

Introduction to ARIMA Models

Auto Regressive Integrated Moving Average' is actually a class of models that 'explains' a given time series based on its own past values, that is, its own lags and the lagged forecast errors, so that equation can be used to forecast future values. Any 'non-seasonal' time series that exhibits patterns and is not a random white noise can be modeled with ARIMA models.

$$y'_t = c + \phi_1 y'_{t-1} + \cdots + \phi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t,$$

ARIMA model

- An ARIMA model is characterized by 3 terms: p , d , q
- p is the order of the AR term
- q is the order of the MA term
- d is the number of differencing required to make the time series stationary

What is D?

To make a series stationary?

- The most common approach is to difference it. That is, subtract the previous value from the current value. Sometimes, depending on the complexity of the series, more than one differencing may be needed.
- The value of d , therefore, is the minimum number of differencing needed to make the series stationary. And if the time series is already stationary, then $d = 0$.

Next, what are the 'p' and 'q' terms?

'p' is the order of the 'Auto Regressive' (AR) term. It refers to the number of lags of Y to be used as predictors. And 'q' is the order of the 'Moving Average' (MA) term. It refers to the number of lagged forecast errors that should go into the ARIMA Model.

Table 8.1: Special cases of ARIMA models.

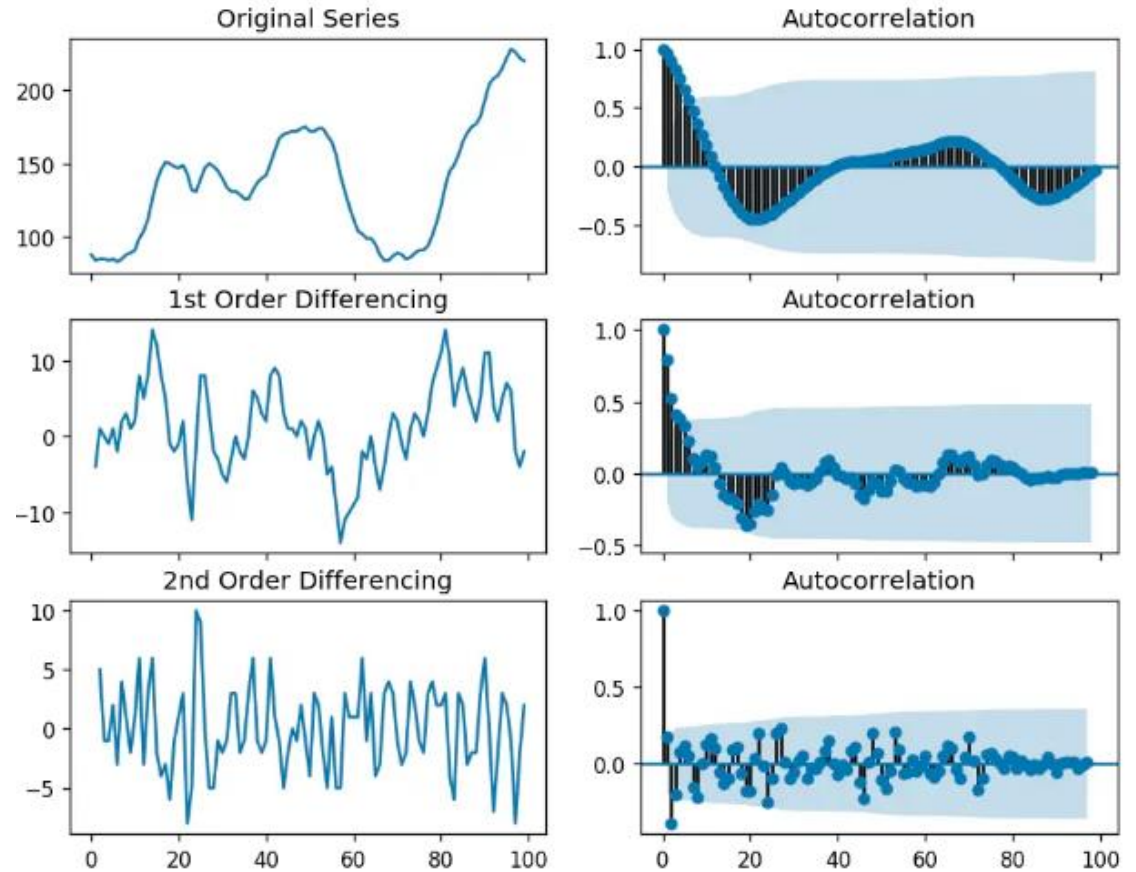
White noise	ARIMA(0,0,0)
Random walk	ARIMA(0,1,0) with no constant
Random walk with drift	ARIMA(0,1,0) with a constant
Autoregression	ARIMA(p,0,0)
Moving average	ARIMA(0,0,q)





How to find the order of differencing (d) in ARIMA model

- The purpose of differencing it to make the time series stationary.
- But you need to be careful to not over-difference the series. Because, an over differenced series may still be stationary, which in turn will affect the model parameters.
- The right order of differencing is the minimum differencing required to get a near-stationary series which roams around a defined mean and the ACF plot reaches to zero fairly quick.
- If the autocorrelations are positive for many number of lags (10 or more), then the series needs further differencing. On the other hand, if the lag 1 autocorrelation itself is too negative, then the series is probably over-differenced.
- In the event, you can't really decide between two orders of differencing, then go with the order that gives the least standard deviation in the differenced series.

.the time series reaches stationarity with two orders of differencing. But on looking at the autocorrelation plot for the 2nd differencing the lag goes into the far negative zone fairly quick, which indicates, the series might have been over differenced.

.So, we can tentatively fix the order of differencing as 1 even though the series is not perfectly stationary (weak stationarity).



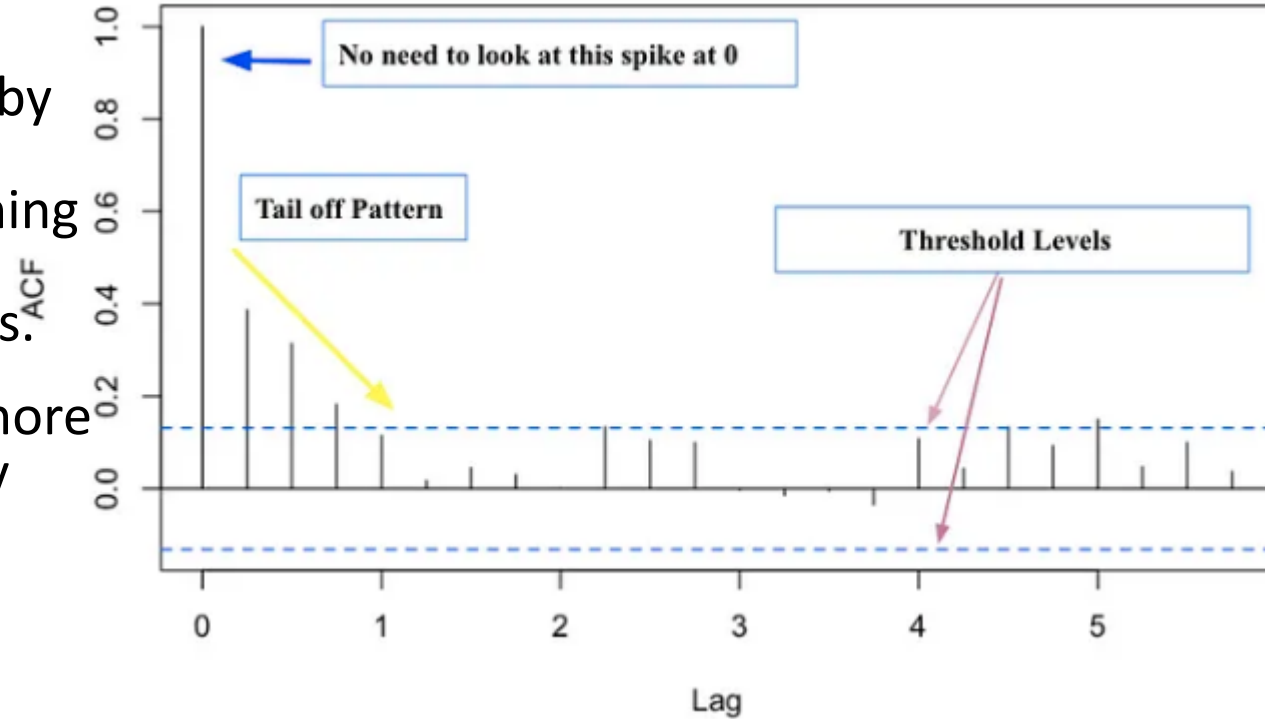
- 
- The basic guideline for interpreting the ACF and PACF plots are as following:
 - Look for tail off pattern in either ACF or PACF.
 - If tail off at ACF \rightarrow AR model \rightarrow Cut off at PACF will provide order p for $AR(p)$.
 - If tail off at PACF \rightarrow MA model \rightarrow Cut off at ACF will provide order q for $MA(q)$.
 - Tail of at both ACF and PACF \rightarrow ARMA model
- 
- 
- 

	ACF	PACF
$AR(p)$	Tails off (trend to zero gradually)	Cuts off after lag p
$MA(q)$	Cuts off after lag q (disappear or zero)	Tails off
$ARMA(p,q)$	Tails off after lag $(q - p)$	Tails off after lag $(q - p)$

ACF

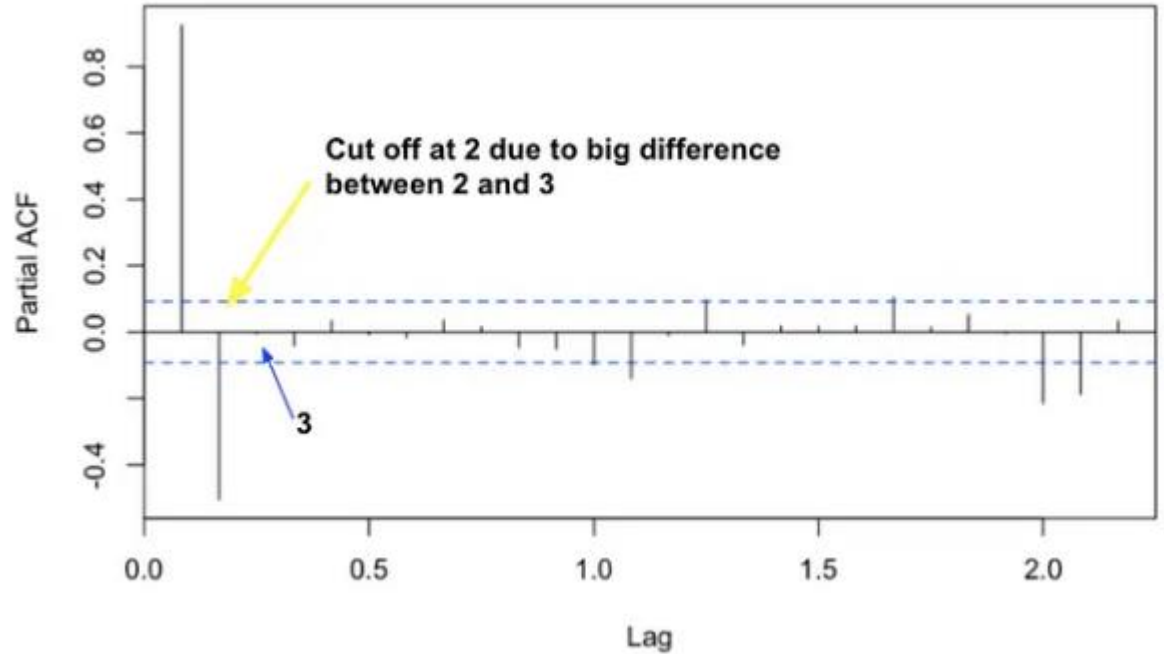
•The two blue dash lines pointed by purple arrows represent the significant threshold levels. Anything that spikes over these two lines reveals the significant correlations.

•When looking at ACF plot, we ignore the long spike at lag 0 (pointed by the blue arrow).



PACF

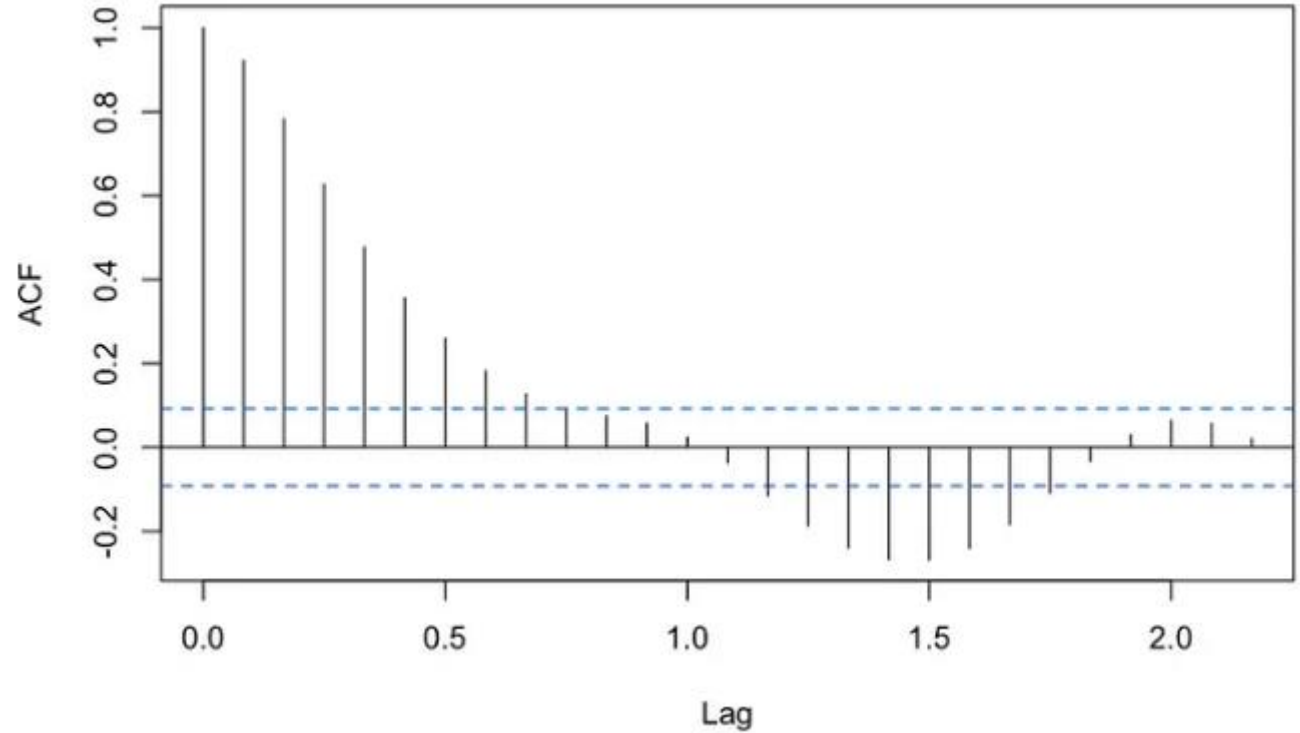
- For PACF, the line usually starts at 1.
- The lag axes will be different depending on the times series data.



How to find the order of the AR term (p)

•Here's the ACF for AR model.

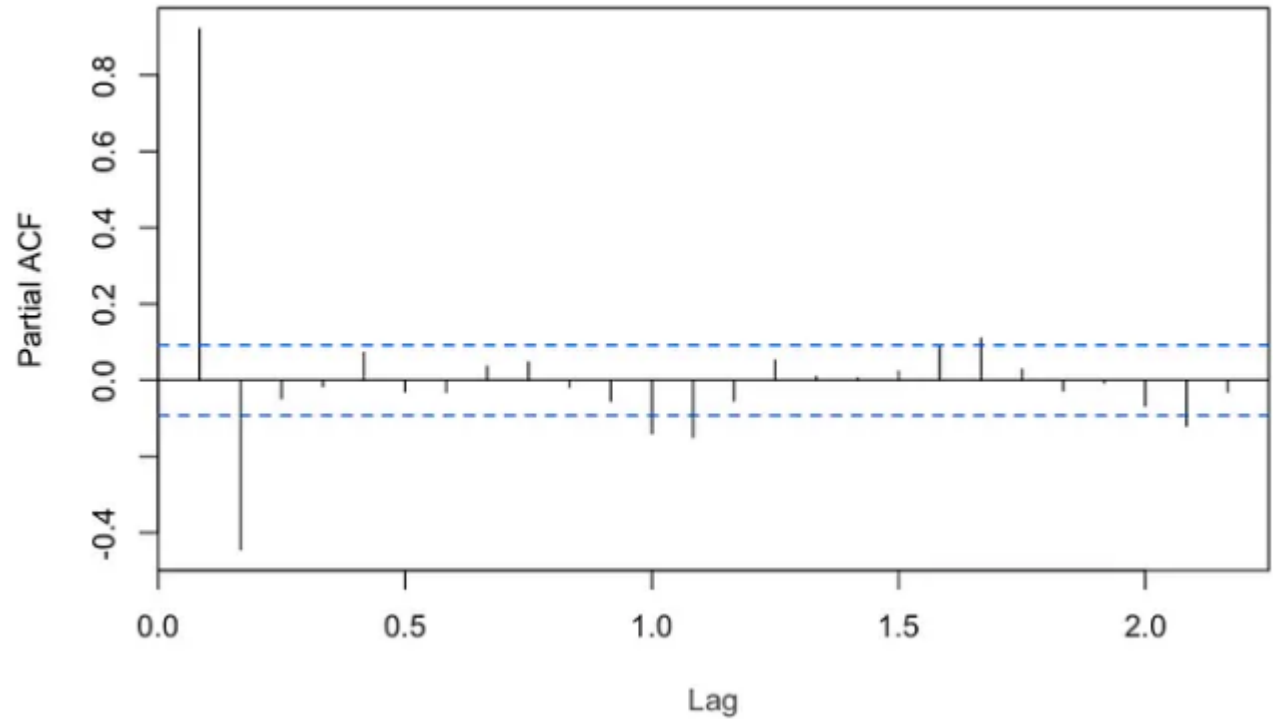
•Tail off is observed at ACF plot. Thus, it's a AR model.



How to find the order of the AR term (p)

• Here's PACF plots of the AR model.

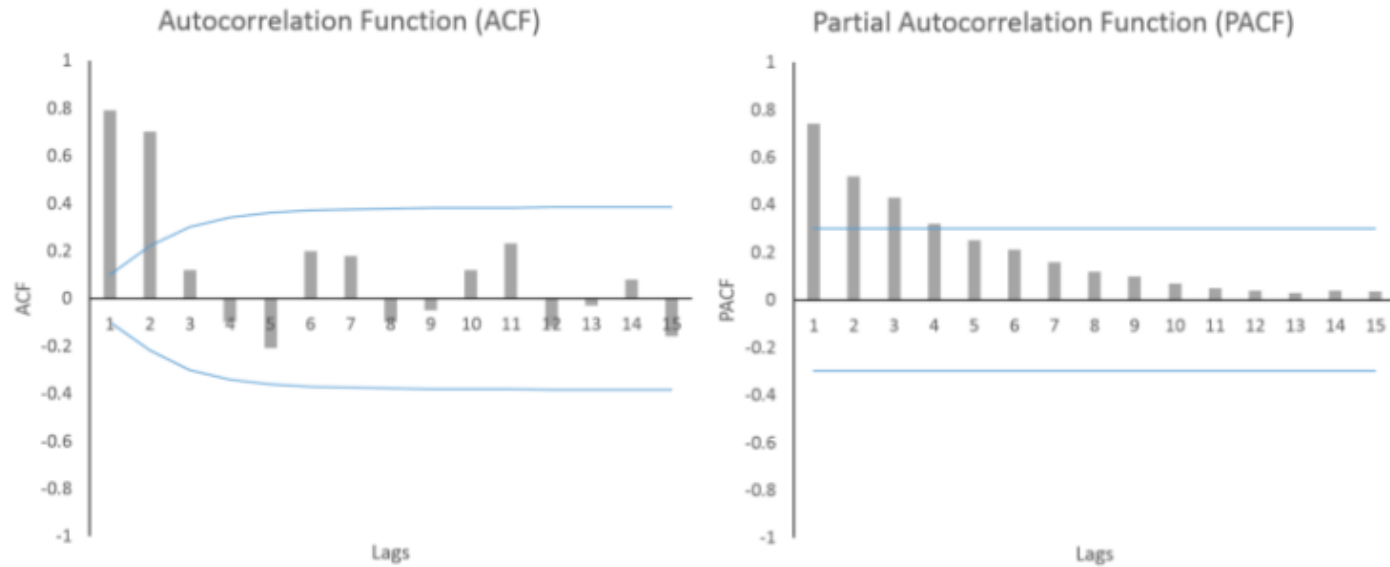
• From PACF, cut off happens at lag 2. Thus, the order is 2. So it should be AR(2) model.



How to find the order of the MA term (q)

•MA process shows a gradually geometrically declining PACF and the ACF has a few significant lags. This is the opposite of the AR process above.

•Here, the PACF is falling geometrically and the ACF has 2 significant lags before dropping. This indicates MA(2) process



ARMA vs ARIMA

- The difference between ARMA and ARIMA is the integration part. The integrated I stands for the number of times differencing is needed to make the times series stationary.

•ARMA Model Equation

• $r(t)=C+\phi r(t-1)+\theta \varepsilon(t-1)+\varepsilon(t)$ where,

• $r(t), r(t-1)$ = current value and value one period ago.

• $\varepsilon(t), \varepsilon(t-1)$ = current error term and one period ago.

• c = baseline constant factor.

• ϕ = value coefficient, what part of the last period value is relevant in explaining the current value.

• θ = error coefficient, what part of the last period value is relevant in explaining the current error value.

2. ARIMA Model Equation

$\Delta r(t)=C+\phi \Delta r(t-1)+\theta \varepsilon(t-1)+\varepsilon(t)$, where,

• $\Delta r(t)= r(t)-r(t-1)$, difference in consecutive period.

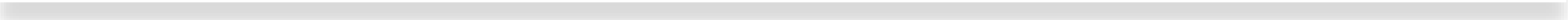
•other is same as the ARMA model.

How to handle if a time series is slightly under or over differenced

- It may so happen that your series is slightly under differenced, that differencing it one more time makes it slightly over-differenced.
- How to handle this case?
- If your series is slightly under differenced, adding one or more additional AR terms usually makes it up. Likewise, if it is slightly over-differenced, try adding an additional MA term.

Vector Autoregression



- The primary difference is those models are uni-directional, where, the predictors influence the Y and not vice-versa. Whereas, Vector Auto Regression (VAR) is bi-directional. That is, the variables influence each other.
- 



How does a VAR model's formula look like?

- In the VAR model, each variable is modeled as a linear combination of past values of itself and the past values of other variables in the system. Since you have multiple time series that influence each other, it is modeled as a system of equations with one equation per variable (time series).
 - That is, if you have 5 time series that influence each other, we will have a system of 5 equations.
-

Equation

Let's suppose, you have two variables (Time series) Y1 and Y2, and you need to forecast the values of these variables at time (t). To calculate Y1(t), VAR will use the past values of both Y1 as well as Y2. Likewise, to compute Y2(t), the past values of both Y1 and Y2 be used.

.For example, the system of equations for a VAR(1) model with two time series (variables `Y1` and `Y2`) is as follows:

.Where, $Y_{1,t-1}$ and $Y_{2,t-1}$ are the first lag of time series Y1 and Y2 respectively.

.The above equation is referred to as a VAR(1) model, because, each equation is of order 1, that is, it contains up to one lag of each of the predictors (Y1 and Y2).

.Since the Y terms in the equations are interrelated, the Y's are considered as endogenous variables, rather than as exogenous predictors.

$$\begin{aligned}Y_{1,t} &= \alpha_1 + \beta_{11,1} Y_{1,t-1} + \beta_{12,1} Y_{2,t-1} + \epsilon_{1,t} \\Y_{2,t} &= \alpha_2 + \beta_{21,1} Y_{1,t-1} + \beta_{22,1} Y_{2,t-1} + \epsilon_{2,t}\end{aligned}$$

•Likewise, the second order VAR(2) model for two variables would include up to two lags for each variable (Y1 and Y2).

$$Y_{1,t} = \alpha_1 + \beta_{11,1} Y_{1,t-1} + \beta_{12,1} Y_{2,t-1} + \beta_{11,2} Y_{1,t-2} + \beta_{12,2} Y_{2,t-2} + \epsilon_{1,t}$$
$$Y_{2,t} = \alpha_2 + \beta_{21,1} Y_{1,t-1} + \beta_{22,1} Y_{2,t-1} + \beta_{21,2} Y_{1,t-2} + \beta_{22,2} Y_{2,t-2} + \epsilon_{2,t}$$

•Can you imagine what a second order VAR(2) model with three variables (Y1, Y2 and Y3) would look like?

$$\begin{aligned}Y_{1,t} &= \alpha_1 + \beta_{11,1}Y_{1,t-1} + \beta_{12,1}Y_{2,t-1} + \beta_{13,1}Y_{3,t-1} + \beta_{11,2}Y_{1,t-2} + \beta_{12,2}Y_{2,t-2} + \beta_{13,2}Y_{3,t-2} + \epsilon_{1,t} \\Y_{2,t} &= \alpha_2 + \beta_{21,1}Y_{1,t-1} + \beta_{22,1}Y_{2,t-1} + \beta_{23,1}Y_{3,t-1} + \beta_{21,2}Y_{1,t-2} + \beta_{22,2}Y_{2,t-2} + \beta_{23,2}Y_{3,t-2} + \epsilon_{2,t} \\Y_{3,t} &= \alpha_3 + \beta_{31,1}Y_{1,t-1} + \beta_{32,1}Y_{2,t-1} + \beta_{33,1}Y_{3,t-1} + \beta_{31,2}Y_{1,t-2} + \beta_{32,2}Y_{2,t-2} + \beta_{33,2}Y_{3,t-2} + \epsilon_{3,t}\end{aligned}$$

As you increase the number of time series (variables) in the model the system of equations

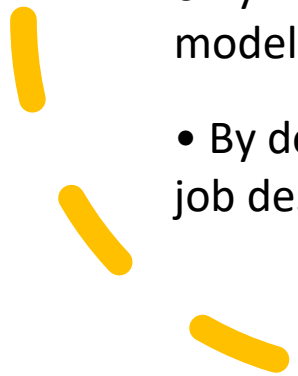


Advantages of Statistical models

- These models are simple and transparent, so they can be understood clearly in terms of their parameters.
- You can apply these models to fairly small data sets and still get good results.
- These simple models and related modifications perform extremely well, even in comparison to very complicated machine learning models. So you get good performance without the danger of overfitting.
- Well-developed automated methodologies for choosing orders of your models and estimating their parameters make it simple to generate these forecasts.



Disdvantages of Statistical models

- Because these models are quite simple, they don't always improve performance when given large data sets. If you are working with extremely large data sets, you may do better with the complex models of machine learning and neural network methodologies.
 - These statistical models put the focus on point estimates of the mean value of a distribution rather than on the distribution. True, you can derive sample variances and the like as some proxy for uncertainty in your forecasts, but your fundamental model offers only limited ways to express uncertainty relative to all the choices you make in selecting a model.
 - By definition, these models are not built to handle nonlinear dynamics and will do a poor job describing data where nonlinear relationships are dominant.
- 

Hidden Markov Model

What are Markov Models?

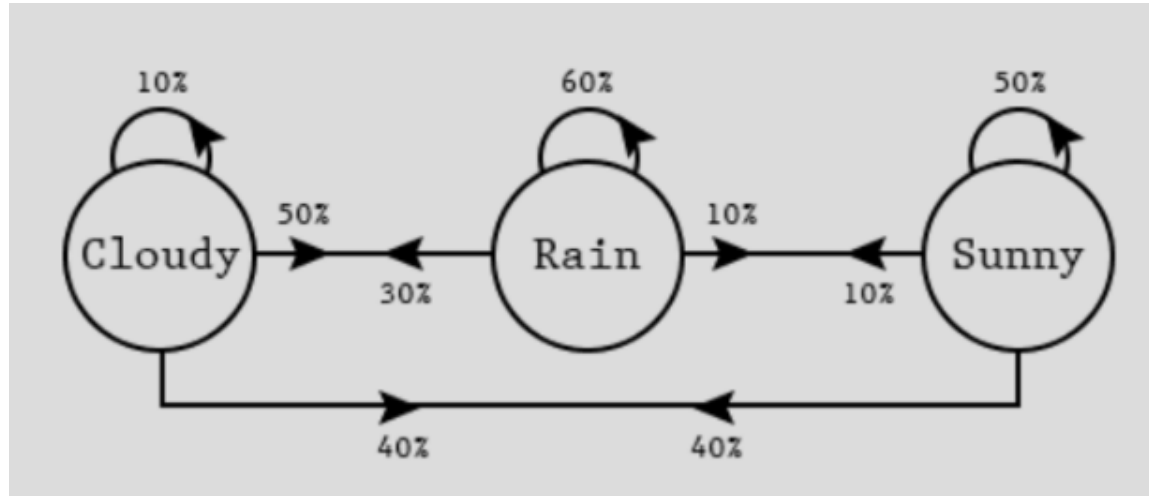
- Markov models are a type of probabilistic model that is used to predict the future state of a system, based on its current state.
- In other words, Markov models are used to predict the future state based on the current hidden or observed states. Markov model is a finite-state machine where each state has an associated probability of being in any other state after one step. They can be used to model real-world problems where hidden and observable states are involved.
- Markov models can be classified into hidden and observable based on the type of information available to use for making predictions or decisions.
- Hidden Markov models deal with hidden variables that cannot be directly observed but only inferred from other observations, whereas in an observable model also termed as Markov chain, hidden variables are not involved.

Markov models

Say you have a bag of marbles that contains four marbles: two red marbles and two blue marbles. You randomly select a marble from the bag, note its color, and then put it back in the bag. After repeating this process several times, you begin to notice a pattern: The probability of selecting a red marble is always two out of four, or 50%. This is because the probability of selecting a particular color of marble is determined by the number of that color of marble in the bag. In other words, the past history (i.e., the contents of the bag) determines the future state (i.e., the probability of selecting a particular color of marble).

What is Markov Chain?

- **Initial probability distribution:** An initial probability distribution over states, π_i is the probability that the Markov chain will start in a certain state i . Some states j may have $\pi_j = 0$, meaning that they cannot be initial states
- **One or more states**
- **Transition probability distribution:** A transition probability matrix A where each a_{ij} represents the probability of moving from state i to state j



•If rainy today, then tomorrow:

- 10% probability for sunny
- 60% probability for rainy
- 30% probability for cloudy

•If rainy today, then tomorrow:

- 10% probability for sunny
- 60% probability for rainy
- 30% probability for cloudy

• If cloudy today, then tomorrow:

- 40% probability for sunny
- 50% probability for rainy
- 10% probability for cloudy

Using this Markov chain, what is the probability that the Wednesday will be cloudy if today is sunny. The following are different transitions that can result in a cloudy Wednesday given today (Monday) is sunny.

- **Sunny – Sunny (Tuesday) – Cloudy (Wednesday)**: The probability to a cloudy Wednesday can be calculated as $0.5 \times 0.4 = 0.2$
- **Sunny – Rainy (Tuesday) – Cloudy (Wednesday)**: The probability of a cloudy Wednesday can be calculated as $0.1 \times 0.3 = 0.03$
- **Sunny – Cloudy (Tuesday) – Cloudy (Wednesday)**: The probability of a cloudy Wednesday can be calculated as $0.4 \times 0.1 = 0.04$

The total probability of a cloudy Wednesday = $0.2 + 0.03 + 0.04 = 0.27$.

Hidden Markov model

HMM is a statistical model in which the system being modeled are Markov processes with unobserved or hidden states. It is a hidden variable model which can give an observation of another hidden state with the help of the Markov assumption. The hidden state is the term given to the next possible variable which cannot be directly observed but can be inferred by observing one or more states according to Markov's assumption.

Hidden Markov models.

- Set of states: $\{s_1, s_2, \dots, s_N\}$
- Process moves from one state to another generating a sequence of states : $s_{i1}, s_{i2}, \dots, s_{ik}, \dots$
- Markov chain property: probability of each subsequent state depends only on what was the previous state:
$$P(s_{ik} \mid s_{i1}, s_{i2}, \dots, s_{ik-1}) = P(s_{ik} \mid s_{ik-1})$$
- States are not visible, but each state randomly generates one of M observations (or visible states) $\{v_1, v_2, \dots, v_M\}$
- To define hidden Markov model, the following probabilities have to be specified: matrix of transition probabilities $A=(a_{ij})$, $a_{ij}= P(s_i \mid s_j)$, matrix of observation probabilities $B=(b_i(v_m))$, $b_i(v_m)= P(v_m \mid s_i)$ and a vector of initial probabilities $\pi=(\pi_i)$, $\pi_i = P(s_i)$. Model is represented by $M=(A, B, \pi)$.