응용통계연구
제 5 권 제 1 호, 1992년

투사지향방법에 의한 판별분석의 모의실험분석†

안윤기* · 이성석**

요 약

다변량 통계분석기법중 하나로 제기된 투사지향방법은 다변량자료를 관심있는 일차원 또는 이차원의 자료로의 선형투사를 찾아 나가는 방법이다. 이 방법은 다변량자료가 갖는 차원의 문제를 해결해 줄 수 있는 유용한 기법으로 제시되었다.

본 연구에서는 투사지향방법을 이용하여 추정한 다변량 확률밀도함수를 사용한 새로운 비모수적인 판별분석방법을 제시하고, 이를 기존의 모수적 판별분석방법중 실제적으로 많이 사용되는 선형판별함수방법, 그리고 기존의 비모수적 판별분석방법중 계산상의 편리성이 많은 K-최인접방법과 컴퓨터 시뮬레이션을 통하여 비교분석하였다.

1. 서 론

판별분석방법은 두가지 접근방법, 즉 모수적(parametric)인 방법과 비모수적(nonparametric or distribution free)인 방법으로 분류할 수 있다(Broffitt, 1982). 전자중에서 공통공분산을 갖는다는 가정하에서 도출된 선형판별함수(linear discriminant function : LDF)와 상이한 공분산을 갖는 다변량 정규분포 가정하에서 도출된 2차형판별함수(quadratic discriminant function : QDF)방법이 실용적으로 널리 이용되는 방법들이다. 그러나 이들 방법들은 다변량 정규분포가정이 위배되거나 공분산이 다른 경우 판별성과가 나빠지는 문제점이 있는 것으로 알려졌다(Krzanowski, 1977 ; Van Ness, 1979). 비모수적인 방법들로는 커널방법(kernel method)을 사용한 분포추정방법(Hand, 1982), K-최인접방법(k-nearest-neighbor method : KNN ; Cover, 1968 ; Anderson and Benning, 1970), 순위방법(rank method : Broffitt, 1982) 및 비모수적 회귀분석방법(Efron, 1975)등 수없이 많이 있지만, 이 방법들을 다변량으로 확장할 때는 차원문제(curse of dimensionality)가 심각하게 되고, 실제로 적용하는 데는 커다란 장애가 된다(Silverman, 1986).

Friedman and Tukey(1974)는 Kruskal(1972)이 제안한 투사지향방법(projection pursuit me-

† 이 논문은 1990년도 문교부 지원 학술진흥재단의 자유공모과제 학술연구조성비에 의하여 연구되었음.
* (120-749) 서울특별시 서대문구 신촌동 134 연세대학교 응용통계학과 교수
** (360-742) 충청북도 청주시 모충동 231 서원대학 응용통계학과 조교수

thod : PP방법)을 사용하여 다변량통계자료들을 분석하는 방법을 제시하였는 바, 이러한 투사지향방법의 장점으로는 기존의 단일변량에 대한 비모수적 확률밀도함수 추정방법(nonparametric density estimation method)들을 다변량으로 확장시키는 데 있어서 커다란 난점이 되는 차원문제를 용이하게 다룰 수 있도록 해주며, 자료가 갖는 비선형구조를 잘 반영해 줄 수 있고, 기존의 모수적 방법과 비모수적 방법들을 위 방법의 특수한 경우로 취급할 수 있다는 점을 들 수 있다(Jones and Sibon, 1987). 투사지향방법을 이용한 비모수적인 판별분석방법은 근본적으로 다변량 확률밀도함수를 추정하는 방법이 가장 주가 되며, 이는 안윤기, 이성석(1990)을 참조하면 된다.

본 연구에서는 투사지향방법을 이용하여 추정한 다변량 확률밀도함수를 사용한 판별분석방법들과 기존의 방법인 LDF와 KNN 방법을 재투입법(resubstitution method)에 의한 명목오류백분율(apparent error percent rate)을 기준으로 컴퓨터 시뮬레이션을 통하여 비교하여 보고자 한다.

2. 모의실험에 의한 비교분석

2.1 비교분석방법

주어진 판별분석방법의 성과를 비교평가하기 위해서는 비교의 기준이 설정되어야 한다. 한가지 중요한 기준으로서 흔히 판별분석의 결과인 오류율을 들 수 있다. 이 오류율은 여러가지로 다르게 정의될 수 있으며, 이 중에서 중요한 것을 살펴보면 다음과 같다(McLachlan, 1989 ; Seber, 1984).

1) 최적 또는 베이즈오류율(optimum or Bayes' error rate)은 사전밀도함수와 조건밀도함수가 완전하게 알려진 경우, 이를 이용하여 베이즈 최적판별규칙을 사용할 때 발생되는 오류확률로서 표본을 무한히 추출할 경우 발생되는 오류율로 생각할 수 있다.

$e_{i,opt}$ = Pr(베어즈 최적판별규칙을 사용할 때 i 집단의 개체를 다른 집단의 개체로 판별할 확률)

$$e_{opt} = q_1 \cdot e_{1,opt} + q_2 \cdot e_{2,opt}$$

여기서 q_i, $i = 1,2$는 집단 i에 속할 확률이다.

2) 실제오류율(actual error rate)은 사전 확률밀도함수와 조건 확률밀도함수가 완전하게 알려진 경우, 이를 이용하여 베이즈 판별규칙이외의 판별규칙을 사용할 때 하나의 표본에서 실제로 발생하는 오류율이다.

$e_{i,act}$ = Pr(어떤 판별규칙을 사용할 때 i 집단의 개체를 다른 집단의 개체로 판별할 확률)

$$e_{act} = q_1 \cdot e_{1,act} + q_2 \cdot e_{2,act}$$

3) 기대 실제오류율(expected actual error rate)은 표본을 무한히 추출할 경우의 실제오류율로

생각할 수 있다.

$$e_{i,exp} = E(e_{i,act})$$
$$e_{exp} = q_1 \cdot e_{1,exp} + q_2 \cdot e_{2,exp}$$

이 중에서 우리가 주로 관심을 쏟는 것은 기대 실제오류율이다. 이를 추정하는 방법은 다음과 같이 여러가지 방법이 있다. 이들의 구분은 하나의 표본이 주어진 경우 판별규칙의 추정에 사용하는 표본(analysis sample)과 추정된 판별규칙을 사용하여 오류율을 추정하려는 표본(training sample)을 어떻게 사용하느냐에 따라 다음과 같이 구분된다(Fukunaga and Kessell, 1971, 1973 ; Gong, 1986 ; Penrod and Wagner, 1979).

1) 재투입방법(resubstitution method)
2) 교차타당성(cross-validation)방법 또는 한개제거(leave-one-out)방법
3) 잭나이프방법(jack-knife method)
4) 붓스트랩방법(bootstrap method)

위의 실제오류율을 추정하는 방법들 중에서 2), 3), 4) 방법은 판별규칙의 추정단계가 표본자료수 또는 그 이상의 횟수만큼씩 반복해야되는 계산상의 단점을 갖고 있다. 따라서 본 연구에서는 1)방법을 이용하여 명목오류백분율(apparent error percent rate)을 구하고 이 명목오류백분율의 편의(bias), 즉 다음과 같이 명목오류백분율과 모의실험에서 사전에 계획한 최적오류백분율(%)과의 차이를 구하여 이를 명목오류백분율 편의(bias of apparent error percent rate : %)라 정의하고 이를 비교분석의 기준으로 삼았다. 즉,

$$\text{명목오류백분율 편의} = \text{명목오류백분율} - \text{최적오류백분율}$$

만일 이 명목오류백분율 편의의 부호가 정(+)이면 명목오류백분율은 최적오류백분율을 과대추정한 것을 의미하고 부(-)이면 과소추정한 것을 의미한다.

2.2 모의실험

본 연구에서는 판별집단이 2개인 경우로 국한하였다. 비교하려는 세가지 판별분석방법은 모수적 판별분석방법중 대표적인 것으로 LDF방법, 비모수적 판별방법중 추정이 용이한 KNN방법(K=1), 그리고 본 논문에서 제기하고 있는 투사지향판별분석방법(PP방법) 등이다. 모의실험의 요인으로는 변수의 수(p), 표본의 수($n_1 = n_2$), 평균벡터간의 거리(Δ), 상관관계의 유무(I 또는 Σ), 분포의 형태(단봉 또는 혼합분포) 등이다. 변수의 수는 $p = 5$ 또는 10일 때로 나누고, 표본의 갯수는 변수가 $p = 5$일 때는 $(n_1, n_2) = (10, 10), (15, 15), (30, 30)$로 세가지 경우이고, 변수가 $p = 10$일 때는 $(n_1, n_2) = (15, 15), (30, 30)$로 두가지 경우로 정하였다. 또한 두 집단의 평균벡터(Δ)간의 거리는 최적오류백분율이 각각 30%, 20%, 그리고 10%에 해당되는 거리로 놓았다. 이상의 상황들은 첫째 두 집단의 공분산행렬이 모두 단위행렬(I)인 경우, 둘째 상관계수행렬(Σ)인 경우, 그리고 셋째는 아래와

같은 혼합분포인 경우로 나뉜다.

첫번째 상황 즉, 공분산행렬이 단위행렬(I)인 경우에 두집단의 평균벡터간의 거리는 $\Delta = (\delta_i, 0, \cdots, 0)$, $i = 1, 2, 3$로 첫번째 변수의 평균만이 다르고 나머지는 0으로 모두 같게 놓았다. 여기서 δ_i, $i = 1, 2, 3$는 각각 1.049, 1.683 그리고 2.563이다.

두번째 상황 즉, 공분산행렬이 상관계수행렬(Σ)인 경우에 두집단의 평균벡터간의 거리는 $\Delta = (\triangle_i, \triangle_i, \cdots, \triangle_i)$, $i = 1, 2, 3$로 놓았고 $\triangle_i$는 $p = 5$일때는 각각 0.75644466, 1.2136286과 1.8482056로 놓고, $p = 10$일 때는 각각 0.71146042, 1.1414661 그리고 1.7383112로 놓는다. 여기서 상관계수행렬(Σ)은 모든 변수들간의 상관관계 $\rho_{ij} = 0.5$, $i \neq j$인 경우로 설정하였다.

세번째 상황에서는 1집단은 공분산행렬이 단위행렬(I)이고, 평균벡터는 각각 $\Delta_1 = (\delta_i, \delta_i, 0, \cdots, 0)$와 $\Delta_2 = (-\delta_i, -\delta_i, 0, \cdots, 0)$인 다변량정규분포의 혼합분포이고, 2집단은 공분산행렬이 단위행렬(I)이고, 평균벡터는 각각 $\Delta_3 = (-\delta_i, \delta_i, 0, \cdots, 0)$와 $\Delta_4 = (\delta_i, -\delta_i, 0, \cdots, 0)$인 다변량정규분포의 혼합분포이다. 여기서 δ_i, $i = 1, 2, 3$은 각각 0.5245, 0.8415, 그리고 1.2825이다.

이와 같은 45개의 모의실험계획하에서 각각에 대하여 모의실험을 이용한 표본추출을 100번씩 생성한 다음 이들에 대하여 판별분석을 수행한 후 명목오류율을 구하여 평균한 값을 판별성과로 삼았다. 한편 이와 같은 다변량확률변수값을 생성하기 위해서 IMSL부프로그램을 이용하였다. 이상의 모의실험 결과는 세가지 상황과 변수의 수($p = 5$, 10)의 조합으로 나누어 표 1부터 표 6으로 요약하였다. 각각의 표는 표본의 수, 두집단의 평균벡터간의 거리 그리고 판별분석방법(KNN, LDF, PP방법)에 따른 명목오류백분율과 명목오류백분율편의를 나타내고 있다.

표 1에서 살펴보면 PP방법은 KNN방법보다 편의(bias)가 적게 나타나고 있으며 특히 변수의 갯수가 적고 상대적으로 표본크기가 큰 경우에는 LDF방법과 거의 유사해지는 것을 볼 수 있다. 예를 들면 $p = 5$, $n = 30$인 경우 PP방법의 명목오류백분율 편의는 −2.68로 LDF방법의 −1.00보다 크지만 KNN방법에 비해 상대적으로 성과가 좋은 것으로 나타났다.

표 2에서는 KNN방법의 명목오류백분율 편의는 약 12%～16%에 고르게 나타나서 집단간의 평균벡터의 거리와는 무관한 것으로 생각되나, PP방법과 LDF방법은 집단간의 평균벡터의 거리뿐만 아니라 표본크기에 따라서 크게 지배되는 것을 볼 수 있다. PP방법은 집단간의 평균벡터의 거리가 가까운 경우에는 명목오류백분율 편의가 매우 심각하지만 거리가 먼 경우에는 적어도 KNN방법보다 명목오류백분율 편의가 적게 나타났다.

표 3에서는 변수간에 상관관계가 존재하는 경우에 PP방법은 LDF방법과 유사한 경향을 보이고 있으나 KNN방법은 상관관계가 없는 단위행렬공분산인 경우보다 편의가 약간 감소됨을 알 수 있다.

표 4에서 보듯이 변수간의 상관관계가 존재하고 변수의 수가 많은 경우에는 KNN방법에 의한 명목오류백분율의 편의가 상당히 감소되고 있으나 PP방법과 LDF방법은 표 2와 비교했을 때 거의 변화되지 않는 것을 알 수 있다.

표 5과 표 6에서 보듯이 KNN방법은 비교적 두집단간 평균벡터의 거리에 상관없이 일정한 명목오류백분율 편의를 수반하고 있으며 LDF는 두집단간 평균벡터의 거리가 떨어질수록 명목오류백분율 편의가 급격하게 증가되고 있는 반면 PP방법은 오히려 급격하게 감소되고 있다. 또한 PP방법은 KNN방법에 대하여 일관되게 명목오류백분율 편의가 적은 것을 볼 수 있고 변수의 수가

p = 5이고 두집단간 평균벡터의 거리가 떨어진 경우에는 명목오류백분율 편의가 특히 낮은 것을 알 수 있다.

이상의 모의실험에서 얻은 결과를 요약해보면 다음과 같다.

(1) 각 집단의 공분산행렬이 같은 다변량 정규분포를 이루는 경우는 LDF방법이 명목오류백분율 편의가 가장 낮은 것으로 나타났다. 그러나 PP방법도 그다지 나쁘지 않은 것으로 나타났다.

(2) 변수간에 상관관계가 존재하는 경우 KNN방법은 두집단간 평균벡터의 거리가 떨어짐에 따라 명목오류백분율 편의가 낮아지는 반면 LDF나 PP방법은 거의 같은 양상을 보이는 것으로 나타났다.

(3) 명목오류백분율 편의는 세가지 방법 모두가 표본크기, 변수의 수, 그리고 두집단간 평균벡터의 거리와 관련이 있으나 LDF나 PP방법에서는 변수간 상관관계의 존재와는 무관한 것으로 나타났다.

표 1. 변수의 수는 5이고 공분산행렬이 단위행렬인 경우

평균간 거리 (Δ)	두집단 표본수 (n_1, n_2)	명목오류백분율			명목오류백분율편의		
		KNN	LDF	PP	KNN	LDF	PP
1.049	10, 10	44.42	19.50	9.15	14.42	-10.50	-20.85
1.049	15, 15	45.20	24.60	15.63	15.20	-5.40	-14.37
1.049	30, 30	41.10	26.95	20.40	11.10	-3.05	-9.60
1.683	10, 10	36.15	13.85	7.70	16.15	-6.15	-12.30
1.683	15, 15	34.86	15.50	10.60	14.86	-4.50	-9.40
1.683	30, 30	31.06	17.82	14.16	11.06	-2.18	-5.84
2.583	10, 10	20.50	5.20	2.15	10.50	-4.80	-7.85
2.583	15, 15	19.13	7.60	5.26	9.13	-2.40	-4.74
2.583	30, 30	17.92	9.00	7.32	7.92	-1.00	-2.68

표 2. 변수의 수는 10이고 공분산행렬이 단위행렬인 경우

평균간 거리 (Δ)	두집단 표본수 (n_1, n_2)	명목오류백분율			명목오류백분율편의		
		KNN	LDF	PP	KNN	LDF	PP
1.049	15, 15	44.60	17.97	2.56	14.60	-12.03	-27.44
1.049	30, 30	42.83	24.32	10.57	12.83	-5.68	-19.43
1.683	15, 15	35.86	10.40	1.76	15.86	-9.60	-18.24
1.683	30, 30	35.02	15.23	7.78	15.02	-4.77	-12.22
2.583	15, 15	22.70	4.13	0.93	12.70	-5.87	-9.07
2.583	30, 30	22.25	7.35	4.20	12.25	-2.65	-5.80

(4) 혼합다변량분포인 경우에는 PP방법에 의한 명목오류백분율 편의는 두집단간 평균벡터의 거리가 멀어질수록 그리고 표본크기가 커질수록 매우 적어지는 것으로 나타났다.

본 연구에서 제시된 방법들을 일차원이 아닌 이차원으로 투사시키는 방법으로 확장시키는 문제와 투사지향방법을 사용한 비모수적 회귀방법의 형태를 적용하여 새로운 판별분석방법을 개발하는 것은 앞으로 본 연구를 확장시키는 좋은 연구가 될 것이다.

표 3. 변수의 수는 5이고 공분산행렬이 상관계수행렬인 경우

평균간 거리 (Δ)	두집단 표본수 (n_1, n_2)	명목오류백분율			명목오류백분율편의		
		KNN	LDF	PP	KNN	LDF	PP
1.049	10, 10	42.65	20.60	10.20	12.65	-9.40	-19.80
1.049	15, 15	40.03	24.00	16.00	10.03	-6.00	-14.00
1.049	30, 30	38.16	26.06	20.56	8.16	-3.94	-9.44
1.683	10, 10	30.05	13.60	6.10	10.05	-6.40	-13.90
1.683	15, 15	30.53	15.46	10.60	10.53	-4.54	-9.40
1.683	30, 30	29.23	17.46	14.25	9.23	-2.54	-5.75
2.583	10, 10	17.95	7.30	3.55	7.95	-2.70	-6.45
2.583	15, 15	17.80	8.30	5.26	7.80	-1.70	-4.74
2.583	30, 30	15.66	8.20	6.86	5.66	-1.80	-3.14

표 4. 변수의 수는 10이고 공분산행렬이 상관계수행렬인 경우

평균간 거리 (Δ)	두집단 표본수 (n_1, n_2)	명목오류백분율			명목오류백분율편의		
		KNN	LDF	PP	KNN	LDF	PP
1.049	15, 15	40.46	16.90	2.43	10.46	-13.10	-27.57
1.049	30, 30	39.58	23.30	11.80	9.58	-6.70	-18.20
1.683	15, 15	29.30	9.60	1.79	9.30	-10.40	-18.21
1.683	30, 30	28.32	14.75	7.37	8.32	-5.25	-12.63
2.583	15, 15	17.80	4.50	0.83	7.80	-5.50	-9.17
2.583	30, 30	15.72	6.80	3.78	5.72	-3.20	-6.22

표 5. 변수의 수는 5이고 공분산행렬이 혼합분포인 경우

평균간 거리 (Δ)	두집단 표본수 (n_1, n_2)	명목오류백분율			명목오류백분율편의		
		KNN	LDF	PP	KNN	LDF	PP
1.049	10,10	53.36	28.75	11.80	23.36	-1.25	-18.20
1.049	15,15	49.03	32.53	18.13	19.03	2.53	-11.87
1.049	30,30	48.40	37.90	25.82	18.40	7.90	-4.18
1.683	10,10	46.80	26.15	9.60	26.80	6.15	-10.40
1.683	15,15	44.63	34.16	15.80	24.63	14.16	-4.20
1.683	30,30	42.88	38.33	22.38	22.88	18.33	2.38
2.583	10,10	37.10	29.70	7.00	27.10	19.70	-3.00
2.583	15,15	31.96	31.76	10.60	21.96	21.76	0.60
2.583	30,30	30.93	38.15	15.02	20.93	28.15	5.02

표 6. 변수의 수는 10이고 공분산행렬이 혼합분포인 경우

평균간 거리 (Δ)	두집단 표본수 (n_1, n_2)	명목오류백분율			명목오류백분율편의		
		KNN	LDF	PP	KNN	LDF	PP
1.049	15,15	51.83	23.43	4.10	21.83	-6.57	-25.90
1.049	30,30	50.58	33.46	14.83	20.58	3.46	-15.17
1.683	15,15	49.43	22.73	3.16	29.43	2.73	-16.84
1.683	30,30	46.68	34.92	11.36	26.68	14.92	-8.64
2.583	15,15	39.10	23.56	2.30	29.10	13.56	-7.70
2.583	30,30	35.22	32.60	7.96	25.22	22.60	-2.04

참 고 문 헌

[1] 안 윤기, 이 성석(1990), "엔트로피지수를 사용한 투사지향방법에 의한 판별분석," 「산업과 경영」, 통권 51호, 111-123.

[2] Anderson, Mashall W. and Benning, Roger D.(1970), "A Distribution-free Discrimination Procedure on Clustering," *IEEE Transactions on Information Theory*, IT-16, 541-548.

[3] Broffitt, J.D.(1982), "Nonparametric Classification," *Handbook of Statistics*, 2, 139-168,

North-Holland, Amsterdam.

[4] Cover, Thomas M. (1968), "Estimation by the Nearest Neighbor Rule," *IEEE Transactions on Information Theory*, IT-14, 50-55.

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

[6] Friedman, Jerome H. and Tukey, John W. (1974), "A Projection Pursuit Algorithm for Exploratory Data Analysis", *IEEE Transactions on Computers*, C23, 881-890.

[7] Fukunaga, Keinosuke. and Kessell, David L. (1971), "Estimation of Classification Error," *IEEE Transactions on Computers*, C-20, 285-293.

[8] Gong, Gail. (1986), "Cross-Validation, the Jackknife, and the Bootstrap : Excess Error Estimation in Forward Logistic Regression," *Journal of the American Statistical Association*, 81, 108-113.

[9] Hand, D.J. (1982), *Kernel Discrimination Analysis*. Research Studies Press, New York.

[10] Jones, M.C. and Sibon, Robin. (1987), "What is Projection Pursuit ? ," *Journal of Royal Statistical Society A*, 150, 1-36.

[11] Krazonowski, W.J. (1977), "The Performance of Fisher's Linear Discriminant Function under Non-optimal Conditions," *Technometrics*, 29, 191-200.

[12] Kruskal, J.B. (1972), "*Linear Transformation of Multivariate Data to Reveal Clustering*," In : Multidimensional Scaling : Theory and Application in the Behavioral Sciences, 1, *Theory*. Seminar Press, New York.

[13] McLachlan, G.J. (1989), "Assessing the Performance of an Allocation Rule," In : S.C. Choi, ed., *Statistical Methods of Discrimination and Classification : Advances Theory and Applications*, Pergamon Press, New York.

[14] Penrod, C.S. and Wagner, T.J. (1979), "Risk Estimation for Nonparametric Discrimination and Estimation Rules : A Simulation Study," *IEEE Transactions of Information Theory*, IT-25, 753-759.

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

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

[17] Van Ness, John. (1979), "On the Effects of Dimension in Discriminant Analysis for Unequal Covariance Populations," *Technometrics*, 21, 119-127.

A Simulation Study on Projection Pursuit Discriminant Analysis†

Yunkee Ahn* and Sung-Suk Rhee**

ABSTRACT

The projection pursuit method has been suggested as a technique for the analysis of the multivariate data. This method seeks out interesting linear projections of the multivariate data onto a line of a plane to solve the curse or dimensionality.

In this paper we developed the discriminant analysis by using the projection method and simulations were used for comparison between this and other existing discriminant analysis methods.

† Research supported in part by the Basic Science Research Institute Program, Ministry of Education, 1990

* Department of Applied Statistics, Yonsei University, 134 Shinchon-dong, Sudaemoon-ku, Seoul 120-749, Korea

** Department of Applied Statistics, Seowon University, 231 Mochung-dong, Chungju, Chungbuk 360-742, Korea