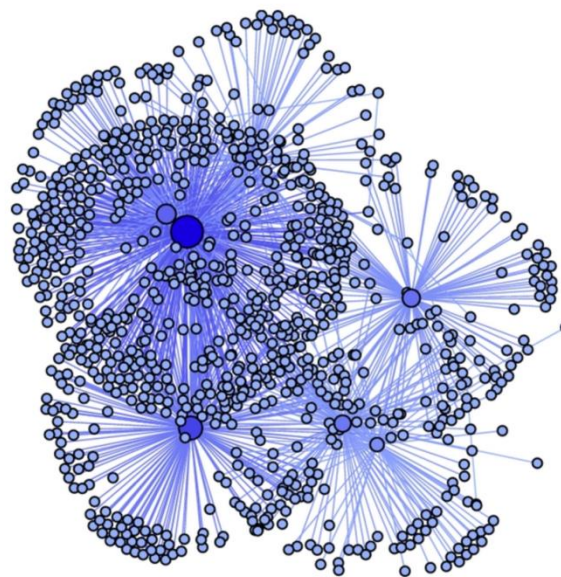
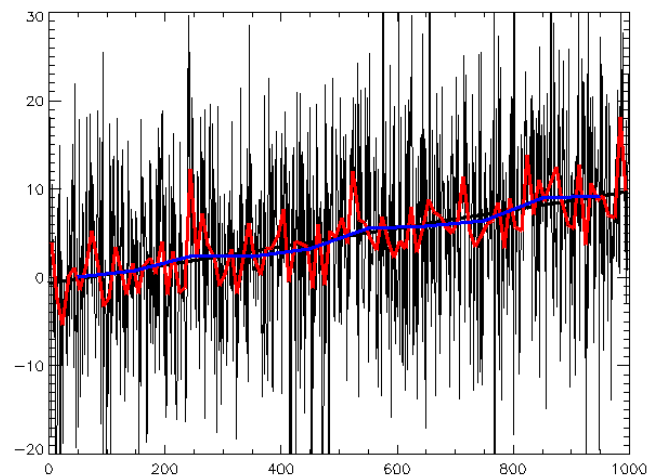
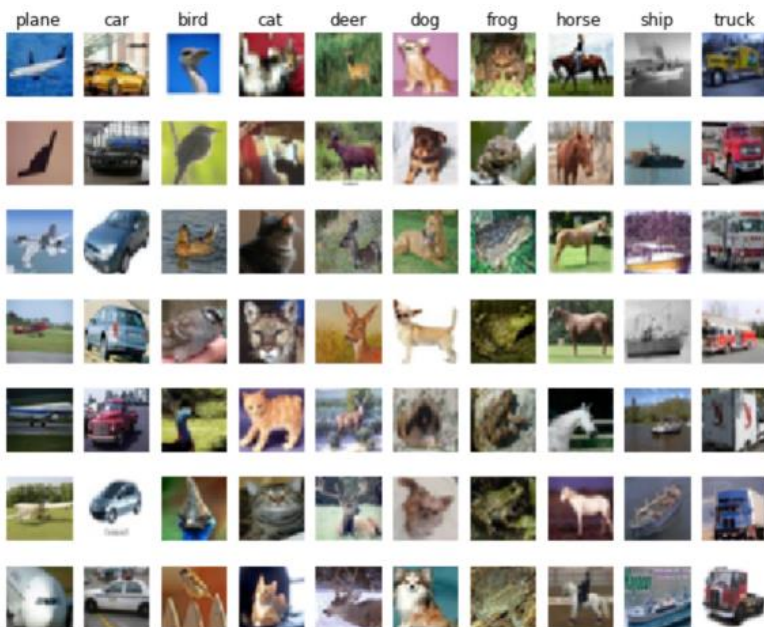


주성분 분석 (Principal Component Analysis)

PCA 개요

- 고차원 데이터를 효과적으로 분석하기 위한 대표적 분석 기법
- 차원축소, 시각화, 군집화, 압축




PCA 개요

- PCA는 n 개의 관측치와 p 개의 변수로 구성된 데이터를 상관관계가 없는 k 개의 변수로 구성된 데이터 (n 개의 관측치)로 요약하는 방식으로, 이 때 요약된 변수는 기존 변수의 선형조합으로 생성됨
- 원래 데이터의 분산을 최대한 보존하는 새로운 축을 찾고, 그 축에 데이터를 사영 (projection) 시키는 기법
- 주요 목적
 - 데이터 차원 축소 ($n \text{ by } p \rightarrow n \text{ by } k, \text{ where } k \ll p$)
 - 데이터 시각화 및 해석
- 일반적으로 PCA는 전체 분석 과정 중 초기에 사용

PCA 개요

| | X_1 | X_2 | ... | X_{p-1} | X_p |
|-------|-------|-------|-----|-----------|-------|
| 1 | | | | | |
| 2 | | | | | |
| ... | | | | | |
| $n-1$ | | | | | |
| n | | | | | |



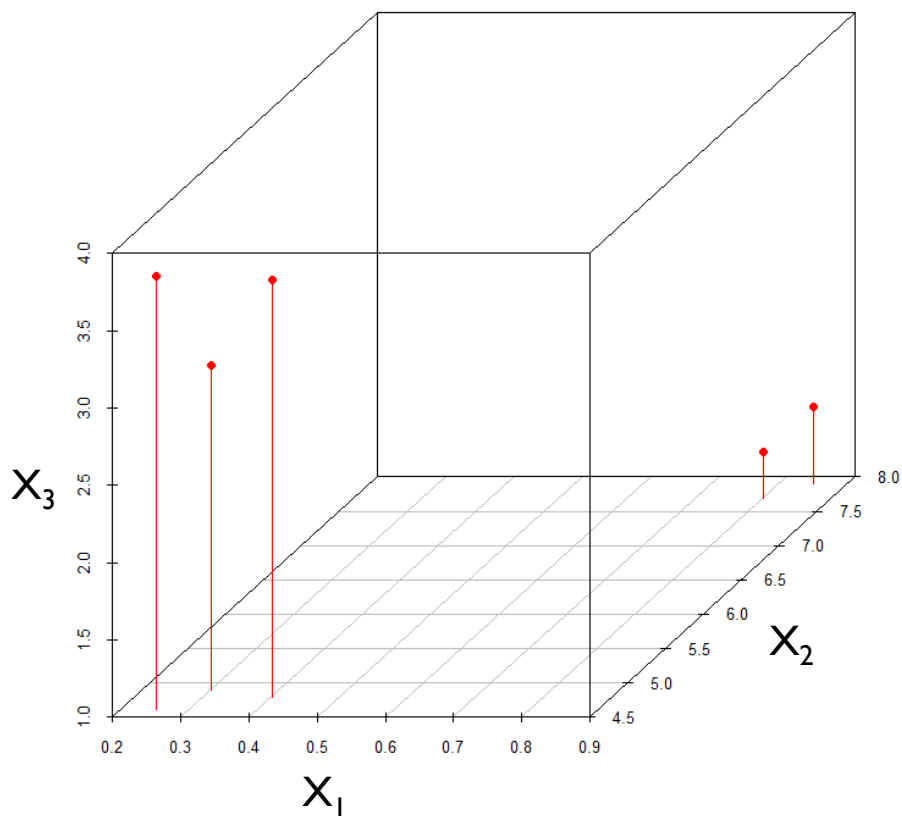
| | Z_1 | Z_2 |
|-------|-------|-------|
| 1 | | |
| 2 | | |
| ... | | |
| $n-1$ | | |
| n | | |

| | Z_1 | Z_2 | Z_3 |
|-------|-------|-------|-------|
| 1 | | | |
| 2 | | | |
| ... | | | |
| $n-1$ | | | |
| n | | | |

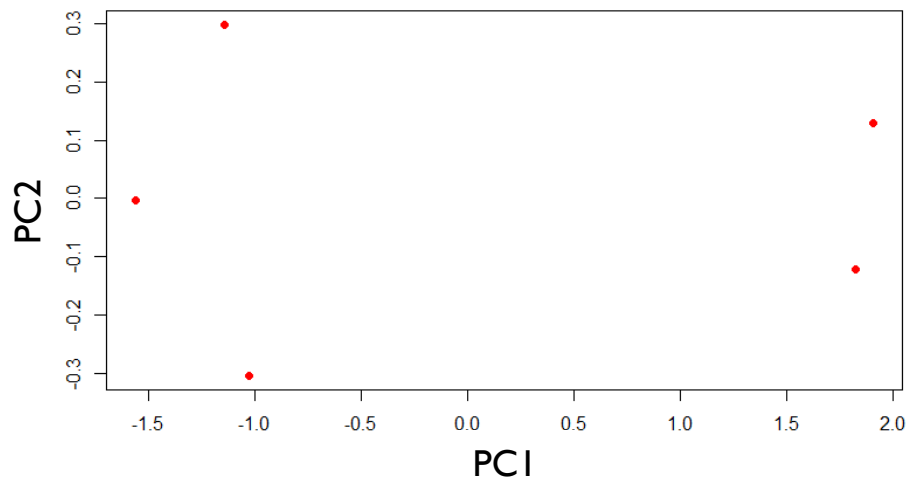
Z_1, Z_2 , 그리고 Z_3 는 기존 변수인 X_1, X_2, \dots, X_p 의 선형조합으로 새롭게 생성된 변수

PCA 개요

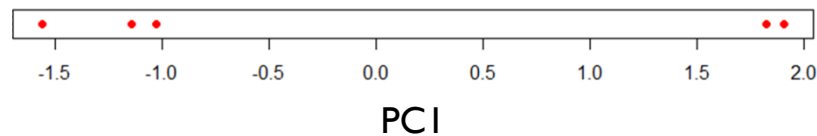
Reduce data from 3D to 2D or 1D



3D data



2D reduction



1D reduction

PCA 개요

Z is a linear combination (선형결합) of the original p variables in X

$$Z_1 = \alpha_1^T X = \alpha_{11}X_1 + \alpha_{12}X_2 + \cdots + \alpha_{1p}X_p$$

$$Z_2 = \alpha_2^T X = \alpha_{21}X_1 + \alpha_{22}X_2 + \cdots + \alpha_{2p}X_p$$

$$\vdots \qquad \qquad \qquad \vdots$$

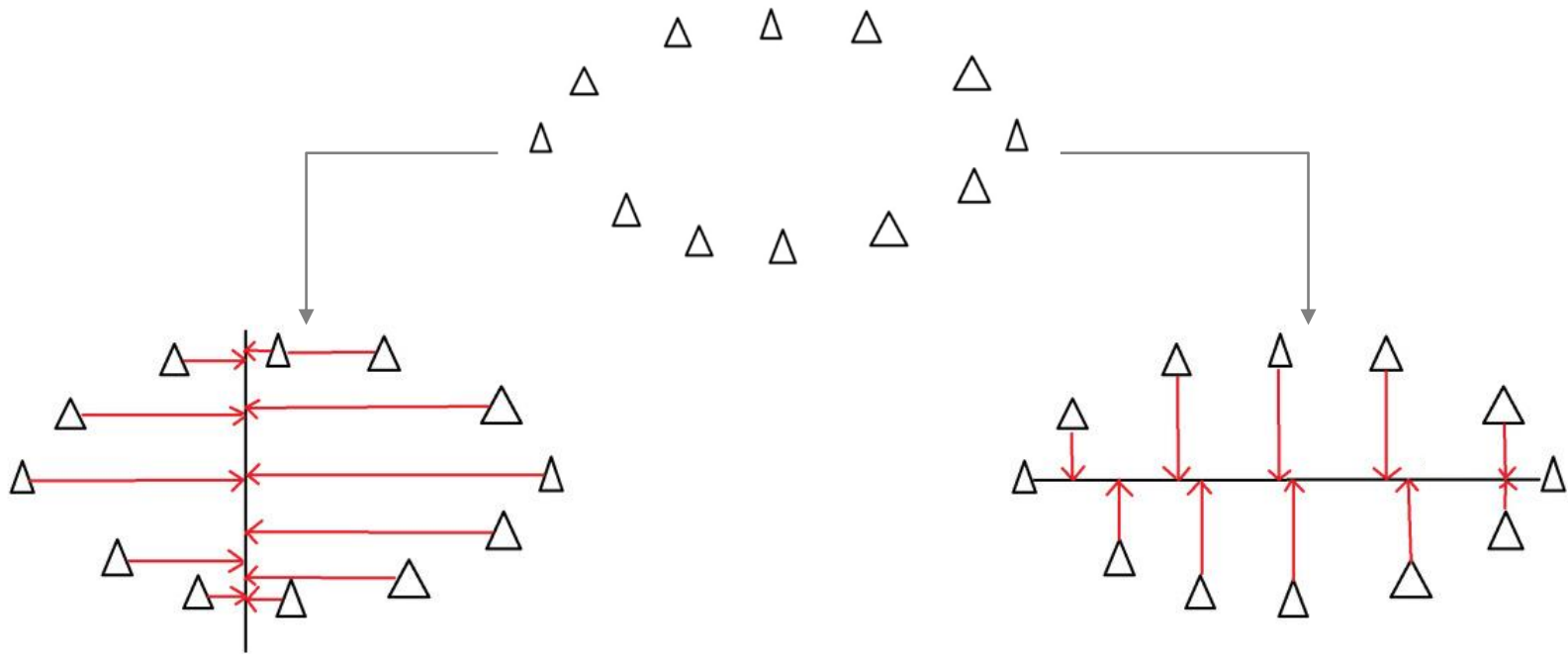
$$Z_p = \alpha_p^T X = \alpha_{p1}X_1 + \alpha_{p2}X_2 + \cdots + \alpha_{pp}X_p$$

- X_1, X_2, \dots, X_p : 원래 변수 (original variable)
- $\alpha_i = [\alpha_{i1}, \alpha_{i2}, \dots, \alpha_{ip}]$: i 번째 기저(basis) 또는 계수 (Loading)
- Z_1, Z_2, \dots, Z_p : 각 기저로 사영된 변환 후 변수 (주성분, Score)

PCA 개요

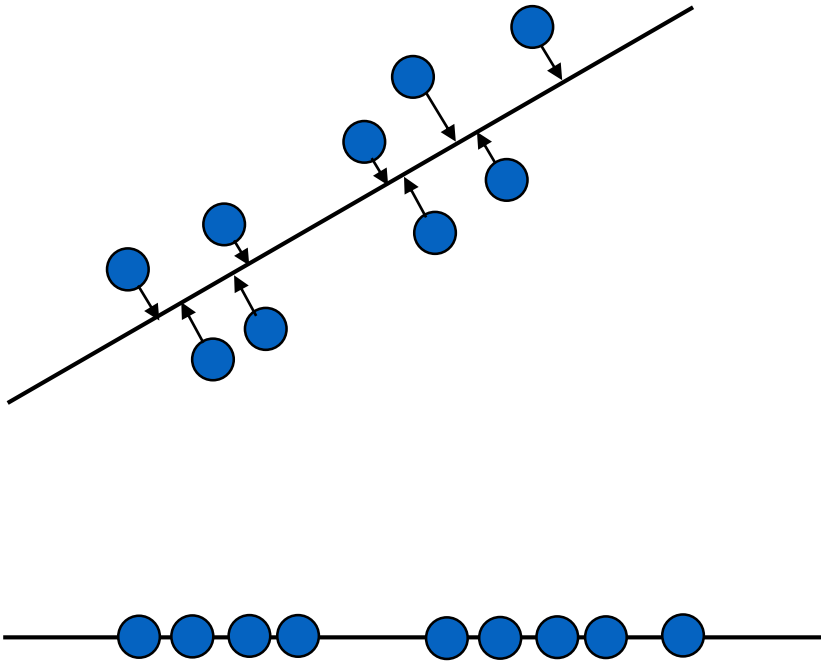
❖ 주성분 분석

- 아래 2차원 데이터를 좌측과 우측 두 개의 축에 사영시킬 경우 우측 기저 (basis)가 좌측 기저에 비해 손실되는 정보의 양(분산의 크기)이 적으므로 상대적으로 선호되는 기저라고 할 수 있음

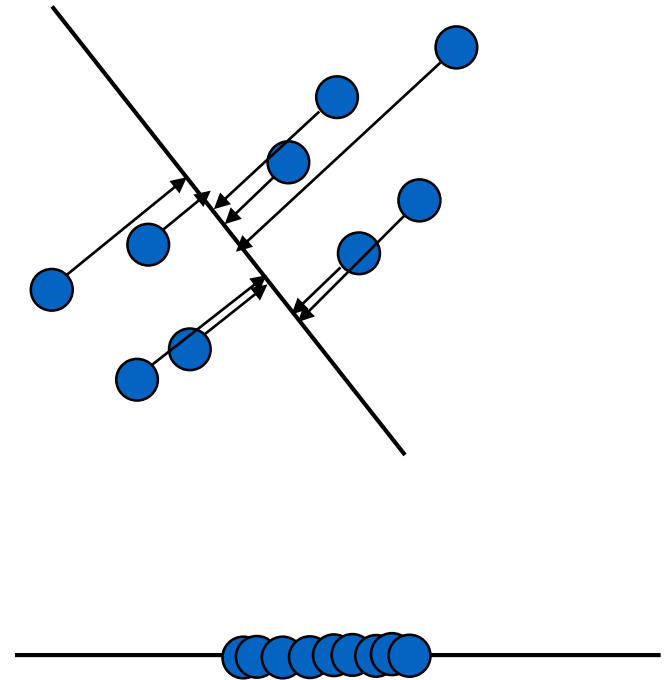


PCA 개요

Find the new axis that
maximizes the variance of data



Find the new axis that
minimizes the variance of data



PCA 수리적 배경

| | X_1 | X_2 | ... | X_{p-1} | X_p |
|-------|-------|-------|-----|-----------|-------|
| 1 | | | | | |
| 2 | | | | | |
| ... | | | | | |
| $n-1$ | | | | | |
| n | | | | | |

| | 1 | 2 | ... | $n-1$ | n |
|-----------|---|---|-----|-------|-----|
| X_1 | | | | | |
| X_2 | | | | | |
| ... | | | | | |
| X_{p-1} | | | | | |
| X_p | | | | | |

• 공분산(Covariance)의 성질

- \mathbf{X} 를 p 개의 변수와 n 개의 개체로 구성된 n by p (or p by n)행렬로 정의할 때 \mathbf{X} 의 공분산 행렬은 다음과 같음

$$Cov(X) = \frac{1}{n} (X - \bar{X})(X - \bar{X})^T$$

- 공분산 행렬의 대각 성분은 각 변수의 분산과 같으며, 비대각행렬은 대응하는 두 변수의 공분산과 같음 (변수 개수: p)

$$C_x = Var[x] = \begin{bmatrix} Var[x_1] & Cov[x_1, x_2] & \dots & Cov[x_1, x_p] \\ Cov[x_2, x_1] & Var[x_2] & \dots & Cov[x_2, x_p] \\ \vdots & \vdots & \ddots & \vdots \\ Cov[x_p, x_1] & Cov[x_p, x_2] & \dots & Var[x_p] \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_{pp} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_p^2 \end{bmatrix}$$

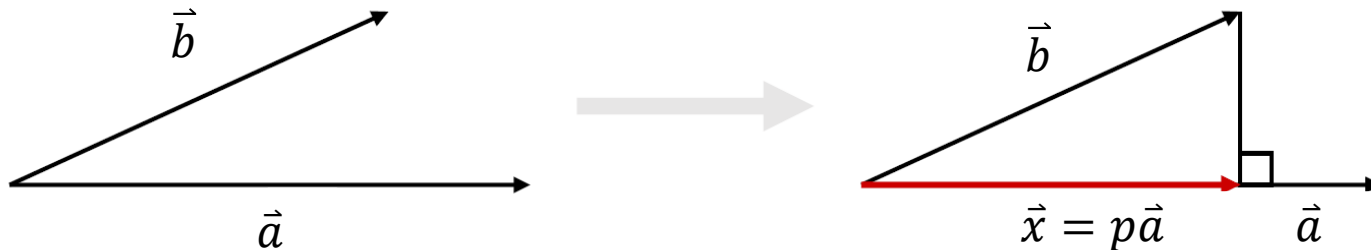
- 데이터의 총분산은 공분산행렬의 대각성분들의 합으로 표현됨

$$tr[Cov(\mathbf{X})] = Cov(\mathbf{X})_{11} + Cov(\mathbf{X})_{22} + Cov(\mathbf{X})_{33} + \dots + Cov(\mathbf{X})_{pp}$$

PCA 수리적 배경

- 사영(Projection)

- 한 벡터 \vec{b} 를 다른 벡터 \vec{a} 에 사영시킨다는 것은 벡터 \vec{b} 로부터 벡터 \vec{a} 에 수직인 점까지의 길이를 가지며 벡터 \vec{a} 와 같은 방향을 갖는 벡터를 찾는다는 것을 의미



$$(\vec{b} - p\vec{a})^T \vec{a} = 0 \Rightarrow \vec{b}^T \vec{a} - p\vec{a}^T \vec{a} = 0 \Rightarrow p = \frac{\vec{b}^T \vec{a}}{\vec{a}^T \vec{a}}$$

$$\vec{x} = p\vec{a} = \frac{\vec{b}^T \vec{a}}{\vec{a}^T \vec{a}} \vec{a}$$

If \vec{a} is unit vector

$$p = \vec{b}^T \vec{a} \Rightarrow \vec{x} = p\vec{a} = (\vec{b}^T \vec{a})\vec{a}$$

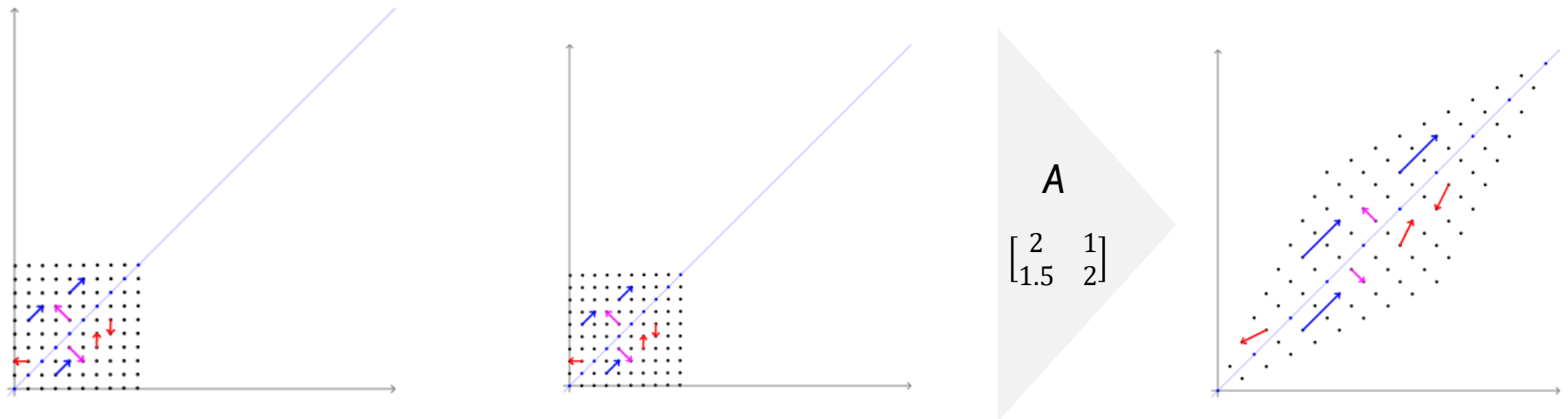
PCA 수리적 배경

- 고유값 및 고유벡터

- 어떤 행렬 \mathbf{A} 에 대해 상수 λ 와 벡터 \mathbf{x} 가 다음 식을 만족할 때, λ 와 \mathbf{x} 를 각각 행렬 \mathbf{A} 의 고유값(eigenvalue) 및 고유벡터(eigenvector)라고 함

$$\mathbf{Ax} = \lambda\mathbf{x} \quad \rightarrow \quad (\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = 0$$

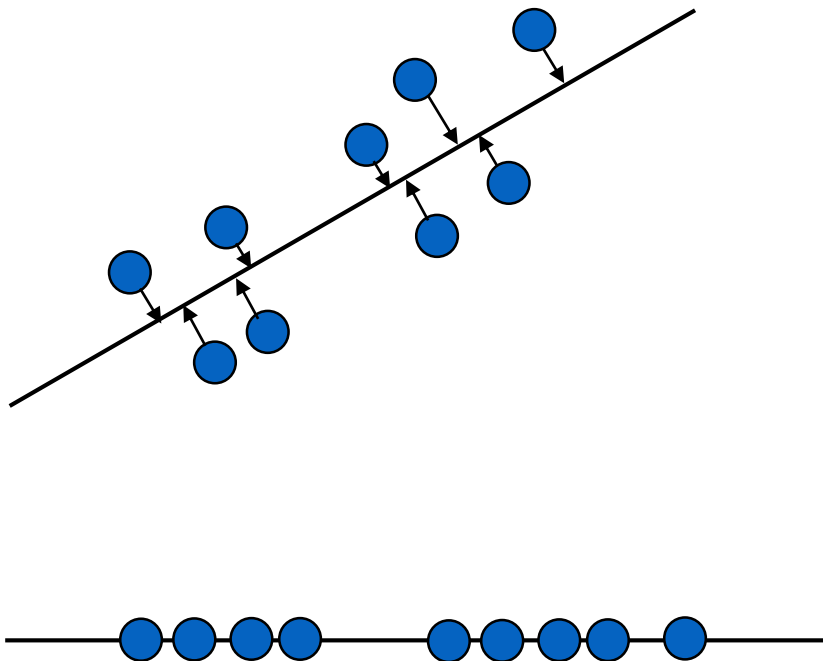
- 벡터에 행렬을 곱한다는 것은 해당 벡터를 선형변환(linear transformation)한다는 의미 \rightarrow 고유벡터는 이 변환에 의해 방향이 변하지 않는 벡터를 의미



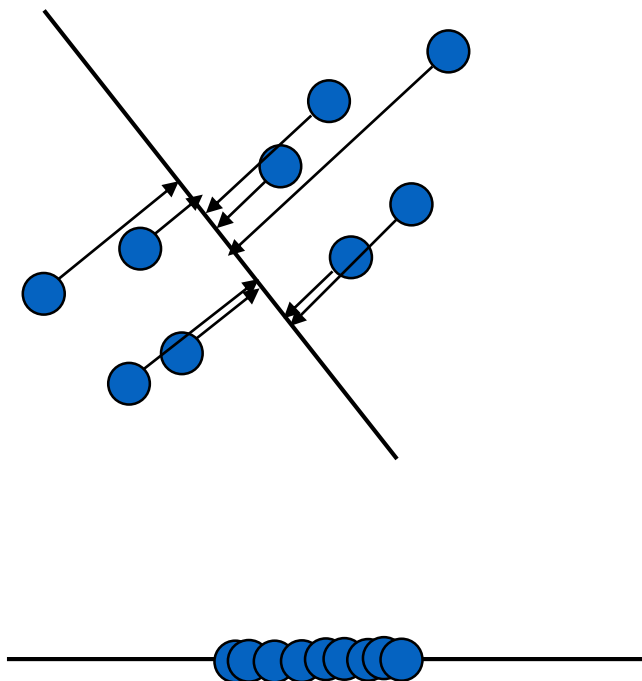
Recall PCA 개요

Principal Component Analysis

Find the new axis that
maximizes the variance of data



Find the new axis that
minimizes the variance of data



PCA 알고리즘 - 주성분 추출

- Assume that we have the centered data (i.e., $\bar{X}_i = 0$, $i = 1, \dots, p$)
- Let \mathbf{X} be an p -dimensional random vector with the covariance matrix Σ
- Let α be an p -dimensional vector of length one (i.e., $\alpha^T \alpha = 1$)
- Let $\mathbf{Z} = \alpha^T \mathbf{X}$ be the projection of \mathbf{X} onto the direction α

The main purpose in PCA is

to find α that produces the largest variance of \mathbf{Z}

$$\text{Max Var}(\mathbf{Z}) = \text{Var}(\alpha^T \mathbf{X}) = \alpha^T \underline{\text{Var}(\mathbf{X})} \alpha = \alpha^T \Sigma \alpha$$

$$\text{s.t. } \|\alpha\| = \alpha^T \alpha = 1$$

PCA 알고리즘 - 주성분 추출

$$\text{Max } \alpha^T \Sigma \alpha = \alpha^T E \Lambda E^T \alpha$$

$$\text{s.t. } \|\alpha\| = 1$$

$$\text{Max } \beta^T \Lambda \beta \text{ where } \beta = E^T \alpha$$

$$\text{s.t. } \|\beta\| = 1$$

$$\text{Max } \lambda_1 \beta_1^2 + \lambda_2 \beta_2^2 + \dots + \lambda_m \beta_m^2$$

$$\text{s.t. } \beta_1^2 + \beta_2^2 + \dots + \beta_m^2 = 1$$

$$\lambda_1 > \lambda_2 > \dots > \lambda_m$$

The Spectral Decomposition

Let A be a $K \times K$ symmetric matrix. Then A can be expressed in terms of its k eigenvalue-eigenvector pairs (λ_i, e_i) , $i = 1, 2, \dots, k$ as

$$A = \sum_{i=1}^k \lambda_i e_i e_i^T.$$

$$\text{ex) } \mathbf{A} = \begin{pmatrix} 2.2 & 0.4 \\ 0.4 & 2.8 \end{pmatrix}$$

$$|A - \lambda I| = \lambda^2 - 5\lambda + 6 = (\lambda - 3)(\lambda - 2).$$

Thus, eigenvalues are $\lambda_1 = 3$ and $\lambda_2 = 2$, and corresponding eigenvec-

$$\text{tors are } \mathbf{e}_1 = \begin{pmatrix} \frac{1}{\sqrt{5}} \\ \frac{2}{\sqrt{5}} \end{pmatrix}, \quad \mathbf{e}_2 = \begin{pmatrix} \frac{2}{\sqrt{5}} \\ -\frac{1}{\sqrt{5}} \end{pmatrix}.$$

$$\text{Finally, } \mathbf{A} = \begin{pmatrix} 2.2 & 0.4 \\ 0.4 & 2.8 \end{pmatrix} = 3 \begin{pmatrix} \frac{1}{\sqrt{5}} \\ \frac{2}{\sqrt{5}} \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} \end{pmatrix} + 2 \begin{pmatrix} \frac{2}{\sqrt{5}} \\ -\frac{1}{\sqrt{5}} \end{pmatrix} \begin{pmatrix} \frac{2}{\sqrt{5}} & -\frac{1}{\sqrt{5}} \end{pmatrix}.$$

$$\beta = E^T \alpha$$

Thus, the optimal value is λ_1 and $\alpha = e_1$

PCA 알고리즘 - 주성분 추출

$$\text{Max Var}(\mathbf{Z}) = \text{Var}(\alpha^T \mathbf{X}) = \alpha^T \text{Var}(\mathbf{X}) \alpha = \alpha^T \Sigma \alpha$$

$$\text{s.t. } \|\alpha\| = \alpha^T \alpha = 1$$

$$\alpha = \mathbf{e}_l$$

$$Y = \alpha^T X = \mathbf{e}_l^T X$$

p x n (p차원)

| | 1 | 2 | ... | n-1 | n |
|------------------|---|---|-----|-----|---|
| X ₁ | | | | | |
| X ₂ | | | | | |
| ... | | | | | |
| X _{p-1} | | | | | |
| X _p | | | | | |

original

$$\begin{matrix} \mathbf{X} \\ p \times n \\ p \text{ 차원} \end{matrix}$$

→

projection

$$\begin{matrix} Y = \alpha^T X \\ 1 \times p \quad p \times n \\ 1 \text{ 차원} \end{matrix}$$

→

reconstruction

$$\begin{matrix} X' = \alpha \alpha^T X \\ p \times 1 \quad 1 \times p \quad p \times n \\ p \text{ 차원} \end{matrix}$$

PCA - 예제

X =

| X_1 | X_2 | X_3 |
|-------|-------|-------|
| 0.2 | 5.6 | 3.56 |
| 0.45 | 5.89 | 2.4 |
| 0.33 | 6.37 | 1.95 |
| 0.54 | 7.9 | 1.32 |
| 0.77 | 7.87 | 0.98 |

X =

| X_1 | X_2 | X_3 |
|---------|---------|---------|
| -1.1930 | -1.0300 | 1.5012 |
| -0.0370 | -0.7647 | 0.3540 |
| -0.5919 | -0.3257 | -0.0910 |
| 0.3792 | 1.0739 | -0.7140 |
| 1.4427 | 1.0464 | -1.0502 |

(normalize X to
 $E(X_i)=0$,
 $Var(X_i)=1$)

Covariance(**X**) =

| | | |
|---------|---------|---------|
| 0.0468 | 0.1990 | -0.1993 |
| 0.1990 | 1.1951 | -1.0096 |
| -0.1993 | -1.0096 | 1.0225 |

Correlation(**X**) =

| | | |
|---------|---------|---------|
| 1 | 0.8417 | -0.8840 |
| 0.8417 | 1 | -0.9133 |
| -0.8840 | -0.9133 | 1 |

Question) $\text{Corr}(X_1, X_2) =$

$\text{Corr}(X_3, X_3) =$

PCA - 예제

The eigenvalue-eigenvector pairs on the correlation matrix, Σ

$$\lambda_1 = 0.0786, \quad e_1^T = [0.2590 \quad 0.5502 \quad 0.7938]$$

$$\lambda_2 = 0.1618, \quad e_2^T = [0.7798 \quad -0.6041 \quad 0.1643]$$

$$\lambda_3 = 2.7596, \quad e_3^T = [0.5699 \quad 0.5765 \quad -0.5855]$$

PCA - 예제

$$\begin{aligned}\lambda_1 &= 0.0786, e_1^T = [0.2590 \quad 0.5502 \quad 0.7938] \\ \lambda_2 &= 0.1618, e_2^T = [0.7798 \quad -0.6041 \quad 0.1643] \\ \lambda_3 &= 2.7596, e_3^T = [0.5699 \quad 0.5765 \quad -0.5855]\end{aligned}$$

, $\mathbf{X} =$

| X_1 | X_2 | X_3 |
|---------|---------|---------|
| -1.1930 | -1.0300 | 1.5012 |
| -0.0370 | -0.7647 | 0.3540 |
| -0.5919 | -0.3257 | -0.0910 |
| 0.3792 | 1.0739 | -0.7140 |
| 1.4427 | 1.0464 | -1.0502 |

(normalize \mathbf{X} to
 $E(X_i)=0$,
 $\text{Var}(X_i)=1$)

$$\lambda_3 > \lambda_2 > \lambda_1$$

$$Z_1 = e_3^T X = 0.5699 \cdot X_1 + 0.5765 \cdot X_2 - 0.5855 \cdot X_3 = 0.5699 \cdot \begin{bmatrix} -1.1930 \\ -0.0370 \\ -0.5919 \\ 0.3792 \\ 1.4427 \end{bmatrix} + 0.5765 \cdot \begin{bmatrix} -1.0300 \\ -0.7647 \\ -0.3257 \\ 1.0739 \\ 1.0464 \end{bmatrix} - 0.5855 \cdot \begin{bmatrix} 1.5012 \\ 0.3540 \\ -0.0910 \\ -0.7140 \\ -1.0502 \end{bmatrix} = \begin{bmatrix} -2.1527 \\ -0.6692 \\ -0.4718 \\ 1.2533 \\ 2.0404 \end{bmatrix}$$

$$Z_2 = e_2^T X = \begin{bmatrix} -0.0615 \\ 0.4912 \\ -0.2798 \\ -0.4703 \\ 0.3204 \end{bmatrix}$$

$$Z_3 = e_1^T X = \begin{bmatrix} 0.3160 \\ -0.1493 \\ -0.4047 \\ 0.1223 \\ 0.1157 \end{bmatrix}$$

$$\therefore Z = \begin{bmatrix} -2.1527 & -0.0615 & 0.3160 \\ -0.6692 & 0.4912 & -0.1493 \\ -0.4718 & -0.2798 & -0.4047 \\ 1.2533 & -0.4703 & 0.1223 \\ 2.0404 & 0.3204 & 0.1157 \end{bmatrix}$$

PCA - 예제

$$Z = \begin{bmatrix} -2.1527 & -0.0615 & 0.3160 \\ -0.6692 & 0.4912 & -0.1493 \\ -0.4718 & -0.2798 & -0.4047 \\ 1.2533 & -0.4703 & 0.1223 \\ 2.0404 & 0.3204 & 0.1157 \end{bmatrix}$$

$$\text{Cov}(Z) = \begin{bmatrix} 2.7596 & 0 & 0 \\ 0 & 0.1618 & 0 \\ 0 & 0 & 0.0786 \end{bmatrix}$$

Question) $\text{Cov}(Z_1, Z_2) =$

$\text{Cov}(Z_2, Z_3) =$

$\text{Cov}(Z_3, Z_1) =$

$\text{Cov}(Z_1, Z_1) =$

주성분(Z)들은 서로 독립!

PCA - 예제

X =

| X_1 | X_2 | X_3 |
|---------|---------|---------|
| -1.1930 | -1.0300 | 1.5012 |
| -0.0370 | -0.7647 | 0.3540 |
| -0.5919 | -0.3257 | -0.0910 |
| 0.3792 | 1.0739 | -0.7140 |
| 1.4427 | 1.0464 | -1.0502 |

Z =

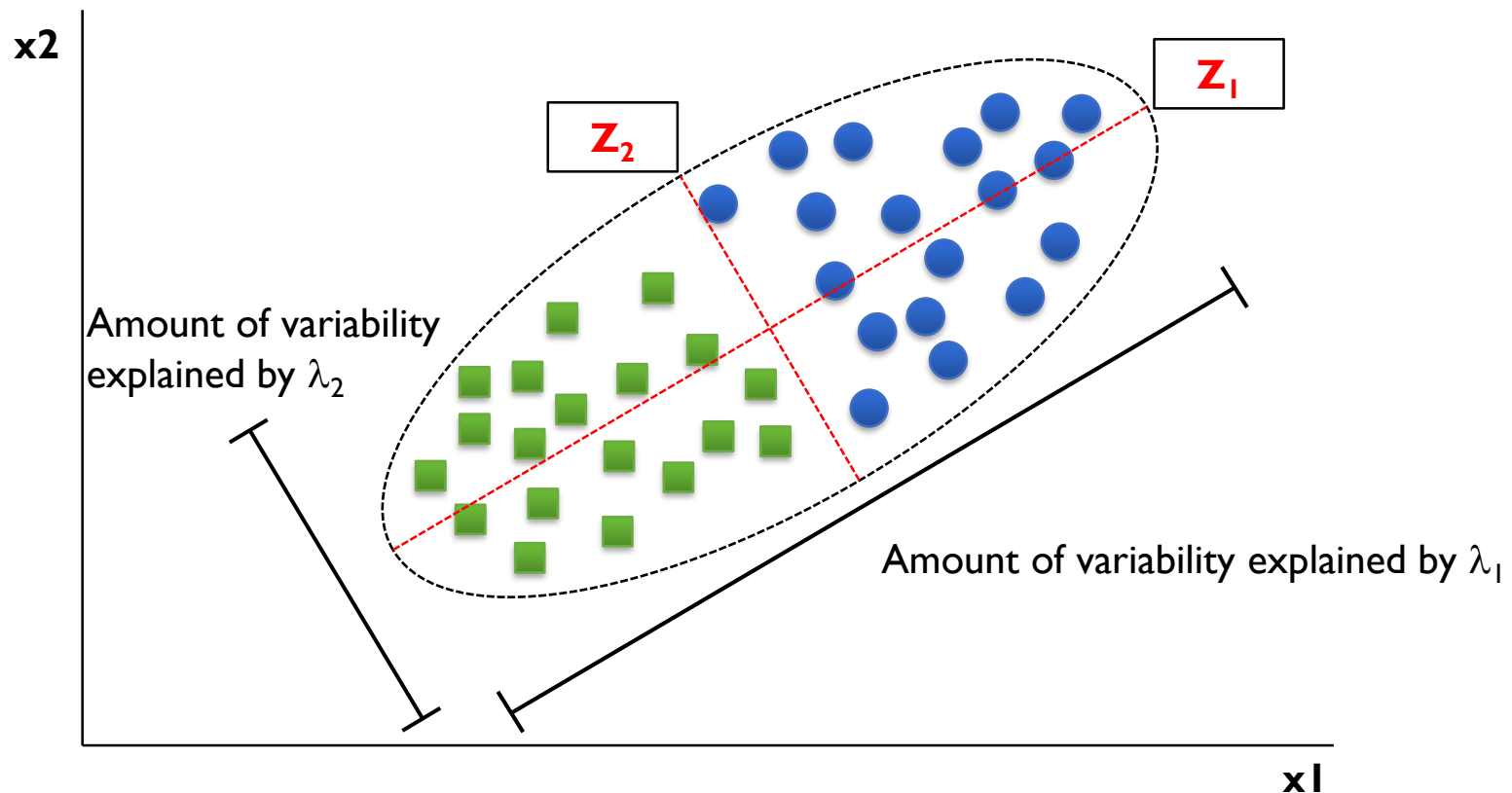
| Z_1 | Z_2 | Z_3 |
|---------|---------|---------|
| -2.1527 | -0.0615 | 0.3160 |
| -0.6692 | 0.4912 | -0.1493 |
| -0.4718 | -0.2798 | -0.4047 |
| 1.2533 | -0.4703 | 0.1223 |
| 2.0404 | 0.3204 | 0.1157 |

몇 개의 주성분을 사용해야 할까?

PCA 개요

Eigenvalues of the covariance matrix

= Variances of each principal component (각 주성분의 분산)



PCA - 예제

Eigenvalues of the covariance matrix ($\lambda_1, \lambda_2, \lambda_3$)

= Variances of each principal component (각 주성분의 분산)

Covariance matrix of principal components

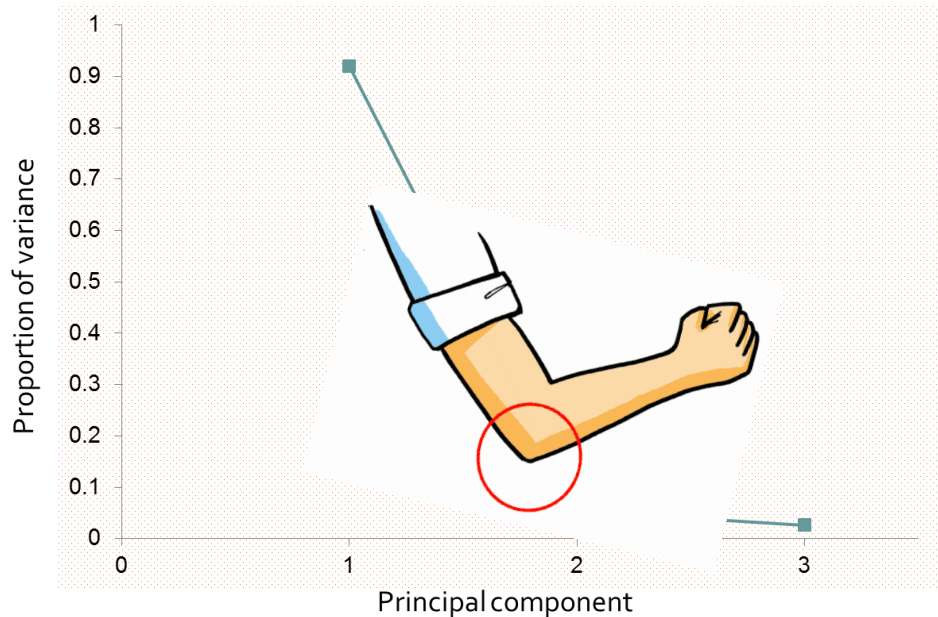
$$\text{Cov}(Z) = \begin{bmatrix} 2.7596 & 0 & 0 \\ 0 & 0.1618 & 0 \\ 0 & 0 & 0.0786 \end{bmatrix}$$

$\text{Var}(Z_1) = 2.7596 = \lambda_3$ (Largest eigenvalue)
 $\text{Var}(Z_2) = 0.1618 = \lambda_2$
 $\text{Var}(Z_3) = 0.0786 = \lambda_1$

Proportion of total population
variance due to the 1st
principal component

$$= \frac{\lambda_3}{\lambda_1 + \lambda_2 + \lambda_3} = \frac{2.7596}{0.0786 + 0.1618 + 2.7596} = 0.920$$

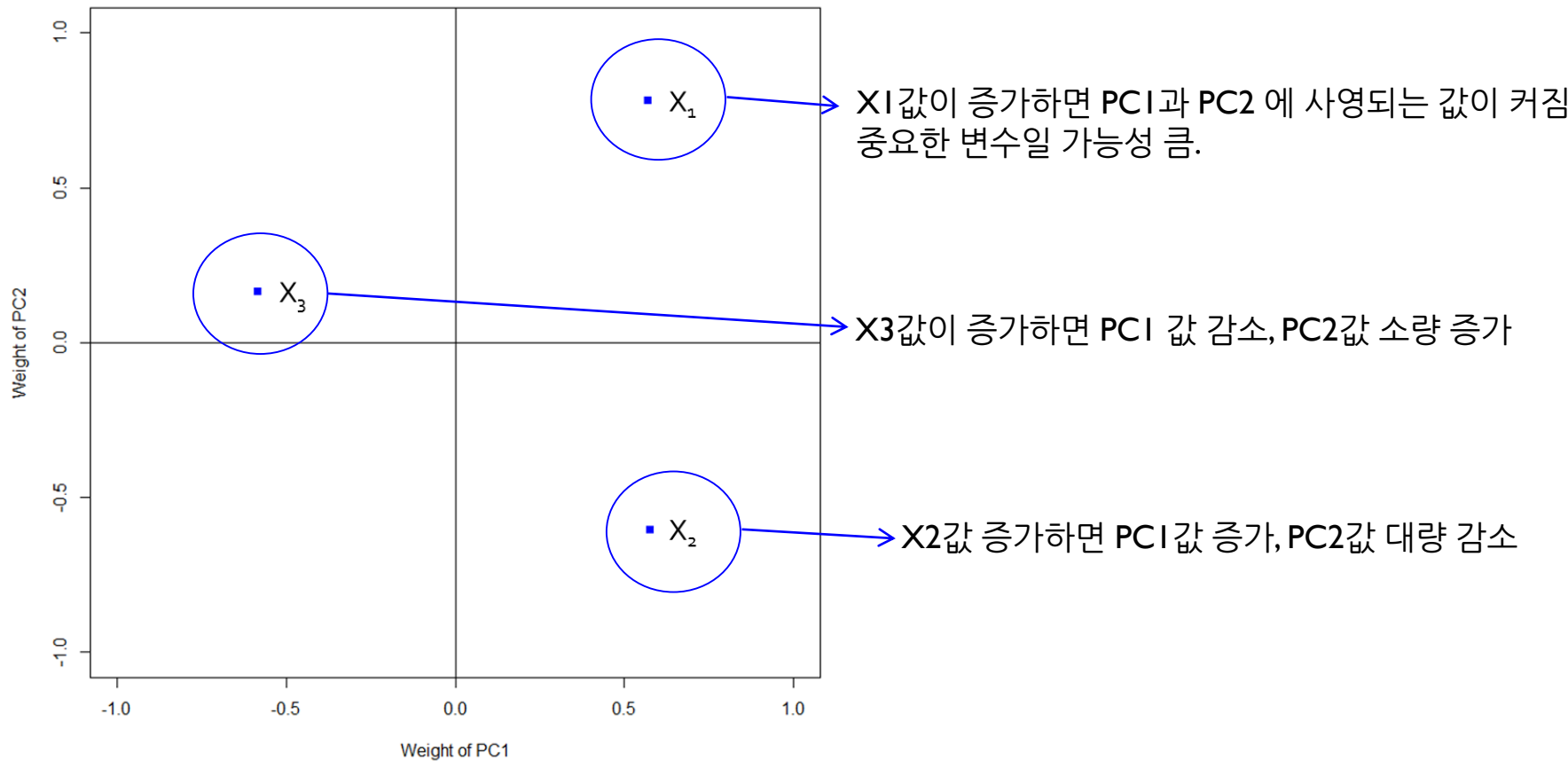
PCA – 예제 (몇 개의 주성분?)



- **선택 방식 1:** 고유값 감소율이 유의미하게 낮아지는 Elbow Point에 해당하는 주성분 수를 선택
- **선택 방식 2:** 일정 수준 이상의 분산비를 보존하는 최소의 주성분을 선택 (보통 70% 이상)

PCA Loading Plot - 예제

PCA Loading: 실제 변수가 주성분 결정에 얼마나 많은 영향을 미쳤는지



PCA 알고리즘 - 요약

Step 1. 데이터 정규화 (mean centering)

Step 2. 기존 변수의 covariance (correlation) matrix 계산

Step 3. Covariance (correlation) matrix로부터 eigenvalue 및 이에 해당하는 eigenvector를 계산

Step 4. Eigenvalue 및 해당하는 eigenvectors 를 순서대로 나열

$$\lambda(1) > \lambda(2) > \lambda(3) > \lambda(4) > \lambda(5)$$

$$e(1) > e(2) > e(3) > e(4) > e(5), e(i), i=1, \dots, 5 \text{ is a vector}$$

Step 5. 정렬된 eigenvector를 토대로 기존 변수를 변환

$$Z_1 = e(1)\mathbf{X} = e_{11} \cdot X_1 + e_{12} \cdot X_2 + \dots + e_{15} \cdot X_5$$

$$Z_2 = e(2)\mathbf{X} = e_{21} \cdot X_1 + e_{22} \cdot X_2 + \dots + e_{25} \cdot X_5$$

$$\dots = \dots$$

$$Z_5 = e(5)\mathbf{X} = e_{51} \cdot X_1 + e_{52} \cdot X_2 + \dots + e_{55} \cdot X_5$$

PCA 한계

- 주성분 분석의 특징

- 공분산 행렬의 고유벡터를 사용하므로 단일 가우시안(unimodal) 분포로 추정할 수 있는 데이터에 대해 서로 독립적인 축을 찾는데 사용할 수 있음

- 한계점 I

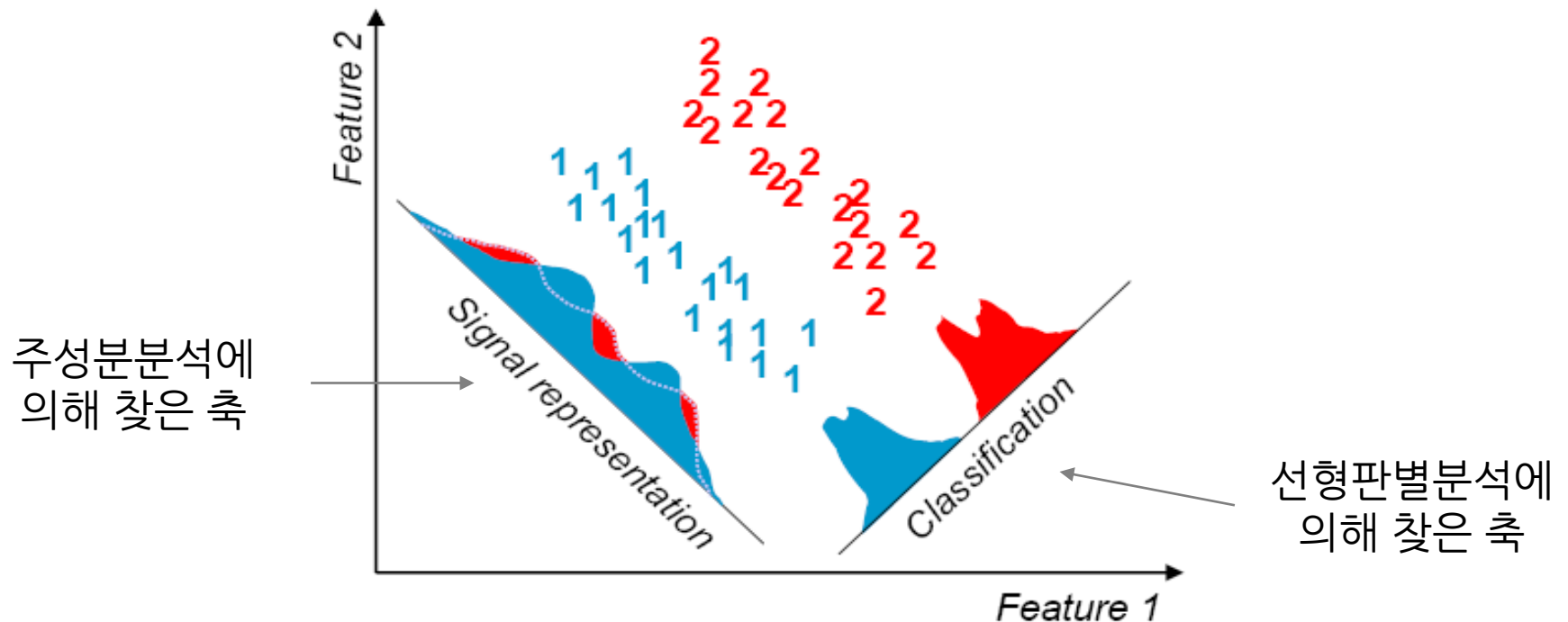
- 데이터의 분포가 가우시안이 아니거나 다중 가우시안 (multimodal) 자료들에 대해서는 적용하기가 어려움
- 대안: 커널 PCA, LLE (Locally Linear Embedding)



PCA 한계

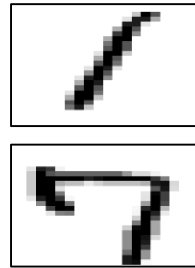
- 한계점 2

- 분류/예측 문제에 대해서 데이터의 범주 정보를 고려하지 않기 때문에 범주간 구분이 잘 되도록 변환을 해주는 것은 아님
 - 주성분분석은 단순히 변환된 축이 최대 분산방향과 정렬되도록 좌표회전을 수행함
 - 대안: Partial Least Square (PLS)

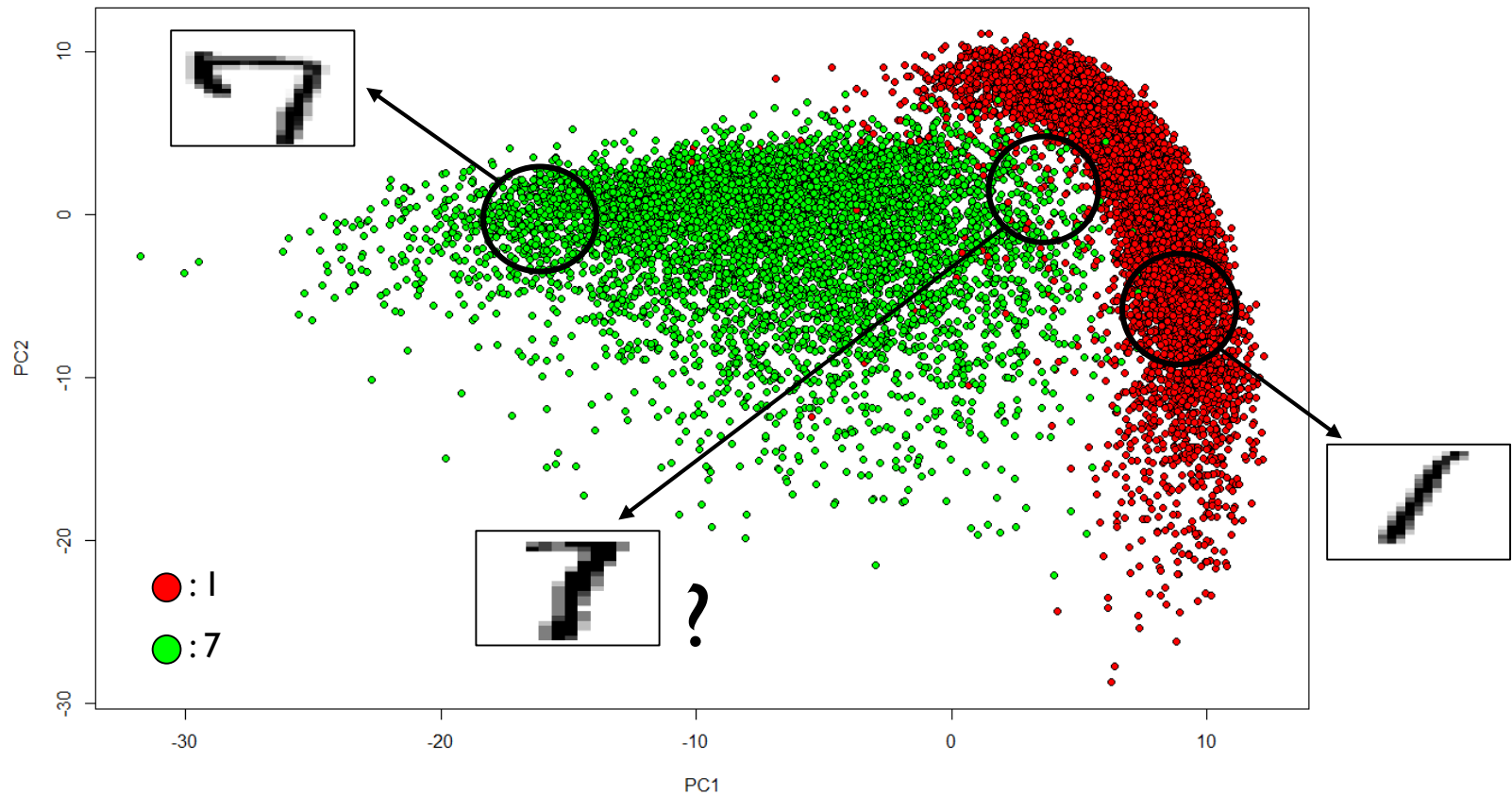


PCA – 실제 예제

약 13,000개의
28 x 28 픽셀 이미지



1? 7?



PCA – 실제 예제

- **IRIS 데이터에 대한 주성분분석**

- IRIS 데이터: 150개의 IRIS에 대해 4개 입력변수, 1개 출력변수 (3 클래스)

<https://archive.ics.uci.edu/ml/datasets/iris>



Iris Versicolor

Iris Setosa

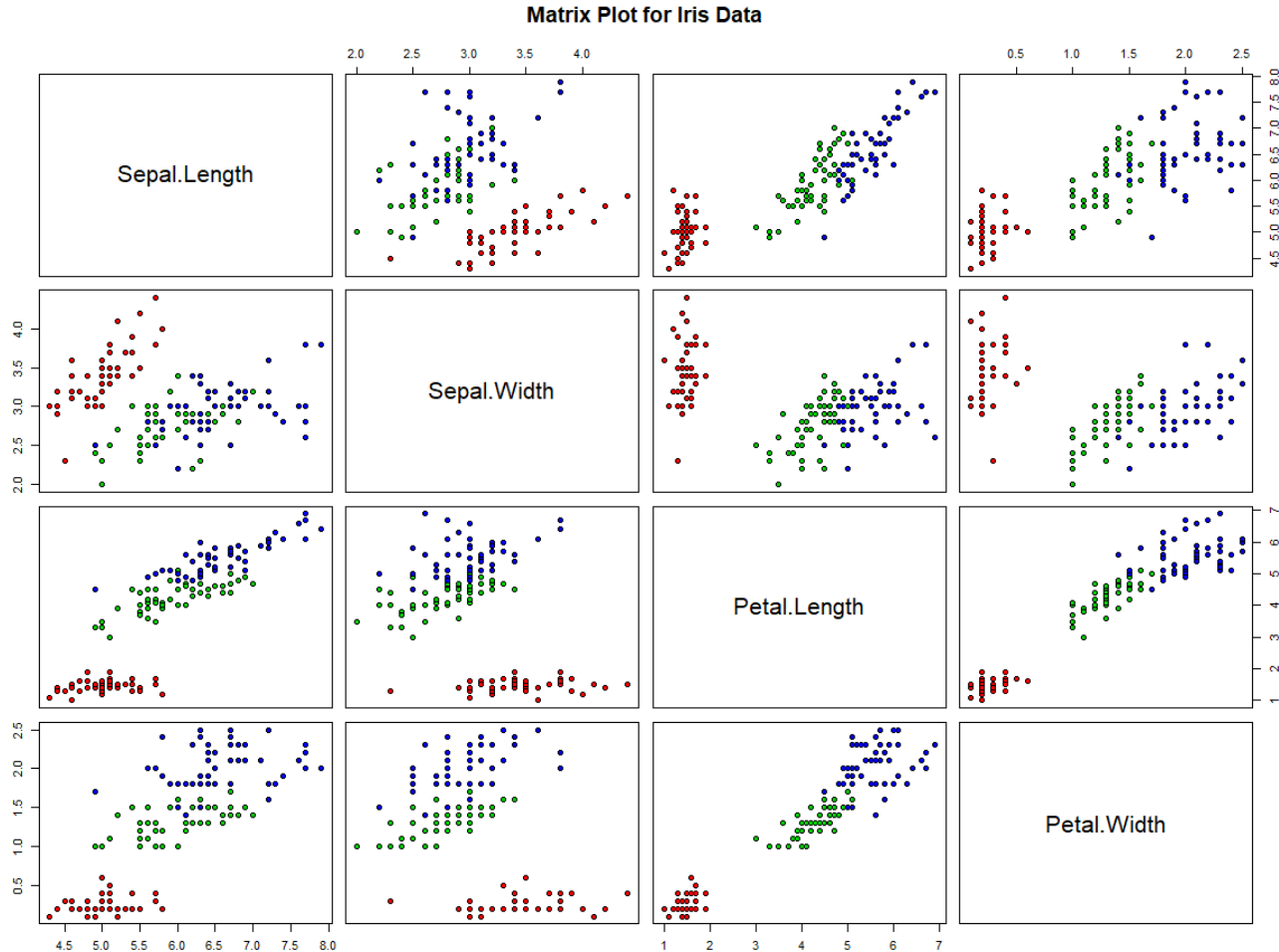
Iris Virginica

| Sepal.Length | Sepal.Width | Petal.Length | Petal.Width | Species |
|--------------|-------------|--------------|-------------|---------|
| 5.1 | 3.5 | 1.4 | 0.2 | 1 |
| 4.9 | 3 | 1.4 | 0.2 | 1 |
| 4.7 | 3.2 | 1.3 | 0.2 | 1 |
| 4.6 | 3.1 | 1.5 | 0.2 | 1 |
| 5 | 3.6 | 1.4 | 0.2 | 1 |
| 5.4 | 3.9 | 1.7 | 0.4 | 1 |
| ... | ... | ... | ... | ... |

Sepal: 꽃받침
Petal: 꽃잎

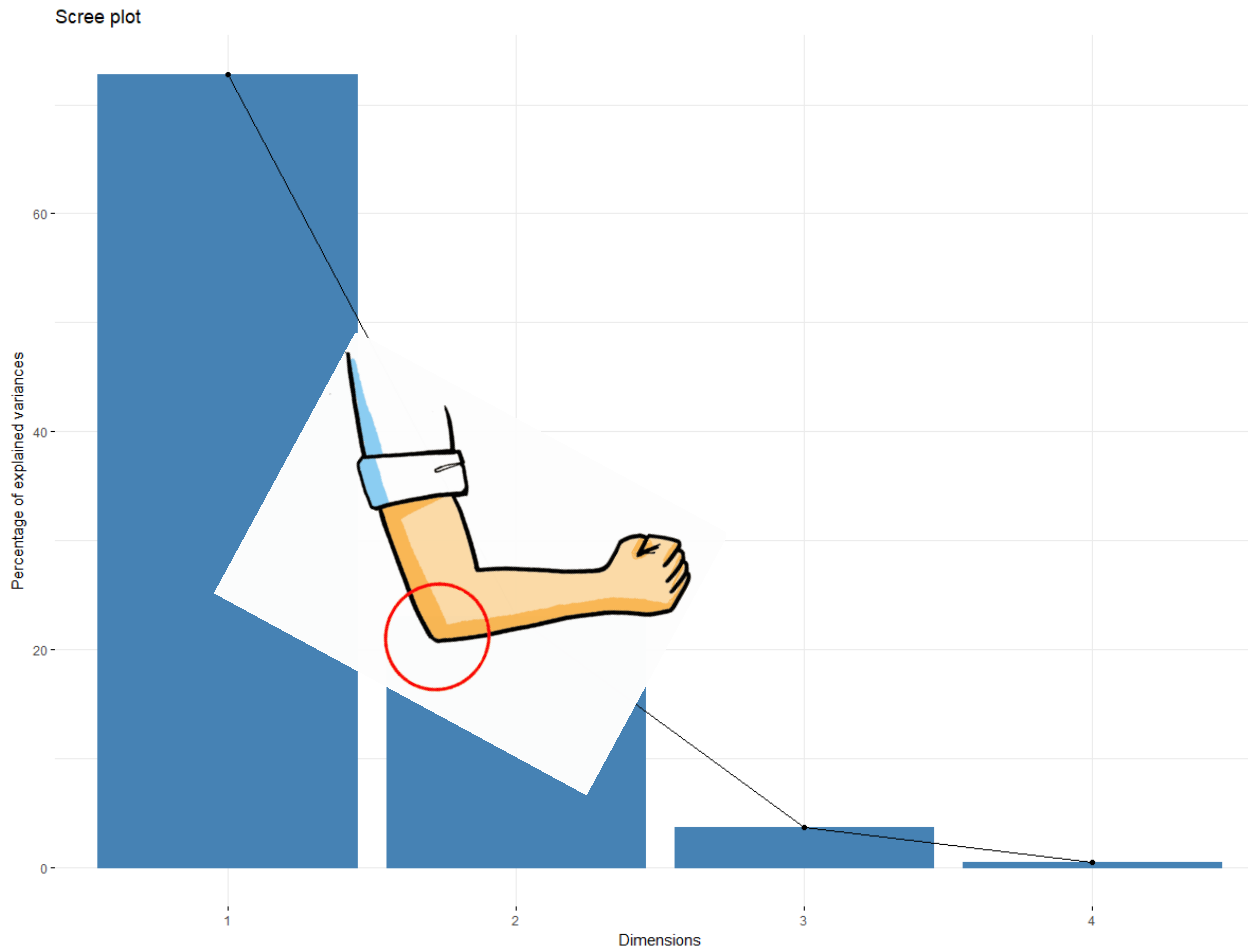
PCA – Matrix Plot

- 입력변수들이 대체로 강한 양의 상관관계를 보이고 있음



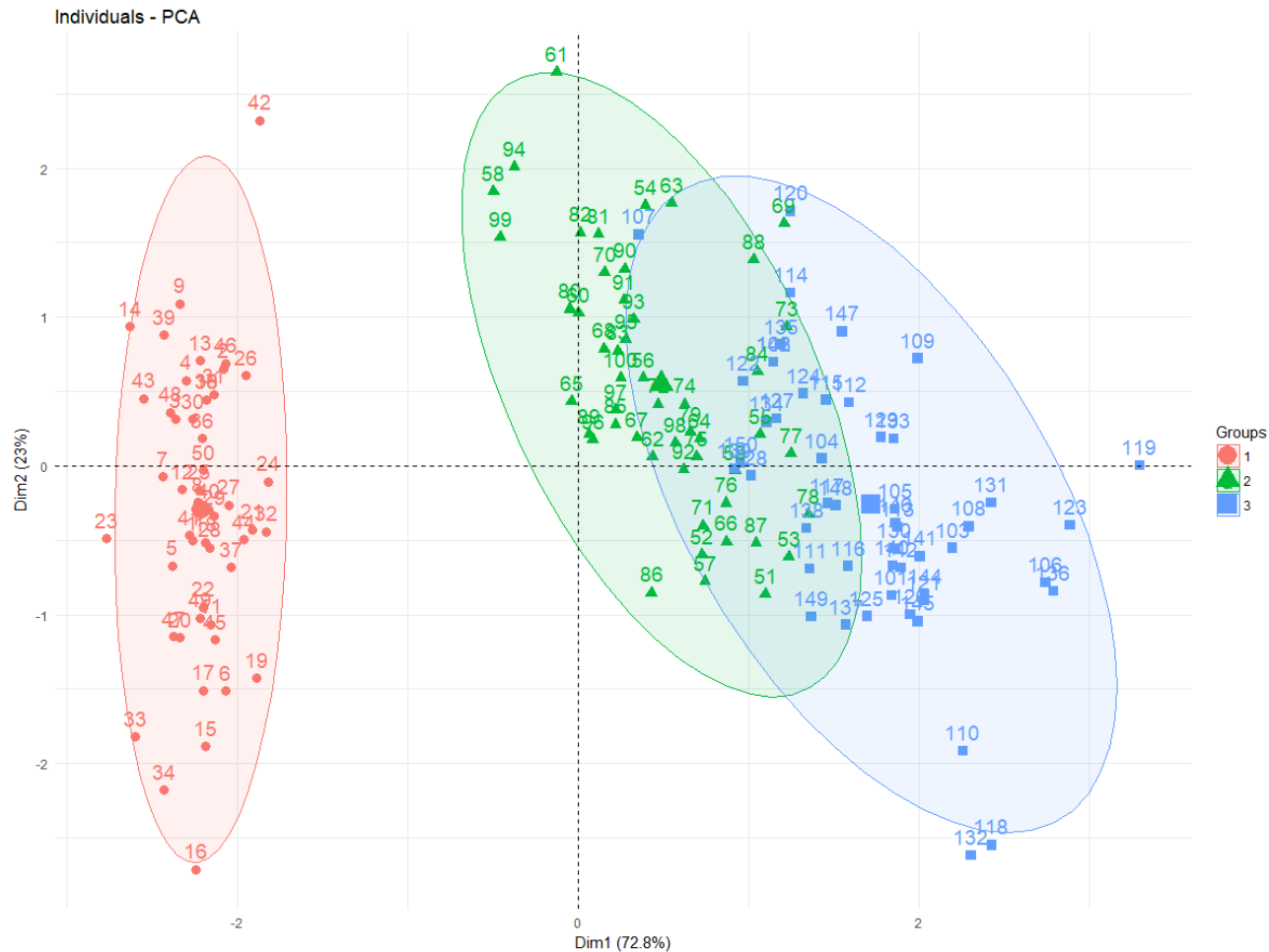
PCA – Scree Plot

- 1-2개의 주성분 (PC)로 충분



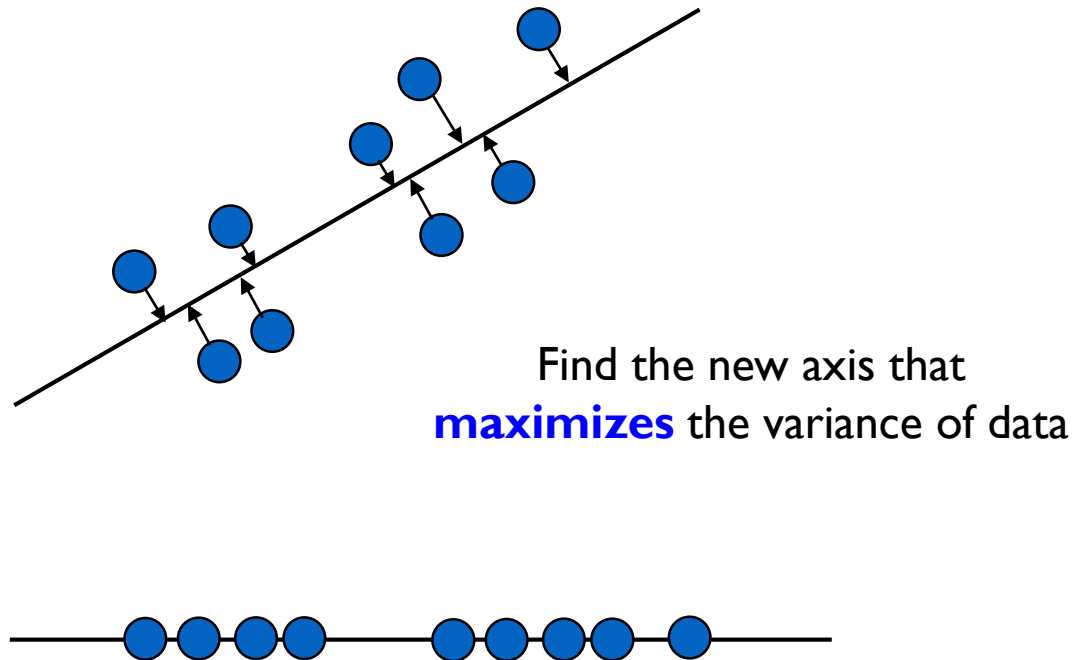
PCA – 2D Score Plot

- PCA의 핵심 그래프
- 같은 품종의 IRIS 가 같은 그룹에 포함



PCA Highlight

- 고차원의 원 데이터의 패턴을 유지하는 저차원 공간을 찾자
- 기존 데이터의 분산을 최대한 보존하는 새로운 차원(축)을 찾자
- 분산이 가장 큰 새로운 차원(축)을 찾자



EOD