Journal of the Korean Data & Information Science Society
2010, **21**(2), 241–250

한국데이터정보과학회지

겹친왜정규혼합분포를 이용한 비대칭 원형자료의 모형화[†]

나종화[1], 장영미[2]

[1]충북대학교 정보통계학과 · [2]한국보건복지정보개발원

접수 2010년 1월 14일, 수정 2010년 3월 10일, 게재확정 2010년 3월 19일

요 약

원형자료에 대한 모형화 분석은 주로 von Mises 분포를 비롯한 대칭형의 경우를 중심으로 많은 연구가 이루어져 왔다. 최근 선형자료의 분석에서 다양한 비대칭의 자료에 적합한 왜정규분포의 활용에 대한 연구가 활발히 수행되고 있다. 본 논문에서는 Pewsey (2000a)에 의해 처음 소개된 겹친왜정규분포를 이용한 비대칭의 원형자료에 대한 적합을 다루었다. 특히 비대칭 다봉형 원형자료의 적합을 위해 겹친왜정규혼합분포를 제안하고, EM 알고리즘을 통한 모수추정 과정을 제시하였다. 모의실험을 통해 EM 알고리즘을 통한 모수추정의 정확성을 확인하고, 실제 지방국도의 일일교통량 자료의 모형화 분석에 적용하였다.

주요용어: 겹친왜정규분포, 원형자료, 혼합분포.

1. 서론

원형자료 (circular data)는 2차원의 방향자료 또는 크기 (magnitude)가 무시된 연속형 자료로 정의될 수 있다. 원형자료는 다양한 방법 - 각 (degree), 이차원의 단위벡터 또는 단위원주상의 점 - 으로 표현된다. 또한 원형자료는 영방향 (또는 기준 방향)과 회전방법 (시계 또는 반시계 방향)에 따라 특정 방향의 값을 여러 개의 다른 값으로 측정될 수도 있다. 따라서 원형자료에 대한 분석은 일반적인 연속형 자료에 적용되는 선형분석 기법을 사용할 수 없으며, 특별히 고안된 통계적 방법을 사용하여야 한다.

원형자료는 주로 생물학, 지질학, 기상학 등의 영역에서 중요하게 다루어져 왔으나 최근에는 의학, 경제학, 심리학 등의 영역에서도 많은 응용이 되고 있다. 원형자료의 통계적 분석과 관련된 문헌에는 Mardia (1972), Batschelet (1981), Fisher (1993), Mardia와 Jupp (1999) 등이 있으며, 최근에는 R을 이용한 분석을 다룬 Jammalamadaka와 SenGupta (2001)가 소개되었다. 원형자료의 모형화 분석과 관련된 지금까지의 연구들은 주로 단봉형의 경우로 von Mises 분포를 비롯한 대칭형의 원형분포를 중심으로 진행되어 왔다. 최근 Jang 등 (2007)은 von Mises 혼합분포를 이용한 대칭형의 다봉형 원형자료에 대한 적합을 수행하였다. 그러나 대부분의 실제 자료는 비대칭의 경우로 주어지며, 이에 대한 연구도 꾸준히 진행되어왔다. 대칭형의 원형분포로는 겹친정규 (wrapped normal), 겹친코쉬 (wrapped Cauchy), Cardioid 및 원형균일 (circular uniform) 분포 등이 있으며, 비대칭의 분포로는 Papakonstantinou (1979), Batschelet (1981) 및 겹친 α -stable 분포 등이 있으나 몇몇 대칭형의 분포 외에는 그 활용성에 대한 연구가 미비하다고 말할 수 있다.

[†] 이 논문은 2009년도 충북대학교 학술연구지원사업의 연구비 지원에 의하여 연구되었음.
[1] 교신저자: (361-763) 충북 청주시 흥덕구 성봉로 410, 충북대학교 정보통계학과, 교수.
E-mail: cherin@chungbuk.ac.kr
[2] (431-080) 경기도 안양시 호계동 903-2 대교빌딩 4층, 한국보건복지정보개발원, 박사.

본 논문에서는 최근 Pewsey (2000a)가 제안한 겹친왜정규분포를 이용한 원형자료의 모형화를 다루기로 한다. 겹친왜정규분포는 선형의 왜정규분포로부터 겹침 (wrapping)의 원리를 통해 유도되는 분포로, 대칭 및 비대칭의 원형자료를 모두 적합할 수 있는 활용성이 큰 분포로 알려져 있다. 본 논문에서는 단봉형의 경우 뿐 아니라 다봉형의 원형자료에 대한 적합을 위해 겹친왜정규혼합분포를 소개하고, EM 알고리즘을 통한 모수추정 방법을 제시하였다. 2절에서는 겹친왜정규분포와 모수추정을 소개하고, 3절에서는 비대칭의 다봉형 원형자료에 대한 적합 모형으로 겹친왜정규혼합분포를 제안하고, 이 분포의 모수추정 방법으로 EM 알고리즘을 제시하였다. 4절에서는 모의실험을 통해 EM 알고리즘을 통한 모수추정의 정확도를 확인하고, 실제자료에 대한 분석을 수행하였다. 5절은 결론으로 구성되었다.

2. 겹친왜정규분포와 모수추정

2.1. 겹친왜정규분포

Azzalini (1985)가 제안한 선형의 왜정규분포는 다음과 같이 정의된다.

$$f(x;\lambda) = 2\phi(x)\Phi(\lambda x), \quad -\infty < x < \infty, \quad -\infty < \lambda < \infty. \tag{2.1}$$

여기서 λ는 왜도모수를 나타내며, $\phi(\cdot)$와 $\Phi(\cdot)$는 각각 $N(0,1)$의 밀도함수와 분포함수를 나타낸다. 위 분포는 $X \sim SN(\lambda)$으로 나타내며, $\lambda = 0$인 경우 표준정규분포가 되며, $\lambda \to \infty$일 때 $X = |Z|, Z \sim N(0,1)$과 동일한 특징을 가진다. 왜정규분포와 관련된 그 외의 성질들은 Azzalini (1985), Azzalini와 Capitanio (1999)를 참고하기 바란다.

왜정규분포는 위치-척도변환 $Y_D = \xi + \eta X$을 통해 다음과 같이 일반화 된다.

$$f(y;\xi,\eta,\lambda) = \frac{2}{\eta}\phi\left(\frac{y-\xi}{\eta}\right)\Phi\left\{\lambda\left(\frac{y-\xi}{\eta}\right)\right\}, -\infty < y < \infty, -\infty < \xi < \infty, \eta > 0. \tag{2.2}$$

위 분포를 $Y_D \sim SN_D(\xi,\eta,\lambda)$으로 표기하고, (ξ,η,λ)를 직접모수 (direct parameters)라 부르기로 한다.

선형의 왜정규분포로부터 모듈러변환 $\Theta_{D=}Y_D(mod 2\pi)$을 통해 밀도함수가 다음과 같이 주어지는 겹친왜정규분포가 유도된다.

$$f(\theta;\xi,\eta\lambda) = \frac{2}{\eta}\sum_{r=-\infty}^{\infty}\phi\left(\frac{\theta+2\pi r-\xi}{\eta}\right)\Phi\left\{\lambda\left(\frac{\theta+2\pi r-\xi}{\eta}\right)\right\}, \quad 0 \le \theta < 2\pi \tag{2.3}$$

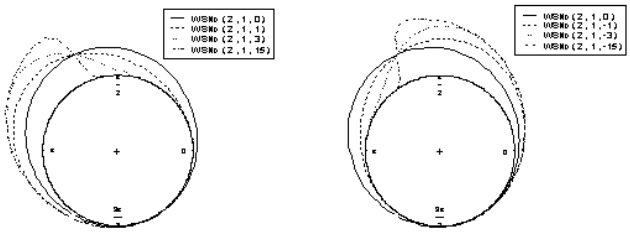
위 분포를 편의상 $\Theta_D \sim WSN_D(\xi,\eta,\lambda)$으로 표기한다.

아래의 그림 2.1은 모수의 변화에 따른 겹친왜정규분포의 형태를 나타낸다. 특히 왜도모수 λ의 값에 따라 치우침의 형태가 달라짐을 알 수 있다.

2.2. 모수추정

직접모수를 가지는 왜정규분포에 대한 ML (Maximum Likelihood) 추정에는, 정규분포 ($\lambda = 0$)의 경우, 모수잉여 (parameter redundancy)에 따른 Fisher 정보행렬의 비정칙 (singular)의 문제로 인해 가능도방정식의 해가 유일하지 않게 되는 문제가 발생한다 (Azzalini, 1985; Catchpole과 Morgan, 1997; Pewsey, 2000b). 이 문제에 대한 해결책으로 Azzalini (1985)는 다음의 중심화변환

$$Y_C = \mu + \sigma\left(\frac{X - E(X)}{\sqrt{Var(X)}}\right)$$



(a) $\lambda > 0$일 때 (b) $\lambda < 0$일 때

그림 2.1 $WSN_D(\xi, \eta, \lambda)$ 분포의 모양($\xi = 2, \eta = 1$)

을 통해 중심화모수 (centered parameters)가 (μ, σ, γ_1)인 왜정규분포를 새로이 정의하고, 이 분포를 $Y_C \sim SN_C(\mu, \sigma, \gamma_1)$으로 나타내었다. 여기서 μ와 σ는 각각 확률변수 Y_C의 평균과 표준편차가 되며, γ_1은 X의 왜도를 나타내는 계수로 $-0.99527 < \gamma_1 < 0.99527$을 만족한다. 직접모수 (ξ, η, λ)와 중심화모수 (μ, σ, γ)사이에는

$$\xi = \mu - c\gamma_1^{1/3}\sigma, \quad \eta = \sigma\sqrt{1 + c^2\gamma_1^{2/3}}, \quad \lambda = \frac{c\gamma_1^{1/3}}{\sqrt{b^2 + c^2(b^2-1)\gamma_1^{2/3}}} \tag{2.4}$$

또는

$$\mu = \xi + \lambda\eta\sqrt{\frac{b^2}{1+\lambda^2}}, \quad \sigma = \eta\sqrt{\frac{1-\lambda^2(b^2-1)}{1+\lambda^2}}, \quad \gamma_1 = \left\{\frac{\lambda^2 b^2}{c^2(1-\lambda^2(b^2-1))}\right\}^{3/2}$$

의 관계가 성립된다. 단, $b = \sqrt{2/\pi}$이고 $c = \{2/(4-\pi)\}^1/3$이다. 식 (2.2)에 식 (2.4)의 관계를 적용하면 Y_C의 확률밀도함수는

$$f(y;\mu,\sigma,\gamma_1) = \frac{2}{\sigma\sqrt{1+c^2\gamma_1^{2/3}}}\phi\left[\frac{1}{\sqrt{1+c^2\gamma_1^{2/3}}}\left\{\left(\frac{y-\mu}{\sigma}\right) + c\gamma_1^{1/3}\right\}\right]$$

$$\times\Phi\left[\frac{c\gamma_1^{1/3}}{\sqrt{\left\{b^2 + c^2(b^2-1)\gamma_1^{2/3}\right\}(1+c^2\gamma_1^{2/3})}}\left\{\left(\frac{y-\mu}{\sigma}\right) + c\gamma_1^{1/3}\right\}\right]$$

이 된다.

직접모수의 경우와 마찬가지로 변환 $\Theta_C = Y_C(mod 2\pi)$을 통해 다음과 같이 중심화모수 (μ, σ, γ_1)을 가지는 겹친왜정규분포를 정의할 수 있다.

$$f(\theta;\mu,\sigma,\gamma_1) = \frac{2}{\sigma\sqrt{1+c^2\gamma_1^{2/3}}} \sum_{r=-\infty}^{\infty} \phi\left\{\frac{1}{\sqrt{1+c^2\gamma_1^{2/3}}}\left(\frac{\theta+2\pi r-\mu}{\sigma}+c\gamma_1^{1/3}\right)\right\} \tag{2.5}$$
$$\times\Phi\left[\frac{c\gamma_1^{1/3}}{\sqrt{\left\{b^2+c^2(b^2-1)\gamma_1^{2/3}\right\}(1+c^2\gamma_1^{2/3})}}\left(\frac{\theta+2\pi r-\mu}{\sigma}+c\gamma_1^{1/3}\right)\right],\ \theta\in[0,2\pi).$$

위 분포를 편의상 $\Theta_C \sim WSN_C(\mu,\sigma,\gamma_1)$으로 표기하기로 한다.

겹친왜정규분포는 σ가 0에 가까울수록 $(\eta \to 0)$ 점분포에 근사하며 σ가 커질수록 $(\eta \to \infty)$ 순환 균등분포에 근사한다. 또한 $\gamma_1 = 0$ $(\lambda = 0)$이면 겹친정규분포와 동일한 분포가 되며, $\gamma_1 \to \pm 0.99527(\lambda \to \pm\infty)$이면 겹친반정규 (wrapped half-normal) 또는 겹친음반정규 (wrapped negative half-normal)분포로 수렴한다.

겹친왜정규분포의 모수에 대한 ML추정을 위해 $WSN_C(\mu,\sigma,\gamma_1)$으로부터 추출한 확률표본을 $\boldsymbol{\theta} = (\theta_1,\cdots,\theta_n)$이라 하자. 식 (2.5)로 부터 $\boldsymbol{\theta}$에 대한 로그우도함수는

$$l(\mu,\sigma,\gamma_1) = nln2 - n\ln\sigma - \frac{n}{2}\ln(1+c^2\gamma_1^{2/3}) \tag{2.6}$$
$$+\sum_{i=1}^{n}\ln\sum_{r=-\infty}^{\infty}\phi\left\{\frac{1}{\sqrt{1+c^2\gamma_1^{2/3}}}\left(\frac{\theta_i+2\pi r-\mu}{\sigma}+c\gamma_1^{1/3}\right)\right\}$$
$$\times\Phi\left[\frac{c\gamma_1^{1/3}}{\sqrt{\left\{b^2+c^2(b^2-1)\gamma_1^{2/3}\right\}(1+c^2\gamma_1^{2/3})}}\left(\frac{\theta_i+2\pi r-\mu}{\sigma}+c\gamma_1^{1/3}\right)\right]$$

으로 주어진다. 식 (2.6)은 무한합으로 정의되나 대부분의 경우 $r\in(-3,3)$이면 로그우도 함수를 정의하기에 충분하며 산포도가 큰 자료에 대해서는 $r\in(-7,7)$이면 적당하다. 본 연구에서는 r의 값을 한정하지 않고 항의 합이 수렴하는 값을 구하여 로그우도함수에 적용하였다.

중심화모수를 가지는 식 (2.6)은 수치적 방법을 통해 최대화 시킬 수 있다. 본 논문의 모의실험에서는 Nelder와 Mead (1965)의 심플렉스 방법을 사용하였다. 수치적 방법에 의해 구해진 중심화모수에 대한 최대가능도추정치를 식 (2.4)의 관계를 이용하여 직접모수에 대한 추정치를 구할 수 있다.

3. 겹친왜정규혼합분포의 모수추정

3.1. 겹친왜정규혼합분포

분포 f_l $(l=1,2,\cdots,k)$을 모수가 $\boldsymbol{\lambda}_l$인 겹친왜정규분포라 할 때, 겹친왜정규혼합분포는 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 혼합분포로 부터의 표본을 $\boldsymbol{\theta} = \{\theta_1, \theta_2, \cdots, \theta_n\}$이라 하자. 이때, 로그우도함수는

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

으로 주어진다. 위 식은 로그함수의 인자가 합의 형태를 취하고 있어, 미분을 통한 최대화에 어려움이 따른다. 이에 대한 해결책으로 본 논문에서는 EM 알고리즘을 통해 수치적인 방법으로 식 (3.2)를 최대화하기로 한다. EM 알고리즘을 통한 ML추정의 과정은 다음과 같다.

먼저 관측자료 $\boldsymbol{\theta}$를 불완비자료로, 관측되지 않은 $\boldsymbol{\psi} = \{\psi_1, \psi_2, \cdots, \psi_n\}$를 잠재자료라고 하자. 각 i에 대해 $\psi_i \in 1, 2, \cdots, k$ 이고, i 번째 관측치 θ_i가 l번째 분포 f_l로부터 생성되었다면 $\psi_i = l$로 주어진다. 잠재자료 $\boldsymbol{\psi}$의 을 아는 경우 완비자료 $(\boldsymbol{\theta}, \boldsymbol{\psi})$의 로그우도함수는

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

으로 표현되며, 위 식에 대한 최대화는 쉽게 수행될 수 있게 된다. 그러나 일반적으로 각 자료 θ에 대응하는 잠재 자료 ψ의 값을 모르기 때문에 $\Psi|(\theta, \boldsymbol{\lambda})$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 완비가능도수에 대한 조건부 기댓값은 다음과 같이 주어진다.

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

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

3.2. EM알고리즘을 통한 모수추정

3.2.1. 혼합비율의 추정

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

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

으로 추정된다.

3.2.2. 중심화모수의 추정

혼합분포에서 l번째 분포의 중심화모수를 $\boldsymbol{\lambda}_l = (\mu_l, \sigma_l, \gamma_{1l})$이라 할 때, 식 (3.4)의 두 번째 항은

$$T = \sum_{i=1}^{n}\sum_{l=1}^{k}\left\{ ln2 - \ln\sigma_l - \frac{1}{2}\ln(1 + c^2\gamma_{1l}^{2/3})\right\} p(l|\theta_i, \boldsymbol{\lambda}^{\boldsymbol{g}})$$

$$+ \sum_{i=1}^{n}\sum_{l=1}^{k}\ln\sum_{r=-\infty}^{\infty}\phi\left\{\frac{1}{\sqrt{1+c^2\gamma_{1l}^{2/3}}}\left(\frac{\theta_i + 2\pi r - \mu_l}{\sigma_l} + c\gamma_{1l}^{1/3}\right)\right\}$$

$$\times\Phi\left[\frac{c\gamma_{1l}^{1/3}}{\sqrt{\left\{b^2 + c^2(b^2-1)\gamma_{1l}^{2/3}\right\}(1+c^2\gamma_{1l}^{2/3})}}\left(\frac{\theta_i + 2\pi r - \mu_l}{\sigma_l} + c\gamma_{1l}^{1/3}\right)\right] p(l|\theta_i, \boldsymbol{\lambda}^{\boldsymbol{g}})$$

으로 표현된다. 위 식을 최대로 하는 $\boldsymbol{\lambda}_l$은 심플렉스 방법을 통해 구할 수 있으며, E-단계와 M-단계를 반복하여 추정치를 개선해 나간다. 직접모수에 대한 추정은 중심화모수의 추정 결과를 식 (2.4)에 대입하여 구할 수 있다.

4. 모의실험과 실증분석

4.1. 모의실험

이 절에서는 앞 절에 다룬 겹친왜정규혼합분포에 대한 ML 추정에 대한 모의실험을 수행한다. 모의실험은 편의상 $k = 2$인 경우를 가정하고, 두 모집단의 직접모수 $(\xi,\ \eta,\ \lambda)$가 각각 $(1, 2, 10)$와 $(5, 1, 2)$이며, $\alpha = 0.7$인 경우를 고려하였다. 즉, 다음의 모수

$$\alpha = 0.7,\ \boldsymbol{\xi}' = (1, 5),\ \boldsymbol{\eta}' = (2, 1),\ \boldsymbol{\lambda}' = (10, 2)$$

를 가지는 혼합분포로부터 표본의 크기가 $n = 50, 100, 500, 1000$인 경우에 대해 모의실험을 수행한 결과는 표 4.1과 같다. 모수 $\boldsymbol{\lambda}$에 대해 초기값으로는 $\boldsymbol{\lambda}^0 = (\alpha, \xi_1, \xi_2, \eta_1, \eta_2, \lambda_1, \lambda_2) = (0.65, 0.8, 4.7, 2.1, 0.8, 9.7, 2.2)$을 사용하였으며, Q함수의 최대값의 차이가 $Q(\boldsymbol{\lambda}^g + 1; \boldsymbol{\lambda}^g) - Q(\boldsymbol{\lambda}^g; \boldsymbol{\lambda}^g - 1) \leq 10^-6$을 만족할 때까지 E-단계와 M-단계를 반복 수행하였다.

아래의 그림 4.1은 표 4.1의 적합결과를 그림으로 나타낸 것이다. 전체적으로 ML 추정의 결과가 자료를 잘 적합함을 알 수 있다.

표 4.1 겹친왜정규혼합분포의 ML 추정 결과

모 수	α	ξ_1	ξ_2	η_1	η_2	λ_1	λ_2
참 값	0.7	1	5	2	1	10	2
$n = 50$	0.68	1.01	4.66	1.65	0.89	10.17	1.84
$n = 100$	0.80	0.85	5.36	1.81	0.64	10.22	0.33
$n = 500$	0.72	1.01	5.25	1.94	0.64	10.16	0.74
$n = 1000$	0.70	0.95	5.00	2.05	0.88	10.08	2.27

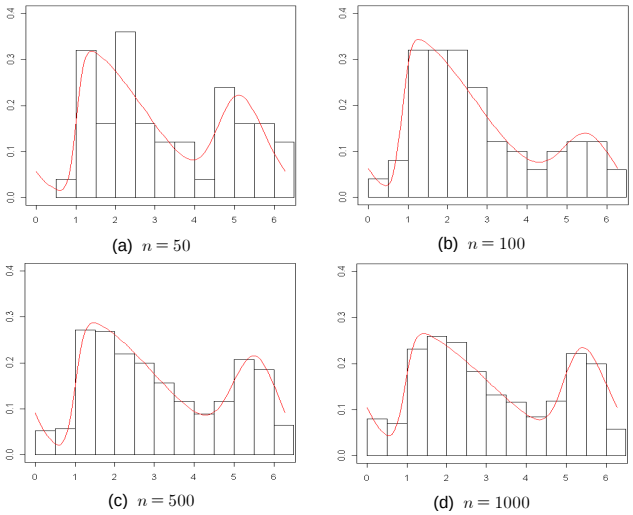


그림 4.1 겹친왜정규혼합분포의 ML 추정 결과

4.2. 실증분석

분석에 사용된 자료는 2004년 1월 1일부터 12월 31일까지의 335번 지방국도 (경기도 용인시 삼계면 오포리간 4차선 도로)의 시간대별 교통량 자료이다. 하루 24시간을 $0 \sim 1, 1 \sim 2, \cdots, 23 \sim 24$의 시구간별로 측정된 교통량 (빈도) 자료는, $0 \sim 2\pi$를 24개 구간으로 나눈 원형자료로 변환될 수 있다. 시간대별 교통량 자료의 특성은 보통 요일별로 다른 패턴을 보이며, 본 논문에서는 특히 토요일 자료만을 분석의 대상으로 삼았다. 아래의 그림 4.2는 분석 자료를 Rose Diagram과 히스토그램으로 요약한 것이다. Rose Diagram의 좌측상단의 수치는 335도로의 연평균 일일교통량 (AADT; Annual Average Daily Traffic)을 나타내며, 실선과 점선은 각각 교통량의 최빈값 (시간)과 평균값을 나타낸다. 여기서 평균과 최빈값을 비롯한 원형자료에 대한 기술통계치에 대한 정의는 Jammalamadaka와 SenGupta (2001)를 참고하기 바란다.

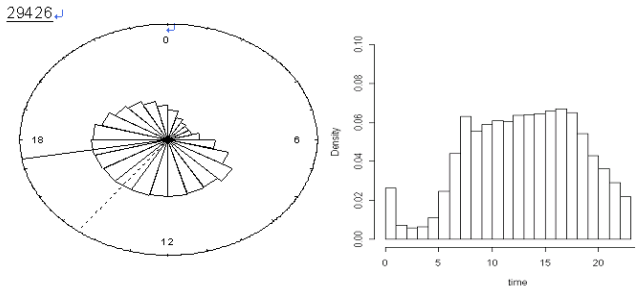


그림 4.2 시간대별 교통량 자료의 요약

아래의 표 4.2는 여러 가지 원형분포를 적합한 결과를 나타낸다. 로그우도값의 비교를 통해 단봉형에 비해 이봉형의 모형이 자료를 보다 잘 적합함을 알 수 있으며, 그 가운데 겹친왜정규혼합분포의 적합결과가 가장 우수한 것으로 나타났다.

표 4.2 원형분포 적합 결과

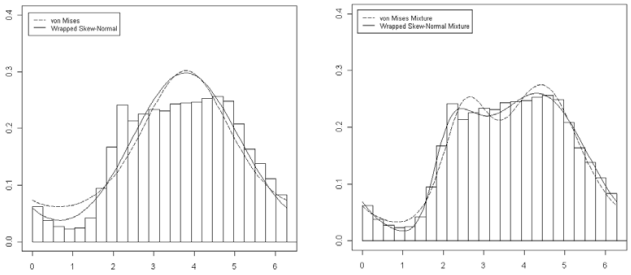
(a) 단일모형 적합 결과

적합모형	모수			로그우도
von Mises	μ	κ		-49999.06
	3.80	0.79		
Wrapped Skew-Normal	ξ	η	λ	-49474.90
	3.15	1.50	0.68	

(b) 혼합모형 적합 결과

적합모형	모수							로그우도
vMM	α	μ_1	μ_2	κ_1	κ_2			-49014.44
	0.27	2.54	4.43	3.45	1.17			
WSNM	α	ξ_1	ξ_2	η_1	η_2	λ_1	λ_2	-48877.73
	0.25	1.82	4.41	0.89	1.17	2.46	-0.03	

아래의 그림 4.3은 표 4.2의 모형 적합결과를 그림으로 나타낸 것이다. 이 그림에서 단봉형의 경우에는 적합이 잘 이루어지지 않는 반면, 이봉형의 경우는 von Mises 혼합분포와 겹친왜정규분포는 모두 만족스러운 적합 결과를 보여주고 있다. 특히 겹친왜정규혼합분포는 von Mises 혼합분포에서의 적합결여 부분을 잘 개선하고 있음을 보여준다. 일반적으로 겹친왜정규혼합분포는 비대칭의 경우를 포함하므로 von Mises 혼합분포보다 다양한 패턴의 자료를 적합할 수 있다.



(a) 단일모형 적합 결과 (b) 혼합모형 적합 결과

그림 4.3 교통량 자료에 대한 원형분포 적합 결과

5. 결론

본 논문에서는 최근 많은 연구가 이루어지고 있는 선형에서의 왜정규분포로부터 유도되는 겹친왜정규분포를 이용한 원형자료에 대한 분석을 다루었다. 비대칭 다봉형의 원형자료에 대한 적합모형으로 겹친왜정규혼합분포를 제안하고, EM 알고리즘을 이용한 모수추정법을 제시하였다. 이 결과는 기존의 Jang 등 (2007)에 의한 대칭형의 원형자료에 대한 분석과 함께 비대칭 자료에 대한 적합에 유용할 것으로 생각된다. 특히 겹친왜정규분포는 대칭의 분포뿐 아니라 다양한 왜도를 가지는 자료 형태를 적합할 수 있어, 기존의 von Mises 분포를 중심으로 한 대칭형의 모형에 비해 활용성이 뛰어날 것으로 기대된다.

참고문헌

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, **12**, 171-178.

Azzalini, A. and Capitanio, A. (1999). Statistical applications of the multivariate skew-normal distribution. *Journal of the Royal Statistical Society*, Series B, **91**, 579-602.

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

Catchpole, E. A. and Morgan, B. J. T. (1997). Detecting parameter redundancy. *Biometrika*, **84**, 187-196.

Fisher, N. I. (1993). *Statistical analysis of circular data*, Cambridge University Press.

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.

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

Papakonstantinou, V. (1979). *Bietrage zur zirkularen statistik*, Ph. D. Dissertation, University of Zürich, Switzerland.

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

Pewsey, A. (2000b). Problems of inference for Azzalini's skew-normal distribution. *Journal of Applied Statistics*, **27**, 859-870.

Journal of the Korean Data & Information Science Society
2010, **21**(2), 241–250

한국데이터정보과학회지

Modeling on asymmetric circular data using wrapped skew-normal mixture

[1]Jong-Hwa Na[2] · Young-Mi Jang[3]

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

Received 14 January 2010, revised 10 March 2010, accepted 19 March 2010

Abstract

Over the past few decades, several studies have been made on the modeling of circular data. But these studies focused mainly on the symmetrical cases including von Mises distribution. Recently, many studies with skew-normal distribution have been conducted in the linear case. In this paper, we dealt the problem of fitting of non-symmetrical circular data with wrapped skew-normal distribution which can be derived by using the principle of wrapping. Wrapped skew-normal distribution is very flexible to asymmetical data as well as to symmetrical data. Multi-modal data are also fitted by using the mixture of wrapped skew-normal distributions. To estimate the parameters of mixture, we suggested the EM algorithm. Finally we verified the accuracy of the suggested algorithm through simulation studies. Application with real data is also considered.

Keywords: Circular data, mixture distribution, wrapped skew-normal.

[1] This work was supported by the research grant of the Chungbuk National University in 2009.
[2] Corresponding author: Professor, Department of Information and Statistics, Chung buk National University, Cheong-ju, Chungbuk 361-763, Korea. Email: cherin@chungbuk.ac.kr
[3] Doctor of philosophy, KHWIS, DaeKyo Bldg 4F, 903-2 Hogye-dong, Dongan-gu, Anyang-si, Gyeonggi-do, Korea.