

Computational Data Analysis

Machine Learning

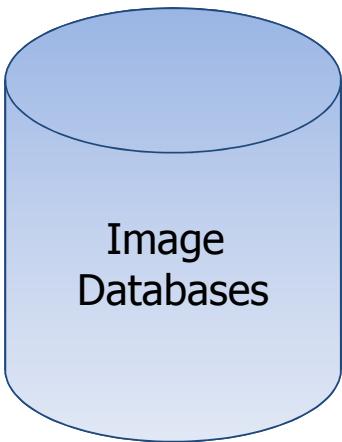
Yao Xie, Ph.D.

Associate Professor

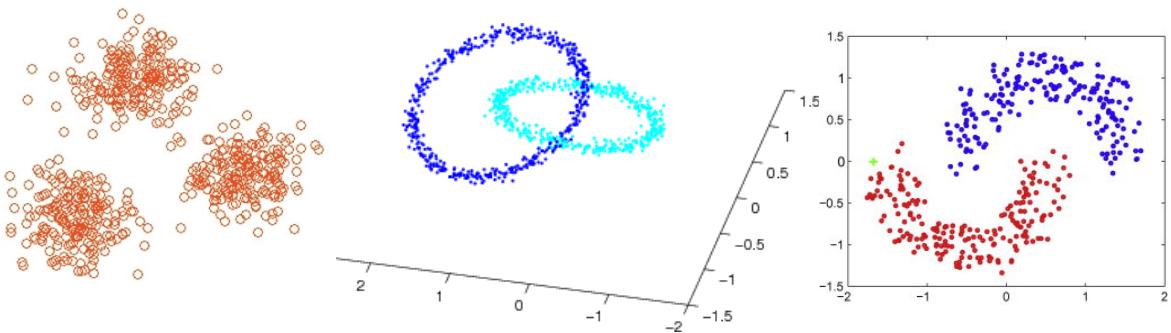
Harold R. and Mary Anne Nash Early Career Professor
H. Milton Stewart School of Industrial and Systems
Engineering

Dimensionality Reduction
Principal Component Analysis





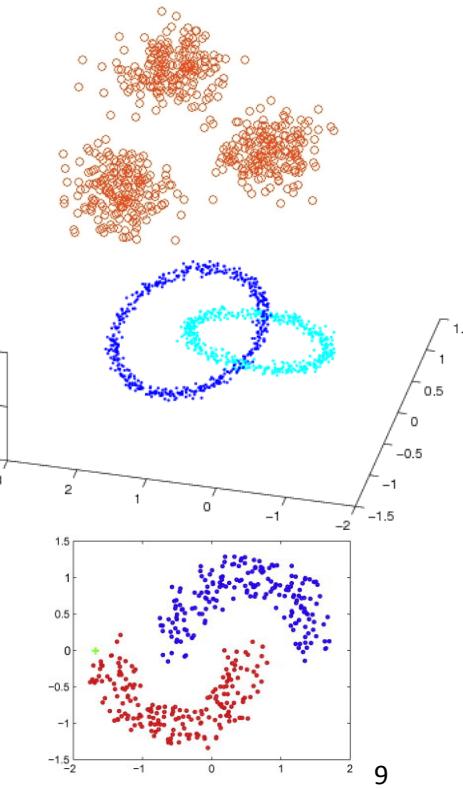
What are the relations
between data points?



Handwritten digits

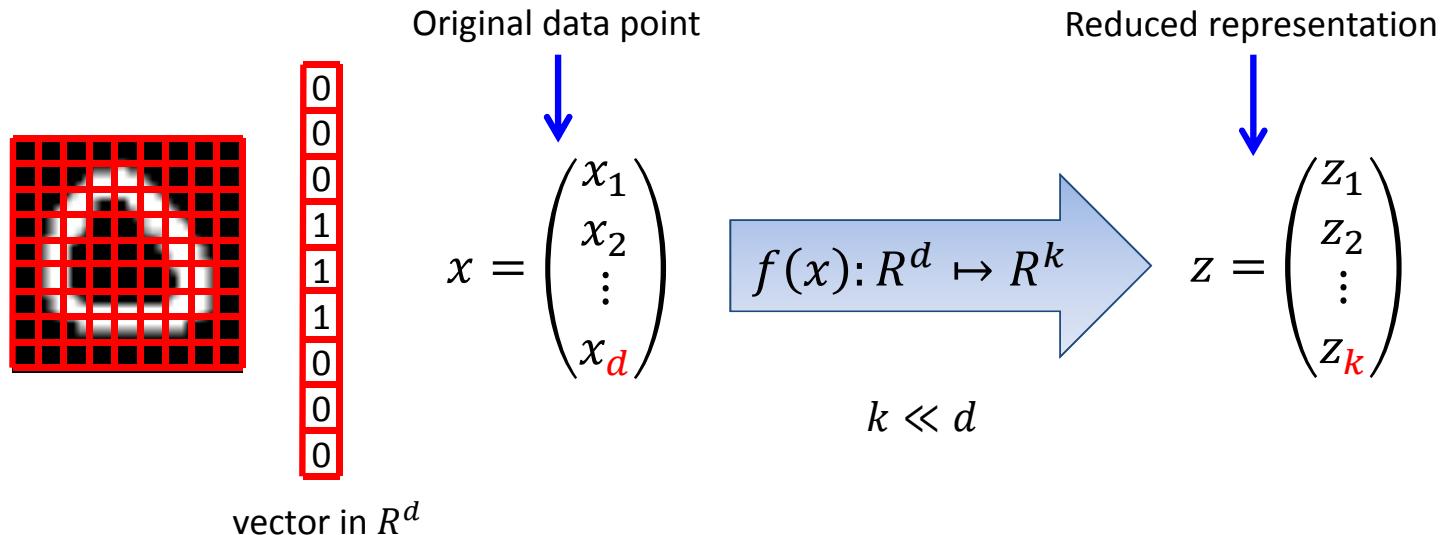
```
72104149590690159784966540740131  
34727121174235124463556041957843  
74043070291732971627847361369314  
176960549921948713974449254767905  
85665781016467317182029355156034  
46546545144723271818185089250111  
09031642361113952945939036557227  
1284173388792415987230442419577  
28268577918180301994182129759264  
15429204002847124027433003196525  
17930420711215339786361381051315  
56185179462250656372088541140337  
61621928619525442838245031773797  
192119292049148184598837600302661  
9533239126805666382758961841269  
19754089914523789406395213136571  
22632654897130383193446421825488  
40023277687447969098046063548339  
33378087170654380963809968685786  
02402231975108462479309822927359  
18020511376712580371409186774399  
19317397691332336128585114431077  
07944855408215845040615326726931  
46251206217341054311749948402451  
16471942415538314568941538032512  
83440883317359632613607217142821  
79611248177480231310770355276692  
83522560829288887493066321322930  
05781446029147473988471212232323  
91740355868267663279117564951334  
78911691445406223151203812671623  
9012089
```

What are the relations between data points?



What is dimensionality reduction?

- The process of reducing the number of random variables under consideration
 - One can combine, transform or select variables
 - One can use linear or nonlinear operations



Why dimensionality reduction and how to think

- The dimension-reduced data can be used for
 - Visualizing, exploring and understanding the data
 - Aggregating weak signals in the data
 - Cleaning the data
 - Speeding up subsequent learning task
 - Building simpler model later
- Key questions of a dimensionality reduction algorithm
 - What is the criterion for carrying out the reduction process?
 - What are the algorithm steps?

Principal component analysis

- Given m data points, $\{x^1, x^2, \dots, x^m\} \in R^d$, with mean
- Step 1: Estimate the mean and covariance matrix from data

$$\mu = \frac{1}{m} \sum_{i=1}^m x^i \quad \text{and} \quad C = \frac{1}{m} \sum_{i=1}^m (x^i - \mu)(x^i - \mu)^\top$$

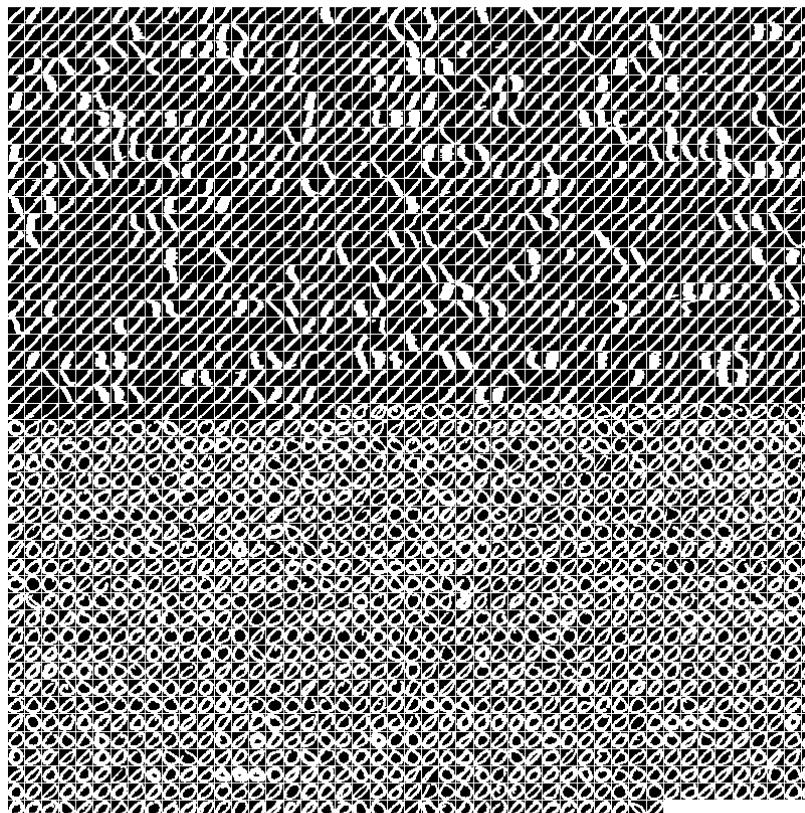
“weights”
to combine
features:
**Principal
directions**

-
- Step 2: Take the eigenvectors w^1, w^2, \dots of C corresponding to the largest eigenvalue λ_1 , the second largest eigenvalue λ_2 ...
 - Step 3: Compute reduced representation (**principle components**)

$$z^i = \begin{pmatrix} w^1^\top (x^i - \mu) / \sqrt{\lambda_1} \\ w^2^\top (x^i - \mu) / \sqrt{\lambda_2} \\ \vdots \end{pmatrix}$$

Run demo PCA_digits.m

digit 1 and 0



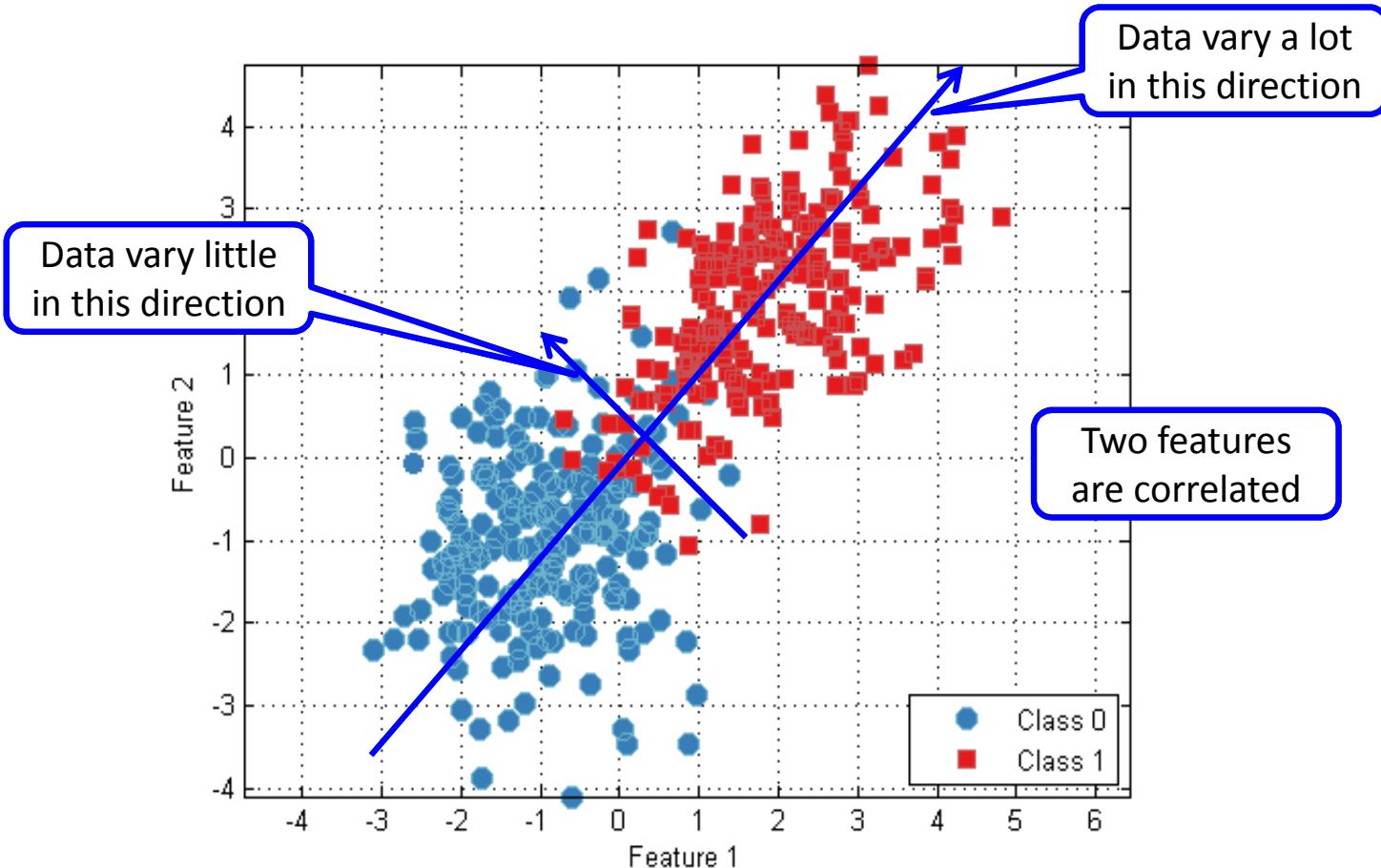
Run demo PCA_leaf.m



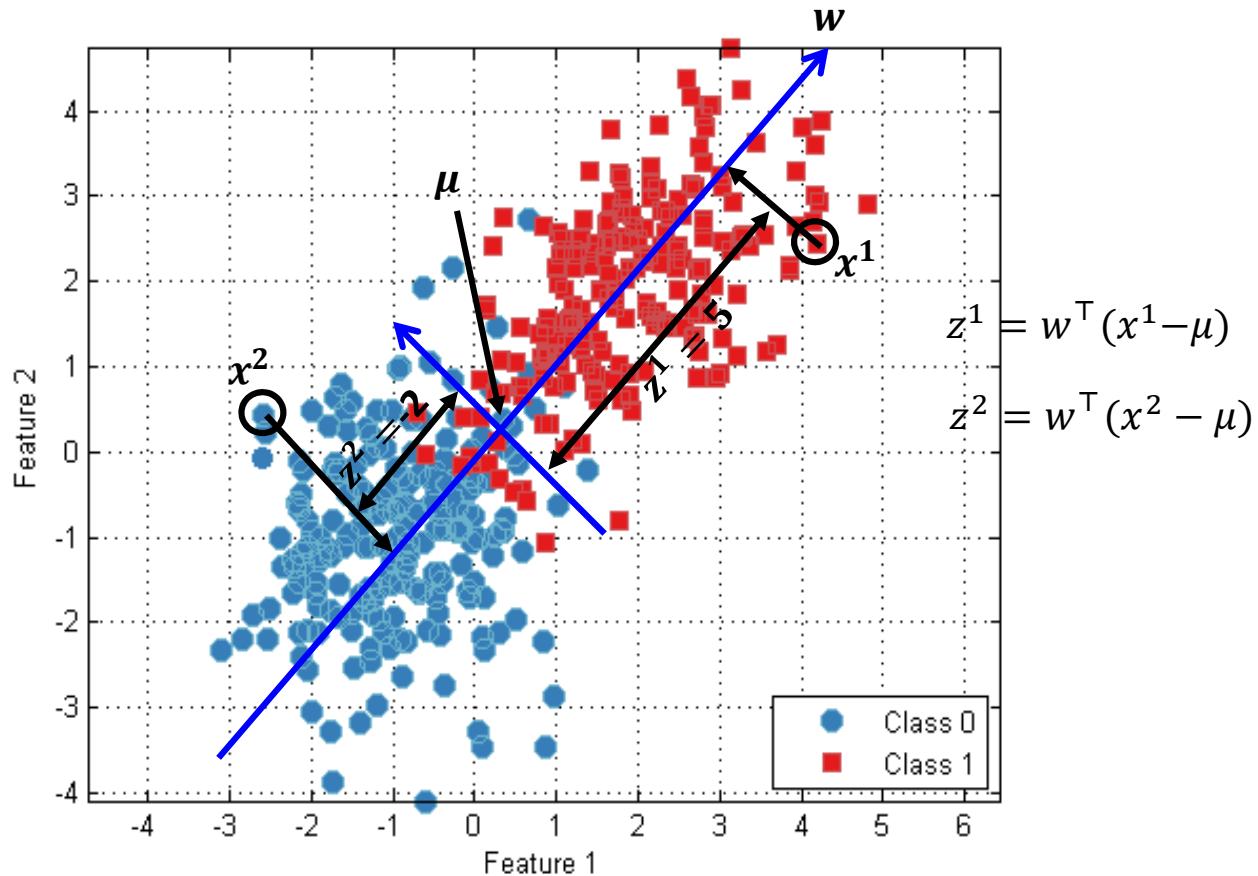
Use what criterion for reduction?

- There are many criteria (geometric based, information theory based, etc.)
- One useful criterion: want to capture **variation** in data
 - variations are “signals” or information in the data
 - need to normalize each variables first
- In the process, also discover variables or dimensions highly **correlated**
 - represent highly related features
 - combine them to form a stronger signal
 - lead to simpler presentation

An example



Reduce to 1 dimension



How to formulate the problem

- Given m data points, $\{x^1, x^2, \dots, x^m\} \in R^n$, with their mean $\mu = \frac{1}{m} \sum_{i=1}^m x^i$
- Find a direction $w \in R^n$ where $\|w\| \leq 1$
- Such that the variance (or variation) of the data along direction w is maximized

$$\max_{w: \|w\| \leq 1} \frac{1}{m} \sum_{i=1}^m (w^\top x^i - w^\top \mu)^2$$



variance

Is it an easy optimization problem?

- Manipulate the objective with linear algebra

$$\begin{aligned}& \frac{1}{m} \sum_{i=1}^m (w^\top x^i - w^\top \mu)^2 \\&= \frac{1}{m} \sum_{i=1}^m (w^\top (x^i - \mu))^2 \\&= \frac{1}{m} \sum_{i=1}^m w^\top (x^i - \mu)(x^i - \mu)^\top w \\&= w^\top \left(\frac{1}{m} \sum_{i=1}^m (x^i - \mu)(x^i - \mu)^\top \right) w\end{aligned}$$



covariance matrix C

Landscape of the optimization problem

- Suppose the data has two dimension ($n = 2$)
- C is a diagonal matrix

$$C = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$$

- The optimization problem becomes

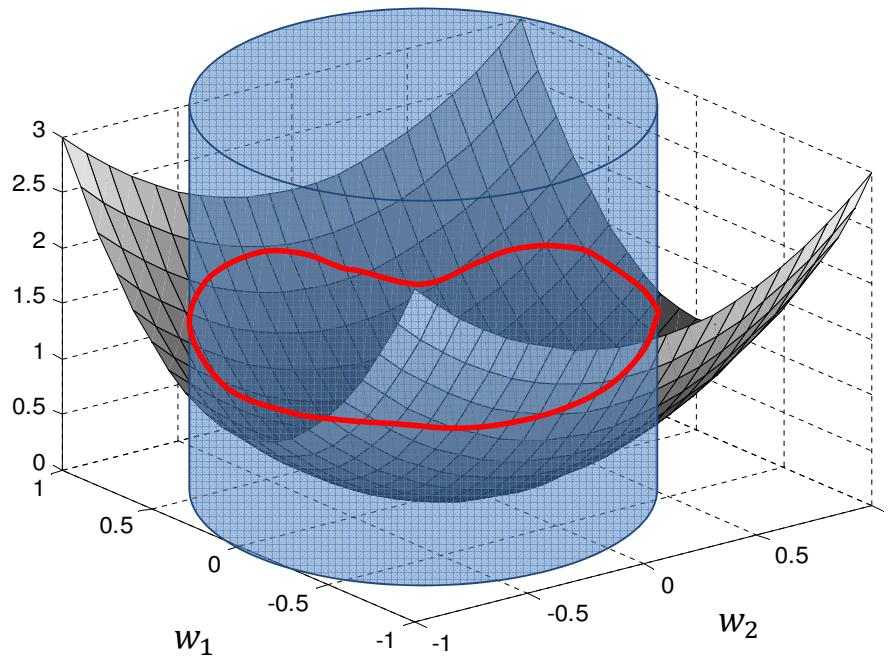
$$\max_{w: \|w\| \leq 1} w^T C w$$

$$= \max_{w: \|w\| \leq 1} (w_1, w_2) \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}$$

$$= \max_{w: \|w\| \leq 1} w_1^2 + 2w_2^2$$

Landscape of the optimization problem

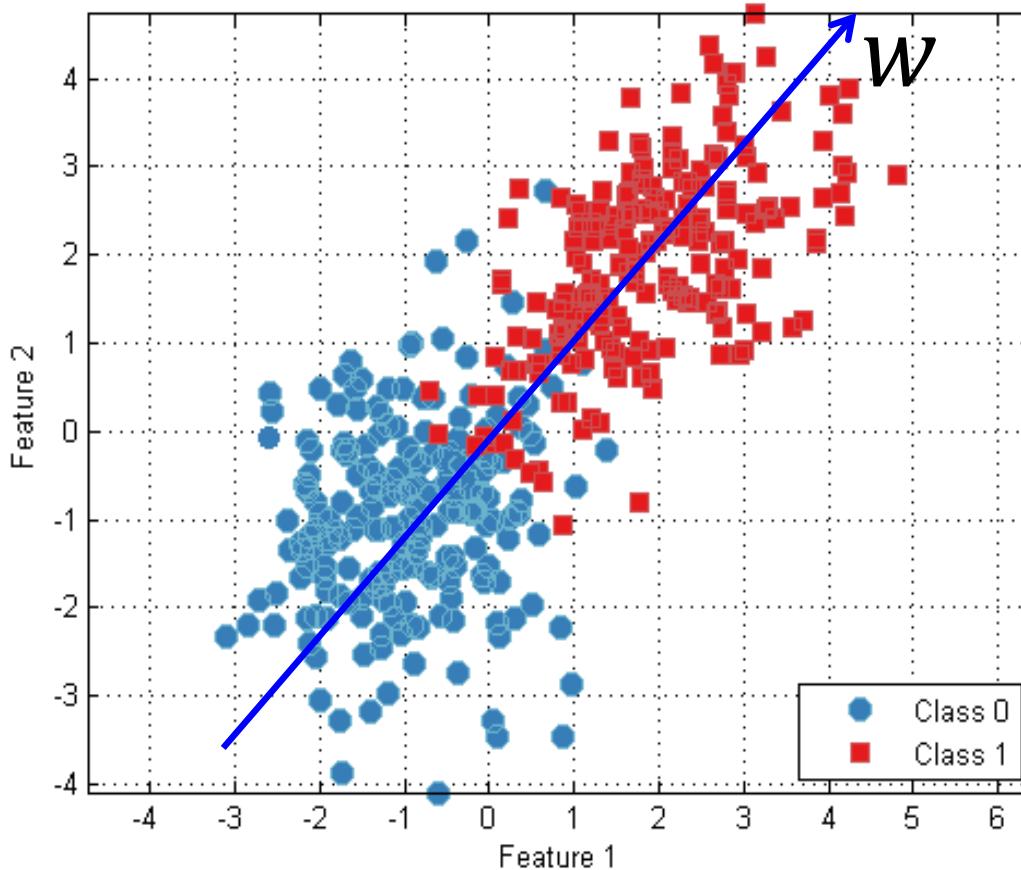
- $f(w_1, w_2) = w_1^2 + 2w_2^2$



Eigen-value problem

- Eigen-value problem
 - Given a symmetric matrix $C \in R^{n \times n}$
 - Find a vector $w \in R^n$ and $\|w\| = 1$
 - Such that
$$Cw = \lambda w$$
- There will be multiple solution of w^1, w^2, \dots with different $\lambda_1, \lambda_2, \dots$
 - They are ortho-normal: $w^{i^\top} w^i = 1, w^{i^\top} w^j = 0$

Principal direction of the data



Variance of in the principal direction

- Principal direction w satisfies

$$Cw = \lambda w$$

- Variance in principal direction is

$$w^\top C w$$

$$= \lambda w^\top w$$

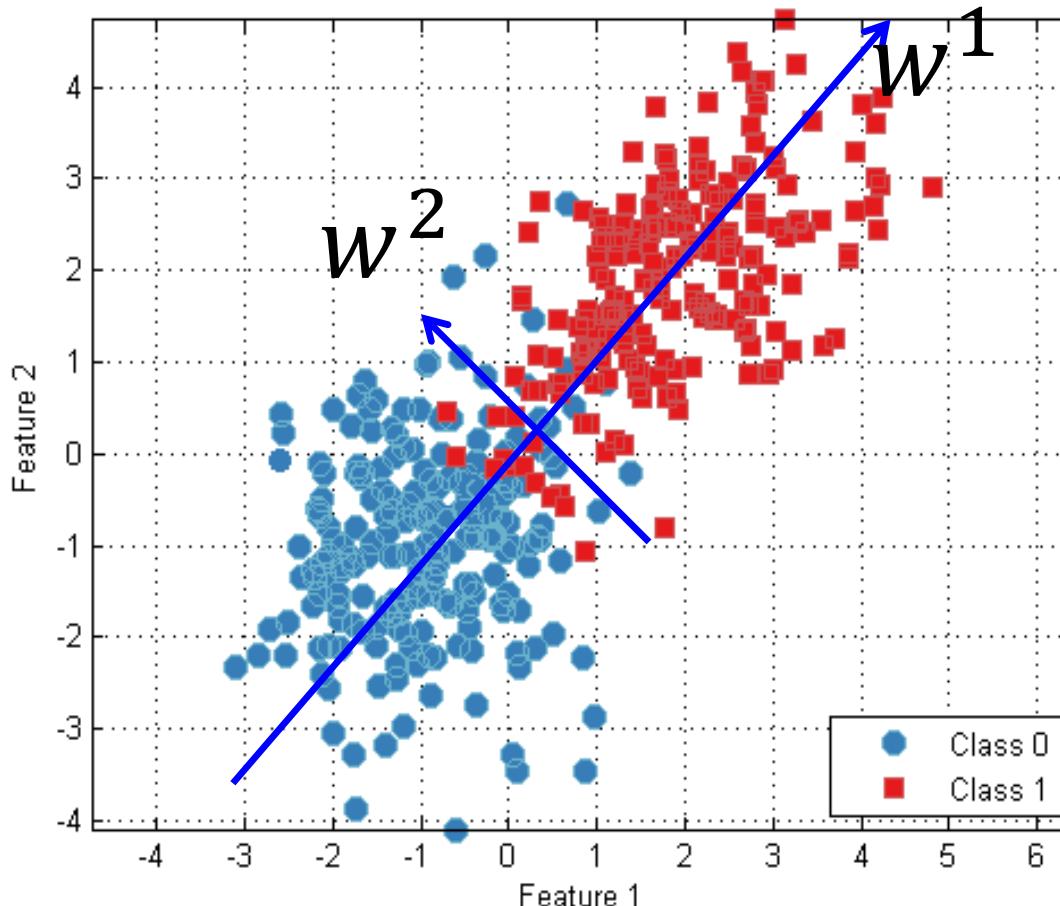
$$= \lambda$$

eigen-value

Find multiple principal directions

- Directions w^1, w^2, \dots which has
 - the largest variances
 - but are **orthogonal** to each other
- Take the eigenvectors w^1, w^2, \dots of C corresponding to
 - the largest eigenvalue λ_1 ,
 - the second largest eigenvalue λ_2
 - ...

Principal direction of the data



Principal component analysis

- Given m data points, $\{x^1, x^2, \dots, x^m\} \in R^d$, with mean
- Step 1: Estimate the mean and covariance matrix from data

$$\mu = \frac{1}{m} \sum_{i=1}^m x^i \quad \text{and} \quad C = \frac{1}{m} \sum_{i=1}^m (x^i - \mu)(x^i - \mu)^T$$

Principal directions

- Step 2: Take the eigenvectors w^1, w^2, \dots of C corresponding to the largest eigenvalue λ_1 , the second largest eigenvalue λ_2 ...
- Step 3: Compute reduced representation

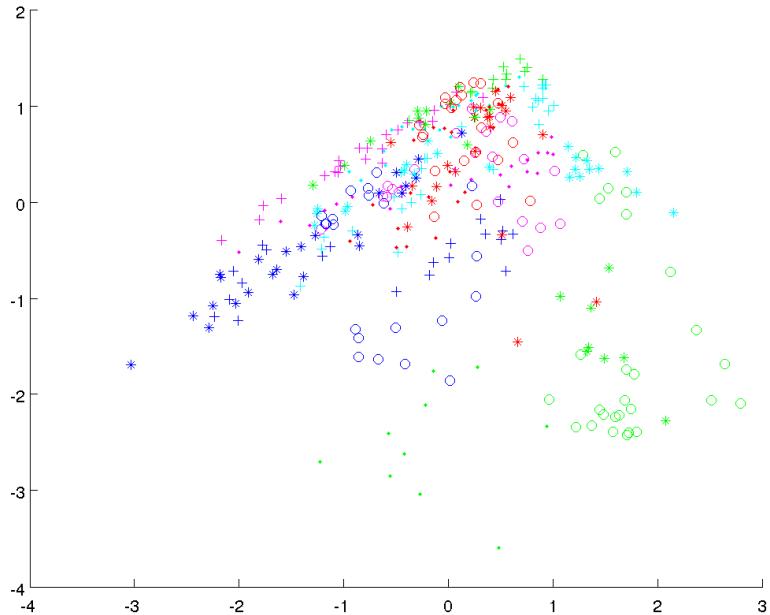
$$z^i = \begin{pmatrix} w^{1T}(x^i - \mu) / \sqrt{\lambda_1} \\ w^{2T}(x^i - \mu) / \sqrt{\lambda_2} \\ \vdots \end{pmatrix}$$

Normalize by
standard deviation

Look more into PCA_leaf.m



Interpreting the reduced representation



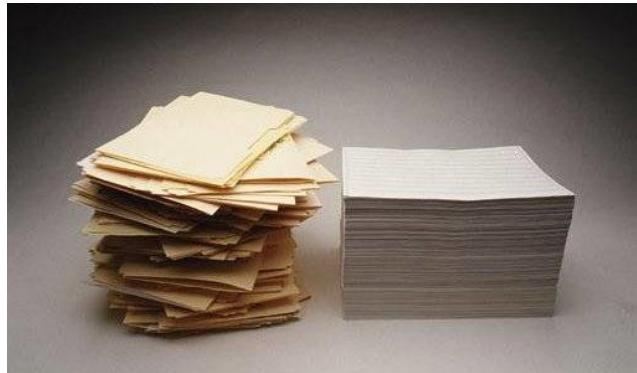
Texture
features

Principal direction:
 $W =$

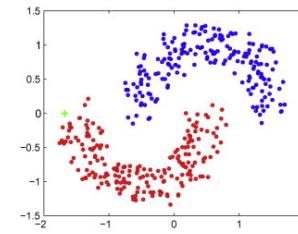
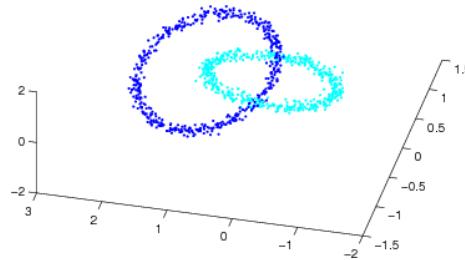
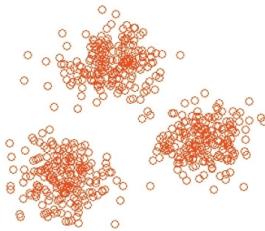
0.0938	0.1924
0.1902	0.0253
0.2266	-0.1800
-0.1850	0.4084
-0.1600	0.3825
-0.2063	0.3488
0.1940	-0.4037
0.2150	-0.3566
-0.3723	-0.2001
-0.3657	-0.1974
-0.3602	-0.2037
-0.3175	-0.1886
-0.3056	-0.1243
-0.3482	-0.1829

Shape
features

Documents collections

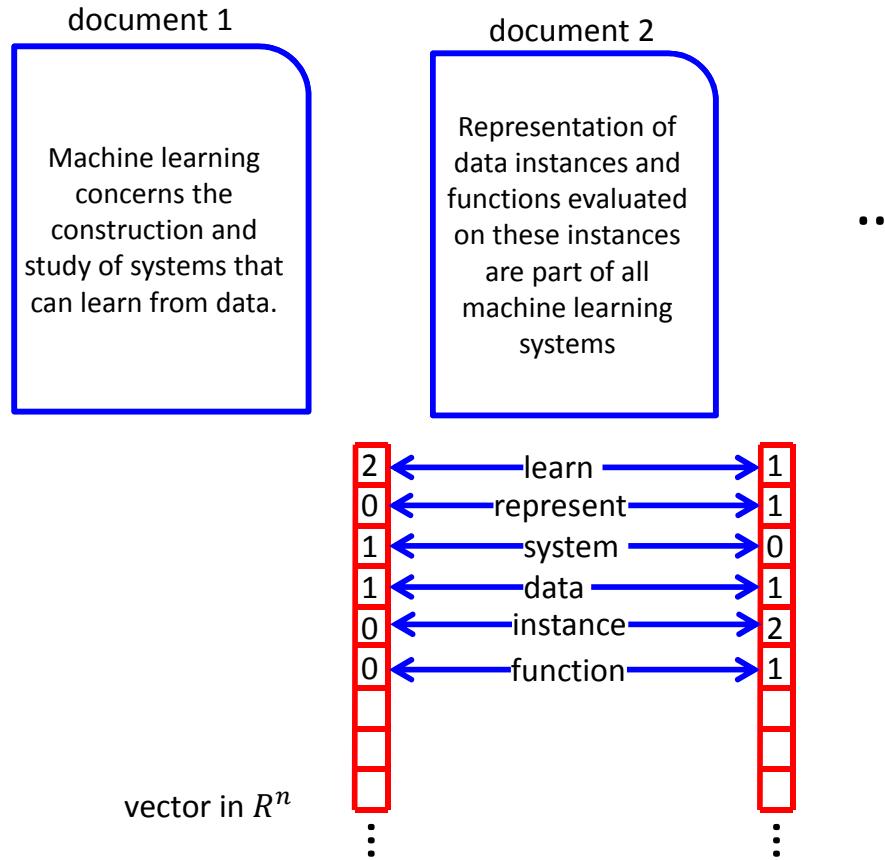


What are the relations
between data points?



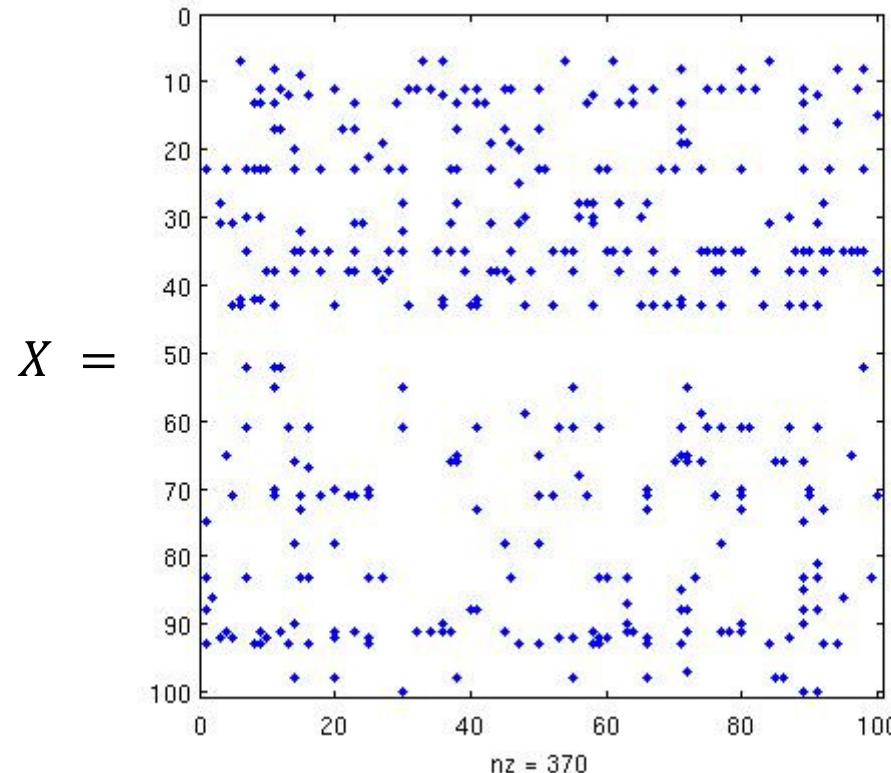
<http://submitpapers.com/misc/docCollection.html>

Bag of words representation



Experiments with 20 news groups

- Bag-of-words, or term-document matrix



Singular Value Decomposition

- Singular value decomposition, known as SVD, is a factorization of a real matrix with applications in calculating pseudo-inverse, rank, solving linear equations, and many others.
- For a matrix $M \in \mathbb{R}^{m \times n}$ assume $n \leq m$
 - $M = U\Sigma V^T$ where $U \in \mathbb{R}^{m \times m}$, $V^T \in \mathbb{R}^{n \times n}$, $\Sigma \in \mathbb{R}^{m \times n}$
 - The m columns of U , and the n columns of V are called the left and right singular vectors of M . The diagonal elements of Σ , Σ_{ii} are known as the singular values of M .
 - Let v be the i^{th} column of V , and u be the i^{th} column of U , and σ be the i^{th} diagonal element of Σ

$$Mv = \sigma u \quad \text{and} \quad M^T u = \sigma v$$

Singular Value Decomposition - II

$$\bullet M = [u_1 \ u_2 \ \dots \ u_n] \begin{bmatrix} \Sigma_{11} & & & 0 \\ \vdots & \ddots & & \vdots \\ 0 & & \ddots & \Sigma_{mn} \end{bmatrix} [v_1 \ v_2 \ \dots \ v_n]^T$$

principal directions

Scaling factor

Projection in principal directions

- Singular value decomposition is related to eigenvalue decomposition
 - Suppose $M = [x_1 - \mu \ x_2 - \mu \ \dots \ x_m - \mu] \in \mathbb{R}^{m \times n}$
 - Then covariance matrix is $C = \frac{1}{m} MM^T$
 - Starting from singular vector pair
 - $M^T u = \sigma v$
 - $\Rightarrow MM^T u = \sigma M v$
 - $\Rightarrow MM^T u = \sigma^2 u$
 - $\Rightarrow Cu = \lambda u$

