The Korean Communications in Statistics Vol. 15, No. 1, 2008, pp. 1–11

혼합정규분포의 모수 추정에서 구간도수 EM 알고리즘의 실행 속도 개선

오창혁[1)]

요 약

혼합정규분포로부터 얻은 자료의 크기가 크면 EM 알고리즘으로 모수를 추정하는 경우 추정에 많은 시간이 걸리며 이는 실시간 음성인식 분야 등에서는 적용이 어렵게 되는 문제가 발생한다. 대용량 자료를 구간도수로 요약하여 구간도수 EM 알고리즘을 적용하면 표준 EM 알고리즘에 비해 실행속도가 획기적으로 개선되며 더욱이 구간도수 EM 알고리즘에서의 추정치의 효율성이 표준 EM 알고리즘에 근접함을 시뮬레이션 실험을 통하여 보였다.

주요용어: 구간도수 EM 알고리즘, 모의실험, 실행 속도, 혼합정규분포.

1. 서론

EM 알고리즘은 결측치를 가지는 불완전 자료에 대하여 반복적 절차로 최우추정치를 구하는 방법이며 Dempster 등 (1977)에 의하여 그 이름이 붙여졌다. 이 알고리즘의 기본 생각은 불완전 자료에서 결측치를 추정하여 완전 자료의 형태를 만들고, 이 추정된 완전 자료에 대한 우도함수를 최대화하는 방법으로 모수의 최우추정치를 구하는 것이다. EM 절차는 다양한 모형에 적용될 수 있으며, 혼합모형에서의 모수 추정에서도 유용하게 사용되고 있다 (McLachlan과 Krishnan, 1997).

혼합모형은 패턴인식 분야를 포함한 여러 곳에서 널리 사용되고 있다. 패턴인식 분야에서 많이 사용되는 숨은 마코브 모형의 경우 관측값에 대한 분포로써 흔히 혼합정규모형을 가정한다 (Rabiner와 Juang, 1993, p. 350). 패턴인식의 한 분야인 음성인식에서는 음성의 피치를 추정하는 문제 (Zolfaghari와 Robinson, 1996), 포만트 보코더를 설계하는 문제 (Zolfaghari와 Robinson, 1997), 음성의 특성치 벡터를 구하는 문제 (Stuttle과 Gales, 2001) 등에서 일변량 혼합정규모형이 활용되고 있다.

1) (712-749) 경북 경산시 대동 214-1, 영남대학교 통계학과, 교수. E-mail: choh@yu.ac.kr

혼합분포에서 얻은 자료는 각 관측값을 생성한 성분에 관한 정보가 결측된 불완전 자료로 간주될 수 있다. 이러한 경우에 대하여 Dempster 등 (1977)은 모수의 최우추정치를 구하기 위한 EM 알고리즘의 적용 방법을 예시하였다.

한편 혼합분포모형에서 얻은 원자료에 대하여 표본공간을 구간으로 분할하고 각 구간에서의 도수 값으로 축약하는 경우는, 성분에 관한 정보 외에 또 하나의 숨겨진 정보를 추가하는 것이다. 이러한 구간 도수자료에 대한 구간 도수 EM 알고리즘은 McLachlan과 Jones (1988)에 의해 연구되었다. 원자료를 구간자료로 변환하여 모수를 추정할 필요성이 발생하는 경우는 원자료 수가 너무 많아 계산 시간이 너무 많이 걸려 실용성이 떨어질 때이다. 이러한 경우는 실시간 음성인식에서 음성자료의 프리에 변환에 의한 스펙토그램을 혼합정규분포로 추정하고자 하는 경우에 발생할 수 있다. 스펙토그램을 히스토그램으로 간주하여 주파수 특성을 추정하는 경우에는 원자료의 개수는 수십만 개 이상이 될 수도 있다 (Fu 등, 2004).

EM 알고리즘의 단점의 하나로 지적되는 것은, 완만한 수렴성에 의해 요구되는 과다한 반복계산으로 인한 시간적 부담이다. 특히 영상이나 음성과 관련된 자료는 그 크기가 매우 크므로 계산의 부하가 커서 원자료에 EM 알고리즘을 적용하는 표준 EM 알고리즘은 실용적이 되지 못한다고 지적되고 있다. 이에 대한 해법의 하나로 Fu 등 (2004)은 원자료를 구간 도수자료로 변환하여 구간 도수 EM 알고리즘을 적용하는 방법을 제시하였고, 시뮬레이션 실험을 통하여, 구간의 넓이와 개수를 적절히 선택하면 추정치의 효율성이 유지되면서 계산의 부하를 대폭 감소할 수 있음을 보였다.

한편으로는 Cadez 등 (2002)은 이변량 혼합정규모형에 대하여 원자료에 대한 표준 EM 알고리즘 보다 구간 도수자료에 대한 구간 도수 EM 알고리즘이 추정치를 구하는 데 시간이 더 많이 걸린다는 주장을 하였다. 구간 알고리즘의 계산 부하는 주로 E 단계에서 구간별 기대치를 구하는 과정에서 발생하는 수치적분 계산에서 기인함을 지적하였다.

Cadez 등 (2002)은 표본공간을 적절한 구간으로 분할하는 경우에는 구간 도수자료에 의한 추정치가 원자료에 의한 추정치에 필적하거나 역설적이지만 참 모형에 더 가까운 경우도 있다는 것을 시뮬레이션으로 보였다. 그러나 실행 속도 면에 대하여는 Cadez 등 (2002)과 Fu 등 (2004)은 상반되는 결과를 제시하였다. 즉, 구간 도수 알고리즘에 대하여, Fu 등 (2004)은 실행 속도가 향상된다는 결과를, Cadez 등 (2002)은 그 반대의 결과를 제시하였다.

Samé 등 (2006)은 구간 도수 자료에 대한 EM 알고리즘을 이용하여 분류 문제를 다루었으며 성분함수의 분산공분산행렬이 대각행렬이 되도록 모형화하여 McLachlan과 Jones (1988)의 구간화 EM 알고리즘의 계산 속도를 개선하는 방법

을 제시하였다.

본 논문에서는 일변량 혼합정규모형에 대하여 원자료를 구간 도수자료로 변환할 때 표본공간을 분할하는 구간의 폭과 개수의 선택에 따른 모수 추정치의 효율성과 추정치를 얻기 위하여 걸리는 시간을 살펴본다. 2절에서는 원자료에 대한 EM 알고리즘과 구간 도수자료에 대한 EM 알고리즘을 소개하고, 구간의 폭을 좁게 하면 구간도수 EM 알고리즘에 의한 추정치가 원자료 EM 알고리즘에 의한 추정치로 수렴함을 증명한다. 그리고 3절에서는 시뮬레이션으로 다양한 구간화에 대하여 계산 속도와 추정치의 효율성을 비교하고, 4절에서는 토의와 결론을 제시한다.

2. 혼합정규분포의 최대우도법

양의 정수 g개의 성분을 가지는 일변량 혼합정규분포모형

$$f(x;\Psi) = \sum_{i=1}^{g} w_i f_i(x;\theta_i) \tag{2.1}$$

를 생각하자. 여기서 $w_i > 0$는 각 성분에 대한 가중치 $(w_1+w_2+\cdots+w_g = 1)$, f_i는 일변량 정규분포를 따르는 성분 확률밀도함수이며, 모수는 평균 μ_i와 분산 σ_i^2에 대하여 $\theta_i = (\mu_i, \sigma_i^2)$이다. 한편, $\Psi = \{\omega, \theta\} = \{w_1, \ldots, w_{g-1}, \theta_1, \ldots, \theta_g\}$는 모든 모수를 포함하는 집합으로 나타낸다.

여기에서 $x_1, \ldots, x_n$는 밀도함수 $f(x;\Psi)$를 따르는 확률표본의 관측값이라고 하자. EM 절차에서는 관측 자료 $x_1, \ldots, x_n$은 각 관측값 x_k를 발생시킨 성분에 관한 정보 $z_k = (z_{k1}, \ldots, z_{kg})$가 결측된 불완전 자료로 간주되며, 성분정보를 포함시킨 자료 $(x_1, z_1), \ldots, (x_n, z_n)$을 완전자료라고 한다. 여기서 k번째 관측값에 대한 성분 정보 $z_{k1}, z_{k2}, \ldots, z_{kg}$는 관측값 x_k를 발생시킨 성분에 대하여 1의 값을 나머지에 대하여는 0의 값을 가진다. EM 알고리즘은 관측된 불완전자료와 주어진 확률모형을 이용하여 관측되지 않은 부분을 추정하여 완전자료를 추정하는 E 단계와 관측 자료가 주어졌다는 조건 하에서 추정된 완전자료에 대한 우도를 최대화시켜 추정치를 얻는 M 단계로 구성된다.

표준 EM 알고리즘의 t번째 반복에서 모수의 현재 추정치를 $\Psi^{(t)} = (w^{(t)}, \theta^{(t)})$라고 하자. 그러면, E 단계에서 알고리즘의 t번째에서 z_{ki}에 관한 추정치는

$$\tau_i(x_k;\Psi^{(t)}) = w_i^{(t)} f_i(x_k;\theta_i^{(t)}) / f(x_k;\Psi^{(t)}) \tag{2.2}$$

로 주어진다. 간략히 나타내기 위해 $c_i^{(t)} = \sum_{k=1}^{n} \tau_i(x_k;\Psi^{(t)})$로 나타내었을 때, $(t+1)$번째의 M 단계에서 모수의 추정치는 다음과 같이 주어진다 (McLachlan과

Krishnan, 1997, p. 70).

$$\begin{aligned} w_i^{(t+1)} &= c_i^{(t)}/n, \\ \mu_i^{(t+1)} &= \frac{1}{c_i^{(t)}}\sum_{k=1}^{n} x_k \tau_i(x_k;\Psi^{(t)}), \\ \sigma_i^{2(t+1)} &= \frac{1}{c_i^{(t)}}\sum_{k=1}^{n} (x_k-\mu_i^{(t+1)})^2 \tau_i(x_k;\Psi^{(t)}). \end{aligned} \tag{2.3}$$

한편 전체 표본공간 H 를 v개의 서로 소인 부분 공간 $H_j, j=1,\ldots,v$, 즉, v개의 구간으로 분할하고 각 부분 공간에 관측되는 자료의 도수를 기록한다고 하자. 구간 도수자료는 $n_1, n_2, \ldots, n_v$로 주어진다고 하자. 각 구간 H_j에 대한 확률분포는

$$h_j(x;\Psi) = f(x;\Psi)/\int_{H_j} f(x;\Psi)dx \tag{2.4}$$

로 주어진다. 도수자료에 대한 로그 우도는 McLachlan과 Jones (1988)에 자세히 기술되어 있다.

구간 도수 EM 알고리즘의 t번째 반복에서 모수의 현재 추정치를 $\hat{\Psi}^{(t)} = (\hat{w}^{(t)}, \hat{\theta}^{(t)})$라고 하자. 구간 H_j에서의, 임의의 연속인 유계함수 $q(x)$의 기대값을

$$E_j^{(t)}[q(X)] = \int_{H_j} h_j(x;\hat{\Psi}^{(t)})q(x)dx \tag{2.5}$$

라고 정의하자. E 단계는

$$\zeta_i^{(t)}(x) = \hat{w}_i^{(t)} f_i(x;\hat{\theta}_i^{(t)})/f(x;\hat{\Psi}^{(t)}), i=1,\ldots,g \tag{2.6}$$

에 대하여

$$E_j^{(t)}[\zeta_i^{(t)}(X)], E_j^{(t)}[X\zeta_i^{(t)}(X)], E_j^{(t)}[(X-\hat{\mu}_i^{(t+1)})^2\zeta_i^{(t)}(X)], j=1,2,\ldots,v \tag{2.7}$$

를 구하게 된다. 여기서 $n_j E_j^{(t)}[\zeta_i^{(t)}(X)]$는 구간 j에서 성분 i에 대한 기대도수로 해석될 수 있다. 따라서

$$d_i^{(t)} = \sum_{j=1}^{v} n_j E_j^{(t)}[\zeta_i^{(t)}(X)] \tag{2.8}$$

는 성분 i에 의해 생성된 자료의 기대도수로 해석될 수 있다. 이들을 이용한 EM 알고리즘의 $(t+1)$번째 M 단계에서의 모수 추정치는 다음과 같이 주어진다

(McLachlan과 Jones, 1988).

$$\begin{aligned}\hat{w}_i^{(t+1)} &= d_i^{(t)}/n,\\ \hat{\mu}_i^{(t+1)} &= \frac{1}{d_i^{(t)}}\sum_{j=1}^{v} n_j E_j^{(t)}[X\zeta_i^{(t)}(X)],\\ \hat{\sigma}_i^{2(t+1)} &= \frac{1}{d_i^{(t)}}\sum_{j=1}^{v} n_j E_j^{(t)}[(X-\hat{\mu}_i^{(t+1)})^2\zeta_i^{(t)}(X)].\end{aligned} \tag{2.9}$$

원자료에 대한 표준 EM 알고리즘과는 달리 구간 도수 EM 알고리즘에서는 구간에서의 기대값 $E_j^{(t)}[\cdot]$을 이용하여 모수를 추정한다.

구간 도수 EM 알고리즘은 기본적으로 식 (2.5)에서의 기대값 $E_j^{(t)}[q(X)]$에 의존한다. 이 기대값은 식 (2.4)의 혼합정규분포에서 유도된 분포에 관한 것이므로 기대값의 실제 계산은 수치적분에 의해 이루어져야 한다. 따라서 구간 도수 EM 알고리즘의 계산 속도는 수치적분을 위한 구간의 세분화에 의존한다. 따라서 수치적분을 위해 구간을 보다 잘게 나누면 계산 양이 많아져서 계산 속도가 저하되게 된다. 표준 및 구간 도수 EM 알고리즘에 대한 각각의 모수 추정식 (2.3)과 (2.9)는 형태적으로 동일함을 알 수 있으며, 더욱이 구간도수 EM 알고리즘에서 구간의 폭을 충분히 작게 하면 추정치는 표준 EM 알고리즘의 추정치로 수렴하게 된다. 즉, 표준 및 구간 도수 EM 알고리즘의 t번째 단계에서 두 알고리즘에 의한 현재의 모수 추정치가 같은 경우에 구간 도수 EM 알고리즘에서 구간의 수를 늘리면서, 가장 넓은 구간의 폭을 0으로 수렴시키면 구간 도수 EM 알고리즘은 표준 EM 알고리즘과 같아진다. 이를 정리로 요약하면 다음과 같다.

정리 2.1 주어진 t에 대하여 $w_i^{(t)} = \hat{w}_i^{(t)}, \mu_i^{(t)} = \hat{\mu}_i^{(t)}, \sigma_i^{2(t)} = \hat{\sigma}_i^{2(t)}$라고 하자. 그러면, $v \to \infty$일 때, 구간 $H_j, j = 1, \ldots, v$의 최대 구간의 폭이 0에 수렴하면 $\hat{w}_i^{(t+1)} \to w_i^{(t+1)}, \hat{\mu}_i^{(t+1)} \to \mu_i^{(t+1)}, \hat{\sigma}_i^{2(t+1)} \to \sigma_i^{2(t+1)}$.

증명: 유한개의 자료 $x_1, \ldots, x_n$이 주어진 경우, 구간의 수 v를 충분히 크게 하면서 가장 넓은 구간의 폭을 충분히 작게 하면 v개의 구간 중 n개의 구간의 도수는 1이 되고 나머지 $n-v$개의 구간에서의 도수는 0이 된다. 만약 구간 H_j가 단 하나의 관측값 x_k를 포함하는 구간이라고 한다면, $n_j = 1$이므로,

$$\begin{aligned}n_j E_j^{(t)}[\zeta_i^{(t)}(X)] &= \int_{H_j} \frac{\hat{w}_i^{(t)} f_i(x;\hat{\theta}_i^{(t)})}{f(x;\hat{\Psi}^{(t)})} h_j(x;\hat{\Psi}^{(t)})dx\\ &\approx \frac{w_i^{(t)} f_i(x_k;\theta_i^{(t)})}{f(x_k;\Psi^{(t)})}\int_{H_j} h_j(x;\Psi^{(t)})dx = \tau_i(x_k;\Psi^{(t)})\end{aligned}$$

이다. 따라서 최대 구간의 폭이 0에 가까워지면, $n_j E_j^{(t)}[\zeta_i^{(t)}(X)] \to \tau_i(x_k;\Psi^{(t)})$임을 알 수 있다. 한편 구간 H_j의 폭이 0으로 수렴하면, $n_j E_j^{(t)}[X\zeta_i^{(t)}(X)] \to x_k\tau_i(x_k;\Psi^{(t)})$이며, $n_j E_j^{(t)}[(X-\hat{\mu}_i^{(t+1)})^2\zeta_i^{(t)}(X)] \to (x_k-\mu_i^{(t+1)})^2\tau_i(x_k;\Psi^{(t)})$이 됨도 마찬가지로 쉽게 보일 수 있다. □

자료의 구간화에 따른 정리의 의미는, 같은 초기값을 사용하고 최대 구간의 폭이 충분히 좁은 경우에 구간도수 EM 알고리즘에 의한 추정치는 원자료에 대한 표준 EM 알고리즘에 의한 추정치로 수렴한다는 것이다.

3. 시뮬레이션 실험

Fu 등 (2004)은 구간 도수 EM 알고리즘은 원자료에 대한 표준 EM 알고리즘에 비해 실행 속도가 빠르다는 결과를 제시하였다. 한편, Cadez 등 (2002)은 표준 EM 알고리즘에서는 수치적분의 필요가 없으므로 구간도수 EM 알고리즘에 비해 빠를 것이라고 반대의 주장을 하였다. 실행 속도의 차이는 표본공간을 구간화하는 구간의 개수와 각 구간에서의 수치적분의 방식 등에 따라 결정된다. 한편 Cadez 등 (2002)은 적정한 구간 개수에 대하여 구간도수 EM 알고리즘에 의한 추정치의 효율성이 표준 EM 알고리즘에 의한 추정치에 필적함을 보였으나, Fu 등 (2004)에서는 추정치 효율성에 관련한 표준 EM 알고리즘과의 비교를 발견할 수 없다.

여기서는 각 구간에서의 수치적분을 위하여 사다리꼴 공식을 적용하는 경우, 구간의 개수에 대비한 실행 속도와 추정치의 특성을 살펴본다. 한편, 시뮬레이션을 위하여 사용하는 관측값에 대한 분포는 일변량 혼합정규분포이다.

매 실험에 대하여, 주어진 성분의 개수가 $g = 3$개인 혼합정규분포로부터 $n = 500, 1000, 2000$개의 자료를 생성한 후, 이를 구간화 작업을 통하여 구간별 자료로 변환한다. 구간의 개수에 따른 실행 속도를 비교하기 위한 것이므로 구간의 개수를 고정한다. 구간화를 위하여 자료의 최소값부터 최대값 까지의 표본영역을 $B = 10, 50, 100, 200$개의 등구간으로 분할한다. 원자료에 대하여는 표준 EM 알고리즘을, 구간별 자료에 대하여는 구간도수 EM 알고리즘을 적용하였다. 추정량의 품질을 측정하기 위하여 참 모형과 추정모형 사이의 거리는 Kullback-Leibler 거리 즉, K-L 거리로 나타내었다. 표에 제시된 K-L 거리는 각 실험에서 1,000개의 반복에 대한 K-L 거리의 평균이다. 한편, 표에는 표준 EM 알고리즘의 평균 실행시간과 이에 대한 구간도수 EM 알고리즘의 평균 실행시간의 비를 괄호 안에 나타내었다.

표준 EM 알고리즘에 대한 시뮬레이션 결과는 구간도수 EM 알고리즘과의 비교를 위하여 제시하였다. 표준 EM 알고리즘에서 표의 값은 매 모수집합 추정에

표 3.1: 모의실험 결과. 평균 K-L거리와 평균 실행시간 (괄호 안), 반복횟수 1,000, 성분의 개수 $g = 3$, $(w_1,\ w_2,\ w_3) = (.33, .33, .34)$, $(v_1,\ v_2,\ v_3) = (1, 1, 1)$. 구간도수 EM 방법에서의 값은 표준 EM 방법에서의 값에 대한 비율임.

(μ_1, μ_2, μ_3)	n	표준 EM	구간도수 EM			
			$B = 10$	$B = 50$	$B = 100$	$B = 200$
	500	0.011	1.446	1.030	1.001	1.000
		(0.796)	(0.291)	(0.354)	(0.473)	(0.560)
$(-2, 0, 2)$	1,000	0.005	1.308	1.032	0.997	0.996
		(1.034)	(0.121)	(0.174)	(0.238)	(0.265)
	2,000	0.002	1.399	1.025	1.004	0.995
		(1.293)	(0.063)	(0.082)	(0.122)	(0.130)
	500	0.012	1.368	1.035	1.007	1.003
		(0.437)	(0.389)	(0.356)	(0.504)	(0.528)
$(-3, 0, 3)$	1,000	0.006	1.422	1.039	1.005	1.003
		(0.801)	(0.216)	(0.175)	(0.259)	(0.264)
	2,000	0.003	1.537	1.050	1.005	1.005
		(1.264)	(0.121)	(0.088)	(0.130)	(0.133)
	500	0.012	1.514	1.040	1.006	1.004
		(0.104)	(0.553)	(0.359)	(0.491)	(0.497)
$(-4, 0, 4)$	1,000	0.006	1.598	1.051	1.004	1.001
		(0.170)	(0.299)	(0.177)	(0.236)	(0.241)
	2,000	0.003	1.835	1.080	1.005	0.999
		(0.294)	(0.162)	(0.083)	(0.119)	(0.124)

서의 평균 K-L 거리이며 괄호 안은 추정에 걸리는 평균시간이다. 한편, 구간도수 EM 알고리즘에서는 표의 값은 표준 EM 알고리즘의 평균 K-L 거리에 대한 구간도수 EM 알고리즘의 평균 K-L 거리의 비율이다. 이 값이 1보다 크면 표준 EM 알고리즘이 더 좋은 추정을 제시하며, 1보다 적은 경우에는 그 반대이다. 괄호 안은 표준 EM 알고리즘과 구간도수 EM 알고리즘의 추정의 평균 실행시간에 대한 비율이다. 표 3.1은 혼합정규분포의 성분의 개수가 $g = 3$이며 성분 분포에 대한 가중치와 분산이 $(w_1,\ w_2,\ w_3) = (.33, .33, .34)$와 $(v_1,\ v_2,\ v_3) = (1, 1, 1)$인 경우에 대하여 성분분포의 평균 $(\mu_1,\ \mu_2,\ \mu_3) = (-2, 0, 2), (-3, 0, 3), (-4, 0, 4)$로 변화시켜 가면서 시뮬레이션 실험한 결과이다. 표 3.1에서 구간의 개수가 작을수록 평균 실행시간이 짧아지는 경향을 보이고 있다. 각 모평균 집합에 대하여 구간의 개수가 $B = 100$ 혹은 200인 경우에 구간도수 EM 알고리즘의 K-L 거리가 매우 근접함을

표 3.2: 모의실험 결과. 평균 K-L거리와 평균 실행시간 (괄호 안), 반복횟수 1,000, 성분의 개수 $g = 3$, $(w_1,\ w_2,\ w_3) = (.3, .5, .2), (v_1,\ v_2,\ v_3) = (1, 1, 1)$. 구간도수 EM 방법에서의 값은 표준 EM 방법에서의 값에 대한 비율임.

(μ_1, μ_2, μ_3)	n	표준 EM	구간도수 EM			
			$B = 10$	$B = 50$	$B = 100$	$B = 200$
	500	0.011	1.400	1.011	0.998	0.991
		(0.761)	(0.282)	(0.342)	(0.436)	(0.519)
$(-2, 0, 2)$	1,000	0.005	1.376	1.022	0.986	0.986
		(1.095)	(0.126)	(0.166)	(0.205)	(0.239)
	2,000	0.002	1.372	1.029	0.990	0.989
		(1.425)	(0.057)	(0.075)	(0.107)	(0.117)
	500	0.013	1.612	1.031	1.004	0.996
		(0.452)	(0.417)	(0.335)	(0.472)	(0.495)
$(-3, 0, 3)$	1,000	0.006	1.459	1.023	0.995	0.989
		(0.834)	(0.207)	(0.158)	(0.224)	(0.227)
	2,000	0.003	1.520	1.028	0.989	0.982
		(1.620)	(0.099)	(0.073)	(0.107)	(0.109)
	500	0.012	1.558	1.030	1.008	1.002
		(0.120)	(0.482)	(0.302)	(0.426)	(0.427)
$(-4, 0, 4)$	1,000	0.006	1.593	1.045	1.002	0.999
		(0.224)	(0.250)	(0.148)	(0.213)	(0.212)
	2,000	0.003	1.740	1.080	1.003	0.999
		(0.431)	(0.140)	(0.075)	(0.105)	(0.107)

나타낸다. 원자료의 개수가 2,000개인 경우에 구간의 개수 $B = 100$으로 하여 구간도수 EM 알고리즘을 적용하면 표준 EM 알고리즘과 비교하여 K-L 거리가 비슷하면서 실행시간은 88 % 정도 절약할 수 있음을 보이고 있다.

표 3.2는 표 3.1에서 성분분포의 가중치를 $(w_1,\ w_2,\ w_3) = (.3, .5, .2)$로 바꾸어 시뮬레이션 실험한 결과이다. 표준 EM 알고리즘에 대비한 구간도수 EM 알고리즘의 평균 실행시간은 표 3.1의 경우와 비하여 더 짧다. 구간의 개수가 $B = 100$ 혹은 200인 경우에 구간도수 EM 알고리즘의 평균 K-L 거리는 표준 EM 알고리즘의 평균 K-L 거리와 비슷하다. 더욱이 경우에 따라서는 구간도수 EM 알고리즘의 평균 K-L 거리가 더 짧게 나타나고 있다. 이는 Cadez 등 (2002)의 시뮬레이션에서도 관측된 결과이며 자료의 구간화에 따른 히스토그램의 평활에 의한 효과로 여겨지고 있다.

4. 결론 및 논의

본 논문에서 제시된 실험 결과에 의하면, 일변량 혼합정규분포에서 얻은 자료를 구간화하여 모수를 추정하는 경우에 구간도수 EM 알고리즘의 추정치에 대한 K-L 거리는 원자료의 개수보다는 구간의 개수에 더 영향을 받는다. 또한, 구간도수 EM 알고리즘은 속도를 향상시킬 뿐 아니라 K-L 거리에 의한 추정치의 효율성은 표준 EM 알고리즘에 필적한다.

실시간 음성인식과 같은 경우에서는, 추정치를 구하기 위한 실행시간이 중요한 요인의 하나이므로 음성신호자료에 대하여 혼합정규분표를 가정하여 구간도수 EM 알고리즘을 적용하는 것은 긍정적으로 고려되어질 수 있다. 실시간 음성인식에서, 음성신호자료에 대한 구간도수 EM 알고리즘의 적용 가능 여부는 추가적 연구에 의해 밝혀질 수 있을 것이다.

시뮬레이션에서 구간도수 EM 알고리즘에 대한 평균 K-L 거리가 원자료에 대한 표준 EM 알고리즘에 비해 더 짧은 현상에 대하여는 그 기저를 밝히는 추가적인 연구가 필요하다고 판단된다.

참고문헌

Cadez, I. V., McLachlan, G. J. and McLaren, C. E. (2002). Maximum likelihood estimation of mixture densities for binned and truncated multivariate data. *Machine Learning*, **47**, 7–34.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, Ser. B, **39**, 1–38.

Fu, Z., Yang, J., Hu, W. and Tan, T. (2004). Mixture clustering using multidimensional histogram for skin detection. In *Proceedings of the 17th International Conference on Pattern Recognition*, **4**, 549–552.

McLachlan, G. J. and Jones, P. N. (1988). Fitting mixture models to grouped and truncated data via the EM algorithm. *Biometrics*, **44**, 571–578.

McLachlan, G. J. and Krishnan, T. (1997). *The EM Algorithm and Extensions.* John Wiley & Sons, New York.

Rabiner, L. and Juang, B. (1993). *Fundamentals of Speech Recognition*, Prentice Hall, New Jersey.

Samé, A., Ambroise, C. and Govaert, G. (2006). A classification EM algorithm for binned data. *Computational Statistics & Data Analysis*, **51**, 466–480.

Stuttle, M. N. and Gales, M. J. F. (2001). A mixture of Gaussians front end for speech recognition. In *Proceedings Eurospeech 2001.*

Zolfaghari, P. and Robinson, T. (1996). Formant analysis using mixtures of Gaussians. In *Proceedings ICSLP 96: Fourth International Conference on Spoken Language Processing.*

Zolfaghari, P. and Robinson, T. (1997). A segmental formant vocoder based on linearly varying mixture of Gaussians. In *Proceedings Eurospeech '97.*

[2007년 5월 접수, 2007년 9월 채택]

Speedup of EM Algorithm by Binning Data for Normal Mixtures

Chang Hyuck Oh[1)]

Abstract

For a large data set the high computational cost of estimating the parameters of normal mixtures with the conventional EM algorithm is crucially impedimental in applying the algorithm to the areas requiring high speed computation such as real-time speech recognition. Simulations show that the binned EM algorithm, being compared to the standard one, significantly reduces the cost of computation without loss in accuracy of the final estimates.

Keywords: Binned EM algorithm; execution time; normal mixtures; simulation.

1) Professor, Department of Statistics, Yeungnam University, Dae-dong 214-1, Gyungsan, Gyungbuk 712-749, Korea. E-mail: choh@yu.ac.kr