데이터마이닝

- 차원축소기법 -

1 주성분분석 및 요인분석

본 챕터의 모든 코드를 오류없이 실행시키기 위해서는 다음의 패키지가 필요함.

```
install.packages(c('mvtnorm','ade4','psych','nFactor'))
```

참고 본 강의자료는

- "R을 이용한 데이터마이닝 (개정판), 박창이, 김용대, 김진석, 송종우, 최호식, 2015, 교우사"
- "An Introduction to Statistical Learning with Applications in R, 7th ed., Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani, 2013, Springer"

를 일부 참고하여 작성되었음.

주성분분석 및 요인분석은 출력변수(반응변수, 목적변수)의 존재를 가정하지 않는다는 측면에서 비지도학습(unsupervised learning)의 범주에 속한다 할 수 있으며, 분석대상이 되는 변수의 수를 줄일 수 있다는 측면에서는 차원축소기법의 일종으로 볼 수 있다.

1.1 주성분분석 (Principal component anaylsis)

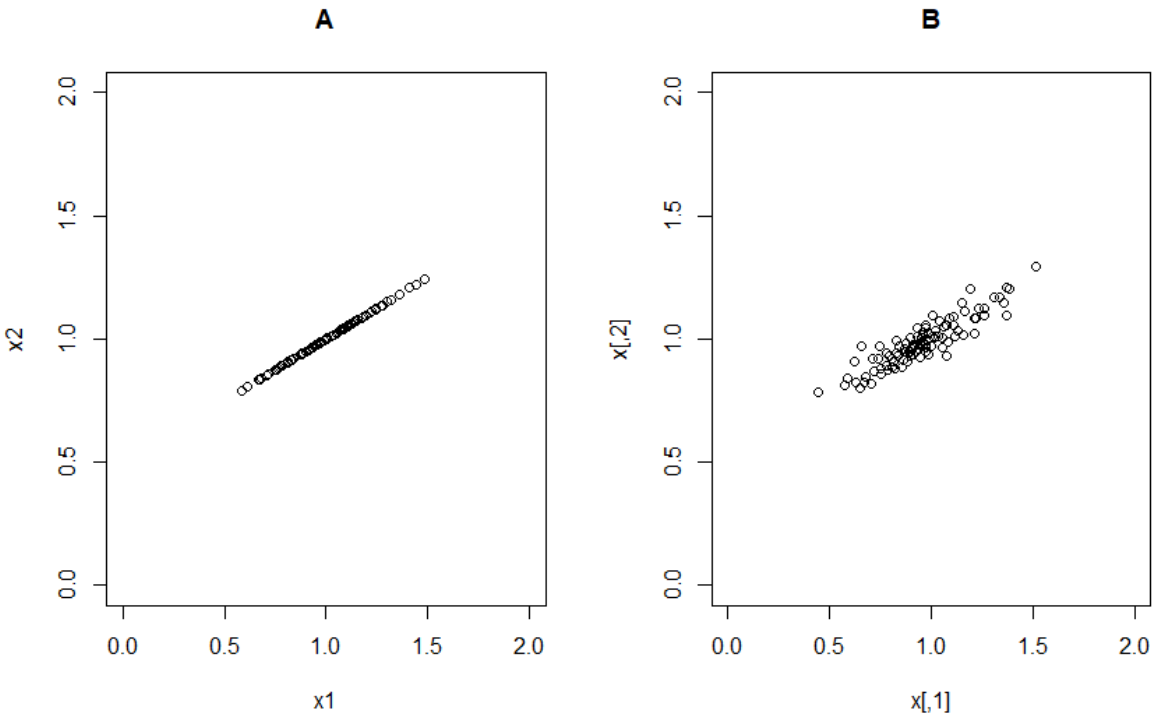
1.1.1 Motivating example

다음과 같은 이변량($p = 2$) 자료를 생각해 보자.

예제 차원축소예시

```
n <- 100
x1 <- rnorm(n,1,0.2)
x2 <- 0.5 + 0.5*x1
par(mfrow=c(1,2))
plot(x1,x2,xlim=c(0,2),ylim=c(0,2),main="A")
```

```r
library(MASS)
# 이변량 정규분포에서 자료생성
mu <- c(1,1)
sigma <- matrix(c(0.04,0.018,0.018,0.01),2,2)
x <- mvrnorm(n,mu,sigma)
plot(x,xlim=c(0,2),ylim=c(0,2),main="B")
```



- 그림 "A"에서는 $x_2 = 0.5 + 0.5x_1$이 성립하므로, 이 경우 데이터를 적절히 변환하면 하나의 변수로 완벽히 표현이 가능하다.
- $(x_1, x_2) \to \left(\dfrac{2(x_1 - 1) + (x_2 - 1)}{\sqrt{5}}, 0 \right) = (z_1, z_2)$
- 그림 "B"에서는 두 변수 사이에 완전한 함수관계가 성립하지는 않지만, 적절한 변환을 거쳐 서로 직교하는 새로운 좌표축 (z_1, z_2)로 나타낼 수 있다.
- 만약 z_1에 비해 z_2의 정보량이 상대적으로 매우 작다면, z_2를 배제함으로써 약간의 정보손실을 감수하는 대신 차원을 2에서 1로 축소할 수 있게 된다.

1.1.2 주성분분석

기본 개념

주성분분석은 직교선형변환을 통해 기존 자료로부터 주성분이라 불리우는 새로운 변수들의 관측값들을 생성해 내는 과정을 말한다.

원자료는 상관성이 존재하는 변수들로부터 관측되었더라도, 주성분들은 직교변환을 통해 얻어지기 때문에 상관관계가 없게 된다.

어떤 방식으로 변환할 것인가? → 분산이 가장 큰 방향을 순차적으로 찾아나가는 방식으로 변환

주성분의 정의

X를 데이터행렬이라 하자. 즉, X의 각 행이 p차원의 관측치 x_i $(i = 1, 2, \ldots, n)$로 구성되어 있다. 즉, X는 $n \times p$ 행렬이고,이 행렬의 k번째 열은 k번째 변수에 대한 관측치들을 의미한다.

편의상 X는 중심화된 것으로 간주한다. 즉, 각 관측치 벡터 x_i 로부터 표본평균 벡터 $\bar{x} = n^{-1} \sum_{i=1}^{n} x_i$ 를 뺀 것으로 생각하면 된다.

1. 제 1 주성분 v_1은 다음과 같이 정의된다.

$$v_1 = \underset{\|w\|=1}{\arg\max} \sum_{i=1}^{n} (x_i^\top w)^2 = \underset{\|w\|=1}{\arg\max}\ w^\top (X^\top X) w$$

   즉, 중심화된 데이터를 변환하였을 때 분산이 최대가 되도록 하는 방향을 제 1 주성분으로 선택하게 된다.

2. 제 2 주성분은 제 1 주성분과 직교하면서 가장 분산을 최대화 하는 방향으로 선택한다.

3. 위 과정을 반복하면 총 p개의 직교하는 새로운 좌표축에 의해 원 데이터를 재 표현할 수 있다.

주성분의 탐색

주성분의 탐색 및 결정은 데이터 행렬의 공분산 구조에 의존하게 된다. 즉, $X^\top X$로부터 주성분을 추출해 낼 수 있다.

1. 제 1 주성분 v_1은 $X^\top X$의 최대 고유값(eigenvalue) λ_1에 대응되는 고유벡터(eigenvector)이다.

2. 제 k 주성분은 k번째로 큰 고유값 λ_k 에 대응되는 고유벡터 v_k이다.

주성분 점수

주성분에 의해 재 표현된 관측치 값들을 주성분 점수(principal component score)라 한다.

적재계수

적재계수(loading)란 각 변수들이 주성분에 기여하는 정도를 나타내는 것으로 고유벡터의 성분값들이 곧 적재계수가 된다.

주성분의 분산

주성분은 가장 분산이 큰 방향부터 순차적으로 결정되므로, 분산은 제 1 주성분에서 가장 크고 점차 감소하게 된다.

$$\sum_{i=1}^{n}(x_i^\top v_1)^2 \geq \sum_{i=1}^{n}(x_i^\top v_2)^2 \geq \cdots \geq \sum_{i=1}^{n}(x_i^\top v_p)^2 \geq 0$$

또한, k번째 주성분에서의 분산은 고유치와 일치한다.

$$\sum_{i=1}^{n}(x_i^\top v_k)^2 = \lambda_k$$

주성분 개수의 선택

1. p차원의 자료에 대해서 직교선형변환을 하여 새로운 p차원의 자료를 얻어냈을 때 자료의 총 변동에는 변함이 없다.

$$\text{원자료에서의 변수들의 총 변동성} = tr(X^\top X) = \sum_{k=1}^{p}\lambda_k$$

2. q번째 주성분까지의 누적 변동성은 q번째 고유값까지의 합이다.

$$\sum_{k=1}^{q}\lambda_k$$

3. 즉, q번째 주성분까지만 선택하여 차원축소를 할 경우 보존되는 정보량은

$$\sum_{k=1}^{q}\lambda_k \Big/ \sum_{k=1}^{p}\lambda_k$$

   와 같다.

4. 보통 총 변동의 80%~90%를 보존할 수 있도록 주성분의 개수를 선택하나 (Scree plot), 적용분야 및 데이터의 특성 등에 따라 달라질 수 있다.

5. 축소된 차원에서의 주성분(혹은 주성분점수)들을 또 다른 분석(회귀분석,군집분석 등)에 활용하기도 하며, 이러한 경우 주성분 개수의 선택은 그 활용 목적에 따라 결정될 수 있는 여지가 있다.

참고

주성분분석은 공분산행렬이 핵심역할을 하게 된다. 공분산은 자료의 단위(scale)에 의존하는 값이므로 만약 변수들간의 척도의 차이가 매우 클 경우에는 적절히 표준화시킨후 분석을 시행하거나 공분산행렬 대신 상관행렬을 분석에 활용하는 편이 좋다.

1.1.3 시각화

각 주성분의 분산을 도표로 나타낸 scree plot, 처음 두 개의 주성분점수에 대한 산점도와 k번째 변수의 적재계수를 함께 나타낸 biplot이 대표적이다. scree plot은 주성분의 개수를 선택하기 위한 시각적인 방법으로도 활용된다 (elbow method).

1.1.4 실습

R 함수

1. 유용한 함수 : princomp (MASS 패키지)
2. 용법

princomp(x, cor=FALSE, scores=TRUE)

- x: 주성분분석을 위한 데이터. matrix나 data frame의 형태.
- cor : 공분산행렬/상관행렬 중 어떤 것을 분석할 것인지 특정
- socres : 주성분점수를 계산할지 여부를 특정

3. 결과값
- sdev : 주성분의 표준편차
- loadings : 적재계수(고유벡터 성분값)
- scores : 주성분점수

예제 1.1 USArrests 자료: 1973년 미국 50개 주의 100,000명당 체포된 강력범죄수(assault,murder,rape) 및 도시거주인구비율(UrbanPop).

```r
data("USArrests"); USArrests
#산점도
plot(USArrests)
#요약
summary(USArrests)
#패키지 로딩
p1 <- princomp(USArrests, cor=TRUE) # 상관행렬에 의한 주성분분석 시행
ls(p1) # 어떤 object들이 포함되어 있는지 확인
p1$loadings  # 각 변수들의 주성분에의 기여도 (고유벡터 집합)
p1$sdev      # 각 주성분의 표준편차
summary(p1)  # 주성분의 표준편차, 각 주성분의 변동의 상대비율, 누적변동비율
screeplot(p1,type="lines")  # scree plot
pve <- cumsum(sort(p1$sdev^2,decreasing = T))/sum(p1$sdev^2)
plot(pve, xlab="Principal Component", ylab="Proportion of Variance Explained",
     ylim=c(0,1),type='b')
biplot(p1)                  # biplot : 제 1,2 주성분점수에 의한 산점도 + 적재계수
plot(p1$scores[,1],p1$scores[,2]) # 제 1,2 주성분점수에 의한 산점도
```

주성분의 해석

주성분분석은 차원축소를 위한 기법으로 좋은 해석을 염두에 두고 고안된 방법은 아니다. 따라서, 주성분이 항상 해석이 가능한 것은 아니다. 그러나, 해석이 가능한 경우 데이터 구조의 이해에 도움이 될 수 있다.

위 예제의 경우

1. 제 1 주성분의 적재계수를 살펴보면 범죄관련 변수에 대해서 비슷한 가중치를 보이고 있고 인구관련 변수에 대한 가중치는 상대적으로 작다. 따라서, 제 1 주성분은 전반적인 범죄율에 대응되는 것으로 해석할 수 있다.

2. 제 2 주성분의 경우 인구관련 변수에 매우 큰 가중치를 두고 있으므로, 도시화 수준을 나타내는 것으로 간주할 수 있다.

3. biplot을 살펴보면, UrbanPop변수가 다른 변수와 상대적으로 멀리 떨어져 있음을 볼 수 있다. 즉, 범죄관련 변수들끼리는 서로 좀 더 상관성이 있고, 인구관련 변수는 상대적으로 상관성이 약한 것을 확인할 수 있다.

예제 1.2 "ade4" 패키지의 olympic 데이터 : 올림픽 10종경기 종목에서 33명의 선수 기록.

```
library(ade4)
data("olympic"); olympic
#요약
summary(olympic$tab)
p2 <- princomp(olympic$tab, cor=TRUE) # 상관행렬에 의한 주성분분석 시행
p2$loadings  # 각 변수들의 주성분에의 기여도 (고유벡터 집합)
p2$sdev      # 각 주성분의 표준편차
summary(p2)  # 주성분의 표준편차, 각 주성분의 변동의 상대비율, 누적변동비율
screeplot(p2,type="lines")  # scree plot
pve <- cumsum(sort(p2$sdev^2,decreasing = T))/sum(p2$sdev^2)
plot(pve, xlab="Principal Component", ylab="Proportion of Variance Explained",
     ylim=c(0,1),type='b')
biplot(p2)                  # biplot : 제 1,2 주성분점수에 의한 산점도 + 적재계수
plot(p2$scores[,1], olympic$score)  # 제 1 주성분 점수와 전체 10종경기 총점 사이의 산
점도
```

1. 제 1 주성분은 단거리 종목에 양의 가중치를 둔 반면 던지기와 뛰기 종목에 대해서는 음의 가중치를 주고 있다. 즉, 나머지 종목 대비 단거리 종목의 능력을 전반적으로 보여주는 변수로 해석해 볼 수 있다.

2. 제 2 주성분은 직관적으로 뚜렷한 해석을 찾기 어렵다.

3. biplot을 살펴보면, 달리기관련, 던지기관련, 뛰기관련 종목 변수들이 서로 가까이에 위치하여 서로 상관이 높다는 사실을 확인할 수 있다.

4. 한 가지 흥미로운 사실은 제 1 주성분 점수와 총점 사이에 반비례 관계가 성립하는 것처럼 보인다는 점이다.

연습 "MASS" 패키지의 Boston 자료 : 1970년 보스턴 지역에서 조사한 census 자료로, 집값의 중위수를 비롯하여 범죄율, 환경관련변수 등 다양한 변수 포함.

위 데이터에 대하여 medv, chas, rad를 제외한 나머지 변수들에 대하여 주성분분석을 해 보고 결과를 해석하여 보아라.

1.2 인자분석 (Factor analysis)

인자분석은 잠재변수의 존재를 전제로 한다. 관측된 변수들의 값들이 어떤 공통요인들의 영향을 받아 실현되었다는 가정을 기초로 한다. 보통 더 작은 수의 잠재변수를 식별해 냄으로써 차원축소를 가능케 하고 이 소수의 변수들간의 공분산구조를 파악해 내는 데에 목적이 있다. 이 때, 이 잠재변수를 일컬어 인자(factor) 혹은 요인이라 한다.

흔히 변수들이 상관성이 높은 것끼리 서로 그룹화되어 있다고 가정하게 되는데, 이 경우 한 그룹에 속하는 변수들은 하나의 인자에 의해 관찰된 속성으로 이들의 높은 상관성은 이 잠재인자에 의해 설명된다고 할 수 있다.

인자분석에 의해 차원축소의 실현이 가능할 수 있으며, 선택된 요인들에 의해 큰 영향을 받지 않는 변수가 있다면 제거할 수도 있다. 설문조사분석에서 여러 문항들이 동일한 개념을 측정하고 있는지를 조사하는 타당성 분석이 좋은 예이다. 또한, 향후 얻어진 인자점수 등을 이용하여 후속분석에의 활용도 가능하다.

1.2.1 Motivating example

불어, 영어, 수학, 체육 등의 시험점수로부터 얻는 자료가 있다고 가정하자. 그렇다면 "지능"이라는 인자에 의해 체육을 제외한 나머지 변수들의 높은 상관성을 설명할 수 있고 "체력"이라는 또 다른 인자에 의하여 체육점수를 설명할 수 있을 것이다. 이처럼 숨겨져 있는 소수의 인자를 이용하여 자료구조를 단순화하여 파악하는 것이 가능하다.

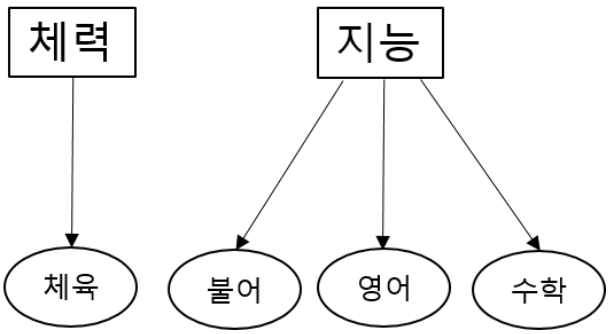


Figure 1.1: 인자분석 예시

1.2.2 인자분석 모형

모형

p개의 관측가능한 변수 $X_1, X_2, \ldots, X_p$의 기댓값(평균)이 $\mu_1, \ldots, \mu_p$라 하자. 만약, $q < p$개의 관측되지 않는 잠재변수 $F_1, F_2, \ldots, F_q$가 존재하여 다음을 만족하는 선형결합이 존재한다고 가정하자.

$$\begin{aligned} X_1 - \mu_1 &= l_{11}F_1 + \cdots + l_{1q}F_q + \epsilon_1 \\ \vdots &= \quad \vdots \\ X_p - \mu_p &= l_{p1}F_1 + \cdots + l_{pq}F_q + \epsilon_p \end{aligned}$$

즉, $X_j,\ j = 1, \ldots, p$들이 q개의 공통인자의 선형결합 (+오차항) 으로 표현된다.

- l_{jk}들을 j번째 변수의 k번째 인자에 대한 인자적재계수(factor loading)이라 한다.
- 위 식을 행렬로 간단히 표현할 수 있다.

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

$X = (X_1, \ldots, X_p)^\top,\ \mu = (\mu_1, \ldots, \mu_p)^\top,\ L = (l_{jk}),\ F = (F_1, \ldots, F_q)^\top,\ \epsilon = (\epsilon_1, \ldots, \epsilon_p)^\top$

- 인자적재계수의 계산을 위해서는 몇 가지 가정이 필요하다.
- 인자적재형렬 L은 유일하게 결정되지 않는다. 임의의 직교행렬 Q에 대하여

$$LF = LQ^\top QF = L^*F^*$$

이 성립하기 때문이다. $(LQ^\top = L^*, QF = F^*, QQ^\top = Q^\top Q = I_p)$

가정

1. 잠재변수들은 중심화, 직교정규화되어 있다.

$$E(F) = 0,\ Cov(F) = I_q$$

2. 오차항은 서로 독립이다.

$$E(\epsilon) = 0,\ Cov(\epsilon) = \Psi = diag(\psi_1, \ldots, \psi_p)$$

3. F와 ϵ은 서로 독립이다.

위 가정들로부터 X의 공분산 행렬을 다음과 같이 표현할 수 있다.

$$\Sigma = Cov(X) = Cov(LF + \epsilon, LF + \epsilon) = LL^\top + \Psi$$

또한, X와 F의 공분산은 다음과 같다.

$$Cov(X, F) = L$$

공통성 및 모형의 설명력

j번째 변수 X_j의 분산은

$$Var(X_j) = \sigma_j^2 = h_j^2 + \psi_j$$

로 분해된다. 여기서 $h_j^2 = \sum_{k=1}^{q} l_{jk}^2$으로 정의되며 인자 $F_1, \ldots, F_q$들이 X_j에 기여하는 비율로 생각할 수 있다. 이를 공통성(commuality)이라고 부른다. 공통인자들에 의해 설명이 가능한 변동이라는 의미로 볼 수 있다. ψ_j는 X_j의 오차분산 혹은 특정분산(specific variance)이라 한다. 이 부분은 인자에 의해 설명되지 않는 변동이다.

위 관계로부터 인자분석 모형으로부터 X_j가 설명되는 비율을

$$\frac{h_j^2}{\sigma_j^2} = 1 - \frac{\psi_j}{\sigma_j^2}$$

으로 정량화할 수 있다.

또한, 모든 변수 $X_j,\ j = 1, \ldots, p$에 존재하는 변동이 k번째 인자에 의해 설명되는 비율(설명력)은

$$\frac{\lambda_k^2}{\sum_{j=1}^{p} \sigma_j^2}$$

으로 정량화할 수 있다. 주의할 것은 여기서 $\lambda_k^2 = \sum_{j=1}^{p} l_{jk}^2$으로 공통성의 정의와는 다르다는 점이다. 보통 이 비율을 필요한 인자의 개수를 결정하는데 이용한다.

변수의 변동은 공통성과 특정분산으로 이루어져 있다. 만약, 특정변수의 공통성이 크다면 다른 변수들과 공유되는 성질(상관성)이 크다는 의미로 볼 수 있을 것이다.

참고 상관행렬에 의한 인자분석시에는 $\sum_{j=1}^{p} \sigma_j^2 = p$가 된다.

1.2.3 추정

인자분석에서 추정대상은 인자적재행렬, 특정분산, 인자점수(factor score) 등이다.

주축인자법

주축인자법(principal factor method)은 주성분분석을 이용하여 인자를 추정하는 방식이다. 이는 Σ가 스펙트럴 분해에 따라 다음과 같이 표현될 수 있음을 이용한 것이다.

$$\Sigma = V\Lambda^{1/2}(V\Lambda^{1/2})^\top = LL^\top + 0_p,\ L = V\Lambda^{1/2}$$

여기서 V는 고유벡터행렬, $\Lambda^{1/2}$은 고유값들의 양의 제곱근을 대각원소로 가지는 행렬이다. 위 식에 의해 k번째 인자에 대한 적재계수들은 가중화된 고유벡터 $\sqrt{\lambda_k}v_k$로 주어진다 (λ_k, v_k는 각각 Σ의 고유값, 고유벡터). 또한, L_q를 L의 처음 q개의 열로만 이루어진 행렬로 정의하여 인자적재행렬을 추정한다.

인자적재형렬에 대한 추정이 이루어지면 특정분산은

$$\Psi_q = diag(\Sigma - L_q L_q^\top) = diag(\sigma_j^2 - h_j^2)$$

으로 추정할 수 있다.

수정 주축인자법

공통성을 얼마나 잘 추정하는지가 인자분석에서 매우 중요한 요소이다. 수정 주축인자법은 반복적으로 특정분산을 수정해 나감으로써, 추정을 개선하고자 하는 목적으로 개발되었다.

만약, 특정분산에 대한 초기추정치 $\hat{\Psi}^{(0)}$가 주어졌다고 하면

$$\Sigma^* = \Sigma - \hat{\Psi}^{(0)}$$

로 수정된 공분산 행렬 Σ^*를 얻는다. 여기에 주축인자법을 적용하면 공통성과 특정분산을 추정할 수 있다. 즉, 업데이트된 특정분산 행렬 $\hat{\Psi}^{(1)}$을 얻게 되는 셈이다. 이 과정을 수행하되 공통성(혹은 특정분산)의 값이 거의 변화가 없을 때까지 반복하여 구하는 방식이다.

최대우도추정법

자료가 다변량 정규분포를 따른다는 가정 하에서, 최대우도추정(Maximum likelihood estimation) 원리에 따라 인자모형의 모수를 추정하는 방법이다. 확률분포에 기초한 방법이므로 모형의 적합도 등을 정량적으로 계산하고 검정해 볼 수 있다는 장점이 있다.

다변량 정규분포 외에 다른 분포를 가정한 분석법도 연구되고 있다.

1.2.4 인자회전 및 인자점수

인자회전

전술한 바대로 인자모형에서 인자(혹은 인자적재행렬)는 유일하게 결정되지 않는다. 따라서, 해석이 좀 더 용이하도록 적절한 직교변환을 실시할 수 있다. 이를 인자회전(factor rotation)이라 한다.

- 직교행렬 Q에 대해서
$$\hat{L}\hat{L}^\top + \hat{\Psi} = \hat{L}^*\hat{L}^{*\top} + \hat{\Psi}, \quad \hat{L}^* = \hat{L}Q.$$
즉, 인자의 회전에 의해서는 특정분산과 공통성은 변하지 않아 적합도 등에는 아무런 영향을 주지 않는다.
- 따라서, 인자의 회전은 특정 기준을 최적화하는 방식으로 하게 되며 보통 해석이 용이한 방향으로 하게 된다.
- 빈번이 쓰이는 회전방식은 배리맥스(varimax), 프로맥스(promax), 사각(oblique)회전 등이 있다.

인자점수

인자분석에서는 주성분분석과 달리 인자를 원 변수들의 선형결합으로 표현할 수 없다. 대안으로 인자점수를 고려하게 된다. 인자모형이 회귀모형과 비슷한 형태라는 사실로부터 원 변수 행렬을 종속변수, 인자적재계수들을 설명변수로 간주한다면 인자점수는 회귀모형에서 회귀계수에 대응되는 것으로 생각할 수 있다. 이를 이용하여 인자점수의 추정이 가능하다.

1.2.5 인자분석과 주성분분석의 차이

두 분석법은 차원축소를 한다는 측면에서 비슷해 보이지만 다음과 같은 차이가 있다.

1. 주성분모형은 선형직교변환에 기초한 수학적 모형이지만 인자모형은 오차를 고려한 통계적 모형이다.

2. 주성분들은 원 변수들의 선형결합으로 표현되며, 인자들은 원 변수들에 영향을 미치는 잠재요인으로 가정된다.

3. 보통 주성분분석은 차원의 축소, 인자분석은 구조의 단순화에 초점을 맞추는 경향이 있다.

1.2.6 실습

R 함수

1. 유용한 함수 : factanal (nFactor 패키지), principal (psych 패키지)

2. 용법 및 결과값

   - factanal 함수 (최대우도추정법)

     factanal(x, factors, covmat, scores, rotation)

     - x : 인자분석을 위한 데이터. matrix나 data frame의 형태.
     - factors : 인자의 개수
     - covmat : x의 공분산/상관행렬을 직접 지정할 수도 있음
     - scores : 인자점수 추정방법 ("regression", "Bartlett")
     - rotation : 회전방법 ("none", "varimax", "promax", "oblimin", ...)

   * 결과값
     - loadings : 인자적재계수
     - scores : 인자점수
     - uniqueness : 특정분산
     - correlation : 사용된 상관행렬

   - principal 함수 (주축인자법)

     principal(r, nfactors, rotate, covar, scores, method)

- r : 인자분석을 위한 데이터 행렬 혹은 상관행렬
- nfactors : 인자의 개수
- rotate : 회전방법 ("none", "varimax", "promax", "oblimin", ...)
- covar : FALSE면 r로부터 상관행렬을 계산하거나, 공분산행렬을 상관행렬로 변환
- scores : 인자점수 추정여부 TRUE/FALSE
- method : 인자점수 추정방법 ("regression", "Bartlett")

* 결과값
  - values : 고유값
  - commuality : 공통성
  - loadings : 인자적재계수
  - scores : 인자점수

예제 1.3 Harman 자료 : 시카고 교외의 145명 학생들에게 실시한 24가지의 심리테스트 점수. factanal(nFactors) 함수.

```
library(psych)
data("Harman74.cor"); Harman74.cor
#패키지 로딩
library(nFactors)
Harman74.FA <- factanal(factors = 1, covmat = Harman74.cor) # 1-factor model
Harman74.FA
Harman74.FA$loadings  # 요인의 설명력
Harman74.FA$uniquenesses # 특정분산 추정치
Harman74.FA$scores
```

1. 위 모형에서는 1개의 인자로 원 자료의 변동을 설명하고자 하였다.
2. 1개의 인자로는 전체변동의 $7.438/24 = 0.310$, 즉 약 31%가 설명될 수 있다.
3. 'factanal'은 최대우도추정법에 의한 추정을 제시한다. 요약결과의 p-value는 다변량정규분포에 의한 모형의 적합도에 대한 검정이다. 본 결과에서는 모형이 적합하지 않음을 알 수 있다.

```
for(factors in 2:5) print(update(Harman74.FA, factors = factors)) # 2 to 5 factors
Harman74.FA <- factanal(factors = 5, covmat = Harman74.cor,
                        rotation = "none")
(LL <- Harman74.FA$loadings)
rowSums(LL^2)       # commuality
```

1. 인자를 2개부터 5개까지 늘리며 순차적으로 시행해 보았을 때, 유의수준 0.05에서는 최소한 5개의 인자를 가지는 모형이 필요함을 알 수 있다.
2. 5개의 인자로는 전체변동의 약 50%가 설명됨을 알 수 있다.

```
Harman74.FA.rotation <- factanal(factors = 5, covmat = Harman74.cor,
                        rotation = "promax")  # promax 회전
Harman74.FA.rotation
plot(Harman74.FA.rotation$loadings)
text(Harman74.FA.rotation$loadings, labels = colnames(Harman74.cor$cov),cex=0.5)
```

1. promax에 의해 인자회전을 하였을 때, 전체 분산의 설명비율이 회전 전과 비교하여 같음을 알 수 있다.
2. 회전된 인자적재행렬의 처음 2개의 인자계수에 대한 산점도를 통하여 비교적 자연스러운 군집이 인지능력별로 형성되었음을 알 수 있다.

예제 1.4 Harman 자료

principal(psych) 함수

```
pc <- principal(Harman74.cor$cov,nfactors=5,rotate="none") # 주축인자법
pc$loadings
pc$communality
pc.rot <- principal(Harman74.cor$cov,nfactors=5,rotate="promax")
pc.rot$loadings
pc.rot$communality
```

```r
plot(pc.rot$loadings)
text(pc.rot$loadings, labels = colnames(Harman74.cor$cov),cex=0.5)
```

참고 대부분의 함수에서는 인자점수를 추정하는 옵션을 두고 있다. 인자점수를 추정하기 위해서는 공분산(상관)행렬 뿐 아니라 원자료 행렬이 필요하다. 위에서 소개한 예에서는 자료의 상관행렬만을 제공하고 있으므로 인자점수의 추정이 불가하다.

연습 "MASS" 패키지의 Boston 자료 : 1970년 보스턴 지역에서 조사한 census 자료로, 집값의 중위수를 비롯하여 범죄율, 환경관련변수 등 다양한 변수 포함.

위 데이터에 대하여 medv, chas, rad를 제외한 나머지 변수들에 대하여 인자분석을 실시하고 결과를 해석하여라.