Principal Component Analysis (PCA)

It is an easy, nonparametric method of collecting related information from unclear datasets. Statistical process that uses an orthogonal transformation to change a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variable called principal components. PCA was invented in 1901 by Karl Pearson who developed it independently and was named by Harold Hoteling in the 1930s. First applied in ecology by Goodall (1954) under the name "factor analysis" It is a way of recognizing patterns in data, and conveying the data in such a way as to compare them. Since patterns in data can be difficult to find in data of high dimension, where the comfort of graphical representation is not available, PCA is a great tool for analyzing data. The other major advantage of PCA is that once you have established these patterns in the data, and you shrink the data, i.e., by reducing the number of dimensions, without losing much information.

Variables are treated similarly and they are not split into dependent and independent variables. In understandable terms, PCA modifies the original interrelated variables into a new set of uncorrelated variables known as Principal Components. An advantage of principal components to researchers is that the difficulty in interpretation that can be caused due to large number of interrelated variables can be decreased by utilizing only the first few principal components that explain a large part of the total variation. PCA can be utilized to test for normality. If the principal components are not normally distributed, then the original data weren't either.

A primary concept of PCA is to decrease the number of variables or reduce dimensionality. A significant choice that the researcher must make when using PCA is to decide the number of principal components to utilize. This decision has no fixed rules, and the decision may seem subjective at times.

2D example

First, consider a dataset in only two dimensions, like (height, weight). This dataset can be plotted as points in a plane. But if we want to tease out variation, PCA finds a new coordinate system in which every point has a new (x,y) value. The axes don't actually mean anything physical; they're combinations of height and weight called "principal components" that are chosen to give one axes lots of variation

PCA Toy Example: Consider the following 3D points

$$\begin{bmatrix}1\\2\\3\end{bmatrix} \quad \begin{bmatrix}2\\4\\6\end{bmatrix} \quad \begin{bmatrix}4\\8\\12\end{bmatrix} \quad \begin{bmatrix}3\\6\\9\end{bmatrix} \quad \begin{bmatrix}5\\10\\15\end{bmatrix} \quad \begin{bmatrix}6\\12\\18\end{bmatrix}$$

If each component is stored in a byte, we need 18 = 3 x 6 bytes

Looking closer, we can see that all the points are related geometrically: they are all the same point, scaled by a factor:

$$\begin{bmatrix}1\\2\\3\end{bmatrix} = 1 * \begin{bmatrix}1\\2\\3\end{bmatrix} \quad \begin{bmatrix}2\\4\\6\end{bmatrix} = 2 * \begin{bmatrix}1\\2\\3\end{bmatrix} \quad \begin{bmatrix}4\\8\\12\end{bmatrix} = 4 * \begin{bmatrix}1\\2\\3\end{bmatrix}$$

$$\begin{bmatrix}3\\6\\9\end{bmatrix} = 3 * \begin{bmatrix}1\\2\\3\end{bmatrix} \quad \begin{bmatrix}5\\10\\15\end{bmatrix} = 5 * \begin{bmatrix}1\\2\\3\end{bmatrix} \quad \begin{bmatrix}6\\12\\18\end{bmatrix} = 6 * \begin{bmatrix}1\\2\\3\end{bmatrix}$$

They can be stored using only 9 bytes (50% savings!): Store one point (3 bytes) + the multiplying constants (6 bytes)

Role of PCA: PCA captures major (principal) variability present in the data set and neglects smaller variability. It helps in finding Eigen Values and Eigen Vectors. •

Significant Eigen Values and Eigen Vectors are taken for determining PCs. • PCA forms new coordinate system defined by the significant Eigen vectors. (new coordinates will have lower dimensions) Mapping data to the new space.

Principle of PCA: It works on Linear projection method to reduce the number of variables. Transfer a set of correlated variables into a new set of uncorrelated variables. Map the data into a space of lower dimensionality. PCA rotates existing axes to new positions in the space defined by original variable

Importance: With data of high dimensions, where graphical representation is difficult, PCA is a strong multivariate statistical tool for analyzing data and discovering patterns in it. Mapping of data compression is also feasible using PCA.

Interpretation: PCA plots should be understood by looking at points relative to the origin. Points that are in similar directions or closed positions are positively correlated. Points that are on opposite sides of the origin are correlated negatively. Points that are far from the origin or more perpendicular less correlated.

Application: Sensory evaluation and quality control in post-harvest technologies. 2. Applicable to Stock assessments in fisheries resources. 3. Applicable in genetics (Genomics) and related experiments. 4. To assess the nutritional status of fishes and shellfishes with respect to biochemical compositions, growth parameters and dietary aspects (Nutrients). 5. Monitoring environmental studies and other parameters in aquaculture.

Advantages: Both objective and subjective attributes can be used. It can be done precisely (only) with the help of Statistical software. Direct inputs from treatments. There is flexibility in naming and using dimensions. PCA is helpful for discovering new, more informative and uncorrelated components. It reduces dimensionality by declining lower variance components.

Disadvantages: Usefulness depends on the researchers' ability to develop a complete and accurate set of attributes. If significant characteristics are missed, accuracy of the procedure is reduced accordingly. Naming of the factors (independent variables) can be difficult - multiple attributes can be highly correlated with no sensible reason. If the observed variables are completely unrelated, PCA analysis is unable to produce a relevant pattern.

PCA can be achieved via a series of steps

Step By Step Computation Of PCA

The below steps need to be followed to perform dimensionality reduction using PCA:

1. Standardization of the data - Standardization is all about scaling your data in such a way that all the variables and their values lie within a similar range. Standardization is carried out by subtracting each value in the data from the mean and dividing it by the overall deviation in the data set.
2. Computing the covariance matrix - A covariance matrix expresses the correlation between the different variables in the data set. The covariance value denotes how co-dependent two variables are with respect to each other. If the covariance value is negative, it denotes the respective variables are indirectly proportional to each other. A positive covariance denotes that the respective variables are directly proportional to each other
3. Calculating the eigenvectors and eigenvalues - Eigenvectors and eigenvalues are the mathematical constructs that must be computed from the covariance matrix in order to determine the principal components of the data set.
4. Computing the Principal Components - Once we have computed the Eigenvectors and eigenvalues, all we have to do is order them in the descending order, where the eigenvector with the highest eigenvalue is the most significant and thus forms the first principal component. The principal components of lesser significances can thus be removed in order to reduce

the dimensions of the data. The final step in computing the Principal Components is to form a matrix known as the feature matrix that contains all the significant data variables that possess maximum information about the data.

5. Reducing the dimensions of the data set - The last step in performing PCA is to re-arrange the original data with the final principal components which represent the maximum and the most significant information of the data set.