■ 연구논문

성장곡선모형의 판별분석에서 균형이차분류법의 적용

심규박
동국대학교 전산통계학과

An Application of the Balanced Quadratic Classification Rule on the Discriminant Analysis in Growth Curve Model

Kyu-Bark Shim
Dept. of Computer Science and Statistics, Dongguk University

Abstract

The problem considered here is to find the optimal discriminant analysis method in growth curve model. It has been studied how to find correct prior probability for the effective classification in discriminant analysis. We use the balanced condition to calculate prior probability. From the informative simulation study, new classification rule for the growth curve model is suggested. The suggested classification rule has better classification result than the other previously suggested method in terms of error rate criterion.

1. 서론

여러가지 사회현상에 대한 분석이나, 생물 의학적인 진단에서 다양한 특성을 가진 자료의 이용이 늘어나고 있다. 성장곡선모형(growth curve model)도 그 가운데 하나인데, 동식물의 성장이나 수출입 물량의 증가등에 관한 자료들은 성장곡선모형을 따른다고 할 수 있다. 이와 같은 자료들을 분석하는 한 방법으로 판별분석(discriminant analysis)이 있는데, 이에 대해서는 Fisher(1936)의 연구 이후 활발하게 진행되어 오고 있다.

2그룹 판별분석(two-group discriminant analysis)이란 각 모집단의 특성을 나타내는 다변량 자료를 바탕으로 개체(individual)들을 2개의 모집단 가운데 하나의 모집단으로 분류하는 분석방법이다. 판별분석에서 가장 일반적인 가정은 v가 관측치들의 p차원 벡터이고 π_i, $i=2$로 부터의 표본이라면, 이것은 평균벡터 μ_i, 공분산 행렬 $\sum_i$를 가진 다변량

정규분포를 따른다는 것이다. 관측치 v를 2개의 공분산이 다른 모집단 π_1, π_2에 분류하는 데 가장 많이 쓰는 규칙은

$$Q = (v-\mu_2)^T\sum_2^{-1}(v-\mu_2)-(v-\mu_1)^T\sum_1^{-1}(v-\mu_1)-\log\frac{|\sum_1|}{|\sum_2|} \geq 2\log\frac{q_2}{q_1} \qquad (1.1)$$

일 경우 v를 π_1에 분류하고, 그렇지 않은 경우 π_2에 분류하는 것이다. 여기서, q_1, q_2는 하나의 개체가 각각 π_1, π_2에서 나올 사전확률이다. 이 법칙을 이차분류법칙(quadratic classification rule : QCR)이라 하며, $\sum_1=\sum_2$일 경우 선형분류법칙(linear classification rule : LCR)이라 한다.

이들 분류법칙에서는 다변량 정규모집단 구성원에 대한 미지의 사전확률 q_i, $i=1, 2$를 고려해야 한다. 이 때, 오분류의 총확률은

$$q_1 Pr(V\in\pi_2 \mid V\in\pi_1) + q_2 Pr(V\in\pi_1 \mid V\in\pi_2), \quad q_1+q_2=1 \qquad (1.2)$$

로 정의되며, $Pr(V\in\pi_i \mid V\in\pi_j)$는 실제로 π_j에 속하는 개체를 π_i에 잘못 할당할 확률이다. q_i값이 주어지면 이 확률은 식 (1.1)을 만족할 때 마다 개체를 π_1에 할당하는 이차분류법칙에 의해 최소화 되며, π_2에 분류하였을 경우에도 같은 원리이다.

따라서, q_i값은 오분류의 총오차를 최소화 하는데 중요한 역할을 한다. 표본으로부터의 추정치를 사용하여 위 분류법칙을 다시 정의하면

$$Q = (v-\bar{x}_2)^T S_2^{-1}(v-\bar{x}_2)-(v-\bar{x}_1)^T S_1^{-1}(v-\bar{x}_1)-\log\frac{|S_1|}{|S_2|} \geq 2\log\frac{\hat{q}_2}{\hat{q}_1} \qquad (1.3)$$

이며, 이 경우 v를 π_1에 분류한다. 여기서, $\bar{x}_i$와 S_i는 각각 μ_i와 $\sum_i$, $i=1, 2$의 표본추정량을 의미하며, $\hat{q}_i$는 q_i의 추정치이다. 사전확률 q_i, $i=1, 2$가 미지일 때, 식 (1.3)에서 절단점 $C=2\log(\hat{q}_2/\hat{q}_1)$는 여러가지 방법에 따라 선택할 수 있으며, 판별분석에서 이에 대한 연구는 여러학자들에 의해 진행되어져 오고 있다. Goldstein et al.(1978)은 2개 모집단들의 상대적 크기를 고려한 직관적 분류법칙(intuitive classification rule ; ICR)을 제안하였고, Johnson et al.(1992)은 모집단들의 상대적 크기가 잘 알려져 있지 않은 경우 $\hat{q}_1=\hat{q}_2=1/2$로 놓아 $C=0$가 되는 값을 보편적 절단점으로 사용한 바 있다. 김혜중(1995)은 다변량 정규분포 하에서 π_1으로 부터의 관측값이 π_2로 부터의 관측값으로 오분류된 확률값이 같은 경우, 균형조건(balanced condition)을 이용한 균형분류법칙(balanced classification rule)을 사용하여 q_i, $i=1, 2$들을 결정한 바 있다. 그러나, 이제까지의 연구들은 모두 모집단의 분포가 다변량 정규분포라는 가정하에서 이루어진 것으로서, 모집단의 분포가 성장곡선모형을 할 경우에서의 연구는 아직 이루어지지 않고 있다. 따라서, 본 논문에서는 모집단의 분포가 성장곡선모형을 할 경우 이차분류법칙에서 오분류의 총

확률을 최소화하는 사전확률값 q_i, $i=1, 2$를 균형분류법칙에 따라 계산하여 새로운 절단점 C를 찾아 보았다.

2. 성장곡선모형

성장곡선모형은 Potthoff(1964)등에 의해 처음 제안된 후 Rao(1965, 1966), Geisser (1970) 및 Lee(1975, 1982)등에 의해 연구 되어져온 다변량 분산분석 모형이며, 다음과 같은 형태를 가졌다.

$$\underset{p\times N}{Y} = \underset{p\times m}{X}\ \underset{m\times r}{\tau}\ \underset{r\times N}{A} + \underset{p\times N}{\varepsilon} \tag{2.1}$$

여기서, τ는 미지모수의 행렬이며, 계획행렬 X와 상수행렬 A는 각각의 계수(rank)가 $m<p$, $r<N$ 임을 전제로 한다. 일반적으로 성장곡선은 시간의 변화에 따른 다항식으로 나타나기 때문에, 계획행렬 X와 상수행렬 A는 각각 다음과 같이 표시된다.

$$X = \begin{bmatrix} 1 & t_1 & t_1^2 & \cdots & t_1^{m-1} \\ 1 & t_2 & t_2^2 & \cdots & t_2^{m-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & t_p & t_p^2 & \cdots & t_p^{m-1} \end{bmatrix} \tag{2.2}$$

$$A^{T}_{r\times N} = \begin{bmatrix} E_1 & O_2 & O_3 & \cdots & O_r \\ O_1 & E_2 & O_3 & \cdots & O_r \\ O_1 & O_2 & E_3 & \cdots & O_r \\ \vdots & & \vdots & & \vdots \\ O_1 & O_2 & O_3 & \cdots & E_r \end{bmatrix} \tag{2.3}$$

여기서, E_i, $i=1, 2, \cdots, r$는 $N_i\times 1$ 단위벡터이고 O_i, $i=1, 2, \cdots, r$는 크기 N_i인 영벡터이며, $N=\sum_{i=1}^{r} N_i$.

이 때, 오차항 행렬 ε의 각 열이 서로 독립이고 평균 0, 공분산 행렬이 Σ인 p차원 다변량 정규분포를 가정하면 다음의 관계가 성립한다.

$$G(Y\,|\,\tau, \Sigma) = N(Y;\ X\tau A, \Sigma\otimes I_N) \tag{2.4}$$

단, $G(\cdot)$는 누적분포함수이고, $\otimes$ 는 Kronecker 곱이다.

Lee(1975)는 성장곡선을 베이지안의 관점에서 처음 연구하였는데, K개의 모집단으로 발생되는 서로 독립인 K개의 성장곡선은

$$G(Y_k \mid \tau_k, \sum_k, \pi_k) = N(Y_k; X\tau_k A_k, \sum_k \otimes I_{N_k}) \tag{2.5}$$

인 분포를 따른다고 가정할 때, 미래 관측치행렬 $V_{p\times q}$의 분포는 각 성장곡선 하에서 아래와 같다.

$$G(V \mid \tau_k, \sum_k, \pi_k) = N(V; X\tau_k F_k, \sum_k \otimes I_q),\ k = 1, 2, \cdots, K \tag{2.6}$$

여기서, F_k는 A_k의 첫번째 q개 열들로 구성된 $r\times q$ 행렬.

위와 같은 분포를 가정하여서 Lee는 모수 τ의 사후확률분포 및 베이즈 추정량을 도출하였다. 또한, 임의의 양정치 행렬 Σ에 대해, Rao(1966)는 모수 τ의 점근적 최우 추정량(asymptotic maximum likelihood estimator)을 다음과 같은 형태로 제시하였다.

$$\hat{\tau} = (X^T S^{-1} X)^{-1} X^T S^{-1} Y A^T (AA^T)^{-1} \tag{2.7}$$

여기서,

$$S = Y(I - A^T(AA^T)^{-1}A)Y^T$$

한편, Lee(1982)는 총오분류확률을 기준으로 성장곡선모형의 판별을 위해 최적분류법칙을 식 (2.6)으로부터 유도하였다.

정리 2-1 : $K = 2$이고, q_1과 q_2가 각각 성장모형의 사전확률일 때, 총오분류 확률 기준에 의한 최적분류법칙은 아래 부등식이 성립하면 $p\times q$ 반응행렬 V를 (성장모형 1)에 분류한다.

$$q\{\ln|\sum_2| - \ln|\sum_1|\} + tr(V - X\tau_2 F_2)^T \sum_2^{-1} (V - X\tau_2 F_2)$$
$$- tr(V - X\tau_1 F_1)^T \sum_1^{-1} (V - X\tau_1 F_1) \geq 2\ln\left(\frac{q_2}{q_1}\right) \tag{2.8}$$

〈증명〉 Lee(1982) 참조.

모수 τ_k와 $\sum_k$가 미지일 경우 이들의 추정값으로 $\hat{\tau}_k$와 $\hat{\sum}_k$를 사용하여 아래와 같이 추정한다.

$$q\{\ln|\hat{\sum}_2| - \ln|\hat{\sum}_1|\} + tr(V - X\hat{\tau}_2 F_2)^T \sum_2^{-1} (V - X\hat{\tau}_2 F_2)$$
$$- tr(V - X\hat{\tau}_1 F_1)^T \sum_1^{-1} (V - X\hat{\tau}_1 F_1) \geq 2\ln\left(\frac{\hat{q}_2}{\hat{q}_1}\right) \tag{2.9}$$

이 때, $\hat{\tau}_k$와 $\hat{\Sigma}_k$는 다음과 같이 정의 된다.

$$\begin{aligned}\hat{\tau}_k &= (X^T S_k^{-1} X)^{-1} X^T S_k^{-1} Y_k A_k^T (A_k A_k^T)^{-1} \\ \hat{\Sigma}_k &= N_k^{-1} (Y_k - X\hat{\tau}_k A_k)(Y_k - X\hat{\tau}_k A_k)^T \\ \text{여기서, } S_k &= Y_k (I - A_k^T (A_k A_k^T)^{-1} A_k) Y_k^T\end{aligned} \tag{2.10}$$

따름 정리 2-1 : 만약 $\Sigma_1 = \Sigma_2 = \Sigma$이고 $q=1$인 경우, 정리 (2-1)의 분류법칙은 다음과 같이 정의된다.

$$V^T \Sigma^{-1} X(\tau_1 F_1 - \tau_2 F_2) - \frac{1}{2}(\tau_1 F_1 + \tau_2 F_2)^T X^T \Sigma^{-1} X(\tau_1 F_1 - \tau_2 F_2) \geq \ln\left(\frac{q_2}{q_1}\right) \tag{2.11}$$

〈증명〉

식 (2.8)에서, $\Sigma_1 = \Sigma_2 = \Sigma$와 $q=1$을 대입하면 식 (2.11)을 얻는다.

모수가 未知일 경우 그들의 추정량을 사용하여 분류법칙 식 (2.11)을 아래와 같이 추정한다.

$$V^T \hat{\Sigma}^{-1} X(\hat{\tau}_1 F - \hat{\tau}_2 F_2) - \frac{1}{2}(\hat{\tau}_1 F_1 + \hat{\tau}_2 F_2)^T X^T \hat{\Sigma}^{-1} X(\hat{\tau}_1 F_1 - \hat{\tau}_2 F_2) \geq \ln\left(\frac{\hat{q}_2}{\hat{q}_1}\right) \tag{2.12}$$

이 때, τ_k및 Σ의 추정량은 아래와 같다.

$$\begin{aligned}\hat{\tau}_k &= (X^T S_k^{-1} X)^{-1} X^T S_k^{-1} Y_k A_k^T (A_k A_k^T)^{-1} \\ \hat{\Sigma} &= (N_1 + N_2)^{-1} (N_1 \hat{\Sigma}_1 + N_2 \hat{\Sigma}_2)\end{aligned} \tag{2.13}$$

3. 균형분류법칙

연속형 확률밀도 $f_i(v)$를 가진 2개의 다변량 모집단 π_i, $i=1, 2$가 있다고 하자. 여기서, v는 어떤 모집단 π_i로 부터 나온 p차원 벡터 관측치이고, 미지의 사전확률 $q_i \neq 0$가 있으며, $v \in \pi_i$이다. 또한, $U\{f_i(v), v\}$는 관측치 v의 선택에 따라 밀도함수 $f_i(v)$를 택하는데 관계있는 효용(utility)을 나타내는 실수값을 갖는 함수라 하자. 이 때, 이 함수의 기대효용은

아래와 같이 정의할 수 있다.

$$EU\{f_i(v), v\} = \int U\{f_i(v), v\} f(v) dv \tag{3.1}$$

여기서, $f(v) = \sum_{i=1}^{2} q_i f_i(v)$ 이다.

Buehler(1971)과 Good(1971)은 적절한 효용함수의 예를 몇가지 언급한 바가 있다. 그러나, Bernardo(1979)는 확률밀도함수 $f_i(v)$의 적절한 활용을 위해 대수의 개념을 도입하여 아래와 같은 형태의 효용함수를 제안하였다.

정리 3-1 : 만약 효용함수 U가 단조함수라면, 어떤상수 A와 함수 B에 대하여 아래 식이 성립한다.

$$U\{f_i(v), v\} = A \log f_i(v) + B(v), \; i = 1, 2 \tag{3.2}$$

〈증명〉 Bernardo(1979) 참조.

효용함수 (3.2)를 사용하여 아래 식이 성립한다면, 관측치 v를 분류하는데 있어 $f_j(v)$의 기대효용이 $f_i(v)$의 기대효용보다 효용이 더 크다고 할 수 있다.

$$EU\{f_j(v), v\} - EU\{f_i(v), v\} = \int \log \frac{f_j(v)}{f_i(v)} \cdot f(v) dv > 0 \tag{3.3}$$

분류실험을 실시할 때, 모집단 내의 개체들이 동일한 기대효용을 갖도록하는 모집단 분포들이 필요한데, v가 모집단 내의 개체들과 유사하다는 사실을 근거로 π_i, $i = 1, 2$에 분류하는 것이다.

정의 3-1 : 만약 모집단 π_i, $i = 1, 2$에 의해 값을 갖는 확률밀도 $f_i(v)$의 기대효용이 동일한 값을 갖는다면, 2 그룹 분류실험에 대해 균형(balance)이라 한다. 즉,

$$EU\{f_1(v), v\} - EU\{f_2(v), v\} = 0$$

혹은,

$$E \log \{ \frac{f_1(v)}{f_2(v)} \} = 0 \tag{3.4}$$

이다.

분류실험에 대한 균형조건 하에서 다음과 같은 최적분류법칙이 성립한다.

정리 3-2 : $v_{p\times 1}$는 사전확률이 q_i, $\sum_{i=1}^{2} q_i = 1$인 모집단 π_i들 중 하나로 부터 얻은 관측치이고, 오분류 비용이 C_{ij}, $i, j = 1, 2, i \neq j$ 라 하자. 균형계획

$$\int \log \frac{f_1(v)}{f_2(v)} \sum_{i=1}^{2} q_i f_i(v) dv = 0 \tag{3.5}$$

하에서, 최소위험 결정법칙은

$$q_2 C_{12} f_2(v) \geq q_1 C_{21} f_1(v) \tag{3.6}$$

을 만족하는 경우 v를 π_1에 할당하는 것이다. 이 때, $f_i(v)$는 모집단 π_i의 확률밀도 함수이다.

〈증명〉 김혜중(1995) 참조.

오분류 비용 C_{ij}가 모두 같다면 균형계획 식 (3.5) 하에서 최소위험 결정법칙은

$$q_2 f_2(v) \leq q_1 f_1(v) \tag{3.7}$$

을 만족하는 경우 v를 π_1에 할당하는 것이다.

만약 식(3.7)이 균형조건을 만족하지 않는다면, 최적확률분류법칙(optimal probabilistic classification rule)이라 한다.

균형계획조건인 식 (3.5)에서 q_i들에 대해 아래와 같은 2개의 선형독립 방정식들을 정의할 수 있다.

$$E[\log f_1(v) - \log f_2(v)] = 0$$

$$\sum_{i=1}^{2} q_i = 1$$

위의 식들이 q_i들에 대해 선형이므로, q_i들에 대한 유일해(unique solution)를 갖는다.

따름 정리 3-1 : 2 그룹 분류실험에서 균형조건을 이용하여 q_i, $0 < q_i < 1$, $i = 1, 2$를 구할 수 있다.

〈증명〉

2개 모집단 경우에 대해, 식 (3.7)은 아래와 같이 쓸 수 있다.

$$q_1 \int \log \frac{f_1(v)}{f_2(v)} \cdot f_1(v)dv - q_2 \int \log \frac{f_2(v)}{f_1(v)} \cdot f_2(v)dv = 0$$

$$q_1 + q_2 = 1$$

위의 관계를 이용하면,

$$q_1 = \frac{\int \log \frac{f_2(v)}{f_1(v)} \cdot f_2(v)dv}{\int \log \frac{f_1(v)}{f_2(v)} \cdot f_1(v)dv + \int \log \frac{f_2(v)}{f_1(v)} \cdot f_2(v)dv} \tag{3.8}$$

이고, $q_2 = 1 - q_1$이다. 여기서, q_i, $0 < q_i < 1$, $i = 1, 2$.

4. 균형이차분류법칙을 이용한 성장곡선모형의 판별

실제로 확률밀도함수(pdf) $f_i(v)$, $i = 1, 2$는 좀처럼 알기 힘들다. 본 장에서는 pdf들이 성장곡선 모형을 따른다고 하고, 오분류의 비용들은 모두 동일하다고 가정한다.

정리 4-1 : 관측치 V의 pdf가 $\prod_i$, $i = 1, 2$ 라고 가정하자.

여기서, $\prod_i = f(V \mid \Theta_i)$

$= N(X\tau_i F_i, \sum_i \otimes I_N), \sum_i > 0, i = 1, 2$

이다.

이 때, 균형이차분류법칙(balanced quadratic classification rule : BQC)은 아래 조건을 만족하는 경우 V를 π_1에 분류하는 것이다.

$$(V - X\tau_2 F_2)^T \sum_2^{-1} (V - X\tau_2 F_2) - (V - X\tau_1 F_1)^T \sum_1^{-1} (V - X\tau_1 F_1) - \log \frac{|\sum_1|}{|\sum_2|} \geq 2 \log \frac{q_2}{q_1} \tag{4.1}$$

이 때, q_1과 q_2는 방정식

$$\log \frac{|\sum_1|}{|\sum_2|} = \sum_{i=1}^{2} q_i \psi_{2i} \text{과} \sum_{i=1}^{2} q_i = 1 \tag{4.2}$$

의 해이다.

여기서, $\psi_{2i} = (\tau_i F_i - \tau_2 F_2)^T X^T \sum_2^{-1} X(\tau_i F_i - \tau_2 F_2)$

$$- (\tau_i F_i - \tau_1 F_1)^T X^T \sum_1^{-1} X(\tau_i F_i - \tau_1 F_1) - tr(\sum_i \sum_2^{-1}) - tr(\sum_i \sum_1^{-1}) \quad (4.3)$$

〈증명〉

오분류의 비용이 같다면 식 (3.6)은 아래와 같다.

$$q_2 f_2(v) \leq q_1 f_1(v) \quad (4.4)$$

성장곡선모형의 최소위험 결정법칙은 균형계획

$$\int \log \frac{f_1(v|\Theta)}{f_2(v|\Theta_2)} \sum_{i=1}^{2} q_i f_i(v|\Theta_i) dv = 0 \quad (4.5)$$

하에서 아래와 같이 쓸 수 있다.

$$q_2 f_2(v|\Theta_1) \leq q_1 f_1(v|\Theta_2) \quad (4.6)$$

성장곡선 모형의 확률밀도함수를 윗 식에 대입함으로서 결과를 얻을 수 있다.

정리 4-1에서 쓴 첨자들은 v를 π_i, $i=1, 2$에 분류하는데 대한 순번을 나타낸다. 따라서, 균형이차분류법칙에서 식 (4.2)의 q_i, $i=1, 2$는 아래 정리에 따라 구할 수 있다.

정리 4-2 : 모수를 알 수 있는 균형 2 그룹 정규분류 하에서, 사전확률 q_i는 아래 식의 해이다.

$$Q = \Psi^{-1}\Phi \quad (4.7)$$

여기서, $\Phi = (1, \phi_2)^T$

$Q = (q_1, q_2)^T$

$\phi_2 = \log|\sum| - \log|\sum_2|$

이고, 2×2 행렬 $\Psi \equiv \{\psi_{ki}\}$ $i, k=1, 2$는 행렬요소로 $\psi_{11} = \psi_{12} = 1$을 가지며, ψ_{2i}는 $i=1, 2$일 때 식 (4.2)에서 정의한 것과 같다.

〈증명〉 김혜중(1995) 참조.

식 (4.7)의 해가 부등식의 제한조건 $0 < q_i < 1$, $i=1, 2$을 항상 만족하므로 아래 정리를 얻는다.

정리 4-3 : 두개의 성장곡선모형이 아래의 분포를 따른다고 하자.

$$G(V|\tau_i, \sum_i, \prod_i) \sim N(V|X\tau_i F_i, \sum_i \otimes I_k)$$

균형이차분류법칙은 아래 조건을 만족하는 경우 V를 π_i에 분류하는 것이다.

$$(V-X\tau_2 F_2)^T\sum_2^{-1}(V-X\tau_2 F_2)-(V-X\tau_1 F_1)^T\sum_1^{-1}(V-X\tau_1 F_1) \\ -\log\frac{|\sum_1|}{|\sum_2|} \geq 2\log\frac{q_2}{q_1} \tag{4.8}$$

여기서,

$$q_1 = \frac{\log\frac{|\sum_2|}{|\sum_1|}+p-(\tau_1 F_1-\tau_2 F_2)^T X^T\sum_1^{-1}X(\tau_1 F_1-\tau_2 F_2)-tr(\sum_2\sum_1^{-1})}{2p-(\tau_1 F_1-\tau_2 F_2)^T X^T(\sum_1^{-1}+\sum_2^{-1})X(\tau_1 F_1-\tau_2 F_2)-tr(\sum_1\sum_2^{-1})-tr(\sum_2\sum_1^{-1})} \tag{4.9}$$

이다.

〈증명〉

정리 4-1과 정리 4-2를 이용하여 증명할 수 있다.

만약, $\sum_1=\sum_2=\sum$이면 식 (4.8)은 아래와 같이 된다.

$$V^T\sum^{-1}(\tau_1 F_1-\tau_2 F_2)-\frac{1}{2}(\tau_1 F_1+\tau_2 F_2)^T X^T\sum^{-1}X(\tau_1 F_1-\tau_2 F_2) \\ \geq \log\frac{q_2}{q_1} \tag{4.10}$$

여기서,

$$q_1 = \frac{p-(\tau_1 F_1-\tau_2 F_2)^T X^T\sum^{-1}X(\tau_1 F_1-\tau_2 F_2)-1}{2p-2(\tau_1 F_1-\tau_2 F_2)^T X^T\sum^{-1}X(\tau_1 F_1-\tau_2 F_2)-2} \tag{4.11}$$

이 되어, $q_1=q_2=\frac{1}{2}$ 이다.

그러나, 모수 τ_1, τ_2 및 Σ_1, Σ_2가 미지인 경우 그들의 최소제곱추정량 식 (2.10)을 이용할 수 있다.

정리 4-4 : 식 (4.8)과 (4.9)에 대해 표본으로부터 추정하여 사용한 균형이차분류법칙에 따라 아래 조건을 만족하는 경우 V를 π_i에 분류한다.

$$(V-X\hat{\tau}_2F_2)^T\hat{\Sigma}_2^{-1}(V-X\hat{\tau}_2F_2)-(V-X\hat{\tau}_1F_1)^T\hat{\Sigma}_2^{-1}(V-X\hat{\tau}_1F_1)$$
$$-\log\frac{|\hat{\Sigma}_1|}{|\hat{\Sigma}_2|}\geq 2\log\frac{\hat{q}_2}{\hat{q}_1} \tag{4.12}$$

여기서,

$$\hat{q}_1=\frac{\log\frac{|\hat{\Sigma}_2|}{|\hat{\Sigma}_1|}+p-(\hat{\tau}_1F_1-\hat{\tau}_2F_2)^TX^T\hat{\Sigma}_1^{-1}X(\hat{\tau}_1F_1-\hat{\tau}_2F_2)-tr(\hat{\Sigma}_2\hat{\Sigma}_1^{-1})}{2p-(\hat{\tau}_1F_1-\hat{\tau}_2F_2)^TX^T(\hat{\Sigma}_1^{-1}+\hat{\Sigma}_2^{-1})X(\hat{\tau}_1F_1-\hat{\tau}_2F_2)-tr(\hat{\Sigma}_1\hat{\Sigma}_2^{-1})-tr(\hat{\Sigma}_2\hat{\Sigma}_1^{-1})} \tag{4.13}$$

이고,

$$\hat{q}_2=1-\hat{q}_1$$

이며, 성장곡선모형의 판별분석에서 분류법칙에 대한 절단점은 $C=2\log(\hat{q}_2/\hat{q}_1)$ 이다.

〈증명〉

정리 4-3에서 모수 τ_1, τ_2 및 Σ_1, Σ_2 대신 그들의 최소제곱추정량을 사용함으로서 도출할 수 있다.

5. 모의실험

성장곡선모형에서 사전확률 식 (4.13)을 사용한 균형이차분류법칙(BQC)과 q_i의 직관적 추정치 $\hat{q}_i=\frac{N_i}{N_1+N_2}$, $i=1$, 2를 사용한 직관적이차분류법칙(intuitive quadratic classification rule : IQC)을 비교하기 위하여 Monte Carlo simulation을 실시하였다.

이를 위해, 공분산 행렬들이 서로 다른 2개의 성장모형 $Y_k\sim N(X\tau_kA_k, \Sigma_k\otimes I_N)$, $k=1$, 2을 가정하고, 각 성장모형의 모수들을 다음과 같이 설정하였다.

$\{X, \tau_1, \tau_2, A_1, A_2, \Sigma_1, \Sigma_2, p, N_1, N_2\}$

2개의 서로 다른 공분산 행렬을 생성하기 위해 $\Sigma_1 = H I_p H^T$ 및 $\Sigma_2 = H D_p H^T$가 되는 정칙행렬을 사용하였다. 즉, Σ_1은 항등행렬을 이용한 정칙행렬이고, Σ_2는 d_i, $i=1, 2, \cdots, p$를 대각요소로 갖는 대각행렬을 이용한 정칙행렬이다. 이 때, 모의실험을 위해 다음의 상황을 설정하였다.

〈 표 5-1 〉 모의실험상황

p	I_p	(N_1, N_2)
2	I_2	(10, 15) (15, 15) (15, 20)
3	I_3	(10, 15) (15, 15) (15, 20)

Case 1 : $D = \text{diag}\left(d - \frac{4d+1}{2d+1} \quad d - \frac{4d+1}{2d+1}\right)$

Case 2 : $D = \text{diag}(d \quad d+1)$

Case 3 : $D = \text{diag}(2d+3 \quad 2d+3)$

Case 4 : $D = \text{diag}\left(d - \frac{4d+1}{2d+1} \quad d - \frac{4d+1}{2d+1} \quad d - \frac{4d+1}{2d+1}\right)$

Case 5 : $D = \text{diag}(d \quad d+1 \quad d+2)$

Case 6 : $D = \text{diag}(2d+3 \quad 2d+3 \quad 2d+3)$

I_p : p차원 항등행렬.

X : 식 (2.2)에서 정의한 계획행렬.

A : 식 (2.3)에서 정의한 상수행렬.

또한, 모수 τ_i, $i=1, 2$는 행렬 T를 사용하여 $T\tau_1 T^T = I_p$ 및 $T\tau_2 T^T = D_p$가 되는 정칙선형변환행렬을 사용하였다.

제안된 분류법칙의 우수성을 판단하는 기준으로서 판별에 대한 오분류 오차비를 계산하였다. 이 때, 오분류 오차비를 계산하기 위해, $N(X\tau_k A_k, \Sigma_k \otimes I_N)$, $k=1, 2$로 부터 한 쌍의 표본을 생성한 후 분류법칙 BQC와 IQC에 따라 분류하였다.

이 때, τ_1, τ_2, Σ_1, Σ_2, p, N_1 및 N_2 값의 각 집합으로 부터 표본의 쌍들에 대해 각각 1000번의 실험을 수행하였다. 모의실험을 위한 프로그램은 SAS/IML을 사용하여 작성하였다.

1000회 모의실험에 대한 오분류 오차비를 보면 BQC에 근거한 결과는 대체로 잘 수행된 것 같다 〈표 5-2 참조〉. BQC의 오차비는 모든 경우에서 IQC의 오차비에 비해 낮았다. 이것은 모집단의 크기에 대한 비율을 사전확률로 단순히 사용하는 것보다, 이를 추정하기 위해 균형조건 식 (4.2)를 사용하여 얻은 결과가 우수함을 의미한다. 기대한 바와 같이 동일한 자료의 크기에 대해, 차수 p가 낮은 경우 오분류 오차비가 상대적으로 작았으며, 동일 차수 하에서도 자료의 크기가 클수록 오분류 오차비가 작았다. 이러한 경향은

자료의 크기가 증가할수록 뚜렷하리라 생각된다. 그리고, BQC의 수행은 실험표본이 동일한가의 여부에 관계없이 IQC의 결과보다 좋았으며, 실험표본의 비율을 역으로하여 유도된 동일한 표본실험도 유사한 결과를 나타내었다. 그러나, 공분산행렬에 따라 모양이 달라지는 성장곡선모형의 특성상 대각요소인 d_i, $i=1, 2, \cdots, p$값에 따라 판별력은 큰 차이를 보인다는 것도 알 수 있다.

〈 표 5-2 〉 1000회 simulation에 대한 오분류 오차비(d=2인 경우)

p	(N_1, N_2)	Case	BQC	IQC
2	(10, 15)	1	0.145	0.179
		2	0.184	0.196
		3	0.121	0.147
	(15, 15)	1	0.136	0.164
		2	0.173	0.189
		3	0.116	0.133
	(15, 20)	1	0.104	0.120
		2	0.152	0.171
		3	0.096	0.107
3	(10, 15)	4	0.149	0.173
		5	0.176	0.189
		6	0.127	0.151
	(15, 15)	4	0.141	0.169
		5	0.165	0.181
		6	0.123	0.158
	(15, 20)	4	0.121	0.136
		5	0.134	0.176
		6	0.109	0.120

6. 결론

성장곡선모형에 균형법칙을 적용하여 사전확률값 식 (4.13)을 추정하여 절단점 C를 계산하였다. 그 결과 두개 모집단의 크기가 다른 경우로 부터의 실험표본일지라도 균형이차분류법칙(BQC)을 적용하는 것이 오분류오차를 감소시키는 효과가 있음을 알았다. 물론 어떤 특수한 상황에 직면할 경우 결과가 달라질 수도 있겠으나, 실험표본의 크기가 같거나 약간 다른 경우 BQC는 IQC보다 더 우수하다고 할 수 있겠다. 그룹의 수가 K개로 확장하였을 경우에도 식 (3.5)와 (3.6)을 K개의 그룹으로 일반화 함으로서 해결할 수 있다. 따라서, 오분류비용이 동일한 경우 일반화된 균형이차분류법칙은 균형계획조건

$$\int \log \frac{f_i(v)}{f_j(v)} \sum_{k=1}^{K} q_k f_k(v)\, dv = 0, \quad \text{for all} \quad i \neq j,\ i, j = 1, 2, \cdots, K \tag{6.1}$$

하에서,

$$q_i f_i(v) \geq q_j f_j(v), \quad \text{for all} \quad i \neq j,\ i, j = 1, 2, \cdots, K \tag{6.2}$$

을 만족하는 경우 v를 π_i에 할당하는 것이다.

그러나, 균형조건으로부터 얻은 q_i들의 R^K개 집합은 부등식의 제한조건 $0 < q_i < 1$ $i = 1, 2, \cdots, K$을 만족하지 않을 수도 있으므로 균형계획조건을 일부 수정한 다중이차분류법칙(multiple quadratic classification rule)이 필요하다. 수정에 관계된 연구는 다음의 과제로 남겨두고자 한다.

참고문헌

[1] Anderson, T. W. (1984), *An Introduction to Multivariate Statistical Methods*, 2nd ed. Wiley & Sons, New York.

[2] Bernardo, J M. (1979), "Expected information as expected utility," *The Annals of Statistics*, Vol. 7, No. 3, pp. 686−690.

[3] Box, G. E. P. and Tiao, G. C. (1973), *Bayesian Inference in Statistical Analysis*, Addison-Wesley Publishing Company, Massachusetts.

[4] Buehler, R. J. (1971), *Measuring information and uncertainty*, *In Foundations of Statistical Inference*, Ed. by Godambe and Sprott, Holt, Rinehart & Winston, Toronto.

[5] Geisser, S. (1964), "Posterior Odds for Multivariate Normal Classifications," *J. Roy. Statist. Soc. Ser. B*, Vol. 1, pp. 69−76.

[6] Geisser, S. and Eddy, W. F. (1979), "A Predictive approach to model selection," *Journal of American Statistical Association*, Vol. 74, pp. 153−160.

[7] Geisser, S. (1980), "Growth Curve Analysis". *Handbook of Statistics*, Vol. 1, pp. 88−115.

[8] Gilbert, E. S. (1969), "The effect of unequal variance-covariance materices on Fisher's linear discriminant function," *Biometrics*, Vol. 35, pp. 505−514.

[9] Goldstein, M. and Dillon, W. R. (1978), *Discrete Discriminant Analysis*, Wiley & Sons, New York.

[10] Good, I. J. (1971), In discussion of R. J. Buehler (1971).

[11] Johnson, R. A. and Wichern, D. W. (1992), *Applied Multivariate Statistical Analysis*, 3rd ed., Prentice Hall, New Jersey.

[12] Lee, L. C. and Geisser, S. (1975), "Applications of growth curve predication," *Sankhya*, Vol. 37, pp. 239−256.

[13] Lee, L. C. (1982), "Classification of Growth Curves," *Handbook of Statistics*, pp. 121−137.

[14] Kim, H. J. (1995), "On a Balanced Quadratic Classification Rule," *Communication Statistics*, Vol. 24, pp. 607−623.

[15] Press, S. J. (1982), *Applied Multivariate Analysis: Using Bayesian and Frequentist Method of Inference*, Robert E. Krieger, Florida.

[16] Potthoff, R. R and Roy, S. N. (1964), "A generalized multivariate analysis of variance model useful especially for growth curve problems," *Biometrika*, Vol. 51, pp. 313−326.

[17] Rao, C. R. (1965), "The theory of least squares when the parameters are stochastic and its application to the analysis of growth curve," *Biometrika*, Vol. 52, pp. 447−458.

[18] Rao, C. R. (1966), "Covariance adjustment and related problems in multivariate analysis," *Multivariate Analysis* Ⅱ, Academic Press, Krishnaiah, P.R., edition, New York, pp. 87−103.