**What is Covariance?**

suppose we have two random variables *X* and *Y* that we’re particularly interested in. I’m not going to make any specific assumptions about *how* they’re distributed except to say they are jointly distributed according to some function *f*(*x*, *y*). In such cases, it’s interesting to consider the extent to which *X* and *Y* vary together, and this is precisely what *covariance* measures: it is a measure of the *joint variability* of two random variables.

**Covariance is a statistical measure** that **indicates the direction of the linear relationship between two variables**. It assesses how much two variables change together from their mean values.

Types of Covariance:

* **Positive Covariance**: When one variable increases, the other variable tends to increase as well, and vice versa.
* **Negative Covariance**: When one variable increases, the other variable tends to decrease.
* **Zero Covariance**: There is no linear relationship between the two variables; they move independently of each other.

Covariance is calculated by taking the average of the product of the deviations of each variable from their respective means. It is useful for understanding the direction of the relationship but not its strength, as its magnitude depends on the units of the variables.

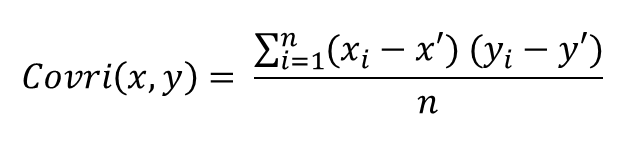
It is an essential tool for understanding how variables change together and is widely used in various fields, including finance, economics, and science.

**Covariance:**

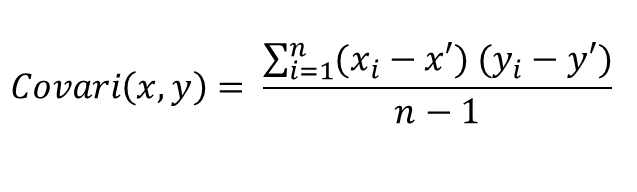
1. It is the relationship between a pair of random variables where a change in one variable causes a change in another variable.
2. It can take any value between – infinity to +infinity, where the negative value represents the negative relationship whereas a positive value represents the positive relationship.
3. It is used for the linear relationship between variables.
4. It gives the direction of relationship between variables.

**Covariance Formula**

**For Population:**

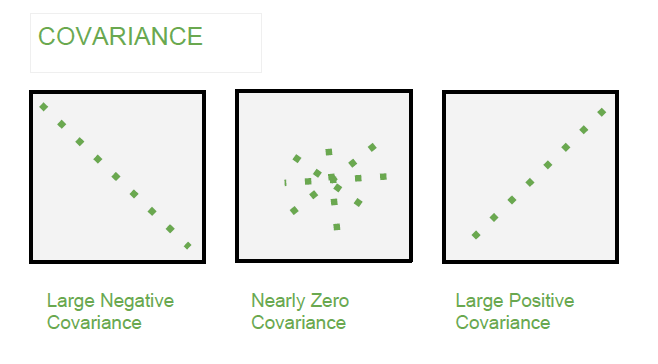


**For Sample:**



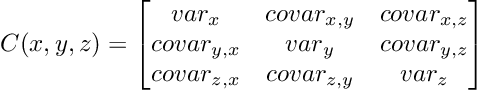
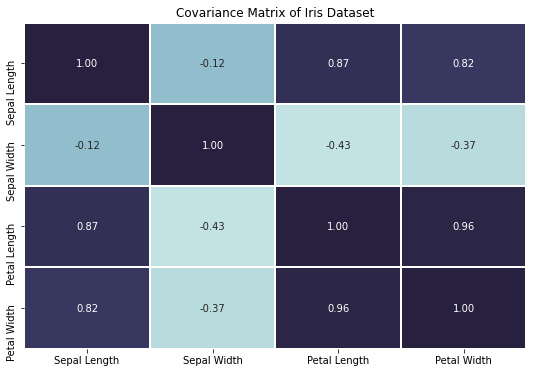
Here, x’ and y’ = mean of given sample set n = total no of sample xi and yi = individual sample of set

**Example –**



**Applications of Covariance and Correlation**

**Applications of Covariance**

* **Portfolio Management in Finance: Covariance is used to measure how different stocks or financial assets move together, aiding in portfolio diversification to minimize risk.**
* **Genetics: In genetics, covariance can help understand the relationship between different genetic traits and how they vary together.**
* **Econometrics: Covariance is employed to study the relationship between different economic indicators, such as the relationship between GDP growth and inflation rates.**
* **Signal Processing: Covariance is used to analyze and filter signals in various forms, including audio and image signals.**
* **Environmental Science: Covariance is applied to study relationships between environmental variables, such as temperature and humidity changes over time.**
* **Covariance and the covariance matrix**
* Assume, we have a dataset with two features and we want to describe the different relations within the data. The concept of covariance provides us with the tools to do so, allowing us to measure the variance between two variables.
* We can calculate the covariance by slightly modifying the equation from before, basically computing the variance of two variables with each other.
* 
* If we mean-center our data before, we can simplify the equation to the following:
* 
* Once simplified, we can see that the calculation of the covariance is actually quite simple. It is just the dot product of two vectors containing data.
* Now imagine, a dataset with three features x, y, and z. Computing the covariance matrix will yield us a 3 by 3 matrix. This matrix contains the covariance of each feature with all the other features and itself. We can visualize the covariance matrix like this:
* 
* Example based on [Implementing PCA From Scratch](https://towardsdatascience.com/implementing-pca-from-scratch-fb434f1acbaa)
* The covariance matrix is symmetric and feature-by-feature shaped. The diagonal contains the variance of a single feature, whereas the non-diagonal entries contain the covariance.
* We already know how to compute the covariance matrix, we simply need to exchange the vectors from the equation above with the mean-centered data matrix.
* 
* Once calculated, we can interpret the covariance matrix in the same way as described earlier, when we learned about the correlation coefficient.
* **Applying our knowledge**
* Now that we’ve finished the groundwork, let’s apply our knowledge.
* For testing purposes, we will use the iris dataset. The dataset consists of 150 samples with 4 different features (*Sepal Length, Sepal Width, Petal Length, Petal Width)*
* Let’s take a first glance at the data by plotting the first two features in a scatterplot.
* 
* An overview of the iris dataset by plotting the first two features [Image by Author]
* Our goal is to *‘manually’* compute the covariance matrix. Hence, we need to mean-center our data before. In order to do that, we define and apply the following function:
* ***Note****: We standardize the data by subtracting the mean and dividing it by the standard deviation.*
* Running the code above, standardizes our data and we obtain a mean of zero and a standard deviation of one as expected.
* Next, we can compute the covariance matrix.
* ***Note:****The same computation can be achieved with NumPy’s built-in function numpy.cov(x).*
* Our covariance matrix is a 4 by 4 matrix, shaped feature-by-feature. We can visualize the matrix and the covariance by plotting it like the following:
* 

Covariance matrix plotted as a heatmap [Image by Author]

We can clearly see a lot of correlation among the different features, by obtaining high covariance or correlation coefficients. For example, the petal length seems to be highly positively correlated with the petal width, which makes sense intuitively — if the petal is longer it is probably also wider.

* **Conclusion**
* The covariance matrix plays a central role in the [principal component analysis](https://towardsdatascience.com/implementing-pca-from-scratch-fb434f1acbaa). Implementing or computing it in a more manual approach ties a lot of important pieces together and breathes life into some linear algebra concepts.

Simply, covariance measures the extent to which the values of one variable are related to the values of another variable, which can either be positive or negative. A *positive* covariance indicates that the two variables tend to move in the same direction. For example, if large values of *X* tend to coincide with the large values of *Y*, then the covariance is positive. The same applies if lower values coincide, too. However, a *negative* covariance indicates that values tend to move in opposite directions: this would occur if large values of *X* correspond with low values of *Y*, for example.

A useful property of covariance is that its sign describes the tendency of the *linear relationship*between *X* and *Y.*That being said, the actual units it’s expressed in are somewhat less useful. Recall that we’re taking products between *X* and *Y* so the measure itself is also in units of *X* × *Y.*This can make comparisons between data difficult because the scale of measurement matters.

[Covariance and Correlation in Machine Learning: Practical Applications and Insights | by Rayan Yassminh | Sep, 2024 | Medium](https://medium.com/@ryassminh/covariance-and-correlation-in-machine-learning-practical-applications-and-insights-1dbe8cee1e1a)