The Korean Communications in Statistics Vol. 15, No. 1, 2008, pp. 65–76

행렬도에서 군집분석의 활용†

최용석[1)], 김형영[2)]

요 약

행렬도 (biplot)는 이원표 자료행렬 (two-way data matrix)의 행과 열을 그래프에 동시에 나타내어 이들의 관계를 살피려는 다변량 그래프적 분석기법이다 (Gower와 Hand, 1996; 최용석, 2006, 1장). 그래프적 분석기법은 그 특성상 대용량 자료를 해석하는 데는 어려움이 따른다. 따라서, 자료를 효과적으로 줄일 수 있는 군집분석을 활용하여 원자료와 변수간의 행렬도가 아닌 각 군집과 변수간의 행렬도 분석을 수행함으로써, 기존의 행렬도에서 해석의 어려웠던 대용량 자료에 대한 해석이 가능하게 되며, 자료에 대한 정보를 쉽게 파악할 수 있는 장점을 가진다.

주요용어: 행렬도; K-평균 군집분석.

1. 서론

행렬도는 복잡한 다변량 분석의 결과를 보다 쉽게 파악할 수 있기 때문에 최근 여러 분야에서 행렬도에 대해서 활발한 연구와 응용을 하고 있다. 행렬도는 Gabriel (1971, 1981)에 의해서 주로 개발되었으며, Bradu와 Gabriel (1978) 그리고 Gabriel (1981)은 모형을 진단하는데 행렬도를 이용하기도 하였다. 국내에선 Choi (1991)가 행렬도를 처음으로 소개하였고, Biplot을 행렬도라 부른 것은 허명회 (1993, 5장)가 국내에선 처음이었으며, 통계상담의 기법으로 행렬도의 활용을 강조하였다. 하지만, 행렬도는 대용량 자료에서는 해석이 힘든 문제점을 가지고 있다. 이러한 한계를 극복하기 위한 방법으로 본 연구에서는 군집분석을 행렬도에 활용하는 방법을 제안하고자 한다. 이를 위해서 2장에서 행렬도를 소개하고 위계적 군집분석을 통해 적절한 군집의 수를 정하는 방법들을 살피려 한다. 다음으로, 대용량 자료의 군집화에 주로 이용되는 비위계적 군집분석인 K-평균 군집분석을 통해 군집들의 중심값을 얻고 이를 행렬도에 활용하는 사례를 보이고자 한다.

† 이 논문은 2007년도 부산대학교 기초과학연구원 기초과학연구기반 조성연구비 지원에 의하여 연구되었음.

1) (609-735) 부산광역시 금정구 장전 2동 산 30, 부산대학교 통계학과, 교수.
Correspondence: yschoi@pusan.ac.kr
2) (609-735) 부산광역시 금정구 장전 2동 산 30, 부산대학교 통계학과, 석사과정.

2. 행렬도에서 군집분석의 활용

2.1. 행렬도의 소개

행과 열의 수가 각각 n과 p인 이원표 자료행렬을

$$X = (x_{ij}) = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{12} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}, \quad i = 1, \ldots, n;\ j = 1, \ldots, p$$

라 하자. 각 열을 나타내는 변수들의 평균 $\bar{x}_{\cdot j} = \sum_{i=1}^{n} x_{ij}/n$을 뺀 새로운 자료행렬을 $Y = (x_{ij} - \bar{x}_{\cdot j})$라 하자. 이때 계수 (rank) r인 자료행렬 Y의 비정칙치분해 (sigular value decomposition)는 다음과 같다.

$$\begin{aligned} Y &= UD_\lambda V' \\ &= \sum_{k=1}^{r} \mathbf{u}_k \lambda_k \mathbf{v}_k' \\ &= \sum_{k=1}^{r} (\mathbf{u}_k \lambda_k^m)(\lambda_k^{1-m} \mathbf{v}_k'). \end{aligned} \tag{2.1}$$

여기서, 크기가 $n \times r$과 $p \times r$행렬 $U = (\mathbf{u}_1, \ldots, \mathbf{u}_r)$와 $V = (\mathbf{v}_1, \ldots, \mathbf{v}_r)$는 직교 열 벡터와을 갖고 있다. 그리고, 대각행렬 $D_\lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_r)$는 $\lambda_1 \geq \cdots \geq \lambda_r > 0$관계를 갖는 비정칙치를 대각원소로 하고 있으며, $0 \leq m \leq 1$이다.

$D_\lambda = D_\lambda^m D_\lambda^{1-m}$을 만족하도록 다음과 같이 정의하자.

$$D_\lambda^m = \mathrm{diag}(\lambda_1^m, \ldots, \lambda_r^m) \text{과} \quad D_\lambda^{1-m} = \mathrm{diag}(\lambda_1^{1-m}, \ldots, \lambda_r^{1-m}).$$

그리고, 크기가 $n \times r$과 $p \times r$행렬을 각각

$$A = (\mathbf{u}_1 \lambda_1^m, \ldots, \mathbf{u}_r \lambda_r^m) = UD_\lambda^m \tag{2.2}$$

이고,

$$B = (\mathbf{v}_1 \lambda_1^{1-m}, \ldots, \mathbf{v}_r \lambda_r^{1-m}) = VD_\lambda^{1-m} \tag{2.3}$$

라 하자. 이때, 식 (2.1)의 비정칙치분해는 다음과 같이 인자분해 (factorization)된다.

$$Y = AB'. \tag{2.4}$$

du Toit 등 (1986, p. 108)은 이를 비정칙치인자분해 (singular value factorization)라 하였다. 식 (2.4)의 이원표 자료행렬 Y의 계수가 2인 근사적 인자분해는 다음과 같다.

$$Y_{(2)} = A_{(2)}B_{(2)}{}'.$$

여기서, $A_{(2)}$와 $B_{(2)}$는 행 벡터로 $\mathbf{a}_i^{(2)} = (u_{i1}\lambda_1^m, u_{i2}\lambda_2^m)'$, $(i = 1, \ldots, n)$와 $\mathbf{b}_j^{(2)} = (v_{j1}\lambda_1^{1-m}, v_{j2}\lambda_2^{1-m})'$, $(j = 1, \ldots, p)$를 각각 갖는 크기가 $n \times 2$이고 $p \times 2$인 행렬이다. Gabriel (1971)은 두 행렬 $A_{(2)}$와 $B_{(2)}$를 행렬도를 위한 인자 (factor)라 하였다. 대개 $Y_{(s)}$에 의한 자료행렬 Y의 근사정도를 전통적으로 근사적합도 (goodness-of-fit of the approximation)

$$fit = (\lambda_1^2 + \lambda_2^2 + \cdots + \lambda_s^2) / \sum_{k=1}^{r} \lambda_k^2 \tag{2.5}$$

에 의해서 측정한다 (Gabriel, 1971; Jolliffe, 1986, p. 78; du Toit 등, 1986, pp. 107–127). 사실, 식 (2.5)의 근사적합도는 원자료에 대한 s차원 행렬도의 설명력이라 할 수 있다.

덧붙여, 식 (2.2)와 (2.3)에서 상수 $m(0 \le m \le 1)$에 따라 여러 가지 행렬도가 있을 수 있다. Gabriel (1971)은 $m = 0, 1/2$ 그리고, 1인 세 가지 전형적인 행렬도를 소개하고 있다. $m = 0$인 경우를 "주성분행렬도"라 하는데, 이는 가장 보편적인 행렬도로 주성분분석 (principal component analysis)의 성질을 가지고 있는 행렬도이다. 그리고, $m = 1/2$인 경우를 "대칭행렬도 (symmetric biplot)"라 부르고 있다. 특히, $m = 0$과 $m = 1$인 경우를 식 (2.4)의 인자분해 (factorization) 관점에서 "GH'행렬도"와 "JK'행렬도"라 부르고 있다.

이러한 기본적인 행렬도 뿐만 아니라 하나의 그림에 관측치와 변수에 관한 정보를 나타내는 관점에서 주성분분석, 인자분석, 대응분석, 다차원척도법 등도 행렬도의 범주에서 살펴볼 수 있다 (최용석, 2006).

대칭행렬도는 변수 사이에 높은 상관이 있는 경우, 매우 좁은 범위에 변수들이 위치하는 경향이 있으며, JK'행렬도는 변수에 관한 벡터가 행렬도 중심에 매우 가깝게 몰려있기 때문에 해석이 종종 어려울 때가 있다. 따라서, 본 연구에서는 가장 보편적이면서 해석하기가 쉬운 주성분행렬도를 이용한다.

2.2. 군집분석의 활용

이미 서론에서 언급했듯이, 행렬도가 자료의 정보를 시각적으로 표현하여 쉽게 정보를 파악할 수 있는 장점을 가지고 있지만, 대용량 자료에서는 해석이 힘든 문

제점을 가지고 있다. 따라서, 위계적 군집분석을 통해 군집의 수를 결정하고, 대용량 자료의 군집분석에 적합한 K-평균 군집분석을 수행하여 행렬도에 활용하는 것을 제안한다.

군집화란, 관찰치들을 유사한 특성을 지닌 몇 개의 소그룹으로 분할하는 작업을 말한다. 작업의 특성이 분류작업과 흡사하지만 분석하고자 하는 자료의 분류가 포함되지 않다는 점에서 차이가 있으며, 다른 분석을 위한 선행 작업으로서의 역할을 수행하는 경우가 많다.

즉, 군집분석은 주어진 관찰치 중에서 유사한 것들을 몇몇의 집단으로 그룹화하여 각 집단의 성격을 파악함으로써 자료 전체의 구조에 대한 이해를 돕고자 하는 분석이다. 대용량의 자료에서 개개의 관찰치를 요약하는 것보다는 전체를 유사한 관찰치들의 군집 (cluster)으로 구분하여 복잡한 전체보다는 그를 잘 대표하는 군집들을 관찰함으로써 전체 자료에 대한 의미 있는 정보를 얻어낼 수 있을 것이다. K-평균 군집분석은 계산이 비교적 간단하여 특히, 큰 자료의 개체 군집화에 효율적이고 여러 실험적 상황의 수행평가에서 상당히 좋은 결과를 내는 것으로 알려져 있다.

K-평균 군집분석의 알고리즘은 4단계로 다음과 같다.

[단계 1] K개 초기 군집들로 분할한다.

[단계 2] 모든 개체를 가장 가까운 중심점을 갖는 군집에 할당한다.

[단계 3] 군집 중심을 계산한다.

[단계 4] [단계 2]와 [단계 3]을 할당이 일어나지 않을 때 까지 반복한다.

이 때, 군집의 수 K와 초기 시드점을 정하기 위해 Sharama (1996, pp. 221–232)는 위계적 방법에서 나온 군집의 수를 K개 초기 군집의 수로 두고 초기 시드점으로 위계적 방법에서의 군집 중심을 사용하고 있다.

위계적 군집방법에는 단일연결, 완전연결, 중심연결, 평균연결, 와드 (WARD) 연결 등 많은 방법이 있다. 본 연구에서는 위계적 군집방법 중 가장 보편적으로 사용되는 평균연결, 중심연결, 와드연결을 사용하여 군집분석을 수행한다.

군집의 수 K를 결정하는 방법으로는 평균제곱 표준편차근 (root-mean-square standard deviation; RMSSTD), 반부분 R^2 (semipartial r-square; SPRSQ), R^2 (r-square; RSQ), CCC (cubic clustering criterion), pseudo-F (PSF), pseudo-t^2 (PST2) 등이 있다.

i번째 군집의 RMSSTD는

$$\sqrt{\frac{\sum_{k \in C_i} ||X_k - \bar{X}_i||^2}{p(N_i - 1)}}$$

으로 나타내어진다. 여기서 p는 변수의 수, N_i는 i번째 군집 C_i의 개체 수, X_k는 i번째 군집 C_i에 속해 있는 개체 $(k = 1, 2, \ldots, N_i)$, $\bar{X}_i$는 i번째 군집 C_i에서의 평균이다. 군집분석의 목적은 군집 내에서는 동질적이어야 하므로 RMSSTD값이 작아야 한다. 따라서 군집의 수에 대응되는 RMSSTD의 값을 그려서 급격한 감소가 발생하는 곳에서 대응되는 군집의 수를 정할 수 있다.

제곱합에 근거한 판정기준에 의해 R^2은 다음과 같이 정의된다.

$$R^2 = \frac{SS_b}{SS_t} \tag{2.6}$$

식 (2.6)에서 SS_t는 전체제곱합, SS_w는 군집내 (within-cluster)제곱합, SS_b는 군집간 (between-cluster)제곱합으로 다음과 같이 표현된다.

$$\begin{aligned} SS_t &= \sum_{i=1}^{g} \sum_{j=1}^{N_i} (X_{ij} - \bar{X})(X_{ij} - \bar{X})' \qquad (2.7) \\ &= \sum_{i=1}^{g} \sum_{j=1}^{N_i} (X_{ij} - \bar{X}_i)(X_{ij} - \bar{X}_i)' + \sum_{i=1}^{g} N_i (\bar{X}_i - \bar{X})(\bar{X}_i - \bar{X})' \\ &= SS_w + SS_b. \end{aligned}$$

식 (2.7)에서 g는 군집의 수를 말하며, X_{ij}는 i번째 군집에서 j번째 개체를 뜻한다. $\bar{X}_i$는 i번째 군집에서의 평균이며, $\bar{X}$는 모든 개체들의 평균이다 $(i = 1, 2, \ldots, g;\ j = 1, 2, \ldots, N_i)$. 일단 자료가 주어지면 SS_t는 고정되므로 SS_b와 SS_w의 관계로부터 군집내 제곱합에 비해서 군집간 제곱합이 크도록 하는 판정기준으로 생각할 수 있다. 따라서, R^2의 값이 급격한 증가가 발생한 곳에서 대응되는 군집의 수를 정할 수 있다.

SPRSQ는 군집분석에서 집단간 유사성을 측정하는 통계량으로, 특정 Y축 값 이내 (예: 0.1)에 구분이 되지 않는 경우는 그룹 내는 동질적이라고 한다. R^2의 경우와 반대로 급격한 감소가 발생한 곳에서 대응되는 군집의 수를 정할 수 있다.

CCC는 각각의 군집이 단일분포라고 가정할 때, 군집분석을 수행함으로써 군집내부의 분포가 단일분포와 달라질수록 좋다는 것을 이용하며, 관찰된 R^2와 단

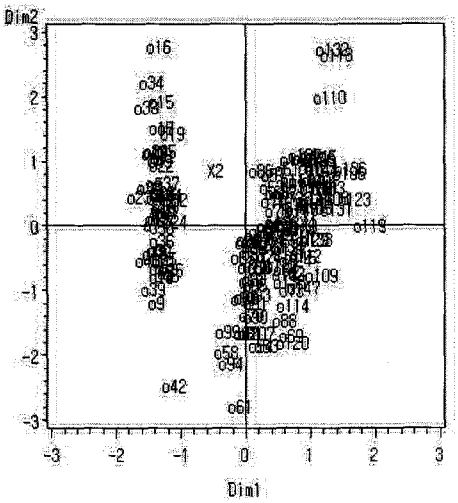

그림 3.1: 붓꽃자료에 대한 행렬도

일분포를 가정했을 때 R^2의 비를 기준 (criterion)으로 사용한다. 군집의 수와 CCC의 산점도를 그려 그 값이 3이상이면서 최대값인 경우 그 때의 군집의 수를 선택하는 방법이다.

pseudo-t^2 검정 통계량은 두 집단간 다변량 평균의 차이를 보는 통계량이다. 개체의 군집간 평균의 차이가 유의하지 않으면 두 군집을 합치고, 유의하면 군집을 그대로 유지하는 방법이다. pseudo-t^2 값이 크다는 것은 군집간 거리가 멀다는 것을 의미하므로 군집을 나누는 것이 좋고, 반대의 경우는 합치는 것이 좋다. pseudo-F 통계량도 이와 유사하지만, 보통 pseudo-t^2를 이용해 군집의 수를 결정한다.

3. 활용사례

피셔의 붓꽃자료는 세 종류의 붓꽃 50포기씩을 임의로 추출하여 X1(꽃받침 길이), X2(꽃받침 폭), X3(꽃잎 길이), X4(꽃잎 폭)을 측정한 자료이다 (Fisher, 1970). 본 연구에서는 세 종류의 붓꽃의 사전 분류 정보를 무시하고, 150포기의 붓꽃에 대해 행렬도분석을 실시하였고, 그 결과는 다음 그림 3.1과 같다. 각 축의 설명력은 제1축이 72.95%이고, 제2축이 22.82%로 모두 95.78%의 설명력을 가진다. 네 변수가 열 좌표점으로, 150개 붓꽃 (o1~o150)이 행 좌표점으로 표시되어 있지만, 사실상 붓꽃과 변수간의 해석은 거의 불가능함을 알 수 있다.

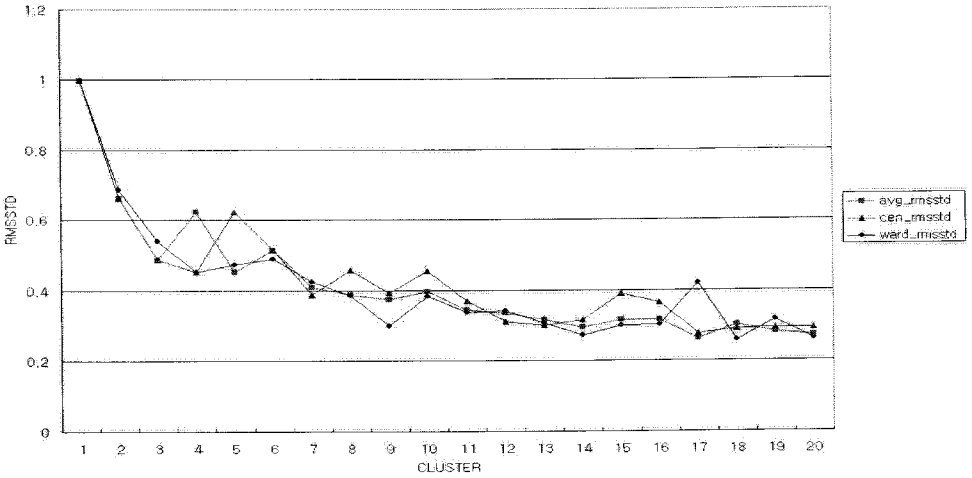

그림 3.2: 붓꽃자료에 대한 RMSSTD

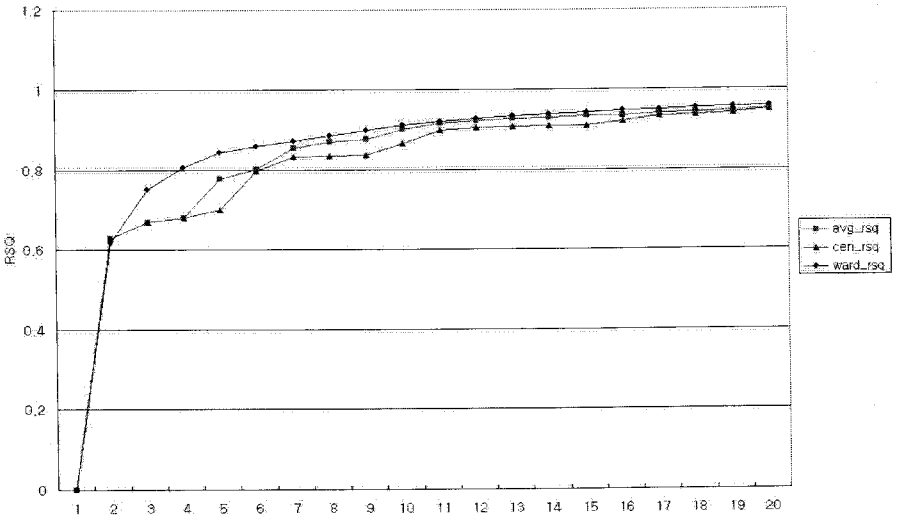

그림 3.3: 붓꽃자료에 대한 RSQ

피셔의 붓꽃자료를 이용하여 위계적 군집방법 중 평균연결(avg), 중심연결(cen), 와드연결(ward)을 수행하고 얻어지는 통계량으로 군집의 수를 정해보기로 하자.

표 3.1: 붓꽃자료에 대한 군집의 수

통계량	군집수
RMSSTD	2,3
RSQ	2,3
SPRSQ	2,3
CCC	2,3
PST2	2,3

표 3.2: 붓꽃자료에 대한 각 군집의 중심값

군집	x_1	x_2	x_3	x_4
CL1	0.88	−0.13	0.88	0.89
CL2	−0.73	0.11	−0.73	−0.74

그림 3.2와 그림 3.3은 RMSSTD와 RSQ 값들에 대한 평균연결, 중심연결, 와드연결 그림을 순서대로 보여주고 있다. 먼저 그림 3.2에서는 군집의 수에 대응되는 RMSSTD의 값을 그려서 급격한 감소가 발생하는 곳에서 대응되는 군집의 수를 정할 수 있으므로, 2개 혹은 3개 군집이 적당할 것으로 보인다. 그림 3.3에서는 RSQ의 값이 급격한 증가가 발생한 부분까지 군집의 수를 정할 수 있으므로, 2개 또는 3개 군집이 적당하다. 덧붙여 SPRSQ, CCC, PST2의 경우도 이와 대동소이한 결과를 나타내었다. 지금까지 군집의 수 K를 결정하는 방법들의 결과를 다시 정리하면 표 3.1과 같다.

원자료를 표준화하여 위계적 군집분석을 수행한 후 얻어진 표 3.1을 바탕으로 군집의 수로 가장 많이 선택됨과 동시에 가장 적은 수인 2개를 초기 군집의 수로 정하고, 초기 시드점으로 군집 중심을 사용하여 K-평균 군집분석을 수행한다. 그 결과, 표 3.2와 같이 변수별 각 군집의 중심값을 얻을 수 있다.

표 3.2의 군집중심값을 행렬도에 적용하면, 다음의 그림 3.4과 같은 행렬도를 그릴 수 있다. 이는 앞의 그림 3.1에서 나타난 행 좌표점과 열좌표들에 대한 해석상의 어려움을 각각의 군집의 군집중심값을 이용함으로써 해석을 가능하게 한다. 행렬도분석 결과, 그림 3.4에서 각 축의 설명력은 제1축이 99.9%으로 제1축만 고려하여도 설명력이 충분히 높음을 알 수 있다. 네 변수가 열 좌표점으로 150개 붓꽃들에 대한 군집 (CL1~CL2)이 행 좌표점으로 표시되어 있다. 열 좌표점의 경우 X1(꽃받침 길이), X3(꽃잎 길이), X4(꽃잎 폭)가 거의 같은 좌표점에 위치하고 있어서 서로 상관이 높은 변수임을 보여주고 있다. 행 좌표점인 군집에 대해 살펴보

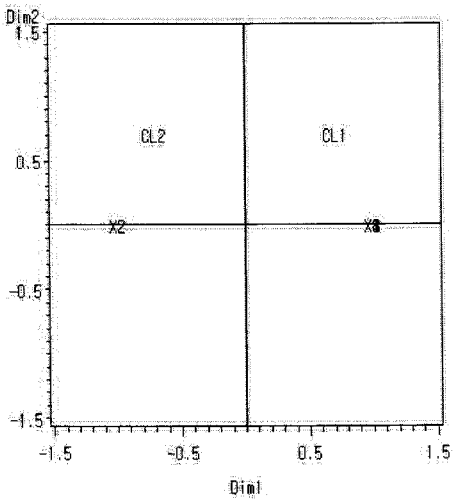

그림 3.4: 붓꽃자료의 군집중심값을 활용한 행렬도

표 3.3: 붓꽃자료의 각 군집에 속하는 붓꽃

군집	군집에 속한 붓꽃								
1군집 (CL1)	51	52	53	55	57	59	64	66	71
	73	75	76	77	78	79	84	86	87
	92	98	101	102	103	104	105	106	108
	109	110	111	112	113	114	115	116	117
	118	119	121	122	123	124	125	126	127
	128	129	130	131	132	133	134	135	136
	137	138	139	140	141	142	143	144	145
	146	147	148	149	150				
2군집 (CL2)	1	2	3	4	5	6	7	8	9
	10	11	12	13	14	15	16	17	18
	19	20	21	22	23	24	25	26	27
	28	29	30	31	32	33	34	35	36
	37	38	39	40	41	42	43	44	45
	46	47	48	49	50	54	56	58	60
	61	62	63	65	67	68	69	70	72
	74	80	81	82	83	85	88	89	90
	91	93	94	95	96	97	99	100	107
	120								

면, 제1축을 기준으로 CL1은 X2(꽃받침 넓이)를 제외한 다른 변수들과 같은 방향으로 놓여져 있어 CL2 군집에 비해 상대적으로 X1, X3, X4 변수의 값이 큰 군집임을 알 수 있다. CL2는 X2변수 방향으로 놓여져 있기 때문에 X2 변수를 제외한 모든 변수의 값이 상대적으로 작은 값을 가진다. 즉, CL1은 꽃받침 폭을 제외한 모든 변수가 상대적으로 값이 큰 군집이며, CL2는 X2 변수만 큰 값을 가지는 군집 특성을 가지고 있다.

따라서, 피셔의 세 종류의 붓꽃 (세토사, 베르시칼라, 비르지니카)에 대한 원자료에서 세 종류의 붓꽃 분류에 대한 정보를 무시하였을때, 각각의 붓꽃은 표 3.3과 같이 2군집으로 고려할 수 있다. 세토사는 X2 변수가 상대적으로 크고, 나머지 변수에 대해서는 작은 값을 가짐을 알 수 있으며, 비르지니카는 세토사와는 정반대의 특성을 가짐을 알 수 있었다. 끝으로 베르시칼라는 세토사와 비르지니카가 속해 있는 군집 모두에 포함되어 있음을 알 수 있다.

4. 결론

행렬도가 자료의 정보를 시각적으로 표현하여 쉽게 정보를 파악할 수 있는 장점을 가지고 있지만 대용량 자료에서는 해석이 힘든 문제점을 가지고 있다. 이러한 한계를 극복하기 위해 군집분석을 행렬도에 활용하고자 하였고, 위계적 방법에서 얻어지는 군집의 수 K와 군집 중심을 초기 시드점으로 한 K-평균 군집분석을 수행한다. 따라서, 원자료의 행렬도가 아닌 각 군집과 변수간의 행렬도 분석을 수행함으로써, 기존에는 행렬도에서 해석의 어려움이 있었던 대용량 자료에 대한 해석이 가능하게 되었다.

K-평균 군집분석을 통해 얻어진 각 군집의 중심값 만으로 각 군집에 대한 특징을 파악하는데 별 무리가 따르지 않는다. 여기에 행렬도를 응용함으로써 각 변수들 간의 관계와 이러한 변수들과 군집들 간의 관계를 추가적으로 알 수 있고, 정보의 파악이 한결 쉬워졌다.

하지만 많은 군집이 필요한 대용량의 자료의 경우, 또다시 행렬도를 통한 2차원 그림 상에서의 해석에 제약이 따를 것으로 생각된다. 또한, 행렬도가 SAS/IML을 이용한 프로그램의 형태를 띠고 있어 대용량 자료를 이용하는데 불편함이 존재한다. 따라서, 많은 군집을 이용할 경우에 행렬도의 표현 방법에 대한 연구가 필요할 것으로 보이며, 프로그램 또한 각 군집에 대한 행렬도를 나타내거나, 군집과 변수간의 행렬도를 바로 출력하는 형태로 개발을 하는 것이 필요하다.

참고문헌

최용석 (2006). <행렬도 분석>. 기초과학 총서 2권, 부산대학교 기초과학연구원.
허명회 (1993). <統計相談의 이해>. 자유아카데미, 서울.
Bradu, D. and Gabriel, K. R. (1978). The biplot as a diagnostic tool for models of two-way tables. *Technometrics,* **20**, 47–68.
Choi, Y. S. (1991). Resistant principal component analysis, biplot and correspondence analysis. Unpublished Ph.D. Dissertation, Department of Statistics, Korea University.
du Toit, S. H. C., Steyn, A. G. W. and Stumpf, R. H. (1986). *Graphical Exploratory Data Analysis.* Springer-Verlag, New York.
Gabriel, K. R. (1971). The biplot graphics display of matrices with applications to principal component analysis. *Biometrika,* **58,** 453–467.
Gabriel, K. R. (1981). Biplot display of multivariate matrices for inspection of data and diagnosis. In *Interpreting Multivariate Data* (Barnett, V., ed), 147–173, Wiley, New York.
Gower, J. C. and Hand, D. J. (1996). *Biplots.* Chapman & Hall/CRC, London.
Jolliffe, I. T. (1986). *Principal Component Analysis.* Springer-Verlag, New York.
Fisher, R. A. (1970). *Statistical Methods for Research Workers.* 14th ed. (originally published 1925), Edinburgh, Oliver and Boyd.
Sharama, S. (1996). *Applied Multivatiate Techniques.* Wiley, New York.

[2007년 9월 접수, 2007년 10월 채택]

Applications of Cluster Analysis in Biplots†

Yong-Seok Choi[1)], Hyoung-Young Kim[2)]

Abstract

Biplots are the multivariate analogue of scatter plots. They approximate the multivariate distribution of a sample in a few dimensions, typically two, and they superimpose on this display representations of the variables on which the samples are measured(Gower and Hand, 1996, Chapter 1). And the relationships between the observations and variables can be easily seen. Thus, biplots are useful for giving a graphical description of the data. However, this method does not give some concise interpretations between variables and observations when the number of observations are large. Therefore, in this study, we will suggest to interpret the biplot analysis by applying the K-means clustering analysis. It shows that the relationships between the clusters and variables can be easily interpreted. So, this method is more useful for giving a graphical description of the data than using raw data.

Keywords: Biplots; K-means cluster analysis.

† This work was supported by Research Institute for Basic Sciences, Pusan National University(2007).

1) Professor, Department of Statistics, Pusan National University, Busan 609-735, Korea. Correspondence : yschoi@pusan.ac.kr

2) Graduate Student, Department of Statistics, Pusan National University, Busan 609-735, Korea.