Practical Data Science

LVC 3: Time Series

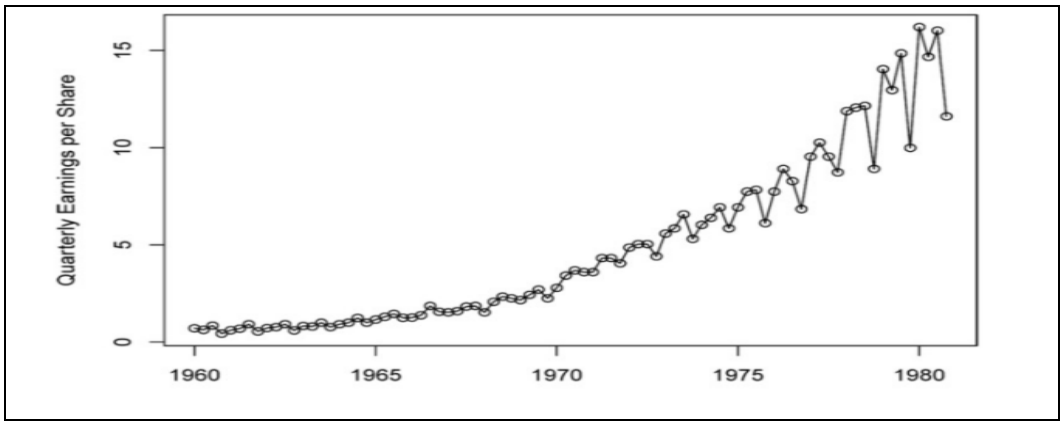
A time series is a collection of observations of well-defined data items obtained through repeated measurements over time. For example, measuring the value of retail sales each month of the year would comprise a time series. This is because sales revenue is well defined, and consistently measured at **equally spaced intervals**. Data collected irregularly or only once are not time series.

Time series data is everywhere since time is a constituent of everything observable. As our world gets increasingly instrumented, sensors and systems are constantly emitting a relentless stream of time series data. Such data has numerous applications across various industries.

Let's put this in context through some examples. Time series data can be useful for a variety of scenarios:

- Tracking daily, hourly, or weekly weather data
- Tracking changes in application performance
- Medical devices to visualize vitals in real-time

Let's look at some of the examples in more details,



The above graph shows the quarterly (that is, four times yearly) observations of the earnings of Johnson & Johnson corporation from 1960 to 1980 recorded at equally spaced time intervals. There are 84 observations, one for each quarter over 21 years. So these values are not random; each of the values is dependent on previous values. Generally, the study of time-series data involves two fundamental questions: what happened (description), and what will happen next (forecasting)?

For the Johnson & Johnson data, you might ask:

- Is the price of Johnson & Johnson shares changing over time?
- Are there quarterly effects, with share prices rising and falling regularly throughout the year?
- Can you then forecast what the future share prices will be?

Why is the time series different?

- Because data points in time series are collected at adjacent time periods, they are dependent on previous observations. In the language of probability, these are dependent random variables. This is one of the features that distinguishes time-series data from other kinds of data.
- The statistical characteristics of time series data often violate the assumptions of conventional statistical methods. For example, the logistic regression or random forest, or any other algorithm assumes the samples are independent of each other. Because of this, analyzing time-series data requires a unique set of tools and methods.
- Two major reasons for variations in the time series data are:
    - Trend
    - Seasonality

Trend

The trend shows a general direction of the time series data over a long period of time. A trend can be increasing (upward), decreasing (downward), or horizontal (stationary). In the Johnson & Johnson quarterly earnings, we can see an upward trend.

Seasonality

The seasonal component tells if there is a regularly repeating pattern of highs and lows related to calendar time such as seasons, quarters, months, days of the week, and so on. Some examples include an increase in water consumption in summer due to hot weather conditions, or an increase in the number of airline passengers during holidays each year.

What is a Stochastic Process?

A stochastic process is a set or collection of random variables $\{X_t\}$ (not necessarily independent), where the index 't' takes values in a certain set, this set is ordered and corresponds to the moment of time. Time series is the realization of the stochastic process. Realization is a unique function of time different from the others. The process is characterized by the joint probability distribution of the random variables X_1, X_2,...., X_T, for any value of T.

Obtaining the probability distributions of the process is possible in some situations, for example with climatic variables, where we can assume that each year a realization of the same process is observed, or techniques that can be generated in a laboratory. Nevertheless, in many situations of interest, such as with economic or social variables, we can only observe one realization of the process.

For example, if we observe the series of yearly growth in the wealth of a country it is not possible to go back in time to generate another realization. The stochastic process exists conceptually, but it is not possible to obtain successive samples or independent realizations of it.

Stationarity

To tackle the above problem we introduce the concept of the stationary stochastic process. There are two types of stationary processes:

1. Strong sense stationarity
2. Wide sense stationarity

Strong sense stationarity

Strong stationarity requires the shift-invariance (in time) of the stochastic process. This means the joint probability distribution is the same if you shift the data. For example, the distribution of X_t, $X_{(t+1)}$, $X_{(t+2)}$... is the same as $X_{(t+h)}$, $X_{(t+h+1)}$, $X_{(t+h+2)}$... for any h.

It's a very strong condition, we must have the joint distributions for any set of variables in the process to prove it. Generally, it is hard to verify, so we are going with a weaker notion, which is wide sense. Before entering into a wide sense of stationary, we need to understand the mean and autocovariance.

Mean of the process:

It's the expected value of a stochastic process X_t, which is given as,

$$\mu_t = E(X_t)$$

Autocovariance:

Autocovariance is defined as the covariance between the random variable X_{t1} and X_{t2}. Autocovariance (auto means itself) of X_{t1} and X_{t2} is defined as the covariance between the same time series at different time periods. It's denoted as,

$$R_X(t_1, t_2) = E((X_{t_1} - \mu_{t_1})(X_{t_2} - \mu_{t_2}))$$

Wide sense stationarity

The conditions for wide sense are

1. $E(X_t) = \mu_t$
2. $R_X(t_1, t_2) = R_X(t_1 - t_2) = R_X(t_2 - t_1)$

For a wide sense, the process should have the same mean at all time points, and the covariance between the values at any two-time points, t, and $t - k$, depend only on k, the difference between the two times, and not on the location of the points along the time axis. For example, the covariance of X_1 and X_5 should be the same as X_3 and X_7.

Testing Stationarity

Before modeling, we need to check if a given series is wide sense stationary or not. To find whether the given series is stationary or not we need to compute

- The sample mean for each λ

$$\hat{\mu} = \frac{1}{N-\lambda} \sum_{i=\lambda}^{N-1} X_i$$

Where N is the total number of samples and λ is any point you have selected.

Let's understand with an example, we have some value in the table

X0	11
X1	13
X2	14
X3	12
X4	14

First, will do for λ = 0, then mean is the sum of all samples by the number of samples

Mean = (11+13+14+12+14)/5

= 12.8

Then increase the value of λ = 1, then mean is

$$= \frac{1}{5-1} \sum_{1}^{4} X_i$$

= (13+14+12+14)/4 =13.25

Then increase the value of λ = 2, then mean is

$$= \frac{1}{5-2} \sum_{2}^{4} X_i$$

= (14+12+14)/3 = 13.33

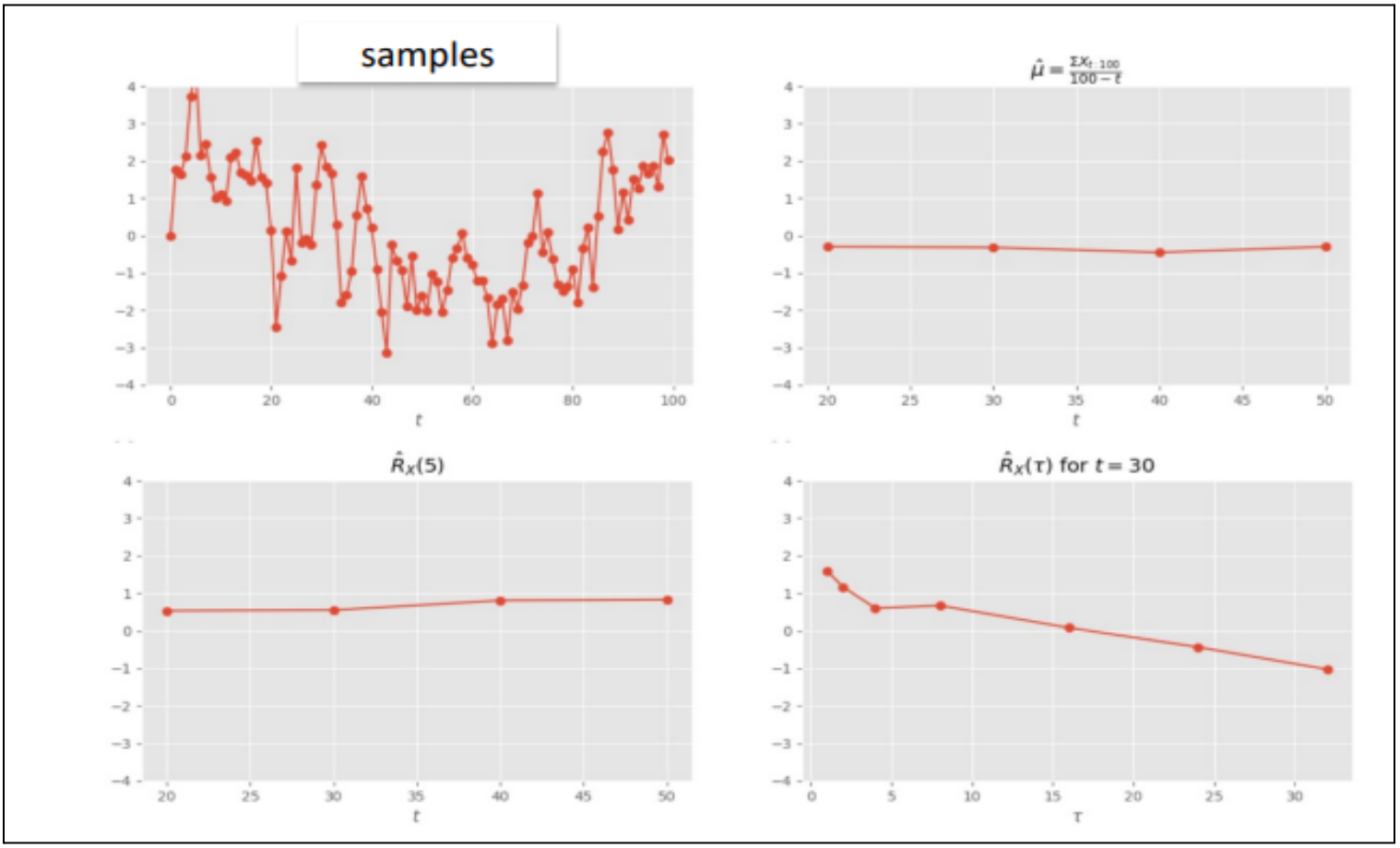
This should be repeated until λ = $N - 1$ and for each of the λ, we need to get the constant mean which is not the case for the above example.

- Compute sample autocovariance for each λ which is

$$\widehat{R}_X(\tau) = \frac{1}{N-\lambda} \sum_{i=\lambda}^{N-1} (X_i - \widehat{\mu})(X_{i+\tau} - \widehat{\mu})$$

The process is the same, go with some value for λ and calculate the covariance between two variables. In the end, you should not see the larger variation over λ.

Let's understand with an example,



Here, we have generated the data of 100 samples as we can see from the first figure. The first step for checking stationarity is to check the empirical mean, as we can see in figure 2, the mean is constant which has been calculated at different values of λ. So, one of the conditions is satisfied.

Next thing is to check the autocovariance for every difference which is $t_1 - t_2 = \tau$ (symbol is called Tau). The tau value will be from 1 to the total number of data points.

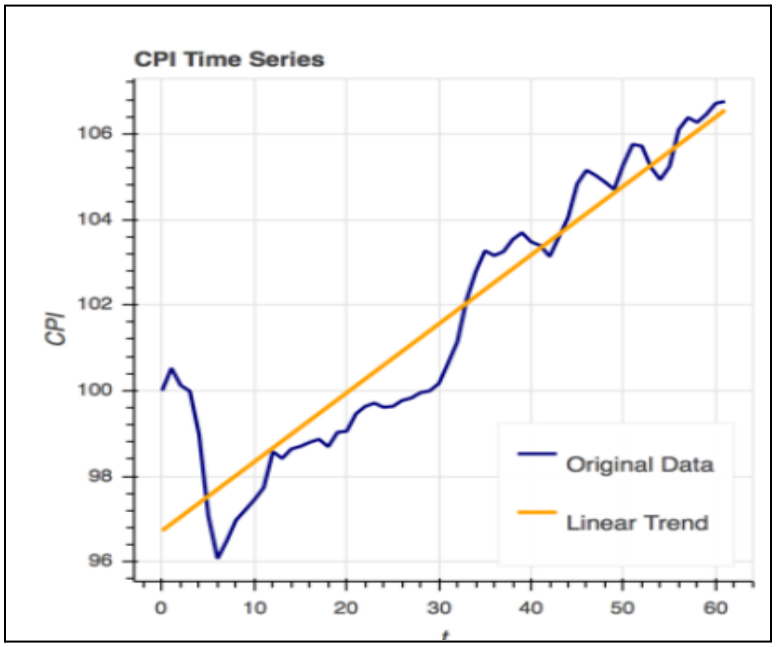
In the third figure, the autocovariance has been calculated for Tau = 5 and different values of λ. It's also constant.

We need to check for every value of Tau and for any of the values of Tau if it violates the condition, then it's not stationary. Figure 4 is not the same for different values of λ so it's not a stationary time series.

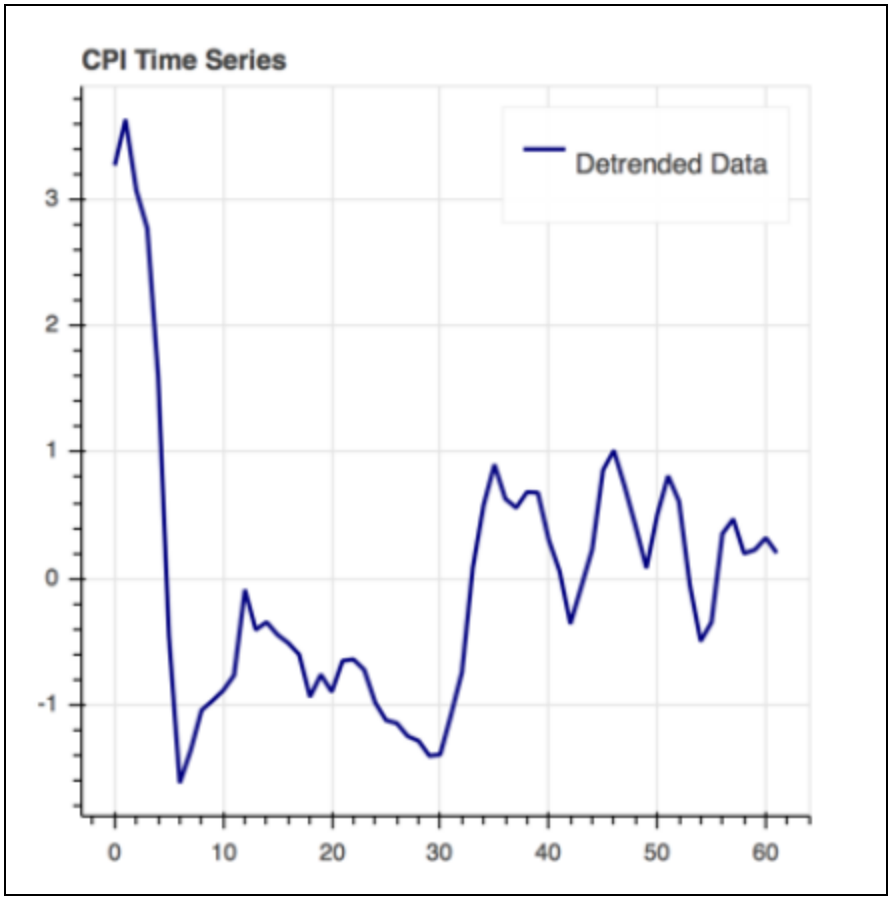
Detrend

The detrending of time series is a process of removing the trend from a non-stationary time series. If we have time series with only trend components, in that case the **detrended** time series is known as a stationary time series, while a time series with a trend is non-stationary time series. A stationary time series oscillates about the horizontal line. If a series does not have a trend or we remove the trend successfully, the series is said to be trend stationary. Elimination of the trend component may be thought of as rotating the trend line to a horizontal position.

For example, the below graph shows the consumer price index. We can see the trend which is increasing as shown by the orange color.



Once we remove the trend, the graph looks like the one below. So, there is no trend in the data.



Similarly, we also have deseasonalizing which is removing the seasonality from the data to make it stationarity.

Models of time series

White Noise

Consider we have a time series W_t. If the elements of the time series have,

1. Mean of zero, i.e., $E(W_t) = 0$

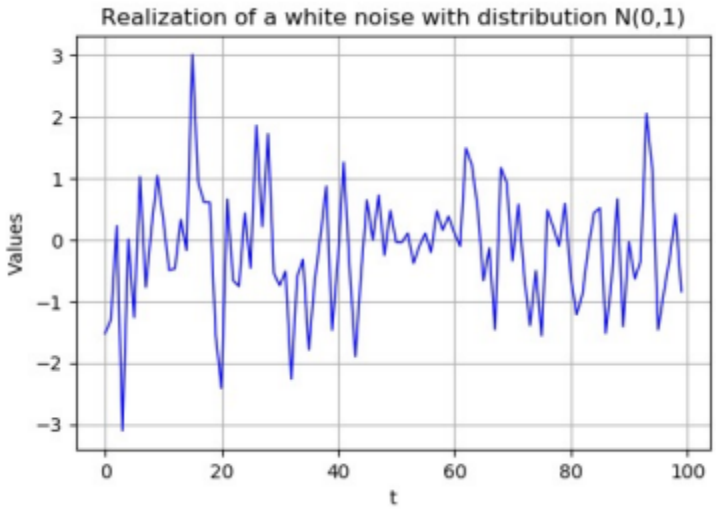
2. Autocovariance between two random variables (if both are not the same) is zero. Considering we have two values $W_{t1,}W_{t2}$, then the autocovariance should be zero. If W_{t1}, W_{t2} are the same, then the autocovariance is equal to the variance.

$$\text{Autocovariance} = E(W_{t1,}W_{t2}) = \sigma^2\delta(t_1 - t_2)$$

$$\delta(t_1 - t_2) = \left(1 \;\; if \;\; t_1 = t_2 \; otherwise \; 0\right)$$

An example of white noise is tossing a coin at a time t. It is white noise because it doesn't correlate with previous trials.

The graph below is white noise with a mean of 0.



White noise is a series that's not predictable, as it's a sequence of random numbers. If you build a model and its residuals (the difference between predicted and actual) values look like white noise, then you know you did everything to make the model as good as possible. On the opposite side, there's a better model for your dataset if there are visible patterns in the residuals.

Random Walk

A random walk is another time series model where the current observation is equal to the previous observation with a random step up or down. It is formally defined as

$$X_t = X_{t-1} + W_t$$

- X_t is the current value
- X_{t-1} is the previous value
- W_t is white noise

Just like white noise, random walk series also isn't predictable on the basis of the past values. In the stock market, it means that short-run changes in stock prices are unpredictable.

Autoregressive Models

An autoregressive model is when a value from a time series is regressed on previous values from that same time series.

$$X_t = \sum_{i=1}^{p} a_i X_{t-i} + W_t$$

- X_t is the current value
- X_{t-i} is the previous values
- W_t is the white noise

In this autoregressive model, the response variable in the previous time period has become the predictor and the errors have our usual assumptions about errors in a simple linear regression model. The order (p) of an autoregression is the number of previous values in the series that are used to predict the value at the present time.

So, if the model is a first-order autoregression, written as AR(1), the equation will be written as,

$$X_t = a_1 X_{t-1} + W_t$$

AR(p) Model

The AR(p) process is given by,

$$X_t = \sum_{i=1}^{p} a_i X_{t-i} + W_t$$

For determining whether the given process is stationary, we want roots of the polynomial $(1 - a_1 z_1 - a_2 z^2 - \ldots\ldots a_p z^p)$ to be inside the unit disc or should be less than 1. If it is >1, then it's not a well-defined AR process.

For example, if we have a random walk $X_t = X_{t-1} + W_t$, then the roots of the polynomial equation will be

$$P(z) = (1 - z)$$

So, if $P(z) = 0$, then z = 1. Here, it violates our condition.

If we write the random walk equation differently,

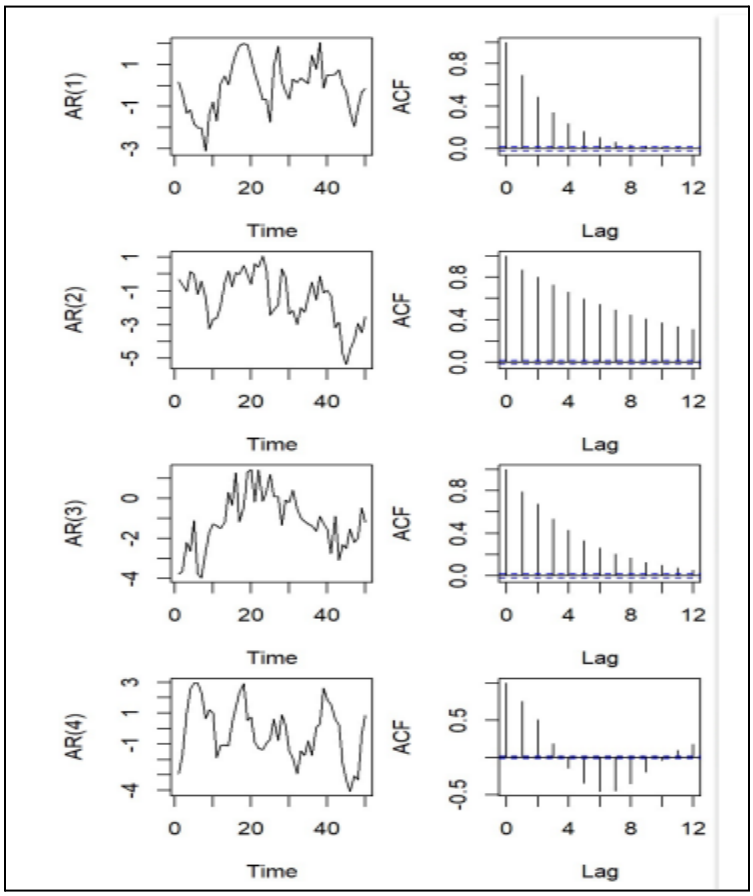
$$X_t - X_{t-1} = W_t$$

Then, it's stationary because it's white noise. We know white noise is a stationary process. The term $X_t - X_{t-1}$ is called first differencing which is used to convert a non-stationary time series to stationary time series.

The **Autocovariance** of the AR(p) process is given by,

$$R_X(\tau) = \sum_{i=1}^{p} a_i R_X(\tau - i) + \sigma^2 \delta(\tau)$$

So, one of the properties of AR models is that the autocovariance decays exponentially. If you calculate the autocovariance using the above formula, then you can find that it decays exponentially.



In the above image, for the AR(1) model, the autocovariance function has been decaying exponentially, and similarly for AR(2) and AR(3), we can see the autocovariance function has been decaying exponentially. For AR(4), the autocovariance function is decaying exponentially, but with a lot of fluctuations.

Ultimately, it's going to decay. It's a property of the AR model that ACF decays exponentially.

Moving Average Models (MA(q))

A Moving Average model is when a value from a time series is regressed on past errors from that same time series. A moving average term in a time series model is a past error (multiplied by a coefficient). So, the equation is given as,

$$X_t = \sum_{i=0}^{q} b_i w_{t-i}$$

- X_t is the current value
- W_t is the white noise
- b_i are coefficients
- q is the order, i.e., how many past errors are to be considered

The **autocovariance** for the MA model is given as,

$$R_X(\tau) = \sigma^2 \sum_{j=0}^{q} b_i b_{j-\tau}$$

And autocovariance for MA is 0 for Tau greater than q. For example, if we have MA(1) model, then the only nonzero value in the ACF is for lag 1. All other autocorrelations are 0. Thus a sample ACF with a significant autocorrelation only at lag 1 is an indicator of a possible MA(1) model.

Autoregressive Moving Average (ARMA (p,q))

ARMA is a model of forecasting in which the methods of autoregression (AR) and moving average (MA) both are applied to time-series data. It accounts for past values as well as past errors. We can write it as,

$$X_t = \sum_{i=1}^{p} a_i X_{t-i} + \sum_{j=0}^{q} b_j w_{t-j}$$

So, the first part of the equation corresponds to the AR process and the second part of the equation corresponds to the MA process.

Integrated Process

For AR, MA, and ARMA modeling, we require the given times series to be stationary. In most cases, we don't have a stationary time series. So, we need to convert the non-stationary time series to stationary time series.

To make time-series stationary one such approach is differencing. Differencing can help stabilize the mean of the time series by removing changes in the level of a time series, and so eliminating (or reducing) trend and seasonality. Differencing is performed by subtracting the previous observation from the current observation. The number of times that differencing is performed is called the difference order.

This process is called an integrated process.

For example,

1. First-order: $X_t - X_{t-1}$
2. Second-order: $X_t - 2X_{t-1} + X_{t-2}$

ARIMA

Arima is the combination of AR, MA, and Integrated. So, let's understand each component of ARIMA in detail,

- The "AR" in ARIMA stands for autoregression indicating that the outcome of the model depends on the past values.
- The "I" stands for integrated to make stationary.
- The "MA" stands for moving average model, indicating that the outcome of the model depends on past errors.

The ARIMA model takes in three parameters:

1. p is the order of the AR term.
2. q is the order of the MA term.
3. d is the number of times we do differencing to make stationary.

Learning Time Series

AR(p)

So far we have understood what AR is and how it works. Just to recap, we know that every value is regressed on past p values. Now, we are going to understand how to estimate the coefficients.

Let's say we have a time series X_0, X_1,....., X_N, assuming we know the value of p, so the equation of the model is,

$$X_t = \sum_{i=1}^{p} a_i X_{t-i} + w_t$$

In the equation, we know X_t's. The unknowns in the equation are $a_i's$. To estimate them, we are going to define,

$$X = (x_{p+1}, x_{p+2}, x_{p+3}, \ldots\ldots\ldots, x_N)'$$

The vector X takes the values starting from the $p + 1$ to N. Consider having p equal to 3, i.e., the current value depends on past 3 values. For x_1, x_2, and x_3, there are not 3 past values for these records. So, we cannot write the equation for them but for x_4, we have 3 past values x_3, x_2, and x_1. This is the reason we are taking the starting value as $p + 1$. We can assume this as a dependent variable.

Let a be the vector of coefficients that the model wants to learn,

$$a = (a_1, a_2, a_3, \ldots\ldots\ldots, a_p)'$$

Now, we are going to define Matrix A,

$$A = \begin{pmatrix} x_p & x_{p-1} & \cdots & x_1 \\ x_{p+1} & x_p & \cdots & x_2 \\ \vdots & \vdots & \vdots & \vdots \\ x_N & x_{N-1} & \cdots & x_{N-m} \end{pmatrix}$$

In the matrix, each row acts as regressors. If we multiply the first row of the matrix A with a vector, the equation will be,

$$x_p a_1 + x_{p-1} a_2 + x_{p-2} a_3 + \ldots\ldots\ldots\ldots$$

The above equation is used for predicting x_{p+1}.

We can think in the sense of regression, where X represents predictors and "$A.a$" represents regressors. So, we need to find the vector "a" such that it minimizes the error. It's a kind of least squares problem. We are finding the best fit over here similar to regression, with the only difference that the values are independent over there, but in time series, the values are dependent. So, we can write this as,

$$\min_{\mathbf{a}} \|\mathbf{x} - A\mathbf{a}\|$$

Order Estimation

The order estimation is one of the most important tasks in time series. There are many ways to estimate the order, some of them are:

1. We are going to divide the data, build multiple models with different orders and compare their errors.
2. Adding penalty to the model's complexity. Models are scored both on their performance on the training dataset and based on the complexity of the model.
    a. **Model Performance**: How well a candidate model has performed on the training dataset?
    b. **Model Complexity**: How complicated the trained candidate model is after training?

Some of the techniques are AIC (Akaike Information Criterion), and MDL (Minimum Description Length). To use AIC for model selection, we simply choose the model giving the smallest AIC over the set of models considered.

We have a quick way of determining the order using ACF plots. As we know, for MA models, the ACF becomes zero after Tau is greater than q. So, we can use the ACF for determining the order of MA, but for AR, the ACF is decaying exponentially, so we cannot decide the order of AR models from the plot. So for that, we are going to introduce another concept called PACF.

PACF (Partial Autocorrelation Function) :

A partial correlation is a conditional correlation. It is the correlation between two variables under the assumption that we know and take into account the values of some other set of variables. For instance, consider a regression context in which y is the response variable and x_1, x_2, and x_3 are predictor variables. The partial correlation between y and x_3 is the correlation between the variables determined taking into accounts how both y and x_3 are related to x_1 and x_2.

For example, if we want to find the covariance between X_t and X_{t+k},

$$X_t \left| X_{t+1}, X_{t+2}, X_{t+3.....,X_{t+k-1}} \right| X_{t+k}$$

We are going to subtract the projection of X_t from X_{t+1} to X_{t+k-1}, i.e., for future values and similarly, for X_{t+k} subtract the projection of past values. So, the Autocovariance between them is,

$$\gamma(k) = E(X_t - P_t)E(X_{t+k} - P_{t+k})$$

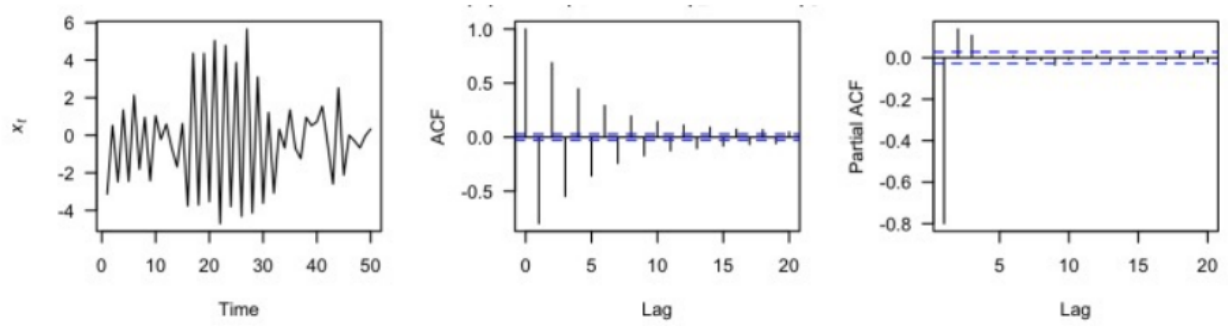
The PACF determines the partial correlation between the time period t and $t - k$. It doesn't take into consideration all the time lags between t and $t - k$. For example, today's stock price might be dependent on 3 days prior stock price but it might not take into consideration yesterday's stock price closure. Hence, we consider only the time lags having a direct impact on future time periods by neglecting the insignificant time lags in between the two-time slots t and $t - k$.

So for AR modes, the PACF becomes zero after Tau is greater than p. So, using PACF, we are going to estimate the order of p.

$$\text{For AR(p)} : \gamma(k) = 0 \ \forall \ k \geq p$$

Let's look at some examples for determining the order p for AR models.

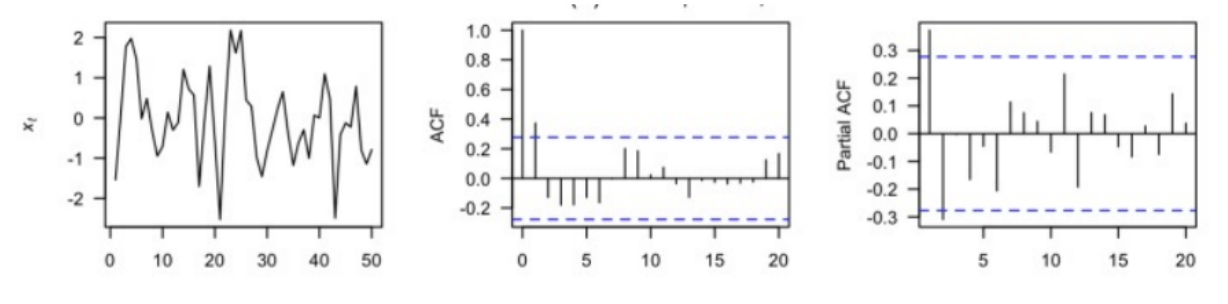
$$X_t = -.7\,X_{t-1} + .2\,X_{t-2} + .1\,X_{t-3}$$



The graph in the above image corresponds to the AR(3) model. As we can see in the equation, we are using the past three values to predict the current value. We observe that the ACF plot is decaying exponentially, so from this plot, we cannot say the order of p but in the PACF plot, we observe that after 3 lags, it's not significant, i.e., the particular autocorrelation is not significant for lags after 3. So, using the PACF plot, we can say that it's an AR(3) process.

Let's take a look at another example,

$$X_t = .7\,W_{t-1} + W_t$$



The graph in the above image corresponds to the MA(2) model. As we can see in the equation, we are using the past two errors to predict the current value. We can observe that the PACF is decaying exponentially, so from this plot, we cannot say the order of q

but in the ACF plot, we observe that after 2 lags, it's not significant, i.e., the autocorrelation is not significant after lag 2. So, using the ACF plot, we can say that it's an MA(2) process.

ACF vs PACF

	ACF	**PACF**
AR(p)	Decays	Zero for $h > p$
MA(q)	Zero for $h > p$	Decays
ARMA(p, q)	Decays	Decays

To conclude, for AR models, we have to look at the PACF plot, and for MA models, we have to look at the ACF plot. For ARMA models, both ACF and PACF decay exponentially, we can get a slight idea, but we need techniques such as AIC to compare different models.