Principal Component Analysis

Principal Component Analysis

- Motivation of Dimension Reduction:
  - Imagine a dataset with 30+ features, how would you understand the key features?
  - Visualization and Data Analysis have limitations when the number of feature dimensions increases.

Principal Component Analysis

- Important Note:
  - Dimensionality Reduction algorithms such as PCA **do not** simply choose a subset of the existing features.
  - They create **new** dimensional components that are combinations of proportions of the existing features.

Principal Component Analysis

- Dimension Reduction
  - Helps visualize and understand complex data sets.
  - Can also act as a simpler data set for training data for machine learning algorithms.
    - Reduce dimensions then train ML Algorithm on smaller data set.

Principal Component Analysis

- Variance Explained
  - We've often seen that certain features are more important or have more explanatory power than other features.
  - For example, size of a house is probably much more important than the color of a house when explaining the price of a house for sale.

Principal Component Analysis

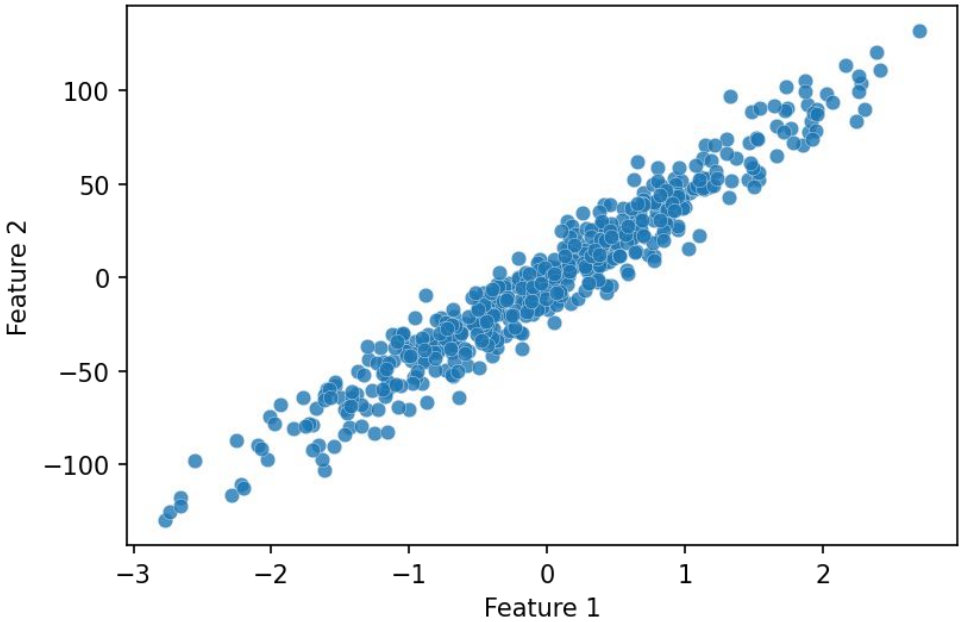
- Variance Explained
  - This idea of more important features is easy to understand when we can directly correlate features to a known label. But what about unlabeled data?
  - What measurement can we use to determine feature “importance”?

Principal Component Analysis

- Variance Explained
  - Measure the proportion to which each feature accounts for dispersion in the data set.

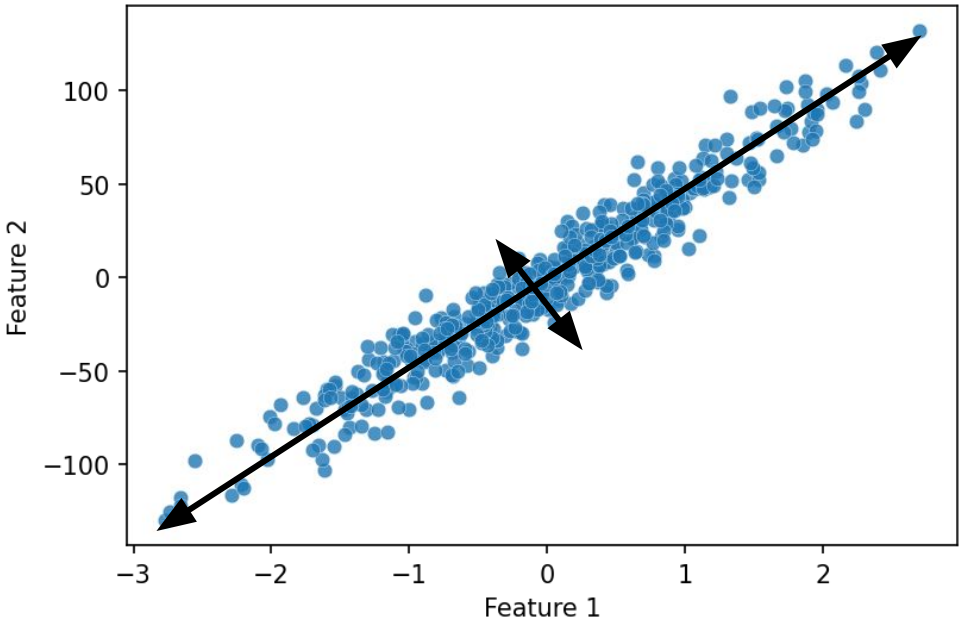
Principal Component Analysis

- Variance Explained



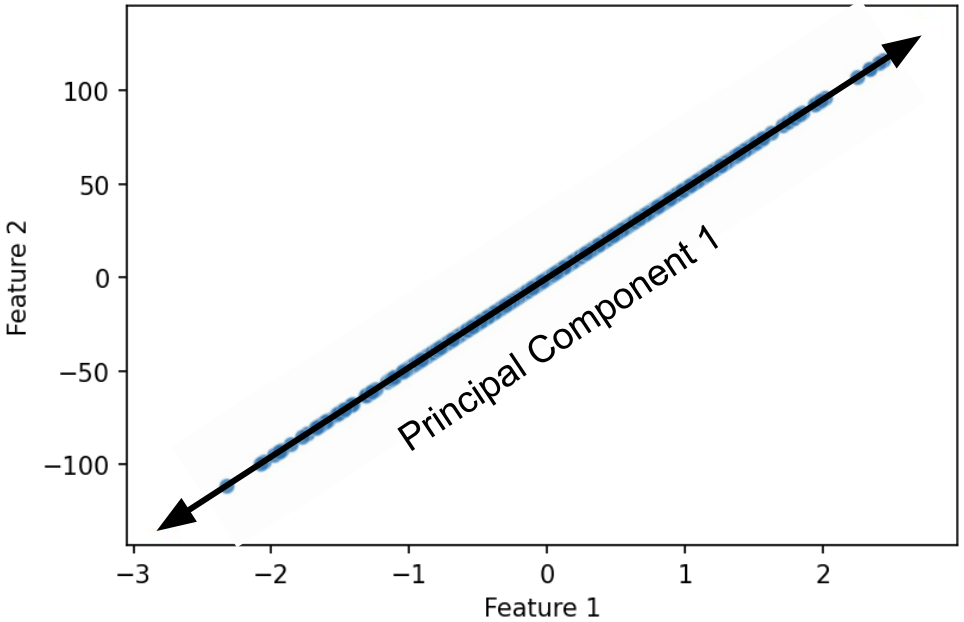
Principal Component Analysis

- Variance Explained



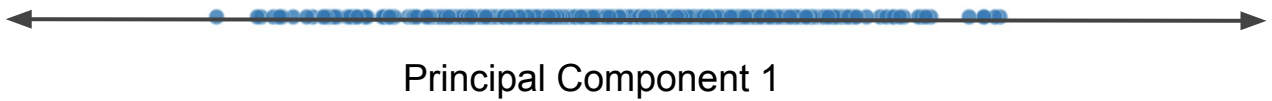
Principal Component Analysis

- Variance Explained



Principal Component Analysis

- Variance Explained



Principal Component Analysis

- Variance Explained
  - Principal Component is a linear combination of original features.
  - The more variance the original feature accounts for, the more influence it has over the principal components.

Principal Component Analysis

- Variance Explained
  - Here we went from 2 features down to 1 principal component.
  - This single principal component can “explain” some percentage of the original data, for example 90% of variance explained by principal component.

Principal Component Analysis

- Variance Explained
  - 100% of the variance in the data is explained by all the original features.
  - We trade off some of the explained variance for less dimensions.
  - This can be significant savings for data sets with many dimensions, but only a few strong features.

Principal Component Analysis

- Principal Component Analysis: