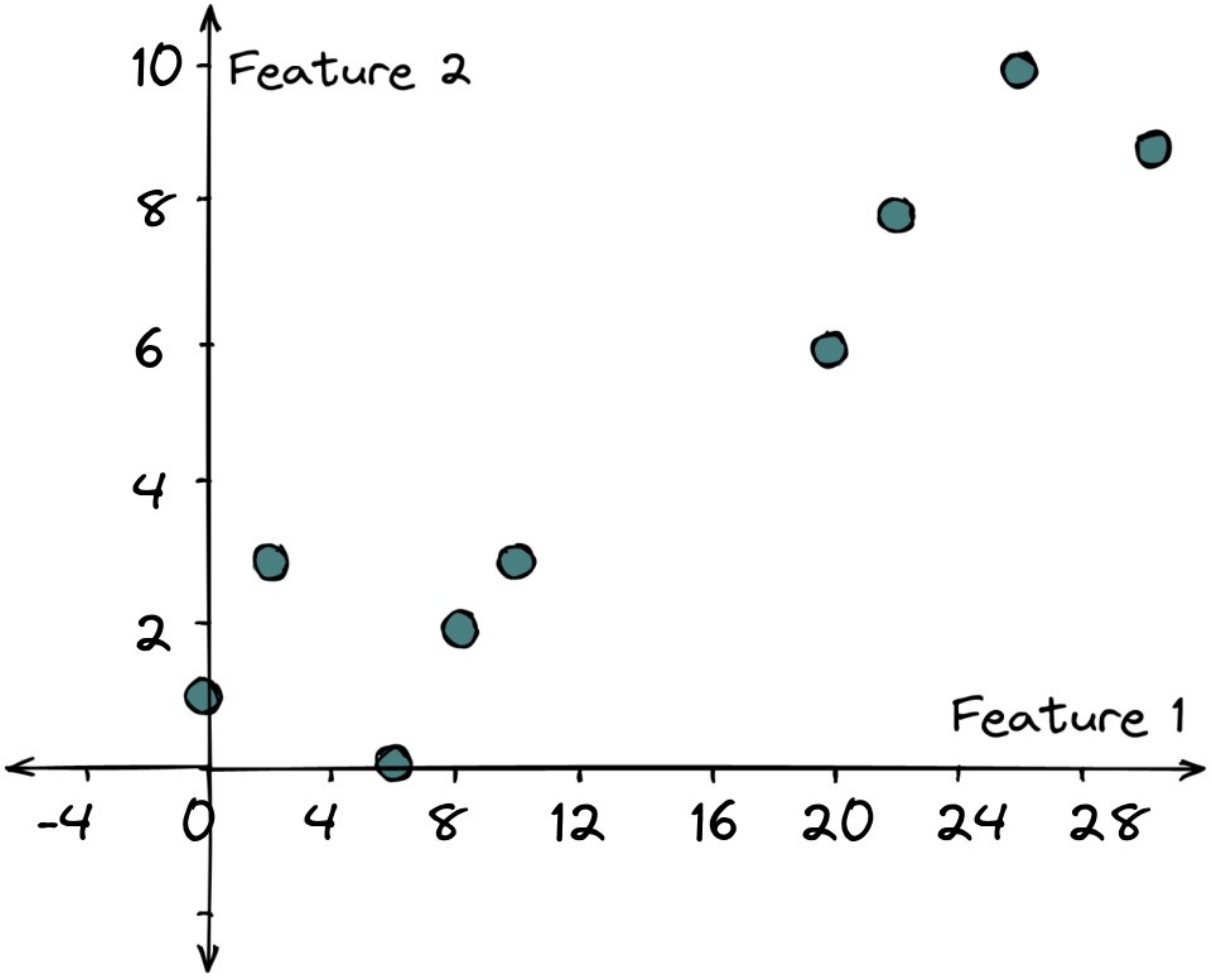
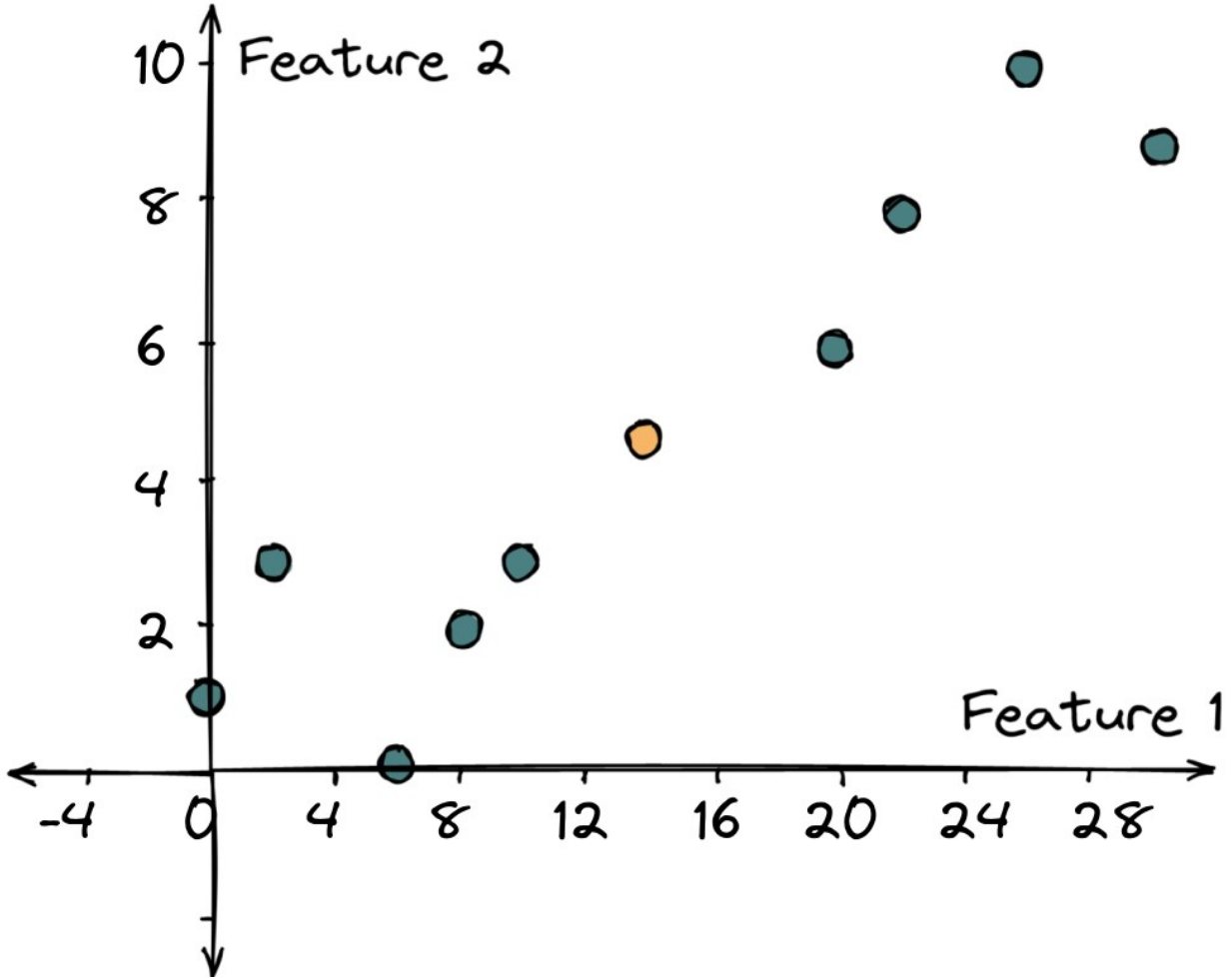
PCA. Step1: Plot Data

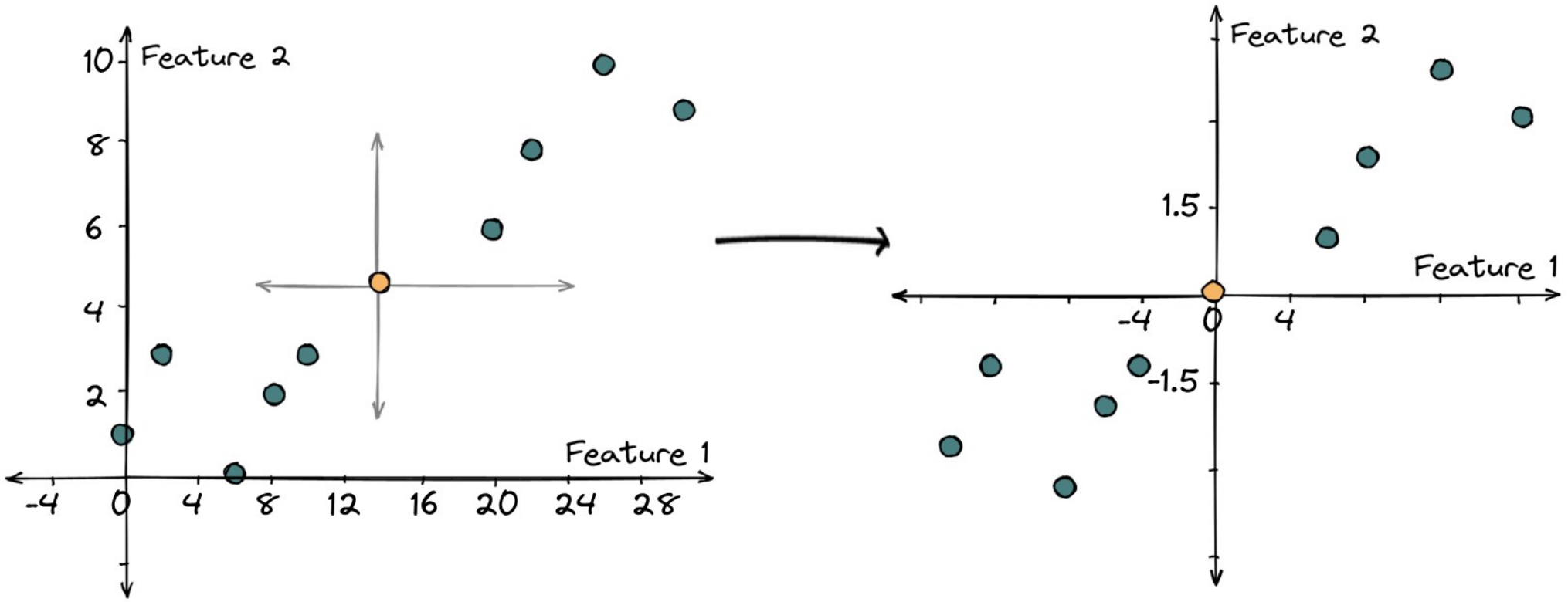


Feature 1	Feature 2
0	1
2	3
6	0
8	2
10	3
20	6
22	8
26	10
30	9

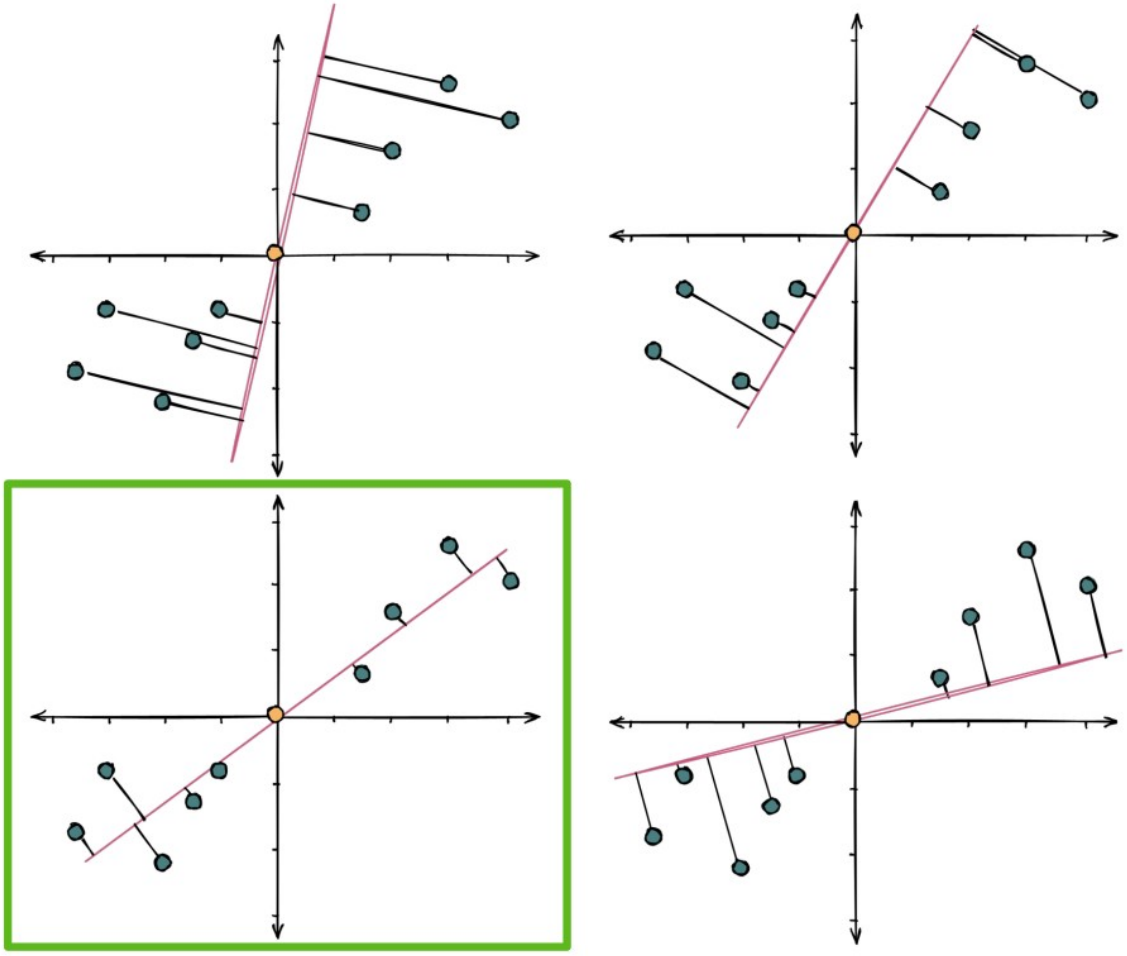
PCA. Step2: Plot Center of Data



PCA. Step3: Recenter Data



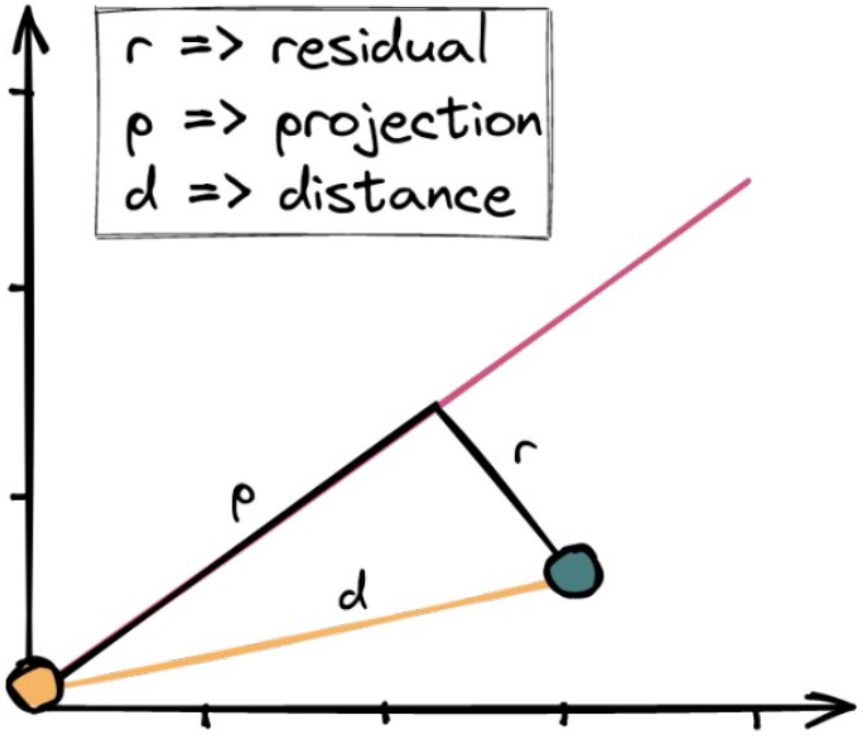
PCA. Step4a: Find the Best Fit Line

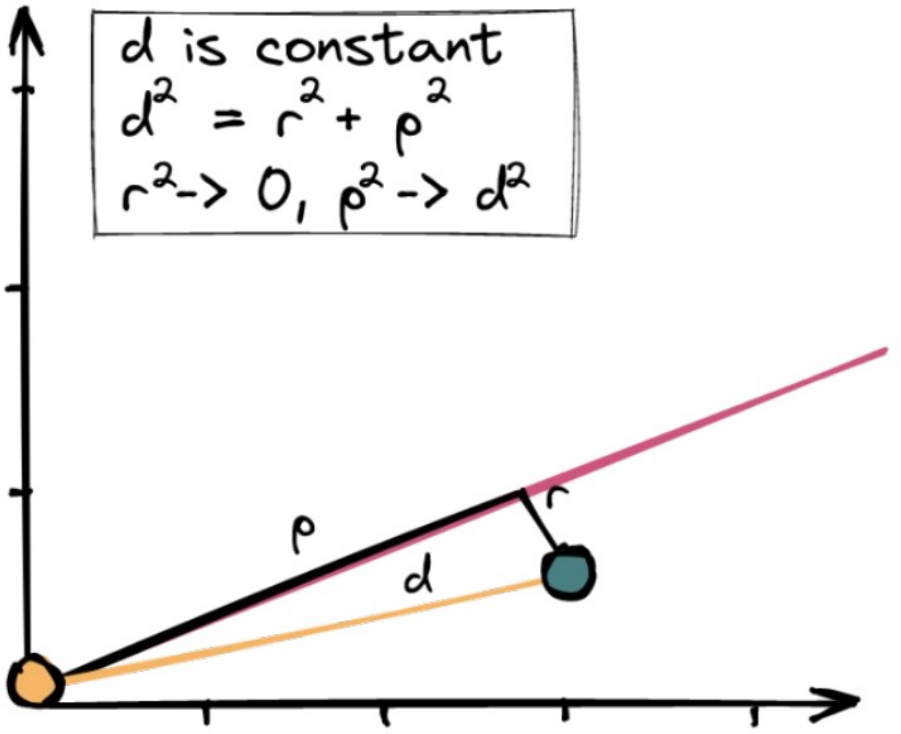


We look for the best fitting line. Classic way is to use least square method.

The resulted line is the first principle component (PC) => PC1

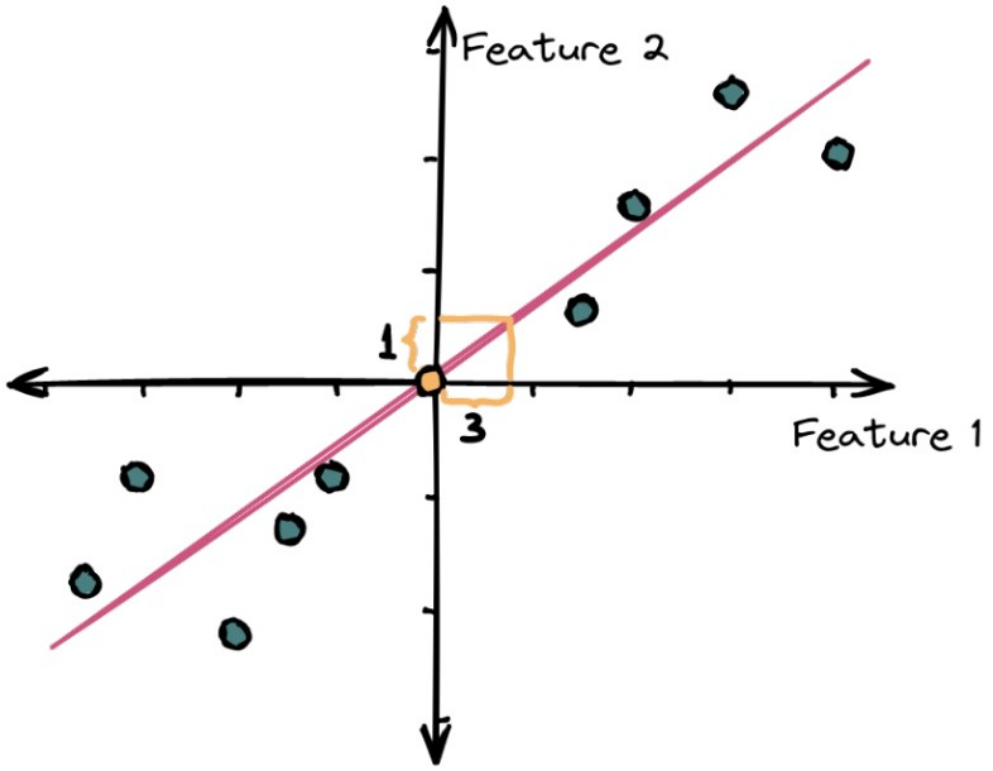
PCA. Step4b: Find the Best Fit Line

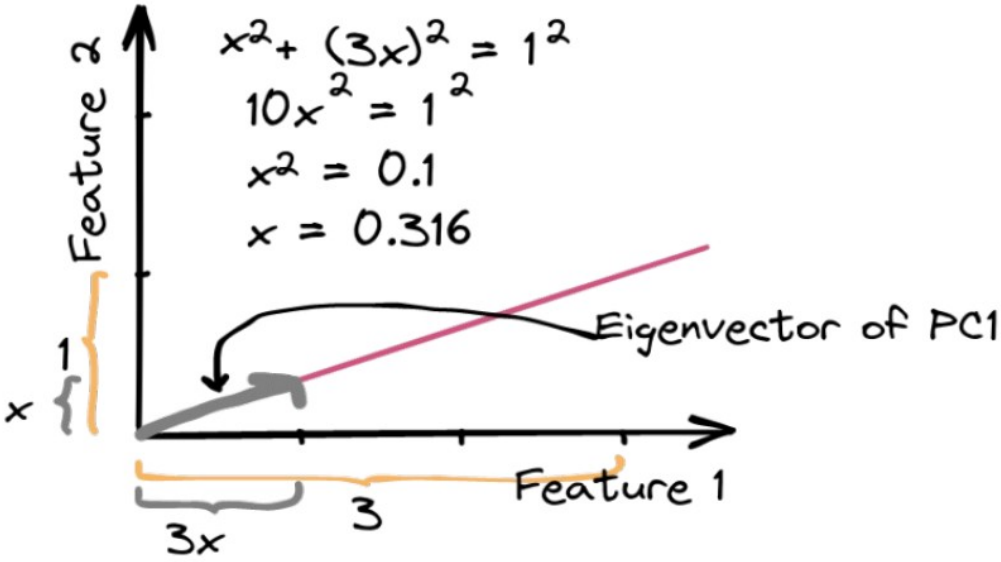




Instead of looking for minimal **sum**(r_i^2), we can look for maximum **sum**(ρ_i^2)

PCA. Step5: Eigenvector & Eigenvalue

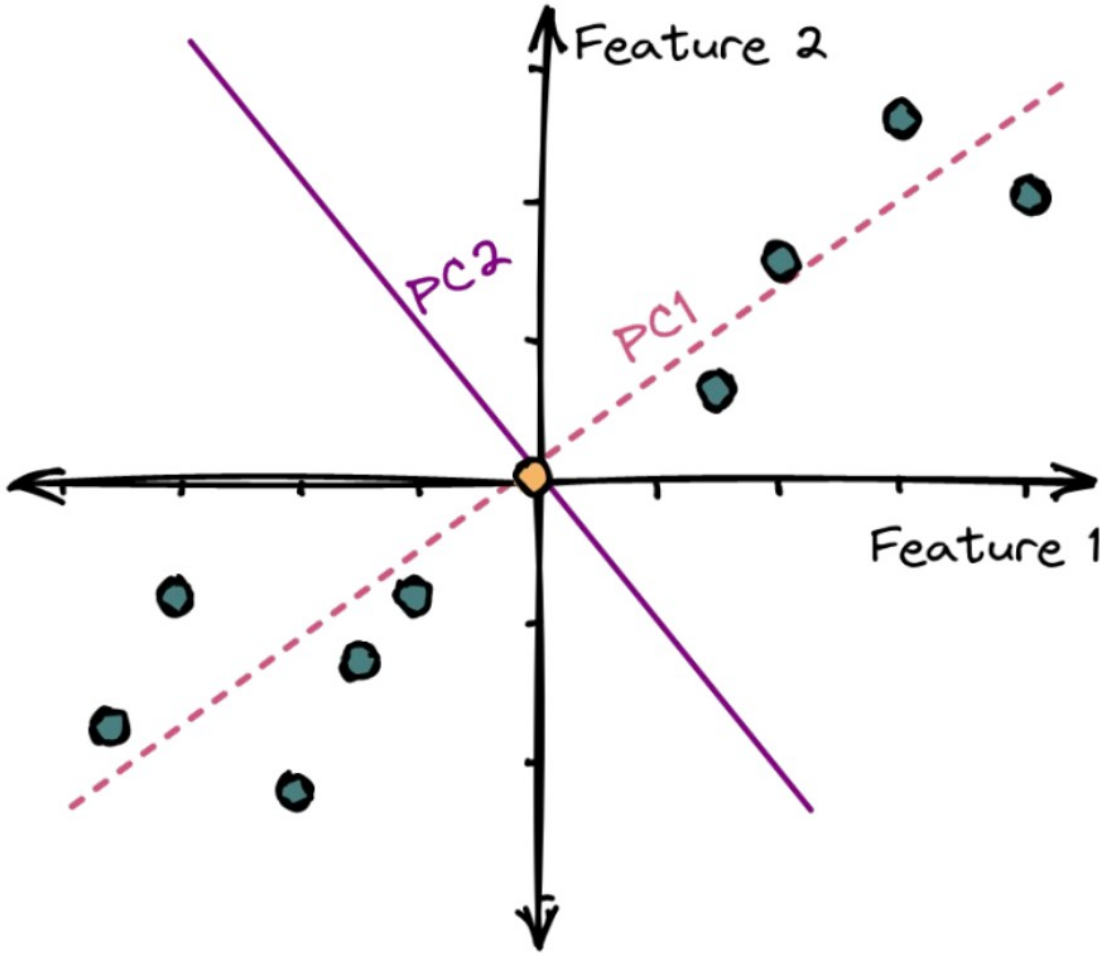




PC1 consist of:
0.316 of Feature 2
0.316*3 = 0.948 of Feature 1

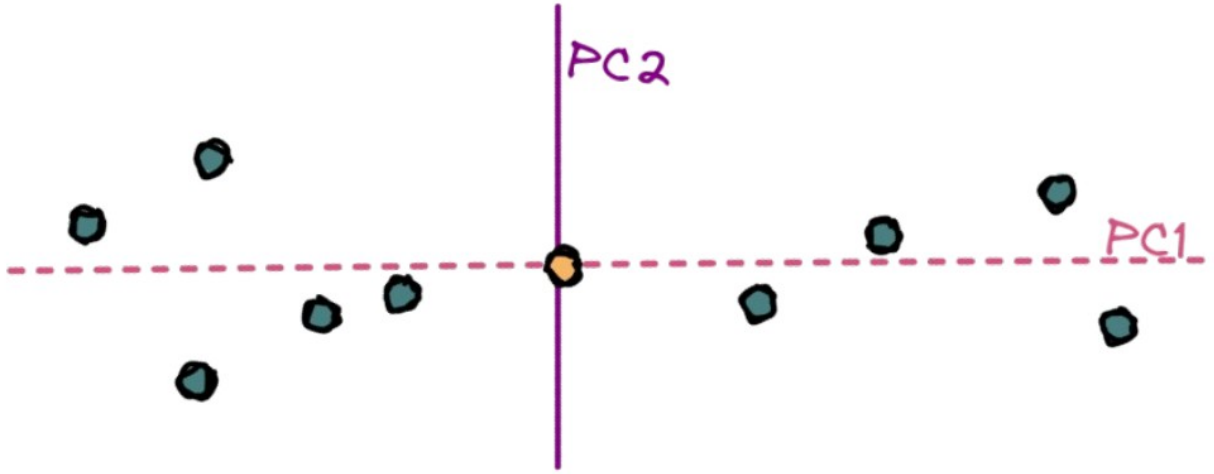
Eigenvalue => sum of ρ^2
Single value = $\sqrt{\text{Eigenvalue}}$

PCA. Step6: Principal Component 2



Second principle component (**PC2**) can be found as a perpendicular to **PC1**

PCA. Step7: Variations



$$\frac{\text{sum of } \rho^2 \text{ (PC1)}}{n-1} = \text{(variation for PC1)} = 130$$

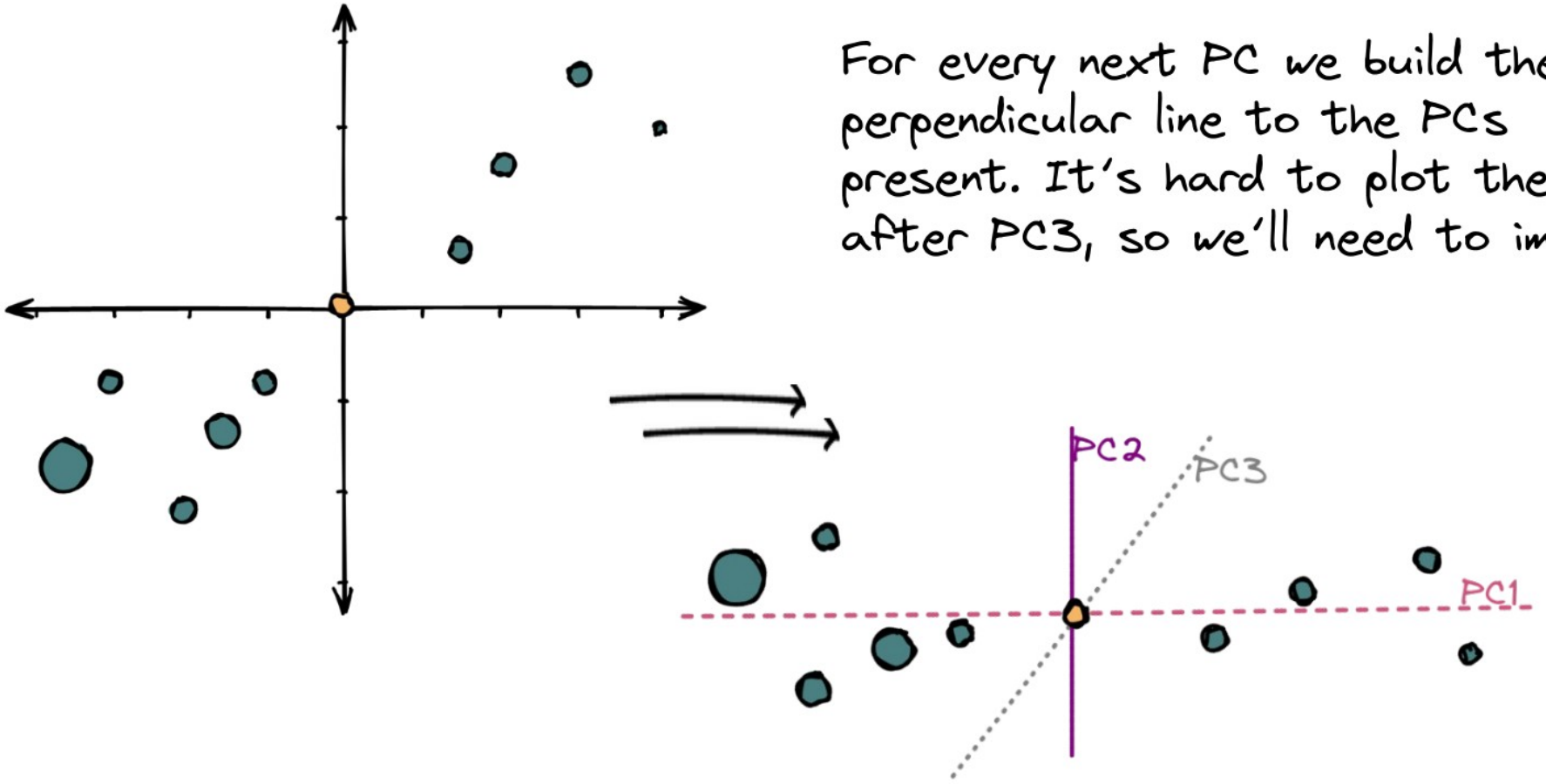
$$\frac{\text{sum of } \rho^2 \text{ (PC2)}}{n-1} = \text{(variation for PC2)} = 1.6$$

PC1 variation = 130/(130+1.6) = 0.987 = 98.7%

Now we don't need Feature1 and Feature2 anymore. By rotating PC1-PC2 we result in the plot we're used to.

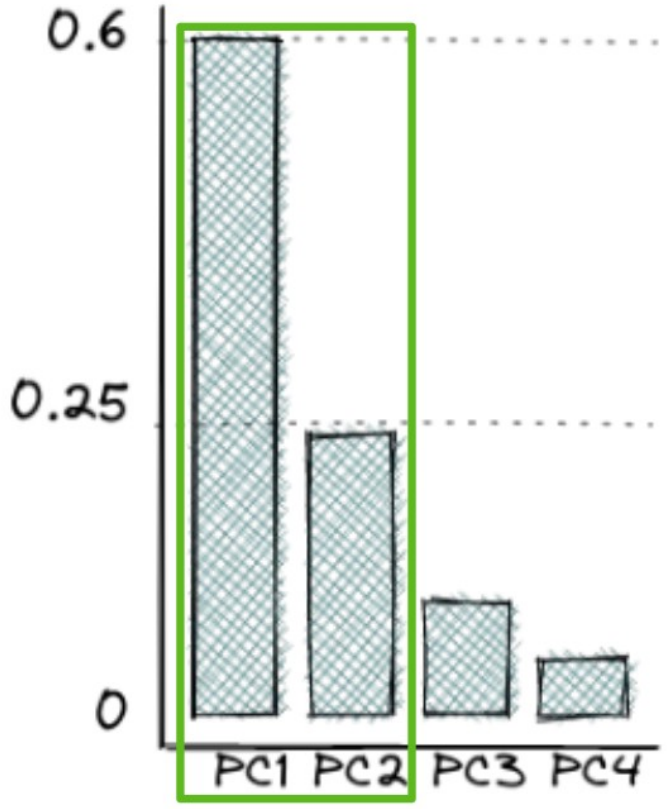
The PC variations allows us to understand importance of principal components in explaining data

PCA. Step8: PC3, PC4, PCn, etc



For every next PC we build the perpendicular line to the PCs present. It's hard to plot the PCs after PC3, so we'll need to imagine.

PCA. Step9: PC Importance



Imagine you've built 4 PCs.
The resulted variations are:
PC1 = 0.6
PC2 = 0.25
PC3 = 0.1
PC4 = 0.05

Makes sense to leave only 2PC, as they represent 85% of the variation.
Now, instead of 4+ features we've reduced their number to 2

https://www.linkedin.com/in/reznikovivan/