The Korean Communications
in Statistics Vol. 9, No. 1, 2002
pp. 87-100

A Study on the Poorly-posed Problems in the Discriminant Analysis of Growth Curve Model[1)]

Kyu-Bark Shim[2)]

Abstract

Poorly-posed problems in the balanced discriminant analysis was considered. We restrict consideration to the case of observations and the number of variables are the same and small. When these problems exist, we do not use a maximum likelihood estimates(MLE) to estimate covariance matrices. Instead of MLE, an alternative estimate for the covariance matrices are proposed. This alternative method make good use of two regularization parameters, λ and γ. A new test rule for the discriminant function is suggested and examed via limited but informative simulation study.

From the simulation study, it is shown that the suggested test rule gives better test result than other previously suggested method in terms of error rate criterion.

Keywords : discriminant analysis, growth curve model, poorly-posed, regularization parameter

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 이 모형을 따른다고 할 수 있다. 성장곡선모형은 Potthoff(1964) 등에 의해 처음 제안된 후 Rao (1966), Geisser(1980) 및 Lee(1982)등에 의해 연구 되어져온 다변량 분산분석모형이며 다음과 같은 형태를 가졌다.

$$\underset{pxN}{Y} = \underset{pxm}{X}\ \underset{mxr}{\tau}\ \underset{rxN}{A} + \underset{pxN}{\varepsilon} \qquad \cdots (1-1)$$

여기서, τ는 미지모수의 행렬이고, ε은 오차항 행렬이며, 계획행렬 X와 상수행렬 A는 각각의 계수가 $m<p$, $r<N$ 임을 전제로 한다. 위 모형을 따르는 자료들을 분석하는 한 방법으로 판별분석(discriminant analysis)이 있는데, 이에 대해서는 Fisher(1936)의 연구 이후 활발하게 진행되어 오고 있다. 2그룹 판별분석이란 각 모집단의 특성을 나타내는 다변량 자료를 바탕으로 개체들을 2개의 모집단 가운데 하나의 모집

1) 본 연구는 2002학년도 동국대학교 전문학술지 논문 게재연구비 지원으로 이루어 졌음.
2) Associate Professor, Department of Information and Statistics, Dongguk University, 707 Suckjang-dong, Kyungju 780-714
E-mail : gpshim@dongguk.ac.kr

단으로 분류하는 분석방법이다. 판별분석에서 가장 일반적인 가정은 V가 관측치들의 p차원 벡터이고 모집단 π_i, $i=1, 2$ 로 부터의 표본이라면, 이것은 평균벡터 μ_i, 공분산 행렬 Σ_i를 가진 다변량 정규분포를 따른다는 것이다.

Kim(1995)은 성장곡선모형에 대한 2그룹 판별분석에서 관측치 V를 모집단 π_i, $i=1, 2$ 에 판별할 때 기준이 되는 절단점(cut off point)의 추정에서 균형이차분류법(balanced quadratic classification : BQC)을 적용한 바 있다. 그는 절단점 계산의 바탕이 되는 사전확률들의 계산에 BQC를 적용시킨 결과 사전확률을 표본의 크기에 비례하여 결정하는 기존의 직관적분류법(intuitive classification rule : ICR)에 비해 오분류오차가 감소됨을 보였다.

그러나, 위의 연구들은 표본의 차수에 비해 자료의 개수가 많은 경우에서 주로 이루어 졌으며, 만약 표본의 개수가 표본의 차수에 비해 작거나 같은 경우가 발생하면 분석의 결과는 큰 오차를 가질 수 밖에 없을 것이다. 이와 같이 표본의 개수가 표본의 차수에 비해 작은 경우를 판별분석에서 ill-posed의 문제라 하고 같은 경우를 poorly-posed의 문제라 하며, 판별분석에서 이들 문제가 발생하면 모수 추정치들의 유효성이 낮아져 올바른 결과를 기대할 수 없다. 이에 대한 연구로서 Titterington(1985)과 O'Sullivan(1986)은 이들 문제를 해결하는 방법으로 조정화 모수를 사용한 조정화 방법을 제안하였으며, Friedman(1989)은 소표본으로부터 모집단 공분산 행렬을 추정할 경우 발생할 수 있는 표본공분산 행렬의 비정칙 문제를 해결하는 방편으로 조정화 판별분석(regularized discriminant analysis : RDA)법을 제안하였다.

본 논문에서는 균형이차분류법을 적용한 성장곡선모형의 판별분석에서 발생할 수 있는 poorly-posed의 문제를 해결하는 방법으로 조정화 방법을 사용하여 보았다.

2. 성장곡선모형의 판별분석

일반적으로 성장곡선은 시간의 변화에 따른 다항식으로 나타나기 때문에 식(1-1)의 모형에서 계획행렬 X와 상수행렬 A는 각각 다음과 같이 표시된다.

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

$$A^T_{r\times N} = \begin{bmatrix} E_1 & 0_2 & 0_3 & \cdots & 0_r \\ 0_1 & E_2 & 0_3 & \cdots & 0_r \\ 0_1 & 0_2 & E_3 & \cdots & 0_r \\ \vdots & \vdots & \vdots & & \vdots \\ 0_1 & 0_2 & 0_3 & \cdots & E_r \end{bmatrix} \quad \cdots (2-2)$$

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 영벡

터이며, $N = \sum_{i=1}^{r} N_i$.

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

$$G(Y|\tau, \Sigma) = N(Y; X\tau A, \Sigma \otimes I_N) \quad \cdots (2-3)$$

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

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

$$G(Y_k|\tau_k, \Sigma_k, \pi_k) = N(Y_k; X\tau_k A_k, \Sigma_k \otimes I_{N_k}) \quad \cdots (2-4)$$

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

$$G(V|\tau_k, \Sigma_k, \pi_k) = N(V; X\tau_k F_k, \Sigma_k \otimes I_q), \; k = 1, 2, \cdots, K \quad \cdots (2-5)$$

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 최우추정량을 다음과 같은 형태로 제시하였다.

$$\hat{\tau} = (X^T S^{-1} X)^{-1} X^T S^{-1} Y A^T (A A^T)^{-1} \quad \cdots (2-6)$$

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)으로 부터 유도하였다.

정리 2-1 : K=2 이고, q_1 과 q_2가 각각 성장모형의 사전확률일 때, 총오분류 확률기준에 의한 최적분류법칙은 아래 부등식이 성립하면 $p\times q$반응행렬 V를 (성장모형 1)에 분류하는 것이다.

$$q\{\ln|\Sigma_2| - \ln|\Sigma_1|\} + tr(V - X\tau_2 F_2)^T \Sigma_2^{-1} (V - X\tau_2 F_2)$$

$$- tr(V - X\tau_1 F_1)^T \Sigma_1^{-1} (V - X\tau_1 F_1) \geq 2\ln\left(\frac{q_2}{q_1}\right) \quad \cdots (2-7)$$

(증명) Lee(1982) 참조.

이 때, 분류의 기준이 되는 값 $2\ln(q_2 / q_1)$을 판별분석에서는 절단점이라 하는데, 이 기준의 精度는 사전 확률값 q_1, q_2의 적절한 선택에 달려있다.

모수 τ_k 와 Σ_k가 미지일 경우 이들의 추정값으로 $\hat{\tau}_k$와 $\hat{\Sigma}_k$ 를 사용하여 아래와 같이 추정한다.

$$q\{\ln|\hat{\Sigma}_2| - \ln|\hat{\Sigma}_1|\} + tr(V - X\hat{\tau}_2F_2)^T\ \hat{\Sigma}_2^{-1}(V - X\hat{\tau}_2F_2)$$
$$- tr(V - X\hat{\tau}_1F_1)^T\ \hat{\Sigma}_1^{-1}(V - X\hat{\tau}_1F_1) \geq 2\ln(\frac{\hat{q}_2}{\hat{q}_1}) \quad \cdots (2-8)$$

이 때, q는 임의의 상수이며, $\hat{\tau}_k$와 $\hat{\Sigma}_k$ 는 다음과 같이 정의된다.

$$\hat{\tau}_k = (X^TS_k^{-1}X)^{-1}X^TS_k^{-1}Y_kA_k^T(A_kA_k^T)^{-1}$$
$$\hat{\Sigma}_k = N_k^{-1}(Y_k - X\hat{\tau}_kA_k)(Y_k - X\hat{\tau}_kA_k)^T \quad \cdots (2-9)$$

여기서, $S_k = Y_k(I - A_k^T(A_kA_k^T)^{-1}A_k)Y_k^T$

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

$$V^T\Sigma^{-1}X(\tau_1F_1 - \tau_2F_2) - \frac{1}{2}(\tau_1F_1 + \tau_2F_2)^TX^T\Sigma^{-1}X(\tau_1F_1 - \tau_2F_2)$$
$$\geq \ln(\frac{q_2}{q_1}) \quad \cdots (2-10)$$

(증명) 식 (2-8)에서, $\Sigma_1 = \Sigma_2 = \Sigma$ 와 $q = 1$을 대입하면 식 (2-10)을 얻는다.

모수가 未知일 경우 그들의 추정량을 사용하여서 분류법칙 식 (2-10)을 아래와 같이 추정한다.

$$V^T\hat{\Sigma}^{-1}X(\hat{\tau}_1F_1 - \hat{\tau}_2F_2) - \frac{1}{2}(\hat{\tau}_1F_1 + \hat{\tau}_2F_2)^TX^T\hat{\Sigma}^{-1}X(\hat{\tau}_1F_1 - \hat{\tau}_2F_2)$$
$$\geq \ln(\frac{\hat{q}_2}{\hat{q}_1}) \quad \cdots (2-11)$$

이 때, τ_k 및 Σ의 추정량은 아래와 같다.

$$\hat{\tau}_k = (X^T S_k^{-1} X)^{-1} X^T S_k^{-1} Y_k A_k^T (A_k A_k^T)^{-1}$$
$$\hat{\Sigma} = (N_1 + N_2)^{-1}(N_1 \hat{\Sigma}_1 + N_2 \hat{\Sigma}_2) \quad \cdots (2-12)$$

Bernardo(1979)는 모집단 분포의 기대효용의 개념을 사용하여 오분류의 총확률을 강제로 최소화 시키는 형태의 이차분류법칙을 제안하였다. 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. Kim(1995)은 Bernardo가 제안한 이차분류법칙에 균형의 개념을 도입하여 균형이차분류법에 의한 판별분석을 실시한 바 있다.

두개의 성장곡선모형이 아래의 분포를 따른다고 하자.

$$G(V|\tau_i, \Sigma_i, \pi_i) \sim N(V|X\tau_i F_i, \Sigma_i \otimes I_k), \quad i=1,2$$

균형이차분류법칙은 아래 조건을 만족하는 경우 V를 모집단 π_i에 분류하는 것이다.

$$(V - X\tau_2 F_2)^T \Sigma_2^{-1}(V - X\tau_2 F_2) - (V - X\tau_1 F_1)^T \Sigma_1^{-1}(V - X\tau_1 F_1)$$
$$- \log \frac{|\Sigma_1|}{|\Sigma_2|} \geq 2\log \frac{q_2}{q_1} \quad \cdots (2-13)$$

여기서,

$$q_1 = \frac{\log \frac{|\Sigma_2|}{|\Sigma_1|} + p - (\tau_1 F_1 - \tau_2 F_2)^T X^T \Sigma_1^{-1} X(\tau_1 F_1 - \tau_2 F_2) - tr(\Sigma_2 \Sigma_1^{-1})}{2p - (\tau_1 F_1 - \tau_2 F_2)^T X^T (\Sigma_1^{-1} + \Sigma_2^{-1}) X(\tau_1 F_1 - \tau_2 F_2) - tr(\Sigma_1 \Sigma_2^{-1}) - tr(\Sigma_2 \Sigma_1^{-1})}$$

$\cdots (2-14)$

이다.

만약, $\Sigma_1 = \Sigma_2 = \Sigma$이면 식 (2-13)은 아래와 같이 된다.

$$V^T \Sigma^{-1}(\tau_1 F_1 - \tau_2 F_2) - \frac{1}{2}(\tau_1 F_1 + \tau_2 F_2)^T X^T \Sigma_1^{-1} X(\tau_1 F_1 - \tau_2 F_2) \geq \log \frac{q_2}{q_1}$$

$\cdots (2-15)$

여기서,

$$q_1 = \frac{p-(\tau_1 F_1-\tau_2 F_2)^T X^T \Sigma^{-1} X(\tau_1 F_1-\tau_2 F_2)-1}{2p-2(\tau_1 F_1-\tau_2 F_2)^T X^T \Sigma^{-1} X(\tau_1 F_1-\tau_2 F_2)-2} \quad \cdots (2-16)$$

이 되어, $q_1 = q_2 = \frac{1}{2}$ 이다.

그러나, 모수 τ_1, τ_2 및 Σ_1, Σ_2 가 미지인 경우 그들의 최소제곱추정량 식 (2-12)를 이용하여 분류할 수 있다.

정리 2-2 : 식 (2-13)과 (2-14)에 대해 표본으로부터 추정하여 사용한 균형이차분류법칙에 따라 아래 조건을 만족하는 경우 V를 π_i에 분류한다.

$$(V-X\hat{\tau}_2 F_2)^T \hat{\Sigma}_2^{-1}(V-X\hat{\tau}_2 F_2)-(V-X\hat{\tau}_1 F_1)^T \hat{\Sigma}_1^{-1}(V-X\hat{\tau}_1 F_1) - \log\frac{|\hat{\Sigma}_1|}{|\hat{\Sigma}_2|} \geq 2\log\frac{\hat{q}_2}{\hat{q}_1} \quad \cdots (2-17)$$

여기서,

$$\hat{q}_1 =$$

$$\frac{\log\frac{|\hat{\Sigma}_2|}{|\hat{\Sigma}_1|}+p-(\hat{\tau}_1 F_1-\hat{\tau}_2 F_2)^T X^T \hat{\Sigma}_1^{-1} X(\hat{\tau}_1 F_1-\hat{\tau}_2 F_2)-tr(\hat{\Sigma}_2 \hat{\Sigma}_1^{-1})}{2p-(\hat{\tau}_1 F_1-\hat{\tau}_2 F_2)^T X^T(\hat{\Sigma}_1^{-1}+\hat{\Sigma}_2^{-1})X(\hat{\tau}_1 F_1-\hat{\tau}_2 F_2)-tr(\hat{\Sigma}_1 \hat{\Sigma}_2^{-1})-tr(\hat{\Sigma}_2 \hat{\Sigma}_1^{-1})} \quad \cdots (2-18)$$

이고,

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

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

3. 균형판별분석에서 조정화 판별법의 적용

Haff(1980) 및 Dey et al.(1985)는 여러가지의 손실함수를 이용하여 최소 위험 공분산 추정량을 고유근 추정에 연계해서 제시하였다. 그러나, 그들이 제안한 공분산 추정량은 표본공분산이 정칙행렬임을 전제로 하였다. 추정 모수의 차원이 관측치의 수와 유사하면 모수의 추정량들은 변동의 폭이 커져서 편의가 발생하고, 유효성이 감소되어서 비정칙 문제가 발생하게 될 수도 있다. 이 문제점을 해결하기 위해서 Cornfield(1967)

는 James-Stein 수축(shrinkage)을 이용하여 모집단 모수의 참값에 대한 편의의 강도를 조절하는 조정화 모수(regularization parameter)기법을 사용한 공분산 추정법을 제안하였다. 이 때, 사용되는 조정화모수는 편의 정도의 크고 작음에 따라 적절한 값이 부여된다. Friedman(1989)은 이 방법을 이용하여 판별분석에서의 공분산 추정에 사용하였고, 그 결과 오분류오차가 감소함을 보였다.
Friedman이 판별분석에 적용한 조정 추정법 중에서 가장 간단한 방법은 각 표본의 합동표본공분산행렬에 자료의 총 수로 나눈 형태를 들 수 있고, 그 추정량은 다음과 같이 정의된다.

$$\hat{R}_k = \frac{S}{n} \qquad \cdots \quad (3-1)$$

$$\text{여기서, } S = \sum_{k=1}^{K} S_k \ , \quad n : \text{자료의 총 수}$$

식 (3-1)에서 제시된 추정량은 공분산행렬의 추정값을 의도적으로 감소시켜 소표본의 경우 좋은 결과를 가져다 줄 때도 있으나, 이 현상이 항상 일어나지 않는다는 점에서 현재 널리 사용되지 않고 있다. 이와는 달리 식(3-1)에 조정화 모수 $\lambda\ (0 \le \lambda \le 1)$를 사용하여 추정한 공분산 행렬의 형태는 다음과 같다.

$$\hat{R}_k(\lambda) = \frac{S_k(\lambda)}{n_k(\lambda)} \qquad \cdots \quad (3-2)$$

$$\text{여기서,} \quad S_k(\lambda) = (1-\lambda)\, S_k + \lambda\, S \qquad \cdots \quad (3-3)$$

$$n_k(\lambda) = (1-\lambda)\, n_k + \lambda\, n \qquad \cdots \quad (3-4)$$

이고, n_k와 S_k는 각각 k번째 표본의 자료 수와 공분산 행렬이다.

식 (3-2,3,4)에서 조정화모수 λ는 합동표본공분산행렬을 모집단 공분산행렬의 추정량으로 사용함으로서 발생할 수 있는 각 그룹 공분산행렬 추정량들 사이의 비정칙성을 조절하는 역할을 한다. 이 때, 조정화의 의미는 λ를 이용해 합동표본공분산행렬의 비정칙성의 정도를 조절한다는 의미이고, 이러한 비정칙성의 정도에 따라 λ값이 변하는데, 이 값을 수축도(degree of shrinkage)라 한다. 특히, $\lambda = 1$일 경우 식 (3-2,3,4)로 부터 선형판별분석(LDA)의 결과와 일치하게 됨을 알 수 있으며, 이 경우 $\hat{R}_k(1) = \hat{R}(1)$으로 분석을 실시할 수 있다.
Titterington(1985)은 식 (3-2)에서 정의된 추정량에 적절한 조정화모수를 사용하면 합동표본공분산이 가진 많은 정도의 비정칙성을 제거할 수 있어 추론의 정도에 매우 큰 효과를 가져다 줄 수 있다고 주장하였다. 그러나, 이 정의된 추정량을 사용할 경우 비정칙 문제가 발생할 때나, 공분산 행렬들이 항등행렬의 배수가 되는 경우에는 심각한 편의문제를 초래할 수도 있음을 Friedman이 보였다. 위의 문제를 해결하기 위해 Friedman은 조정화선형판별의 경우 다음과 같은 새로운 공분산 추정량을 제안하였다.

$$\hat{R}(\lambda, \gamma) = (1-\gamma)\hat{R}(\lambda) + \frac{\gamma}{p} tr(\hat{R}(\lambda))I \quad \cdots (3-5)$$

위 추정량은 식(3-2)에서 사용된 λ 이외에 새로운 조정모수 $\gamma(0 \leq \gamma \leq 1)$를 도입하여서 비정칙의 문제를 해결하도록 한 것이다. 이 때, γ는 항등행렬의 곱에 대한 수축의 정도를 조정하는 역할을 지닌 조정화 모수가 된다.
이 추정량을 이용하여 적절한 γ 값으로 다변량 정규분포 가정 하에서 조정화 선형판별분석을 수행할 때 판별의 기준이 되는 k 번째 그룹의 판별점수는 아래와 같다.

$$d_k(X) = (X - \overline{X}_k)^T \hat{R}^{-1}(\lambda, \gamma)(X - \overline{X}_k) + \ln|\hat{R}(\lambda, \gamma)| - 2\ln\pi_k \quad \cdots (3-6)$$

여기서, π_k는 그룹 k에 할당될 비조건부 사전확률로서 $\pi_k = n_k / n$ 이다.
Friedman이 다변량 정규분포 하에서 모의실험을 실시해 본 결과 조정화 선형판별분석에서 사용된 조정화 모수 γ의 값은 오분류오차 기준에 의해 정하는 것이 바람직함을 보였다.

정리 3-1 : K=2 이고, q_k , k=1,2 가 각 성장모형의 사전확률일 때 $p \times q$ 표본행렬 V의 새로운 분류법칙의 분류영역은

$$R_1 : \frac{g_1(V)}{g_2(V)} \geq \frac{q_2}{q_1}$$

$$R_2 : \frac{g_1(V)}{g_2(V)} < \frac{q_2}{q_1} \quad \cdots (3-7)$$

이다.

여기서, $g_k(V) = (2\pi)^{-\frac{pq}{2}} |R_k(\lambda, \gamma)|^{\frac{q}{2}}$

$$\times \exp\left\{-\frac{1}{2} tr R_k^{-1}(\lambda, \gamma)(V - X\tau_k F_k)(V - X\tau_k F_k)^T\right\} \quad \cdots (3-8)$$

그러므로, 분류법칙은 V가 R_1에 포함되면 첫 번째 성장모형으로 분류하고, R_2에 포함될 경우는 두 번째 성장모형으로 분류한다.

(증명) 식 (2-7)에 Σ_k 대신 조정화 표본공분산행렬 $R_k(\lambda, \gamma)$로 대치시키면 분류법칙 식 (3-1)이 유도된다.

식 (3-1)의 양변에 자연대수를 취하면 아래와 같이 표현된다.

$$q\{\ln|R_2(\lambda,\gamma)| - \ln|R_1(\lambda,\gamma)|\} + tr(V - X\tau_2F_2)^T R_1(\lambda,\gamma)(V - X\tau_2F_2) - tr(V - X\tau_1F_1)^T R_1^{-1}(\lambda,\gamma)(V - X\tau_1F_1) \geq 2\ln(\frac{q_2}{q_1}) \quad \cdots (3-9)$$

두 성장모형에 대한 새로운 분류법칙은 식 (3-9)의 부등식이 성립되면 V를 (그룹 1)에 분류하는 것이다.

따름 정리 3-1 : 만약 $R_1(\lambda,\gamma) = R_2(\lambda,\gamma) = R(\lambda,\gamma)$ 이고 $q = 1$ 이면, 식 (3-9)는 다음과 같이 간단히 쓸 수 있다.

$$V^T R^{-1}(\lambda,\gamma)X(\tau_1F_1 - \tau_2F_2) - \frac{1}{2}(\tau_1F_1 + \tau_2F_2)^T X^T R^{-1}(\lambda,\gamma)X(\tau_1F_1 - \tau_2F_2) \geq \ln(\frac{q_2}{q_1}) \quad \cdots (3-10)$$

(증명) 식 (3-9)에서, $R_1(\lambda,\gamma) = R_2(\lambda,\gamma) = R(\lambda,\gamma)$ 와 $q = 1$ 을 대입하면 식 (3-10)을 얻을 수 있다.

위 분류법칙에 포함된 모수가 未知일 경우 그들의 추정량를 사용하여서 분류법칙을 추정한다. τ_k 과 $R(\lambda,\gamma)$ 의 추정량은 아래와 같다.

$$\hat{\tau}_k = (X^T S_k^{-1} X)^{-1} X^T S_k^{-1} Y_k A_k^T (A_k A_k^T)^{-1} \quad \cdots (3-11)$$

$$\hat{R}(\lambda,\gamma) = (N_1 + N_2)^{-1}(N_1\ \hat{R}_1(\lambda,\gamma) + N_2\ \hat{R}_2(\lambda,\gamma)) \quad \cdots (3-12)$$

정리 3-2 : 두개의 성장곡선모형이 아래의 분포를 따른다고 하자.

$$G(V|\tau_i, R_i(\lambda,\gamma), \pi_i) \sim N(V|X\tau_iF_i,\ R_i(\lambda,\gamma)\otimes I_{N_i}),\ i = 1, 2$$

조정화 균형이차분류법칙은 아래 조건을 만족하는 경우 V를 π_i에 분류하는 것이다.

$$(V-X\tau_2F_2)^T R_2(\lambda,\gamma)^{-1}(V-X\tau_2F_2)-(V-X\tau_1F_1)^T R_1(\lambda,\gamma)^{-1}(V-X\tau_1F_1)$$
$$-\log\frac{|R_1(\lambda,\gamma)|}{|R_2(\lambda,\gamma)|}\geq 2\log\frac{q_2}{q_1} \qquad \cdots (3-13)$$

여기서,

$$q_1 = \frac{M1}{M2} \qquad \cdots (3-14)$$

이고, $q_2 = 1 - q_1$ 이다. 단,

$$M1 = \log\frac{|R_2(\lambda,\gamma)|}{|R_1(\lambda,\gamma)|} + p-(\tau_1F_1-\tau_2F_2)^T X^T R_1(\lambda,\gamma)^{-1}X(\tau_1F_1-\tau_2F_2)$$
$$-tr(R_2(\lambda,\gamma)R_1(\lambda,\gamma)^{-1}) \qquad \cdots (3-15)$$

$$M2 = 2p-(\tau_1F_1-\tau_2F_2)^T X^T(R_1(\lambda,\gamma)^{-1}+R_2(\lambda,\gamma)^{-1})X(\tau_1F_1-\tau_2F_2)$$
$$-tr(R_1(\lambda,\gamma)R_2(\lambda,\gamma)^{-1})-tr(R_2(\lambda,\gamma)R_1(\lambda,\gamma)^{-1}) \qquad \cdots (3-16)$$

(증명) 식 (2-13)과 (2-14)에서 Σ_i , $i=1,2$ 대신 조정화 표본공분산행렬 $R_i(\lambda,\gamma)$, $i=1,2$ 로 대치시키 분류법칙 식 (3-13)이 유도된다.

정리 3-3 : 식 (3-13)과 (3-14)에 대해 표본으로부터 추정하여 사용한 조정화 균형이차분류법칙에 따라 아래 조건을 만족하는 경우 V를 π_i에 분류한다.

$$(V-X\hat{\tau}_2F_2)^T\ \widehat{R}_2(\lambda,\gamma)^{-1}(V-X\hat{\tau}_2F_2)$$
$$-(V-X\hat{\tau}_1F_1)^T\ \widehat{R}_1(\lambda,\gamma)^{-1}(V-X\hat{\tau}_1F_1)-\log\frac{|\widehat{R}_1(\lambda,\gamma)|}{|\widehat{R}_2(\lambda,\gamma)|}\geq 2\log\frac{\hat{q}_2}{\hat{q}_1}$$
$$\cdots (3-17)$$

여기서, $$\hat{q}_1 = \frac{\widehat{M1}}{\widehat{M2}} \qquad \cdots (3-18)$$

이고, $\hat{q}_2 = 1 - \hat{q}_1$ 이다. 단,

$$\widehat{M1} = \log \frac{|\widehat{R}_2(\lambda, \gamma)|}{|\widehat{R}_1(\lambda, \gamma)|} + p - (\hat{\tau}_1 F_1 - \hat{\tau}_2 F_2)^T X^T \widehat{R}_1(\lambda, \gamma)^{-1} X(\hat{\tau}_1 F_1 - \hat{\tau}_2 F_2) - tr(\widehat{R}_2(\lambda, \gamma)\ \widehat{R}_1(\lambda, \gamma)^{-1}) \quad \cdots (3-19)$$

$$\widehat{M2} = 2p - (\hat{\tau}_1 F_1 - \hat{\tau}_2 F_2)^T X^T (\widehat{R}_1(\lambda, \gamma)^{-1} + \widehat{R}_2(\lambda, \gamma)^{-1}) X(\hat{\tau}_1 F_1 - \hat{\tau}_2 F_2) - tr(\widehat{R}_1(\lambda, \gamma)\ \widehat{R}_2(\lambda, \gamma)^{-1}) - tr(\widehat{R}_2(\lambda, \gamma)\ \widehat{R}_1(\lambda, \gamma)^{-1}) \quad \cdots (3-20)$$

(증명) 정리 (3-2)에서 모수 τ_1 , τ_2및 $R_1(\lambda, \gamma)$, $R_2(\lambda, \gamma)$ 대신 그들의 최소제곱추정량을 사용함으로서 도출할 수 있다.

이 때, 균형법칙을 이용한 성장곡선모형의 조정화 판별분석에서 분류법칙에 대한 절단점은 $C = 2\log(\hat{q}_2 / \hat{q}_1)$ 이다.

4. 모의실험

성장곡선모형에서 본 논문에서 제안한 식(3-18)을 이용한 조정화 균형이차분류법칙(RBQC)와 사전확률 식(2-18)을 사용한 기존의 균형이차분류법칙(BQC)을 비교하기 위하여 모의실험을 실시하였다.
이를 위해, 공분산 행렬들이 서로 다른 2개의 성장모형 $Y_k \sim N(X\tau_i A_i\ ,\ R_i(\lambda, \gamma) \otimes I_{N_i})$, $i = 1, 2$ 를 가정하고, 각 성장모형의 모수들을 다음과 같이 설정하였다.

$$\{ X, \tau_1, \tau_2, A_1, A_2, R_1(\lambda, \gamma), R_2(\lambda, \gamma), p \}$$

2개의 서로 다른 공분산 행렬을 생성하기 위해 식(3-1)에서 통계량 S_1과 S_2의 모수 Σ_1 과 Σ_2 는 각각 $\Sigma_1 = H I_p H^T$ 및 $\Sigma_2 = H D_p H^T$ 가 되는 정칙행렬을 사용하였다. 즉, Σ_1 은 항등행렬을 이용한 정칙행렬이고, Σ_2 는 $d_i, i = 1, 2, \cdots, p$ 를 대각요소로 갖는 대각행렬을 이용한 정칙행렬이다. 그리고, poorly-posed의 문제가 야기되는 상황을 설정하기 위하여 $N_1 = N_2 = J$로 하였을 경우 $p = J$가 되게 하였다.

(표 4-1) 모의실험 상황

p	I_p	J
2	I_2	2
3	I_3	3
4	I_4	4

I_p : p차원 항등행렬 D_p : p차원 대각행렬

X : 식 (2-1)에서 정의한 계획행렬.

A : 식 (2-2)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.

그리고, RBQC와 BQC의 오분류오차를 비교하기 위하여 조정화 모수는 Friedman(1989)이 사용한 λ와 γ값을 아래와 같이 설정하였다.

(표 4-2) λ와 γ값의 설정

	Case 1	Case 2	Case 3	Case 4	Case 5
λ	0.35	0.125	0.354	0.50	0.65
γ	0.14	0.25	0.50	0.50	0.75

제안된 분류법칙의 우수성을 판단하는 기준으로서 판별에 대한 오분류오차비를 계산하였다. 이 때, 오분류오차비를 계산하기 위해, $N(X\tau_i A_i , R_i(\lambda , \gamma)\otimes I_{N_i})$, $i = 1, 2$ 로 부터 한 쌍의 표본을 생성한 후 분류법칙 RBQC와 BQC에 따라 그룹 1에 분류하는 경우에 대해 계산하였다.

이 때, τ_1 , τ_2 , Σ_1 , Σ_2 , p , J 값의 각 집합으로 부터 표본의 쌍들에 대해 각각 100번의 실험을 수행하였다. 모의실험을 위한 프로그램은 SAS/IML을 사용하여 작성하였다.

100회 모의실험에 대한 오분류 오차비를 보면 BQC에 근거한 결과는 대체로 잘 수행된 것 같다(표 4-3 참조). RBQC의 오차비는 모든 경우에서 BQC의 오차비에 비해 낮았다.

(표 4-3) 모의실험 결과

p = J	분석방법	Case 1	Case 2	Case 3	Case 4	Case 5
2	RBQC	0.36	0.32	0.31	0.28	0.25
	BQC	0.37	0.35	0.38	0.40	0.33
3	RBQC	0.09	0.07	0.10	0.09	0.07
	BQC	0.34	0.28	0.34	0.34	0.28
4	RBQC	0.02	0.03	0	0.01	0.01
	BQC	0.06(38)	0.05(24)	0.08(32)	0.11(32)	0.05(35)

(註) $p = J = 4$ 인 경우 BQC 결과에서 ( )의 수치는 실제로 분석이 이루어진 회수임.

차수가 $p = 4$이고 표본의 개수 $J = 4$인 경우 표본공분산행렬이 비정칙행렬이 되어 분석이 이루어지지 않았다. 이는 추정모수의 차원이 관측치의 수와 유사하면 모수 추정량들의 변동의 폭이 커져서 발생하는 현상으로 이문제의 해결 방안으로 Friedman(1989) 은 조정화 판별법을 제안한 바 있다. 성장곡선모형의 경우 Shim(1994)은 Friedman의 방법이 효과가 있음을 입증하였으며, 그는 이 모형의 특성상 분석이 되는 회수는 사용한 계획행렬에 따라 다소 차이가 있다고 하였다.

5. 결론

Shim(1995)은 성장곡선모형에 균형법칙을 적용하여 사전확률값 식(2-15)를 추정하여 절단점 C를 계산한 바있다. 그 결과 두개 모집단의 크기가 매우 다른 경우로 부터의 실험표본일지라도 균형이차분류법칙(BQC)을 적용하는 것이 오분류오차를 감소시키는 효과가 있음을 알았다. 그러나, 그의 연구는 표본의 크기가 10 ~15인 소표본일 때 이므로 표본의 크기와 차수가 같은 경우, 즉 poorly-posed의 문제가 발생하는 경우에도 BQC가 좋은 분류법이라 할 수는 없다. 본 논문에서는 poorly-posed의 문제가 발생하는 경우 Friedman(1989)이 사용한 조정화 모수를 이용하면 공분산행렬의 비정칙성을 제거함과 동시에 판별에 대한 오분류오차를 줄이는 효과가 있음을 보였다. 표(4-3)의 결과에서 알 수 있듯이 조정화 모수에 따라 오분류오차가 다소 달라 질 수도 있으나, 본 논문에서 제안한 RBQC 방법이 기존의 BQC 방법보다 대체적으로 우수한 판별력을 나타내고 있다. 다만, Friedman(1989)이 지적했듯 적절한 조정화 모수를 선택하는 일이 이 분석방법의 가장 큰 어려움이라 하겠다.

참고문헌

[1] Bernardo,J.M.(1979). Expected information as expected utility, The Annals of Statistics, Vol.7,No.3,686-690.

[2] Cornfield,J.(1967). Discriminant Function. Review of the Intenational Statistical Institute , 35, 142-153.

[3] Dey,D.K. and Srmivasan,C.(1985). Estimation of a Covariance Matrix Under Stein's Loss. The Annals of Statistics,13,pp.1581-1591.

[4] Fisher,R.A.(1936). The use of multiple measurements in taxonomic problems .Ann.Eugen., 7, 179-188.

[5] Friedman, J.H.(1989). Regularized Discriminant Analysis. Journal of American Statistical Associationn, 84, 165-175.

[6] Geisser,S. (1980). Growth Curve Analysis. Handbook of Statistics, Vol. 1, 88-115.

[7] Haff,L.R.(1980). Empirical Bayes Estimation of Multivariate Normal Covariance Matrix. The Annals of Statistics , 8, 586-597.

[8] Kim,H.J.(1995). On a Balanced Quadratic Classification Rule, Communication Statistics, Vol. 24.

[9] Lee,L.C,and Geisser,S.(1975), Applications of growth curve prediction Sankhya, 37, 239-256.

[10] Lee,L.C.(1982).Classification of Growth Curves, Handbook of Statistics, Vol. 121-137.

[11] O'Sullivan,F.(1986). A Statistical Prespective on Ill-Posed Inverse Problems. Statistical Science, 1, 502-517.

[12] Potthoff,R.R and Roy,S.N.(1964), A generalized multivariate analysis of Covariance model useful especially for growth curve problems ,Biometrika, 51, 313-326.

[13] Rao,C.R.(1966). The Theory of Least Squares When the Parameters are Stochastic and Its Application to the Analysis of Growth Curve, Biometrika, 52, 447-458.

[14] Shim, K,B.(1995). An Application of the Balanced Quadratic Classification Rule on the Discriminant Analysis in Growth Curve Model, Journal of the Korean Society for Quality Management.

[15] Titterington,D.M.(1985). Common Structure of Smmothing Tichniques in Statistics. International Statistical Review, 53, 141-170.

[2001년 9월 접수, 2002년 1월 채택]