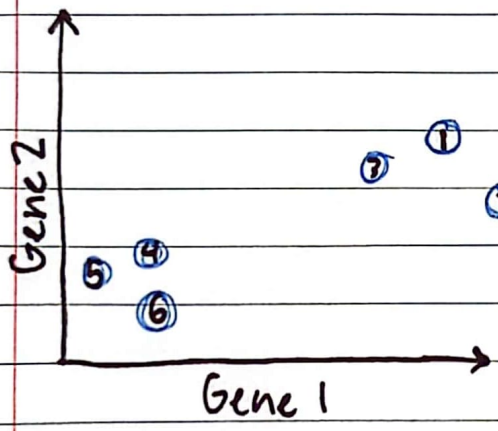
PRINCIPAL COMPONENT ANALSIS (PCA)

PCA is useful for reducing the dimensions of high dimensional data. It transform the data onto axes that account for the most variation, which are linear combinations of the original data.

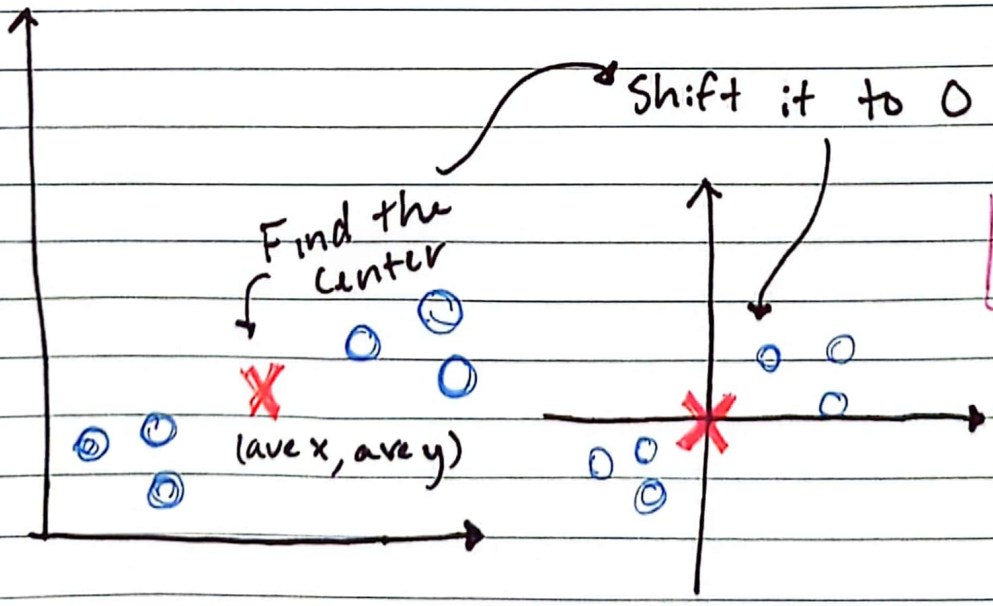
Example/Steps:

	Sample 1	sample 2	sample 3	sample 4	sample 5	sample 6
Gene 1	10	11	8	3	1	2
Gene 2	6	4	5	3	2.8	1



There can always be more dimensions i.e. gene 3, 4, etc

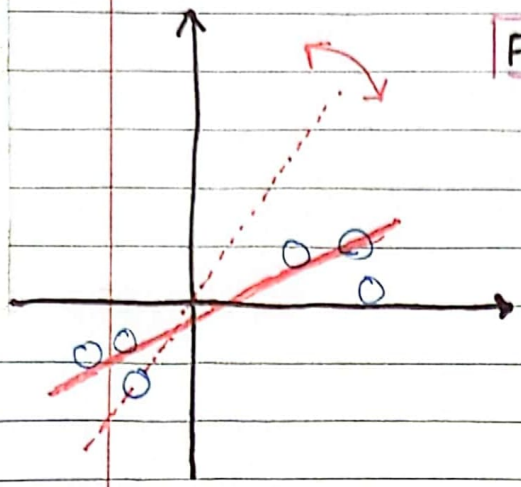
Similar samples cluster PCA can help decide which variable seperates them the most.



Centering Data

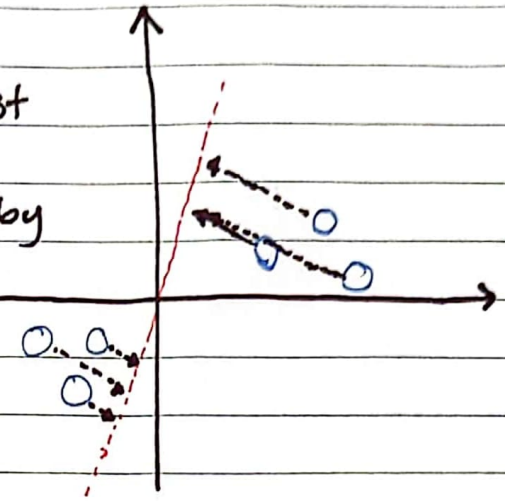
* Example from Statquest

CamScanner

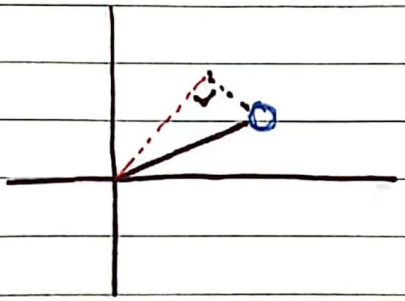


Fit a line through the points
it has to go through the origin

The best line is Chosen by projecting the points onto the line

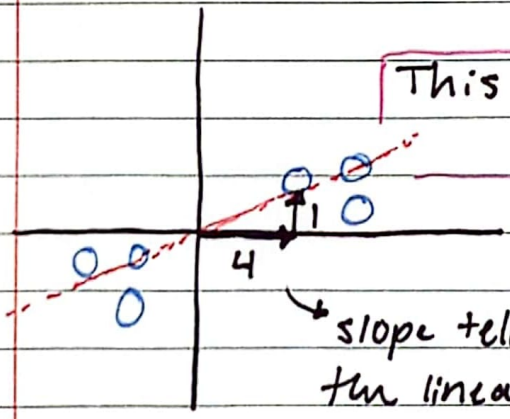


The best line minimizes the distances from the point to the line / maximizes the distance from the origin



The distances to the origin:

$d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2 + d_6^2$ = sum of squared distances

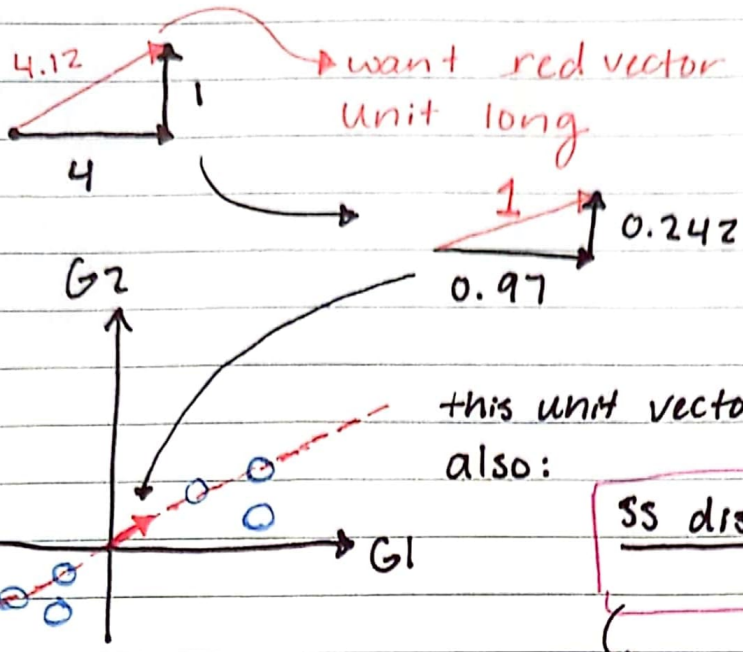


This first line is principal component 1 : PC1

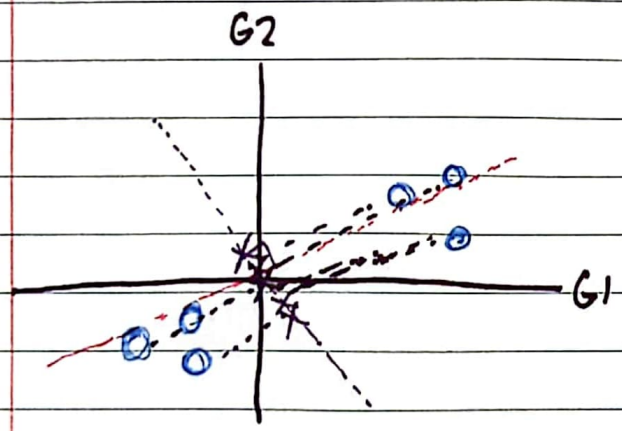
slope tells you about the linear combination

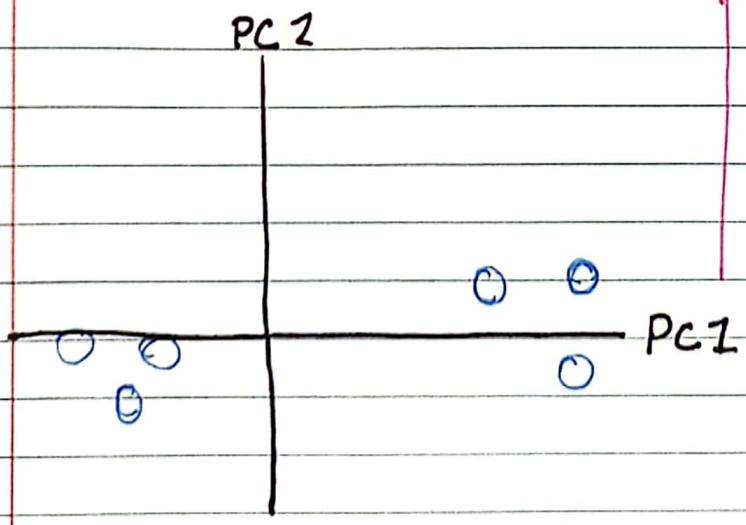
4 Gene 1
1 Gene 2

CamScanner



On a 2D graph, PC2 is just the line ⊥ to PC1





Back to the eigenvalues:

$\frac{\text{Sum Squared dist PC1}}{n-1}$ = eigenvalue = variation for PC1 = 15

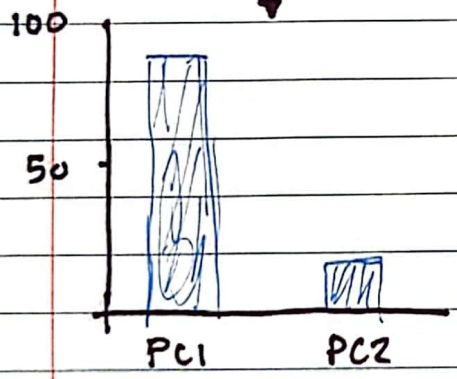
$\frac{\text{Sum Squared dist PC2}}{n-1}$ = eigenvalue = variation for PC2 = 3

So, $15+3=18$ $\quad \frac{15}{18} = 83\%$ $\quad \frac{3}{18} = 17\%$

PC1 accounts for 83% of the variance

PC2 accounts for 17% of the variance

You can plot these on a Scree plot

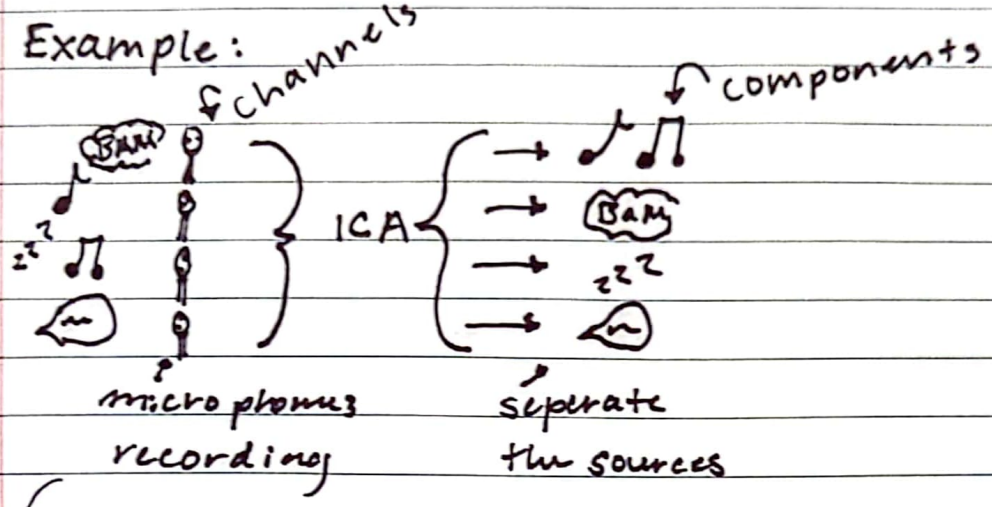


For higher order data, the steps are the same but it's harder to visualize → there will be 1 PC for each dimension. You take the PCs that account for most of the variation & use that as your plot.

INDEPENDENT COMPONENT ANALYSIS (ICA)

ICA is used to seperate linearly mixed signals from multiple sources. It finds components that are statistically independent from each other.

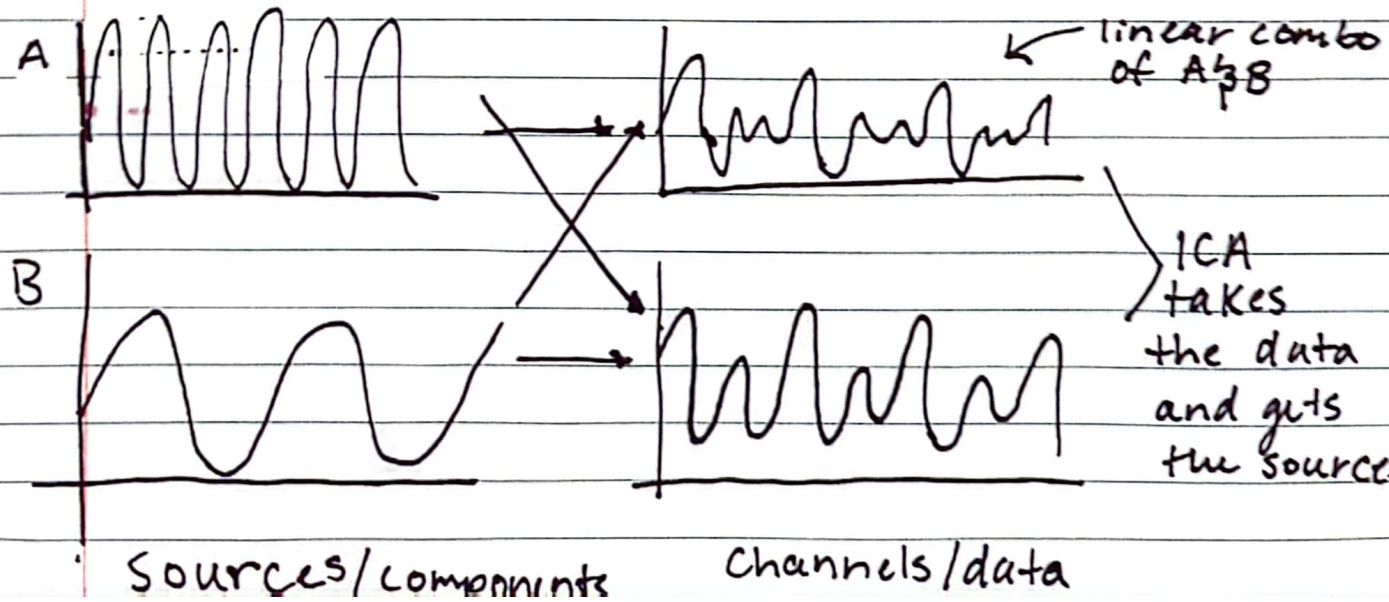
Example:



So, the goal is to uncover the original version by using matrices to "unmix" the data

$$U = WX$$

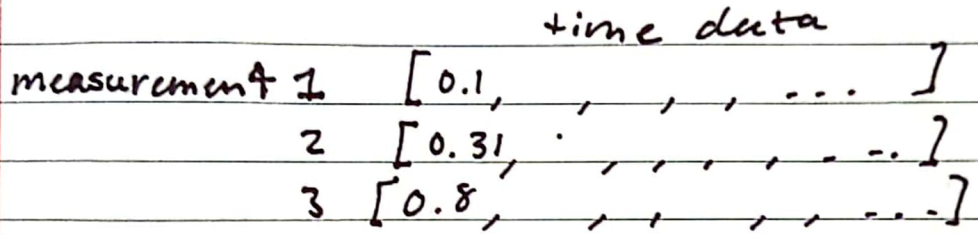
X : the data (channels x time)
U : the ICA source activities (component x time)
W : the unmixing matrix (component x channel)



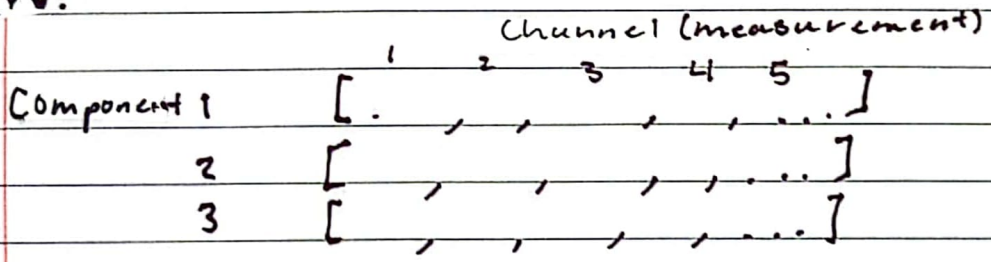
Before doing ICA, you need to pre-process the data. The first step is to whiten the data, this means removing correlations in the data → different channels are made to be uncorrelated

ICA can be preformed on high # of dimensions. The whitened data is rotated to minimize the gaussianity of the projection onto each axis.

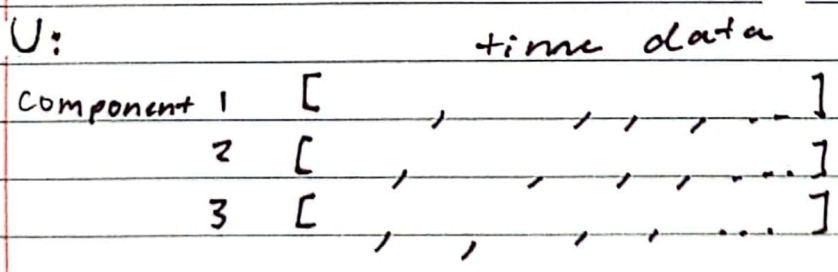
X:



W:



U:



Notes:

- can only be used on linearly mixed sources
- perfect gaussian sources cannot be seperated
- even if sources are not completely independent ICA will find how they maximinally are independent.

CamScanner