Proceedings of Joint Conference of Korean Data And Information Science Society and The Korean Data Analysis Society, April 30 - May 1, 2004, pp. 127-130

Density Estimation of Mixture Normal Distribution with Binned Data Using Nonlinear Regression

나영호[1] · 오창혁[2]

요 약

혼합정규분포에서 얻어진 히스토그램 자료에서 모수의 추정은 EM 알고리즘 혹은 스프라인 방법이 흔히 이용되고 있다. 본 논문에서는 히스토그램 자료를 비선형회귀모형으로 적합하는 방법을 제시하고, 시뮬레이션으로 제시된 방법과 EM 알고리즘 방법을 비교하였다.

Keywords : 혼합정규분포, 히스토그램 자료, 비선형회귀모형.

1. 서론

혼합정규분포에서 얻어진 표본의 경우 모수의 추정은 EM 알고리즘으로 구하는 것이 잘 알려져 있다. Dempster 외(1977) 참조. 혼합정규분포에서 얻어진 표본을 구간별 도수로 분류한 히스토그램 자료에 대하여서도 MacLachlan 외(1988)는 EM 알고리즘을 적용한 모수추정방법을 제시하고 있다. 한편 Koopersberg 와 Stone(1991) 등은 스프라인 방법으로 분포가정을 하지 않은 상태에서 밀도함수의 추정을 하는 방법을 제시하였다. 본 논문에서는 혼합정규분포 가정 하에서 히스토그램 자료를 적절한 회귀모형으로 적합하여 모수를 추정하는 방법을 생각하고 이 방법과 MacLachlan 외(1988)의 방법을 시뮬레이션으로 비교하여 제시된 방법이 우수함을 보인다. 2절에서는 비선형회귀모형과 최소제곱법으로 모수 추정하는 알고리즘을 소개하고, 3절에서는 시뮬레이션 실험결과를 보인다.

2. 비선형회귀모형을 이용한 혼합정규분포의 모수추정

확률변수 X는 혼합정규확률밀도함수

$$f(x;\theta)=\sum_{k=1}^{r}\omega_k\frac{1}{\sigma_k\sqrt{2\pi}}\exp\left(-\frac{(x-\mu_k)^2}{2\sigma_k^2}\right)$$

를 가진다고 하자. 다음과 같은 r+2개의 구간을 생각하자.

[1]712-749, 경북 경산시 대동 214-1, 영남대학교 수학통계학부 통계학 전공 박사과정 수료.
[2]712-749, 경북 경산시 대동 214-1, 영남대학교 수학통계학부 통계학 전공 교수.

$$B_0=\left(-\infty, b_0+\frac{b_1-b_0}{2}\right)$$

$$B_{i+1}=\left(b_i-\frac{b_i-b_{i-1}}{2}, b_i+\frac{b_{i+1}-b_i}{2}\right), i=1,2,\ldots,r,$$

$$B_{r+1}=\left(b_{r+1}-\frac{b_{r+1}-b_r}{2}, \infty,\right)$$

단, $b_0<b_1<\ldots<b_{r+1}$ 이며, 이들 값은 구간의 중앙값으로 생각할 수 있다. 확률변수 X에 대한 독립표본에서 이들 구간 B_j, $(j=0,1,\ldots r+1)$에 떨어지는 표본의 개수 n_j 를 기록한다고 하자. 이렇게 하여 얻어진 자료를

$$(b_0, n_0),(b_1, n_1),\ldots,(b_{r+1}, n_{r+1}),$$

라고 나타내자. 이 자료에 대하여 비선형 회귀모형

$$n_j=\sum_{k=1}^{g}\omega_k\frac{1}{\sigma_k\sqrt{2\pi}}\exp\left(-\frac{(b_j-\mu_k)^2}{2\sigma_k^2}\right)+\varepsilon_j,\ j=0,1,2,\ldots,r+1$$

을 가정하자. $\varepsilon_0, \varepsilon_1, \ldots \varepsilon_{r+1}$은 평균이 0이며 유한분산을 가지는 iid 변수임을 가정한다. 모수 $(\omega_k, \mu_k, \sigma_k)$, k=1, 2,..., g 는

$$SSE=\sum_{j=0}^{r+1}\left(n_j-\sum_{k=1}^{g}\omega_k\frac{1}{\sigma_k\sqrt{2\pi}}\exp\left(-\frac{(b_j-\mu_k)^2}{2\sigma_k^2}\right)\right)^2,$$

를 최소로 하는 값으로 추정하려고 한다. 그러나 SSE를 최소화하는 대수적 해를 구하는 것은 난망하다. 따라서 수치적 방법으로 최소화 해를 구하는 알고리즘을 제시한다.

알고리즘

단계 1. 모수들의 초기값을 다음과 같이 둔다.

$$\omega_k=\frac{1}{g}, \mu_k=b_0+(b_{r+1}-b_0)\frac{k+1}{g+1}, \sigma_k=\frac{\sqrt{b_{r+1}-b_0}}{g\times 1.5}, \text{ k=1,2, ..., g}$$

단계 2. SSE를 최소로 하는 평균의 추정. 가중치와 분산 ω_k, σ_k, k=1,2,...,g를 고정한다. 각각의 k=1, 2, .., g에 대하여, k 번째 평균값 μ_k를 제외한 나머지 평균값을 고정한 채로 μ_k의 값을 b_0, b_1, ..., b_{r+1} 까지 증가시켜 가면서 SSE를 최소로 하는 값을 구하여 그 값을 μ_k 의 새로운 추정값으로 한다.

단계 3. SSE를 최소로 하는 가중치의 추정. 가중치와 분산 μ_k, σ_k, k=1,2,...,g를 고정한다. 각각의

k=1, 2, .., g에 대하여, k 번째 가중치 ω_k 값을 0.01(0.01)0.99로 증가시켜 가면서 SSE를 최소로 하는 값을 구하여 그 값을 ω_k의 새로운 추정값으로 한다. 여기서는 가중치의 합이 1이 되는 조건을 만족하도록 한다.

단계 4. SSE를 최소로 하는 분산의 추정. 가중치와 분산 ω_k, μ_k, k=1,2,...,g를 고정한다. 각각의 k=1, 2, .., g에 대하여, k 번째 분산 σ_k^2를 제외한 나머지 분산값을 고정한 채로 σ_k^2의 값을 0.01(0.01) ∞로 증가시켜 가면서 SSE가 감소하다가 증가하는 최소점의 값을 σ_k^2의 새로운 추정값으로 한다.

단계 5. 반복의 종료. 새롭게 추정된 모수의 값들이 이전 모수추정값과 차이가 정해진 범위 내에 도달하면 종료하고 그렇지 않으면 단계 2로 간다.

단계 4에서는 분산값을 증가시켜 가면 SSE가 감소하다가 증가하는 형태를 띄게 된다. 따라서 최소점이 되는 값을 추정값으로 정하는 규칙을 따르도록 한다.

3. 시뮬레이션

앞의 2절에 주어진 알고리즘과 MacLachlan 외(1988)의 EM 방법을 비교하기 위하여, 구성원소의 수가 3인 경우에 여러 가지 모수값에 대하여 시뮬레이션 실험을 하였다. 평균이 충분히 떨어진 경우, 즉 평균의 차이가 분산의 3배 이상인 경우 두 가지 방법은 유사하였다. 다음 표 1에는 평균값의 차이를 줄여 가면서 실험한 결과이다. 표에는 분포수 생성에 사용된 가중치, 평균, 분산의 값이 상단에 있으며, LSE 행은 제시된 방법에 대한 추정치의 표본평균과 표본분산, MJ 행은 EM 알고리즘에 대한 추정치와 표본평균과 표본분산이다. 제시된 표의 경우에는 본 논문에서 제시된 방법이 우수함을 보이고 있다.

표 1. 추정치의 표본평균과 표본분산

표 1. 추정치의 표본평균과 표본분산									
	weight			mean			variance		
	0.500	0.300	0.200	-7.000	0.000	7.000	2.000	1.000	1.000
LSE	0.500	0.300	0.200	-7.004	0.004	7.002	2.035	1.029	1.013
	(0.000)	(0.000)	(0.000)	(0.006)	(0.006)	(0.009)	(0.034)	(0.014)	(0.027)
MJ	0.500	0.299	0.201	-7.014	0.004	7.021	2.056	0.997	1.058
	(0.000)	(0.000)	(0.000)	(0.004)	(0.003)	(0.005)	(0.018)	(0.007)	(0.011)

	weight			mean			variance		
	0.500	0.300	0.200	-7.000	0.000	5.000	2.000	1.000	1.000
LSE	0.500	0.300	0.200	-7.000	0.002	5.000	2.046	1.019	1.014
	(0.000)	(0.000)	(0.000)	(0.007)	(0.006)	(0.008)	(0.036)	(0.014)	(0.028)
MJ	0.500	0.296	0.204	-7.011	-0.044	4.987	2.057	1.167	1.124
	(0.000)	(0.001)	(0.001)	(0.004)	(0.262)	(0.079)	(0.020)	(4.751)	(0.400)

	weight			mean			variance		
	0.500	0.300	0.200	-7.000	0.000	4.000	2.000	1.000	1.000
LSE	0.500	0.299	0.202	-7.004	-0.005	3.996	2.035	1.007	1.053
	(0.000)	(0.000)	(0.000)	(0.007)	(0.007)	(0.010)	(0.036)	(0.020)	(0.041)
MJ	0.497	0.266	0.237	-7.013	-0.846	3.718	2.020	2.324	1.585
	(0.000)	(0.008)	(0.009)	(0.004)	(6.084)	(0.599)	(0.026)	(24.739)	(1.804)

	weight			mean			variance		
	0.500	0.300	0.200	-7.000	0.000	3.000	2.000	1.000	1.000
LSE	0.496	0.270	0.234	-7.003	-0.137	2.752	2.033	0.920	1.394
	(0.001)	(0.001)	(0.003)	(0.009)	(0.965)	(0.286)	(0.042)	(0.060)	(0.350)
MJ	0.491	0.075	0.435	-7.012	-3.535	1.496	1.925	7.781	2.955
	(0.000)	(0.010)	(0.010)	(0.005)	(16.280)	(0.421)	(0.027)	(68.050)	(0.676)

	weight			mean			variance		
	0.500	0.300	0.200	-7.000	0.000	2.000	2.000	1.000	1.000
LSE	0.443	0.162	0.395	-7.100	-1.882	0.868	1.875	1.524	1.833
	(0.005)	(0.004)	(0.012)	(0.043)	(33.026)	(0.266)	(0.095)	(16.250)	(0.316)
MJ	0.483	0.021	0.496	-7.011	-4.741	0.797	1.911	12.663	2.011
	(0.001)	(0.000)	(0.000)	(0.006)	(7.925)	(0.006)	(0.024)	(85.175)	(0.021)

참고문헌

[1] Dempster, A. P., Laird, N. M., and Rubin, D. R.(1977) Maximum likelihood from incomplete data. *Journal of the Royal Statistical Society,* Series B 39, 1-38.

[2] Koopersberg, C. and Stone, C. J.(1991) A study of logspline density estimation. *Computational Statistics & Data Analysis,* 12, 327-347.

[3] MacLachlan G. J. and Jones, P. N. (1988) Fitting mixture models to grouped and truncated data via the EM algorithm. *Biometrics,* 44, 57-578.