Journal of the KSQC
Vol. 8, No. 1 (1980)

분류방법과 그의 전산화에 관한 연구 *
-정준판별분석법을 중심으로-
Pattern Recognition and It's Computer Program
(By Canonical Discriminant Analysis)

김 재 주, 김 성 주**

Abstract

There are many methods of pattern recognition. In this paper we assume that the responses of independent m groups are described by p-variate normal random variables with distinct mean vectors and a common covariance matrix. Under the assumption we give pattern recognition of m groups by means of canonical discrininant analysis and it's computer program. An example is presented.

1. 서 론

이미 구별하여 설정한 m개의 집단이 있어 그 집단 성원의 특성이 p변량으로 주어 진다고 하자.

어느 집단에 속하는 가를 모르는 개체가 있을 때, 그 개체에 같은 p변량 특성을 관측하여 이 개체가 m개의 집단 중 어느 집단에 속하는 가를 판정하는 문제를 판별의 문제라고 한다[8].

그러나 실제로 문제가 되는 것은 m개의 집단을 구별하여 구성하는 것이 문제이다.

판별에서 말하는 집단은 통계적으로 모집단을 지정하고 있지만 실제의 데이타 해석에서는 각 집단마다 p변량 특성의 관측치가 유한개 얻어지므로 판별의 전 단계로 m개의 집단의 특징을 먼저 파악하여 두는 것이 중요하다.

즉, 각 집단의 차이와 유사성은 될수 있는 한 명확히 관측 특성상에 파악하는 문제를 분류의 문제라 정의하고 판별의 문제와 구별한다[9].

집단의 수가 2개 일때는 지시표본(index sample or initial sample)을 이용하여 판별함수를 추측 함으로서 이와같은 분류를 행하고 2개의 집단간의 차이로 추측한 판별함수를 써서 고찰할 수 있다[3][4].

일반적으로 집단의 수가 m개 관측특성의 수가 p라고 하면 m개의 집단을 p개의 관측특성에 의하여 특징을 부여하고 분류하려면 p개의 변량특성으로부터 되는 새로운 q개의 인자($q<p$)를 찾고 그 위에서 집단간의 차이를 보고자 하는 문제를 취급하는 것이 정준 판별분석법이다[5][6][7].

본 논문은 정준 판별분석법의 모델을 설정하고 추정과 검정은 어떻게 행하며 그것을 전산화하여 실례를 해결하는데 그 목적이 있다.

2. 정준판별분석법 모델

분류된 m개의 모집단을(Π_1, Π_2, ……, Π_m)라 하고, 제 g 모집단 Π_g는 p변량정규 모집단으로 그의 평균벡타를 μ_g, 분산행렬을 Σ_g라 한다.
($g=1, 2, \cdots\cdots m$).

* 1979학년도 문교부 학술연구 조성비에 의하여 일부지원된 것임.
** 서울대학교 자연과학대학 교수

$\Sigma_1=\Sigma_2=\cdots\cdots=\Sigma_m=\Sigma$

라 가정하고 Σ를 모집단내 분산행렬로 두어 Σ_w라 쓴다.

m개 모집단간의 변동의 상이를 정의하기 위하여 m개의 모평균의 공통평균을

$$(2,1)\quad \underset{\sim}{\mu}_{\cdot}=\frac{1}{m}\sum_{g=1}^{m}\underset{\sim}{\mu}_g$$

라 두고 모집단간 분산행렬

$$(2,2)\quad \Sigma_b=\sum_{g=1}^{m}(\underset{\sim}{\mu}_g-\bar{\underset{\sim}{\mu}}_{\cdot})(\underset{\sim}{\mu}_g-\bar{\underset{\sim}{\mu}}_{\cdot})'$$

을 정의한다.

문제는 p개의 변량에 무게(weight)를 부여하여 특징지우고 그상에서 Σ_b를 크게 하는 것이 의미가 있으므로 무게벡타를 $\underset{\sim}{a}'=(a_1, a_2\cdots\cdots a_p)$라 하고 $p\times 1$ 관측벡타 $\underset{\sim}{x}$에 대하는 변환

$$(2,3)\quad y=\underset{\sim}{a}'\underset{\sim}{x}=\sum_{i=1}^{p}ai\ xi$$

을 생각한다.

실제 이용의 편리를 위해

$$(2,4)\quad y=\underset{\sim}{a}'(\underset{\sim}{x}-\bar{\underset{\sim}{\mu}}_{\cdot})=\sum_{i=1}^{p}ai(xi-\bar{\mu}i)$$

로 둔다.

다음은 이와같은 변환하에서 Σ_w는

$$(2,5)\quad \mathrm{Var}(y)=\underset{\sim}{a}'\Sigma_w\underset{\sim}{a}$$

로 변환되고 Σ_b는

$$(2,6)\quad v_g=\underset{\sim}{a}'(\underset{\sim}{\mu}_g-\bar{\underset{\sim}{\mu}}_{\cdot}),\quad \bar{v}=\frac{1}{m}\sum_{g=1}^{m}v_g$$

라 둘때,

$$(2,7)\quad {}_B\mathrm{Var}(y)=\frac{1}{m}\sum_{g=1}^{m}(v_g-\bar{v}_{\cdot})^2=\underset{\sim}{a}'\Sigma_b\underset{\sim}{a}$$

가 된다.

문제는 $\mathrm{Var}(y)$와 ${}_B\mathrm{Var}(y)$와의 비가 최대가 되도록 표준조건

$$(2,8)\quad \mathrm{Var}(y)=\underset{\sim}{a}'\Sigma_w\underset{\sim}{a}=1$$

하에서 $\underset{\sim}{a}$를 정하는 것이다.

이것은 Lagrange의 승수 λ를 써서

$$(2,9)\quad F=\underset{\sim}{a}'\Sigma_B\underset{\sim}{a}-\lambda(\underset{\sim}{a}'\Sigma_w\underset{\sim}{a}-1)$$

을 최대로 하면 된다.

따라서 방정식

$$(2,10)\quad \frac{1}{2}\frac{\partial F}{\partial \underset{\sim}{a}}=\Sigma_B\underset{\sim}{a}-\lambda\Sigma_w\underset{\sim}{a}=(\Sigma_B-\lambda\Sigma_w)\underset{\sim}{a}=0$$

가 얻어진다.

이것에서 $\underset{\sim}{a}$가 (2, 10)식의 0아닌 해를 가지기 위해서는

$$(2,11)\quad |\Sigma_B-\lambda\Sigma_w|=0$$

가 되는 것이 필요하다.

결국 λ는 행렬 Σ_B의 행렬 Σ_w에 관한 고유치이고, $\underset{\sim}{a}$는 이 고유치에 대응하는 고유벡타이다.

(2, 10)식과 (2, 11)식을 보면,

고유치 λ와 고유벡타 $\underset{\sim}{a}$는 일반적으로 p개 얻어진다.

고유치를 큰것으로부터 차례로 $\lambda(1)$, $\lambda(2)\cdots\cdots$, $\lambda(p)$라 하고, 그것에 대응하는 고유벡타를 각각 $\underset{\sim}{a}(1)$, $\underset{\sim}{a}(2)\cdots\cdots$, $\underset{\sim}{a}(p)$라 하면 (2, 4)식의 y도 p개 구해진다. 결국

$$(2,12)\quad yi=\underset{\sim}{a}'(i)\quad(\underset{\sim}{x}-\bar{\underset{\sim}{\mu}}_{\cdot})\quad i=1,2,\cdots\cdots,p$$

가 되고 이것을 제 i정준변량이라 부르고 $\underset{\sim}{a}$를 제 i 정준판별계수벡타라고 부른다.

이와 같이 하여 얻어진 p개의 정준변량의 성질을 보다 명확히 하기 위하여 (2, 10) (2, 11) 식의 해법을 생각해 보자. (2, 10) 식에서와 $\underset{\sim}{a}$를 구하는 것은 실제로는

$$(2,13)\quad (\Sigma_w^{-1}\Sigma_B=\lambda I)\underset{\sim}{a}=0$$

의 해를 구하는 것을 의미한다.

$\Sigma_w^{-1}\Sigma_B$는 대칭행렬이 아니므로 간단히 해를 구할 수 없다.

Σ_w는 정정치 부호행렬이라 가정되므로

$$(2,14)\quad L'\Sigma_w L=I$$

이 되는 정측행렬 L가 존재한다.

여기에서 I는 $p\times p$ 항등행렬이다.

$$(2,15)\quad \underset{\sim}{a}=L\underset{\sim}{b}$$

라 두면 (2, 10) 식에서

$$(2,16)\quad [(L'\Sigma_B L)-\lambda I]\underset{\sim}{b}=0$$

이 성립한다. 제약조건 (2, 8)은

$$(2,17)\quad \underset{\sim}{b}'\underset{\sim}{b}=1$$

을 의미하므로

(2, 10)을 푸는 것은 대칭행렬$(L'\Sigma_B L)$의 고유치

와 고유벡타를 구하는 것으로 환원된다.

그러기 위해서는 L를 구하지 않으면 안되지만 우선 $\sum_w$의 고유치와 고유벡타를

(2,18) $(\sum_w - rI)\ \underline{f} = 0,\ \underline{f}'\underline{f} = 1$

로 부터 구한다.

이것을 풀어서 얻어지는 p개의 고유치($r_1, r_2 \cdots\cdots r_p$)을 대각 성분으로 하는 행렬 p와 그것에 대응하는 고유벡타 $f_1, f_2 \cdots\cdots f_p$을 나열하여 되는 행렬$F$로 부터

(2,19) $L = F r^{-1/2}$

이 얻어진다. (2,16) 식을 보면

L는 $p \times p$ 정측행렬이므로 이것을 풀어서 얻어지므로 양의 고유치 λ의 갯수는 $\sum_B$의 계수(rank)와 같다.

$\sum_B$의 계수는 $(m-1)$과 p의 적은 것과 같아진다.

따라서, $(m-1) > p$이면 모든 고유치 $\lambda(1), \cdots\cdots \lambda(p)$은 양이 되지만, $(m-1) \leqq p$이면 $(m-1)$ 개의 고유치 $\lambda(1)\cdots\cdots, \lambda(m-1)$은 양이되고, 나머지 $(p-m+1)$개의 고유치 $\lambda(m)\cdots\cdots, \lambda(p)$는 모두 0이 된다.

현실적인 문제에 있어서는 정준변량은 반드시 $(m-1)$과 p의 적은수 만큼 얻어지지만 이들 모두를 해석할 필요는 없다.

그러기 위해서는 전체의 모집단간 변동을 r 개 ($r \leqq \min(m-1, p)$의 정준변량으로 어느정도 나타낼 수 있는가를 나타내는 측도로 분류의 기여율

(2,20) $\dfrac{\lambda(1)+\lambda(2)+\cdots\cdots+\lambda(r)}{\lambda(1)+\lambda(2)+\cdots\cdots+\lambda(k)} \times 100$단

$k = m-1$ 혹은 p를 사용한다.

이 식으로 부터 알 수 있는 바와 같이 기여율은 정준변량의 수 r을 늘리면 점점 더 커지고 $r = \min(m-1, p)$일때 100%가 된다.

그러나 각 모집단의 특징을 명확히하여 분류한다는 입장에서 생각하면 정준변량의 수 r는 적지만 기여율을 높이는데 정준판별 분석의 목적이 있다.

특히 $m = 2$ 정준판별분석법의 정준변량은 판별함수와 일치한다[9].

3. 정준판별 분석법에 있어서 추측

1) 추 정

m개의 모집단에 있어서 모평균 $\underline{\mu}_1, \underline{\mu}_2, \cdots\cdots, \underline{\mu}_m$와 모분산행렬 $\sum_w$는 모르는 경우가 많아 실지 적용에 있어서는 표본으로부터 추정하지 않으면 안된다.

이제 제 g모집단 Π_g $(g = 1, 2, \cdots\cdots, m)$로 부터의 표본을 추출하여 그 크기를 n_g라 하고, 이들 n_g 개의 표본으로부터 되는 표본집단을 R_g라 한다.

이들 관측치 벡타를 $p \times 1$ 벡타로

$\underline{x}_{gl} = [x_{g1l}, x_{g2l} \cdots\cdots, x_{gpl}]'$ $(g = 1, 2, \cdots\cdots, m;\ l = 1, 2, \cdots\cdots, n_g)$

라 하고,

(3,1) $n = n_1 + n_2 + \cdots\cdots + n_m$

(3,2) $\underline{\bar{x}}_{g\cdot} = \dfrac{1}{n_g}\sum_{l=1}^{n_g} \underline{x}_{gl} \quad g = 1, 2, \cdots\cdots, m$

(3,3) $\underline{\bar{x}}_{\cdots} = \dfrac{1}{m}\sum_{g=1}^{m} \underline{\bar{x}}_{g\cdot} = \dfrac{1}{n}\sum_{g=1}^{m}\sum_{l=1}^{n_g} \underline{x}_{gl}$

라 하면

이때 표본집단내 분산행렬은

(3,4) $u_g = \dfrac{1}{n_g - 1}\sum_{l=1}^{n_g} (\underline{x}_{gl} - \underline{\bar{x}}_{g\cdot})(\underline{x}_{gl} - \underline{\bar{x}}_{g\cdot})'$

$g = 1, 2, \cdots\cdots, m$로 추정된다.

따라서 모든 표본집단에 공통의 집단내분산행렬은

(3,5) $u_w = \dfrac{1}{n-m}\sum_{g=1}^{m} (n_g - 1)\, u_g$

로 추정된다.

같은 방법으로 표본집단간 분산행렬은

(3,6) $v_B = \dfrac{1}{n}\sum_{g=1}^{m} n_g (\underline{\bar{x}}_{g\cdot} - \underline{\bar{x}}_{\cdot\cdot})(\underline{\bar{x}}_{g\cdot} - \underline{\bar{x}}_{\cdot\cdot})'$

로 추정된다.

Seal [6]과 浅野[7]는 전분산행렬을

$$u_t = \frac{1}{n-1}\sum_{g=1}^{m}\sum_{l=1}^{n_g} (\underline{x}_{gl} - \underline{\bar{x}}_{\cdot\cdot})(\underline{x}_{gl} - \underline{\bar{x}}_{\cdot\cdot})'$$

로 추정하고 이것을 (3,5) 식을 써서 집단간 분산행렬을

$$v_B = \frac{1}{m-1}\{(n-1)\, u_t - (n-m)\, u_w\}$$

로 정의했다.

이들 추정량 $\underline{\bar{x}}_{g\cdot}$, $\underline{\bar{x}}_{\cdot\cdot}$, u_g, u_w 및 v_B는 각각 $\underline{\mu}_g$, $\bar{\mu}_\cdot$, $\sum_g$, $\sum_w$ 및 $\sum_B$의 일치추정량이고 v_B 이외는

불편 추정량이기도 하다.

그러므로 $\sum_B$의 불편추정량으로 (3,6)식을 조금 수정하여

(3,7) $u_B = v_B - (\frac{m-1}{n})\, u_{\bar{w}}$

을 정의한다.

$\sum_w$와 $\sum_B$ 대신에 표본으로 부터 u_w와 v_B가 추정되면 이것을 (2,10)식과 (2,11)식에 대입하여,

(3,8) $(v_B - \hat{\lambda}\, u_{\bar{w}})\ \hat{\underline{a}} = 0$

(3,9) $| v_B - \hat{\lambda}\, u_{\bar{w}} | = 0$

을 얻고 정준판별계수 벡타 $\hat{\underline{a}}$가 추정된다.

이것의 해법은 (2,14)식에서 (2,19)에는 계산과정과 같이 하면 된다.

제 i정준변량은

(3,10) $y_i = \hat{\underline{a}}_i' (\underline{x} - \bar{\underline{x}}..)\quad i=1,2,\cdots\cdots,r$

로 추정되고 기여율도,

(3,11) $\dfrac{\hat{\lambda}\,(i)}{\hat{\lambda}\,(1)+\hat{\lambda}\,(2)+\cdots\cdots+\hat{\lambda}\,(r)} \times 100$

$i=1,2,\cdots\cdots,r$ 로 얻어진다.

단, $r=\min(m-1,\ p)$이고 $\hat{\lambda}\,(1),\ \hat{\lambda}\,(2),\cdots\cdots,\hat{\lambda}\,(r)$는 (3,9)식의 근의 큰 순서로 r개 나열한 것이다.

2) 정준변량상의 모평균의 신뢰한계

정준판별 계수벡타 $\hat{\underline{a}}(1),\ \hat{\underline{a}}(2),\cdots\cdots,\hat{\underline{a}}(r)$와 정준변량 (3,10)식이 추정되면 다음은 정준변량상에서 m개의 집단의 분류 상태를 알고져 한다. 결국 m개의 집단중 어느것들이 가까이 있으며, 어느것이 떨어져 있는가를 정준변량상에서 파악코져 한다.

그 결과로 몇개의 집단간을 분류하는데 어느 관측 특성이 영향을 하는가를 명확하게 알 수 있게 된다.

이와 같은 목적으로 여기서는 정준변량상에서 각 모집단에서 평균벡타의 신뢰한계를 구하여 본다.

정준변량의 각 모집단의 평균은 (2,6)식의 $v_g\,(g=1,2,\cdots\cdots,m)$로 정의되어 있다.

이것에 의하여 제 i정준변량상의 각 모집단의 모평균을 $v_{ig}\,(i=1,2,\cdots\cdots,r:\ g=1,2,\cdots\cdots,m)$라 하고, 각 관측치 벡타 $\underline{x}_{gl}(g=1,2,\cdots\cdots,m;\ l=1,2,\cdots\cdots,n_g)$에 대하는 제 i정준변량 (3,10)식의 값을

(3,12) $y_{igl} = \hat{\underline{a}}'_i\,(\underline{x}_{gl} - \bar{\underline{x}}..)$

라 두고,

다시 제 i정준변량상의 모평균을

(3,13) $\bar{v}_{i\cdot} = \dfrac{1}{m}\sum_{g=1}^{m} v_{ig},\quad i=1,2,\cdots\cdots,r$

라 두면 v_{ig}와 $\bar{v}_{i\cdot}$는 표본에서 각각

(3,14) $\bar{y}_{ig\cdot} = \dfrac{1}{n_g}\sum_{l=1}^{n_g} y_{igl}\quad \bar{y}_{i\cdot\cdot} = \dfrac{1}{m}\sum_{g=1}^{m}\bar{y}_{ig\cdot}$ 로 추정된다.

그런데 임의의 관측벡타 $\underline{x}_{gl}\,(g=1,2,\cdots\cdots,m;\ l=1,2,\cdots\cdots,m)$에 대하여 제 정준변량상에서의 y_{igl}와 제 j정준변량상에서의 $y_{igl}(i\neq j)$는 각각 모평균 v_{ig}와 v_{jg}를 가지고 서로 독립이고, 그의 분산은 표준화 조건(2,8)에 의하여 다같이 1이 된다.

또한 m개의 모든 모집단은 모두 p변량 정규분포를 따른다고 가정했으므로 y_{igl}와 y_{jgl}는 정규분포를 따른다고 생각해도 좋다.

따라서 $y_{ig\cdot}$는 $N(v_{ig},\ \frac{1}{n_g})$을 따르므로

$\sqrt{n_g}\,(\bar{y}_{ig\cdot} - v_{ig})$는 $N(0,1)$을 따른다.

이상으로부터 모평균 $v_{ig}\,(i=1,2,\cdots\cdots,r;\ g=1,2,\cdots\cdots,m)$의 신뢰계수 $100(1-\alpha)\%$의 신뢰구간은,

$\sqrt{n_g}\,|\,\bar{y}_{ig\cdot} - v_{ig}\,| \leq k_\alpha$

혹은

(3,15) $\bar{y}_{ig\cdot} - \dfrac{1}{\sqrt{n_g}}\,k_\alpha \leq v_{ig} \leq \bar{y}_{ig\cdot} + \dfrac{1}{\sqrt{n_g}}\,k_\alpha$

로 구해진다. 단, k_α는 정규분포의 $100_\alpha\%$ 점이다.

(3,15)식으로 주어지는 모평균 v_{ig}의 신뢰 구간은 개개의 정준변량상에 독립으로 구한 것이다. 그러나 실제문제에서는 정준변량$(y_1, y_2,\cdots\cdots, y_r)$ 상에서 모평균 벡타$(v_{1g}, v_{2g},\cdots\cdots, v_{rg})$의 동시 신뢰한계를 구하여 고찰하는 편이 자연스럽다.

그러기 위하여 다음은$(v_{1g}, v_{2g},\cdots\cdots, v_{rg})$의 동시 신뢰한계를

$\sqrt{n_g}\,(\bar{y}_{ig\cdot} - v_{ig})\ (i=1,2,\cdots\cdots,r;g=1,2,\cdots,m)$

$N(0,1)$을 따름으로,

(3,16) $\chi^2 r(g) = \sum_{i=1}^{r}\{\sqrt{n_g}\,(\bar{y}_{ig\cdot} - v_{sg})\}^2$

$= n_g \sum_{i=1}^{r}(\bar{y}_{ig\cdot} - v_{ig})^2$

는 자유로 r의 χ^2-분포를 따른다.

그러므로$(v_{1g}, v_{2g},\cdots\cdots, v_{rg})$에 관한 신뢰계수$100(1-\alpha)\%$의 동시 신뢰한계는

$$(3,17)\quad \sum_{i=1}^{r}(\bar{y}_{ig}-v_{ig})^2 \leqq \frac{1}{n_g}\chi_r^2(\alpha)$$

$g=1,2,\cdots\cdots,m$

로 주어진다. 단, $\chi_r^2(\alpha)$는 자유도 r의 χ^2분포의 $100\alpha\%$점이다.

특히 $r=2$ 인 경우는 (3,17)식은

$$(3,18)\quad (\bar{y}_{1g\cdot}-v_{1g})^2+(\bar{y}_{2g}-v_{2g})^2 \leqq \frac{1}{n_g}\chi_2^2(\alpha)$$

로 되고, 이것은 제 1 정준변량 y_1과 제 2 정준변량 y_2로 되는 직교좌표 축(y_1, y_2)상에 중심을$(\bar{y}_{1g\cdot}, \bar{y}_{2g\cdot})$로 하는 반경 $\chi_2(\alpha)/\sqrt{n_g}$의 원주상과 내부를 의미한다.

이와같은 관점으로부터 제 1 정준변량 y_i를 제 i 정준축이라 부르기도 한다.

3) 정준변량의 유의성 검정.

정준변량의 수를 r개, 즉 $\min(m-1,p)$개 구했지만 이들 모두가 m개의 집단의 분류에 의미를 가지고 있다고는 볼 수 없다.

실제로 r개의 정준변량 가운데 0에 가까운 고유치에 대응하는 정준변량은 기여율도 낮고 m개의 집단의 분류에 공헌하지 못한다.

그래서 몇개의 정준변량이 m개의 집단의 분류에 기여하는가를 보기 위하여 이들 정준변량에 대응하는 고유치를 큰것으로부터 검토하는 것이 보통이다.

여기에 Bartlett〔1〕의 검정이 이용된다.

이것은 r개의 고유치 가운데 최초의 k개 까지는 0가 아니지만 제 $k+1$이후의 고유치가 0이라는 가설을 검정한다.

결국 귀무 가설

$HO: \lambda(k+1)=\lambda(k+2)=\cdots\cdots=\lambda(r)=0,\ r=\min(m-1,\ p)$

이 검정된다. 이것은 통계량

$$(3,19)\quad \chi_0^2(k)=\{(n-1)-\frac{1}{2}(p+m)\}\ 1og_e\{\prod_{j=k+1}^{r}1+\hat{\phi}(j)\}$$

가 근사적으로 자유도$(p-k)(m-k-1)$ χ^2 분포를 따른다는 것을 이용한다.

단, $\hat{\phi}(i)$ $(i=1.2.\cdots\cdots,r)$는 $\hat{\lambda}(i)$와

$$(3,20)\quad \hat{\phi}(j)=\frac{n}{n-m}\hat{\lambda}(j)$$

의 관계가 있다.

이 검정을 k에 관하여 축차 반복하여가면 몇개의 정준변량이 유의가 되는가. 다시 말해서 m개의 집단의 분류에 유의로 공헌하는 정준축의 갯수가 몇개인가를 알 수 있다.

4. 예 제

30인의 고객에 의뢰하여 3종류의 남성화장품 (로숀)의 사용감을 조사했다.

사용한 조사대상은 A사의 P제품, B사의 Q 제품 및 C 사의 R제품이다.

30인의 고객을 렌덤으로 3군으로 나누고 각 10사람에게 한 종류의 화장품을 사용하게하여 다음의 5항목,

1. 향 기.
2. 전체적인 느낌.
3. 기분좋은 자극감.
4. 점 도.
5. 피부의 긴축성.

에 관하여 (2, 1, 0, −1, −2)의 5단계 평가를 하여 표 4,1과 같은 데이타를 얻었다. 이들 데이타로 부터 정준판별 분석법을 사용하여 P제품, Q제품 및 R제품의 사용특성을 조사하여 보자.

표 4,1 남성화장품의 사용감 조사

	고객	항목				
		1	2	3	4	5
A 사 P 제 품	1	−1	−1	0	0	0
	2	−1	0	−1	0	−1
	3	−1	0	0	1	1
	4	−1	−1	−1	−1	1
	5	0	1	0	−1	−1
	6	1	1	1	1	0
	7	0	1	0	1	0
	8	−1	0	2	0	−1
	9	1	1	−1	0	−1
	10	−1	0	−1	0	0
B 사	1	2	2	1	1	1
	2	2	0	2	0	1
	3	0	2	1	1	0
	4	1	1	2	2	0

Q 제 품	5	0	−1	0	0	1
	6	0	−1	0	0	1
	7	1	1	1	0	0
	8	2	0	1	0	0
	9	2	−1	1	0	2
	10	1	0	1	0	2
C 사	1	−1	−1	0	−1	−1
	2	0	−1	−1	−1	0
	3	0	−1	0	−1	−2
	4	−1	0	−1	0	−1
	5	0	−2	−1	−1	0
	6	−2	−1	−2	−2	−2
R 제 품	7	0	0	−1	−1	0
	8	1	0	0	0	−1
	9	−2	−2	−2	−2	−1
	10	−1	−2	−1	−2	−1

우선 (3,2)식과 (3,3)식으로부터 평균벡타를 추정한다. 이것은 표 4,2와 같다.

표 4,2 평균 벡타의 추정.

		$\bar{x}_{P\cdot}$	$\bar{x}_{Q\cdot}$	$\bar{x}_{R\cdot}$	$\bar{x}_{\cdot\cdot}$
성 분	1	−0.4	1.1	−0.6	0.03
	2	0.2	0.3	−1.0	−0.17
	3	−0.1	1.0	−0.9	0
	4	0.1	0.4	−1.1	−0.2
	5	−0.5	0.8	−0.9	−0.2

표본 집단내 분산행렬은 (3,4)식과 (3,5)식을 써서 추정된다. 이것을 표 4,3에 나타낸다.

표 4,3 표본집단내 분산행렬 U_w

	1	2	3	4	5
1	0.7233	0.2833	0.2733	0.1800	0.1267
2	0.2833	0.7900	0.2067	0.3533	−0.1133
3	0.2733	0.2067	0.5933	0.2400	0.0133
4	0.1800	0.3533	0.2400	0.4733	0.1133
5	0.1267	−0.1133	0.0133	0.1133	0.5667

표본 집단간 분산행렬은 (3,6)식에 의하여 표 3,4와 같이 된다.

표 4,4 표본 집단간 분산행렬 V_B

	1	2	3	4	5
1	0.5756	0.2889	0.5600	0.3600	0.5467
2	0.2889	0.3489	0.3933	0.3800	0.3133
3	0.5600	0.3933	0.6067	0.4600	0.5533
4	0.3600	0.3800	0.4600	0.4200	0.3800
5	0.5467	0.3133	0.5533	0.3800	0.5267

이것을 (3,8)식과 (3,9)식에 대입하여 $\hat{\lambda}$, $\hat{a}$ 을 추정해야 하지만 이것은 5×5 행렬방정식이 되므로 손 계산으로는 거의 불가능하므로 부록의FORTRAN IV 프로그램을 써서 구한다. (부록 참조)

이제 $p=5$, $m=3$에서 정준변량은 2개만 구해진다.

그 결과,

$$\Gamma = \begin{bmatrix} 1.2803 & 0 & 0 & 0 & 0 \\ 0 & 0.6242 & 0 & 0 & 0 \\ 0 & 0 & 0.4063 & 0 & 0 \\ 0 & 0 & 0 & 0.3423 & 0 \\ 0 & 0 & 0 & 0 & 0.1401 \end{bmatrix}$$

$$F = \begin{bmatrix} 0.4469 & 0.2896 & 0.5856 & 0.5319 & 0.3010 \\ 0.5988 & -0.4539 & -0.2873 & 0.3371 & -0.4891 \\ 0.4126 & 0.1013 & 0.4519 & 0.7064 & -0.3410 \\ 0.5167 & 0.0876 & -0.4875 & -0.3015 & 0.6299 \\ 0.0666 & 0.8320 & -0.3642 & 0.1165 & -0.3964 \end{bmatrix}$$

$$L = F\Gamma^{-\frac{1}{2}} = \begin{bmatrix} 0.3950 & 0.2560 & 0.5176 & 0.4701 & 0.2661 \\ 0.7579 & -0.5745 & -0.3636 & 0.4267 & -0.6191 \\ 0.6473 & 0.1589 & 0.7089 & -1.1083 & -0.5350 \\ 0.8832 & 0.1497 & -0.8333 & -0.5154 & 1.0768 \\ 0.1779 & 2.2222 & -0.9727 & 0.3112 & -1.0587 \end{bmatrix}$$

$$L'V_BL = \begin{bmatrix} 1.2111 & 0.8375 & -0.0820 & -0.2539 & -0.4477 \\ 0.8375 & 0.8278 & 0.1694 & -0.0936 & -0.3851 \\ -0.0820 & 0.1694 & 0.2111 & 0.0917 & -0.0383 \\ -0.2539 & -0.0936 & 0.0917 & 0.8027 & 0.0690 \\ -0.4477 & -0.3851 & -0.0383 & 0.0690 & 0.1884 \end{bmatrix}$$

가 된다.

이것으로부터 (2,16)식의 $L'\sum_B L$에 $L'V_BL$ 을 대입하여 풀면

$\hat{\lambda}(1)=2,100$ $\hat{\lambda}(2)=0.4187$ 및

$$\hat{\underline{b}}(1)=\begin{bmatrix}0.7399\\0.5892\\0.0204\\-0.1296\\-0.2971\end{bmatrix}\qquad \hat{\underline{b}}(2)=\begin{bmatrix}0.3834\\-0.4857\\-0.7087\\-0.3279\\0.0860\end{bmatrix}$$

이 얻어진다.

따라서 (2, 15) 식에서 정준판별 계수벡타는

$$\hat{\underline{a}}(1)=L\hat{\underline{b}}(1)=\begin{bmatrix}0.1702\\0.3575\\0.7869\\-0.0457\\0.9412\end{bmatrix}\quad \hat{\underline{a}}(2)=L\hat{\underline{b}}(2)=\begin{bmatrix}-0.9065\\0.5000\\-0.1072\\0.9710\\-0.2403\end{bmatrix}$$

으로 추정된다. 고유치의 크기로 부터 제 1 정준변량의 기여율은

$$\frac{2.1000}{2.1000\times 0.4187}\times 100\fallingdotseq 83.38$$

이 되고, 제 2 정준변량의 기여율은 100−83.38= 16.62로 주어진다. 이것으로부터 여기서 조사한 3 종류의 남성화장품은 제 1 정준변량상에 거의 분류되고, 제 1 정준판별 계수벡타 $\hat{\underline{a}}(1)$의 성분의 크기로부터 제 3 항목의 "기분좋은 자극감"과 제 5 항목의 "피부의 긴축성"이 이 분류에 크게 공헌하고 있음을 안다.

다음은 3 개의 집단의 모평균의 신뢰한계를 구하여 보자. 정준판별계수 $\hat{a}(1)$와 $\hat{a}(2)$는 이미 구해졌으므로 정준변량상의 각 관측치의 값 즉 정준평점은(3 ,12) 식에서 구 할 수 있다. 이것은 표 4 , 5 와 같이 된다.

표4.5 정준평점 $y_{igl}(i=1,2\cdots;g=P,Q,R;l=1,2\cdots10)$

	P		Q		R	
	y_{1pl}	y_{2pl}	y_{1ql}	y_{2ql}	y_{1rl}	y_{2rl}
1	−0.2631	0.5923	2.7599	0.3175	−1.1149	−0.2116
2	−1.5466	1.3885	2.8567	−1.7511	−0.7838	−1.1389
3	0.9236	1.8646	1.5611	2.1382	−1.8657	−0.7771
4	−2.7333	0.1162	2.0455	1.8101	−1.5466	1.3885
5	−0.2956	0.0699	0.7864	−0.4330	−1.1187	−0.6975
6	1.3755	0.8743	0.7864	−0.4330	−3.5527	−0.0242
7	0.5078	1.7676	1.4238	−0.1594	−0.4488	−0.6704
8	0.6084	1.0949	1.2382	−1.4234	−0.5296	−0.3002
9	−0.9130	0.2662	2.7035	−2.3316	−2.9875	−0.7226
10	−0.6465	1.1587	2.8891	−1.0876	−2.1198	−1.6159

$\bar{y}_{ig}\cdot(i=1,2;g=P,Q,R)$ 는 각각

$\bar{y}_{1p}\cdot=-0.2938$ $\quad$ $\bar{y}_{2p}\cdot=0.9053$

$\bar{y}_{1q}\cdot=1.9051$ $\quad$ $\bar{y}_{2q}\cdot=-0.3373$

$\bar{y}_{1r}\cdot=-1.6068$ $\quad$ $\bar{y}_{2r}\cdot=-0.5680$

이다.

$\alpha=0.05$라 두면 $\chi^2_2(0.05)=5.99$이므로 (3 , 18) 식에서

$$(-0.2983-v_{1p})^2+(0.9053-v_{2p})^2\leq\frac{1}{\sqrt{10}}\,5.99\fallingdotseq 0.5991$$

$$(1.9051-v_{1q})^2+(-0.3373-v_{2q})^2\leq 0.5991$$

$$(-1.6068-v_{1r})^2+(-0.5680-v_{2r})^2\leq 0.5991$$

을 얻는다. 이것을 그림을 그려 살펴보면 제 1 정준변량 상에서는 Q제품과 R제품은 많이 떨어져 있으므로 이들 두 제품은 우선 정준 판별계수의 무게를 합쳐 생각하면"기분좋은 자극감"과 "피부의 긴축성"에서 상당한 차이가 있다고 볼 수 있고, P제품은 Q제품과 R제품의 중간에 위치하고 있지만 R제품에 가깝다. 같은 방법으로 제 2 정준 변량상에서는 생각하여 분석해야 되지만 이 경우의 기여율은 16.62 %에 지나지 않으므로 이 예에서는 제 1 정준변량에서만 해석 하고 각 제품이 차지하는 위치를 보는것이 타당하다.

마지막으로 정준변량의 유의성 검정을 행하여 보자. 이 예에서는

$n=30, m=3, r=2$, $\hat{\lambda}(1)=2.1000$, $\hat{\lambda}(2)=0.4187$이므로

(3 , 20) 식에서

$$\hat{\phi}(1)=\frac{30}{30-3}(2.1000)=2.3333$$

$$\hat{\phi}(2)=\frac{30}{30-3}(0.4187)=0.4652$$

가 얻어진다. 우선 가설

H_o: $\lambda(1)=\lambda(2)=0$

에 대한 검정을 실시한다.

(3 , 19) 식에서

$$\chi^2_0(0)=\{(30-1)-\tfrac{1}{2}(5+3)\}\log_e(1+2.3333)(1+0.4652)\fallingdotseq 39.6300$$

이 값은 자유로 10의 χ^2 분포의 5 %의 점 18.31을 얻으므로 가설 H_o을 기각한다.

결국 2 개의 정준변량 모두를 사용하면 3 개의 집단의 분류가 유의하게 얻어 난다는 것을 말하고 있다.

다음으로 가설

H_o: $\lambda(2)=0$

을 검정하여 보자.

(3, 19)식에서 $K=1$

$\chi_0^2(1) \fallingdotseq 9.63$

이다. 이것도 자유도 4의 χ^2 분포의 5%점 9.49을 넘으므로 H_o를 기각한다.

결국 제1정준변량만으로 분류하더라도 유의가 된다. 그러나 제2정준변량의 분류에 미치는 기여율이 적으므로 분류의 특징을 해석하는데는 제1정준변량만으로도 충분하다고 본다.

5. 결 론

본 논문의 이론과 그 전산화 결과는 본 논문에서 취급한 예제 이외에도 의사가 환자를 진단할 때 환자의 임상 소견성적으로(p변량특성) 질환군을 구성하는 문제, 공산품의 품질특성(p변량특성)으로 제품의 등급을 구성하는 문제, 조립공장에서는 불량롯트를 조기 발견하기 위하여 원료특성 중간제품특성을 관측하여 양품군과 불량품군을 구성하는 문제등에 이용될 것으로 믿으나 본 논문의 모델 설정에 있어서 모집단을 p변량 정규분포를 가정한데 문제점이 있으며 예제에 있어서 정서적인 항목을 5단계 평가(2, 1, 0, −1, −2)로 한데도 문제점이 있는 것으로 본다.

참고문헌

(1) Bartlett, M. S. "The general canonical correlation distribution. "Ann. Math. Statist. Vol 18, pp. 1, 1947.

(2) Cooley, W. W. and Lohnes, P. S. "Multibiriate Procedures for the Behavioral Sciences," John Wiley & Sons, 1962.

(3) Lachenbruch, P. A. "Discriminant Analysis when the inistial Samples are misclassified", Technometrics Vol 8, 1966.

(4) Lachenbruch, P. A. "Estimation of error rates in discriminant Analysis" Technometrics, Vol 10, 1968.

(5) Rao, C. R. "Advanced Statistical Methods in Biometric Research," John Wiley & Sons, New York, 1952.

(6) Seal, H. "Multivariate Statistical Analysis for Biologists," Methen, 1964.

(7) 浅野長一郎. "因子分析通論," 共立出版, 1971.

(8) 奥野忠一, 芳賀敏郎"多変量解析法" 日科技連出版社 1971.

(9) 後藤昌司, "多変量テ"ータの 解析法"科学情 報社 1972.

부록 : 포트란 프로그램(FORTRAN Program)

이 정준판별 분석법의 포트란프로그램은 전분산 공분산행렬을 군간 및 군내 분산공분산행렬로 나누어 정준계수 및 정준평점을 계산한다. 이 프로그램에서는 주 프로그램(Main Program) 과 고유치와 고유벡타를 계산하기 위한 파우워(POWER)라는 서브프로그램(Subprogram)으로 되어 있다. 입력데이타(Input Data)의 양이 많을 경우를 생각해서 카드에 펀취된 입력데이타를 일단 마그네틱 테이프(Magnetic Tape)에 수록해서 처리 하기 위해 이 프로그램에서는 2개의 마그네틱 테이프가 사용되어진다.

또한 입력조정카드(Input Control Card)가 한장 사용되는데 여기서는 변량의 갯수, 그룹(group)수 전체표본수 각 그룹의 표본수 최대 반복수(Maximum Number of Iteration)를 지정한다.