응용통계연구 (2011)
24(6), 1213–1224
DOI: http://dx.doi.org/10.5351/KJAS.2011.24.6.1213

이분산 상황 하에서 정규혼합모형 기반 군집분석의 변수선택

김승구[1]

[1]상지대학교 컴퓨터데이터정보학과

(2011년 9월 접수, 2011년 9월 채택)

요 약

관측치의 개수보다 변량의 개수가 더 많은 다변수 상황에서 정규혼합모형을 이용하여 군집분석을 하기 위해서는 비정보적인 변수들을 제거하는 과정이 필수적으로 요구된다. 이와 같은 변수선택과 군집의 동시 처리를 위한 기존 연구의 대부분은 군집별 등분산 가정 하에서 이루어져 왔으며, 비정보적인 변수를 제거하기 위해 주로 벌점화 우도 기법이 이용되었다. 본 연구에서는 약간 변형된 정규혼합모형을 기반으로 비현실적인 등분산 가정을 탈피하면서 효율적으로 비정보적인 변수를 제거하는 새로운 방법을 제공한다. 이 모형에 대한 타당성을 설명하였고, 모수 추정을 위한 EM 알고리즘을 유도하였다. 그리고 모의실험 및 실자료 실험을 통해 제안된 방법의 유효성을 보였다.

주요용어: 정보적 변수, 변수선택, 군집분석, EM 알고리즘, 마이크로어레이 유전자 발현.

1. 서론

군집분석은 많은 분야에서 중요한 도구로 사용되는 통계학적 도구라 할 수 있다. 특히 마이크로어레이 유전자 발현 자료분석에서는 질병에 대한 유전자 경로를 발견하거나 이해하는데 큰 역할을 한다 (Kim, 2006; McLachlan 등, 2006). 이런 응용분야의 특징은 흔히 소수의 관측치와 다수의 변수로 이루어진다. 그러나 다행히 이러한 응용문제에서 변수들 중 상당수는 군집분석에 기여하지 않는 경우가 많다. 예를 들어 그림 1.1에서 변수 y_1에 대해서 3개의 군집은 각 군집 평균 $\bar{y}_1, \bar{y}_2, \bar{y}_3$들이 차이를 보이며 잘 분리되어 있어 정보적(informative)이지만, 변수 y_2에서는 군집평균의 차이가 거의 없어 비정보적(noninformative)이다. 이와 같이 본 연구에서 변수의 군집에 대한 정보력은 오직 군집의 평균 차이에 기인하는 것으로 한정할 것이다.

g-성분 정규혼합모형(normal mixture model; NMM) 기반 군집분석의 경우, 모형을 적합하는 중에 p개의 변수들 중에서 이러한 비정보적인 변수를 제외시킴으로써 자동적으로 정보적인 변수들을 선택하게 하여 관측치들을 군집하는 방식으로 이루어진다. Raftery와 Dean (2006)은 회귀분석의 단계별(stepwise) 변수선택과 유사한 방법을 사용하여 이 문제에 접근하였다. 그러나 이 방법은 다변수 상황 하에서는 비실용적이라고 비판을 받았다. 보다 실용적인 접근법으로서 최근 대표적인 연구는 Pan과 Shen (2006)의 벌점화 우도(penalized likelihood) 추정 군집 기법이다. 이 방법은 (사전에 변수 자료별 중심화가 되어 있다는 전제 하에) 모평균 $\mu_{ik}\,(i = 1, \ldots, g;\ k = 1, \ldots, p)$에 대해 그리고 어떤 벌점상수 $\lambda > 0$에 대해 우도에 벌점항

$$\lambda \sum_{i=1}^{g} \sum_{k=1} |\mu_{ik}| \tag{1.1}$$

본 연구는 상지대학교 2010년 교내연구비 지원에 의해 수행되었음.
[1](220-702) 강원도 원주시 우산동 83, 상지대학교 컴퓨터데이터정보학과, 교수. E-mail: sgukim@sangji.ac.kr

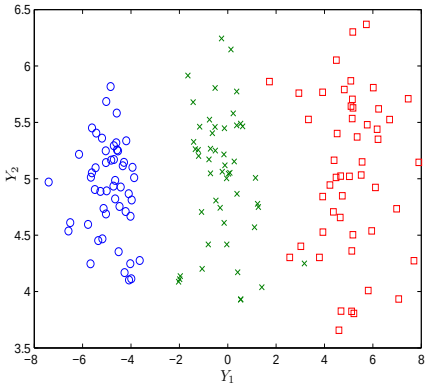

그림 1.1. 정보적 변수와 비정보적 변수

을 두어 최우추정법으로 적합한다. 이때 결과적으로 비정보적인 변수의 평균 추정치 $\hat{\mu}_{ik}$는 적절히 크게 주어진 벌점상수 λ에 대해 정확히 0이 되어 군집판별에 영향을 미치지 못하게 한다. 그러나 이 방법의 단점은 비정보적 변수 y_k의 g개의 군집 평균 추정치 $\hat{\mu}_{1k}, \ldots, \hat{\mu}_{gk}$가 동시에 0이 되지 못하는 경향이 있다. 이와 같은 "개별적 축퇴(shrinkage)"라는 부자연스러운 이 특성은 사후에 y_k가 비정보적인 변수인지를 판정하기에도 어려운 점이 있다. 이러한 문제점을 극복하기 위해 Wang과 Zhu (2008)은 벌점항으로 $\lambda \sum_{k=1}^{p} \max_{\{1,\ldots,g\}} |\mu_{ik}|$의 사용을 제안하였다. 만약 가장 큰 $\hat{\mu}_{ik}$가 0이면 나머지 $(g-1)$개의 추정치들도 자동적으로 0이 되도록 하려는 의도이다. 그러나 위 두 방법은 각 군집에 대응하는 성분의 분산이 $\sigma_{1k}^2 = \cdots = \sigma_{gk}^2 \stackrel{\text{let}}{=} \sigma_k^2 = 1$과 같이 모든 변수들에 대해 군집에 걸친 등분산 가정을 하고 있다. 이 가정은 비현실적으로서 특별한 경우 외에는 상당한 사용의 제약을 받을 수 있다. 그래서 Xie 등 (2008)은 이분산 모형 하에서 (사전에 변수 자료별 표준화가 되어 있다는 전제 하에) 두 개의 벌점항으로

$$\lambda_1 \sum_{k=1}^{p} |\mu_{ik}| + \lambda_2 \sum_{k=1}^{p} |\sigma_{ik}^2 - 1| \tag{1.2}$$

을 사용할 것을 제안하였다. 그러나 이 방법 역시 개별적 축퇴의 특성에서 자유롭지 못하며, 2개의 벌점상수 λ_1과 λ_2를 동시에 조율해야 한다는 불편함을 가지고 있다. 이상의 방법론들은 벌점항의 설계를 바탕으로 (EM 알고리즘을 통해) 벌점화 우도를 최대화는 추정법이라는 특징이 있다.

본 논문에서는 벌점화 우도 기법을 탈피하여, 특정 변수가 정보적일 확률을 정의하고 이를 통해 모형을 통제하는 간단한 착상을 통해 개별적 축퇴 문제와 이분산 문제를 극복하면서 효과적으로 변수선택과 군집을 통시에 수행하는 방법을 제시한다. 다음 절에서는 공식을 사용하여 비정보적 변수의 특징을 좀 더 구체적으로 설명하며, 제안된 모형을 제시하고 그 특징을 살펴본다. 3절에서는 제안 모형을 적합하기 위한 대안적 EM 알고리즘을 제공하며, 4절에서는 모의실험을 통해 제안된 방법의 유효성을 보인다. 5절에서는 결론과 추가적 연구돼야 할 사항을 논의한다.

2. 기본 개념과 모형 제안

2.1. 전제 및 가정들

n개의 p-변량 표본자료들이 관측되었고, 각 변량 자료별로 표준화되어 있다고 하자. 그리고 관측치 $\boldsymbol{y}_j = (y_{j1}, \ldots, y_{jp})^T \, (j = 1, \ldots, n)$은 g개의 "참"군집 중 하나로부터 왔다고 하자. 이를 모형화하여

$\boldsymbol{y}_j$는 g-성분 유한 NMM

$$f(\boldsymbol{y}_j) = \sum_{i=1}^{g} \pi_i f_i\left(\boldsymbol{y}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \boldsymbol{\eta}\right) = \sum_{i=1}^{g} \pi_i \phi\left(\boldsymbol{y}_j; \boldsymbol{\mu}_i^*, \boldsymbol{\Sigma}_i^*\right) \tag{2.1}$$

으로부터 독립적으로 관측된 관측치라 하자. 여기서 π_i들은 $0 < \pi_i < 1$이며 $\sum_{i=1}^{g} \pi_i = 1$을 만족하는 혼합비율이고, $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{ip})^T$는 i번째 군집을 특성화하는 p-변량 정규분포의 평균벡터이며, $\boldsymbol{\Sigma}_i$는 대응하는 분산-공분산행렬로서 이 논문에서는

$$\boldsymbol{\Sigma}_i = \mathrm{diag}\left(\sigma_{i1}^2, \ldots, \sigma_{ip}^2\right), \quad i = 1, \ldots, g$$

을 가정한다. 즉, 관측치 $y_{j1}, \ldots, y_{jp}$들은 서로 무상관이며, 각 군집에서 서로 다른 대각 분산-공분산행렬을 가진다고 가정한다. 이것은 Pan과 Shen (2006) 및 Wang과 Zhu (2008)에서 군집에 걸쳐 등분산을 가정한 경우와 차별된다. 그리고 $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_p)^T$이며, η_k는 변수 y_k가 정보적 변수일 확률로서

$$\mu_{ik}^* = \eta_k \mu_{ik} + (1 - \eta_k)(0) \quad \text{및} \quad \sigma_{ik}^{*2} = \eta_k \sigma_{ik}^2 + (1 - \eta_k)(1), \quad k = 1, \ldots, p \tag{2.2}$$

와 같이 y_k의 평균과 분산에 결합되어 있는 0과 1사이의 상수라 정의한다. 그리고 식 (2.1)의 $\phi(\boldsymbol{y}_j; \boldsymbol{\mu}_i^*, \boldsymbol{\Sigma}_i^*)$는 식 (2.2)의 평균과 분산 $\boldsymbol{\mu}_i^*, \boldsymbol{\Sigma}_i^*$을 가지는 다변량 정규분포를 나타내며, 그리고 y_k가 정보적 변수일 확률 η_k는 본 연구의 목적상 주어진 어떤 값 $\beta\,(> 0)$에 대해

$$\begin{aligned}
\eta_k &= 1 - \Pr\left\{|\bar{y}_{1k}| \le \beta, \ldots, |\bar{y}_{gk}| \le \beta\right\} \\
&= 1 - \Pr\left\{|\bar{y}_{1k}| \le \beta\right\} \times \cdots \times \Pr\left\{|\bar{y}_{gk}| \le \beta\right\} \\
&= 1 - \prod_{i=1}^{g} \left[\Phi\left(\frac{\beta - \mu_{ik}^*}{\sigma_{ik}^*/\sqrt{n_i}}\right) - \Phi\left(\frac{-\beta - \mu_{ik}^*}{\sigma_{ik}^*/\sqrt{n_i}}\right)\right]
\end{aligned} \tag{2.3}$$

와 같이 정의하면 적절할 것이다. 그 이유는, $\bar{y}_{ik}$는 i번째 군집에서 변수 y_k의 표본평균인데, 식 (2.3)의 첫 번째 식은, 만약 y_k가 정보적 변수라면 g개의 군집평균 절대값 $|\bar{y}_{1k}|, \ldots, |\bar{y}_{gk}|$들 중 최소한 1개는 어떤 β 값보다 큰 값일 것이기 때문이다. 두 번째 식은, 관측치들이 독립이라면 g개로 분할된 관측치 그룹들도 역시 독립이기 때문이며, 세 번째 식은 $\bar{y}_{ik} \sim \mathcal{N}(\mu_{ik}^*, \sigma_{ik}^{*2}/n_i)$을 따르기 때문이다. 단, n_i는 i번째 군집의 관측치 개수이다.

마지막으로, 앞에서도 그러했듯이 앞으로도 본 논문에서는 확률변수와 그 실현치를 나타낼 때 대소문자를 구분하지 않고 모두 소문자로 표기할 것이다. 그리고 식 (2.1)의 혼합모형의 "성분(component)" $\pi_i \phi(\boldsymbol{y}_j)$와 이것의 랜덤표본인 자료들의 "군집(cluster)"은 엄격히 다르지만 본 논문에서는 두 단어를 동의어처럼 사용할 것이다. 따라서 상황에 따른 독자들의 어느 정도의 분별이 요구된다.

2.2. 비정보적 변수에 대한 식별 원리

식 (2.1)의 NMM 하에서 관측치 $\boldsymbol{y}_j$가 i번째 참군집으로부터의 표본일 사후확률은

$$\tau_{ij} = \frac{\pi_i \prod_{k=1}^{p} \phi\left(y_{jk}; \mu_{ik}^*, \sigma_{ik}^{*2}\right)}{\sum_{h=1}^{g} \pi_h \prod_{k=1}^{p} \phi\left(y_{jk}; \mu_{hk}^*, \sigma_{hk}^{*2}\right)}, \quad i = 1, \ldots, g;\ j = 1, \ldots, n \tag{2.4}$$

로서 주어지는데, $\tau_{1j}, \ldots, \tau_{gj}$ 중 가장 큰 것이 $\tau_{i^0 j}$라 한다면 관측치 $\boldsymbol{y}_j$을 i^0번째 군집에 할당한다. 이 때 만약 예를 들어 1번째 변수가 비정보적이어서 $\eta_1 = 0$이면, $\mu_{i1}^* = 0$이 $\sigma_{i1}^{*2} = 1$이 되고 대응하는 밀도

$\phi(y_{j1};0,1)$은 군집에 대한 첨자를 상실해서 식 (2.4)의 분자/분모 항에서 약분되어 소거된다. 즉, τ_{ij}를 계산하는데 전혀 영향을 미치지 않으므로 변수 y_1은 군집 결정에 기여하지 못한다.

또한 $\eta_1 = 0$이면 모든 군집($i = 1, \ldots, g$)의 원 평균 μ_{ik}과 원 분산 σ_{ik}^2의 크기에 상관없이 모든 g개의 군집밀도가 $\phi\left(y_{j1};0,1\right), \ldots, \phi\left(y_{j1};0,1\right)$으로서 우도에서 상수가 되어 모수의 추정에 영향을 미치지 않는다. 따라서 비정보적 변수 y_1이 어떤 특정 군집에서는 $\mu_{i1}^* = 0$, $\sigma_{i1}^{*2} = 1$이 아닌 추정치를 생산하게 되는 "개별적 축퇴"는 원천적으로 나타날 수 없다.

2.3. 식별가능성(identifiability)의 충분조건

NMM을 다룰 때 보통은 언급되지 않지만 필수적인 가정이 있다. 그것은 모수들의 순서집합

$$\mathcal{T}_i = \left\{\pi_i, \mu_{i1}, \ldots, \mu_{ip}, \sigma_{i1}^2, \ldots, \sigma_{ip}^2\right\}, \quad i = 1, \ldots, g$$

들은 (고정된 자료 하에서) g개의 성분에 걸쳐 서로 달라야 한다는 것이다. 그래야 주어진 자료 하에서 성분들을 서로 다르게 구분할 수 있게 된다. 그렇지 않으면 NMM의 성분들을 식별할 수 없고, 모형 적합이 불가능한 상황이 발생하여 결국 사후확률 τ_{ij}의 추정치로부터의 군집분석은 실패하게 된다. 만약 g개의 순서집합 $\mathcal{T}_1, \ldots, \mathcal{T}_g$들이 서로 다르다면, $0 < \eta_1, \ldots, \eta_p \leq 1$에 대해 $\mathcal{T}_i^* = \left\{\pi_i, \mu_{i1}^*, \ldots, \mu_{ip}^*, \sigma_{i1}^{*2}, \ldots, \sigma_{ip}^{*2}\right\}$들 역시 성분에 걸쳐 서로 다르다. 왜냐하면 μ_{ik}^*와 σ_{ik}^{*2}들은 서로 다른 μ_{ik}와 σ_{ik}^{*2}들에 성분에 걸쳐 일정한 η_k들을 곱하거나 뺀 것들이기 때문이다. 더욱이 $\eta_1, \ldots, \eta_p$ 중에 최소한 1개만이라도 0이 아니라면, 즉 p개의 변수 중 단 한 개라도 비정보적 변수가 아니면 모수 순서집합 $\mathcal{T}_1^*, \ldots, \mathcal{T}_g^*$들은 같을 수 없다. 즉 (2.1)의 NMM은 식별가능하다. 그런데 만약 $\eta_1 = \cdots = \eta_p = 0$이면 모든 성분에 걸쳐 모수 순서집합 $\mathcal{T}_i^*$들은 $\{\pi_i, (0, \ldots, 0), (1, \ldots, 1)\}$이 됨으로써 π_i만 빼고 모든 군집의 모수들이 모두 같게 된다. 이런 경우도 모형은 식별가능하지 않다. 그러나 이 상황은 우리가 가진 p개의 변수 모두 비정보적임을 의미하는 것인데, 이러한 상황은 현실에서는 거의 발견할 수 없을 것이며, 관측치들이 g개의 참군집 중 하나로부터 왔다는 전제의 모순이 된다. 따라서 식 (2.1)의 모형이 식별 가능하지 않은 경우는 현실에서 없다고 할 것이다.

이상을 정리하면, 모형 (2.1)이 식별 가능하기 위한 충분조건은 "$\eta_1, \ldots, \eta_p$ 중 최소한 1개는 0이 아니다"이다.

3. 모형 추정

모형 (2.1)에 대한 로그-우도는

$$L(\boldsymbol{\theta}) = \sum_{j=1}^{n} \log \left\{ \sum_{i=1}^{g} \pi_i \prod_{k=1}^{p} \phi\left(y_{jk}; \mu_{ik}^*, \sigma_{ik}^{*2}\right) \right\} \tag{3.1}$$

이다. 여기서 $\boldsymbol{\theta} = \{\sigma_{ik}^2, \pi_i, \mu_{ik}, \eta_k,\ j = 1, \ldots, n;\ i = 1, \ldots, g;\ k = 1, \ldots, p\}$는 모수들의 집합인데, β는 조율상수로 사용할 것이므로 $\boldsymbol{\theta}$에 포함하지 않았다. 일반적으로 $\boldsymbol{\theta}$를 추정하기 위해 식 (3.1)의 $L(\boldsymbol{\theta})$를 직접 최대화 하지 않고, EM 알고리즘을 이용한 간접적인 방법을 사용한다. 그래서 먼저 $z_{ij} = (\boldsymbol{z}_j)_i$는 관측치 $\boldsymbol{y}_j$가 i번째 군집으로부터 왔다면 1, 그렇지 않으면 0을 나타내는 지시변수라 하자. 여기서 $\boldsymbol{z}_j = (z_{1j}, \ldots, z_{gj})^T$은 오직 한 원소만 1이며 나머지는 0인 벡터이다. 그리고 앞으로 $\boldsymbol{y} = (\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_n^T)^T$, $\boldsymbol{z} = (\boldsymbol{z}_1^T, \ldots, \boldsymbol{z}_n^T)^T$로 쓰자. 이때 완비자료 $(\boldsymbol{y}, \boldsymbol{z})$에 대한 로그-우도는

$$L_c(\boldsymbol{\theta}) = \log f(\boldsymbol{y}, \boldsymbol{z}) = \sum_{j=1}^{n} z_{ij} \sum_{k=1}^{p} \log \phi\left(y_{jk}; \mu_{ik}^*, \sigma_{ik}^{*2}\right) + \sum_{j=1}^{n} z_{ij} \log \pi_i$$

와 같이 얻을 수 있다. EM 알고리즘은 $(t+1)$번째 반복의 E-step에서 조건부 기댓값

$$Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}\right) = \mathrm{E}\left[L_c(\boldsymbol{\theta})|\boldsymbol{y}, \boldsymbol{\theta}^{(t)}\right]$$

을 계산하고, M-step에서 $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$를 $\boldsymbol{\theta}$에 관해 최대화하는 반복과정을 수렴할 때까지 수행한다. 여기서 $\boldsymbol{\theta}^{(t)}$는 이전 반복과정에서 얻은 모수 추정치이다.

3.1. E-step

E-step에서는 결국 사후확률

$$\begin{aligned}\tau_{ij}^{(t+1)} &= \mathrm{E}\left[z_{ij}|\boldsymbol{y}_j, \boldsymbol{\theta}^{(t)}\right] = \Pr\left\{z_{ij}=1|y_{1k},\ldots,y_{nk},\boldsymbol{\theta}^{(t)}\right\} \\ &= \frac{\pi_i^{(t)}\prod_{k=1}^{p}\phi\left(y_{jk};\mu_{ik}^{*(t)},\sigma_{ik}^{*2(t)}\right)}{\sum_{h=1}^{g}\pi_h^{(t)}\prod_{k=1}^{p}\phi\left(y_{jk};\mu_{hk}^{*(t)},\sigma_{hk}^{*2(t)}\right)}, \quad i=1,\ldots,g;\ j=1,\ldots,n \end{aligned} \tag{3.2}$$

를 계산하는 것으로 귀결된다.

그런데 앞 2.2절에서도 설명하였듯이 $\eta_k = 0$인 비정보적 변수들에 대한 밀도들은 어차피 분자/분모 항에서 약분되므로 밀도 값 $\phi_i(y_{jk};\mu_{ik}^{*(t)},\sigma_{ik}^{*2(t)})$들을 계산할 필요가 없다. 따라서 식 (3.2)는

$$\tau_{ij}^{(t+1)} = \frac{\pi_i^{(t)}\prod_{k\in\mathcal{A}}\phi\left(y_{jk};\mu_{ik}^{*(t)},\sigma_{ik}^{*2(t)}\right)}{\sum_{h=1}^{g}\pi_h^{(t)}\prod_{k\in\mathcal{A}}\phi\left(y_{jk};\mu_{hk}^{*(t)},\sigma_{hk}^{*2(t)}\right)}, \quad i=1,\ldots,g;\ j=1,\ldots,n \tag{3.3}$$

의 결과와 같을 것이다. 단, $\mathcal{A} = \{k : \eta_k \neq 0, k = 1,\ldots,p\}$이다. 비정보적 변수가 많을 경우 식 (3.3)은 식 (3.2)에 비해 상당하게 계산시간을 줄여줄 수 있을 뿐 아니라 (유효숫자 상실에 기인한) 계산의 부정확성도 줄여 줄 수 있을 것이다.

3.2. M-step

M-step에서는 $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$를 $\boldsymbol{\theta}$에 관해 최대화해야 한다. π_i들은 $\mu_{ik}^*, \sigma_{ik}^{*2}$ 및 η_k들과 분리되어 있어서

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

과 같이 추정치를 유도하는 것은 어렵지 않다 (McLAchlan과 Peel, 2000).

그러나 $(\mu_{ik}^*, \sigma_{ik}^{*2}, \eta_k)$를 추정할 때 우리는 큰 어려움에 봉착하게 된다. 왜냐하면 $(\mu_{ik}^*, \sigma_{ik}^{*2})$는 η_k와 결합되어 있고, 또 식 (2.3)의

$$\eta_k = 1 - \prod_{i=1}^{g}\left[\Phi\left(\frac{\beta-\mu_{ik}^*}{\sigma_{ik}^*/\sqrt{n_i}}\right) - \Phi\left(\frac{-\beta-\mu_{ik}^*}{\sigma_{ik}^*/\sqrt{n_i}}\right)\right] \tag{3.5}$$

를 만족하도록 최대화해야 한다. 이를 위해 식 (3.5)의 제약 하에서 3개의 우도 방정식 $\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})/\partial\eta_k = 0$, $\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})/\partial\mu_{ik}^* = 0$ 및 $\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})/\partial\sigma_{ik}^{*2} = 0$으로부터 동시에 해를 구해야 하는데, 이것은 거의 불가능하게 보인다.

그래서 본 논문에서는 이 문제를 회피하는 대안적 방법으로 Meng과 Rubin (1993)의 ECM(expectation-conditional maximization) 알고리즘을 이용할 것이다. 즉, η_k를 고정하고 $(\mu_{ik}^*, \sigma_{ik}^{*2})$를 추정하고, 다시 $(\mu_{ik}^*, \sigma_{ik}^{*2})$를 고정하고 η_k를 추정하는 것이다.

(1) 우선 $\eta_k = \eta_k^{(t)}$로 고정되어 있다하고, $\partial Q(\mu_{ik}^*|\boldsymbol{\theta}^{(t)})/\partial\mu_{ik}^* = 0$ 및 $\partial Q(\sigma_{ik}^{2*}|\boldsymbol{\theta}^{(t)})/\partial\sigma_{ik}^{*2} = 0$으로부터 $(\mu_{ik}^*, \sigma_{ik}^{*2})$를 추정하자. 이때 유의해야 할 사항은

$$\mu_{ik}^* = \eta_k^{(t)}\mu_{ik} + \left(1-\eta_k^{(t)}\right)0 \quad \text{및} \quad \sigma_{ik}^{*2} = \eta_k^{(t)}\sigma_{ik}^2 + \left(1-\eta_k^{(t)}\right)1$$

이기 때문에 μ_{ik}^*와 σ_{ik}^{*2}의 모수 공간은 각각

$$(\min\{0,\mu_{ik}\}, \max\{0,\mu_{ik}\}) \text{ 이며,} \quad \left(\min\left\{1,\sigma_{ik}^2\right\}, \max\left\{1,\sigma_{ik}^2\right\}\right)$$

라는 점이다. 이런 제한 하에서 추정하는 방법은 두 극단점에서 추정치를 구한 후 그 사이의 내점을 취하는 것이다. 즉, 먼저 $\eta_k^{(t)} = 1$일 때, 즉, $\mu_{ik}^* = \mu_{ik}$일 때, $\partial Q(\mu_{ik}|\boldsymbol{\theta}^{(t)})/\partial\mu_{ik} = 0$으로부터

$$\mu_{ik}^{(t+1)} = \frac{\sum_{j=1}^{n}\tau_{ij}^{(t+1)}y_{jk}}{\sum_{j=1}^{n}\tau_{ij}^{(t+1)}} \tag{3.6}$$

를 얻는다. 그 다음 $\eta_k^{(t)} = 0$일 때 즉 $\mu_{ik}^* = 0$일 때는 주어진 자료에 관계없이 $\mu_{ik}^{*(t+1)} = 0$이 $Q(0|\boldsymbol{\theta}^{(t)})$을 최대로 한다. 결국

$$\mu_{ik}^{*(t+1)} = \eta_k^{(t)}\mu_{ik}^{(t+1)} + \left(1-\eta_k^{(t)}\right)0, \quad i=1,\ldots,g;\ k=1,\ldots,p \tag{3.7}$$

이다.

비슷한 방법으로 $\eta_k^{(t)} = 1$일 때

$$\sigma_{ik}^{2(t+1)} = \frac{\sum_{j=1}^{n}\tau_{ij}^{(t+1)}\left(y_{jk}-\mu_{ik}^{(t+1)}\right)^2}{\sum_{j=1}^{n}\tau_{ij}^{(t+1)}} \tag{3.8}$$

을 얻고, $\eta_k^{(t)} = 0$일 때 $\sigma_{ik}^{*2(t+1)} = 1$을 얻어서

$$\sigma_{ik}^{*2(t+1)} = \eta_k^{(t)}\sigma_{ik}^{2(t+1)} + (1-\eta_k^{(t)})1, \quad i=1,\ldots,g;\ k=1,\ldots,p \tag{3.9}$$

을 구한다.

(2) 이제 우리는 $(\mu_{ik}^*, \sigma_{ik}^{*2}) = (\mu_{ik}^{*(t)}, \sigma_{ik}^{*2(t)})$로서 주어졌다하고 식 (3.5)의 제약 하에서 $Q(\eta_k|\boldsymbol{\theta}^{(t)})$를 최대화해야 한다. 물론 복잡하긴 하더라도 수치해석적인 방법을 동원하면 가능하리라 판단되지만 EM 알고리즘 반복 과정 내에 또 다른 많은 시간이 소요되는 수치적 반복과정을 두는 것은 썩 바람직하지 않다.

여기서 저자는 GEM(generalized EM)의 원리를 이용하여 좀 더 명시적인 추정치를 얻고자 한다. 즉, 추정치로서 $Q(\eta_k|\boldsymbol{\theta}^{(t)})$를 최대화하는 $\eta_k^{\max}$만큼 Q-함수를 증가시키진 못할지라도 단지

$$Q\left(\eta_k|\boldsymbol{\theta}^{(t)}\right) \geq Q\left(\eta_k^{(t)}|\boldsymbol{\theta}^{(t)}\right) \tag{3.10}$$

를 만족하는 어떤 η_k를 추정치로서 고려하는 것이다. 그러한 후보 중에 우리는

$$\eta_k^{(t+1)} = 1 - \prod_{i=1}^{g}\left[\Phi\left(\frac{\beta - \mu_{ik}^{*(t+1)}}{\sigma_{ik}^{*(t+1)}/\sqrt{n_i^{(t+1)}}}\right) - \Phi\left(\frac{-\beta - \mu_{ik}^{*(t+1)}}{\sigma_{ik}^{*(t+1)}/\sqrt{n_i^{(t+1)}}}\right)\right], \quad k = 1, \ldots, p \tag{3.11}$$

를 사용할 것이다. 여기서 $n_i^{(t+1)} = \sum_{j=1}^{n} \tau_{ij}^{(t+1)}$을 나타낸다. 식 (3.11)의 추정치는 단순히 식 (3.5)에 $(\mu_{ik}^{*(t)}, \sigma_{ik}^{*(t)})$ 대신 $(\mu_{ik}^{*(t+1)}, \sigma_{ik}^{*(t+1)})$를 대입한 것이다. 당연히 이 추정치는 식 (3.5)의 제약을 만족하면서 $(\mu_{ik}^{*(t)}, \sigma_{ik}^{*2(t)})$ 대신 한 단계 개량된 추정치들이 사용되므로 자연스럽게 부등식 (3.10)을 만족한다.

혹자는 너무 간접적인 EM 방법들이 동원돼서 수렴 속도가 지나치게 느려지는 것을 우려할 수 있다. 그러나 다음 절에서 다룰 실험에서 우리의 알고리즘은 거의 매번 10회 반복 이내로 수렴하는 결과를 보였다.

3.3. 성분의 개수 결정

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

$$-2L\big(\hat{\theta}\big) + \nu(g)\log(n) \tag{3.12}$$

을 최소로 하는 g로서 결정한다. 여기서 $L(\hat{\theta})$은 $\hat{\theta}$에서 계산된 식 (3.1) 로그-우도 값이며 $\nu(g)$는 자유모수의 개수를 나타낸다. 우리의 모형에서 $\eta_1, \ldots, \eta_p$ 중에서 0이 아닌 개수가 q라 할 때, g개 성분에서 추정해야할 자유 모수의 개수는 $\nu(g) = (g-1) + gq + gq$이 된다.

4. 실험

4.1. 모의 실험

이 절에서는 제안된 방법을 Pan과 Shen (2006)의 벌점화 우도 기법 (앞으로 "P-S 기법"이라 부르겠음)과 비교하면서 4가지 경우의 모의 실험을 통해 그 성능을 알아 볼 것이다.

우선 $n = 150$개의 관측치와 $p = 500$개의 변수를 생성하여 전형적인 다변수 상황을 만들었는데, 정보적 변수 $p_1 = 15$개 및 비정보적 변수 $p_2 = 485$개인 데이터 세트 "DataSet-15"와 $p_1 = 30$개 그리고 $p_2 = 470$인 데이터 세트 "DataSet-30"를 준비하였다. 그리고 두 데이터 세트 각각에 대해 "低-분산 상황"과 "高-분산 상황"을 만들었다. 그래서 총 4개의 데이터 세트에 대해 각각 50번씩의 모의 실험을 할 것이다.

한편, 우리는 $g = 3$개의 군집을 생성할 것이다. 이때 각 3군집에서 평균은 $\boldsymbol{\mu}_1 = ((-5)\mathbf{1}_{p_1}^T, (5)\mathbf{1}_{p_2}^T)^T$, $\boldsymbol{\mu}_2 = ((0)\mathbf{1}_{p_1}^T, (5)\mathbf{1}_{p_2}^T)^T$ 및 $\boldsymbol{\mu}_3 = ((5)\mathbf{1}_{p_1}^T, (5)\mathbf{1}_{p_2}^T)^T$와 같이 하였다. 따라서 처음 p_1개 변수는 정보적 변수이며, 다음 p_2개의 변수는 비정보적 변수가 되도록 하였다. 그리고 분산 행렬은 다음과 같이 정하였다. 우선 $p = p_1 + p_2 = 500$개 변수들의 분산 $\sigma_1^2, \ldots, \sigma_p^2$을 $(2, 3)$ 사이의 서로 다른 실수값으로 취한 다음, 공통 분산행렬을 $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2)$과 같이 정하고, 3군집에서 각각의 분산행렬을 서로 다른 $s_i\,(i = 1, 2, 3)$에 대해

$$\boldsymbol{\Sigma}_i = s_i \times \boldsymbol{\Sigma}, \quad i = 1, 2, 3$$

과 같이 이분산 구조를 만들었다. 이때, 저-분산인 경우 $s_1 = 1, s_2 = 2, s_3 = 2.5$로 하였으며, 고-분산인 경우는 $s_1 = 1, s_2 = 10, s_3 = 20$으로 하여 분산 자체도 크지만 그 차이가 크도록 하였다. 그리고

표 4.1. 모의 실험 결과: 조율상수를 제외하고 모든 수치는 50번의 실험에 대한 평균이며, 괄호 내의 수치는 표준편차를 나타낸다. 변수선택 오차 항에서 정보적 항과 비정보적 항의 개수만큼 각각 정보적변수와 비정보적 변수를 식별하지 못하였음을 나타낸다. 군집할당 오류 항에서의 수치는 150개 관측치 중 3군집으로 잘못 할당된 개수를 나타낸다. 그리고 개별적 축퇴 항의 수치는 3군집의 평균 추정치가 모두 0이 아닌 개수를 나타낸다.

기법	데이터 세트	분산	변수선택 오차		군집할당 오류	개별적 축퇴	조율상수
			정보적	비정보적			
제안된 기법	Data-15	저-분산	0.0200 (0.1414)	0.0600 (0.2399)	0 (0)	0 (0)	$\beta = 0.32$
		고-분산	1.1800 (1.2728)	0.3800 (0.7796)	4.4600 (2.2516)	0 (0)	
	Data-30	저-분산	10.0000 (0)	10.6400 (0.8514)	0 (0)	0 (0)	
		고-분산	10.5200 (0.7351)	10.6000 (1.1066)	0.5800 (0.6728)	0 (0)	
P-S 기법	Data-15	저-분산	0.0600 (0.2399)	0.8000 (0.9258)	0 (0)	15.8600 (0.9478)	$\lambda = 20$
		고-분산	0.9400 (3.5937)	1.5000 (1.1473)	31.3600 (21.2994)	15.6400 (4.1441)	
	Data-30	저-분산	10.0000 (0)	22.7000 (3.1639)	0 (0)	42.7000 (3.1639)	$\lambda = 15$
		고-분산	10.0000 (0)	24.5600 (4.2769)	3.3200 (1.4490)	44.5600 (4.2769)	

$k\,(= 1, \ldots, 50)$번째 실험 마다 매번 $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$, $\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ 및 $\mathcal{N}(\boldsymbol{\mu}_3, \boldsymbol{\Sigma}_3)$으로부터 각각 $n_1 = n_2 = n_3 = 50$개씩의 자료를 생성하였다.

그리고 생성된 각 자료에 대해, 성분의 개수를 결정하기 위해 식 (3.12)의 BIC를 사용하지 않고, $g = 3$으로 하여 식 (2.1)의 NMM을 적합하였다.

표 4.1에 그 결과를 수록하였는데, 두 기법 모두 정보적 변수가 많을 때보다 적을 때 그리고 고-분산일 때보다 저-분산일 때 더욱 좋은 결과를 보였다. P-S 기법은 변수의 식별 능력만을 볼 때 비판받을 만큼 나쁜 방법으로 보이지는 않는다. 다만, 알려진 바 대로 P-S 기법은 평균 추정치들의 성격이 군집에 걸쳐 일치되지 않는 성질인 "개별적 축퇴"가 항상 나타났으며, 그것도 정보적 변수의 개수보다 크게 나타났다 (반면, 제안된 방법에서는 예상대로 전혀 나타나지 않았다). 그래서 정보적 변수가 알려져 있지 않은 현실에서는 어떤 변수가 정보적 변수인지를 판단하는데 어려움이 따를 것이다.

한편 제안된 방법은 (Data-15의 고-분산 상황에서 정보적 변수선택 능력을 제외하면) 모든 평가 항목에서 P-S 기법보다 다소 좋거나 매우 좋은 것으로 나타났다. 특히 고-분산 상황하에서 관측치들에 대한 군집 할당 능력은 P-S 기법보다 우수함을 보이고 있다. 그러나 저-분산 상황에서는 두 기법 모두 완전한 군집 할당 능력을 보여주고 있다. 그렇지만 이것은 P-S 기법의 자료 적합능력이 좋아서라기 보다는 평균의 위치상 그런 결과가 나올 수 밖에 없음을 그림 4.1에 표현된 정보적 변수 y_4의 적합 결과를 보며 설명하겠다.

그림의 첫 행은 왼쪽부터 차례대로 Data-15의 저-분산 상황에서 참모수에 의한 적합 그리고 제안된 방법과 P-S 기법의 추정치에 의한 적합을 나타낸다. 제안된 방법은 참모수에 의한 적합과 거의 유사한 수준의 모형적합 결과를 보여주고 있다. 반면, P-S 기법은 등분산을 가정하고 있으므로 분산 추정치로서 합동분산 추정치를 사용하게 된다. 그러한 이유로 그 적합 상태가 제안된 방법에 의한 적합보다 분명히 좋지는 않다. 그러나 평균 추정치들의 위치의 정확성과 그리고 비중의 유사함에 기인하여 (실험에서

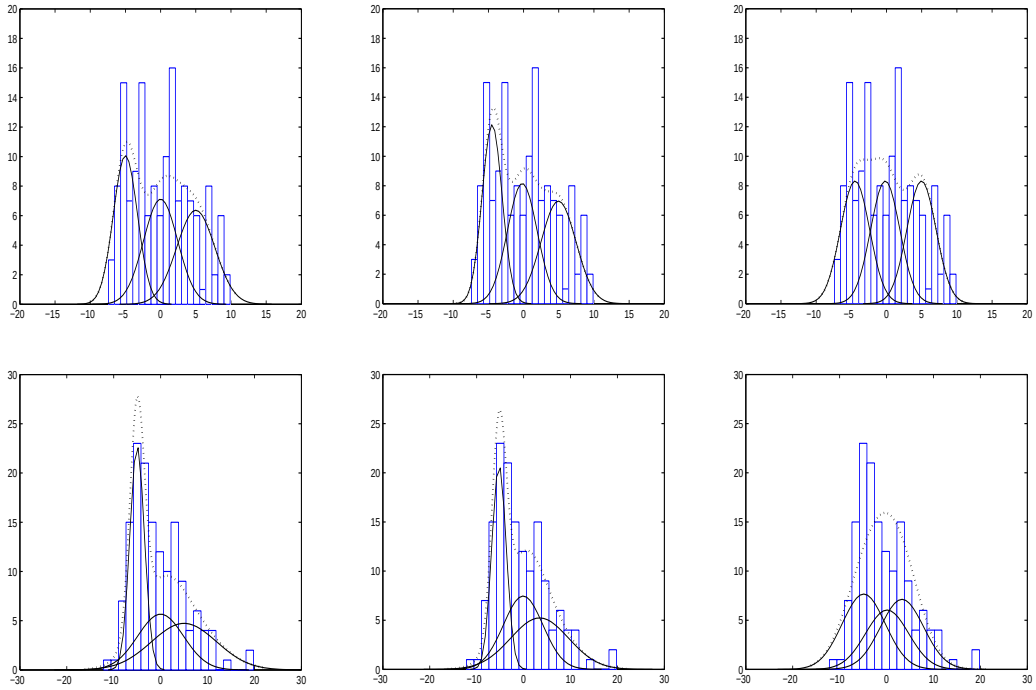

그림 4.1. 정보적 변수 y_k의 적합 결과 (첫째 행: 저-분산 상황, 둘째 행: 고-분산 상황. 각 열의 왼쪽부터, 참모수에 의한 적합, 제안된 방법에 의한 적합, P-S 기법에 의한 적합. 히스토그램은 자료 $(y_{1,4}, \ldots, y_{150,4})$를 그린 것이며, 실선은 왼쪽부터 순서대로 세 성분 추정치 $\hat{\pi}_1\phi(y_4, \hat{\mu}^*_{14}, \hat{\sigma}^{*2}_{14})$, $\hat{\pi}_2\phi(y_4, \hat{\mu}^*_{24}, \hat{\sigma}^{*2}_{24})$, $\hat{\pi}_3\phi(y_4, \hat{\mu}^*_{34}, \hat{\sigma}^{*2}_{34})$을 나타내고, 점선은 이들의 합을 나타낸다)

$n_1 = n_2 = n_3 = 50$으로서 결국 $\pi_1 = \pi_2 = \pi_3 = 1/3$임을 상기하자) 관측치들을 확률적으로 분류하는 능력이 크게 다르지 않은 것이다. 그러나 그림의 둘째 행의 고-분산 상황 하에서는 단순히 평균 추정치의 위치만으로는 관측치들을 분류하는데 한계를 보이고 있다. 그 결과가 표 4.1의 모의실험 결과에 그대로 나타난 것이라 하겠다.

마지막으로, P-S 기법에서 사용된 조율상수 $\lambda = 15$ 및 20은 여러번 실험의 시행착오 후에 대략 최적이라고 판단하여 결정된 값이다. 반면 제안된 방법의 $\beta = 0.32$는 50번의 실험 중 처음 실험의 단 한개의 자료에 대한 사전분석 후 결정된 것으로서 이후 모두 동일하게 사용하였다. 따라서 이 값은 50번의 실험에 속한 모든 자료에 대해 최적이 아니다. 즉, 각 자료에 대해 더 좋은 결과를 제공하는 β를 찾을 수 있을 것이다. 그러나 이렇게 한 이유는 "그럼에도 불구하고 P-S 기법보다 좋음"을 보이기 위해서이다.

4.2. 실자료 실험

본 실험에서는 Golub 등 (1999)의 백혈병 마이크로어레이 유전자발현 자료에 대해 제안된 방법과 P-S 기법의 성능을 비교한다. 이 자료는 27개의 ALL(acute lymphoblastic leukemia) 표본과 11개의 AML(acute myeloid leukemia) 표본 (합계 $n = 38$개)과 $p = 7139$개의 유전자(변수)로 이루어져 있다. 여기서 우리 목적은 작은 군집할당 오류율을 가지는 가능한 한 적은 개수의 유전자를 선택하는 것이다.

표 4.2에 최소 개수의 선택된 변수를 가지는 S-P 기법의 결과와 제안된 방법의 다양한 조율상수에 따른 결과를 제공하였다. S-P 기법의 경우 $\lambda = 8.8$에서 선택된 유전자의 개수는 187개로서 최소였다 (유전

표 **4.2.** 선택된 유전자의 개수와 대응하는 군집할당 오류율: 오류율에서 비율 a/b은 b개 중 a개의 오할당 개수를 나타낸다.

기법	조율상수	선택된 유전자 개수	오류율	
			ALL	AML
S-P	$\lambda = 8.8$	187	0/27	8/11
제안된 기법	$\beta = 0.25$	56	0/27	1/11
	$\beta = 0.28$	42	0/27	5/11
	$\beta = 0.30$	28	0/27	5/11
	$\beta = 0.32$	24	0/27	5/11
	$\beta = 0.34$	12	2/27	2/11

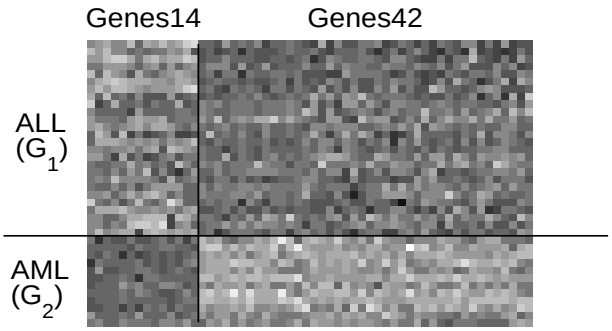

그림 4.2. $\beta = 0.25$에서 제안된 방법에 의해 선택된 56개 유전자들의 두 군집 G_1(ALL) 및 G_2(AML)에서의 유전자 발현도 (밝은 점: 고발현, 어두운 점: 저발현)

자의 개수를 줄이기 위해 λ를 더 크게 하면 모든 평균 추정치들이 0으로 축퇴되어 모든 변수를 비정보적 변수로 만들게 된다). 선택된 187개의 유전자들은 ALL 집단을 정확하게 식별하지만, AML 집단에 대해서는 매우 나쁜 식별력을 보여주고 있다.

한편 제안된 방법의 경우 β를 0.25에서 0.34까지 변화시키면서 결과를 살펴보았다. 모든 경우에 S-P 기법보다 훨씬 적은 개수의 유전자를 선택하여 보다 낮은 군집할당 오류율을 보였다. 그림 4.2는 $\beta = 0.25$에서 식별한 56개의 유전자들의 두 군집 G_1(ALL) 및 G_2(AML)에서의 발현자료를 gray map으로 나타낸 것이다 (밝은 점: 고발현, 어두운 점: 저발현). 대체적으로 처음 14개의 유전자들(Genes14)은 ALL에서 양성(AML에서 음성)을 보이고, 나머지 42개의 유전자들(Genes42)은 AML에서 양성(ALL에서 음성)을 보이고 있다.

한편 군집의 개수는 주어진 조율 상수에서 식 (3.12)의 BIC를 사용하여 선택하였는데, 모든 경우 $g = 2$를 나타내었다.

5. 결론 및 추후연구에 대한 토의

본 논문에서는 관측치의 개수보다 변수가 더 많은 상황에서 NMM에 기반한 군집분석을 위한 정보적 변수선택을 위해 새로운 접근법을 제안하였다. 특히 제안된 모형은 성분들이 이분산을 가지는 상황에 적용이 가능하도록 하기 위해 개별 변수들이 정보적 변수일 확률을 정의하고, 이를 NMM에 결합하여 비정보적 변수들에 대해서는 모수들이 상수가 되도록 설계하였다. 이 모형에 대한 정보적 변수의 식별 원리를 검토하였고, 식별가능성의 충분조건을 제공하였다. 또한 모수에 대한 최우추정을 위한 EM 알고리즘을 유도할 때 발생하는 문제점을 해결하는 방법을 제시하였다. 그리고 모의 실험과 실자료 실험을 통

해 Pan과 Shen (2006)의 기법과 비교하여 그 우위적 성능을 보였다.

1절에서 소개한 이분산 상황 하에서 변수선택을 위해 식 (1.2)를 벌점항으로 하는 Xie 등 (2008)의 방법을 모의실험에서 함께 비교하지 못한 점은 아쉽다. 솔직히 말해 그들의 논문에 수록된 방법론에 대한 설명만으로는 정확히 프로그래밍할 수가 없었다.

본 논문를 포함하여 지금까지의 이 분야의 연구들은 아직 변수들 사이의 독립성을 가정하고 있다. 이 역시 비현실적이라 하지 않을 수 없다. 현실에서는 변수들 사이에 상관성이 없을 수 없다. 특히 마이크로어레이 발현 자료에서 유전자 프로파일들 사이에는 큰 상관성을 가지는 것으로 알려져 있다. Ng 등 (2006)은 군집-특성 랜덤효과를 평균 모형에 고려함으로써 유전자들 사이의 상관성을 추정에 반영하면서 군집을 수행하였다. 이들의 논문은 정보적 변수의 선택을 목적으로 하지 않고 있기 때문에, 이들의 방법에 벌점항을 두고 추정한다든지 혹은 본 논문에서 제안된 방법을 적용해 볼 수도 있을 것 같다. 아무튼 많은 연구자들이 현재 이 문제의 해결책을 연구하고 있을 것이다.

참고문헌

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 andclass prediction by gene expression monitoring, *Science*, **286**, 531–537.

Kim, S.-G. (2006). Use of factor analyzer normal mixture model with mean pattern modeling on clustering genes, *Communications Korean Statistical Society*, **13**, 113–123. (Korean with English abstract)

McLachlan, G. J., Bean, R. W. and Jones, B.-T. (2006). A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays, *Bioinformatics*, **22**, 1608–1615.

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

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.

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

Raftery, A. E. and Dean, N. (2006). Variable selection for model-based clustering, *Journal of the American Statistical Association*, **101**, 168–178.

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.

Variable Selection in Normal Mixture Model Based Clustering under Heteroscedasticity

Seung-Gu Kim[1]

[1]Department of Data and Information, Sangji University

(Received September 2011; accepted September 2011)

Abstract

In high dimensionality where the number of variables are excessively larger than observations, it is required to remove the noninformative variables to cluster observations. Most model-based approaches for variable selection have been considered under the assumption of homoscedasticity and their models are mainly estimated by a penalized likelihood method. In this paper, a different approach is proposed to remove the noninformative variables effectively and to cluster based on the modified normal mixture model simultaneously. The validity of the model was provided and an EM algorithm was derived to estimate the parameters. Simulation studies and an experiment using real microarray dataset showed the effectiveness of the proposed method.

Keywords: Informative variables, variable selection, clustering, EM algorithm, microarray gene expression.

This research was supported by a Sangji Research Fund in 2010.

[1]Professor, Department of Data and Information, Sangji University, 83 Usan-Dong, Kang-Won do, Wonju 122-807, Korea. E-mail: sgukim@sanji.ac.kr