인 자 분 석

이재용,임요한

서 울 대 학 교
통 계 학 과

2017년 8월

노트. 다루는 내용 I

1. Everitt and Hothorn (2011) 5장의 요약이다.
2. 외부 강의를 할 때 학생들이 관심을 많이 보이는 주제인 것 같다. 학생들 중에 문과 사람들이 많기 때문이다.
3. 인자분석의 2가지 종류 : 이 이야기는 하지 않는다.
   - 3.1 탐색적 인자분석(exploratory factor analysis) : 관측변수와 짐재변수의 관계를 탐색한다. 잠재변수에 조건을 걸지 않는다.
   - 3.2 확증적 인자분석(confirmatory factor analysis) : 주어진 인자 모형이 주어진 자료를 잘 설명하는지 검정

인자분석 I

인자분석이란?

1. 관측할 수 없는 잠재변수(latent variable)들 (예. 지능, 사회적 계층)과 관측된 변수들의 관계를 밝히는 분석
2. 관계는 중회귀분석인데 관측변수가 반응변수가 되고 잠재변수가 예측변수가 된다.
3. 잠재변수는 공통인자(common factor)라고 하고, 회귀계수는 인자적재값(factor loading)이라 한다.

인자분석 II

한 개 인자모형의 예 (Spearman, 1904)

classics (x_1), french (x_2), english (x_3) 세 개의 변수가 관측

상관행렬은

$$R = \begin{pmatrix} 1.00 & & \\ 0.83 & 1.00 & \\ 0.78 & 0.67 & 1.00 \end{pmatrix}$$

일개인자모형

$$\begin{aligned} x_1 &= \lambda_1 f + u_1 \\ x_2 &= \lambda_2 f + u_2 \\ x_3 &= \lambda_3 f + u_3. \end{aligned}$$

여기서 f는 인자이고, $\lambda_1, \lambda_2, \lambda_3$는 인자적재(factor loading) 이다.

인자분석 III

노트.

1. 스피어만은 최초로 인자모형을 제안하였다.
2. 세 개의 관측값, classics, french, english가 모후 한개의 인자에 영향을 받는다. 이 인자는 관측되지 않고 지능이라 해석된다.

인자분석 IV

k인자 모형

$$
\begin{aligned}
x_1 &= \lambda_{11} f_1 + \lambda_{12} f_2 + \ldots + \lambda_{1k} f_k + u_1 \\
x_2 &= \lambda_{21} f_1 + \lambda_{22} f_2 + \ldots + \lambda_{2k} f_k + u_2 \\
&\vdots \\
x_q &= \lambda_{q1} f_1 + \lambda_{q2} f_2 + \ldots + \lambda_{qk} f_k + u_q
\end{aligned}
$$

$\mathbf{x} = (x_1, \ldots, x_q)^T$: 관측변수

$\mathbf{f} = (f_1, \ldots, f_k)^T$, $k < q$: 공통인자

가정

1. u_i는 서로 상관이 없고(not correlated)
2. $\mathbf{u}, \mathbf{f}$는 서로 상관이 없다.
3. f_i도 서로 상관이 없다.

인자분석 V

k인자 모형 : 행렬식

$$\mathbf{x} = \Lambda\mathbf{f} + \mathbf{u}$$

$$\Lambda == \begin{bmatrix} \lambda_{11} & \lambda_{12} & \cdots & \lambda_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{11} & \lambda_{12} & \cdots & \lambda_{1k} \end{bmatrix}, \mathbf{f} = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_k \end{bmatrix}, \mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_q \end{bmatrix}$$

가정

$\mathbf{x} = (x_1, \ldots, x_q)^T$: 관측변수

$\mathbf{f} = (f_1, \ldots, f_k)^T$, $k < q$: 공통인자

$$\begin{aligned} \mathbf{f} &\sim N(0, I) \\ \mathbf{u} &\sim N(0, \Psi), \ \Psi = diag(\psi_1, \ldots, \psi_k). \end{aligned}$$

동일한 식을 뒤 쪽은 행렬식으로 쓴 것이다.

인자분석 VI

k인자 모형의 분산들

x_i의 분산

$$\mathbb{V}ar(x_i) = \sigma_i^2 = \sum_{j=1}^{k} \lambda_{ij}^2 + \psi_i$$

$h_i = \sum_{j=1}^{k} \lambda_{ij}$: 공통성(communality). x_i의 분산 중 다른 x_j 들과 공유하는 인자 때문에 생기는 분산

ψ_i : 특정분산(specific or unique variance). x_i 고유의 분산.

로딩값의 가로.세로 제곱합에 대한 설명

인자분석 VII

x_i와 x_j의 공분산

$$\mathbb{C}ov(x_i, x_j) = \sigma_{ij} = \sum_{l=1}^{k} \lambda_{il}\lambda_{jl}$$

x 분산의 행렬식

$$\mathbb{V}ar(\mathbf{x}) = \Sigma = \Lambda\Lambda^T + \Psi$$

노트. 추정

Σ는 표본분산 S로 추정이 되고 이를 이용해 Λ, Ψ, k를 추정해야한다. 상관행렬만 가지고도 인자분석을 수행할 수 있다.

노트. 상관행렬을 이용한 인자분석과 공분산행렬을 이용한 인자분석

노트. k인자 모형의 척도불변성(scale invariance)

$\mathbf{x}$가 적재값이 Λ, 특정분산이 Ψ인 인자모형을 만족하면, $\mathbf{x}$에 대각행렬 C를 곱한 $\mathbf{y} = C\mathbf{x}$도 인자모형을 만족하고 이의 적재값이 $C\Lambda$, 특정분산이 $C\Psi$가 된다.

인자분석 VIII

노트. 상관행렬을 이용한 인자분석

$\mathbf{y}$가 관측치일 때, $\mathbf{x} = C^{-1}\mathbf{y} = (x_1/s_1, \ldots, x_q/s_q)$라고 정의하자. 여기서 $C = diag(s_1, \ldots, s_q)$, s_i는 y_i의 표준편차이다. $\mathbf{x}$가 인자모형

$$\mathbf{x} = \Lambda\mathbf{f} + \mathbf{u}$$

를 만족하면 $\mathbf{y}$ 역시 인자모형

$$\mathbf{y} = C\Lambda\mathbf{f} + C\mathbf{u}$$

를 만족하게 된다.
따라서 공분산행렬을 이용해 인자분석을 하고 상관행렬의 인자분석으로 바꿀수도 있고, 상관행렬의 인자분석을 해서 이를 공분산행렬로 바꿀 수도 있다.
이 부분은 너무 수리적이다. 메시지는 "공분산행렬을 이용해 인자분석을 하고 상관행렬의 인자분석으로 바꿀수도 있고, 상관행렬의 인자분석을 해서 이를 공분산행렬로 바꿀 수도 있다. "

노트. 척도불변성의 증명

k인자 모형은 다음과 같다.

$$\mathbf{x} = \Lambda\mathbf{f} + \mathbf{u}, \ \mathbb{V}ar(\mathbf{x}) = \Lambda\Lambda^T + \Psi$$

$\mathbf{x}$에 대각행렬 $C = diag(c_1, \ldots, c_q)$를 곱한 것을 $\mathbf{y}$라고 하면,

$$\mathbf{y} = C\mathbf{x} = C\Lambda f + C\mathbf{u}$$

가 되고, $\mathbf{y}$의 분산은

$$\mathbb{V}ar(\mathbf{y}) = C\Lambda\Lambda^T C^T + C\Psi C^T$$

인자분석 IX

가 된다. 따라서 $\mathbf{y}$의 경우, 적재값이 $C\Lambda$, 특정분산이 $C\Psi$인 인자모형을 만족하게 된다. 즉,

$$\lambda_{ij} \longrightarrow c_i \lambda_{ij}$$
$$\psi_i \longrightarrow c_i^2 \psi_i$$

가 된다.

인자분석 X

인자분석의 단계

1. 파라미터의 추정 : 최대가능도 인자분석, 주성분인자분석
2. 인자개수의 추정
3. 인자의 회전 : varimax 등
4. 인자의 해석
5. 인자 점수(factor score)의 추정

인자분석 XI

모수의 추정

1. **최대가능도 인자분석** 모수의 추정방법으로 최대가능도 방법을 쓰는 것이다.
2. **주성분 인자분석** Λ과 Ψ를 반복적으로 추정하는데, Λ를 추정할 때 주성분분석 방법을 쓴다.

노트. 주성분인자분석

다음과 같은 알고리듬으로 Λ와 Ψ를 추정한다.

Step 1. $\psi_i^{(0)}$를 $\mathbb{V}ar(x_i)$ 빼기 x_i와 $\mathbf{x}_{-i}$ 사이의 다중상관계수 혹은 $\max_{j \neq i} Corr(x_i, x_j)$로 놓는다.

Step 2. $l = 1, 2, \ldots$

Step 1..1 $S^{(l)} = S - \Psi^{(l-1)}$라 놓는다.

Step 2..2 $S^{(l)}$의 고유치값이 큰 k개의 고유벡터로 적재값을 추정한다. 즉 $\lambda_1, \ldots, \lambda_k$가 $S^{(l)}$의 고유벡터라면,

$$\Lambda = [\lambda_1, \lambda_2, \ldots, \lambda_k]$$

라 놓는다.

인자분석 XII

Step 3..3 $\psi_i^{(l)} = s_i^2 - \sum_{j=1}^{k} \lambda_{ij}^2$이라 놓는다.

$\mathbf{x}_{-i}$는 $\mathbf{x}$에서 x_i를 뺀 모든 변수를 말한다.

노트. 헤이우드 경우

위의 알고리듬을 돌리면 $\psi_i < 0$인 경우가 나오는데, 이를 헤이우드 경우(Heywood case, Heywood 1931)이라고 한다. 이 때 그냥 0으로 놓으면 되나?

노트. 참고. 다중상관계수

x_i와 $\mathbf{x}_{-i}$ 사이의 다중상관계수는 x_i와 $\mathbf{x}_{-i}$의 선형조합(linear combination)사이의 상관계수 중 최대값을 말한다. 그리고 이는 x_i를 반응변수 $\mathbf{x}_{-i}$를 예측변수로 한 회귀모형의 R^2값의 제곱근 값이다.

노트. 최대가능도인자분석

가능도함수는 $-\frac{1}{2}nF$와 x의 함수로 이루어진다. 여기서,

$$F = \log|\Lambda\Lambda^T + \Psi| + tr(S|\Lambda\Lambda^T + \Psi|^{-1}) - \log|S| - q$$

이다. 최대가능도 방법은 F를 최소화하는 Λ와 Ψ를 찾는 것이다.
몇 가지 반복적 방법이 존재한다. 이 경우도 헤이우드 경우가 생길 수 있다.

인자분석 XIII

인자 개수의 추정

인자 개수의 적재값에의 영향

인자의 개수가 작게 추정되면 적재값이 너무 커지고, 인자개수가 너무 크게 추정되면 적재값이 작게 나뉘어져 해석이 어려워진다.

추정방법

k_0를 1부터 늘려가면 $H_0 : k = k_0$를 검정하고, H_1이 유의하지 않을 때 멈춘다. 이 때의 k_0 값을 인자의 개수로 정한다.

인자분석 R 코드 I

factanal(x, factors, data = NULL, covmat = NULL, n.obs = NA, subset, na.action, start = NULL, scores = c("none", "regression", "Bartlett"), rotation = "varimax", control = NULL, ...)

```
> sapply(1:3, function(f) factanal(life, factors = f)$PVAL)
   objective    objective    objective
1.879555e-24 1.911514e-05 4.578204e-01

> factanal(life, factors = 3, method ="mle")
Call:
factanal(x = life, factors = 3, method = "mle")

Uniquenesses:
   m0   m25   m50   m75    w0   w25   w50   w75
0.005 0.362 0.066 0.288 0.005 0.011 0.020 0.146
```

```
Loadings:
    Factor1 Factor2 Factor3
m0  0.964   0.122   0.226
m25 0.646   0.169   0.438
m50 0.430   0.354   0.790
m75         0.525   0.656
w0  0.970   0.217
w25 0.764   0.556   0.310
w50 0.536   0.729   0.401
w75 0.156   0.867   0.280

                Factor1 Factor2 Factor3
SS loadings       3.375   2.082   1.640
Proportion Var    0.422   0.260   0.205
Cumulative Var    0.422   0.682   0.887

Test of the hypothesis that 3 factors are sufficient.
The chi square statistic is 6.73 on 7 degrees of freedom.
The p-value is 0.458
```

노트.

1. 첫 명령은 인자의 개수를 정하는 것이다.
   - 1.1 sapply는 벡터나 리스트의 원소에 함수를 적용하는 함수이다.
   - 1.2 여기서 함수는 인자의 개수를 인자(argument)로 받아들여 그 개수의 인자모형을 적합해서 P 값을 구하는 것이다.
   - 1.3 factanal은 인자분석을 수해하는 명령어이다. method는 언제나 mle 라고 한다.

인자분석 R 코드 II

1.4 결과는 각 인자의 개수가 1, 2, 3인 모형 3개를 적합하고 각 모형의 p 값을 출력하는 명령이다. 여기서 1:3을 넣어서 돌아갔는데, 4이상이 들어가면 에러가 발생한다. 이를 기반으로 factor의 개수가 3이라고 결론을 낸다.

2. 다음의 모형을 적합했다.

$$X = \Lambda f + u, \ u \sim N(0, \Psi).$$

여기서 X는 $k = 8$차원 랜덤벡터이고, Λ는 8×3 인자적재(factor loading)이다. 인자는 $f = (f_1, f_2, f_3)$로 나타낸다.

3. 결과에서 uniqueness는 특정분산으로 u의 분산 대각행렬 Ψ의 대각원소이다.

4. **Loadings** 는 Λ 행렬을 나타낸다. 빈칸이 있는데 0을 의미하는 것 같다.

인자분석 R 코드 III

5. **SS loadings** 는 각 인자의 공통성 h_i이다. cummulative var는 공통성과 유일성을 합했을 때 공통성의 비율이다. 즉,

$$h_1 + h_2 + h_3 + \sum_{j=1}^{k} \psi_j$$

   를 100으로 했을 때 h_i들의 비율이다.

6. 마지막 부분은 가설검정인데 H_0 : 적합한 모형이 옳다. 에 대한 가설이다. p-value가 작으면 이 모형이 적당하지 않다는 강한 증거이다. 여기서는 0.458이므로 이 모형이 틀리다는 뚜렷한 증거는 없는 것이므로, 이 모형이 적당하다는 뜻이다.

인자의 해석을 위한 방법 I

인자의 회전

인자적재 Λ를 $k \times k$ 직교행렬 M으로 바꾸어도 자료는 동일하게 설명된다. 즉, 임의의 직교행렬 M에 대해 적재행렬과 특정분산 $(\Lambda^* = \Lambda M, \Psi)$를 가진 인자모형은 (Λ, Ψ)를 가진 인자모형과 동일하게 자료를 설명할 수 있다.

노트. 인자의 회전의 증명

$$\mathbf{x} = \Lambda \mathbf{f} + \mathbf{u}$$

의 인자모형이 성립한다고 하자. 그리고 M은 $k \times k$ 직교행렬이라고 하자. 위의 식을

$$\begin{aligned} \mathbf{x} &= (\Lambda M)(M^T \mathbf{f}) + \mathbf{u} \\ &= \Lambda^* \mathbf{f}^* + u \end{aligned}$$

와 같이 나타낼 수 있다. 여기서 $\Lambda^* = \Lambda M$, $\mathbf{f}^* = M^T \mathbf{f}$이다.

인자의 해석을 위한 방법 II

x의 분산도

$$\Sigma = \Lambda\Lambda^T + \Psi = \Lambda M M^T \Lambda^T + \Psi = \Lambda^* \Lambda^{*T} + \Psi$$

와 같이 동일하게 나타낼 수 있다.

제약

Λ와 Ψ의 추정량을 유일하게 존재하게 하기 위해 보통 제약을 주는데 첫번째 인자가 가장 많이 자료를 설명하고 두번째 인자가 두번째로 많이 설명하도록 제한한다. 이것은

$$G = \Lambda \Psi^{-1} \Lambda$$

가 대각행렬이 되도록 제한하는 것과 같다.
보다 좋은 해석을 위해서 인자를 회전시키기도 한다.

인자의 해석을 위한 방법 III

Thurstone (1931)의 간단한 구조(simple structure)

Thurstone이 인자적재가 만족하면 해석이 쉬워지는 조건을 제안했다.

1. 적재행렬의 모든 로우는 0을 포함한다.
2. 적재행렬의 모든 컬럼은 적어도 k개의 0을 포함한다.
3. 임의의 두 개의 컬럼을 뽑았을 대 한쪽에선 모두 0 다른 한쪽에선 0이 아닌 로우들이 여러 개 있어야 한다.
4. 인자의 개수가 4개이상이면 임의의 두 개의 컬럼을 뽑았을 때, 모두 0인 로우들이 많이 있어야 한다.
5. 임의의 두 개의 컬럼을 뽑으면 둘 다 0이 아닌 로우는 매우 적은 수이어야 한다.

인자의 해석을 위한 방법 IV

회전의 방법

1. 직교회전의 방법
   - 1.1 varimax rotation (Kaiser 1958). 적재값을 몇 개의 큰 값과 많은 0 근처의 값으로 만들고자 함.
   - 1.2 quartimax rotation (Carroll 1953). 한 개의 x_i가 한 개의 f_j와 크게 상관이 있고, 나머지 인자는 상관이 없도록 함.
2. 사각회전의 방법
   - 2.1 oblimin 회전 (Jennrich and Sampson 1966). 인자간의 상관관계의 정도를 조절하는 파라미터를 통해 간단한 구조를 찾으려 함. 이 파라미터를 정하는 것이 쉽지 않다.
   - 2.2 promax 회전. 직교회전의 해에 승을 올리고 이를 통해 간단한 구조를 찾으려 한다.

노트.

1. 사각 : 90도의 배가 되지 않는 각
2. 인자의 해석을 위해 사각회전을 인정하기도 한다.

인자의 해석을 위한 방법 V

인자(점수)의 추정: 회귀분석 방법

정규 가정하에서

$$\mathbf{f}|\mathbf{x} \sim N(\Lambda^T \Sigma^{-1}\mathbf{x}, (\Lambda^T \Psi^{-1}\Lambda + I)^{-1})$$

이므로

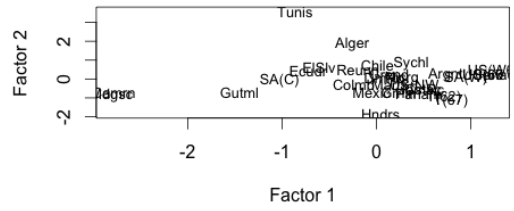
$$\hat{\mathbf{f}} = \Lambda^T \Sigma^{-1}\mathbf{x}$$

으로 **f**를 추정한다. 다른방법으로 ML method가 있다. "option=Bartlett"

인자점수의 추정이 필요한 이유

인자는 관측자료의 parsimonious summary이고 이는 이후의 분석에 사용될 수 있다.

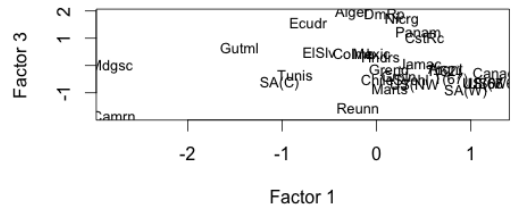
R 코드 I

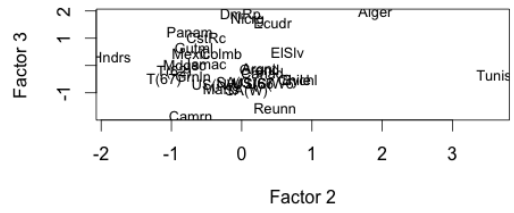



```
scores <- factanal(life, factors = 3, method = "mle", scores = "regressior
```

scores = "regression")$scores

```
cex <- 0.8
plot(scores[,1], scores[,2], type = "n", xlab = "Factor 1", ylab = "Factor
text(scores[,1], scores[,2], abbreviate(rownames(life), 5), cex = cex)
plot(scores[,1], scores[,3], type = "n", xlab = "Factor 1", ylab = "Factor
text(scores[,1], scores[,3], abbreviate(rownames(life), 5), cex = cex)
plot(scores[,2], scores[,3], type = "n", xlab = "Factor 2", ylab = "Factor
text(scores[,2], scores[,3], abbreviate(rownames(life), 5), cex = cex)
```





인자분석과 주성분이 차이

인자분석과 주성분분석 모두 다변량 자료를 작은 차원으로 설명하려는 시도이다.
두 분석의 차이는 다음과 같다.

1. 요약된 변수의 차원을 m에서 $m+1$로 한 차원 늘리면, 처음 m개의 주성분은 변하지 않지만, 인자는 변한다.
2. 주성분점수는 계산이 쉽지만 인자점수는 계산이 쉽지 않아 여러방법이 존재한다.
3. 주성분분석은 자료의 작은 차원으로의 "근사" 인 반면 인자분석은자료에 대한 작은 차원의 "정확한 모형" (자료의 공분산 행렬 대한 모형)에 기반한 분석이다.

참고문헌 I

1. Brian Everitt and Torsten Hothorn. *An introduction to applied multivariate analysis with R*.
   Springer, 2011.