Journal of Korean
Data & Information Science Society
2008, Vol. 19, No. 3, pp. 751 ~ 759

Demension reduction for high-dimensional data via mixtures of common factor analyzers - an application to tumor classification[1)]

Jangsun Baek[2)]

Abstract

Mixtures of factor analyzers (MFA) is useful to model the distribution of high-dimensional data on much lower dimensional space where the number of observations is very large relative to their dimension . Mixtures of common factor analyzers (MCFA) can reduce further the number of parameters in the specification of the component covariance matrices as the number of classes is not small. Moreover, the factor scores of MCFA can be displayed in low-dimensional space to distinguish the groups. We propose the factor scores of MCFA as new low-dimensional features for classification of high-dimensional data. Compared with the conventional dimension reduction methods such as principal component analysis (PCA) and canonical covariates (CV), the proposed factor score was shown to have higher correct classification rates for three real data sets when it was used in parametric and nonparametric classifiers.

Keywords: Canonical covariates; Classification; Common factor loadings; Gene expression data; Mixtures of factor analyzers; Principal component analysis.

1. 서론

고차원 변수의 차원축소를 위해서 지금까지 가장 일반적으로 사용되어온 직교변환 방법은 주성분분석 (PCA)이다. PCA는 총 특징 변동을 가능한 많이 설명할 수 있도록 소수의 특징변수를 추출하는 방법이다. PCA는 추출되는 성분의 분산을 최대화할 수 있도록 축차적으로 원래 특징변수들의 직교선형결합을 새로운 특징변수로 추출하

1) 이 논문은 2005년도 전남대학교 연구년 교수연구비 지원에 의하여 연구되었음.
2) 광주광역시 북구 용봉동 300번지 전남대학교 통계학과 교수
E-mail : jbaek@chonnam.ac.kr

는 과정이다. 이 절차는 $\boldsymbol{Y}$ 를 p-차원 특징변수라 할 때, 다음 식을 만족하는 가중치 벡터 $\boldsymbol{a_m}$을 찾는 과정이다.

$$\boldsymbol{a_m} = \arg \max_{\boldsymbol{a'a}=1} Var(\boldsymbol{Ya}), \quad m = 1,2,...,q.$$

각 단계의 해는 $\boldsymbol{S}$를 표본공분산행렬이라 할 때, 다음의 직교성을 만족해야한다.

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

해 $\boldsymbol{a_m}$은 $\boldsymbol{S}$의 m번째 고유값 λ_m에 대응하는 고유벡터이다. m번째 주성분은 원래 특징변수들의 선형결합인 $\boldsymbol{Ya_m}$이며, 이렇게 추출된 q개의 주성분 점수들이 판별분석방법에 입력되는 새로운 특징변수 값이다. Bicciato *et al.* (2003)은 유전자 자료에 대하여 PCA로 차원축소를 하여 암의 표식 유전자를 찾고 분류하였다.

만약 고차원 자료의 각 관측값들이 g개의 집단 ($G_1, G_2, ..., G_g$) 중 어느 집단으로부터 발생했는지 그 소속이 알려져 있는 경우, 집단간제곱합과 교차곱행렬 ($\boldsymbol{B}$)과 집단내제곱합과 교차곱행렬 ($\boldsymbol{W}$)은 각각 다음과 같이 정의할 수 있다.

$$\boldsymbol{B} = \frac{1}{g-1}\sum_{i=1}^{g} n_i(\overline{\boldsymbol{y_i}} - \overline{\boldsymbol{y}})(\overline{\boldsymbol{y_i}} - \overline{\boldsymbol{y}})'$$

$$\boldsymbol{W} = \sum_{i=1}^{g}\sum_{j=1}^{n} z_{ij}(\boldsymbol{y_j} - \overline{\boldsymbol{y}})(\boldsymbol{y_j} - \overline{\boldsymbol{y}})'$$

이 때 n_i는 i번째 집단에 속한 관측값들의 개수이며, $\overline{\boldsymbol{y}}$, $\overline{\boldsymbol{y_i}}$는 각각 관측값들의 전체 평균과 i번째 집단에 속한 관측값들의 평균이다. z_{ij}는 $\boldsymbol{y_j}$가 i번째 집단 G_i에 속해있는지 아닌지에 따라 1 혹은 0 값을 취한다. $\boldsymbol{S_p} = \boldsymbol{W}/(n-g)$라고 할 때, p차원 관측값 $\boldsymbol{y}$에 대한 q차원 정준변량(canonical covariates: CV) 벡터 $\boldsymbol{v}$는 다음과 같이 정의된다.

$$\boldsymbol{v} = \boldsymbol{\Gamma_q}\boldsymbol{y},$$

이 때 $\boldsymbol{\Gamma_q} = (\boldsymbol{\gamma_1}, ..., \boldsymbol{\gamma_q})'$ 이며, $\boldsymbol{\gamma_1}$ 은 $\boldsymbol{\gamma}'\boldsymbol{B}\boldsymbol{\gamma}/\boldsymbol{\gamma}'\boldsymbol{S_p}\boldsymbol{\gamma}$ 비율을 최대화하는 벡터이다. $k = 2, ..., q$ 에 대하여 $\boldsymbol{\gamma_k}$는 $\boldsymbol{\gamma_k}'\boldsymbol{S_p}\boldsymbol{\gamma_h} = 0$ $(h = 1, ..., k-1)$ 조건하에서 $\boldsymbol{\gamma}'\boldsymbol{B}\boldsymbol{\gamma}/\boldsymbol{\gamma}'\boldsymbol{S_p}\boldsymbol{\gamma}$ 를 최대화하는 벡터이다. 따라서 h번째 정준변량 $\boldsymbol{\gamma_h}'\boldsymbol{y}$ 와 k번째 정준변량 $\boldsymbol{\gamma_k}'\boldsymbol{y}$ 사이의 상관계수가 0 이므로 서로 직교한다. 실제적으로 $\boldsymbol{\gamma_k}$ 는 $\boldsymbol{S_p}^{-1}\boldsymbol{B}$ 의 k번째로 큰 고유값에 대응하는 고유벡터이다. McLachlan (1992)는 CV에 의한 판별에 있어서 분리 양상에 대하여 자세히 다루고 있다.

본 논문의 제2장에서 공통요인 분석자 혼합모형에 따른 차원축소 방법을 요약하고, 제3장에서는 실제 유전자 발현자료인 두 가지 백혈병자료와 잡음이 추가된 한 가지 고차원 화학 성분 자료에 PCA, CV 그리고 요인점수를 모수적 판별분석방법인 LDA

와 QDA, 그리고 비모수적 판별분석 방법인 국소선형 로지스틱 판별분석에 각각 적용한 결과를 살펴보겠다. 마지막으로 결론 및 토의를 제4장에 정리한다.

2. 공통요인 분석자 혼합모형 차원축소

$\boldsymbol{Y}=(Y_1,...,Y_p)'$을 p-차원의 연속형 다변량 특징벡터라고 하자. 분석 자료 관측값들의 소속집단이 알려져 있지 않은 경우, 즉 몇 개의 집단으로부터 표본자료가 그 구성비율이 알려지지 않은 채 관측되었다면 몇 개의 각기 다른 분포들의 혼합분포로 이루어진 유한혼합모형 (finite mixture model)으로 군집분석을 수행할 수 있으며 (McLachlan and Peel (2000)), 이러한 접근방법을 모형 기반 군집화 (model based clustering) 이라한다. 이 때 혼합분포의 성분 분포로서 주로 정규분포를 이용하며, 그 경우 정규혼합모형 (Gaussian Mixture Model: GMM)이라 한다. 따라서 GMM에서는 $\boldsymbol{Y}$의 분포함수를 다음과 같이 g개의 다변량정규분포들의 혼합분포로 가정한다.

$$f(\boldsymbol{y},\boldsymbol{\Psi})=\sum_{i=1}^{g}\pi_i\phi(\boldsymbol{y};\boldsymbol{\mu_i},\boldsymbol{\Sigma_i}). \tag{2.1}$$

이 때, $\phi(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma})$는 평균 $\boldsymbol{\mu}$와 공분산행렬 $\boldsymbol{\Sigma}$를 가진 p-차원 다변량정규분포 밀도함수이며, 벡터 $\boldsymbol{\Psi}$는 혼합비율 π_i, 성분평균 $\boldsymbol{\mu_i}$, 그리고 성분공분산행렬 $\boldsymbol{\Sigma_i}$의 성분들로 이루어진 미지의 모수벡터이다 $(i=1,...,g)$. 관측표본이 $\boldsymbol{y_1},...,\boldsymbol{y_n}$이라 하면 $\boldsymbol{\Psi}$에 대한 로그우도함수는 $\log L(\boldsymbol{\Psi})=\sum_{j=1}^{n}\log f(\boldsymbol{y_j};\boldsymbol{\Psi})$로 유도되며 최우추정량 $\hat{\boldsymbol{\Psi}}$는 $\partial\log L(\boldsymbol{\Psi})/\partial\boldsymbol{\Psi}=0$ 의 해로서 EM (Expectation-Maximization) 알고리즘 (Demster *et al.* (1997))을 이용하여 구해진다. 모수추정 후 관측값에 대한 각 혼합성분의 사후확률을 추정하여 가장 높은 사후확률 추정값을 제공하는 성분집단에 할당함으로써 군집분석을 수행할 수 있다. Banfileld and Raftery (1993)는 GMM을 이용한 군집분석을 제안했으며, 또한 판별분석과 분포함수추정에도 GMM이 사용되었다 (Fraley and Raftery (2004)).

식 (2.1)의 GMM은 차원 p가 클 경우 각 $\boldsymbol{\Sigma_i}$가 $d=p(p+1)/2$개의 모수로 이루어져 있으므로 매우 모수가 많은 모형이다 $(i=1,...,g)$. Banfileld and Raftery (1993)은 모수의 수를 줄이기 위해 $\boldsymbol{\Sigma_i}$에 대하여 다양한 스펙트럼 분해에 기초한 모형화를 제안하였다. 그러나 차원 p가 표본크기 n에 비하여 매우 큰 경우 이러한 분해를 적용한다하더라도 여전히 성분공분산 행렬의 모형에 대한 적절한 추론이 가능하지 않을 수 있다. 모수 추정이 가능하더라도 n에 비해 p가 큰 경우 성분공분산 행렬의 추정치가 비정칙 (singular)에 가까울 수 있다.

성분공분산 행렬의 모수 수를 줄이기 위하여 Ghahramani and Hinton (1996)과 McLachlan and Peel (2000) 등은 요인분석자혼합모형(Mixtures of Factor Analyzers: MFA)을 제안하였다. MFA는 관측벡터에 대하여 각 성분별로 다음과 같이 요인분석 모형을 가정한다.

$$\boldsymbol{Y}_j = \boldsymbol{\mu}_i + \boldsymbol{A}_i \boldsymbol{V}_{ij} + \boldsymbol{e}_{ij}, \quad i = 1,...,g, \quad j = 1,...,n. \tag{2.2}$$

이 때 $\boldsymbol{V}_{ij}$는 q-차원 $(q < p)$의 요인이라고 불리는 비관측 잠재벡터이며 다변량정규분포 $N_q(\boldsymbol{0}, \boldsymbol{I}_q)$를 따른다. $\boldsymbol{A}_{ij}$는 $p \times q$차원의 요인적재행렬이다. 오차벡터 $\boldsymbol{e}_{ij}$는 $\boldsymbol{V}_{ij}$와 독립이며 대각행렬 $\boldsymbol{D}_i$을 공분산행렬로 가지는 $N_p(\boldsymbol{0}, \boldsymbol{D}_i)$를 따른다. MFA에 따르면 성분공분산 행렬은 $\boldsymbol{\Sigma}_i = \boldsymbol{A}_i \boldsymbol{A}_i' + \boldsymbol{D}_i$로 표현된다. 따라서 p보다 충분히 작은 q를 선택한다면 MFA의 모수 수는 GMM의 그것보다도 매우 적게 된다. 참고로 MFA의 모수 수는 $d_1 = (g-1) + 2gp + g(pq - q(q-1)/2)$이다.

MFA가 GMM에 비해 획기적으로 모수의 수를 줄일 수 있으나 p가 매우 크고 또한 성분 즉 집단의 수 g 역시 작지 않을 경우 여전히 모수추정에 있어 어려움이 발생할 수 있다. Baek and McLachlan (2008)은 MFA보다 더 모수의 수를 줄일 수 있는 공통요인분석자혼합모형 (Mixtures of Common Factor Analyzers: MCFA)을 제안하였다. MCFA는 각 성분별로 공통의 요인적재 행렬을 가정하며 요인벡터는 서로 다른 평균과 공분산 행렬을 가진 요인모형을 가정한다. 즉,

$$\boldsymbol{Y}_j = \boldsymbol{A}\boldsymbol{U}_{ij} + \boldsymbol{e}_{ij}, \quad i = 1,...,g, \quad j = 1,...,n \tag{2.3}$$

이다. 이 때 $\boldsymbol{U}_{ij}$는 평균이 $\boldsymbol{\xi}_i$이고 공분산행렬이 $\boldsymbol{\Omega}_i$인 다변량정규분포 $N_q(\boldsymbol{\xi}_i, \boldsymbol{\Omega}_q)$를 따른다. $\boldsymbol{A}$는 $p \times q$차원의 공통요인적재행렬이다. 오차벡터 $\boldsymbol{e}_{ij}$는 $\boldsymbol{U}_{ij}$와 독립이며 대각행렬 $\boldsymbol{D}$를 공분산행렬로 가지는 $N_p(\boldsymbol{0}, \boldsymbol{D})$를 따른다. MCFA에 따르면 성분공분산 행렬은 $\boldsymbol{\Sigma}_i = \boldsymbol{A}\boldsymbol{\Omega}_i\boldsymbol{A}' + \boldsymbol{D}$로 표현된다. MCFA의 모수 수는 $d_2 = (g-1) + p + q(p+g) + (g-1)q(q+1)/2$이며 MFA의 그것보다 적다. 예를 들어 $p = 1000$, $q = 2$인 경우 $g = 2$에서 $g = 4$로 증가한 경우 MFA의 전체 모수 수는 $d_1 = 7999$에서 $d_1 = 15999$로 약 두 배 증가하지만 MCFA에서는 $d_2 = 3008$ 개에서 $d_2 = 3020$으로 거의 비슷하다.

본 연구에서는 고차원 자료의 판별이 목적이므로 훈련자료를 이용하여 MCFA의 모수들에 대한 최우추정량 $\hat{\boldsymbol{\Psi}} = \{\hat{\pi}_i, \hat{\boldsymbol{\xi}}_i, \hat{\mathrm{A}}, \hat{\boldsymbol{\Omega}}_i, \hat{\mathrm{D}}\}$을 EM 알고리즘에 의해 추정하고, 그렇게 추정된 모수들을 이용하여 차원 축소된 저차원의 요인점수를 계산하여 새로운 특징변수로 판별에 사용하는 것을 제안한다. 식 (2.2)의 MFA에서는 모든 집단의 요인($\boldsymbol{V}_{ij}$)들이 동일한 평균 $\boldsymbol{0}$을 갖는다고 가정되었으므로 요인점수들은 각 집단별로 구분할 수 없다. 반면 식 (2.3)의 MCFA에서는 요인($\boldsymbol{U}_{ij}$)들은 각 집단별로 각기 다른 평균($\boldsymbol{\xi}_i$)들을 갖기 때문에 집단별로 구분할 수 있어 새로운 판별 특징벡터 변수로 활용할 수 있다. 만약 관측값 $\boldsymbol{y}_j$가 i번째 집단에 속해있다면 $E(\boldsymbol{U}_{ij} | \boldsymbol{y}_j) = \boldsymbol{\xi}_i + \boldsymbol{\gamma}_i'(\boldsymbol{y}_j - \boldsymbol{A}\boldsymbol{\xi}_i)$ 이므로, 관측값 $\boldsymbol{y}_j$의 요인점수 벡터의 추정량으로서 $\hat{\boldsymbol{U}}_{ij} = \hat{\boldsymbol{\xi}}_i + \hat{\boldsymbol{\gamma}}_i'(\boldsymbol{y}_j - \hat{\boldsymbol{A}}\hat{\boldsymbol{\xi}}_i)$를 사용한다 (이 때 $\boldsymbol{\gamma}_i = (\boldsymbol{A}\boldsymbol{\Omega}_i\boldsymbol{A}' + \boldsymbol{D})^{-1}\boldsymbol{A}\boldsymbol{\Omega}_i$ 이며 최우추정량 $\hat{\boldsymbol{\Psi}}$ 에 대한 유도는 Baek and McLachlan (2008) 참조).

3. 고차원 자료의 판별분석

고차원 자료의 차원 축소된 새로운 특징변수로서 PCA와 CV 그리고 공통요인 분석자 혼합모형의 요인점수 등의 판별 성능을 비교하기 위하여 우리는 세 가지 판별분석 방법을 이용하였다. 즉, 모수적 판별분석방법인 선형판별분석 (Linear Discriminant Analysis: LDA), 이차판별분석 (Quadratic Discriminant Analysis: QDA)와 비모수적 판별분석인 국소선형 로지스틱 판별분석(Baek and Son (2006))을 이용하여 두 가지 유전자 발현자료에 대한 종양을 판별하고, 마지막으로 베트남 자료라고 불리는 고차원의 화학성분 자료를 판별함으로써 차원축소 방법들의 성능을 비교하였다. 각 차원축소 방법의 성능은 세 가지 실험자료 모두에 대하여 판별분석방법을 시행했을 때 교차타당성 (leave-one-out cross validation) 방법에 의한 분류결과에 의하여 판단하였다.

3.1 ALL-AML 백혈병 유전자발현 자료

첫 번째 분석 자료는 Golub *et al.* (1999)의 두 가지 ($g=2$) 형태의 백혈병 (ALL: acute lymphoblastic leukemia, AML: acute myeloid leukemia) 유전자 발현자료이다. 이 자료는 72 개의 세포조직 표본으로부터 추출한 7129개의 인간 유전자발현 측정값으로 이루어져있다. 우리는 Dudoit *et al.* (2002)에서 적용한 방법과 동일하게 사전처리를 시행한 후 Nguyen and Rocke (2002)에서와 같이 t-통계량에 의해서 유전자들을 두 집단 간 차이가 많이 나는 순서대로 순위화하여 $p=100$ 개의 유전자들을 선택하였다.

우리는 이렇게 선택된 100개의 유전자 발현값들에 대하여 PCA, MCFA 요인점수들을 계산하고 이들을 새로운 특징변수로 LDA, QDA, 국소선형 로지스틱 판별기에 입력하였다. 이 자료의 경우 차원 $p=100$에 비하여 표본크기가 $n=72$ 로서 더 작아 집단내제곱합과 교차곱행렬 ($\boldsymbol{W}$)이 비정칙 행렬이 되어 CV를 계산할 수가 없다. 따라서 이 자료의 경우 PCA와 MCFA 요인점수의 성능만 비교하였다.

<표 1>은 총 72개의 세포조직에 대하여 각각 $q=1,2$ 차원의 PCA와 MCFA 요인점수를 사용하여 각 판별방법을 시행했을 때 교차타당성 (leave-one-out cross validation) 방법에 의한 정분류 결과를 나타내고 있다. 세 가지 판별분석방법 모두

<표 1> 차원축소 방법별 ALL-AML 백혈병 유전자 자료의 정분류율

차원 축소(q)	차원 축소방법	LDA	QDA	국소선형 로지스틱
1	PCA	0.9722	0.9722	0.9722
	MCFA	0.9861	0.9861	0.9722
2	PCA	0.9583	0.9722	0.9583
	MCFA	0.9722	0.9722	0.9722

MCFA 요인점수의 정분류율이 PCA의 그것보다 더 높거나 동일한 결과를 나타낸다. 참고로 이 자료의 경우 PCA와 MCFA 요인점수 모두 $q=1$ 차원의 차원축소된 특징변수가 $q=2$ 차원의 차원축소 특징변수보다 분류 성능이 더 우수하거나 동일하다.

3.2 소아 백혈병 유전자발현 자료

두 번째 분석 자료는 Yeoh *et al.* (2002)의 $g=7$ 가지 형태의 소아 백혈병 유전자 발현자료이다. Yeoh *et al.* (2002)에서는 X^2과 t-통계량 등에 의하여 선택된 유전자 자료들을 이용하여 $n=327$ 개의 세포조직 표본에 대하여 위계적 군집분석을 시행하였다. 우리는 7 가지 각각의 소아 백혈병 집단마다 가장 큰 X^2 통계량값을 갖는 20 개의 유전자들을 추출하였다. 그 중에는 각 소아 백혈병 집단에서 중복되게 선택된 유전자들이 있으므로 최종적으로 $p=132$ 개의 서로 다른 유전자들이 선택되었다. 집단의 수는 $g=7$ 로서 작지 않다.

우리는 이렇게 선택된 $p=132$ 차원의 유전자 발현값에 대하여 PCA, CV, MCFA 요인점수들을 계산하고 이들을 새로운 특징변수로 LDA, QDA, 국소선형 로지스틱 판별기에 입력하였다. <표 2>은 총 $n=327$ 개의 세포조직에 대하여 각각 $q=1,2,3,4$ 차원의 PCA, CV, MCFA 요인점수를 사용하여 각 판별방법을 시행했을 때 교차타당성 방법에 의한 정분류 결과를 나타내고 있다. $q=1$ 차원에서는 세 가지 분류방법 모두에서 PCA가 다른 두 가지 차원축소 방법보다 분류능력이 우수하나 다른 차원에 비하여 가장 낮은 정분류율을 나타낸다.

<표 2> 차원축소 방법별 소아백혈병 유전자 자료의 정분류율

차원 축소(q)	차원 축소방법	LDA	QDA	국소선형 로지스틱
1	PCA	0.64832	0.63914	0.68502
	CV	0.40979	0.40367	0.42813
	MCFA	0.54434	0.60856	0.59021
2	PCA	0.77982	0.70031	0.78899
	CV	0.63914	0.6422	0.69113
	MCFA	0.69113	0.84098	0.79205
3	PCA	0.81346	0.80122	0.78593
	CV	0.78899	0.8104	0.77676
	MCFA	0.86544	0.94801	0.90214
4	PCA	0.85015	0.8318	0.78899
	CV	0.93578	0.92661	0.8685
	MCFA	0.89602	0.98165	0.92966

QDA와 국소선형 로지스틱 판별방법의 경우 $q=2,3,4$ 차원에서 MCFA 요인점수가 PCA나 CV에 비하여 향상된 성능을 보여주고 있다. PCA는 $q=4$ 차원에서 LDA 방법을 이용할 때 가장 높은 정분류율 0.85015을 달성했으며, CV 역시 $q=4$ 차원에서 LDA 방법을 이용할 때 정분류율이 0.93578 로서 가장 높았다. 그러나 PCA와 CV의 최고 정분류율은 $q=3$ 과 $q=4$ 의 MCFA의 요인점수를 QDA에 각각 적용했을 때의 정분류율 0.94801 과 0.98165 보다 더 낮다.

3.3 베트남 화학성분 농도 자료

세 번째 분석 자료는 Smyth *et al.* (2006)에서 분석한 소위 베트남 자료이다. 이 자료는 $n=224$ 명의 베트남인들이 $g=6$ 가지 집단으로 나누어져있으며 각 사람마다 머리카락 속에 들어있는 17가지 화학성분의 농도와 고차원 자료의 군집분석을 위하여 인공적으로 추가된 네 가지 형태의 잡음 변수들로 이루어져있다. 본 연구에서는 $p=67$ 개의 변수 (17 개 화학성분 농도 변수와 인공적인 50 개 정규 잡음 변수)로 이루어진 자료를 이용하였다. 이 자료 역시 $g=6$ 로서 집단의 수가 작지 않다.

<표 3>은 총 $n=224$ 명의 베트남인들에 대하여 각각 $q=1,2,3$ 차원의 PCA, CV, MCFA 요인점수를 사용하여 각 판별방법을 시행했을 때 교차타당성 방법에 의한 정분류 결과를 나타내고 있다. $q=1$차원의 LDA를 제외하고는 세 가지 분류방법 모두에서 MCFA 요인점수의 분류성능이 가장 우수하고 그 다음이 CV, 그리고 PCA가 가장 낮았다.

<표 3> 차원축소 방법별 베트남 화학성분 농도 자료의 정분류율

차원 축소(q)	차원 축소방법	LDA	QDA	국소선형 로지스틱
1	PCA	0.4375	0.4911	0.5268
	CV	0.6473	0.6205	0.6786
	MCFA	0.6250	0.7009	0.6830
2	PCA	0.8036	0.8036	0.7768
	CV	0.8839	0.8973	0.8795
	MCFA	0.9375	0.9955	0.9955
3	PCA	0.8661	0.8482	0.7813
	CV	0.9821	0.9777	0.9152
	MCFA	0.9911	1.0000	0.9866

4. 결론 및 토의

유전자 발현자료나 이미지 자료 등 고차원 자료는 종종 특징변수의 수가 표본 크기

보다 훨씬 더 많은 경우가 대부분이다. 판별분석 방법을 이용하여 이러한 고차원 자료를 분류하고자 할 때 필연적으로 차원의 저주가 발생하여 이를 극복하여야 한다. 본 연구에서는 PCA, CV, 공통요인 분석자 혼합모형의 요인점수 등 차원축소 방법을 고려하였으며, 그것들을 새로운 특징변수로서 모수적 혹은 비모수적 판별분석 방법에 사용하였을 때 어느 것이 가장 우수한 성능을 가지고 있는지 알아보고자 하였다. 두 가지 실제 유전자 발현 자료와 잡음이 추가된 한 가지 고차원 화학 성분 자료에 적용한 결과 공통요인 분석자 혼합모형의 요인점수를 차원 축소된 특징변수로 사용한 경우 대부분의 차원에서 PCA나 CV보다 향상된 분류 능력을 나타내고 있음을 확인하였다.

PCA는 성분점수를 추출할 때 집단표지(class label)에 대한 정보를 사용하지 않으며, CV 역시 집단 내 변동에 비해 집단 간 변동이 크도록 하는 선형 결합을 유도함으로써 각 집단의 분포 구조를 활용한 판별 정보를 추출하는데 미흡하다. 이 들에 비해 공통요인 분석자 혼합모형의 요인점수는 각 집단 별 고차원변수 분포를 서로 다른 평균과 공분산 구조를 갖는 요인분석 모형으로 적합시킴으로써 집단 별로 서로 상이한 분포 구조를 갖도록 추출된 저차원의 특징변수로서 분류를 위하여 보다 효과적이다.

참고문헌

1. Baek, J. and McLachlan, G. J. (2008). Mixtures of factor analyzers with common factor loadings for the clustering and visualisation of high-dimensional data. *Technical Report NI08018-SCH*, Preprint Series of the Isaac Newton Institute for Mathematical Sciences, Cambridge.
2. Baek, J. and Son, Y. S. (2006). Local linear logistic discriminant analysis with partial least square components, *Lecture Notes in Artificial Intelligence*, 4093, 574-581.
3. Banfield, J. D. and Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering, *Biometrics*, 49, 803-821.
4. Bicciato, S., Luchini, A. and Di Bello, C. (2003). PCA disjoint models for multiclass cancer analysis using gene expression data, *Bioinformatics*, 19, 571-578.
5. Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977), Maximum likelihood from incomplete data via the EM algorithm (with discussion), *Journal of the Royal Statistical Society*: Series B, 39, 1-38.
6. Dudoit, S. et al. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data, *Journal of the American Statistical Association*, 97, 77-87.
7. Fraley, C. and Raftery, A. E. (2004). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97, 611-631.
8. Ghahramani, Z. and Hinton, G. E. (1996). The EM algorithm for mixture

of factor analyzers, *Technical Report CRG-TR-96-1*, 8, University of Toronto, Canada.
9. 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.
10. McLachlan, G. J. (1992). *Discriminant Analysis and Statistical Pattern Recognition,* Wiley, New York.
11. McLachlan, G. J. and Peel D. (2000). *Finite Mixture Models, Wiley*, New York.
12. Nguyen, D., Rocke, D. (2002). Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, 18, 39-50.
13. Smyth, C., Coomans, D., and Everingham, Y. (2006). Clustering noisy data in a reduced dimension space via multivariate regression trees, *Pattern Recognition,* 39, 424-431.
14. Yeoh, E. and Ross, M. E., et al, (2002). Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling, *Cancer Cell*, 1, 133-143.

[접수일(2008년 7월 19일), 수정일(2008년 8월 6일), 게재확정일(2008년 8월 8일)]