

제 2 장  
주성분분석

Principal Component Analysis

## 2.1 선형변환과 주성분

- 주성분분석은 차원의 단순화를 통해 서로 상관되어 있는 변수들 간의 복잡한 구조를 분석하는 데 그 목적을 두고 있으며, 이를 위하여 관찰변수들을 선형변환시켜 ‘주성분’(principal component)이라고 불리는 서로 상관되어 있지 않은(혹은 독립적인) 새로운 인공변수들을 유도한다.

obs	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$		평균	주성분
1	3	33	73	8	12	→ 차원 축소	25.8	16.0
2	3	30	59	28	20		28.0	21.4
3	35	83	91	32	34		55.0	45.4
4	35	83	85	33	32		53.6	44.8
5	15	40	55	68	52		46.0	43.5
6	3	53	76	10	8		30.0	18.8
7	68	83	85	48	50		66.8	62.4
8	15	47	77	76	76		56.2	53.6
9	46	60	83	83	68		67.8	65.1
10	98	83	91	80	72		84.8	84.9
평균	32.1	59.5	77.5	46.6	42.4		51.4	45.6
표준편차	31.6	22.0	12.4	28.6	25.0		19.3	22.2

## 주성분점수 (Principal Component Score)

- 선형변환 (Linear Transformation)

- ✓  $Y = a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4 + a_5x_5$
- ✓ 주성분점수는 동일한 제약조건을 가지는 모든 가능한 선형결합 중 가장 변이가 크다는(즉, 개체들의 상대 위치를 멀리 떨어뜨려 놓는다는) 점에서 최적의 성질을 가진다.
- ✓ 주성분점수는 관찰변수들과 최대 다중상관계수를 가진다.

(상관계수)	국어	영어	제2외국어	수학	과학
국어	1.000				
영어	0.784	1.000			
제2외국어	0.683	0.860	1.000		
수학	0.559	0.212	0.138	1.000	
과학	0.610	0.309	0.279	0.973	1.000
(표준편차)	31.6	22.0	12.4	28.6	25.0

- ✓ 평균점수 =  $0.20 x_1 + 0.20 x_2 + 0.20 x_3 + 0.20 x_4 + 0.20 x_5$
- ✓ 주성분점수 =  $0.30 x_1 + 0.15 x_2 + 0.07 x_3 + 0.25 x_4 + 0.23 x_5$

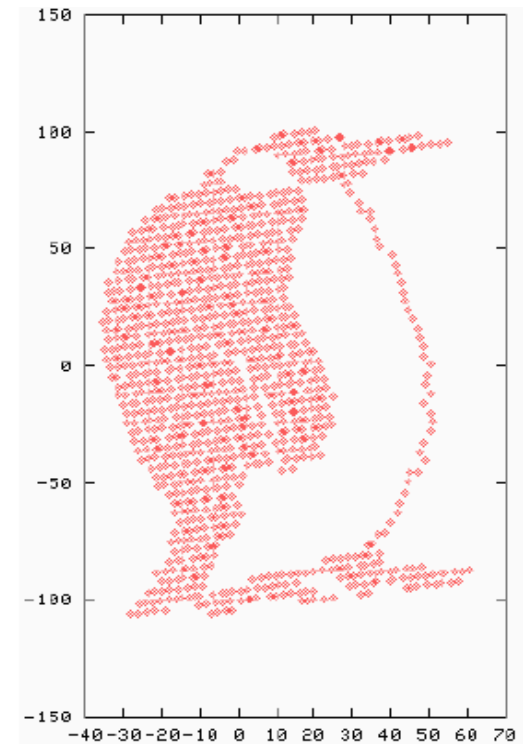
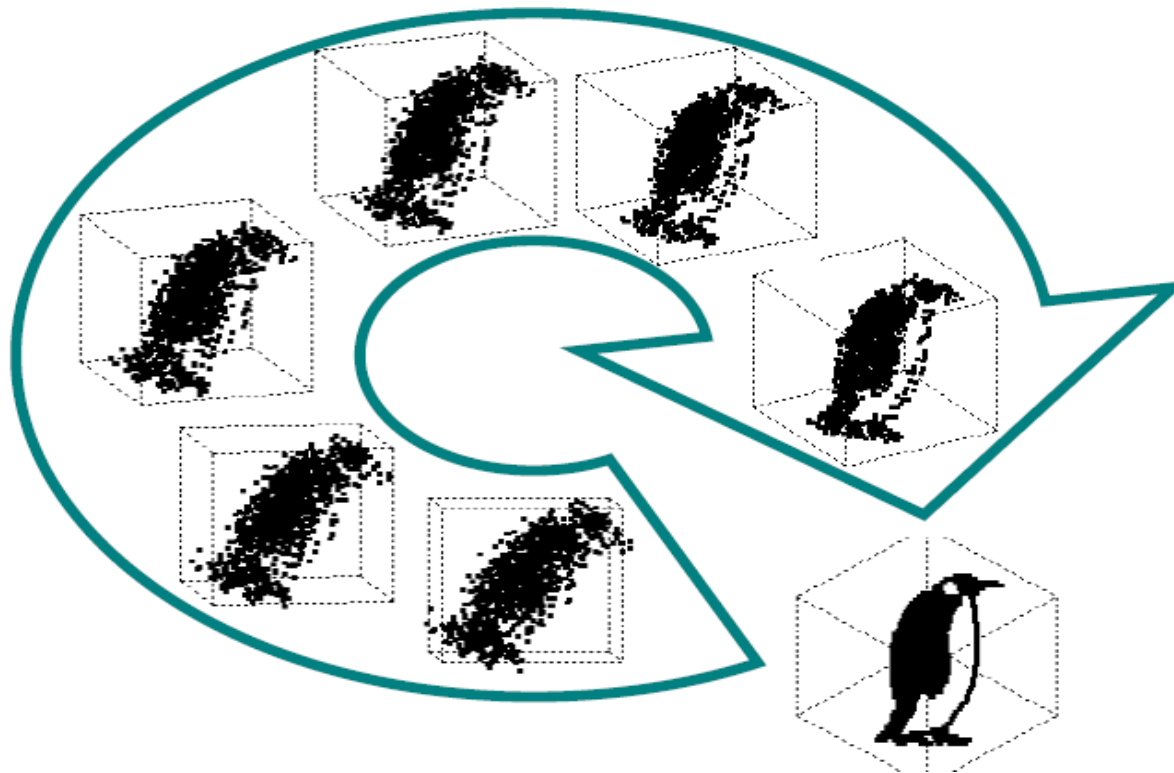
## 참조: 지수에 대한 가중치(중요도) 산출

---

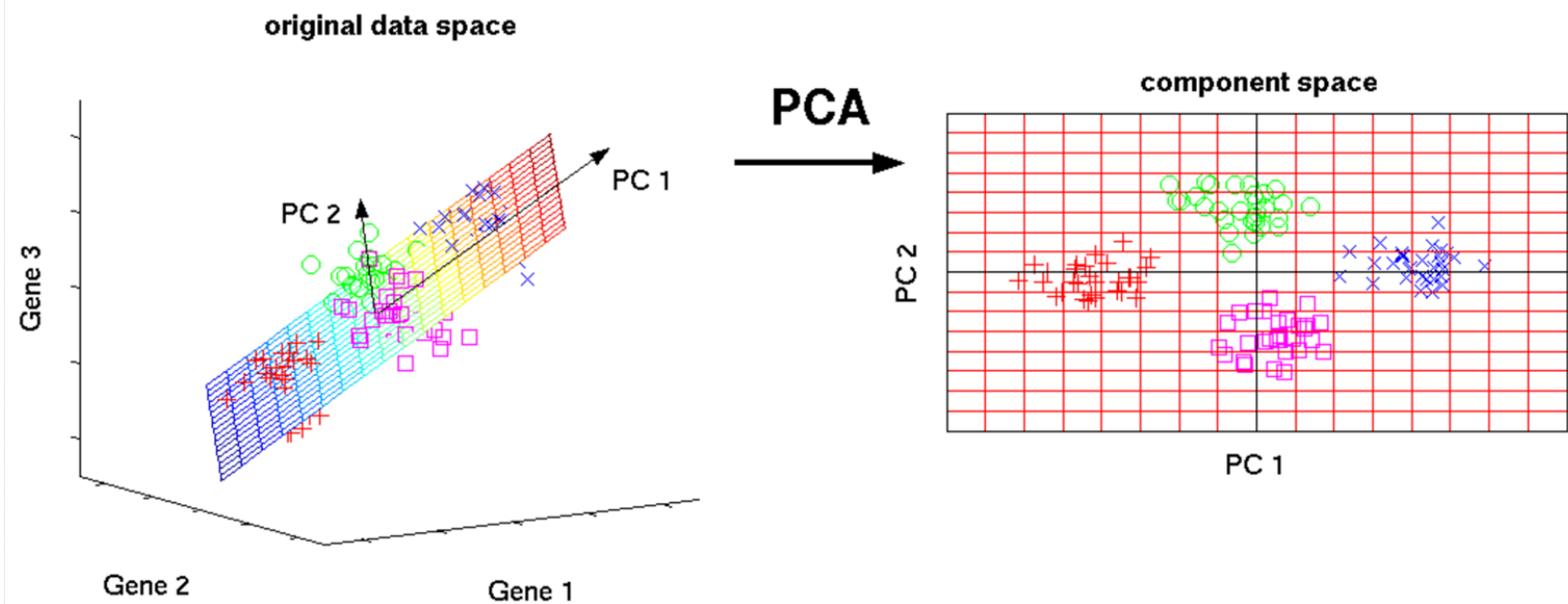
- **응답자로부터 가중치를 직접 얻는 방법**
  - ✓ 설문 응답자로부터 중요도의 점수 혹은 순위를 얻고 이의 상대적 크기를 이용.
  - ✓ 각 항목에 동일한 기본 가중치를 부여한 후, 응답자의 중요도를 반영하기도 함.
- **전문가 집단으로부터 가중치를 얻는 방법**
  - ✓ 전문가들의 의견을 종합하여 가중치를 부여함.
  - ✓ ‘델파이법’(delphi method) 혹은 ‘앙케이트 수렴법’이라고도 함.
- **준거변수를 이용하는 방법**
  - ✓ 상관분석을 통해 얻은 상관계수의 상대적 크기를 이용.
  - ✓ 회귀분석을 통해 얻은 (표준화)회귀계수의 상대적 크기를 이용.
- **관측 항목들의 연관성을 이용하는 방법**
  - ✓ 관측 항목들의 단순평균을 취함.
  - ✓ 주성분분석을 통해 얻은 주성분계수의 상대적 크기를 이용.
- **구조방정식모형을 이용하는 방법**
  - ✓ 이론적 타당성을 가지는 모형을 구축한 후, 이를 통해 가중치를 계산.
  - ✓ ACSI, NCSI, KSCI 등에서 활용.

## 주성분 분석의 간단한 예제 1.








- 아래의 그림은 3차원 자료를 2차원으로 투영한 경우이다. 처음에는 자료 구름의 연장 방향을 제외하고는 자료 점들의 명확한 구조가 보이지 않으나, 적절한 회전을 하게 되면 숨겨진 구조가 드러나게 된다. 이는 3차원 자료 주위를 돌면서 가장 적절한 관찰방향을 찾는 것과 같다.
- PCA는 그러한 숨겨진 구조를 찾는 데 도움을 준다. PCA는 주어진 자료에서의 대부분의 분산이 처음 몇 개의 차원에 표현되도록 회전을 시키는 것이다.



## 주성분 분석의 간단한 예제 2.



## 주성분 (Principal Component)

변수		A	B	C	D	주성분
		0.82	0.01	0.16	0.01	
		0.55	0.41	0.02	0.02	
		0.65	0.09	0.25	0.01	
		0.26	0.09	0.64	0.01	
		0.07	0.44	0.47	0.02	
		0.06	0.63	0.30	0.01	
		0.27	0.64	0.08	0.01	
	합계	2.68	2.31	1.93	0.10	고유값
	%	38.3%	27.4%	33.0%	1.3%	

\* 색광의 삼원색 : 빨강, 녹색, 파랑

## 2.2 prcomp 명령어 이용한 “고객만족 데이터” 분석

고객만족  
데이터

obs	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	성별	연령
1	1	2	4	1	1	여자	10대
2	1	2	3	2	1	여자	10대
3	2	5	5	2	2	여자	20대
4	2	5	5	2	2	여자	20대
5	1	2	3	4	3	여자	30대
6	1	3	4	1	1	남자	30대
7	4	5	5	3	3	남자	40대
8	1	3	4	4	4	남자	40대
9	3	3	5	5	4	남자	50대
10	5	5	5	4	4	남자	50대
평균	2.10	3.50	4.30	2.80	2.50		

- 변수:

$x_1$ (가격),  $x_2$ (성능),  $x_3$ (편리성),  $x_4$ (디자인),  $x_5$ (색상)

- 변수값:

1=‘매우 만족하지 않는다.’, 2=‘만족하지 않는다.’, 3=‘보통이다.’,

4=‘만족한다.’, 5=‘매우 만족한다.’



## [고객만족 데이터] 데이터 불러오기

---

```
Satis<- read.table("satis.txt",header=T) # 데이터 읽기
```

```
Satis<- Satis[,-1] # 첫번째 변수 (subject) 삭제
```

```
colnames(Satis)<-  
c("Gender","Age","Price","Function","easy","design","color") #  
변수 이름 설정
```

```
summary(Satis[,3:7])
```

## [고객만족 데이터] 기초 통계량과 상관계수

---

```
apply(Satis[,3:7], 2, sd) # 2는 column 을 의미, 각  
column 의 standard deviation 을 구하기  
apply(Satis[,3:7], 2, sum) # 각 column 의 합을 구하기  
apply(Satis[,3:7], 2, length) # 각 column 의 데이터 수  
apply(Satis[,3:7], 2, mean)  
apply(Satis[,3:7], 2, min)  
apply(Satis[,3:7], 2, max)
```

```
cor(Satis[,3:7]) # 피어슨 상관계수행렬
```

```
# 주성분 분석을 하기 전에 R function 둘러보기 #  
? prcomp
```

## [고객만족 데이터] 고유값과 고유벡터

아래의 식을  
이용하여  
고유값과  
고유벡터를  
구한다

$$S \underline{e}_1 = \delta_1 \underline{e}_1$$

```
> cov(Satis[,3:7]) # 공분산 행렬
```

	Price	Function	easy	design	color
Price	2.1000000	1.3888889	0.8555556	0.9111111	1.0555556
Function	1.3888889	1.8333333	0.9444444	0.1111111	0.5000000
easy	0.8555556	0.9444444	0.6777778	0.1777778	0.3888889
design	0.9111111	0.1111111	0.1777778	1.9555556	1.6666667
color	1.0555556	0.5000000	0.3888889	1.6666667	1.6111111

$S$

```
> eigen(cov(Satis[,3:7]))
```

\$values

$\delta_1$  [1] 5.08214264 2.45818249 0.44411043 0.14197348 0.05136875

\$vectors

	[,1]	[,2]	[,3]	[,4]	[,5]
[1,]	-0.5737259	-0.2558755	0.77345688	-0.04239327	0.0730312
[2,]	-0.4101979	-0.5814345	-0.50999927	-0.46583216	-0.1287354
[3,]	-0.2600413	-0.2859829	-0.22302049	0.87797754	-0.1732328
[4,]	-0.4523733	0.6017424	-0.08005497	-0.08610099	-0.6476437
[5,]	-0.4799096	0.3906169	-0.29243507	0.05427026	0.7270775

$\underline{e}_1$

## [고객만족 데이터] 주성분 분석 (Principal Component Analysis)

```
# 주성분 분석 (공분산 행렬에 기초)
> PCA.model1<-prcomp(Satis[,3:7],center=TRUE,scale=FALSE) #
covariance matrix 이용.
> PCA.model1
Standard deviations:
[1] 2.2543608 1.5678592 0.6664161 0.3767937 0.2266467

Rotation:
      PC1      PC2      PC3      PC4      PC5
Price 0.5737259 0.2558755 0.77345688 -0.04239327 -0.0730312
Function 0.4101979 0.5814345 -0.50999927 -0.46583216 0.1287354
easy 0.2600413 0.2859829 -0.22302049 0.87797754 0.1732328
design 0.4523733 -0.6017424 -0.08005497 -0.08610099 0.6476437
color 0.4799096 -0.3906169 -0.29243507 0.05427026 -0.7270775
>
> # 주성분 분석 결과보기
> ls(PCA.model1) # 모든 리스트 보기
[1] "center" "rotation" "scale" "sdev" "x"
```

## [고객만족 데이터] 주성분 분석 (Principal Component Analysis)

> `summary(PCA.model1)` # 일반적으로 누적기여율이 80~85% 이상을 설명하는 최소한의 주성분까지 사용. 여기서는 2개.

Importance of components:

	PC1	PC2	PC3	PC4	PC5
Standard deviation	2.2544	1.5679	0.66642	0.37679	0.22665
Proportion of Variance	0.6215	0.3006	0.05431	0.01736	0.00628
Cumulative Proportion	0.6215	0.9221	0.97636	0.99372	1.00000

> `PCA.model1$rotation` # Eigenvector(고유행렬) 보기

	PC1	PC2	PC3	PC4	PC5
Price	0.5737259	0.2558755	0.77345688	-0.04239327	-0.0730312
Function	0.4101979	0.5814345	-0.50999927	-0.46583216	0.1287354
easy	0.2600413	0.2859829	-0.22302049	0.87797754	0.1732328
design	0.4523733	-0.6017424	-0.08005497	-0.08610099	0.6476437
color	0.4799096	-0.3906169	-0.29243507	0.05427026	-0.7270775

> `PCA.model1$sdev` # Eigenvalue(고유값)의 제곱근 보기. 즉 이 값에 제곱을 하면 고유값이 됨.

```
[1] 2.2543608 1.5678592 0.6664161 0.3767937 0.2266467
```

> `PCA.model1$sdev^2` # 고유값.

```
[1] 5.08214264 2.45818249 0.44411043 0.14197348 0.05136875
```

## [고객만족 데이터] 주성분 (Principal Component)

---

- 제1 주성분 (P1) 구하기

$$P1 = \underline{X}e_1$$

$\underline{X} = (x_1, x_2, x_3, x_4, x_5)$  는 데이터 행렬. 5개의 문항이 각각  $x_1, \dots, x_5$ 로 표시됨.

- 제1 주성분은 일종의 ‘전반적인 만족도’를 나타낸다고 할 수 있다.

$$P1 = (0.574 x_1 + 0.410 x_2 + 0.260 x_3 + 0.452 x_4 + 0.480 x_5)$$

- 제2 주성분 (P2)은 가격, 성능, 편리성 등 제품의 ‘내형적 요인’과 디자인, 색상 등 ‘외형적 요인’의 차이를 나타낸다고 해석할 수 있다.

$$P2 = \underline{X}e_2$$

$$P2 = (0.256 x_1 + 0.581 x_2 + 0.286 x_3) - (0.602 x_4 + 0.391 x_5)$$

## [고객만족 데이터] 주성분 점수 (Principal Component Score)

```
> Satis[,3:7] # raw data
  Price Function easy design color
1      1         2    4       1     1
2      1         2    3       2     1
3      2         5    5       2     2
4      2         5    5       2     2
5      1         2    3       4     3
6      1         3    4       1     1
7      4         5    5       3     3
8      1         3    4       4     4
9      3         3    5       5     4
10     5         5    5       4     4
```

주성분 점수는 각각의 개인에 대한 특성을 말해준다.

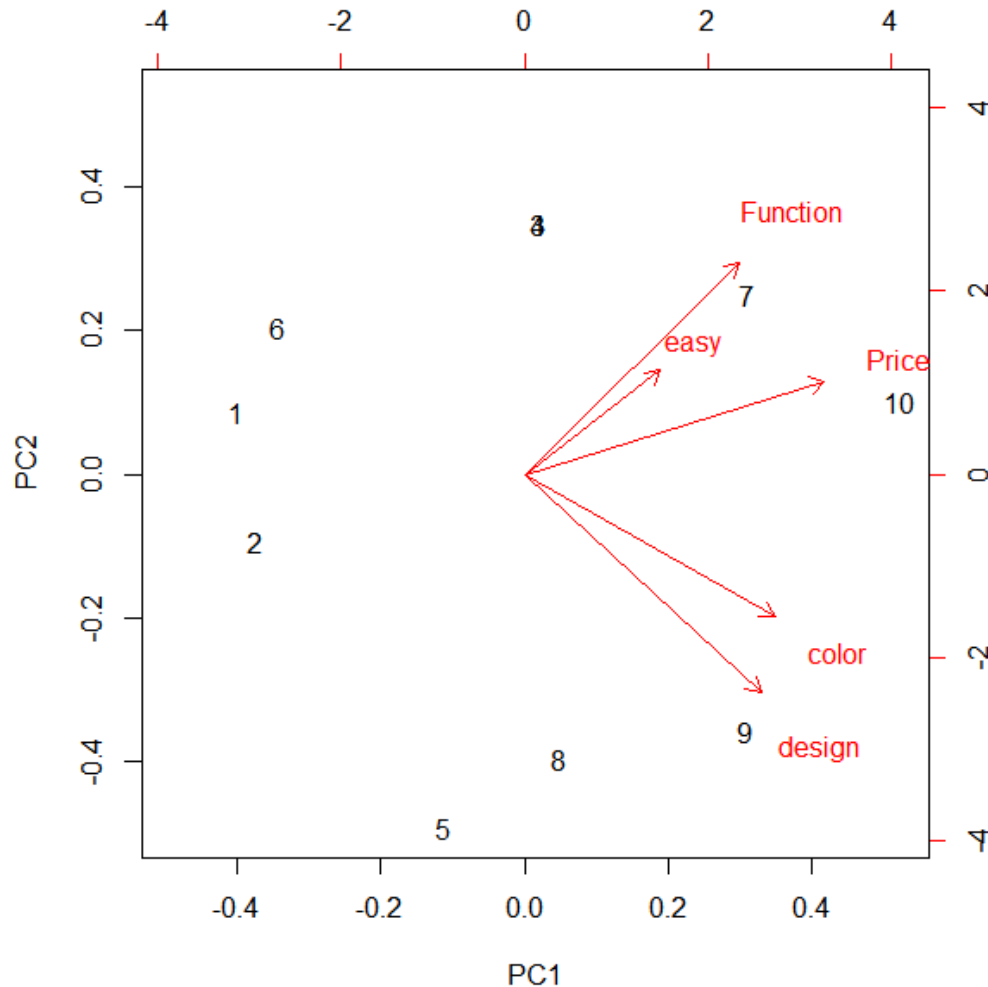
ex) 1번 고객은 전반적인 만족도가 낮다.  
4번 고객은 내형적 요인의 만족도가 외형적 요인의 만족도에 비해 높다.  
5번 고객은 4번 고객과 반대  
10번 고객은 전반적인 만족도가 높다.

```
> PCA.model1$x # 주성분 점수 보기
```

	PC1	PC2	PC3	PC4	PC5
[1,]	-2.8585440	0.4296520	0.56385404	0.555563982	-0.23988087
[2,]	-2.6662120	-0.4580732	0.70681956	-0.408514552	0.23453003
[3,]	0.1380997	1.7234546	-0.78819742	-0.038178976	0.16709291
[4,]	0.1380997	1.7234546	-0.78819742	-0.038178976	0.16709291
[5,]	-0.8016464	-2.4427918	-0.03816052	-0.472176023	0.07566233
[6,]	-2.4483461	1.0110865	0.05385477	0.089731817	-0.11114552
[7,]	2.2178345	1.2428462	0.38622629	-0.154796250	-0.05840333
[8,]	0.3485024	-1.9659914	-1.06361535	-0.005760388	-0.34944707
[9,]	2.2083689	-1.7699999	0.18022294	0.701329621	0.32536699
[10,]	3.7238432	0.5063624	0.78719312	-0.229020255	-0.21086838

## 그래프로 주성분 보기 (공분산행렬에 기초)

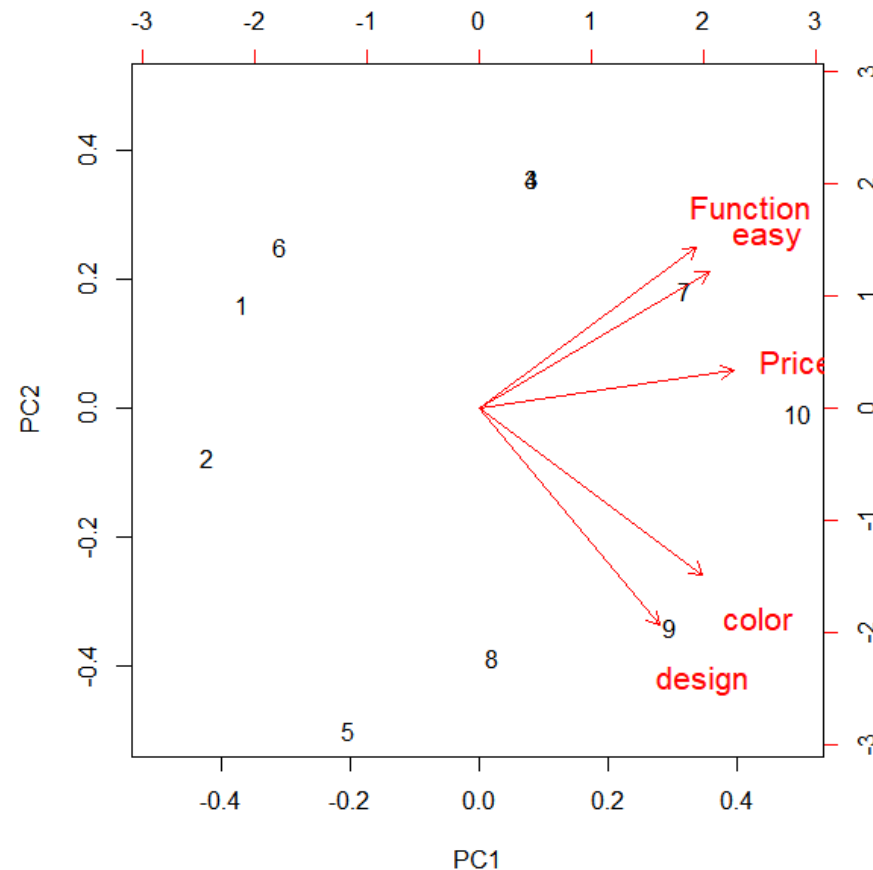
```
> biplot(PCA.model1,cex=c(1,1))
```





## 그래프로 주성분 보기 (상관행렬에 기초)

```
> PCA.model2<-prcomp(Satis[,3:7],center=TRUE,scale=TRUE) #  
correlation matrix 이용.  
> biplot(PCA.model2,cex=c(1,1.3))
```



## 2.3 보유 주성분 개수에 관한 판정

- 전체변이에 대한 공헌도

$$✓ C(m) = \begin{cases} 100(\hat{\delta}_1 + \hat{\delta}_2 + \cdots + \hat{\delta}_m)/\text{tr}(\mathbf{S}), & (\mathbf{S} \text{를 사용하는 경우}) \\ 100(\hat{\delta}_1 + \hat{\delta}_2 + \cdots + \hat{\delta}_m)/p, & (\mathbf{R} \text{을 사용하는 경우}) \end{cases}$$

✓  $C(m) > c^*$ 를 만족하는 최소 정수값  $m$

> # 전체변이에 대한 공헌도 (공분산 행렬에 기초)

> 100\*PCA.model1\$sdev^2/sum(PCA.model1\$sdev^2) #  
sum(PCA.model1\$sdev^2) 는 tr(S) 이다.

[1] 62.1457659 30.0592967 5.4306982 1.7360887 0.6281504

# 전체변이에 대한 공헌도 (상관행렬에 기초)

> 100\*PCA.model2\$sdev^2/sum(PCA.model2\$sdev^2) #  
sum(PCA.model2\$sdev^2) 는 p (=변수의 개수) 이다.

[1] 61.4862423 30.0577368 4.9430539 2.9232072 0.5897599

## 2.3 보유 주성분 개수에 관한 판정

- 전체변이에 대한 공헌도

✓ 
$$C(m) = \begin{cases} 100(\hat{\delta}_1 + \hat{\delta}_2 + \cdots + \hat{\delta}_m)/\text{tr}(\mathbf{S}), & (\mathbf{S} \text{를 사용하는 경우}) \\ 100(\hat{\delta}_1 + \hat{\delta}_2 + \cdots + \hat{\delta}_m)/p, & (\mathbf{R} \text{을 사용하는 경우}) \end{cases}$$

✓  $C(m) > c^*$ 를 만족하는 최소 정수값  $m$

> `summary(PCA.model1)` # 일반적으로 누적기여율이 80~85% 이상을 설명하는 최소한의 주성분까지 사용. 여기서는 2개.

Importance of components:

	PC1	PC2	PC3	PC4	PC5
Standard deviation	2.2544	1.5679	0.66642	0.37679	0.22665
Proportion of Variance	0.6215	0.3006	0.05431	0.01736	0.00628
Cumulative Proportion	0.6215	0.9221	0.97636	0.99372	1.00000

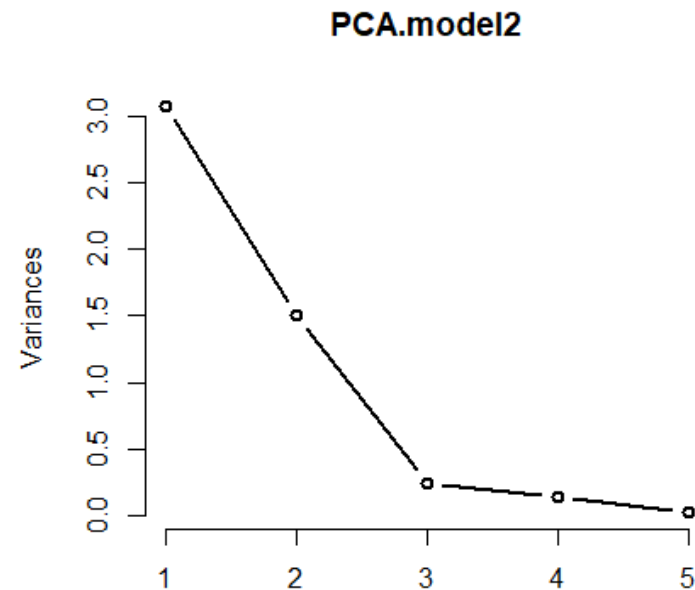
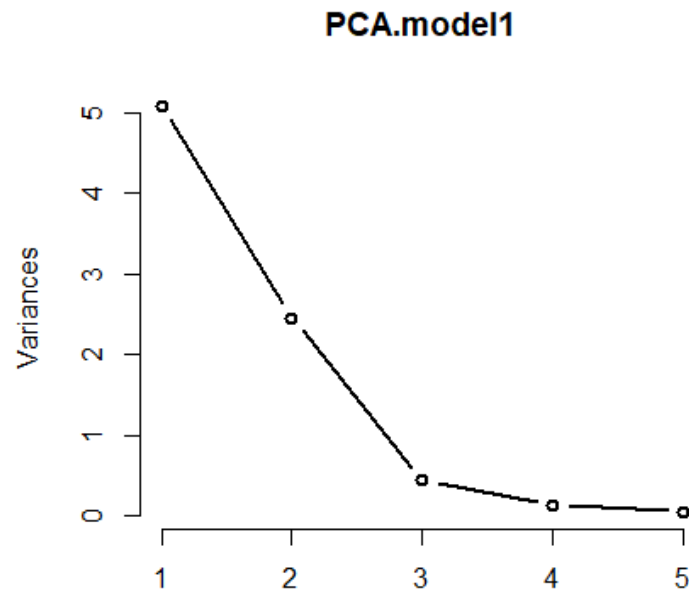
## 2.3 보유 주성분 개수에 관한 판정

- 고유값의 크기 (Kaiser의 규칙)
  - ✓ 주성분이 상관행렬에 기초하고 있다면 상관행렬의 대각원소가 1이므로 모든 주성분의 분산은 1이 된다. 따라서 1보다 작은 고유값을 가지는 주성분은 원래 반응변수 중의 어느 하나 보다 작은 정보를 가지므로 보유할 가치가 없다고 하겠다.
  - ✓ 상관행렬에 기초하여 분석을 수행하는 경우, 고유값이 1이상인 주성분을 보유하는 'Kaiser(1960)의 규칙'을 기준으로 사용할 수 있다.

## 2.3 보유 주성분 개수에 관한 판정

- Scree Plot 을 보고 판단함

```
> par(mfrow=c(1,2))  
> screeplot(PCA.model1,type='l',lwd=2)  
> screeplot(PCA.model2,type='l',lwd=2)
```



## 2.4 주성분 분석의 이론적 배경.

- $\Sigma$ : 공분산행렬 ( $p \times p$ )
- $\delta_1 \geq \delta_2 \geq \dots \geq \delta_k \geq \dots \geq \delta_{p-1} \geq \delta_p$ : 고유값  
 $\vdots$   
 $\mathbf{e}_k = (e_{1k}, e_{2k}, \dots, e_{pk})'$ : 고유벡터,  $k=1, \dots, p$
- 선형결합  $y = \mathbf{a}'\mathbf{x} = a_1x_1 + a_2x_2 + \dots + a_px_p$ 에 대하여
  - ①  $\text{Var}(y) = \text{Var}(\mathbf{a}'\mathbf{x})$ 를 최대로 하는  $\mathbf{a}$ 를 구하면?  
 $\rightarrow \mathbf{a} = \mathbf{e}_1, \text{Var}(\mathbf{e}_1'\mathbf{x}) = \text{Var}(y_1) = \delta_1, y_1$ : 첫번째 주성분
  - ②  $\text{Cov}(\mathbf{e}_1'\mathbf{x}, \mathbf{a}'\mathbf{x}) = 0$ 이고,  $\text{Var}(y) = \text{Var}(\mathbf{a}'\mathbf{x})$ 를 최대로 하는  $\mathbf{a}$ 를 구하면?  
 $\rightarrow \mathbf{a} = \mathbf{e}_2, \text{Var}(\mathbf{e}_2'\mathbf{x}) = \text{Var}(y_2) = \delta_2, y_2$ : 두번째 주성분
  - ③  $\text{Cov}(\mathbf{e}_1'\mathbf{x}, \mathbf{a}'\mathbf{x}) = \text{Cov}(\mathbf{e}_2'\mathbf{x}, \mathbf{a}'\mathbf{x}) = 0$ 이고,  $\text{Var}(y) = \text{Var}(\mathbf{a}'\mathbf{x})$ 를 최대로 하는  $\mathbf{a}$ 를 구하면?  
 $\rightarrow \mathbf{a} = \mathbf{e}_3, \text{Var}(\mathbf{e}_3'\mathbf{x}) = \text{Var}(y_3) = \delta_3, y_3$ : 세번째 주성분
  - ...

## 고유값과 고유벡터의 성질

$\delta_1 \geq \delta_2 \geq \dots \geq \delta_k \geq \dots \geq \delta_{p-1} \geq \delta_p$ : 고유값은 항상 0보다 크다.

고유벡터는 0 이 아니고, 고유벡터의 Norm 은 항상 1이다.

- 즉,  $\mathbf{e}_k = (e_{1k}, e_{2k}, \dots, e_{pk})'$  일때  $\sqrt{e_{1k}^2 + e_{2k}^2 + \dots + e_{pk}^2} = 1$

# 고유벡터들은 Norm 의 값이 모두 1이다.

```
> sum(PCA.model1$rotation[,1]^2)
```

```
[1] 1
```

```
> sum(PCA.model1$rotation[,2]^2)
```

```
[1] 1
```

```
> sum(PCA.model1$rotation[,3]^2)
```

```
[1] 1
```

```
> sum(PCA.model1$rotation[,4]^2)
```

```
[1] 1
```

```
> sum(PCA.model1$rotation[,5]^2)
```

```
[1] 1
```

## 대수적 최적성과 통계적 의미

- 전체 분산의 합계 :  $\text{tr}(\Sigma) = \sum_{i=1}^p \sigma_i^2 = \sum_{i=1}^p \delta_i$
- $k$  번째 주성분  $y_k$ 가 전체 데이터 변동을 설명하는 부분 :  $\delta_k / (\delta_1 + \cdots + \delta_p)$
- 첫  $m$ 개의 주성분  $y_1, y_2, \dots, y_m$ 에 의해 설명되는 부분 :  $(\delta_1 + \cdots + \delta_m) / \text{tr}(\Sigma)$

예제 1)  $\Sigma (= \mathbf{P}) = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$

✓  $\delta_1 = 1 + \rho, \delta_2 = 1 - \rho$

✓  $\mathbf{e}'_1 = (1/\sqrt{2}, 1/\sqrt{2}), \mathbf{e}'_2 = (1/\sqrt{2}, -1/\sqrt{2})$

✓  $y_1 = \mathbf{e}'_1 \mathbf{x} = (x_1 + x_2)/\sqrt{2}, y_2 = \mathbf{e}'_2 \mathbf{x} = (x_1 - x_2)/\sqrt{2}$

✓  $\rho \approx 1 \rightarrow \delta_1 \approx 2, \delta_2 \approx 0$

✓  $\rho \approx 0 \rightarrow \delta_1 \approx 1, \delta_2 \approx 1$



## 예제 1 풀이 (앞 페이지 예제)

---

$$S \underline{e}_1 = \delta_1 \underline{e}_1 \quad \delta_1 = \text{첫번째 고유값} \quad \underline{e}_1 = \text{첫번째 고유벡터}$$

위의 식을 앞 페이지 예제에 대입해보면 다음과 같다.

$$S = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

$$(S - \delta_1 E) \underline{e}_1 = 0$$

위에서, 고유벡터는 정의에 의해 0 이 아니다. 따라서  $(S - \delta_1)$  의 역행렬이 존재하면 안된다. 만약  $(S - \delta_1)$  의 역행렬이 존재한다면 양변에 역행렬을 곱해주어  $\underline{e}_1 = 0$  이 성립하기 때문에 정의에 위배됨.

## 예제 1 풀이 ( 첫번째 고유값 구하기)

---

역행렬이 존재하지 않기 위해서는  $\det(S - \delta_1) = 0$  을 만족해야함. ( $\det$  = determinant)

$$\det \begin{pmatrix} a & b \\ c & d \end{pmatrix} = ad - bc$$

$$\det(S - \delta_1 E) = \det \begin{pmatrix} 1 - \delta_1 & \rho \\ \rho & 1 - \delta_1 \end{pmatrix} = (1 - \delta_1)^2 - \rho^2$$

$(1 - \delta_1)^2 - \rho^2 = (1 - \delta_1 - \rho)(1 - \delta_1 + \rho) = 0$  를 풀어 고유값 구함.

$$\delta_1 = 1 + \rho$$

( $\delta_1 = 1 - \rho$  도  $\det$ 를 영으로 만들지만  $1 + \rho$ 보다 작으므로 두번째 고유값이 된다. 즉,  $\delta_2 = 1 - \rho$  )

## 예제 1 풀이 (첫번째 고유벡터 구하기)

$$(S - \delta_1 E)\underline{e}_1 = \begin{pmatrix} 1 - \delta_1 & \rho \\ \rho & 1 - \delta_1 \end{pmatrix} \begin{pmatrix} e_1 \\ e_2 \end{pmatrix} = \begin{pmatrix} -\rho & \rho \\ \rho & -\rho \end{pmatrix} \begin{pmatrix} e_1 \\ e_2 \end{pmatrix} = \mathbf{0}$$

위의 식은 다음과 같이 정리된다.

$$-\rho e_1 + \rho e_2 = 0 \quad (1)$$

$$\rho e_1 - \rho e_2 = 0 \quad (2)$$

(1)번식에서 (2)번식을 빼면  $-2\rho e_1 + 2\rho e_2 = 0$  또는  $e_1 = e_2$

고유벡터의 Norm 은 1 이므로

$$\sqrt{e_1^2 + e_2^2} = \sqrt{e_1^2 + e_1^2} = \sqrt{2e_1^2} = 1 \text{ 또는 } e_1 = 1/\sqrt{2}$$

$$\text{따라서 첫번째 고유벡터} = \underline{e}_1 = \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}$$

## 예제 1 풀이 ( 두번째 고유값 구하기 )

역행렬이 존재하지 않기 위해서는  $\det(S - \delta_2) = 0$  을 만족해야함. (det = determinant)

$$\det \begin{pmatrix} a & b \\ c & d \end{pmatrix} = ad - bc$$

$$\det(S - \delta_2 E) = \det \begin{pmatrix} 1 - \delta_2 & \rho \\ \rho & 1 - \delta_2 \end{pmatrix} = (1 - \delta_2)^2 - \rho^2$$

$(1 - \delta_2)^2 - \rho^2 = (1 - \delta_2 - \rho)(1 - \delta_2 + \rho) = 0$  를 풀어 고유값 구함.

$$\delta_2 = 1 - \rho$$

( $\delta_2 = 1 + \rho$  도 det를 영으로 만들지만  $1 - \rho$ 보다 크기 때문에 첫번째 고유값이 된다. 즉,  $\delta_1 = 1 + \rho$  )

## 예제 1 풀이 (두번째 고유벡터 구하기)

$$(S - \delta_2 E)\underline{e}_2 = \begin{pmatrix} 1 - \delta_2 & \rho \\ \rho & 1 - \delta_2 \end{pmatrix} \begin{pmatrix} e_1 \\ e_2 \end{pmatrix} = \begin{pmatrix} \rho & \rho \\ \rho & \rho \end{pmatrix} \begin{pmatrix} e_1 \\ e_2 \end{pmatrix} = \mathbf{0}$$

위의 식은 다음과 같이 정리된다.

$$\rho e_1 + \rho e_2 = 0 \quad (1)$$

$$\rho e_1 + \rho e_2 = 0 \quad (2)$$

(1)번식에서 (2)번식을 더하면  $2\rho e_1 + 2\rho e_2 = 0$  또는  $e_1 = -e_2$

고유벡터의 Norm 은 1 이므로

$$\sqrt{e_1^2 + e_2^2} = \sqrt{2e_1^2} = 1 \text{ 또는 } e_1 = 1/\sqrt{2}, e_2 = -1/\sqrt{2}$$

$$\text{따라서 두번째 고유벡터} = \underline{e}_2 = \begin{pmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix}$$

## 예제 1 풀이 (첫번째, 두번째 주성분 구하기)

---

$$y_1 = \mathbf{e}'_1 \mathbf{x} = (x_1 + x_2)/\sqrt{2}$$

$$y_2 = \mathbf{e}'_2 \mathbf{x} = (x_1 - x_2)/\sqrt{2}$$

여기서  $\mathbf{x}$  는 데이터의 설명변수.

$y_1$ : 첫번째 주성분

$y_2$ : 두번째 주성분

## 연습문제 1

---

모집단 공분산행렬이 다음과 같이 주어졌을 때,

$$A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$$

1. 고유값을 구하시오.
2. 고유벡터를 구하시오.
3. 첫번째와 두번째 주성분을 구하시오.

## 공분산행렬 대 상관행렬

- 상관행렬을 이용하는 주요 이유

- ✓ 상관행렬에 기초한 분석결과가 공분산행렬의 경우보다 더 직접적으로 비교하기가 좋다. 공분산행렬에서 얻어진 주성분은 관찰변수에 대해 사용된 측정단위를 달리함에 따라 완전히 다른 결과를 주지만, 상관행렬은 그의 척도불변성으로 인해 측정단위와는 무관하게 되기 때문이다.
- ✓ 관찰변수들의 분산들 사이에 많은 차이가 있다면 첫 주성분은 가장 큰 분산을 가진 변수들에 의해 결정되는 경향이 있다. 따라서 측정단위의 선택에 자의성이 개재되어 있을 경우에는, 각 변수에 동일한 변이성을 부여하는 상관행렬에 기초한 주성분분석이 선호된다.

- 공분산행렬을 이용하는 주요 이유

- ✓ 공분산행렬에 기초하여 얻은 표본 주성분은 그의 분포가 상관행렬의 그것보다 훨씬 간단하기 때문에 표본 주성분을 이용하여 모집단 주성분에 관한 통계적 추론을 시도할 경우 공분산행렬의 이용이 선호된다.
- ✓ 모든 관찰변수들이 동일한 단위로 측정될 경우 공분산행렬의 이용이 선호된다. 이 때 상관행렬을 주성분에서 이용하는 것은 관찰변수들을 측정할 때 임의의 단위를 선택하는 것과 동일하다는 점에서 비판될 수 있다.



## 공분산행렬 대 상관행렬 (변수의 측정단위가 다른 데이터)

	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	$x_6$	$x_7$	$x_8$
변 수	RBOOLD	PLATE	WBLOOD	NEUT	LYMPH	BILIR	SODIUM	POTASS
표준편차	0.37	41.25	1.94	0.08	0.08	4.04	2.73	0.30

$$\mathbf{R} = \begin{pmatrix} 1 & & & & & & & \\ 0.29 & 1 & & & & & & \\ 0.20 & 0.42 & 1 & & & & & \\ -0.06 & 0.29 & 0.42 & 1 & & & & \\ -0.11 & -0.38 & -0.52 & -0.88 & 1 & & & \\ -0.25 & -0.35 & -0.44 & -0.08 & 0.21 & 1 & & \\ -0.23 & -0.16 & -0.15 & 0.02 & 0.03 & 0.19 & 1 & \\ 0.06 & -0.13 & -0.08 & -0.13 & 0.15 & 0.08 & 0.42 & 1 \end{pmatrix}$$

$$\mathbf{S} = \begin{pmatrix} 0.14 & & & & & & & \\ 4.43 & 1701.56 & & & & & & \\ 0.14 & 33.61 & 3.76 & & & & & \\ -0.00 & 0.96 & 0.07 & 0.01 & & & & \\ -0.00 & -1.25 & -0.08 & -0.01 & 0.01 & & & \\ -0.37 & -58.33 & -3.45 & -0.03 & 0.07 & 16.32 & & \\ -0.23 & -18.02 & -0.79 & 0.00 & 0.01 & 2.10 & 7.45 & \\ 0.01 & -1.61 & -0.05 & -0.00 & 0.00 & 0.10 & 0.34 & 0.09 \end{pmatrix}$$

## 표본상관행렬 대 표본공분산행렬

변수\주성분	$y_1$	$y_2$	$y_3$	$y_4$
$x_1$	0.19	-0.42	0.41	0.65
$x_2$	0.40	-0.15	0.17	0.05
$x_3$	0.46	0.00	0.16	-0.27
$x_4$	0.43	0.47	-0.17	0.17
$x_5$	-0.49	-0.36	0.08	-0.19
$x_6$	-0.32	0.32	-0.27	0.64
$x_7$	-0.18	0.54	0.41	-0.17
$x_8$	-0.17	0.25	0.71	0.09
고유값	2.80	1.53	1.25	0.78
설명비율(%)	35.01	19.10	15.56	9.75

공분산행렬에 기초한 첫 4개의 주성분

상관행렬에 기초한 첫 4개의 주성분

변수\주성분	$y_1$	$y_2$	$y_3$	$y_4$
$x_1$	0.00	-0.02	-0.02	0.00
$x_2$	1.00	0.04	0.00	-0.01
$x_3$	0.02	-0.19	0.01	0.98
$x_4$	0.00	0.00	0.00	0.02
$x_5$	0.00	0.00	0.00	-0.02
$x_6$	-0.03	0.96	-0.20	0.19
$x_7$	-0.01	0.19	0.98	0.03
$x_8$	0.00	0.01	0.05	0.00
고유값	1704.45	15.06	6.98	2.64
설명비율(%)	98.56	0.87	0.40	0.15

변수들의 단위(unit)가  
다를 때에는 상관행렬을 사용.