Dimension Reduction

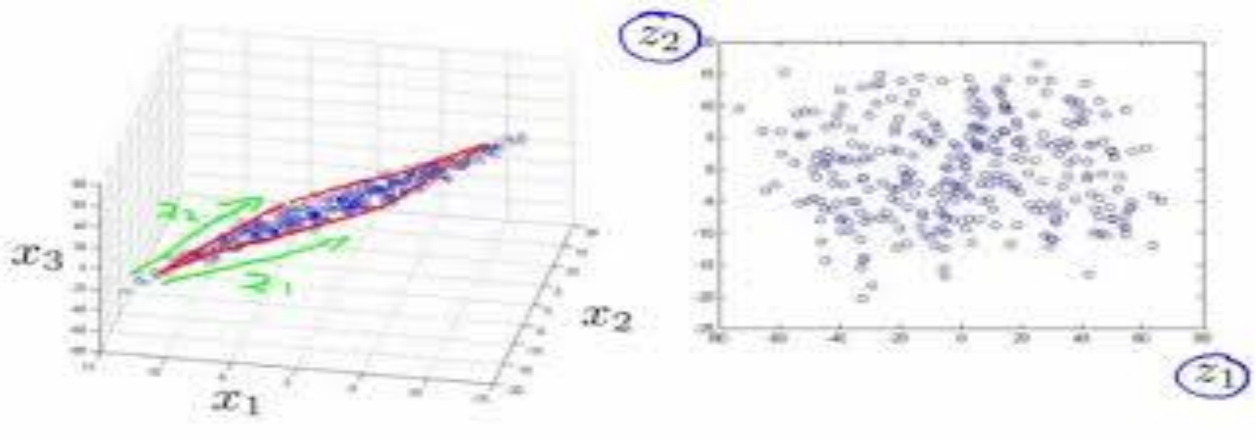
講者：Isaac

Outline

- SVD
- PCA
- T-SNE

What's Dimension Reduction

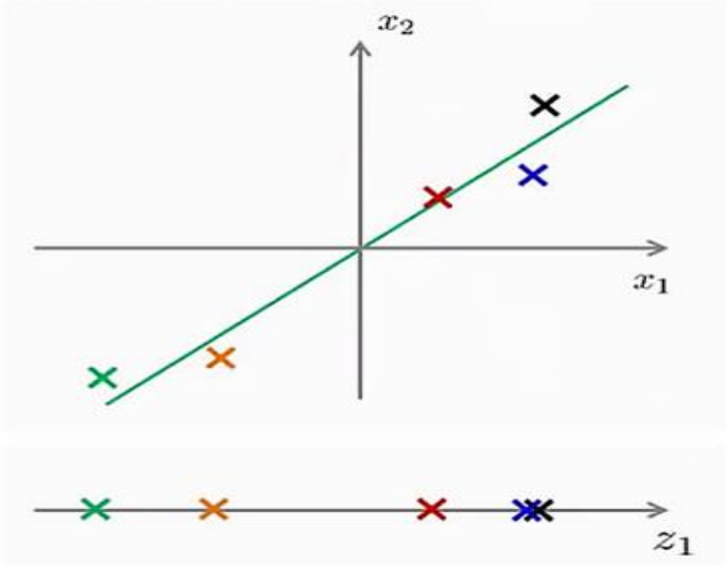
- Dimension Reduction is an unsupervised learning
- Can use to data compression
  - Make learning faster(Use less RAM)
- Help visualize data



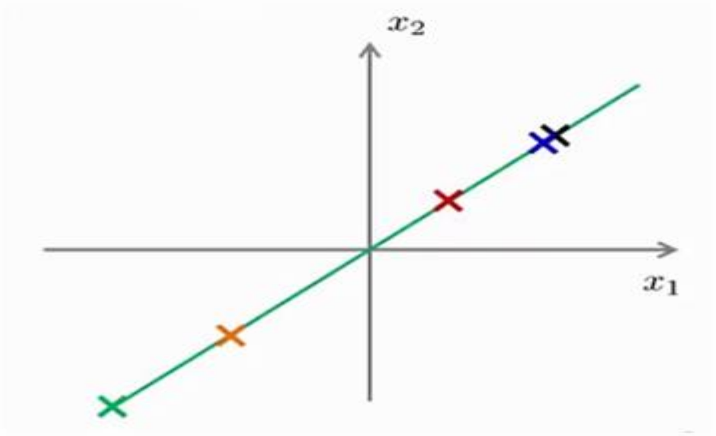
Why Reduce Dimensionality?

- Reduces time complexity
  - Less computation
- Reduces space complexity
  - Less parameters
- Saves the cost of observing the features
- Data visualization if plotted in 2 or 3 dimensions

Reduce Dimensionality Illustraion

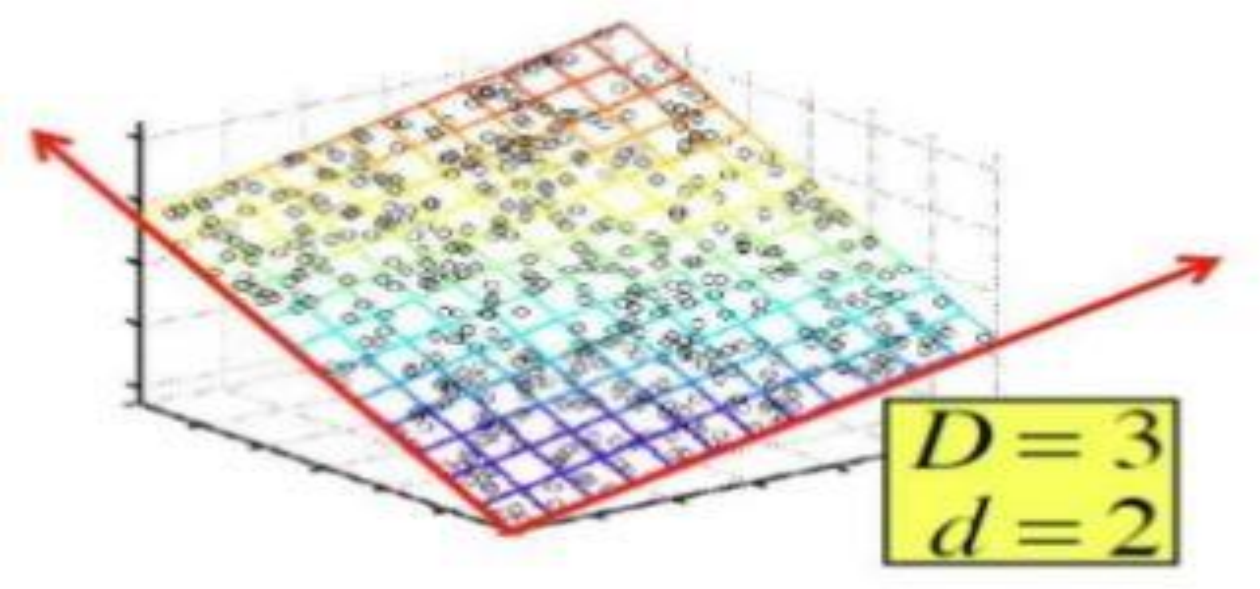


Principal Component Analysis



Reconstruction

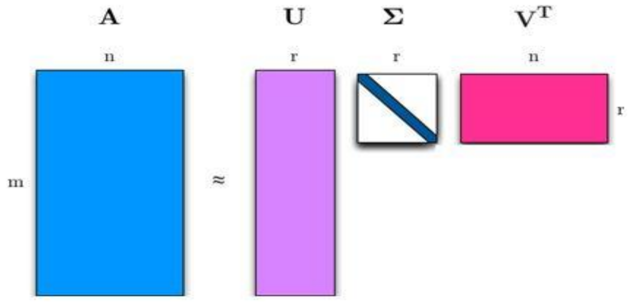
Reduce Dimensionality Illustraion



SVD

What's SVD

- singular-value decomposition(SVD) is a factorization of a real/complex matrix
  - can be used to do dimension reduction
- Every real matrix can be decomposed as $U\Sigma V^T$
  - $U^TU = I, V^TV = I$
  - Σ is a diagonal matrix with non-negative real number on diagonal
  - rank(A) = r



m*n m*r r*r r*n

Matrix rank

- Rank = # of linearly independent columns/rows in A

$$rank\left(\begin{bmatrix} 1 & 2 & 1 \\ -2 & -3 & 1 \\ 3 & 5 & 0 \end{bmatrix}\right) = 2$$

$$rank\left(\begin{bmatrix} 1 & 2 & -2 & 3 \\ 2 & 4 & -2 & 6 \\ 3 & 6 & -6 & 19 \end{bmatrix}\right) = 3$$

$$rank\left(\begin{bmatrix} 1 & 0 & 2 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\right) = 3$$

$$rank\left(\begin{bmatrix} 1 & 2 & -2 & 3 \\ 2 & 4 & -2 & 6 \\ 3 & 6 & -4 & 9 \end{bmatrix}\right) = 2$$

SVD Example

$$\begin{bmatrix} 3 & 1 & 1 \\ -1 & 3 & 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \sqrt{12} & 0 \\ 0 & \sqrt{10} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ \frac{2}{\sqrt{5}} & \frac{-1}{\sqrt{5}} & 0 \end{bmatrix}$$

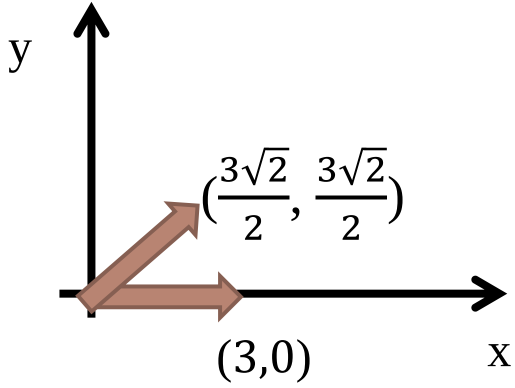
What's SVD

each matrix is a transformation

$$M = \frac{\sqrt{2}}{2}\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$$

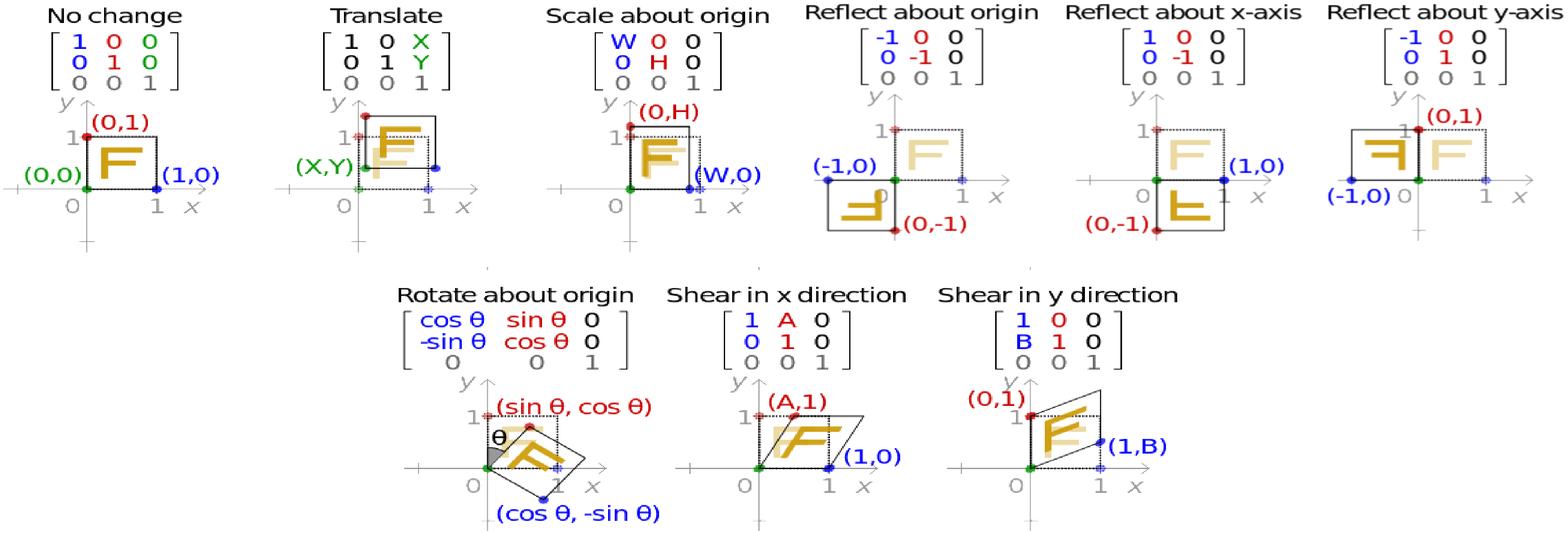
$$X = \begin{bmatrix} 3 \\ 0 \end{bmatrix}$$

$$MX = \frac{\sqrt{2}}{2}\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} 3 \\ 0 \end{bmatrix} = \begin{bmatrix} \frac{3\sqrt{2}}{2} \\ \frac{3\sqrt{2}}{2} \end{bmatrix}$$

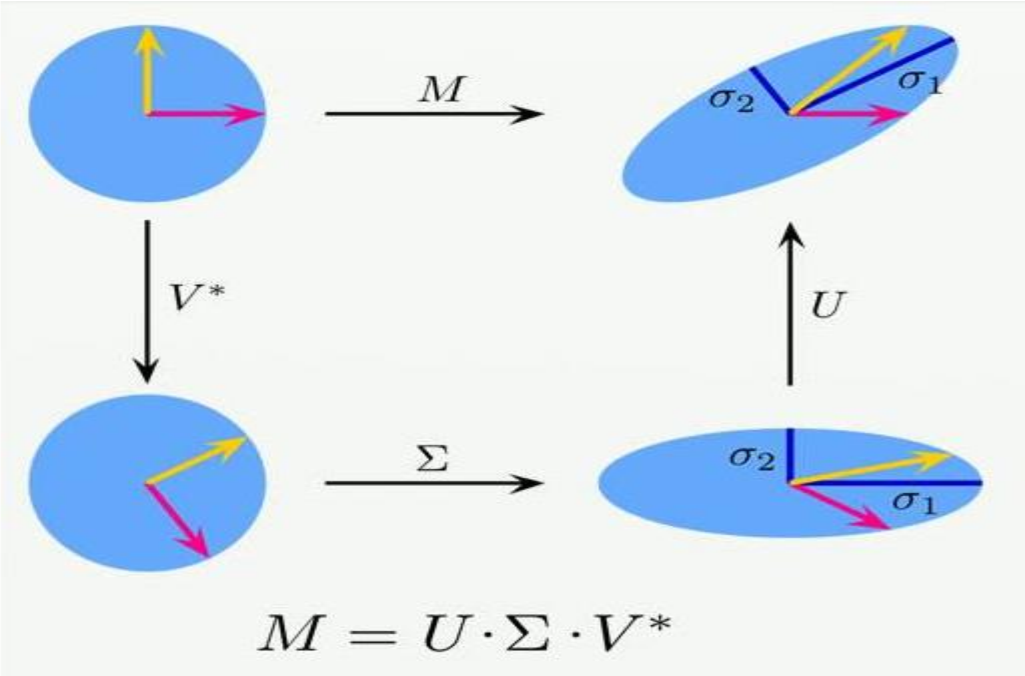


What’s SVD

each matrix is a transformation



What's SVD



Example

	Matrix	Alien	Star Wars	Casablanca	Titanic
Joe	1	1	1	0	0
Jim	3	3	3	0	0
John	4	4	4	0	0
Jack	5	5	5	0	0
Jill	0	2	0	4	4
Jenny	0	0	0	5	5
Jane	0	1	0	2	2

Example

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 4 & 4 & 4 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 2 & 0 & 4 & 4 \\ 0 & 0 & 0 & 5 & 5 \\ 0 & 1 & 0 & 2 & 2 \end{bmatrix} =$$

$$M'$$

$$\begin{bmatrix} .13 & .02 & -.01 \\ .41 & .07 & -.03 \\ .55 & .09 & -.04 \\ .68 & .11 & -.05 \\ .15 & -.59 & .65 \\ .07 & -.73 & -.67 \\ .07 & -.29 & .32 \end{bmatrix} \begin{bmatrix} 12.4 & 0 & 0 \\ 0 & 9.5 & 0 \\ 0 & 0 & 1.3 \end{bmatrix} \begin{bmatrix} .56 & .59 & .56 & .09 & .09 \\ .12 & -.02 & .12 & -.69 & -.69 \\ .40 & -.80 & .40 & .09 & .09 \end{bmatrix}$$

$$U \qquad \Sigma \qquad V^T$$

Example

Assume we drop the smallest singular value

$$
\begin{bmatrix}
1 & 1 & 1 & 0 & 0 \\
3 & 3 & 3 & 0 & 0 \\
4 & 4 & 4 & 0 & 0 \\
5 & 5 & 5 & 0 & 0 \\
0 & 2 & 0 & 4 & 4 \\
0 & 0 & 0 & 5 & 5 \\
0 & 1 & 0 & 2 & 2
\end{bmatrix} =
$$

$$M'$$

$$
\underbrace{\begin{bmatrix}
.13 & .02 & \cancel{-.01} \\
.41 & .07 & \cancel{-.03} \\
.55 & .09 & \cancel{-.04} \\
.68 & .11 & \cancel{-.05} \\
.15 & -.59 & \cancel{.65} \\
.07 & -.73 & \cancel{-.67} \\
.07 & -.29 & \cancel{.32}
\end{bmatrix}}_{U}
\underbrace{\begin{bmatrix}
12.4 & 0 & 0 \\
0 & 9.5 & 0 \\
\cancel{0} & \cancel{0} & \cancel{1.3}
\end{bmatrix}}_{\Sigma}
\underbrace{\begin{bmatrix}
.56 & .59 & .56 & .09 & .09 \\
.12 & -.02 & .12 & -.69 & -.69 \\
\cancel{.40} & \cancel{-.80} & \cancel{.40} & \cancel{.09} & \cancel{.09}
\end{bmatrix}}_{V^T}
$$

Example

$$
\begin{bmatrix} .13 & .02 \\ .41 & .07 \\ .55 & .09 \\ .68 & .11 \\ .15 & -.59 \\ .07 & -.73 \\ .07 & -.29 \end{bmatrix}
\begin{bmatrix} 12.4 & 0 \\ 0 & 9.5 \end{bmatrix}
\begin{bmatrix} .56 & .59 & .56 & .09 & .09 \\ .12 & -.02 & .12 & -.69 & -.69 \end{bmatrix}
$$

$$
= \begin{bmatrix} 0.93 & 0.95 & 0.93 & .014 & .014 \\ 2.93 & 2.99 & 2.93 & .000 & .000 \\ 3.92 & 4.01 & 3.92 & .026 & .026 \\ 4.84 & 4.96 & 4.84 & .040 & .040 \\ 0.37 & 1.21 & 0.37 & 4.04 & 4.04 \\ 0.35 & 0.65 & 0.35 & 4.87 & 4.87 \\ 0.16 & 0.57 & 0.16 & 1.98 & 1.98 \end{bmatrix}
$$

Example

$$\begin{bmatrix} \mathbf{1} & \mathbf{1} & \mathbf{1} & 0 & 0 \\ \mathbf{3} & \mathbf{3} & \mathbf{3} & 0 & 0 \\ \mathbf{4} & \mathbf{4} & \mathbf{4} & 0 & 0 \\ \mathbf{5} & \mathbf{5} & \mathbf{5} & 0 & 0 \\ 0 & \mathbf{2} & 0 & \mathbf{4} & \mathbf{4} \\ 0 & 0 & 0 & \mathbf{5} & \mathbf{5} \\ 0 & \mathbf{1} & 0 & \mathbf{2} & \mathbf{2} \end{bmatrix} \approx \begin{bmatrix} \mathbf{0.92} & \mathbf{0.95} & \mathbf{0.92} & 0.01 & 0.01 \\ \mathbf{2.91} & \mathbf{3.01} & \mathbf{2.91} & -0.01 & -0.01 \\ \mathbf{3.90} & \mathbf{4.04} & \mathbf{3.90} & 0.01 & 0.01 \\ \mathbf{4.82} & \mathbf{5.00} & \mathbf{4.82} & 0.03 & 0.03 \\ 0.70 & \mathbf{0.53} & 0.70 & \mathbf{4.11} & \mathbf{4.11} \\ -0.69 & 1.34 & -0.69 & \mathbf{4.78} & \mathbf{4.78} \\ 0.32 & \mathbf{0.23} & 0.32 & \mathbf{2.01} & \mathbf{2.01} \end{bmatrix}$$

Example

$\boldsymbol{assume\ new\ sample\ data\ [1, 1, 0, 3, 2]}$

$$[1,1,0,3,2]\begin{bmatrix} .56 & .12 \\ .59 & -.02 \\ .56 & .12 \\ .09 & -.69 \\ .09 & -.69 \end{bmatrix} = [1.599 \quad -3.35]$$

from 5D to 2D

How many singular values to keep?

$$\begin{array}{cc} & \longleftarrow m \longrightarrow \\ n \updownarrow & \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 4 & 4 & 4 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 2 & 0 & 4 & 4 \\ 0 & 0 & 0 & 5 & 5 \\ 0 & 1 & 0 & 2 & 2 \end{bmatrix} \end{array} = \sigma_1 \quad u_1 \quad v^T{}_1 \quad + \quad \sigma_2 \quad u_2 \quad v^T{}_2 \quad + \dots$$

Assume: $\sigma_1 \geq \sigma_2 \geq \sigma_3 \geq \dots$

keep 80-90% of energy

How many singular values to keep?

$$
\underset{U}{\begin{bmatrix} .13 & .02 & -.01 \\ .41 & .07 & -.03 \\ .55 & .09 & -.04 \\ .68 & .11 & -.05 \\ .15 & -.59 & .65 \\ .07 & -.73 & -.67 \\ .07 & -.29 & .32 \end{bmatrix}}
\underset{\Sigma}{\begin{bmatrix} 12.4 & 0 & 0 \\ 0 & 9.5 & 0 \\ 0 & 0 & 1.3 \end{bmatrix}}
\underset{V^T}{\begin{bmatrix} .56 & .59 & .56 & .09 & .09 \\ .12 & -.02 & .12 & -.69 & -.69 \\ .40 & -.80 & .40 & .09 & .09 \end{bmatrix}}
$$

$$\frac{12.4 + 9.5}{12.4 + 9.5 + 1.3} = 0.944$$

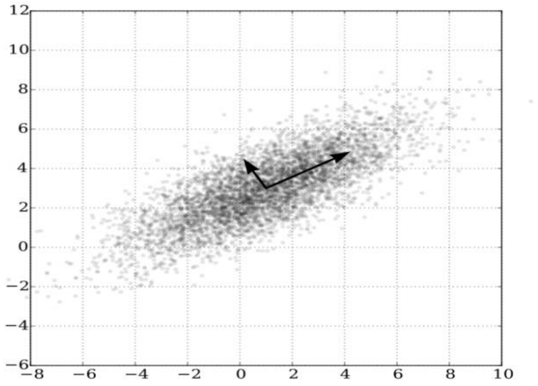
(補充)SVD

- How to calculate SVD on a matrix
  - http://www.d.umn.edu/~mhampton/m4326svd_example.pdf
- SVD calculator
  - https://m.wolframalpha.com/input/?i=SVD+%7B%7B1%2C+0%2C+-1%7D%2C+%7B-2%2C+1%2C+4%7D%7D&lk=3

PCA

What's PCA

- principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables

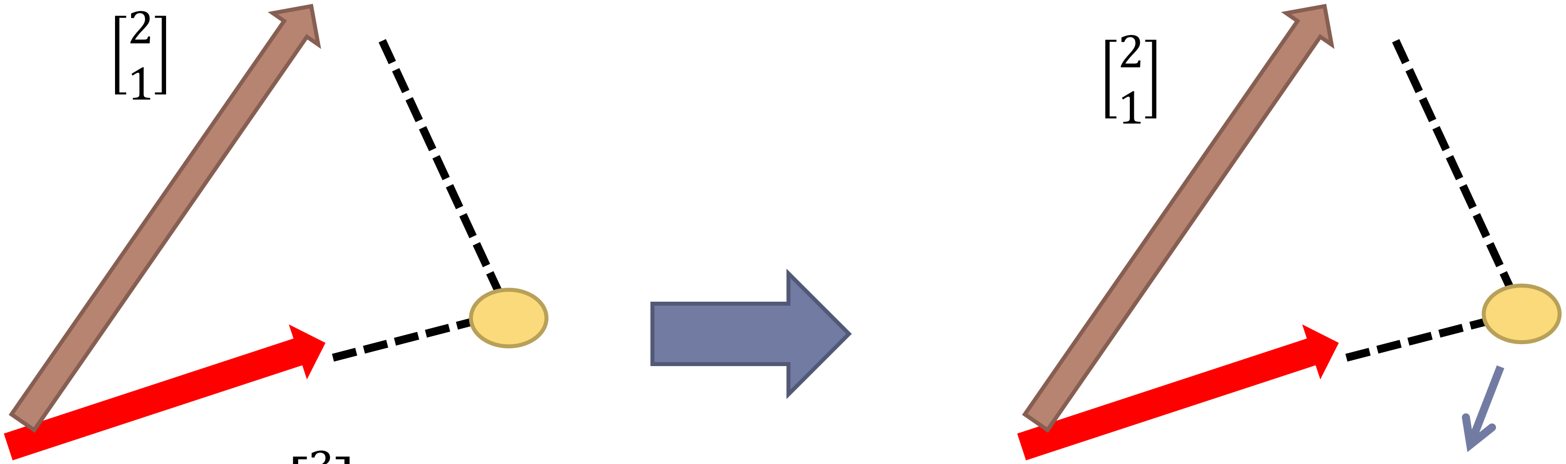


- Find a projection matrix w from d-dimensional to k-dimensional vectors that keeps error low

$$z = \mathbf{w}^T \mathbf{x}$$

unit vector

Example



$$\begin{bmatrix}2\\1\end{bmatrix}$$

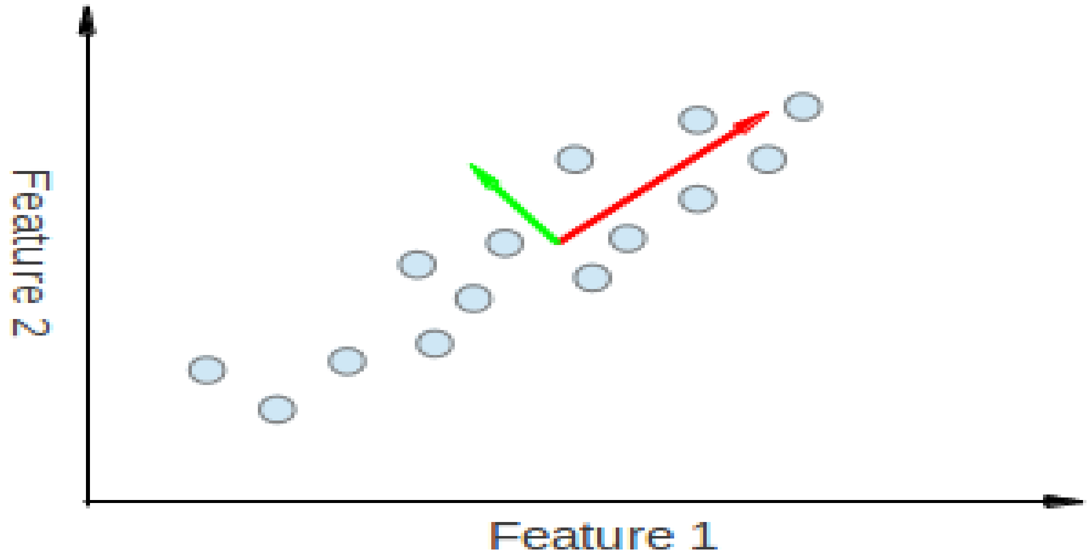
$$\begin{bmatrix}\frac{3}{5}\\\frac{4}{5}\end{bmatrix}$$

$$\begin{bmatrix}2\\1\end{bmatrix}$$

$$\begin{bmatrix}\frac{3}{5}\\\frac{4}{5}\end{bmatrix}^T * \begin{bmatrix}2\\1\end{bmatrix} = 2$$

coordinate of yellow point under red axis

Which projection axis is better?



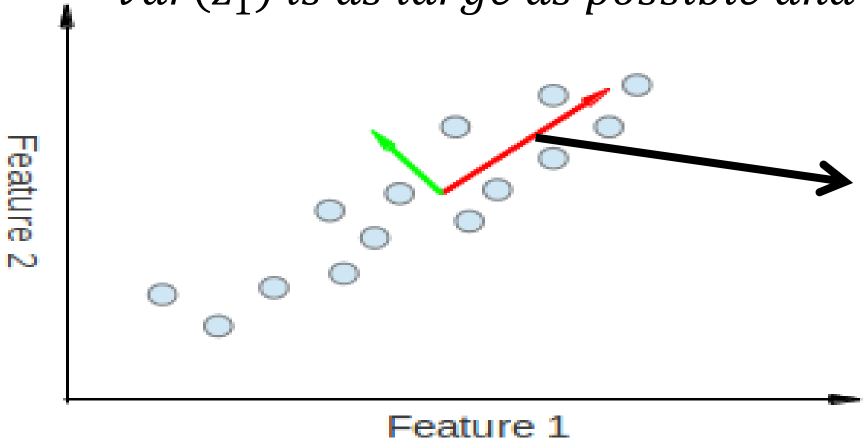
red arrow have larger variance

PCA

Project all the data points x onto **first** axis w1 and obtain a set of z1

What we want:

$$var(z_1)\ is\ as\ large\ as\ possible\ and\ \|w^1\| = 1$$



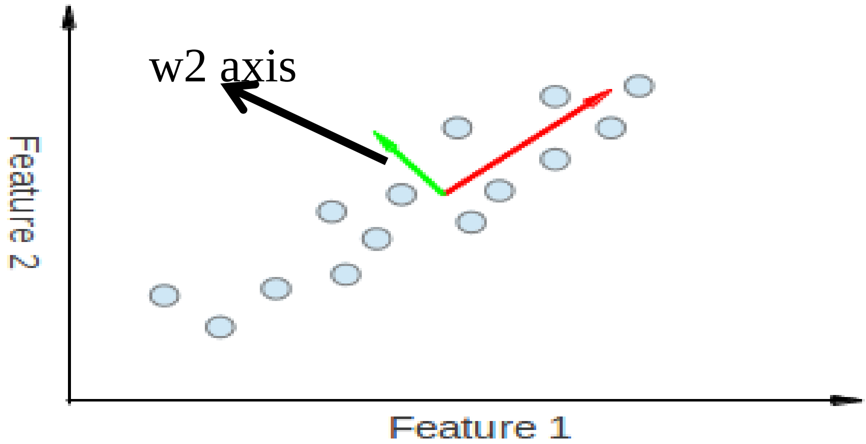
w1 axis

PCA

Project all the data points x onto second axis w2 and obtain a set of z2

What we want:

$$var(z_2)\ is\ as\ large\ as\ possible\ and\ \|w^2\| = 1, \boxed{w^1 \cdot w^2 = 0}$$



should find the new axis which is orthogonal with all of previous axis

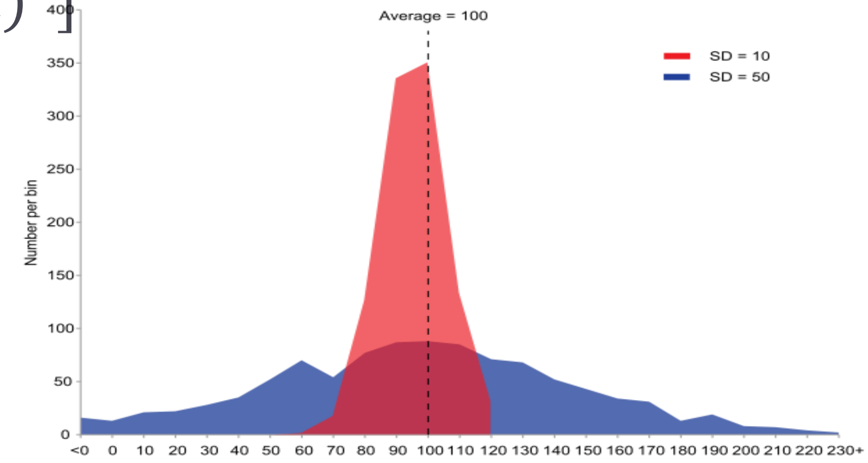
PCA concept

- Choose directions such that a total variance of data will be maximum
  - Maximize Total Variance
- Choose directions that are orthogonal
  - Minimize correlation
- Choose k<d orthogonal directions which maximize total variance

Variance and Covariance

- Variance
  - Measure of how far a set of numbers are spread out from their average value
  - $Var(X) = \sigma_X{}^2 = E[(X-\mu)^2]$

standard deviation



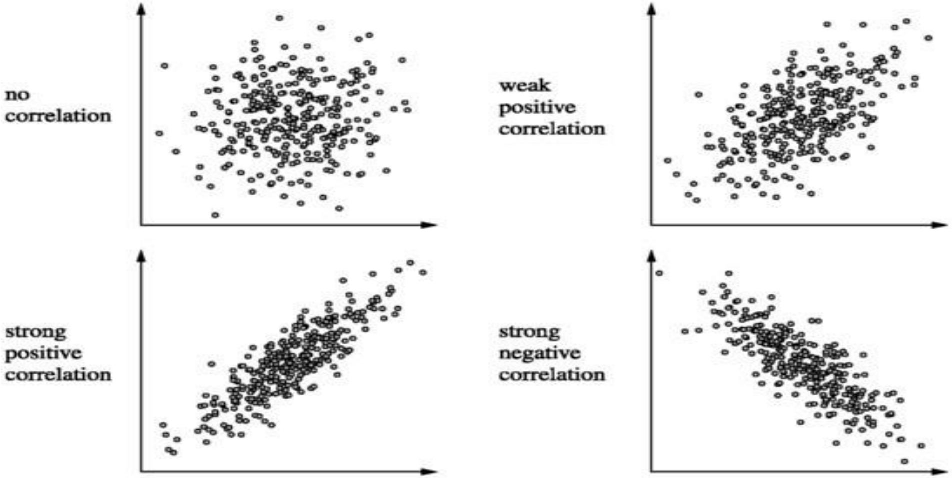
Variance and Covariance

- covariance
  - measure of the joint variability of two random variables
  - $\text{cov}(X, Y) = E[(X - \mu_x)(Y - \mu_y)]$
  - $\text{cov}(X, X) = Var(X)$

- Magnitude of covariance is meaningless but normalized covariance(correlation) meaningful

Variance and Covariance

- Correlation
  - $\rho_{X,Y}$ (correlation)
  - = covariance with normalization
  - $= \dfrac{cov(X,Y)}{\sigma_X \sigma_Y}$



Covariance Matrix

$$\mathrm{Cov}[X,Y] = \begin{bmatrix} \mathrm{E}[(X_1 - \mathrm{E}[X_1])(Y_1 - \mathrm{E}[Y_1])] & \mathrm{E}[(X_1 - \mathrm{E}[X_1])(Y_2 - \mathrm{E}[Y_2])] \\ \mathrm{E}[(X_2 - \mathrm{E}[X_2])(Y_1 - \mathrm{E}[Y_1])] & \mathrm{E}[(X_2 - \mathrm{E}[X_2])(Y_2 - \mathrm{E}[Y_2])] \\ \mathrm{E}[(X_3 - \mathrm{E}[X_3])(Y_1 - \mathrm{E}[Y_1])] & \mathrm{E}[(X_3 - \mathrm{E}[X_3])(Y_2 - \mathrm{E}[Y_2])] \end{bmatrix}$$

$$= \begin{bmatrix} \mathrm{Cov}[X_1, Y_1] & \mathrm{Cov}[X_1, Y_2] \\ \mathrm{Cov}[X_2, Y_1] & \mathrm{Cov}[X_2, Y_2] \\ \mathrm{Cov}[X_3, Y_1] & \mathrm{Cov}[X_3, Y_2] \end{bmatrix}$$

Covariance Matrix Example

X1	X2
3	7
2	4

$$\begin{bmatrix} var(X1) & conv(X1, X2) \\ conv(X1, X2) & var(X2) \end{bmatrix}$$

$$\bar{x}_1 = \frac{3+2}{2} = \frac{5}{2}$$

$$\bar{x}_2 = \frac{7+4}{2} = \frac{11}{2}$$

$$var(x_1) = (3 - \frac{5}{2})^2 + (2 - \frac{5}{2})^2$$

$$var(x_2) = (7 - \frac{11}{2})^2 + (4 - \frac{11}{2})^2$$

$$cov(x_1, x_2) = \left(3 - \frac{5}{2}\right)\left(7 - \frac{11}{2}\right) + \left(2 - \frac{5}{2}\right)\left(4 - \frac{11}{2}\right)$$

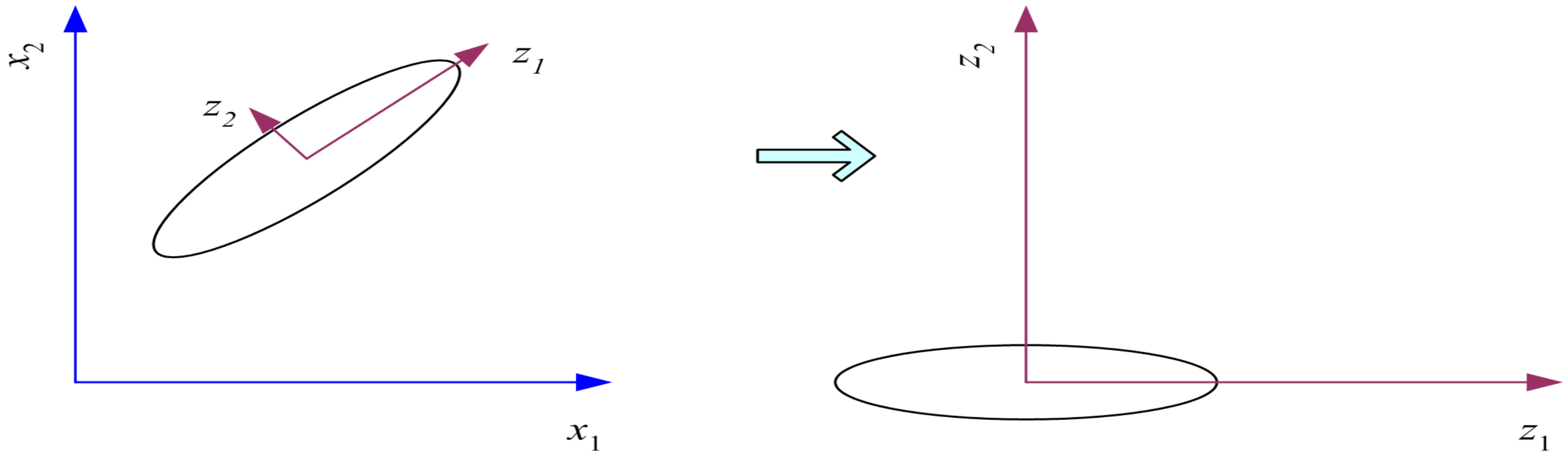
Covariance Matrix Example

X1	X2
3	7
2	4

$$\begin{bmatrix} var(X1) & conv(X1,X2) \\ conv(X1,X2) & var(X2) \end{bmatrix}$$

$$\begin{bmatrix} \frac{1}{4} & \frac{3}{4} \\ \frac{3}{4} & \frac{9}{4} \end{bmatrix}$$

What PCA does



- Step 1: Get some data
- Step 2: Subtract the mean
- Step 3: Calculate the covariance matrix
- Step 4: Calculate the eigenvectors and eigenvalues of the covariance matrix
- Step 5: Choosing components and forming a feature vector

Example

Data =

x	y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9

DataAdjust =

x	y
.69	.49
-1.31	-1.21
.39	.99
.09	.29
1.29	1.09
.49	.79
.19	-.31
-.81	-.81
-.31	-.31
-.71	-1.01

Example

DataAdjust =

x	y
.69	.49
-1.31	-1.21
.39	.99
.09	.29
1.29	1.09
.49	.79
.19	-.31
-.81	-.81
-.31	-.31
-.71	-1.01

calculate covariance matrix

$$cov = \begin{pmatrix} .616555556 & .615444444 \\ .615444444 & .716555556 \end{pmatrix}$$

Example

$$eigenvalues = \begin{pmatrix} .0490833989 \\ 1.28402771 \end{pmatrix}$$

$$eigenvectors = \begin{pmatrix} -.735178656 & -.677873399 \\ .677873399 & -.735178656 \end{pmatrix}$$

Find eigenvalues and eigenvectors

Example

$$eigenvalues = \begin{pmatrix} \cancel{.0490833989} \\ 1.28402771 \end{pmatrix}$$

$$eigenvectors = \begin{pmatrix} -.735178656 & -.677873399 \\ .677873399 & -.735178656 \end{pmatrix}$$

Leave eigenvalue that is larger

Note: Larger eigenvalue mean large variance on axis

Example

x	y
.69	.49
-1.31	-1.21
.39	.99
.09	.29
1.29	1.09
.49	.79
.19	-.31
-.81	-.81
-.31	-.31
-.71	-1.01

$$\begin{bmatrix} -0.6778 \\ -0.7351 \end{bmatrix}$$

Transformed Data (Single eigenvector)

x
-.827970186
1.77758033
-.992197494
-.274210416
-1.67580142
-.912949103
.0991094375
1.14457216
.438046137
1.22382056

from 2D to 1D

Another PCA Interpretation

Demo

- PCA demo
  - http://setosa.io/ev/principal-component-analysis/

Example and Practice

- Example
  - PCA
    - example/dimension reduction
- Practice
  - Try to reduce dimension on each data to 3D(note that please drop the last feature in this dataset)
    - dataset/seeds_dataset.csv
    - practice/dimension reduction

t-SNE

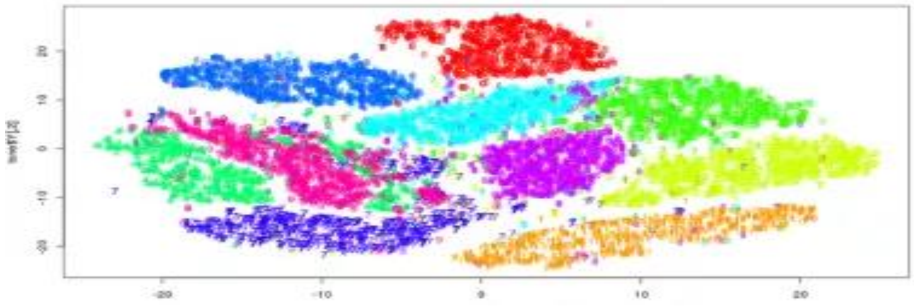
What's t-SNE

- T-distributed Stochastic Neighbor Embedding (t-SNE) is a machine learning algorithm for visualization
  - a nonlinear dimensionality reduction technique
  - Usually used to map N-D data into 2D/3D

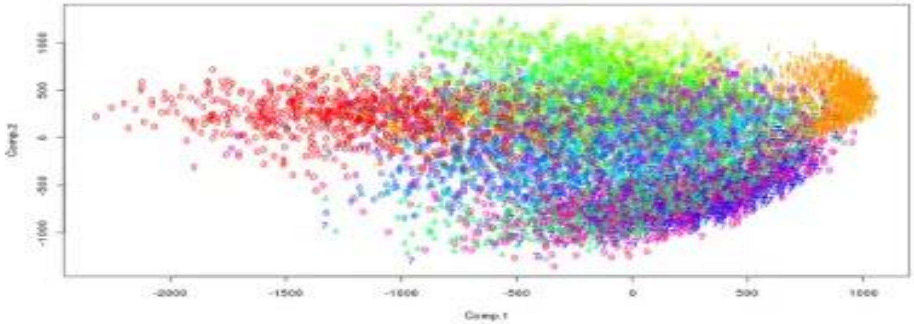
What’s t-SNE

- Advantage of t-SNE
  - Some classification group should be closer
  - Different classification group should be farther(main different compared with PCA)

T-SNE



PCA



t-SNE concept

X(High dimension space) → Z(2D/3D space)

$$P(x^j|x^i) = \frac{S(x^i, x^j)}{\sum_{k \neq i} S(x^i, x^k)}$$

$$Q(z^j|z^i) = \frac{S'(z^i, z^j)}{\sum_{k \neq i} S'(z^i, z^k)}$$

Define similarity function and calculate similarity between all data pair

How to choose similarity function?

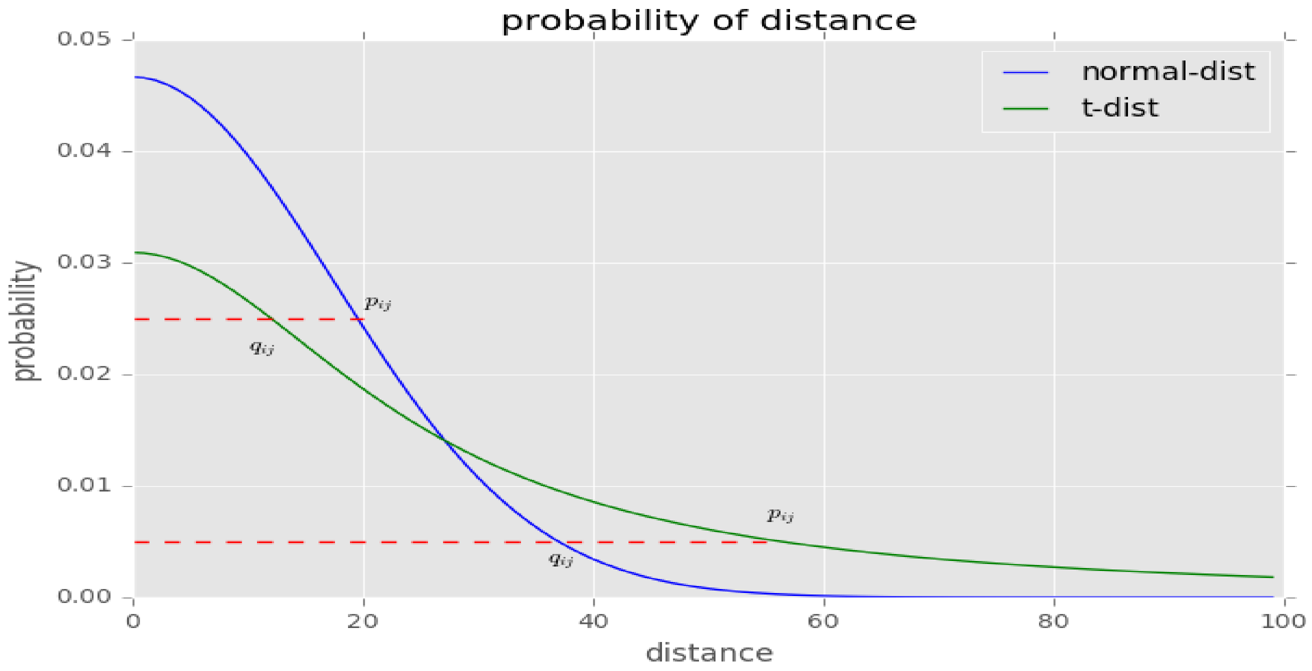
SNE

$$S(x^i, x^j) = e^{-\|x^i - x^j\|} \qquad S'(z^i, z^j) = e^{-\|z^i - z^j\|}$$

t-SNE

$$S(x^i, x^j) = e^{-\|x^i - x^j\|} \qquad S'(z^i, z^j) = \frac{1}{1 + \|z^i - z^j\|}$$

How to choose similarity function?



t-SNE concept

$$loss = \sum_{i} KL(P_i || Q_i) = \sum_{i \neq j} P\left(x^j \middle| x^i\right) \log\left(\frac{P\left(x^j \middle| x^i\right)}{Q\left(z^j \middle| z^i\right)}\right)$$

use gradient descent to optimization

t-sne reference

- t-sne result
  - https://lvdmaaten.github.io/tsne/
- t-sne source code
  - https://github.com/oreillymedia/t-SNE-tutorial
- Other reference
  - https://medium.com/d-d-mag/%E6%B7%BA%E8%AB%87%E5%85%A9%E7%A8%AE%E9%99%8D%E7%B6%AD%E6%96%B9%E6%B3%95-pca-%E8%88%87-t-sne-d4254916925b