응용통계연구 (2011)
24(3), 559–566
DOI: 10.5351/KJAS.2011.24.3.559

편정준상관 행렬도

염아림[1] · 최용석[2]

[1]부산대학교 통계학과, [2]부산대학교 통계학과

(2011년 4월 접수, 2011년 5월 채택)

요 약

행렬도는 이원표 자료행렬의 행과 열을 탐색하기에 유용한 그래프적 방법이다. 특히, 정준상관 행렬도는 정준상관분석의 결과를 이용하여 두 변수군과 개체간의 관계를 기하적으로 살펴볼 수 있다. 그 반면에 자료의 성격에 따라 세 개 이상의 변수군이 존재하는 경우에는 정준상관분석의 개념에서 확장한 일반화 정준상관분석을 이용하여 일반화 정준상관 행렬도를 고려할 수 있다. 그러나 자료의 성격에 따라 두 변수군 외에 이들 두 변수군에 선형적 영향을 미치는 공변량변수로 이루어진 다른 한 변수군이 존재하는 경우에, 일반화 정준상관 행렬도를 적용한다면 공변량변수군의 영향력 때문에 주 관심인 두 변수군에 대하여 잘못 해석할 수 있다. 따라서 본 연구에서는 Rao (1969)의 공변량변수군의 영향력을 제거한 편정준상관분석을 살펴보고, 이를 기하적으로 해석하기 위한 편정준상관 행렬도를 제안한다.

주요용어: 행렬도, 공변량변수군, 편정준상관분석, 편정준상관 행렬도.

1. 서론

행렬도(biplot)는 Gabriel (1971)에 의해서 주로 개발되었고, 국내에서는 Choi (1991)가 저항성 행렬도(resistant biplot)를 연구하면서 본격적으로 행렬도를 소개하였으며, 허명회 (1993, 5장)가 처음으로 biplot을 행렬도(行列圖)라 불렀다. 행렬도는 이원표 자료행렬(two-way data matrix)의 행과 열을 그래프에 동시에 나타내어 복잡한 다변량 분석의 결과를 보다 쉽게 파악할 수 있기 때문에 최근 여러 분야에서 행렬도에 대한 많은 연구와 응용을 하고 있다.

그중에서도 정준상관 행렬도(canonical correlation biplot)는 정준상관분석(canonical correlation analysis)을 통해 두 변수 집단에 의해서 측정된 다변량 자료의 변수군 사이의 관계와 개체들의 관계를 탐색하기 위한 2차원 그림이다. 일반적으로 정준상관분석은 Hotelling (1936)에 의해 개발되었고, 국내에서는 Park과 Huh (1996)가 정준상관분석에서 수량화 방법(quantification method) 관점을 이용하여 정준상관 행렬도를 제안하였다. 최근 이 행렬도를 응용한 연구를 살펴보면, 최태훈과 최용석 (2008)은 2006년도 KLPGA 선수 중 상금 순위 상위 50명을 대상으로 정준상관 행렬도를 통해 기술요인 변수군과 경기성적요인 변수군 간의 관련성과 더불어 군집분석의 활용을 가미하였다. 최태훈 등 (2009)은 테니스 그랜드슬램대회의 선수특성요인과 경기요인에 대한 정준상관 행렬도에서 프로크러스티즈 분석을 통하여 행렬도의 형상 비교를 하였고, 더 나아가 최태훈과 최용석 (2010)은 2004년 대한테니스협회(KTA)에 등록된 랭킹 100위권 이내의 선수 50명을 대상으로 세 변수군인 체격요인변수군, 체력요인변수군 그리고 기초기술요인변수군의 상호 연관성을 살펴보기 위해 일반화 정준상관 행렬

이 논문은 부산대학교 자유과제 학술연구비(2년)에 의하여 연구되었음.
[2]교신저자: (609-735) 부산시 금정구 장전동 산30, 부산대학교 통계학과, 교수. E-mail: yschoi@pusan.ac.kr

도(generalized canonical correlation biplot)를 활용하였다. 홍현욱 등 (2010)은 결측값이 있는 정준상관 행렬도의 형상변동에 관한 연구를 하였다.

특히, 최태훈과 최용석 (2010)처럼 세 개 이상의 변수군이 존재하는 경우에는 정준상관분석의 개념에서 확장한 Rao (1969)의 일반화 정준상관분석을 이용하여 일반화 정준상관 행렬도를 고려할 수 있다. 그 반면에 자료의 성격에 따라 두 변수군 외에 이들 두 변수군에 선형적 영향을 미치는 공변량변수(covariate variables)로 이루어진 다른 한 변수군이 존재하는 경우에, 일반화 정준상관 행렬도를 적용한다면 공변량변수군의 영향력 때문에 주 관심인 두 변수군에 대하여 잘못 해석할 수 있다. 이 경우 공변량변수군의 선형적 영향을 제거한 두 변수군에 대한 편(partial)정준상관분석을 이용해야 한다. 이 분석 또한 Rao (1969)가 정준상관분석의 개념을 응용한 것으로, 본 연구에서는 편정준상관분석을 살펴보고 이를 위한 편정준상관 행렬도를 제안하려 한다. 2절에서는 정준상관 행렬도 및 편정준상관 행렬도의 대수적인 면을 설명하고, 3절에서는 활용 사례를 제시하려 한다.

2. 편정준상관 행렬도

2.1. 정준상관 행렬도 및 기하적 해석

2.1절에서는 최용석 (2006)을 참고로 하여 정준상관 행렬도 및 기하적 해석에 대하여 정리하고, 2.2절에서는 Timm (2002)을 참고로 하여 편정준상관분석의 기초 이론을 요약하고, 이를 위한 시각적 도구인 편정준상관 행렬도를 제안하려한다.

일반적으로 정준상관분석은 두 변수군 사이의 관계를 분석하는 다변량 기법이다. 이 기법은 두 변수군의 선형결합(linear combination)간의 상관관계를 가장 크게 만드는 알고리즘을 통해 이루어진다. p개의 변수와 q개의 변수로 이루어진 두 변수군 $\mathbf{x} = (x_1, \ldots, x_p)'$와 $\mathbf{y} = (y_1, \ldots, y_q)'$는 각각 평균 $\boldsymbol{\mu}_{\mathbf{x}} = (\mu_{x_1}, \ldots, \mu_{x_p})'$와 $\boldsymbol{\mu}_{\mathbf{y}} = (\mu_{y_1}, \ldots, \mu_{y_q})'$를 가지며 공분산행렬 Σ_{xx}, Σ_{yy}, $\Sigma_{xy} = \Sigma'_{yx}$을 가지는 확률벡터이다. 이들에 의해 측정된 n명의 자료에서 표본공분산행렬은 식 (2.1)과 같다.

$$\mathbf{S} = \begin{pmatrix} \mathbf{S}_{xx} & \mathbf{S}_{xy} \\ \mathbf{S}_{yx} & \mathbf{S}_{yy} \end{pmatrix}. \tag{2.1}$$

임의의 계수벡터 $\mathbf{u}$와 $\mathbf{v}$에 대한 두 변수군 각각의 선형결합은 다음과 같고,

$$\hat{x} = u_1 x_1 + \cdots + u_1 x_p = \mathbf{u}'\mathbf{x}, \qquad \hat{y} = v_1 y_1 + \cdots + v_1 y_q = \mathbf{v}'\mathbf{y}. \tag{2.2}$$

식 (2.2)의 두 선형결합 $\hat{x}$와 $\hat{y}$의 상관은

$$\hat{\rho}_{\hat{x}\hat{y}} = \frac{\sum_{i=1}^{n} \hat{x}_i \hat{y}_i}{\sqrt{\sum_{i=1}^{n} \hat{x}_i^2} \sqrt{\sum_{i=1}^{n} \hat{y}_i^2}} = \frac{\mathbf{u}'\mathbf{S}_{xy}\mathbf{v}}{\sqrt{\mathbf{u}'\mathbf{S}_{xx}\mathbf{u}}\sqrt{\mathbf{v}'\mathbf{S}_{yy}\mathbf{v}}} \tag{2.3}$$

이다. 여기서 계수벡터 $\mathbf{u}$와 $\mathbf{v}$는 식 (2.3)의 두 선형결합의 상관을 최대화하는 알고리즘을 통하여 구할 수 있다. 이 알고리즘은 $\hat{x}$와 $\hat{y}$의 분산이 1인 제약 조건을 $\mathbf{u}'\mathbf{S}_{xx}\mathbf{u} = 1$과 $\mathbf{v}'\mathbf{S}_{yy}\mathbf{v} = 1$을 두고 $\mathbf{u}'\mathbf{S}_{xy}\mathbf{v}$를 최대화하는 계수벡터 $\mathbf{u}$와 $\mathbf{v}$를 찾는 것과 동일하다. 이는 라그랑즈승수법(Lagrange multiplier method)을 이용하면 고유체계 문제로 유도하여 풀 수 있다. 또한 이 알고리즘은 정준계수벡터와 정준상관을 대수적으로 한꺼번에 제공하는 비정칙값분해(singular value decomposition)

$$\mathbf{S}_{xx}^{-\frac{1}{2}}\mathbf{S}_{xy}\mathbf{S}_{yy}^{-\frac{1}{2}} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}'$$

를 이용하면 간편하게 구할 수 있다. 여기서 $r\,(\leq \min(p,q))$은 $\mathbf{S}_{xx}^{-1/2}\mathbf{S}_{xy}\mathbf{S}_{yy}^{-1/2}$의 계수(rank)이고 $\mathbf{U}=(\mathbf{u}_1,\ldots,\mathbf{u}_r)$와 $\mathbf{V}=(\mathbf{v}_1,\ldots,\mathbf{v}_r)$은 크기가 각각 $p\times r$과 $q\times r$인 정준계수벡터로 이루어진 직교행렬이며, 대각행렬 $\mathbf{\Lambda}=\mathrm{daig}(\lambda_1,\ldots,\lambda_r)$는 $\lambda_1\geq\cdots\geq\lambda_r>0$의 관계를 갖는 비정칙값을 대각원소로 하고 있다. 그리고 이 비정칙값이 정준상관에 해당한다. 이를 통하여 i번째 정준상관계수벡터를 구하면 각각 식 (2.4)와 같고,

$$\boldsymbol{\alpha}_i=\mathbf{S}_{xx}^{-\frac{1}{2}}\mathbf{u}_i,\qquad \boldsymbol{\beta}_i=\mathbf{S}_{yy}^{-\frac{1}{2}}\mathbf{v}_i,\quad i=1,\ldots,r. \tag{2.4}$$

이들 정준상관계수벡터에 의해서 구성된 정준상관계수행렬을 $\mathbf{A}=(\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_r)$와 $\mathbf{B}=(\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_r)$로 정의하자.

따라서 두 변수군에 대해 n명을 측정한 크기가 각각 $n\times p$와 $n\times q$인 자료행렬 $\mathbf{X}$와 $\mathbf{Y}$에 대한 정준상관 행렬도의 행 좌표행렬과 열 좌표행렬은 각각 식 (2.5)와 식 (2.6)에 주어져있다.

$$\mathbf{R}_{\mathbf{X}}=\mathbf{X}\mathbf{A}\mathbf{\Lambda},\qquad \mathbf{C}_{\mathbf{X}}=\mathbf{A}\mathbf{\Lambda}. \tag{2.5}$$

$$\mathbf{R}_{\mathbf{Y}}=\mathbf{Y}\mathbf{B}\mathbf{\Lambda},\qquad \mathbf{C}_{\mathbf{Y}}=\mathbf{B}\mathbf{\Lambda}. \tag{2.6}$$

s차원의 정준상관 행렬도는 식 (2.5)와 식 (2.6)의 행렬에서 처음 s개의 열을 고려한 부행렬로 이루어지며 이 s차원의 정준상관 행렬도의 근사적합도는 다음과 같다.

$$\left(\sum_{k=1}^{s}\lambda_k^2\Big\langle\sum_{k=1}^{r}\lambda_k^2\right)\times 100\%.$$

더불어 정준상관 행렬도의 기하적 성질을 살펴보자. 이는 일반적으로 행렬도에서 공통적으로 적용되는 성질이기도 하다. 정준상관 행렬도에서 각 축을 기준으로 서로 반대편에 있는 좌표점은 서로 다른 집단이라 볼 수 있다. 그리고 행 좌표점 사이의 거리는 유클리드거리로 가까이 있다면 이와 관련된 개체들이 비슷한 경향을 갖는 집단임을 나타내며, 변수들을 나타내는 열 좌표점인 두 벡터 $\mathbf{c}_k$와 $\mathbf{c}_j$의 사이각을 θ_{kj}라 하고 각 벡터의 길이(norm)을 $\|\mathbf{c}_k\|$와 $\|\mathbf{c}_j\|$라 하면

$$\cos\theta_{kj}=\frac{\mathbf{c}_k'\mathbf{c}_j}{\|\mathbf{c}_k\|\,\|\mathbf{c}_j\|} \tag{2.7}$$

이다. 식 (2.7)의 코사인 값은 실제로 두 벡터 $\mathbf{c}_k$와 $\mathbf{c}_k$ 간의 상관관계를 근사적으로 나타낸다. 또한 행과 열 좌표점의 상대적인 위치는 특정한 열 좌표점에 대해서 어떤 행들이 큰 값, 평균값, 작은 값을 갖는지를 나타낸다.

2.2. 편정준상관 행렬도

2.1절에서 p개의 변수와 q개의 변수로 이루어진 두 변수군 $\mathbf{x}=(x_1,\ldots,x_p)'$와 $\mathbf{y}=(y_1,\ldots,y_q)'$에 대한 정준상관분석과 행렬도에 대해서 살펴보았다. 그러나 가끔은 이들 두 변수군에 영향을 미치는 m개의 공변량변수로 이루어진 한 변수군 $\mathbf{z}=(z_1,\ldots,z_m)'$가 추가되어 있는 자료를 접하곤 한다. 이 경우를 공변량변수군이라 하고 이것의 선형적 영향을 제거한 두 변수군 $\mathbf{x}$와 $\mathbf{y}$에 대한 편정준상관분석을 살펴보고 이와 관련된 편정준상관 행렬도를 제안하려한다.

세 변수군은 각각 평균벡터로 $\boldsymbol{\mu}_{\mathbf{x}}=(\mu_{x_1},\ldots,\mu_{x_p})'$와 $\boldsymbol{\mu}_{\mathbf{y}}=(\mu_{y_1},\ldots,\mu_{y_q})'$ 그리고 $\boldsymbol{\mu}_{\mathbf{z}}=(\mu_{z_1},\ldots,\mu_{z_m})'$를 가지며 공분산행렬 Σ_{xx}, Σ_{yy}, Σ_{zz}, $\Sigma_{xy}=\Sigma_{yx}^{'}$, $\Sigma_{xz}=\Sigma_{zx}^{'}$, $\Sigma_{yz}=\Sigma_{zy}^{'}$을 가지는 확률벡터이

다. 이들에 의해 측정된 n명의 자료에서 표본공분산행렬은 식 (2.8)과 같다.

$$\mathbf{S} = \begin{pmatrix} \mathbf{S}_{xx} & \mathbf{S}_{xy} & \mathbf{S}_{xz} \\ \mathbf{S}_{yx} & \mathbf{S}_{yy} & \mathbf{S}_{yz} \\ \mathbf{S}_{zx} & \mathbf{S}_{zy} & \mathbf{S}_{zz} \end{pmatrix}. \tag{2.8}$$

이 경우 한 변수군 $\mathbf{z} = (z_1, \ldots, z_m)'$가 공변량변수군이므로 두 변수군에 대한 정준상관분석을 실시하는 것보다 식 (2.8)로부터 공변량변수군 $\mathbf{z}$의 효과를 제거한 식 (2.9)의 조건부 표본공분산행렬 $\widetilde{\mathbf{S}}$를 이용하여 정준상관분석을 하는 것이 바람직하다.

$$\begin{aligned} \widetilde{\mathbf{S}} &= \begin{pmatrix} \mathbf{S}_{xx|z} & \mathbf{S}_{xy|z} \\ \mathbf{S}_{yx|z} & \mathbf{S}_{yy|z} \end{pmatrix}, \\ &= \begin{pmatrix} \mathbf{S}_{xx} - \mathbf{S}_{xz}\mathbf{S}_{zz}^{-1}\mathbf{S}_{zx} & \mathbf{S}_{xy} - \mathbf{S}_{xz}\mathbf{S}_{zz}^{-1}\mathbf{S}_{zy} \\ \mathbf{S}_{yx} - \mathbf{S}_{yz}\mathbf{S}_{zz}^{-1}\mathbf{S}_{zx} & \mathbf{S}_{yy} - \mathbf{S}_{yz}\mathbf{S}_{zz}^{-1}\mathbf{S}_{zy} \end{pmatrix}. \end{aligned} \tag{2.9}$$

이를 편정준상관분석이라 하며, 이를 위한 알고리즘은 2절의 정준상관분석과 유사하고 다음의 두 고유체계로 이루어진다.

$$\begin{aligned} \left\{\mathbf{S}_{xx|z}^{-1}\mathbf{S}_{xy|z}\mathbf{S}_{yy|z}^{-1}\mathbf{S}_{yx|z} - \tilde{\lambda}^2\mathbf{I}\right\}\tilde{\mathbf{u}} &= 0, \\ \left\{\mathbf{S}_{yy|z}^{-1}\mathbf{S}_{yx|z}\mathbf{S}_{xx|z}^{-1}\mathbf{S}_{xy|z} - \tilde{\lambda}^2\mathbf{I}\right\}\tilde{\mathbf{v}} &= 0. \end{aligned} \tag{2.10}$$

식 (2.10)에서 고유값 $\tilde{\lambda}^2$은 공통으로 편정준상관의 제곱이다. 따라서 고유체계의 성질에 따라 k번째 편정준상관의 제곱인 $\tilde{\lambda}_k^2$에 대응하는 고유벡터 $\tilde{\mathbf{u}}_k$와 $\tilde{\mathbf{v}}_k$를 얻게 된다. 이들 고유벡터를 편정준계수벡터라 한다. 이들은 비정칙값분해

$$\mathbf{S}_{xx|z}^{-\frac{1}{2}}\mathbf{S}_{xy|z}\mathbf{S}_{yy|z}^{-\frac{1}{2}} = \tilde{\mathbf{U}}\tilde{\mathbf{\Lambda}}\tilde{\mathbf{V}}' \tag{2.11}$$

를 이용하면 대수적으로 쉽게 제공된다. 식 (2.11)에서 $r(\leq \min(p,q))$은 $\mathbf{S}_{xx|z}^{-1/2}\mathbf{S}_{xy|z}\mathbf{S}_{yy|z}^{-1/2}$의 계수(rank)이고 $\tilde{\mathbf{U}} = (\tilde{\mathbf{u}}_1, \ldots, \tilde{\mathbf{u}}_r)$와 $\tilde{\mathbf{V}} = (\tilde{\mathbf{v}}_1, \ldots, \tilde{\mathbf{v}}_r)$은 크기가 각각 $p\times r$과 $q\times r$인 정준계수벡터로 이루어진 직교행렬이며, 대각행렬 $\tilde{\mathbf{\Lambda}} = \mathrm{daig}(\tilde{\lambda}_1, \ldots, \tilde{\lambda}_r)$는 $\tilde{\lambda}_1 \geq \cdots \geq \tilde{\lambda}_r > 0$의 관계를 갖는 비정칙값을 대각원소로 하고 있다. 그리고 이 비정칙값이 편정준상관에 해당한다. 따라서 두 변수군 $\mathbf{x} = (x_1, \ldots, x_p)'$와 $\mathbf{y} = (y_1, \ldots, y_q)'$에 대하여 측정한 n명의 자료행렬을 각각 크기가 $n \times p$와 $n \times q$인 $\mathbf{X}$와 $\mathbf{Y}$라 하고 이들은 중심화되어 있다고 하면, i번째 편정준상관계수벡터는 각각 식 (2.12)와 같고,

$$\tilde{\boldsymbol{\alpha}}_i = \mathbf{S}_{xx|z}^{-\frac{1}{2}}\tilde{\mathbf{u}}_i, \qquad \tilde{\boldsymbol{\beta}}_i = \mathbf{S}_{yy|z}^{-\frac{1}{2}}\tilde{\mathbf{v}}_i, \quad i = 1, \ldots, r. \tag{2.12}$$

이들로부터 구성된 편정준상관계수행렬은 $\tilde{\mathbf{A}} = (\tilde{\boldsymbol{\alpha}}_1, \ldots, \tilde{\boldsymbol{\alpha}}_r)$이고 $\tilde{\mathbf{B}} = (\tilde{\boldsymbol{\beta}}_1, \ldots, \tilde{\boldsymbol{\beta}}_r)$가 된다. 이들에 의해서 자료행렬 $\mathbf{X}$에 대한 편정준상관 행렬도의 행 좌표행렬과 열 좌표행렬은

$$\tilde{\mathbf{R}}_{\mathbf{X}} = \mathbf{X}\tilde{\mathbf{A}}\tilde{\mathbf{\Lambda}}, \qquad \tilde{\mathbf{C}}_{\mathbf{X}} = \tilde{\mathbf{A}}\tilde{\mathbf{\Lambda}} \tag{2.13}$$

으로, 자료행렬 $\mathbf{Y}$에 대한 행 좌표행렬과 열 좌표행렬은

$$\tilde{\mathbf{R}}_{\mathbf{Y}} = \mathbf{Y}\tilde{\mathbf{B}}\tilde{\mathbf{\Lambda}}, \qquad \tilde{\mathbf{C}}_{\mathbf{Y}} = \tilde{\mathbf{B}}\tilde{\mathbf{\Lambda}} \tag{2.14}$$

으로 정의하며, 이들에 의해 나타나는 행렬도를 편정준상관 행렬도라 하겠다.

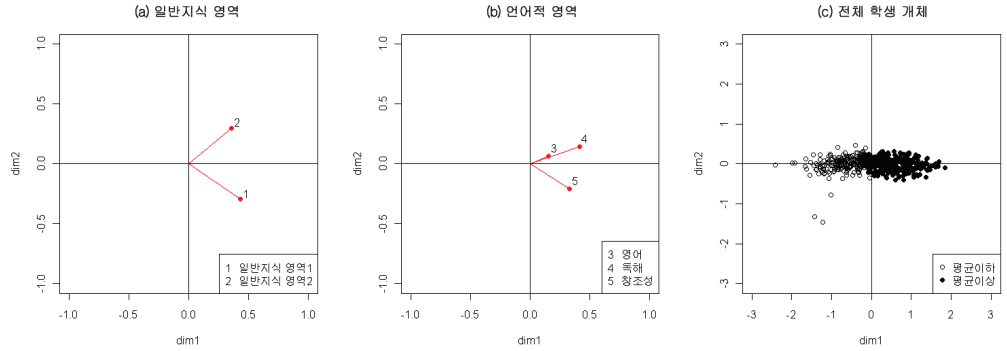

그림 3.1. 재능 자료의 정준상관 행렬도

s차원의 편정준상관 행렬도는 식 (2.13)와 식 (2.14)의 행렬에서 처음 s개 열을 고려한 부행렬로 이루어지며 이 s차원의 편정준상관 행렬도의 근사적합도는 다음과 같다.

$$\left(\sum_{k=1}^{s}\tilde{\lambda}_k^2 \Big\backslash \sum_{k=1}^{r}\tilde{\lambda}_k^2\right)\times 100\%.$$

편정준상관 행렬도는 공변량변수 $\mathbf{Z}$를 독립변수로 하고 $\mathbf{X}$와 $\mathbf{Y}$를 종속변수로 하여, 각각 회귀분석한 후 얻어지는 잔차행렬(residual matrix) $\mathbf{E_X}$와 $\mathbf{E_Y}$에 대한 정준상관 행렬도를 구하는 것과 동일한 결과를 얻는다.

3. 활용 사례

Cooley와 Lohnes (1971)의 재능(talent) 자료는 미국 전역의 고등학교 3학년 학생 505명을 대상으로 일반지식, 언어, 비언어, 흥미도의 4개 영역을 측정한 자료이다. 일반지식 영역은 일반지식 영역1(information test, part1), 일반지식 영역2(information test, part1)의 2개 항목을 측정하였으며, 언어적 영역은 영어(English test), 독해(reading comprehension test), 창조성(creativity test)의 3개 항목을, 비언어적 영역은 기계적 추론(mechanical reasoning test), 추상화 추론(abstract reasoning test), 수학(Mathematics test)의 3개 항목을 측정하였다. 그리고 흥미도는 사교성(sociability inventory), 자연과학 흥미도(physical science interest inventory), 사무 흥미도(office work interest inventory)의 3개 항목을 검사한 점수이다. 이들은 $\mathbf{X}=\{$일반지식 영역1, 일반지식 영역2$\}$, $\mathbf{Y}=\{$영어, 독해, 창조성$\}$, $\mathbf{Y}=\{$기계적 추론, 추상화 추론, 수학, 사교성, 자연과학 흥미도, 사무 흥미도$\}$의 세 부분으로 나눌 수 있다.

재능 자료의 일반지식 영역($\mathbf{X}$)과 언어적 영역($\mathbf{Y}$)에 대한 정준상관 행렬도를 통하여 두 변수군 간에 관계를 살펴볼 수도 있지만, 일반지식 영역과 언어적 영역의 두 변수군에 영향을 줄 수 있는 비언어적 영역과 흥미도($\mathbf{Z}$)의 효과를 제거한 후 두 변수군의 관계를 비교하는 방법이 요구된다. 본 연구에서는 일반지식 영역과 언어적 영역의 편정준상관 행렬도를 정준상관 행렬도와 비교하여 두 변수군의 관계를 살펴보았다.

그림 3.1의 (a)와 (b)는 각각 재능 자료에 대한 일반지식 영역, 언어적 영역의 변수군의 2차원 정준상관 행렬도이며, (c)는 일반지식 영역의 변수군에 의한 전체 학생의 행렬도이다. 언어적 영역의 변수군에 의한 전체 학생의 행렬도 또한 비슷한 결과를 제공하므로 이는 생략하였다. 제1정준상관은 0.757이

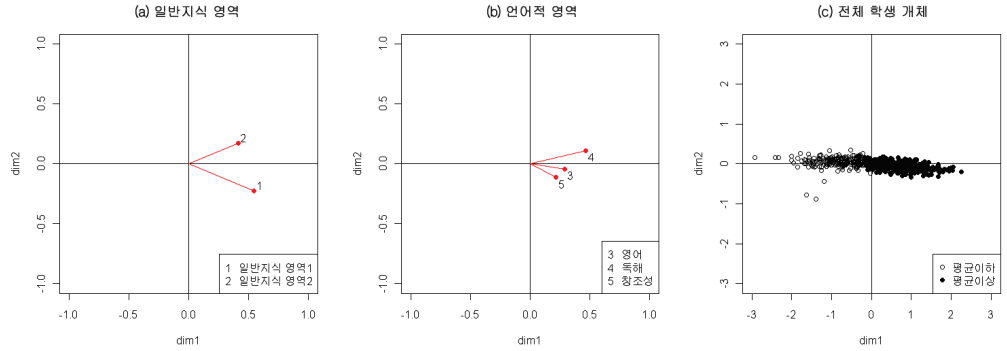

그림 3.2. 재능 자료의 편정준상관 행렬도

고 제2정준상관은 0.177이며, 근사적합도는 제1정준축이 94.80%이고 제2정준축이 5.20%이므로 1차원만 고려하여도 원자료를 잘 설명할 수 있다.

그림 3.1(a) 일반지식 영역의 행렬도에서 일반지식 영역1과 일반지식 영역2 모두 오른편에 위치하고 있으며 두 사이각의 코사인값은 0.292로 두 변수가 거의 무관하게 나타나고 있다. 그림 3.1(b) 언어적 영역의 행렬도에서도 영어, 독해, 창조성 모두 오른편에 위치하고 있으며 세 변수 모두 양의 상관이 존재하고 특히 영어와 독해는 매우 유사한 성격으로 나타나고 있다. 그림 3.1(c) 전체 학생들의 행렬도는 오른쪽에 주로 평균이상인 학생들이 나타나고 왼쪽에 평균이하인 학생들이 나타나고 있다. 전체 행렬도를 비교해서 살펴보면 일반지식 영역2는 영어, 독해와 관련이 높고 일반지식 영역1은 창조성과 관련이 높으며, 개체들의 행렬도에서 변수들이 존재하는 방향인 오른편에 위치하고 있는 학생들이 이 변수들에 대해 성적이 좋은 학생들이다.

다음으로는 일반지식 영역과 언어적 영역의 두 변수군에 영향을 줄 수 있는 비언어적 영역과 흥미도의 효과를 고려하여 그 영향력을 제거한 일반지식 영역과 언어적 영역의 두 변수군의 편정준상관 행렬도를 살펴보도록 한다.

그림 3.2의 (a)와 (b)는 재능 자료에 대한 각각 일반지식 영역, 언어적 영역의 변수군에 대한 2차원 편정준상관 행렬도이고, (c)는 일반지식 영역의 변수군에 의한 전체 학생의 행렬도이다. 편정준상관 행렬도에서도 언어적 영역의 변수군에 의한 전체 학생의 행렬도는 일반지식 영역의 변수군에 의한 전체 학생의 행렬도와 비슷한 결과를 제공하므로 이는 생략하였다. 제1정준상관은 0.566, 제2정준상관은 0.102이고 근사적합도는 제1정준축이 96.86%, 제2정준축이 3.14%이므로 1차원만 고려하여도 원자료를 잘 설명할 수 있다.

그림 3.2(a) 일반지식 영역의 행렬도에서 일반지식 영역1과 일반지식 영역2 모두 오른편에 위치하고 있으며 두 사이각의 코사인값은 0.719로 그림 3.1(a)에서 두 변수가 거의 무관하게 나타나고 있었던 것과 다르게 양의 상관관계가 존재하는 것으로 나타난다. 그림 3.2(b) 언어적 영역의 행렬도에서도 영어, 독해, 창조성 모두 오른편에 위치하고 있으며 세 변수 모두 양의 상관이 존재하고 그림 3.1(b)에서 영어가 독해와 매우 유사하게 나타났던 것에 반해, 편정준상관 행렬도에서는 영어가 독해와 창조성이 이루는 각을 거의 이등분하고 있으므로 독해와 밀접한 관련이 있기보다는 독해, 창조성 모두와 비슷하게 관련이 높다. 그림 3.2(c) 전체 학생들의 행렬도는 그림 3.1(c)와 비슷한 결과를 나타내고 있으며 제2축에 대해서 중심으로 더 많이 모인 모습이다. 전체 행렬도를 비교해서 살펴보면 일반지식 영역2는 독해와 관련이 높고 일반지식 영역1은 창조성과 관련이 높으며, 개체들의 행렬도에서 변수들이 존재하는 방향인 오른편에 위치하고 있는 학생들이 이 변수들에 대해 성적이 좋은 학생들이다.

표 **3.1.** 행렬도에 따른 프로크러스티즈 통계량

		편정준상관 행렬도		
		일반지식 영역	언어적 영역	전체 학생 개체
정준상관 행렬도	일반지식 영역	0.037		
	언어적 영역		0.051	
	전체 학생 개체			15.531

끝으로 표 3.1의 프로크러스티즈 통계량을 통해 정준상관 행렬도와 편정준상관 행렬도의 형상변동을 살펴보자. 일반적으로 두 행렬도의 형상이 일치하는 경우 이 통계량은 0이 되어 변동차이가 없음을 의미한다. 일반지식 영역의 행렬도에서는 0.037의 변동이 있었고 언어적 영역의 행렬도에서는 0.051, 일반지식 영역에 의한 전체 학생 개체의 행렬도에서는 15.531의 변동이 나타났다. 이는 두 행렬도의 형상 간에 차이가 있음을 의미하고 특히 좌표수가 가장 많은 전체 학생 개체의 행렬도에서 변동이 가장 높게 나타났다. 이전의 각 행렬도에 대한 해석을 통하여, 일반지식 영역1과 일반지식 영역2 모두 기본적인 지식 정도를 측정하는 시험이므로 두 변수가 무관하게 나타난 정준상관 행렬도 보다는 서로 관련이 높게 나타나는 편정준상관 행렬도가 더 적절함을 알 수 있다. 특히, 대상이 미국 학생임을 고려하면 영어가 특정한 변수와 유사하게 나타나는 것보다 모든 변수와 적절히 높은 상관을 나타내는 편정준상관 행렬도를 고려하는 것이 더욱 바람직하다.

결과적으로, 두 변수군에 영향을 미치는 공변량변수군이 존재하는 자료를 접하게 될 때에는 공변량변수군의 효과 때문에 잘못된 해석 또는 불분명한 해석을 하게 되므로 편정준상관 행렬도를 고려할 필요가 있다.

참고문헌

최용석 (2006). <행렬도 분석>, 부산대학교 기초과학연구원 기초과학총서 2, 부산대학교 출판부.

최태훈, 최용석 (2008). 정준상관 행렬도와 군집분석을 응용한 KLPGA 선수의 기술과 경기성적요인에 대한 연관성 분석, <응용통계연구>, **21**, 429–439.

최태훈, 최용석 (2010). 일반화 정준상관 행렬도와 프로크러스티즈 분석을 응용한 대한테니스협회 등록 선수의 체격요인, 체력요인 및 기초기술요인에 대한 분석연구, <한국통계학논문집>, **17**, 917–925.

최태훈, 최용석, 신상민 (2009). 테니스 그랜드슬램대회의 선수특성요인과 경기용인에 대한 분석연구 -정준상관 행렬도와 프로크러스티즈 분석의 응용-, <응용통계연구>, **22**, 855–864.

허명회 (1993). <통계상담의 이해>, 자유아카데미, 서울.

홍현욱, 최용석, 신상민, 강창완 (2010). 결측값이 있는 정준상관 행렬도의 형상변동 연구, <응용통계연구>, **23**, 955–966.

Choi, Y. S. (1991). *Resistant Principal Component Analysis, Biplot and Correpondence Anaysis*, Unpublished Ph.D. Dissertation, Department of Statistics, Korea University.

Cooley, W. W. and Lohnes, P. R. (1971). *Multivariate Data Analysis*, Wiley, New York.

Gabriel, K. R. (1971). The biplot graphics display of matrices with applications to principal analysis, *Biometrika*, **58**, 453–467.

Hotelling, H. (1936). Relations between two sets of variates, *Biometrika*, **28**, 321–377.

Park, M. R. and Huh, M. H. (1996). Canonical correlation biplot, *The Korea Communications in Statistics*, **3**, 11–19.

Rao, B. (1969). Partial canonical correlations, *Trabajos de Estudestica y de Investigacion Oerativa*, **20**, 211–219.

Timm, N. H. (2002). *Applied Multivariate Analysis*, Springer, New York.

Partial Canonical Correlation Biplot

Ah-Rim Yeom[1] · Yong-Seok Choi[2]

[1]Department of Statistics, Pusan National University
[2]Department of Statistics, Pusan National University

(Received April 2011; accepted May 2011)

Abstract

Biplot is a useful graphical method to explore simultaneously rows and columns of two-way data matrix. In particular, canonical correlation biplot is a method for investigating two sets of variables and observations in canonical correlation analysis graphically. For more than three sets of variables, we can apply the generalized canonical correlation biplot in generalized canonical correlation analysis which is an expansion of the canonical correlation analysis. On the other hand, we consider the set of covariate variables which is affecting the linearly two sets of variables. In this case, if we apply the generalized canonical correlation biplot, we cannot clearly interpret the other two sets of variables due to the effect of the set of covariate variables. Therefor, in this paper, we will apply the partial canonical correlation analysis of Rao (1969) removing the linear effect of the set of covariate variables on two sets of variables. We will suggest the partial canonical correlation biplot for inpreting the partial canonical correlation analysis graphically.

Keywords: Biplot, set of covariate variables, partial canonical correlation analysis, partial canonical correlation biplot.

This work was supported by a 2-Year Research Grant of Pusan National University.
[2]Corresponding author: Professor, Department of Statistics, Pusan National University, Jangjeon-Dong, Geumjeong-Gu, Pusan 609-735, Korea. E-mail: yschoi@pusan.ac.kr