Principal Component Analysis

Conor McCormick

Advisor: Dr. Muhammad Islam

What is Principal Component Analysis?

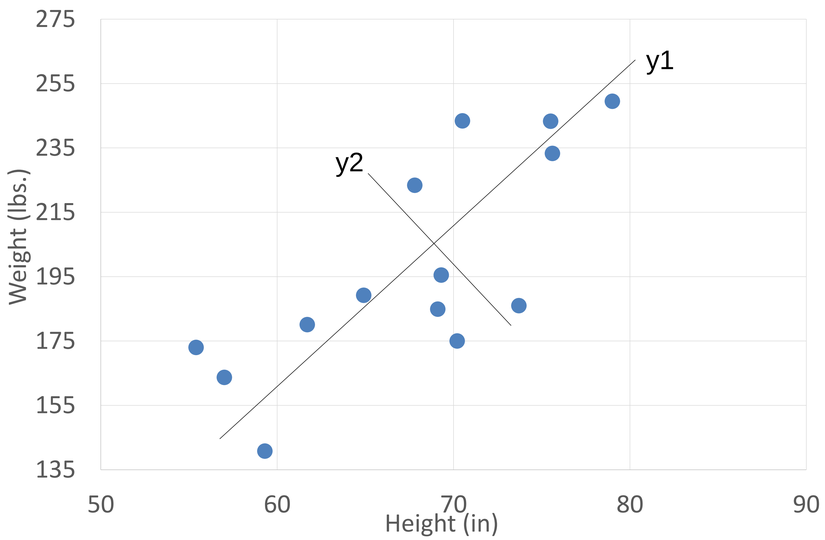
Principal Component Analysis (PCA) a mathematical method that allows us to more easily understand large multivariate sets of data.

Goal of PCA

The goal of using PCA is to reduce the amount of dimensions in a set of data with high dimension without a loss of significant information to help better understand the correlations between the variables.

A Simple Example

A random sample of 14 young adults males is taken. All of them are asked for their height (in) and weight (lbs.). There are two ways to do PCA, (1) using the covariance matrix and (2) using singular value decomposition. For the simplicity of this example, I will be using the first method.



Original Data

Height	Weight
69.3	195.5
70.2	175
61.7	180.1
75.5	243.3
69.1	184.9
57	163.7
73.7	186
67.8	223.4
59.3	140.8
75.6	233.3
79	249.5
55.4	173
64.9	189.2
70.5	243.4

Adjusted Data

Height	Weight
1.51	-3.15
2.41	-23.65
-6.09	-18.55
7.71	44.65
1.31	-13.75
-10.79	-34.95
5.91	-12.65
0.01	24.75
-8.49	-57.85
7.81	34.65
11.21	50.85
-12.39	-25.65
-2.89	-9.45
2.71	44.75

1. Create Covariance Matrix

- First find the mean of the data, in our case this is 67.79 in for height and 198.65 for weight.
- Next subtract the mean from the data, call this matrix X.
- Then the Covariance Matrix, S, is equal to $(1/N\text{-}1)X^TX$, where N is the sample size.

$$S = \begin{bmatrix} 48.97 & 177.02 \\ 177.02 & 1069.49 \end{bmatrix}$$

2. Find Eigenvalues and Eigenvectors

- Find the eigenvalues of eigenvectors of S.
- Note: these must be unit eigenvectors, meaning that the length is 1.

$\lambda_1 = 1099.32$ $\qquad$ $\lambda_2 = 19.14$

$$u_1 = \begin{bmatrix} 0.166 \\ 0.986 \end{bmatrix} \qquad u_2 = \begin{bmatrix} -0.986 \\ 0.166 \end{bmatrix}$$

3. Find the Principal Components

- Each eigenvalue corresponds to a principal component, going in order of size. Meaning the first principal component (PC) corresponds with the largest eigenvalue.
- Each principal component is a new variable created using the corresponding unit eigenvector. For example the first principal component is $y_1 = 0.116x_1 + 0.986x_2$ where x_1 and x_2 are the height and weight respectfully.

4. Reduce Dimensions and Analyze

- The final step is to see how many of these PC's are actually needed to analyze the data.
- We need to use as many PC's as we can to get to at least 90% of the variance of the data.

$$\frac{1099.32}{1118.46} = 98.3\% \qquad \frac{19.14}{1118.46} = 1.71\%$$

- Thus we can conclude that the first PC is an accurate representation of the data and that our data is essentially one-dimensional.

Applications in the Real World

- Demographic studies for businesses
- Predicting the rise and fall of stocks based upon the current economy
- Scientific research, finding the what factors effect others.
- Image processing, such as object or facial recognition.