The Korean Communications in Statistics **Vol. 14** No. 1, 2007, pp. 205–213

Normal Mixture Model with General Linear Regressive Restriction: Applied to Microarray Gene Clustering

Seung-Gu Kim[1)]

Abstract

In this paper, the normal mixture model subjected to general linear restriction for component-means based on linear regression is proposed, and its fitting method by EM algorithm and Lagrange multiplier is provided. This model is applied to gene clustering of microarray expression data, which demonstrates it has very good performances for real data set. This model also allows to obtain the clusters that an analyst wants to find out in the fashion that the hypothesis for component-means is represented by the design matrices and the linear restriction matrices.

Keywords: Normal mixture model; general linear restriction; EM algorithm; microarray; gene clustering.

1. 서론

정규혼합모형 (normal mixture model: NMM) 은 최근 다양한 응용분야에서 다변량 자료의 모형-기반 군집을 위한 표준적 기법으로 사용되고 있다. 본 연구에서는 혼합성분의 성분평균에 대해 주어진 계획행렬에 관한 일반선형회귀모형을 고려하되 회귀계수에 대한 선형 (등식) 제약 하에서 정규혼합모형의 적합을 유도할 것이다. 이러한 접근법은 최근 응용통계분야의 중요한 토픽인 마이크로어레이 유전자 군집 (microarray gene clustering) 에 유용하게 사용될 수 있다. 김승구 (2006) 는 성분평균의 선형회귀모형 하에서 다양한 계획행렬의 설계를 통해 분석자가 원하는 형태의 마이크로어레이 유전자 군집을 유도할 수 있음을 보였으며, 김승구 (2007) 는 단순선형제약 하에서 NMM에 의한 유전자 군집기법을 다루었다. 본 연구는 김승구 (2007) 의 모형을 일반화된 제약으로 확장한 것이라 할 수 있다. 이 기법은 분석자가 원하는 형태의 유전자 군집 크기를 주어진 어떤 수준으로 통제할 수 있게 하는 유용성을 제공하게 될 것이다.

다음 절에서는 성분평균에 대한 일반선형제약 하의 정규혼합모형을 소개하며, 3 절에서는 제안된 모형의 적합을 EM 알고리즘을 바탕으로 유도하고, 4 절에서는 제안된

1) Professor, Department of Data Information, Sangji University, WooSan-Dong, Wonju, KangWon 220-702, Korea.
E-mail : sgukim@sangji.ac.kr

방법이 유전자 군집에 매우 효과적으로 사용될 수 있음을 보일 것이다. 그리고 마지막 5절에서는 결론을 정리하였다.

2. 일반선형제약 정규혼합모형

주어진 $n \times p$ 다변량 자료행렬 $\boldsymbol{Y}$에 대해 j번째 $p-$차원 관측치 $\boldsymbol{y}_j = (y_{j1}, \ldots, y_{jp})^T$는 독립적으로

$$f(\boldsymbol{y}_j; \boldsymbol{\Theta}) = \sum_{i=1}^{g} \pi_i \phi(\boldsymbol{y}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \quad j = 1, \ldots, n \tag{2.1}$$

단,

$$\boldsymbol{\mu}_i = E(\boldsymbol{Y}_j | j \in \mathcal{G}_i) = \boldsymbol{X}_i \boldsymbol{\beta}_i, \quad i = 1, \ldots, g \tag{2.2}$$

과 같은 $g-$성분 정규혼합모형을 따른다고 가정하자. 여기서 $\pi_i (\pi_1 + \cdots + \pi_g = 1)$, $\boldsymbol{\mu}_i$ 및 $\boldsymbol{\Sigma}_i$는 각각 i번째 성분의 혼합비율 및 다변량 정규분포 ϕ의 $p \times 1$ 평균벡터와 $p \times p$ 공분산 행렬을 나타내며, $\boldsymbol{X}_i$와 $\boldsymbol{\beta}_i$는 각각 $p \times q_i$ 완전 위수 (full-rank) 계획행렬과 $q_i \times 1$ 선형회귀계수벡터를 나타내고 $\mathcal{G}_i$는 i번째 모집단을 의미한다. 그리고 모수벡터 $\boldsymbol{\Theta}$는 식 (2.1)–(2.2) 의 모수 $\{\pi_i\}, \{\boldsymbol{\beta}_i\}, \{\boldsymbol{\Sigma}_i\}$를 모함하는 벡터를 나타낸다. 이제 $q_i \times r_i$ (단, $r_i < q_i$) 크기의 제약행렬 $\boldsymbol{A}_i$ (단, $\mathrm{rank}(\boldsymbol{A}_i) = r_i$) 과 상수벡터 $\boldsymbol{\delta}_i$에 대해 회귀계수들은

$$\boldsymbol{A}_i^T \boldsymbol{\beta}_i = \boldsymbol{\delta}_i, \quad i = 1, \ldots, g \tag{2.3}$$

의 선형관계를 만족한다 하자.

여기서 본 연구의 목적은 식 (2.3) 의 제약하에서 $\boldsymbol{\Theta}$의 MLE를 구하고, 이를 바탕으로 사후확률 $\tau_{ij} = \Pr\{j \in \mathcal{G}_i | \boldsymbol{y}_j\}$의 추정치 $\hat{\tau}_{ij}$를 구하여, j번째 관측치를 성분 $\arg\max_i \hat{\tau}_{ij}$에 할당하는 방식으로 n개의 관측치를 g개의 군집으로 분할하는 것이라 하겠다. 이때, 각 군집은 모수 $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_g^T)^T$에 대한 식 (2.2) 의 관계와 식 (2.3) 의 제약을 가진 정규모집단의 표본으로 판단할 수 있을 것이다.

한편, 식 (2.1) 으로부터 관측치에 대한 $\boldsymbol{\Theta}$의 로그–우도는

$$\log L(\boldsymbol{\Theta}) = \sum_{j=1}^{n} \log \left\{ \sum_{i=1}^{g} \pi_i \phi(\boldsymbol{y}_j; \boldsymbol{X}_i \boldsymbol{\beta}_i, \boldsymbol{\Sigma}_i) \right\} \tag{2.4}$$

와 같다. 이때 식 (2.3) 의 제약하에서 식 (2.4) 를 $\boldsymbol{\Theta}$에 관하여 직접 최대화하는 것은 쉽지 않다. 이를 위해 다음 장에서 EM (expectation-maximization) 알고리즘과 Lagrange multiplier를 이용하여 모형적합과정을 유도할 것이다.

3. EM 알고리즘에 의한 적합

EM 알고리즘의 구성을 위해, 먼저 $\boldsymbol{y}_j$가 $\mathcal{G}_i$로부터 왔다면 1, 그렇지않으면 0을 나타내는 성분지시변수 $z_{ij} = (\boldsymbol{z}_j)_i$를 고려하여, $\boldsymbol{y}$를 불완비자료, $\boldsymbol{z} = (\boldsymbol{z}_1^T, \ldots, \boldsymbol{z}_n^T)^T$를 결측

자료 그리고 $(\boldsymbol{z}^T, \boldsymbol{y}^T)^T$를 완비자료로서 정의하자. 이때 완비자료에 대한 로그-우도는

$$\log L_c(\boldsymbol{\Theta}) = \sum_{j=1}^{n}\sum_{i=1}^{g} z_{ij} \log\left\{\pi_i \phi(\boldsymbol{y}_j; \boldsymbol{X}_i\boldsymbol{\beta}_i, \boldsymbol{\Sigma}_i)\right\}$$

과 같이 얻을 수 있다. 이때, EM 알고리즘은 $(k+1)$ 번째 단계에서 조건부 기대값

$$Q(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(k)}) = E[\log L_c(\boldsymbol{\Theta})|\boldsymbol{\Theta}^{(k)}, \boldsymbol{y}] \tag{3.1}$$

의 최대화를 요구한다. 이는 결국 E-step에서 사후확률 추정치

$$\tau_{ij}^{(k+1)} = E(Z_{ij}|\boldsymbol{\Theta}^{(k)}, \boldsymbol{y}_j) = \frac{\pi_i \phi(\boldsymbol{y}_j; \boldsymbol{X}_i\boldsymbol{\beta}_i, \boldsymbol{\Sigma}_i)}{\sum_{h=1}^{g} \pi_h \phi(\boldsymbol{y}_j; \boldsymbol{X}_h\boldsymbol{\beta}_h, \boldsymbol{\Sigma}_h)} \tag{3.2}$$

를 계산하고, M-step에서는 $i = 1, \ldots, g$에 대하여

$$\pi_i^{(k+1)} = \frac{\sum_{j=1}^{n} \tau_{ij}^{(k+1)}}{n}, \tag{3.3}$$

$$\boldsymbol{\beta}_i^{(k+1)} = \boldsymbol{b}_i^{(k+1)} + \boldsymbol{c}_i^{(k+1)}(\boldsymbol{\delta}_i - \boldsymbol{A}_i^T \boldsymbol{b}_i^{(k+1)}), \tag{3.4}$$

$$\boldsymbol{\Sigma}_i^{(k+1)} = \frac{\sum_{j=1}^{n} \tau_{ij}^{(k+1)} (\boldsymbol{y}_j - \boldsymbol{\mu}_i^{(k+1)})(\boldsymbol{y}_j - \boldsymbol{\mu}_i^{(k+1)})^T}{\sum_{j=1}^{n} \tau_{ij}^{(k+1)}} \tag{3.5}$$

을 계산하는 것으로 귀결된다. 단,

$$\boldsymbol{c}_i^{(k+1)} = (\boldsymbol{X}_i^T \boldsymbol{\Sigma}_i^{(k)^{-1}} \boldsymbol{X}_i)^{-1} \boldsymbol{A}_i \{\boldsymbol{A}_i^T (\boldsymbol{X}_i^T \boldsymbol{\Sigma}_i^{(k)^{-1}} \boldsymbol{X}_i)^{-1} \boldsymbol{A}_i\}^{-1}$$

그리고

$$\boldsymbol{b}_i^{(k+1)} = (\boldsymbol{X}_i^T \boldsymbol{\Sigma}_i^{(k)^{-1}} \boldsymbol{X}_i)^{-1} \boldsymbol{X}_i^T \boldsymbol{\Sigma}_i^{(k)^{-1}} \bar{\boldsymbol{y}}_i^{(k+1)}, \tag{3.6}$$

$$\bar{\boldsymbol{y}}_i^{(k+1)} = \frac{\sum_{j=1}^{n} \tau_{ij}^{(k+1)} \boldsymbol{y}_j}{\sum_{j=1}^{n} \tau_{ij}^{(k+1)}} \tag{3.7}$$

이며 $\boldsymbol{\mu}_i^{(k+1)} = \boldsymbol{X}_i \boldsymbol{\beta}_i^{(k+1)}$ 을 나타낸다. 여기서 식 (3.2), (3.3) 및 (3.5) 의 유도과정은 순수 NMM에서의 그것과 동일하므로 생략하기로 하고 (McLachlan과 Peel (2000) 을 참조바람), 식 (3.4) 의 유도과정은 부록에 수록하였다.

한편 관측치의 개수 n에 비해 변량의 차원 p가 상대적으로 큰 경우, 공분산 추정치 식 (3.5) 로 인해 과적합의 위험성이 있다. 이를 방지하기 위해 $d(<p)$ 차원의 인자적재 (factor loading) $\boldsymbol{B}_{(p\times d)}$와 유일성 (uniqueness) $\boldsymbol{D}_{(p\times p)}$를 바탕으로

$$\boldsymbol{\Sigma}_i^{(k+1)} = \boldsymbol{B}_i^{(k+1)}\boldsymbol{B}_i^{(k+1)^T} + \boldsymbol{D}_i^{(k+1)}, \quad i = 1, \ldots, g \tag{3.8}$$

의 모형을 고려한 알고리즘을 구성할 수 있다. 이 모형을 factor analyzer NMM 모형이라 하는데, EM 알고리즘 구현시 식 (3.5) 대신 식 (3.8) 을 추정하는 conditional M-step을 추가로 두고 있다. 이것에 대한 구현방법에 대해서는 McLachlan과 Bean (2003) 및 김승구 (2006) 에서 상세히 설명하고 있으므로 여기서는 생략하기로 한다. 다만 본 연구에서는 다음 장의 사례적용에 식 (3.5) 대신 식 (3.8) ($d = 5$) 을 사용하였다.

만약 이 모형으로부터 제약이 없는 순수 NMM의 적합을 원한다면 $\boldsymbol{X}_1 = \cdots = \boldsymbol{X}_g = \boldsymbol{I}_p$ 그리고 $\boldsymbol{c}_1 = \cdots = \boldsymbol{c}_g = \boldsymbol{0}_{(p\times p)}$으로 정의하면 된다. 이때, $\boldsymbol{A}_i$와 $\boldsymbol{\delta}_i$의 정의에 관계없이 $\boldsymbol{\mu}_i^{(k)} = \bar{\boldsymbol{y}}_i^{(k)}$이 된다.

4. 마이크로어레이 유전자 군집에의 응용

4.1. 단순선형제약 군집: 유전자 선별

이 절에서는 Alon 등 (1999) 의 cDNA 대장암 마이크로어레이 자료를 이용하여, 본 연구에서 제안한 방법으로 다양한 폴드 변이 (fold change) (즉, 계급간 평균차이) 수준에 대응한 유전자 군집을 유도할 것이다. 원래 Alon의 대장암 마이크로어레이 자료 (I2000) $\boldsymbol{Y}$는 2000개 행의 유전자와 62개 열의 조직표본으로 이루어진 마이크로어레이 발현자료이지만, 본 연구에서는 845×56 크기의 부분행렬만을 사용하여 실험하기로 한다. 이때 1-35열은 대장암 계급 (C_1) 그리고 36-56열은 정상조직 계급 (C_2) 이다. 여기서 i번째 정규모집단 $\mathcal{G}_i$에서 두 계급 C_1, C_2의 모평균을 각각 β_{i1}, β_{i2}라 하자.

첫번째 실험은 대장암에 대해 양성, 음성 및 무특성에 대응하는 정규모집단의 유전자 군집 G_1, G_2, G_3를 찾는 것으로서, 이것은 유전자 선별 (gene selection) 을 목적으로 이용될 수 있을 것이다. 3개의 모집단 $\mathcal{G}_i$를 각각 귀무가설 $H_0^{(i)} : \beta_{i1} - \beta_{i2} = \delta_i (i = 1, 2, 3)$의 단순선형제약식 (즉, $r_i = 1$) 으로 표현하기로 하자. 이때 양성과 음성 및 무특성 모집단은 δ_i를 각각 양수, 음수, 0으로 정의할 수 있을 것이다.

이를 위해 $\boldsymbol{\beta}_i = (\beta_{i1}, \beta_{12})^T$ 그리고 계획행렬을

$$\boldsymbol{X} = \boldsymbol{X}_1 = \boldsymbol{X}_2 = \boldsymbol{X}_3 = \begin{pmatrix} 1\ 1 \cdots 1\ 0\ 0 \cdots 0 \\ 0\ 0 \cdots 0\ 1\ 1 \cdots 1 \end{pmatrix}^T \tag{4.1}$$

와 같이 제1열은 1-35행을 그리고 제2열은 36-56행을 1로 하고 나머지 원소는 0 인 행렬 그리고 $\boldsymbol{A} = \boldsymbol{A}_1 = \boldsymbol{A}_2 = \boldsymbol{A}_3 = (1, -1)^T$로 정의하자. 이렇게 함으로써 $\boldsymbol{\mu}_i = \boldsymbol{X}\boldsymbol{\beta}_i = (\beta_{i1}, \ldots, \beta_{i1}, \beta_{i2}, \ldots, \beta_{i2})^T$ 이고 $\boldsymbol{A}^T\boldsymbol{\beta}_i = \beta_{i1} - \beta_{i2}$ 임을 알 수 있다.

두번째 실험에서는 양성집단과 음성집단에 대해 다양한 수준의 δ_i를 부여함으로써 구체적으로 보다 유의적이거나 혹은 덜 유의적인 상이발현 유전자 (differentially expressed

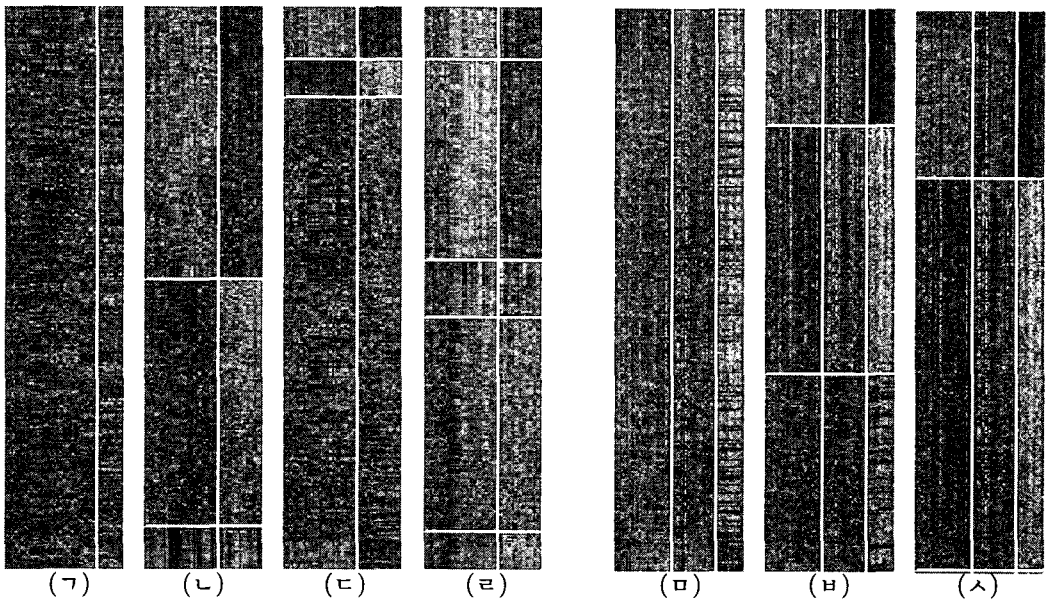

그림 4.1: 선형제약군집 실험.

Note: (ㄱ)-(ㄹ): Alon 자료 군집 그리고 (ㅁ)-(ㅅ): van'tVeer 자료 군집. (ㄱ): 군집전 발현자료, (ㄴ): $\boldsymbol{\delta}_1 = 1, \boldsymbol{\delta}_2 = -1, \boldsymbol{\delta}_3 = 0$, (ㄷ): $\boldsymbol{\delta}_1 = 3, \boldsymbol{\delta}_2 = -3, \boldsymbol{\delta}_3 = 0$ 및 (ㄹ): $\boldsymbol{\delta}_1 = 2, \boldsymbol{\delta}_2 = 1, \boldsymbol{\delta}_3 = 0, \boldsymbol{\delta}_4 = -1, \boldsymbol{\delta}_5 = -2$. (ㅁ): 군집전 발현자료, (ㅂ): $\boldsymbol{\delta}_1 = [1,2]^T, \boldsymbol{\delta}_2 = [-1,-2]^T, \boldsymbol{\delta}_3 = [0,0]^T$, 및 (ㅅ): $\boldsymbol{\delta}_1 = [0,1]^T, \boldsymbol{\delta}_2 = [0,-1]^T, \boldsymbol{\delta}_3 = [0,0]^T$.

genes) 집단으로 통제하려 한다. 즉, 5개의 군집의 폴드 변이가 각각 $\delta_1 > \delta_2 > \delta_3 = 0 > \delta_4 > \delta_5$이 되도록 군집 G_1, G_2, G_3, G_4, G_5로 분할할 것이다. 이것은 양성과 음성 집단을 세부적으로 고양성, 저양성 그리고 저음성, 고음성 집단으로 구분하고자 하는 것이다. 이 경우 $\boldsymbol{X}_3 = \boldsymbol{X}_4 = \boldsymbol{X}$와 $\boldsymbol{A}_3 = \boldsymbol{A}_4 = \boldsymbol{A}$를 추가하면 될 것이다.

그림 4.1에 군집실험 결과를 나타내었다. 발현자료는 256레벨의 graymap으로 나타내었는데, 큰 값 일수록 흰색 그리고 작은 값 일수록 검은색으로 표현되었다. 한편, 세로선은 계급을 구분한 것이며, 가로선은 군집의 구분을 나타낸다 (위부터 $G_1, G_2, \ldots$). 그림 (ㄱ) 은 군집 전의 대장암 마이크로어레이 발현 자료를 나타낸다 (왼쪽: 환자계급, 오른쪽:정상계급). 그림 (ㄴ) 은 $\boldsymbol{\delta}_1 = 1, \boldsymbol{\delta}_2 = -1, \boldsymbol{\delta}_3 = 0$의 제약을 주어 군집한 결과로서, 대장암에 대해 양성 (G_1), 음성 (G_1) 그리고 무특성 (G_3) 유전자 집단을 잘 표현해 주고 있다. 또 $\boldsymbol{\delta}_1 = 3, \boldsymbol{\delta}_2 = -3, \boldsymbol{\delta}_3 = 0$으로 양성 및 음성에 대해 두 계급 평균차이의 제약을 강화하였을 때, 예상대로 그 크기가 줄어든 유전자 군집을 얻을 수 있었다 (그림 (ㄷ)). 그리고 그림 (ㄹ) 은 $\boldsymbol{\delta}_1 = 2, \boldsymbol{\delta}_2 = 1, \boldsymbol{\delta}_3 = 0, \boldsymbol{\delta}_4 = -1, \boldsymbol{\delta}_5 = -2$의 제약을 통해 점차 양성에서 음성으로 변화되는 유전자 군집을 찾은 결과이다. 한편, 각 계급의 평균 추정치 $\boldsymbol{\beta}_i$를 표 4.1 (상단) 에 나타내었다. 추정의 결과로부터 주어진 제약이 만족됨을 확인 할 수 있다.

표 4.1: 모수 $\boldsymbol{\beta}_i$ 추정결과

	실험종류		군집 G_1	G_2	G_3	G_4	G_5
Alon 자료	(ㄴ)	$\boldsymbol{\delta}_i$	1	-1	0	-	-
		β_1	0.3595	-0.4525	0.2223	-	-
		β_2	-0.6405	0.5475	0.2223	-	-
	(ㄷ)	$\boldsymbol{\delta}_i$	3	-3	0	-	-
		β_1	1.0162	-0.9415	-0.0646	-	-
		β_2	-1.9838	2.0585	-0.0646	-	-
	(ㄹ)	$\boldsymbol{\delta}_i$	2	1	0	-1	-2
		β_1	0.7004	0.3998	-0.0234	-0.4611	-0.6733
		β_2	-1.2996	-0.6002	-0.0234	0.5389	1.3267
van'tVeer 자료	(ㅂ)	$\boldsymbol{\delta}_i$	$[1,2]^T$	$[-1,-2]^T$	$[0,0]^T$	-	-
		β_1	0.8573	-0.8754	-0.0737	-	-
		β_2	-0.1427	0.1246	-0.0737	-	-
		β_3	-2.1427	2.1246	-0.0737	-	-
	(ㅅ)	$\boldsymbol{\delta}_i$	$[0,1]^T$	$[0,-1]^T$	$[0,0]^T$	-	-
		β_1	0.1918	-0.2397	-0.0851	-	-
		β_2	0.1918	-0.2397	-0.0851	-	-
		β_3	-0.8082	0.7603	-0.0851	-	-

4.2. 복합선형제약 군집

이 절에서 실험할 자료는 van't Veer 등 (2002) 의 올리고뉴클레오티드 유방암 마이크로어레이 발현자료로서 98 개의 조직 (열) 과 24,881 개 유전자 (행) 으로 이루어져 있다. 그런데 본 실험에서는 사전 선별된 898개의 유전자만 사용하도록 하겠다. 한편 환자 조직 (열) 은 3 그룹으로 구성되는데, 1–44 열은 돌발성 (sporadic) 유방암 환자 중 예후가 좋은 그룹 (C_1), 45–78 열은 예후가 나쁜 그룹 (C_2) 이며 나머지 79–98 열은 BRCA1 및 BRCA2 germline mutation에 기인한 유전적 (hereditary) 소인을 가지는 그룹 (C_3) 을 나타낸다.

우선 3개의 계급의 모평균을 각각 $\boldsymbol{\beta}_{i1}, \boldsymbol{\beta}_{i2}, \boldsymbol{\beta}_{i3} (i = 1, \ldots, 3)$이라 하자. 이때 각 성분의 귀무가설 $H_0^{(1)} : \beta_{11} > \beta_{12} > \beta_{13}$, $H_0^{(2)} : \beta_{21} < \beta_{22} < \beta_{23}$ 및 $H_0^{(3)} : \beta_{31} = \beta_{32} = \beta_{33}$에 대응하는 군집을 찾고자 한다. 여기서 계획행렬을

$$\boldsymbol{X} = \boldsymbol{X}_1 = \boldsymbol{X}_2 = \boldsymbol{X}_3 = \begin{pmatrix} 1 \cdots 1 & 0 \cdots 0 & 0 \cdots 0 \\ 0 \cdots 0 & 1 \cdots 1 & 0 \cdots 0 \\ 0 \cdots 0 & 0 \cdots 0 & 1 \cdots 1 \end{pmatrix}^T \tag{4.2}$$

과 같이 정의하고, 제약식은

$$\boldsymbol{A}^T \boldsymbol{\beta}_i = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} \beta_{i1} \\ \beta_{i2} \\ \beta_{i3} \end{pmatrix} = \boldsymbol{\delta}_i = \begin{pmatrix} \delta_{i1} \\ \delta_{i2} \end{pmatrix}, \quad i = 1, 2, 3 \tag{4.3}$$

와 같은 복합제약 (즉, $r_i \geq 2$) 으로 구성할 수 있다.

그림 4.1의 (ㅁ) 은 군집 전 유방암 발현자료로서, 왼쪽부터 C_1, C_2, C_3 계급을 나타낸다. 그림 (ㅂ) 은 제약값 $\boldsymbol{\delta}_1 = [1,2]^T, \boldsymbol{\delta}_2 = [-1,-2]^T, \boldsymbol{\delta}_3 = [0,0]^T$을 통해 $H_0^{(1)}$: $\beta_{11} - \beta_{12} = 1$ and $\beta_{12} - \beta_{13} = 2$, $H_0^{(2)}$: $\beta_{11} - \beta_{12} = -1$ and $\beta_{12} - \beta_{13} = -2$ 및 $H_0^{(3)}$: $\beta_{11} = \beta_{12} = \beta_{13}$에 대응하는 군집을 찾은 것이다. 그 군집 (특히 G_1, G_2) 는 시각적으로 제약의 특성을 잘 반영하고 있음을 보여주고 있다. 그리고 그림 (ㅅ) 은 $\boldsymbol{\delta}_1 = [0,1]^T, \boldsymbol{\delta}_2 = [0,-1]^T, \boldsymbol{\delta}_3 = [0,0]^T$의 보다 완화된 제약을 주어 군집한 결과로서, 대다수의 유전자는 G_1, G_2에 속하며, 단지 5개의 유전자만이 무특성 군집에 속한 결과를 얻었다. 한편, 표 4.1 (하단) 에 수록한 각 계급에서의 평균 추정치들을 통해 추정치들은 복합선형제약을 만족하고 있음을 알 수 있다.

지금까지 본 절에서 실험한 결과로부터, 성분평균 선형제약 하에서 NMM 적합을 통해 매우 효과적인 유전자 군집을 얻을 수 있음을 보았다. 이것은 분석자가 관심을 가지는 형태의 유전자들를 손쉽게 찾아내준다는 점에서 특별히 의미있다 하겠다. 특히 $\boldsymbol{X}_i = \boldsymbol{I}_p (i = 1, \ldots, g)$라 하고 식 (4.3) 의 제약행렬을 확장하여

$$\boldsymbol{A}_i^T = \begin{pmatrix} 1 & -1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & -1 & 0 & \cdots & 0 \\ & & & \ddots & & \\ 0 & 0 & & \cdots & 1 & -1 \end{pmatrix}_{(p-1)\times p}$$

와 같이 정의하면, DeRisi 등 (1997) 의 yeast 자료와 같은 time–course 발현자료에 적용하여, 융합 시간에 따라 증가하거나 감소하는 유전자 집단이나 전혀 변화가 없는 유전자집단을 발견할 수 있을 것이다.

한편, 본 실험에서는 계획행렬 $\boldsymbol{X}_i$와 제약행렬 $\boldsymbol{A}_i$를 모든 성분 $(i = 1, \ldots, g)$ 에 대해 동일하게 놓았지만, 반드시 같을 필요는 없고 각 성분의 실험계획에 따라 얼마든지 서로 다르게 정의할 수 있다. 단, 성분별 서로다른 $\boldsymbol{X}_i$와 $\boldsymbol{A}_i$를 정의할 때 반드시 군집의 식별성 (identifiability) 을 충족하도록 해야 한다. 식별성을 위한 충분조건 문제는 본 연구에서 다루지 못하였다. 다만, 제약이 식별성을 충족하지 못할 때, 제안된 EM 알고리즘은 단조 수렴하지 못하여 예상밖의 군집결과가 나타나기 때문에 제약의 바르게 정의되었는지 혹은 그렇지 않은지를 실용적으로 판단할 수는 있다.

5. 결론 및 제한점 토의

본 논문에서는 성분평균에 관해 선형회귀 일반선형제약 하에서 NMM 모형을 제안하고, 그 적합과정을 유도하였으며, 제안된 모형을 마이크로어레이 발현 자료의 유전자 군집 문제에 적용하여 그 유용성을 확인하였다. 그 과정에서 혼합 성분평균에 대한 다양한 가설들은 계획행렬과 제약행렬로서 표현될 수 있음을 보이려 하였다. 대부분의 군집 기법들이 군집된 결과로부터 사후에 군집의 평균 특성을 판단할 수 밖에 없도록 한다면, 제안된 군집기법은 분석자가 사전에 평균 구조에 관해 정의한 형태의 군집을 발

견하게 하여 준다는 장점을 가진다.

반면, 대다수의 군집 기법들이 그러하듯 제안된 기법도 관측치들 사이의 독립성을 가정하고 있다. 그런데 본 연구의 실험 대상인 마이크로어레이 자료에서 유전자들은 완전 독립적으로 발현하지 않는다. 이것은 유전자 군집에서 큰 장애물 중 하나라 할 수 있는데, 그러나 유전자의 개체수가 크고, 잘 적합된 군집 결과를 얻었다면 유전자 군집은 의미가 있다 하겠다 (McLachlan 등, 2004). 아무튼 이 문제를 근본적으로 해결하는 기법이 개발되어야 하겠다.

또한, 본 연구에서는 선형제약이 등식을 취하고 있는데, 때에 따라서 실용에서는 부등식 선형제약이 더 가치가 있을 것으로 판단된다. 그러나 부등식 선형제약은 EM 알고리즘 내에서 표현가능식 (explicit form) 으로 존재하지 않으며 구현 자체가 매우 어렵다. 이에 대해 추가 연구가 이루어져야 하겠다.

부록

식 (3.4) 의 유도: 먼저 $\boldsymbol{\beta}$의 표현에 대해 약간의 혼동이 있겠지만, 표기의 절약을 위해 성분을 나타내는 아래첨자 i와 EM 알고리즘의 단계를 나타내는 위첨자 (k) 는 생략하기로 하자.

식 (2.3) $\boldsymbol{A}^T\boldsymbol{\beta} = \boldsymbol{\delta}$의 제약하에 $Q(\boldsymbol{\beta}|\boldsymbol{\beta}^{(k)})$의 최대화는 Lagrange multiplier 벡터 $\boldsymbol{\lambda}_{r\times 1}$에 관하여 함수 $Q_{\boldsymbol{\lambda}}(\boldsymbol{\beta}) = -\sum_j \tau_j(\boldsymbol{y}_j - \boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_j - \boldsymbol{X}\boldsymbol{\beta}) + \boldsymbol{\lambda}^T(\boldsymbol{A}^T\boldsymbol{\beta} - \boldsymbol{\delta})$의 최대화를 의미한다. 먼저 $\boldsymbol{0} = \partial Q_{\boldsymbol{\lambda}}(\boldsymbol{\beta})/\partial\boldsymbol{\beta} = 2\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\sum_j \tau_j(\boldsymbol{y}_j - \boldsymbol{X}\boldsymbol{\beta}) + \boldsymbol{A}\boldsymbol{\lambda}$로부터

$$\boldsymbol{\beta} = (\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\bar{\boldsymbol{y}} + (1/2)\boldsymbol{A}\boldsymbol{\lambda}) = \boldsymbol{b} + (1/2)(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{A}\boldsymbol{\lambda} \quad \text{(A.1)}$$

를 얻을 수 있고, 이것을 $\boldsymbol{0} = \partial Q_{\boldsymbol{\lambda}}(\boldsymbol{\beta})/\partial\boldsymbol{\lambda} = \boldsymbol{A}^T\boldsymbol{\beta} - \boldsymbol{\delta}$에 대입하여

$$\boldsymbol{\lambda} = 2\{\boldsymbol{A}^T(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{A}\}^{-1}(\boldsymbol{\delta} - \boldsymbol{A}^T\boldsymbol{b}) \quad \text{(A.2)}$$

의 관계를 얻을 수 있다. 이제 식 (A.2) 의 결과를 식 (A.1) 에 대입하면 식 (3.4)임을 알 수 있다.

참고문헌

김승구 (2006). Use of Factor Analyzer Normal Mixture Model with Mean Pattern Modeling on Clustering Genes. <한국통계학회논문집>, **13**, 113–123.

김승구 (2007). Gene Clustering using Normal Mixture Model restricted Fold Change in Microarray data. *Journal of the Korean Data Anlysis*, **9**, 127–135.

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 tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences U.S.A.*, **96**, 6745–6750.

DeRisi, J. L., Iyer, V. R. and Brown, P. O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680–686.

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

McLachlan, G. J., Peel, D. and Bean, R. W. (2003) Modelling high-dimensional data by mixtures of factor analyzers. *Computational Statistics & Data Analysis*, **41**, 379–388.

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

van't Veer, L. J., van de Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R. and Friend, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–536.

[Received January 2007, Accepted February 2007]