CORRELATION ANALYSIS

Pan, Lei

DOI: 10.1615/AtoZ.c.correlation_analysis

The term "correlation" is used to indicate the degree of interrelation between two or more variables. The procedure of calculating quantitatively the degree of the interrelation is called correlation analysis. Correlation analysis can be carried out for both continuous variables and discrete data, and the analysis for discrete data most often found in engineering practice and digital calculations is described below.

Autocorrelation function. Assuming a discrete time-series with finite number of samples of N and an average value of $\overline{x}$,

$$x_n = \{x_0, x_1, \ldots, x_{N-1}\}, \tag{1}$$

the autocorrelation function of which is defined as:

$$R_{xx}(k) = \sum_{n=0}^{N-1-k} (x_n - \overline{x})(x_{n+k} + \overline{x})\ (k = 0, 1, \ldots, N-1) \tag{2}$$

which effectively averages all possible products of the time-series and its time-shifted version separated by a time lag k. In practice, formula (2) is preferred in its normalized form

$$\rho_{xx}(k) = \frac{R_{xx}(k)}{R_{xx}(0)} = \frac{\sum_{n=1}^{N-1-k} (x_n - \overline{x})(x_{n+k} - \overline{x})}{\sum_{n=1}^{N-1} (x_n - \overline{x})^2}, \quad (k = 0, 1, \ldots, N-1). \tag{3}$$

The value of ρ_{xx} is such that $-1 \le \rho_{xx} \le 1$. The autocorrelation function is an average measure of the time-domain properties of the time-series, and is related to the power spectral density function in the frequency domain by the Fourier transform (see Spectral Analysis). If the magnitude of the autocorrelation function ρ_{xx} decreases with increasing time lag k, there is some degree of randomness in the time series. If the ρ_{xx} changes sign at regular time intervals, then the time-series is periodic, and a combination of the two may imply that the time-series is quasi-periodic, which is often the case in real engineering problems.

Cross-Correlation Function

The cross-correlation function for two sets of time-series data

$$x_n = \{x_0, x_1, \ldots, x_{N-1}\} \text{ and } y_n = \{y_0, y_1, \ldots, y_{N-1}\}$$

is defined as

$$R_{xy}(k) = \sum_{n=0}^{N-1-k} (x_n - \overline{x})(y_{n+k} - \overline{y}) \quad (k = 0, 1, \ldots, N-1) \tag{4}$$

The correlation function and cross-spectral function are equivalent measures in time and frequency domains which are related to each other by the Fourier transform (see Spectral Analysis).

Correlation Coefficient

The correlation coefficient is defined as the normalized version of formula (4) and is given by

$$\rho_{xy}(k) = \frac{R_{xy}(k)}{\sqrt{R_{xx}(0) R_{yy}(0)}} = \frac{\sum_{n=1}^{N-1} (x_n - \overline{x})(x_{n+k} - \overline{x})}{\sqrt{\sum_{n=1}^{N-1} (x_n - \overline{x})^2 \sum_{n=1}^{N-1} (y_n - \overline{y})^2}} \quad (k = 0, 1, \ldots, N-1), \tag{5}$$

the value of which at a particular time corresponding to k is a measure of similarity of the strength of components in x_n and y_n at that time. The value of ρ_{xy} is such that $-1 \leq \rho_{xy} \leq 1$, and the larger the ρ_{xy} the more strongly correlated are the x_n and y_n at a given time.

It should be emphasized that the concept of correlation is different from that of regression. The procedure of finding a best fit curve is called regression, whereas the accuracy of the regression curve is measured by correlation.

References

1. Gardner, W. A. (1988) *Statistical Spectral Analysis, a Non-probabilistic Theory* Prentic-Hall, Inc.