Principal Component Analysis (PCA) & Dimensionality Reduction

Why, When, and How

Mahdi Mastani

Why Do We Need Feature Selection?

1. Real-world datasets often contain many features — not all are useful.

2. Remove irrelevant or redundant features

3. Feature selection helps identify the most informative variables for learning tasks.

Selection vs. Reduction: What's the Difference?

- Feature Selection: Choose a subset of existing features (e.g., filter, wrapper methods).
- Dimensionality Reduction: Transform features into a lower-dimensional space (e.g., PCA, t-SNE).
- PCA doesn't remove features — it creates new ones that are combinations of the original.

Why Reduce Dimensions?

- Improve model performance and generalization
- Reduce training time and computational cost
- Enhance visualization (e.g., 2D/3D plots)
- Remove multicollinearity
- Make patterns in the data more detectable

Principal Component Analysis (PCA)

PCA steps

1. Data is first normalized.
2. A covariance matrix is computed.
3. Eigenvectors and eigenvalues are then determined.
4. Principal components are subsequently selected.
5. Data is transformed into a new dimensional space.

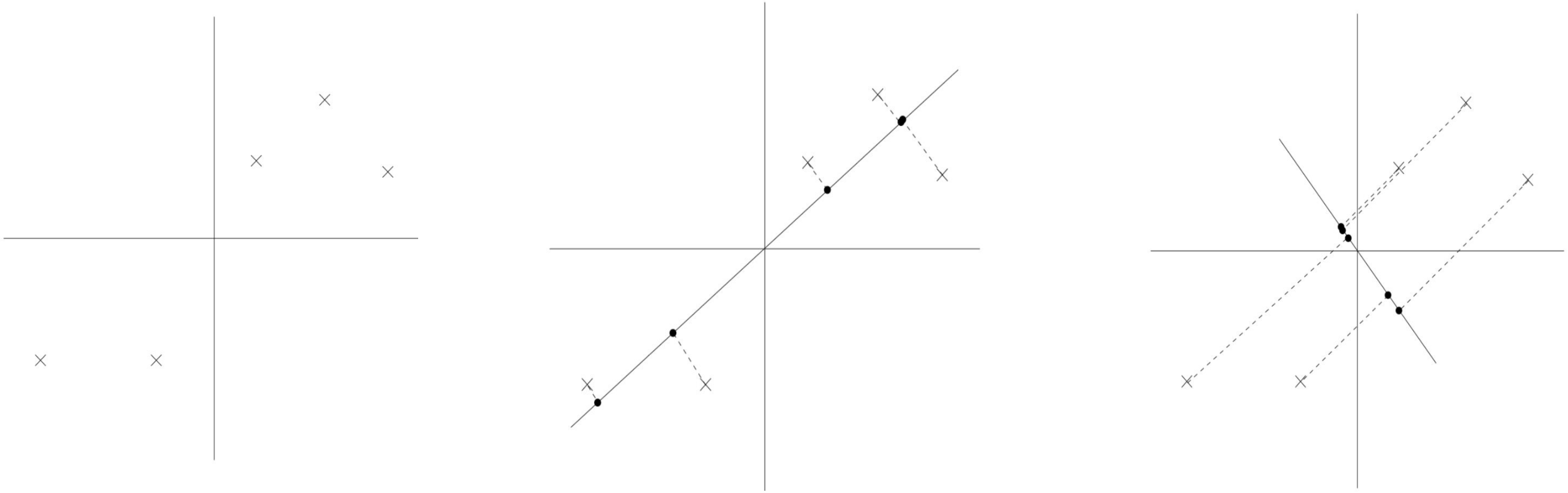
Why Normalize Data Before PCA?

- PCA is sensitive to the scale of features.
- It relies on the covariance matrix, which is affected by the magnitude of the data.
- Features with larger scales (e.g., income in thousands, age in years) will dominate the principal components.
- Without normalization:
    - PCA might give undue importance to high-variance features simply due to their units, not their true informativeness.
- Normalization ensures that:
    - Each feature contributes equally to the analysis
    - PCA captures true variance patterns, not scale effects

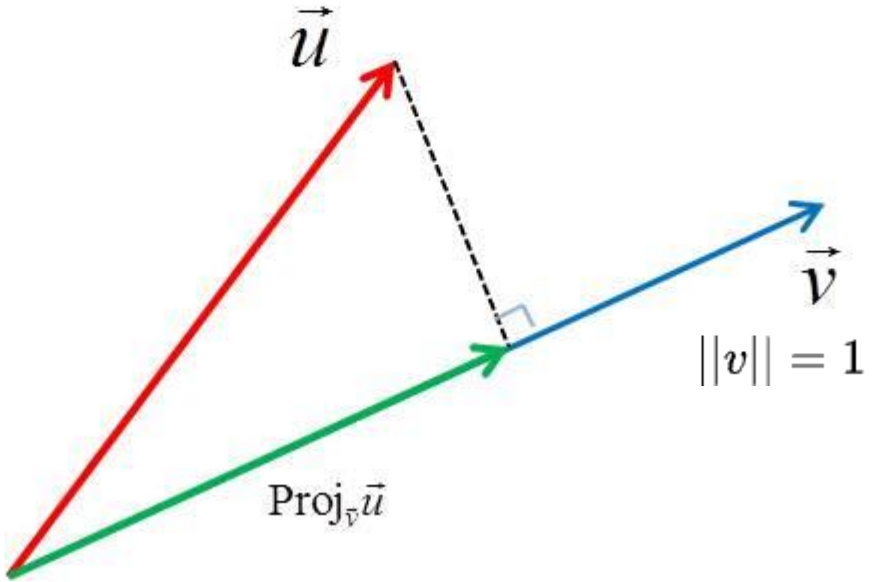
How normalize?

- $$\mu_j = \frac{1}{n}\sum_{i=1}^{n} x_j^{(i)}$$
- $$\sigma_j^2 = \frac{1}{n}\sum_{i=1}^{n}(x_j^{(i)} - \mu_j)^2$$
- $$x_j^{(i)} \leftarrow \frac{x_j^{(i)} - \mu_j}{\sigma_j}$$

Which one?



Projection of point on vector



$$proj_v(u) = v^T u$$

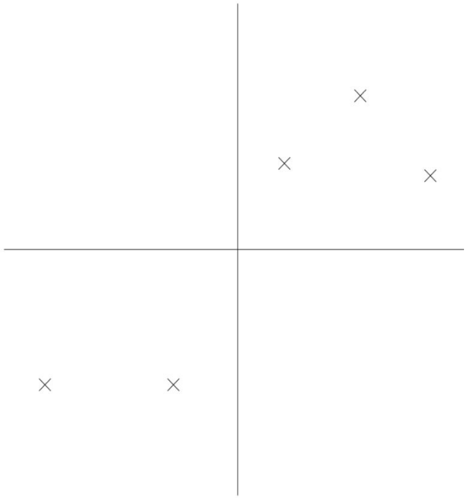
What Is the Purpose of PCA?

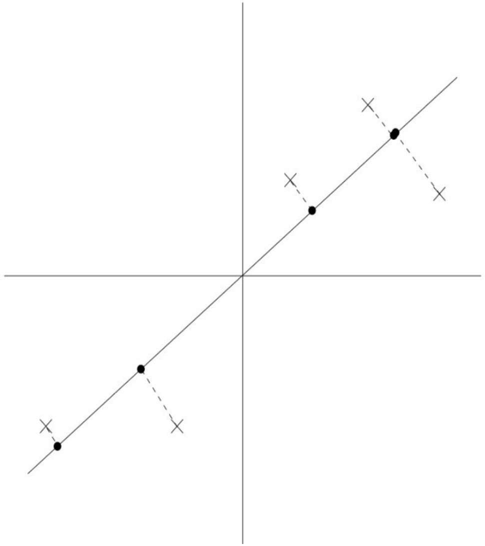
- The main goal of Principal Component Analysis (PCA) is to:

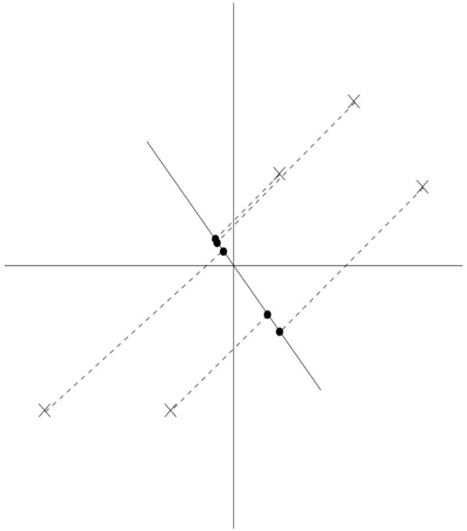
  Find a new set of axes (directions) that best capture the structure of the data.

- Mathematically, PCA finds directions (principal components) that:
  - Are orthogonal (uncorrelated)
  - Capture the maximum variance in the data

Which one?







PCA formulation

$$Var_{project} = \frac{1}{n}\sum_{i=1}^{n}\left(x^{(i)T}u\right)^2 = \frac{1}{n}\sum_{i=1}^{n}u^T x^{(i)} x^{(i)T} u = u^T\left(\frac{1}{n}\sum_{i=1}^{n}x^{(i)}x^{(i)T}\right)u$$

$$\Sigma = \frac{1}{n}\sum x^{(i)}x^{(i)T}$$

$$\max_{||u||=1} \; u^T \Sigma u$$

Link to Linear Algebra: A Classic Optimization Problem

The PCA optimization problem:

$$\max_{||u||=1} \; u^T \Sigma u$$

is a well-known problem in linear algebra.

Its solution is found by solving the eigenvalue problem of the covariance matrix \Sigma:

- The direction u that maximizes the variance is the eigenvector corresponding to the largest eigenvalue of \Sigma.
- This eigenvector becomes the first principal component.
- Subsequent components correspond to the next largest eigenvalues and are orthogonal to each other.

Projecting Data onto Principal Components

After computing the top k principal components $u_1, u_2, \ldots, u_k$, we can represent each original data point $x^{(i)}$ in a reduced-dimensional space:

$$y^{(i)} = \begin{bmatrix} u_1^T x^{(i)} \\ u_2^T x^{(i)} \\ \vdots \\ u_k^T x^{(i)} \end{bmatrix} \in \mathbb{R}^k$$