The Korean Communications
in Statistics Vol. 9, No. 2, 2002
pp. 543-553

Results of Discriminant Analysis with Respect to Cluster Analyses Under Dimensional Reduction

Seong-San Chae[1)]

Abstract

Principal component analysis is applied to reduce p-dimensions into q-dimensions ($q \leq p$). Any partition of a collection of data points with p and q variables generated by the application of six hierarchical clustering methods is re-classified by discriminant analysis. From the application of discriminant analysis through each hierarchical clustering method, correct classification ratios are obtained. The results illustrate which method is more reasonable in exploratory data analysis.

Keywords : Clustering Method, Principal Component Analysis, Discriminant Analysis

1. 서론

p-차원 ($p \geq 2$)의 다변량자료에서, N개의 개체를 성질이 다른 것은 이질적인 집단으로, 성질이 같은 것은 유사한 집단으로 분류하는 구조적 단순화는 통계적 자료분석의 과정에 선행되어야 할 과제이다. 자료분석의 대상이 되는 변수의 수가 셋 이하 ($p \leq 3$)인 경우 관찰값들을 p-차원으로 표현하여 분류된 결과를 관찰하는 것도 바람직하나, 변수의 수가 늘어남에 따라 상당한 제약을 받는 것도 사실이다. 이때, 관련이 있는 변수들을 q-차원 ($q < p$)으로 축소하여 생성된 새로운 자료에 통계적인 기법을 적용하면 어떻게 될까? 물론, 관련이 있는 변수들에 대하여 의미있는 정보를 가진다는 것은 다변량자료에 대한 구조적 단순화 내지 요약이라는 측면에서 상당히 중요한 의미를 갖게된다.

다변량 정규분포의 가정 하에서 김혜중(1992)은 공분산행렬이 서로 다른 경우 변수선택문제를 논의한바 있으며, Silverman(1986)은 판별분석법의 문제를 다변량으로 확장하였을 경우 차원문제를 고려하여야 하고 실제적인 문제에 적용 시 장애가 된다고 하였으며, 김기영 · 전명식(1990)은 다른 변수의 선형결합으로 표현될 수 있는 변수는 그 변수만의 독립적인 판별정보를 제공하지 못할 뿐만 아니라 분석에서 필요한 행렬 조작 시 문제를 발생시킨다고 언급하였다.

주성분점수에 대한 이용은 김기영 · 전명식(1989)을 참고하기 바라며, 최승배 · 강창환(2001)은 주

1) Associate Professor, Department of Information and Statistics, Daejeon University, Daejeon, 300-716, Korea
E-mail : chae@dju.ac.kr

성분점수를 감도분석에, 강창환 · 김대학(2000)은 주성분분석과 군집분석의 결과를 차기분석에 이용하였다.

본 연구에서는 관련이 있는 변수들을 p-차원보다 적은 q-차원 $(q \leq p)$으로 축소할 때 주성분분석을 사용하였으며, 차원축소의 결과로 얻어지는 주성분점수를 판별분석 및 군집분석에 대한 입력자료로 이용하였다. 부분집단이 사전에 규정되어 있지 않은 경우 개체들간의 유사성에 근거하는 탐색적 통계적방법인 군집분석(Everitt : 1974 ; Hartigan : 1975)을 p-차원 자료와 축소된 q-차원 자료에 적용하였다. 군집분석의 결과로서 생성된 집단의 정보에 대하여 다시 판별분석을 실시하여 적정분류율을 계산하였으며, 이를 통하여 차원 축소된 자료에 대한 판별분석과 군집방법의 역할에 대하여 비교 연구하였다.

2. 판별분석법과 군집방법

본 연구의 목적을 위하여 g개의 부모집단에 대하여 추출한 $N = \sum_{h=1}^{g} n_h$개의 p-차원 연습표본을 X라 하면, 모든 개체에 대해서 크기 N인 표본의 $(N \times p)$ 자료행렬, X, 는

$$X_{(N\times p)} = \begin{bmatrix} X_1 \\ X_2 \\ \cdots \\ \cdots \\ \cdots \\ X_g \end{bmatrix} = \begin{bmatrix} x_{111} & x_{112} & \cdots & x_{11p} \\ x_{121} & x_{122} & \cdots & x_{12p} \\ \cdots & \cdots & \cdots & \cdots \\ x_{1n_11} & x_{1n_12} & \cdots & x_{1n_1p} \\ x_{211} & x_{212} & \cdots & x_{21p} \\ \cdots & \cdots & \cdots & \cdots \\ x_{2n_21} & x_{2n_22} & \cdots & x_{2n_2p} \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ x_{g11} & x_{g12} & \cdots & x_{g1p} \\ \cdots & \cdots & \cdots & \cdots \\ x_{gn_g1} & x_{gn_g2} & \cdots & x_{gn_gp} \end{bmatrix}$$

이다. 이때, x_{hij}를 h-번째 집단에 속하는 i-번째 개체의 j-번째 변수에 대한 측정값이며, $\overline{X}$는 p개 변수의 전체표본평균벡터, $\overline{x_h}$와 S_h는 h-집단에 대한 $p*1$ 표본평균벡터와 표본-공분산행렬 이라 정의하면, $S_p = [\sum_{h=1}^{g}(n_h - 1)S_h]/[\sum_{h}^{g}(n_h - 1)]$ 은 합동 표본-공분산행렬이다.

사전확률은 각 부모집단에 속하는 개체들의 수에 따라 결정되도록 하고, 각 부모집단에서 h 와 k 에 소속하는 개체를 서로 다른 집단으로 오분류하였을 때 소요되는 비용이 같다는 가정 하에서, 판별분석방법의 적용을 요약하면 다음과 같다.

분류하여야 할 집단이 2 개(g = 2)인 경우에, 선형판별함수를 이용한 판별분석은,

$$(\overline{x_1} - \overline{x_2})' S_p^{-1} x_0 > \frac{1}{2}(\overline{x_1} - \overline{x_2})' S_p^{-1}(\overline{x_1} + \overline{x_2})$$

이면 개체 x_0 를 집단 1 로 분류하고, 그렇지 않은 경우에는 개체 x_0 를 집단 2 로 분류한다. 본 연구에서는 집단의 공분산행렬을 동일하게 설정하였으며, 선형판별함수를 적용하여 적정분류율을 계산하였다.

군집방법은 Lance와 Williams(1966, 1967)의 공식으로부터 DuBien과 Warde(1987)가 두 가지 제한 조건하에서 유도한 공식,

$$d_{(ij)k}=\frac{1-\beta-2\pi}{2}d_{ik}+\frac{1-\beta+2\pi}{2}d_{jk}+\beta d_{ij}$$

에서 두 개의 모수, $(\beta,\ \pi)$, 로 정의되는 계통적 군집방법을 고려하였다.

위에서, d_{ij}, d_{ik}, d_{jk} 는 군집 $i,\ j,\ k$ 들 중 임의의 두 군집간의 거리를, $d_{(ij)k}$는 군집 i 와 j 를 결합하여 형성되는 새로운 군집과 군집 k 간의 거리를 나타내며, $d_{ij} < d_{ik} < d_{jk}$ 이고, 이들은 거리개념을 만족한다. 우선적으로 선택된 계통적 군집방법은 다음과 같으며, 채와 Warde(1991)의 연구에서 존재하는 군집수의 예측에 적합하다고 평가된 군집방법들이다.

(1) $\beta=\ 0.0$ 일 때, $\pi=\ 0.0,\ \ 0.5$;
(2) $\beta=-0.25$ 일 때, $\pi=-0.25,\ \ 0.0$;
(3) $\beta=-0.5$ 일 때, $\pi=-0.0,\ \ 0.25$;

위의 방법 중에서, (.0, .0)은 평균연결법, (.0, .5)는 최장연결법, 그리고 (-.25, .0)과 (-.5, .0)은 유동연결법이다. 최단거리법(0.0, -0.5)도 고려의 대상이었으나, 변수들 간의 상관관계가 있는 경우(폭이 좁고 긴 사슬형태의 군집형태)의 자료에 적합하다고 알려진 최단거리법의 특성 때문에 모의실험을 통한 연구에서 다음과 같은 문제점, 즉 군집내 개체의 수가 변수의 수보다 적은 경우가 발생하여, 군집방법의 적용으로 생성되는 군집들의 공분산행렬이 정칙이 아닌 경우도 발생하였다. 따라서 여러 가지 계통적 군집방법과의 동시 적용을 고려한 본 연구에서는 최단거리법을 제외한 6개의 군집방법만을 사용하였다.

3. 모의실험

연습표본 X를 생성하기 위하여 IMSL의 부프로그램을 이용하였고, 연습표본에 대하여 설정된 모수는 $N=90,\ p=9,\ g=3$ 이며, N은 X에 있는 개체의 수, p는 변수의 수, g는 부모집단의 수이며, 다변량 정규분포를 따르는 자료구조는 다음과 같다.

$$\underline{X}_{hi}\ \sim\ N_p(\underline{\mu}_h,\ \Sigma_h),$$

$i=1,2,..,n_h,\ h=1,2,3$, 각 부모집단에는 $30-30-30(n_h: n_1-n_2-n_3)$개의 개체로 구성하고,

각 부모집단 평균은 $\underline{\mu_1}' = (.0\ .0\ \delta\ .0\ .0\ \delta\ .0\ .0\ .0)'$

$\underline{\mu_2}' = (.0\ .0\ \delta\ .0\ .0\ .0\ .0\ .0\ \delta)'$

$\underline{\mu_3}' = (.0\ .0\ .0\ .0\ .0\ \delta\ .0\ .0\ \delta)'$

공분산행렬은

$$\Sigma_h = \Sigma = \begin{pmatrix} A & B & B \\ B & A & B \\ B & B & A \end{pmatrix}, \quad A_{3*3} = \begin{pmatrix} 1.0 & \rho_1 & \rho_1 \\ \rho_1 & 1.0 & \rho_1 \\ \rho_1 & \rho_1 & 1.0 \end{pmatrix}, \quad B_{3*3} = \begin{pmatrix} \rho_2 & \rho_2 & \rho_2 \\ \rho_2 & \rho_2 & \rho_2 \\ \rho_2 & \rho_2 & \rho_2 \end{pmatrix}$$

$\rho_1 = 0.0, 0.4, 0.8,\ \rho_2 = 0.0, 0.2, 0.4$ 이다.

이때, 변수군집의 관점에서 $\rho_1 \geq \rho_2$를 제한하여 설정하였다. 각 부모집단의 평균간 거리는, $\delta_g = \sqrt{4 \times \delta^2} = 2.0,\ 4.0$ 으로 서로 동등함을 유지하도록 하였으며, 이러한 값을 얻도록 $\delta = 1.0,\ 2.0$ 를 정의하였다. δ의 설정 시, 부모집단의 평균간 거리가 큰 경우는 부모집단이 잘 구분되기 때문에 군집분석을 적용한다는 것이 큰 의미를 갖지 못하고, 부모집단의 평균간 거리가 작은 경우에는 판별분석을 실시하지 않는 것이 일반적이라는 것을 고려하였으며, 다음과 같은 단계를 통하여 모의실험을 진행하였다.

일차적으로 $g=3$개의 부모집단을 같는 연습표본 X에 대하여, 선형판별함수와 군집분석을 적용하여 각 방법들이 생성된 부모집단에 속하는 개체들을 얼마나 잘 분류하고 있는지 살펴보았다. 판별분석의 적용시, 3개의 부모집단을 갖도록 생성된 연습표본 X에 대하여 부모집단 h 와 k 에 소속하는 개체를 서로 다른 집단으로 오분류하였을 때 소요되는 비용은 같다고 가정하고, 사전확률은 각 부모집단에 속하는 개체들의 수에 따라 결정되도록 하였다. 또한, 군집방법의 적용에 의하여 형성되는 각 선행군집에서 각 군집의 공분산행렬은 정칙행렬인 경우로 제한하였다.

이차적으로 부모집단의 구분이 명확하지 않다는 가정 하에 연습표본 X에 군집분석을 적용하고, 이의 결과로 생성된 $g=3$개의 군집을 갖는 복원된 연습표본 X'에 대하여 선형판별함수를 적용하였다.

삼차적으로 p-차원으로 구성되는 연습표본 X의 구조를 파악하기 위하여 주성분분석을 적용하고, 관련이 있는 변수들을 p-차원 보다 적은 q-차원 $(q \leq p)$으로 축소하였다. 차원축소의 과정에서 축약의 적정성에 대한 기준은, 설정된 공분산행렬의 구조 및 표 1에 제시된 주성분의 누적기여율 평균이 0.8에 근접하게 되는 경우에 따라서 각 경우에 대응하는 주성분점수로 이루어진 새로운 축소된 연습표본 Y를 생성하였다. 예를 들면, $\delta_g = 2.0,\ \rho_1 = .0, \rho_2 = .0$ 일 경우 주성분의 수를 $q=6$, $\delta_g = 2.0,\ \rho_1 = .4,\ \rho_2 = .0, .2$ 일 경우는 주성분의 수를 $q=5$, $\delta_g = 2.0, 4.0,\ \rho_1 = .8$ 일 경우에는 주성분의 수를 $q=3$으로 변화하여 연구를 진행하였다.

다음단계로 q-차원 $(q \leq p)$으로 생성된, 축소된 Y에 대해 판별함수와 군집분석을 적용하여

각 방법들이 생성된 부모집단에 속하는 개체들을 얼마나 잘 분류하고 있는지 살펴보았다.

표 1. 주성분의 수(q)에 따른 누적기여율 평균

δ_g	ρ_1	0.0	0.4		0.8		
	q / ρ_2	0.0	0.0	0.2	0.0	0.2	0.4
2.0	1	.2003	.2498	.3050	.3362	.3872	.4919
	2	.3632	.4529	.4890	.5994	.6304	.6804
	3	.4963	.6110	.6298	.8133	.8174	.8256
	4	.6123	.7121	.7245	.8772	.8803	.8843
	5	.7136	.7970	.8035	.9287	.9293	.9296
	6	.8025	.8630	.8652	.9521	.9522	.9526
	7	.8784	.9176	.9194	.9714	.9714	.9715
	8	.9450	.9636	.9648	.9873	.9873	.9875
4.0	1	.2895	.3021	.2838	.3389	.3184	.3662
	2	.5313	.5530	.5277	.6096	.5899	.6229
	3	.6288	.6833	.7149	.7816	.8044	.8340
	4	.7151	.7761	.7916	.8750	.8857	.8987
	5	.7900	.8500	.8538	.9483	.9485	.9492
	6	.8549	.8994	.9002	.9653	.9655	.9660
	7	.9111	.9390	.9404	.9792	.9794	.9795
	8	.9598	.9732	.9733	.9909	.9910	.9910

마지막으로, 부모집단의 구분이 명확하지 않다는 가정 하에 Y에 군집분석을 적용하고, 생성된 3개의 군집을 갖는 축소-복원된 $Y^{'}$에 대하여 판별함수를 적용하였다.

이러한 과정을 각 모수 설정에 따라 형성된 자료를 이용하여 100번의 결과가 얻어질 때까지 반복수행하여 적정분류율의 평균을 계산하였으며, 다음과 같이 그 결과들을 비교 검토하였다.

4. 분석 및 결과

전체 개체 중에서 원래의 집단(X)으로 적합하게 분류된 개체의 비율 및 6개의 군집방법의 적용에 의하여 생성된 선행집단($X^{'}$)으로 적합하게 분류된 개체의 비율을 적정분류율이라 하고, 다음과 같이 정의하였다. 즉,

$$\frac{\sum_{h=1}^{g} n_{hh}}{N},$$

n_{hh}는 h 집단에 속하는 개체들이 h 집단으로 분류된 개체들의 수이다. 이때, 설정된 모수 $(\rho_1, \rho_2, \delta_g)$에 대하여 각각 100번의 결과가 얻어질 때까지 반복이 이루어졌으며, 이들의 평균을

구하여 적정분류율의 평균을 계산하였다.

여기서 적정분류율의 평균은 이미 집단이 잘 분류되어 있다는 가정 하에 적용된 판별분석의 적정분류성과, 모집단에 대한 정보가 알려져 있지 않다고 가정하고 6개의 군집방법을 통하여 분류한 선행집단에 판별분석을 적용한 결과의 적정분류성에 대한 정보를 제공한다. 이러한 정보에 근거하여, 차원축소된 자료에 대한 판별분석의 적용 및 (β, π) 평면상에 정의된 6개 군집분석방법의 활용성을 비교 검토하였다.

우선적으로, $g=3$개의 부모집단을 같는 연습표본 X에 대하여 선형판별함수와 6개의 군집분석방법을 적용하고, 각 방법들이 생성된 부모집단에 속하는 개체들을 얼마나 잘 분류하고 있는지를 표 2에 주어진 적정분류율의 평균을 통하여 살펴보았으며, 판별분석이 적용된 6개의 군집분석보다 생성된 부모집단에 속하는 개체들을 잘 분류하고 있음을 알 수 있다. 여기서 원자료에 대한 군집분석방법들의 적정분류율이 판별분석의 적정분류율 보다 낮은 것을 알 수 있다.

이차적으로 부모집단의 구분이 명확하지 않다는 가정 하에 연습표본 X에 군집분석을 적용하고, 이의 결과로 생성된 $g=3$개의 군집을 갖는 복원된 연습표본 X'에 대하여 선형판별함수를 적용하여 적정분류율을 계산하여 그 결과를 표 3에 정리하였다.

표 3에서 원자료에 적용한 판별분석의 적정분류율은 $\rho_2=0.0$일 때 $\rho_1=(.0, .4, .8)$의 변화에 따라 (.7888, .8237, .9583 : $\delta_g=2.0$), (.9726, .9872, .9999 : $\delta_g=4.0$)이며, 필연적인 것은 아니지만 집단간 거리와 상관의 정도가 큰 경우 판별함수의 판별력을 증가시킨다(Mardia, Kent, Bibby : 1979)고 할 수 있다. 그러나, $\delta_g=2.0$일 때, $\rho_{1g}=0.8$인 경우를 제외하고는 군집분석방법을 적용한 후 판별분석을 적용한 경우가 판별분석을 직접적으로 적용한 경우보다 적정분류율이 상당히 큰 것으로 나타났다. 따라서, 부모집단의 구분이 명확하지 않다면 군집분석을 통하여 부모집단을 구분하고, 판별분석을 적용하는 것이 타당할 것으로 판단된다.

여기서 표 2의 결과와 표 3의 결과를 다시 살펴보면, 표 2의 결과에서 군집분석의 적용에 의한 적정분류율이 판별분석의 적정분류율보다 작다고 하더라도, 표 3의 결과에서 6개의 군집분석방법을 적용하여 복원된 자료에 판별분석을 실시하면 적정분류율이 상당한 수준으로 증가하였다. $\delta_g=2.0$인 경우, 군집분석의 적용으로 복원된 자료에 대한 판별분석의 적정분류율이 증가되었다는 사실에서, 군집분석의 적용 결과로 생성되는 군집(즉, 부모집단)이 원자료에 존재하는 부모집단과는 다르지만 좋은 적정분류율을 나타내는 집단으로 구분되었음을 의미한다.

다음 단계로, p-차원으로 구성되는 연습표본 X의 구조를 파악하기 위하여 주성분분석을 적용하고, 관련이 있는 변수들을 주성분의 누적기여율 평균이 0.8에 근접하게 되는 모수설정의 각 경우에 따라서 p-차원을 q-차원 $(q \leq p)$으로 축소하였다. q개의 주성분점수로 이루어진 연습표본 Y를 생성하였으며, $p=9$로 생성된 자료에 대한 분석과 동일하게 q-차원 $(q \leq p)$으로 축소된 Y에 대해 선형판별함수와 군집분석을 적용하여 각 방법들이 생성된 부모집단에 속하는 개체들을 얼마나 잘 분류하고 있는지를 표 2와 대응하여 표 4에서 살펴보았다. 여기서, $\rho_1=\rho_2=0.0$인 경우 주성분분석을 통한 차원축소는 의미가 없게되어 분석의 유용성이 제한되고 있음을 밝히며, 차원축소의 문제에서 $\rho_1=\rho_2=0.0$에 대한 분석은 더 이상 고려하지 않았다.

표 2. 원자료에 대한 판별분석 및 군집분석의 적정분류율

δ_g	ρ_1	0.0	0.4		0.8		
	$(\beta,\ \pi)/\ \rho_2$	0.0	0.0	0.2	0.0	0.2	0.4
2.0	판별분석	.7888	.8237	.8341	.9583	.9578	.9589
	(0.0 , 0.0)	.7110	.6510	.6631	.6228	.6137	.6278
	(0.0 , 0.5)	.6239	.6051	.5970	.5770	.5896	.5800
	(-0.25,-0.25)	.6269	.5969	.5879	.5708	.5720	.5752
	(-0.25 ,0.0)	.6130	.5817	.5789	.5408	.5518	.5563
	(-0.5 , 0.0)	.6018	.5639	.5637	.5283	.5296	.5419
	(-0.5 ,-0.25)	.6018	.5558	.5571	.5262	.5249	.5271
4.0	판별분석	.9726	.9872	.9890	.9999	.9999	.9999
	(0.0 , 0.0)	.8887	.8670	.8478	.8067	.7948	.7621
	(0.0 , 0.5)	.8822	.8512	.8101	.7371	.7111	.6904
	(-0.25,-0.25)	.8909	.8930	.8851	.8833	.8717	.8273
	(-0.25 ,0.0)	.9082	.9039	.9077	.8822	.8831	.7997
	(-0.5 , 0.0)	.9070	.9123	.9184	.8932	.9054	.8542
	(-0.5 ,-0.25)	.9050	.9108	.9127	.8889	.8867	.8053

표 3. 원자료와 복원된 자료 ($p=9$)에 대한 판별분석의 적정분류율

δ_g	ρ_1	0.0	0.4		0.8		
	$(\beta,\ \pi)/\ \rho_2$	0.0	0.0	0.2	0.0	0.2	0.4
2.0	판별분석	.7888	.8237	.8431	.9583	.9578	.9589
	(0.0 , 0.0)	.9356	.9346	.9433	.9370	.9448	.9539
	(0.0 , 0.5)	.8911	.9116	.9224	.9139	.9333	.9390
	(-0.25,-0.25)	.9144	.9260	.9281	.9368	.9423	.9459
	(-0.25 ,0.0)	.9238	.9292	.9358	.9372	.9478	.9497
	(-0.5 , 0.0)	.9191	.9288	.9320	.9334	.9434	.9452
	(-0.5 ,-0.25)	.9119	.9192	.9273	.9356	.9336	.9404
4.0	판별분석	.9726	.9872	.9890	.9999	.9999	.9999
	(0.0 , 0.0)	.9666	.9638	.9624	.9507	.9531	.9582
	(0.0 , 0.5)	.9553	.9428	.9513	.9282	.9309	.9467
	(-0.25,-0.25)	.9644	.9659	.9632	.9566	.9606	.9659
	(-0.25 ,0.0)	.9700	.9633	.9696	.9519	.9589	.9643
	(-0.5 , 0.0)	.9727	.9676	.9694	.9504	.9599	.9614
	(-0.5 ,-0.25)	.9706	.9639	.9656	.9514	.9549	.9542

표 4에서 차원 축소된 자료에 직접적으로 적용한 판별분석의 적정분류율은 ρ_1이 고정되었을 때 ρ_2가 증가함에 따라 평균간 거리에 관계없이 증가하지만, 군집분석의 적정분류율은 평균간 거리 및 군집방법에 따라 증가 혹은 감소하는 경향을 보이고 있다.

표 4를 표 2와 대응하여 살펴보면, $\rho_1=\rho_2=0.0$인 경우를 제외하고는 평균간 거리에 관계없이 주성분분석을 통하여 축소된 자료에 대한 판별분석과 군집분석의 적정분류율은, 원자료에 대한 적정분류율 보다 작아지는 경향이 있으며, ρ_1이 증가하면 감소하는 경향을 보이고 있다.

마지막으로, 자료에 대한 집단 정보가 주어지지 않은 경우를 가정하고 Y에 군집분석을 적용하여 생성된 3개의 군집을 갖는 축소-복원된 Y^{*}에 대하여 선형판별함수를 적용하였고, 이에 대한 결과를 표 5에 요약 정리하였다.

표 5를 살펴보면, 평균간 거리가 $\delta_g=2.0$인 경우, ρ_1이 고정되었을 때 ρ_2가 증가하면 적정분류율이 증가하는 양상을 보이며, 축소-복원된 자료에 군집분석방법을 적용한 후 판별분석을 적용한 경우가 판별분석을 직접적으로 적용한 경우보다 적정분류율이 상당히 큰 것으로 나타났다. 이것은 표 2와 표 3의 결과에서 언급한 바와 같이 군집분석방법의 적정분류율이 판별분석의 적정분류율보다 작다고 하더라도, 군집분석의 적용 결과로 생성되는 부모집단이 원자료에 존재하는 부모집단과는 다르지만 부모집단의 구분이 보다 좋은 적정분류율을 나타내는 집단으로 구분되었음을 나타낸다.

평균간 거리가 $\delta_g=2.0$인 경우 표 3과 대응하여 표 5를 살펴보면, $\rho_1=0.8$, $\rho_2=0.2$인 경우를 제외하고는 주성분분석을 이용하여 차원축소된 자료에 군집방법을 적용하고 순차적으로 판별분석을 적용한 경우의 적정분류율이 원자료에 대한 적정분류율 보다 증가한 것으로 나타났으나, 그 차이는 별로 크지 않음을 알 수 있다.

표 2-표 5의 결과에서, 분석대상으로 주어진 다차원 자료에 대해 부모집단의 구분이 명확하지 않거나 부모집단에 대한 정보가 알려져 있지만 집단간 평균 거리가 작은 경우에는, 주성분분석에 의하여 p-차원을 q-차원으로 축소하고, 군집분석으로 복원된 자료에 대하여 판별분석을 실시하는 것도 하나의 통계적인 방법임을 제시하였다.

다차원 자료에 군집분석을 적용하여 판별분석을 적용한 경우와 주성분분석을 통하여 차원축소된 자료에 군집분석과 판별분석을 단계적으로 적용한 경우의 적정분류율에 약간의 차이가 있음을 알 수 있었으며, 주성분분석 및 군집분석이 차기분석의 하나인 판별분석에 대한 일련의 분석과정에서 선행단계의 역할을 할 수 있음을 보여주었다.

위의 결과들에 대한 해석에 있어서, 6개의 계통적 군집분석방법의 적용에 의하여 생성된 선행군집이 $g=3$으로 타당하게 분류되었음을 전제로 하였으며, 서로 다른 3가지 통계적 분석기법이 단계적으로 한 자료에 적용됨에 따라 그들이 적정분류율에 미치는 결과가 교락되어 있으므로, 표 4와 표 5에 제시된 적정분류율에 대한 평가 및 이용 시 주의가 요구된다.

표 4. 축소된 자료에 대한 판별분석과 군집분석의 적정분류율

	ρ_1	0.0	0.4		0.8		
δ_g	$(\beta,\ \pi)/\ \rho_2$	0.0	0.0	0.2	0.0	0.2	0.4
2.0	판별분석	.7624	.7862	.7977	.5501	.5878	.6564
	(0.0 , 0.0)	.6844	.6503	.6303	.5921	.6002	.5916
	(0.0 , 0.5)	.6471	.6036	.5783	.5509	.5632	.5622
	(-0.25,-0.25)	.6436	.5889	.5799	.5468	.5401	.5610
	(-0.25 ,0.0)	.6410	.5696	.5583	.5383	.5252	.5276
	(-0.5 , 0.0)	.6252	.5586	.5380	.5177	.5040	.5027
	(-0.5 ,-0.25)	.6301	.5580	.5313	.5187	.4972	.4997
	proportion	.8025	.7970	.8035	.8133	.8174	.8256
	주성분수(q)	6	5	5	3	3	3
4.0	판별분석	.9647	.9661	.9802	.8600	.9193	.9731
	(0.0 , 0.0)	.9059	.8494	.8568	.7447	.7422	.6923
	(0.0 , 0.5)	.9167	.8442	.8278	.7488	.7147	.6443
	(-0.25,-0.25)	.9123	.8662	.8800	.7506	.7327	.6862
	(-0.25 ,0.0)	.9197	.8809	.9061	.7574	.7470	.6877
	(-0.5 , 0.0)	.9222	.8874	.9157	.7710	.7679	.6964
	(-0.5 ,-0.25)	.9222	.8920	.9157	.7582	.7627	.6886
	proportion	.7900	.7761	.7916	.7816	.8044	.8340
	주성분수(q)	5	4	4	3	3	3

표 5. 축소된 자료와 축소-복원된 자료에 대한 판별분석의 적정분류율

	ρ_1	0.0	0.4		0.8		
δ_g	$(\beta,\ \pi)/\ \rho_2$	0.0	0.0	0.2	0.0	0.2	0.4
2.0	판별분석	.7624	.7862	.7977	.5501	.5878	.6564
	(0.0 , 0.0)	.9391	.9390	.9493	.9321	.9378	.9626
	(0.0 , 0.5)	.9121	.9203	.9380	.9144	.9308	.9563
	(-0.25,-0.25)	.9256	.9316	.9366	.9322	.9352	.9570
	(-0.25 ,0.0)	.9247	.9381	.9416	.9378	.9392	.9589
	(-0.5 , 0.0)	.9239	.9363	.9356	.9337	.9397	.9593
	(-0.5 ,-0.25)	.9246	.9278	.9428	.9234	.9366	.9567
	proportion	.8025	.7970	.8035	.8133	.8174	.8256
	주성분수(q)	6	5	5	3	3	3
4.0	판별분석	.9647	.9661	.9802	.8600	.9193	.9731
	(0.0 , 0.0)	.9738	.9532	.9587	.9366	.9409	.9507
	(0.0 , 0.5)	.9699	.9493	.9456	.9466	.9304	.9380
	(-0.25,-0.25)	.9752	.9567	.9596	.9424	.9423	.9519
	(-0.25 ,0.0)	.9758	.9581	.9648	.9504	.9463	.9501
	(-0.5 , 0.0)	.9772	.9617	.9639	.9459	.9399	.9491
	(-0.5 ,-0.25)	.9743	.9599	.9627	.9431	.9407	.9440
	proportion	.7900	.7761	.7916	.7816	.8044	.8340
	주성분수(q)	5	4	4	3	3	3

5. 결론

본 연구에서는 주성분분석을 통하여 p-차원 자료를 q-차원으로 축소한 자료를 이용하여 계통적 군집분석방법과 일반적 가정하의 판별분석에 대한 두 분석방법간의 상호보완적인 활용성 및 3가지 서로 다른 분석기법이 한 자료에 적용됨에 따라 나타나는 결과를 비교 검토하였다.

원자료에 대한 부모집단의 정보를 가지고 있는 p-차원 자료와 주성분분석을 적용하여 q-차원으로 축소된 자료에 적용한 군집분석의 적정분류율이 판별분석의 적정분류율 보다 낮은 수준이었다. 그러나, 군집분석의 적용 결과로 새로운 부모집단으로 구성되는 복원된 자료에 판별분석을 적용한 결과, 판별분석의 적정분류율이 증가되었음을 확인할 수 있었다. 이것은 p-차원 자료 및 q-차원으로 축소된 자료에 군집분석을 적용하였을 때, 그 적용 결과로 생성되는 부모집단이 원자료에 존재하는 부모집단과는 다르지만 부모집단의 구분이 보다 좋은 적정분류율을 나타내는 집단으로 구분되었음을 의미하였다.

본 연구에 있어서, 서로 다른 3가지 통계적 분석기법인 주성분분석, 군집분석, 그리고 판별분석이 단계적으로 한 자료에 적용됨에 따라 그들이 적정분류율에 미치는 결과가 교락되어 있음으로, 제시된 적정분류율에 대한 평가 및 이용 시 주의가 요구된다. 물론, 주성분분석과 군집분석을 선행적인 탐색적 분석방법으로 사용하여 차기의 통계적 분석방법을 적용할 경우에도 고려하여야할 사항으로 생각된다.

참고문헌

[1] 강창환, 김대학 (2000), 쌀 예상 생산량 추정방법에 대한 연구, 「응용통계연구」, 제 13권 2호, 329-341

[2] 김기영, 전명식 (1989), SAS 주성분분석, 자유아카데미

[3] 김기영, 전명식 (1990), SAS 판별 및 분류분석, 자유아카데미

[4] 채성산, 황정연 (1994), 집락분석과 판별분석의 활용성연구, 「품질경영학회지」, 제22권 2호, 143-153.

[5] 최승배 · 강창환 (2001), 주성분점수를 이용한 이변량 공간자료에 대한 감도분석, 「응용통계연구」, 제 14권 2호, 415-427.

[6] Chae, S. S. and Warde, W. D. (1991), A Method to Predict the Number of Clusters, *Journal of the Korean Statistical Society, Vol.* 20, 162-176.

[7] DuBien J. L. and Warde, W. D. (1987), A Comparison of Agglomerative Clustering Method with respect to Noise, *Communication of Statistics, Theory Method*, Vol. 16, 1433-1460.

[8] Everitt, B. S. (1974), *Cluster Analysis*, Wiley, New York.

[9] Hartigan, J. A. (1975), *Clustering Algorithms*. Wiley, New York.

[10] Lance, G. N. and Williams, W. T. (1966), A Generalized Sorting Strategy for Computer Classification, *Nature*, Vol. 212, 218.

[11] Lance, G. N. and Williams, W. T. (1967), A General Theory of Classificatory Sorting Strategies, 1. Hierarchical Systems. *The Computer Journal*, Vol. 9, 373-380.
[12] Madia, K. V., Kent, J. T. and Bibby, J. M. (1979), *Multivariate Analysis*, Academic Press.
[13] Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, New York.

[2002년 1월 접수, 2002년 5월 채택]