Finamex®

Casa de Bolsa

PCA

What is a PCA?:

Principal component analysis (PCA) is a technique used to emphasize variation and bring out strong patterns in a dataset. It's often used to make data easy to explore and visualize.

"How do I take all of the variables I've collected and focus on only a few of them?"

By reducing the dimension of your feature space, you have fewer relationships between variables to consider and you are less likely to overfit your model.

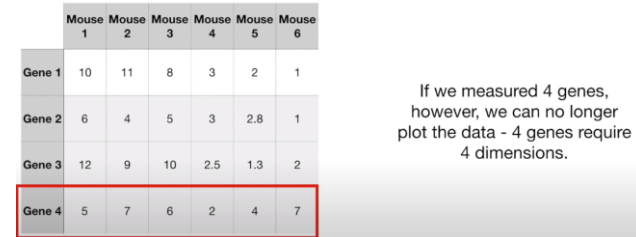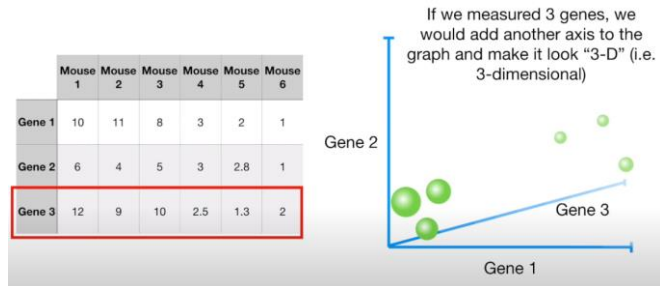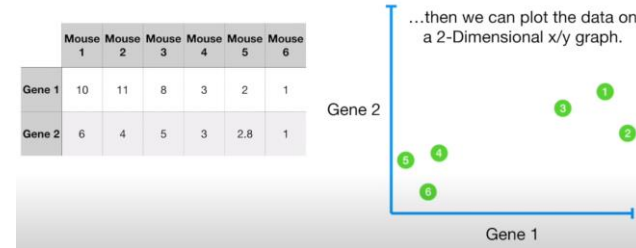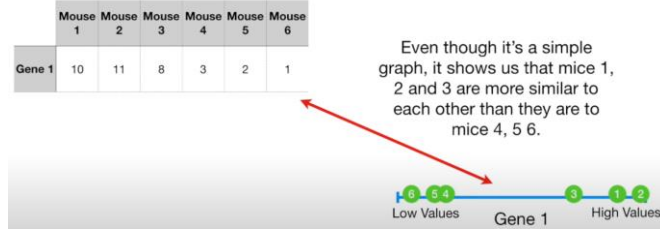Reducing the dimension of the feature space is called "dimensionality reduction".

When should I use PCA?:

- **Do you want to reduce the number of variables, but aren't able to identify variables to completely remove from consideration?**
- **Do you want to ensure your variables are independent of one another?**
- **Are you comfortable making your independent variables less interpretable?**

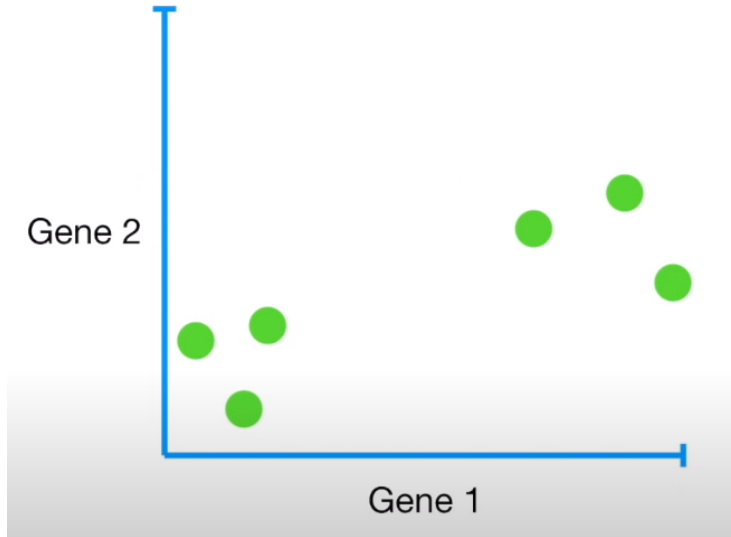**If you answered "NO" to question 3, <u>you should not use PCA.</u>**

PCA

**There are as many principal components as there are variables in the data.**

Even though it's a simple graph, it shows us that mice 1, 2 and 3 are more similar to each other than they are to mice 4, 5 6.

...then we can plot the data on a 2-Dimensional x/y graph.

If we measured 3 genes, we would add another axis to the graph and make it look "3-D" (i.e. 3-dimensional)

If we measured 4 genes, however, we can no longer plot the data - 4 genes require 4 dimensions.

PCA

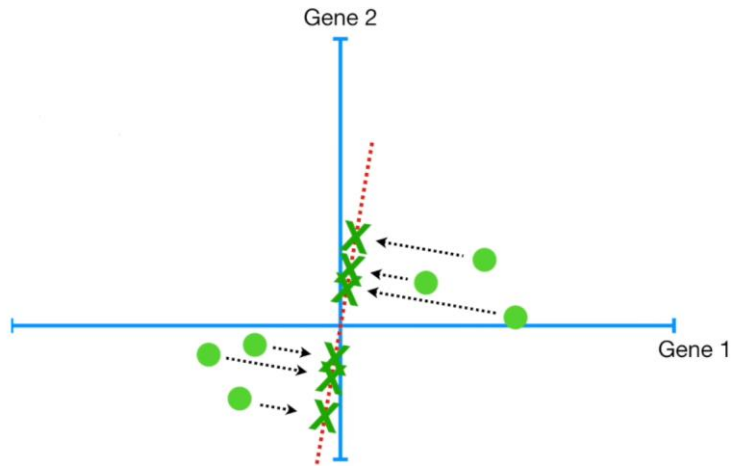- **Step 1:** **Centering the data** (Normalization)

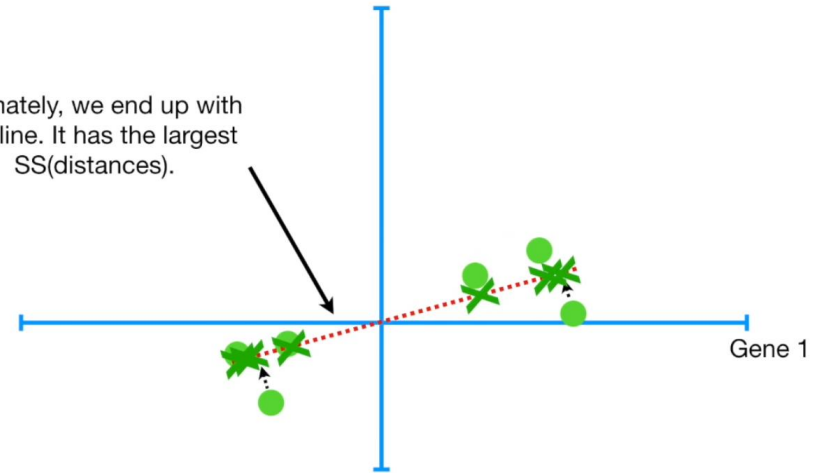NOTE: Shifting the data did not change how the data points are positioned *relative to each other*.
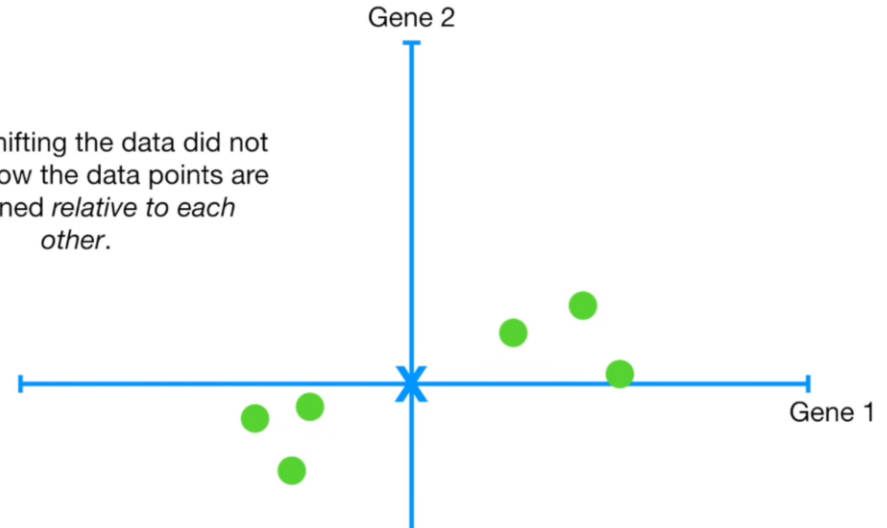
PCA

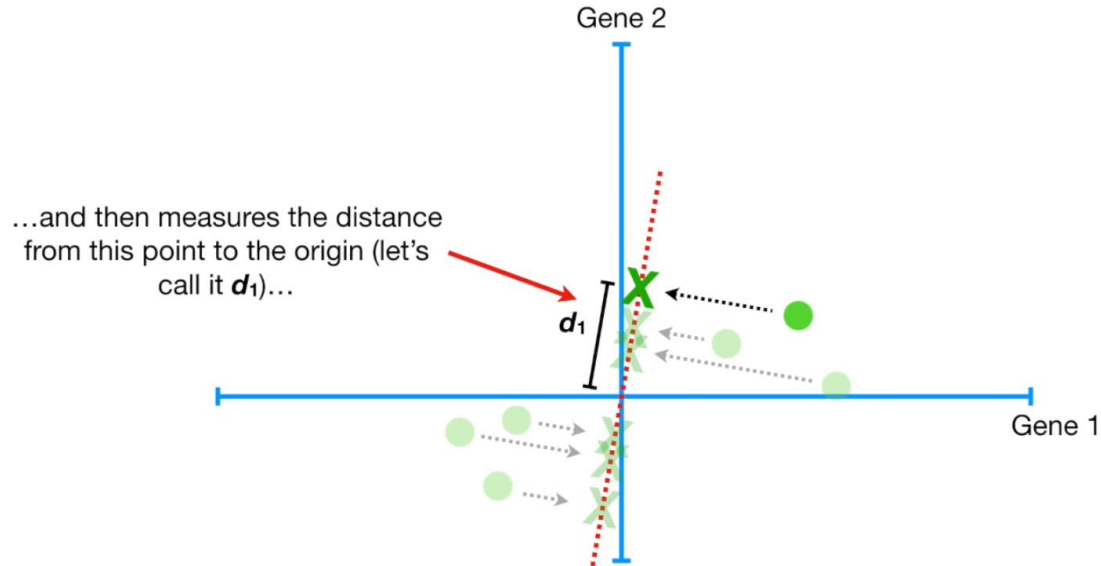- **Step 2:** **Make a line that maximizes the Sum of Squared distances.**
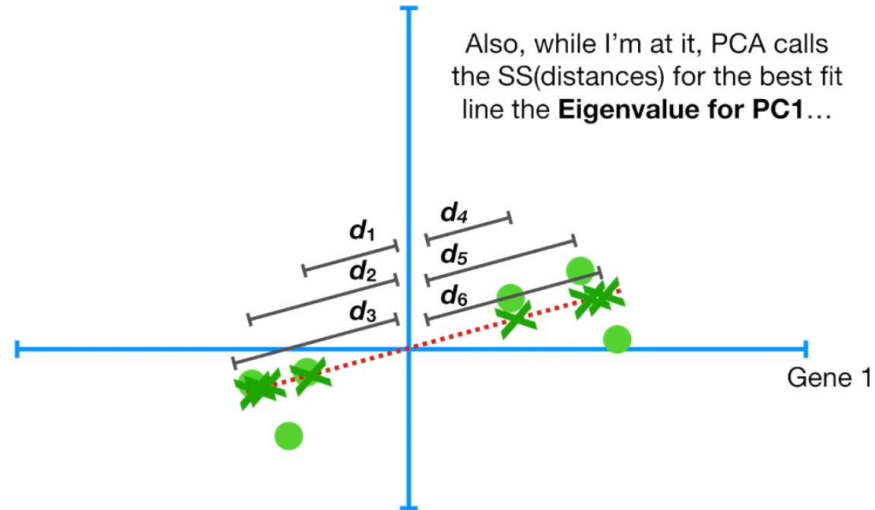
PCA

**Distance** (d_1) **: The distance from the projected value to the center.**

PCA

Sum of Squared distances: Called Eigenvalue.

$$d_1{}^2 + d_2{}^2 + d_3{}^2 + d_4{}^2 + d_5{}^2 + d_6{}^2 = \text{sum of squared distances} = \text{SS(distances)}$$

Also, while I'm at it, PCA calls the SS(distances) for the best fit line the **Eigenvalue for PC1**...

PCA

Second PCA: For any other PCA is simply a line through the origin that is perpendicular to the rest of PCAs.
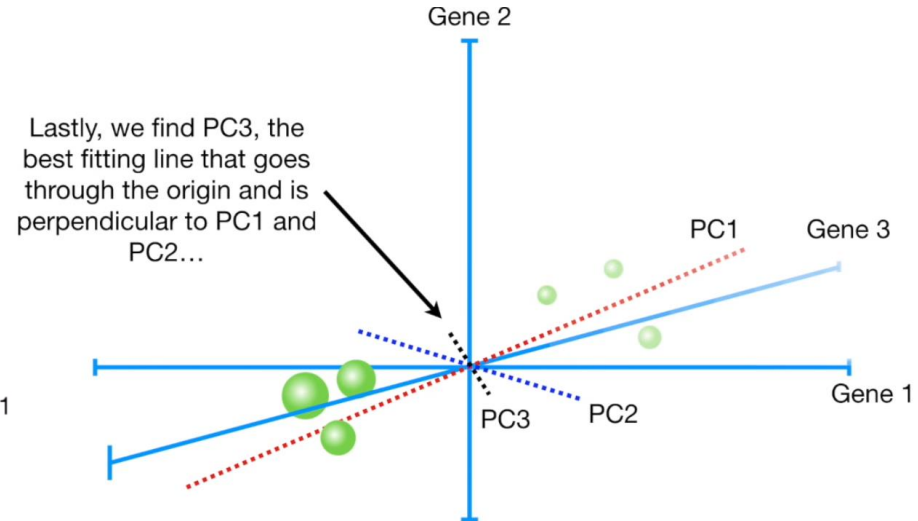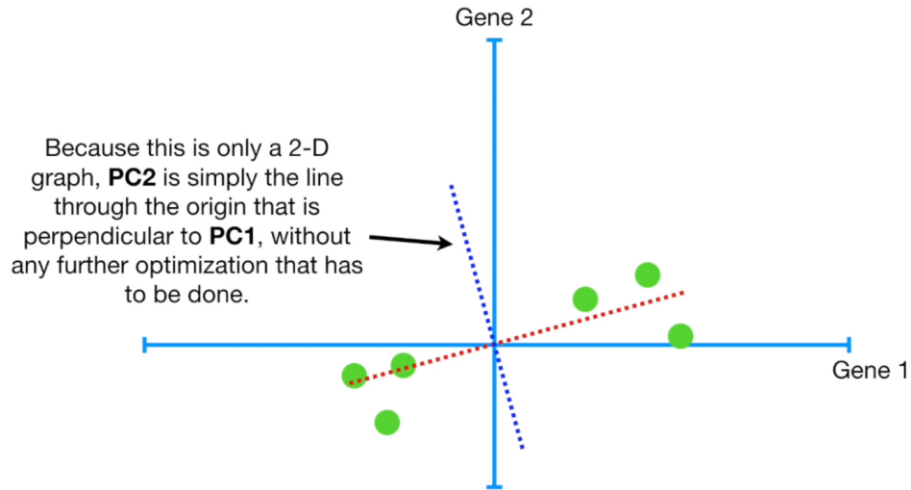
PCA

VPC = Variation for PC1

SSPC1 = SS(distances for PC1)

n = Number of data points

PVE = Porcentage of Variance Explained

$$VPC1 = \frac{SSPC1}{n-1}$$

$$PVE = \frac{VPC1}{Sum\ of\ VPC}$$

PCA

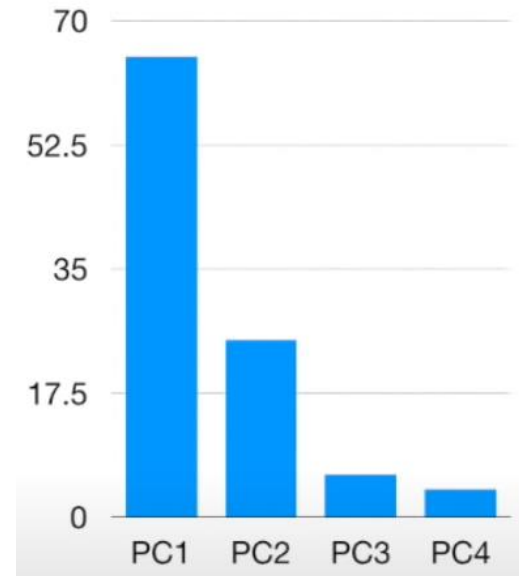Let's suppouse that we have the next values of PVE

PC1 = 69%

PC2 = 21%

PC3 = 6%

PC4 = 4%

With those values we can use just the PC1 and PC2
Instead of the 4 variables.

References:

- StatQuest: Principal Component Analysis (PCA), Step-by-Step
- A One-Stop Shop for Principal Component Analysis
- A STEP BY STEP EXPLANATION OF PRINCIPAL COMPONENT ANALYSIS
- Principal component analysis (PCA): Explained and implemented

Disclaimers