Journal of the Korean Data & Information Science Society
2010, **21**(5), 863–871

한국데이터정보과학회지

겹친라플라스 혼합분포를 통한 첨 · 다봉형 비대칭 원형자료의 모형화

나종화[1] · 장영미[2]

[1]충북대학교 정보통계학과
[2]한국보건복지정보개발원

접수 2010년 6월 27일, 수정 2010년 8월 19일, 게재확정 2010년 8월 28일

요 약

지금까지 원형자료의 적합에 대한 연구는 주로 von Mises, 겹친왜정규 분포를 비롯하여 주로 완만한 봉우리를 가지는 대칭 및 비대칭의 경우에 대해 수행되어 왔다. 본 논문에서는 뾰족한 봉우리를 가지며 정점을 중심으로 비대칭의 경향이 심한 첨봉형의 비대칭 원형자료에 대한 적합을 다루었다. 최근 Jammalamadaka와 Kozubowski (2003)가 소개한 겹친라플라스 분포와 그의 혼합분포를 중심으로 단봉형 및 다봉형의 원형자료에 대한 모형화 과정을 다루었다. 특히 혼합분포의 모수추정을 위해 EM 알고리즘을 사용하였으며, 모의실험을 통해 그 정확도를 확인하였다.

주요용어: 겹친라플라스 분포, 원형자료, 혼합분포.

1. 서론

원형자료 (circular data)는 2차원의 방향자료 (directional data)를 의미하며, 각각의 자료는 크기가 1인 단위원 상의 점 또는 원점으로부터의 단위벡터로 표현되거나 하나의 각 ($0^\circ \sim 360^\circ$ 또는 $0 \sim 2\pi$)으로 표현될 수 있다. 원형자료는 많은 응용분야에서 나타나며, 최근에는 시계열 등 특정자료의 원형자료로의 변환을 통한 분석도 활발히 이루어지고 있다. 원형 및 방향자료와 관련된 대표적 문헌으로는 Mardia (1972), Fisher (1993), Mardia와 Jupp (1999), Jammalamadaka와 SenGupta (2001)를 들 수 있다.

원형자료에 대한 적합모형으로는 대칭형의 경우 von Mises 분포를 중심으로 겹친정규, 겹친코쉬, Cardioid 및 원형균일 분포 등이 있으며, 비대칭의 경우로는 Papakonstantinou (1978), Batschelet (1981), 겹친 α-stable 분포와 최근 Pewsey (2000)에 의해 소개된 겹친왜정규 분포 등이 있다. 그러나 이들 분포들은 실재의 응용에서 자주 발생되는 뾰족한 봉우리를 가지는 자료의 적합에는 한계를 가진다.

본 논문에서는 첨봉형 (sharply peaked)의 비대칭 원형자료에 대한 적합을 다루고자 한다. 이를 위해 최근 Jammalamadaka와 Kozubowski (2003)가 제안한 겹친라플라스 (wrapped Laplace) 분포를 소개하고, 이의 모수추정 문제를 다루고자 한다. 또한 제안된 분포의 혼합분포를 통해 다봉형의 원형자료에 대한 적합을 수행한다. 다봉형의 원형자료의 적합과 관련된 응용으로는 von Mises 혼합분포에 대한 Mooney 등 (2003), Jang 등 (2007)과 겹친왜정규 혼합분포에 대한 Pewsey (2006), Na와 Jang (2010a)의 연구가 있다.

[1] 교신저자: (361-763) 충북 청주시 흥덕구 개신동 12, 충북대학교 정보통계학과, 교수.
E-mail: cherin@cbu.ac.kr
[2] (100-705) 서울시 중구 충무로 3가 60-1 극동빌딩 21층, 한국보건복지정보개발원, 박사.

2절에서는 겹친라플라스 분포와 모수 추정에 대해 다루었으며, 3절에서는 위 분포의 혼합분포를 통한 다봉형 자료의 적합과 EM 알고리즘을 이용한 모수추정 과정을 다루었다. 4절에서는 모의실험을 통해 모수추정의 정도를 확인하였다.

2. 겹친라플라스 분포와 모수추정

2.1. 겹친라플라스 분포 소개

겹친라플라스 분포는 선형 상의 (대칭인) 라플라스 분포를 비대칭의 경우로 확장한 뒤, 이에 겹침(wrapping)의 원리를 적용하여 얻어지는 원형분포로 비대칭의 원형자료의 적합에 용이하다. 겹친라플라스 분포는 겹친지수와 겹친음지수 분포의 혼합분포로 다음과 같이 정의된다.

$$f_w(\theta) = p\frac{\lambda_1 e^{-\lambda_1\theta}}{1-e^{-2\pi\lambda_1}} + (1-p)\frac{\lambda_2 e^{\lambda_2\theta}}{e^{2\pi\lambda_2}-1}, \quad \theta \in [0, 2\pi). \tag{2.1}$$

위 식에서 $\lambda_1, \lambda_2 > 0$이고 $p \in [0,1]$이며, 이 분포를 $\Theta \sim WL^*(\lambda_1, \lambda_2)$로 나타내기로 한다. 특히, Jammalamadaka와 Kozubowski (2003)는 식 (2.1)에서 다음의 모수관계

$$p = \frac{1}{\kappa^2+1}, \quad \lambda_1 = \lambda\kappa, \quad \lambda_2 = \lambda/\kappa \tag{2.2}$$

를 만족하는 겹친라플라스 분포를 다음과 같이

$$f_w(\theta) = \frac{\lambda\kappa}{1+\kappa^2}\left(\frac{e^{-\lambda\kappa\theta}}{1-e^{-2\pi\lambda\kappa}} + \frac{e^{(\lambda/\kappa)\theta}}{e^{2\pi\lambda/\kappa}-1}\right), \quad \theta \in [0, 2\pi) \tag{2.3}$$

으로 정의하고, 이를 $\Theta \sim WL(\lambda, \kappa)$로 표현하였다. 위 식에서 $\lambda, \kappa > 0$이고, κ는 왜도 모수를 나타낸다. 또한 $\kappa = 1$ (즉, $p = 1/2$)일 때 대칭인 형태가 되며, 단봉형의 유일한 최빈값을 가진다. 위치모수 $\eta \in [0, 2\pi)$가 고려된 $f_w(\theta - \eta)$에서 위치모수 η는 최빈값에 대응된다.

다음의 그림 2.1은 모수 λ와 κ에 따른 겹친 라플라스 분포의 형태를 나타낸다. 이 그림에서 동일한 κ값에 대해 모수 λ가 커짐에 따라 봉우리가 점점 뾰족해지는 형태를 띠며, κ의 값이 1로부터 벗어날 때 비대칭의 정도가 심해지는 특징을 가짐을 알 수 있다.

(a) $\lambda = 3$, $\kappa = 0.5, 1, 3, 7$ 일 때 **(b)** $\lambda = 5$, $\kappa = 0.5, 1, 3, 7$ 일 때

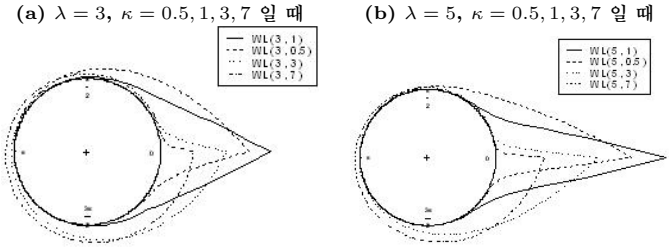

그림 **2.1** $WL(\lambda, \kappa)$분포의 형태

2.2. 모수추정

여기서는 겹친라플라스 분포의 최대가능도 (ML) 추정 문제를 다루기로 한다. 먼저 원형자료 $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_n)$에 대해 로그가능도함수는, 식 (2.1)과 (2.2)의 관계로부터 ($p = \lambda_2/(\lambda_1+\lambda_2)$),

$$l = \ln L(\lambda_1, \lambda_2 | \theta_1, \cdots, \theta_n) = n \ln \left(\frac{\lambda_1 \lambda_2}{\lambda_1 + \lambda_2} \right) + \sum_{i=1}^{n} \ln \left\{ \frac{e^{-\lambda_1 \theta_i}}{1 - e^{-2\pi\lambda_1}} + \frac{e^{\lambda_2 \theta_i}}{e^{2\pi\lambda_2} - 1} \right\}$$

으로 주어진다.

위 식을 최대로 하는 모수값은 Nelder-Mead의 심플렉스 방법 (Nelder와 Mead, 1965)을 사용하거나, 다음의 가능도방정식

$$\frac{\partial}{\partial \lambda_1} l = n \left(\frac{1}{\lambda_1} - \frac{1}{\lambda_1 + \lambda_2} \right) + \sum_{i=1}^{n} \frac{(e^{2\pi\lambda_2} - 1) \left\{ (\theta_i - 2\pi) e^{-2\pi\lambda_1 - \lambda_1 \theta_i} - \theta_i e^{-\lambda_1 \theta_i} \right\}}{(1 - e^{-2\pi\lambda_1}) \left\{ e^{-\lambda_1 \theta_i} (e^{2\pi\lambda_2} - 1) + e^{\lambda_2 \theta_i} (1 - e^{-2\pi\lambda_1}) \right\}} = 0,$$

$$\frac{\partial}{\partial \lambda_2} l = n \left(\frac{1}{\lambda_2} - \frac{1}{\lambda_1 + \lambda_2} \right) + \sum_{i=1}^{n} \frac{(1 - e^{-2\pi\lambda_1}) \left\{ (\theta_i - 2\pi) e^{2\pi\lambda_2 + \lambda_2 \theta_i} - \theta_i e^{\lambda_2 \theta_i} \right\}}{(e^{2\pi\lambda_2} - 1) \left\{ e^{-\lambda_1 \theta_i} (e^{2\pi\lambda_2} - 1) + e^{\lambda_2 \theta_i} (1 - e^{-2\pi\lambda_1}) \right\}} = 0$$

을 Newton-Raphson 방법을 통해 구할 수 있다. 다만 Newton-Raphson 반복식에 사용되는 2차 미분식은 다음과 같이 주어진다.

$$\frac{\partial^2}{\partial \lambda_1^2} l = n \left(-\frac{1}{\lambda_1^2} + \frac{1}{(\lambda_1 + \lambda_2)^2} \right) + \sum_{i=1}^{n} \frac{\partial^2}{\partial \lambda_1^2} \ln \left(\frac{e^{-\lambda_1 \theta_i}}{1 - e^{-2pi\lambda_1}} + \frac{e^{\lambda_2 \theta_i}}{e^{2\pi\lambda_2} - 1} \right),$$

$$\frac{\partial^2}{\partial \lambda_2^2} l = n \left(-\frac{1}{\lambda_2^2} + \frac{1}{(\lambda_1 + \lambda_2)^2} \right) + \sum_{i=1}^{n} \frac{\partial^2}{\partial \lambda_2^2} \ln \left(\frac{e^{-\lambda_1 \theta_i}}{1 - e^{-2pi\lambda_1}} + \frac{e^{\lambda_2 \theta_i}}{e^{2\pi\lambda_2} - 1} \right),$$

$$\frac{\partial^2}{\partial \lambda_1 \lambda_2} l = \frac{n}{(\lambda_1 + \lambda_2)^2} + \sum_{i=1}^{n} \frac{\partial^2}{\partial \lambda_1 \lambda_2} \ln \left(\frac{e^{-\lambda_1 \theta_i}}{1 - e^{-2pi\lambda_1}} + \frac{e^{\lambda_2 \theta_i}}{e^{2\pi\lambda_2} - 1} \right).$$

여기서, $\ln \left(\frac{e^{-\lambda_1 \theta_i}}{1 - e^{-2pi\lambda_1}} + \frac{e^{\lambda_2 \theta_i}}{e^{2\pi\lambda_2} - 1} \right)$을 A라 두면

$$\frac{\partial^2}{\partial \lambda_1^2} A = -\frac{(b_1 + c_1)^2}{a^2} + \frac{d_1 + (2\pi + \theta_i) b_1}{a} + \frac{\theta_i (b_1 + c_1)}{a},$$

$$\frac{\partial^2}{\partial \lambda_2^2} A = -\frac{(b_2 + c_2)^2}{a^2} + \frac{d_2 + (2\pi + \theta_i) b_2}{a} + \frac{\theta_i (b_2 + c_2)}{a},$$

$$\frac{\partial^2}{\partial \lambda_1 \partial \lambda_2} A = \frac{\partial^2}{\partial \lambda_2 \partial \lambda_1} A = \frac{(b_1 + c_1)(b_2 + c_2)}{a^2}$$

와 같다. 위 식에서 $a, b_1, c_1, d_1, b_2, c_2, d_2$는 다음과 같다.

$$a = \frac{e^{-\lambda_1 \theta_i}}{1 - e^{-2pi\lambda_1}} + \frac{e^{\lambda_2 \theta_i}}{e^{2pi\lambda_2} - 1},$$

$$b_1 = \frac{2\pi e^{-2pi\lambda_1 - \lambda_1 \theta_i}}{(1 - e^{-2\pi\lambda_1})^2}, c_1 = \frac{\theta_i e^{-\lambda_1 \theta_i}}{1 - e^{-2pi\lambda_1}}, d_1 = \frac{8\pi^2 e^{-4pi\lambda_1 - \lambda_1 \theta_i}}{(1 - e^{-2\pi\lambda_1})^3},$$

$$b_2 = -\frac{2\pi e^{2pi\lambda_2 + \lambda_2 \theta_i}}{(e^{2\pi\lambda_2} - 1)^2}, c_2 = \frac{\theta_i e^{\lambda_2 \theta_i}}{e^{2pi\lambda_2} - 1}, d_2 = \frac{8\pi^2 e^{4pi\lambda_2 + \lambda_2 \theta_i}}{(e^{2\pi\lambda_2} - 1)^3}.$$

이 과정에서 모수들에 대한 초기치는 다음의 적률추정량 (Jammalamadaka와 Kozubowski, 2003)으로부터 구해질 수 있다.

$$\widehat{\lambda} = \left(\frac{12r_1^2 - r_2^2}{3r_1^2 r_2^2 - 4r_2^2 + r_1^2} \right)^{1/4},$$

$$\widehat{\kappa} = \sqrt{\frac{\widehat{\lambda}^4 - r_1^2(1+\widehat{\lambda}^4) + \sqrt{(r_1^2 + r_1^2\widehat{\lambda}^4 - \widehat{\lambda}^4)^2 - 4\widehat{\lambda}^4 r_1^4}}{2\widehat{\lambda}^2 r_1^2}}.$$

위 식에서 r_1과 r_2는 각각

$$r_1 = \sqrt{\left(\frac{1}{n}\sum_{j=1}^{n} \cos\theta_j \right)^2 + \left(\frac{1}{n}\sum_{j=1}^{n} \sin\theta_j \right)^2},$$

$$r_2 = \sqrt{\left(\frac{1}{n}\sum_{j=1}^{n} \cos(2\theta_j) \right)^2 + \left(\frac{1}{n}\sum_{j=1}^{n} \sin(2\theta_j) \right)^2}$$

으로 정의된다.

3. 겹친라플라스 혼합분포와 모수추정

3.1. 혼합분포와 EM 알고리즘

일반적으로 k개의 밀도함수로 구성되는 혼합분포를 다음과 같이 정의된다.

$$f(\theta|\boldsymbol{\Lambda}) = \sum_{l=1}^{k} \alpha_l f_l(\theta|\boldsymbol{\lambda_l}). \tag{3.1}$$

여기서 $\boldsymbol{\lambda}$는 모수집합으로 $\boldsymbol{\lambda} = \{\alpha_1, \alpha_2, \cdots, \alpha_k, \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \cdots, \boldsymbol{\lambda}_k\}$이고, α_l은 각 단일분포의 혼합비율로 $\sum_{l=1}^{k} \alpha_l = 1$을 만족한다. 위 혼합분포로부터의 표본을 $\boldsymbol{\theta} = \{\theta_1, \theta_2, \cdots, \theta_n\}$이라 하자. 이때, 로그가능도함수는

$$\ln(L(\boldsymbol{\lambda}|\boldsymbol{\theta})) = \ln \prod_{i=1}^{n} f(\theta_i|\boldsymbol{\lambda}) = \sum_{i=1}^{n} \ln \left(\sum_{l=1}^{k} \alpha_l f_l(\theta_i|\boldsymbol{\lambda_l}) \right) \tag{3.2}$$

으로 주어진다.

위 식을 모수에 대해 직접 최대화하는 것은 매우 어려우므로, 본 연구에서는 EM 알고리즘 (Dempster 등, 1977)을 통해 최대화 과정을 수행하기로 한다. 그 과정은 다음과 같다. 먼저 관측자료 $\boldsymbol{\theta}$를 불완비자료로, 관측되지 않은 $\boldsymbol{\psi} = \{\psi_1, \psi_2, \cdots, \psi_n\}$을 잠재자료라고 하자. 각 i에 대해 $\psi_i \in 1, 2, \cdots, k$이고, i번째 관측치 θ_i가 l번째 분포 f_l로부터 생성되었다면 $\psi_i = l$로 주어진다. 잠재자료 $\boldsymbol{\psi}$의 값을 아는 경우 완비자료 $(\boldsymbol{\theta}, \boldsymbol{\psi})$의 로그완비가능도함수는

$$\ln(L(\boldsymbol{\lambda}|\boldsymbol{\theta},\boldsymbol{\psi})) = \ln(f(\boldsymbol{\theta},\boldsymbol{\psi}|\boldsymbol{\lambda})) = \sum_{i=1}^{n} \ln(\alpha_{\psi_i} f_{\psi_i}(\theta_i|\boldsymbol{\lambda}_{\boldsymbol{\psi_i}})) \tag{3.3}$$

으로 표현되며, 위 식에 대한 최대화는 쉽게 수행될 수 있게 된다. 그러나 일반적으로 자료 θ_i에 대응하는 잠재 자료 ψ_i의 값을 모르기 때문에 $\Psi|(\theta,\lambda)$의 분포를 추정한 후 (E-단계), 이 분포에 대한 식 (3.3)의 (조건부) 기대값을 최대화 (M-단계)하는 모수를 찾게 된다.

여기서 $\Psi|(\theta,\lambda)$ 의 분포는 $\boldsymbol{\lambda}^g = (\alpha_1^g, \alpha_2^g, \cdots, \alpha_k^g, \boldsymbol{\lambda}_1^g, \boldsymbol{\lambda}_2^g, \cdots, \boldsymbol{\lambda}_k^g)$가 주어질 때, 각 i와 l에 대해 다음과 같이 주어진다.

$$p(l|\theta_i, \boldsymbol{\lambda}^g) = \frac{\alpha_l^g f_l(\theta_i|\boldsymbol{\lambda}_l^g)}{f(\theta_i|\boldsymbol{\lambda}^g)} = \frac{\alpha_l^g f_l(\theta_i|\boldsymbol{\lambda}_l^g)}{\sum_{l=1}^k \alpha_l^g f_l(\theta_i|\boldsymbol{\lambda}_l^g)}.$$

또한 로그완비가능도함수에 대한 조건부 기댓값은 다음과 같이 주어진다.

$$\begin{aligned} Q(\boldsymbol{\lambda},\boldsymbol{\lambda}^{\boldsymbol{g}}) &= \sum_{i=1}^n E_{\Psi_i|\theta_i,\boldsymbol{\lambda}^{\boldsymbol{g}}}[\ln \alpha_{\psi_i} f_{\psi_i}(\theta_i|\boldsymbol{\lambda}_{\boldsymbol{\psi_i}})] \\ &= \sum_{l=1}^k \sum_{i=1}^n \ln(\alpha_l f_l(\theta_i|\boldsymbol{\lambda_l}))p(l|\theta_i, \boldsymbol{\lambda^g}) \\ &= \sum_{l=1}^k \sum_{i=1}^n \ln(\alpha_l) p(l|\theta_i, \boldsymbol{\lambda^g}) + \sum_{l=1}^k \sum_{i=1}^n \ln(f_l(\theta_i|\boldsymbol{\lambda_l}))p(l|\theta_i, \boldsymbol{\lambda^g}). \end{aligned} \tag{3.4}$$

EM 알고리즘은 조건부 기댓값인 식 (3.4)를 구하는 단계 (E-단계)와 식 (3.4)의 Q 함수를 최대로 하는 모수 $\boldsymbol{\lambda}$를 구하는 과정 (M-단계)을 반복적으로 수행해 나감으로써 모수의 추정치를 개선해 나가는 일종의 반복 알고리즘이다. 이때, 반복횟수는 Q 함수의 최대값의 변화가 충분히 작아질 때 까지 수행한다.

EM 알고리즘에 대한 보다 자세한 내용은 McLachlan과 Krishnan (1977), Tanner (1996), 혼합분포에 대한 통계분석에 대해서는 Titterington 등 (1985)을 참고하기 바란다.

3.2. 겹친라플라스 혼합분포의 모수추정

먼저 혼합비율 $\alpha_1, \alpha_2, \cdots, \alpha_k$에 대한 추정은 다음과 같다. 식 (3.4)에서 혼합비율은 첫 번째 항에만 관련되며, 제약조건 $\sum_{l=1}^k \alpha_l = 1$ 하에서 라그랑지 방법을 적용하면

$$\hat{\alpha}^{new} = \frac{1}{n}\sum_{i=1}^n p(l|\theta_i, \boldsymbol{\lambda^g}) \tag{3.5}$$

으로 추정된다. 나머지 모수들의 추정은 식 (3.4)의 두 번째 항에만 관련되며, 이 항은

$$\begin{aligned} T &= \sum_{i=1}^n \sum_{l=1}^k (\ln f_l(\theta_i|\boldsymbol{\lambda_l}))p(l|\theta_i, \boldsymbol{\lambda^g}) \\ &= \sum_{i=1}^n \sum_{l=1}^k \ln\left(\frac{\lambda_{1l}\lambda_{2l}}{\lambda_{1l}+\lambda_{2l}}\right) p(l|\theta_i, \boldsymbol{\lambda^g}) + \sum_{i=1}^n \sum_{l=1}^k \ln\left\{\frac{e^{-\lambda_{1l}\theta_i}}{1-e^{-2\pi\lambda_{1l}}} + \frac{e^{\lambda_{2l}\theta_i}}{e^{2\pi\lambda_{2l}}-1}\right\} p(l|\theta_i, \boldsymbol{\lambda^g}) \end{aligned}$$

으로 표현된다.

위의 식 T를 최대로 하는 모수는 다음의 연립방정식

$$\frac{\partial}{\partial\lambda_{1l}}T=\sum_{i=1}^{n}\left(\frac{1}{\lambda_{1l}}-\frac{1}{\lambda_{1l}+\lambda_{2l}}\right)p(l|\theta_i,\boldsymbol{\lambda}^{\boldsymbol{g}})$$
$$+\sum_{i=1}^{n}\frac{(e^{2\pi\lambda_{2l}}-1)\left\{(\theta_i-2\pi)e^{-2\pi\lambda_{1l}-\lambda_{1l}\theta_i}-\theta_i e^{-\lambda_{1l}\theta_i}\right\}}{(1-e^{-2\pi\lambda_{1l}})\left\{e^{-\lambda_{1l}\theta_i}(e^{2\pi\lambda_{2l}}-1)+e^{\lambda_{2l}\theta_i}(1-e^{-2\pi\lambda_{1l}})\right\}}p(l|\theta_i,\boldsymbol{\lambda}^{\boldsymbol{g}})=0,$$
$$\frac{\partial}{\partial\lambda_{2l}}T=\sum_{i=1}^{n}\left(\frac{1}{\lambda_{2l}}-\frac{1}{\lambda_{1l}+\lambda_{2l}}\right)p(l|\theta_i,\boldsymbol{\lambda}^{\boldsymbol{g}})$$
$$+\sum_{i=1}^{n}\frac{(1-e^{-2\pi\lambda_{1l}})\left\{(\theta_i-2\pi)e^{2\pi\lambda_{2l}+\lambda_{2l}\theta_i}-\theta_i e^{\lambda_{2l}\theta_i}\right\}}{(e^{2\pi\lambda_{2l}}-1)\left\{e^{-\lambda_{1l}\theta_i}(e^{2\pi\lambda_{2l}}-1)+e^{\lambda_{2l}\theta_i}(1-e^{-2\pi\lambda_{1l}})\right\}}p(l|\theta_i,\boldsymbol{\lambda}^{\boldsymbol{g}})=0$$

의 해를 Newton-Raphson 방법을 이용하여 구하거나, Nelder-Mead의 심플렉스 방법을 통해 구할 수 있다. 다만 Newton-Raphson 방법의 경우에는 T항에 대한 이차 편미분식이 요구되며, 이는 단일분포에서와 유사하게 계산되어지며 자세한 식은 생략하기로 한다.

이상의 방법을 통해 l번째 분포에 대한 개선된 추정량 $\boldsymbol{\lambda}^{g+1}=(\widehat{\lambda}_{1l},\widehat{\lambda}_{2l})$을 구할 수 있으며, 이로부터 모수 p_l의 추정치는, 식 (2.2)의 관계로부터,

$$\widehat{p_l}=\widehat{\lambda}_{2l}/(\widehat{\lambda}_{1l}+\widehat{\lambda}_{2l})$$

으로 구해지며, 식 (2.2)에 의해 λ_l과 κ_l의 추정치 역시

$$\widehat{\lambda}_l=\sqrt{\widehat{\lambda}_{1l}\widehat{\lambda}_{2l}},\quad \widehat{\kappa}_l=\sqrt{\frac{\widehat{\lambda}_{1l}}{\widehat{\lambda}_{2l}}}$$

으로 개선된다. 이 과정을 정지조건이 만족될 때까지 반복하여 최대가능도 추정치를 얻게 된다.

4. 모의실험

이 절에서는 겹친라플라스와 그의 혼합분포에 대한 최대가능도추정을 모의실험을 통해 확인하였다. 모의실험에는 $WL(2,0.5)$ 분포가 고려되었다. 이 분포는 식 (2.2)의 관계로부터 $(\lambda_1,\lambda_2)=(1,4)$이며, p는 0.8 $(=\lambda_2/(\lambda_1+\lambda_2))$인 분포에 해당된다. 아래의 표 4.1은 위 분포로부터 $n=100,500,1000$인 난수를 발생하고, Nelder-Mead의 심플렉스 방법으로 최대가능도추정치를 구한 것이다.

표 4.1 겹친라플라스 분포의 모수추정

모수		참값	$n=100$	$n=500$	$n=1000$
$WL^*(\lambda_1,\lambda_2)$	p	0.8	0.74	0.78	0.78
	λ_1	1	1.18	0.96	1.00
	λ_2	4	3.77	3.84	4.38
$WL(\lambda,\kappa)$	λ	2	2.11	1.92	2.09
	κ	0.5	0.56	0.50	0.48

아래의 그림 4.1은 $n=1000$인 경우의 추정결과를 그림으로 나타낸 것으로 위치모수는 편의상 $\eta=3$을 고려하였다.

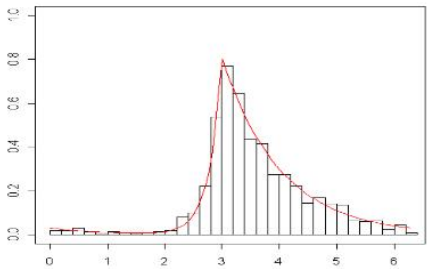

그림 4.1 겹친라플라스 분포의 적합결과 (n=1000)

다음으로 다봉형의 혼합분포에 대한 모의실험 결과는 아래의 표 4.2와 같다. 모의실험에는 다음의 모수값을 가지는 이봉형의 혼합분포를 고려하였다.

$$\alpha = 0.7,\ \boldsymbol{\lambda}_1{}' = (1,8),\ \boldsymbol{\lambda}_2{}' = (4,2),\ \boldsymbol{\eta}' = (3,5).$$

모의실험은 표본의 수가 $n = 50, 100, 500, 1000$인 경우에 대해 실시하였으며, 혼합분포의 최대가능도 추정을 위해 앞 절에서 소개된 EM 알고리즘을 사용하였으며, M-단계에서의 최대화 과정에는 Nelder-Mead의 심플렉스 방법이 사용되었다. 반복은 정지조건 $Q(\boldsymbol{\lambda}^{g+1};\boldsymbol{\lambda}^{g}) - Q(\boldsymbol{\lambda}^{g};\boldsymbol{\lambda}^{g-1}) \le 10^{-6}$을 만족할 때까지 수행하였다.

표 4.2 EM 알고리즘을 통한 겹친 라플라스 혼합모형의 모수추정

모 수	α	λ_{11}	λ_{12}	λ_{21}	λ_{22}	η_1	η_2
참 값	0.7	1	8	4	2	3	5
$n = 50$	0.71	1.03	7.38	3.78	1.38	2.98	4.95
$n = 100$	0.68	1.17	7.49	3.58	1.45	3.03	5.01
$n = 500$	0.66	1.06	7.34	3.73	2.01	3.04	5.01
$n = 1000$	0.70	0.95	7.50	3.70	1.79	3.01	5.03

아래의 그림 4.2는 $n = 1,000$인 경우에 대해 추정된 혼합분포를 나타낸 그림이다. 이 결과 EM 추정을 통한 적합모형이 자료를 매우 잘 적합함을 알 수 있다.

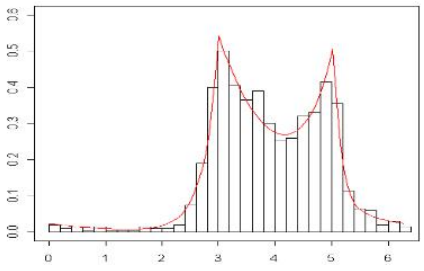

그림 4.2 겹친라플라스 혼합분포의 EM 추정 결과 (n=1000)

5. 맺음말

본 논문에서는 비대칭 첨봉형의 원형자료의 적합을 위해 겹친라플라스 분포에 대한 모수추정과정을 다루었다. 단봉형의 경우와 다봉형의 경우를 모두 고려하였으며, 특히 다봉형의 경우 EM 알고리즘을 통한 모수추정 과정을 구체적으로 제시하였다. 모의실험을 통해 제시된 모수추정 방법이 매우 유용함을 확인하였다. 본 연구에서 다룬 겹친라플라스 관련 분포는 비대칭이며, 특히 정점부근에서 첨봉형을 띠는 원형자료의 적합에 특히 유용함을 알 수 있다. 본 논문에서 다루어진 겹친 라플라스 모형의 적합 예로, 지방부 (또는 도시부) 도로의 특정지점에서 측정된 일일 시간대별 교통량 자료의 경우, 출 · 퇴근 시간대를 중심으로 교통량이 급증하며 (보통 이봉형의 모형이 적합), 그 패턴 역시 비대칭을 이루는 경우가 빈번하여 겹친라플라스 혼합모형이 유용하다 (Na와 Jang, 2010b).

참고문헌

Batschelet, E. (1981). *Circular statistics in biology*, Academic Press, London.

Dempster, Laird and Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, **39**, 1-38.

Jammalamadaka, S. R. and Kozubowski, T. J. (2003). A new family of circular models: The wrapped Laplace distributions. *Advances and Application in Statistics*, **3**, 77-103.

Jammalamadaka, S. R. and SenGupta, A. (2001). *Topics in circular statistics*, World Scientific.

Jang, Y. M., Yang, D. Y., Lee, J. Y. and Na, J. H. (2007). Modelling on multi-modal circular data using von Mises mixture distribution. *The Korean Communications in Statistics*, **14**, 517-530.

Mardia, K. V. (1972). *Statistics of directional data*, Academic Press, New York.

Mardia, K. V. and Jupp, P. E. (1999). *Directional statistics*, Wiley.

McLachlan, G. J. and Krishnan, T. (1997). *The EM algorithm and extensions*, Wiley.

Mooney, J. A., Helms, P. J. and Jolliffe, I. T. (2003). Fitting mixtures of von Mises distributions: A case study involving sudden infant death syndrome. *Computational Statistics and Data Analysis*, **41**, 505-513.

Na, J. H. and Jang, Y. M. (2010a). Modeling on asymmetric circular data using wrapped skew-normal mixture. *Journal of the Korean Data & Information Science Society*, **21**, 241-250.

Na, J. H. and Jang, Y. M. (2010b). Modeling on daily traffic volume of local state road using circular mixture distributions. *Unpublished Manuscript*.

Nelder, J. A. and Mead, R. (1965). A simplex method for function minimization. *Computer Journal*, **7**, 308-313.

Papakonstantinou, V. (1979). *Bietrge zur zirkulren statistik*, PhD Dissertation, University of Zurich, Switzerland.

Pewsey, A. (2000). The wrapped skew-normal distribution on the circle. *Communications in Statistics: Theory and Methods*, **29**, 2459-2472.

Pewsey, A. (2006). Modelling asymmetrically distributed circular data using the wrapped skew-normal distribution. *Environmental and Ecological Statistics*, **13**, 257-269.

Tanner, M. A. (1996). *Tools for statistical inference*, Springer.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985). *Statistical analysis of finite mixture distributions*, Wiley, Chichester.

Journal of the Korean Data & Information Science Society
2010, **21**(5), 863–871

한국데이터정보과학회지

Modeling sharply peaked asymmetric multi-modal circular data using wrapped Laplace mixture

Jong-Hwa Na[1] · Young-Mi Jang[2]

[1]Department of Information and Statistics, Chungbuk National University
[2]Korea Health and Welfare Information Service

Received 27 June 2010, revised 19 August 2010, accepted 28 August 2010

Abstract

Until now, many studies related circular data are carried out, but the focuses are mainly on mildly peaked symmetric or asymmetric cases. In this paper we studied a modeling process for sharply peaked asymmetric circular data. By using wrapped Laplace, which was firstly introduced by Jammalamadaka and Kozbowski (2003), and its mixture distributions, we considered the model fitting problem of multi-modal circular data as well as unimodal one. In particular we suggested EM algorithm to find ML estimates of the mixture of wrapped Laplace distributions. Simulation results showed that the suggested EM algorithm is very accurate and useful.

Keywords: Circular data, mixture model, wrapped Laplace.

[1] Corresponding author: Professor, Department of Information and Statistics, Chungbuk National University, Cheong-ju, Chungbuk 361-763, Korea. E-mail: cherin@cbu.ac.kr
[2] Doctor of philosophy, KHWIS, KukDong Bldg 21F, 60-1 Chungmu-ro 3ga, Jung-gu, Seoul 100-705, Korea.