PCA :

Principal Component Analysis

12.08.2021

Hannah Do

0. Introduction

Dimension reduction technique is commonly used to pre-process the dataset.

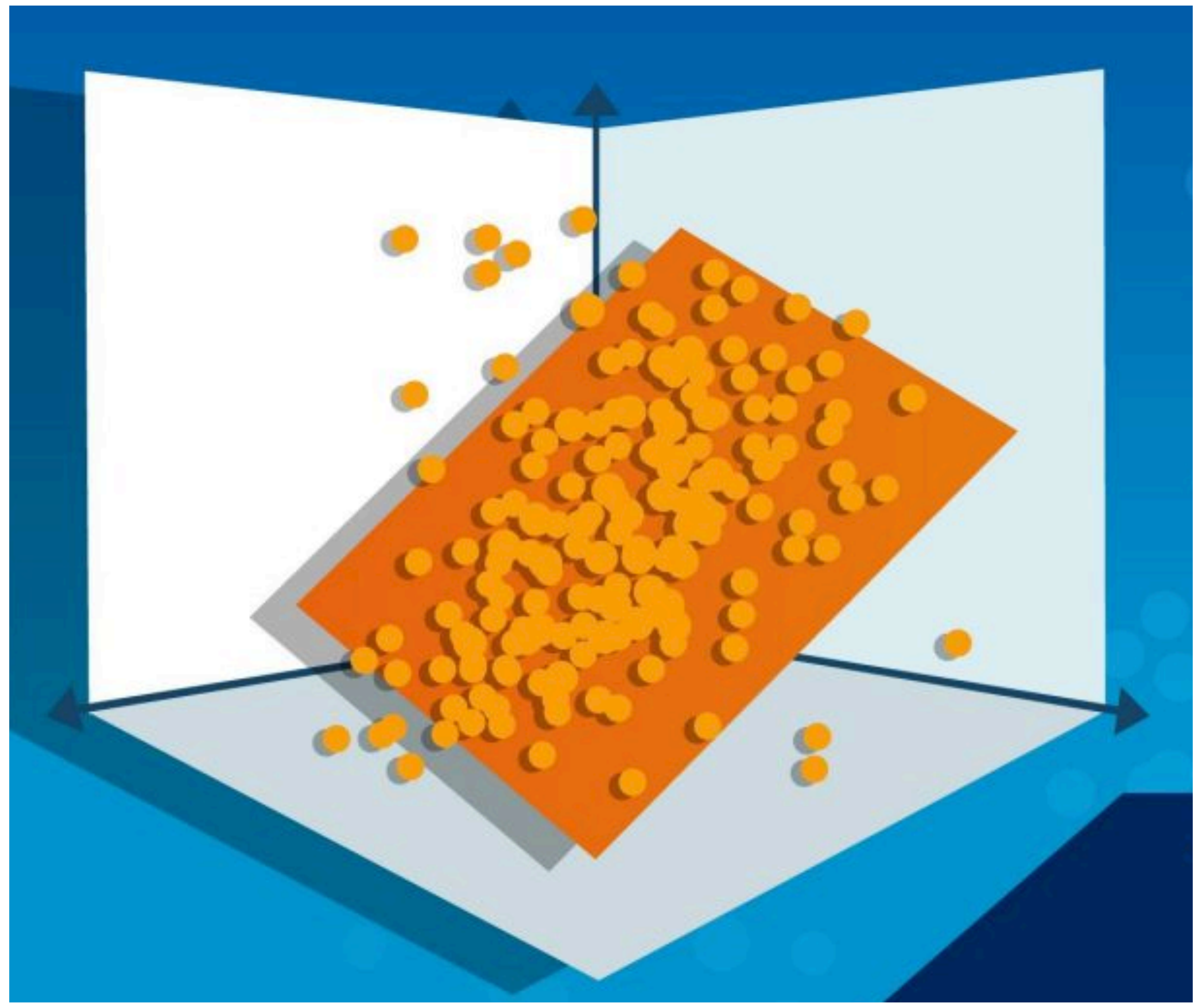
Why is dimension reduction used?

- Takes less computation or training time
- Takes care of multicollinearity by removing redundant features.
- Helps in visualizing data.

0. Introduction

Principal Component Analysis

Statistical procedure that allows you to **summarize the information content** in large data tables **by means of** a smaller set of “**summary indices**” that can be more easily visualized and analyzed.

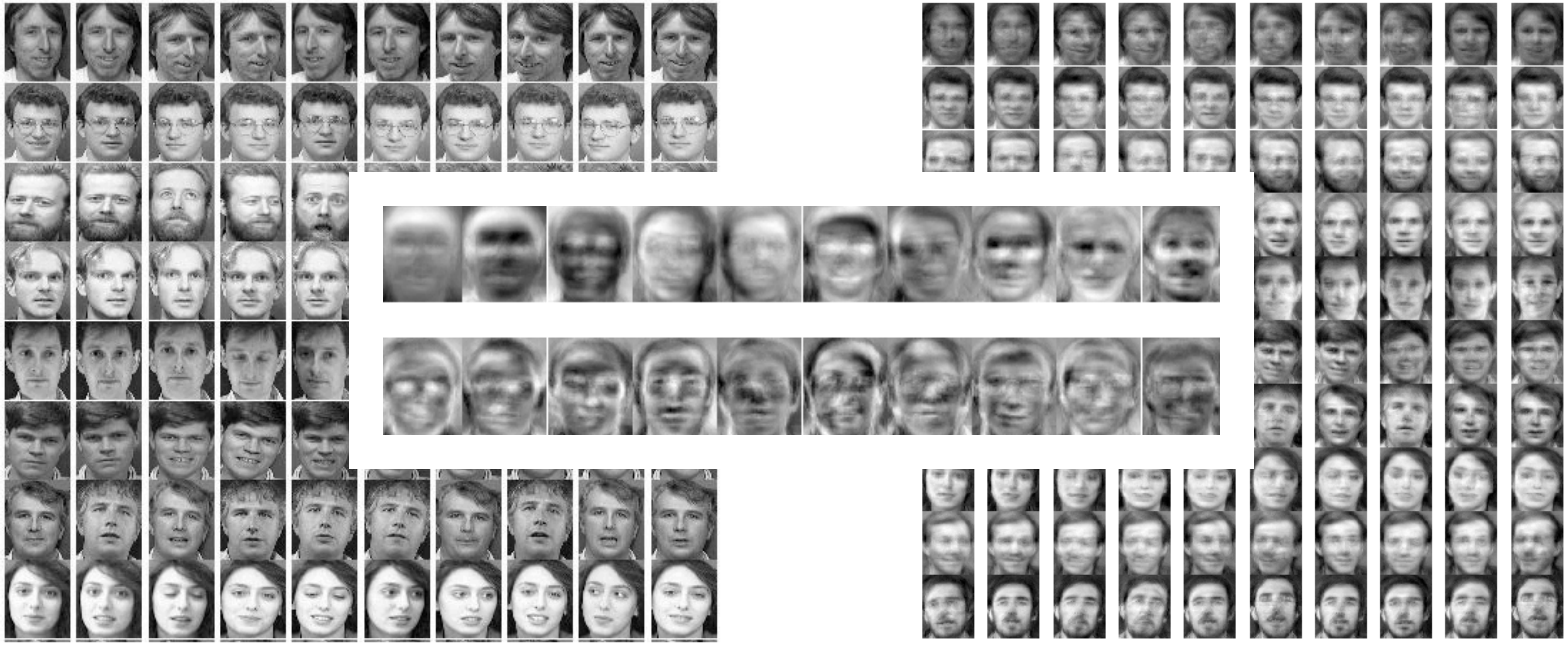


What is: PCA? (365datascience.com)

https://www.sartorius.com/en/knowledge/science-snippets/what-is-principal-component-analysis-pca-and-how-it-is-used-507186

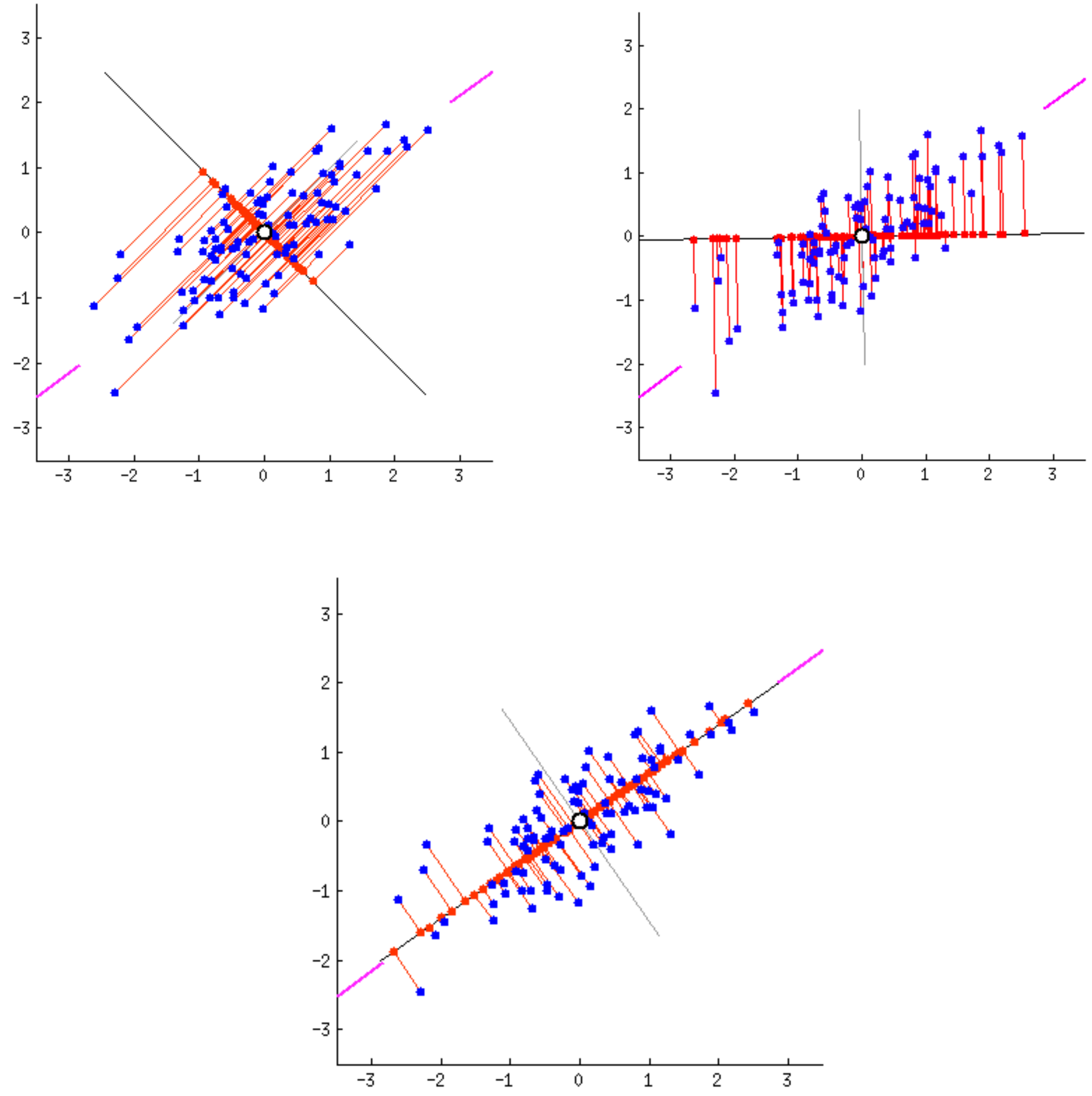
0. Introduction

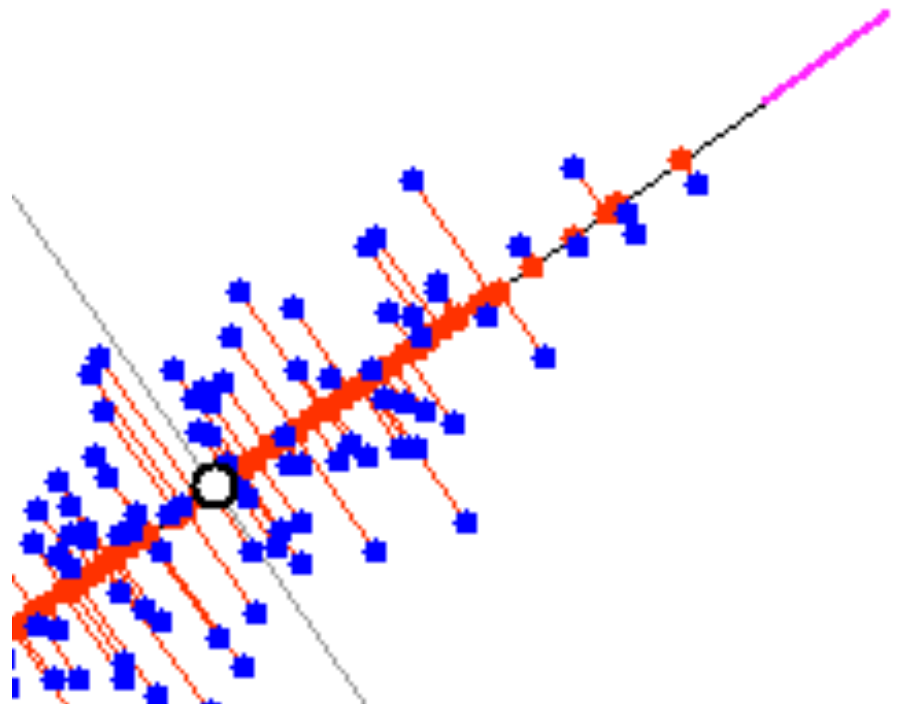
PCA can be used to reduce dimensions of a dataset into a smaller set of dimensions by using eigenvectors of the matrices to recreate the matrices in smaller dimension.



Sebastian Norena (2018) PCA applied to images of faces

1. PCA : Extracting Principal Components





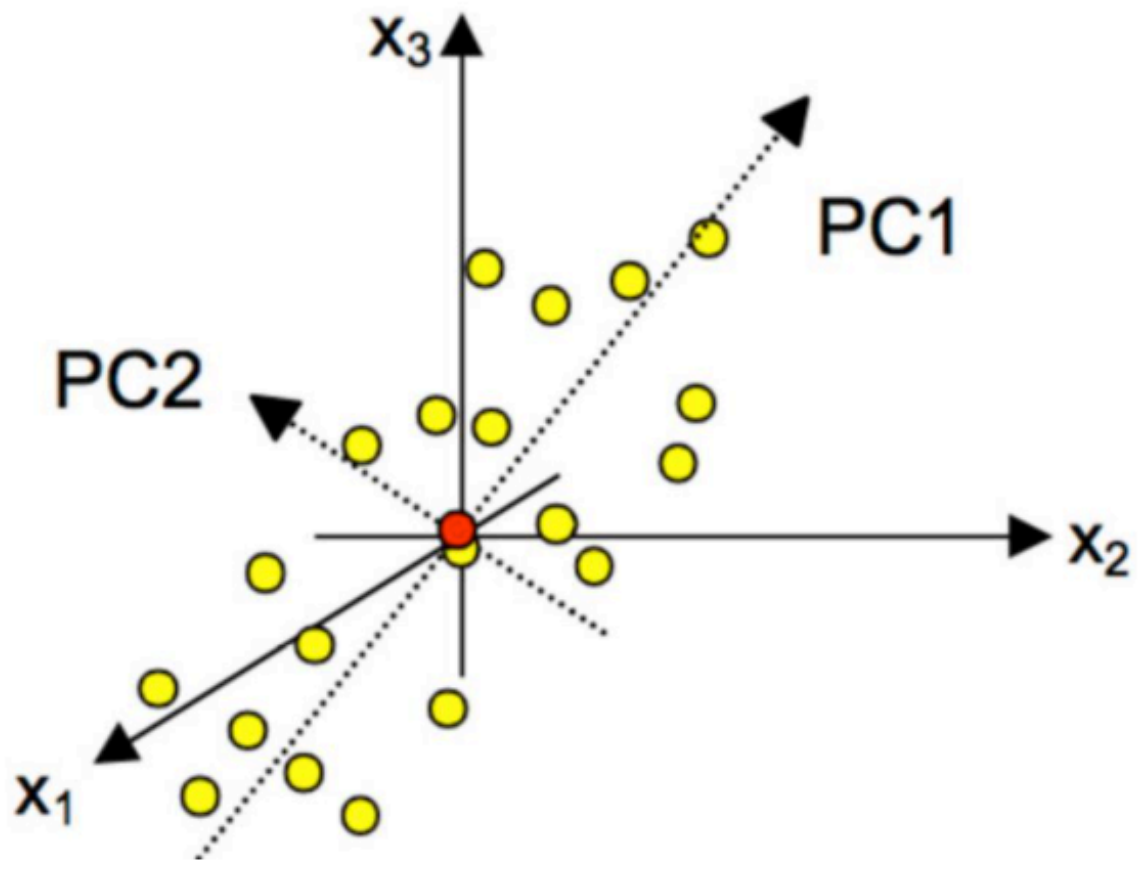
Eigenvector :
Points from mean-center
to projected red dots

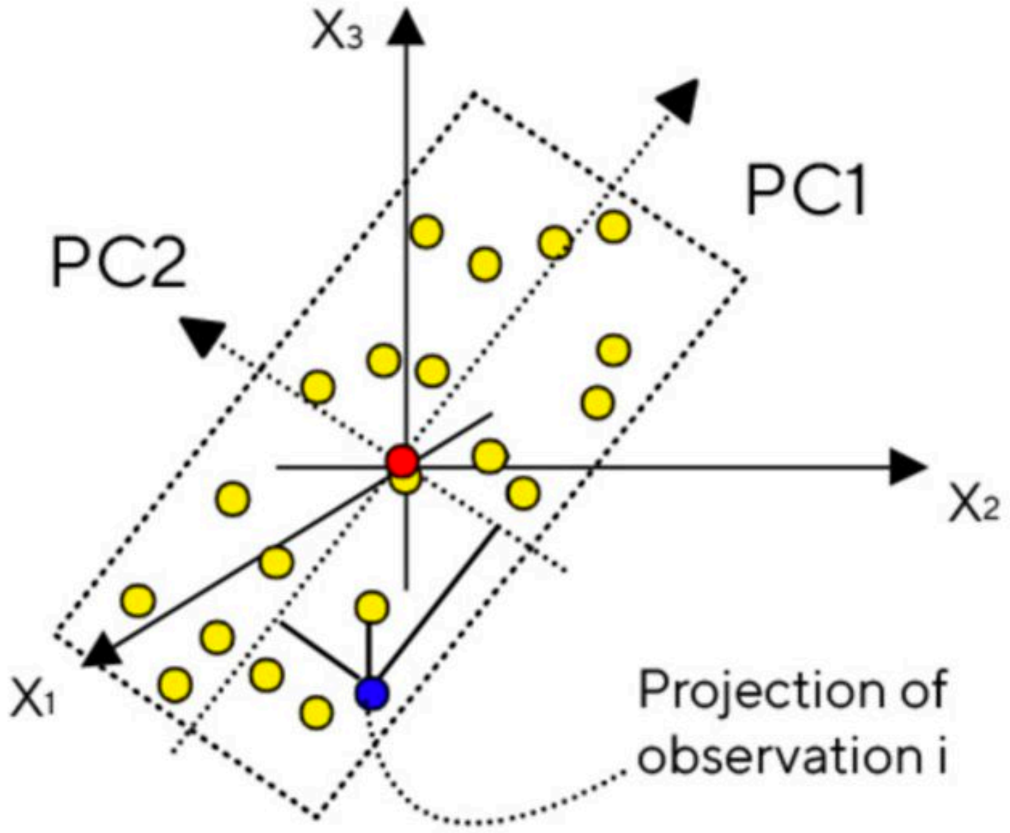
Extracting the first principal component

- Let’s say there is a 2-dimensional graph with feature X and Y
- In order to find the first Principal Component, we need to find a mean-center and scale the rows to unit variance.
- Then a line is selected that has the most variance (spread of dots) - keeping the characteristic of the original features at its best.
- The eigenvectors are then calculated (vectors that start from the center to the red dot) and saved as a new feature - PC1

https://i.stack.imgur.com/Q7HIP.gif
https://stats.stackexchange.com/questions/2691/making-sense-of-principal-component-analysis-eigenvectors-eigenvalues/140579

1. PCA : Extracting Principal Components





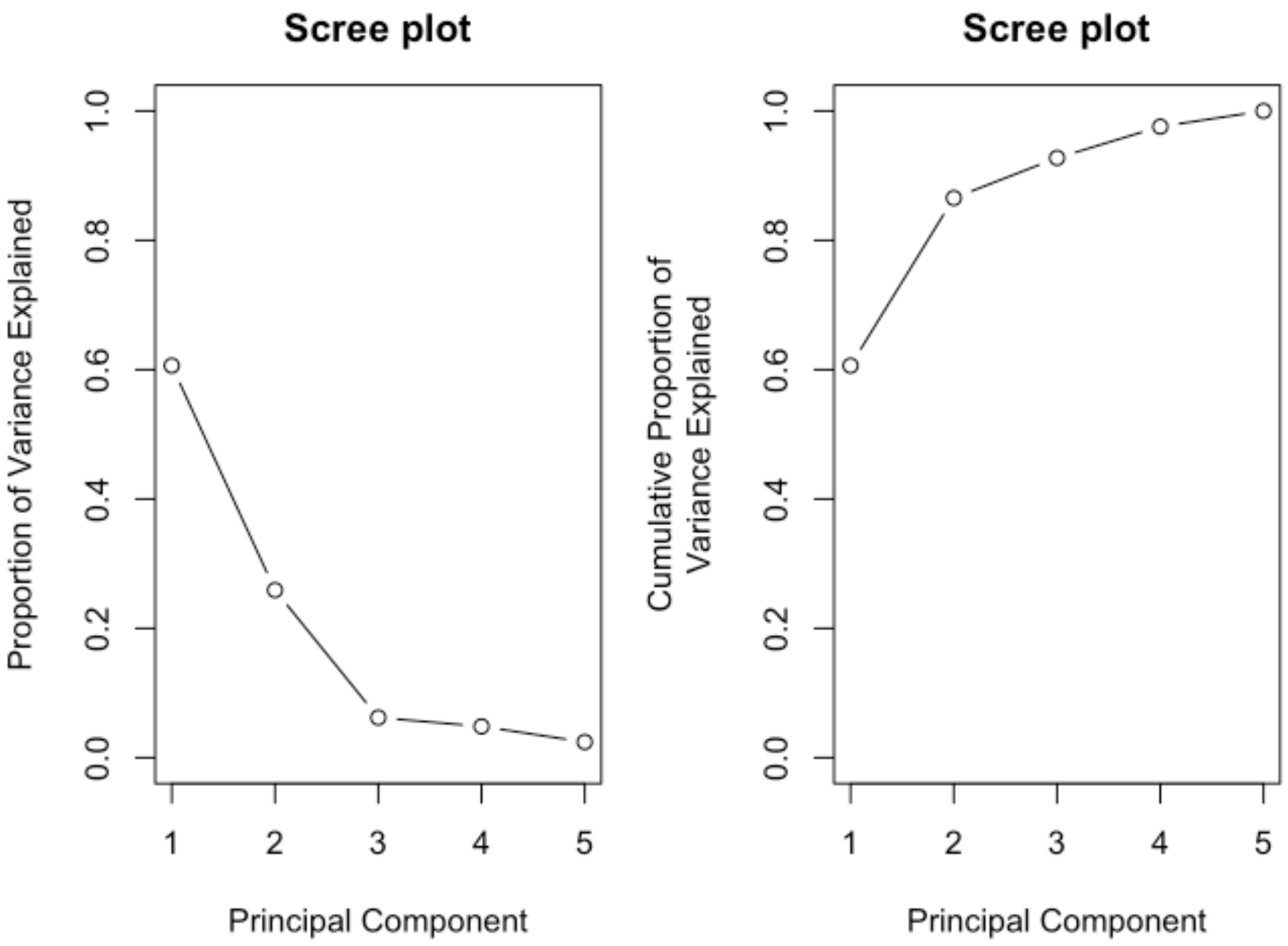
Extracting other principal components

- The second principal component (PC2) is oriented such that it reflects the second largest source of variation in the data while being orthogonal to the first PC. PC2 also passes through the average point.

- Consequently the k-th component is calculated by finding another orthogonal principal components from the original matrix.

https://i.stack.imgur.com/Q7HIP.gif
https://stats.stackexchange.com/questions/2691/making-sense-of-principal-component-analysis-eigenvectors-eigenvalues/140579

2. PCA : Choosing the most relevant principal components

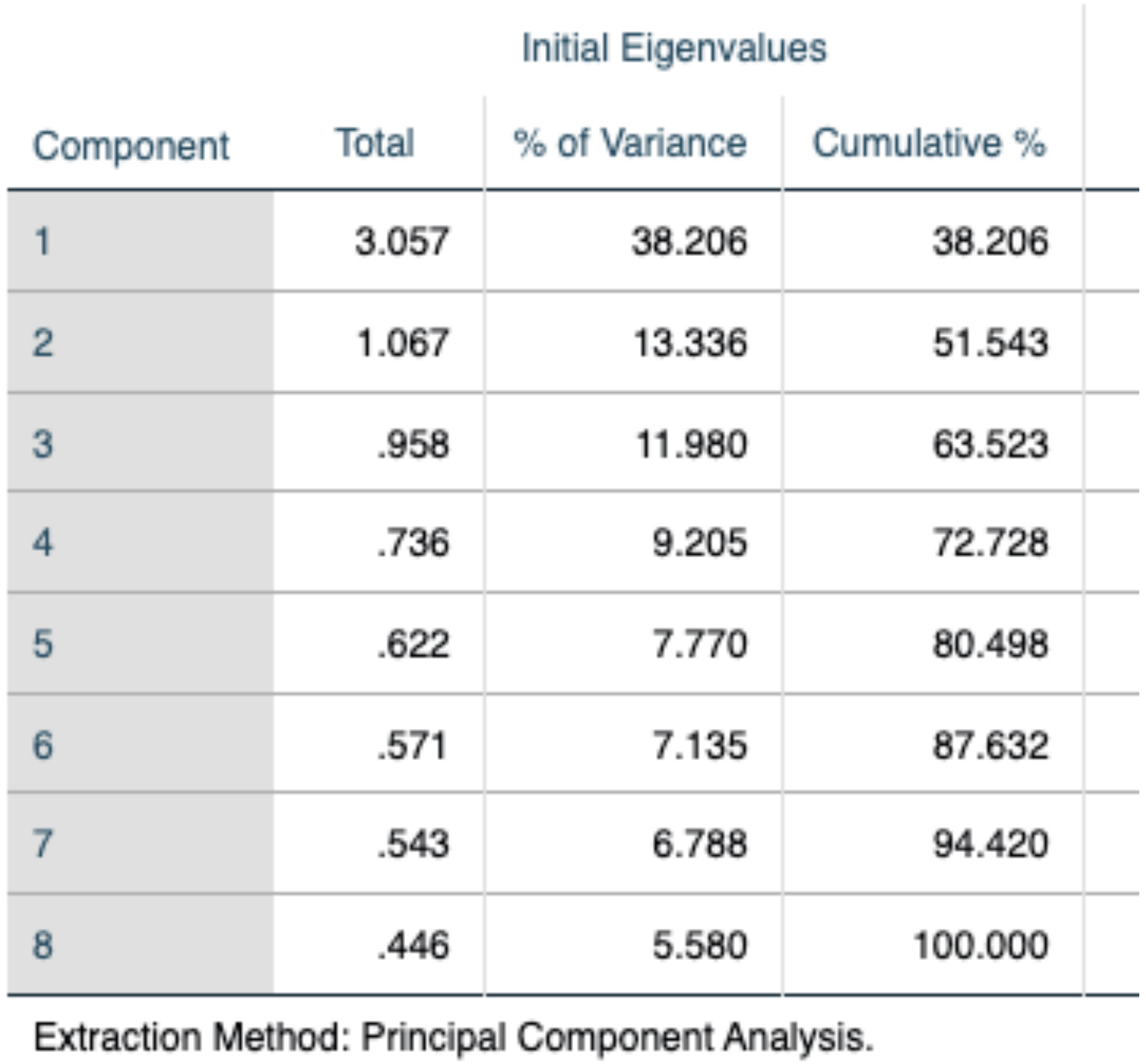


How do we choose the number of principal components?

A widely applied approach is to decide on the number of principal components by examining a **scree plot**. By eyeballing the scree plot, and looking for a point at which the proportion of variance explained by each subsequent principal component drops off. This is often referred to as an *elbow* in the scree plot.

https://www.geo.fu-berlin.de/en/v/soga/Geodata-analysis/Principal-Component-Analysis/principal-components-basics/Choose-principal-components/index.html

2. PCA : Choosing the most relevant principal components



	Initial Eigenvalues		
Component	Total	% of Variance	Cumulative %
1	3.057	38.206	38.206
2	1.067	13.336	51.543
3	.958	11.980	63.523
4	.736	9.205	72.728
5	.622	7.770	80.498
6	.571	7.135	87.632
7	.543	6.788	94.420
8	.446	5.580	100.000

Extraction Method: Principal Component Analysis.

https://stats.idre.ucla.edu/spss/seminars/efa-spss/

Kaiser rule

Other than the scree plot (elbow method) that addresses the explained variance of the principal components, we have **Kaiser's rule** to determine the number of PCs.

Kaiser rule chooses PCs with eigenvalues that are greater than 1.

3. Summary

PCA

- Powerful in dealing with multicollinearity
- Used when variables outnumber the samples ($d>n$)
- Highly affected by outliers
- Can decide on how much variance to preserve using eigenvalues

1. References

1. PRINCIPAL COMPONENTS (PCA) AND EXPLORATORY FACTOR ANALYSIS (EFA) WITH SPSS, UCLA Statistical Consulting
   https://stats.idre.ucla.edu/spss/seminars/efa-spss/\

2. Principal Component Analysis - The basics, Freie University Berlin.
   https://www.geo.fu-berlin.de/en/v/soga/Geodata-analysis/Principal-Component-Analysis/principal-components-basics/index.html

3. What Is Principal Component Analysis (PCA) and How It Is Used?, Sartorius - Data Analytics
   https://www.sartorius.com/en/knowledge/science-snippets/what-is-principal-component-analysis-pca-and-how-it-is-used-507186