Principle Component Analysis

CS 556

High Dimensional Data

- Nowadays data are high dimensional
- Example
  - 300x300 image, each pixel is a tuple (Red, Green, Blue)
  - House price datasets can contains tens or hundreds of features

Challenges of High Dimensional Data

- Hard to analyze
- Interpretation is difficult
- Impossible visualization
- Computationally expensive
- Lie on lower dimensional space

Mean

Mean denoted by μ is the average value in a collection of numbers.

$$\boxed{\mu = \frac{1}{n}\sum_{i=1}^{n} x_i}$$

$$X = \{3,7,5\}$$

$$\mu = \frac{3+7+5}{3} = 5$$

Variance

Variance denoted by σ^2 is a statistical measurement of the spread between the numbers in a data set.

$$\boxed{\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2,\ \mu = E[X]}$$

$$X = \{3,7,5\}$$

$$\mu = \frac{3+7+5}{3} = 5$$

$$\sigma^2 = \frac{1}{3}((3-5)^2 + (7-5)^2 + (5-5)^2) = \frac{8}{3}$$

Covariance

Covariance is a statistical measure of the strength and sign of the linear relationship between two variables in the scale of the original data.

$$Cov[X, Y] = \frac{1}{n-1} \sum_{i=1}^{n} [(x - \mu_x)(y - \mu_y)]$$

Covariance Matrix

$$\begin{bmatrix} Var[X] & Cov[X,Y] \\ Cov[Y,X] & Var[Y] \end{bmatrix}$$

How to construct Covariance Matrices

Assume we have the following dataset:

	Study Time(ST)	Exam Score (ES)
Student 1	10	90
Student 2	6	80
Student 3	20	100

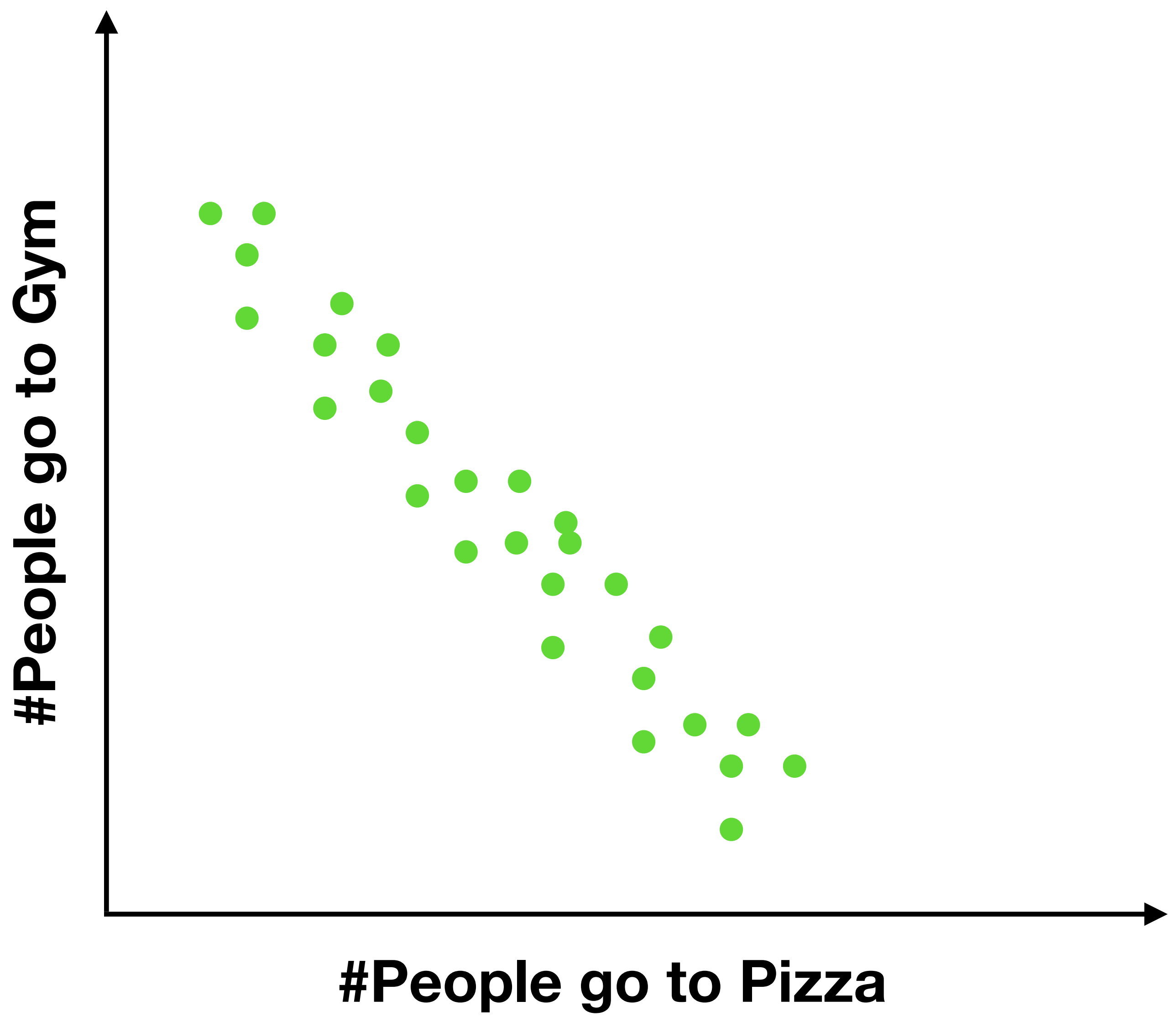
$$D = \begin{bmatrix} 10 & 90 \\ 6 & 80 \\ 20 & 100 \end{bmatrix} \qquad X = \begin{bmatrix} -2 & 0 \\ -6 & -10 \\ 8 & 10 \end{bmatrix}$$

$$COV = \begin{bmatrix} Var(ST) & Cov(ST, ES) \\ Cov(ES, ST) & Var(ES) \end{bmatrix}$$

$$COV = \frac{1}{n-1} X^T X = \frac{1}{n-1} \begin{bmatrix} -2 & -6 & 8 \\ 0 & -10 & 10 \end{bmatrix} \begin{bmatrix} -2 & 0 \\ -6 & -10 \\ 8 & 10 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 104 & 140 \\ 140 & 200 \end{bmatrix} = \begin{bmatrix} 52 & 70 \\ 70 & 100 \end{bmatrix}$$

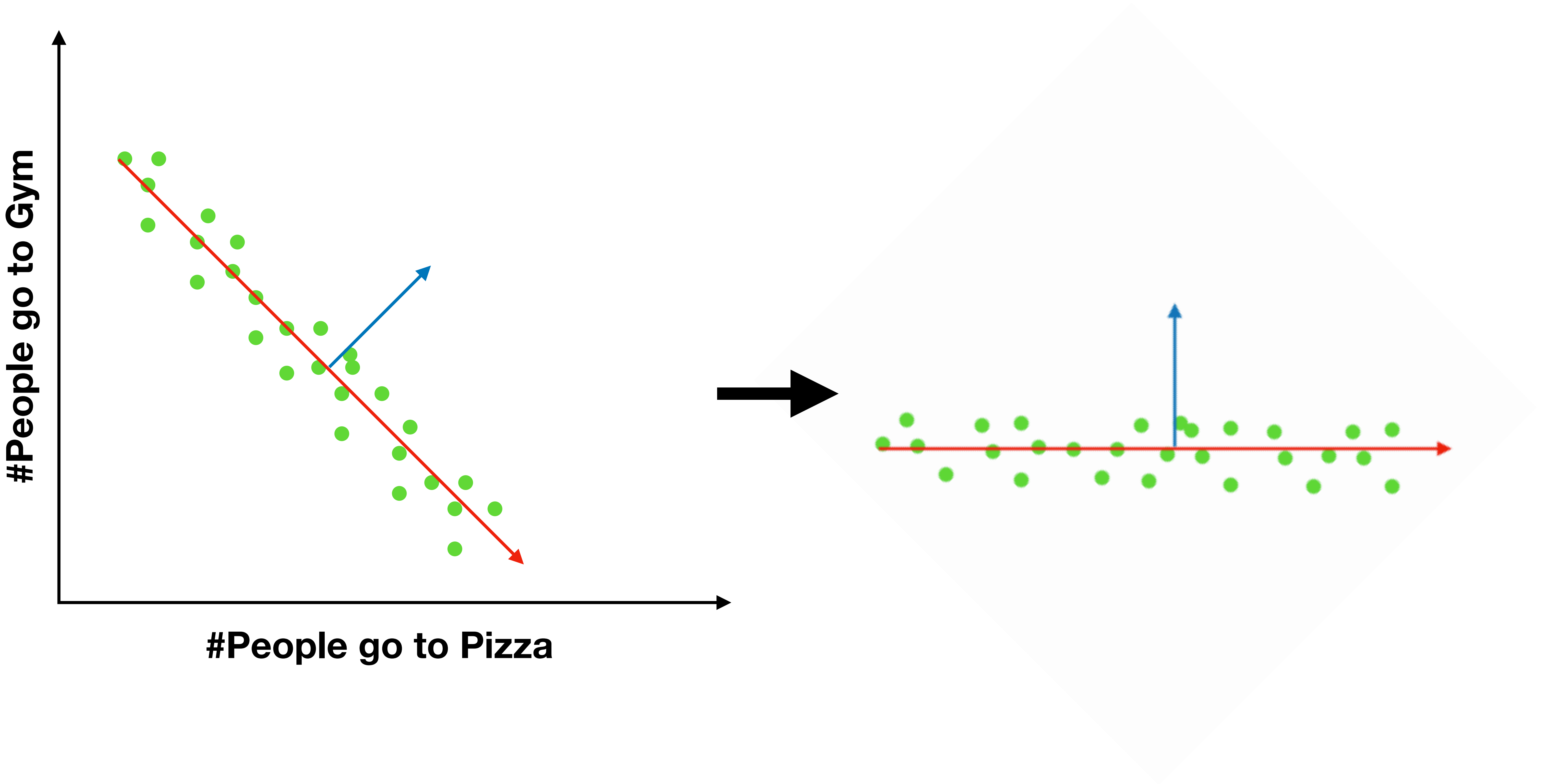


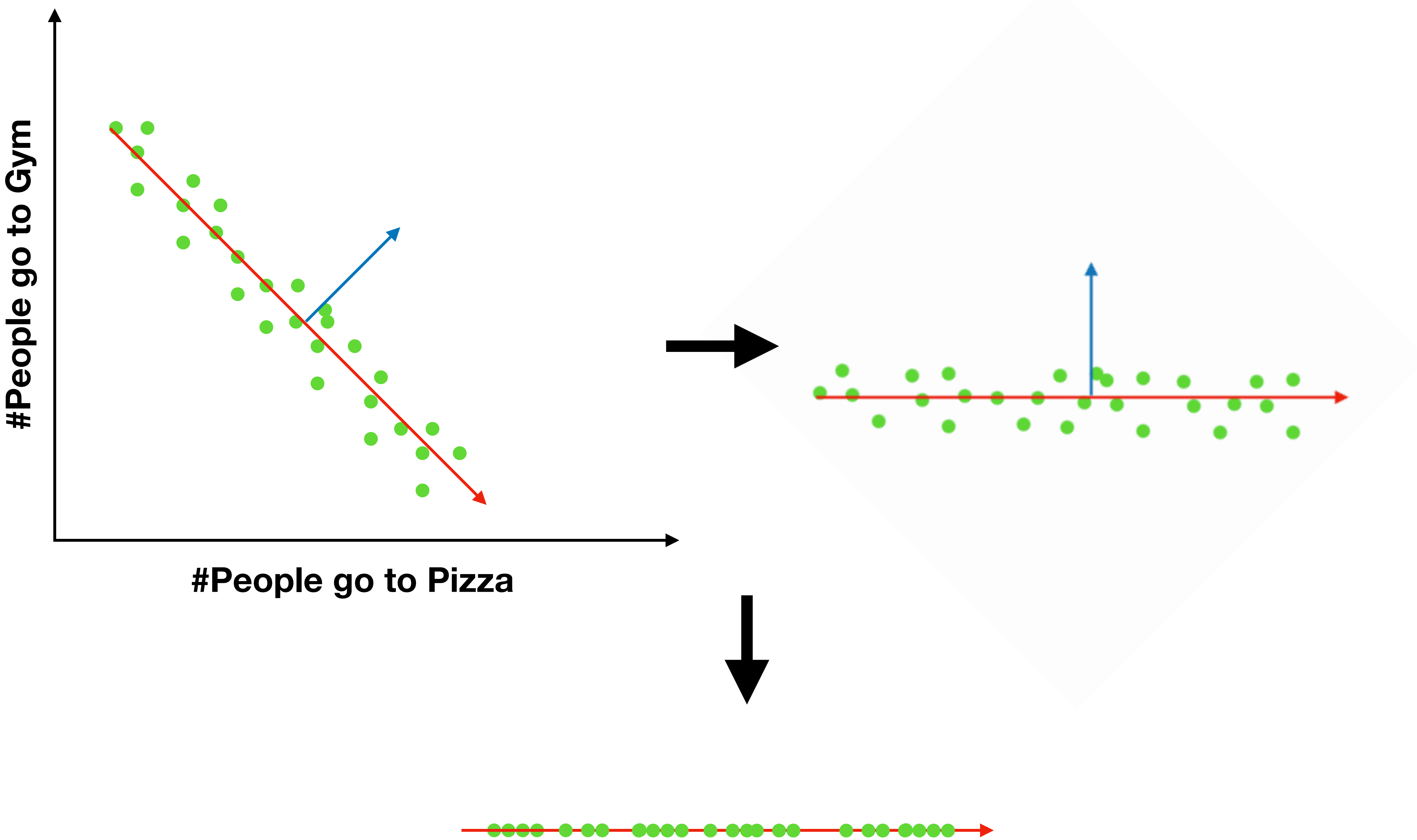
GYM


#People go to Gym
#People go to Pizza


#People go to Gym
#People go to Pizza

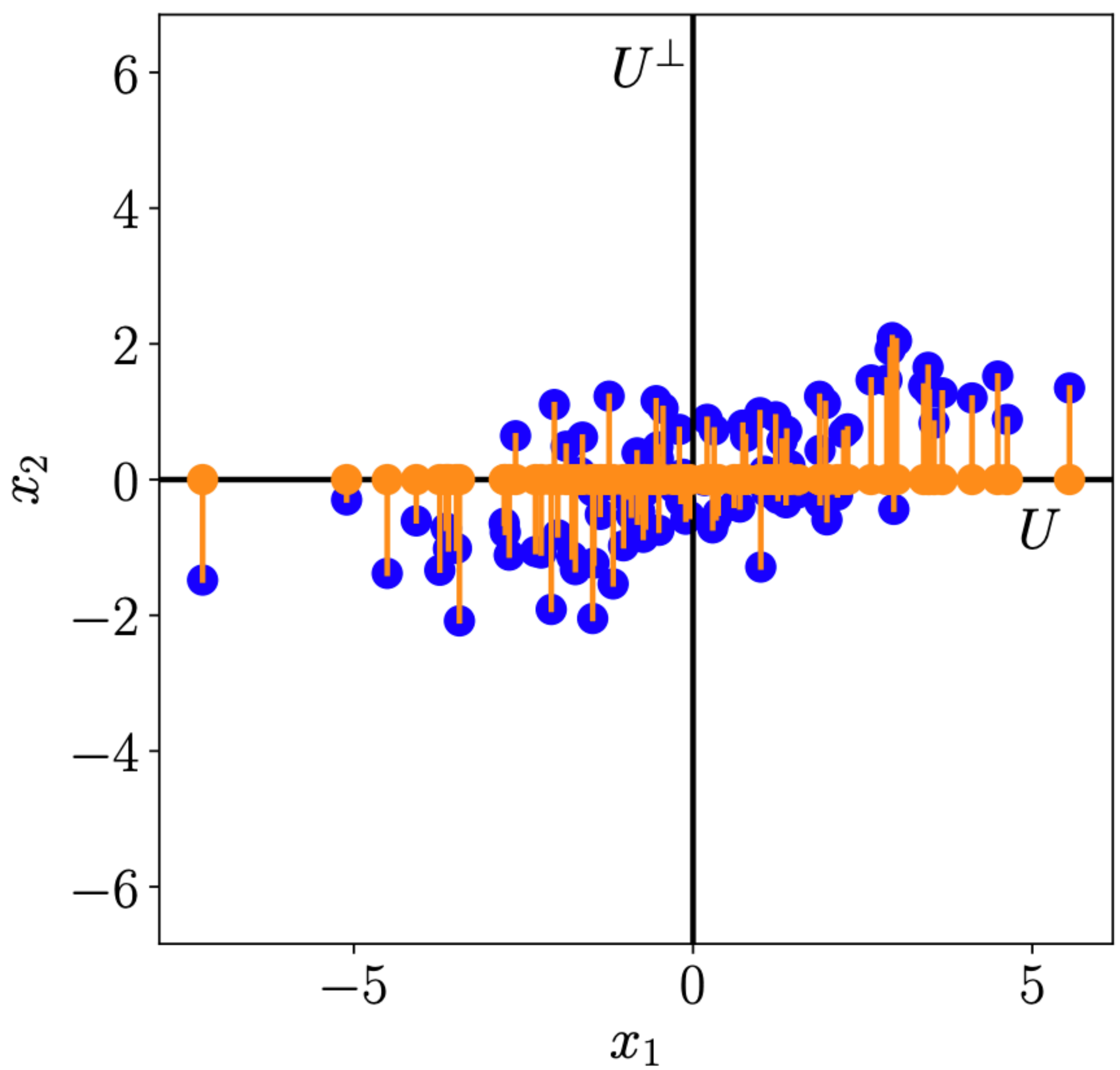

#People go to Gym
#People go to Pizza


#People go to Gym
#People go to Pizza


#People go to Gym
#People go to Pizza

PCA: Key Idea

Use orthogonal projections to find lower dimensional representations of data that retain as much information as possible.



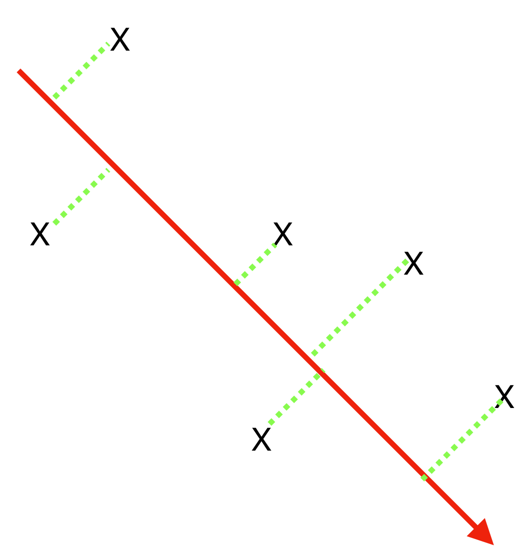
Why PCA?

- Visualize data in a lower-dimensional space
- Understand the sources of variability in the data
- Understand correlations between different coordinates of the data points

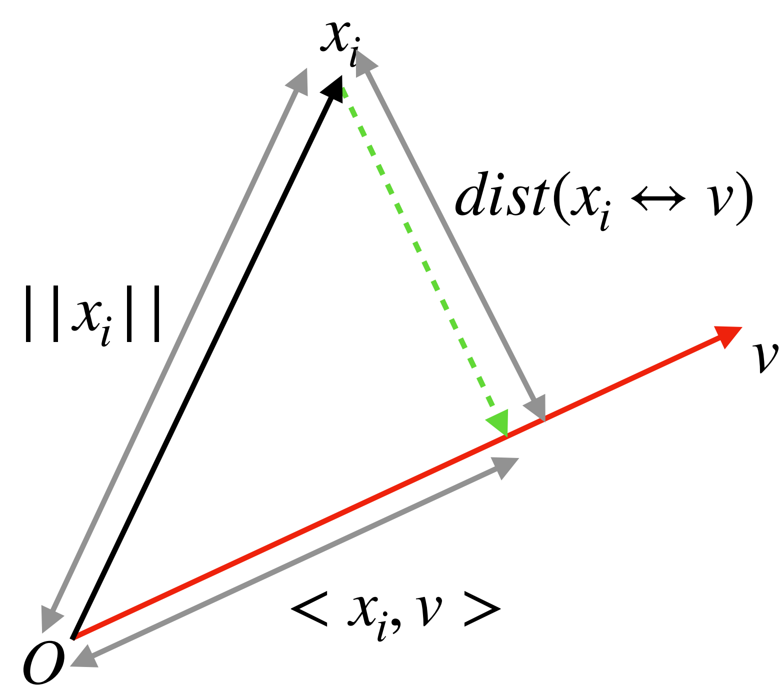
Objective Function

For a given data set and parameter k, to goal of PCA is to compute the k-dimensional subspace that minimizes the average squared distance between the points and the subspace.

$$\underset{k-dim\ spaces\ S}{argmin} \frac{1}{n} \sum_{i=1}^{n} ((distance\ of\ x_i\ from\ S)^2)$$



Objective Function



$$\underset{\mathbf{v}:\|\mathbf{v}\|=1}{\operatorname{argmin}} \frac{1}{m} \sum_{i=1}^{m} \left((\text{distance between } \mathbf{x}_i \text{ and line spanned by } \mathbf{v})^2 \right)$$

$$(\operatorname{dist}(\mathbf{x}_i \leftrightarrow \text{line}))^2 + \langle \mathbf{x}_i, \mathbf{v} \rangle^2 = \|\mathbf{x}_i\|^2$$

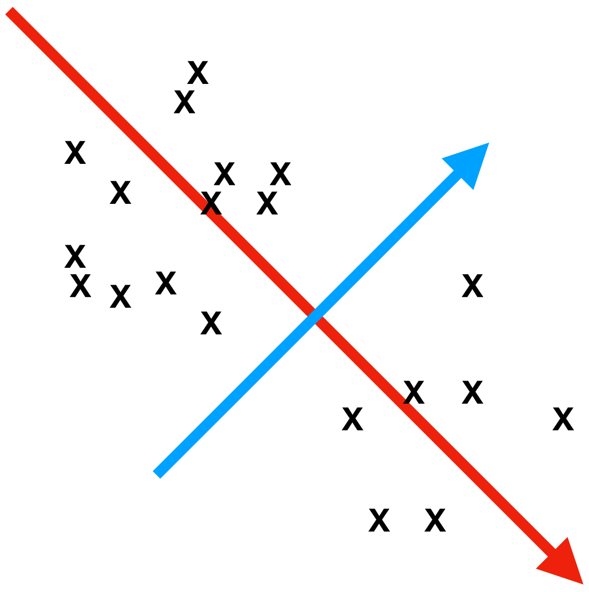
$$\underset{\mathbf{v}:\|\mathbf{v}\|=1}{\operatorname{argmax}} \frac{1}{m} \sum_{i=1}^{m} \langle \mathbf{x}_i, \mathbf{v} \rangle^2$$

Objective Function

Given $x^1, \ldots, x^m \in \mathbb{R}^n$ and a parameter $k \geq 1$, compute orthonormal vectors $v_1, \ldots, v_k \in \mathbb{R}^n$ to maximize:

$$\frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{k} \langle x_i, v_j \rangle^2$$

The resulting k orthonormal vectors are called the top k principal components of the data.



Which is the best principle component?

Characterizing PCs

$x_1, x_2, \ldots x_m \in \mathbb{R}^n$

$k \in \{1,2,\ldots,n\}$

$$\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{k}\langle x_i, v_j\rangle^2$$

$$X = \begin{bmatrix} -x_1- \\ -x_2- \\ \vdots \\ -x_m- \end{bmatrix}, ||v|| = 1 \rightarrow Xv = \begin{bmatrix} \langle x_1, v\rangle \\ \langle x_2, v\rangle \\ \vdots \\ \langle x_m, v\rangle \end{bmatrix}$$

$$For\ k = 1,\ \ argmax_{v:||v||=1}\frac{1}{m}\sum_{i=1}^{m}\langle x_i, v\rangle^2$$

$$argmax_{v:||v||=1}\frac{1}{m}(Xv)^T Xv = argmax_{v:||v||=1} v^T X^T Xv = argmax_{v:||v||=1} v^T Av$$

A is the symmetric covariance matrix, from spectral theorem $\rightarrow argmax_{v:||v||=1} v^T QDQ^T v$.

The direction of v that maximizes the objective function is the direction of maximum stretch which corresponds to the eigenvector with the highest eigenvalue.

$$\boxed{\text{Principle components corresponds to the k eigenvectors of the covariance matrix that have the largest eigenvalues.}}$$

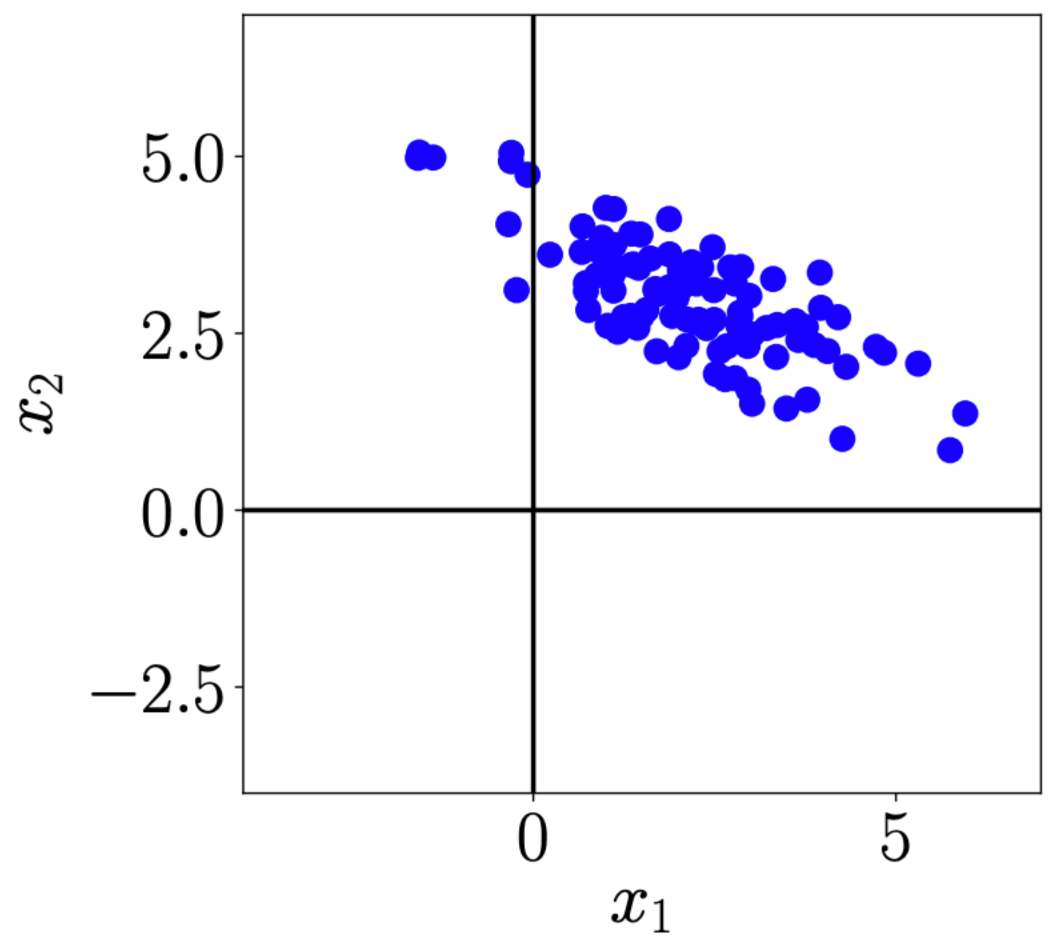
Finding principal components

1. Calculate the eigenvalues and unit eigenvectors of the covariance matrix and order the eigenvectors in descending order with respect to the corresponding eigenvalues.
2. The unit vectors $u_1, u_2, \ldots u_n$ of the covariance matrix represent the principle components of the data. The corresponding eigenvalues give the variance of the principle components.
3. Pick top k principle components $u_1, u_2, \ldots, u_k$ and construct

$$Q_k = [u_1, u_2, \ldots, u_k]$$

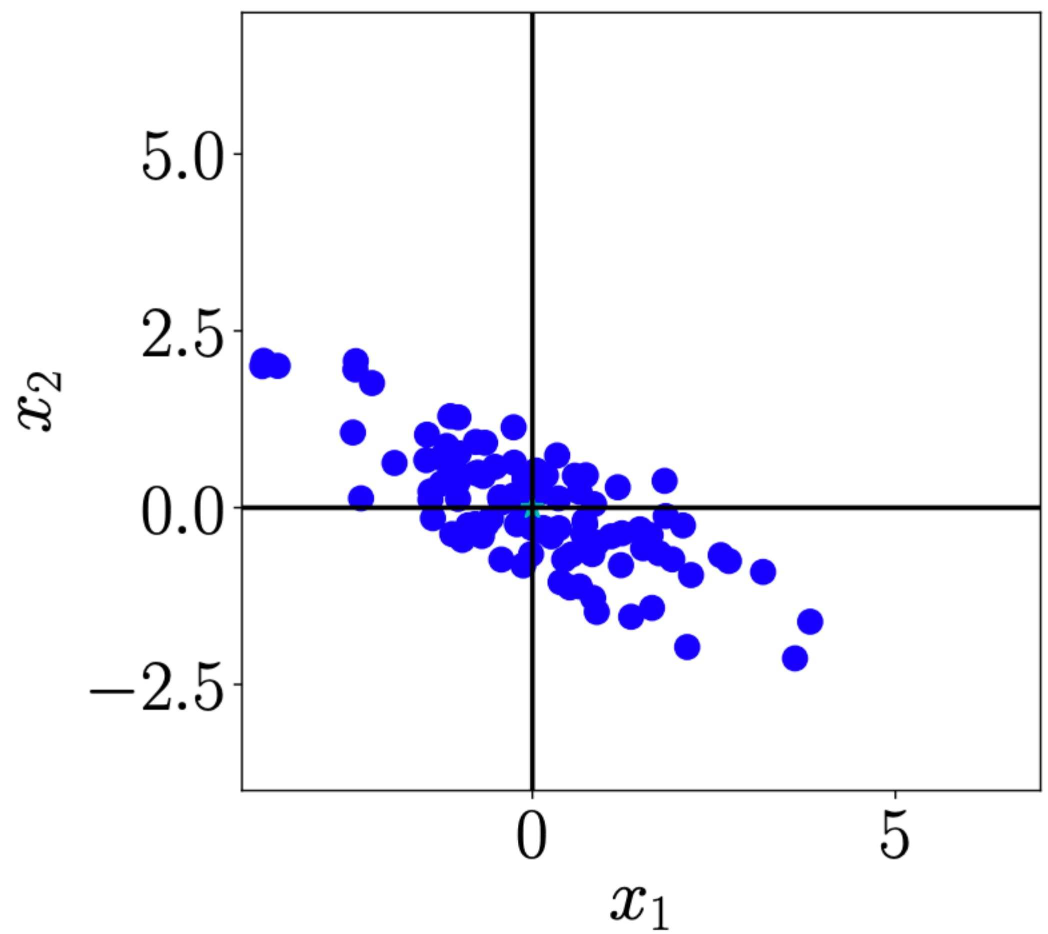
4. Each observation $x \in \mathbb{R}^n$ will be represented as $Q_k^T x$ in the lower dimensional space $\mathbb{R}^k$.

PCA in Practice



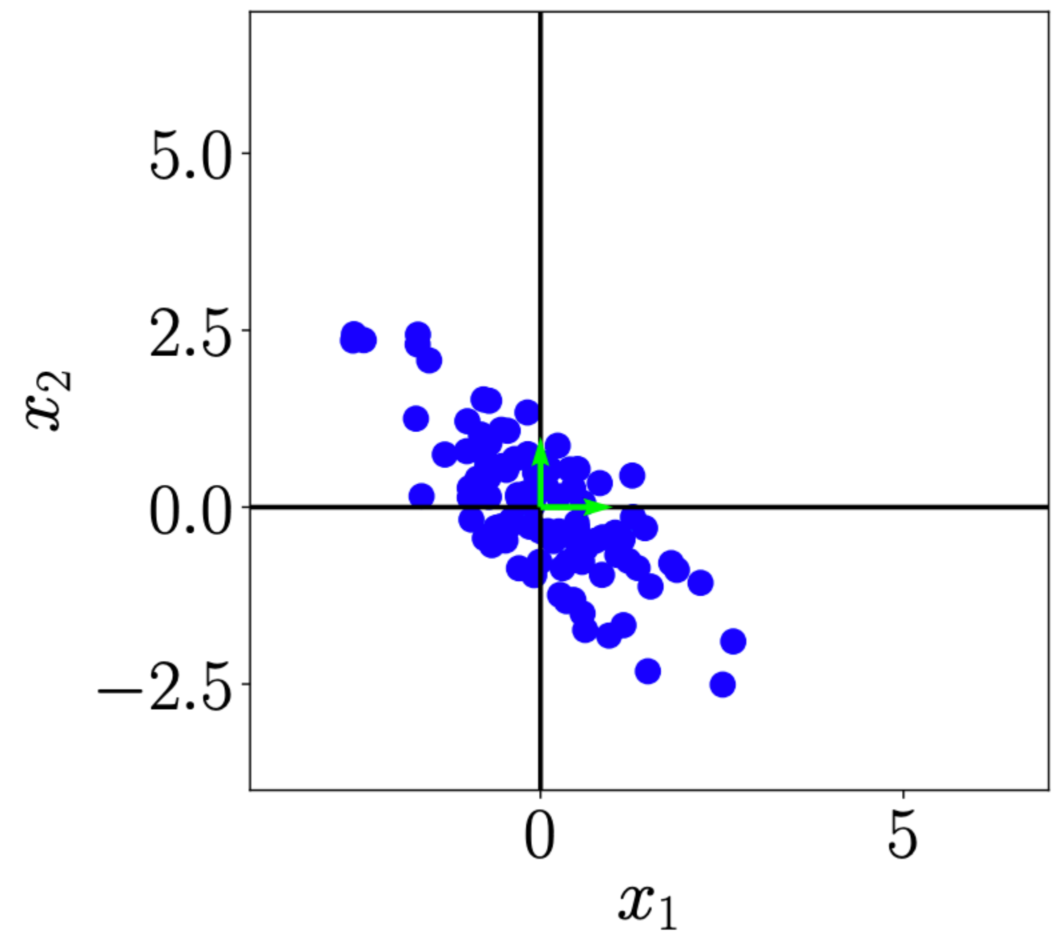
(a) Original dataset.

PCA in Practice



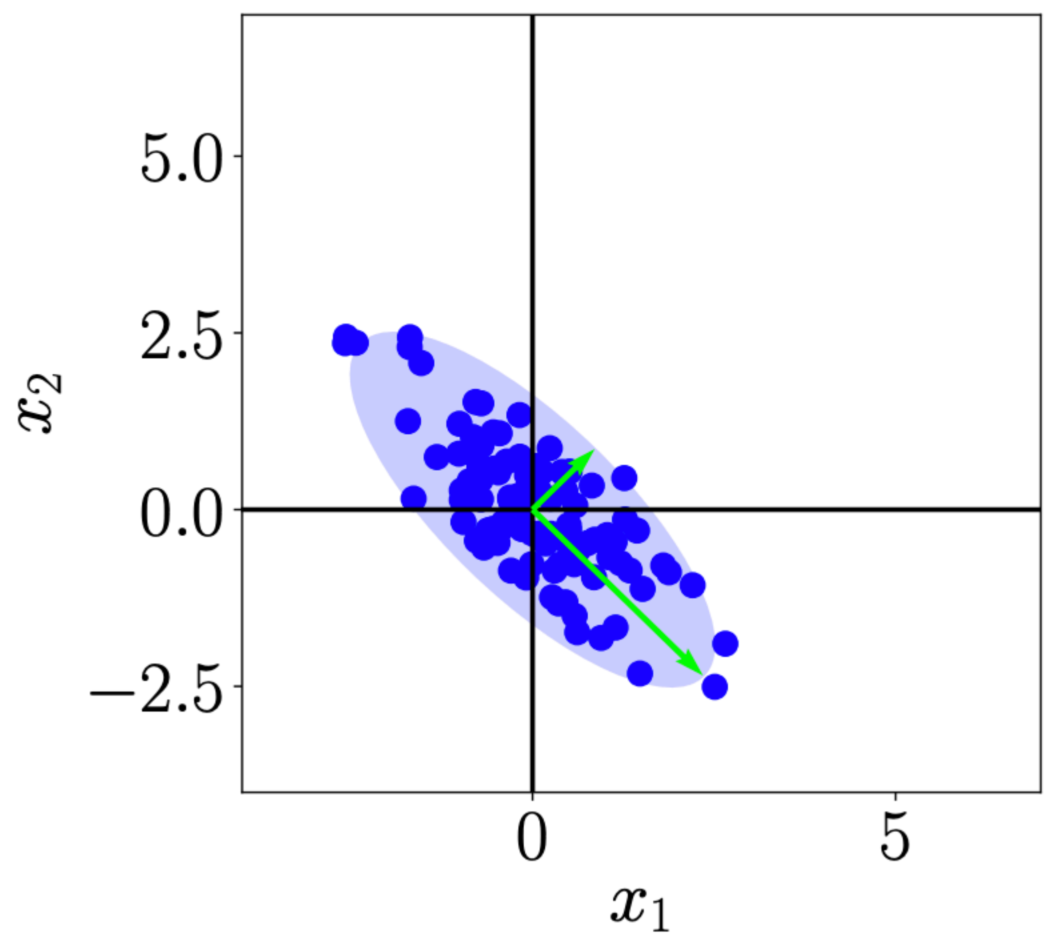
(b) Step 1: Centering by subtracting the mean from each data point.

PCA in Practice



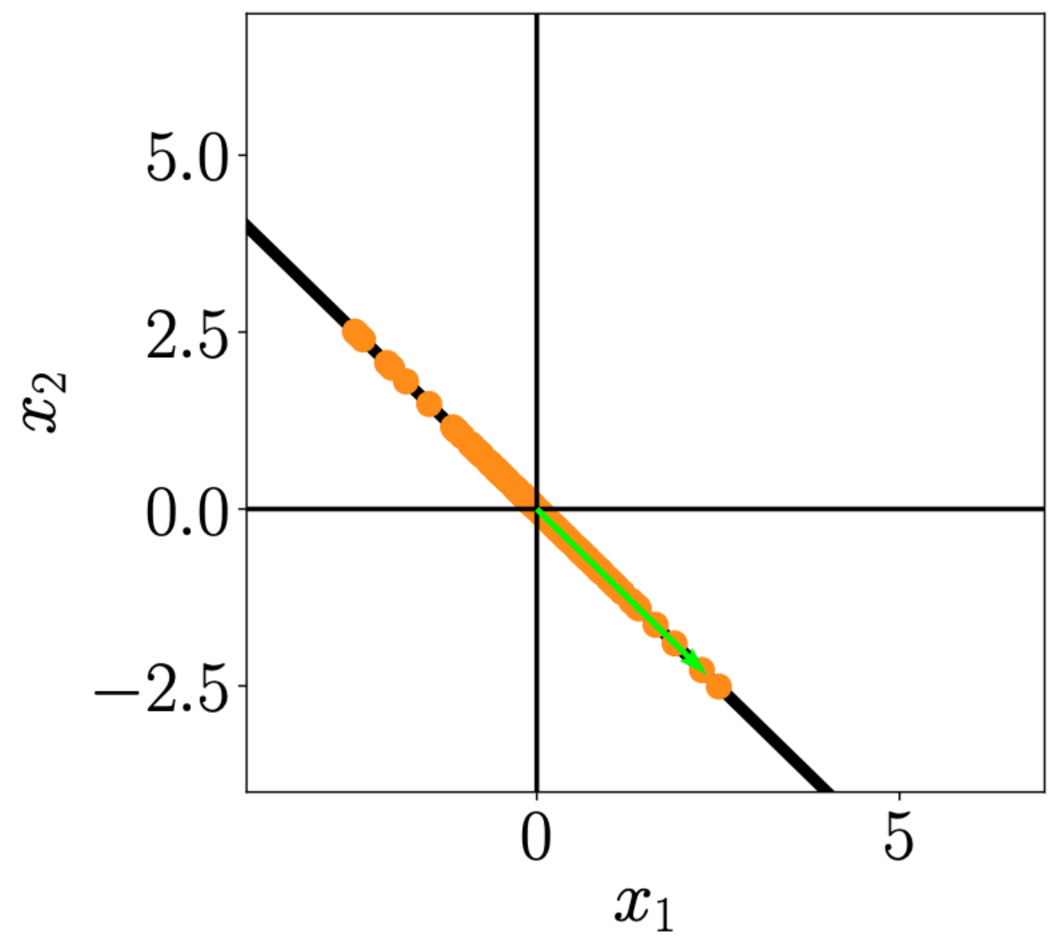
(c) Step 2: Dividing by the standard deviation to make the data unit free. Data has variance 1 along each axis.

PCA in Practice



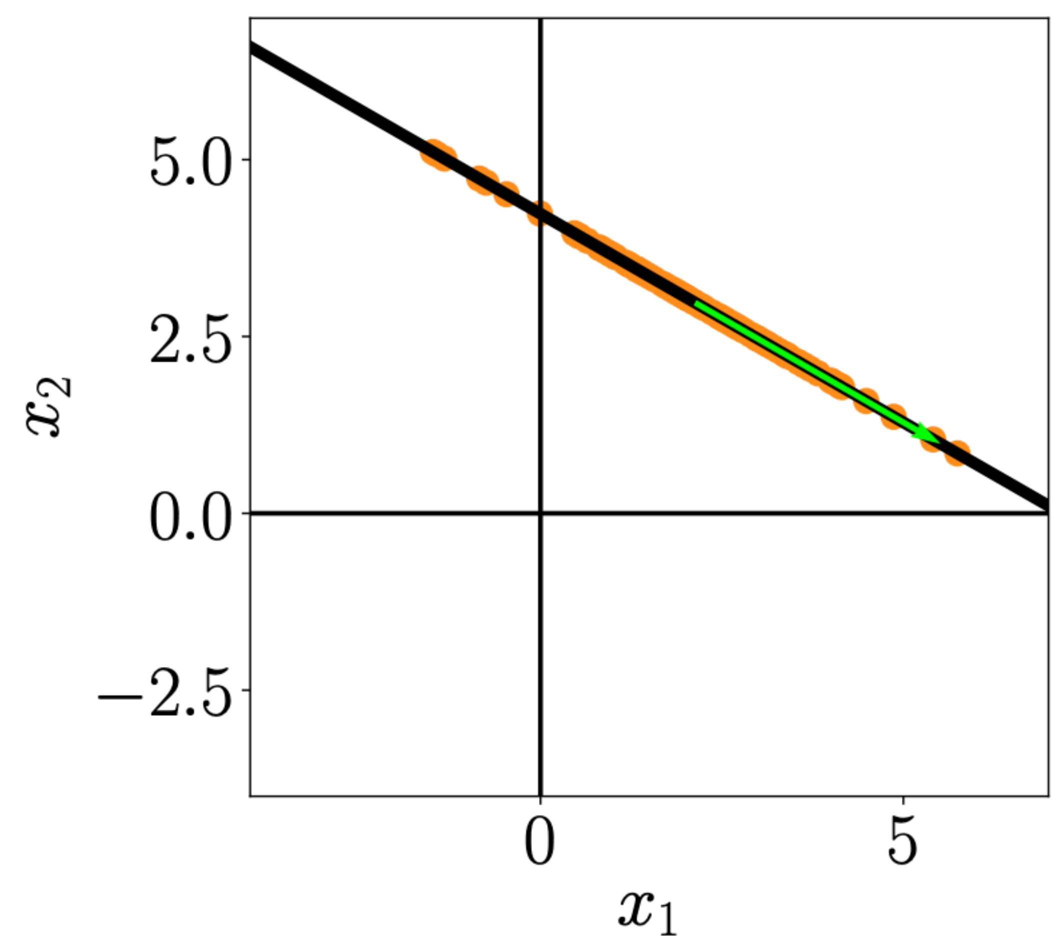
(d) Step 3: Compute eigenvalues and eigenvectors (arrows) of the data covariance matrix (ellipse).

PCA in Practice



(e) Step 4: Project data onto the principal subspace.

PCA in Practice



(f) Undo the standardization and move projected data back into the original data space from (a).

Successful Applications

- Novembre, John, et al.
  "Genes mirror geography within Europe."
  Nature 456.7218 (2008): 98-101.

- Turk, Matthew, and Alex Pentland.
  "Eigenfaces for recognition."
  Journal of cognitive neuroscience 3.1 (1991): 71-86.

Failure Cases

- Wrong scaling/normalization
- Non linear structure in your data
- Non orthogonal structure

Extra Materials

- https://web.stanford.edu/class/cs168/l/l7.pdf
- https://web.stanford.edu/class/cs168/l/l8.pdf
- https://www.youtube.com/watch?v=g-Hb26agBFg&t=1121s