Cross-correlation function

Lag features

Contents

WHAT IS THE CCF

HOW TO USE CCF TO IDENTIFY USEFUL LAGS

Cross-correlation function

- So far we have shown the correlation between y_t and a lagged value of itself y_{t-k}.

- Now we show the correlation between y_t and the lagged value of a feature x_{t-k}.

- The cross-correlation function between x_t and y_t is defined as [1]:

$$corr(x_t, y_{t+k}) = \frac{cov(x_t, y_{t+k})}{\sigma_y \sigma_x}$$

- This is equivalent to $corr(x_{t-k}, y_t)$

	Target y_t ↓	Feature x_t ↓
Date	**Sales**	**Ad spend**
2020-02-12	23	100
2020-02-13	30	120
2020-02-14	35	90
2020-02-15	30	80
2020-02-16	?	100

[1] - George, E. P. "Box. Time series analysis: forecasting and control." (1970).

Cross-correlation function

- We want to know which lags x_{t-k} are most predictive of y_t.

- We can look at the cross-correlation function between x_t and y_t to do this.

- When the correlation between x_{t-k} and y_t does not depend on time we can write:

$$\rho_{xy}(k) = corr(x_t, y_{t+k}) = \frac{cov(x_t, y_{t+k})}{\sigma_x \sigma_y}$$

Target y_t ↓

Feature x_t ↓

Date	Sales	Ad spend
2020-02-12	23	100
2020-02-13	30	120
2020-02-14	35	90
2020-02-15	30	80
2020-02-16	?	100

Cross-correlation function

- We want to know which lags x_{t-k} are most predictive of y_t.

- We can look at the cross-correlation function between x_t and y_t to do this.

$$\rho_{xy}(k) = corr(x_t, y_{t+k}) = \frac{cov(x_t, y_{t+k})}{\sigma_x \sigma_y}$$

- It can be shown this is equivalent to:

$$\rho_{yx}(-k) = corr(y_t, x_{t-k}) = \frac{cov(y_t, x_{t-k})}{\sigma_y \sigma_x}$$

Target y_t ↓

Feature x_t ↓

Date	Sales	Ad spend
2020-02-12	23	100
2020-02-13	30	120
2020-02-14	35	90
2020-02-15	30	80
2020-02-16	?	100

Cross-correlation function

- We want to know which lags x_{t-k} are most predictive of y_t.
- We can look at the cross-correlation function between x_t and y_t to do this.

$$\rho_{xy}(k) = \frac{\sum_{t=1}^{N-k}(x_t - \bar{x})(y_{t+k} - \bar{y})}{\sqrt{\sum_{t=1}^{N}(x_t - \bar{x})^2}\sqrt{\sum_{t=1}^{N}(y_t - \bar{y})^2}}$$

	Target y_t ↓	Feature x_t ↓
Date	**Sales**	**Ad spend**
2020-02-12	23	100
2020-02-13	30	120
2020-02-14	35	90
2020-02-15	30	80
2020-02-16	?	100

Cross-correlation function

- We want to know which lags x_{t-k} are most predictive of y_t.
- We can look at the cross-correlation function between x_t and y_t to do this.

$$\rho_{xy}(k) = \frac{\sum_{t=1}^{N-k}(x_t - \bar{x})(y_{t+k} - \bar{y})}{\sqrt{\sum_{t=1}^{N}(x_t - \bar{x})^2}\sqrt{\sum_{t=1}^{N}(y_t - \bar{y})^2}}$$

Target y_t ↓

Feature x_t ↓

Date	Sales	Ad spend
2020-02-12	23	100
2020-02-13	30	120
2020-02-14	35	90
2020-02-15	30	80
2020-02-16	?	100

Disclaimer: This only captures the **linear** relationship between y_t and x_{t-k}. If a **non-linear** relationship exists (e.g., $y_t \sim x_{t-k}^2$) it won't necessarily be picked up by the CCF.

Cross-correlation function

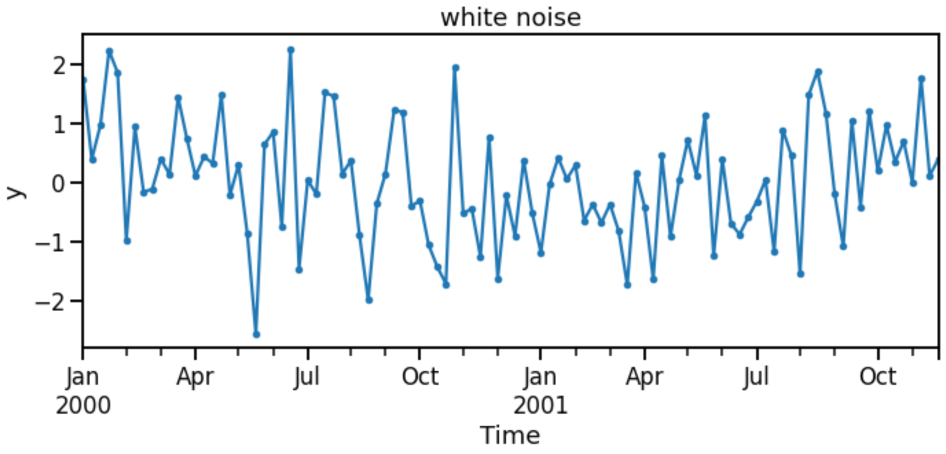
- We want to know which lags x_{t-k} are most predictive of y_t
- We can look at the cross-correlation function between x_t and y_t to do this.

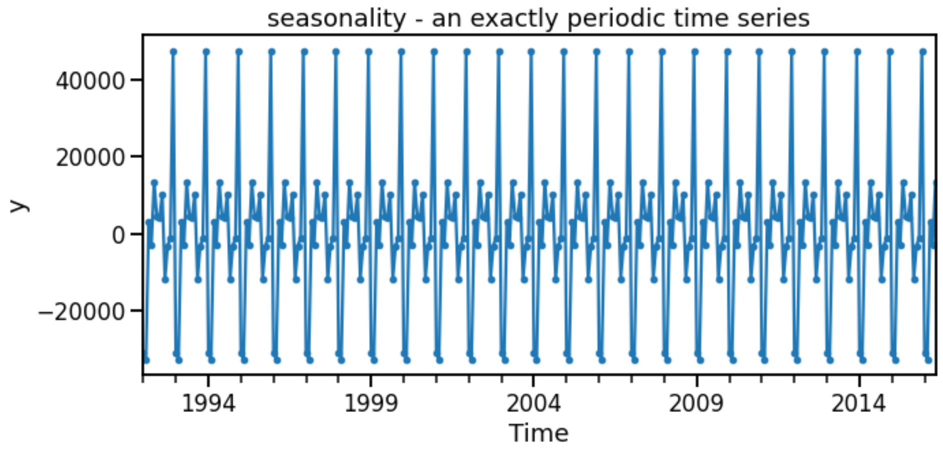
$$\rho_{xy}(k) = \frac{\sum_{t=1}^{N-k}(x_t - \bar{x})(y_{t+k} - \bar{y})}{\sqrt{\sum_{t=1}^{N}(x_t - \bar{x})^2}\sqrt{\sum_{t=1}^{N}(y_t - \bar{y})^2}}$$

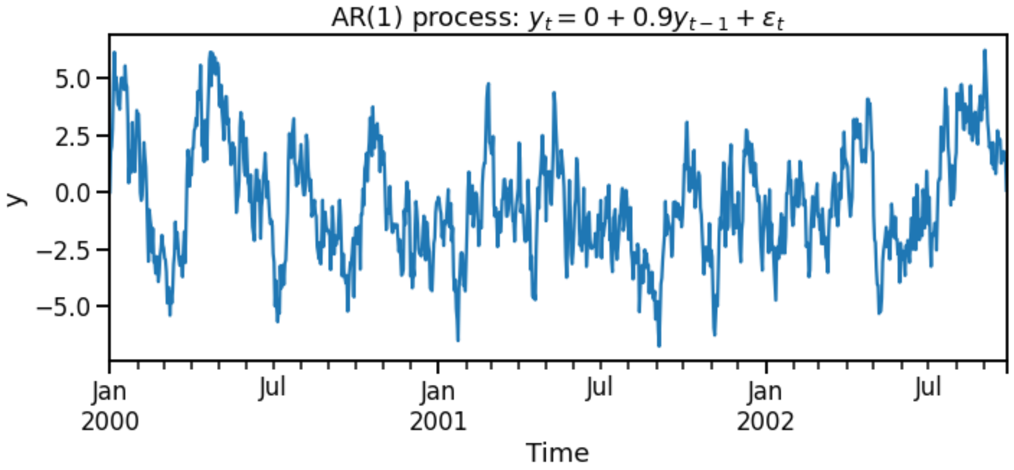
	Target y_t ↓	Feature x_t ↓
Date	**Sales**	**Ad spend**
2020-02-12	23	100
2020-02-13	30	120
2020-02-14	35	90
2020-02-15	30	80
2020-02-16	?	100

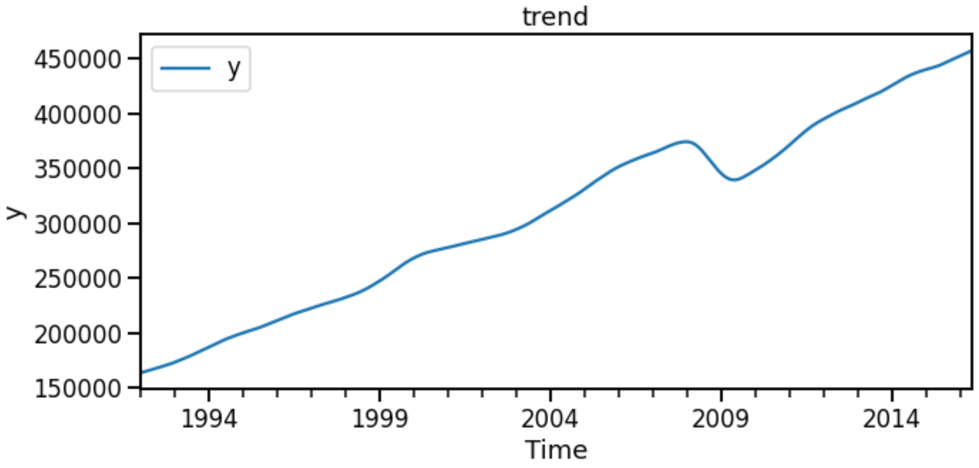
Disclaimer: If a **non-linear** relationship is suspected (e.g., $y_t \sim x_{t-k}^2$) then transform the feature before plugging it into the CCF.

Let’s look at the CCF for different time series

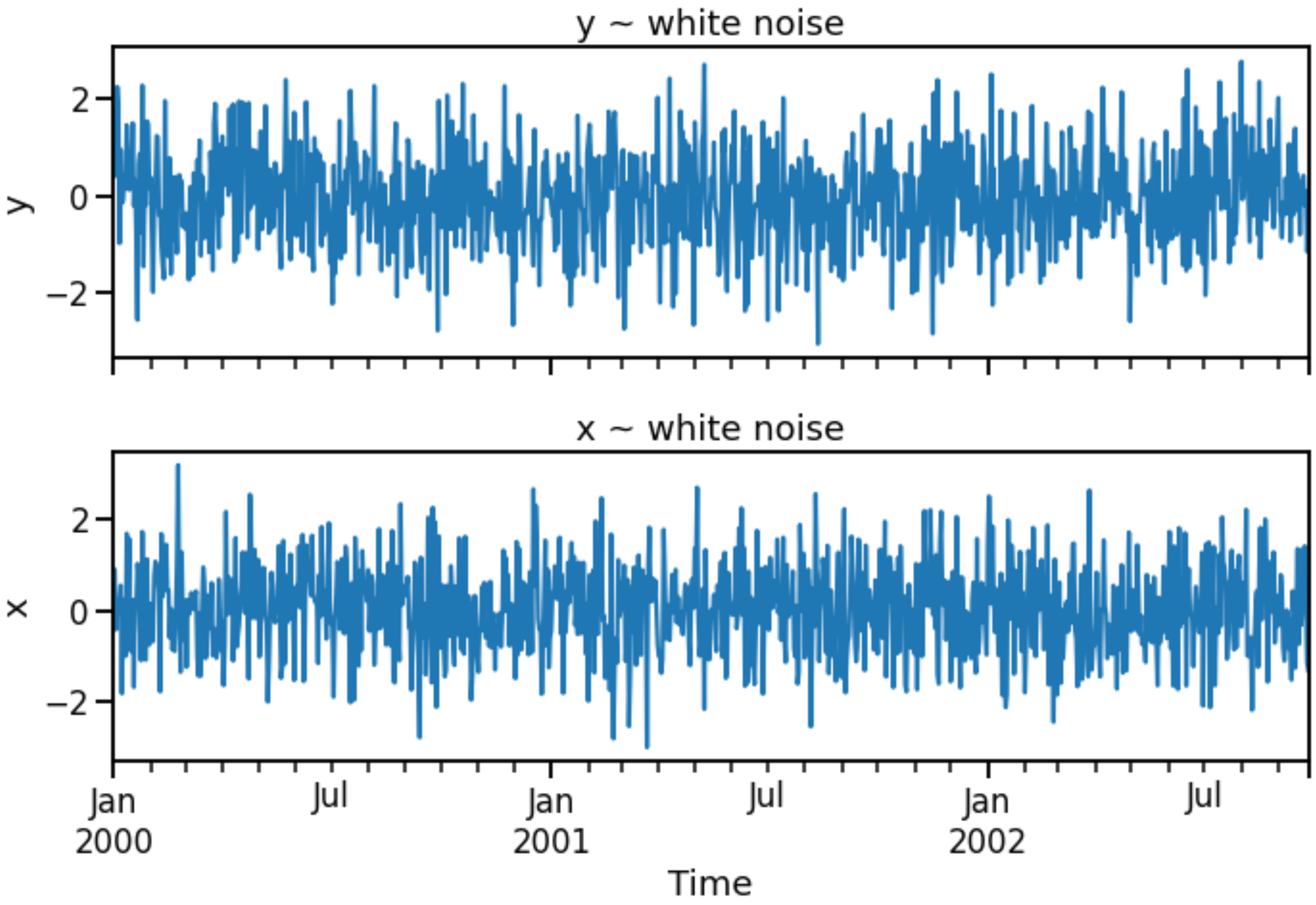








White noise: two time series



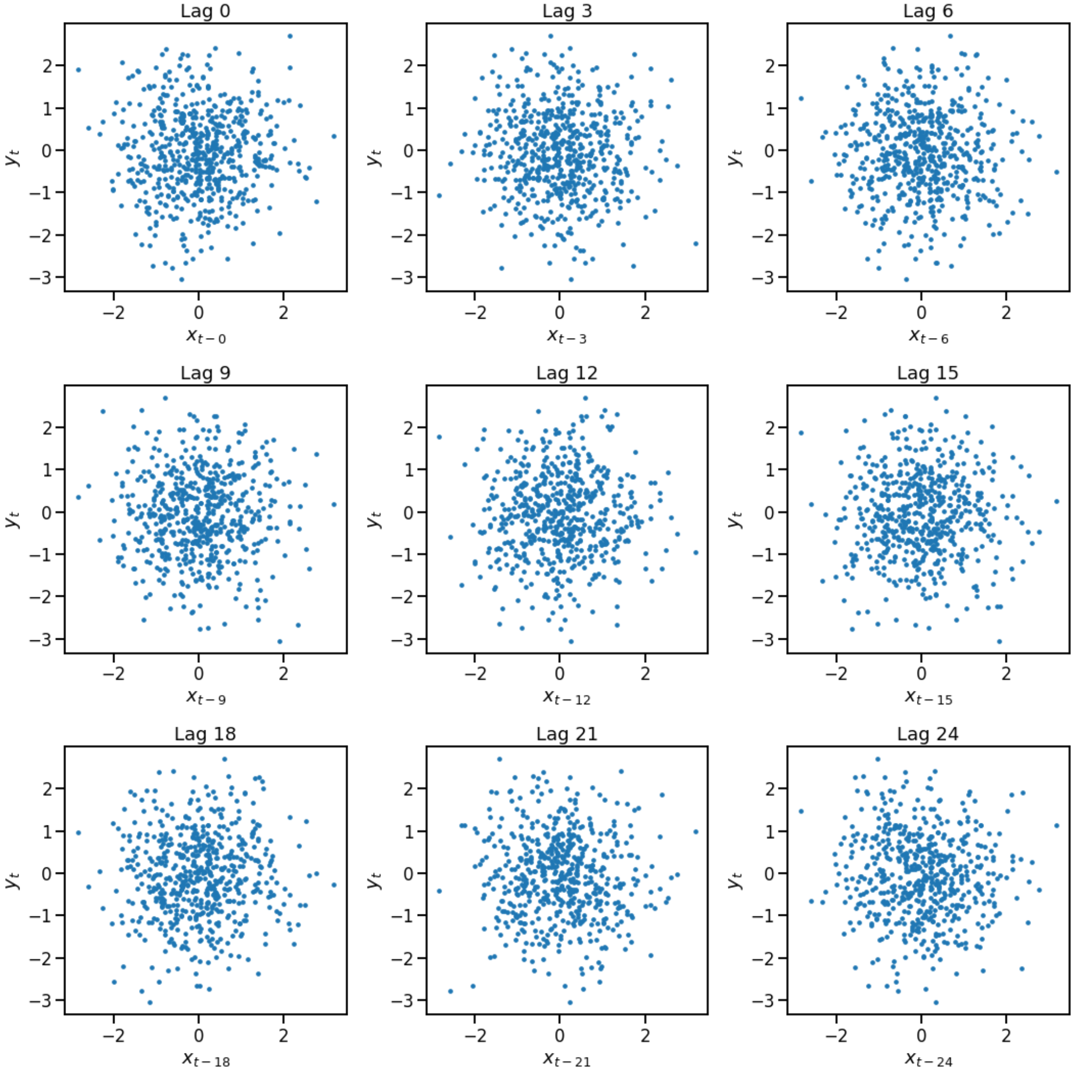
$y_t = \epsilon_t$ where $\epsilon_t \sim N(0,1)$

No correlation within or between time series

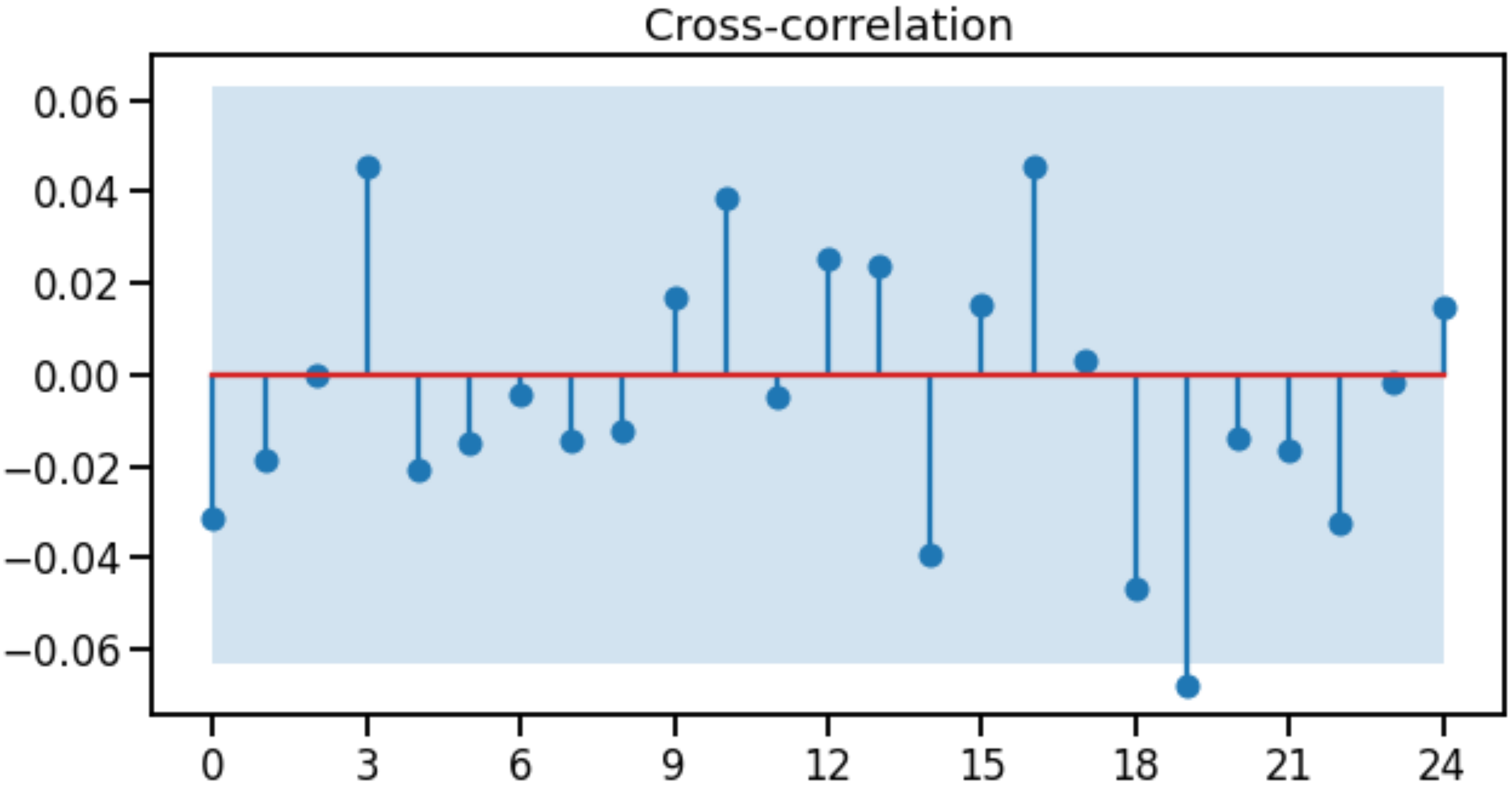
$x_t = \epsilon_t$ where $\epsilon_t \sim N(0,1)$

White noise: lag plots

The lag plots show that we don’t expect a large correlation at any of the lags shown.

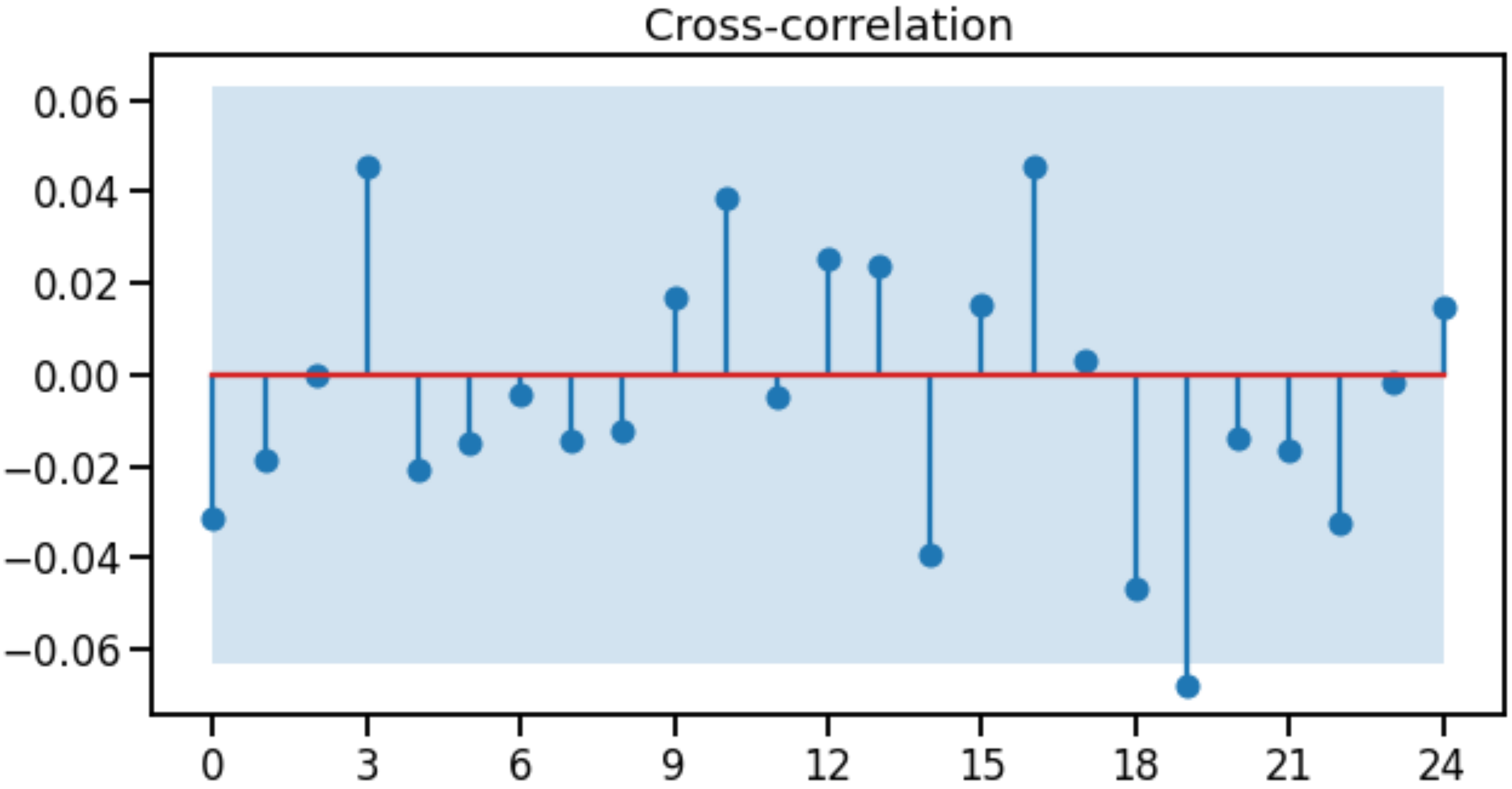


White noise: CCF



The CCF shows very small correlations as expected for two uncorrelated time series. We expect some significant lags just by chance so these are not anything to be concerned with.

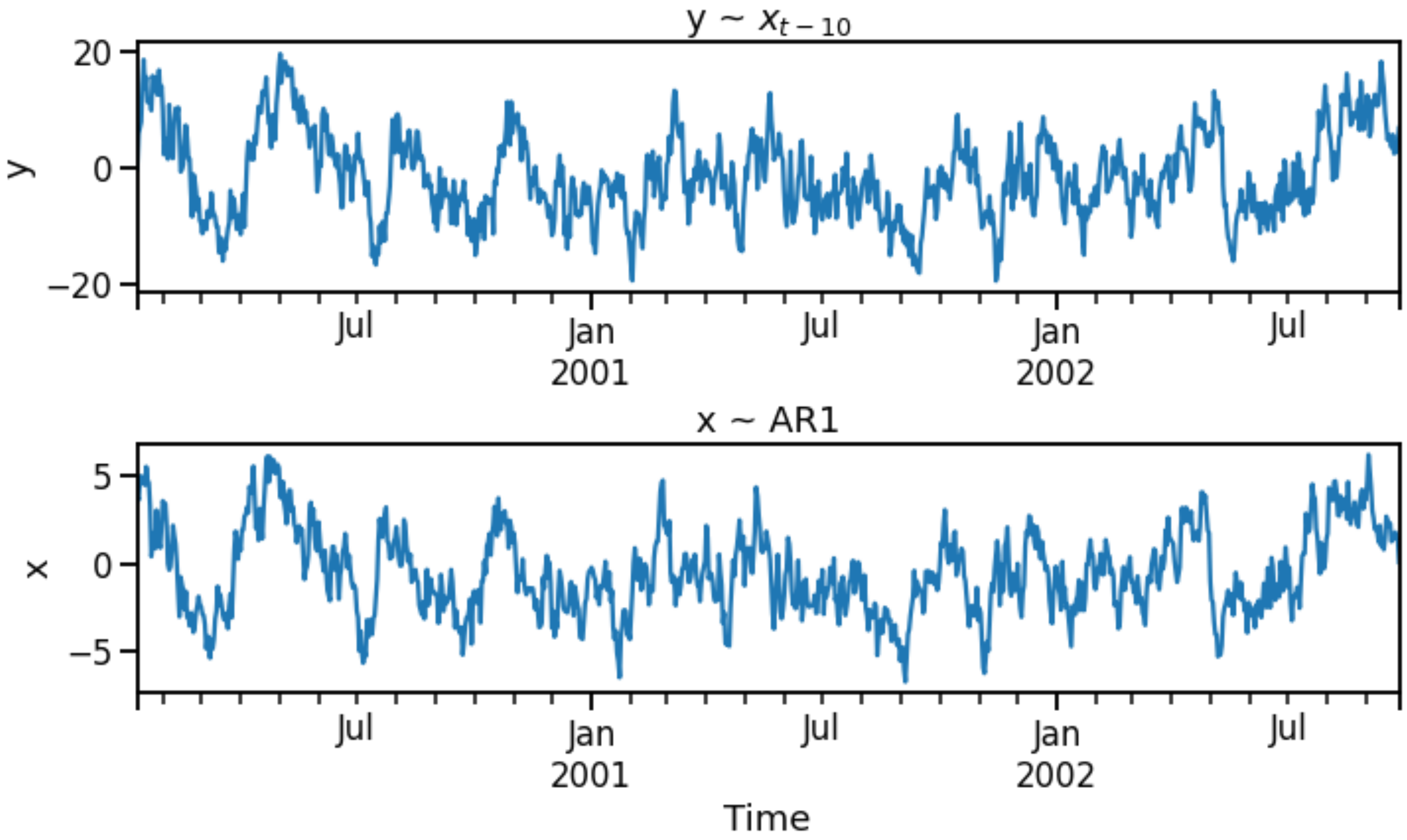
White noise: CCF



The 95% confidence interval (the blue shaded area) for the cross-correlation is given by: $2/\sqrt{n}$ where n is the length of the time series. The null hypothesis is that there is no correlation. [1]

[1]- Box, George EP, et al. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015. Page 462

AR(1) process: Two correlated time series

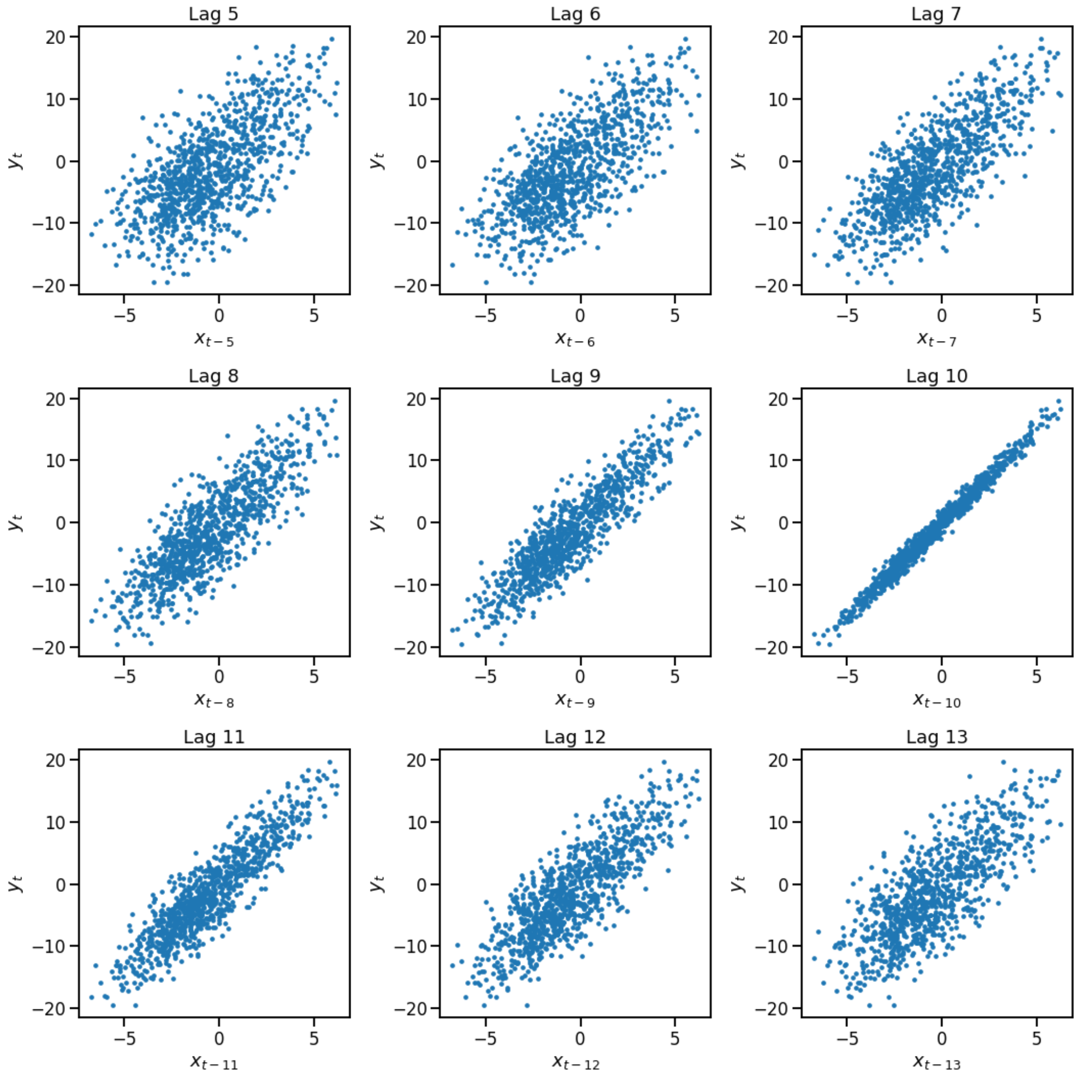


$y_t = 3x_{t-10} + \epsilon_t$ where $\epsilon_t \sim \mathrm{N}(0,1)$

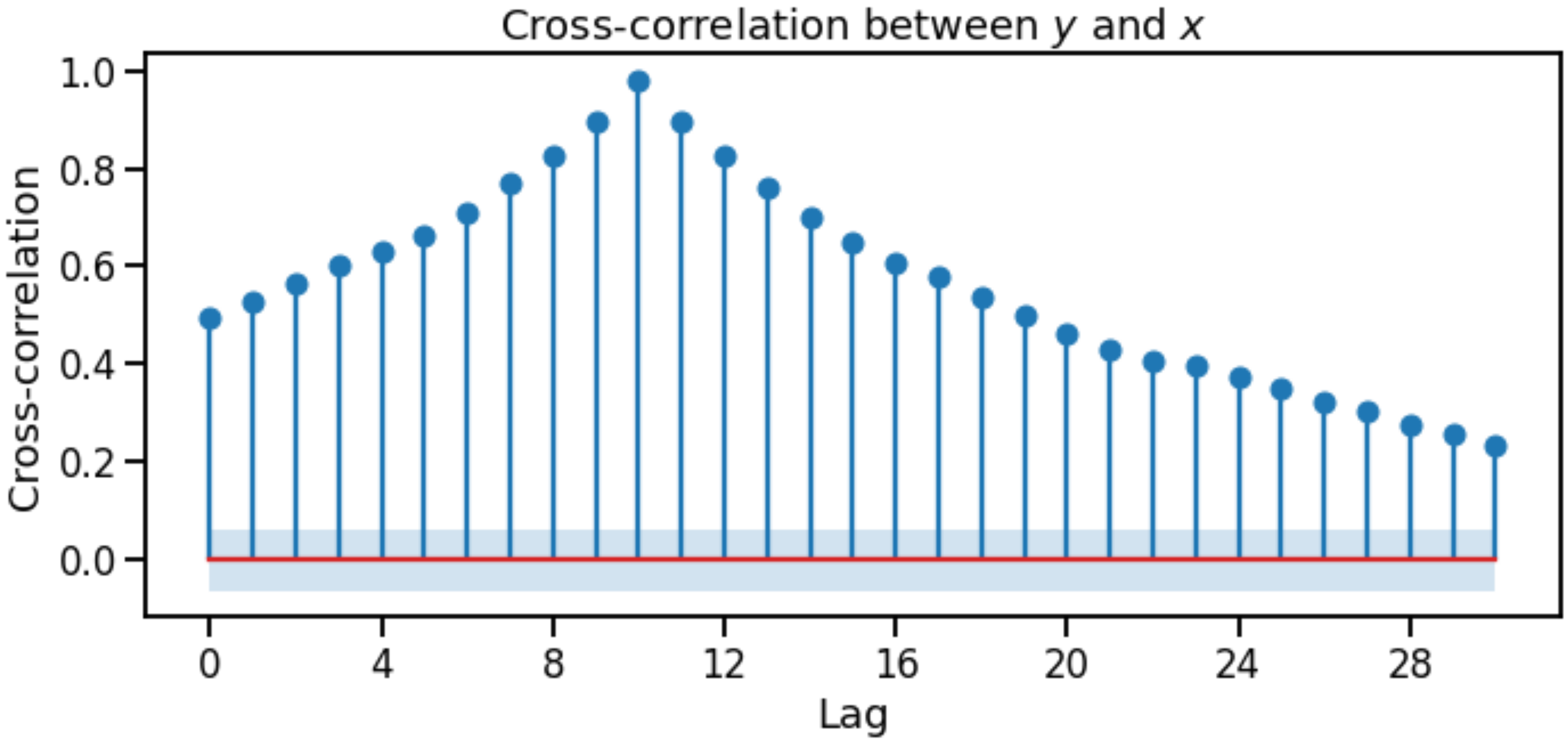
$x_t = 0.9x_{t-1} + \epsilon_t$ where $\epsilon_t \sim \mathrm{N}(0,1)$

AR(1) process: Lag plots

We see that the largest cross-correlation occurs at a lag of 10 as expected.

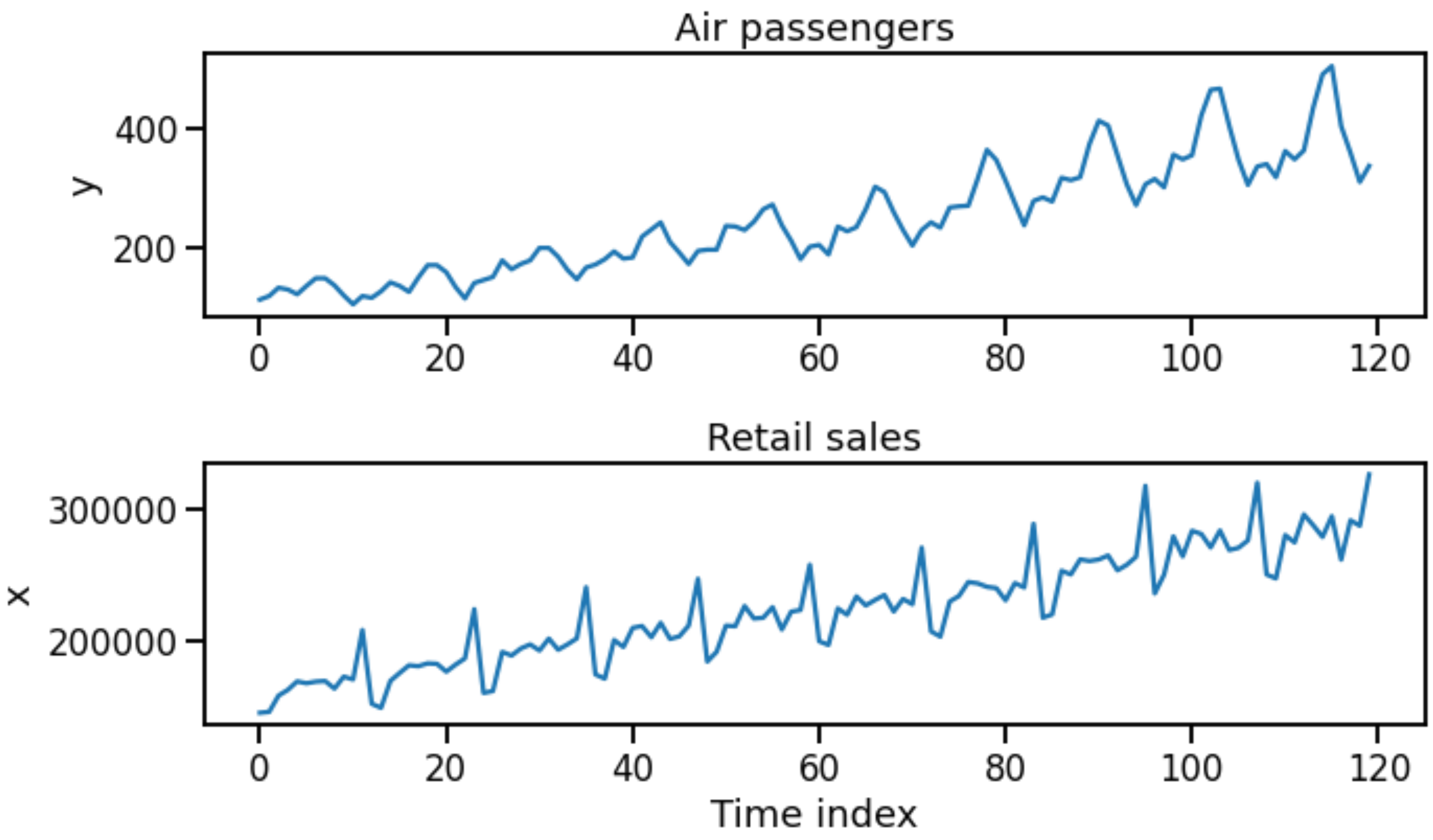


AR1 process: CCF



We see many significant lags even though only one is important (lag 10). Nevertheless, we see that the CCF peaks at the lag of 10 allowing us to determine that it is an important lag.

Two time series with trend and seasonality



These two time series measure very different quantities and were recorded over different decades (the first in 40s-60s and the latter in the 90s-00s).

The index is just for alignment.

We will show how just having a trend component will create a correlation between these two series.

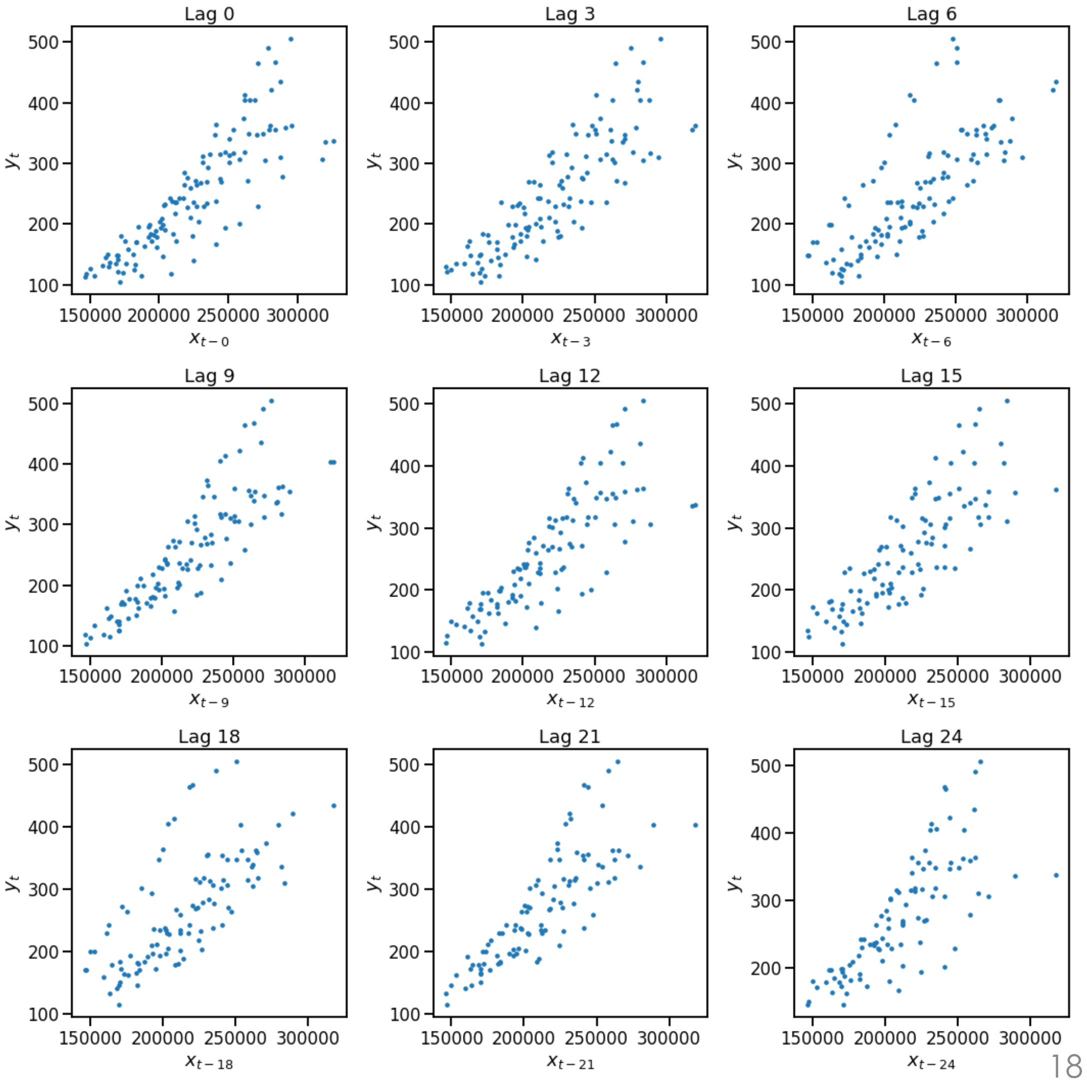
They both have yearly seasonality.

Two time series with trend and seasonality: Lag plot

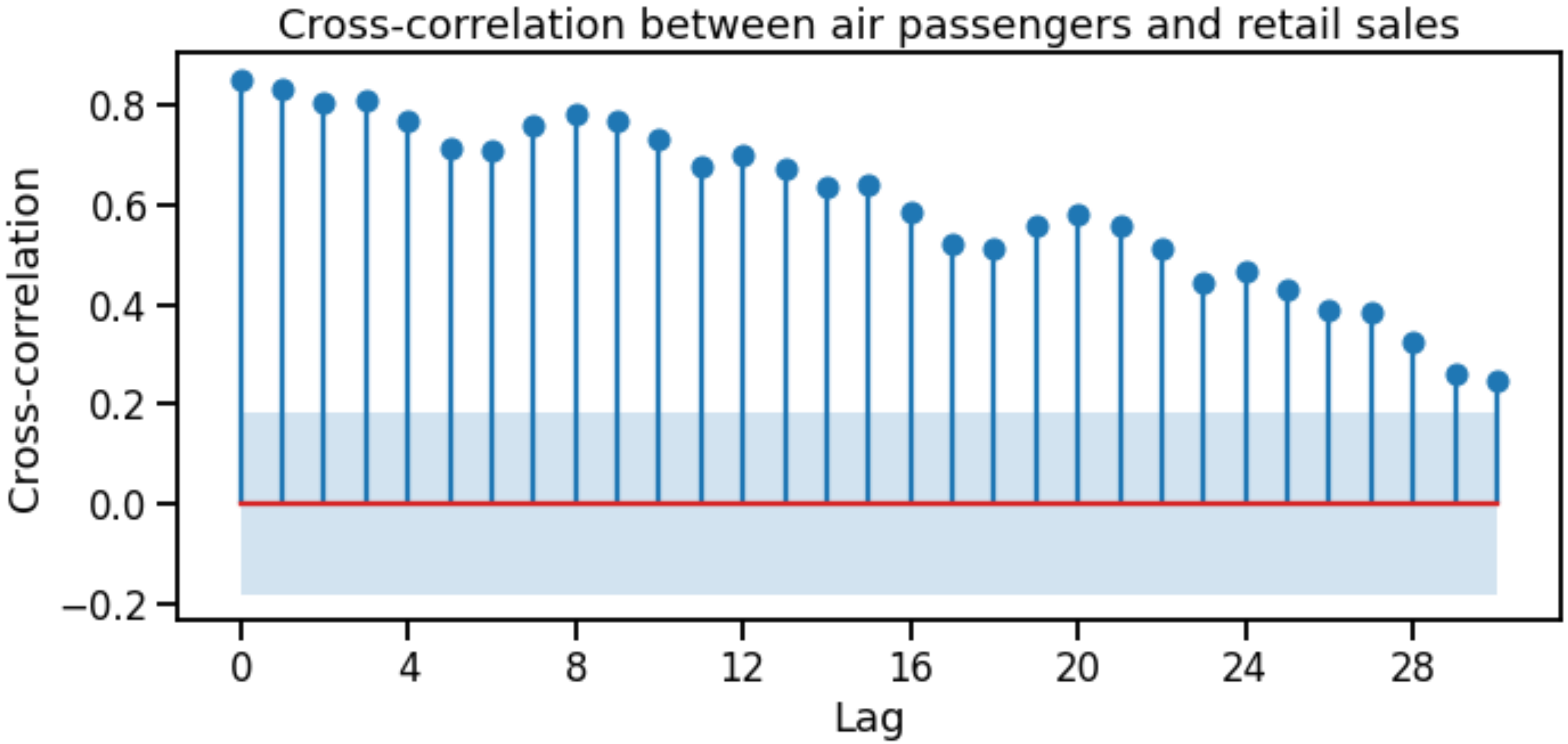
- Due to the trend we see that when one time series is small so is the other, when one is large so is the other.

- As a result the time series are correlated at many different lags even though we know these two time series are not causally linked.

- In fact any two time series with a trend would be cross-correlated. Therefore, the CCF is not useful for these cases.

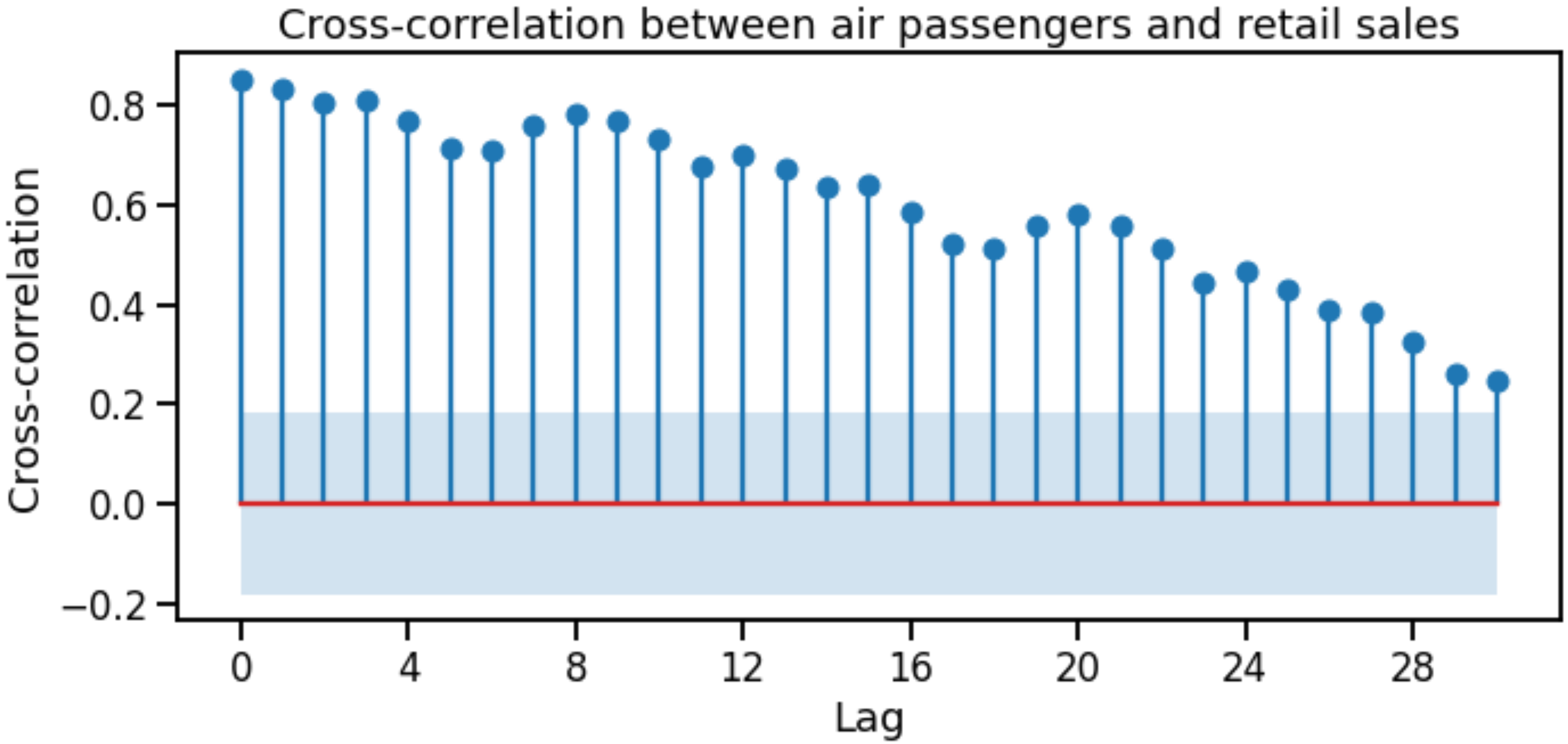


Two time series with trend and seasonality: CCF



We see many significant lags in the CCF due to the trend. We also see oscillations in the CCF as both time series have seasonality.

Two time series with trend and seasonality: CCF



Take away message: De-trend and de-seasonalise (i.e., try to make stationary) data with strong trend and seasonality for the purposes of measuring the CCF between two time series.

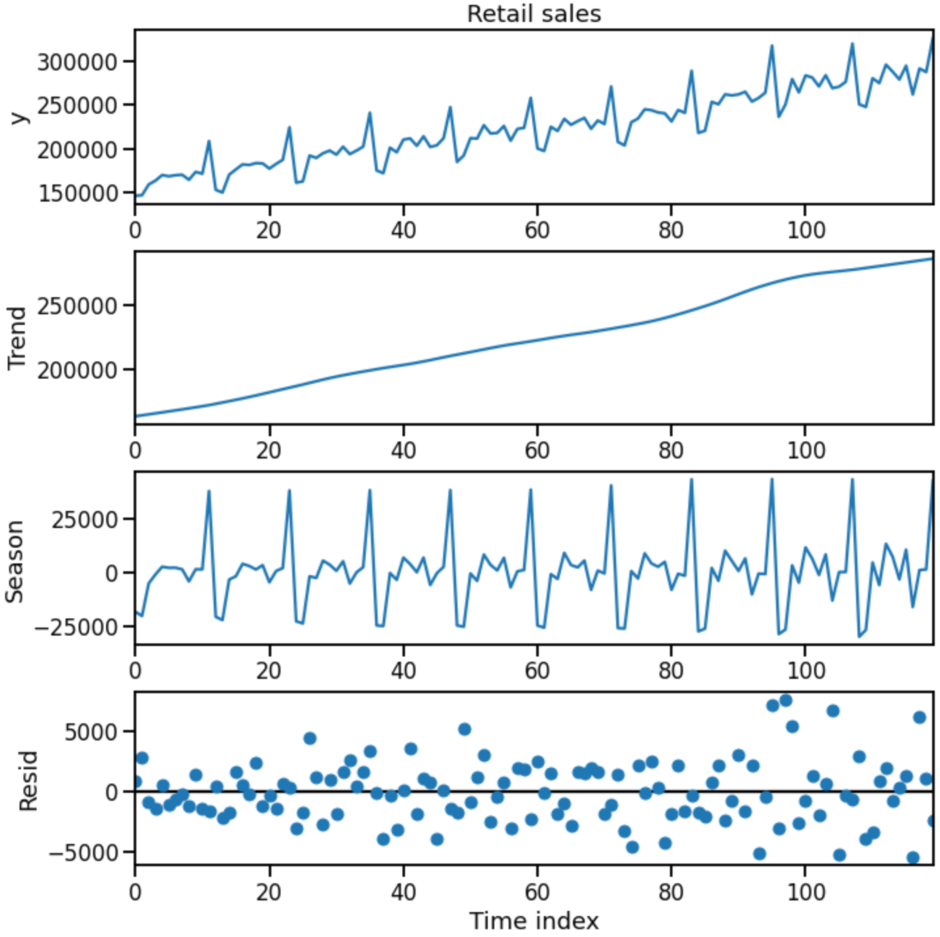
De-seasonalize and de-trend the data

- This is the STL decomposition of the retail sales dataset.

- The residual component is equivalent to y – trend – seasonality.

- This means the residual component is equivalent to de-trending and de-seasonalizing the data.

- This should result in a more stationary-looking time series.



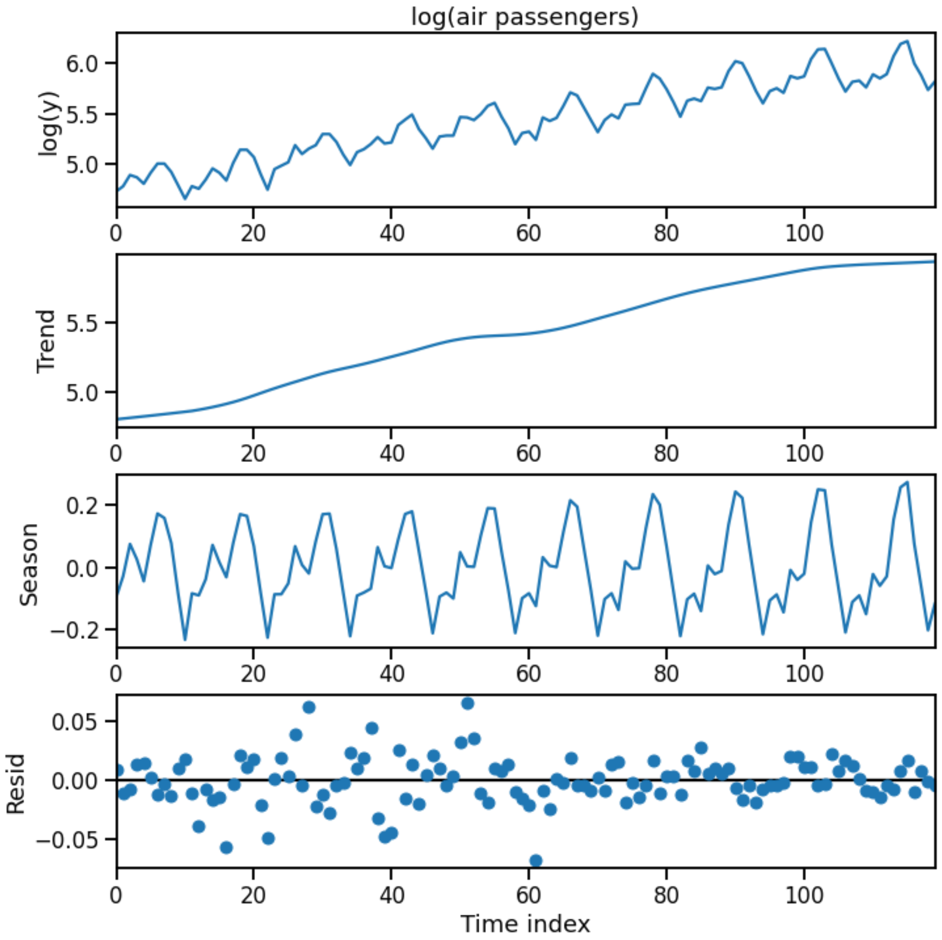
De-seasonalize and de-trend the data

- This is the STL decomposition of the log of the air passengers dataset.

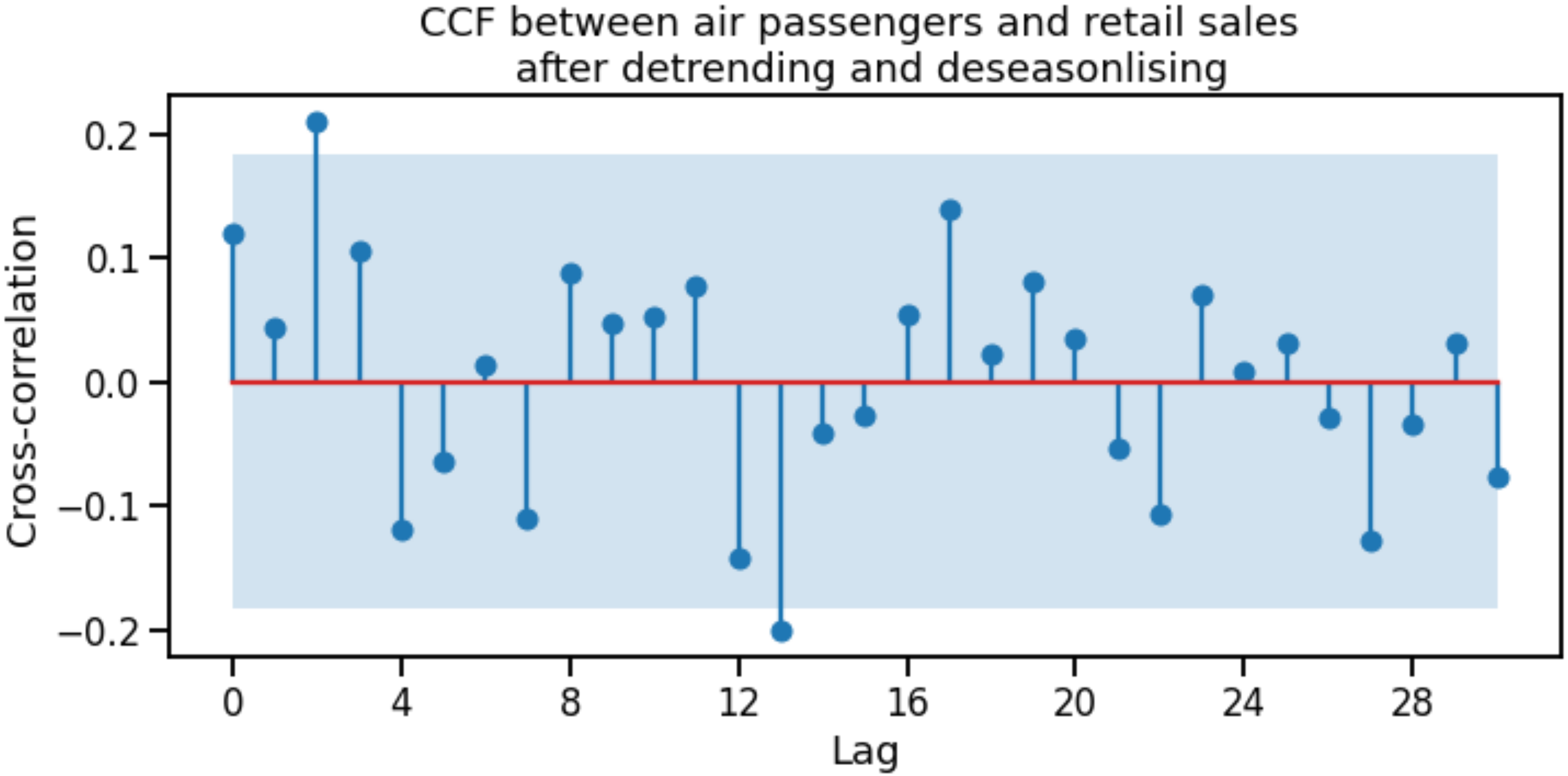
- The log is taken to as the air passenger dataset appears to be multiplicative and STL requires that the time series is additive.

- The residual component is equivalent to
  residual = y – trend – seasonality.

- When comparing with the retail dataset we will convert this back to the original units: exp(residual).

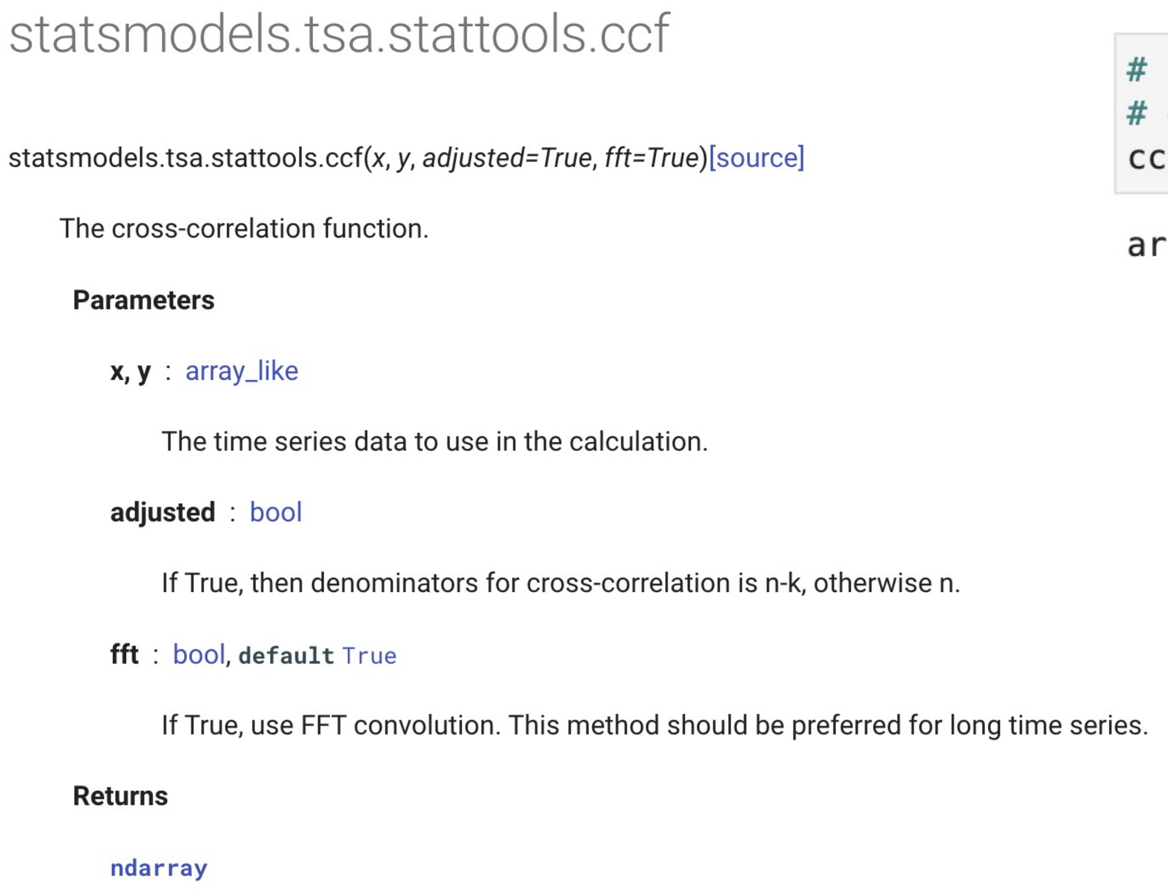


CCF after de-trending and de-seasonalising



The CCF shows much less correlation now. There are no large significant lags. This shows that there is not much information in one of these time series to predict the other, as we would expect.

CCF implementation in Statsmodels



statsmodels.tsa.stattools.ccf

statsmodels.tsa.stattools.ccf(*x*, *y*, *adjusted=True*, *fft=True*)[source]

The cross-correlation function.

Parameters

x, y : array_like

The time series data to use in the calculation.

adjusted : bool

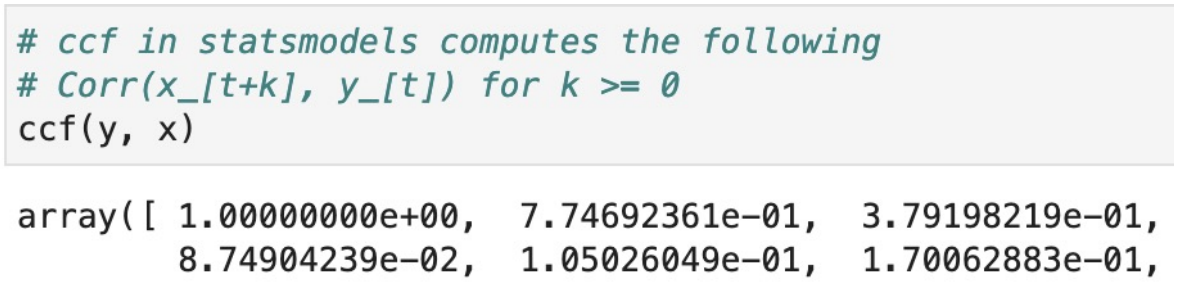
If True, then denominators for cross-correlation is n-k, otherwise n.

fft : bool, **default** True

If True, use FFT convolution. This method should be preferred for long time series.

Returns

ndarray



```
# ccf in statsmodels computes the following
# Corr(x_[t+k], y_[t]) for k >= 0
ccf(y, x)
```

```
array([ 1.00000000e+00,  7.74692361e-01,  3.79198219e-01,
        8.74904239e-02,  1.05026049e-01,  1.70062883e-01,
```

Summary

Cross-correlation function (CCF) measures how correlated y_t is with another variable at some lag x_{t-k}.

Large values in the CCF can help identify useful lags of a feature to use for forecasting.

Trend and seasonality can create spurious correlations. Try to ensure the data is stationary before using the CCF.