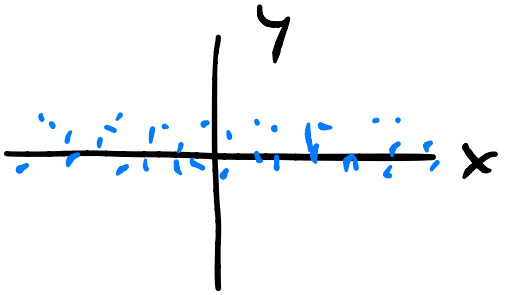


Could we just toss
Out y dimension here
without running into problems?

- 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

⇓

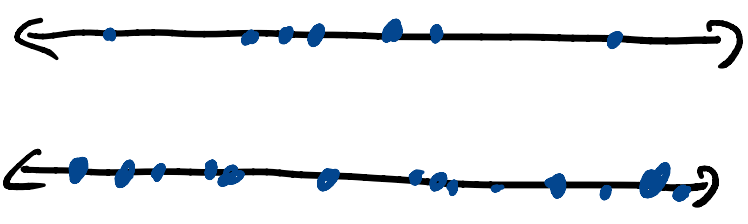
Run analysis on 30-D data, not 10,000-D.

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

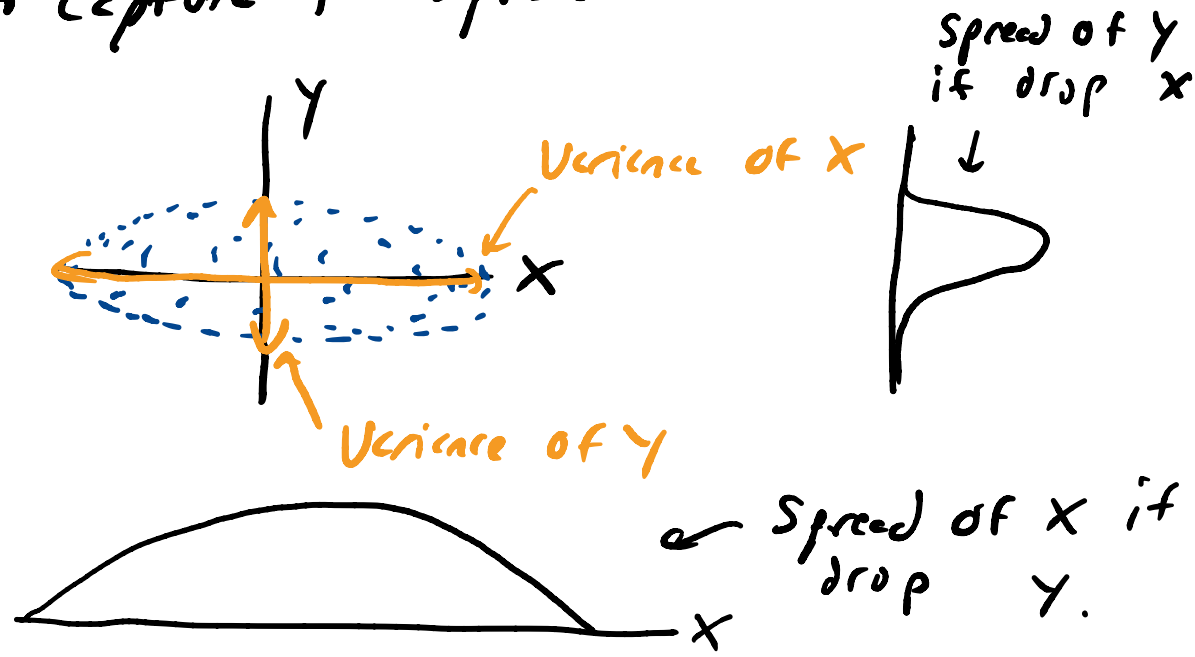
PCA depends on something called the covariance matrix — Let's introduce this concept 1st before diving into PCA.

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



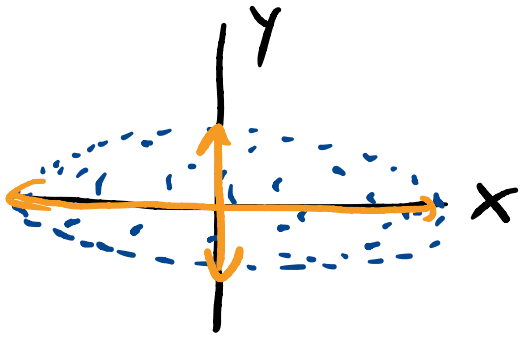
Clearly variance(X) > variance(Y).

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

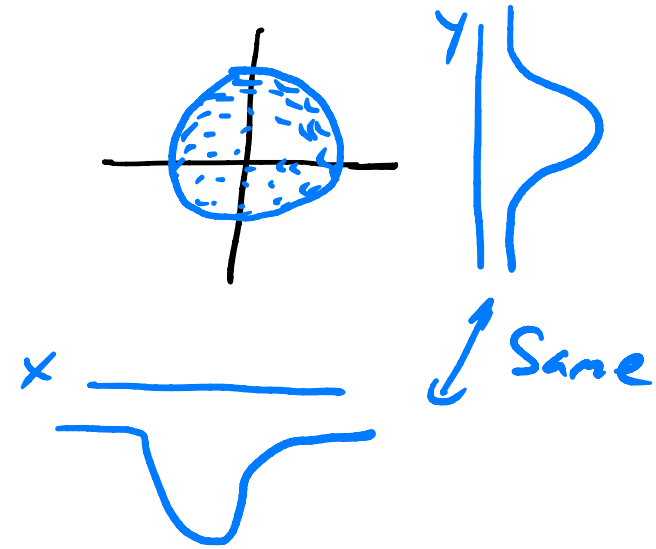
$Cov(y, x) = 0$

[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

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