다변량 공분산분석 행렬도

정수미[1], 최용석[2], 현기홍[3]

ABSTRACT

Biplot is a graphical display of the rows and columns of an $n \times p$ data matrix. In particular, Gabriel(1981) 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 exists 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.

KEY WORDS : Covariates, BIPLOT, MANCOVA, MANOVA. SVD

1. 서론

Gabriel(1981)은 변량 자료의 각 처리별로 변수들의 평균값을 구하여 행렬도의 입력자료로 사용하고, 대수적으로는 비정칙치분해(Singular Value Decomposition:SVD)를 하는 다변량 분산분석 행렬도(MANOVA BIPLOT)를 제안했다. 그러나 그의 논문에서는 외생변수의 영향이 있는 자료에 대한 연구가 이루어지지 않았다.

본 논문에서는 Smith와 Cornell(1993), 장대홍(1996)이 다반응값 자료 및 다변량 회귀분석에서 추정된 계수행렬값을 비정칙치분해하여, 반응변수와 설명변수간의 관계를 파악했던 것에 착안하여, 다변량 분산분석 모형의 모수 추정치를 사용하는 다변량 분산분석 행렬도를 제안하였다. 또 공변량의 효과를 고려한 다변량 공분산분석 행렬도도 이와 같은 방법으로 제안하였다.

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

$$\begin{aligned} y_{ij} &= \mu + \alpha_i + \varepsilon_{ij} \\ &= \mu_i + \varepsilon_{ij}. \end{aligned}$$

와 같다. 이 때, $y_{ij} \sim N_p(\mu_i, \Sigma)$는 독립된 평균벡터와 공통된 공분산행렬을 가지는 다변량 정규분포를 따르며, μ_i는 i번째 모평균벡터이고, α_i는 i번째 처리효과이며, $\varepsilon_{ij} \sim N_p(0, \Sigma)$는 오차벡터로 평균벡터가 0이고, 공분산행렬이 Σ 인 정규 확률변수이다. 이를 행렬로 나타

1) 부산대학교 자연과학대학 통계학과 석사과정 졸업.
2) 부산대학교 자연과학대학 통계학과 교수.
3) 부산대학교 자연과학대학 통계학과 박사수료.

내면 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, Y는 $n\times p$의 관측치행렬이고, $\mathbf{X}=\begin{bmatrix} 1_{n_1} & 0 & \cdots & 0 \\ 0 & 1_{n_2} & \cdots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \cdots & 1_{n_k} \end{bmatrix}$의 $n\times k$의 다변량 분산분석 설계행렬이며, $\mathbf{B}=\begin{bmatrix} \mu_1' \\ \mu_2' \\ \vdots \\ \mu_k' \end{bmatrix}$는 $k\times p$의 모평균행렬이고, $\boldsymbol{E}=\begin{bmatrix} \varepsilon_1' \\ \varepsilon_2' \\ \vdots \\ \varepsilon_n' \end{bmatrix}$는 $n\times p$의 오차행렬이다.

모수행렬 B의 추정치는 식 (2.2)와 같이 나타낼 수 있다.

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

다변량 분산분석 모형의 추정된 모수행렬인 식 (2.2)를 비정칙치분해하여 행렬도를 그린다. 각 처리간 종속변인의 차이점을 알고자하는 다변량 분산분석의 목적과 관측치간의 관계를 보고자 하는 JK 행렬도의 특징이 일치하므로 이를 이용한다. 모형의 추정치는 통계패키지의 모듈로부터 쉽게 구할 수 있으므로, Gabriel의 직접 프로그램하는 방법보다 용이하게 같은 결과를 얻을 수 있는 장점이 있다.

다변량 공분산분석 모형은 위의 다변량 분산분석과 다변량 회귀모형이 결합된 형태로

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

와 같이 나타낼 수 있다.

X는 다변량 분산분석 설계행렬이고, Z는 h개의 공변량이 포함된 다변량 회귀 모형의 행렬이다. 식 (2.3)는

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

와 같이 다시 표현할 수 있다. 단, $\mathbf{A}=[\mathbf{X},\mathbf{Z}]$이고, $\boldsymbol{\Theta}=\begin{bmatrix} \mathbf{B} \\ \boldsymbol{\Gamma} \end{bmatrix}$이다.

Timm(2002)과 같이 $\mathbf{Q}=\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$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. 식 (2.4)의 추정치에서 $\hat{\boldsymbol{\Gamma}}$에 대한 추정치만을 비정칙치분해하여 공변량과 종속변인에 대

한 GH 행렬도를 작성한다. 공변량벡터를 종속변수들의 벡터에 투영시켰을 때 정사영의 크기를 곱하면 이들의 관계를 살펴볼 수 있다. 여기서 효과가 미미한 공변량은 제거하고, 유의한 공변량만을 모형에 추가하여 식 (2.4)의 추정치를 다시 계산한다. 그런 다음 $\hat{\mathbf{B}}$과 $\hat{\boldsymbol{\Gamma}}$ 각각에 관하여 행렬도를 그리면 다변량 공분산분석 행렬도를 구현할 수 있게 된다.

3. 활용사례

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 실시하여 점수를 채점하였다.

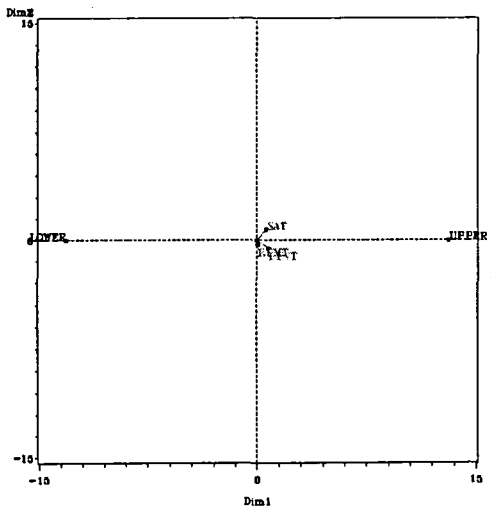
먼저 다변량 분산분석을 실시한 결과 유의확률값이 0.0001보다 작았다. 그러므로 두 집단(UPPER, LOWER)간 종속변인의 평균벡터가 다르다. 이것을 행렬도로 살펴보자. 2절의 식 (2.1)의 다변량 분산분석 모형에 대한 모수행렬 추정치를 <표 1>과 같이 얻을 수 있고, 이를 비정칙치분해하여 <그림 1>과 같은 행렬도를 그릴 수 있다. 제 1축을 기준으로 봤을 때, UPPER 그룹과 LOWER 그룹이 반대방향에 놓여있다. 행렬도에서 행은 좌표점으로 그려지며 이들 사이의 거리는 마할라노비스(Mahalanobis) 거리를 나타낸다. 이는 두 그룹이 다른 경향의 군집을 이룸을 의미하므로 그룹간 평균벡터에 차이가 있다고 할 수 있다. 또, 세 종속변인 PPVT, RPMT, SAT 모두 UPPER 그룹과 같은 방향에 놓여 양의 상관을 가지므로, 사회경제적 지위가 높을수록 표준화 검사 점수가 높게 나타나는 것으로 보인다.

<표 1> 다변량 분산분석 모형의 모수 추정행렬 $\hat{\mathbf{B}}$

집단	$\hat{\mathbf{B}}$		
	PPVT	RPMT	SAT
UPPER	83.0938	15.0000	47.6562
LOWER	62.6486	13.2432	31.2703

* UPPER : 높은 지위 집단, LOWER : 낮은 지위 집단

다변량 공분산분석 행렬도



<그림 1> 다변량 분산분석 행렬도

이제 공변량(PA검사결과)을 고려한 다변량 공분산분석의 결과를 살펴보자. 5개공변량 모두 종속변인의 결과에 유의한 영향을 미치는지 검정해본 결과, NS와 NA만이 유의확률 0.0047과 0.0012로 유의한 영향을 미치는 것으로 나타났다. 행렬도에서는 어떻게 나타나는지 살펴보자.

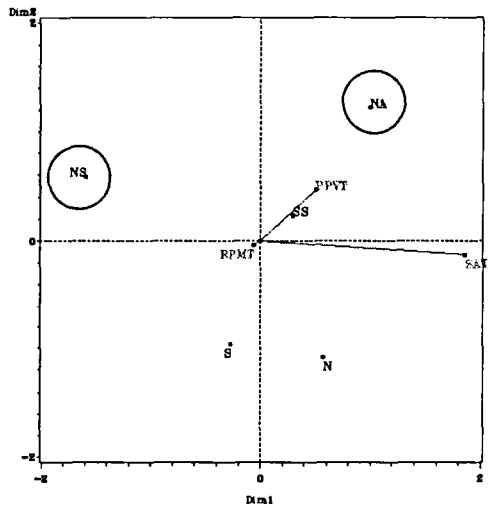
<표 2> 전체 공변량에 관한 모수 추정행렬 T

공변량	T		
	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 action, NA : named still, SS : sentence still

<그림 2>를 보면 표준화테스트 PPVT, SAT는 공변량 NS, SS 그리고 N과 같은 방향에 놓여 있는데, 이는 양의 상관관계를 가짐을 의미한다. 또, NS, S와는 반대 방향에 놓여 있다. 이는 음의 상관관계를 가짐을 의미한다. RPMT는 거의 원점에 가까이 놓여 있어 특정한 공변량과 관련성을 말하기는 힘들다. 그러나 좌표점이 NS와 S의 방향에 가까우므로 그들과 양의 상관을 가진다고 할 수 있다. 그리고 종속변인 PPVT, RPMT는 양의 상관으로는 공변량 NA가, 음의 상관으로는 공변량 NS의 벡터크기가 가장 크므로 그들의 영향력이 가장 크며, 종속변인 SAT는 양의 상관으로 NS가, 음의 상관으로 NA벡터의 크기가 가장 크므로 그들의 영향력이 가장 크다. 즉, 종속변인에 공변량 NS와 NA가 가장 큰 영향을 미치므로 5개의 공변량 중 NS와 NA만을 고려하여 다변량 공분산분석을 실시하였다. 다변량 공분산분석의 결과 유의확률값이 0.0001보다 작게 나타났고, 이는 다변량 분산분석의 결과와 같다. 공변량 효과가 존재하지만, 전

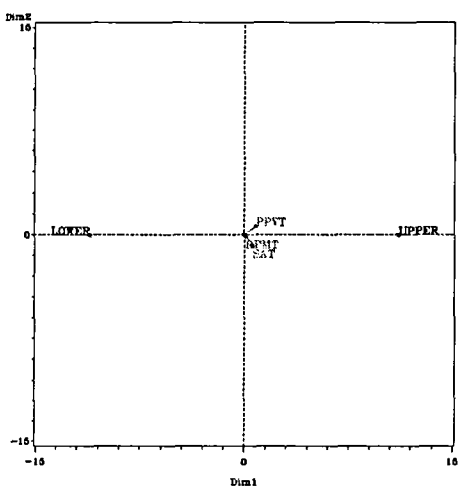
체적 결론은 공변량의 영향이 마치 없는 것처럼 나타났다. 행렬도에서는 이런 유의성 검정 방법과는 다르게 공변량의 영향을 확인할 수 있다.



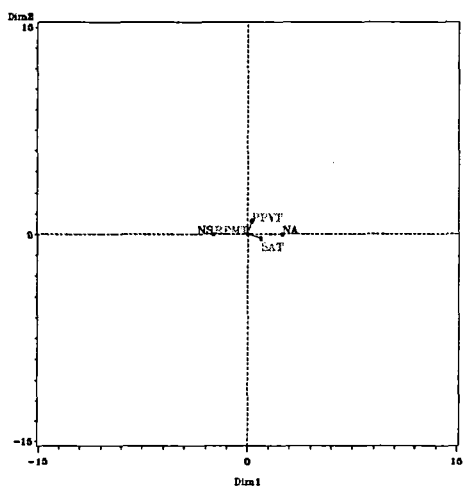
<그림 2> 전체 공변량과 종속변수간 행렬도

<표 3> 수정된 공분산분석 모형에 관한 모수 추정행렬 $\hat{B}$와 $\hat{T}$

		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



<그림 3> 유의한 공변량과 종속변수간 행렬도



<그림 4> 다변량 공분산분석 행렬도

추정치 $\hat{T}$으로 그린 행렬도가 <그림 3>이고, $\hat{B}$으로 그린 행렬도는 <그림 4>이다. 두 그림을 마치 겹쳐놓은 그림이라 생각하여 보자. 두 그룹 UPPER와 LOWER가 반대방향에 놓여있어, 앞서 말한 것과 같이 그룹간 평균벡터에 차이가 있음을 나타낸다. 그러나 두 그룹간의 거리가 <그림 1>의 다변량 분산분석 행렬도에 비해서 좁혀졌다. 이는 두 그룹의 차이가 다변량 분산분석에 비해서는 작음을 나타낸다. 이와 같은 결과에 공변량들이 미친 영향을 알아보면, 공변량 NA는 종속변인 PPVT, SAT에 양의 영향을, NS는 음의 영향을 미친다. 그리고 거의 원점에 위치해 있는 RPMT에는 공변량이 큰 영향을 주지 않는다. 이런 공변량의 영향 때문에 두 그룹간의 거리가 다변량 분산분석 행렬도에 비해 좁혀진 것으로 해석할 수 있다.

4. 결론

다변량 분산분석 및 공분산분석의 모형으로부터 모수행렬을 추정하여, 이를 비정칙치분해함으로써 행렬도를 그릴 수 있다. 공변량이 있는 자료의 경우, 공변량의 효과를 행렬도로 나타내어 다변량 공분산분석의 결과값에 어떤 영향을 미치는지 한 눈에 파악할 수 있다. 그러나 본 논문에서 다루어진 다변량 분산분석 및 공분산분석 행렬도는 일원배치와 완전계수인 경우에 국한되어 있다. 차후에는 이원배치인 경우 및 완전계수가 아닌 자료에 대한 다변량 분산분석 및 공분산분석 행렬도의 연구가 더 이루어져야 할 것으로 생각된다.

참고문헌

장대홍 (1996). "다반응값 자료에 대한 Biplot의 활용에 관한 연구" 한국통계학회논문집 3(1), 1-9.

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

Gabriel, K.R. (1981). Biplot display of multivariate matrices for inspection of data and diagnosis *In V. Barnett(Ed.) Interpreting Multivariate Data*, 147-173, Wiley, London.

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

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