7장 주성분 분석

덕성여자대학교 정보통계학과 김 재희

7.1 서론

주성분 분석 분산-공분산 구조를 변수들의 선형결합식(주성분: principal component)으로 설명하고자 하는 접근방법.

주성분 분석의 목적

(1)차원 축소

(2)변동이 큰 축 탐색

(3)주성분을 통한 데이터의 해석 등.

반응변수들의 선형결합식으로 형성된 주성분은 서로 독립적인 새로운 변수.

p개 변수에 포함된 전체 변동을 m개($m \leq p$) 주성분으로 대신하여 설명한다.

▶ 주성분분석은 Pearson(1901)에 의해 처음 연구되기 시작.

▶ Hotelling(1933) 이래 더욱 발전.

▶ Pearson(1901)은 p-차원 공간에 놓이는 데이터에 대해 가장 잘 적합하는 직선이나 평면을 찾는데 관심이 있었고 기하학적 최적화 문제로 접근하였다.

▶ 32년 후 Hotelling(1933)은 원래 p개 변수의 값을 대신 결정할 수 있는 저차원의 독립변수들의 집합이 존재할 것이라는 아이디어로부터 출발하여 원래 변수들의 분산을 최대로 설명하는 성분을 구하게 되었고 이와 같은 성분들을 '주성분(principal component)'이라고 부르게 되었다.

▶ Girshick(1939)은 표본주성분 계수에 대한 근사적인 분포를 밝혔으며 그 이후로 주성분분석의 이론과 응용이 활발히 진행되어 농학, 생물학, 화학, 기후학, 인구학, 생태학, 경제학, 식품영양학, 지질학, 기상학, 해양학, 심리학 등 여러 분야에서 꾸준하게 응용되고 있다.

7.2 주성분의 정의와 개념

$p \times 1$ 확률벡터 $\boldsymbol{X} = (X_1, X_2, ..., X_p)'$ 가 모평균벡터 μ 와 모공분산행렬 Σ 를 가진다고 하자.

: 수학적으로 p-차원에 놓이게 되며 원래 변수의 선형결합 또는 회전 변환을 통해 p개의 새로운 좌표축을 형성. 데이터의 변동을 최대로 설명해주는 동시에 공분산 구조에 대한 해석을 용이하게 하도록 만들어질 수 있는데 이것을 주성분(principal component)이라 한다.

▶ 첫 번째 주성분(제1 주성분): 변동(variability)을 최대로 설명해주는 방향으로 변수들의 선형결합식.

▶ 두 번째 주성분(제2 주성분): 첫 번째 주성분 다음으로 변동을 가장 많이 설명해주는 변수들의 선형결합식이며 첫 번째 주성분과는 독립이다.

▶ 이와 같이 찾아진 p개의 주성분들은 새로운 축을 형성하며, 주성분과 이들이 설명하는 변동량, 주성분, 주성분점수 등은 변수들로 표현된 시스템에 대한 이해를 돕는다.

7.2.1 주성분의 정의

확률벡터 $X' = (X_1, X_2, ..., X_p)$ 의 공분산행렬 Σ 는

고유값 $\lambda_1 \geqq \lambda_2 \geqq \cdots \geqq \lambda_p \geqq 0$ 고유값에 해당하는 고유벡터 $e_1, e_2, ..., e_p$를 가진다.

다음과 같은 p개의 선형결합식을 생각해 보자.

$$
\begin{aligned}
Y_1 &= l_1'X = l_{11}X_1 + l_{21}X_2 + ... + l_{p1}X_p \\
Y_2 &= l_2'X = l_{12}X_1 + l_{22}X_2 + ... + l_{p2}X_p \\
&\vdots \qquad\qquad\qquad \vdots \\
Y_i &= l_i'X = l_{1i}X_1 + l_{2i}X_2 + ... + l_{pi}X_p \\
&\vdots \qquad\qquad\qquad \vdots \\
Y_p &= l_p'X = l_{1p}X_1 + l_{2p}X_2 + ... + l_{pp}X_p
\end{aligned}
$$

여기서 $l_i' = (l_{1i}, l_{2i}, ..., l_{pi})$이며 선형결합식들의 분산은

$$Var(Y_i) = l_i'\Sigma l_i, \quad i = 1, ..., p$$

$$Cov(Y_i, Y_k) = l_i'\Sigma l_k, \; i, k = 1, ..., p$$

■ 주성분 구하는 과정

(1) 첫 번째 주성분은 $l_1'l_1 = 1$을 만족하는 $p \times 1$ 벡터 l_1에 대해

$Var(l_1'X)$를 최대로 하는 선형결합식 $l_1'X$로 구한다.

(2) 두 번째 주성분은 $l_2'l_2 = 1$을 만족하는 $p \times 1$ 벡터 l_2에 대해

$Var(l_2'X)$를 최대로 하며 또한 $Cov(l_1'X, l_2'X) = 0$인 선형결합식 $l_2'X$로 구한다.

(i) i번째 주성분은 $l_i'l_i = 1$을 만족하는 n 벡터 l_i에 대해

$Var(l_i'X)$를 최대로 하며, 앞에서 구한 $(i-1)$개의 주성분들과는 직교하도록

$Cov(l_i'X, l_j'X) = 0,\ j = 1, \cdots, i-1$ 인 선형결합식 $l_i'X$로 구한다.

⋮

(p) 마지막으로 p번째 주성분은 $l_p'l_p = 1$을 만족하는 $p \times 1$ 벡터 l_p에 대해

$Var(l_p'X)$를 최대로 하며, 앞에서 구한 $(p-1)$개의 주성분들과는 직교하도록

$Cov(l_p'X, l_p'X) = 0,\ j = 1, \cdots, p-1$ 인 선형결합식 $l_p'X$로 구한다.

정리 7.1 확률벡터 $\boldsymbol{X}' = (X_1, X_2, ..., X_p)$의 공분산행렬 $\boldsymbol{\Sigma}$는 고유값 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$을 갖고 각 고유값에 해당하는 고유벡터 $\boldsymbol{e}_1, \boldsymbol{e}_2, ..., \boldsymbol{e}_p$를 갖는다.

i번째 주성분은

$$Y_i = \boldsymbol{e}_i{}' \boldsymbol{X} = e_{1i}X_1 + e_{2i}X_2 + ... + e_{pi}X_p$$

이고 i번째 주성분의 분산, 다른 주성분과의 공분산은

$$Var(Y_i) = \boldsymbol{e_i}' \boldsymbol{\Sigma} \boldsymbol{e_i} = \lambda_i, \quad i = 1, 2, ..., p$$

$$Cov(Y_i, Y_j) = \boldsymbol{e}_i{}' \boldsymbol{\Sigma} \boldsymbol{e}_j = 0, \quad i \neq j \quad .$$

정리 7.2 전체변이는 공분산행렬 Σ의 고유값들의 합으로 표현된다. 즉,

$$\sum_{i=1}^{p} Var(X_i) = \sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp} = \lambda_1 + \lambda_2 + \cdots + \lambda_p = \sum_{i=1}^{p} Var(Y_i)$$

그러므로 $_i$번째 주성분에 의해 설명되는 전체분산의 비율은

$$\frac{i\text{번째 주성분에 의해 설명되는 분산}}{\text{전체 분산}} = \frac{\lambda_i}{\lambda_1 + \ldots + \lambda_p}, \quad i = 1,2,\ldots,p$$

정리 7.3 Y_i($_i$번째 주성분)와 X_k(원래 데이터의 k번째 확률변수)의 상관계수는

$$\rho_{Y_i, X_k} = \frac{e_{ki}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}}, \qquad i,k = 1,\ldots,p$$

이다. 여기서 e_{ki}는 $\boldsymbol{e}_i$의 k번째 성분이다.

≪예제 7.1≫ 두 변수 X_1과 X_2에 대해 관측값을 얻은 후 표본공분산행렬 $S=\begin{pmatrix}7&1\\1&7\end{pmatrix}$를 얻었을 때 주성분을 구해보자.

S의 고유값을 구하기 위해

$$|S-\lambda I|=\begin{vmatrix}7-\lambda&1\\1&7-\lambda\end{vmatrix}=(7-\lambda)^2-1=\lambda^2-14\lambda+48=(\lambda-6)(\lambda-8)=0$$

을 풀면 고유값은 $\lambda_1=8$과 $\lambda_2=6$로, 이에 대한 고유벡터를 다음과 같이 얻는다.

(i) $\lambda_1=8$

$$S-\lambda_1 I=\begin{pmatrix}7&1\\1&7\end{pmatrix}-\begin{pmatrix}8&0\\0&8\end{pmatrix}=\begin{pmatrix}-1&1\\1&-1\end{pmatrix},\qquad \begin{pmatrix}-1&1\\1&-1\end{pmatrix}\begin{pmatrix}y\\z\end{pmatrix}=\begin{pmatrix}0\\0\end{pmatrix}$$

을 풀면 된다. 즉 연립방정식 $y-z=0$을 풀어 고유벡터 $x_1=\begin{pmatrix}y\\z\end{pmatrix}=\begin{pmatrix}1\\1\end{pmatrix}$을 얻는다.

$L_{x_1}=\sqrt{2}$ 이므로 단위고유벡터는 $e_1=\begin{pmatrix}1/\sqrt{2}\\1/\sqrt{2}\end{pmatrix}$ 가 된다.

(ii) $\lambda_2 = 6$

$$S - \lambda_2 I = \begin{pmatrix} 7 & 1 \\ 1 & 7 \end{pmatrix} - \begin{pmatrix} 6 & 0 \\ 0 & 6 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \quad \text{이므로} \quad \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

즉 $y + z = 0$ 을 풀면 된다. $z = -y$ 이므로 고유벡터 $x_2 = \begin{pmatrix} y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$ 을 얻는다.

$L_{x_2} = \sqrt{2}$ 이므로 단위 고유벡터는 $e_2 = \begin{pmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix}$ 가 된다.

첫 번째와 두 번째 고유벡터는 각각 $e_1 = (0.707, 0.707)'$, $e_2 = (0.707, -0.707)'$이므로

첫번째 주성분 : $Y_1 = 0.707X_1 + 0.707X_2$ 두변수의가중평균형태
두번째 주성분 : $Y_2 = 0.707X_1 - 0.707X_2$ 두변수의대비$(contrast)$

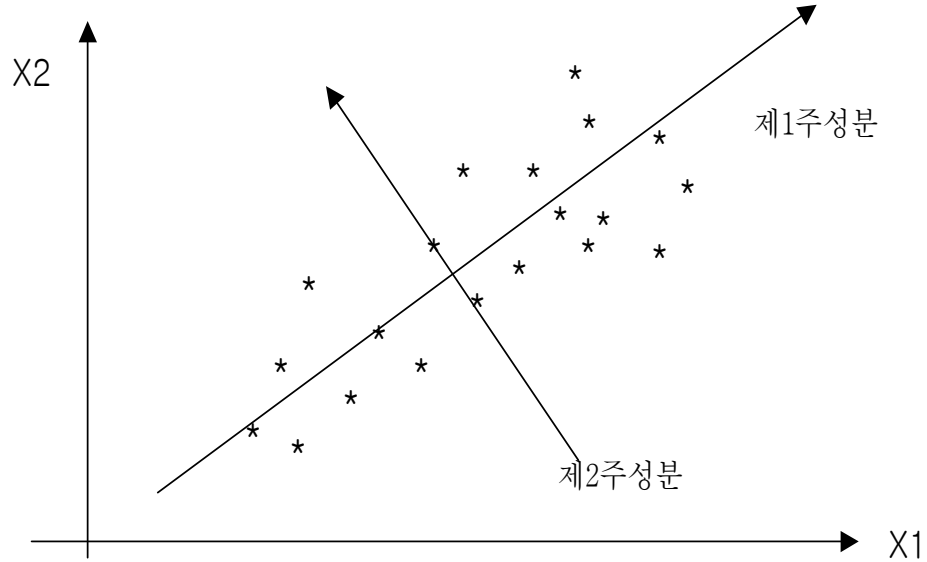
첫 번째 주성분이

$$\frac{8}{8+6} = 0.571$$

으로 전체 분산의 57% 정도를 설명한다.

7.2.2 주성분의 기하학적 의미

- 제1주성분 축의 데이터 변이에 비해 제2주성분 축의 데이터 변이가 상대적으로 작다.
- 주성분 좌표축의 데이터는 원래 좌표축의 변수에 비해 랜덤하게 펼쳐지게 된다.
- 서로 직교하는 주성분들은 통계적인 추론과 해석에 많은 편리한 점을 제공한다.



[그림 7.1] 데이터와 주성분

▶ 다변량 정규분포를 따르는 확률변수의 경우 주성분의 의미를 살펴보자.

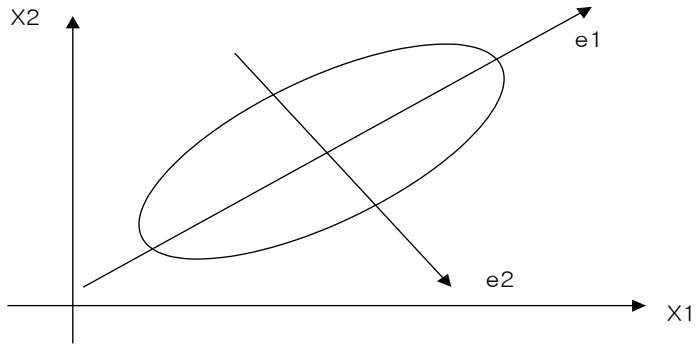
$X \sim N_p(\mu, \Sigma)$를 따를 때

$(x-\mu)'\Sigma^{-1}(x-\mu) = c^2$는 μ를 중심으로 축이 $\pm c\sqrt{\lambda_i}\, e_i$

인 타원체(ellipsoid). 일반성을 잃지 않고 $\mu = 0$ 으로 놓자.

$$c^2 = x'\Sigma^{-1}x = \frac{1}{\lambda_1}(e_1'x)^2 + \frac{1}{\lambda_2}(e_2'x)^2 + \ldots + \frac{1}{\lambda_p}(e_p'x)^2 = \sum_{i=1}^{p}\frac{y_i^2}{\lambda_i}$$

여기서 $y_i = e_i'x$, $e_1, \ldots, e_p$ 방향으로의 축. λ_1이 가장 큰 고유값일 때 주축은 e_1 방향이 된다.



[그림 7.2] $\rho = 0.5$인 경우 등고선 타원에서의 제1 주성분 e_1과 제2 주성분 e_2

7.3 상관행렬을 이용한 주성분분석

▶ 공분산행렬 S 를 사용할 경우 분산이 큰 변수가 주성분의 압도적인 비중을 차지할 수 있으므로 분석의 균형을 유지하기 위해서도 표본상관행렬 R 을 이용할 필요가 있다.

▶ 변수 표준화

$$Z_1 = \frac{X_1 - \mu_1}{\sqrt{\sigma_{11}}}$$
$$Z_2 = \frac{X_2 - \mu_2}{\sqrt{\sigma_{22}}}$$
$$\vdots$$
$$Z_p = \frac{X_p - \mu_p}{\sqrt{\sigma_{pp}}}$$

벡터로 표현하면

$$Z = (V^{1/2})^{-1}(X-\mu)$$

여기서

$$V^{1/2} = \begin{pmatrix} \sqrt{\sigma_{11}} & 0 & \cdots & 0 \\ 0 & \sqrt{\sigma_{22}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{\sigma_{pp}} \end{pmatrix}$$

$$Y_i = e_i' Z = e_i' (V^{1/2})^{-1}(X-\mu)$$

$E(Z) = 0,\ Cov(Z) = I$

$V^{1/2}\rho V^{1/2} = \Sigma \qquad\qquad \rho = V^{-1/2}\Sigma V^{-1/2}$

즉, Σ는 분산의 대각행렬 $V^{1/2}$과 상관행렬 ρ 로부터 얻어진다.

정리 7.4 표준화 변수들로 구성한 벡터 $Z' = (Z_1, ..., Z_p)$에 대해 주성분을 구하면

i번째 주성분은

$$Y_i = e_i' Z = e_i' (V^{1/2})^{-1} (X - \mu)$$

이 된다. 주성분에 대한 총분산은

$$\sum_{i=1}^{p} Var(Y_i) = \sum_{i=1}^{p} Var(Z_i) = 1 + \cdots + 1 = p$$

이며, i번째 주성분 Y_i와 k번째 표준화 변수 Z_k의 상관계수는

$$\rho_{Y_i, Z_k} = \frac{e_{ki}\sqrt{\lambda_i}}{\sqrt{Var(Z_k)}} = e_{ki}\sqrt{\lambda_i}$$

이 된다. 또한 표준화 변수로부터 구한 주성분에 대하여 i번째 주성분에 의해 설명되는 전체분산의 비율은

$$\frac{i\text{번째 주성분에 의해 설명되는 분산}}{\text{전체 분산}} = \frac{\lambda_i}{p}, \quad i = 1, 2, ..., p$$

▶ 주성분분석에서 표본상관행렬 R 을 사용할 경우 몇 가지 주의점을 살펴보자.

(i) R 에 의한 주성분과 S에 의한 주성분이 설명하는 분산의 양이 다르다.

(ii) R 에 의한 주성분과 S에 의한 주성분 계수가 다르다.

(iii) R 자체가 척도 불변이므로 R 에 의한 주성분은 척도 불변이다(scale invariant).

(iv) R 에 의한 주성분은 유일하지 않다.

≪예제 7.2≫ 다음의 이변량 데이터에 대해 표본상관행렬 R 을 이용하여 주성분을 구해보자.

$$R=\begin{pmatrix}1 & r\\ r & 1\end{pmatrix}, \qquad r>0$$

고유값을 구하면 $\lambda_1=1+r,\ \lambda_2=1-r$이고

첫 번째 고유벡터는 $\quad e_1=(0.707,0.707)',$

두 번째 고유벡터는 $\quad e_2=(0.707,-0.707)'$

첫 번째 주성분, 두 번째 주성분은 각각

$$Y_1=0.707\left(\frac{X_1-\overline{X_1}}{s_1}\right)+0.707\left(\frac{X_2-\overline{X_2}}{s_2}\right)$$

$$Y_2=0.707\left(\frac{X_1-\overline{X_1}}{s_1}\right)-0.707\left(\frac{X_2-\overline{X_2}}{s_2}\right)$$

주성분을 구성하는 각 계수는 r에 의존하지 않는다. 즉 $r=0.01$ 이든 $r=0.99$ 이든 주성분의 계수에는 영향을 미치지 않고 분산의 설명비율에만 영향을 미친다.

상관행렬 R 로부터의 주성분은 상관계수들의 상대적인 비(relative ratio)에 의존한다.

≪예제 7.3≫ 두 변수 X_1과 X_2를 갖는 이변량 데이터에 대해

표본공분산행렬 $S=\begin{pmatrix}1 & 6\\ 6 & 100\end{pmatrix}$, 표본상관행렬 $R=\begin{pmatrix}1 & 0.6\\ 0.6 & 1\end{pmatrix}$ 을 얻었을 때 각각 주성분을 구하고 비교해보자.

S의 고유값을 구하면

$$|S-\lambda I|=\begin{vmatrix}1-\lambda & 6\\ 6 & 100-\lambda\end{vmatrix}=(1-\lambda)(100-\lambda)-36=\lambda^2-101\lambda+64=0$$

을 풀어 고유값과 고유벡터를 구하면

$$\lambda_1=100.363,\ e_1{}'=(0.060,0.998)$$
$$\lambda_2=0.64,\qquad e_2{}'=(0.998,-0.060)$$

X_2의 분산이 X_1의 분산의 100배이므로 첫 번째 주성분인 $Y_1=0.060X_1+0.998X_2$ 가 전체분산에서 설명하는 비율은

$$\frac{\lambda_1}{\lambda_1+\lambda_2}=\frac{100.363}{101}=0.9937$$

$\boldsymbol{R}$의 고유값을 구하기 위해

$$|\boldsymbol{R}-\lambda \boldsymbol{I}| = \begin{vmatrix} 1-\lambda & 0.6 \\ 0.6 & 1-\lambda \end{vmatrix} = (1-\lambda)^2 - 0.36 = \lambda^2 - 2\lambda + 0.64 = 0$$

을 풀어 고유값과 고유벡터를 구하면

$$\begin{aligned} \delta_1 &= 1.6, \qquad \boldsymbol{f}_1{}' = (0.707, \quad 0.707) \\ \delta_2 &= 0.4, \qquad \boldsymbol{f}_2{}' = (0.707, -0.707) \end{aligned}$$

이다. 표준화된 X_1의 분산과 표준화된 X_2의 분산이 같으며 첫 번째 주성분인 $U_1 = 0.707Z_1 + 0.707Z_2$ 가 전체분산에서 설명하는 비율:

$$\frac{\delta_1}{\delta_1 + \delta_2} = \frac{1.6}{2} = 0.8$$

표본공분산행렬로부터 구한 첫 번째 주성분의 설명 비율보다 낮아지며 차이가 난다.

주성분 U_1과 U_2를 원래 변수 X_1과 X_2로 표현하면

$$U_1 = 0.707\left(\frac{X_1 - \overline{X_1}}{1}\right) + 0.707\left(\frac{X_2 - \overline{X_2}}{10}\right)$$
$$= 0.707X_1 + 0.07X_2 - 0.707\left(\overline{X_1} + \frac{\overline{X_2}}{10}\right),$$

$$U_2 = 0.707X_1 - 0.07X_2 - 0.707\left(\overline{X_1} - \frac{\overline{X_2}}{10}\right)$$

과 같이 구할 수 있어 S로부터 구한 주성분과는 차이가 남을 알 수 있다.

또한 $0.707(0.707) - 0.07(0.07) = 0.49 \neq 0$ 으로

원래 변수로 표현했을 때 두 주성분이 직교하지 않음을 알 수 있다.

7.4 주성분에 의한 표본 변동 설명

▶ 모평균벡터 μ 와 모공분산행렬 Σ 를 가진 모집단으로부터

n개의 p-차원 확률표본 $X' = (X_1, X_2, ..., X_p)$에 대한 표본주성분을 구하고자 한다.

▶ 분석의 목적 : 표본주성분을 이용해 모집단 주성분의 추정치를 구하여 모집단에 대한 해석을 하는 것이다.

▶ 추정 과정: 모집단 공분산행렬 Σ 의 추정량인 표본공분산행렬 S 를 이용

또한 모집단 상관행렬 ρ의 추정량인 표본상관행렬 R 을 주성분분석에 이용한다.

▶ 이렇게 추정된 표본주성분은 확률변수가 되어 모집단 주성분으로의 확률적 수렴 등의 통계적 성질을 밝힐 수 있다.

▶ 확률표본 $X_1, X_2, ..., X_n$에 대한 표본공분산행렬 $S = \{s_{ij}\}$는 고유값

$\hat{\lambda}_1 > \hat{\lambda}_2 > \cdots > \hat{\lambda}_p > 0$와 각 고유값에 해당하는 고유벡터 $\hat{e}_1, \hat{e}_2, ..., \hat{e}_p$를 가진다.

즉 S의 고유값-고유벡터를 다음과 같이 짝지을 수 있다: $(\hat{\lambda}_1, \hat{e}_1), ..., (\hat{\lambda}_p, \hat{e}_p)$.

정리 7.5 i번째 표본주성분은

$$Y_i = \hat{\boldsymbol{e}}_i' \boldsymbol{X} = \hat{e}_{1i} X_i + \hat{e}_{2i} X_2 + ... + \hat{e}_{\pi} X_p, \qquad i = 1,2,...,p$$

이고, i번째 표본주성분의 분산, 다른 주성분과의 공분산은

$$Var(Y_i) = \hat{\lambda}_i, \quad i = 1,2,...,p,$$

$$Cov(Y_i, Y_j) = 0, \qquad i \neq j$$

표본총분산 $= \sum_{i=1}^{p} s_{ii} = \hat{\lambda}_1 + \hat{\lambda}_2 + ... + \hat{\lambda}_p$ 가 되며,

i번째 표본주성분 Y_i 와 X_k (원래 데이터의 k 번째 확률변수)의 상관계수는

$$Corr(Y_i, X_k) = \frac{\hat{e}_{ki}\sqrt{\hat{\lambda}_i}}{\sqrt{s_{kk}}}, \qquad i,k = 1,2,...,p$$

이다.

7.5 표본고유값과 표본고유벡터의 대수적 성질

7.5.1 표본고유값과 표본고유벡터의 분포

- 표본주성분은 근사적으로 모집단 주성분을 추정한다.
- 모집단 주성분은 Σ 또는 ρ 로부터 얻어진 (λ_i, e_i)에 의존한다.
- 표본주성분은 S 또는 R 로부터 얻어진 $(\hat{\lambda}_i, \hat{e}_i)$에 의존한다.
- Anderson (1963)과 Girshick(1939): 표본고유값과 표본고유벡터에 대한 성질 규명

정리 7.6 (표본고유값과 표본고유벡터에 대한 성질)

확률표본 $\boldsymbol{X}_1,\boldsymbol{X}_2,...,\boldsymbol{X}_n$은 공분산행렬 $\boldsymbol{\Sigma}$를 가지는 p-변량 정규분포를 따른다.

$\boldsymbol{\Sigma}$는 고유값 $\lambda_1 > \lambda_2 > \cdots > \lambda_p > 0$와 각 고유값에 해당하는 고유벡터 $\boldsymbol{e}_1,\boldsymbol{e}_2,...,\boldsymbol{e}_p$

(i) $\boldsymbol{\lambda}' = (\lambda_1,...,\lambda_p)$일 때 $n \to \infty$이면 근사적으로

$$\sqrt{n}(\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}) \sim N_p(\boldsymbol{0}, 2\boldsymbol{\Lambda}^2)$$

이다. 여기서

$$\boldsymbol{\Lambda} = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & \cdots & \lambda_p \end{pmatrix}$$

근사적으로

$$\hat{\lambda}_i \sim N\left(\lambda_i, \frac{2\lambda_i^2}{n-1}\right)$$

을 따른다.

(ii) $n \to \infty$ 이면, 근사적으로

$$\sqrt{n}(\hat{e}_i - e_i) \sim N_p(0, E_i)$$

여기서 $E_i = \lambda_i \sum_{k=1, k \neq i}^{p} \frac{\lambda_k}{(\lambda_k - \lambda_i)^2} e_k e_k'$ 이다.

(iii) 표본고유값 $\hat{\lambda}_i$와 표본고유벡터 $\hat{e}_i$는 서로 독립이다.

(iv) 표본고유값벡터 $\hat{\lambda}$와 표본고유벡터 $\hat{e}$는 서로 독립이다.

(v) 모집단 고유값 λ_i에 대한 $100(1-\alpha)\%$ 신뢰구간은

$$\frac{\hat{\lambda}_i}{1+z_{\alpha/2}\sqrt{\frac{2}{n-1}}} \leq \lambda_i \leq \frac{\hat{\lambda}_i}{1-z_{\alpha/2}\sqrt{\frac{2}{n-1}}} \quad .$$

왜냐하면 근사적으로

$$\frac{\hat{\lambda}_i-\lambda_i}{\lambda_i\sqrt{\frac{2}{n-1}}} \sim N(0,1)$$

이 되므로

$$P\left(\left| \frac{\hat{\lambda}_i-\lambda_i}{\lambda_i\sqrt{\frac{2}{n-1}}} \right| \leqq z_{\alpha/2} \right) = 1-\alpha$$

여기서 z_α는 $N(0,1)$을 따르는 확률변수 Z에 대해 $P(Z \geqq z_\alpha) = \alpha$ 를 만족하는 값이다.

7.5.2 등상관구조에 대한 검정

두 변수간의 상관관계가 ρ로 동일한 경우, Σ의 고유값이 두 종류로만 구해지며 중복 고유값이 여러 개 있게 되어, 지금까지의 주성분에 대한 결과들이 유효하지 않다. 식(7.32)의 귀무가설을 만족하는 등상관구조의 공분산행렬을 가진다면 고유값은

$$\sigma^2[1+(p-1)\rho] = \lambda_1 > \lambda_2 = \lambda_3 = \ldots = \lambda_p = \sigma^2(1-\rho)$$

이 되어 $(p-1)$개의 고유값이 같아지므로 첫 번째 주성분이 중요 변동을 설명하고 나머지 주성분은 같은 양의 분산을 설명하는 '잡음(noise)'으로 간주되기도 한다.

▶ 등상관구조(equal correlation structure)를 가지는지 검정

$$H_0 : \rho = \rho_{0_{p\times p}} = \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix} \text{에 대해 } H_1 : \rho \neq \rho_0$$

가설검정을 하기 위해 우선 표본상관행렬

$$\boldsymbol{R} = \begin{pmatrix} 1 & r_{12} \cdots r_{1p} \\ r_{21} & 1 & \cdots r_{2p} \\ \vdots & \vdots & \vdots \\ r_{p1} & r_{p2} \cdots & 1 \end{pmatrix}$$

비대각선 요소들의 평균

$$\overline{r_k} = \frac{1}{p-1} \sum_{i=1}^{p} r_{ik}, \qquad i \neq k$$

전체 평균

$$\bar{r} = \frac{2}{p(p-1)} \sum\sum_{i<k} r_{ik} \ , \qquad \hat{\gamma} = \frac{(p-1)^2[1-(1-\bar{r})^2]}{p-(p-2)(1-\bar{r})^2}$$

구한 후

검정통계량

$$T = \frac{(n-1)}{(1-\bar{r})^2}\left[\sum\sum_{i<k}(r_{ik}-\bar{r})^2 - \hat{\gamma}\sum_{k=1}^{p}(\overline{r_k}-\bar{r})^2\right]$$

은 H_0 하에서 근사적으로 카이제곱분포를 따른다.

유의수준 α에서 검정법은

$T \geqq \chi^2_{(p+1)(p-2)/2}(\alpha)$이면 H_0를 기각한다.

[참고] $\rho = \rho_{0_{p\times p}} = \begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix}$

일 경우 고유값을 구하면 $\lambda_1 = 1+2\rho,\ \lambda_2 = \lambda_3 = 1-\rho$ (중근)이 된다.

7.6 주성분 그래프

▶ 2차원 좌표축에 표현할 수 있는 첫 번째 주성분과 두 번째 주성분은 원래 데이터에 대한 중요한 특성을 나타낼 수 있다.

▶ 주성분 그래프(principal component graph)로부터 두 주성분간의 관계와 패턴을 도출할 수 있으며 또한 전체 데이터가 주성분을 통해 변화되어 나타내는 관계도 알 수 있다.

7.7 주성분 개수의 선택

(1) 전체 변이에의 공헌도(percentage of total variance)

: 전체 변이의 70 ~ 90%가 되도록 주성분의 수를 결정한다

i번째 주성분이 설명하는 분산 양을 λ_i라 하자.

▶ m개의 주성분을 결정하고자 할 때,

(i) 공분산행렬을 이용할 경우

$$100 \times \frac{\sum_{i=1}^{m} \lambda_i}{\sum_{i=1}^{p} \lambda_i} = 100 \times \frac{\sum_{i=1}^{m} \lambda_i}{\sum_{i=1}^{p} s_{ii}}$$

(ii) 상관행렬을 이용한 경우

$$\frac{100}{p} \times \sum_{i=1}^{m} \lambda_i$$

이 70 ~ 90%가 되는 m으로 정한다.

(2) 평균 고유값(average eigenvalues)

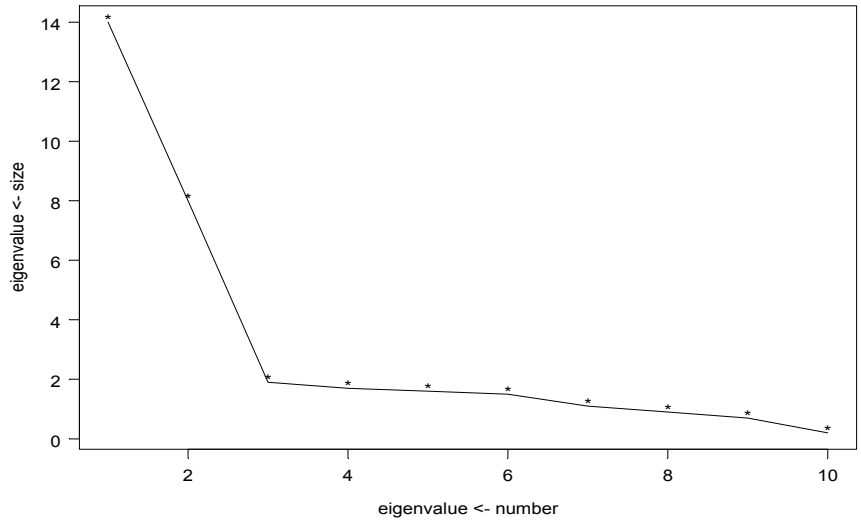
고유값들의 평균 $\bar{\lambda} = \sum_{i=1}^{p} \frac{\lambda_i}{p}$ 을 구한 후 고유값이 평균값 이상이 되는 주성분을 설정한다.

▷ 상관행렬을 사용한 경우 평균 고유값은 1 이 된다.

(3) 스크리 그래프(scree graph)

▲ 2차원 좌표축에 (고유값순서, 고유값크기)으로 (i, λ_i) 점을 찍고 점간을 선분으로 연결

▲ 가파른 정도를 보고 큰 고유값과 작은 고유값을 구분하여 자연스럽게 적절한 개수를 정한다.



[그림 7.3] 이상적인 스크리그래프

7.8 마지막 주성분으로부터의 정보

예를 들어, 마지막 주성분이 설명하는 분산의 양이 0 또는 거의 0이라고 하자. 이들 변수들 간의 선형 관계를 나타내는 주성분이 상수값을 가진다는 의미이고 변수들간의 공선성(collinearity)을 나타내어 유용한 정보로 활용될 수 있다. 즉 X_i 가 표준화 된 변수로 평균이 0이라면 Y_5의 분산설명량이 0일 때

$$Y_5 = e_{51}X_1 + e_{52}X_2 + \cdots + e_{55}X_5 = 0$$

이 되어 X_5가 X_1, X_2, X_3, X_4에 의존함을 알 수 있다.

7.9 주성분에 대한 해석

▶ 주성분은 변수들의 선형결합식으로 구해지기 때문에 주성분을 구성하는 변수들의 계수 구조를 파악하여 적절하게 해석되어야하며 명확한 정형화된 해석 방법이 있는 것은 아니다.

≪예제 7.4≫ 다음의 [표 7.1]은 38명의 학생(여자 20명, 남자 18명)을 대상으로 신체적 크기를 조사하고 표본상관행렬을 이용하여 주성분분석을 한 결과 얻은 주성분계수들이다.

[표 7.1] 신체 해부학적 자료에 대한 주성분분석결과

	여자				남자		
주성분번호	1	2	3	주성분번호	1	2	3
손(hand)	0.33	0.56	0.03	손(hand)	0.23	0.62	0.64
손목(wrist)	0.26	0.62	0.11	손목(wrist)	0.29	0.53	-0.42
키(height)	0.40	-0.44	-0.00	키(height)	0.43	-0.20	0.04
앞팔(forearm)	0.41	-0.05	-0.12	앞팔(forearm)	0.33	-0.53	0.38
머리(head)	0.27	-0.19	0.80	머리(head)	0.41	-0.09	-0.51
가슴(chest)	0.45	-0.26	-0.55	가슴(chest)	0.44	0.08	-0.01
허리(waist)	0.47	0.03	-0.03	허리(waist)	0.46	-0.07	0.09
고유값	3.72	1.37	0.97	고유값	4.17	1.26	0.66
누적변동비율 (%)	53.2	72.7	86.5	누적변동비율 (%)	59.6	73.6	87.0

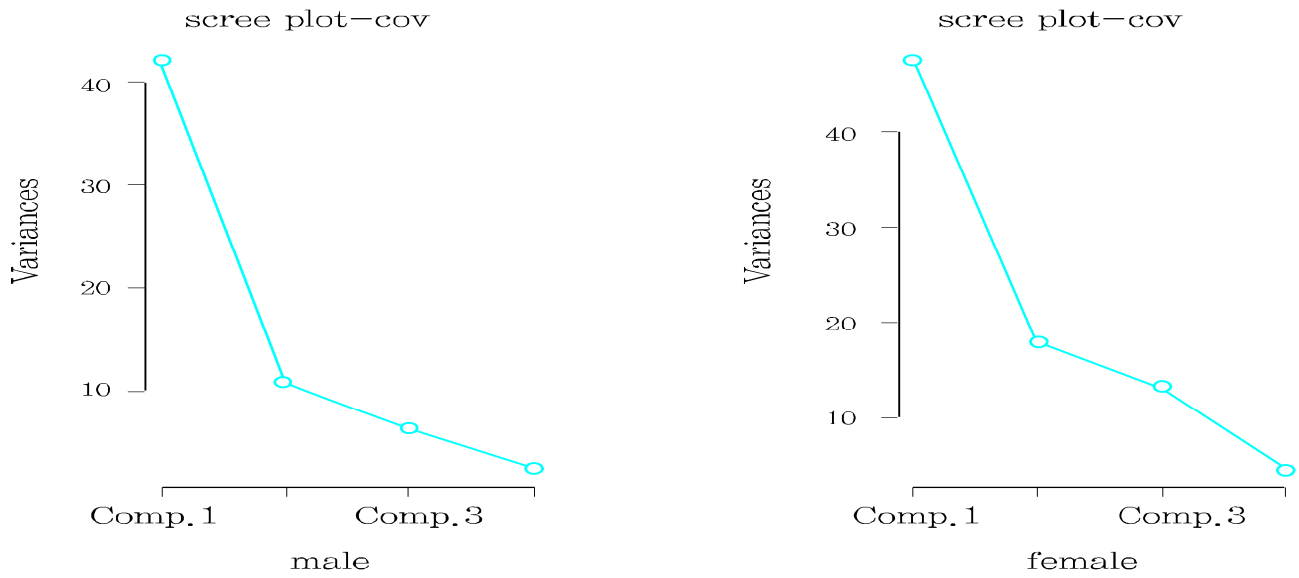
[표 7.2] [표 7.1] 결과에 대한 간략표

	여자				남자		
주성분번호	1	2	3	주성분번호	1	2	3
손(hand)	+	+		손(hand)	+	+	+
손목(wrist)	+	+		손목(wrist)	+	+	−
키(height)	+	−		키(height)	+	(−)	
앞팔(forearm)	+			앞팔(forearm)	+	−	+
머리(head)	+	(−)	+	머리(head)	+		−
가슴(chest)	+	(−)	−	가슴(chest)	+		
허리(waist)	+			허리(waist)	+		
누적변동비율 (%)	53.2	72.7	86.5	누적변동비율 (%)	59.6	73.6	87.0

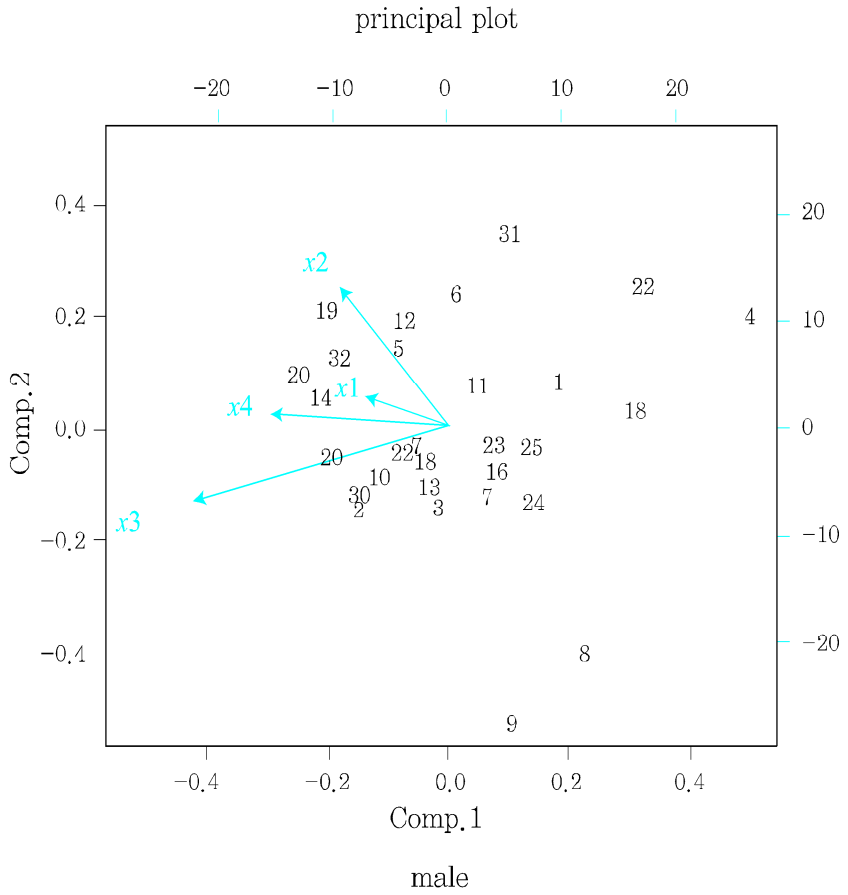
[표 7.3] 심리자료에 대한 주성분분석 결과

주성분번호	여자 1	2	3	4	주성분번호	남자 1	2	3	4
X_1	0.217	0.273	0.373	0.860	X_1	0.237	0.205	-.004	0.949
X_2	0.388	0.621	0.465	-.498	X_2	0.312	0.851	-.331	-.263
X_3	0.681	0.171	-.708	0.081	X_3	0.756	-.476	-.441	-.088
X_4	0.582	-.715	0.378	-.084	X_4	0.525	0.086	0.834	-.146
고유값	48.96	18.46	13.54	4.81	고유값	43.56	11.14	6.47	2.52
누적변동 비율(%)	57.07	78.60	94.38	100	누적변동 비율(%)	68.39	85.88	96.04	100

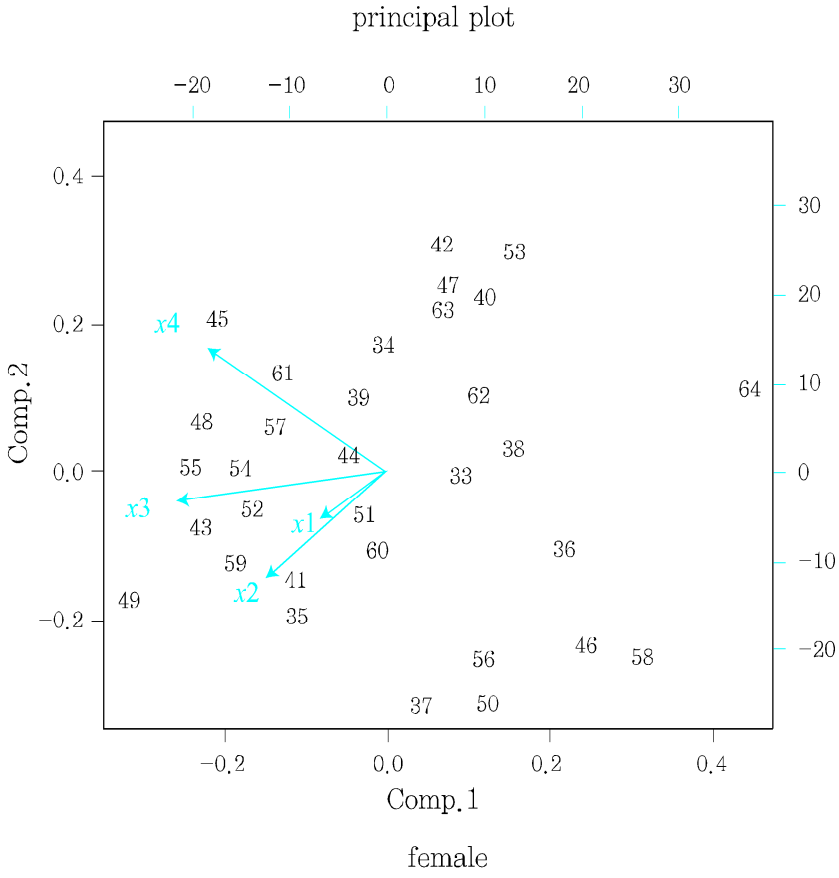
주성분번호	전체 1	2	3	4
X_1	0.274	-.002	0.327	0.904
X_2	0.284	0.185	0.854	-.394
X_3	0.856	-.409	-.271	-.163
X_4	0.333	0.894	-.300	0.009
고유값	72.72	16.11	13.11	4.29
누적변동 비율(%)	68.45	83.61	95.96	100



[그림 7.7] 남녀 집단별 스크리 그래프



male



female

[그림 7.8](a) 남자 집단 주성분 그래프 (b) 여자 집단 주성분 그래프

7.10 R을 이용한 주성분분석

≪예제 7.6≫ 형제들의 머리 크기(head size in brothers) 자료를 [표 7.4]와 같이 얻었다. 주성분을 구하고 의미를 해석하고자 한다.

[표 7.4] 형제들의 머리 크기 자료

첫 번째 아들	두 번째 아들	첫 번째 아들	두 번째 아들
191	179	195	201
181	185	183	188
176	171	208	192
189	190	197	189
188	197	192	187
179	186	183	174
174	186	190	195
188	187	163	161
195	183	186	173
181	182	175	165
192	185	174	178
176	176	197	200
190	187		

● R에서 주성분분석: princopm() 함수를 이용

library(graphics)와 함께 screeplot()를 이용해 스크리 그래프를 그리며

biplot()을 이용해 주성분 축으로 한 이변량 그림을 그릴 수 있다.

▶ 상관행렬을 이용한 경우 표본 주성분

첫 번째 주성분:

$$Y_1 = 0.707 \times firstson \text{ 머리크기} + 0.707 \times secondson \text{ 머리크기}$$

두 번째 주성분:

$$Y_2 = -0.707 \times firstson \text{ 머리크기} + 0.707 \times secondson \text{ 머리크기}$$

첫 번째 주성분이 전체분산의 85% 정도를 설명하고 있다.

Y_1은 두 변수의 가중 평균을 나타내며, Y_2 는 두 변수의 차를 나타내는 성분이다.

[프로그램 7.1] 형제들의 머리 크기 자료에 대한 주성분분석

```
son=read.csv("C:/data/son[redacted]header=T)
son
attach(son)

m=mean(son)
S=cov(son)
R=cor(son)
eigen(S)     # eigen values and vectors of S
eigen(R)     # eigen values and vectors of R

p_cor=princomp(son, cor=TRUE) # with correlation matrix
summary(p_cor)
attributes(p_cor)
p_cor$sdev      # the standard deviations of the principal
components
p_cor$loadings  # the matrix of variable loadings
p_cor$scores    # the scores of the principal components for data
```

```
all=cbind(son, p_cor$scores)
all
all[ order(p_cor$scores[,1]), ]  # sort by the first princomp

###### scatterplot, scree plot and Biplot #####
plot(first, second, pch="*", main="scatterplot of head sizes")      #그림 7.4
library(graphics)
screeplot(p_cor,npcs=2, type="lines",main="scree plot-correlation") #그림 7.5
biplot(p_cor)                                                       #그림 7.6

p_cov=princomp(son)   # with covariance matrix
summary(p_cov)
p_cov$sdev
p_cov$loadings
p_cov$scores
```

[결과 7.1] 형제들의 머리 크기 자료에 대한 주성분분석 결과

(1)

```
> m=mean(son) ; m
 first second
185.72 183.88
> S=cov(son)  ; S
          first    second
first  95.29333  69.17333
second 69.17333 100.94333
> R=cor(son)  ; R
           first    second
first  1.0000000 0.7052915
second 0.7052915 1.0000000
```

(2)

```
> eigen(S)     # eigen values and vectors of S$values
[1] 167.34933  28.88734
$vectors
          [,1]       [,2]
[1,] 0.6925296 -0.7213894
[2,] 0.7213894  0.6925296

> eigen(R)     # eigen values and vectors of R$values
[1] 1.7052915 0.2947085
$vectors
          [,1]       [,2]
[1,] 0.7071068 -0.7071068
[2,] 0.7071068  0.7071068
```

(3)

```
> p_cor=princomp(son, cor=TRUE)
> summary(p_cor)
Importance of components:
                                Comp.1
 Comp.2
Standard deviation     1.3058681
0.5428706
Proportion of Variance 0.8526457
0.1473543
Cumulative Proportion  0.8526457
1.0000000
```

(4)

```
> p_cor$sdev
   Comp.1    Comp.2
1.3058681 0.5428706

> p_cor$loadings  # the matrix of variable
Loadings:
       Comp.1 Comp.2
first   0.707 -0.707
second  0.707  0.707
               Comp.1 Comp.2
SS loadings       1.0    1.0
Proportion Var    0.5    0.5
Cumulative Var    0.5    1.0
```

```
> p_cor$scores
            Comp.1      Comp.2
 [1,]  0.03981389 -0.74088227
 [2,] -0.26849706  0.42939800
 [3,] -1.64377574 -0.20658508
 [4,]  0.68209332  0.19711540
 [5,]  1.11097919  0.77386040
 [6,] -0.34452540  0.64908790
 [7,] -0.71417321  1.01873570
 [8,]  0.39267142  0.05555263
 [9,]  0.62285524 -0.74927741
[10,] -0.48398939  0.21390567
[11,]  0.54472811 -0.38382717
 ⋮
[25,]  1.99183757  0.32398668
```

(5)

```
> all[ order(p_cor$scores[,1]), ]
   first second      Comp.1      Comp.2
21   163    161 -3.32316780  0.03619144
23   175    165 -2.14868996 -0.56364018
3    176    171 -1.64377574 -0.20658508
24   174    178 -1.28881942  0.44408948
12   176    176 -1.28462185  0.15256881
19   183    174 -0.91077648 -0.50859967
22   186    173 -0.76081858 -0.80221913
7    174    186 -0.71417321  1.01873570
10   181    182 -0.48398939  0.21390567
6    179    186 -0.34452540  0.64908790
2    181    185 -0.26849706  0.42939800
1    191    179  0.03981389 -0.74088227
15   183    188  0.09485440  0.49703121
8    188    187  0.39267142  0.05555263
13   190    187  0.54053054 -0.09230650
11   192    185  0.54472811 -0.38382717
9    195    183  0.62285524 -0.74927741
4    189    190  0.68209332  0.19711540
18   192    187  0.68838967 -0.24016562
5    188    197  1.11097919  0.77386040
20   190    195  1.11517676  0.48233972
17   197    189  1.20169902 -0.46615187
14   195    201  1.91580923  0.54367658
25   197    200  1.99183757  0.32398668
16   208    192  2.23041653 -1.06388471
```

(5)는 첫 번째 주성분에 의해 개체들을 오름차순으로 정렬한 결과:

주성분과의 관계, 원래 변수들과의 관계에 대한 정보를 얻을 수 있다.

```
    (6)
> p_cov=princomp(son)   # with covariance matrix
> summary(p_cov)
Importance of components:
                            Comp.1    Comp.2
Standard deviation     12.6749894 5.2661034
Proportion of Variance  0.8527934 0.1472066
Cumulative Proportion   0.8527934 1.0000000
> p_cov$sdev
   Comp.1    Comp.2
12.674989  5.266103
> p_cov$loadings
Loadings:
       Comp.1 Comp.2
first   0.693 -0.721
second  0.721  0.693
               Comp.1 Comp.2
SS loadings       1.0    1.0
Proportion Var    0.5    0.5
Cumulative Var    0.5    1.0
```

(6)은 표본공분산행렬을 이용하여 주성분분석을 한 결과

첫 번째 주성분 :

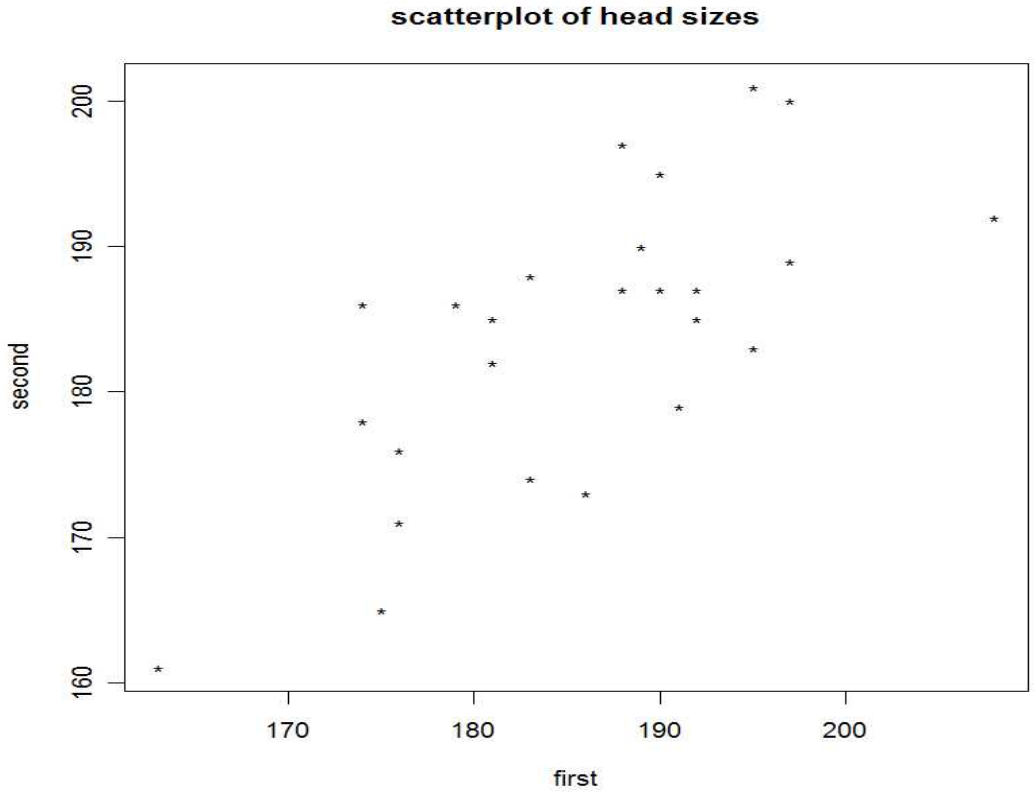
$$Y_1 = 0.693 \times firstson \text{ 머리크기} + 0.721 \times secondson \text{ 머리크기}$$

두 번째 주성분 :

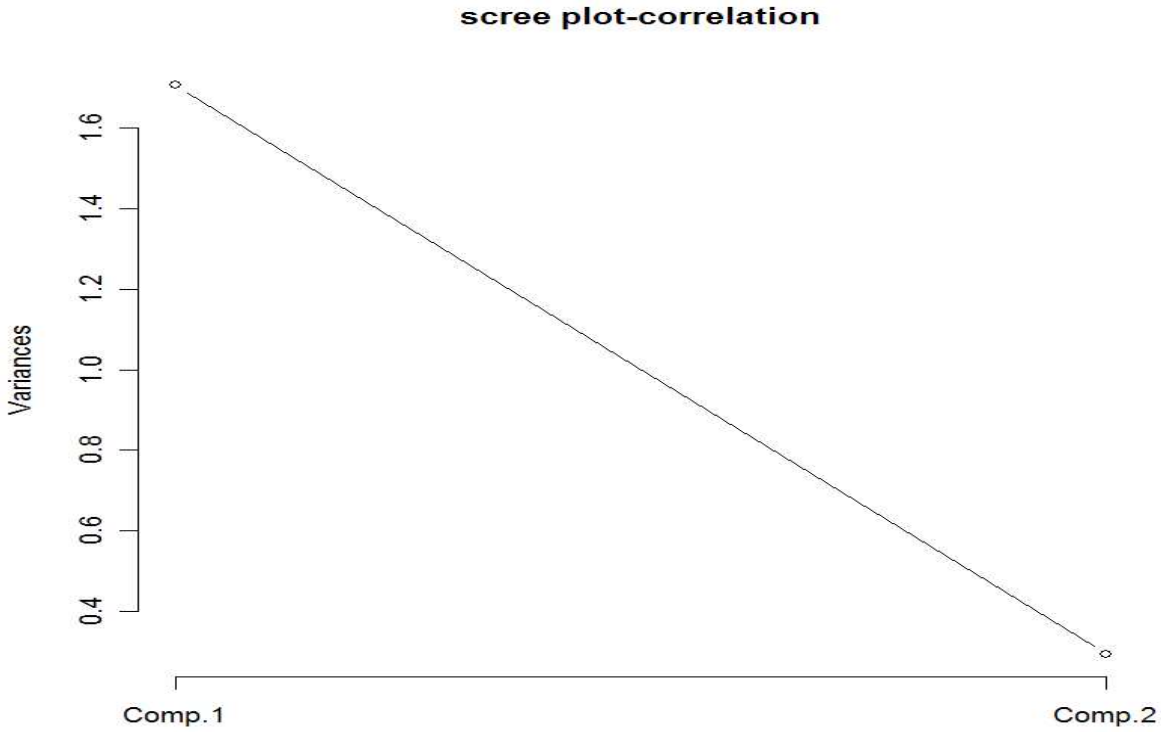
$$Y_2 = -0.721 \times firstson \text{ 머리크기} + 0.693 \times secondson \text{ 머리크기}$$

첫 번째 주성분이 전체 분산의 85.28%정도를 설명하고 있다.

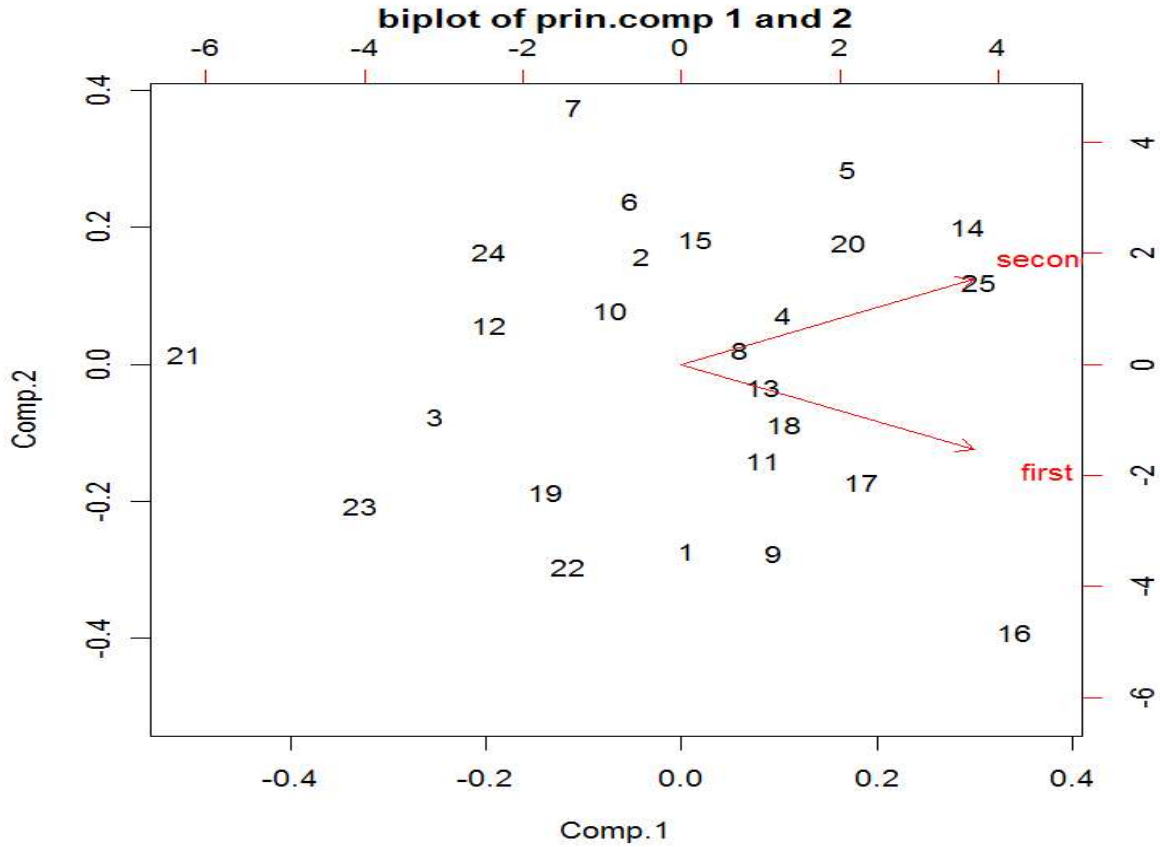
이 경우와 같이 두 변수의 단위가 일치할 때는 주성분분석 결과에 큰 차이가 없으나, 변수들의 단위가 다를 경우에는 상관행렬을 이용하는 것이 해석하기 편리하다.



[그림 7.4] 머리 크기 자료의 원래 변수의 산점도



[그림 7.5] 머리 크기 자료의 상관행렬에 대한 스크리 그래프



[그림 7.6] 머리 크기 자료의 상관행렬이용한 주성분 그래프

≪예제 7.6≫ [표 5.3]의 심리검사자료에 대해 공분산행렬을 이용한 주성분분석을 하고자 한다. ≪예제 5.2≫에서 보면 남녀 집단간에 반응평균벡터간에 통계적으로 유의한 차이가 있으므로 집단별 주성분분석을 하고자 한다.

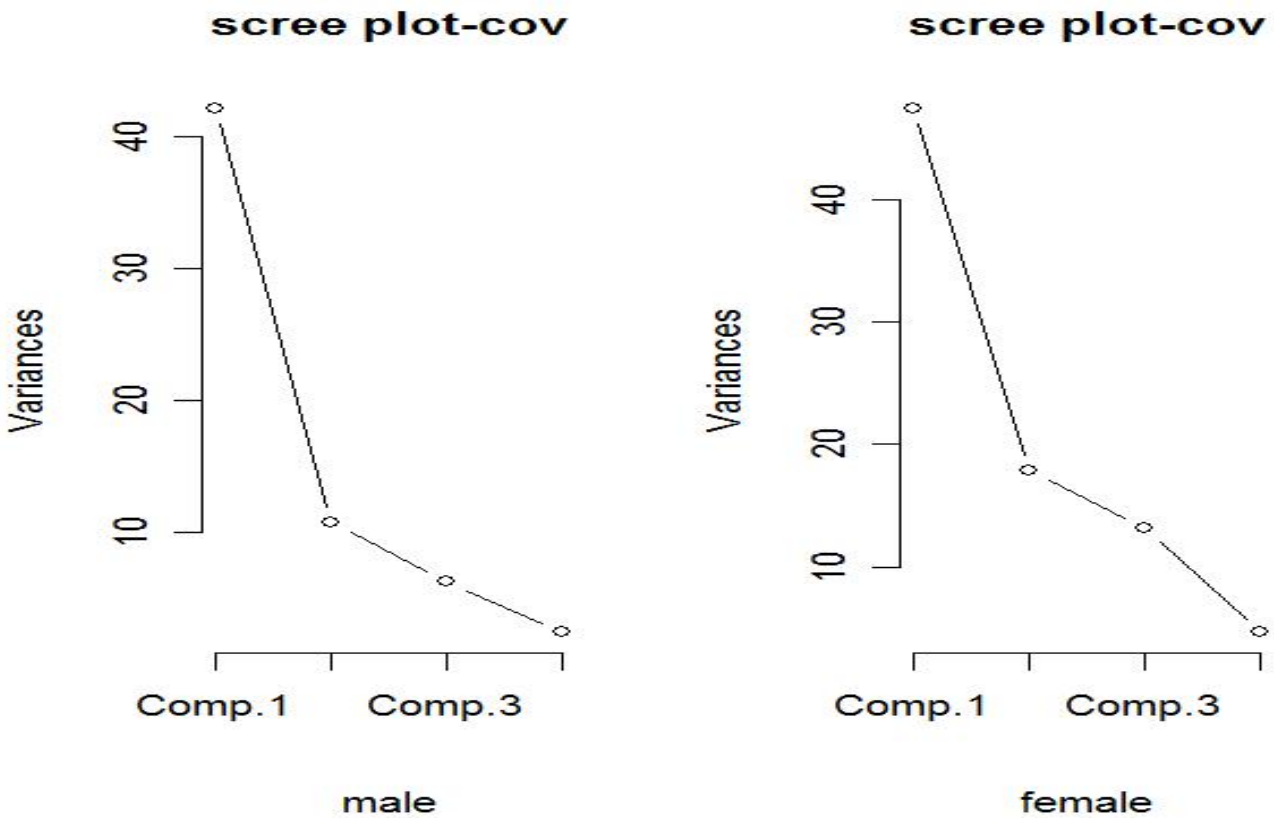
[표 7.4]에서 보면

- 첫 번째 주성분: 여자/남자, 전체자료의 경우 모두 측정변수들의 가중평균형태.
- 두 번째 주성분: 성별에 따른 차이가 있으며 전체에 대한 결과와도 차이가 난다.
  - 여자의 경우: X_2와 X_4의 대비
  - 남자의 경우: X_2와 X_3의 대비
  - 전체 자료의 경우: X_3와 X_4의 대비로 나타난다.

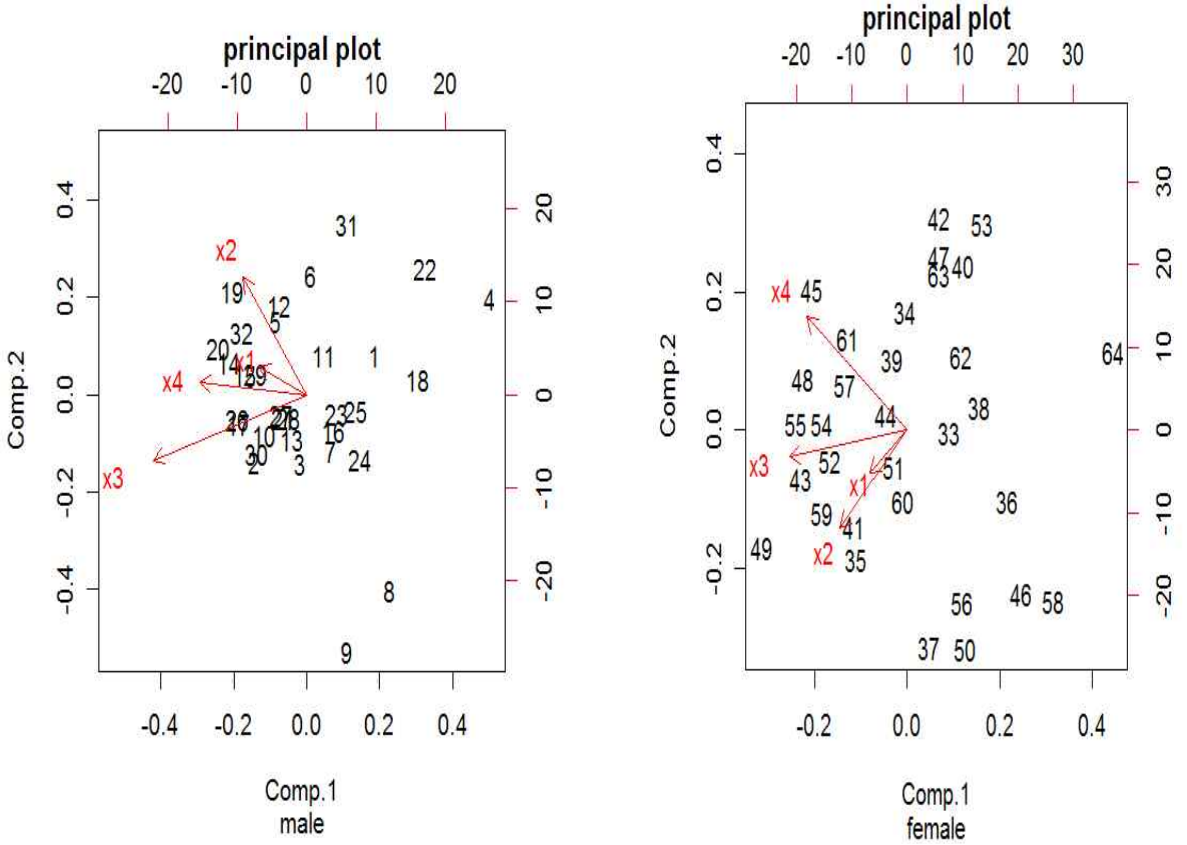
▶ 표 7.4 심리자료에 대한 주성분분석 결과

주성분번호	여자 1	2	3	4	주성분번호	남자 1	2	3	4
X_1	-0.217	-0.273	0.373	0.860	X_1	-0.237	0.205		0.949
X_2	-0.388	-0.621	0.465	-0.498	X_2	-0.312	0.851	-0.331	-0.263
X_3	-0.681	-0.171	-0.708		X_3	-0.756	-0.476	-0.441	
X_4	-0.582	0.715	0.378		X_4	- 0.525		0.834	-0.146
고유값	48.96	18.46	13.54	4.81	고유값	43.56	11.14	6.47	2.52
누적변동 비율(%)	57.07	78.60	94.38	100	누적변동 비율(%)	68.39	85.88	96.04	100

주성분번호	전체			
	1	2	3	4
X_1	-0.274		0.327	0.904
X_2	-0.284	0.185	0.854	-.394
X_3	-0.856	-.409	-.271	-.163
X_4	-0.333	0.894	-.300	
고유값	72.72	16.11	13.11	4.29
누적변동 비율(%)	68.45	83.61	95.96	100



[그림 7.7] 남녀 집단별 스크리 그래프



[그림 7.8] 남녀 집단별 주성분 그래프

[프로그램 7.2] 남녀 심리 자료에 대한 주성분분석

```
pschy=read.csv("C:/data/pschy.c          header=T)
pschy
attach(pschy)
 ps=pschy[, -1]   # except gender column
ps1=ps[pschy$gender==1,]    # male
ps2=ps[pschy$gender==2,]    # female

p_cov1=princomp(ps1, cor=FALSE)   # with cov matrix  : male
summary(p_cov1)
attributes(p_cov1)
p_cov1$sdev      # the standard deviations of the principal components
p_cov1$loadings  # the matrix of variable loadings
p_cov1$scores    # the scores of the principal components for data
###### scatterplot, scree plot and Biplot #####
library(graphics)
screeplot(p_cov1,npcs=4,type="l                    plot-cov",sub="male") #그림7.7
biplot(p_cov1)    #그림 7.8
```

```
p_cov2=princomp(ps2, cor=FALSE)   # with cov  matrix : female
summary(p_cov2)
attributes(p_cov2)
p_cov2$sdev      # the standard deviations of the principal components
p_cov2$loadings  # the matrix of variable loadings
p_cov2$scores    # the scores of the principal components for data
###### scatterplot, scree plot and Biplot #####
library(graphics)
screeplot(p_cov2,npcs=4,type="l[redacted]plot-cov",sub="female")   #그림7.7
biplot(p_cov2)   #그림 7.8

p_cov=princomp(ps)   # with covariance matrix  : All male and female
summary(p_cov)
p_cov$sdev
p_cov$loadings
p_cov$scores
```

[결과 7.2] 심리 자료에 대한 주성분분석 결과

```
> p_cov1$loadings

Loadings:
   Comp.1 Comp.2 Comp.3 Comp.4
x1 -0.237  0.205         0.950
x2 -0.312  0.851 -0.331 -0.263
x3 -0.756 -0.476 -0.441
x4 -0.525         0.834 -0.146

               Comp.1 Comp.2 Comp.3 Comp.4
SS loadings      1.00   1.00   1.00   1.00
Proportion Var   0.25   0.25   0.25   0.25
Cumulative Var   0.25   0.50   0.75   1.00
> summary(p_cov1)
Importance of components:
                          Comp.1    Comp.2    Comp.3     Comp.4
Standard deviation     6.4963237 3.2848316 2.5041099 1.56274496
Proportion of Variance 0.6839343 0.1748660 0.1016216 0.03957813
Cumulative Proportion  0.6839343 0.8588003 0.9604219 1.00000000
```

```
> p_cov2$loadings
Loadings:
   Comp.1 Comp.2 Comp.3 Comp.4
x1 -0.217 -0.273  0.373  0.860
x2 -0.388 -0.621  0.466 -0.498
x3 -0.681 -0.171 -0.708
x4 -0.582  0.715  0.378

                Comp.1 Comp.2 Comp.3 Comp.4
SS loadings      1.00   1.00   1.00   1.00
Proportion Var   0.25   0.25   0.25   0.25
Cumulative Var   0.25   0.50   0.75   1.00
> summary(p_cov2)
Importance of components:
                           Comp.1    Comp.2    Comp.3     Comp.4
Standard deviation     6.8866378 4.2292468 3.6214532 2.16043839
Proportion of Variance 0.5707436 0.2152547 0.1578310 0.05617076
Cumulative Proportion  0.5707436 0.7859982 0.9438292 1.00000000
```

```
> summary(p_cov)
Importance of components:
                          Comp.1    Comp.2    Comp.3     Comp.4
Standard deviation     8.4605675 3.9823127 3.5929697 2.05607785
Proportion of Variance 0.6844839 0.1516474 0.1234444 0.04042438
Cumulative Proportion  0.6844839 0.8361312 0.9595756 1.00000000

> p_cov$loadings
Loadings:
   Comp.1 Comp.2 Comp.3 Comp.4
x1 -0.274          0.327  0.904
x2 -0.284  0.185  0.854 -0.394
x3 -0.856 -0.409 -0.271 -0.163
x4 -0.333  0.894 -0.300

               Comp.1 Comp.2 Comp.3 Comp.4
SS loadings      1.00   1.00   1.00   1.00
Proportion Var   0.25   0.25   0.25   0.25
Cumulative Var   0.25   0.50   0.75   1.00
```