (0.183)
(.6, .4)	(0,3)	0.616 (0.115)	0.384 (0.115)	2.021 (0.369)	0.964 (0.284)
	(0,4)	0.606 (0.070)	0.394 (0.070)	2.010 (0.287)	0.999 (0.182)
	(0,5)	0.603 (0.055)	0.397 (0.055)	2.003 (0.228)	0.999 (0.164)
	(0,6)	0.601 (0.050)	0.399 (0.050)	2.007 (0.204)	1.003 (0.164)
	(0,7)	0.599 (0.049)	0.401 (0.049)	2.000 (0.186)	1.002 (0.160)
	cMLE	0.600 (0.049)	0.400 (0.049)	1.997 (0.178)	0.999 (0.161)
(.5, .5)	(0,3)	0.517 (0.113)	0.483 (0.113)	2.031 (0.417)	0.962 (0.198)
	(0,4)	0.506 (0.066)	0.494 (0.066)	2.025 (0.312)	0.999 (0.159)
	(0,5)	0.502 (0.054)	0.498 (0.054)	2.011 (0.248)	1.003 (0.149)
	(0,6)	0.500 (0.051)	0.500 (0.051)	1.999 (0.224)	1.002 (0.142)
	(0,7)	0.501 (0.050)	0.499 (0.050)	2.006 (0.211)	1.004 (0.144)
	cMLE	0.499 (0.050)	0.501 (0.050)	2.002 (0.205)	1.002 (0.143)

* 완전자료에 대한 최우추정치

성분 값을 이용하여 혼합분포의 모수인 혼합가중치와 포아송 성분 분포의 모수를 추정하는 EM 절차를 제시하였다. 한편 모수추정절차에서 모수의 초기값을 정하는 방법

<표 5> 모의실험 결과. g=3, (π_1, π_2, π_3)=(.33, .33, .34), $(\Theta_1, \Theta_2, \Theta_3)$=(1, 1, 1)인 경우의 표본평균(표준오차)

(k_1, k_2, k_3)	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\pi}_3$	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}_3$
(0, 3, 6)	0.334 (0.045)	0.331 (0.047)	0.336 (0.043)	1.013 (0.209)	1.024 (0.226)	1.004 (0.150)
(0, 4, 8)	0.330 (0.039)	0.331 (0.039)	0.339 (0.040)	1.004 (0.165)	1.007 (0.164)	1.000 (0.144)
(0, 5, 10)	0.329 (0.038)	0.330 (0.039)	0.340 (0.040)	1.002 (0.149)	1.003 (0.149)	1.001 (0.140)
cMLE*	0.330 (0.039)	0.330 (0.038)	0.340 (0.039)	1.001 (0.143)	1.003 (0.145)	0.998 (0.140)

* 완전자료에 대한 최우추정치

을 제시하였는데, 이 방법의 효율성에 대한 추가적 고찰이 필요할 것이다. 한편 본 연구에서는 이동모수가 기지인 경우를 다루었으나, 보다 일반적인 경우인 이동모수가 미지인 경우에 모수의 추정 방법에 대한 연구가 이루어져야 할 것이다. 초기값의 선택과 이동모수의 추정에 관한 것은 추후 연구로 남겨 두었다. 제시된 모수 추정법의 효율성을 살펴보기 위하여 모의 실험을 하였으며, 비교를 위하여 완전 자료를 이용한 모의실험도 시행하였다. 실험 결과 이동모수가 적절히 떨어져 있는 경우 불완전 자료에 대한 제시된 모수 추정 절차에 대한 추정치가 완전자료에 대한 최우추정치에 필적하였다. Liu 외(2006)의 네트워크 트래픽 자료에서처럼 혼합 포아송분포의 적용이 만족스럽지 않은 상황에서 혼합 이동 포아송분포의 적용을 시도해 볼 수 있을 것이다.

참고문헌

1. 김인영, 박수범, 김병수, 박태규 (2006). 포아송 분포의 혼합모형을 이용한 기부 횟수 자료 분석, 응용통계연구, 제 19권 1호, 1-12.
2. Church, K. W. and Gale, W. A. (1995). Poisson mixtures. *Natural Language Engineering*, Vol. 1, No. 2, 163-190.
3. 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.
4. 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.
5. Hasselblad, V. (1969). Estimation of finite mixtures from the exponential family. *Journal of the American Statistical Association* Vol. 64, 1459-1471.
6. Lia, J. and Zha, H. (2006). Two-way Poisson mixture models for simultaneous document classification and word clustering. *Computaional Statistics and Data Analysis*, Vol. 50, 163-180.

7. 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.
8. McLachlan, G. J. and Peel, D. (2001). *Finite Mixture Models*. John Wiley & Sons, Inc.
9. Schlattman, P. (2005). On bootstrapping the number of components in finite mixtures of Poisson distributions. *Statistical Computing*, Vol. 15, 179-188.
10. Yavuz, M. and Fessler, J. A. (1998). Statistical image reconstruction methods for random-precorrected PET scans. *Medical Image Analysis*, Vol. 2, No. 4, 369-378.

[2006년 6월 접수, 2006년 8월 채택]