Principal Component Analysis (PCA)

Learn How to Manage Data Dimensionality

Without Losing Information

Contents

The Need For Data Reduction

Explosion in information collection techniques and data availability has resulted in data scientists facing a unique requirement – Need for reducing data.

But how can more information be a problem?

Not all data fields are useful.

Variables can be highly correlated or at times redundant Most data analysis algorithms work on columns(variables).

Large number of variables may result in slowing down of the process.

- Does this mean simply removing data? Definitely not.
- Key rule is - Lose data but not information !

What Is High Dimensional Data ?

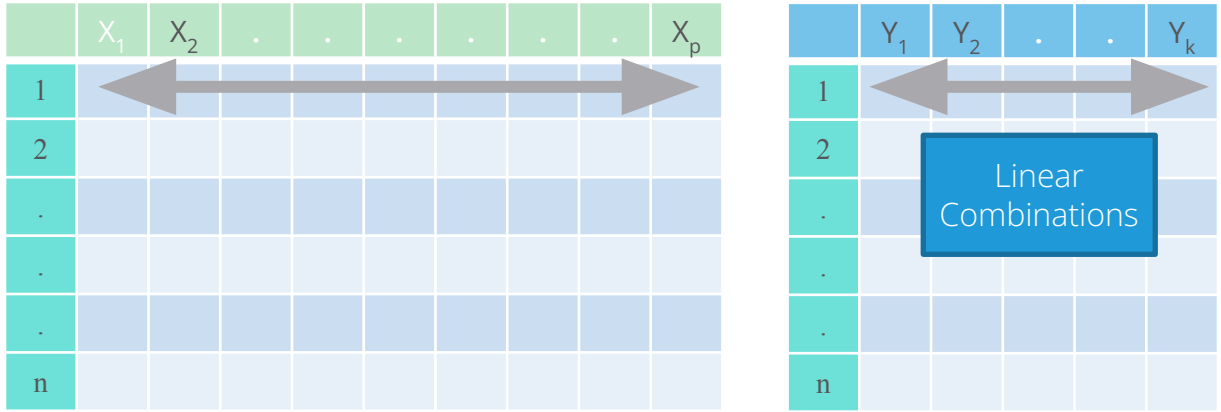
Simply put, high dimensional data can either be data for several observations with several variables or data with several layers of information.

- **Majority of the real world analytics** - in the fields of medical sciences, ecology, biology, finance, etc. - deals with datasets having a few dozens to hundreds and at times thousands of variables. Also the number of cases can go up several million depending upon the industry.

- Consider data maintained by a bank. For each customer, there can be numerous data which can be categorized into transaction data, customer demographics and customer call centre data.
- Further, there can be millions of such customers.

Data Reduction

- Summary of data with p variables by a smaller set of (k) derived variables.
- These k derived variables are **linear combinations of original p variables.**



- In short, **n * p** matrix is **reduced to n * k** matrix.

Variation and Correlation-Key Data Features

- Variation in the data is nothing but the information that the data gives.
- Variables in the data can be highly correlated
- Example: Employee satisfaction survey asks the following questions:

  - Do you think you are rewarded regularly for your work?
  - Do you think the system monitoring your performance is flawed?
  - Do you get timely encouragement from your supervisors regarding your work?

- The above three questions measure the robustness of employee work review. Having three questions to represent a common area can be considered redundant.

- Reduction is justified when there is such Correlation present in the data.

Principal Component Analysis (PCA)

- PCA is probably the most widely-used and well-known of the “Standard” multivariate methods.

  - PCA uses correlation structure of original p variables and derives p linear combinations which are uncorrelated.
  - Each Principal Component provides unique information about the data.
  - Although ‘p' principal components are derived, first 'k' principal components are expected to explain most of the variability in the data.

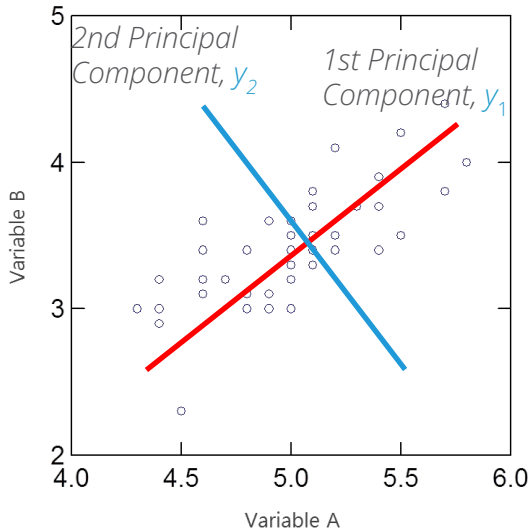
- The method was invented by **Pearson** (1901) and **Hotelling** (1933).
- The method was first applied in ecology by Goodall (1954) under the name “Factor Analysis”. (“Principal Factor Analysis” is a synonym of PCA).

PCA-General Concept

Consider 2 variables A & B (Assume that A & B are features of an object)

A & B have approximately same variance and are highly correlated

Fig. 01



- Now suppose we pass a vector through the scatter points such that it is a good linear fit to the data points. This could be considered a new feature of the object which is a linear combination of A & B.

- We then plot a second line which is perpendicular to the vector, such that both lines pass through the centroid of the data. This becomes our second linear combination.

PCA – General Concept

From p original variables: $x_1, x_2, \ldots, x_p$, derive p new variables $y_1, y_2, \ldots, y_p$

$$
\begin{aligned}
y_1 &= a_{11}x_1 + a_{12}x_2 + \ldots + a_{1p}x_p \\
y_2 &= a_{21}x_1 + a_{22}x_2 + \ldots + a_{2p}x_p \\
&\vdots \\
y_p &= a_{p1}x_1 + a_{p2}x_2 + \ldots + a_{pp}x_p
\end{aligned}
$$

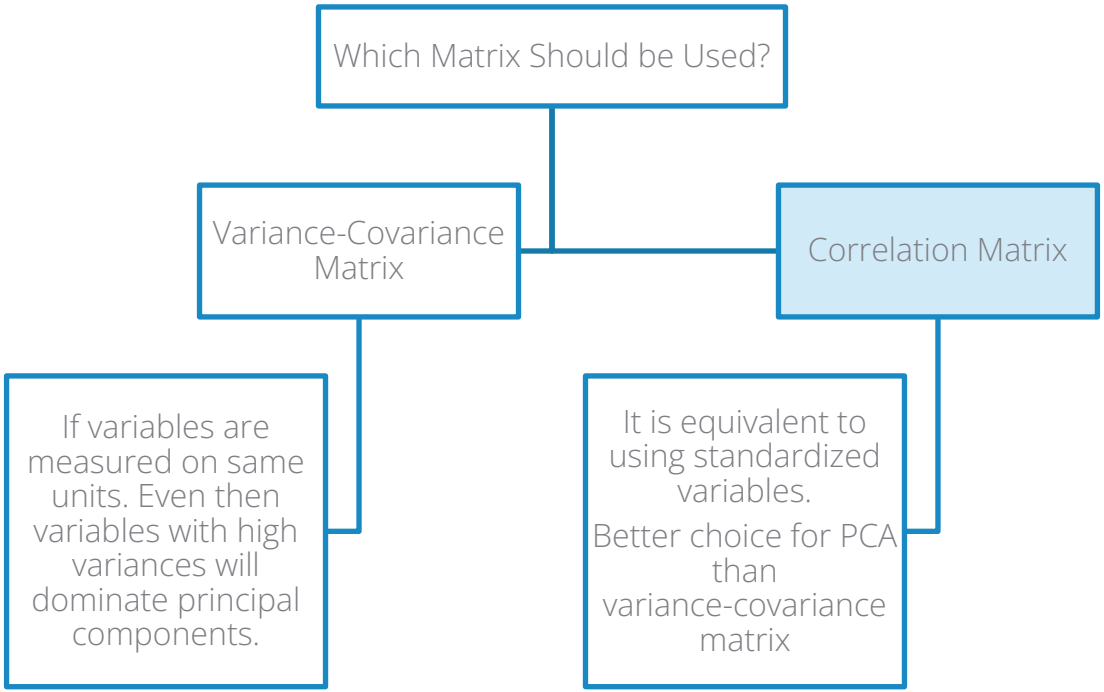
y_k's are Principal Components

Although p principal components are derived, data reduction is achieved with first k principal components.

Properties of Principal Components :

- y_k's are uncorrelated (Orthogonal).
- y_1 explains as much as possible of original variance in data set.
- y_2 explains as much as possible of remaining variance. And so on.

PCA – Choice of Matrix Analysis



* Correlation matrix of original variables = Variance-covariance matrix of standardized variables.

Principal Components – Definition

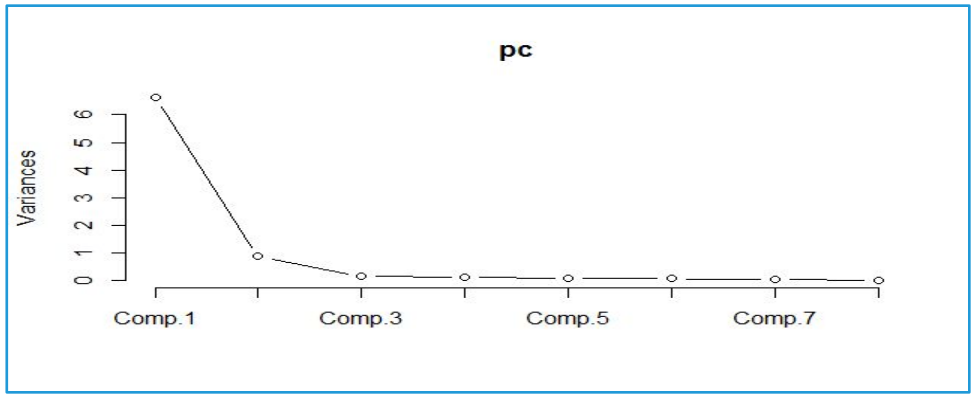
Component	Definition
First Component	A linear combination $a_{11}x_1 + a_{12}x_2 + ... + a_{1p}x_p$ which has maximum variance among all possible linear combinations, subject to $a_{11}^2 + a_{12}^2 + ... + a_{1p}^2 = 1$.
Second Component	A linear combination $a_{21}x_1 + a_{22}x_2 + ... + a_{2p}x_p$ which has maximum variance among all remaining linear combinations, subject to $a_{21}^2 + a_{22}^2 + ... + a_{2p}^2 = 1$ and zero correlation with first principal component.
Third Component *And so on..*	Similarly third Principal Component is obtained which will be uncorrelated with first and second principal components.

How Many Principal Components to Retain

P principal components are derived but data reduction is achieved with first k PC's.

Retain principal components associated with Eigen Value more than '1'.
Variance of PCs is equal to associated Eigen Value. This is known as "**Kaiser Criterion**".

Look at a **Scree plot** and check for "**Elbow**" to determine the correct number of PCs to use.
A scree plot shows how much variation each PC captures from the data. The y axis is eigenvalues, which essentially stand for the variation.
Example of **Scree Plot** below shows that, **first principal component** is **sufficient** to explain variation in the data.



Quick Recap

Data Reduction and PCA	• In the age of big data, data reduction is necessary for analysis • Principal Component Analysis is most popular data reduction method.
PCA – General Approach	• Principal Components are linear combinations of original variables • These components are uncorrelated • Retention of components is based on Eigenvalues.