Cross correlation

In signal processing, cross-correlation is a measure of *similarity* of two series as a function of the displacement of one relative to the other. This is also known as a sliding dot product or sliding inner product. It is commonly used for searching a long signal for a shorter, known feature.

In an autocorrelation, which is the cross-correlation of a signal with itself, there will always be a peak at a lag of zero, and its size will be the signal energy.

Cross correlation

$$f(t) = \int_{-\infty}^{\infty} g(\tau)\, h(\tau + t)\, d\tau$$

$$= \int_{-\infty}^{\infty} g(\tau - t)\, h(\tau)\, d\tau$$

$$\hat{f}(\omega) = \int e^{-i\omega t} f(t)\, dt$$

$$= \int e^{-i\omega t} g(\tau - t)\, h(\tau)\, d\tau\, dt$$

$$= \int e^{-i\omega \tau} h(\tau)\, d\tau\; e^{-i\omega \underset{t-\tau}{(t-\tau)}} \underset{\tau - t}{g(\tau - t)}\, dt$$

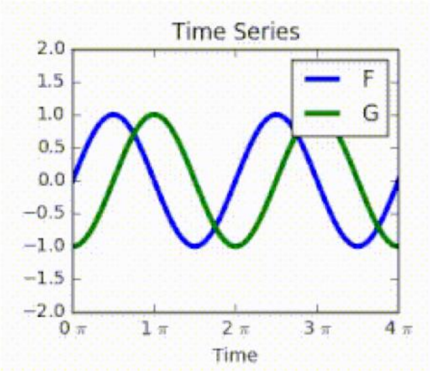
$$= \hat{h}(\omega)\, \hat{g}(-\omega)$$

$$= \hat{h}(\omega)\, \hat{g}^*(\omega) = \hat{g}^*(\omega)\, \hat{h}(\omega)$$

Note the similarity and difference between convolution & cross correlation

Below is the picture behind the mathematical definition, with an illustration.

As an example, consider two real valued functions g and h differing only by an unknown shift along the *x*-axis. One can use the cross-correlation to find how much g must be shifted along the *x*-axis to make it identical to h. The formula essentially slides the g function along the *x*-axis, calculating the integral of their product at each position. When the two functions match, the value of $(g * h)$ is maximized. This is because when peaks (positive areas) are aligned, they make a large contribution to the integral. Similarly, when troughs (negative areas) align, they also make a positive contribution to the integral because the product of two negative numbers is positive.



Two real valued functions differing only by an unknown shift along the *x*-axis. This shift can be determined by computing their cross correlation.

Convolution

Convolution is a mathematical operation on two functions (g and h) that produces a third function (f) expressing how the shape of one is modified by the other. The term convolution refers to both the result function and to the computing process. It is defined as the integral of the product of the two functions after one is reversed and shifted.

Convolution $(g*h)(t)$

$$\int_{-\infty}^{\infty} g(\tau)\, h(t-\tau)\, d\tau = f(t)$$

$g = s$, signal, data stream
$h = r$, response function
peaked & narrowly distributed.

Given t and $r(t-\tau)$,
τ has to be close to t if r is peaked at and distributed around $t-\tau=0$.

$$\int_{-\infty}^{\infty} g(\tau)\, h(t-\tau)\, d\tau$$

is mostly contributed from τ close to t.
This is to smear the signal g according to the response h.

$$f(t) = \int_{-\infty}^{\infty} g(\tau)\, h(t-\tau)\, d\tau$$

$$\hat{f}(\omega) = \int f(t)\, e^{-i\omega t}\, dt$$

$$= \int g(\tau)\, h(t-\tau)\, e^{-i\omega t}\, d\tau\, dt$$

$$= \int g(\tau)\, e^{-i\omega\tau}\, d\tau \; h(t-\tau)\, e^{-i\omega(t-\tau)}\, dt$$

$$= \hat{g}(\omega)\, \hat{h}(\omega)$$

Convolution to be efficiently done via Fourier transform

Below is a list of a few applications of convolution.

- In probability theory, the probability distribution of the sum of two independent random variables is the convolution of their individual distributions.

Proof: Suppose that the PDF of X is $f(x)$ and the PDF of Y is $g(y)$. Then the PDF of $Z = X + Y$ is given by $h(z) = \iint dxdy\delta(z-x-y) f(x)g(y) = \int dx f(x)g(z-x)$, which is the convolution of f and g.

- In statistics, a weighted moving average is a convolution.

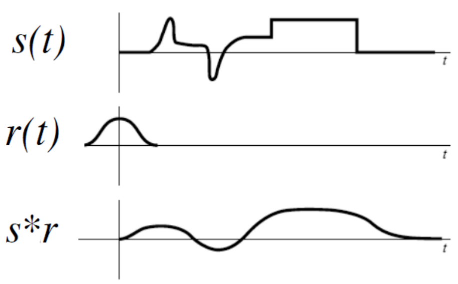
In financial applications a simple moving average is the unweighted mean of the previous n data. An example of a simple equally weighted running mean for an n-day sample of closing price is the mean of the previous n days' closing prices.

$$\bar{p}_{\mathrm{SM}} = \frac{p_M + p_{M-1} + \cdots + p_{M-(n-1)}}{n} = \frac{1}{n}\sum_{i=0}^{n-1} p_{M-i}\,.$$

In technical analysis of financial data, a weighted moving average (WMA) has the specific meaning of weights that decrease in arithmetical progression. In an n-day WMA the latest day has weight n, the second latest $n-1$, etc., down to one.

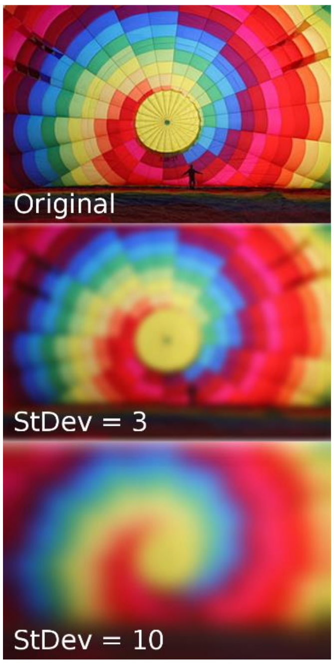
$$\mathrm{WMA}_M = \frac{np_M + (n-1)p_{M-1} + \cdots + 2p_{M-(n-2)} + p_{M-(n-1)}}{n + (n-1) + \cdots + 2 + 1}\,.$$

- Mathematically, applying a Gaussian blur to an image is the same as convolving the image with a Gaussian function. The response function is a Gaussian function, given by $\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}$ in one dimension. In two dimensions, it is the product of two Gaussian functions, one for each dimension.



The signal $s(t)$ is convolved with a response function $r(t)$

– Since the response function is broader than some features in the original signal, these are smoothed out in the convolution



From a one-dimensional illustration to a two-dimensional application.

Cross correlation between stocks

Price $P_i(t)$ of stock i at time t

$$R_i(t, \Delta t) = \ln \frac{P_i(t+\Delta t)}{P_i(t)} \quad \text{or} \quad \ln \frac{P_i(t)}{P_i(t-\Delta t)}$$

$\Delta t = 1$ day.

$$r_i(t, \Delta t) = \frac{R_i - \langle R_i \rangle}{\sigma_i}$$

where $\sigma_i = \sqrt{\langle R_i^2 \rangle - \langle R_i \rangle^2}$ is the SD.

$\langle \cdot \rangle = \frac{1}{N} \sum_{t=1}^{N} \cdot$ for time averaging over N days.

$R_i - \langle R_i \rangle$: fluctuation around the mean.

$r_i = (R_i - \langle R_i \rangle)/\sigma_i$: normalized fluctuation

$$C_{ij} = \langle r_i(t, \Delta t)\, r_j(t, \Delta t) \rangle = \frac{1}{N} \sum_{t=1}^{N} r_i(t, \Delta t)\, r_j(t, \Delta t)$$

$i = j$, $\quad C_{ii} = \langle r_i r_i \rangle = \dfrac{\langle R_i^2 \rangle - \langle R_i \rangle^2}{\sigma_i^2} = 1$

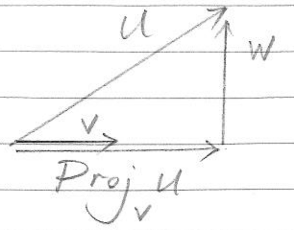
$C_{ij} \in [-1, 1]$ if $j \neq i$.

$$\langle r_i r_j \rangle = \frac{1}{N} \sum_{t=1}^{N} r_i(t)\, r_j(t), \quad \Delta t \text{ skipped}$$

$$= \frac{1}{N}\, r_i \cdot r_j$$

$r_i \cdot r_j = \sum_{t=1}^{N} r_i(t)\, r_j(t)$, inner product of $r_i(t)$ & $r_j(t)$.

Schwarz inequality



$$u = \mathrm{Proj}_v\, u + w = \frac{u\cdot v}{v\cdot v}\, v + w$$

$$w = u - \frac{u\cdot v}{v\cdot v}\, v$$

$$w\cdot v = u\cdot v - u\cdot v = 0 \qquad w \perp v$$

$$u\cdot u = \frac{(u\cdot v)^2}{v\cdot v} + w\cdot w \geq \frac{(u\cdot v)^2}{v\cdot v}$$

$$(u\cdot v)^2 \leq (u\cdot u)(v\cdot v)$$
$$|u\cdot v| \leq \sqrt{u\cdot u}\,\sqrt{v\cdot v}$$

Applied to r_i and r_j

$$|r_i\cdot r_j| \leq \sqrt{r_i\cdot r_i}\,\sqrt{r_j\cdot r_j}$$

Using $r_k\cdot r_l = N\langle r_k r_l\rangle$ between inner product and time averaging, we have

$$|\langle r_i\, r_j\rangle| \leq \sqrt{\langle r_i\, r_i\rangle}\,\sqrt{\langle r_j\, r_j\rangle} = 1$$
$$C_{ij} \in [-1, 1]$$
and $C_{ii} = 1$

The distribution of $C_{ij} = C_{ji}$ reveals the correlation of stocks.