응용통계연구 제 11권 2호, 1998년, pp. 435-449

MDA에서 판별변수 선택을 위한 베이즈 기준

김혜중[1] 유희경 [2]

요 약

본 연구는 다중판별분석(MDA)에서 필요한 변수선택기준을 베이즈접근법으로 제안하였다. 이 베이즈판별변수 선택기준은 여러 정규모집단분포의 평균벡터에 대한 등질성 검정에 필요한 디폴터형태의 베이즈요인을 객관적 베이즈방법으로 유도하여 설정하였다. 디폴트베이즈요인(default Bayes factor)은 Spiegelhalter와 Smith (1982)가 개발한 가상적트레이닝표본법(imaginary training sample method)을 사용하여서 도출하였다. 또한 제안된 베이즈판별변수선택기준이 지닌 분포의 성질을 이용하여, 추가 판별변수(또는 변수군)가 MDA에 기여하는 부가적인 판별력에 대한 검정법 및 추가판별변수(또는 변수군)의 선택기준에 대해서도 논하였다. 본 연구에서 새로이 얻은 변수선택기준은 최적부분집합선택법(optimal subset selection method)뿐 아니라 단계적방법(stepwise method)의 변수선택기준으로 사용될 수 있으며, 두그룹 판별분석에도 사용이 가능하다는 점에서 표본이론에 의해 여러형태로 개발된 기존의 판별변수 선택기준들을 하나로 통합시킬 수 있는 기능을 지니고 있다. 모의실험을 실시하여 최적부분집합선택법과 단계적방법하에서 제안된 판별변수선택기준이 가진 효용성을 평가하였다.

1. 서론

다중판별분석(multiple discriminant analysis; MDA)은 관측벡터가 X인 임의의 표본단위를 이미 정의된 세 개이상의 그룹이나 모집단에 정확히 분류하는 문제에 관련된 다변량 분석기법이다. MDA에서 K개 모집단을 $\Pi_1, \cdots, \Pi_K$라 할 때, Rencher (1995)는 오분류확률의 기대값을 최소로 하는 분류절차는 관측벡터 X를 가진 표본단위를 다음의 조건하에서 Π_g에 분류시키는 것임을 보였다.

$$\pi(k)P_k(X) \leq \pi(g)P_g(X), \quad k = 1, \cdots, K; k \neq g. \tag{1.1}$$

여기서, $\pi(k)$는 관측벡터가 Π_k에 속할 사전확률이며, $P_k(X)$는 Π_k의 확률밀도함수를 나타낸다. Kim (1995, 1996)외 여러 연구들에 의해 위 절차는 특정한 손실함수하에서 베이즈 절차가 되며, 모든 분류규칙들 중에서 오분류 위험(misclassification risk)을 최소화시키는 것임이 밝혀졌다.

특히 다변량 정규모집단의 경우 (즉, $\Pi_k \sim N(\mu_k, \Sigma), k = 1, \cdots, K$) 식 (1.1)은

$$(\bar{X}_g - \bar{X}_k)'S^{-1}(X - \frac{1}{2}(\bar{X}_g + \bar{X}_k)) \geq ln\frac{\pi(k)}{\pi(g)}, \quad k \neq g \tag{1.2}$$

1) (100-715) 서울시 중구 필동 3가 26, 동국대학교 통계학과, 교수
2) (245-080) 강원도 삼척시, 삼척산업대학교 컴퓨터과학과, 부교수

로 추정된다. 여기서 $\bar{X}_g$, $\bar{X}_k$와 S는 μ_g, μ_k 및 Σ의 불편 추정량을 각각 나타낸다.

효과적인 MDA를 수행하기 위하여 분석자는 분류규칙의 도출에 앞서 X의 성분인 판별변수들이 MDA에 미치는 영향을 평가한 후 판별에 기여도가 없거나 아주 낮은 변수들은 분석으로부터 제거시키는 절차가 필요하다. 이것을 판별 변수선택절차라 하며, 이 절차는 변수의 측정비용절감 및 계산시간의 단축에 효율적일 뿐아니라 MDA의 오분류율을 낮추는 데에도 유용한 것이다. 판별변수선택법에는 크게 단계적 방법과 최적부분집합 선택법(optimal subset selection method)으로 구분된다. 그리고 전자는 후자와 달리 변수 선택에 필요한 계산이 간단하도록 고안되었으나 이 방법으로는 X로부터 최적의 판별변수부분집합을 찾아낼 수 없는 문제점이 McLachlan (1992)에 의해 제기되었다. 한편, MDA의 판별변수선택의 기준으로는 Wilks의 Λ 통계량, Mahalanobis 거리제곱, Evans와 Schwager (1994)의 추정된 베이즈 위험의 차이, 그리고 Ganeshanomdam과 Krzanowsk (1989)에서 정의된 외견오분류율(apparent error rate)이나 실오분류율(actual error rate)에 의해 추정되는 오분류율이 주로 사용되고 있다.

앞에서 열거된 변수선택기준들은 오분류율을 제외하고는 모두 단계적 방법에 사용되는 변수 선택기준이며, 최적부분집합 선택법에는 계산이 복잡한 오분류율기준 외에는 제안된 것이 없는 실정이다(Johnson과 Witchern (1992) 참조). 본 논문에서는 이점에 착안하여 새로운 판별변수 선택기준을 제안하고자 한다. 새로이 제안될 선택기준은 계산이 복잡한 오분류율을 대체할 수 있을 뿐아니라, 기존의 것들과 달리 최적부분집합 선택법및 단계적 방법에 공통적으로 사용할 수 있는 판별변수 선택기준이다. 이 기준은 다변량 정규모집단하에서 가정된 판별모형들의 주변우도비 (marginal likelihood ratio)인 베이즈 요인(Bayes factor)에 의해서 정의된다. 2장에서는 다변량 정규판별모형의 모수들에 대한 객관적 사전확률분포(objective prior) 하에서 유도된 베이즈요인은 Kass와 Raftery (1995)가 언급한 베이즈요인의 임의성 문제가 발생됨을 보이고, Spigelhalter와 Smith (1982)가 제안한 가상적 트래이닝표본기법(imaginary training sample method)으로 그 문제를 해결하여 도출한 새로운 판별변수선택기준을 제안한다. 3장은 제안된 판별변수 선택기준의 분포이론에 대한 논의와 함께 선택된 판별변수의 판별력 평가에 필요한 검정통계량을 유도한다. 4장에서는 모의 실험을 통하여 제안된 변수 선택기준의 유용성을 보이고, 이에 대한 평가 및 결론과 차후 연구과제에 대한 논의를 5장에서 할 것이다.

2. MDA의 새로운 판별변수 선택 기준

Jeffreys (1935)가 베이즈 요인을 사용하여 두 대응 모형의 비교방법을 제안한 후, 베이즈 요인은 여러 분야에서 사용되는 통계적 모형의 설정 및 적합도검정에 응용되어 왔다(Kass와 Raftery (1995) 참조). 이 장에서는 다변량 정규모집단 가정하에서 MDA의 판별변수선택에 유용한 베이즈요인을 유도하고, 이 때 필연적으로 발생되는 베이즈요인의 임의성을 Spiegelhalter와 Smith (1982)가 제안한 가상적 트레이닝표본법으로 제거시켜서 얻은 새로운 MDA 변수선택기준을 제시한다.

2.1. 베이즈요인

MDA에서 가정한 K개 서로 독립인 정규모집단을 $\Pi_1, \cdots, \Pi_K$라 하고, 각 모집단의 분포는 $\Pi_k \sim N(\mu_k, \Sigma)$, $k = 1, 2, \cdots, K$이며, $p \times 1$ 평균벡터인 μ_k와 $p \times p$ 공분산 행렬인 Σ는 미지의 모수라 하자. 이러한 MDA 모형의 가정하에서 판별변수선택에 사용될 베이즈요인을 얻기 위해서는 기존의 선택기준과 같이 K개 모집단의 평균벡터들에 대한 다음과 같은 두 대응 모형(M_1과 M_2)의 설정이 필요하다.

$$M_1 : \mu_1 = \cdots = \mu_K = \mu \text{ versus } M_2 : \mu_1 \neq \cdots \neq \mu_K. \tag{2.1}$$

따라서, 모형 M_1은 p개 판별변수들이 MDA에서 판별력이 없다는 것을 나타내고, 이와 대응되는 의미를 지닌 모형은 M_2가 된다. 이와 더불어, p개 판별변수들 중에서 각 변수 또는 부분변수군이 MDA에서 기여하는 부분적 판별력 향상의 정도를 측정하는 모형의 설정 및 측정방법도 판별변수의 선택에서 중요한 과제다. 이에 대한 논의는 다음 장에서 별도로 할 예정이다.

$X_1(k), X_2(k), \cdots, X_{N_k}(k)$를 모집단 Π_k로부터 독립적으로 얻은 표본크기 N_k인 p개 판별변수에 대한 표본이라 하고, K개 모집단으로부터 얻은 K개 독립표본들을 모두 D로 나타내면 모형 M_1과 M_2에서 얻어지는 D의 확률밀도함수는 각각

$$P(D|\mu, \Sigma, M_1) = (2\pi)^{-\frac{Np}{2}} |\Sigma|^{-\frac{N}{2}} \exp\left\{ -\frac{1}{2} \text{tr}[\Sigma^{-1}\Omega] \right\} \tag{2.2}$$

와

$$P(D|\mu_1, \cdots, \mu_K, \Sigma, M_2) = (2\pi)^{-\frac{Np}{2}} |\Sigma|^{-\frac{N}{2}} \exp\left\{ -\frac{1}{2} \text{tr}\left[\Sigma^{-1} \sum_{k=1}^{K} \Omega_k\right] \right\}$$

이다. 여기서, $\Omega = V + N(\mu - \bar{X})(\mu - \bar{X})'$, $\Omega_k = V_k + N_k(\mu_k - \bar{X}(k))(\mu_k - \bar{X}(k))'$, $\bar{X}(k) = \sum_{j=1}^{N_k} X_j(k)/N_k$, $\bar{X} = \sum_{k=1}^{K} N_k \bar{X}(k)/N$, $N = \sum_{k=1}^{K} N_k$, $V_k = \sum_{j=1}^{N_k} (X_j(k) - \bar{X}(k))(X_j(k) - \bar{X}(k))'$, $V = \sum_{k=1}^{K} \sum_{j=1}^{N_k} (X_j(k) - \bar{X})(X_j(k) - \bar{X})'$이다.

여기에서의 관심 사항은 미지모수에 대한 추론이 아니라, 모형 M_1과 M_2중 어느 모형이 D에 적합한 것 인지에 대한 판단이기 때문에, 객관적 베이즈 방법으로 모형의 적합도를 검정하는 것이 타당하다. 따라서 적합도 검정에 사용될 사전확률 분포는 각 모형하에서의 부적절 사전확률분포(improper prior distribution)로서 다변량 정규-역 위셔트 공액사전확률분포(normal-inverted Wishart conjugate prior)의 극한 형태를 사용하였고, 그 형태는 다음과 같이 설정된다(DeGroot (1970) 참조).

$$\begin{aligned} P_0(\mu, \Sigma|M_1) &= P_0(\mu|\Sigma, M_1) P_0(\Sigma|M_1) = c_1 |2\pi\Sigma|^{-\frac{1}{2}} |\Sigma|^{-\frac{p+1}{2}}, \\ P_0(\mu_1, \cdots, \mu_k, \Sigma|M_2) &= P_0(\mu_1, \cdots, \mu_k|\Sigma, M_2) P_0(\Sigma|M_2) \\ &= c_2 |2\pi\Sigma|^{-\frac{K}{2}} |\Sigma|^{-\frac{p+1}{2}}. \end{aligned} \tag{2.3}$$

여기서, c_1과 c_2는 정규화상수(normalizing constants)를 나타낸다.

(2.2)에서 주어진 D의 확률밀도함수와 부적절 사전확률분포하에서 Kass와 Raftery (1995)의 정의에 따라 구한 표본 D의 주변우도함수는 모형 M_1과 모형 M_2의 조건하에서

$$P(D|M_1) = \int\int P(D|\mu,\Sigma,M_1)P_0(\mu,\Sigma|M_1)d\mu d\Sigma$$
$$= c_1 N^{-\frac{p}{2}}\Delta_p|V|^{-\frac{N}{2}} \tag{2.4}$$

과

$$P(D|M_2) = \int\cdots\int P(D|\mu_1,\cdots,\mu_k,\Sigma,M_2)P_0(\mu_1,\cdots,\mu_k\Sigma|M_2)(\prod_{k=1}^{K}d\mu_k)d\Sigma$$
$$= c_2\Big(\prod_{k=1}^{K}N_k\Big)^{-\frac{p}{2}}\Delta_p\left|\sum_{k=1}^{K}V_k\right|^{-\frac{N}{2}} \tag{2.5}$$

이 된다. 단, $\Delta_p = \pi^{p(p-2N-1)/4}\prod_{i=1}^{p}\Gamma\{(N-i+1)/2\}$.

그러므로, 모수의 부적절 사전확률분포를 (2.3)로 가정할때, 표본 D가 가진 모형 M_1과 모형 M_2간에 조건부 확률 승산비인 베이즈요인은

$$B_{12} = \frac{P(D|M_1)}{P(D|M_2)} = \frac{c_1}{c_2}\Big(\prod_{k=1}^{K}N_k/N\Big)^{\frac{p}{2}}\Big(|V|\Big/\Big|\sum_{k=1}^{K}V_k\Big|\Big)^{-\frac{N}{2}} \tag{2.6}$$

이 된다.

베이즈요인값에 대한 구체적이고 구간별로 이루어진 해석은 Jefferys (1961) 및 Kass와 Raftery (1995)에서 제안하고 있으나, 일반적으로 $B_{12} > 1$이면, 모형 M_2보다 모형 M_1이 표본 D에 더 적합한 모형으로 판단한다.

한편, (2.6)에서 도출된 베이즈요인에는 정의되지 않은 정규화상수인 c_1과 c_2의 비가 포함되어 있다. 이로 인해 객관적 베이즈 방법에서 필연적으로 발생되는 베이즈요인의 임의성 문제를 (2.6)이 앓고 있어서 모형 M_1과 모형 M_2의 적합성 검정에 (2.6)을 사용할 수 없게 된다. c_1과 c_2의 비를 모수들의 사후확률분포에 의해 추정함으로서 베이즈요인의 임의성 문제를 해결하려는 노력이 많은 연구에 의해 행해졌다. 이들 방법 중 대표적인 것으로는 Aitkin (1991)의 방법, Berger와 Perrichi (1996)의 고유 베이즈요인(intrinsic Bayes factors), 그리고 O'Hagan (1995)의 단편적 베이즈요인(fractional Bayes factors)을 들 수 있다.

그러나 이들 방법은 베이즈요인의 계산에 주어진 표본의 일부 또는 전부를 반복하여 사용하는 것으로, 사전확률분포의 정의 및 베이즈이론의 논리와 대치되는 것으로 평가되었다(Berger와 Perrichi (1996) 참조). 한편, Spiegelhalter와 Smith (1982)가 제안한 가상적 트레이닝표본법은 이러한 문제와 관계없이 c_1/c_2를 구할 수 있는 방법이다.

2.2. 가상적 트레이닝표본방법에 의한 판별변수 선택기준

가상적 트레이닝표본방법이란 Good (1947)가 제안한 방법의 한 변형으로, 이를 사용하면 (2.6)에 포함된 c_1/c_2값을 계산해 낼 수 있다. 가상적 트레이닝표본이란 다음과 같이 정의된다.

정의 1 (Spiegelhalter와 Smith (1982)). 베이즈요인에 의한 두 모형 M_1과 M_2의 비교에서
(i) 두 모형 M_1과 M_2의 비교에 필요한 최소한의 자료로 구성되며,
(ii) 모형 M_1에 대한 최대의 적합도를 보이면서 베이즈요인 값이 $B_{12} \approx 1$이 되도록 하는 자료가 존재하면, 이 자료를 가상적 트레이닝표본이라 한다.

위 정의에 의하면, 식 (2.6)에 포함된 c_1/c_2의 값을 다음과 같은 절차로 얻을 수 있다.

첫째, 정의 1의 조건 (i)로 부터 (2.6)의 계산에 필요한 최소한의 자료수는 K개 정규모집단의 평균벡터 $\mu_1, \cdots, \mu_K$의 추정에 각각 1개씩의 자료가 필요하고, 공분산행렬 Σ의 추정에 최소한 p개의 자료가 더 필요하다. 그러므로, 가상적 트레이닝표본의 크기는 $N_k = 1$과 $N_{k'} = p + 1$이다. 단, $k = 1, \cdots, K;\ k \neq k', 1 \leq k' \leq K$.

둘째, (2.6)에서 $\left(|V| \Big/ \left|\sum_{k=1}^{K} V_k\right|\right)^{-\frac{N}{2}}$을 λ로 표기하면, λ는 귀무가설 M_1을 검정하기 위한 우도비 검정통계량과 같다 (Anderson (1984) 참조). 한편 λ_0를 첫째 절차로 부터 얻은 크기가 $p+K$인 가상적 트레이닝표본을 λ의 식에 적용시켜서 얻은 우도비 검정통계량이라 하자. 그러면 우도비 검정통계량의 성질에 의해 $0 \leq \lambda_0 \leq 1$이 되며, 모형 M_1에 대한 최대 적합도는 $\lambda_0 = 1$의 값을 가질 때 이루어진다 (정의 1의 조건 (ii) 참조).

마지막으로, 가상적 트레이닝표본에 의해서 계산된 식 (2.6)의 왼쪽 항을 조건 (ii)에 따라 1로 두고 오른쪽 항에는 위에서 얻은 두 결과를 대입하여 얻은 등식을 c_1/c_2에 대해 풀면

$$\frac{c_1}{c_2} = \left(\frac{p+K}{p+1}\right)^{\frac{p}{2}} \tag{2.7}$$

가 된다.

위 절차로부터 얻은 c_1/c_2값을 (2.6)에 대입시키면 다음의 정리를 얻는다.

정리 2.1 부적절 사전확률분포인 (2.3)를 이용한 모형 M_1과 모형 M_2의 비교에서 가상적 트레이닝표본방법으로 도출한 베이즈요인은

$$B_{12}^{I} = \left(\frac{p+K}{p+1}\right)^{\frac{p}{2}} \left(\frac{\prod_{k=1}^{K} N_k}{N}\right)^{\frac{p}{2}} U^{\frac{N}{2}} \tag{2.8}$$

이다. 여기서, $U = \left(|V| \Big/ \left|\sum_{k=1}^{K} V_k\right|\right)^{-1}$은 M_1의 검정에 사용되는 우도비 검정 기준이며, $U_{p,\alpha,\beta}$ $(\alpha = K-1, \beta = N-K)$의 분포를 따른다 (Anderson 1984 참조). 그리고, B_{12}^{I}에서 윗첨자 I는 베이즈요인의 상수항(c_1/c_2)을 가상적 트레이닝표본법으로 구했음을 나타낸다.

(2.8)에서 제안된 베이즈 요인의 유용성은 (i) 이미 열거된 기존의 c_1/c_2 계산 방법들에서 얻어질 베이즈요인과는 달리 계산이 편리한 폐쇄형태(closed form)를 가진 디폴트 베이즈요인(default Bayes factor)이며, (ii) 공액(또는 적절)사전확률분포를 사용했을 때 발생되는 초모수(hyperparameters)의 평가작업이 필요없으며, (iii) 제안된 베이즈 요인은 모형 M_1과 M_2간에 검정을 위하여 p와 K의 값에 따라 표본분포이론에 의해 여러 형태로 제안된 기존의 검정통계량들과 함수관계를 가지고 있다. 따라서, 제안된 베이즈 요인은 기존의 검정통

계량들을 한개의 통일된 검정 기준으로 통합시키는 역할을 한다. p와 K의 값에 따른 베이즈 요인과 기존의 검정통계량들과의 관계는 다음과 같다.

$p=1$일 때 B_{12}^I는 아래와 같이 일원배치 분산분석의 F 검정통계량의 함수 형태를 가진다.

$$B_{12}^I = \left(\frac{K+1}{2}\right)^{\frac{1}{2}} \left(\frac{\prod_{k=1}^K N_k}{N}\right)^{\frac{1}{2}} \left(1+\frac{K-1}{N-K}F\right)^{-\frac{N}{2}}, \tag{2.9}$$

단, $F = \frac{N-K}{K-1} \frac{\sum_{k=1}^K N_k(\bar{X}(k)-\bar{X})^2}{\sum_{k=1}^K \sum_{j=1}^{N_k}(X_j(k)-\bar{X}(k))^2}$이며 $F_{(K-1,N-K)}$ 분포를 따른다.

또한, $K=2$ 인 경우 B_{12}^I 는 귀무가설 M_1과 대립가설 M_2의 검정에 사용되는 Hotelling의 T^2 검정통계량의 함수 형태로 주어진다.

$$B_{12}^I = \left(\frac{p+2}{p+1}\right)^{\frac{p}{2}} \left(\frac{N_1N_2}{N}\right)^{\frac{p}{2}} \left(1+\frac{T^2}{N-2}\right)^{-\frac{N}{2}}, \tag{2.10}$$

여기서, $T^2 = \frac{N_1N_2}{N}(\bar{X}(1)-\bar{X}(2))'S^{-1}(\bar{X}(1)-\bar{X}(2))$ 이고 $S = \frac{1}{N-2}(V_1+V_2)$ 이다.

그리고, (2.10)의 결과를 $p=1$인 경우에 적용하면, B_{12}^I는 귀무가설 M_1과 대립가설 M_2의 검정에 필요한 두 표본 t 검정통계량의 함수 형태인

$$B_{12}^I = \left(\frac{3N_1N_2}{2(N_1+N_2)}\right)^{\frac{1}{2}} \left(1+\frac{t^2}{N_1+N_2-2}\right)^{-\frac{N_1+N_2}{2}}$$

가 된다. 여기서, t는 $t_{N_1+N_2-2}$의 분포를 따른다.

표 2.1은 베이즈 요인 B_{12}^I의 여러 임계값 (Kass와 Raftery (1995) 참조) 들인 1, $10^{-\frac{1}{2}}$, 0.1, 0.01로 각각 놓았을 때, 이에 대응하는 $U_{p,\alpha,\beta}(p\geq 2, K\geq 3$인경우), Hotelling의 $T^2(p\geq 2, K=2$인경우), $F(p=1, K\geq 3$인경우), 및 t 검정통계값$(p=1, K=2$인경우)으로부터 계산된 P-값(probability value)을 판별변수의 수 p, 그룹의 수 K 및 표본의 크기 $N^* = N_1 = ,\cdots, = N_K$ 값의 변화에 따라 계산하여 기록한 것이다. 여기서 $U_{p,\alpha,\beta}$ 검정통계량으로 부터 구한 P-값은 U 통계량의 카이제곱분포 근사(Anderson (1984) 참조)인

$$-\{\alpha+\beta-(p+\alpha+1)/2\}\log U_{p,\alpha,\beta} \approx \chi^2_{(\alpha p)} \tag{2.11}$$

를 이용하여 구한 값이다.

표 2.1에 나열된 B_{12}^I의 임계값과 기존의 검정통계값 간에 관계를 요약하면 다음과 같다. K개 모집단으로부터 추출된 표본들의 집합이 소표본일때 p, K, 및 N^*값이 작을 경우가 값이 클 경우보다 P-값이 항상 높다. 중표본(moderate sample)일 경우 B_{12}^I의 임계값 $10^{-\frac{1}{2}}, 0.1, 0.01$에 대응하는 기존의 검정통계량의 P-값은 주어진 조건 (p, K, N^*의값)에 관계없이 0.06 미만으로 나타나고 있어서, B_{12}^I에 의한 모형 M_1과 M_2의 적합도 검정기준으로 임계값을 $10^{-\frac{1}{2}}$로 설정할 경우 대체적으로 기존의 검정보다 더 엄격한 검정기준(conservative test criterion)이 된다. 그러므로, 표 2.1의 결과는 Berger와 Sellke (1987)의 결과를 뒷받침해 주는 새로운 수치적 예로 볼 수있다. 한편, 대표본일 경우, B_{12}^I가 M_1보다 복잡한 구조를

표 2.1: B_{12}^I의 임계값에 대응하는 기존 검정통계량들의 P-값

			B_{12}^I의 임계값			
p	K	N^*	1	$10^{-\frac{1}{2}}$	0.1	0.01
1	2	10	.1841	.0510	.0160	.0017
		20	.1110	.0301	.0088	.0008
	4	10	.1179	.0465	.0179	.0025
		20	.0431	.0158	.0057	.0007
	8	10	.0435	.0201	.0090	.0016
		20	.0062	.0026	.0011	.0002
2	2	10	.1993	.0549	.0281	.0039
		20	.0910	.0313	.0108	.0012
	4	10	.0827	.0391	.0178	.0034
		20	.0151	.0053	.0026	.0004
	8	10	.0133	.0068	.0034	.0008
		20	.0004	.0001	.0000	.0000
5	2	20	.0537	.0241	.0105	.0019
		30	.0213	.0084	.0039	.0005
	4	10	.0429	.0245	.0137	.0039
		20	.0011	.0005	.0002	.0000
	8	10	.0008	.0004	.0002	.0000
		20	.0000	.0000	.0000	.0000
10	2	20	.0341	.0158	.0098	.0026
		30	.0019	.0061	.0030	.0014
	4	10	.0303	.0260	.0130	.0052
		20	.0000	.0000	.0000	.0000
	8	10	.0000	.0000	.0000	.0000
		20	.0000	.0000	.0000	.0000

가진 모형 M_2를 선호하려면 B_{12}^I는 기존의 검정기준보다 매우 낮은 P-값을 필요로 하고 있다. 이는 "Lindley paradox"(Lee (1988) 참조) 현상이 B_{12}^I에도 적용되고 있음을 보여주고 있다. 이를 종합하면, B_{12}^I는 기존의 검정기준과 달리 모형 M_1과 M_2의 검정에서 표본의 크기 및 모형의 구조를 검정에 반영시켜서 유의수준을 검정상황에 맞게 자동적으로 조절하여 정하는 기능을 갖춘 기준임을 표 2.1에서 보여주고 있다.

따라서, 앞에서 제안한 베이즈 요인 B_{12}^I은 모형 M_1과 M_2의 비교에서, 표본분포이론에 의해 유도된 여러 검정통계량들을 대체할 수 있는 베이즈 기준이 됨을 보여준다. 또한, 표본분포이론하에서는 검정통계량의 분포가 p와 K의 값에 따라 $U_{p,\alpha,\beta}$ 분포, Hotelling의 T^2-분포, F-분포 그리고 t-분포로 변화하는 반면에 B_{12}^I는 이들을 한개의 검정기준으로 통합할 수

있는 것이 될 수 있다. 이는 B_{12}^I가 다변량 정규모집단 가정하의 MDA 또는 두 그룹 판별분석에서 주어진 판별변수들의 베이즈 선택기준이며, p와 K값에 따라 달라지는 기존의 판별변수 선택기준들을 하나의 통일된 기준으로 대신할 수 있는 것임을 의미한다.

3. 추가 판별변수 선택 기준

이 장에서는 판별변수 선택 기준 B_{12}^I에 의해 이미 선택된 판별변수군에 새로운 판별변수를 추가시켰을 때 추가된 판별변수가 MDA에서 판별에 미치는 부가적인 기여도를 검정할 수 있는 베이즈 기준을 정의하고자 한다. MDA에서 이미 선택한 판별변수들을 중요도 순서로 재배열하여서 $x_1, \cdots, x_{p-q}$로 나타낸 변수군에 새로운 변수 $x_{p-q+1}, \cdots, x_p$를 추가시키는 경우를 가정하고, 이에 따라, 2장에서 이미 정의한 판별변수들을 중요도 순서로 재배열한 벡터 $X = (x_1, \cdots, x_p)'$의 관측값 벡터 $X_j(k)$, 그리고 이로부터 계산된 그룹내 제곱합 행렬 V_k 및 총제곱합 행렬 V을 다음과 같이 분할하였다고 하자.

$$X_j(k) = \begin{bmatrix} X_j^{(p-q)}(k) \\ X_j^{(q)}(k) \end{bmatrix}, \quad V_k = \begin{bmatrix} V_{11}(k) & V_{12}(k) \\ V_{21}(k) & V_{22}(k) \end{bmatrix}, \quad V = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}, \tag{3.1}$$

$k = 1, \cdots, K$; $j = 1, \cdots, N_k$. 여기서 $X_j^{(p-q)}(k)$은 $(p-q) \times 1$ 벡터이며, $V_{11}(k)$와 V_{11}은 이에 대응되게 분할된 $(p-q) \times (p-q)$ 행렬들이다. $\sum_{k=1}^K V_k$를 A로 놓으면, 위의 분할에 의해 (2-8)에서 정의된 우도비 검정통계량 U는

$$U = \frac{|A|}{|V|} = \frac{|A_{11}||A_{22\bullet 1}|}{|V_{11}||V_{22\bullet 1}|} \tag{3.2}$$

으로 인자분해 된다. 그리고 U로 부터 분해된 인자 $|A_{22\bullet 1}|/|V_{22\bullet 1}|$은 추가된 판별변수들인 $x_{p-q+1}, \ldots, x_p$가 MDA에서 판별력 향상에 기여도가 없다는 귀무가설

$$H_0 : E[(x_{p-q+1}, \ldots, x_p)'|(x_1, \cdots, x_{p-q})', \Pi_k] = E[(x_{p-q+1}, \ldots, x_p)'|(x_1, \cdots, x_{p-q})', \Pi_l], \tag{3.3}$$

$k \neq l; k, l = 1, \cdots, K$ 하에서 아래의 분포를 따른다(Kshirsagar (1972) 참조).

$$\frac{|A_{22\bullet 1}|}{|V_{22\bullet 1}|} \sim \prod_{i=p-q+1}^{p} W_i, \quad n = N - K. \tag{3.4}$$

단, W_i들은 서로 독립이며, $(W_i^{-1} - 1)(n - i + 1)/(K - 1)$은 자유도가 $(K-1)$과 $(n-i+1)$인 F-분포를 따른다. 여기서, $A_{11} = \sum_{k=1}^K V_{11}(k)$, $A_{22} = \sum_{k=1}^K V_{22}(k)$, $A_{12} = A_{21}' = \sum_{k=1}^K V_{12}(k)$, $A_{22\bullet 1} = A_{22} - A_{21}A_{11}^{-1}A_{12}$, $V_{22\bullet 1} = V_{22} - V_{21}V_{11}^{-1}V_{12}$.

McKay (1977)는 분포 (3.4)를 사용하여 추가된 판별변수들인 $x_{p-q+1}, \ldots, x_p$가 MDA에서 기여한 부가적 판별력을 동시에 검정할 수 있는 동시검정통계량(simultaneous test statistic)을 다음과 같이 제안하였다.

$$\frac{|A_{22\bullet 1}|}{|V_{22\bullet 1}|} \sim U_{p,K-1,n}. \tag{3.5}$$

한편, (2.8)과 (3.2)에 의하면 인자 $|A_{22\bullet 1}|/|V_{22\bullet 1}|$은 판별변수의 수가 p인 경우에 정의되는 모형 M_1과 $p-q$인 경우에 정의되는 모형 M_1을 검정하기 위해 각각 사용되는 우도비 검정 통계량간에 비로 구할 수 있어 다음의 정의를 가능하게 한다.

정의 2 K개 동분산 정규모집단에 대한 MDA에서 판별변수의 수가 p일 때 (2.8)에서 정의된 베이즈 요인을 $B_{12}^I(x_1,\cdots,x_p)$, 그리고 변수의 수가 $p-q$일 때의 베이즈 요인을 $B_{12}^I(x_1,\cdots,x_{p-q})$라 하면, 기존에 선택된 판별변수군인 $\{x_1,\cdots,x_{p-q}\}$에 추가된 판별변수들인 $x_{p-q+1},\ldots,x_p$가 MDA에서 기여하는 부가적인 판별력은

$$BR(x_{p-q+1},\ldots,x_p)=B_{12}^I(x_1,\cdots,x_{p-q})/B_{12}^I(x_1,\cdots,x_p),\ \ q=1,\ldots,p-1 \tag{3.6}$$

이며, 이것을 추가판별변수 $x_{p-q+1},\ldots,x_p$에 대한 베이즈 선택기준이라 한다.

정의 2에 의하면 $BR(x_{p-q+1},\ldots,x_p)>1$ 일 경우 새로이 추가된 판별변수들이 MDA의 판별력을 향상시킨다고 볼 수 있으며, 다음의 정리로 그 판별력 향상에 대한 유의성을 검정할 수 있다.

정리 3.1 $1\le q\le p-1$에 대해서 $0<BR(x_{p-q+1},\ldots,x_p)<\infty$이며, 귀무가설 (3-3)하에서

$$(\alpha BR(x_{p-q+1},\ldots,x_p))^{-\frac{2}{N}}\sim U_{p,K-1,n},\ \ n=N-K \tag{3.7}$$

이다. 여기서 $n=N-K$이며 $\alpha=\left(\frac{(p-q+1)(p+K)}{(p+1)(p-q+K)}\right)^{\frac{p-q}{2}}\left(\frac{p+K}{p+1}\right)^{q/2}\left(\frac{\prod_{k=1}^K N_k}{N}\right)^{q/2}$이다.

증명: (2.8)에 유도된 베이즈 요인은 모든 $p\ge 1$에 대하여 $B_{12}^I>0$를 성립하므로 정의 2로부터 $0<BR(x_{p-q+1},\ldots,x_p)<\infty$이다. 그리고, (2.8)과 (3.3)으로 부터

$$\frac{|A_{22\bullet 1}|}{|V_{22\bullet 1}|}=\left(\alpha^{-1}\frac{B_{12}^I(x_1,\cdots,x_p)}{B_{12}^I(x_1,\cdots,x_{p-q})}\right)^{\frac{2}{N}}=(\alpha BR(x_{p-q+1},\ldots,x_p))^{-\frac{2}{N}}$$

의 관계식을 유도해 낼 수 있으며, 이것에 (3.5)를 이용하여 (3.7)을 구한다. □

식 (3.7)에 정의된 통계량의 분포인 $U_{p,K-1,n}$분포는 식 (2.11)으로 근사 시킬 수 있으며, 특히 $q=1$인 경우 검정통계량은

$$W_p=(\alpha BR(x_p))^{-2/N},\ \ \alpha=\left(\frac{p(p+K)}{(p+1)(p+K-1)}\right)^{(p-1)/2}\left(\frac{p+K}{p+1}\right)\left(\frac{\prod_{k=1}^K N_k}{N}\right)^{1/2} \tag{3.8}$$

이고 W_p의 분포는

$$(W_p^{-1}-1)(n-p+1)/(K-1)\ \sim\ F_{(K-1,n-p+1)} \tag{3.9}$$

가 되어 단계적방법에서 한 개의 판별변수 x_p가 추가로 기여하는 판별력의 검정에 사용된다.

4. 모의 실험

2장과 3장에서 제안된 베이즈요인을 이용한 판별변수 및 추가판별변수 선택기준이 MDA의 판별변수선택에 유용하게 사용될 수 있는 것 인지를 평가하기 위해서 모의 실험을 실시하였다. K개 모집단들의 다변량 정규성과 분산-공분산 행렬의 동일성 가정하에서 평균 벡터 및 분산-공분산 행렬의 값 변화에 따른 베이즈 판별변수 선택기준들의 효용성을 평가하기 위하여 이장에서는 경험적 자료분석보다 모의 실험을 택하였다.

모의 실험은 세 그룹($K = 3$) 판별분석의 경우로 한정하였으며, 실험에 보편성을 부여하기 위해서 판별변수 X_1, X_2, X_3, X_4에 대한 각 그룹의 분포를 아래와 같이 가정하였다.

$$\Pi_k \sim N(\mu_k, \sigma^2 I_4), \quad k = 1, 2, 3. \tag{4.1}$$

단, $\mu_1 = (1, -1, 1, -1)$, $\mu_2 = (\delta, -\delta, 1, -1)$, $\mu_3 = (-\delta, \delta, 1, -1)$.

따라서, (4.1)에서 정의된 판별변수 X_3, X_4의 조건부 기대값들은

$$E[(X_3\ X_4)' | (X_1\ X_2)', \Pi_k] = (1, -1)', \ k = 1, 2, 3$$

이 되어, $\{X_1, X_2\}$가 최적의 판별변수집합이 되도록 모의실험을 설계하였다. 또한, σ^2값이 주어졌을 때 세 그룹 상호간에 Mahalanobis거리는 δ만의 함수가 되도록 모집단의 평균벡터들을 설정하였다.

σ^2과 δ값의 변화에 따라 정의된 (4.1)의 분포로 부터 각각 크기가 $N_1 = N_2 = N_3 = N^* = 20, 50$인 표본을 발생시켜서 베이즈요인 B_{12}^I와 추가판별변수(1개)의 베이즈 선택기준인 $BR(X_p)$를 SAS/IML로 계산하였고, 이 실험을 100회 반복하여 B_{12}^I와 $BR(X_p)$값의 평균 및 표준편차를 얻었다. 이와 더불어, 제안된 기준들의 효용성을 기존의 판별변수 선택기준에 대비시켜 평가하기 위해서 SAS/STAT을 이용하여 매회의 실험마다 교차타당성법(cross validation method)으로 AER(actual error rate)의 추정값을 구하여 그들의 평균 및 표준편차도 함께 계산하였다.

표 4.1는 위와 같은 방법으로 행해진 모의실험에서 본 논문에서 제안한 베이즈기준과 기존의 기준인 AER을 사용하여 최적부분집합 선택법과 단계적 선택법을 실시한 결과를 나타낸 것이다. 표의 최적변수란에는 주어진 판별변수의 개수에 대해 최소의 B_{12}^I값을 가진 판별변수군을 찾아서 나타낸 것으로, 이들이 가진 B_{12}^I값과 AER추정값의 평균 및 표준편차(괄호안의 값)을 실었다. 또한, 단계적 변수선택법에서는 추가변수($X_p\ p = 1, \ldots, 4$)의 판별력향상에 대한 유의성검정을 (3.9)의 검정통계량으로 실시하였고, 검정으로부터 얻은 P-값을 표에 함께 실었다. 추가변수(X_p)의 판별력향상에 있어서 $BR(X_p) < 1$인 경우는 유의성검정이 불필요하나, $BR(X_p)$의 값과 P-값과의 관계를 보이기 위해서 참고로 표에 나타내었다.

5. 실험결과 분석 및 결론

본 논문에서는 베이즈 추론법을 이용하여 MDA에서 표본이론에 의해 여러형태로 개발된 변수선택기준들을 한 개의 기준으로 대체 시킬 수 있는 새로운 변수선택기준을 제안하

표 4.1: B_{12}^I에 의해 판별변수의 개수별로 얻은 최적변수군 (괄호속의 값은 표준편차를 나타냄)

최적 변수군	$\delta = 1.5$ B_{12}^I	AER	P-값	최적 변수군	$\delta = -1.5$ B_{12}^I	AER	P-값
$(N^* = 20, \sigma^2 = 0.1)$							
X_2	4.672E-34	0.1418	-	X_2	6.270E-35	0.1388	-
	(0.0000)	(0.0455)			(0.0000)	(0.0422)	
X_1X_2	2.945E-41*	0.0953	0.0000	X_1X_2	1.935E-42*	0.0932	0.0000
	(0.0000)	(0.0397)			(0.0000)	(0.0372)	
$X_1X_2X_4$	1.707E-40	0.0977	0.4862	$X_1X_2X_3$	1.781E-41	0.0937	0.7429
	(0.0000)	(0.0384)			(0.0000)	(0.0364)	
$X_1X_2X_3X_4$	6.240E-40	0.1003	0.3350	$X_1X_2X_3X_4$	6.022E-41	0.0982	0.3123
	(0.0000)	(0.0398)			(0.0000)	(0.0369)	
$(N^* = 50, \sigma^2 = 0.1)$							
X_1	9.110E-33	0.1430	-	X_2	2.226E-35	0.1372	-
	(0.0000)	(0.0197)			(0.0000)	(0.0214)	
X_1X_2	1.630E-42*	0.0893	0.0000	X_1X_2	6.603E-43*	0.0870	0.0000
	(0.0000)	(0.0181)			(0.0000)	(0.0157)	
$X_1X_2X_4$	5.400E-42	0.0909	0.1122	$X_1X_2X_3$	4.988E-42	0.0873	0.2490
	(0.0000)	(0.0182)			(0.0000)	(0.0165)	
$X_1X_2X_3X_4$	1.397E-41	0.0928	0.0927	$X_1X_2X_3X_4$	1.820E-41	0.0887	0.1290
	(0.0000)	(0.0209)			(0.0000)	(0.0177)	
$(N^* = 20, \sigma^2 = 0.5)$							
X_1	5.635E-16	0.2861	-	X_2	3.688E-16	0.2650	-
	(5.049E-15)	(0.0732)			(2.172E-15)	(0.0572)	
X_1X_2	4.310E-22*	0.2358	0.0000	X_1X_2	1.868E-22*	0.2110	0.0000
	(0.0000)	(0.0535)			(0.0000)	(0.0598)	
$X_1X_2X_4$	4.567E-21	0.2405	0.8453	$X_1X_2X_3$	1.205E-21	0.2187	0.5363
	(0.0000)	(0.0518)			(0.0000)	(0.0607)	
$X_1X_2X_3X_4$	2.062E-20	0.2487	0.4051	$X_1X_2X_3X_4$	1.014E-20	0.2287	0.7095
	(0.0000)	(0.0583)			(0.0000)	(0.0619)	
$(N^* = 50, \sigma^2 = 0.5)$							
X_1	1.295E-15	0.2698	-	X_2	5.537E-16	0.2695	-
	(6.070E-15)	(0.0279)			(3.038E-15)	(0.0289)	
X_1X_2	4.040E-21*	0.2130	0.0000	X_1X_2	1.789E-21*	0.2133	0.0000
	(0.0000)	(0.0266)			(0.0000)	(0.0251)	
$X_1X_2X_4$	3.163E-20	0.2142	0.2578	$X_1X_2X_3$	1.025E-20	0.2155	0.1906
	(0.0000)	(0.0267)			(0.0000)	(0.0274)	
$X_1X_2X_3X_4$	3.659E-19	0.2175	0.3906	$X_1X_2X_3X_4$	4.448E-20	0.2177	0.1523
	(0.0000)	(0.0269)			(0.0000)	(0.0262)	
$(N^* = 20, \sigma^2 = 1.0)$							
X_1	1.292E-9	0.3485	-	X_2	4.504E-10	0.3385	-
	(1.039E-8)	(0.0708)			(3.754E-9)	(0.0888)	
X_1X_2	5.092E-15*	0.2788	0.0000	X_1X_2	6.013E-15*	0.2672	0.0000
	(4.578E-14)	(0.0516)			(5.155E-14)	(0.0604)	
$X_1X_2X_4$	5.041E-15	0.2888	0.0962	$X_1X_2X_3$	1.930E-14	0.2807	0.2828
	(3.916E-14)	(0.0587)			(1.327E-13)	(0.0636)	
$X_1X_2X_3X_4$	5.104E-14	0.2987	0.8381	$X_1X_2X_3X_4$	4.371E-14	0.2920	0.2177
	(4.339E-13)	(0.0612)			(2.374E-13)	(0.0678)	
$(N^* = 50, \sigma^2 = 1.0)$							
X_2	9.984E-10	0.3376	-	X_2	1.206E-9	0.3394	-
	(6.234E-9)	(0.0284)			(5.540E-9)	(0.0281)	
X_1X_2	1.488E-12*	0.2711	0.0005	X_1X_2	2.791E-15*	0.2720	0.0000
	(1.487E-1)	(0.0273)			(1.503E-14)	(0.0262)	
$X_1X_2X_4$	4.141E-12**	0.2726	0.0948	$X_1X_2X_3$	1.453E-14	0.2736	0.1737
	(4.139E-11)	(0.0269)			(1.158E-13)	(0.0260)	
$X_1X_2X_3X_4$	2.994E-12	0.2756	0.0272	$X_1X_2X_3X_4$	2.446E-14	0.2757	0.0613
	(2.981E-11)	(0.0287)			(1.211E-13)	(0.0274)	

였다. 이것을 위하여 MDA에서 가정한 정규모집단들 간에 평균 벡터들의 차를 검정할 수 있는 베이즈기준을 제안하였다. 이 기준은 베이즈 추론에서 모형들의 적합도 검정에 주로 사용되는 객관적 사전분포를 이용한 베이즈요인에 의해 도출되었으며, 객관적 사전분포인 부적절 사전분포하에서 베이즈요인을 유도해 낼 때 발생되는 베이즈요인의 임의성문제는 가상적 트레이닝표본법으로 해결하였다. 또한, 단계적 선택법에서 필요한 추가판별변수의 선택기준은 제안된 베이즈 기준의 함수형태로 정의하였으며, 이것의 분포이론 및 추가판별변수(또는 추가판별변수군)의 판별력향상에 대한 유의성 검정법도 개발하였다.

$\{X_1, X_2\}$를 최적의 판별변수군으로 설정하고 행한 모의실험에서 새로이 제안한 베이즈 기준 B_{12}^I로 최적 부분집합선택법을 사용하여 변수선택을 한 결과, 표 4.1에서 **표기가 된 경우를 제외하고는 B_{12}^I가 *표기에서와 같이 $\{X_1, X_2\}$를 최적의 판별변수군으로 정확하게 판단하였다. 한편, **표기가 된 경우(즉, B_{12}^I가 최적의 판별변수군을 선택하지 못한 경우)는 $BR(x_p)$에 의해 추가 판별변수의 유의성을 (3.9)로 검정하였다. 이와 더불어 모의실험에서 계산한 기존의 변수선택 기준인 AER의 추정값과 제안한 베이즈 기준을 비교하면 두 선택기준들이 모두 $\{X_1, X_2\}$을 최적의 판별변수군으로 판단하고 있다. 이는 최적 부분집합선택법에 AER기준을 적용할 경우, 판별변수(p 개)에 의해 얻어지는 $2^p - 1$개의 부분집합에 대한 AER값의 추정에 교차타당성법이 모두 필요하여 계산이 복잡한 반면에, 폐쇄형(closed form)인 B_{12}^I기준은 판별변수선택에서 고려되는 $2^p - 1$개 부분 판별변수집합의 B_{12}^I만 계산하여 비교하면 간단히 최적의 판별변수군을 얻을수 있다. 이는 AER값의 추정에 필요한 복잡한 계산(특히 대표본이나 p 및 그룹의 수 K가 큰 경우)를 고려할 때 B_{12}^I기준은 최적 부분집합선택법에서 유용한 기준이 될 수 있음을 보여준다.

모의실험에서 함께 실시한 단계적선택법에서는 제안된 추가판별변수 선택기준 ($BR(X_p)$)가 모든 경우에서 $\{X_1, X_2\}$를 최적의 판별변수군으로 선택하고 있다(표 4.1의 P-값 참조). 그러므로, 제한적이지만 4장에서 행한 모의실험은 본 논문에서 제안한 베이즈기준이 MDA에서 판별변수를 효과적이고 정확하게 선택하고 있음을 보여주고 있다. 또한 이 기준은 일반성을 유지하고 있어서, $K = 2$로 대입하면 두 그룹 판별분석의 변수선택에 바로 적용된다.

본 논문에서는 동일한 분산-공분산행렬의 가정하에서 K개 정규모집단들에 대한 MDA에서 필요한 베이즈 판별변수선택기준에 대해서만 연구하였다. 그래서, 분산-공분산행렬들이 서로 다른 경우에 적용될 수 있는 베이즈 판별변수선택기준에 대한 연구도 시도해 볼 만한 가치가 있으며 이에 대한 것은 앞으로의 연구과제로 두었다.

참고문헌

[1] Aitkin, M. (1991). Posterior Bayes factors, *Journal of the Royal Statistical Society,* Ser. B, 53, 111-142.

[2] Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis,* John Wiley, New York .

[3] Berger, J. O. and Pericchi, L. R. (1996). The intrinsic Bayes factor for model selection and prediction, *Journal of the American Statistical Association,* Vol. 91, 109-122.

[4] Berger, J. O. and Sellke, T. (1987). Testing a point null hypothesis: The reconciliability of P-values and evidence, *Journal of the American Statistical Association,* Vol. 82, 112-122.

[5] DeGroot, M. H. (1970). *Optimal Statistical Decisions,* McGraw-Hill, New York.

[6] Evans, J. C. and Schwager, S. J. (1984). A test of additional accuracy for selecting groups of allocation variables, *Technometrics,* Vol. 36, 202-211.

[7] Geneshanandam, S. and Krzanowski, W. J. (1989). On selecting variables and assessing their performance in linear discriminant analysis, *Australian Journal of Statistics,* Vol. 32, 443-447.

[8] Good, I. J. (1947). *Probability and the Weighing of Evidence,* Haffner, New York.

[9] Jeffreys, H. (1935). Some tests of significance treated by the theory of probability. *Proceedings of the Cambridge Philosphy Society,* Vol. 31, 203-222.

[10] Jeffreys, H. (1961). *Theory of Probability,* Oxford University Press.

[11] Johnson, R. A. and Wichern, D. W. (1992). *Applied Multivariate Statistical Analysis,* Prentice Hall, New Jersey.

[12] Kass, R. E. and Raftery, A. E. (1995). Bayes factors, *Journal of the American Statistical Association,* Vol. 90, 773-795.

[13] Kim, H. J. (1995). On a balanced quadratic classification rule, *Communications in Statistics: Theory and Methods,* Vol. 24, 607-623.

[14] Kim, H. J. (1996). On a constrained optimal rule for classification with unknown prior individual group membership, *Journal of Multivariate Analysis,* Vol. 59, 166-186.

[15] Kshirsagar, A. M. (1972). *Multivariate Analysis,* Marcel Dekker, New York.

[16] Lee, P. M. (1988). *Bayesian Statistics: An Introduction,* John Wiely, New York.

[17] Mckay, R. J. (1977). Simultaneous procedures for variable selection in multiple discrimant analysis. *Biometrika,* Vol. 64, 283-290.

[18] McLachlan, G. J. (1992). *Discriminant Analysis and Statistical Pattern Recognition,* John Wiely, New York.

[19] O'Hagan, A. (1995). Fractional Bayes Factors for Model Comparisons, *Journal of the*

Royal Statistical Society, B, Vol. 57, 99-138.

[20] Rencher, A. C. (1995). *Methods of Multivariate Analysis,* John Wiley, New York.

[21] Spiegelhalter, D. J. and Smith, A. F. M. (1982). Bayes Factors for Linear and Log-linear Models with Vague Prior Information, *Journal of the Royal Statistical Society,* B, Vol. 44, 377-387.

[1998년 3월 접수, 1998년 6월 최종수정]

A Bayes Criterion for Selecting Variables in MDA

Hea-Jung Kim[1] Hee-Kyung Yoo[2]

ABSTRACT

In this article we have introduced a Bayes criterion for the variable selection in multiple disriminant analysis (MDA). The criterion is a default Bayes factor for the comparision of homo/heteroscadasticity of the multivariate normal means. The default Bayes factor is obtained from a development of the imaginary training sample method introduced by Spiegelhalter and Smith (1982). Based on the criterion, we also provided a test for additional discrimination in MDA. The advantage of the criterion is that it is not only applicable for the optimal subset selection method but for the stepwise method. Moreover, the criterion can be reduced to that for two-group discriminant analysis. Thus the criterion can be regarded as an unified alternative to variable selection criteria suggested by various sampling theory approaches. To illustrate the performance of the criterion, a numerical study has been done via Monte Carlo experiment.

1) Professor, Department of Statistics, Dongguk University, Seoul 100-715, Korea
2) Associate Professor, Department of Computer Science, Samchok National University, Samchok-Si, Kangwon Do 245-080, Korea