한국통계학회논문집 제2권 2호
1995년 12월 pp. 249-265

커널 판별분석의 오분류확률에 대한 붓스트랩 조정[1)]

백 장 선[2)]

요 약

본 논문에서는 확률분포가 알려져 있지 않은 두 모집단 중 어느 하나로 새로운 관측치를 분류할 때 오분류확률이 분석자에 의해 사전에 정해진 수준에 부합할 수 있도록 커널 판별함수의 임계치를 결정하였다. 정해진 오분류확률을 만족시키기 위한 판별함수의 임계치는 붓스트랩(bootstrap)기법을 판별함수에 적용시켜 계산된다. 본 논문에서 제시된 방법은 모집단에 대한 모수적 가정이 없으므로 어느 분포에도 적용가능하며, 모집단이 정규분포, 대수정규분포, 이산형과 연속형 변수가 혼합된 분포의 경우 모의실험을 통하여 그 성능에 대한 검증을 하였다.

1. 서론

두 모집단에 대한 고전적인 판별및 분류분석(discriminant and/or classification analysis)방법들은 주로 두 모집단의 확률밀도함수들의 비율에 근거하여 판별함수를 구성한다. 이러한 비율에 근거한 분류 방법(classification rule)은 총오분류확률(total probability of misclassification) 혹은 총오분류비용(total cost of misclassification)을 최소화시킨다는 의미에서 최적(optimal)이다(Welch,1939; Anderson,1984). 확률밀도함수들을 정규분포로 가정하고 필요한 모수를 표본으로부터 추정하여 새로운 관측치를 분류하기 위한 판별함수를 구성하는 것을 모수적 접근방법(parametric approaches)이라 한다. Fisher(1936)는 두 모집단의 공분산 행렬이 서로 같다는 가정하에서 선형판별함수(lenear discriminant function)를 맨 처음 제안하였다. 또한 Anderson(1958)과 John(1960,1963)은 가설검정방법에 의한 판별분석절차를 고안하였다. 두 공분산 행렬이 서로 다른 경우에는 이차판별함수(quadratic discriminant function)가 제안되었다(Seber,1984,p297; Anderson,1984,p235).

모수적 접근방법에 반하여 확률밀도함수의 형태에 대한 정보가 전혀없을 때, 자료로부터 비모수적 밀도함수추정(nonparametric density estimation)에 의하여 직접 그들의 확률밀도함수를 추정하여 판별식을 구성하는 비모수적 접근방법(nonparametric approaches)이 있다(Hand,1982). Remme, Habbema, and Hermans (1980)은 비모수적 방법과 고전적인 방법을 여러 모델에 적용하여 커널을 이용한 비모수적 방법이 최선 혹은 최선에 가까운 결과를 나타낸다고 밝혔다.

대부분의 분류방법은 두 확률밀도함수의 비율로 구성된 판별식이 임계치 c보다 크고 작음에 따라 새로운 관측치를 두 모집단 중의 한 모집단에 각각 속한다고 분류한다. 이 때 임계치 c 는 보통 각각의 모집단으로부터 관측치가 추출될 확률이나 오분류비용등에 의해 결정된다. 이처럼 구성

1) 이 논문은 1993년도 전남대학교 학술연구비에 의하여 연구되었음.
2) (500-757) 광주광역시 북구 용봉동 전남대학교 통계학과.

된 분류방법은 두가지 오분류확률들(misclassification probabilities)이 발생한다. 만약 두가지 오분류확4률들 중에서 어느 하나가 나머지보다 분석자에게 상대적으로 매우 중요한 의미를 가질 때, 그 오분류확률을 사전에 정해진 수준만큼 낮게 고정시켜줄 수 있는 임계치가 결정되어야 한다. 판별변수가 다변량 정규분포를 따르고 분산공분산 행렬이 동일한 경우 Anderson(1973)과 Kanazawa(1979)는 정해진 오분류확률을 만족시키는 임계치를 계산할 수 있도록 판별 함수의 점근분포를 유도하였다. 최근에 Baek, Gray and Woodward(1993)에 의해서 정규분포 이외의 임의의 모수적 분포에 대하여도 오분류확률을 조정할 수 있는 붓스트랩을 적용한 분류방법이 개발되었다.

본 연구의 목적은 모집단 분포의 형태에 대한 정보가 없는 경우, 비모수적 확률밀도함수 추정방법을 이용하여 관심 오분류확률을 분석자가 원하는 수준만큼 작게 조정할 수 있는 판별식을 개발하는데 있다. 두 확률밀도함수들은 커널 평활법(kernel smoothing technique)에 의하여 추정되며, 이렇게 추정된 두 확률밀도함수의 비율에 대하여 붓스트랩을 적용하여 임계치를 결정하게 된다. 다음절에서는 커널을 이용하여 판별함수를 추정하는 절차를 논의한 후 판별함수에 붓스트랩을 적용하여 오분류확률을 조정할 수 있는 임계치를 어떻게 결정하는가를 설명한다. 제 3절은 본 연구에서 제안된 방법을 세가지 경우에 적용한 것이다. 먼저 모수적 판별분석에서 흔히 가정하는 정규분포를 이용하여 제안된 비모수적 방법이 미리 설정된 오분류확률에 근사하는지 여러가지 표본크기에 따라 검사한다. 다음으로는 모수적 판별분석을 적용하기 부적합한 경우로서 모집단분포가 대수정규분포(lognormal distribution)일 때와 이산형과 연속형의 혼합분포일 때, 제안된 비모수적 방법의 성능을 각각 위와같이 측정하고, 검정력을 추정한다. 마지막으로 제 4절에서 결론을 기술한다.

2. 붓스트랩을 이용한 커널판별분석

$\{X_1,X_2,\cdots,X_n\}$, $\{Y_1,Y_2,\cdots,Y_m\}$이 각각 모집단 π_1과 π_2로부터 추출된 표본들일 때 새로운 관측치 Z가 π_1에 속하는지 π_2에 속하는지를 결정하는 Z에 대한 분류문제를 고려하자. 모집단 π_1의 확률밀도함수를 f_1, 모집단 π_2의 확률밀도함수를 f_2라 할 때, 여러가지 최적기준들(총오분류확률 또는 총오분류비용의 최소화 등)에 의한 분류기준은 두 확률밀도함수의 비율에 의해 정해진다. 즉 새로운 관측치 Z 에 대하여,

$$\frac{f_1(Z)}{f_2(Z)} > c \qquad (1)$$

이면, Z를 π_1에 속한다고 판단한다.

식(1)에 의해 표시된 분류기준은 Z가 속할 수 있는 두 개의 영역(region), R_1과 R_2로 나타낼 수 있다. 즉, $R_1=\{Z : f_1(Z)/f_2(Z)>c\}$, $R_2=\{Z : f_1(Z)/f_2(Z)\le c\}$이다. 이때 우리는 두가지 오류를 범할 수 있는데, 한가지는 $Z\in\pi_1$ 임에도 불구하고 $Z\in\pi_2$ 라고 잘못 분류하는 것이고, 나머지는 $Z\in\pi_2$임에도 $Z\in\pi_1$이라 잘못 분류하는 것이다. 각각의 오분류확률은 $P(2\mid 1)=\int_{R_2} f_1(Z)dZ$ 와 $P(1\mid 2)=\int_{R_1} f_2(Z)dZ$ 이다. 만약 이 두가지 오류 중 어느 한가지가

다른 것에 비하여 분석자에게 매우 중요한 의미를 가질 때, 그 중요한 오분류확률을 분석자가 원하는 수준만큼 낮게 고정할 필요가 있다. 예를들면 어느 치명적인 질병에 대하여 의사는 환자가 그 질병에 걸렸음에도 불구하고 그렇지않다고 잘못 진단할 확률을 가능한한 일정 수준으로 작게 고정시킬 수 있는 판별식을 원할 것이다.

2.1 커널을 이용한 판별함수 추정

판별식 (1)에 나타나 있는 확률밀도함수 f_i, $i=1,2$, 에 대하여 정규분포와 같은 모수적 가정을 사용하지 않고 단지 표본자료만을 이용하여 f_i를 추정하는 것을 비모수적 밀도함수 추정이라 한다 (Silverman,1986). 비모수적 밀도함수 추정방법은 여러가지가 있으나(kernel, splines, orthogonal series estimator) 본 연구에서는 가장 많이 쓰이는 커널방법을 사용하였다.

확률변수 X의 확률밀도함수가 $f(x)$로 표시될 때, $X=x$ 에서의 확률밀도함수 값, $f(x)$를 추정하기위해 표본 $\{X_1,X_2,\cdots,X_n\}$을 사용한다고 하자. 커널추정방법은 각 관측치 X_i들이 x로 부터 떨어져 있는 거리에 따라 추정치 $\hat{f}(x)$를 구하는데 공헌을 달리하며, 이러한 각 관측치들의 공헌도를 평균함으로써 $\hat{f}(x)$를 계산하는 방법이다. 구체적으로

$$\hat{f}(x;h)=\frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x-X_i}{h}\right).$$

이다. 여기서 $K(\cdot)$는 대칭인 임의의 확률밀도함수로서 커널(kernel)이라 불리며, h는 평활의 양을 좌우하는 평활모수(smoothing parameter,bandwidth)로서 $n\to\infty$에 따라 $h\to 0$ 와 $nh\to\infty$를 만족한다. 주로 많이 이용되는 커널로서는 Epanechnikov 커널 $K(u)=(3/4)(1-u^2)$, $|u|<1$ 와 Gaussian 커널 $K(u)=(1/\sqrt{2\pi})e^{-(1/2)u^2}$, $-\infty<u<\infty$ 등이 있다. $\hat{f}(x;h)$ 의 일치성(consistency)과 점근적 정규성(asymptotic normality)은 Parzen(1962)에 의해서 규명되었다. 자료로부터 최적평활모수를 결정하는 방법으로는 최소제곱교차타당성 방법(least squares cross validation; Rudemo, 1982, Bowman, 1984), 편의교차타당성 방법(biased cross validation), 평활교차타당성 방법(smoothed cross validation), 대입법(plug-in method) 등이 있으며 Park(1991), Cao, Cuevas and Manteiga(1994)에 잘 정리되어 있다. 본 연구에서는 최소제곱교차타당성 방법을 이용하여 최적평활모수를 추정한다.

커널평활법은 다변량 확률변수에도 사용가능하다. $\mathbf{X}_i$들이 p차원 연속형 관측치들일때 $\mathbf{X}=\boldsymbol{x}$ 에서의 $f(\boldsymbol{x})$는 다음과 같이 추정된다.

$$\hat{f}(\boldsymbol{x};h)=\frac{1}{nh^p}\sum_{i=1}^{n}K_1\left(\frac{\boldsymbol{x}-\mathbf{X}_i}{h}\right). \qquad (2)$$

여기서 $K_1(\cdot)$는 p차원 커널(p-dimensional kernel) 이다. 다변량 커널 추정치에 대한 점근적 성질은 Cacoullos(1966)에 잘 나와 있다.

새로운 관측치 Z가 Z=z 로 관측되었을 때 판별식 $T(z)=f_1(z)/f_2(z)$ 를 추정하기위하여 먼

저 $f_1(z)$와 $f_2(z)$를 커널평활법을 적용하여 다음과 같이 추정한다.

$$\hat{f}_1(z;h_1)=\frac{1}{nh_1}\sum_{i=1}^{n}K\left(\frac{z-X_i}{h_1}\right),$$

$$\hat{f}_2(z;h_2)=\frac{1}{nh_2}\sum_{i=1}^{m}K\left(\frac{z-Y_i}{h_2}\right).$$

판별식 $T(z)$에 대한 추정치는 $\hat{f}_1(z;h_1)$과 $\hat{f}_2(z;h_2)$ 의 비율로서 다음과 같다.

$$\hat{T}(z)=\frac{\hat{f}_1(z;h_1)}{\hat{f}_2(z;h_2)} \tag{3}$$

2.2 붓스트랩에 의한 임계치 결정

판별식에 대한 추정치가 위와 같이 구해지면 z에 대한 분류절차는 다음과 같다. 어떤 상수 c에 대하여 $\hat{T}(z) > c$ 이면 $z\in\pi_1$으로 분류하고, 그렇지 않으면 $z\in\pi_2$로 분류한다. 그런데 $z\in\pi_1$임에도 불구하고 $z\in\pi_2$라고 잘못 분류할 오분류확률 $P(2|1)$은 바로 $P(\hat{T}(z)\le c \mid z\in\pi_1)$이다. 마찬가지로 $P(1|2) = P(\hat{T}(z)>c \mid z\in\pi_2)$ 임을 알 수 있다. 그러므로 $z\in\pi_i$에 따른 $\hat{T}(z)$의 분포만 안다면 원하는 수준의 오분류 확률에 대응하는 c를 결정할 수 있다. f_i들에 대한 어떠한 분포적 가정을 하지 않았으므로 $\hat{T}(z)$에 대한 정확한 분포를 구하는 것은 무리다. 본 연구에서는 $\hat{T}(z)$에 대한 분포를 붓스트랩 방법(Efron 1979, 1982)에 의해 근사적으로 구하여 원하는 수준의 오분류확률을 달성할 수 있는 판별식의 임계치 c를 다음과 같이 결정한다.

조정하려는 오분류 확률이 $P(2|1)$이고 이것을 α $(0<\alpha<1)$로 고정하려고 한다. $P(2|1)=\alpha$ 를 만족하는 임계치 c는 먼저 반복해서 추출된 붓스트랩 표본들을 이용하여 $z\in\pi_1$인 상황하의 커널 판별통계량 $\hat{T}(z)=\hat{f}_1(z;h_1)/\hat{f}_2(z;h_2)$ 의 분포를 근사적으로 추정하고, 그 추정된 분포로 부터 결정된다. 모집단 분포 f_1으로부터 z와 크기 n인 독립표본을 추출하고, 다른 모집단분포 f_2로부터 크기 m인 독립표본을 추출하여 커널판별 통계량 $\hat{T}(z)$값을 계산하는 것을 여러번 반복함으로써 $z\in\pi_1$하에서의 커널 판별통계량 분포를 근사적으로 구할 수 있을 것이다. 물론 분포 f_i들의 형태에 대한 정보가 없는 상황이므로 완벽하게 독립표본들을 각 분포로부터 생성할 수는 없지만 미지의 분포 f_i에 대한 정보를 담고있는 확률표본 $\{X_1,X_2,\cdots,X_n\}$과 $\{Y_1,Y_2,\cdots,Y_m\}$으로부터 각각 무작위 반복 추출에 의해서 단순 붓스트랩 표본을 추출함으로써 그것으로 원하는 독립표본을 가름할 수 있겠다. 이렇게 단순한 붓스트랩 방법은 모수적 가정을 제거하는 장점도 있으나 원래 표본에 나타나 있지 않는 관측치는 결코 붓스트랩 표본에 포함되지 않게되어 만약 원래 표본이 왜곡된 구조를 가지고 있는 경우에는 그러한 왜곡된 구조가 계속해서 재생될 위험도 있게 된다. Silverman(1986)은 이러한 단점을 극복하기위해 원래표본을 이용하여 미지의 확률밀도함수 f_i

에 대한 비모수적 추정치 $\hat{f}_i$를 구하여 분포구조를 파악한 후 $\hat{f}_i$로부터 붓스트랩에 의한 독립표본을 추출하도록 제안하였다. 비음(nonnegative)의 커널 K를 이용해서 밀도함수를 추정하여 독립표본을 얻고자 할 때, 실제로 $\hat{f}_i$를 계산할 필요는 없으며 그것을 가능케하는 붓스트랩 표본추출방법은 다음과 같다. 확률표본 $\{X_1, X_2, \cdots, X_n\}$으로부터 평활모수 h를 이용한 커널 K를 사용하여 크기 n인 붓스트랩표본 $\{X_1^*, X_2^*, \cdots, X_n^*\}$을 추출하는 절차는 다음과 같다 :

(i) $\{1,2,\cdots,n\}$으로부터 무작위 복원 추출에 의해 I를 선택.
(ii) 확률밀도함수 K로부터 난수 ε 발생.
(iii) $X^* = X_I + h\varepsilon$.
(iv) (i) - (iii)을 n번 반복.

원래의 표본 $\{X_1, X_2, \ldots, X_n\}$ 과 $\{Y_1, Y_2, \ldots, Y_m\}$ 로부터 위의 절차에 따라 각각 붓스트랩 표본 $\{X_1^*, X_2^*, \ldots, X_{n+1}^*\}$과 $\{Y_1^*, Y_2^* \ldots, Y_m^*\}$을 구성한다. z, $\{X_1, X_2, \ldots, X_n\}$, $\{Y_1, Y_2, \ldots, Y_m\}$ 대신 각각 X_{n+1}^*, $\{X_1^*, X_2^*, \ldots, X_n^*\}$, $\{Y_1^*, Y_2^* \ldots, Y_m^*\}$을 사용하여 식 (3)으로 부터 붓스트랩 표본에 대한 $\hat{T}(z)$를 계산한다. 이것을 $\hat{T}^*$라 표기하자. 이러한 절차를 독립적으로 B번 반복해서 매번 $\hat{T}^*$를 계산하면 $\{\hat{T}_1^*, \hat{T}_2^*, \ldots, \hat{T}_B^*\}$를 얻게 되며, 이것을 이용하여 $z \in \pi_1$일 때의 $\hat{T}(z)$의 분포를 근사적으로 얻을 수 있다. 구체적으로 말하면 $\{\hat{T}_1^*, \hat{T}_2^*, \ldots, \hat{T}_B^*\}$의 경험적 α 백분위수(αth empirical quantile), c_α^*는 n과 m이 큰 수이고 $B \to \infty$이면 $z \in \pi_1$ 일때의 $\hat{T}(z)$의 진실된 α 백분위수 c에 근사된다 (붓스트랩의 백분위과정에 대한 근사이론은 Bickel and Freedman(1981) 참조.) 그러므로 근사적으로 구해진 c_α^*를 이용하여 z를 분류한다. 즉 $\hat{T}(z) > c_\alpha^*$ 이면 z를 π_1 으로 분류하게 되며 이 절차는 오분류 확률 $P(2|1) = \alpha$ 를 근사적으로 충족시킨다.

3. 오분류확률에 대한 붓스트랩 조정 방법의 적용

판별변수 Z는 그 형태가 이산형, 연속형, 혹은 이산형과 연속형의 혼합형으로서, i번째 모집단에 대한 확률밀도함수 혹은 확률질량함수 f_i 를 따른다, i=1, 2. 첫번째와 두번째 모집단으로부터 각각 그에 대한 확률표본 $\{X_1, X_2, \cdots, X_n\}$과 $\{Y_1, Y_2, \cdots, Y_m\}$이 관측되었다고 하자. 오분류확률을 주어진 수준에 맞추어 분류분석을 행하고자 할 때 다음의 세가지 일반적인 경우에 대하여 각각 본 연구에서 제안한 붓스트랩을 이용한 커널판별기법을 적용한다. 세가지 경우 모두 대표본에 대하여 정해진 수준의 오분류확률을 달성할 수 있음을 확인하게 된다. 판별변수 Z가 다변량 변수일 때에도 다변량 커널을 이용, 마찬가지로 붓스트랩 방법에 의해 오분류확률을 조정할 수 있

으나 시간적 절약을 위해 일변량인 경우로 한정하여 모의실험을 수행하였다.

3.1 동일분산을 가진 두 개의 정규분포 모집단

두 모집단이 각각 정규분포 $N(\mu_1,\sigma_1^2)$ 와 $N(\mu_2,\sigma_2^2)$ 를 따른다고 할 때 $f_1(\cdot;\mu_1,\sigma_1^2)$와 $f_2(\cdot;\mu_2,\sigma_2^2)$를 각각의 확률밀도함수라고 하자. 만약 $\sigma_1^2=\sigma_2^2=\sigma^2$ 이면 판별식 $f_1(z;\mu_1,\sigma^2)/f_2(z;\mu_2,\sigma^2)$ 에 미지의 모수 μ_1, μ_2, σ^2 에 대한 추정치들을 대입하므로써 모수적 다변량 판별분석 방법중 하나인 Anderson의 W 통계량을 얻게 된다. 즉 μ_1과 μ_2에 대한 추정치는 각각 $\overline{X}=\sum_{i=1}^{n}X_i/n$, $\overline{Y}=\sum_{i=1}^{m}Y_i/m$ 이며 $S_1^2=\sum_{i=1}^{n}(X_i-\overline{X})^2/(n-1)$, $S_2^2=\sum_{i=1}^{m}(Y_i-\overline{Y})^2/(m-1)$ 이라 하면 동일 분산 σ^2에 대한 추정치는 $S_p^2=((n-1)S_1^2+(m-1)S_2^2)/(n+m-2)$ 이다. 따라서 일변량인 경우 Anderson 의 W 통계량은

$$W = \frac{\{ z-(1/2)(\overline{X}+\overline{Y})\}(\overline{X}-\overline{Y})}{S_p^2}$$

이다.

α를 조정하고자 하는 오분류확률 $P(2|1)$의 목표수준이라 하자. Anderson(1973)은 $z\in\pi_1$일때 통계량 W의 점근적 정규분포를 유도하여 그로부터 근사적 임계치 W_α를 얻었다. $D=\sqrt{(\overline{X}-\overline{Y})^2/S_p^2}$ 이고 $N(0,1)$의 누적분포함수 $\Phi(\cdot)$ 에 대하여 $\Phi(u_0)=\alpha$인 α백분위수 u_0에 대하여 표본크기 n,m 이 충분히 크면 일변량인 경우

$$W_\alpha = D\left[\frac{1}{2}D+u_0\left\{\frac{2n+1}{2n}+\frac{1}{4(n+m-2)}(1+u_0^2)\right\}\right]$$

이다. Kanazawa(1979)은 다른 모수적 판별통계량의 하나인 John의 Z 통계량에 대하여 역시 오분류확률을 조정할 수 있도록 근사적 임계치를 얻었다.

판별통계량의 점근분포를 유도하지 않고서도 2.2에 기술된 바와같이 커널밀도함수추정에 의한 판별통계량 $\hat{T}(z)$의 임계치를 붓스트랩에 의해서 근사적으로 결정할 수 있다. 식 (3)의 통계량 $\hat{T}(z)$를 계산하기위해서는 $\hat{f}_i(z;h_i)$의 계산이 선행되어야 하며 이를 위해서 우리는 Epanechnikov커널과 최소제곱교차타당성에 의해서 결정된 최적 평활모수를 사용하였다. 표본 $\{X_1,X_2,\cdots,X_n\}$, $\{Y_1,Y_2,\cdots,Y_m\}$으로부터 각각 붓스트랩 표본을 추출할 때 사용된 커널이 Epanechnikov 커널인 경우 2.2의 붓스트랩표본 추출절차 (ii)를 다음과 같이 비교적 간단하게 수행할 수 있다 (Devroye and Györfi, 1985.)

(ii)′ 균일분포(-1,1)로부터 세개의 난수 u_1, u_2, u_3를 발생시킨 후, 만약 $|u_3|\geq|u_2|$ 이고 $|u_3|\geq|u_1|$ 이면 $\varepsilon=u_2$ 로 하고, 그렇지 않으면 $\varepsilon=u_3$ 로 한다.

또한 관측 표본의 크기가 크면 표본적률로 모집단 분포의 적률을 추정할 수 있으므로 생성되는 붓스트랩 표본도 관측표본과 동일한 제 1차, 제 2차 적률을 갖도록 붓스트랩 표본을 추출할 수 있다. 이를위해서는 단계 (iii)을

$$(iii)'\quad X^* = \overline{X} + (X_I - \overline{X} + h\varepsilon)/(1 + h^2\sigma_k^2/S_x^2)^{1/2}$$

으로 대체해야하며, 이 때 $\overline{X}$와 S_x^2은 각각 관측표본의 평균과 분산이고, σ_k^2는 커널 K의 분산이다.

주어진 수준 α에 대하여 통계량 W와 $\hat{T}$ 를 그들 각각의 근사적 임계치 W_α, c_α^*와 함께 사용했을 때, 모집단 π_1으로 분류될 새로운 관측치를 모집단 π_2로 잘못 분류할 오분류확률을 각각 $P_W(2|1)$, $P_{\hat{T}}(2|1)$이라 하자. 따라서 $P_W(2|1)=P(W\le W_\alpha|\pi_1)$이며 $P_{\hat{T}}(2|1)=P(\hat{T}\le c_\alpha^*|\pi_1)$ 이다. 이제 동일분산을 가진 두개의 정규분포 모집단에 대하여 $P_W(2|1)$과 $P_{\hat{T}}(2|1)$이 미리 설정된 희망 오분류확률 $\alpha=P(2|1)$ 에 과연 얼마나 가까운지 모의실험을 통해서 검사해 보기로 한다.

$N(\mu_1,\sigma_1^2)$ 와 $N(\mu_2,\sigma_2^2)$ 로부터 확률표본 $\{z_i,X_{i1},X_{i2},\cdots,X_{in}\}_{i=1}^M$ 과 $\{Y_{i1},Y_{i2},\cdots,Y_{im}\}_{i=1}^M$ 을 추출한다. 이때 $\mu_1=0$, $\sigma_1^2=\sigma_2^2=1$ 을 가정하였고 세가지 다른 경우의 $\mu_2=1, 2, 3$ 을 시도하였다. 각각의 $i=1,2,\cdots,M$ 에 대하여 $\{z_i,X_{i1},X_{i2},\cdots,X_{in}\}$과 $\{Y_{i1},Y_{i2},\cdots,Y_{im}\}$을 사용하여 통계량 W와 $\hat{T}$의 계산 값 W_i와 $\hat{T}_i$를 먼저 구한 후 그것들을 고정된 α에 대응하는 각각의 임계치 $W_{i\alpha}, c_{i\alpha}^*$와 비교한다. 이때 $c_{i\alpha}^*$를 계산하기 위하여 B=499개의 붓스트랩 표본이 사용되었다. 이렇게 행해진 M번의 시행에 대한 각각의 통계량 값이 그들의 임계치보다 작거나 같은 경우의 비율로서 $P_W(2|1)$과 $P_{\hat{T}}(2|1)$를 추정한다. 따라서 $P_W(2|1)$과 $P_{\hat{T}}(2|1)$에 대한 추정치는 각각 $\hat{P}_W(2|1) = \sum_{i=1}^{M} I(W_i\le W_{i\alpha})/M$, $\hat{P}_{\hat{T}}(2|1) = \sum_{i=1}^{M} I(\hat{T}_i\le c_{i\alpha}^*)/M$ 이며, 이때 $I(\cdot)$는 지수함수 (indicator function)이다. $\hat{P}_W(2|1)$은 비율에 대한 추정치이므로 그것의 표준편차는 $\sqrt{\hat{P}_W(2|1)(1-\hat{P}_W(2|1))/M}$ 로 추정할 수 있다. $\hat{P}_{\hat{T}}(2|1)$에 대한 표준편차 역시 비슷하게 추정될 수 있다. 표 1의 처음 부분에는 $\alpha=0.05$, M=1000일 때 모수적 판별통계량 W와 붓스트랩을 이용한 비모수적 커널 판별통계량 $\hat{T}$를 이용한 경우 오분류확률에 대한 추정치와 그것의 표준편차에 대한 추정치가 각기 다른 표본크기에 따라 계산되어 있다. 표본크기가 n=m=75, n=m=100, n=m=200 인 경우 모두 붓스트랩에 의한 커널판별분석방법이 모수적 방법에 비하여 손색없이 목표 오분류확률 $\alpha=0.05$ 를 근사적으로 달성하고 있음을 알 수 있다.

다음으로는 $P(2|2)$ 를 W와 $\hat{T}$에 대하여 비교해 보자. $P(2|2)$는 π_2에서 나온 새로운 관측치가 π_2로 올바르게 분류될 정분류확률(correct classification probability)이다. 앞에서 가정한 같은

값의 모수 $\mu_1, \sigma_1^2, \mu_2, \sigma_2^2$를 사용하여 확률표본 $\{X_{i1}, X_{i2}, \cdots, X_{in}\}_{i=1}^{M}$ 과 $\{z_i, Y_{i1}, Y_{i2}, \cdots, Y_{im}\}_{i=1}^{M}$ 을 각각 $N(\mu_1, \sigma_1^2)$, $N(\mu_2, \sigma_2^2)$ 으로부터 추출한다. $P_W(2|2)$와 $P_T(2|2)$에 대한 추정치 $\hat{P}_W(2|2)$, $\hat{P}_T(2|2)$는 $\hat{P}_W(2|1)$와 $\hat{P}_T(2|1)$을 얻는 방법과 같은 방식으로 계산된 것이다. $\alpha=0.05$, $n=m=200$, $M=1000$, $B=499$ 에 대하여 정분류확률에 대한 추정치들이 표 1의 두번째 부분에 나와있으며 π_2가 π_1에서 멀리 떨어져 있을수록 $\hat{P}_W(2|2)$와 $\hat{P}_T(2|2)$가 증가함으로써 π_2에 속해있는 새로운 관측치 z를 π_2로 정확하게 분류함을 나타낸다. $\mu_2=1, 2, 3$인 경우 모두 $\hat{P}_W(2|2)$와 $\hat{P}_T(2|2)$가 서로 비슷한 값을 보여주며 이는 모수적 분포의 형태에 대한 정보를 전혀 이용하지 않은 비모수적 분류방법이 대표본인 경우 모수적 방법에 비하여 성능면에서 하등 뒤지지않음을 보여준다.

표1. 오분류확률 $P(2|1)=0.05$ 에 대한 추정치와 정분류확률 $P(2|2)$에 대한 추정치.

	μ_2		
	1	2	3
		n = m = 75	
$\hat{P}_W(2\|1)$	0.054	0.045	0.045
	(0.007)	(0.007)	(0.007)
$\hat{P}_T(2\|1)$	0.051	0.044	0.044
	(0.007)	(0.006)	(0.006)
		n = m = 100	
$\hat{P}_W(2\|1)$	0.061	0.062	0.055
	(0.008)	(0.008)	(0.007)
$\hat{P}_T(2\|1)$	0.055	0.059	0.048
	(0.007)	(0.007)	(0.007)
		n = m = 200	
$\hat{P}_W(2\|1)$	0.046	0.050	0.053
	(0.007)	(0.007)	(0.007)
$\hat{P}_T(2\|1)$	0.052	0.056	0.052
	(0.007)	(0.007)	(0.007)
		n = m = 200	
$\hat{P}_W(2\|2)$	0.253	0.658	0.909
	(0.014)	(0.015)	(0.009)
$\hat{P}_T(2\|2)$	0.252	0.655	0.904
	(0.014)	(0.015)	(0.009)

3.2 두개의 대수정규분포 모집단

판별및 분류분석이 적용되는 두 모집단 자료가 분산이 동일하든지 그렇지 않든지간에 정규분포를 따르지 않는다면 정규분포 가정하에서 유도된 모수적 판별및 분류기법들은 그 신뢰성이 감소

될 수 밖에 없다. 정규분포를 따르지않는 모집단 분포로서 그 분포가 대칭적이지 못하고 한쪽으로 기울어진 분포인 대수정규분포일 때 과연 커널 밀도함수추정에 의한 판별함수에 붓스트랩 방법을 적용하여 오분류확률을 조정할 수 있는지와, 또한 분포의 기울어진 정도가 그리 심하지 않은 상황에서 비모수적 방법과 Anderson 판별함수의 근사적 임계치를 적용한 경우를 간단한 모의 실험을 통하여 비교하고자 한다.

커널을 이용하여 계산된 $X=x$ 에서의 밀도함수 추정치 $\hat{f}(x)$는 $X=x$ 로부터 평할모수 h 이내에 속한 관측치 X_i 와 x와의 차이들에 대한 일종의 가중평균치이다. 만약 동일한 h를 모든 x에서의 밀도함수 추정치에 적용한다면 저밀도의 x에서는 가용 관측치들이 얼마되지않아 매끈한 추정치를 얻기 힘들다. 따라서 대수정규분포와 같이 기울어진 분포인 경우 밀도가 높은 곳에서는 작은 h를, 낮은 곳에서는 상대적으로 큰 h를 적용하며 밀도함수를 추정하는 적응커널추정방법 (adaptive kernel estimation method; Silverman,1986) 이 바람직하다. 확률표본 $\{X_1,X_2,\cdots,X_n\}$ 을 이용하여 $X=x$에서의 밀도함수추정을 위한 일반적인 적응커널추정방법은 다음과 같다 :

(v) 모든 $i=1,2,\cdots,n$ 에 대하여 $\tilde{f}(X_i)>0$을 만족하는 예비추정치(pilot estimate) $\tilde{f}(x)$를 구한다.

(vi) $\tilde{f}(X_i)$의 기하평균을 g라 할 때 (즉, $\log g=(1/n)\sum\log\tilde{f}(X_i)$), 국소평할모수요인 (local bandwidth factor) λ_i를 $\lambda_i = \{\tilde{f}(X_i)/g\}^{-\alpha}$ 로 계산한다. 이 때 α는 민감모수 (sensitivity parameter)로서 $0\le\alpha\le1$ 이다.

(vii) 적응커널추정치 $\hat{f}(x)$는 $\hat{f}(x) = (1/n)\sum_{i=1}^{n}(1/(h\lambda_i))K\{(x-X_i)/(h\lambda_i)\}$ 로 정의된다. 이 때 K는 커널함수이고 h는 평할모수이다.

Breiman, Meisel and Purcell (1977) 과 Abramson(1982)에 의하면 적응커널추정방법은 예비추정치에 그리 민감하지 않으므로 시간이 많이 소요되는 교차타당성과 같은 방법에 의해 계산된 평할모수로 예비추정치를 계산하는 것보다는 참고분포(reference distribution)등을 이용하여 (Silverman, 1986) 간단히 계산할 수 있다. 적응커널추정방법은 민감모수의 값이 만약 $\alpha=0$ 이면 동일한 평할모수 h를 사용하는 커널추정방법과 같게되고, α가 클수록 적응커널추정방법은 예비추정치들의 변화에 민감하게 반응하게 된다. $\alpha=1/2$ 의 민감모수를 사용했을 때 대체로 만족할 만한 커널추정치를 얻을 수 있으며 이에 대한 이론 및 실제적 근거는 Abramson(1982)에 나와있다.

첫번째 모집단 π_1이 대수정규분포 $LOGN(\mu,\sigma^2)$ 를 따르고 두번째 모집단 π_2는 π_1으로부터 u만큼 $(u>0)$ 이동하여 동일하게 분포하고 있다고 하자. 즉 각각의 확률밀도함수 f_1와 f_2는 $x\in\pi_1$에 대하여 $f_1(x;\mu,\sigma^2) = (1/(x\sigma\sqrt{2\pi}))\exp\{-(\ln x-\mu)^2/2\sigma^2\}$, $0<x<\infty$ 이며, $y\in\pi_2$ 에 대하여 $f_2(y;u,\mu,\sigma^2) = (1/((y-u)\sigma\sqrt{2\pi}))\exp\{-\ln(y-u)-\mu)^2/2\sigma^2\}$, $u<y<\infty$이다. 모집단이 정규분포가 아닌 대수정규분포를 따르는 경우에도 붓스트랩을 이용하여 커널 판별식의 오분류확률

을 조정할 수 있음을 확인하기 위하여 3.1에서 행했던 바와 같이 f_1과 f_2로 부터 각각 독립 표본들을 여러번 추출하여 $\hat{P}_{\gamma}(2|1)$을 구한 후 정해진 오분류확률 α에 근접하는지 살핀다. 또한 동일한 표본들을 이용하여 Anderson 통계량 W의 $P(2|1)$에 대한 추정치 $\hat{P}_{W}(2|1)$를 계산하고 $\hat{P}_{\gamma}(2|1)$와 비교한다. $f_1(x;\mu,\sigma^2)$로부터 $\{ z_i, X_{i1}, X_{i2}, \cdots, X_{in} \}_{i=1}^{M}$를 $f_2(y;u,\mu,\sigma^2)$로부터 $\{ Y_{i1}, Y_{i2}, \cdots, Y_{im} \}_{i=1}^{M}$을 추출한다. 이때 $\mu=0$, $\sigma^2=(0.5)^2$를 가정하였고 u는 $u=0.5, 1.0, 1.5, 2.0$을 선택하였다. $i=1,2,\cdots,M$ 에 대하여 $\{ z_i, X_{i1}, X_{i2}, \cdots, X_{in} \}$과 $\{ Y_{i1}, Y_{i2}, \cdots, Y_{im} \}$을 사용하여 커널 판별 통계량 $\hat{T}_i(z_i)=\hat{f}_1(z_i)/\hat{f}_2(z_i)$ 을 계산할 때 위에 설명된 적응커널추정방법을 적용하여 $\hat{f}_1$과 $\hat{f}_2$를 구한다. 이때 적응커널추정방법의 단계 (v)에서 참고분포를 이용하여 예비 추정치를 구할 때, $A=\min$(표준편차,사분위수/1.34)라 하면 예비추정치 $\tilde{f}(x)$를 위한 평할모수 h는 $h=0.9An^{-1/5}$ 로 (이 때 n은 표본 크기이다) 계산되었으며, 이렇게 계산된 h는 t-분포, 대수정규분포, 혼합정규분포등 비정규분포의 밀도함수추정에 유효하다(Silverman, 1984, pp45-48.) 단계 (vi)의 민감모수 α는 $\alpha=1/2$ 이 사용되었고 단계 (vii)에서 적용된 평할모수는 계산의 절약을 위해 단계(v)에서 구해진 h값을 그대로 이용하였다. 이렇게 계산된 $\hat{T}_i$와 비교하기 위한 붓스트랩 임계치 c^*_{α}를 결정하기 위해 2.2의 절차 (i) - (iv)를 따라 붓스트랩 표본을 B개 추출한다. 그러나 밀도함수 추정이 적응커널추정방법에 따라 행해졌으므로 (각 관측치마다 다른 평할모수 $h\lambda_i$가 사용됨), 2.2의 절차(iii)은

(iii)'' $X^* = X_I + h\lambda_I\varepsilon$

으로 대체되어야 한다. 각 표본으로 부터 이렇게 추출된 붓스트랩 표본을 $\{ z^*_{ij}, X^*_{ij1}, X^*_{ij2}, \cdots, X^*_{ijn} \}_{j=1}^{B}$, $\{ Y^*_{ij1}, Y^*_{ij2}, \cdots, Y^*_{ijm} \}_{j=1}^{B}$ 라 하자. 추출된 붓스트랩 표본은 원래의 표본처럼 기울어진 분포특성을 갖고 있을 것이므로 역시 적응커널추정방법에 의해 각 붓스트랩 표본에 대한 밀도함수 추정치 $\hat{f}_1(z^*_{ij})$, $\hat{f}_2(z^*_{ij})$ 를 계산한 후 판별함수 추정치 $\hat{T}^*_j = \hat{f}_1(z^*_{ij})/\hat{f}_2(z^*_{ij})$ 를 구한다. 이때 붓스트랩 표본의 각 관측치에 대한 적응커널추정방법의 평할모수는 붓스트랩 표본추출을 (iii)'' 와 같이 추출한 이상 또다시 독립적으로 계산할 필요없이 X^*를 구성하는 X_I에 대응한 $h\lambda_I$를 그대로 사용하면 된다. 그림 1 에는 $n=m=400$ 의 확률표본을 $f_1(x;\mu=0,\sigma^2=(0.5)^2)$과 $f_2(y;u=1.5,\mu=0,\sigma^2=(0.5)^2)$ 로부터 독립적으로 추출한 후 구간 (0,4)과 (1.5, 5.5)내에 각각 속한 200개의 등간격 x, y 점들에 대한 적응커널추정방법에 의한 밀도함수추정치 $\hat{f}_1$, $\hat{f}_2$ 들과, 각 확률표본들로 부터 추출한 붓스트랩 표본들에 대하여 위의 평할모수 $h\lambda_I$를 적용하여 계산한 밀도함수 추정치 $\hat{f}^*_1$, $\hat{f}^*_2$ 들이 진실된 밀도함수 f_1, f_2 와 함께 그려져 있다. $n=m=400$의 대표본이므

로 j=1, 2에 대하여 $\hat{f}_j^*$는 $\hat{f}_j$와 매우 비슷하며 또한 그들 모두 f_j의 분포특성을 잘 표현하고 있다. 따라서 i번째 두 표본 $\{X_{i1}, X_{i2}, \cdots, X_{in}\}$, $\{Y_{i1}, Y_{i2}, \cdots, Y_{im}\}$에 대한 커널판별통계량 $\hat{T}_i = \hat{f}_1/\hat{f}_2$의 임계치 $c_{i\alpha}^*$는 B번 추출된 붓스트랩 표본에 의해서 계산된 $\{\hat{T}_1^*, \hat{T}_2^*, \cdots, \hat{T}_B^*\}$의 경험적 α 백분위수로 결정된다.

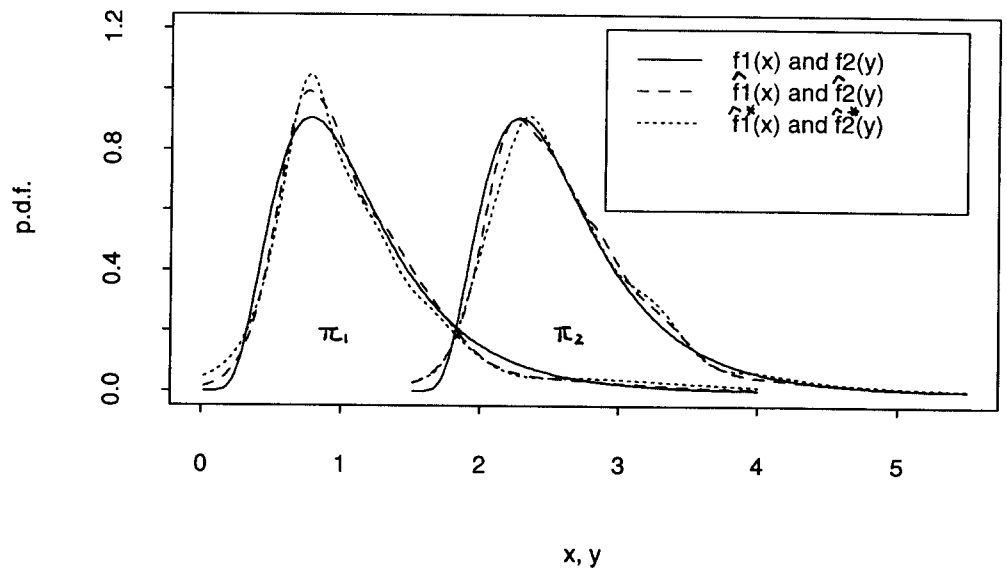

그림 1. 확률밀도함수 $f_1(x)$, $f_2(y)$와 적응커널추정방법에 의한 밀도함수추정치 $\hat{f}_1(x)$, $\hat{f}_2(y)$, 그리고 붓스트랩표본에 의한 밀도함수추정치 $\hat{f}_1^*(x)$, $\hat{f}_2^*(y)$.

$\hat{P}_T(2|1)$과 $\hat{P}_W(2|1)$을 3.1에서와 같이 통계량 $\hat{T}$과 W에 의해 M번 시행중 잘못분류한 비율로 표시된, 고정된 오분류확률 $P(2|1)=\alpha$에 대한 각각의 추정치라 하자. $\alpha=0.5$, $M=1000$, $B=499$ 일때 각 통계량에 대한 오분류확률 추정치들이 $u=0.5$, 1.0, 1.5, 2.0 에 대하여 표 2의 처음부분에 나타나 있다. 각 u에 대하여 $\hat{P}_T(2|1)=0.044, 0.059, 0.052, 0.050$ 으로서 $H_0 : P_T(2|1)=\alpha$ 를 기각할 수 없으며 따라서 커널판별통계량 $\hat{T}$에 의하여 원하는 오분류확률 $P(2|1)=\alpha$ 가 달성되었다고 볼 수 있다. 반면에 Anderson 통계량 W는 u = 0.5, 1.0, 2.0 인 경우 $\hat{P}_W(2|1)$ = 0.068, 0.064, 0.073 으로서 $Z = |(\hat{P}_W(2|1)-\alpha)/\sqrt{\alpha(1-\alpha)/1000}| > 1.96$ 이므로 유의수준 5%하에서 $H_0 : P_W(2|1)=\alpha$ 는 기각된다. $\hat{T}$에 의해 $z \in \pi_2$ 인 z를 π_2로 바르게 분류할 정분류확률에 대한 추정치 $\hat{P}_T(2|2)$가 표 2의 밑부분에 계산되어 있다. π_2가 π_1으로 부터 멀

리 떨어져 있을수록 바르게 분류함을 알 수 있다.

표 2. 대수정규분포 모집단일 때 오분류확률 $P(2|1)=0.05$ 에 대한 추정치와 정분류확률 $P(2|2)$에 대한 추정치

	u			
	0.5	1.0	1.5	2.0
$\hat{P}_T(2\|1)$	0.044 (0.006)	0.059 (0.007)	0.052 (0.007)	0.050 (0.007)
$\hat{P}_W(2\|1)$	0.068 (0.008)	0.064 (0.008)	0.055 (0.007)	0.073 (0.008)
$\hat{P}_T(2\|2)$	0.097 (0.009)	0.300 (0.014)	0.700 (0.014)	0.979 (0.005)

3.3 이산형과 연속형 변수의 혼합

판별분석이 적용되는 다변량 관측변수중 일부분의 변수는 이산형 혹은 범주형인 경우가 많이 있다. 이처럼 관측변수가 이산형이거나 이산형과 연속형 변수의 혼합형인 경우에도 비모수적인 커널판별분석이 확장 적용될 수 있다.

Aitchison and Aitken(1976)은 각 요소변수가 0 혹은 1 값을 갖는 다변량 이진자료(multivariate binary data)에 대하여 그 확률분포를 추정할 수 있는 커널을 제시하였다. 길이가 k인 다변량 이진확률벡터의 분포는 2^k개의 가능한 결과들에 대한 확률로서 기술된다. 만약 관측변수의 갯수 k가 10이상인 경우에는 관측자료의 수가 아주 많지않은 이상 특정 결과에 대한 확률을 그 결과가 전체자료중에서 발생하는 비율로 정의되는 최우추정치에 의해서 계산할 수가 없다. 다변량 이진관측자료의 공간 $\{0,1\}^k$를 B^k라 할 때, B^k에 속한 관측치들을 평활함으로써 표본에 관측되지 않은 결과에 대한 확률도 커널을 이용하여 추정하게 된다. B^k에 속해있는 다변량 이진벡터 $\boldsymbol{X}, \boldsymbol{X}_i$ 에 대하여 $d(\boldsymbol{X}, \boldsymbol{X}_i)=(\boldsymbol{X}-\boldsymbol{X}_i)'(\boldsymbol{X}-\boldsymbol{X}_i)$ 라 정의하면 $d(\boldsymbol{X}, \boldsymbol{X}_i)$ 는 $\boldsymbol{X}$와 $\boldsymbol{X}_i$의 대응되는 요소들 중에서 일치하지 않은 것들의 갯수임을 알 수 있다. Aitchison and Aitken (1976)은 $1/2 \le \lambda \le 1$을 만족하는 λ에 대하여 다음과 같은 커널 K_2를 제시하였다.

$$K_2(\boldsymbol{X}|\boldsymbol{X}_i;\lambda) = \lambda^{k-d(\boldsymbol{X},\boldsymbol{X}_i)}(1-\lambda)^{d(\boldsymbol{X},\boldsymbol{X}_i)} . \tag{4}$$

확률표본 $\{\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_n\}$이 표본공간 B^k를 가진 분포 p로부터 추출되었다면 $\boldsymbol{X}=\boldsymbol{x}$ 에서의 p에 대한 커널 추정치 $\hat{p}$는

$$\hat{p}(\boldsymbol{x}) = (1/n)\sum_i K_2(\boldsymbol{x}|\boldsymbol{X}_i;\lambda)$$

로 계산된다. 이 때 λ는 평활모수이며 평활정도를 나타낸다. 자료로부터 λ를 결정할 수 있는 방법으로는 $\hat{p}_i(\boldsymbol{X}_i)$를 $\boldsymbol{X}_i$를 제외한 자기제외 추정치라 할 때 $\sum \log \hat{p}_i(\boldsymbol{X}_i)$를 최대화시키는 λ를 찾는 최우교차타당성방법(maximum likelihood cross-validation method)이 많이 쓰인다.

이산형인 관측벡터를 확장하여 k_1개의 이산형 이진변수들과 또한 그들과 독립적인 k_2개의 연속형 변수들로 구성된 혼합형 관측벡터를 고려한다. 즉 확률벡터 $\boldsymbol{X}$에 대하여 $\boldsymbol{X}' = (\boldsymbol{X}_1', \boldsymbol{X}_2')$ 이며 이 때 $\boldsymbol{X}_1$은 k_1개의 이진 변수들로 이루어진 $k_1 \times 1$ 벡터이고, $\boldsymbol{X}_2$는 k_2개의 연속형 변수들로 이루어진 $k_2 \times 1$ 벡터이다. 혼합형 관측벡터 표본 $\{ \boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_n \}$이 모집단으로 부터 추출되었고 모집단의 밀도함수를 $f(\boldsymbol{X})$라 하자. 따라서 i번째 관측치 $\boldsymbol{X}_i$는 $\boldsymbol{X}_i' = (\boldsymbol{X}_{1i}', \boldsymbol{X}_{2i}')$ 이며 $\boldsymbol{X}_{1i}$는 $k_1 \times 1$ 이진 변수 벡터이고 $\boldsymbol{X}_{2i}$는 $k_2 \times 1$ 연속형 변수 벡터이다. 식 (2) 에 사용된 다변량 커널 K_1과 식 (4) 로 정의된 이산형 자료에 대한 커널 K_2을 결합하여 다음과 같이 혼합자료의 밀도함수추정을 위한 커널을 정의할 수 있다 :

$$K_3(\boldsymbol{X}|\boldsymbol{X}_i;\lambda,h) = \lambda^{k_1 - d(\boldsymbol{X}_1, \boldsymbol{X}_{1i})}(1-\lambda)^{d(\boldsymbol{X}_1, \boldsymbol{X}_{1i})} K_1\left(\frac{\boldsymbol{X}_2 - \boldsymbol{X}_{2i}}{h}\right)/h^{k_2} . \tag{5}$$

따라서 $\boldsymbol{X} = \boldsymbol{x}$ 에 대한 밀도함수 추정치는 $\hat{f}(\boldsymbol{x};\lambda,h) = (1/n)\sum_{i=1}^{n} K_3(\boldsymbol{x}|\boldsymbol{X}_i;\lambda,h)$ 로 계산된다.

혼합형 자료인 경우 커널판별식에 붓스트랩 방법을 적용하여 목적하는 오분류확률을 달성하는지와 정분류확률을 추정해봄으로써 커널 판별분석의 오분률확률에 대한 붓스트랩 조정방법의 타당성을 검정해 보기로 한다. 베르누이분포 *Bernoulli*(p)를 따르는 이산형 변수와 그와 독립적인 정규분포 $N(\mu,\sigma^2)$를 따르는 연속형 변수로 이루어진 혼합형 자료를 가정한다. π_1으로부터 추출된 확률표본은 $\{ \boldsymbol{X}_i' = (X_{1i}, X_{2i}) \}_{i=1}^{n}$ 이고 π_2로부터 추출된 확률표본은 $\{ \boldsymbol{Y}_i' = (Y_{1i}, Y_{2i}) \}_{i=1}^{m}$ 이 추출되었으며 이때 $X_{1i} \sim$ *Bernoulli*(p_1), $Y_{1i} \sim$ *Bernoulli*(p_2), $X_{2i} \sim N(\mu_1, \sigma_1^2)$, $Y_{2i} \sim N(\mu_2, \sigma_2^2)$ 이다. 따라서 새로운 관측치 $\boldsymbol{z} = (z_1, z_2)$ 에 대한 식 (5)의 커널 K_3를 이용한 커널 판별식 추정치는

$$\hat{T}(\boldsymbol{z}) = \frac{\hat{f}_1(\boldsymbol{z};\lambda_1,h_1)}{\hat{f}_2(\boldsymbol{z};\lambda_2,h_2)} = \frac{(1/nh_1)\sum_{i=1}^{n}\lambda_1^{1-(z_1 - X_{1i})^2}(1-\lambda_1)^{(z_1 - X_{1i})^2} K((z_2 - X_{2i})/h_1)}{(1/mh_2)\sum_{i=1}^{m}\lambda_2^{1-(z_1 - Y_{1i})^2}(1-\lambda_2)^{(z_1 - Y_{1i})^2} K((z_2 - Y_{2i})/h_2)} \tag{6}$$

이다.

먼저 목표오분류확률이 $\alpha = P(2|1) = 0.05$ 일때 위의 커널판별식에 붓스트랩 방법을 적용하여 구한 임계치 c_α^*를 결정하여 그 목표 오분류확률을 달성할 수 있는지 확인하기 위해 $p_1 = 0.3$, $\mu_1 = 0$, $\sigma_1^2 = \sigma_2^2 = 1$, $n = m = 150$ 을 가정하였고, $\mu_2 = 1, 2$ 일때 $p_2 = 0.5, 0.7, 0.9$ 인 경우 각각 오분류확률을 추정하였다. 식 (6) 의 커널판별식을 구할 때 이산형 변수에 대한 평활모수 λ_1과 λ_2는 최우교차타당성방법에 의하여 결정하였고 연속형 변수에 대한 평활모수 h_1과 h_2는 최소제곱교차

타당성방법이 적용되었으며 사용된 커널은 Epanechnik- ov 커널이다.

임계치 결정을 위한 붓스트랩 표본 추출은 연속형 변수에 대해서는 2.2 의 단계 (i), (iv)와 3.1 의 단계 (ii)', (iii)'이 적용되었다. 이산형 변수에 대한 붓스트랩 추출은 2.2의 단계 (i) - (iv)를 따르되 단계 (ii)와 (iii)은 다음과 같이 다시 고려되어야 한다. π_1의 이산형 표본 $\{X_{11}, X_{12}, \cdots, X_{1n}\}$의 한 관측치 X_{1l}에 대응하는 붓스트랩 관측치를 X^*라 하자. X_{1l}는 이진자료이므로 X^* 역시 이진 값을 가져야 하며, X^*의 X_{1l}에 대한 커널함수는 $\varepsilon = 1-(X^*-X_{1l})^2$ 로 놓을 때

$$\begin{aligned} K_2(X^*|X_{1l};\lambda_1) &= \lambda_1^{1-d(X^*,X_{1l})}\ (1-\lambda_1)^{d(X^*,X_{1l})} \\ &= \lambda_1^{1-(X^*-X_{1l})^2}\ (1-\lambda_1)^{(X^*-X_{1l})^2} \\ &= \lambda_1^{\varepsilon}(1-\lambda_1)^{1-\varepsilon} \end{aligned}$$

임을 알 수 있다. 따라서 위의 K_2는 *Bernoulli*(λ_1) 분포의 확률질량함수이다. $\varepsilon = 1-(X^*-X_{1l})^2$ 으로 부터 $\varepsilon = 0 \leftrightarrow X^* = 1-X_{1l}$ 이고 $\varepsilon = 1 \leftrightarrow X^* = X_{1l}$ 임을 알 수 있으므로 *Bernoulli*(λ_1)으로 부터 무작위 추출된 ε 로 부터 X^*를 다음과 같이 계산할 수 있다.

(ii)''' *Bernoulli*(λ_1) 으로 부터 난수 ε 발생.

(iii)''' $X^* = X_{1l}\varepsilon + (1-X_{1l})(1-\varepsilon)$.

π_2의 이산형 표본 $\{Y_{11}, Y_{12}, \cdots, Y_{1n}\}$에 대한 붓스트랩 표본도 λ_1대신 λ_2를 사용하여 위의 절차에 따라 동일하게 추출된다.

$\{\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_n\}$으로부터 추출된 크기 n+1인 혼합형 붓스트랩 표본을 $\{\boldsymbol{z}^*, \boldsymbol{X}_1^*, \boldsymbol{X}_2^*, \cdots, \boldsymbol{X}_n^*\}$라 하고 $\{\boldsymbol{Y}_1, \boldsymbol{Y}_2, \cdots, \boldsymbol{Y}_m\}$으로부터 추출된 붓스트랩 표본을 $\{\boldsymbol{Y}_1^*, \boldsymbol{Y}_2^*, \cdots, \boldsymbol{Y}_m^*\}$라 하자. $\{\boldsymbol{z}^*, \boldsymbol{X}_1^*, \boldsymbol{X}_2^*, \cdots, \boldsymbol{X}_n^*\}$와 $\{\boldsymbol{Y}_1^*, \boldsymbol{Y}_2^*, \cdots, \boldsymbol{Y}_m^*\}$를 식 (6)에 대입하여 통계량 값을 계산하고 이를 $B=499$회 반복하여 임계치 c_α^*를 전과같이 결정한다. 표본 $\{\boldsymbol{z}, \boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_n\}$와 $\{\boldsymbol{Y}_1, \boldsymbol{Y}_2, \cdots, \boldsymbol{Y}_m\}$을 각각 π_1과 π_2로 부터 $M=1000$회 독립적으로 추출하여 각 표본에 대한 식 (6)의 통계량 값이 그것의 임계치 c_α^* 보다 작거나 같은 비율로서 추정된, 오분류확률 $\alpha = P(2|1) = 0.05$ 에 대한 추정치들이 표 3 에 나와있다. π_1의 모수들과 다른 값을 갖는 π_2에 대하여 오분류확률 추정치 $\hat{P}_\gamma(2|1)$가 유의수준 0.05 하에서 모두 목표오분류확률 $\alpha = 0.05$ 과 다르지 않다는 것을 확인할 수 있다. $\boldsymbol{z} \in \pi_2$ 일 때 정분류확률 $P(2|2)$에 대한 추정치 $\hat{P}_\gamma(2|2)$가 그림 2에 나와있다. π_2의 연속형 변수의 평균 μ_2가 π_1의 그것의 평균 μ_1으로

부터 멀어질수록, 또한 p_2가 p_1과 많이 다를수록 정분류확률은 증가한다. 그러나 π_2의 연속형 분포가 π_1의 연속형 분포와 구분이 확연한 경우 (μ_2=3.0) 에는 p_2 값에는 그리 큰 영향을 받지 않음을 알 수 있다.

표 3. 혼합형 자료일 때 오분류확률 $P(2|1) = 0.05$ 에 대한 추정치

$\hat{P}_{\gamma}(2\|1)$	p_2		
	0.5	0.7	0.9
μ_2=1	0.054 (0.007)	0.058 (0.007)	0.054 (0.007)
μ_2=2	0.062 (0.008)	0.057 (0.007)	0.048 (0.007)

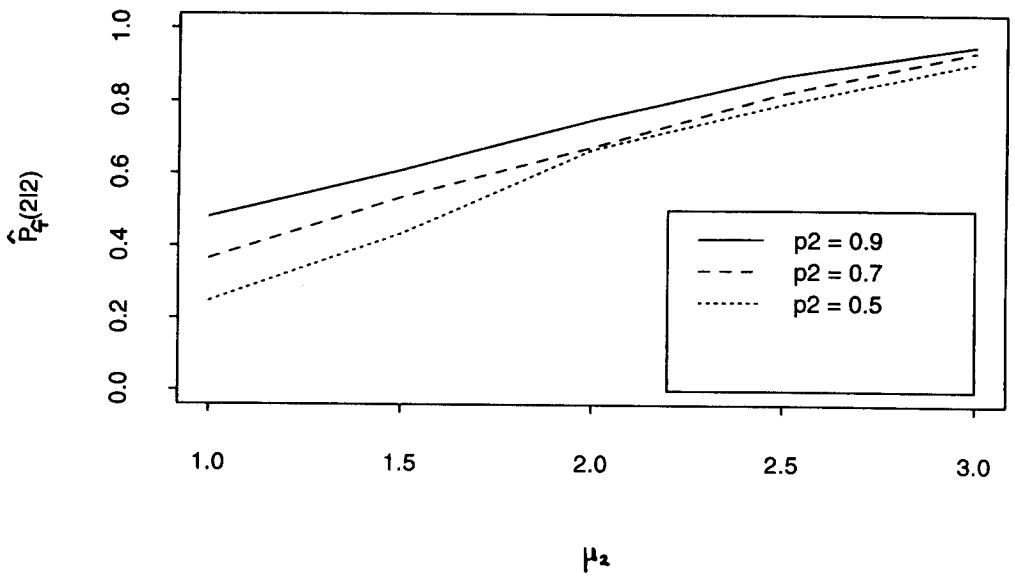

그림 2. 이산형과 연속형 변수들의 혼합자료에 대한 정분류확률의 추정치 $\hat{P}_{\gamma}(2|2)$.
$\mu_1 = 0$, $\sigma_1^2 = \sigma_2^2 = 1$, $p_1 = 0.3$

4. 결론

관심 오분류확률을 분석자가 원하는 수준에 맞추어 새로운 관측치를 두 모집단 중 어느 하나에 비모수적인 방법으로 분류할 때, 본 논문에서 제시된 붓스트랩 기법을 이용한 커널 판별분석방법이 유용하다는 것이 밝혀졌다. 커널 밀도함수 추정방법에 의해 두 모집단의 확률밀도함수들을 추정하여 그 추정치들의 비율을 판별(분류) 통계량으로 구성하고, 오분류확률을 조정할 수 있는 통계량 임계치를 붓스트랩 기법에 의해 근사적으로 결정하였다.

분산이 동일한 두 개의 정규분포 모집단의 경우 붓스트랩을 이용한 커널 판별 통계량은 표본크기가 그리 크지 않을 때에도 (n=m=75) 그 효능에 있어서 점근분포가 알려져 있는 Anderson의 W 통계량과 거의 대등하다는 것이 보여졌다. 본 논문에 제시된 방법은 모집단들이 정규분포를 따르지 않은 경우와, 이진의 이산형 변수와 정규분포를 따르는 연속형 변수가 혼합된 경우에도 물론 목표 오분류확률을 달성하였다. 더욱이 이러한 방법은 이산형과 연속형 변수가 혼합된 어떠한 혼합자료에도 적용될 수 있다. 커널 판별 통계량이 목표하는 오분류확률을 정확하게 달성하기 위해서는 표본크기 n,m 은 물론 붓스트랩 표본 크기 B가 모두 커야 한다. 분류변수의 차원이 증가함에 따라 그에 적절한 표본크기에 관한 연구도 가치가 있을 것이다.

참고문헌

[1] Abramson, I. S. (1982). On Bandwidth Variation in Kernel Estimates - A Square Root Law, *Annals of Statistics*, Vol. 10, 1217-1223.

[2] Aitchison, J. and Aitken, C. G. G. (1976). Multivariate Binary Discrimination by the Kernel Method, *Biometrika*, Vol. 63, 413-420.

[3] Anderson, T. W. (1958). *An Introduction to Multivariate Statistical Analysis*, John Wiley & Sons, New York.

[4] Anderson, T. W. (1973). An Asymptotic Expansion of the Distribution of the Standardized Classification Statistic W, *Annals of Statistics*, Vol. 1, 964 - 972.

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

[6] Baek, J., Gray, H. L. and Woodward, W. A (1993). A Bootstrap Generalized Likelihood Ratio Test in Discriminant Analysis, to appear in *Computational Statistics & Data Analysis*.

[7] Bickel, P. J. and Freedman, D. A. (1981). Some Asymptotic Theory for the Bootstrap, *Annals of Statistics*, Vol. 6, 1196-1217.

[8] Bowman, A. (1984). An Alternative Method of Cross-Validation for the Smoothing of Density Estimates, *Biometrika*, Vol. 71, 353-360.

[9] Breiman, L., Meisel W. and Purcell, E. (1977). Variable Kernel Estimates of Multivariate Densities, *Technometrics*, Vol. 19, 135-144.

[10] Cacoullos, T. (1966). Estimation of a Multivariate Density, *Annals of Institute of Statistical Mathematics*, Vol. 18, 179 - 189.

[11] Cao, R. Cuevas, A. and Manteiga, W. G. (1994). A Comparative Study of Several Smoothing Methods in Density Estimation, *Computational Statistics & Data Analysis*, 17, 153-176.

[12] Devroye, L. and Györfi, L. (1985). *Nonparametric Density Estimation, The* L_1 *View*, New York, Wiley.

[13] Efron, B. (1979). Bootstrap Method: Another Look at the Jackknife, *Annals of Statistics*, Vol. 7, 1-26.

[14] Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*, SIAM, Philadelphia.

[15] Fisher, R, A. (1936). The Use of Multiple Measurements in Taxonomic Problems, *Annals of Eugenics*, Vol. 7, 179 -188.

[16] Hand, D. J. (1982). *Kernel Discriminant Analysis*, Research Studies Press, Chichester.

[17] John, S. (1960). On Some Classification Problems, *Sankhya, Series A* 22, 301-308.

[18] John, S. (1963). On Classification by the Statistics R and Z, *Annals of Institute of Statistical Mathematics*, 14, 237-246.

[19] Kanazawa, M. (1979). The Asymptotic Cut-off Point and Comparison of Error Probabilities in Covariate Discriminant Analysis, *Journal of Japan Statistical Society*, Vol. 9, 7-17.

[20] Park, B. U. (1991). Advances in Data-Driven Bandwidth Selection, *Journal of the Korean Statistical Society*, 1-28.

[21] Parzen, E. (1962). On Estimation of a Probability Density Function and Mode, *Annals of Mathematical Statistics*, 33, 1065-1076.

[22] Remme, J., Habbema, J. D. F. and Hermans, J. (1980). A Simulative Comparison of Linear, Quadratic and Kernel Discrimination, *Journal of Statistical Computation and Simulation*, Vol. 11, 87-106.

[23] Rudemo, M. (1982). Empincal Choice of Histograms and Kernel Density Estimators, *Scandinavian Journal of Statistics*, Vol. 9. 65-78.

[24] Seber, G. A. F. (1984). *Multivariate Observation*, John Wiley & Sons, New York.

[25] Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*, Chapman & Hall, London.

[26] Welch, B. L. (1939). Note on Discriminant Functions, *Biometrika*, Vol. 31, 218-220.