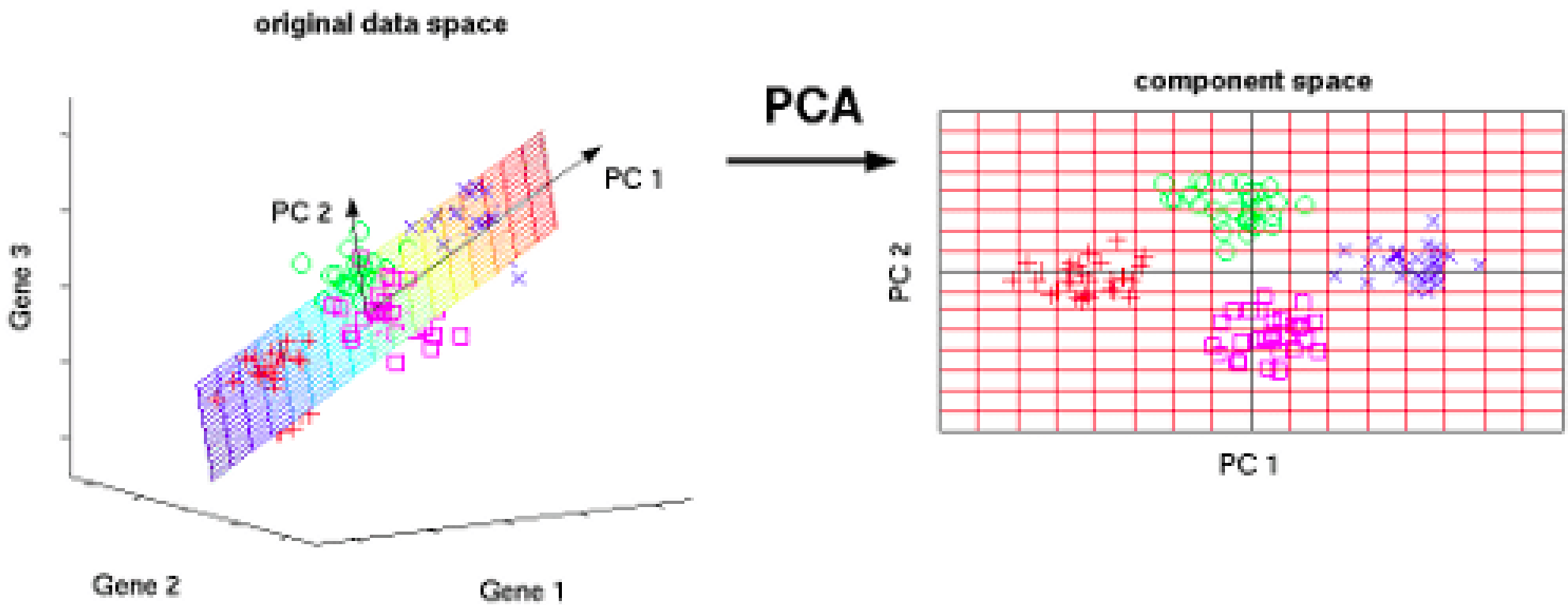
Analytics Vidhya

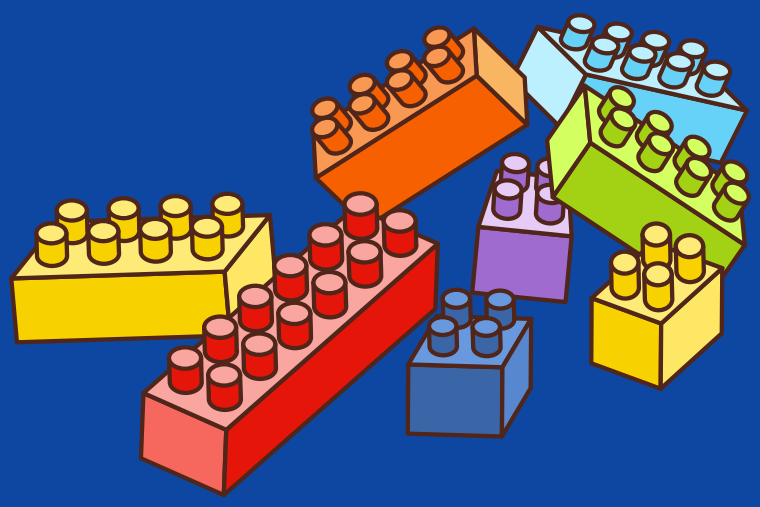
A Guide to

PCA

Principal Component Analysis



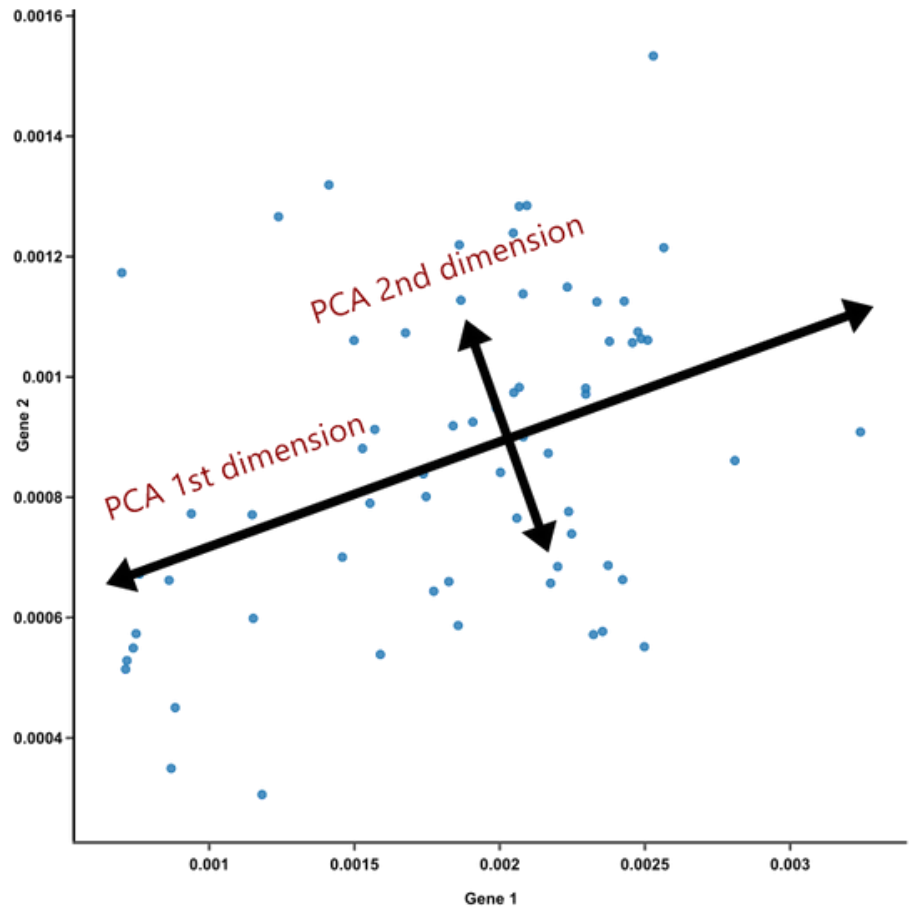
Imagine you have a big box of colorful blocks, each representing a different kind of information. However, sometimes there are so many blocks that it becomes hard to understand what they mean.



Principal Component Analysis (PCA) is like having a magic power that helps you organize these blocks and find the most important ones.

What is Principal Component Analysis?

Principal Component Analysis (PCA) is an unsupervised learning algorithm that is used for reducing the dimensionality of a dataset while preserving maximum variation. It transforms the original variables into a new set of linearly uncorrelated variables called principal components.



Properties of principal components

- The resulting principal components should be a **linear combination of the original features** without any outliers that can distort the covariance matrix and affect the identification of the most significant characteristics.

- These components should be **independent of one another,** meaning that there is no correlation between pairs of variables.

- As we move from the first to the last principal component, their relative importance decreases, **with the first one being the most important and the last one being the least significant**.

Analytics Vidhya

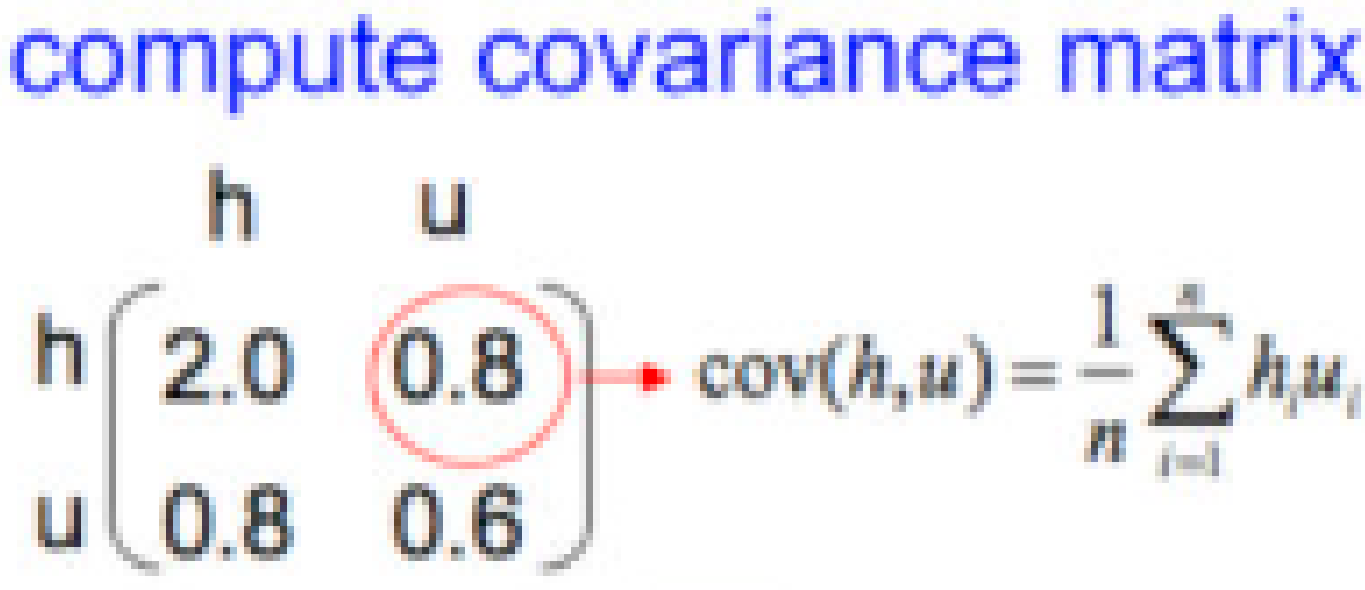
Steps for PCA Algorithm

Standardize the data

If the variables in the dataset have different scales, it is important to standardize them to have zero mean and unit variance.

Calculate the covariance matrix or correlation matrix

Depending on the specific goals of the analysis, you can choose to calculate either the covariance matrix or the correlation matrix.



Compute the eigenvectors and eigenvalues

The next step is to calculate the eigenvectors and eigenvalues of the covariance or correlation matrix. These represent the directions (eigenvectors) and the amount of variance (eigenvalues) explained by each principal component.

eigenvectors + eigenvalues

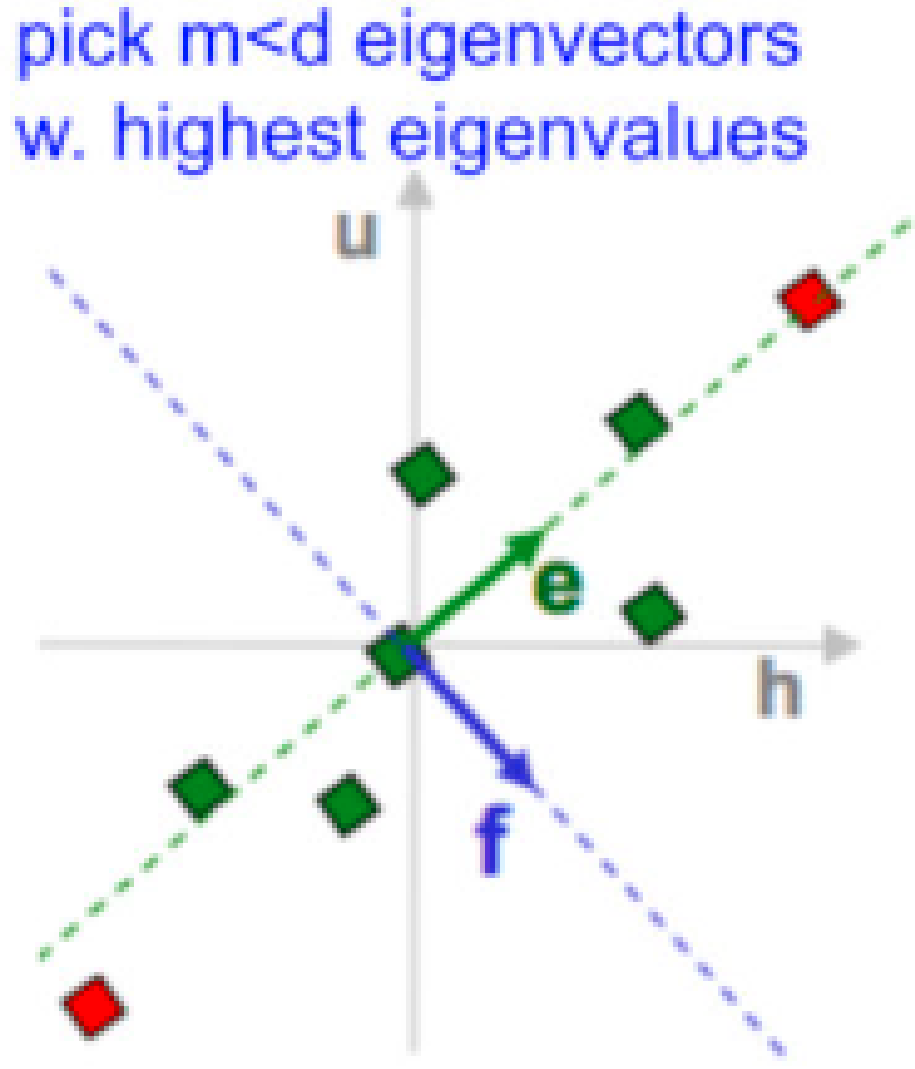
$$\begin{pmatrix} 2.0 & 0.8 \\ 0.8 & 0.6 \end{pmatrix} \begin{pmatrix} e_h \\ e_u \end{pmatrix} = \lambda_e \begin{pmatrix} e_h \\ e_u \end{pmatrix}$$

$$\begin{pmatrix} 2.0 & 0.8 \\ 0.8 & 0.6 \end{pmatrix} \begin{pmatrix} f_h \\ f_u \end{pmatrix} = \lambda_f \begin{pmatrix} f_h \\ f_u \end{pmatrix}$$

```
eig(cov(data))
```

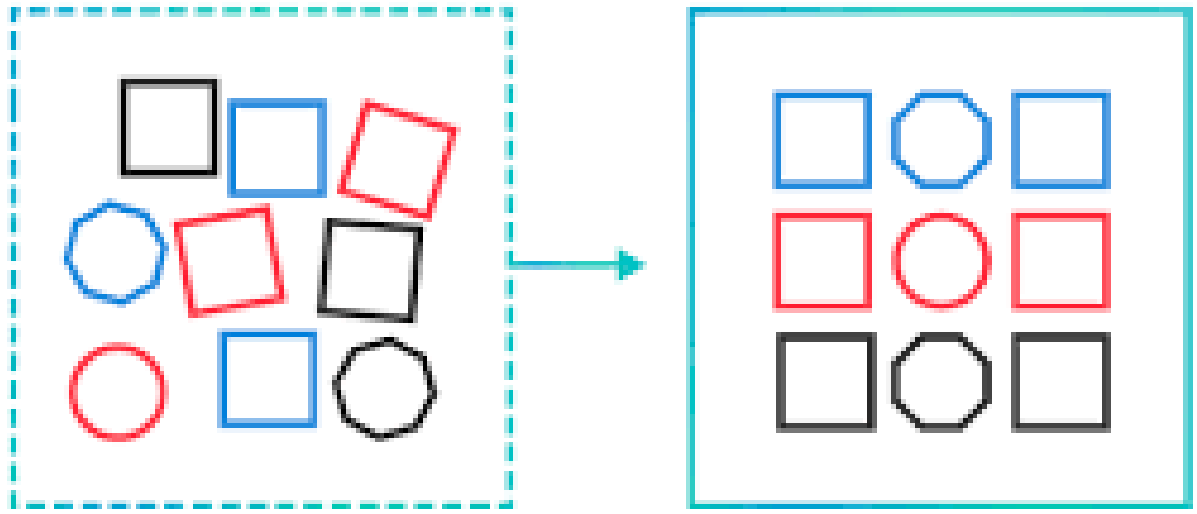
Select the principal components

Sort the eigenvalues in descending order and select the top-k eigenvectors corresponding to the highest eigenvalues. These are the principal components that capture the most significant variability in the data.



Transform the data

Take the selected eigenvectors and transform the standardized data onto the new feature subspace. This is achieved by computing the dot product of the data with the eigenvectors to obtain the new set of principal components.



Analytics Vidhya

Advantages of PCA

Dimensionality reduction: simplifies complex data by identifying the most significant features, aiding visualization and analysis in high-dimensional datasets.

Feature Extraction: Generates new, meaningful features from correlated or noisy data, enhancing interpretability and insights.

Data visualization: Enables visualization of high-dimensional data, revealing hidden patterns and clusters in a lower-dimensional space.

Noise Reduction: By locating the underlying signal or pattern in the data, PCA can also be used to lessen the impacts of noise or measurement errors in the data.

Multicollinearity: When two or more variables are strongly correlated, there is multicollinearity in the data. PCA can lessen the impacts of multicollinearity on the analysis by identifying the most crucial features or components.

Analytics Vidhya

Disadvantages of PCA

Interpretability: PCA simplifies data and reveals patterns, but interpretability of resulting components can be challenging compared to the original features.

Information loss: PCA reduces dimensionality, but interpretability of components varies, risking information loss if crucial features are not represented.

Outliers: Data anomalies can distort PCA outcomes, hindering the identification of crucial characteristics due to their impact on the covariance matrix.

Scaling: Scaling and centralization ensure accurate PCA representation; without it, principal components may be misleading.

Computing complexity: Computing eigenvectors and eigenvalues for large datasets can be computationally expensive, limiting the scalability and applicability of PCA in certain cases.

Analytics Vidhya

That's a wrap. Was this post Helpful?

Follow us for more!

Analytics Vidhya