Roberto Pagliarini

DIMENSIONALITY REDUCTION: PRINCIPAL COMPONENT ANALYSIS

Introduction to Dimensionality Reduction

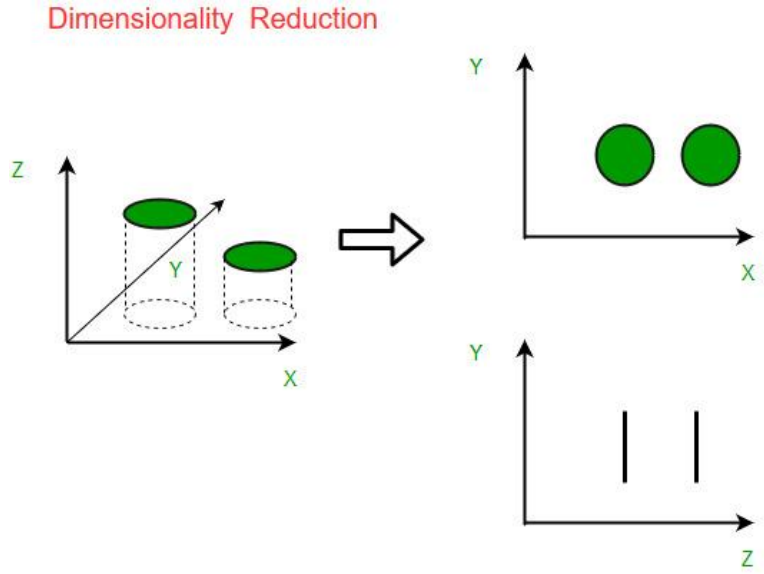
- When working with mathematical models, datasets with too many features can cause issues like slow computation and overfitting.

- Dimensionality reduction helps to reduce the number of features while retaining key information.
  - Principal Component Analysis (PCA)
  - Singular value decomposition (SVD)
  - Linear discriminant analysis (LDA)

- These techniques convert data into a lower-dimensional space while preserving important details.

Let us consider an example

- Imagine a dataset where each data point exists in a 3D space defined by axes X, Y and Z
- If most of the data variance occurs along X and Y then the Z-dimension may contribute very little to understanding the structure of the data.
  - These new features don't overlap with each other and the first few keep most of the important differences found in the original data.



What is Principal Component Analysis (PCA)?

- It helps you to reduce the number of features in a dataset while keeping the most important information.

- It changes your original features into new features.
  - These new features don't overlap with each other and the first few keep most of the important differences found in the original data.

- PCA is commonly used for data preprocessing for use with machine learning algorithms. It helps to remove redundancy, improve computational efficiency and make data easier to visualize and analyze especially when dealing with high-dimensional data.

How PCA works

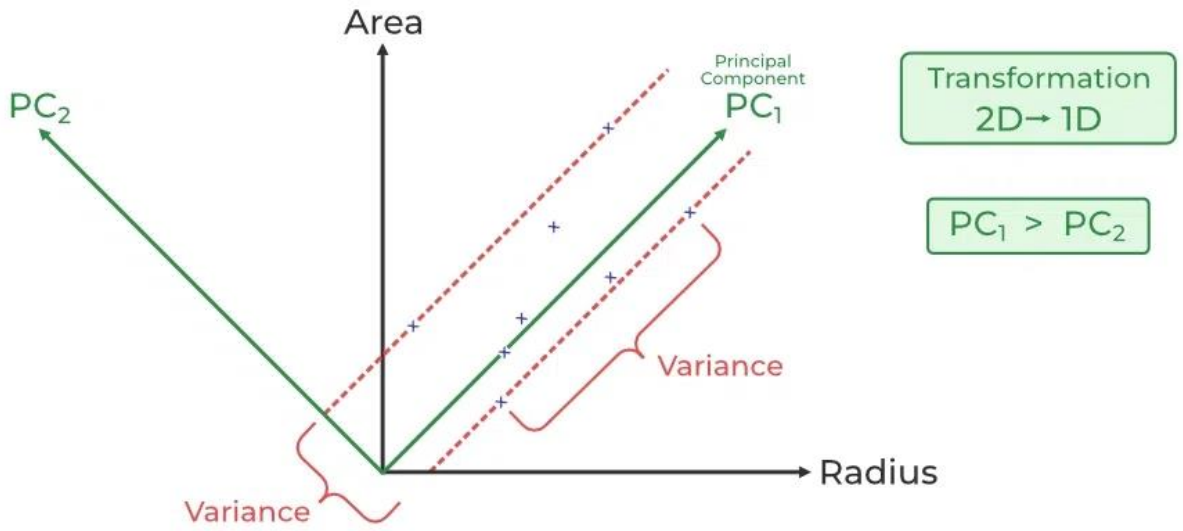
- It uses linear algebra to transform data into new features, the **principal components**

1. ***Standardize data***: different features may have different units and scales. After that, each feature has mean 0 and standard deviation 1.
2. ***Calculate covariance matrix:*** to see how features relate to each other whether they increase or decrease together.
3. ***Find the principal components:*** PCA identifies, by solving a system of equations, new axes where the data spreads out the most (PC1 and PC2).
4. **Pick the top directions & transform data:**
    1. Select the top k components hat capture most of the variance like 95%.
    2. Transform the original dataset by projecting it onto these top components.

- We reduced the number of features while keeping the important patterns in the data.

Example: a dataset with two features: "Radius" and "Area"

- PCA identifies two new directions: PC_1 and PC_2 which are the principal components.
- These new axes are rotated versions of the original ones. PC_1 captures the maximum variance in the data meaning it holds the most information while PC_2 captures the remaining variance and is perpendicular to PC_1.

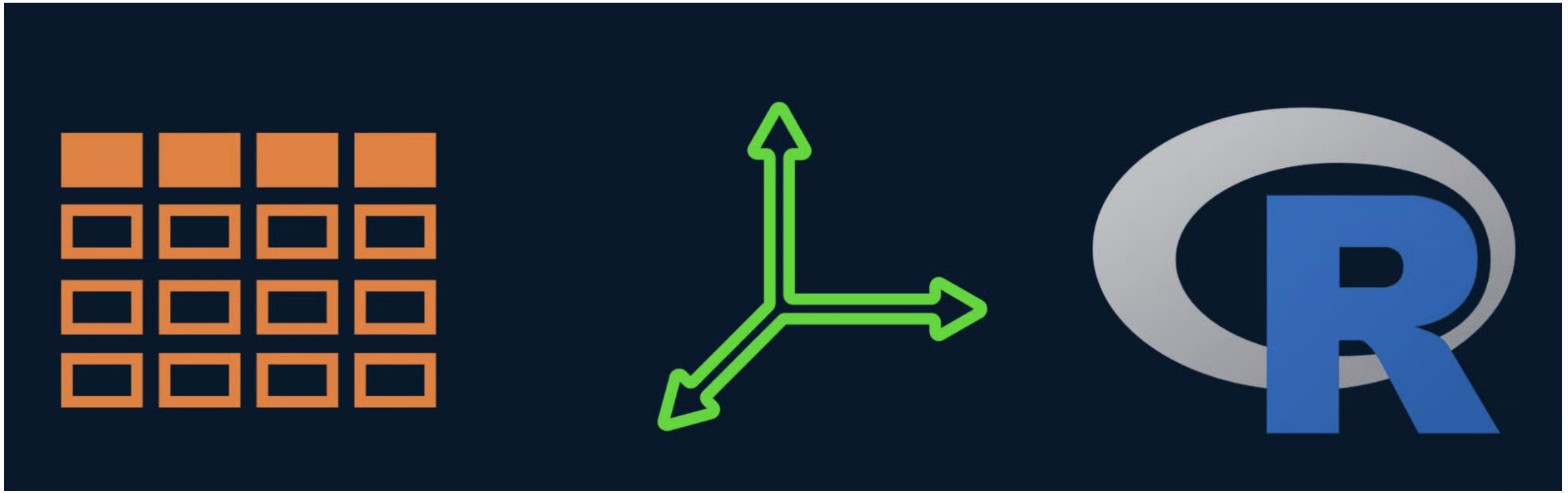


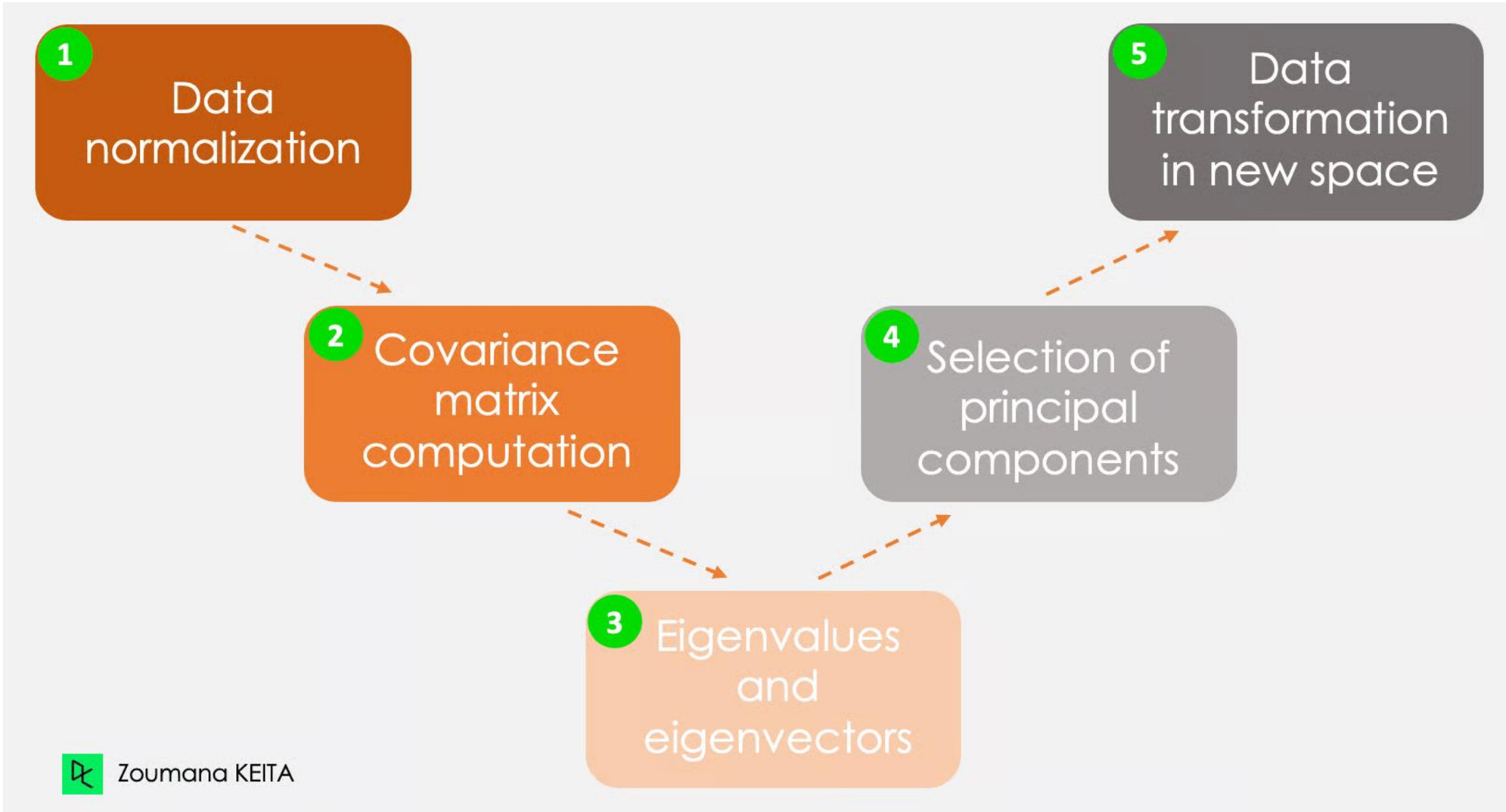
Advantages of Principal Component Analysis

1. **Multicollinearity Handling:** creates new, uncorrelated variables to address issues when original features are highly correlated.
2. **Noise Reduction:** eliminates components with low variance enhance data clarity.
3. **Data Compression:** represents data with fewer components reduce storage needs and speeding up processing.
4. **Outlier Detection:** identifies unusual data points by showing which ones deviate significantly in the reduced space.

Disadvantages of Principal Component Analysis

1. **Interpretation Challenges:** the new components are combinations of original variables which can be hard to explain.
2. **Data Scaling Sensitivity:** requires proper scaling of data before application or results may be misleading.
3. **Information Loss:** reducing dimensions may lose some important information if too few components are kept.
4. **Assumption of Linearity:** works best when relationships between variables are linear and may struggle with non-linear data.
5. **Computational Complexity:** can be slow and resource-intensive on very large datasets.
6. **Risk of Overfitting:** using too many components or working with a small dataset might lead to models that don't generalize well.




1
Data normalization
2
Covariance matrix computation
3
Eigenvalues and eigenvectors
4
Selection of principal components
5
Data transformation in new space
Zoumana KEITA

Violent Crime Rates by US State

- **Description:** this data set contains statistics, in arrests per 100,000 residents for assault, murder, and rape in each of the 50 US states in 1973. Also given is the percent of the population living in urban areas.
- **Format:** a data frame with 50 observations on 4 variables.

  - [,1] Murder: numeric Murder arrests (per 100,000)
  - [,2] Assault: numeric Assault arrests (per 100,000)
  - [,3] UrbanPop: numeric Percent urban population
  - [,4] Rape: numeric Rape arrests (per 100,000)