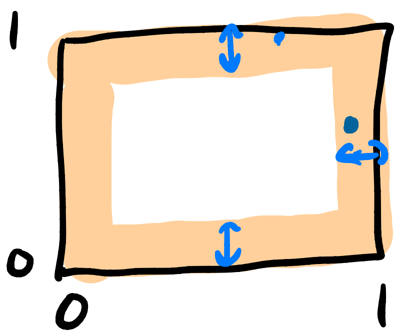
Lecture 17

Example: 10,000-D data.



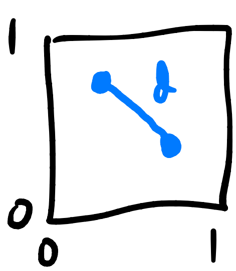
prob a point is within 0.001 of border. = 0.004

But in 10,000 D hypercube ⇒ > 0.9999!

⇒ Virtually all data becomes very extreme!

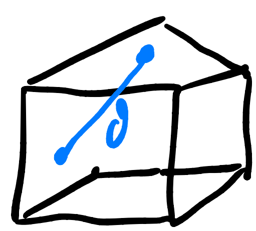
Avg distance between 2 random pts

2D:



≈ 0.5

3D:



≈ 0.66

10,000 D hypercube

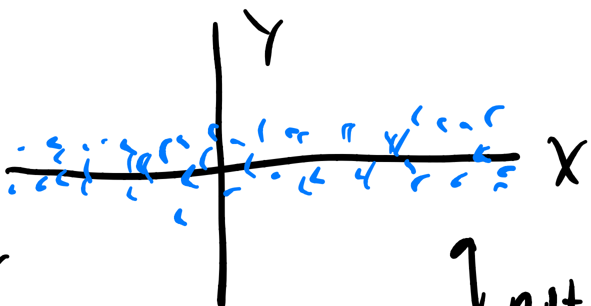
≈ 408.25 !!

higher dimension, on avg, things are more spaced out

Sparse

Curse of dimensionality: Techniques that work in lower dimensions fail in high dimensional spaces.

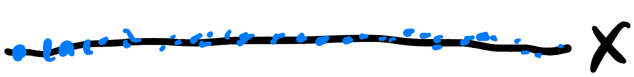
Work around:



would I lose all that much if I dropped Y coordinate/dim?

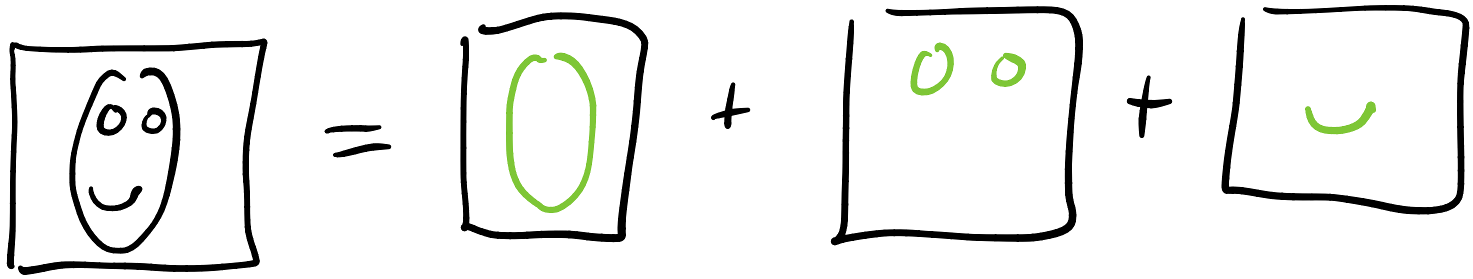
not much meaningful variation in Y, might be ok to drop.

drop Y

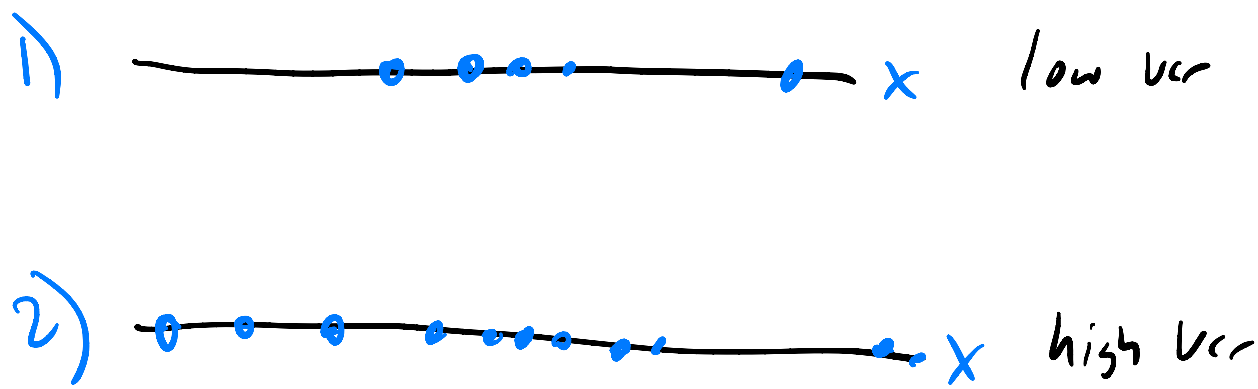


When we have 10,000 Dimension, maybe 30 variables are actually meaningful — find out how to reduce

$10{,}000\,D \xrightarrow[\text{project}]{} 30\,D$ → run analysis on 30 meaningful vars principal component analysis (PCA): popular technique to reduce dimensionality — algorithm to tell us which variables are important.

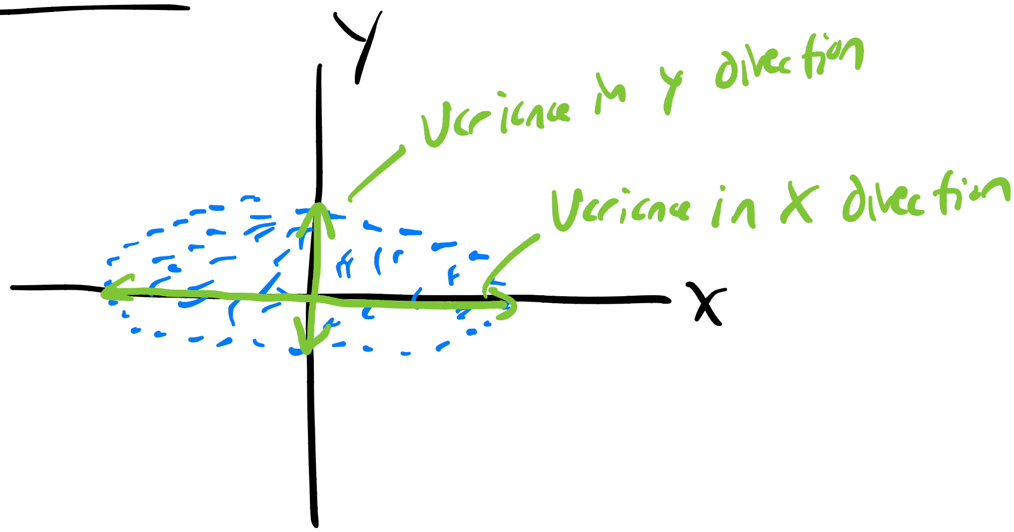


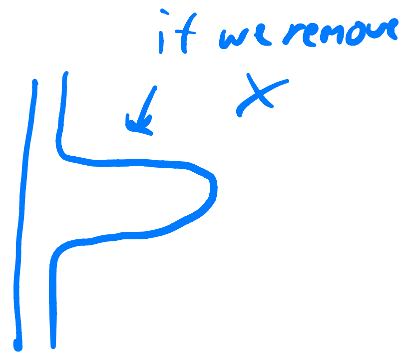
Covariance matrix



Variance: measure of spread/dispersion in data
scalar value

2D data





$$Var(x) > Var(y)$$

Covariance matrix: Store info about variances among variables

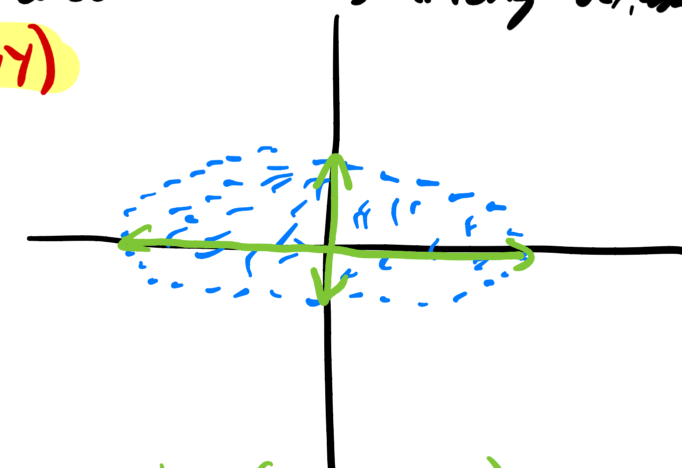
σ_x^2 $\quad$ $cov(x,y)$

$$\Sigma = \begin{array}{c|cc} & x & y \\ \hline x & 2 & 0 \\ y & 0 & 1 \end{array}$$

σ_y^2

$cov(y,x)$

Variance within each variable (x, y, etc)



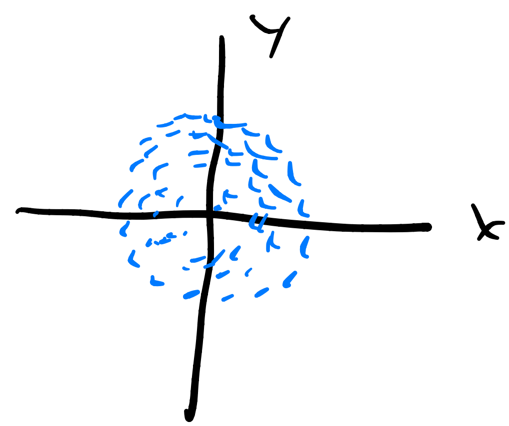
Off diagonals: How vars vary together jointly

Covariance: $\underbrace{cov(x,y)}_{\text{above it is } 0}$

★ $\boxed{Cov(X,Y) = Cov(Y,X)}$

Covariance matrix always must be symmetry

$$\Sigma = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}$$



$$\Sigma = \begin{bmatrix} 3 & 1 \\ 0 & 3 \end{bmatrix}$$

✓ impossible! not symmetry

$$\Sigma = \begin{bmatrix} 20 & +5 \\ +5 & 1 \end{bmatrix}$$