

주성분 분석(PCA)

<http://adnoctum.tistory.com/977>

자동차 1만대에 대한 특성을 파악하고자 다음과 같은 데이터를 수집하였다.

➔ 차의 가격, 무게, 색상, 최대탑승인원수, 배기용량, 제조회사 본사의 위도와 경도, 운전하는 사람의 키를 측정한 데이터, 자동차 문의 수, 자동차 바퀴의 수, 자동차 핸들의 모양, 차체의 높이.

위에서 측정한 특성 중 어느 변수가 1만개의 데이터를 가장 잘 나타낼 수 있을까?
즉, 각 데이터의 차이가 가장 잘 두드러지게 하는 특성은 무엇일까?

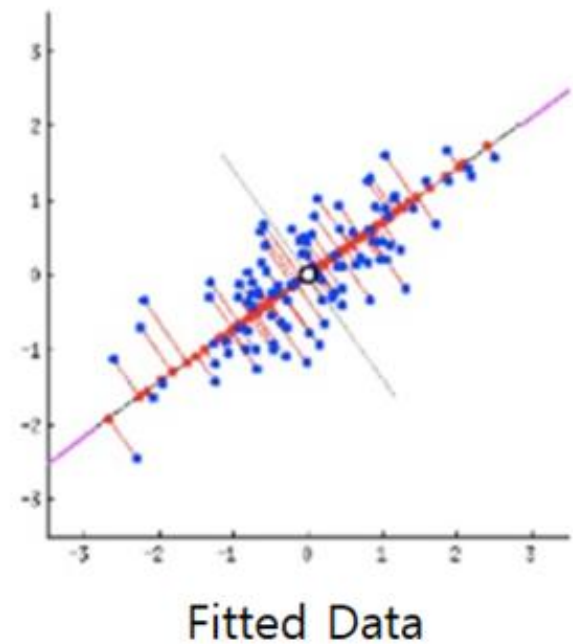
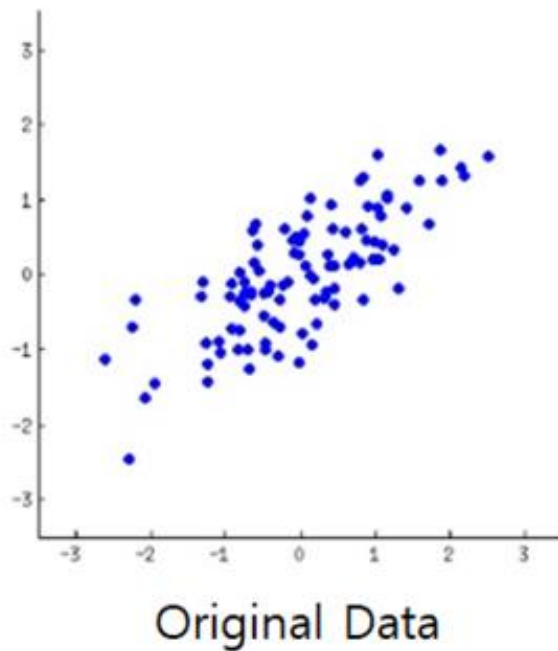
자동차 핸들의 모양, 문의 수 등으로는 각 자동차를 다른 자동차와 구분하기 좋은 특성은 아닐 것이다(동일한 값을 갖는 자동차들이 많기 때문에)반면 가격/무게/차체의 높이 등은 각 자동차가 서로 다른 값을 갖는 경우가 많기 때문에 이 값들의 적절한 조합은 각 차를 다른 차와 구별시켜 주는 특징이 될 수 있을 것이다.

➔ 큰 분산 값을 활용

➔ 분산 값이 크다는 건 평균에 벗어나 있으므로 특성을 파악하기 좋은 변수가 될 수 있다. ➔ 특성을 잘 표현하는 대표 값이 다양하게 나온다는 의미를 포함

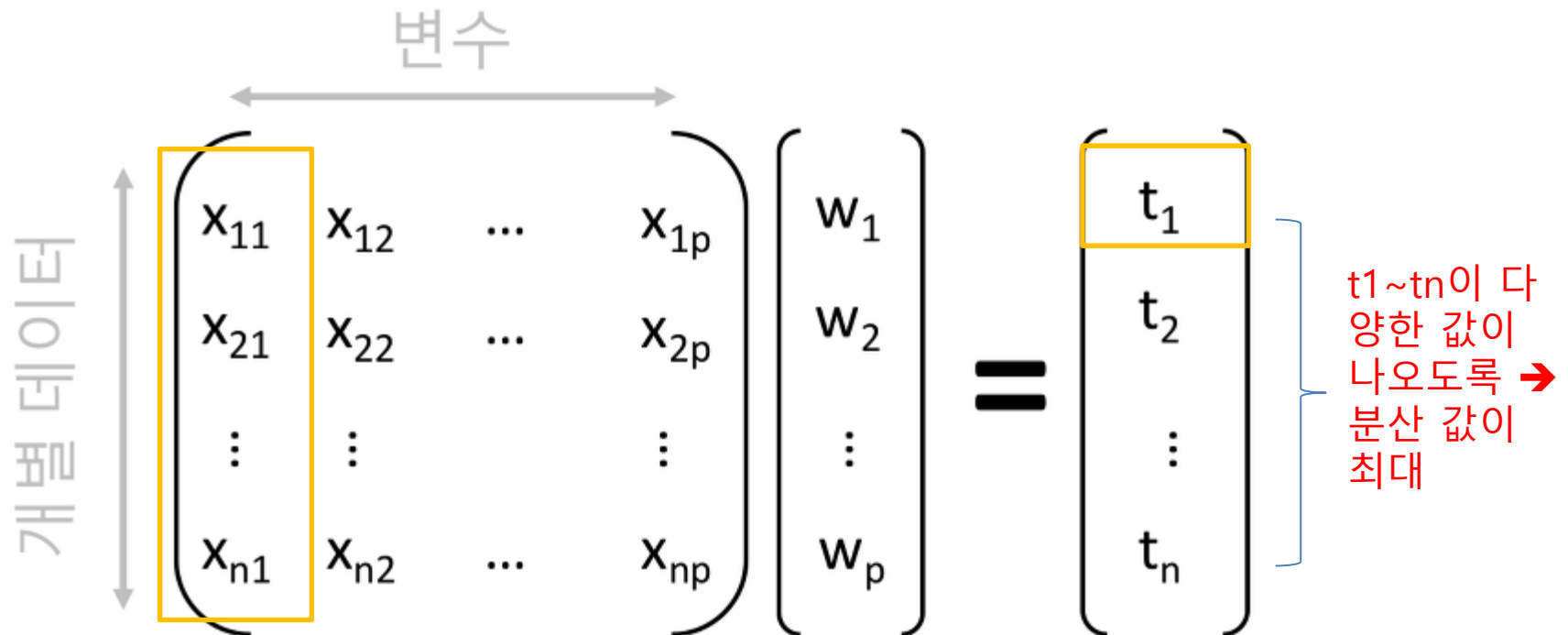
주성분 분석(PCA)

분산을 크게 한다는 의미 : 평균으로부터 넓게 퍼지게 되도록, 분산을 잘 설명하는 벡터를 주성분벡터라고 볼 수 있다.



주성분 분석(PCA)

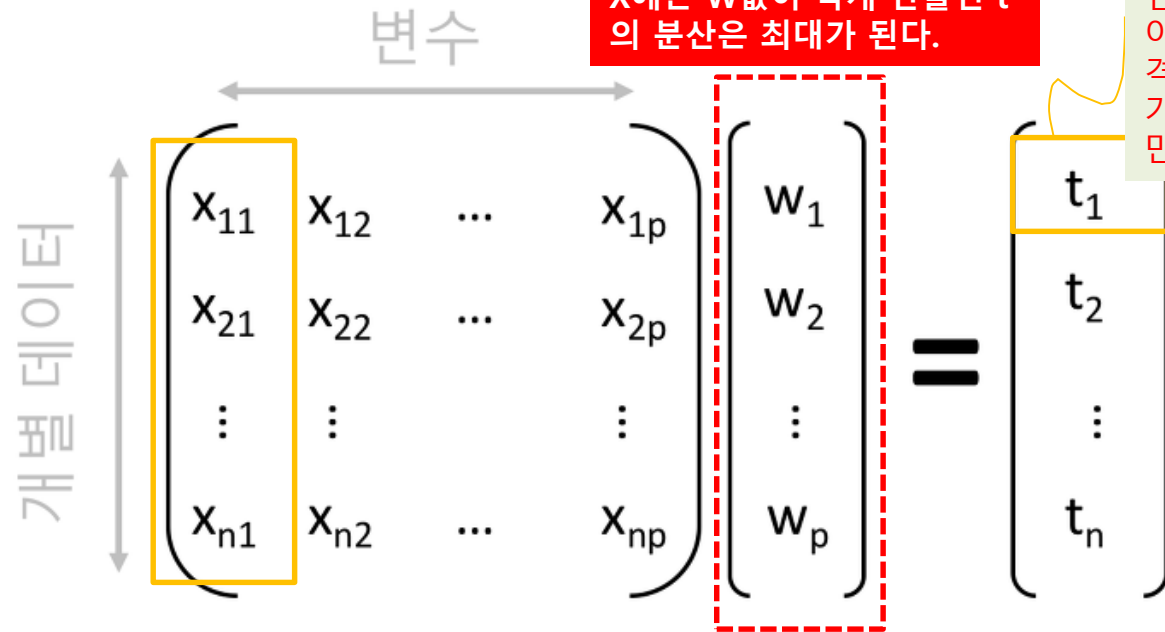
주성분 분석은 이러한 특징(특성이 강한=분산 값이 큰)들을 선형 결합시킬 때 데이터들이 가장 다양한 값을 갖도록 하는 특징의 가중치를 찾아 주는 방법이다.



The diagram illustrates the PCA equation:
$$\begin{matrix} \text{개별 데이터} \updownarrow \\ \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \end{matrix} \begin{matrix} \xleftarrow{\text{변수}} \\ \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_p \end{pmatrix} \end{matrix} = \begin{pmatrix} t_1 \\ t_2 \\ \vdots \\ t_n \end{pmatrix}$$
 The matrix of individual data points is highlighted with a yellow box. The resulting principal components t_1, t_2, \dots, t_n are also highlighted with a yellow box. A blue bracket on the right side of the equation points to the principal components, with the text: $t_1 \sim t_n$ 이 다양한 값이 나오도록 → 분산 값이 최대

주성분 분석의 핵심은 이렇게 찾아진 가중치 w 에 의해 변환된 새로운 값인 t_1, \dots, t_n 이 최대의 분산을 갖도록 w 를 찾는다는 점

주성분 분석(PCA)



T값들의 분산이 최대가 되게 하려면 분산 값이 큰 X에는 W값을 크게, 분산 값이 작은 X에는 W값이 작게 만들면 t의 분산은 최대가 된다.

$t_1 \sim t_n$ 이 다양한 값이 나오도록 하려면 분산 값이 큰 변수의 가중치의 영향력 즉 w_i 의 값이 커진다. 즉, 평균값에 가까운 값이 아닌 특별한 즉, 평균값이 먼 성격을 가진 즉, 분산 값이 큰 변수가 큰 w (가중치)를 갖는 벡터가 만들어 진다.

$$\begin{aligned} t_1 &= x_{11} \cdot w_1 + x_{12} \cdot w_2 + \dots + x_{1p} \cdot w_p \\ t_2 &= x_{21} \cdot w_1 + x_{22} \cdot w_2 + \dots + x_{2p} \cdot w_p \\ &\vdots \\ t_n &= x_{n1} \cdot w_1 + x_{n2} \cdot w_2 + \dots + x_{np} \cdot w_p \end{aligned}$$

- p 는 변수(특성)의 개수이다.
- 각각의 개별 데이터는 주성분 분석에 의하여 찾아진 각 변수에 대한 가중치를 이용하여 선형 결합된다.

주성분 분석(PCA)

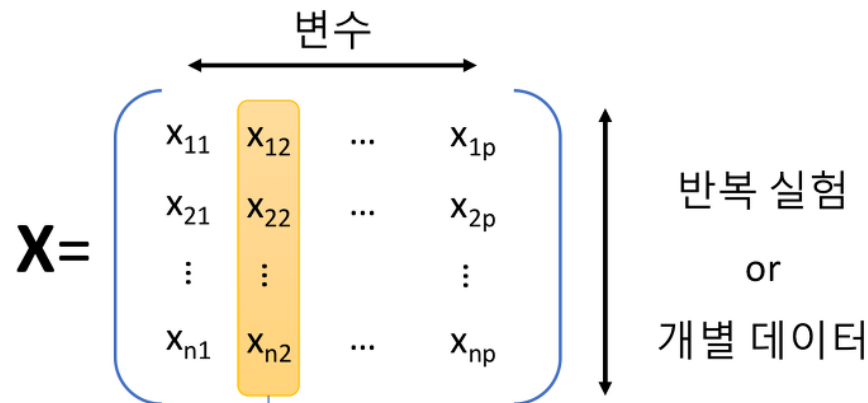
각 데이터는 주성분 분석에 의하여 찾아진 가중치 $w=(w_1, \dots, w_p)$ 에 의하여 새로운 데이터 $t=(t_1, \dots, t_n)$ 으로 변형된다(1). 이 때, 각 데이터는 처음에는 p -차원의 데이터였으나 w 에 의하여 1차원 값인 스칼라로 변경되었다는 점을 주의 깊게 보아야 한다. 주성분 분석의 핵심은 이렇게 찾아진 가중치 w 에 의해 변환된 새로운 값인 t_1, \dots, t_n 이 최대의 분산을 갖도록 w 를 찾는다는 점(2)이다. 최대의 분산을 갖는다는 것의 의미는 데이터 값이 가장 다양하게 된다는 것으로, 1만개의 데이터가 될 수 있으면 같은 값을 갖는 데이터가 없게 된다는 것과 일맥 상통한다. 만약 자동차의 바퀴 수, 로만 생각해 보면 1만개 중 대략 80%~90%가 4를 갖게 될 것이다. 그러나 차체의 높이와 무게는 바퀴의 수보다는 다양할 것이다. 따라서 이 경우 주성분 분석을 하게 되면 바퀴에 해당하는 가중치는 차체의 높이와 무게에 해당하는 가중치보다 작게 나올 것이다.

주성분 분석(PCA)의 수학적 분석

PCA 는 주어진 데이터 \mathbf{X} 에 대하여

$$\mathbf{T} = \mathbf{X}\mathbf{w}$$

인 $\mathbf{T}=(\mathbf{t}_1, \dots, \mathbf{t}_m)$ 에 대하여 각 \mathbf{t}_i 가 최대의 분산을 갖도록 \mathbf{w} 를 찾는 방법이다.
이 때 \mathbf{X} 는 다음과 같다.



평균을 0으로 맞춤.

원본 데이터에서 컬럼 벡터의 평균을 0 으로 맞추어 준다. 그 후 주성분 분석은 $\mathbf{X}^T\mathbf{X}$ 의 eigenvector 의 크기 순으로 정렬하여 각각에 해당하는 eigenvector 를 구하면 그 순서에 해당하는 가중치가 된다. 실제 k-번째 주성분은 원본 데이터를 k-번째 가중치에 내적하면 얻어 진다.

주성분 분석(PCA)의 수학적 분석

PCA 는 주어진 데이터 \mathbf{X} 에 대하여

$$\mathbf{T} = \mathbf{X}\mathbf{w}$$

인 $\mathbf{T}=(\mathbf{t}_1, \dots, \mathbf{t}_m)$ 에 대하여 각 \mathbf{t}_i 가 최대의 분산을 갖도록 \mathbf{w} 를 찾는 방법이다.
이 때 \mathbf{X} 는 다음과 같다.

$$\begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \begin{pmatrix} w_{1(1)} & w_{1(2)} & \dots & w_{1(m)} \\ w_{2(1)} & w_{2(2)} & \dots & w_{2(m)} \\ \vdots & \vdots & & \vdots \\ w_{p(1)} & w_{p(2)} & \dots & w_{p(m)} \end{pmatrix} = \begin{pmatrix} t_{(1)1} & t_{(1)2} & \dots & t_{(1)m} \\ t_{(2)1} & t_{(2)2} & \dots & t_{(2)m} \\ \vdots & \vdots & & \vdots \\ t_{(n)1} & t_{(n)2} & \dots & t_{(n)m} \end{pmatrix}$$

$w_{(1)} \quad w_{(2)} \quad w_{(m)}$ $t_1 \quad t_2 \quad t_m$

앞에서는 \mathbf{w} 를 마치 1개의 열벡터인 것처럼 말했으나 실제로는 자신이 찾고자 하는 개수 m 개인 열로 된 벡터이다. 그래서 $\mathbf{T}=\mathbf{X}\mathbf{w}$ 를 행렬식으로 표현하면 위와 식과 같다.

주성분 분석(PCA)의 수학적 분석

제 1 주성분은 다음과 같이 표현된다.

$$\begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \begin{pmatrix} w_{1(1)} \\ w_{2(1)} \\ \vdots \\ w_{p(1)} \end{pmatrix} = \begin{pmatrix} t_{(1)1} \\ t_{(1)2} \\ \vdots \\ t_{(1)n} \end{pmatrix}$$

제 1 주성분 $\mathbf{t}_1 = (t_{1(1)}, t_{2(1)}, \dots, t_{n(1)})$ 의 분산을 최대화 시키도록 $\mathbf{w}_{(1)} = (w_{1(1)}, w_{2(1)}, \dots, w_{p(1)})$ 을 찾는 것이다. 마찬가지로 그 다음의 제 2 주성분은 다음과 같이 표현된다.

주성분 분석(PCA)의 수학적 분석

이런 식으로 원하는 m 개의 주성분을 찾을 수 있다. 일반적으로 그래프로 표현하기 위해 제 1 과 제 2 주성분을 찾아 x 와 y 축으로 표현하거나 아니면 제 3 주성분까지 찾아서 z 축으로 표현해서 데이터를 시각화하곤 한다. 또한, 각 주성분이 원래의 데이터의 변화량의 몇 %를 대변하는지를 계산할 수도 있다. 또한, 각 $\mathbf{w}_{(i)}$ 는 벡터의 길이가 1 인 unit vector 라는 제약 조건을 둔다.

$$\begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \begin{pmatrix} \mathbf{w}_{1(1)} & \mathbf{w}_{1(2)} \\ \mathbf{w}_{2(1)} & \mathbf{w}_{2(2)} \\ \vdots & \vdots \\ \mathbf{w}_{p(1)} & \mathbf{w}_{p(2)} \end{pmatrix} = \begin{pmatrix} t_{(1)1} & t_{(2)1} \\ t_{(1)2} & t_{(2)2} \\ \vdots & \vdots \\ t_{(1)n} & t_{(2)n} \end{pmatrix}$$

주성분 분석(PCA)의 활용

Principal Component Analysis (PCA)

- 고차원의 데이터를 저차원의 데이터로 환원시키는 차원축소기법
- p 개의 설명변수(x_1, x_2, \dots, x_p)가 있을 때,
설명변수들의 변동을 가장 잘 설명하는 새로운 낮은 차원의 변수($z_1, z_2, \dots, z_m, m < p$)를 구하는 방법
- 차원의 단순화를 통해 서로 상관되어 있는 변수들 간의 복잡한 구조를 분석하는 것이 목적

주성분 분석 실습

- 주성분의 개념이해

<과목별 시험성적>

```
x1 <-c(26,46,57,36,57,26,58,37,36,56,78,95,88,90,52,56)
x2 <-c(35,74,73,73,62,22,67,34,22,42,65,88,90,85,46,66)
x3 <-c(35,76,38,69,25,25,87,79,36,26,22,36,58,36,25,44)
x4 <-c(45,89,54,55,33,45,67,89,47,36,40,56,68,45,37,56)
```

```
score <- cbind(x1,x2,x3,x4)
colnames(score) <- c("국어","영어","수학","과학")
rownames(score) <- 1:16
head(score)
```

국어 영어 수학 과학

1	26	35	35	45
2	46	74	76	89
3	57	73	38	54
4	36	73	69	55

주성분 분석 실습 데이터

```
x1 <-c(26,46,57,36,57,26,58,37,36,56,78,95,88,90,52,56)
x2 <-c(35,74,73,73,62,22,67,34,22,42,65,88,90,85,46,66)
x3 <-c(35,76,38,69,25,25,87,79,36,26,22,36,58,36,25,44)
x4 <-c(45,89,54,55,33,45,67,89,47,36,40,56,68,45,37,56)
```

주성분 분석 실습

- 주성분의 개념

```
result <- prcomp(score)
```

```
result # 주성분 벡터
```

```
Standard deviations (1, .., p=4):
```

```
[1] 30.122748 27.052808 9.076140 6.152386
```

```
Rotation (n x k) = (4 x 4):
```

	PC1	PC2	PC3	PC4
국어	0.6093268	-0.39286407	-0.6126773	-0.3146508
영어	0.7185749	-0.09337973	0.6200124	0.3008572
수학	0.2624323	0.73573272	0.1052861	-0.6154198
과학	0.2085672	0.54372366	-0.4786711	0.6570680

주성분1

주성분2

주성분 분석 실습

• 주성분의 개념이해

```
summary(result) # 요약
Importance of components%s:

                PC1      PC2      PC3      PC4
Standard deviation 30.1227 27.0528 9.07614 6.15239
Proportion of Variance 0.5157 0.4159 0.04682 0.02151
Cumulative Proportion 0.5157 0.9317 0.97849 1.00000
```

summary(result)에 의해

- Cumulative Proportion 이 70~80%이상까지 주성분선택
- 주성분 2개(93.17%)선택

result에 의한 주성분 함수

- 주성분1 = $0.61 \times \text{국어} + 0.72 \times \text{영어} + 0.26 \times \text{수학} + 0.21 \times \text{과학}$
- 주성분2 = $-0.39 \times \text{국어} + -0.09 \times \text{영어} + 0.74 \times \text{수학} + 0.54 \times \text{과학}$
- 주성분1=문과성향 / 주성분2=이과성향

주성분 분석 실습

• 주성분 분석의 목적과 개념

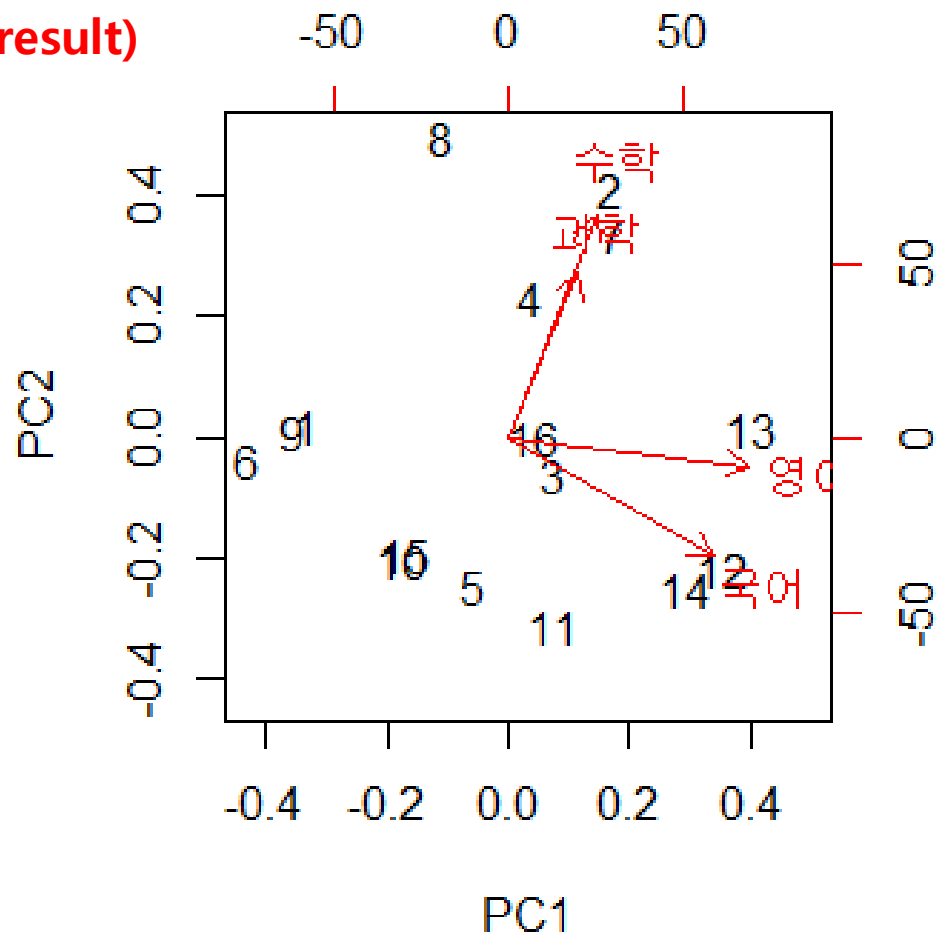
```
##국어, 영어,수학,과학
#16명의 국어점수
x1<-c(26,46,57,36,57,26,58,37,36,56,78,95,88,90,52,56)
x2<-c(35,74,73,73,62,22,67,34,22,42,65,88,90,85,46,66)
x3<-c(35,76,38,69,25,25,87,79,36,26,22,36,58,36,25,44)
x4<-c(45,89,54,55,33,45,67,89,47,36,40,56,68,45,37,56)

score<-cbind(x1,x2,x3,x4)
colnames(score)<-c("국어","영어","수학","과학")
rownames(score)<-1:16
head(score)
#주성분분석(PCA)
result<-prcomp(score)
result
summary(result)
biplot(result)
screeplot(result,npcs=4,type="lines",main="Score")
```

주성분 분석 실습

- 결과화면 해석

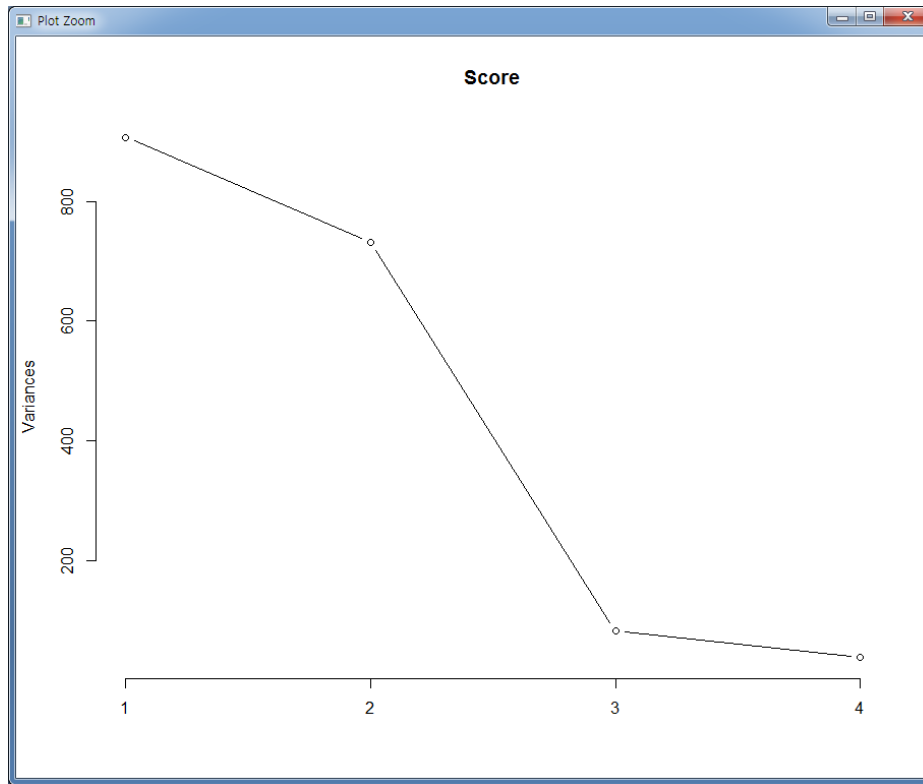
biplot(result)



주성분 분석 실습

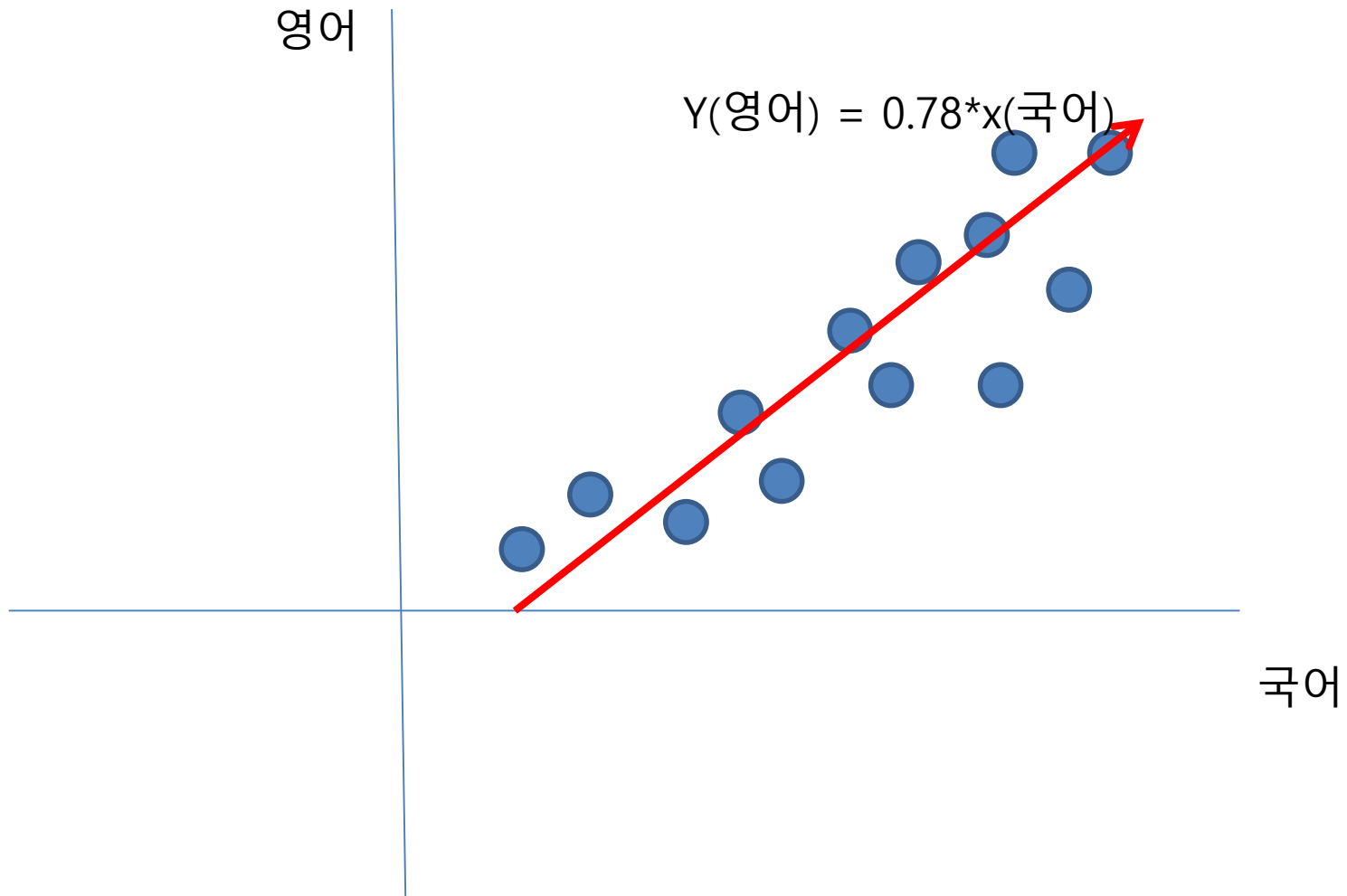
• 결과화면 해석

`screeplot(result,npcs=4,type="lines",main="Score")`



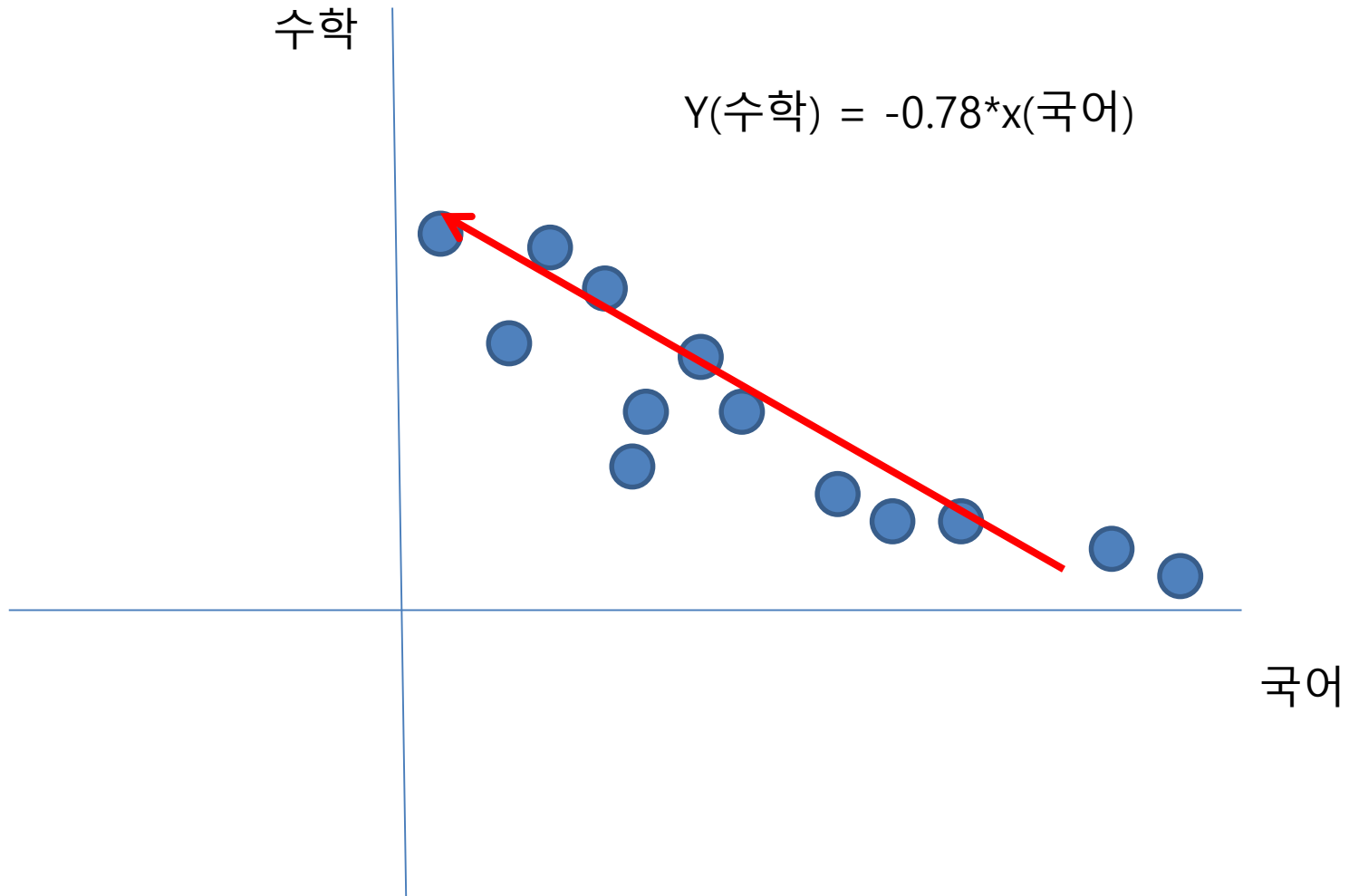
주성분 분석 실습

• 결과화면 해석

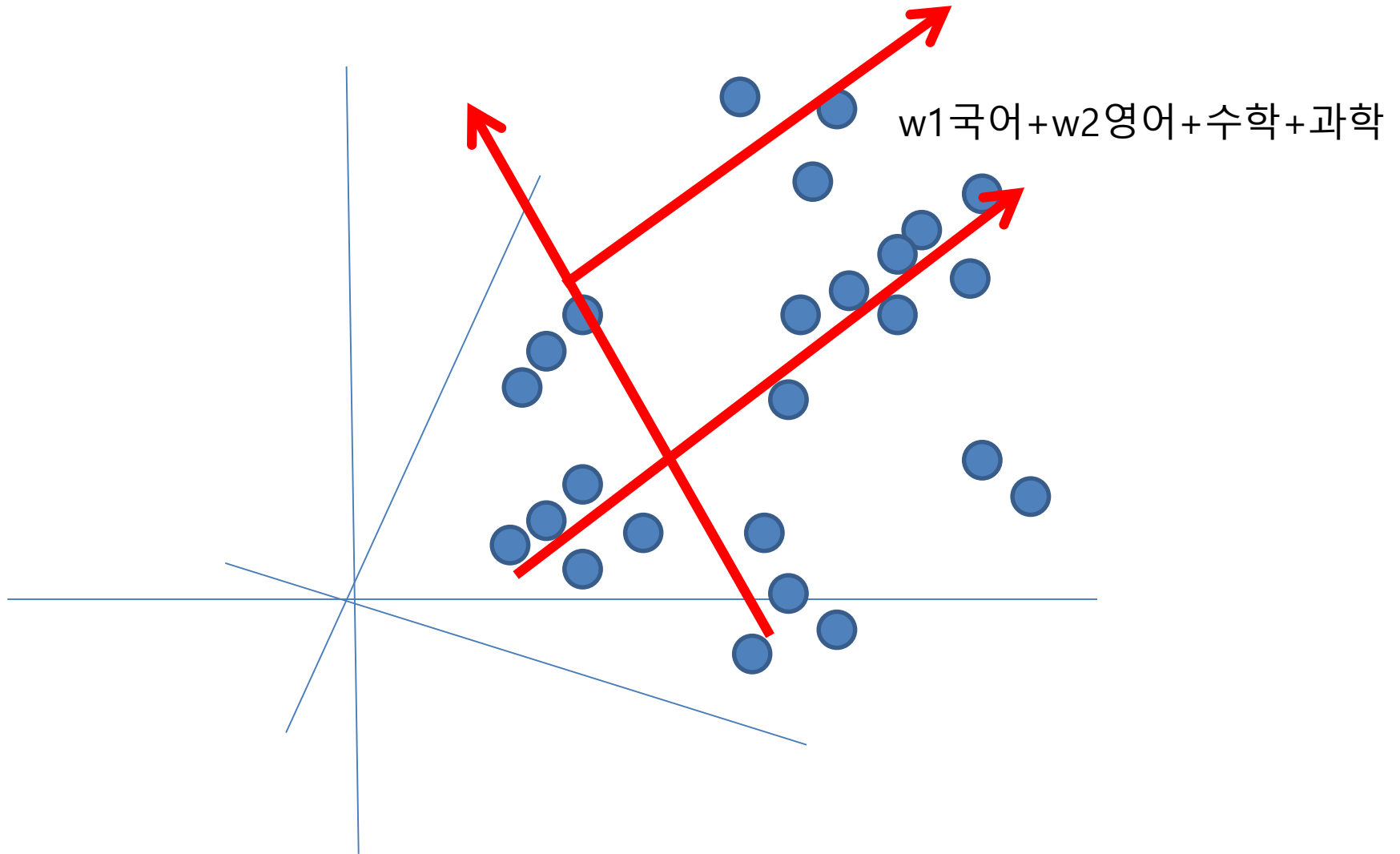


주성분 분석 실습

• 결과화면 해석

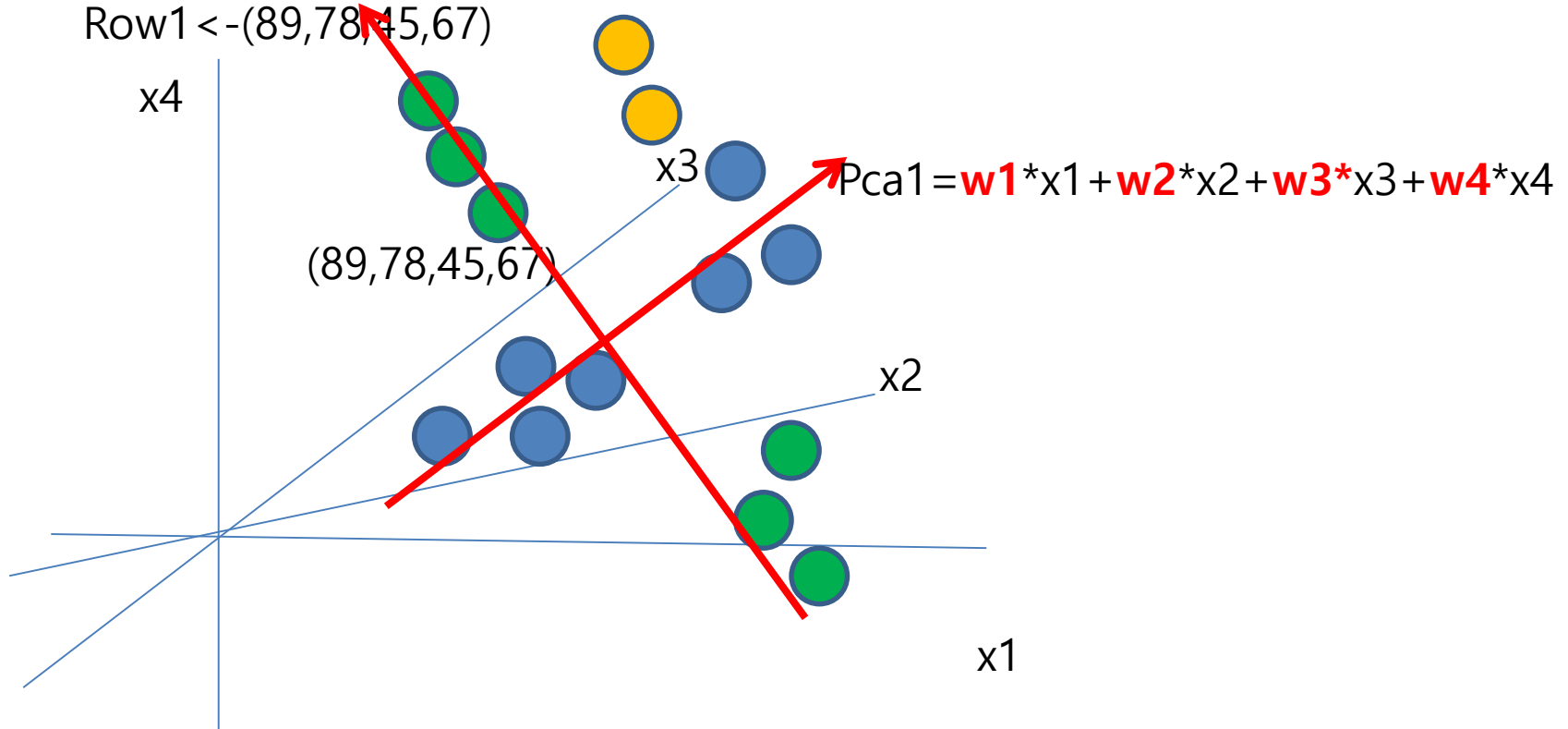


주성분 분석 실습



주성분분석(PCA)

x_1 (vector), x_2 , x_3 , x_4 → 상관관계 → 공간좌표계에 시각화(4차원)
`Row1 <- (89,78,45,67)`

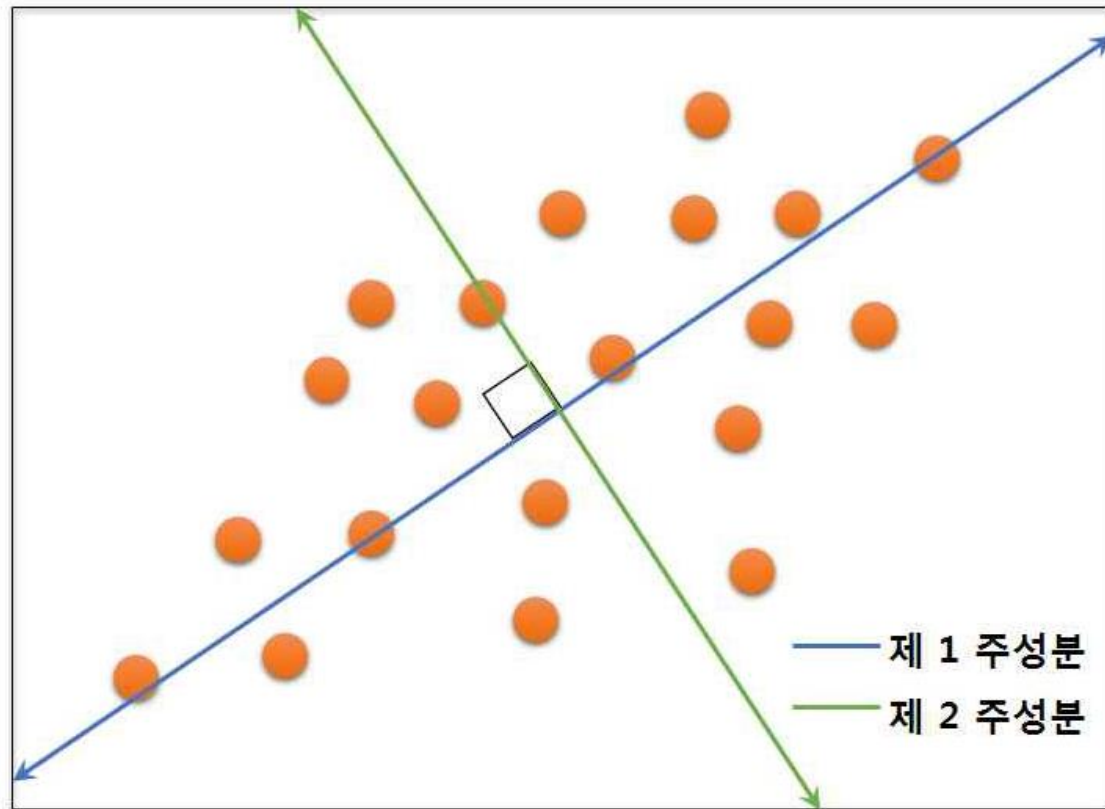


비율적 거리(상대적거리) = 분산영향을 주기때문에
→ 정규화
→ `Prcomp(x, scale=TRUE)`

주성분 분석 실습

• 주성분 분석의 목적과 개념

- 차원 축소



<주성분 분석(PCA)의 예>

주성분 분석(PCA) 요약정리

특성이 될만한 요인을 찾아 선형식으로 표현해 내는 것
= 여러 요인 중 주성분 추출

예) 이미지를 특성 짓게 만들어 주는 요인은 무엇일까?

	x1	x2										← 요인
이미지1												
이미지2												
이미지3												
.....												
이미지n												

배경에 해당하는 요인일 경우 거의 255로 동일한 색상정보를 가짐

→ 결국 이러한의미없는 정보는 평균값에 가까운 값들이 대부분 차지하는 벡터가 되며

→ 특성있는 요인 벡터는 분산값이 크고, 다양한 값을 가지는 특성이 보임

→ 이러한 특성을 가진 요인 벡터를 찾아 새로운 선형벡터로 만들어내는 것이 PCA

주성분 분석(PCA)

- 주성분 분석(principal component analysis, PCA)은 변수 간의 상관관계가 있는 다차원의 데이터를 효율적으로 저차원의 데이터로 요약하는 방법 중 하나이다.

- 데이터에 존재하는 패턴을 인식하는 데에서 많은 양의 정보는 세밀한 인식을 가능하게 하지만, 계산 속도를 떨어뜨리고 주요하지 않은 사소한 패턴에 과한 집중을 하게 해 오히려 잘못된 패턴을 인식하게 하는 문제가 있다. 중요한 변수를 구분하고 계산을 단순화하며 데이터 시각화를 위해 차원을 축소하는 데이터 처리 기법이 중요하다. 생체 시스템은 구성요소 간의 제어 연관성이 있고 이들은 강한 상관관계를 보인다. 따라서 어떤 하나의 구성요소의 값을 알면 이와 강한 상관관계를 가지는 다른 구성요소들의 값은 추정 가능하므로, 다차원 데이터를 탐구하는 데에서 하나의 구성요소를 탐구하면 된다는 것이 주성분 분석에서 차원 축소(dimension reduction)의 개념이다.

주성분 분석(PCA)

- 데이터에 존재하는 상관관계를 효율적으로 동정하려면 구성요소의 상관 행렬에 고유값 분해(eigenvalue decomposition)를 적용해서 고유벡터를 구한다. 고유벡터는 구성요소들의 선형조합(linear combination)으로 정의된 상관관계가 가장 큰 축들을 의미하며 주성분(principal component) 로딩/loading)이라고 한다(그림 1). 데이터에 있는 상관관계를 유의미하게 설명하는 고유벡터의 숫자가 원래 측정된 유전자 숫자보다 적어지게 되어 차원 감소가 일어난다. 데이터 행렬에 고유벡터를 곱하면 주성분 축들로 변환된 데이터가 되며 이를 주성분 분석 스코어(PCA score)라고 한다(그림 1). 이 스코어가 주성분 축으로 표현되는 주성분 공간에서의 데이터 값이 된다. 주성분 스코어와 로딩을 외적하면 원래 데이터 행렬을 복원할 수 있고, 로딩의 숫자가 원래 변수인 유전자 숫자보다 작기 때문에 복원된 데이터 행렬은 상관관계가 약한 노이즈는 없어진다.

주성분 분석(PCA)

그림 1. 주성분 분석의 예시. x 는 측정된 변수(유전자) 개수이고, 파란색 점은 n 차원에서의 샘플을 의미한다. w_1 은 첫 주성분 분석 로딩이고, 연두색 선은 본래 차원 샘플이 주성분 공간으로 투사가 되는 변환을 나타낸다. 이와 같이 주성분 분석을 통해 차원감소를 할 수 있다.

