Covariance & Correlation of Random Variables

Young W Lim

Sep 10 2021

Based on
Probability, Random Variables and Random Signal Principles,
P.Z. Peebles,Jr. and B. Shi

Outline

1. Covariance of Random Variables
2. Correlation of Random Variables
3. Pearson Correlation Coefficient
4. Comparisons

Covariance (1)

a measure of the joint variability of two random variables.

- the **positive covariance**
  the variables tend to show **similar behavior**

  the greater values of one variable
  the greater values of the other variable

  the lesser values of one variable
  the lesser values of the other variable

- the **negative covariance**
  the variables tend to show **opposite behavior**

  the greater values of one variable
  the lesser values of the other variable

  the lesser values of one variable
  the greater values of the other variable

https://en.wikipedia.org/wiki/Covariance

Covariance (2)

- the **sign** of the covariance

  the tendency in the linear relationship
  between the variables.

- the **magnitude** of the covariance

  not normalized and depends on
  the magnitudes of the variables.

- the **correlation coefficient**

  the **normalized magnitude** of the covariance
  the strength of the linear relation.

https://en.wikipedia.org/wiki/Covariance

Covariance (3)

must distinguish between

1. the **covariance** of two **random variables**,

   a **population parameter** that can be seen as
   a property of the **joint probability distribution**

2. the **sample covariance**

   serves as a **descriptor** of the **sample**
   serves as an estimated value of the **population parameter**.

https://en.wikipedia.org/wiki/Covariance

Covariance definition (1)

- For two jointly distributed **real-valued random variables** X and Y with finite **second moments**, the **covariance** is defined as the **expected value** (or mean) of the product of their deviations from their individual expected values

$$cov(X, Y) = E[(X - E[X])(Y - E[Y])]$$

https://en.wikipedia.org/wiki/Covariance

Covariance definition (2)

$$cov(X,Y) = E[(X - E[X])(Y - E[Y])]$$
$$= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)(x - E(X))(y - E(Y))dxdy$$

- where $E[X]$ is the expected value of X, also known as the mean of X.
- The covariance is also sometimes denoted σ_{XY} or $\sigma(X,Y)$, in analogy to variance.

https://en.wikipedia.org/wiki/Covariance

Covariance definition (3)

- By using the **linearity** property of expectations, this $cov(X, Y)$ can be simplified to the expected value of their product $E[XY]$ minus the product of their expected values $E[X]E[Y]$:

$$
\begin{aligned}
cov(X, Y) &= E[(X - E[X])(Y - E[Y])] \\
&= E[XY - XE[Y] - E[X]Y + E[X]E[Y]] \\
&= E[XY] - E[X]E[Y] - E[X]E[Y] + E[X]E[Y] \\
&= E[XY] - E[X]E[Y]
\end{aligned}
$$

https://en.wikipedia.org/wiki/Covariance

Covariance definition (4) Complex Random Variables

- The covariance between two **complex** random variables Z, W is defined as

$$cov(Z, W) = E[(Z - E[Z])(\overline{W - E[W]})] \\ = E[Z\overline{W}] - E[Z]E[\overline{W}]$$

https://en.wikipedia.org/wiki/Covariance

Covariance definition (5) Discrete Random Variables

- If the (real) **discrete random variable** pair (X, Y) can take on the values (x_i, y_i) for $i = 1, \ldots, n$, with equal probabilities

$$p_i = 1/n$$

then the covariance can be equivalently written in terms of the means $E[X]$ and $E[Y]$ as

$$cov(X, Y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - E[X])(y_i - E[Y])$$

https://en.wikipedia.org/wiki/Covariance

Covariance definition (6) Discrete Random Variables

- if there are n possible **realizations** of (X, Y), namely (x_i, y_i) but with possibly unequal probabilities p_i for $i = 1, \ldots, n$, then the covariance is

$$\mathrm{cov}(X, Y) = \sum_{i=1}^{n} p_i (x_i - E(X))(y_i - E(Y))$$
$$= \sum_{m} \sum_{n} p(x_m, y_n)(x_m - E(X))(y_n - E(Y))$$

https://en.wikipedia.org/wiki/Covariance

Statistical population

- a **population** is a set of similar items or events which is of interest for some question or experiment.
- A **statistical population** can be
  - a group of existing objects (e.g. the set of all stars within the Milky Way galaxy)
  - a hypothetical and potentially infinite group of objects conceived as a *generalization* from experience (e.g. the set of all possible hands in a game of poker).
- A common purpose of **statistical analysis** is to produce information about some chosen population.

https://en.wikipedia.org/wiki/Statistical_population

Sample Mean and Covariance

- The **sample mean** (empirical mean) and the **sample covariance** are statistics are computed from a sample of data on one or more random variables.
- the **sample mean** and **sample covariance** are widely used in statistics
  - to represent the location and dispersion of the distribution of values in the **sample**
  - to estimate the values for the **population**

https://en.wikipedia.org/wiki/Sample_mean_and_covariance

Sample Mean

- The **sample mean** (empirical mean) are statistics computed from a sample of data on one or more random variables.
- The **sample mean** can be used to refer to a vector of average values when the statistician is looking at the values of several variables in the sample

  e.g. the sales, profits, and employees of a sample of Fortune 500 companies.

https://en.wikipedia.org/wiki/Sample_mean_and_covariance

Sample Covariance

for the values of several variables in the sample :

- a **sample variance** for each variable
- a **sample variance-covariance matrix**
  (or simply **sample covariance matrix**)

  the relationship between each pair of variables.

  eg a 3x3 matrix when 3 variables are being considered.

https://en.wikipedia.org/wiki/Sample_mean_and_covariance

Sample Covariance

- The **sample covariance** is useful
  - in judging the **reliability** of the **sample means** as estimators
  - as an estimate of the **population covariance matrix**.

https://en.wikipedia.org/wiki/Sample_mean_and_covariance

Sample Mean Definition

- The **sample mean**
  the average of the values of a **variable** in a **sample**
  the sum of those values divided by the number of values
- if a **sample** of N observations on **variable** X
  is taken from the population, the **sample mean** is:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i .$$

https://en.wikipedia.org/wiki/Sample_mean_and_covariance

Sample Covariance Definition

- The sample covariance matrix is a K-by-K matrix $\boldsymbol{Q} = \left[q_{jk}\right]$ with entries

$$q_{jk} = \frac{1}{N-1} \sum_{i=1}^{N} \left(x_{ij} - \bar{x}_j\right)\left(x_{ik} - \bar{x}_k\right)$$

- where q_{jk} is an estimate of the covariance between the *j-th* variable and the *k-th* variable of the **population** underlying the data.

https://en.wikipedia.org/wiki/Sample_mean_and_covariance

Correlation (1)

- **correlation** or **dependence** is any statistical relationship, whether causal or not, between two **random variables** or **bivariate data**.
- in the broadest sense **correlation** is any statistical association, though it commonly refers to the degree to which **a pair of variables** are linearly related.

https://en.wikipedia.org/wiki/Correlation

Correlation (2)

familiar examples of dependent phenomena include

- the **correlation** between the height of parents and their offspring,
- the **correlation** between the price of a good and the quantity the consumers are willing to purchase, as it is depicted in the so-called demand curve.

https://en.wikipedia.org/wiki/Correlation

Correlation (3)

- **Correlations** are useful
  because they can indicate a predictive relationship
  that can be exploited in practice.
  - mild weather, less electricity consumption
  - extreme weather, more electricity consumption
  - the **correlation** between electricity demand and weather.
  - a causal relationship
- However, in general,
  the presence of a **correlation** is not sufficient
  to infer the presence of a **causal relationship**
  (i.e., **correlation** does not imply causation).

https://en.wikipedia.org/wiki/Correlation

Correlation (4)

- Formally, **random variables** are **dependent** if they do not satisfy a mathematical property of **probabilistic independence**.
- **correlation** is synonymous with **dependence**.

https://en.wikipedia.org/wiki/Correlation

Correlation (5)

- However, when used in a technical sense, **correlation** refers to any of several specific types of mathematical operations between the tested variables and their respective expected values.
- Essentially, **correlation** is the measure of how two or more **variables** are related to one another.

https://en.wikipedia.org/wiki/Correlation

Correlation (6)

- There are several **correlation coefficients**, often denoted r, measuring the **degree** of **correlation**.
- The most common of these is the **Pearson correlation coefficient**, which is sensitive only to a linear relationship between two variables (which may be present even when one variable is a nonlinear function of the other).
- Mutual information can also be applied to measure **dependence** between two variables.

https://en.wikipedia.org/wiki/Correlation

Correlation

$$\rho_{X,Y} = \mathrm{corr}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$$

$$= \frac{\mathrm{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

https://en.wikipedia.org/wiki/Covariance_and_correlation

Pearson correlation coefficient (1)

- **Pearson's correlation coefficient** is
  the **covariance** of the two variables
  divided by the product of their **standard deviations**.

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

https://en.wikipedia.org/w/index.php?title=Pearson_correlation_coefficient&action=edit&section=4

Pearson correlation coefficient (2)

- The form of the definition involves a "**product moment**", that is, the **mean** (the first moment about the origin) of the **product** of the **mean-adjusted** random variables; hence the modifier product-moment in the name.

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

$$\rho_{X,Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

https://en.wikipedia.org/w/index.php?title=Pearson_correlation_coefficient&action=edit&section=4

Pearson correlation coefficient (3)

Pearson correlation coefficient

$$\rho_{X,Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

$$\begin{aligned} E[(X-\mu_X)(Y-\mu_Y)] &= E[(X-E[X])(Y-E[Y])] \\ &= E[XY]-E[X]E[Y] \end{aligned}$$

$$\rho_{X,Y} = \frac{E[XY]-E[X]E[Y]}{\sqrt{E[X^2]-(E[X])^2}\ \sqrt{E[Y^2]-(E[Y])^2}}$$

https://en.wikipedia.org/w/index.php?title=Pearson_correlation_coefficient&action=edit&section=4

Pearson correlation coefficient (4)

Pearson correlation coefficient

$$\rho_{X,Y} = \frac{E[X\,Y] - E[X]E[Y]}{\sqrt{E[X^2] - (E[X])^2}\,\sqrt{E[Y^2] - (E[Y])^2}}$$

$$\mu_X = E[X]$$
$$\mu_Y = E[Y]$$
$$\sigma_X^2 = E[(X - [X])^2] = E[X^2] - (E[X])^2$$
$$\sigma_Y^2 = E[(Y - E[Y])^2] = E[Y^2] - (E[Y])^2$$

https://en.wikipedia.org/w/index.php?title=Pearson_correlation_coefficient&action=edit&section=4

Pearson correlation coefficient (5)

for a sample

- Pearson's correlation coefficient,
  when applied to a sample,
  is commonly represented by r_{xy}
  and may be referred to as
  the sample correlation coefficient or
  the sample Pearson correlation coefficient
- We can obtain a formula for r_{xy}
  by substituting estimates of the **covariances** and **variances**
  based on a sample into the formula above.

https://en.wikipedia.org/w/index.php?title=Pearson_correlation_coefficient&action=edit&section=4

Pearson correlation coefficient (6)

for a sample

- Given paired data $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ consisting of n pairs, r_{xy} is defined as

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

https://en.wikipedia.org/w/index.php?title=Pearson_correlation_coefficient&action=edit&section=4

Pearson correlation coefficient (7)

for a sample

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- n is sample size
- x_i, y_i are the individual sample points indexed with i
- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (the sample **mean**)
- $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ (the sample **mean**)

https://en.wikipedia.org/w/index.php?title=Pearson_correlation_coefficient&action=edit&section=4

Pearson correlation coefficient (8)

for a sample

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

rearranging gives us this formula for

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\ \sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

.

https://en.wikipedia.org/w/index.php?title=Pearson_correlation_coefficient&action=edit&section=4

Pearson correlation coefficient (9)

Properties of the Correlation Coefficient

- The correlation coefficient ranges from -1 to 1.
- A value of $+1$ implies that a linear equation describes the relationship between X and Y perfectly, with all data points lying on a line for which Y increases as X increases.
- A value of -1 implies that all data points lie on a line for which Y decreases as X increases.
- A value of 0 implies that there is no linear correlation between the variables

https://en.wikipedia.org/w/index.php?title=Pearson_correlation_coefficient&action=edit&section=4

Pearson correlation coefficient (10)

Properties of the Correlation Coefficient

- Note that $(X_i - X)(Y_i - Y)$ is positive
  if and only if X_i and Y_i lie on the same side
  of their respective means.
- Thus the correlation coefficient is positive
  if X_i and Y_i tend to be simultaneously greater than,
  or simultaneously less than, their respective means.
- The correlation coefficient is negative (anti-correlation)
  if X_i and Y_i tend to lie on opposite sides of their respective means.
- Moreover, the stronger is either tendency,
  the larger is the absolute value of the correlation coefficient.

https://en.wikipedia.org/w/index.php?title=Pearson_correlation_coefficient&action=edit&section=4

Covariance vs Correlation (1)

- covariance of two random variables

$$\mathrm{cov}_{XY} = \sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)]$$

- correlation of two random variables

$$\mathrm{corr}_{XY} = \rho_{XY} = E[(X - \mu_X)(Y - \mu_Y)]/(\sigma_X \sigma_Y)$$

https://en.wikipedia.org/wiki/Covariance_and_correlation

Covariance vs Correlation (2)

- In probability theory and statistics, the mathematical concepts of covariance and correlation are very similar.
- Both describe the degree to which two random variables or sets of random variables tend to **deviate** from their **expected values** in similar ways.
- **correlation** is dimensionless
- **covariance** is in units obtained
- by multiplying the units of the two variables.

https://en.wikipedia.org/wiki/Covariance_and_correlation

Time series analysis (1)

- In the case of a **time series** (random processes) which is **stationary** in the **wide sense**, both the means and variances are constant over time
  - $E[X_{n+m}] = E[X_n] = \mu_X$
  - $E[Y_{n+m}] = E[Y_n] = \mu_Y$
  - $var(X_{n+m}) = var(X_n)$
  - $var(Y_{n+m}) = var(Y_n)$
- In this case the **cross-covariance** and **cross-correlation** are functions of the time difference: m

https://en.wikipedia.org/wiki/Covariance_and_correlation

Time series analysis (2)

- cross-covariance in a WSS time series

$$\mathrm{cov}_{XY}(m) = \sigma_{XY}(m) = E[(X_n - \mu_X)(Y_{n+m} - \mu_Y)]$$

- cross-correlation in a WSS time series

$$\mathrm{corr}_{XY}(m) = \rho_{XY}(m) = E[(X_n - \mu_X)(Y_{n+m} - \mu_Y)]/(\sigma_X \sigma_Y)$$

https://en.wikipedia.org/wiki/Covariance_and_correlation

Time Series analysis (3)

- auto-covariance in a WSS time series

$$\mathrm{cov}_{XX}(m) = \sigma_{XX}(m) = E[(X_n - \mu_X)(X_{n+m} - \mu_X)]$$

- auto-correlation in a WSS time series

$$\mathrm{corr}_{XX}(m) = \rho_{XX}(m) = E[(X_n - \mu_X)(X_{n+m} - \mu_X)]/(\sigma_X^2)$$

https://en.wikipedia.org/wiki/Covariance_and_correlation

Auto-correlation/covariance for Random Processes

- auto-correlation for random processes

$$\mathrm{R}_{XX}(t_1, t_2) = \mathrm{E}\left[X_{t_1}\overline{X}_{t_2}\right]$$

- auto-covariance for random processes

$$\mathrm{K}_{XX}(t_1, t_2) = \mathrm{E}\left[(X_{t_1} - \mu_{X,t_1})\overline{(X_{t_2} - \mu_{X,t_2})}\right] = \mathrm{E}\left[X_{t_1}\overline{X}_{t_2}\right] - \mu_{X,t_1}\overline{\mu}_{X,t}$$

https://en.wikipedia.org/wiki/Covariance_and_correlation

Cross-correlation/covariance for Random Processes

- cross-correlation for random processes

$$\mathrm{R}_{XY}(t_1, t_2) = \mathrm{E}\left[X_{t_1}\overline{Y}_{t_2}\right]$$

- cross-covariance for random processes

$$\mathrm{K}_{XY}(t_1, t_2) = \mathrm{E}\left[(X_{t_1} - \mu_{X,t_1})\overline{(Y_{t_2} - \mu_{Y,t_2})}\right]$$

https://en.wikipedia.org/wiki/Covariance_and_correlation

Auto-correlation/covariance for WSS Processes

- auto-correlation for WSS processes

$$\mathrm{R}_{XX}(\tau) = \mathrm{E}\left[X_t \overline{X}_{t+\tau}\right]$$

- auto-covariance for WSS processes

$$\mathrm{K}_{XX}(\tau) = \mathrm{E}\left[(X_t - \mu_X)\overline{(X_{t+\tau} - \mu_X)}\right] = \mathrm{E}\left[X_t \overline{X}_{t+\tau}\right] - \mu_X \overline{\mu}_X$$

https://en.wikipedia.org/wiki/Covariance_and_correlation

Cross-correlation/covariance for WSS Processes

- cross-correlation for WSS processes

$$\mathrm{R}_{XY}(\tau) = \mathrm{E}\left[X_t \overline{Y}_{t+\tau}\right]$$

- cross-covariance for WSS processes

$$\mathrm{K}_{XY}(\tau) = \mathrm{E}\left[(X_t - \mu_X)\overline{(Y_{t+\tau} - \mu_Y)}\right]$$

https://en.wikipedia.org/wiki/Covariance_and_correlation