The Korean Communications
in Statistics Vol. 12 No. 3, 2005
pp. 705-712

MANCOVA Biplot

Yong-Seok Choi[1], Gee Hong Hyun[2], and Su Mi Jung[3]

Abstract

Biplot is a graphical display of the rows and columns of an $n \times p$ data matrix. In particular, Gabriel (1995) suggested the MANOVA biplot using singular value decomposition (SVD) with the averages of response variables according to treatment groups. But his biplot may cause wrong results by disregarding them when there exist covariate effects. In this paper, we will provide the MANCOVA biplot based on the SVD with the parameter estimates for MANCOVA model when there exist covariate effects.

Keywords : Covariates, Biplot, MANOVA, MANCOVA, SVD

1. 서론

행렬도 (biplot)는 다변량 자료의 행과 열을 한 그림에 동시에 나타내어 이들의 관계를 한 눈에 알아볼 수 있게 하는 탐색적 방법의 그림도구이다. 행렬도에는 주성분 (principal component), 대칭 (symmetric), JK 행렬도 등이 있다 (최용석, 1999, 1장). 특히, Gabriel (1995)은 주어진 변량 자료로부터 각 처리별 변수들의 평균값을 구하고 대수적으로는 비정칙값분해 (singular value decomposition, SVD)에 의한 JK 행렬도를 활용하여 다변량 분산분석 행렬도 (MANOVA biplot)를 제안했다. 그러나 그의 알고리즘은 계산이 복잡하고 때때로 해석상 안정적이지 못했다. 이에 최용석 외 2인 (2005)은 Smith와 Cornell (1993)과 장대홍 (1996)이 다반응값 자료 및 다변량 회귀분석에서 추정된 계수행렬값을 비정칙값분해하여 반응변수와 설명변수간의 관계를 파악했던 것에 착안하여 다변량 분산분석 모형의 모수 추정치를 사용하는 다변량 분산분석 행렬도를 제안하였다. 이 방법은 행렬도를 그리기 위한 알고리즘과 프로그램 작성에서의 복잡성을 줄일 수 있고 모형을 해석하는데 더 나은 결과를 주었다. 그러나 이들의 연구에서는 공변량 (covariate)의 영향이 있는 자료에 대한 연구가 이루어지지 않았다.

따라서 본 논문에서는 공변량의 효과를 고려한 다변량 공분산분석 행렬도 (MANCOVA biplot)를 제안하려 한다.

1) Professor, Department of Statistics, Pusan University, Busan, 609-735. Korea.
E-mail: yschoi@pusan.ac.kr
2) Doctorial Course, Department of Statistics, Pusan University, Busan, 609-735. Korea.
3) Master Degree, Department of Statistics, Pusan University, Busan, 609-735. Korea.

2. 다변량 공분산분석 행렬도

k개의 서로 독립된 관측치 벡터($i=1,2,\cdots,k$, $j=1,2,\cdots,n_i$)의 일원 다변량 분산분석 모형은

$$\begin{aligned} \boldsymbol{y}_{ij} &= \boldsymbol{\mu} + \boldsymbol{\alpha}_i + \boldsymbol{\varepsilon}_{ij} \\ &= \boldsymbol{\mu}_i + \boldsymbol{\varepsilon}_{ij.} \end{aligned}$$

와 같다. 이 때, $\boldsymbol{y}_{ij} \sim \boldsymbol{N}_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$는 독립된 평균벡터와 공통된 공분산행렬을 가지는 다변량 정규분포를 따르며, $\boldsymbol{\mu}_i$는 i번째 모평균벡터이고, $\boldsymbol{\alpha}_i$는 i번째 처리효과이며, $\boldsymbol{\varepsilon}_{ij} \sim \boldsymbol{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$는 오차벡터로 평균벡터가 $\mathbf{0}$이고, 공분산행렬이 $\boldsymbol{\Sigma}$ 인 정규 확률변수이다. 이를 행렬로 나타내면 다변량 분산분석 모형은 식 (2.1)과 같다.

$$\mathbf{Y} = \mathbf{XB} + \boldsymbol{E}. \tag{2.1}$$

단, $n = \sum_{i=1}^{k} n_i$일 때, $\mathbf{Y}$는 $n \times p$의 관측치행렬이고, $\mathbf{X}$는 $n \times k$의 다변량 분산분석 설계행렬이며, $\mathbf{B}$는 $k \times p$의 모평균행렬이며, $\mathbf{E}$는 $n \times p$의 오차행렬이고 다음과 같다.

$$\mathbf{X} = \begin{bmatrix} \mathbf{1}_{n_1} & 0 & \cdots & 0 \\ 0 & \mathbf{1}_{n_2} & \cdots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \cdots & \mathbf{1}_{n_k} \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} \boldsymbol{\mu}_1' \\ \boldsymbol{\mu}_2' \\ \vdots \\ \boldsymbol{\mu}_k' \end{bmatrix}, \quad \boldsymbol{E} = \begin{bmatrix} \boldsymbol{\varepsilon}_1' \\ \boldsymbol{\varepsilon}_2' \\ \vdots \\ \boldsymbol{\varepsilon}_n' \end{bmatrix}.$$

모수행렬 $\mathbf{B}$의 추정치는 식 (2.2)와 같이 나타낼 수 있다.

$$\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \begin{bmatrix} \boldsymbol{y}_{1.}' \\ \boldsymbol{y}_{2.}' \\ \vdots \\ \boldsymbol{y}_{k.}' \end{bmatrix}. \tag{2.2}$$

다변량 공분산분석 모형은 식 (2.1)의 다변량 분산분석모형과 다변량 회귀모형이 결합된 형태로

$$\mathbf{Y} = \mathbf{XB} + \mathbf{Z}\boldsymbol{\Gamma} + \boldsymbol{E} \tag{2.3}$$

와 같이 나타낼 수 있다. 여기서 $\mathbf{X}$와 $\mathbf{B}$는 식 (2.1)의 다변량 분산분석의 것과 동일하고, $\mathbf{Z}$는 다변량 회귀모형에서 h개의 공변량으로 이루어진 크기가 $n \times h$인 행렬이고, $\boldsymbol{\Gamma}$는 크기가 $h \times p$인 모

수행렬이다. 식 (2.3)은

$$\mathbf{Y}=\mathbf{A}\boldsymbol{\Theta}+\boldsymbol{E}$$

와 같이 다시 표현할 수 있다. 단, $\mathbf{A}=[\mathbf{X},\mathbf{Z}]$이고, $\boldsymbol{\Theta}=[\mathbf{B}',\boldsymbol{\Gamma}']'$이다. 그리고 Timm(2002, p. 226)과 유사하게 $\mathbf{Q}=\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$를 정의하면, $\boldsymbol{\Theta}$의 추정량은

$$\hat{\boldsymbol{\Theta}}=\begin{bmatrix}\hat{\mathbf{B}}\\ \hat{\boldsymbol{\Gamma}}\end{bmatrix}=\begin{bmatrix}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{Y}-\mathbf{Z}\hat{\boldsymbol{\Gamma}})\\ (\mathbf{Z}'\mathbf{Q}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Q}\mathbf{Y}\end{bmatrix} \tag{2.4}$$

이다.

공분산분석 행렬도를 구현하기 위해서는 먼저 공변량들의 효과를 살펴볼 필요가 있다. 이를 위하여 식 (2.4)의 추정치에서 $\hat{\boldsymbol{\Gamma}}$에 대한 추정치만을 비정칙값분해하여 공변량과 종속변인에 대한 주성분 행렬도를 작성한다. 이 행렬도의 특성은 공변량과 종속변인에 대한 관계를 잘 보여준다. 여기서 효과가 미미한 공변량은 제거하고, 유의한 공변량만을 모형에 추가하여 식 (2.4)의 추정치를 다시 계산한다. 다음으로 최용석 외 2인 (2005)에서 언급한 것처럼 각 처리간 종속변인의 차이점을 알기 위하여 $\hat{\mathbf{B}}$에 대해서는 JK 행렬도를 이용한다. 그리고 $\hat{\boldsymbol{\Gamma}}$에 대해서는 주성분 행렬도를 그려, 두 행렬도의 비교를 통하여 다변량 공분산분석 행렬도를 구현할 수 있게 된다. 끝으로 다변량 공분산분석 행렬도를 위한 알고리즘의 단계를 요약하면 다음과 같다.

<다변량 공분산분석 행렬도의 알고리즘>

[1 단계] 모수행렬의 추정치 $\hat{\Theta}$을 구한다.

[2 단계] 1 단계에서 구한 추정치 $\hat{\Gamma}$의 비정칙값분해를 계산한다.

[3 단계] 비정칙값 인자분해를 통해 공변량과 종속변인간의 주성분 행렬도를 작성하고, 공변량 효과를 검토한다.

[4 단계] 영향력이 큰 공변량을 고려하여 $\hat{\mathbf{B}}$에 대한 JK 행렬도와 $\hat{\Gamma}$에 대한 주성분 행렬도를 각각 그린다.

[5 단계] 4 단계에서 얻어진 두 행렬도의 비교를 통하여 다변량 공분산분석의 결과를 해석한다.

3. 활용사례

3.1 부모의 지위에 따른 유치원 자녀의 표준화 검사

사회경제적 지위가 높은 부모의 유치원 자녀 32명과 지위가 낮은 부모의 유치원 자녀 37명을 대상으로 3가지 표준화 검사 (Peabody Picture Vocabulary Test: PPVT, Raven Progressive Matrices Test: RPMT, Student Achievement Test: SAT)를 실시하였다. 표준화 검사 전 아이들의 수행능력을 어느 정도 예측할 수 있는 PA (paired-associate) 검사를 5개 영역에 대해 실시하

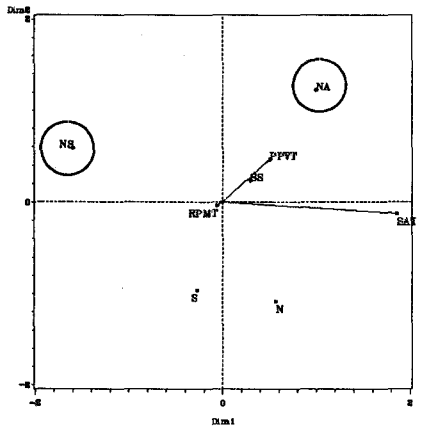
여 점수를 채점하였다 (Timm, 2002, Chapter 4).

<표 1> 전체 공변량에 관한 모수 추정행렬 $\hat{\Gamma}$

공변량	$\hat{\Gamma}$		
	PPVT	RPMT	SAT
N	0.002	0.015	1.605
S	-0.351	0.181	0.026
NS	-0.299	0.112	-2.627
NA	1.294	-0.010	2.106
SS	0.479	-0.004	0.930

* N : named, S : still, NS : named still, NA :named action, SS : sentence still

먼저 <표 1>은 5개의 공변량(PA검사결과)을 고려한 다변량 공분산분석에서 그들에 관한 모수 추정행렬의 결과이다. 이에 대한 주성분 행렬도 <그림 1>을 보면 표준화테스트 PPVT와 SAT는 공변량 NA, SS 그리고 N과 같은 방향에 놓여 있고 NS와 S와는 반대 방향에 놓여 있다. 이는 각각 양과 음의 상관관계를 나타낸다. RPMT는 거의 원점에 가까이 놓여 있어 특정한 공변량과 관련성을 말하기는 힘들다. 그러나 RPMT는 NS와 S의 방향에 있으므로 양의 상관을 가진다. 특히 5개의 공변량 중 NA와 NS의 원점으로부터의 벡터크기가 상대적으로 크므로 그들의 영향력이 가장 큼을 알 수 있다. 실제로 5개 공변량 모두 종속변인의 결과에 유의한 영향을 미치는지 검정해 본 결과, NS와 NA만이 유의확률 0.0047과 0.0012로 유의한 영향을 미치는 것으로 나타났다. 따라서 공변량 NS와 NA만을 고려하여 다변량 공분산분석을 실시하여 <표 2>의 모수 추정행렬 $\hat{B}$과 $\hat{\Gamma}$을 얻었다.

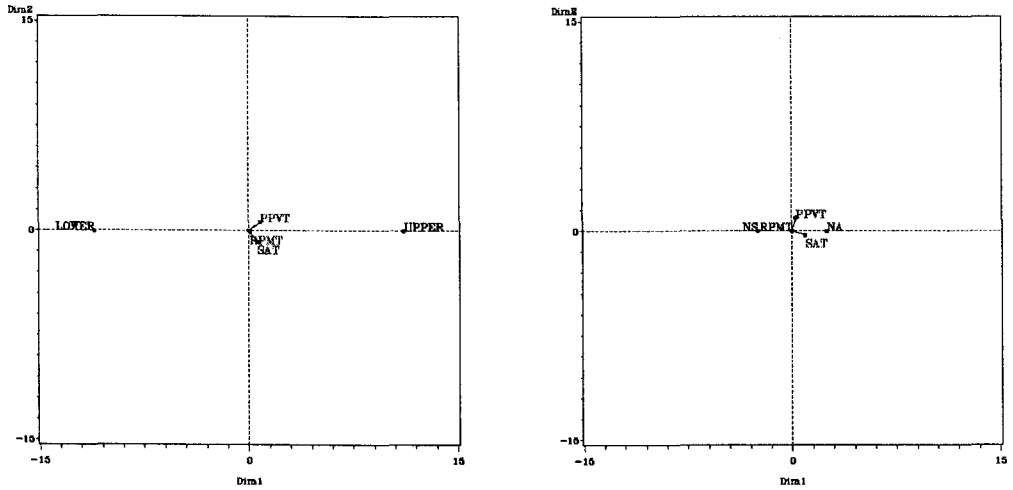


<그림 1> 전체 공변량에 관한 모수 추정행렬 $\hat{\Gamma}$ 의 주성분행렬도

<표 2> 수정된 공분산분석 모형에 관한 모수 추정행렬 $\hat{B}$과 $\hat{T}$

		추정치		
		PPVT	RPMT	SAT
$\hat{B}$	UPPER	81.735	14.873	45.829
	LOWER	63.824	13.353	32.851
$\hat{T}$	NS	-0.117	0.104	-1.937
	NA	1.371	0.068	2.777

<표 2>의 추정치 $\hat{T}$의 주성분 행렬도가 <그림 2>의 (가)이고, $\hat{B}$의 JK 행렬도는 <그림 2>의 (나)이다. 두 그림을 비교하자면 두 그룹 UPPER와 LOWER가 반대방향에 놓여있어 이들 두 그룹 간 평균벡터에 차이가 있음을 나타낸다. 실제로 두 그룹간의 거리가 최용석 외 2인 (2005)의 다변량 분산분석 행렬도에 비해서 좁혀졌다. 이는 공변량 NA와 NS의 영향 때문으로 여겨진다.



(가) (나)

<그림 2> (가) $\hat{T}$의 주성분 행렬도, (나) $\hat{B}$의 JK 행렬도

3.2 신생아 가사 예방약의 효능

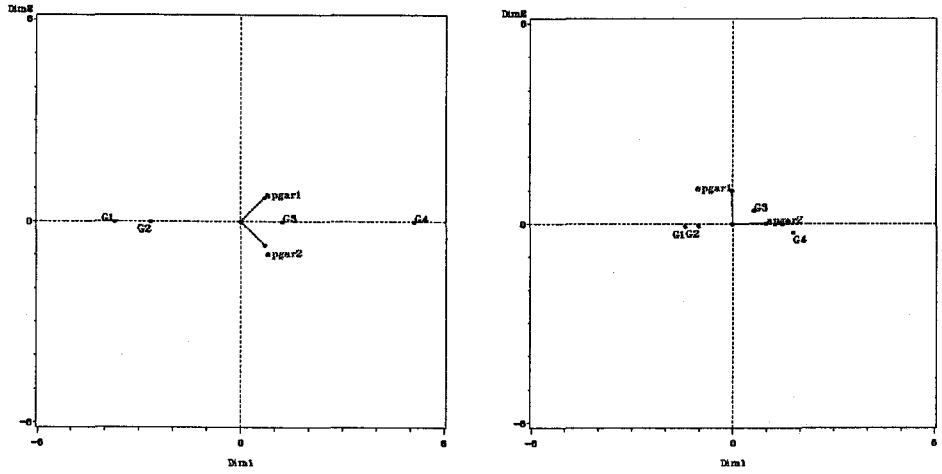
신생아 가사(분만 직후의 신생아에서 심장의 박동은 있으나 호흡이 곤란하거나 또는 정지되어 있는 상태)는 심박수 · 호흡 · 근육수축 · 피부색 및 자극에 대한 반응의 5개 항목에 관한 아프가 점수(apgar score)로 관찰, 판단한다. 가사에 대한 4가지 약의 효과를 측정하기 위해서 쌍둥이를 가진 산모를 8명씩 4개의 집단으로 분류하고, 서로 다른 4개의 가사 예방약을 출산직전 각 집단의 산모들에게 투여한 후, 쌍둥이가 태어난 후 아프가 점수를 각각 채점하였다. 그리고 산도의 압박

과 연관이 있는 산모의 몸무게를 공변량으로 생각하여 측정하였다 (Timm, 1997).

특이한 점은 공변량이 하나인 경우는 종속변인에 대한 공변량의 효과를 Υ의 행렬도로는 검토할 수는 없다. 그러나 공변량을 고려하지 않은 다변량 분산분석 행렬도와 공변량을 고려한 다변량 공분산분석 행렬도를 통하여 공변량의 효과를 볼 수 있다. 먼저 공변량을 고려하지 않은 다변량 분산분석 행렬도를 최용석 외 2인 (2005)로부터 참고로 <그림 3>의 (가)에서 보여주고 있다. 다음으로 공변량의 영향을 고려하여 얻어진 다변량 공분산분석 모형의 모수 추정행렬의 결과 <표 3>에 대한 다변량 공분산분석 행렬도가 <그림 3>의 (나)이다.

<표 3> 다변량 공분산분석 모형의 모수 추정행렬 $\hat{B}$과 Υ

		추정치	
		apgar 1	**apgar 2**
$\hat{B}$	**G1**	5.3101	3.9556
	G2	5.3257	4.3597
	G3	5.7674	6.0227
	G4	5.0969	7.1618
Υ	**WT**	0.1679	0.0791



(가) (나)

<그림 3> (가) 다변량 분산분석 행렬도, (나) 다변량 공분산분석 행렬도

<그림 3>의 (가)를 보면 제 1축(Dim1)을 기준으로 G1과 G2는 음의 방향, G3와 G4는 양의 방향으로 놓여있어 약의 종류에 따라 효능에 차이가 있는 것으로 판단할 수 있다. 실제로 다변량 분

산분석을 실시한 결과 유의확률값이 0.0001보다 작으므로 약의 효능에는 차이가 없다라는 가설을 기각함을 알 수 있었다. 특히, 다변량 분산분석 행렬도 (가)에서는 종속변인인 apgar 1과 apgar 2가 G3와 G4 방향으로 놓여 있어 두 가지 약 G3와 G4를 사용했을 때 이들 아프가 점수가 높게 나타남을 보이고 있다. 더군다나 G3보다는 G4에서 두 종속변인인 apgar 1과 apgar 2 점수가 큰 것으로 나타났다. 따라서, G1과 G2의 효과보다는 G3와 G4의 효과가 더 효과적이라고 할 수 있다. 그러나 다변량 공분산분석 행렬도 (나)에서는 제 1축(Dim1)을 기준으로 볼때 (가)와 같이 약의 효능에 차이가 있어 보이지만 각 그룹들의 좌표점 G1, G2, G3, G4가 원점 주위로 옮겨온 것으로 보아 그 차이는 (가)보다 매우 작음을 보여준다. 실제로 다변량 공분산분석을 한 결과, 다변량 분산분석의 결과와 다르게 유의확률값이 0.05보다 커서 각 약의 효능에는 차이가 없다라는 가설을 기각할 수 없었다. 또한 각 그룹의 좌표가 원점 가까이에 위치한다는 것은 평균의 편차가 작아졌음을 의미한다. 더군다나 종속변인인 apgar 2는 (가)와 경향이 비슷하나 apgar 1은 오히려 제 1축(Dim1)의 관점에서 보면 거의 원점에 있어 약의 종류에 따라 영향을 받지 않음을 의미한다. 이는 공변량으로 산모의 몸무게를 고려했을 때 생기는 효과라고 여겨진다.

4. 결론

공변량이 고려된 다변량 공분산분석의 모형으로부터 추정된 모수행렬의 행렬도는 그래프적으로 공변량의 효과를 한 눈에 파악할 수 있도록 하였다. 특히, 공변량이 고려되지 않은 다변량 분산분석 행렬도와의 비교를 통하여 확연히 공변량의 효과를 눈으로 확인 할 수 있었다. 그러나 본 논문에서 다루어진 공분산분석 행렬도는 일원배치와 완전계수인 경우에 국한되어 있다. 차후에는 이원배치인 경우 및 완전계수가 아닌 자료에 대한 다변량 공분산분석 행렬도의 연구가 더 이루어져야 할 것으로 생각된다.

참고문헌

[1] 장대홍 (1996). 다반응값 자료에 대한 Biplot의 활용에 관한 연구, 「한국통계학회논문집」, 제 3권 1호, 1-9.

[2] 최용석 (1999). 「행렬도의 이해와 응용」, 부산대학교출판부, 부산.

[3] 최용석, 현기홍, 정수미 (2005). 다변량 분산분석에서 추정된 모수 행렬의 행렬도, *Journal of the Korean Data Analysis Society*, 7권 3호, 851-858.

[4] 허명회 (1994). 「SAS 분산분석」, 자유아카데미, 서울.

[5] Gabriel, K. R. (1995). MANOVA biplots for two-way contingency tables *In W. J. Krazanowski (Ed.), Recent Advances in Descriptive Multivariate Analysis*, 227-268. Clarendon Press. Oxford.

[6] Smith, W. F. and Cornell, J. A.(1993). Biplot displays for looking at multiple response data in maxture experiments, *Technometrics*, Vol. 35, No. 4, 337-350.

[7] Timm, N. H. (1997). *Univariate and Multivariate General Linear Models: Theory and Application using SAS Software,* SAS, North Carolina.

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

[2005년 7월 접수, 2005년 10월 채택]