Covariance Part 2

At last session we looked at variance for 1 Dim Data Set.

In the session, we looking at variance for higher Dim. data sets.

The intuition and definition of variance we had earlier does not really work in the same way in higher dimensions, and squaring vectors is not really defined.

Assuming we have a 2 Dim dataset, we can now Compute variance in the x direction and y direction.

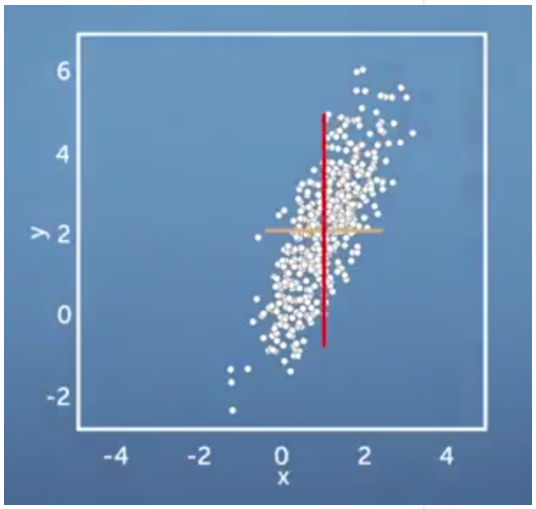
But the variance are insufficient to describe what is going on in the dataset.

In particular, we only have the variation of the data in either direction independent of the other direction, but now we also be interested in the relationship between the x and y variables.

And this is where the concept of the Co-variance between the Components come into play.

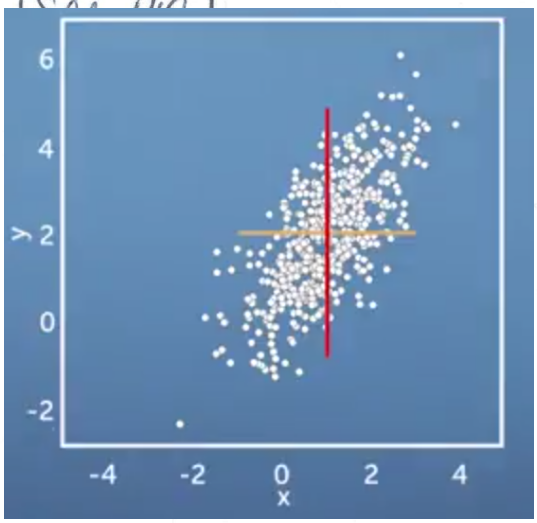
Lets have a look at an example in 2 Dim's:

For this data set we can compute the variance in x-direction, and variance in x direction, which is indicated by vertical and horizontal Bar: (as pic)

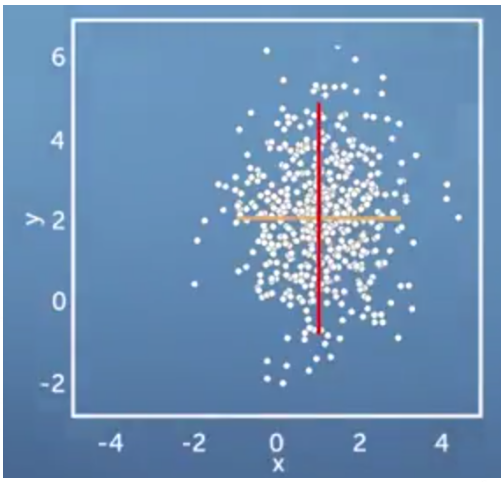


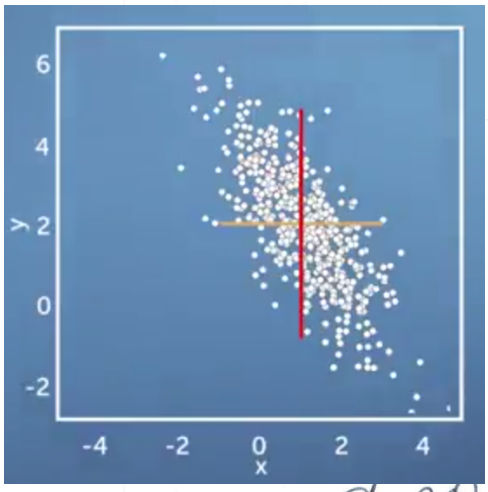
But this can be insufficient, because we can look at other eg's where the variance in x,y direction are same, but data set look very different

If we look at this particular example, we have a different shape of data set, but variance in x, y direction are exactly the same, (so are)



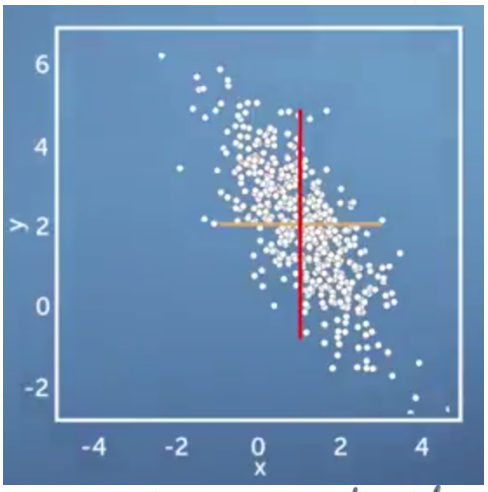
and mean values are also identical.
And I can look at different sets like this one here, and this one (these 4 data set looks very different)





You've seen 4 different eg's, with 4 different pairs or shapes of data set, but variance in x and y, and mean value are identical.
If we exclusively focus on the horizontal and vertical spread of data, we can't explain any correlation between, x and y.

In the last figure we can clearly see, that on average if the x value of data point increases, then on average the y-value decreases (see pic)



So that x, and y are negatively correlated.

The correlation can be captured by extending the notion of the variance to what is called the co-variance of data.

The Co-variance between x and y is defined as follows:

σ the expected value of x means μ (mean) in x direction times y means μ (mean) in y direction"

" μ_x is the expected value in x direction"

μ_y is the expected value in y direction"

or y coordinate

$$Cov[x,y] = E[(x-\mu x)(y-\mu y)]$$

$$\mu x = E[x]$$

$$\mu y = E[y]$$

For the 2D Data, we can therefore obtain, 4 quantities of interest

"we obtain the variance of x, variance of y

$\therefore Var[x]$

$\therefore VAR[y]$

we also have the Co-variance term:

"Covariance between x and y"

"Covariance between y and x"

$\therefore Cov[x,y]$

$\therefore Cov[y,x]$

we can summarize these values in this matrix, called the Co-variance matrix, with 4 entries; in top left corner we have variance in x direction; then Co-variance term between x and y in top right corner; Co-variance between y and x in bottom left corner; and variance of y in bottom right corner.

$$\begin{bmatrix} \text{Var}[x] & \text{Cov}[x,y] \\ \text{Cov}[y,x] & \text{Var}[y] \end{bmatrix}$$

If Covariance between x, y is positive, then on average the y value increases if we increase x and if Covariance between x, y is negative, then the y-value decreases, if we increase x on average.

If Co-variance between x and y is 0, then x, y have nothing to do with each other, they're uncorrelated.

The co-variance Matrix is always symmetrical, positive definitive matrix, with variances on the diagonal.

If we now, look at D Dimensional dataset, we have dataset consisting of n vectors

$$\Delta = \{x_1, \ldots, x_N\}, \text{ and every } x_i \text{ is in } R^D$$
$$, x_i \in \mathbb{R}^D$$

Then we can compute the variance of this dataset as 1 over N time Sum, $i = 1$ to Capital N, x_i minus μ (mu), time x_i minus μ transpose where μ is the mean of dataset

$$\therefore \text{VAR}[\Delta] = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)(x_i - \mu)^T$$

$$\downarrow$$

$D \times D$ matrix