응용통계연구 제 18권 3호, 2005년, pp. 655-660

추가정보검정법 적용의 개선

김명근[1)]

요 약

두 그룹 판별분석에서 자주 사용되는 Rao의 추가정보검정법의 적용에서 변수들의 모든 조합중에 일부분만을 고려하여 변수를 선택하는 방법을 제시하고, 예를 들어 제시된 방법의 효용성을 살펴본다.

주요용어: Rao의 추가정보검정법, 변수선택, 판별분석

1. 서론

판별분석에서의 변수선택에 대한 많은 연구가 있다. 자세한 방법은 McKay와 Campbell (1982a,b)의 연구조사논문에서 찾을 수 있으며, 이 이후에 제시된 방법은 McLachlan (1992)의 책에서 살펴 볼 수 있다. 제시된 변수선택방법 중에서 변수들의 모든 조합의 각각에 Rao(1973)의 추가정보검정(test for additional information)을 적용하여 변수선택을 하는 방법이 있으며, McKay (1978)는 이러한 변수선택을 위한 그래픽방법을 제시하였다. 그러나, 이 방법을 적용하기 위하여 변수들의 모든 조합을 고려해야 하므로, 변수의 수가 조금만 많아도 엄청난 계산을 필요로 할 수 있다.

본 연구에서는 두 그룹 판별분석에서 자주 사용되는 Rao의 추가정보검정법의 적용에서 변수들의 모든 조합중에 일부분만을 고려하여 변수를 선택하는 방법을 제시하고, 예를 들어 살펴본다.

2. 변수선택의 절차

같은 공분산행렬을 갖는 p-변량의 정규분포 $N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$와 $N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$에서 각각 크기가 n_1, n_2인 확률 표본을 추출하여, 각 표본에 입각한 $\boldsymbol{\mu}_i$와 $\boldsymbol{\Sigma}$의 최우추정량을 $\bar{\boldsymbol{x}}_i$ 와 $\boldsymbol{S}_i$ $(i = 1, 2)$로 표시한다. 그리고, $\boldsymbol{\Sigma}$의 합동표본공분산행렬을 $\boldsymbol{S} = (n_1\boldsymbol{S}_1 + n_2\boldsymbol{S}_2)/(n_1 + n_2 - 2)$로 나타낸다. 이 이후에 나오는 모수들의 추정량은 $\boldsymbol{\mu}_i$와 $\boldsymbol{\Sigma}$를 $\bar{\boldsymbol{x}}_i$와 $\boldsymbol{S}$로 대치하여 얻으며, $\hat{}$ 기호로 표시한다.

1) (361-742) 충북 청주시 모충동 231, 서원대학교 정보분석학과, 교수
E-mail: mgkim@seowon.ac.kr

2.1. Rao의 추가정보검정법

먼저 Rao의 추가정보검정법 (Rao, 1973, p. 568)을 살펴 본다. 두 그룹 $N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$와 $N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ 사이의 Mahalanobis 거리는

$$\Delta_p^2 = \boldsymbol{\delta}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\delta}$$

로 정의된다. 여기서, $\boldsymbol{\delta} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$이다. $\boldsymbol{\delta}$와 $\boldsymbol{\Sigma}$를 다음과 같이 분할한다:

$$\boldsymbol{\delta} = \begin{bmatrix} \boldsymbol{\delta}_1 \\ \boldsymbol{\delta}_2 \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}.$$

여기서 $\boldsymbol{\delta}_1$은 k 개의 원소를 가지는 것으로 가정하고, $\boldsymbol{\Sigma}$는 $\boldsymbol{\delta}$의 분할에 맞추어서 분할한다. 처음 k 개의 확률변수에 입각한 두 그룹사이의 Mahalanobis 거리는

$$\Delta_k^2 = \boldsymbol{\delta}_1^T \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\delta}_1$$

와 같이 표현할 수 있다. 두 그룹의 판별분석이 p 개의 확률변수 중에서 처음 k 개의 확률변수로 충분하다는 가설은

$$H_k : \Delta_p^2 = \Delta_k^2$$

로 나타낸다. 이 가설이 성립하면, 나머지 $p-k$ 개의 확률변수는 두 그룹의 판별분석에 필요한 추가정보를 주지 못하며, 따라서, 이 판별분석에 기여하는 바가 없다. 이러한 가설 H_k의 검정을 Rao의 추가정보검정이라 부르며, 필요한 검정통계량은

$$F = \frac{n_1 + n_2 - p - 1}{p - k} \left(\frac{\hat{\Delta}_p^2 - \hat{\Delta}_k^2}{m + \hat{\Delta}_k^2} \right)$$

로 주어진다. 여기서, $m = (n_1 + n_2)(n_1 + n_2 - 2)/n_1 n_2$이다. 좀더 상세한 내용은 Mardia et al. (1979)의 책에서 찾을 수 있다. H_k가 성립할 때, 검정통계량 F 는 분자, 분모의 자유도가 각각 $p-k$, $n_1 + n_2 - p - 1$인 F-분포를 따른다. F의 값이 유의하게 클 때 H_k를 기각한다.

최적의 변수선택을 하기 위해서 $k = 1$인 경우부터 $k = p - 1$인 경우를 고려해야 하므로, Rao의 검정을 $2^p - 2$번 실시해야 한다. 패턴인식과 같은 분야에서 사용되는 변수의 수는 매우 많다. 따라서, 이러한 분야에서 변수들의 모든 조합을 고려하여 변수선택을 하는 것은 거의 불가능하다.

2.2. 추가정보검정의 적용을 위한 개선된 절차

먼저 $k = 1$, 즉 Rao의 추가정보검정의 대상이 되는 첫 번째 변수를 선택하는 경우를 살펴 본다. $\hat{\boldsymbol{\delta}}$의 원소를 대각원소로 갖는 대각행렬을 $\boldsymbol{D}(\hat{\boldsymbol{\delta}})$로 하고, $p \times p$ 행렬 $\boldsymbol{C}$를

$$\boldsymbol{C} = \boldsymbol{D}(\hat{\boldsymbol{\delta}}) \hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{D}(\hat{\boldsymbol{\delta}}) \tag{2.1}$$

로 정의한다. p 차원 단위벡터 $\boldsymbol{v}$의 함수인 이차형태 $\boldsymbol{v}^T\boldsymbol{C}\boldsymbol{v}$의 최대값은 $\boldsymbol{C}$의 제일 큰 특성치 (λ_1이라 한다)로 주어지며, 이때 단위벡터 $\boldsymbol{v}$는 $\boldsymbol{C}$의 제일 큰 특성치에 해당하는 특성벡터가 된다 (Schott, 1997, p.105). 즉, 임의의 단위벡터 $\boldsymbol{v}$에 대하여

$$\boldsymbol{v}^T\boldsymbol{C}\boldsymbol{v} \leq \lambda_1$$

이 성립한다. 이 부등식에 의해, 모든 원소가 $1/\sqrt{(p)}$인 p 차원의 단위벡터를 $\mathbf{1}_p$로 나타내면, 두 그룹사이의 Mahalanobis 거리는

$$\hat{\Delta}_p^2 = p\mathbf{1}_p^T\boldsymbol{C}\mathbf{1}_p \leq p\lambda_1$$

를 만족한다. 여기서 단위벡터 $\mathbf{1}_p$의 의미는 p 개의 확률변수가 같은 정도로 Δ_p^2의 값에 기여함을 뜻한다. 즉, 모든 변수가 같은 정도로 변수선택에 대한 정보를 제공할 때, 이차형태 $\boldsymbol{v}^T\boldsymbol{C}\boldsymbol{v}$는 $\hat{\Delta}_p^2$의 값을 결정한다. 이에 비추어서, 두 그룹사이의 Mahalanobis 거리 $\hat{\Delta}_p^2$에 가장 큰 기여를 하는 변수는 $\boldsymbol{C}$의 제일 큰 특성치에 해당하는 특성벡터의 성분의 절대치가 가장 큰 것에 해당하는 변수로 택한다. 이러한 변수를 $k=1$인 경우에 Rao의 추가정보검정의 대상이 되는 변수로 선택한다.

다음은 처음 k 개의 변수가 판별분석에 사용되는 변수로 이미 선택되었고, 나머지 $p-k$ 개의 변수 중에서 어떤 하나의 변수가 판별분석에 사용되는 변수로 추가 선택될 수 있는지를 살펴 본다. $k=1$인 경우와 같은 맥락으로 전개된다. $\hat{\boldsymbol{\delta}}_{2.1} = \hat{\boldsymbol{\delta}}_2 - \hat{\boldsymbol{\Sigma}}_{21}\hat{\boldsymbol{\Sigma}}_{11}^{-1}\hat{\boldsymbol{\delta}}_1$로 하고, $(p-k)\times(p-k)$ 행렬 $\boldsymbol{C}_{2.1}$를

$$\boldsymbol{C}_{2.1} = \boldsymbol{D}(\hat{\boldsymbol{\delta}}_{2.1})(\hat{\boldsymbol{\Sigma}}_{22} - \hat{\boldsymbol{\Sigma}}_{21}\hat{\boldsymbol{\Sigma}}_{11}^{-1}\hat{\boldsymbol{\Sigma}}_{12})^{-1}\boldsymbol{D}(\hat{\boldsymbol{\delta}}_{2.1}) \tag{2.2}$$

로 정의한다. $\hat{\Delta}_p^2$와 $\hat{\Delta}_k^2$의 차이는

$$\hat{\Delta}_p^2 - \hat{\Delta}_k^2 = (p-k)\mathbf{1}_{p-k}^T\boldsymbol{C}_{2.1}\mathbf{1}_{p-k} \leq (p-k)\nu_1$$

를 만족한다. 여기서, ν_1은 $\boldsymbol{C}_{2.1}$의 제일 큰 특성치이다. $\hat{\Delta}_p^2 - \hat{\Delta}_k^2$는 나머지 $p-k$ 개의 변수가 주는 변수선택에 대한 추가정보가 된다. 해당되는 가설 H_k가 기각되지 않으면, $\hat{\Delta}_k^2$는 $\hat{\Delta}_p^2$를 근사화하므로, $\boldsymbol{C}_{2.1}$은 변수선택에 대한 추가정보를 주지 않는다. 그러므로, 나머지 $p-k$ 개의 변수 중에서 어떠한 것도 판별분석에 추가로 사용되지 않는 것으로 결론 내린다. 그러나, 가설 H_k가 기각되면, 나머지 $p-k$ 개의 변수 중에서 $\hat{\Delta}_p^2 - \hat{\Delta}_k^2$에 가장 큰 기여를 하는 변수는 $\boldsymbol{C}_{2.1}$의 제일 큰 특성치에 해당하는 특성벡터의 성분의 절대치가 가장 큰 것에 해당하는 변수로 택하고, 이 변수를 판별분석에 추가로 사용될 하나의 변수로 결정한다.

위의 내용을 정리하면, 변수선택을 위한 다음의 반복적인 절차를 얻는다. I를 선택되는 변수들의 지수집합으로 나타낸다. 지수집합 I에 해당되는 변수들이 $\boldsymbol{\delta}_1$에 대응한다. 물론 제일 처음 단계에서 지수집합 I는 공집합이다.

단계 1 : 식 (2.2)에서 정의된 행렬 $\boldsymbol{C}_{2.1}$의 제일 큰 특성치에 해당하는 특성벡터의 성분의 절대치가 가장 큰 것에 해당하는 변수를 지수집합 I에 추가한다. 첫 변수를 선택하는 단계에서만 $\boldsymbol{C}_{2.1}$대신에 식 (2.1)에서 정의된 행렬 $\boldsymbol{C}$를 사용한다.

단계 2 : 단계 1의 지수집합 I에 대응되는 δ_1을 택하여, 이에 해당되는 가설 H_k를 설정하고, Rao의 추가정보검정을 실시한다. 가설 H_k가 성립하면 변수선택절차를 종료한다. 만약 가설 H_k가 기각되면, 단계 1을 반복하여 지수집합 I를 갱신하고, 단계 2를 실시하여, 해당되는 가설 H_k가 성립할 때까지 반복한다.

3. 예제

SPLUS에 내장된 함수 *rnorm*을 이용하여, 4 변량 정규분포에서 크기가 30인 두 개의 모의표본을 추출한다. 이렇게 하기 위하여, 먼저 단변량 정규분포에서 240개의 모의자료를 추출하여, 크기가 120인 두 개의 표본으로 나누어, 각기 30×4인 자료행렬이 되도록 한다. 이 두개의 자료행렬을 다음의 평균과 공분산행렬을 갖도록 변환한다.

$$\boldsymbol{\mu}_1 = \begin{bmatrix} 4.0 \\ 4.0 \\ 2.5 \\ 2.5 \end{bmatrix}, \quad \boldsymbol{\mu}_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} 1.0 & 0.6 & 0.8 & 0.7 \\ 0.6 & 1.0 & 0.2 & 0.3 \\ 0.8 & 0.2 & 1.0 & 0.6 \\ 0.7 & 0.3 & 0.6 & 1.0 \end{bmatrix}.$$

이 변환된 자료의 처음 두 개의 변수가 정확히 가설 H_2를 만족한다.

앞에서 제시한 반복적인 방법을 사용하여 처음 두 개의 변수가 선택되는지 살펴 본다. 표 3.1은 각 변수선택 단계에서의 지수집합 I, Rao의 추가정보검정통계량 F와 해당되는 p-값을 보여준다.

표 3.1: 반복적인 추가정보 검정

단계	I	F	p-값
1	1	4.56	0.006
2	1, 2	0	1.0

표 3.1에 의하면, 단계 2에서 처음 두 개의 변수가 선택되고, 변수선택과정은 종료된다. 처음 두 개의 변수에 입각해서 Fisher의 선형판별함수를 사용하면 위의 모의자료를 정확하게 판별한다. 위에서 기술한 바와 같이 모의자료를 30회 반복하여 얻어서, 매회 앞에서 제시한 반복적인 방법을 사용하여 변수선택을 실시해도, 표 3.1의 결과와 마찬가지로 2단계에서 처음 두 개의 변수가 선택되었다.

참고문헌

Mardia, K. V., Kent, J. T. and Bibby, J. M.(1979). *Multivariate Analysis,* Academic Press, New York.

McKay, R. J. (1978). A graphical aid to selection of variables in two-group discriminant analysis, *Applied Statistics*, **27**, 259–263.

McKay, R.J. and Campbell, N.A. (1982a). Variable selection techniques in discriminant analysis I, Description, *British Journal of Mathematical and Statistical Psychology*, **35**, 1–29.

McKay, R.J. and Campbell, N.A. (1982b). Variable selection techniques in discriminant analysis II, Allocation, *British Journal of Mathematical and Statistical Psychology*, **35**, 30–41.

McLachlan, G.J. (1992). *Discriminant Analysis and Statistical Pattern Recognition,* Wiley, New York.

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications,* 2nd ed. Wiley, New York.

Schott, J.R. (1997). *Matrix Analysis for Statistics.* Wiley, New York.

[2005년 3월 접수, 2005년 8월 채택]

Improvement in Performing a Test for Additional Information

Myung Geun Kim[1)]

ABSTRACT

An iterative method that greatly reduces a burden of computation in performing Rao's test for additional information in two-group discriminant analysis is suggested. A numerical example is provided for illustration.

Keywords: Discriminant analysis, Rao's test for additional information, Selection of variables

1) Professor, Dept. of Applied Statistics, Seowon University, 231 Mochung-Dong, Cheongju, Chung-Buk 361-742, Korea
E-mail: mgkim@seowon.ac.kr