The Korean Journal of Applied Statistics (2013) DOI: http://dx.doi.org/10.5351/KJAS.2013.26.5.821
26(5), 821–834

Variable Selection in Clustering by Recursive Fit of Normal Distribution-based Salient Mixture Model

Seung-Gu Kim[a,1]

[a]Department of Data and Information, Sangji University

(Received August 26, 2013; Revised October 18, 2013; Accepted October 21, 2013)

Abstract

Law *et al.* (2004) proposed a normal distribution based salient mixture model for variable selection in clustering. However, this model has substantial problems such as the unidentifiability of components and the inaccurate selection of informative variables in the case of a small cluster size. We propose an alternative method to overcome problems and demonstrate a good performance through experiments on simulated data and real data.

Keywords: Saliency parameter, variable selection, clustering, normal mixture model, EM algorithm.

1. 서론

혼합모형을 기반으로 한 군집분석에서 변수선택(variable selection or feature selection)은 최근 통계학뿐만 아니라 인공지능 및 정보생물학 분야에서 큰 관심사이다. 혼합모형기반 군집분석에서 변수선택은 크게 두 가지 접근법으로 이루어져 왔는데, 하나는 L_1 벌점화 우도 접근법으로서 Pan과 Shen (2006), Wang과 Zhu (2008) 및 Xie 등 (2008)에 의해 연구되었다. 나머지 하나는 본 연구에서 다루고자 하는 접근법으로서 "두각 모수(saliency parameter)"를 도입한 혼합모형기반 군집분석 기법이다.

두각모수 접근법은 Law 등 (2004)에 의해 처음 제안되었다. 이후 Graham과 Miller (2006)에 의해 성분 개수의 결정에 대해 연구되었고, Bouguila와 Ziou (2006), Boutemedjet 등 (2009), Bouguila 등 (2012) 및 Elguebaly와 Bouguila (2013)에 의해 베이지안 접근법으로 확장되었다. 한편 Li와 Hua (2008), Li 등 (2009)은 기존의 전역적 두각성(global saliency) 즉 성분 공통적 정보적 변수 선택 방식의 비현실성을 지적하고 성분별 두각성을 고려한 국소적 변수선택 방식을 제안하였다. 그러나 본 연구에서는 전역적 변수선택 방식으로 고려할 것이다. 성분 별로 어떤 변수가 더 정보적인지를 설명할 수 있다면 매우 유용한 것임에 분명하다. 그러나 전역적 변수선택이 여전히 유효한 이유는 첫째, 국소적 두각성의 고려는 모형에서 모수 개수를 크게 늘리게 되므로 자료가 제한적인 경우 과적합의 위험성이 있고, 둘째, 마이크로어레이 분석 등과 같은 응용문제에서는 성분을 구분해 주는 전역적 변수를 찾는 것이 더욱 중요하기 때문이라 하겠다.

본 연구의 목적은 Law 등 (2004)의 두각 모수 혼합모형의 적합상의 문제점과 변수선택의 오류 가능성을 보완하고자 하는 것이다. 다음 절에서는 그들의 모형 설명과 EM 알고리즘에 의한 추정법을 가능한 한 통계학 친화적으로 (그들의 논문에 제시된 EM 알고리즘은 통계학자들에게 다소 불친절한 방식으로

[1]Professor, Department of Data and Information, Sangji University, 83 Usan-Dong, Wonju 220-702, Korea. E-mail: sgukim@sangji.ac.kr

제공되고 있으므로) 그리고 상세하게 설명하되 그들이 언급하지 않았던 내용을 포함하여 제공할 것이다. 그리고 3절에서는 언급된 문제점들을 지적하면서 문제해결 방법을 제시하고, 4절에서는 다양한 모의실험 자료와 실자료를 이용하여 제안 방법의 실효성 보일 것이다. 5절에서는 결론을 정리하고 몇 가지 문제에 대한 토의할 것이다.

2. 모형 소개

2.1. 정규분포 두각 혼합모형: 개념 및 용어 정의

n개의 관측들이 서로 독립인 j번째 p-변량 관측치 $\boldsymbol{y}_j = (y_{1j}, \ldots, y_{pj})^T$의 g-성분 혼합모형

$$f\left(\boldsymbol{y}_j; \boldsymbol{\Theta}\right) = \sum_{i=1}^{g} \pi_i \prod_{k=1}^{p} \left[\eta_k \phi\left(y_{jk}; \mu_{ik}, \sigma_{ik}^2\right) + (1-\eta_k)\, \phi\left(y_{jk}; \mu_k, \sigma_k^2\right)\right], \quad j = 1, \ldots, n \tag{2.1}$$

을 고려하자. 여기서 $\boldsymbol{\Theta}$는 모형에 포함된 모든 모수들의 집합을 나타내며, π_i는 i번째 성분의 혼합비율(mixing proportion)이며, $\phi(\cdot\,; \mu, \sigma^2)$은 평균과 분산이 (μ, σ^2)인 단변량 정규분포 밀도를 나타낸다. 그리고 $\eta_k\,(k = 1, \ldots, p)$는 k번째 변수가 "정보적 변수(informative variable or relevant feature)"일 확률로서 "두각 모수(saliency parameter)"라 부른다.

$\eta_k \to 1$이면 k번째 관측치 y_{jk}는 성분밀도 $\phi(y_{jk}; \mu_{ik}, \sigma_{ik}^2)\,(i = 1, \ldots, g)$ (모수들에 성분 첨자 i가 있음을 유의)의 표본일 가능성이 커져서 g개의 성분을 구분하는데 기여하므로 정보적이지만, $\eta_k = 0$이면 성분밀도 $\phi(y_{jk}; \mu_k, \sigma_k^2)$ (모수들에 성분 첨자 i가 없음을 유의)의 표본이므로 성분을 구분하는데 비정보적임을 의미한다. 특히 $\eta_k = 1$일 때 k번째 변수는 "완전하게 정보적"이라 부를 것이다. 완전하게 정보적인 변수란 군집들을 온전하게 구분하는데 꼭 필요한 변수라 정의할 수 있다.

앞으로 $\phi(y_{jk}; \mu_{ik}, \sigma_{ik}^2)$와 같이 구분된 성분 모수를 가지는 밀도를 "유관밀도(relevant density)"라, $\phi(y_{jk}; \mu_k, \sigma_k^2)$와 같이 성분 공통 모수를 가지는 밀도를 "무관밀도(irrelevant density)"라 부를 것이다.

2.2. 모형에서의 가정들

모형 (2.1)은 성분 i가 주어졌을 때 p개 변수들은 서로 독립인 조건부 독립을 가정하고 있다. 이 가정은 현실의 많은 문제에서 꽤 만족스럽게 받아들여진다.

Z_{ij}를 j번째 관측치 $\boldsymbol{y}_j$가 i번째 성분에서 왔으면 1 그렇지 않으면 0, 그리고 X_k를 k번째 변수 Y_k가 정보적이면 1 그렇지 않으면 0을 나타내는 지시변수라 하자. 이때

$$\eta_k = P_{\boldsymbol{\Theta}}\left\{X_k = 1 | Z_{ij} = 1\right\} = P_{\boldsymbol{\Theta}}\left\{X_k = 1\right\}$$

임을 가정하고 있다. 즉, 성분의 소속과 정보적 변수 여부는 서로 독립이다. 그러나 $\boldsymbol{y}_j$가 주어졌을 때는

$$P_{\boldsymbol{\Theta}}\left\{X_k = 1 | Z_{ij} = 1, \boldsymbol{y}_j\right\} \neq P_{\boldsymbol{\Theta}}\left\{X_k = 1, \boldsymbol{y}_j\right\}$$

임을 가정하고 있다. 즉, 두각성은 관측전에는 성분 소속과 독립이지만 관측 후에는 종속인 조건부 종속성을 가정하고 있다.

2.3. 완비자료

관측치 $\boldsymbol{y} = (\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_n^T)^T$이 주어졌을 때, 로그-우도는

$$\log L(\boldsymbol{\Theta}) = \sum_{j=1}^{n} \log \left\{ \sum_{i=1}^{g} \pi_i \prod_{k=1}^{p} \left[\eta_k \phi\left(y_{jk}; \mu_{ik}, \sigma_{ik}^2\right) + (1-\eta_k)\phi\left(y_{jk}; \mu_k, \sigma_k^2\right)\right] \right\} \tag{2.2}$$

이다. 이제 미관측 결측자료 $\boldsymbol{z}_j = (z_{1j}, \ldots, z_{gj})^T$ 및 $\boldsymbol{x} = (x_1, \ldots, x_p)^T$라 쓰자. 이때 완비자료 $(\boldsymbol{y}_j, \boldsymbol{x}, \boldsymbol{z}_j)$의 결합밀도는

$$f_c(\boldsymbol{y}_j, \boldsymbol{x}, \boldsymbol{z}_j; \boldsymbol{\Theta}) = \prod_{i=1}^{g} \pi_i^{z_{ij}} \left[\prod_{k=1}^{p} \eta_k^{x_k} (1-\eta_k)^{1-x_k} \prod_{k=1}^{p} \phi\left(y_{jk}; \mu_{ik}, \sigma_{ik}^2\right)^{x_k} \phi\left(y_{jk}; \mu_k, \sigma_k^2\right)^{1-x_k} \right]^{z_{ij}} \tag{2.3}$$

와 같이 얻을 수 있다. 실제로 모든 $z_{ij}\,(i = 1, \ldots g; j = 1, \ldots, n)$와 모든 $x_k\,(k = 1, \ldots, p)$에 대해 $f_c(\boldsymbol{y}_j, \boldsymbol{x}, \boldsymbol{z}_j; \boldsymbol{\Theta})$를 합산하면 식 (2.1)의 $f(\boldsymbol{y}_j; \boldsymbol{\Theta})$임으로 보일 수 있다.

2.4. E-Step

관측치 $\boldsymbol{y} = (\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_n^T)^T$과 $\boldsymbol{\Theta}$의 (이전 단계) 추정치 $\widehat{\boldsymbol{\Theta}}$가 주어졌을 때, 완비자료 로그-우도의 조건부 기대값은

$$\begin{aligned} Q\left(\boldsymbol{\Theta}|\widehat{\boldsymbol{\Theta}}\right) &= E\left[\sum_{j=1}^{n} \log f_c\left(\boldsymbol{y}_j, \boldsymbol{x}, \boldsymbol{z}_j, u_j; \boldsymbol{\Theta}\right) | \boldsymbol{y}_j, \widehat{\boldsymbol{\Theta}}\right] \\ &\propto Q_1\left(\boldsymbol{\pi}|\widehat{\boldsymbol{\Theta}}\right) + Q_2\left(\boldsymbol{\eta}|\widehat{\boldsymbol{\Theta}}\right) + Q_3\left(\boldsymbol{\mu}_{[1]}, \boldsymbol{\sigma}_{[1]}^2|\widehat{\boldsymbol{\Theta}}\right) + Q_4\left(\boldsymbol{\mu}_{[0]}, \boldsymbol{\sigma}_{[0]}^2|\widehat{\boldsymbol{\Theta}}\right) \end{aligned} \tag{2.4}$$

와 같이 얻게 된다. 단,

$$Q_1\left(\boldsymbol{\pi}|\widehat{\boldsymbol{\Theta}}\right) = \sum_{j=1}^{n}\sum_{i=1}^{g} E_{\widehat{\boldsymbol{\Theta}}}\left(Z_{ij}|\boldsymbol{y}_j\right) \log \pi_i,$$

$$Q_2\left(\boldsymbol{\eta}|\widehat{\boldsymbol{\Theta}}\right) = \sum_{j=1}^{n}\sum_{i=1}^{g}\sum_{k=1}^{p} \left[E_{\widehat{\boldsymbol{\Theta}}}(Z_{ij}X_k|\boldsymbol{y}_j) \log \eta_k + \left\{E_{\widehat{\boldsymbol{\Theta}}}(Z_{ij}|\boldsymbol{y}_j) - E_{\widehat{\boldsymbol{\Theta}}}(Z_{ij}X_k|\boldsymbol{y}_j\right\} \log(1-\eta_k)\right],$$

$$Q_3\left(\boldsymbol{\mu}_{[1]}, \boldsymbol{\sigma}_{[1]}^2|\widehat{\boldsymbol{\Theta}}\right) = \sum_{j=1}^{n}\sum_{i=1}^{g}\sum_{k=1}^{p} \left[-\frac{1}{2} E_{\widehat{\boldsymbol{\Theta}}}(Z_{ij}X_k|\boldsymbol{y}_j) \log \sigma_{ik}^2 - \frac{1}{2} E_{\widehat{\boldsymbol{\Theta}}}(Z_{ij}X_k|\boldsymbol{y}_j) \frac{(y_{jk}-\mu_{ik})^2}{\sigma_{ik}^2}\right],$$

$$\begin{aligned} Q_4\left(\boldsymbol{\mu}_{[0]}, \boldsymbol{\sigma}_{[0]}^2|\widehat{\boldsymbol{\Theta}}\right) = \sum_{j=1}^{n}\sum_{i=1}^{g}\sum_{k=1}^{p} \Bigg[&-\frac{1}{2}\left\{E_{\widehat{\boldsymbol{\Theta}}}(Z_{ij}|\boldsymbol{y}_j) - E_{\widehat{\boldsymbol{\Theta}}}(Z_{ij}X_k|\boldsymbol{y}_j)\right\} \log \sigma_{ik}^2 \\ &- \frac{1}{2}\left\{E_{\widehat{\boldsymbol{\Theta}}}(Z_{ij}|\boldsymbol{y}_j) - E_{\widehat{\boldsymbol{\Theta}}}(Z_{ij}X_k|\boldsymbol{y}_j)\right\} \frac{(y_{jk}-\mu_k)^2}{\sigma_k^2}\Bigg] \end{aligned}$$

그리고 $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_g)^T$, $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_p)^T$, $\boldsymbol{\mu}_{[1]} = (\mu_{11}, \ldots, \mu_{gp})^T$, $\boldsymbol{\sigma}_{[1]}^2 = (\sigma_{11}^2, \ldots, \sigma_{gp}^2)^T$, $\boldsymbol{\mu}_{[0]} = (\mu_1, \ldots, \mu_p)^T$ 및 $\boldsymbol{\sigma}_{[0]}^2 = (\sigma_1^2, \ldots, \sigma_p^2)^T$를 나타낸다.

이때 2가지 종류의 결측자료의 조건부 기대값 $E_{\widehat{\boldsymbol{\Theta}}}\left(Z_{ij}|\boldsymbol{y}_j\right)$, $E_{\widehat{\boldsymbol{\Theta}}}(Z_{ij}X_k|\boldsymbol{y}_j)$을 계산해야 한다. 우선

$$\begin{aligned} \tau_{ij} = E_{\widehat{\boldsymbol{\Theta}}}\left(Z_{ij}|\boldsymbol{y}_j\right) &= P_{\widehat{\boldsymbol{\Theta}}}\left\{Z_{ij} = 1|\boldsymbol{y}_j\right\} \\ &= \frac{\hat{\pi}_i \prod_{k=1}^{p} \left[\hat{\eta}_k \phi\left(y_{jk}; \hat{\mu}_{ik}, \hat{\sigma}_{ik}^2\right) + (1-\hat{\eta}_k)\phi\left(y_{jk}; \hat{\mu}_k, \hat{\sigma}_k^2\right)\right]}{\sum_{h=1}^{g} \hat{\pi}_h \prod_{k=1}^{p} \left[\hat{\eta}_k \phi\left(y_{jk}; \hat{\mu}_{hk}, \hat{\sigma}_{hk}^2\right) + (1-\hat{\eta}_k)\,\phi\left(y_{jk}; \hat{\mu}_k, \hat{\sigma}_k^2\right)\right]}, \end{aligned} \tag{2.5}$$

$$\begin{aligned} \gamma_{ijk} &= P_{\widehat{\boldsymbol{\Theta}}}\left\{X_k = 1|Z_{ij} = 1, \boldsymbol{y}_j\right\} \\ &= \frac{\hat{\eta}_k \phi\left(y_{jk}; \hat{\mu}_{ik}, \hat{\sigma}_{ik}^2\right)}{\hat{\eta}_k \phi\left(y_{jk}; \hat{\mu}_{hk}, \hat{\sigma}_{hk}^2\right) + (1-\hat{\eta}_k)\,\phi\left(y_{jk}; \hat{\mu}_k, \hat{\sigma}_k^2\right)} \end{aligned} \tag{2.6}$$

이다. τ_{ij}는 j번째 관측치 $\boldsymbol{y}_j$가 i번째 속할 사후확률이며, γ_{ijk}는 j번째 관측치 $\boldsymbol{y}_j = (y_{j1}, \ldots, y_{jk}, \ldots, y_{jp})^T$가 i번째 성분에 속할 때 y_{jk}가 정보적 변수의 관측치일 사후확률이다.

따라서

$$\begin{aligned}\kappa_{ijk} = E_{\widehat{\boldsymbol{\Theta}}}(Z_{ij}X_k|\boldsymbol{y}_j) &= P_{\widehat{\boldsymbol{\Theta}}}\left\{X_k = 1, Z_{ij} = 1|\boldsymbol{y}_j\right\} \\ &= P_{\widehat{\boldsymbol{\Theta}}}\left\{X_k = 1|Z_{ij} = 1, \boldsymbol{y}_j\right\} P_{\widehat{\boldsymbol{\Theta}}}\left\{Z_{ij} = 1|\boldsymbol{y}_j\right\} \\ &= \gamma_{ijk}\tau_{ij} \end{aligned} \tag{2.7}$$

로서 계산된다. 한편 관측치 $\boldsymbol{y}_j$가 주어졌을 때, y_{jk}가 정보적 변수의 관측치일 (비조건부) 확률은

$$P_{\widehat{\boldsymbol{\Theta}}}\left\{X_k = 1|\boldsymbol{y}_j\right\} = \sum_{i=1}^{g} P_{\widehat{\boldsymbol{\Theta}}}\left\{X_k = 1, Z_{ij} = 1|\boldsymbol{y}_j\right\} = \sum_{i=1}^{g}\kappa_{ijk} = \sum_{i=1}^{g}\tau_{ij}\gamma_{ijk} \stackrel{\text{let}}{=} \bar{\gamma}_{jk} \tag{2.8}$$

와 같이 나타낼 수 있을 것이다.

2.5. M-Step

앞서 2가지의 결측자료의 조건부 기대값 τ_{ij} 및 κ_{ijk}를 구하였다. 이들을 바탕으로 우리는 식 (2.4)의 $Q(\boldsymbol{\Theta}|\widehat{\boldsymbol{\Theta}})$를 최대화하는 추정치를 구한 결과는 다음과 같다. 즉, $i = 1, \ldots, g$ 및 $k = 1, \ldots, p$에 대하여

$$\begin{aligned}
\hat{\pi}_i &= \frac{\sum_{j=1}^{n}\tau_{ij}}{n}, \\
\hat{\eta}_k &= \frac{\sum_{j=1}^{n}\sum_{i=1}^{g}\kappa_{ijk}}{\sum_{j=1}^{n}\sum_{i=1}^{g}\tau_{ij}} = \frac{\sum_{j=1}^{n}\sum_{i=1}^{g}\tau_{ij}\gamma_{ijk}}{n} = \frac{\sum_{j=1}^{n}\bar{\gamma}_{jk}}{n}, \\
\hat{\mu}_{ik} &= \frac{\sum_{j=1}^{n}\kappa_{ijk}y_{jk}}{\sum_{j=1}^{n}\kappa_{ijk}} = \frac{\sum_{j=1}^{n}\tau_{ij}\gamma_{ijk}y_{jk}}{\sum_{j=1}^{n}\tau_{ij}\gamma_{ijk}}, \\
\hat{\sigma}_{ik}^2 &= \frac{\sum_{j=1}^{n}\kappa_{ijk}(y_{jk} - \hat{\mu}_{ik})^2}{\sum_{j=1}^{n}\kappa_{ijk}} = \frac{\sum_{j=1}^{n}\tau_{ij}\gamma_{ijk}(y_{jk} - \hat{\mu}_{ik})^2}{\sum_{j=1}^{n}\tau_{ij}\gamma_{ijk}}, \\
\hat{\mu}_k &= \frac{\sum_{j=1}^{n}\sum_{i=1}^{g}(\tau_{ij} - \kappa_{ijk})y_{jk}}{\sum_{j=1}^{n}\sum_{i=1}^{g}(\tau_{ij} - \kappa_{ijk})} = \frac{\sum_{j=1}^{n}\sum_{i=1}^{g}\tau_{ij}(1 - \gamma_{ijk})y_{jk}}{\sum_{j=1}^{n}\sum_{i=1}^{g}\tau_{ij}(1 - \gamma_{ijk})}, \\
\hat{\sigma}_k^2 &= \frac{\sum_{j=1}^{n}\sum_{i=1}^{g}(\tau_{ij} - \kappa_{ijk})(y_{jk} - \hat{\mu}_k)^2}{\sum_{j=1}^{n}\sum_{i=1}^{g}(\tau_{ij} - \kappa_{ijk})} = \frac{\sum_{j=1}^{n}\sum_{i=1}^{g}\tau_{ij}(1 - \gamma_{ijk})(y_{jk} - \hat{\mu}_k)^2}{\sum_{j=1}^{n}\sum_{i=1}^{g}\tau_{ij}(1 - \gamma_{ijk})}
\end{aligned} \tag{2.9}$$

을 얻는다.

이상의 결과들을 바탕으로 E-Step과 M-Step를 모든 추정치들이 적절한 종료 기준 하에서 충분히 수렴할 때까지 반복한다.

3. 문제점 및 대응책

이 절에서는 2절에서 소개된 Law 등 (2004) 모형의 문제점과 그 해결방법을 제시한다.

3.1. 군집의 식별성(Identifiability)

두 정보적 변수 (y_1, y_2) 공간에 $g = 4$의 군집이 잘 분리되어 있다하자. 그리고 이를 모형화하여 $\pi_1 = \pi_2 = \pi_3 = \pi_4 = 1/4$이고 $\sigma_{11}^2 = \cdots = \sigma_{42} = 1$이며, $\eta_1 = \eta_2 = 1/2$로서 4개의 성분은 오직 평균 $\boldsymbol{\mu}_i = (\mu_{i1}, \mu_{i2})^T\,(i = 1, 2, 3, 4)$에 의해 구분된다하자. Figure 3.1(a)의 경우 4개 성분의 평균들이 y_1-축과 y_2-축의 동일 선상에 위치하는 성분이 없어서 4개의 성분 밀도 모두에서 무관성분이 없다. 결국 $\eta_1 = \eta_2 = 1$이 되어 4개의 성분은 온전하게 식별가능하다.

반면, Figure 3.1(b)의 경우를 보자. 이때 y_1과 y_2가 완전하게 정보적이어야만 즉 $\eta_1 = \eta_2 = 1$이어야만 4개의 성분을 온전하게 식별할 수 있다. 그러나 각 성분의 평균들은 y_1-축과 y_2-축에서 두 개씩 서로 동일 선상에 위치해 있다. 예를 들어, 성분 1은 변수 y_1에 대하여 성분 3을 $1 - \eta_1 = 1/2$로서 무관성분으로 하고 있으며, 변수 y_2에 대하여 성분 2를 $1 - \eta_2 = 1/2$로서 무관성분으로 하고 있다. 결국 어떤 한 성분은 나머지 두 개 성분을 각 축에서 $1 - \eta_k = 1/2\,(k = 1, 2)$으로서 무관성분으로 하고 있어서 $\eta_1 = \eta_2 = 1$일 수 없으므로 4개의 성분은 유일하게 식별할 수 없게 된다. Figure 3.1(b)에서 하나를 뺀 3개의 성분을 고려해도 비슷한 결과가 나타난다.

정리하여 말하면, p-차원 상에서 만약 어떤 성분 평균의 p개의 축 모두에 다른 성분이 걸쳐 있으면 그 성분들을 온전하게 식별할 수 없는 성질을 가지게 된다. 결국 두각 혼합모형은 표준적인 혼합모형과는 다르게, g개의 성분에 걸쳐 성분밀도의 모수들이 모두 다르더라도, 성분평균들의 위치에 따라 모형이 식별가능하지 않을 수 있다.

저자가 아는 한 지금까지 두각혼합모형을 다루는 모든 문헌에서 이 문제점에 대한 언급이 없다. 본 연구에서는 이 문제를 해결하기 위해 (비록 완전한 해법은 아닐지라도) 아주 간단한 해결방법을 제안하고자 한다. 그것은 무관분포의 평균을 $\eta_k = 1\,(k = 1, \ldots, p)$일 때의 모든 성분 평균들의 중심에 있다고 제약하는 것이다. 이 제약으로부터 간단히

$$\hat{\mu}_k = \frac{1}{g}\sum_{i=1}^{g}\left(\frac{\sum_{j=1}^{n}\tau_{ij}y_{jk}}{\sum_{j=1}^{n}\tau_{ij}}\right), \quad k = 1, \ldots, p \tag{3.1}$$

을 얻을 수 있다. 이렇게 하면 모든 유관성분들의 평균 위치는 어떠한 경우에도 무관분포의 평균위치와 동일 선상에 위치하지 않게 된다.

이제 제안된 방법이 추론 상의 문제가 없음을 설명하기 위해 먼저 식 (2.9)에서 $\hat{\eta}_k$를 살펴보자. k번째 변수가 완전하게 정보적이어서 $\hat{\eta}_k = \sum_{j=1}^{n}\sum_{i=1}^{g}\tau_{ij}\gamma_{ijk}/n = 1$이면, 이는 곧 모든 $\gamma_{11k} = \cdots = \gamma_{ijk} = \cdots = \gamma_{gnk} = 1$임을 의미한다.

k번째의 모든 $\gamma_{ijk} = 1$이면 식 (2.4)의

$$\begin{aligned}
\frac{\partial Q_4\left(\boldsymbol{\mu}_{[0]}, \boldsymbol{\sigma}_{[0]}^2|\widehat{\boldsymbol{\Theta}}\right)}{\partial\mu_{ik}} &\propto -\sum_{j=1}^{n}\sum_{i=1}^{g}\tau_{ij}(1-\gamma_{ijk})\frac{(y_{jk}-\mu_k)^2}{\sigma_k} = 0, \\
\frac{\partial Q_4\left(\boldsymbol{\mu}_{[0]}, \boldsymbol{\sigma}_{[0]}^2|\widehat{\boldsymbol{\Theta}}\right)}{\partial\sigma_{ik}^2} &\propto \sum_{j=1}^{n}\sum_{i=1}^{g}\tau_{ij}\left[(1-\gamma_{ijk})\sigma_{ik}^2 - (1-\gamma_{ijk})(y_{jk}-\mu_k)^2\right] = 0
\end{aligned} \tag{3.2}$$

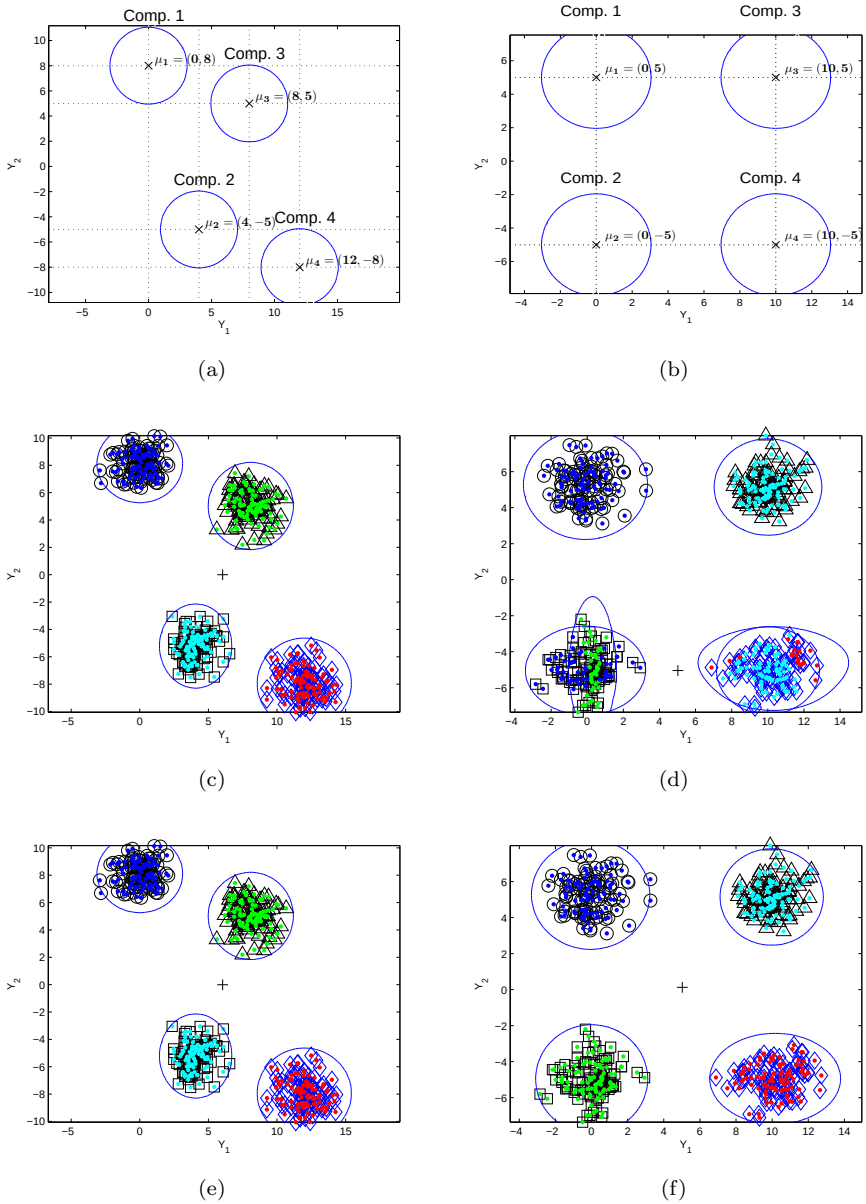


Figure 3.1. Identifications of Components: (a)–(b) Diagrams of components location, (c)–(d) Results of clustering by Law *et al.*'s method, (e)–(f) Results of clustering by proposed method. Different shapes indicate data points of true cluster, and different colored points indicate the results of clustering. Maker '+' indicates the location of the estimates of mean parameter in irrelevant density.

이므로 (μ_k, σ_k^2)에 대한 무한 개의 해가 존재하게 됨을 알 수 있다. 다시 말해서, 이 경우 해로서 어떠한 μ_k와 σ_k^2을 선택할 수 있는데, 그 중에 본 연구에서는 식 (3.1)의 $\hat{\mu}_k$으로 선택한 것 뿐이다. 그런데 우리의 해는 성분 평균들의 중심에 위치하므로 성분 평균의 식별성을 만족시키는 성질을 가진다. 그리고

$\hat{\sigma}_k^2$로서 0에 가까운 아주 작은 값으로 정할 것인데, 이렇게 함으로써 무관 성분밀도를 평균으로 퇴화시키는 형태를 취하게 하여 무관밀도의 형태를 손쉽게 알아 볼 수 있을 것이다.

Figure 3.1(c)와 (d) 그리고 Figure 3.1(e)와 (f)는 각각 성분별로 100개 씩 자료를 생성하여 Law 등 (2004)과 수정된 방법으로 적합한 결과를 나타낸 것이다. 생성된 자료는 군집별로 각각 다른 형태의 마커(원, 사각형, 삼각형, 다이아몬드)를 사용하여 나타내었고 군집의 결과는 각각 다른 색상의 점 (•)으로 표현하였다. Figure 3.1(c)와 (e)의 결과는 Law 등 (2004)의 방법이나 제안된 방법은 거의 동일하게 좋은 군집 결과를 보여주고 있다. 그러나 Figure 3.1(d)를 보면 Law 등 (2004)의 방법은 실패한 군집 결과를 보여주고 있는데, 그 이유는 '+'로 표시된 무관밀도의 평균 방향으로 군집들이 서로를 무관 성분으로 간주하기 때문 임을 잘 보여주고 있다. 반면, Figure 3.1(f)를 보면 제안된 방법은 매우 만족스러운 군집 결과를 제공하고 있다. 무관밀도의 평균 위치가 자료의 중심에 위치하여 4개의 군집들이 서로를 무관 성분으로 할 수 없도록 제약한 결과이다.

3.2. 변수 선택의 문제점

자료의 군집들이 잘 분리되어 있어도 자료의 산포가 어느정도 있는 경우 $\hat{\eta}_k$들은 확실하게 0이나 1로 잘 드러나지 않는 특성이 있다. 이를 개선하기 위해 Law 등 (2004)에서는 M-Step에서 $\boldsymbol{\eta}$에 대하여

$$J(\boldsymbol{\eta}) = -gR\sum_{k=1}^{p}\log\eta_k - S\sum_{k=1}^{p}\log(1-\eta_k)$$

의 벌점 (penalty)를 두고 추정하였다. 여기서 R과 S는 각각 유관밀도와 무관밀도에 포함된 모수의 개수로서 우리의 모형에서는 각각 2이다. 이로부터 얻은 개량된 추정치는

$$\hat{\eta}_k = \frac{\max\left\{\sum_{i=1}^{g}\sum_{j=1}^{n}\tau_{ij}\gamma_{ijk} - \frac{gR}{2}, 0\right\}}{\max\left\{\sum_{i=1}^{g}\sum_{j=1}^{n}\tau_{ij}\gamma_{ijk} - \frac{gR}{2}, 0\right\} + \max\left\{\sum_{i=1}^{g}\sum_{j=1}^{n}\tau_{ij}(1-\gamma_{ijk}) - \frac{S}{2}, 0\right\}} \tag{3.3}$$

이다. 사실 벌점항의 지수 $\exp\{J(\boldsymbol{\eta})\} = \prod_{k=1}^{p}\eta_k^{-gR}(1-\eta_k)^{-S}$은 완비자료에 대한 $\boldsymbol{\eta}$의 우도의 공액 사전분포이다. 따라서 자료의 개수가 충분히 클 때 식 (3.3)의 정확성을 보장한다고 보아야 한다. 군집들이 잘 분리되어 있더라도 군집당 자료의 개수 작은 경우(예를 들어 개수가 30 이하)에는 여전히 식 (3.3)의 $\hat{\eta}_k$들이 0 혹은 1로 잘 부각되지 않는다. 이 경우 저자는 $c_1 = gR$과 $c_2 = S$를 조율 상수로 취급하여 c_1, c_2를 다양하게 조율하면 종종 만족스러운 결과를 얻을 수 있었다. 그러나 이 방법은 매우 자료 종속적이어서 자료가 조금만 달라져도 많은 시행착오를 거쳐 조율상수를 다시 수정해야하는 어려움 때문에 일반적으로 적용하기에 어려움이 있다. 더욱이 조율상수가 2개여서 상당한 번거로움을 피할 수 없다.

본 연구에서는 우선 두 개의 조율상수 c_1과 c_2를 $c_1 = c_2 = c$와 같이 1개로 줄이고, 결과가 c에 크게 민감하지 않은 방법을 찾고자 한다. 이를 위해 표본의 크기에 민감하지 않으면서 다만 변수가 정보적이면 아주 크게 그리고 비정보적이면 아주 작게 나타나는 다음 통계량의 사용을 제안한다. 그것은

$$F_k = \frac{\sum_{i=1}^{g}\hat{\pi}_i(m_{ik} - \bar{m}_k)^2}{\sum_{i=1}^{g}\hat{\pi}_i\left[\hat{\eta}_k\frac{\hat{\sigma}_{ik}^2}{\hat{n}_i} + (1-\hat{\eta}_k)\frac{\hat{\sigma}_k^2}{\hat{n}_i}\right]}, \quad k = 1, \ldots, p \tag{3.4}$$

인데, 여기서 $m_{ik} = \hat{\eta}_k\hat{\mu}_{ik} + (1-\hat{\eta}_k)\hat{\mu}_k$, $\bar{m}_k = \sum_{i=1}^{g}\hat{\pi}_i m_{ik}$ 그리고 $\hat{n}_i = \sum_{j=1}^{n}\tau_{ij}$를 나타낸다. 통계량 F_k는 k번째 변수의 성분 평균 $\mu_{1k}, \ldots, \mu_{gk}$들이 모두 같다는 귀무가설 하에서의 검정 통계량으로부터 아이디어를 얻은 것이다. 그러나 $\boldsymbol{Y}_j$의 분포는 정규분포 2-성분 혼합 곱의 g-성분 혼합이므로 k번째 원소 Y_{jk}가 결코 정규분포를 따르지 않는다. 따라서 F_k가 F-분포를 따른다고 볼 수는 없다. 그러나 대략 어떤 F_0 값에 대해 $F_k < F_0$이면 변량 Y_k를 비정보적 변수으로 판단할 수 있을 것이다.

3.3. 조율 상수 사용에 대하여

제안된 방법에서 조율상수는 c와 F_0로서 여전히 2개 이다. 그러나 본 연구의 모든 실험에서는 $F_0 = 5$를 고정하여 사용할 것이다. 본 논문에 기록하지 않은 다양한 상황에서 많은 실험을 통해 가장 적절하다고 판단하였기 때문이다. 사실 분산분석에서 자유도를 고려하지 않을 때 F-통계량 값 5는 귀무가설을 분명하게 기각하기에 꽤 모호한 값이다.

이제 우리는 상수 c를 조율하면 된다. 저자의 경험으로 단일 조율상수 c을 사용하면 두 개의 조율상수 c_1과 c_2을 사용하였을 때보다 훨씬 덜 민감한 결과를 제공한다. 보통 $1 \le c \le 4$ 사이에서 사용하되 보통의 상황이라면 $c = 2$를 추천한다. 다음 절의 실험에서는 특별한 언급이 없는 한 $c = 2$로 고정하여 실험할 것이다.

이제 본 연구에서 다음과 같은 순환적(recursive) EM-알고리즘을 제안한다.

(1) 초기 설정

- 조율상수 c를 정한다.
- 초기치 $\boldsymbol{\Theta}_0$를 정한다.

(2) 순환적 EM-알고리즘

- E-Step: τ_{ij} 및 γ_{ijk}를 계산한다.
- M-Step: 식 (2.9)의 모든 추정치들를 계산하되, $\hat{\mu}_k = (1/g)\sum_{i=1}^{g}(\sum_{j=1}^{n}\tau_{ij}y_{jk}/\sum_{j=1}^{n}\tau_{ij})$ 및 $\hat{\sigma}_k^2 \approx 0$으로 계산한다.
- E-Step과 M-Step을 연속된 로그-우도의 값의 차이가 10^{-7} 이하가 될 때까지 충분히 반복한다.

(3) 비정보적 변수 식별

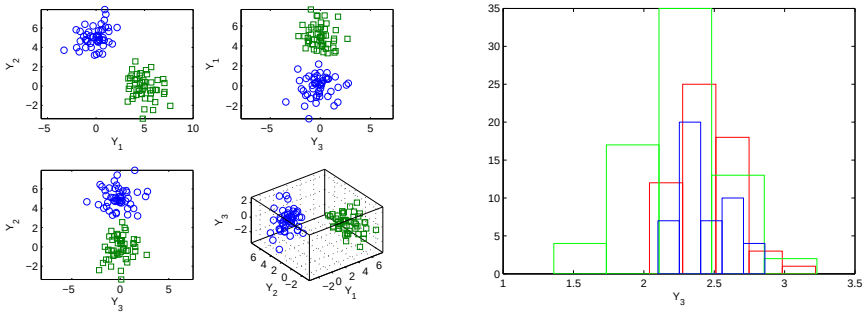
- $K = \{k : F_k < F_0\}$를 만족하는 변수 집합을 구한다.
- 만약 K가 공집합이 아니면, 모든 $k \in K$에 대해 $\eta_k = 0$으로 치환하고 (2)로 돌아간다.
- 만약 K가 공집합이면 순환적 알고리즘을 종료한다.

3.4. 성분의 개수 결정

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

$$-2\log L\left(\widehat{\boldsymbol{\Theta}}\right) + \ell\log(n) \tag{3.5}$$

을 최소로 하는 g로서 결정한다. 여기서 $\log L(\widehat{\boldsymbol{\Theta}})$은 $\widehat{\boldsymbol{\Theta}}$에서 계산된 식 (2.2)의 로그-우도 값이며 ℓ은 자유모수의 개수를 나타낸다. 우리의 모형에서 $\eta_1, \ldots, \eta_p$ 중에서 0이 아닌 개수가 q라 할 때, g개 성분에서 추정해야 할 자유모수의 개수는 $\ell = (g-1) + 2gq + 2q + q$이 된다.



(a) (b)

Figure 4.1. (a) Plots of simulated data when the size of cluster is 50. Different shapes indicate data points of true cluster. Y_1 and Y_2 are informative to identify two clusters while Y_3 is noninformative. (b) Histogram of three cultivars of Y_3 for wine data set.

4. 실험

이 절에서는 모의자료와 실자료를 이용하여 제안 방법의 실효성을 보일 것이다.

4.1. 모의자료 실험

본 실험에서는 5변량 관측치 $\boldsymbol{y}_j = (y_{1j}, \ldots, y_{5j})^T$으로 이루어진 잘 분리된 2개의 군집을 생성하되 각각의 크기가 100, 50, 30, 15인 경우에 대해 실험한다. 그리고 처음 두 변량 Y_1, Y_2는 정보적이며, 나머지 세 변량 Y_1, Y_2, Y_3는 비정보적 변량으로 하였다. 첫 번째 군집과 두 번째 군집의 평균은 각각 $\boldsymbol{\mu}_1 = (0, 5, 0, 0, 0)^T$ 및 $\boldsymbol{\mu}_1 = (5, 0, 0, 0, 0)^T$으로 했으며 분산은 $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{I}_5$으로 하여 자료를 생성하였다.

Figure 4.1(a)에 군집의 크기가 50일 때의 (Y_1, Y_2, Y_3)에 대하여 산점도를 예로써 나타내었다. 그림에서 보는 바와 같이 변량 Y_1, Y_2가 정보적인 반면 변량 Y_3는 비정보적이다. 또한 정보적 변량인 (Y_1, Y_2) 공간 상에서 두 군집의 평균 위치가 동일 선상에 위치하지 않토록 하여 Law 등 (2004)의 기법이 앞 절에서 지적한 식별성의 문제가 발생하지 않토록 하였다. 이렇게 한 이유는 오직 표본의 크기에 따른 Law 등 (2004)의 기법의 정확성을 검토하기 위함이다.

모의 실험은 각 경우에 대해 50번 씩 반복하였으며 Table 4.1에 얻은 결과들의 평균과 표준편차를 제공하였다. 군집의 크기가 100 이상일 때는 제안된 방법이나 Law 등 (2004)의 방법은 거의 차이 없이 정확하게 정보적 변수와 비정보적 변수를 식별하며 군집 오류율도 매우 낮게 제공하고 있다. 그러나 군집의 크기가 작아질수록 Law 등 (2004)의 기법은 점점 나쁜 결과를 보이고 있는데, 특히 크기가 15일 때 두각모수 추정치의 평균이 대략 0.5이고 표준편차 역시 대략 0.5인 것으로 보아, 5개의 두각성 추정치 $\hat{\eta}_i$ $(i = 1, \ldots, 5)$들이 모두 대략 0 혹은 모두 1로 얻어진 것으로 보인다. 따라서 매우 나쁜 군집 식별력을 제공할 수 밖에 없는데, 실제 오류율의 평균은 50%나 되는 것으로 나타났다.

반면 제안된 방법은 군집의 크기와 상관없이 거의 확실하게 정보적 변수와 비정보적를 식별하며, 오류율 역시 작은 군집에서도 급격히 나빠지는 결과를 보이지 않음을 알 수 있다.

4.2. 실자료 실험

4.2.1. Wine data set $n = 178$의 자료로 wine 데이터 세트는 $g = 3$개의 품종의 포도를 구분하기 위해 구성성분 $p = 13$의 변량 (Alcohol, Malic acid, Ash, Alcalinity of ash, Magnesium, Total

Table 4.1. Estimates of Saliency parameters and Error rates. Means and standard deviations in the parenthesis through 50 repetitions of experiments.

Size of Cluster	Method	Saliency Parameters					Error Rate
		$\hat{\eta}_1$	$\hat{\eta}_2$	$\hat{\eta}_3$	$\hat{\eta}_4$	$\hat{\eta}_5$	
100	Proposed	0.9992	0.9998	0	0	0	0.0004
		(0.0038)	(0.0052)	(0.0000)	(0.0000)	(0.0000)	(0.0020)
	Law *et al.*'s	0.9991	0.9984	0	0	0	0.0004
		(0.0040)	(0.0058)	(0.0000)	(0.0000)	(0.0000)	(0.0020)
50	Proposed	0.9993	0.9998	0	0	0	0.0004
		(0.0053)	(0.0080)	(0.0000)	(0.0000)	(0.0000)	(0.0028)
	Law *et al.*'s	0.9789	0.9774	0	0	0	0.0204
		(0.9789)	(0.9774)	(0.0000)	(0.0000)	(0.0000)	(0.1414)
30	Proposed	**0.9985**	**0.9989**	0	0	0	**0.0013**
		(0.0062)	(0.0060)	(0.0000)	(0.0000)	(0.0000)	(0.0066)
	Law *et al.*'s	**0.7384**	**0.7375**	0.0183	0	0.0200	**0.2613**
		(0.4422)	(0.4417)	(0.1291)	(0.0000)	(0.1414)	(0.4423)
15	Proposed	**0.9800**	**0.9755**	0	0	0	**0.0213**
		(0.1414)	(0.1417)	(0.0000)	(0.0000)	(0.0000)	(0.1415)
	Law *et al.*'s	**0.4729**	**0.4808**	0	0	0.0027	**0.5080**
		(0.4985)	(0.4917)	(0.0000)	(0.0000)	(0.0191)	(0.4954)

Table 4.2. Estimates of Saliency Parameters

Method	Estimates of Saliency Parameter												
	$\hat{\eta}_1$	$\hat{\eta}_2$	$\hat{\eta}_3$	$\hat{\eta}_4$	$\hat{\eta}_5$	$\hat{\eta}_6$	$\hat{\eta}_7$	$\hat{\eta}_8$	$\hat{\eta}_9$	$\hat{\eta}_{10}$	$\hat{\eta}_{11}$	$\hat{\eta}_{12}$	$\hat{\eta}_{13}$
Proposed	0.93	0.72	**1.0**	1.0	0.89	0.97	1.0	0.71	0.95	0.89	1.00	1.0	0.85
Law *et al.*'s	0.99	0.79	**0.0**	1.0	0.84	0.94	1.0	0.66	0.93	0.81	0.86	1.0	0.89

phenols, Flavanoids, Nonflavanoid phenols, Proanthocyanins, Color intensity, Hue, OD280/OD315 of diluted wines, Proline)을 사용한다.

제안된 방법의 BIC = 7017.9로서 $g = 3$을 추천하였다. 그리고 Law 등 (2004)의 기법에 대해서는 BIC를 고려하지 않고 군집의 개수를 3으로 할 것이다. 이로써 Law 등 (2004)의 기법에 보다 유리한 상황을 만들었다.

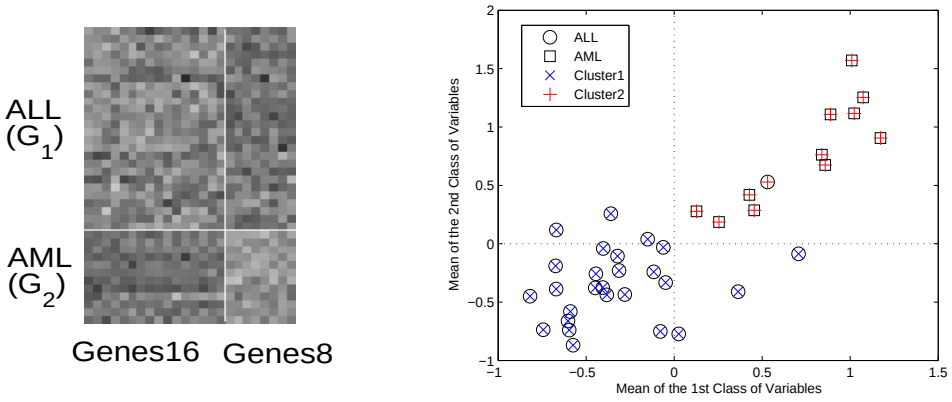
Table 4.2에 두 방법으로 두각모수들의 추정치를 제공하였다. 두 방법 모두 거의 비슷한 결과를 나타내었다. 다만 제안된 방법은 $\hat{\eta}_3 = 1$로서 변량 Y_3 (Ash)를 완전한 정보적 변수로 취급하고 있는 반면 Law 등 (2004)의 기법은 $\hat{\eta}_3 = 0$으로서 비정보적 변수로 추정하고 있다. Figure 4.1(b)에 제공한 변량 Y_3 자료의 군집별 히스토그램을 보면 3개의 참 군집의 평균이 각각 $\bar{y}_1 = 2.4556$, $\bar{y}_2 = 2.2448$ 및 $\bar{y}_3 = 2.4371$로서 뚜렷하게 차이를 보이지 않아서 Law 등 (2004) 기법의 주장이 일견 타당해 보이지만 실제 참군집에 대한 Y_3 변량의 ANOVA 검정을 해보면 검정통계량 값은 13.31이고, 비모수적으로도 Kruskal-Wallis 검정통계량 값은 23.13로서 두 경우 모두 p-value ≈ 0이다. 따라서 3 군집의 분포가 같다는 귀무가설은 거의 사실이 아니므로 제안된 방법의 주장이 더 설득력이 있다.

한편 두 방법의 오류 분할표를 Table 4.3에 제공하였다. 제안된 방법이 좀 더 정확한 군집 할당을 보이고 있는데, 그 이유는 Y_3를 정보적 변수로 취급하는가 혹은 그렇지 않은가의 차이인 것으로 볼 수 있다.

McLachlan과 Peel (2000, Chapter 8, p.255)에 변량들의 종속성을 고려한 표준적인 정규혼합모형을 적합하여 분할한 결과를 제공하고 있는데 오류의 개수가 7개 이다. 정규분포기반 두각 혼합모형이 표준

Table 4.3. Error Table

Cluster	Proposed			Law *et al.*'s		
	1	2	3	1	2	3
1	58	**1**	0	58	**1**	0
2	0	70	**1**	0	68	**3**
3	0	0	48	0	0	48
# of error		2			4	



(a) (b)

Figure 4.2. (a) Gray level hit map of reordered 24 informative genes. (b) Plot of averages for two gene groups.

적인 정규혼합 모형보다 더 좋은 결과를 보인 이유는 보다 적은 모수를 사용하므로 과적합 문제에서 좀 더 자유롭고, 이 자료의 경우 변량들의 조건부 독립성이 잘 만족하고 있기 때문일 것이다.

4.2.2. Golub's data set 본 실험에서는 Golub 등 (1999)의 백혈병 마이크로어레이 유전자발현 자료에 대해 제안된 방법의 성능을 보인다. 이 자료는 27개의 ALL(acute lymphoblastic leukemia) 표본과 11개의 AML(acute myeloid leukemia) 표본 (합계 $n = 38$개)과 7139개의 유전자 (변수)로 이루어져 있다. 이 중에서 우리는 Kim (2011)의 방법을 선행하여 60개의 유전자를 선별하였다. 이때 군집 오류의 개수는 ALL을 AML로 잘 못 분류한 1개 (표본번호 2번) 였다. 이제 우리는 60개의 유전자 중 가능한 한 적은 개수의 정보적인 유전자를 선별하여 38개의 표본을 BIC가 최소가 되도록 군집하는 것이다. 원자료는 스케일이 너무 커서 다루기 불편함으로 실험하기 전에 우선 마이크로어레이 자료행렬에 대해 행 (표본)표준화 후 다시 열 (유전자)표준화하였다. 이렇게 하여도 제안된 방법의 일반성을 잃지 않는다.

저자는 여러번의 사전 시행을 통해 조율상수 $c = 2$ 대신에 $c = 3.3$이 적합다고 보고 제안된 방법을 수행하였다. 그 결과 BIC $= 2526.2$로서 $g = 2$를 추천하였다. 이것은 참 군집의 개수와 일치한다. 이때 오류의 개수는 1개 (표본번호 2번) 그대로 였지만 정보적 변수는 24개로서 비정보적 변수를 무려 36개나 줄일 수 있었다. 할당 오류를 나타낸 1개의 표본을 제외하면 두 군집 G_1, G_2가 ALL과 AML 집단과 거의 일치하고 있다.

이 24개의 정보적 변수 (즉 정보적 유전자)에 대한 유전자 프로파일을 평균값이 큰 순위로 Figure 4.2(a)에 gray level map (흰색: 고발현, 검정색: 저발현)으로 나타내었다. 시각적으로도 분명하게

Gene16과 Gene8은 두 집단 ALL과 AML에 대해 양성과 음성 반응을 보이고 있다. 따라서 두 유전자군의 평균은 ALL과 AML을 진단하는데 유용할 것이다. Figure 4.2(b)는 두 변수군의 평균으로 ALL과 AML 표본 집단의 산점도를 나타낸 것이다. 이로부터 만약 두 유전자군의 평균이 제1사분면에 속하면 ALL로 그리고 제3사분면에 속하면 AML로 진단할 수 있을 것이다.

한편, Law 등 (2004)의 기법은 41개의 변수를 선별하였고 이때 오류는 2개 였다. 그리고 $g = 2$에서 BIC = 2959.0으로서 최소였는데, 선택된 정보적 변수의 개수, 오류율 및 적합도에서 제안된 방법보다 열위에 있다. 또한 L_1 벌점화 우도 접근법을 사용하는 Wang과 Zhu (2008)는 이 자료에 대해 25개 정보적 유전자를 얻어 1개의 할당오류를 보였다. 이로써 제안된 방법이 그들의 방법보다 효과적인 변수선택이 이루어졌음을 알 수 있다.

5. 결론 및 토의

본 논문에서는 정규분포기반 두각 혼합모형(normal distribution-based salient mixture model)의 사용에 대해 다루었다. 이 모형은 Law 등 (2004) 이후 인공지능 및 패턴인식 등 무감독형 학습(unsupervised learning)에서 군집과 특징 선택을 위해 폭넓게 사용되고 있다. 그러나 본문에서 지적했던 바와 같이 모형 자체에 몇 가지 문제를 가지고 있다. 즉, 모형의 성분들이 특별한 위치에 있을 때 성분의 식별성 문제에 문제가 있으며, 기존에 사용된 EM 알고리즘으로는 군집의 크기가 작을 때 변수선택이 잘 이루어지지 않는다는 것이다. 본 논문에서는 이 문제를 수정하여 개량된 순환적 알고리즘을 제안하였다. 모의자료 실험과 실자료 실험을 통해 제안된 방법이 기존의 방법보다 우수함을 보였다.

두각혼합모형은 본 논문에서 지적한 문제점 말고도 다양한 문제가 있을 수 있다. 그 중 하나는 변수들이 독립일 때 다변량 밀도가 단변량 밀도의 곱으로 표현되는 모형에 대해서만 사용할 수 있다는 것이다. 따라서 이것이 성립하지 않는 다변량 t-분포는 사용될 수 없다. t-분포 모형을 적용할 수 없다는 것은 매우 큰 취약점이라 할 수 있다. 관련 문헌을 고찰해 보건데, 두각혼합모형은 주로 공학도들에 의해 발전되어왔다. 그러한 이유인지는 몰라도 지금까지 이루어온 성과에 비해 이론적 엄밀성이 다소 취약하다. 이 모형에 대해 통계학자들의 관심이 모아졌으면 한다.

References

Bouguila, N., Almakadmeh, K. and Boutemedjet, S. (2012). A finite mixture model for simultaneous high-dimensional clustering, localized feature selection and outlier rejection, *Expert Systems with Applications*, **39**, 6641–6656.

Bouguila, N. and Ziou, D. (2006). A hybrid SEM algorithm for high-dimensional unsupervised learning using a finite generalized Dirichlet mixture, *IEEE Transactions on Image Processing*, **15**, 2657–2668.

Boutemedjet, S., Bouguila, N. and Ziou, D. (2009). A hybrid feature extraction selection approach for high-dimensional non-Gaussian data clustering, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **31**, 1429–1443.

Elguebaly, T. and Bouguila, N. (2013). Simultaneous Bayesian clustering and features election using RJMCMC-based learning of finite generalized Dirichlet mixture models, *Signal Processing*, **93**, 1531–1546.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A. and Bloomfield, C. D. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring, *Science*, **286**, 531–537.

Graham, M. W. and Miller, D. J. (2006). Unsupervised learning of parsimonious mixtures on large spaces with integrated feature and component selection, *IEEE Transactions on Signal Processing*, **54**, 1289–1303.

Kim, S. G. (2011). Variable selection in normal mixture model based clustering under heteroscedasticity, *The Korean Journal of Applied Statistics*, **24**, 1–12.

Law, M. H. C., Figueiredo, M. A. T. and Jain, A. K. (2004). Simultaneous feature selection and clustering using mixture models, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **26**, 1154–1166.

Li, M. D. Y. and Hua, J. (2008). Localized feature selection for clustering, *Pattern Recognition Letters*, **29**, 10–18.

Li, Y., Dong, M. and Hua, J. (2009). Simultaneous localized feature selection and model detection for Gaussian mixtures, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **31**, 953–960.

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*, Wiley, New York.

Pan, W. and Shen, X. (2006). Penalized model-based clustering with application to variable selection. *Journal of Machine Learning Research*, **8**, 1145–1164.

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

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

Xie, B., Pan, W. and Shen, X. (2008). Variable selection in penalized model-based clustering via regularization on grouped parameters, *Biometrics*, **64**, 921–930.

정규분포기반 두각 혼합모형의 순환적 적합을 이용한 군집분석에서의 변수선택

김승구[a,1]

[a]상지대학교 컴퓨터데이터정보학과

(2013년 8월 26일 접수, 2013년 10월 18일 수정, 2013년 10월 21일 채택)

요 약

Law 등 (2004)은 군집분석에서 변수선택을 위해 정규분포기반 "두각 혼합모형(salient mixture model)"의 사용을 제안하였다. 본 논문에서는 이 모형의 적합 상의 문제점과 변수선택의 결함을 지적하고 그 대안을 제시한다. 모의자료와 실자료를 바탕으로 제안된 방법이 기존의 방법보다 유용함을 보였다.

주요용어: 두각 모수, 변수선택, 군집분석, 정규혼합모형, EM 알고리즘.

[1](220-702) 강원도 원주시 우산동 83, 상지대학교 이공대학 컴퓨터데이터정보학과, 교수.
E-mail: sgukim@sangji.ac.kr