SOON CHUN HYANG UNIVERSITY

PCA(Principal Component Analysis)

SCH Univ.
Dept. of AI and Bigdata
Lee Howoo

개요

1. 주성분 분석 개요

2. 주성분 분석 수리적 배경

3. 주성분 분석 알고리즘

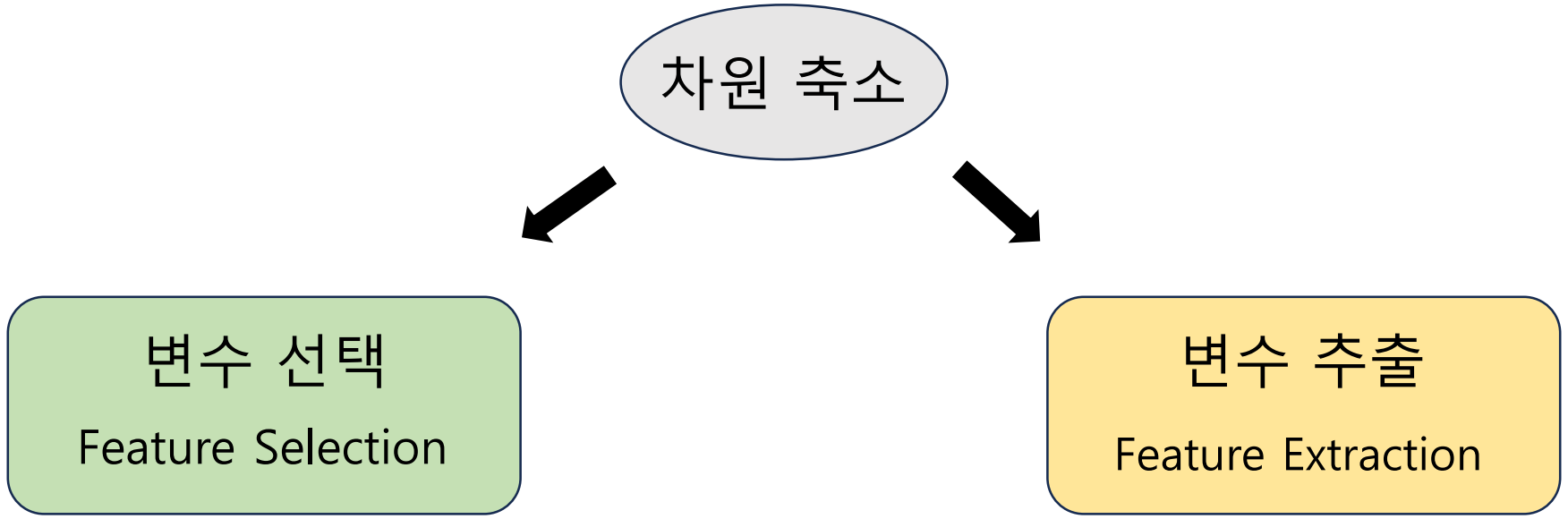
4. 주성분 분석 예제

주성분 분석 개요 – 고차원 데이터

변수 / 관측치	X_1	…	X_i	…	X_p
N_1	X_{11}	…	X_{1i}	…	X_{1p}
…	…	…	…	…	…
N_i	X_{i1}	…	X_{ii}	…	X_{ip}
…	…	…	…	…	…
N_n	X_{n1}	…	X_{ni}	…	X_{np}

- 3차원 이상의 데이터 → 시각적 표현 한계
- 차원이 클수록 계산복잡도 ↑, 모델링 비효율적
- 따라서 중요한 변수들로 차원을 줄일 필요가 있음

주성분 분석 개요



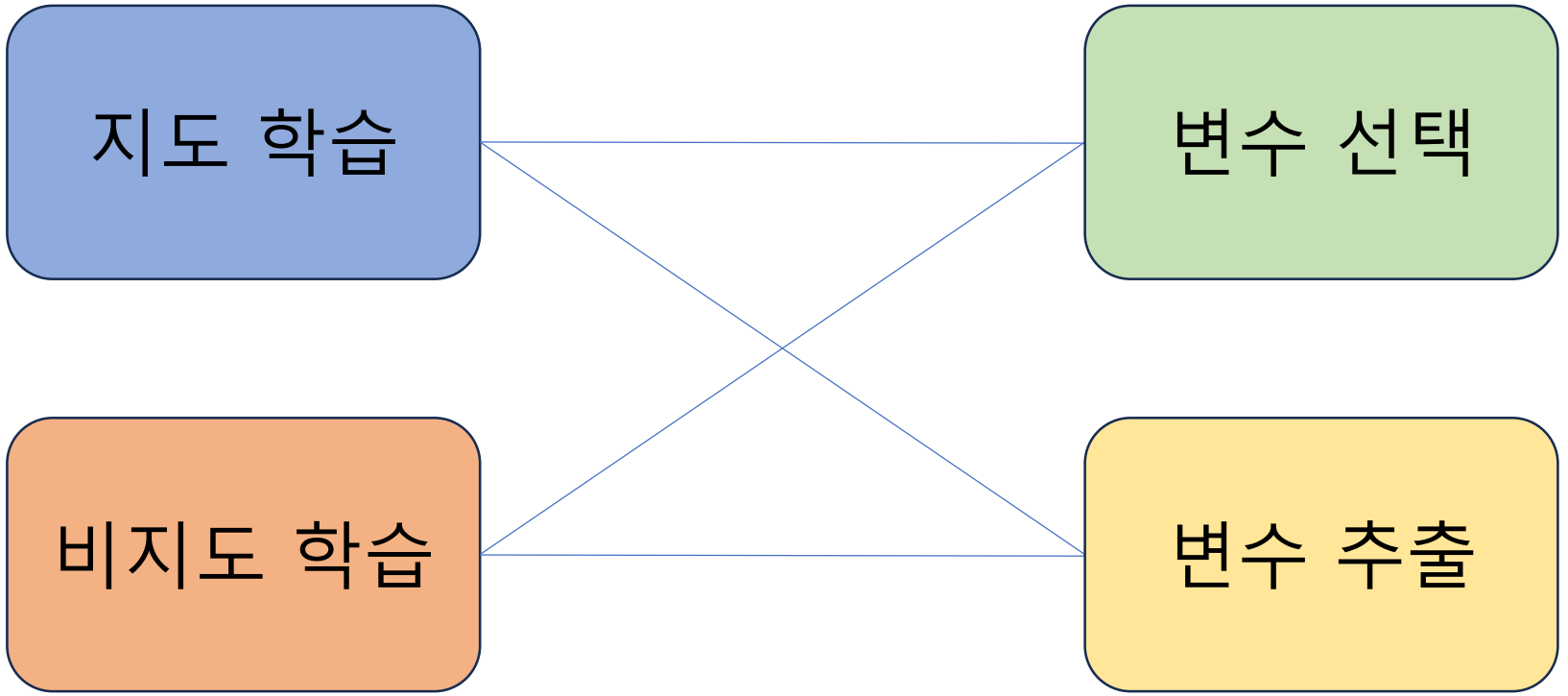
장점: 선택한 변수들의 해석이 용이함

단점: 변수간 상관관계 고려가 어려움

장점: 변수간 상관관계 고려함

단점: 추출된 변수의 해석이 어려움

주성분 분석 개요



주성분 분석 개요

Supervised feature selection: Information Gain, Stepwise regression, LASSO, Genetic algorithm

Supervised feature extraction: Partial least squares (PLS)

Unsupervised feature selection: PCA loading

Unsupervised feature extraction: Principal component analysis (PCA), Wavelets transforms, Autoencoder

주성분 분석 개요

Supervised feature selection: Information Gain, Stepwise regression, LASSO, Genetic algorithm

Supervised feature extraction: Partial least squares (PLS)

Unsupervised feature selection: PCA loading

Unsupervised feature extraction: Principal component analysis (PCA), Wavelets transforms, Autoencoder

주성분 분석 개요

PCA의 개념: 고차원 데이터를 효과적으로 분석하기 위한 대표적 분석 기법

목적) 차원 축소, 군집화, 시각화, 압축 등

PCA는 n개 관측치, p개 변수로 구성된 데이터를 상관관계가 없는 k개의 변수로 구성된 데이터로 요약하는 방식으로, 이 때 요약된 변수는 기존 변수의 선형 조합으로 생성됨

원래 데이터의 분산을 최대한 보존하는 새로운 축을 찾고, 그 축에 데이터를 사영 시키는 기법

주요 목적은 데이터 차원 축소 (n by p → n by k, where k << p)와 데이터 시각화 및 해석

전체 분석 과정 중 주로 초기에 사용됨 (변수 많을 때)

주성분 분석 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, , Z_3은 기존 변수인 X_1, X_2,…, X_p의 선형 조합으로 새롭게 생성된 변수

주성분 분석 개요

Z is linear combination (선형 결합) of the original all p variables in X

$$
\begin{aligned}
Z_1 &= \alpha_1^T X = \alpha_{11}X_1 + \alpha_{12}X_2 + \cdots + \alpha_{1p}X_p \\
Z_2 &= \alpha_2^T X = \alpha_{21}X_1 + \alpha_{22}X_2 + \cdots + \alpha_{2p}X_p \\
&\vdots \qquad\qquad\qquad\qquad \vdots \\
Z_p &= \alpha_p^T X = \alpha_{p1}X_1 + \alpha_{p2}X_2 + \cdots + \alpha_{1p}X_p
\end{aligned}
$$

$\boldsymbol{X_1}$, $\boldsymbol{X_2}$,⋯, $\boldsymbol{X_p}$: 원래 변수 (Original Variable)

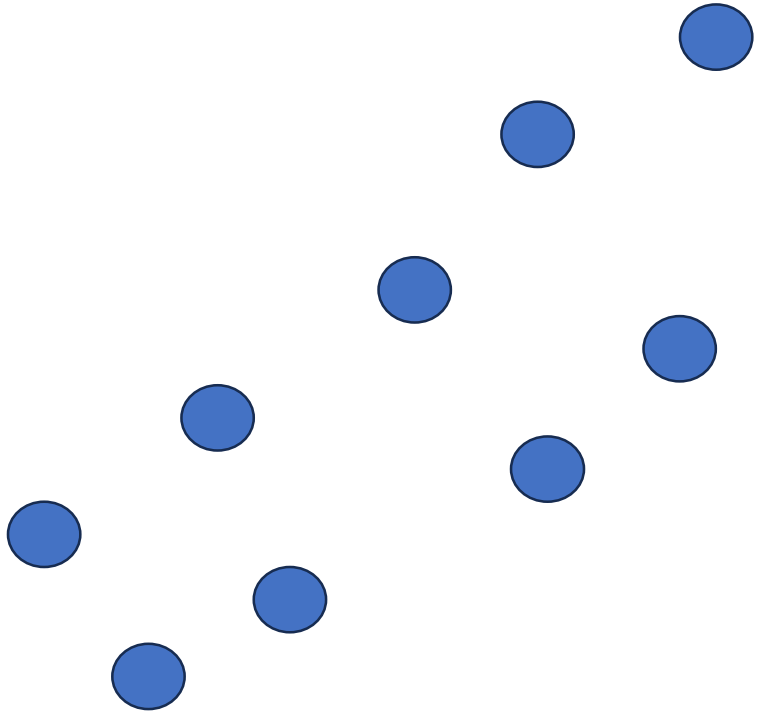
$\alpha_i = [\alpha_{i1}, \alpha_{i2},\cdots,\alpha_{ip}]$: i번째 기저(basis) 또는 계수 (Loading)

$\boldsymbol{Z_1}$, $\boldsymbol{Z_2}$,⋯, $\boldsymbol{Z_p}$: 각 기저로 사영 변환 후 변수 (주성분, Score)

주성분 분석 개요

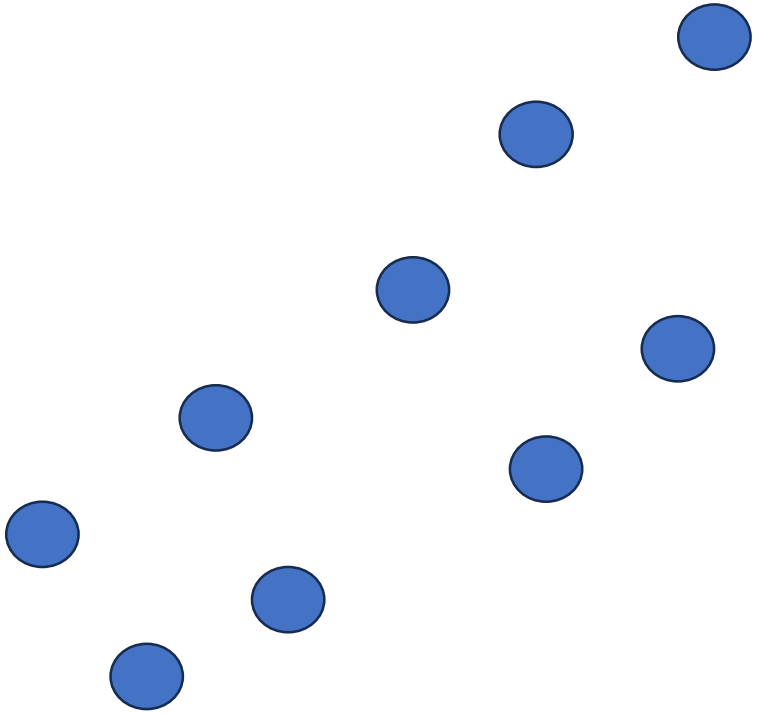
Find the new axis that

maximizes the variance of data



Find the new axis that

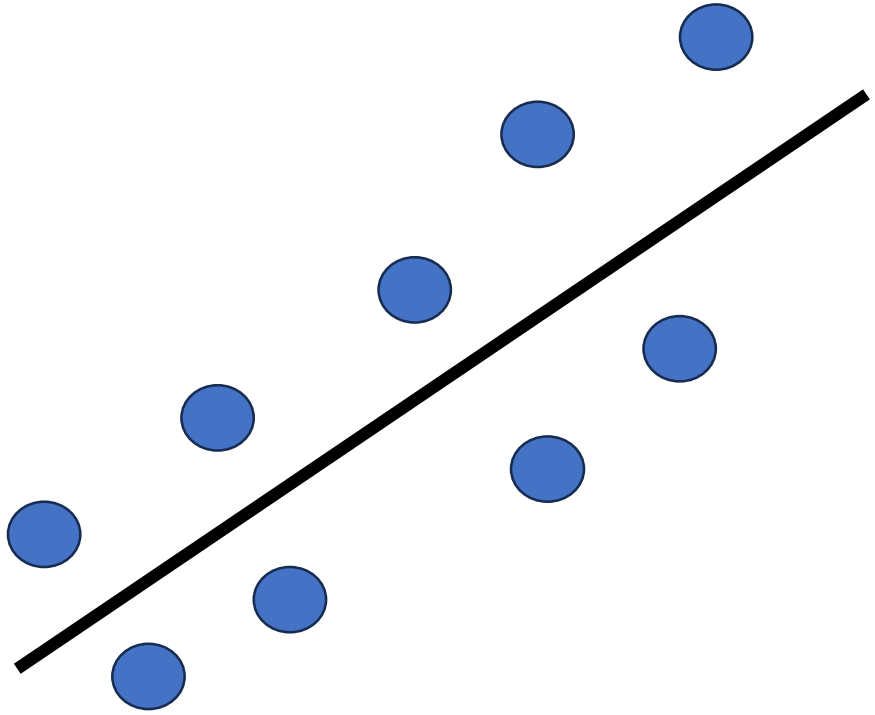
minimizes the variance of data



주성분 분석 개요

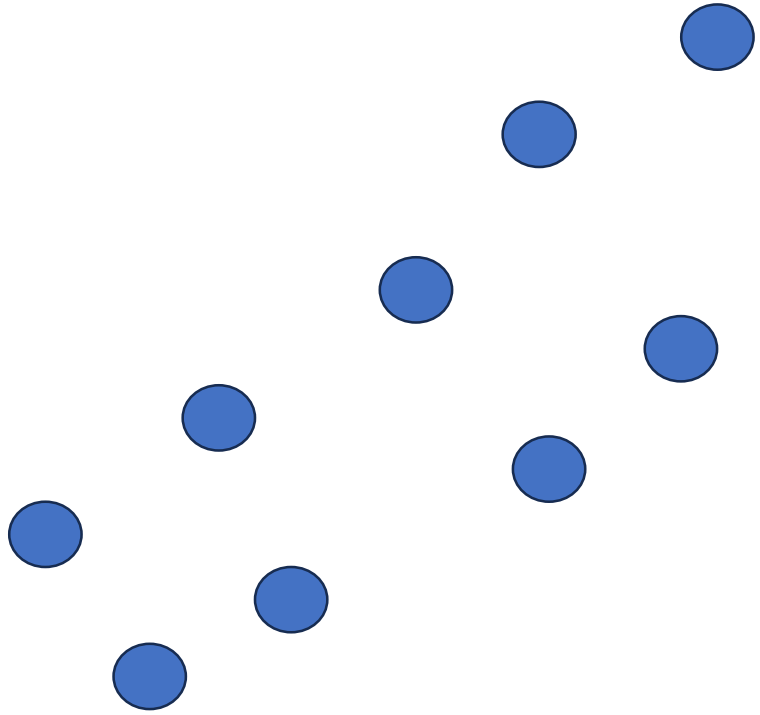
Find the new axis that

maximizes the variance of data



Find the new axis that

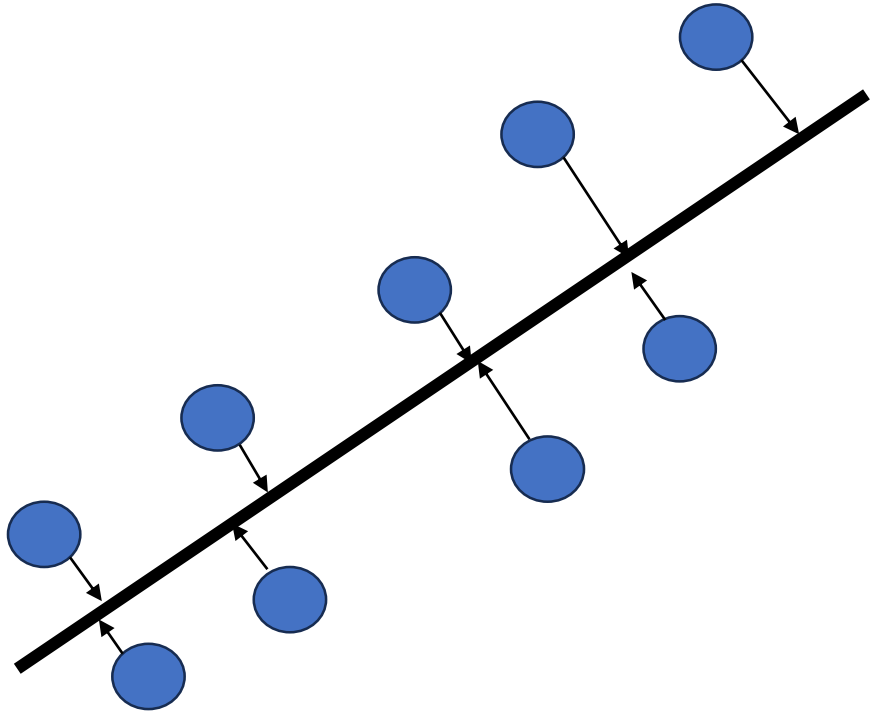
minimizes the variance of data



주성분 분석 개요

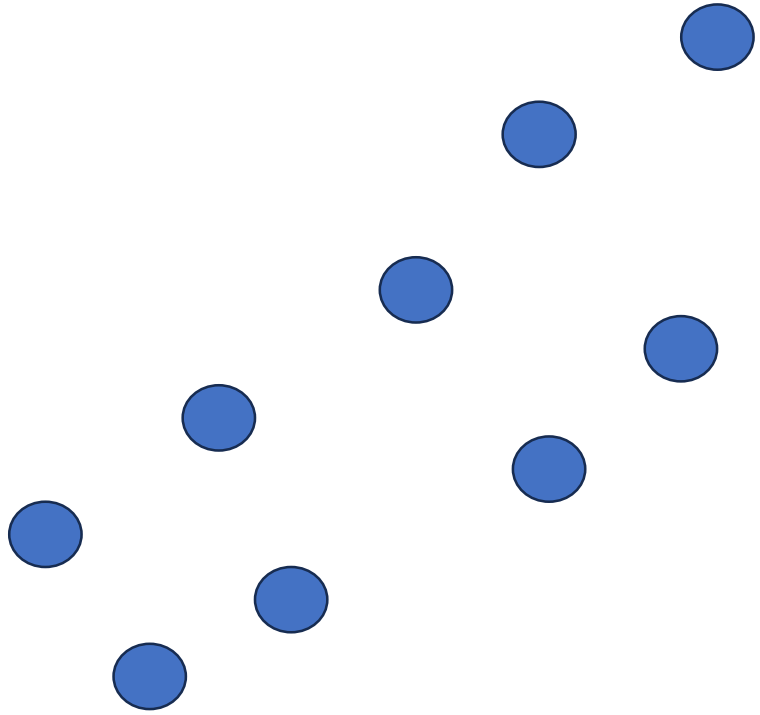
Find the new axis that

maximizes the variance of data



Find the new axis that

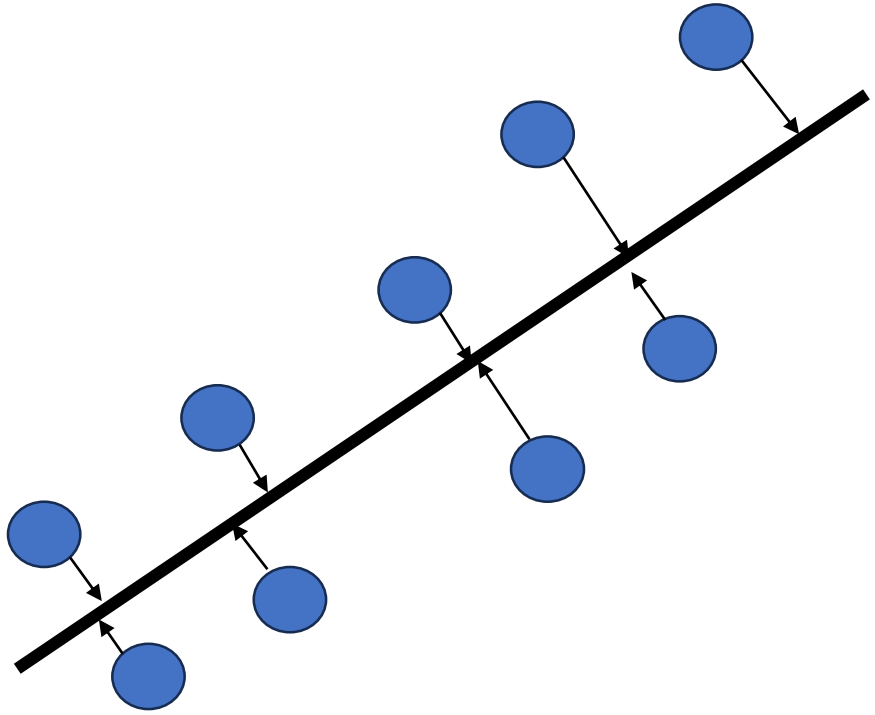
minimizes the variance of data

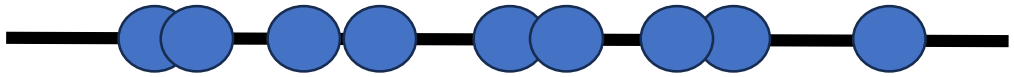


주성분 분석 개요

Find the new axis that

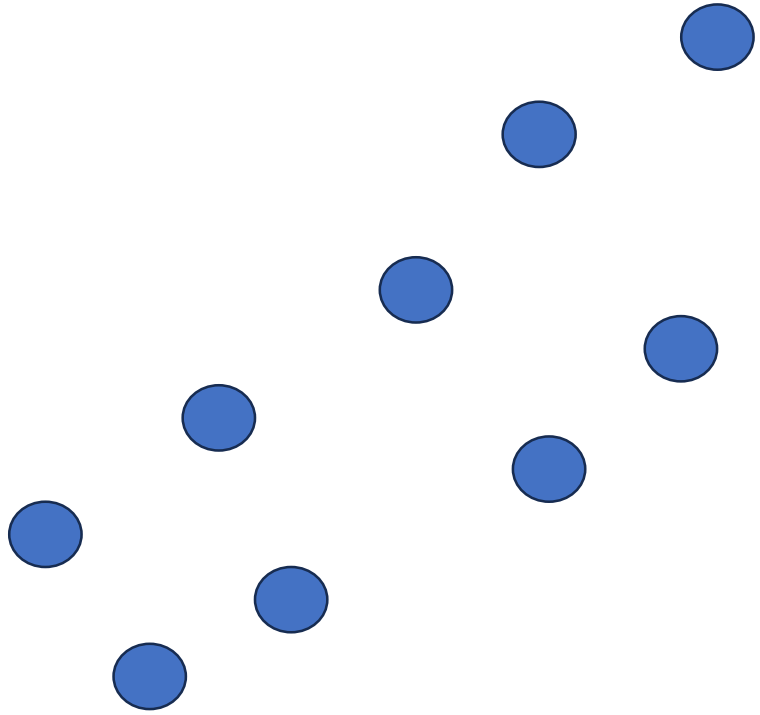
maximizes the variance of data





Find the new axis that

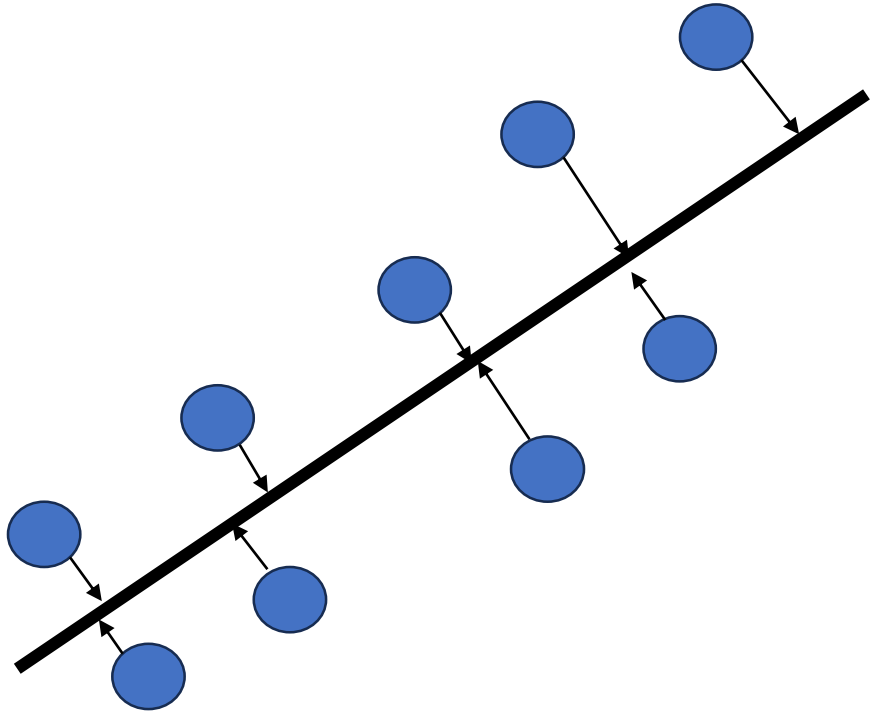
minimizes the variance of data

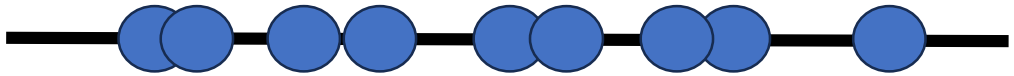


주성분 분석 개요

Find the new axis that

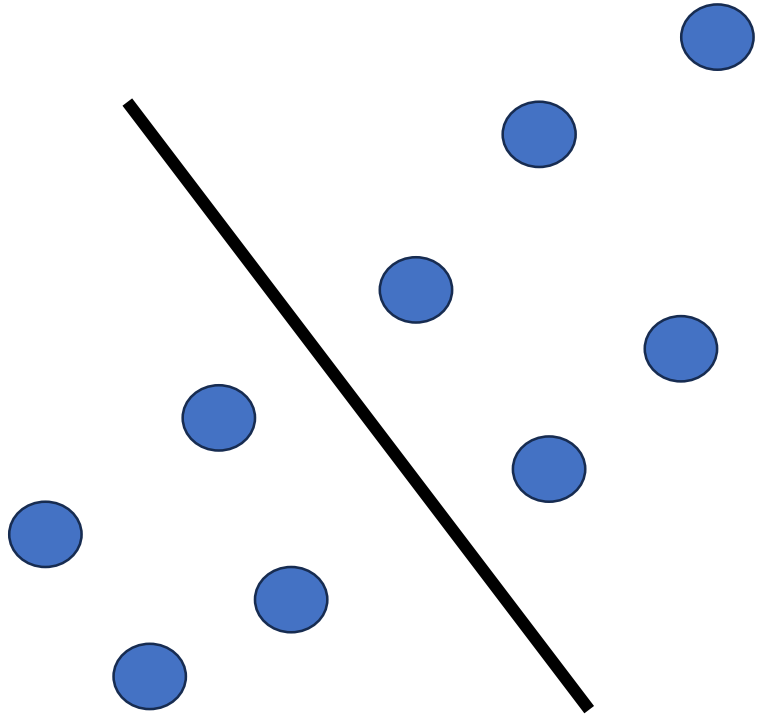
maximizes the variance of data





Find the new axis that

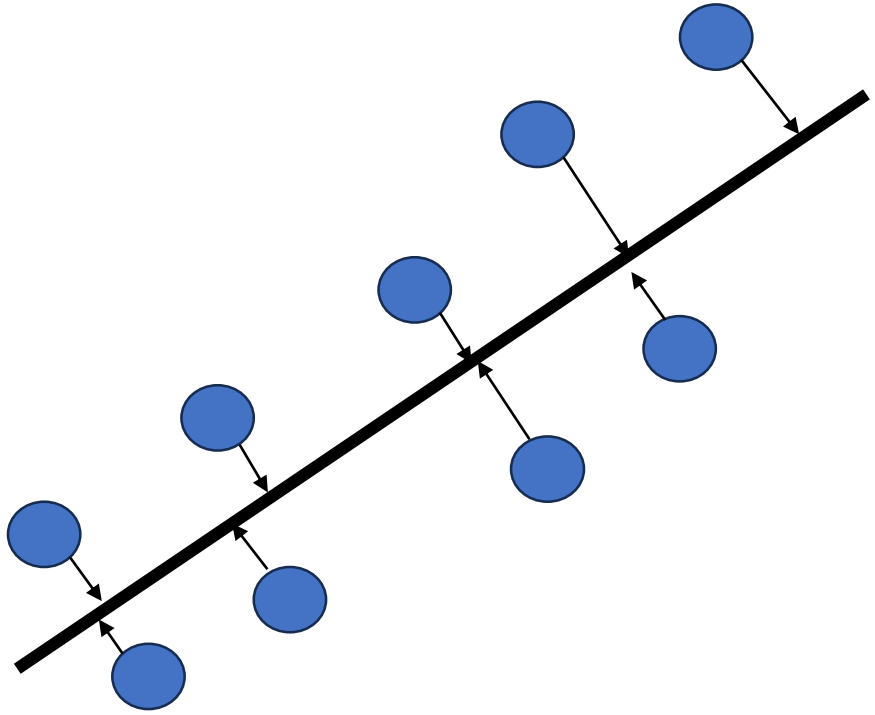
minimizes the variance of data

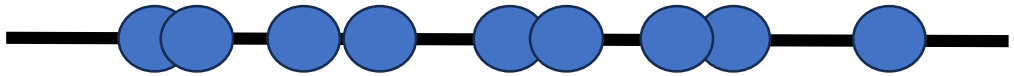


주성분 분석 개요

Find the new axis that

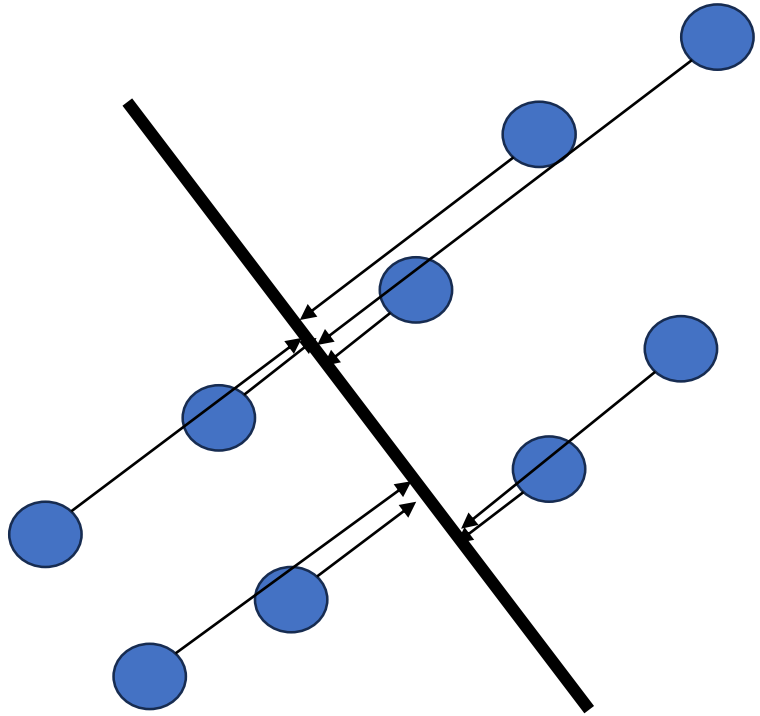
maximizes the variance of data





Find the new axis that

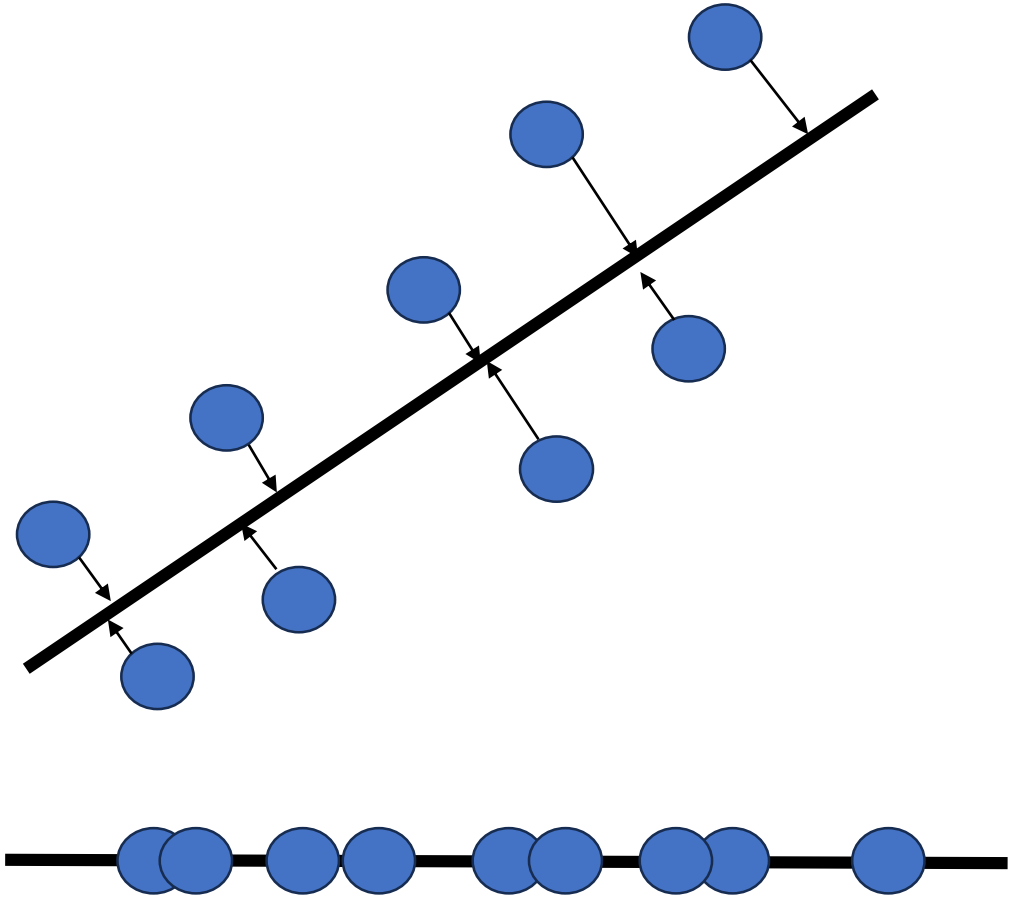
minimizes the variance of data



주성분 분석 개요

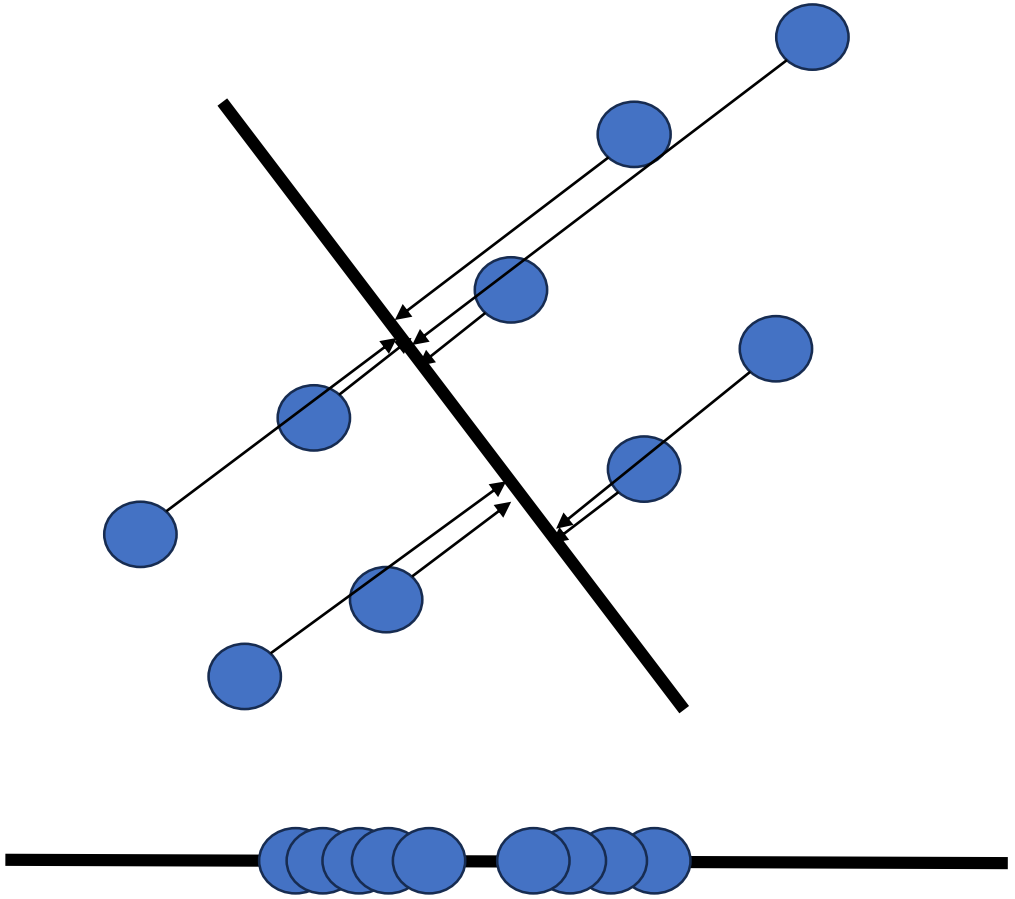
Find the new axis that

maximizes the variance of data



Find the new axis that

minimizes the variance of data



주성분 분석 수리적 배경

변수 / 관측치	X_1	⋯	X_i	⋯	X_p
N_1	X_{11}	⋯	X_{1i}	⋯	X_{1p}
⋯	⋯	⋯	⋯	⋯	⋯
N_i	X_{i1}	⋯	X_{ii}	⋯	X_{ip}
⋯	⋯	⋯	⋯	⋯	⋯
N_n	X_{n1}	⋯	X_{ni}	⋯	X_{np}

$$\overline{X} = \begin{bmatrix} \overline{X}_1 \\ \overline{X}_2 \\ \cdots \\ \overline{X}_p \end{bmatrix}$$

Mean vector

$$C_n = \begin{bmatrix} S_{11} & \cdots & S_{1p} \\ \vdots & \ddots & \vdots \\ S_{p1} & \cdots & S_{pp} \end{bmatrix}$$

Covariance Matrix

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

Correlation Matrix

주성분 분석 수리적 배경

- 공분산 (Covariance)의 성질
  - X를 p개의 변수와 n개의 개체로 구성된 n by p 행렬로 정의할 때 X의 공분산 행렬은 다음과 같음

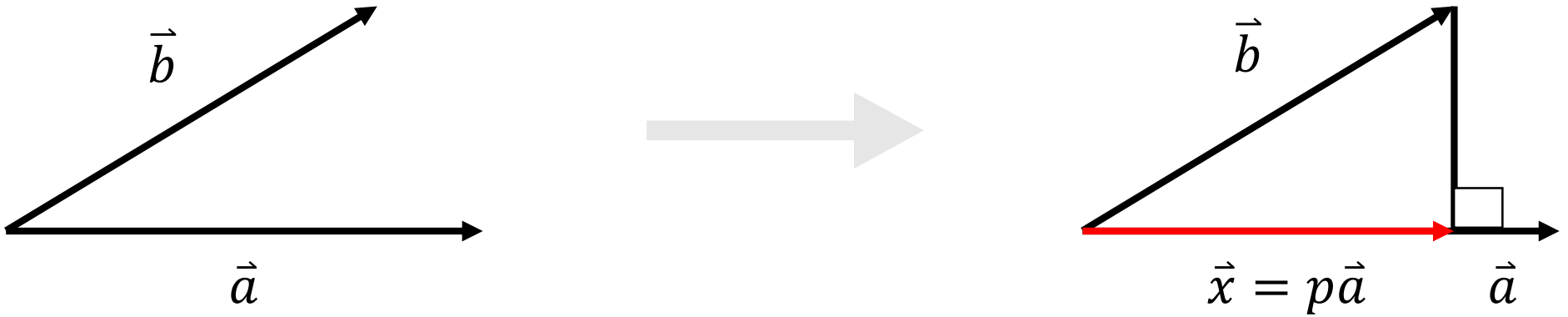
$$Cov(X) = \frac{1}{n}(X - \bar{X})\,(X - \bar{X})^T$$

  - 공분산 행렬의 대각 성분은 각 변수의 분산과 같으며, 비대각행렬은 대응하는 두 변수의 공분산과 같음

  - 데이터의 총분산은 공분산행렬의 대각성분들의 합으로 표현됨

주성분 분석 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

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

주성분 분석 수리적 배경

- 고유값(eigenvalue) 및 고유벡터(eigenvector)
  - 어떤 행렬 A에 대해 상수 λ와 벡터 x가 다음 식을 만족할 때,

    λ와 x를 각각 행렬 A의 고유값, 고유벡터라고 함

$$Ax = \lambda x \rightarrow (A - \lambda I)x = 0$$

  - 벡터에 행렬을 곱한다는 것은 해당 벡터를 선형변환(linear transformation)

    한다는 의미 → 고유벡터는 이 변환에 의해 방향이 변하지 않는 벡터를 의미

주성분 분석 알고리즘 – 주성분 추출

- Assume that we have the centered data (i.e., $\bar{X}_i = 0,\ \mathrm{i} = 1,\cdots,\mathrm{p}$)
- Let 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 $\mathrm{Z} = \alpha^T\mathrm{X}$ be the projection of X onto the direction α

The main purpose in PCA is

to find α that produces the largest variance of Z

$$Max\ Var(Z) = Var(\alpha^T X) = \alpha^T \mathrm{Var(x)}\alpha = \alpha^T \Sigma \alpha$$
$$s.t. \|\alpha\| = \alpha^T\alpha = 1$$

주성분 분석 알고리즘 – 주성분 추출

$$\Sigma = E\Lambda E^T$$

$$Max\ \alpha^T \Sigma \alpha = \alpha^T E \Lambda E^T \alpha$$

$$s.t. \|\alpha\| = 1$$

eigenvalues and eigenvector of Σ

$$[E \wedge V] = svd(\Sigma)$$

$$\lambda_1 \geq \cdots \lambda_p \geq 0$$

$$e_1, \cdots, e_p$$

$$\Lambda = diag(\lambda_1, \cdots, \lambda_p)$$

$$Max\ \beta^T \Lambda\, \beta\ where\ \beta = E^T \alpha$$

$$s.t. \|\beta\| = 1$$

$$Max\ \lambda_1 {\beta_1}^2 + \lambda_2 {\beta_2}^2 + \cdots + \lambda_p {\beta_p}^2$$

$$s.t.\ {\beta_1}^2 + {\beta_2}^2 + \cdots + {\beta_p}^2 = 1$$

$$\lambda_1 > \lambda_2 > \cdots > \lambda_p$$

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

주성분 분석 알고리즘 – 주성분 추출

$$\Sigma = E\Lambda E^T$$

$$Max\ \alpha^T \Sigma \alpha = \alpha^T E \Lambda E^T \alpha$$

$$s.t. \|\alpha\| = 1$$

eigenvalues and eigenvector of Σ

$$[E \wedge V] = svd(\Sigma)$$

$$\lambda_1 \geq \cdots \lambda_p \geq 0$$

$$e_1, \cdots, e_p$$

$$\Lambda = diag(\lambda_1, \cdots, \lambda_p)$$

$$Max\ \beta^T \Lambda\, \beta\ where\ \beta = E^T \alpha$$

$$s.t. \|\beta\| = 1$$

$$Max\ \lambda_1 {\beta_1}^2 + \lambda_2 {\beta_2}^2 + \cdots + \lambda_p {\beta_p}^2$$

$$s.t.\ {\beta_1}^2 + {\beta_2}^2 + \cdots + {\beta_p}^2 = 1$$

$$\lambda_1 > \lambda_2 > \cdots > \lambda_p$$

$$\begin{bmatrix} \lambda_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \lambda_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \lambda_p \end{bmatrix}$$

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

주성분 분석 알고리즘 – 주성분 추출

$$\Sigma = E\Lambda E^T$$

$$Max\ \alpha^T \Sigma \alpha = \alpha^T E \Lambda E^T \alpha$$

$$s.t. \|\alpha\| = 1$$

eigenvalues and eigenvector of Σ

$$[E \wedge V] = svd(\Sigma)$$

$$\lambda_1 \geq \cdots \lambda_p \geq 0$$

$$e_1, \cdots, e_p$$

$$\Lambda = diag(\lambda_1, \cdots, \lambda_p)$$

$$Max\ \beta^T \Lambda\, \beta\ where\ \beta = E^T \alpha$$

$$s.t. \|\beta\| = 1$$

$$Max\ \lambda_1 {\beta_1}^2 + \lambda_2 {\beta_2}^2 + \cdots + \lambda_p {\beta_p}^2$$

$$s.t.\ {\beta_1}^2 + {\beta_2}^2 + \cdots + {\beta_p}^2 = 1$$

$$\lambda_1 > \lambda_2 > \cdots > \lambda_p$$

$$\beta_1 = 1$$
$$\beta_2 = 0$$
$$\vdots$$
$$\beta_p = 0$$

$$\begin{bmatrix} \lambda_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \lambda_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \lambda_p \end{bmatrix}$$

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

주성분 분석 알고리즘 – 주성분 추출

$$\Sigma = E\Lambda E^T$$

$$Max\ \alpha^T \Sigma \alpha = \alpha^T E \Lambda E^T \alpha$$

$$s.t. \|\alpha\| = 1$$

eigenvalues and eigenvector of Σ

$$[E \wedge V] = svd(\Sigma)$$

$$\lambda_1 \geq \cdots \lambda_p \geq 0$$

$$e_1, \cdots, e_p$$

$$\Lambda = diag(\lambda_1, \cdots, \lambda_p)$$

$$Max\ \beta^T \Lambda\, \beta\ where\ \beta = E^T \alpha$$

$$s.t. \|\beta\| = 1$$

$$Max\ \lambda_1 {\beta_1}^2 + \lambda_2 {\beta_2}^2 + \cdots + \lambda_p {\beta_p}^2$$

$$s.t.\ {\beta_1}^2 + {\beta_2}^2 + \cdots + {\beta_p}^2 = 1$$

$$\lambda_1 > \lambda_2 > \cdots > \lambda_p$$

$$\beta_1 = 1,\ \beta_2 = 0,\ \vdots,\ \beta_p = 0 \qquad \alpha = E\beta$$

$$\begin{bmatrix} \lambda_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \lambda_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \lambda_p \end{bmatrix}$$

Thus, the optimal value is $\boldsymbol{\lambda_1}$ and $\boldsymbol{\alpha = e_1}$

주성분 분석 알고리즘 – 주성분 추출

$$\Sigma = E\Lambda E^T$$

$$Max\ \alpha^T \Sigma \alpha = \alpha^T E \Lambda E^T \alpha$$

$$s.t. \|\alpha\| = 1$$

eigenvalues and eigenvector of Σ

$$[E \wedge V] = svd(\Sigma)$$

$$\lambda_1 \geq \cdots \lambda_p \geq 0$$

$$e_1, \cdots, e_p$$

$$\Lambda = diag(\lambda_1, \cdots, \lambda_p)$$

$$Max\ \beta^T \Lambda\, \beta\ where\ \beta = E^T \alpha$$

$$s.t. \|\beta\| = 1$$

$$Max\ \lambda_1 {\beta_1}^2 + \lambda_2 {\beta_2}^2 + \cdots + \lambda_p {\beta_p}^2$$

$$s.t.\ {\beta_1}^2 + {\beta_2}^2 + \cdots + {\beta_p}^2 = 1$$

$$\lambda_1 > \lambda_2 > \cdots > \lambda_p$$

$$\beta_1 = 1,\ \beta_2 = 0,\ \cdots,\ \beta_p = 0$$

$$\alpha = E\beta = [e_1 \quad \cdots \quad e_p] \begin{bmatrix} \beta_1 \\ \cdots \\ \beta_p \end{bmatrix} \begin{bmatrix} \lambda_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \lambda_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \lambda_p \end{bmatrix}$$

Thus, the optimal value is $\boldsymbol{\lambda_1}$ and $\boldsymbol{\alpha = e_1}$

주성분 분석 예제

$\mathbf{X} =$

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

$\mathbf{X} =$

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 X to

$\mathrm{E}(X_i) = 0$

$\mathrm{Var}(X_i) = 1$

주성분 분석 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

$\mathrm{E}(X_i) = 0$

$\mathrm{Var}(X_i) = 1$

$\mathrm{Covariance}(X) =$

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

$\mathrm{Correlation}(X) =$

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

주성분 분석 예제

The eigenvalue-eigenvector pairs on the correlation matrix, Σ

$$[E \wedge V] = svd(\Sigma)$$

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

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

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

주성분 분석 예제

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

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

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

$\lambda_3 > \lambda_2 > \lambda_1$

$\mathbf{X} =$

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 X to

$\mathrm{E}(X_i) = 0$

$\mathrm{Var}(X_i) = 1$

주성분 분석 예제

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

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

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

$\mathbf{X} =$

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 X to

$\mathrm{E}(X_i) = 0$

$\mathrm{Var}(X_i) = 1$

$\lambda_3 > \lambda_2 > \lambda_1$

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

주성분 분석 예제

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

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

$\mathbf{X} =$

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 X to

$\mathrm{E}(X_i) = 0$

$\mathrm{Var}(X_i) = 1$

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

$$Z_2 = e_2{}^T X = \begin{bmatrix} -0.0615 \\ 0.4912 \\ -0.2798 \\ -0.4703 \\ 0.3204 \end{bmatrix} \qquad Z_3 = e_1{}^T X = \begin{bmatrix} 0.3160 \\ -0.1493 \\ -0.4047 \\ 0.1223 \\ 0.1157 \end{bmatrix}$$

주성분 분석 예제

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

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

$\mathbf{X} =$

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 X to

$\mathrm{E}(X_i) = 0$

$\mathrm{Var}(X_i) = 1$

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

$$Z_2 = {e_2}^T X = \begin{bmatrix} -0.0615 \\ 0.4912 \\ -0.2798 \\ -0.4703 \\ 0.3204 \end{bmatrix} \qquad Z_3 = {e_1}^T X = \begin{bmatrix} 0.3160 \\ -0.1493 \\ -0.4047 \\ 0.1223 \\ 0.1157 \end{bmatrix}$$

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

주성분 분석 예제

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

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

주성분 (Z)들은 서로 독립

주성분 분석 예제

$\mathbf{X} =$

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

$\mathbf{Z} =$

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

주성분 몇 개 사용?

주성분 분석 예제

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

= Variance of each principal component (각 주성분의 분산)

Covariance matrix of principal components

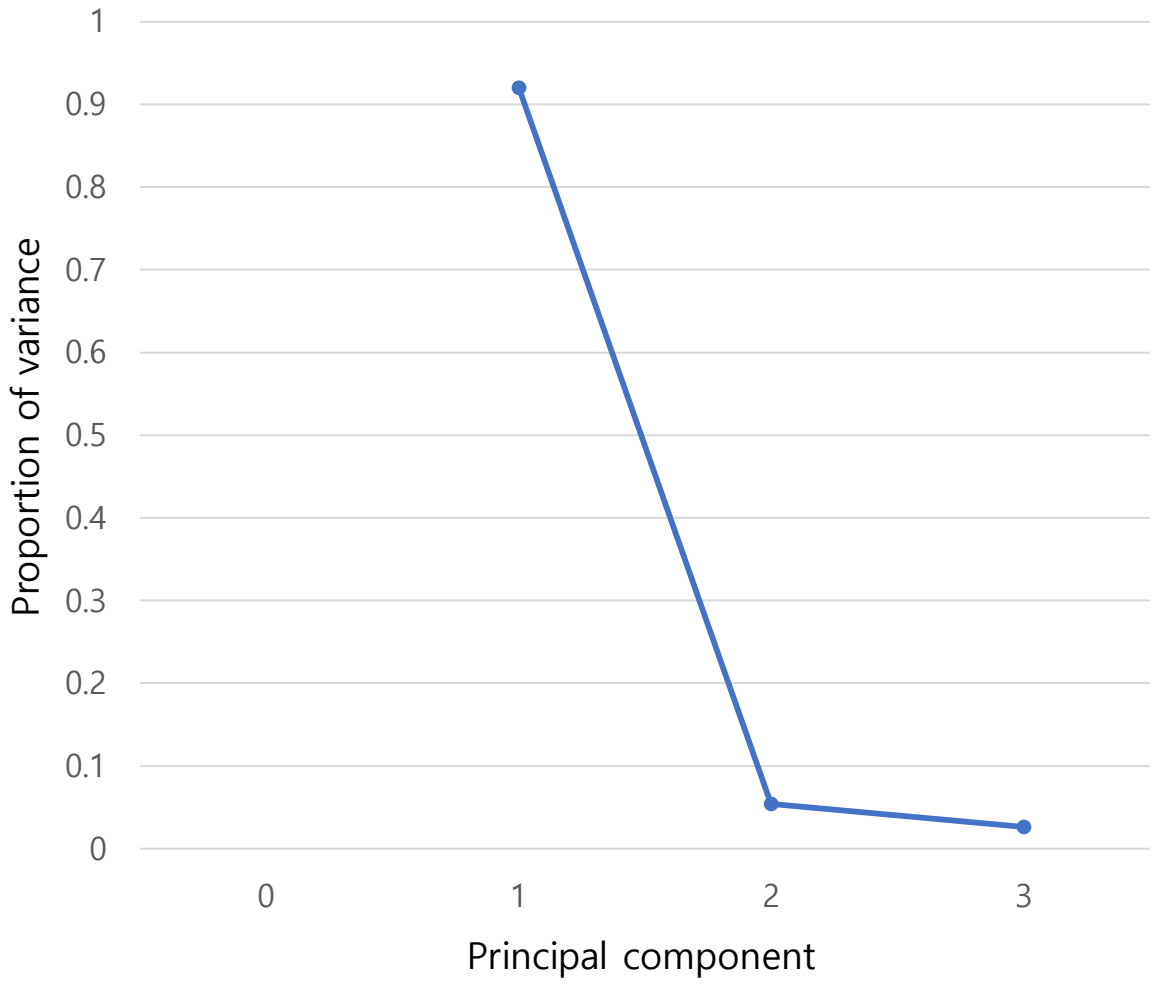
$$Cov(Z) = \begin{bmatrix} 2.7596 & 0 & 0 \\ 0 & 0.1618 & 0 \\ 0 & 0 & 0.0786 \end{bmatrix}$$

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

Proportion of total population Variance due to the 1st Principal component

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

주성분 분석 예제



선택방식 1: 고유값 감소율이 유의미하게 낮아지는 Elbow Point에 해당하는 주성분 수를 선택

선택방식 2: 일정 수준 이상의 분산비를 보존하는 최소의 주성분을 선택 (70% 이상)

주성분 분석 알고리즘 요약

Step 1. 데이터 정규화 **(mean centering)**

Step 2. 기존 변수의 **covariance (correlation) matrix** 계산

Step 3. **Covariance (correlation) matrix**로부터 **eigenvalue** 및 이에 해당되는 **eigenvector**를 계산

Step 4. **eigenvalue** 및 해당되는 **eigenvectors**를 순서대로 나열

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

$$\mathrm{e}(1) > e(2) > e(3) > e(4) > e(5), \quad e(i), \quad i = 1, \cdots, 5 \; is \; a \; vector$$

Step 5. 정렬된 **eigenvector**를 토대로 기존 변수를 변환

$$Z_1 = e(1)X = e_{11} \cdot X_1 + e_{12} \cdot X_2 + \cdots + e_{15} \cdot X_5$$
$$Z_2 = e(2)X = e_{21} \cdot X_1 + e_{22} \cdot X_2 + \cdots + e_{25} \cdot X_5$$
$$\cdots \qquad = \qquad \cdots$$
$$Z_5 = e(5)X = e_{51} \cdot X_1 + e_{52} \cdot X_2 + \cdots + e_{55} \cdot X_5$$

Thank you