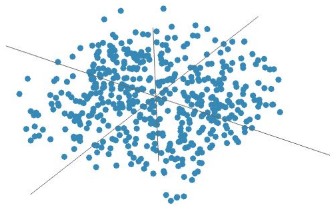
Dimensional Reduction

Principal Component Analysis

Goals

- Understand the need and purpose of dimensionality reduction algorithms.
- Understand and learn the details of Principal Component Analysis (PCA). Including its strengths and weaknesses.
- See concrete applications of using PCA in context.
- Show other dimensionality reduction algorithms.

Point Cloud Data



A point cloud is a collection of data points in $\mathbb{R}^n$

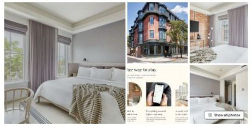
- Each point is represented according to its coordin $(X_1, X_2, X_3, \ldots, X_m)$
- Each coordinate may represent a different feature or characteristic:
  - Hotels in a city can be represented in $\mathbb{R}$6 according to user rating of characteristics: cleanliness, accuracy, communication, check-in, location, value
  - A 28x28 pixel grayscale image can be represented in $\mathbb{R}$784: each pixel is represented with a unique number according to a scale from black to white
  - Samples of expressions of N genes can be represented in $\mathbb{R}$N

Example 1: Hotel Listings

- Each data point in $\mathbb{R}^6$ corresponds to a hotel
- Hotels are ranked according to 6 categories:
- Each individual hotel can be represented by a point in

$$= \begin{bmatrix} 3.8 \\ 4.6 \\ 4.8 \\ 4.2 \\ 4.7 \\ 4.3 \end{bmatrix}$$

$$= \begin{bmatrix} 4.8 \\ 5 \\ 5 \\ 4.8 \\ 4.7 \\ 4.8 \end{bmatrix}$$



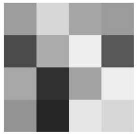
$$= \begin{bmatrix} 4.9 \\ 4.5 \\ 4.7 \\ 4.7 \\ 4.9 \\ 4.5 \end{bmatrix}$$

$$= \begin{bmatrix} 4.6 \\ 4.7 \\ 4.7 \\ 4.5 \\ 4.1 \\ 4.2 \end{bmatrix}$$

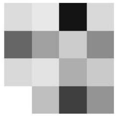
$$= \begin{bmatrix} 4.6 \\ 4.7 \\ 4.7 \\ 4.7 \\ 4.9 \\ 4.5 \end{bmatrix}$$

Example 2: Grayscale Images

- Each data point corresponds to an image of resolution 4x4
- Each of the 16 pixels is represented with a number from 0 (black) to 1 (white)
- Each image can be represented by a point in $\mathbb{R}^{16}$



$$\rightarrow \begin{pmatrix} 0.62 & 0.84 & 0.65 & 0.61 \\ 0.3 & 0.67 & 0.93 & 0.35 \\ 0.66 & 0.19 & 0.64 & 0.93 \\ 0.58 & 0.15 & 0.9 & 0.84 \end{pmatrix} \rightarrow \begin{pmatrix} 0.62 \\ 0.84 \\ 0.65 \\ 0.61 \\ 0.3 \\ 0.67 \\ 0.93 \\ 0.35 \\ 0.66 \\ 0.19 \\ 0.64 \\ 0.93 \\ 0.58 \\ 0.15 \\ 0.9 \\ 0.84 \end{pmatrix}$$



$$\rightarrow \begin{pmatrix} 0.86 & 0.91 & 0.08 & 0.85 \\ 0.4 & 0.63 & 0.8 & 0.55 \\ 0.85 & 0.89 & 0.68 & 0.79 \\ 1. & 0.75 & 0.25 & 0.58 \end{pmatrix} \rightarrow \begin{pmatrix} 0.86 \\ 0.91 \\ 0.08 \\ 0.85 \\ 0.4 \\ 0.63 \\ 0.8 \\ 0.55 \\ 0.85 \\ 0.89 \\ 0.68 \\ 0.79 \\ 1. \\ 0.75 \\ 0.25 \\ 0.58 \end{pmatrix}$$

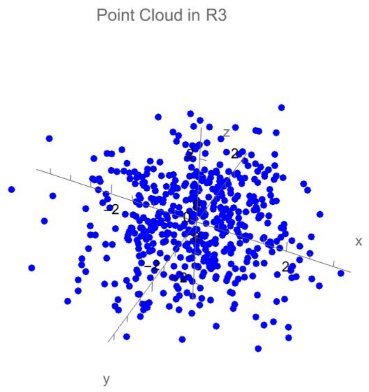
Example 3: Gene Expression

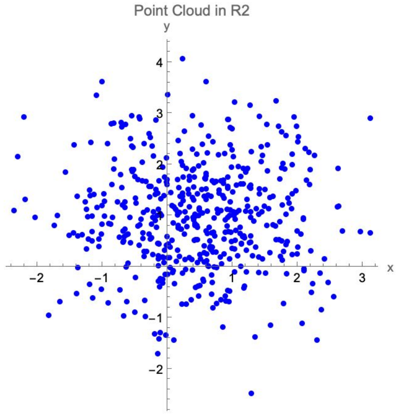
- Each sample measures expressions of N genes in distincts cells
- Each individual cell can be represented by a point in $\mathbb{R}^N$

Sample ID	Gene1	Gene2	Gene3	...	GeneN
Sample1	5.2	0.1	3.4	...	7.6
Sample2	4.9	0.0	3.8	...	6.8

Point Cloud Data: Goal

- Is there a way to visualize higher dimensional data?
- If so, how much is it representative of the original data?
- Are there any features that are more important than others?
- Are there any **combinations of features** that are more important than others?





Why Dimensional Reduction?

Reasons are practical from both a computational and statistical point of view:

- Reduce computational complexity
- Reduce redundancy and noise
- Reduce overfitting (improve performance)
- Enable visualization
- Find correlations between input features

What do we need?

- Statistics
- Linear Algebra

Statistics Basics

Statistics Measurements

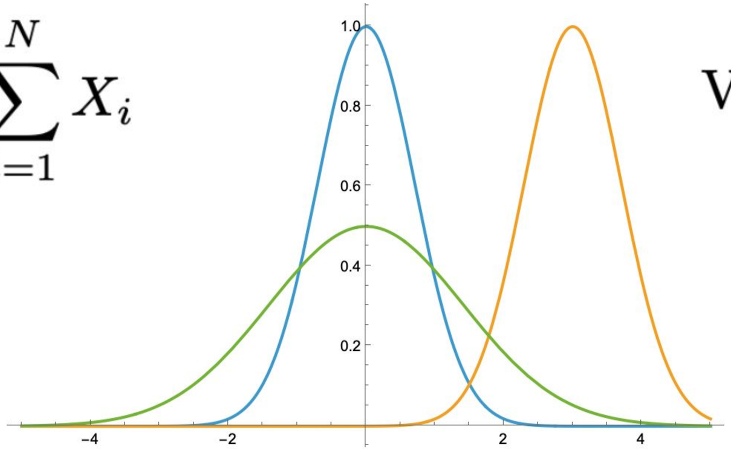
Suppose we have N measurements of a certain feature: $X_1, X_2, \ldots, X_N$

The **mean** is the central tendency or "average" of a set of numbers:

$$\overline{X} = \frac{1}{N} \sum_{i=1}^{N} X_i$$

The **variance** measures how spread out the values are around the mean:

$$\mathrm{Var}(X) = \frac{1}{N} \sum_{i=1}^{N} (X_i - \overline{X})^2$$



Statistics Measurements

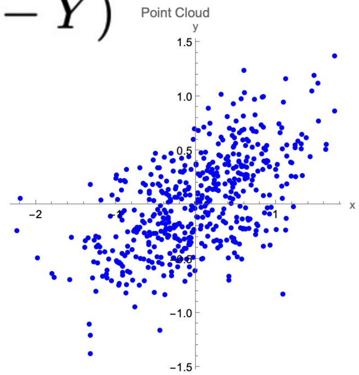
Suppose we have N measurements of two features:

$$X_1, X_2, \ldots, X_N$$

$$Y_1, Y_2, \ldots, Y_N$$

The **covariance** is a measure of how two variables change together—whether they tend to increase or decrease at the same time.

$$\mathrm{Cov}(X, Y) = \frac{1}{N} \sum_{i=1}^{N} (X_i - \overline{X})(Y_i - \overline{Y})$$



Correlation is the standardized version of covariance

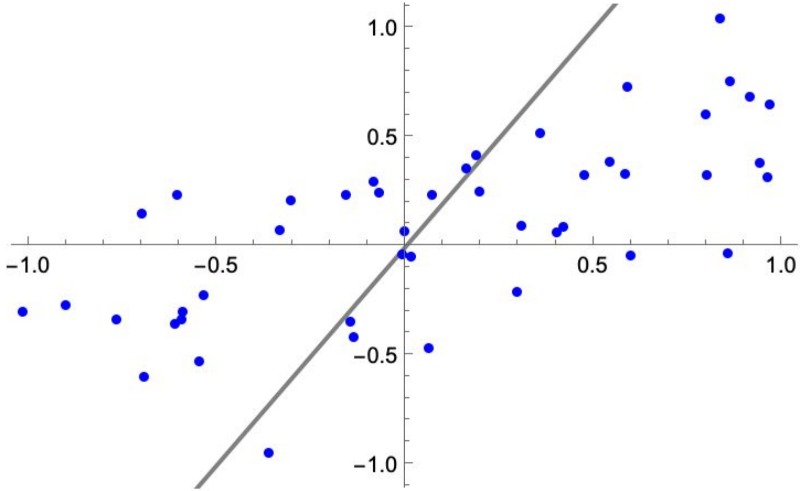
Linear Algebra Basics

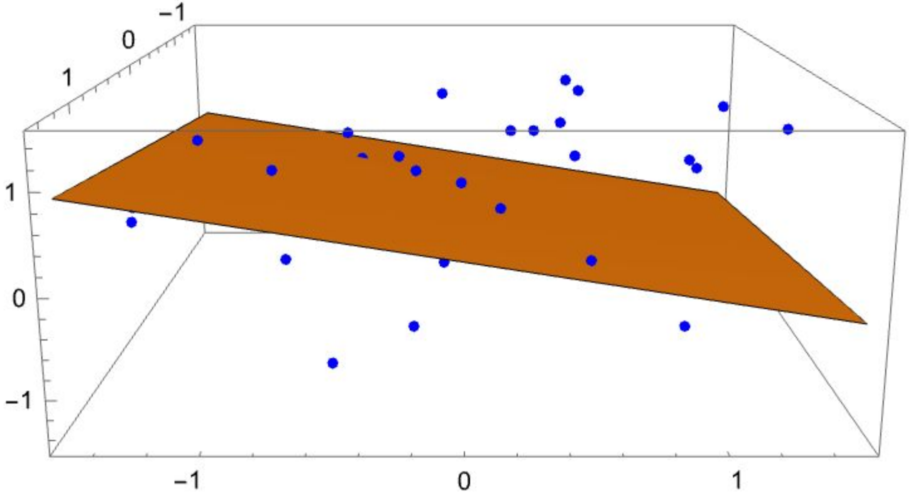
Linear Transformations

Suppose we have N measurements of a certain feature: $X_1, X_2, \ldots, X_N$

Orthogonal Projections

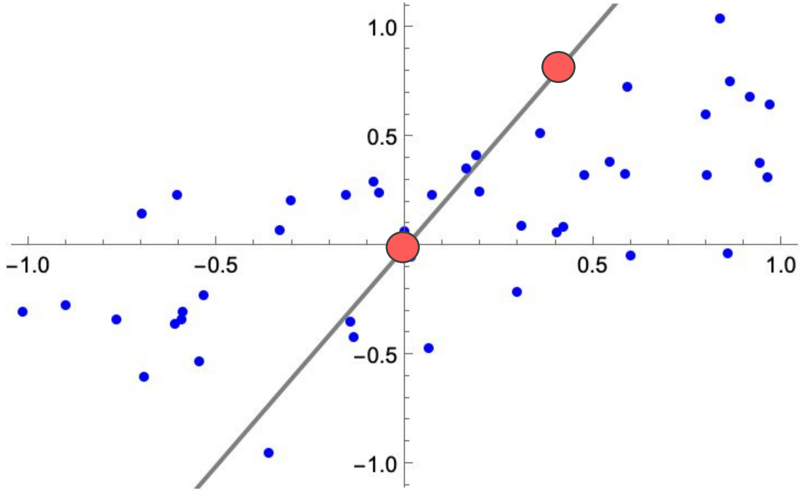
- Orthogonal Projections are a type of Linear Transformation
- Let V be a n-dimensional subspace of Rn (line in R2, line or plane in R3, etc.)
- ORthogonal projections minimize distance between points and projections.

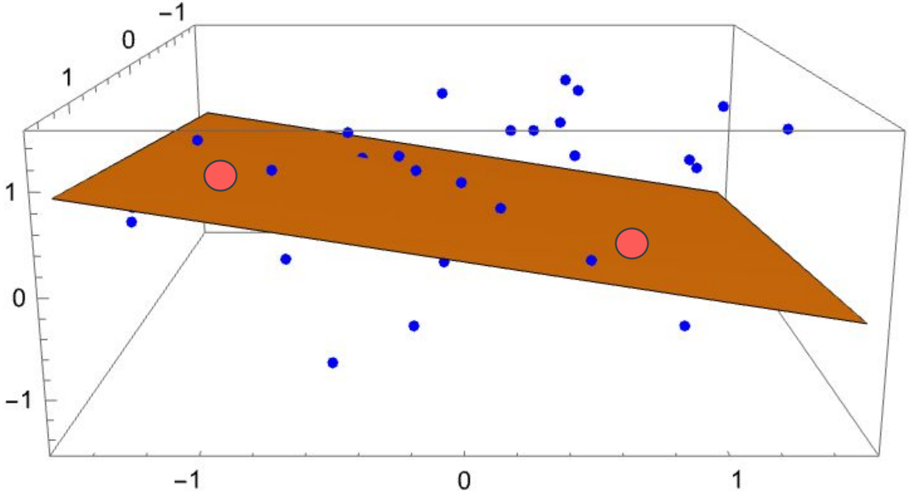




Orthogonal Projections

- Orthogonal Projections have special points:
  - Points who’s value doesn’t chage
  - Points who’s value becomes 0





Eigenvalues & Eigenvectors

Eigenvectors and eigenvalues are specific properties of square (nxn) matrices.
Definition is not super important.

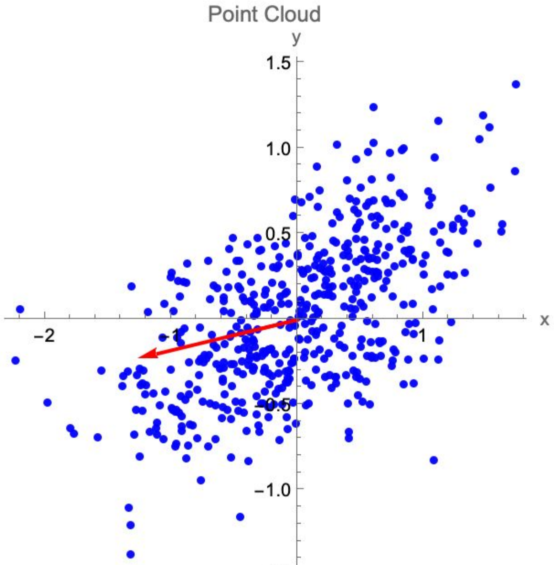
Suppose A is an $x \times x$ matrix. A nonzero vector $\vec{v}$ in $\mathbb{R}^n$ is an **eigenvector** of A of **eigenvalue** λ if

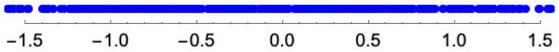
$$A\vec{v} = \lambda\vec{v}.$$

Principal Component Analysis

Principal Component Analysis

- PCA **finds the direction(s) in which the data varies the most** (i.e., is most spread out), and
- **projects the data onto those directions** to reduce dimensionality while preserving as much variance as possible.





Step 0: Gathering The Data

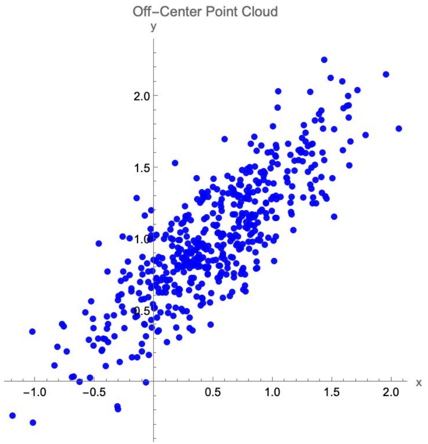
- Consider a multidimensional dataset consisting of **N** observations of **m** different characteristics X_i :

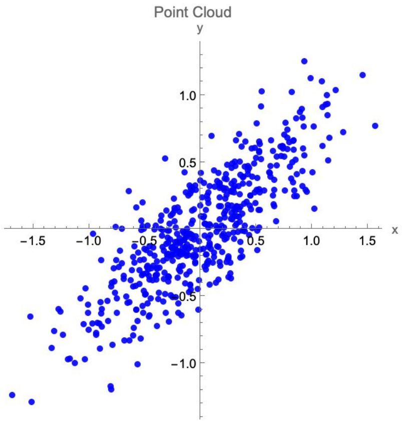
$$(X_1^{(1)}, X_2^{(1)}, X_3^{(1)}, \dots, X_m^{(1)})$$
$$(X_1^{(2)}, X_2^{(2)}, X_3^{(2)}, \dots, X_m^{(2)})$$
$$\vdots$$
$$(X_1^{(N)}, X_2^{(N)}, X_3^{(N)}, \dots, X_m^{(N)})$$

- This data lives in a high-dimensional space $\mathbb{R}^m$ that is “impossible” for us to visualize

Step 1: Standardizing The Data

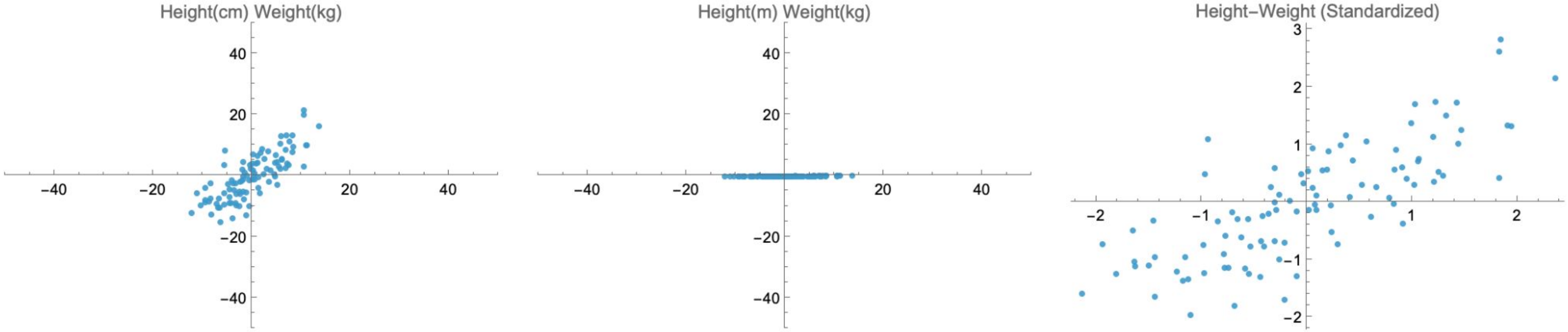
- Center the data: $X_i - \overline{X}$
- Standardize (typically): $\dfrac{X_i - \overline{X}}{\sigma}$
- Point cloud before/after centering





Step 1: Standardizing The Data

- Visual: why is it important to standardize?
  - Remove dependency on units
  - Get rid of scaling differences



Height vs weight of a 100 person sample

Step 2: Finding Covariance Matrix

- Find the covariance matrix:

$$\mathrm{Cov}(\vec{X}) = \begin{bmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) & \cdots & \mathrm{Cov}(X_1, X_m) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}(X_2) & \cdots & \mathrm{Cov}(X_2, X_m) \\ \vdots & & \ddots & \vdots \\ \mathrm{Cov}(X_m, X_1) & \mathrm{Cov}(X_m, X_2) & \cdots & \mathrm{Var}(X_m) \end{bmatrix}$$

- Computational shortcut: if M is the matrix of your standardized data. Then

$$\mathrm{Cov}(\vec{X}) = \frac{1}{N} M^T M$$

Step 3: Finding Eigenvectors and Eigenvalues

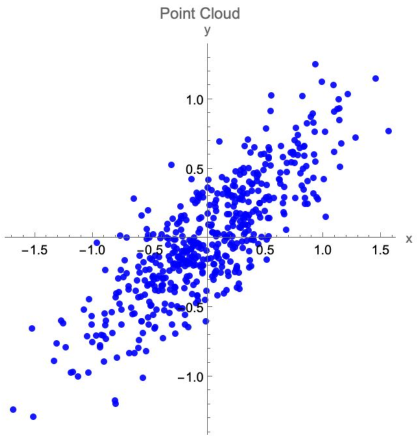
If the data is standardized, the eigenvalues of the covariance matrix $\mathrm{Cov}(\vec{X})$ measure the proportion of the variance in the direction of the corresponding eigenvectors.

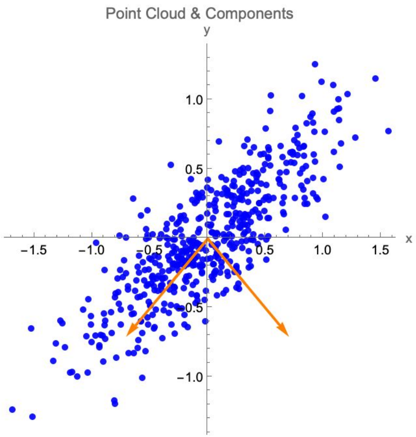
$$\sum_{i=1}^{N} \lambda_i = \sum_{i=1}^{N} \mathrm{Var}(X_i) = N$$

The eigenvector of largest eigenvalue will determine the first principal component,
The eigenvector of second largest eigenvalue will be the second principal component,
Etc.

Step 4: Choose Number of Principal Components

- Choose a number N of principal components
- Pick the eigenvectors with the largest N eigenvalues





First two principal components

Step 4: Choosing Number of Principal Components

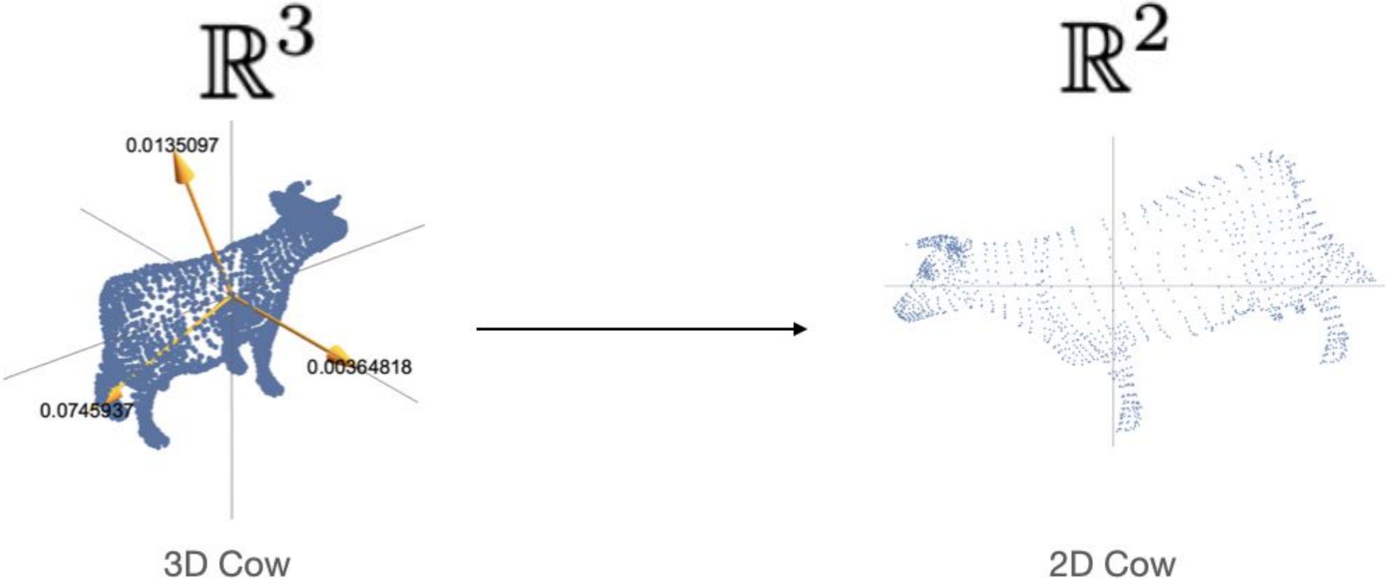
Deciding the number of components onto which the

- For visualization purposes 2 or 3 (obvious reasons)
- Elbow Rule
- Scree Plot

Feature Extraction

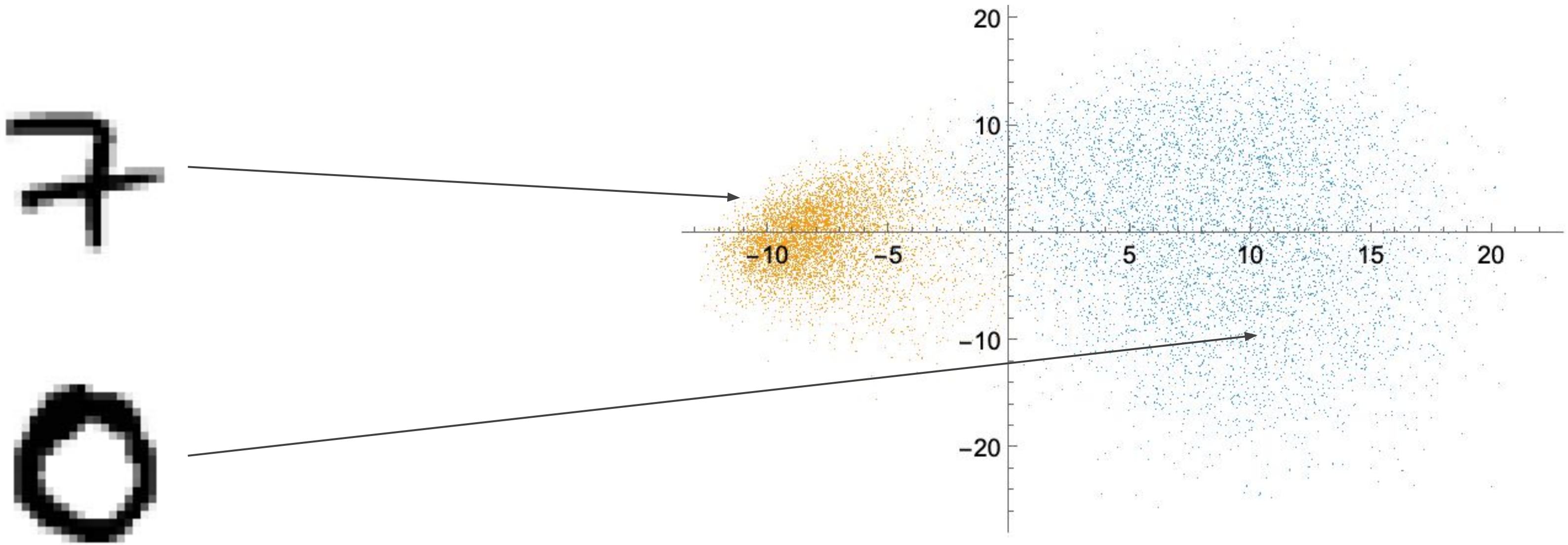
Example: From $\mathbb{R}^3$ to $\mathbb{R}^2$

Projection onto the **first two** principal components.



Example: MNIST Classification

$$\mathbb{R}^{784} \longrightarrow \mathbb{R}^2$$



Example: Medical Imaging

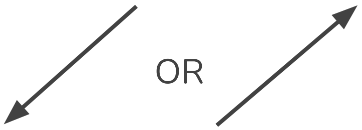
$$\mathbb{R}^{784} \longrightarrow \mathbb{R}^2$$

Principal Component Analysis: Summary

- Standardize (or center) each feature
- Compute covariance matrix
- Find eigenvectors and eigenvalues of the covariance matrix
  - The eigenvalues represent the proportion of overall variance in the direction of the eigenvector
  - Select a number of eigenvectors according to their eigenvalues
  - Project the data onto those eigenvectors
  - Find combinations of feature that are more relevant to the overall variance

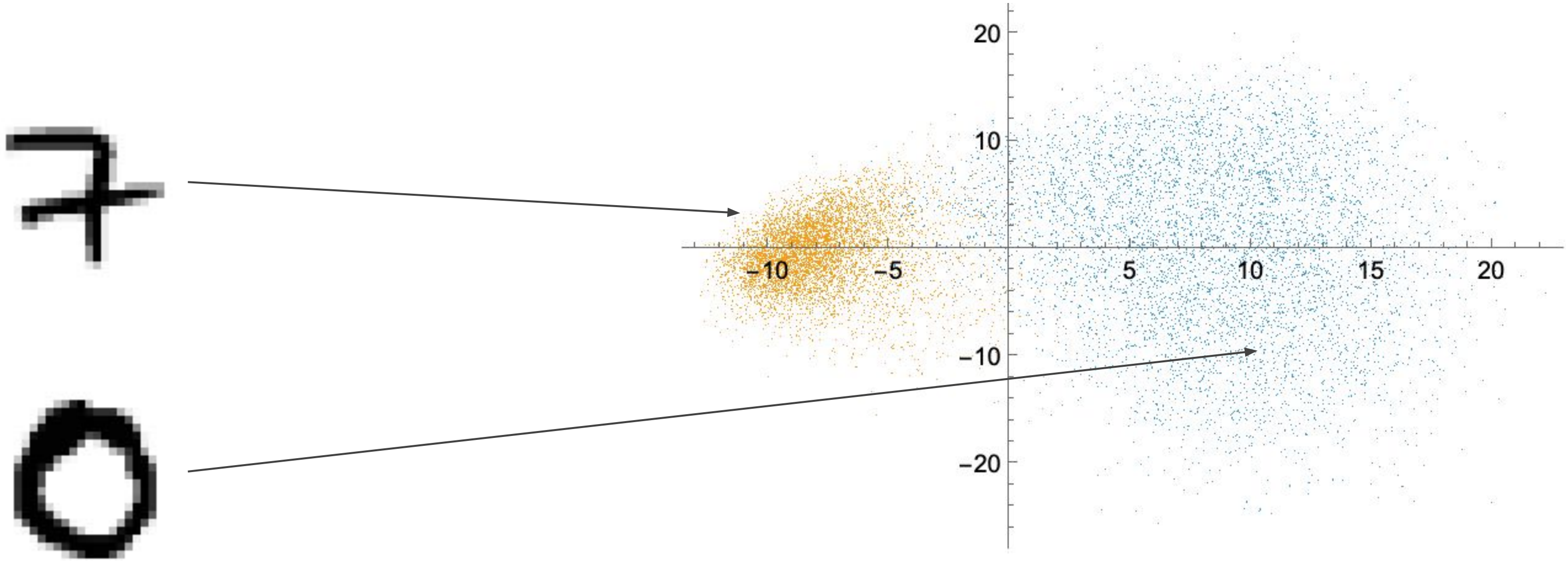
Subtleties, Remarks, and Coding

- Built-in algorithms will center your data, but (typically) won’t standardize it.
- There is a sign ambiguity when choosing the eigenvectors.
- There are some rules about how many principal components eigenvectors to choose.



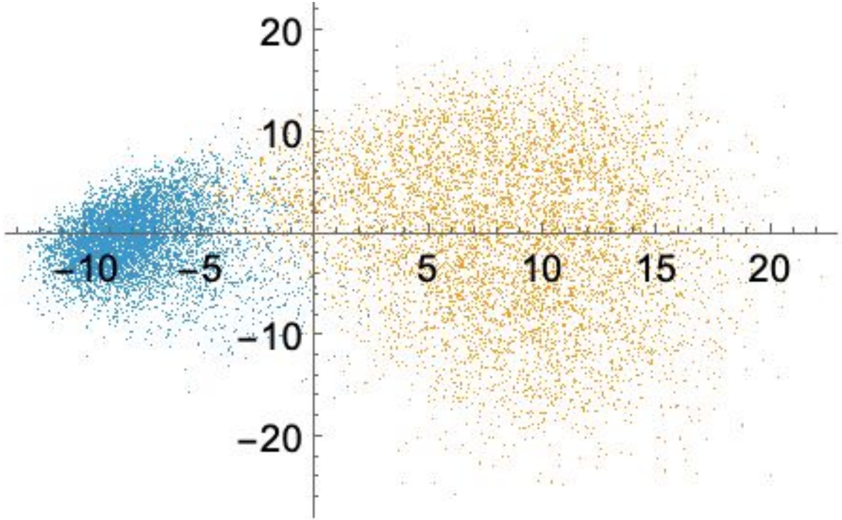
PCA + Other Algorithms

Running Example: Classifying 0s and 7s

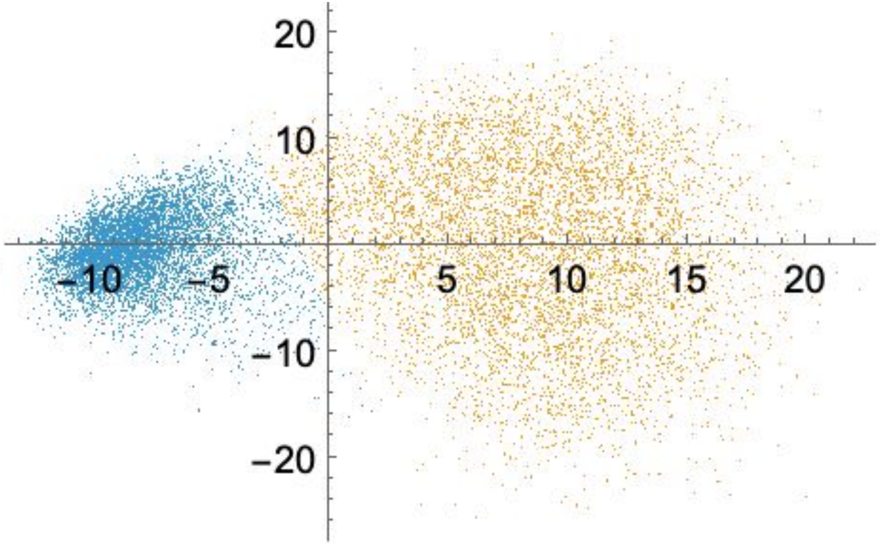


PCA + Logistic Regression

- First run Logistic Regression
- Then apply PCA
- Timing: 27s

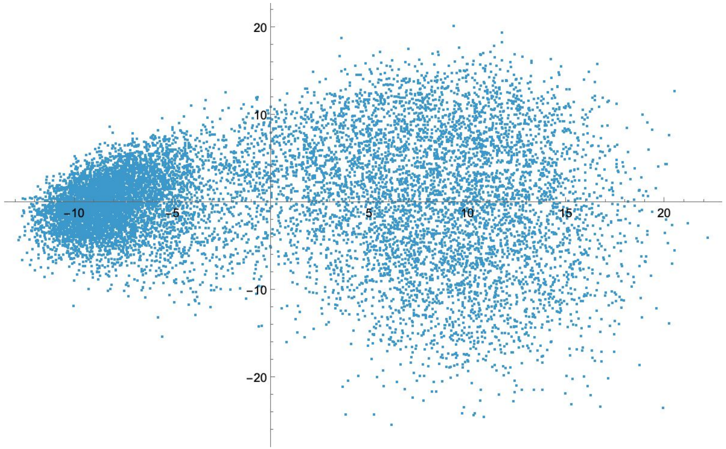


- First apply PCA
- Then run Logistic Regression
- Timing: 1.9s

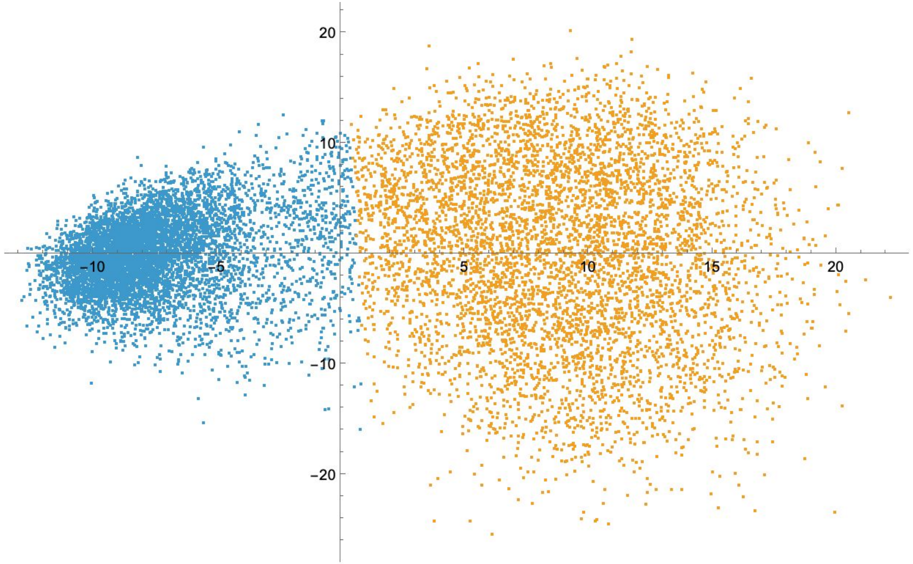


PCA + Clustering

- First find 2 clusters
- Apply PCA
- Timing: 2.6s
- Finds 1 cluster (and 1 singleton)



- First apply PCA
- Find 2 clusters
- Timing: 0.5s



Other Dimensional Reduction Algorithms