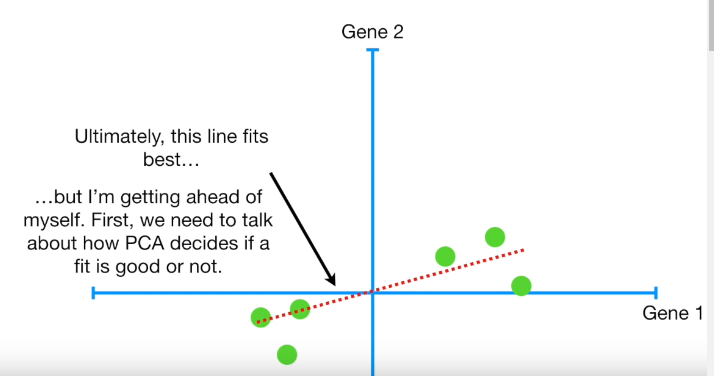
Principal Component Analysis- Using Singular Value Decomposition

Explanation from Statquest

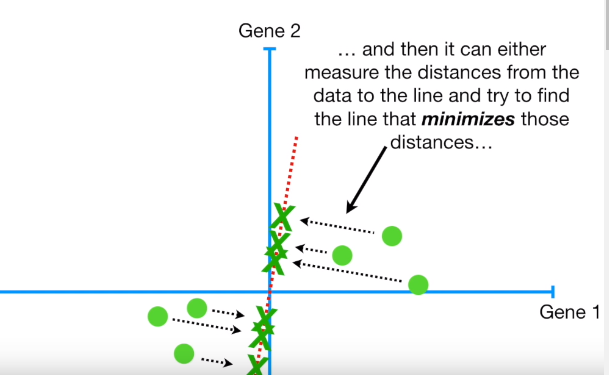
* Similar observation cluster together.
* PCA tells us which variable is the most valuable to cluster the data.

With the average values we can calculate the center of the data. Shift the data so that the center is on top of the origin (0,0) in the graph. Shifting the data did not change how the data points are positioned relative to each other.

After is centered on the origin we can fit a line to it. (A random line that goes through the origin and then it rotates so it fits the data.

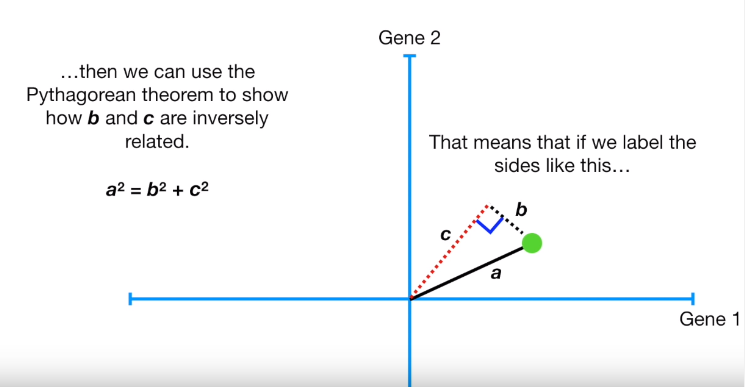


To quantify how good this line fits the data PCA projects the data onto it and then it can either measure the distances from the data to the line and try to find the line that minimizes those distances or it can try to find the line that maximizes the distances from the projected points to the origin.



These distances shrink when the line fits better.

The distance from the point to the origin doesn’t change when the line changes. When we project the points we get a right angle between the line that is projected to the origin and the one that is projected to the fitting line.



Since a (and thus a^2 doesn’t change if b gets bigger then c must get smaller. Likewise if c gets bigger then b must get smaller. Thus PCA can either minimize the distance to the line or maximize the distance from the projected point to the origin.

But it is easier to calculate c, the distance from the projected point to the origin, so PCA finds the best fitting line by **maximizing the sum of the squared distances from the projected points to the origin.**

We measure all the distances from the data points and then we squared so negative values don’t cancel out positive values then we sum up all squared distances

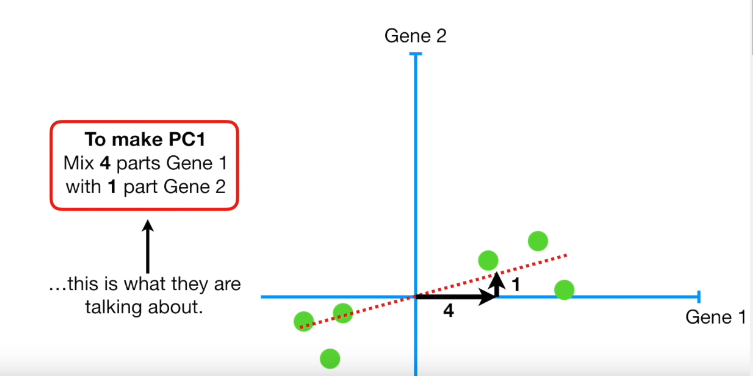
=sum of squared distances SS(distances)

We repeat until we end up with the line with the largest sum of squared distances between the projected points and the origin.

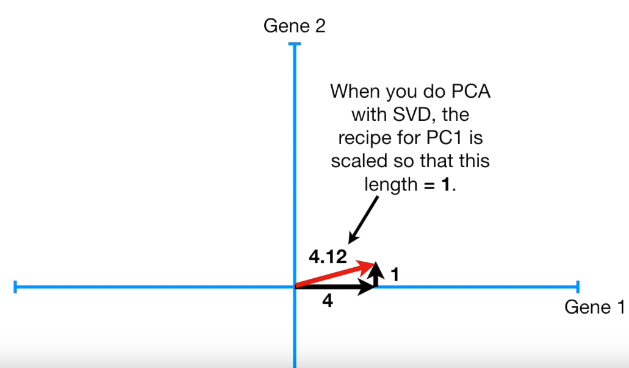
This line is called PC1

The ratio of one factor to another (if it is bigger than 2) tells you that the factor is more important when it comes to describing how the data are spread out.

PC1 is a linear combination of variables.



When using SVD then the length of the sides are scaled so that



You have to divide all three sides by 4.12 to make that side equal to one. The new values change the recipe.

* The data
* The best fitting line
* The unit vector that we calculated

This one unit long vector is called the “**Singular vector”** or the “**Eigenvector”** for **PC1.**

**SS(distances for PC1) = Eigenvalue for PC1**

**(Eigenvalue for PC1)^(1/2) = Singular value for PC1**

PC2 is the line through the origin that is perpendicular to PC1, without any further optimization that has to be done.

**Principal Component with more than 2 variables**

Center the data

Find the best fitting line that goes through the origin. For PC2 find that line that is perpendicular to PC1, and then for PC3 find the line that is perpendicular to both PC1 and PC2. If we have more variables we’d just keep on finding more and more principal component by adding perpendicular lines and rotating them.

In theory, there is one PC per variable, but in practice, the number of PCs is either number of variables or the number of samples, whichever smaller.

Once you have all the principal components figured out, you can use the eigenvalues (SS(distance) to determine the proportion of variation that each PC accounts for.

