Principal Component Analysis

Created Date: 2022-11-17

> Metadata 🧰
>
> - Title: Principal Component Analysis
> - Author: Andrew Jones
> - Reference: Data Science Infinity

> Links & Tags 🖇️
>
> - Index: Course Note Index
> - Atomic Tag: #datascience
> - Subatomic Tags: #machinelearning #pca #unsupervisedlearning #supervisedlearning

---

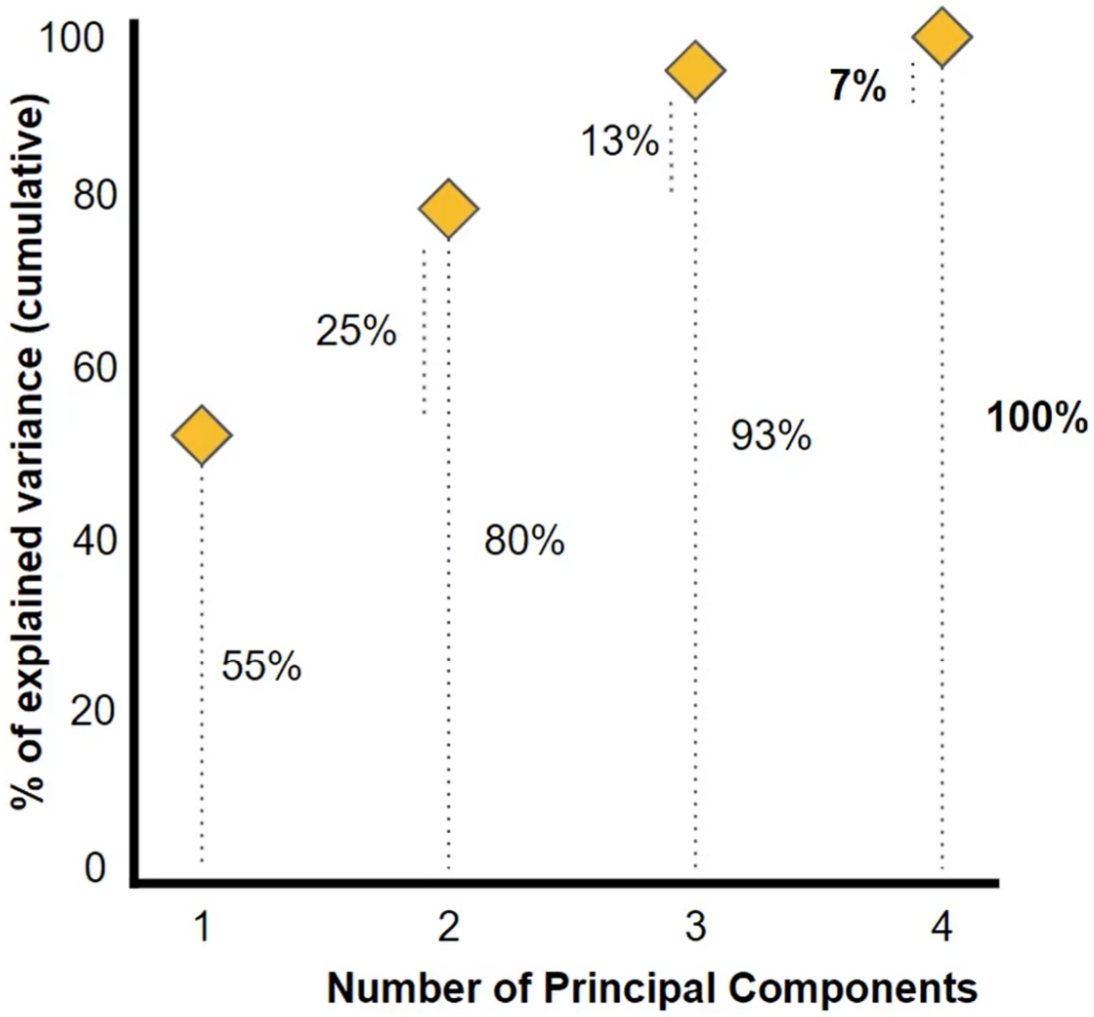
High-Level Overview

Jupyter Notebook: Advanced PCA Template

> Principal Component Analysis (PCA) is often used as a *Dimensionality Reduction* technique that can reduce a large set of variables down to a smaller set that still contains most of the original information.

- Aims to simplify the space we are operating in
- Applying models to a vast amount of features (variables) can be computationally expensive
- PCA allows us to reduce the computation expense while maintaining the key insights
- Applicable to both unsupervised and supervised learning

- We choose how many components we want to reduce the original feature set to
- To determine the optimal amount of principal components, we measure the % of explained variance for each component



Principal Component Analysis (PCA)

Var_1	Var_2	Var_3	Var_4	Var_5	Var_6	Var_7	Var_8	Var_9	Var_10
0.83	0.07	0.52	0.63	0.54	0.12	0.57	0.26	0.02	0.43
0.54	0.59	0.16	0.39	0.07	0.30	0.22	0.87	0.34	0.21
0.28	0.56	0.07	0.49	0.51	0.20	0.90	0.59	0.77	0.10
0.12	0.15	0.83	0.91	0.53	0.98	0.06	0.35	0.22	0.21

PC1	PC2	PC3
0.71	-0.57	-0.14
0.34	-0.07	0.14
0.01	0.12	-0.07
-0.12	0.22	0.33

Principal Components

Considerations when Applying PCA

- Always apply feature scaling, using standardization over normalization
- We will lose some of the information/variance contained in the original data
- It is more difficult to interpret the outputs based on component values versus original variables