Covariance

Covariance provides a measure of the strength of the correlation between two or more sets of random variates. The covariance for two random variates X and Y, each with sample size N, is defined by the expectation value

$$\operatorname{cov}(X, Y) = \langle (X - \mu_X)(Y - \mu_Y) \rangle \tag{1}$$
$$= \langle X\, Y \rangle - \mu_X\, \mu_y \tag{2}$$

where $\mu_x = \langle X \rangle$ and $\mu_y = \langle Y \rangle$ are the respective means, which can be written out explicitly as

$$\operatorname{cov}(X, Y) = \sum_{i=1}^{N} \frac{(x_i - \bar{x})(y_i - \bar{y})}{N}. \tag{3}$$

For uncorrelated variates,

$$\operatorname{cov}(X, Y) = \langle X\, Y \rangle - \mu_X\, \mu_Y = \langle X \rangle \langle Y \rangle - \mu_X\, \mu_Y = 0, \tag{4}$$

so the covariance is zero. However, if the variables are correlated in some way, then their covariance will be nonzero. In fact, if $\operatorname{cov}(X, Y) > 0$, then Y tends to increase as X increases, and if $\operatorname{cov}(X, Y) < 0$, then Y tends to decrease as X increases. Note that while statistically independent variables are always uncorrelated, the converse is not necessarily true.

In the special case of $Y = X$,

$$\operatorname{cov}(X, X) = \langle X^2 \rangle - \langle X \rangle^2 \tag{5}$$
$$= \sigma_X^2, \tag{6}$$

so the covariance reduces to the usual variance $\sigma_X^2 = \operatorname{var}(X)$. This motivates the use of the symbol $\sigma_{X\,Y} = \operatorname{cov}(X, Y)$, which then provides a consistent way of denoting the variance as $\sigma_{X\,X} = \sigma_X^2$, where σ_X is the standard deviation.

The derived quantity

$$\operatorname{cor}(X, Y) = \frac{\operatorname{cov}(X, Y)}{\sigma_X\, \sigma_Y} \tag{7}$$
$$= \frac{\sigma_{X\,Y}}{\sqrt{\sigma_{X\,X}\, \sigma_{Y\,Y}}}, \tag{8}$$

is called **statistical correlation (Pearson)** of X and Y.

The covariance is especially useful when looking at the variance of the sum of two random variates, since

$$\operatorname{var}(X + Y) = \operatorname{var}(X) + \operatorname{var}(Y) + 2\operatorname{cov}(X, Y). \tag{9}$$

The covariance is symmetric by definition since

$$\operatorname{cov}(X, Y) = \operatorname{cov}(Y, X). \tag{10}$$

Given n random variates denoted X_1, ..., X_n, the covariance $\sigma_{ij} \equiv \operatorname{cov}(X_i, X_j)$ of X_i and X_j is defined by

$$\begin{aligned} \operatorname{cov}(X_i, X_j) &= \langle (X_i - \mu_i)(X_j - \mu_j) \rangle && (11) \\ &= \langle X_i X_j \rangle - \mu_i \mu_j, && (12) \end{aligned}$$

where $\mu_i = \langle X_i \rangle$ and $\mu_j = \langle X_j \rangle$ are the means of X_i and X_j, respectively. The matrix (V_{ij}) of the quantities $V_{ij} = \operatorname{cov}(X_i, X_j)$ is called the covariance matrix.

The covariance obeys the identities

$$\begin{aligned} \operatorname{cov}(X + Z, Y) &= \langle (X + Z) Y \rangle - \langle X + Z \rangle \langle Y \rangle && (13) \\ &= \langle X Y \rangle + \langle Z Y \rangle - (\langle X \rangle + \langle Z \rangle) \langle Y \rangle && (14) \\ &= \langle X Y \rangle - \langle X \rangle \langle Y \rangle + \langle Z Y \rangle - \langle Z \rangle \langle Y \rangle && (15) \\ &= \operatorname{cov}(X, Y) + \operatorname{cov}(Z, Y). && (16) \end{aligned}$$

By induction, it therefore follows that

$$\operatorname{cov}\left(\sum_{i=1}^{n} X_i, Y\right) = \sum_{i=1}^{n} \operatorname{cov}(X_i, Y) \tag{17}$$

$$\begin{aligned} \operatorname{cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right) &= \sum_{i=1}^{n} \operatorname{cov}\left(X_i, \sum_{j=1}^{m} Y_j\right) && (18) \\ &= \sum_{i=1}^{n} \operatorname{cov}\left(\sum_{j=1}^{m} Y_j, X_i\right) && (19) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{m} \operatorname{cov}(Y_j, X_i) && (20) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{m} \operatorname{cov}(X_i, Y_j). && (21) \end{aligned}$$