응용통계연구 제 19권 3호, 2006년, pp. 587–598

직교요인을 이용한 국소선형 로지스틱 마이크로어레이 자료의 판별분석*

백장선[1] 손영숙[2]

요 약

본 논문에서는 마이크로어레이(microarray) 자료에 판별분석을 적용 시 나타나는 고차원 및 소표본 문제의 해결방법으로서 직교요인을 새로운 특징변수로 사용한 비모수적 국소선형 로지스틱 판별분석을 제안한다. 제안된 방법은 국소우도에 기반한 것으로서 다범주 판별분석에 적용될 수 있으며, 고려된 직교인자는 주성분 요인, 부분최소제곱 요인, 인자분석 요인 등이다. 대표적인 두 가지 실제 마이크로어레이 자료에 적용한 결과 직교요인들 중에서 부분최소제곱 요인을 특징변수로 사용한 경우 고전적인 통계적 판별분석보다 향상된 분류 능력을 나타내고 있음을 확인하였다.

주요용어: 마이크로어레이, 국소우도, 로지스틱 판별분석, 부분 최소제곱 요인, 주성분분석 요인, 인자분석 요인

1. 서론

마이크로어레이 기술로 생물표본에 있는 거의 모든 유전자의 발현수준을 혼합과정을 통하여 신속하게 측정할 수 있다. 마이크로어레이 자료에 나타나는 유전자 발현 패턴을 분석함으로써 이미 다양한 문제를 해결하는데 도움을 주고 있으며, 특히 임상적인 신약 개발뿐만 아니라 생물학의 근본 의문을 해결하는데 공헌하고 있다. Golub *et al.*(1999)는 급성 백혈병의 분류를 위하여 마이크로어레이 유전자 발현 자료를 이용하였고, Alon *et al.*(1999)는 결장암 세포와 정상 세포들에 대하여 유전자 발현 자료를 이용한 군집분석을 시행하였다. Dudoit *et al.*(2002)는 선별된 유전자 자료를 이용하여 여러 가지 판별분석 방법들에 대한 비교 연구를 시행하였다. 마이크로어레이 자료에 대한 분석 도구들은 통계적 판별분석 방법, 베이지안 접근 방법, 그리고 Boosting, Bagging, Support vector machines, Neural network 등과 같은 기계학습 방법 등 여러 가지가 적용되고 있다.

마이크로어레이 연구에서 한 가지 뚜렷한 특징은 수집된 표본 수 n이 표본의 유전자 수 p(보통 수천 개)보다 훨씬 작다는 것이다. 통계적 관점에서 이 문제를 살펴보면 변수(유

* 본 연구는 산업자원부 지방기술혁신사업(RTI04-03-03) 지원으로 수행되었음.

1) (500-757) 광주광역시 북구 용봉동 300, 전남대학교 자연과학대학 통계학과, 교수
E-mail: jbaek@chonnam.ac.kr
2) (500-757) 광주광역시 북구 용봉동 300, 전남대학교 자연과학대학 통계학과, 교수
E-mail: ysson@chonnam.ac.kr

전자)의 수가 표본(마이크로어레이)수보다 비교할 수 없을 정도로 많기 때문에, 변수 선택 혹은 차원축소를 시행하지 않으면 고전적인 통계적 판별분석 방법을 적용하기가 곤란하다. 예를 들어 $n < p+1$이면 선형 판별 함수의 합동집단 내 표본공분산 행렬이 비정칙(singular)이 되는 문제점이 발생된다. 위에 언급된 연구들에서는 판별분석을 적용하기 전에 필요한 유전자들의 수를 줄이기 위하여 모두 일변량적인 방법들을 사용하였다. West *et al.* (2001)은 n개의 암세포들에 대한 p 개의 유전자 발현 프로파일을 열로 갖는 $p \times n$ 행렬에 대하여 SVD(singular value decomposition) 방법을 적용하여 유전자 수를 줄였다. Bicciato *et al.*(2003)은 유전자 자료에 대한 PCA(principal component analysis)로 차원축소를 하여 암의 표식 유전자를 찾고 분류하였다. Nguyen and Rocke(2002)는 차원 축소 방법으로 PLS(partial least squares)방법의 적용을 제안하고, 그렇게 추출된 요소들을 로지스틱(logistic), 이차판별분석(QDA: quadratic discriminant analysis)에 적용하였다. Martella (2006)과 McLachlan *et al.*(2002) 등은 그들의 혼합모형(mixture model)에 인자분석(FA: factor analysis) 모형을 삽입함으로써 차원축소를 통한 마이크로어레이 자료에 대한 판별분석과 군집분석을 수행하였다. Antoniadis *et al.*(2003)은 차원 축소 방법으로서 Xia *et al.*(2002)가 제안한 MAVE(minimum average variance estimation) 방법을 적용하였다.

Support vector machines, Neural network 등과 같은 기계적인 판별분석 방법은 여러 분야에 성공적으로 적용됨에도 불구하고 새로운 관측치의 분류집단 범주는 예측해내고 있으나 그에 대한 분류확률은 추정하지 못한다. 반면에 선형판별분석(LDA: linear discriminant analysis), QDA, 로지스틱 등 통계적인 분류방법들은 분류확률을 제공해 주는 장점이 있다. 한편 LDA, QDA, 로지스틱 판별분석은 선형 혹은 이차식의 판별경계에 의하여 관측치 들을 분류하므로 집단 간 경계가 일차 혹은 이차함수가 아닌 임의의 연속적인 함수에 의해 규정되는 경우 판별능력이 저하된다. Baek and Son(2006)은 판별집단이 여러 개인 경우 소속사후확률이 단조증가함수뿐만이 아니라 임의의 연속적인 매끄러운 함수일 때도 적용이 가능한 비모수적인 국소선형 로지스틱 판별분석 방법을 제안하였다.

비모수적 함수추정 방법에 의한 판별방법은 분포에 대한 가정 등을 필요로 하지 않아 일반적인 경우에 적용이 가능하지만 특징변수의 차원이 증가함에 따라 막대한 양의 자료가 소요된다 (차원의 저주, curse of dimensionality). 마이크로어레이 자료는 그 특성상 고차원의 특징변수를 가지고 있으나 표본의 수는 변수차원에 비하여 매우 작으므로 마이크로어레이 자료의 판별분석에 국소선형 로지스틱 방법을 적용하기 의해서는 차원축소가 필요하다. 본 연구에서는 마이크로어레이 자료의 판별분석을 위한 차원축소 방법으로서 PCA, PLS, FA 등 직교변환을 고려하였으며, 각 직교변환 요인들을 새로운 특징변수로서 국소선형 로지스틱 판별분석에 사용할 경우 어느 요인들이 더욱 판별에 유용한지를 많은 선행연구에서 분석된 두 가지 실제 마이크로어레이 자료에 대한 분석을 통하여 밝히고자 한다.

본 논문의 제2장에서 국소선형 로지스틱 판별분석 방법을 요약하고, 제3장에서 차원축소 방법으로서 PCA, PLS, FA 등 세 가지 직교변환 방법들을 정리하겠다. 제4장에서 실제 마이크로어레이 자료인 백혈병 자료와 결장암 자료에 대하여 각 차원축소 요인들을 사용한 LDA, QDA, 로지스틱, 국소선형 로지스틱 판별분석 결과를 살펴보겠다. 마지막으로 결론 및 토의를 제5장에 정리한다.

2. 국소선형 로지스틱 판별분석

분류해야 할 K개의 집단 $(C_1, C_2, \ldots, C_K)$이 있고, n 개의 훈련자료 $\{(\boldsymbol{y_i}, \boldsymbol{x_i}); i = 1, 2, \ldots, n\}$이 있다고 하자. 이 때 $\boldsymbol{x_i} = (x_{i1}, x_{i2}, \ldots, x_{ip})'$ 는 p-차원의 특징벡터이고, $\boldsymbol{y_i} = (y_{i1}, y_{i2}, \ldots, y_{i(k-1)})'$ 는 $\boldsymbol{x_i}$의 소속을 알려주는 $(K-1)$-차원의 베르누이 확률변수이다. 즉, 만약 $\boldsymbol{x_i}$가 l번째 집단 C_l에 속해 있으면 $\boldsymbol{y_i} = (y_{i1}, y_{i2}, \ldots, y_{i(k-1)})'$ 는 $y_{il} = 1$ 과 모든 $j \neq l$에 대하여 $y_{ij} = 0$ 값을 갖는다. 그리고 만약 $\boldsymbol{x_i}$가 K번째 집단 C_K에 속해 있으면 $\boldsymbol{y_i} = (0, 0, \ldots, 0)'$이다.

$\pi_l(\boldsymbol{x}) = P(C_l|\boldsymbol{x})$를 특징변수 값이 $\boldsymbol{x}$인 관측치가 집단 C_l에 속할 확률이라 하자. C_K를 기준 범주라고 하면 로지스틱 회귀모형은 로그승산 $\log(\pi_l(\boldsymbol{x})/\pi_K(\boldsymbol{x}))$을 특징변수들 $x_1, x_2, \ldots, x_p$ 의 함수 $\eta_l(\boldsymbol{x})$로 가정한다. 그러면 관측벡터 $\boldsymbol{x}$를 갖는 관측치가 집단 C_l과 C_K에 속할 확률은 각각 다음과 같다.

$$\begin{aligned} \pi_l(\boldsymbol{x}) &= \exp(\eta_l(\boldsymbol{x}))/\{1 + \textstyle\sum_{k=1}^{K-1} \exp(\eta_k(\boldsymbol{x}))\}, \quad l = 1, 2, \ldots, K-1, \\ \pi_K(\boldsymbol{x}) &= 1/\{1 + \textstyle\sum_{k=1}^{K-1} \exp(\eta_k(\boldsymbol{x}))\}. \end{aligned} \tag{2.1}$$

보통의 로지스틱 회귀모형에서는 $\eta_l(\boldsymbol{x})$를 링크함수라 부르며 다음과 같이 선형함수로 가정한다.

$$\eta_l(\boldsymbol{x}) = \beta_{l0} + \beta_{l1} x_1 + \cdots + \beta_{lp} x_p, \quad l = 1, 2, \ldots, K-1.$$

이차 로지스틱모형(Anderson, 1975)에서는 $\eta_l(\boldsymbol{x}) = \beta_{l0} + \beta_{l1} x_1 + \beta_{l2} x_2 + \cdots + \beta_{lp} x_p + \sum_{1 \leq i,j \leq p} \beta_{lij} x_i x_j$로 가정하기도 하지만 추정해야 할 모수의 수가 증가하므로 실제 사용에 있어서 제한적이다.

$\boldsymbol{y}$가 다항분포를 따르므로 로그우도함수 $l = \log L$는

$$l = \sum_{i=1}^{n} \left(\sum_{k=1}^{K-1} y_{ik} \log\{\pi_k(\boldsymbol{x_i})\} + (1 - \sum_{k=1}^{K-1} y_{ik}) \log\{1 - \sum_{k=1}^{K-1} \pi_k(\boldsymbol{x_i})\} \right)$$

이다. 모수 $\boldsymbol{\beta_l}' = (\beta_{l0}, \beta_{l1}, \ldots, \beta_{lp})$는 로그우도함수 l을 최대화하여 추정한다. 소속확률 $\pi_l(\boldsymbol{x})$ 는 $\boldsymbol{\beta_l}$를 최우추정치 $\hat{\boldsymbol{\beta}}_{\boldsymbol{l}}$로 대체한 $\hat{\pi}_l(\boldsymbol{x}; \hat{\boldsymbol{\beta}}_{\boldsymbol{l}})$로 추정하며, 새로운 관측치는 가장 큰 추정확률을 갖는 집단으로 분류한다. 이러한 판별분석을 (선형) 로지스틱 판별분석이라 한다.

선형 로지스틱 판별분석의 장점은 모형 모수 $\boldsymbol{\beta_l}$에 대한 해석이 쉽다는 점이다. 그러나 만약 링크함수가 선형과 거리가 멀다면 선형 로지스틱모형은 자료를 잘 적합 시키지 못할 것이다. 곡면의 링크함수관계를 잘 설명하면서도 모수에 대한 해석을 손쉽게 할 수 있는 한 가지 대안으로서 링크함수를 국소다항식에 의하여 근사화하는 비모수적인 방법을 고안할 수 있다. 판별집단이 두개인 경우 국소우도에 의한 로지스틱 접근방법이 Loader(1999)에 예시되어 있으며, Fan and Gijbels(1996)은 이항회귀모형의 국소적합을 연구하였다.

특징벡터 $\boldsymbol{x_0}$ 에서 링크함수 $\eta_l(\boldsymbol{x})$ 의 일차 미분치가 존재한다고 할 때, 차수 1인 다항식에 의하여 미지의 링크함수 $\eta_l(\boldsymbol{x})$ 를 국소적으로 근사할 수 있다. 만약 특징벡터 $\boldsymbol{x_0}$를 갖는 관측치를 분류하고자 할 때, $\boldsymbol{x_0}$의 이웃에 있는 $\boldsymbol{x}$에 대하여 링크함수에 대한 테일러 전개는

다음과 같다.

$$\begin{aligned}\eta_l(\boldsymbol{x}) &\approx \eta_l(\boldsymbol{x_0}) + \frac{\partial \eta_l(\boldsymbol{x_0})}{\partial \boldsymbol{x}'}(\boldsymbol{x}-\boldsymbol{x_0}) \\ &= \beta_{l0} + \beta_{l1}(x_1 - x_{01}) + \beta_{l2}(x_2 - x_{02}) + \cdots + \beta_{lp}(x_p - x_{0p}) \\ &= \boldsymbol{\beta_l}'\boldsymbol{z}, \quad l = 1, 2, \ldots, K-1.\end{aligned}$$

이 때 $\boldsymbol{z}' = (1, x_1 - x_{01}, \ldots, x_p - x_{0p})$ 이며, $\beta_{l0} = \eta_l(\boldsymbol{x_0})$, $\beta_{lj} = \partial\eta_l(\boldsymbol{x_0})/\partial x_j$, $j = 1, 2, \ldots, p$ 에 대하여 $\boldsymbol{\beta_l}' = (\beta_{l0}, \beta_{l1}, \ldots, \beta_{lp})$ 이다. 다항식의 계수 $\boldsymbol{\beta_T} = (\boldsymbol{\beta_1}', \boldsymbol{\beta_2}', \ldots, \boldsymbol{\beta_{K-1}}')'$는 다음의 국소 로그우도함수 $l^*(\boldsymbol{\beta_T})$를 최대화함으로써 추정할 수 있다;

$$l^*(\boldsymbol{\beta_T}) = \sum_{i=1}^{n} K_B(\boldsymbol{z_i}) \left(\sum_{k=1}^{K-1} y_{ik} \log\{\pi_k(\boldsymbol{z_i})\} + (1 - \sum_{k=1}^{K-1} y_{ik}) \log\{\pi_k(\boldsymbol{z_i})\} \right). \tag{2.2}$$

이 때 $K_B(\boldsymbol{u}) = K(B^{-1}\boldsymbol{u})/|B|$는 $p-$차원의 커널함수이며 B는 정칙 $p \times p$ 평활모수행렬이다. 본 연구에서는 다변량 표준정규분포를 커널 $K(\cdot)$로 사용하였으며, $K_B(\cdot)$는 국소우도함수에서 가중치로서 역할을 수행한다.

국소 스코어 통계량을 $U(\boldsymbol{\beta_T}) = \partial l^*(\boldsymbol{\beta_T})/\partial \boldsymbol{\beta_T}$ 로 정의하며, 국소 스코어 방정식 $U(\boldsymbol{\beta_T}) = 0$의 해로서 모수들을 추정한다. 방정식의 해는 다음의 피셔(Fisher) 스코어 알고리즘에 의하여 반복적인 방법으로 얻어진다.

$$\hat{\boldsymbol{\beta}}_T^{(s+1)} = \hat{\boldsymbol{\beta}}_T^{(s)} + I(\hat{\boldsymbol{\beta}}_T^{(s)})^{-1} U(\hat{\boldsymbol{\beta}}_T^{(s)}), \quad s = 0, 1, 2, \ldots.$$

이 때 $I(\boldsymbol{\beta}_T)$는 피셔 정보행렬로서 식 (2.3)과 같다. 모든 모수가 추정이 되면 링크함수 값은 $\hat{\eta}_l(\boldsymbol{x_0}) = \hat{\beta}_{l0}$로서 추정되며, 특징벡터 값 $\boldsymbol{x_0}$을 갖는 관측치는 $\hat{\eta}_l(\boldsymbol{x_0}) = \hat{\beta}_{l0}$에 의하여 계산되는 $\hat{\pi}_l(\boldsymbol{x_0})$ 값들 중 최대값을 가진 집단으로 분류된다. 이러한 분류방법을 국소 선형 로지스틱 판별분석이라 부르기로 한다(Baek and Son, 2006).

$\boldsymbol{z_i}' = (1, x_{i1} - x_{01}, x_{i2} - x_{02}, \ldots, x_{ip} - x_{0p})$ 이라하면, π의 $\boldsymbol{\beta_l}$에 대한 미분치는 다음과 같다.

$$\begin{aligned}\partial\pi_l/\partial\boldsymbol{\beta_l} &= \pi_l(1-\pi_l)\boldsymbol{z} \\ \partial\pi_k/\partial\boldsymbol{\beta_l} &= -\pi_k\pi_l\boldsymbol{z}, \quad k \neq l,\ k = 1, 2, \ldots, K.\end{aligned}$$

또한 스코어 통계량 $U(\boldsymbol{\beta_T}) = (U_1(\boldsymbol{\beta_1})', U_2(\boldsymbol{\beta_2})', \ldots, U_{K-1}(\boldsymbol{\beta_{K-1}})')'$의 l번째 요소 $U_l(\boldsymbol{\beta_l}) = \partial l^*(\boldsymbol{\beta_T})/\partial\boldsymbol{\beta_l}$는 식 (2.2)를 편미분하여 구하면 $U_l(\boldsymbol{\beta_l}) = \sum_{i=1}^{n} K_B(\boldsymbol{z_i})\{y_{il} - \pi_l(\boldsymbol{z_i})\}\boldsymbol{z_i}$이다.

피셔 정보행렬은

$$I(\boldsymbol{\beta_T}) = E\left(-\frac{\partial^2 l^*(\boldsymbol{\beta_T})}{\partial\boldsymbol{\beta_T}\partial\boldsymbol{\beta_T}'}\right) \tag{2.3}$$

로서 정의되며, 자료가 모수에 대한 정보를 얼마나 많이 담고 있는지를 의미한다. 모수 $\boldsymbol{\beta_T}$가 $(p+1)(K-1) \times 1$ 차원 벡터이므로 피셔 정보행렬 $I(\boldsymbol{\beta_T})$은 $(p+1)(K-1) \times (p+1)(K-1)$ 차원이다. 피셔 정보행렬의 (j,k) 요소는 $I_{jk}(\boldsymbol{\beta_T}) = E(-\partial U_j(\boldsymbol{\beta_j})/\partial\boldsymbol{\beta_k})$이므로 다음과 같이

구해진다.

$$\begin{aligned} I_{jk}(\boldsymbol{\beta_T}) &= \sum_{i=1}^{n} K_B(\boldsymbol{z_i})\pi_j(\boldsymbol{z_i})\{1-\pi_j(\boldsymbol{z_i})\}\boldsymbol{z_i}\boldsymbol{z_i}', \quad j=k, \\ &= -\sum_{i=1}^{n} K_B(\boldsymbol{z_i})\pi_j(\boldsymbol{z_i})\pi_k(\boldsymbol{z_i})\boldsymbol{z_i}\boldsymbol{z_i}', \quad j\neq k. \end{aligned} \tag{2.4}$$

$\boldsymbol{Z}$를 자료행렬 즉, $\boldsymbol{Z}=(\boldsymbol{z_1}', \boldsymbol{z_2}', \ldots, \boldsymbol{z_n}')'$이라하고, $\boldsymbol{W_j}$를 대각요소 $w_{ji}=K_B(\boldsymbol{z_i})\pi_j(\boldsymbol{z_i})\times\{1-\pi_j(\boldsymbol{z_i})\}$를 가진 대각행렬이라 하면, 식 (2.4)의 $I_{jj}(\boldsymbol{\beta_T})$는 다음과 같이 $\boldsymbol{Z}$와 $\boldsymbol{W_j}$의 행렬 곱으로 표현된다.

$$I_{jj}(\boldsymbol{\beta_T}) = \boldsymbol{Z}'\boldsymbol{W_j}\boldsymbol{Z}, \quad j=1,2,\ldots,K-1.$$

$K_B(\boldsymbol{z_i})>0$, $0<\pi_j(\boldsymbol{z_i})<1$이고 대각행렬 $\boldsymbol{W_j}$이 정칙이므로 $rank(\boldsymbol{W_j}\boldsymbol{Z})=rank(\boldsymbol{Z})$ 이며, 따라서

$$rank(I_{jj}(\boldsymbol{\beta_T})) = rank(\boldsymbol{Z}'\boldsymbol{W_j}\boldsymbol{Z}) \leq \min(rank(\boldsymbol{Z}'), rank(\boldsymbol{W_j}\boldsymbol{Z})) = rank(\boldsymbol{Z})$$

이다. 그러므로 $\boldsymbol{Z}$이 완전계수(full rank)를 갖지 못하면 j번째 정보행렬 $I_{jj}(\boldsymbol{\beta_T})$의 역행렬이 존재하지 않으며 따라서 피셔 정보행렬 $I(\boldsymbol{\beta_T})$의 역행렬 또한 존재하지 않는다. 이 경우 피셔 스코어 알고리즘에 의한 모수 추정이 불안정하게 된다. 자료행렬 $\boldsymbol{Z}$이 완전계수 $(p+1)$를 갖지 못하게 되는 것은 특징변수들 간에 다중공선성이 존재하는 경우 발생하며 특히 특징변수의 수가 많은 경우 발생할 가능성이 높다. 특징변수 간 다중공선성을 제거하고 동시에 비모수적 판별분석에 대한 차원의 저주를 회피하기 위한 차원축소의 접근방법으로는 변수선택(variable selection) 방법과 특징변환(feature transformation) 방법이 있다. 우리는 특징변환의 하나인 직교변환(orthogonal transformation) 방법들을 고려할 것이다.

3. 직교변환 방법

특징변수의 차원이 증가함에 따라 비모수 국소추정방법으로 정확하게 모수를 추정하기 위해서는 막대한 양의 자료가 필요하다. 따라서 비단 피셔 정보행렬의 비정칙성(singularity)을 해결하기 위해서 뿐만 아니라 차원의 저주를 극복하기 위해서도 차원의 축소가 불가피하다. 차원축소의 목적은 높은 $p-$차원의 원 특징(original feature) 공간을 낮은 $M-$차원의 요인(component) 공간으로 감소시키는 것이다 ($M<<p$). 이것은 미리 정해진 목적기준을 최적화시킬 수 있는 M개의 요인들을 추출함으로써 달성할 수 있다. 우리는 차원축소 방법으로서 원 자료의 정보를 거의 손실 없이 내포하는 동시에 새롭게 추출된 요인들로 이루어진 자료행렬의 정칙성을 보장할 수 있는 직교변환 방법을 고려하기로 한다.

특징차원의 축소를 위해서 지금까지 가장 일반적으로 사용되어온 직교변환 방법은 주성분분석 방법(PCA)이다. PCA는 총 특징 변동을 가능한 많이 설명할 수 있도록 소수의 특징요인을 추출하는 방법이다. PCA는 추출되는 요인의 분산을 최대화할 수 있도록 축차적으로 원래 특징변수들의 직교선형결합을 새로운 특징 요인으로 추출하는 과정이다. 이 절차는 다음 식을 만족하는 가중치 벡터 $\boldsymbol{a}_m$를 찾는 과정이다.

$$\boldsymbol{a}_m = \arg\max_{\boldsymbol{a}'\boldsymbol{a}=1} Var(\boldsymbol{X}\boldsymbol{a}), \quad m=1,2,\ldots,M.$$

각 단계의 해는 $\boldsymbol{S} = \boldsymbol{X}'\boldsymbol{X}$이라 할 때, 다음의 직교성을 만족해야 한다.

$$\boldsymbol{a}_m'\boldsymbol{S}\boldsymbol{a}_j = 0, \quad 1 \le j < m.$$

해 $\boldsymbol{a}_m$은 $\boldsymbol{S}/(n-1)$의 m번째 고유값 λ_m에 대응하는 고유벡터이다. m번째 주요인(principal component)은 원래 특징변수들의 선형결합인 $\boldsymbol{X}\boldsymbol{a}_m$이며, 이렇게 추출된 M개의 주요인 점수들이 판별분석 방법에 입력되는 새로운 특징자료이다.

두 번째 직교변환 방법은 부분최소제곱(PLS) 방법이다. PCA와는 달리 PLS는 집단변수와 원래특징변수의 선형결합 간 공분산이 최대가 되도록 요인들을 추출해가는 방법이다. $\boldsymbol{X}$를 표준화된 $n \times p$ 원 자료행렬이라 하고, $\boldsymbol{Y} = (\boldsymbol{y_1}, \ldots, \boldsymbol{y_n})'$를 $n \times (K-1)$ 다항 집단표시행렬(multinomial class indicator matrix)이라 하자. PLS는 다음을 만족하는 가중치 벡터 $\boldsymbol{b_m}$를 축차적으로 찾아가는 과정이다.

$$\boldsymbol{b_m} = \arg \max_{\boldsymbol{b'b}=1, \boldsymbol{c'c}=1} Cov^2(\boldsymbol{Xb}, \boldsymbol{Yc}), \quad m = 1, 2, \ldots, M.$$

이 때 $\boldsymbol{b}$, $\boldsymbol{c}$는 단위벡터(unit vector)들이며, $\boldsymbol{S} = \boldsymbol{X}'\boldsymbol{X}$이라 할 때 각 단계의 해는 다음의 직교성을 만족해야 한다.

$$\boldsymbol{b_m}'\boldsymbol{S}\boldsymbol{b_j} = 0, \quad 1 \le j < m.$$

m번째 PLS 요인은 원래 특징들의 선형결합 $\boldsymbol{X}\boldsymbol{b_m}$이다. PLS 방법은 요인과 집단 변수 간 상관관계가 최대가 되도록 요인들을 추출하는 것이다.

인자분석(FA) 방법은 다변량 자료의 특징변수들 간의 상관관계를 설명하기 위하여 많이 사용될 뿐만 아니라 차원축소를 위해서도 사용된다. Martella (2006)과 McLachlan *et al.*(2002) 등은 마이크로어레이 자료에 대한 판별분석과 군집분석에서 그들의 혼합모형에 인자분석 모형을 삽입하여 차원축소를 통한 분석을 수행하였다. 요인분석 모형에서 $\boldsymbol{x}_i$는 다음과 같이 모형화된다.

$$\boldsymbol{x}_i = \boldsymbol{\mu} + \boldsymbol{B}\boldsymbol{u}_i + \boldsymbol{e}_i, \quad , i = 1, \ldots, n.$$

이 때 $\boldsymbol{u}_i$는 M차원의 ($M << p$)의 인자(factor)라 불리는 잠재변수(latent variable)이며, $\boldsymbol{B}$는 $p \times M$ 인자적재행렬이다. 보통 $\boldsymbol{u}_i$는 동일하고 독립적인 정규분포 $N(\boldsymbol{0}, \boldsymbol{I}_M)$를 따르며, 오차 $\boldsymbol{e}_i$와는 서로 독립적이라고 가정한다. 이 때 $\boldsymbol{I}_M$는 단위행렬이며, $\boldsymbol{e}_i$는 $\boldsymbol{D} = diag(\sigma_1^2, \ldots, \sigma_p^2)$이 공분산 행렬인 $N(\boldsymbol{0}, \boldsymbol{D})$의 정규분포를 따른다고 가정한다. 따라서 $\boldsymbol{x}_i$는 $N(\boldsymbol{\mu}, \Sigma = \boldsymbol{B}\boldsymbol{B}' + \boldsymbol{D})$를 따르며, 인자분석은 $\boldsymbol{x}_i$의 공분산행렬을 가장 잘 모형화 하는 $\boldsymbol{B}$와 $\boldsymbol{D}$를 찾는 것이다. $E(\boldsymbol{u}|\boldsymbol{x}) = \boldsymbol{B}'\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu})$이므로 인자 $\boldsymbol{u}$는 미지의 모수들인 $\boldsymbol{B}$, Σ, $\boldsymbol{\mu}$들을 각각의 추정량으로 대체하여 추정한다. 이렇게 추정된 인자들을 인자점수(factor score)라 부르며 새로운 특징 요인으로 삼는다.

4. 마이크로어레이 자료의 판별분석

4.1. 백혈병 마이크로어레이 자료

첫 번째 분석 자료는 두 가지 형태의 백혈병(ALL: acute lymphoblastic leukemia, AML:

acute myeloid leukemia) 유전자 발현에 대한 연구 자료이다(Golub *et.al.*, 1999). 이 자료는 72개의 세포조직 표본으로부터 추출한 7129개의 인간 유전자발현 측정값으로 이루어져있다. 72개 표본자료는 다시 38개의 훈련자료(27 ALL, 11 AML)와 34개의 시험자료(20 ALL, 14 AML)로 나뉘어져있다.

측정기계로 부터 생산된 유전자 발현측정값들은 칩(chip) 효과, 배경 밀도, RNA 추출 변이, 염료 효율 등 실험과 관련하여 발생된 변동 요인을 제거하기 위하여 측정값에 대한 변환이 필요하다. 여러 가지 표준화와 변환 방법이 제안되었지만(Alizadeh *et al.*, 2000; Bolstard *et al.*, 2003; Yeung *et al.*, 2001) 아직 최선의 방법이 존재하는 것은 아니다.

우리는 McLachlan *et al.*(2002)와 Dudoit *et al.*(2002)에서 적용한 방법과 유사한 아래와 같은 사전처리 방법을 시행하였다. 이 자료에 대한 이전 연구 결과들과 직접적인 비교를 하기 위하여 동일한 방법을 적용한 것이다. 사전처리 단계는 다음과 같다:

1. 모든 유전자 발현값들을 (100, 16,000) 범위 안으로 제한시킨다. 즉, 발현값이 16,000 이상이면 16,000으로 삭감시킨다. 왜냐하면 이 값 이상에서는 형광 포화상태가 발생하며, 따라서 본래의 측정값이 정확하게 측정될 수가 없기 때문이다. 100 미만의 측정값들은 100으로 규정되며 이는 분석에서 100의 측정값들과 동일한 역할을 하기 때문이다.

2. 어느 특정 유전자의 표본 내 최대값과 최소값을 $\max, \min$ 이라 할 때, $(max/min) \leq 5$ 혹은 $(\max - \min) \leq 500$인 유전자 자료들을 제거한다. 즉 분석하는 표본 내 변동이 작은 유전자들을 분석에서 제외시킨다.

3. $\log_{10}$ 변환을 시행한다.

4. 유전자별로 평균 0, 표준편차 1을 갖도록 자료를 표준화한다. 그 다음 표본별로도 동일한 표준화를 시행한다. 이렇게 함으로써 체계적인 변동을 제거할 수 있다(Dudoit and Fridlyand, 2003).

위와 같이 원래 유전자들에 대하여 사전처리를 하여도 여전히 표본 수보다 많은 수천개의 유전자들이 특징변수로 남게 되며 이들 모두를 직교변환 할 수도 있지만, 이들 모두가 판별에 유효한 것은 아니다. 따라서 또 한 번의 변수선택이 필요하며, 우리는 Nguyen and Rocke(2002)에서와 같이 다음과 같은 t 통계량에 의해서 유전자들을 두 집단 간 차이가 많이 나는 순서대로 순위화하여 일정 개수의 유전자들을 선택하였다.

$$t = \frac{\bar{x}_0 - \bar{x}_1}{\sqrt{s_0^2/n_0 + s_1^2/n_1}}.$$

이 때 $n_k, \bar{x}_k$와 s_k^2은 각각 k번째 집단의 표본 수, 표본평균, 표본분산이다, $k = 1, 2$. 각 유전자에 대하여 t통계량 값을 계산하고 가장 큰 양의 값을 갖는 유전자 $p^*/2$개와 가장 작은 음의 값을 갖는 유전자 $p^*/2$개를 골라냄으로써 모두 p^*개의 유전자들을 선택하였다. 우리는 이렇게 선택된 $p^* = 50$개의 원래 유전자 발현값들에 대하여 각 직교변환을 실시하여 가

표 4.1: 직교요인별 백혈병 마이크로어레이 시험자료 분류결과

p^*	요인개수(M)	요인	LDA	QDA	로지스틱	국소선형로지스틱
		PLS	33	33	34	33
	1	PCA	33	33	33	33
		FA	33	33	33	34
		PLS	34	34	34	34
50	2	PCA	33	33	33	34
		FA	34	33	32	34
		PLS	34	34	34	34
	3	PCA	33	33	33	33
		FA	33	33	32	34

장 설명력이 우수한 처음 M개의 직교요인들을 계산하고 이들을 새로운 특징변수로 LDA, QDA, 로지스틱, 국소선형 로지스틱 판별기에 입력하였다.

표 4.1은 34개의 시험자료에 대하여 각각 $M = 1, 2, 3$ 개의 직교요인을 사용하여 각 판별방법을 시행했을 때의 분류결과를 나타내고 있다. 국소선형 로지스틱 방법에 사용된 커널은 다변량 표준 정규분포로서 각 변수별 동일한 평활모수를 사용하였다. 최적 평활모수는 $h = 0.1$에서부터 $h = 2$까지 100개의 평활모수 중 가장 최선의 분류능력을 가진 것을 선택하였다. $M = 2, 3$개의 직교요인을 사용했을 때 모든 직교요인에 대하여 국소선형 로지스틱 판별방법이 기존의 통계적 판별방법보다 우수하거나 동일한 분류결과를 보여주며 특히 $M = 2$개를 사용했을 때 어느 직교요인을 사용하던지 모든 시험자료를 정 분류함으로써 판별력이 가장 높았다. Nguyen and Rocke(2002)는 PLS를 몇 개 사용했는지는 보고하지 않았지만 로지스틱 방법과 QDA에 적용하여 최대 33개의 정 분류율을 기록하였다. 한편 직교요인 별로 비교해보면 역시 $M = 2, 3$에 대하여 PLS 요인이 다른 직교요인들 보다 어느 판별기를 사용하든지 가장 우수한 판별능력을 가지고 있음을 확인할 수 있다. 국소선형 판별기의 경우 이 자료에서는 FA가 PCA보다 우수하며 PLS와 동일한 판별능력을 보여준다.

4.2. 결장 마이크로어레이 자료

두 번째 분석 자료는 Alon *et al.*(1999)의 결장자료(colon data)로서 2,000개의 유전자에 근거하여 세포조직을 분류하고자 한다. 원 자료는 40개의 암세포 조직과 22개의 정상세포 조직에 대한 6,500개 인간 유전자 발현의 절대 관측값으로 구성되어 있다. Alon *et al.*(1999)에서는 표본 내 가장 높은 최소 휘도(highest minimal intensity)를 가진 2,000개의 유전자만을 분석하였으며 본 연구의 분석 자료도 이 2,000개의 유전자들로 구성되어 있다. 따라서 자료행렬은 $p = 2,000$개의 열과 $n = 62$개의 행들로 이루어져있다. 이 자료 역시 다음과 같은 사전 처리가 필요하다. 먼저 자료행렬의 모든 발현값에 대하여 자연로그를 취하

표 4.2: 직교요인별 결장 마이크로어레이 자료 분류결과

p^*	요인개수(M)	요인	LDA	QDA	로지스틱	국소선형로지스틱
		PLS	55	55	56	56
	1	PCA	55	55	56	56
		FA	56	56	56	56
		PLS	57	57	57	58
30	2	PCA	56	55	56	56
		FA	53	55	55	56
		PLS	57	57	57	58
	3	PCA	56	55	56	55
		FA	55	55	55	55
		PLS	54	55	55	55
	1	PCA	53	54	55	55
		FA	53	53	55	55
		PLS	57	57	57	59
50	2	PCA	55	55	56	56
		FA	53	54	55	56
		PLS	58	58	58	59
	3	PCA	56	55	56	56
		FA	54	53	54	55

고, 각 열들이 평균 0, 표준편차 1이 되도록 표준화를 시행하였다. 마지막으로 이렇게 표준화된 행렬에 대하여 각 행들이 평균 0, 표준편차 1이 되도록 표준화를 실시하였다. 이 자료 역시 백혈병 자료와 같이 t통계량 점수에 의하여 동일한 방법으로 $p^* = 30, 50$ 개의 유전자들을 선별하여 각각 직교변환을 실시하였다.

표 4.2는 총 62개의 자료에 대하여 각각 $M = 1, 2, 3$개의 직교요인을 사용하여 각 분류방법을 시행했을 때 교차타당성(leave-one-out cross validation) 방법에 의한 분류결과를 나타내고 있다. 국소선형 로지스틱 방법은 역시 다변량 표준 정규분포를 커널로 사용하였으며, 교차타당성에 의하여 $h = 0.1$에서부터 $h = 2$ 사이에서 최적 평활모수를 선택하였다. $p^* = 30$일 때보다 $p^* = 50$일 때가 전반적으로 정 분류율이 높다. $M = 1, 2, 3$개의 직교요인을 사용했을 때 모든 직교요인에 대하여 국소선형 로지스틱 판별방법이 기존의 통계적 판별방법보다 우수하거나 동일한 분류결과를 보여주며 특히 PLS를 2개 혹은 3개를 사용했을 때 59개 정 분류를 하여 가장 우수한 판별력을 보였다. 이 자료에 대하여 Nguyen and Rocke(2002)는 PLS를 로지스틱 방법에 적용했을 때 가장 높은 58개의 정 분류를 보고하였다. 한편 직교요인 별로 비교해보면 역시 $M = 1, 2, 3$ 모두에서 PLS 요인이 다른 직교요인

들 보다 어느 판별기를 사용하든지 가장 우수한 판별능력을 보여주었다.

5. 결론 및 토의

비모수적 국소선형 로지스틱 판별분석은 국소 우도에 기반한 분류방법으로서 로지스틱 방법의 장점을 갖고 있으면서도 집단 간 경계가 특정 모수적 함수가 아닌 경우에도 잘 분류할 수 있는 방법이다. 그러나 특징변수의 수가 표본의 수보다 훨씬 많은 특징을 가지고 있는 마이크로어레이 자료의 분류에 이 방법을 사용할 경우 차원의 저주를 극복하여야 한다. 본 논문에서는 주성분분석, 부분최소제곱방법, 인자분석 등 직교변환에 의한 차원축소 방법을 고려하였으며 그것들을 새로운 특징변수로서 국소선형 로지스틱 판별방법에 사용하였을 때 어느 것이 가장 효과적인지를 알아보고자 하였다. 대표적인 두 가지 실제 마이크로어레이 자료에 적용한 결과 직교요인들 중에서 부분최소제곱 요인을 특징변수로 사용한 경우 고전적인 통계적 판별분석보다 향상된 분류 능력을 나타내고 있음을 확인하였다.

주성분 요인이 다른 요인들보다 판별정보가 부족한 이유는 요인을 추출할 때 집단표지(class label)에 대한 정보를 사용하지 않았기 때문이다. 반면에 부분최소제곱 방법은 원래 특징변수의 선형결합으로서 그 요인을 구성할 때 집단표지에 대한 정보를 활용하였기 때문에 주성분분석 보다 집단을 더 잘 예측할 수 있는 선형결합 가중치들을 부여하게 된다. 또한 인자분석은 적은 수의 잠재변수들을 이용하여 고차원 자료의 공분산을 모형화한 것으로서 주로 혼합모형에 삽입하여 마이크로어레이 자료를 분류하였다. 인자 요인 역시 집단표지 정보의 활용 없이 구성되었기 때문에 실제 자료 분석에서도 확인할 수 있듯이 부분최소제곱 요인보다 판별정보를 덜 가진 것으로 생각된다. 부분최소제곱 요인을 포함하여 이 세 가지 직교 요인들에 대한 단점으로는 모든 원래 특징변수들에 대하여 가중치 적재값이 존재하기 때문에 요인에 대한 해석이 쉽지 않다는 점이다.

참고문헌

Alizadeh, A. A. *et al.* (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling, *Nature*, **403**, 491-492.

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, *Proc. Nat. Acad. Sci.*, **96**, 6745-6750.

Anderson, J. A. (1975). Quadratic logistic discrimination, *Biometrika*, **62**, 149-154.

Antoniadis, A., Lambert-Lacroix, S. and Leblanc, F. (2003). Effective dimension reduction methods for tumor classification using gene expression data, *Bioinformatics*, **19**, 563-570.

Baek, J. and Son, Y. S. (2006). Local linear logistic discriminant analysis with partial least square components, To appear in *Lecture Notes in Artificial Intelligence (LNAI 4093)*.

Bicciato, S., Luchini, A. and Di Bello, C. (2003). PCA disjoint models for multiclass cancer analysis using gene expression data, *Bioinformatics*, **19**, 571-578.

Bolstard B. M. *et al.* (2003). A comparison of normalization methods for high density oligonucleotide array data based on bias and variance, *Bioinformatics*, **19**, 185-193.

Dudoit, S. *et al.* (2002). Comparison of discrimination methods for the classification of tumors using gene expression data, *J. Am. Stat. Assoc.*, **97**, 77-87.

Dudoit, S. and Fridlyand, J. (2003). Classification in microarray experiments, In Speed, T.P., *Statistical analysis of gene expression microarray data*, Chapman and Hall-CRC, New York.

Fan, J. and Gijbels, I. (1996). *Local polynomial modeling and its applications*, London: Chapman & Hall.

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

Loader, C. (1999). *Local regression and likelihood*, New York: Springer.

Martella, F. (2006). Classification of microarray data with factor mixture models, *Bioinformatics*, **22**, 202-208.

McLachlan, G. J. *et al.* (2002). A mixture model-based approach to the clustering of microarray expression data, *Bioinformatics*, **18**, 413-422.

Nguyen, D. and Rocke, D. (2002). Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, **18**, 39-50.

West, M., Blanchette, C., Dressman, H., Huang, F., Ishida, S., Spang, R., Zuzan, H., Olason, J., Marks, I., Nevins, J. (2001). Predicting the clinical status of human breast cancer by using gene expression profiles. *PNAS*, **98**, 11462-11467.

Xia, Y., Tong, H., Li, W. K. and X, Z. L. (2002). An adaptive estimation of dimension reduction space, *J. R. Statist. Soc. B.*, **64**, 363-410.

Yeung K. Y. *et al.* (2001). Model-based clustering and data transformations for gene expression data, *Bioinformatics*, **17**, 977-987.

[2006년 6월 접수, 2006년 7월 채택]

Local Linear Logistic Classification of Microarray Data Using Orthogonal Components*

Jangsun Baek[1] Young Sook Son[2]

ABSTRACT

The number of variables exceeds the number of samples in microarray data. We propose a nonparametric local linear logistic classification procedure using orthogonal components for classifying high-dimensional microarray data. The proposed method is based on the local likelihood and can be applied to multi-class classification. We applied the local linear logistic classification method using PCA, PLS, and factor analysis components as new features to Leukemia data and colon data, and compare the performance of the proposed method with the conventional statistical classification procedures. The proposed method outperforms the conventional ones for each component, and PLS has shown best performance when it is embedded in the proposed method among the three orthogonal components.

Keywords: Microarray, local likelihood, logistic classification, partial least squares component, principal component, factor analysis component

* This work was supported by grant No. RTI04-03-03 from the Regional Technology Innovation Program of the Ministry of Commerce, Industry and Energy(MOCIE).

1) Professor, Department of Statistics, Chonnam National University, Gwangju 500-757, Korea
E-mail: jbaek@chonnam.ac.kr

2) Professor, Department of Statistics, Chonnam National University, Gwangju 500-757, Korea
E-mail: ysson@chonnam.ac.kr