Updating Formulas for Correlations

May 20, 2016

Here, we collect updating formulas for correlations between time series:

1 General Time Series

The standard case is to compute the correlation between two time series $x_t(i)$, $t = 1, \ldots, T$, $i = 1, \ldots, N$, and $y_t(i)$, $t = 1, \ldots, T$, $i = 1, \ldots, N$. Additionally, it is possible that weights are given for each time step, i.e. there are non-negative number w_t, $t = 1, \ldots, T$. Our goal then is to compute the (unnormalized) correlation

$$C(i,j) = \sum_{t=1}^{T} w_t \left(x_t(i) - \overline{x}(i)\right)\left(y_t(j) - \overline{y}(j)\right),$$

where $\bar{x}(i)$, $\bar{y}(j)$ denote the weighted mean values of the time series, i.e.

$$\begin{aligned} \overline{x}(i) &= \frac{1}{W_T}\sum_{t=1}^{T} w_t x_t(i), \\ W_T &= \sum_{t=1}^{T} w_t. \end{aligned}$$

We are interested in computing the correlation $C(i,j)$ in chunks. That means we split the data into, say, two blocks $x_t(i)$, $t = 1, \ldots, T_1$, and $x_t(i)$, $t = T_1 + 1, \ldots, T_2 = T$, and the same for y_t. We would then like to compute the correlation of each chunk separately, sum them up and add a correction term. Let us introduce the following notation

$$\overline{x_{T_1}}(i) = \frac{1}{w_{T_1}} \sum_{t=1}^{T_1} w_t x_t, \tag{1.1}$$

$$\overline{x_{T_2}}(i) = \frac{1}{W_{T_2}} \sum_{t=T_1+1}^{T_2} w_t x_t \tag{1.2}$$

$$W_{T_1} = \sum_{t=1}^{T_1} w_t \tag{1.3}$$

$$W_{T_2} = \sum_{t=T_1+1}^{T_2} w_t \tag{1.4}$$

$$S_{T_1}(i,j) = \sum_{t=1}^{T_1} \left(x_t(i) - \overline{x_{T_1}}(i)\right) \left(y_t(j) - \overline{y_{T_1}}(j)\right) \tag{1.5}$$

$$S_{T_2}(i,j) = \sum_{t=T_1+1}^{T_2} \left(x_t(i) - \overline{x_{T_2}}(i)\right) \left(y_t(j) - \overline{y_{T_2}}(j)\right). \tag{1.6}$$

Now, the calculations from section 5 show that the full correlation $C(i,j)$ can be computed as

$$C(i,j) = S_{T_1}(i,j) + S_{T_2}(i,j) + \frac{W_{T_1} W_{T_2}}{W_T} \left(\overline{x_{T_2}}(i) - \overline{x_{T_1}}(i)\right) \left(\overline{y_{T_2}}(j) - \overline{y_{T_1}}(j)\right) \tag{1.7}$$

2 Symmetrization

In some cases, a symmetric correlation matrix is desired, for example if y_t is a time-lagged version of x_t. This can be achieved by redefining the means

$$\overline{x}(i) = \frac{1}{2W_T} \left[\sum_{t=1}^{T} w_t x_t(i) + \sum_{t=1}^{T} w_t y_t(i) \right],$$

and defining the symmetrized correlation by

$$\begin{aligned} C_s(i,j) &= \sum_{t=1}^{T} w_t \left(x_t(i) - \overline{x}(i)\right) \left(y_t(j) - \overline{x}(j)\right) \\ &\quad + \sum_{t=1}^{T} w_t \left(y_t(i) - \overline{x}(i)\right) \left(x_t(j) - \overline{x}(j)\right). \end{aligned}$$

Using the analogues of Eqs. (1.1)-(1.6), we arrive at the updating formula

$$C_s(i,j) = S_{T_1}(i,j) + S_{T_2}(i,j) + \frac{2W_{T_1}W_{T_2}}{W_T}\left(\overline{x_{T_2}}(i) - \overline{x_{T_1}}(i)\right)\left(\overline{x_{T_2}}(j) - \overline{x_{T_1}}(j)\right) \quad (2.1)$$

see again section 5. Please note that for time-lagged data, T_1 and T_2 must be changed to $T_1 - \tau$ and $T_2 - \tau$, such that the first τ steps of every chunk only appear in x_t, while the last τ steps only appear in y_t.

3 Time-lagged Data without Symmetrization

If we assume to be given a time-series $\tilde{x}_t(i)$, $t = 1, \dots, T + \tau$, and define the time-lagged time-series $x_t(i) = \tilde{x}_t(i)$, $t = 1, \dots T$ and $y_t(i) = \tilde{x}_{t+\tau}$, $t = 1, \dots T$. If we do not wish to symmetrize the correlations, it seems most consistent to use the weights of the first T steps, w_t, $t = 1, \dots, T$, only. The means are thus defined by

$$\begin{aligned}
\overline{x}(i) &= \frac{1}{W_T}\sum_{t=1}^{T} w_t x_t(i) \\
\overline{y}(i) &= \frac{1}{W_T}\sum_{t=1}^{T} w_t y_t(i) \\
&= \frac{1}{W_T}\sum_{t=\tau}^{T+\tau} w_{t-\tau}\tilde{x}_t \\
W_T &= \sum_{t=1}^{T} w_t.
\end{aligned}$$

The asymmetric correlation then becomes

$$C_a(i,j) = \sum_{t=1}^{T} w_t \left(x_t(i) - \overline{x}(i)\right)\left(y_t(j) - \overline{y}(j)\right).$$

Using the analogues of Eqs. (1.1)-(1.6), we find the updating formula for time-lagged data to be the same as Eq. (1.7):

$$C_a(i,j) = S_{T_1}(i,j) + S_{T_2}(i,j) + \frac{W_{T_1}W_{T_2}}{W_T}\left(\overline{x_{T_2}}(i) - \overline{x_{T_1}}(i)\right)\left(\overline{y_{T_2}}(j) - \overline{y_{T_1}}(j)\right) \quad (3.1)$$

4 Conclusions

We have shown that mean-free correlations can be easily computed in chunks for arbitrary time series x_t, y_t, including time-dependent weights. Moreover, symmetrized mean-free correlations can be computed for arbitrary time-series, which can also be time-lagged copies. Finally, we found that for time-lagged time series which are not supposed to be symmetrized, it seems to make sense to compute the means using the weights of the first T steps.

5 Proofs

First, we determine an expression for the full correlation in terms of the partial sums S_{T_1}, S_{T_2} and a correction term for all cases considered here. We will see then that the correction term can be expressed in the forms given in Eqs. (1.7), (2.1) and (3.1). Let us consider the standard case:

$$\begin{aligned}
C(i,j) &= \sum_{t=1}^{T} w_t \left(x_t(i) - \overline{x}(i)\right)\left(y_t(j) - \overline{y}(j)\right) && (5.1)\\
&= \sum_{t=1}^{T_1} w_t \left(x_t(i) - \overline{x}(i)\right)\left(y_t(j) - \overline{y}(j)\right)\\
&\quad + \sum_{t=T_1+1}^{T_2} w_t \left(x_t(i) - \overline{x}(i)\right)\left(y_t(j) - \overline{y}(j)\right) && (5.2)\\
&= \sum_{t=1}^{T_1} w_t \left(\left(x_t(i) - \overline{x_{T_1}}(i)\right) - \gamma_1^x(i)\right)\left(\left(y_t(j) - \overline{y_{T_1}}(j)\right) - \gamma_1^y(j)\right)\\
&\quad + \sum_{t=T_1+1}^{T_2} w_t \left(\left(x_t(i) - \overline{x_{T_2}}(i)\right) - \gamma_2^x(i)\right)\left(\left(y_t(j) - \overline{y_{T_2}}(j)\right) - \gamma_2^y(j)\right) && (5.3)
\end{aligned}$$

where $\gamma_k^x(i) = \overline{x}(i) - \overline{x_{T_k}}(i)$ and $\gamma_k^y(i) = \overline{y}(i) - \overline{y_{T_k}}(i)$. We proceed to find

$$\begin{aligned}
C(i,j) &= \sum_{t=1}^{T_1} w_t \left(x_t(i) - \overline{x_{T_1}}(i)\right)\left(y_t(j) - \overline{y_{T_1}}(j)\right) - \gamma_1^x(i)\left(y_t(j) - \overline{y_{T_1}}(j)\right)\\
&\quad -\gamma_1^y(j)\left(x_t(i) - \overline{x_{T_1}}(i)\right) + \gamma_1^x(i)\gamma_1^y(j)\\
&\quad + \sum_{t=T_1+1}^{T_2} w_t \left(x_t(i) - \overline{x_{T_2}}(i)\right)\left(y_t(j) - \overline{y_{T_2}}(j)\right) - \gamma_2^x(i)\left(y_t(j) - \overline{y_{T_2}}(j)\right)\\
&\quad -\gamma_2^y(j)\left(x_t(i) - \overline{x_{T_2}}(i)\right) + \gamma_2^x(i)\gamma_2^y(j) && (5.4)\\
&= S_{T_1}(i,j) + S_{T_2}(i,j) + W_{T_1}\gamma_1^x(i)\gamma_1^y(j) + W_{T_2}\gamma_2^x(i)\gamma_2^y(j). && (5.5)
\end{aligned}$$

It remains to deal with the term:

$$
\begin{aligned}
W_{T_1}\gamma_1^x(i)\gamma_1^y(j) + W_{T_2}\gamma_2^x(i)\gamma_2^y(j) &= W_{T_1}\left(\overline{x}(i)\overline{y}(j) - \overline{x}(i)\overline{y_{T_1}}(j) - \overline{x_{T_1}}(i)\overline{y}(j) + \overline{x_{T_1}}(i)\overline{y_{T_1}}(j)\right) \\
&\quad + W_{T_2}\left(\overline{x}(i)\overline{y}(j) - \overline{x}(i)\overline{y_{T_2}}(j) - \overline{x_{T_2}}(i)\overline{y}(j) + \overline{x_{T_2}}(i)\overline{y_{T_2}}(j)\right) \qquad (5.6) \\
&= \left(W_{T_1} + W_{T_2}\right)\overline{x}(i)\overline{y}(j) + W_{T_1}\overline{x_{T_1}}(i)\overline{y_{T_1}}(j) \\
&\quad + W_{T_2}\overline{x_{T_2}}(i)\overline{y_{T_2}}(j) - \overline{x}(i)\left(W_{T_1}\overline{y_{T_1}}(j) + W_{T_2}\overline{y_{T_2}}(j)\right) \\
&\quad - \overline{y}(j)\left(W_{T_1}\overline{x_{T_1}}(i) + W_{T_2}\overline{x_{T_2}}(i)\right). \qquad (5.7)
\end{aligned}
$$

Now, we use that $W_{T_1}\overline{x_{T_1}}(i) + W_{T_2}\overline{x_{T_2}}(i) = W_T\overline{x}(i)$ to find:

$$
\begin{aligned}
&= W_{T_1}\overline{x_{T_1}}(i)\overline{y_{T_1}}(j) + +W_{T_2}\overline{x_{T_2}}(i)\overline{y_{T_2}}(j) \\
&\quad - \overline{x}(i)\left(W_{T_1}\overline{y_{T_1}}(j) + W_{T_2}\overline{y_{T_2}}(j)\right) \qquad (5.8) \\
&= \frac{1}{W_T}\left[W_T\left(W_{T_1}\overline{x_{T_1}}(i)\overline{y_{T_1}}(j) + +W_{T_2}\overline{x_{T_2}}(i)\overline{y_{T_2}}(j)\right)\right] \\
&\quad - \frac{1}{W_T}\left[W_T\overline{x}(i)\left(W_{T_1}\overline{y_{T_1}}(j) + W_{T_2}\overline{y_{T_2}}(j)\right)\right] \qquad (5.9) \\
&= \frac{W_{T_1}W_{T_2}}{W_T}\left[\overline{x_{T_1}}(i)\overline{y_{T_1}}(j) + \overline{x_{T_2}}(i)\overline{y_{T_2}}(j) - \overline{x_{T_1}}(i)\overline{y_{T_2}}(j) - \overline{x_{T_2}}(i)\overline{y_{T_1}}(j)\right] \qquad (5.10)
\end{aligned}
$$

This completes the proof of Eq. (1.7). For the symmetric case, the procedure from Eqs. (5.1)-(5.5) can be repeated to come up with the expression

$$
\begin{aligned}
C_s(i,j) &= S_{T_1}(i,j) + S_{T_2}(i,j) + W_{T_1}\left(\gamma_1(i)\gamma_1(j) + \gamma_1(j)\gamma_1(i)\right) \\
&\quad + W_{T_2}\left(\gamma_2(i)\gamma_2(j) + \gamma_2(j)\gamma_2(i)\right),
\end{aligned}
$$

where $\gamma_k(i) = \overline{x}(i) - \overline{x_{T_k}}(i)$. Then, the steps of Eqs. (5.6)-(5.9) can be repeated in the same way. For the asymmetric case, Eqs. (5.1)-(5.5) yield the expression

$$
\begin{aligned}
C_a(i,j) &= S_{T_!}(i,j) + S_{T_2}(i,j) + W_{T_1}\gamma_1^x(i)\gamma_1^y(j) + W_{T_2}\gamma_2^x(i)\gamma_2^y(j) \\
&\quad - \gamma_1^x(i)\sum_{t=1}^{T_1} w_t\left(y_t(j) - \overline{y_{T_1}}(j)\right) - \gamma_1^y(j)\sum_{t=1}^{T_1} w_t\left(x_t(i) - \overline{x_{T_1}}(i)\right) \\
&\quad - \gamma_2^x(i)\sum_{t=T_1+1}^{T_2} w_t\left(y_t(j) - \overline{y_{T_2}}(j)\right) - \gamma_2^y(j)\sum_{t=T_1+1}^{T_2} w_t\left(x_t(i) - \overline{x_{T_2}}(i)\right).
\end{aligned}
$$

Here, we have used $\gamma_k^x(i) = \overline{x}(i) - \overline{x_{T_k}}(i)$, $\gamma_k^y(i) = \overline{y}(i) - \overline{y_{T_k}}(i)$. The cross-terms cancel out and the expression $W_{T_1}\gamma_1^x(i)\gamma_1^y(j) + W_{T_2}\gamma_2^x(i)\gamma_2^y(j)$ can be reformulated through Eqs. (5.6)-(5.10) to end up with Eq. (3.1).