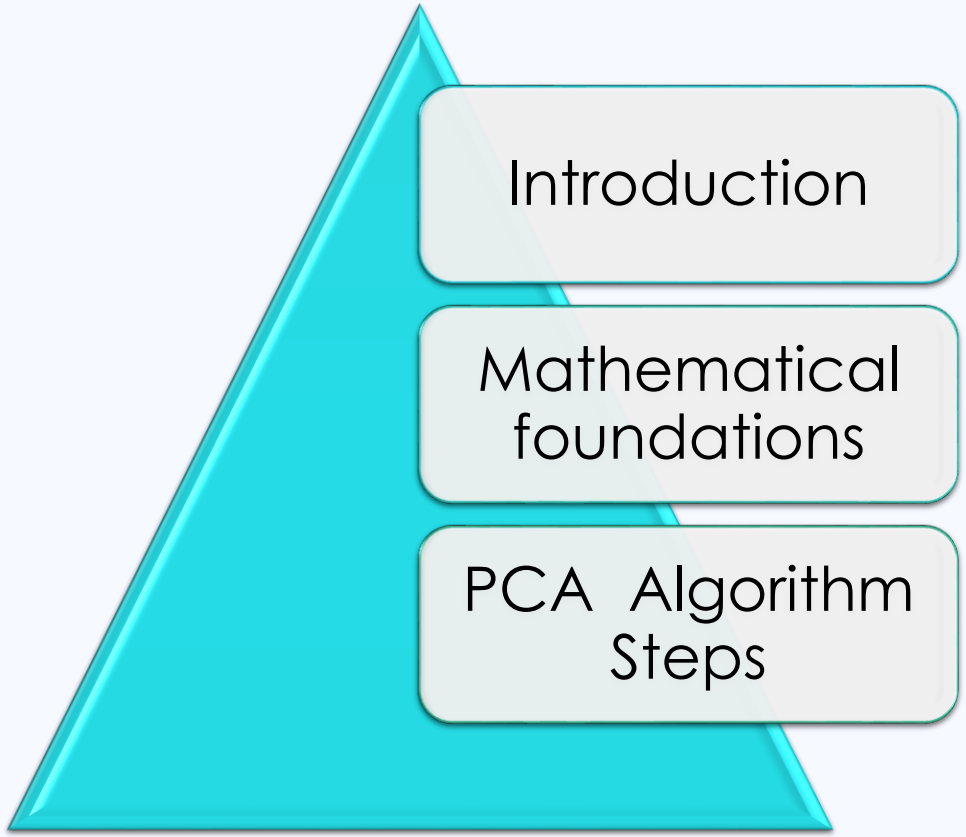
PRINCIPAL COMPONENT ANALYSIS (PCA)

ZAINEB MEZDOUD

OUTLINE


Introduction
Mathematical foundations
PCA Algorithm Steps

INTRODUCTION

PCA is a useful statistical technique that has found application in fields such as face recognition and image compression, and is a common technique for finding patterns in data of high dimension.
Before getting to a description of PCA, this tutorial first introduces mathematical concepts that will be used in PCA. It covers standard deviation, covariance, eigenvectors and eigenvalues.
This background knowledge is meant to make the PCA section very straightforward.

MATHEMATICAL FOUNDATIONS

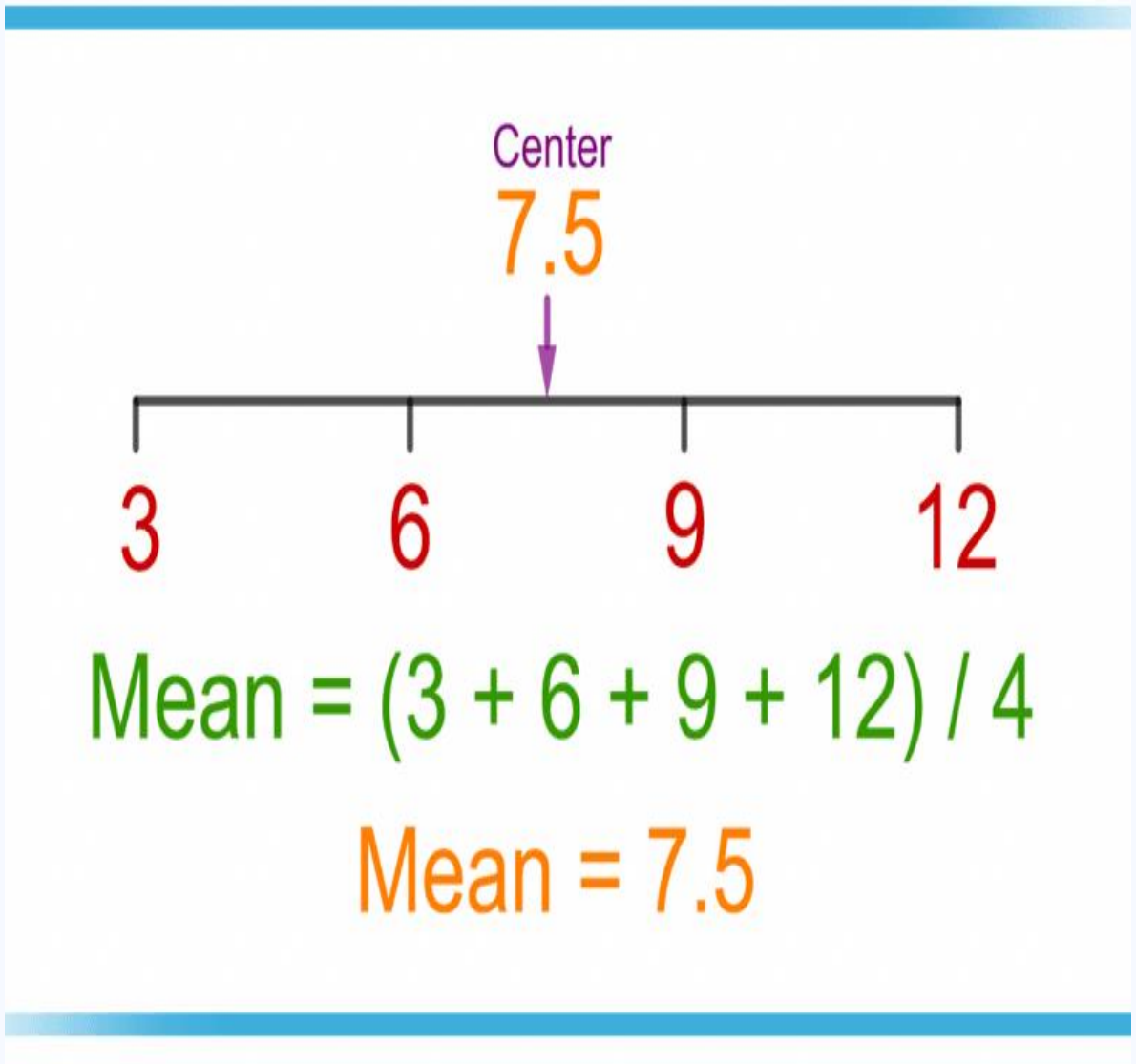
Statistics

Statistics revolves around the fundamental concept of analyzing large sets of data to uncover relationships between individual data points. In this context, I will explore several key measures that can be applied to a dataset and discuss what insights they provide about the data as a whole.

The mean

The **mean**, often referred to as the **average**, is one of the most commonly used measures in statistics.
It is calculated by adding up all the values in a dataset and then dividing that total by the number of values.
The mean provides a central value that represents the overall distribution of the data, giving you a sense of the "typical" or "expected" value within the set.

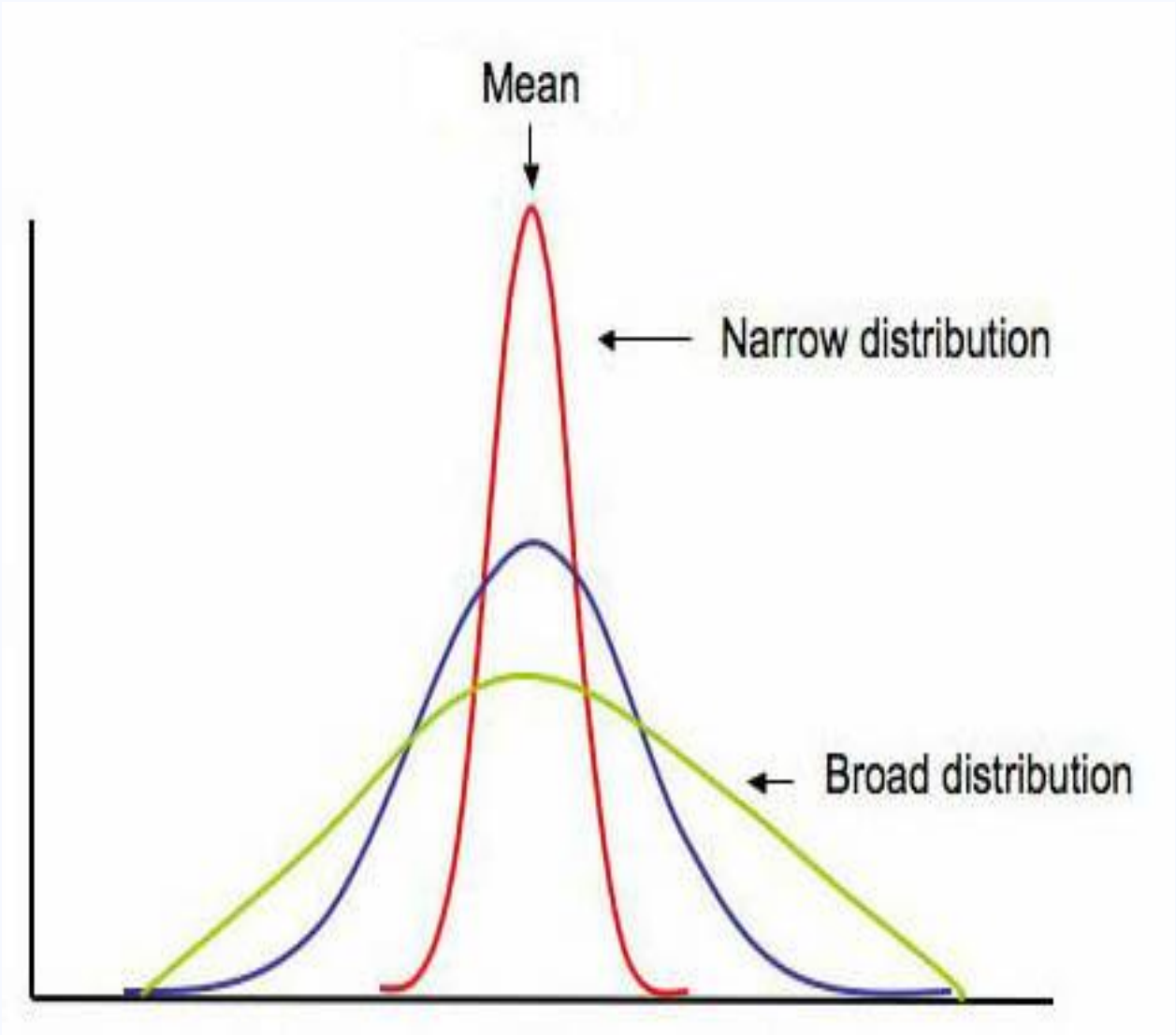
$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$



unfortunately, the mean doesn't tell us a lot about the data except for a sort of middle point. For example, these two data sets have exactly the same mean (10), but are obviously quite different:

$$[\,0\ 8\ 12\ 20\,] \; and \; [\,8\ 9\ 11\ 12\,]$$

So what is different about these two sets?
It is the ***spread*** of the data that is different.



The Standard Deviation (SD)

The Standard Deviation (SD) of a data set is a measure of how spread out the data is.

In simple terms, the SD is the **average distance** between each data point and the mean of the dataset.

The formula for the population standard deviation is:

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2}$$

And for the sample standard deviation:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Where:

- x_i = each individual data point
- μ = population mean
- $\bar{x}$ = sample mean
- n = number of data points

The way to calculate it is to compute the squares of the distance from each data point to the mean of the set, add them all up, divide by $n-1$ $(or\ n)$, and take the positive square root

Set 1:

X	$(X-\bar{X})$	$(X-\bar{X})^2$
0	-10	100
8	-2	4
12	2	4
20	10	100
Total		208
Divided by (n-1)		69.333
Square Root		8.3266

Set 2:

X_i	$(X_i-\bar{X})$	$(X_i-\bar{X})^2$
8	-2	4
9	-1	1
11	1	1
12	2	4
Total		10
Divided by (n-1)		3.333
Square Root		1.8257

Note: Variance is another measure of the spread of data in a data set. In fact it is almost identical to the standard deviation.

$$\sigma^2 = \frac{\Sigma(x-\mu)^2}{N}$$

Covariance

Standard deviation and **variance** are measures that operate on **a single dimension** of data. This means that if your dataset has multiple dimensions (or variables), you can only compute the standard deviation or variance **for each dimension independently**, without considering how the dimensions interact.

However, in many cases, it's important to understand how **two dimensions vary together** for example, how a change in one variable might relate to changes in another.

This is where **covariance** comes in.

Covariance Formula For Population

$$Cov(X,Y) = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{n}$$

Covariance Formula For Sample

$$Cov(X,Y) = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}$$

Covariance measures the degree to which two variables **change together**.

A **positive covariance** indicates that the variables tend to increase or decrease together, while a **negative covariance** means that as one increases, the other tends to decrease.

An interesting property of covariance is that when you compute the covariance of a variable **with itself**, you get the **variance**.

In this sense, variance is a special case of covariance.

Correlation

While **covariance** tells us how two variables change together, it doesn't give us a sense of the **strength** or **consistency** of that relationship, especially because its value depends on the scale of the variables. To address this, we use **correlation**.

Correlation is a standardized version of covariance that measures both the **strength** and **direction** of a linear relationship between two variables. It is calculated by dividing the covariance by the product of the standard deviations of the two variables.

Population Correlation Coefficient

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\sum (xi - \bar{x})(yi - \bar{y})}{\sqrt{\left(\sum (xi - \bar{x})^2\right)\left(\sum (yi - \bar{y})^2\right)}}$$

Where,

- $\sigma_x, \sigma_y \rightarrow$ Population Standard Deviation
- $\sigma_{xy} \rightarrow$ Population Covariance
- $\bar{X}, \bar{Y} \rightarrow$ Population Mean

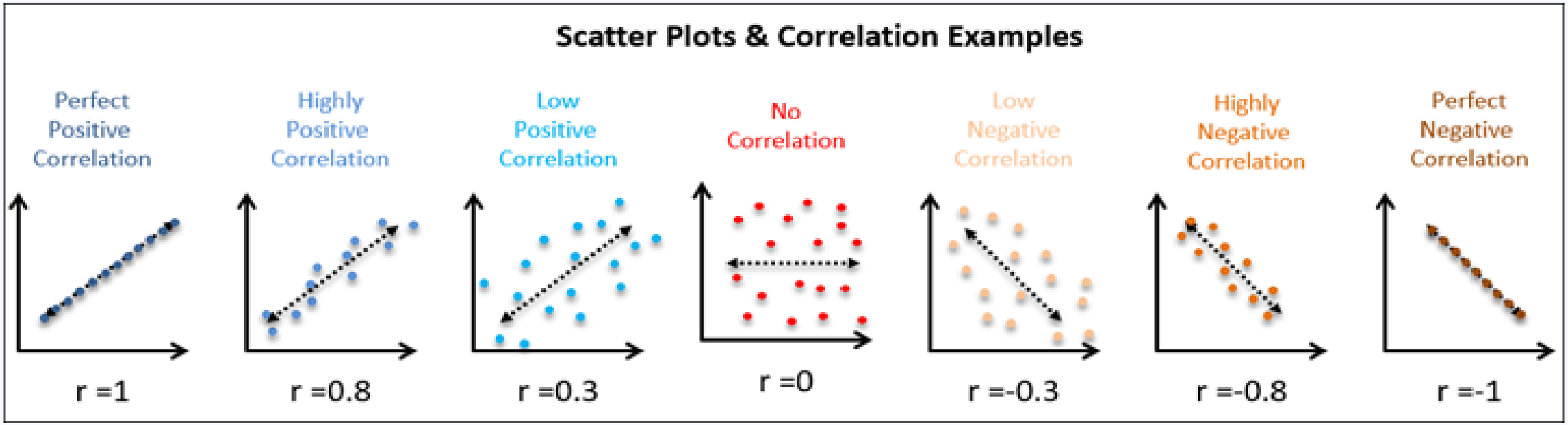
Sample Correction, coefficient between x and y

$$r_{xy} = \frac{S_{xy}}{S_x S_y} = \frac{\sum (xi - \bar{x})(yi - \bar{y})}{\sqrt{\left(\sum (xi - \bar{x})^2\right)\left(\sum (yi - \bar{y})^2\right)}}$$

Where,

- $S_x, S_y \rightarrow$ Sample Standard Deviation
- $S_{xy} \rightarrow$ Sample Covariance
- $\bar{X}, \bar{Y} \rightarrow$ SampleMean

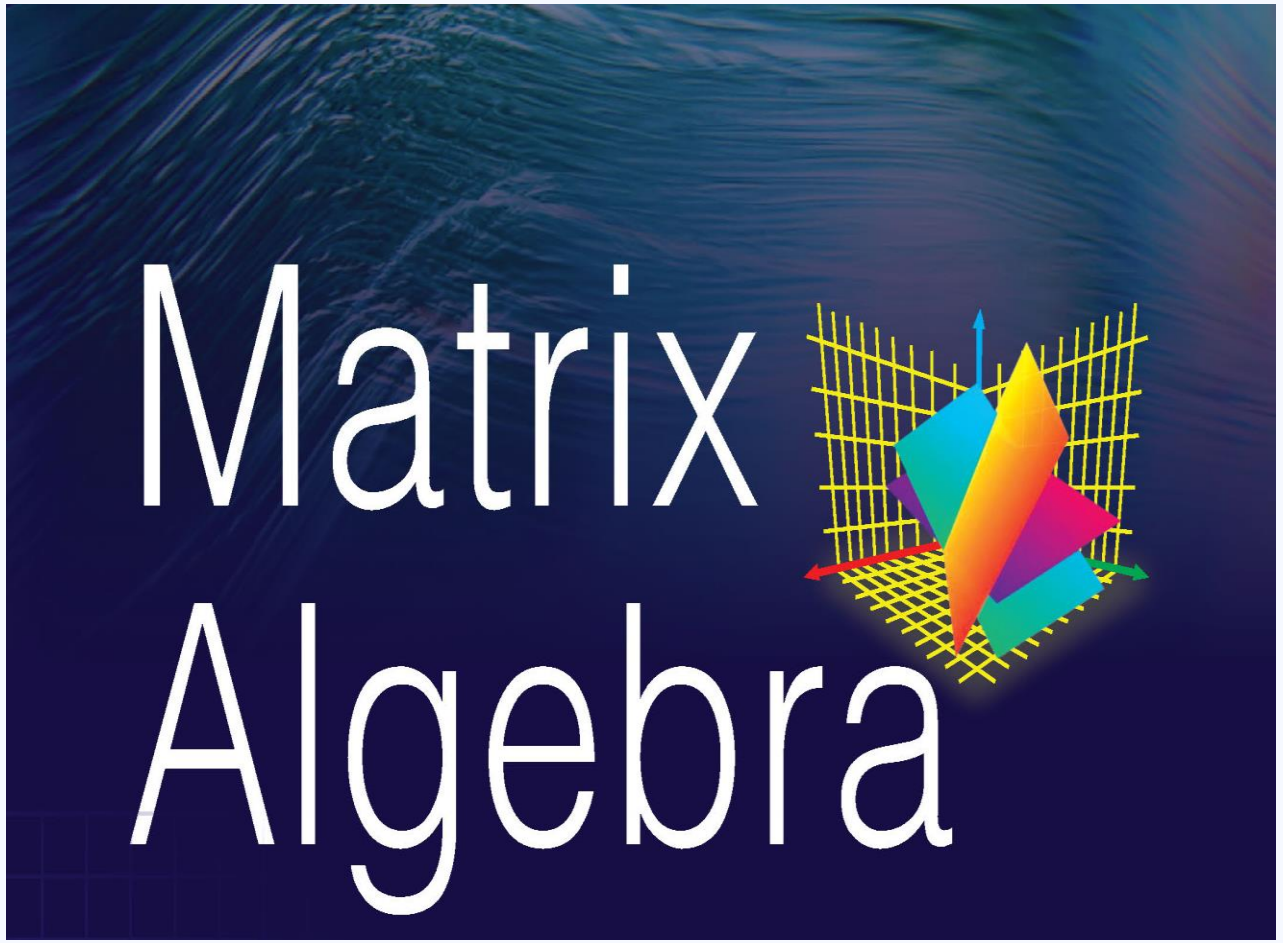
The result is a value called the **correlation coefficient**, usually denoted as r (or ρ), which always falls between **-1 and 1.**



Note: Correlation provides a unit-less measure, allowing you to easily compare the strength of relationships between different pairs of variables, regardless of their original units of measurement.

MATRIX ALGEBRA

See chapter 01



The covariance Matrix

Recall that covariance is always measured between 2 dimensions. If we have a data set with more than 2 dimensions, there is more than one covariance measurement that can be calculated.

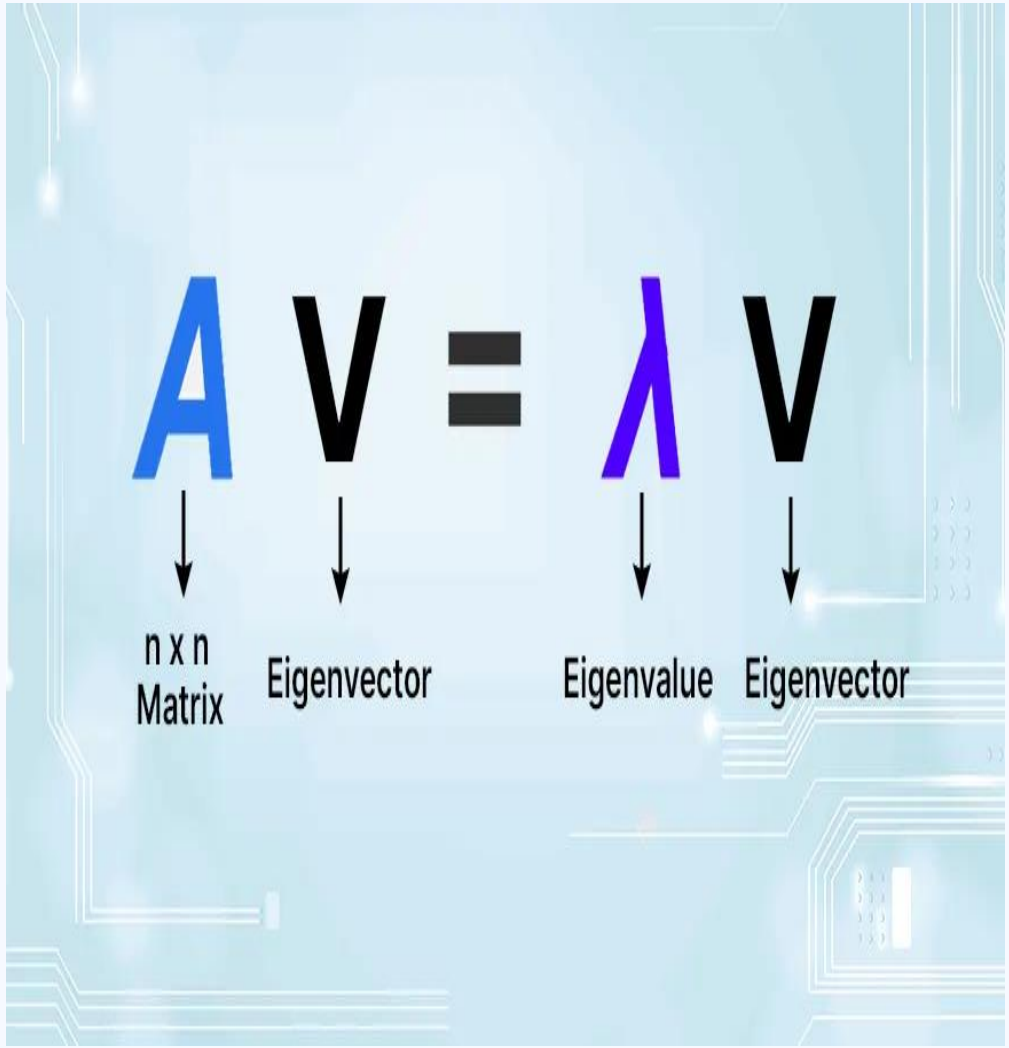
The **covariance matrix** is a square matrix that summarizes the covariances between **each pair of variables** in the dataset.
Each element in the matrix at position (i, j) represents the **covariance between variable i and variable j**.

Example structure of a 3-variable covariance matrix

$$\begin{bmatrix} Var(x_1) & \cdots & Cov(x_1, x_n) \\ \vdots & \ddots & \vdots \\ Cov(x_n, x_1) & \cdots & Var(x_n) \end{bmatrix}$$

$$\Sigma = \begin{pmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) & \mathrm{Cov}(X_1, X_3) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}(X_2) & \mathrm{Cov}(X_2, X_3) \\ \mathrm{Cov}(X_3, X_1) & \mathrm{Cov}(X_3, X_2) & \mathrm{Var}(X_3) \end{pmatrix}$$

Eigenvectors



Eigenvectors and eigenvalues are fundamental concepts in linear algebra, particularly in the analysis of linear transformations. Given a square matrix A, an eigenvector is a non-zero vector $\vec{v}$ that changes only in magnitude (not in direction) when A is applied to it.

Mathematically, this is expressed as $A\vec{v} = \lambda\vec{v}$, where λ is a scalar known as the **eigenvalue** corresponding to the eigenvector $\vec{v}$.

The eigenvalue represents the factor by which the eigenvector is stretched or compressed during the transformation. These quantities provide deep insight into the structure and behavior of linear systems, making them essential in numerous scientific fields including physics (e.g., quantum mechanics), engineering (e.g., vibrations and stability analysis), data science (e.g., principal component analysis), and machine learning. Eigenvectors indicate the invariant directions under a transformation, while eigenvalues quantify the effect of the transformation along those directions.

PCA ALGORITHM STEPS

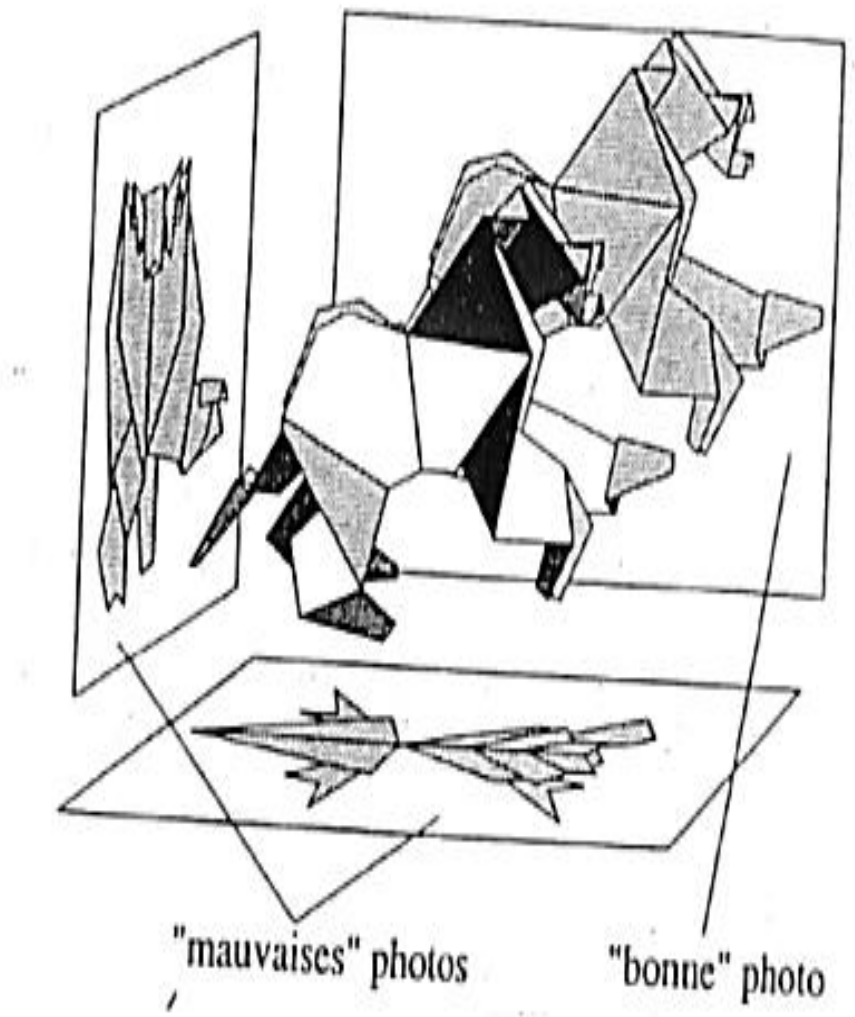
Principal Components Analysis (PCA)

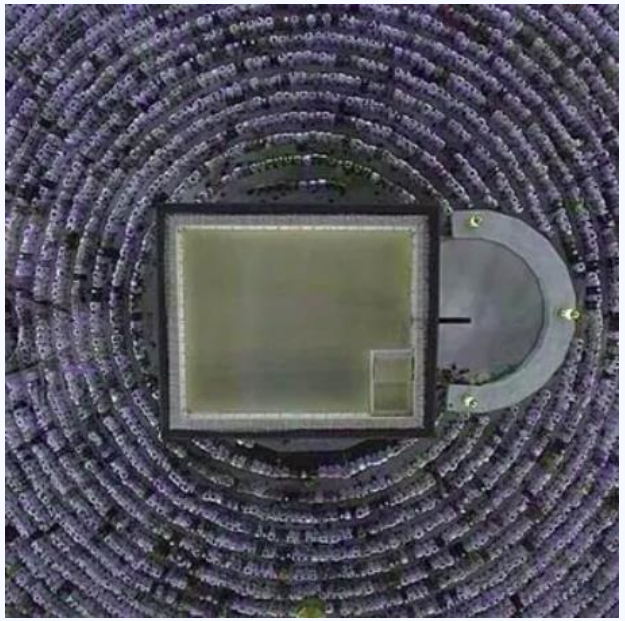
What is it? It is a way of identifying patterns in data, and expressing the data in such a way as to highlight their similarities and differences. Since patterns in data can be hard to find in data of high dimension, where the luxury of graphical representation is not available, PCA is a powerful tool for analysing data.

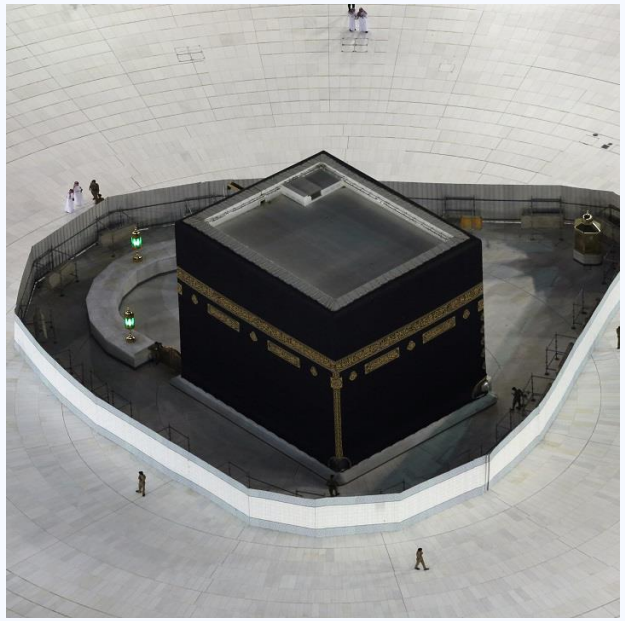
The other main advantage of PCA is that once you have found these patterns in the data, and you compress the data, ie. by reducing the number of dimensions, without much loss of information

PCA

Principal Component Analysis (PCA) is a common statistical technique for identifying and re-referencing the data by linear mapping, which transforms a number of possibly correlated variables into a smaller number of uncorrelated variables known as principal components.



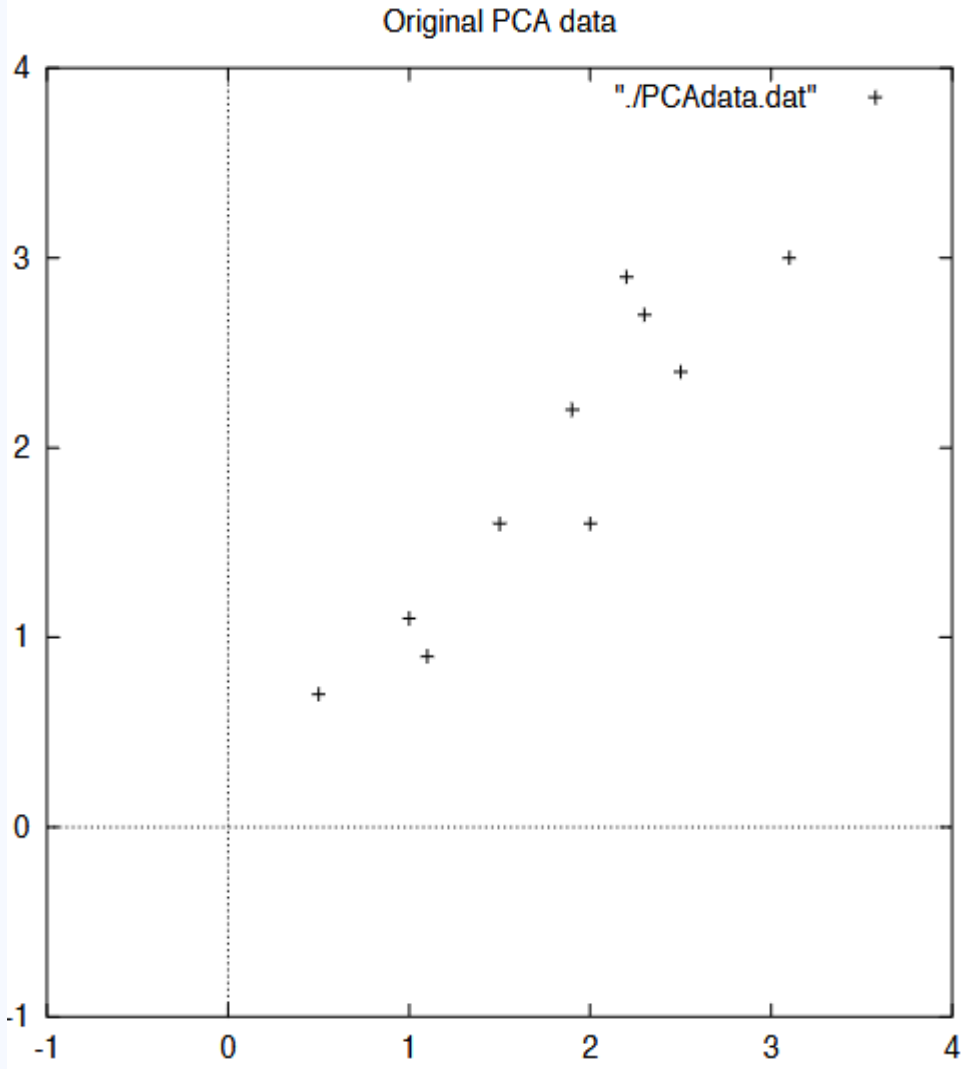




Method

Step 1: Get some data

In my simple example, I am going to use my own made-up data set. It's only got 2 dimensions, and the reason why I have chosen this is so that I can provide plots of the data to show what the PCA analysis is doing at each step.



Data =

x	y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9

Method

Step 2: Subtract the mean

For PCA to work properly, you have to subtract the mean from each of the data dimensions. The mean subtracted is the average across each dimension. So, all the x values have $\bar{x}$(the mean of the x values of all the data points) subtracted, and all the y values have $\bar{y}$ subtracted from them. This produces a data set whose mean is zero.

$$X_c = X - \bar{X}$$

DataAdjust =

x	y
.69	.49
-1.31	-1.21
.39	.99
.09	.29
1.29	1.09
.49	.79
.19	-.31
-.81	-.81
-.31	-.31
-.71	-1.01

Step 3: Calculate the covariance matrix (or correlation matrix)

Calculate **the covariance matrix** is done in exactly the same way as was discussed before.

Since the data is 2 dimensional, the covariance matrix will be 2 × 2. So give you the result:

$$cov = \begin{pmatrix} .616555556 & .615444444 \\ .615444444 & .716555556 \end{pmatrix}$$

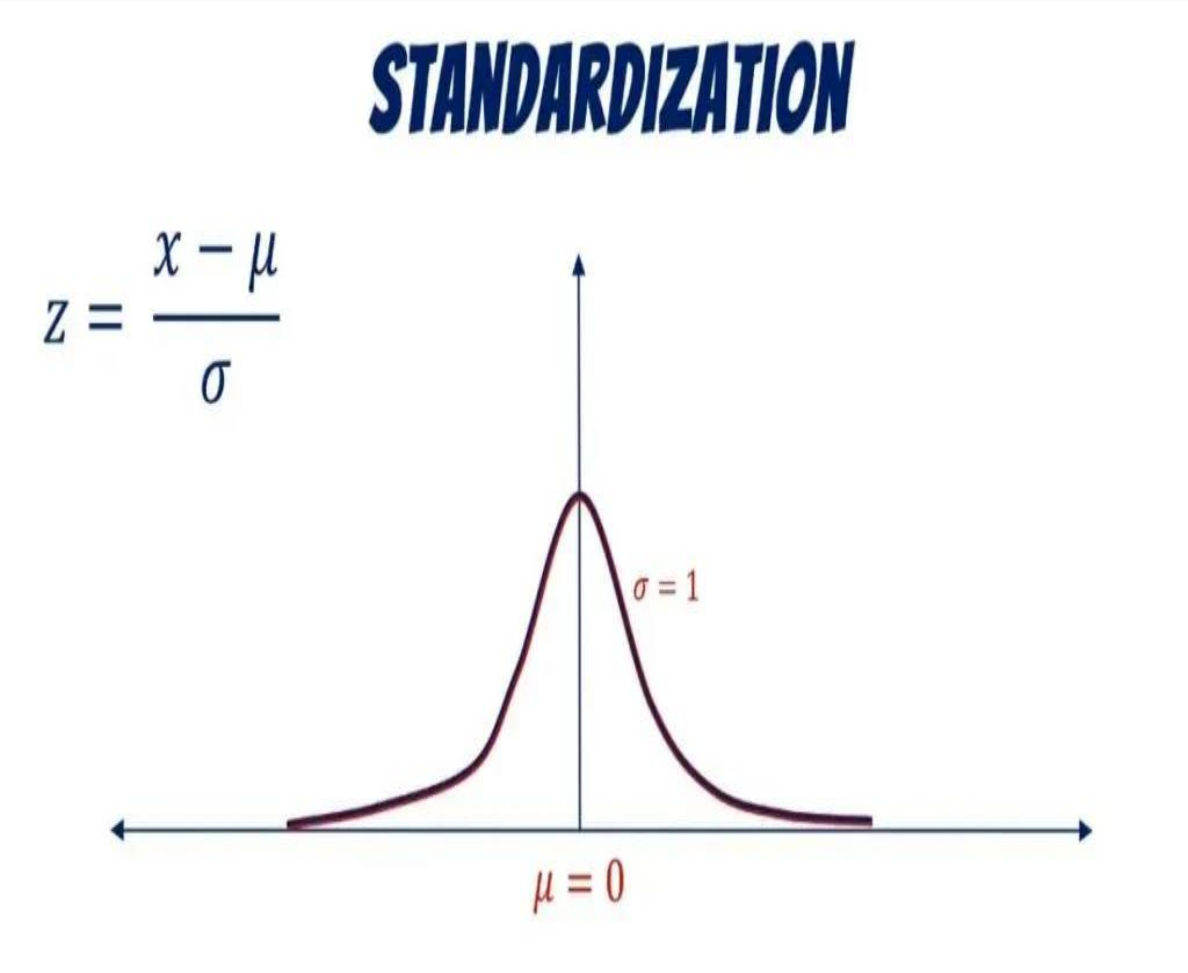
$$\Sigma = \begin{bmatrix} Var(x_1) & \dots & Cov(x_1, x_n) \\ \vdots & \ddots & \vdots \\ Cov(x_n, x_1) & \dots & Var(x_n) \end{bmatrix}$$

$$\Sigma = \frac{1}{n} X_c^T X_c$$

The calculation of **the correlation matrix** involves first **standardizing** the dataset to ensure that each variable has a mean of zero and a standard deviation of one.

$$Z = \frac{X - X_c}{\sigma}$$

This step removes the influence of differing scales or units across variables, making them comparable.



Once **standardization** is complete, **the correlation matrix** is computed by multiplying the transpose of the standardized data matrix by the matrix itself and dividing the result by n.

$$\mathrm{R} = \frac{1}{n} Z^T Z$$

The resulting matrix contains **Pearson correlation coefficients**, where each element represents the linear relationship between a pair of variables, ranging from $-\mathbf{1}$ (perfect negative correlation) to $+\mathbf{1}$ (perfect positive correlation), with $\mathbf{0}$ indicating no linear correlation. The diagonal elements of the matrix are always equal to 1, as each variable is perfectly correlated with itself.

$$Correlation = \frac{Cov\,(x, y)}{\sigma x * \sigma y}$$

In Principal Component Analysis (PCA), the choice between using the covariance matrix or the correlation matrix depends on the nature and scale of the variables in the dataset. When the variables are measured on the same scale and in the same units, the covariance matrix is typically used, as it reflects the actual variances and covariances between variables. This approach preserves the original magnitudes of variability and emphasizes directions in the data space where the absolute variance is maximized. However, when the variables differ in scale or units—for example, when one variable is measured in kilograms and another in meters—the correlation matrix becomes more appropriate. The correlation matrix is computed from standardized variables, ensuring that each variable contributes equally to the analysis regardless of its original scale. This standardization process makes PCA more robust to scale differences and focuses on the structure of relationships among variables rather than their absolute variances. Thus, the correlation matrix is used when equal weighting of variables is desired, while the covariance matrix is preferred when maintaining the true variances is important.

Step 4: Calculate the eigenvectors and eigenvalues of the covariance Matrix (or correlation matrix)

Since the covariance matrix is square, we can calculate the eigenvectors and eigenvalues for this matrix.

These are rather important, as they tell us useful information about our data. In the meantime, here are the eigenvectors and eigenvalues:

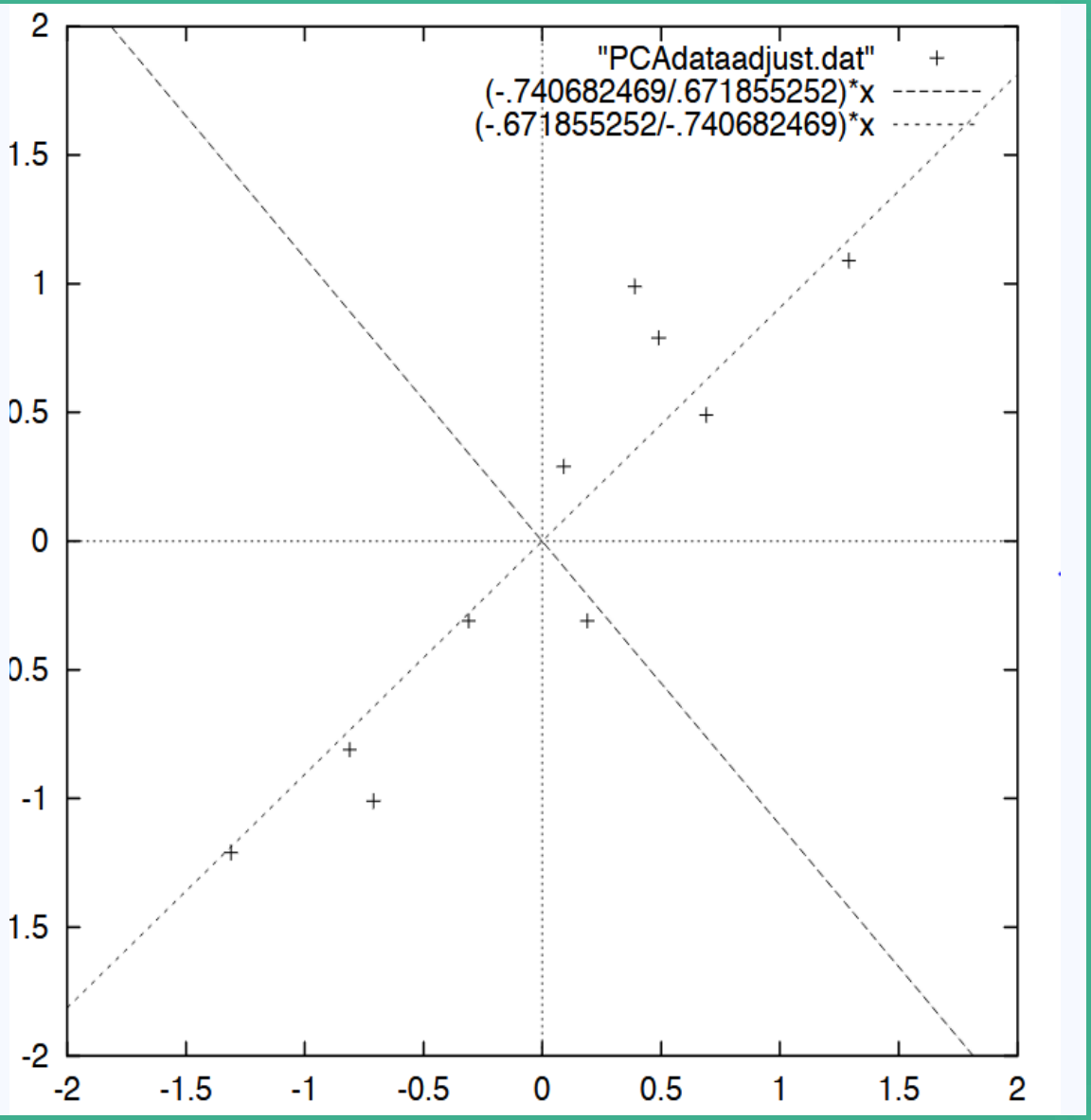
$$eigenvalues = \begin{pmatrix} .0490833989 \\ 1.28402771 \end{pmatrix}$$

$$eigenvectors = \begin{pmatrix} -.735178656 & -.677873399 \\ .677873399 & -.735178656 \end{pmatrix}$$

Step 5: Choosing components and forming a feature vector

Here is where the notion of data compression and reduced dimensionality comes into it. If you look at the eigenvectors and eigenvalues from the previous section, you will notice that the eigenvalues are quite different values. In fact, it turns out that the eigenvector with the highest eigenvalue is the principle component of the data set.

In our example, the eigenvector with the larges eigenvalue was the one that pointed down the middle of the data. It is the most significant relationship between the data dimensions.



In general, once eigenvectors are found from the covariance matrix, the next step is to order them by eigenvalue, highest to lowest.

This gives you the components in order of significance. Now, if you like, you can decide to ignore the components of lesser significance. You do lose some information, but if the eigenvalues are small, you don't lose much.

If you leave out some components, the final data set will have less dimensions than the original. To be precise, if you originally have n dimensions in your data, and so you calculate n eigenvectors and eigenvalues, and then you choose only the first p eigenvectors, then the final data set has only p dimensions.

Sort the eigenvalues in descending order: $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$.

Decide how many principal components to retain based on the sorted eigenvalues using one of these methods:

1. ***Variance Explained Criterion : (most recommended approach in theory)***

   - Calculate Total Variance : The total variance in the dataset is given by: $\text{Total Variance} = \sum_{i=1}^{p} \lambda_i$.
   - Calculate Proportion of Variance Explained: The proportion of variance explained by each principal component i is:

     $$\text{Proportion of Variance for PC}_i = \frac{\lambda_i}{\text{Total Variance}} = \frac{\lambda_i}{\sum_{j=1}^{p} \lambda_j}.$$

   - Calculate Cumulative Explained Variance : The cumulative explained variance for the first \(k\) principal components is:

     $$\text{Cumulative Explained Variance}_k = \sum_{i=1}^{k} \frac{\lambda_i}{\sum_{j=1}^{p} \lambda_j}.$$

     This can also be expressed as:

     $$\text{Cumulative Explained Variance}_k = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{j=1}^{p} \lambda_j}$$

   - Choose the number of components where the cumulative explained variance reaches a satisfactory threshold (commonly 80% to 90%).

Kaiser Criterion : Retain components with eigenvalues greater than . This rule is based on the idea that each component should explain at least as much variance as a single original variable. Components with eigenvalues less than 1 contribute less information than one of the original variables and are often discarded.

$$\begin{pmatrix} -.677873399 \\ -.735178656 \end{pmatrix}$$

Create a matrix V of the selected eigenvectors (principal components). This matrix will be used to transform the original data.

Step 6: Deriving the new data set

This the final step in PCA, and is also the easiest. Once we have chosen the components (eigenvectors) that we wish to keep in our data and formed a feature vector, we simply take the transpose of the vector and multiply it on the left of the original data set, transposed.

Transformed Data (Single eigenvector)

x
-.827970186
1.77758033
-.992197494
-.274210416
-1.67580142
-.912949103
.0991094375
1.14457216
.438046137
1.22382056

2. ***Project the original data onto the new feature space*** by multiplying the centered data matrix by the matrix of selected eigenvectors. This gives you the principal component scores.

$$\boldsymbol{C = \Sigma V \quad or \quad (C = ZV)}$$

where:

$\boldsymbol{C: is\ the\ matrix\ of\ principal\ component\ scores.}$

$\boldsymbol{\Sigma\ or\ (Z): is\ the\ centered\ (or\ centred\ and\ reduced)\ data\ matrix.}$

$\boldsymbol{V: is\ the\ matrix\ of\ selected\ eigenvectors.}$

3. Evaluate Component Contributions: Look at the loadings of each variable on the principal components to understand which variables contribute most to each component.
4. Perform Further Analysis: Depending on your goals, you can use the principal component scores for various analyses, such as clustering, regression, or visualization, while potentially reducing the dimensionality of the dataset.

THANK YOU