한국통계학회논문집
2011, 18권, 3호, 391–401

DOI: 10.5351/CKSS.2011.18.3.391

합성된 평균과 분산을 가진 군집 식별

김승구[1,a]

[a]상지대학교 컴퓨터데이터정보학과

요 약

본 논문에서는 자료 내의 군집 중에 '父 군집'과 '母 군집'이라 부르는 두 군집 사이에, 합성된 평균과 분산을 가지는 '합성군집' 즉 '자식 군집'이라 부르는 한 군집이 있을 경우에 주목하여, 그들의 관계를 평균과 분산에 관해 모형화하고 각각의 군집을 식별하는 방법을 제공하였다. 관측치는 정규혼합모형을 따른다고 가정하고, EM 알고리즘을 통해 모형 추정을 시도하였다. 추정 과정에 여러 난제가 있었으나, 근사적 방법으로 비교적 잘 극복할 수 있었다. 그리고 수치실험을 통해 제안 방법은 성공적으로 주어진 세 군집 즉 '군집族'을 식별할 수 있음을 보였다.

주요용어: 합성 군집, 파생강도, 군집 족, 정규혼합모형, EM 알고리즘.

1. 서론

군집분석 후 분류된 군집들을 면밀히 살펴보면, 서로 멀지 않은 두 군집 사이 중간에 끼여 있는 한 군집을 종종 발견하게 된다. 저자는 이 세 군집이 모수에 관해 합성된 어떤 관계로 이루어진 것일 수도 있다는 생각을 해 보았다. 특히, 만약 그 중간 군집의 산포정도가 두 군집 중 최소한 한 개의 군집의 그것보다 작아 보이면서 가까이 있는 쪽과 유사해 보인다면 다음과 같은 관계로서 세 군집을 설명해 볼 수 있을 것이다. 즉, 두 군집 G_1, G_2는 각각 정규분포 $N(\mu_1, \sigma_1^2)$과 $N(\mu_2, \sigma_2^2)$으로부터 생성되었고, 두 군집 사이에 낀 중간 군집 G_3는 평균과 분산이 각각 $\mu_3 = \alpha\mu_1 + (1-\alpha)\mu_2$ 및 $\sigma_3^2 = \alpha^2\sigma_1^2 + (1-\alpha)^2\sigma_2^2$인 $N(\mu_3, \sigma_3^2)$으로부터 생성되었다고 가정하는 것이다. 여기서 α는 $(0, 1)$ 사이의 값이다. 이로써 평균 μ_3는 μ_1과 μ_2 사이 값을 가지게 될 것이고, 분산 σ_3^2도 역시 $\max(\sigma_1^2, \sigma_2^2)$보다 작은 값을 가지게 될 것이며, α가 1에 가까울수록 G_1에 그리고 0에 가까울수록 G_2의 산포와 비슷할 것이다. 앞으로 편의를 위해 G_1과 G_2를 각각 '父 군집'과 '母 군집'이라 하고, G_3를 '자식 군집'이라 부르겠으며, 이 세 군집을 합쳐 '군집 族'이라 하겠다. 그리고 α를 '파생 강도'라 부를 것이다. 이를테면, $\alpha = 0.4$인 경우 자식 군집은 부 군집과 모 군집으로부터 각각 40%와 60%의 영향을 받아 (혹은 닮아서) 부모 군집의 평균과 분산에 관해 합성된 특성을 가지며 파생된 집단이라고 해석할 수 있을 것이다. 본 연구의 목적은 자료로부터 이와 같은 군집 족의 세 군집을 동시에 식별하고 파생강도를 추정하는 것이다. 여러 군집 중에서 군집 족을 파악하고, 이에 더하여 파생강도를 제공할 수 있다면 군집분석으로부터 얻을 수 있는 정보를 더욱 풍부하게 해 줄 것으로 기대한다.

정규혼합모형(normal mixture model)은 거의 모든 응용분야에서, 특히 마이크로어레이 유전자 발현 자료분석에서 군집분석의 표준적 도구로 사용되고 있다 (Wang과 Zhu, 2008; Ng 등, 2006; 김승구, 2007). 본 논문에서도 역시 정규혼합모형 위에 군집 족을 식별하기 위한 설계를 하게 될 것이다. 그리고 모형을 추정하기 위해 Dempster 등 (1977)의 EM(expectation- maximization) 알고리즘을 이용했다.

본 연구는 상지대학교 2010년도 연구비 지원에 의해 수행되었음.

[1] (110-302) 강원도 원주시 우산동 660, 상지대학교 컴퓨터데이터정보학과, 교수. E-mail: sgukim@sangji.ac.kr

저자는 본 논문의 연구 목적과 관련된 사전연구를 거의 찾을 수 없었다. 굳이 비슷한 경우를 들자면 Wang과 Cohen (2005) 및 Levin 등 (2008)의 영상 매팅(image matting) 연구 사례에서 찾을 수 있었는데, 이것은 영상의 전경영역(forward area) G_1과 후경영역(backward area) G_2 사이의 경계부분인 G_3에서 투명도 α를 화소별로 추정하는 것이 목적이다. 그러나 영상매팅에서는 영상이 오직 세 화소집단 G_1, G_2, G_3로 분할되어 있어야 하며, 각 집단은 사전에 구체적으로 파악되어 있어야 한다는 점이 본 연구의 목적과 차이가 있다.

다음 절에서는 우리의 군집 족을 식별하기 위한 정규혼합모형을 설명하며, 이 모형에 대한 諸 조건을 설명한다. 3절에서는 모형 추정을 위한 EM 알고리즘을 제공하며, 4절에서는 수치실험을 통해 제안된 방법의 유효성을 제공할 것이다. 5절에서는 결론을 정리하고, 제안된 방법의 한계점과 추가 연구되어야 할 점들을 토의한다.

2. 모형

독립적으로 관측된 n개의 자료 $y_1, \ldots, y_n$은 g개의 군집으로 분할될 수 있고, 그 중 3개가 부 군집, 모 군집 및 자식 군집이라 하자. 이를 모형으로 표현하기 위해

$$f(y_j;\theta) = \sum_{j=1}^{2} \pi_i \phi\left(y_j;\mu_i,\sigma_i^2\right) + \pi_3 \phi\left(y_j;\mu_3,\sigma_3^2,\alpha\right) + \sum_{i=4}^{g} \pi_i \phi\left(y_j;\mu_i,\sigma_i^2\right), \quad j = 1,\ldots,n \tag{2.1}$$

과 같은 g-성분 정규혼합모형으로 나타내자. 여기서 $\theta = \left\{\{\pi_i\},\{\mu_i\},\left\{\sigma_i^2\right\},\alpha\right\}$인 모수 집합이며, $\pi_i(>0)$는 혼합비율로서 $\pi_1 + \cdots + \pi_g = 1$을 만족한다. 그리고

$$\mu_3 = \alpha\mu_1 + (1-\alpha)\mu_2, \tag{2.2}$$

$$\sigma_3^2 = \alpha^2\sigma_1^2 + (1-\alpha)^2\sigma_2^2 \tag{2.3}$$

의 관계를 가진다. 단, $\alpha \in (0,1)$을 만족한다. 한편, 모형 (2.1)에서 부 군집, 모 군집 및 자식 군집에 대응한 혼합성분을 각각 제1, 제2, 제3 성분에 위치하게 했는데, 이렇게 설정해도 본 논문에서 일반성을 잃지 않는다. 따라서 부 군집, 모 군집, 자식 군집은 각각 제1, 제2, 제3 군집으로 식별하게 될 것이다.

2.1. 가정들

본 논문에서는 $\mu_1 \neq \mu_2$를 가정한다. 왜냐하면 $\mu_1 = \mu_2 \overset{\text{let}}{=} \mu$라면 조건 (2.2)로부터 파생 강도 α와 관계없이 $\mu_3 = \mu$가 되어 연구의 목적이 성립되지 않기 때문이다. 다시 말해서 우리의 문제는 μ_1과 μ_2 사이의 중간 군집이 존재해야 하는데, $\mu_1 - \mu_2 = 0$이라면 문제의 의미를 상실하게 된다. 아울러 우리는 $\alpha \neq 0$ 그리고 $\alpha \neq 1$을 가정하고 있다. 그 이유는 만약 $\alpha = 0$이면 조건 (2.2)–(2.3)으로부터 $\mu_3 = \mu_2$가 되고 $\sigma_3^2 = \sigma_2^2$이 되는데, 이 경우 식 (2.1)의 모형은 식별가능하지 않게 된다. 그리고 $\alpha = 1$인 경우도 마찬가지이다. 혼합모형의 식별가능성(identifiability)에 대한 자세한 내용은 Titterington 등 (1994)를 참조하기 바란다.

2.2. 성분의 개수 결정

본 논문에서 정규혼합모형의 성분의 개수는 BIC(Bayesian Information Criterion; Schwarz, 1978)

$$-2L\left(\hat{\theta}\right) + \nu(g)\log(n) \tag{2.4}$$

을 최소로 하는 g로서 결정하게 될 것이다. 여기서 $L(\hat{\theta})$은 $\hat{\theta}$에서 계산된 로그-우도이며 (식 (3.1)에서 정의될 것임), $\nu(g)$는 자유모수의 개수로서, 우리의 모형은 표준 정규혼합모형에 비해 (2.2)와 (2.3)의

두 제약이 있고, α가 추가되므로 표준 정규혼합모형의 자유모수의 개수에서 1를 빼야 할 것이다. 즉, $\nu(g) = 3g - 2$이다.

3. EM 알고리즘에 의한 모형 추정

조건 (2.2)–(2.3)과 $0 < \alpha < 1$의 제약 하에서, 모수 θ의 최우추정치를 구하기 위해 로그-우도

$$L(\theta) = \sum_{j=1}^{n} \log \sum_{i=1}^{g} \pi_i \phi\left(y_j; \mu_i, \sigma_i^2, \alpha\right) \tag{3.1}$$

를 직접 최대화하는 것은 쉽지 않다. 그래서 보통 관측치 y_j가 i번째 성분으로부터의 표본이면 $(\boldsymbol{z}_j)_i = z_{ij} = 1$ 그렇지 않으면 0을 나타내는 지시변수라 놓고, $\boldsymbol{z} = (\boldsymbol{z}_1^T, \ldots, \boldsymbol{z}_n^T)^T$를 결측 자료 그리고 $(\boldsymbol{y}^T, \boldsymbol{z}^T)^T$를 완비 자료로 설정하는데, 이때 EM 알고리즘에서는 반복 $(t+1)$에서 불완비 자료 $\boldsymbol{y}$에 대한 완비 자료의 로그-우도의 조건부 기대값

$$Q\left(\theta|\theta^{(t)}\right) = \sum_{j=1}^{n} \log \pi_i \sum_{i=1}^{g} \tau_{ij} \log \phi\left(y_j; \mu_i, \sigma_i^2, \alpha\right) \tag{3.2}$$

을 최대화 한다. 여기서

$$\begin{aligned} \tau_{ij} &= E\left(Z_{ij}|y_j, \theta^{(t)}\right) = \Pr_{\theta^{(t)}}\left\{Z_{ij} = 1\middle|y_j\right\} \\ &= \frac{\pi_i \phi\left(y_j; \mu_i, \sigma_i^2, \alpha\right)}{\sum_{h=1}^{g} \pi_h \phi\left(y_j; \mu_h, \sigma_h^2, \alpha\right)}, \quad i = 1, \ldots, g \end{aligned} \tag{3.3}$$

로서 관측치 y_j가 i번째 성분에 속하게 될 사후확률을 의미하게 된다. 이 최대화 과정을 충분히 반복하면 $\theta^{(t)}$는 (국소) 최우추정치에 도달하게 된다.

이제 주어진 목표는 조건 (2.2)–(2.3)과 $0 < \alpha < 1$의 제약 하에서 식 (3.3)를 최대화 하는 추정치 $\pi_i^{(t+1)}, \mu_i^{(t+1)}, \sigma_i^{(t+1)}$ 및 $\alpha^{(t+1)}$를 구하는 것인데, 우선 제약과 무관한 혼합비율 추정치는

$$\pi_i^{(t+1)} = \frac{\sum_{j=1}^{n} \tau_{ij}^{(t+1)}}{n}, \quad i = 1, \ldots, g \tag{3.4}$$

단,

$$\tau_{ij}^{(t+1)} = \frac{\pi_i^{(t)} \phi\left(y_j; \mu_i^{(t)}, \sigma_i^{2(t)}, \alpha^{(t)}\right)}{\sum_{h=1}^{g} \pi_h^{(t)} \phi\left(y_j; \mu_h^{(t)}, \sigma_h^{2(t)}, \alpha^{(t)}\right)}, \quad i = 1, \ldots, g;\ j = 1, \ldots, n \tag{3.5}$$

과 같이 얻을 수 있다 (McLachlan과 Peel, 2000). 그리고 나머지 추정치에 대해서는 아래에 소절로 나누어 차례로 제공할 것이다.

3.1. μ_i의 추정

잠시 표기의 간편함을 위해 반복 첨자 (t)는 생략하기로 하자. 제약 $\alpha\mu_1 + (1-\alpha)\mu_2 - \mu_3 = 0$ 하에서 목적함수 (2.2)의 μ_i에 관한 최대화는 라그랑지 배수(Lagrange's multiplier) λ에 대하여

$$Q_\lambda\left(\mu_1, \ldots, \mu_g, \lambda\middle|\theta^{(t)}\right) \propto -\frac{1}{2} \sum_{j=1}^{n} \sum_{i=1}^{g} \tau_{ij} \frac{(y_j - \mu_i)^2}{\sigma_i^2} + \lambda \sum_{i=1}^{3} v_i \mu_i, \tag{3.6}$$

을 (μ_i, λ)에 관하여 최대화 하는 문제가 된다. 단, $v_1 = \alpha$, $v_2 = 1-\alpha$ 및 $v_3 = -1$을 나타낸다. 우선 $\mu_4, \ldots, \mu_g$에 대해서는 제약이 없으므로 $\lambda = 0$으로 생각하면 된다. 이때 $0 = \partial Q_0/\partial \mu_i$으로부터

$$\mu_i = \frac{\sum_{j=1}^n \tau_{ij} y_j}{\tau_{i\cdot}} \overset{\text{let}}{=} \bar{y}_i, \quad i = 4, \ldots, g \tag{3.7}$$

을 얻을 수 있다. 단, $\tau_{i\cdot} = \sum_{j=1}^n \tau_{ij}$을 나타낸다.

이제 μ_1, μ_2, μ_3에 대하여, $0 = \partial Q_\lambda/\partial \mu_i \propto \sum_{j=1}^n \tau_{ij}(y_j - \mu_i) + \lambda v_i \sigma_i^2 = \sum_{j=1}^n \tau_{ij} y_j + \lambda v_i \sigma_i^2 - \mu_i \tau_{i\cdot}$으로부터

$$\mu_i = \frac{\sum_{j=1}^n \tau_{ij} y_j}{\tau_{i\cdot}} + \lambda v_i \left(\frac{\sigma_i^2}{\tau_{i\cdot}} \right) = \bar{y}_i + \lambda v_i \xi_i \tag{3.8}$$

의 先 결과를 얻을 수 있다. 여기서 $\xi_i = \sigma_i^2/\tau_{i\cdot}$를 나타낸다. 그리고 $0 = \partial Q_\lambda/\partial \lambda = \sum_{i=1}^3 v_i \mu_i = \sum_{i=1}^3 v_i \bar{y}_i + \lambda \sum_{i=1}^3 v_i^2 \xi_i$으로부터 $\lambda = -\sum_{i=1}^3 v_i \bar{y}_i / \sum_{i=1}^3 v_i^2 \xi_i$을 얻게 되는데, 이것을 (3.8)의 선 결과에 대입하면,

$$\mu_i = \bar{y}_i - v_i \xi_i \left(\frac{\sum_{h=1}^3 v_h \bar{y}_h}{\sum_{h=1}^3 v_h^2 \xi_h} \right), \quad i = 1, 2, 3 \tag{3.9}$$

을 얻게 된다. 이 결과로부터 (양변에 v_i를 곱하고 $\sum_{i=1}^3$을 취하면) 제약 조건 $v_1\mu_1 + v_2\mu_2 + v_3\mu_3 = 0$이 만족 됨을 확인 할 수 있다. 결국 식 (3.7)과 (3.9)로부터 $(t+1)$번째 반복에서

$$\mu_i^{(t+1)} = \bar{y}_i^{(t+1)} - v_i^{(t+1)} \xi_i^{(t+1)} \left(\frac{\sum_{h=1}^3 v_h^{(t+1)} \bar{y}_h^{(t+1)}}{\sum_{h=1}^3 \left(v_h^{(t+1)} \right)^2 \xi_h^{(t+1)}} \right), \quad i = 1, 2, 3 \tag{3.10}$$

$$= \bar{y}_i^{(t+1)}, \quad i = 4, \ldots, g \tag{3.11}$$

를 계산한다. 단, $\xi_i^{(t+1)} = \sigma_i^{2^{(t+1)}}/\tau_{i\cdot}^{(t+1)}$을 나타낸다.

3.2. σ_i^2의 추정

잠시 표기의 간편함을 위해 반복 첨자 (t)는 생략하기로 하자. 제약 $\alpha^2\sigma_1^2 + (1-\alpha)^2\sigma_2^2 - \sigma_3^2 = 0$ 하에서 목적함수 (2.2)의 σ_i^2에 관한 최대화는 라그랑지 배수 λ에 대하여

$$Q_\lambda \left(\sigma_1^2, \ldots, \sigma_g^2, \lambda \middle| \theta^{(t)} \right) \propto -\frac{1}{2} \sum_{j=1}^n \sum_{i=1}^g \tau_{ij} \log \sigma_i^2 - \frac{1}{2} \sum_{j=1}^n \sum_{i=1}^g \tau_{ij} \frac{(y_j - \mu_i)^2}{\sigma_i^2} + \lambda \sum_{i=1}^3 w_i \sigma_i^2, \tag{3.12}$$

을 (σ_i^2, λ)에 관하여 최대화 하는 문제가 된다. 여기서 $w_1 = \alpha^2$, $w_2 = (1-\alpha)^2$ 및 $w_3 = -1$을 나타낸다. 먼저 $\sigma_4^2, \ldots, \sigma_g^2$에 대해서는 제약이 없으므로 $\lambda = 0$으로 생각하고 $0 = \partial Q_0/\partial \sigma_i^2$으로부터

$$\sigma_i^2 = \frac{\sum_{j=1}^n \tau_{ij} (y_j - \mu_i)^2}{\tau_{i\cdot}} \overset{\text{let}}{=} S_i^2, \quad i = 4, \ldots, g \tag{3.13}$$

을 얻을 수 있다. 단, $\tau_{i\cdot} = \sum_{j=1}^n \tau_{ij}$을 나타낸다.

다음으로 $\sigma_1^2, \sigma_2^2, \sigma_3^2$에 대하여 우도방정식은

$$0 = \frac{\partial Q_\lambda}{\partial \sigma_i^2} = -\tau_{i\cdot}\sigma_i^2 + \sum_{j=1}^{n} \tau_{ij}(y_j - \mu_i)^2 + 2\lambda w_i \left(\sigma_i^2\right)^2$$
$$= \left(\frac{2\lambda w_i}{\tau_{i\cdot}}\right)\left(\sigma_i^2\right)^2 - \sigma_i^2 + S_i^2 \tag{3.14}$$

과 같이 σ_i^2의 2차 방정식이므로 근의 공식을 이용하여 두 해 $\sigma_i^2(\lambda) = \{1 \pm \sqrt{1 - 8\lambda w_i S_i^2/\tau_{i\cdot}}\}/\{4\lambda w_i/\tau_{i\cdot}\}$ 중 양의 해를 선택하여 구할 수 있을 것이다. 그러나 두 해 중에 양의 실수해를 명시적으로 결정한다는 것은 도대체 쉬워 보이지 않으며, 혹시 결정할 수 있다 하더라도 라그랑지 배수 λ가 분모와 분자에 서로 다른 차원으로 존재하고 있기 때문에 최종해는 표현 가능하게 풀리지 않을 것이 분명하다. 그래서 본 연구에서는 대안으로 발견적 방법으로서 근사적인 기술을 적용하고자 한다. 즉, Green (1990)의 OSL(one-step-late) 추정 방식과 유사하게 제약항의 (편)미분 항에 $(\sigma_i^2)^2$ 대신 (3.14)을 근사적으로 만족하는 이전 단계에서 양의 값을 가지는 어떤 $A^{(t)}$를 고려하자. 즉,

$$0 = \frac{\partial Q_\lambda}{\partial \sigma_i^2} \approx \left(\frac{2\lambda w_i}{\tau_{i\cdot}}\right) A^{(t)} - \sigma_i^2 + S_i^2 \tag{3.15}$$

로부터 선 결과를 얻는 것이다. 결국,

$$\sigma_i^2 = \lambda 2 w_i \eta_i^{(t)} + S_i^2 \tag{3.16}$$

를 얻게 된다. 단, $\eta_i^{(t)} = A^{(t)}/\tau_{i\cdot}$를 나타낸다. 그리고 $0 = \partial Q_\lambda/\partial\lambda = \sum_{i=1}^3 w_i\sigma_i^2 = 2\lambda \sum_{i=1}^3 w_i^2\eta_i^{(t)} + \sum_{i=1}^3 w_i S_i^2$으로부터

$$\lambda = -\frac{1}{2}\frac{\sum_{i=1}^3 w_i S_i^2}{\sum_{i=1}^3 w_i^2 \eta_i^{(t)}}$$

를 얻게 되고, 이것을 식 (3.16)에 대입하면,

$$\sigma_i^2 = S_i^2 - w_i\eta_i^{(t)}\left(\frac{\sum_{h=1}^3 w_h S_h^2}{\sum_{h=1}^3 w_h^2\eta_h^{(t)}}\right), \quad i = 1, 2, 3 \tag{3.17}$$

와 같이 결정된다. 이 결과로부터 (양변에 w_i를 곱하고 $\sum_{i=1}^3$을 취하면) 제약 조건 $w_1\sigma_1^2 + w_2\sigma_2^2 + w_3\sigma_3^2 = 0$이 만족 됨을 확인 할 수 있다. 그리고 EM 알고리즘이 수렴한다면, $\sigma_i^2 \to S_i^2$이 되고 $w_1S_1^2 + w_2S_2^2 + w_3S_3^2 = 0$을 만족하게 됨으로써, $\lambda = 0$이 실현되며 결국 식 (3.17)에서 $\sigma_i^2 = S_i^2 > 0$이 되어 비음의 조건을 만족하게 된다.

이제 문제는 $A^{(t)}$의 선택이다. 후보로서 $\sigma_i^{2(t)}$이나 $\mu_i^{4(t)}$, 심지어 $|\mu_i^{(t)}|$도 후보가 될 수 있다. 이 모든 후보들에 대해 EM 알고리즘은 원하는 값은 아닐지라도 어떤 한 점으로 수렴하였다. 직관적으로는 $(\sigma_i^{2(t)})^2$이 가장 유력한 후보였지만 예상외로 $A^{(t)} = S_i^{2(t)}$일 때 가장 정확한 값으로 수렴하였다.

이상의 결과를 정리하면 다음과 같다. 식 (3.13)과 (3.17)으로부터, $(t+1)$번째 반복에서

$$\sigma_i^{2^{(t+1)}} = S_i^{2^{(t+1)}} - w_i^{(t+1)}\eta_i^{(t)}\left(\frac{\sum_{h=1}^3 w_h^{(t+1)} S_h^{2^{(t+1)}}}{\sum_{h=1}^3 w_h^{2^{(t+1)}}\eta_h^{(t)}}\right), \quad i = 1, 2, 3 \tag{3.18}$$

$$= S_i^{2^{(t+1)}}, \quad i = 4, \ldots, g \tag{3.19}$$

를 계산한다. 단, $\eta_i^{(t)} = S_i^{2^{(t)}}/\tau_{i\cdot}^{(t+1)}$을 나타낸다.

3.3. α의 추정

잠시 표기의 간편함을 위해 반복 첨자 (t)는 생략하기로 하자. 이제 α에 대한 로그-우도는

$$Q\left(\alpha\middle|\theta^{(t)}\right) \propto -\frac{1}{2}\sum_{j=1}^{n}\tau_{3j}\log\left\{\sigma_3^2(\alpha)\right\} - \frac{1}{2}\frac{\sum_{j=1}^{n}\tau_{3j}\left\{y_j-\mu_3(\alpha)\right\}^2}{\sigma_3^2(\alpha)} \tag{3.20}$$

과 같이 주어진다. 단, $\mu_3(\alpha)=\alpha\mu_1+(1-\alpha)\mu_2$ 및 $\sigma_3^2(\alpha)=\alpha^2\sigma_1^2+(1-\alpha)^2\sigma_2^2$을 나타낸다. 이제 우리는 $0<\alpha<1$의 제약하에서 식 (3.20)를 α에 관하여 최대화 해야한다. 이때 다소 지루한 과정을 거쳐

$$\begin{aligned} 0 = \frac{\partial Q\left(\alpha\middle|\theta^{(t)}\right)}{\partial\alpha} &\propto -(\mu_1-\mu_2)\left\{\bar{y}_3-\mu_3(\alpha)\right\} + \left\{\alpha\left(\sigma_1^2+\sigma_2^2\right)-\sigma_2^2\right\}\left[1-\frac{S_3^2(\alpha)}{\sigma_3^2(\alpha)}\right] \\ &\overset{\text{let}}{=} g(\alpha) \end{aligned} \tag{3.21}$$

과 같이 편미분을 얻을 수 있다. 그런데 $g(\alpha)$는 α의 비선형 함수로서 명시적 해를 제공하지 않는다. 그래서 본 연구에서는 먼저 다음과 같이 Newton 알고리즘을 고려하였다. 즉, EM 알고리즘의 $(t+1)$번째 반복에서 $\alpha^{(t:0)}=\alpha^{(t)}$를 초기치로 하여

$$\alpha^{(t:\ell+1)} = \left(\alpha^{(t:\ell)} - \frac{g(\alpha^{(t:\ell)})}{g'(\alpha^{(t:\ell)})}\right)_+ \tag{3.22}$$

를 충분히 반복한 후 $\alpha^{(t+1)}=\alpha^{(t:\ell+1)}$을 얻었다. 여기서 $(x)_+=\max(0_+,\min(x,1_-))$로서 $0<\alpha<1$의 제약을 충족시키기 위함이다. 여기서 도함수 $g'(\alpha)$는

$$g'(\alpha) = (\mu_1-\mu_2)^2 + \left(\sigma_1^2+\sigma_2^2\right)\left[1-\frac{S_3^2(\alpha)}{\sigma_3^2(\alpha)}\right] - \left\{\alpha\left(\sigma_1^2+\sigma_2^2\right)-\sigma_2^2\right\}\left(\frac{S_3^2(\alpha)}{\sigma_3^2(\alpha)}\right)'$$

이며,

$$\left(\frac{S_3^2(\alpha)}{\sigma_3^2(\alpha)}\right)' = -\frac{2}{\sigma_3^2(\alpha)}\left[(\mu_1-\mu_2)\left\{\bar{y}_3-\mu_3(\alpha)\right\} + \left\{\alpha\left(\sigma_1^2+\sigma_2^2\right)-\sigma_2^2\right\}\frac{S_3^2(\alpha)}{\sigma_3^2(\alpha)}\right]$$

이다.

이 Newton 알고리즘은 경우에 따라 매우 많은 반복을 요구하였고, 종종 그릇된 수렴값을 제공하였다. 또한 보통 EM 알고리즘 내에 추가적인 반복 알고리즘 두는 것을 피하는 경향이 있으므로, 본 연구에서는 반복을 필요로 하지 않는 근사적인 방법을 고려해 보았다. 식 (3.21)에서 $S_3^2(\alpha)/\sigma_3^2(\alpha)\approx S_3^{2(t)}/\sigma_3^{2(t)}$를 대입하여 α에 대해 풀면,

$$\alpha^{(t+1)} = \frac{\left(\mu_1^{(t)}-\mu_2^{(t)}\right)\left(\bar{y}_3^{(t)}-\mu_2^{(t)}\right)+\sigma_2^{2(t)}\left[1-S_3^{2(t)}/\sigma_3^{2(t)}\right]}{\left(\mu_1^{(t)}-\mu_2^{(t)}\right)^2+\left(\sigma_1^{2(t)}+\sigma_2^{2(t)}\right)\left[1-S_3^{2(t)}/\sigma_3^{2(t)}\right]} \tag{3.23}$$

와 같은 해를 얻을 수 있다. 이 추정치는 EM 알고리즘이 수렴하면 함께 수렴하며, 만약 자식 군집의 중심 위치가 부 군집과 모 군집 사이에 있다면, 그 수렴값은 제약 조건 $0<\alpha^{(t+1)}<1$를 만족한다. 이를 증명해 보자.

$\bar{y}_3^{(t)} = \mu_3^{(t)}$이고 $S_3^{2(t)} = \sigma_3^{2(t)}$를 만족할 정도로 EM 알고리즘이 충분히 수렴하였다하자. 이때 가정에 의해 $\mu_1^{(t)} - \mu_2^{(t)} \neq 0$이므로, 식 (3.23)는

$$\alpha^{(t+1)} = \frac{\mu_3^{(t)} - \mu_2^{(t)}}{\mu_1^{(t)} - \mu_2^{(t)}} \tag{3.24}$$

$$= \frac{\left\{\alpha^{(t)}\mu_1^{(t)} + \left(1-\alpha^{(t)}\right)\mu_2^{(t)}\right\} - \mu_2^{(t)}}{\mu_1^{(t)} - \mu_2^{(t)}} = \alpha^{(t)}\frac{\mu_1^{(t)} - \mu_2^{(t)}}{\mu_1^{(t)} - \mu_2^{(t)}} = \alpha^{(t)} \tag{3.25}$$

이 된다. 우선 식 (3.25)로부터 $\alpha^{(t+1)} - \alpha^{(t)} = 0$이므로 결국 수렴하게 됨을 알 수 있다. 또한 자식 군집의 위치가 부 군집과 모 군집 사이에서 위치하고 있다면 $\mu_3^{(t)}$는 두 점 $(\mu_1^{(t)}, \mu_2^{(t)})$ 사이의 내점이므로, 식 (3.24)에서 $\mu_1^{(t)}$과 $\mu_2^{(t)}$의 대소 관계와 상관없이 $0 < \alpha^{(t+1)} < 1$을 만족한다.

식 (3.23) 추정치는 Newton 반복값을 사용하였을 때보다 훨씬 안정적이며, 정확하고 빠른 수렴 속도를 보였다. 따라서 다음 절 실험에서는 α의 추정을 위해 식 (3.23)를 이용하였다.

지금까지 이 절에서 제공한 $\mu_i^{(t+1)}$, $\sigma_i^{(t+1)}$ 및 $\alpha^{(t+1)}$에 대한 추정 방식은 동시에 결정하였다기 보다는 나머지들이 주어졌을 때 조건부로 추정한 것이다. 따라서 엄밀히 말해 우리의 EM 알고리즘은 Meng과 Rubin (1993)의 ECM(expectation-conditional maximization) 알고리즘이라 할 것이다. 이 알고리즘 역시 단조 수렴성을 보장한다.

4. 수치 실험

이 절에서는 부 군집, 모 군집 및 자식 군집 그리고 한 개의 기타 군집의 정규모집단으로부터 각각 생성한 자료에서 제안된 방법이 군집 족을 얼마나 잘 식별하는지를 실험할 것이다. 부 군집의 모평균과 모분산은 각각 $\mu_1 = -3.5, -7, -14$ 및 $\sigma_1^2 = 1$, 그리고 모 군집에 대해서는 $\mu_2 = 3.5, 7, 14$ 및 $\sigma_2^2 = 4$으로 하였다. 그래서 부군집과 모군집과의 거리가 "가까울 때", "보통 일 때", "멀 때"를 각각 $\mu_2 - \mu_1 = 7, 14, 28$로 정하여 실험한다. 이때 자식 군집의 모평균과 모분산은 $\mu_3 = \alpha(-7) + (1-\alpha)(7)$과 $\sigma_3^2 = \alpha^2(1) + (1-\alpha)^2(4)$로 계산되며, α 값에 따라 자식 군집으로서 다양한 자료가 만들어지게 될 것이다. 그리고 기타 군집의 모평균과 모분산을 $\mu_4 = \mu_2 + 8$ 및 $\sigma_4^2 = 5$로 하여 위치와 산포를 모 군집과 비슷하게 함으로써 구분을 어렵게 하였다. 자료의 개수 $n = 1000$개로 하되, $\pi_1 = \pi_2 = 0.30$, $\pi_3 = 0.25, \pi_4 = 0.15$로 하여 부 군집과 모 군집의 각 모집단에서 각각 300개씩, 자식 군집의 모집단으로부터 250개 그리고 기타 군집의 모집단으로부터 150개의 표본을 생성하였다. 이제 파생 강도 α를 0.2, 0.5, 0.8에 대해 제안된 방법이 군집 족을 얼마나 잘 식별하는지 실험에 보도록 할 것이다.

정규혼합모형의 적합을 통해 우리는 사후확률 추정치(엄격히 말하면, 결측 자료 z_{ij}의 대체값) $\hat{\tau}_{hj}$를 얻을 수 있다. 이때 최대 사후확률 추정치의 레이블로서 관측치들을 g개의 군집으로 분류할 수도 있고(outright clustering), 혹은 $\hat{\tau}_{hj}$를 자체를 제공하여 관측치 j가 군집 i에 속할 정도를 파악하게 할 수도 있다(probabilistic clustering). 저자는 후자의 방법을 선택할텐데, 본 연구에서는 단변량 자료를 다루고 있으므로 $n \cdot g$개의 $\hat{\tau}_{hj}$를 모두 나열하는 대신, 자료의 경험분포에서 추정된 정규혼합모형 (2.1)이 얼마나 잘 적합되었는지 그리고 부 군집, 모 군집, 자식 군집의 자료를 얼마나 잘 식별하고 있는지를 살피는 것으로 충분할 것이다.

설명을 위해 그림 1의 첫 번째에 파생 강도 $\mu_2 - \mu_1 = 14$이며 $\alpha = 0.5$일 때 생성된 자료들의 경험분포를 나타내었다. 군집 족의 구성원 각각과 기타 군집이 비교적 잘 분리되어 있어 보인다 (Father: 부 군집, Mother: 모 군집, Son: 자식 군집, Other: 기타 군집). 그러나 현실에서는 4개의 군집 중 어느 3군집이 군집 족인지 알기는 어렵다. 아마 군집 족을 구성할 수 있는 가능한 가지 수는 8가지가 될 것이

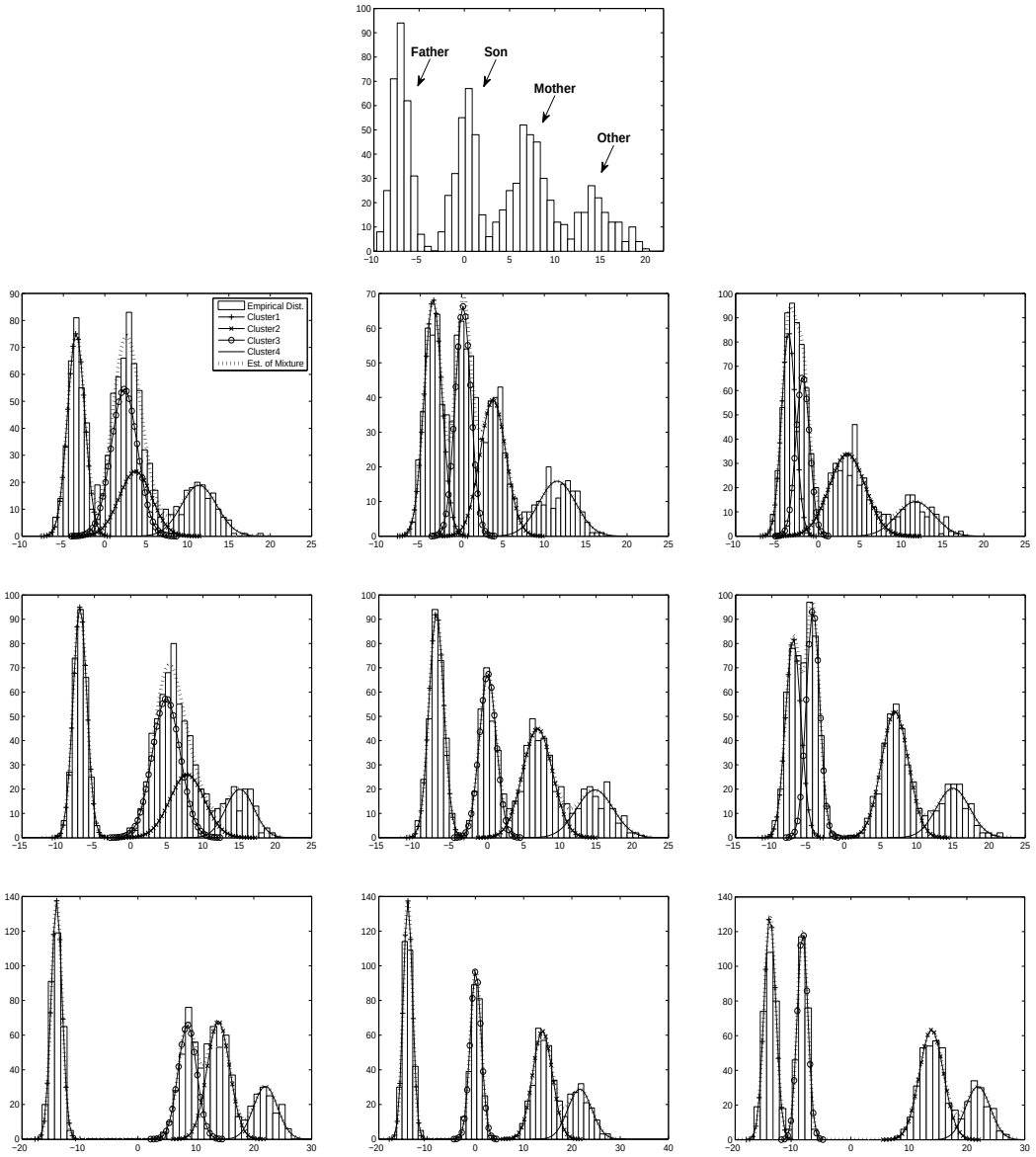

그림 1: *군집 족 식별 결과. 제1행은* $\mu_2 - \mu_1 = 14$ *및* $\alpha = 0.5$*일 때 생성된 예시 자료. 제2, 3, 4행은 각각 모 군집과 부 군집의 모평균 차이가* $\mu_2 - \mu_1 = 7, 14, 28$*일 때의 결과이며, 제1, 2, 3열은 파생강도가 각각* $\alpha = 0.2, 0.5, 0.8$*일 때의 결과들이다. 마커* $+, \circ, \times$*를 가진 실선 및 마커가 없는 실선은 각각* $\hat{\pi}_i \phi(y; \hat{\mu}_i, \hat{\sigma}_i^2)\,(i = 1, 2, 3, 4)$*를 그린 것으로서 각각 식별된 부 군집, 모 군집, 자식 군집 및 기타 군집을 나타내며, 검정색 점선은 이들의 합으로서 추정된 혼합모형을 나타낸다.*

며, 그 중 하나만이 참 군집 족이다. 자료 내에 군집이 많을 경우, 표준적인 군집분석 후 탐색적 방법으로 참 군집 족을 식별하기란 매우 어려울 것으로 판단된다.

그림 1의 제2, 3, 4행은 각각 모 군집과 부 군집의 모평균 차이가 $\mu_2 - \mu_1 = 7, 14, 28$일 때의 결과이

표 1: 모수 추정 결과. G_1: 부 군집, G_2: 모 군집, G_3: 자식 군집, G_4: 기타 군집. 군집별 (부 군집 평균, 모 군집 평균): (μ_1, μ_2)= $(-3.5, 3.5)$, $(-7, 7)$ 및 $(-14, 14)$이며, $\mu_3 = \alpha(\mu_1) + (1-\alpha)(\mu_2)$이고 $\mu_4 = \mu_2 + 8$이다. 군집별 모분산(σ_i^2): $1, 4, \alpha^2(1) + (1-\alpha)^2(4)$, 및 5이다. 군집별 혼합비율($\pi_i$): 0.3, 0.3 0.25, 0.15이다. 자료의 총 개수 단, $n = 1000$이다.

평균사이 거리 $\lvert\mu_2 - \mu_1\rvert$	파생강도 α	추정치	군집 G_1	G_2	G_3	G_4	α 추정치
7	0.2	$\hat{\mu}_i$	−3.4579	3.7148	2.3626	11.3916	
		$\hat{\sigma}_i^2$	0.8817	4.0556	2.5594	3.5656	0.1885
		$\hat{\pi}_i$	0.3106	0.1852	0.3444	0.1598	
	0.5	$\hat{\mu}_i$	−3.4435	3.7993	0.1876	11.5684	
		$\hat{\sigma}_i^2$	1.1561	2.4572	0.9051	4.9279	0.4987
		$\hat{\pi}_i$	0.3146	0.2640	0.2702	0.1511	
	0.8	$\hat{\mu}_i$	−3.5983	3.5005	−1.8832	11.7411	
		$\hat{\sigma}_i^2$	0.7182	4.7797	0.6921	5.1824	0.7584
		$\hat{\pi}_i$	0.3064	0.3182	0.2357	0.1397	
14	0.2	$\hat{\mu}_i$	−7.0003	7.8098	4.9167	15.0767	
		$\hat{\sigma}_i^2$	0.9096	5.7105	3.7321	3.7686	0.1953
		$\hat{\pi}_i$	0.3000	0.2054	0.3663	0.1283	
	0.5	$\hat{\mu}_i$	−7.0425	6.9263	0.0604	14.8529	
		$\hat{\sigma}_i^2$	1.0467	4.2573	1.3536	6.4332	0.4915
		$\hat{\pi}_i$	0.2999	0.2929	0.2490	0.1582	
	0.8	$\hat{\mu}_i$	−7.0447	7.0047	−4.3109	15.0263	
		$\hat{\sigma}_i^2$	1.1334	3.4537	0.8660	5.2346	0.8054
		$\hat{\pi}_i$	0.2747	0.3033	0.2752	0.1467	
28	0.2	$\hat{\mu}_i$	−14.0280	13.9031	8.5491	22.0364	
		$\hat{\sigma}_i^2$	0.8762	3.8171	2.5262	4.2544	0.1917
		$\hat{\pi}_i$	0.3000	0.3086	0.2444	0.1471	
	0.5	$\hat{\mu}_i$	−13.9752	13.9059	0.0831	21.6295	
		$\hat{\sigma}_i^2$	0.9294	4.1922	1.2943	6.0426	0.4958
		$\hat{\pi}_i$	0.3000	0.2889	0.2500	0.1611	
	0.8	$\hat{\mu}_i$	−13.9939	13.9023	−8.3628	21.9121	
		$\hat{\sigma}_i^2$	1.0324	4.3721	0.8358	4.5267	0.7981
		$\hat{\pi}_i$	0.2994	0.3012	0.2506	0.1488	

며, 제1, 2, 3열은 파생강도가 각각 $\alpha = 0.2, 0.5, 0.8$일 때의 결과들을 나타낸 것이다. 마커 +(파랑색), ∘(빨강색), ×(초록색)를 가진 실선 및 마커가 없는 실선(검정색)은 각각 $\hat{\pi}_i\phi(y; \hat{\mu}_i, \hat{\sigma}_i^2)$ $(i = 1, 2, 3, 4)$를 그린 것으로서 각각 식별된 부 군집, 모 군집, 자식 군집 및 기타 군집을 나타내며, 검정색 점선은 이들의 합으로서 추정된 혼합모형을 나타낸다.

모든 경우에 있어서 군집 족의 구성원을 각각 잘 식별되고 있으며, 자료의 경험분포도 추정된 혼합모형으로 양호하게 적합되고 있음을 알 수 있다. 표 1은 $\mu_2 - \mu_1 = 7, 14, 28$ 각각의 경우에 대하여 각각 $\alpha = 0.2, 0.5$, 및 0.8에서의 모수 추정의 결과를 제공하고 있다. 부 군집과 모 군집의 거리가 가까운 $\mu_2 - \mu_1 = 7$일 때 $\alpha = 0.2$ 및 0.8에서 다소 정확성이 떨어지기는 하지만, 대체로 모든 모수들을 거의 정확하게 추정하고 있음을 확인 할 수 있다. 또한 자식 군집의 평균 추정치와 분산 추정치는 제약한 대로 $\hat{\mu}_3 = \hat{\alpha}\hat{\mu}_1 + (1-\hat{\alpha})\hat{\mu}_2$과 $\hat{\sigma}_3 = \hat{\alpha}^2\hat{\sigma}_1^2 + (1-\hat{\alpha})^2\hat{\sigma}_2^2$을 만족한다.

본 수치 실험에서는 모든 경우 식 (2.4)의 BIC는 군집의 개수 $g = 4$를 추천하였다. 다만, $\alpha < 0.1$ 혹은 > 0.9일 때 BIC는 $g = 3$ 혹은 4 이상을 제안하곤 하였는데, 이것은 그렇게 놀랄 일은 아닐 것이다. 왜냐하면 그런 상황은 자식 군집이 부 군집 혹은 모 군집에 완전히 통합되어 한 개의 정규분포나 다수개의 정규분포의 혼합으로 인식될 수 있기 때문이다. $\alpha < 0.1$ 혹은 $\alpha > 0.9$ 경우는 실험에서 제외시켰

는데, 사실 현실에서는 이러한 경우 자식군집이 있는 것으로 보이지 않을 것이기 때문에 분석자는 군집 족 식별을 시도하지도 않을 것이다.

5. 결론과 토의

본 논문에서는 자료 내의 군집 중에 '父 군집'과 '母 군집'이라 부르는 두 군집 사이에, 합성된 평균과 분산을 가지는 '자식 군집'이라 부르는 한 군집이 있을 경우에 주목하여, 그들의 관계를 평균과 분산에 관해 모형화하고 각각의 군집을 식별하는 방법을 제공하였다. 관측치는 정규혼합모형을 따른다고 가정하고, EM 알고리즘을 통해 모형 추정을 시도하였다. 추정 과정에 여러 난제가 있었으나, 근사적 방법으로 비교적 잘 극복할 수 있었다. 그리고 수치실험을 통해 제안 방법은 성공적으로 주어진 세 군집 즉 '군집 族'을 식별할 수 있음을 보였다.

그러나 제안된 방법은 중요한 문제점을 가지고 있는데, 부 군집과 모 군집 사이에 군집이 2개 이상이면 잘 식별하지 못하였다. 이와 같은 문제점의 이유를 알아내서 해결하는 것을 포함하여, 다변량 자료로 확장해 보는 것이 추후 과제라 할 것이다. 한편, 모형 $\sigma_3 = \alpha\sigma_1^2 + (1-\alpha)\sigma_2^2$은 우리의 분산 모형 $\sigma_3 = \alpha^2\sigma_1^2 + (1-\alpha)^2\sigma_2^2$ 보다 자식 군집이 부 군집 혹은 모 군집에 가까이 갈 때 다소 덜 빠르게 닮은 형태를 취한다. 따라서 전자의 모형도 한 번 고려해 볼만하다 하겠다.

참고 문헌

김승구 (2007). Normal mixture model with general linear regressive restriction: Applied to Microarray Gene Clustering, <한국통계학회논문집>, **14**, 205–213.

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

Green, P. J. (1990). On use of the EM algorithm for penalized likelihood estimation, *Journal of Royal Statistical Society B*, **52**, 443–452.

Levin, A., Lischinski, D. and Weiss, Y. (2008). A closed form solution to natural image matting, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **30**, 228–242.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*, John Wiley & Sons, Inc.

Meng, X.-L. and Rubin, D. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework, *Biometrika*, **80**, 267–278.

Ng, S. K., McLachlan, G. J., Wang, K., Ben-Tovim, L. and Ng, S. W. (2006). A Mixture model with random-effects components for clustering correlated gene-expression profiles, *Bioinformatics*, **22**, 1745–1752.

Schwarz, G. (1978). Estimating the dimension of a model, *Annals of Statistics*, **6**, 461–464.

Titterington, D. M., Smith, A. F. and Makov, U. E. (1994). *Statistical Analysis of Finite Mixture Distributions*, John Wiely & Sons.

Wang, J. and Cohen, M. F. (2005). An interactive optimization approach for unified image segmentation and matting, *ICCV 2005*, 936–943.

Wang, S. and Zhu, J. (2008). Variable selection for model-based high-dimensional clustering and its application to Microarray data, *Bioinformatics*, **64**, 440–448.

2011년 3월 접수; 2011년 4월 채택

Identification of Cluster with Composite Mean and Variance

Seung-Gu Kim[1,a]

[a]Department of Data & Information, Sangji University

Abstract

Consider a cluster, so called a 'son cluster', whose mean and variance is composed of the means and variances of both clusters called as a 'father cluster' and a 'mother cluster'. In this paper, a method for identifying each of three clusters is provided by modeling the relationship with father and mother clusters.

Under the normal mixture model, the parameters are estimated via EM algorithm. We were able to overcome the problems of estimation using ECM approximation. Numerical examples show that our method can effectively identify the three clusters, so called a 'family of clusters'.

Keywords: Composite cluster, strength of derivation, family of clusters, normal mixture model, EM algorithm.

This research was supported by Sangji Research Fund 2010.

[1] Professor, Department of Data & Information, Sangji University, 660 Woosan-Dong, Wonju, KangWon-Do 220-702, Korea. E-mail: sgukim@sangji.ac.kr