Principal Component Analysis (PCA)

Namwoo Kang

Smart Design Lab
CCS Graduate School of Green Transportation
KAIST

Reference

❑ 강의 슬라이드 및 실습코드는 아래의 링크에서 받으실 수 있습니다

- http://www.smartdesignlab.org/dl_aischool_2021.html
- Contributors: 김성신, 유소영, 이성희, 김은지

❑ 강의 소스

- Andrew Ng의 ML Class (www.holehouse.org/mlclass/)
- Fei-Fei Li & Justin Johnson & Serena Yeung, CS231n: Convolutional Neural Networks for Visual Recognition, Stanford (http://cs231n.stanford.edu/)
- Stefano Ermon & Aditya Grover, CS 236: Deep Generative Models , Stanford (https://deepgenerativemodels.github.io/)
- 모두를 위한 딥러닝 (https://hunkim.github.io/ml/)
- 모두를 위한 딥러닝 시즌 2 (https://deeplearningzerotoall.github.io/season2/lec_tensorflow.html)
- 이활석, Autoencoders (https://www.slideshare.net/NaverEngineering/ss-96581209)
- 최윤제, 1시간만에 GAN(Generative Adversarial Network) 완전 정복하기 (https://www.slideshare.net/NaverEngineering/1-gangenerative-adversarial-network?qid=c53ce33f-6643-4437-8e93-88776c9cebb1&v=&b=&from_search=5)
- 김성범, [핵심 머신러닝] Principal Component Analysis (PCA, 주성분 분석) (https://youtu.be/FhQm2Tc8Kic)

SMART DESIGN LAB KAIST

Contents

- **Ch1: Introduction to Unsupervised Learning Part I** → Probability & Maximum Likelihood
- **Ch2: Introduction to Unsupervised Learning Part II** → Generative Model & Dimensionality Reduction
- **Ch3: Principal Component Analysis (PCA)** → Machine Learning Model
- **Ch4: Autoencoder & Anomaly Detection**
  + 실습
- **Ch5: Variational AutoEncoder (VAE)**
  + 실습
- **Ch6: Generative Adversarial Network (GAN)**
  + 실습

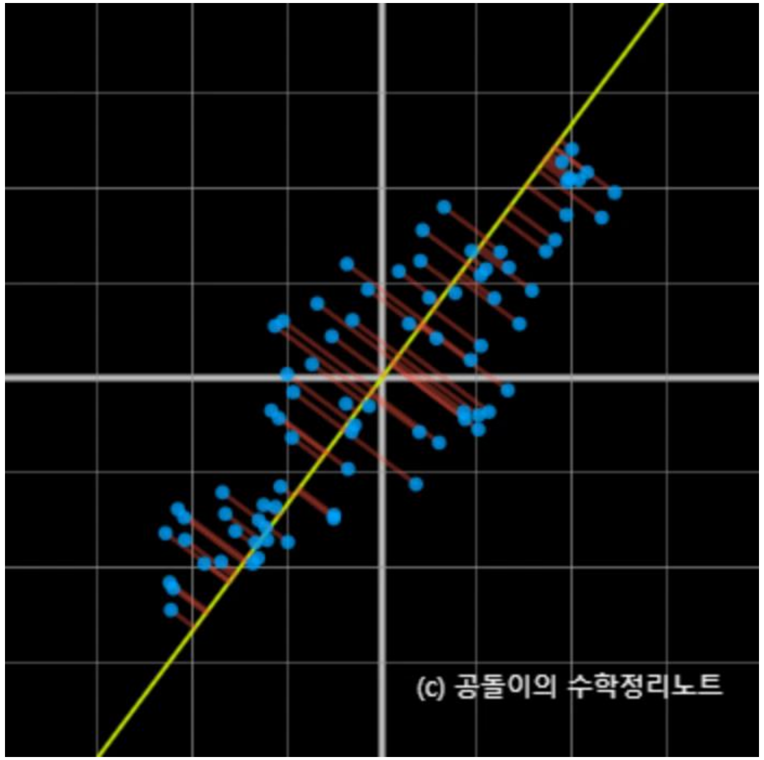
→ Deep Learning Models

- **Ch7: Application: Mechanical Design + AI** → CAD/CAM/CAE/Design Optimization + AI

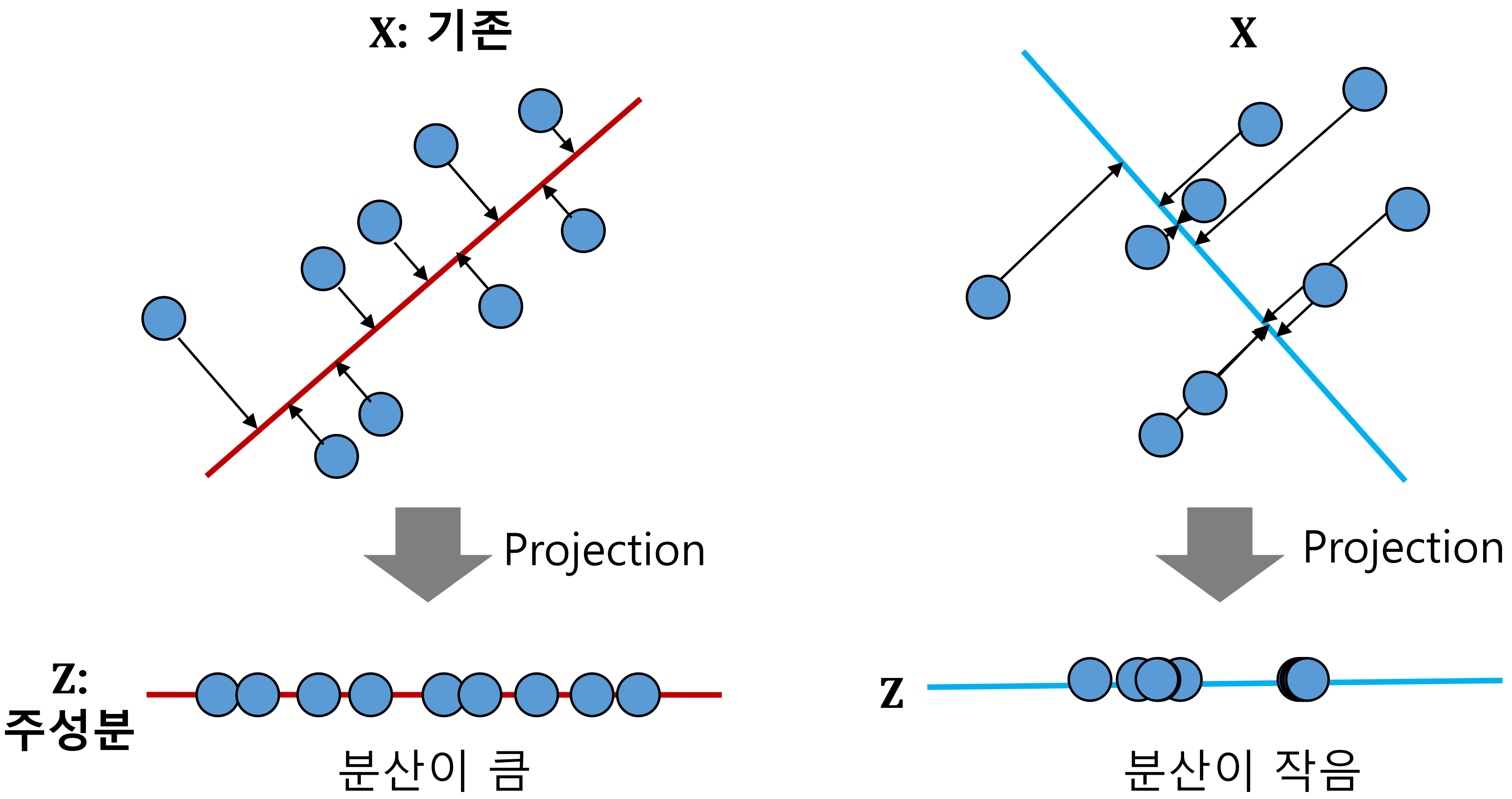
Concept of PCA

❖ Principal Component Analysis (PCA), 주성분 분석

- 원래 데이터의 **<u>분산을 최대한 보존하는(구조를 잘 유지하는)</u>** 새로운 축을 찾고, 그 축에 데이터를 사영(Projection) 시키는기법



(https://angeloyeo.github.io/2019/07/27/PCA.html)



Maximize the variance of data

Concept of PCA

- $\mathbf{z}$ is a linear combination (**<u>선형결합</u>**) of the original p variables in $\mathbf{x}$

$$\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p] \rightarrow \mathbf{z} = [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_p]$$

$$\mathbf{z}_1 = \mathbf{x}\boldsymbol{\alpha}_1 = \alpha_{11}\mathbf{x}_1 + \alpha_{12}\mathbf{x}_2 + \cdots + \alpha_{1p}\mathbf{x}_p$$

$$\mathbf{z}_2 = \mathbf{x}\boldsymbol{\alpha}_2 = \alpha_{21}\mathbf{x}_1 + \alpha_{22}\mathbf{x}_2 + \cdots + \alpha_{2p}\mathbf{x}_p$$

$$\vdots$$

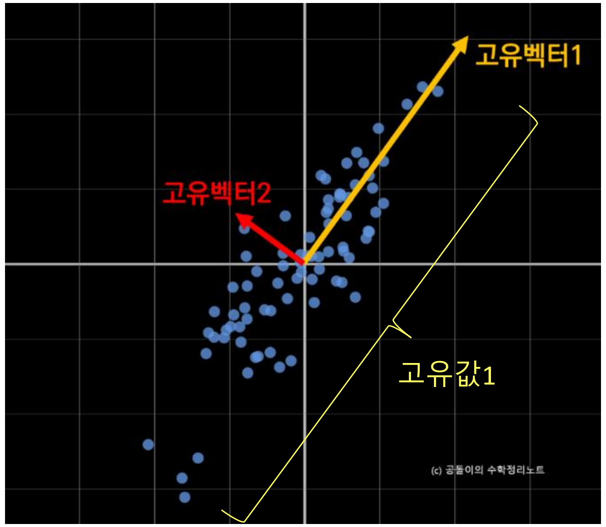
$$\mathbf{z}_p = \mathbf{x}\boldsymbol{\alpha}_p = \alpha_{p1}\mathbf{x}_1 + \alpha_{p2}\mathbf{x}_2 + \cdots + \alpha_{pp}\mathbf{x}_p$$

$\mathbf{x}$: original variables

$\boldsymbol{\alpha}_i$: i-th 기저(basis) 또는 계수(loading)

$\mathbf{z}$: 기저로 사영된 변환 후 변수 (주성분, Score)

① $\mathbf{z}$의 분산을 최대화할 수 있는 $\boldsymbol{\alpha}$ 찾기
→ $\mathbf{x}$ 의 **<u>공분산행렬의 고유벡터(eigenvector)</u>**

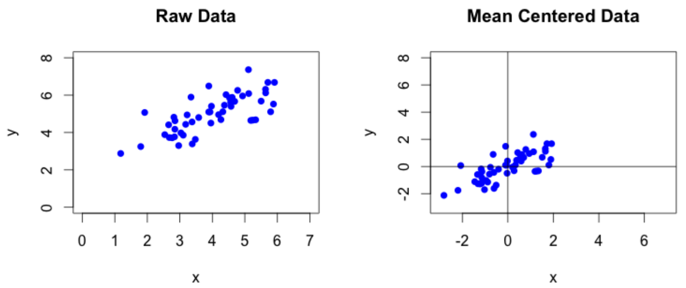


(https://angeloyeo.github.io/2019/07/27/PCA.html)

②고유값(eigenvalue)의 비율로 $\mathbf{z}$(주성분)의 갯수 결정 → **<u>차원축소</u>**

Process of PCA

- **Step 1: 기존 데이터 $\mathbf{x}$의 mean centering**



- **Step 2: Mean centered 데이터의 covariance matrix 계산**

$$Cov(\mathbf{x}) = \frac{1}{n}\mathbf{x}^T\mathbf{x}$$

- **Step 3: Covariance matrix로부터 eigenvalue를 구하고, eigenvalue 크기 순서대로 이에 해당되는 eigenvector를 정렬**

$$\lambda_1 = 2.7596 \qquad e_1^T = [0.5699, 0.5765, -0.5855]$$

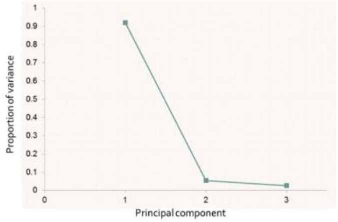
$$\lambda_2 = 0.1618 \qquad e_2^T = [0.7798, -0.6041, 0.1643]$$

- **Step 4: 정렬된 eigenvectors를 mean centered 데이터에 선형결합하여 $\mathbf{z}$로 변환**

$$z_1 = \mathbf{x}\mathbf{e}_1 = e_{11}x_1 + e_{12}x_2 + e_{13}x_3$$

- **Step 5: Eigenvalue 비율을 바탕으로 주성분($\mathbf{z}$) 갯수 결정**

$$\frac{\lambda_1}{\lambda_1+\lambda_2+\lambda_3} = 0.92\ (92\%)$$



Example of PCA

3개의 features

$\mathbf{X}$ =

5개의 관측치

$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$
0.20	5.6	3.56
0.45	5.89	2.40
0.33	6.37	1.95
0.54	7.9	1.32
0.77	7.87	0.98

Step1: Mean-centering

$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$
-1.19	-1.03	1.50
-0.04	-0.76	0.35
-0.59	-0.33	-0.09
0.38	1.07	-0.71
1.44	1.05	-1.05

Step 2: Covariance matrix

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

$$\begin{bmatrix} Var(\mathbf{x}_1) & Cov(\mathbf{x}_1,\mathbf{x}_2) & Cov(\mathbf{x}_1,\mathbf{x}_3) \\ Cov(\mathbf{x}_2,\mathbf{x}_1) & Var(\mathbf{x}_2) & Cov(\mathbf{x}_2,\mathbf{x}_3) \\ Cov(\mathbf{x}_3,\mathbf{x}_1) & Cov(\mathbf{x}_3,\mathbf{x}_2) & Var(\mathbf{x}_3) \end{bmatrix}$$

$$Cov(\mathbf{x}) = \frac{1}{n}\mathbf{x}^T\mathbf{x}$$

Step 3: Eigenvalue & eigenvector

$$\lambda_1 = 2.7596 \qquad e_1^T = [0.5699, 0.5765, -0.5855]$$

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

$$\lambda_3 = 0.0786 \qquad e_3^T = [0.2590, 0.5502, 0.7938]$$

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

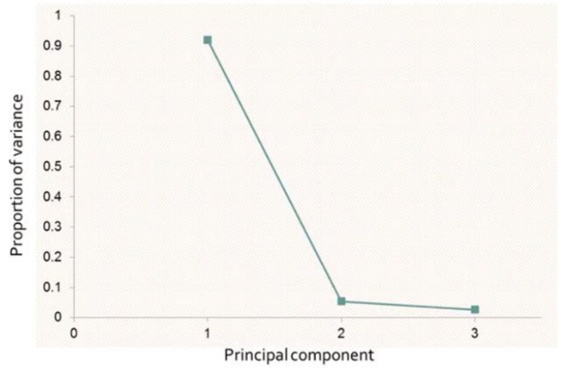
Step 4: z로 선형결합 변환

$$z_1 = \mathbf{xe}_1 = e_{11}x_1 + e_{12}x_2 + e_{13}x_3$$

$$= 0.5699\begin{bmatrix}-1.19\\-0.04\\-0.59\\0.38\\1.44\end{bmatrix} + 0.5765\begin{bmatrix}-1.03\\-0.76\\-0.33\\1.07\\1.05\end{bmatrix} - 0.5855\begin{bmatrix}1.50\\0.35\\-0.09\\-0.71\\-1.05\end{bmatrix} = \begin{bmatrix}-2.1527\\-0.6692\\-0.4718\\1.2533\\2.0404\end{bmatrix}$$

$$z_2 = \mathbf{xe}_2 = \begin{bmatrix}-0.0615\\0.4912\\-0.2798\\-0.4703\\0.3204\end{bmatrix} \qquad z_3 = \mathbf{xe}_3 = \begin{bmatrix}0.3160\\-0.1493\\-0.4047\\0.1223\\0.1157\end{bmatrix}$$

Step 5: 주성분 갯수 선택



$$\frac{\lambda_1}{\lambda_1+\lambda_2+\lambda_3} = 0.92\ (92\%)$$

SMART DESIGN LAB KAIST

What Questions Do You Have?

nwkang@kaist.ac.kr

www.smartdesignlab.org

SMART DESIGN LAB

KAIST