Principal Component Analysis (PCA)

PCA

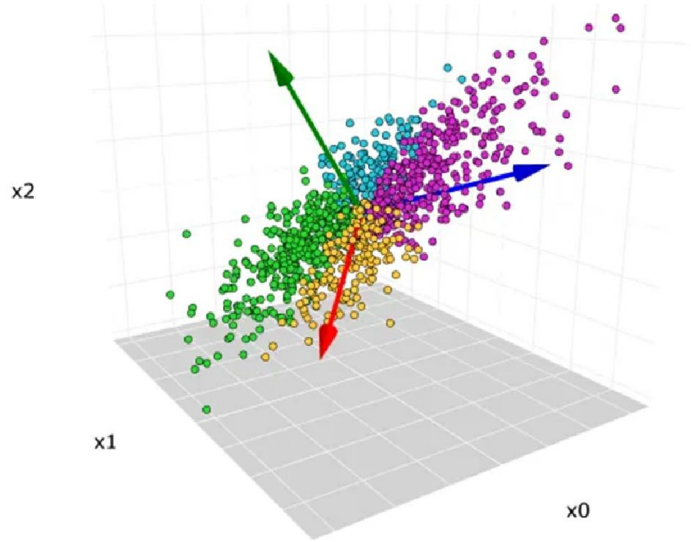
- **Unsupervised** learning method

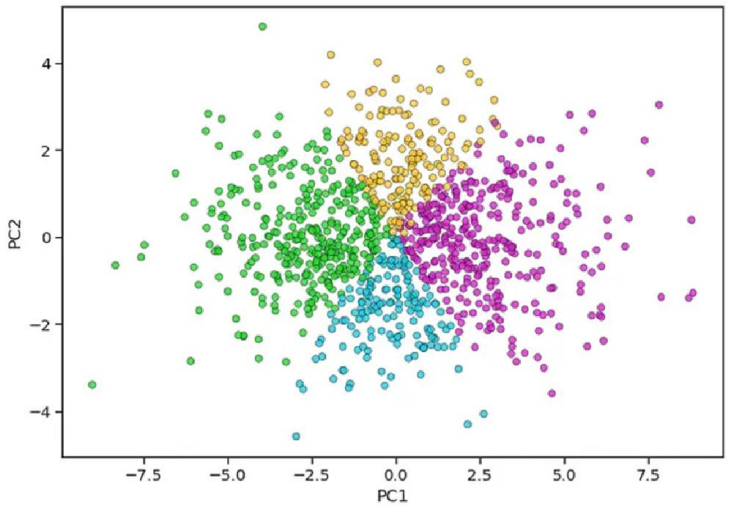
- Visualization and Dimensionality Reduction
  - Curse of dimensionality

- Extremely useful when working with data sets that have a lot of features.
  - image processing, compression, genome research etc.

PCA (cont.)

- Transforms high-dimensions data into lower-dimensions while retaining as much information as possible.





How PCA Works

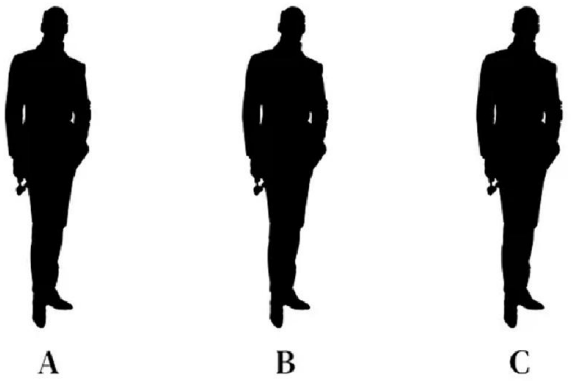
- Two-step Process
  - Understand which part of our data is important. How?
  - Mathematically quantify the amount of information embedded within the data. How?
- Answer: through the *Variance*
- The greater the variance, the more the information (and vice versa).

$$S^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

Variance vs. Information

Person	Height
Alex	145
Ben	160
Chris	185

Person	Height
Daniel	172
Elsa	171
Fernandez	173



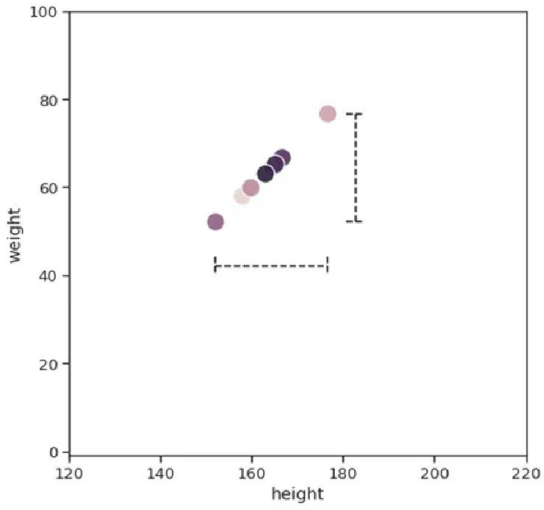
Variance vs. Information (cont.)

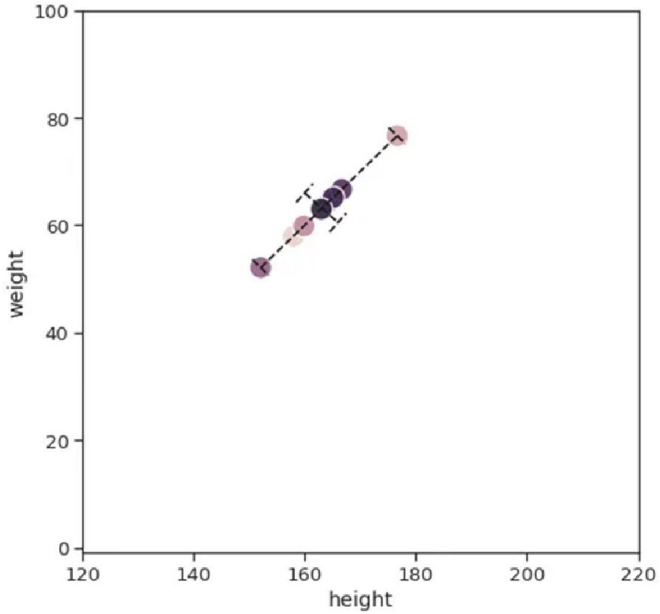
Person	Height (cm)	Weight (kg)
Alex	145	68
Ben	160	67
Chris	185	69

- The weight variance is so small (little information), so it doesn't help differentiate our friends at all.
  - Still rely mostly on height to make guesses (variable with higher variance).
  - Reducing our data from 2-dimensions to 1-dimension.

Principal Components

- What if height and weight have the same variance?
  - Can we keep both?
  - Can we combine them?
- The maximum amount of variance lies not in the x-axis, nor in the y-axis, but a diagonal line across.
  - The second-largest variance would be a line 90 degrees that cuts through the first.
- To represent these 2 lines, PCA combines both height and weight to create two new variables. It could be 30% height and 70% weight or any other combinations depending on the data.
- These two new variables are called the **first principal component (PC1)** and the **second principal component (PC2)**.

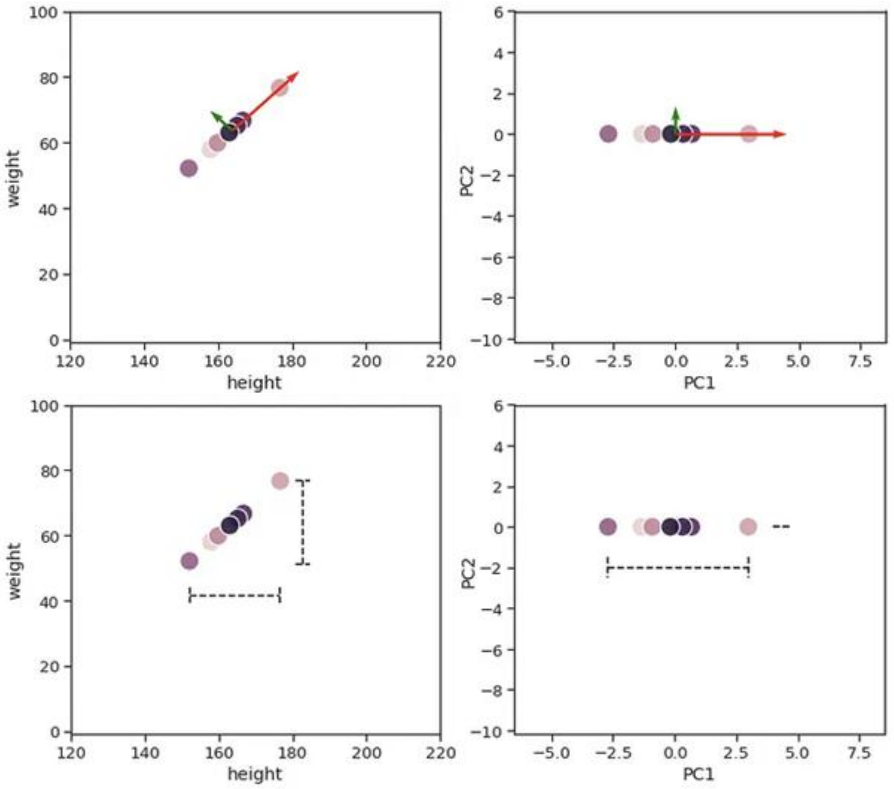




Principal Components (cont.)

- Instead of using height and weight on the two axes, we can use PC1 and PC2 respectively.
- PC1 alone can capture the total variance of Height and Weight combined.
  - We can safely remove PC2 and know that our new data is still representative of the original data.

- When it comes to real data, we won't get a principal component that captures 100% of the variances.
- We generally choose the least number of principal components that would explain the most amount of our original data.



Feature	Variance	Feature	Variance
Height	1.11	PC1	2.22
Weight	1.11	PC2	0
Total	2.22	Total	2.22

Computing Principal Components

- Principal components are basically vectors that are linearly uncorrelated and have a variance within data.
  - From the principal components top *p* is picked which have the most variance.

- Eigenvectors and Eigenvalues
  - Property of a matrix which satisfies the following equation:

$$Ax = \lambda x$$

where A denotes the matrix, x denotes the eigenvector and λ denotes the eigenvalues.

- Are principal components the eigenvectors of the covariance matrix?
- By finding the eigenvalues and eigenvectors of the covariance matrix, we find that **the eigenvectors with the largest eigenvalues correspond to the dimensions that have the strongest correlation in the dataset**. This is the principal component.

Principal Components Analysis

- How to do a Principal Component Analysis

1. Standardize the range of continuous initial variables
2. Compute the covariance matrix to identify correlations
3. Compute the eigenvectors and eigenvalues of the covariance matrix to identify the principal components
4. Create a feature vector to decide which principal components to keep
5. Recast the data along the principal components axes

1- Standardization

- If there are large differences between the ranges of variables, those variables with larger ranges will dominate over those with small ranges
- To Prevent biased results

$$z = \frac{value - mean}{standad\ deviation}$$

2- Covariance Matrix

- To Identify correlations between variables, the covariance matrix is used.
- For instance, for a 3-dimensional dataset: $\begin{bmatrix} Var(x) & Cov(x,y) & Cov(x,z) \\ Cov(y,x) & Var(y) & Cov(y,z) \\ Cov(z,x) & Cov(z,y) & Var(z) \end{bmatrix}$
  - It's actually the sign of the covariance that matters
    - If positive then: the two variables increase or decrease together (correlated)
    - If negative then: one increases when the other decreases (Inversely correlated)

3- Identify Principal Components

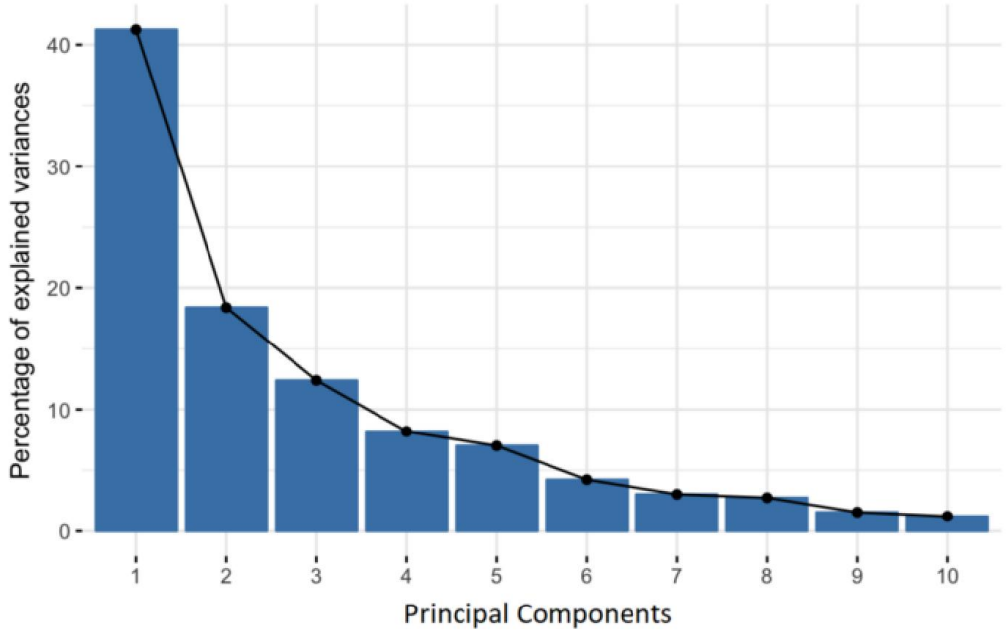
- We need to compute Eigenvectors and Eigenvalues of the covariance matrix to determine the principal components of the data.

- Principal components are new variables that are constructed as linear combinations of initial variables.

- These combinations are done in such a way that the principal components are uncorrelated and most of the information within the initial variables is squeezed into the first component, then maximum remaining information in the second and so on.

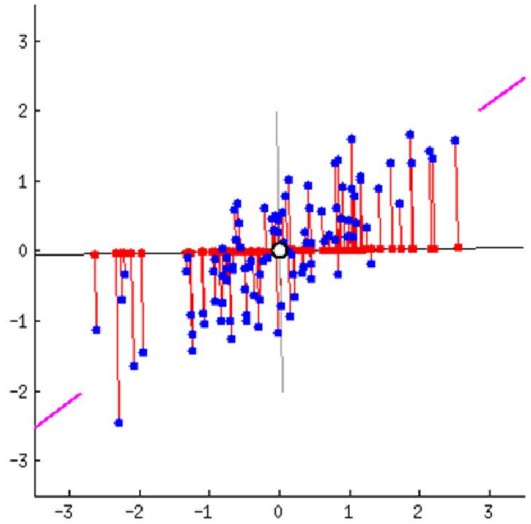
3- Identify Principal Components (cont.)

- Geometrically speaking, principal components represent the directions of the data that explain a **maximal amount of variance**, that is to say, the lines that capture most information of the data.



3- Identify Principal Components (cont.)

- Can you guess the first principal component within the data?
  - It's approximately the line that matches the purple marks because it goes through the origin and it's the line in which the projection of the points (red dots) is the most spread out.
  - It's the line that maximizes the variance (the average of the squared distances from the projected points (red dots) to the origin).



3- Identify Principal Components (cont.)

- The second principal component is calculated in the same way, with the condition that it is uncorrelated with (i.e., perpendicular to) the first principal component and that it accounts for the next highest variance.

- This continues until a total of p principal components have been calculated, equal to the original number of variables.

4- Feature Vector

- Let's suppose that our data set is 2-dimensional with 2 variables ***x, y*** and that the eigenvectors and eigenvalues of the covariance matrix are as follows:

$$v1 = \begin{bmatrix} 0.6778736 \\ 0.7351785 \end{bmatrix} \qquad \lambda_1 = 1.284028$$

$$v2 = \begin{bmatrix} -0.7351785 \\ 0.6778736 \end{bmatrix} \qquad \lambda_2 = 0.04908323$$

- If we rank the eigenvalues in descending order, we get $\lambda_1 > \lambda_2$, which means that the eigenvector that corresponds to the first principal component (PC_1) is v_1 and the one that corresponds to the second principal component (PC_2) is v_2.

- PC_1 and PC_2 carry respectively $\frac{\lambda_1}{\lambda_1+\lambda_2} = 96\%$ and $\frac{\lambda_2}{\lambda_1+\lambda_2} = 4\%$ of the variance of the data.

4- Feature Vector (cont.)

- We can either form a feature vector with both of the eigenvectors v_1 and v_2:

$$\begin{bmatrix} 0.6778736 & -0.7351785 \\ 0.7351785 & 0.6778736 \end{bmatrix}$$

Or discard the eigenvector v_2, which is the one of lesser significance, and form a feature vector with v_1 only:

$$\begin{bmatrix} 0.6778736 \\ 0.7351785 \end{bmatrix}$$

- Discarding the eigenvector v_2 will reduce dimensionality by 1, and will consequently cause a loss of information in the final data set. But given that v_2 was carrying only 4 percent of the information, the loss will be therefore not important and we will still have 96 percent of the information that is carried by v_1.

5- Recast Data

- In this step, the aim is to use the feature vector formed using the eigenvectors of the covariance matrix, to reorient the data from the original axes to the ones represented by the principal components.
- This can be done by multiplying the transpose of the original data set by the transpose of the feature vector.

$$FinalDataSet = FeatureVector^T * StandardizedOriginalDataSet^T$$

Source

- https://towardsdatascience.com/
- https://medium.com/
- https://builtin.com/data-science/step-step-explanation-principal-component-analysis