Lecture 11

Dimensionality Reduction

AI/ML Foundation Course with Python

Finding a Penny Problem

Let's say you have a straight line 100 yards long and you dropped a penny somewhere on it.

Now let's say you have a square 100 yards on each side and you dropped a penny somewhere on it. I

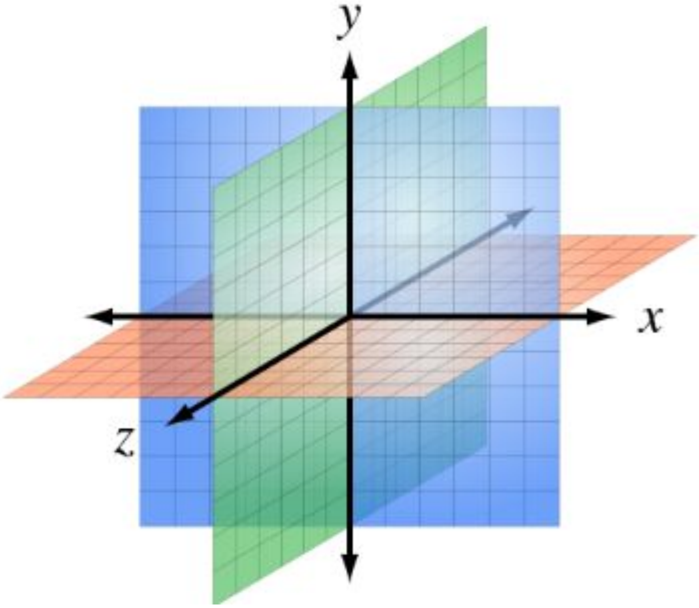
Now a cube of 100 yards across.

The difficulty of searching through the space gets a lot harder as you have more dimensions.

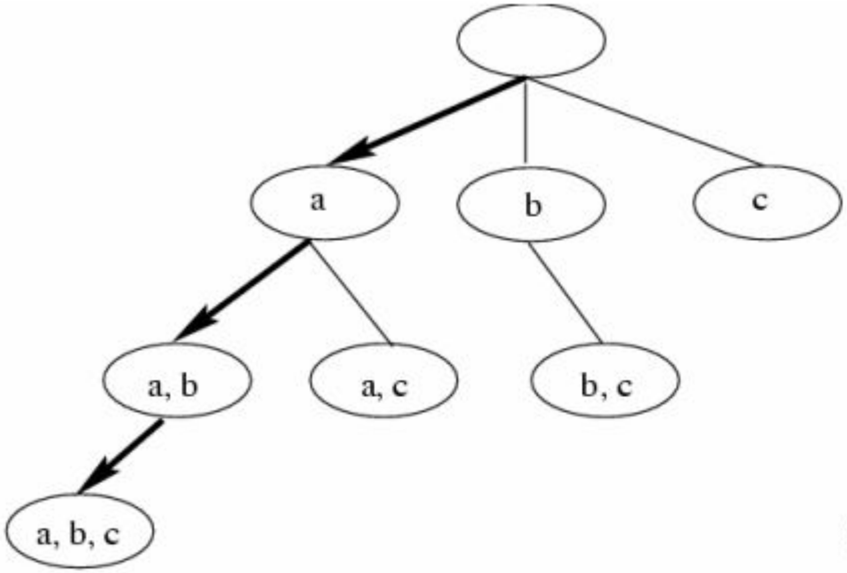
The Curse of Dimensionality

In machine learning, “dimensionality” simply refers to the number of features (i.e. input variables) in your dataset.

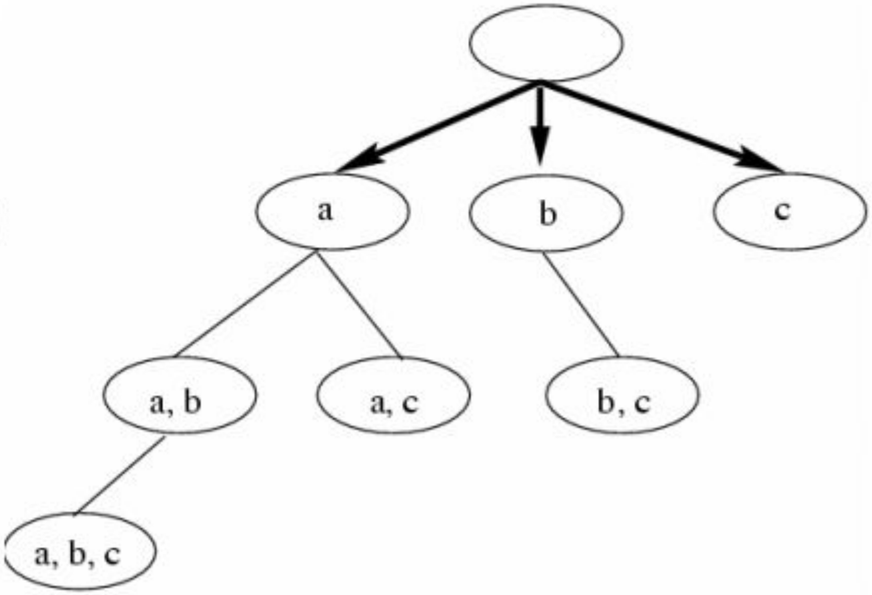
When the number of features is very large relative to the number of observations in your dataset, *certain* algorithms struggle to train effective models. This is called the “Curse of Dimensionality,”



Illustrations of Search Strategies



Depth-first search



Breadth-first search

★ Dimensionality reduction is used to reduce the number of random variables to consider.

★ By obtaining a set of principal variables.

★ It involves feature selection and feature extraction.

★ Dimensionality reduction makes analyzing data much easier and faster for machine learning algorithms without extraneous variables to process.

★ Thus making machine learning algorithms faster and simpler.

Feature selection

★ Feature selection techniques find a smaller subset of multi dimensional data to create a data model

★ Techniques include -

  ○ Wrapper
  ○ Filter
  ○ Embedded

Feature extraction

- Feature extraction techniques involves transforming high dimensional data into spaces of fewer dimensions.

- Techniques include -

    - Principal Component Analysis
    - Kernel PCA
    - Linear Discriminant Analysis (LDA)
    - Generalized Discriminant Analysis

Difference

- Feature selection is for filtering irrelevant or redundant features from your dataset.
- Feature selection keeps a subset of the original features
- While feature extraction creates brand new ones.

Example 1

- ★ A classification problem that relies on both humidity and rainfall can be collapsed into just one underlying principal feature since both of these features are highly correlated.

Example 2

- ★ Email classification Problem
  - ○ We need to class whether an email is spam or not.
  - ○ May involve large number of features to decide whether a particular email is spam or not
  - ○ Email subject
  - ○ Email content
  - ○ Email attachments
  - ○ Email template
  - ○ Email sender

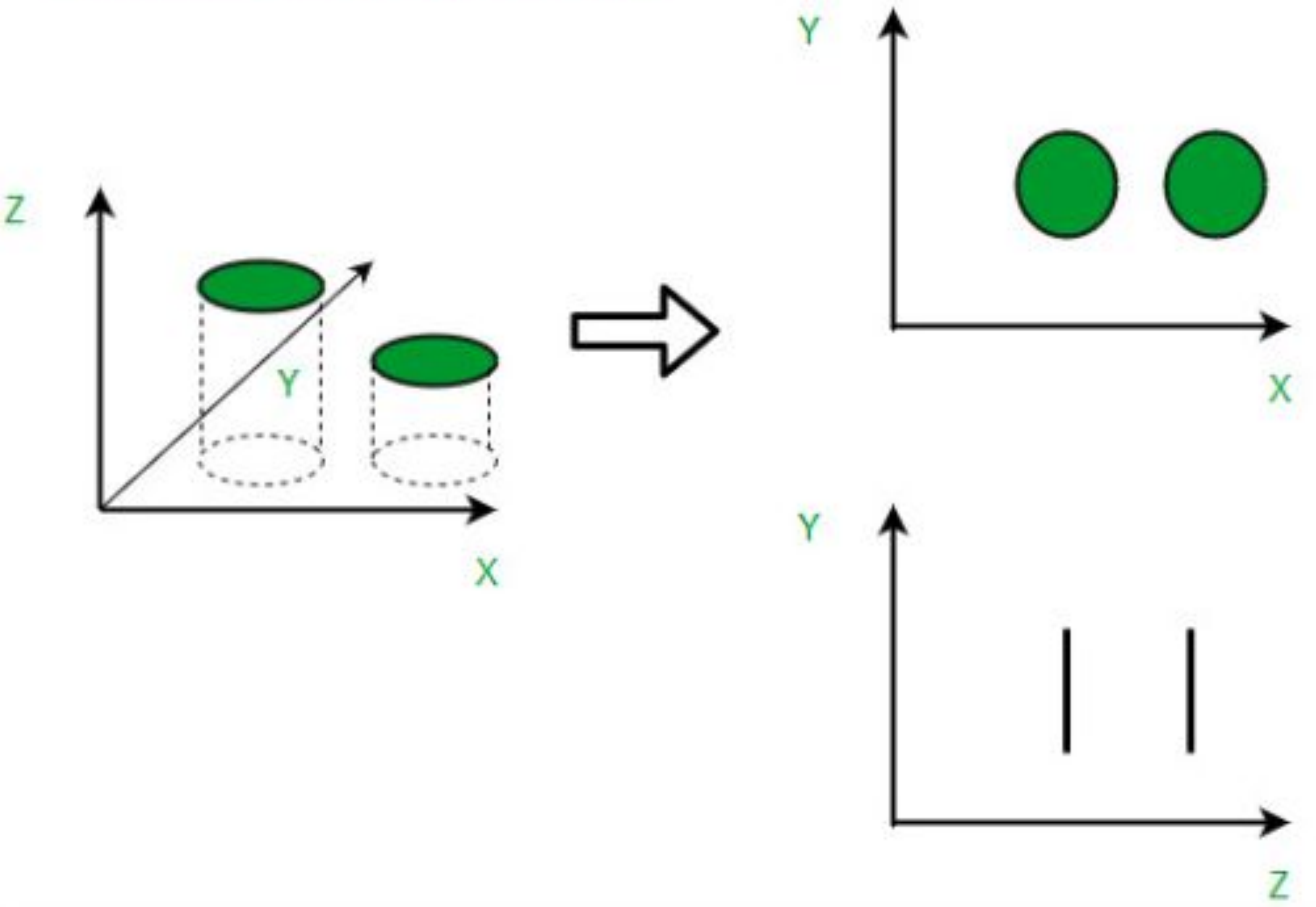
Some features may overlap or be redundant. And can be exterminated from the feature set.

Why

- ★ 3-D classification problem can be hard to visualize
- ★ 2-D one can be mapped to a simple 2 dimensional space
- ★ 1-D problem to a simple line.

Lesser no. of dimensionality to deal with results in simpler models and easy to comprehend visualizations

Dimensionality Reduction



Advantages

- It helps in data compression, and hence reduced storage space.
- It reduces computation time.
- It also helps remove redundant features, if any.

Disadvantages

- It may lead to some amount of data loss.
- PCA tends to find linear correlations between variables, which is sometimes undesirable.
- PCA fails in cases where mean and covariance are not enough to define datasets.
- We may not know how many principal components to keep- in practice, some thumb rules are applied.

Principal Component Analysis

1. PCA reduces the dimensions of a data set by projecting the data onto a lower-dimensional subspace.

2. For example, a two dimensional data set could be reduced by projecting the points onto a line.

   ⇒ Each instance in the data set would then be represented by a single value rather than a pair of values

Principal Component Analysis

1. A three-dimensional dataset could be reduced to two dimensions by projecting the variables onto a plane.

2. Formal Definition

   ⇒ Any n-dimensional dataset can be reduced by projecting the dataset onto a k-dimensional subspace, where k is less than N such that the projected vectors retain the greatest proportion of the original data set's variance.

Example

1. Imagine that you are a photographer for a gardening supply catalog.

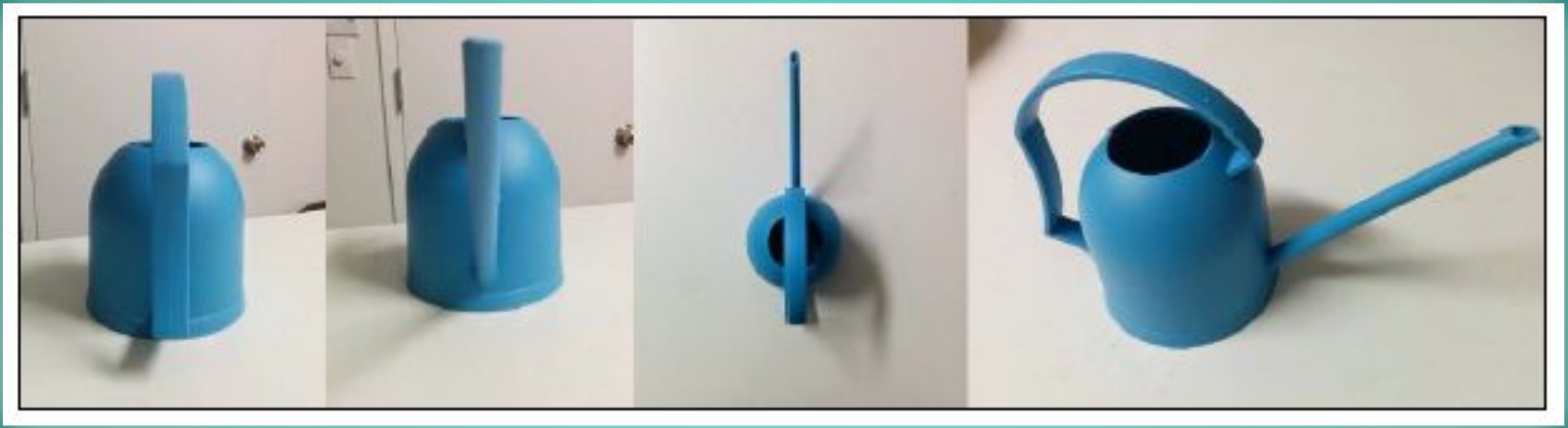
2. You are tasked with photographing a watering can.

3. The watering can is three-dimensional

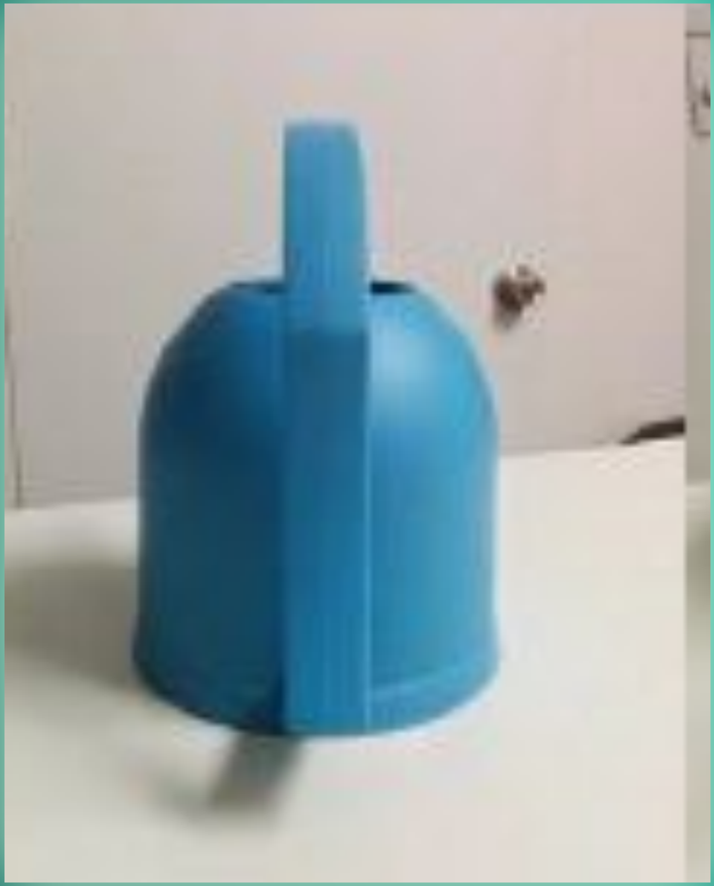
4. But the photograph is two-dimensional

Task : You must create a two-dimensional Representation that describes as much of the watering can as possible.

Possible Pictures



First Photograph

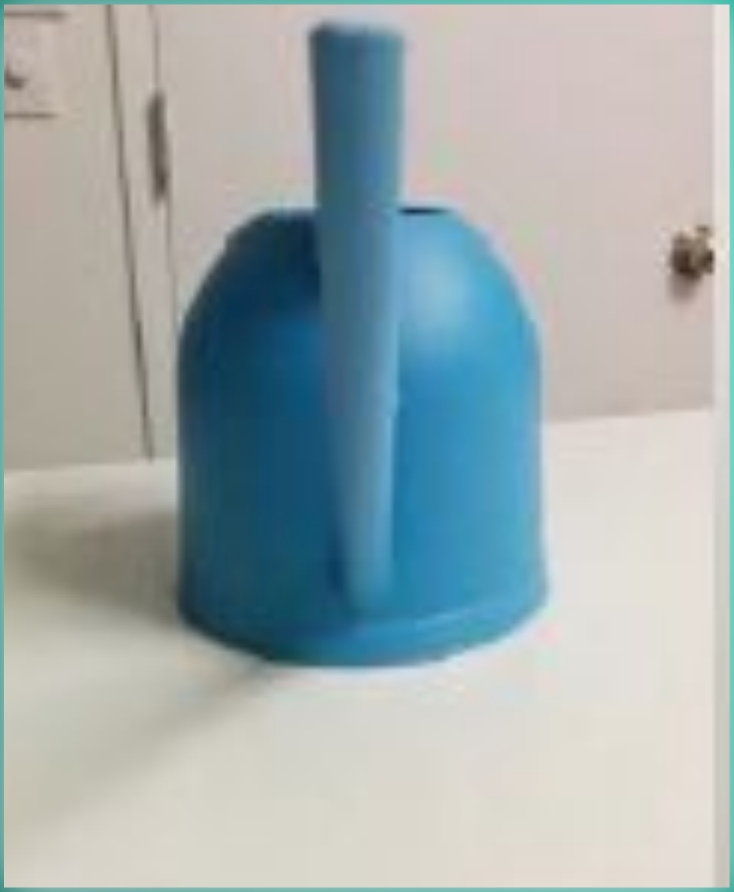


The back of the watering can is visible

But, the Front cannot be seen

Can be considered an accurate representation of the watering can???

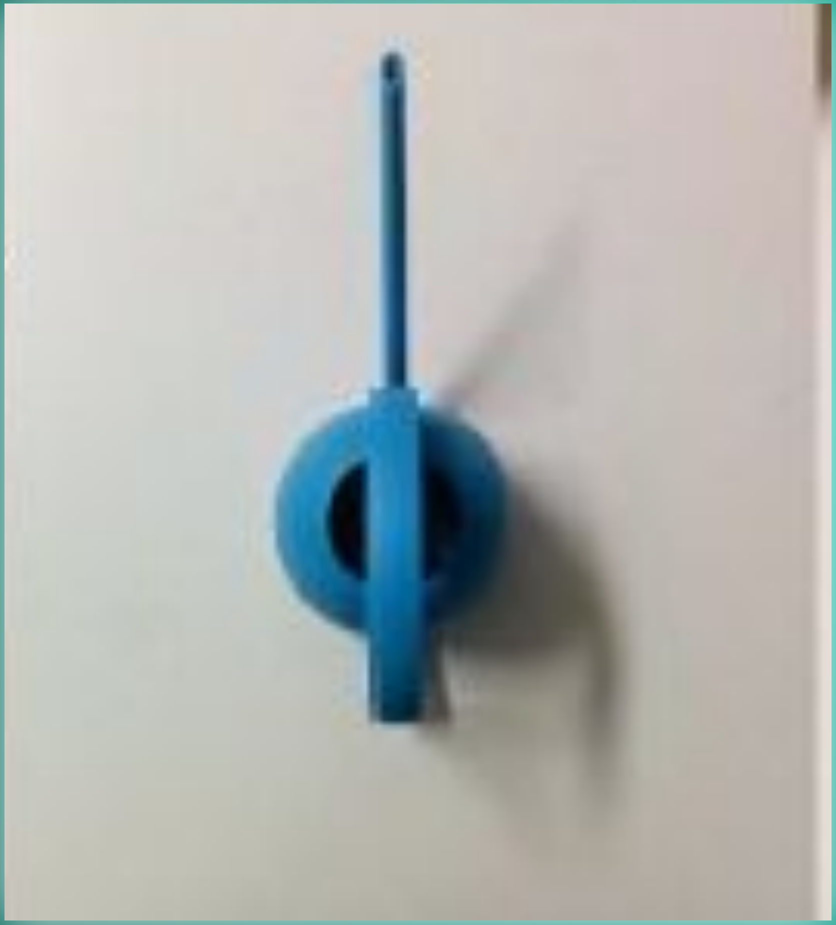
Second Photograph



The second picture is angled to look directly down the spout of the watering
can

This picture provides information about the front of the can that was not visible in the first photograph. But now the handle is hidden!

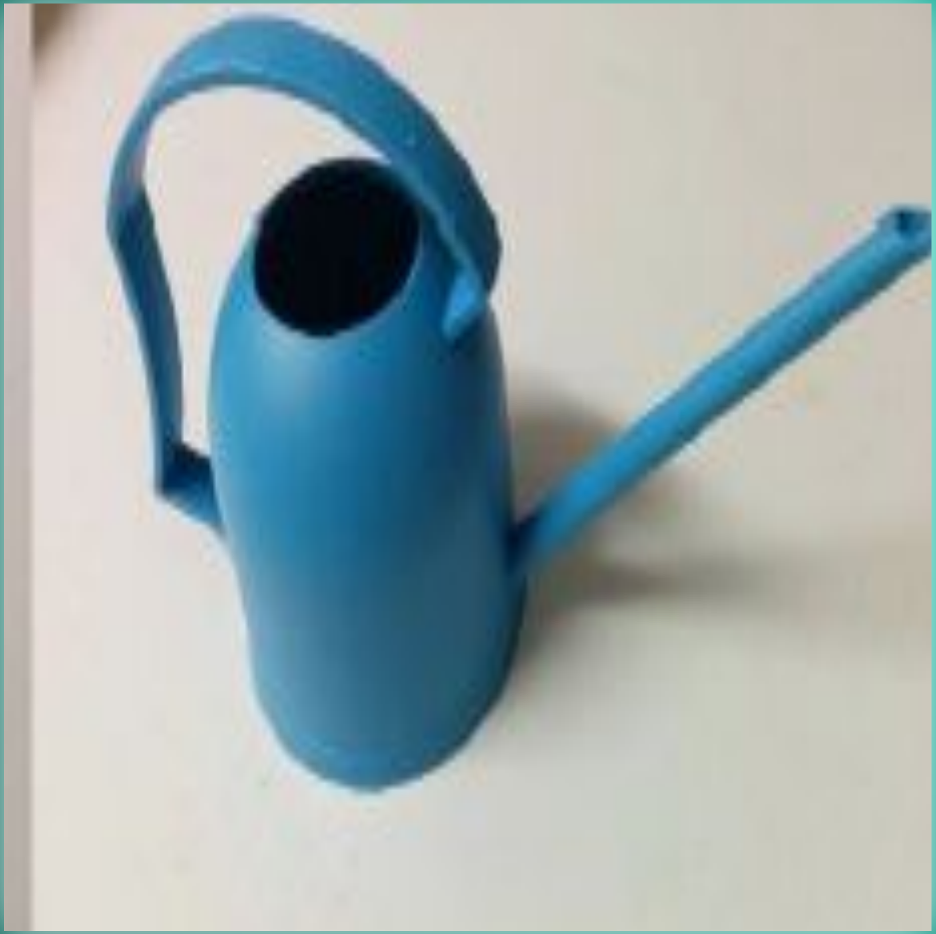
Third Photograph



Provides a Top view of the image

But, the *height* of the watering can cannot be discerned from the bird's eye view of the third picture.

Fourth Photograph



The fourth picture is the obvious choice for the catalog.

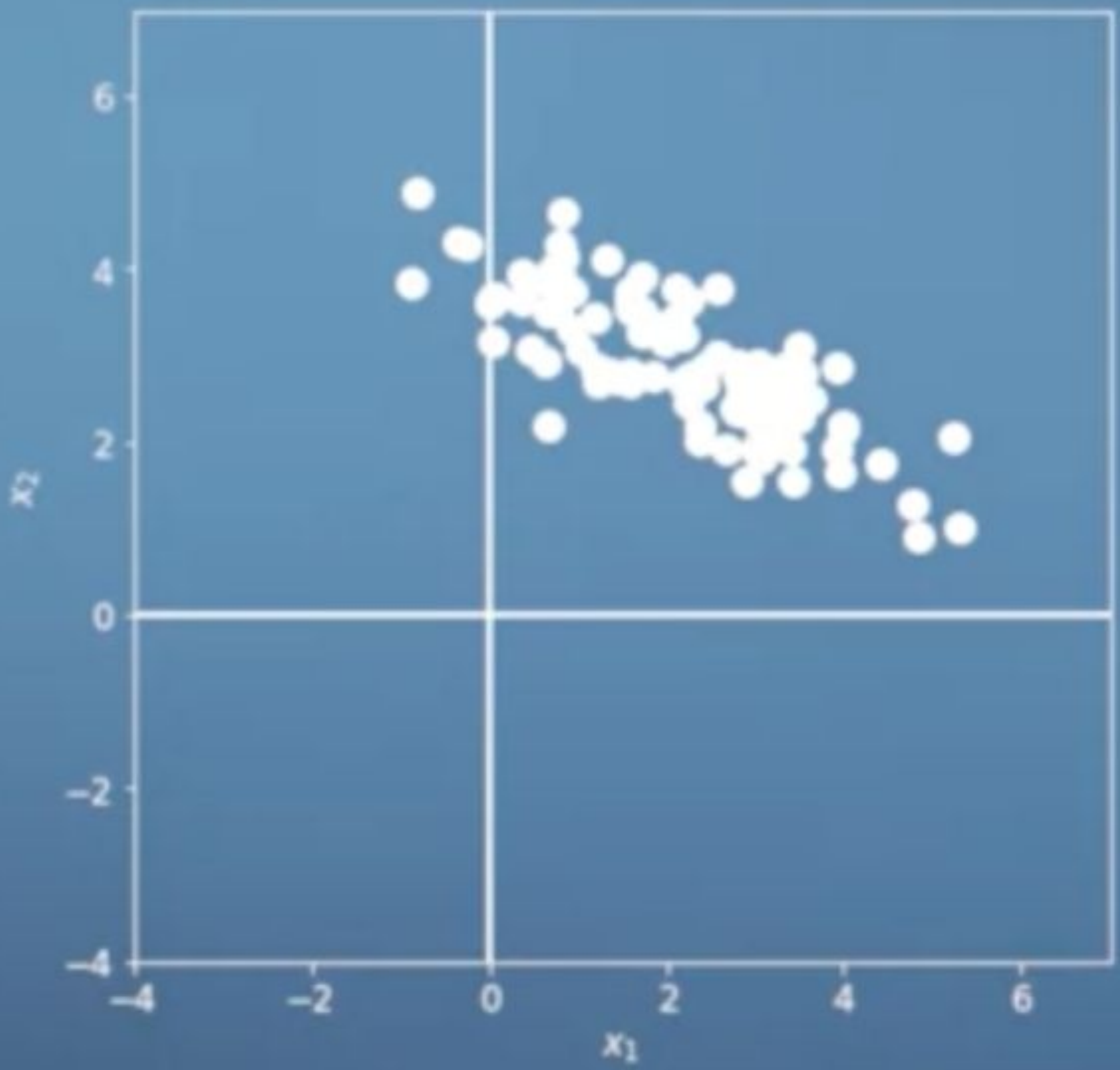
Why?

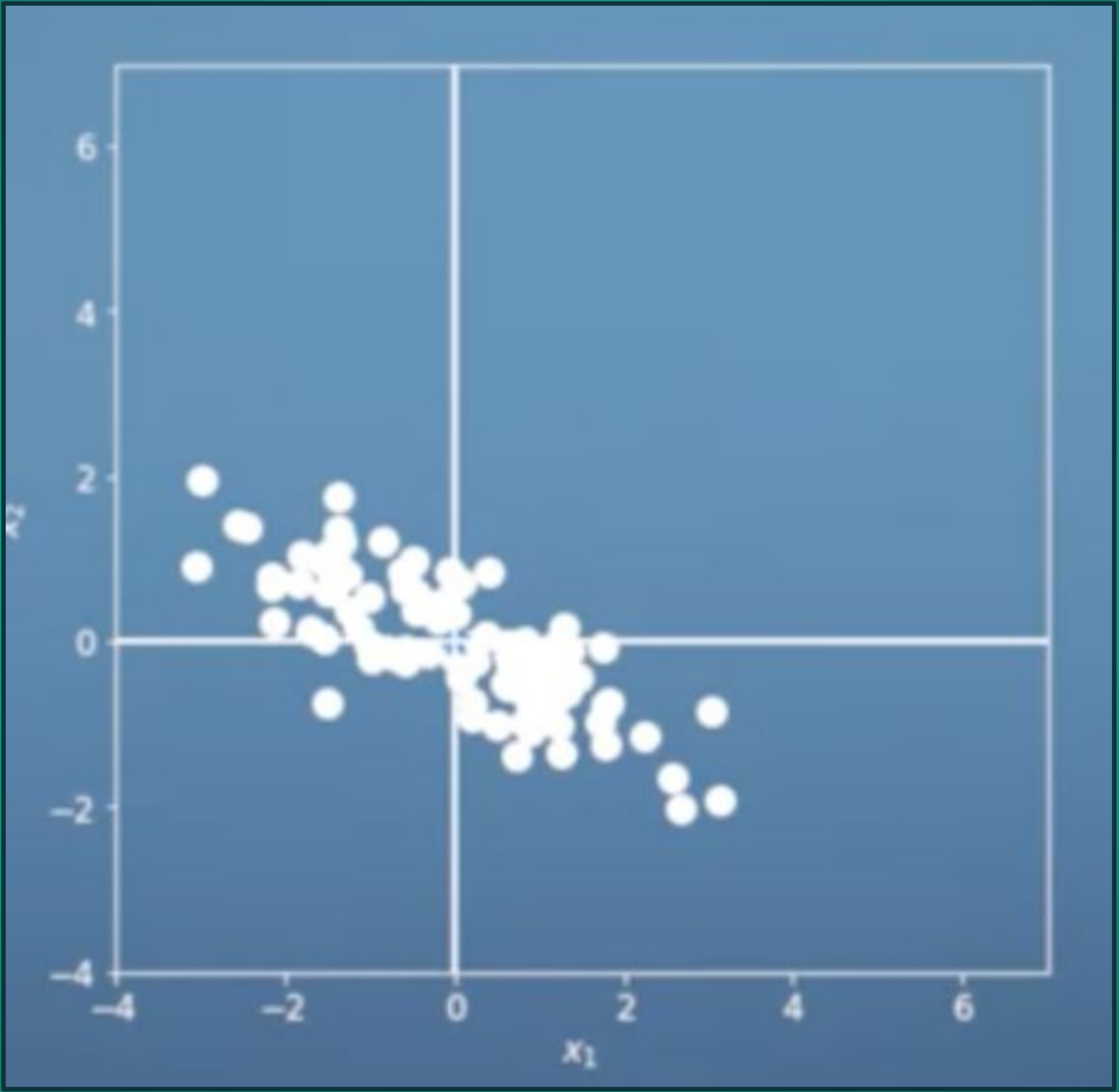
The watering can‘s height, top, spout/nozzle, and handle are all deducible in this image.

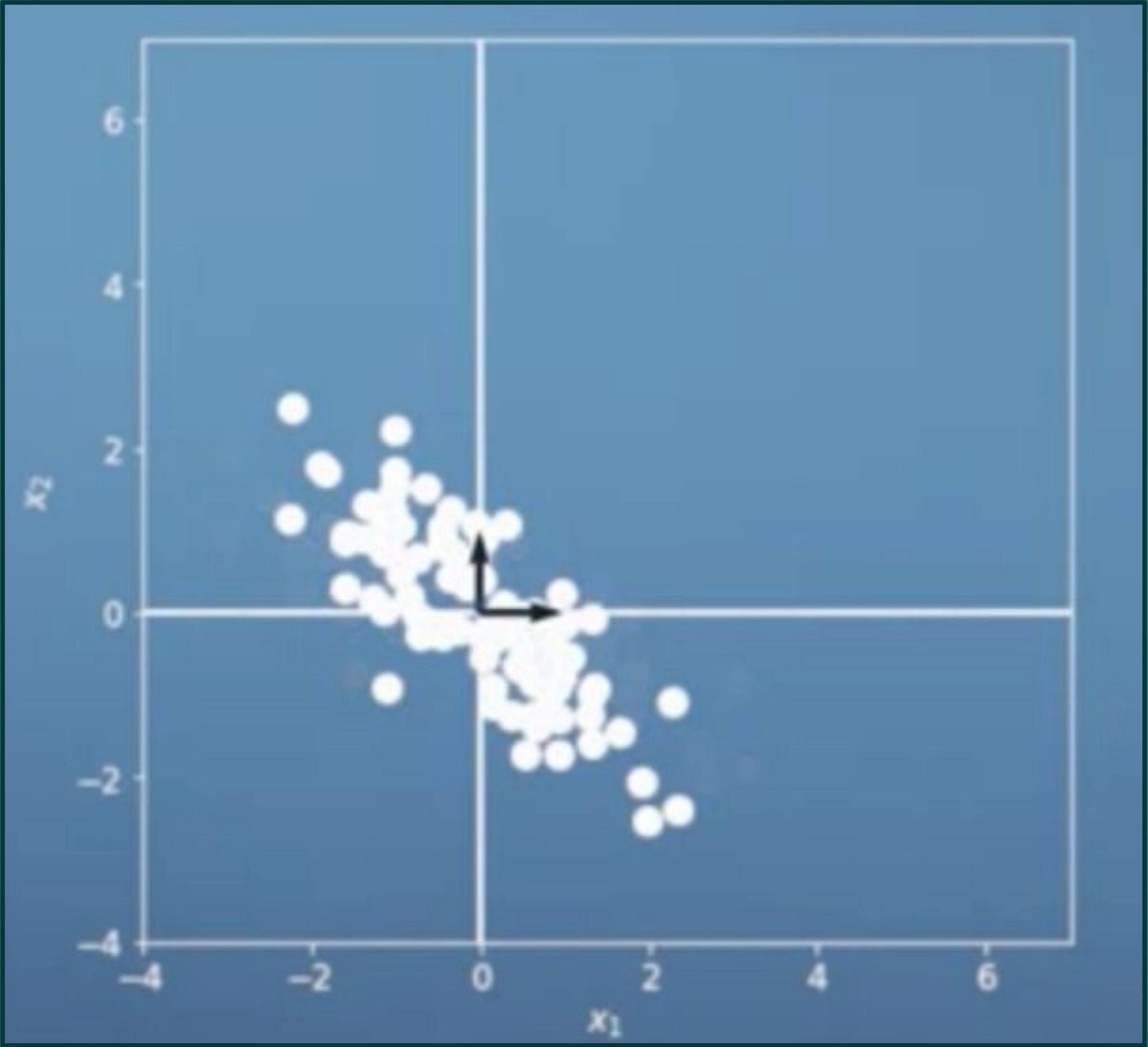
Motivation for PCA

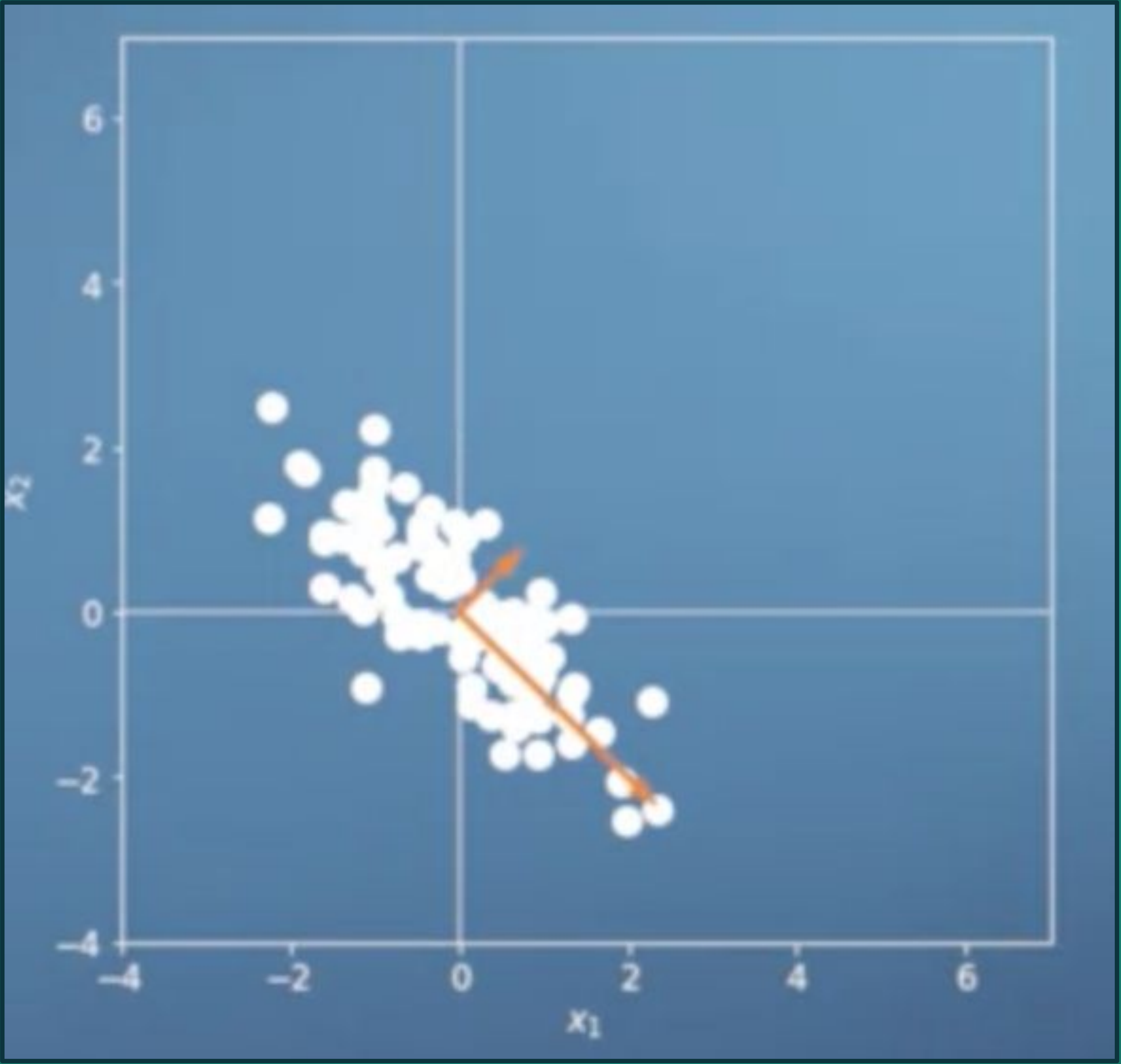
So, It can project data in a high-dimensional Space to a lower-dimensional space that retains as much of the variance as possible.

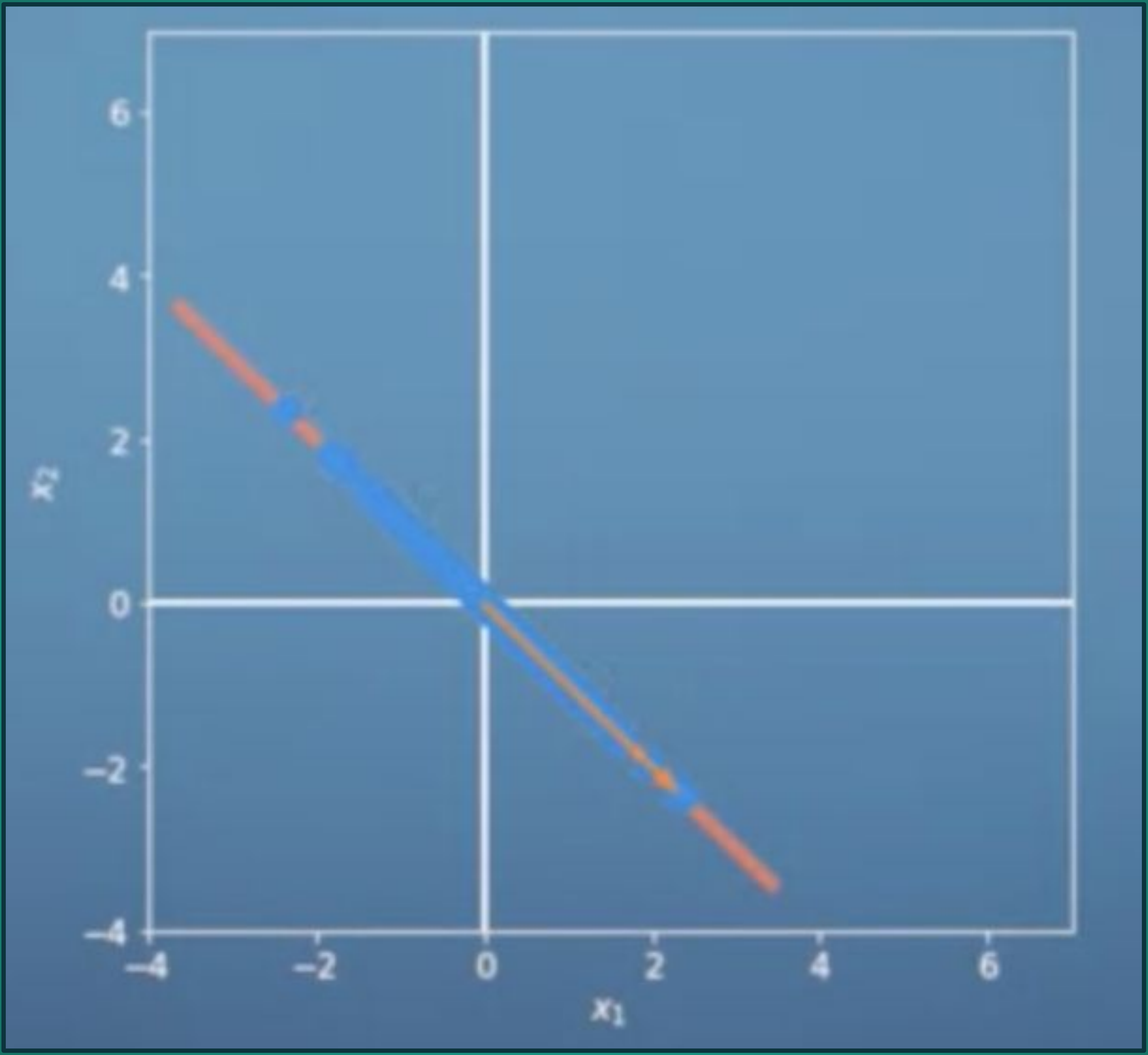
PCA rotates the data set to align with its principal components to maximize the variance
contained within the first several principal components.

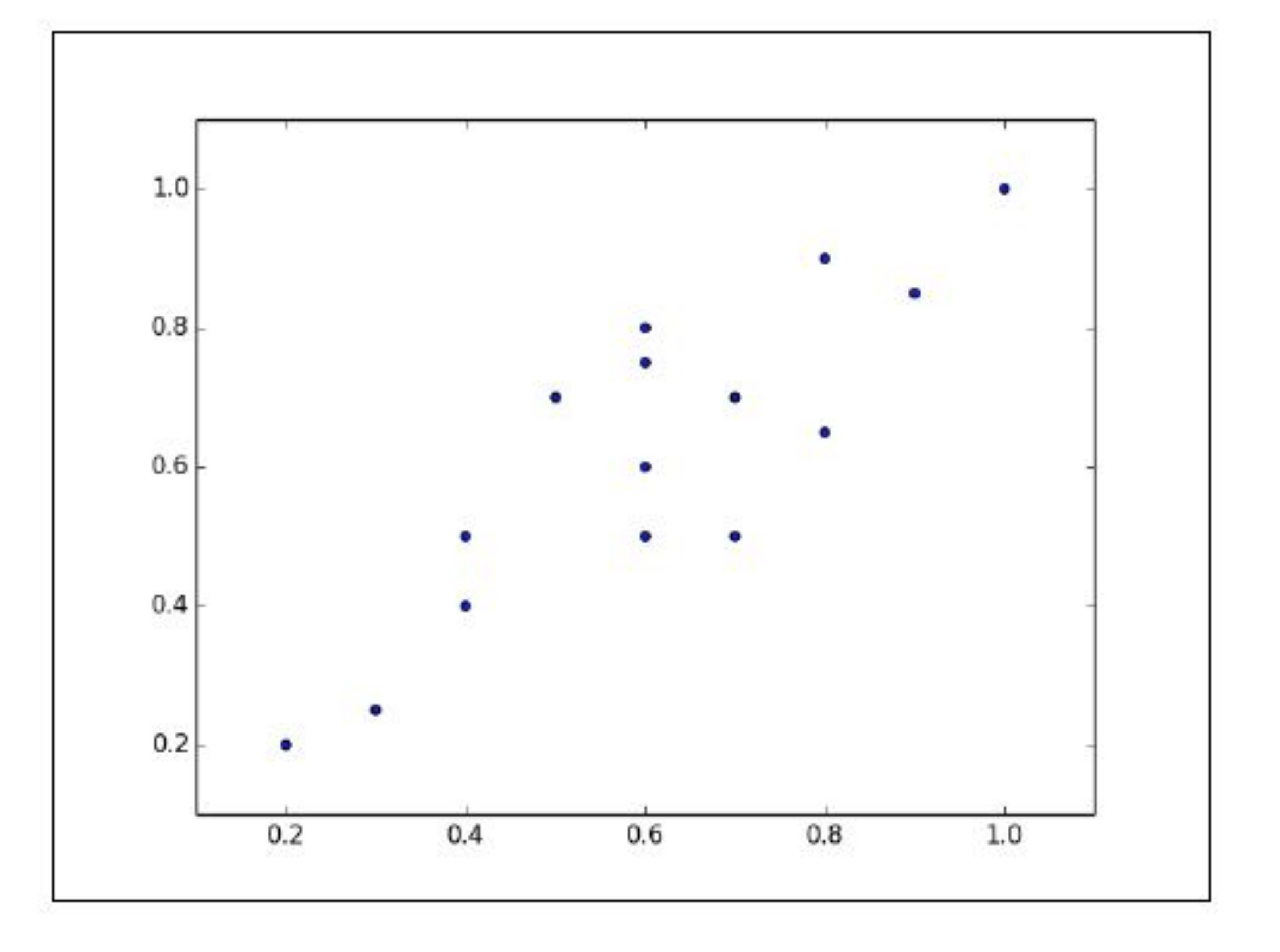

6
4
2
0
−2
−4
x_2
−4
−2
0
2
4
6
x_1


6
4
2
0
−2
−4
x_2
−4
−2
0
2
4
6
x_1

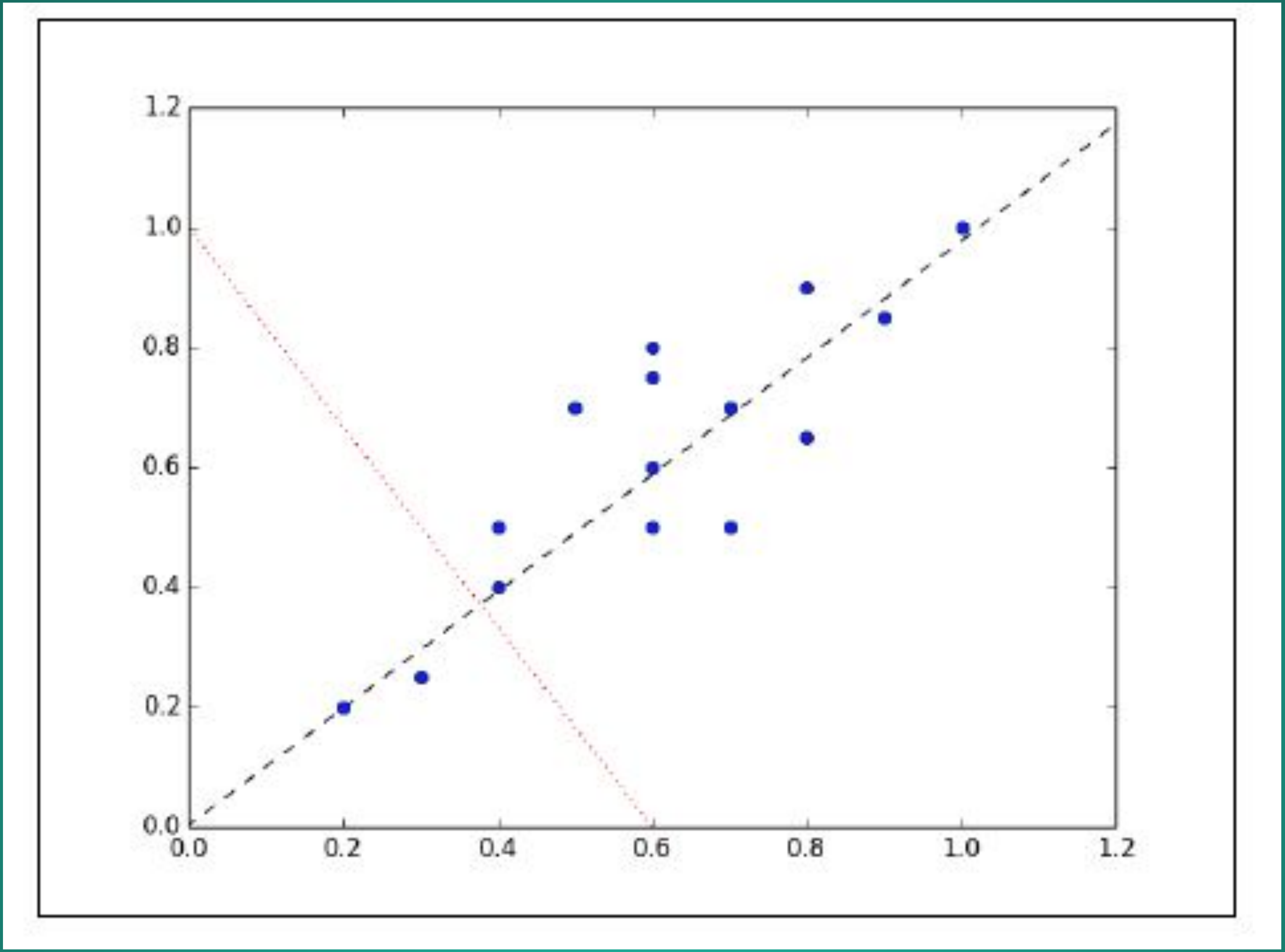

6
4
2
0
−2
−4
x_2
−4
−2
0
2
4
6
x_1


6
4
2
0
−2
−4
x_2
−4
−2
0
2
4
6
x_1

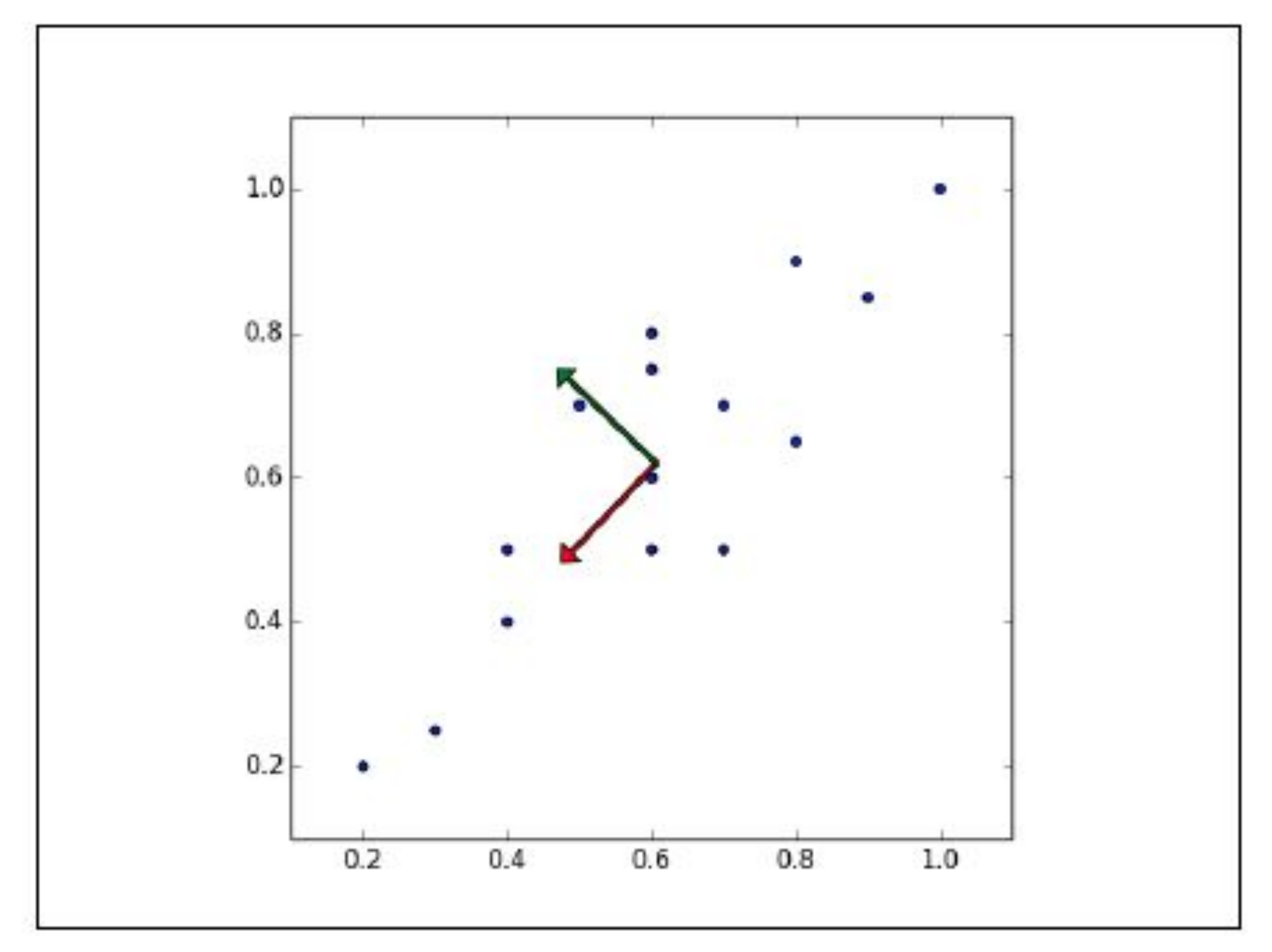

6
4
2
0
−2
−4
x_2
−4
−2
0
2
4
6
x_1



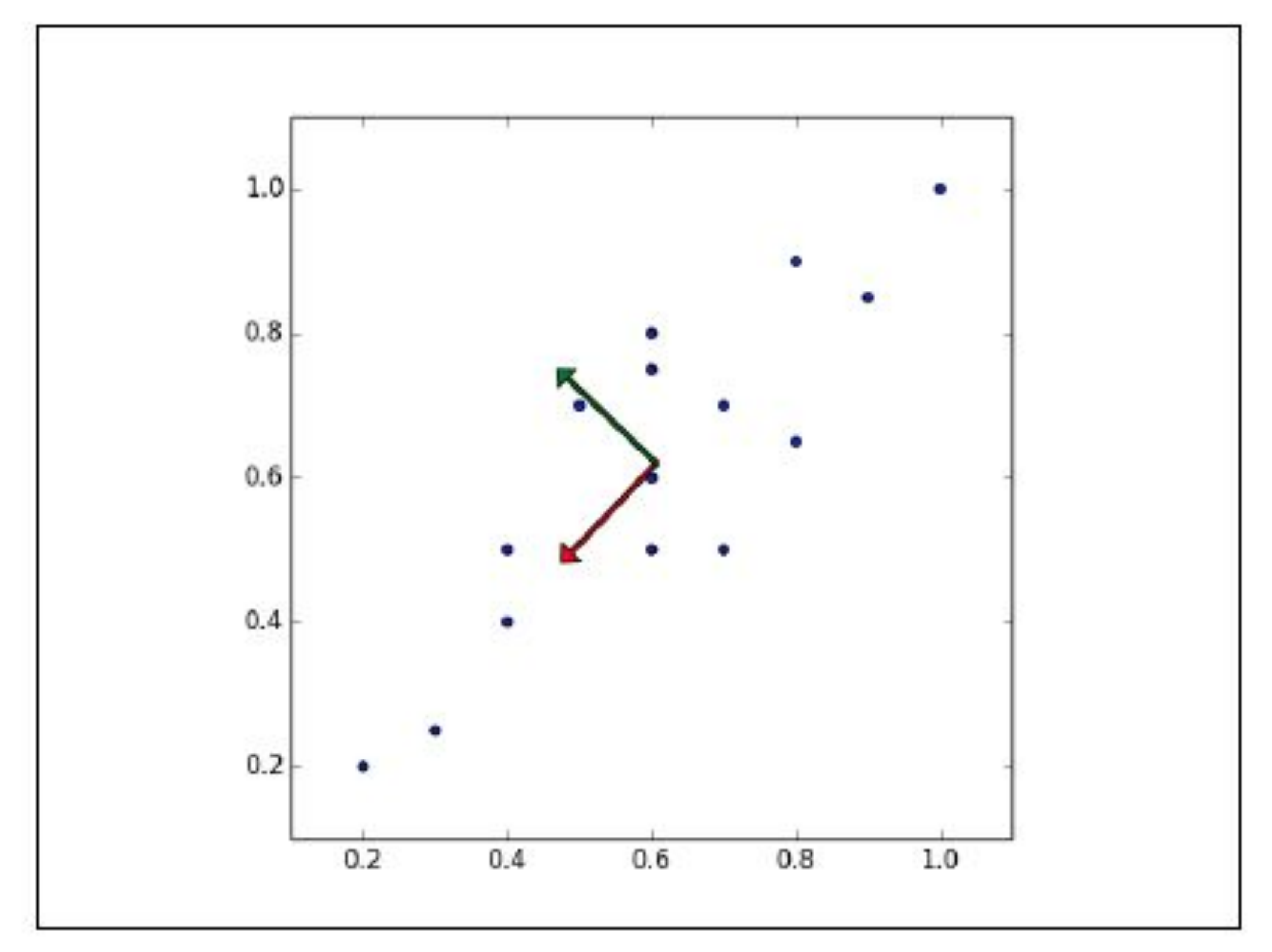
The instances approximately form a long, thin ellipse stretching from the origin to the top right of the plot.



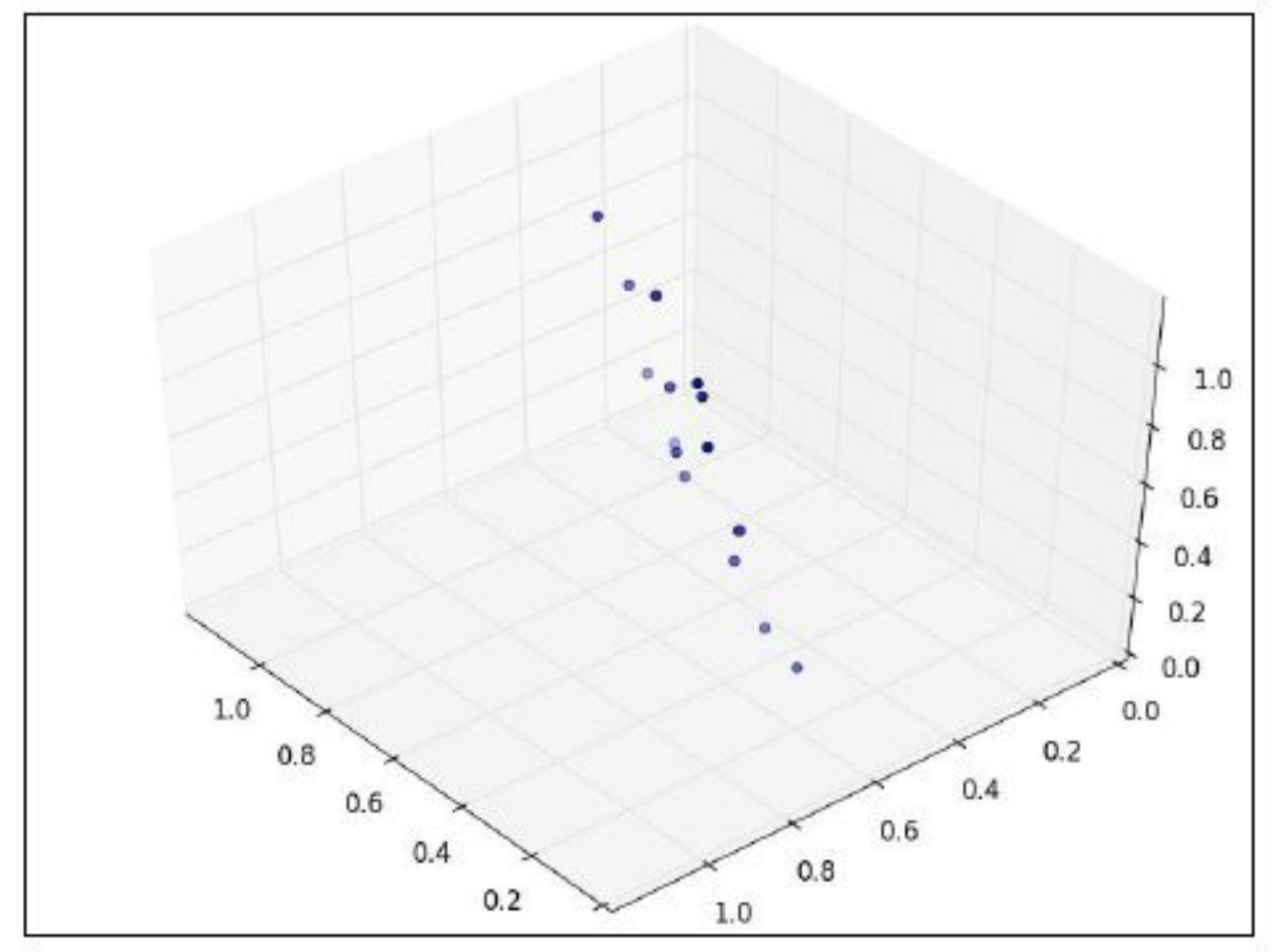
To reduce the dimensions of this data set, we must project the points onto a line. The following are two lines that the data could be projected onto. Along which line do the instances vary the most?



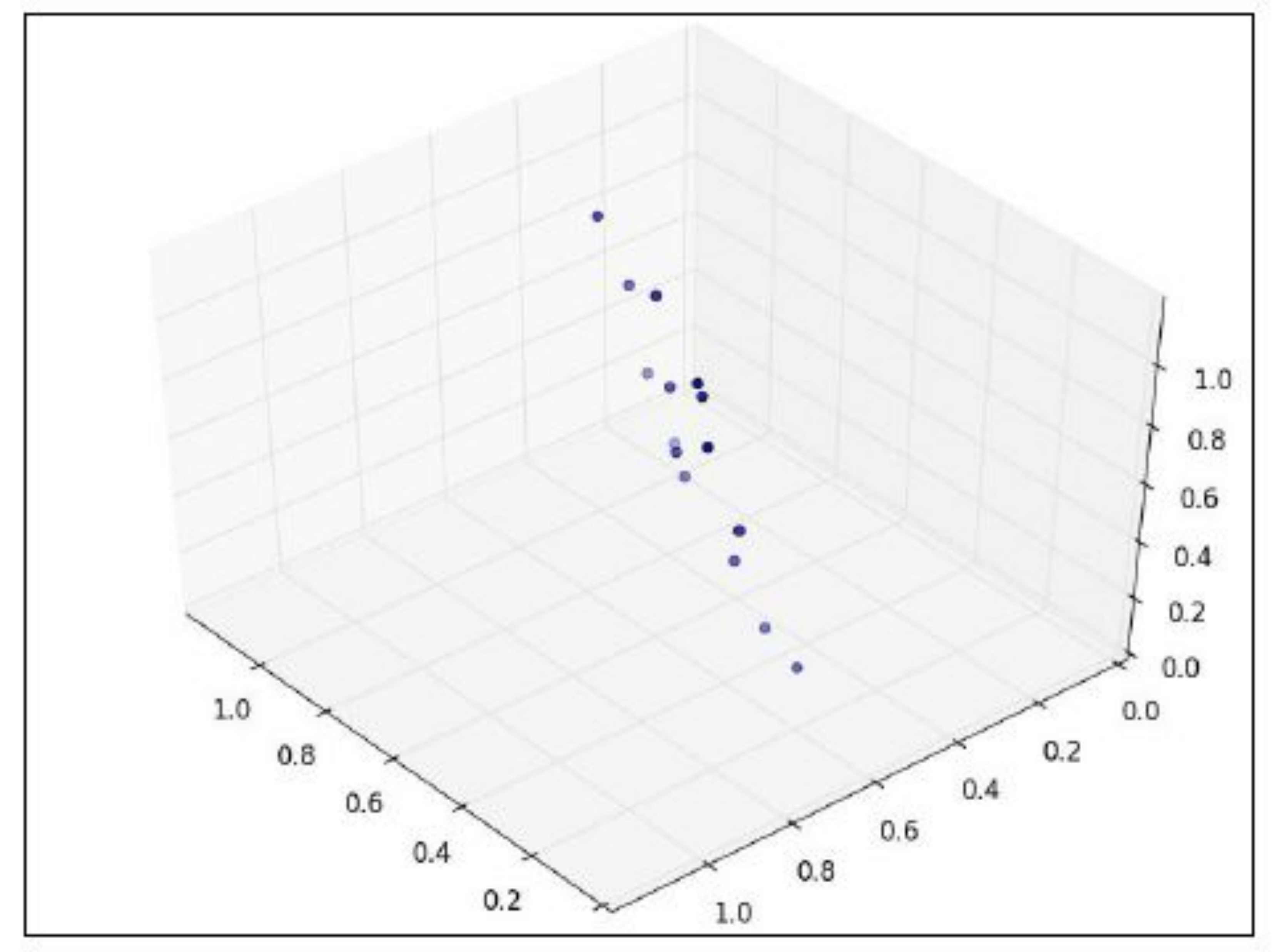
The instances vary more along the dashed line than the dotted line. In fact, the dashed line is the first principal component. The second principal component must be orthogonal (perpendicular)/ statistically independent to the first principal component.



Each subsequent principal component preserves the maximum amount of the remaining variance; the only constraint is that each must be orthogonal to the other principal components.



Now assume that the data set is three dimensional. The scatter plot of the points looks like a flat disc that has been rotated slightly about one of the axes.



The points can be rotated and translated such that the tilted disk lies almost exactly in two dimensions. The points now form an ellipse; the third dimension contains almost no variance and can be discarded.

CONSTRAINTS

PCA is most useful when the variance in a data set is distributed unevenly across the dimensions.

PCA cannot be used effectively with any data set which has equal variance in each dimension.

None of the dimensions can be discarded without losing a significant amount of information.

It is easy to visually identify the principal components of data sets with only two or three dimensions.

In the next section, we will discuss how to calculate the principal components of high-dimensional data.

VARIANCE, COVARIANCE, COVARIANCE MATRICES

$$s^2 = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}{n-1}$$

$$\operatorname{cov}(X,Y) = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)\left(Y_i - \bar{y}\right)}{n-1}$$

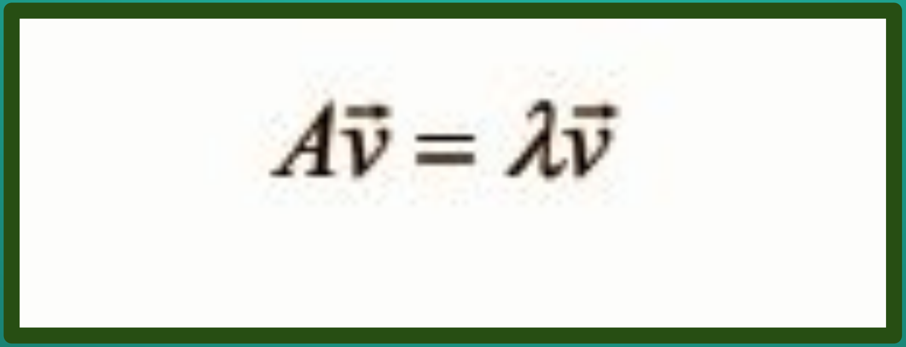
$$C = \begin{bmatrix} \operatorname{cov}(x_1,x_1) & \operatorname{cov}(x_1,x_2) & \operatorname{cov}(x_1,x_3) \\ \operatorname{cov}(x_2,x_1) & \operatorname{cov}(x_2,x_2) & \operatorname{cov}(x_2,x_3) \\ \operatorname{cov}(x_3,x_1) & \operatorname{cov}(x_3,x_2) & \operatorname{cov}(x_3,x_3) \end{bmatrix}$$

EIGEN VALUES AND EIGEN VECTORS

The principal components of a matrix are the eigenvectors of its covariance matrix, ordered by their corresponding eigenvalues.

The eigenvector with the greatest eigenvalue is the first principal component

The second principal component is the eigenvector with the second greatest eigenvalue, and so on.



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

EIGEN VALUES AND EIGEN VECTORS

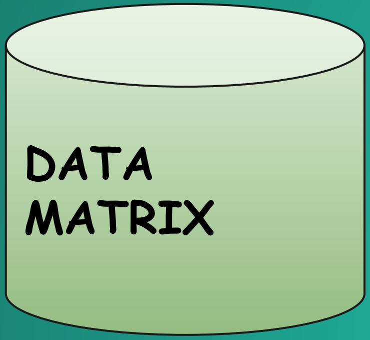
In the preceding equation, v is an eigenvector, A is a square matrix, and lambda is a scalar called an eigenvalue.

The direction of an eigenvector remains the same after it has been transformed by A, only its magnitude has changed, as indicated by the eigenvalue.

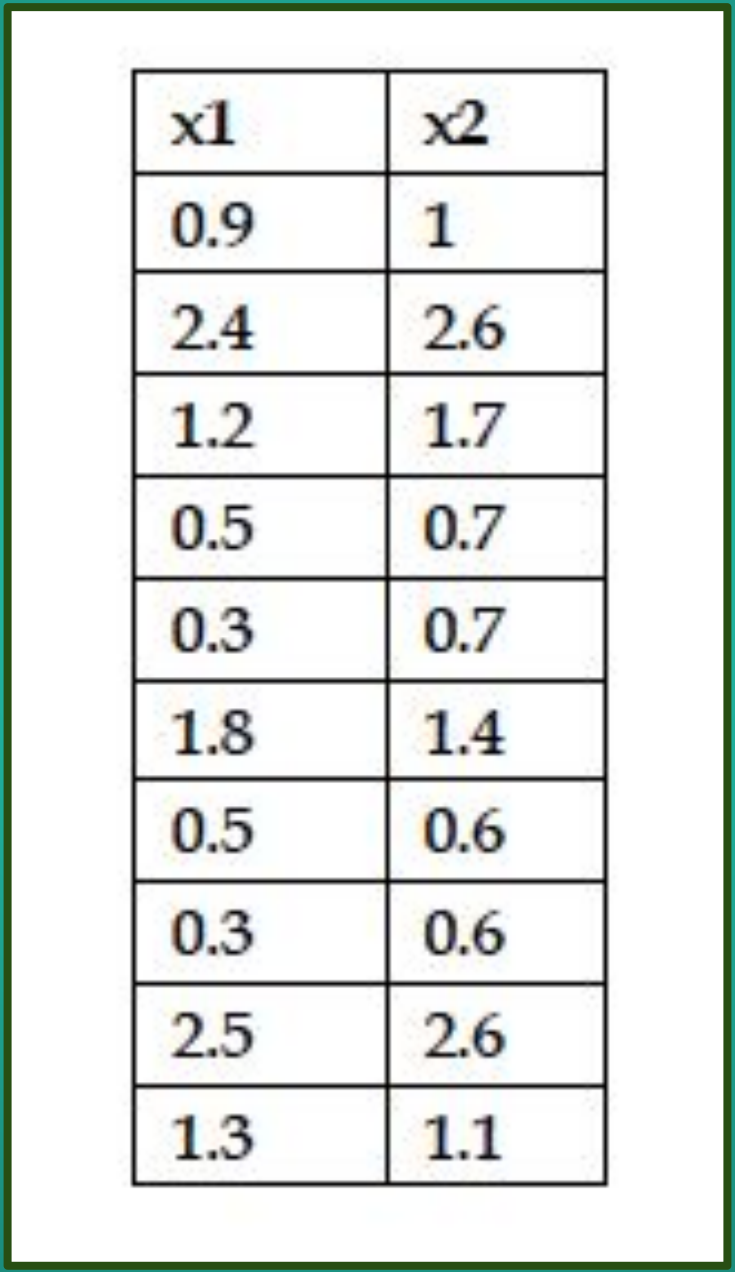
Thus, multiplying a matrix by one of its eigenvectors is equal to scaling the eigenvector.

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

Finding Principal Components

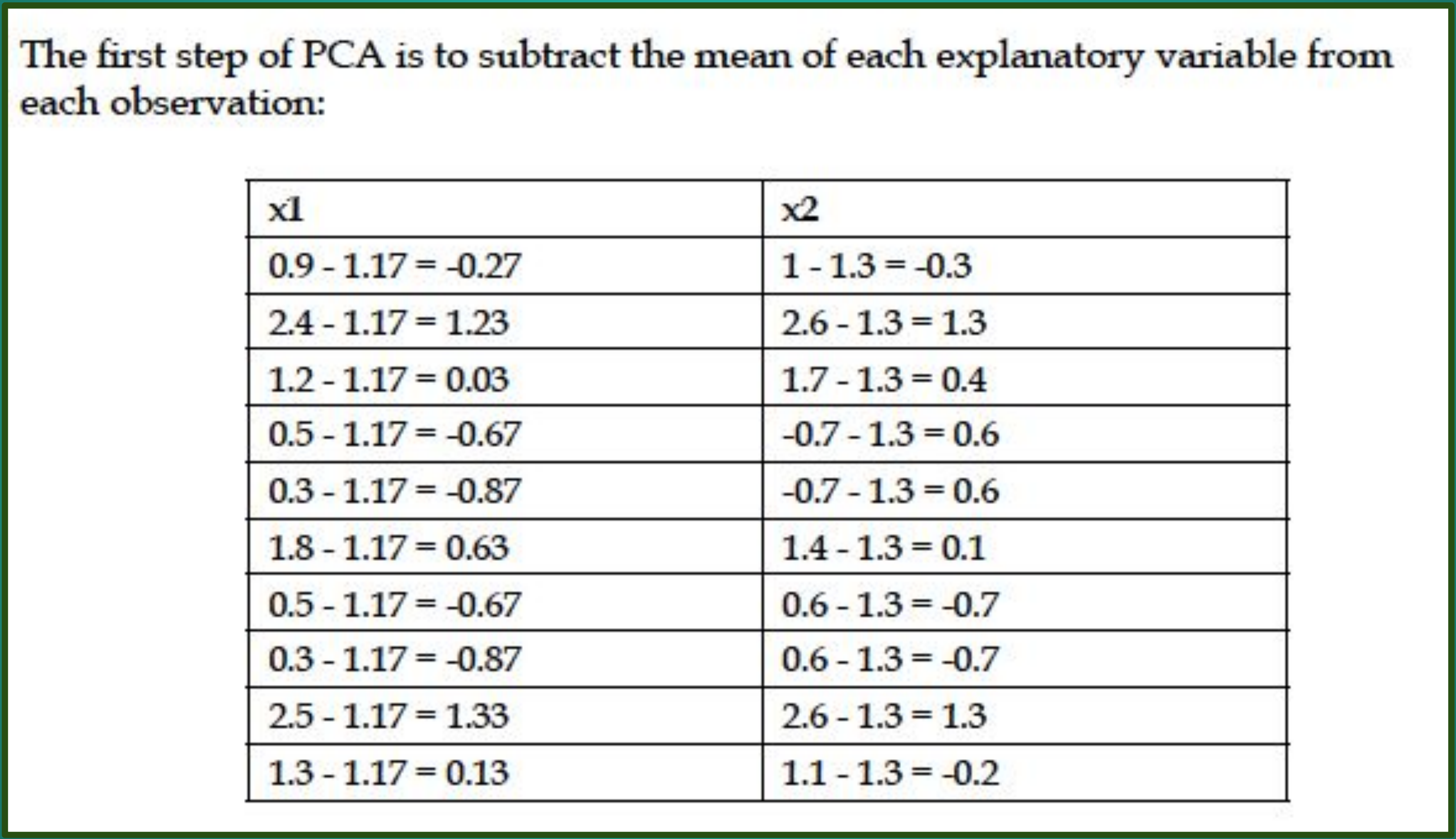


DATA
MATRIX



x1	x2
0.9	1
2.4	2.6
1.2	1.7
0.5	0.7
0.3	0.7
1.8	1.4
0.5	0.6
0.3	0.6
2.5	2.6
1.3	1.1

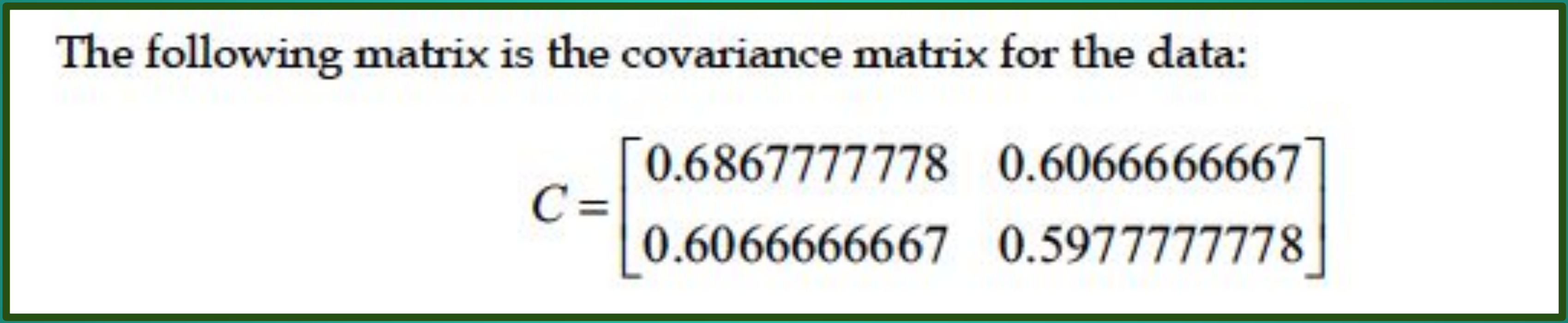
Finding Principal Components



The first step of PCA is to subtract the mean of each explanatory variable from each observation:

x1	x2
0.9 - 1.17 = -0.27	1 - 1.3 = -0.3
2.4 - 1.17 = 1.23	2.6 - 1.3 = 1.3
1.2 - 1.17 = 0.03	1.7 - 1.3 = 0.4
0.5 - 1.17 = -0.67	-0.7 - 1.3 = 0.6
0.3 - 1.17 = -0.87	-0.7 - 1.3 = 0.6
1.8 - 1.17 = 0.63	1.4 - 1.3 = 0.1
0.5 - 1.17 = -0.67	0.6 - 1.3 = -0.7
0.3 - 1.17 = -0.87	0.6 - 1.3 = -0.7
2.5 - 1.17 = 1.33	2.6 - 1.3 = 1.3
1.3 - 1.17 = 0.13	1.1 - 1.3 = -0.2

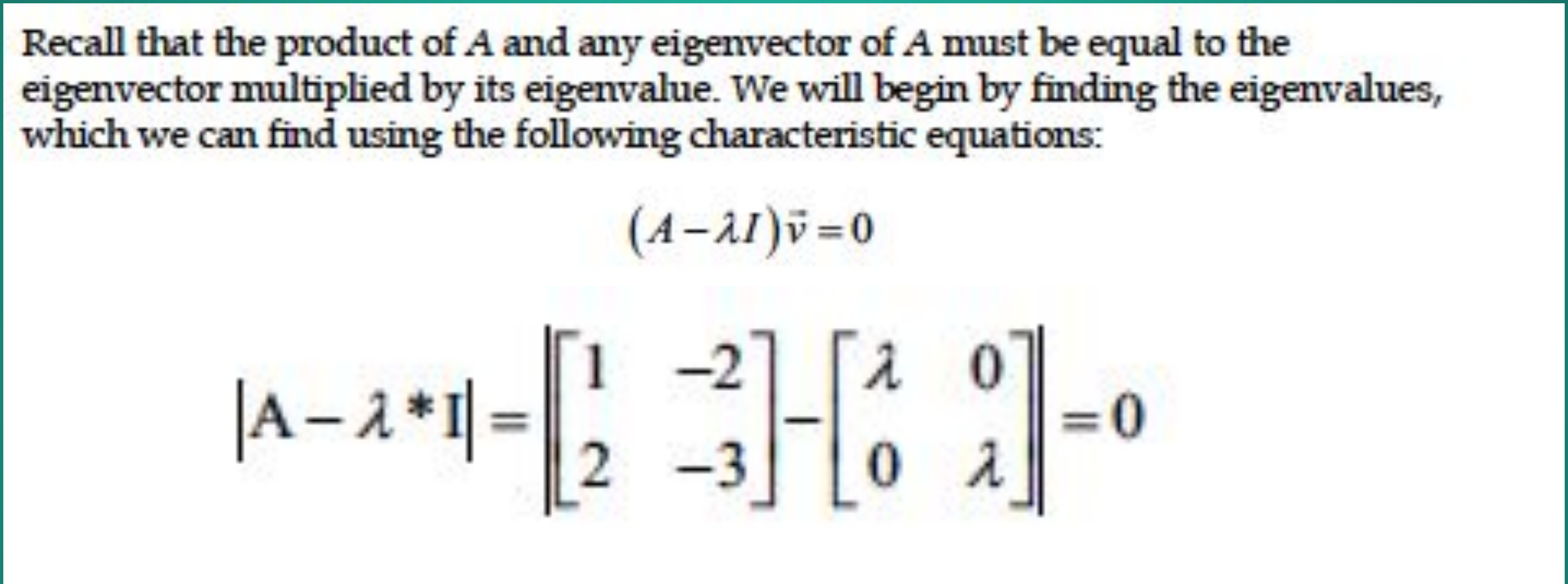
Finding Principal Components



The following matrix is the covariance matrix for the data:

$$C = \begin{bmatrix} 0.6867777778 & 0.6066666667 \\ 0.6066666667 & 0.5977777778 \end{bmatrix}$$

Finding Principal Components

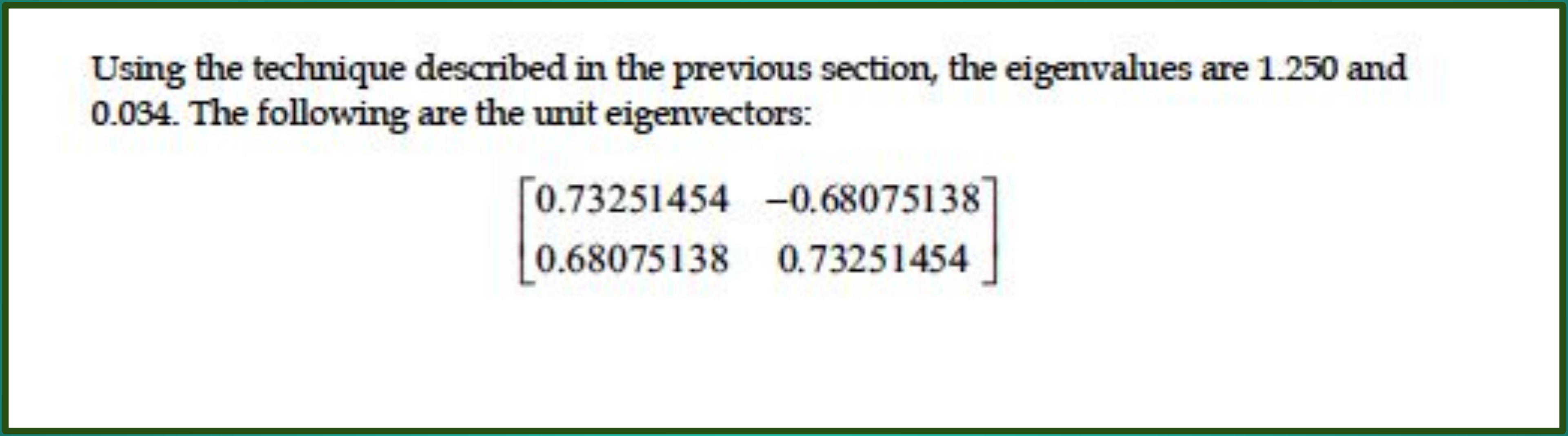


Recall that the product of *A* and any eigenvector of *A* must be equal to the eigenvector multiplied by its eigenvalue. We will begin by finding the eigenvalues, which we can find using the following characteristic equations:

$$(A-\lambda I)\vec{v}=0$$

$$|A-\lambda * I| = \left|\begin{bmatrix}1 & -2\\ 2 & -3\end{bmatrix}-\begin{bmatrix}\lambda & 0\\ 0 & \lambda\end{bmatrix}\right| = 0$$

Finding Principal Components



Using the technique described in the previous section, the eigenvalues are 1.250 and 0.034. The following are the unit eigenvectors:

$$\begin{bmatrix} 0.73251454 & -0.68075138 \\ 0.68075138 & 0.73251454 \end{bmatrix}$$

Next Steps.....

Project the data onto the principal components.......

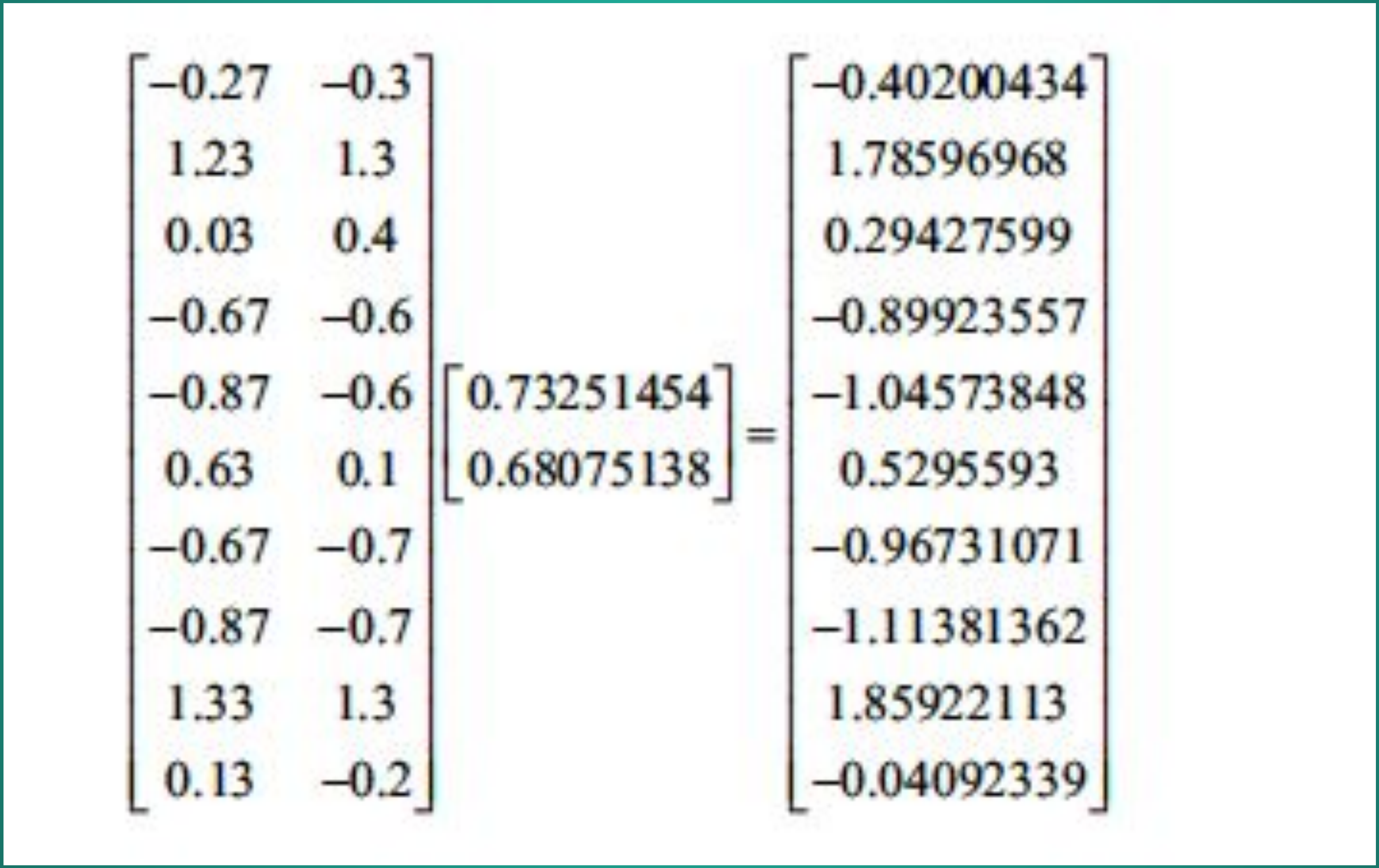
First, we will build a transformation matrix in which each column of the matrix is the eigenvector for a principal component.

The first eigenvector has the greatest eigenvalue and is the first principal component.

If we were reducing a five-dimensional data set to three dimensions, we would build a matrix with three columns. In this example, we are projecting our two-dimensional data set onto one dimension, so we will use only the Eigen vector for the first principal component.

Finally, we will find the dot product of the data matrix and transformation matrix.

Next Steps.....



$$\begin{bmatrix} -0.27 & -0.3 \\ 1.23 & 1.3 \\ 0.03 & 0.4 \\ -0.67 & -0.6 \\ -0.87 & -0.6 \\ 0.63 & 0.1 \\ -0.67 & -0.7 \\ -0.87 & -0.7 \\ 1.33 & 1.3 \\ 0.13 & -0.2 \end{bmatrix} \begin{bmatrix} 0.73251454 \\ 0.68075138 \end{bmatrix} = \begin{bmatrix} -0.40200434 \\ 1.78596968 \\ 0.29427599 \\ -0.89923557 \\ -1.04573848 \\ 0.5295593 \\ -0.96731071 \\ -1.11381362 \\ 1.85922113 \\ -0.04092339 \end{bmatrix}$$

Summary

1. Dimensionality Reduction can be used to mitigate problems caused by the curse of dimensionality.

2. Dimensionality reduction can be used to compress data while minimizing the amount of information that is lost.

3. Understanding the structure of data with hundreds of dimensions can be difficult; data with only two or three dimensions can be visualized easily.

Thanks!