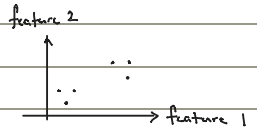
Principal Component Analysis

- Using Singular Value Decomposition (SVD)
- Get deeper insights into data
- Anything > 3D cannot be plotted
- Can be used to identify feature most valuable for clustering
  - ↳ which feature produces most clear split
- PCA can tell how accurate 2D PCA plot is.
- n features yields n principal components OR number of data samples whichever is smaller

How PCA works (2 features)

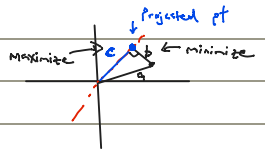


1) Plot all samples
2) Find average wrt both features $\bar{X}_1$ & $\bar{X}_2$
3) $(\bar{X}_1, \bar{X}_2)$ represents center of data
4) shift center to origin
5) fit line to data
   - 5.1) start with random line passing origin
   - 5.2) Rotate & repeat until best fit line (must go through origin)

Best fit: 1) Measure dist from data to line & minimize
→ 2) Measure dist from origin to projected points & maximize } Both seek to get best fit line.

PCA uses this technique because easier to find dist from origin to projected point

$\sum_{i=1}^{n} (d_i - 0)^2$ d_i = projected point on line.



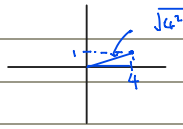
if ↓
$a^2 = b^2 + c^2$
dist. from point to origin (fixed).
then c ↑. (aka maximize)

6) The best fit line for $(\bar{X}_1, \bar{X}_2)$ is known as Principal Component 1

Best fit line: $y = 0.25x$ ← Linear combination of X_1 & X_2

$X_2 = 0.25 X_1$

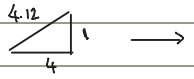
1 unit up = 4 unit up. (data spread out in X_1) but more compact in X_2.



$\sqrt{4^2 + 1^2}$ = 4.12 (PC1) (more impt feature to describe spread of data)

* People usually say PCA is a linear combination of features.
* PC line always passes through origin & scaled to 1.
* typically scaled to 1 by dividing all sides with calculated magnitude.

PC 1 (original) PC 1 (scaled) * Ratio unchanged.



Singular vector (Eigenvector for PC1)

* Eigenvector is a unit vector in the direction of the Principal Component

To make PC1 :	← Loading scores
$0.97 X_1$	
$0.242 X_2$	

Singular Value for PC1

Singular value = $\sqrt{\sum_{i=1}^{n} d_i}$

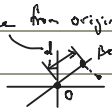
↑ Sum of squared dist for PC1 best fit

Eigenvalue for PC1

- Average of the sum of squared dist for best fit line

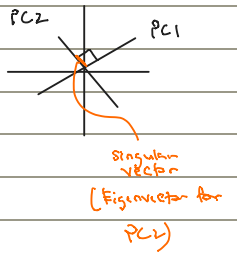
$\sum_{i=1}^{n} d_i \cdot \frac{1}{n-1}$ = Eigenvalue for PC1

↑ distance from origin to projected point



After finding PC1 : Find PC2.

if only 2 features, PC2 is simply the orthogonal of PC1
- still has to pass the origin



To make PC2:

$-0.242\ X_1$
$0.97\ X_2$
} Loading score for PC2 : $X_2 \approx 4\times$ more impt than X_1

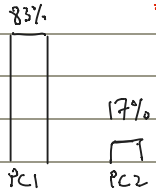
Final PCA plot

1) Rotate such that PC1 is horizontal
2) use projected points to plot data on PCA plot
   - were all previously on best fit line.

* Eigenvalues are measures of variation ← used to determine proportion of variation for each PC.

Eigenvalue 1 = variation for PC1 etc.

if Eigen Val 1 = 15 (83% of variation)
Eigen Val 2 = 3 (17% of variation)
Total variation for PC = 15 + 3
= 18.



← Scree plot : Graph for variation (in percentage) for each PC.