Principal Component Analysis

Background

Section 10.2 of

Introduction to Statistical Learning

By Gareth James, et al.

Background

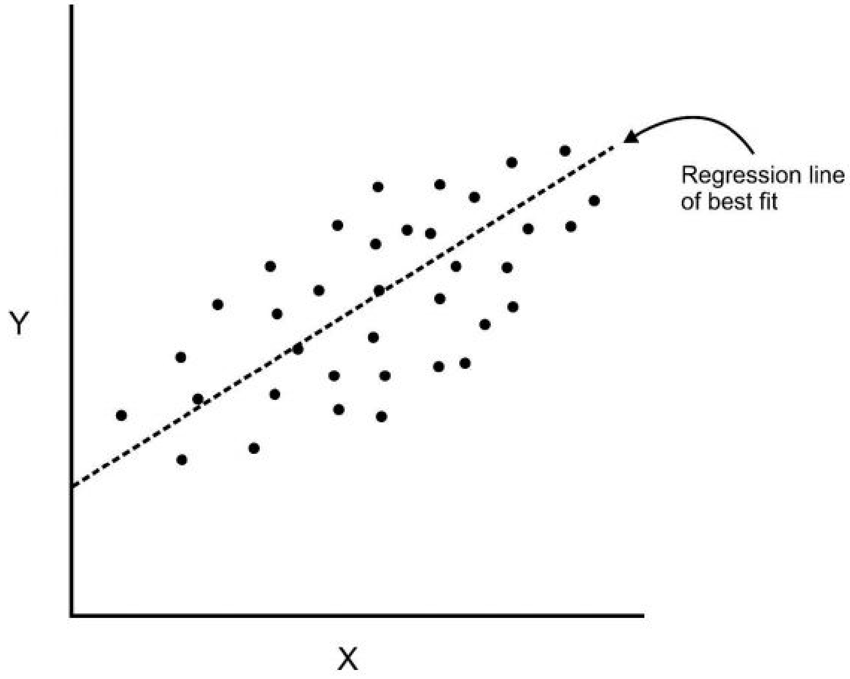
- Let’s discuss the basic idea behind principal component analysis.
- It is an unsupervised statistical technique used to examine the interrelations among a set of variables in order to identify the underlying structure of those variables.
- It is also known sometimes as a general **factor analysis.**

Background

- Where regression determines a line of best fit to a data set, factor analysis determines several orthogonal lines of best fit to the data set.
- Orthogonal means “at right angles”.
    - Actually the lines are perpendicular to each other in n-dimensional space.
- n-Dimensional Space is the variable sample space.
    - There are as many dimensions as there are variables, so in a data set with 4 variables the sample space is 4-dimensional.

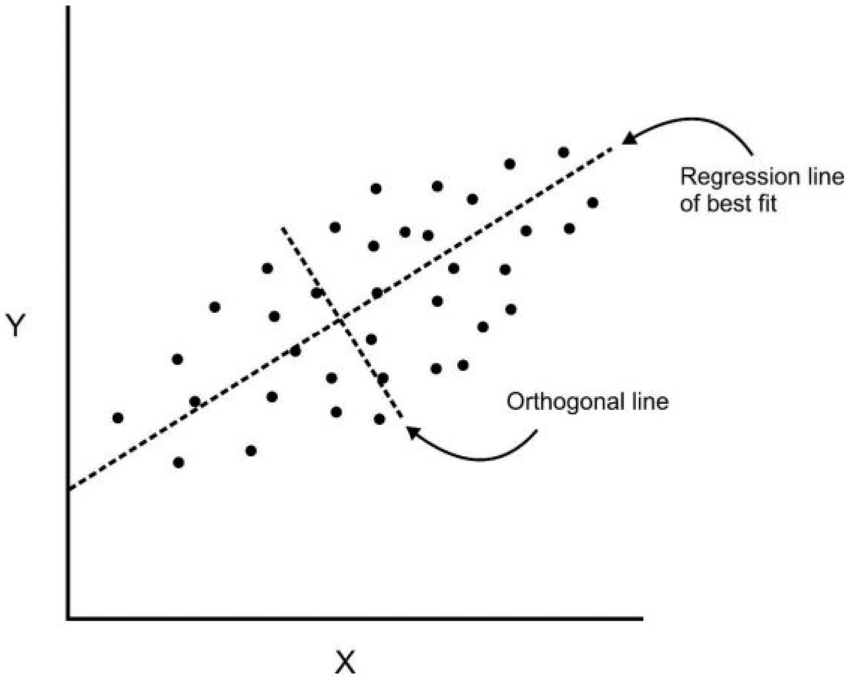
Background

- Here we have some data plotted along two features, x and y.



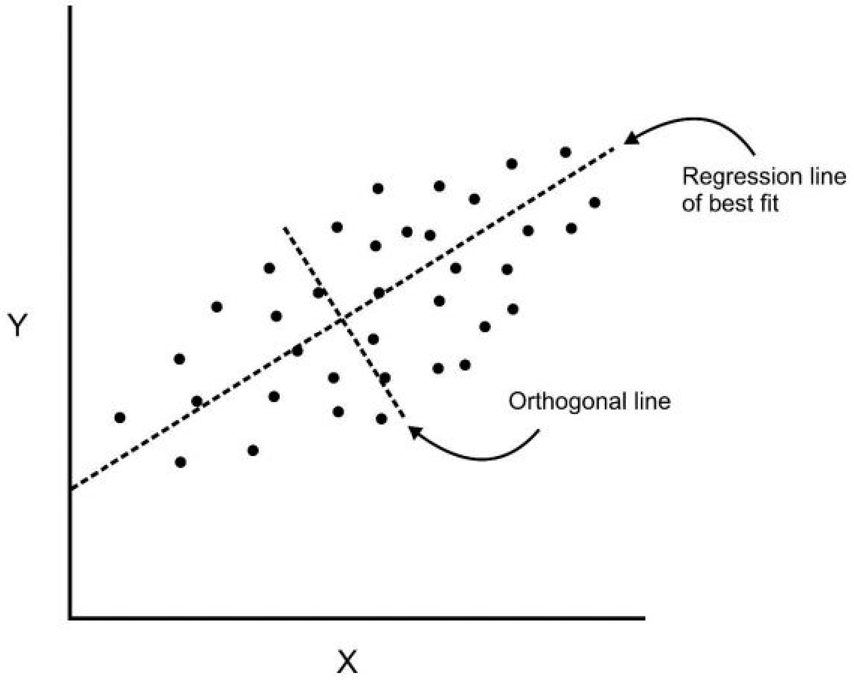
Background

- We can add an orthogonal line.
- Now we can begin to understand the components!



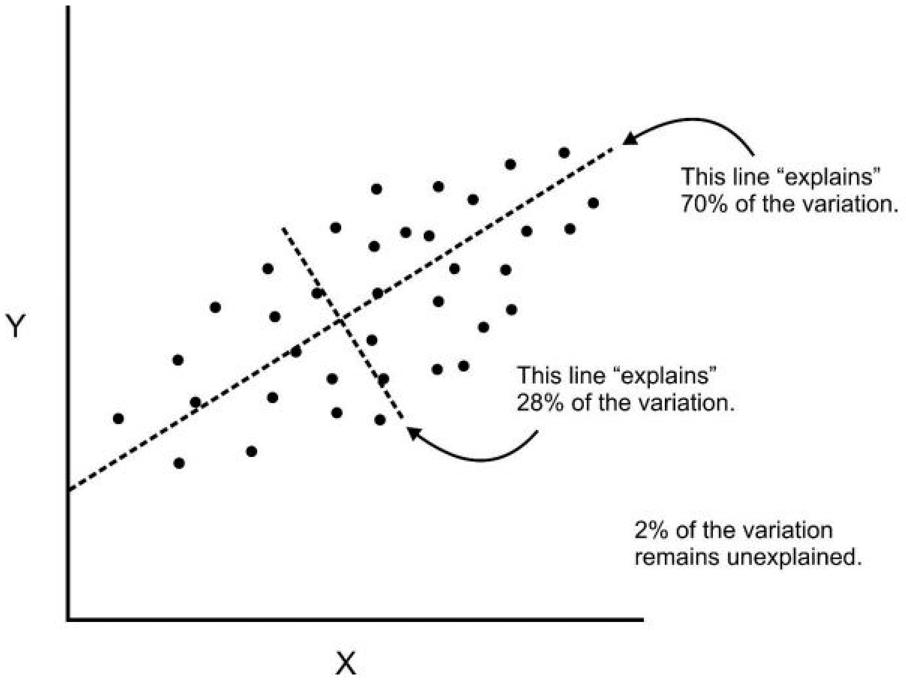
Background

- Components are a linear transformation that chooses a variable system for the data set such that the greatest variance of the data set comes to lie on the first axis



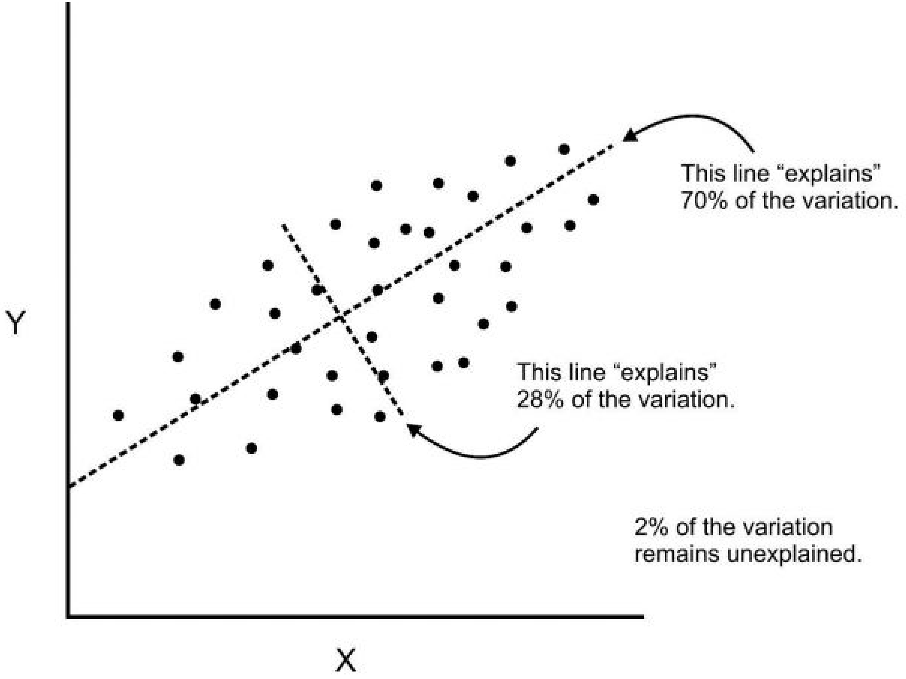
Background

- The second greatest variance on the second axis, and so on …
- This process allows us to reduce the number of variables used in an analysis.


Y
X
This line "explains" 70% of the variation.
This line "explains" 28% of the variation.
2% of the variation remains unexplained.

Background

- Note that components are uncorrelated, since in the sample space they are orthogonal to each other.



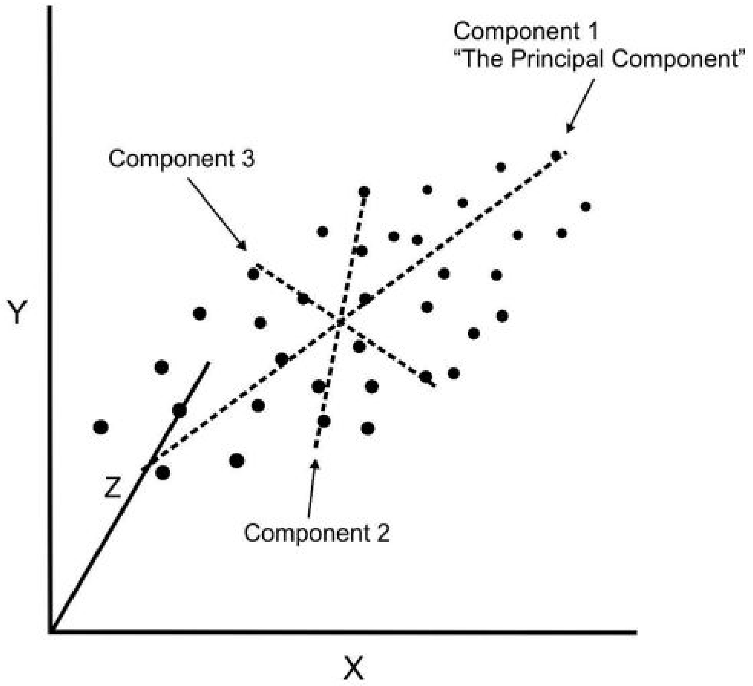
This line "explains" 70% of the variation.

This line "explains" 28% of the variation.

2% of the variation remains unexplained.

Background

- We can continue this analysis into higher dimensions



Background

- If we use this technique on a data set with a large number of variables, we can compress the amount of explained variation to just a few components.
- The most challenging part of PCA is interpreting the components.

Background

- For our work with Python, we’ll walk through an example of how to perform PCA with scikit learn.
- We usually want to standardize our data by some scale for PCA, so we’ll cover how to do this as well.
- Since this algorithm is used usually for analysis of data and not a fully deployable model, there won’t be a portfolio project for this topic.

Example with Python