Partial autocorrelation function

Lag features

Contents

WHAT IS THE PACF

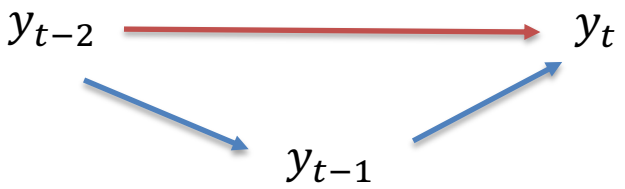
HOW TO USE PACF TO IDENTIFY USEFUL LAGS

Partial autocorrelation: motivation

- We want to know which lags y_{t-k} are most predictive of y_t.
- Imagine a process where only the previous lag matters (e.g., AR(1) process):

$$y_t = \phi_1 y_{t-1} + \epsilon_t \text{ where } \epsilon_t \sim \mathrm{N}(0,1)$$

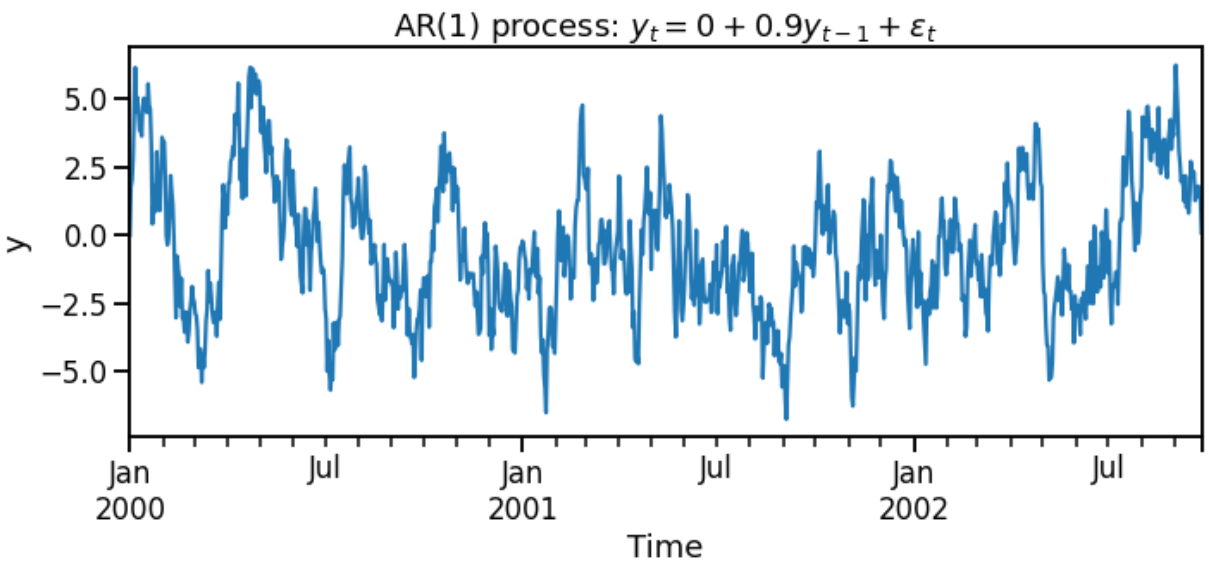
- y_{t-2} will influence y_{t-1} which influences y_t:

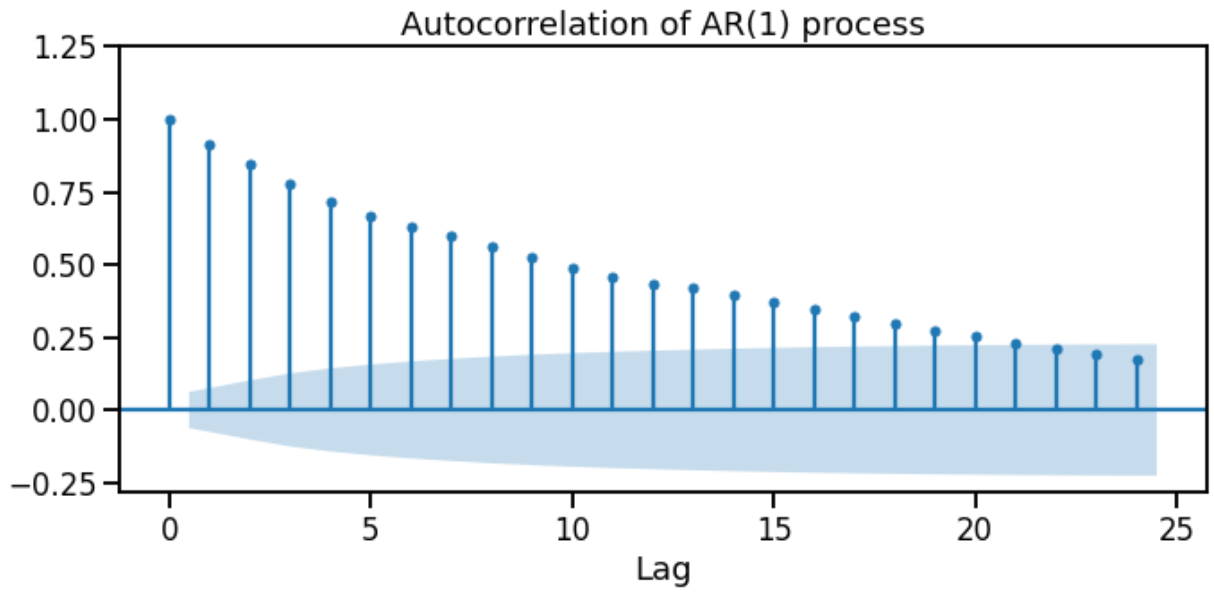


- This means that the autocorrelation between y_t and y_{t-2} will be non-zero even though only a lag of 1 matters.

AR(1) process

$$y_t = 0.9y_{t-1} + \epsilon_t \text{ where } \epsilon_t \sim \mathrm{N}(0,1)$$





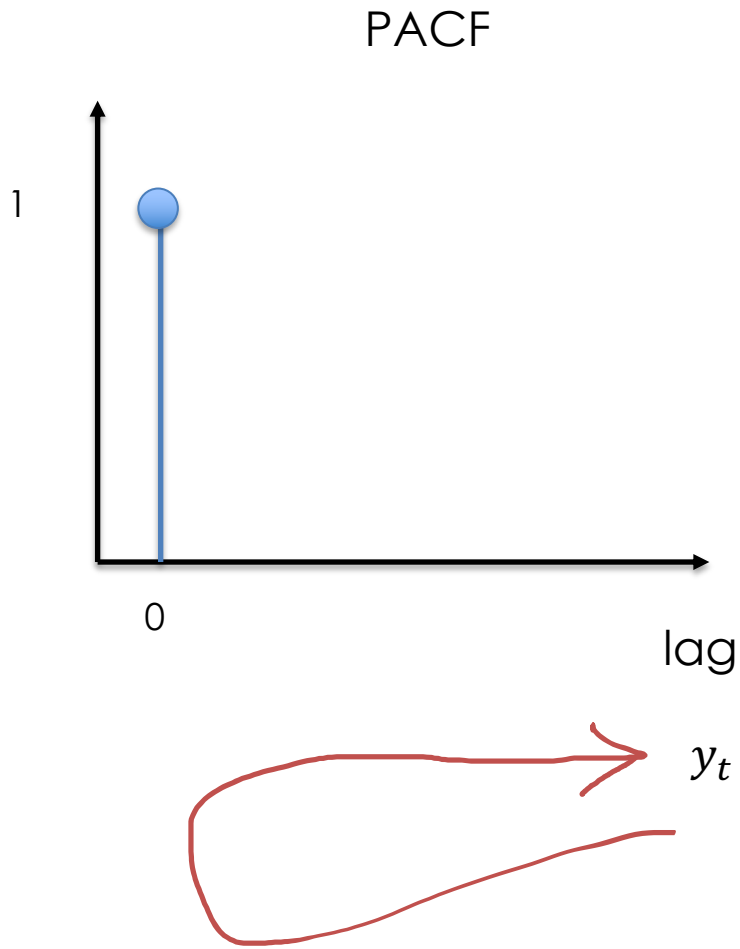
Even though the time series only depends on the previous lag the ACF has multiple significant lags for k > 1. We shall see the PACF picks up that lag 1 is special.

Partial autocorrelation function: main idea

- The partial autocorrelation function (PACF) measures the correlation between y_t and y_{t-k} after removing the correlation introduced by intermediate lags on y_t & y_{t-k}.

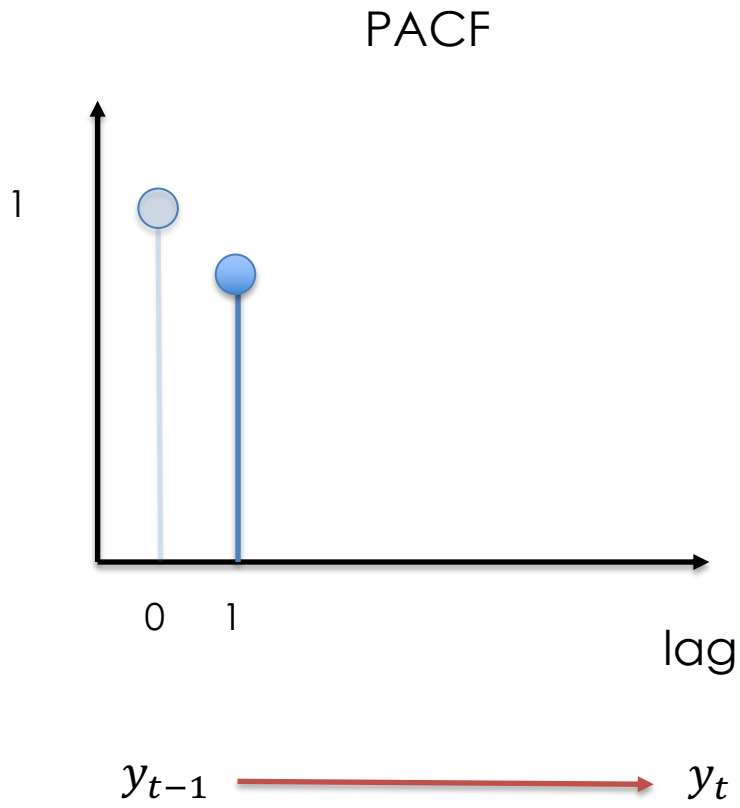
Partial autocorrelation function: main idea

- The partial autocorrelation function (PACF) measures the correlation between y_t and y_{t-k} after removing the correlation introduced by intermediate lags on y_t & y_{t-k}.



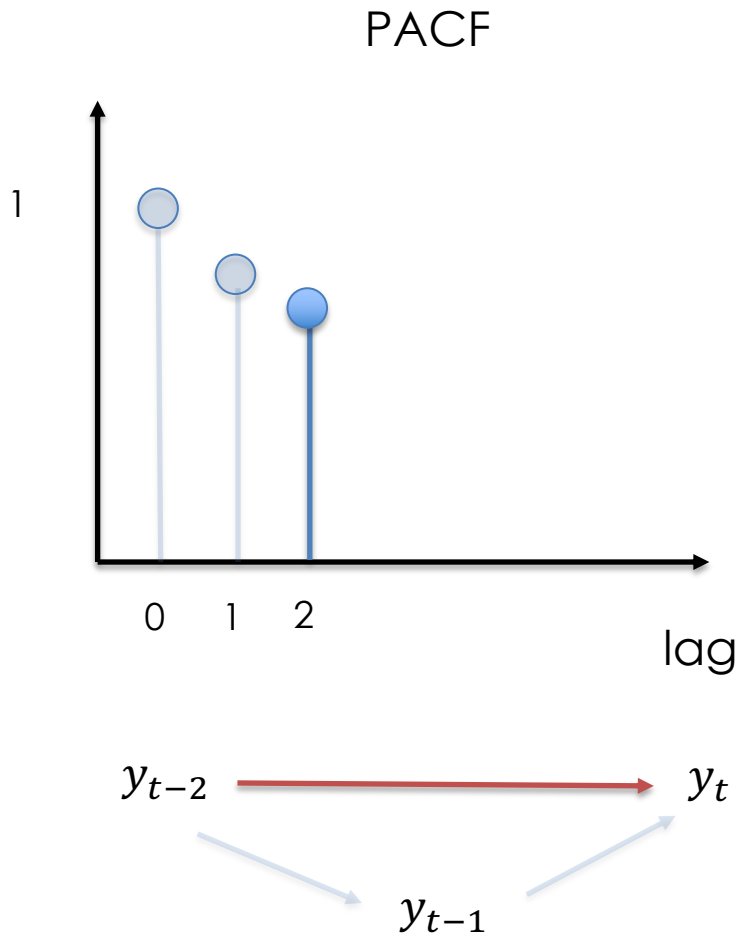
Partial autocorrelation function: main idea

- The partial autocorrelation function (PACF) measures the correlation between y_t and y_{t-k} after removing the correlation introduced by intermediate lags on y_t & y_{t-k}.
- The PACF at **lag 1** is the correlation between $\boldsymbol{y_t}$ **&** $\boldsymbol{y_{t-1}}$.



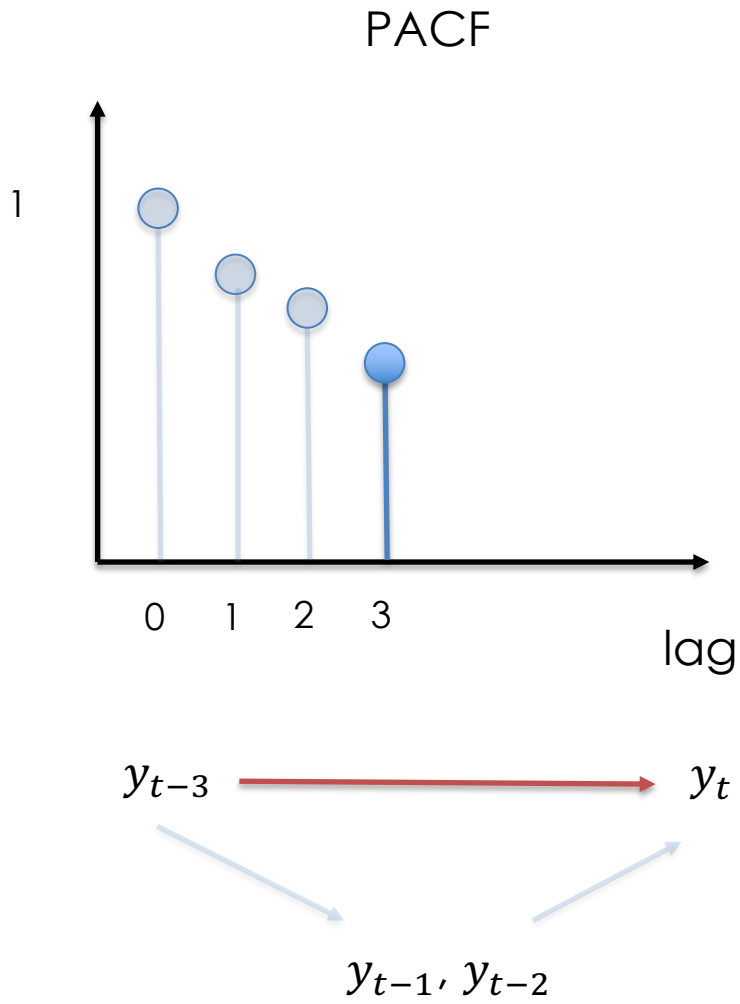
Partial autocorrelation function: main idea

- The partial autocorrelation function (PACF) measures the correlation between y_t and y_{t-k} after removing the correlation introduced by intermediate lags on y_t & y_{t-k}.
- The PACF at lag 1 is the correlation between y_t & y_{t-1}.
- The PACF at **lag 2** is the correlation between $\boldsymbol{y_t}$ **&** $\boldsymbol{y_{t-2}}$ after **removing the effects of** $\boldsymbol{y_{t-1}}$.



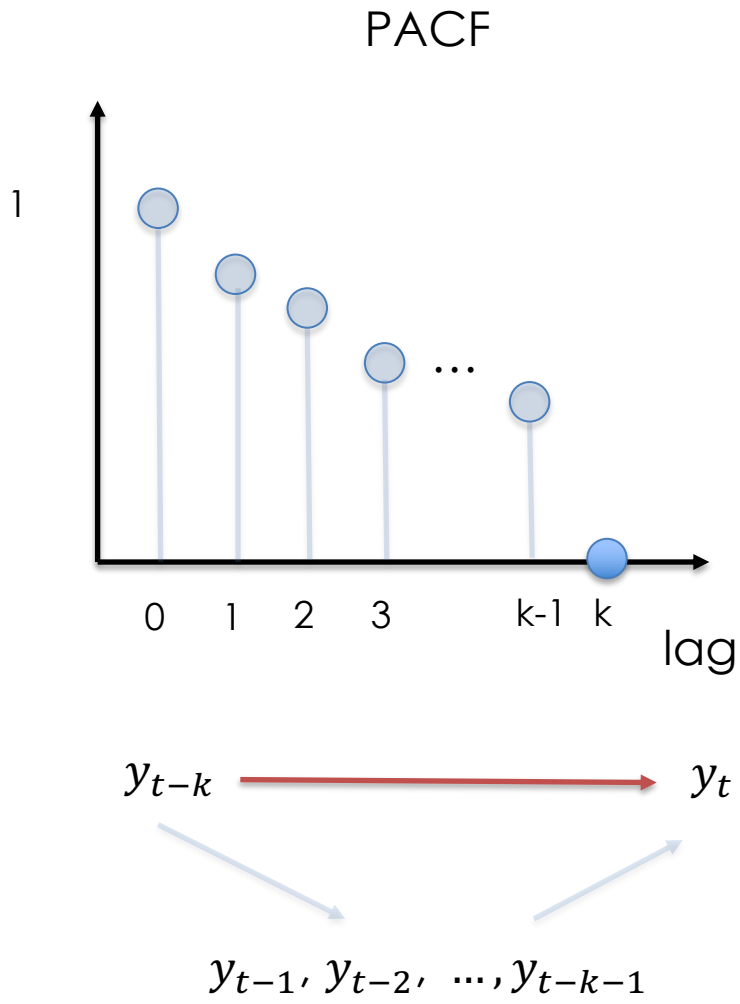
Partial autocorrelation function: main idea

- The partial autocorrelation function (PACF) measures the correlation between y_t and y_{t-k} after removing the correlation introduced by intermediate lags on y_t & y_{t-k}.
- The PACF at lag 1 is the correlation between y_t & y_{t-1}.
- The PACF at lag 2 is the correlation between y_t & y_{t-2} after removing the effects of y_{t-1}.
- The PACF at **lag 3** is the correlation between $\boldsymbol{y_t}$ **&** $\boldsymbol{y_{t-2}}$ after **removing the effects of** $\boldsymbol{y_{t-1}}$ **and** $\boldsymbol{y_{t-2}}$.



Partial autocorrelation function: main idea

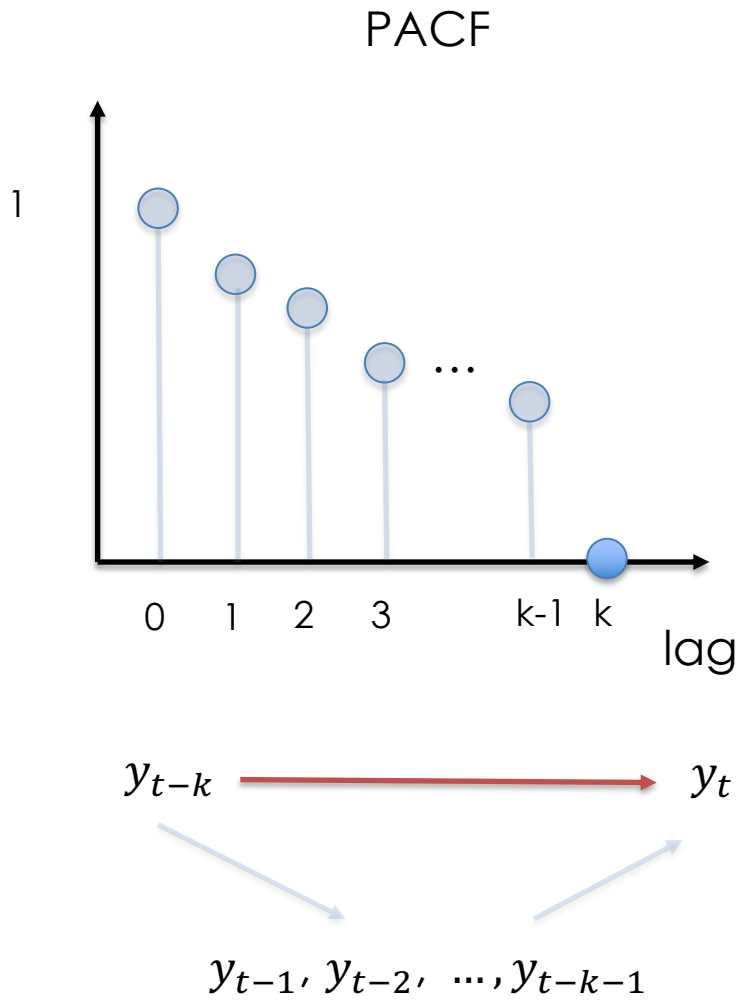
- The partial autocorrelation function (PACF) measures the correlation between y_t and y_{t-k} after removing the correlation introduced by intermediate lags on y_t & y_{t-k}.
- The PACF at lag 1 is the correlation between y_t & y_{t-1}.
- The PACF at lag 2 is the correlation between y_t & y_{t-2} after removing the effects of y_{t-1}.
- The PACF at lag 3 is the correlation between y_t & y_{t-2} after removing the effects of y_{t-1} and y_{t-2}.
- The PACF at **lag k** is the correlation between $\boldsymbol{y_t}$ **&** $\boldsymbol{y_{t-k}}$ after **removing the effects of** $\boldsymbol{y_{t-1}, y_{t-2}, \cdots, y_{t-k-1}}$.



Partial autocorrelation function: main idea

- The partial autocorrelation function (PACF) measures the correlation between y_t and y_{t-k} after removing the correlation introduced by intermediate lags on y_t & y_{t-k}.

- This means that the correlation at lag k is high only if it adds additional information that is not already accounted for by all the lags prior to it.



PACF: how is it calculated?

- We want to compute a correlation between y_t and y_{t-k} which accounts for the correlation introduced from intermediate lags $\{y_{t-1}, y_{t-2}, \ldots, y_{t-k+1}\}$.
- How do we remove the effects of intermediate lags? By subtracting the the linear impact of the intermediate lags on y_t and y_{t-k} as given by a linear regression.

Regress y_t on $\{y_{t-1}, y_{t-2}, \ldots, y_{t-k+1}\}$

$$\hat{y}_t = \beta_0 + \beta_1 y_{t-1} + \cdots + \beta_{k+1} y_{t-k+1}$$

Regress y_{t-k} on $\{y_{t-1}, y_{t-2}, \ldots, y_{t-k+1}\}$

$$\hat{y}_{t-k} = \alpha_0 + \alpha_1 y_{t-1} + \cdots + \alpha_{k+1} y_{t-k+1}$$

PACF: how is it calculated?

The PACF is given by:

$$corr(y_t, y_{t-1})\ ;\ k = 1$$ ---------- Same as ACF(k=1)

$$corr(y_t - \hat{y}_t, y_{t-k} - \hat{y}_{t-k})\ ;\ k > 2$$

where

$$\hat{y}_t = \beta_0 + \beta_1 y_{t-1} + \cdots + \beta_{k+1} y_{t-k+1}$$

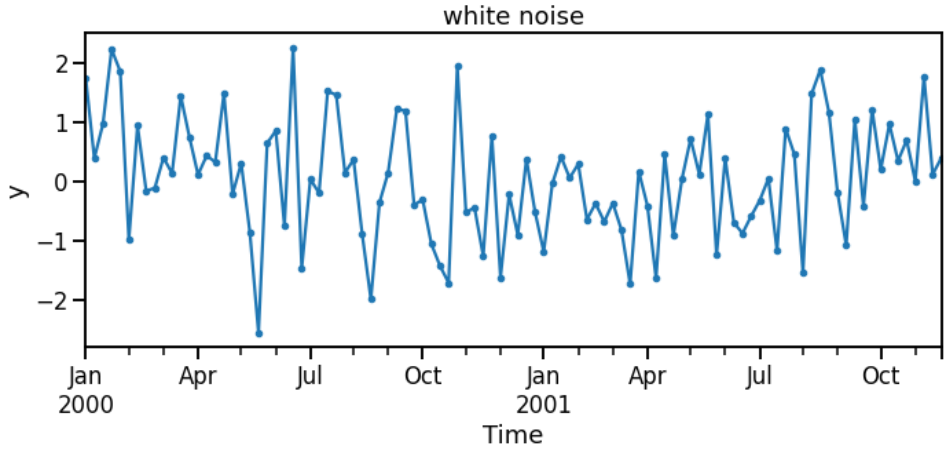
$$\hat{y}_{t-k} = \alpha_0 + \alpha_1 y_{t-1} + \cdots + \alpha_{k+1} y_{t-k+1}$$

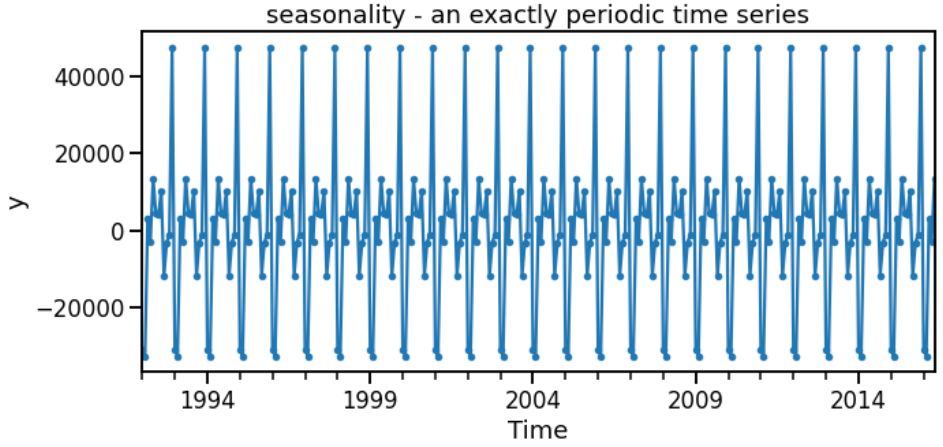
PACF: how is it calculated?

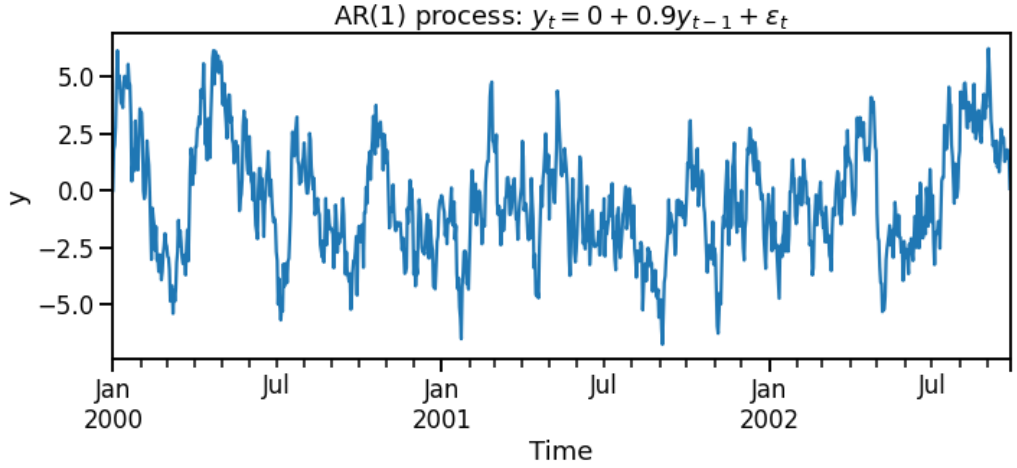
- In practice, software packages implement more efficient ways of calculating the PACF [1].
- An assumption when calculating the PACF is that the time series, y_t, is stationary which means the following should **not change with time**:
  - Mean: try de-trending the data if needed.
  - Variance: log transform the data to stabilize variance if needed.
  - Autocorrelation: this means that the correlation between y_t and y_{t-k} should not depend on t. There are no simple transforms of the data to try to enforce this.
  - We can still get some information from the PACF even when these assumptions are not met exactly, however, it can make it difficult to interpret the PACF.

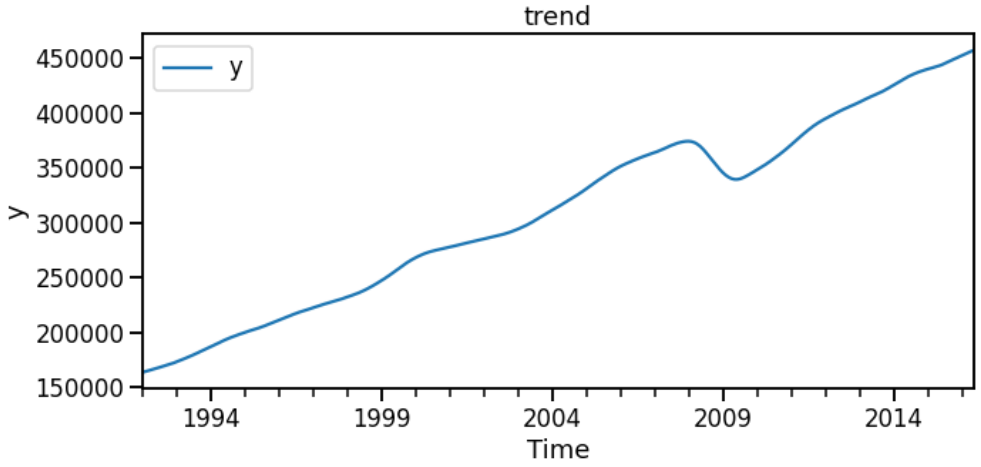
[1] - George, E. P. "Box. Time series analysis: forecasting and control." (1970).

Let's look at the PACF for different time series



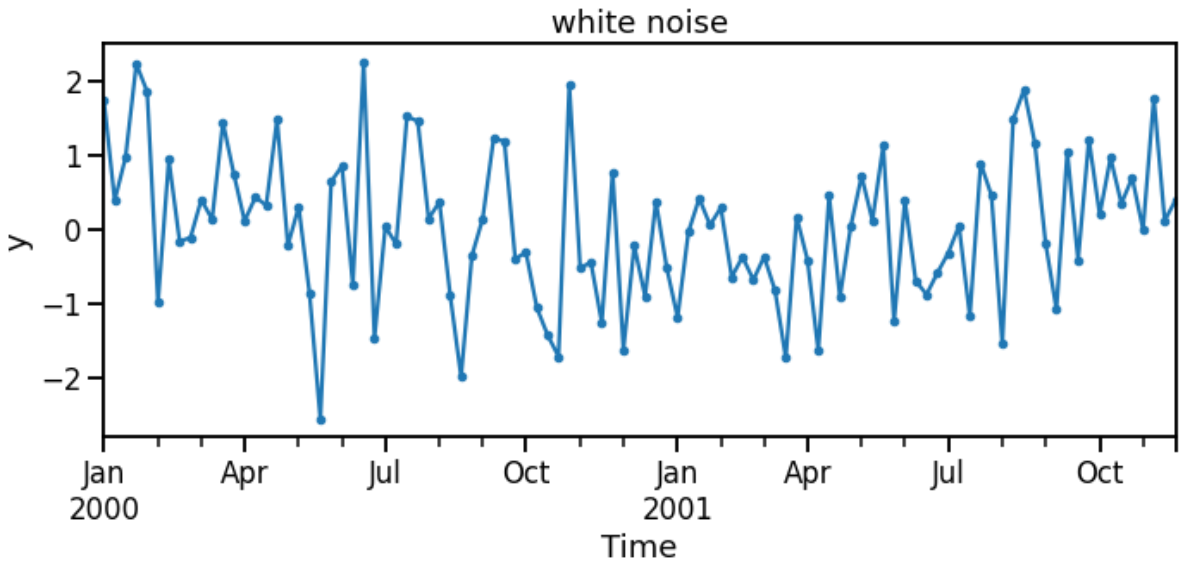


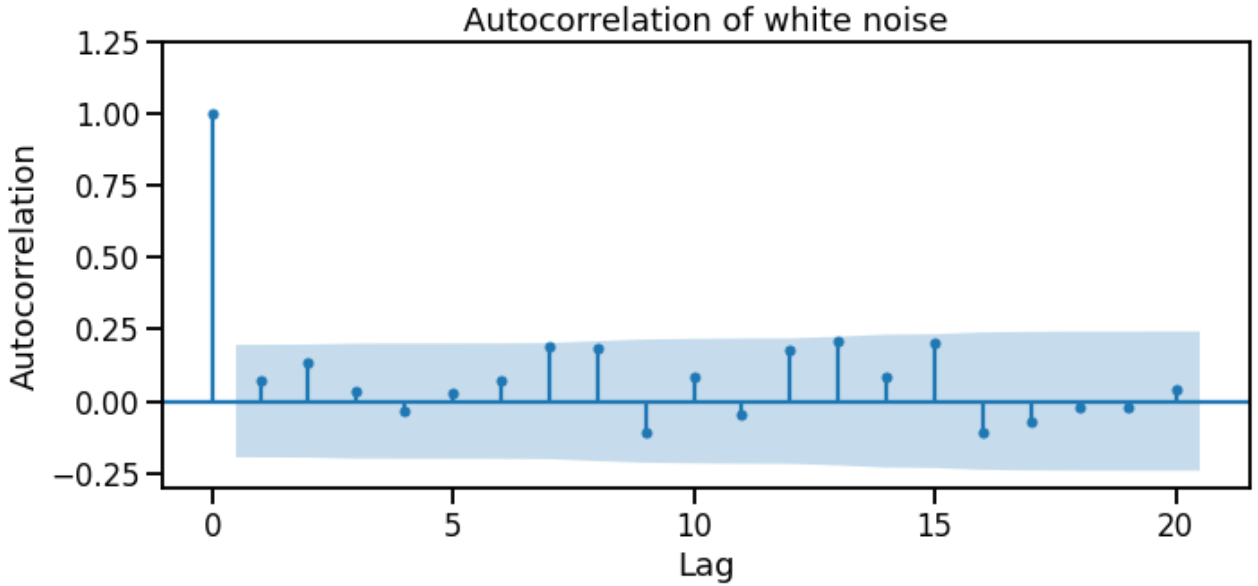




White noise

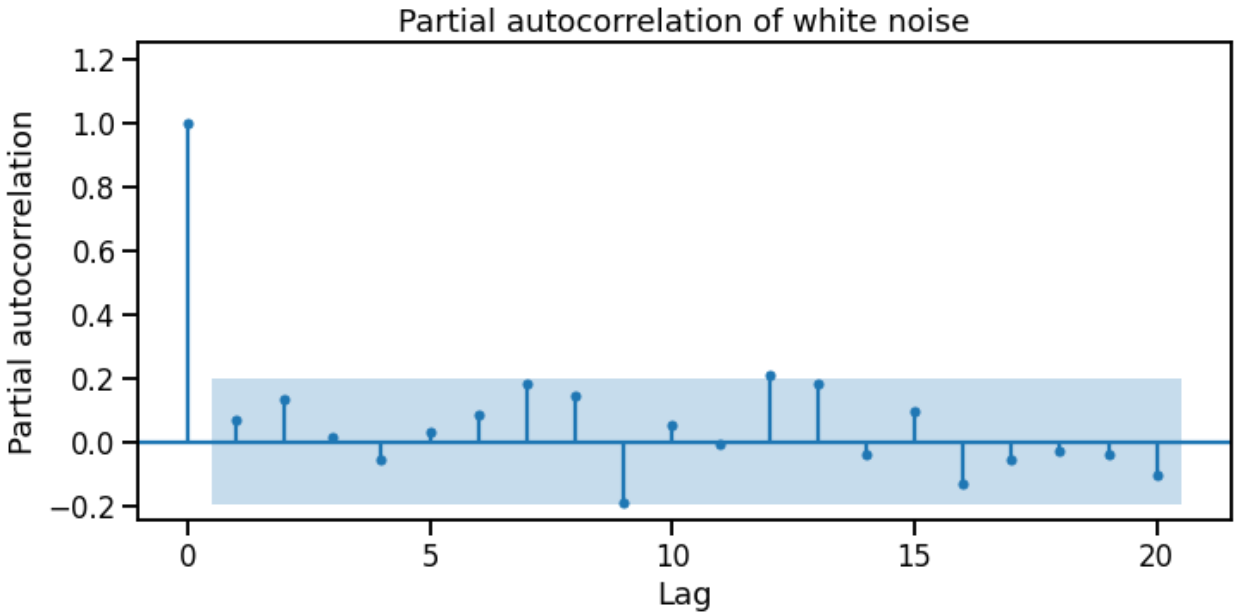
$y_t = \epsilon_t$ where $\epsilon_t \sim \mathrm{N}(0,1)$

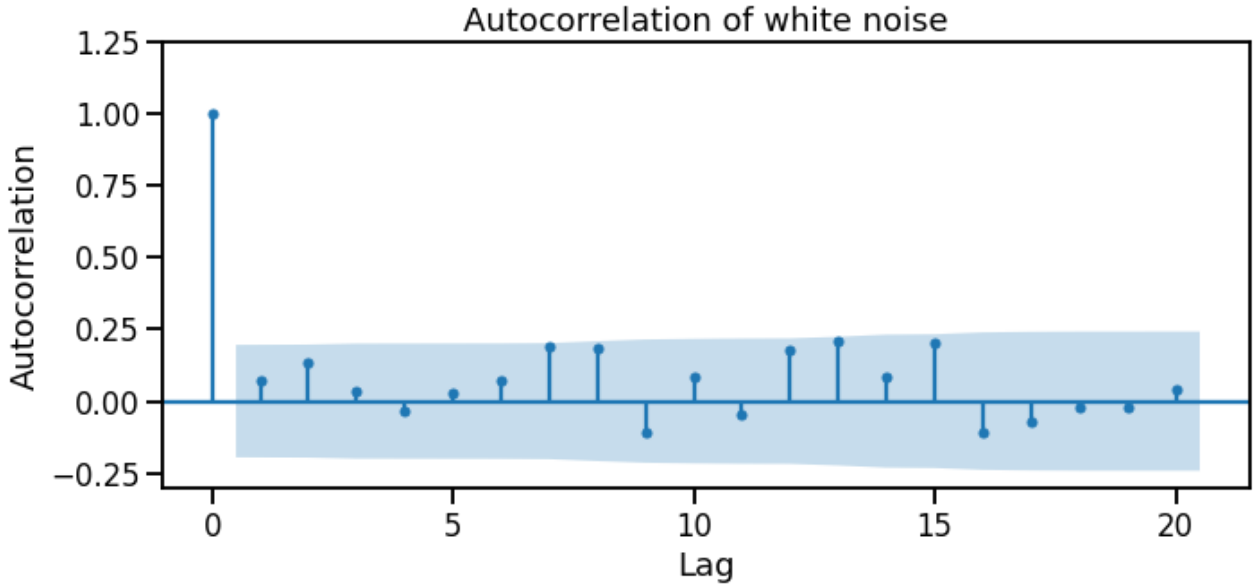




White noise

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

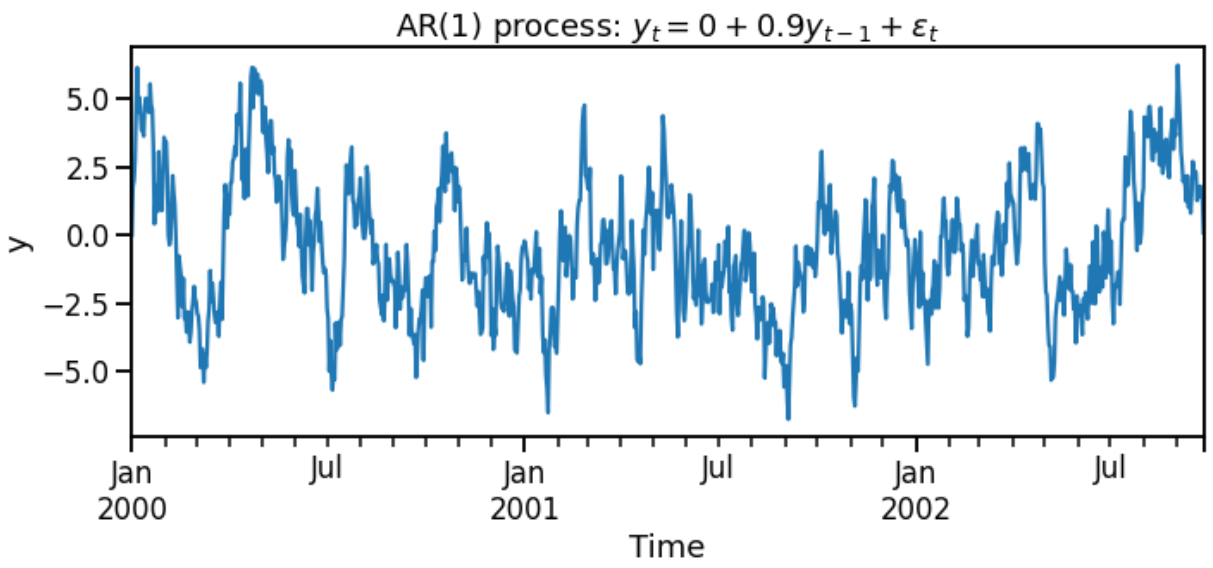


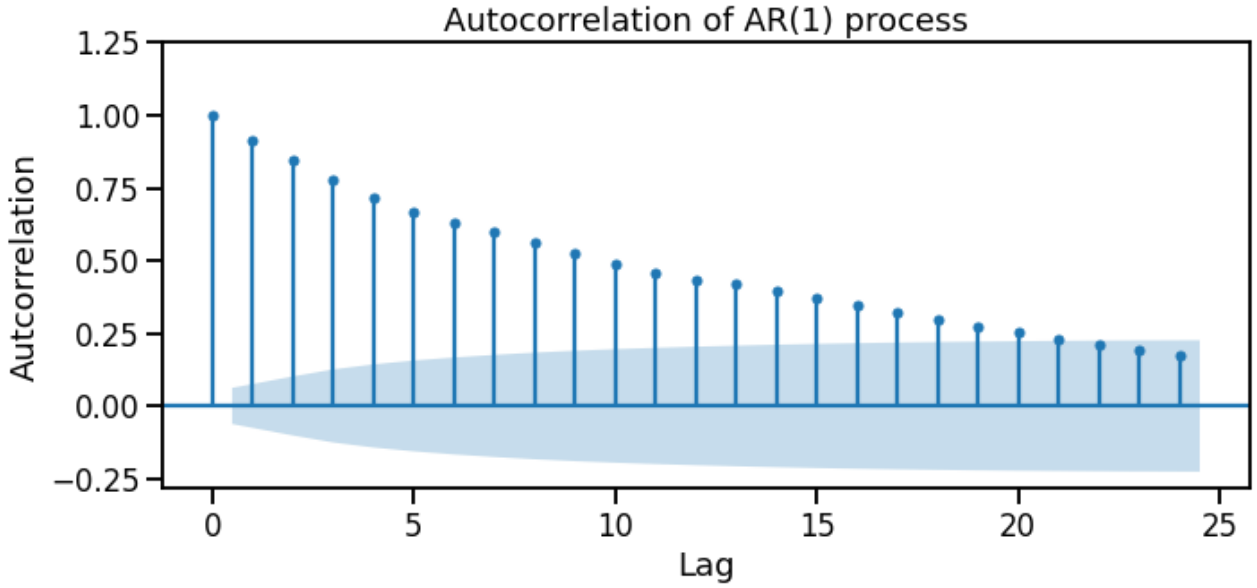


The PACF for white noise shows no large significant partial autocorrelations at any lag as we would expect.

AR(1) process

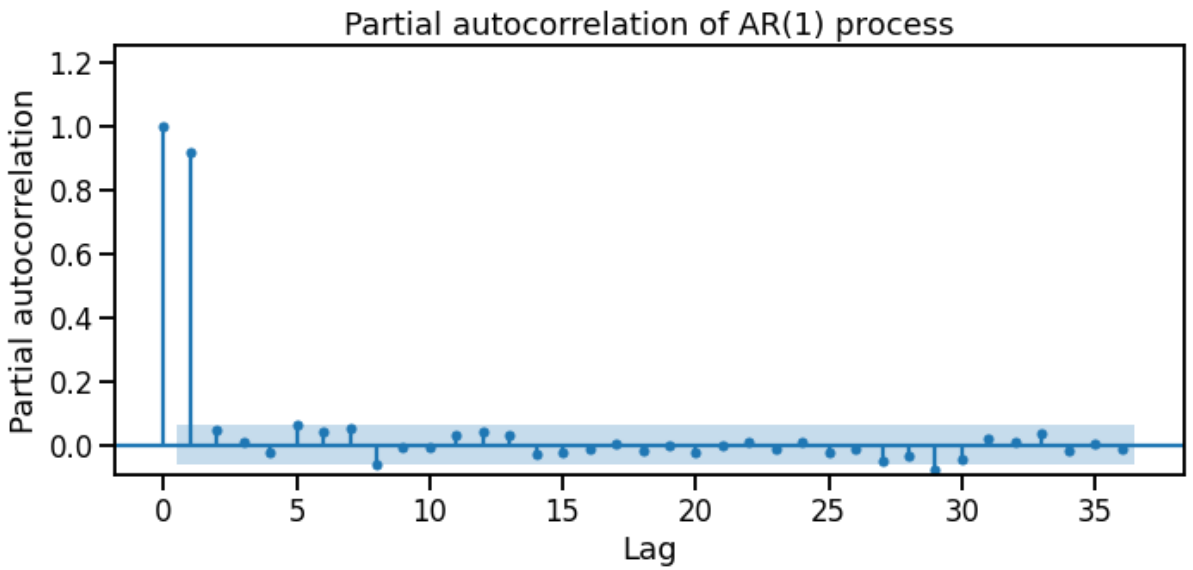
$$y_t = 0.9y_{t-1} + \epsilon_t \text{ where } \epsilon_t \sim \mathrm{N}(0,1)$$

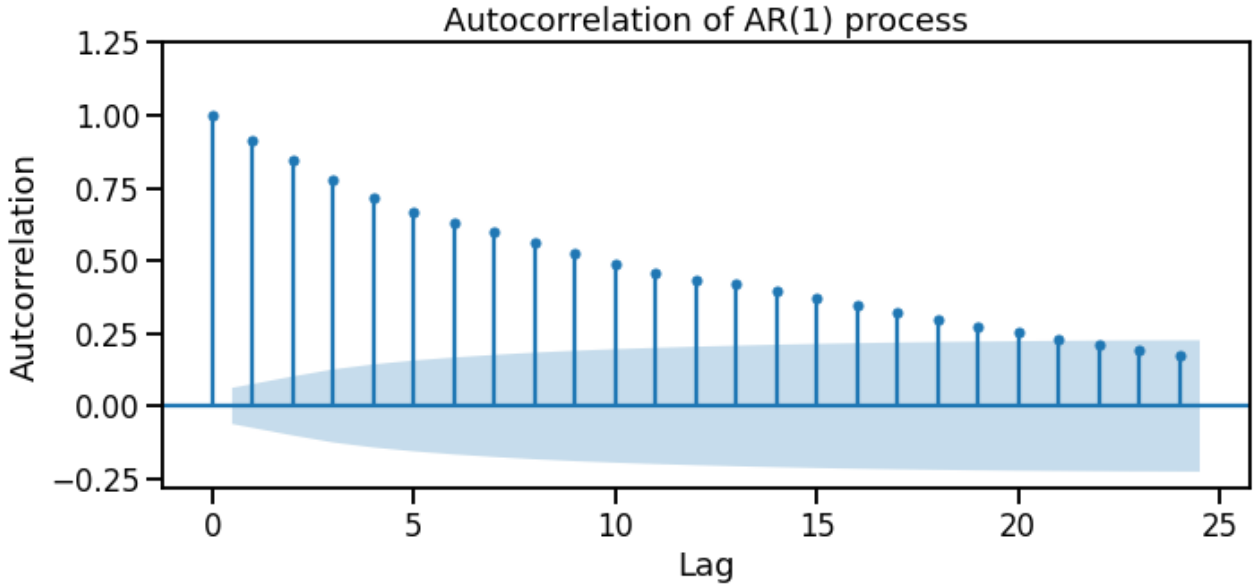




AR(1) process

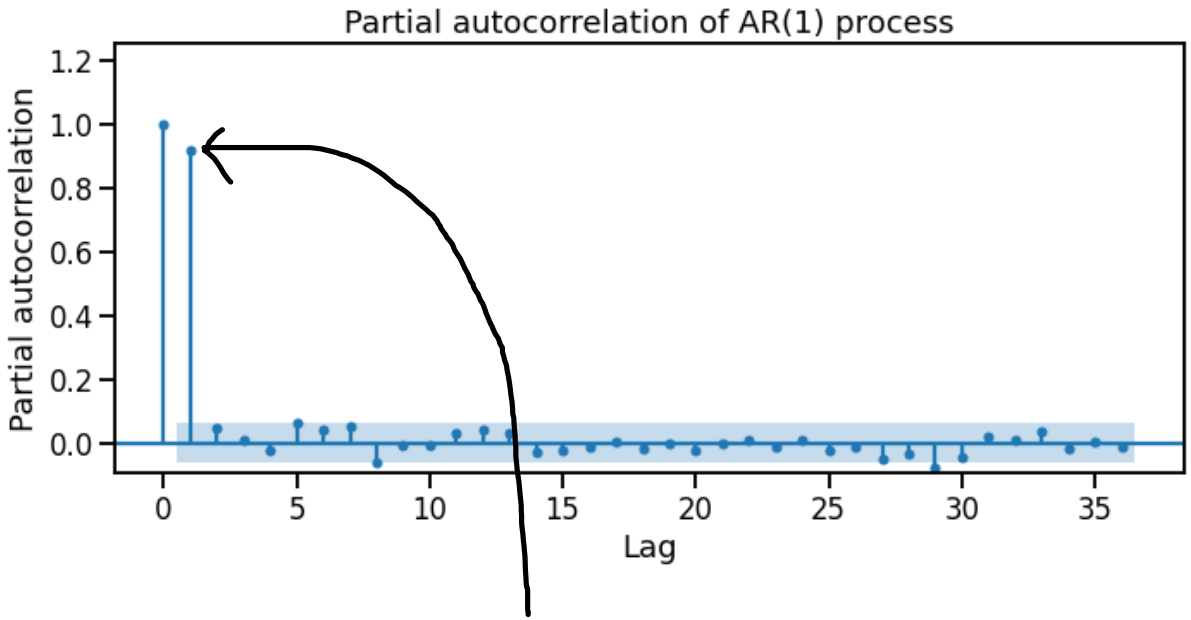
$$y_t = 0.9y_{t-1} + \epsilon_t \text{ where } \epsilon_t \sim \mathrm{N}(0,1)$$

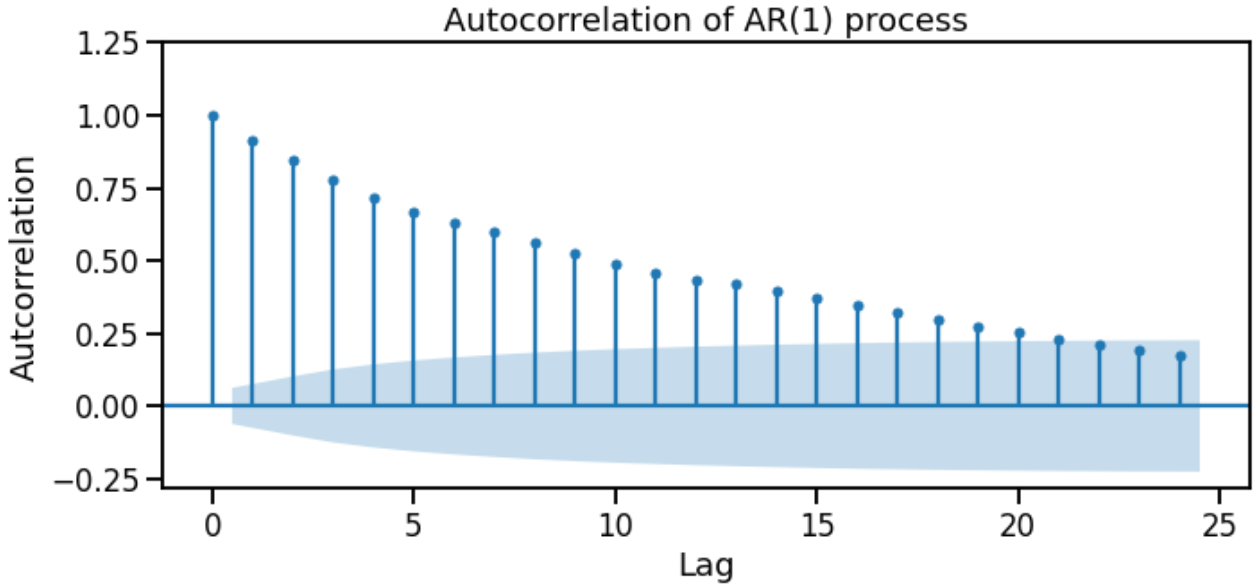




AR(1) process

$$y_t = 0.9y_{t-1} + \epsilon_t \text{ where } \epsilon_t \sim \mathrm{N}(0,1)$$

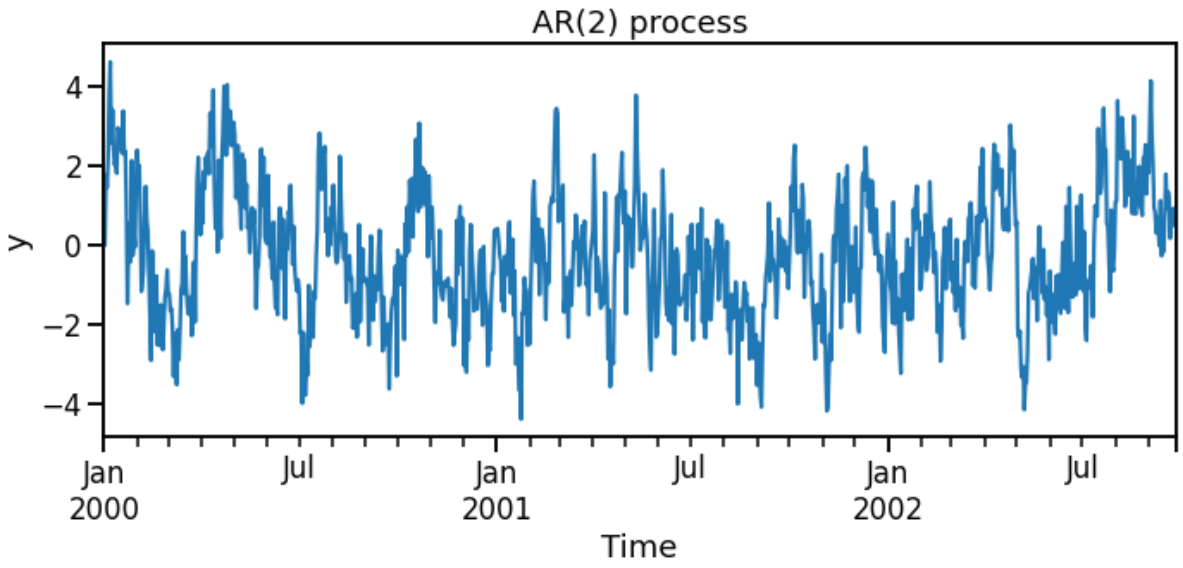


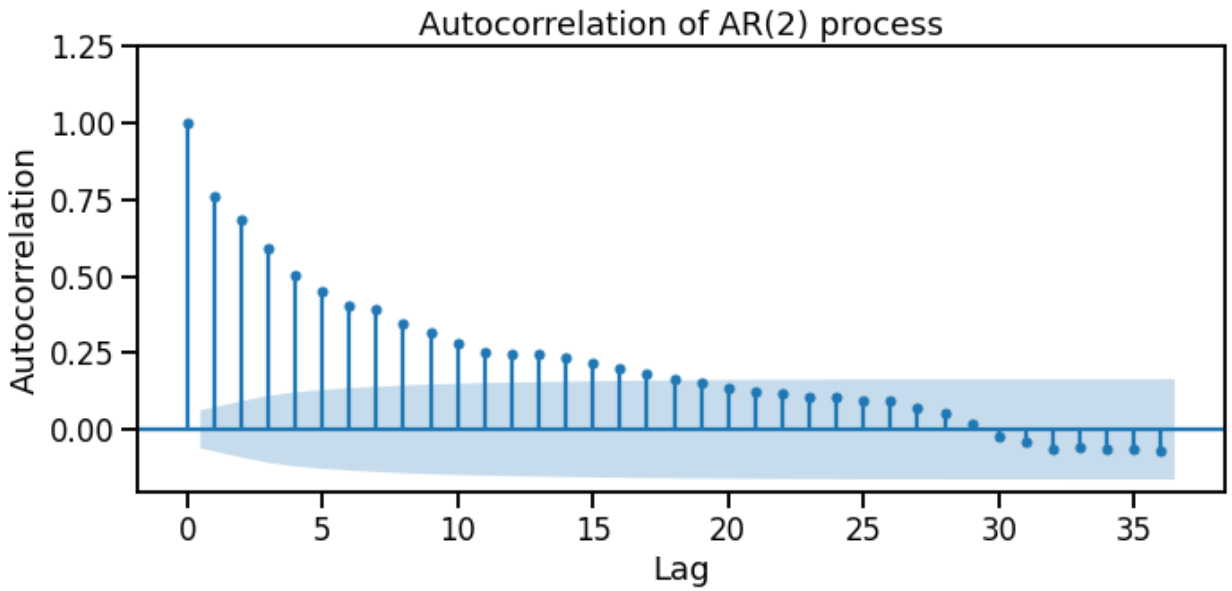


The PACF is large for k = 1 and not significant for k > 2 as we wanted!

AR(2) process

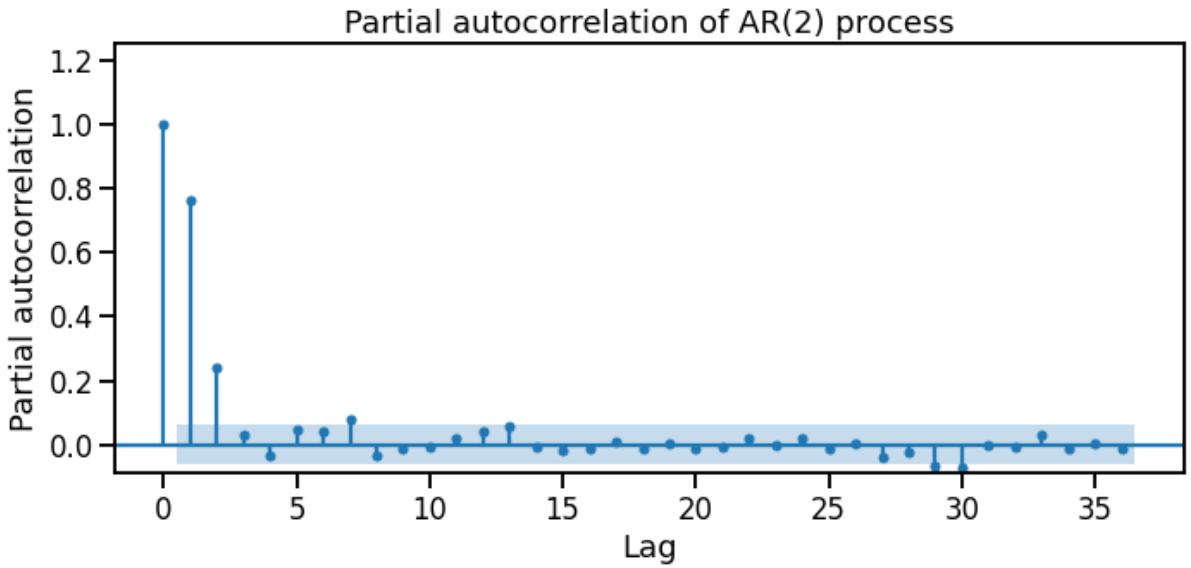
$$y_t = 0.6y_{t-1} + 0.2y_{t-2} + \epsilon_t \text{ where } \epsilon_t \sim \mathrm{N}(0,1)$$

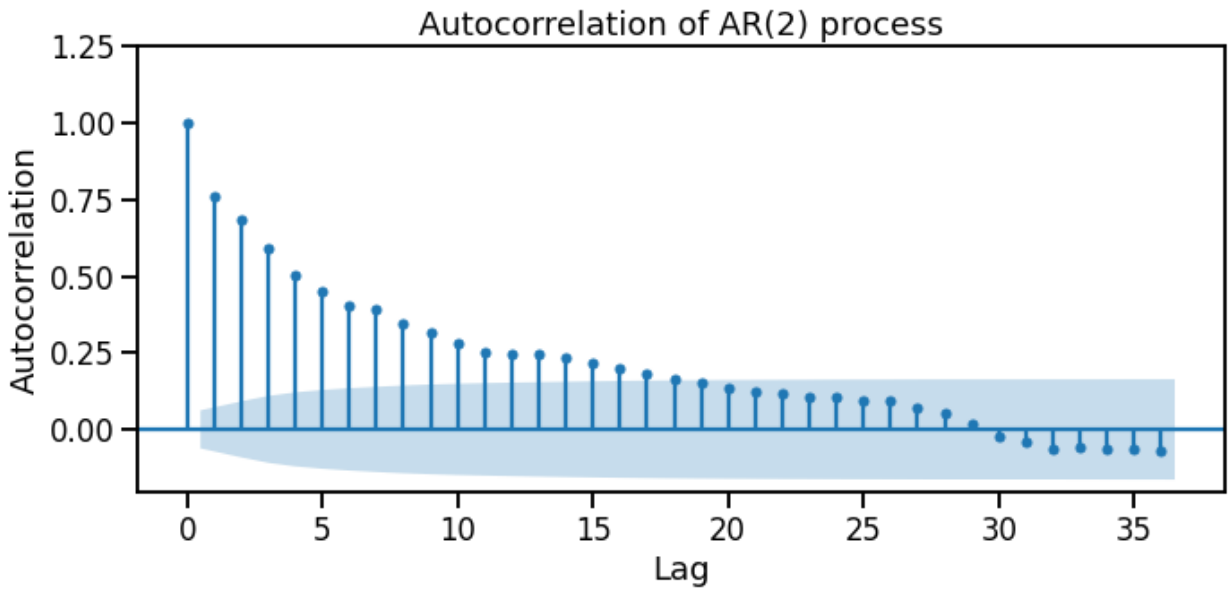




AR(2) process

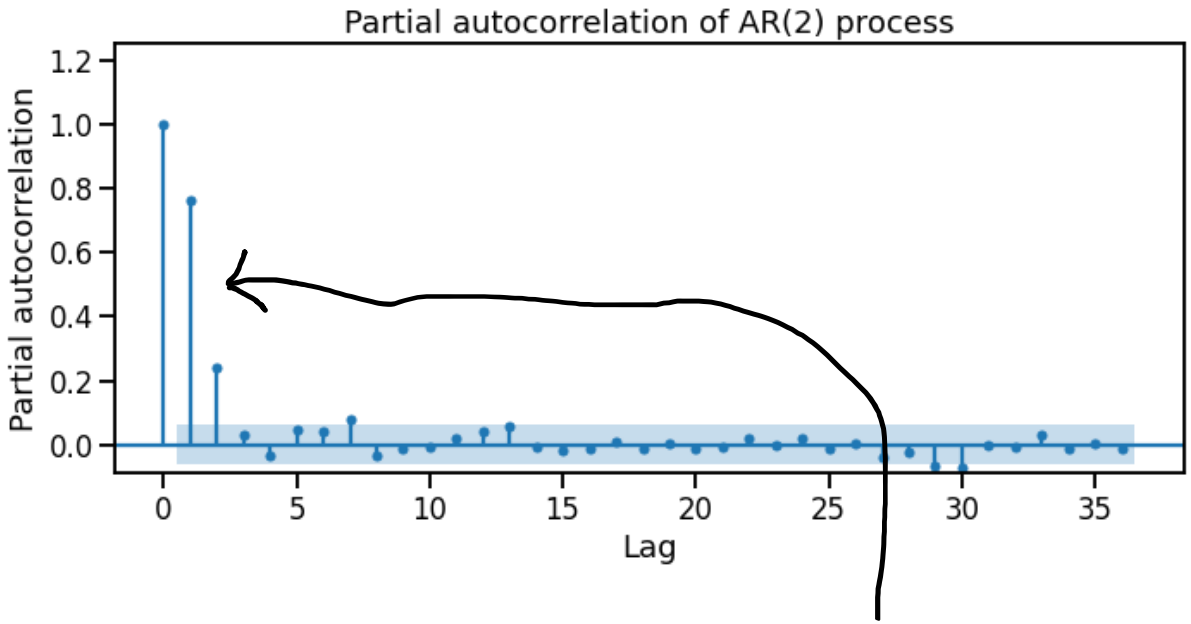
$$y_t = 0.6y_{t-1} + 0.2y_{t-2} + \epsilon_t \text{ where } \epsilon_t \sim \mathrm{N}(0,1)$$

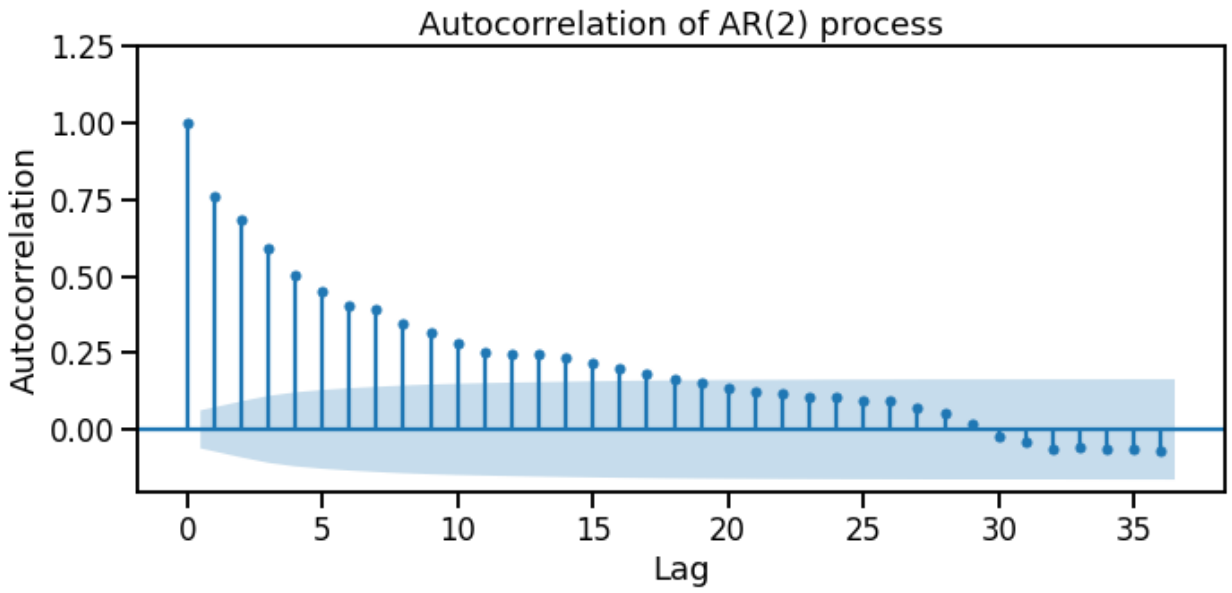




AR(2) process

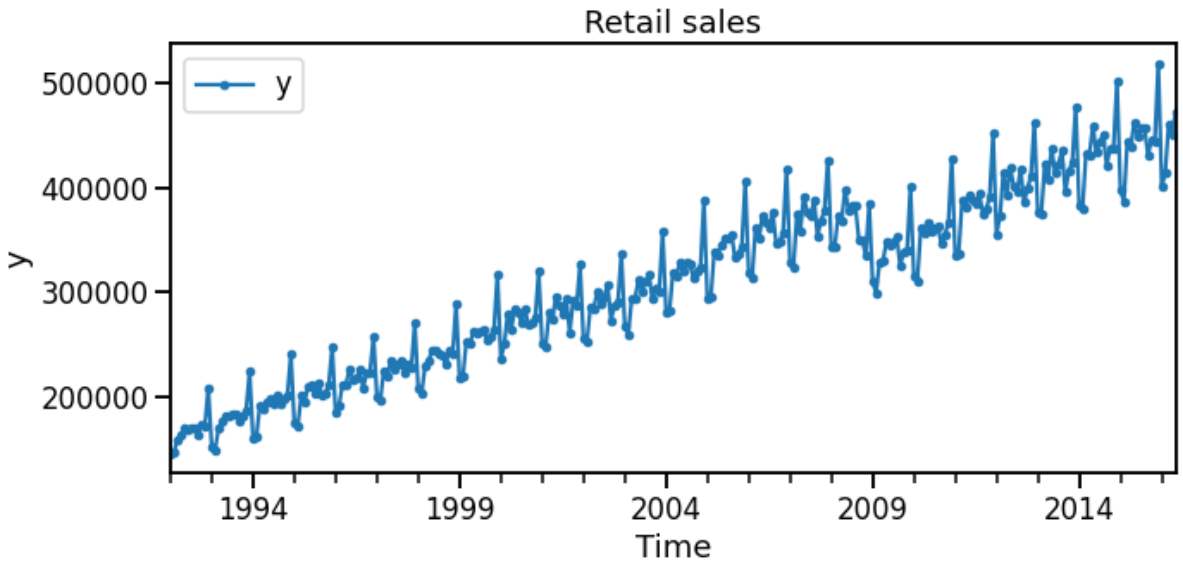
$$y_t = 0.6y_{t-1} + 0.2y_{t-2} + \epsilon_t \text{ where } \epsilon_t \sim \mathrm{N}(0,1)$$

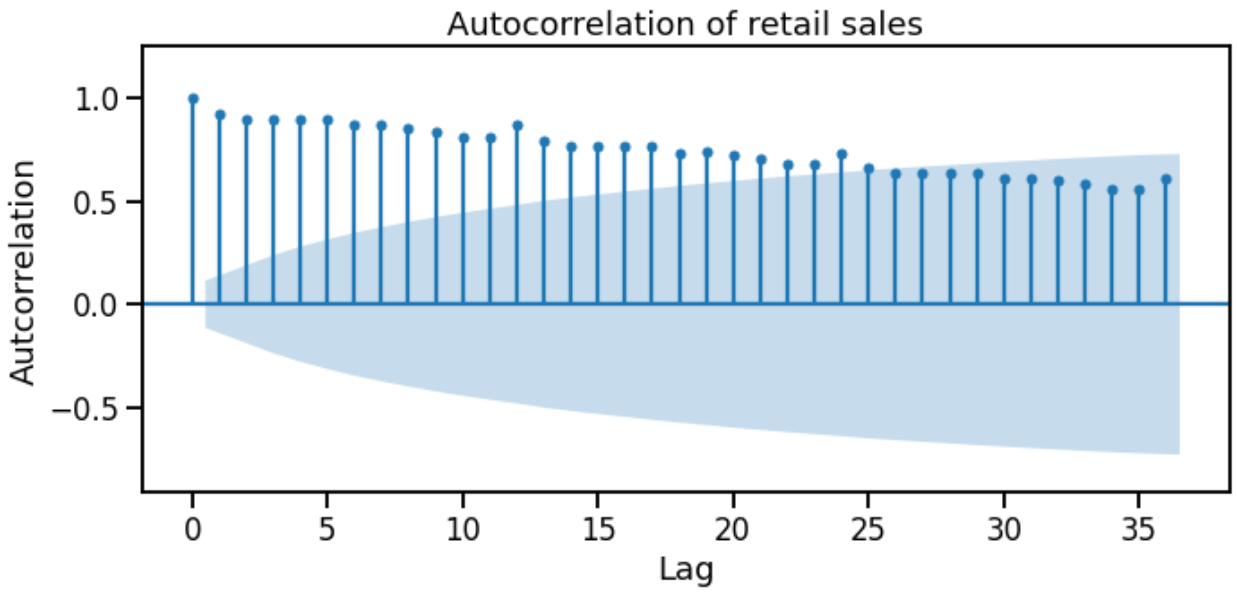




- The PACF has significant lags at k=1 and k=2 as we wanted!

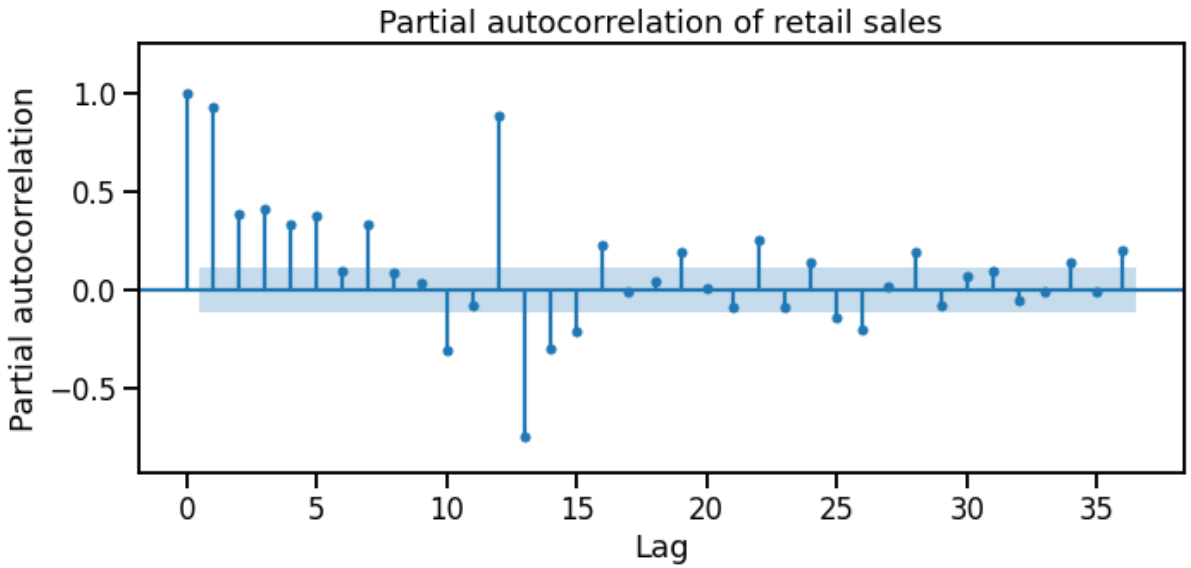
Time series with trend and seasonality

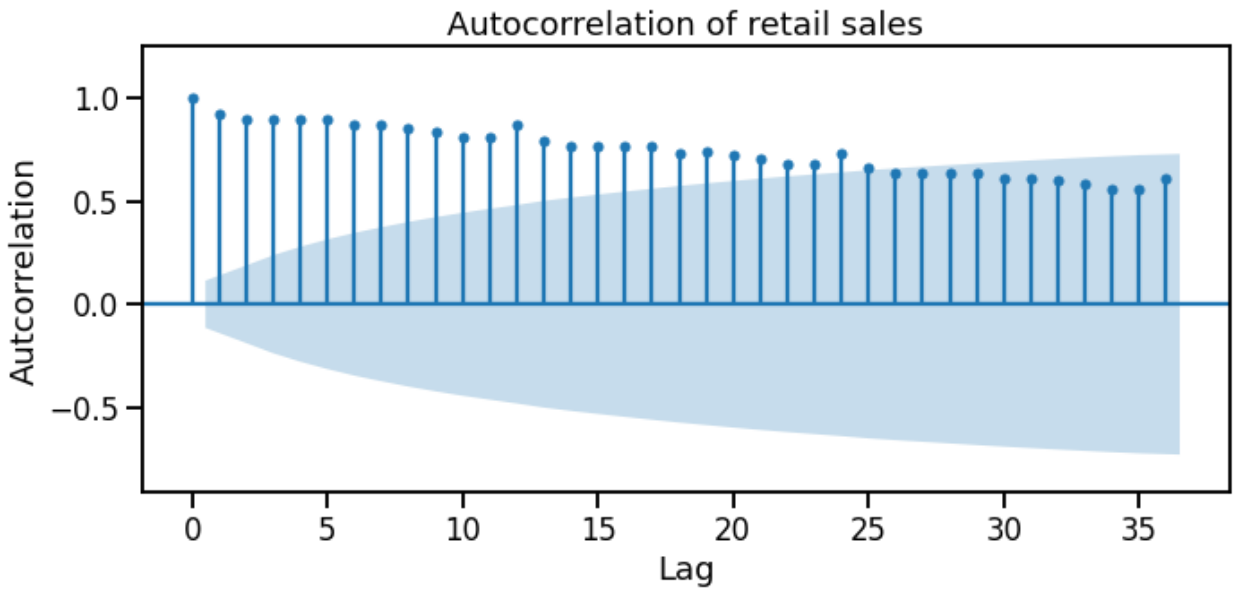




Note: This time series is clearly not stationary. The trend means that the mean changes in time.

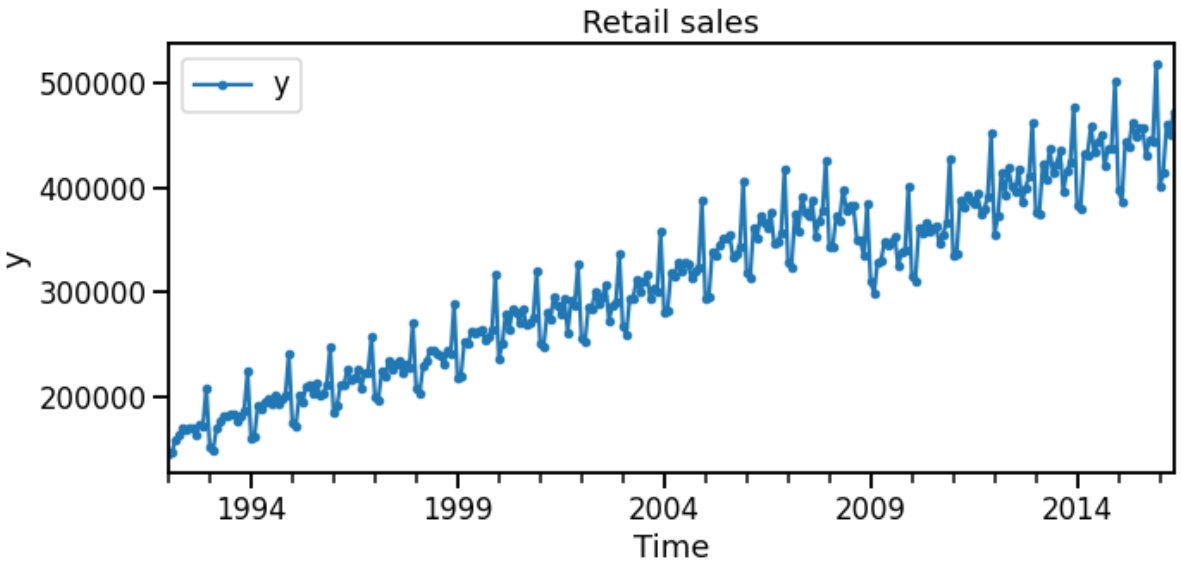
Time series with trend and seasonality





- Despite the lack of stationarity we can still see a strong peak at k=12 and smaller lags (i.e., ~ k=1 to 7). There are also other lags which are more difficult to interpret (k=10, 13).
- We know that the PACF assumes the time series is stationary. So let's try removing trend and seasonality.

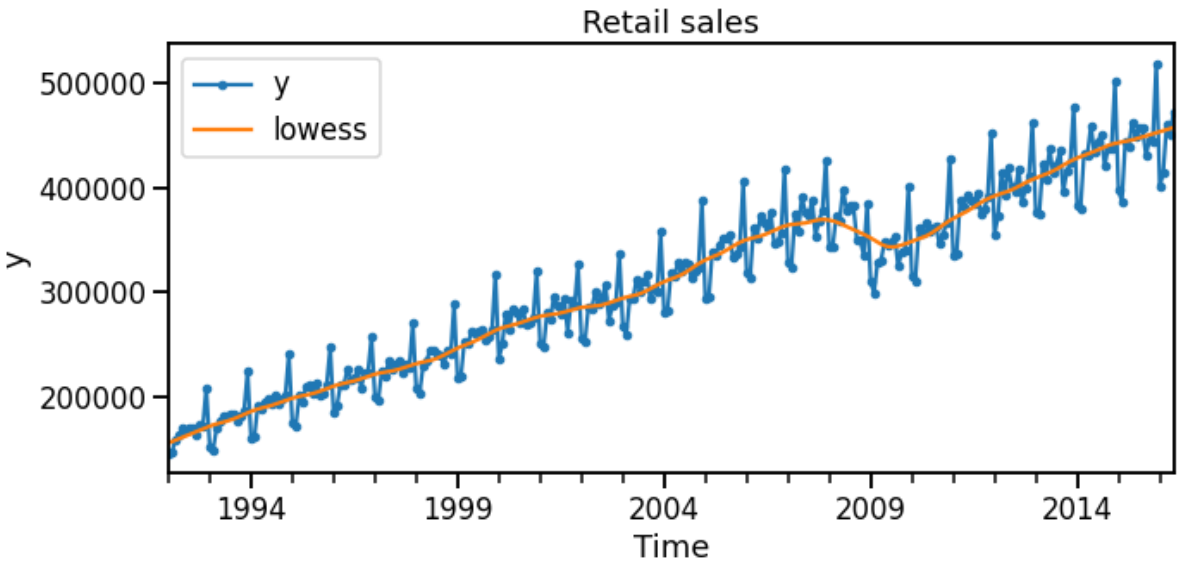
De-trended with seasonality

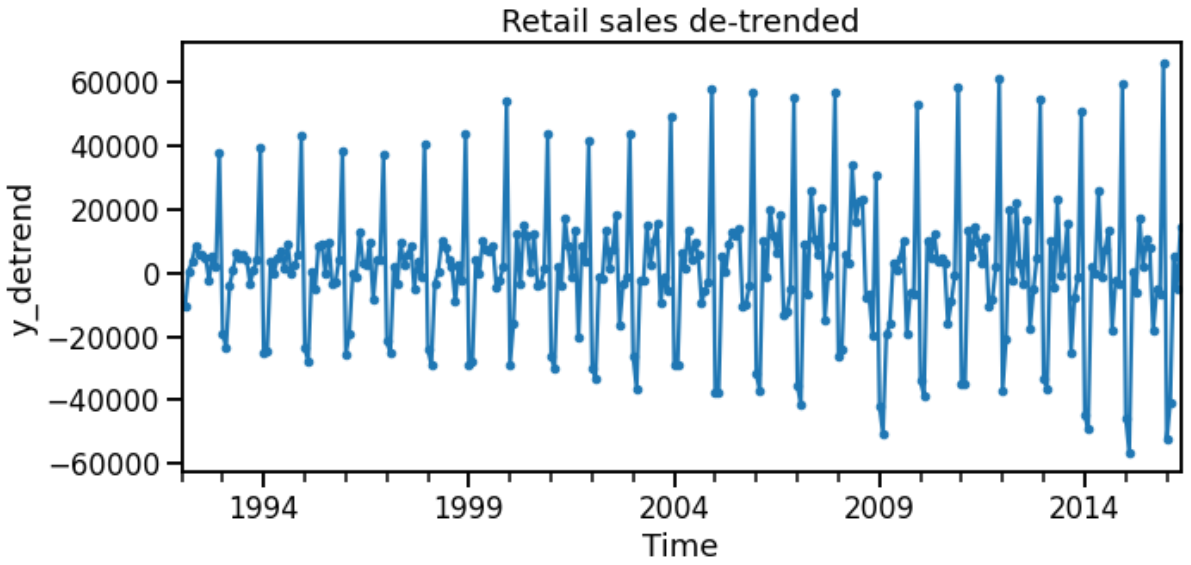


This time series is clearly not stationary. The trend implies that the mean changes in time.

Let's try de-trending the data

De-trended with seasonality

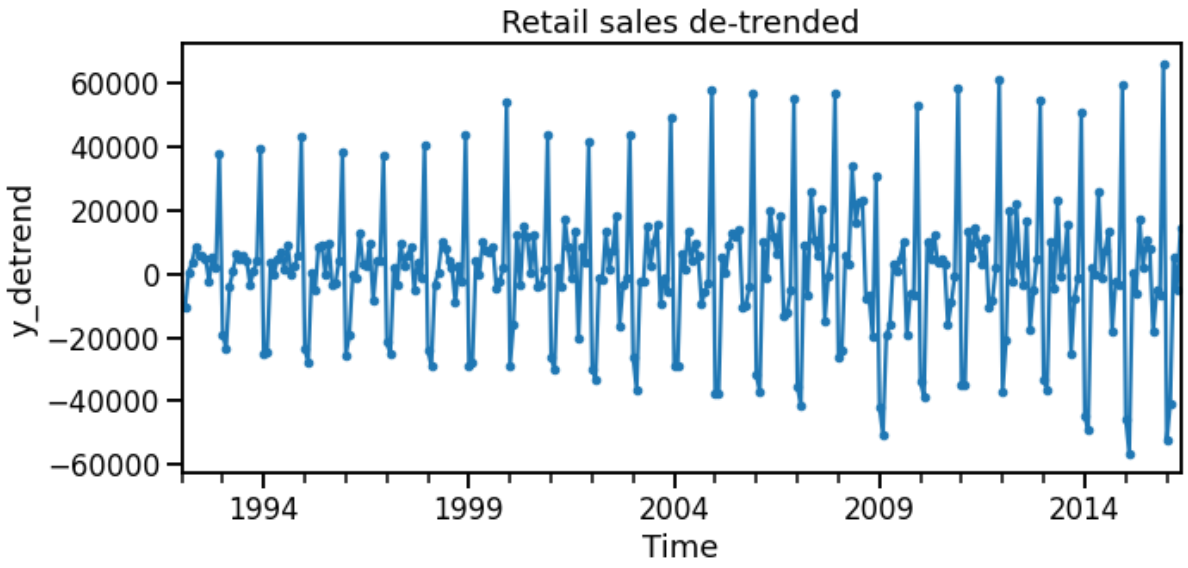


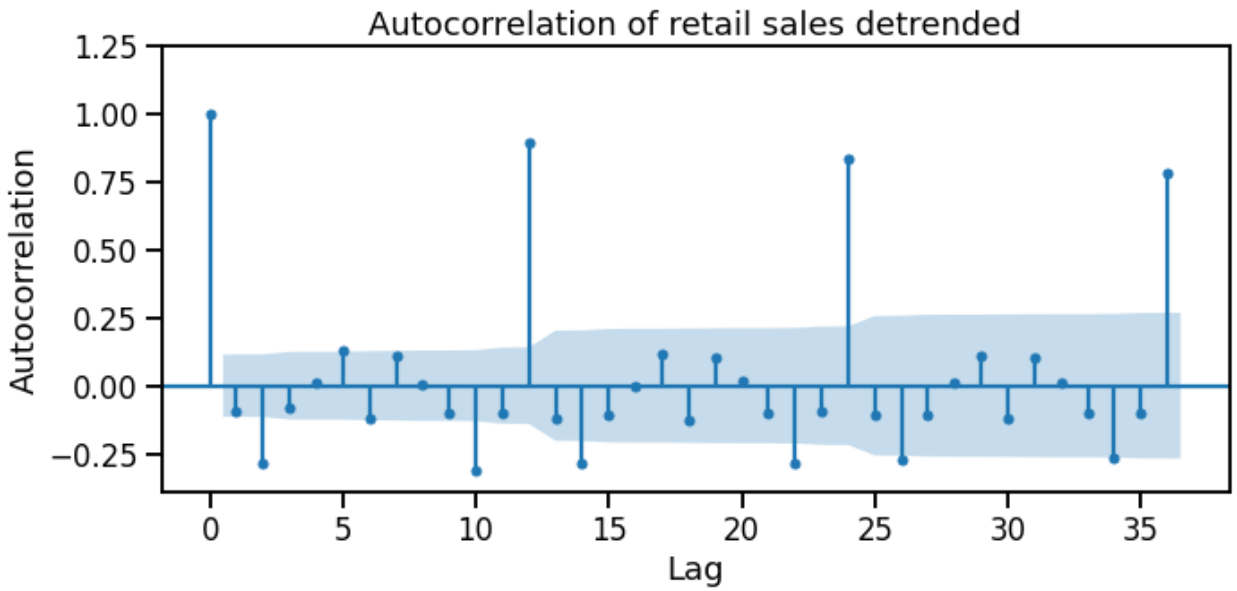


This time series is clearly not stationary. The trend implies that the mean changes in time.

Let's try de-trending the data.

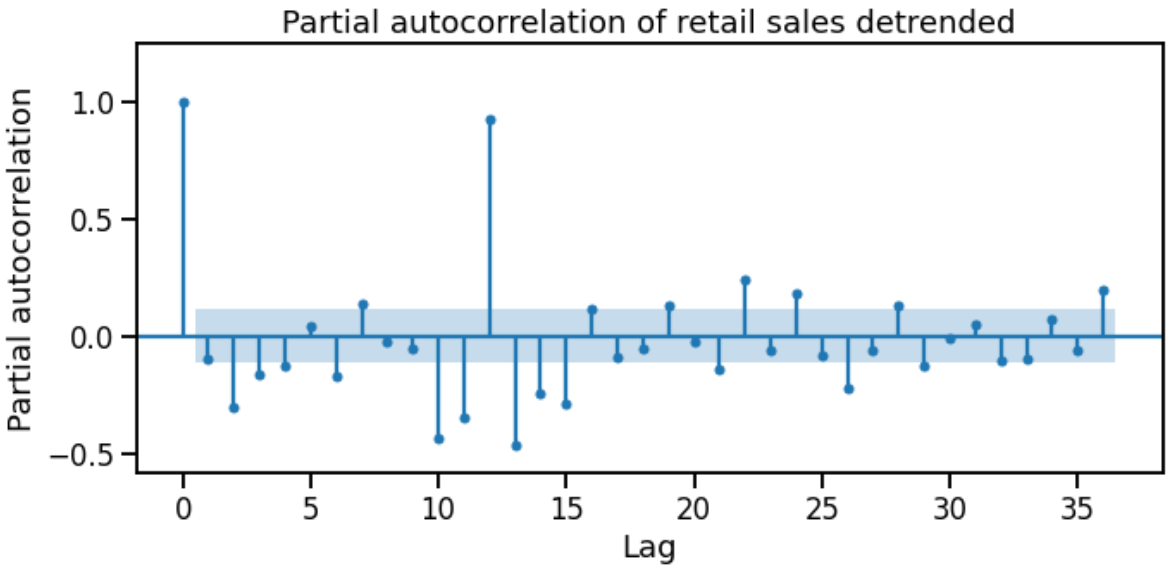
De-trended with seasonality

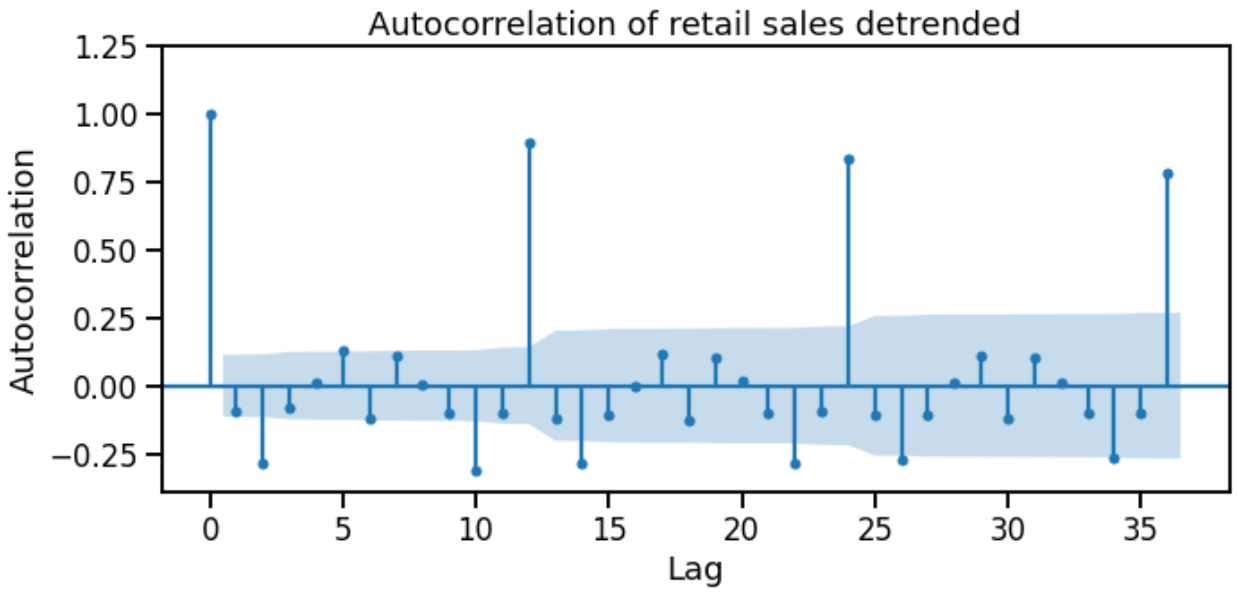




- As we have seen previously, the ACF shows a clear signal of the seasonality.
- Note: the seasonality means that the variance still changes with time. So this is not considered stationary either. We address this later by also de-seasonalsing.

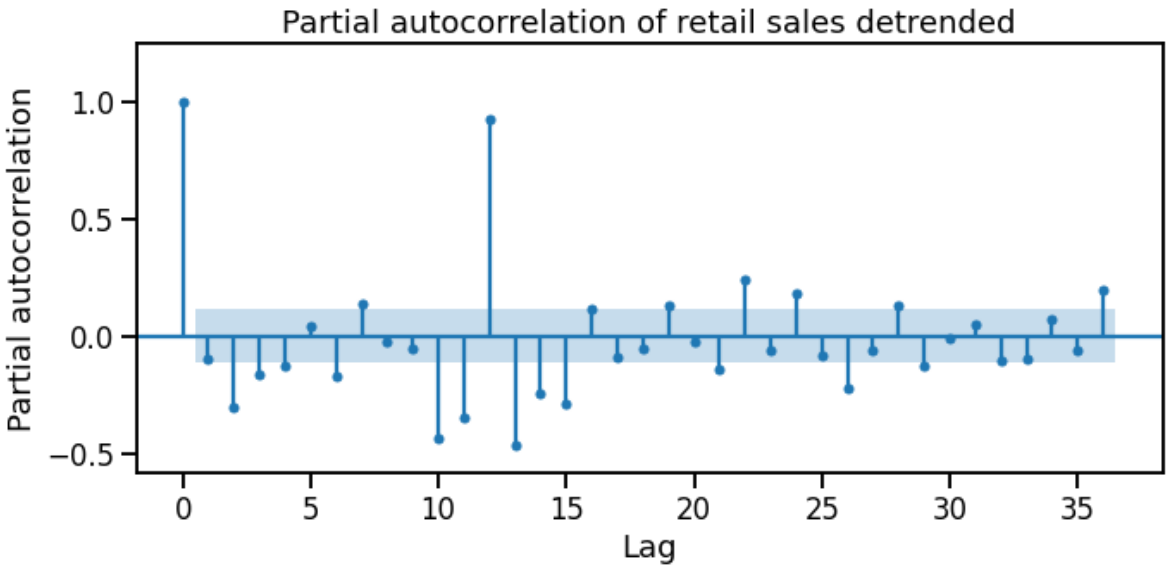
De-trended with seasonality

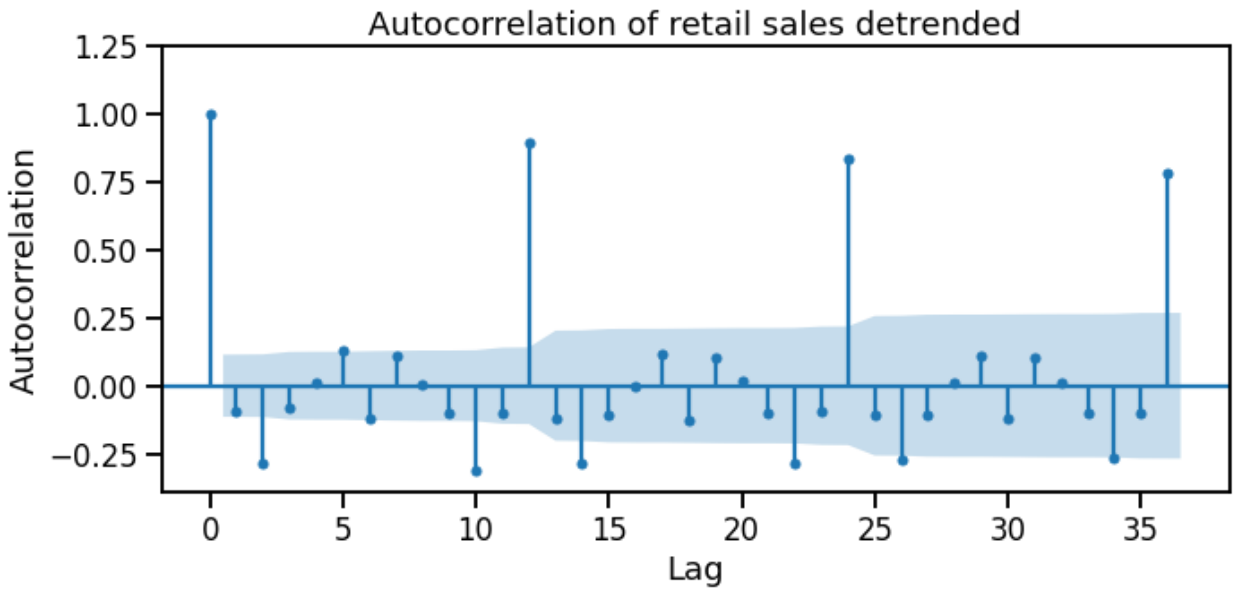




The PACF shows a strong lag at 12 but not at multiples of 12 afterwards. This means that most of the predictive information from the seasonal lag is captured at lag 12 as expected!

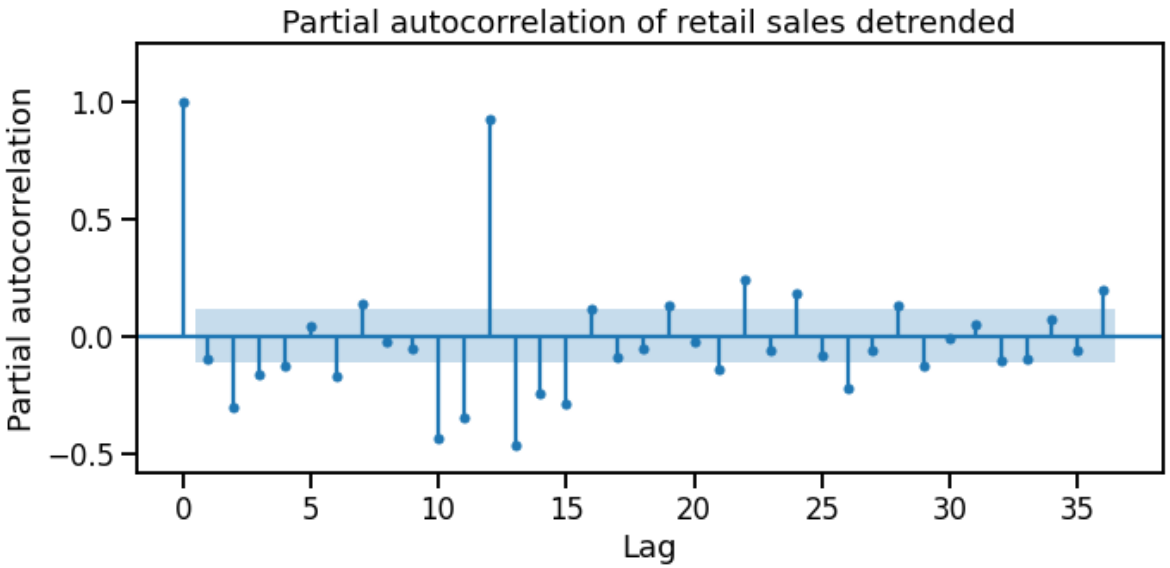
De-trended with seasonality

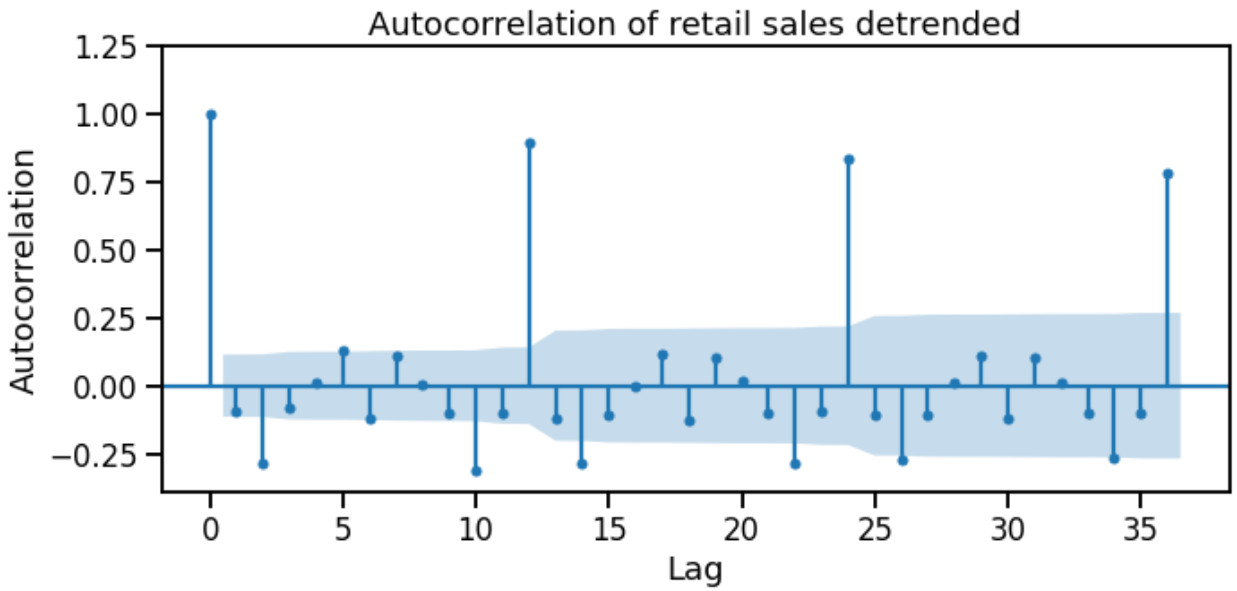




It is harder to interpret whether the other significant lags should be included. In practice this is enough evidence to try some of the smaller lags (e.g., k=1,2,3) and measure the impact in modelling.

De-trended with seasonality

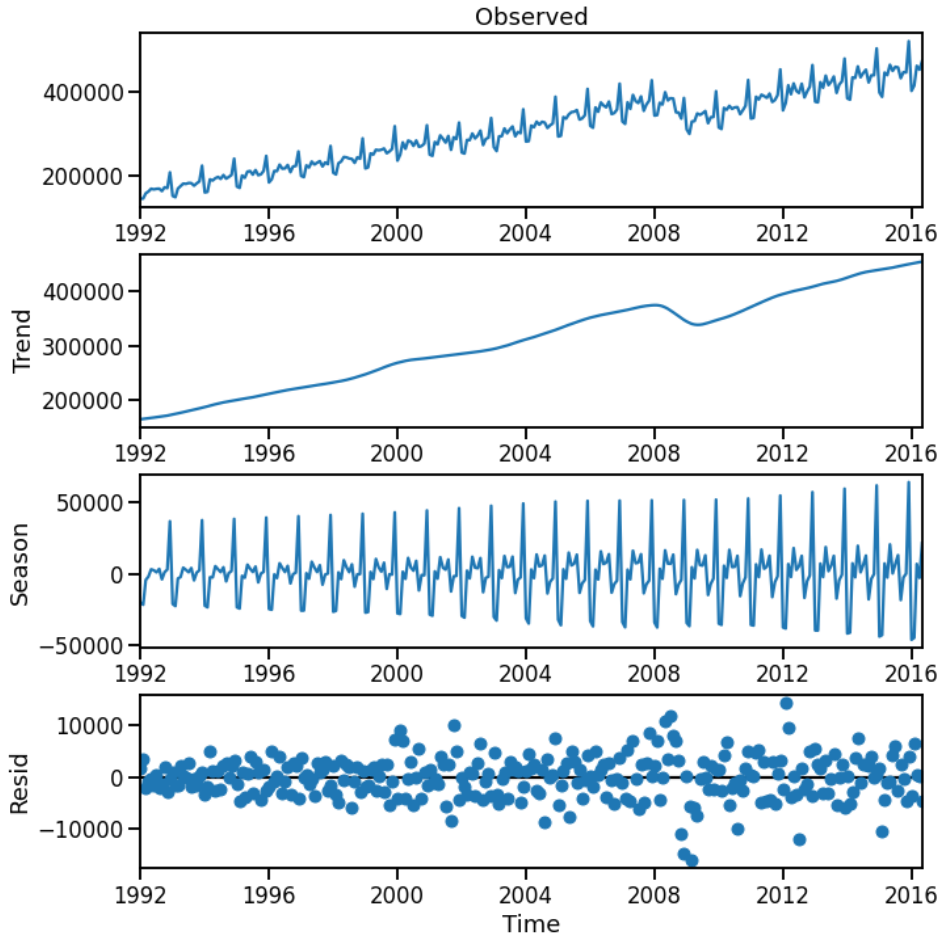




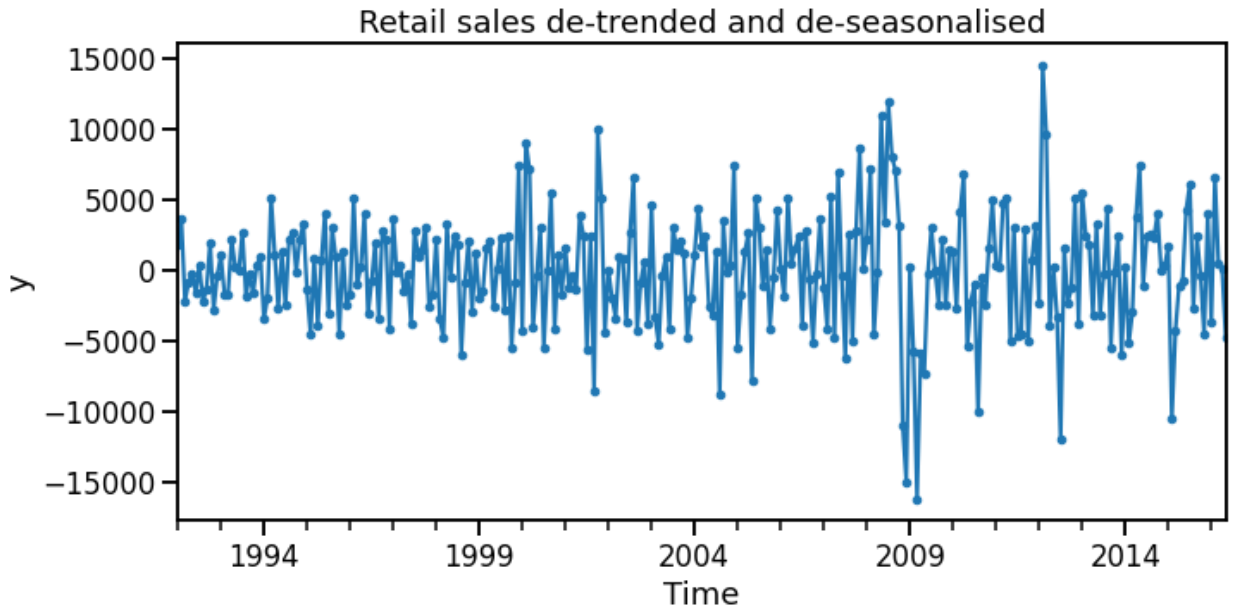
The other larger lags such as lag 10, 11, or 13 are more difficult to interpret. One reason for not including them is that you do not see peaks at multiples of 10, 11, or 13 in the ACF. Also from domain knowledge (this is retail sales) it is highly unlikely that lags much beyond 12 months will be relevant.

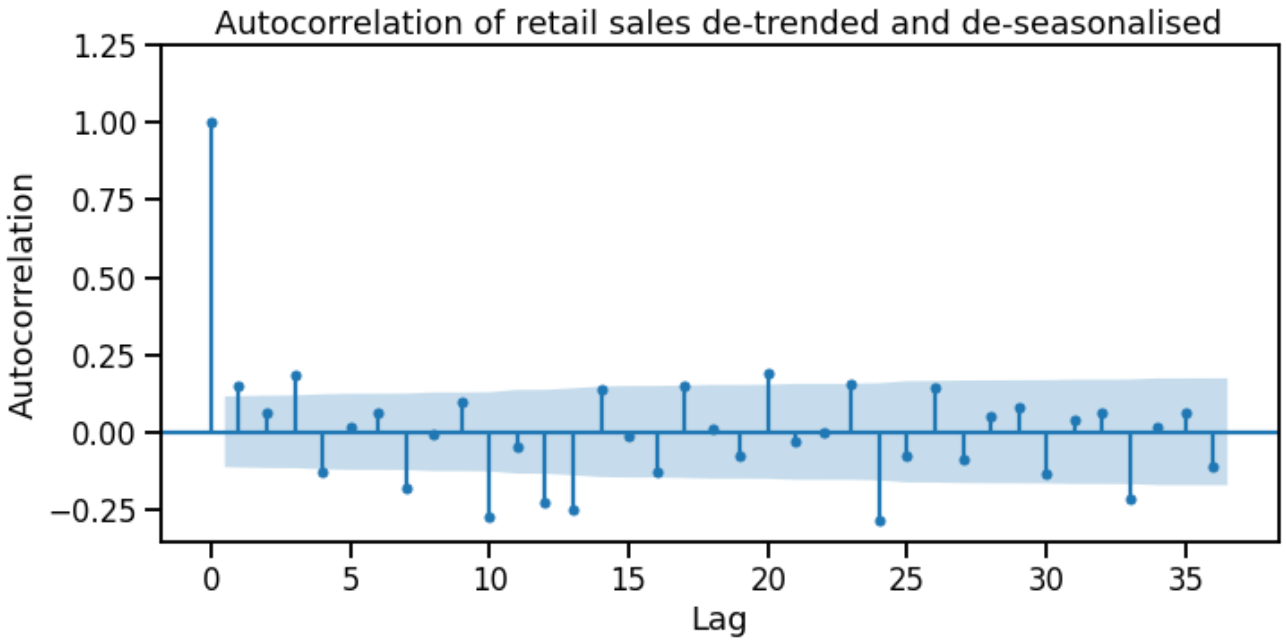
De-trending & de-seasonalising the data

- 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.

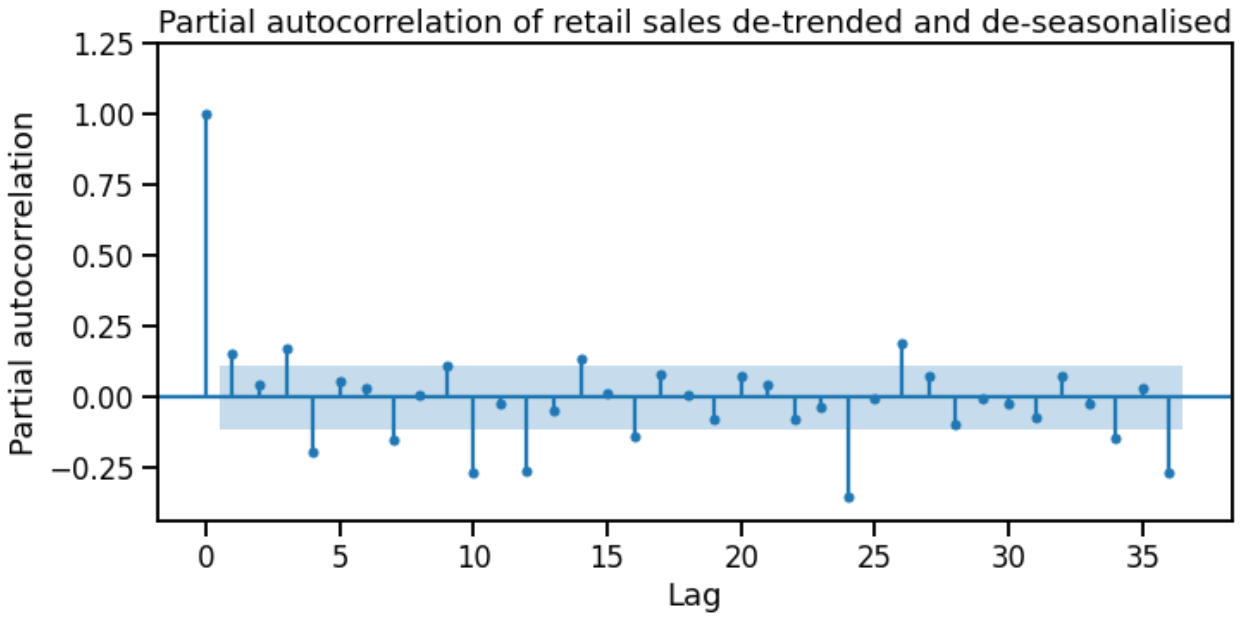


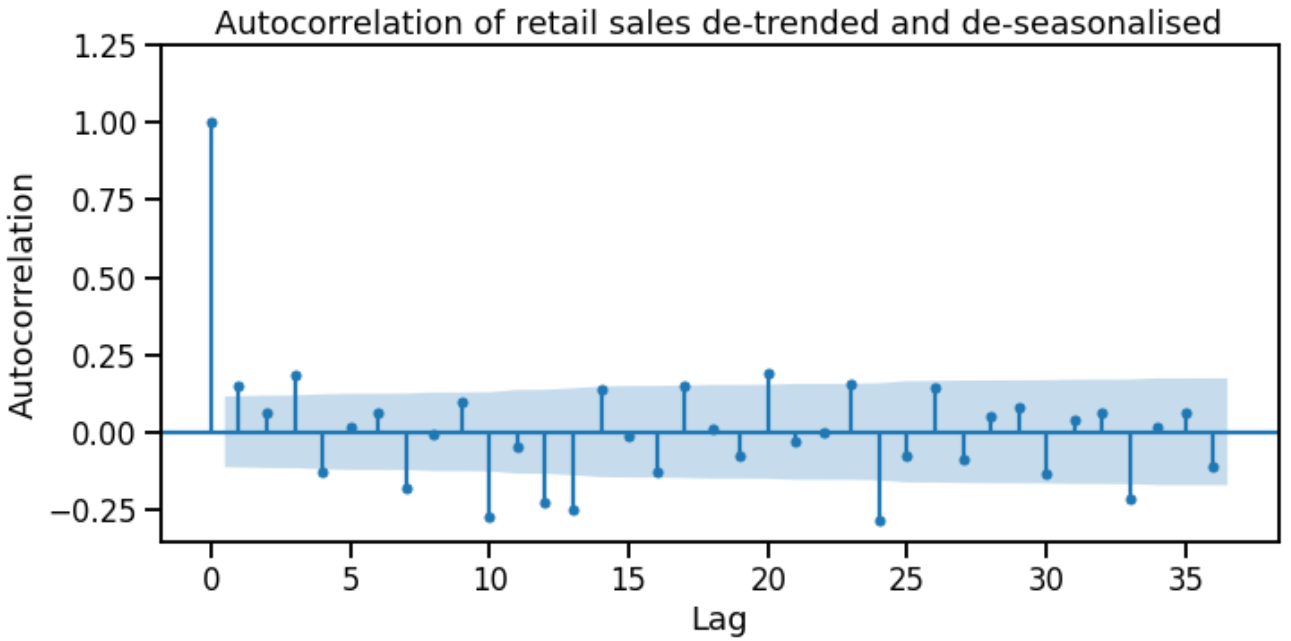
De-trending & de-seasonalising the data





De-trending & de-seasonalising the data





- There aren't any very large significant lags other than potentially k=4, 10, 12, and 24.
- There are significant lags at multiples of 12 which suggests seasonal component is still in the residuals and was not perfectly extracted by STL. May need to tune STL `seasonal` parameter.
- Practically speaking from looking at this plot there wouldn't be an additional lag beyond 1 or 2 that we would want to add for feature engineering purposes.

PACF implementation in Statsmodels

statsmodels.tsa.stattools.pacf

statsmodels.tsa.stattools.pacf(*x*, *nlags=None*, *method='ywadjusted'*, *alpha=None*)[source]

Partial autocorrelation estimate.

Parameters

x : array_like

Observations of time series for which pacf is calculated.

nlags : `int`, **`optional`**

Number of lags to return autocorrelation for. If not provided, uses min(10 * np.log10(nobs), nobs // 2 - 1). The returned value includes lag 0 (ie., 1) so size of the pacf vector is (nlags + 1,).

method : `str`, **`default`** "ywunbiased"

Specifies which method for the calculations to use.

- "yw" or "ywadjusted" : Yule-Walker with sample-size adjustment in denominator for acovf. Default.

```
pacf(df['y'], nlags=12, method='ywmle')
```

```
array([ 1.        ,  0.9199895 ,  0.33026945,  0.30965707,  0.20432541,
        0.1867679 , -0.11743539,  0.11980535, -0.10428794, -0.05842467,
       -0.18727931,  0.10993277,  0.53566757])
```

"ymle" (Yuke-Walker maximum likelihood estimate) is the recommended fitting method Statsmodels notes based on experiments. For more details about the fitting method see [1].

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

PACF implementation in Statsmodels

statsmodels.graphics.tsaplots.plot_pacf

statsmodels.graphics.tsaplots.plot_pacf(*x*, *ax=None*, *lags=None*, *alpha=0.05*, *method='ywm'*, *use_vlines=True*, *title='Partial Autocorrelation'*, *zero=True*, *vlines_kwargs=None*, ***kwargs*)[source]

Plot the partial autocorrelation function

Parameters

x : array_like

Array of time-series values

ax : `AxesSubplot`, **optional**

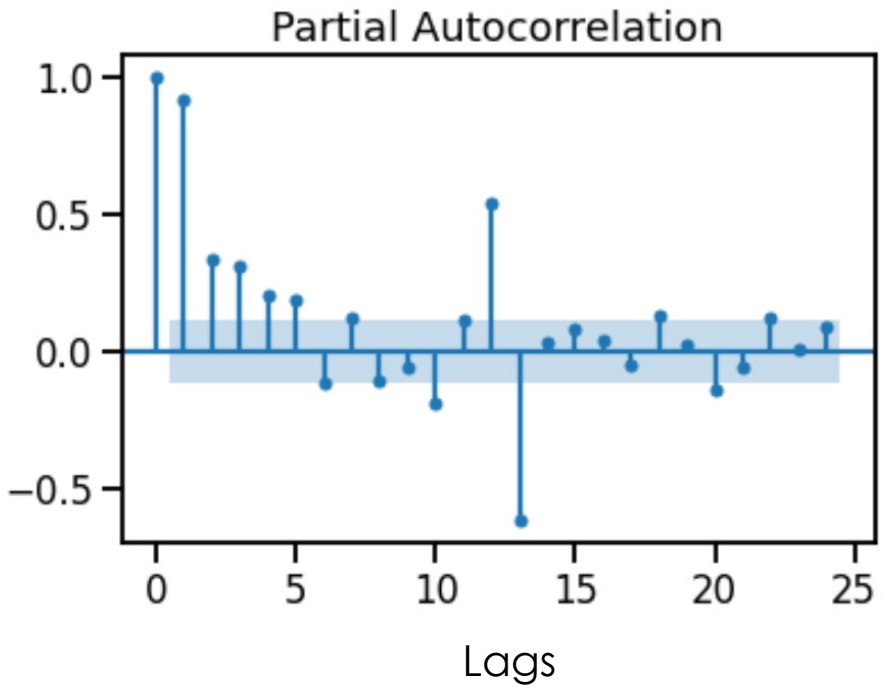
If given, this subplot is used to plot in instead of a new figure being created.

lags : {`int`, array_like}, **optional**

An int or array of lag values, used on horizontal axis. Uses np.arange(lags) when lags is an int. If not provided, `lags=np.arange(len(corr))` is used.

alpha : `float`, **optional**

```
plot_pacf(df['y'], lags=24, method='ywmle');
```



Summary

Partial autocorrelation function (PACF) measures how correlated a y_t is with itself at lags y_{t-k} after removing the effect of intermediate lags.

The PACF for an AR(P) process has non-zero lags up to P and zero afterwards.

Methods to estimate the PACF assume the time series is stationary.

The PACF can help identify lags which may be helpful for forecasting.