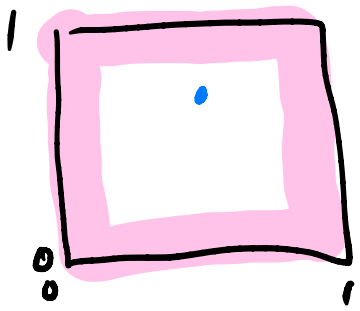
Project 4: Principal Component Analysis

Curse of dimensionality

Data behaves very differently in a high dimensional space (e.g. 10,000-D) compared to 2D/3D.

- **prevalence of extreme values**



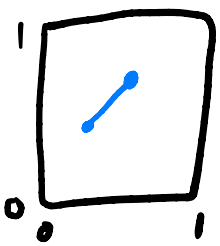
2D unit square:

probability pt within 0.001 of border = 0.004

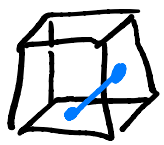
10,000-D hypercube: same probability >0.9999!

⇒ In high dimensional spaces, almost all points are "extreme".

- **Distance inflation:**



2D unit square: avg dist between 2 pts ≈ 0.5



3D unit cube: avg dist ≈ 0.66

[You can simulate this to verify!]

10,000-D hypercube: avg dist between 2 pts

$\approx$ 408.25!

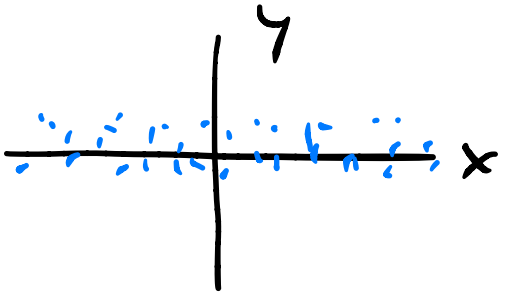
$\Rightarrow$ Things are much farther apart in high dimensional spaces!

↳ Space is sparsely filled with data

- Curse of dimensionality: Refers to these problems.

Difficult to analyze data in high dimensional spaces — techniques that find patterns in 2D/3D may not generalize to 10,000-D.

Work around for Curse of dimensionality:



Could we just toss out y dimension here without running into problems?

- Real data not evenly distributed across all variables — Not all 10,000 vars carry equal weight — Maybe only 30 do.

Project data on those 30

↓

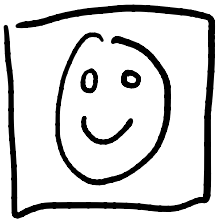
run analysis on those

↓

Run analysis on 30-D data, not 10,000-D.

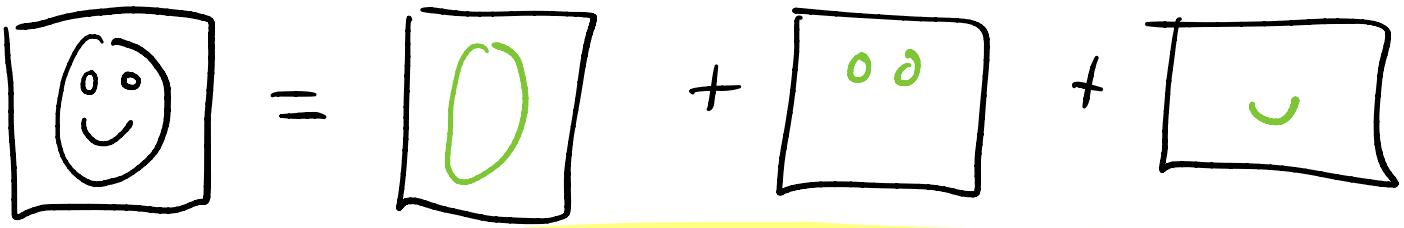
principal component analysis (PCA) is a popular technique for doing this dimensionality reduction.

In project, you will apply PCA to images of human faces:



PCA allows us, without reducing # pixels, to discover intrinsic variables in image data that are important (e.g., eyes, mouth, head)

⇒ Can express an image as linear combination of these variables:



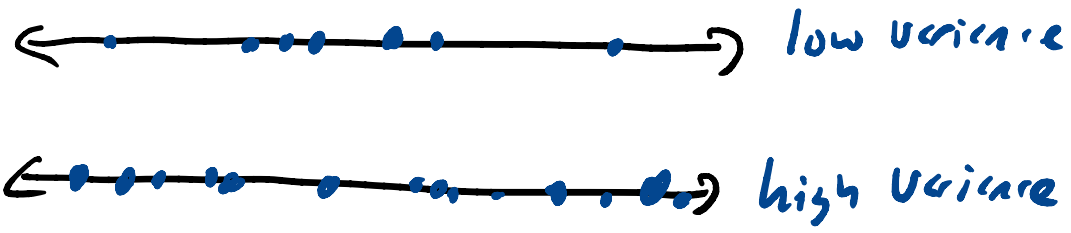
[Show Eigenface examples]

Application: Facial recognition

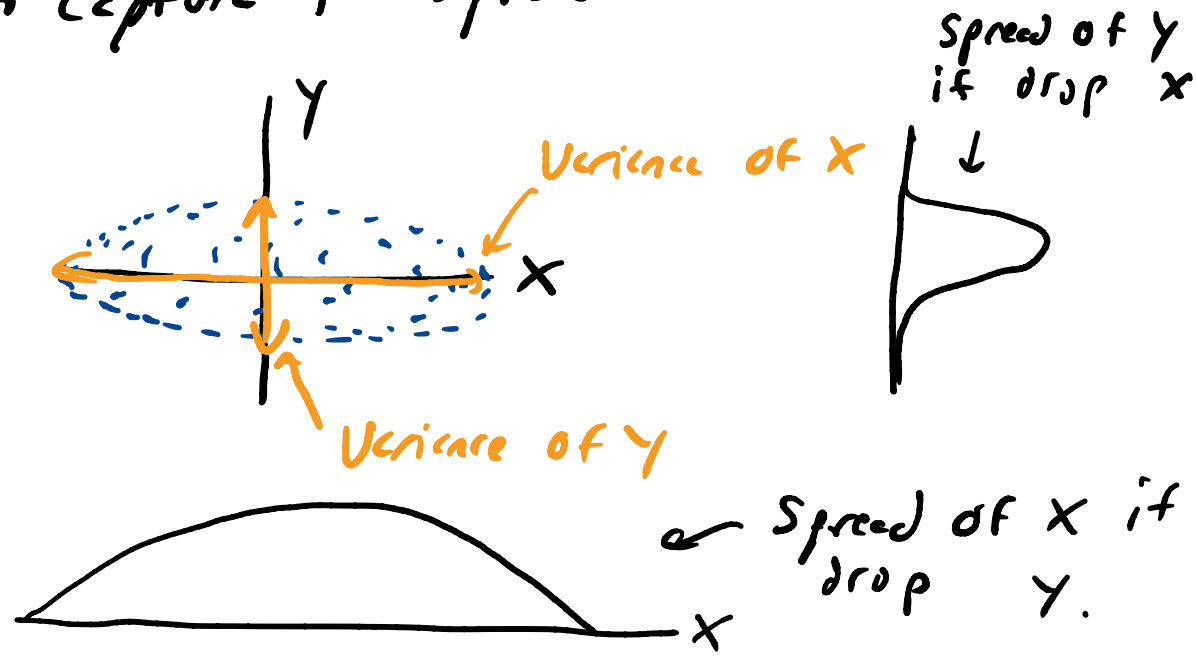
PCA depends on something called the **covariance matrix** — Let's introduce this concept 1st before diving into PCA.

Covariance Matrix

For 1D data, the ordinary sample variance describes the spread of data:



For 2D data, we can't just have a single variance to capture the spread



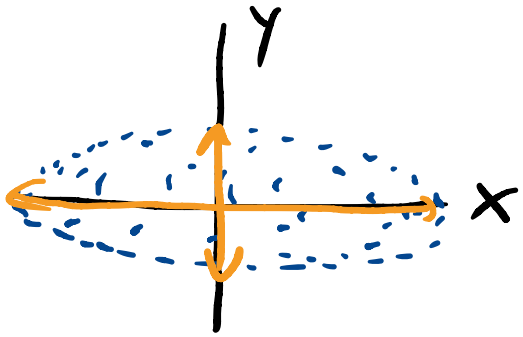
Clearly Variance (X) > Variance (Y).

Instead of a scalar to describe the variance of 2D data or higher dimensional data we use a matrix — called <u>covariance matrix</u> Σ.

plausible covariance matrix for the above 2D data:

$$\Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} \begin{matrix} x \\ y \end{matrix}$$

$$\begin{matrix} x & y \end{matrix}$$



<u>Main diagonal</u>: variance <u>within</u> each variable
- e.g. 2 is Var(x)
  1 is Var(y)

$\Rightarrow$ Var(x) = 2 Var(y)

<u>Off-main diagonal</u>: variance <u>between</u> pairs of vars called <u>covariance</u>.

e.g. row 0, col 1

$Cov(x, y) = 0$

[x does not vary with y]

e.g. row 1, col 0

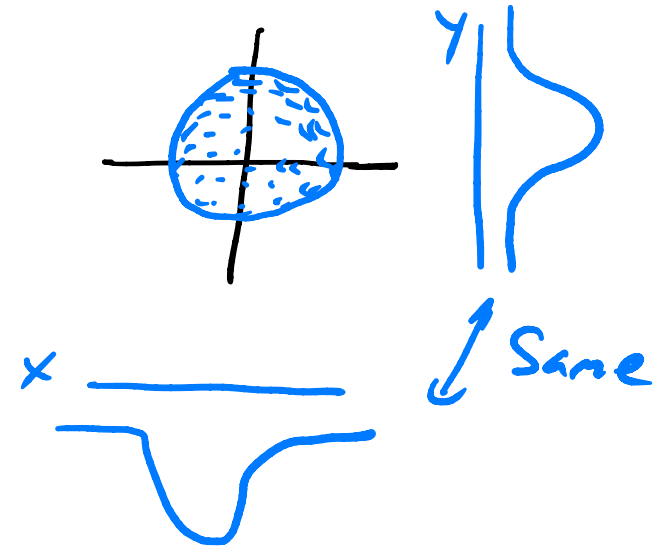
$Cov(y, x) = 0$

[y does not vary with x]

Σ, the covariance matrix, is always **Symmetric** about the main diagonal b/c $Cov(x,y) = Cov(y,x)$

picture for these Σ ?

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



$$\Sigma = \begin{bmatrix} 3 & 1 \\ 0 & 3 \end{bmatrix} \Rightarrow$$ impossible, not symmetric

positive correlation in x + y

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