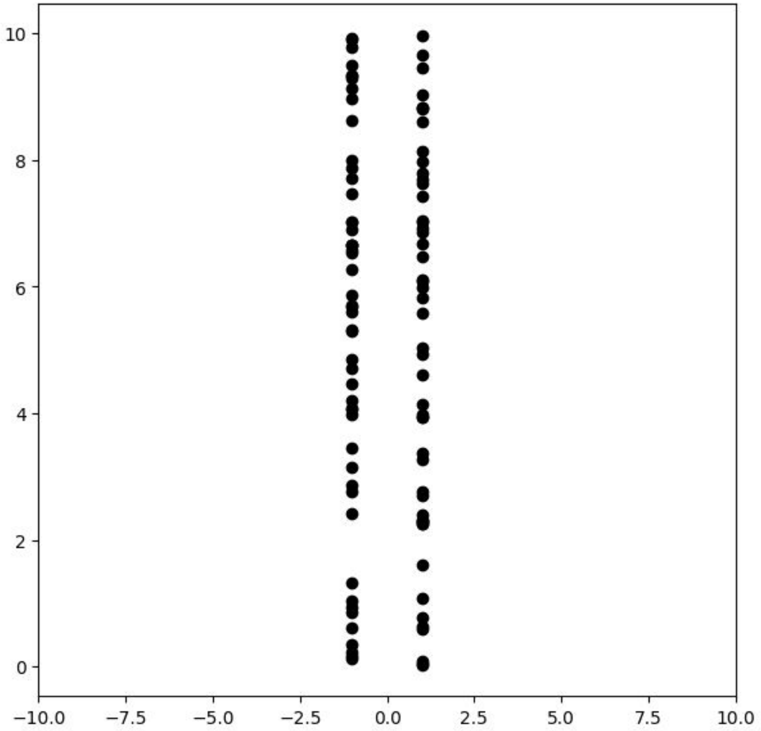
Dimensionality Reduction and Principal Component Analysis

CISC 3225
Spring 2024
PDSH 45, DSFS 3

Principal Component Analysis

Question: If you had to remove a dimension in the figure to the right, which would you remove?

Goal: Preserve *as much information about the distribution of the data as possible.*

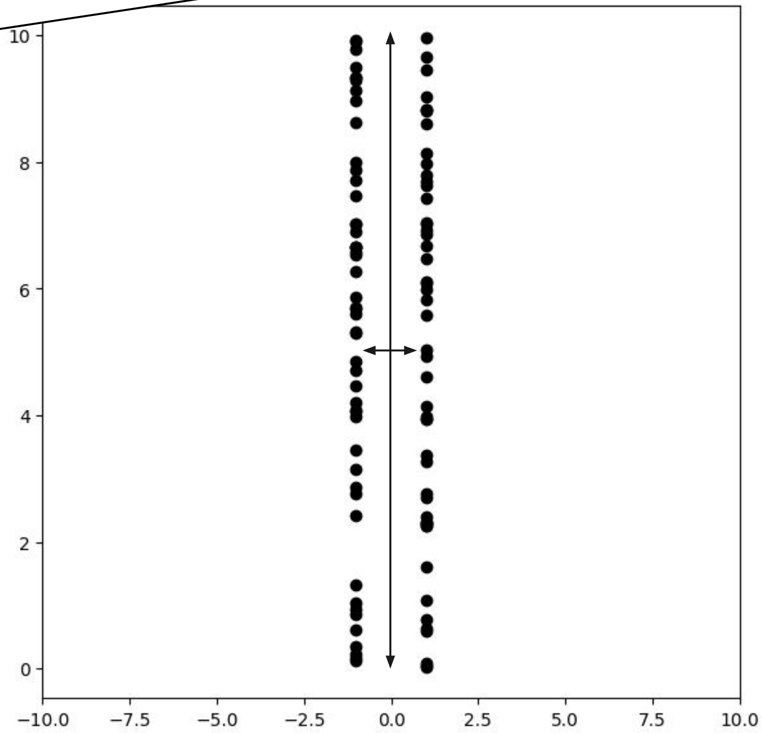


Principal Component Analysis

Question: If you had to remove a dimension in the figure to the right, which would you remove?

Goal: Preserve *as much information about the distribution of the data as possible.*

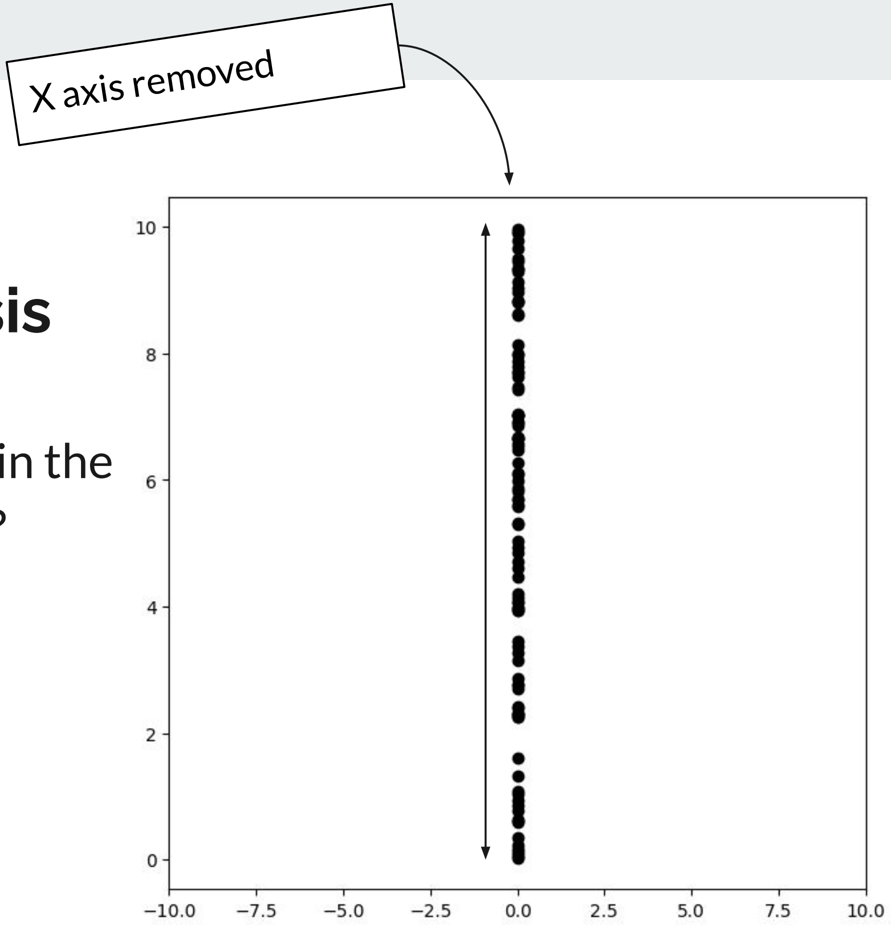
There is significantly more *variance* in the Y axis than the X axis



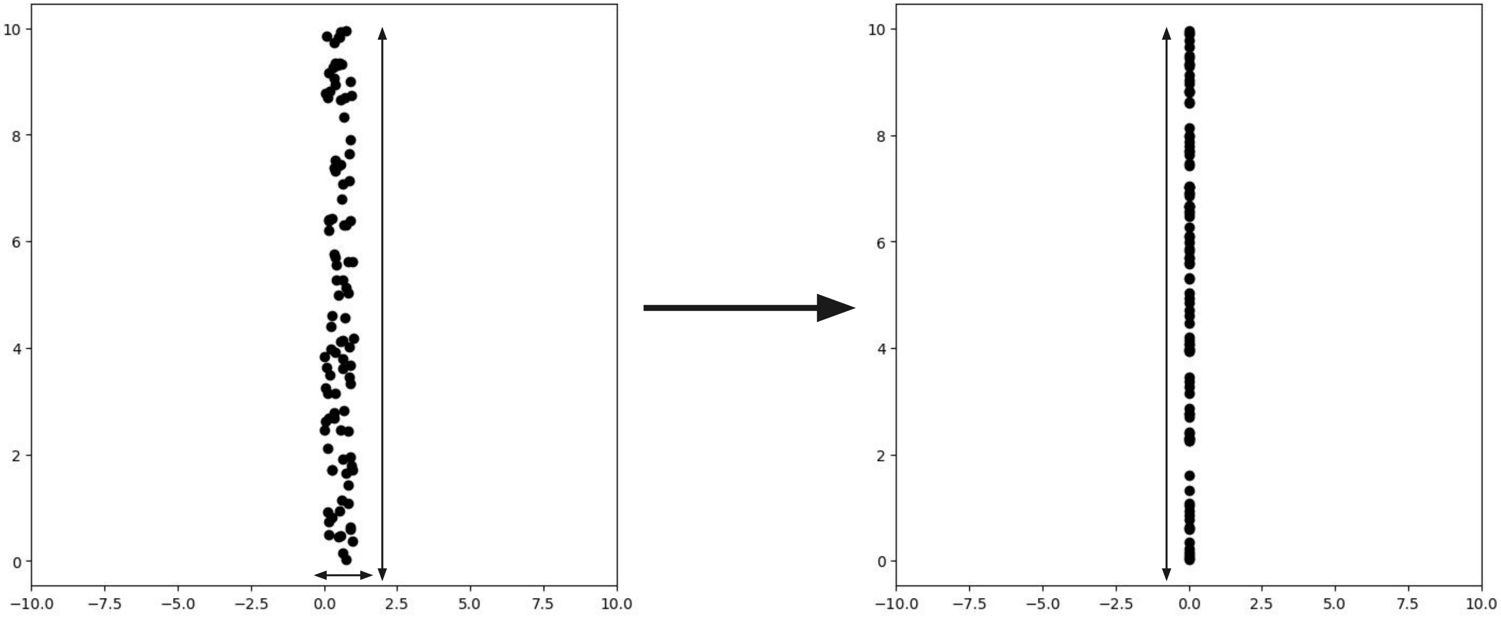
Principal Component Analysis

Question: If you had to remove a dimension in the figure to the right, which would you remove?

Goal: Preserve *as much information about the distribution of the data as possible.*


X axis removed

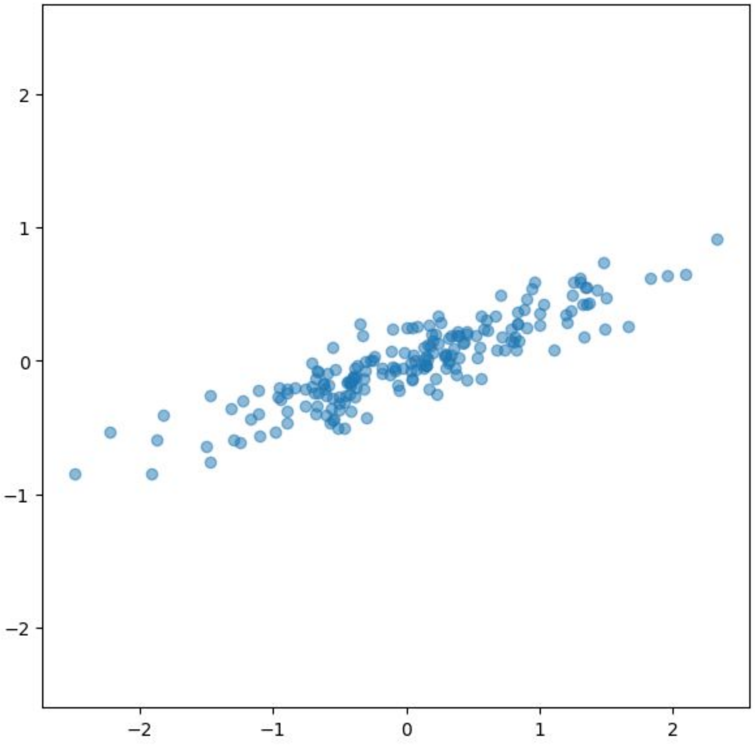
Principal Component Analysis



Principal Component Analysis

Question: If you had to remove a dimension in the figure to the right, which would you remove?

Goal: Preserve *as much information about the distribution of the data as possible.*

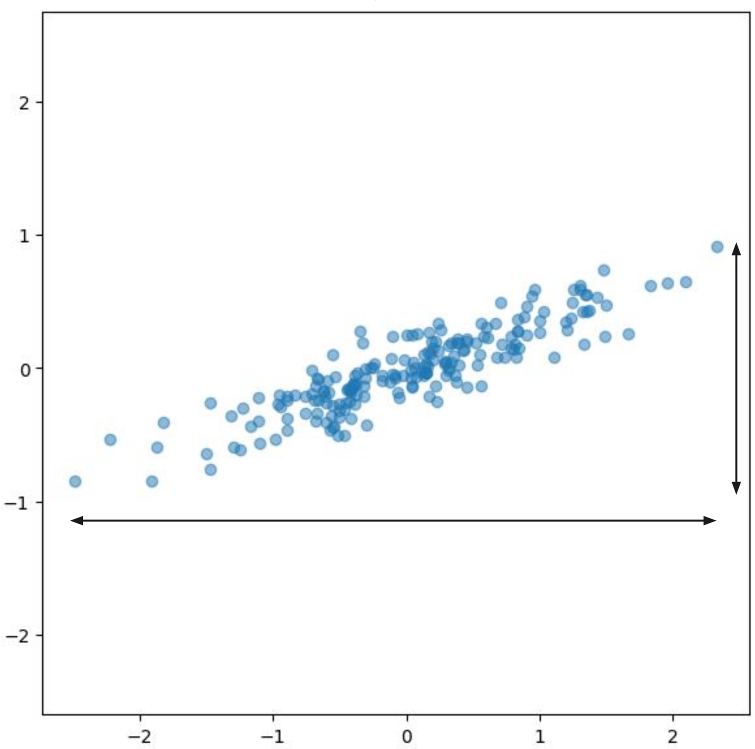


Principal Component Analysis

Question: If you had to remove a dimension in the figure to the right, which would you remove?

Goal: Preserve *as much information about the distribution of the data as possible.*

There is significantly more *variance* in the X axis than the Y axis

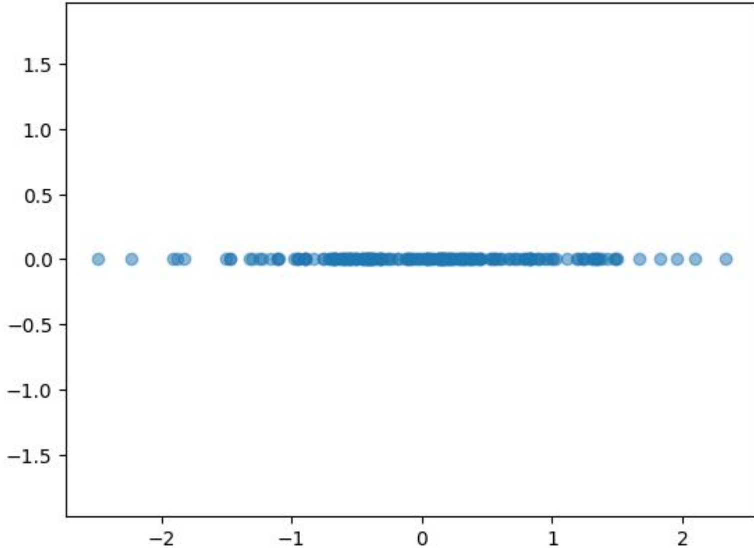


Principal Component Analysis

Question: If you had to remove a dimension in the figure to the right, which would you remove?

Goal: Preserve *as much information about the distribution of the data as possible.*

There is significantly more *variance* in the X axis than the Y axis



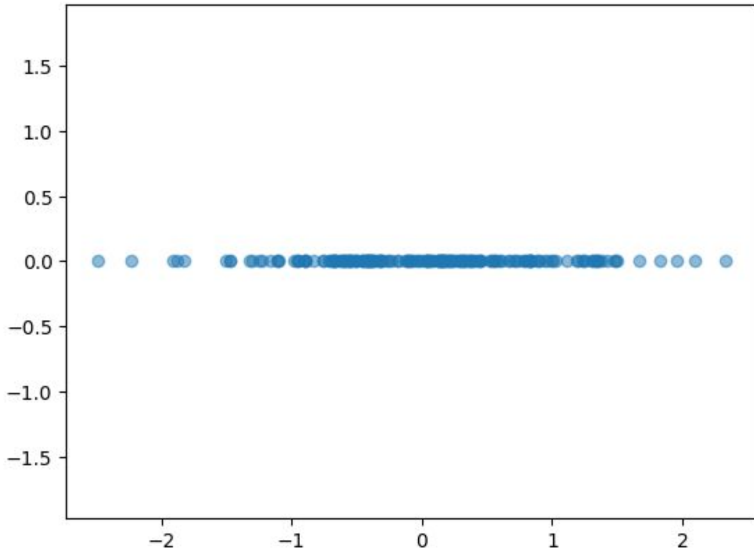
Principal Component Analysis

Question: If you had to remove a dimension in the figure to the right, which would you remove?

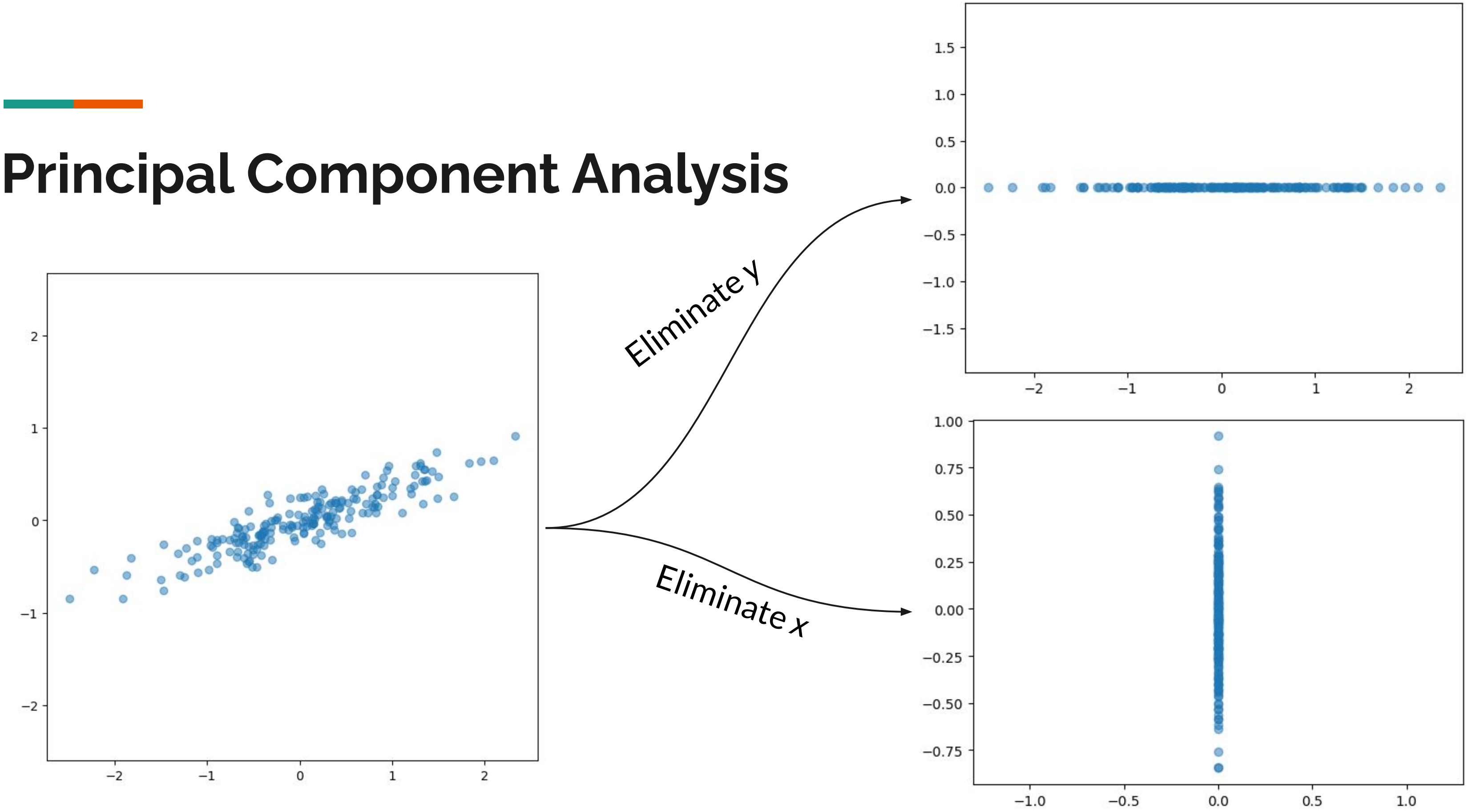
Goal: Preserve *as much information about the distribution of the data as possible.*

Did we really preserve as much information as we can?

There is significantly more *variance* in the X axis than the Y axis



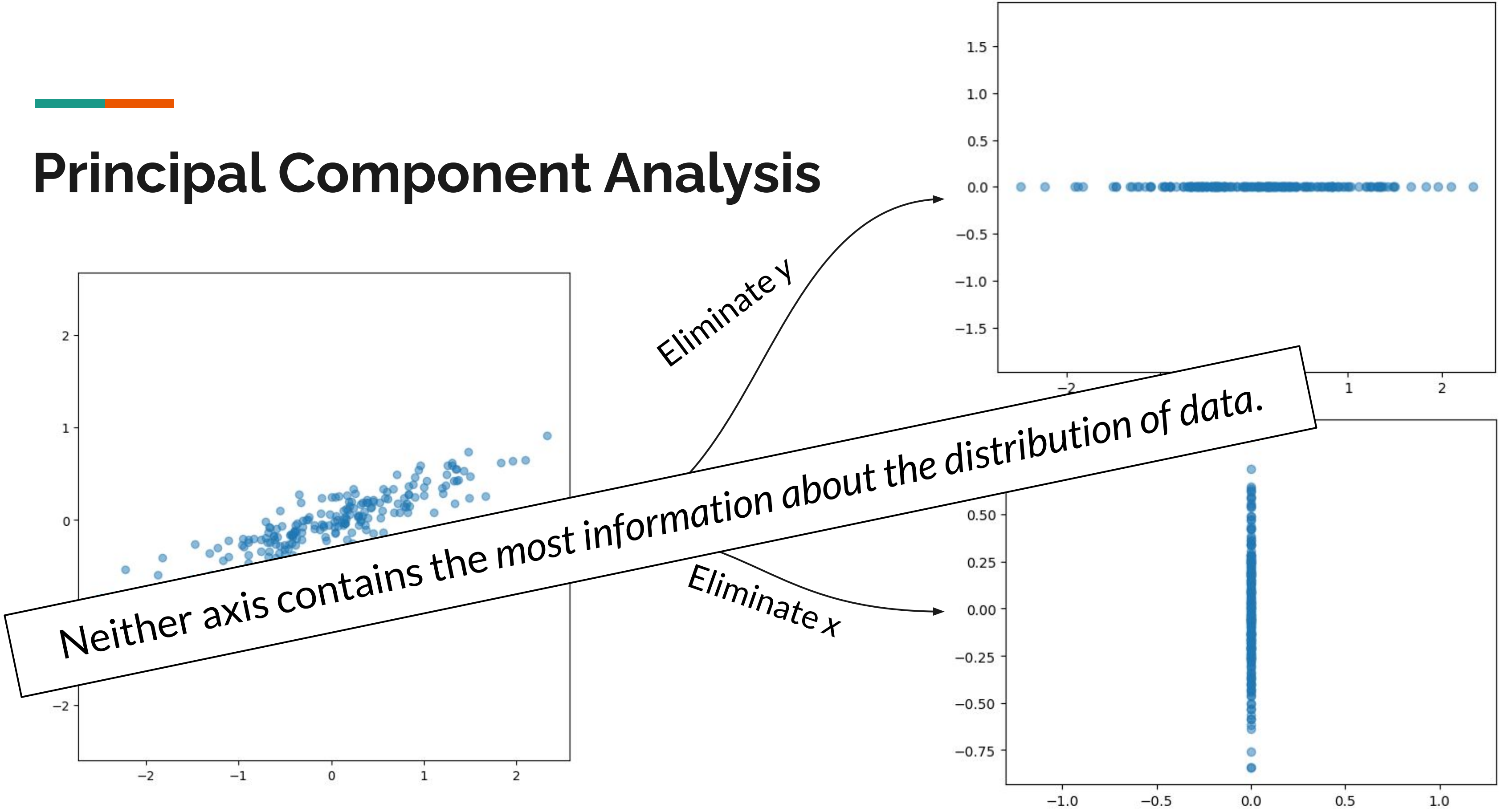
Principal Component Analysis



Eliminate y

Eliminate x

Principal Component Analysis



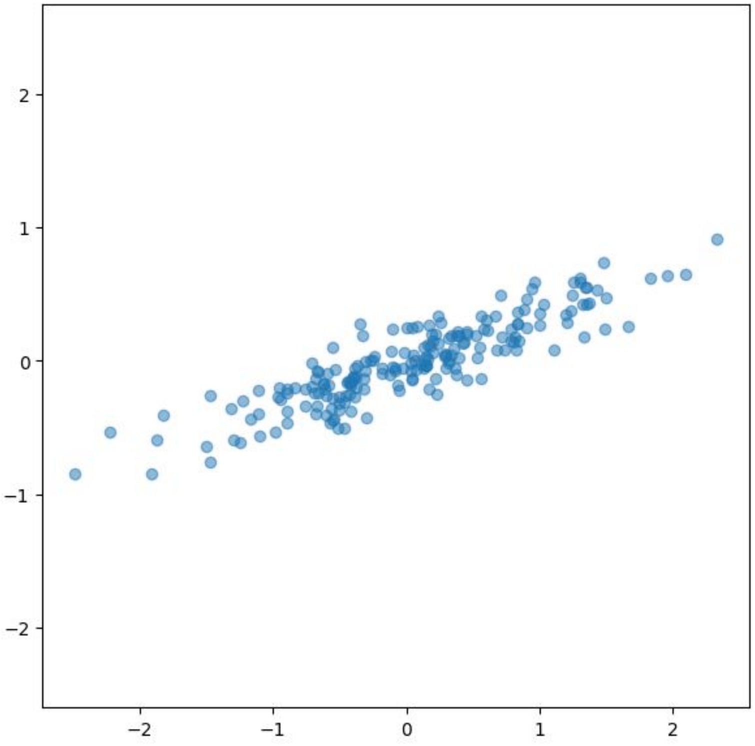
Eliminate y

Eliminate x

Neither axis contains the most information about the distribution of data.

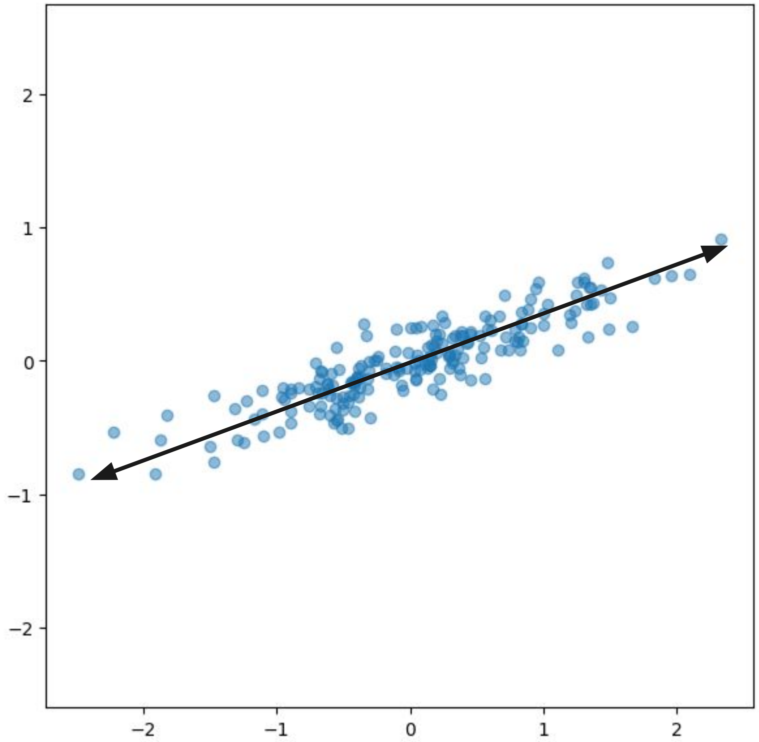
Principal Component Analysis

Idea: Find the *direction* of the highest variance, not the *axis* with the highest variance.



Principal Component Analysis

Idea: Find the *direction* of the highest variance, not the *axis* with the highest variance.

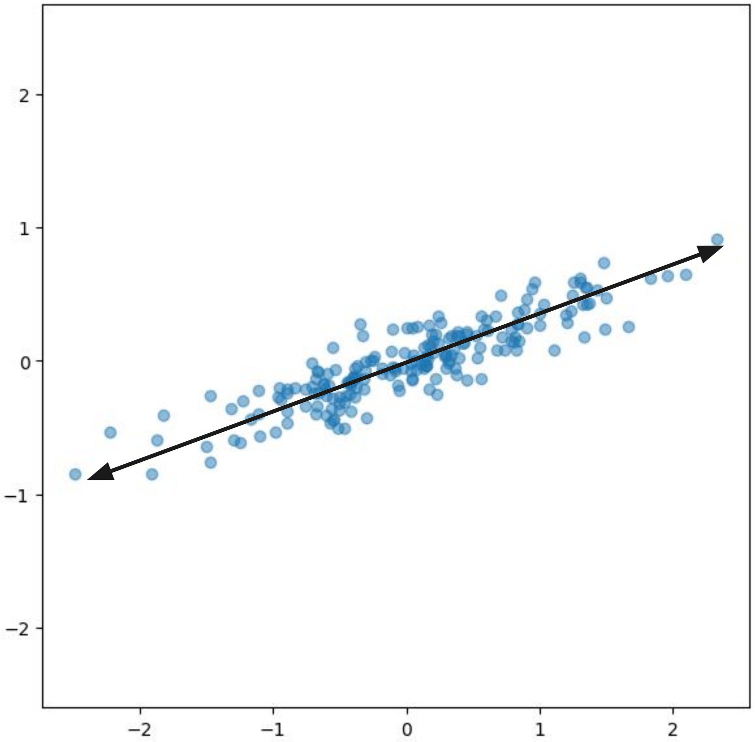


Principal Component Analysis

Idea: Find the *direction* of the highest variance, not the *axis* with the highest variance.

The arrow:

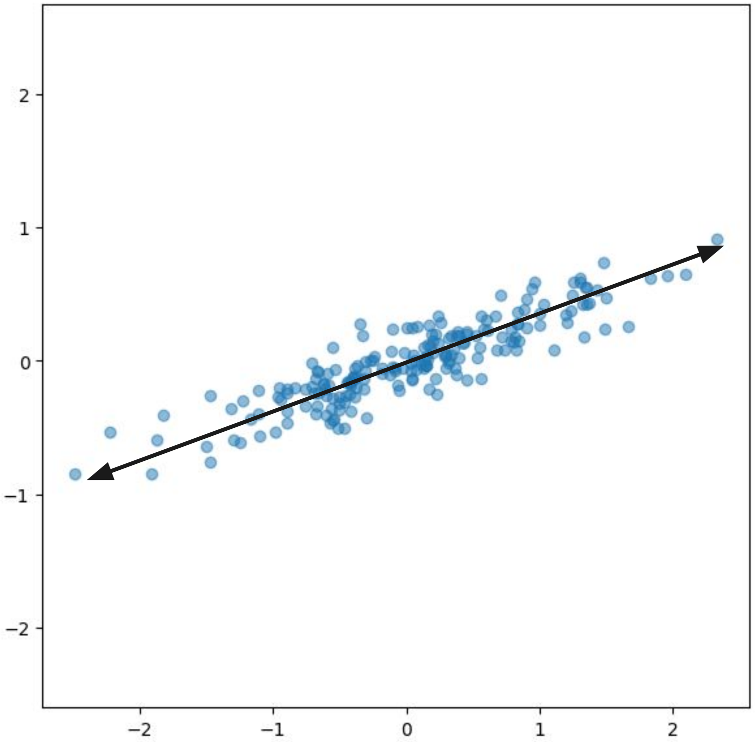
- Is a *principal axis* of the data
- Shows how *important* the axis is (i.e., how much variance the axis contains)



Principal Component Analysis

Principal component analysis (PCA) algorithm:

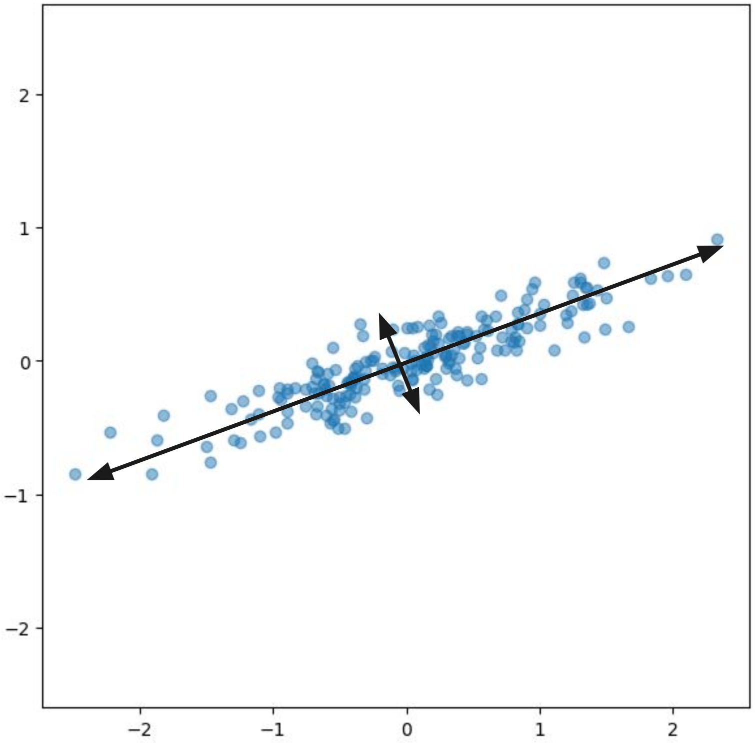
1. Find the principal axis that contains the most variance.
2. Eliminate this axis from consideration
3. Go to 1 until the number of principal axes is equal to the dimensionality of the input data.



Principal Component Analysis

Principal component analysis (PCA) algorithm:

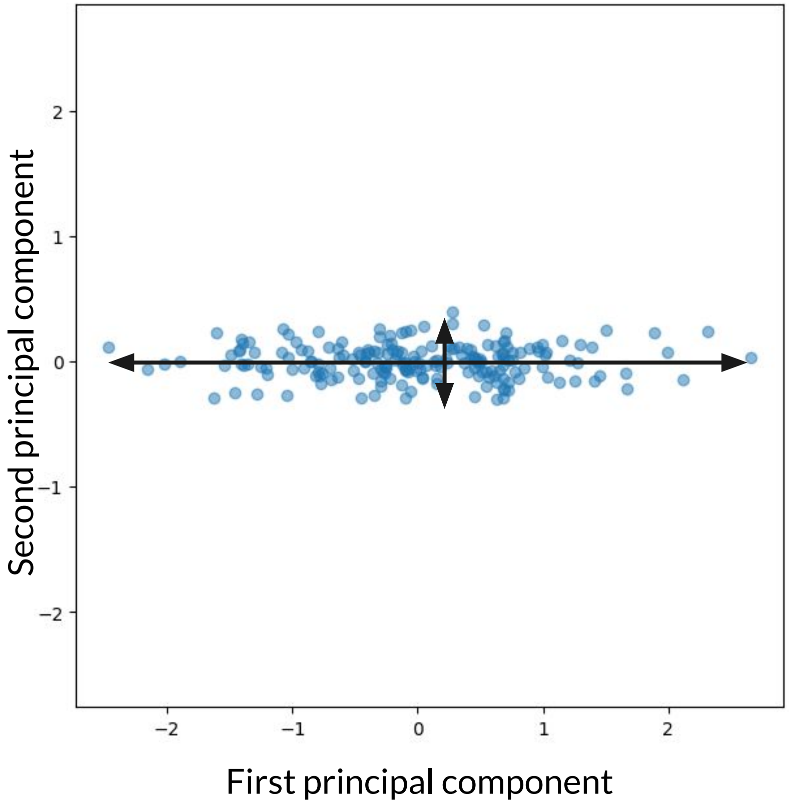
1. Find the principal axis that contains the most variance.
2. Eliminate this axis from consideration
3. Go to 1 until the number of principal axes is equal to the dimensionality of the input data.



Principal Component Analysis

Principal component: A projection of each data point on to the principal axis.

We can plot the principal components:



Demo: PCA and highly dimensional data

scikit-learn PCA documentation:
https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

Uses of PCA

- Visualization
    - Visualize high-dimensional data that is otherwise unvisualizable
        - Automatically find and shows the most important principal axes
    - Demonstrate that there are groups of related instances in your data
    - Justify future clustering/*k*-NN/ML work
- Machine learning
    - Focuses model: Removes low-variance data
    - Memory constraints: Reduce the dimensionality of your data to train faster and save memory