Journal of Korean
Data & Information Science Society
2004, Vol. 15, No. 2, pp. 297~306

Improving Interpretability of Multivariate Data Through Rotations of Artificial Variates[1)]

S.Y. Hwang[2)] · A.M. Park[3)]

Abstract

It is usual that multivariate data analysis produces related (small number of) artificial variates for data reduction. Among them, refer to MDS(multidimensional scaling), MDPREF(multidimensional preference analysis), CDA(canonical discriminant analysis), CCA(canonical correlation analysis) and FA(factor analysis). Varimax rotation of artificial variables which is originally invented in FA for easy interpretations is applied to diverse multivariate techniques mentioned above. Real data analysisis is performed in order to manifest that rotation improves interpretations of artificial variables.

Key words : Artificial variable, Interpretation, Varimax rotation,

1. 서론

다변량 자료 분석의 주요 관심사 중의 하나는 분석 후 나온 결과에서 축이 어떠한 의미를 갖고 있는가 하는 것이다. 인자분석, 다차원척도법, 다차원선호도분석, 정준판별분석, 정준상관분석이 그 대표적인 예라고 할 수 있다. 그러나 최초로 구해진 결과들은 때때로 해석하기 어려운 경우가 있다. 이러한 문제점을 해결하기 위한 방법으로 인자분석에서는 축을 회전하는 방법을 사용하여 왔다. 인자를 회전하는 방법에는 직교회전(Orthogonal Rotation)과 사각회전(Oblique Rotation)이 있다. 직교회전은 인자들이 서로 독립적일 때, 회전 후 인자가 직각이 되도록 회전시키는 방법으로 베리맥스 회전(Varimax Rotation)이 주로 사용된다. 또한 사각회전은 인자들이 상관이 있다는 가정 하에서 회전후의 인자가 비직각이 되도록 인자를 회전하는 방법으로

---

1) This work was supproted by a research grant 2003 from Sookmyung Women's University.

2) First Author : Professor, Department of Statistics, Sookmyung Women's Univ., Seoul, 140-742, Korea. Email : shwang@sookmyung.ac.kr

3) Graduate student, Sookmyung Women's University, Seoul, 140-742, Korea.

Promax, Harris-Kaiser 방법 등이 있으나 이 회전방법은 인자 회전 후 인자의 상관관계를 다시 해석하여야 하는 문제가 발생한다(성웅현, 2000).

인자분석에서 베리맥스 회전이란 p-변량, m 개 인자로 구성된 초기인자적재행렬 L 에 직교행렬 T 를 곱해서 얻은 p x m 행렬 L^* , 즉 $L^*=L\times T$ 에서

$$V=\frac{1}{p}\sum_{j=1}^{m}[\sum_{i=1}^{p} l_{ij}^{*4}-(\sum_{i=1}^{p} l_{ij}^{*2})^2/p] \tag{1.1}$$

를 가장 크게 하는 직교회전행렬 T 를 찾는 것이다. 여기서 l_{ij}^* 는 L^* 행렬의 원소를 표시한다.

지금까지 위의 베리맥스 회전은 인자분석의 경우에만 적용되어왔다. 하지만 인자분석 이외의 다변량 분석기법들(예를 들어 다차원척도법, 다차원선호도분석, 정준판별분석, 정준상관분석 등(최용석,정광모(2001) 참고))에서도 축에 내재되어있는 의미를 찾기를 원한다. 인자분석 이외의 나머지 분석기법들에서도 분석자들의 자료에 관한 지식, 사전정보, 경험에 의해 축이 해석되어져 왔다. 하지만 이는 분석자들의 주관이나 지식이 많이 개입될 수 있고, 또한 축의 해석이 어려운 경우가 남아있을 수 있다. 본 논문에서는 이러한 문제점들을 해결하기 위하여 인자분석에서 가장 많이 이용되고 있는 베리맥스 회전을 인자분석 이외의 다변량 자료 분석에서도 적용하여 보고, 이를 구현할 수 있는 방법을 제시하고자 한다.

2. 직교 회전행렬

2차원 $(m=2)$ 의 경우 직교회전행렬(cf. Johnson and Wichern(2002))은

$$T=\begin{cases}\begin{bmatrix}\cos\theta & \sin\theta\\ -\sin\theta & \cos\theta\end{bmatrix}, & \text{시계 방향 회전}\\ \begin{bmatrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta\end{bmatrix}, & \text{시계 반대 방향 회전}\end{cases} \tag{2.1}$$

의 두 가지 경우가 있다. 본 논문에서는 시계 반대 방향으로 회전하는 행렬을 사용하기로 한다. SAS/IML을 이용하여 회전 각도를 1 도씩 늘려가며 V를 최대로 하는 직교회전행렬을 찾도록 하였다.

3차원 $(m=3)$ 인 경우에는 여러 가지의 직교회전행렬이 존재할 수 있으나, 여기서는 먼저 x - y 축으로 θ 만큼 회전한 후, x - z 축으로 ρ 만큼 회전한다고 생각하여 다음과 같은 직교회전행렬을 적용하기로 한다.

[1 단계] x - y 축으로 θ 만큼 회전(z 좌표를 중심으로)

$$\begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} x\cos\theta - y\sin\theta \\ x\sin\theta + y\cos\theta \\ z \end{bmatrix}$$

[2단계] 1단계후 x - z 축으로 ρ 만큼 회전(y 좌표를 중심으로)

$$\begin{bmatrix} x'' \\ y'' \\ z'' \end{bmatrix} = \begin{bmatrix} \cos\rho & -\sin\rho & 0 \\ 0 & 1 & 0 \\ \sin\rho & \cos\rho & 0 \end{bmatrix} \begin{bmatrix} x' \\ y' \\ z' \end{bmatrix}$$

$$= \begin{bmatrix} \cos\rho & -\sin\rho & 0 \\ 0 & 1 & 0 \\ \sin\rho & \cos\rho & 0 \end{bmatrix} \begin{bmatrix} x\cos\theta - y\sin\theta \\ x\sin\theta + y\cos\theta \\ z \end{bmatrix}$$

$$= \begin{bmatrix} x\cos\rho\cos\theta - y\cos\rho\sin\theta - z\sin\rho \\ x\sin\theta + y\cos\theta \\ x\sin\rho\cos\theta - y\sin\rho\sin\theta + z\cos\rho \end{bmatrix}$$

따라서 본 논문에서는 3차원 직교회전행렬로

$$T = \begin{bmatrix} \cos\rho\cos\theta & -\cos\rho\sin\theta & -\sin\rho \\ \sin\theta & \cos\theta & 0 \\ \sin\rho\cos\theta & -\sin\rho\sin\theta & \cos\rho \end{bmatrix} \tag{2.2}$$

를 이용하기로 한다. 2차원에서와 마찬가지로 SAS/IML을 이용하여 θ 와 ρ 를 모두 1°씩 회전시켜 베리맥스 회전의 V를 최대화 시켜주는 직교회전행렬을 찾도록 하였다.

3. 다변량 분석기법에의 적용

이번 장에서는 앞 장에서 제시했던 베리맥스 회전을 다차원척도법, 다차원선호도 분석, 정준판별분석에 각각 적용하여 회전한 후 만들어진 축이 어떠한 의미를 가지고 있는지 살펴보겠다. 각 적용사례에 대한 자세한 프로그램은 저자들에게 연락하면 얻을 수 있다.

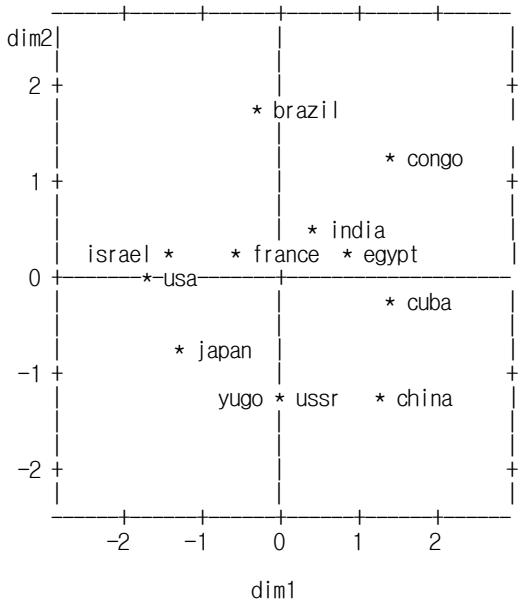
3.1 다차원 척도법(Multidimensional Scaling)

자료 : 12개 국가의 유사성에 대한 평가 (출처: 성웅현, 2000).

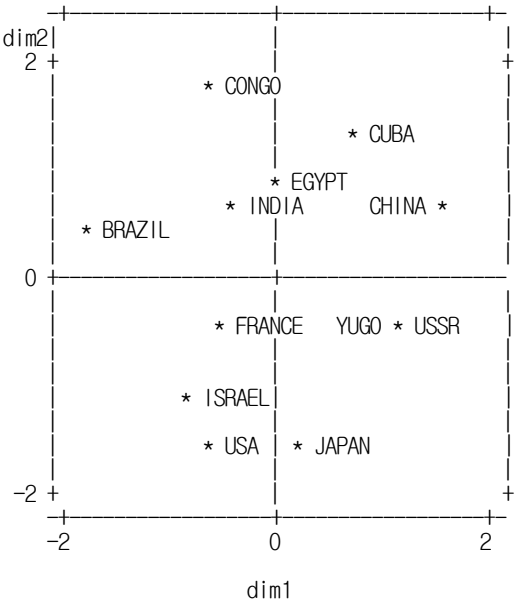
18명의 학생들이 12개 국가를 각각 두 나라씩 짝지은 66가지 경우에 대해 1(매우 다르다)~9(매우 유사하다)의 척도로 응답하여 그 평균으로 12개국의 유사성행렬을 얻었다. 자료의 내용은 성웅현(2000, p. 362)를 참조하기 바란다.

<표 3.1-1> 다차원 척도법 분석결과

		회전하기 전		회전한 후	
		dim1	dim2	dim1	dim2
인조변수행렬	brazil	-0.33606	1.77663	-1.77349	0.35223
	congo	1.41020	1.28037	-0.66058	1.78652
	cuba	1.42931	-0.27069	0.78523	1.22459
	egypt	0.78777	0.31172	0.00526	0.84718
	france	-0.61103	0.31007	-0.51596	-0.45088
	india	0.40025	0.60868	-0.41500	0.59872
	israel	-1.45195	0.34994	-0.86720	-1.21596
	japan	-1.30390	-0.78164	0.23772	-1.50153
	china	1.26301	-1.18007	1.56658	0.73049
	ussr	0.00538	-1.23044	1.14330	-0.45484
	usa	-1.66362	0.03021	-0.64974	-1.53179
	yugo	0.07063	-1.20478	1.14388	-0.38473
회전각도	$\theta = 292°$				



<그림 3.1-1> 회전하기 전



<그림 3.1-2> 회전한 후

<그림 3.1-1>을 보면 축의 의미를 바로 해석해 내기 어렵다. 하지만 <그림 3.1-2>를 보면 즉, 제1차원축을 정치적노선(친서방국가-친공산국가), 제2차원축을 경제개발

(선진국-개발도상국가) 차원축으로 해석할 수 있는 장점이 있다.

3.2 다차원선호도 분석

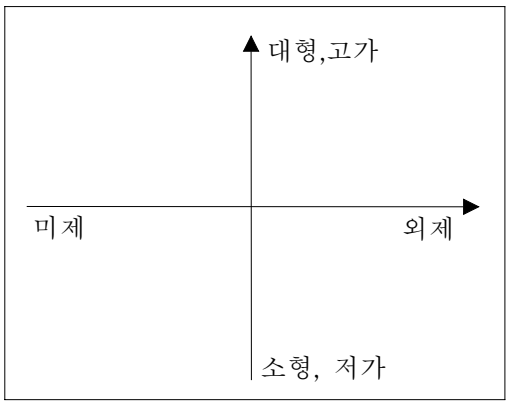
다차원선호분석(multidimensional preference analysis; MDPREF)은 이원도(biplot)의 일종으로서(Gabriel, 1971), 여러 상품들과 이들 상품에 대한 개인 또는 그룹의 선호도를 알아보기 위한 분석 방법으로 결과는 보통 2차원 그림으로 제공된다. 다차원선호분석은 선호도 자료를 중심으로 한 분석기법으로, 이 선호도 자료는 소비자들의 제품에 대한 선호, 혹은 제품 속성에 대한 각 제품별 평가를 말한다. 즉, 다차원선호분석은 이러한 두 종류의 대상을 동시에 공통공간에 기하학적인 구조로 나타내 준다.

자료 : 17개 자동차에 대한 선호도 (자료 출처: 이경일 외, 1993)

다음 자료는 미국 승용차 시장에서 17개의 모델(평가대상)에 대하여 25명의 평가자가 각각에 대한 선호도를 나타낸 것이다. 0~9까지 등급으로 0점으로 갈수록 매우 선호하지 않음을 의미하고, 9점으로 갈수록 매우 선호함을 뜻한다. 결과를 보면, 최적 회전각도는 22° 로 판명되었으며 <그림 3.2-1>과는 달리, <그림 3.2-2>의 dim1 축을 보면 좌측에 있는 승용차의 메이커는 미국이며, 우측에 있는 승용차들은 미국 이외의 나라들 제품이다. 즉, 이 두 그룹의 차이는 미국산과 외제로 볼 수 있다. dim2축의 경우 위측에는 대형, 고가격 종류의 차종이고 아래측에는 소형, 저가격 종류의 차종임을 알 수 있다. 즉, dim2축에 따른 두 그룹의 차이는 크기와 가격이라고 볼 수 있다.

<표 3.2-1> 17개 자동차에 대한 선호도 자료

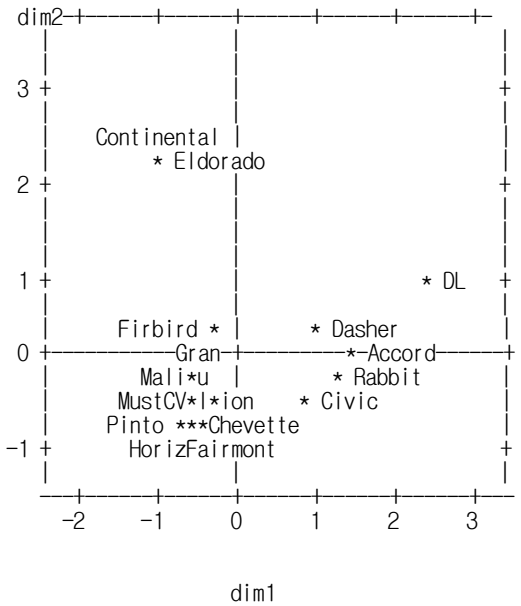
제조회사	모델명	선호도
Cadilac	Cadilac	80079904912405089710938 09
Chevrolet	Chevrolet	00512004234510430035156 98
Chevrolet	Chevrolet	40533058141616435447477 95
…	…	…
Volkswagen	Volkswagen	48585097096957954878850 00
Volvo	Volvo	99899989099999879899190 00



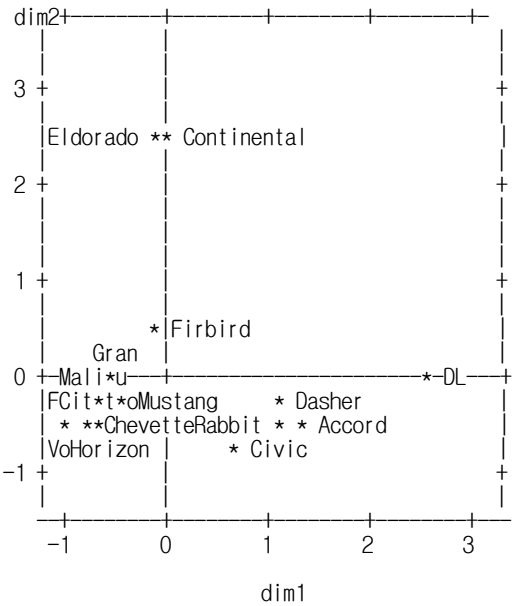
<그림 3.2-1> 분석 결과

<표 3.2-2> 다차원 선호도 분석 결과

	회전하기 전		회전한 후	
		dim1	dim2	dim1	dim2
인조변수행렬	Eldorado	-1.06117	2.35377	-0.102351	2.579892
	Chevette	-0.55718	-0.79036	-0.81264	-0.524148
	Citation	-0.44200	-0.72088	-0.679828	-0.502863
	Malibu	-0.53592	-0.55493	-0.704752	-0.313812
	Fairmont	-0.48902	-0.73090	-0.727175	-0.494542
	Mustang	-0.32998	-0.48664	-0.488229	-0.327631
	Pinto	-0.70438	-0.81702	-0.959117	-0.493732
	Accord	1.38591	0.04048	1.3001895	-0.481544
	Civic	0.88338	-0.46144	0.6462491	-0.758716
	Continental	-0.98082	2.32094	-0.04014	2.519355
	Gran Fury	-0.56490	-0.20849	-0.601871	0.018259
	Horizon	-0.51026	-0.68787	-0.730757	-0.446687
	Volare	-0.55451	-0.62874	-0.749631	-0.375286
	Firebird	-0.30397	0.37177	-0.142602	0.458553
	Dasher	1.05781	0.20650	1.0581602	-0.204721
	Rabbit	1.30380	-0.13210	1.1594188	-0.610812
	DL	2.40321	0.92593	2.5750747	-0.041567
회전각도	$\theta = 22°$				



<그림 3.2-2> 회전하기 전



<그림 3.2-3> 회전한 후

3.3 정준 판별 분석

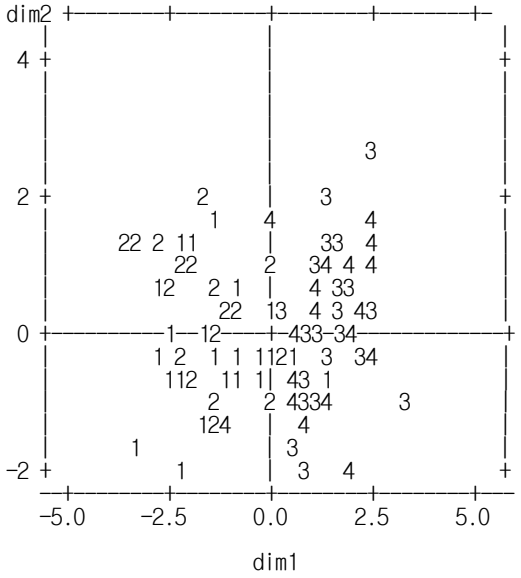
정준판별분석(canonical discriminant analysis)은 소속모집단에 대한 가정을 하지 않고 단지 판별변수의 적절한 선형결합을 통하여 그룹을 분리하고자 하는 다변량 분석기법이다. 판별분석의 주요목적은 여러 집단 사이의 중복을 최소화하면서 구분할 수 있는 판별함수를 유도하는 것이다. 그러나 표본의 크기가 아주 작지 않는 한, 각 개체에 대한 판별점수를 살펴보거나 집단별 평균만을 고려함으로써는 충분한 정보를 얻지 못할 것이다. 이때, 각 개체의 판별점수를 판별공간에 구성함으로써 집단별 판별점수 평균의 위치와 평균을 중심으로 흩어진 정도가 개괄적으로 파악될 수 있을 것이다.

자료 : 태생에 따른 연어 판별 (자료 출처: 최용석, 정광모(2001))

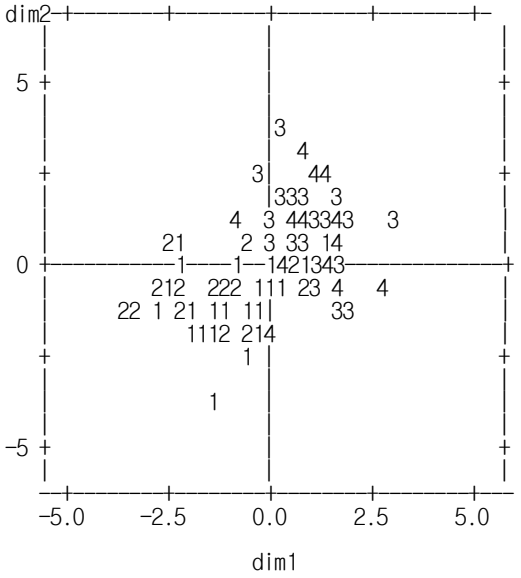
연어의 태생 장소에 따른 몇 가지 특징 중 하나는 생장 고리에 있으며 대체로 알래스카 태생이 캐나다 산에 비해 지름이 작은 특징이 있다. 판별변수는 2개로서 강(fresh water)과 바다(marine)에서의 생장고리이며 집단은 모두 4개로서 집단1; 알래스카 태생 male, 집단2; 알래스카 태생 female, 집단3; 캐나다 태생 male, 집단4; 캐나다 태생 female 이다.

<표 3.3-2> 정준 판별 분석 결과

	회전하기 전			회전한 후		
표준정준계수		dim1	dim2		dim1	dim2
	freshwater	0.916766	0.399425	freshwater	0.26880	1.49237
	marine	-0.831183	0.555999	marine	-1.32867	0.32718
회전각도	$\theta = 321°$					



<그림 3.3-1> 회전하기 전



<그림 3.3-2> 회전한 후

<그림 3.3-1>에서 dim1축을 따라서 알라스카산 연어는 주로 좌측에, 캐나다산 연어는 주로 우측에 위치하고 있음을 알 수 있다. 하지만 dim2축에 대해서는 모두 알래스카산과 캐나다산의 구분이 없음을 알 수 있다. 하지만, 회전된 결과인 <그림 3.3-2>에서는 주로 1사분면(dim1축과 dim2축 모두 양 (+) 의 부호)에 캐나다산 연어가 위치해 있고, 3사분면(dim1축과 dim2축 모두 음 (−) 의 부호)에 알래스카산 연어가 위치하고 있음을 알 수 있다.

3.4 인자분석을 이용한 3차원 적용사례

자료 : 6개의 시험과목에 대한 평가 (자료 출처: Lawley 외, 1971).

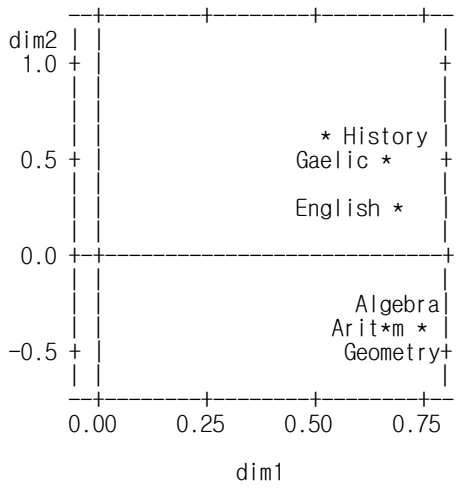
<표 3.4-1>은 220명의 남학생들의 6과목(게일어, 영어, 역사, 산술, 대수, 기하)의 시험점수에 관한 상관행렬을 보여 주고 있다.

<표 3.4-1> 6개의 시험과목 평가 자료

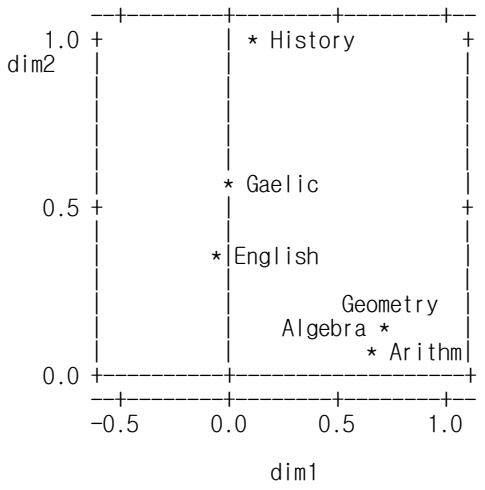
Gaelic	1.000	0.439	0.410	0.288	0.329	0.248
English	0.439	1.000	0.351	0.354	0.320	0.329
History	0.410	0.351	1.000	0.164	0.190	0.181
Arithmetic	0.288	0.354	0.164	1.000	0.595	0.470
Algebra	0.329	0.320	0.190	0.595	1.000	0.464
Geometry	0.248	0.329	0.181	0.470	0.464	1.000

<표 3.4-2> 인자분석 결과

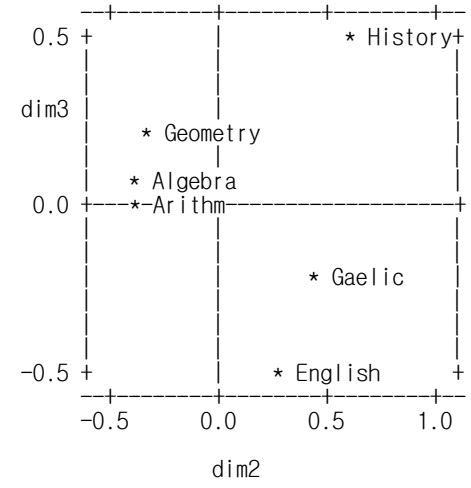
	회전하기 전				회전한 후		
인자적재 행렬		dim1	dim2	dim3	dim1	dim2	dim3
	Gaelic	0.65782	0.44905	-0.18658	-0.02043	0.57775	-0.57877
	English	0.68842	0.29039	-0.50965	-0.06204	0.33108	-0.83936
	History	0.51737	0.63734	0.52656	0.13693	0.96454	0.04522
	Arithm	0.73831	-0.41303	0.00065	0.68539	0.04930	-0.49347
	Algebra	0.74388	-0.37545	0.10674	0.71734	0.12697	-0.41834
	Geometry	0.67831	-0.35501	0.17880	0.70353	0.14203	-0.32092
회전각도	$\theta = 329°,\ \rho = 42°$						



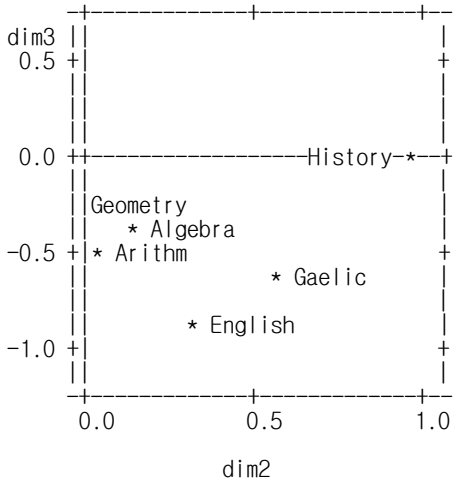
<그림 3.4-1> 회전하기 전(dim2*dim1)



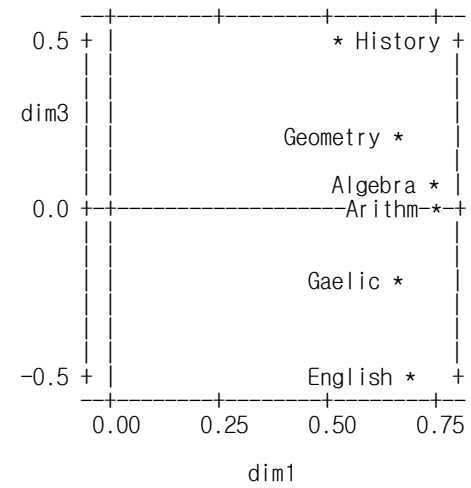
<그림 3.4-2> 회전한 후(dim2*dim1)



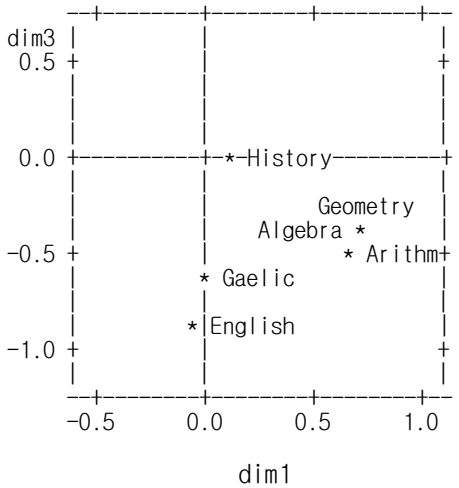
<그림 3.4-3> 회전하기 전(dim3*dim2)



<그림 3.4-4> 회전한 후(dim3*dim2)



<그림 3.4-5> 회전하기 전(dim3*dim1)



<그림 3.4-6> 회전한 후(dim3*dim1)

먼저 회전하기 전 인자적재 행렬을 보면 dim1의 경우 모든 교과목에 대해 비교적 유사한 값을 나타내며, dim2는 어휘력과 수리력인자간의 대비를 나타낸다고 볼 수 있겠다. 2절에서 소개한 3차원 회전인 식(2.2)를 적용한 결과 최적각도가 $\theta=319$ 도, $\rho=42$ 도 인 것으로 나타났으며 회전한 후의 dim1은 산술과 대수, 기하에서, dim2는 역사에 높은 인자적재를 나타내고 있음을 알 수 있다. 그리고, 회전한 후의 dim3은 게일어와 영어에서 높은 인자적재를 가지고 있음을 알 수 있다. 즉, 정리하여 보면 dim1은 수리력인자, dim2는 역사인자, dim3은 어휘력인자라고 할 수 있겠다. <그림 3.4-1>부터 <그림 3.4-6>까지를 살펴보아도 인자적재행렬에서 찾아낸 사실을 그대로 시각적으로 보여 주고 있음을 알 수 있다.

지금까지의 적용 사례에서 보았듯이 베리맥스 방법을 사용하여 회전한 후 축을 해석하는 것이 조금은 더 편리하고 객관적 판단을 내릴 수 있도록 도움을 주는 것을 알 수 있었다. 물론 각 분석별로 많은 사례에 적용하여 본 것이 아니므로 축의 해석이 어려운 모든 경우에 있어서 축의 회전이 가장 좋은 방법이라고는 할 수 없다. 그러나 본 논문을 통해서 축에 내재되어 있는 의미를 보다 잘 해석할 수 있는 하나의 방법을 제시한 점에 큰 의미를 부여하고 싶다. 또한, 앞으로 각 분석에 있어서 회전하기 전과 회전한 후의 성질들에는 어떠한 변화가 있는가 하는 부분에 있어서 추가적인 연구가 필요할 것이라고 사료된다.

참 고 문 헌

1. 성웅현 (2000), 응용 다변량 분석, 탐진.
2. 이경일, 박종규 (1993), 다차원 척도법(MDS)과 컨조인트분석 활용과 결과해석, 홍릉과학출판사.
3. 최용석, 정광모 (2001), SAS를 활용한 응용 다변량 자료분석, 교우사.
4. Gabriel, K.R (1971), The biplot graphics display of matrices with applications to principal component analysis, *Biometrika*, 58, 3, pp 453-467.
5. Johnson, R.A. and Wichern, D.W. (2002), *Applied Multivariate Statistical Analysis*, Prentice Hall.
6. Lawley, D.N. and Maxwell, A.E. (1971). *Factor Analysis as a Statistical Methods,* Elsevier, N.Y.

[2003년 11월 접수, 2004년 3월 채택]