Removing Redundancies From Data: Principle Component Analysis

COMS21202, Part III

• Song Liu (song.liu@bristol.ac.uk), Lecturer in Data Science and A.I.

1

Objectives

- Understand potential harm of high dimensionality of dataset
- Use Principle Component Analysis (PCA) to remove "redundant" dimensions from data.

• Song Liu (song.liu@bristol.ac.uk), Lecturer in Data Science and A.I.

2

High Dimensionality, Good? Bad?

- $X = \{\boldsymbol{x}_i\}_{i=1}^{n}, \boldsymbol{x} \in R^d$.
- Is a large d always a good thing?
  - ☺ We have more info as d grows!
  - ☹ LS does not work when $d > n$
  - ☹ Large d causes overfitting
  - More ☹ ?

• Song Liu (song.liu@bristol.ac.uk), Lecturer in Data Science and A.I.

3

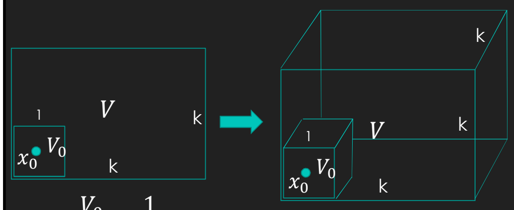
Curse of Dimensionality (CoD)

- CoD is a generic term referring to the fact that many machine learning algorithms scale very poorly with d, in terms of performance.
  - Many geometry concepts work differently in higher dimensional space.
  - One of those concepts is "locality".

• Song Liu (song.liu@bristol.ac.uk), Lecturer in Data Science and A.I.

4

The Vanishing Neighborhood



$$\frac{V_0}{V} = \frac{1}{k^2}$$

$$\frac{V_0}{V} = \frac{1}{k^3}$$

V_0: Neighborhood volume of x_0

$$\lim_{d\to\infty} \frac{V_0}{V} = \frac{1}{k^d} = 0$$

• Song Liu (song.liu@bristol.ac.uk), Lecturer in Data Science and A.I.

5

The Vanishing Neighborhood

- The neighborhood cube quickly vanishes as d increases.
- As a result, your k-nearest neighbors are **no longer** in the neighborhood V_0.
- These neighbors are no longer good at predicting the label of x_0.

• Song Liu (song.liu@bristol.ac.uk), Lecturer in Data Science and A.I.

6

- Can we reduce the dimensionality of X without losing too much information?

• Song Liu (song.liu@bristol.ac.uk), Lecturer in Data Science and A.I.

7

Reduce the Dimensionality using Feature Transform

- We want to find a **feature transform** $\boldsymbol{f}(\boldsymbol{x}) \in R^m$, where $m \ll d$.
  - $\boldsymbol{f}$ transforms original input $\boldsymbol{x}$ to a subspace as $R^m \subset R^d$.
  - We assume our dataset is **centered**:
    - $\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i = \boldsymbol{0}$
    - **If dataset X' is not centered:**
    - **Centering:** $\forall_i \, \boldsymbol{x}_i = \boldsymbol{x}'_i - \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}'_i$

• Song Liu (song.liu@bristol.ac.uk), Lecturer in Data Science and A.I.

8

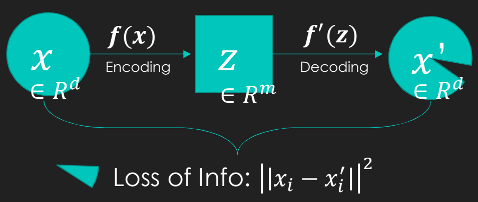
Reduce the Dimensionality using Feature Transform

- What is the optimal strategy of selecting $\boldsymbol{f}(\boldsymbol{x})$?
- Want to reduce dimension using $\boldsymbol{f}$.
  - while preserving **as much info as possible**!
- Let's look at this problem from data compression perspective!

• Song Liu (song.liu@bristol.ac.uk), Lecturer in Data Science and A.I.

9

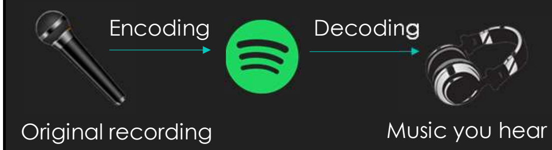
Encoder and Decoder



• Song Liu (song.liu@bristol.ac.uk), Lecturer in Data Science and A.I.

10

Codec



• Song Liu (song.liu@bristol.ac.uk), Lecturer in Data Science and A.I.

11

Linear Codec

- Suppose $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{B}\boldsymbol{x}^\top$, $\boldsymbol{B} \in R^{m \times d}$.
- Suppose $\boldsymbol{f}'(\boldsymbol{z}) = \boldsymbol{B}'\boldsymbol{z}^\top$, $\boldsymbol{B}' \in R^{d \times m}$.
- We can learn a codec by
- $\min_{\boldsymbol{B},\boldsymbol{B}'} \sum_{i=1}^{n} \left|\left|\boldsymbol{x}_i^\top - \boldsymbol{B}'\boldsymbol{B}\boldsymbol{x}_i^\top\right|\right|^2$
  - However, there are so many possible candidates $\boldsymbol{B}$ and $\boldsymbol{B}'$!
  - Solving above problem is **hard**.

• Song Liu (song.liu@bristol.ac.uk), Lecturer in Data Science and A.I.

12

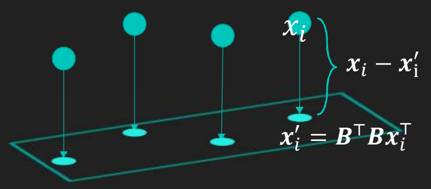
Linear Codec

- We need to put **constraints** on the $\boldsymbol{B}$ and $\boldsymbol{B}'$ to make our problem easier.
- One possible constraint is:
  - $\boldsymbol{B}' = \boldsymbol{B}^\top$
  - $\boldsymbol{B}\boldsymbol{B}' = \boldsymbol{B}\boldsymbol{B}^\top = \boldsymbol{I}$
- Such a codec actually defines an **orthogonal projection of $\boldsymbol{X}$.**
  - Show $\boldsymbol{B}'\boldsymbol{B}$ is an orth. projection matrix

• Song Liu (song.liu@bristol.ac.uk), Lecturer in Data Science and A.I.

13

Orthogonal Projection

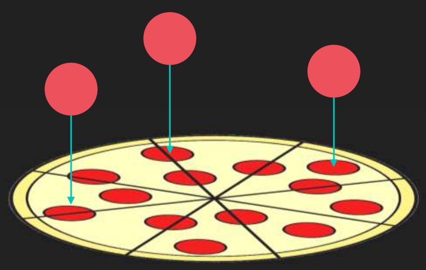


$\boldsymbol{z}_i = \boldsymbol{f}(\boldsymbol{x}_i) = \boldsymbol{B}\boldsymbol{x}_i^\top$ is called an **embedding** of $\boldsymbol{x}_i$, $\boldsymbol{B}$ is called embedding matrix.

• Song Liu (song.liu@bristol.ac.uk), Lecturer in Data Science and A.I.

14

A Pizza Topping Analogy of Embedding



• Song Liu (song.liu@bristol.ac.uk), Lecturer in Data Science and A.I.

15

Minimizing Projection Error

- $\min_{B, BB^\top = I} \sum_{i=1}^{n} \left|\left| x_i^\top - B^\top B x_i^\top \right|\right|^2$
  - We minimize square error between original data points and its projection.
- The above problem is equivalent to:
  - $\max_{B, BB^\top = I} \mathrm{tr}(BX^\top XB^\top)$
  - Live demonstration

Song Liu (song.liu@bristol.ac.uk), Lecturer in Data Science and A.I.

16

Minimizing Projection Error

- $\max_{B, BB^\top = I} \mathrm{tr}(BX^\top XB^\top)$
- Remarkably, this seemingly complex optimization has an analytical solution:
- Let $[(\lambda_1, v_1), \dots, (\lambda_m, v_m)]$ be **sorted** eigenvalue and eigenvec of $X^\top X$.
  - $\lambda_1 \geq \lambda_2 \dots \geq \lambda_m$
  - $\hat{B} = [v_1, v_2, \dots, v_m]^\top$ is an optimal solution, suppose v_i is a **column vector**.

Song Liu (song.liu@bristol.ac.uk), Lecturer in Data Science and A.I.

17

Principle Component Analysis

- As X is a centered dataset,
  - $X^\top X = n \cdot \mathrm{cov}[x]$ (PC: show it!)
- Computing $\hat{B}$ via computing sorted eigenvectors of $\mathrm{cov}[x]$ is called Principle Component Analysis (PCA).
- Finally, embedding $\hat{f}(x_i) = \hat{B}x_i^\top \in R^m$ is called **PCA embedding** of x_i.
  - m dimensional "compression" we want!

Song Liu (song.liu@bristol.ac.uk), Lecturer in Data Science and A.I.

18

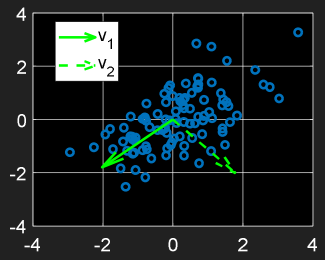
Refresh: Eigenvectors and Eigenvalues

- Given a square $n \times n$ matrix $\boldsymbol{A}$, If there exists non-zero vector $\boldsymbol{v}$ such that
- $\boldsymbol{A}\boldsymbol{v} = \lambda \boldsymbol{v}, \boldsymbol{v} \in \boldsymbol{R}^n$
- Then λ is an eigenvalue and $\boldsymbol{v}$ is an eigenvector of $\boldsymbol{A}$.

- Song Liu (song.liu@bristol.ac.uk), Lecturer in Data Science and A.I.

19

Example, One Cluster



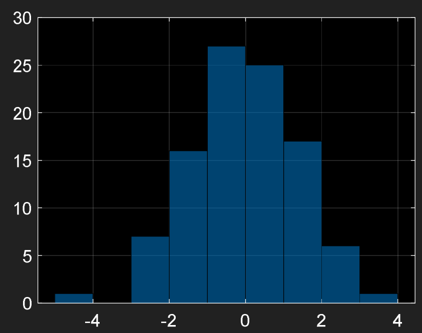
v_1 always points at the direction where your dataset has the largest variance!

PC: Intuitively explain why.

- Song Liu (song.liu@bristol.ac.uk), Lecturer in Data Science and A.I.

20

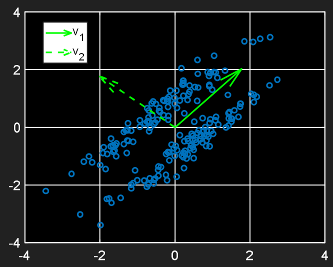
Example, Embedding $z = v_1^\top x^\top$



- Song Liu (song.liu@bristol.ac.uk), Lecturer in Data Science and A.I.

21

Example, Two Clusters

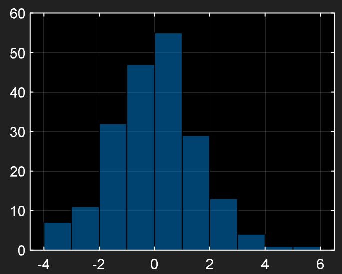


However, PCA embedding does **not necessarily** preserve clustering information.

- Song Liu (song.liu@bristol.ac.uk), Lecturer in Data Science and A.I.

22

Example, Embedding $z = v_1^\top x^\top$



Cluster information **lost** after embedding!
Will address this issue in the next lecture.

- Song Liu (song.liu@bristol.ac.uk), Lecturer in Data Science and A.I.

23

Conclusion

- Curse of Dimensionality
  - d increases, performance may decrease.
- Principle Component Analysis
  - Finding an embedding matrix $\hat{\boldsymbol{B}}$ by computing sorted eigenvalue/vectors of $\mathrm{cov}[x]$.
  - PCA Embedding: $\hat{\boldsymbol{f}}(\boldsymbol{x}_i) = \hat{\boldsymbol{B}}\boldsymbol{x}^\top$.

- Song Liu (song.liu@bristol.ac.uk), Lecturer in Data Science and A.I.

24