10장 판별분석과 분류

덕성여자대학교 정보통계학과 김 재희

10.1 서론

판별(discrimination) 및 분류(classification) 분석 :

집단에 대한 정보로부터 집단을 구별할 수 있는 판별함수(discriminant function) 또는 판별규칙(discriminant rule)을 만들고, 새로운 개체에 대해 어느 집단에 속하는지를 판별하여 분류하는 다변량 기법으로 집단에 대한 정보를 이용한 탐색적인 통계 기법.

▶ 판별분석(discriminant analysis)에 대한 개념의 시발점: 1920년대 영국 통계학자인 Karl Pearson이 인종 그룹간 데이터의 거리(distance)를 나타내는 지수(index)로 인종경향계수(coefficient of racial likeness: CRL)를 제안.

▶ 1920년대 G. M. Morant가 CRL을 확장 연구하였으며,

▶ 1930년대 인도(India) 출신의 통계학자인 P. C. Mahalanobis가 거리 측도 제안.

▶ 1930년대 R. A. Fisher는 다변량 그룹간 거리(multivariate intergroup distance)를 그룹간 판별을 목적으로 변수들의 선형조합으로 변환.

▶ 1940년대 Fisher의 아이디어를 확장하고 발전시킨 연구.

판별분석은 초기 생물학이나 의학 분야에서 주로 분류를 위해 적용되었으나 이제는 경영학, 경제학, 교육학, 공학, 심리학 등의 여러 분야에서 비중있게 적용되는 방법이 되었다.

특히 1950-60년대 하버드 대학(Havard University)에서 교육학과 심리학 분야에서 판별분석 방법을 적용한 연구 결과를 발표하여 방법론 확장에 많은 기여를 하였다.

30-40년 동안 판별분석의 목적은 주로 새로운 개체에 대해 그룹을 판단하는 예측판별분석(predictive discriminant analysis)이었으나 1960년대 이후에는 Fisher의 선형판별함수(linear discriminant function: LDF)를 다변량 분산분석(MANOVA)에서 그룹간의 효과를 드러내주는 함수로 해석하기 시작했으며 다변량 분산분석 개념이 판별분석 방법 개발에 활용.

▶ 이 장에서는 모집단을 나타내는 용어로 '그룹(group)'을 사용하기로함.

▶ 그룹간의 차이를 크게 해주는 그룹 판별함수(그룹간의 거리를 최대로 멀리하도록 만든 변수들의 선형결합식) 구함.

▶ 구한 판별함수 이용해 기존의 개체 분류하여 오분류율(misclassification rate)계산.

▶ 새로운 개체에 대해서는 판별함수를 이용하여 속하는 그룹을 추정할 수 있게 한다.

판별과 분류분석의 목적

(1) 몇 개의 알려진 그룹으로부터 그룹의 특성을 나타내주고 구별해주는 함수를 결정한다.

(2) 결정된 판별함수를 이용하여 새로운 관측치를 판별하여 개체를 분류한다.

10.2 두 개 그룹의 판별

10.2.1 Fisher의 방법

두 개의 그룹 G_1, G_2 판별할 수 있는 변수들의 선형조합으로 판별함수를 구하고자 한다.
Fisher는 다변량벡터를 일변량 변수로 변환하여 그룹을 판별하는 방법 제안.
정규성(normaliry) 가정은 필요하지 않다.

G_1에 속하는 n_1개의 관측된 확률벡터 $\boldsymbol{X}_{1i}=(X_{1i1},X_{1i2},...,X_{1ip})'$, $i=1,2,...,n_1$ 는 평균벡터 μ_1을 가지며 공분산행렬 Σ를 갖는다.
G_2에 속하는 n_2개의 관측된 확률벡터 $\boldsymbol{X}_{2i}=(X_{2i1},X_{2i2},...,X_{2ip})'$, $i=1,2,...,n_2$는 평균벡터 μ_2와 공분산행렬 Σ를 갖는다.
G_1으로부터의 확률표본 $\boldsymbol{X}_{11},\boldsymbol{X}_{12},...,\boldsymbol{X}_{1n_1}$과 G_2으로부터의 확률표본 $\boldsymbol{X}_{21},\boldsymbol{X}_{22},...,\boldsymbol{X}_{2n_2}$를 얻었다고 하자.

그룹을 판별하기 위해서 G_1, G_2를 가능한 한 떨어져 구별되게 하는 선형판별함수 $Y = l'\boldsymbol{X}$를 구하고자 한다.

두 그룹 G_1, G_2는 공통 공분산행렬 Σ 를 가지며 각 그룹에서 판별함수의 기대값과 분산

$$
\begin{aligned}
\mu_{1Y} &= E(Y|G_1) = l'\boldsymbol{\mu}_1 \\
\mu_{2Y} &= E(Y|G_2) = l'\boldsymbol{\mu}_2 \\
\sigma_Y^2 &= Var(Y) = Var(l'\boldsymbol{X}) = l'\Sigma l
\end{aligned}
\tag{10.1}
$$

분산을 고려하여 두 그룹의 평균 차를 고려하기위해

$$
\begin{aligned}
\frac{(\text{두 그룹 평균의 차이})^2}{Y\text{ 의 분산}} &= \frac{(\mu_{1Y} - \mu_{2Y})^2}{\sigma_Y^2} \\
&= \frac{l'(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'l}{l'\Sigma l} \\
&= \frac{(l'\delta)^2}{l'\Sigma l}
\end{aligned}
\tag{10.2}
$$

가 최대가 되도록 하며 그룹을 분리할 수 있는 함수를 구하고자한다. $\delta = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$이다.

$\frac{(l'\delta)^2}{l'\Sigma l}$를 최대화하도록 계수벡터 l을 구하기 위해 2장의 최대화정리(maximization lemma)를 이용하면 다음의 결과

$$l = c\Sigma^{-1}\delta = c\Sigma^{-1}(\mu_1 - \mu_2) \tag{10.3}$$

를 얻게 된다.

▶ 특히 $c = 1$로 놓을 때, Fisher의 선형판별함수는

$$Y = \boldsymbol{l}'\boldsymbol{X} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma^{-1}\boldsymbol{X}. \tag{10.4}$$

▶ 표본에 대한 Fisher의 선형판별함수는

$$y = \hat{l}'\boldsymbol{X} = (\overline{\boldsymbol{X}}_1 - \overline{\boldsymbol{X}}_2)'\boldsymbol{S}_{pl}^{-1}\boldsymbol{X} \tag{10.5}$$

여기서 각 그룹의 표본평균벡터, 표본공분산행렬과 합동공분산행렬을 구하면

$$\overline{\boldsymbol{X}}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1}\boldsymbol{X}_{1i}, \quad \overline{\boldsymbol{X}}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2}\boldsymbol{X}_{2i},$$

$$\boldsymbol{S}_1 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(\boldsymbol{X}_{1i}-\overline{\boldsymbol{X}}_1)(\boldsymbol{X}_{1i}-\overline{\boldsymbol{X}}_1)',$$

$$\boldsymbol{S}_2 = \frac{1}{n_2-1}\sum_{i=1}^{n_2}(\boldsymbol{X}_{2i}-\overline{\boldsymbol{X}}_2)(\boldsymbol{X}_{2i}-\overline{\boldsymbol{X}}_2)',$$

$$\begin{aligned}\boldsymbol{S}_{pl} &= \left[\frac{(n_1-1)}{(n_1-1)+(n_2-1)}\right]\boldsymbol{S}_1 + \left[\frac{(n_2-1)}{(n_1-1)+(n_2-1)}\right]\boldsymbol{S}_2 \\ &= \frac{(n_1-1)\boldsymbol{S}_1+(n_2-1)\boldsymbol{S}_2}{n_1+n_2-2}\end{aligned} \qquad (10.6)$$

각 그룹의 평균점은 $\overline{y_1} = \hat{\boldsymbol{l}}'\overline{\boldsymbol{X}}_1$과 $\overline{y_2} = \hat{\boldsymbol{l}}'\overline{\boldsymbol{X}}_2$ 이고 이들의 중간점은

$$\widehat{m} = \frac{1}{2}(\overline{y_1} + \overline{y_2}) = \frac{1}{2}(\overline{\boldsymbol{X}}_1 - \overline{\boldsymbol{X}}_2)'\boldsymbol{S}_{pl}^{-1}(\overline{\boldsymbol{X}}_1 + \overline{\boldsymbol{X}}_2) \qquad (10.7)$$

이 된다. 새로운 관측벡터 x_0에 대해 다음의 판별함수

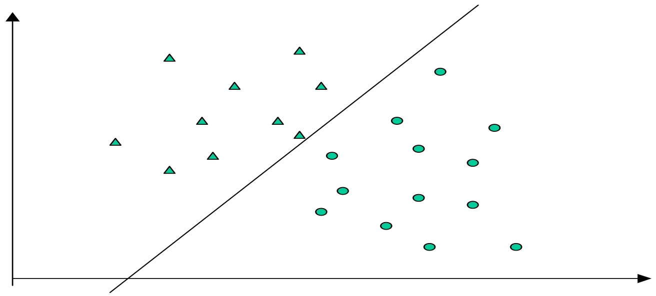
$$y_0 = (\overline{\boldsymbol{X}}_1 - \overline{\boldsymbol{X}}_2)'\boldsymbol{S}_{pl}^{-1}\boldsymbol{X}_0 \qquad (10.8)$$

를 이용한 판별규칙(discriminant rule)은:

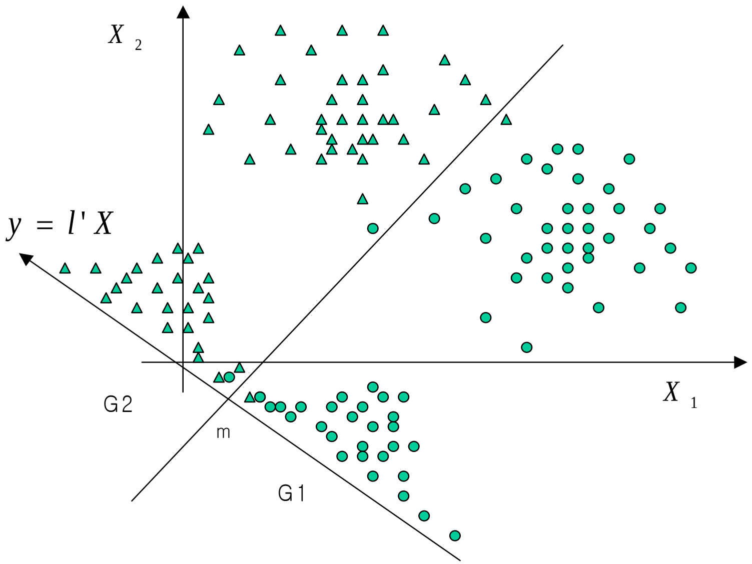
$$y_0 > \widehat{m}\text{이면 } \boldsymbol{X}_0\text{가 } G_1\text{에 속하고,}$$

$$y_0 \leqq \widehat{m}\text{이면 } \boldsymbol{X}_0\text{가 } G_2\text{에 속한다.} \qquad (10.9)$$

변수가 2개인 경우 선형판별함수는 [그림 10.1]과 같이 그려질 수 있다. [그림 10.2]는 Fisher의 선형판별함수(정준판별함수)에 대한 이해를 위한 그림으로, 선형판별함수에 의해 새로운 y 축이 형성되어 두 그룹의 중간점 $\widehat{m}$을 기준으로 두 그룹이 나뉘어 분포됨을 보여주고 있다.



[그림 10.1] 두 그룹을 분리하는 선형판별함수



[그림 10.2] Fisher의 선형판별함수와 분류

≪예제 10.1≫ 2개 그룹으로부터의 확률표본이 다음과 같을 때

$$G_1: \begin{pmatrix}1\\5\end{pmatrix} \quad \begin{pmatrix}0\\3\end{pmatrix}\begin{pmatrix}2\\4\end{pmatrix}$$
$$G_2: \begin{pmatrix}-3\\5\end{pmatrix}\begin{pmatrix}-2\\7\end{pmatrix}\begin{pmatrix}-1\\6\end{pmatrix}$$

Fisher의 판별함수를 구하고자 한다. 우선 각 그룹의 평균벡터와 공분산행렬을 계산하면

$$\overline{\boldsymbol{X}}_1 = \begin{pmatrix}1\\4\end{pmatrix}, \qquad \overline{\boldsymbol{X}}_2 = \begin{pmatrix}-2\\6\end{pmatrix}$$

$$\boldsymbol{S}_1 = \begin{pmatrix}1 & 0.5\\0.5 & 1\end{pmatrix} = \boldsymbol{S}_2$$

이 되며 합동공분산행렬

$$\boldsymbol{S}_{pl} \equiv \frac{(n_1-1)\boldsymbol{S}_1 + (n_2-1)\boldsymbol{S}_2}{n_1+n_2-2} = \begin{pmatrix}1 & 0.5\\0.5 & 1\end{pmatrix}$$

이 구해진다.

주어진 표본에 대한 Fisher의 선형판별함수는

$$y = \hat{l}'\boldsymbol{X} = (\overline{\boldsymbol{X}}_1 - \overline{\boldsymbol{X}}_2)'\boldsymbol{S}_{pl}^{-1}\boldsymbol{X}$$

$$= (3\ ,\ \ -2)\begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}^{-1}\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = (3\ ,\ \ -2)\begin{pmatrix} \frac{4}{3} & -\frac{2}{3} \\ -\frac{2}{3} & \frac{4}{3} \end{pmatrix}\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$$

$$= (5.33\ ,\ \ -4.67)\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \ 5.33X_1 - 4.67X_2$$

으로 구할 수 있다. 그룹의 평균점은 각각

$$\overline{y_1} = \hat{l}'\overline{\boldsymbol{X}}_1 = (5.33,\ -4.67)\begin{pmatrix} 1 \\ 4 \end{pmatrix} = -13.35$$

$$\overline{y_2} = \hat{l}'\overline{\boldsymbol{X}}_2 = (5.33,\ -4.67)\begin{pmatrix} -2 \\ 6 \end{pmatrix} = -38.68$$

이들의 중간점: $$\widehat{m} = \frac{1}{2}(\overline{y_1} + \overline{y_2}) = \frac{1}{2}(\overline{\boldsymbol{X}}_1 - \overline{\boldsymbol{X}}_2)'\boldsymbol{S}_{pl}^{-1}(\overline{\boldsymbol{X}}_1 + \overline{\boldsymbol{X}}_2)$$

$$= \frac{1}{2}(-13.35 - 38.68) = -26.015$$

새로운 관측벡터 $\boldsymbol{X}_0$에 대해

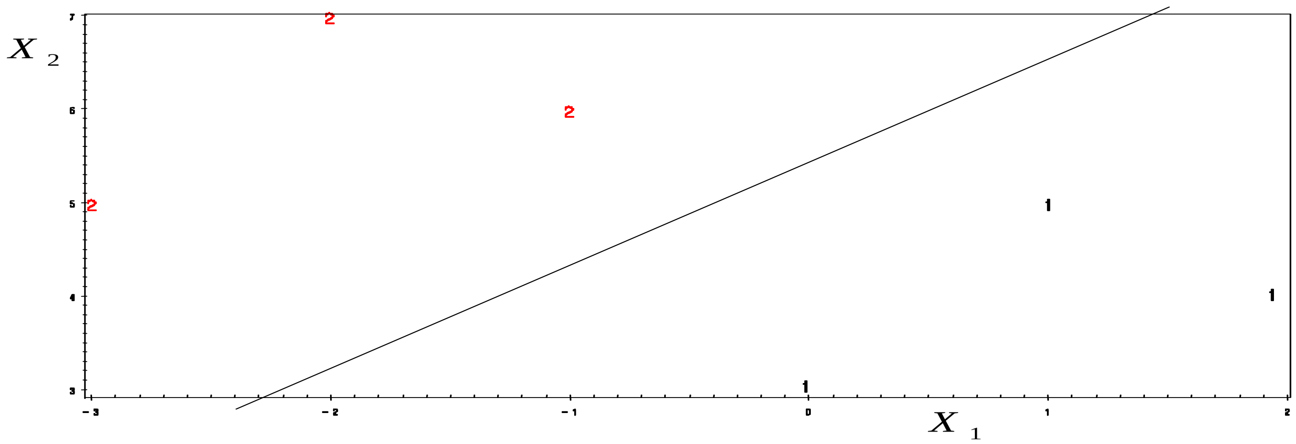
$$y_0 = (\overline{\boldsymbol{X}}_1 - \overline{\boldsymbol{X}}_2)' \boldsymbol{S}_{pl}^{-1} \boldsymbol{X}_0$$

를 계산하고 이에 대한 판별규칙은:

$y_0 > -26.015$이면 $\boldsymbol{X}_0$가 G_1에 속하고,

$y_0 \leqq -26.015$이면 $\boldsymbol{X}_0$가 G_2에 속한다.

주어진 데이터와 구한 선형판별함수를 그려보면 [그림 10.1]과 같으며 두 그룹을 분리하는 직선 $5.33X_1 - 4.67X_2 = -26.015$으로 판별함수가 정의됨을 알 수 있다.



[그림 10.3] ≪예제 10.1≫에서 구한 선형판별함수 그래프

10.2.2 다변량 정규분포를 따르며 두 그룹의 공분산행렬이 같은 경우

두 그룹의 공분산행렬을 각각Σ_1, Σ_2라 하고,$\Sigma_1 = \Sigma_2$라고 가정하자.

$f(\boldsymbol{X}|G_1)$은 G_1그룹으로부터 발생한 $\boldsymbol{X}$의 확률밀도함수이고 $f(\boldsymbol{X}|G_2)$은 G_2그룹으로부터 발생한 $\boldsymbol{X}$의 확률밀도함수로 표기한다. (조건부 확률의 표현이 아님을 유의한다.)

p_1과 p_2는 각각 $\boldsymbol{X}$가 G_1, G_2 그룹에서 발생할 사전확률(prior probability)이다. 그러므로 $p_2 = 1 - p_1$의 관계를 갖는다.

정리 10.1 오분류(misclassification) 확률을 최소화하는 최적 분류규칙(optimal classification rule)은

$$p_1 f(\boldsymbol{X}|G_1) > p_2 f(\boldsymbol{X}|G_2)$$ 이면 $\boldsymbol{X}$ 를 G_1그룹에 분류하고,

$$p_1 f(\boldsymbol{X}|G_1) \leqq p_2 f(\boldsymbol{X}|G_2)$$ 이면 $\boldsymbol{X}$ 를 G_2그룹에 분류한다. (10.10)

만약 $p_1 = p_2$ 이면 최적 분류규칙은

$$f(\boldsymbol{X}|G_1) > f(\boldsymbol{X}|G_2)$$ 이면 $\boldsymbol{X}$ 를 G_1그룹에 분류한다.

이며 이는 최대우도(maximum likelihood)를 이용한 규칙이 된다.

이번에는 특히 $\boldsymbol{X}$가 다변량 정규분포를 따를 경우를 고려하자.
$f(\boldsymbol{X}|G_1) \equiv N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$이고 $f(\boldsymbol{X}|G_2) \equiv N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$일 때 확률밀도함수

$$f(\boldsymbol{X}|G_i) = (2\pi)^{-p/2}|\boldsymbol{\Sigma}|^{-1/2} e^{-(\boldsymbol{X}-\boldsymbol{\mu}_i)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu}_i)/2}, \quad i = 1,2 \tag{10.11}$$

를 가진다.

정리 10.1로부터 그룹분류를 위한 다음의 부등식

$$\frac{f(\boldsymbol{X}|G_1)}{f(\boldsymbol{X}|G_2)} > \frac{p_2}{p_1} \tag{10.12}$$

이 성립하며 확률밀도함수를 이용하여 계산하면

$$\begin{aligned}\frac{f(\boldsymbol{X}|G_1)}{f(\boldsymbol{X}|G_2)} &= e^{-(\boldsymbol{X}-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu}_1)/2+(\boldsymbol{X}-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu}_2)/2} \\ &= e^{(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}\boldsymbol{X}-(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)/2}\end{aligned} \tag{10.13}$$

이 되어 양변에 자연로그 $\ln$함수를 취하여 식(10.13)에 대입하면, 최적 분류규칙은:

$$(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}\boldsymbol{X} > \frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)+\ln\left(\frac{p_2}{p_1}\right) \tag{10.14}$$

이면 $\boldsymbol{X}$를 G_1그룹에 분류하고, 그렇지 않으면 $\boldsymbol{X}$를 G_2그룹에 분류한다.

▶ 일반적으로 μ_1, μ_2와 Σ는 모르는 경우이고 이 때 추정량으로 $\overline{\boldsymbol{X}}_1$, $\overline{\boldsymbol{X}}_2$와 $\boldsymbol{S}_{pl}$을 사용한다. 여기서

$\overline{\boldsymbol{X}}_1$는 G_1그룹에서 얻어진 n_1개 표본 평균벡터,

$\overline{\boldsymbol{X}}_2$는 G_2그룹에서 얻어진 n_2개 표본 평균벡터,

$\boldsymbol{S}_{pl}$는 두 집단의 표본으로부터 추정된 합동공분산행렬이다.

그러므로 새로이 관측된 벡터 $\boldsymbol{X}$가 속하는 그룹에 대한 선형판별 규칙(linear classification rule)은 다음과 같다.

$$(\overline{\boldsymbol{X}}_1 - \overline{\boldsymbol{X}}_2)'\boldsymbol{S}_{pl}^{-1}\boldsymbol{X} > \frac{1}{2}(\overline{\boldsymbol{X}}_1 - \overline{\boldsymbol{X}}_2)'\boldsymbol{S}_{pl}^{-1}(\overline{\boldsymbol{X}}_1 + \overline{\boldsymbol{X}}_2) + \ln\left(\frac{p_2}{p_1}\right) \tag{10.15}$$

이면 $\boldsymbol{X}$를 G_1그룹에 분류하고, 그렇지 않으면 $\boldsymbol{X}$를 G_2그룹에 분류한다.

만약 $p_1 = p_2$이면, 정규성으로부터 유도된 선형판별규칙은 Fisher의 판별규칙과 같아진다. 식 (10.15) 이용한 분류규칙은 근사적으로 최적(asymptotically optimal) 판별규칙이 된다.

≪예제 10.2≫ 모집단이 다변량 정규분포를 따르고 공분산행렬이 같다고 할 수 있으며 그룹1에 대한 사전확률은 0.6, 그룹2에 대한 사전확률은 0.4라 하자. ≪예제 10.1≫의 자료에 대해 판별 규칙을 정하고자 한다.

$$\frac{1}{2}(\overline{\boldsymbol{X}}_1-\overline{\boldsymbol{X}}_2)'\boldsymbol{S}_{pl}^{-1}(\overline{\boldsymbol{X}}_1+\overline{\boldsymbol{X}}_2)+\ln\left(\frac{p_2}{p_1}\right)=-26.015-0.406=-25.609$$

이므로 식(10.15)를 이용하면 판별 규칙은

$$(\overline{\boldsymbol{X}}_1-\overline{\boldsymbol{X}}_2)'\boldsymbol{S}_{pl}^{-1}\boldsymbol{X}>-25.609$$

이면 그룹1로 분류하고, 그렇지 않으면 그룹2로 분류한다. 즉

$5.33X_1-4.67X_2>-25.609$ 이면 그룹1로 분류하고 그렇지 않으면 그룹2로 분류한다.

10.2.3 다변량 정규분포를 따르며 두 그룹의 공분산행렬이 다른 경우

두 그룹 G_1, G_2의 공분산행렬을 각각Σ_1, Σ_2이고,$\Sigma_1 \neq \Sigma_2$라고 가정하자. 우도비를 구하면

$$\frac{f(\boldsymbol{X}|G_1)}{f(\boldsymbol{X}|G_2)} = \frac{|\Sigma_2|^{1/2}}{|\Sigma_1|^{1/2}} e^{-(\boldsymbol{X}-\mu_1)'\Sigma_1^{-1}(\boldsymbol{X}-\mu_1)/2 + (\boldsymbol{X}-\mu_2)'\Sigma_2^{-1}(\boldsymbol{X}-\mu_2)/2} \quad (10.16)$$

이고 우도비에 자연로그함수를 취하면

$$Q(\boldsymbol{X}) = \ln\frac{f(\boldsymbol{X}|G_1)}{f(\boldsymbol{X}|G_2)} = \frac{1}{2}\ln\left(\frac{|\Sigma_2|}{|\Sigma_1|}\right) - \frac{1}{2}(\mathrm{X}-\mu_1)'\Sigma_1^{-1}(\mathrm{X}-\mu_1) + \frac{1}{2}(\boldsymbol{X}-\mu_2)'\Sigma_2^{-1}(\boldsymbol{X}-\mu_2)$$

$$= \frac{1}{2}\ln\left\{\frac{|\Sigma_2|}{|\Sigma_1|}\right\} - \frac{1}{2}\left(\mu_1'\Sigma_1^{-1}\mu_1 - \mu_2'\Sigma_2^{-1}\mu_2\right) + (\mu_1'\Sigma_1^{-1} - \mu_2'\Sigma_2^{-1})\boldsymbol{X} - \frac{1}{2}\boldsymbol{X}'(\Sigma_1^{-1} - \Sigma_2^{-1})\boldsymbol{X} \quad (10.17)$$

식(10.17)의 마지막 항 $X'(\Sigma_1^{-1} - \Sigma_2^{-1})X$은 X의 이차형식을 취하므로 $Q(X)$를 이차판별함수(quadratic discriminant function) 또는 이차분류함수(quadratic classification function)라고 한다.

▶ $Q(X)$를 이용한 최적 분류규칙은:

$$Q(X) > \ln\left(\frac{p_2}{p_1}\right) \tag{10.18}$$

이면 X를 G_1그룹에 분류하고, 그렇지 않으면 X를 G_2그룹에 분류한다.

각 그룹에서의 표본을 이용할 때의 분류 규칙을 정하기 위해 μ_1, μ_2로 표본평균 $\overline{X}_1$, $\overline{X}_2$를 이용하고 Σ_1, Σ_2에 대해서는 각각 공분산추정량 S_1, S_2를 사용하면 $Q(X)$의 표본함수 $Q_s(X)$는

$$Q_s(\boldsymbol{X}) = \frac{1}{2} ln\left\{\frac{|\boldsymbol{S}_2|}{|\boldsymbol{S}_1|}\right\} - \frac{1}{2}\left(\overline{\boldsymbol{X}}_1{}'\boldsymbol{S}_1^{-1}\overline{\boldsymbol{X}}_1 - \overline{\boldsymbol{X}}_2{}'\boldsymbol{S}_2^{-1}\overline{\boldsymbol{X}}_2\right) \tag{10.19}$$

$$+(\overline{\boldsymbol{X}}_1{}'\boldsymbol{S}_1^{-1} - \overline{\boldsymbol{X}}_2{}'\boldsymbol{S}_2^{-1})\boldsymbol{X}$$

$$-\frac{1}{2}\boldsymbol{X}'(\boldsymbol{S}_1^{-1} - \boldsymbol{S}_2^{-1})\boldsymbol{X}$$

이 되며, 이는 다음과 같은 이차형식으로 표현할 수 있다.

$$Q_s(\boldsymbol{X}) = b + \boldsymbol{c}'\boldsymbol{X} - \boldsymbol{X}'\boldsymbol{A}\boldsymbol{X}$$

$Q_s(\boldsymbol{X})$를 이용한 분류규칙은:

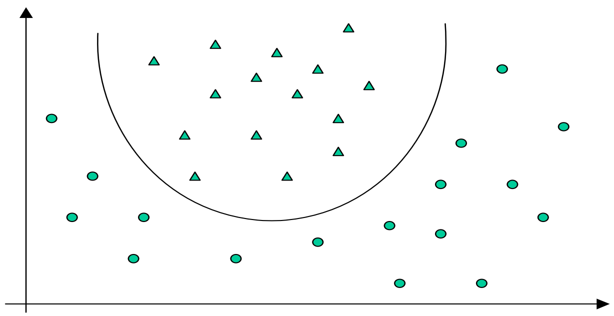
$$Q_s(\boldsymbol{X}) \ > \ \ln\left(\frac{\mathrm{p}_2}{\mathrm{p}_1}\right) \tag{10.20}$$

이면 $\boldsymbol{X}$를 G_1그룹에 분류하고, 그렇지 않으면 $\boldsymbol{X}$를 G_2그룹에 분류한다.

▶ $\Sigma_1 \neq \Sigma_2$인 경우 $Q_s(\boldsymbol{X})$를 이용한 분류규칙은 근사적으로 최적규칙은 아니다. 표본의 크기가 작을 경우에는 $\boldsymbol{S}_i$가Σ_i에 대한 안정적인 추정량이 아니므로, 즉 표본에 따라 변동이 심하므로, 이차판별함수보다 선형판별함수를 사용하는 것이 더 좋을 수 있다.

▶ 표본의 크기가 클 경우에는Σ_1과Σ_2의 차이가 클수록 이차판별함수를 사용하는 것이 더 좋다고 알려져 있다.

▶ 실제 판별분석에서 판별함수를 구하고자할 때, $H_0 : \Sigma_1 = \Sigma_2$에 대한 가설검정후 H_0를 기각하지 못할 경우에는 선형판별함수를, H_0를 기각할 경우에는 이차판별함수를 선택할 수 있다.



[그림 10.4] 두 그룹을 분리하는 이차판별함수

≪예제 10.3≫ 다변량 정규분포를 따르는 2개 그룹으로부터의 확률표본이 다음과 같다:

$$G_1 : \begin{pmatrix} 1 \\ 5 \end{pmatrix} \quad \begin{pmatrix} 0 \\ 3 \end{pmatrix} \quad \begin{pmatrix} 2 \\ 4 \end{pmatrix} \quad \begin{pmatrix} 5 \\ 6 \end{pmatrix}$$

$$G_2 : \begin{pmatrix} 2 \\ 1 \end{pmatrix} \quad \begin{pmatrix} 8 \\ 20 \end{pmatrix} \quad \begin{pmatrix} 14 \\ 16 \end{pmatrix} \quad \begin{pmatrix} 15 \\ 7 \end{pmatrix}$$

그룹을 구별할 수 있는 판별함수를 구하고자 한다. 우선 각 그룹의 표본평균벡터는

$$\overline{\boldsymbol{X}}_1 = \begin{pmatrix} 2 \\ 4.5 \end{pmatrix}, \ \overline{\boldsymbol{X}}_2 = \begin{pmatrix} 9.75 \\ 11 \end{pmatrix}$$

와 같이 구해지며, 표본공분산행렬은 각각

$$\boldsymbol{S}_1 = \begin{pmatrix} 4.67 & 2.33 \\ 2.33 & 1.67 \end{pmatrix}, \quad \boldsymbol{S}_2 = \begin{pmatrix} 36.25 & 20.67 \\ 20.67 & 74 \end{pmatrix}$$

이며 두 집단의 공분산행렬이 같다고 할 수 없다고 한다. 한편 일반화분산을 구해보면 $|\boldsymbol{S}_1| = 2.37, \ |\boldsymbol{S}_2| = 2255.25$이고

$$\boldsymbol{S}_1^{-1} = \begin{pmatrix} 0.70 & -0.98 \\ -0.98 & 1.97 \end{pmatrix}, \quad \boldsymbol{S}_2^{-1} = \begin{pmatrix} 0.03 & -0.01 \\ -0.01 & 0.02 \end{pmatrix}$$

으로 계산된다. 또한

$$\overline{\boldsymbol{X}}_1{}'\boldsymbol{S}_1^{-1} = (2,\ 4.5)\begin{pmatrix} 0.70 & -0.98 \\ -0.98 & 1.97 \end{pmatrix} = (-3.01,\ 6.905)$$

$$\overline{\boldsymbol{X}}_2{}'\boldsymbol{S}_2^{-1} = (9.75,\ 11)\begin{pmatrix} 0.03 & -0.01 \\ -0.01 & 0.02 \end{pmatrix} = (0.183,\ 0.123)$$

이므로

$$\overline{\boldsymbol{X}}_1{}'\boldsymbol{S}_1^{-1} - \overline{\boldsymbol{X}}_2{}'\boldsymbol{S}_2^{-1} = (-3.193,\ 6.782)$$

$$\overline{\boldsymbol{X}}_1{}'\boldsymbol{S}_1^{-1}\overline{\boldsymbol{X}}_1 = (2,\ 4.5)\begin{pmatrix} 0.70 & -0.98 \\ -0.98 & 1.97 \end{pmatrix}\begin{pmatrix} 2 \\ 4.5 \end{pmatrix} = 0.920$$

$$\overline{\boldsymbol{X}}_2{}'\boldsymbol{S}_2^{-1}\overline{\boldsymbol{X}}_2 = (9.75,\ 11)\begin{pmatrix} 0.03 & -0.01 \\ -0.01 & 0.02 \end{pmatrix}\begin{pmatrix} 9.75 \\ 11 \end{pmatrix} = 3.137$$

이며 $\boldsymbol{S}_1^{-1} - \boldsymbol{S}_2^{-1} = \begin{pmatrix} 0.67 & -0.97 \\ -0.97 & 1.95 \end{pmatrix}$이다.

표본통계량을 이용하여 식(10.19)의 판별함수를 구하면

$$
\begin{aligned}
Q_s(\boldsymbol{X}) &= \frac{1}{2}\ln\left\{\frac{|\boldsymbol{S}_2|}{|\boldsymbol{S}_1|}\right\} - \frac{1}{2}\left(\overline{\boldsymbol{X}}_1{}'\boldsymbol{S}_1\overline{\boldsymbol{X}}_1 - \overline{\boldsymbol{X}}_2{}'\boldsymbol{S}_2^{-1}\overline{\boldsymbol{X}}_2\right) \\
&\quad + (\overline{\boldsymbol{X}}_1{}'\boldsymbol{S}_1^{-1} - \overline{\boldsymbol{X}}_2{}'\boldsymbol{S}_2^{-1})\boldsymbol{X} - \frac{1}{2}\boldsymbol{X}'(\boldsymbol{S}_1^{-1} - \boldsymbol{S}_2^{-1})\boldsymbol{X} \\
&= \frac{1}{2}ln\left\{\frac{2255.25}{2.37}\right\} - \frac{1}{2}(0.920 - 3.137) + (-3.193 \quad 6.782)\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \\
&\quad - \frac{1}{2}(X_1 \quad X_2)\begin{pmatrix} 0.67 & -0.97 \\ -0.97 & 1.95 \end{pmatrix}\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}
\end{aligned}
$$

와 같이 이차형식의 함수로 주어진다.

10.3 세 개 이상의 그룹의 판별

그룹이 3개 이상인 경우의 판별분석은 g개 그룹의 차이를 가장 크게 하도록 하는 변수들의 선형결합식을 찾는 것이다. 두 그룹에 대한 판별분석을 세 개 이상의 그룹으로 확장하는 것은 분산분석법을 이용하자는 것이다.

이제 g개 그룹으로부터의 확률표본을 얻었으며 i번째 그룹에서는 n_i개의 관측벡터를 얻었다고 하자.

10.3.1 Fisher의 방법

i번째 그룹에서 j번째 관측벡터 $\boldsymbol{X}_{ij}$를 판별계수벡터 $\boldsymbol{a}$를 통해 $Z_{ij}=a'X_{ij}$, $j=1,2,...,n_i$로 변환하고자 하며 각 그룹평균은 $\overline{Z}_i=\boldsymbol{a}'\overline{\boldsymbol{X}}_i$이 된다. $\overline{Z}_1, \overline{Z}_2, ..., \overline{Z}_g$를 분리해내는 식(10.2)의 기준을 g개 그룹에 대해 확장하기 위하여 분산분석에서의 그룹간행렬 $\boldsymbol{B}$와 그룹내행렬 $\boldsymbol{E}$를 이용하여 표현한다.

$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'$대신 $\boldsymbol{B}$ 를, Σ 대신 $\boldsymbol{E}$ 를 이용하게 되어

$$F = \frac{\boldsymbol{a}'\boldsymbol{B}\boldsymbol{a}}{\boldsymbol{a}'\boldsymbol{E}\boldsymbol{a}} \tag{10.21}$$

를 정의한다. 여기서 그룹내 제곱합(within sum of squares)은

$$\begin{aligned} \boldsymbol{E} &= \sum_{l=1}^{g}\sum_{j=1}^{n_l}(\boldsymbol{X}_{lj} - \overline{\boldsymbol{X}_l})(\boldsymbol{X}_{lj} - \overline{\boldsymbol{X}_l})' \\ &= (n_1 - 1)\boldsymbol{S}_1 + (n_2 - 1)\boldsymbol{S}_2 + ... + (n_g - 1)\boldsymbol{S}_g \end{aligned} \tag{10.22}$$

그룹간 제곱합(between sum of squares)은

$$\boldsymbol{B} = \sum_{l=1}^{g} n_l(\overline{\boldsymbol{X}}_l - \overline{\boldsymbol{X}})(\overline{\boldsymbol{X}}_l - \overline{\boldsymbol{X}})' \ . \tag{10.23}$$

F-비를 최대로 하는 $\boldsymbol{a}$는 $\boldsymbol{E}^{-1}\boldsymbol{B}$의 최대 고유값 λ_1에 해당하는 고유벡터 $\boldsymbol{a}_1$이고 이 때의 F값은 λ_1이 된다.

그러므로 판별함수 $Z_1 = \boldsymbol{a}_1'\boldsymbol{Y}$는 그룹평균들의 차이를 가장 크게 해주는 즉 그룹을 구별해주는

함수가 된다. $s = \min(g-1, p)$이며 $\boldsymbol{E}^{-1}\boldsymbol{B}$는 s개의 고유값 $\lambda_1, \lambda_2, \ldots, \lambda_s$와 이에 해당하는 고유벡터 $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_s$를 가지므로 s개의 판별함수

$$Z_i = \boldsymbol{a}_i{}' \boldsymbol{Y}, \quad i = 1, 2, \ldots, s \tag{10.24}$$

를 얻게 되고 서로 독립이며, 이 판별함수들은 $\overline{Z}_1, \overline{Z}_2, \ldots, \overline{Z}_g$의 차이를 크게 나도록 해주는 차원 또는 방향을 나타낸다. 그러므로 $q(\leq s)$개의 정준벡터 $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_q$에 대응하는 정준판별함수 $Z_i = \boldsymbol{a}_i{}' \boldsymbol{Y}, \quad i = 1, 2, \ldots, q$ 를 얻게 되며 다음과 같은 분류규칙이 유도된다 :

새로운 개체 $\boldsymbol{X}_0$에 대하여 모든 $j \neq k$에 대해

$$\sum_{i=1}^{q} \boldsymbol{a}_i{}' (\boldsymbol{X}_0 - \overline{\boldsymbol{X}}_k)^2 \leq \sum_{i=1}^{q} \boldsymbol{a}_i{}' (\boldsymbol{X}_0 - \overline{\boldsymbol{X}}_j)^2 \text{이면}$$

$\boldsymbol{X}_0$를 k번째 그룹인 G_k그룹으로 분류한다. (10.25)

즉 $j = 1, 2, \ldots, g$에 대해 $\sum_{i=1}^{q} \boldsymbol{a}_i{}' (\boldsymbol{X}_0 - \overline{\boldsymbol{X}}_j)^2$을 가장 작게 하는 그룹으로 분류한다.

이와 같이 구해진 q개의 정준판별함수를 사용할 경우 판별력은 누적고유값의 비인

$$\frac{\sum_{i=1}^{q} \lambda_i}{\sum_{i=1}^{s} \lambda_i} \tag{10.26}$$

가 되며 이것을 고려해 판별함수의 개수를 결정하면 된다.

10.3.2 공분산행렬이 모두 같은 경우

$p_1, p_2, \ldots, p_g$는 관측벡터 $\boldsymbol{X}$가 각각 그룹 $G_1, G_2, \ldots, G_g$로부터 발생할 사전확률이고 각 집단의 공분산행렬은 모두 같다고 가정한다. 즉 $\Sigma_1 = \Sigma_2 = \cdots = \Sigma_g$인 경우를 고려한다. 각 그룹으로부터의 확률표본의 개수는 $n_1, n_2, \ldots, n_g$이며 전체 표본크기는 $N = \sum_{i=1}^{g} n_i$이다.

확률밀도함수를 아는 경우 최적 판별규칙은:

$$p_i f(\boldsymbol{X}|G_i) \geq p_j f(\boldsymbol{X}|G_j), \quad j = 1, 2, \ldots, g \tag{10.27}$$

이면 $\boldsymbol{X}$ 를 G_i그룹에 분류한다.

즉 $p_i f(\boldsymbol{X}|G_i) = \max_j p_j f(\boldsymbol{X}|G_j)$이면 $\boldsymbol{X}$를 그룹에 분류한다.

$\boldsymbol{X}$가 다변량 정규분포를 따를 경우를 고려하자. 즉

$$f(\boldsymbol{X} | G_j) = N_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}), \quad j = 1, 2, \ldots, g$$

일 때 $p_i f(\boldsymbol{X} | G_i)$를 최대화하는 것은 $\ln p_i f(\boldsymbol{X} | G_i)$를 최대화하는 것과 동치이므로 $\ln p_i f(\boldsymbol{X} | G_i)$를 구해보면

$$\ln p_i f(\boldsymbol{X} | G_i) = \ln p_i - \frac{1}{2} p \ln(2\pi) - \frac{1}{2} \ln |\boldsymbol{\Sigma}| - \frac{1}{2} (\boldsymbol{X} - \boldsymbol{\mu}_i)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{X} - \boldsymbol{\mu}_i) \qquad (10.28)$$

이다. 여기서 $\boldsymbol{\Sigma}$는 공통공분산행렬이고 p는 $\boldsymbol{X}$를 구성하는 변수 개수이며 p_i는 G_i에서 발생할 수 있는 사전확률이다. 어떤 그룹에서 발생하든지 $\frac{1}{2} p \ln(2\pi)$와 $\frac{1}{2}(\boldsymbol{X}' \boldsymbol{\Sigma}^{-1} \boldsymbol{X})$는 같은 값을 갖게 되므로 식(10.28)에서 중요한 부분은

$$\ln p_i + \boldsymbol{\mu}_i' \boldsymbol{\Sigma}^{-1} \boldsymbol{X} - \frac{1}{2} \boldsymbol{\mu}_i' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_i \qquad (10.29)$$

이 된다. 표본으로부터의 추정량을 이용하면 다음의 선형함수

$$L_i(\boldsymbol{X}) = \ln p_i + \overline{\boldsymbol{X}}_i{}' \boldsymbol{S}_{pl}^{-1} \boldsymbol{X} - \frac{1}{2} \overline{\boldsymbol{X}}_i{}' \boldsymbol{S}_{pl}^{-1} \overline{\boldsymbol{X}}_i \qquad (10.30)$$

를 얻게 된다. 여기서 $\boldsymbol{S}_{pl} = \boldsymbol{E}/(N-g)$은 공통공분산행렬의 추정량이다. 그러므로 판별규칙은:

새로운 관측벡터 $\boldsymbol{X}$를 $L_i(\boldsymbol{X})$를 최대화하는 그룹에 분류한다. (10.31)

그리고 이와 같은 판별규칙은 근사적으로 최적인 분류규칙이 된다.

만약 $p_1 = p_2 = \cdots = p_g$으로 그룹에 대한 사전확률이 같다면 식(10.27)의 판별규칙은 우도함수만을 이용하게 되므로 최대우도를 이용한 규칙(maximum likelihood rule)이 된다.

10.3.3 공분산행렬이 모두 같지 않은 경우

$p_1, p_2, \ldots, p_g$는 관측벡터 $\boldsymbol{X}$가 각각 그룹 $G_1, G_2, \ldots, G_g$로부터 발생할 사전 확률이고 각 집단의 공분산 행렬은 모두 같다고 할 수 없는 경우를 고려해 보자.

$\boldsymbol{X}$가 다변량 정규분포를 따를 경우 즉 $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, $j = 1, 2, \ldots, g$일 때

$$\ln p_i f(\boldsymbol{X} | G_i) = \ln p_i - \frac{1}{2} p \ln(2\pi) - \frac{1}{2} ln|\boldsymbol{\Sigma}_i| - \frac{1}{2}(\boldsymbol{X} - \boldsymbol{\mu}_i)' \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{X} - \boldsymbol{\mu}_i) \qquad (10.32)$$

이다. 여기서 $\boldsymbol{\Sigma}_i$ 는 G_i그룹의 공분산행렬이고 p는 $\boldsymbol{X}$를 구성하는 변수 개수이며 p_i 는 G_i에서 발생할 수 있는 사전확률이다. 식(10.28)에서 중요한 부분은

$$\begin{aligned} &\ln p_i - \frac{1}{2} ln|\boldsymbol{\Sigma}_i| - \frac{1}{2}(\boldsymbol{X} - \boldsymbol{\mu}_i)' \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{X} - \boldsymbol{\mu}_i) \qquad (10.33) \\ &= \ln p_i - \frac{1}{2} ln|\boldsymbol{\Sigma}_i| - \frac{1}{2} \boldsymbol{\mu}_i' \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\mu}_i \\ &\qquad - \frac{1}{2} \boldsymbol{X}' \boldsymbol{\Sigma}_i^{-1} \boldsymbol{X} + \boldsymbol{\mu}_i' \boldsymbol{\Sigma}_i^{-1} \boldsymbol{X} \end{aligned}$$

표본으로부터의 추정량을 이용하면 다음의 이차형식함수

$$Q_i(\boldsymbol{X}) = \ln p_i - \frac{1}{2} ln|\boldsymbol{S}_i| - \frac{1}{2}\overline{\boldsymbol{X}}_i{}'\boldsymbol{S}_i^{-1}\overline{\boldsymbol{X}}_i \qquad (10.34)$$

$$- \frac{1}{2}\boldsymbol{X}'\boldsymbol{S}_i^{-1}\boldsymbol{X} + \overline{\boldsymbol{X}}_i{}'\boldsymbol{S}_i^{-1}\boldsymbol{X}$$

를 얻는다. 그러므로 판별규칙은:

새로운 관측벡터 $\boldsymbol{X}$를 $Q_i(\boldsymbol{X})$를 최대화하는 그룹에 분류한다. (10.35)

그리고 이와 같은 판별규칙은 근사적으로 최적인 분류규칙이다. 만약 $p_1 = p_2 = \cdots = p_g$이라면 각 그룹에서 $\ln p_i$는 같으므로 식(10.34)은

$$Q_i(\boldsymbol{X}) = - \frac{1}{2} ln|\boldsymbol{S}_i| - \frac{1}{2}\overline{\boldsymbol{X}}_i{}'\boldsymbol{S}_i^{-1}\overline{\boldsymbol{X}}_i \qquad (10.36)$$

$$- \frac{1}{2}\boldsymbol{X}'\boldsymbol{S}_i^{-1}\boldsymbol{X} + \overline{\boldsymbol{X}}_i{}'\boldsymbol{S}_i^{-1}\boldsymbol{X}$$

10.4 오분류율 계산

분류함수의 능력을 판단하기 위한 오분류(misclassification)의 확률 계산

(1) 재대입 분류에 의한 오류율 계산

(2) 표본분할에 의한 오류율 계산

(3) 교차타당성에 의한 오류율 계산

(4) 붓스트랩 이용한 오류율 계산

10.4.1 재대입 분류에 의한 오류율 계산

데이터로부터 유도된 판별함수를 다시 데이터에 적용하는 재대입(resubstution)분류에 의해 오분류율을 계산할 수 있다.

[표 10.1] 2개 그룹 분류표

		예측 그룹	
실제 그룹	**표본의 수**	**1**	**2**
1	n_1	n_{11}	n_{12}
2	n_2	n_{21}	n_{22}

$$\text{명백한 오류율} = \frac{n_{12}+n_{21}}{n_1+n_2} = \frac{n_{12}+n_{21}}{n_{11}+n_{12}+n_{21}+n_{22}} \qquad (10.37)$$

$$\text{정확한 분류율} = \frac{n_{11}+n_{22}}{n_1+n_2} \qquad (10.38)$$

$$\text{정확한 분류율} = 1 - \text{오류율}$$

이번에는 3개 그룹의 경우 [표 10.2]와 같은 분류표를 만들고 오류율을 정의해 보자.

[표 10.2] 3개 그룹 분류표

		예측 그룹		
실제 그룹	**표본의 수**	**1**	**2**	**3**
1	n_1	n_{11}	n_{12}	n_{13}
2	n_2	n_{21}	n_{22}	n_{23}
3	n_3	n_{31}	n_{32}	n_{33}

$$\text{정확한 분류율} = \frac{n_{11}+n_{22}+n_{33}}{n_1+n_2+n_3} \tag{10.39}$$

$$\begin{aligned}\text{명백한 오류율} &= \frac{n_{12}+n_{21}+n_{21}+n_{23}+n_{31}+n_{32}}{n_1+n_2+n_3} \\ &= 1-\text{정확한 분류율}\end{aligned} \tag{10.40}$$

≪예제 10.4≫ 판별 분석 결과 다음의 [표 10.3]과 같은 분류결과를 얻었다.

[표 10.3] 2개 그룹 분류표 예

실제 그룹	표본의 수	예측 그룹 1	예측 그룹 2
1	32	28	4
2	32	4	28

$$\text{명백한 오류율} = \frac{n_{12}+n_{21}}{n_1+n_2} = \frac{4+4}{32+32} = 0.125$$

$$\text{정확한 분류율} = \frac{n_{11}+n_{22}}{n_1+n_2} = \frac{28+28}{32+32} = 0.875$$

10.4.2 표본분할에 의한 오류율 계산

▶ 재대입 분류에 의해 오류율을 계산하면 실제 편의(bias)보다 적게 계산될 수 있다.

▶ 편의를 줄이는 방법으로 표본에 대해 표본의 크기를 반으로 하는 두 부분으로 나누어, 한 부분은 판별함수를 만드는데 이용하고 나머지 부분은 만든 판별함수를 이용해 분류한 후 판별함수의 판별능력을 평가해 보는 것이다.

[단점]

(1) 표본을 두 부분으로 나누어야하므로 비교적 큰 표본의 크기가 요구된다.

(2) 실제로 사용할 판별함수에 대해서는 평가할 수 없다.

10.4.3 교차타당성에 의한 오류율 계산

▶ 표본분할 방법보다 진보된 방법으로 교차타당성(cross-validation)에 의한 방법.

▶ 한 개만의 표본을 제외한 나머지 표본으로 판별함수를 계산하고 구해진 판별함수를 이용해 제외되었던 표본을 분류한다. 이와 같은 과정을 전체 표본의 크기만큼 시행하여 오류율을 구한다. 한 번에 한 개의 표본을 제외하고 시행하므로 한 개씩 제거 방법(leaving-one-out method)라고 불리기도 한다.

▶ 표본이 커지면 계산횟수와 시간이 엄청나게 늘어나는 불편이 있지만 컴퓨터의 발달로 해결되는 문제이기도 하다.

▶ 교차타당성에 의해 구한 오류율은 재대입분류로 구한 오류율보다 커지는 경향이 있다.

10.5 판별함수의 표준화

▶ 판별함수에 기여하는 변수의 상대적 비중은 판별함수계수(discriminant function coefficient)에 나타난다.

▶ 변수값들의 단위가 다르면 동등한 비교를 할 수 없으므로 단위가 다를 때는 표준화 변수를 이용하여 판별함수를 구한 후 판별함수의 계수를 비교.

▶ 두 개의 그룹이 있으며 공분산행렬이 같다고 할 수 있는 경우를 생각해 보자.

•원래 변수를 이용해 구한 선형판별함수:

$$Y = \boldsymbol{a}'\boldsymbol{X} = a_1 X_1 + a_2 X_2 + \cdots + a_p X_p \quad . \qquad (10.41)$$

•표준화 변수를 이용할 경우 그룹1에서 i번째 관측벡터에 대해 판별함수를 구하면

$$Z_{1i} = \boldsymbol{a}^{*\prime}\boldsymbol{X} = a_1^* \frac{X_{1i1} - \overline{X_{11}}}{s_1} + a_2^* \frac{X_{1i2} - \overline{X_{12}}}{s_2} + \cdots + a_p^* \frac{X_{1ip} - \overline{X_{1p}}}{s_p} \qquad (10.42)$$

여기서 $\overline{\boldsymbol{X}}_1 = (\overline{X_{11}}, \overline{X_{12}}, \ldots, \overline{X_{1p}})'$는 그룹1의 평균벡터, $i = 1, 2, \ldots, n_1$.

그룹2에서 i/번째 관측벡터에 대해 판별함수를 구하면, $i = 1,2,...,n_2$에 대해

$$Z_{2i} = \boldsymbol{a}^{*\prime}\boldsymbol{X} = a_1^* \frac{X_{2i1} - \overline{X_{21}}}{s_1} + a_2^* \frac{X_{2i2} - \overline{X_{22}}}{s_2} + \cdots + a_p^* \frac{X_{2ip} - \overline{X_{2p}}}{s_p} \qquad (10.43)$$

여기서 $\overline{\boldsymbol{X}}_2 = (\overline{X_{21}}, \overline{X_{22}}, ..., \overline{X_{2p}})'$는 그룹2의 평균벡터이다.

s_r은 두 그룹의 합동공분산행렬 $\boldsymbol{S}_{pl}$에서 r번째 대각선상에 놓인 r번째 변수의 표준편차이다.

$$a_r^* = s_r a_r, \; r = 1,2,...,p. \qquad (10.44)$$

의 관계를 갖게 되며, 벡터로 표현하면

$$\boldsymbol{a}^* = (diag\boldsymbol{S}_{pl})^{1/2}\boldsymbol{a} \;. \qquad (10.45)$$

▶ 여러 개의 그룹이 있으며 공분산행렬이 같다고 할 수 있는 경우를 생각해보자. 두 개 그룹의 경우와 마찬가지로 $m=1,2,...,g,\ r=1,2,...,p$에 대해 표준화계수는

$$a_{mr}^{*}=s_{r}a_{mr} \tag{10.46}$$

의 관계를 만족하며 여기서 s_r은 합동공분산행렬 $\boldsymbol{S}_{pl}$에서 r번째 대각선상에 놓인 r번째 변수의 표준편차이다.

10.6 판별함수의 유의성 검정

Fisher의 판별함수는 집단 또는 그룹간의 평균차가 최대가 되도록 판별함수를 구하는 것. 판별함수의 유의성 검정을 위해서는 다변량 정규성 가정이 필요하다.

10.6.1 두 개 그룹의 경우 판별함수의 유의성 검정

두 그룹의 분리를 위래 두 그룹간 거리가 최대가 되도록 만든 판별함수의 계수벡터는

$$l = \Sigma^{-1}(\mu_1 - \mu_2) \tag{10.47}$$

판별함수 계수벡터의 유의성에 대한 귀무가설

$$H_0 : l = 0 \tag{10.48}$$

=> $$H_0 : \mu_1 = \mu_2 \tag{10.49}$$

두 그룹의 평균비교에 대한 가설검정은 Hotelling의 T^2 검정을 이용.

10.6.2 여러 개 그룹의 경우 판별함수의 유의성 검정

그룹이 여러 개 있을 경우, s개 그룹(모집단)의 판별함수를 $a_1'X, a_2'X, \ldots, a_s'X$로 표기하자. 판별함수 계수벡터의 유의성에 대한 귀무가설은

$$H_0 : \boldsymbol{a}_1 = \boldsymbol{a}_2 = \cdots = \boldsymbol{a}_s = \mathbf{0} \qquad (10.50)$$

10.3.1절의 내용을 다시 살펴보면 a는 $\Sigma^{-1}\Omega$의 고유벡터가 되며 여기서 $\Omega = \sum_{l=1}^{g} n_l(\boldsymbol{\mu}_l - \overline{\boldsymbol{\mu}})(\boldsymbol{\mu}_l - \boldsymbol{\mu})'$이다.

H_0가 사실이면, 즉 고유값이 모두 0인 행렬은 $\mathbf{0}$ 행렬밖에 없으므로 $\Sigma^{-1}\Omega = 0$이 되며 $\Sigma \neq 0$이므로 $\Omega = 0$이 된다. 즉 $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_g$인 경우가 되며 6장에서 다룬 그룹간 평균벡터 비교 문제와 같은 문제가 된다.

10.3.1절의 내용을 다시 살펴보면 판별기준에 해당하는

$$F=\frac{a' Ba}{a' Ea} \tag{10.51}$$

는 $E^{-1}B$의 최대 고유값 λ_1에 해당하는 고유벡터 a_1이고 이 때의 F값은 λ_1이 되며 나머지 $\lambda_2,\ldots,\lambda_s$에 해당하는 고유벡터와 더불어 판별차원(discriminant dimensions)을 구성한다. 여기서 고유값은 6장에서 다룬 여러 개 집단에서 평균벡터 비교에 대한 검정문제에서 Wilks lambda 통계량 계산에 쓰이는 고유값과 같다. 그러므로

Wilks lambda 통계량

$$\Lambda_1=\frac{|E|}{|B+E|}=\prod_{i=1}^{s}\frac{1}{1+\lambda_i} \tag{10.52}$$

을 이용하여 판별함수에 대한 유의성 검정을 할 수 있다.

$N=\sum_{i=1}^{g} n_i$ 일 때, 유의수준 α에서의 검정법은 $\Lambda_1 \leq \Lambda_{p,g-1,N-p-1}(\alpha)$이면

$H_0 : \mu_1=\mu_2=\cdots=\mu_g$를 기각한다.

H_0가 기각되고 s개의 고유값으로 판별차원을 형성한다면

$\overline{\boldsymbol{X}_1}, \overline{\boldsymbol{X}_2}, \ldots, \overline{\boldsymbol{X}_g}$ 개의 평균벡터를 s-차원으로 분할한다는 의미를 갖는다.

그러나 H_0를 기각하는 것이 s개의 판별차원을 말해주는 것이 아니므로 평균벡터의 차이가 유의한 평균벡터들의 부분집합으로 다시 그룹을 만들어 구별이 유의한 판별차원을 찾을 수 있다.

Λ_1에 대해

$$
\begin{aligned}
V_1 &= -\left[\nu_E - \frac{1}{2}(p - \nu_B + 1)\right] \ln \Lambda_1 \qquad (10.53) \\
&= -\left[N - g - \frac{1}{2}(p - g + 2)\right] \ln \prod_{i=1}^{s} \frac{1}{1+\lambda_i} \\
&= \left[N - g - \frac{1}{2}(p + g)\right] \sum_{i=1}^{s} \ln(1+\lambda_i)
\end{aligned}
$$

는 근사적으로 $\chi^2_{p(g-1)}$분포를 따르므로 V_1을 이용하여 카이제곱검정을 할 수 있다.

만약 H_0를 기각한다면 적어도 한 개의 고유값은 0이 아니며

첫 번째 고유값에 해당하는 첫 번째 판별함수

$$Z_1 = \boldsymbol{a}_1{}'\boldsymbol{X} \tag{10.54}$$

는 그룹을 구별하는데 의미있는 함수가 된다.

그러면 첫 번째 고유값을 제외한 나머지 고유값에 대한 검정을 해보자.

$\lambda_2,\ldots,\lambda_p$에 대한 유의성 검정을 위해

Wilks lambda 통계량

$$\Lambda_2 = \prod_{i=2}^{s} \frac{1}{1+\lambda_i} \tag{10.55}$$

$$V_2 = \left[N - g - \frac{1}{2}(p+g)\right] \ln \sum_{i=2}^{s} (1+\lambda_i) \tag{10.56}$$

는 근사적으로 $\chi^2_{(p-1)(g-2)}$ 분포를 따르므로 V_2를 이용하여 카이제곱검정을 할 수 있다.

만약 H_0를 기각한다면 적어도 한 개의 고유값은 0이 아니며

두 번째 고유값에 해당하는 두 번째 판별함수

$$Z_2 = \boldsymbol{a}_2{}'\boldsymbol{X} \tag{10.57}$$

는 그룹을 구별하는데 의미있는 함수가 되며 판별차원의 축이 된다.

이와 같은 방법으로 판별함수에 대한 유의성 검정을 계속할 수 있다.

m번째 단계에서는 $\lambda_m,...,\lambda_p$에 대한 유의성 검정을 위해 Wilks lambda 통계량

$$\Lambda_m = \prod_{i=m}^{s} \frac{1}{1+\lambda_i} \tag{10.58}$$

를 계산할 수 있고

$$V_m = \left[N - g - \frac{1}{2}(p+g)\right] \ln \sum_{i=m}^{s} (1+\lambda_i) \tag{10.59}$$

는 근사적으로 $\chi^2_{(p-m+1)(g-m)}$ 분포를 따르므로 V_m을 이용하여 카이제곱검정을 할 수 있다.

또한 Λ_i에 대해 근사적인 F-검정을 이용할 수 있다. Λ_1에 대해서

$$F = \frac{1-\Lambda_1^{1/t}}{\Lambda_1^{1/t}} \frac{df_2}{df_1} \tag{10.60}$$

$$t = \sqrt{\frac{p^2(g-1)^2-4}{p^2+(g-1)^2-5}}, \qquad w = N-1-\frac{1}{2}(p+g) \tag{10.61}$$

$$df_1 = p(g-1), \qquad df_2 = wt - \frac{1}{2}[p(g-1)-2].$$

$\Lambda_m, \ m=2,3,\ldots,s$에 대해서는

$$F=\frac{1-\Lambda_m^{1/t}}{\Lambda_m^{1/t}}\frac{df_2}{df_1} \tag{10.62}$$

p를 $p-m+1$로, $k-1$을 $k-m$으로 대치하면

$$t=\sqrt{\frac{(p-m+1)^2(g-m)^2-4}{(p-m+1)^2+(g-m)^2-5}}, \qquad w=N-1-\frac{1}{2}(p+g)$$

$$df_1=(p-m+1)(g-m), \qquad df_2=wt-\frac{1}{2}[(p-m+1)(g-m)-2]$$

(10.63) F-통계량은 근사적으로 F_{df_1,df_2} 분포를 따르므로 F-검정을 할 수 있다.

10.7 판별함수의 변수선택 : 단계별 변수선택

10.7.1 변수선택방법

판별함수에 사용될 수 있는 변수가 여러 개 있을 경우 적절한 변수선택을 통해 판별함수를 찾으면 판별차원을 줄일 뿐만 아니라 오분류율을 낮출 수 있다.

▶ 전진적 선택(forward selection) 방법:

- 그룹간의 거리를 최대로하는 한 개의 변수를 선택한 후 그 다음으로 그룹간의 거리를 최대로하는 다른 한 개의 변수를 선택.
- 부분 F-검정통계량의 값을 최대로하는 변수가 선택된다.
- 이와 같은 방법으로 변수가 더 이상 선택되지 않을 때까지 진행한다.

▶ 후진적 선택(backward selection) 방법

- 모든 변수로부터 시작한다.
- 부분 F-검정통계량의 값이 가장 작은 변수부터 제거해가며 더 이상 변수가 제거되지 않을 때까지 진행한다.

▶ 단계적 선택(stepwise selection) 방법

•전진적 선택방법과 후진적 선택방법의 혼합형.

•변수가 선택되는 매 단계마다 선택된 변수가 기여하는 바를 검정하여 결정한다.
매단계 MANOVA 과정을 수행하는 것이다.

•변수 선택 과정이 끝나면 선택된 변수를 이용하여 판별함수를 구하게 된다.

▶ 이와 같이 유의한 변수를 선택하여 판별함수를 만드는 작업을 단계적 판별분석(stepwise discriminant analysis).

10.7.2 부분 F-통계량을 이용한 단계별 변수선택

▶ 전진적 변수 선택과 후진적 변수 선택의 절충형인 단계별 변수 선택방법에 대해 부분 F-통계량(partial F-statistic)을 이용한 변수 선택과정을 설명하기로 한다.

▶ 일변량에 대한 분산분석의 F-통계량을 비교하여 가장 큰 값을 갖는 변수가 첫 번째로 선택된다. 첫 번째로 X_1이 선택되었다고 하자. X_1에 대한 Wilks lambda $\Lambda(X_1)$을 구한다. 그리고 각 변수 X_r, $r=2,3,\ldots,p$에 대해 부분(partial) Λ 통계량

$$\Lambda(X_r|X_1)=\frac{\Lambda(X_1,X_r)}{\Lambda(X_1)} \tag{10.64}$$

와 부분 F-통계량

$$F=\frac{1-\Lambda(X_r|X_1)}{\Lambda(X_r|X_1)}\frac{\nu_E-1}{\nu_B} \tag{10.65}$$

을 계산하여 가장 큰 부분 F-값을 갖는 변수를 선택한다. 여기서 $\nu_E=N-g$, $\nu_B=g-1$이다.

이 단계에서 X_2가 선택되었다고 하자.

각 변수 $X_r,\ r=3,4,...,p$ 에 대해 부분 Λ 통계량

$$\Lambda(X_r|X_1,X_2)=\frac{\Lambda(X_1,X_2,X_r)}{\Lambda(X_1,X_2)} \tag{10.66}$$

부분 F-통계량

$$F=\frac{1-\Lambda(X_r|X_1,X_2)}{\Lambda(X_r|X_1,X_2)}\ \frac{\nu_E-2}{\nu_B} \tag{10.67}$$

을 계산하여 가장 큰 부분 F-값을 갖는 변수를 선택한다.

이 단계에서 X_3가 선택되었다고 하자.

모형에 포함된 변수들에 대해 부분 Λ 통계량

$$\Lambda(X_1|X_2,X_3),\ \Lambda(X_2|X_1,X_3),\ \Lambda(X_3|X_1,X_2)$$

를 계산하고 비교하여 모형에의 기여정도가 유의하지 않은 변수는 제거된다.

이와 같은 방법으로 더 이상 변수가 선택되거나 제거되지 않는 단계에서 과정은 멈추게 된다.

10.8 R을 이용한 판별분석

≪예제 10.4≫ 두 시대의 이집트인의 두개골에 대한 자료, 다음과 같은 4개의 변수를 가지고 있다. 각 모집단은 다변량 정규분포를 따르고, 같은 분산을 갖는다는 가정 하에 이 자료를 이용하여 판별분석을 수행해 보고자 한다.

[표 10.4]에서 X_1, X_2, X_3, X_4는 각각 MB, BH, BL, NH, Year를 나타내며 각각의 구체적인 의미는 다음과 같다.

Year = 두개골 형성 시기의 대략적인 연도로 기원전 4000년, 기원후 150년

MB = Maximum Breadth(㎜) : 머리뼈의 둘레 중 가장 큰 둘레의 길이

BH = Basibregmatic Height(㎜) : 기저시상봉합과 관상봉합의 접합점의 크기로 머리뼈의 정수리부터 눈썹뼈 위까지의 길이

BL = Basialveolar Length(㎜) : 기조치조의 길이로 맨 앞의 치아부터 혀가 닿는 끝부분까지의 길이

NH = Nasal Height of Skull(㎜) : 코의 높이

[표 10.4] 두개골 자료

번호	X_1	X_2	X_3	X_4	연대	번호	X_1	X_2	X_3	X_4	연대
1	131	138	89	49	-4000	1	137	123	91	50	150
2	125	131	92	48	-4000	2	136	131	95	49	150
3	131	132	99	50	-4000	3	128	126	91	57	150
4	119	132	96	44	-4000	4	130	134	92	52	150
5	136	143	100	54	-4000	5	138	127	86	47	150
6	138	137	89	56	-4000	6	126	138	101	52	150
7	139	130	108	48	-4000	7	136	138	97	58	150
8	125	136	93	48	-4000	8	126	126	92	45	150
9	131	134	102	51	-4000	9	132	132	99	55	150
10	134	134	99	51	-4000	10	139	135	92	54	150
11	129	138	95	50	-4000	11	143	120	95	51	150
12	134	121	95	53	-4000	12	141	136	101	54	150
13	126	129	109	51	-4000	13	135	135	95	56	150
14	132	136	100	50	-4000	14	137	134	93	53	150
15	141	140	100	51	-4000	15	142	135	96	52	150
16	131	134	97	54	-4000	16	139	134	95	47	150
17	135	137	103	50	-4000	17	138	125	99	51	150
18	132	133	93	53	-4000	18	137	135	96	54	150
19	139	136	96	50	-4000	19	133	125	92	50	150
20	132	131	101	49	-4000	20	145	129	89	47	150
21	126	133	102	51	-4000	21	138	136	92	46	150
22	135	135	103	47	-4000	22	131	129	97	44	150
23	134	124	93	53	-4000	23	143	126	88	54	150
24	128	134	103	50	-4000	24	134	124	91	55	150
25	130	130	104	49	-4000	25	132	127	97	52	150
26	138	135	100	55	-4000	26	137	125	85	57	150
27	128	132	93	53	-4000	27	129	128	81	52	150
28	127	129	106	48	-4000	28	140	135	103	48	150
29	131	136	114	54	-4000	29	147	129	87	48	150
30	124	138	101	46	-4000	30	136	133	97	51	150

● library(MASS)와 lda() 함수로 선형 판별분석을 수행하여 선형판별함수를 얻을 수 있으며 qda() 함수를 이용하여 이차판별분석을 수행한다.

ldahist() 함수에서 옵션으로 type=c("histogram", "density", "both") 을 이용하여 각 변수의 그룹별 분포에 대한 히스토그램, 확률밀도함수, 히스토그램과 확률밀도함수 곡선을 동시에 그릴 수 있다.

정규성을 가정하고 합동공분산행렬을 이용하여 구했을 때 선형판별함수가

$$Y_1 = 0.11876404\, X_1 \pm 0.08410818\, X_2 - 0.10842595\, X_3 + 0.02565593\, X_4$$

임을 보여준다.

[프로그램 10.1] 두개골 자료에 대한 선형 판별분석

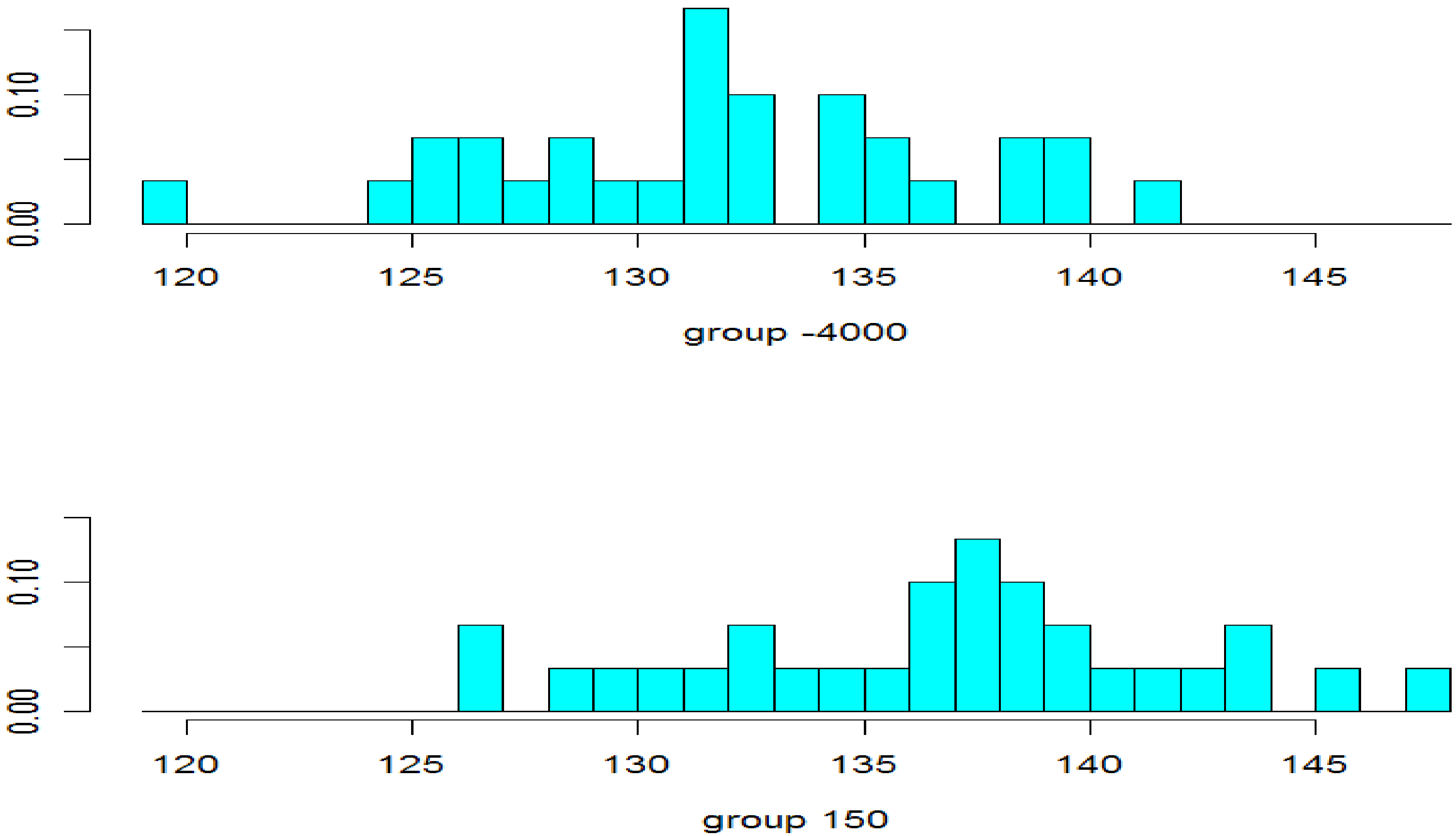
```
skull=read.csv("C:/data/skull.csv", header=T)
skull
attach(skull)
n=dim(skull)[[1]]
n
x=skull[,2:5]
x

#Simple histogram by the grouping variable
ldahist(data = x1, g = year, type="histogram")
ldahist(data = x2, g = year, type="density")
ldahist(data = x3, g = year, type="both")
ldahist(data = x4, g = year)
```

```
library(MASS)
ld = lda(year ~ x1+x2+x3+x4, data=skull)
                              # linear discriminat analysis
ld
pc=predict(ld, skull)$class
pc=as.numeric(pc)
pc
pc[(pc==1)]=-4000
pc[(pc==2)]=150
pc

res=cbind(year, pc)
correct= res[(year==pc),]    # match
correct.rate= dim(correct)[[1]]/n
correct.rate
error.rate=1-correct.rate
error.rate
```

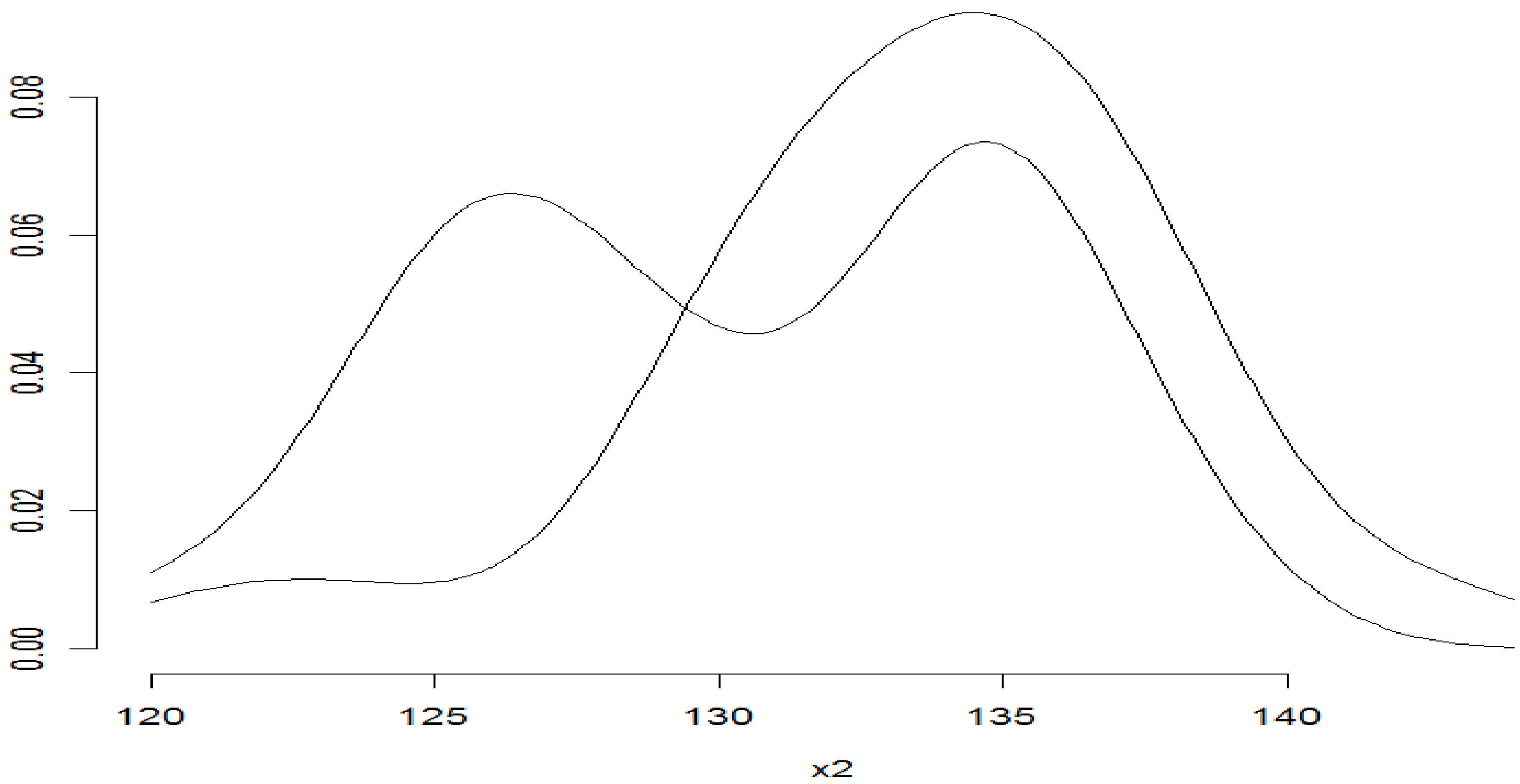
[결과 10.1] 두개골 자료에 대한 선형 판별분석 결과

```
> library(MASS)
> ld = lda(year ~ x1+x2+x3+x4, data=skull)  # linear discriminat analysis
> ld
Call:
lda(year ~ x1 + x2 + x3 + x4, data = skull)
Prior probabilities of groups:
-4000   150
  0.5   0.5
Group means:
            x1       x2       x3       x4
-4000 131.3667 133.6000 99.16667 50.53333
150   136.1667 130.3333 93.50000 51.36667
Coefficients of linear discriminants:
           LD1
x1   0.11876404
x2 -0.08410818
x3 -0.10842595
x4   0.02565593
```

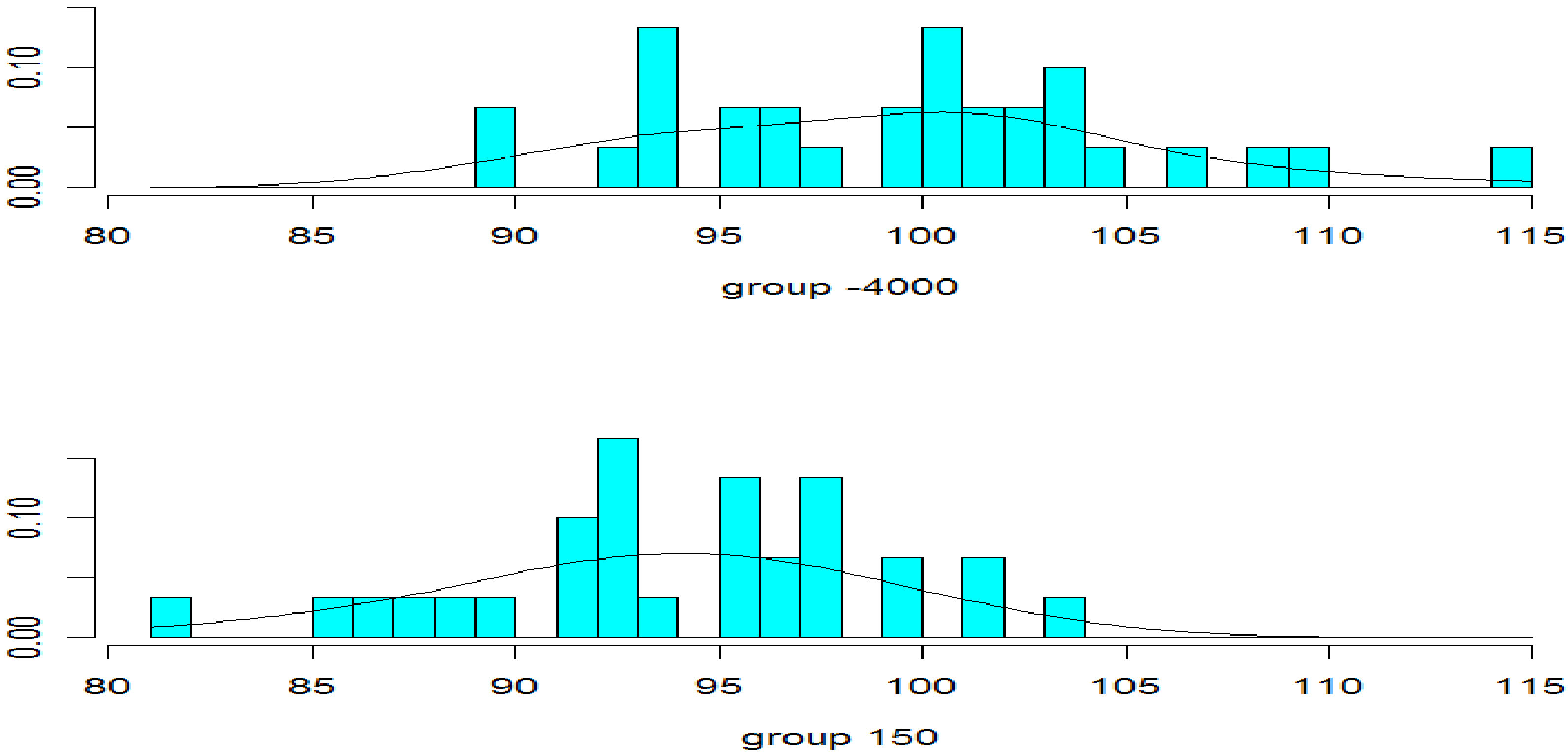
```
> pc=predict(ld, skull)$class
> pc=as.numeric(pc)
> pc[(pc==1)]=-4000
> pc[(pc==2)]=150
> pc

> res=cbind(year, pc)
> correct= res[(year==pc),]    # match
> correct.rate= dim(correct)[[1]]/n
> correct.rate
[1] 0.8
> error.rate=1-correct.rate
> error.rate
[1] 0.2
```



[그림10.4] x1에 대한 연도별 히스토그램



[그림10.5] x2에 대한 연도별 확률밀도함수 추정선



[그림10.6] x3에 대한 연도별 히스토그램과 확률밀도함수 추정선

[프로그램 10.2] 두개골 자료에 대한 이차 판별분석

```
library(MASS)
n=dim(skull)[[1]]
qd = qda(x, year)      # Quadratic discriminant analysis
qd
qc=predict(qd)$class
qc=as.numeric(qc)
qc
qc[(qc==1)]=-4000
qc[(qc==2)]=150
qc

resq=cbind(year, qc)
correctq= resresq[(year==qc),]    # match
correctq.rate= dim(correctq)[[1]]/n
correctq.rate
errorq.rate=1-correctq.rate
errorq.rate
```

[결과 10.2] 두개골 자료에 대한 이차 판별분석 결과

```
> qd = qda(x, year)     # Quadratic discriminant analysis
> qd
Call:
qda(x, year)
Prior probabilities of groups:
-4000   150
  0.5   0.5
Group means:
            x1       x2       x3       x4
-4000 131.3667 133.6000 99.16667 50.53333
150   136.1667 130.3333 93.50000 51.36667
> qc=predict(qd)$class
> qc=as.numeric(qc)
```

```
> qc
 [1] 1 1 1 1 1 2 1 1 1 1 1 2 1 1 1 1 1 1 2 1 1 1 2 1 1 1 1 1 1 1 2 2 2 1 2 1 2 2 2 2 2 1
2 2 2 2 2 2 2 2 2 2 2 2 1 2 2 2 2 2
> qc[(qc==1)]=-4000
> qc[(qc==2)]=150
> qc
> resq=cbind(year, qc)
> correctq= resq[(year==qc),]    # match
> correctq.rate= dim(correctq)[[1]]/n
> correctq.rate
[1] 0.8666667
> errorq.rate=1-correctq.rate
> errorq.rate
[1] 0.1333333
```

[프로그램 10.3] 선형 판별분석시 leave-one-out cross validation 사용하는 경우

```
##### with CV:leave-one-out cross validation
ldc = lda(year ~ x1+x2+x3+x4, data=skull, CV = TRUE, prior=c(1/2,1/2))
names(ld)
results=data.frame(year, ldc$class, ldc$posterior)
results[1:10,]      # 그룹에 대한 사후 확률 리스트

#Summarize crossvalidation
class.table= table(year, ldc$class)
class.table
```

```
#One could make the summary of the classifications nicer by
# writing a function
summarize.class= function(original, classify) {
    class.table<-table(original, classify)
    numb<-rowSums(class.table)
    prop<-round(class.table/numb,4)
    list(class.table = class.table, prop = prop)
                  #return a list object type
}

summarize.class(original = year, classify = ldc$class)
```

[결과 10.3] [프로그램 10.3] 수행 결과

```
> ldc = lda(year ~ x1+x2+x3+x4, data=skull, CV = TRUE, prior=c(1/2,1/2))
> names(ld)
[1] "class"     "posterior" "terms"     "call"      "xlevels"
> results<-data.frame(year, ldc$class, ldc$posterior)
> results[1:10,]
   year ldc.class    X.4000       X150
1  -4000       150 0.4534631 0.54653691
2  -4000     -4000 0.6613711 0.33862894
3  -4000     -4000 0.7177983 0.28220171
4  -4000     -4000 0.9397626 0.06023737
5  -4000     -4000 0.7807131 0.21928691
6  -4000       150 0.1076276 0.89237244
7  -4000     -4000 0.6189429 0.38105712
8  -4000     -4000 0.8152152 0.18478485
9  -4000     -4000 0.8340035 0.16599651
10 -4000     -4000 0.6498000 0.35020005
```

```
> #Summarize crossvalidation
> class.table<-table(year, ldc$class)
> class.table
year    -4000 150
  -4000    23   7
  150       8  22
> summarize.class(original = year, classify = ldc$class)
$class.table
        classify
original -4000 150
   -4000    23   7
   150       8  22
$prop
        classify
original  -4000    150
   -4000 0.7667 0.2333
   150   0.2667 0.7333
```