The Korean Communications in Statistics **Vol. 14** No. 3, 2007, pp. 687–698

Detection of Differentially Expressed Genes by Clustering Genes Using Class-Wise Averaged Data in Microarray Data

Seung-Gu Kim[1)]

Abstract

A normal mixture model with which dependence between classes is incorporated is proposed in order to detect differentially expressed genes. Gene clustering approaches suffer from the high dimensional column of microarray expression data matrix which leads to the over-fit problem. Various methods are proposed to solve the problem. In this paper, use of simple averaging data within each class is proposed to overcome the various problems due to high dimensionality when the normal mixture model is fitted. Some experiments through simulated data set and real data set show its availability in actuality.

Keywords: Class-dependence; differentially expressed gene; microarray; normal mixture model.

1. 서론

마이크로어레이 통계분석에서 사실상 첫번째 주목적은 수천개의 유전자 중에서 기지인 계급들에 걸쳐 상이 발현하는 유전자 즉, DE 유전자(differentially expressed gene)들을 찾는 것이라 하겠다. 일반적으로 마이크로어레이 발현 자료는 $n \times p$의 행렬로서 나타낼 수 있는데, 행과 열은 각각 유전자와 (조직)표본을 나타낸다. 그리고 열들은 g개의 계급으로 분할되어 있다. 이때 DE 유전자를 찾는 통계적 접근법에는 크게 두 가지를 고려할 수 있다.

첫 번째 접근법은 개별 유전자 발현자료에 대해(즉 각 행에 대해) 유의성 검정을 수행하는 것이다. 다시 말해 g개의 계급 모평균이 같다는 귀무가설 $H_0 : \mu_1 = \cdots = \mu_g$ 하에서 고전적 t-통계량이나 이를 개선한 Tusher 등 (2001)의 SAM 통계량 혹은 일원 분산분석의 F-통계량 등을 사용하여 통계적 검정을 수행하는 것이다. 그리고 최근에는 경험적 베이지안 기법(empirical Bayesian approach)이 주류를 이루고 있는데, Allison 등 (2002), Efron (2004) 및 McLachlan 등 (2006) 등이 그 예로서, 고전적 검정 통계량의 유

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

의확률을 이용한 단변량 2-성분 혼합모형을 바탕으로 하고 있다. 그러나 이러한 DE 유전자 식별 기법들이 역시 주류적 방법이기는 하지만 계급들 사이의 종속성을 반영하지 못한다는 결정적 문제점을 가지고 있다.

두 번째 접근법은 다변량 군집기법을 이용하는 것이라 하겠다. 이 경우 유전자 프로파일(행)들을 p-변량 관측치로 취급하여 이들을 DE 유전자 그룹($\mathcal{G}_1$)와 non-DE 유전자 그룹($\mathcal{G}_0$)으로 군집하는 것이다. 이러한 접근법은 (조직)표본들을 관측치가 아닌 변량으로 취급한다는 점에서 다소 문제의 소지를 가지고는 있지만, 다중검정 기법에서 해결할 수 없었던 표본들 사이의 종속성을 모형에서 반영할 수 있다는 장점을 가진다. 그러나 혼합모형에 의한 군집기법은 몇가지 문제점을 가지고 있다. 첫째, (조직)표본의 개수 즉 변량의 차원 p가 지나치게 크면 추정해야 할 성분 공분산 행렬의 모수의 개수($p(p+1)/2$)가 급격히 증가하므로 과적합의 문제점으로 인해 사실상 적용이 불가능하다는 문제점이 있다. McLachlan 등 (2003)는 성분-공분산 행렬에 대해 인자모형을 적용하여 과적합 문제를 회피하는 방법을 제안하였으나 과도한 계산시간에 의한 비실용성이 문제가 될 수 있다. 또한 He 등 (2006)처럼 두 표본계급 사이의 독립성을 가정하는 등 추정해야할 모수의 개수를 줄이는 방법을 고려할 수는 있으나 지나친 가정은 현실성을 떨어뜨린다. 마지막으로, 어쩌면 가장 중요한 문제로서, 혼합모형의 성분분포를 사전에 파악하기란 거의 불가능하다는 점이다. 사전에 발견적 방법으로 대략적인 분포형태를 파악할 수는 있으나 실용에서는 거의 제한적 일 수 밖에 없다.

본 연구에서는 계급내의 몇 개의 표본들을 평균하여 사용하는 단순한 착상으로부터, 계급간 종속성은 반영하면서 앞서 언급한 다변량 혼합모형이 가지는 문제점을 해소하는 방법을 고안하여 보았다. 이에 대해서는 다음 절에서 자세히 소개하며, 3절에서는 모의실험과 실자료 실험을 통해 제안된 방법의 타당성을 보였다. 마지막으로 4절에서는 결론을 정리하였다.

2. 제안된 방법

2.1. 계급별 평균 관측벡터

$n \times p$ 크기의 마이크로어레이 자료행렬 $\boldsymbol{Y}$의 열들은 g개의 계급 $C_1, \ldots, C_g$로 분할되어 있고, 계급 C_i의 원소의 개수는 각각 $p_i (, p_1 + \cdots + p_g = p)$라 하자. 이때 행(유전자 프로파일)들 $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$은 서로 독립이라 가정한다. 여기서 각 계급 C_i를 g_i개의 부분계급 $C_{i1}, \ldots, C_{i,g_i}$로 분할한다 하자. 각 부분계급 C_{ik}의 원소수는 $p_{ik} (, p_{i1} + \cdots + p_{i,g_1} = p_i)$라 하자. 그리고 $j (= 1, \ldots, n)$번째 관측치 $\boldsymbol{y}_j$의 원소들을 부분계급별로

$$z_{jik} \stackrel{\text{def}}{=} \bar{y}_{jik} = \frac{1}{p_{ik}} \sum_{\ell \in C_{ik}} y_{ji\ell}, \quad k = 1, \ldots, g_i;\ i = 1, \ldots, g;\ j = 1, \ldots, n \qquad (2.1)$$

를 평균을 계산하여 p-변량 자료벡터 $\boldsymbol{y}_j = [y_{j1}, \ldots, y_{jp}]^T$ 대신 q-변량 자료벡터 $\boldsymbol{z}_j = [z_{j1}, \ldots, z_{jq}]^T$를 사용한다 하자. 단, $q = g_1 + \cdots + g_g$ 이다.

예를들어, $p = 16$, $g = 3$ 일 때, 각 계급 C_1, C_2, C_3를 각각 $g_1 = 2$, $g_2 = 2$ 및 $g_3 = 1$개의 부분계급으로 하여 아래와 같이 분할한다 하자. 그리고 부분계급별로 평균자료를 계산하였다. 이 경우 우리는 16-변량 자료벡터 $\boldsymbol{y}_j = [1, 2, \ldots, 2, 4]^T$ 대신 5-변량 평균자료벡터 $\boldsymbol{z}_j = [2.0, 1.25, 4.0, 6.0, 4.0]^T$를 사용한다.

주어진 계급	C_1				C_2		C_3
분할된 부분계급	C_{11}		C_{12}		C_{21}	C_{22}	C_{31}
원소 개수 p_{ik}	3		4		2	3	4
$\boldsymbol{y}_j$	1 2 3		1 2 1 1		4 4	5 6 7	6 4 2 4
$\boldsymbol{z}_j$	2.0		1.25		4.0	6.0	4.0

2.2. 부분계급 개수에 대하여

각 계급내에서 부분 계급의 개수를 몇 개로 해야하는지 정하는 기준은 로그-우도 증분이나 BIC(Bayesian information criterion) 등의 기준을 사용할 수 있지만, 이들은 모두 많은 계산시간을 요하는 모형 추정 후의 사후취득 요약 통계량들이다. 본 연구에서는 좀 더 실용적인 방법으로서, 설명된 분산비

$$\gamma = \frac{\text{trace}(\boldsymbol{P}\boldsymbol{S}_Z\boldsymbol{P})}{\text{trace}(\boldsymbol{S}_Y)} \times 100\% \tag{2.2}$$

를 사용하도록 하겠다. 여기서 $\boldsymbol{P} = \text{diag}\left\{\sqrt{p_1}, \ldots, \sqrt{p_q}\right\}$이며, $\boldsymbol{S}_Z$와 $\boldsymbol{S}_Y$는 각각 사전에 자료행렬 $\boldsymbol{Z}$와 $\boldsymbol{Y}$로부터 계산된 공분산 행렬을 나타낸다. 식 (2.2)의 분자는 부분계급내 자료의 크기로 척도화된 변이를 다시 되돌린 것이다. 처음 각 계급에서 부분계급의 개수를 각각 1개씩으로 시작하여 계속 분할하여 나가되, 설명된 분산비의 증가가 크지 않으면 멈춘다. 지나친 분할은 부분계급 내의 원소개수를 줄이게 되므로 정규분포 근사에 문제를 초래한다. 따라서 부분계급의 원소 개수가 어느 수준 이하(예: 15 이하)이면 더 이상 분할하지 않는 것이 좋겠다.

2.3. 계급별 주성분 점수 사용에 대하여

계급별 g_i개의 표본평균자료 $\bar{y}_{ik}$ 대신 각 계급에서의 몇 개의 주성분(principal components) 점수의 사용을 고려할 수도 있을 것이다. 즉, $\boldsymbol{Z}_i = \boldsymbol{Y}_i\boldsymbol{E}_i$를 사용하는 것이다. 여기서, $\boldsymbol{Y}_i$는 열별로 표준화된 i번째 계급의 $n \times p_i$ 자료행렬이며, $\boldsymbol{D}_i$는 $\text{cov}(\boldsymbol{Y}_i)\boldsymbol{E}_i = \boldsymbol{E}_i\boldsymbol{D}_i$를 만족하는 $g_i(\leq p_i)$개의 지배 고유치를 원소로 하는 대각행렬이며, $\boldsymbol{E}_i$는 그에 대응하는 $p_i \times g_i$ 크기의 고유벡터 행렬이다.

주성분 점수의 사용은 꽤 설득력 있어 보이나 추천하고 싶지 않다. 그 이유는 다음과 같다. 첫째, 저자의 경험 상 대부분의 마이크로어레이 자료들은 계급내의 표본들이 극단적인 상관성을 가지는 경우가 흔하다. 이 경우 보통 1개의 고유치가 거의 지배하게 되며, 이때 고유벡터의 원소는 거의 균질하게 $1/p_i$이 되어 계급평균을 사용하는 것과 큰

차이가 없다. 둘째, 반대로 계급내의 표본이 거의 독립적인 경우 차원축소의 능력이 떨어져 원자료와 거의 동일한 차원의 성분을 사용해야 한다. 이 경우 다차원의 문제점을 해결하지 못할 뿐 아니라, p_i가 클 때, 주성분 점수 변량이 정규분포를 따르는지 알 수 없다는 문제점이 있다. 그러나 $\bar{y}_{ik}$는 이 경우 오히려 보다 효율적으로 정규분포의 근사를 보장한다.

2.4. non-DE 유전자

앞 절에서도 언급하였듯이 일반적으로 귀무가설 $H_0 : \mu_1 = \cdots = \mu_g$를 기각할 수 없는 유전자를 non-DE 유전자라 한다. 그래서, 앞의 예에서 μ_{ik}를 부분계급 C_{ik}에서의 모평균이라 할 때, $H_0^* : \mu_{11} = \cdots = \mu_{31}$를 기각할 수 없다면 non-DE 유전자가 분명하지만, H_0^*를 기각할 수 있다고 해서 DE 유전자를 의미하지 않는다. 왜냐하면 DE 유전자란 "부분계급들이 아닌 계급들" 사이에 걸쳐 다르게 발현하는 유전자를 의미하기 때문이다. 각 계급내의 원소의 배치가 임의적이라면 $\mu_{11} = \mu_{12} = \mu_1$, $\mu_{21} = \mu_{22} = \mu_2$ 그리고 $\mu_{31} = \mu_3$이므로, 이는 곧 non-DE 유전자란 귀무가설

$$H_0 : \frac{1}{2}(\mu_{11} + \mu_{12}) = \frac{1}{2}(\mu_{21} + \mu_{22}) = \mu_{31}$$

의 제약을 기각하지 못하는 유전자를 의미한다. 그리고 이 제약은

$$\boldsymbol{X}\boldsymbol{\mu} = \boldsymbol{C}\boldsymbol{D}^{-1}\boldsymbol{\mu} = \begin{pmatrix} 1 & 1 & -1 & -1 & 0 \\ 0 & 0 & 1 & 1 & -1 \end{pmatrix} [\mathrm{diag}(2,2,2,2,1)]^{-1} \begin{pmatrix} \mu_{11} \\ \mu_{12} \\ \mu_{21} \\ \mu_{22} \\ \mu_{31} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad (2.3)$$

으로 표현할 수 있다. 여기서 $\boldsymbol{C}$는 대비행렬(contrast matrix), $\boldsymbol{D}$는 대응하는 부분계급의 개수를 원소로 하는 대각행렬 그리고 $\boldsymbol{\mu}$은 부분계급의 모평균 벡터를 나타낸다. 그리고 행렬 $\boldsymbol{X} = \boldsymbol{C}\boldsymbol{D}^{-1}$를 나타낸다. $C_{ik} \subset C_i (k = 1, \ldots, g_i; i = 1, \ldots, g)$ 인 일반적 상황에서, 본 연구에서는

$$\boldsymbol{X}_{((g-1)\times q)}\boldsymbol{\mu}_{(q\times 1)} = \boldsymbol{C}_{((g-1)\times q)}\boldsymbol{D}^{-1}_{(q\times q)}\boldsymbol{\mu}_{(q\times 1)} = \boldsymbol{0}_{((g-1)\times q)} \quad (2.4)$$

을 만족하는 모평균 벡터를 가진 q-변량 확률분포로부터의 표본을 non-DE 유전자로 정의하겠다. 여기서 $q = \sum_{k=1}^{g} g_i$를 나타낸다.

마이크로어레이 분석에서 흔히 계급의 수는 2개이다. 이때 각 계급내에 부분계급의 수 $g_1 = g_2 = 1$로 한다면, $\boldsymbol{D} = \boldsymbol{I}$이므로 결국 $\boldsymbol{X} = \boldsymbol{C} = [1, -1]$로 축소된다. 그리고 부분계급의 개수를 p개 그리고 모든 부분계급의 크기 p_{ik}를 1로 하면, $H_0 : \mu_1 = \cdots = \mu_p$를 제약하는 것이 된다.

2.5. 정규혼합모형

관측값 $z_k (k = 1, \ldots, q)$들은 부분계급내에서 표본자료들의 평균값이므로 부분계급의 크기가 충분히 크다면 관측벡터 $\boldsymbol{z} = [z_1, \ldots, z_q]^T$의 q-변량 정규분포 $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$를 상

정할 수 있을 것이다. 단, $\boldsymbol{\Sigma}$는 관측벡터 $\boldsymbol{z}$의 $q \times q$ 공분산 행렬이다. 그래서 j번째 관측벡터(유전자 프로파일) $\boldsymbol{z}_j$는 $\pi_0 = \Pr\{j \in \mathcal{G}_0\}$의 확률로 분포 $N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$의 표본이거나, $\pi_1 = \Pr\{j \in \mathcal{G}_1\}$의 확률로 분포 $N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$의 표본이라 할 수 있다(단, $\pi_0 + \pi_1 = 1$). 단, $\boldsymbol{\mu}_0$는 제약 (2.4)를 만족하는 평균벡터이다. 이것은 자연스럽게

$$f(\boldsymbol{z}_j; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \pi_0 \phi_0(\boldsymbol{z}_j; \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) + \pi_1 \phi_1(\boldsymbol{z}_j; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), \quad j = 1, \ldots, n \tag{2.5}$$

과 같은 정규혼합모형을 고려하게 한다. 여기서 ϕ_0와 ϕ_1은 g-변량 정규분포 밀도로서 각각 $\boldsymbol{z}_j$의 귀무밀도(null density) 및 비귀무밀도(non-null density)를 나타낸다.

2.6. EM 알고리즘에 의한 정규혼합모형 추정

관측치 $(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)$에 대한 모수 벡터 $\boldsymbol{\Theta}$의 로그-우도는

$$\log L(\boldsymbol{\Theta}) = \sum_{j=1}^{n} \log \left\{ \sum_{h=0}^{1} \pi_h \phi(\boldsymbol{z}_j; \boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h) \right\} \tag{2.6}$$

와 같다. 여기서 $\boldsymbol{\Theta}$는 식 (2.4)의 모수 $\{\pi_h\}, \{\boldsymbol{\mu}_h\}, \{\boldsymbol{\Sigma}_h\}$를 포함하는 벡터를 나타낸다. 이때 식 (2.3)의 $\boldsymbol{X}\boldsymbol{\mu}_0 = \boldsymbol{0}$의 제약하에서 식 (2.4)를 $\boldsymbol{\Theta}$에 관하여 직접 최대화하여 MLE(maximum likelihood estimator)를 구하는 것은 쉽지 않기 때문에 EM(expectation-maximization) 알고리즘을 경유하여 MLE를 구할 것이다. 단, $\boldsymbol{\mu}_0$에 관하여 최대화 시에 Lagrange multiplier를 이용할 것이다. 그런데, 이것은 김승구 (2007)의 특수한 경우이므로 그의 논문을 참조하기를 바라며, 여기서는 자세한 유도과정 없이 그 결과만 수록한다.

EM 알고리즘은 결국 E-step에서 사후확률 추정치

$$\tau_{0j}^{(t+1)} = \frac{\pi_0^{(t)} \phi_0(\boldsymbol{z}_j; \boldsymbol{\mu}_0^{(t)}, \boldsymbol{\Sigma}_0^{(t)})}{\sum_{h=0}^{1} \pi_h^{(t)} \phi(\boldsymbol{z}_j; \boldsymbol{\mu}_h^{(t)}, \boldsymbol{\Sigma}_h^{(t)})} \tag{2.7}$$

및 $\tau_{1j}^{(t+1)} = 1 - \tau_{0j}^{(t+1)}$를 계산하고, M-step에서는

$$\pi_0^{(t+1)} = \frac{\sum_{j=1}^{n} \tau_{0j}^{(t+1)}}{n} \quad \text{및} \quad \pi_1^{(t+1)} = 1 - \pi_0^{(t+1)}, \tag{2.8}$$

$$\bar{\boldsymbol{z}}_h^{(t+1)} = \frac{\sum_{j=1}^{n} \tau_{hj}^{(t+1)} \boldsymbol{z}_j}{\sum_{j=1}^{n} \tau_{hj}^{(t+1)}}, \quad h = 0, 1, \tag{2.9}$$

$$\boldsymbol{\mu}_0^{(t+1)} = (\boldsymbol{I} - \boldsymbol{\Sigma}_0^{(t)} \boldsymbol{X}^T (\boldsymbol{X} \boldsymbol{\Sigma}_0^{(t)} \boldsymbol{X}^T)^{-1} \boldsymbol{X}) \bar{\boldsymbol{z}}_0^{(t+1)} \quad \text{및} \quad \boldsymbol{\mu}_1^{(t+1)} = \bar{\boldsymbol{z}}_1^{(t+1)}, \tag{2.10}$$

$$\boldsymbol{\Sigma}_h^{(t+1)} = \frac{\sum_{j=1}^{n} \tau_{hj}^{(t+1)} (\boldsymbol{z}_j - \boldsymbol{\mu}_h^{(t+1)})(\boldsymbol{z}_j - \boldsymbol{\mu}_h^{(t+1)})^T}{\sum_{j=1}^{n} \tau_{hj}^{(t+1)}}, \quad h = 0, 1 \tag{2.11}$$

을 계산하는 것으로 귀결된다. 그리고 E-step과 M-step을 모수들이 충분히 수렴할 때까지 반복한다.

이상 제안된 모형의 장점은 다음과 같다. 첫째, 우리는 자료행렬로서 $\boldsymbol{Y}_{(n\times p)}$ 대신 열의 크기가 훨씬 작은 $\boldsymbol{Z}_{(n\times q)}$를 사용한다. 실용에서 조직표본의 개수 p는 보통 50개 이상이 흔한 반면 계급의 개수 g는 기껏 2 혹은 3개 정도다. 계급 당 1-3개의 부분계급을 가진다 한다면, 최대 9-변량 자료를 사용하게 된다. 따라서 다변량 정규혼합모형에서 흔히 발생하는 과적합문제를 회피하기 위해 비현실적인 가정(예: 계급간 독립 혹은 표본간 동일한 상관성 등)을 할 필요가 없다. 둘째, 따라서 처리속도가 매우 빠르다. 셋째, 제안된 모형은 조직표본의 개별 특성을 반영하는 것이 아니라 표본계급의 특성의 반영한다(예: 계급내 평균, 변이 및 계급간 상관성 등). 따라서 "계급간" 상이하게 발현하는 DE 유전자의 의미에 더 부합한다. 넷째, 각 부분계급내의 개체수 p_{ik}가 크면 클수록 원자료의 분포와 거의 무관하게 "정규분포" 혼합모형을 상정할 수 있다. 마지막으로, 무엇보다도 부분 표본에 대한 평균화는 계산이 쉽고 이해하기 쉬우며 직관적이라는 점이다.

3. 의사결정 및 오류율

관측 유전자 $\boldsymbol{z}_j$가 귀무 유전자(즉, non-DE 유전자)일 사후확률은

$$\tau_{0j} = \frac{\pi_0\phi_0(\boldsymbol{z}_j;\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)}{f(\boldsymbol{z}_j;\boldsymbol{\mu},\boldsymbol{\Sigma})}, \quad j=1,\ldots,n \tag{3.1}$$

이다. 여기서 어떤 상수 $c_0 \in [0,1]$에 대해 만약

$$\tau_{0j} \le c_0 \tag{3.2}$$

이면 유전자 j를 DE 유전자 군집(G_1)에 할당하고 그렇지 않으면 non-DE 유전자 군집(G_0)에 할당하기로 하자. 이와같은 의사결정은 할당 위험(allocation risk) Risk $= (1-c_0)\pi_0\epsilon_{01} + c_0\pi_1\epsilon_{10}$ 을 최소화하는 베이즈 의사결정으로 알려져 있다. 여기서 ϵ_{01} 및 ϵ_{10}은 각각 위양성(false positive)과 위음성(false negative) 오류 범할 확률이며, $(1-c_0)$와 c_0는 각각에 대응하는 비용을 나타낸다. 만약 $c_0 = 0.5$ 정하면, 이것은 최대사후확률(maximum *a posteriori*: MAP) 의사결정이 된다.

Efron과 Tibshirani (2002)은 사후확률 τ_{0j}를 국소 FDR(local false discovery rate)이라 하였다. 이것은 측도로서 각 유전자에 대한 귀무적 준거인 양성판정 오류율을 측정한다. 다시 말해서, τ_{0j}는 유전자 j를 DE-유전자라 판정했을때 실제 DE 유전자가 아닐 확률인 것이다. 한편, McLachlan 등(2004)은 전역적 FDR의 추정치를

$$\widehat{\mathrm{FDR}} = \frac{\sum_{j=1}^{n}\hat{\tau}_{0j}I_{[0,c_0]}(\hat{\tau}_{0j})}{n_r} \tag{3.3}$$

과 같이 제안하였다. 여기서 n_r은 선별된 유전자의 개수이며 $I_A(x)$는 $x \in A$이면 1 그렇지 않으면 0 인 지시함수이다. 한편, 전역적 FDR의 정확한 추정치를 얻기 위해서는 귀무확률 π_0 및 귀무밀도 ϕ_0의 정확한 추정치를 확보해야 한다. 그런데 π_0에 대한 몇몇 비모수적 추정방법은 편의를 발생시키는 것으로 알려져 있다. Pawitan 등 (2005)은 FDR에 대한 비모수적 추정법보다 혼합모형에 의한 추정이 편의를 덜 발생한다고 하였다.

4. 실험

4.1. 모의자료실험

본 실험에서는 McLachlan 등 (2006)의 방법을 사용하여 고전적 (합동분산) t-검정통계량을 바탕을 하는 EB(empirical Bayesian) 기법의 문제점 특히 계급간 상관성이 강할 때 DE 유전자 검출의 문제점을 보이고자 한다. 본 실험에서는 계급간 종속성이 있을 때의 상황을 모의실험을 한다. $n \times p = 1000 \times 60$ 크기의 마이크로어레이 모의자료를 생성하되, 첫 500행과 나머지 500행을 각각 귀무그룹($\mathcal{G}_0$)과 비-귀무그룹($\mathcal{G}_1$)이 되도록하였고, 첫 $30(p_1)$열과 나머지 $30(p_2)$을 각각 제1, 2계급으로 하였다. 그리고 각 그룹별로 2-변량 정규분포 자료쌍을 생성하여 각각 제1계급과 제2계급에 나누어 배치하였다. 귀무그룹에서는 평균 $\mu_1 = \mu_2 = 0$ 그리고 비-귀무그룹에서는 $\mu_1 = 0, \mu_2 = 1.5$으로하였고, 표준편차는 두 그룹 모두 $\sigma_1 = 1.0, \sigma_2 = 3.0$로 하였다. 한편, 대응표본쌍의 상관계수를 $\rho_0 = 0.9, 0.2, -0.2 - 0.9$의 4가지 경우를 실험하였다.

그림 4.1(ㄱ), (ㄴ)은 계급 사이에 상관성이 다소 약할 때($\rho_0 = 0.2, -0.2$) EB 기법(임계값 $c_0 = 0.3$)의 검출 결과를 보인 것이다. 그림 (ㄱ), (ㄴ)에서 참 DE 유전자(+)와 참 non-DE 유전자(○)들은 다소 모호하기는 하나 시각적으로 고유의 분포에 대응하여 구분된 위치에서 분포하고 있다. 이때 EB 기법은 대다수의 참 DE 유전자를 식별(•)하고는 있지만, 참 DE 유전자를 식별하지 못하는 위음성오류와 참 non-DE 유전자를 검출하는 위양성 오류를 범하고 있다. 어떤 검출 기법도 이 두 오류는 피할 수는 없다. 그러나 소수이기는 할지라도 두 집단의 경계지역이 아닌 귀무집단의 외곽 영역에서 위양성 유전자를 검출하는 경향은 특별히 문제라 하겠다. 특히 이런 위양성 유전자들이 매우 유의적으로 검출되었을 때는 더욱 그러할 것이다. 이 상황은 그림 4.1(ㄹ)에서 처럼 계급간 강한 음의 상관성이 존재할 때 더욱 심해지게 된다. 이러한 위양성 오류의 원인은 EB 기법이 t-검정통계량을 경유하여 평균차 $|\bar{y}_1 - \bar{y}_2|$에 크게 의존하기 때문에 일 것으로 판단된다.

물론 이 경우 임계값 c_0를 아주 작게 함으로써 문제의 위양성 오류를 줄일 수는 있겠지만, 그럴수록 위음성 오류가 증가하여 생의학자들이 찾고자 하는 많은 참 DE 유전자를 발견할 가능성이 낮아지게 될 것이다. 한편, 제안된 방법($\hat{\tau}_{0j} \leq 0.5$)은 어떠한 경우에서건 귀무집단과 비귀무집단을 잘 식별하였으며(타원: 제안된 방법의 추정치에 의한 Mahalnobis-거리의 95% 신뢰영역), 특히 그림 (ㄷ)과 (ㄹ)에서는 각각 0, 2개의 오류만을 보였다.

4.2. 대장암 자료실험

여기서는 잘 알려진 Alon 등(1999)의 대장암 마이크로어레이 자료로부터 DE 유전자를 찾고자 한다. Alon의 대장암 마이크로어레이 자료(I2000) $\boldsymbol{Y}$는 2000개 행의 유전자와 62개 열의 조직표본으로 이루어진 마이크로어레이 발현자료인데, 사전에 log를 취하고 열별로 (중위수와 절대평균편차를 사용하여) 표준화 작업을 하였다. $\boldsymbol{Y}$의 1-22열은 정상조직 계급(C_1)이며 23-62열은 대장암조직 계급(C_2)이다. 이 자료는 22개의 대응표본을

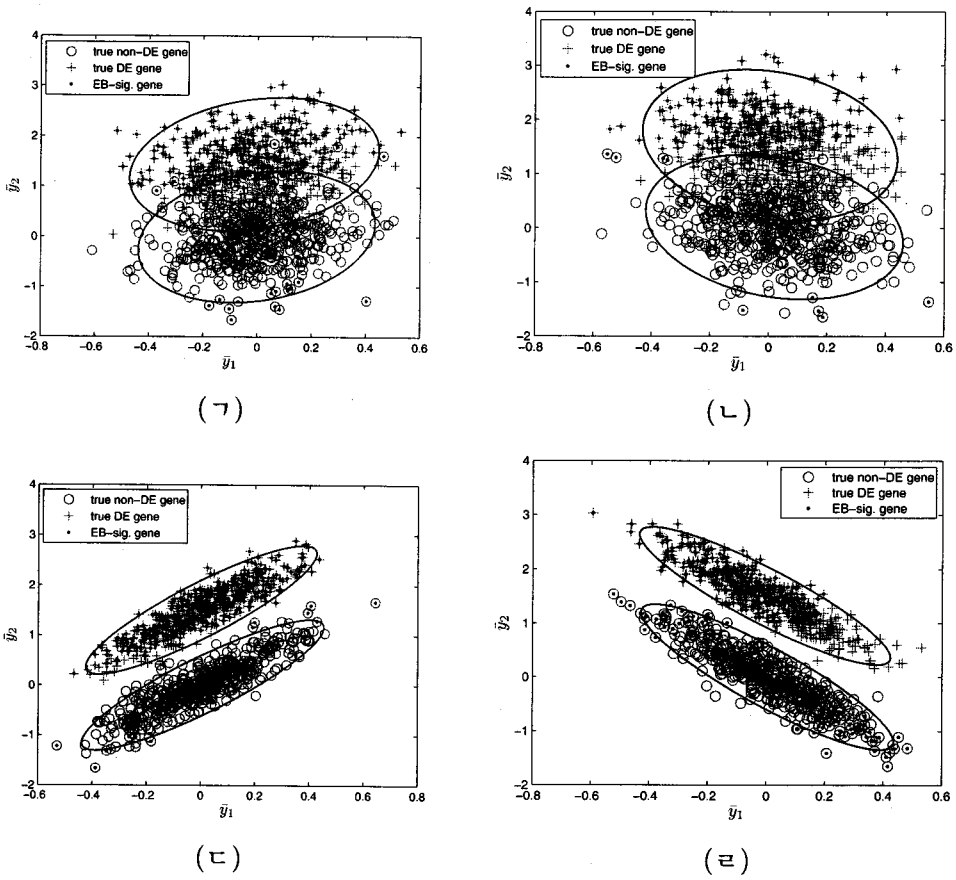

그림 4.1: **DE 유전자 검출 결과.** (ㄱ) $\rho_0 = 0.2$ 일 때, (ㄴ) $\rho_0 = -0.2$ 일 때, (ㄷ) $\rho_0 = 0.9$ 일 때, (ㄹ) $\rho_0 = -0.9$ 일 때.

가지고 있고 각 계급내의 원소 사이 뿐 아니라 계급간의 원소 사이에도 강한 양의 상관관계를 가지고 있다. 더구나 유전자(행)들 간의 독립성도 보장할 수 없는 것으로 알려져 있다. 그림 4.2를 보면 전체 자료 뿐 아니라 C_1, C_2별 자료는 1개의 고유치에 의해 거의 지배되고 있음을 알 수 있다. 따라서 사실 각 계급의 분할이 필요하지 않을 수도 있다. 분할 없이 C_1, C_2의 평균자료를 사용하였을 때, 설명된 분산비는 $\gamma_2 = 76.44\%$ 였다. 그리고 C_1은 그대로 두고 계급 C_2를 C_{21}(23-44열) 그리고 C_{22}(45-62열)로 분할하였을 때, 설명된 분산비는 $\gamma_3 = 77.15\%$ 였다. 그 이상 분할하면 설명된 분산비는 완만히 증가하나 부분계급의 크기가 너무 작아지기 때문에 더 이상의 분할은 하지 않았다. 이제 3개의 부분계급의 평균값을 원소로 하는 3-변량 자료벡터 $\boldsymbol{z}_j = [z_{j1}, z_{j2}, z_{j3}]^T$를 사용할 것이다.

그림 4.3의 (ㄱ)-(ㄴ)은 제안된 방법($c_0 = 0.5, \widehat{\text{FDR}} = 0.1743$)에 의한 검출 결과(○:

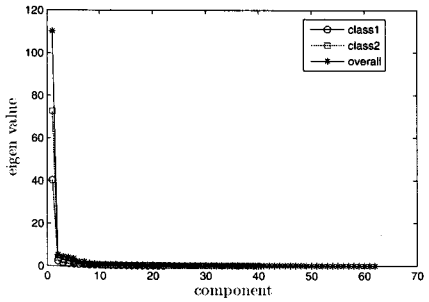

그림 4.2: 계급별 자료의 고유치

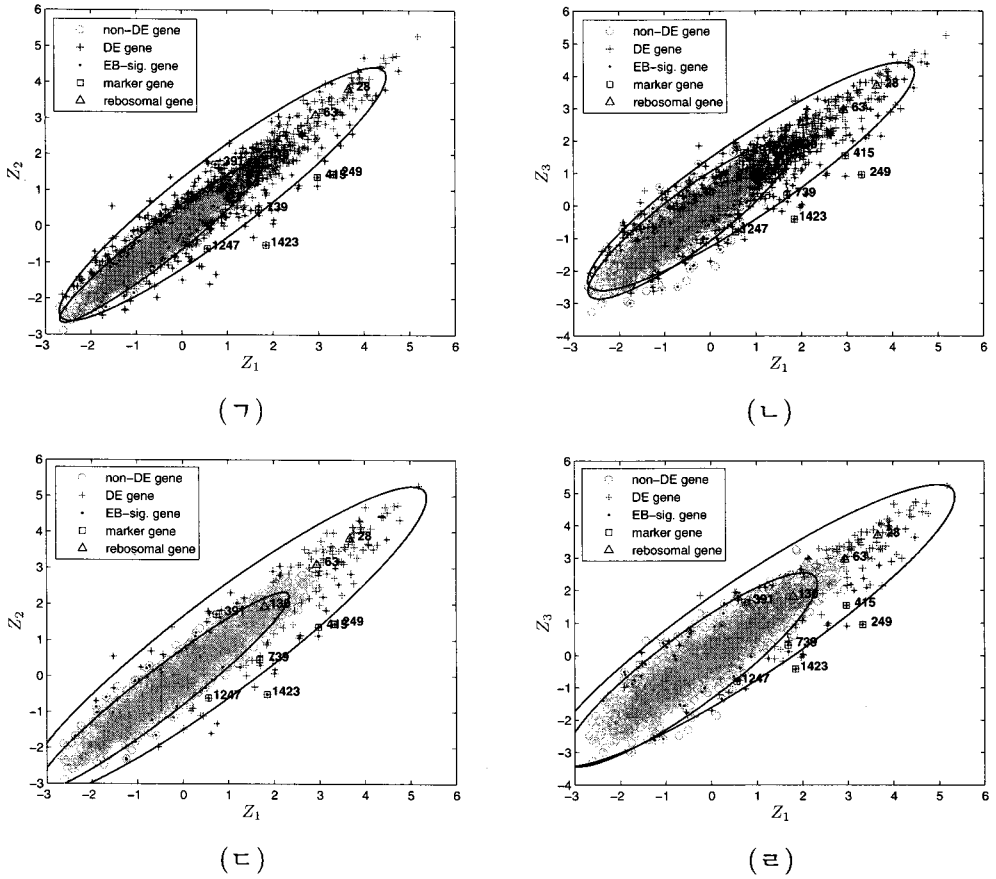

그림 4.3: 2차원 주변 공간에 표시된 DE 유전자 검출 결과.

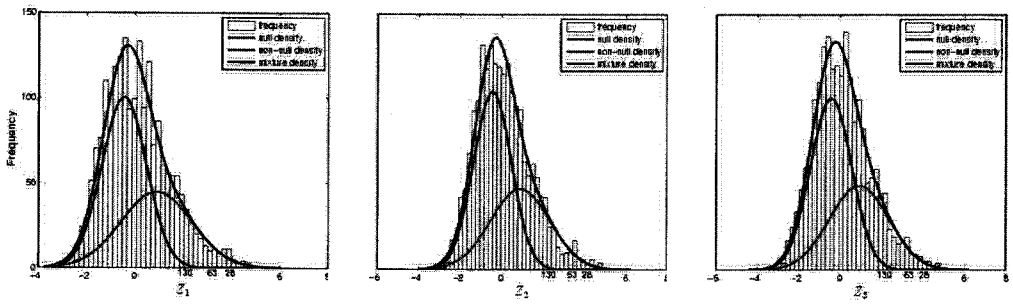

그림 4.4: 주변경험분포의 적합 결과

DE 유전자, +: non-DE 유전자)와 EB 기법에 의한 검출 결과(•)를 각각 계급 C_1 대 C_{11} 그리고 계급 C_1 대 C_{12}에 대응한 2차원 주변공간에 나타낸 것이다. 1247, 1423, 249, 739, 415, 391번 유전자(각각 유전자-ID X74295, J02854, M63391, X12369, T60155, D31885)들(□)은 대장벽의 연근육(smooth muscle) 조직과 관련을 가진 것들로서 흔히 대장암의 표지자(maker gene)로 이용된다. 두 기법 모두 이들을 DE 유전자로 검출하고 있다. 이들의 제안된 방법에 의한 귀무사후확률 추정치는 표 4.1에 정리하였다. 다만, EB 기법은 두 계급 C_1, C_2 사이의 평균값 차이가 일정한 대각선을 따라 검출하는 경향을 띄는 반면, 제안된 방법은 3-차원 공간에서 서로 다른 모수에 의한 정규분포의 차이를 바탕으로 검출한다. 그래서 EB 기법에서 발견하지 않은 많은 유전자를 DE 유전자로 검출하고 있다. 특히, 63, 28, 130번 유전자(각각 유전자-ID H77302, T63484, R85464)들은 리보좀 유전자(ribosomal gene)로서 최초 Alon 등 (1999)에 의해 정상계급(C_1)과 대장암계급(C_2) 사이에 상이발현하는 유전자로 분류된 것들이다. 그러나 McLachlan 등(2006) 및 Do 등(2005)은 그들의 EB 기법을 통해 이들을 귀무사후확률을 0.4 이상인 non-DE 유전자로 분류하였다. 그것은 어쩌면 당연하다. 그림 4.3을 보면 이 세 유전자의 계급간 평균차이는 거의 없어보이기 때문이다. 그러나 제안된 방법은 이들 중 28번과 63번 유전자를 매우 유의한 유전자로 식별하고 있다. 평균차이는 없지만 귀무분포로부터 매우 떨어져 있기 때문이다. 그림 4.3의 (ㄷ)-(ㄹ)은 두 기법의 임계값 $c_0 = 0.01(\widehat{\text{FDR}} = 0.0019)$로 하였을 때 검출결과를 나타낸 것이다(신뢰타원 한계 99%). 이때 역시 28번과 63번 유전자는 귀무분포보다는 비귀무분포의 표본으로 보인다. 그림 4.4는 3-변량 정규혼합모형 적합결과 각 변량 Z_k의 주변분포로 나타내어 주변경험분포의 적합상태를 보인 것인데, 적합결과는 좋은 것으로 판단된다. 특히, 각 주변 경험분포의 오른쪽 꼬리부분을 보자. 이것은 단일 정규분포의 꼬리라고 볼 수 없을 만큼 길다. 다시말해 그 부분은 추가적 정규분포 즉 비귀무분포의 혼합이 이루어져야 할 수 밖에 없다. 따라서 28번과 63번 유전자는 단일 정규분포 즉 귀무분포의 표본이라고 보기는 어렵다.

한편, 이 자료에 대해 제안된 방법의 추정치는 $\hat{\pi}_0 = 0.5845, \hat{\pi}_1 = 0.4155$, $\hat{\mu}_0 =$

표 4.1: 몇몇 유의 유전자의 귀무사후확률

	유전자 번호(I2000)	유전자 ID	귀무사후확률 추정치 $\hat{\pi}_{0j}$
marker gene	1423	J02854	0.00000
	249	M63391	0.00000
	415	T60155	0.00000
	739	X12369	0.00020
	1247	X74295	0.00228
	391	D31885	0.00335
ribosomal gene	28	T63484	0.00047
	63	H77302	0.00573
	130	R85464	0.12861

$[-0.4477, -0.4457, -0.4496]^T, \hat{\mu}_1 = [0.8998, 0.8975, 0.9105]^T$ 및

$$\hat{\mathbf{\Sigma}}_0 = \begin{pmatrix} 0.8407 & 0.8032 & 0.8086 \\ 0.8032 & 0.8360 & 0.8369 \\ 0.8086 & 0.8369 & 0.9728 \end{pmatrix}, \quad \hat{\mathbf{\Sigma}}_1 = \begin{pmatrix} 2.1590 & 1.9674 & 1.9305 \\ 1.9674 & 2.0733 & 2.0344 \\ 1.9305 & 2.0344 & 2.0599 \end{pmatrix}$$

과 같이 얻었다. 특히 귀무평균은 모형에서 제약한 대로 $\hat{\mu}_{01} = -0.4477 = (-0.4457 - 0.4496)/2 = (\hat{\mu}_{02} + \hat{\mu}_{03})/2$를 만족하고 있다. EM 알고리즘을 위한 추정치는 다양한 $\hat{\pi_0}^{(0)}$를 사용하였으나 동일한 결과를 얻었다. 처리시간은 약 4.12초(Intel Pentium(R)4, 2.53GHz) 였으며, 프로그램은 MATLAB으로 하였다.

5. 결론 및 제한점

본 논문에서는 $n \times p$ 크기의 마이크로어레이 자료 행렬에서 g개의 계급을 부분계급으로 분할하고, 부분계급의 평균자료를 사용함으로써 계급간 상관성을 유지하면서 g-변량 정규혼합모형을 이용해 효과적으로 DE 유전자를 식별할 수 있음을 보였다. 부분계급 평균화는 다차원에 의한 여러 문제점을 근본적으로 해결할 뿐만 아니라 혼합모형의 정규성을 보장해 준다는 이론적 근거를 제공한다. DE 유전자란 p개의 조직표본이 아닌 g개 계급에서 상이발현하는 유전자를 의미한다는 점에서 본 연구에서는 부분계급들의 평균이 이를 충족하도록 제약하면서 추정치를 얻는 EM 알고리즘을 제공하였다. 무엇보다도 제안된 방법은 처리속도가 빠르며, 모의자료와 실제 자료를 통해 실용적 타당성을 갖추었음을 보였다.

본 연구에서는 보다 광범위한 모의실험을 제공하지 못하였다. 특히, 계급내 개체들의 상관성이 반영된 실험이 이루어지 않았다. 또한, 제안된 방법은 유전자들 사이의 국소적 상관성을 반영하고 있지 않다. 이 문제는 DE 유전자 검출 분야에서 해결해야할 숙제로서 여전히 남아 있다.

참고문헌

김승구 (2007). Use of factor analyzer normal mixture model with mean pattern modeling on clustering genes. <한국통계학회논문집>, **13**, 113–123.

Allison, D. B. Gadbury, G. L., Heo, M., Ferández, J. R., Lee, C.-K., Prolla, T. A. and Weindruch, R. (2002). A mixture model approach for the analysis of microarray gene expression data. *Computational Statistics and Data Analysis*, **39**, 1–20.

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, 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. In *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 6745–6750.

Do, K.-A, Mueller, P. and Tang, F. (2005). A nonparametric Bayesian mixture model for gene expression. *Applied Statistics*, **54**, 1–18.

Efron, B. and Tibshirani, R. (2002). Empirical Bayes methods and false discovery rates for microarrays. *Genetic Epidemiology*, **23**, 70–86.

Efron, B. (2004). Large-scale simultaneous hypothesis testing : the choice of a null hypothesis. *Journal of the American Statistical Association*, **99**, 96–104.

He, Y., Pan, W. and Lin, J. (2006). Cluster analysis using multivariate normal mixture models to detect differential gene expression with microarray data. *Computational Statistics and Data Analysis*, **51**, 641–658.

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

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

Tusher, V. G., Tibshirani, R. and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. In *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 5116–5121.

Pawitan, Y., Murthy, K. R. K., Michiels, S. and Ploner, A. (2005). Bias in the estimation of false discovery rate in microarray studies. *Bioinformatics*, **21**, 3865–3872.

[Received October 2007, Accepted December 2007]