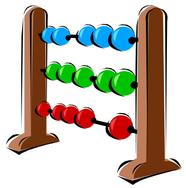
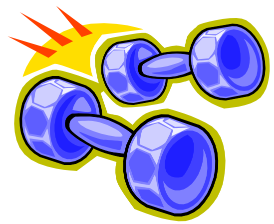
# Summary

- There are 3 topics in learning resea

  - Understanding human brains
  - Developing learning machines
  - Clarifying essence of learning mathematically

- There are 3 types of learning.

  - Supervised learning
  - Unsupervised learning
  - Reinforcement learning

- Topics of unsupervised learning:

  - Dimensionality reduction
  - Data clustering
  - Blind source separation
  - Outlier/novelty detection


# Dimensionality Reduction

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

- High-dimensional data is too complex to analyze: ➡ "Curse of dimensionality"
- We want to reduce the dimensionality of the data while preserving the intrinsic "information" in the data.
- Dimensionality reduction is also called
  - Embedding
  - Data visualization (if the dimension is reduced up to 3)

# What Kind of “Information” Do We Want to Preserve?

- There are several criteria: For example, embed the data such that
  - Original data is preserved
  - Local structure is maintained
  - Interesting components are extracted
- Which criterion is the best?
- It depends on the data, purpose…

# Methods to be Dealt With

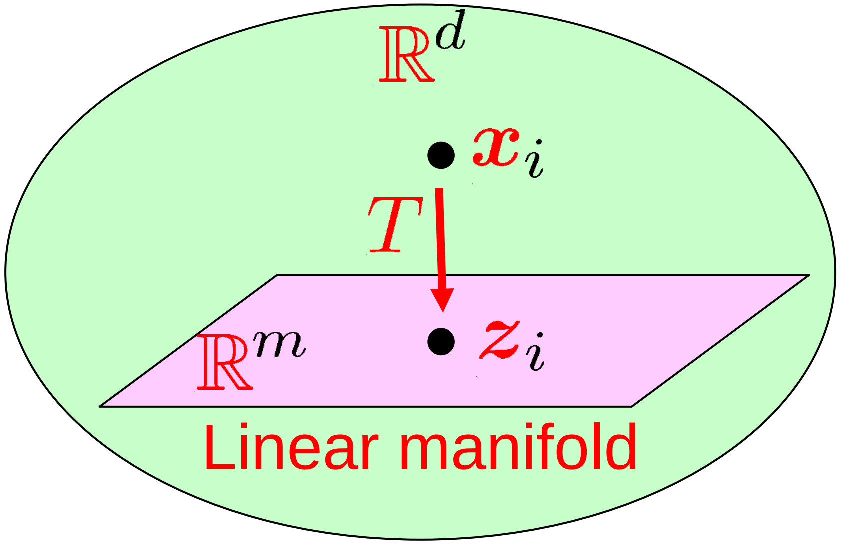
- Original data is preserved:
  - Principal component analysis (PCA)
  - Kernel PCA
- Local structure is maintained:
  - Locality preserving projection
  - Laplacian eigenmap embedding
- Interesting components are extracted:
  - Projection pursuit
  - Non-Gaussian component analysis

# Notation

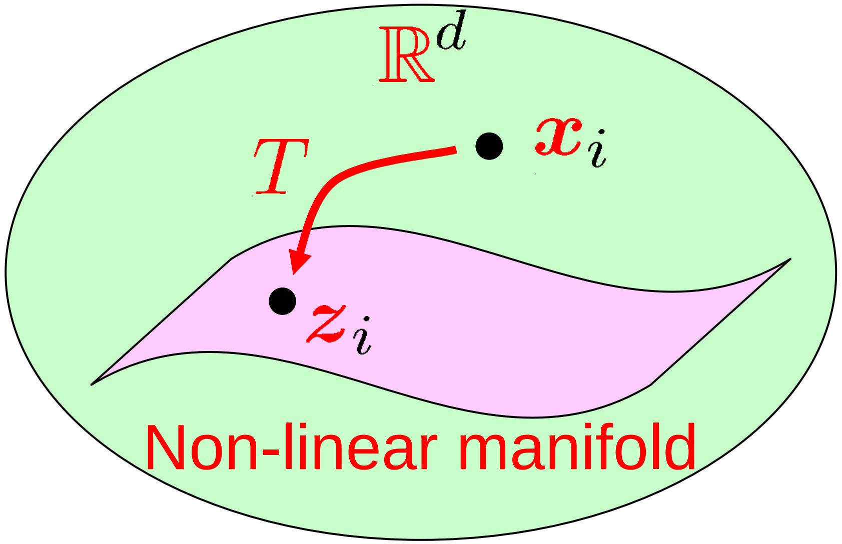
- Data: $\{\boldsymbol{x}_i\}_{i=1}^{n}, \quad \boldsymbol{x}_i \in \mathbb{R}^d, \quad d \gg 1$
- Mapping: $T : \mathbb{R}^d \rightarrow \mathbb{R}^m, \quad 1 \leq m \ll d$

  (Note: $T$ is generally non-linear)
- Embedded data: $\{\boldsymbol{z}_i\}_{i=1}^{n}, \quad \boldsymbol{z}_i = T\boldsymbol{x}_i \in \mathbb{R}^m$

Linear embedding




Non-linear embedding



# Preserving Original Data

- We want to embed the data so that embedded data is as "close" to the original data as possible.
- $d = 2,\ m = 1$

$$\{\boldsymbol{x}_i\}_{i=1}^n = \left\{ \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 0.1 \\ 2 \end{pmatrix}, \begin{pmatrix} -0.1 \\ 3 \end{pmatrix} \right\}$$

- For better description of data, throwing the first element away would be good.
- How about

$$\{\boldsymbol{x}_i\}_{i=1}^n = \left\{ \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \begin{pmatrix} 2 \\ 4.1 \end{pmatrix}, \begin{pmatrix} 3 \\ 5.9 \end{pmatrix} \right\}$$

# Principal Component Analysis

- Principal component analysis (PCA) tries to automatically find the best description of data.
- Assume
  - Mapping $T: \boldsymbol{x}_i \longrightarrow \boldsymbol{z}_i$ is linear.
  - Data is centered:

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

- Linear mapping is essentially a projection onto a subspace.
- Find the subspace which best describes data!

# PCA Criterion

- Projection: $\widetilde{\boldsymbol{x}}_i = \boldsymbol{P}\boldsymbol{x}_i$
- “Closeness” :Squared Euclidean distance

$$\|\widetilde{\boldsymbol{x}}_i - \boldsymbol{x}_i\|^2$$

- PCA criterion:

$$\min_{\boldsymbol{P} \in \mathcal{P}_m} \left[ \sum_{i=1}^{n} \|\widetilde{\boldsymbol{x}}_i - \boldsymbol{x}_i\|^2 \right]$$

- $\mathcal{P}_m$:Set of all orthogonal projection matrices with rank $m$

# How to Obtain Solution

- $\{\boldsymbol{b}_i\}_{i=1}^m$ : Orthonormal basis in a subspace

$$\boldsymbol{P} = \sum_{i=1}^m \boldsymbol{b}_i \boldsymbol{b}_i^\top = \boldsymbol{B}^\top \boldsymbol{B}$$

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

- PCA criterion is equivalently expressed as

$$\boldsymbol{B}_{PCA} = \underset{\boldsymbol{B} \in \mathbb{R}^{m \times d}}{\mathrm{argmin}} \left[ J_{PCA}(\boldsymbol{B}) \right]$$

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

$$J_{PCA}(\boldsymbol{B}) = \sum_{i=1}^n \|\boldsymbol{B}^\top \boldsymbol{B} \boldsymbol{x}_i - \boldsymbol{x}_i\|^2$$

# How to Obtain Solution (cont.)

- $J_{PCA}(\boldsymbol{B}) = -\mathrm{tr}(\boldsymbol{B}^\top \boldsymbol{B}\boldsymbol{C}) + \mathrm{tr}(\boldsymbol{C})$

$$\boldsymbol{C} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top$$

- PCA criterion is equivalently expressed as

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

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

# Lemma

- $\boldsymbol{B}$ : $m \times d, (1 \leq m \leq d)$
- $\boldsymbol{C}$ : $d \times d$, positive, symmetric
- Problem:

$$\boldsymbol{B}_{max} = \operatorname*{argmax}_{\boldsymbol{B} \in \mathbb{R}^{m \times d}} \left[ \mathrm{tr}(\boldsymbol{B}\boldsymbol{C}\boldsymbol{B}^\top) \right]$$

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

- Solution:

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

- $\{\lambda_i, \boldsymbol{\psi}_i\}_{i=1}^m$ : Eigenvalues and eigenvectors of $\boldsymbol{C}\boldsymbol{\psi} = \lambda \boldsymbol{\psi}$

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

# Proof

- Lagrangian:

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

(Note: $\boldsymbol{\Lambda}$ is symmetric)

- Stationary point:

$$\frac{\partial L}{\partial \boldsymbol{B}} = 2\boldsymbol{B}\boldsymbol{C} - 2\boldsymbol{\Lambda}\boldsymbol{B} = 0$$

$$\Longrightarrow \quad \boldsymbol{C}\boldsymbol{B}^\top = \boldsymbol{B}^\top\boldsymbol{\Lambda}$$

$$\Longrightarrow \quad \mathcal{R}(\boldsymbol{C}\boldsymbol{B}^\top) = \mathcal{R}(\boldsymbol{B}^\top\boldsymbol{\Lambda}) \subset \mathcal{R}(\boldsymbol{B}^\top)$$

# Proof (cont.)

- $\mathcal{R}(\boldsymbol{C}\boldsymbol{B}^\top) \subset \mathcal{R}(\boldsymbol{B}^\top)$

  ⇒ $\mathcal{R}(\boldsymbol{B}^\top)$ is an invariant subspace of $\boldsymbol{C}$

  ⇒ $\mathcal{R}(\boldsymbol{B}^\top) = \mathrm{span}(\{\boldsymbol{\psi}_{k_j}\}_{j=1}^m)$

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

  ⇒ $\mathrm{rank}\left(\boldsymbol{B}^\top\right) = m$

  ⇒ All $\{k_j\}_{j=1}^m$ are distinct

# Proof (cont.)

- $\boldsymbol{B}^\top \boldsymbol{B}$ is the orthogonal projection onto $\mathrm{span}(\{\boldsymbol{\psi}_{k_j}\}_{j=1}^m)$ because

$$(\boldsymbol{B}^\top \boldsymbol{B})^2 = \boldsymbol{B}^\top \boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{B} = \boldsymbol{B}^\top \boldsymbol{B}$$

$$(\boldsymbol{B}^\top \boldsymbol{B})^\top = \boldsymbol{B}^\top \boldsymbol{B}$$

- Since $\{\boldsymbol{\psi}_{k_j}\}_{j=1}^m$ are orthonormal,

$$\boldsymbol{B}^\top \boldsymbol{B} = \sum_{j=1}^m \boldsymbol{\psi}_{k_j} \boldsymbol{\psi}_{k_j}^\top$$

# Proof (cont.)

- Eigendecomposition:

$$\boldsymbol{C} = \sum_{i=1}^{d} \lambda_i \boldsymbol{\psi}_i \boldsymbol{\psi}_i^\top$$

- $\mathrm{tr}(\boldsymbol{B}\boldsymbol{C}\boldsymbol{B}^\top) = \mathrm{tr}(\boldsymbol{C}\boldsymbol{B}^\top\boldsymbol{B}) = \sum_{j=1}^{m} \lambda_{k_j}$

- $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$

$\Rightarrow$ $k_j = j$ gives a solution

# PCA Embedding

$$\boldsymbol{B}_{PCA} = \underset{\boldsymbol{B}\in\mathbb{R}^{m\times d}}{\operatorname{argmax}} \left[\operatorname{tr}(\boldsymbol{B}^\top \boldsymbol{B}\boldsymbol{C})\right]$$

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

- $\{\lambda_i, \boldsymbol{\psi}_i\}_{i=1}^m$ :Eigenvalues and eigenvectors of $\boldsymbol{C}\boldsymbol{\psi} = \lambda\boldsymbol{\psi}$

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

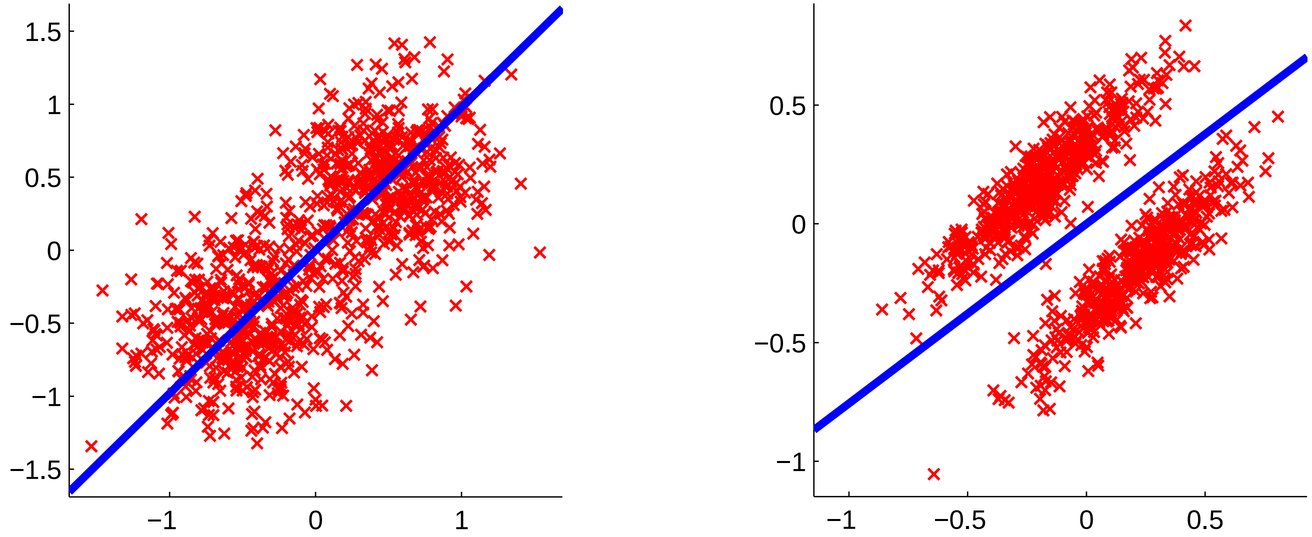
- PCA solution:

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

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

$$\boldsymbol{z}' = \boldsymbol{B}_{PCA}\boldsymbol{x}'$$

# Examples




- Data is well described
- PCA is intuitive, easy to implement, analytic solution available, and fast.
- However, PCA does not necessarily preserve interesting information such as clusters.

# Homework

- $\boldsymbol{B}$ : $m \times d, (1 \leq m \leq d)$
- $\boldsymbol{C}, \boldsymbol{D}$ : $d \times d$ , positive, symmetric
- $\boldsymbol{B}_{min} = \underset{\boldsymbol{B} \in \mathbb{R}^{m \times d}}{\operatorname{argmin}} \left[ \operatorname{tr}(\boldsymbol{B}\boldsymbol{C}\boldsymbol{B}^\top) \right]$

  subject to $\boldsymbol{B}\boldsymbol{D}\boldsymbol{B}^\top = \boldsymbol{I}_m$
- Prove:

$$\boldsymbol{B}_{min} = (\boldsymbol{\psi}_{d-m+1} | \boldsymbol{\psi}_{d-m+2} | \cdots | \boldsymbol{\psi}_d)^\top$$

- $\{\lambda_i, \boldsymbol{\psi}_i\}_{i=1}^m$ : Generalized eigenvalues and eigenvectors of $\boldsymbol{C}\boldsymbol{\psi} = \lambda \boldsymbol{D}\boldsymbol{\psi}$

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

# Homework (cont.)

- Read the following article for upcoming classes:
- X. He & P. Niyogi: Locality preserving projections, In *Advances in Neural Information Processing Systems 16*, MIT Press, Cambridge, MA, 2004.