The Korean Communications in Statistics **Vol. 14** No. 3, 2007, pp. 517–530

Modelling on Multi-modal Circular Data using von Mises Mixture Distribution*

Young-Mi Jang,[1] Dong-Yoon Yang,[2] Jin-Young Lee[3] and Jonghwa Na[4]

Abstract

We studied a modelling process for unimodal and multimodal circular data by using von Mises and its mixture distribution. In particular we suggested EM algorithm to find ML estimates of the mixture model. Simulation results showed the suggested methods are very accurate. Applications to two kinds of real data sets are also included.

Keywords: Circular data; EM algorithm; von Mises distribution; mixture model.

1. 서론

연속형의 다변량 자료는 크기 (magnitude)와 방향 (direction)의 정보를 함께 가지고 있다. 그러나 많은 응용분야에서 자료의 크기와는 관계없이 방향의 정보만이 주요 관심사가 되는 경우가 발생한다. 예를 들어, 기상학에서는 바람의 방향, 생물학에서는 동물의 이동방향, 지질학에서는 고자기 (paleomagnetic)의 방향을 연구하는 것 등이 이에 해당된다. 최근에는 물리학의 뇌파연구에서 phase synchronization 영역이나 microarray 자료에서의 유전자들 간의 연관방향성에 대한 연구에서 방향자료에 대한 분석 기법들이 시도되고 있다. 또한 의학에서 특정 질병에 따른 월별사망률이나 시간의 경과에 따른 경제시계열 자료 등도 시간의 주기를 방향의 자료로 전환하여 분석할 수 있다.

일반적으로 p-차원의 방향자료 (directional data)는 원점을 중심으로 크기가 1인 p-차원의 초평면 (hypersphere)상의 점으로 나타낼 수 있다. 이 때, $p = 2$인 경우의 방향자료를 순환자료 (circular data) 또는 각자료 (angular data)라고 하고, $p = 3$인 경우의 방향자료를 구형자료 (spherical data)라고 한다. 또한 순환자료에서 방향성을 무시할 경우 즉, $\theta(0 < \theta \leq 180°)$와 $\theta + 180°$를 동일하게 취급하는 경우의 자료를 축자료 (axial data) 또는 귀속자료 (orientation data)로 구분한다. 이차원의 방향자료 즉, 순환자료에

* This work was supported by the research grant of the Chungbuk National University in 2007.

1) Researcher, Ph. D., Korea Center for Disease Control & Prevention, Seoul 122-701, Korea.
2) Researcher, Ph. D., Korea Institute of Geoscience & Mineral Resources, Daejeon 305-350, Korea.
3) Researcher, Ph. D., Korea Institute of Geoscience & Mineral Resources, Daejeon 305-350, Korea.
4) Professor, Department of Information & Statistics, Chungbuk National University, Cheong-ju, Chungbuk 361–763, Korea.
Correspondence : cherin@cbu.ac.kr

대한 대표적인 문헌에는 Mardia (1972), Batschelet (1981)과 Fisher (1993), Mardia와 Jupp (1999)등을 들 수 있으며, 최근 Jammalamadaka와 SenGupta (2001)는 R언어와 연계한 여러 가지 순환자료의 분석법을 소개하고 있다. 특히 순환자료의 모형화 분석과 관련하여, 지금까지의 연구는 von Mises 분포를 중심으로 하는 단봉형의 대칭자료를 중심으로 진행되어 왔다. 최근에는 비대칭 순환분포 모형으로 Pewsey (2000, 2006)와 Jammalamadaka와 Kozubouski (2003)는 각각 겹친 왜-정규 (wrapped skew-normal)와 겹친 라플라스 (wrapped Laplace) 분포에 대한 연구를 수행한 바 있다.

본 논문에서는 실제의 자료분석 과정에서 매우 빈번하게 발생하는 이봉형 또는 다봉형의 순환자료에 대한 적합을 다루기로 한다. 혼합정규분포를 포함한 선형에서의 여러 혼합분포들에 대한 연구는 지금까지 매우 활발하게 진행되어 왔다 (Titterington 등, 1985). 그 가운데 Everitt (1984)과 Do와 McLachlan (1984)은 EM 알고리즘을 통해 일변량과 다변량 정규분포의 모수를 추정하는 방법에 대해 연구하였다. 선형의 혼합분포에 비해 순환분포의 혼합모형에 대한 모수추정은 분포함수의 복잡성으로 인한 계산상의 제약으로 인해 혼합-von Mises 분포를 중심으로 제한적으로 이루어져 왔다. 혼합-von Mises 분포의 최우추정에 대한 연구로는 Jones와 James (1972), Mardia와 Sutton (1975)의 연구를 들 수 있다. Jones와 James (1972)는 이봉형의 방향자료에 대해 최대경사법을 사용하여 모수를 추정하였으며, Mardia와 Sutton (1975)은 역시 이봉형 (bimodal)에 국한된 혼합-von Mises 분포에 대해 Newton 알고리즘을 사용한 모수추정을 제안하였다. 본 논문에서는 일반적인 다봉형의 순환자료에 대한 적합모형으로 혼합-von Mises 분포를 제안하고, 이에 대한 최우추정법으로 EM알고리즘을 제시하고자 한다. 2절에서는 EM 알고리즘을 통한 혼합-von Mises 분포에서의 모수추정 과정을 다루고 모의실험을 수행하였다. 3절에서는 두 가지 유형의 실제 자료분석을 수행하였다. 먼저 단봉형 순환 분포 모형의 적용 예로 물리학 분야에서 중요 관심사 중 하나인 위상차 자료의 분석을 다루었다. 다음으로 문화재 자료의 사면경사방향에 대한 모형화 과정에 본 연구에서 중요하게 취급되는 EM 알고리즘을 통한 혼합-von Mises 문포의 적합을 실시하였다.

2. 혼합-von Mises 분포의 모수추정

2.1. EM 알고리즘 소개

Dempster 등 (1977)에 의해 소개된 EM 알고리즘은 특정분포의 모수에 대한 최대우도 추정에 유용한 알고리즘이다. 특히, 관측 과정의 제약에 의해 결측 자료가 발생하는 경우, 또는 추정해야 할 모수의 수가 많아지거나 우도함수의 형태가 복잡한 경우에는 최대우도 추정에 효과적인 해결책이 된다. EM 알고리즘은 간단히 결측자료를 관측자료에 보완하여 최대화가 보다 쉬운 우도함수를 정의하고, E-단계 (Expectation step)와 M-단계 (Maximization step)로 불리는 두 단계를 반복적으로 수행하여 모수를 추정하는 수치해석적 방법이라 말할 수 있다. 여기서 결측자료는 관측되지 않은 잠재 확률변수 (hidden or latent random variable)의 값 (이를 잠재치 (latent values)라고 함.) 또는 모수값이 될 수 있으며, 관측된 원자료에 결측자료를 포함시킨 보완된 자료를 완비자료

(complete data) 또는 증대자료 (augmented data)라 부른다. 따라서 수치적인 난점을 해결하기 위해 새로 정의된 우도함수는 곧 완비자료의 우도함수가 되며 이를 완비우도함수 (complete likelihood function)라 한다.

EM 알고리즘을 좀 더 구체적으로 설명하면 다음과 같다. 관측자료 또는 불완비 자료를 X, 잠재자료를 Y라 하면 완비자료는 (X,Y)로 표현할 수 있으며, 추정해야 할 모수의 집합을 Θ 라고 하면 관측자료의 우도함수는 $f(X|\Theta)$로, 완비자료의 우도함수는 $f(X,Y|\Theta)$로 나타낼 수 있다. EM 알고리즘의 E-단계는 관측된 우도함수를 보완하기 위해 결측자료를 보완 (augment)하는 단계로 다음과 같이 정의되는 조건부 기대치인 Q함수를 계산한다.

$$Q(\Theta,\Theta^r) \equiv E_Y[\ln f(X,Y|\Theta)|X,\Theta^r]. \tag{2.1}$$

여기서, Θ^r은 현 단계에 주어진 모수추정치의 집합을 나타내며(Θ^0는 초기값임), 정의된 Q함수는 관측치와 모수추정치가 주어진 경우에 잠재변수가 추가된 완비자료의 로그-우도함수에 대한 조건부 기대값을 의미한다. M-단계에서는 $Q(\Theta;\Theta^r)$를 최대로 하는 Θ를 계산하여 새로운 모수 추정치 Θ^{r+1}로 가정한다. 이상의 E-단계와 M-단계의 반복을 다음의 정지조건 즉,

$$Q(\Theta^{r+1},\Theta^r) - Q(\Theta^r,\Theta^{r-1}) \le \varepsilon, \quad \forall \varepsilon > 0$$

을 만족할 때까지 반복 수행하여 최우추정치를 얻을 수 있다. EM 알고리즘에 대한 보다 자세한 내용은 Tanner (1996), McLachlan과 Krishnan (1997) 또는 Titterington 등 (1985)을 참고하기 바란다.

2.2. 혼합모형에서의 혼합비율 추정

이 절에서는 일반적인 혼합분포 모형으로부터의 EM 알고리즘을 통한 혼합비율에 대한 추정에 대해 알아본다. 이 방법은 다음 절에서 다루게 될 혼합-von Mises 분포의 경우에도 동일하게 적용된다.

먼저 $X = \{x_1,\ldots,x_n\}$을 다음의 혼합분포로부터 독립적으로 관측된 자료라고 하자 (다음 절에서 다루는 순환확률변수의 경우에는 x 대신 ϕ를 사용하고 있다).

$$f(x|\Theta) = \sum_{l=1}^{k} \alpha_l f_l(x|\theta_l). \tag{2.2}$$

여기서 Θ는 모수집합으로 $\Theta = \{\alpha_1,\ldots,\alpha_k,\theta_1,\ldots,\theta_k\}$이고, f_l은 모수가 $\theta_l (l=1,\ldots,k)$인 l번째 밀도함수를 나타내며, α_l은 각 단일분포의 혼합비율로 $\sum_{l=1}^{k}\alpha_l = 1$을 만족한다. 위 혼합 분포로 부터의 로그우도 함수는

$$\ln(L(\Theta|X)) = \ln\prod_{i=1}^{n} f(x_i|\Theta) = \sum_{i=1}^{n}\ln\left(\sum_{l=1}^{k}\alpha_l f_l(x_i|\theta_l)\right) \tag{2.3}$$

으로 표현되며, 위 식의 최대화를 통해 최우추정치를 구할 수 있다. 그러나 위 식은 로그함수의 인자가 합의 형태를 취하고 있어 직접 최대화를 수행하는데는 계산상 어려움

이 따른다. 따라서 본 연구에서는 최우추정의 방법으로 EM 알고리즘을 제안하고, 혼합-von Mises 분포에서의 적용과정을 구체적으로 제시한다.

먼저 혼합비율 α_l에 대한 추정과정은 다음과 같다. 관측자료 X를 불완비자료, 관측되지 않은 $Y = \{y_i\}_{i=1}^n$를 잠재자료라고 하자. 각 i에 대해 $y_i \in 1, \ldots, k$이고, i번째 표본이 h번째 분포로부터 생성되었다면 $y_i = h$라 한다. Y의 값을 이미 알고 있다면 관측자료의 로그우도 함수는

$$\ln(L(\Theta|X,Y)) = \ln(f(X,Y|\Theta)) = \sum_{i=1}^{n} \ln(\alpha_{y_i} f_{y_i}(x_i|\theta_{y_i})) \tag{2.4}$$

으로 표현될 수 있으며 다양한 방법을 통해 모수추정치를 산출할 수 있다. 그러나 대부분의 경우 잠재자료 Y의 값을 모르기 때문에 확률 벡터로 간주하여 $Y|(X,\Theta)$의 분포를 추정한 후 조건부 기대치 (식 (2.1))의 최대화를 통해 모수를 추정하게 된다.

다음으로 $Y|(X,\Theta)$의 분포는 $\Theta^g = (\alpha_1^g, \ldots, \alpha_k^g, \theta_1^g, \ldots, \theta_k^g)$가 주어질 때, 각 i와 h에 대해 다음과 같이 주어진다.

$$p(h|x_i, \Theta^g) = \frac{\alpha_h^g f_h(x_i|\theta_h^g)}{f(x_i|\Theta^g)} = \frac{\alpha_h^g f_h(x_i|\theta_h^g)}{\sum_{l=1}^k \alpha_l^g f_l(x_i|\theta_l^g)}.$$

따라서 완비우도함수에 대한 조건부 기대값인 Q함수는 다음의 식

$$\begin{aligned} Q(\Theta, \Theta^g) &= \sum_{i=1}^{n} E_{p(y_i|x_i,\Theta)}[\ln \alpha_{y_i} f_{y_i}(x_i|\theta_{y_i})] \\ &= \sum_{l=1}^{k}\sum_{i=1}^{n} \ln(\alpha_l f_l(x_i|\theta_l)) p(l|x_i, \Theta^g) \\ &= \sum_{l=1}^{k}\sum_{i=1}^{n} \ln(\alpha_l) p(l|x_i, \Theta^g) + \sum_{l=1}^{k}\sum_{i=1}^{n} \ln(f_l(x_i|\theta_l)) p(l|x_i, \Theta^g) \end{aligned} \tag{2.5}$$

으로 주어진다.

위의 식 (2.5)를 최대화하는 Θ^g의 값이 곧 혼합분포의 모수 $\Theta = \{\alpha_1, \ldots, \alpha_k\, \theta_1, \ldots, \theta_k\}$에 대한 개선된 추정치가 된다. 식 (2.5)에서 α_l은 첫 번째 항에만 관련되며, θ_l은 두 번째 항에만 관련되어 있으므로 각 항별 최대화를 통해 α_l과 θ_l에 대한 개선된 추정치를 각각 구할 수 있다.

먼저 $\sum_l^k \alpha_l$의 제약하에 혼합비율 α_l의 추정을 실시하자. 라그랑즈 승수 (Lagrange multiplier)를 λ로 하는 라그랑즈 목적함수를 α_l로 미분하여 0으로 놓은 다음 방정식

$$\frac{\partial}{\partial \alpha_l}\left[\sum_{l=1}^{k}\sum_{i=1}^{n} \ln(\alpha_l) p(l|x_i, \Theta^g) + \lambda\left(\sum_{l=1}^{k} \alpha_l - 1\right)\right] = 0$$

또는

$$\sum_{i=1}^{n} \frac{1}{\alpha_l} p(l|x_i, \Theta^g) + \lambda = 0 \tag{2.6}$$

의 해를 통해 식 (2.5)의 첫 번째 항을 최대화하는 모수추정치 $\hat{\alpha}_l$을 얻을 수 있다. 위의 식 (2.6)은

$$\alpha_l \lambda = -\sum_{i=1}^{n} p(l|x_i, \Theta^g)$$

으로 고쳐 쓸 수 있으며, 양변에 모든 l에 대해 합을 취하면

$$\sum_{l=1}^{k} \alpha_l \lambda = -\sum_{i=1}^{n}\sum_{l=1}^{k} p(l|x_i, \Theta^g)$$

이 된다. 위 식은 $\sum_{l=1}^{k} \alpha_l = 1$, $\sum_{l=1}^{k} p(l|x_i, \Theta^g) = 1$이므로 $\lambda = -n$이 되어 α_l에 대한 추정치는 다음과 같이 계산된다.

$$\hat{\alpha_l}^{new} = \frac{1}{n}\sum_{i=1}^{n} p(l|x_i, \Theta^g). \tag{2.7}$$

식 (2.7)의 혼합비율 α_l에 대한 추정치는 다음 절에서 다루게 되는 혼합 von-Mises 분포에도 동일하게 적용된다.

2.3. 혼합-von Mises 분포의 모수추정

이 절에서는 식 (2.5)의 두 번째 항으로부터, θ_l에 대한 추정치를 혼합-von Mises 분포하에서 구하기로 한다. 혼합-von Mises 분포의 확률밀도함수는 다음과 같이 $l \in 1, \ldots, k$에 대하여 모수가 $\theta_l = (\mu_l, \kappa_l)$ 인 von Mises 분포의 선형결합으로 정의할 수 있다.

$$f(\phi|\Theta) = \sum_{l=1}^{k} \alpha_l \left\{ \frac{1}{2\pi I_0(x_l)} e^{\kappa_l \cos(\phi - \mu_l)} \right\}.$$

단, $\Theta = \{\alpha_1, \ldots, \alpha_k, \theta_1, \ldots, \theta_k\}$, $\alpha \geq 0$, $\sum_{l=1}^{k} \alpha_l = 1$이다.

식 (2.5)에서 정의된 Q함수의 두 번째 항을 T라 두면

$$\begin{aligned} T &= \sum_{i=1}^{n}\sum_{l=1}^{k} (\ln f_l(\phi_i|\theta_l{}^g)) p(l|\phi_i, \Theta^g) \\ &= \sum_{i=1}^{n}\sum_{l=1}^{k} \{ -\ln 2\pi - \ln(I_0(k_l)) + k_l \cos(\phi_i - \mu_l) \} p(l|\phi_i, \Theta^g) \end{aligned}$$

으로 표현되며, 이를 μ_l과 κ_l로 미분하여 0으로 놓으면 최대우도 방정식은

$$\frac{\partial T}{\partial \mu_l} = \sum_{i=1}^{n} k_l \sin(\phi_i - \mu_l) p(l|\phi_i, \Theta^g) = 0, \tag{2.8}$$

$$\frac{\partial T}{\partial \kappa_l} = \sum_{i=1}^{n} \left\{ -\frac{I_0'(\kappa_l)}{I_0(\kappa_l)} + \cos(\phi_i - \mu_l) \right\} p(l|\phi_i, \Theta^g) = 0 \tag{2.9}$$

이 된다. 위 식 (2.8)은

$$\frac{\sin \mu_l}{\cos \mu_l} = \frac{\sum_{i=1}^{n} \sin \phi_i p(l|\phi_i, \Theta^g)}{\sum_{i=1}^{n} \cos \phi_i p(l|\phi_i, \Theta^g)}$$

으로 표현될 수 있다. 따라서 주어진 모수 Θ^g로부터 μ_l의 새로운 추정치는

$$\widehat{\mu_l}^{new} = \arctan^{\star}\left(\frac{\sum_{i=1}^{n} \sin \phi_i p(l|\phi_i, \Theta^g)}{\sum_{i=1}^{n} \cos \phi_i p(l|\phi_i, \Theta^g)}\right)$$

으로 주어지며, arctan*는 다음과 같이 유일하게 정의될 수 있다.

$$\arctan^{\star}(S/C) = \begin{cases} \arctan(S/C), & \text{if } C > 0,\ S \geq 0, \\ (\pi/2), & \text{if } C = 0,\ S > 0, \\ \arctan(S/C) + \pi, & \text{if } C < 0, \\ \arctan(S/C) + 2\pi, & \text{if } C \geq 0,\ S < 0, \\ \text{undefined}, & \text{if C=0, S=0.} \end{cases}$$

또한 κ_l의 새로운 추정치는 식 (2.9)로부터

$$\frac{I_1(\hat{\kappa}_l^{new})}{I_0(\hat{\kappa}_l^{new})} = \frac{\sum_{i=1}^{n} \cos(\phi_i - \hat{\mu}_l^{new}) p(l|\phi_i, \Theta^g)}{\sum_{i=1}^{n} p(l|\phi_i, \Theta^g)}$$

의 해로 정리된다.

2.4. 모의실험

앞 절에서 다봉형의 순환자료에 대해 혼합-von Mises 분포의 적합을 위한 EM 알고리즘을 제시하였다. 여기서는 EM 알고리즘의 효율을 확인하기 위한 모의실험을 수행한다. 모의실험에서 사용된 혼합 모집단의 분포는 이봉형의 경향이 비교적 뚜렷한 다음의 혼합-von Mises 분포를 가정하였다.

$$0.6 \cdot \nu M(45°, 2) + 0.4 \cdot \nu M(225°, 3).$$

위의 분포로부터 표본의 수가 $n = 50, 100, 500, 1000$인 난수를 발생하고, 각각의 경우에 대해 EM 알고리즘을 이용한 ML 추정을 실시하였다. 모든 모의실험은 R언어를 이용하였다.

모수 $\Theta = (\alpha_1, \alpha_2, \mu_1, \mu_2, \kappa_1, \kappa_2)$에 대해 초기값은 $\Theta^0 = (0.55, 0.45, 0°, 90°, 1.2, 2.5)$을 사용하였으며, 조건 $Q(\Theta^{r+1}; \Theta^r) - Q(\Theta^r; \Theta^{r-1}) \leq 10^{-6}$을 만족할 때까지 EM 알고리즘을 반복수행 하였다. EM 알고리즘을 통한 모수추정의 결과는 <표 2.1>과 같다. 이 결과에서 알 수 있듯이 표본의 수가 증가할수록 EM알고리즘을 통한 ML 추정치가 참값에 근사함을 알 수 있다.

표 2.1: EM 알고리즘을 통한 혼합-von Mises 분포의 모수추정

모수 추정 결과 (괄호안은 radian임)						
모수	α_1	α_1	μ_1	μ_2	κ_1	κ_2
참 값	0.6	0.4	45°(0.79)	225°(3.93)	2	3
$n = 50$	0.532	0.468	53.8°(0.94)	225.2°(3.93)	1.875	2.263
$n = 100$	0.631	0.369	44.6°(0.78)	219.2°(3.82)	1.674	3.180
$n = 500$	0.591	0.409	44.8°(0.78)	224.8°(3.92)	1.878	3.394
$n = 1000$	0.626	0.374	44.5°(0.78)	224.6°(3.92)	1.861	3.141

3. 실증분석

3.1. 위상차 자료의 모형화 분석

동일한 주파수를 가지는 두 신호간의 위상차 (phase difference)는 한 신호주기의 시작점에서 다른 신호에 앞서거나 뒤처지는 값으로 생각할 수 있다. 예를 들어, y축이 진폭이고 x축을 시간으로 하는 *sine*-곡선 형태의 파동그래프를 생각하자. 만약 두 신호 A와 B가 영(0)에서 출발하여, +, 0, − 의 값에서 다시 0의 값으로 돌아오는데 걸린 시간이 동일하다면, 두 신호는 동일한 주파수를 가지며 이 경우 두 신호 간에는 위상차가 없으며 "in phase" 상태라고 말한다. 역으로, 만약 두 신호가 동일한 주파수를 가지지만 (0의 값에서 다시 0으로 돌아오는 x축 상의 거리가 동일), 한 신호가 좀 더 빨리 혹은 늦게 시작되는 경우 두 신호 간에는 시간차가 존재하며 "out of phase" 상태라고 한다. 위상차는 0°에서 360°사이의 값 또는 라디안 (radians)으로 표현된다. 만약 그 차이가 180°이면 두 신호는 "in antiphase" 상태 즉, 크기는 동일하나 방향이 반대인 상태라고 하며 이 경우 두 신호 값의 합은 0으로 일정하다. 만약 위상차가 90°이면 두 신호는 "in quadrature" 상태라고 한다. 다음의 미분방정식의 형태로 표현되는 위상모형을 생각하자.

$$\frac{d\phi_1}{dt} = \omega_1 + a\sin\phi_1 + \gamma\sin(\phi_2 - \phi_1) + \xi_1,$$
$$\frac{d\phi_2}{dt} = \omega_2 + a\sin\phi_2 + \gamma\sin(\phi_1 - \phi_2) + \xi_2.$$

위 모형에서 ϕ_1, ϕ_2는 두 진동자의 위상을 나타내며, ω_1, ω_2는 두 진동자의 각 주기를 나타내며, γ는 진동자 간의 coupling의 수준을 의미한다. 또한 ξ_1, ξ_2는 각각 $N(0, \sigma^2)$을 따르는 백색잡음을 의미한다. 아래의 <그림 3.1>은 위의 모형에서 $\omega_1 = 1.0, \omega_2 = 0.4, a = 0.2, \sigma = 0.01$ 그리고 $\gamma = 0.3$인 경우에 대해 두 위상 ϕ_1과 ϕ_2를 각각 시간(t)의 경과에 따라 10,000개를 생성하고 처음 100초까지의 자료를 그림으로 나타낸 결과이다. <그림 3.1>에서 두 번째와 세 번째 그림은 본 연구의 주된 관심사인 두 진동자 간의 위상차 $\phi_1 - \phi_2$와 $\phi_1 - \phi_2 (\mathrm{mod}\ 2\pi)$ 를 그린 것이다.

전체 10,000개의 위상차 자료를 $(0, 2\pi)$구간 상에서 Rose Diagram으로 표시하면 <그림 3.2>와 같다.

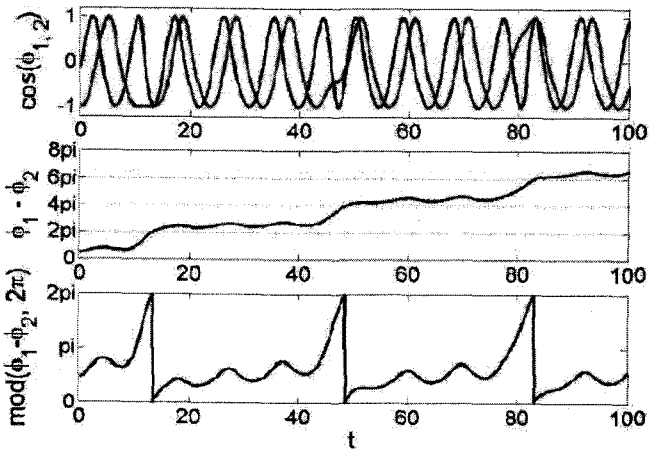

그림 3.1: 위상 모형으로부터 생성된 위상자료의 일부분

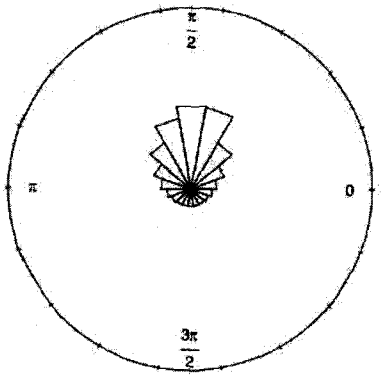

그림 3.2: 위상차 자료에 대한 Rose Diagram

이제 위의 위상차 자료에 대해 통계적 모형을 적합하기로 한다. 본 연구에서는 대표적인 대칭 단봉형의 순환모형인 von Mises와 Wrapped Normal 및 Wrapped Cauchy 분포를 적합하기로 한다. 모수들에 대한 최대우도 추정치와 적합결과는 <표 3.1>, <그림 3.3>과 같다. 이 결과에서 알 수 있듯이 비교적 대칭성이 뚜렷한 순환자료에 대해서 von Mises 분포 등이 효과적임을 알 수 있다. 각 순환분포의 ML 추정에 대해서는 Jammalamadaka와 SenGupta (2001)를 참고하기 바란다.

3.2. 가마터 자료의 모형화 분석

특정 문화유적의 입지 변인들에 대한 통계분석은 특정 문화재에 대한 이해를 제고

표 3.1: 위상차 자료에 대한 각 분포의 ML 추정 결과

순환 분포 모형	모수추정치 (괄호안은 radian 임)		
	$\widehat{\alpha}$	$\widehat{\rho}$	$\widehat{\kappa}$
von Mises	90.07°(1.572)	–	1.996
Wrapped Normal	90.70°(1.583)	0.652	–
Wrapped Cauchy	87.55°(1.528)	0.653	–

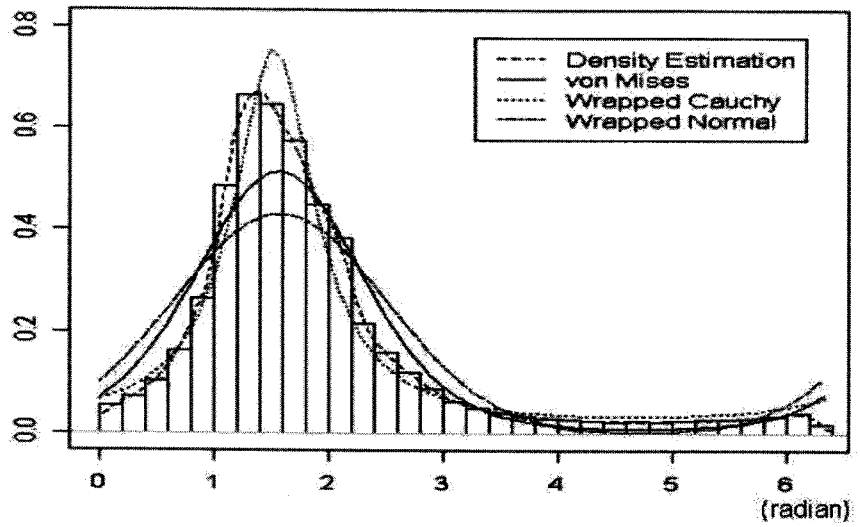

그림 3.3: 위상차 자료에 대해 적합된 순환분포모형

하는 것은 물론 문화재의 발굴 및 보전에도 크게 기여할 수 있다. 본 논문에서는 중요한 문화유적 가운데 하나인 가마터 유적에 대한 분석을 실시하기로 한다. 가마터 유적의 입지조건에는 형성된 가마터의 고도, 경사, 수계로부터의 거리, 도로와의 거리 등 많은 자연적 변인을 고려할 수 있다. 본 논문에서는 가마터 유적의 중요한 입지조건 가운데 하나인 가마터의 사면경사방향에 대한 분석을 실시하기로 한다. 조사된 가마터의 사면경사방향은 동쪽(E)을 기준방향인 0°로 하여, 시계방향으로 남쪽(S)은 90°, 서쪽(W)은 180°, 북쪽(N)은 270°로 측정되었다.

분석에 사용될 사면의 경사방향 자료는 지금까지 남한의 1,280개 지역에서 조사된 총 183,406개의 자료 (자료제공: 한국지질자원연구원)이다. 이 자료를 Rose Diagram과 히스토그램으로 요약한 결과는 <그림 3.4>와 같다.

위의 전체 가마터 자료로부터 많은 양의 자료를 확보하고 있는 두 종류의 가마터 즉, 청자가마터 ($n_1 = 21,516$)와 백자가마터 ($n_2 = 69,278$)에 대한 Rose Diagram과 히스토

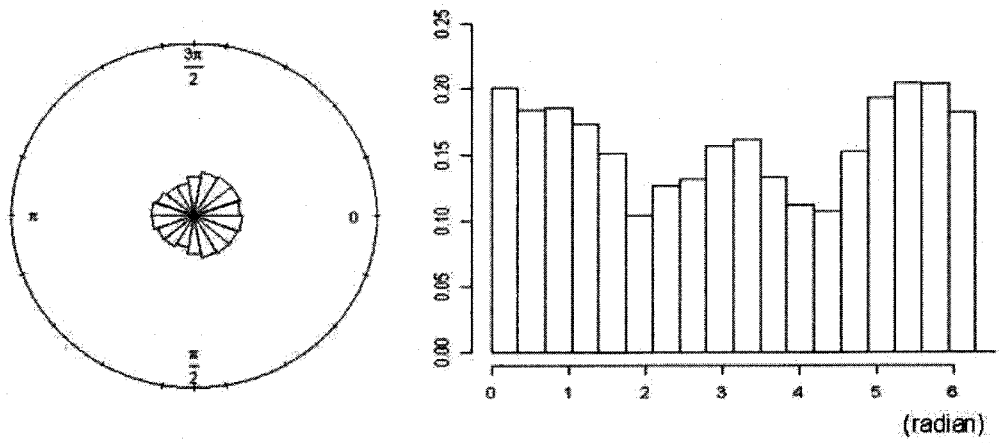

그림 3.4: 전체가마의 사면 경사방향에 대한 Rose Diagram과 히스토그램

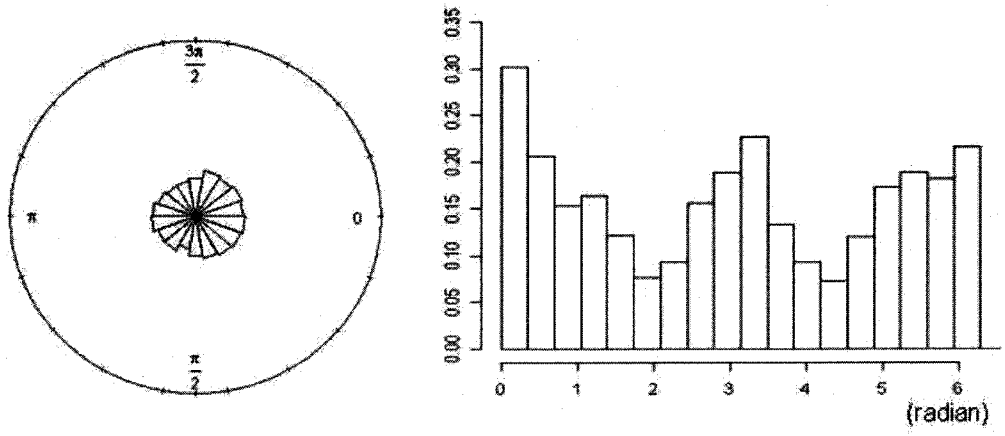

그림 3.5: 청자가마터의 사면 경사방향에 대한 Rose Diagram과 히스토그램

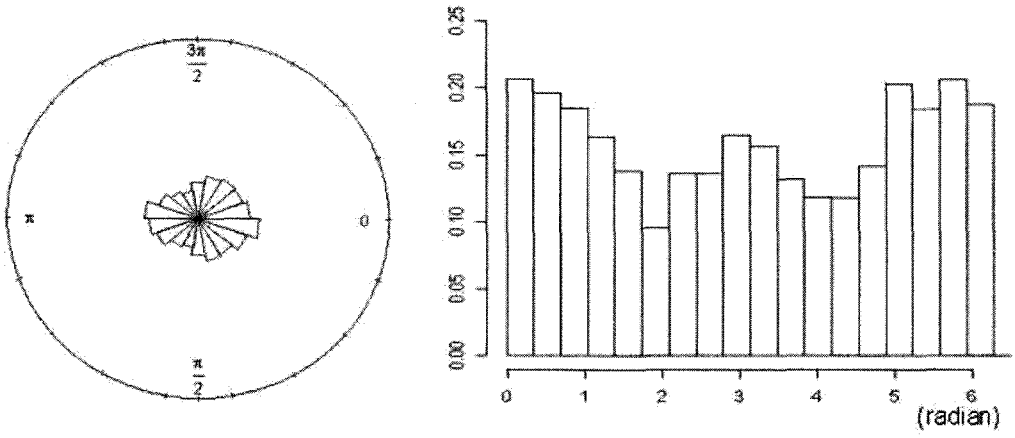

그림 3.6: 백자가마터의 사면 경사방향에 대한 Rose Diagram과 히스토그램

그램의 요약은 각각 <그림 3.5>와 <그림 3.6>과 같다.

위의 그림으로부터 전체가마, 청자가마 및 백자가마의 자료가 모두 뚜렷한 이봉형의 경향을 나타내고 있음을 알 수 있다. 따라서 경사방향 자료에 대한 적합모형으로 혼합-von Mises 분포를 제안하고, 2장에서 소개한 EM 알고리즘을 사용하여 모수추정을 실시하였다. 혼합-von Mises 분포모형의 ML 추정 결과를 요약하면 다음의 <표 3.2> 와 같다.

표 3.2: EM 알고리즘을 통한 혼합-von Mises 분포의 모수추정

가마터	표본수	모수추정 결과 (괄호안은 radian 임)				
		$\widehat{\alpha}_1$	$\widehat{\mu}_1$	$\widehat{\mu}_2$	$\widehat{\kappa}_1$	$\widehat{\kappa}_2$
전체가마	183,406	0.661	357.7°(6.24)	179.7°(3.14)	1.157	1.734
청자가마	21,516	0.808	359.9°(6.28)	179.4°(3.13)	0.784	5.341
백자가마	69,278	0.546	351.1°(6.13)	170.7°(2.98)	1.055	0.672

위의 적합결과를 그림으로 나타내면 아래의 <그림 3.7>~<그림 3.9>와 같다. 이 결과는 지금까지 가마터 유적의 자연 입지환경에 대한 분석이 주로 연속 및 범주형의 변인들에 대해 수행되어 온 점에 비추어 볼 때, 방향 (사면의 경사방향) 변인에 대한 통계적 모형을 제시한 것에 의미를 둘 수 있다. 특히 순환분포모형을 통한 방향변인에 대한 모형화 과정은 가마터 유적의 경우처럼 방향의 변인이 중요하게 취급되는 기타의 문화재 자료에도 매우 유용하게 사용될 것으로 기대된다.

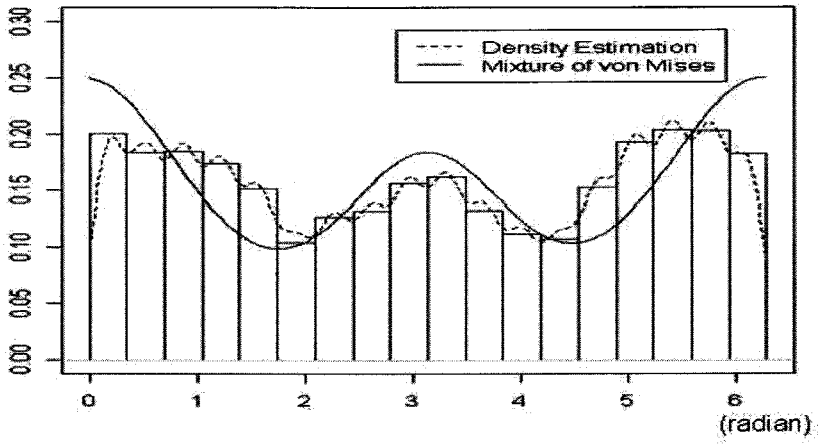

그림 3.7: 전체 가마터에 대한 혼합-von Mises 분포의 적합결과

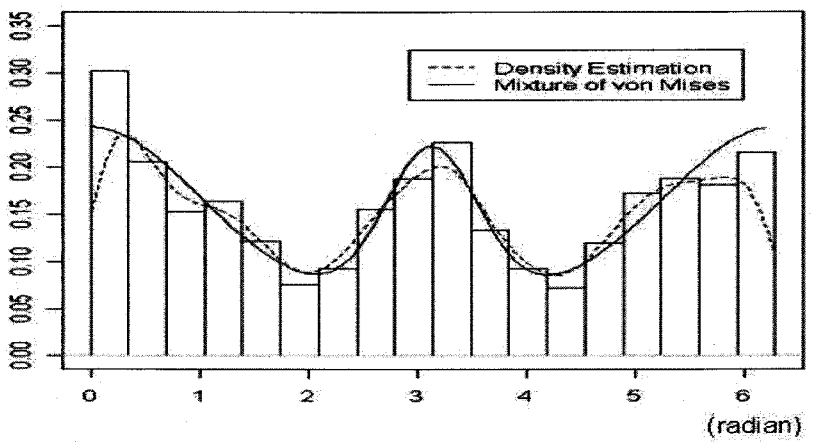

그림 3.8: 청자가마터에 대한 혼합-von Mises 분포의 적합결과

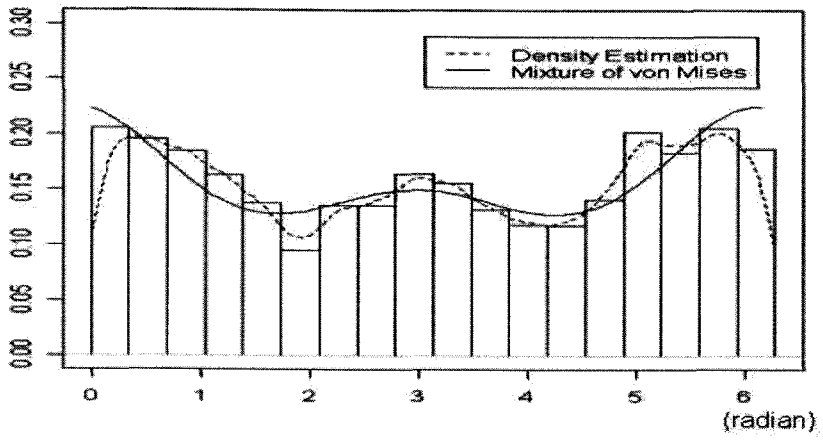

그림 3.9: 백자가마터에 대한 혼합-von Mises 분포의 적합결과

4. 맺음말

일반적으로 순환자료에 대한 분석법은 선형의 경우와는 매우 다르다. 본 논문에서는 다봉형의 순환자료에 대한 적합모형으로 혼합-von Mises 분포를 제안하고, 이에 대한 모수추정법으로 EM 알고리즘을 제안하였다. 또한 모의실험 및 실제의 자료분석을 통해 그 효용성을 확인하였다. 본 논문에서 사용된 혼합-von Mises 분포의 모수추정 알고리즘은 보다 다양한 대칭의 혼합-순환분포는 물론, 최근 많은 연구가 이루어지고 있는 겹친 왜-정규분포를 포함한 비대칭 순환분포의 혼합모형에도 적용할 수 있다.

참고문헌

Batschelet, E. (1981). *Circular Statistics in Biology*, Academic Press, London.

Do, K. and McLanchlan, G. J. (1984). Estimation of mixing proportions : a case study. *Applied Statistics*, **33**, 134–140.

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.

Everitt, B. S. (1984). Maximum likelihood estimation of the parameters in a mixture of two univariate normal distributions; a comparison of different algorithms. *The Statistician*, **33**, 205–215.

Fisher, N. I. (1993). *Statistical Analysis of Circular Data.* Cambridge University Press, New York.

Jammalamadaka, S. R. and SenGupta, A. (2001). *Topics in Circular Statistics.* World Scientific, NJ.

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.

Jones, M. C. and James, W. R. (1972). Analysis of bimodal orientation data. *Mathematical Geology*, **1**, 129–135.

Mardia, K. V. (1972). *Statistics of Directional Data.* Academic Press, New York.

Mardia, K. V. and Jupp, P. E. (1999). *Directional Statistics.* John Wiley & Sons, New York.

Mardia, K. V. and Sutton, T. W. (1975). On the modes of a mixture of two von Mises distributions. *Biometrika*, **62**, 699–701.

McLachlan, G. J. and Krishnan, T. (1997). *The EM Algorithm and Extensions.* 2nd ed., John Wiley & Sons, New York.

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. *Enviromental and Ecological Statistics*, **13**, 257–269.

Tanner, M. A. (1996). *Tools for Statistical Inference: : Methods for the Exploration of Posterior Distributions and Likelihood Functions.* Springer-Verlag, New York.

Titterington, D. M., Smith, A. F. M. and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions.* John Wiley & Sons, Chichester.

[Received July 2007, Accepted August 2007]