Covariance

From Wikipedia, the free encyclopedia

In probability theory and statistics, covariance is a measure of how much two random variables change together. If the greater values of one variable mainly correspond with the greater values of the other variable, and the same holds for the lesser values, i.e., the variables tend to show similar behavior, the covariance is positive.[1] For example, as a balloon is blown up it gets larger in all dimensions. In the opposite case, when the greater values of one variable mainly correspond to the lesser values of the other, i.e., the variables tend to show opposite behavior, the covariance is negative. If a sealed balloon is squashed in one dimension then it will expand in the other two. The sign of the covariance therefore shows the tendency in the linear relationship between the variables. The magnitude of the covariance is not easy to interpret. The normalized version of the covariance, the correlation coefficient, however, shows by its magnitude the strength of the linear relation.

A distinction must be made between (1) the covariance of two random variables, which is a population parameter that can be seen as a property of the joint probability distribution, and (2) the sample covariance, which serves as an estimated value of the parameter.

Contents

Definition

The covariance between two jointly distributed real-valued random variables X and Y with finite second moments is defined as[2]

$$\operatorname{cov}(X,Y) = \mathbf{E}\left[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])\right],$$

where E[X] is the expected value of X, also known as the mean of X. By using the linearity property of expectations, this can be simplified to

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

However, when $\operatorname{E}[XY] \approx \operatorname{E}[X]\operatorname{E}[Y]$, this last equation is prone to catastrophic cancellation when computed with floating point arithmetic and thus should be avoided in computer programs when the data has not been centered before.[3] Numerically stable algorithms should be preferred in this case.

For random vectors $\mathbf{X} \in \mathbb{R}^m$ and $\mathbf{Y} \in \mathbb{R}^n$, the *m*×*n* cross covariance matrix (also known as dispersion matrix or variance–covariance matrix,[4] or simply called covariance matrix) is equal to

$$\begin{aligned}\operatorname{cov}(\mathbf{X},\mathbf{Y}) &= \operatorname{E}\left[(\mathbf{X} - \operatorname{E}[\mathbf{X}])(\mathbf{Y} - \operatorname{E}[\mathbf{Y}])^{\mathrm{T}}\right] \\ &= \operatorname{E}\left[\mathbf{X}\mathbf{Y}^{\mathrm{T}}\right] - \operatorname{E}[\mathbf{X}]\operatorname{E}[\mathbf{Y}]^{\mathrm{T}},\end{aligned}$$

where m^{T} is the transpose of the vector (or matrix) m.

The (*i*,*j*)-th element of this matrix is equal to the covariance $\operatorname{Cov}(X_i, Y_j)$ between the *i*-th scalar component of *X* and the *j*-th scalar component of *Y*. In particular, $\operatorname{Cov}(Y, X)$ is the transpose of $\operatorname{Cov}(X, Y)$.

For a vector $\mathbf{X} = \begin{bmatrix} X_1 & X_2 & \dots & X_m \end{bmatrix}^{\mathrm{T}}$ of *m* jointly distributed random variables with finite second moments, its covariance matrix is defined as

$$\Sigma(\mathbf{X}) = \sigma(\mathbf{X}, \mathbf{X}).$$

Random variables whose covariance is zero are called uncorrelated. Similarly, random vectors whose covariance matrix is zero in every entry outside the main diagonal are called uncorrelated.

The units of measurement of the covariance $\operatorname{Cov}(X, Y)$ are those of *X* times those of *Y*. By contrast, correlation coefficients, which depend on the covariance, are a dimensionless measure of linear dependence. (In fact, correlation coefficients can simply be understood as a normalized version of covariance.)

Discrete variables

If each variable has a finite set of equal-probability values, x_i and y_i respectively for $i = 1, \dots, n$, then the covariance can be equivalently written in terms of the means $E(X)$ and $E(Y)$ as

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

It can also be equivalently expressed, without directly referring to the means, as[5]

$$\operatorname{cov}(X,Y) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{2}(x_i - x_j)\cdot(y_i - y_j) = \frac{1}{n^2}\sum_{i}\sum_{j>i}(x_i - x_j)\cdot(y_i - y_j).$$

Properties

- Variance is a special case of the covariance when the two variables are identical:

$$\operatorname{cov}(X, X) = \operatorname{Var}(X) \equiv \sigma^2(X).$$

- If *X*, *Y*, *W*, and *V* are real-valued random variables and *a*, *b*, *c*, *d* are constant ("constant" in this context means non-random), then the following facts are a consequence of the definition of covariance:

$$\begin{aligned}
\sigma(X, a) &= 0 \\
\sigma(X, X) &= \sigma^2(X) \\
\sigma(X, Y) &= \sigma(Y, X) \\
\sigma(aX, bY) &= ab\,\sigma(X, Y) \\
\sigma(X + a, Y + b) &= \sigma(X, Y) \\
\sigma(aX + bY, cW + dV) &= ac\,\sigma(X, W) + ad\,\sigma(X, V) + bc\,\sigma(Y, W) + bd\,\sigma(Y, V)
\end{aligned}$$

For a sequence X_1, ..., X_n of random variables, and constants a_1, ..., a_n, we have

$$\sigma^2\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2 \sigma^2(X_i) + 2 \sum_{i,j\,:\,i<j} a_i a_j \sigma(X_i, X_j) = \sum_{i,j} a_i a_j \sigma(X_i, X_j)$$

- A useful identity to compute the covariance between two random variables X, Y is the Hoeffding's Covariance Identity:[6]

$$\operatorname{cov}(X, Y) = \int_{\mathbb{R}} \int_{\mathbb{R}} F_{XY}(x, y) - F_X(x) F_Y(y) dx dy$$

where $F_{XY}(x, y)$ is the joint distribution function of the random vector (X, Y) and $F_X(x), F_Y(y)$ are the marginals.

A more general identity for covariance matrices

Let $\mathbf{X}$ be a random vector with covariance matrix $\Sigma(\mathbf{X})$, and let A be a matrix that can act on $\mathbf{X}$. The covariance matrix of the vector $A\mathbf{X}$ is:

$$\Sigma(A\mathbf{X}) = A\,\Sigma(\mathbf{X})\,A^{\mathrm{T}}.$$

This is a direct result of the linearity of expectation and is useful when applying a linear transformation, such as a whitening transformation, to a vector.

Uncorrelatedness and independence

If *X* and *Y* are independent, then their covariance is zero. This follows because under independence,

$$\mathbf{E}[XY] = \mathbf{E}[X] \cdot \mathbf{E}[Y].$$

The converse, however, is not generally true. For example, let *X* be uniformly distributed in [-1, 1] and let $Y = X^2$. Clearly, *X* and *Y* are dependent, but

$$\begin{aligned}\sigma(X,Y) &= \sigma(X, X^2) \\ &= \mathbf{E}\left[X \cdot X^2\right] - \mathbf{E}[X] \cdot \mathbf{E}\left[X^2\right] \\ &= \mathbf{E}\left[X^3\right] - \mathbf{E}[X]\mathbf{E}\left[X^2\right] \\ &= 0 - 0 \cdot \mathbf{E}\left[X^2\right] \\ &= 0.\end{aligned}$$

In this case, the relationship between *Y* and *X* is non-linear, while correlation and covariance are measures of linear dependence between two variables. This example shows that if two variables are uncorrelated, that does not in general imply that they are independent. However, if two variables are jointly normally distributed (but not if they are merely individually normally distributed), uncorrelatedness *does* imply independence.

Relationship to inner products

Many of the properties of covariance can be extracted elegantly by observing that it satisfies similar properties to those of an inner product:

1. bilinear: for constants *a* and *b* and random variables *X*, *Y*, *Z*, σ(*aX* + *bY*, *Z*) = *a* σ(*X*, *Z*) + *b* σ(*Y*, *Z*);
2. symmetric: σ(*X*, *Y*) = σ(*Y*, *X*);
3. positive semi-definite: $\sigma^2(X) = \sigma(X, X) \geq 0$ for all random variables *X*, and σ(*X*, *X*) = 0 implies that *X* is a constant random variable (*K*).

In fact these properties imply that the covariance defines an inner product over the quotient vector space obtained by taking the subspace of random variables with finite second moment and identifying any two that differ by a constant. (This identification turns the positive semi-definiteness above into positive definiteness.) That quotient vector space is isomorphic to the subspace of random variables with finite second moment and mean zero; on that subspace, the covariance is exactly the L^2 inner product of real-valued functions on the sample space.

As a result for random variables with finite variance, the inequality

$$|\sigma(X,Y)| \leq \sqrt{\sigma^2(X)\sigma^2(Y)}$$

holds via the Cauchy–Schwarz inequality.

Proof: If $\sigma^2(Y) = 0$, then it holds trivially. Otherwise, let random variable

$$Z = X - \frac{\sigma(X,Y)}{\sigma^2(Y)}Y.$$

Then we have

$$\begin{aligned}0 \leq \sigma^2(Z) &= \sigma\left(X - \frac{\sigma(X,Y)}{\sigma^2(Y)}Y, X - \frac{\sigma(X,Y)}{\sigma^2(Y)}Y\right) \\ &= \sigma^2(X) - \frac{(\sigma(X,Y))^2}{\sigma^2(Y)}.\end{aligned}$$

Calculating the sample covariance

The sample covariance of *N* observations of *K* variables is the *K*-by-*K* matrix $\overline{\overline{q}} = [[q_{jk}]]$ with the entries

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

which is an estimate of the covariance between variable j and variable k.

The sample mean and the sample covariance matrix are unbiased estimates of the mean and the covariance matrix of the random vector $\mathbf{X}$, a row vector whose *j*th element (*j = 1, ..., K*) is one of the random variables. The reason the sample covariance matrix has $N-1$ in the denominator rather than N is essentially that the population mean $E(X)$ is not known and is replaced by the sample mean $\bar{\mathbf{X}}$. If the population mean $E(X)$ is known, the analogous unbiased estimate is given by

$$q_{jk} = \frac{1}{N} \sum_{i=1}^{N} \left(X_{ij} - E(X_j)\right) \left(X_{ik} - E(X_k)\right)$$

Comments

The covariance is sometimes called a measure of "linear dependence" between the two random variables. That does not mean the same thing as in the context of linear algebra (see linear dependence). When the covariance is normalized, one obtains the correlation coefficient. From it, one can obtain the Pearson coefficient, which gives the goodness of the fit for the best possible linear function describing the relation between the variables. In this sense covariance is a linear gauge of dependence.

Applications

In genetics and molecular biology

Covariance is an important measure in biology. Certain sequences of DNA are conserved more than others among species, and thus to study secondary and tertiary structures of proteins, or of RNA structures, sequences are compared in closely related species. If sequence changes are found or no changes at all are found in noncoding RNA (such as microRNA), sequences are found to be necessary for common structural motifs, such as an RNA loop.

In financial economics

Covariances play a key role in financial economics, especially in portfolio theory and in the capital asset pricing model. Covariances among various assets' returns are used to determine, under certain assumptions, the relative amounts of different assets that investors should (in a normative analysis) or are predicted to (in a positive analysis) choose to hold in a context of diversification.

In meteorological and oceanographic data assimilation

The covariance matrix is important in estimating the initial conditions required for running weather forecast models. The 'forecast error covariance matrix' is typically constructed between

perturbations around a mean state (either a climatological or ensemble mean). The 'observation error covariance matrix' is constructed to represent the magnitude of combined observational errors (on the diagonal) and the correlated errors between measurements (off the diagonal).

In feature extraction

The covariance matrix is used to capture the spectral variability of a signal.[7]

See also

- Algorithms for calculating covariance
- Analysis of covariance
- Autocovariance
- Correlation and dependence
- Covariance function
- Covariance mapping
- Covariance matrix
- Covariance operator
- Distance covariance, or Brownian covariance.
- Eddy covariance
- Law of total covariance
- Propagation of uncertainty

References

1. http://mathworld.wolfram.com/Covariance.html
2. Oxford Dictionary of Statistics, Oxford University Press, 2002, p. 104.
3. Donald E. Knuth (1998). *The Art of Computer Programming*, volume 2: *Seminumerical Algorithms*, 3rd edn., p. 232. Boston: Addison-Wesley.
4. W. J. Krzanowski, *Principles of Multivariate Analysis*, Chap. 7.1, Oxford University Press, New York, 1988
5. Yuli Zhang,Huaiyu Wu,Lei Cheng (June 2012). *Some new deformation formulas about variance and covariance*. Proceedings of 4th International Conference on Modelling, Identification and Control(ICMIC2012). pp. 987–992.
6. Papoulis (1991). *Probability, Random Variables and Stochastic Processes*. McGraw-Hill.
7. Sahidullah, Md.; Kinnunen, Tomi (March 2016). "Local spectral variability features for speaker verification". *Digital Signal Processing*. 50: 1–11. doi:10.1016/j.dsp.2015.10.011.

External links

- Hazewinkel, Michiel, ed. (2001), "Covariance", *Encyclopedia of Mathematics*, Springer, ISBN 978-1-55608-010-4
- MathWorld page on calculating the sample covariance (http://mathworld.wolfram.com/Covariance.html)
- Covariance Tutorial using R (http://www.r-tutor.com/elementary-statistics/numerical-measures/covariance)
- Covariance and Correlation (http://itfeature.com/probability/covariance-and-correlation)

Look up *covariance* in Wiktionary, the free dictionary.

Categories: Covariance and correlation | Algebra of random variables

---