Plan:

1. Define Dimensionality Reduction
2. Explain when dimensionality reduction is useful

Dimensionality Reduction

Shannon E. Ellis, Ph.D
UC San Diego

Department of Cognitive Science
sellis@ucsd.edu

Dimensionality Reduction Outline

- Definition
- When to Use
- Mathematical Overview
- Key Concepts
- Examples
  - Diet in the UK
  - Genetics around the world

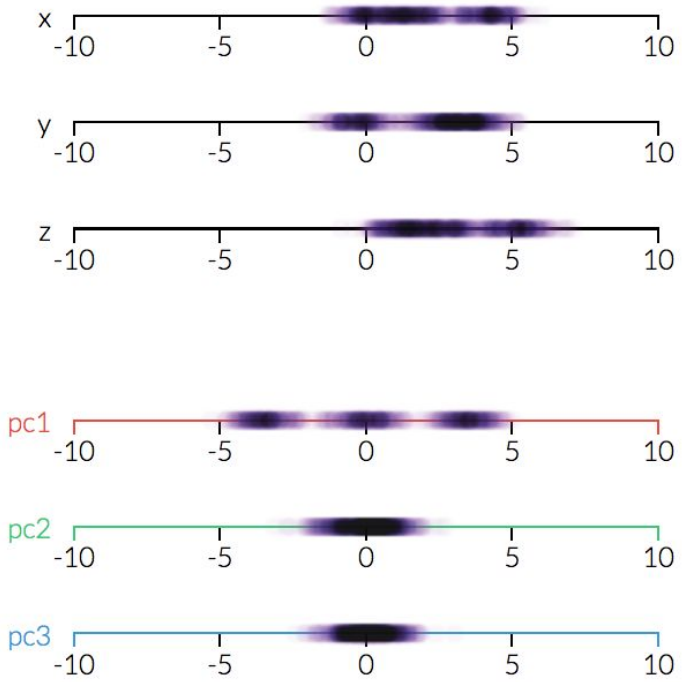
Dimensionality Reduction

A mathematical process to reduce the number of random variables to consider

Discuss: why may we want to do this?

Dimensionality Reduction

- Reduce the dimension of quantitative data to a more manageable set of variables
- Reduced set can then be input to reveal underlying patterns in the data and/or as inputs in a model (regression, classification, etc.)



Use Cases for Dimensionality Reduction

- Thousands of sensors used to monitor an industrial process
  - Reducing the data from these 1000s of sensors to a few features, we can then build an interpretable model
  - Goal : predict process failure from sensors
- Understanding diet around the world
  - Amount of foods eaten among populations across the world
  - Goal: identify diet similarity among populations
- Identify genetic ancestry
  - Determine ancestral origins based on genetic variation
  - Goal: Learn more about our genetic history

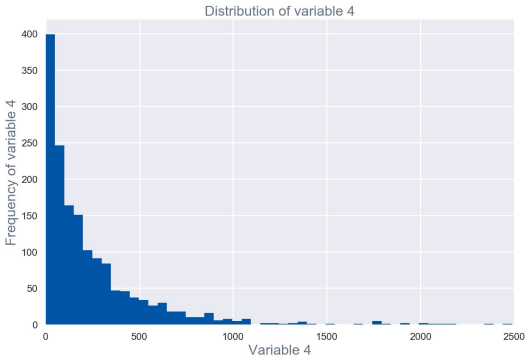
As an extension of EDA

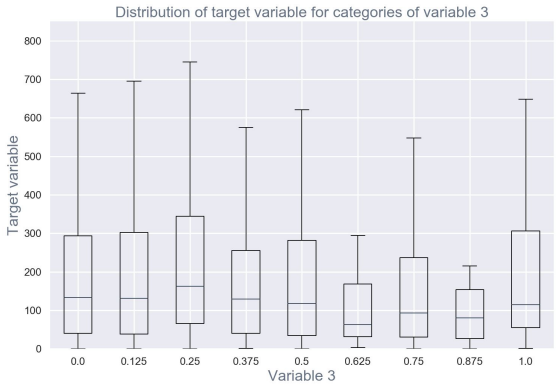
- Gain insight into a set of data
- Understand how different variables relate to one another

EDA Approaches to “Get a Feel for the Data”

Exploratory

Understanding the relationship between variables in your dataset





Univariate

understanding a single variable

i.e.: histogram, densityplot, barplot

Bivariate

understanding relationship between 2 variables

i.e.: boxplot, scatterplot, grouped barplot, boxplot

Dimensionality Reduction

projecting high-D data into a lower-D space

i.e.: PCA, ICA, Clustering

Image source: https://www.svds.com/value-exploratory-data-analysis/

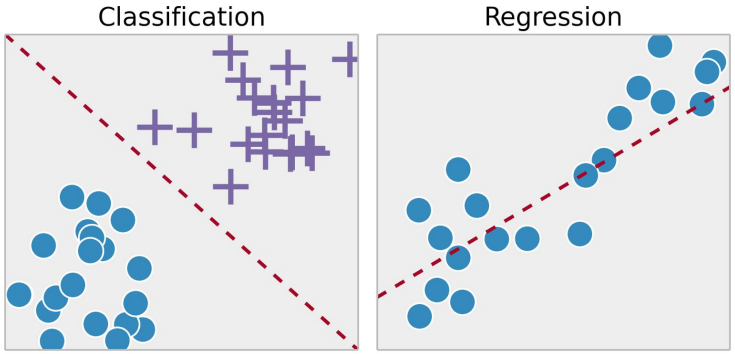
As an extension of EDA

- Gain insight into a set of data
- Understand how different variables relate to one another

Note: PCA/Dimensionality reduction can also be used for modeling & prediction

Approaches to machine learning

Supervised Learning

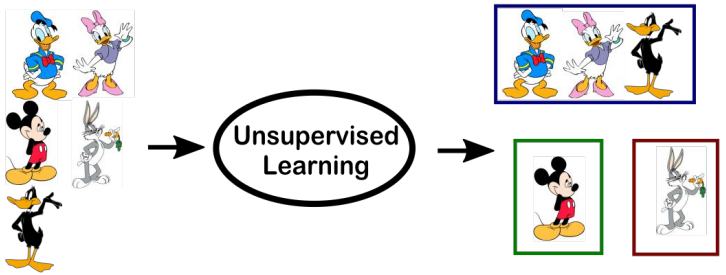


categorical variables

continuous variables

Prediction accuracy dependent on training data

Unsupervised Learning



Clustering (categorical) & dimensionality reduction (continuous)

can automatically identify structure in data

Image source: https://towardsdatascience.com/supervised-vs-unsupervised-learning-14f68e32ea8d

Principal Component Analysis (PCA)

Key Terms:

- **Principal Component (PC)** - a linear combination of the predictor variables
- **Loadings** - the weights that transform the predictors into components (aka weights)
- **Screeplot** - variances of each component plotted

Principal Component Analysis (PCA)

Goal : combine multiple **numeric predictor** variables into a smaller set of variables. Each variable in this smaller set is a weighted linear combination of the original set.

This smaller set of variables -- the *principal components* (PCs) - “explain” most of the variability of the full set of variables....but uses many fewer dimensions to do so.

The weights (loadings) used to form the PCs explain the relative contributions of the original variables to the new PCs.

Source: Practical Statistics for Data Scientists (Bruce & Bruce)

“Simple” PCA : Two predictor variables (X_1 and X_2)

For two variables, X_1 and X_2, there are two principal components Z_i ($i = 1$ or 2):

$$Z_i = w_{i,1}X_1 + w_{i,2}X_2$$

$w_{i,1}$ and $w_{i,2}$: weightings (*loadings*)

- Transform the original variables into principal components

Z_1 : the first principal component (PC1)

- The linear combination that best explains the total variance

Source: Practical Statistics for Data Scientists (Bruce & Bruce)

Stock Price returns for Chevron (CVX) and ExxonMobil (XOM)

PC1 and PC2 are the dotted lines on the plot

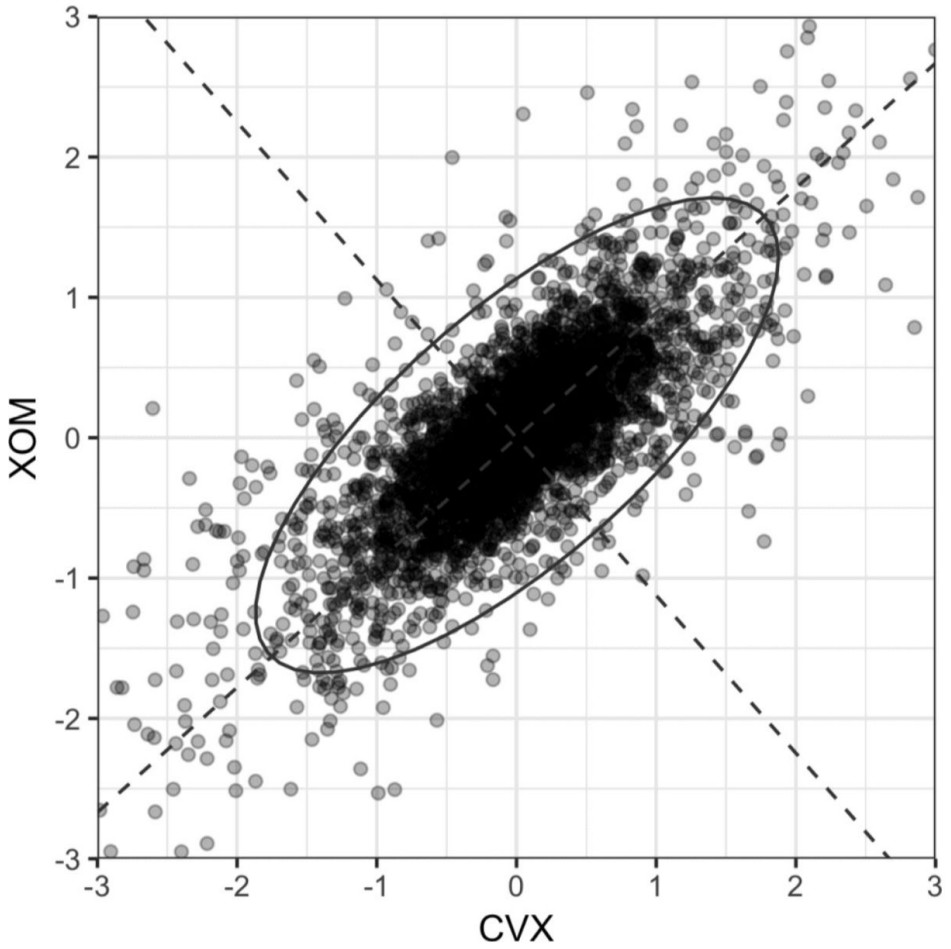


Figure 7-1. The principal components for the stock returns for Chevron and ExxonMobil

Source: Practical Statistics for Data Scientists (Bruce & Bruce)

Principal Component Analysis (PCA)

But....PCA shines when you're dealing with high-dimensional data. So we have to move *beyond* two predictors to many predictors....

Step 1: Combine all predictors in linear combination

Step 2: Assign weights that optimize the collection of the covariation to the first PC (Z_1) (maximizes the % total variance explained)

Step 3: Repeat Step 2 to generate new predictor Z_2 (second PC) with different weights. By definition Z_1 and Z_2 are uncorrelated. Continue until you have as many new variables (PCs) as original predictors

Step 4: Retain as many components as are needed to account for *most* of the variance.

Source: Practical Statistics for Data Scientists (Bruce & Bruce)