Module 1

Covariances Part 2

at last session we looked at variance for 1 Dim Data set.

In this session, we looking at variance for higher Dim. data set.

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 data set, we can now compute variances in the x direction and y direction.

But these variances are insufficient to describe what's going on in the dataset.

In particular, we only have the variation of the data in either direction independent of the other direction, but ~~mean~~ we also be interested in the relationship between the x and y variable.
And this is where the concept of the co-variance between those components come into play.

Let's have a look at an example in 2 Dim's:
For this data set we can compute the variance in y-direction, and variance in x direction, which is indicated by vertical and horizontal Bar: (See 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 3 shape of data set, but variances in x, y direction are exactly the same, (see pic)

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 properties or shapes of data sets, but variances in x and y, and mean values 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.
This 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:
"as the expected value of x minus μ (mean) in x direction times y minus μ (mean) in y direction"
"μx is the expected value in x direction"
"μy is the expected value in y direction"
or y coordinate

$$\mathrm{Cov}[x,y] = E\left[(x-\mu x)(y-\mu y)\right]$$

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

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

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

"we obtain the variance of x, variance of y

$$\mathrm{Var}[x]$$

$$\mathrm{VAR}[y]$$

we also have the Co-variance term:

"Covariance between x and y"

"Covariance between y and x"

$$\therefore \mathrm{Cov}[x,y]$$

$$\therefore \mathrm{Cov}[y,x]$$

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.

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

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 are 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 data set consisting of n vectors

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

Then we can compute the variance of this dataset as 1 over N times sum, i=1 to capital N, x_i minus μ (mu), times x_i minus μ transpose

where μ is the mean of dataset

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

↓

D×D matrix