韓國 OA學會 論文誌
第5卷 第3號, 2000. 9.

2000-5-3-4-2

범주형 다변량 데이터의 상관관계분석에 관한 기초적 연구(II)

노 형 진*

A Study on the Correlation Analysis about Categorical Multivariate Data(II)

Hyung-Jin Rho*

요 약

범주형 다변량 데이터의 상관관계분석을 위하여 개발한 수량화이론 III류나 대응분석 등의 기법은 다차원 공간상에서 점간의 거리로써 두 요소집합간의 관련성을 설명하는 데 있어서 매우 유용하다. 본 연구에서는 상관관계분석을 위한 대응분석의 특성을 수량화이론 III류와 비교하여 설명하고 그 유용성을 논하기로 한다. 이 기법은 사회과학 분야의 상관관계분석에 널리 활용될 것으로 기대된다.

Abstract

The purpose of this study is to suggest usefulness of correlation analysis about categorical multivariate data to business administration and other social sciences. Some examples of their application are presented.

* 경기대학교 경영학부 교수, 경영학박사

본고에서는 상관관계분석을 위한 대응분석의 특성을 수량화이론 Ⅲ류와 비교하여 설명하고 그 유용성을 논하기로 한다. 이 기법은 사회과학 분야의 상관관계분석에 널리 활용될 것으로 기대된다.

Ⅰ. 서 론

수량화이론(quantification theory)이라고 불리는 것은 L. Guttman의 예측이론으로부터 출발하여 일본의 하야시(林知己夫)에 의해서 전개된 독창적인 이론이다. 결과적으로는 더미변수법과 극히 유사한 것인데, 더미변수가 아직 보급되어 있지 않았던 1950년대 초에 이미 오늘날의 체계적인 형태를 갖추고 있었다는 점에서 높이 평가할 만하다. 더미변수의 경우와 마찬가지로 양적변수를 질적변수와 나란히 함께 쓸 수가 있다.

특히 수량화이론 Ⅲ류에 대해서는, 하야시의 흐름과는 독립적으로 프랑스에서는 J. P. Benzecri 및 그의 제자들이 대응분석(correspondence analysis)이라고 하는 수법을 개발하여 다방면에 걸쳐 응용하고 있다는 점에서 주목할 가치가 있다.

그런데 수량화이론 Ⅲ류에서는 얻어진 최적득점(고유벡터)을 좌표값으로 보고 요소집합 내의 요소를 유크리드 공간의 점으로 표현하는데 있어서, 행요소와 열요소의 산포도가 각각 별도로 그려지므로 때로 해석상의 어려움이 따른다. 또한 두 개의 산포도의 척도가 다르므로 공간상에서 점간의 거리로써 두 요소집합간의 관련성을 설명하는 데 있어서 곤란한 경우가 있다. 또한 수량화이론 Ⅲ류에서는 기여율(contribution) 및 누적기여율(cumulative contribution)과 관련해서, Gower(1966)의 주좌표분석(principal coordinates analysis)에서의 고유치전개에 그 이론적 기반을 두고 있다. 그러나 Gower는 그의 논문에서 주좌표분석을 주성분분석과 같은 맥락에서 논하고 있다. 즉 모든 변량이 양적일 때는 주성분분석이 그대로 그의 제안에 이용될 수 있다고 했다.

그렇다면 수량화이론 Ⅲ류에서의 기여율 및 누적기여율에 대한 정의는 자연스럽지 못하다. Tenenhaus & Young(1985)은 하야시(林知己夫)의 수량화이론 Ⅲ류를 분산분석법으로 분류했고, 주성분분석에 의한 접근방법은 Benzecri 등의 대응분석을 들고 있다. 실제로 Benzecri 등은 이러한 면에서 현저한 성과를 거두고 있다(Benzecri, 1973, 1977 ; Greenacre, 1984).

Ⅱ. 주성분분석에 의한 수량화

1. χ^2 거리의 도입

여기에서는 먼저 Gower가 생각했던 $\mathbf{Ax} = \lambda\mathbf{x}$ 형태의 고유치전개(固有値展開)를 확장하여 일반화된 고유치문제 $\mathbf{Ax} = \lambda\mathbf{Bx}$의 고유치전개에 상당하는 것을 정식화하기로 한다. 단, 여기에서 $A=(a_{ij})$는 n차의 반정치(半正値) 대칭행렬이고 rank $A = r \leq n$, $B=(b_{ij})$는 n차의 정치 대칭행렬이라고 한다. 모든 고유치를 크기순으로 $\lambda_1, \lambda_2, \cdots, \lambda_r$, 각각에 대응하는 고유벡터를 $\mathbf{x}_1$, $\mathbf{x}_2$, $\cdots$, $\mathbf{x}_r$이라 하고, $X=(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_r)$, $\Lambda=\mathrm{diag}(\lambda_i)$ (i=1, 2, $\cdots$, r)이라 놓으면,

$$AX = BA\Lambda \qquad (2.1)$$

라고 표현할 수 있다. 여기에서 고유벡터는 $X'BX = I$와 같이 직교화해 놓는다. 다음에 행렬 B의 고유치를 크기순으로 $\mu_1, \mu_2, \cdots, \mu_n$이라 하고, 대응하는 고유벡터를 $\mathbf{w}_1$, $\mathbf{w}_2$, $\cdots$, $\mathbf{w}_n$이라 한다. W = ($\mathbf{w}_1$, $\mathbf{w}_2$, $\cdots$, $\mathbf{w}_n$), M = diag (μ_l) (l=1, 2, $\cdots$, n)라고 행렬표시해서 $WW' = W'W = I$와 같이 정규직교화해 놓으면, $B = WMW'$라고 표현할 수 있다. $Y = B^{1/2}X$로 놓는다. 여기에서 $B^{1/2} = WM^{1/2}W'$, $M^{1/2}$ = diag($\sqrt{\mu_l}$)이다. 그런데 식 (2.1)로부터

$$AX = B^{1/2}B^{1/2}X\Lambda = B^{1/2}Y\Lambda,\ B^{1/2}AX = Y\Lambda$$

따라서 $B^{-1/2}AB^{-1/2} = Y\Lambda Y^{-1}$, 그런데 $Y'Y = X'B^{1/2}B^{1/2}X = X'BX = I$이므로 $Y' = Y^{-1}$. 그러므로 식 (2.1)은 $B^{-1/2}AB^{-1/2} = Y\Lambda Y'$라고 쓸 수 있으므로, $Y = B^{1/2}X$를 앞 식에 대입하여 변형하면

$$B^{-1}AB^{-1}=X\Lambda X' \tag{2.2}$$

를 얻는다. $B^{-1}=WM^{-1}W$이므로 B^{-1}의 $(i,\ j)$ 요소를 b^{ij}라고 하면, $b^{ij}=\sum_{l=1}^{n}\frac{w_{il}w_{jl}}{\mu_l}$라고 쓸 수 있다. 식 (2.2)를 요소의 형태로 쓰면,

$$\sum_{k=1}^{n}\sum_{k'=1}^{n}b^{ik}b^{jk'}a_{kk'}=\sum_{l=1}^{r}\lambda_l x_{il}x_{jl}$$

$$(i,j=1,2,\cdots,n) \tag{2.3}$$

이 된다. 이것이 Gower의 $a_{ij}=\sum_{l=1}^{r}\lambda_l x_{il}x_{jl}$

$(\lambda_l>0\ ;\ j=1,2,\cdots,r)$에 상당하는 것이다. 식 (2.3)으로부터

$$\sum_{k=1}^{n}\sum_{k'=1}^{n}(b^{ik}b^{ik'}-2b^{ik}b^{jk'}+b^{jk}b^{jk'})a_{kk'}$$

$$=\sum_{l=1}^{r}\lambda_l(x_{il}-x_{jl})^2\quad(i,j=1,2,\cdots,n) \tag{2.4}$$

이 도출된다. 이것은 Gower의

$a_{ii}-2a_{ij}+a_{jj}=\sum_{l=1}^{r}\lambda_l(x_{il}-x_{jl})^2$에 상당한다.

$u_{il}=\sqrt{\lambda_l}\,x_{il}$이라고 놓으면 식 (2.4)의 우변은 두 점 $i,\ j$간의 r차원 유크리드 평방거리를 나타내고 있다. $b^{ij}=\sum_{l=1}^{n}\frac{w_{il}w_{jl}}{\mu_l}$를 대입해서 식 (2.4)를 바꾸어 쓰면,

$$\sum_{l=1}^{n}\sum_{l'=1}^{n}\frac{1}{\mu_l\mu_{l'}}(w_{il}w_{il'}-2w_{il}w_{jl'}+w_{jl}w_{jl'})$$

$$\sum_{k=1}^{n}\sum_{k'=1}^{n}w_{kl}w_{k'l}a_{kk'}$$

$$=\sum_{l=1}^{r}\lambda_l(x_{il}-x_{jl})^2\quad(i,j=1,2,\cdots,n) \tag{2.5}$$

을 얻는다.

좌변에 있어서의 $\sum_{k=1}^{n}\sum_{k'=1}^{n}w_{kl}w_{k'l}a_{kk'}$는 A가 반정치이므로 반드시 0이 아닌 값을 취한다. 식 (2.4) 혹은 (2.5)는 r차원 유크리드 공간에 있어서의 n개의 점 중에서 임의의 두 점 $i,\ j$간의 가중치가 붙은 유크리드 평방거리가, A 및 B의 요소에 의해서 어떻게 표현되는가를 나타내고 있다.

이제 수량화이론 Ⅲ류에 있어서의 고유치전개로부터 얻어지는 제관계를 알아 보기로 한다. 수량화이론 Ⅲ류에 있어서는 식 (2.1) 중의 A, B에 해당하는 것은

$$P_{YX}{P_X}^{-1}P_{XY}\tilde{\boldsymbol{y}}=r_{XY}{}^2P_Y\tilde{\boldsymbol{y}}$$

$$P_{XY}{P_Y}^{-1}P_{YX}\tilde{\boldsymbol{x}}=r_{XY}{}^2P_X\tilde{\boldsymbol{x}}$$ 로부터 각각 행렬

$$P_{YX}{P_X}^{-1}P_{XY}=\left(\sum_{i=1}^{N_1}\frac{p_{ij}p_{ij'}}{p_{i\cdot}}\right) \tag{2.6}$$

$$P_Y=\mathrm{diag}\,(p_{\cdot j})$$

또는

$$P_{XY}{P_Y}^{-1}P_{YX}=\left(\sum_{j=1}^{N_2}\frac{p_{ij}p_{i'j}}{p_{\cdot j}}\right) \tag{2.7}$$

$$P_X=\mathrm{diag}\,(p_{i\cdot})$$

이다. 여기에서도 $N_1\geq N_2$라고 생각한다. 수량화이론 Ⅲ류의 고유방정식을 풀어서 구해지는 제 l고유치를 λ_l이라 하고, 그에 대응하는 고유벡터를 식 (2.6)의 경우

$\mathrm{yl}'=(y_{1l},\cdots,y_{jl},\cdots,y_{N_2l})$,

(2.7)의 경우

$\mathrm{xl}'=(x_{1l},\cdots,x_{il},\cdots,x_{N_1l})$

라고 한다.

식 (2.6)의 경우에는

$a_{kk'}=\sum_{i=1}^{N_1}p_{ik}p_{ik'}/p_{i\cdot}$, $b^{ij}=\delta_{ij}/p_{\cdot j}$이므로 식 (2.3)의 좌변은

$$\sum_{k=1}^{N_1}\sum_{k'=1}^{N_2}\frac{\delta_{ik}}{p_{\cdot k}}\frac{\delta_{ik'}}{p_{\cdot k'}}\sum_{s=1}^{N_1}\frac{p_{sk}p_{sk'}}{p_{s\cdot}}$$

$$=\frac{1}{p_{i\cdot}p_{\cdot j}}\sum_{s=1}^{N_1}\frac{p_{si}p_{sj}}{p_{s\cdot}}$$

이 된다. 따라서

$$\frac{1}{p_{\cdot j}p_{\cdot j'}}\sum_{i=1}^{N_1}\frac{p_{ij}p_{ij'}}{p_{i\cdot}}=\sum_{l=1}^{N_2}\lambda_l y_{jl}y_{j'l}$$

$$(j,\ j'=1,2,\cdots,N_2) \tag{2.8}$$

를 얻는다. 똑같은 방법으로

$$\frac{1}{p_{i\cdot}p_{i'\cdot}}\sum_{j=1}^{N_2}\frac{p_{ij}p_{i'j}}{p_{\cdot j}}=\sum_{l=1}^{N_2}\lambda_l x_{il}x_{i'l}$$

$$(i,\ i'=1,2,\cdots,N_1) \tag{2.9}$$

이 얻어진다.

또한 두 점간의 거리를 나타내는 관계식 (2.4)에 상당하는 것으로서는, 다음과 같은 것을 쉽게 도출할 수 있다. 식 (2.8)로부터

$$\sum_{i=1}^{N_1}\frac{1}{p_{i\cdot}}\left\{\frac{p_{ij}}{p_{\cdot j}}-\frac{p_{i'j}}{p_{\cdot j}}\right\}^2=\sum_{l=1}^{N_2}\lambda_l(y_{jl}-y_{j'l})^2 \tag{2.10}$$

식 (2.9)로부터

$$\sum_{j=1}^{N_2}\frac{1}{p_{\cdot j}}\left\{\frac{p_{ij}}{p_{i\cdot}}-\frac{p_{i'j}}{p_{i'\cdot}}\right\}^2=\sum_{l=1}^{N_2}\lambda_l(x_{il}-x_{i'l})^2 \tag{2.11}$$

여기에서 두 점 i, i'간의 거리를 나타내는 식 (2.11)의 우변에 주목하여 이를 유크리드 거리로 정의하고, 이를 출발점으로 해서 수량화이론 III류를 재고하고자 하는 것이 본연구의 핵심이다. 즉,

$$d^2(i,i')=\sum_{j=1}^{N_2}\frac{1}{p_{\cdot j}}\left\{\frac{p_{ij}}{p_{i\cdot}}-\frac{p_{i'j}}{p_{i'\cdot}}\right\}^2$$

$$=\sum_{j=1}^{N_2}\left\{\frac{p_{ij}}{p_{i\cdot}\sqrt{p_{\cdot j}}}-\frac{p_{i'j}}{p_{i'\cdot}\sqrt{p_{\cdot j}}}\right\}^2 \tag{2.12}$$

식 (2.10), (2.11) 사이에는 행렬의 성질에 의해서 쌍대성(duality)이 성립하므로 여기에서는 식 (2.11)만을 생각하기로 한다.

식 (2.12)에서는 이 점간 거리가 가능한 한 식별될 수 있도록 가까운 것은 가까이, 먼 것은 멀리 공간의 배치를 한다고 하는 것이 목표이다. 이것은,

$$\zeta_{ij}=\frac{p_{ij}}{p_{i\cdot}\sqrt{p_{\cdot j}}} \tag{2.13}$$

로 놓고, 이의 N_2차원 공간 내 N_1개의 점의 분포를 생각해서 이것의 주성분분석을 실시하는 것에 상당한다.

ζ_{ij} $(i=1,2,\cdots,N_1\ ;\ j=1,2,\cdots,N_2)$를 주어진 데이터로 생각하고, 이것의 주성분석을 실시한다. 행렬 $Z=(\zeta_{ij})$라고 놓고 열요소군(列要素群)에 대한 요소의 평균과 요소간의 분산공분산행렬을 만든다.

제 j요소의 평균은

$$\overline{\zeta_j}=\sum_{i=1}^{N_1}p_{i\cdot}\zeta_{ij}=\sum_{i=1}^{N_1}p_{i\cdot}\frac{p_{ij}}{p_{i\cdot}\sqrt{p_{\cdot j}}}=\sqrt{p_{\cdot j}} \tag{2.14}$$

이 된다. 또 이것을 요소로 하는 평균벡터를 $\overline{\zeta}'=(\overline{\zeta_1},\overline{\zeta_2},\cdots,\overline{\zeta_j},\cdots,\overline{\zeta_{N_2}})$로 나타낸다.

다음에 Z의 열요소군 N_2개의 요소간 분산공분산행렬을 만들면,

$$V=(Z-\overline{Z})'P_X(Z-\overline{Z}) \tag{2.15}$$

여기에서

$$Z=P_X{}^{-1}P_{XY}P_Y{}^{-1/2}$$

$$\overline{Z}=1_{N_1}\overline{\zeta}'$$

이 때, V의 요소인 분산 v_{ij}, 공분산 v_{jk}는 다음과 같다.

$$v_{ij}=\sum_{i=1}^{N_1}p_{i\cdot}(\zeta_{ij}-\overline{\zeta_j})^2=\sum_{i=1}^{N_1}p_{i\cdot}\left(\frac{p_{ij}}{p_{i\cdot}\sqrt{p_{\cdot j}}}-\sqrt{p_{\cdot j}}\right)^2$$

$$=\sum_{i=1}^{N_1}\left(\frac{p_{ij}-p_{i\cdot}p_{\cdot j}}{\sqrt{p_{i\cdot}p_{\cdot j}}}\right)^2=\sum_{i=1}^{N_1}\frac{(p_{ij}-p_{i\cdot}p_{\cdot j})^2}{p_{i\cdot}p_{\cdot j}} \tag{2.16}$$

$$v_{jk}=\sum_{i=1}^{N_1}p_{i\cdot}(\zeta_{ij}-\overline{\zeta_j})(\zeta_{ik}-\overline{\zeta_k})$$

$$=\sum_{i=1}^{N_1}p_{i\cdot}\left(\frac{p_{ij}}{p_{i\cdot}\sqrt{p_{\cdot j}}}-\sqrt{p_{\cdot j}}\right)\left(\frac{p_{ik}}{p_{i\cdot}\sqrt{p_{\cdot k}}}-\sqrt{p_{\cdot k}}\right)$$

$$=\sum_{i=1}^{N_1}\left(\frac{p_{ij}-p_{i\cdot}p_{\cdot j}}{\sqrt{p_{i\cdot}p_{\cdot j}}}\right)\left(\frac{p_{ik}-p_{i\cdot}p_{\cdot k}}{\sqrt{p_{i\cdot}p_{\cdot k}}}\right)$$

$$=\sum_{i=1}^{N_1}\frac{p_{ij}p_{ik}}{p_{i\cdot}\sqrt{p_{\cdot j}p_{\cdot k}}}-\sqrt{p_{\cdot j}p_{\cdot k}} \tag{2.17}$$

그런데 행렬 $V=(v_{ij})$에서

$$trV=\sum_{j=1}^{N_2}v_{ij}=\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\frac{(p_{ij}-p_{i\cdot}p_{\cdot j})^2}{p_{i\cdot}p_{\cdot j}}$$

$$=\frac{1}{N}\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\frac{(f_{ij}-\hat{f}_{ij})^2}{\hat{f}_{ij}}=\frac{1}{N}\chi^2 \tag{2.18}$$

여기에서

$$\hat{f}_{ij}=\frac{f_{i\cdot}f_{\cdot j}}{N},\quad f_{i\cdot}=\sum_j f_{ij},\quad f_{\cdot j}=\sum_i f_{ij},$$

$$p_{ij}=\frac{f_{ij}}{N},\quad N=\sum_i\sum_j f_{ij}$$

따라서 $N\,trV=N\sum_{j=1}^{N_2}v_{ij}=\chi^2$는 바로 Pearson의 χ^2통계량이다. 그러므로 식 (2.12)에서 정의한 것을 χ^2거리라고 부르고 있는 것이다(노형진, 1990, pp. 207~211).

2. 수량화이론 Ⅲ류에 대한 재고

전술한 내용을 정리하면 다음과 같다. 출현확률을 요소로 하는 $P_{XY}=(p_{ij})$로부터 만든 $Z=(\zeta_{ij})$를 새로운 데이터표로 생각한다. 이 Z의 분산공분산행렬 V의 주성분분석을 실시하고자 하므로, 결국 행렬 V의 고유치문제

$$V\mathbf{u} = \lambda\mathbf{u}\ ,\ \text{단}\ \mathbf{u}'\mathbf{u} = 1 \qquad (2.19)$$

을 풀면 된다. 식 (2.19)를 요소의 형태로 쓰면

$$\sum_{k=1}^{N_2} v_{jk}u_k=\lambda u_j \quad (j=1,2,\cdots,N_2) \qquad (2.20)$$

위의 식에 식 (2.17)을 대입하면

$$\sum_{k=1}^{N_2}\left\{\sum_{i=1}^{N_1}\frac{p_{ij}p_{ik}}{p_{i\cdot}\sqrt{p_{\cdot j}p_{\cdot k}}}-\sqrt{p_{\cdot j}p_{\cdot k}}\right\}u_k=\lambda u_j$$

$$(j=1,2,\cdots,N_2) \qquad (2.21)$$

여기에서 $\lambda_1=0$에 대한 고유벡터

$\mathrm{u}_1'=(\sqrt{p_{\cdot 1}},\cdots,\sqrt{p_{\cdot k}},\cdots,\sqrt{p_{\cdot N_2}})$는 자명한 해(trvial solution)이다. 즉

$$\sum_{k=1}^{N_2}\left\{\sum_{i=1}^{N_1}\frac{p_{ij}p_{ik}}{p_{i\cdot}\sqrt{p_{\cdot j}p_{\cdot k}}}-\sqrt{p_{\cdot j}p_{\cdot k}}\right\}\sqrt{p_{\cdot k}}$$

$$=\sum_{k=1}^{N_2}\sum_{i=1}^{N_1}\frac{p_{ij}p_{ik}}{p_{i\cdot}\sqrt{p_{\cdot j}}}-\sum_{k=1}^{N_2}\sqrt{p_{\cdot j}p_{\cdot k}}$$

$$=\sum_{i=1}^{N_1}\frac{p_{ij}p_{i\cdot}}{p_{i\cdot}\sqrt{p_{\cdot j}}}-\sqrt{p_{\cdot j}}$$

$$=\frac{1}{\sqrt{p_{\cdot j}}}\sum_{i=1}^{N_1}p_{ij}-\sqrt{p_{\cdot j}}=\sqrt{p_{\cdot j}}-\sqrt{p_{\cdot j}}=0$$

또 직교조건으로부터 $\mathbf{u}_l'\mathbf{u}_1=0 \quad (l\neq 1)$

$\mathbf{u}_l'=(u_1,u_2,\cdots,u_j,\cdots,u_{N_2})$

$\mathbf{u}_1'=(\sqrt{p_{\cdot 1}},\sqrt{p_{\cdot 2}},\cdots,\sqrt{p_{\cdot j}},\cdots,\sqrt{p_{\cdot N_2}})$

$\mathbf{u}_l'\mathbf{u}_1=u_1\sqrt{p_{\cdot 1}}+\cdots+u_j\sqrt{p_{\cdot j}}$

$+\cdots+u_{N_2}\sqrt{p_{\cdot N_2}}=\sum_{j=1}^{N_2}u_j\sqrt{p_{\cdot j}}=0$

이것을 식 (2.21)에 대입하면,

$$\sum_{k=1}^{N_2}\left\{\sum_{i=1}^{N_1}\frac{p_{ij}p_{ik}}{p_{i\cdot}\sqrt{p_{\cdot j}p_{\cdot k}}}-\sqrt{p_{\cdot j}p_{\cdot k}}\right\}u_k$$

$$=\sum_{k=1}^{N_2}\left(\sum_{i=1}^{N_1}\frac{p_{ij}p_{ik}}{p_{i\cdot}\sqrt{p_{\cdot j}p_{\cdot k}}}\right)u_k-\sqrt{p_{\cdot j}}\sum_{k=1}^{N_2}u_k\sqrt{p_{\cdot k}}$$

$$=\sum_{k=1}^{N_2}\left(\sum_{i=1}^{N_1}\frac{p_{ij}p_{ik}}{p_{i\cdot}\sqrt{p_{\cdot j}p_{\cdot k}}}\right)u_k=\lambda u_j$$

따라서 식 (2.21)은 다음과 같이 쓸 수 있다.

$$\sum_{k=1}^{N_2}\left(\sum_{i=1}^{N_1}\frac{p_{ij}p_{ik}}{p_{i\cdot}\sqrt{p_{\cdot j}p_{\cdot k}}}\right)u_k=\lambda u_j \quad (j=1,2,\cdots,N_2) \qquad (2.22)$$

여기에서

$$\frac{p_{ij}p_{ik}}{p_{i\cdot}\sqrt{p_{\cdot j}p_{\cdot k}}}=\frac{p_{ij}}{\sqrt{p_{i\cdot}p_{\cdot j}}}\cdot\frac{p_{ik}}{\sqrt{p_{i\cdot}p_{\cdot k}}}$$

라고 쓸 수 있으므로,

$$q_{ij}=\frac{p_{ij}}{\sqrt{p_{i\cdot}p_{\cdot j}}}\left(=\frac{f_{ij}/N}{\sqrt{f_{i\cdot}/N\times f_{\cdot j}/N}}=\frac{f_{ij}}{\sqrt{f_{i\cdot}f_{\cdot j}}}\right) \qquad (2.23)$$

라고 놓으면, 식 (2.22)는

$$\sum_{k=1}^{N_2}\left(\sum_{i=1}^{N_1}q_{ij}q_{ik}\right)u_k=\lambda u_j \quad (j=1,2,\cdots,N_2) \qquad (2.24)$$

이 된다. $Q=(q_{ij})$로 놓으면,

$$Q=P_X{}^{-1/2}P_{XY}P_Y{}^{-1/2} \qquad (2.25)$$

로 해서 $S=Q'Q$의 고유치문제 $\mathrm{Su}=\lambda\mathrm{u}$가 되어, 고유방정식 $\det(S-\lambda I_{N_2})=0$를 푸는 것으로 된다.

그런데 여기에서

$$S=Q'Q$$

$$=(P_X{}^{-1/2}P_{XY}P_Y{}^{-1/2})'(P_X{}^{-1/2}P_{XY}P_Y{}^{-1/2})$$

$$=P_Y{}^{-1/2}P_{YX}P_X{}^{-1}P_{XY}P_Y{}^{-1/2}$$

으로 된다. 따라서 S의 고유치문제는 바로 수량화이론 Ⅲ류의 기본방정식을 푸는 것과 같은 결과를 가져온다.

이 때의 고유치는

$\lambda_1=1\geq\lambda_2\geq\cdots\geq\lambda_l\geq\cdots\geq\lambda_r\geq 0 \quad (r=\mathrm{rank}\ S)$이 된다. λ_l에 대한 고유벡터를 ul 라 하고 그리고 또한 $\mathrm{U}=(\mathbf{u}_1,\ \mathbf{u}_2,\ \cdots,\ \mathbf{u}_l,\ \cdots,\ \mathbf{u}_r)$라고 놓으면, 주성분득점 즉 행요소 $i(i=1,2,\cdots,N_1)$에 대한 수치(득점)는 다음과 같이 구해진다.

$$\mathrm{F_{N1}}=\mathrm{ZU}=P_X{}^{-1}P_{XY}P_Y{}^{-1/2}U \qquad (2.26)$$

이 $\mathrm{F_{N1}}$의 제 $(i,\ l)$ 요소, 즉 제 i요소에 대한 제 l 주성분의 수치는 다음과 같다.

$$F_l(i)=\sum_{j=1}^{N_2}u_{jl}\zeta_{ij}=\sum_{j=1}^{N_2}u_{jl}\left(\frac{p_{ij}}{p_{i\cdot}\sqrt{p_{\cdot j}}}\right) \qquad (2.27)$$

여기에서 u_{jl}은 $\mathbf{u}_l$의 제j요소이다. 이제 $y_{jl}=u_{jl}/\sqrt{p_{\cdot j}}$로 놓고 이것을 요소로 하는 벡터 $\mathbf{y}^{(l)}$을 만들면,

$$\mathbf{y}^{(l)}=P_Y^{-1/2}\mathbf{u}_l \tag{2.28}$$

이때, $\mathbf{y}^{(l)'}P_Y\mathbf{y}^{(l)}=\mathbf{u}_l'P_Y^{-1/2}P_YP_Y^{-1/2}\mathbf{u}_l=\mathbf{u}_l'\mathbf{u}_l=1$ 이 된다. 또

$$\text{평균}: \sum_{j=1}^{N_2}p_{\cdot j}y_{jl}=\sum_{j=1}^{N_2}\sqrt{p_{\cdot j}}\,u_{jl}=0$$

$$\text{분산}: \sum_{j=1}^{N_2}p_{\cdot j}y_{jl}^2=\sum_{j=1}^{N_2}p_{\cdot j}\frac{u_{jl}^2}{p_{\cdot j}}=\sum_{j=1}^{N_2}u_{jl}^2=0$$

$(\because \mathbf{u}_l'\mathbf{u}_l=1)$

따라서 $Y=(\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \cdots, \mathbf{y}^{(l)}, \cdots, \mathbf{y}^{(r)})$로 놓으면 식 (2.28)은 행렬로 표시하여 다음과 같이 된다.

$$Y=P_Y^{-1/2}U \tag{2.29}$$

그리고 식 (2.26)은

$$F_{N1}=P_X^{-1}P_{XY}P_Y^{-1/2}U=P_X^{-1}P_{XY}Y \tag{2.30}$$

로 쓸 수 있다.

이제 열요소 $j(j=1,2,\cdots,N_2)$에 대한 수치(득점)를 산출해 보기로 한다. 앞에 나왔던 $Q=(q_{ij})$를 이용해서 $S^*=QQ'$로 놓고 이것의 고유치문제를 생각한다. 즉 $S^*\mathbf{w}=\lambda\mathbf{w}$, $\mathbf{w}'\mathbf{w}=1$. 행요소 i에 대한 계산과 똑같은 방법으로 행요소 j에 대한 수치는 다음 식으로 주어진다.

$$F_{N2}=P_Y^{-1}P_{YX}P_X^{-1/2}W \tag{2.31}$$

여기에서, $W=(\mathbf{w}_l, \cdots, \mathbf{w}_l, \cdots, \mathbf{w}_r)$

$r\le\min\{N_1-1, N_2-1\}$

또한 식 (2.28)에서와 마찬가지로

$$\mathbf{x}^{(l)}=P_X^{-1/2}\mathbf{w}_l \tag{2.32}$$

라고 변환하면, $X=P_X^{-1/2}W$이 되어 식 (2.31)은

$$F_{N2}=P_Y^{-1}P_{YX}P_X^{-1/2}W=P_Y^{-1}P_{YX}X \tag{2.33}$$

이 된다.

이제 행렬의 성질로부터 $S=Q'Q$의 고유치 λ, 고유벡터 $\mathbf{u}$, $S^*=QQ'$의 고유치 λ, 고유벡터 $\mathbf{w}$에 대해서 $\mathbf{u}'\mathbf{u}=1$, $\mathbf{w}'\mathbf{w}=1$에 주의하면 다음의 관계가 성립한다.

$$\mathbf{w}=\frac{1}{\sqrt{\lambda}}Q\mathbf{u},\ \mathbf{u}=\frac{1}{\sqrt{\lambda}}Q'\mathbf{w} \tag{2.34}$$

이것을 고려하면 식 (2.32)는

$$\mathbf{x}^{(l)}=P_X^{-1/2}\mathbf{w}_l$$

$$=P_X^{-1/2}\left(\frac{1}{\sqrt{\lambda_l}}Q\mathbf{u}_l\right)$$

$$=P_X^{-1/2}\cdot\frac{1}{\sqrt{\lambda_l}}P_X^{-1/2}P_{XY}P_X^{-1/2}\mathbf{u}_l$$

$$=\frac{1}{\sqrt{\lambda_l}}P_X^{-1}P_{XY}P_X^{-1/2}\mathbf{u}_l=\frac{1}{\sqrt{\lambda_l}}P_X^{-1}P_{XY}\mathbf{y}^{(l)}$$

따라서 다음의 관계를 얻는다.

$$\mathbf{x}^{(l)}=\frac{1}{\sqrt{\lambda_l}}P_X^{-1}P_{XY}\mathbf{y}^{(l)} \tag{2.35}$$

또한 마찬가지로,

$$\mathbf{y}^{(l)}=P_Y^{-1/2}\mathbf{u}_l$$

$$=P_Y^{-1/2}\left(\frac{1}{\sqrt{\lambda_l}}Q'\mathbf{w}_l\right)=\frac{1}{\sqrt{\lambda_l}}P_Y^{-1}P_{YX}P_X^{-1/2}\mathbf{w}_l$$

따라서 다음의 관계를 얻는다.

$$\mathbf{y}^{(l)}=\frac{1}{\sqrt{\lambda_l}}P_Y^{-1}P_{YX}\mathbf{x}^{(l)} \tag{2.36}$$

Ⅲ. 대응분석의 실제

대응분석(correspondence analysis)은 앙케트의 질문에 대한 회답의 패턴에 주목하여, 패턴이 비슷한 회답자(개인 또는 집단)와 비슷하지 않은 회답자를 분류하기 위한 기법이다. 이 기법은 질문항목끼리의 관계도 동시에 분석할 수 있다.

대응분석은 다음과 같은 세 가지 타입의 데이터표에 적용할 수 있다.

(1) 분할표

(2) 아이템 카테고리형 데이터표

(3) (0, 1)형 데이터표

대응분석은 다음과 같이 두 가지의 기법으로 나눌 수 있다.

(A) 단순대응분석(일반적으로는 이것을 대응분석이라고 부른다)

(B) 다중대응분석

분할표를 처리할 때는 (A)의 단순대응분석을 이용한다. 아이템·카테고리형 데이터표나 (0, 1)형 데이터표를 처리할 때는 (B)의 다중대응분석을 이용한다.

여기에서 간단한 예를 통하여 대응분석의 원리와 분석방법을 살펴보기로 한다. 어떤 상품에 대한 만족도를 5단계 평가로 400명에게 질문했다. 그 회답결과를 학생, 회사원, 주부, 사업가의 네 그룹으로 나누어 집계하여 다음과 같은 분할표로 정리했다.

	불만	약간 불만	어느쪽도 아님	약간 만족	만족
학 생	5	14	40	28	13
회사원	30	40	15	10	5
주 부	5	10	15	30	40
사업가	10	18	25	40	7

SPSS에서 대응분석을 실시하려면 SPSS Categories라고 하는 옵션 제품이 필요하다. SPSS Categories에는 다음과 같은 분석기법이 포함되어 있다.

① 대응분석
② 동질성분석
③ 비선형 주성분분석
④ 비선형 정준상관분석

이 분할표를 대응분석으로 분석하면 다음과 같은 결과를 얻게 된다.

(1) 차원에 따른 정보 제공력

Dimension	Singular Value	Inertia	Proportion Explained	Cumulative Proportion
1	.18982	.23993	.668	.668
2	.31319	.09809	.273	.941
3	.14548	.02117	.059	1.000
Total		.35918	1.000	1.000

위의 표는 각 차원에 따른 정보 제공력을 나타낸다. 2차원으로 출력했을 경우에 94.1%의 정보를 나타내며, 이것으로 5.9%(=100% - 94.1%)의 정보손실이 발생했음을 알 수 있다. 3차원으로 나타낼 경우에는 원래의 자료가 가지고 있는 정보를 100% 제공할 수 있다는 것을 알 수 있다.

(2) 행 요인의 각 차원에 대한 공헌도 및 차원의 각 행 요인에 대한 공헌도

Contribution of row points to the inertia of each dimension:

그룹	Marginal Profile	Dim 1	2
1 학생	.250	.039	.281
2 회사원	.250	.631	.105
3 주부	.250	.330	.419
4 사업가	.250	.001	.195
		1.000	1.000

Contribution of dimensions to the inertia of each row point:

그룹	Marginal Profile	Dim 1	2	Total
1 학생	.250	.201	.600	.802
2 회사원	.250	.934	.064	.998
3 주부	.250	.659	.341	1.000
4 사업가	.250	.004	.617	.621

(3) 열 요인의 각 차원에 대한 공헌도 및 차원의 각 열 요인에 대한 공헌도

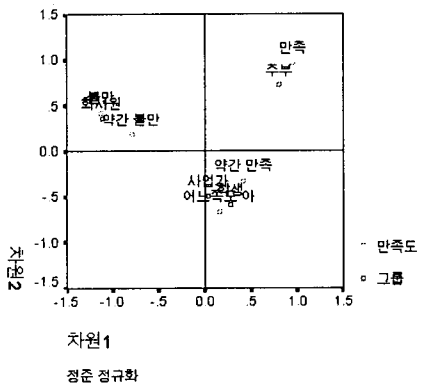
Contribution of column points to the inertia of each dimension:

만족도	Marginal Profile	Dim 1	2
1 불만	.125	.325	.071
2 약간 불만	.205	.265	.021
3 어느쪽도 아님	.238	.012	.335
4 약간 만족	.270	.095	.092
5 만족	.163	.303	.482
		1.000	1.000

Contribution of dimensions to the inertia of each column point:

만족도	Marginal Profile	Dim 1	2	Total
1 불만	.125	.917	.082	.999
2 약간 불만	.205	.969	.031	1.000
3 어느쪽도 아님	.238	.063	.745	.808
4 약간 만족	.270	.528	.207	.735
5 만족	.163	.600	.390	.991

(4) 행 및 열 점수에 대한 산포도



행 및 열 점수의 그래프를 보면, 주부는 만족도가 높고 회사원은 불만도가 높은 것을 알 수 있다. 사업가와 학생은 비슷한 만족도의 평가를 하고 있다.

'어느쪽도 아님' 이라고 하는 중간회답은 '약간 만족'에 가까운 위치에 있다.

Ⅳ. 결론

본 연구는 이미 개발되어 여러 분야에 응용되어 왔던 수량화이론 Ⅲ류를 대응분석의 원리에 의거하여 재정식화했으며, 몇 가지 문제점을 분명히 하는 데 목적이 있었다.

그러나 본 연구에서는 범주형 데이터 행렬에서의 결측치(missing data)의 처리와 분할표의 독립성검정에 대해서는 연구가 이루어지지 않았다. 종래의 수량화이론 Ⅲ류에서는 행요소와 열요소의 분포에 선험적으로 분포함수를 가정하지 않고, 얻어진 데이터만으로부터 출발하는 기술적인 입장을 취하며 취급하는 데이터도 주로 반응형 데이터만을 대상으로 해왔기 때문에 위와 같은 연구가 거의 이루어지지 않았다. 그러나 분할표, 도수표 등의 정리된 데이터 행렬을 주로 대상으로 하는 대응분석이나 보다 다양한 데이터 행렬을 분석대상으로 하는 쌍대척도법에서는 이에 관한 연구가 꾸준히 진행되어 오고 있다.

본 연구에서 제안하고 있는 재정식화된 수량화이론 Ⅲ류에서도 반응형 데이터 뿐만 아니라 정리된 다양한 데이터 행렬을 분석대상으로 하고 있기 때문에, 앞으로 이에 대한 연구가 뒤따라야 할 것으로 생각된다. 그리고 분석대상의 데이터도 심리학 분야에서 주로 다루어지고 있는 일대비교 데이터(paired comparison data), 반응범주에 순위가 있는 경우의 계차범주 데이터(successive categorical data), 또는 여러 형태의 다차원표(multidimensional table) 등에 대한 연구도 병행되어져야 할 것이다.

참고문헌

〔1〕 노형진, 다변량해석-질적 데이터의 수량화-, 석정, 1990.

〔2〕 노형진, 한글 SPSSWIN에 의한 조사방법 및 통계분석, 형설출판사, 1999.

〔3〕 노형진, 한글 SPSSWIN에 의한 알기 쉬운 다변량분석, 형설출판사, 1999.

〔4〕 노형진, 한글 SPSSWIN에 의한 다변량 데이터의 통계분석, 석정, 1999.

〔5〕 岩坪秀一, n-way質的データの多變量解析手法の研究, 電子技術總合研究所研究報告, 第801號, 1979年 10月.

〔6〕 岩坪秀一, 數量化法の基礎, 朝倉書店, 1987.

〔7〕 林知己夫, 數量化の方法, 東洋經濟新報社, 1974a.

〔8〕 林知己夫, データ解析の方法, 東洋經濟新報社, 1974b.

〔9〕 駒澤勉·橋口捷久, パソコン數量化分析, 朝倉書店, 1988

〔10〕 Benzecri, J. P., L'analyse des donnees ; T. 2, l'analyse des correspondences, Paris, Dunod, 1973.

〔11〕 Benzecri, J. P., Sur l'analyse des tableaux binaires associes a une correspondences multiple, Les Cahiers de l'analyse des Donnees, 2, 1977, pp. 55~71.

〔12〕 Gower, J. C., Some distance properties of latent root vector method used in multivariate analysis, Biometrika, 53, 1966, pp. 325~338.

〔13〕 Greenacre, M. J., Theory and applications of correspondence analysis, London, Academic, Press, 1984.

〔14〕 Tenenhaus, M., and F. W. Young, An analysis and synthesis of multiple

correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data, Psychometrika, 1985, 50, No. 1, pp. 91~119.

저 자 소 개

노형진

서울대학교 공과대학 졸업(공학사)
고려대학교 대학원 수료(경영학 박사), 일본 쓰쿠바 대학 대학원 수료(경영공학 박사 과정)
일본 동경대학 객원 교수
현재, 경기대학교 경영학부 교수
관심분야 : 품질경영, 다변량 분석 6시그마, Single PPM 품질혁신