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

DiL
Data-driven Intelligence & Learning Lab

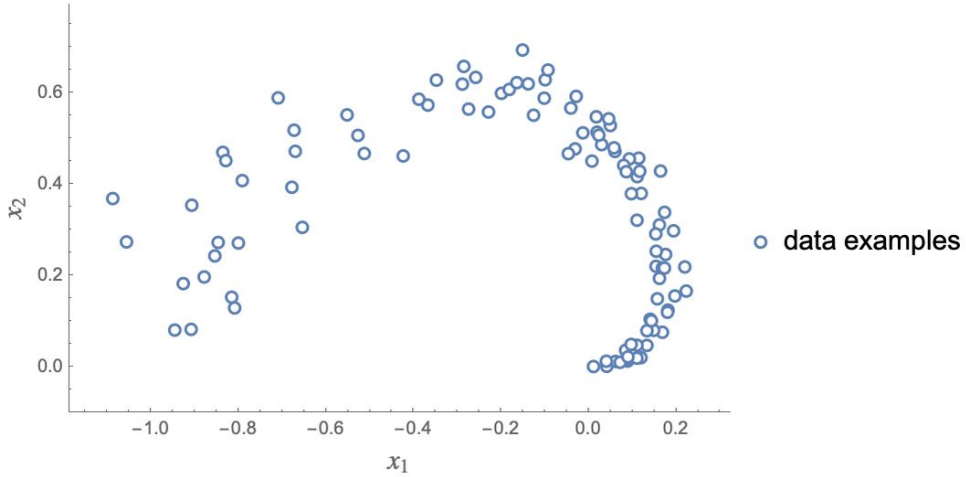
So far in FoML

- 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



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

Figure credits

भारतीय सांकेतिक विज्ञान संस्थ हैदराबाद
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

DiL
Data-driven Intelligence & Learning Lab

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?

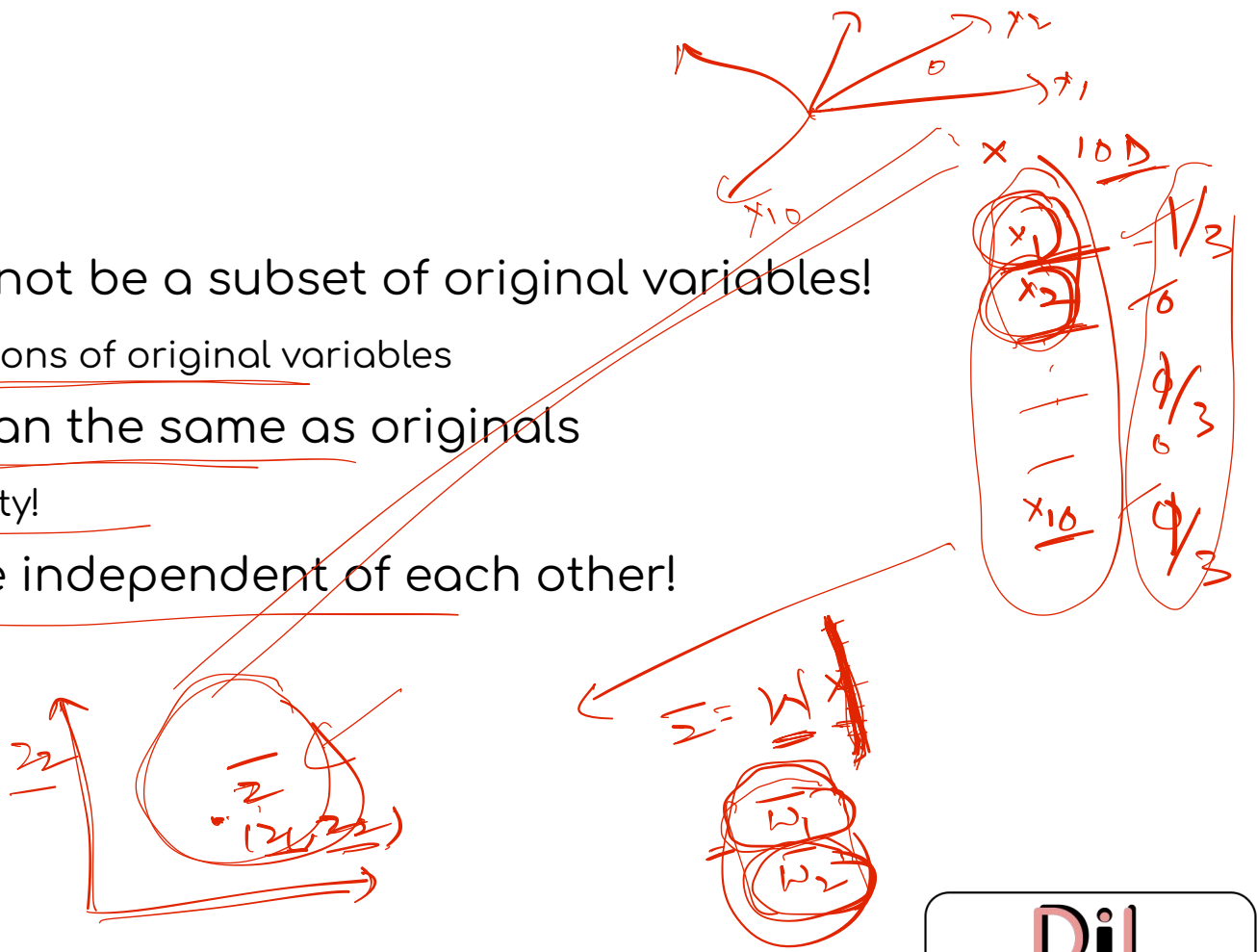
- Tool to summarize a large set of variables (dimensions) with a smaller set
  - Of ‘representative’ variables that explain the ‘variability’ in the original set

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

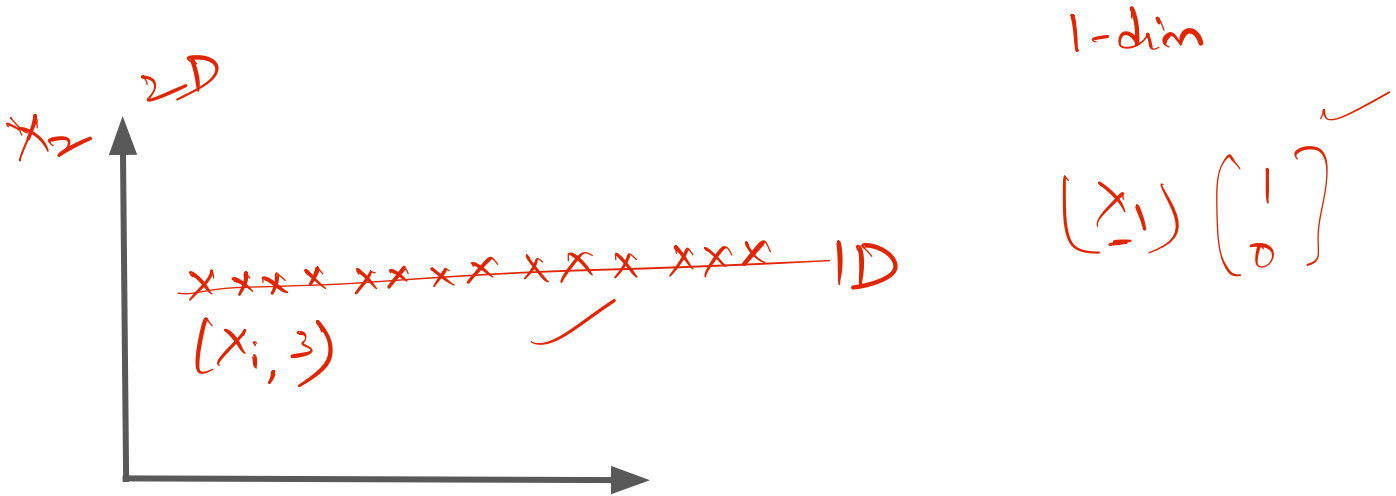
What is PCA?

- 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 $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ $\mathbf{x}_i \in \mathbb{R}^D$

- Aim: project data onto M dimensional space (M < D) maximizing the variance of the projected data

PCA

- Mean $\bar{\mathbf{x}} = \frac{1}{N} \sum_{i=1}^{N} x_i$
- Covariance $\mathbf{S} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})(x_i - \bar{x})^T$

Symmetric

PSD

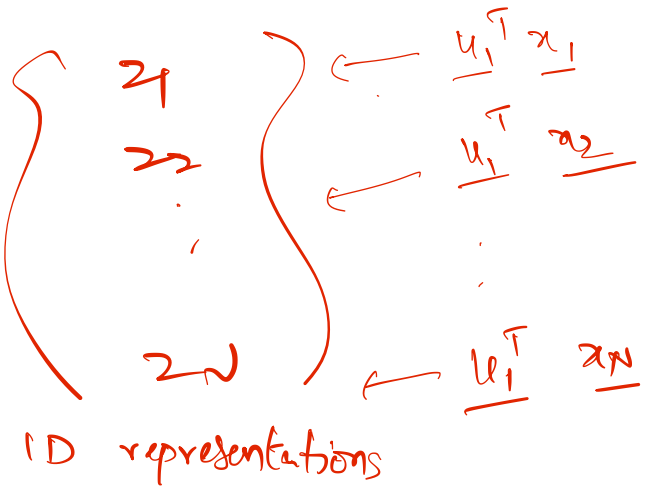
$\lambda_i \geq 0$

Lin ind Eig vecs

1D representation using PCA

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

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



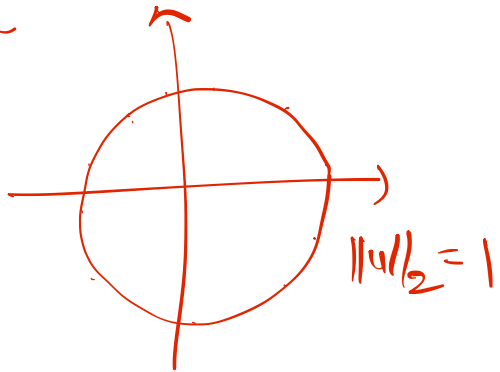
1D representation using PCA

- Direction of u_1 is important → consider unit vector in that direction

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

But not its magnitude

w/o this constraint, we can't optimize



1D representation using PCA

- Consider the variance in the new subspace

$$\mathrm{Var}[\mathbf{z}] = \frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})^2 = \frac{1}{N}\sum_{i=1}^{N}(u_1^T x_i - u_1^T \bar{x})^2$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left[u_1^T(x_i - \bar{x})\right]^2 = \frac{1}{N}\sum_{i=1}^{N} u_1^T(x_i - \bar{x})(x_i - \bar{x})^T u_1$$

$$= \frac{1}{N} u_1^T \underbrace{\sum_{i=1}^{N}(x_i - \bar{x})(x_i - \bar{x})^T}_{S} u_1 = u_1^T S u_1$$

1D representation using PCA

1st principal component — PC1 ↓ u_1

- Let's find the direction (u_1) that maximizes the variance in z_i

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

using Lagrange multiplier

$$\arg\max_{u_1, \lambda} \; u_1^T S u_1 + \lambda(1 - u_1^T u_1) = 0$$

$$\frac{\partial}{\partial u_1} \downarrow = 0$$

$$2Su_1 - 2\lambda u_1 = 0 \qquad Su_1 = \lambda u_1$$

$\nabla f(x) + \lambda \nabla g(x) = 0$

$f(x) + \lambda g(x)$

u_1 is eig. vector
λ is corresponding eig. value of S

PCA via maximum variance

$(x_i - z_i u_1)$ — $d \times 1$

- 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)

$$u_1^T \Sigma u_1 = u_1^T \lambda u_1 = \lambda u_1^T u_1 = \lambda$$

The variance captured by a PC is its eigen value (λ_i)

(of the projected data) (u_i)

PCA - Eigen decomposition

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

PCA - Eigen decomposition

- Variance is Tr(S)

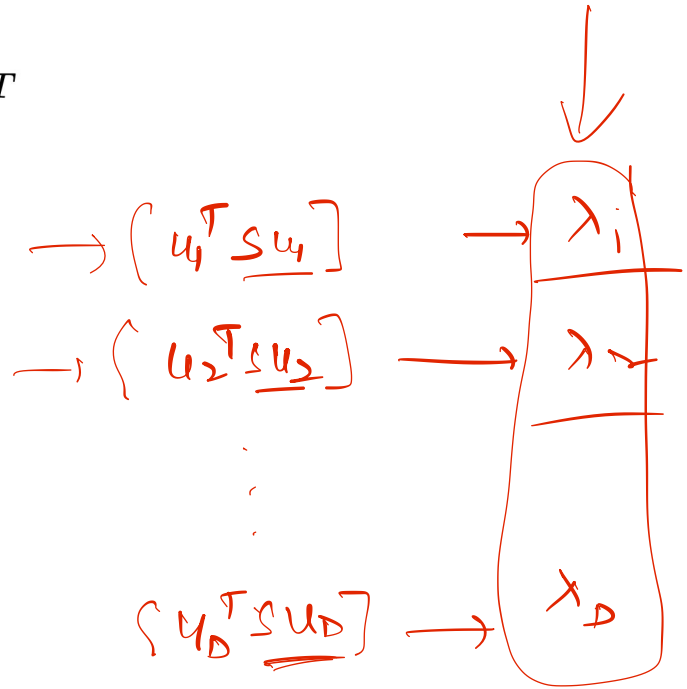
$$\mathbf{S} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$$

of the projected data

$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_D$

$Tr[Cov[z]]$ = variance of data along each of the new dims

$= \lambda_1 + \lambda_2 + \cdots + \lambda_D$



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

mean subtraction and variance normalization

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

P = no. of dimensions in the original space

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

assume mean subtracted

PCA - Proportion of the Variance Explained

- Total variance

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

- Variance explained by the 'm'th PC

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

↳ linear transformation doesn't disturb the mean centering

PCA - Proportion of the Variance Explained

- Total variance

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

- Variance explained by the 'm'th PC

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

upto m ✓

$$\sum_{i=1}^{m} \lambda_m$$

PVE of the 'm'th PC = ↑

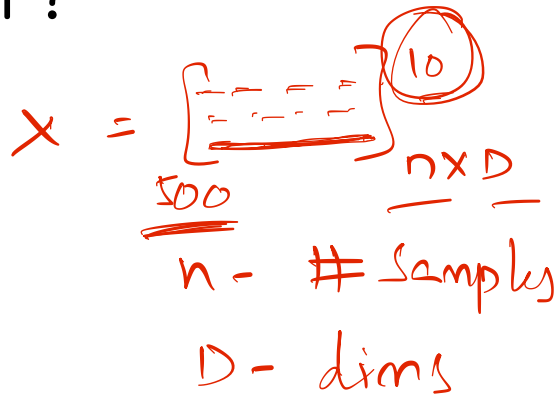
$$\frac{\lambda_m}{\sum_{i=1}^{D} \lambda_i}$$

$$\frac{\sum_{i=1}^{m} \lambda_m}{\sum_{i=1}^{P} \lambda_m}$$

PCA - How many PCs to consider?

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

Mean centering makes the rows dependent, hence n-1



n > D generally

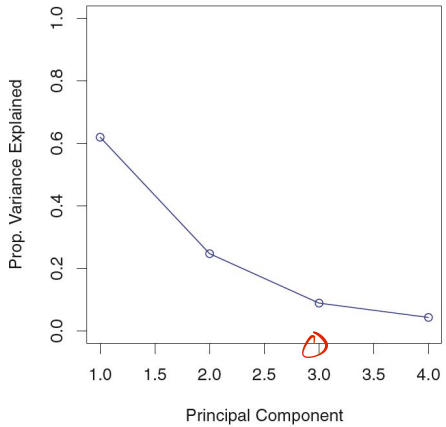
n < D rarely

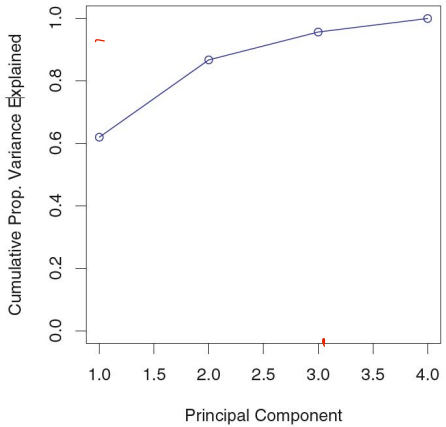
PCA - How many PCs to consider?

- For an n X p data matrix
  - min(n-1, p) PCs are possible
- Not all of them may be interesting (leads to acceptable loss of information)

PCA - How many PCs to consider?

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





PCA

(not feature selection, rather feature extraction)

- 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)

'p' and 'D' both indicate the original dimension of the data

Applications of PCA

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

Next

- PCA continued

on the constraint set (green level set)

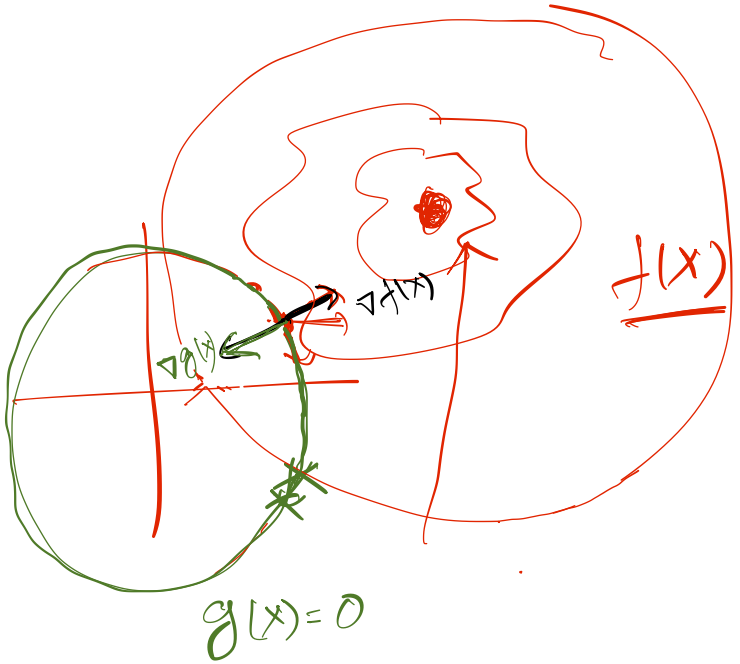
$\nabla g(x)$ is $\perp^r$ to the level set surface

$g(x+\epsilon) = g(x) + \epsilon^T \nabla g(x)$

$g(x+\epsilon) = g(x)$ on the level set

$\Rightarrow \quad \epsilon^T \nabla g(x) = 0 \quad \nabla g(x) \perp \epsilon$

ϵ is the movement along the levelset



$\nabla f(x)$ will also be $\perp^r$ to the level set at x^*. Otherwise we can increase $f(x)$ by moving along the $g(x)$