Principle Component Analysis

Dr. Jyotismita Chaki

Introduction to PCA

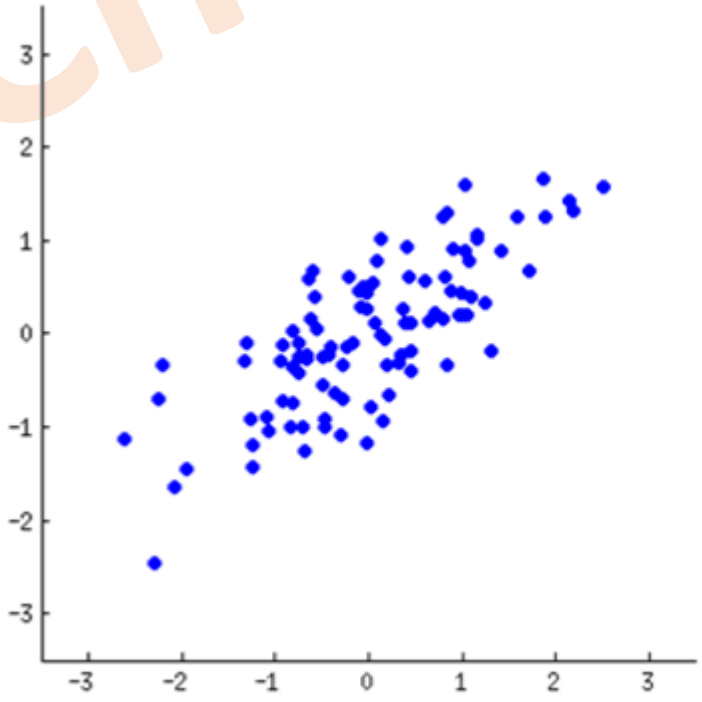
- PCA is particularly handy when you're working with "wide" data sets. But why is that?
- In such cases, where many variables are present, you cannot easily plot the data in its raw format, making it difficult to get a sense of the trends present within.
- PCA allows you to see the overall "shape" of the data, identifying which samples are similar to one another and which are very different.
- This can enable us to identify groups of samples that are similar and work out which variables make one group different from another.
- The basics of PCA are as follows: you take a dataset with many variables, and you simplify that dataset by turning your original variables into a smaller number of "Principal Components".

Introduction to PCA

- Principal Components are the underlying structure in the data.
- They are the directions where there is the most variance, the directions where the data is most spread out.
- This means that we try to find the straight line that best spreads the data out when it is projected along it.
- This is the first principal component, the straight line that shows the most substantial variance in the data.
- PCA is a type of linear transformation on a given data set that has values for a certain number of variables (coordinates) for a certain amount of spaces.
- This linear transformation fits this dataset to a new coordinate system in such a way that the most significant variance is found on the first coordinate, and each subsequent coordinate is orthogonal to the last and has a lesser variance.
- In this way, you transform a set of x correlated variables over y samples to a set of p uncorrelated principal components over the same samples.

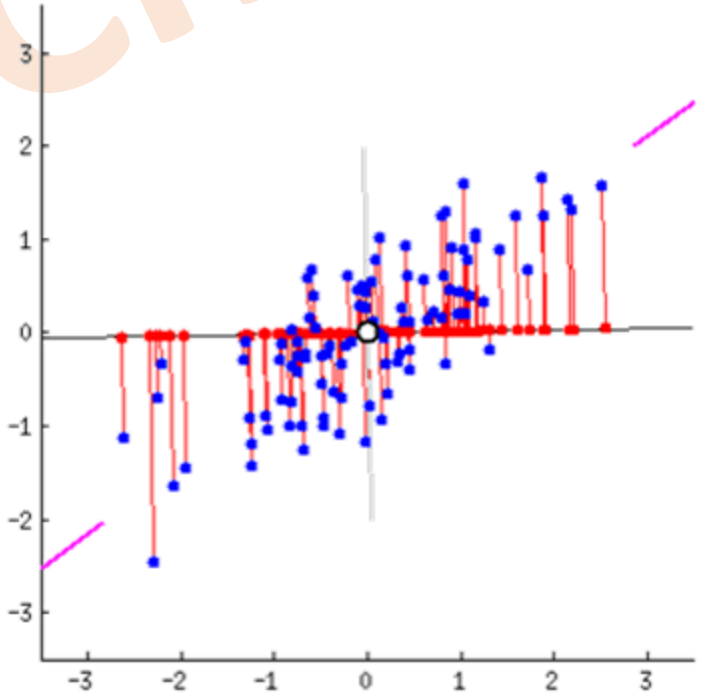
Introduction to PCA

- Let us pick two wine characteristics, perhaps wine darkness and alcohol content.
- Here is what a scatter plot of different wines could look like.
- Each dot in this "wine cloud" shows one particular wine. You see that the two properties (x and y on this figure) are correlated.
- A new property can be constructed by drawing a line through the center of this wine cloud and projecting all points onto this line.
- This new property will be given by a linear combination w1x + w2y where each line corresponds to some particular values of w1 and w2.



Introduction to PCA

- Here is what these projections look like for different lines (red dots are projections of the blue dots).
- PCA will find the "best" line according to two different criteria of what is the "best".
- First, the variation of values along this line should be maximal.
- Pay attention to how the "spread" (we call it "variance") of the red dots changes while the line rotates (you can see when it reaches maximum).
- Second, if we reconstruct the original two characteristics (position of a blue dot) from the new one (position of a red dot), the reconstruction error will be given by the length of the connecting red line (you can see when the total length reaches minimum).
- Observe how the length of these red lines changes while the line rotates.
- If you stare at this animation for some time, you will notice that "the maximum variance" and "the minimum error" are reached at the same time, namely when the line points to the magenta ticks.
- This line corresponds to the new wine property that will be constructed by PCA.



Introduction to PCA

- Where many variables correlate with one another, they will all contribute strongly to the same principal component.
- Each principal component sums up a certain percentage of the total variation in the dataset.
- Where your initial variables are strongly correlated with one another, you will be able to approximate most of the complexity in your dataset with just a few principal components.
- As you add more principal components, you summarize more and more of the original dataset.
- Adding additional components makes your estimate of the total dataset more accurate, but also more unwieldy.

Introduction to PCA: Eigenvalues and Eigenvectors

- Eigenvectors, and eigenvalues come in pairs: every eigenvector has a corresponding eigenvalue.
- An eigenvector is a direction, such as "vertical" or "45 degrees", while an eigenvalue is a number telling you how much variance there is in the data in that direction.
- The eigenvector with the highest eigenvalue is, therefore, the first principal component.
- The number of eigenvalues and eigenvectors that exits is equal to the number of dimensions the data set has. So if the data set was two-dimensional, that means that there are two eigenvectors and eigenvalues.
- Similarly, you'd find three pairs in a three-dimensional data set.
- We can reframe a dataset in terms of these eigenvectors and eigenvalues without changing the underlying information.
- Note that reframing a dataset regarding a set of eigenvalues and eigenvectors does not entail changing the data itself, you’re just looking at it from a different angle, which should represent the data better.