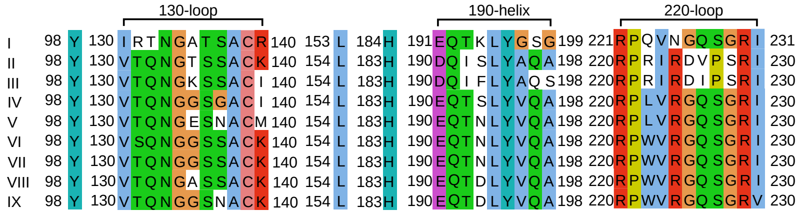
Principal Component Analysis

Jiahui Chen
Department of Mathematical Sciences
University of Arkansas

UNIVERSITY OF ARKANSAS

Motivation

- High-Dimensions = Lot of Features
  - Influenza A: 13,500 bp
  - SARS-CoV-2: 29903 bp
  - Spotify or Netflix
  - Image data
  - …



	Movie 1	Movie 2	Movie 3	Movie 4	Movie 5	Movie 6	Movie 7
Tom	4	2	?	3	?	3	
John	5	0	2	?	2	3	

Introduction

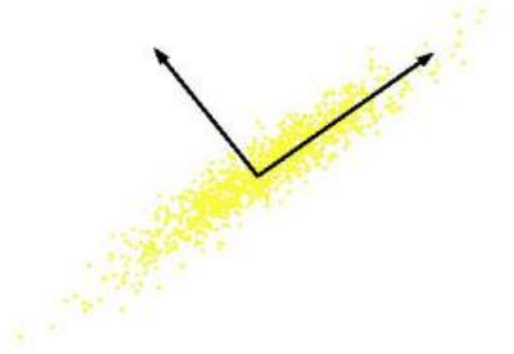
- High-Dimensional Data
- Useful to learn lower dimensional representations of the data
- Powerful unsupervised learning techniques for extracting hidden (potentially lower dimensional) structure for high dimensional datsets

Useful for

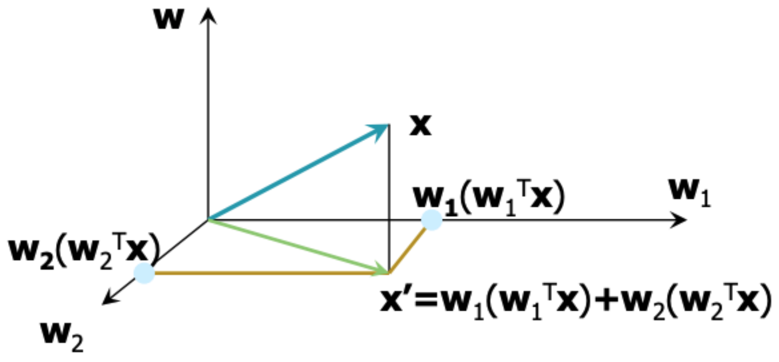
- Visualization
- More efficient use of resources (e.g., time, memory)
- Statistical: fewer dimensions -> better generalization
- Noise removal
- Further processing by machine learning algorithms

Principal Component Analysis

What is PCA: Unsupervised technique for extracting variance structure from high dimensional dataset

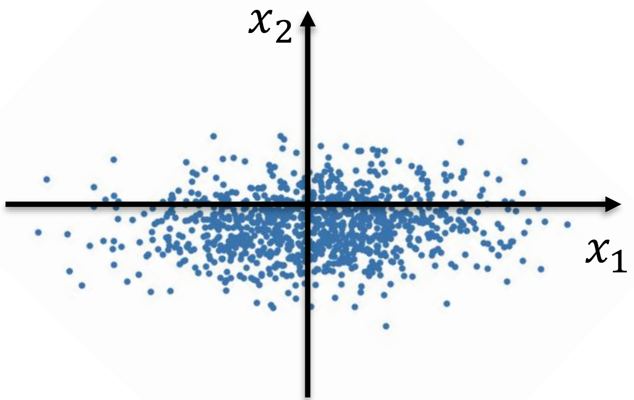


PCA is an orthogonal projection or transformation of the data into a (possibly lower dimensional) subspace so that the variance of the projected data is maximized.



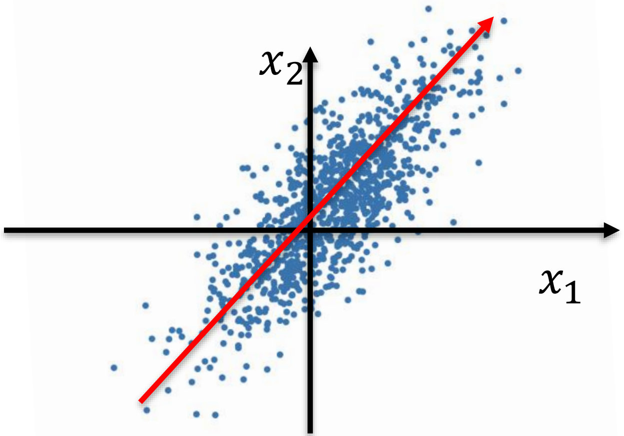
Principal Component Analysis

Intrinsically lower dimensional than the dimension of the ambient space.



Only one relevant feature

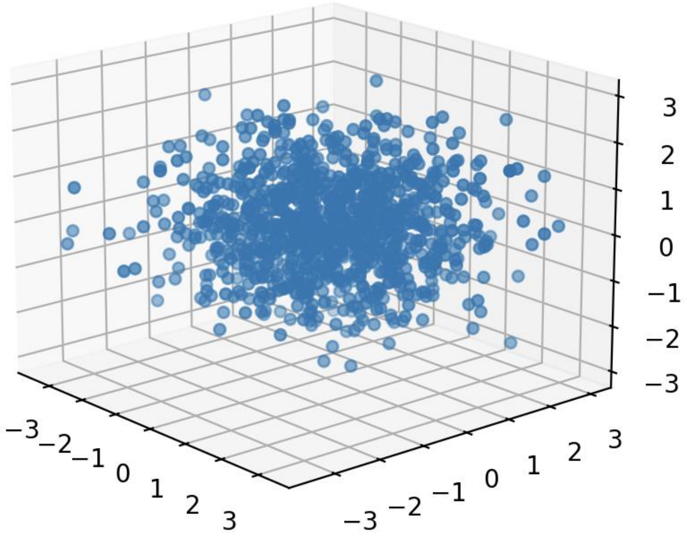
If we rotate data, again only one coordinate is more important.

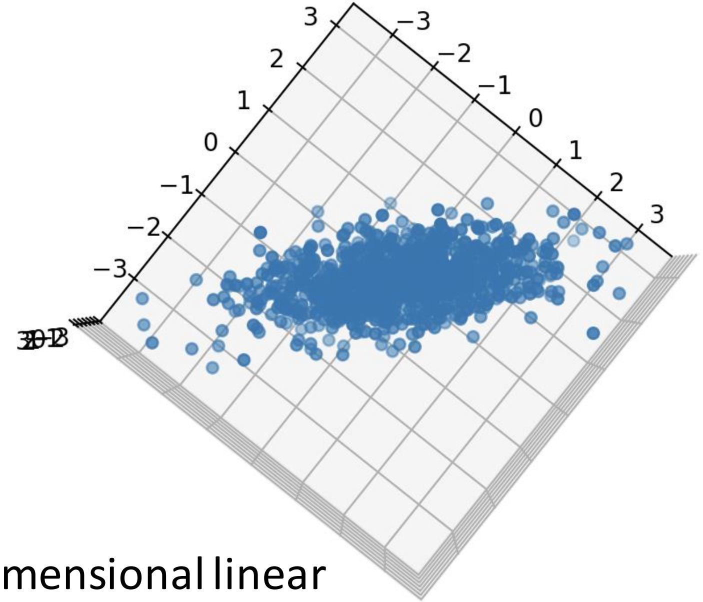


Both features are relevant

Question: Can we transform the features so that we only need to preserve one latent feature?

Principal Component Analysis





In case where data lies on or near a low d-dimensional linear subspace, axes of this subspace are an effective representation of the data.

Identifying the axes is known as Principal Components Analysis, and can be obtained by using classic matrix computation tools (Eigen or Singular Value Decomposition).

PCA method 1

Give the centered data $\{\mathbf{x}_1, \dots, \mathbf{x}_m\}$, compute the principal vectors:

$\mathbf{w}_1 = \underset{||\mathbf{w}||=1}{\operatorname{argmax}} \frac{1}{m} \sum_{i=1}^{m} (\mathbf{w}_1^T \mathbf{x}_i)^2$, first PCA vector

We maximize the variance of projection of $\mathbf{x}$

$\mathbf{w}_k = \underset{||\mathbf{w}||=1}{\operatorname{argmax}} \frac{1}{m} \sum_{i=1}^{m} (\mathbf{w}_k^T (\mathbf{x}_i - \sum_{j=1}^{k-1} \mathbf{w}_j \mathbf{w}_j^T \mathbf{x}_i))^2$, k-th PCA vector

We maximize the variance of the projection in the residual subspace

Questions: Have you ever seen a similar process?

Gram-Schmidt Process: to find a set of orthogonal vectors

PCA method 1

How to solve the maximize problem?

Lagrange multipliers

Maximize $f(\mathbf{w}) = \frac{1}{m}\sum_{i=1}^{m}(\mathbf{w}^T\,\mathbf{x}_i)^2$

Subject to $g(\mathbf{w}) = ||\mathbf{w}||^2 - 1 = 0$

1. Construct the Lagrange multiplier: $L(\mathbf{w},\lambda) = f(\mathbf{w}) - \lambda g(\mathbf{w})$
2. Take the Derivative $\frac{\partial L}{\partial \mathbf{w}}$ and $\frac{\partial L}{\partial \lambda}$
3. Solve the System $\frac{\partial L}{\partial \mathbf{w}} = \mathbf{0}$ and $\frac{\partial L}{\partial \lambda} = 0$

PCA method 2

Covariance Matrix

Give the centered data $\{\mathbf{x}_1, \dots, \mathbf{x}_m\}$, compute the covariance matrix M

$$M = \frac{1}{m}\sum_{i=1}^{m}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$$

where $\bar{\mathbf{x}}$ is the average of the centered data

- PCA basis vectors = the eigenvectors of M
- Larger eigenvalues -> more important eigenvectors
- Are method 1 and method 2 equivalent?

Eigenvalues & Eigenvectors

Why the Eigenvectors?

Maximize $\mathbf{w}^T \mathbf{X}\mathbf{X}^T \mathbf{w}$, s.t. $||\mathbf{w}|| = \mathbf{1}$

Construct Lagrangian:

$$\mathbf{w}^T \mathbf{X}\mathbf{X}^T \mathbf{w} - \lambda||\mathbf{w}||^2$$

Vector of partial derivatives set to zero

$$\mathbf{X}\mathbf{X}^T \mathbf{w} - \lambda \mathbf{w} = 0$$

which is an eigenvalue problem

$$(\mathbf{X}\mathbf{X}^T - \lambda \mathbf{I})\mathbf{w} = 0$$

- Why the eigenvectors are orthogonal?
- Why all eigenvalues are real numbers?
- Why all eigenvalues are non-negative?

Discussion

Advantages of PCA

- **Noise Reduction:** PCA can help in noise reduction by eliminating the components with lower variance, which are likely to represent noise.
- **Feature Extraction:** PCA can be considered a feature extraction technique as it generates new features (principal components) that better capture the underlying structure in the data.
- **Visualization:** By reducing dimensionality to two or three principal components, PCA allows for the visualization of complex and high-dimensional data.

Limitations of PCA

- **Linear Assumptions**: PCA assumes that the principal components are linear combinations of the original features, which may not capture complex structures.
- **Variance Equals Information**: PCA assumes that features with higher variance are more important, which is not always the case.
- **Sensitivity to Scaling**: The results of PCA are sensitive to the scaling of the features, which is why standardization is a critical step.