

10장 비지도학습

서울대학교
통계학과

2018년 8월

1. 군집분석

노트. 이 장에서 다룰 내용

1. 자율학습과 지도학습의 소개
2. K평균 군집분석
3. 계층군집화

군집분석 I

1. 자료를 동질적인 군집 혹은 부분으로 나누는 분석.
2. 각 군집을 쉽게 설명하므로 전체를 설명하고자 한다.
3. 군집분석의 예
 - 3.1 유방암 환자들의 자료를 이용해 알려지지 않은 암의 subtype을 찾으려고 한다.
 - 3.2 사람들의 자료를 이용하여 시장을 분할하고, 특정 광고에 잘 반응할 subgroup을 찾으려 한다.

자율학습(unsupervised Learning)

1. 반응변수 Y 는 없고 변수들 X_1, \dots, X_p 만 있는 경우이다.
2. 예측에는 관심이 없고, 변수들 사이의 특별한 관계가 있는가? 관측치들에 그룹이 있는가 등의 질문에 관심이 있다.
3. 보통 탐색적 자료분석의 일부분이다.
4. 자료분석의 결론이 성능이 좋은지 안좋은지 판단하기가 어렵고, 주관적인 측면이 강하다.

지도학습(supervised learning)

1. 자료가 반응변수 Y 와 설명변수 X_1, \dots, X_p 로 구성되어 있다.
2. 설명변수 X_1, \dots, X_p 가 주어져 있을 때, 반응변수 Y 의 예측에 목적이 있다.


K 평균(K means) 군집분석 | k-means clustering가

군집

C_1, \dots, C_K 를 군집들이라 한다. 이는 다음의 조건을 만족한다.

$$\bigcup_{k=1}^K C_k = \{1, 2, \dots, n\} \quad C_i \cap C_j = \emptyset, \quad i \neq j.$$

목표

$$\sum_{k=1}^K W(C_k)$$


ex)

를 최소화하는 C_1, \dots, C_K 를 찾고자 한다.

K 평균(K means) 군집분석 II

여기서 $W(C_k)$ 는 군집 C_k 의 변동을 측정하는 척도이다. 즉, C_k 의 동질성이 커지면 작아지는 값으로 대표적인 예로는

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,j \in C_k} \|x_i - x_j\|^2$$

를 생각한다.



: $d(x_i, x_j)$

가

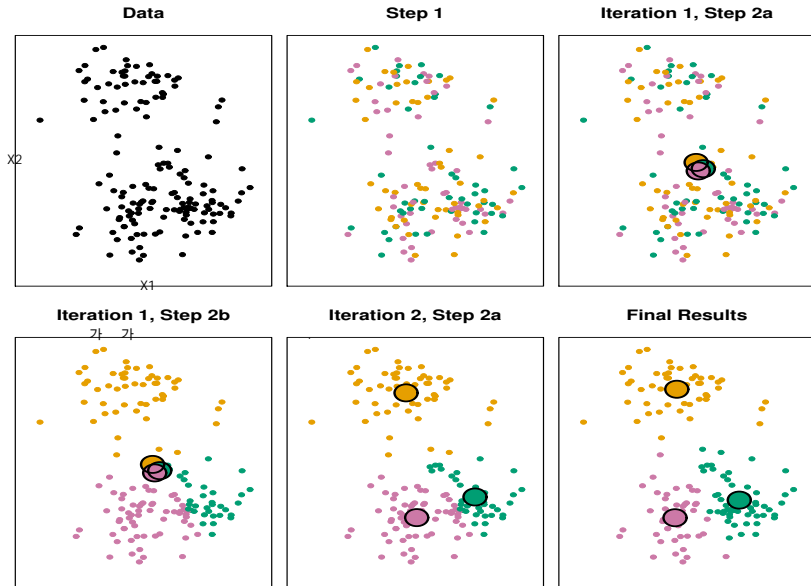
K 평균(K means) 군집분석 III

알고리즘

1. 관측치들을 C_1, \dots, C_K 에 랜덤하게 할당한다.
2. 군집이 바뀌지 않을 때까지 다음을 반복한다.
 - 2.1 각 군집마다 군집의 중앙(centroid)을 계산한다.
 - 2.2 모든 관측치를 가장 가까운 중심의 군집에 할당한다.

K 평균(K means) 군집분석 IV

step1 ()



K 평균 군집화 R 코드 I

자료의 생성

```
set.seed(2)
x=matrix(rnorm(50*2), ncol=2)
x[1:25,1]=x[1:25,1]+3
x[1:25,2]=x[1:25,2]-4
```

50개의 자료를 생성하고, 첫 25개의 x_1, x_2 값을 바꾸었다.

K-means

1. Initial

-kmeans++

-nstart

2. cluster

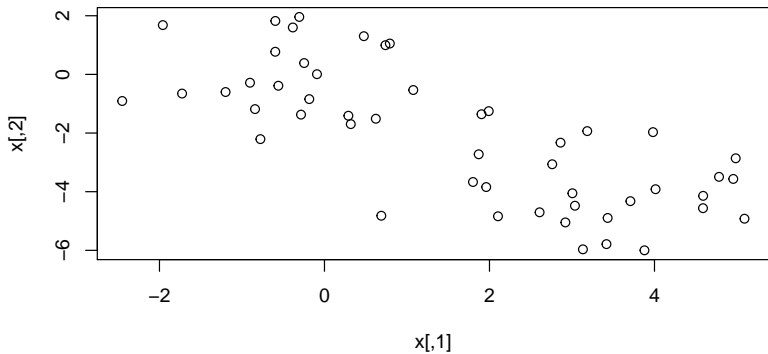
3. meta data (dist(Xi, Xj))

-center

-center PAM(Partition Around Medoids,)


K 평균 군집화 R 코드 II

```
plot(x)
```



K 평균 군집화 R 코드 III

K 평균 군집화의 수행

 `x` numeric
`km.out=kmeans(x,2,nstart=20)`

1. $K = 2$ 인 군집화이다. 2는 군집의 개수를 의미한다.
2. 종종 초기값으로 쓰인 군집의 중앙값에 군집화의 결과가 다를 수 있다. `nstart=20`은 20개의 초기값으로 군집화를 한 후에 가장 좋은 결과를 보고하라는 뜻이다.
3. 초기 중앙값을 랜덤하게 선택하므로 수행할 때마다 결과가 다를 수 있다. 이를 방지하기 위해 `set.seed` 함수를 사용하는 것이 좋다.

1) k-means

EM

([https://ko.wikipedia.org/wiki/K-%ED%8F%89%EA%B7%A0_%EC%95%8C%EA%B3%A0%EB%A6%AC%EC%A6%98#EM_%EC%95%8C%EA%B3%A0%EB%A6%AC%EC%A6%98\[25\]\)](https://ko.wikipedia.org/wiki/K-%ED%8F%89%EA%B7%A0_%EC%95%8C%EA%B3%A0%EB%A6%AC%EC%A6%98#EM_%EC%95%8C%EA%B3%A0%EB%A6%AC%EC%A6%98[25])))

2)

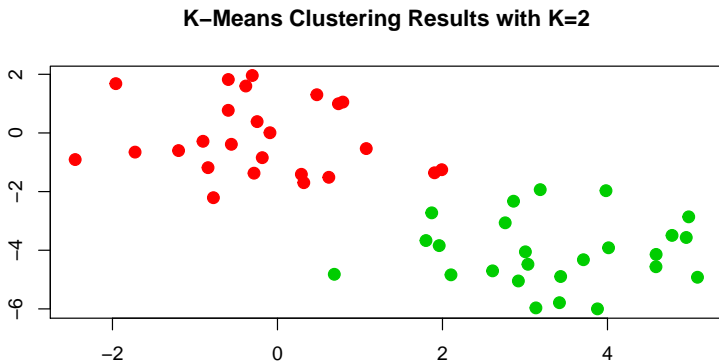
가

ex)

...?

K 평균 군집화 R 코드 IV

```
plot(x, col=(km.out$cluster+1),  
main="K-Means Clustering Results with K=2", xlab="", ylab="", pch=20, cex=2)
```



K 평균 군집화 R 코드 V

```
km.out
```

```
## K-means clustering with 2 clusters of sizes 25, 25
##
## Cluster means:
##           [,1]      [,2]
## 1 -0.1956978 -0.1848774
## 2  3.3339737 -4.0761910
##
## Clustering vector:
## [1] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1 1
## [36] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##
## Within cluster sum of squares by cluster:
## [1] 65.40068 63.20595
## (between_SS / total_SS =  72.8 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
## [5] "tot.withinss" "betweenss"    "size"         "iter"
## [9] "ifault"
```

K 평균 군집화 R 코드 VI

클러스터 벡터를 보면 정확하게 첫 25개와 후반 25개로 군집으로 나눈것을 알 수 있다.

아래는 세 개의 군집으로 K 평균 알고리즘을 수행한 결과이다.

```
set.seed(4)
km.out=kmeans(x,3,nstart=20)
```

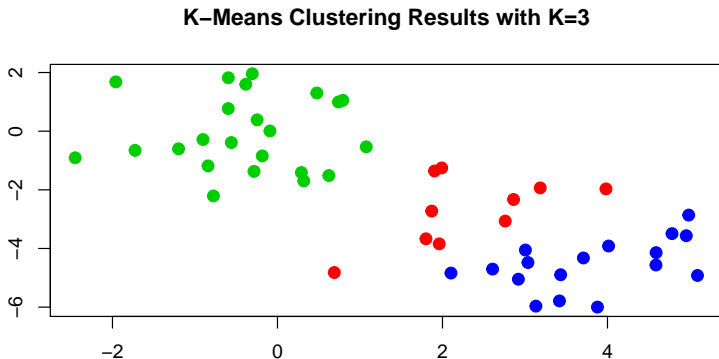
K 평균 군집화 R 코드 VII

```
km.out

## K-means clustering with 3 clusters of sizes 10, 23, 17
##
## Cluster means:
##      [,1]      [,2]
## 1  2.3001545 -2.69622023
## 2 -0.3820397 -0.08740753
## 3  3.7789567 -4.56200798
##
## Clustering vector:
## [1] 3 1 3 1 3 3 3 1 3 1 3 1 3 1 3 3 3 3 3 1 3 3 3 2 2 2 2 2 2 2 2
## [36] 2 2 2 2 2 2 2 2 1 2 1 2 2 2 2
##
## Within cluster sum of squares by cluster:
## [1] 19.56137 52.67700 25.74089
## (between_SS / total_SS =  79.3 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
## [5] "tot.withinss" "betweenss"    "size"         "iter"
## [9] "ifault"

plot(x, col=(km.out$cluster+1), main="K-Means Clustering Results with K=3", xlab="", ylab="",
     pch=20, cex=2)
```


K 평균 군집화 R 코드 VIII



K 평균 군집화 R 코드 IX

k "elbow analysis"
total within sum of square가 가

```
set.seed(3)
km.out=kmeans(x,3,nstart=1)
km.out$tot.withinss

## [1] 104.3319

km.out=kmeans(x,3,nstart=20)
km.out$tot.withinss

## [1] 97.97927
```

초기값을 1개 사용한 것과 20개 사용한 것을 비교하였다.
20개의 초기값을 사용한 것이 내부제곱합이 더 작다. 항상
초기값을 20 혹은 50 같이 큰 값을 사용하는 것을 추천한다.

계층군집화(hierarchical clustering) I

bottom-up 가 가 top-down 가
 , 가 가 2 가 가

1. Bottom-Up 절차 vs Top-Down 절차
2. 계층군집화의 결론으로 덴도그램이 생성된다.
3. 덴도그램(dendrogram)

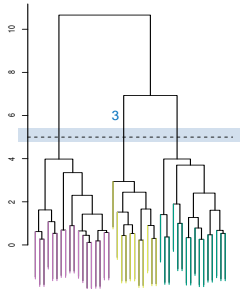
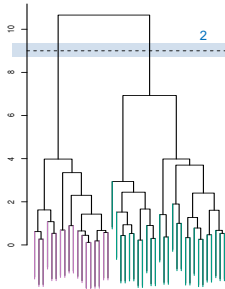
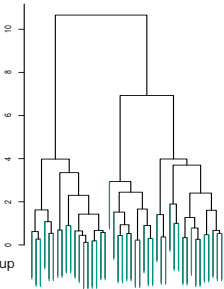
top-down

bottom-up

top-down



bottom-up



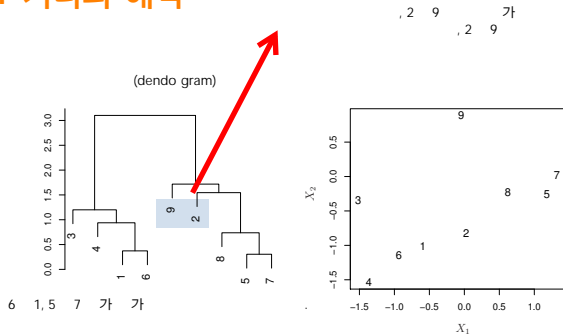
$$\begin{aligned}
 d(\{x1, x2\}, x3) &= \min(d(x1, x3), d(x2, x3)) \\
 &= \max(d(x1, x3), d(x2, x3)) \\
 &= \text{Aug}(d(x1, x3), d(x2, x3)) \\
 &: d(C1, Xj) \quad C1 \quad \{X1, X2\}
 \end{aligned}$$

계층군집화(hierarchical clustering) II

1. 중간그림. $y = 9$ 에서 자르면 2개의 군집이 나타난다.
2. 오른쪽 그림 $y = 5$ 에서 자르면 3개의 군집이 나타난다.
3. 바닥에서 합쳐진 관측치들은 가깝고, 상부에서 합쳐진 관측치들은 거리가 멀다.

계층군집화(hierarchical clustering) III

주의점 : 거리의 해석



그림을 보면 9가 2와 제일 가깝고 5,7,8과는 멀다고 보이는데, 그렇게 해석하면 안된다. 2, 5, 7,8 부분의 덴도그림을 표현하는 방법은 매우 많다. 덴도그램의 x축은 가까움을 나타내지 않는다.

계층군집화(hierarchical clustering) IV

계층군집화가 적당하지 않은 자료

계층군집화 방법의 계층구조는 매우 매력적이지만 어떤 자료에는 적합하지 않다. 예를 들면 남성과 여성과 함께 백인 흑인 황인으로 나누어진 자료가 있다고 한다. 이 그룹은 네스티드되어 있지 않기 때문에 계층군집은 자료를 잘 나타내지 못하고 오히려 K 평균 방법이 더 나을 수 있다.

계층군집화(hierarchical clustering) V

계층군집화의 알고리즘

1. 한 개의 관측치가 포함된 n 개의 군집으로 시작한다.
 n 개의 군집간 거리를 계산한다.
2. $i = n, n - 1, \dots, 2$
 - 2.1 i 개의 군집 간의 거리를 재서 가장 거리가 작은 군집 2개를 합친다.
 - 2.2 $i - 1$ 개의 군집간 거리를 계산한다.

계층군집화(hierarchical clustering) VI

연결법(linkage) : 군집간의 거리를 계산하는 방법

1. complete

가

$$d(C_1, C_2) = \max_{i \in C_1, j \in C_2} d(x_i, x_j)$$

2. single

가

$$d(C_1, C_2) = \min_{i \in C_1, j \in C_2} d(x_i, x_j)$$

100

가
1

99

3. average

$$d(C_1, C_2) = \text{ave}_{i \in C_1, j \in C_2} d(x_i, x_j)$$

4. centroid

$$d(C_1, C_2) = d(\bar{x}_1, \bar{x}_2), \quad \bar{x}_i = \text{ave}_{j \in C_i} (x_j).$$

계층군집화(hierarchical clustering) VII

노트

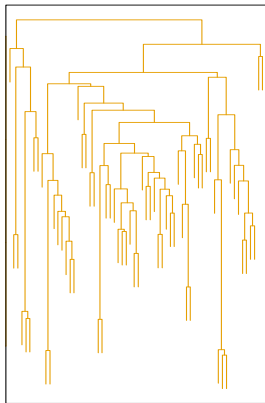
1. 보통 complete이나 average가 선호된다. 왜냐하면 balanced dendrogram을 만든다고 한다. 이유는 정확히 이해 못했다.
2. centroid는 genomics에서 종종 사용되는데 inversion이 생길 수 있는게 문제이다. inversion이 무엇인가?
3. 덴도그램은 연결법에 따라 매우 다르게 나타난다. 아래의 그림을 참조
4. 거리 측도도 군집의 형성에 많은 영향을 미친다. 유클리디언 거리외에 두 벡터 사이의 상관계수도 많이 쓰인다.

계층군집화(hierarchical clustering) VIII

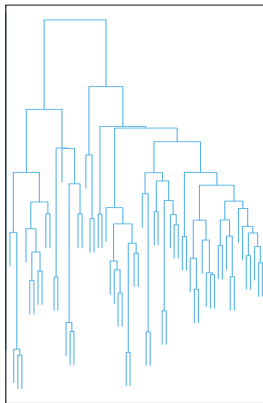
연결법에 따른 덴도그램의 차이 덴도그램은 연결법에 따라 매우 다르게 나타난다.

?

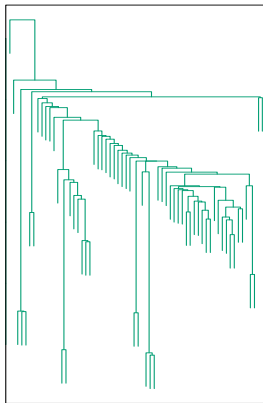
Average Linkage



Complete Linkage




Single Linkage



계층군집화(hierarchical clustering) IX

군집분석에서 결정해야할 문제들

1. 변수들을 표준화해야 하는가?
2. 계층 군집의 경우
 - 2.1 거리는 어떤 것을 사용해야하는가?
 - 2.2 연결법은?  2.1, 2.2, 2.3
 - 2.3 덴도그램은 어디서 잘라야 하는가? k (cluster) 가 !
3. K 평균 방법에서 K는 몇 개를 해야하나?

답변

위의 질문들에 대해 명확한 답은 없다. 보통 여러 개를 시도해보고 이 중 가장 해석이 좋은 것을 선택한다.

계층 군집화 R 코드 I

n x n

, numeric

가

가

dist.x = as.dist(1-cor(x))

```
hc.complete=hclust(dist(x), method="complete")
hc.average=hclust(dist(x), method="average")
hc.single=hclust(dist(x), method="single")
```

계층군집화를 수행하는 함수는 hcluster이다. dist(x)는 자료들 간의 거리를 구해주는 함수이다. method는 연결방법을 지정한다.

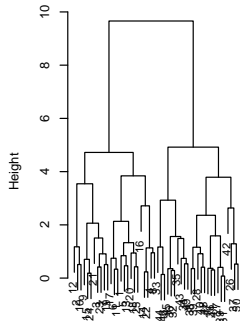
```
dist(x[1:4,])
```

```
##           1           2           3
## 2 3.099491
## 3 2.500046 2.979541
## 4 2.126855 1.534550 3.281285
```

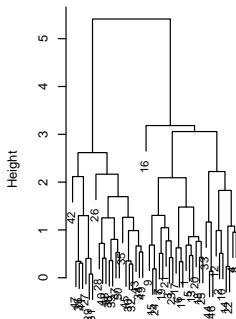
```
par(mfrow=c(1,3))
plot(hc.complete,main="Complete Linkage", xlab="", sub="", cex=.9)
plot(hc.average, main="Average Linkage", xlab="", sub="", cex=.9)
plot(hc.single, main="Single Linkage", xlab="", sub="", cex=.9)
```

계층 군집화 R 코드 II

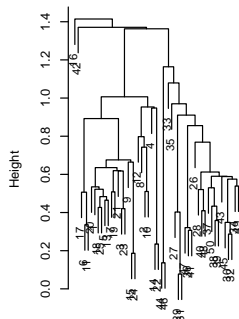
Complete Linkage



Average Linkage



Single Linkage



그림을 그릴 때는 plot 함수를 사용한다.

계층 군집화 R 코드 III

`cutree(hc.complete, 2)`

2

```
## [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2
## [36] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
```

`cutree(hc.average, 2)`

```
## [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 1 2 2
## [36] 2 2 2 2 2 2 2 2 1 2 1 2 2 2 2
```

`cutree(hc.single, 2)`

```
## [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [36] 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

`cutree(hc.single, 4)`

```
## [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 3 3 3 3 3 3 3 3 3 3
## [36] 3 3 3 3 3 3 4 3 3 3 3 3 3 3
```

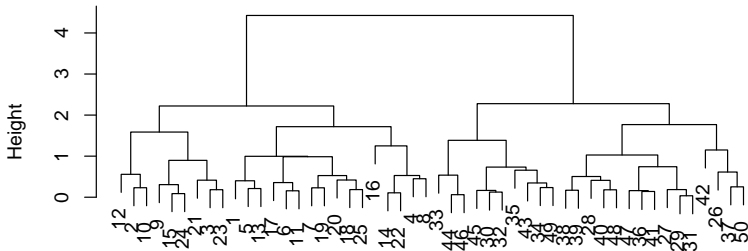
계층 군집화 R 코드 IV

군집의 인덱스를 구하는 함수는 `cutree`이다. `k=2`는 군집의 개수를 지정한다.

```
xsc=scale(x)
plot(hclust(dist(xsc), method="complete"),
     main="Hierarchical Clustering with Scaled Features")
```

계층 군집화 R 코드 V

Hierarchical Clustering with Scaled Features



```
dist(xsc)  
hclust (*, "complete")
```

변수를 표준화해서 계층군집화를 수행하였다.

계층 군집화 R 코드 VI

```
x=matrix(rnorm(30*3), ncol=3)
```

```
dd=as.dist(1-cor(t(x)))
```

abs()가

(abs(1-cor(t(x))))

correlation

1 가

"

,

"

0 가

"

,

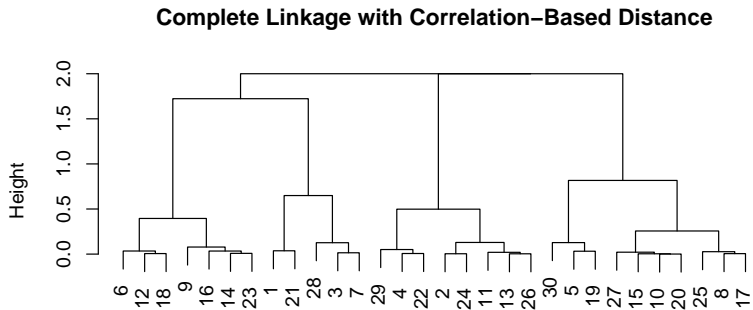
"

as.dist는 주어진 행렬을 거리행렬로 바꾸어준다.

```
plot(hclust(dd, method="complete"),
```

```
main="Complete Linkage with Correlation-Based Distance", xlab="", sub="")
```

계층 군집화 R 코드 VII



2. 주성분분석

노트. 다루는 내용 I

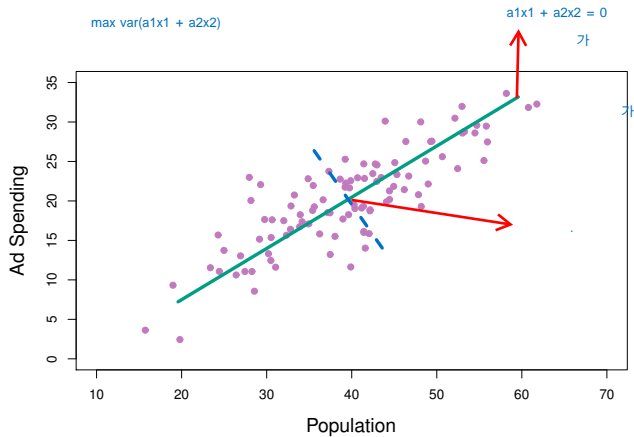
1. 주성분분석
2. 주성분회귀분석

주성분분석 I

주성분분석(principal component analysis)의 목적

1. p 개의 설명변수 X_1, X_2, \dots, X_p 가 있을 때, 설명변수들의 변동을 가장 크게 설명하는 새로운 저차원 변수들을 구하는 방법을 주성분분석이라 한다.
2. 새로운 변수들을 이용해 설명변수들 전체의 변동을 좀 더 쉽게 이해할 수 있다.

주성분분석 II



ex) (X_1, \dots, X_p) (Z_1, Z_2, Z_3) ..

주성분분석 III

노트.

1. 2차원 자료가 그림에 주어져 있다. 이 2차원 자료의 변동을 가장 잘 설명하는 1차원 설명변수가 무엇일까 물어본다.
2. 자료는 조그만 도시들의 인구수와 특정한 제품의 광고비로 이루어져 있다.

주성분의 정의

1. 평균이 0인 설명변수들 X_1, X_2, \dots, X_p 의 first principal component 첫번째 주성분은

$$\text{Score } \underline{Z_1} = \phi_{11}X_1 + \phi_{21}X_2 + \dots + \phi_{p1}X_p, \quad \sum_{j=1}^p \phi_{j1}^2 = 1$$

와 같은 X_1, X_2, \dots, X_p 의 선형조합 중 분산이 가장 큰 선형조합이다.

2. $\underline{\phi_1} = (\phi_{11}, \phi_{21}, \dots, \phi_{p1})'$ 를 적재값(loadings)이라 한다. (principal component)

$$\text{cov}(X) = VDV^T = (d_1 * v_1 * v_1^T) + (d_2 * v_2 * v_2^T) + \dots + (d_p * v_p * v_p^T)$$

주성분분석 V

$$\|X - Z\Phi\|$$

$n \times p$ $n \times q$ $q \times p$

3. 위의 그림의 경우

$$Z_1 = 0.839 \times (pop - p\bar{o}p) + 0.544 \times (ad - \bar{a}d)$$

가 첫번째 주성분이다. 평균이 0이 아닌 변수는 평균으로 빼준다.

4. 두번째 주성분은

$$Z_2 = \phi_{12}X_1 + \phi_{22}X_2 + \dots + \phi_{p2}X_p, \quad \sum_{j=1}^p \phi_{j2}^2 = 1, \quad \phi_1 \perp \phi_2$$

를 만족하는 X_1, X_2, \dots, X_p 의 선형조합 중 분산이 가장 큰 선형조합이다. 여기서 $\phi_2 = (\phi_{12}, \phi_{22}, \dots, \phi_{p2})'$ 는 적재값이다.

5. 이와 같이, 세번째에서 p 번째 주성분 Z_3, \dots, Z_p 를 정의한다.

주성분분석 VI

미국의 강력범죄 자료

주어진 자료는 USArrest 자료로 미국 50개 주의 도시인구비율(UrbanPop), 폭행(Assault), 살인(Murder), 강간(Rape)의 인구 10만명당 체포수를 포함한 자료이다.

	PC1	PC2
Murder	0.5358995	-0.4181809
Assault	0.5831836	-0.1879856
UrbanPop	0.2781909	0.8728062
Rape	0.5434321	0.1673186

주성분분석 VII

위의 표에서 첫번째 열의 벡터는 첫번째 주성분의 적재값, 두번째 열의 벡터는 두번째 주성분의 적재값이다. 첫번째 행의 벡터는 첫번째 변수 Murder의 주성분점수(principal component score), 두번째 행의 벡터는 두번째 변수 Assault의 주성분 점수이다.

노트. 주성분점수의 해석.

주성분 점수로 각 변수가 각 주성분에 어떻게 영향을 미치는지 해석할 수 있다. 예를 들면 UrbanPop은 pc1에는 큰 영향을 미치지 않지만 pc2에는 많은 영향을 미친다. pc1은 살인, 강도, 강간에 평균적으로 영향을 받고, pc2는 인구수에 주로 영향을 받는다. pc1은 범죄를, pc2는 인구수를 나타낸다고 할 수 있다. 즉, 살인, 강도, 강간의 체포수와 인구수로 이루어진 전체자료의 변동은 크게 범죄와 인구수의 변동으로 설명할 수 있다.

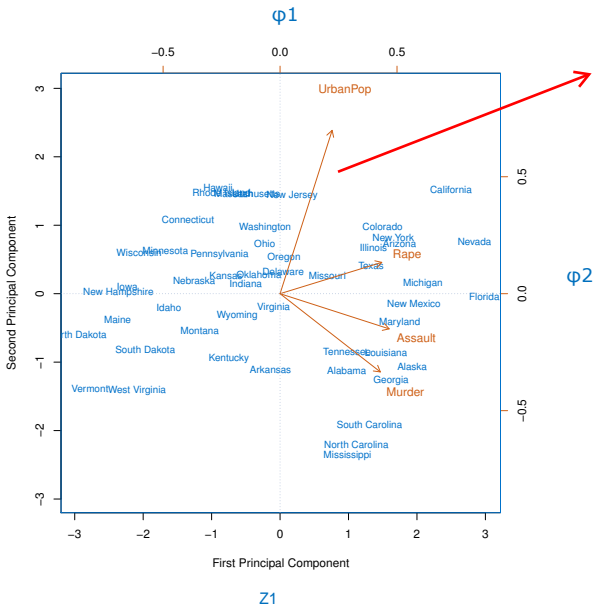
주성분분석 VIII

loading score

biplot



Z2



가

ϕ_2

Z1

주성분분석 IX

노트.

첫번째 주성분과 두번째 주성분을 그린 그림이다.

노트. 그림에 대한 해석.

1. 그림의 파란 점은 각 관측치에 해당하는 주성분 즉

$$z_{1i} = \phi_{11}x_{1i} + \dots + \phi_{p1}x_{pi}$$

$$z_{2i} = \phi_{12}x_{1i} + \dots + \phi_{p2}x_{pi}$$

를 계산하여 (z_{1i}, z_{2i}) 를 그린 것이다. 즉, **각 관측치를 주성분의 좌표로 그린 것이다.**

주성분분석 X

2. 그림의 오렌지 화살표는 각 변수에 해당하는 적재값을 우측과 위에 새로운 축을 그려서 그린 것이다. 예를 들면 x_1 에 해당하는 화살표는 (ϕ_{11}, ϕ_{12}) 를 그린 것이다. 이 값을 x_1 의 주성분점수(principal component scores)라 한다. 즉, 오렌지 화살표는 각 변수가 각 주성분에 미치는 영향을 나타낸다. 즉, UrbanPop의 경우 x 축의 좌표는 0 근처, y 축의 좌표는 1 근처인데 이는 UrbanPop이 pc1에는 영향을 안미치고, pc2에 주로 영향을 미친다는 것을 알 수 있다.
3. 이와 같은 그림을 쌍도(biplot)라 한다. 왜냐하면 이 그림이 주성분 적재값과 주성분점수(principal component scores)를 함께 그리기 때문이다.

주성분분석 XI

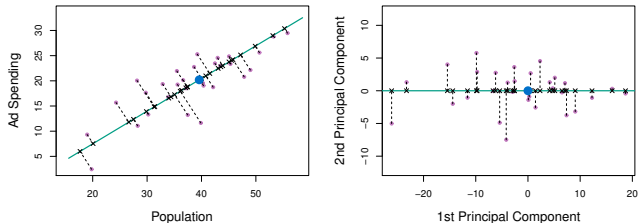
- 오렌지 화살표를 보면 변수 x_j 들이 주성분과 어떻게 관계되는지 알 수 있다. UrbanPop은 pc1에는 큰 영향을 미치지만 pc2에는 큰 영향을 미치지 않는다. Rape, Assault, Murder는 pc1에 비슷한 정도로 영향을 미친다. 또한 범죄 관련 변수들 Rape, Assault, Murder가 서로 가까이 위치하기 때문에 이들이 상관관계가 있다는 것을 알 수 있다. UrbanPop은 이들 범죄관련 변수들과 상관관계가 낮다.
- 파란점을 보면 각 주가 pc1과 pc2의 좌표로 어떻게 표시되는지 보인다. 예를 들면 Florida는 pc1은 크지만 pc2의 값은 거의 0에 가깝다. pc1은 주로 범죄관련 변수들로 설명되고 pc2는 주로 도시인구비율로 설명되므로, Florida의 전반적 범죄율이 높은 것을 알 수 있다. California, Nevada도 마찬가지이다. 반면에 North Dakota는 전반적 범죄율이 낮다.

주성분분석 XII

주성분분석의 또다른 해석

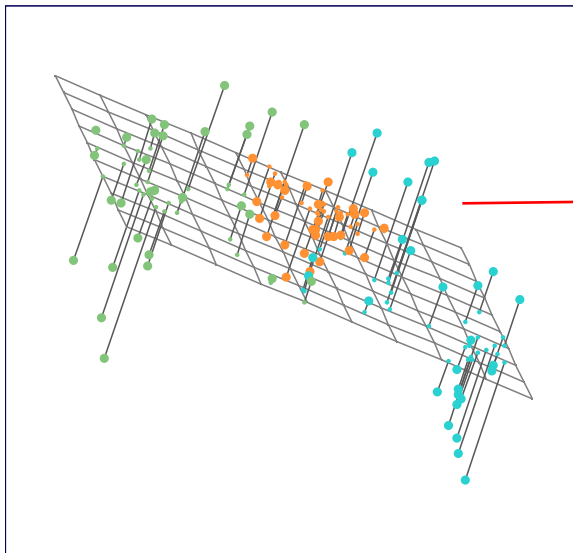
1. 첫번째 주성분은 X_1, \dots, X_p 의 공간에 하나의 벡터로 표현할 수 있다. 즉, 주성분 적재값은 변수들의 공간에서 하나의 벡터를 나타낸다. 모든 관측치를 이 벡터의 직선에 사영을 하고, 사영된 점과 원래 관측치 사이의 거리를 잔차라 하면, 첫번째 주성분은 이 잔차제곱합을 최소로 하는 벡터이다. 즉, 첫번째 적재값 벡터는 관측치들을 가장 잘 요약하는 벡터이다.

주성분분석 XIII



2. 첫번째와 두번째 주성분은 위와 같이 두 적재값 벡터를 X_1, \dots, X_p 의 공간에 평면으로 나타낼 수 있다. 마찬가지로 관측치를 이 평면에 사영하고 그 차이를 잔차라 하면 이 평면은 잔차를 최소로 하는 평면이다. 즉, 첫번째와 두번째 적재값 벡터로 이루어지는 평면은 관측치들을 가장 잘 요약하는 평면이다.

주성분분석 XIV



x-zφ가
가

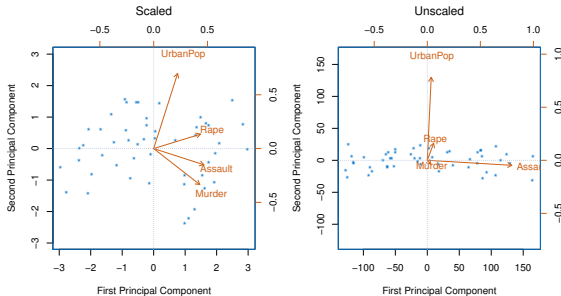
주성분분석 XV

3. 3, 4, ... p 차원에 대해서도 동일한 해석을 할 수 있다.

주성분분석의 이슈들 I

척도화(scaling)

1. 변수들을 척도화 했을 때와 척도화하지 않았을 때, 주성분분석의 결과가 달라진다.



주성분분석의 이슈들 II

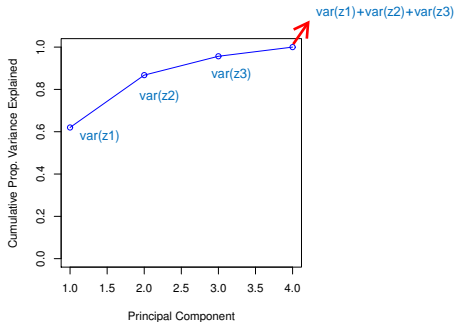
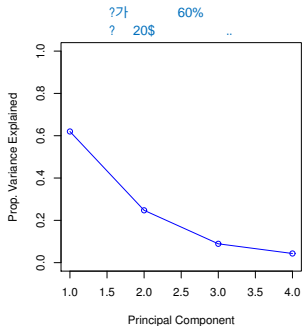
2. 왼쪽은 척도화했을 때, 오른쪽은 척도화하지 않을 때의 쌍도이다. Murder, Rape, Assault, UrbanPop의 분산은 각각 18.97, 87.73, 6943.16, 209.5이다. Assault의 분산이 가장 크기 때문에 첫번째 주성분은 Assault의 변동만 표현하게 된다. 오른쪽 그림이 이를 나타낸다.
3. 주성분분석하기 전에 대개 변수들을 척도화하는 것을 추천한다. 모든 변수가 동일한 단위일 때는 척도화하지 않아도 된다.

주성분의 유일성

주성분은 적재벡터의 부호를 제외하면 유일하다. 부호가 달라도 동일한 직선을 나타낸다.

주성분분석의 이슈들 III

산비탈그림(scree plot, 스크리그림) : 설명된 분산 비율



주성분분석의 이슈들 IV

왼쪽 그림은 전체 분산합 중에서 주성분의 분산의 비율을,
오른쪽 그림은 주성분 분산의 누적합을 그린 것이다.

노트.

1. 왼쪽 그림과 오른쪽 그림은 각각 다음을 그린다.

$$\frac{\text{Var}(Z_j)}{\sum_{j=1}^p \text{Var}(Z_j)} \text{와 } \frac{\sum_{j=1}^k \text{Var}(Z_j)}{\sum_{j=1}^p \text{Var}(Z_j)}.$$

2. $\sum_{j=1}^p \text{Var}(Z_j) = \sum_{j=1}^p \text{Var}(X_j)$ 이다.

주성분분석의 이슈들 V

몇 개의 주성분을 사용하나?

산비탈그림을 보고 판단한다. 감소하는 추세가 느껴지는 부분에서 끊는다.

노트.

위의 그림은 2개의 주성분을 선택하는 것이 좋을 것 같다.
세번째 주성분의 분산은 전체의 10% 미만을 설명한다.

주성분분석 R 코드 I

```
str(USArrests)
```

```
## 'data.frame': 50 obs. of 4 variables:  
## $ Murder : num 13.2 10 8.1 8.8 9 7.9 3.3 5.9 15.4 17.4 ...  
## $ Assault : int 236 263 294 190 276 204 110 238 335 211 ...  
## $ UrbanPop: int 58 48 80 50 91 78 77 72 80 60 ...  
## $ Rape : num 21.2 44.5 31 19.5 40.6 38.7 11.1 15.8 31.9 25.8 ...
```

```
head(USArrests)
```

```
##           Murder Assault UrbanPop Rape  
## Alabama      13.2      236        58 21.2  
## Alaska       10.0      263        48 44.5  
## Arizona       8.1      294        80 31.0  
## Arkansas      8.8      190        50 19.5  
## California    9.0      276        91 40.6  
## Colorado      7.9      204        78 38.7
```

미국 각 주의 범죄관련 자료이다. 50개의 관측치와 4개의 변수로 구성되어 있다.

주성분분석 R 코드 II

```
apply(USArrests, 2, mean)
```

```
##      Murder  Assault UrbanPop      Rape  
##      7.788   170.760   65.540   21.232
```

```
apply(USArrests, 2, sd)
```

```
##      Murder  Assault UrbanPop      Rape  
##  4.355510  83.337661  14.474763  9.366385
```

```
summary(USArrests)
```

```
##      Murder      Assault      UrbanPop      Rape  
##  Min.   : 0.800   Min.   : 45.0   Min.   :32.00   Min.   : 7.30  
##  1st Qu.: 4.075   1st Qu.:109.0   1st Qu.:54.50   1st Qu.:15.07  
##  Median : 7.250   Median :159.0   Median :66.00   Median :20.10  
##  Mean   : 7.788   Mean   :170.8   Mean   :65.54   Mean   :21.23  
##  3rd Qu.:11.250   3rd Qu.:249.0   3rd Qu.:77.75   3rd Qu.:26.18  
##  Max.   :17.400   Max.   :337.0   Max.   :91.00   Max.   :46.00
```

주성분분석 R 코드 III

각 변수들의 평균과 분산은 매우 다르다. 주성분분석을 할 때, 척도화(scaling)를 하는 것이 필요하다.

```
pr.out=prcomp(USArrests, scale=TRUE)
str(pr.out)

## List of 5
## $ sdev      : num [1:4] 1.575 0.995 0.597 0.416
loading <-- ## $ rotation: num [1:4, 1:4] -0.536 -0.583 -0.278 -0.543 0.418 ...
## ..- attr(*, "dimnames")=List of 2
## .. ..$ : chr [1:4] "Murder" "Assault" "UrbanPop" "Rape"
## .. ..$ : chr [1:4] "PC1" "PC2" "PC3" "PC4"
## $ center    : Named num [1:4] 7.79 170.76 65.54 21.23
## ..- attr(*, "names")= chr [1:4] "Murder" "Assault" "UrbanPop" "Rape"
## $ scale     : Named num [1:4] 4.36 83.34 14.47 9.37
score <-- ## ..- attr(*, "names")= chr [1:4] "Murder" "Assault" "UrbanPop" "Rape"
## $ x         : num [1:50, 1:4] -0.976 -1.931 -1.745 0.14 -2.499 ...
## ..- attr(*, "dimnames")=List of 2
## .. ..$ : chr [1:50] "Alabama" "Alaska" "Arizona" "Arkansas" ...
## .. ..$ : chr [1:4] "PC1" "PC2" "PC3" "PC4"
## - attr(*, "class")= chr "prcomp"
```

주성분분석 R 코드 IV

척도화(scale=TRUE)를 해서 주성분분석을 하였다.
주성분분석의 결과는 5개의 리스트(sdev, rotation, center, scale, x)로 이루어져있다.

노트.

결과의 각 값들은 아래에 설명한다.

주성분분석 R 코드 V

```
pr.out$center
```

```
##      Murder  Assault UrbanPop      Rape  
##      7.788   170.760   65.540    21.232
```

```
pr.out$scale
```

```
##      Murder    Assault  UrbanPop      Rape  
##      4.355510  83.337661 14.474763   9.366385
```

5개의 리스트 중 center와 scale은 주성분분석을 수행하기 전 각 변수들의 평균과 표준편차를 의미한다.

주성분분석 R 코드 VI

```
pr.out$rotation
```

##		PC1	PC2	PC3	PC4
##	Murder	-0.5358995	0.4181809	-0.3412327	0.64922780
##	Assault	-0.5831836	0.1879856	-0.2681484	-0.74340748
##	UrbanPop	-0.2781909	-0.8728062	-0.3780158	0.13387773
##	Rape	-0.5434321	-0.1673186	0.8177779	0.08902432

rotation은 적재 벡터를 포함하고 있다. 즉,

$$\begin{aligned}pc_1 &= -0.535 \times \textit{Murder} - 0.583 \times \textit{Assault} \\ &\quad - 0.278 \times \textit{UrbanPop} - 0.543 \times \textit{Rape} \\ pc_2 &= 0.418 \times \textit{Murder} + 0.187 \times \textit{Assault} \\ &\quad - 0.872 \times \textit{UrbanPop} - 0.167 \times \textit{Rape}\end{aligned}$$

와 같이 나타낼 수 있다.

주성분분석 R 코드 VII

```
dim(pr.out$x)

## [1] 50  4

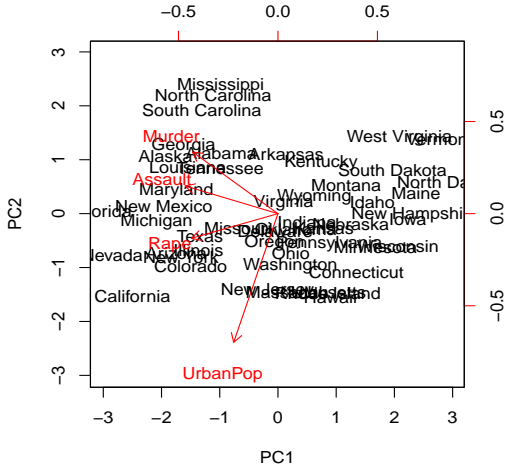
str(pr.out$x)

##  num [1:50, 1:4] -0.976 -1.931 -1.745 0.14 -2.499 ...
##  - attr(*, "dimnames")=List of 2
##    ..$ : chr [1:50] "Alabama" "Alaska" "Arizona" "Arkansas" ...
##    ..$ : chr [1:4] "PC1" "PC2" "PC3" "PC4"
```

x는 주성분을 포함하고 있다. 즉, x[,1]은 첫번째 주성분, x[,2]는 두번째 주성분이다. 각 관측치마다 주성분의 값을 계산하여서 x의 행은 관측치의 개수와 같고 열은 주성분의 개수와 같다.

주성분분석 R 코드 VIII

```
biplot(pr.out, scale=0)
```



주성분분석 R 코드 IX

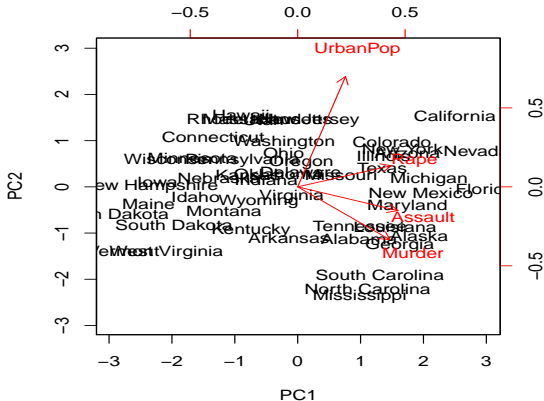
쌍도를 그린 그림이다. 이 그림은

```
plot(pr.out$x[,1], pr.out$x[,2])
```

를 그린 것과 같다. 두번째와 세번째 주성분의 그림을 그리고 싶으면 옵션 `choices = c(2,3)`을 쓰면 된다. 변수는 λ^{scale} 와 같이 관측치는 $\lambda^{1-scale}$ 로 표시된다. `scale=0`는 있는 그대로 척도를 맞추는 것이다. λ 는 주성분분석의 특이값(singular value)를 의미한다.

주성분분석 R 코드 X

```
pr.out$rotation=-pr.out$rotation  
pr.out$x=-pr.out$x  
biplot(pr.out, scale=0)
```



주성분분석 R 코드 XI

주성분의 사인을 바꾼것이다.

```
pr.out$sdev

## [1] 1.5748783 0.9948694 0.5971291 0.4164494

pr.var=pr.out$sdev^2
pr.var

## [1] 2.4802416 0.9897652 0.3565632 0.1734301

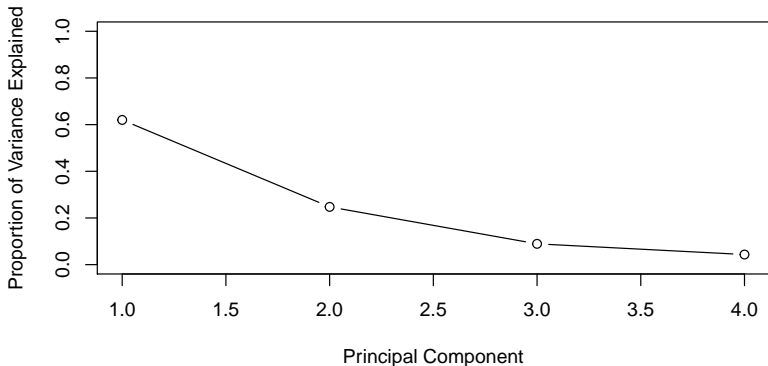
pve=pr.var/sum(pr.var)
pve

## [1] 0.62006039 0.24744129 0.08914080 0.04335752
```

sdev는 주성분의 표준편차이다. 주성분의 분산을 계산하였다.
각 주성분의 설명 비율이다.

주성분분석 R 코드 XII

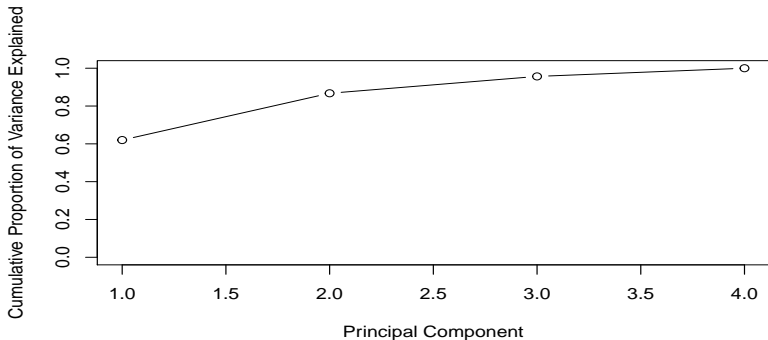
```
plot(pve, xlab="Principal Component",  
     ylab="Proportion of Variance Explained", ylim=c(0,1),type='b')
```



각 주성분의 설명하는 분산 비율의 그림이다.

주성분분석 R 코드 XIII

```
plot(cumsum(pve), xlab="Principal Component",  
     ylab="Cumulative Proportion of Variance Explained", ylim=c(0,1),type='b')
```



주성분의 누적 분산 비율이다.

주성분회귀분석(principal component regression) I

목적

1. Y 를 반응변수로 X_1, \dots, X_p 를 설명변수로 회귀모형을 적합하고자 한다. 이 때, 설명변수의 개수 p 가 클 때, 변수의 개수를 줄이는 방법으로 사용한다.
2. 설명변수들 사이에 공선형성(collinearity, 공선성)이 있을 때 공선형성을 없애는 방법으로 사용되기도 한다.

방법

1. 설명변수 X_1, \dots, X_p 에 주성분분석을 적용한 후 $M \leq p$ 개의 주성분 Z_1, \dots, Z_M 을 새로운 설명변수로 선택한다.
2. 선택된 M 개의 주성분 Z_1, \dots, Z_M 를 설명변수로 Y 를 반응변수로 회귀모형을 적합한다.

주성분회귀분석(principal component regression)

II

가정

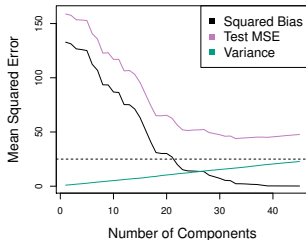
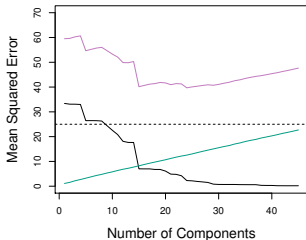
1. 주성분회귀분석은 설명변수들의 분산이 큰 방향이 반응변수와 상관이 높은 방향이라는 가정하에서 사용된다.
2. 이 가정이 항상 옳은 것은 아니지만 종종 성립한다.

주성분회귀분석(principal component regression)

III

주성분 개수의 선택

1. 주성분의 개수 M 이 커질수록 주성분회귀모형의 유연성이 커진다. 따라서 M 이 커질수록 분산은 커지고 편이는 작아진다.
2. M 은 교차검증을 통해 선택한다.



주성분회귀분석(principal component regression)

IV

노트.

능선회귀와 주성분회귀분석은 밀접하게 연결되어 있다.
능선회귀를 주성분회귀모형의 연속형 버전으로 생각할 수
있다고 한다. 이에 대한 자세한 내용은

Friedman, J., Hastie, T., & Tibshirani, R. (2001). The
elements of statistical learning (Vol. 1). Springer, Berlin:
Springer series in statistics.

에 나와 있다고 한다. 무슨 얘기인지 알아보는 것이 좋겠다.

주성분회귀분석 R 코드 I

```
library(pls)

##
## Attaching package: 'pls'
##
## The following object is masked from 'package:stats':
##
##   loadings

library(ISLR)
```

노트.

Partial Least Squares Regression (PLSR), Principal Component Regression (PCR) and Canonical Powered Partial Least Squares (CPPLS)가 있는 패키지이다.

주성분회귀분석 R 코드 II

```
str(Hitters)
head(Hitters)
```

노트.

Hitters는 타자들의 연봉에 대한 자료로 322명의 타자들에 대한 20개의 변수를 포함하였다.

```
set.seed(2)
pcr.fit=pcr(Salary~., data=Hitters,scale=TRUE,validation="CV")
```

```
summary(pcr.fit)
```

주성분회귀분석 R 코드 III

```
> summary(pcr.fit)
Data:      X dimension: 263 19
Y dimension: 263 1
Fit method: svdpc
Number of components considered: 19
```

VALIDATION: RMSEP

Cross-validated using 10 random segments.

	(Intercept)	1 comps	2 comps	3 comps	4 comps
CV	452	348.9	352.2	353.5	352.8
adjCV	452	348.7	351.8	352.9	352.1

....

TRAINING: \% variance explained

	1 comps	2 comps	3 comps	4 comps	5 comps
X	38.31	60.16	70.84	79.03	84.29
Salary	40.63	41.58	42.17	43.22	44.90

....

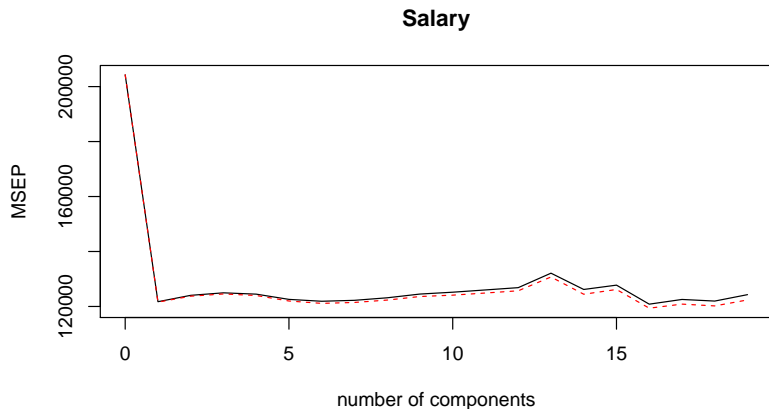
주성분회귀분석 R 코드 IV

함수 pcr의 사용법은 lm 함수와 비슷하다. scale은 설명변수를 척도화해서 주성분분석을 돌리는 것이다. validation="CV"는 주성분의 개수를 정하는데 10겹 교차검증을 사용하게 한다.

```
validationplot(pcr.fit, val.type="MSEP")
```

Mean Squared Error Prediction

주성분회귀분석 R 코드 V



교차검증에러를 그림으로 그린 것이다. $M = 16$ 일 때
교차검증에러가 가장 작지만 $M = 1$ 일 때 이미 에러가 작다.

주성분회귀분석 R 코드 VI

(가 R)

```
set.seed(1)
pcr.fit=pcr(Salary~., data=Hitters,subset=train,scale=TRUE, validation="CV")
validationplot(pcr.fit,val.type="MSEP")
pcr.pred=predict(pcr.fit,x[test,],ncomp=7)
mean((pcr.pred-y.test)^2)
pcr.fit=pcr(y~x,scale=TRUE,ncomp=7)
summary(pcr.fit)
```

훈련자료에 주성분회귀분석을 적합하고 시험오차를 계산하였다. 예측값을 계산할 때는 predict 함수를 사용한다.

3. 인자분석

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 잠재변수의 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

$$x_1 = \lambda_1 f + u_1$$

$$x_2 = \lambda_2 f + u_2$$

$$x_3 = \lambda_3 f + u_3.$$

여기서 f 는 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 모두 한개의 인자에 영향을 받는다. 이 인자는 관측되지 않고 지능이라 해석된다.

인자분석 IV

k인자 모형

$$x_1 = \lambda_{11}f_1 + \lambda_{12}f_2 + \dots + \lambda_{1k}f_k + u_1$$

$$x_2 = \lambda_{21}f_1 + \lambda_{22}f_2 + \dots + \lambda_{2k}f_k + u_2$$

$$\vdots$$

$$x_q = \lambda_{q1}f_1 + \lambda_{q2}f_2 + \dots + \lambda_{qk}f_k + u_q$$

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

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

인자분석 V

가정

1. u_i 는 서로 상관이 없고(not correlated)
2. \mathbf{u}, \mathbf{f} 는 서로 상관이 없다.
3. f_i 도 서로 상관이 없다.

인자분석 VI

k인자 모형 : 행렬식

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

$$\Lambda = \begin{bmatrix} \lambda_{11} & \lambda_{12} & \dots & \lambda_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{q1} & \lambda_{q2} & \dots & \lambda_{qk} \end{bmatrix}, \mathbf{f} = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_k \end{bmatrix}, \mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_q \end{bmatrix}$$

1. $\mathbf{x} = (x_1, \dots, x_q)^T$: 관측변수
2. $\mathbf{f} = (f_1, \dots, f_k)^T, k < q$: 공통인자

인자분석 VII

가정

$$\mathbf{f} \sim N(0, I)$$

$$\mathbf{u} \sim N(0, \Psi), \Psi = \text{diag}(\psi_1, \dots, \psi_k).$$

인자분석 VIII

k인자 모형의 분산들

x_i 의 분산

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

$h_i = \sum_{j=1}^k \lambda_{ij}^2$: 공통성(communality). x_i 의 분산 중 다른 x_j 들과 공유하는 인자 때문에 생기는 분산

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

인자분석 IX

x_i 와 x_j 의 공분산

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

\mathbf{x} 분산의 행렬식

$$\text{Var}(\mathbf{x}) = \Sigma = \Lambda \Lambda^T + \Psi$$

추정

Σ 는 표본분산 S 로 추정이 되고 이를 이용해 Λ, Ψ, k 를 추정해야한다.
상관행렬만 가지고도 인자분석을 수행할 수 있다.

Non-identifiability/Non-uniqueness

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** Λ 과 Ψ 를 반복적으로 추정하는데, Λ 를 추정할 때 주성분분석 방법을 쓴다.

인자분석 XII

노트. 주성분인자분석

다음과 같은 알고리즘으로 Λ 와 Ψ 를 추정한다.

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

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

Step 1.1 $S^{(l)} = S - \Psi^{(l-1)}$ 라 놓는다.

Step 2.2 $S^{(l)}$ 의 고유치값이 큰 k 개의 고유벡터로 적재값을 추정한다. 즉 $\lambda_1, \dots, \lambda_k$ 가 $S^{(l)}$ 의 고유벡터라면,

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

라 놓는다.

Step 3.3 $\psi_i^{(l)} = s_i^2 - \sum_{j=1}^k \lambda_{ij}^2$ 이라 놓는다.

인자분석 XIII

\mathbf{x}_{-i} 는 \mathbf{x} 에서 x_i 를 뺀 모든 변수를 말한다.

노트. 헤이우드 경우

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

인자분석 XIV

노트. 최대가능도인자분석

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

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

이다. 최대가능도 방법은 F 를 최소화하는 Λ 와 Ψ 를 찾는 것이다.
몇 가지 반복적 방법이 존재한다. 이 경우도 헤이우드 경우가 생길 수 있다.

인자분석 XV

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_0 를 1부터 늘려가면 $H_0 : k = k_0$ 를 검정하고, H_1 이 유의하지 않을 때 멈춘다. 이 때의 k_0 값을 인자의 개수로 정한다.

인자분석 R 코드 I

```
> 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

인자분석 R 코드 II

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

인자분석 R 코드 III

노트.

1. 첫 명령은 인자의 개수를 정하는 것이다.
 - 1.1 `sapply`는 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 `factanal`은 인자분석을 수행하는 명령어이다. `method`는 언제나 `mle`라고 한다.
 - 1.4 결과는 각 인자의 개수가 1, 2, 3인 모형 3개를 적합하고 각 모형의 p 값을 출력하는 명령이다. 여기서 1:3을 넣어서 돌아갔는데, 4이상이면 들어가면 에러가 발생한다. 이를 기반으로 `factor`의 개수가 3이라고 결론을 낸다.

인자분석 R 코드 IV

2. 다음의 모형을 적합했다.

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

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

3. 결과에서 uniqueness는 특정분산으로 u 의 분산 대각행렬 Ψ 의 대각원소이다.
4. **Loadings** 는 Λ 행렬을 나타낸다. 빈칸이 있는데 0을 의미하는 것 같다.

인자분석 R 코드 V

5. **SS loadings** 는 각 인자의 공통성 h_i 이다. cumulative 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_i 들의 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 Λ 를 $k \times k$ 직교행렬 M 으로 바꾸어도 자료는 동일하게 설명된다. 즉, 임의의 직교행렬 M 에 대해 적재행렬과 특정분산 ($\Lambda^* = \Lambda M, \Psi$)를 가진 인자모형은 (Λ, Ψ) 를 가진 인자모형과 동일하게 자료를 설명할 수 있다.

노트. 인자의 회전의 증명

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

인자의 해석을 위한 방법 II

의 인자모형이 성립한다고 하자. 그리고 M 은 $k \times k$ 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.
 \mathbf{x} 의 분산도

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

와 같이 동일하게 나타낼 수 있다.

인자의 해석을 위한 방법 III

제약

Λ 와 Ψ 의 추정량이 유일하게 존재하게 하기 위해 필요한 만큼의 제약을 준다. 최대우도 추정량의 경우 계산의 편의성을 위하여

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

가 대각행렬이 되도록 제한한다.

계산된 추정량의 보다 좋은 해석을 위해서 인자를 회전시켜 해석하기 쉬운 추정량값을 찾는다.

인자의 해석을 위한 방법 IV

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 k 개의 0을 포함한다.
3. 임의의 두 개의 컬럼을 뽑았을 때 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 수 이어야 한다.

인자의 해석을 위한 방법 V

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_i 가 한 개의 f_j 와 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.

인자의 해석을 위한 방법 VI

노트.

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

인자(점수)의 추정

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 \mathbf{f} 를 추정한다.

인자의 해석을 위한 방법 VII

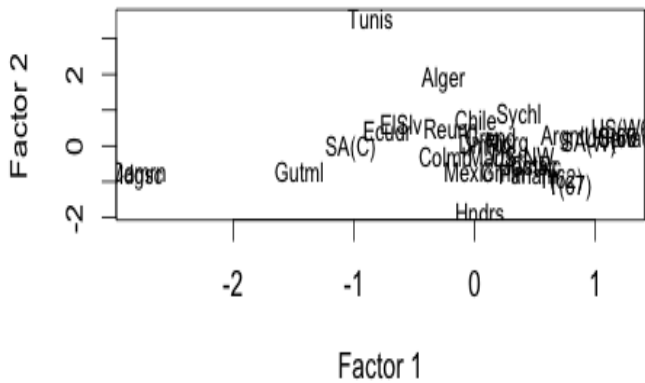
인자점수의 추정이 필요한 이유

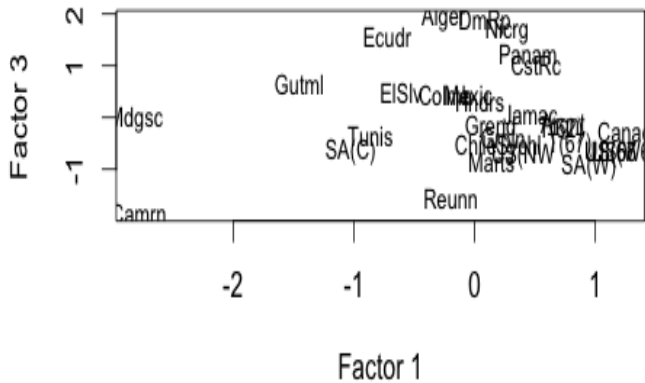
인자는 관측자료의 parsimonious summary이고 이는 이후의 분석에 사용될 수 있다. 2 개의 이유가 더 나와 있는데 무슨 얘기인지 잘 모르겠다.

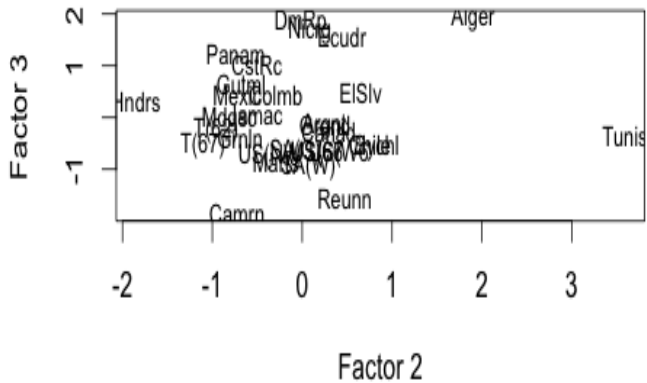
R 코드 I

```
scores <- factanal(life, factors = 3, method = "mle",
                  scores = "regression")$scores

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







참고문헌

아래의 책에서 제공하는 그림들을 사용하였다.

1. Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. [An introduction to statistical learning.](#) Springer, 2013.