dimensional reduction ~ unsupervised learning

차원 축소

주 성분 분석

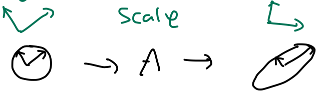
Inclass 18: Principal Component Analysis (PCA)

SVD

[SCS4049] Machine Learning and Data Science

vector → $\boxed{\text{Matrix}}$ → vector

scale



Seongsik Park (s.park@dgu.edu)

School of AI Convergence & Department of Artificial Intelligence, Dongguk University

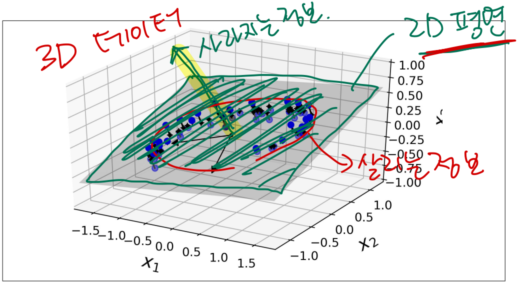


Figure 8-2. A 3D dataset lying close to a 2D subspace

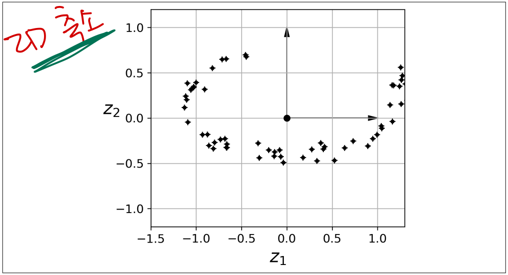


Figure 8-3. The new 2D dataset after projection

- 차원 축소

→ 데이터의 정보, 분포를 유지하는 방향

principal component.

데이터가 주로 분포한 축. 정보

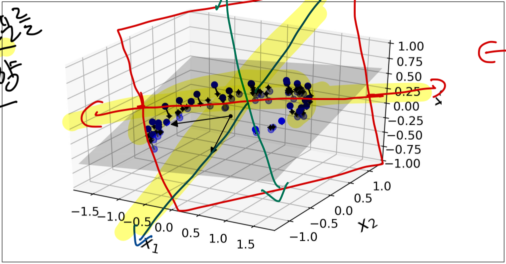
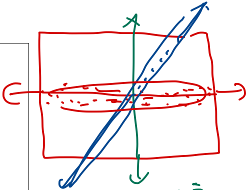


Figure 8-2. A 3D dataset lying close to a 2D subspace



의미없는 축.

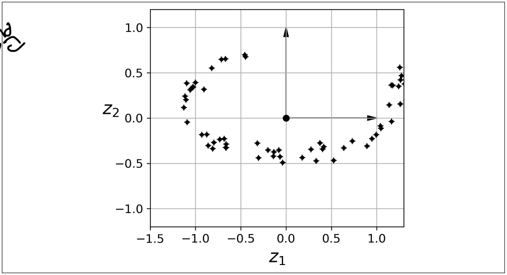


Figure 8-3. The new 2D dataset after projection

covariance → covariance matrix.

공 분산

$X-\mu_X$	X	Y	$Y-\mu_Y$
-2	-2	1	-1
-1	-1	2	0
0	0	4	2
1	1	2	0
2.	2	1	-1

$$\sigma_{XY} = E[(X-\mu_X)(Y-\mu_Y)]$$

Covariance matrix $\in$ vector, multidimension

$$X = [x_1, x_2, x_3, x_4]^T$$

$$C \in \mathbb{R}^{4\times 4}$$

$$C = \begin{bmatrix} \sigma_{x_1x_1} & \sigma_{x_1x_2} & \sigma_{x_1x_3} & \sigma_{x_1x_4} \\ \sigma_{x_2x_1} & \sigma_{x_2x_2} & \sigma_{x_2x_3} & \sigma_{x_2x_4} \\ & \vdots & & \end{bmatrix}^{4\times 4}$$

Covariance matrix

Covariance measures the strength of the linear relationship between two variables

$$\sigma_{xy} = \mathrm{E}[(x - \mu_x)(y - \mu_y)] \tag{1}$$

Covariance matrix C for multivariate random variable X

$$C_{ij} = \mathrm{E}[(x_i - \mu_i)(x_j - \mu_j)] \tag{2}$$

Preserving the variance

2D → 1D 압축, 축소.

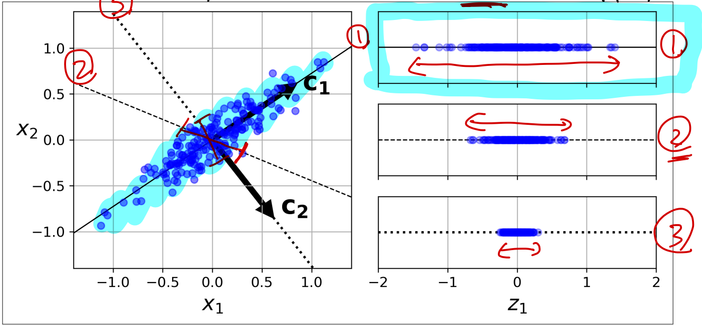


Figure 8-7. Selecting the subspace onto which to project

Principal component analysis (PCA)

For given data $x_1, x_2, \ldots, x_N \in \Re^D$ D차원, N sample

1. create a matrix $X \in \Re^{\cancel{D \times N}\ N \times D}$ with one column vector per each sample
2. covariance matrix $\cancel{\Sigma}\ C = \mathrm{E}\left[(X - \mathrm{E}(X))(X - \mathrm{E}(X))^T\right] \in \Re^{D \times D}$
3. find singular vectors and singular values of $\cancel{\Sigma}\ C$
4. principal components = largest singular values and vectors

covariance

$$C = U \Sigma V^T$$

$$= \begin{bmatrix} | & | & \\ u_1, & u_2 & \cdots \\ | & | & \end{bmatrix} \begin{bmatrix} \sigma_1 & & \\ & \sigma_2 & \\ & & \ddots \end{bmatrix} \begin{bmatrix} - v_1^T - \\ - v_2^T - \\ \vdots \end{bmatrix}$$

내림차순

input

$\mathbb{X} \in \mathbb{R}^{N \times D}$

$D \longrightarrow D'$ 축소. PCA.

$\mathbb{C} \in$ $\mathbb{X}$의 covariance

$\mathbb{C} = U \Sigma V^T$

$U = \begin{bmatrix} | & & | \\ u_1 & \cdots & u_r \\ | & & | \end{bmatrix}$ $\xrightarrow{\quad\quad} D'$

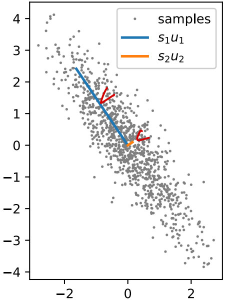
$U' = \begin{bmatrix} | & | & & | \\ u_1 & u_2 & \cdots & u_{D'} \\ | & | & & | \end{bmatrix} \in \mathbb{R}^{D \times D'}$

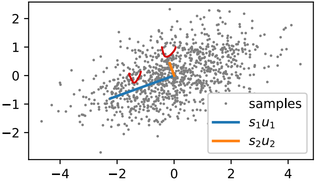
$\mathbb{Z} = \underset{N \times D}{\mathbb{X}} \; \underset{D \times D'}{U'} \in \mathbb{R}^{N \times D'}$

output

Principal component analysis (PCA)

① 2D random sample → SVD → PCA





② MNIST, 784 차원 → PCA → 축소 → classification

D9

DT