What is PCA and how it works

Principal Component Analysis, or PCA, is a ***dimensionality-reduction*** method that is often used to reduce the dimensionality of large data sets, by ***transforming a large set of variables into a smaller one*** that still contains most of the information in the large set.

Reducing the number of variables of a data set naturally comes at the expense of accuracy, but the trick in dimensionality reduction is to ***trade a little accuracy for simplicity***. Because smaller data sets are easier to explore and visualize and make analyzing data ***much easier*** and ***faster*** for machine learning algorithms without extraneous variables to process.

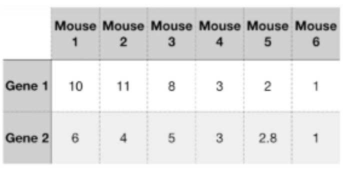
So to sum up, the idea of PCA is simple — reduce the number of variables of a data set, while preserving as much information as possible.

For example:

Lets start with simple data set:

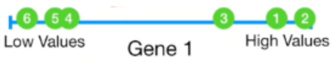
we've measured transcription of 2 genes, gene 1 & gene 2 in 6 differenet mice.

we can think about the ***mice as indvidual samples*** and their ***genes as variables*** that we measure for each sample.

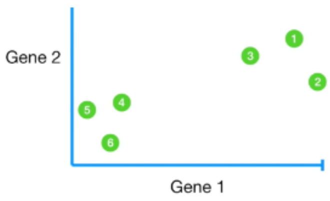


	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1

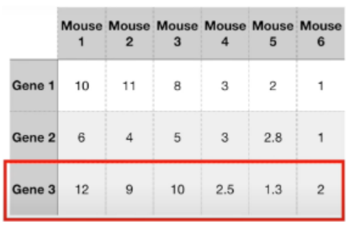
If we only measure 1 gene, we can plot the data on a number line:



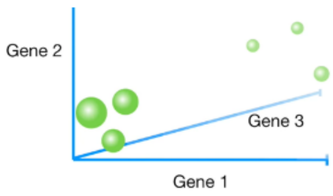
If we only measure 2 genes, we can plot the data on a 2-Dimensional x/y graph:



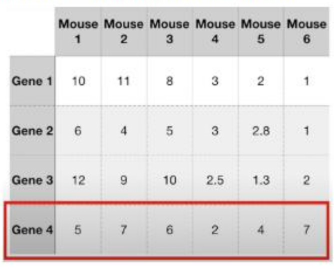
If we measured 3 genes, we can plot the data on a 3-Dimensional x/y/z graph:



	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1
Gene 3	12	9	10	2.5	1.3	2



But if we measured 4 genes, we can no longer plot the data - 4 genes require 4 dimensions:

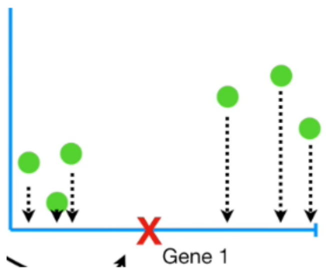


	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1
Gene 3	12	9	10	2.5	1.3	2
Gene 4	5	7	6	2	4	7

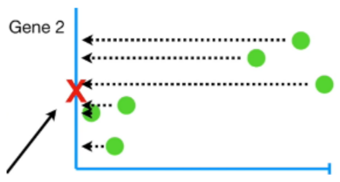
How PCA can take 4 or more gene measurements, and make it a 2-D PCA plot:

Let's go back to the dataset that only had 2 genes:

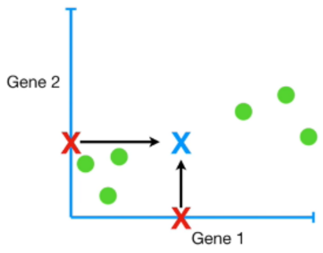
1) Calculate the average measurment for gene 1:



2) Calculate the average measurment for gene 2:



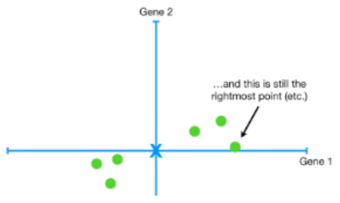
3) With the average value, we can calculate the center of the data:



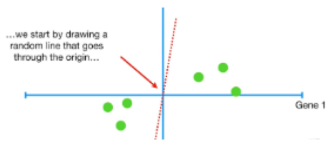
4) Shift the data so that the center is on top of the origin (0,0) in the graph:

Note:

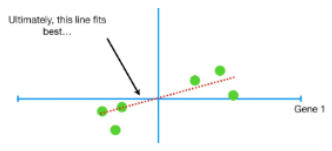
Shifting the data did'nt change how the data points are positioned *relative to each other*.



5) Draw a random line that goes through the origin:


...we start by drawing a random line that goes through the origin...
Gene 1

6) Rotate the line until it fits the data as well as is it:



Step by Step Explanation of PCA:

`1) STANDARDIZATION:`

The aim of this step: Standardize the range of the ***continuous initial variables*** so that each one of them contributes equally to the analysis.

For example, a variable that ranges between ***0 and 100*** will dominate over a variable that ranges between ***0 and 1***.

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

`2) COVARIANCE MATRIX COMPUTATION:`

The aim of this step: Understand how the variables of the ***input data set are varying from the mean with respect to each other***, or in other words, to see if there is any ***relationship between them***.

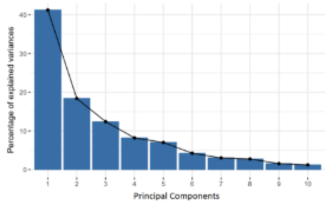
The covariance matrix is a p × p symmetric matrix (where p is the number of dimensions) that has as entries the covariances associated with all possible pairs of the initial variables. For example, for a 3-dimensional data set with 3 variables x, y, and z, the covariance matrix is a 3×3 matrix of this from.

$$\begin{bmatrix} Cov(x,x) & Cov(x,y) & Cov(x,z) \\ Cov(y,x) & Cov(y,y) & Cov(y,z) \\ Cov(z,x) & Cov(z,y) & Cov(z,z) \end{bmatrix}$$

`3) COMPUTE THE EIGENVECTORS AND EIGENVALUES OF THE COVARIANCE MATRIX TO IDENTIFY THE PRINCIPAL COMPONENTS:`

Eigenvectors and eigenvalues are the linear algebra concepts that we need to compute from the covariance matrix in order to determine the ***principal components*** of the data. Before getting to the explanation of these concepts, let's first understand what do we mean by principal components.

The idea is 10-dimensional data gives you 10 principal components, but PCA tries to put maximum possible information in the first component, then maximum remaining information in the second and so on, until having something like shown in the scree plot below.



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. The relationship between variance and information here, is that, the larger the variance carried by a line, the larger the dispersion of the data points along it, and the larger the dispersion along a line, the more the information it has. To put all this simply, just think of principal components as new axes that provide the best angle to see and evaluate the data, so that the differences between the observations are better visible.