The Korean Journal of Applied Statistics (2012)
25*(3), 521–529*
DOI: http://dx.doi.org/10.5351/KJAS.2012.25.3.521

Semi-Partial Canonical Correlation Biplot

Bo-Hui Lee[1] · Yong-Seok Choi[2] · Sang-Min Shin[3]

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

(Received January 25, 2012; Revised February 12, 2012; Accepted April 13, 2012)

Abstract

Simple canonical correlation biplot is a graphical method to investigate two sets of variables and observations in simple canonical correlation analysis. If we consider the set of covariate variables that linearly affects two sets of variables, we can apply the partial canonical correlation biplot in partial canonical correlation analysis that removes the linear effect of the set of covariate variables on two sets of variables. On the other hand, we consider the set of covariate variables that linearly affect one set of variables but not the other. In this case, if we apply the simple or partial canonical correlation biplot, we cannot clearly interpret other two sets of variables. Therefore, in this study, we will apply the semi-partial canonical correlation analysis of Timm (2002) and remove the linear effect of the set of covariate variables on one set of variables but not the other. And we suggest the semi-partial canonical correlation biplot for interpreting the semi-partial canonical correlation analysis. In addition, we will compare shapes and shape the variabilities of the simple, partial and semi-partial canonical correlation biplots using a procrustes analysis.

Keywords: Biplot, set of covariate variables, semi-partial canonical correlation analysis, Procrustes analysis.

1. 서론

Gabriel (1971)에 의해서 개발된 행렬도(biplot)는 이원표 자료행렬(two-way data matrix)의 행과 열을 한 그림에 동시에 나타내는 탐색적 기법이다. 행렬도에 대한 연구는 다양한 분야에서 활발하게 진행되어 왔는데, 특히 Park과 Huh (1996a, 1996b)는 두 변수 집단 간의 통계적 관계를 탐색하기 위한 다변량 분석 방법인 정준상관분석(canonical correlation analysis)의 수량화 방법(quantification method) 관점을 이용하여 정준상관 행렬도(canonical correlation biplot)를 제안하였고, 세 변수 집단 이상인 경우까지 확장한 정준상관분석의 일반화를 시도하였다.

최근에 정준상관 행렬도를 활용하여 Choi과 Choi (2008)는 2006년도 한국여자골프협회(KLPGA) 선수 중 상금 순위 상위 50명을 대상으로 기술요인과 경기성적요인간의 관련성을 살펴보고 군집분석을 활용하여 각 선수들의 군집을 시도하였고, Choi 등 (2009)은 테니스 그랜드슬램대회의 선수특성요인과 경기요인에 대한 정준상관 행렬도에서 프로크러스티즈 분석(Procrustes analysis)을 통하여 행렬도의 형상을 비교하였다. 또한 Choi과 Choi (2010)는 2004년 대한테니스협회(KTA)에 등록된 랭킹

[2]Corresponding author: Professor, Department of Statistics, Pusan National University, Jangjeon-Dong, Geumjeong-Gu, Busan 609-735, Korea. E-mail: yschoi@pusan.ac.kr

100위권 이내의 선수 50명을 대상으로 체격요인, 체력요인 그리고 기초기술요인의 세 변수군이 존재하는 경우, 이들의 상호 연관성을 살펴보기 위해 일반화 정준상관분석을 이용하여 일반화 정준상관 행렬도(generalized canonical correlation biplot; GCCB)를 고려하였다. 더 나아가 Yeom과 Choi (2011)는 두 변수군에 선형적 영향을 미치는 공변량변수(covariate variables)로 이루어진 한 변수군이 존재하는 경우, 공변량변수군의 선형적 영향을 제거한 두 변수군에 대한 편정준상관분석을 이용하여 편정준상관 행렬도(partial canonical correlation biplot; PCCB)를 제안하고 응용의 예를 보였다.

이와 달리 공변량변수군이 하나의 변수군에는 영향을 주지만 다른 한 변수군에는 영향을 주지 않는 경우, 단순정준상관 행렬도(simple canonical correlation biplot; SCCB)나 일반화 정준상관 행렬도, 혹은 편정준상관 행렬도를 적용한다면 주 관심 대상인 두 변수군에 대하여 잘못 해석할 수 있다. 이러한 경우에는 영향을 미치는 한 변수군에 대해서만 공변량변수군의 효과를 제거한 뒤 두 변수군에 대한 준편정준상관분석을 활용하여야 할 것이다.

따라서 본 연구에서는 준편정준상관분석을 응용하여 이를 그래프적으로 살펴볼 수 있는 준편정준상관 행렬도(semi-partial canonical correlation biplot; SPCCB)를 제안하고자 한다. 이에 2절에서는 단순정준상관 행렬도와 편정준상관 행렬도, 그리고 준편정준상관 행렬도의 기초 이론에 대해 설명하고, 행렬도의 형상변동 차이를 비교하기 위해 활용한 프로크러스티즈 분석을 추가적으로 소개하려 한다. 3절에서는 준편정준상관 행렬도의 활용 사례를 보이고, 단순정준상관 행렬도와 편정준상관 행렬도의 결과와 이들의 형상변동 차이를 비교하고자 한다.

2. 준편정준상관 행렬도와 프로크러스티즈 분석

2.1. 단순정준상관 행렬도

이 절에서는 Choi (2006, Chapter 2)를 참고로 하여 기존의 단순정준상관 행렬도를 소개하기로 하자. 단순정준상관분석은 두 변수군 사이의 관계를 분석하는 다변량 기법이다. 일반적으로 이 기법은 두 변수군의 선형 결합간의 상관에 관심을 두며, 이 상관관계를 가장 크게 만드는 알고리즘을 통해 그 관계를 분석하는 것이 목적이다.

먼저 p개의 변수와 q개의 변수로 이루어진 두 변수군 $\mathbf{x} = (x_1, \ldots, x_p)'$와 $\mathbf{y} = (y_1, \ldots, y_q)'$는 각각 평균벡터 $\mu_{\mathbf{x}} = (\mu_{x_1}, \ldots, \mu_{x_p})'$와 $\mu_{\mathbf{y}} = (\mu_{y_1}, \ldots, \mu_{y_q})'$를 가지며 공분산행렬 Σ_{xx}, Σ_{yy}, $\Sigma_{xy} = \Sigma_{yx}'$를 가지는 확률벡터라 하자. 그러면 이들에 의해 측정된 n명의 자료에서 표본공분산행렬은 다음과 같다.

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

여기서 임의의 계수벡터 $\mathbf{u}$와 $\mathbf{v}$에 대한 두 변수군의 선형결합 $\hat{x}$과 $\hat{y}$을 각각

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

라 할 때, 이들은 다차원의 변수군을 1차원으로 축소하여 나타내며, 이들의 상관은

$$\hat{\rho}_{\hat{x}\hat{y}} = \frac{\sum_{i=1}^{n} \hat{x}_i \hat{y}_i}{\sqrt{\sum_{i=1}^{n} \hat{x}_i^2} \sqrt{\sum_{i=1}^{n} \hat{y}_i^2}} = \frac{\mathbf{u}'\mathbf{S}_{xy}\mathbf{v}}{\sqrt{\mathbf{u}'\mathbf{S}_{xx}\mathbf{u}}\sqrt{\mathbf{v}'\mathbf{S}_{yy}\mathbf{v}}} \tag{2.1}$$

으로 정의될 수 있다. 여기서 계수벡터 $\mathbf{u}$와 $\mathbf{v}$는 식 (2.1)의 두 선형결합의 상관을 최대화하는 알고

리즘에 의해서 구할 수 있으며, 이는 $\hat{x}$과 $\hat{y}$의 분산이 1인 제약 조건 $\mathbf{u}'\mathbf{S}_{xx}\mathbf{u} = 1$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. 이 알고리즘은 라그랑즈 승수법(Lagrange multiplier method)을 이용하면 고유체계(eigensystem) 문제로 유도하여 풀 수 있고, 또한 단순정준계수벡터와 단순정준상관을 대수적으로 한꺼번에 제공하는 비정칙치분해(singular value decomposition)

$$\mathbf{S}_{xx}^{-\frac{1}{2}}\mathbf{S}_{xy}\mathbf{S}_{yy}^{-\frac{1}{2}} = \mathbf{UDV}' \tag{2.2}$$

를 이용하면 간편하게 구할 수 있다. 여기서 $\mathbf{S}_{xx}^{-1/2}\mathbf{S}_{xy}\mathbf{S}_{yy}^{-1/2}$의 계수(rank)는 $r(\leq \min(p,q))$이고, $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_r)$와 $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_r)$는 크기가 각각 $p \times r$과 $q \times r$인 단순정준계수벡터로 이루어진 직교행렬이다. 그리고 $\mathbf{D} = \text{diag}\left(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_r}\right)$는 $\sqrt{\lambda_1} \geq \cdots \geq \sqrt{\lambda_r} > 0$의 관계를 갖는 비정칙치를 대각원소로 하는 대각행렬이고, 이 비정칙치가 단순정준상관에 해당한다.

따라서 두 변수군 $\mathbf{x}$와 $\mathbf{y}$에 대해 n명을 측정한 크기가 각각 $n \times p$와 $n \times q$인 중심화 자료행렬을 $\mathbf{X}$와 $\mathbf{Y}$라 하면, i번째 단순정준상관계수벡터는 각각 다음과 같고,

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

이들에 의해서 구성된 단순정준상관계수행렬은 $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_r)$와 $\mathbf{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_r)$로 정의될 수 있다. 이를 통해 계산된 중심화 자료행렬 $\mathbf{X}$와 $\mathbf{Y}$에 대한 단순정준상관 행렬도의 행 좌표행렬 $\mathbf{R}_\mathbf{X}, \mathbf{R}_\mathbf{Y}$와 열 좌표행렬 $\mathbf{C}_\mathbf{X}, \mathbf{C}_\mathbf{Y}$는 각각 다음과 같다.

$$\begin{aligned} \mathbf{R}_\mathbf{X} &= \mathbf{XAD}, \qquad \mathbf{C}_\mathbf{X} = \mathbf{AD}, \\ \mathbf{R}_\mathbf{Y} &= \mathbf{YBD}, \qquad \mathbf{C}_\mathbf{Y} = \mathbf{BD}. \end{aligned}$$

2.2. 준편정준상관 행렬도

이 절에서는 Yeom과 Choi (2011)의 편정준상관 행렬도를 간략히 언급하고, Timm (2002, Chapter 8)을 참고로 하여 준편정준상관분석의 기초 이론을 요약하고, 이를 위한 시각적 도구인 준편정준상관 행렬도를 제안하려 한다.

먼저 2.1절의 단순정준상관분석에서 두 변수군 $\mathbf{x} = (x_1, \ldots, x_p)'$와 $\mathbf{y} = (y_1, \ldots, y_q)'$에 선형적 영향을 미치는 m개의 변수로 이루어진 공변량변수군 $\mathbf{z} = (z_1, \ldots, z_m)'$가 존재하는 경우, 이것의 영향을 제거한 두 변수군 $\mathbf{x}$와 $\mathbf{y}$에 대한 관계를 살펴보는 편정준상관분석을 이용해야 한다. 그러나 때로는 공변량변수군 $\mathbf{z}$가 하나의 변수군 $\mathbf{x}$에는 영향을 주지만 다른 한 변수군 $\mathbf{y}$에는 영향을 주지 않는 자료를 접할 수 있다. 이 경우 단순정준상관분석이나 편정준상관분석을 이용한다면 주 관심 대상인 두 변수군에 대하여 잘못된 해석을 할 수 있어 영향을 받는 한 변수군에서만 공변량변수군의 영향을 제거한 후 두 변수군의 관계를 파악하는 준편정준상관분석을 살펴보고자 한다.

세 변수군 $\mathbf{x}, \mathbf{y}, \mathbf{z}$는 각각 평균벡터로 $\mu_\mathbf{x} = (\mu_{x_1}, \ldots, \mu_{x_p})'$, $\mu_\mathbf{y} = (\mu_{y_1}, \ldots, \mu_{y_q})'$, $\mu_\mathbf{z} = (\mu_{z_1}, \ldots, \mu_{z_m})'$를 가지며, 공분산행렬 Σ_{xx}, Σ_{yy}, Σ_{zz}, $\Sigma_{xy} = \Sigma_{yx}'$, $\Sigma_{xz} = \Sigma_{zx}'$, $\Sigma_{yz} = \Sigma_{zy}'$를 가지는 확률벡터이다. 이들에 의해 측정된 n명의 자료에 대하여 표본공분산행렬은

$$\mathbf{S} = \begin{pmatrix} \mathbf{S}_{xx} & \mathbf{S}_{xy} & \mathbf{S}_{xz} \\ \mathbf{S}_{yx} & \mathbf{S}_{yy} & \mathbf{S}_{yz} \\ \mathbf{S}_{zx} & \mathbf{S}_{zy} & \mathbf{S}_{zz} \end{pmatrix} \tag{2.3}$$

이다. 여기서 편정준상관분석의 경우, 변수군 $\mathbf{z}$가 공변량변수군이므로 두 변수군에서 공변량변수군 $\mathbf{z}$의 효과를 제거한 조건부 표본공분산행렬

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

를 이용하는 것이 바람직하며, 편정준상관 행렬도의 대수적인 이론은 준편정준상관분석과 유사하므로 생략하기로 한다.

다음으로 공변량변수군 $\mathbf{z}$가 한 변수군 $\mathbf{x}$에만 영향을 주는 경우, 식 (2.3)으로부터 영향을 받는 변수군 $\mathbf{x}$에서만 공변량변수군 $\mathbf{z}$의 효과를 제거한 조건부 표본공분산행렬은

$$\begin{aligned}\mathbf{S}_{(x.z)y} &= \begin{pmatrix}\mathbf{S}_{xx.z} & \mathbf{S}_{xy.z}\\ \mathbf{S}_{yx.z} & \mathbf{S}_{yy}\end{pmatrix}\\ &= \begin{pmatrix}\mathbf{S}_{xx}-\mathbf{S}_{xz}\mathbf{S}_{zz}^{-1}\mathbf{S}_{zx} & \mathbf{S}_{xy}-\mathbf{S}_{xz}\mathbf{S}_{zz}^{-1}\mathbf{S}_{zy}\\ \mathbf{S}_{yx}-\mathbf{S}_{yz}\mathbf{S}_{zz}^{-1}\mathbf{S}_{zx} & \mathbf{S}_{yy}\end{pmatrix}\end{aligned}$$

이며, 이 조건부 표본공분산행렬을 이용한 정준상관분석을 준편정준상관분석이라 한다. 이때 두 선형결합의 상관을 최대화하는 알고리즘은 2.1절의 단순정준상관분석처럼 대수적으로 비정칙치분해

$$\mathbf{S}_{xx.z}^{-\frac{1}{2}}\mathbf{S}_{xy.z}\mathbf{S}_{yy}^{-\frac{1}{2}} = \mathbf{U}^*\mathbf{D}^*\mathbf{V}^{*\prime} \tag{2.4}$$

를 이용하면 된다. 여기서 $\mathbf{S}_{xx.z}^{-1/2}\mathbf{S}_{xy.z}\mathbf{S}_{yy}^{-1/2}$의 계수는 $r(\leq \min(p,q))$이고, 크기가 각각 $p\times r$과 $q\times r$인 $\mathbf{U}^* = (\mathbf{u}_1^*,\ldots,\mathbf{u}_r^*)$와 $\mathbf{V}^* = (\mathbf{v}_1^*,\ldots,\mathbf{v}_r^*)$는 준편정준계수벡터의 직교행렬이다. 그리고 대각행렬 $\mathbf{D}^* = \mathrm{diag}\left(\sqrt{\lambda_1^*},\ldots,\sqrt{\lambda_r^*}\right)$는 $\sqrt{\lambda_1^*}\geq\cdots\geq\sqrt{\lambda_r^*}>0$의 관계를 갖는 준편정준상관을 대각원소로 하고 있다.

따라서 두 변수군 $\mathbf{x}$와 $\mathbf{y}$에 대해 측정한 n명의 중심화 자료행렬 $\mathbf{X}$와 $\mathbf{Y}$를 각각 크기가 $n\times p$와 $n\times q$라 하면, i번째 준편정준상관계수벡터는 각각 다음과 같다.

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

여기서 이들로 구성된 준편정준상관계수행렬은 $\mathbf{A}^* = (\mathbf{a}_1^*,\ldots,\mathbf{a}_r^*)$와 $\mathbf{B}^* = (\mathbf{b}_1^*,\ldots,\mathbf{b}_r^*)$가 된다. 따라서 중심화 자료행렬 $\mathbf{X}$와 $\mathbf{Y}$에 대한 행 좌표행렬 $\mathbf{R}_{\mathbf{X}}^*, \mathbf{R}_{\mathbf{Y}}^*$와 열 좌표행렬 $\mathbf{C}_{\mathbf{X}}^*, \mathbf{C}_{\mathbf{Y}}^*$는 각각

$$\begin{aligned}\mathbf{R}_{\mathbf{X}}^* &= \mathbf{X}\mathbf{A}^*\mathbf{D}^*, \qquad \mathbf{C}_{\mathbf{X}}^* = \mathbf{A}^*\mathbf{D}^*\\ \mathbf{R}_{\mathbf{Y}}^* &= \mathbf{Y}\mathbf{B}^*\mathbf{D}^*, \qquad \mathbf{C}_{\mathbf{Y}}^* = \mathbf{B}^*\mathbf{D}^*\end{aligned}$$

로 정의되며, 이들에 의해 나타나는 행렬도를 준편정준상관 행렬도라 하자.

2.3. 프로크러스티즈 분석

이 절에서는 Choi과 Hyun (2006, Chapter 2), Gower와 Dijksterhuis (2004, Chapter 4), Choi 등 (2005)을 참고로 하여 행렬도의 형상변동 차이를 비교하기 위해 활용한 프로크러스티즈 분석에 대하여 소개하려 한다.

프로크러스티즈 분석이란 기하적 공간상에서 개체간의 형상비교를 위해 한 개체를 다른 개체 쪽으로 적

Table 3.1. Multiple correlation coefficients for three sets of variables

	Competition score factors				Skill factors			
	200m	1km	Winning rate	Average score	Front racing	Overtaking	Bending back	Marking
Physique factors	0.482	0.529	0.435	0.465	0.303	0.274	0.343	0.371
p-value	0.065	0.022	0.152	0.089	0.633	0.743	0.468	0.355

합시키는 방법이다. 자료행렬 $\mathbf{X}$에 대한 두 종류의 행렬도를 위한 각각의 크기가 $p \times s$인 중심화 좌표행렬을 $\mathbf{C_X}$와 $\mathbf{C_X^*}$라 하면, 이들간의 제곱 유클리드거리는 다음과 같다.

$$O(\mathbf{C_X}, \mathbf{C_X^*}) = \left\|\mathbf{C_X^*} - \mathbf{C_X}\boldsymbol{\Gamma} - \mathbf{1}_p\mathbf{t}'\right\|^2. \tag{2.5}$$

여기서 $\|\mathbf{C}\| = \sqrt{\mathrm{tr}(\mathbf{C}'\mathbf{C})}$이고, $\boldsymbol{\Gamma}$는 크기가 $s \times s$인 직교회전행렬(orthogonal rotation matrix)이다. 그리고 $\mathbf{1}_p$는 모든 원소가 1인 $p \times 1$벡터이며, $\mathbf{t}$는 $s \times 1$인 위치모수(location parameter)벡터이다. 그러면 비정칙치분해 $\mathbf{C_X^*}'\mathbf{C_X} = \mathbf{P\Lambda Q}'$를 이용하여 식 (2.5)를 최소화하는 $\hat{\boldsymbol{\Gamma}} = \mathbf{QP}'$와 $\hat{\mathbf{t}} = \mathbf{0}$을 찾을 수 있고, 이는 두 형상 $\mathbf{C_X}$와 $\mathbf{C_X^*}$가 잘 일치되도록 하게하는 정보를 제공한다.

이를 이용하여 두 형상의 변동성을 평가하는 측도인 프로크러스티즈 통계량

$$\mathrm{PS}(\mathbf{C_X}, \mathbf{C_X^*}) = \mathrm{tr}(\mathbf{C_X'C_X}) + \mathrm{tr}(\mathbf{C_X^*}'\mathbf{C_X^*}) - 2\mathrm{tr}(\boldsymbol{\Lambda})$$

를 얻을 수 있고, 일반적으로 두 행렬도의 형상이 일치하는 경우 프로크러스티즈 통계량 값이 0이 되어 변동차이가 없음을 의미한다.

3. 활용 사례

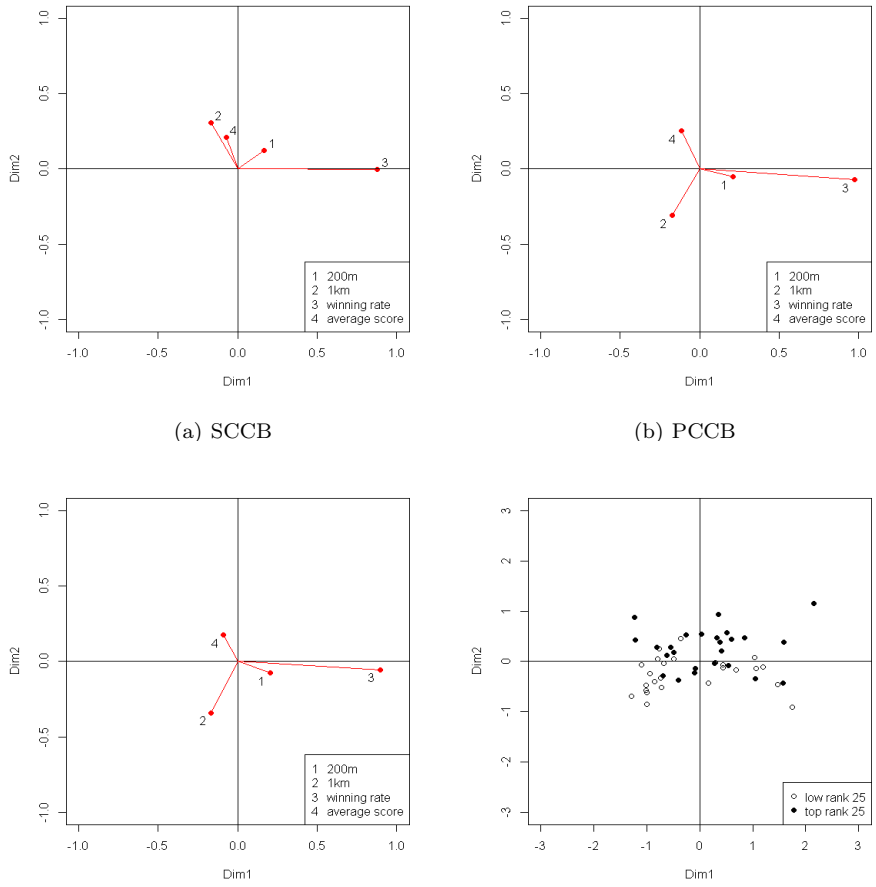
경륜 자료는 경륜운영본부 홈페이지(http://www.kcycle.or.kr)에서 제공하는 자료로 2011년 기준 전체 성적 순위 100위권 이내의 선수들 중 결측값이 없는 상위 선수 50명에 대해 정리한 것이다. 해석의 편의상 단순정준상관 행렬도, 편정준상관 행렬도, 그리고 준편정준상관 행렬도를 1절에서 언급한 바와 같이 각각 SCCB, PCCB, SPCCB로 나타내기로 하겠다.

본 연구에서는 경기기록요인 $\mathbf{x}$ = {200m 기록(초), 1km 기록(초), 승률(%), 평균득점}, 기술요인 $\mathbf{y}$ = {선행, 추입, 젖히기, 마크}, 체격요인 $\mathbf{z}$ = {나이, 신장(cm), 체중(kg), 가슴둘레(cm), 대퇴(cm), 하퇴(cm)}의 세 변수군으로 나누고, 이들의 다중상관계수(multiple correlation coefficient)를 살펴보았다. Table 3.1에서 체격요인($\mathbf{z}$)과 경기기록요인($\mathbf{x}$)의 상관계수는 비록 승률이 다소 상관이 떨어지나 전체적으로 모두 0.4 이상으로 유의한 상관이 존재하고, 체격요인($\mathbf{z}$)과 기술요인($\mathbf{y}$)의 상관계수는 p-값이 모두 0.3 이상으로 유의하지 않게 나타났다. 따라서 체격요인을 경기기록요인에 더 큰 영향을 미치는 공변량변수군으로 고려하는 것이 옳다고 판단하고, 경기기록요인에서만 체격요인의 효과를 제거한 후 경기기록요인과 기술요인의 두 변수군의 관계를 살펴보는 SPCCB를 적용할 것이다. 덧붙여, SCCB와 PCCB와의 비교를 통해 그 차이점을 설명하고자 한다.

기술요인 변수를 좀 더 자세히 설명하면 경륜은 바람에 의한 선두선수의 불리함을 피하기 위해 선두유도원이 선수들 앞에서 일정구간 끌어주다가 퇴장하는데, 이때 선두유도원이 퇴장한 후 마지막 코너 전에 선두로 나서는 전법이 선행이다. 추입은 다른 선수들 뒤에서 바람을 피해 달리다가 마지막 직선코스에서 역전하는 전법이고, 젖히기는 마지막 바퀴까지 중간이나 끝에 있다가 순간적으로 젖히고 나가는 전법이다. 마크는 자신의 기량이 부족하거나 상대의 능력이 강하다고 생각될 때 1등을 포기하고 2, 3등을 노리는 전법이다.

Table 3.2. Goodness-of-fits of the approximation for three biplots

	Canonical axis	Canonical correlation	Goodness-of-fits(%)	Cumulative Goodness-of-fits(%)
SCCB	1	0.887	82.26	93.51
	2	0.328	11.25	
PCCB	1	0.883	73.16	88.14
	2	0.399	14.98	
SPCCB	1	0.820	72.20	88.13
	2	0.385	15.93	

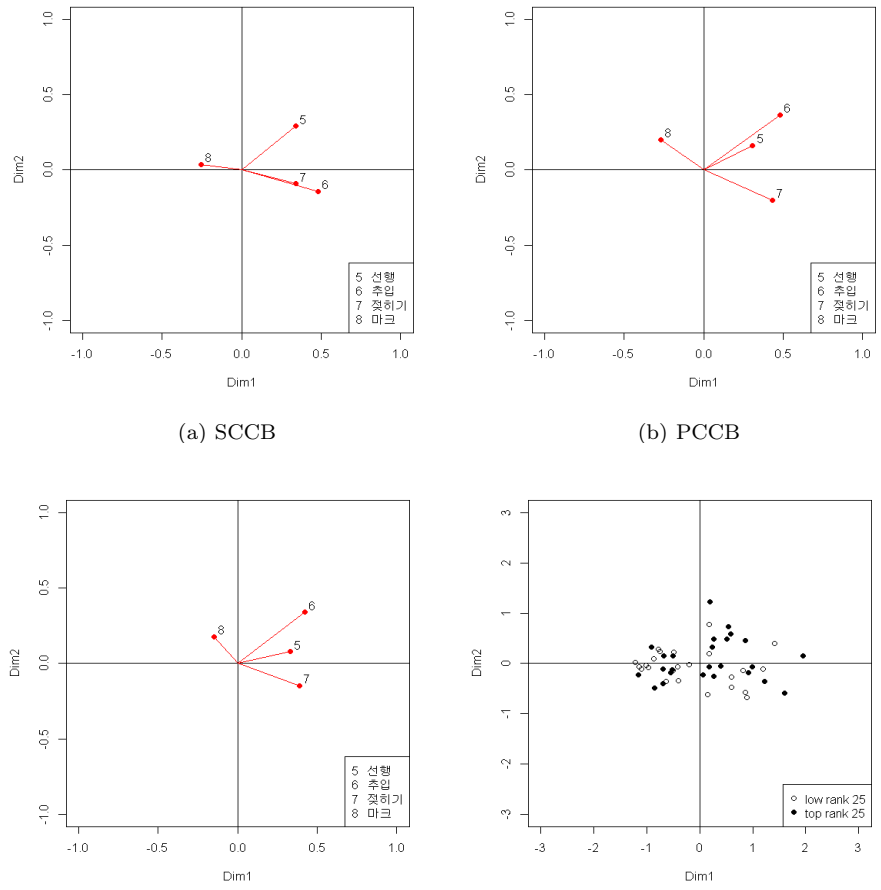


(a) SCCB (b) PCCB

(c) SPCCB (d) Total players (SPCCB)

Figure 3.1. Biplots for competition score factors: SCCB(simple canonical correlation biplot), PCCB(partial canonical correlation biplot), SPCCB(semi-partial canonical correlation biplot)

먼저 Figure 3.1과 Figure 3.2의 (a)는 경기기록요인 변수군과 기술요인 변수군에 대한 2차원 SCCB이고, (b)는 공변량변수군인 체격요인 변수군의 영향을 제거한 경기기록요인 변수군과 기술요인 변수군에



(a) SCCB (b) PCCB

(c) SPCCB (d) Total players (SPCCB)

Figure 3.2. Biplots for skill factors: SCCB(simple canonical correlation biplot), PCCB(partial canonical correlation biplot), SPCCB(semi-partial canonical correlation biplot)

대한 2차원 PCCB이다. 그리고 (c)는 경기기록요인 변수군에서만 공변량변수군의 영향을 제거한 두 변수군에 대한 2차원 SPCCB이다. 이 세 행렬도의 근사적합도(goodness-of-fit of the approximation)는 Table 3.2에 나타나있으며, 이는 식 (2.2)와 (2.4)에서 얻어지는 비정칙치의 제곱의 총합에 대한 처음 s개 비정칙치의 제곱의 합이 차지하는 비율로 나타낸다. 값을 살펴보면 SCCB는 93.51%, PCCB는 88.14%, SPCCB는 88.13%로, 전체적으로 2차원 행렬도의 적합도는 80% 이상으로 원자료를 잘 설명한다고 여겨진다.

행렬도의 기하적 해석은 Choi (2006, 1장)에 자세히 설명되어 있으며, 이를 참고로 하여 Figure 3.1에서 경기기술요인 변수군에 대한 (a) SCCB, (b) PCCB 그리고 (c) SPCCB를 비교해보면 특히 (a) SCCB에서 다른 경향을 보인다. 먼저 200m 기록과 1km 기록이 평균득점과 갖는 관계를 살펴보면 (a) SCCB에서는 양의 상관을 보이는 반면, (b) PCCB에서는 음의 상관을 보이며 (c) SPCCB에서는 음의 상관이 조금 더 커지는 것을 알 수 있다. 이는 200m와 1km 기록 시간이 낮을수록 평균득점이 높아지게

Table 3.3. Procrustes statistics for comparing shape variabilities of three biplots

	Competition score factors			Skill factors		
	SCCB	PCCB	SPCCB	SCCB	PCCB	SPCCB
SCCB	0.000			0.000		
PCCB	0.228	0.000		0.294	0.000	
SPCCB	0.152	0.014	0.000	0.275	0.031	0.000

되므로 음의 상관을 갖는 것이 바람직하다고 할 수 있다. 그리고 200m 기록과 1km 기록의 관계를 살펴보면 두 사이각의 코사인값이 (a) SCCB에서는 0.127로 약한 양의 상관을, (b) PCCB에서는 −0.273으로 음의 상관을, (c) SPCCB에서는 −0.096으로 거의 무관하게 나타난다. 이는 200m는 단거리 기록이고 1km는 장거리 기록이므로 두 변수는 독립적으로 나타나는 것이 올바르다. 따라서 공변량변수군을 전혀 고려하지 않은 (a) SCCB는 변수들 사이의 상관관계를 잘못 해석하게 되고, (b) PCCB와 (c) SPCCB는 거의 비슷한 결과를 보인다. 다만 (c) SPCCB에서 조금 더 바르게 해석이 가능하며, (b) PCCB에 비해 각 열 좌표점 벡터들의 길이가 짧아져 변수들의 분산이 작아짐을 알 수 있다. 또한 승률과 평균득점의 관계는 일반적으로 양의 상관을 가질 것 같아 보이지만, 경륜은 기록이 아닌 순위 경주이기 때문에 세 행렬도에서 음의 상관을 보인다. 즉, 승률은 1등 횟수만으로 계산되어지지만, 평균득점은 1등과 7등을 번갈아가며 하는 기복이 심한 선수보다 3등과 4등의 횟수가 많은 선수가 더 높게 나오기 때문이다.

Figure 3.2에서 기술요인 변수군에 대한 (a) SCCB, (b) PCCB 그리고 (c) SPCCB를 비교하면 모두가 조금 다른 경향을 보인다. 먼저 세 행렬도 모두 동일하게 제1정준축(Dim1)의 오른편에는 선두나 역전을 통해 1등을 노리는 선행, 추입, 젖히기가 존재하고, 왼편에는 2, 3등을 노리는 마크가 존재한다. (a) SCCB에서는 추입과 젖히기가 강한 양의 상관을 가지며, (b) PCCB에서는 선행과 추입이 강한 양의 상관을 가진다. 반면에, (c) SPCCB에서는 선행, 추입, 젖히기의 세 변수가 특정한 변수와 유사하게 나타나기보다 모든 변수가 적절히 높은 상관을 가진다. 이는 경륜 기술들을 각각 독립적인 변수로 더욱 확실하게 구분하여 주는 (c) SPCCB의 해석이 더 명확하며, 마찬가지로 각 열 좌표점 벡터들의 길이가 짧아져 변수들의 분산이 작아짐을 알 수 있다.

Figure 3.1과 Figure 3.2의 (d)는 SPCCB에서 두 변수군에 대한 전체 선수의 행렬도이고, 이는 SCCB나 PCCB에서도 거의 같은 결과를 나타낸다. 먼저 Figure 3.1의 (d)에서는 순위를 결정하는 평균득점이 존재하는 위쪽에 주로 상위랭킹 선수들이 나타나고, 아래쪽에 하위랭킹 선수들이 나타나고 있다. 이와 달리 Figure 3.2의 (d)에서는 상위랭킹 선수와 하위랭킹 선수가 섞여있는 것을 볼 수 있는데, 자료를 살펴보면 제1정준축의 왼편에 중위권 선수들이 위치하고 있음을 확인할 수 있다. 이들은 마크와 같은 방향에 위치하고 있으며, 1등보다는 2, 3등을 많이 하여 승률은 낮지만 득점은 비교적 높은 선수들이다.

Figure 3.1과 Figure 3.2의 전체 행렬도를 비교해서 살펴보면 제1정준축의 오른편에는 선행, 추입, 젖히기가 스퍼트를 내어 1등을 노리는 전법이므로 200m 기록, 승률과 관련이 높고, 왼편에는 마크가 지속적으로 2, 3등을 노리는 전법으로 1km 기록, 평균득점과 관련이 높게 나타난다.

끝으로 Table 3.3의 프로크러스티즈 통계량을 통해 Figure 3.1과 Figure 3.2에서 살펴본 경기기록요인 변수군과 기술요인 변수군별 SCCB, PCCB 그리고 SPCCB의 형상변동을 살펴보면, 두 변수군 모두 SCCB와 PCCB가 제일 변동(variability)이 높고, 다음으로 SCCB와 SPCCB가 변동이 높다. 이는 Figure 3.1의 세 행렬도의 비교에서 지적한 것과 같이 처음부터 공변량요인을 고려하지 않은 SCCB의 문제점으로 여겨진다. 또한, 기술요인 변수군의 통계량이 경기기록요인에 비해 상대적으로 크게 나타나

공변량요인의 영향력이 다소 크다는 것을 알 수 있다.

4. 결론

본 연구에서는 공변량변수군이 한 변수군에는 영향을 주지만 다른 한 변수군에는 영향을 주지 않는 자료의 경우, 영향을 받는 한 변수군에서만 공변량변수군의 영향를 제거한 후 두 변수군의 관계를 파악하는 준편정준상관분석을 이용하여 이를 기하적으로 해석하기 위한 SPCCB를 제안하였다. 그리고 이를 활용하여 경륜운영본부에서 제공하는 2011년 기준 전체 성적 순위 100위권 이내의 선수 50명에 대한 자료에서 경기기록요인과 기술요인의 두 변수군 중 경기기록요인에 더 큰 영향을 주는 체격요인 변수군을 공변량변수군으로 고려하여 SPCCB의 예를 보이고, 추가적으로 SCCB와 PCCB와의 비교를 통해 그 차이점을 설명하였다. 결과적으로 SCCB보다는 PCCB가, PCCB보다는 SPCCB가 요인들의 성격을 더욱 두드러지게 나타내었고, 보다 나은 해석을 가능하게 해주었다. 그리고 이러한 행렬도간의 형상변동 차이를 프로크러스티즈 분석을 활용하여 비교한 결과에서도 유사한 해석을 할 수 있었다.

References

Choi, T.-H. and Choi, Y.-S. (2008). A study on the relationship between skill and competition score factors of KLPGA players using canonical correlation biplot and cluster analysis, *The Korean Journal of Applied Statistics*, **21**, 429–439.

Choi, T.-H. and Choi, Y.-S. (2010). A study on the relationship between physique, physical fitness and basic skill factors of tennis players in the Korea Tennis Association using the generalized canonical correlation biplot and procrustes analysis, *Communications of the Korean Statistical Society*, **17**, 917–925.

Choi, T.-H., Choi, Y.-S. and Shin, S. M. (2009). A study on the relationship between player characteristic factors and competitive factors of tennis grand slams competition using canonical correlation biplot and procrustes analysis, *The Korean Journal of Applied Statistics*, **22**, 855–864.

Choi, Y. S. (2006). *Biplot Analysis*, Research Institute for Basic Sciences, Series 2 of Basic Sciences, Pusan National University Press.

Choi, Y. S. and Hyun, G. H. (2006). *Understanding and Application of Statistical Shape Analysis - Study and Development of Resistant Version of Procrustes Analysis-*, Free Academy, Seoul.

Choi, Y. S., Hyun, G. H. and Yun, W. J. (2005). Biplots' variability based on the Procrustes analysis, *Journal of the Korean Data Analysis Society,* **7**, 1925–1933.

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

Gower, J. C. and Dijksterhuis, G. B. (2004). *Procrustes Problems,* Oxford: University Press.

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

Park, M. and Huh, M. H. (1996b). Quantification plots for several sets of variables, *Journal of the Korea Statistical Society,* **25**, 599–601.

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

Yeom, A.-R. and Choi, Y.-S. (2011). Partial canonical correlation biplot, *The Korean Journal of Applied Statistics*, **24**, 559–566.