# Lecture 12. Principle Component Analysis

Xin Chen

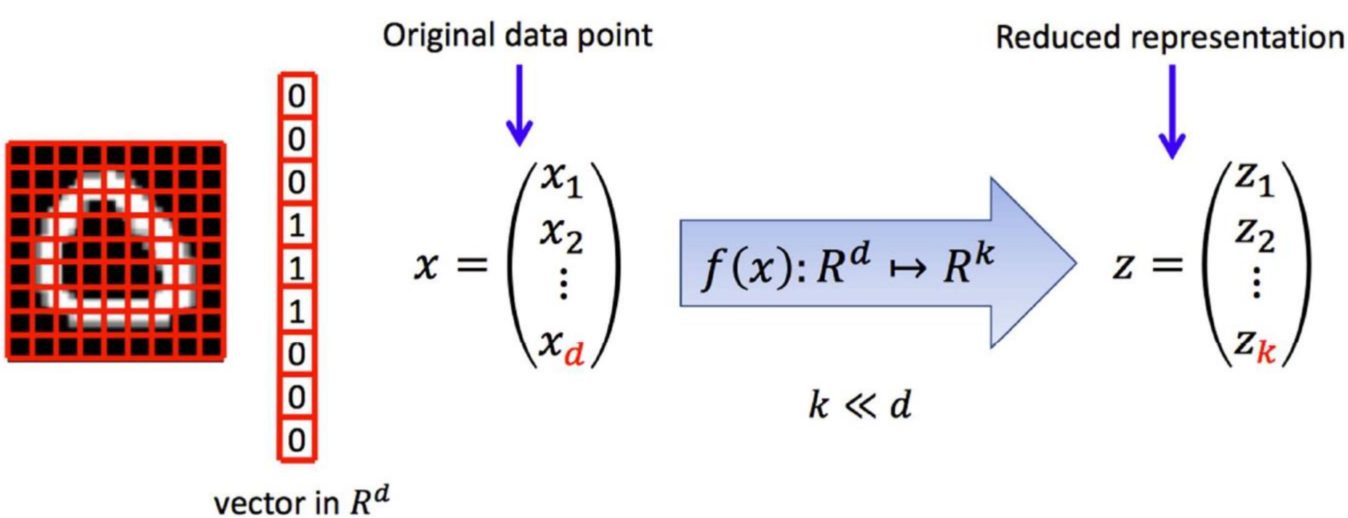
These slides are based on slides from Mahdi Roozbahani

# Outline

- Overview
- Main idea of Principle Component Analysis (PCA)
- PCA algorithm
- PCA and SVD
- Summary

# What is Dimension reduction?

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



# Applications of the dimension reduction

- 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 tasks
  - 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?

# Outline

- Overview
- Main idea of Principle Component Analysis (PCA)
- PCA algorithm
- PCA and SVD
- Summary

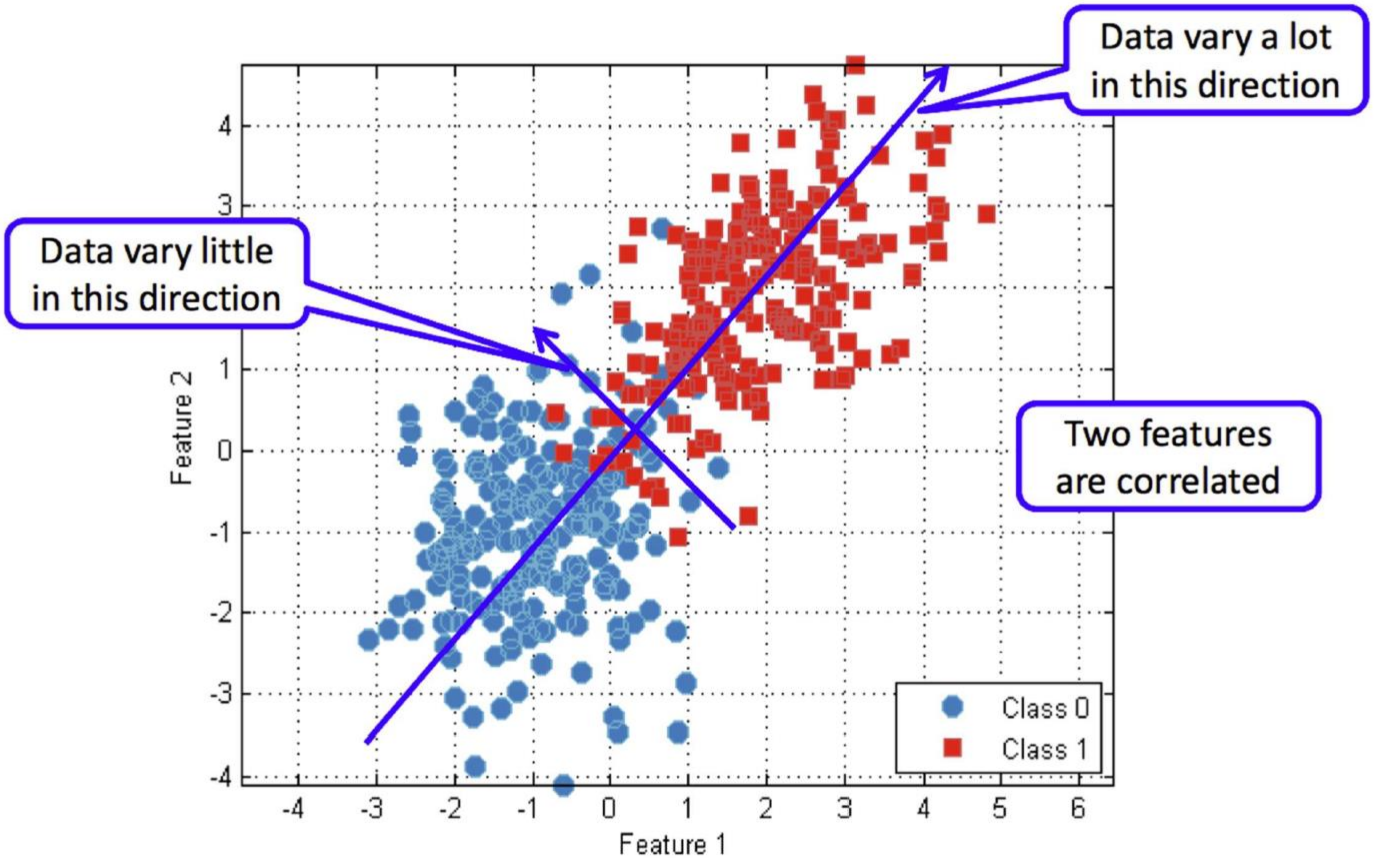
# PCA: Dimension reduction by capturing variation

- There are many criteria, geometric based, information theory based, etc.

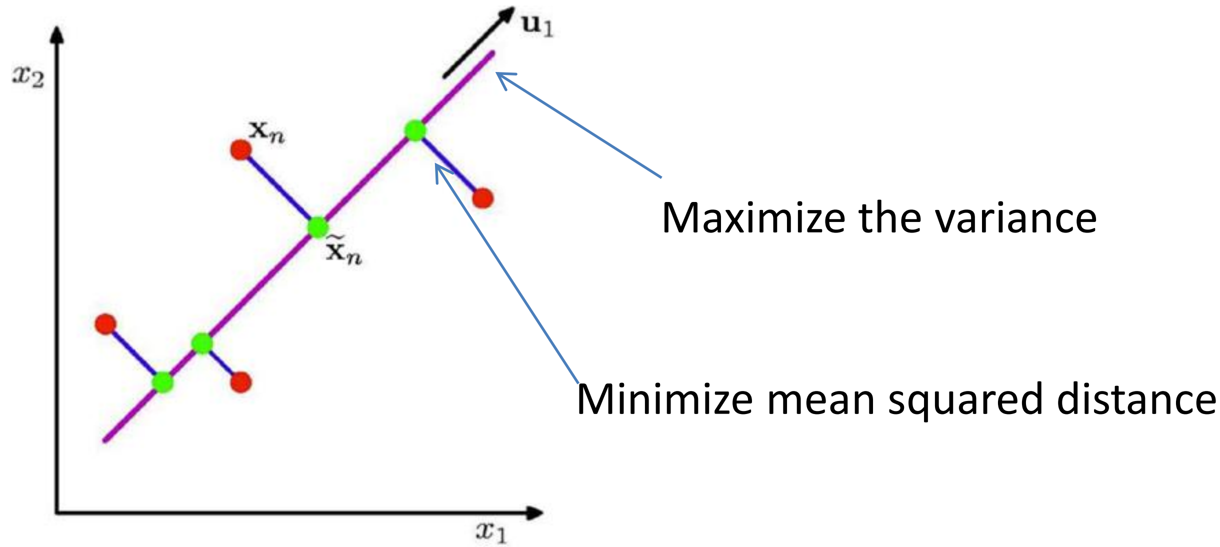
- One criterion: want to capture variation in data
  - Variations are “signals” or “useful” information in the data
  - Need to normalize each variable first

- In the process, also discover variables or dimensions highly correlated
  - Represent highly related phenomena
  - Combine them to form a stronger signal
  - Lead to simpler presentation

# Capture Variation in Data



# Two perspective of Principal Component Analysis (PCA)




- Orthogonal projection of the data onto a lower-dimension linear space that
  - Maximize variance of projected data
  - Minimize mean squared distance between the data points and projections.

# Outline

- Overview
- Main idea of Principle Component Analysis (PCA)
- PCA algorithm
- PCA and SVD
- Summary

# Formulating the problem

Given $n$ data points, $\{x_1, x_2, x_3, \ldots, x_n\} \in R^d$, with their mean $u = \frac{1}{n}\sum_{i=1}^{n} x_i$

Find a direction $\mathrm{w} \in R^d$, where $||\mathrm{w}|| = \sqrt{\sum_{j\in d} \omega_j{}^2} = 1$

We constrain the norm of $w$ to be equal to 1, to avoid having very large variance in each new dimension.

# Formulating the problem

Given $n$ data points, $\{x_1, x_2, x_3, \dots, x_n\} \in R^d$, with their mean $u = \frac{1}{n}\sum_{i=1}^{n} x_i$

$$||w|| = \sqrt{\sum_{j \in d} \omega_j{}^2} = 1$$

Optimization target: the variance of the data along direction $w$ is maximized. $max\,\frac{1}{n}\sum_{i=1}^{n}(x_i w - uw)^2$

Variance in new feature space.

# Formulate it as an optimization problem

Manipulate the objective with linear algebra

$$\frac{1}{n}\sum_{i=1}^{n}(x_i w - \mu w)^2 \quad = \frac{1}{n}\sum_{i=1}^{n}((x_i - \mu)w)^2 \; =$$

$$= \frac{1}{n}\sum_{i=1}^{n}\big((\underbrace{x_i - \mu}_{A})\underbrace{w}_{B}\big)^T((x_i - \mu)w) = \frac{1}{n}\sum_{i=1}^{n} w^T(x_i - \mu)^T\,(x_i - \mu)w$$

$$(AB)^T = B^T A^T$$

$$w^T \boxed{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^T\,(x_i - \mu)\right)} w \qquad = w^T C w$$

Covariance matrix

# Equivalence to the eigenvalue problem

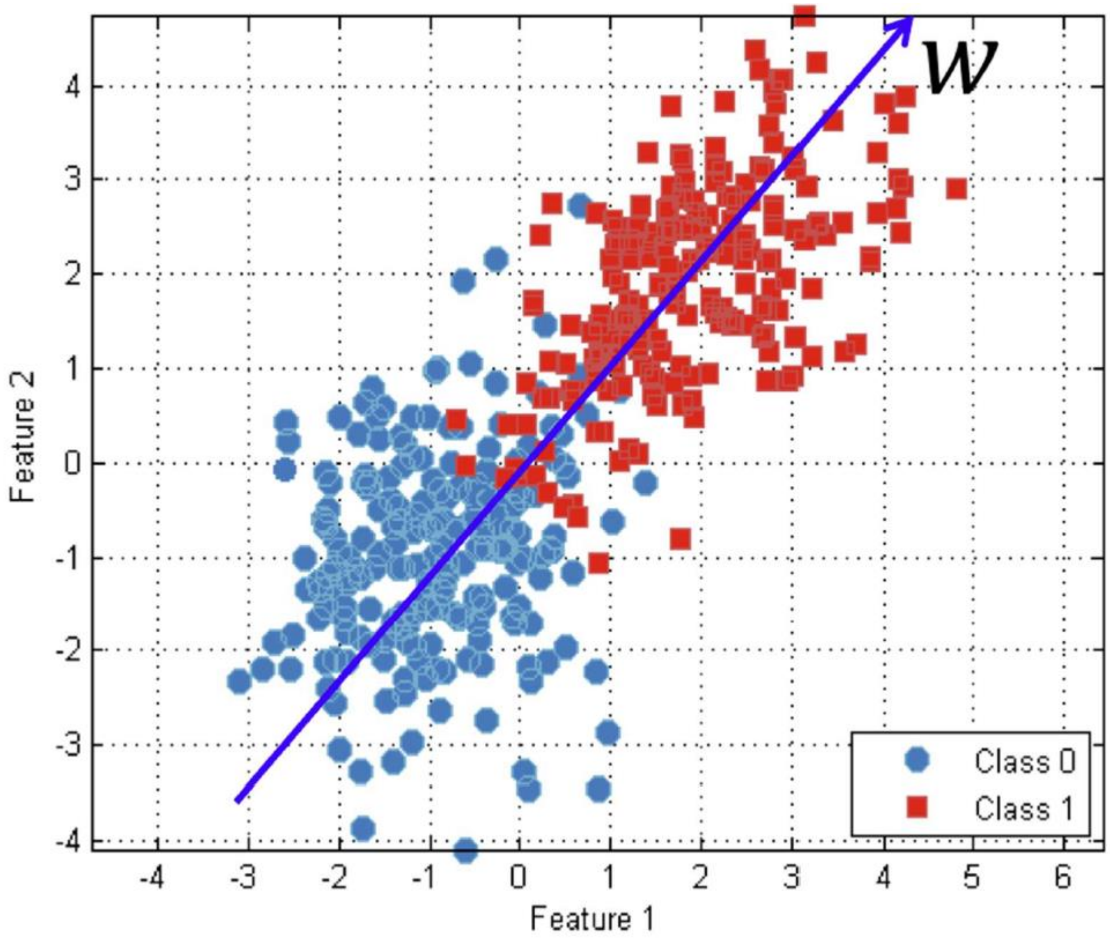
- Claim $\max w^T C w$
- Form lagrangian function of the optimization problem $L(w, \lambda) = w^T C w + \lambda(1 - w^T w)$
- If $w$ is a maximum of the original optimization problem, then there exists a $\lambda$, where $(\mathrm{w}, \lambda)$ is a stationary point of $\mathrm{L}(\mathrm{w}, \lambda)$
- This implies that $\frac{\partial L}{\partial w} = 0 = 2Cw - 2\lambda w \quad \Rightarrow \quad Cw = \lambda w$

# Eigen value problem

- Eigen-value problem
  - Given a symmetric matrix $C \in R^{d \times d}$
  - Find a vector w $\in R^d$ and $||w|| = 1$
  - Such that $Cw = \lambda w$

- There will be multiple solutions of the eigenvectors $w_1, w_2, \ldots$ of $C$ corresponding to the largest eigenvalue $\lambda_1, \lambda_2, \ldots, \lambda_d$
  - They are ortho-normal: ${w_i}^T w_i = 1, {w_i}^T w_j = 0$

# Principle direction of the data



# Variance in the principle direction

- Principle direction $w$ satisfies

$$Cw = \lambda w = w\lambda$$

- Variance in principle direction is

$$w^T C w = w^T w \lambda = \lambda$$

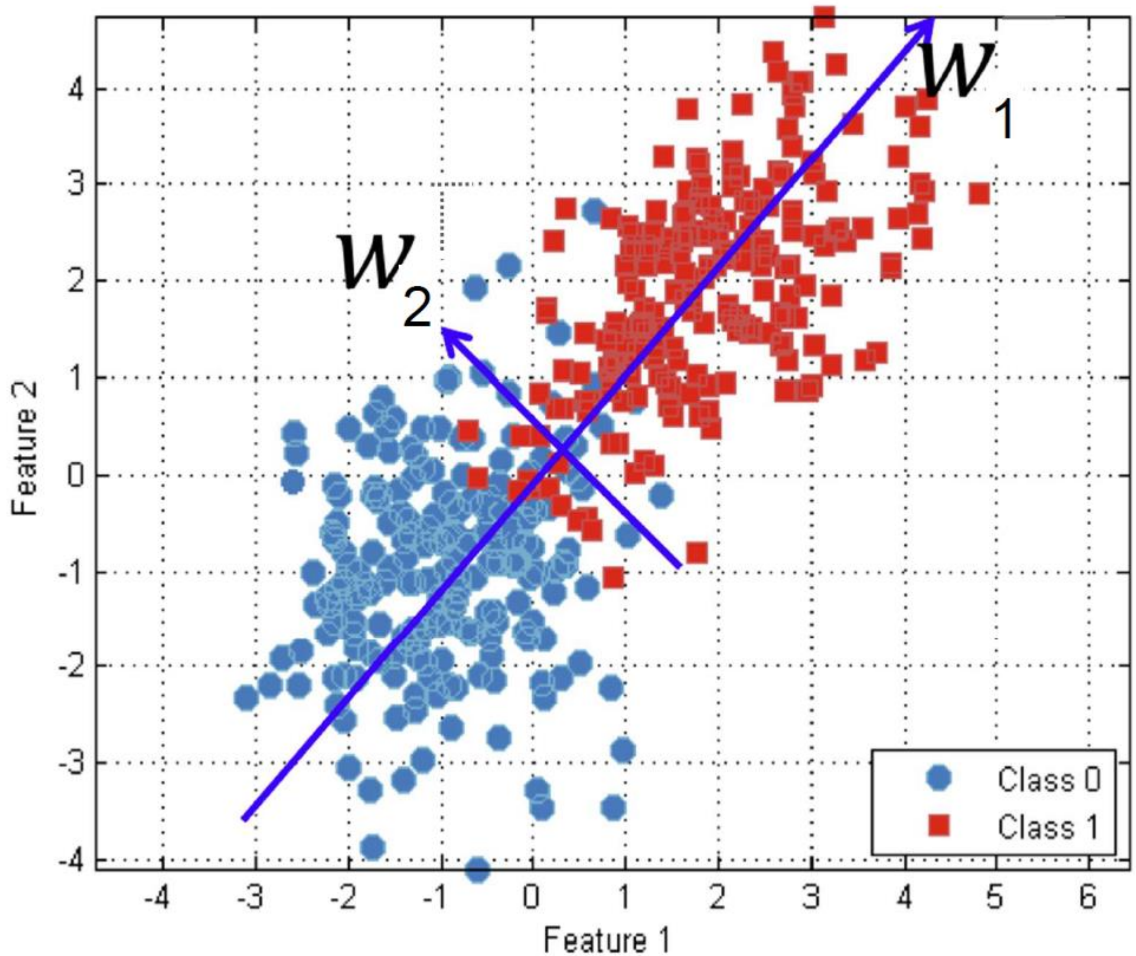
Eigen value

# Multiple principle directions

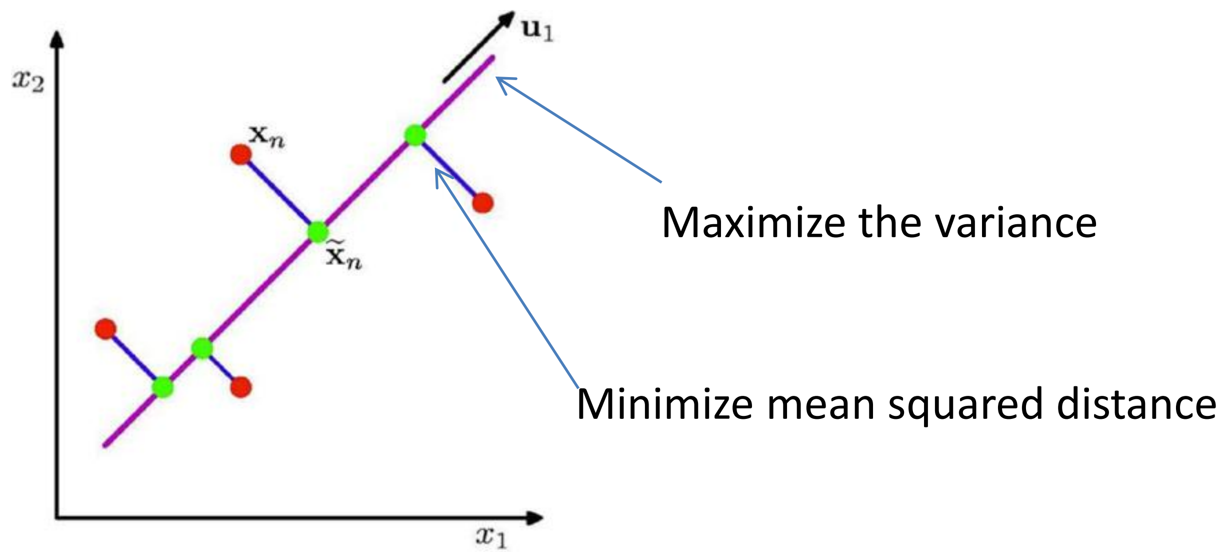
- Directions $w_1, w_2, \ldots$ which has
  - The largest variances
  - But are orthogonal to each other

- Take the eigenvectors $w_1, w_2, \ldots$ of $C$ corresponding to
  - The largest eigenvalue $\lambda_1$
  - The second largest eigenvalue $\lambda_2$
  - …

# Extra principle directions



# Remember the two perspectives




$$MSE(\vec{w}) = \frac{1}{n}\sum_{i=1}^{n} \|\vec{x_i}\|^2 - (\vec{w}\cdot\vec{x_i})^2$$

$$= \frac{1}{n}\left(\sum_{i=1}^{n} \|\vec{x_i}\|^2 - \sum_{i=1}^{n}(\vec{w}\cdot\vec{x_i})^2\right)$$

$$\frac{1}{n}\sum_{i=1}^{n}(\vec{w}\cdot\vec{x_i})^2 = \left(\frac{1}{n}\sum_{i=1}^{n}\vec{x_i}\cdot\vec{w}\right)^2 + \mathrm{Var}\left[\vec{w}\cdot\vec{x_i}\right]$$

# Relations between principle components

- Principle component #1: points in the direction of largest variance.

- Each subsequent principle component
  - Is orthogonal to the previous ones, and
  - Points in the directions of the largest variance of the residual subspace.

# The PCA algorithm

Given $n$ data points, $\{x_1, x_2, x_3, \ldots, x_n\} \in R^d$, with their mean $u = \frac{1}{n}\sum_{i=1}^{n} x_i$

Step 1: Estimate the mean and covariance matrix from data,
$C = \frac{1}{n}\sum_{i=1}^{n}(x_i - u)^T(x_i - u)$

Step 2: Take the eigenvectors $w_1, w_2, \ldots$ of $C$ corresponding to the largest eigenvalue $\lambda_1$, the second largest eigenvalue $\lambda_2$, …

Step 3: Compute reduced representation
$z_i = (\frac{(x_i - u_1)}{\sigma_1} w_1 \ \frac{(x_i - u_2)}{\sigma_2} w_2 \ldots)$

$z = n \times k$
$k < d$

# Outline

- Overview
- Main idea of Principle Component Analysis (PCA)
- PCA algorithm
- PCA and SVD
- Summary

# Singular Value Decomposition

$X_{n \times d}$

n: instances
d: dimensions
X is a centered matrix

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

$$U_{n \times n} \rightarrow unitary\ matrix \rightarrow U \times U^T = I$$

$$\Sigma_{n \times d} \rightarrow diagonal\ matrix$$

$$\mathrm{V}_{d \times d} \rightarrow unitary\ matrix \rightarrow V \times V^T = I$$

Principle direction

$$X = \underbrace{\begin{bmatrix} u_{1\times1} & \cdots & \cdots & \cdots & u_{1\times n} \\ \vdots & \ddots & \cdots & \cdots & \vdots \\ \vdots & \vdots & \ddots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \ddots & \vdots \\ u_{1\times1} & \cdots & \cdots & \cdots & u_{n\times n} \end{bmatrix}}_{U} \times \underbrace{\begin{bmatrix} \sum_{1\times1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sum_{d\times d} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}}_{\Sigma} \times \underbrace{\begin{bmatrix} \boxed{v_{1\times1} \quad \cdots \quad \cdots \quad \cdots \quad v_{1\times d}} \\ \vdots \quad \ddots \quad \cdots \quad \cdots \quad \vdots \\ \vdots \quad \vdots \quad \ddots \quad \cdots \quad \vdots \\ \vdots \quad \vdots \quad \cdots \quad \ddots \quad \vdots \\ v_{d\times1} \quad \cdots \quad \cdots \quad \cdots \quad v_{d\times d} \end{bmatrix}}_{V^T}$$

$$d < n$$

According to PCA $\Rightarrow Cw = \lambda w = w\lambda$

$$\text{Covariance } C_{d\times d} = \frac{1}{n}\sum_{i=1}^{n}\left(x^i - \mu\right)^T \underbrace{\left(x^i - \mu\right)}_{\text{Centering X}} = \frac{X^T X}{n}$$

$$\left.\begin{array}{l} X = U\Sigma V^T \\ \\ C = \dfrac{X^T X}{n} \end{array}\right\} C = \frac{V\Sigma^T U^T U\Sigma V^T}{n} = \frac{V\Sigma^2 V^T}{n}$$

$$C = \frac{V\Sigma^2 V^T}{n} = V\frac{\Sigma^2}{n}V^T$$

$$CV = V\frac{\Sigma^2}{n}V^TV = V\frac{\Sigma^2}{n}$$

According to Eigen-decomposition definition ➔ $CV = V\Lambda$

$V$ is the eigen vectors of covariance (Principal directions)

$\lambda_i = \frac{\sigma_i^2}{n}$ ➔ The eigenvalues of covariance matrix

Let's project the data (X) on principal directions:

$$XV = U\Sigma V^TV = U\Sigma$$

$\boldsymbol{XV}$ is independent linear combinations of the original data

Projection of one instance $(x)$ on the first principal direction using k dimensions

$$\mathrm{p}_1 = [u_{1\times1}\Sigma_{1\times1}\,,\, u_{1\times2}\Sigma_{2\times2}\,,\, \dots\,,\, u_{1\times k}\Sigma_{k\times k}]$$
$$\mathrm{p}_2 = [u_{2\times1}\Sigma_{1\times1}\,,\, u_{2\times2}\Sigma_{2\times2}\,,\, \dots\,,\, u_{2\times k}\Sigma_{k\times k}]$$

$U \Rightarrow n \times k$

$\Sigma \Rightarrow k \times k$

Upper left corner

Eigen values $\lambda = \frac{\Sigma^2}{m}$

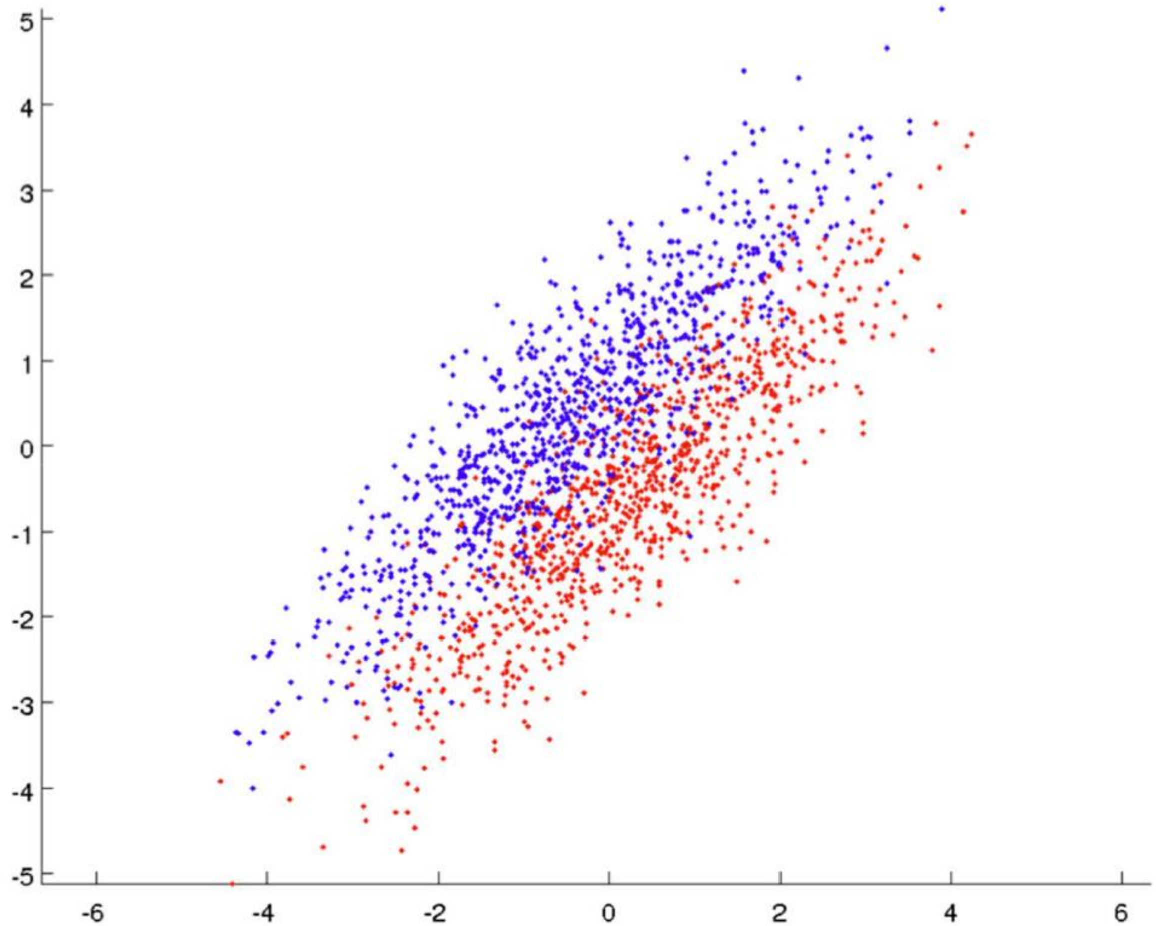
Eigenvectors (principal directions) V

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

Principal components (Scores) or projections on principal directions

- In fact, using the SVD to perform PCA makes better sense numerically than performing the covariance matrix, since the calculating $x^T x$ can cause loss of precision.

# Are principal components good for classification?



# Why PCA potentially works in classification?

- The dimension with the largest variance corresponds to the dimension and thus encodes the most information (information theory).
- The smallest eigenvectors often simply represent noise components.