Principal Component analysis

Dimesionality Reduction

Motivations

What are dimensions?

- feature space

· We want to reduce this for

Time Complexity

1. Efficiency
2. Visualization

Interpretation

Is it different than feature selection?

$a_1 \quad a_2$ —— a_m (original feature)

z_1 . —— z_p (transformed feature space)

X Matrix

$$\begin{array}{cccc} x_1 & & & x_m \\ x_{11} & \cdots & & x_{1m} \\ x_{21} & \cdots & & x_{2m} \\ x_n & \cdots & & x_{nm} \end{array}$$

* In PCA we want to create new variables such that they are linear combos of the original variables

$$Z_1 = \phi_1 \times [X_1] + \phi_2 [X_2] + \ldots + \phi_m [X_m]$$

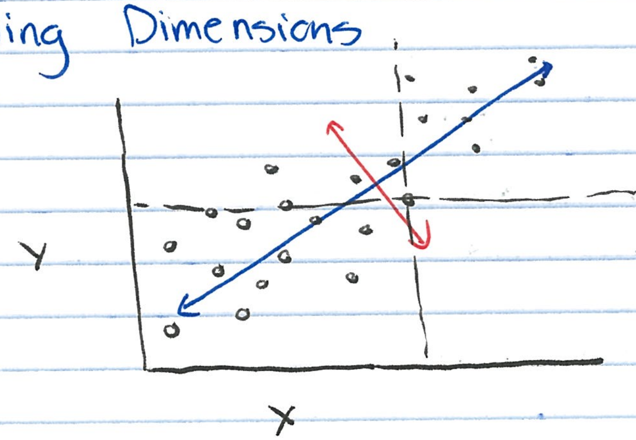
we want to create such p features

Property of new features

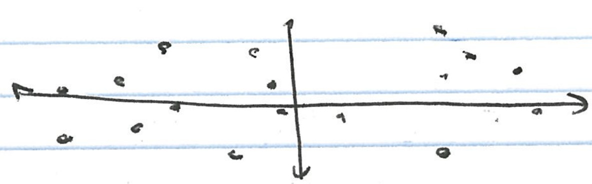
- we want to preserve variability of the data
- we want to find linear combinations of original features such that they explain maximum variability
- such components are called principal components
- these components are uncorrelated to each other & thus orthogonal
- first principal component explains maximum variability, second explains 2nd highest maximum variability

Finding Dimensions

- instead of using original axis, we rotate axis along it's variability



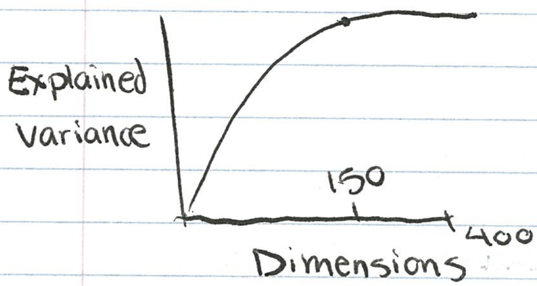
$\Rightarrow$



Properties

- to fully capture the variability we need m principal components

So where is the reduction?



- you may be able to explain variability using 150 dimensions rather than 400

Mathematical Intuition

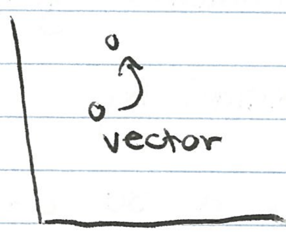
Covariance Matrix - represents variability of the data which is a square matrix

eigenvalue

$$Av = \lambda v$$

matrix

eigenvalue



- you can apply different transformation on the point & it changes

- multiply original vector by transformation matrix to get new vector

FIVE STAR.

EA

Eigenvalue & Eigenvector

- such a spatial vector such that when you transform the vector it doesn't change the vector but rather scales it by λ

Understanding with Iris

x_1 x_2 x_3 x_4 class

- look at variability of data given by covariance

1. find eigenvalues & eigenvalue
   - tells you how much of the variability is explained by eigenvectors
   - you can see how much of the principal components you need
2. identify where data is most explained by
3. Calculate first principal component & so on

When to use PCA

1. When you want to reduce number of variables but aren't able to identify variables to complety remove from consideration

2. You want to ensure variables are indep

3. Are you comfortable making indep less interpretable

Remarks

- as this deals with variance, scale matters
  ex: age vs salary, thus Standardize data

- Use correlation matrix

FIVE STAR