Advanced Data Analysis: Principal Component Analysis

Masashi Sugiyama (Computer Science)

W8E-505, sugi@cs.titech.ac.jp

http://sugiyama-www.cs.titech.ac.jp/~sugi

Curse of Dimensionality

$$\{\boldsymbol{x}_i\}_{i=1}^n, \quad \boldsymbol{x}_i \in \mathbb{R}^d, \quad d \gg 1$$

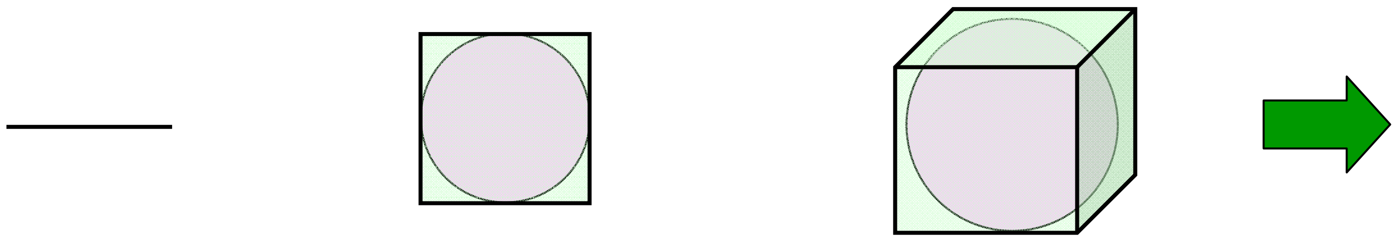
- If your data samples are high-dimensional, they are often too complex to directly analyze.
- Usual geometric intuitions are often only applicable to low-dimensional spaces; such intuitions could be even misleading in high-dimensional spaces.

Curse of Dimensionality (cont.)

- When the dimensionality increases,
  - Volume of unit hyper-cube V_c is always 1.
  - Volume of inscribed hyper-sphere V_s goes to 0.
- Relative size of hyper-sphere gets small!

$$\frac{V_s}{V_c} \to 0$$

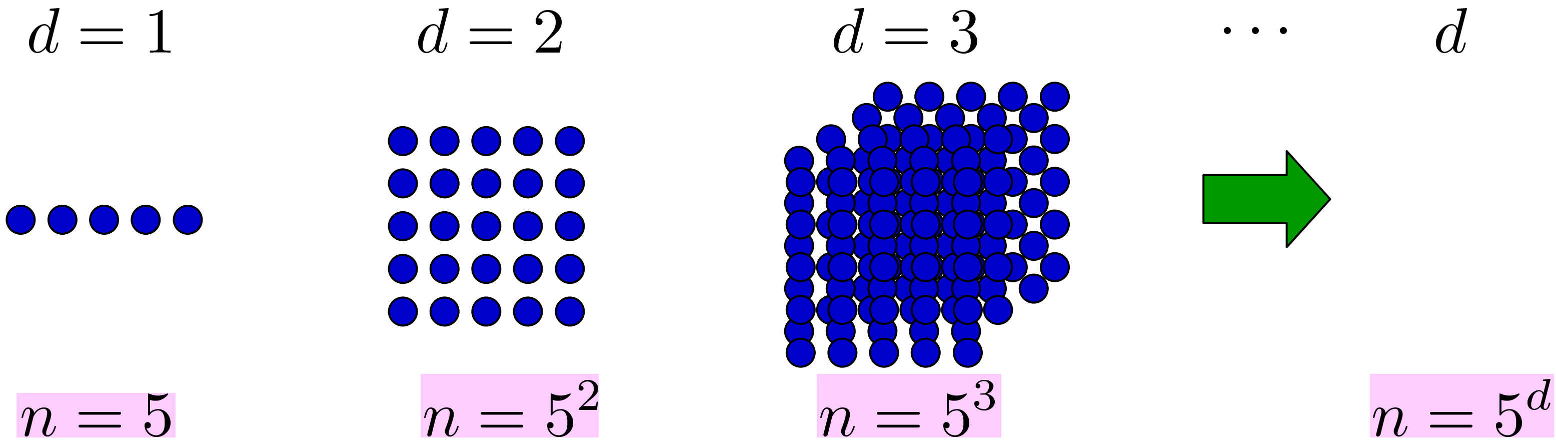
$d = 1$ $\quad$ $d = 2$ $\quad$ $d = 3$ $\quad \cdots \quad$ $d = \infty$



1 $\quad$ $\pi(0.5)^2 \approx 0.79$ $\quad$ $4\pi(0.5)^3/3 \approx 0.52$ $\quad$ 0

Curse of Dimensionality (cont.)

- Grid sampling requires an exponentially large number.



- Unless you have an exponentially large number of samples, your high-dimensional samples are never dense.

Dimensionality Reduction

- We want to reduce the dimensionality of the data while preserving the intrinsic "information" in the data.
- Dimensionality reduction is also called embedding; if the dimension is reduced up to 3, it is also called data visualization.
- Basic assumption (or belief) behind dimensionality reduction: your high-dimensional data is redundant in some sense.

Notation: Linear Embedding

- Data samples:

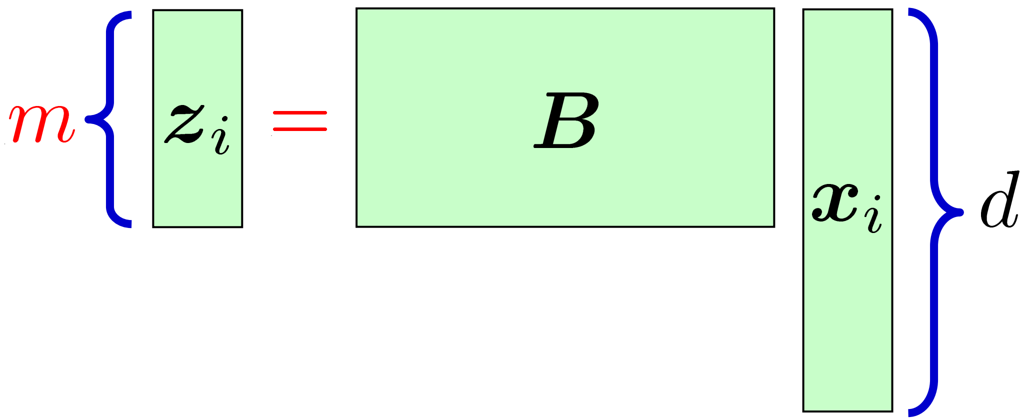
$$\{\boldsymbol{x}_i\}_{i=1}^n, \quad \boldsymbol{x}_i \in \mathbb{R}^d, \quad d \gg 1$$

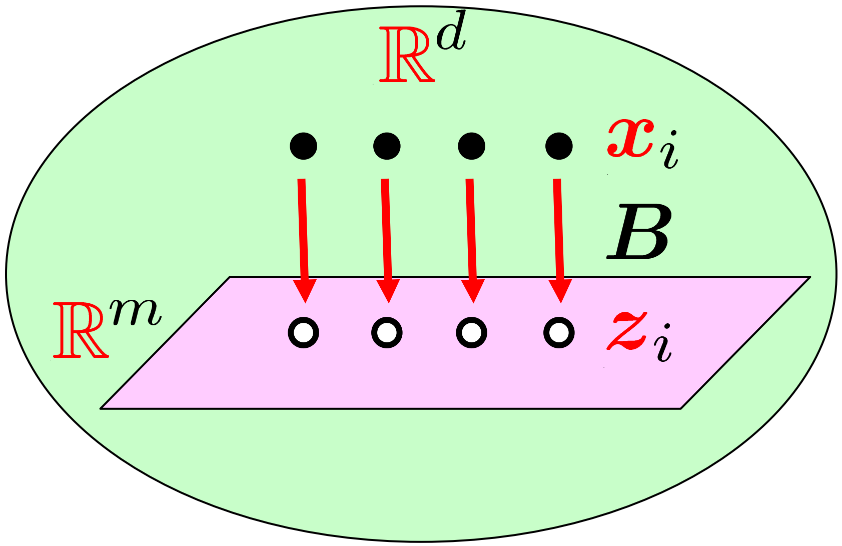
- Embedding matrix:

$$\boldsymbol{B} \in \mathbb{R}^{m \times d}, \quad 1 \leq m \ll d$$

- Embedded data samples:

$$\{\boldsymbol{z}_i\}_{i=1}^n, \quad \boldsymbol{z}_i = \boldsymbol{B}\boldsymbol{x}_i \in \mathbb{R}^m$$



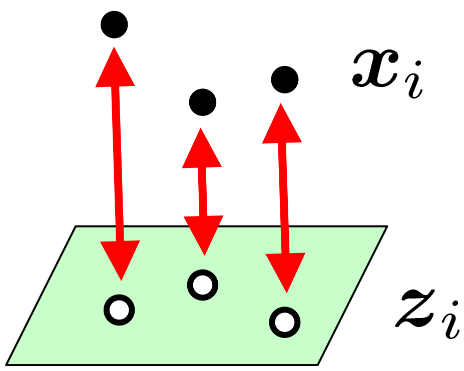


Principal Component Analysis (PCA)

- Idea: We want to get rid of a redundant dimension of the data samples

$$\begin{pmatrix}1\\0\end{pmatrix}, \begin{pmatrix}2\\0.1\end{pmatrix}, \begin{pmatrix}3\\-0.1\end{pmatrix}$$

- This could be achieved by minimizing the distance between embedded samples and original samples.



Data Centering

- We center the data samples by

$$\overline{\boldsymbol{x}}_i = \boldsymbol{x}_i - \frac{1}{n}\sum_{j=1}^{n}\boldsymbol{x}_j$$

$$\frac{1}{n}\sum_{i=1}^{n}\overline{\boldsymbol{x}}_i = 0$$

- In matrix,

$$\overline{\boldsymbol{X}} = \boldsymbol{X}\boldsymbol{H}$$

$$\overline{\boldsymbol{X}} = (\overline{\boldsymbol{x}}_1 | \overline{\boldsymbol{x}}_2 | \cdots | \overline{\boldsymbol{x}}_n)$$

$$\boldsymbol{X} = (\boldsymbol{x}_1 | \boldsymbol{x}_2 | \cdots | \boldsymbol{x}_n)$$

$$\boldsymbol{H} = \boldsymbol{I}_n - \frac{1}{n}\boldsymbol{1}_{n\times n}$$

$\boldsymbol{I}_n$: n-dimensional identity matrix

$\boldsymbol{1}_{n\times n}$: $n \times n$ matrix with all ones

Orthogonal Projection

- $\{\boldsymbol{b}_i \ (\in \mathbb{R}^d)\}_{i=1}^m$: Orthonormal basis in m -dimensional embedding subspace

$$\langle \boldsymbol{b}_i, \boldsymbol{b}_j \rangle = \delta_{i,j} = \begin{cases} 1 & (i = j) \\ 0 & (i \neq j) \end{cases}$$

- In matrix, $\boldsymbol{B}\boldsymbol{B}^\top = \boldsymbol{I}_m$

$$\boldsymbol{B} = (\boldsymbol{b}_1 | \boldsymbol{b}_2 | \cdots | \boldsymbol{b}_m)^\top$$

- Orthogonal projection of $\overline{\boldsymbol{x}}_i$ is expressed by

$$\sum_{j=1}^m \langle \boldsymbol{b}_j, \overline{\boldsymbol{x}}_i \rangle \boldsymbol{b}_j \quad \left(= \boldsymbol{B}^\top \boldsymbol{B} \overline{\boldsymbol{x}}_i \right)$$

PCA Criterion

- Minimize the sum of squared distances.

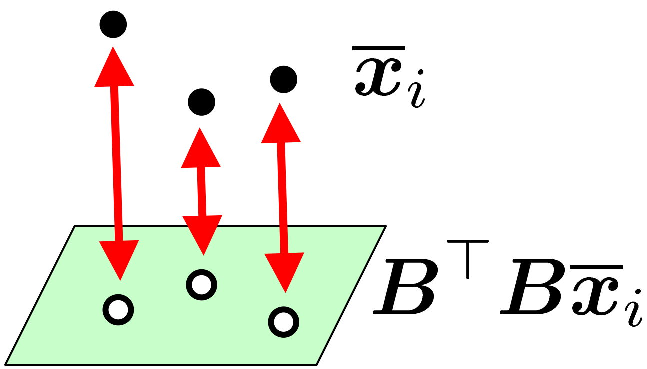
$$\sum_{i=1}^{n} \|\boldsymbol{B}^\top \boldsymbol{B}\overline{\boldsymbol{x}}_i - \overline{\boldsymbol{x}}_i\|^2 \quad \left(= -\mathrm{tr}(\boldsymbol{B}\overline{\boldsymbol{C}}\boldsymbol{B}^\top) + \mathrm{tr}(\overline{\boldsymbol{C}})\right)$$

$$\overline{\boldsymbol{C}} = \sum_{i=1}^{n} \overline{\boldsymbol{x}}_i \overline{\boldsymbol{x}}_i^\top = \overline{\boldsymbol{X}}\,\overline{\boldsymbol{X}}^\top$$

- PCA criterion:

$$\boldsymbol{B}_{PCA} = \underset{\boldsymbol{B}\in\mathbb{R}^{m\times d}}{\mathrm{argmax}}\, \mathrm{tr}(\boldsymbol{B}\overline{\boldsymbol{C}}\boldsymbol{B}^\top)$$

$$\text{subject to } \boldsymbol{B}\boldsymbol{B}^\top = \boldsymbol{I}_m$$



PCA: Summary

- A PCA solution:

$$\boldsymbol{B}_{PCA} = (\boldsymbol{\psi}_1 | \boldsymbol{\psi}_2 | \cdots | \boldsymbol{\psi}_m)^\top$$

$\{\lambda_i, \boldsymbol{\psi}_i\}_{i=1}^m$:Sorted eigenvalues and normalized eigenvectors of $\overline{\boldsymbol{C}}\boldsymbol{\psi} = \lambda\boldsymbol{\psi}$

$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$ $\quad\quad$ $\langle \boldsymbol{\psi}_i, \boldsymbol{\psi}_j \rangle = \delta_{i,j}$

- PCA embedding of a sample $\boldsymbol{x}$:

$$\overline{\boldsymbol{z}} = \boldsymbol{B}_{PCA}(\boldsymbol{x} - \frac{1}{n}\boldsymbol{X}\boldsymbol{1}_n)$$

$\boldsymbol{1}_n$: n-dimensional vector with all ones

Proof

- Lagrangian:

$$L(\boldsymbol{B},\boldsymbol{\Delta}) = \mathrm{tr}(\boldsymbol{B}\overline{\boldsymbol{C}}\boldsymbol{B}^\top) - \mathrm{tr}((\boldsymbol{B}\boldsymbol{B}^\top - \boldsymbol{I}_m)\boldsymbol{\Delta})$$

$\boldsymbol{\Delta}$:Lagrange multipliers (symmetric)

- Stationary point (necessary condition):
  - $$\frac{\partial L}{\partial \boldsymbol{B}} = 2\boldsymbol{B}\overline{\boldsymbol{C}} - 2\boldsymbol{\Delta}\boldsymbol{B} = 0$$

    → $\overline{\boldsymbol{C}}\boldsymbol{B}^\top = \boldsymbol{B}^\top\boldsymbol{\Delta}$ (1)
  - $$\frac{\partial L}{\partial \boldsymbol{\Delta}} = \boldsymbol{B}\boldsymbol{B}^\top - \boldsymbol{I}_m = 0$$

    → $\boldsymbol{B}\boldsymbol{B}^\top = \boldsymbol{I}_m$ (2)
- Eigendecomposition:

$\boldsymbol{\Delta} = \boldsymbol{T}\boldsymbol{\Gamma}\boldsymbol{T}^\top$ (3)

$\boldsymbol{T}$: orthogonal matrix
$\boldsymbol{\Gamma}$: diagonal matrix

$\boldsymbol{T}^{-1} = \boldsymbol{T}^\top$

Proof (cont.)

- (1) & (3) ⇒ $\overline{\boldsymbol{C}}\boldsymbol{B}^\top = \boldsymbol{B}^\top\boldsymbol{T}\boldsymbol{\Gamma}\boldsymbol{T}^\top$ (4)

  ⇒ $\overline{\boldsymbol{C}}\boldsymbol{B}^\top\boldsymbol{T} = \boldsymbol{B}^\top\boldsymbol{T}\boldsymbol{\Gamma}$

  ⇒ $\overline{\boldsymbol{C}}\boldsymbol{F} = \boldsymbol{F}\boldsymbol{\Gamma}$ (5) $\boldsymbol{F} = \boldsymbol{B}^\top\boldsymbol{T}$
- (5) is an eigensystem

  ⇒ $\mathcal{R}(\boldsymbol{F}) = \mathrm{span}(\{\boldsymbol{\psi}_{k_i}\}_{i=1}^m)$ (6)

  $\boldsymbol{\Gamma} = \mathrm{diag}(\lambda_{k_1}, \lambda_{k_2}, \ldots, \lambda_{k_m})$ (7)

  $k_i \in \{1, 2, \ldots, d\}$
- $\mathcal{R}(\boldsymbol{F}) = \mathcal{R}(\boldsymbol{B}^\top\boldsymbol{T}) = \mathcal{R}(\boldsymbol{B}^\top)$ (8)
- (6) & (8) ⇒ $\mathcal{R}(\boldsymbol{B}^\top) = \mathrm{span}(\{\boldsymbol{\psi}_{k_i}\}_{i=1}^m)$
- A solution is expressed as

$$\boldsymbol{B} = (\boldsymbol{\psi}_{k_1}|\boldsymbol{\psi}_{k_2}|\cdots|\boldsymbol{\psi}_{k_m})^\top$$

Proof (cont.)

- (2) ⇒ $\mathrm{rank}(\boldsymbol{B}) = m$
  ⇒ all $\{k_i\}_{i=1}^{m}$ are distinct
- We should choose the best $\{k_i\}_{i=1}^{m}$ that maximizes $\mathrm{tr}(\boldsymbol{B}\overline{\boldsymbol{C}}\boldsymbol{B}^\top)$.
- (4) & (7) ⇒

$$\begin{aligned}\mathrm{tr}(\boldsymbol{B}\overline{\boldsymbol{C}}\boldsymbol{B}^\top) &= \mathrm{tr}(\boldsymbol{B}\boldsymbol{B}^\top\boldsymbol{T}\boldsymbol{\Gamma}\boldsymbol{T}^\top) \\ &= \mathrm{tr}(\boldsymbol{T}\boldsymbol{\Gamma}\boldsymbol{T}^\top) \\ &= \mathrm{tr}(\boldsymbol{\Gamma}\boldsymbol{T}^\top\boldsymbol{T}) \\ &= \sum_{i=1}^{m}\lambda_{k_i}\end{aligned}$$

- $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$
  ⇒ $k_i = i$ gives a solution.

(Q.E.D.)

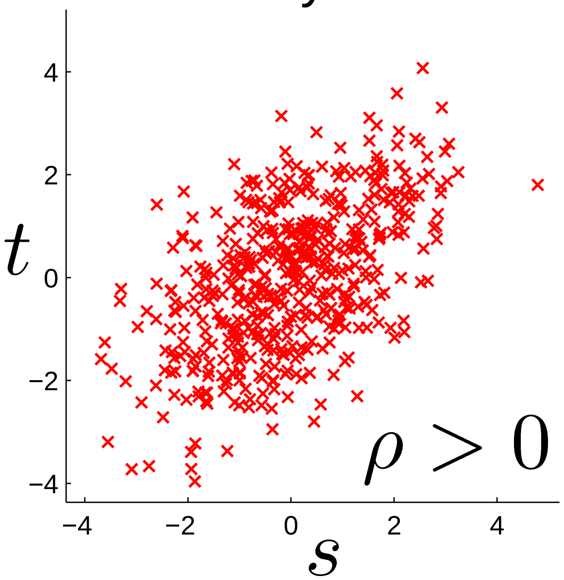
Correlation

- Correlation coefficient for $\{s_i, t_i\}_{i=1}^{n}$:

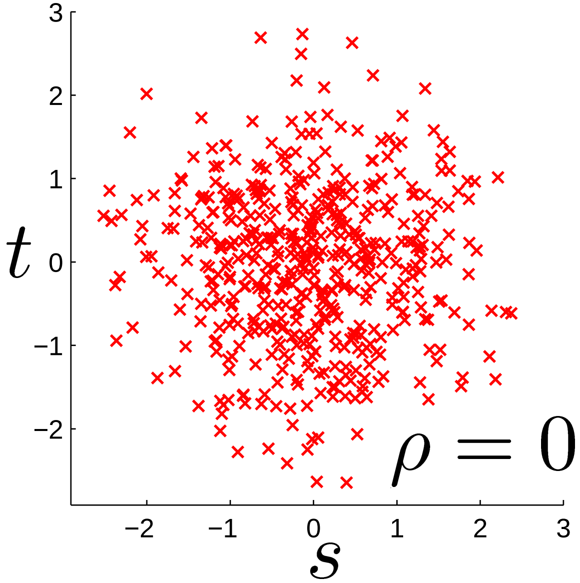
$$\rho = \frac{\sum_{i=1}^{n}(s_i - \overline{s})(t_i - \overline{t})}{\sqrt{\left(\sum_{i=1}^{n}(s_i - \overline{s})^2\right)\left(\sum_{i=1}^{n}(t_i - \overline{t})^2\right)}}$$

$$\overline{s} = \sum_{i=1}^{n} s_i \qquad \overline{t} = \sum_{i=1}^{n} t_i$$

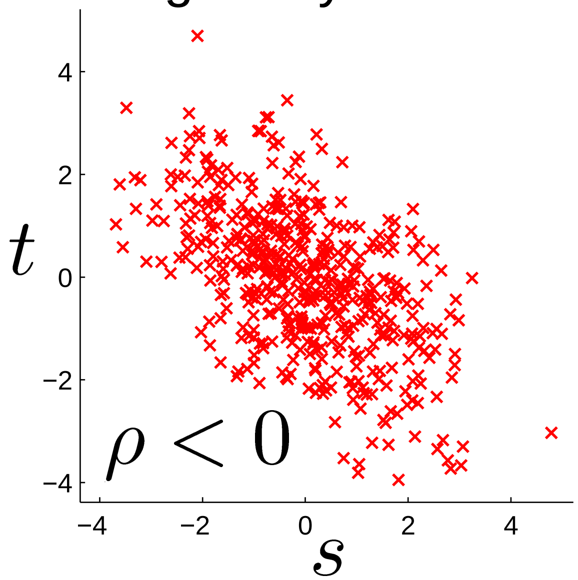
Positively correlated



Uncorrelated



Negatively correlated



PCA Uncorrelates Data

$$\boldsymbol{B}_{PCA} = (\boldsymbol{\psi}_1 | \boldsymbol{\psi}_2 | \cdots | \boldsymbol{\psi}_m)^\top$$

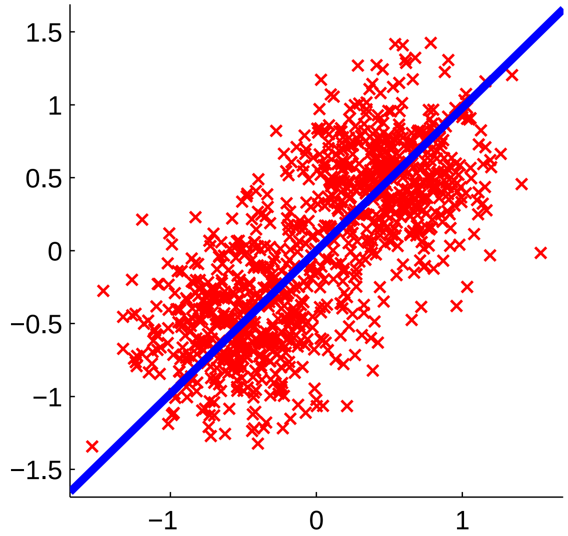
- Covariance matrix of the PCA-embedded samples is diagonal.

$$\frac{1}{n}\sum_{i=1}^{n} \overline{\boldsymbol{z}}_i \overline{\boldsymbol{z}}_i^\top = \mathrm{diag}\left(\lambda_1, \lambda_2, \ldots, \lambda_m\right)$$

(Homework)

→ Each element in $\overline{\boldsymbol{z}}$ is uncorrelated!

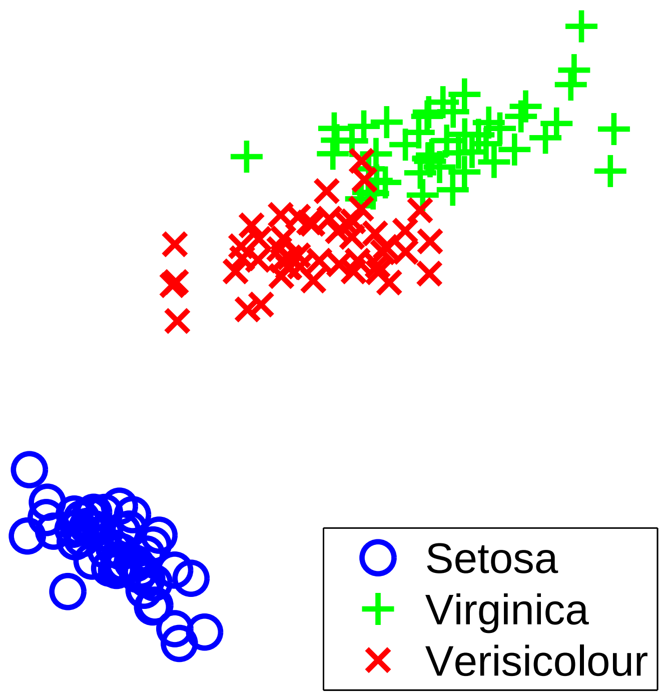
Examples



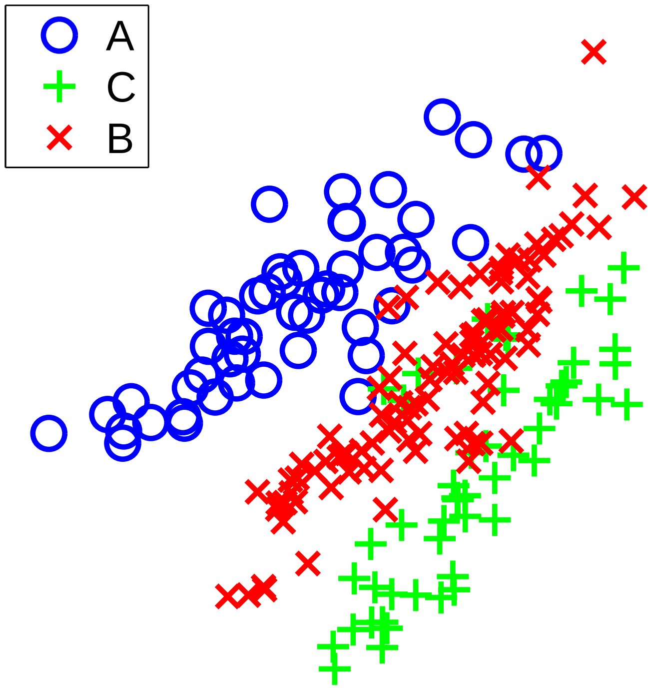
- Data is well described
- PCA is intuitive, easy to implement, analytic solution available, and fast.

Examples (cont.)

Iris data (4d->2d)

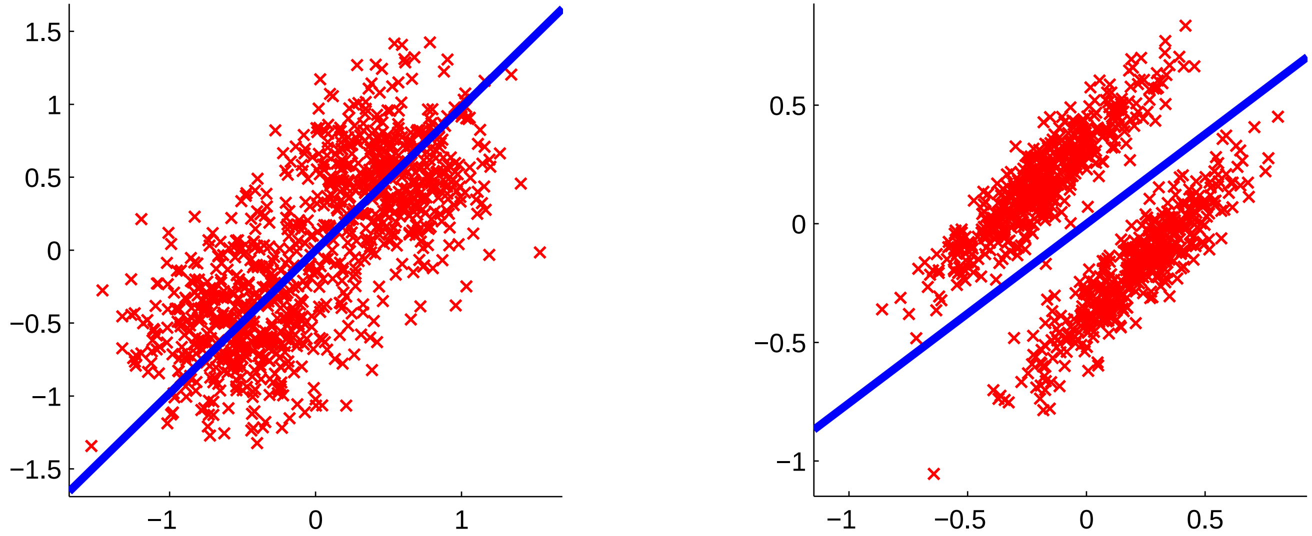


Letter data (16d->2d)



- Embedded samples seem informative.

Examples (cont.)



- However, PCA does not necessarily preserve interesting information such as clusters.

Important Notice

- There will be no class on May 3rd.
- Instead, you are expected to do some exercise (at home).
- Information will be put on the web.
  - See http://www.ocw.titech.ac.jp/ for details.