

## 7.1 7.2 주성분 정의&개념

2019년 6월 14일 금요일      오후 10:21

주성분분석 : 다변량 데이터에 대해 분산-공분산 구조를 변수들의 선형결합식으로 설명하고자 하는 접근방법.

주성분분석의 목적 1) 차원축소 2) 변동이 큰 축 탐색 3) 주성분을 통한 데이터의 해석

주성분은 서로 독립적인 새로운 변수.  $P$ 개의 변수에 포함된 전체 변동을  $m$ 개의 주성분으로 대신하여 설명함. ( $m < P$ )

차원축소 결과로 얻어지는 주성분점수는 원래 변수에 대한 정보 축약과 함께 새로운 변수  $\{x_i\}$ 으로 또 다른 통계분석에 이용됨.

$P \times 1$  확률벡터  $x$ 가 모평균벡터와 모공분산행렬을 가진다고 하자.

이러한 확률벡터는 수학적으로 차원에 놓이게 되며 원래 변수의 선형결합 또는 회전변환을 통해  $P$ 개의 새로운 좌표축을 형성할 수 있음.

이때 생기는 새로운 축은 데이터의 변동을 최대로 설명해주는 동시에 공분산 구조에 대한 해석을 놓이하게 하도록 만들어질 수 있는데, 이것을 주성분이라 함.

첫번째 주성분은 변동을 최대로 설명해주는 방향으로 변수들의 선형결합식이다.

두번째 주성분은 첫번째 주성분 다음으로 변동을 가장 많이 설명해주는 변수들의 선형결합식이며, 첫번째 주성분과는 독립이다.

이와 같이 찾아진  $P$ 개의 주성분들은 새로운 축을 형성하며, 주성분과 이들이 설명하는 변동량, 주성분, 주성분점수 등을 변수들로 표현된 시스템에 대한 이해를 돋는다.

### 7.2.1 주성분의 정의

학률벡터  $X$ 의 공분산행렬  $\Sigma$ 는 고유값을 갖고 각 고유값에 해당하는 고유벡터  $e$ 를 가진다.

$$Y_i = l_i' X = l_{1i} X_1 + l_{2i} X_2 + \dots + l_{pi} X_p \quad (i=1, \dots, p)$$

$$\text{Var}(Y_i) = l_i' \Sigma l_i$$

$$\text{Cov}(Y_i, Y_k) = l_i' \Sigma l_k$$

주성분은 아래의 과정들로 구할 수 있다.

(1) 첫번째 주성분은  $l_1' l_1 = 1$  을 만족하는  $p \times 1$  벡터  $l_1$ 에 대해  $\text{Var}(l_1' X)$ 를 최대로 하는 선형결합식  $l_1' X$ 로 구한다.

(2) 두번째 주성분은  $l_2' l_2 = 1$  을 만족하는  $p \times 1$  벡터  $l_2$ 에 대해  $\text{Var}(l_2' X)$ 를 최대로 하며  $\text{Cov}(l_1' X, l_2' X) = 0$  인 선형결합식  $l_2' X$ 로 구한다.

(i)  $i$ 번째 주성분은  $l_i' l_i = 1$  을 만족하는  $p \times 1$  벡터  $l_i$ 에 대해  $\text{Var}(l_i' X)$ 를 최대로 하며 앞에서 구한  $(i-1)$ 개의 주성분들과는 직교하도록  $\text{Cov}(l_i' X, l_j' X) = 0$  인 선형결합식  $l_i' X$ 로 구한다.

(p)  $P$ 번째 주성분은  $l_p' l_p = 1$  을 만족하는  $p \times 1$  벡터  $l_p$ 에 대해  $\text{Var}(l_p' X)$ 를 최대로 하며 앞에서 구한  $(p-1)$ 개의 주성분들과는 직교하도록  $\text{Cov}(l_p' X, l_j' X) = 0$  인 선형결합식  $l_p' X$ 로 구한다.

주성분을 구하기위해 Lagrange 기법 이용.

$l_i = e_i$ , 즉  $\Sigma$ 의 첫번째 고유벡터로 구하여 첫번째 주성분의 계수가 되며 설명하는 분산의 양은 첫번째 고유값으로 나타난다.

$l_i = e_i$ 이며  $p$ 개의 서로 독립적인 주성분은 가능한 선형결합들 중 주성분들로 인한 분산의 합이 최소가 되도록 한다.

### 정리 7.1 주성분 구하기

혹들벡터  $X' = (X_1, \dots, X_p)$ 의 공분산행렬  $\Sigma$ 는 고유값  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$  을 갖고 각 고유값에 해당하는 고유벡터  $e_1, e_2, \dots, e_p$  를 갖는다.

i번째 주성분은  $Y_i = e_i' X = e_{i1} X_1 + e_{i2} X_2 + \dots + e_{ip} X_p$  이고

i번째 주성분의 분산, 다른 주성분과의 공분산은  $\text{Var}(Y_i) = e_i' \Sigma e_i = \lambda_i$ ,  $\text{Cov}(Y_i, Y_j) = e_i' \Sigma e_j = 0$  이 된다.

[증명]  $\max_{a \neq 0} \frac{a' \Sigma a}{a'a} = \lambda_i$  을 얻고  $a = e_i$  으로 구해진다.

$$\text{또한 } e_i' e_i = 1 \text{ 이므로 } \max_{a \neq 0} \frac{a' \Sigma a}{a'a} = \lambda_i = \frac{e_i' \Sigma e_i}{e_i' e_i} = \text{Var}(Y_i)$$

$$\max_{a \perp e_1, \dots, e_k} \frac{a' \Sigma a}{a'a} = \lambda_{k+1} \text{에서 } a = e_{k+1} \text{로 선택하면 } e_{k+1}' e_i = 0$$

$$\frac{e_{k+1}' \Sigma e_{k+1}}{e_{k+1}' e_{k+1}} = e_{k+1}' \Sigma e_{k+1} = \lambda_{k+1} = \text{Var}(Y_{k+1}) \text{ 를 얻는다.}$$

$i \neq k$  에 대해서  $\text{Cov}(Y_i, Y_k) = e_i' \Sigma e_k = e_i' \lambda_k e_k = \lambda_k e_i' e_k = 0$  이 되어 주성분들이 서로 직교한다.

### 정리 7.2 주성분을 이용해도 원래 변수들의 충분산이 변하지 않음.

전체변이는 공분산행렬  $\Sigma$ 의 고유값들의 합으로 표현된다.

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

[증명] 대각행렬  $\Lambda$  와 고유벡터로 이루어진 직교행렬  $P = [e_1 \ e_2 \ \dots \ e_p]$  에 대해서  $\Sigma = P \Lambda P'$  로 표현된다.

$P P' = P' P = I$  이며  $\Sigma$ 는 양정치행렬이므로 스펙트럼 분석에 의해

$$\text{tr}(\Sigma) = \text{tr}(P \Lambda P') = \text{tr}(\Lambda P P') = \text{tr}(\Lambda) = \lambda_1 + \dots + \lambda_p \text{ 이므로 정리 7.1의 결과를 얻는다.}$$

그러므로 i번째 주성분에 의해 설명되는 전체분산의 비율은

$$\frac{i\text{번째 주성분에 의해 설명되는 분산}}{\text{전체 분산}} = \frac{\lambda_i}{\lambda_1 + \dots + \lambda_p}$$

### 정리 7.3 상관계수

$Y_i$  ( $i$ 번째 주성분) 와  $X_k$  (원래 데이터의  $k$ 번째 확률변수)의 상관계수는

$$\rho_{Y_i, X_k} = \frac{e_{ki} \sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}} \text{ 이다. 여기서 } e_{ki} \text{ 는 } e_i \text{ 의 } k\text{번째 성분이다.}$$

[증명]  $k$ 번째 성분이 1인 벡터  $\ell_k' = [0 \ 0 \ 0 \ \cdots \ 0 \ . \ 1 \ . \ 0 \ . \ \cdots \ 0]$  를 이용해  $X_k$ 를  $X_k = \ell_k' X$ 로 표현한 후

$\sum e_i = \lambda_i e_i$  인 사실을 이용하여 공분산을 구하면

$$\begin{aligned} \text{Cov}(X_k, Y_i) &= \text{Cov}(\ell_k' X, e_i' X) = \ell_k' \text{Cov}(X, Y) e_i \\ &= \ell_k' \sum e_i = \ell_k' \lambda_i e_i = \lambda_i \ell_k' e_i = \lambda_i e_{ki} \text{ 이 된다.} \end{aligned}$$

상관계수 정의에 의하여

$$\text{Corr}(X_k, Y_i) = \frac{\text{Cov}(X_k, Y_i)}{\sqrt{\text{Var}(X_k)} \sqrt{\text{Var}(Y_i)}} = \frac{\lambda_i e_{ki}}{\sqrt{\sigma_{kk}} \sqrt{\lambda_i}} = \frac{e_{ki} \sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}} \text{ 이 된다.}$$

예제 7.1 두 변수  $X_1$  과  $X_2$ 에 대해 관측값을 얻은 후 표본공분산행렬  $S = \begin{pmatrix} 7 & 1 \\ 1 & 7 \end{pmatrix}$  를 얻었을 때 주성분을 구해보자.

① 고유값 찾기

$$|S - \lambda I| = (\lambda - 6)(\lambda - 8) = 0$$

$$\lambda_1 = 8, \lambda_2 = 6$$

② 고유벡터 구하기

$$\text{i) } \lambda_1 = 8$$

$$\text{ii) } \lambda_2 = 6$$

$$(S - \lambda_1 I) X_1 = 0$$

$$(S - \lambda_2 I) X_2 = 0$$

$$X_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

$$X_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

$$e_1 = \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}$$

$$e_2 = \begin{pmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix}$$

③ 주성분

$$Y_1 = e_1' X = 0.707 X_1 + 0.707 X_2 \quad ; \text{ 두 변수의 가중평균}$$

$$Y_2 = e_2' X = 0.707 X_1 - 0.707 X_2 \quad ; \text{ 두 변수의 대비}$$

#### ④ 분산 설명량

첫번째 주성분이 의해  $\frac{8}{8+6} = 0.571$  으로 전체분산의 57.1% 정도를 설명한다.

#### 7.2.2 주성분의 기하학적 의미

모집단 주성분 분석 모집단 공분산행렬을 고려하여 최적의 직교 선형결합식을 찾는 것이다.

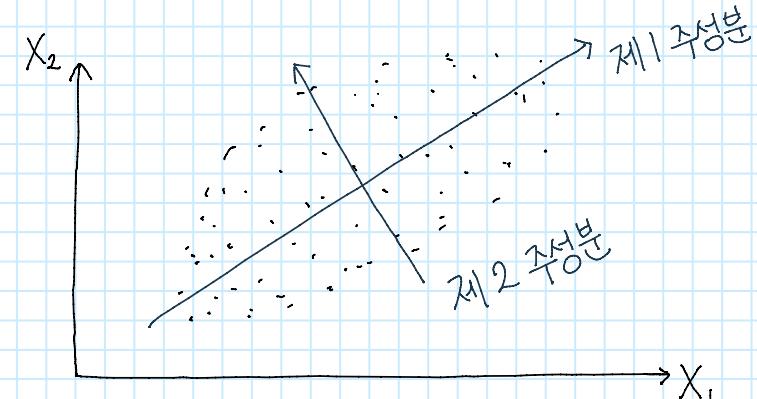
산점도를 보면 두 변수 사이의 선형상관성이 높음을 알 수 있다.

공분산행렬을 고려하여 주성분축을 찾으면 원래 변수들의 좌표축이 직교회전변환되어 나타나며

제 1 주성분 축의 데이터 벡터에 비해 제 2 주성분 축의 데이터 벡터가 상대적으로 작음을 알 수 있다.

또한 주성분 좌표축의 데이터는 원래 좌표축의 변수에 비해 독립하게 펼쳐지게 된다.

따라서 원래의 변수들에 비해 새로운 선형결합이며 서로 직교하는 주성분들은 통계적인 추론과 해석에 많은 편리한 점을 제공한다.

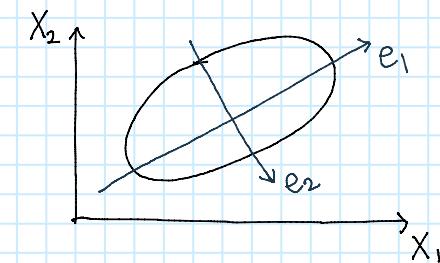


다변량 정규분포를 따르는 확률변수의 경우 주성분의 의미.

$X \sim N_p(\mu, \Sigma)$  을 따를 때  $(X-\mu)' \Sigma^{-1} (X-\mu) = C^2$  는  $\mu$ 를 중심으로 축이  $\pm C\sqrt{\lambda_i} e_i$  인 타원체이다.

일반성을 잃지 않고  $\mu=0$  으로 놓아보자. 그러면  $C^2 = x' \Sigma^{-1} x = \frac{1}{\lambda_1} (e_1' x)^2 + \dots + \frac{1}{\lambda_p} (e_p' x)^2 = \sum_{i=1}^p \frac{y_i^2}{\lambda_i}$

여기서  $y_i = e_i' x$  으로  $e_1, \dots, e_p$  방향으로의 축이 된다.  $\lambda_i$  이 가장 큰 고유값일 때 주축은  $e_1$  방향이 된다.



## 7.3 상관행렬을 이용한 주성분분석

2019년 6월 15일 토요일 오후 4:52

일반적으로 데이터의 본래의 의미를 해석하려면 표본공분산행렬  $S$ 를 이용해 주성분분석을 하게 된다.

그러나 변수의 단위가 다르거나 또는 분산의 차이가 큰 경우, 표본 공분산행렬  $S$  대신 표본상관행렬  $R$ 을 이용해 주성분을 구하면 해석하는데 편리한 이점이 있다.

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

주성분 계수는 척도변에 따라 달라지므로, 공분산행렬을 이용한 경우의 주성분 결과와 상관행렬을 이용해 얻은 주성분 결과에 차이가 나며 설명하는 변동의 비율도 차이가 난다.

$$\text{각 변수 표준화. } Z_p = \frac{x_p - \mu_p}{\sqrt{\sigma_{pp}}}$$

$$\text{벡터로 표현하면 } z = (V^{1/2})^{-1}(X - \mu) \sim N_p(0, I)$$

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

$$E(z) = 0, \text{ Cov}(z) = I, V^{1/2} \rho V^{-1/2} = \sum, \rho = V^{-1/2} \sum V^{-1/2}$$

→  $\sum$ 는 분산의 대각행렬과 상관행렬로부터 얻어진다.

### 정리 7.4

표준화 변수들로 구성한 벡터  $z' = (z_1, \dots, z_p)'$ 에 대해 주성분을 구하면  $i$ 번째 주성분은  $y_i = e_i' z = e_i' (V^{1/2})^{-1}(X - \mu)$  이 된다.

주성분에 대한 총분산은  $\sum \text{Var}(y_i) = \sum \text{Var}(z_i) = 1 + \dots + 1 = p$  이며,

$$i\text{번째 주성분 } y_i \text{ 와 } k\text{번째 표준화 변수 } z_k \text{ 의 상관계수는 } \rho_{y_i z_k} = \frac{e_{ik} \sqrt{\lambda_i}}{\sqrt{\text{Var}(z_k)}} = e_{ik} \sqrt{\lambda_i} |0| \text{ 된다.}$$

또한 표준화변수로부터 구한 주성분에 대하여  $i$ 번째 주성분에 의해 설명되는 전체분산의 비율은  $\frac{i\text{번째 주성분에 의해 설명되는 분산}}{\text{전체 분산}} = \frac{\lambda_i}{p}$  이다.

\* 표본상관행렬  $R$ 을 사용할 경우 주의점.

- ①  $R$ 에 의한 주성분과  $S$ 에 의한 주성분이 설명하는 분산의 양이 다르다.
- ②  $R$ 에 의한 주성분과  $S$ 에 의한 주성분 계수가 다르다.
- ③  $R$  자체가 척도불변이므로  $R$ 에 의한 주성분은 척도 불변이다.
- ④  $R$ 에 의한 주성분은 유일하지 않다.

예제 7.2 아변량 데이터에 대해 표본상관행렬  $R$ 을 이용하여 주성분을 구해보자.

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

고유값  $\lambda_1 = 1+r \quad \lambda_2 = 1-r$

고유벡터  $e_1 = (0.707, 0.707)'$

$$e_2 = (0.707, -0.707)'$$

주성분  $y_1 = 0.707 \cdot \left( \frac{x_1 - \bar{x}_1}{s_1} \right) + 0.707 \cdot \left( \frac{x_2 - \bar{x}_2}{s_2} \right)$

$$y_2 = 0.707 \cdot \left( \frac{x_1 - \bar{x}_1}{s_1} \right) - 0.707 \cdot \left( \frac{x_2 - \bar{x}_2}{s_2} \right)$$

↳ 주성분을 구하는 각 계수는  $r$ 에 의존하지 않는다.

↑이 몇이든 주성분의 계수에는 영향을 미치지 않고 분산의 설명비율에만 영향을 미친다.

## 7.4 주성분에의한 표본변동설명

2019년 6월 15일 토요일      오후 6:52

표본 주성분을 이용해 모집단 주성분의 추정치를 구하여 모집단에 대한 해석.

획득 표본  $X_1, X_2, \dots, X_n$ 에 대한 표본공분산행렬  $S$ 는 고유값  $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \dots \geq \hat{\lambda}_p > 0$  와 각 고유벡터  $\hat{e}_1, \hat{e}_2, \dots, \hat{e}_p$ 를 가진다.

정리 7.5

i번째 표본주성분은  $Y_i = \hat{e}_i' X = \hat{e}_{1i} X_1 + \hat{e}_{2i} X_2 + \dots + \hat{e}_{pi} X_p$  이고

i번째 표본주성분의 분산, 다른 주성분과의 공분산은  $\text{Var}(Y_i) = \hat{\lambda}_i$ ,  $\text{Cov}(Y_i, Y_j) = 0$

표본 총분산 =  $\sum S_{ii} = \hat{\lambda}_1 + \hat{\lambda}_2 + \dots + \hat{\lambda}_p$  가 되며,

i번째 표본주성분  $Y_i$ 와  $X_k$ 의 상관계수는  $\text{Corr}(Y_i, X_k) = \frac{\hat{e}_{ki} \sqrt{\hat{\lambda}_i}}{\sqrt{S_{kk}}}$  이다.

## 7.5 표본고유값과 표본고유벡터의 대수적성질

2019년 6월 15일 토요일 오후 7:04

### 7.5.1 표본고유값과 표본고유벡터의 분포

모집단 주성분은  $\Sigma$  또는  $P$ 로부터 얻어진  $(\lambda_i, e_i)$ 에 의존한다. 표본 주성분은  $S$  또는  $R$ 로부터 얻어진  $(\hat{\lambda}_i, \hat{e}_i)$ 에 의존한다.

정리 7.6 표본 고유값과 표본 고유벡터에 대한 성질

획을 표본  $X_1, \dots, X_n$ 은 공분산행렬  $\Sigma$ 를 가지는  $P$ -변량 정규분포를 따른다.

$\Sigma$ 는 고유값과 고유벡터를 가질 때,

i)  $\lambda' = (\lambda_1, \dots, \lambda_p)$  일 때  $n \rightarrow \infty$  이면 근사적으로  $\sqrt{n}(\hat{\lambda} - \lambda) \sim N_p(0, 2\Lambda^2)$  이다.

또한 근사적으로  $\hat{\lambda}_i \sim N(\lambda_i, \frac{2\lambda_i^2}{n-1})$  을 따른다.

$$* \Lambda = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}$$

ii)  $n \rightarrow \infty$  이면 근사적으로  $\sqrt{n}(\hat{e}_i - e_i) \sim N_p(0, E_i)$

$$\text{여기서 } E_i = \lambda_i \sum_{\substack{k=1 \\ k \neq i}}^n \frac{\lambda_k}{(\lambda_k - \lambda_i)^2} \cdot e_k e_k'$$

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

iv) 표본 고유값 벡터  $\hat{\lambda}$  와 표본 고유벡터  $\hat{e}$ 는 서로 독립이다.

v) 모집단 고유값  $\lambda_i$ 에 대하는  $100(1-\alpha)\%$  신뢰구간  $\frac{\hat{\lambda}_i}{1 + z_{\alpha/2} \cdot \sqrt{\frac{2}{n-1}}} \leq \lambda_i \leq \frac{\hat{\lambda}_i}{1 - z_{\alpha/2} \cdot \sqrt{\frac{2}{n-1}}}$   
근사적으로  $\frac{\hat{\lambda}_i - \lambda_i}{\lambda_i \sqrt{\frac{2}{n-1}}} \sim N(0, 1)$  이 되기 때문.

### 7.5.2 등상관구조에 대한 검정

두 변수간의 상관관계  $P$ 로 동일한 경우,  $\Sigma$ 의 고유값이 두 종류로만 구해지며 중복 고유값이 여러개 있음을.

(다변량해석 책 p167-168)

## 7.6 7.7 주성분 그래프 & 개수

2019년 6월 15일 토요일 오후 8:09

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

따라서 주성분 그래프로부터 두 주성분간의 관계와 패턴을 도출할 수 있으며 또한 전체 데이터가 주성분을 통해 변화되어 나타내는 관계도 알 수 있음.

처음 몇개의 주성분은 분산 또는 공분산 구조를 왜곡시키는 이상점의 영향을 많이 받는다. 그리고 마지막 몇개의 주성분은 인공적인 차원을 유도하는 이상값에 의해 민감하게 변한다.

주성분 분석시 원래 데이터에 대한 정보의 손실을 최소화하기 위해서 적절한 개수의 주성분을 선택해야 함.

전체 대변이의 대부분을 설명하기 위하여 설정해야 할 주성분의 개수를 결정하는 몇 가지 기준들.

### 1) 전체 대변이의 공헌도 (percentage of total variance)

전체 대변이의 70~90%가 되도록 주성분의 수를 결정한다.

i 번째 주성분이 설명하는 분산 양 =  $\lambda_i$

m개의 주성분을 결정할 때

#### (1) 공분산 행렬을 이용할 경우

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

#### (2) 상관행렬을 이용할 경우

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

## 2) 평균 고유값 ( average eigenvalues )

고유값들의 평균을 구한 후 고유값이 평균값 이상이 되는 주성분을 설정한다.

상관행렬 사용한 경우 평균 고유값은 1이 된다.

$$\bar{\lambda} = \frac{\sum_{i=1}^p \lambda_i}{p}$$

## 3) 스크리 그래프 ( scree graph )

2차원 좌표축에 ( 고유값 순서, 고유값 크기 ) 를 점을 찍고 거리간을 선분으로 연결함.

값이 큰 고유값부터 크기 순으로 점을 찍을 때 값의 차이가 크면 가파른 경사가 되고 고유값의 변화가 작으면 내리막 경사가 완만해짐.

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

## 7.8 7.9 마지막 주성분으로부터의 정보 & 주성분에 대한 해석

2019년 6월 16일 일요일 오전 3:05

일반적으로 전체 벡터의 대부분을 설명하는 처음 몇 개의 주성분에 대해 강조하게 되고 해석에 관심을 기울인다.

그러나 마지막 몇 개의 주성분, 즉 전체 벡터를 거의 설명하지 않는 주성분으로부터도 유용한 정보를 얻을 수 있다.

마지막 주성분이 설명하는 분산의 양이 거의 0일 때, 벡터들 간의 선형관계를 나타내는 주성분이 상수값을 가지다는 의미이고 벡터들 간의 공선성을 나타내어 유용한 정보를 활용될 수 있다.

즉  $x_5$  가 표준화된 벡터로 평균이 0이라면  $y_5$  의 분산설명량이 0일 때  $x_5 \leftarrow x_1, x_2, x_3, x_4$  에 의존함을 알 수 있다.

$$y_5 = e_{51}x_1 + e_{52}x_2 + \dots + e_{55}x_5 = 0$$

주성분 척도 불변이 아니다.

따라서 공분산 행렬을 이용하여 주성분분석을 수행한 결과와 상관행렬을 이용하여 주성분분석을 수행한 결과가 차이가 나므로 해석에도 차이가 난다.

일반적으로 벡터들의 측정단위가 다르거나 분산값이 크게 차이가 날 때는 공분산행렬보다는 상관행렬을 이용하여 주성분 분석을 하는 것이 데이터를 해석하는데 편리하다.

<예제 7.4>

38명의 학생을 대상으로 신체적 크기를 조사하고 표본상관행렬을 이용하여 주성분분석을 한 결과 얻은 주성분 계수들이다. 표 7.1의 결과로부터 주성분을 해석해보자. (책 172쪽)

먼저 짐만행렬을 살펴보면, 첫 번째 주성분과 두 번째 주성분은 주성분 계수의 경향이 비슷하여 세 번째 주성분은 계수의 부호에 차이를 보인다.

첫번째 주성분 : 신체크기들의 가중평균을 나타내며 전반적인 신체 크기 면동의 주요 면동을 나타내는 축

두번째 주성분 : 여자의 경우 키에 비해 큰 손을 가질수록 두번째 주성분값이 커짐

남자의 경우 키와 앞팔에 비해 큰 손을 가질수록 두번째 주성분값이 커짐

=> 키와 손의 대비를 나타내는 축

세번째 주성분 : 여자의 경우 머리와 가슴의 대비축

남자의 경우 머리와 손목 그리고 앞팔과 손의 대비

=> 남녀의 신체적 특성