응용통계연구 (2010)
23(5), 955–966
DOI: 10.5351/KJAS.2010.23.5.955

결측값이 있는 정준상관 행렬도의 형상변동 연구

홍현욱[1] · 최용석[2] · 신상민[3] · 강창완[4]

[1]부산대학교 통계학과, [2]부산대학교 통계학과, [3]부산대학교 통계학과, [4]동의대학교 정보통계학과

(2010년 6월 접수, 2010년 8월 채택)

요 약

정준상관 행렬도는 두 변수군 사이에 연관성이 있는 데이터 행렬을 시각적으로 묘사하고 데이터가 가진 패턴을 찾는데 유용하고, 분석의 더욱 정형화된 방법으로써 결과를 보여주기에도 유용하다. 그럼에도 불구하고, 자료에 결측값이 존재하는 경우에 대부분의 행렬도는 바르게 적용되지 않는다. 이 문제를 해결하기 위해, 결측률에 따라 중앙값과 평균, EM알고리즘, MCMC대체법을 사용해서 결측 자료를 추정한다. 완전하지 않은 자료의 행렬도의 결측값을 추정하더라도, 대체법과 결측률에 따라 행렬도의 모양이 달라진다. 따라서 Shin 등 (2008)에서 제안한 RMS(root mean square)와 원 행렬도와 추정된 행렬도간의 형상 변동을 측정하고 비교하기 위한 PS(Procrustes statistic)를 사용한다.

주요용어: 정준상관 행렬도, 형상 변동, 프로크러스티즈, 결측 구조, 대체법.

1. 서론

정준상관 행렬도(canonical correlation biplot)는 일반적인 자료 분석 방법에서 나아가, 데이터가 가진 패턴을 발견하고 두 변수군 사이의 연관성을 포함한 데이터 행렬의 시각적 기술을 제공하는데 유용하게 사용된다. 국내에선 Park과 Huh (1996)가 정준상관분석에서 수량화 방법(quantification method) 관점에서 2차원 그림을 제안하였고 이를 정준상관 행렬도라 하였다.

행렬도는 Gabriel (1971)에 의해서 개발되었고 국내에서 Choi (1991)가 저항성 버전을 개발하면서 처음으로 소개하였다. 행렬도는 복잡한 다변량 분석의 결과 해석의 용이성을 통해 활발한 연구와 응용이 여러 분야에서 이루어지고 있다. Shin 등 (2008)은 결측값을 가진 행렬도에 관한 연구를 진행하였으며, 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. 또한 최태훈 등 (2009)은 테니스 그랜드 슬램 대회의 선수특성 요인과 경기 요인에 대한 행렬도에서 프로크러스티즈 분석(Procrustes analysis)을 통하여 행렬도의 형상 비교를 하였다.

정준상관 분석에서 변수군의 일부 데이터가 결측이 되면, 그 결측 자료를 포함한 자료의 행 전체가 분석에서 제외되어 분석의 정확성이 떨어지게 된다. 따라서 본 논문에서는 이러한 정보의 손실을 없애고, 원 데이터로부터 도출된 변수들 간의 관계를 잘 나타낼 수 있도록 하는 대체방법들에 대하여 각 결측률에 따른 형상변동을 비교 분석하였다. 평균, 중위수대체, EM알고리즘과 MCMC방법을 사용한 다중대체를 통한 결측값을 추정한 자료행렬을 생성하였다. 기존의 주성분, 판별분석, 대응분석 행렬도 등과는

이 논문은 부산대학교 자유과제 학술연구비(2년)에 의하여 연구되었음.
[2]교신저자: (609-735) 부산시 금정구 장전동 산 30, 부산대학교 통계학과, 교수. E-mail: yschoi@pusan.ac.kr

다르게 두 변수군의 형상의 변동을 동시적으로 확인하는 것이 필요하여, 추정된 자료행렬을 이용한 정준상관 행렬도를 원 자료의 정준상관 행렬도에 적합 시키는 프로크러스티즈 분석을 활용하였다. 2절에서는 정준상관 행렬도의 소개와 더불어 형상변동 측정과 프로크러스티즈 분석을 설명하고자 한다. 3절에서는 결측 구조와 다중대체법에 대한 소개를 다루었으며, EM알고리즘법, MCMC방법을 통한 다중대체법을 정리하였다. 4절에서는 적용사례를 제시하려 한다.

2. 정준상관 행렬도와 형상변동

2.1. 정준상관 행렬도

이 절에서는 최용석 (2006)을 참고로 하여 정준상관 분석과 관련된 정준상관 행렬도에 대하여 정리하기로 한다.

정준상관분석은 두 변수군 사이의 관계를 분석하는 다변량 기법이다. 두 변수군의 선형 조합(linear combination)간의 상관관계를 가장 크게 만드는 최대의 선형 조합을 찾아내어 그 관계를 분석하는 것이 목적이다. q개의 변수로 이루어진 두 변수군 $\mathbf{x} = (x_1, \ldots, x_p)^T$와 $\mathbf{y} = (y_1, \ldots, y_q)^T$는 각각 평균 $\mu_{\mathbf{x}} = (\mu_{x_1}, \ldots, \mu_{x_p})^T$와 $\mu_{\mathbf{y}} = (\mu_{y_1}, \ldots, \mu_{y_q})^T$를 가지며 공분산행렬 Σ_{xx}, Σ_{yy}, $\Sigma_{xy} = \Sigma_{yx}^T$을 가지는 확률벡터이다.

임의의 계수벡터 $\mathbf{u}$와 $\mathbf{v}$에 대하여 두 변수군 각각의 선형결합

$$\mathbf{Z_x} = u_1x_1 + \cdots + u_1x_p = \mathbf{u}^T\mathbf{x}, \qquad \mathbf{Z_y} = v_1y_1 + \cdots + v_1y_q = \mathbf{v}^T\mathbf{y} \tag{2.1}$$

에서 식 (2.1)의 두 선형결합 $\mathbf{Z_x}$와 $\mathbf{Z_y}$의 상관은

$$r_{\mathbf{Z_xZ_y}} = \frac{\sum_{i=1}^{n} z_{x_i} z_{y_i}}{\sqrt{\sum_{i=1}^{n} z_{x_i}^2}\sqrt{\sum_{i=1}^{n} z_{y_i}^2}} = \frac{\mathbf{u}^T\mathbf{S}_{xy}\mathbf{v}}{\sqrt{\mathbf{u}^T\mathbf{S}_{xx}\mathbf{u}}\sqrt{\mathbf{v}^T\mathbf{S}_{yy}\mathbf{v}}} \tag{2.2}$$

이다. 여기서 계수벡터 $\mathbf{u}$와 $\mathbf{v}$는 식 (2.2)의 두 선형 결합의 상관을 최대화하는 알고리즘을 통하여 구할 수 있다. 상관을 최대화하는 알고리즘은 $\mathbf{Z}_x$와 $\mathbf{Z}_y$의 분산이 1인 제약조건을 $\mathbf{u}^T\mathbf{S}_{xx}\mathbf{u} = 1$과 $\mathbf{v}^T\mathbf{S}_{yy}\mathbf{v} = 1$을 두고 $\mathbf{u}^T\mathbf{S}_{xy}\mathbf{v}$를 최대화하는 계수벡터 $\mathbf{u}$와 $\mathbf{v}$를 찾는 것과 동일하다. 이는 라그랑즈 승수(Lagrange multiplier) 방법을 이용하면 고유체계 문제로 유도하여 풀 수 있다. 또한 이 알고리즘은 정준계수벡터와 정준상관을 대수적으로 한꺼번에 제공하는 비정칙값분해(singular value decomposition)

$$\mathbf{S}_{xx}^{-\frac{1}{2}}\mathbf{S}_{xy}\mathbf{S}_{yy}^{-\frac{1}{2}}\mathbf{S}_{yx} = \mathbf{U}\mathbf{D}_{\sqrt{\lambda}}\mathbf{V}^T$$

를 이용하면 간편하게 구할 수 있다. 여기서 크기가 $p \times r$과 $q \times r$행렬 $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_r)^T$와 $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_r)^T$은 정준계수벡터의 직교행렬이며 대각행렬 $\mathbf{D}_{\sqrt{\lambda}} = \mathrm{daig}\left(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_r}\right)$는 $\sqrt{\lambda_1} \geq \cdots \geq \sqrt{\lambda_r} > 0$의 관계를 갖는 비정칙값이 정준상관에 해당하고 이를 대각원소로 하고 있다. 이를 통하여, i번째 정준상관 계수벡터를 구하면 각각

$$\mathbf{a}_i = \mathbf{S}_{xx}^{-\frac{1}{2}}\mathbf{u}_i, \qquad \mathbf{b}_i = \mathbf{S}_{yy}^{-\frac{1}{2}}\mathbf{v}_i, \quad i = 1, \ldots, r$$

이다. 이들에 의해 구성된 정준상관계수행렬은 $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_r)$과 $\mathbf{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_r)$이 된다. 이를 통해 계산된 자료행렬 $\mathbf{X}$에 대한 정준상관 행렬도의 행 좌표 행렬과 열 좌표 행렬은

$$\mathbf{R_X} = \mathbf{X}\mathbf{A}\mathbf{D}_{\sqrt{\lambda}}, \qquad \mathbf{C_X} = \mathbf{A}\mathbf{D}_{\sqrt{\lambda}}$$

이 되며, 자료행렬 $\mathbf{Y}$에 대한 정준상관 행렬도의 행 좌표 행렬과 열 좌표 행렬은

$$\mathbf{R_Y} = \mathbf{YBD}_{\sqrt{\lambda}}, \qquad \mathbf{C_Y} = \mathbf{BD}_{\sqrt{\lambda}}$$

가 된다.

2.2. 형상변동의 측정

원 자료의 정준상관 행렬도와 각각의 결측 대체방법을 적용한 정준상관 행렬도와의 비교를 위하여 형상변동을 측정한다. 본 연구에서는 결측이 대체된 자료행렬을 통한 정준상관 행렬도의 변수군의 형상변동을 측정하기 위해 Shin 등 (2008)이 제안한 RMS(root mean squared)방법을 사용하였다. 이 외에도 형상변동 측정과 관련하여 Kim 등 (2010a, 2010b)의 최근 연구도 참고할 만하다. 최종적으로 생성된 $\mathbf{X}, \mathbf{Y}$ 개체군과 변수군의 자료 행렬을 $\widehat{\mathbf{R}}_\mathbf{X}, \widehat{\mathbf{R}}_\mathbf{Y}, \widehat{\mathbf{C}}_\mathbf{X}, \widehat{\mathbf{C}}_\mathbf{Y}$라고 하자. $\widehat{\mathbf{R}}_\mathbf{X}, \widehat{\mathbf{R}}_\mathbf{Y}$는 $n \times 2$ 행렬이고, $\widehat{\mathbf{C}}_\mathbf{X}, \widehat{\mathbf{C}}_\mathbf{Y}$는 각각 $p \times 2$, $q \times 2$ 행렬이다. 여기서 개체군과 변수군의 좌표행렬을 세로로 병합한 행렬을 각각 $\widehat{\mathbf{R}}, \widehat{\mathbf{C}}$라 하면 다음과 같다.

$$\widehat{\mathbf{R}} = \begin{pmatrix} \widehat{\mathbf{R}}_\mathbf{X} \\ \widehat{\mathbf{R}}_\mathbf{Y} \end{pmatrix}, \qquad \widehat{\mathbf{C}} = \begin{pmatrix} \widehat{\mathbf{C}}_\mathbf{X} \\ \widehat{\mathbf{C}}_\mathbf{Y} \end{pmatrix}$$

행렬 $\widehat{\mathbf{R}}$은 $2n \times 2$, $\widehat{\mathbf{C}}$은 $(p+q) \times 2$ 행렬이 되며, 이들 행렬을 벡터화 하면 $\mathrm{vec}(\widehat{\mathbf{R}})$는 $4n \times 1$, $\mathrm{vec}(\widehat{\mathbf{C}})$는 $2(p+q) \times 1$인 벡터가 된다.

$$\mathrm{vec}\left(\widehat{\mathbf{R}}\right) = \left(\hat{\mathbf{r}}_1^T, \ldots, \hat{\mathbf{r}}_{4n}^T\right)^T, \qquad \mathrm{vec}\left(\widehat{\mathbf{C}}\right) = \left(\hat{\mathbf{c}}_1^T, \ldots, \hat{\mathbf{c}}_{2(p+q)}^T\right)^T.$$

이를 이용한 원 자료 정준상관 행렬도의 변수군의 좌표와 추정된 자료의 정준상관 행렬도 좌표간 거리의 RMS

$$\mathrm{RMS}_\mathbf{R} = \sqrt{\frac{1}{4n}\left\|\mathrm{vec}\left(\widehat{\mathbf{R}} - \mathbf{R}\right)^2\right\|}, \qquad \mathrm{RMS}_\mathbf{C} = \sqrt{\frac{1}{2(p+q)}\left\|\mathrm{vec}\left(\widehat{\mathbf{C}} - \mathbf{C}\right)^2\right\|}$$

가 0에 가까울수록 형상변동이 작아 추정이 잘 이루어 졌음을 보여준다. 하지만, 정준상관 행렬도의 형상변동에서는 각각의 변수 군에서 동시적인 변동을 고려해야 하므로, 단순한 변동에 대한 측도인 RMS는 한계를 가진다. 따라서 원 자료의 정준상관 행렬도의 변수들의 형상과 대체된 자료의 정준상관 행렬도의 변수들의 형상의 비교를 위해 프로크러스티즈 분석을 추가적으로 고려하였다. 프로크러스티즈 분석이란 기하학적 공간상에서 형상점(landmark)에 의해서 나타낸 개체들의 형상을 측정하고, 기술하며 비교하는 형상분석에서 개체간의 형상비교를 위해 한 개체를 다른 개체 쪽으로 적합 시키는 방법이다 (최용석과 현기훙, 2006). 원 자료 행렬은 $\mathbf{W}_s$, 대체된 자료 행렬은 $\mathbf{W}_r$라고 하고, 두 중심화 형상좌표 행렬의 크기는 $k \times m$이라고 하자. 벡터 $\mathbf{t}$와 직교행렬 $\mathbf{R}$에 의해서 $\mathbf{W}_r$의 l번째 점 $\mathbf{w}_{r(l)}$의 좌표점을 변환한 $\mathbf{R}\mathbf{w}_{r(l)} + \mathbf{t}$을 고려하자. 그러면 $\mathbf{W}_s$의 l번째 점 $\mathbf{w}_{s(l)}$과 $\mathbf{R}\mathbf{w}_{r(l)} + \mathbf{t}$간의 제곱거리합

$$\sum_{l=1}^{k} \left(\mathbf{w}_{s(l)}\mathbf{R}\mathbf{w}_{r(l)} - \mathbf{t}\right)^T \left(\mathbf{w}_{s(l)}\mathbf{R}\mathbf{w}_{r(l)} - \mathbf{t}\right) \tag{2.3}$$

을 생각할 수 있으며, 식 (2.3)을 최소화 하는 $\mathbf{R}$과 $\mathbf{t}$를 찾는 것이 두 형상 $\mathbf{W}_s$과 $\mathbf{W}_r$가 잘 일치되도록 하는 정보를 제공한다. 비정칙치분해(singular value decomposition) $\mathbf{W}_r^T\mathbf{W}_s = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T$를 이

용하여 $\widehat{\mathbf{R}} = \mathbf{V}\mathbf{U}^T$과 $\hat{\mathbf{t}} = \mathbf{0}$를 제공받을 수 있으며 이를 이용하여 두 형상의 일치성을 평가하는 측도(measure)인 프로크러스티즈 통계량

$$\mathrm{PS}\left(\mathbf{W}_r, \mathbf{W}_s\right) = \mathrm{tr}\left(\mathbf{W}_r^T\mathbf{W}_r\right) + \mathrm{tr}\left(\mathbf{W}_s^T\mathbf{W}_s\right) - 2\mathrm{tr}(\Lambda) \tag{2.4}$$

을 얻을 수 있다. 식 (2.4)의 프로크러스티즈 통계량 값이 0이면 두 형상은 일치한다고 평가할 수 있다. 본 연구에서는 크기가 $(p+q)\times 2$인 변수군의 형상좌표를 사용하여 변수들의 형상변동을 측정하였고, 앞서 사용된 RMS와 함께 형상변동을 나타내고자 한다.

3. 결측구조와 다중대체법

3.1. 결측구조(Missing Mechanism)

결측구조에 대한 자료행렬 $\mathbf{Y} = \{y_{ij}\}$는 결측값을 가진 $N \times p$ 행렬이고, y_{ij}가 결측되었을 때는 1, 그렇지 않을 때는 0을 가지는 $N \times p$ 지시 행렬을 $\mathbf{R} = \{r_{ij}\}$이라고 하자. $\Pr\{r_{ij} = 0|y_{ij}\} = \Pr\{y_{ij}\,\text{observed}|y_{ij}\} = p_{ij}$라고 정의할 때, 자료행렬 $\mathbf{Y}$는 미지의 모수 θ에 대해 $P(\mathbf{Y}|\theta)$를 가지고, $\mathbf{R}$은 미지의 모수 ξ에 의해 영향을 받는 확률 분포 $P = (\mathbf{R}|\xi, \mathbf{Y})$를 가진다고 할 수 있다. 또한 $\mathbf{Y} = (\mathbf{Y}_{obs}, \mathbf{Y}_{mis})$로 표현되고, $\mathbf{Y}_{obs}$는 관찰된 값이며, $\mathbf{Y}_{mis}$는 결측치를 뜻한다. 이러한 가정 하에 반응변수들과 결측 지시변수들의 결합확률분포(joint probability distribution)는 다음과 같이 표현된다.

$$P(\mathbf{Y}, \mathbf{R}|\theta, \xi) = P(\mathbf{Y}|\theta)P(\mathbf{R}|\xi, \mathbf{Y})$$

여기에서 Little과 Rubin (2002)에서는 조건부 분포인 $P(\mathbf{R}|\xi, \mathbf{Y})$에 기초하여 결측구조를 3가지로 범주화하였다. 첫 번째, 완전임의결측(MCAR; missing completely at random)은 결측의 발생여부가 주어진 데이터에 전혀 의존하지 않는 것을 의미한다. 즉

$$P(\mathbf{R}|\xi, (\mathbf{Y}_{obs}, \mathbf{Y}_{mis})) = P(\mathbf{R}|\xi)$$

으로 표현할 수 있다. 두 번째로 결측의 발생여부가 관찰된 원소($\mathbf{Y}_{obs}$)에만 의존할 때, 조건적 임의결측(MAR; missing at random)이라고 정의한다. 즉 결측이 $\mathbf{Y}$의 결측값에 영향을 받지 않으나 관측값에는 영향을 받아 모든 $\mathbf{Y}_{mis}$에 대하여

$$P(\mathbf{R}|\xi, (\mathbf{Y}_{obs}, \mathbf{Y}_{mis})) = P(\mathbf{R}|\xi, \mathbf{Y}_{obs})$$

가 성립함을 뜻한다. 마지막으로 결측 발생이 관찰되지 않는 값($\mathbf{Y}_{mis}$)에 의존할 경우를 NMAR(not missing at random)으로 정의하고 결측이 임의로 이루어지지도 않으며, 모형의 변수들을 가지고 특정한 변수의 값을 예측할 수 없는 경우를 뜻한다. 그리고 MCAR과 MAR을 무시할 수 있는 결측 구조라고 하였다. 본 연구의 경우 결측이 결측된 값에 영향을 받지 않고, 관찰된 값에 영향을 받는 MAR 상황을 가정하여 진행하였다.

3.2. 다중대체법(Multiple Imputation)

대체법은 불완전한 데이터를 채워 넣는 일반적인 방법으로, 각각의 결측값은 예측분포와 여러 방법을 통하여 추출된 값으로 채워진다. 하지만 이러한 과정에서 대체된 값은 고려되지 않은 불확실성에 기인한 변동을 일으킨다. 다중대체법은 반복적인 수행을 통하여 결측값을 대체하여 변동성을 줄이고자 하는데 그 의의가 있다. Rubin (1987)에 따르면 다중대체의 스텝은 아래와 같다.

Step 1. 대체 방법들을 통하여 완전한 m개의 데이터 셋을 만든다.

Step 2. m개의 대체 데이터 셋으로부터 모수 θ의 추정량 $\hat{\theta}_j$를 산출하고 $\hat{\theta}_j$의 분산을 $V(\hat{\theta}_j)$표기 한다($j = 1, \ldots, m$).

Step 3. $\hat{\theta}_1, \ldots, \hat{\theta}_m$을 결합하여 θ에 대한 추정량과 그것의 분산을 산출한다.

$$\bar{\theta}_{MI} = \frac{1}{m}\left(\hat{\theta}_1 + \cdots + \hat{\theta}_m\right), \quad V(\bar{\theta}_{MI}) = \frac{1}{m}\sum_{j=1}^{m} V\left(\hat{\theta}_j\right) + \left(1 + \frac{1}{m}\right)\frac{1}{m-1}\sum_{j=1}^{m}\left(\hat{\theta}_j - \bar{\theta}_{MI}\right)^2.$$

3.3. EM알고리즘법

EM의 각 반복 단계는 E(Expectation)-단계와 M(Maximization)-단계로 구성된다. 먼저, $\mathbf{Y} = (\mathbf{Y}_{obs}, \mathbf{Y}_{mis})$라고 하면, $\mathbf{Y}_{obs}$는 관측된 데이터이고, $\mathbf{Y}_{mis}$는 결측이 발생한 것을 뜻한다. 불완전한 데이터 문제에서 완전한 데이터의 분포는 다음과 같이 분해될 수 있다 (Schafer, 1997, pp. 37–39).

$$P(\mathbf{Y}|\theta) = P(\mathbf{Y}_{obs}|\theta)P(\mathbf{Y}_{mis}|\mathbf{Y}_{obs}, \theta). \tag{3.1}$$

식 (3.1)의 양변에 log-likeligood를 취하면 다음과 같다.

$$l(\mathbf{Y}|\theta) = l(\mathbf{Y}_{obs}|\theta) + \log P(\mathbf{Y}_{mis}|\mathbf{Y}_{obs}, \theta) + C. \tag{3.2}$$

식 (3.2)의 $l(\mathbf{Y}|\theta) = \log P(\mathbf{Y}|\theta)$는 완전한 데이터의 로그 우도이며 $l(\mathbf{Y}_{obs}|\theta) = \log L(\theta|\mathbf{Y}_{obs})$는 관측된 데이터의 로그 우도이며 C는 임의의 상수이다.

E-단계에서 식 (3.2)에 기댓값을 $P(\mathbf{Y}_{mis}|\mathbf{Y}_{obs}, \theta^{(t)})$에 대해 취하여 $Q(\theta|\theta^{(t)})$를 계산하면 아래와 같은 결과를 얻을 수 있다.

$$Q\left(\theta|\theta^{(t)}\right) = l(\theta|\mathbf{Y}_{obs}) + H\left(\theta|\theta^{(t)}\right) + C,$$

여기에서

$$Q\left(\theta|\theta^{(t)}\right) = \int l(\theta|\mathbf{Y}) P\left(\mathbf{Y}_{mis}|\mathbf{Y}_{obs}, \theta^{(t)}\right) d\mathbf{Y}_{mis}$$

이고,

$$H\left(\theta|\theta^{(t)}\right) = \int \log P(\mathbf{Y}_{mis}|\mathbf{Y}_{obs}, \theta) P\left(\mathbf{Y}_{mis}|\mathbf{Y}_{obs}, \theta^{(t)}\right) d\mathbf{Y}_{mis}$$

이다. M-단계에서는 $Q(\theta|\theta^{(t)})$를 최대화 하는 θ^{t+1}을 찾는다. 또한 여기에서 θ^{t+1}는

$$l\left(\theta^{(t+1)}|\mathbf{Y}_{obs}\right) \geq l\left(\theta^{(t)}|\mathbf{Y}_{obs}\right)$$

를 항상 만족한다. 이를 다시 표현 하면,

$$Q\left(\theta^{(t+1)}|\theta^{(t)}\right) - Q\left(\theta^{(t)}|\theta^{(t)}\right) - \left[H\left(\theta^{(t+1)}|\theta^{(t)}\right) - H\left(\theta^{(t)}|\theta^{(t)}\right)\right] \geq 0$$

이고, 모든 θ에 대해 아래의 식이 성립한다.

$$Q\left(\theta^{(t+1)}|\theta^{(t)}\right) \geq Q\left(\theta^{(t)}|\theta^{(t)}\right)$$

그리고,

$$H\left(\theta^{(t+1)}|\theta^{(t)}\right) - H\left(\theta^{(t)}|\theta^{(t)}\right) = \int \log \frac{P\left(\mathbf{Y}_{mis}|\mathbf{Y}_{obs},\theta^{(t+1)}\right)}{P\left(\mathbf{Y}_{mis}|\mathbf{Y}_{obs},\theta^{(t)}\right)} P\left(\mathbf{Y}_{mis}|\mathbf{Y}_{obs},\theta^{(t)}\right) d\mathbf{Y}_{mis}$$

이므로, $\log x \leq x-1$ 혹은 Jensen의 부등식을 이용하면

$$H\left(\theta^{(t)}|\theta^{(t)}\right) \geq H\left(\theta^{(t+1)}|\theta^{(t)}\right)$$

를 만족한다. 이러한 E-단계와 M-단계를 추정치가 수렴할 때까지 계속해서 반복한다.

3.4. MCMC법

MCMC의 한 방법인 자료확대(data augmentation)는 결측 데이터와 모수들을 반복적인 마르코프 연쇄(Markov chain) 시뮬레이션을 통하여 안정적인 분포로 수렴하는 값을 구하여 분포를 추정하는 방법이다 (Tanner와 Wong, 1987). 확률벡터 $\mathbf{z}$가 $\mathbf{z}=(\mathbf{u},\mathbf{v})$와 같이 두 개의 하위 벡터로 분할된다고 가정하면, 결합 확률 분포 $P(\mathbf{Z})$는 쉽게 시뮬레이션 되지 않지만 각각의 조건부 분포인

$$P(\mathbf{u}|\mathbf{v}) = g(\mathbf{u}|\mathbf{v}), \qquad P(\mathbf{v}|\mathbf{u}) = h(\mathbf{v}|\mathbf{u})$$

로 쉽게 계산되어 질 수 있다. 각각의 반복계산에서

$$\begin{aligned}\mathbf{z} &= \left(\mathbf{z}_1^t, \mathbf{z}_2^t, \ldots, \mathbf{z}_m^t\right) \\ &= \left(\left(\mathbf{u}_1^t, \mathbf{v}_1^t\right), \left(\mathbf{u}_2^t, \mathbf{v}_2^t\right), \ldots, \left(\mathbf{u}_m^t, \mathbf{v}_m^t\right)\right)\end{aligned}$$

라고 두면, $P(\mathbf{Z})$와 근사한 분포로부터 추출된 크기가 m인 표본이 된다. 그리고 이 표본은 두 가지 단계에 의해 갱신되어 진다. 첫 번째로

$$\mathbf{u}_i^{(t+1)} \sim g\left(\mathbf{u}|\mathbf{v}_i^{(t)}\right)$$

에 의해 독립적인 $i=1,\ldots,m$ 까지의 추출을 통해

$$\mathbf{U}^{(t+1)} = \left(\mathbf{u}_1^{(t+1)}, \mathbf{u}_2^{(t+1)}, \mathbf{u}_3^{(t+1)}, \ldots, \mathbf{u}_m^{(t+1)}\right)$$

를 생성하고, $h(\mathbf{v}|\mathbf{u}_i^{(t)})$에서의 추출을 통하여

$$\mathbf{V}^{(t+1)} = \left(\mathbf{v}_1^{(t+1)}, \mathbf{v}_2^{(t+1)}, \mathbf{v}_3^{(t+1)}, \ldots, \mathbf{v}_m^{(t+1)}\right)$$

을 생성한다. 이 과정을 거치면서 새로운 표본 $\mathbf{z}^{(t+1)}$이 생성된다.

$$\mathbf{z}^{(t+1)} = \left(\left(\mathbf{u}_1^{t+1}, \mathbf{v}_1^{t+1}\right), \left(\mathbf{u}_2^{t+1}, \mathbf{v}_2^{t+1}\right), \ldots, \left(\mathbf{u}_m^{t+1}, \mathbf{v}_m^{t+1}\right)\right).$$

Tanner와 Wong (1987)은 $\mathbf{Z}^{(t)}$의 분포가 $t\to\infty$임에 따라 $P(\mathbf{Z})$로 수렴하는 것을 보였다. 이러한 과정을 결측 자료 데이터에 적용한다. MCMC방법에서는 I-단계와 P-단계로 나누어진다.

$$\mathbf{Y}_{mis}^{(t+1)} \sim P\left(\mathbf{Y}_{mis}|\mathbf{Y}_{obs},\theta^{(t)}\right), \tag{3.3}$$

$$\theta^{(t+1)} \sim P\left(\theta|\mathbf{Y}_{obs},\mathbf{Y}_{mis}^{(t+1)}\right). \tag{3.4}$$

I-단계인 식 (3.3)에서는 사후 예측 분포 $P(\mathbf{Y}_{mis}|\mathbf{Y}_{obs},\theta^{(t)})$에서 $\mathbf{Y}_{mis}^{(t+1)}$을 추출하고, P-단계인 식 (3.4)에서는 추출된 $\mathbf{Y}_{mis}^{(t+1)}$을 통해 대체된 자료의 사후분포 $P(\theta|\mathbf{Y}_{obs},\mathbf{Y}_{mis}^{(t+1)})$에서 $\theta^{(t+1)}$을 추출한다. 위의 두 단계를 반복적으로 충분히 시행하면 궁극적으로 모수의 사후분포 $P(\theta,\mathbf{Y})$로부터 모수의 추출이 가능해진다.

표 **4.1.** 헬스클럽 자료

Weight	Waist	Pulse	Chin-up	Abdominal	High jump
191	36	50	5	162	60
189	37	52	2	110	60
193	38	58	12	101	101
162	35	62	12	105	37
189	35	46	13	155	58
182	36	56	4	101	42
211	38	56	8	101	38
167	34	60	6	125	40
176	31	74	15	200	40
154	33	56	17	251	250
169	34	50	17	120	38
166	33	52	13	210	115
154	34	64	14	215	105
247	46	50	1	50	50
193	36	46	6	70	31
202	37	62	12	210	120
176	37	54	4	60	25
157	32	52	11	230	80
156	33	54	15	225	73
138	33	68	2	110	43

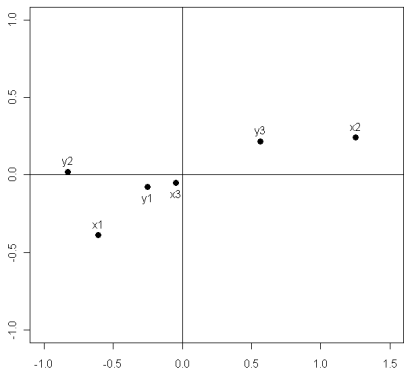

그림 **4.1.** 헬스클럽 자료의 정준상관 행렬도(결측률 0.0%)

4. 활용 사례

표 4.1은 헬스클럽의 중년회원 20명을 대상으로 몸무게(Weight), 허리둘레(Waist), 맥박수(Pulse), 턱걸이(Chin-up), 복근운동(Abdominal), 높이뛰기(High jump)에 대해 측정한 자료이다 (SAS Institute Inc., 1990, Chapter 15).

그림 4.1은 표 4.1의 신체적 변수군(몸무게(x_1), 허리둘레(x_2), 맥박수(x_3))과 체력적 변수군(턱걸이(y_1), 복근운동(y_2), 높이뛰기(y_3))의 정준상관 행렬도를 보여준다. 몸무게(x_1), 맥박수(x_3), 턱걸이(y_1), 복근운동(y_2)은 제 1축(dim1)에 대해 왼쪽편에 위치해 있고 그들이 이루는 사이의 각이 좁아 서

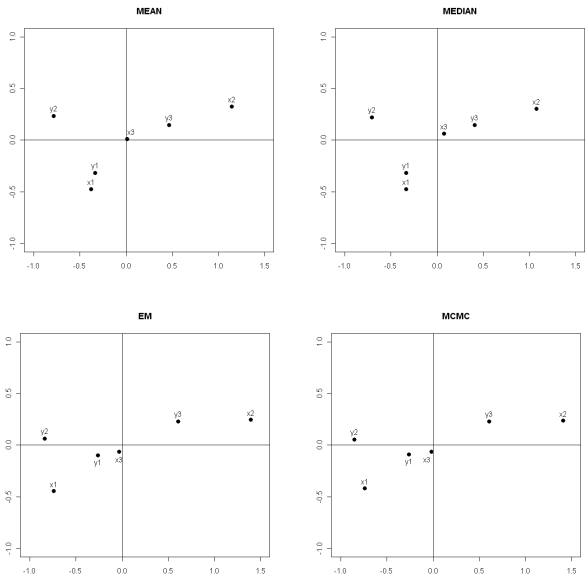

그림 4.2. 헬스클럽 자료의 정준상관 행렬도(결측률 5.0%)

로 상관이 높은 변수임을 보여준다. 허리둘레(x_2)와 높이뛰기(y_3)는 오른편에 위치해 있고, 그들이 이루는 사이의 각이 좁아 서로 상관이 높은 변수임을 보여준다. 외쪽 편의 변수군은 근력적인 부분의 의미가 강하며, 오른쪽의 변수군은 유연성과 관련된다고 볼 수 있다. 이러한 변수군 사이의 관계의 시각적인 측면에서, 원 자료에서 결측이 발생한 경우를 가정하여, 각각의 대체법에 따른 변수군의 관계에 대한 그림을 살펴보도록 한다. 결측은 임의로 5, 10, 15, 20%로 나누어 각각의 대체법에 따른 형상을 살펴보았다.

그림 4.2는 표 4.1의 헬스클럽 자료에서 결측률 5.0%의 상황 하에서 각각의 대체법을 적용한 정준상관 행렬도를 나타낸다. 원 자료의 정준상관 행렬도에서 몸무게(x_1), 맥박수(x_3), 턱걸이(y_1), 복근운동(y_2)은 왼쪽 편에 위치하고 있으나, 평균과 중위수대체의 결과에서는 그 변수들 중, 맥박수(x_3)가 오른쪽 편으로 이동한 결과를 보여주고 있다. 하지만 전체적인 형상에서는 평균, 중위수 대체는 원 자료의 정준상관 행렬도와 비슷한 형상을 유지하고 있는 것으로 나타났다. EM알고리즘법과 MCMC법을 이용한 다중대체의 경우, 원 데이터의 형상을 거의 비슷하게 나타내는 것으로 확인 할 수 있다.

그림 4.3은 표 4.1의 헬스클럽 자료에서 결측률 10.0%의 상황 하에서 각각의 대체법을 적용한 정준상관 행렬도를 나타낸다. 원 자료의 정준상관 행렬도에서 몸무게(x_1), 맥박수(x_3), 턱걸이(y_1), 복근운동(y_2)은 왼쪽 편에 위치하고 허리둘레(x_2)와 높이뛰기(y_3)는 오른쪽 편에 위치하나, 평균, 중위수대체를 통한 정준상관 행렬도에서는 맥박수(x_3), 복근운동(y_2)가 왼쪽 편에 위치하고 나머지는 오른편에 위치함을 보여준다. EM알고리즘법과 MCMC법을 통한 대체의 결과에서는 원 자료의 정준상관 행렬도의 변수들의 위치에 비하여 변동이 적게 나타났음을 알 수 있다. 복근운동(y_2), 맥박수(x_3)가 위로 올라간 형태로 나타났고, 턱걸이(y_1), 높이뛰기(y_3)는 아래로 내려간 형태로 나타났다.

그림 4.4는 표 4.1의 헬스클럽 자료에서 결측률 15.0%의 상황 하에서 각각의 대체법을 적용한 정준상관 행렬도를 나타낸다. 평균, 중위수대체를 통한 정준상관 행렬도에서는 원 자료의 정준상관 행렬도와 비교하여 변수들의 위치에 큰 변화가 생겼다. 턱걸이(y_1), 복근운동(y_2)가 왼쪽 편에 위치하고 나머지는

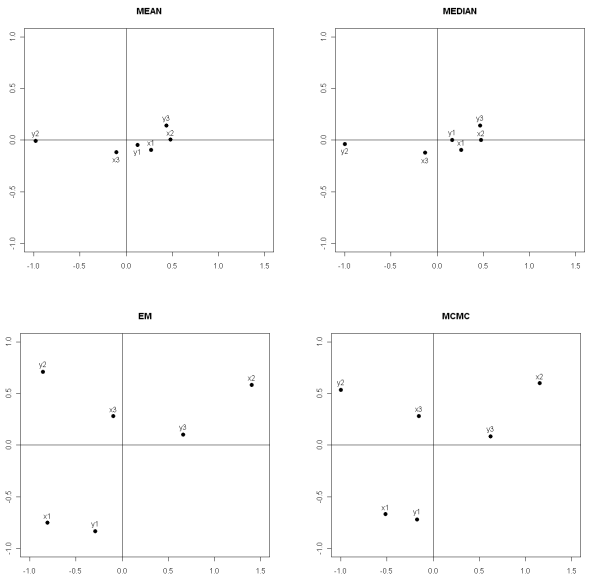

그림 4.3. 헬스클럽 자료의 정준상관 행렬도(결측률 10.0%)

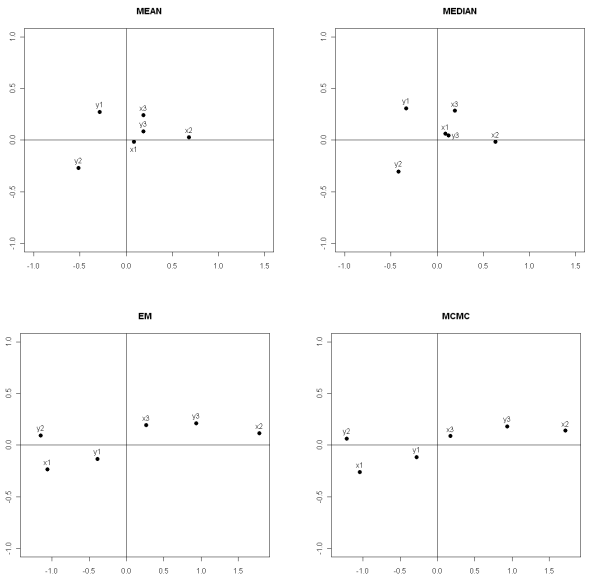

그림 4.4. 헬스클럽 자료의 정준상관 행렬도(결측률 15.0%)

오른쪽 편에 위치함을 나타내 형상변동이 커졌음을 의미한다. EM알고리즘법과 MCMC법을 이용한 다중대체의 결과에서는 기존의 형상은 유지하지만, 왼쪽 편에 있던 맥박수(x_3)가 오른쪽 편으로 위치했다. 결측률이 15%일 때, EM알고리즘법과 MCMC법을 이용한 다중대체의 결과는 어느 정도 원 자료의 형상에 근접하는 것을 확인할 수 있다.

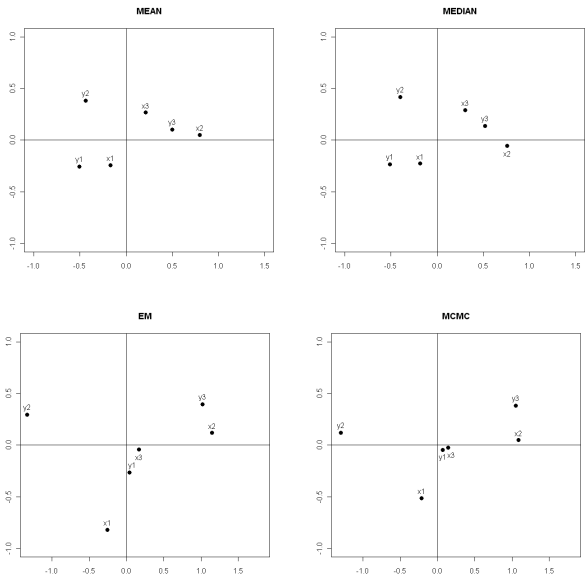

그림 4.5. 헬스클럽 자료의 정준상관 행렬도(결측률 20.0%)

표 4.2. 결측률에 따른 **X**, **Y** 개체군과 변수의 평균제곱근(RMS)

대체법	개체군				변수군			
	결측률				결측률			
	5.0%	10.0%	15.0%	20.0%	5.0%	10.0%	15.0%	20.0%
MEAN	0.182	0.361	0.437	0.511	0.134	0.380	0.365	0.291
MEDIAN	0.188	0.367	0.466	0.578	0.134	0.380	0.365	0.291
EM	0.136	0.357	0.394	0.486	0.058	0.353	0.267	0.273
MCMC	0.126	0.387	0.329	0.476	0.059	0.306	0.269	0.280

그림 4.5는 표 4.1의 헬스클럽 자료에서 결측률 20.0%의 상황 하에서 각각의 대체법을 적용한 정준상관 행렬도를 나타낸다. 평균과 중위수대체를 통한 정준상관 행렬도의 변수들의 형상은 비슷한 경향을 나타낸다. 맥박수(x_3)를 제외한 턱걸이(y_1), 복근운동(y_2)은 몸무게(x_1)는 원 자료의 행렬도와 같이 왼쪽 편에 위치함을 보인다. EM알고리즘법과 MCMC법을 이용한 다중대체의 결과에서는 기존의 형상과 비슷하게 나타났지만, 턱걸이(y_1), 맥박수(x_3)의 움직임이 오른쪽 편으로 위치함에 따라 변수간의 상관관계를 해석함에 있어 어려움을 나타내고 있다. 각각의 결측률에 따라 여러 대체법을 사용한 결과를 비교해보면, 대체적으로 EM알고리즘법과 MCMC법을 이용한 다중대체법의 결과가 원 자료의 정준상관 행렬도와 유사한 형상을 나타낸 것을 확인할 수 있었다. 평균제곱근(RMS)와 프로크러스티즈 통계량으로 각각의 변동성을 측정한 결과를 살펴보자. 표 4.2는 결측률에 따른 개체군과 변수의 RMS를 나타내고 있다.

표 4.2에 따르면 각각의 결측률에 따라 전반적으로 EM알고리즘법과 MCMC법에서 RMS가 낮게 나타나고 있음을 확인 할 수 있다. 또한 프로크러스티즈 통계량을 보여주는 표 4.3에서도 평균과 중위수대체법에 비하여 EM알고리즘법 그리고 MCMC법을 통한 다중대체법이 각각의 결측률에 따라 형상변동이 작게 일어났음을 보여준다. RMS의 결과와 프로크러스티즈 통계량(PS)값을 동시적으로 확인한 결

표 **4.3.** 결측률에 따른 프로크러스티즈 통계량

대체법	결측률			
	5.0%	10.0%	15.0%	20.0%
MEAN	0.208	1.732	1.468	0.950
MEDIAN	0.271	1.780	1.694	1.081
EM	0.041	1.496	0.767	0.896
MCMC	0.040	1.087	0.796	0.943

과 EM알고리즘법과 MCMC법을 통한 다중대체에서 정준상관 행렬도의 변동이 작았음을 알 수 있었다. 각각의 결측률에 따른 형상의 변동을 나타낸 그림과 RMS, 프로크러스티즈 통계량을 확인한 결과 EM알고리즘법과 MCMC법에서 평균과 중위수대체의 방법을 통한 결과보다 원 자료의 정준상관 행렬도와의 형상변동이 적었음을 알 수 있었다.

참고문헌

최용석 (2006). <행렬도 분석>, 부산대학교 기초과학연구원, 부산대학교 출판부, 83–86.

최용석, 현기흥 (2006). <통계적 형상분석의 이해와 응용>, 자유아카데미, 서울.

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

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

Choi, Y. S. (1991). *Resistant Principal Component Analysis*, *Biplot and Corresponding Analysis*, 고려대학교, 박사학위 논문, 서울.

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

Kim, J. G., Choi, Y. S. and Lee, N. Y. (2010a). Unbalanced ANOVA for testing shape variability in statistical shape analysis, *The Korean Journal of Applied Statistics*, **23**, 317–323.

Kim, J. G., Choi, Y. S. and Shin, S. M. (2010b). Shape variability and classification using PS, MPS and RMS in statistical shape analysis, *Far East Journal of Applied Mathematics*, **42**, 49–60.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, Willey, New York.

Park, M. R. and Huh, M. H. (1996). Canonical correlation biplot, *Journal of the Korea Statistical Society*, **3**, 11–19.

Rubin, D. (1987). *Multiple Imputation for Nonresponse in Survey*, Wiley & Sons, New York.

SAS Institute Inc. (1990). *SAS/STAT User's Guide*, **1**, 4/e, SAS Institute Inc., Cary NC.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*, Chapman & Hall, London.

Shin, S. M., Choi, Y. S. and Lee, N. Y. (2008). Comparison of shape variability in principal component biplot with missing values, *The Korean Journal of Applied Statistics*, **21**, 1109–1116.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distribution by data augmentation, *Journal of the American Statistical Association*, **82**, 528–540.

A Study on Shape Variability in Canonical Correlation Biplot with Missing Values

Hyun Uk Hong[1] · Yong-Seok Choi[2] · Sang Min Shin[3] · Chang Wan Kang[4]

[1]Department of Statistics, Pusan National University
[2]Department of Statistics, Pusan National University
[3]Department of Statistics, Pusan National University
[4]Department of Data Information Science, Dongeui University

(Received June 2010; accepted August 2010)

Abstract

Canonical correlation biplot is a useful biplot for giving a graphical description of the data matrix which consists of the association between two sets of variables, for detecting patterns and displaying results found by more formal methods of analysis. Nevertheless, when some values are missing in data, most biplots are not directly applicable. To solve this problem, we estimate the missing data using the median, mean, EM algorithm and MCMC imputation methods according to missing rates. Even though we estimate the missing values of biplot of incomplete data, we have different shapes of biplots according to the imputation methods and missing rates. Therefore we use a RMS(root mean square) which was proposed by Shin *et al.* (2007) and PS(procrustes statistic) for measuring and comparing the shape variability between the original biplots and the estimated biplots.

Keywords: Canonical correlation biplot, shape variability, procrustes, missing mechanism, imputation methods.

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