성신여자대학교 교수학습지원센터
2013학년도 1학기 연구방법론 워크샵

Exploratory Factor Analysis

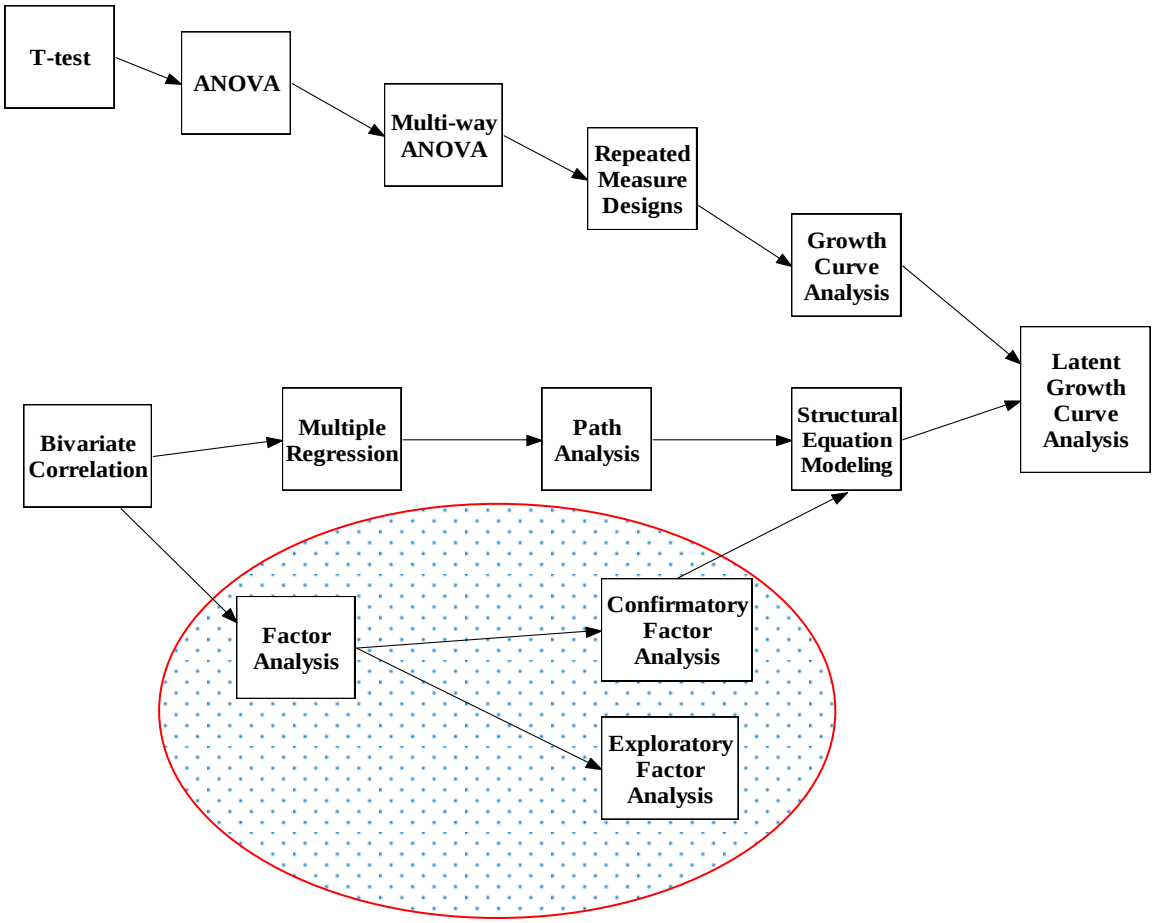
I. 2013년 5월 2일 (목) 7-9PM
II. 2013년 5월 9일 (목) 7-9PM

강사: 강태훈 (성신여대 교육학과)

< 목 차 >

1. 들어가며

➢ 통계적 분석 방법 중 요인분석의 위치



➢ 요인분석의 역사

- ✓ 요인분석은 심리학으로부터 탄생한 몇 안 되는 다변량(multivariate) 분석 기법 중 하나로서, Charles Spearman이 1904년 발표한 논문에서 인간 지능의 구조를 연구하기 위한 방법론(이요인 이론, two-factor theory)으로 제시하였다: "General intelligence, objectively determined and measured" (American Journal of Psychology)

- ✓ Louis Leon Thurstone (1887-1955) 는 요인분석의 다요인이론(multi-factor theory)을 개발하고 보편화시키는 업적을 세웠으며, 오늘날 우리가 요인분석이라고 부르는 것은 바로 이 "다요인이론"을 의미한다.

- ✓ Karl Joreskog 은 요인에 대한 선행 이론이 존재할 경우 이를 자료를 통해 확인할 수 있는 확인적 요인분석을 제시하였다. 또한 이를 확장하여 '구조방정식 모형'(structural equation modeling)을 개발하는 데 있어서도 독보적인 공헌을 하였다.

  Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika, 34,* 183–202.

  Jöreskog, K. G. (1970). "A general method for analysis of covariance structures". *Biometrika , 57*, 239–251.

"~ 보이는 것은 나타난 것으로 말미암아 된 것이 아니니라"
(히브리서 11장 3절)

➢ 세 가지 대표적 요인분석 이론

✓ Spearman's Two-factor theory (or g-factor theory)

인간의 지능에서는 일반요인(general factor; g)이 거의 모든 인간의 행동 영역에 중요한 작용을 하며, 이 요인에 의해 설명되는 것 외에 구체적 혹은 독특한 여러 요인들(specific factors 혹은 unique factors; 언어, 수리, 공간, 기계 등등)이 존재한다.

사실 여기서 말하는 unique factor는 각각의 지능검사 문항에 개별적으로 영향을 미치는 잠재적 특성으로 본다.

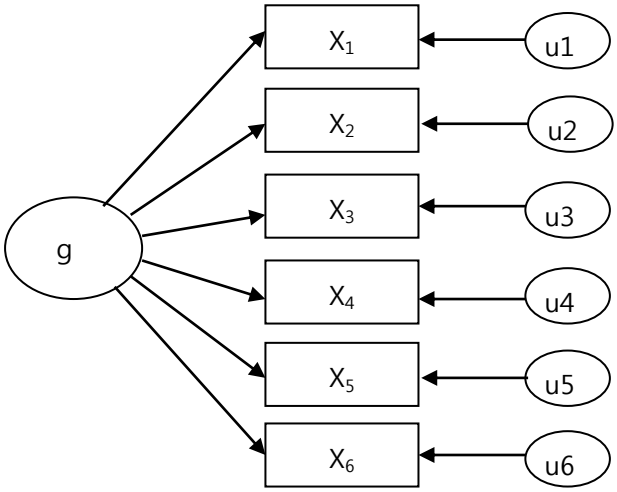
✓ Thurstone's multiple-factor theory 혹은 common-factor theory

Thurstone의 이론은 Spearman의 g-factor 대신 복수의 요인들이 존재한다고 본다. 즉 모든 행동 영역에 영향을 미치는 잠재요인(g-factor)이란 존재하지 않으며 몇 개로 나위어진 집단요인들(group factors)이 각 문항에 관련된 unique factor와 함께 해당 문항들에 대한 수행을 설명한다는 것이다.

❖ 오늘날, 우리가 (탐색적) 요인분석이라고 부르는 것은 바로 이 Thurstone's common-factor theory에 기반한 것이다. 이 경우, 물론, group factor의 수가 하나이면 Spearman의 모형이 된다. 따라서, Spearman의 모형은 Thurstone 모형의 특수한 형태라고 볼 수 있다.

✓ H*o*lzinger's bi-factor theory

이 이론은 Spearman의 이론을 약간 변형한 것으로서, 관찰변수 혹은 문항의 수만큼 존재하는 unique factor들보다는 이들을 몇 개의 그룹으로 묶고, 이러한 집단요인들(group factors)과 g-요인이 함께 피험자의 문항 수행을 결정한다고 본다.

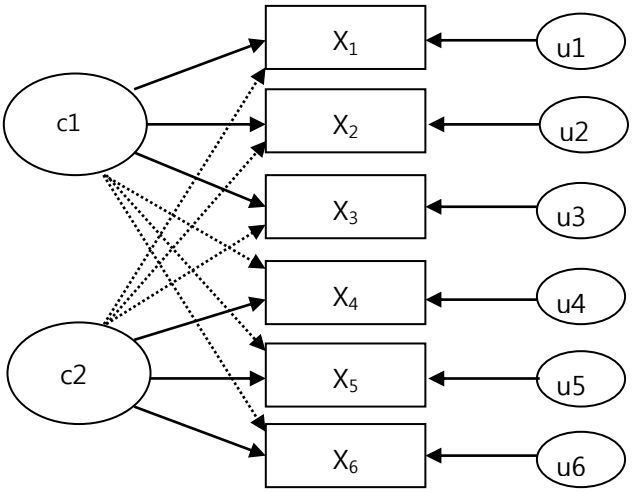


Spearman

two-factor theory

혹은

g-factor theory

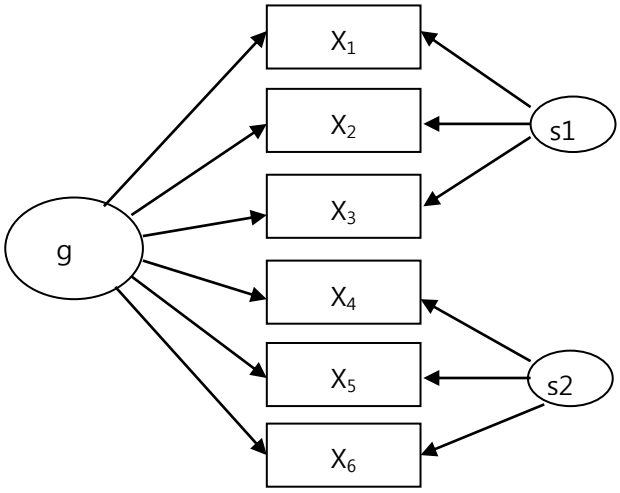


Thurstone

multiple-factor theory

혹은

Common-factor theory



Holzinger

bi-factor theory

➢ 탐색적 요인분석과 확인적 요인분석

✓ 연구자가 주어진 자료의 구조에 관련된 이론적 가설을 가지고 있는 경우가 있다. 그러나, 이러한 사전적 가설이 없으며 연구의 목적이 그와 같은 구조를 탐구해 보는 데에 있다면, 우리는 "탐색적 요인분석" (Exploratory factor analysis)을 실시하는 것이다. 모든 요인이 모든 관찰변수에 계수를 갖도록 하는 공통요인모형은 바로 탐색적 요인분석을 실시하기 위한 방법이다.

✓ 자료의 구조에 대한 이론적 가설을 가지고 있을 때, 실시할 수 있는 것이 "확인적 요인분석" (Confirmatory factor analysis)이다. 일반적으로 확인적 요인분석을 실시할 때에는 요인부하량 행렬의 한 요소가 연구자에 의해서 0으로 결정된 채 분석이 시작된다. 이 0은 해당 요인이 해당 관찰변수와 아무런 관련이 없다는 이론적 가설이 있기 때문에 설정되는 것이다. 즉 확인적 요인분석의 목적은 이미 가지고 있는 이론을 주어진 자료를 통해 확인하는 데에 있다.

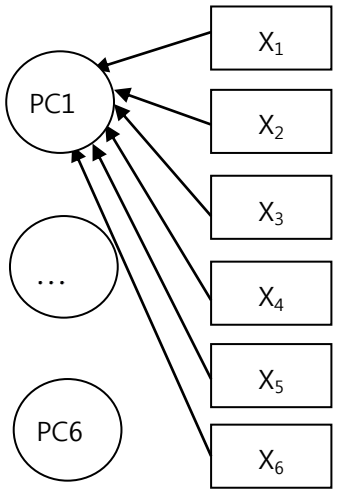
✓ 탐색적 요인분석은 SAS나 SPSS 등의 프로그램으로 실시하는 경우가 대부분이며, 확인적 요인분석은 LISREL이나 AMOS 등의 프로그램을 이용해서 실시할 수 있다.

➢ 주성분분석과 요인분석

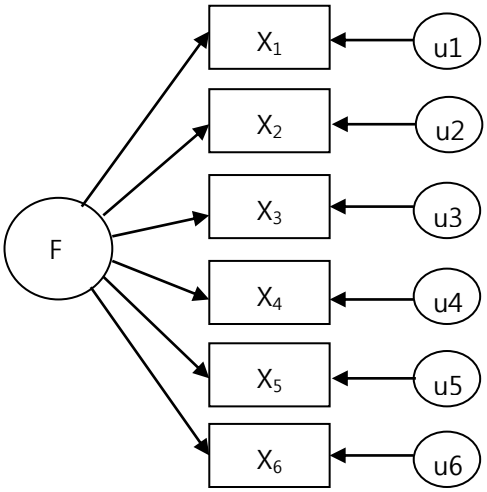
✓ 주성분분석에서는 주성분이 관찰변수의 선형 결합으로 표현되지만, 요인분석에서는 잠재변수(요인)의 선형결합으로 각 관찰변수가 구성된다고 본다.

✓ 요인분석의 경우, 주성분분석과 달리, 관찰변수의 측정에 있어서 오차가 존재할 수 있음을 인정한다. 다시 말해서, 주성분분석 하에서 모든 관찰변수 분산이 설명되지만 요인분석 하에서는 잠재변수들에 의해서 설명될 수 있는 분산이 주된 관심의 대상이다. 각 관찰변수 분산을 공통성(communality)과 독특성(uniqueness)으로 나누어 생각한다.

➔ 요인분석은 상관행렬을 주재료로 하여 적용된다. 그런데 상관행렬이 그대로 사용되는 것이 아니라, 이른바 축소상관행렬(reduced correlation matrix)이 사용된다. 이는 보통의 상관행렬에서처럼 대각원소가 1이 되는 것이 아니라 공통성(communalities)이 된다. 이는 각 관찰변수 측정에 있어서 모형에 포함된 요인들(factors; 잠재변수)로 설명되는 않는 unique 한 부분이 있다는 것을 고려하기 위함이다. 이 부분은 또한 측정의 오차와 혼합되어(confounded) 있다.

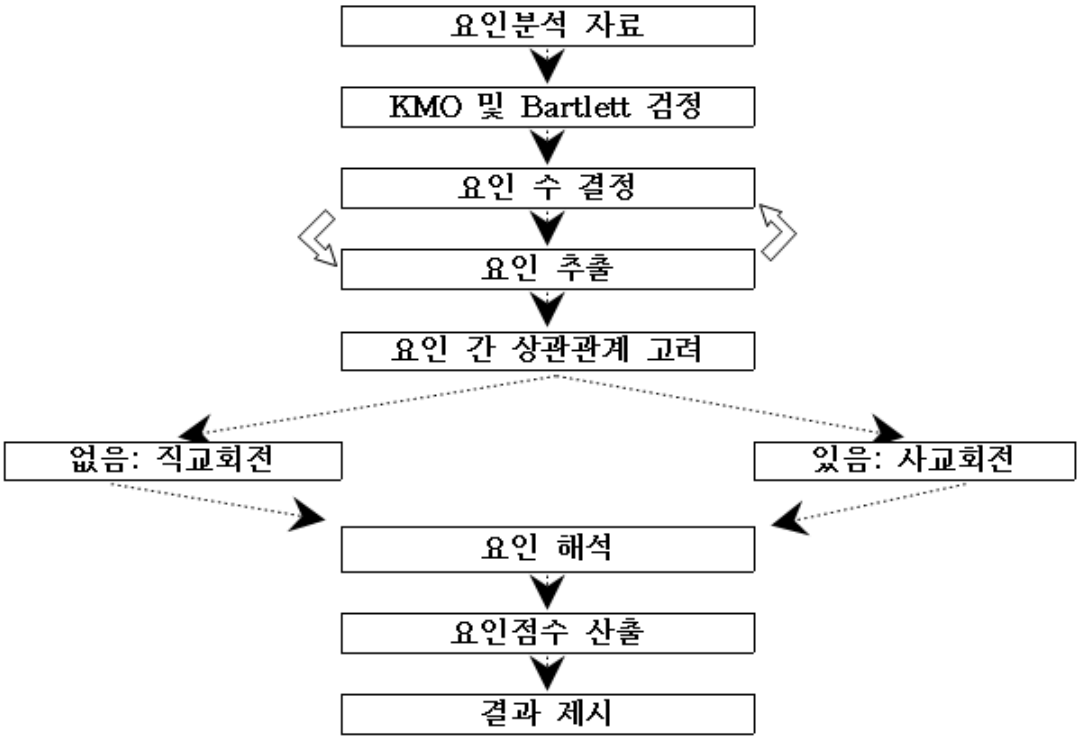


주성분분석



요인분석

➢ 탐색적 요인분석 실행 절차



➢ 탐색적 요인분석 사용 현황 연구 (강태훈 등, 2013년 현재 연구 진행 중)

✓ 연구대상: 한국연구재단 (NRF : National Research Foundation of Korea)의 등재지 또는 등재후보지로 등재된 학회 중 탐색적 요인분석에 관련되어 게재된 학회지 가운데 제목과 주제어 목록에 '탐색적 요인분석'이 포함된 연구를 대상으로 2005년부터 2011년 최근 7년까지 국내학회지 중에서 연구에 적합하다고 판단되는 연구 논문 259편을 분석함

전공분야	EFA 사용 논문
교육심리 및 상담	112편 (43.2)
청소년교육	64편 (24.7)
평생교육 및 성인교육	34편 (13.1)
유아교육	28편 (10.8)
사회복지 및 장애인교육	10편 (3.9)
기타	11편 (4.2)
합계	259편 (100)

✓ 상관행렬 계산을 위한 피험자 수

피험자수	50 이하	100 이하	100 -150	151 -200	201 -250	251 -300	301 -350	351 -400	401 이상
빈도	1	7	15	18	42	29	20	12	115
%	.4	2.7	5.8	6.9	16.2	11.2	7.7	4.6	44.4

✓ 구형성 검증 및 요인의 수 결정 과정에 대한 언급 여부 (%)

	KMO 및 Bartlett 검증 결과 제시		요인의 수 결정 과정 제시	
	있음	없음	있음	없음
빈도(%)	108(41.7)	151(58.3)	211(81.47)	48(18.53)

✓ 요인 간의 상관관계에 대한 제시 방법

	상관관계		
	있다.		언급된사항 없다.
빈도 (%)	237 (91.5%)		22 (8.5%)
	제시된 형식	**빈도 (%)**	
	이론적배경	20 (7.7%)	
	상관계수	210 (81.1%)	
	이론적배경과 상관계수	7 (2.7%)	
	합계	**237 (91.5%)**	**22 (8.5%)**
합 계		259 (100.0%)	

✓ 요인 축의 회전 방법

	회전방식				
	직교회전		사교회전		언급없음
빈도 (%)	147 (56.8%)		77 (29.7%)		35 (13.5%)
	유형	**빈도 (%)**	**유형**	**빈도 (%)**	
	베리멕스	141 (54.4%)	프로멕스	18 (6.9%)	
	기타	1 (0.4%)	디렉트오블리민	41 (15.8.%)	
	언급없음	5 (1.9%)	기타	3 (1.2%)	
			언급없음	15 (5.8%)	
	합계	147 (56.8%)		77 (29.7%)	35 (13.5%)
합 계			279 (100.0%)		

✓ 요인 간 상관관계와 회전 방법 적용

회전방식

		직교회전	사교회전	
요인 간 상관관계	있다	133(59.3%)	73(32.5%)	206(91.6%)
	없다	14(6.2%)	4(1.7%)	18(8.0%)
		147(65.6%)	77(34.3%)	224(100.0%)

✓ 탐색적 요인분석 결과 도출을 위한 요인추출방법

	ML	주성분분석	주축분해법	기타	언급사항없음
빈도 (%)	31 (12.0%)	118 (45.6%)	59 (22.8%)	10 (3.9%)	41 15.8)

✓ 요인추출방법과 회전방식 (%)

		회전방식			
		직교회전	사교회전	언급 없음	
요인추출방법	주성분분석	101	14	3	118(45.6)
	최대우도법	1	28	2	31(12.0)
	주축분해법	21	31	7	59(22.8)
	기타	5	4	1	10(3.9)
	언급 없음	19	0	22	41(15.8)
		147(56.8)	77(29.7)	35(13.5)	259(100.0)

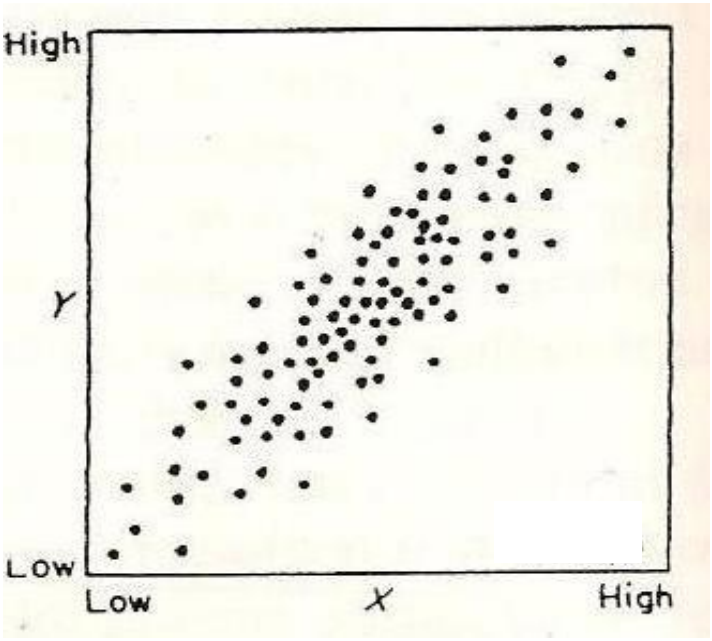
➢ 공분산과 상관계수

✓ 상관계수(correlation coefficient): 의미와 공식

- 공분산은 두 변수가 함께 변화하는 정도라고 정의할 수 있다. 상관도 역시 두 변수가 함께 변화하는 정도를 의미한다.
- 다른 점은 표준화된 공분산이라는 점이다. 즉 각각의 변수를 표준점수로 바꾼 뒤에 두 변수간 공분산을 구하면 이는 상관계수와 같게 된다.

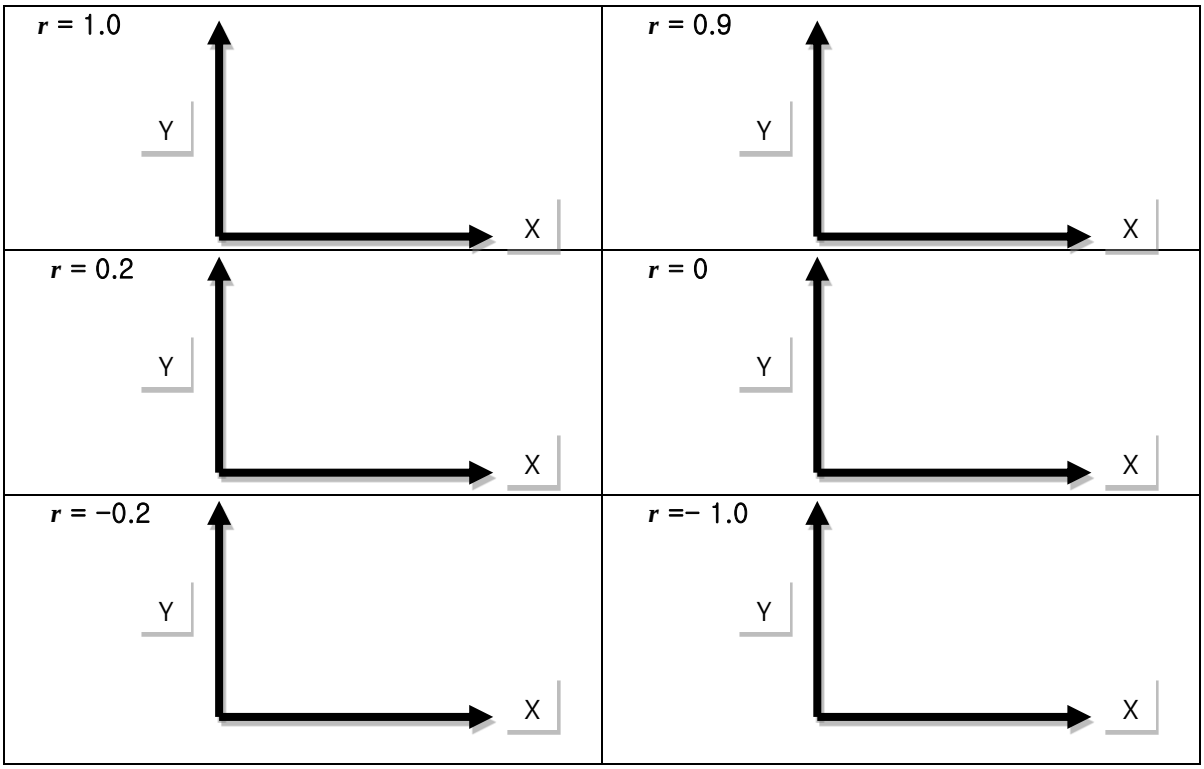
$$\rho_{XY} = \mathrm{cov}(Z_X, Z_Y) = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}(X)\,\mathrm{var}(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

- 산포도 (scatter plot): 두 변수간의 관계를 알아보기 위하여 그리는 도표이다.

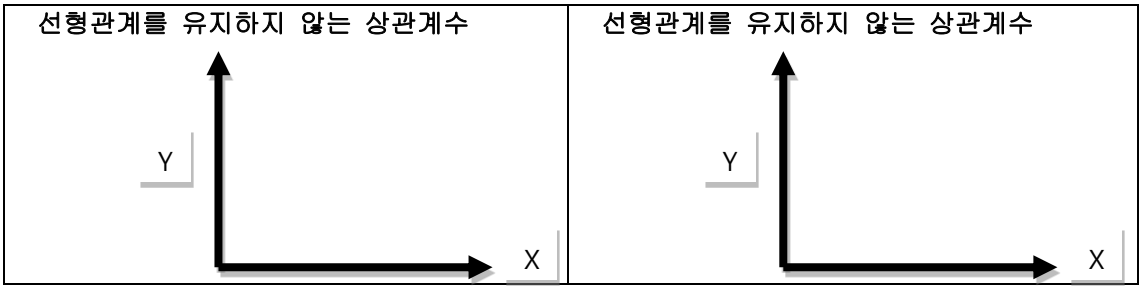


(a) Math ability(X) and IQ(Y)

- 산포도 연습: 상관계수($-1 \le r \le 1$)와 그에 따른 산포도



✓ Pearson 의 적률상관계수 (product moment correlation coefficient): 엄밀히 말해 두 변수가 **선형적**으로 함께 변화하는 정도를 의미한다.



✓ 두 변수간 "통계적 혹은 확률적 독립"과 "상관계수=0"의 의미는?

✓ 모수치에 의한 상관계수는 ρ(rho)로 표현하고, 통계치에 의한 표현은 보통 r로 표현한다.

- 상관계수의 특징
  - ✓ 지역무관성(location-free): 일정한 수를 한 변수에 더하거나 빼도 상관계수는 변하지 않는다.
  - ✓ 척도독립성(scale-free): 일정한 수를 한 변수에 곱하거나 나누어도 상관계수는 변하지 않는다.
    - ◆ 확인: EXCEL의 COVAR와 CORREL 함수를 각각 이용하여 각 경우의 공분산과 상관계수를 구해 보세요.

	원점수		변환점수 1		변환점수 2	
	X	Y	X+2	Y-1	10 X	5 Y
	2	1	4	0	20	5
	3	2	5	1	30	10
	5	4	7	3	50	20
	3	3	5	2	30	15
	7	5	9	4	70	25
공분산						
상관계수						

✓ 상관계수의 해석 (언어적 표현)

- 절대적이지는 않지만 대략 아래와 같은 기준에 따라 상관관계를 표현한다.

상관계수 범위	상관관계의 언어적 표현
.00 – .20	상관이 매우 낮다
.20 – .40	상관이 낮다
.40 – .60	상관이 있다
.60 – .80	상관이 높다
.80 – 1.00	상관이 매우 높다

- 또한 SPSS를 이용하여 하나의 상관계수가 통계적으로 유의미한지에 대한 검정도 가능하다. 이 때의 영가설은 “두 변수간 상관계수는 0이다” 이므로 이를 기각하게 되면 (p-value가 0.05보다 작으면) 두 변수간에 통계적으로 유의미한 상관이 있다는 의미이다.

✧ 연습문제: 학생들의 주당 학습시간과 학점과의 상관이 있는지를 알아보기 위하여 다음과 같은 자료를 수집하였다. SPSS를 사용하여 다음과 같은 분석을 실시하자.

학생	주당 학습시간	학점
1	27	3.6
2	6	2.2
3	16	3.1
4	21	2.95
5	25	3.8
6	10	2.93
7	6	2.7
8	12	2.0
9	6	3.2
10	7	2.8
11	11	2.99
12	8	2.15
13	20	3.5
14	10	3.2
15	13	3.18

- 산포도 그리기
- 각 변수의 평균과 표준편차 구하기
- 두 변수간 공분산과 상관계수 구하기
- 두 변수의 관계에 대하여 해석하기

✓ 상관관계 행렬과 탐색적 요인분석

- 여러 관찰변수들 간의 상관계수를 구했을 때 변수들 간의 상호 관계를 이러한 정보에 기초하여 요약하고자 한다. 이 때, 관찰변수의 수가 많다면 효과적으로 정보를 요약하는 것이 쉽지 않을 것이다 → 탐색적 요인분석 필요

	Classic	French	English	Math	Discover	Music
Classic	1	.83	.78	.7	.66	.63
French		1	.67	.67	.65	.57
English			1	.64	.54	.51
Math				1	.45	.51
Discover					1	.4
Music						1

2. 주성분분석

➢ 목적 및 활용

- ✓ 서로 상관관계가 있는 변수들 사이의 복잡한 관계를 좀더 간편하고 이해하기 쉽게 설명하기 위해 사용하는 분석기법(독립변수와 종속변수의 구분 없음)이다.
- ✓ 다시 말해서, 분석 대상 변수들이 많은 경우 변수들 사이의 상관 구조가 복잡(예를 들어, p=20인 경우 unique한 상관계수는 p(p-1)/2개 만큼 존재)하기 때문에 이를 체계적으로 요약하는 기법이다.
- ✓ 변수를 요약하여 그 수를 줄이면
  - 자료를 해석하기 용이
  - 중다회귀분석의 독립변수 자료로 이용시 다중공선성 문제 해결 (각 주성분에 대한 해석은 물론 연구자의 몫이다)

➢ 수학적 모형

- ✓ 주어진 자료(X)의 분산-공분산 행렬(Σ_X)에 이 행렬의 고유벡터로 이루어진 행렬(변환 행렬; transformation matrix)을 앞뒤로 곱하여 상호 상관이 없는 변수들(Y: 주성분, principal component or PC) 간의 분산-공분산 행렬($\Sigma_Y = \Lambda$)로 변환: 스펙트럼 분해 참고

$$V'\Sigma_X V = \Lambda \;\rightarrow\; \Sigma_X = V\Lambda V'$$

- ✓ $X_1, X_2, \ldots, X_p$ 변수 집합이 정칙행렬인 Σ_X 를 갖는다면, 언제나 상호 상관이 없는 $Y_1, Y_2, \ldots, Y_p$ 변수 집합으로 유도할 수 있다.

$$Y' = X'V$$

$$(nXp)=(nXp)(pXp)$$

- ✓ $Y = V'X$ $(pXn)=(pXp)\,(pXn)$: 실제 자료 형태 (피험자 j의 주성분 값)

$$y_{1j} = a_{11}x_{1j} + a_{21}x_{2j} + a_{31}x_{3j} + \cdots + a_{p1}x_{pj}$$
$$y_{2j} = a_{12}x_{1j} + a_{22}x_{2j} + a_{32}x_{3j} + \cdots + a_{p2}x_{pj}$$
$$y_{3j} = a_{13}x_{1j} + a_{23}x_{2j} + a_{33}x_{3j} + \cdots + a_{p3}x_{pj}$$
$$\ldots$$
$$y_{pj} = a_{1p}x_{1j} + a_{2p}x_{2j} + a_{3p}x_{3j} + \cdots + a_{pp}x_{pj}$$

- 주성분분석 결과를 제시할 때는 주어진 자료를 가장 많이 설명하는 주성분부터 순서대로 (즉 고유값의 크기에 따라) 언급한다.
- 즉, 변수들의 선형 결합을 통해 변수들이 가진 전체 정보를 최대한 설명하는 인공변수들(Y or PC)을 유도 (이 때 위에서 확인하였듯이 분산의 합은 변하지 않음)
- 위에서 사용된 선형결합의 계수 a는 가중치(weight)라고 부른다. 즉 정규화된 고유벡터 (즉 자신과 내적했을 때 그 값이 1)의 각 원소가 이들 가중치가 된다.
- 위에서 관찰변수들(X)을 선형결합하여 주성분을 만들 때 오차항이 없음: 주성분 분석에서는 X 변수들의 측정에 오차가 없다고 가정함에 주의 (따라서, 주로 정확한 측정이 가능한 자연과학 쪽에서 사용됨)
- 요인분석에서는 잠재변수(요인; factors)들의 선형결합(+error)을 통해 각 관찰변수를 표현하며 이 때 사용되는 계수가 부하량(loading)이다.

✓ 예를 들어,

$\Sigma_X = \begin{bmatrix} 4 & -7 & 8 \\ -7 & 13 & -17 \\ 8 & -17 & 28 \end{bmatrix}$ 이면, eigenvalues를 구하려면 $|\Sigma_X - \lambda I| = 0$

이를 풀면 $\lambda = 42, 3, 0$ 이므로 해당 eigenvectors를 구하면 아래와 같다.

(참고: 원자료 분산의 합 = $tr(\Sigma_X) = \sum \lambda = 45$)

$V = \begin{bmatrix} .26726 & -.57735 \\ -.53452 & .57735 \\ .80178 & .57735 \end{bmatrix}$. 그러므로,

$V'\Sigma_X V = \begin{bmatrix} 42 & 0 \\ 0 & 3 \end{bmatrix} = \Lambda$

```
% MATLAB

SigX = [4 -7 8; -7 13 -17; 8 -17 28];
[V, D] = eig(SigX)
```

➢ 상관행렬을 이용한 주성분분석

✓ 변수들의 분산은 측정 단위에 따라서 크게 달라질 수 있다. 예를 들어, 같은 돈이라도 원화와 dollar 로 사용될 때 각각 그 분산 값이 다르게 된다. 이 때 변수들의 측정 단위가 연구자의 해석에 도움이 되는 경우는 관계없지만, 오히려 방해하거나 도움이 되지 않는 경우 미리 모든 변수를 평균 0, 표준편차 1 로 표준화시킨 다음에 분석을 하는 것이 바람직하다.

✓ 표준화변수의 분산-공분산 행렬은 상관행렬이므로, 이 경우 상관행렬을 이용한 주성분분석이 된다.

✓ 표준화변수를 활용한 주성분분석의 특징은 다음과 같다.
- 주성분들의 분산의 합은 p이다.
- 주성분 y_i가 설명할 수 있는 비율은 해당 고유값을 p로 나눈 것이다.
- 처음 k개 주성분들이 설명할 수 있는 누적비율은 $\sum_{i=1}^{k} \frac{\lambda_i}{p}$ 이다.
- 주성분분석에서는 자료 변수들 분산의 대부분을 설명할 수 있는 소수의 주성분을 선택하게 된다.

➢ 주성분분석의 유용성

✓ 앞에서 언급된 바와 같이, 주어진 자료 속에 많은 수의 관찰변수가 존재할 때 전체 분산의 상당 부분을 설명하는 적은 수의 변수(주성분)들로 요약 설명하고자 할 때 → 관찰변수들간 공분산 혹은 상관이 존재할 때 이들 간에 중복적으로 설명하는 요소가 있기 때문에 이러한 redundancy 로부터 자유로운 parsimony 추구!

✓ 보다 이론적인 측면에서, 관찰 대상들을 이해할 수 있는 주된 방법을 찾으려는 의도를 충족하기 위해서 → 한 가지 구인(construct)를 재고 있는 여러 관찰변수들이 과연 각각 어느 정도의 비중(혹은 가중치)으로 그 구인을 재고 있는 지를 파악할 수 있다. 예를 들어, 국어, 고전문학, 불어, 영어 등에 대한 능력을 재는 여러 척도들이 있을 때 이들 대부분이 비슷한 구인을 재고 있다면 이들 척도들은 하나의 주성분으로 설명될 수 있을 것이다.

☆ 연습문제: 다음은 220명 학생으로부터 얻은 Everitt(1984)의 자료이다.

X_1: French Score
X_2: English Score
X_3: History Score
X_4: Arithmetic Score
X_5: Algebra Score
X_6: Geometry Score

이 자료로부터 구한 상관행렬은 다음과 같다.

$$R = \begin{bmatrix} 1.0 & & & & & \\ .44 & 1.0 & & & & \\ .41 & .35 & 1.0 & & & \\ .29 & .35 & .16 & 1.0 & & \\ .33 & .32 & .19 & .59 & 1.0 & \\ .25 & .33 & .18 & .47 & .46 & 1.0 \end{bmatrix}$$

주성분분석을 MATLAB 을 사용하여 실시하고 그 결과를 해석해 보자.

```
% MATLAB
clear all

R = [1.0 .44 .41 .29 .33 .25;
     .44  1.0 .35 .35 .32 .33;
     .41 .35  1.0 .16 .19 .18;
     .29 .35 .16  1.0 .59 .47;
     .33 .32 .19 .59  1.0 .46;
     .25 .33 .18 .47 .46  1.0];
[V, D] = eig(R)

------------------------------------------------------------------
V =

   -0.1384   -0.6288   -0.4554    0.2073   -0.4184    0.4000
    0.1601    0.3832    0.3473    0.6774   -0.2727    0.4168
   -0.0303    0.2825    0.1453   -0.6619   -0.6020    0.3126
   -0.6926    0.3361   -0.2334   -0.0227    0.3925    0.4453
    0.6889    0.1300   -0.3937   -0.1695    0.3508    0.4491
   -0.0069   -0.4981    0.6643   -0.1757    0.3333    0.4105

D =

    0.4019         0         0         0         0         0
         0    0.5225         0         0         0         0
         0         0    0.6028         0         0         0
         0         0         0    0.6153         0         0
         0         0         0         0    1.1288         0
         0         0         0         0         0    2.7287
```

1st PC= $0.4000Z_1 + 0.4168Z_2 + 0.3126Z_3 + 0.4453Z_4 + 0.4491Z_5 + 0.4105Z_6$

2nd PC= $-0.4184Z_1 - 0.2727Z_2 - 0.6020Z_3 + 0.3925Z_4 + 0.3508Z_5 + 0.3333Z_6$

3rd PC= $0.2073Z_1 + 0.6774Z_2 - 0.6619Z_3 - 0.0227Z_4 - 0.1695Z_5 - 0.1757Z_6$

4th PC= $-0.4554Z_1 + 0.3473Z_2 + 0.1453Z_3 - 0.2334Z_4 - 0.3937Z_5 + 0.6643Z_6$

5th PC= $-0.6288Z_1 + 0.3832Z_2 + 0.2825Z_3 + 0.3361Z_4 + 0.1300Z_5 - 0.4981Z_6$

6th PC= $-0.1384Z_1 + 0.1601Z_2 - 0.0303Z_3 - 0.6926Z_4 + 0.6889Z_5 - 0.0069Z_6$

➔ 위와 같이 이론적으로는 6개의 관찰변수가 있으면 6개의 주성분을 구할 수 있다. 그러나, 적절한 양이 설명되었을 때 멈추는 것이 바람직하기 때문에 주성분의 수는 관찰변수의 수보다 적게 결정된다.

➢ 주성분 개수의 결정

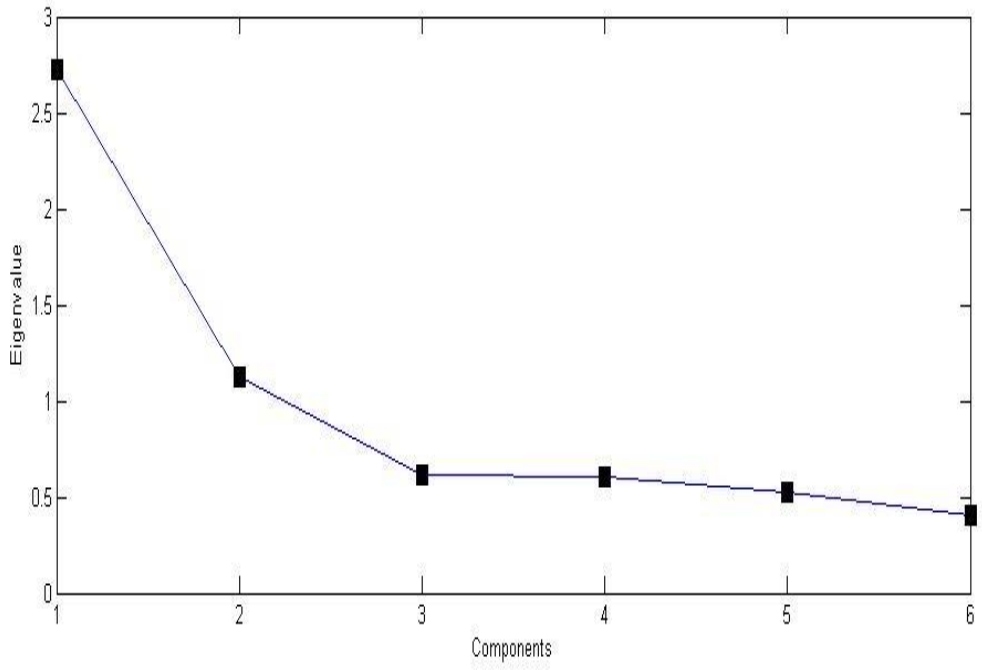
✓ 관찰자료를 적절히 요약하기 위한 주성분의 개수를 결정해야 하며, 이러한 주성분의 수는 절대 관찰변수의 수보다 클 수 없다.

✓ 각 주성분의 분산은 eigenvalue이기 때문에 각 주성분의 상대적 중요도(즉 전체 분산 중에서 설명하는 비율)를 결정하는 데에 이를 유용하게 사용할 수 있다.

✓ 가능한 방법

- Scree Plot
- Kaiser 판단: Only keep PCs having eigenvalues greater than 1 (when analyzing a correlation matrix)
- Set a variance-accounted for threshold (e.g., extract PCs until 80% of the variance is accounted for)

```
EV = [2.7287 1.1288  0.6153  0.6028 0.5225 0.4019];
        PC = [1 2 3 4 5 6];
        plot(PC, EV,'-bs','LineWidth',1,...
              'MarkerEdgeColor','k',...
              'MarkerFaceColor','k',...
              'MarkerSize',10)
        xlabel('Components')
        ylabel('Eigenvalue')
```



- 주성분의 해석
  - ✓ 각 관찰변수가 하나의 주성분을 정의하는 데에 있어서 기여하는 정도를 살펴봄으로써 주성분을 해석할 수 있다. 다시 말해서, (음의 값이든 양의 값이든) 상대적으로 큰 가중치(a)를 가지는 관찰변수가 해당 주성분의 분산을 많이 설명한다고 볼 수 있다.
  - ✓ 해석을 보다 용이하기 위해서 다음과 같이 주성분과 관찰변수간 상관계수를 구할 수 있다.

$$r(PC_i, X_j) = a_{ij}\sqrt{\lambda_i}$$

예를 들어, 앞의 연습문제에서 1st PC와 첫번째 관찰변수와의 상관계수:

$$r(PC_1, X_1) = a_{11}\sqrt{\lambda_1} = 0.4 \times \sqrt{2.7287} = 0.6608$$

- 여기서 고유값(λ)은 관찰변수들의 상관행렬을 스펙트럼 분해하여 얻은 것이며, 이러한 주성분과 관찰변수간 상관을 해당 주성분의 '성분 부하량(component loading)이라고 부른다.

  - ✓ SPSS 를 사용할 때 결과로 제공되는 성분행렬(component matrix)는 이러한 부하량을 담고 있다. 따라서 SPSS 를 통해서 실제로 주성분분석 결과로서의 가중치를 계산하려면 각 부하량을 해당 주성분의 고유값 제곱근으로 나누어주어야 한다.

$$a_{ij} = \frac{r(PC_i, X_j)}{\sqrt{\lambda_i}}$$

  - ✓ 하나의 주성분에 대하여 그 관련 가중치를 제곱하여 모두 더하면 1 이 되며, 관련 부하량을 제곱하여 모두 더하면 그 주성분의 고유값이 된다. 예를 들어, 앞의 Everitt(1984) 자료에 대한 분석 결과를 보면

$$\sum_{j=1}^{p} a_{ij}^2 = 0.4000^2 + 0.4168^2 + 0.3126^2 + 0.4453^2 + 0.4491^2 + 0.4105^2 = 1$$

$$\sum_{j=1}^{p} r_{1j}^2 = \sum_{j=1}^{p} a_{1j}^2\lambda_1 = \lambda_1\sum_{j=1}^{p} a_{1j}^2 = \lambda_1 = 2.7287$$

- 주성분 점수 (Principal Component Scores)

  - 각 개인 피험자의 주성분 점수는 모든 관찰변수의 표준화 값과 함께 가중치(고유벡터)를 구하면 쉽게 계산할 수 있다.
  - 앞의 연습문제처럼 상관행렬을 자료로 하여 주성분분석을 실시할 경우, 개인 피험자의 주성분 점수를 계산하는 것은 불가능하다.
  - 원자료를 사용하여 주성분분석을 하고 주성분분석을 실시하여 모든 개인의 주성분 점수를 파악하고 나면, 이를 이용하여 회귀분석 등의 다른 통계적 분석에 이용할 수 있다.

➢ 원자료를 사용하여 주성분분석 실시하기 (SPSS 사용)

✓ 다음은 15개 병원으로부터 얻은 자료이다.

X_1: 하루 평균 환자수

X_2: 월 평균 X-선 검사 수

X_3: 월 평균 입원환자 수

X_4: 지역인구 (1,000명)

병원	X_1	X_2	X_3	X_4
1	44.02	2048	9.5	696.82
2	20.42	3940	12.8	1033.15
3	18.74	6505	36.7	1603.62
4	49.20	5723	35.7	1611.37
5	44.92	11520	24.0	1613.27
6	55.48	5779	43.3	1854.17
7	59.28	5969	46.7	2160.55
8	94.39	8641	78.7	2305.58
9	128.02	20106	180.5	3503.93
10	96.00	13313	60.9	3571.89
11	131.42	10771	103.7	3741.40
12	127.21	15543	126.8	4026.52
13	252.90	36194	157.7	10343.81
14	409.20	34703	169.4	11732.17
15	463.70	39204	331.4	15414.94

✓ MATLAB을 이용하여 고유값과 고유벡터 구하기

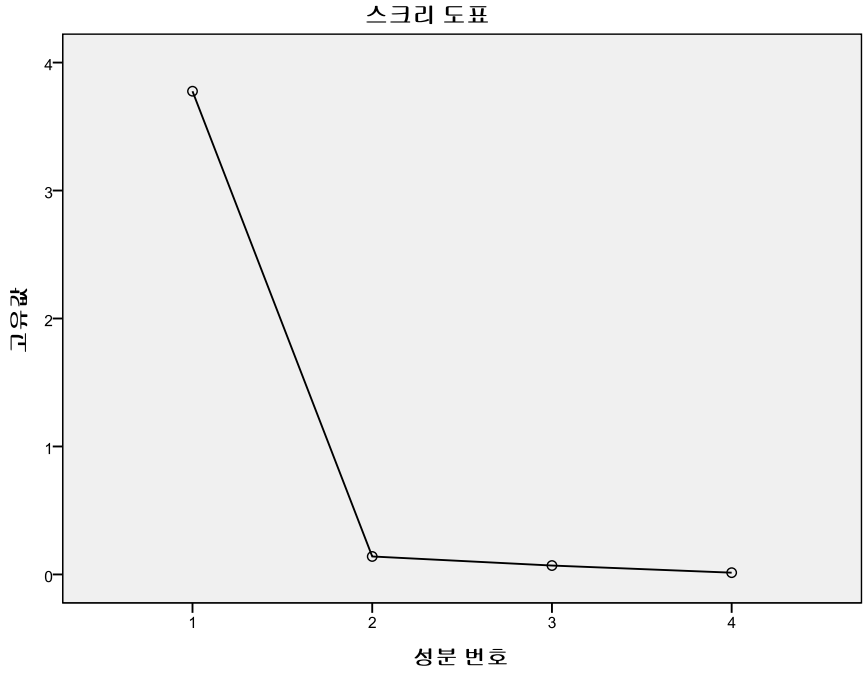
✓ 독립변수들에 대한 상관행렬을 구하고 해석하여라

✓ SPSS 메뉴를 사용하여 주성분 분석을 실시하기

✓ 주성분분석을 실시할 때 선택되는 주성분의 개수는? 그 설명량은?

✓ SPSS syntax를 사용하여 주성분 분석을 실시하기 (공분산행렬 사용?)

```
% MATLAB

load hospital.txt
R=corr(hospital)
[V, D] = eig(R)
-------------------------------------------------------------------------
R =

    1.0000    0.9330    0.8968    0.9814
    0.9330    1.0000    0.8840    0.9595
    0.8968    0.8840    1.0000    0.8945
    0.9814    0.9595    0.8945    1.0000


V =

    0.5766    0.5981    0.2341    0.5050
    0.2482   -0.7762    0.2926    0.5003
   -0.0418   -0.0288   -0.8723    0.4863
   -0.7773    0.1974    0.3140    0.5082


D =

    0.0137         0         0         0
         0    0.0699         0         0
         0         0    0.1408         0
         0         0         0    3.7756
```



```
MATRIX DATA VARIABLES = NumDP NumXr NumMP NumLoc
    /FORMAT = FREE LOWER
    /N=220
    /CONTENTS=CORR.

BEGIN DATA
1
.933 1
.897 .884 1
.981 .960 .895 1
END DATA.

FACTOR MATRIX = IN(CORR*)
  /PRINT ALL
  /CRITERIA FACTORS(4)
  /EXTRACTION PC
  /ROTATION NOROTATE
  /PLOT EIGEN.
```

공통성

	초기	추출
NumDP	1.000	1.000
NumXr	1.000	1.000
NumMP	1.000	1.000
NumLoc	1.000	1.000

추출 방법: 주성분 분석.

성분행렬[a]

	성분			
	1	2	3	4
NumDP	.981	−.087	−.159	.067
NumXr	.972	−.111	.204	.030
NumMP	.945	.327	.009	−.005
NumLoc	.988	−.117	−.051	−.092

요인추출 방법: 주성분 분석.

a. 추출된 4 성분

3. 탐색적 요인분석의 개념과 모형

➢ 수학적 모형

✓ Spearman's 일반요인모형 (g-factor model) 혹은 단일공통요인모형(single common-factor model)

$$X_j = \mu_j + \mathbf{l}_j F + e_j \qquad (j = 1, \cdots, p)$$
$$z_j = l_j F + e_j = l_j F + d_j U_j$$

✓ Thurstone's 공통요인모형

$$z_j = l_{j1}F_1 + l_{j2}F_2 + l_{j3}F_3 + \cdots + l_{jm}F_m + e_j$$
$$= l_{j1}F_1 + l_{j2}F_2 + l_{j3}F_3 + \cdots + l_{jm}F_m + d_j U_j \qquad (j = 1, \cdots, p)$$

X_j = 무선 관찰변수 j

z_j = 관찰변수 j의 표준화된 값

p = 관찰변수의 개수

m = 모형에서 사용된 공통요인(common factors)의 개수

F_k = 전체 m 개의 요인 중 k 번째 요인 $(k = 1, \cdots, m)$

U_j = 관찰변수 j에만 영향을 주는 unique factor

l_{jk} = 관찰변수 j의 요인 k 에 대한 요인 부하량

e_j = 무선 오차 항

d_j = 관찰변수 j의 unique factor에 대한 요인 부하량

"요인분석의 가정"	$E(z_j) = E(F_k) = E(e_j) = E(U_j) = 0,$ $\text{var}(z_j) = \text{var}(F_k) = \text{var}(U_j) = 1, \ \text{var}(e_j) = \psi_j = d_j^2$ $\text{cov}(F_k, F_{k'}) = 0$ 여기서 $k \neq k'$ → 요인간 상호직교 $\text{cov}(e_j, F_{k'}) = 0$ 공통요인과 무선 오차는 상관이 0 $\text{cov}(e_a, e_b) = 0$ 두 다른 관찰변수 관련 오차간 상관이 0

➢ 요인분석은 관찰변수(manifest variables)와 관련하여 다음 세 가지 행렬을 다룬다.

✓ 분산-공분산 행렬

S = 관찰변수들로 계산된 표본 분산-공분산 행렬

Σ_{XX} = 관찰변수들의 모집단 수준에서의 분산-공분산 행렬

$\hat{\Sigma}_{XX}$ = 요인분석 결과 추정된 값들로 계산된 추정된 혹은 적합된 (fitted) 분산-공분산 행렬

✓ 상관 행렬

R = 관찰변수들로 계산된 표본 상관행렬

Σ_{zz} = 관찰변수들의 모집단 수준에서의 상관 행렬

$\hat{\Sigma}_{zz}$ = 요인분석 결과 추정된 값들로 계산된 추정된 혹은 적합된 (fitted) 상관 행렬

✓ 각 개인이 가진 잠재변수 즉 요인 점수를 직접 관찰할 수 없기 때문에, 요인분석 모형을 적합하는 것은 단순회귀나 중다회귀분석의 상황과는 다르다. 공통요인(common factors)에 대한 직접적인 관찰 없이 요인분석을 가능하도록 하는 중요한 가정은, 관찰변수와 잠재변수간 선형적 관계를 갖는다는 것이다. 다음, 앞의 "요인분석의 가정"은 다음처럼 각 적합된 행렬의 원소가 표현될 수 있도록 필요하다!

· 단일공통요인모형

$$\begin{aligned}\text{var}(z_j) &= \text{var}(l_j F + e_j) = l_j{}^2 + \psi_j \\ \text{cov}(z_a, z_b) &= \text{cov}(l_a F + e_a, l_b F + e_b) \\ &= l_a l_b \,\text{cov}(F,F) + l_a \,\text{cov}(F,e_b) + l_b \,\text{cov}(F,e_a) + \text{cov}(e_a,e_b) \\ &= l_a l_b\end{aligned}$$

· 공통요인모형

$$\begin{aligned}\text{var}(z_j) &= \text{var}(l_{j1}F_1 + l_{j2}F_2 + l_{j3}F_3 + \cdots + l_{jm}F_m + e_j) = l_{j1}{}^2 + l_{j2}{}^2 + l_{j3}{}^2 + \cdots + l_{jm}{}^2 + \psi_j \\ \text{cov}(z_a, z_b) &= \text{cov}(l_{a1}F_1 + l_{a2}F_2 + \cdots + l_{am}F_m + e_a,\ \ l_{b1}F_1 + l_{b2}F_2 + \cdots + l_{bm}F_m + e_b) \\ &= l_{a1}l_{b1} + l_{a1}l_{b2} + \cdots l_{am}l_{bm}\end{aligned}$$

✓ 따라서, 단일공통요인모형이 주어진 자료를 잘 적합한다고 할 때 Σ_{zz}의 대각원소는 $l_j^{\,2}+\psi_j$가 되고 비대각 원소는 $l_a l_b$이 될 것이다.
다른 말로 하면, 이들 값을 추정하고 난 뒤 $\hat{l}_j^{\,2}+\hat{\psi}_j$과 $\hat{l}_a\hat{l}_b$을 계산하면 적합된 행렬 $\hat{\Sigma}_{zz}$을 얻을 수 있다.

✓ 모형의 적합도("goodness of fit")를 평가하려면, 적합된 상관행렬을 표본 상관행렬과 비교해 볼 수 있다.

Redidual Matrix = $R-\hat{\Sigma}_{zz}$ 혹은 $(=S-\hat{\Sigma}_{XX})$

➢ 공통성 (communality)

✓ 요인들간 상호 상관이 없다(orthogonality)는 가정 하에서 다음과 같은 유도가 가능하다.

$$\mathrm{var}(z_j)=l_{j1}^2+l_{j2}^2+l_{j3}^2+\cdots+l_{jm}^2+d_j^2=\mathbf{1}$$

여기서 common factors의 요인부하량 제곱을 모두 더한 것을 공통성 (communality; 아래에서 h_j^2)이라고 부른다.

$$h_j^2=l_{j1}^2+l_{j2}^2+l_{j3}^2+\cdots+l_{jm}^2 \;\rightarrow\; h_j^2+d_j^2=\mathbf{1}$$

✓ 다음 uniqueness에 의한 분산을 의미하는 d_j^2는 다시 해당 문항이 정말로 구체적으로 특수하게 재고 있는 부분(specificity; 아래에서 b_j^2)과 측정의 오차에 의한 분산(아래에서 e_j^2)으로 나누어 질 수 있다.

$$h_j^2+d_j^2=h_j^2+b_j^2+e_j^2=1$$

교육측정이론에서 한 검사 *A*의 측정의 표준오차 제곱(측정의 오차분산)은 "검사 점수의 분산X(1-신뢰도)"로 구한다: $\sigma_e^2 = \sigma_A^2(1 - r_{AA'})$
따라서, 표준화된 한 X 변수의 측정에 대한 오차 분산은 다음과 같이 표현될 수 있다.

$$e_j^2 = 1 - r_{jj'} \quad \rightarrow \quad r_{jj'} = 1 - e_j^2 \quad \rightarrow \quad h_j^2 = (1 - e_j^2) - b_j^2 = r_{jj'} - b_j^2$$

결론적으로, 한 변수의 communality가 가지는 상한계는 신뢰도이다.

$$h_j^2 \le r_{jj'} \qquad (\because \ b_j^2 \ge 0)$$

또한, 각각의 관찰변수에 대해서 다른 모든 변수들을 독립변수로 하여 중다회귀분석을 한 뒤 중다상관(R_j^2)을 구할 수 있을 것이다. 이 때, 공통성과의 관계는 다음과 같다.

$$R_j^2 \le h_j^2 \le r_{jj'}$$

❖ 위와 같은 지식 등은 축소상관행렬을 추정하기 위해 필요하다. 그러나, SPSS 등 통계 프로그램들은 축소상관행렬을 구하는 알고리듬이 포함되어 있으므로, 이를 연구자가 따로 제공할 필요는 없다. 즉 원자료나 상관행렬을 입력하면 된다. 단, SPSS의 옵션에서 요인추출을 위한 방법으로 주성분방법(principal component method)을 선택하면 주성분분석을 위한 방법이 된다 (물론, 주성분분석의 결과를 얻으려면 SPSS가 제공한 요인부하량 결과를 가중치로 변환해 주어야 한다: $l \rightarrow a$)

- 요인분석에서 요인부하량의 해석

  - 앞의 L을 요인 부하량 행렬 혹은 요인 유형 행렬(factor *pattern matrix*)이라고 부른다. 이 행렬의 각 원소는 해당 관찰변수가 요인으로부터 영향을 받는 정도를 나타낸다고 해석할 수 있다.

  - 요인간 상호 상관이 0 인 경우, 요인 부하량은 각 관찰점수와 요인간 상관계수를 나타낸다. 예를 들어, l_{jk}은 관찰변수 j 와 요인 k 간의 상관을 의미한다.

  - 앞에서 본 바와 같이 한 관찰변수와 관련된 요인부하량을 제곱하여 모두 더하면 이는 그 관찰변수의 공통성(communality; h_j^2)이며, 한 요인과 관련된 요인부하량을 제곱하여 모두 더하면 고유값(eigenvalue)가 된다.

  - 주성분분석과 달리 요인분석은 축소상관행렬을 원자료로 보기 때문에 이들 고유값은 $R-\Psi$을 스펙트럼 분해할 때 얻어지는 값이다.

❖ 요인분석을 이해하기 위한 또다른 접근: 요인분석을 통해서 위에서와 같이 잔차행렬의 원소를 0 에 가깝게 줄여나가는 시도 속에서 우리가 생각해 볼 수 있는 하나의 개념은 "조건적 선형 독립성의 원칙" (the principal of conditional linear independence)이다.

이 원칙을 이해하기 위해서는 부상관(partial correlation)에 대해서 생각해 볼 필요가 있다. 즉 두 변수 X_1과 X_2 간 상관을 구할 때 또다른 변수 X_3이 주어진 채 다시 말해서 X_3이 고정된 상수와 같은 상황 속에서 상관을 구하는 것이다. 이를 $\rho(X_1, X_2 \mid X_3)$이라고 표기한다.

요인분석이란 잠재변수 즉 공통요인들로 관찰변수의 분산을 대부분 설명하기 위한 시도이므로, 우리가 요인분석을 통해 이루고자 하는 것은 $\rho(X_a, X_b \mid F_1, F_2, \cdots, F_m) = 0$을 가능하면 적은 요인들을 사용하여 만족시키는 것이다.

☆ 연습문제: Everitt(1984)의 자료를 가지고 표본 상관행렬을 구한 뒤 단일공통요인모형으로 적합하였다. 추정된 모형 모수들은 다음과 같다.

$\hat{l}_1 = .50,\ \hat{l}_2 = .54,\ \hat{l}_3 = .35,\ \hat{l}_4 = .73,\ \hat{l}_5 = .73,\ \hat{l}_6 = .62$

$\hat{\psi}_1 = .75,\ \hat{\psi}_2 = .71,\ \hat{\psi}_3 = .87,\ \hat{\psi}_4 = .47,\ \hat{\psi}_5 = .47,\ \hat{\psi}_6 = .62$

✓ 잔차행렬을 구하면?

$$R = \begin{bmatrix} 1.0 & & & & & \\ .44 & 1.0 & & & & \\ .41 & .35 & 1.0 & & & \\ .29 & .35 & .16 & 1.0 & & \\ .33 & .32 & .19 & .59 & 1.0 & \\ .25 & .33 & .18 & .47 & .46 & 1.0 \end{bmatrix} \qquad \hat{\Sigma}_{zz} = \begin{bmatrix} & & & & & \\ .27 & & & & & \\ .18 & .19 & & & & \\ .36 & .39 & .25 & & & \\ .37 & .39 & .26 & .53 & & \\ .31 & .33 & .22 & .45 & .45 & \end{bmatrix}$$

$$R - \hat{\Sigma}_{zz} = \begin{bmatrix} & & & & & \\ .17 & & & & & \\ .24 & .16 & & & & \\ -.07 & -.04 & -.09 & & & \\ -.04 & -.07 & -.07 & .06 & & \\ -.06 & .00 & -.04 & .02 & .01 & \end{bmatrix}$$

✓ 같은 자료에 대해서 공통요인모형(두 개 요인 포함)을 사용하여 적합한 결과 추정된 모형 모수 추정치는 다음과 같다. 잔차행렬을 구하면?

$$\hat{L} = \begin{bmatrix} .56 & .43 \\ .57 & .29 \\ .39 & .45 \\ .74 & -.28 \\ .72 & -.21 \\ .60 & -.13 \end{bmatrix} \qquad diag(\hat{\Psi}) = \begin{bmatrix} .51 \\ .59 \\ .64 \\ .38 \\ .44 \\ .63 \end{bmatrix}$$

$$R - \hat{\Sigma}_{zz} = \begin{bmatrix} .001 & & & & \\ .000 & -.002 & & & \\ -.003 & .010 & -.004 & & \\ .018 & -.029 & .002 & .001 & \\ -.025 & .030 & .007 & -.006 & .005 \end{bmatrix}$$

✓ Everitt(1984)의 자료에 대한 공통요인모형(두 개 요인) 분석 결과를 통해 공통성과 고유값(eigenvalue)을 계산하면 다음과 같다.

$$\hat{L} = \begin{bmatrix} .56 & .43 \\ .57 & .29 \\ .39 & .45 \\ .74 & -.28 \\ .72 & -.21 \\ .60 & -.13 \end{bmatrix}$$

관찰변수 X_1의 공통성 = $.56^2 + .43^2 = .49$, uniqueness=.51
관찰변수 X_2의 공통성 = $.57^2 + .29^2 = .41$, uniqueness=.59
관찰변수 X_3의 공통성 = $.39^2 + .45^2 = .36$, uniqueness=.64
관찰변수 X_4의 공통성 = $.74^2 + (-.28)^2 = .62$, uniqueness=.38
관찰변수 X_5의 공통성 = $.72^2 + (-.21)^2 = .56$, uniqueness=.44
관찰변수 X_6의 공통성 = $.60^2 + (-.13)^2 = .37$, uniqueness=.63

공통성의 합= 약 2.8 특이성의 합= 약 3.2

Factor1의 고유값 = $.56^2 + .57^2 + .39^2 + .74^2 + .72^2 + .60^2 = 2.20$

Factor2의 고유값 = $.43^2 + .29^2 + .45^2 + (-.28)^2 + (-.21)^2 + (-.13)^2 = 0.60$

```
MATRIX DATA VARIABLES = French English History Arithmetic Algebra
Geometry
    /FORMAT = FREE LOWER
    /N=220
    /CONTENTS=CORR.

BEGIN DATA
1
.44 1
.41 .35 1
.29 .35 .16 1
.33 .32 .19 .59 1
.25 .33 .18 .47 .46 1
END DATA.

FACTOR MATRIX = IN(CORR*)
  /PRINT ALL
  /CRITERIA FACTORS(1)
  /EXTRACTION ML
  /ROTATION NOROTATE
  /PLOT EIGEN.

FACTOR MATRIX = IN(CORR*)
  /PRINT ALL
  /CRITERIA FACTORS(2)
  /EXTRACTION ML
  /ROTATION NOROTATE
  /PLOT EIGEN.
```

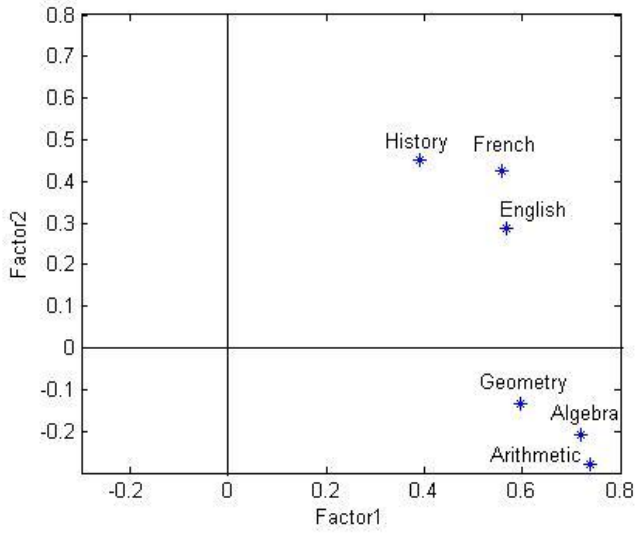
4. 요인분석: 직교회전과 사교회전

➢ 요인 축의 회전 (Rotation Methods)

✓ 탐색적 요인분석에서 요인의 구조를 탐구하기 위하여 흔히 고려되는 것이 요인 축의 회전이다. 하나의 요인 추출 방법에 의해서 추정된 요인부하량 행렬은 회전을 통해 똑같은 "적합된 상관행렬"을 만들어낼 수 있는 행렬로 변환될 수 있고, 이런 식으로 무한개 수의 요인부하량 행렬을 만들어 내는 것이 가능하다. 그 중 요인의 구조를 해석하는 데에 가장 용이한 것을 고르는 것이 요인 축 회전 방법(Rotation Method)의 핵심이다.

✓ Everitt(1984)의 자료

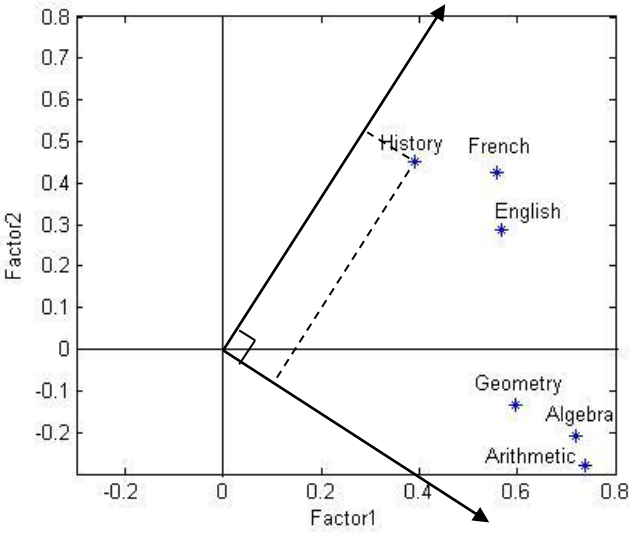
```
MATRIX DATA VARIABLES = French English History Arithmetic Algebra Geometry
    /FORMAT = FREE LOWER
    /N=220
    /CONTENTS=CORR.

BEGIN DATA
1
.44 1
.41 .35 1
.29 .35 .16 1
.33 .32 .19 .59 1
.25 .33 .18 .47 .46 1
END DATA.

FACTOR MATRIX = IN(CORR*)
  /PRINT ALL
  /CRITERIA FACTORS(2)
  /EXTRACTION ML
  /ROTATION NOROTATE
/PLOT EIGEN.
```



< Unrotated Loadings >

	Factor1	Factor2
French	.558	.425
English	.569	.286
History	.392	.450
Arithmetic	.738	-.279
Algebra	.718	-.209
Geometry	.595	-.133

✓ 모든 축이 상호 90도가 됨을 유지하면서 회전할 때 이를 **직교 회전** (orthogonal rotation)이라고 한다 → 회전 후에도 요인 간 상관은 0이다. Everitt 자료에 대한 분석 결과에서 보듯이 90도로 적절히 회전한다고 해도 좋은 해석이 가능한 요인부하량을 얻기 힘들 것으로 보인다.



✓ 각도에 상관없이 각 축이 최적의 해석을 찾아 90도가 아닌 각도로 회전할 때 이를 **사교 회전**(oblique rotation)이라고 한다 → 회전 후 요인 간 상관은 0이 아니게 된다.

➢ 요인 회전의 이유 및 논리

✓ 최대우도방법이나 주축요인방법 등의 요인추출법으로 최초로 구한 요인부하량은 때때로 해석하기 어려운 경우가 발생한다. 요인분석은 복잡한 자료를 요약하여 의미있는 해석이 가능한 요인부하량들을 유도하는 것을 주목적으로 한다.

✓ 이러한 요인 해석의 문제점을 해결하기 위하여 요인의 축을 회전시키는 방법을 사용한다. 즉 요인부하량 값들을 각 해당 요인을 나타내는 축 상의 좌표를 나타내는 값이라고 보고, 축을 원점을 기준으로 해서 회전했을 때 새로운 축 상에서의 좌표값을 구하는 것이다.

✓ 각 축을 어느 정도 회전하는가는 철저히 해석의 용이성 즉 보다 명확한 해석이 가능한 구조의 단순화를 지향하면서 이루어진다. 요인들 간의 단순구조(simple structure)가 의미하는 바는, 회전 후 새로 구한 요인부하량 행렬에서 각 관찰변수에 대응하는 요인부하량들이 대개 하나의 요인에는 큰 값들을 갖고 나머지 요인들에 대해서는 매우 작은 값을 갖게 되는 것을 말한다.

✓ Everitt(1984)의 자료 (PROMAX 방법을 이용한 사교 회전)

```
FACTOR MATRIX = IN(CORR*)
  /PRINT ALL
  /CRITERIA FACTORS(2)
  /EXTRACTION ML
  /ROTATION PROMAX
/PLOT EIGEN.
```

패턴 행렬[a]

	요인	
	1	2
French	.038	.680
English	.17	.524
History	−.112	.651
Arithmetic	.823	−.065
Algebra	.743	.009
Geometry	.579	.053

요인추출 방법: 최대 우도.
회전 방법: Kaiser 정규화가 있는 프로멕스.

<사교회전 후 **요인부하량** 해석을 위해서 사용하는 행렬>

구조행렬

	요인	
	1	2
French	.412	.700
English	.462	.620
History	.246	.589
Arithmetic	.787	.387
Algebra	.748	.418
Geometry	.608	.371

요인추출 방법: 최대 우도.
회전 방법: Kaiser 정규화가 있는 프로멕스.

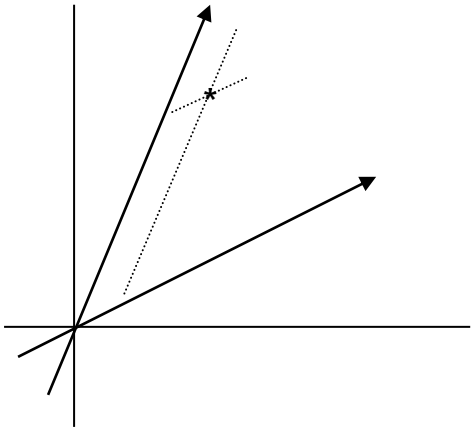
<사교회전 후 요인과 관찰변수 간 **상관계수**를 보여주는 행렬>

❖ 패턴 행렬에서 볼 수 있는 바와 같이 사교회전 후 요인부하량을 해석하여 두 요인과 관찰변수들 간의 관계를 요약 설명하기가 훨씬 수월해졌다.

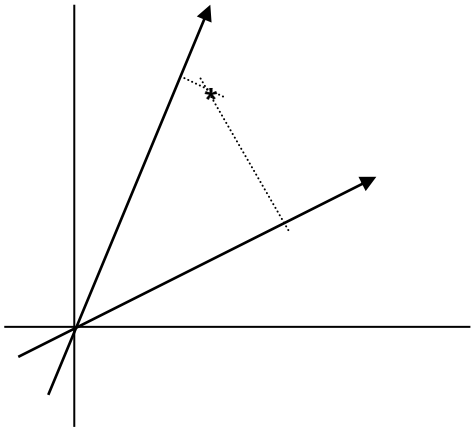
요인 상관행렬

요인	1	2
1	1.000	.550
2	.550	1.000

❖ 사교회전 후 패턴 행렬과 구조 행렬의 원소 값이 의미하는 바



패턴 행렬 (Pattern Matrix)



구조 행렬 (Structure Matrix)

✓ 회전이 가능한 것은 다음과 같은 요인의 비결정성(factor indeterminacy)이 존재하기 때문이다. 요인 축의 회전 결과 새로 얻은 요인부하량 행렬을 $\hat{L}_A$로 부른다면 다음의 식이 성립한다: $\hat{L}\hat{L}' = \hat{L}_A\hat{L}_A'$

$$\Sigma_{zz} = \rho = LL' + \Psi \text{와 동등한 요인분석 결과} \rightarrow$$
$$\Sigma_{zz} = \rho = LMM'L' + \Psi = LM(LM)' + \Psi$$
$$= L_A L_A' + \Psi$$

여기서 M은 $m \times m$ 행렬로서 요인 축들이 회전하는 정도를 결정한다. LM은 새로운 요인부하량 행렬(혹은 factor loading matrix)이며, 앞에서 말한대로 이와 같은 행렬은 무한대로 만들어 낼 수 있다.

➢ 요인의 회전방법 정리

✓ 직교회전은 요인들이 서로 독립적일 때 회전 후 요인들도 역시 서로 독립이 되도록 요인 축을 회전할 때 직각으로 회전하는 방법이다. 여러 직교회전 방법 중 주로 Varimax 방법이 사용된다.
회전을 위해서는, 변환형렬인 M $(m \times m)$를 이용한다. 직교 변환행렬은 $MM' = M'M = I$를 만족하며 이 행렬을 곱하여 회전된 요인부하량행렬은 $\hat{L}_A = \hat{L}M$ 이 된다.

$$\hat{L}\hat{L}' = \hat{L}MM'\hat{L}' = \hat{L}_A\hat{L}_A$$
$$\text{표본상관행렬} = \hat{L}\hat{L}' + \hat{\phi} = \hat{L}_A\hat{L}_A' + \hat{\phi}$$

결국 회전 후에도, 각 관찰변수의 공통성($\hat{h}_j^2$)은 변하지 않고, 이에 따라서 공통성들의 합($\sum_{j=1}^{p}\hat{h}_j^2$)과 요인 고유값들의 합도 당연히 변하지 않는다.

요인이 두 개일 때를 상정하면 원하는 회전의 각도에 따라 요인 축의를 어떻게 회전을 어떻게 구현하는 지 쉽게 알 수 있다. 변환행렬 M는 다음과 같다.

$$M = \begin{bmatrix} \cos x & \sin x \\ -\sin x & \cos x \end{bmatrix}$$ → 시계 방향으로 x 각도만큼 회전

$$M = \begin{bmatrix} \cos x & -\sin x \\ \sin x & \cos x \end{bmatrix}$$ → 시계 반대 방향으로 x 각도만큼 회전

Varimax 방법은 이러한 M 행렬을 구하는 방법이다. 여기서 사용되는 논리는, 회전 후에 각 요인에 연관된 요인부하량들의 분산 합이 최대가 되도록 하는 각도를 구하는 것이다. 분산이 클수록 비슷한 값보다는 작은 값부터 큰 값까지 다양하게 존재하므로 단순구조에 더 가까워질 수 있다.

✓ 사교회전은 연구자가 요약해서 최종 결과물로 얻게 되는 요인들이 서로 상관이 있다는 가정 하에서 요인 축들이 상호 비직각이 되도록 회전하는 방법이다. 최초 독립성을 가정하고 구한 $\hat{L}$ 을 사교회전 한 뒤에는 $\hat{L}_A$ 에서 나타내고 있는 요인간의 상관관계를 다시 해석해야 한다. 사교회전 방법으로는 Promax, Harris-Kaiser 방법 등이 있다.

☆ 다음은 철인 10종 경기(decathlon) 결과 230명의 선수들로부터 얻은 상관계수 자료이다. Track 및 field events (100m, 멀리뛰기, 포환던지기, 높이뛰기, 400m, 110m 허들, 원반던지기, 장대높이뛰기, 창던지기, 1500m)들로 이루어진 이 경기는 각 종목의 기록을 환산하는 표가 있으며 이들 환산된 점수의 합으로 최종 성적을 산출한다.

```
MATRIX DATA VARIABLES = hund lj sp hj fourth h110 disc pv jav r1500
   /FORMAT = FREE LOWER
   /N=230
   /CONTENTS=CORR.

BEGIN DATA
1
.59 1
.35 .42 1
.34 .51 .38 1
.63 .49 .19 .29 1
.40 .52 .36 .46 .34 1
.28 .31 .73 .27 .17 .32 1
.20 .36 .24 .39 .23 .33 .24 1
.11 .21 .44 .17 .13 .18 .34 .24 1
-.07 .09 -.08 .18 .39 .00 -.02 .17 .10 1
END DATA.

FACTOR MATRIX = IN(CORR*)
  /CRITERIA FACTORS(2)
  /EXTRACTION PAF
  /ROTATION PROMAX
  /PLOT EIGEN ROTATION(1,2).
```

✓ NOROTATE, VARIMAX, PROMAX에 따른 요인부하량의 변화

무 회전

요인행렬[a]

	요인	
	1	2
hund	.636	.254
lj	.755	.228
sp	.745	−.593
hj	.596	.104
fourth	.583	.488
h110	.612	.085
disc	.596	−.446
pv	.451	.064
jav	.374	−.231
r1500	.124	.265

직교 회전

회전된 요인행렬[a]

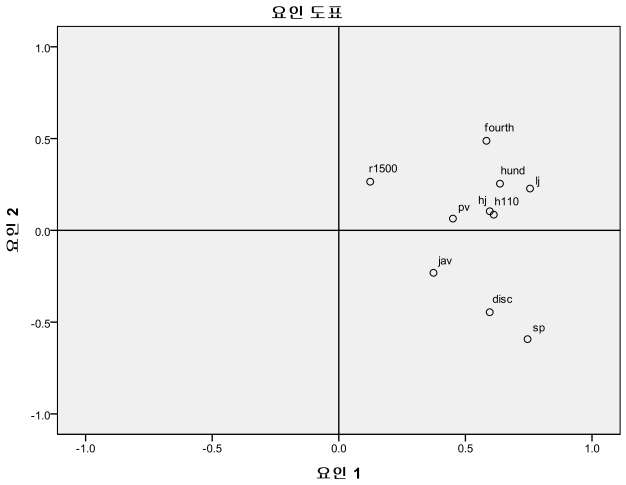
	요인	
	1	2
hund	.634	.261
l	.700	.363
sp	.122	.945
hj	.500	.341
fourth	.758	.056
h110	.498	.365
disc	.117	.735
pv	.368	.268
jav	.107	.426
r1500	.274	−.104

사교 회전

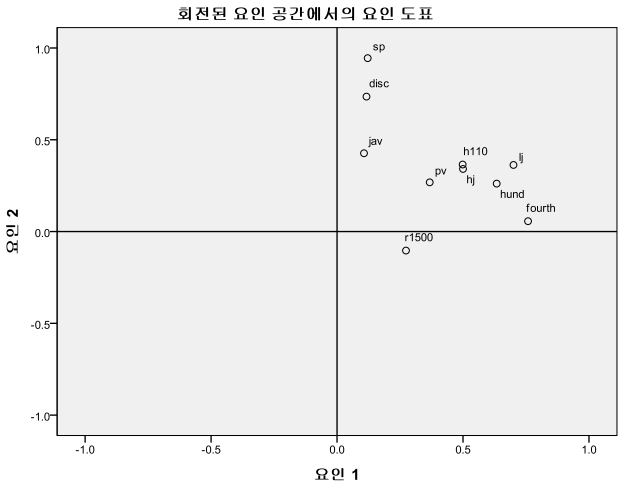
패턴 행렬[a]

	요인	
	1	2
hund	.648	.069
lj	.698	.158
sp	−.097	.997
hj	.478	.203
fourth	.839	−.199
h110	.470	.230
disc	−.051	.768
pv	.347	.168
jav	.015	.432
r1500	.334	−.208

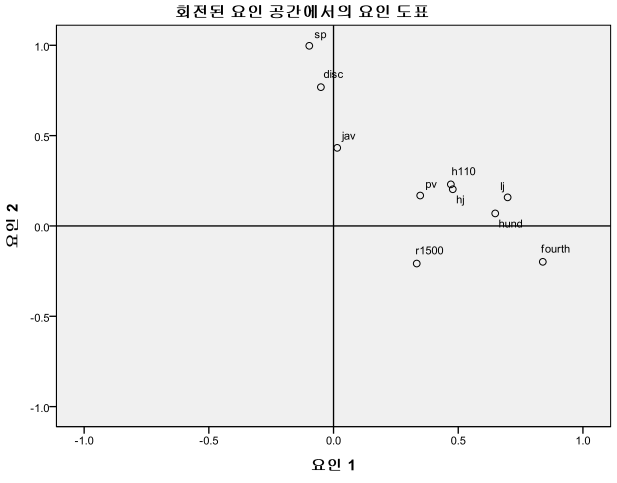
NOROTATE:



VARIMAX:



PROMAX:



✓ 세 경우에 있어서 /PRINT ALL 명령을 이용해 재연된 상관계수 행렬과 잔차 행렬 계수를 구해보면?

➢ **다양한 회전 방법**

✓ 직교회전

- VARIMAX: 요인부하량 제곱들로 계산되는 분산을 최대화 (분산을 최대화하려면 요인부하량이 큰 것은 더욱 크게 작은 것은 더욱 작게 되는 것이 유리하기 때문에)
- QUARTIMAX: 각 관찰변수가 오직 하나의 큰 요인부하량을 가지도록 회전
- EQUIMAX: 단순구조를 지향 (한 관찰변수가 가능한한 적은 요인 수에 큰 요인부하량을 가지고 나머지 요인에 대해서는 0에 가까운 요인부하량을 갖도록)

✓ 사교회전

- PROMAX
- OBLIMAX: 중간 크기의 요인부하량을 줄이기 위해
- QUARTIMIN: 구조 행렬의 벡터들의 내적의 합이 최소가 되도록
- COVARIMIN: 사교회전이면서 VARIMAX의 논리를 따름
- BIQUARTIMIN: QUARTIMIN과 COVARIMIN을 함께 고려
- OBLIMIN: QUARTIMIN과 COVARIMIN을 다른 방식으로 함께 고려

❖ 단순 구조 (Simple Structure): 요인 유형 행렬이 다음과 같은 특징들을 가질 때 단순 구조라고 말한다.

(1) 각 행이 적어도 하나의 0를 가질 때

(2) 각 열이 적어도 공통 요인 수만큼 0을 가질 때

(3) 각각의 두 열(each pair of columns)에서, 하나는 0에 가깝고 다른 하나는 0이 아닌 경우의 관찰변수가 적어도 공통 요인 수만큼 있을 때

➔ 단순 구조의 기본적인 아이디어는 각 관찰변수가 다른 변수들과 가지는 상관이 가능하면 적은 수의 요인으로 잘 설명되도록 하는 것이다.

예를 들어,

$$\begin{bmatrix} .7 & 0 & .5 \\ .6 & .5 & 0 \\ .5 & 0 & 0 \\ 0 & .8 & .4 \\ .7 & .5 & 0 \\ 0 & .4 & 0 \\ 0 & 0 & .8 \\ 0 & .6 & .3 \\ .4 & 0 & .7 \end{bmatrix}$$

☆ 다음은 위에서 본 철인 10종 경기(decathlon) 자료를 주축분해법(PAF: 요인추출을 위한 추정방법 중 하나)과 직교회전(VARIMAX)을 사용하여 분석했을 때 나오는 다양한 SPSS 결과이다.

✓ 직교회전

```
FACTOR MATRIX = IN(CORR*)
  /CRITERIA FACTORS(2)
  /EXTRACTION PAF
  /ROTATION VARIMAX
  /PLOT EIGEN ROTATION(1,2).
```

공통성

	초기	추출
hund	.606	.470
lj	.530	.622
sp	.635	.907
hj	.394	.366
fourth	.601	.578
h110	.370	.382
disc	.545	.554
pv	.236	.207
jav	.232	.193
r1500	.375	.086

추출 방법: 주축요인추출.

설명된 총분산

요인	초기 고유값			추출 제곱합 적재값			회전 제곱합 적재값		
	합계	% 분산	% 누적	합계	% 분산	% 누적	합계	% 분산	% 누적
1	3.790	37.902	37.902	3.315	33.147	33.147	2.215	22.153	22.153
2	1.486	14.855	52.757	1.051	10.506	43.653	2.150	21.500	43.653
3	1.165	11.654	64.411						
4	.926	9.257	73.668						
5	.680	6.797	80.466						
6	.602	6.017	86.483						
7	.526	5.256	91.738						
8	.382	3.815	95.554						
9	.235	2.355	97.908						
10	.209	2.092	100.000						

추출 방법: 주축요인추출.

요인행렬[a]

	요인	
	1	2
hund	.636	.254
lj	.755	.228
sp	.745	-.593
hj	.596	.104
fourth	.583	.488
h110	.612	.085
disc	.596	-.446
pv	.451	.064
jav	.374	-.231
r1500	.124	.265

요인추출 방법: 주축 요인추출.
a. 추출된 2 요인 19의 반복계산이 요구됩니다.

회전된 요인행렬[a]

	요인	
	1	2
hund	.634	.261
lj	.700	.363
sp	.122	.945
hj	.500	.341
fourth	.758	.056
h110	.498	.365
disc	.117	.735
pv	.368	.268
jav	.107	.426
r1500	.274	-.104

요인추출 방법: 주축 요인추출.
회전 방법: Kaiser 정규화가 있는 베리멕스.
a. 3 반복계산에서 요인회전이 수렴되었습니다.

요인 변환행렬

요인	1	2
1	.717	.697
2	.697	-.717

요인추출 방법: 주축 요인추출.
회전 방법: Kaiser 정규화가 있는 베리멕스.

☆ 다음은 위에서 본 철인 10종 경기(decathlon) 자료를 주축분해법(PAF: 요인추출을 위한 추정방법 중 하나)과 사교회전(PROMAX)을 사용하여 분석했을 때 나오는 다양한 SPSS 결과이다.

✓ 사교회전: 직교회전과 각 공통성 값이 같고, 고유값의 합이 같다.

```
FACTOR MATRIX = IN(CORR*)
  /CRITERIA FACTORS(2)
  /EXTRACTION PAF
  /ROTATION PROMAX
  /PLOT EIGEN ROTATION(1,2).
```

공통성

	초기	추출
hund	.606	.470
lj	.530	.622
sp	.635	.907
hj	.394	.366
fourth	.601	.578
h110	.370	.382
disc	.545	.554
pv	.236	.207
jav	.232	.193
r1500	.375	.086

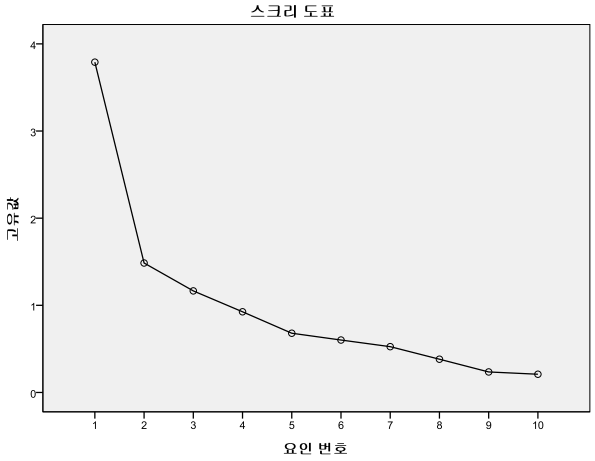
추출 방법: 주축요인추출.

설명된 총분산

요인	초기 고유값			추출 제곱합 적재값			회전 제곱합 적재값[a]
	합계	% 분산	% 누적	합계	% 분산	% 누적	합계
1	3.790	37.902	37.902	3.315	33.147	33.147	2.854
2	1.486	14.855	52.757	1.051	10.506	43.653	2.628
3	1.165	11.654	64.411				
4	.926	9.257	73.668				
5	.680	6.797	80.466				
6	.602	6.017	86.483				
7	.526	5.256	91.738				
8	.382	3.815	95.554				
9	.235	2.355	97.908				
10	.209	2.092	100.000				

추출 방법: 주축요인추출.

a. 요인이 상관된 경우 전체 분산을 구할 때 제곱합 적재값이 추가될 수 없습니다.



요인행렬[a]

	요인	
	1	2
hund	.636	.254
lj	.755	.228
sp	.745	-.593
hj	.596	.104
fourth	.583	.488
h110	.612	.085
disc	.596	-.446
pv	.451	.064
jav	.374	-.231
r1500	.124	.265

요인추출 방법: 주축 요인추출.
a. 추출된 2 요인 19의 반복계산이 요구됩니다.

패턴 행렬[a]

	요인	
	1	2
hund	.648	.069
lj	.698	.158
sp	-.097	.997
hj	.478	.203
fourth	.839	-.199
h110	.470	.230
disc	-.051	.768
pv	.347	.168
jav	.015	.432
r1500	.334	-.208

요인추출 방법: 주축 요인추출.
회전 방법: Kaiser 정규화가 있는 프로멕스.
a. 3 반복계산에서 요인회전이 수렴되었습니다.

구조행렬

	요인	
	1	2
hund	.683	.391
lj	.777	.505
sp	.398	.949
hj	.579	.441
fourth	.740	.218
h110	.585	.464
disc	.331	.743
pv	.431	.341
jav	.229	.440
r1500	.230	-.042

요인추출 방법: 주축 요인추출.
회전 방법: Kaiser 정규화가 있는 프로멕스.

요인 상관행렬

요인	1	2
1	1.000	.497
2	.497	1.000

요인추출 방법: 주축 요인추출.
회전 방법: Kaiser 정규화가 있는 프로멕스.

☆ 요인분석 연습: 첨부한 위기청소년 진단 척도 (청소년 상담원)를 보면 다음 여섯 개 영역에 대하여 총 81개 문항이 존재한다. 탐색적 요인분석을 통하여 의도한바 대로 해당 문항들이 여섯 개 구인으로 묶이는 지를 검토해 보자.

- 심리: 문항13-34
- 개인: 문항35-49
- 가정: 문항50-69
- 학교: 문항70-84
- 또래: 문항85-89
- 지역사회: 문항90-93

5. 요인분석: 모수 추정

- **요인부하량 행렬(L)의 추정**: $\rho = LL' + \Psi$ 에서 L과 Ψ에 대해서는 알려져 있지 않으므로 이들 요인부하량 값들과 d_j^2 값들은 주어진 자료(표본)로부터 추정되어야 한다. 아래에서는 SPSS에서 사용되는 다양한 요인부하량 추정 방법에 대해서 살펴본다.

  각 추정 방법을 통하여 사용하는 요인분석 모형의 모수를 추정하고, 이를 통해 재생 상관 행렬(reproduced correlation matrix) 및 잔차행렬(residual matrix)을 계산하여 적합도(goodness-of-fit)를 검증한다. 적합도를 검증하는 또다른 방법은 잔차제곱평균제곱근(root mean square residual; RMSR)을 계산하는 것이다. *RMSR*은 잔차 행렬에서 각 요소를 제곱하여 더한 뒤에 이를 계산에 사용한 요소의 수, 즉 p(p-1)/2로 나눈 뒤에 제곱근을 씌워서 구하게 된다.
  예를 들어서, 지난 시간에 다룬 decathlon 자료에서 두 개의 요인을 사용하고 PAF 방법으로 요인을 추출하고 PROMAX로 요인 축을 회전한 경우에 얻은 잔차행렬과 *RMSR*은 다음과 같다.

잔차행렬										
	hund	lj	sp	hj	fourth	h110	disc	pv	jav	r1500
hund		.051	.026	−.066	.135	−.011	.014	−.103	−.069	−.216
lj	.051		−.008	.036	−.061	.038	−.039	.005	−.020	−.064
sp	.026	−.008		−.003	.045	−.046	.021	−.058	.024	−.015
hj	−.066	.036	−.003		−.108	.086	−.039	.115	−.029	.079
fourth	.135	−.061	.045	−.108		−.059	.040	−.064	.025	.188
h110	−.011	.038	−.046	.086	−.059		−.007	.049	−.029	−.099
disc	.014	−.039	.021	−.039	.040	−.007		.000	.014	.024
pv	−.103	.005	−.058	.115	−.064	.049	.000		.086	.097
jav	−.069	−.020	.024	−.029	.025	−.029	.014	.086		.115
r1500	−.216	−.064	−.015	.079	.188	−.099	.024	.097	.115	

$$RMSR = \sqrt{\frac{1}{45}(.051^2 + .026^2 + \cdots + .115^2)} = .073$$ (rough guideline: 0.05)

➢ **주성분방법** (Principal Component Method): d_j^2이 거의 0이거나 무시할만큼 작아서 $\rho = LL'$로 볼 수 있는 경우(다시 말하면, 주성분분석에서의 가정)에 사용되는 방법. 상관행렬을 스펙트럼 분해할 경우 다음과 같이 표시된다.

$$\rho = V\Lambda V' = V\begin{bmatrix} \lambda_1 & & & & \\ & \ddots & & 0 & \\ & & \lambda_m & & \\ & 0 & & \ddots & \\ & & & & \lambda_p \end{bmatrix} V' = \lambda_1 e_1 e_1' + \cdots + \lambda_m e_m e_m' + \cdots + \lambda_p e_p e_p' = LL'$$

✓ 주성분분석에서 eigenvalue가 1보다 큰 경우라든가 여러 방법을 이용하여 분석에 사용할 주성분의 수를 결정하게 된다. 만약 m개의 주성분으로 관찰변수들의 상관구조가 대부분 설명될 수 있다면 (즉, 첫번째부터 m번째까지의 λ가 상당히 크고 나머지는 작다), 상관행렬 $\{\rho = LL' : (p\times p) = (p\times p)(p\times p)\}$은 다음과 같이 근사하게 된다.

$$\hat{\rho} \cong \hat{L}\hat{L}' : (p\times p) = (p\times m)(m\times p)$$

$$\hat{\rho} \cong \hat{\lambda}_1 e_1 e_1' + \cdots + \hat{\lambda}_m e_m e_m' = \begin{bmatrix} \sqrt{\hat{\lambda}_1}e_1 & \sqrt{\hat{\lambda}_2}e_2 & \cdots & \sqrt{\hat{\lambda}_m}e_m \end{bmatrix} \begin{bmatrix} \sqrt{\hat{\lambda}_1}e_1' \\ \sqrt{\hat{\lambda}_2}e_2' \\ \vdots \\ \sqrt{\hat{\lambda}_m}e_m' \end{bmatrix} = \hat{L}\hat{L}'$$

따라서 추정된 요인부하량 행렬 $\hat{L}$ ($p \times m$)은 다음과 같다.

$$\hat{L} = \left[\sqrt{\hat{\lambda}_1}\hat{e}_1 \quad \sqrt{\hat{\lambda}_2}\hat{e}_2 \quad \cdots \quad \sqrt{\hat{\lambda}_m}\hat{e}_m \right]$$

- ✓ 표준화 관찰변수를 활용한 주성분분석에서, 선택된 m개 주성분들이 설명할 수 있는 누적비율은 전체 관찰변수 개수 분의 고유값의 합, 즉 $\sum_{i=1}^{m} \frac{\hat{\lambda}_i}{p}$ 이 된다.

- **비가중최소자승법** (Unweighted Least Squares Method; ULS): 관찰상관행렬과 재생상관행렬 간의 차이를 최소화(즉 *RMSR*을 최소화)할 수 있는 L과 Ψ의 모수를 추정해내는 방법이다.
  - ✓ 장점: 통계적 모형의 모수를 추정하는 데에 있어서 가장 간단한 방법이며 직관적이며, 그 결과를 이해하기 쉽다.
  - ✓ 단점: 추정 과정이 끝난 후 제대로 모수 추정이 되었는지에 대한 통계적 검증(test of significance) 방법이 없다는 점이 단점이다.

- **가중최소자승법** (Weighted Least Squares Method; WLS): ULS와 거의 비슷한 방법이지만, 차이점은 잔차행렬의 각 요소에 가중치를 부여한다는 점이다. 즉 관찰변수간 상관들이 모두 같은 정확도로 산출되는 것은 아니기 때문에 이 점을 이용하여 가중치를 부여한다.
  - ✓ 장점: ULS에 비하여 일부 상관값들의 중요성을 강조할 수 있다는 점이며, 또한 표집 크기가 클 경우 카이제곱 검증을 통하여 적합도에 대한 유의도 검증을 실시할 수 있다는 점이 장점이다.
    영가설 (Ho) : The model fits the given data.
    대립가설 (H_A) : The model does not fit.
  - ✓ 단점: 추정 과정 및 계산이 매우 복잡하다. 또한 표집 크기가 매우 큰 경우 카이제곱 검증은 실제 적합도와 관계없이 영가설을 부정하는 경우가 많기 때문에, 이 경우 모형 부적합의 증거가 되지 못한다.

➢ **주축요인법** (Principal Factor Method; 주축분해법, Principal Axis Factoring): 앞에서 공부한 바와 같이, 주성분분석과 달리 요인분석에서는 d_j^2이 0과 유의미하게 다른 값으로 존재한다고 보다. 그러나 주성분방법은 이러한 특성을 반영하고 있지 않다.

✓ 주축요인법에서는 상관행렬의 주대각선 값을 1.0 대신 이를 h_j^2으로 대치시킨 축소상관행렬(reduced correlation matrix; 이하 RCM)을 구하여 요인분석을 진행하게 한다. RCM을 구하는 절차는 다음과 같다. 이 과정에서 공통요인(common factor)으로 뽑히게 될 잠재변수들만 사용하기 때문에 추출된 요인들은 공통성(h_j^2)을 최대한 설명하게 된다.

Step 1.

RCM의 대각원소 초기값으로, 각 관찰변수마다 다른 $p-1$개 관찰변수를 독립변수로 하여 회귀분석을 실시하고 구한 $R_j^2 (= h_j^{2(0)})$를 사용한다. 이 $RCM^{(0)}$을 사용하여 요인부하량 행렬을 구하면 다음과 같다.

$RCM^{(0)} = L^{(0)} L'^{(0)} : (p \times p) = (p \times p)(p \times p)$

❖ 요인간 독립을 가정하는 이 초기 상황에서, 주성분분석에서의 가중치(weight; w_{ij})와 각 요인부하량(a_{ij})의 관계는 다음과 같다:

$a_{ij} = w_{ij} \cdot \sqrt{\lambda_i} \quad (i = 1,...,m;\ j = 1,...p)$

다음, 일반적으로 주성분방법을 통하여 (고유값, 스크리도표 등) 결정된 요인의 수 m을 이용하여 다음과 같은 근사를 얻게 된다.

$RCM^{(0)} \cong L^{(0)} L'^{(0)} : (p \times p) = (p \times m)(m \times p)$

이렇게 얻은 요인부하량 행렬을 이용하여 다시 공통성을 계산하고 이를 $h_j^{2(1)}$ 라고 표시한다.

Step 2.

Step 1에서 얻은 $h_j^{2(1)}$ 으로 표본상관행렬의 대각원소를 다시 대치시킨 후에 이를 $RCM^{(1)}$이라고 부른다. 이를 이용하여 다시 $L^{(1)}$과 $h_j^{2(2)}$를 순차적으로 구한다.

Step 3.

위의 Step 1과 Step 2를 N 번 계속 반복하여 $h_j^{2(N)}$과 $h_j^{2(N-1)}$의 차이가 굉장히 미미하다면 중단하고 여기서의 RCM을 최종 축소상관행렬로 결정한다.

✓ 장점: PAF의 경우, 요인의 분산을 최대화하려는 기본적인 의도 하에서 공통성을 추정하려는 목표로 개발되었기 때문에, 주성분분석을 시도하려 했던 관점에서 요인분석을 하고자 할 때 가장 우선적으로 고려될 수 있는 요인추출방법이다. 다시 말해서, PAF는 축소상관행렬을 사용해서 주성분분석을 하는 것이라고 말할 수 있다 (축소상관행렬의 스펙트럼 분해)

✓ 최대우도법 (Maximum Likelihood Method): 앞의 두 방법에서는 $Z = LF + \varepsilon$에서 관찰변수의 고유한 부분이자 측정의 오차를 표현하고 있는 ε의 분포에 대한 가정을 따로 하지 않았다. 이는 이 식에서 무선성분(random component)이라고 할 수 있는 관찰점수들의 분산에 대해 달리 가정을 하지 않았다는 의미이기도 하다.

✓ 반면에 최대우도법에서는 관찰변수들의 분포가 다변량정규분포를 따른다는 추가적 가정을 한다.

$$X \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad Z \sim N(\mathbf{0}, \boldsymbol{\Sigma}_Z)$$

그 다음, 다변량정규분포에서의 확률밀도함수를 이용하여 어느 경우에(즉 $\Sigma_Z = \rho = LL' + \Psi$ 에서 L과 Ψ가 어떤 값을 가질 때) 주어진 관찰자료가 나타났을 가능성이 가장 클까를 생각해볼 수 있다.

✓ 이런 가능도를 나타내는 함수를 가능도함수(혹은 우도 함수; likelihood function)라고 한다. 이 가능도함수를 최대화하는 L과 Ψ을 구하려면 이들 값으로 가능도함수를 편미분한 뒤에 그 미분된 함수를 0으로 놓고 구하면 된다.

이 때, 주성분분석 등을 통하여 요인의 수(m)를 미리 정한 뒤에 이러한 과정이 전개되므로 L은 $(p \times m)$ 행렬이다.

✓ 장점

- 통계적 유의도 검증 가능 (최대우도법을 사용하면 요인분석 모형의 적합성에 대한 통계적 검정 가능하다. 이는 사용된 요인의 수가 통계적으로 적합한 지를 검정할 수 있다는 의미이므로 유용하다)
  영가설 (Ho): $\rho = LL' + \Psi$
  → 모형에 포함된 요인의 수는 m개면 충분
  대립가설 (H_A): $\rho \neq LL' + \Psi$
  → 모형에 m보다 더 많은 요인 필요

- 표집 크기가 클 때 다른 방법에 비하여 비교적 더 정확한 모수 추정
- 다변량정규분포에 대한 가정이 충족된다면, 어떤 방법보다도 더 정확한 재생상관행렬을 산출할 수 있다.

✓ 단점

- 추정 과정이 매우 복잡하고 시간이 많이 걸린다는 점. 그러나 SPSS 프로그램 등의 발전으로 이 점은 큰 문제가 되지 않음.
- 관찰변수의 수가 많은 경우 improper solution 이 발생할 가능성이 높다: 가능도함수의 편미분된 결과를 가지고 반복적 방법을 통하여 L과 ϕ의 해를 찾아가는 과정에서 추정된 공통성 $\hat{h}_j^2$ (remember $h_j^2 + d_j^2 = \mathbf{1}$)의 값이 1 보다 커져서 uniqueness 의 분산 $\hat{d}_j^2$ (즉 $\hat{\Psi}$ 의 j 번째 대각원소)이 0 보다 작게 되는 상황이 발생할 수 있다.

  → 이를 Heywood case 라고 부른다.

❖ Heywood case는 어느 추정 방법을 사용하건 발생할 가능성이 있다.

✓ Heywood 상황이 발생하는 원인을 정리해 보면 다음과 같다.

(1) 공통성의 초기값이 부적합한 경우
(2) 고려된 요인의 수가 너무 많거나 작은 경우
(3) 신뢰성 있는 추정치를 구하기에는 피험자 수가 너무 적은 경우
(4) 요인분석 모형으로 적합되기에 부적절한 자료인 경우

✓ Heywood case가 발생하는 경우의 예: 관찰된 상관행렬이 다음과 같을 때, 이를 요약하는 하나의 요인만을 고려(i.e., single common-factor model)하여 요인을 추출하는 경우

$$R = \begin{bmatrix} 1.0 & & \\ 0.9 & 1.0 & \\ 0.4 & 0.9 & 1.0 \end{bmatrix}$$

단일 공통요인 모형을 통해서 다음과 같은 여섯 가지 식을 유도할 수 있다. 먼저, 관찰변수 간 상관을 통해 다음 세 개의 식을 구할 수 있다.

$0.9 = l_1 l_2,\ 0.9 = l_2 l_3,\ 0.4 = l_1 l_3$

다음, 관찰변수의 분산을 통해 다음 세 개의 식을 구할 수 있다.

$\hat{l}_1^2 + \psi_1 = 1.0,\ \hat{l}_2^2 + \psi_2 = 1.0,\ \hat{l}_3^2 + \psi_3 = 1.0$

이를 이용해 각 요인부하량을 계산해 보면

$l_1 = .63$
$\hat{l}_2 = 1.43$
$\hat{l}_3 =$

그러므로, 두 번째 변수의 경우 그 공통성(λ_2^2)이 1보다 크게 되고 특이성(uniqueness)은 음수 값이 된다: 요인분석 결과에 뭔가 문제가 있다는 경고!

➔ 대책: ML을 사용해도 추정된 값들이 MLE가 되지 않는 등 improper solution이 될 수 있으므로, 다른 요인추출 방법 등을 시도해 보거나 R^2 같은 다른 대안적 방법으로 얻은 값들을 공통성으로 사용할 수 있다.

➢ 기타 SPSS에서 사용되는 요인분석 모수추정 방법

- ✓ Alpha factoring (알파 분해법): 이 방법은 검사이론과 밀접하게 관련되어 있다. 즉 검사의 문항들을 대상으로 하여 요인분석을 실시한다고 할 때, 각각의 문항이 문항의 전집으로부터 무선적으로 표집된 것이라고 보고 이들 문항들로 이루어진 검사의 신뢰도(Cronbach's alpha)가 최대가 되도록 요인부하량을 추출한다.
  달리 말하면, 우리가 가지고 있는 자료의 변수들이 전체 변수들의 모집단으로부터 표집된 것이라고 보고, 이들 표집된 변수들로부터 설정한 요인분석 결과가 모집단 수준에서도 합당한 추정이 될 지를 결정하는 것이다.

- ✓ Image factoring (이미지 분해법): 이 방법은 변수의 "이미지"라는 개념에 기반을 두는데, 한 변수를 종속변수로 하고 다른 변수들을 독립변수로 했을 때 계산된 중다상관(R^2)을 공통성으로 하여 분석을 실시한다.

➢ 각 요인분석 모수추정 방법의 특징

- ✓ 표집 크기: GLS의 경우 사례 수가 200 이상일 때만 유의도 검증을 실시할 수 있다.
- ✓ 변수의 수: 일반적으로 변수의 숫자가 증가할수록 여러 다른 방법들로부터 구한 모수추정 결과가 더 유사해진다.
- ✓ 공통성의 양: 상관행렬을 자료로 할 때 모든 공통성의 값이 1에 가까울 때, 어떤 추정 방법을 쓰던 매우 유사한 결과가 산출된다.
- ✓ 공통성의 변산: 모든 변수의 공통성 값이 비슷할수록 주성분분석과 PAF는 비슷한 결과를 산출한다.

✓ 앞에서 본 바와 같이 decathlon 자료(100m, 멀리뛰기, 포환던지기, 높이뛰기, 400m, 110m 허들, 원반던지기, 장대높이뛰기, 창던지기, 1500m)에서 두 개의 요인을 사용할 때 *RMSR*은 결과가 0.05를 넘어서 만족스러운 적합도를 보이지 않았다. 다음은 세 개의 요인을 사용할 때 얻을 수 있는 재생산상관행렬 및 잔차행렬이다.

재연된 상관계수

		hund	lj	sp	hj	fourth	h110	disc	pv	jav	r1500
재연된 상관계수	hund	.675	.613	.313	.397	.513	.465	.242	.262	.109	−.061
	lj	.613	.623	.431	.453	.529	.483	.348	.329	.209	.100
	sp	.313	.431	.848	.391	.158	.408	.686	.314	.424	−.094
	hj	.397	.453	.391	.363	.388	.359	.323	.281	.219	.179
	fourth	.513	.529	.158	.388	.575	.378	.137	.290	.112	.336
	h110	.465	.483	.408	.359	.378	.385	.329	.262	.196	.037
	disc	.242	.348	.686	.323	.137	.329	.557	.263	.352	−.038
	pv	.262	.329	.314	.281	.290	.262	.263	.226	.190	.197
	jav	.109	.209	.424	.219	.112	.196	.352	.190	.245	.107
	r1500	−.061	.100	−.094	.179	.336	.037	−.038	.197	.107	.808
잔차 b	hund		−.023	.037	−.057	.117	−.065	.038	−.062	.001	−.009
	lj	−.023		−.011	.057	−.039	.037	−.038	.031	.001	−.010
	sp	.037	−.011		−.011	.032	−.048	.044	−.074	.016	.014
	hj	−.057	.057	−.011		−.098	.101	−.053	.109	−.049	.001
	fourth	.117	−.039	.032	−.098		−.038	.033	−.060	.018	.054
	h110	−.065	.037	−.048	.101	−.038		−.009	.068	−.016	−.037
	disc	.038	−.038	.044	−.053	.033	−.009		−.023	−.012	.018
	pv	−.062	.031	−.074	.109	−.060	.068	−.023		.050	−.027
	jav	.001	.001	.016	−.049	.018	−.016	−.012	.050		−.007
	r1500	−.009	−.010	.014	.001	.054	−.037	.018	−.027	−.007	

$$RMSR = \sqrt{\frac{1}{45}((-.023)^2 + .037^2 + \cdots + (-.007)^2} = .048$$

따라서 좋은 적합도를 보이고 있다.

하지만, 위의 잔차행렬에서 절대값이 0.05를 넘는 요소를 갖는 경우가 28% (=13/45)에 달하여 이론적 배경이 존재한다면 네 개의 요인을 고려해볼 필요도 있다.

2요인 모형 (PAF)

KMO 와 Bartlett 의 검정

표준형성 적절성의 Kaiser-Meyer-Olkin 측도.		.729
Bartlett 의 구형성 검정	근사 카이제곱	833.608
	자유도	45
	유의확률	.000

패턴 행렬[a]

	요인	
	1	2
hund	.648	.069
lj	.698	.158
sp	-.097	.997
hj	.478	.203
fourth	.839	-.199
h110	.470	.230
disc	-.051	.768
pv	.347	.168
jav	.015	.432
r1500	.334	-.208

요인추출 방법: 주축 요인추출.
회전 방법: Kaiser 정규화가 있는 프로멕스.

요인 상관행렬

요인	1	2
1	1.000	.497
2	.497	1.000

요인추출 방법: 주축 요인추출.
회전 방법: Kaiser 정규화가 있는 프로멕스.

3요인 모형 (PAF)

KMO 와 Bartlett 의 검정

표준형성 적절성의 Kaiser-Meyer-Olkin 측도.		.729
Bartlett 의 구형성 검정	근사 카이제곱	833.608
	자유도	45
	유의확률	.000

패턴 행렬[a]

	요인		
	1	2	3
hund	.893	-.130	-.225
lj	.736	.104	-.015
sp	.031	.897	-.098
hj	.418	.238	.130
fourth	.732	-.157	.244
h110	.514	.185	-.046
disc	.009	.740	-.034
pv	.247	.245	.180
jav	-.056	.509	.135
r1500	-.025	.029	.904

요인추출 방법: 주축 요인추출.
회전 방법: Kaiser 정규화가 있는 프로멕스.

요인 상관행렬

요인	1	2	3
1	1.000	.491	.186
2	.491	1.000	-.031
3	.186	-.031	1.000

요인추출 방법: 주축 요인추출.
회전 방법: Kaiser 정규화가 있는 프로멕스.

KMO(Kaiser-Meyer-Olkin) 측정치는 변인 쌍의 상관 관계가 다른 변인에 의해 얼마나 잘 설명되는가의 정도 혹은 '표본의 적절성'을 나타내는 수치이다. 이 수치가 작으면 요인분석을 위한 요인의 선정이 좋지 못한 것이다. 보통 .90이상이면 매우 좋은 편, .50 미만이면 받아들일 수 없는 것으로 판정함.

Bartlett 구상 검정치는 요인분석모형의 적합성 여부를 나타내는 수치로 상관관계행렬이 단위행렬이라는 영가설을 검증하기 위한 측정치이다. 영가설을 기각해야 요인분석모형을 사용할 수 있다.

ULS

패턴 행렬[a]

	요인		
	1	2	3
hund	.863	−.118	−.203
lj	.750	.090	−.026
sp	.029	.900	−.082
hj	.438	.222	.110
fourth	.738	−.161	.226
h110	.526	.173	−.057
disc	.008	.744	−.023
pv	.270	.228	.153
jav	−.043	.502	.124
r1500	−.015	.035	1.006

요인추출 방법: 가중되지 않은 최소제곱.
회전 방법: Kaiser 정규화가 있는 프로멕스.

GLS

패턴 행렬[a]

	요인		
	1	2	3
hund	.921	−.094	−.218
lj	.689	.123	−.016
sp	.033	.946	−.066
hj	.384	.242	.125
fourth	.785	−.143	.264
h110	.497	.154	−.075
disc	.028	.744	−.009
pv	.258	.168	.133
jav	−.030	.485	.115
r1500	−.030	.044	1.006

요인추출 방법: 일반화 최소제곱.
회전 방법: Kaiser 정규화가 있는 프로멕스.

ML

패턴 행렬[a]

	요인		
	1	2	3
hund	.917	−.100	−.219
lj	.670	.136	−.016
sp	.054	.894	−.075
hj	.358	.260	.127
fourth	.789	−.152	.261
h110	.470	.182	−.073
disc	.015	.772	−.011
pv	.233	.202	.136
jav	−.037	.499	.114
r1500	−.027	.046	1.005

요인추출 방법: 최대 우도.
회전 방법: Kaiser 정규화가 있는 프로멕스.

ALPHA

패턴 행렬[a]

	요인		
	1	2	3
hund	.913	−.158	−.243
lj	.746	.100	.009
sp	.012	.904	−.145
hj	.428	.229	.166
fourth	.678	−.147	.224
h110	.521	.195	−.015
disc	.015	.705	−.062
pv	.215	.291	.235
jav	−.088	.526	.150
r1500	−.018	−.011	.804

요인추출 방법: 알파 요인추출.
회전 방법: Kaiser 정규화가 있는 프로멕스.

➢ 요인점수의 추정 (Estimating Factor Scores)

✓ 요인부하량과 특이성에 대한 추정이 요인분석의 주된 관심 사항이긴 하지만 주성분분석에서의 주성분점수(principal component scores)처럼 요인점수(factor scores)를 계산하여 이를 추가적 분석에 사용하는 것이 가능하다.

✓ 요인분석에서는 요인의 선형결합으로 관찰점수를 나타내기 때문에, 주성분분석에서 처럼 관찰변수 값들과 가중치를 이용하여 쉽게 계산할 수 없고 따로 추정을 하기 위한 절차를 따라야 한다.

✓ 요인점수를 계산하기 위한 여러가지 방법이 있으나 여기에서는 두 가지 방법에 대해서만 알아보기로 한다: Regression method & Bartlett's method.

✓ Regression Method: 기본적으로 중다회귀방법을 사용하여 관찰점수의 선형결합으로 요인점수를 추정한다. 당연히 요인점수는 미리 알려진 것이 아니므로, 전에 배운 OLS 방법 등을 통하여 회귀계수를 추정하는 것은 불가능하다 → 따라서 관건은 요인분석을 통하여 추정된 **구조행렬**(structure matrix: 요인과 관찰변수 간 상관. 물론, 요인간 직교할 시에는 요인부하량 행렬과 같다) 등의 정보를 이용하여 어떻게 아래 식에서의 회귀계수를 결정하는가 하는 것이다.

$$\hat{f}_{ij} = \hat{b}_1 z_{i1} + \hat{b}_2 z_{i2} + \cdots + \hat{b}_p z_{ip}$$

이 식을 행렬로 나타내면 다음과 같다: $\hat{F}_{n\times m} = Z_{n\times p} \cdot \hat{B}_{p\times m}$

이 식의 양쪽에 $\frac{1}{n} \cdot Z'$를 곱해주면: $\frac{1}{n} \cdot Z' \cdot \hat{F}_{n\times m} = \frac{1}{n} \cdot Z' \cdot Z_{n\times p} \cdot \hat{B}_{p\times m}$

이 식은 왼쪽은 관찰점수와 요인 간 상관을 나타내고 오른 쪽은 처음 두 항이 관찰점수간 상관을 나타내기 때문에, 아래와 같이 정리된다.

$\Lambda_{p\times m} = R_{p\times p} \cdot \hat{B}_{p\times m}$ 따라서, 최종적으로 계수행렬은 다음과 같다.

$\hat{B}_{p\times m} = R^{-1}_{p\times p} \cdot \Lambda_{p\times m}$ 그러므로, 요인점수는 다음처럼 구할 수 있다.

$$\hat{F}_{n\times m} = Z_{n\times p} \cdot R^{-1}_{p\times p} \cdot \Lambda_{p\times m}$$

✓ Bartlett's method (이 방법은 Maximum Likelihood method 혹은 Weighted least squares method라고도 불린다)

Bartlett(1937)은 각 관찰변수의 unique factor에 해당하는 부분의 제곱 합이 최소가 되도록 하면 의도한 공통 요인들의 의미에 충실한 요인점수들을 구해낼 수 있다고 제안하였다. 다른 방법에 비한 장점으로서, 그의 방법은 진 요인점수에 대한 unbiased estimates를 산출한다고 알려져 있고 이는 최대우도방법에 기초하여 요인점수를 계산하기 때문이다 (Hershberger, 2005).

그가 제안한 방법에 따라서, 요인점수를 계산하기 위한 관찰점수의 가중치 (혹은 계수)는 다음과 같다.

$\hat{B}_{p\times m} = \Psi^{-1}\Lambda(\Lambda'\Psi^{-1}\Lambda)^{-1}$ 그러므로, 요인점수는 다음처럼 계산한다.

$\hat{F}_{n\times m} = Z_{n\times p}\cdot\Psi^{-1}\Lambda(\Lambda'\Psi^{-1}\Lambda)^{-1}$

☆ 요인분석 요인점수 계산 연습: 위기청소년 진단 척도 심리 영역

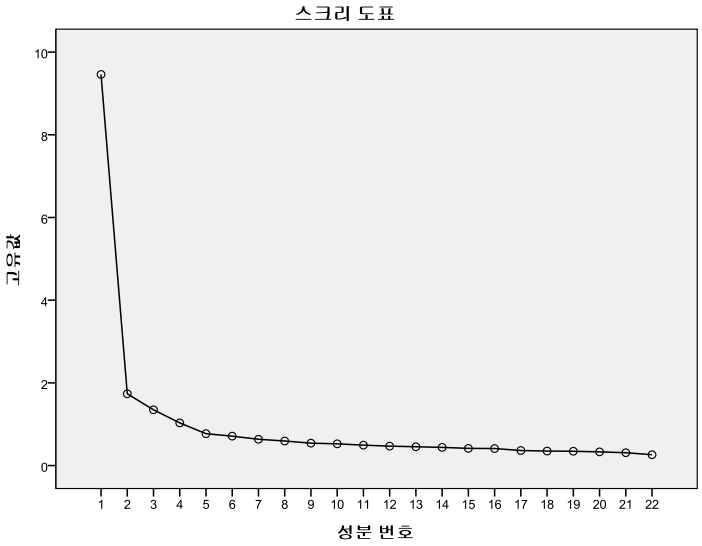
```
FACTOR
  /VARIABLES 심리1 심리2 심리3 심리4 심리5 심리6 심리7 심리8 심리9 심리10 심리11 심리12 심리13 심리14
심리15 심리16 심리17 심리18 심리19
    심리20 심리21 심리22
  /MISSING LISTWISE
  /ANALYSIS 심리1 심리2 심리3 심리4 심리5 심리6 심리7 심리8 심리9 심리10 심리11 심리12 심리13 심리14
심리15 심리16 심리17 심리18 심리19
    심리20 심리21 심리22
  /PRINT INITIAL EXTRACTION
  /PLOT EIGEN
  /CRITERIA FACTORS(1) ITERATE(25)
  /EXTRACTION ALPHA
  /ROTATION NOROTATE
  /SAVE REG(ALL)
  /METHOD=CORRELATION.
```

요인행렬[a]

	요인
	1
심리1	.683
심리2	.740
심리3	.646
심리4	.665
심리5	.560
심리6	.701
심리7	.660
심리8	.735
심리9	.576
심리10	.695
심리11	.651
심리12	.515
심리13	.458
심리14	.544
심리15	.527
심리16	.560
심리17	.658
심리18	.585
심리19	.689
심리20	.689
심리21	.656
심리22	.651

요인추출 방법: 알파 요인추출.

a. 추출된 1 요인 5의 반복계산이 요구됩니다.



- Regression Method 등에 따른 요인 점수는 /SAVE 명령에 따라서 변수로 기록되어 자료 파일에 함께 기록된다.