응용통계연구 제 6권 2호
1993년 9월 pp. 409-419.

공분산행렬이 서로 다를 경우 그래프에 의한 판별분석[1)]

김 성 주[2)], 정 갑 도[3)]

요 약

본 논문은 요즈음 국내외를 막론하고 통계학계에서 활발히 연구하고 있는 그래프에 의한 통계적 방법의 일부로서 그래프에 의한 판별분석법을 다루고 있다. 기존에 알려진 Sammon 그래프와 MV 그래프를 바탕으로 새로운 대안의 가능성을 소개하고 있으며 그룹의 수가 2인 경우 실제자료와 모의실험을 이용하여 3가지 그래프의 특성을 비교·분석하고 있다. 새로운 대안이 해결해야할 차원축소 문제와 로버스트 방법에 대한 앞으로의 과제를 간략히 언급하고 있다.

1. 서 론

요즈음 통계학계에서는 그래프에 의한 통계적 방법을 활발히 연구하고 있으며 국내에서도 금년 들어 한국통계학회가 후원한 2건의 workshops(Cook 교수의 "Graphics and Diagnostics for Regression" 및 Wakimoto 교수의 "Recent Developments in Statistical Graphics")이 개최되어 이 분야의 열기를 더해주고 있다. 본 논문에서는 통계적 방법의 한 분야로서 판별분석법에 대하여 그룹의 수 $g = 2$인 경우 그래프에 의한 방법을 다루고자 하며 기존의 연구결과를 정리하고 이를 바탕으로 새로운 대안의 가능성을 제시하고자 한다.

판별분석법은 여러 분야에서 널리 이용되고 있으나, 통계학적 측면에서 살펴보면 다음과 같이 특징 지워진다. 즉 두 종류의 다변량 관측값들이 주어져 있다고 하자. 한 종류의 관측값들에 대해서는 g개 그룹중에서 어느 그룹에 속하는지를 알고 있으나 다른 종류의 관측값들에 대해서는 어느 그룹에 속하는 지가 알려져 있지 않다고 하자. 첫번째 경우의 관측값들을 연습표본(training samples) 이라고 하고 두번째 경우의 관측값들을 검정표본(test samples) 이라고 한다. 판별분석이란 연습표본을 기초로 하여 검정표본에 있는 관측값들이 어느 그룹에 속하는지를 판별하고자 하는 것이다.

역사적 관점에서 살펴보면, Fisher(1936)가 두 집단의 공분산행렬이 서로 같다는 가정하에서 선형판별함수(linear discriminant function)를 맨처음 제안하였다. 또한 Efron(1975)에 의해 논의된 바와 같이 두 그룹중 어느 그룹에 속하는지를 0 또는 1 값을 갖는 가변수 형태의 반응변수로 나타내면, Fisher가 제안한 선형판별함수는 로지스틱 회귀모형과 대응하게 된다.

두 공분산행렬이 서로 다를 경우에는 이차판별함수(quadratic discriminant function)가 제안되었다. 이차판별함수는 주어진 변수들에 관한 선형식으로 표시되지 않기 때문에, 선형식으로 표시되는 함수 중에서 가장 적절한 판별함수를 찾고자 노력하게 된다. 이러한 문제에 대한 해결책은 Anderson and Bahadur(1962)에 의해 수행되었으며 김혜중(1992)은 공분산행렬이 서로

1) 이 논문은 1991년도 교육부지원 한국학술진흥재단의 자유공모과제 학술연구조성비에 의해 연구되었음.
2) (110-745) 서울시 종로구 명륜동 성균관대학교 통계학과 부교수
3) (137-701) 서울시 서초구 반포동 가톨릭대학교 의과대학 통계학과 조교

다룰 경우 변수선택문제를 논의한 바 있다. 앞에서 언급한 판별분석법들은 주로 다변량정규분포 가정에 기초하고 있으나 다변량정규분포 가정을 떠난 비모수적 내지는 로버스트 판별분석법에 대한 연구도 활발히 수행되어 왔다. 이에는 커널(kernel) 또는 k최인접(kth-nearest-neighborhood)에 기초한 분포추정방법(Silverman 1986), 요즈음 활발히 연구되고 있는 투영추구방법(projection pursuit method) 및 순위(rank)에 의한 방법(Broffitt 1982)등을 들 수 있다.

이러한 여러 형태의 판별분석법 중에서 과연 어떤 방법이 좋은 것인가에 관한 연구는 안윤기·이성석(1992), Marks and Dunn(1974), Wahl and Kronmal(1977) 등이 모의실험에 의해 판별분석법의 오차율(misclassification error rate)을 조사하였다. 그들의 연구결과에 의하면, 당연히 기대했던 대로, 두집단의 공분산행렬의 차이가 심하면 심할수록 선형판별함수의 오차율은 다른 판별분석법보다 증가한다는 사실이 알려져 있다.

본 논문의 주제인 그래프에 의한 판별분석법에서는 연습표본을 가능한 한 잘 분리시킬 수 있는 방법을 집중적으로 모색하여 왔다. 왜냐하면 그래프에 의한 방법은 직관적인 방법으로서 자료분석자(data analyst)에게 자료의 본질을 알려주고자 하는 것이 궁극적인 목적이기 때문이다. 이 분야의 선두 주자는 Sammon(1970)으로서 두 그룹의 경우 일차원 수직선에만 표시되는 선형판별함수를 이차원 평면으로 확장할 수 있는 방법을 제안하였으며 이차원 평면에 표시된 연습표본을 Sammon 그래프라고 한다. Chien(1978)에 의하면 Sammon 그래프는 그리기 쉽고 간단하게 표시되기 때문에 패턴인식 분야에서는 매우 널리 이용되고 있다고 한다. Chang(1987)은 두 공분산행렬이 같다고 가정할 수 없는 경우 Sammon 그래프의 대안으로서 MV 그래프라고 불리는 새로운 방법을 제안하였다.

본 논문에서는 Sammon 그래프와 MV 그래프의 특징을 살펴보고 이를 바탕으로 새로운 대안의 가능성을 제시하고자 하며 극히 제한적이긴 하지만 3가지 그래프의 특성을 비교·분석해 보고자 한다. Sammon 그래프, MV 그래프 및 새로운 대안의 유도과정은 제2장에 요약되어 있으며 제3장에서는 실제자료 및 모의실험을 이용하여 3가지 그래프를 비교해 보고자 한다. 제4장에서는 결론 및 앞으로의 연구과제를 요약하고 있다.

2. 그래프에 의한 판별분석법

g 개 그룹에서 추출한 p 차원 연습표본을 x_{ij} (i = 1, 2, ⋯, g ; j = 1, 2, ⋯, n_i) 라고 하자. 각 그룹에 대하여 사전확률을 π_i, 평균벡타를 μ_i, 공분산행렬을 Σ_i 라고 표시하자. Σ_i 와 μ_i 에 대한 표본추정량을 각각 $S_i = \sum_j (x_{ij} - \overline{x_i})(x_{ij} - \overline{x_i})'/(n_i - 1)$, $\overline{x_i} = \sum_j x_{ij}/n_i$ 라고 하자. 만약 $\Sigma_1 = \Sigma_2 = \Sigma$ 인 경우 Σ 에 대한 표본추정량은 $S_p = \{(n_1-1)S_1+(n_2-1)S_2\}/(n_1+n_2-2)$ 라고 하자. 또한 p차원 공간에 주어진 연습표본 x_{ij}를 i와 j에 대하여 구별하지 않을 때는 첨자를 생략하고 x라고 표시하기로 하자. g = 2인 경우, Fisher(1936)의 선형판별함수는 두 공분산행렬이 같다는 가정하에서 식 (1)을 최대로 하는 투영(projection) l을 찾자는 것이며 그 결과는 수직선에 표시된다.

$$(l'd)^2/l'S_pl, \tag{1}$$

여기서 $d = \overline{x_1} - \overline{x_2}$이며 식 (1)을 최대로 하는 l은 $l_1=S_p^{-1}d$이다. 따라서 연습표본은 새로운 축 $Z_1=l_1'x$에 표시된다. $g \geq 3$인 경우 선형판별함수는 공분산행렬이 모두 같다는 가정하에서 식 (2)를 최대로 하는 투영 l을 찾고자 한다.

$$l' Bl / l' Wl, \tag{2}$$

여기서 $\bar{x} = \sum_i \sum_j x_{ij} / (\sum_i n_i)$, $B = \sum_i (\bar{x_i} - \bar{x})(\bar{x_i} - \bar{x})'$, $W = \sum_i (n_i - 1) S_i$.

g = 2인 경우 두 공분산행렬이 서로 다르다면 식 (3)에 정의된 이차판별함수 $Q(x)$를 생각하게 되며 어떤 상수 c에 대하여 $Q(x) \le c$일 경우 그룹 1로 판별하게 된다.

$$Q(x) = G(x, \bar{x_1}; S_1) - G(x, \bar{x_2}; S_2) , \tag{3}$$

여기서 $G(x, \bar{x_i}; S_i) = (x - \bar{x_i})' S_i^{-1} (x - \bar{x_i})$ 이다. Seber(1984, chap. 6)에 의하면 $Q(x) \le c$ 일 때 그룹 1로 판별하는 이차판별함수는 다변량정규분포라는 가정하에서 여러 특징을 갖는다는 것이 알려져 있다. 첫째 $c = 2\ln(\Pi_1/\Pi_2) - \ln(|S_1|/|S_2|)$이면 이는 오차율의 기대값(expected probability of misclassification)을 최소화하는 판별함수이며, 둘째 $c = 0$이면 우도비 방법(likelihood ratio method)에 의한 판별함수라는 것이다.

2.1 Sammon 그래프

Sammon(1970)은 g = 2인 경우 식 (2)는 식 (1)에 상수를 곱해준 형태로 표시된다는 사실에 착안하여 g = 2인 경우에도 식 (1) 대신에 식 (2)를 최대로 하는 l을 찾고자 하였다. 그 이유는 식 (1)을 최대로 할 경우 오직 한 방향으로의 투영만 가능하지만 식 (2)를 최대로 할 경우 적절한 제약조건 하에서 두 방향 이상으로의 투영도 가능하기 때문이다. 다시 말하면 식 (1)을 최대로 할 경우에는 연습표본은 수직선에 표시되나 식 (2)를 최대로 할 경우에는 연습표본은 이차원 평면 또는 삼차원 공간에 표시될 수 있기 때문이다.

Seber(1984, p526)에 의하면 식 (2)를 최대로 하는 l은 고유방정식 (4)를 만족한다. 여기서 행렬 B의 순위(rank)는 1이므로 고유방정식 (4)의 0이 아닌 고유값은 오직 1개만 존재하며 그 고유벡타를 l_1이라고 하면 첫번째 투영은 l_1에 의해 이루어 진다.

$$(B - \lambda W) l = 0. \tag{4}$$

Sammon(1970)은 두번째 투영을 위하여 l_1과 직교(orthogonal)한다는 제약조건 하에서 식 (2)를 최대로 하는 투영 l_2를 생각하였으며 이를 라그랑지 승수를 이용하여 구하면 아래와 같다.

$$l_2 = r\{W^{-1} - (d' W^{-2} d / d' W^{-3} d) W^{-2}\} d,$$

여기서 r은 표준화상수(normalizing constant)이다. 따라서 Sammon 그래프는 연습표본을 두 축 $Z_1 = l_1' x$ 와 $Z_2 = l_2' x$ 에 의해 (Z_1, Z_2) 평면에 표시한 다음 외관오차율(apparent error rate)을 최소화하는 판별선을 시각적으로 찾자는 것이다.

2.2 MV 그래프

g = 2인 경우 Sammon 그래프는 판별공간을 수직선에서 이차원으로 확장시킨 장점이 있지만 이미 첫번째 축이 평균의 차이를 반영하였기 때문에 두번째 축이 과연 무엇을 의도하는지 의심스럽다. 따라서 Chang(1987)은 Sammon 그래프의 기본가정인 두 공분산행렬이 같다는 가정을 하지 않는 대신에 두번째 축이 두 공분산행렬의 차이를 반영하도록 하는 MV 그래프를 제안하였다. 즉 MV 그래프는 연습표본을 (Z_1, Z_2) 평면에 표시하되 Z_1은 두 평균의 차이를 반영하고 Z_2는 두 공분산행렬의 차이를 반영하자는 것이다.

우선 $y = x - \bar{x_2}$라고 하면, 그룹 1에서 y의 평균은 d이고 그룹 2에서 y의 평균은 0이다. 따라서 그룹 1에서 y와 그의 평균과의 차이는 Mahalanobis 제곱거리(sample Mahalanobis

squared distance)인 $(y-d)' S_1^{-1} (y-d)$로 측정되며 마찬가지로 그룹 2에서는 $y' S_2^{-1} y$로 측정된다. 그런데

$$(y-d)' S_1^{-1}(y-d) = \sup\{l' (y-d)\}^2/l' S_1 l ,$$
$$y' S_2^{-1} y = \sup(l' y)^2/l' S_2 l ,$$

이므로 MV 그래프의 첫번째 축 Z_1은 식 (5)와 같이 정의된다.

$$Z_1 = \{l_c' (y-d)\}^2/l_c' S_1 l_c - (l_c' y)^2/l_c' S_2 l_c . \quad (5)$$

식 (5)에서 l_c을 어떻게 정의할 지에 대하여는 많은 논란이 있을 수 있다. Chang(1987)은 0과 1사이의 값을 갖는 p에 대하여 $S = pS_1+(1-p)S_2$라고 하고 $l_c=S^{-1}d$ 로 정의하였다. 그 이유는 $S_1 = S_2$인 경우 이렇게 정의된 Z_1은 선형판별함수의 선형변환으로 표시될 수 있기 때문이라고 생각되며 보통의 경우 $p = n_1/(n_1+n_2)$가 이용된다.

MV 그래프의 두번째 축 Z_2는 두 공분산행렬의 차이를 반영하도록 정의된다. 즉 평균의 차이인 d와 직교하는 $p-1$개의 $p\times 1$ 벡타 $h_2, h_3, \cdots, h_p$를 생각하고 이들로 구성된 $p\times(p-1)$ 행렬을 H라고 하자. y를 H에 의해 선형변환 시킨 $H'y$ 의 평균은 그룹 1과 2의 경우 모두 0이므로 두번째 축 Z_2는 그룹 1에서 $H'y$ 가 평균에 이르는 Mahalanobis 제곱거리와 그룹 2에서 $H'y$ 가 평균에 이르는 Mahalanobis 제곱거리의 차이인 식 (6)과 같이 정의된다.

$$Z_2 = (H'y)' (H' S_1 H)^{-1}(H'y) - (H'y)' (H' S_2 H)^{-1}(H'y) . \quad (6)$$

실제로 행렬 H를 구할 때는 dd'의 순위(rank)는 1이므로 dd'의 고유값이 0이 되는 $p-1$개 고유벡타로서 구성된다.

앞에서 살펴본 MV 그래프의 두 축 Z_1과 Z_2의 유도과정은 다음과 같은 이유로 인하여 자연스럽지 못하다고 생각된다. 첫째, 식 (5)에서 첫번째 축 Z_1을 결정함에 있어 $l_c = S^{-1}d$로 정의하였다. 그러나 $p = (n_1-1)/(n_1+n_2-2)$이면 $S = S_p$이므로 S는 두 공분산행렬이 같은 경우의 불편추정량인 S_p의 일반화 형태로 이해될 수 있으며 이는 두 공분산행렬이 다르다는 가정하에서 출발한 MV 그래프의 목적에 부합하지 않는다고 생각된다. 둘째, 식 (6)에 정의된 두번째 축 Z_2는 S_1과 S_2의 차이를 상수로 표시하고자 하는 직관적인 방법으로서 이러한 목적으로는 다양한 행렬간의 거리(matrix norm)을 고려해 볼 수 있을 것이다. 그러나 Chang(1987)은 S_1과 S_2가 비례할 경우, 이렇게 정의된 $Z_1 + Z_2$는 식 (3)에 정의된 이차판별함수값과 같아짐을 보였으며 따라서 MV 그래프가 이차판별함수인 $Q(x)$를 근사적으로 이차원 평면에 나타낸 것이라고 주장하였다. 한편 MV 그래프의 판별선은 Sammon 그래프의 경우와 마찬가지로 외관오차율을 최소화 하도록 그려진다.

2.3 새로운 대안의 가능성

MV 그래프가 새로운 축 Z_1과 Z_2에 의해 이차판별함수를 근사 시킨 것이라면, 새로운 대안으로서 이차판별함수 자체를 낮은 차원 ($p \le 3$)에서 나타낼 수 있는 방법에 대하여 관심을 갖게 된다. 식 (3)에서 정의된 이차판별함수 $Q(x)$는 Mahalanobis 제곱거리인 $G(x, \overline{x_i}; S_i)$ 의 차이이므로, 동시대각화(simultaneous diagonalization)에 의해 $G(x, \overline{x_i}; S_i)$ 는 김성주(1987, eq. (3))와 같이 표시될 수 있으며 $Q(x)$ 자체가 대각화 될 수 있다.

이를 정리하면, 일반화 고유방정식 (7)을 만족하는 고유값 e 와 고유벡타 t 를 구하고 고유값

들로 대각행렬 E를 구성하고, 고유벡타들로 행렬 T를 구성하고, 행렬 T에 의한 x의 변환을 식 (8)과 같이 정의하자,

$$S_1 t = e(S_2 t)\ , \tag{7}$$

$$z = T'(x - \overline{x_1}). \tag{8}$$

그러면 $Q(x) \le c$는 식 (9)로 변환되며 이를 정리하면 식 (10)을 얻게 된다.

$$z' E^{-1} z - (z-m)'(z-m) \le c\ . \tag{9}$$

$$(z+F^{-1}m)' F(z+F^{-1}m) \le c^*\ , \tag{10}$$

여기서 $F = E^{-1}-I$, $m = T'(\overline{x_2} - \overline{x_1})$, $c^* = c+m'(F^{-1}+I)m$ 이다. 식 (10)은 중심이 $-F^{-1}m$에 있는 대각화된 이차곡면으로서 Fraleigh and Beauregard(1990, section 7.2)에 의하면 대각행렬 F에서 양의 값을 갖는 대각원소의 수에 따라 그 형태가 달라진다. 예를 들어 p=2인 경우 F의 두 대각원소가 모두 양일 경우 식 (10)은 타원이 되며 한 대각원소만 양일 경우 쌍곡선이 된다. 즉 새로운 그래프는 연습표본을 식 (8)에 의해 변환시키고 변환된 공간에 대각화된 이차판별함수인 식 (10)을 표시하자는 것이다.

여기서 두 평균벡타의 신뢰타원은 $G(x, \overline{x_i}; S_i)$ 의 형태로 표시되는 점에 주목하자. 이 두 신뢰타원은 변환된 공간에서 식 (9)에 있는 첫번째 항과 두번째 항의 형태로 각각 대각화 된다. 따라서 새로운 그래프에서는 이차판별함수는 물론이고 각 그룹의 대각화된 신뢰타원도 나타낼 수 있다. 또한 대각화된 신뢰타원의 장축과 단축의 길이를 결정하는 식 (7)의 고유값들은 두 그룹의 공분산행렬이 같다는 가설을 검정하는데 이용될 수 있다는 점에 주목할 필요가 있다. Seber(1984, p105)에 의하면, 이 가설은 합과 곱의 법칙(union-intersection principle)에 의해 식 (7)의 고유값 중 최대값이 아주 크든지 또는 최소값이 아주 작은 경우 기각되기 때문이다.

3. 실증분석

여기서는 g = 2인 경우 앞에서 살펴본 3가지 그래프의 특성을 실제자료 (예제 1)와 모의실험 (예제 2)을 이용하여 비교·분석하고자 한다. 예제 1과 2에서 Sammon 그래프와 MV 그래프의 판별선은 외관오차율을 최소화 하도록 그렸으며 MV 그래프에서 $p = n_1/(n_1+n_2)$라고 하였다. 새로운 대안에서 어떤 상수 c는 우도비방법에 기초하여 0으로 고정하였으며 두 신뢰타원의 신뢰도는 95%로 고정하였다. 또한 외관오차율을 계산할 때 잘못 판별한 도수를 $f_1 + f_2$로 표시하였으며, 여기서 f_1은 그룹 1을 그룹 2로 잘못 판별한 도수이고 f_2는 그룹 2를 그룹 1로 잘못 판별한 도수이다.

예제 1. Smith(1947)는 정상인 25명(그룹 1)과 정신질환자 25명(그룹 2)에 대하여 두 변수를 관측하였으며 그 자료가 그림 1(a)에 산점도로 나타나 있다. 이 자료는 Mardia, Kent and Bibby(1979, p312)가 선형판별함수와 이차판별함수를 비교할 목적으로 다룬 바 있으며 두 판별함수에 의해 잘못 판별된 도수는 각각 0+4, 2+2이다. 이 자료에 대한 Sammon 그래프, MV 그래프 및 새로운 대안은 각각 그림 1의 (b), (c), (d)에 나타나 있다. 그림 1(a)에 의하면 각 변수의 분산은 그룹 1의 경우 그룹 2에 비해 작아 보이며 두 변수간의 상관관계는 두 그룹 모두 약한 것으로 보인다. 실제로 그룹 1과 그룹 2에서 두 변수간의 상관계수는 각각 -0.3, 0.1이다. 그림 1(b)에 있는 Sammon 그래프에서 판별선은 두번째 축 Z_2에 평행하게 나타나기 때문에 이 예에서 Z_2는 두 그룹을 판별하는데 전혀 영향을 미치지 못하고 있으며 선형판별함수와 마찬가지로 잘못 판별된 도수는 0+4이다. 그림 1(c)에 있는 MV 그래프에서 그룹 1은 원점 근방에 밀집해 있는 반면에 그룹 2는 매우 넓게 흩어져 나타나 있으며 잘못 판별된 도수는 1+2이다.

새로운 대안은 그림 1(d)에 나타나 있으며 이차판별함수는 큰 타원으로, 95% 신뢰타원은 각각 원과 작은 타원으로 나타나 있다. 식 (7)의 고유값은 0.25, 0.10이므로 이에 의해 95% 신뢰타원의 장축과 단축의 길이의 비율이 결정되며 이차판별함수는 타원이 된다. 두 고유값 중 최소값인 0.10은 1보다 상당히 작기 때문에 두 공분산행렬이 같다고 하기는 어렵다. 새로운 대안은 이차판별함수를 변환된 공간에서 대각화 시킨 것이므로 이차판별함수와 마찬가지로 잘못 판별된 도수는 2+2이다.

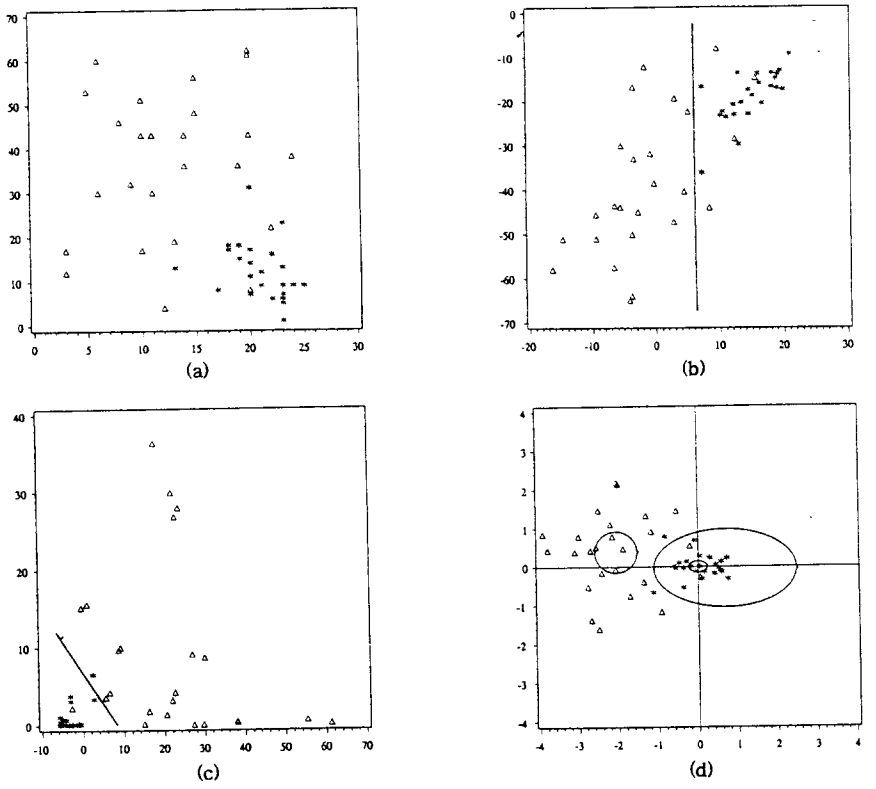

그림 1. Smith(1947) 자료 (그룹 1은 *, 그룹 2는 Δ)에 대한 3가지 그래프 :
(a) 원시자료의 산점도 (b) Sammon그래프
(c) MV그래프 (d) 새로운 대안.

예제 2. 여기서는 아래 두 경우의 모의실험을 이용하여 3가지 그래프의 특성을 비교해 보고자 한다. 모의실험I과 II는 이변량정규모집단을 대상으로 하고 있으며, 모의실험I은 공분산행렬은 같고 평균만 다른 경우이고 모의실험II는 모의실험I에서 공분산행렬을 달리한 경우이다. 우리는 이 예에서 모의실험I과 II를 비교함으로서 공분산행렬이 같은 경우와 같지 않은 경우의 차이를 살펴보고자 한다. 이제 SAS를 이용하여 각 그룹마다 50개씩 관측값을 얻었으며 모의실험I에 대하여 자료의 산점도, Sammon 그래프, MV 그래프 및 새로운 대안은 각각 그림 2의 (a), (b), (c), (d)에 나타나 있으며 마찬가지로 그림 3은 모의실험II에 대하여 보여주고 있다.

모의실험 I : 그룹 1의 $\mu=0$, $\Sigma=I$
그룹 2의 $\mu=[2, 2]'$, $\Sigma=I$
모의실험 II : 그룹 1의 $\mu=0$, $\Sigma=I$
그룹 2의 $\mu=[2, 2]'$, $\Sigma = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$

Chang(1987)에 의해 지적된 바와 같이, 두 공분산행렬이 유사한 경우 Sammon 그래프나 MV 그래프의 판별선은 Z_2축에 평행하게 나타나는 경향이 있으며 이를 그림 2의 (b)와 (c)에서 확인할 수 있다. 따라서 두 그룹의 공분산행렬이 유사한 경우 Sammon 그래프와 MV 그래프의 두번째 축 Z_2는 예상대로 판별분석에 별다른 영향을 미치지 못하고 있다. 모의실험I의 경우 선형판별함수가 최적판별함수일 것으로 기대되며 선형판별함수를 확장한 Sammon 그래프와 선형판별함수를 또다른 측면에서 확장한 MV 그래프에서 잘못 판별한 도수는 모두 0+6이다. 이차판별함수에 기초한 새로운 대안은 그림 2(d)에 나타나 있다. 이 경우 식 (7)의 고유값은 1.85, 0.96이므로 이에 의해 95% 신뢰타원의 장축과 단축의 길이의 비율이 결정되며 이차판별함수는 쌍곡선으로 나타난다. 이 쌍곡선에 의해 잘못 판별된 도수는 2+5로서 그 합은 Sammon 그래프나 MV 그래프의 결과에 근사하다.

한편 모의실험II의 경우 그림 3(b)에 나타난 Sammon 그래프의 판별선은 두번째 축 Z_2에 평행하게 나타나며 잘못 판별된 도수는 3+16이다. 그림 3(c)에 나타난 MV 그래프는 예제 1의 그림 1(c)와 매우 유사하게 나타나며 잘못 판별한 도수는 15+7이다. 새로운 대안은 그림 3(d)에 나타나 있다. 이 경우 식 (7)의 고유값은 0.90, 0.34이므로 이에 의해 95% 신뢰타원의 장축과 단축의 길이의 비율이 결정되며 이차판별함수는 타원이 되며 그 일부가 나타나 있다. 이 타원에 의해 잘못 판별된 도수는 6+13으로서 그 합은 Sammon 그래프의 결과와 동일하고 MV 그래프 보다는 약간 나은 결과이다.

비록 두번 시행한 모의실험 결과이지만, 모의실험I과 II를 비교함으로서 다음 사항을 확인할 수 있었다. 첫째, MV 그래프는 의도한 대로 두 공분산행렬이 유사할 경우 그 결과는 선형판별함수에 근사하지만 두 공분산행렬이 서로 다를 경우 그 결과는 이차판별함수에 비해 나쁘다는 것이다. 둘째, 모의실험II에서 선형판별함수를 확장한 Sammon 그래프가 최적 판별함수일 것으로 기대되는 이차판별함수와 유사한 결과를 갖는 이유는 이미 알려진 대로 선형판별함수는 어느 정도 공분산행렬의 차이에 대해 로버스트하다는 사실로 설명될 수 있을 것이다. 셋째, 공분산행렬이 같은 경우나 다른 경우나 Sammon 그래프의 두번째 축의 효용성에 대해 많은 의심을 갖게 된다.

4. 결론 및 앞으로의 과제

Sammon 그래프는 일차원을 이차원으로 확장시킨 장점이 있지만 이론적으로나 실증적으로 두번째 축 Z_2의 효용성은 의심스럽다. 또한 MV 그래프는 Sammon 그래프와 마찬가지로 시각적 판단에 의해 판별하게 되므로 자료분석자의 주관이 개입될 수 있으며 이는 과학적 자료분석에서는 지양되어야 할 것으로 생각된다. 반면에 새로운 대안은 4차원 이상 자료에는 적용될 수 없다는 결정적인 단점을 갖고 있지만 새로운 대안의 판별곡선은 이차판별함수를 변환된 공간에다 그대로 표시한 것이므로 Sammon 그래프나 MV 그래프에 비해 매우 객관적이라는 장점이 있다. 실제 자료분석에 있어서 두 그룹의 평균의 차이가 작으면 작을수록 그리고 각 변수의 분산이 크면 클수록, Sammon 그래프와 MV 그래프의 판별선을 그리는 것은 매우 어려워지나 새로운 대안에서는 언제나 명확한 판별곡선을 그릴 수 있다.

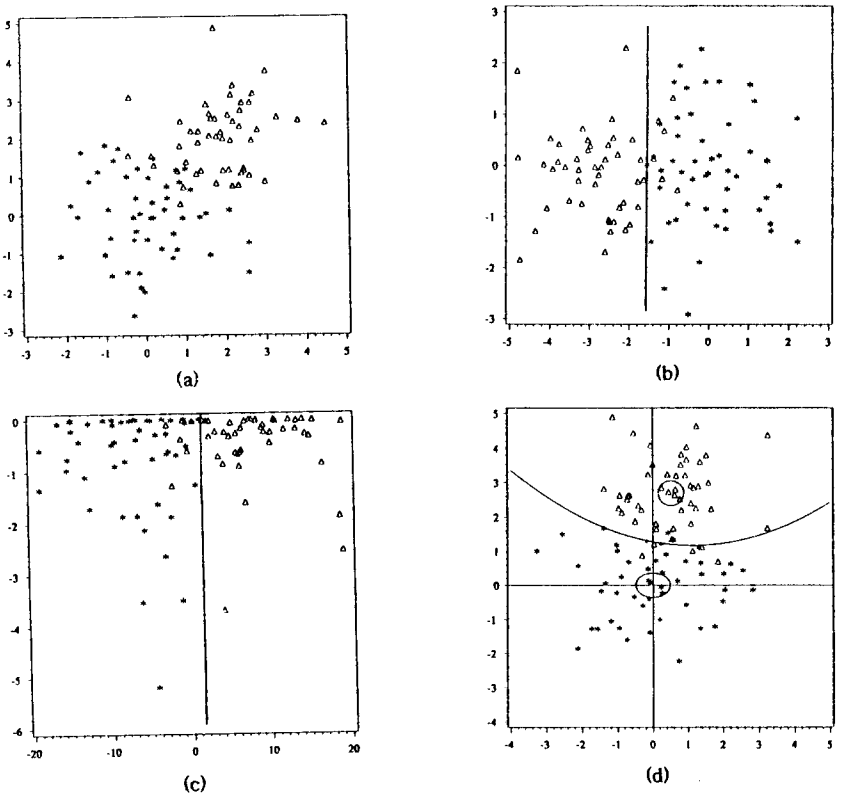

그림 2. 모의실험I (그룹 1은 *, 그룹 2는 Δ)에 대한 3가지 그래프 :
(a) 원시자료의 산점도 (b) Sammon그래프
(c) MV그래프 (d) 새로운 대안.

앞에서 살펴본 실증분석에서 $p = 2$인 경우 이차원 평면에 나타난 3가지 그래프의 특성을 다루었으나 다음과 같은 이유로 인하여 본 논문에서는 $p \geq 3$인 경우는 다루지 못하였다. 첫째, Sammon 그래프는 이차원 평면 또는 삼차원 공간에 표시될 수 있으나 MV 그래프는 이차원 평면에만 표시된다. 새로운 대안은 이차원 평면 또는 삼차원 공간에 표시될 수 있으나 표시되는 공간의 차원은 p의 값과 일치하여야 한다. 따라서 3가지 그래프를 공평하게 비교하기 위해서는 표시되는 공간은 이차원 평면이 적절한 것으로 생각된다. 둘째, 지금으로서는 새로운 대안에 대하여 차원축소 (dimensionality reduction)가 가능하지 않기 때문에 표시되는 공간이 이차원일 경우 새로운 대안은 $p = 2$일 경우만 표시될 수 있다. 따라서 새로운 대안의 차원축소 문제에 대하여 앞으로 계속 연구할 예정이다.

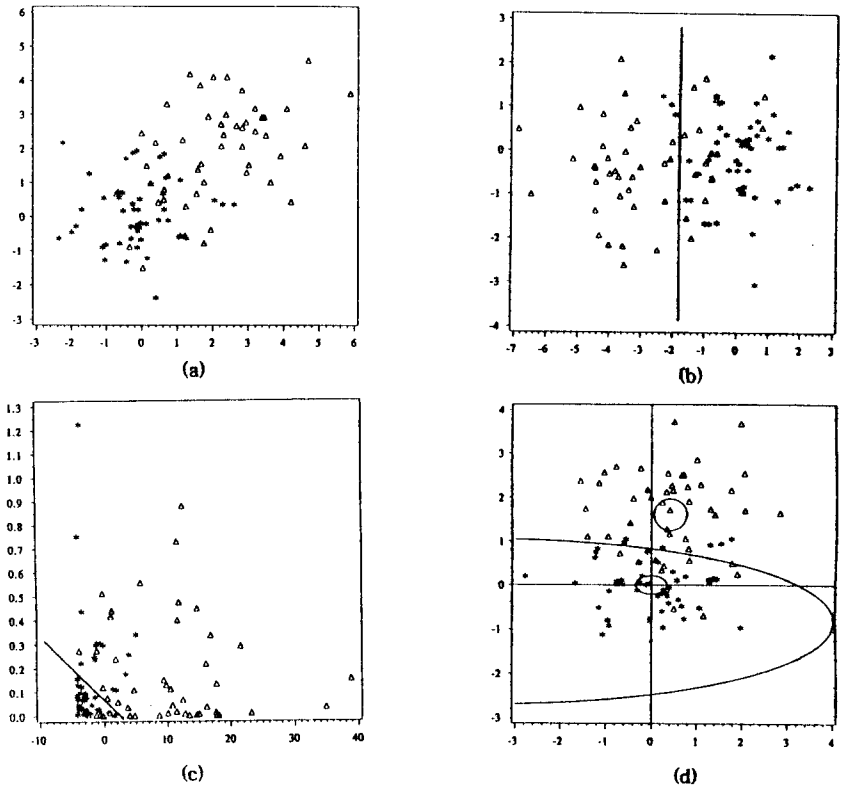

그림 3. 모의실험II (그룹 1은 *, 그룹 2는 Δ)에 대한 3가지 그래프 :
(a) 원시자료의 산점도 (b) Sammon그래프
(c) MV그래프 (d) 새로운 대안.

본 논문에서는 그룹의 수가 $g = 2$인 경우만 다루었으나 3가지 그래프 모두 $g \geq 3$인 경우에도 적용될 수 있다. 왜냐하면 $g \geq 3$인 경우 두 그룹에 대한 그래프를 그리고 이 그래프들을 산점도행렬(scatterplot matrix)로 나타낼 수 있기 때문이다. 본 논문에서 다룬 3가지 그래프는 모두 모수적 방법에 기초하고 있으나 비모수적 방법에 기초한 판별분석법을 그래프로 나타내보는 것은 매우 흥미있는 과제라고 생각된다. 이에는 Broffitt (1982)가 다룬 순위(rank)에 의한 방법을 비롯하여 요즈음 활발히 연구되고 있는 투영추구방법 (projection pursuit method)등이 연구대상이 될 것으로 생각된다.

참 고 문 헌

[1] 김 성주(1987), "On The Condition That Two Hyper-ellipsoids Have No Points in Common," 통계학연구 제16권 제1호, 45-51.

[2] 김 혜중(1992), "A Variable Selection in Heteroscedastic Discriminant Analysis : General Predictive Discriminant Case," 통계학연구 제21권 제1호, 1-13.

[3] 안 윤기, 이 성석(1992), "투사지향방법에 의한 판별분석의 모의실험분석," 응용통계연구 제5권 제1호, 103-111.

[4] Anderson, T. W., and Bahadur, R. R. (1962), "Classification Into Two Multivariate Normal Distributions With Different Covariance Matrices," *The Annals of Mathematical Statistics*, **33**, 420-431.

[5] Broffitt, J. D.(1982), "Nonparametric Classification," in *Handbook of Statistics*, Vol. 2, P. R. Krishnaiah and P. K. Sen (Editors), North-Holland, Amsterdam, 139-168.

[6] Chang, W. C. (1987), "A Graph for Two Training Samples in a Discriminant Analysis," *Applied Statistics*, **36**, 82-91.

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

[8] Efron, B. (1975), "The Efficiency of Logistic Regression Compared to Normal Discriminant Analysis," *Journal of the American Statistical Association*, **70**, 892-898.

[9] Fisher, R. A. (1936), "The Use of Multiple Measurements in Taxonomic Problem," *Annals of Eugenics*, **7** (part 2), 179-188.

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

[11] Mardia, K. V., Kent, J. T., and J. M. Bibby (1979), *Multivariate Analysis*, Academic Press, New York.

[12] Marks, S., and Dunn, O. J. (1974), "Discriminant Functions When Covariance Matrices are Unequal," *Journal of the American Statistical Association*, **69**, 555-559.

[13] Sammon, J. W., Jr. (1970), "An Optimal Discriminant Plane," *IEEE Transactions on Computers*, **C-19**, 826-829.

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

[15] Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*, Chanpman and Hall, New York.

[16] Smith, C. A. B. (1947), "Some Examples of Discrimination," *Annals of Eugenics*, **13**, 272-282.

[17] Wahl, P. W., and Kronmal, R. A. (1977), "Discriminant Functions When Covariances are Unequal and Sample Sizes are Moderate," *Biometrics*, **33**, 479-484.

A Graphical Method for Discriminant Analysis When Covariance Matrices Are Unequal[1)]

Seong-Ju Kim[2)], Kab-Do Chung[3)]

Abstract

This paper concerns graphical methods for discriminant analysis. We discuss Sammon's graph, MV graph and possibility of an alternative. The properties of the three graphs are investigated using real data and simulation studies. Dimensionality reduction for an alternative and robust procedure are discussed.

1) This paper was supported by NON DIRECTED RESEARCH FUND, Korea Research Foundation, 1991

2) Department of Statistics, Sung Kyun Kwan University, Chongro-Ku, Seoul, 110-745 Korea

3) Department of Biostatistics, Catholic University Medical College, Socho-Ku, Seoul, 137-701, Korea