응용통계연구 제 **21**권 *3*호, *2008*년, *pp. 429–439*

정준상관 행렬도와 군집분석을 응용한 KLPGA 선수의 기술과 경기성적요인에 대한 연관성 분석

최태훈[1] 최용석[2]

요 약

정준상관 행렬도(canonical correlation biplot)는 정준상관분석에서 두 변수 집단에 의해서 측정된 다변량 자료에서 변수 집단 간의 관계와 개체들의 관계를 탐색하기 위한 2차원 그림이다. 이는 일반적으로 최용석 (2006, 1장)의 한 변수 집단에 의한 행렬자료에 대한 일반적인 행렬도를 두 변수 집단에 의한 행렬자료로 확장한 것으로 볼 수 있다. 최근에 Choi와 Kim (2008)은 개체들이 많은 대용량 자료에서 행렬도의 해석상 힘든 문제점을 지적하고 이를 극복하는 데 군집분석을 활용하는 방법을 제시하고 있다. 일반적인 행렬도에서 발생하는 대용량 자료에 대한 문제는 정준상관 행렬도에서도 동일하게 발생하곤 한다.

본 연구에서는 2006년도 KLPGA 선수 중 상금 순위 상위 50명을 대상으로 정준상관 행렬도를 통해 기술요인변수군(평균 퍼팅수, 그린 적중율, 파 세이브율, 파 브레이크율)과 경기성적요인변수군(상금, 평균 타수)간의 관련성을 살펴보고 군집분석을 활용하여 각 선수들의 군집을 시도하려한다.

주요용어: 정준상관분석, 정준상관 행렬도, K-평균 군집분석, KLPGA.

1. 서론

정준상관 행렬도(canonical correlation biplot)는 정준상관분석에서 두 변수 집단에 의해서 측정된 다변량 자료에서 변수 집단 간의 관계와 개체들의 관계를 탐색하기 위한 2차원 그림이다. 이는 일반적으로 최용석 (2006, 1장) 의 한 변수 집단에 의한 행렬자료에 대한 일반적인 행렬도를 두 변수 집단에 의한 행렬자료로 확장한 것으로 볼 수 있다.

국내에선 Park (1995), Park과 Huh (1996a, 1996b)가 정준상관분석에서 수량화 방법(quantification method) 관점을 이용하여 정준상관도와 같은 그림을 제안하였다. 이들은 이런 그림을 정준상관 행렬도라 하였고, 세 변수 집단 이상인 경우까지 확장한 정준상관분석의 일반화를 시도하였다.

일반적으로 행렬도(biplot)는 복잡한 다변량 분석의 결과를 보다 쉽게 파악할 수 있기 때문에 최근 여러 분야에서 행렬도에 대해서 활발한 연구와 응용을 하고 있다. 행렬도는 Gabriel (1971)에 의해서 주로 개발되었고, 국내에선 Choi (1991)가 처음으로 소개하였으며, 이를 행

1) (760-709) 경상북도 안동시 서후면 교리 469번지, 안동과학대학 체육계열, 전임강사.
E-mail: thchoi@asc.ac.kr

2) (609-735) 교신저자. 부산 금정구 장전동 산 30, 부산대학교 통계학과, 교수.
E-mail: yschoi@pusan.ac.kr

렬도라 부른 것은 허명회 (1993, 5장)가 처음이었다. 더군다나 최용석 등 (2005a)은 다변량 분산분석 모형의 모수 추정치를 사용하는 MANOVA 행렬도를 제안하였다. 더 나아가 공변량(covariate)의 효과가 있는 경우 MANCOVA 행렬도를 최용석 등 (2005b)은 제안하고 응용의 예를 보였다. 최용석 등 (2005c), Choi 등 (2005a) 그리고 Choi 등 (2005b)는 행렬도에 대한 활용의 폭을 넓히고 있다.

최근에 Choi와 Kim (2008)은 개체들이 많은 대용량 자료에서 행렬도의 해석상 힘든 문제점을 지적하고 이를 극복하는데 군집분석을 활용하는 방법을 제시하고 있다. 일반적인 행렬도에서 발생하는 대용량 자료에 대한 문제는 정준상관 행렬도에서도 동일하게 발생하곤 한다.

본 연구에서는 2006년도 KLPGA(Korea Ladies Professional Golf Association) 선수 중 상금 순위 상위 50명을 대상으로 정준상관 행렬도를 통해 기술요인변수군(평균 퍼팅수, 그린 적중율, 파 세이브율, 파 브레이크율)과 경기성적요인변수군(상금, 평균 타수)간의 관련성을 살펴보고 군집분석을 활용하여 각 선수들의 군집을 시도하려한다. 이미 최태훈 (2006)은 2005년도 KLPGA 선수들의 성적에 대해 기술요인과 랭킹, 평균타수에 대해 스피어만의 상관계수분석과 다중회귀분석(multiple regression analysis)을 실시하였다. 2장에서 정준상관 행렬도를 소개하고 Choi와 Kim (2008)의 방법에 따라 정준상관 행렬도에서 군집분석을 활용하고 3장에서는 실제 분석의 예를 보이고자 한다.

2. 정준상관 행렬도와 군집분석의 활용

2.1. 정준상관 행렬도의 소개

이 절에서는 정준상관 행렬도에 대해 잘 정리된 최용석 (2006, 2.3절)을 참고로 정준상관분석와 관련된 정준상관 행렬도의 대수적인 면을 간단히 설명하기로 하자.

정준상관분석은 두 변수군 사이의 관계를 분석하는 다변량기법이다. 각 변수군은 여러 변수로 구성되어있다. 정준상관분석은 Hotelling (1935)에 의해서 처음 개발된 기법으로 아동들의 읽기변수군(읽기속도, 읽기능력)과 산술변수군(산술속도, 산술능력)간의 관계를 보여주었다.

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 평균 $\mu_{\mathbf{x}} = (\mu_{x_1}, \ldots, \mu_{x_p})'$와 $\mu_{\mathbf{y}} = (\mu_{y_1}, \ldots, \mu_{y_q})'$를 가지며 모집단 공분산행렬 Σ_{xx}, Σ_{yy}, $\Sigma_{xy} = \Sigma'_{yx}$을 가지는 확률벡터이다.

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

$$Z_x = u_1 x_1 + \cdots + u_1 x_p = \mathbf{u}'\mathbf{x}, \qquad Z_y = v_1 y_1 + \cdots + v_1 y_q = \mathbf{v}'\mathbf{y} \tag{2.1}$$

을 생각하자. 이들 선형결합은 변수군의 다차원(p 또는 q 차원)정보를 1차원으로 축소하여 단순 요약 측정치(simple summary measure)를 제공한다. 따라서 식 (2.1)의 두 선형결합 Z_x와

Z_y의 상관은

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

이다. 여기서 S_{xx}, S_{yy}, S_{xy}는 각각 $\Sigma_{xx}, \Sigma_{yy}, \Sigma_{xy}$의 표본 공분산행렬들이다. 이젠 식 (2.1)에서 계수벡터 $\mathbf{u}$와 $\mathbf{v}$는 식 (2.2)의 두 선형결합의 상관을 최대화하는 알고리즘을 통하여 구할 수 있다. 상관을 최대화하는 알고리즘은 Z_x와 Z_y의 분산이 1인 제약조건 $\mathbf{u}' S_{xx} \mathbf{u} = 1$과 $\mathbf{v}' S_{yy} \mathbf{v} = 1$을 두고 $\mathbf{u}' S_{xy} \mathbf{v}$를 최대화하는 계수벡터 $\mathbf{u}$와 $\mathbf{v}$를 찾는 것과 동일하다. 이 알고리즘은 라그랑즈 승수(Lagrange multiplier) 방법을 이용하면 쉽게 다음의 두 고유체계(eigensystem) 문제로 유도할 수 있다.

$$(S_{xx}^{-1} S_{xy} S_{xy}^{-1} S_{yx} - \lambda \mathbf{I})\mathbf{u} = 0, \qquad (S_{yy}^{-1} S_{yx} S_{xx}^{-1} S_{xy} - \lambda \mathbf{I})\mathbf{u} = \mathbf{0}. \tag{2.3}$$

식 (2.3)에서 고유값 λ는 공통으로 정준상관의 제곱이다. 이 고유값의 개수는 $r = \min(p, q)$이며 $\lambda_1 \geq \cdots \geq \lambda_r \geq 0$의 크기 순서로 되어 있다. 따라서 고유체계의 성질에 따라 k번째 정준상관의 제곱인 λ_k에 대응하는 고유벡터 $\mathbf{u}_k$와 $\mathbf{v}_k$를 얻게 된다. 이들 고유벡터를 정준계수벡터라하며 식 (2.1)에 대입하면 정준변수의 한 짝인 $Z_{xk} = \mathbf{u}_k' \mathbf{x}$과 $Z_{yk} = \mathbf{v}_k' \mathbf{y}$를 구하게 된다. 특히, 이들 짝의 상관은 k번째 정준상관을 $\rho_k = \sqrt{\lambda_k}$라 하면 $\rho_k = r_{Z_{xk}Z_{yk}}$를 만족한다.

식 (2.3)의 고유체계 대신에 비정칙값분해(singular value decomposition)

$$S_{xx}^{-\frac{1}{2}} S_{xy} S_{xy}^{-\frac{1}{2}} S_{yx} = U D_{\sqrt{\lambda}} V' \tag{2.4}$$

을 이용하여 정준계수벡터와 정준상관을 대수적으로 한꺼번에 구할 수도 있다. 여기서 크기가 $p \times r$과 $q \times r$행렬 $U = (\mathbf{u}_1, \ldots, \mathbf{u}_r)$와 $V = (\mathbf{v}_1, \ldots, \mathbf{v}_r)$은 정준계수벡터의 직교행렬이며 대각행렬 $D_{\sqrt{\lambda}} = \mathrm{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_r})$는 $\sqrt{\lambda_1} \geq \cdots \geq \sqrt{\lambda_r} > 0$ 관계를 갖는 비정칙값이 정준상관을 대각원소로 하고 있다.

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 X와 Y라 하고 이들은 중심화되어 있다고 하자. 그러면 식 (2.4)로부터 i번째 표준정준상관계수벡터는 각각

$$\mathbf{a}_i = S_{xx}^{-\frac{1}{2}} \mathbf{u}_i, \qquad \mathbf{b}_i = S_{yy}^{-\frac{1}{2}} \mathbf{v}_i, \quad i = 1, \ldots, r \tag{2.5}$$

이다. 이들에 의해서 구성된 표준정준상관계수행렬은 $A = (\mathbf{a}_1, \ldots, \mathbf{a}_r)$이고 $B = (\mathbf{b}_1, \ldots, \mathbf{b}_r)$가 된다. 이들에 의해서 정준상관도의 좌표는 자료행렬 X에 대하여 행좌표행렬과 열좌표행렬은

$$R_X = XAD_{\sqrt{\lambda}}, \qquad C_X = AD_{\sqrt{\lambda}} \tag{2.6}$$

이며 자료행렬 Y에 대하여 행좌표 행렬과 열좌표 행렬은

$$R_Y = XBD_{\sqrt{\lambda}}, \qquad C_Y = BD_{\sqrt{\lambda}} \tag{2.7}$$

이다.

특히 행좌표행렬들은 n명의 각 개체에 대하여 정준변수점수행렬이라 할 수 있다. s차원의 정준상관 행렬도는 식 (2.6)과 (2.7)의 행렬에서 처음 s개 열을 고려한 부행렬로 이루어지며 이 s차원의 정준상관 행렬도의 근사도는 식 (2.4)로부터 얻어지는 전체 정준상관의 제곱값 합에서 s개의 정준상관의 제곱값 합이 차지하는 비율을 이용한다.

2.2. 군집분석의 활용

이 절에서는 Choi와 Kim (2008)의 군집분석에 대한 내용을 요약하기로 하자. 서론에서 언급했듯이, 일반적으로 행렬도가 자료의 정보를 시각적으로 표현하여 쉽게 정보를 파악할 수 있는 장점을 가지고 있지만, 개체 수가 많은 대용량 자료에서는 해석이 힘든 문제점을 가지고 있다. 이를 극복하기 위해 대용량 자료의 개체 군집화에 효율적이고 적합한 K-평균 군집분석을 수행하여 정준상관 행렬도에 활용하는 것을 제안한다. K-평균 군집분석은 비위계적 군집방법(non-hierachical clustering method)으로 그 알고리즘은 4단계로 다음과 같다.

[**단계 1**] K개 초기 군집들로 분할한다.

[**단계 2**] 모든 개체를 가장 가까운 중심점을 갖는 군집에 할당한다.

[**단계 3**] 군집 중심을 계산한다.

[**단계 4**] [단계 2]와 [단계 3]을 할당이 일어나지 않을 때 까지 반복한다.

이 때, Sharma (1996, pp. 221–232)는 위계적 군집방법에서 나온 군집의 수와 군집 중심을 초기 군집의 수 K와 초기 시드점으로 사용하는 방법을 제안하고 있다. 위계적 군집방법에는 단일연결, 완전연결, 중심연결, 평균연결, 와드(WARD)연결 등 많은 방법이 있다.

본 연구에서는 위계적 군집방법 중 가장 보편적으로 사용되는 평균연결, 중심연결, 와드연결을 사용하여 군집분석을 수행하고 평균제곱표준편차근(root-mean-square standard deviation: RMSSTD), 반부분 R^2(semipartial r-square: SPRSQ), R^2(r-square: RSQ), CCC(cubic clustering criterion), pseudo-F(PSF), pseudo-t^2(PST2) 등을 이용하여 군집의 수 K를 결정하려 한다.

먼저 i번째 군집의 RMSSTD는

$$\sqrt{\frac{\sum_{k \in C_i} \|X_k - \bar{X}_i\|^2}{p(N_i - 1)}}$$

이다. 여기서 p는 변수의 수, N_i는 i번째 군집 C_i의 개체 수, X_k는 i번째 군집 C_i에 속해 있는 개체($k = 1, 2, \ldots, N_i$), $\bar{X}_i$는 i번째 군집 C_i에서의 평균이다. 군집분석의 목적은 군

집 내에서는 동질적이어야 하므로 RMSSTD값이 작아야 한다. 따라서 군집의 수에 대응되는 RMSSTD의 값을 그려서 급격한 감소가 발생하는 곳에서 대응되는 군집의 수를 정할 수 있다.

제곱합에 근거한 판정기준에 의해 R^2은 다음과 같이 정의된다.

$$R^2 = \frac{SS_b}{SS_t}, \tag{2.8}$$

식 (2.8)에서 SS_t는 전체제곱합, SS_w는 군집내(within-cluster)제곱합, SS_b는 군집간(between-cluster)제곱합으로 다음과 같이 표현된다.

$$\begin{aligned} SS_t &= \sum_{i=1}^{g}\sum_{j=1}^{N_i}(X_{ij}-\bar{X})(X_{ij}-\bar{X})' \\ &= \sum_{i=1}^{g}\sum_{j=1}^{N_i}(X_{ij}-\bar{X}_i)(X_{ij}-\bar{X}_i)' + \sum_{i=1}^{g} N_i(\bar{X}_i-\bar{X})(\bar{X}_i-\bar{X})' \\ &= SS_w + SS_b. \end{aligned} \tag{2.9}$$

식 (2.9)에서 g는 군집의 수를 말하며, X_{ij}는 i번째 군집에서 j번째 개체를 뜻한다. $\bar{X}_i$는 i번째 군집에서의 평균이며, $\bar{X}$는 모든 개체들의 평균이다. 일단 자료가 주어지면 SS_t는 고정되므로 SS_b와 SS_w의 관계로부터 군집내 제곱합에 비해서 군집간 제곱합이 크도록 하는 판정기준으로 생각할 수 있다. 따라서, R^2의 값이 급격한 증가가 발생한 곳에서 대응되는 군집의 수를 정할 수 있다.

SPRSQ는 군집분석에서 집단 간 유사성을 측정하는 통계량으로, 특정 Y축 값 이내(예: 0.1)에 구분이 되지 않는 경우는 그룹 내는 동질적이라고 한다. R^2의 경우와 반대로 급격한 감소가 발생한 곳에서 대응되는 군집의 수를 정할 수 있다.

CCC는 각각의 군집이 단일분포라고 가정할 때, 군집분석을 수행함으로써 군집내부의 분포가 단일분포와 달라질수록 좋다는 것을 이용하며, 관찰된 R^2와 단일분포를 가정했을 때 R^2의 비를 기준(criterion)으로 사용한다. 군집의 수와 CCC의 산점도를 그려 그값이 3이상이면서 최대값인 경우 그 때의 군집의 수를 선택하는 방법이다.

pseudo-t^2 검정 통계량은 두 집단 간 다변량 평균의 차이를 보는 통계량이다. 개체의 군집간 평균의 차이가 유의하지 않으면 두 군집을 합치고, 유의하면 군집을 그대로 유지하는 방법이다. pseudo-t^2 값이 크다는 것은 군집간 거리가 멀다는 것을 의미하므로 군집을 나누는 것이 좋고, 반대의 경우는 합치는 것이 좋다. pseudo-F 통계량도 이와 유사하지만, 보통 pseudo-t^2를 이용해 군집의 수를 결정한다.

3. KLPGA 선수의 기술요인과 경기성적요인 분석

2절에서 요약한 정준상관분석 행렬도와 군집분석을 응용하기 위한 자료는 한국여자프로골프협회(KLPGA) 홈 페이지(www.klpga.com, 2006)에서 제공하는 2006년도 전체 경기의 기록 결과이다. 특히, 경기기록 측정 항목 중 본 연구에서 사용되는 변수로 평균타수(scoring

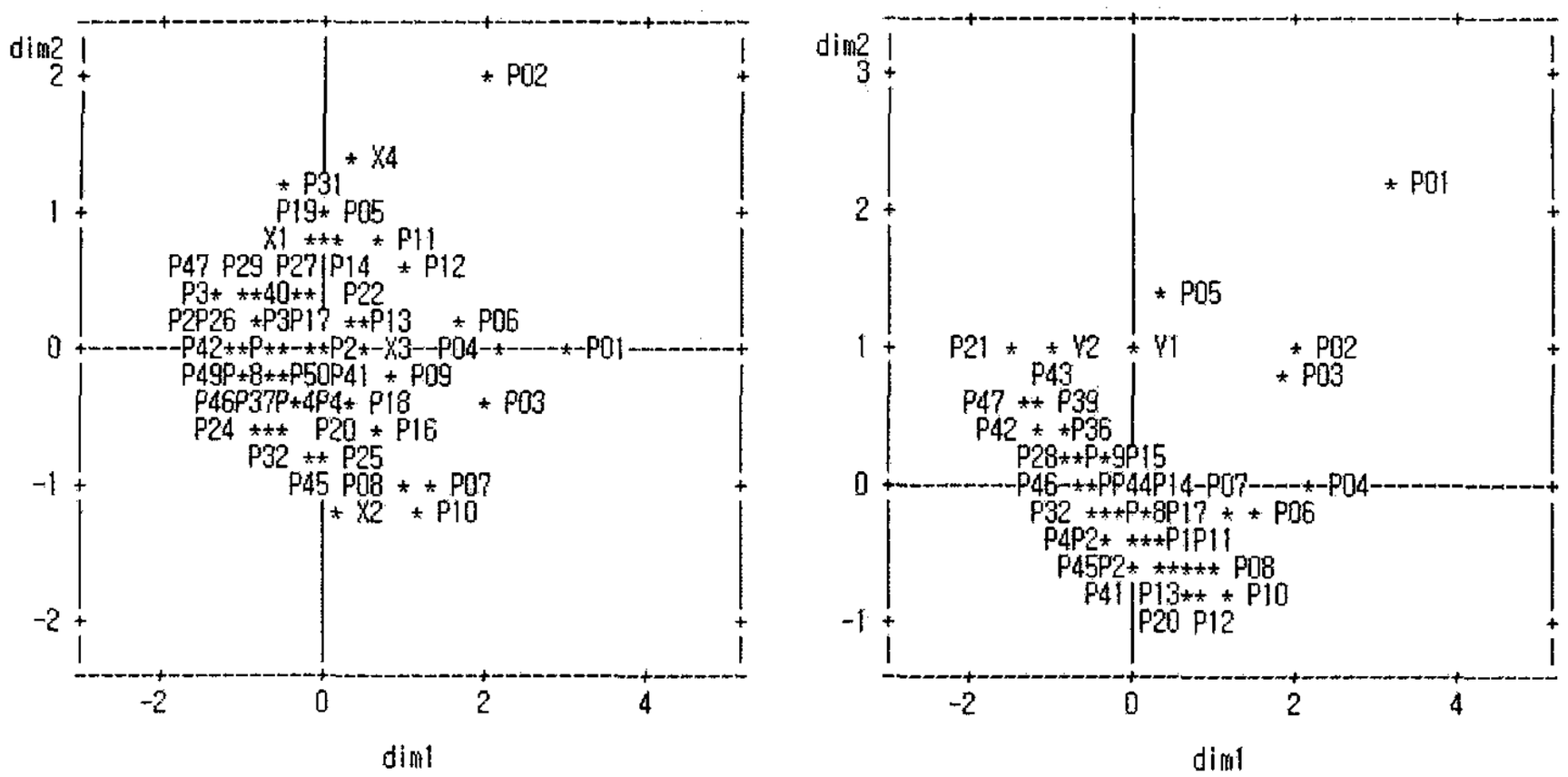


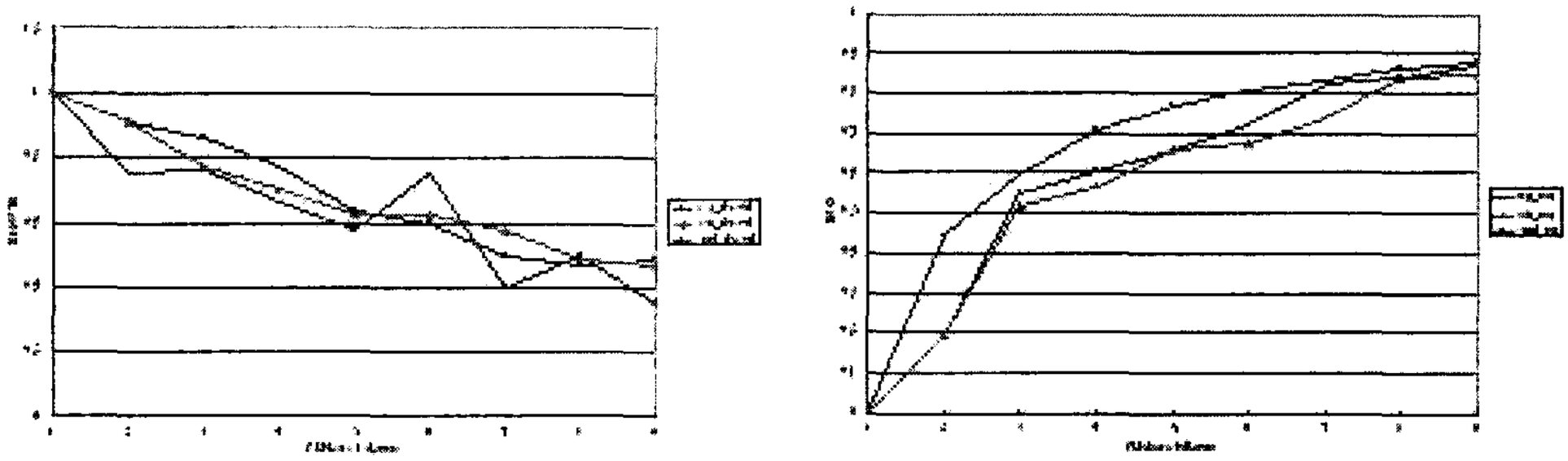
그림 3.1: 정준상관 행렬도 ((a) 기술요인변수군, (b) 경기성적요인변수군)

average), 평균퍼팅 수(putting average), 그린적중 율(green in regulation percentage), 파 세이브 율(par save percentage), 파브레이크 율(par break percentage), 상금(prize)을 선정하였다. 특히, 상금은 선수들의 순위(ranking)를 결정하는 중요 변수이다. 상금외에 이들 변수는 골프경기의 실제상황에 대처하는 중요한 기술 (서은영과 박태섭, 2003; APGA, 1996)로 여러 가지 측면에서 볼 때 최종 스코어에 영향을 미쳐 측정항목들 간의 상대적 기여도를 알아볼 수 있는 장점이 있다. 특별히, 평균 퍼팅수($X1$), 그린 적중률($X2$), 파 세이브율($X3$), 파 브레이크율($X4$)을 기술요인변수군으로 상금($Y1$)과 평균 타수($Y2$)를 경기성적요인변수군으로 고려할 수 있다.

그림 3.1의 (a)와 (b)는 각각 기술요인변수군과 경기성적요인변수군에 대한 2차원 정준상관 행렬도로 적합도는 100%이며 특히, 첫 번째 정준상관이 88.32%로 수평축인 dim1(제1축)에 대한 해석만으로도 충분하다고 여겨진다.

그림 3.1의 (a) 기술요인변수군에 대한 정준상관 행렬도에서 $X2$(그린적중율), $X3$(파세이브율), $X4$(파브레이크율)이 수평축인 dim1(제1축)에 대하여 오른쪽 방향에 놓여 있어 양의 상관이 높고 같은 경향을 나타냄을 알 수 있으며, $X1$(평균퍼팅수)는 나머지와 반대의 경향을 나타내고 있다. 그리고, 상위 10위 이내 선수(P01–P10)들을 포함한 상위권 선수들이 $X2$(그린적중율), $X3$(파세이브율), $X4$(파브레이크율)변수들과 같은 방향으로 놓여 있어 하위권 선수들에 비해 상위권 선수들은 그린적중율, 파세이브율, 파브레이크율이 높음을 알 수 있다. 또한, $X1$(평균퍼팅수)변수와 상위권 선수들은 반대 반향에 놓여 있기 때문에 상위권 선수들은 보통 평균퍼팅수는 작다는 것을 알 수 있다.

다음으로 경기성적요인변수군에 대한 2차원 정준상관 행렬도 그림 3.1의 (b)에서 $Y1$(상금), $Y2$(평균타수)가 dim1(제1축)에 대하여 서로 다른 방향에 놓여 있어 성향이 다른 변수임을 알 수 있다. 특히, 상위권 선수들이 $Y1$(상금)변수와 같은 방향으로 놓여 있어 하위권 선수들에 비해 상위권 선수들이 누적 상금액도 높음을 알 수 있다. 또한, $Y2$(평균타수)변수와 상



(a) (b)

그림 3.2: 군집의 수 ((a) RMSSTD, (b) RSQ)

위권 선수들은 반대 반향에 놓여 있기 때문에 상위권 선수들은 보통 평균타수는 작다는 것을 보여준다.

또한 서로 다른 변수군의 변수들간의 연관성을 엿보기 위해 그림 3.1의 (a)와 (b)를 동시에 비교해 보면된다. 즉, dim1(제1축)에 대하여 그림 3.1-(a)의 오른편에 있는 기술요인변수 $X2$(그린적중율), $X3$(파세이브율), $X4$(파브레이크율)는 그림 3.1-(b)의 경기성적요인변수 $Y1$(상금)과 같은 방향에 놓여 있어 연관성이 높음을 보여준다. 이와는 반대로 그림 3.1-(a)의 $X1$(평균퍼팅수)과 그림 3.1-(b)의 $Y2$(평균타수)가 서로 연관성이 있음을 보여준다. 이는 기술요인변수인 퍼팅이 경기성적 요인변수이면서 최종 스코어인 평균타수에 아주 중요한 영향을 미치고 있음을 의미한다. 최태훈 (2006)에서는 스코어가 낮아야 좋은 기록인 골프 경기에서 퍼팅수가 적을수록 스코어가 낮아지므로, 퍼팅 수를 줄이는 것이 당연하다고 지적하고 있다. 더군다나 골프선수들은 경기 전과 후에도 항상 그린에서 퍼팅연습에 많은 시간을 투자하지만 실질적으로 아마추어 골퍼들은 아이언이나 드라이브 연습에 많은 시간을 투자하기 때문에 결과적으로 기록경기인 골프경기에서의 평균타수를 낮추기 위해서는 퍼팅연습에 많은 시간을 투자하고 연습하는 것이 좋은 평균타수를 기록하는 밑거름이라 생각된다고 지적하고 있다.

덧붙여, 정준상관 행렬도에서 50명의 선수들이 2차원 상에서 표현에 제약이 있음을 알 수 있다. 따라서, 2절에서 언급한 군집 방법을 활용하여 이들을 적절하게 군집화하고 이들 군집의 중심값을 활용하여 군집의 성질을 파악하려 한다.

이를 위하여 먼저 군집의 수를 결정하기로 하자. 그림 3.2의 (a) RMSSTD와 (b) RSQ 값들에 대한 평균연결(ave_rmsstd, ave_rsq), 중심연결(cen_rmsstd, cen_rsq), 와드연결(ward_rmsstd, ward_rsq) 그림을 순서대로 보여주고 있다. 먼저 그림 3.2의 (a)에서는 군집의 수에 대응되는 RMSSTD의 값을 그려서 급격한 감소가 발생하는 곳에서 대응되는 군집의 수를 정할 수 있으므로, 3개-5개 군집이 적당할 것으로 보인다. 그림 3.2의 (b)에서는 RSQ의 값이 급격한 증가가 발생한 부분까지 군집의 수를 정할 수 있으므로, 3개–5개 군집이 적당해 보인다. 또한 SPRSQ, CCC, PST2의 경우도 이와 대동소이한 결과를 나타내어 생략하였다. 지금까지 군집의 수 K를 결정하는 방법들의 결과를 다시 정리하면 표 3.1과 같다.

표 3.1: KLPGA자료에 대한 군집의 수

통계량	군집수
RMSSTD	3 , 4, 5
RSQ	3, 4, 5
SPRSQ	3, 5
CCC	3, 5
PST2	3, 5

표 3.2: KLPGA 자료에서 각 군집의 중심값

변수 / 군집	$X1$	$X2$	$X3$	$X4$	$Y1$	$Y2$
1군집(CL1)	−0.794	1.844	1.881	2.246	2.519	−2.119
2군집(CL2)	0.221	−0.481	−0.566	−0.504	−0.451	0.588
3군집(CL3)	−0.277	0.554	0.772	0.449	0.192	−0.733

표 3.3: 각 군집에 속한 선수

군집	선수
1군집 (CL1)	P01, P02, P03, P04, P062
군집 (CL2)	P14, P15, P17, P19, P21, P23, P24, P25, P26, P27, P28, P29, P30, P31, P32, P33, P34, P35, P36, P37, P38, P39, P40, P41, P42, P43, P44, P45, P46, P47, P48, P49, P50
3군집 (CL3)	P05, P07, P08, P09, P10, P11, P12, P13, P16, P18, P20, P22

원자료를 표준화하여 위계적 군집분석을 수행한 후, 군집의 수에 대한 결정은 표 3.1의 결과를 이용한다. 초기 시드점으로 계층적 방법에서의 군집 중심을 사용하고, 군집의 수로 가장 많이 선택됨과 동시에 가장 적은 수인 3개를 초기 군집의 수로 두고, K-평균 군집분석을 수행한다. 그 결과, 표 3.2는 변수별 각 군집의 중심값과 표 3.3은 각 군집에 어느 선수들이 속한지를 보여주고 있다.

그림 3.3는 표 3.2의 군집중심값에 대한 행렬도로 제1축의 설명력이 98.48%이다. 이는 앞의 그림 3.1의 (a)와 (b)에서 나타난 선수들의 군집과 변수들간의 해석에 도움을 제공한다. 그림 3.3을 보면 $X2$(그린적중율), $X3$(파세이브율), $X4$(파브레이크율), $Y1$(상금)이 dim1(제1축)에 대하여 같은 방향에 놓여 있어 이 변수들 간의 양의 상관이 높아서 같은 경향을 나타냄을 알 수 있으며, $X1$(평균퍼팅수)과 $Y2$(평균타수)는 나머지 변수들과 반대의 경향을 나타내고 있다. 이런 해석은 이미 그림 3.1의 (a)와 (b)를 동시에 비교하여 서로 다른 변수군에 속한 변수들간의 연관성을 엿본 것과 대동소이하다.

다만, 이들 그림에서 알 수 없었던 선수들의 군집화 정보와 그 군집의 특성을 설명하는 변수들간의 관계를 그림 3.3은 보여주고 있다. 즉, 제1군집(CL1)에 속한 선수들은 $X2$, $X3$, $X4$, $Y1$ 변수들과 같은 방향으로 놓여있는 것을 알 수 있다. 즉, 타 군집에 속한 선수들에 비

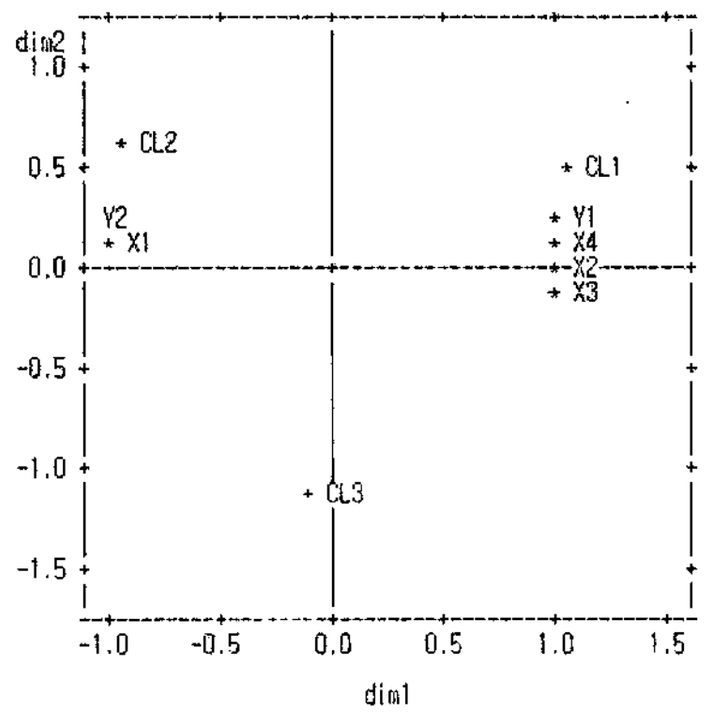


그림 3.3: KLPGA 자료의 각 군집중심값에 대한 행렬도

해 그린적중율, 파세이브율, 파브레이크율, 상금이 높음을 나타내며 이것은 성적이 뛰어난 선수들이 속한 군집임을 알 수 있다. 또한, $X1$와 $Y2$변수들과는 제1군집이 반대 반향으로 놓여 있기 때문에 상위권 선수들은 보통 평균퍼팅수와 평균타수가 타 군집에 비해 작다(이는 골프에서 성적이 좋은 것임)는 것을 알 수 있다. 제2군집(CL2)은 $X1$, $Y2$변수의 방향에 놓여있어 타 군집에 속한 선수들에 비해 성적이 좋지 않은 선수들이 속한 군집임을 알 수 있다. 즉, 게임에서 평균퍼팅수와 평균타수는 높고 그린적중율, 파세이브율, 파브레이크율은 타 군집에 비해 떨어짐을 알 수 있고, 이로 인해 상금 또한 낮음을 알 수 있다. 제3군집(CL3)은 타 군집과 비교하면 가장 평범한 선수들이 속한 군집이다. 전체적으로, 성적이 크게 좋지도 나쁘지도 않은 그만그만한 선수들이 속한 군집이라 말할 수 있다.

4. 결론

정준상관 행렬도가 정준상관분석에서 두 변수 집단에 의해서 측정된 다변량 자료에서 변수 집단 간의 관계와 개체들의 관계를 탐색하기 위한 2차원 그림이다. 이는 서로 다른 변수군의 변수들간의 정보와 개체를 시각적으로 표현하여 쉽게 정보를 파악할 수 있는 장점을 가지고 있지만 개체 수가 많은 대용량 자료에서는 군집화와 그 특성을 파악하기가 쉽지 않은 문제점이 있다. 이러한 한계를 극복하기 위해 군집분석을 행렬도에 활용하고자 하였고, 위계적 군집 방법에서 얻어지는 군집의 수 K와 군집 중심을 초기 시드점으로 한 K-평균 군집분석을 수행하였다. 이를 통하여 정준상관 행렬도에 K-평균 군집분석으로부터 얻어지는 각 군집의 중심값에 대한 행렬도는 각 변수들간의 관계도 통합적으로 보여 주고 이러한 변수들과 군집들간의 관계를 통해 군집의 특성을 파악하는 데 도움을 주었다.

따라서 본 연구의 정준상관 행렬도와 군집분석에서에서 나타난 결과를 요약하자면 다음과 같다.

첫째, 정준상관 행렬도에서 기술요인변수군의 그린적중율, 파세이브율, 파브레이크율이 경기성적요인변수군의 상금과 같은 방향에 놓여 있어 연관성이 높음을 보여주었다. 이와 반대로 평균퍼팅수와 평균타수가 서로 연관성이 있음을 보여준다. 이는 기술요인변수인 퍼팅이

경기성적 요인변수이면서 최종 스코어인 평균타수에 아주 중요한 영향을 미치고 있음을 의미한다.

둘째, 군집분석의 결과 하위권 선수들에 비해 상위권 선수들은 그린적중율, 파세이브율, 파브레이크율이 높은 반면에 평균퍼팅수는 상위권 선수들이 작다는 것을 보여주었다.

참고문헌

서은영, 박태섭 (2003). 여자프로골프선수의 경기력에 미치는 결정요인분석, <한국기록분석학회지>, **1**, 27–38.

최용석 (2006). <행렬도 분석>, 기초과학 총서 2권, 부산대학교 기초과학연구원.

최용석, 현기홍, 정수미 (2005a). 다변량 분산분석에서 추정된 모수행렬의 행렬도, *Journal of the Korean Data Analysis Society*, **7**, 851–858.

최용석, 현기홍, 정수미 (2005b). MANCOVA Biplot, <한국통계학논문집>, **12**, 705–712.

최용석, 강창완, 김경덕 (2005c). 명목형 다항반응 로지스틱회귀모형의 행렬도 분석, *Journal of the Korean Data Analysis Society*, **7**, 839–849.

최태훈 (2006). KLPGA 선수의 기술요인과 랭킹간의 관계, <안동과학대학 논문집>, **28**, 431–439.

허명회 (1993). <통계상담의 이해>, 자유아카데미, 서울.

APGA (1996). *Official Media Guide of the PGA TOUR*, American Professional Golf Association.

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

Choi, Y. S., Hyun, G. H. and Kim, J. G. (2005a). A numerical comparison of map variability in SCA using the procrustes analysis, *Journal of the Korean Data Analysis Society*, **7**, 1531–1538.

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

Choi, Y. S. and Kim, H. Y. (2008). Applications of cluster analysis in biplots, *Communications of the Korean Statistical Society*, **15**, 65–76.

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

Hotelling, H. (1935). The most predictable criterion, *Journal of Educational Psychology*, **26**, 139–142.

Park, M. (1995). *Quantification Plots for Canonical Correlation Analysis*, Unpublished Ph.D. Dissertation, Department of Statistics, Korea University.

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**, 589–601.

Sharma, S. (1996). *Applied Multivariate Techniques*, Wiley, New York.

[2008년 2월 접수, 2008년 3월 채택]

A Study on the Relationship between Skill and Competition Score Factors of KLPGA Players Using Canonical Correlation Biplot and Cluster Analysis

Tae-Hoon Choi[1] Yong-Seok Choi[2]

ABSTRACT

Canonical correlation biplot is 2-dimensional plot for investigating the relationship between two sets of variables and the relationship between observations and variables in canonical correlation analysis graphically. In general, biplot is useful for giving a graphical description of the data. However, this general biplot and also canonical correlation biplot do not give some concise interpretations between variables and observations when the number of observations are large. Recently, for overcoming this problem, Choi and Kim (2008) suggested a method to interpret the biplot analysis by applying the K-means clustering analysis. Therefore, in this study, we will apply their method for investigating the relationship between skill and competition score factors of KLPGA players using canonical correlation biplot and cluster analysis

Keywords: Canonical correlation analysis, biplot, KLPGA, K-means cluster analysis.

1) Instructor, Dept. of Physical Education, Andong Science College, Andong 760-709, Korea.
E-mail: thchoi@asc.ac.kr
2) Corresponding author. Professor, Dept. of Statistics, Pusan National University, Busan 609-735, Korea.
E-mail: yschoi@pusan.ac.kr