

# Foundations of Machine Learning

## AI2000 and AI5000

FoML-29  
PCA

Dr. Konda Reddy Mopuri

Department of AI, IIT Hyderabad  
July-Nov 2025



భారతీయ సాంకేతిక విజ్ఞాన సంస్థ హైదరాబాద్  
भारतीय प्रौद्योगिकी संस्थान हैदराबाद  
Indian Institute of Technology Hyderabad



# So far in FoML

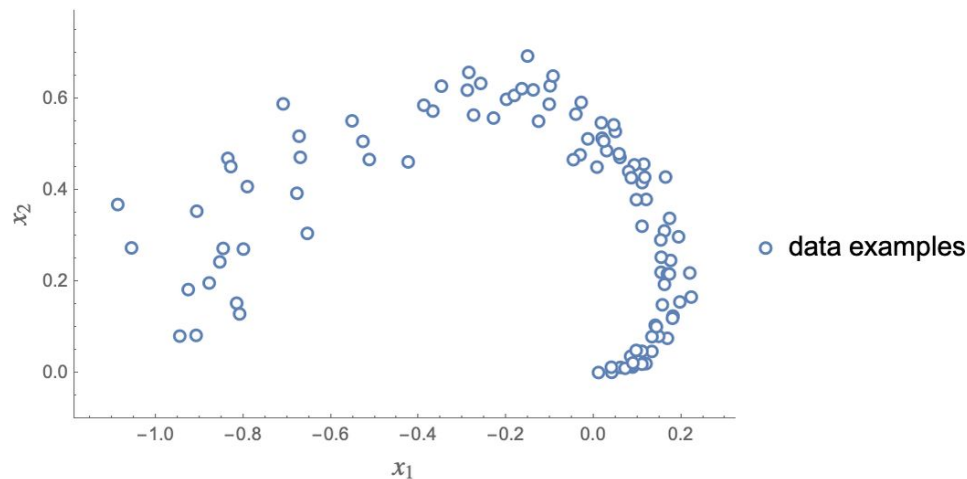
- Intro to ML and Probability refresher
- MLE, MAP, and fully Bayesian treatment
- Supervised learning
  - a. Linear Regression with basis functions
  - b. Bias-Variance Decomposition
  - c. Decision Theory - three broad classification strategies
  - d. Neural Networks
- Unsupervised learning
  - a. K-Means, Hierarchical, and GMM for clustering



# For today

- PCA - Principal Component Analysis (Pearson, 1901) & (Hotelling, 1933)

# Manifold coordinates as Latent variables



$$\{x_1, x_2\} = \{t \cos(3 t), t \sin(3 t)\}$$



# Example - facial image data

- Possible degrees of freedom
  - Skull size
  - Skin color
  - Eye color
  - Facial attributes
  - Horizontal orientation
  - Vertical orientation
  - Mood (e.g., happy)
  - etc.



# Example - facial image data

- Possible degrees of freedom
  - Skull size
  - Skin color
  - Eye color
  - Facial attributes
  - Horizontal orientation
  - Vertical orientation
  - Mood (e.g., happy)
  - etc.

Latent subspace will be a nonlinear transformation of image data



# PCA

- Linear latent subspaces



# What is PCA?

- Tool to summarize a large set of variables (dimensions) with a smaller set





# What is PCA?

- Tool to summarize a large set of variables (dimensions) with a smaller set
  - Of 'representative' variables that explain the 'variability' in the original set



# What is PCA?

- Tool to summarize a large set of variables (dimensions) with a smaller set
  - Of 'representative' variables that explain the 'variability' in the original set



# What is PCA?

- Smaller set need not be a subset of original variables!
  - Rather, combinations of original variables
- They may not mean the same as originals
  - lost interpretability!
- New variables are independent of each other!

# What is PCA?

- Gives the directions along which the data are highly 'variable'
  - Projects linearly such that the variance in the projected space is maximal



# PCA

- Data  $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$   $\mathbf{x}_i \in \mathbb{R}^D$
- Aim: project data onto M dimensional space ( $M < D$ ) maximizing the variance of the projected data

# PCA

- Mean

$$\bar{\mathbf{x}} =$$

- Covariance

$$\mathbf{S} =$$

# 1D representation using PCA

- Project data onto a direction where most of the variance is preserved
- Projecting onto  $\mathbf{u}_1$  gives a scalar  $\rightarrow$  1D representation

$$\mathbf{z}_i = \mathbf{u}_1^T \mathbf{x}_i$$



# 1D representation using PCA

- Direction of  $\mathbf{u}_1$  is important  $\rightarrow$  consider unit vector in that direction

$$\|\mathbf{u}_1\|_2 = 1$$



# 1D representation using PCA

- Consider the variance in the new subspace

$$\text{Var}[\mathbf{z}] =$$



# 1D representation using PCA

- Let's find the direction ( $\mathbf{u}_1$ ) that maximizes the variance in  $z_i$

$$\arg \max_{\mathbf{u}_1} \mathbf{u}_1^T \mathbf{S} \mathbf{u}_1 \quad \text{such that} \quad \mathbf{u}_1^T \mathbf{u}_1 = 1$$



# PCA via maximum variance

- Repeat the procedure for the next  $M-1$  orthogonal vectors
  - Maximize the variance by projecting onto a direction orthogonal to the found ones
  - These are the next  $M-1$  eigenvectors of the covariance matrix ( $S$ )

# PCA - Eigen decomposition

- For the symmetric PSD matrix  $\mathbf{S} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$
- Eigenvectors are orthonormal (contained in  $\mathbf{U}$ )
- Eigenvalues are non-negative (contained in  $\mathbf{\Lambda}$ )



# PCA - Eigen decomposition

- Variance is  $\text{Tr}(\mathbf{S})$

$$\mathbf{S} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$$



# PCA - some notes



# PCA - Scaling the features

- Sensitive to the scales of the features/variables
  - Features with greater range dominate the process of finding the PCs
- Perform standardization to prevent this



# PCA - Proportion of the Variance Explained

- How much of the information is lost by projecting onto PCs?



# PCA - Proportion of the Variance Explained

- Total variance

$$\sum_{j=1}^p \text{Var}(X_j) = \sum_{j=1}^p \frac{1}{n} \sum_{i=1}^n x_{ij}^2$$

# PCA - Proportion of the Variance Explained

- Total variance
- Variance explained by the 'm'<sup>th</sup> PC

$$\sum_{j=1}^p \text{Var}(X_j) = \sum_{j=1}^p \frac{1}{n} \sum_{i=1}^n x_{ij}^2$$

$$\frac{1}{n} \sum_{i=1}^n z_{im}^2$$

# PCA - Proportion of the Variance Explained

- Total variance
- Variance explained by the 'm'<sup>th</sup> PC

$$\sum_{j=1}^p \text{Var}(X_j) = \sum_{j=1}^p \frac{1}{n} \sum_{i=1}^n x_{ij}^2$$

$$\frac{1}{n} \sum_{i=1}^n z_{im}^2$$

PVE of the 'm'<sup>th</sup> PC =



# PCA - How many PCs to consider?

- For an  $n \times p$  data matrix
  - $\min(n-1, p)$  PCs are possible
  - Why?



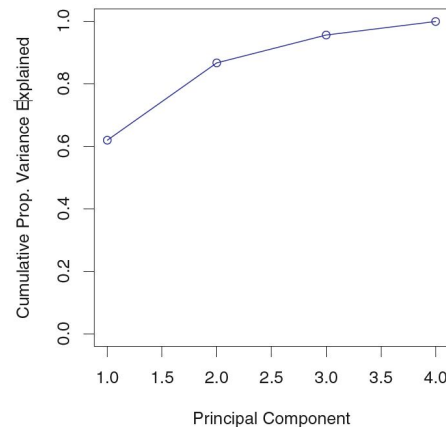
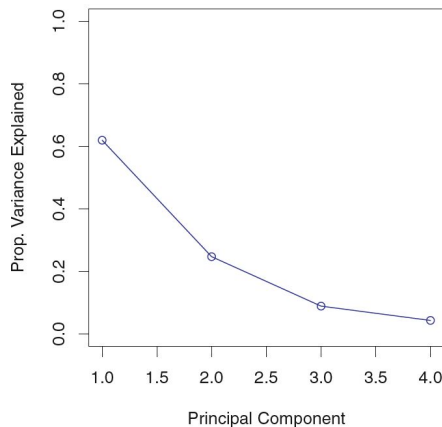
# PCA - How many PCs to consider?

- For an  $n \times p$  data matrix
  - $\min(n-1, p)$  PCs are possible
- Not all of them may be interesting



# PCA - How many PCs to consider?

- Generally, we want the 'smallest' number of them → good understanding of the data
- → scree plot & elbow



# PCA

- Doesn't discard the redundant variables
  - Finds new variables (linear combinations of the 'p' variables) that summarize the data well
  - The 'best' variables (among the all possible linear combinations)
  - Resulting new features are uncorrelated (covariance matrix will be diagonal)

# Applications of PCA

- Dimensionality reduction → tackles curse of dimensionality
- Less compute requirement
- Less prone to overfitting
- Useful preprocessing



# Next

- PCA continued

