한국통계학회논문집 제2권 2호
1995년 12월 pp. 414-422

그래픽스를 이용한 판별분석법[1)]

김 성 주[2)]

요 약

본 논문에서는 그래픽스에 의한 판별분석을 다루고 있다. 본 논문에서 제안하는 새로운 그래프는 표본이차판별함수에 기초하고 있으며 기존의 MV 그래프와 실제자료에 대하여 비교하고 있다. 판별분석에서 공분산행렬이 같지 않은 경우의 3차원 그래프는 처음 시도된 것으로서 이를 위하여 차원축소문제를 논의하고 있다.

1. 서론

판별분석에서 연습표본(training samples)과 판별함수(discriminant function)를 2~3차원 그래프로 표시할 수 있다면 자료분석자에게 매우 유익할 것이다. 왜냐하면 자료분석자는 그래프를 살펴봄으로서 연습표본이 어느 정도 분리되었는지에 관한 요약된 정보를 직관적으로 파악할 수 있기 때문이다. 각 그룹의 공분산행렬이 모두 같다면 우리는 Fisher의 선형판별함수를 이용하여 2~3차원 그래프를 쉽게 얻을 수 있다. 그러나 각 그룹의 공분산행렬이 모두 같지 않은 경우 2~3차원 그래프를 얻는 것은 쉽지 않은 문제로서 그 뿌리는 소위 "Behrens-Fisher Problem"에 귀착된다. 이 경우 어떠한 거리측도(distance measure)를 이용할 것인지는 논란의 대상이 되며 보다 구체적인 사항은 Gnanadesikan (1977, chap. 4), Kim (1992)에서 다루고 있다.

그래픽스에 의한 판별분석은 Sammon (1970)에 의해 맨 처음 시도되었다. Sammon (1970)은 그룹의 수 g=2이고 공분산행렬이 같은 경우 1차원 수직선에 표시되는 선형판별함수를 2차원 평면으로 확장할 수 있는 방안을 제안하였으며 이를 Sammon 그래프라고 한다. 즉 Sammon (1970)은 g=2인 경우 선형판별함수는 $g\geq 3$인 경우 선형판별함수로부터 얻어질 수 있다는 사실에 착안하였다. g=2인 경우 선형판별함수는 1차원 수직선에만 표시되나 $g\geq 3$인 경우에는 2차원 이상에 표시될 수 있으므로 g=2인 경우에도 $g\geq 3$인 경우를 이용하여 첫 번째 축과 직교한다는 조건하에서 두 번째 축을 찾았다. Sammon 그래프의 두 번째 축은 자연스럽지 못하다고 생각되나 Chien (1978)에 의하면 Sammon 그래프는 그리기 쉽고 간단하기 때문에 패턴인식 분야에서 널리 이용된다고 한다.

Sammon 그래프는 기본적으로 공분산행렬이 같다는 가정 하에서 얻어지나, Chang (1987)은 공분산행렬이 같지 않은 경우 Sammon 그래프의 대안으로서 MV 그래프를 제안하였다. 즉 2차원 평면에 표시되는 MV 그래프의 첫 번째 축은 Sammon 그래프와 마찬가지로 선형판별함수에 의해

1) 이 논문은 1993년도 교육부지원 한국학술진흥재단의 대학교수 국비해외파견 연구지원비에 의해 연구된 결과의 일부임.
2) (110-745) 서울시 종로구 명륜동 성균관대학교 통계학과

얻어지나 두 번째 축은 공분산행렬의 차이가 최대가 되도록 얻어진다. Sammon 그래프와 MV 그래프에 관한 보다 구체적인 설명은 김성주 · 정갑도 (1993)에서 다루고 있다.

본 논문에서는 공분산행렬이 같지 않은 경우 이차판별함수를 2~3차원 그래프로 나타낼 수 있는 새로운 방안을 제시하고자 한다. 본 논문에서 다루는 변수의 수 p는 일반적으로 $p>2$이며 그룹의 수 g는 $g=2$ 인 경우를 주로 다루고 있으며 $g\geq 3$인 경우는 $g=2$ 의 결과를 산점도행렬 (scatterplot matrix)로 표시하고자 한다.

제 2절에서는 새로운 그래프를 제안하고 있으며 제 3절에서는 새로운 그래프를 2~3차원에 나타내기 위한 차원축소문제를 다루고 있다. 제 4절에서는 실제자료를 이용하여 새로운 그래프를 MV 그래프와 비교하고 있다.

2. 새로운 그래프

그룹의 수 $g=2$ 인 경우 p차원 연습표본을 x_{ij} ($j=1,\cdots,n_i$; $i=1,2$) 라고 하고 이는 사전확률 Π_i, 모평균벡터 μ_i, 모공분산행렬 Σ_i인 모집단에서 추출된 확률표본이라고 하자. 연습표본에 대하여 보통 하는 대로 표본평균 $\bar{x}_i=\sum_j x_{ij}/n_i$, 표본공분산행렬 $S_i=\sum_j(x_{ij}-\bar{x}_i)(x_{ij}-\bar{x}_i)'\ /\ (n_i-1)$ 라고 정의하자. 또한 어느 그룹에 속하는지 알려져 있지 않은 검정표본(test sample) 을 x라고 하고 x와 $\bar{x}_i$사이의 표본 Mahalanobis 거리를 $G(x,\bar{x}_i;S_i)=(x-\bar{x}_i)'S_i^{-1}(x-\bar{x}_i)$ 라고 표시하자. 표본이차판별함수에 의한 판별분석은 식 (1)에 정의된 R에 대하여 $x\in R$이면 x를 그룹 1로 판별하고 $x\notin R$이면 x를 그룹 2로 판별한다. 식 (1)에서 c는 어떤 상수이며 R을 이차영역 (quadratic region)이라고 부르기로 하자.

$$R=\{x: G(x,\bar{x}_1;S_1)-G(x,\bar{x}_2;S_2)\leq c\}. \tag{1}$$

식 (1)은 여러 의미를 내포하고 있다. Seber (1984, chap. 6)에 의하면 모집단의 분포가 다변량 정규분포인 경우 잘못 판별할 확률(total probability of misclassification)을 최소화하는 판별법은 식 (1)에서 $c=2\ln(\Pi_1/\Pi_2)-\ln(|S_1|/|S_2|)$ 일 때 얻어짐을 알 수 있다. 또한 Gnanadesikan (1977, eq. 58)에 의하면 식 (1)은 다변량 정규분포라는 가정을 하지 않더라도 직관적인 판별 법임을 밝히고 있다. 식 (1)에 나타난 이차영역 R은 표본 Mahalanobis 거리인 $G(x,\bar{x}_1;S_1)$과 $G(x,\bar{x}_2;S_2)$의 차이로 표시되며 $G(x,\bar{x}_i;S_i)$는 모평균벡터 μ_i에 대한 신뢰타원과 관련 있음을 알 수 있다. 즉 모집단이 다변량 정규분포를 따를 경우 μ_i에 대한 $100(1-\alpha)\%$ 신뢰타원은 식 (2)에 있는 E_i이다. 식 (2)에서 c_i는 상수로서 $F(p,n_i-p;\alpha)$ 를 자유도가 p와 n_i-p 인 F분포의 상위 100α 백분위수라고 하면 $c_i=F(p,n_i-p;\alpha)p(n_i-1)/\{n_i(n_i-p)\}$이다.

$$E_i = \{x : G(x, \bar{x}_i ; S_i) \le c_i\} . \tag{2}$$

본 논문의 결과를 소개하기 전에 공분산행렬 S_1과 S_2에 관한 일반화 고유값 문제 (generalized eigenvalue problem)인 식 (3)을 생각해 보자.

$$S_1 t_k = d_k(S_2 t_k), \ k=1,\cdots,p . \tag{3}$$

식 (3)을 만족하는 고유값을 $d_1 \ge d_2 \cdots \ge d_p > 0$라고 하고, $D = diag[d_1,\cdots,d_p]$ 라고 하고, $T = [t_1,\cdots,t_p]$ 라고 하자. 또한 $p\times p$ 단위행렬(identity matrix)을 I라고 하자. 그러면 S_1과 S_2는 다음과 같이 동시에 대각화 된다.

$$T'S_1T = D, \quad T'S_2T = I .$$

본 논문에서는 식 (4)에 정의된 선형변환을 고려하고자 한다. 식 (4)에 의하면 S_1과 S_2는 동시에 대각화됨은 물론 E_1의 중심이 원점에 있게 된다.

$$y = T'(x - \bar{x}_1) . \tag{4}$$

결과 1. 선형변환 $y = T'(x - \bar{x}_1)$에 의해 E_1, E_2 및 R은 다음과 같이 E_1^*, E_2^* 및 R^*로 대각화 된다. 여기서 상수 m과 c^* 그리고 대각행렬 L은 각각 식 (5), (6), (7)에 나타나 있다.

$$E_1^* = \{y : y'D^{-1}y \le c_1\} .$$

$$E_2^* = \{y : (y-m)'(y-m) \le c_2\} .$$

$$R^* = \{y : (y+L^{-1}m)'L(y+L^{-1}m) \le c^*\} .$$

증명. 선형변환 y에 의해 E_1^*, E_2^*는 분명하다. 이제

$$R = \{y : y'D^{-1}y - (y-m)'(y-m) \le c\},$$

$$= \{y : y'(D^{-1}-I)y + 2m'y \le c + m'm\},$$

이므로 위 식을 제곱 꼴로 바꾸면 증명은 끝난다. 여기서

$$m = T'(\bar{x}_2 - \bar{x}_1), \tag{5}$$

$$c^* = c + m'(L^{-1}+I)m, \tag{6}$$

$$L = D^{-1} - I. \tag{7}$$

결과 1에 나타난 E_1^*, E_2^*, R^*에 대해 살펴보자. E_1^*는 중심이 원점에 있는 대각화된 타원 또는 타원체이다. E_2^*는 중심이 m에 있는 원 또는 공이다. R^*는 중심이 $-L^{-1}m$에 있는 대각화된 이차곡면으로서 대각행렬 L의 대각요소의 부호에 따라 그 모양이 달라진다. $p=2$인 경우 L의 대각요소의 부호가 모두 양이면 R^*는 타원이며 하나만 양이면 R^*는 쌍곡선이다. $p=3$인 경우 L의 대각요소의 부호가 모두 양, 두개만 양, 하나만 양이냐에 따라 R^*는 각각 타원체, 1장의 쌍곡면(hyperboloid of one sheet), 2장의 쌍곡면(hyperboloid of two sheets)이 된다. 이차곡면에 관한 자세한 사항은 Fraleigh & Beauregard (1990)에서 찾을 수 있다.

한편 MV 그래프를 제안한 Chang (1987)은 Sammon 그래프에 대한 MV 그래프의 장점으로서 MV 그래프의 축을 (Z_1, Z_2)라고 하면 Z_1+Z_2는 근사적으로 표본이차판별함수를 나타낸다고 주장하고 있다. 만약 MV 그래프가 표본이차판별함수를 근사적으로 나타낸다면 표본이차판별함수 자체를 그래프로 표시하면 어떨까? 본 논문은 이러한 의문으로부터 시작되었다. 본 논문에서 제안하는 새로운 그래프는 결과 1에 나타난 E_1^*, E_2^*, R^* 를 연습표본과 함께 나타내고자 한다.

3. 차원축소

결과 1에 나타난 E_1^*, E_2^*, R^*가 비록 대각화 되어 있다고 하더라도 변수의 수 $p\geq 4$이면 4차원 이상은 현실적으로 표현할 수 없다. 이러한 문제를 해결하기 위하여 차원축소를 논의하고자 한다.

그룹의 수 $g\geq 3$ 이고 모공분산행렬이 모두 같은 경우 Fisher의 선형판별함수를 생각해 보자. 이를 위하여 제2장에서 다룬 $g=2$ 인 경우의 연습표본 $x_{ij}(j=1,\cdots,n_i\ ;i=1,2)$ 를 $g\geq 3$ 인 경우로 확장하여 $x_{ij}\ (j=1,\cdots,n_i\ ;\ i=1,\cdots,g)$ 라고 하자. 총평균 $\bar{x} = \sum_{i=1}^{g}\sum_{j=1}^{n_i} x_{ij}/(\sum_i n_i)$, 표본급간행렬 $B = \sum_{i=1}^{g}(\bar{x}_i-\bar{x})(\bar{x}_i-\bar{x})'$, 표본급내행렬 $W = \sum_{i=1}^{g}(n_i-1)S_i$, 합동공분산행렬 $S_p = W/(\sum_{i=1}^{g} n_i - g)$ 라고 표시하자. 또한 $W^{-1}B$의 고유벡터를 $e_k(k=1,\cdots,p)$ 라고 하면 Fisher의 k번째 선형판별함수는 $e_k'x$ 로 주어진다.

한편 Fisher의 선형판별함수의 효용성을 측정할 수 있는 한가지 방안으로 식 (8)에 정의된 분리측도 (separatory measure) h_1을 생각해 볼 수 있다.

$$h_1 = \sum_{i=1}^{g}(\bar{x}_i-\bar{x})'S_p^{-1}(\bar{x}_i-\bar{x}). \qquad (8)$$

즉 식 (8)에 정의된 분리측도는 각 그룹에서 표본평균과 총평균간의 표본 Mahalanobis 거리의 합이며 이 값이 클수록 각 그룹은 잘 분리되어 있다고 볼 수 있다. 또한 Johnson & Wichern(1992,

chap 11)에 의하면 식 (8)의 h_1은 $W^{-1}B$의 고유값들의 합과 같고 k번째 선형판별함수 $e_k'x$ 에 의해 설명되는 양은 고유벡터 e_k에 해당되는 고유값과 같음을 알 수 있다.

식 (8)의 분리측도는 모공분산행렬이 모두 같다는 가정 하에서 정의된 것이다. 만약 모공분산행렬이 모두 같다고 할 수 없다면, 분리측도는 식 (9)와 같이 정의하는 것이 타당할 것이다.

$$h_2 = \sum_{i=1}^{g}(\bar{x}_i - \bar{x})'S_i^{-1}(\bar{x}_i - \bar{x}). \tag{9}$$

식 (9)에서 $g=2$인 경우를 정리하면 다음과 같다.

$$\begin{aligned} h_2 &= (\bar{x}_2 - \bar{x}_1)'(S_1^{-1}+S_2^{-1})(\bar{x}_2 - \bar{x}_1)/4 \\ &= m'(D^{-1}+I)m/4 \\ &= \sum_{k=1}^{p} b_k, \end{aligned}$$

여기서

$$b_k = (d_k+1)m_k^2/(4d_k). \tag{10}$$

식(10)에 있는 d_k는 식(3)에서 정의한 고유값이고, m_k는 식(5)에서 정의한 p차원 벡터 m의 k번째 성분이다.

결론적으로 새로운 그래프는 식 (4)의 선형변환에 의해 연습표본과 각 모평균에 대한 신뢰타원인 E_1^*, E_2^*와 이차영역인 R^*를 나타내고자 한다. 비록 E_1^*, E_2^*, R^*가 모두 대각화 되어 있지만, $p \geq 4$인 경우에는 현실적으로 4차원이상의 공간에 E_1^*, E_2^*, R^* 를 나타낼 수 없으므로 새로운 그래프는 식 (10)에 있는 b_k값이 큰 순서대로 그에 해당하는 2~3개 축을 선정하여 2차원 평면 또는 3차원 공간에 E_1^*, E_2^*, R^* 를 나타내고자 한다. 예를 들어 제4절 예제 1에 있는 <표1>과 같이 $b_3 > b_1 > b_4 > b_2$라고 하자. 이 경우 새로운 그래프를 2차원 평면에 표시할 때는 선형변환된 공간에서 세 번째 축과 첫 번째 축을 택하여 나타내고, 3차원 공간에 표시할 때는 세 번째 축과, 첫 번째 축과 네 번째 축을 택하여 나타내고자 한다. b_k값이 큰 순서대로 축을 선택하는 이유는 식 (9)에 정의된 분리측도는 b_k의 합이므로 b_k가 클수록 분리측도를 더 많이 설명할 수 있기 때문이다. 이는 모공분산행렬이 모두 같은 경우 Fisher의 선형판별함수에서 식 (8)에 정의된 분리측도는 $W^{-1}B$의 고유값의 합이므로 고유값이 큰 순서대로 그에 해당하는 고유벡터에 의해 축을 결정하는 것과 마찬가지이다.

4. 예제

실제자료를 이용하여 새로운 그래프를 나타내고 이를 기존의 MV 그래프와 비교해 보고자 한다. Chang (1987)에 의하면 MV 그래프가 Sammon 그래프 보다 우수하다고 알려져 있기 때문에 본 논문에서는 Sammon 그래프와의 비교는 생략하였다. 연습표본 중에서 잘못 판별된 관측값의 수를 f_1+f_2 로 표시하자. 여기서 f_1은 그룹 1인데 그룹 2로 잘못 판별된 관측값의 수이며 f_2은 그룹 2인데 그룹 1로 잘못 판별된 관측값의 수이다. 예를 들어 2+3은 연습표본 중에서 5개 관측값이 잘못 판별되었는데 그룹 1인데 그룹 2로 잘못 판별된 관측값의 수는 2이고 그룹 2인데 그룹 1로 잘못 판별된 관측값의 수는 3이다.

예제 1. 4개 변수에 대하여 50번 반복측정한 Fisher의 Iris 데이터에서 Iris versicolor (그룹 1)와 Iris virginica (그룹 2)를 고려해 보자. Iris setosa를 배제한 이유는 Iris setosa는 다른 그룹과 분명히 구별되기 때문이다. MV 그래프는 그림 1에 나타나 있다. 그림 1에 있는 판별선은 시각적으로 두 그룹을 가장 잘 분리할 수 있도록 그려지므로 자료분석자의 주관이 개입될 수 있다. 이는 과학적 자료분석에서는 바람직하지 못하다고 생각된다. 이러한 관점을 단적으로 나타내기 위하여 그래프에 의하지 않고 대수적으로 표본이차판별함수를 적용해 보면 잘못 판별된 관측값의 수는 2+1이다. 그러나 MV 그래프에서 잘못 판별된 관측값의 수는 1+1이다. MV 그래프는 근사적으로 표본이차판별함수를 나타낸다고 하는데 MV 그래프에 의한 방법이 표본이차판별함수에 의한 방법보다 더 나은 결과를 가져다준다는 것은 매우 이해하기 어렵다.

새로운 그래프를 그리기 위하여 d_k, m_k, b_k를 구하면 아래 표와 같다. 첫 번째 축과 세 번째 축에 의한 b_k는 각각 3.17과 3.18이므로 두 축에 의해 분리측도 $h_2=7.92$ 중에서 6.35(100)/7.92=80%가 설명된다. 여기서 네 번째 축을 추가할 경우 b_k는 1.19이므로 분리측도 중에서 7.54(100)/7.92=95%가 설명된다. 따라서 첫 번째, 세 번째, 네 번째 축을 이용하여 3차원 공간에 나타낸 새로운 그래프가 그림 2에 나타나 있다. 여기서 E_1^*와 E_2^*는 μ_1과 μ_2에 대한 90% 신뢰타원을 나타내며 연습표본 중에서 Iris versicolor는 '•' Iris versinica는 '⊠'로 나타낸다. 그림 3에서 잘못 판별된 관측값의 수는 0+8이다.

<표1> Iris 데이터에서 d_k, m_k, b_k

k	d_k	m_k	b_k
1	1.40	2.72	3.17
2	0.88	-0.84	0.38
3	0.65	2.24	3.18
4	0.18	0.85	1.19
			$h_2=7.92$

예제 2. Lubischew (1962)가 다룬 3종류의 수놈 벼룩 (male flea beetles)중에서 3변수 (x_{10}, x_{14}, x_{18})를 고려해 보자. 즉 22마리의 C. hetapotamica (그룹 1), 31마리의 C. heikertingeri (그룹 2), 21마리의 C. concinna (그룹 3)를 연습표본이라고 하자. 3종류의 수놈 벼룩에 대한 새

로운 그래프는 산점도행렬 형태로 그림 3에 나타나 있다. 그림 3에 나타난 3개의 이차영역 R^*는 C. hetapotamica & C. heikertingeri인 경우는 타원체인데 그 일부를, C.hetapotamica & C. concinna인 경우는 2장의 쌍곡면 (hyperboloid of two sheets) 중에서 1장을, C.heikertingeri & C. concinna인 경우는 장구 모양의 1장의 쌍곡면 (hyperboloid of one sheet)인데 그 반쪽을 보여주고 있다. 그림 3에서 모평균에 대한 90% 신뢰타원은 E_1^*와 E_2^*로 나타내며 연습표본은 '•' 또는 '⊠'로 나타낸다. 그림 3에서 잘못 판별된 관측값의 수는 모두 0+0이다. Chang (1987)이 제안한 MV 그래프는 $g=2$ 인 경우만 다루었으므로 $g=3$ 인 이 예제에서는 언급하지 않았다.

참 고 문 헌

[1] 김성주 · 정갑도 (1993). 공분산행렬이 서로 다를 경우 그래프에 의한 판별분석, 「응용통계연구」, 제6권 2호, 409-419.

[2] Chang, W.C. (1987). A graph for two training samples in a discriminant analysis, *Applied Statistics*, Vol. 36, 82-91.

[3] Chien, Y. (1978). *Interactive Pattern Recognition*, Marcel Dekker, New York.

[4] Fraleigh, J.B. and Beauregard, R.A. (1990). *Linear Algebra* (2nd ed.), Addison-Wesley, New York.

[5] Gnanadesikan, R. (1977). *Methods for Statistical Data Analysis of Multivariate Observations*, New York, John Wiley.

[6] Johnson, R.A. and Wichern, D.W. (1992). *Applied Multivariate Statistical Analysis* (3rd ed.), Prentice-Hall, New Jersy.

[7] Kim, S.-J. (1992). A practical solution to the multivariate Behrens-Fisher problem, *Biometrika*, Vol. 79, 171-176.

[8] Lubischew, A.A. (1962). On the use of discriminant functions in taxonomy, *Biometrics*, Vol. 18, 455-477.

[9] Sammon, J.W. and Jr. (1970). An optimal discriminant plane, *IEEE Transactions on Computers*, Vol. C-19, 826-829.

[10] Seber, G.A.F. (1984). *Multivariate Observations*, John Wiley, New York.

그림 1.
Iris versicolor(S)와 Iris virginica(G)에 대한 MV 그래프와 판별선

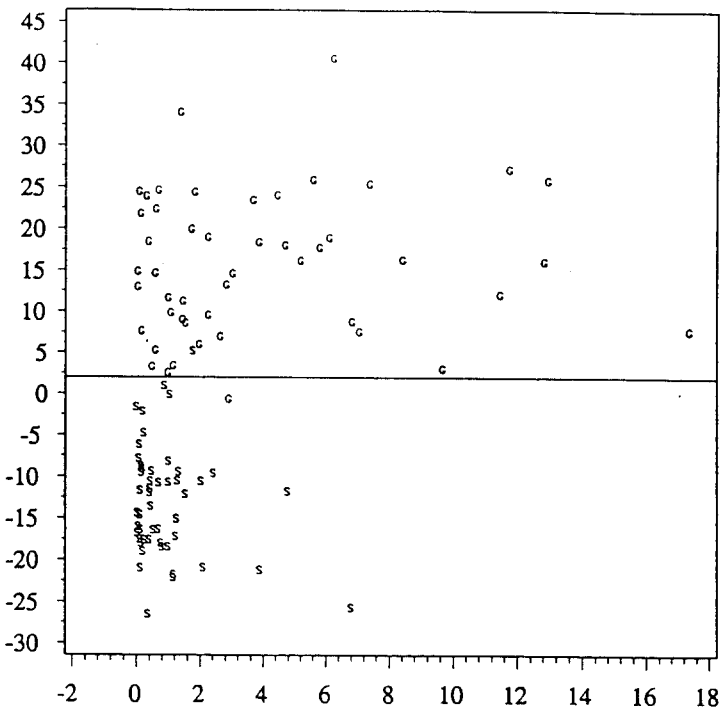


그림 2.
Iris versicolor(•)와 Iris virginica(⊠)에 대한 새로운 그래프

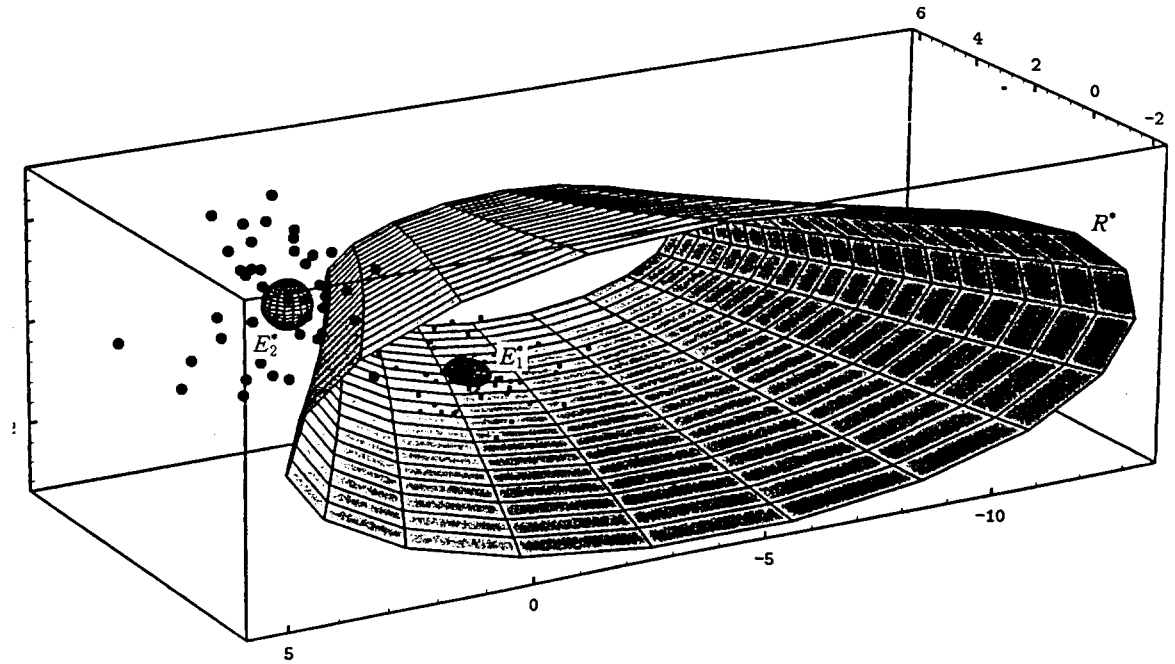
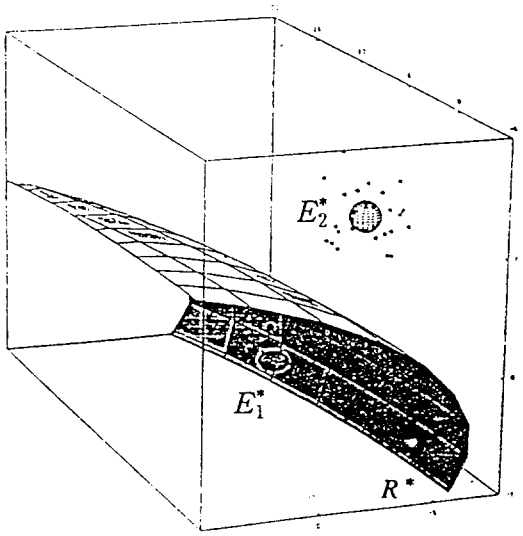
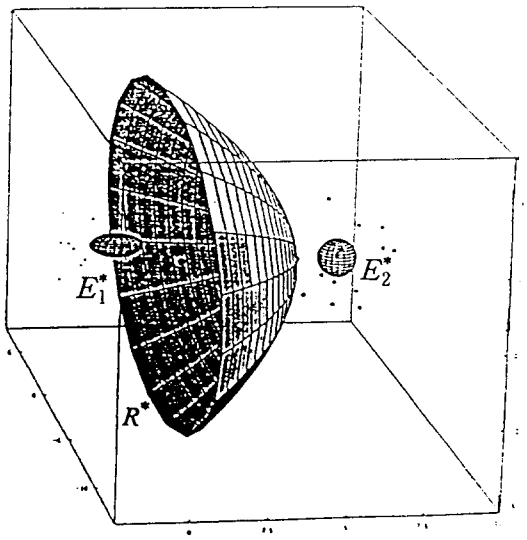


그림 3.
Lubischew(1962)의 3종류의 수놈 벼룩에 대한 새로운 그래프의 산점도행렬

C. hetapotamica (•) & C. heikertingeri (•)



C. hetapotamica (•) & C. concinna (•)



C. heikertingeri (•) & C. concinna (•)