

Unsupervised Learning: Principle Component Analysis

Unsupervised Learning

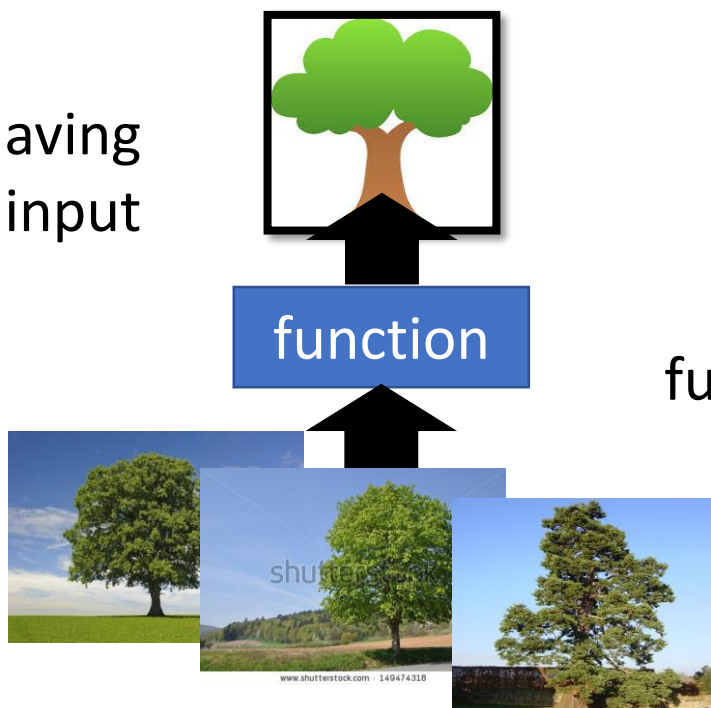
化繁為簡 < clustering

dimension reduction

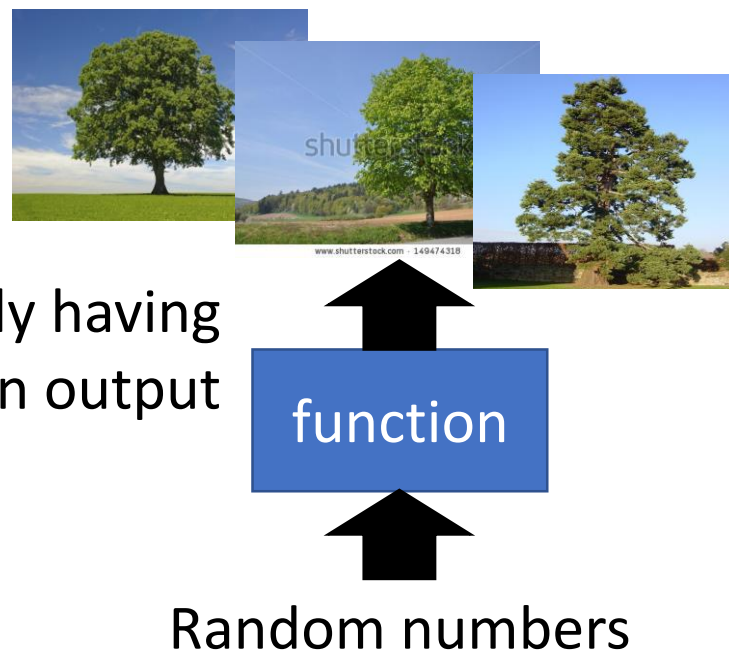
- Dimension Reduction (化繁為簡)

- Generation (無中生有)

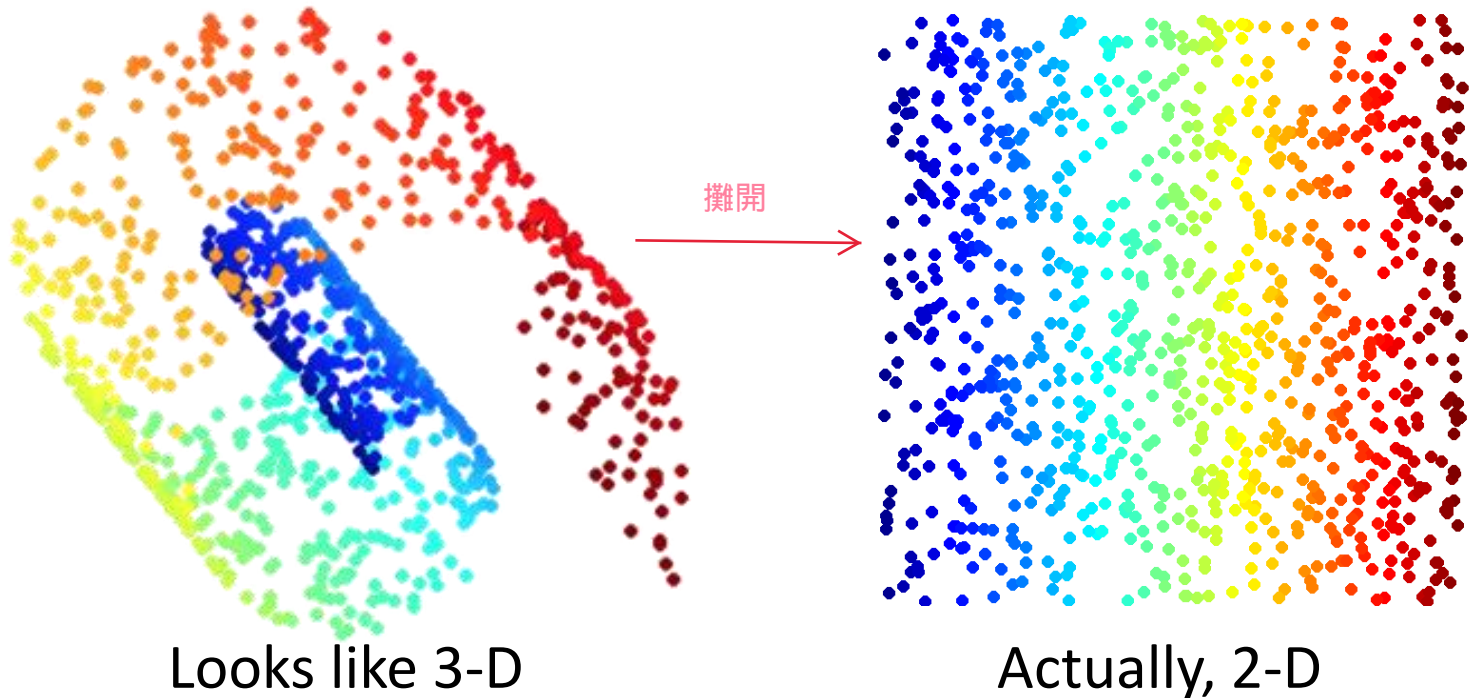
only having
function input



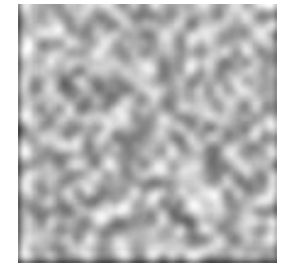
only having
function output



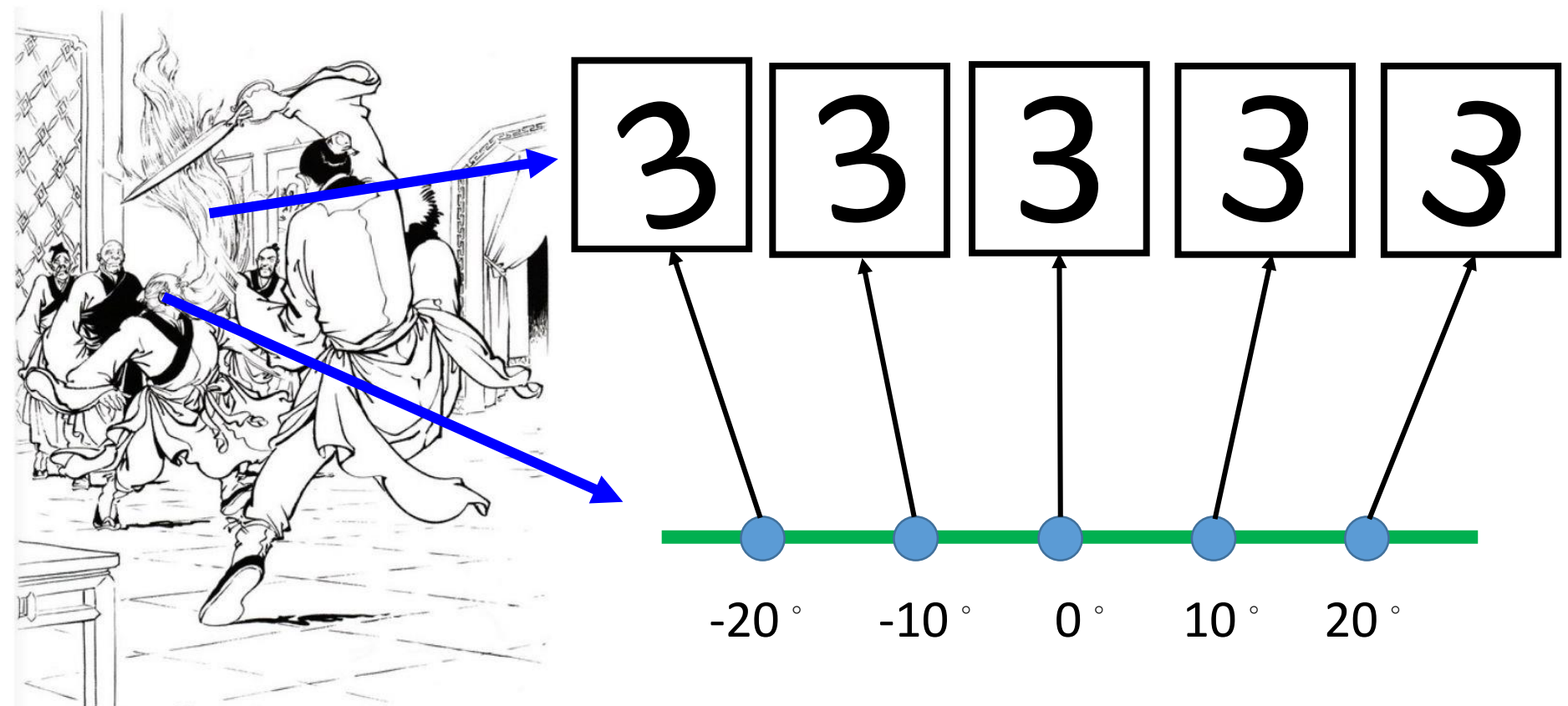
Dimension Reduction



Dimension Reduction



- In MNIST, a digit is 28 x 28 dims.
- Most 28 x 28 dim vectors are not digits



Clustering

$$\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

Cluster 1



Open question: how many clusters do we need?



Cluster 3

$$\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

$$\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

Cluster 2



- K-means

unlabeled data

- Clustering $X = \{x^1, \dots, x^n, \dots, x^N\}$ into K clusters
- Initialize cluster center c^i , $i=1,2, \dots K$ (K random x^n from X)

- Repeat

- For all x^n in X :
$$b_i^n = \begin{cases} 1 & x^n \text{ is most "close" to } c^i \\ 0 & \text{Otherwise} \end{cases}$$

- Updating all c^i :
$$c^i = \sum_{x^n} b_i^n x^n / \sum_{x^n} b_i^n$$

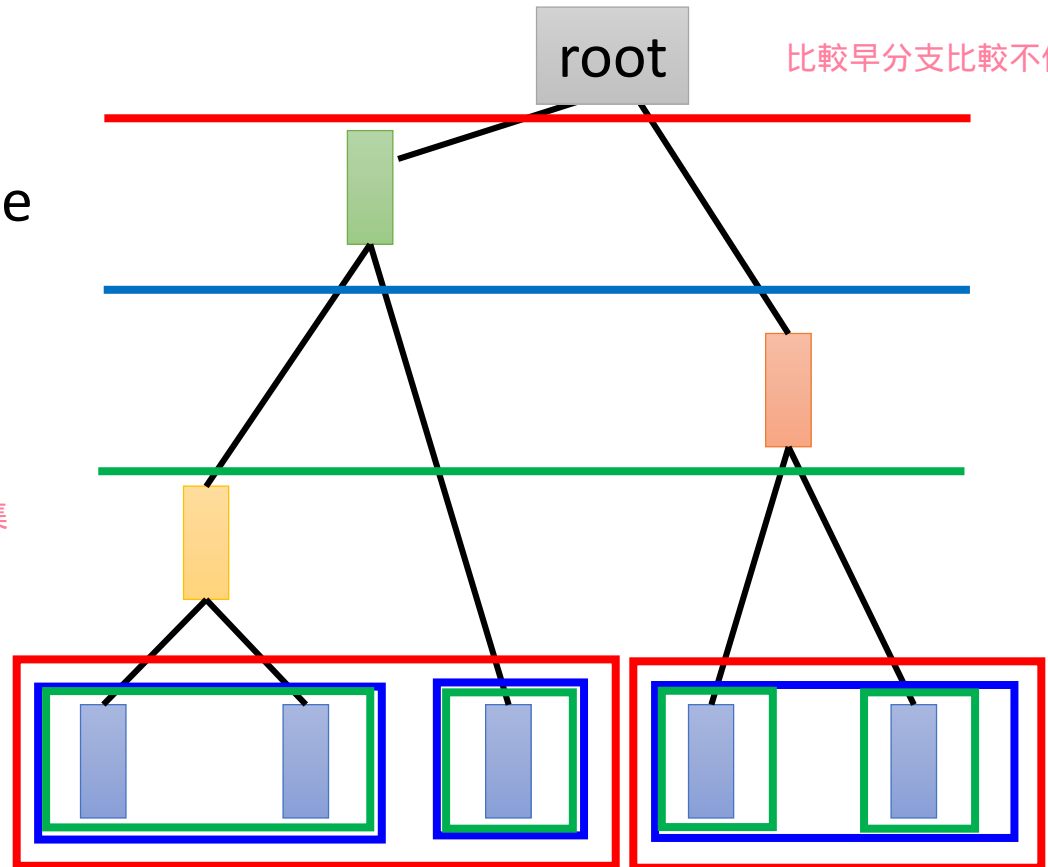
Clustering

- Hierarchical Agglomerative Clustering (HAC)

Step 1: build a tree

Step 2: pick a threshold

切在哪條線，會決定有幾個群集



Distributed Representation

太以偏概全

- Clustering: an object must belong to one cluster

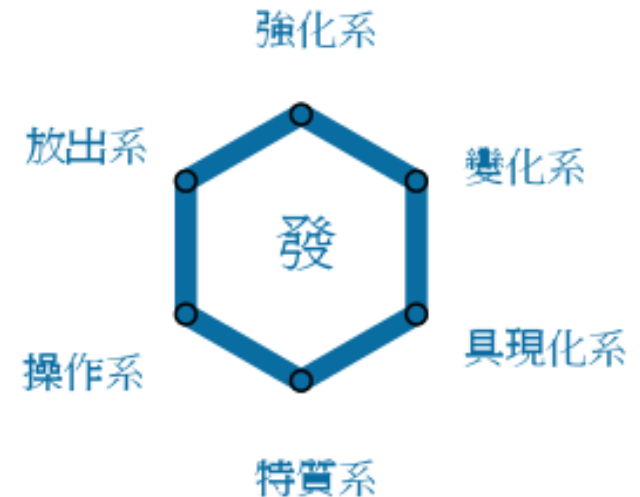
小傑是強化系

- Distributed representation

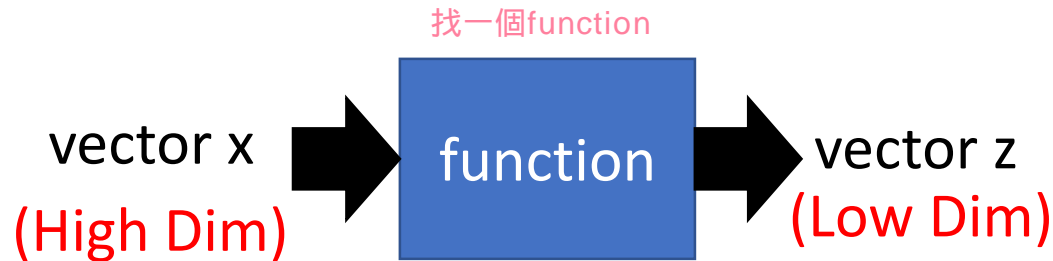
用vector表示

小傑是

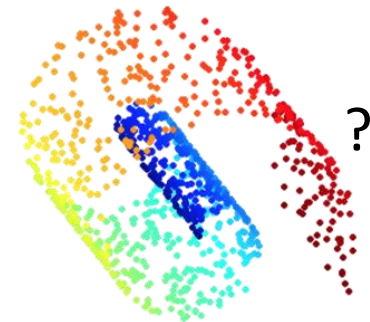
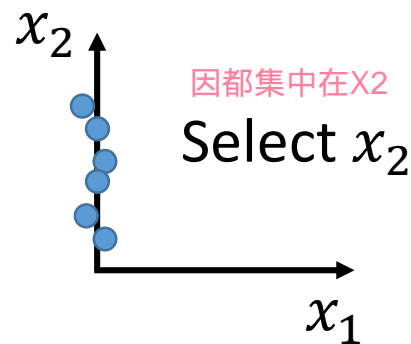
強化系	0.70
放出系	0.25
變化系	0.05
操作系	0.00
具現化系	0.00
特質系	0.00



Distributed Representation



- Feature selection



- Principle component analysis (PCA)
[Bishop, Chapter 12]

liner function

$$z = Wx$$

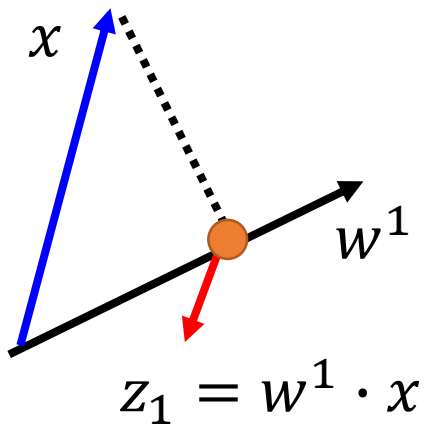
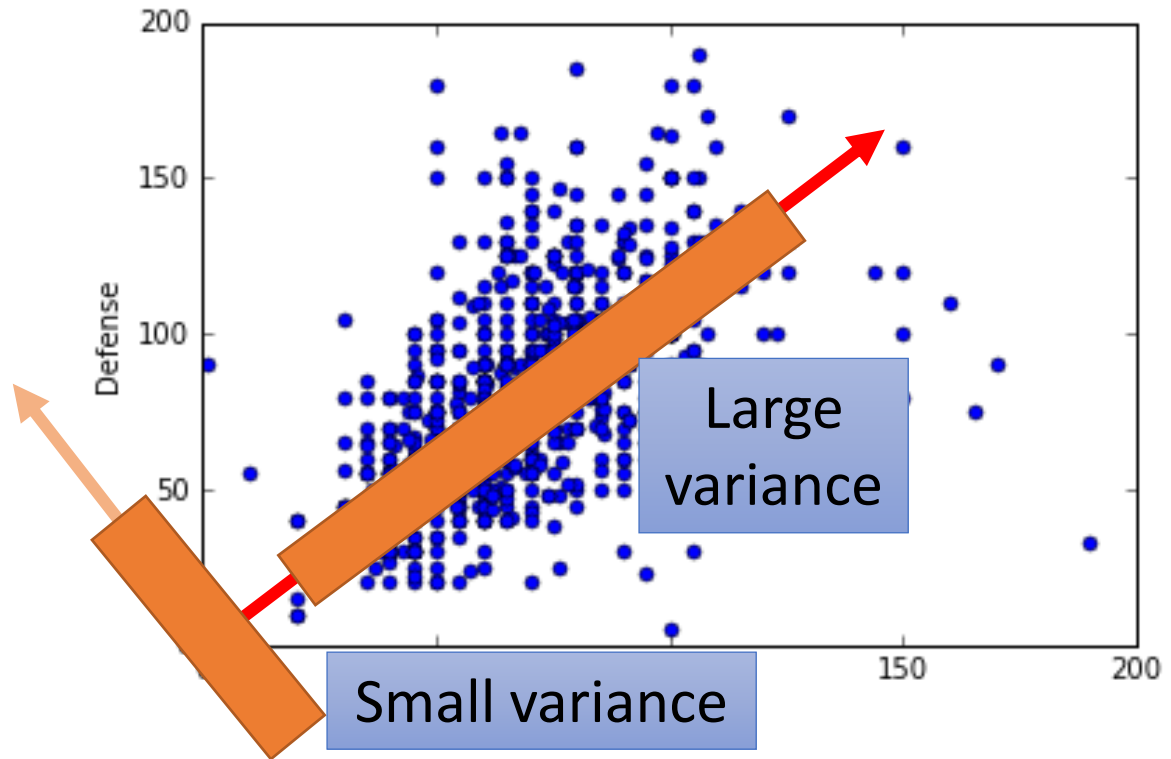
PCA

找W

$$z = Wx$$

Reduce to 1-D:

$$z_1 = w^1 \cdot x$$



Project all the data points x onto w^1 ,
and obtain a set of z_1

We want the variance of z_1 as large as
possible

$$Var(z_1) = \frac{1}{N} \sum_{z_1} (z_1 - \bar{z}_1)^2 \quad \|w^1\|_2 = 1$$

PCA

$$z = Wx$$

Reduce to 1-D:

$$z_1 = w^1 \cdot x$$

$$z_2 = w^2 \cdot x$$

$$W = \begin{bmatrix} (w^1)^T \\ (w^2)^T \\ \vdots \end{bmatrix}$$

Orthogonal
matrix

Project all the data points x onto w^1 ,
and obtain a set of z_1

We want the variance of z_1 as large as
possible

$$Var(z_1) = \frac{1}{N} \sum_{z_1} (z_1 - \bar{z}_1)^2 \quad \|w^1\|_2 = 1$$

We want the variance of z_2 as large as
possible

$$Var(z_2) = \frac{1}{N} \sum_{z_2} (z_2 - \bar{z}_2)^2 \quad \|w^2\|_2 = 1$$
$$w^1 \cdot w^2 = 0$$

Warning of Math

PCA

$$z_1 = w^1 \cdot x$$

$$\bar{z}_1 = \frac{1}{N} \sum z_1 = \frac{1}{N} \sum w^1 \cdot x = w^1 \cdot \frac{1}{N} \sum x = w^1 \cdot \bar{x}$$

$$Var(z_1) = \frac{1}{N} \sum_{z_1} (z_1 - \bar{z}_1)^2$$

$$(a \cdot b)^2 = (a^T b)^2 = a^T b a^T b$$

$$= a^T b (a^T b)^T = a^T b b^T a$$

$$= \frac{1}{N} \sum_x (w^1 \cdot x - w^1 \cdot \bar{x})^2$$

$$= \frac{1}{N} \sum (w^1 \cdot (x - \bar{x}))^2$$

$$= \frac{1}{N} \sum (w^1)^T (x - \bar{x})(x - \bar{x})^T w^1$$

$$= (w^1)^T \left[\frac{1}{N} \sum (x - \bar{x})(x - \bar{x})^T \right] w^1$$

$$= (w^1)^T Cov(x) w^1$$

$$S = Cov(x)$$

Find w^1 maximizing

$$(w^1)^T S w^1$$

$$\|w^1\|_2 = (w^1)^T w^1 = 1$$

Find w^1 maximizing $(w^1)^T S w^1$ $(w^1)^T w^1 = 1$

$S = \text{Cov}(x)$ Symmetric Positive-semidefinite 半正定
(non-negative eigenvalues)

Using Lagrange multiplier [Bishop, Appendix E]

$$g(w^1) = (w^1)^T S w^1 - \alpha((w^1)^T w^1 - 1)$$

$$\left. \begin{array}{l} \partial g(w^1) / \partial w_1^1 = 0 \\ \partial g(w^1) / \partial w_2^1 = 0 \\ \vdots \end{array} \right\} \begin{array}{l} S w^1 - \alpha w^1 = 0 \\ S w^1 = \alpha w^1 \quad w^1 : \text{eigenvector} \\ (w^1)^T S w^1 = \alpha (w^1)^T w^1 \\ = \alpha \quad \text{Choose the maximum one} \end{array}$$

w^1 is the eigenvector of the covariance matrix S

Corresponding to the largest eigenvalue λ_1

Find w^2 maximizing $(w^2)^T S w^2$ $(w^2)^T w^2 = 1$ $(w^2)^T w^1 = 0$

$$g(w^2) = (w^2)^T S w^2 - \alpha((w^2)^T w^2 - 1) - \beta((w^2)^T w^1 - 0)$$

$$\left. \begin{array}{l} \partial g(w^2) / \partial w_1^2 = 0 \\ \partial g(w^2) / \partial w_2^2 = 0 \\ \vdots \end{array} \right\} \begin{array}{l} S w^2 - \alpha w^2 - \beta w^1 = 0 \\ \underline{0} - \alpha \underline{0} - \beta \underline{1} = 0 \\ = ((w^1)^T S w^2)^T = (w^2)^T S^T w^1 \\ = (w^2)^T S w^1 = \lambda_1 (w^2)^T w^1 = 0 \end{array}$$

$$S w^1 = \lambda_1 w^1$$

$$\beta = 0: \quad S w^2 - \alpha w^2 = 0 \quad S w^2 = \alpha w^2$$

w^2 is the eigenvector of the covariance matrix S

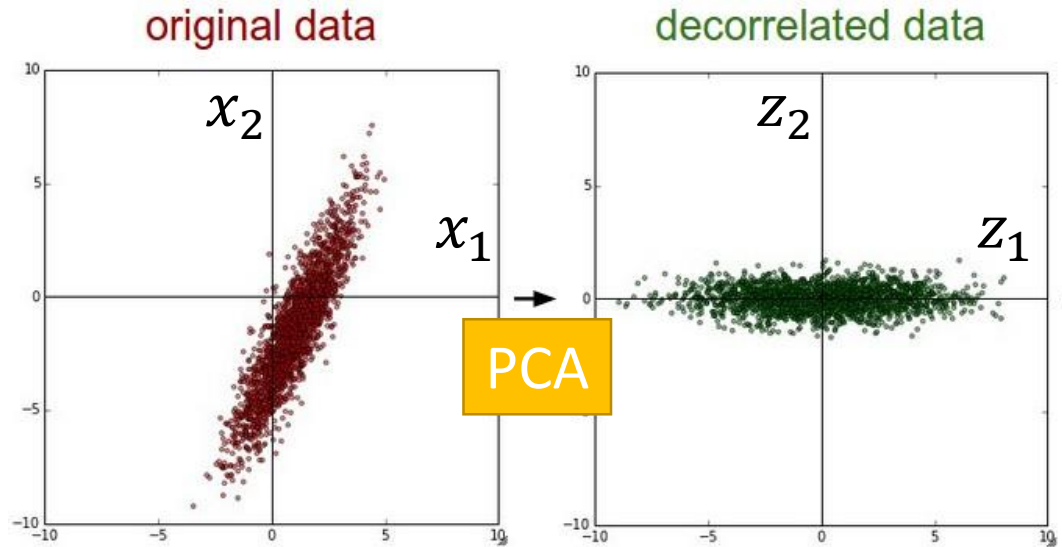
Corresponding to the 2nd largest eigenvalue λ_2

PCA - decorrelation

$$z = Wx$$

$$\text{Cov}(z) = D$$

Diagonal matrix



$$\text{Cov}(z) = \frac{1}{N} \sum (z - \bar{z})(z - \bar{z})^T = W S W^T \quad S = \text{Cov}(x)$$

$$= W S [w^1 \quad \dots \quad w^K] = W [S w^1 \quad \dots \quad S w^K]$$

$$= W [\lambda_1 w^1 \quad \dots \quad \lambda_K w^K] = [\lambda_1 W w^1 \quad \dots \quad \lambda_K W w^K]$$

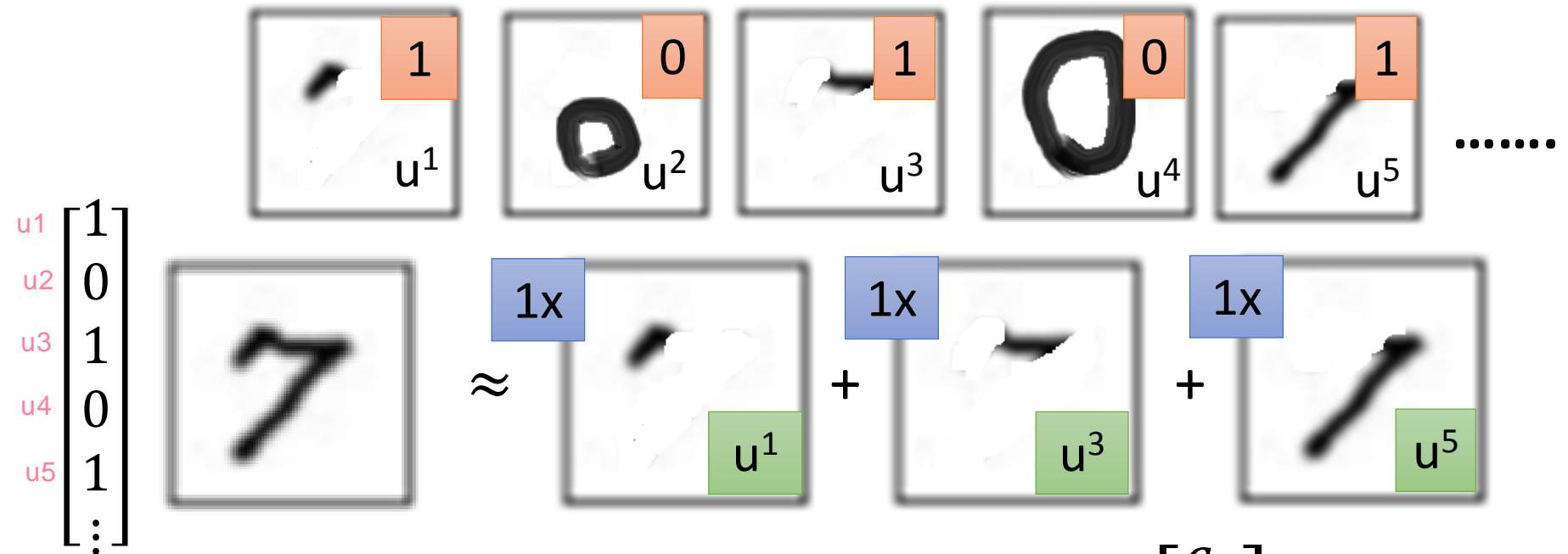
$$= [\lambda_1 e_1 \quad \dots \quad \lambda_K e_K] = D$$

Diagonal matrix

End of Warning

PCA – Another Point of View

Basic Component:



$$x \approx c_1 u^1 + c_2 u^2 + \dots + c_K u^K + \bar{x}$$


Pixels in a
digit image

component

$$\begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_K \end{bmatrix}$$

Represent a
digit image

PCA – Another Point of View

$$x - \bar{x} \approx c_1 u^1 + c_2 u^2 + \dots + c_K u^K = \hat{x}$$


Reconstruction error:

$$\| (x - \bar{x}) - \hat{x} \|_2$$

Find $\{u^1, \dots, u^K\}$ minimizing the error


$$L = \min_{\{u^1, \dots, u^K\}} \sum \left\| (x - \bar{x}) - \underbrace{\left(\sum_{k=1}^K c_k u^k \right)}_{\hat{x}} \right\|_2$$

PCA: $z = Wx$

$$\begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_K \end{bmatrix} = \begin{bmatrix} (w_1)^T \\ (w_2)^T \\ \vdots \\ (w_K)^T \end{bmatrix} x$$

$\{w^1, w^2, \dots, w^K\}$ (from PCA) is the component $\{u^1, u^2, \dots, u^K\}$ minimizing L

Proof in [Bishop, Chapter 12.1.2]

$$x - \bar{x} \approx c_1 u^1 + c_2 u^2 + \dots + c_K u^K = \hat{x}$$


Reconstruction error:

$$\| (x - \bar{x}) - \hat{x} \|_2$$

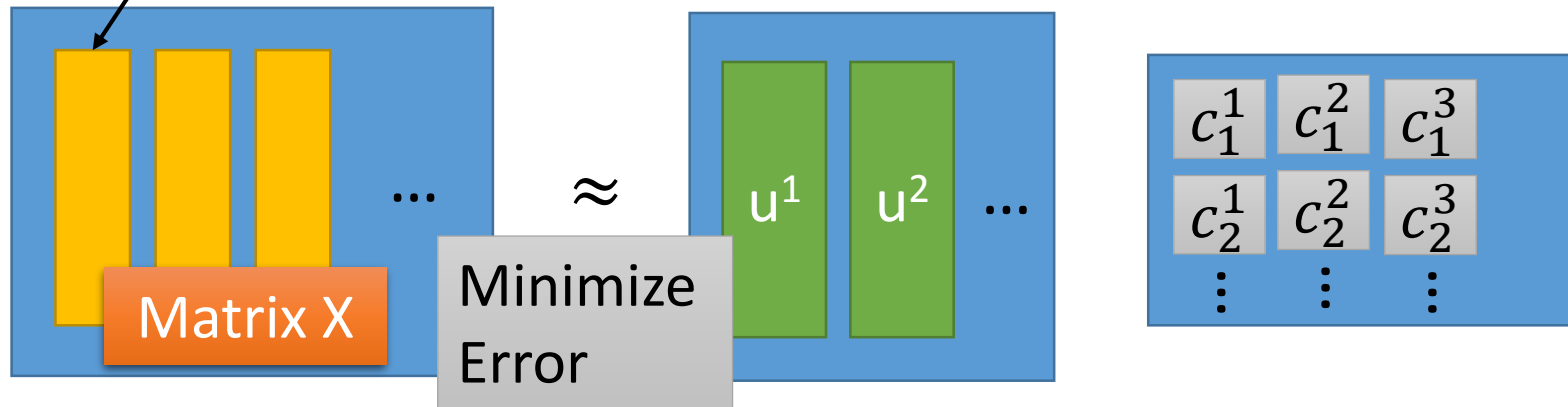
Find $\{u^1, \dots, u^K\}$ minimizing the error

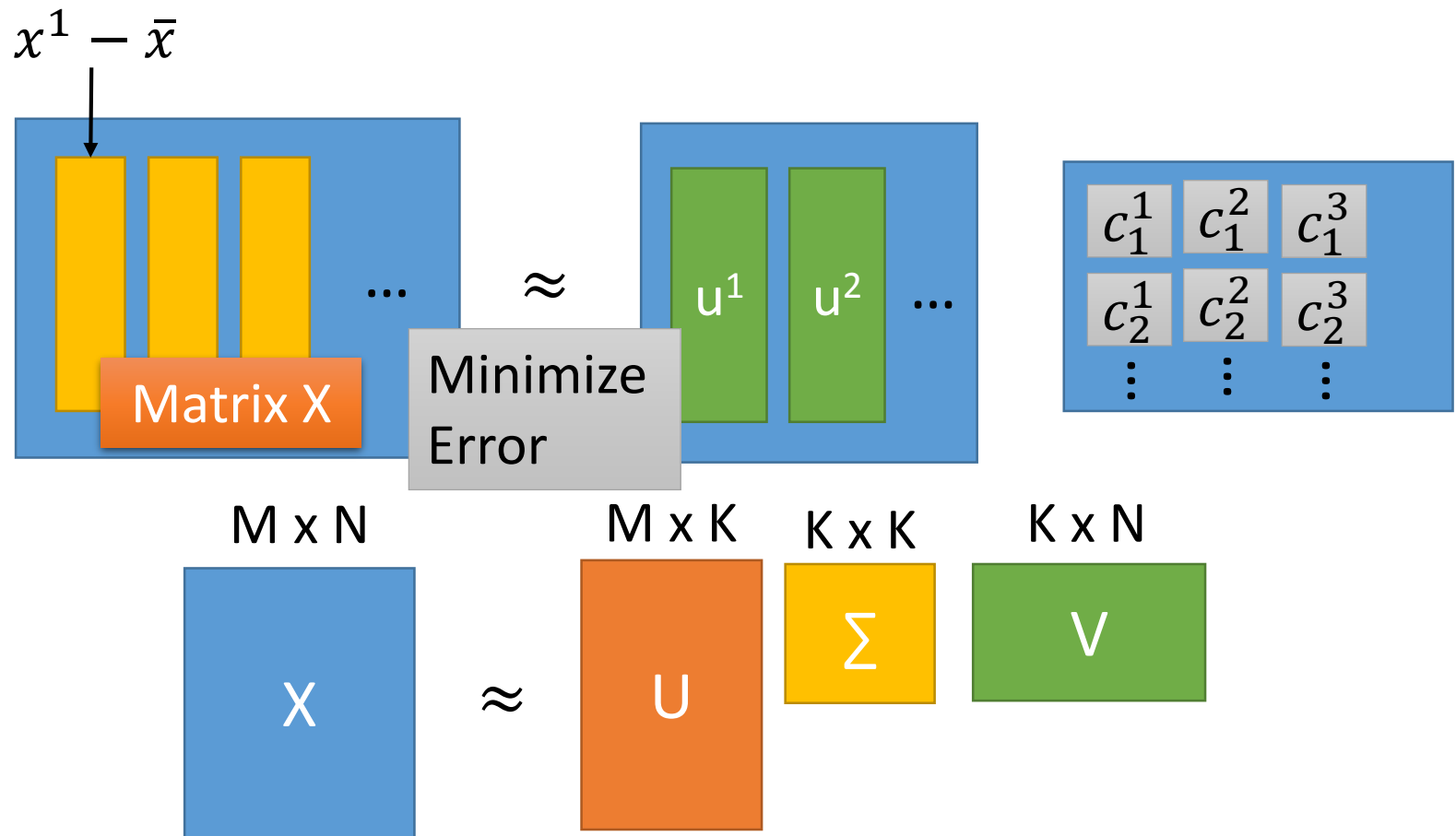
$$\underline{x^1 - \bar{x}} \approx \underline{c_1^1} \underline{u^1} + \underline{c_2^1} \underline{u^2} + \dots$$

$$x^2 - \bar{x} \approx c_1^2 u^1 + c_2^2 u^2 + \dots$$

$$x^3 - \bar{x} \approx c_1^3 u^1 + c_2^3 u^2 + \dots$$

⋮





K columns of U : a set of orthonormal eigen vectors corresponding to the K largest eigenvalues of XX^T

This is the solution of PCA

SVD:

http://speech.ee.ntu.edu.tw/~tlkagk/courses/LA_2016/Lecture/SVD.pdf

PCA looks like a neural network with one hidden layer (linear activation function)

Autoencoder

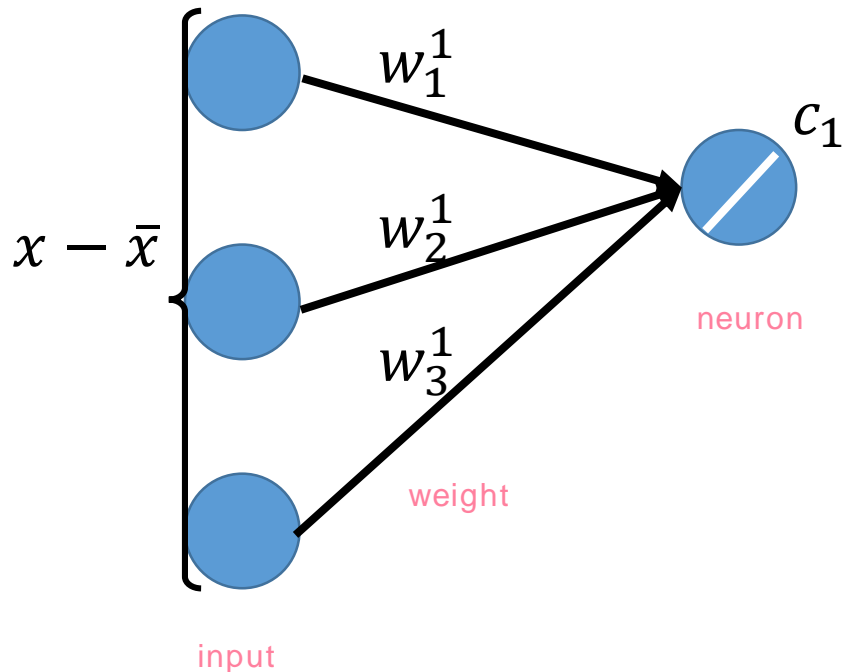
If $\{w^1, w^2, \dots w^K\}$ is the component $\{u^1, u^2, \dots u^K\}$

$$\hat{x} = \sum_{k=1}^K c_k w^k \iff x - \bar{x}$$

To minimize reconstruction error:

$$c_k = (x - \bar{x}) \cdot w^k$$

$K = 2$:



PCA looks like a neural network with one hidden layer (linear activation function)

Autoencoder

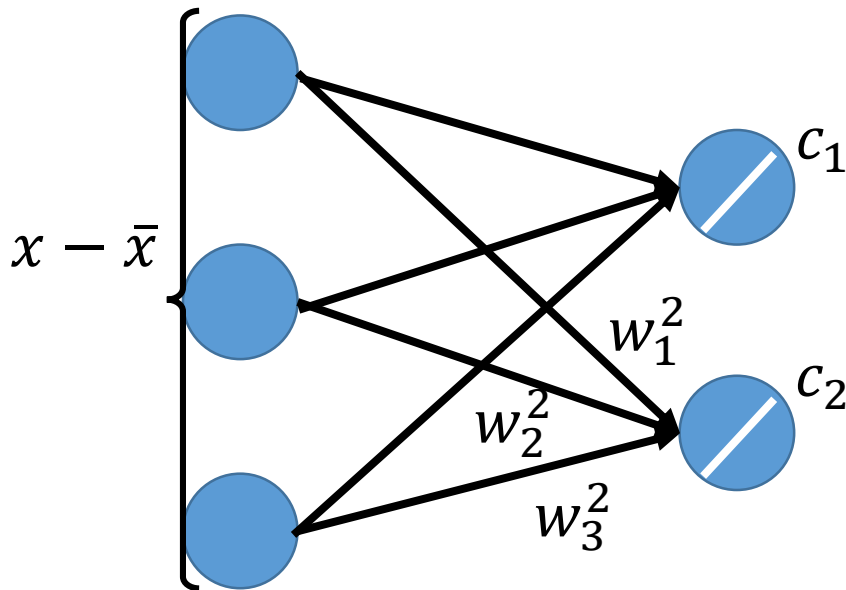
If $\{w^1, w^2, \dots, w^K\}$ is the component $\{u^1, u^2, \dots, u^K\}$

$$\hat{x} = \sum_{k=1}^K c_k w^k \longleftrightarrow x - \bar{x}$$

To minimize reconstruction error:

$$c_k = (x - \bar{x}) \cdot w^k$$

$K = 2$:



PCA looks like a neural network with one hidden layer (linear activation function)

Autoencoder

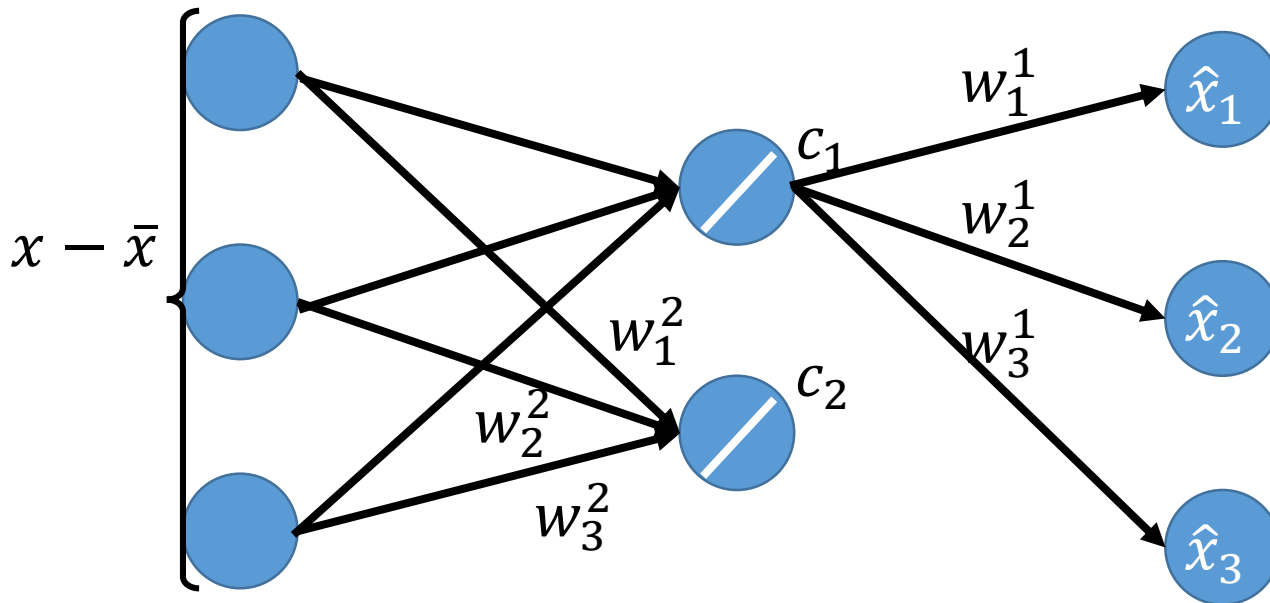
If $\{w^1, w^2, \dots, w^K\}$ is the component $\{u^1, u^2, \dots, u^K\}$

$$\hat{x} = \sum_{k=1}^K c_k w^k \iff x - \bar{x}$$

To minimize reconstruction error:

$$c_k = (x - \bar{x}) \cdot w^k$$

$K = 2$:



PCA looks like a neural network with one hidden layer (linear activation function)

Autoencoder

If $\{w^1, w^2, \dots, w^K\}$ is the component $\{u^1, u^2, \dots, u^K\}$

$$\hat{x} = \sum_{k=1}^K c_k w^k \longleftrightarrow x - \bar{x}$$

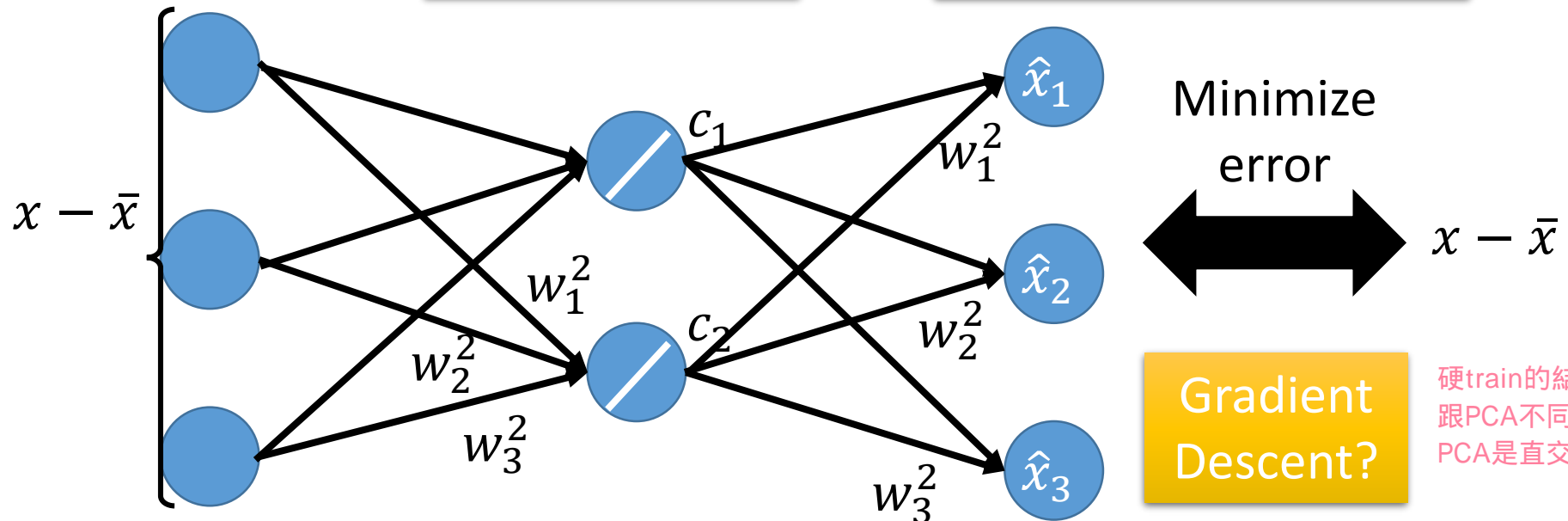
To minimize reconstruction error:

$$c_k = (x - \bar{x}) \cdot w^k$$

$K = 2$:

It can be deep.

Deep Autoencoder



PCA - Pokémon

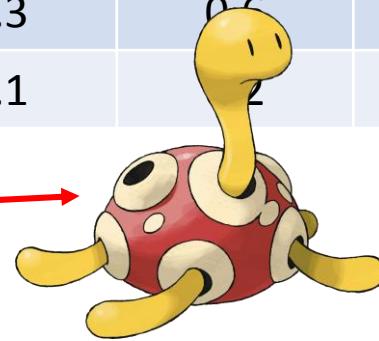
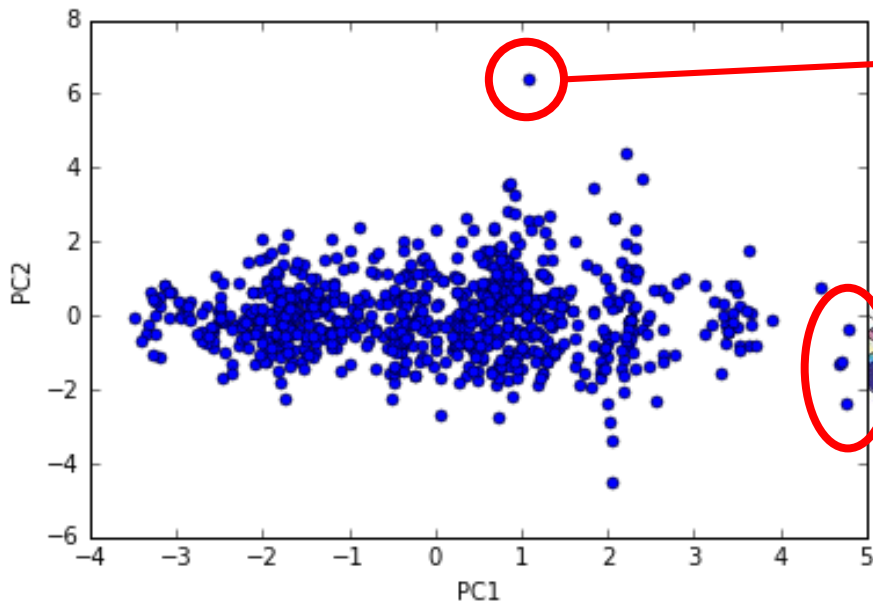
- Inspired from:
<https://www.kaggle.com/strakul5/d/abcsds/pokemon/principal-component-analysis-of-pokemon-data>
- 800 Pokemons, 6 features for each (HP, Atk, Def, Sp Atk, Sp Def, Speed)
- How many principle components? $\frac{\lambda_i}{\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 + \lambda_5 + \lambda_6}$

	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6
ratio	0.45	0.18	0.13	0.12	0.07	0.04

Using 4 components is good enough

PCA - Pokémon

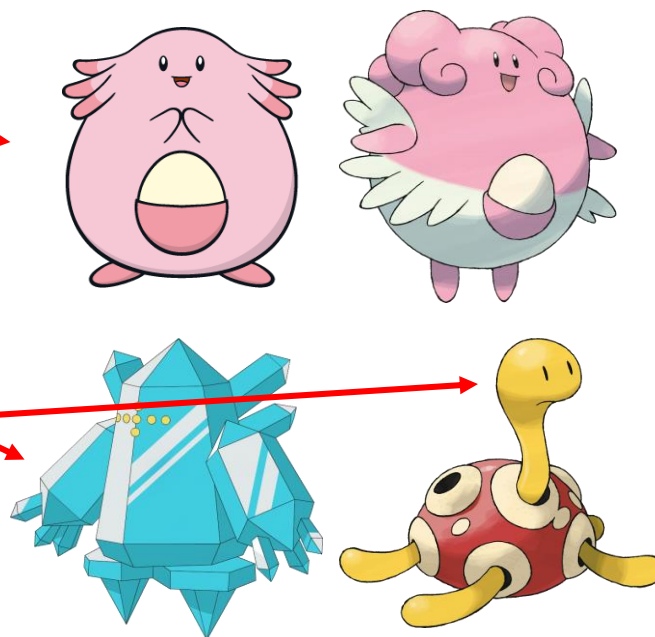
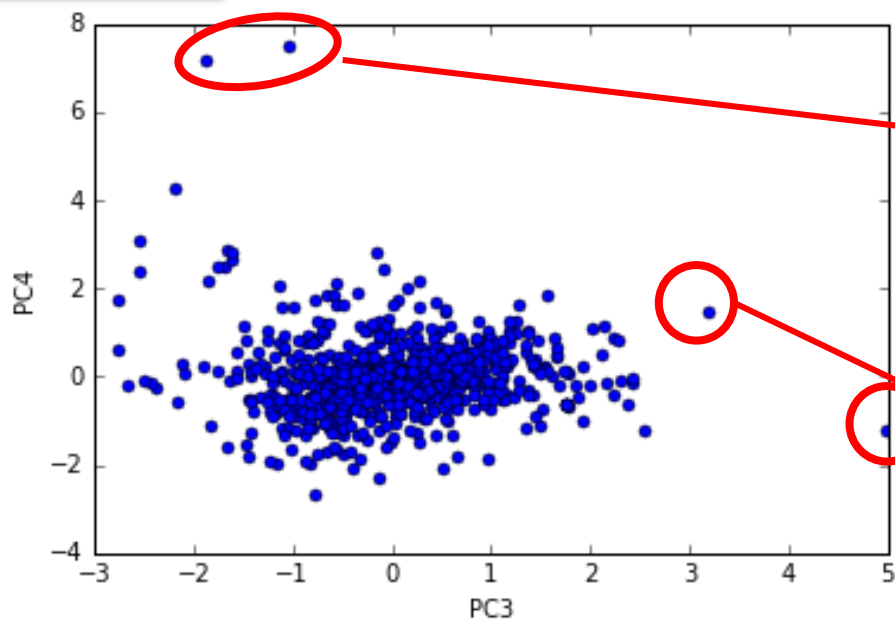
	HP	Atk	Def	Sp Atk	Sp Def	Speed	
PC1	0.4	0.4	0.4	0.5	0.4	0.3	強度
PC2	0.1	0.0	0.6	-0.3	0.2	-0.7	
PC3	-0.5	-0.6	0.1	0.3	0.6		防禦(犠牲速度)
PC4	0.7	-0.4	-0.4	0.1	0.2	-0.3	



PCA - Pokémon

	HP	Atk	Def	Sp Atk	Sp Def	Speed
PC1	0.4	0.4	0.4	0.5	0.4	0.3
PC2	0.1	0.0	0.6	-0.3	0.2	-0.7
PC3	-0.5	-0.6	0.1	0.3	0.6	
生命力強	0.7	-0.4	-0.4	0.1	0.2	

特殊防禦(犧牲
攻擊和生命)



PCA - Pokémon

- <http://140.112.21.35:2880/~tlkagk/pokemon/pca.html>
- The code is modified from
 - <http://jkunst.com/r/pokemon-visualize-em-all/>

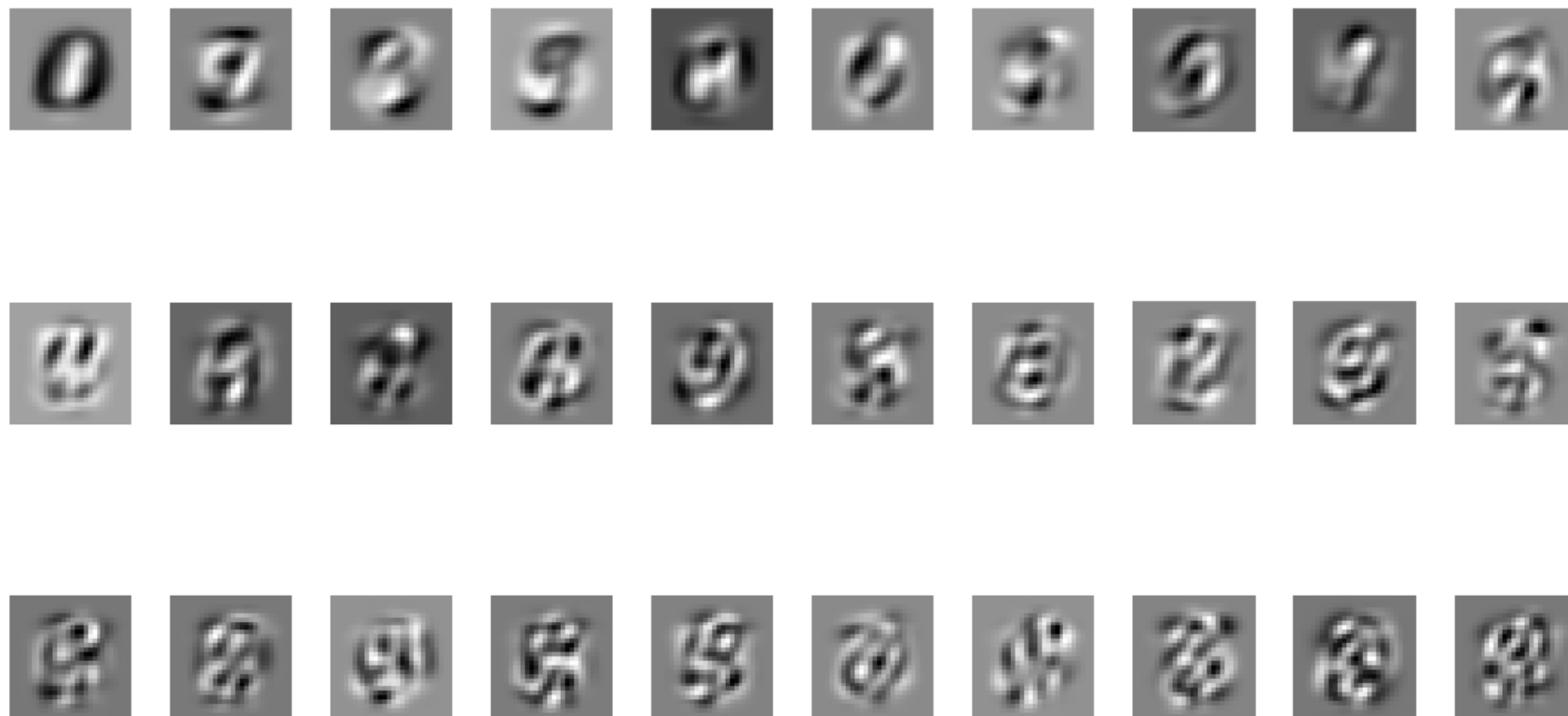
PCA - MNIST



$$= a_1 w^1 + a_2 w^2 + \dots$$

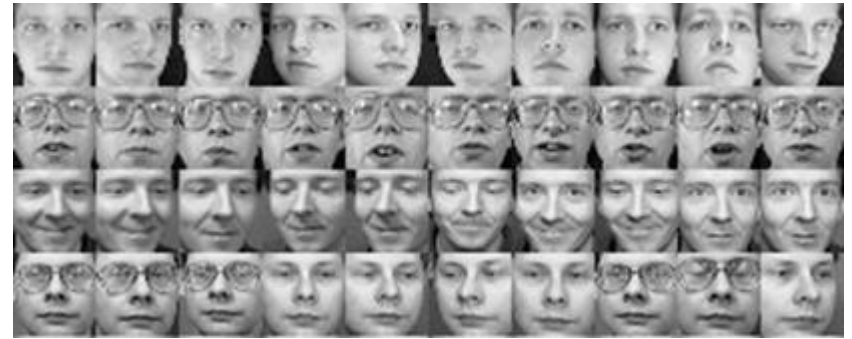
images

30 components:



Eigen-digits

PCA - Face



30 components:

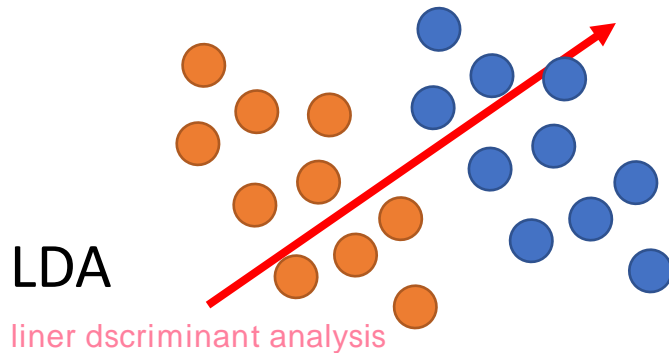
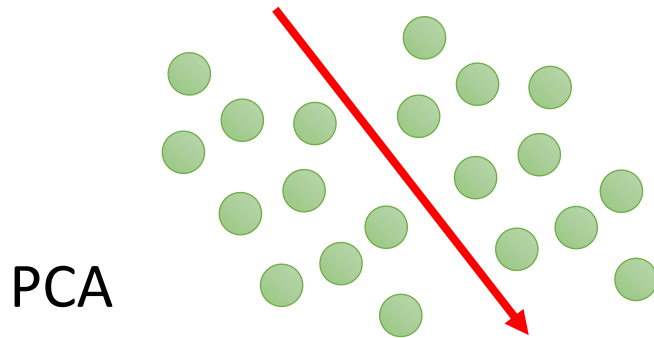


<http://www.cs.unc.edu/~lazebnik/research/spring08/assignment3.html>

Eigen-face

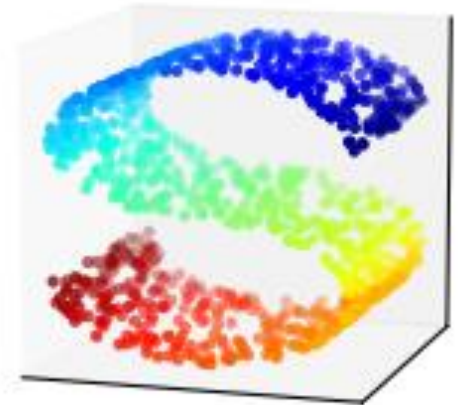
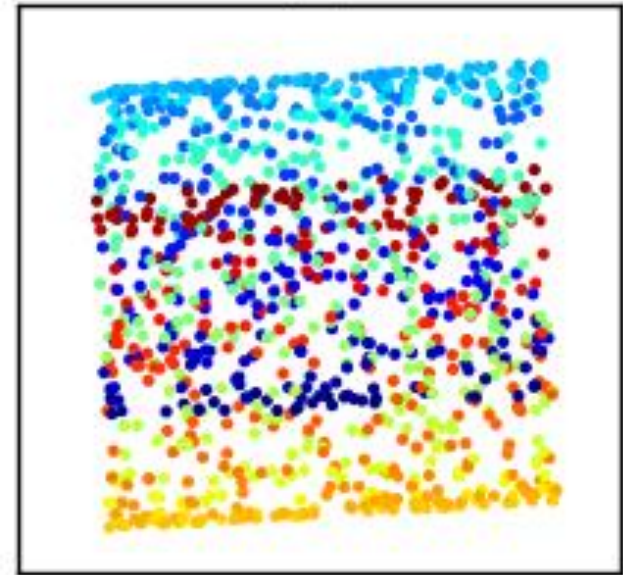
Weakness of PCA

- Unsupervised



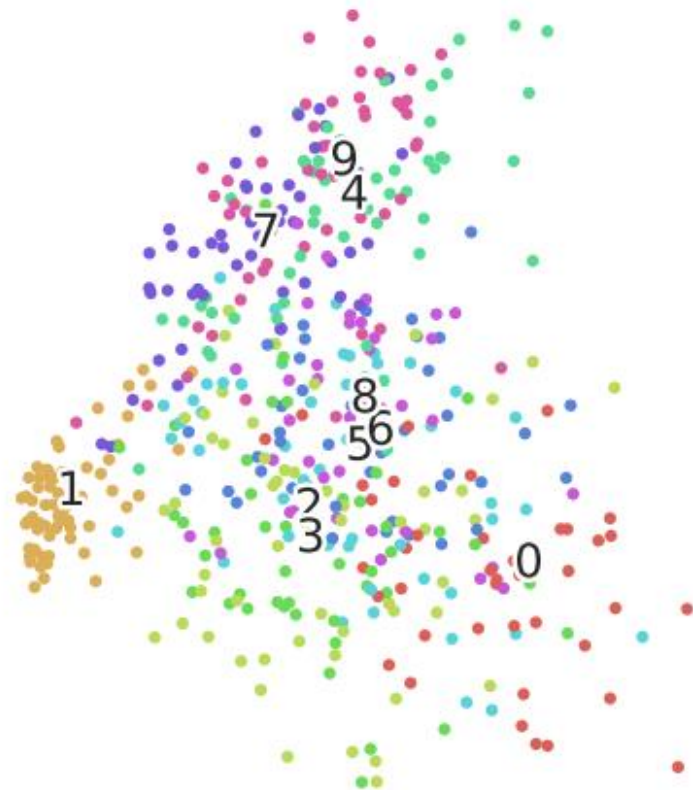
- Linear

無法拉直(非線性轉換)只能打扁

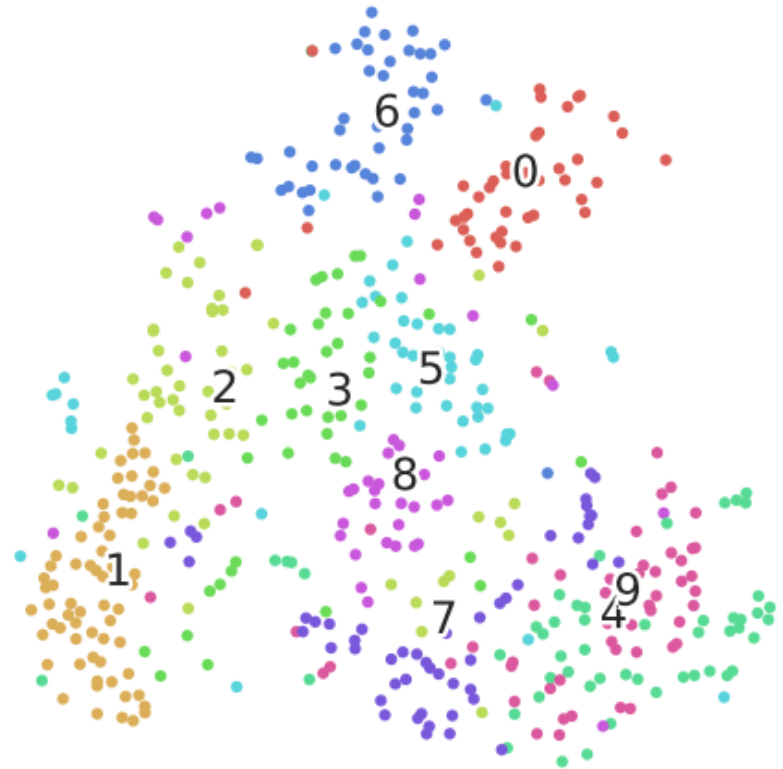


http://www.astroml.org/book_figures/chapter7/fig_S_manifold_PCA.html

Weakness of PCA



Pixel (28x28) -> PCA (2)



Pixel (28x28) -> tSNE (2)

Acknowledgement

- 感謝 彭冲 同學發現引用資料的錯誤
- 感謝 Hsiang-Chih Cheng 同學發現投影片上的錯誤

Appendix

- http://4.bp.blogspot.com/_sHcZHRnXLLE/S9EpFXYjfvI/AAAAAAAAABZ0/_oEQiaR3WVM/s640/dimensionality+reduction.jpg
- https://lvdmaaten.github.io/publications/papers/TR_Dimensionality_Reduction_Review_2009.pdf

