8.1 8.2 다중 인자모형

2019년 6월 16일 일요일 오전 3:06

공분산 구조를 몇개의 관측 불가능한 인자로 설명.

변수들 간에 내재하고 있는 공통의 구조 파악.

데이터의 특성을 몇개의 인자로 축약하여 설명.

여러 개 변수들의 상관성 구조를 나타내는 몇개의 인자로 분석.

변수들에 내재되어 있는 공통의 인자로서 데이터를 설명.

분석에 필요한 변수의 차원을 줄일 수 있음.

주성분 분석 : 주성분은 관측된 변수들의 선형결합식으로 정의됨.

인자 분석 : 변수들이 공통인자들의 선형결합식으로 정의됨.

모든 변수들이 동시에 고려되어 인자를 찾게됨.

각 인자는 다른 변수들과의 관계로 설명됨.

인자는 변수들을 최대한으로 설명할 수 있도록 구함.

각 인자는 전체 관측치의 함수 형태로 전체 변수에 의존적인 변수로 볼 수 있음.

$p \times 1$ 확률벡터 X 가 모평균벡터 μ 와 모공분산행렬 Σ 를 가진다.

또한 X 는 m개의 공통인자 $F_1 \cdots F_m$ 과 특수인자 $\varepsilon_1 \cdots \varepsilon_p$ 의 선형결합으로 표현된다고 가정.

$$X_p - \mu_p = l_{p1}F_1 + l_{p2}F_2 + \cdots + l_{pm}F_m + \varepsilon_p$$

여기서 l_{ij} 는 i번째 변수의 j번째 인자의 적재값.

$F_1 \cdots F_m$ 와 $\varepsilon_1 \cdots \varepsilon_p$ 는 $p+m$ 개의 관측불가능한 확률변수.

m개의 공통인자를 포함한 인자모형 ⇒ m - 인자모형

8.2.1 직교인자모형

m-인자모형은 변수들에 잠재하는 공통인자들과 유일한 부분을 설명하는 특수인자들의 선형결합으로 표현됨.

$$X - \mu = LF + \varepsilon$$

$$\begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix} - \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_p \end{pmatrix} = \begin{pmatrix} l_{11} & l_{12} & \cdots & l_{1m} \\ l_{21} & l_{22} & \cdots & l_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ l_{p1} & l_{p2} & \cdots & l_{pm} \end{pmatrix} \begin{pmatrix} F_1 \\ \vdots \\ F_m \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_p \end{pmatrix}$$

μ_i : X_i 의 기대값

ε_i : i번째 특수인자

F_j : j 번째 공통인자

l_{ij} : i번째 변수의 j번째 인자의 적재값

L : $\{l_{ij}\}$ 인자적재행렬

직교인자모형 가정

$E(F) = 0_{m \times 1}$

$Cov(F) = E(FF') = I_{m \times m}$

$E(\varepsilon) = 0_{p \times 1}$

$Cov(\varepsilon) = E(\varepsilon\varepsilon') = \Psi_{p \times p} = \begin{pmatrix} \psi_1 & & 0 \\ & \psi_2 & \\ 0 & & \ddots \ \psi_p \end{pmatrix}$ ⇒ 원래 변수 X_i들 간의 공분산은 공통인자에만 의존한다고 간주함.

$Cov(\varepsilon, F) = E(\varepsilon F') = 0_{p \times m}$ ⇒ F 와 ε 는 서로 독립. (인자들이 서로 직교)

8.2.2 인자모형에서 공분산행렬의 분해

변수들의 공분산행렬을 구하면 인자적재행렬과 특수인자의 공분산행렬로 인자분해됨.

$$(X-\mu)(X-\mu)' = (LF+\varepsilon)(LF+\varepsilon)' = (LF)(LF)' + \varepsilon(LF)' + LF\varepsilon' + \varepsilon\varepsilon'$$

$$\Sigma = Cov(X) = E(X-\mu)(X-\mu)' = LE(FF')L' + E(\varepsilon F')L' + LE(F\varepsilon') + E(\varepsilon\varepsilon')$$
$$= LL' + \psi$$

변수와 공통인자간의 공분산을 구하면 인자적재행렬로 나타내어짐

$$Cov(X, F) = E[(X-\mu)F'] = LE(FF') + E(\varepsilon F') = L$$

원래 변수의 분산은 공통인자들에의한 공통분산과 특수인자에 의한 특수분산으로 분해되어 표현됨.

$$\sigma_{ii} = Var(X_i) = l_{i1}^2 + l_{i2}^2 + \cdots + l_{im}^2 + \psi_i$$
$$= \underbrace{h_{ii}^2}_{(공통성)} + \psi_i = 공통분산 + 특수분산$$

변수들간의 공분산은 인자적재값의 곱들의 합으로 표현됨.

$$Cov(X_i, X_k) = l_{i1} l_{k1} + l_{i2} l_{k2} + \cdots + l_{im} l_{km}$$

변수와 공통인자간의 공분산은 공통인자의 인자적재값으로 표현됨.

$$Cov(X_i, F_j) = l_{ij}$$

8.2.3 인자적재행렬의 비유일성

인자의 개수가 1보다 많을 경우 인자적재행렬은 mxm 직교행렬 T를 이용하여 표현 가능.

직교행렬 T 는 $TT' = T'T = I$ 를 만족하므로

$$X - \mu = LF + \varepsilon = LTT'F + \varepsilon = L^*F^* + \varepsilon$$

$$\Sigma = LL' + \psi = LTT'L' + \psi = L^*L^{*'} + \psi$$

$L^* = LT$

$F^* = T'F$

$E(F^*) = T'E(F) = 0$

$Cov(F^*) = T'Cov(F)T = T'T = I_{m \times m}$

=> 직교행렬 T의 형태에 따라 인자적재행렬은 변할 수 있으므로 인자적재행렬은 항상 유일하지는 않음

8.2.4 인자모형의 척도불변성

확률벡터 X에 정칙행렬 A를 이용하여 선형변환.

$$Z = AX + b$$

$$Z - \mu_z = AX - A\mu = ALF + A\varepsilon = L_zF + \varepsilon_z$$

$$Cov(Z) = \Sigma_z = A\Sigma A' = ALL'A' + A\psi A' = L_zL_z' + \psi_z$$

$\mu_z = A\mu + b$

$L_z = AL$

$\psi_z = A\psi$

=> 변수에 적용된 선형변환식이 인자적재행렬에도 적용됨.

분산행렬이 $D_\sigma = diag(\sigma_1, \sigma_2, \cdots \sigma_p)$ 인 경우 특히 표준화 변수 $D_\sigma^{-1}(X-\mu)$ 를 이용한 인자모형을 세우면 상관행렬 ρ 는 다음과 같이 표현됨.

$$\rho = L_zL_z' + \psi_z$$

8.2.5 인자모형의 기하학적 표현

인자를 축으로 변수를 벡터로 표현.

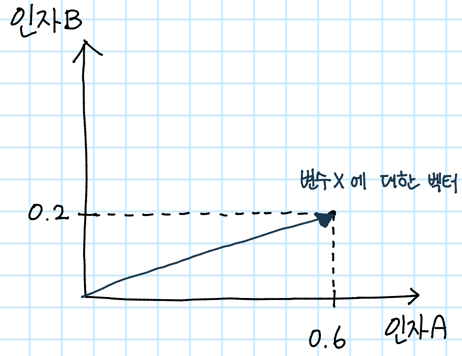
변수가 어떤 인자축에 가까이 놓일수록 변수와 인자간의 관련성이 크다.

벡터로 표현되는 변수의 방향과 길이는 인자패턴에 의존한다.

변수벡터의 길이는 공통인자에 의한 분산 (공통성) 을 나타낸다.

서로 독립인 인자축이 서로 90도의 각을 형성하면 직교한다고 한다.

또한 인자축의 비직교회전을 통해 직교하지 않는 인자축을 형성할 수도 있다.



8.3 인자적재와 공통성에 대한 추정방법

2019년 6월 16일 일요일 오전 3:07

인자분석의 목적은 공통인자를 정하고 인자적재행렬을 구하여 해석하는데 있다.

인자적재행렬을 추정하는 방법 3가지

1) 주성분법
2) 주축인자법
3) 최대우도법

8.3.1 주성분법

인자적재행렬은 공분산행렬의 분해로부터 얻을 수 있으므로 스펙트럼분해 방법을 통한 주성분을 구하는 방법으로 접근할 수 있다.

또한 상관행렬에 대해서도 마찬가지로 적용할 수 있다.

공분산행렬 스펙트럼 분해

$$\Sigma = \lambda_1 e_1 e_1' + \lambda_2 e_2 e_2' + \cdots + \lambda_p e_p e_p' = \left[\sqrt{\lambda_1}\, e_1 \,|\, \sqrt{\lambda_2}\, e_2 \,|\, \cdots \,|\, \sqrt{\lambda_p}\, e_p\right] \begin{bmatrix} \sqrt{\lambda_1}\, e_1' \\ \sqrt{\lambda_2}\, e_2' \\ \vdots \\ \sqrt{\lambda_p}\, e_p' \end{bmatrix}$$

λ_i 는 Σ 의 고유값. e_i 는 Σ 의 고유벡터 $\Rightarrow \Sigma e_i = \lambda_i e_i$

만약 $m=p$ 이면, 즉 $\psi_i = 0$ 이면

$\underset{p\times p}{\Sigma} = LL' + 0 = LL'$ 이 되어 인자적재값과 주성분계수의 값이 일치함.

그러나 인자분석의 목적은 원래변수의 개수보다 적은 몇개의 공통인자로 공분산구조를 설명하고자 하는것.

$$\Sigma \approx \lambda_1 e_1 e_1' + \lambda_2 e_2 e_2' + \cdots + \lambda_m e_m e_m' = \left[\sqrt{\lambda_1} e_1 \,|\, \sqrt{\lambda_2} e_2 \,|\, \cdots \,|\, \sqrt{\lambda_m} e_m\right] \left[\quad\right] = LL'$$

특수성을 추정하려면 공분산행렬이 $\Sigma = LL' + \psi$ 가 되므로

$\psi_i = a_{ii} - \sum_{j=1}^{m} l_{ij}^2$ 으로 특수성을 추정할 수 있음.

표본공분산행렬 S를 이용하는 주성분법에 의한 인자적재행렬 추정방법.

S의 고유값 $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p \geq 0$ 을 갖고 각 고유값에 해당하는 고유벡터 e_i 를 갖는다.

$S = \tilde{L}\tilde{L}' + \tilde{\psi}$ $\quad \tilde{L} = \left[\sqrt{\hat{\lambda}_1} e_1 \,|\, \cdots \,|\, \sqrt{\hat{\lambda}_m} e_m\right]$, $\quad \tilde{\psi} = diag(\tilde{\psi}_1, \cdots \tilde{\psi}_p)$

특수성은 $\tilde{\psi}_i = s_{ii} - \tilde{h}_i^2$ 으로 추정.

공통성은 $\tilde{h}_i^2 = \tilde{l}_{i1}^2 + \cdots + \tilde{l}_{im}^2$ 으로 구함.

i번째 인자에 의해 설명되는 분산의 비율 $\dfrac{\hat{\lambda}_i}{s_{11} + \cdots + s_{pp}}$ (표본상관행렬 R 사용 : $\dfrac{\hat{\lambda}_i}{p}$)

8.3.2 주축인자법

주성분법을 이용하여 구한 초기 인자적재행렬로부터 수정된 인자적재행렬을 구하는 방법

주성분법에 의해 공통성 또는 특수성을 추정한 후 이에 대한 정보를 이용하여 다시 인자적재행렬을 구함.

pxp 인 공분산행렬, 상관행렬로부터 주성분법의 결과를 이용하여 조정된 공분산행렬, 조정된 상관행렬에 대해 다음의 근사적인 관계식 쓸 수 있음.

$S - \hat{\psi} \approx LL'$, $R - \hat{\psi} \approx LL'$ ※ L 은 pxm 인 인자적재행렬

$S - \hat{\psi}$ 의 대각선상의 성분 = 공통성 $\tilde{h}_i^2 = s_{ii} - \tilde{\psi}_i$ (초기추정량 $\tilde{h}_i^2 = s_{ii} - \dfrac{1}{s^{ii}} = s_{ii} \cdot R_i^2$)

$R - \hat{\psi}$ 의 대각선상의 성분 = 공통성 $\tilde{h}_i^2 = 1 - \tilde{\psi}_i$ (초기추정량 $\tilde{h}_i^2 = R_i^2 = 1 - \dfrac{1}{r^{ii}}$)

s^{ii} : S^{-1}의 i번째 대각선상의 성분

r^{ii} : R^{-1}의 i번째 대각선상의 성분

R_i^2 : X_i 와 나머지 $(p-1)$개의 변수들과의 다중상관계수제곱

조정된 공분산행렬 또는 조정된 상관행렬을 이용하여 다시 주성분법을 적용하여 공통성과 특수성을 구하고
이것을 새로운 초기값으로 하고 조정한 후 다시 주성분법을 적용한다.
이와 같은 과정을 공통성의 추정치가 수렴할 때까지 반복한다.

8.3.3 최대우도법

공통인자 F와 특수인자 이 정규분포를 따른다고 가정하면 우도함수를 구할 수 있으므로
인자적재행렬에 대해 최대우도추정량을 구할 수 있다.

n개의 확률표본에 대한 우도함수는 L과 F에 의존하여 표현됨. (우도함수식 : p202)
계산의 편리성을 위해 L에게 유일성 조건을 부여하여 최대우도를 만족시키는
L과 ψ 의 최대우도추정량 $\hat{L}$, $\hat{\psi}$ 을 수치해석적인 알고리즘을 이용하여 반복적
계산을 통해 구할 수 있음.

최대우도방법의 장점 1 ⇒ $\hat{L}$ 이 척도 불변

변수 X_i 가 $c_i X_i$ 로 변화되었을때 원래 인자적재값 $\hat{l}_{i1} \cdots \hat{l}_{im}$ 에도 마찬가지로 $c_i \hat{l}_{i1} \cdots c_i \hat{l}_{im}$ 으로 변환됨.

표본상관행렬 R을 이용하여 인자분석을 할 경우, 각 변수 X_i 에 $1/\sqrt{s_{ii}}$ 를 이용한 척도변환으로 생각하면 되므로
표본공부산행렬 S에 대해 최대우도법을 적용하여 구한 인자적재값에 같은 척도변환을 해주면된다.
그러나 주성분을 이용해 구한 인자적재값에는 성립하지 않는다.

최대우도방법의 장점 2 => m-인자모형의 적합성에 대한 통계적 검정을 할 수 있는것으로 적절한 인자개수 선택에 이용할 수 있다.

8.3.4 heywood 상황

인자에 대한 공통분산은 다중상관계수의 제곱이 되므로 0에서 1사이가 됨.

그러나 최대우도법을 이용할 경우 공통성에 대한 초기추정값은 다중상관계수제곱을 취하지만

반복적 수치적 계산과정에 따르다보면 공통성이 1이거나 1을 초과하는 현상이 나타남.

공통성이 1인 경우 = heywood 상황 / 공통성이 1을 초과한 경우 = ultra-heywood 상황

이런 경우는 특수분산의 추정값이 0 또는 음수가 되는 경우로 인자분석의 결과가 유효하지 않게됨.

(이러한 상황이 발생하는 원인은 p203)

최대우도추정법은 다른 방법에 비해 heywood 상황이 발생하기 쉬운편.

반복적인 계산과정에서 높은 공통성을 가진 변수에게 더 높은 비중이 부여되고

이것은 다시 공통성을 증가시키는 결과를 가져오기 때문.

8.4 인자개수 결정

2019년 6월 16일 일요일 오전 3:07

원래 데이터를 내재하는 공통인자들로 나타낼때 적절한 개수의 인자를 선택해야함. 설명에 필요한 인자의 개수를 결정하는 기준.

1) 총 변이에의 공헌도 (percentage of total variance)

공통인자들에 의해 설명되는 분산의 비율이 전체 변이의 70-90%가 되도록 공통인자의 개수를 결정함.

2) 평균 고유값 (average eigenvalues)

고유값들의 평균을 구한 후 고유값이 평균값 이상이 되는 주성분을 보유함.

상관행렬을 사용한 경우 평균 고유값은 1이 된다.

3) 스크리 그래프

2차원 좌표축에 (고유값 순서, 고유값 크기)로 점을 찍고 점간을 선분으로 연결함.

값이 큰 고유값부터 크기순으로 점이 찍히며 값의 차이가 크면 가파른 경사를 나타내고 고유값의 변화가 작으면 내리막 경사가 완만해짐

큰 고유값과 작은 고유값을 구분하는 방법으로, 가파른 정도를 보고 가파른 부분에 해당하는 고유값까지로 인자의 개수를 정함.

4) 적절한 인자의 개수를 정하고 다음의 검정을 통해 인자의 개수를 선택할 수 있음

관심있는 귀무가설은 인자의 개수가 m개를 만족하는 모형.

데이터에 대해 다변량 정규분포 가정을 할 수 있는 경우, 위의 가설검정 문제에 대해 우도비를 계산함.

귀무가설하에서 최대우도추정량과 전체가설하에서 최대우도추정량을 구한후 우도비를 이용함. (p205)

검정통계량이 귀무가설하에서 근사적으로 카이제곱분포를 따른다는것을 이용해 카이제곱검정을 함. (검정통계량 식은 p205)

유의수준 α에서 검정법 : $T_m \ge \chi^2(\alpha)$ 이면 H_0 기각

귀무가설 기각되지 않으면 $\hat{L}\hat{L}' + \hat{\Psi}$ 은 적절한 인자모형이 되므로 모형에서 가정한 인자의 개수로 인자모형을 설정하고 해석함.

귀무가설이 기각되면 $\hat{L}\hat{L}' + \hat{\Psi}$ 은 부적절한 인자모형이 되므로 인자의 개수를 늘려 다시 검정해야함.

8.5 인자적재의 회전

2019년 6월 16일 일요일 오전 3:07

초기 인자적재행렬에 대해 직교회전을 하더라도 인자적재행렬은 공분산 (또는 상관) 구조에 대해 같은 양을 설명함.

그러므로 적절한 변환을 통해 인자에 대한 해석이 용이한 단순한 구조를 만들 필요가 있으며 이러한 구조를 얻기 위해 회전변환을 시도할 수 있음.

초기 인자적재행렬의 해석상 어려움이 발생할 경우가 있음.

예를들어, 한 변수가 여러개의 인자에 큰 적재값을 갖게 되면 여러개의 인자가 한 변수와 큰 연관성을 갖게 되어 인자에 대한 해석을 복잡하게 만듦.

이때 해석의 편리를 위해 인자회전을 시도하여 단순한 구조가 되도록 변환할 수 있음.

단순한 구조란 변수들이 한 인자에 대표적으로 큰 적재값을 갖게 되어 그 인자를 설명하는데 비중있는 변수가 되도록 만드는것.

회전변환의 기하학적 목적은 인자적재값이 인자축에 놓이게 하는 것.

인자의 축에 놓이는 변수, 즉 그 인자에서만 비중이 높은 변수들을 파악해 인자를 해석할 수 있게됨.

그러나 회전변환을 하더라도 항상 단순한 구조를 얻는 것은 아니므로 여러변환을 시도해보고 인자의 구조를 파악해야함.

8.5.1 직교회전 : 베리맥스 회전 (varimax rotation)

인자적재값들의 제곱을 취해 이들의 분산을 최대화하여 인자적재값들의 차이가 많이 나도록 하는 것.

V를 최대화하는 직교행렬 T를 찾아 변환하는 방법. (자세한 방법은 p208)

8.5.2 비직교회전

공통인자들이 서로 독립이라고 가정하면 직교회전이 적절함.

해석의 편리를 위해 비직교회전을 종종 사용함.

변수들을 겹치지 않는 그룹으로 분류할 수 있다고 가정하면 축이 서로 직교하는 직교회전을 통해 결과를 얻는 것이 좋음.

사각회전은 회전 후 축이 서로 직교하지 않으며 그룹을 통과하는 축으로 형성됨.

사각회전을 통해 비록 회전된 축이 서로 직교하지는 않으나 축 위에 인자적재값들이 놓이게 될 수 있어 해석상 이점을 얻을 수 있음.

사각회전에서 변환행렬 Q는 비직교행렬이며 비정칙행렬이다.

비직교행렬 Q에 의해 변환된 인자 F^*는 독립이 아니다.

$F^* = Q'F$, $L^* = LQ$

$Cov(F^*) = Q'IQ = Q'Q \neq I$

8.6 인자점수

2019년 6월 16일 일요일 오전 3:08

다중인자모형에서 $F_1, F_2, \cdots, F_m$ 은 관측 불가능한 인자이다.

그러나 공통인자를 파악한 후 이들에 대한 점수를 얻고자 할 때, 관측 불가능한 공통인자에 대한 추정으로 인자점수를 구함.

구하는 방법 1) 가중최소제곱법 2) 회귀적 방법

j번째 개체에 대하여 m개 공통인자에 대한 벡터 $f_j = (f_{j1}, \cdots f_{jm})$

이에 대한 추정값을 인자점수라 하고 $\hat{f}_j = (\hat{f}_{j1}, \cdots \hat{f}_{jm})$ 로 표현함.

8.6.1 가중최소제곱법

인자모형 $X - \mu = LF + \varepsilon$ 에 대한 적합으로 인자적재행렬 L과 평균벡터 그리고 $Var(\varepsilon_i) = \psi_i$ 를 구한 후 이들의 추정값을 이용하여 인자점수를 추정할 수 있음.

특수인자의 분산이 서로 다른것에 대한 가중값을 고려한 가중최소제곱법 (p210)

인자점수와 인자가 높은 상관관계를 갖게 되며 추정된 인자점수는 불편추정량이다.

8.6.2 회귀적 방법

인자들을 변수들의 다중회귀식으로 놓고 인자점수를 추정하는 방법으로 회귀적 방법이 있음.

최소제곱추정량을 구하고 인자들과 변수들이 다변량 정규분포를 따를 경우 조건부 기대값을 구함. (p211)

n개의 인자점수벡터가 각 행을 구성하는 행렬 $\hat{f}$ (p212)

8.7 인자분석 전략

2019년 6월 16일 일요일 오전 3:08

(1) 주성분을 이용한 인자분석을 한다.

인자분석의 첫 단계로 적절함.

분석 결과로부터 인자점수를 통해 이상치를 가려내고 인자의 해석을 위해 varimax 직교변환을 모색함.

(2) 최대우도를 이용한 인자분석을 한다.

다변량 정규분포를 가정할 수 있다면 최대우도를 이용한 인자분석을 수행한 후 인자의 해석을 위해 varimax 직교변환을 모색함.

(3) 주성분을 이용한 인자분석 결과와 최대우도를 이용한 결과를 비교한다.

인자적재의 형태가 비슷한가 판단하고 각 개체에 대해 두 방법에 의한 인자점수를 그래프로 나타내어 비교함.

(4) 적절한 공통인자 m개에 대해 (1) (2) (3) 과정을 반복한다.

(5) 데이터의 크기가 큰 경우 데이터를 두 부분으로 나누어 각 부분에서 따로 인자분석을 시행하고 결과를 비교해 인자분석결과의 안정성을 점검한다.

(6) 인자의 해석

인자적재행렬의 열을 살펴보아 어떤 변수가 어떤 인자에서 높은 비중을 차지하는지 봄.

각 인자에서 큰 적재값을 갖는 변수, 즉 높은 비중을 차지하는 변수들을 중심으로 인자에 대한 이름을 붙이고 해석함.

인자분석은 발현된 관측값을 내재하는 인자들로 설명할 수 있도록 도와주며 이런 면에서 인자에 대한 해석은 주관적인 해석이 뒤따름.

<예제 8.3> 신체 측정 데이터에 대해 상관행렬을 이용하여 인자분석을 하여 인자적재값, 공통성과 특수성을 구하여 정리함. (P214)

인자 1 = 길이부분에 더 많이 내재하는 인자로 '길이 인자' 로 볼 수 있음.

인자 2 = 무게와 둘레에 기여하는 부분이 큰 '무게인자' 로 볼 수 있음.

공통인자에 의해 설명되는 분산인 공통성은 각 변수에 대해 2개 인자의 인자적재값 제곱의 합으로 계산됨.

특수인자에 의해 설명되는 분산은 $\psi_i = Var(x_i) - h_i^2 = 1 - h_i^2$ 으로 계산됨.

여기서 상관행렬을 이용했으므로 $Var(X_i) = 1$ 이다.