The Korean Communications in Statistics
Vol. 13 No. 1, 2006, pp. 113-123

Use of Factor Analyzer Normal Mixture Model with Mean Pattern Modeling on Clustering Genes

Seung-Gu Kim[1)]

Abstract

Normal mixture model(NMM) frequently used to cluster genes on microarray gene expression data. In this paper some of component means of NMM are modelled by a linear regression model so that its design matrix presents the pattern between sample classes in microarray matrix. This modelling for the component means by given design matrices certainly has an advantage that we can lead the clusters that are previously designed. However, it suffers from "overfitting" problem because in practice genes often are highly dimensional. This problem also arises when the NMM restricted by the linear model for component-means is fitted. To cope with this problem, in this paper, the use of the factor analyzer NMM restricted by linear model is proposed to cluster genes. Also several design matrices which are useful for clustering genes are provided.

Keywords : Clustering genes; Design matrix; Factor analyzer normal mixture model.

1. 서론

마이크로어레이 유전자 발현자료(microarray gene expression data) 분석에서 유전자 집락분석은 매우 중요한 부분을 차지한다. 왜냐하면 식별된 유전자 집락은 암과 같은 질병의 경로(pathway)나 병인에 대한 생물학적 통찰을 제공하기 때문이다. 예를 들면, 어떤 특별한 분자 경로(molecular pathway)에 속한 유전자를 발견하기 위해 혹은 유전자들의 업스트림 영역(upstream region)에서의 공통된 모티프(motifs)를 찾기 위해 (잠재적으로) 공동 통제되는(coregulated) 유전자 집락 찾기는 중요하다(Segal et al., 2003). 유전자들의 집락기법은 집락분석의 종류만큼이나 다양하다고 할 수 있다. McLachan et al. (2004)은 통계적 유전자 집락기법들을 소개하고 있는데, 본 연구에서는 이들 중 정규혼합모형(normal mixture model: NMM)을 이용한 기법에 관심을 가진다.

특히 Francesco and Chiaromonte (2001)은 유전자 집락을 위해 모수들의 재표현을 통한 정규혼합모형을 설계하고 적용하였다. 이들의 재표현 중 성분-평균(component-mean)

1) Professor, Department of Applied Statistics, SangJi University, KangWon 220-702, Korea. E-mail : sgukim@mail.sangji.ac.kr.

에 대한 선형 모형화가 본 논문의 관심의 대상이다. 성분-평균의 선형 모형화는 사전에 계획된 계획행렬(design matirx)를 이용하여 계획된 집락을 유도하며, 집락에 대한 구체적인 해석을 가능하게 한다. 그러나 Francesco and Chiaromonte (2001)의 정규혼합모형 접근법은 마이크로어레이 표본의 개수가 작지 않을 때는 과적합(overfitting)의 문제점을 안고 있다. 특히 이 과적합 문제의 주된 요인은 다차원 성분-공분산 행렬(component-covariance matrix)을 직접 추정함으로써 비롯된다 할 수 있다. 이에 본 연구에서는 성부-공분산 행렬에 인자모형을 적용한 McLachan et al. (2002)의 인자분석자 혼합모형(factor analyzer normal mixture model: FA-NMM)을 사용하여 이 문제를 해결하고자 한다. 아울러 유전자 집락을 위해 몇 가지 계획행렬을 설계하여 문제에 적용함으로써 그 유용성을 보일 것이다.

다음절에서는 성분-평균에 대해 선형모형으로 제약된 정규혼합모형을 소개할 것이며, 아울러 계획행렬이 유전자 프로파일(gene profile)의 평균 패턴을 어떻게 결정하는지를 설명하고 추가적으로 몇 가지 계획행렬을 제안할 것이다. 3절에서는 선형제약을 가진 FA-NMM을 위한 AECM(alternating expectation-conditional maximization) 알고리즘을 유도하고, 4절에서는 모의자료와 실제 자료를 이용하여 실험하여 제안된 방법의 유용성을 밝히며, 5절에서는 결론과 보완될 점을 정리할 것이다.

2. 선형제약을 가진 정규혼합모형

2.1 모형

j번째 p-차원 관측치(즉, 유전자 프로파일) $\boldsymbol{y}_j = (y_{j1}, \cdots, y_{jp})^T$에 관하여 g개의 성분을 가진 정규혼합모형

$$f(\boldsymbol{y}_j; \Psi) = \sum_{i=1}^{g} \pi_i \, \phi(\boldsymbol{y}_j; \boldsymbol{m}_i, \boldsymbol{S}_i), \; j = 1, \cdots, n; \; i = 1, \cdots, g \tag{2.1}$$

을 가정한다. 단, 위첨자 T는 행렬의 전치를 나타내며, π_i와 $\phi(\boldsymbol{y}_j; \boldsymbol{m}_i, \boldsymbol{S}_i)$는 각각 i번째 그룹 G_i의 혼합비율과 평균벡터 $\boldsymbol{m}_i$ 및 성분-공분산행렬 $\boldsymbol{S}_i$을 가지는 다변량정규밀도를 나타낸다. 특히 $\boldsymbol{m}_i$는

$$\boldsymbol{m}_i = E(\boldsymbol{y}_j | j \in G_i) = \boldsymbol{X}_i \boldsymbol{b}_i \;, \; i = 1, \cdots, g \tag{2.2}$$

의 관계를 가진다고 가정한다.

여기서 $\boldsymbol{X}_i$는 주어진 상수로서 $p \times q_i$(단, $q_i \leq p$) 크기의 열에 대한 완전위수(column full rank) 계획행렬(design matrix)이며, $\boldsymbol{b}_i$는 $q_i \times 1$ 계수벡터를 나타낸다. 앞으로 식 (2.1)-(2.2)를 Francesco and Chiaromonte (2001)의 선형제약을 가진 정규혼합모형(NMM with linear restriction: NMM-LR)이라 부르겠다.

2.2 계획행렬의 설계

계획행렬 $\boldsymbol{X}_i$의 열들을 서로 다르게 함으로써 i번째 집락의 유전자 프로파일의 평균패턴을 계획할 수 있다. 이 절에서는 표기의 단순함을 위해 성분을 나타내는 첨자 i를 생략하고 설명하겠다.

(1) $\boldsymbol{X}=\boldsymbol{I}_p$: 이 경우 $\boldsymbol{m}=(m_1,\cdots,m_p)^T=\boldsymbol{X}\boldsymbol{b}=\boldsymbol{I}_p\boldsymbol{b}=(b_1,\cdots,b_p)^T$, 즉 $m_i=b_i$이므로, 어떠한 제약도 부과하지 않은 것과 같다. 즉, 표준 정규혼합모형의 성분을 의미한다.

(2) $\boldsymbol{X}=1_{1:p}$: 여기서 $1_{a:b}$는 a번째부터 b번째까지의 원소가 모두 1, 그 외에는 모든 0인 $p\times 1$ 벡터를 나타낸다. 이때 $\boldsymbol{m}=(m_1,\cdots,m_p)^T=\boldsymbol{X}\boldsymbol{b}=(b,\cdots,b)^T$ 로서 p개의 특징변수들의 평균 패턴을 모두 동일하도록 제약한다. 이런 제약 하에서 얻은 집락은 p개의 특징변수들의 평균차이가 유의하지 않은 관측치(즉, 유전자 프로파일)들로 이루어질 것이라 기대할 수 있다. 이 제약은 특이 패턴을 갖지 않는 유전자들을 제거하는 필터링에 사용될 수 있을 것이다.

이상은 Francesco and Chiaromonte (2001)가 제공한 계획행렬들이다. 본 연구에서는 마이크로어레이 유전자 집락을 위해 유용한 몇 가지 계획행렬을 다음과 같이 추가로 고안한다.

(3) $\boldsymbol{X}=(1_{1:h},1_{h+1:p})$: 예를 들어, $p=6$이고 $h=3$인 경우

$$\boldsymbol{X}=(1_{1:3},1_{4:6})=\begin{bmatrix}1&1&1&0&0&0\\0&0&0&1&1&1\end{bmatrix}^T\text{이므로 }\boldsymbol{m}=\boldsymbol{X}\boldsymbol{b}=\begin{bmatrix}b_1&b_1&b_1&b_2&b_2&b_2\end{bmatrix}^T$$

와 같은 유전자 프로파일의 평균패턴을 얻을 수 있다. 이 패턴은 마이트로어레이에서 두 표본 집단에 대해 특이 발현하는(differentially expressed) 유전자들로 이루어진 집락을 유도하는데 유용하게 사용될 수 있을 것이다.

(4) $\boldsymbol{X}=(1_{1:h},1_{h+1:k},1_{k+1:p})$: 이 경우는 (3)의 계획행렬을 확장한 것으로서, 예를 들어 $p=6$이고, $h=2$, $k=4$의 경우

$$\boldsymbol{X}=(1_{1:2},1_{3:4},1_{5:6})=\begin{bmatrix}1&0&0\\1&0&0\\0&1&0\\0&1&0\\0&0&1\\0&0&1\end{bmatrix}\text{이므로 }\boldsymbol{m}=\boldsymbol{X}\boldsymbol{b}=\begin{bmatrix}b_1&b_1&b_2&b_2&b_3&b_3\end{bmatrix}^T$$

과 같은 패턴을 나타낸다.

이것은 유전자 프로파일의 특정한 영역에서 특이 발현하는 유전자들로 구성된 집락을 찾아내는데 사용될 수 있을 것이다.

(5) $\boldsymbol{X}=(1_{1:p-1})$: 이것은 p번째 표본의 효과가 0인 유전자들을 식별하는데 사용될

수 있을 것이다.

2.3 EM 알고리즘과 모수추정

$\Psi=\{\{\pi_i\},\{\boldsymbol{S}_i\},\{b_i\}\}$를 우리가 관측치 $\boldsymbol{y}=(\boldsymbol{y}_1^T,\cdots,\boldsymbol{y}_n^T)^T$를 바탕으로 추정해야 할 모수들의 집합이라 하자. 이때 로그-우도는

$$\log L(\Psi)=\sum_{j=1}^{n}\log\left\{\sum_{i=1}^{g}\pi_i\,\phi(\boldsymbol{y}_j;\boldsymbol{X}_i b_i,\boldsymbol{S}_i)\right\} \tag{2.3}$$

과 같이 주어진다. 여기서 $\boldsymbol{y}_j$가 i번째 성분으로부터 왔다면 1 그렇지 않으면 0을 나타내는 성분지시변수 $z_{ij}=(\boldsymbol{z}_j)_i$를 도입하여, 관측치 벡터 $\boldsymbol{y}$를 불완전자료(incomplete data) 그리고 $\boldsymbol{z}=(\boldsymbol{z}_1^T,\cdots,\boldsymbol{z}_n^T)^T$를 결측자료(missing data)로 하여 $(\boldsymbol{y}^T,\boldsymbol{z}^T)^T$를 완전자료(complete data)로 취급함으로써 완전자료에 대한 로그-우도

$$\log L_c(\Psi)=\sum_{i=1}^{g}\sum_{j=1}^{n}z_{ij}\log\pi_i\,\phi(\boldsymbol{y}_j;\boldsymbol{X}_i\boldsymbol{b}_i,\boldsymbol{S}_i) \tag{2.4}$$

를 얻을 수 있다. 이때 EM 알고리즘은 $(k+1)$단계에서 관측치에 대한 조건부 기대값 $Q(\Psi\mid\Psi^{(k)})=E[\log L_c(\Psi)\mid\boldsymbol{y},\Psi^{(k)}]$을 최대화하게 되는데, 결국은 E-step에서 $i(=1,\cdots,g)$에 대해 사후확률

$$\tau_{ij}^{(k+1)}=\tau_i(\boldsymbol{y}_j;\Psi^{(k)})=E(Z_{ij}\mid\boldsymbol{y}_j,\Psi^{(k)})=\pi_i^{(k)}\phi(\boldsymbol{y}_j;\boldsymbol{m}_i^{(k)},\boldsymbol{S}_i^{(k)})\Big/\sum_{h=1}^{g}\pi_h^{(k)}\phi(\boldsymbol{y}_j;\boldsymbol{m}_h^{(k)},\boldsymbol{S}_h^{(k)}) \tag{2.5}$$

을 계산하며, M-step에서는 $i(=1,\cdots,g)$에 대해

$$\pi_i^{(k+1)}=\sum_{j=1}^{n}\tau_{ij}^{(k+1)}/n\quad, \tag{2.6}$$

$$\boldsymbol{b}_i^{(k+1)}=(\boldsymbol{X}_i^T\boldsymbol{S}_i^{(k)^{-1}}\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i^T\boldsymbol{S}_i^{(k)^{-1}}\overline{\boldsymbol{y}}_i\quad\text{단, }\overline{\boldsymbol{y}}_i=\sum_{j=1}^{n}\tau_{ij}^{(k+1)}\boldsymbol{y}_j\Big/\sum_{j=1}^{n}\tau_{ij}^{(k+1)}, \tag{2.7}$$

$$\boldsymbol{m}_i^{(k+1)}=\boldsymbol{X}_i\boldsymbol{b}_i^{(k+1)}, \tag{2.8}$$

$$\boldsymbol{S}_i^{(k+1)}=\sum_{j=1}^{n}\tau_{ij}^{(k+1)}(\boldsymbol{y}_j-\boldsymbol{m}_i^{(k+1)})(\boldsymbol{y}_j-\boldsymbol{m}_i^{(k+1)})^T\Big/\sum_{j=1}^{n}\tau_{ij}^{(k+1)} \tag{2.9}$$

를 계산하는 것으로 귀결됨을 보일 수 있다. EM 알고리즘으로부터 얻은 모수 추정치를 $\hat{\Psi}$로 나타내자. 이때 관측치 $\boldsymbol{y}_j$에 대한 집락할당은 다음과 같은 방법으로 한다. 즉, 사후확률 추정치에 대해 $h=\mathrm{argmax}_k\hat{\tau}_{kj}$라 할 때, $\boldsymbol{y}_j$를 그룹 G_h에 할당한다.

식 (2.2)과 같은 평균에 대한 선형 제약방법은 우리가 원하는 유전자 프로파일의 패턴들로 이루어진 집락을 사전에 설계할 수 있게 하는 반면 통제되지 않은 표준 집락분석은 집락에 대한 해석을 위해 추가적인 분석을 필요로 한다. 선형제약 정규혼합모형은 성분-평균에 대한 (선형) 모형화를 통해 사전에 계획된 집락을 얻을 수 있다는 장점이 있다. 그러나 식 (2.1)의 모형은 다음의 문제점이 있다. 마이크로어레이 집락분

석에서는 사전에 선별된 유전자의 개수 n에 비해 표본의 개수 p가 상대적으로 작지 않은 경우가 많다. 이때 식 (2.8)에서 g개의 성분-공분산 행렬 $\boldsymbol{S}_i$의 모수 $gp(p+1)/2$개만을 추정하는데도 관측치(즉, 유전자)의 수가 너무 작아 과적합(overfittng) 문제가 발생할 수 있다. 과적합은 EM 알고리즘에서 사후확률 추정치 $\tau_i(\boldsymbol{y}_j;\Psi^{(k)})$를 0 혹은 1로 빠르게 수렴시키기 때문에 조기에 종료시키는 효과를 가진다. 다시 말해서 초기 추정치에서 크게 벗어나지 못한다. 이는 결국 알고리즘을 그릇된 추정치로 유도하게 한다. 이런 추정치를 바탕으로 얻은 사후확률 추정치는 의미 없는 집락을 구성하게 된다.

이에 대한 해법은 성분-공분산 추정을 위한 차원축소에 있다고 여겨진다. 차원축소 기법은 많이 있지만, 본 연구에서는 최근 이 분야에서 많이 사용되고 있는 인자분석자 혼합모형(FA-NMM)을 사용할 것이다.

3. 선형모형의 제약을 가진 인자분석자 정규혼합모형

인자분석자 정규혼합모형은 식 (2.1)에서 성분-공분산 행렬에 대해

$$\boldsymbol{S}_i = \boldsymbol{B}_i\boldsymbol{B}_i^T + \boldsymbol{D}_i, \quad i=1,\cdots,g \tag{3.1}$$

의 관계를 가지도록 한 정규혼합모형이다. 여기서 $\boldsymbol{B}_i$는 $p\times r\,(\le p)$ 인자적재(factor loadings) 행렬이며, $\boldsymbol{D}_i$은 $p\times p$ 대각행렬로서 흔히 고유성(uniqueness)이라 부른다. 인자분석자 혼합모형은 식 (3.1)의 관계를 바탕으로 성분-공분산 행렬의 추정 때문에 발생하는 과적합 문제를 해결한 집락분석, 혹은 식 (2.1)의 모형을 바탕으로 한 비선형 인자분석을 수행하려는 목적으로 사용된다. 물론 본 연구에서의 목적은 전자에 해당된다.

인자분석자 정규혼합모형을 적합하기 위해 일반적으로 Meng and Dyk (1997)의 AECM 알고리즘이 적용된다. 이에 대한 상세한 내용은 McLachlan and Peel (2000, 245-247)을 참조하기 바라며, 여기서는 간단한 서술과 이를 바탕으로 하여 성분-평균에 대한 선형제약 하에서의 AECM 알고리즘을 유도하고자 한다. 유도 과정은 논리적으로 단순하므로 가능한 한 수식의 사용은 제한하겠다.

인자분석자 정규혼합모형의 적합을 위한 AECM 알고리즘은 $(k+1)$번째 단계에서 두 개의 과정을 수행하도록 되어있다. 제1과정은 $\boldsymbol{S}_i=\boldsymbol{S}_i^*\,(i=1,\cdots,g)$로서 주어졌다는 조건 하에서 완전자료 $(\boldsymbol{y}^T,\boldsymbol{z}^T)^T$에 대응한 조건부 기대값

$$Q_1(\Psi\,|\,\Psi^{(k)}) = E\left[\log L_c(\pi,\boldsymbol{m})\,|\,\boldsymbol{y},\pi^{(k)},\boldsymbol{m}^{(k)},\boldsymbol{S}^*\right] \tag{3.2}$$

을 $\pi=(\pi_1,\cdots,\pi_g)^T$와 $\boldsymbol{m}=(\boldsymbol{m}_1^T,\cdots,\boldsymbol{m}_g^T)^T$에 관하여 최대화하며, 제2과정은 $\pi=\pi^*$와 $\boldsymbol{m}=\boldsymbol{m}^*$로서 주어졌다는 조건하에서 완전자료 $(\boldsymbol{y}^T,\boldsymbol{z}^T,\boldsymbol{u}^T)^T$에 대응한 조건부 기댓값

$$Q_2(\Psi\,|\,\Psi^{(k)}) = E\left[\log L_c(\pi,\boldsymbol{B},\boldsymbol{D})\,|\,\boldsymbol{y},\boldsymbol{B}^{(k)},\boldsymbol{D}^{(k)},\pi^*,\boldsymbol{m}^*\right] \tag{3.3}$$

을 $\boldsymbol{B}$와 $\boldsymbol{D}$에 관하여 최대화 한다. 단, $\boldsymbol{u}=(\boldsymbol{u}_1^T,\cdots,\boldsymbol{u}_n^T)^T$이며, $\boldsymbol{u}_i$들은 인자 벡터를 나타낸다. 여기서 주목할 것은, 두 과정은 서로 공통으로 하는 (최대화를 위한) 모수

를 가지고 있지 않다는 점이다. 이런 의미에서 두 최대화 과정은 서로 분리되어 있다고 할 수 있다. 따라서 성분-평균이 $m_i = X_i b_i$일 때, $b = (b_1^T, \cdots, b_g^T)^T$는 오직 제1과정에만 포함되어있는 모수이므로 제2과정의 최대화와는 무관하다. 그래서 제1과정은 모수 S만을 제외하면 2.3절에서의 것과 완전히 동일한 최대화 과정이 된다.

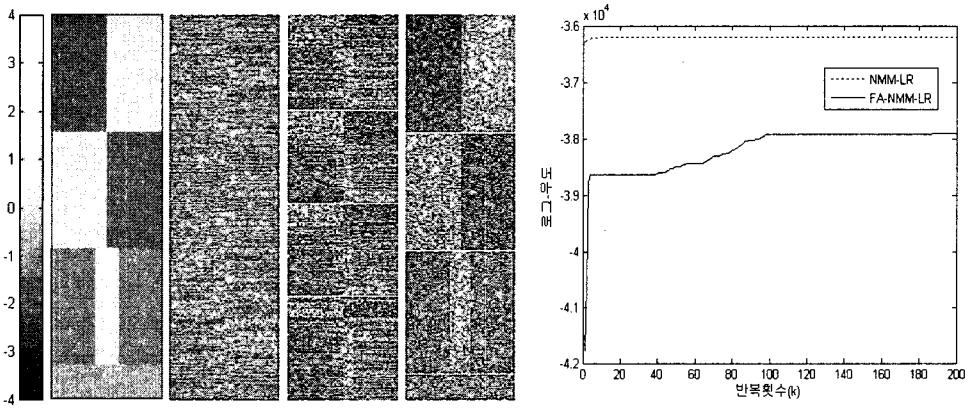
따라서 이 알고리즘은 제약 없는 인자분석자 혼합모형을 위한 AECM 알고리즘과 마찬가지로 제1과정에서 $Q_1(\Psi^* | \Psi^{(k)}) \geq Q_1(\Psi^{(k)} | \Psi^{(k)})$ 과 제2과정에서 $Q_1(\Psi^{(k+1)} | \Psi^*) \geq Q_1(\Psi^* | \Psi^*)$를 만족케 하여, 결국 $Q(\Psi^{(k+1)} | \Psi^{(k)}) \geq Q(\Psi^{(k)} | \Psi^{(k)})$ 따라서 $L(\Psi^{(k+1)}) \geq L(\Psi^{(k)})$를 만족시킴으로써 우도단조증가 알고리즘임을 알 수 있다.

결과적으로 정리하자면, 성분-평균이 $m_i = X_i b_i$일 때 인자분석자 정규혼합모형을 위한 AECM 알고리즘은 제1과정에서 식 (2.4)-(2.7)을 계산하며, 제2과정에서는 제약 없는 인자분석자 정규혼합모형과 동일하게 $B_i^{(k+1)}$와 $D_i^{(k+1)}$을 구한 후 $S_i^{(k+1)} = B_i^{(k+1)} B_i^{(k+1)^T} + D_i^{(k+1)}$을 계산하면 된다. 다음절에서 이 과정을 FA-NMM-LR(FA-NMM with linear restriction)이라 부를 것이다.

4. 실험

4.1 모의자료 실험

모의자료생성: 모의자료는 $n \times p = 330 \times 50$의 마이크로어레이 배열을 나타내는데, 330개의 유전자들은 $g = 4$개의 집락별로 서로 다른 평균벡터와 공분산행렬을 가진 다변량 정규분포로부터 자료를 생성하였다. (<표 4.1> 참조).



(ㄱ) (ㄴ) (ㄷ) (ㄹ) (ㅁ)

<그림 4.1> (ㄱ) 집락별 평균 패턴: 위부터 아래로 집락1부터 집락4를 나타낸다. (ㄴ) 초기 마이크로어레이 배열 (ㄷ) NMM-LR에 의한 집락 (ㄹ) FA-NMM-LR에 의한 집락 (ㅁ) 로그-우도 수렴결과, 그림 (ㄷ)과 (ㄹ)의 흰색 가로선은 집락구분선이다.

<표 4.1> 모의자료 생성을 위한 집락별 평균벡터와 공분산행렬, 자료수 및 적합결과

집락	집락1		집락2		집락3			집락4
표본집단	1-25열	26-50열	1-25열	26-50열	1-20열	21-30열	31-50열	1-50열
평균벡터	$(-2)1_{1:25}$	$(2)1_{26:50}$	$(2)1_{1:25}$	$(-2)1_{25:50}$	$(-1)1_{1:20}$	$(2)1_{21:30}$	$(-1)1_{31:50}$	0_{50}
공분산행렬	$(5)I_{50}$		$(10)I_{50}$		$(5)I_{50}$			$(5)I_{50}$
생성자료수	100		100		100			30
NMM-LR의 추정회귀계수	0.4295	-0.5125	0.2788	-0.1765	-0.3789	0.6772	-0.2090	-0.029
FA-NMM-LR의 추정회귀계수	-1.9521	1.9842	2.1297	-2.0778	-0.9889	1.9683	-1.0293	0.0968

집락1, 집락2는 서로 상반된 특이유전자들의 집단으로서 각각 음성(negative) 및 양성(positive)를 나타내도록 하였다. 특히 집락2는 상대적으로 큰 분산을 주어 표본집단 간 차이를 모호하게 하였다. 집락3은 유전자 프로파일의 중간 10개의 열에서만 고-발현하도록 하였다. 마지막으로 집락4는 제1-50열에 걸쳐 어떠한 패턴도 없는 유전자 집단을 의도한 것이다. <그림 4.1> (ㄱ)에 각 집락의 평균패턴을 나타내었다. 한편, 마이크로어레이 배열에서의 발현값은 일반적으로 red-green 색상표(hitmap)로 표현하지만, 여기서는 (흑백출판을 고려하여) gray-level로 나타내기로 한다. 따라서 밝을수록 고-발현을 어두울수록 저-발현을 나타낸다. 그리고 초기 마이크로어레이 모의자료는 집락별로 자료를 생성한 후 관측치를 임의로 섞어 만들었다. (<그림 4.1>의 (ㄴ)).

집락계획: 혼합모형의 제1성분과 제2성분에는 집락1과 집락2의 평균패턴 구조에 대응하도록 $\boldsymbol{X}_1 = \boldsymbol{X}_2 = (1_{1:25}, 1_{26:50})$를 그리고 제3성분에는 집락3에 평균패턴 구조에 대응하도록 $\boldsymbol{X}_3 = (1_{1:20}, 1_{21:30}, 1_{31:50})$를, 제4성분에는 집락4에 대응하도록 $\boldsymbol{X}_4 = 1_{1:50}$을 적용하였다. 따라서 각 성분의 회귀계수는 각각 $\boldsymbol{b}_1 = (b_{11}, b_{12})^T$, $\boldsymbol{b}_2 = (b_{21}, b_{22})^T$, $\boldsymbol{b}_3 = (b_{31}, b_{32}, b_{33})^T$ 및 $\boldsymbol{b}_4 = b_4$가 될 것이다.

초기추정치 : NMM-LR 및 FA-NMM-LR 기법은 EM 알고리즘이므로 초기추정치가 필요하다. 일반적으로는 k-means와 같은 방법으로 얻은 집락을 바탕으로 초기추정치를 얻지만, 여기서는 초기상태의 자료를 대략 4 등분하여 이를 바탕으로 초기추정치를 얻었다. 이런 임의적인 초기추정치를 사용하는 이유는 두 기법의 성능비교를 위해 초기집락의 효과를 없애기 위해서이다. 한편, FA-NMM-LR을 위한 인자의 개수는 $r = 3 \sim 5$개까지 거의 비슷한 결과를 보였는데, 여기서는 $r = 5$로 정하였다.

NMM-LR의 결과 : <그림 4.1>의 (ㄷ)은 NMM-LR에 의한 집락결과를 나타낸다. 4개의 집락이 모두 비슷한 패턴을 보이고 있고, 초기집락 상태에서 거의 변하지 않았으며, 계획행렬에 의한 집락의 특성이 전혀 반영되고 있지 않다. 그 원인은 NMM-LR

은 $n=330$개의 자료로 성분 공분산 $\boldsymbol{S}_i$의 모수만으로도 1000개 이상을 추정해야 하는 과적합 상황 때문일 것이다. 그럼에도 <그림 4.1>의 (ㅁ)은 NMM-LR의 EM 알고리즘(점선)이 매우 빠르게 수렴되고 있음을 확인할 수 있다.

FA-NMM-LR의 결과 : <그림 4.1>의 (ㄹ)의 FA-NMM-LR에 의한 집락결과는 사전에 설계된 집락계획에 따라 원자료의 집락패턴 상태를 거의 완전하게 복원하고 있다. 단지 집락4에 속했던곤 한다. 이 자료 (I2000)는 원래 62개의 조직표본을 가진 2000개의 유전자를 포함하고 있는데, 본 실험에서는 중복된 8개의 유전자를 먼저 제거한 후, 유의수준 15%하에서 암조직 표본과 정상조직 표본에 대한 개별 t-검정을 통해 421개의 유전자를 선별하였다. 유의수준을 15%로 크게 한 이유는 많은 거짓-양성(false positive) 유전자를 포함시키기 위해서이다. 그리고 $n\times p=421\times 62$크기의 배열을 행별로 그 다음 열별로 표준화를 하였다.

한편, 62개의 조직표본의 <표 4.1>에 나타낸 바와 같이 1-40열은 대장암조직표본(A1)이며, 41-62열은 정상조직표본(A2)을 나타낸다. 그리고 McLachlan et al. (2004)에 따르면 대장암조직표본 중 1-11열(A11)과 12-40열(A12)은 각각 "old 프로토콜" 및 "new 프로토콜"에 의한 서로 다른 검사기법으로 얻었다고 알려져 있다. 6개의 관측치가 집락3으로 오할당 되었음을 밝힌다. 또한 적용된 계획행렬들은 집락의 특성을 잘 반영하고 있음을 확인할 수 있다. <표 4.1>의 마지막 줄에 추정된 회귀계수를 나타내었는데, 원자료의 표본집단별 평균(패턴)에 아주 가깝게 근접해 있음을 보여주고 있다. 아울러 <그림 4.1>의 (ㅁ)으로부터 AECM 알고리즘(실선)이 단조적으로 수렴함을 확인할 수 있다.

4.2 실제자료 실험: Alon의 대장암 자료

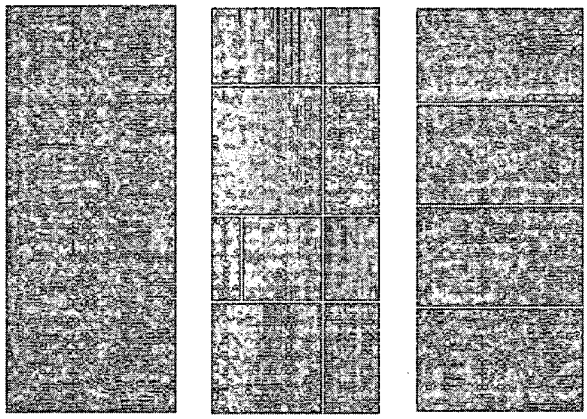
Alon et al. (1999)가 제공한 대장암 마이크로어레이 자료는 벤치마킹 자료로서 많이 사용되고 있어서 이 자료를 통해 제안된 여러 방법들의 신뢰성을 검증하는데 사용되곤 한다. 이 자료 (I2000)는 원래 62개의 조직표본을 가진 2000개의 유전자를 포함하고 있는데, 본 실험에서는 중복된 8개의 유전자를 먼저 제거한 후, 유의수준 15%하에서 암조직 표본과 정상조직 표본에 대한 개별 t-검정을 통해 421개의 유전자를 선별하였다. 유의수준을 15%로 크게 한 이유는 많은 거짓-양성(false positive) 유전자를 포함시키기 위해서이다. 그리고 $n\times p=421\times 62$크기의 배열을 행별로 그 다음 열별로 표준화를 하였다.

한편, 62개의 조직표본의 표4.1에 나타낸 바와 같이 1-40열은 대장암조직표본(A1)이며, 41-62열은 정상조직표본(A2)을 나타낸다. 그리고 McLachlan et al. (2004)에 따르면 대장암조직표본 중 1-11열(A11)과 12-40열(A12)은 각각 "old 프로토콜" 및 "new 프로토콜"에 의한 서로 다른 검사기법으로 얻었다고 알려져 있다.

집락계획 : 집락1과 집락2에서 대장암 조직에 대해 양성 혹은 음성을 나타내는 유전자 집락을 얻기 위해, 혼합모형의 제1-2성분의 계획행렬을 $\boldsymbol{X}_1=\boldsymbol{X}_2=(\mathbf{1}_{1:40},\mathbf{1}_{41:62})$

로 정하였다. 그리고 조직표본 집단 A11, A12에서 특이발현 하는 유전자들이 존재하는지를 알기 위해 제3성분의 계획행렬은 $\boldsymbol{X}_3 = (1_{1:11}, 1_{12:40}, 1_{41:62})$로 하였다. 그리고 집락4에 비-특이발현 유전자들을 모으기 위해 제4성분의 계획행렬은 $\boldsymbol{X}_4 = 1_{1:62}$로 하였다.

<그림 4.2>의 (ㄷ)은 NMM-NR에 의한 집락결과로서 다차원 모수에 대한 과적합으로 인해 의미 없는 집락을 제공하고 있다. <그림 4.2>의 (ㄴ)은 $r=3$으로 한 FA-NMM-NR로서 구한 집락 결과이다. 각 집락에서의 추정된 회귀계수를 <표 4.2>에 나타내었다. 그림과 표로부터 집락1과 집락2는 대체적으로 각각 대장암에 대해 양성과 음성집단으로 판단할 수 있다. 그리고 집락4는 비-특이발현 유전자 집단으로 판단된다. 다만, 그림 상에서 처음 몇 개의 유전자들이 음성으로 보인다. 이는 I2000 자료 자체가 이미 선별된 유전자들을 모아놓는 것이 때문에 특이발현 유전자로 보이는 것이 많기 때문일 것으로 판단한다. 아무튼 이들은 집락4에 귀속될 (사후)확률이 더 컸다.



(ㄱ) (ㄴ) (ㄷ)

<그림 4.2> (ㄱ) 선별된 421×62 마이크로어레이 (ㄴ) FA-NMM- LR에 의한 집락 결과 (ㄷ) NMM-LR에 의한 집락결과. 흰색 가로선은 집락구분선 세로선은 표본구분선을 나타낸다. 그리고 위부터 아래로 집락1부터 집락4를 나타낸다.

<표 4.2> 표본집단별 추정회귀계수

집락	집락1		집락2		집락3			집락4
표본집단	1-40열	41-62열	1-40열	41-62열	1-11열	12-40열	41-62열	1-62열
추정회귀계수	0.2205	-0.7425	-0.4849	0.2763	0.1399	0.4017	-0.6351	-0.0033

마지막으로 집락3은 대장암 조직 중 old 프로토콜(A11)과 new 프로토콜(A12)에 의한 표본 집단들 사이에 특이발현 하는 유전자가 존재하는지를 알기위해 계획한 것이다. 그림과 표에서 보듯이 A12 집단에서 특별히 고-발현하는 유전자들이 상당히 존재

함을 알 수 있다. 특히 X74295 (H.sapiens 2553)은, McLachlan et al. (2004)와 Ben-Dor et al. (2000)에 의하면, 대장벽에 포함된 평활근 조직(soft muscle tissue)에 관련된 유전자로서 양성 혹은 음성으로 분류하기가 어려운 유전자로 알려져 있다. 그런데 본 실험에서는 이 유전자의 (A11, A12, A2)에 대응하는 평균이 (-0.0445, 0.6270, -1.2141) 이며, 높은 사후확률(≈ 1.0)로서 집락3에 포함되었음을 확인하였다. 이와 같은 집락은 표준적인 집락분석기법으로는 찾기 어려울 것이다.

5. 결론 및 논의

본 논문에서는 마이크로어레이 발현자료에서 유전자 집락을 위해 성분-평균을 선형모형으로 제약한 인자분석자 정규혼합모형의 사용을 제안하였다. 실용에서 표준 정규혼합모형은 유전자의 다차원성 때문에 모형의 과적합 문제를 발생시킨다. 이런 문제는 성분-공분산 행렬을 인자모형으로 하는 인자분석자 혼합모형으로 해결될 수 있음을 보였다.

그리고 성분-평균에 대한 선형회귀모형을 사전에 설정함으로써 분석자가 원하는 유전자 프로파일 패턴을 가진 집락을 유도할 수 있다. 이러한 집락의 유도는 표준적인 집락분석으로는 불가능하다. 한편, 본 연구에서는 유전자 프로파일 패턴을 정의하는 몇 가지 계획행렬을 제시하고 모의자료실험과 실제자료실험을 통해 그 유용성을 보였다.

성분-평균에 대한 선형회귀모형에서 회귀계수 혹은 회귀계수의 선형결합에 대한 유의성 검정은 집락의 적합도 혹은 집락의 특성을 세부적으로 분석하는데 도움을 줄 수 있을 것으로 판단된다. 또한 자료가 정규분포를 따르지 않거나 이상치 등이 존재할 때, 정규혼합 모형의 적합은 로버스트(robust) 하다고 할 수 없다. 이에 대한 대안으로 t-분포 혼합모형이나 균등분포 성분이 추가된 정규혼합모형을 고려해 볼 수 있을 것이다.

아울러 저자의 관련지식의 한계로 인해 본 연구에서 얻은 Alon의 대장암 자료의 집락에 대한 보다 구체적인 생물학적 해석을 충분히 할 수 없었다. 이에 대한 추후분석이 있어야 할 것이다.

References

[1] Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybrra, S., Mack, D., and Levine, A.J. (1999). Broad patterns of gene expression revealed by clustering analysis of tomor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences USA 96*, 6745-6750.

[2] Ben-Dor, A., Bruhn, L., Friedman, N., Nachman, I., Schummer, M., and Yakhimi, Z. (2000). Tissue classification with gene expression profiles. *Journal of Computational Biology*, Vol. 7, 559-584.

[3] Francesco and Chiaromonte (2001). Analyzing Gene Expression Data From Microarrays: A Mixture-Based Approach. *ENAR 2001 Spring Meeting*, 25-28 March 2001, Charlotte, North Carolina, USA.

[4] McLachlan, G.J., and Peel, D. (2000). *Finite Mixture Models*. New York: Wiely.

[5] McLachlan, G.J., Bean, R.W., and Peel, D. (2002). A mixture model-based approach to the clustering of micorarray expression data. *Bioinformatics*, Vol. 18, 413-422.

[6] McLachlan, G.J., Do, K-A., Ambroise, C. (2004). *Analyzing Microarray Gene Expression Data.*. Wiely and Sons.

[7] Meng, X.L., and van Dyk (1997). The EM algorithm-an old folk song sung to a fast new tune (with discussion). *Journal of the Royal Statistical Society, Series B*, Vol. 59, 511-567.

[8] Segal, E. Wang, H., and Koller, D. (2003). Discovering molercular pathways from protein interaction and gene expression data. *Bioinformatics*, Vol. 19(Suppl. 1), i264-i272.

[Received October 2005, Accepted December 2005]