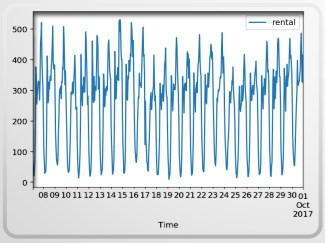
Time series: Autoregressive models

1

Time series



- A time series is a sequential set of data points, measured typically over successive times
- Time series analysis comprises methods for analyzing time series data to extract meaningful statistics and predict future data from past data

2

Categories and Terminologies

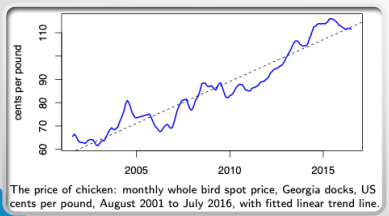
- **univariate vs. multivariate**
  A time series containing records of a single variable is termed as univariate, but if records of more than one variable are considered then it is termed as multivariate
- **linear vs. non-linear**
  A time series model is said to be linear or non-linear depending on whether the current value of the series is a linear or non-linear function of past observations
- **discrete vs. continuous**
  In a continuous time series observations are measured at every instance of time, whereas a discrete time series contains observations measured at discrete points in time

Here: **Univariate, linear, discrete time series**

3

Components of a Time Series

In general, a time series is affected by four components: trend, seasonal, cyclical and irregular components

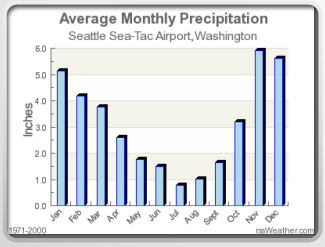


The price of chicken: monthly whole bird spot price, Georgia docks, US cents per pound, August 2001 to July 2016, with fitted linear trend line.

- **Trend**
  The general tendency of a time series to increase, decrease or stagnate over a long period of time.

4

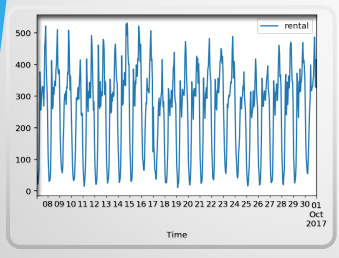
Components of a Time Series



- **Seasonal variation**
  This component explains fluctuations within a year during the season, usually caused by climate and weather conditions, customs, traditional habits, etc.

5

Components of a Time Series



- **Cyclical variation**
  This component describes the medium-term changes caused by circumstances, which repeat in cycles. The duration of a cycle extends over longer periods of time

6

Components of a Time Series

- **Irregular variation**
  Irregular or random variations in a time series are caused by unpredictable influences, which are not regular and also do not repeat in a particular pattern.
  These variations are caused by incidences such as war, strike, earthquake, flood, revolution, etc.
  There is no defined statistical technique for measuring random fluctuations in a time series

7

Modelling using four components

- Considering the effects of these four components, a simple linear model can be defined:

$$Y(t) = T(t) + S(t) + C(t) + I(t)$$

- Assumption: These four components are **independent** of each other

8

Time Series Example: White Noise

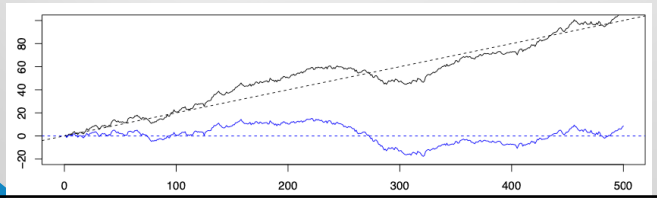
- **White Noise**
  - A simple time series could be a collection of uncorrelated random variables, $\{w_t\}$, with zero mean $\mu = 0$ and finite variance σ^2_w, denoted as $w_t \sim wn(0, \sigma^2_w)$
- **Gaussian White Noise**
  - A particular useful white noise is Gaussian white noise, wherein the w_t are independent normal random variables (with mean 0 and variance σ^2_w), denoted as $w_t \sim iid\ N(0, \sigma^2_w)$
- White noise time series is of great interest because if the stochastic behaviour of all time series could be explained in terms of the white noise model, then classical statistical methods would suffice

9

Time Series Example: Random Walk

- A random walk is the process by which randomly-moving objects wander away from where they started
- Consider a simple 1-D process:
  - The value of the time series at time t is the value of the series at time $t - 1$ plus a completely random movement determined by w_t. More generally, a constant drift factor δ is introduced.

$$X_t = \delta + X_{t-1} + w_t = \delta t + \sum_{i=1}^{t} w_i$$



10

Time Series Analysis

- The procedure of using known data values to fit a time series with suitable model and estimating the corresponding parameters
- It comprises methods that attempt to understand the nature of the time series and is often useful for future forecasting and simulation
- There are several ways to build time series forecasting models, but this lecture focuses on **stochastic process**
  - We assume a time series can be defined as a collection of random variables indexed according to the order they are obtained in time, $X_1, X_2, X_3, \ldots$
  - t will typically be discrete and vary over the integers $t = 0, \pm 1, \pm 2, \ldots$
  - The collection of random variables $\{X_t\}$ is referred to as a **stochastic process**, while the observed values are referred to as a **realization** of the stochastic process

11

Autocovariance for Time Series

- Lack of independence between adjacent values in time series X_s and X_t can be numerically assessed
- **Autocovariance Function**
  - Assuming the variance of X_t is finite, the autocovariance function is defined as the second moment product

$$\gamma(s,t) = \gamma_X(s,t) = cov(X_s, X_t) = E[(X_s - \mu_s)(X_t - \mu_t)]$$

for all s and t

  - Note that $\gamma(s,t) = \gamma(t,s)$ for all time points s and t
- The autocovariance measures the **linear dependence** between two points on the same series observed at different times
  - Very smooth series exhibit autocovariance functions that stay large even when the t and s are far apart, whereas random series tend to have autocovariance functions that are nearly zero for large separations

12

Autocorrelation for Time Series

- Autocorrelation Function (ACF)
  - The autocorrelation function is defined as

$$\rho(s,t) = \frac{\gamma(s,t)}{\sqrt{\gamma(s,s)\gamma(t,t)}}$$

  - It's easy to show that $-1 \le \rho(s,t) \le 1$
- ACF measures the linear predictability of X_t using only X_s
  - If we can predict X_t perfectly from X_s through a linear relationship, then ACF will be either +1 or −1

13

Stationarity of Stochastic Process

- Forecasting is difficult as time series is non-deterministic in nature, i.e., we cannot predict with certainty what will occur in the future
- But the problem could be easier if the time series is stationary: you simply predict its statistical properties will be the same in the future as they have been in the past
  - A stationary time series is one whose statistical properties such as mean, variance, autocorrelation, etc. are all constant over time
- Most statistical forecasting methods are based on the assumption that the time series can be rendered approximately stationary after mathematical transformations

14

Autocorrelation for Stationary Time Series

- The autocovariance $\gamma_X(s, t)$ of stationary time series depends on s and t only through $|s - t|$, thus we can rewrite notation $s = t + h$, where h represents the time shift

$$\gamma_X(t+h, t) = cov(X_{t+h}, X_t) = cov(X_h, X_0) = \gamma(h, 0) = \gamma(h)$$

- **Autocovariance Function of Stationary Time Series**

$$\gamma(h) = cov(X_{t+h}, X_t) = E[(X_{t+h} - \mu)(X_t - \mu)]$$

- **Autocorrelation Function of Stationary Time Series**

$$\rho(h) = \frac{\gamma(t+h, t)}{\sqrt{\gamma(t+h, t+h)\gamma(t, t)}} = \frac{\gamma(h)}{\gamma(0)}$$

15

Partial Autocorrelation

- Another important measure is called **partial autocorrelation**, which is the correlation between X_s and X_t with the linear effect of "everything in the middle" removed
- **Partial Autocorrelation Function (PACF)**
  - For a stationary process X_t, the PACF (denoted as ϕ_{hh}), for h = 1,2,... is defined as

$$\phi_{11} = \text{corr}(X_{t+1}, X_t) = \rho_1$$

$$\phi_{hh} = \text{corr}(X_{t+h} - \hat{X}_{t+h}, X_t - \hat{X}_t), \quad h \geq 2$$

  - With

$$\hat{X}_{t+h} = \beta_1 X_{t+h-1} + \beta_2 X_{t+h-2} + \cdots + \beta_{h-1} X_{t+1}$$

$$\hat{X}_t = \beta_1 X_{t+1} + \beta_2 X_{t+2} + \cdots + \beta_{h-1} X_{t+h-1}$$

16

ARIMA Models

17

ARIMA Models

- ARIMA is an acronym that stands for **A**uto-**R**egressive **I**ntegrated **M**oving **A**verage
  - **AR** Autoregression. A model that uses the dependent relationship between an observation and some number of lagged observations
  - **I** Integrated. The use of differencing of raw observations in order to make the time series stationary
  - **MA** Moving Average. A model that uses the dependency between an observation and a residual error from a moving average model applied to lagged observations
- Each of these components are explicitly specified in the model as a parameter
- Note: **AR** and **MA** are two widely used linear models that work on stationary time series, and **I** is a preprocessing procedure to "stationarize" time series if needed

18

Notations

- A standard notation is used of *ARIMA(p, d, q)* where the parameters are substituted with integer values to quickly indicate the specific ARIMA model being used
  - **p:** The number of lag observations included in the model, also called the **lag order**
  - **d:** The number of times that the raw observations are differenced, also called the **degree of differencing**
  - **q:** The size of the moving average window, also called the **order of moving average**
- A value of 0 can be used for a parameter, which indicates to not use that element of the model
- That is, ARIMA model can be configured to perform the function of an ARMA model, and even a simple AR, I, or MA model

19

Autoregressive Models

- Intuition
  - Autoregressive models are based on the idea that current value of the series, X_t, can be explained as a linear combination of p past values, $X_{t-1}, X_{t-2}, \ldots, X_{t-p}$, together with a random error in the same series
- Definition
  - An autoregressive model of order p, abbreviated $AR(p)$, is of the form

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + w_t = \sum_{i=1}^{p} \phi_i X_{t-i} + w_t$$

- where X_t is stationary, $w_t \sim wn(0,\sigma^2_w)$, and $\phi_1, \phi_2, \ldots, \phi_p$ are **model parameters**
  - The hyperparameter p represents the length of the "direct look back" in the series

20

AR Example: $AR(0)$ and $AR(1)$

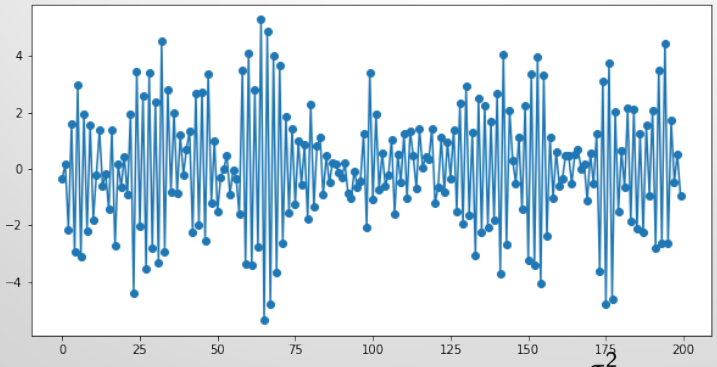
- The simplest AR process is $AR(0)$, which has no dependence between the terms. In fact, $AR(0)$ is essentially white noise
- $AR(1)$ can be given by $X_t = \phi_1 X_{t-1} + w_t$
  - Only the previous term in the process and the noise term contribute to the output
  - If ϕ_1 is close to 0, then the process still looks like white noise
  - If $\phi_1 < 0$, X_t tends to oscillate between positive and negative values
  - If $\phi_1 = 1$ then the process is equivalent to random walk, which is not stationary as the variance is dependent on t (and infinite)

21

AR Examples: $AR(1)$ Process

- Simulated $AR(1)$ process

$$X_t = -0.9 X_{t-1} + w_t$$

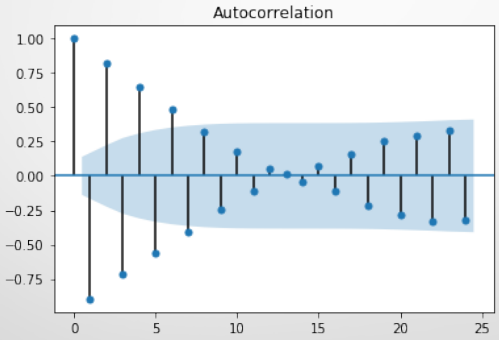


$$E[X_t] = 0 \qquad \text{Var}(X_t) = \frac{\sigma_w^2}{(1-\phi_1^2)}$$

22

AR Examples: $AR(1)$ Process

- Autocorrelation Function (ACF)

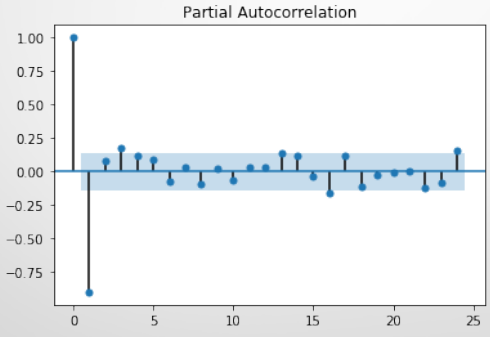


$$\rho_h = \phi_1^h$$

23

AR Examples: $AR(1)$ Process

- Partial Autocorrelation Function (PACF)



$$\phi_{11} = \rho_1 = \phi_1 \qquad \phi_{hh} = 0, \forall h \geq 2$$

24

General $AR(p)$ Process

- An important property of $AR(p)$ models in general is
  - When $h > p$, theoretical **partial autocorrelation function is 0**:

$$\phi_{hh} = \text{corr}(X_{t+h} - \hat{X}_{t+h}, X_t - \hat{X}_t) = \text{corr}(w_{t+h}, X_t - \hat{X}_t) = 0$$

  - When $h \leq p$, ϕ_{pp} is not zero and $\phi_{11}, \phi_{22}, \ldots, \phi_{h-1,h-1}$ are not necessarily zero
- In fact, **identification of an AR model is often best done with the PACF**

25

AR Models: Parameters Estimation

- Note that p is like a hyperparameter for the $AR(p)$ process, thus fitting an $AR(p)$ model presumes p is known and only focusing on estimating coefficients, i.e. $\phi_1, \phi_2, \ldots, \phi_p$
- There are many feasible approaches:
  - Method of moments estimator (e.g. Yule-Walker estimator)
  - Ordinary Least Squares (OLS) estimator
  - **Maximum Likelihood Estimation (MLE) estimator**
- If the observed series is short or the process is far from stationary, then substantial differences in the parameter estimations from various approaches are expected

26

Moving Average Models (MA)

- The name might be misleading: **moving average models** are not be confused with the **moving average smoothing**
- Motivation
  - Recall that in AR models, current observation X_t is regressed using the previous observations $X_{t-1}, X_{t-2}, \ldots, X_{t-p}$ plus an error term w_t at current time point
  - One problem of AR model is the ignorance of correlated noise structures (which is unobservable) in the time series
  - In other words, the imperfectly predictable terms in current time w_t and previous steps, $w_{t-1}, w_{t-2}, \ldots, w_{t-q}$ are also informative for predicting observations

27

Moving Average Models (MA)

- Definition
  - A moving average model of order q, or $MA(q)$, is defined to be

$$X_t = w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + \cdots + \theta_q w_{t-q} = w_t + \sum_{j=1}^{q} \theta_j w_{t-j}$$

where $w_t \sim wn(0, \sigma^2_w)$, and $\theta_1, \theta_2, \ldots, \theta_q$ are parameters

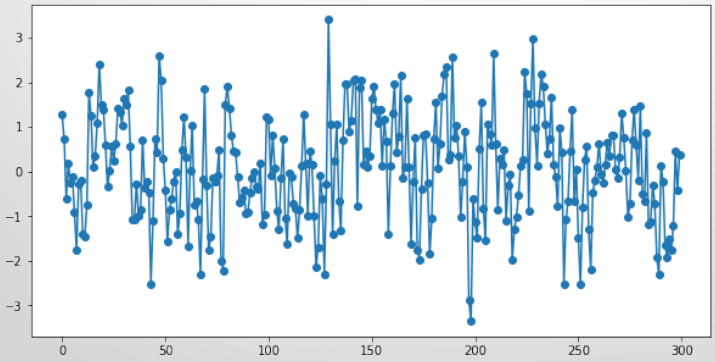
  - Although it looks like a regression model, the difference is that the w_t is **not observable**
- Contrary to AR model, finite MA model is always stationary, because the observation is just a weighted moving average over past forecast errors

28

MA Examples: *MA(1)* Process

- Simulated *MA(1)* process

$$X_t = w_t + 0.8\, w_{t-1}$$

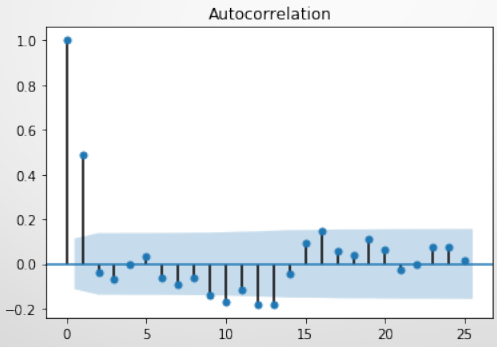


$$E[X_t] = 0 \qquad \mathrm{Var}(X_t) = \sigma_w^2(1 + \theta_1^2)$$

29

MA Examples: *MA(1)* Process

- Autocorrelation Function (ACF)

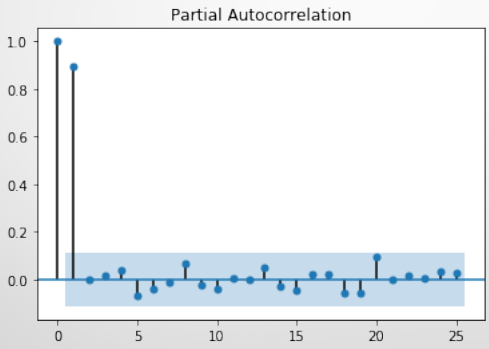


$$\rho_1 = \frac{\theta_1}{1 + \theta_1^2} \qquad \rho_h = 0, \forall h \geq 2$$

30

MA Examples: *MA(1)* Process

- Partial Autocorrelation Function (PACF)



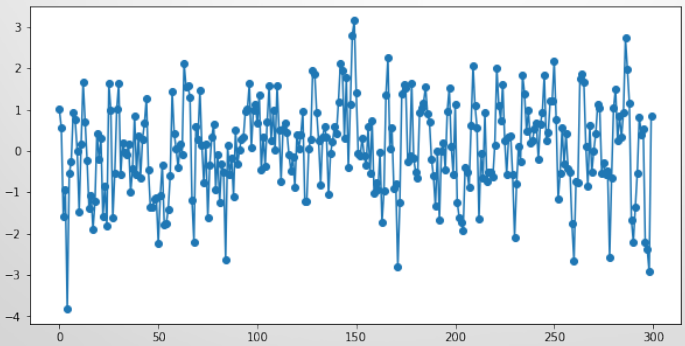
$$\phi_{hh} = -\frac{(-\theta_1)^h(1-\theta_1^2)}{1-\theta_1^{2(h+1)}}, \qquad h \geq 1$$

31

MA Examples: *MA(2)* Process

- Simulated *MA(2)* process

$$X_t = w_t + 0.5\, w_{t-1} + 0.3\, w_{t-2}$$

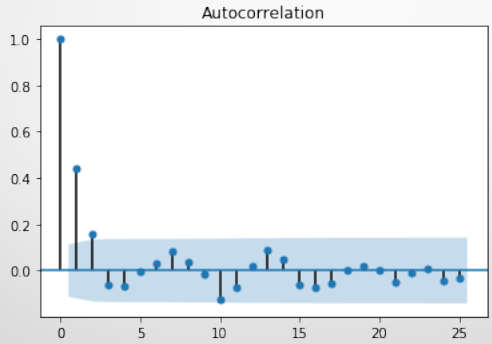


$$E[X_t] = 0 \qquad \mathrm{Var}(X_t) = \sigma_w^2(1+\theta_1^2+\theta_2^2)$$

32

MA Examples: $MA(2)$ Process

- Autocorrelation Function (ACF)



$$\rho_1 = \frac{\theta_1 + \theta_1\theta_2}{1 + \theta_1^2 + \theta_2^2} \qquad \rho_2 = \frac{\theta_2}{1 + \theta_1^2 + \theta_2^2} \qquad \rho_h = 0, \forall h \geq 3$$

33

General $MA(q)$ Process

- An important property of $MA(q)$ models in general is that there are **nonzero autocorrelations** for the first q lags, and $\boldsymbol{\rho_h = 0}$ *for all lags* $h>q$
- The ACF provides a considerable amount of information about the order of the dependence q for $MA(q)$ process
- **Identification of an MA model is often best done with the ACF rather than the PACF**

34

MA Models: Parameters Estimation

- Parameter estimation for MA model is more difficult than AR model
  - One reason is that the lagged error terms are not observable
- We can still use **method of moments** estimators for MA process
  - In fact, since MA process is nonlinear in the parameters, we need iterative non-linear fitting instead of linear least squares
- From a practical point of view, ***modern scientific computing software packages will handle most of the details after given the correct configurations***

35

$ARMA$ Models

- Autoregressive and moving average models can be combined together to form $ARMA$ models
- Definition
  - A time series $\{x_t; t=0, \pm 1, \pm 2, \ldots\}$ is $ARMA(p,q)$ if it is **stationary** and

$$X_t = w_t + \sum_{i=1}^{p} \phi_i X_{t-i} + \sum_{j=1}^{q} \theta_j w_{t-j}$$

Where ϕ_p and θ_q are not 0, and $\sigma^2_w > 0$, $w_t \sim wn(0,\sigma^2_w)$

36

Choosing Model Specification

- ACF and PACF can be used for determining $ARMA$ model hyperparamters p and q

	$AR(p)$	$MA(q)$	$ARMA(p,q)$
ACF	Tails off	Cuts off after lag q	Tails off
PACF	Cuts off after lag p	Tails off	Tails off

  - Note that the selection for p and q is not unique

37

"Stationarize" Nonstationary Time Series

- One limitation of ARMA models is the **stationarity condition**
- In many situations, time series can be thought of as being composed of two components, a non-stationary trend series μ_t and a zero-mean stationary series Y_t

$$X_t = \mu_t + Y_t$$

- Strategies
  - **Detrending**: Subtracting with an estimate for trend and deal with residuals

$$Y'_t = X_t - \mu'_t$$

  - **Differencing**: Recall that random walk with drift is capable of representing trend, thus we can model trend as a stochastic component as well

$$\mu_t = \delta + \mu_{t-1} + w_t$$

$$\nabla X_t = X_t - X_{t-1} = \delta + w_t + (Y_t - Y_{t-1}) = \delta + w_t + \nabla Y_t$$

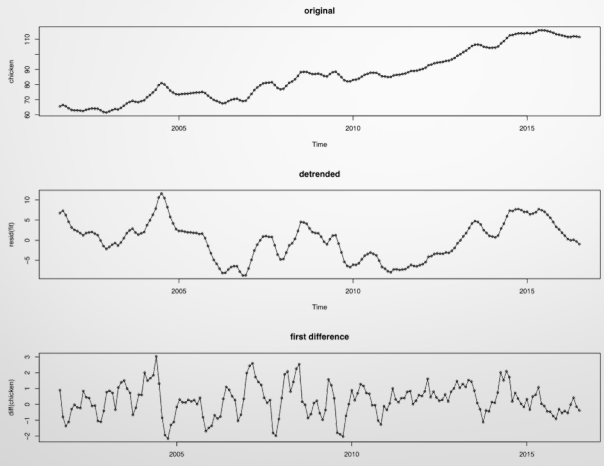
38

Differencing

- One advantage of differencing over detrending for trend removal is that no parameter estimation is required
- In fact, differencing operation can be **repeated**
  - The first difference eliminates a linear trend
  - A second difference, i.e., the difference of first difference, can eliminate a quadratic trend
- This allows us to estimate Y_t and then use it to build X_t

39

Detrending vs. Differencing



40

From *ARMA* to *ARIMA*

- Order of Differencing
  - Differences of order d are defined as

$$\nabla^d = (1-B)^d$$

where $(1-B)^d$ can be expanded algebraically for higher integer values of d

- Definition
  - A process X_t is said to be $ARIMA(p, d, q)$ if

$$\nabla^d X_t = (1-B)^d X_t$$

is $ARMA(p, q)$

41

Box-Jenkins Methods

- As we have seen ARIMA models have numerous parameters and hyper parameters. Box and Jenkins suggests an iterative three-stage approach to estimate an ARIMA model
- Procedures
  1. **Model identification**: Checking stationarity and seasonality, performing differencing if necessary, choosing model specification ARIMA(p, d, q)
  2. **Parameter estimation**: Computing coefficients that best fit the selected ARIMA model using *maximum likelihood estimation* or *non-linear least-squares estimation*
  3. **Model checking**: Testing whether the obtained model conforms to the specifications of a stationary univariate process (i.e. the residuals should be independent of each other and have constant mean and variance). If failed go back to step 1.

42

Prediction evaluation

- Given the predicted values X'_t and the actual values X_t, one can compute the prediction error as
  - *Mean Percentage Error:* $MPE = \frac{100}{N} \sum_{t=1}^{N} \left(\frac{X_t - \widehat{X_t}}{X_t} \right)$
  - *Mean Absolute Percentage Error:* $MAPE = \frac{100}{N} \sum_{t=1}^{N} \left| \frac{X_t - \widehat{X_t}}{X_t} \right|$
  - *Mean Square Error:* $MSE = \frac{1}{N} \sum_{t=1}^{N} (X_t - \widehat{X_t})^2$
    - Also know as *Residual Sum of Squares (RSS)*
  - *Root Mean Square Error:* $RMSE = \sqrt{\frac{1}{N} \sum_{t=1}^{N} (X_t - \widehat{X_t})^2}$
- The smaller they are – the better it is

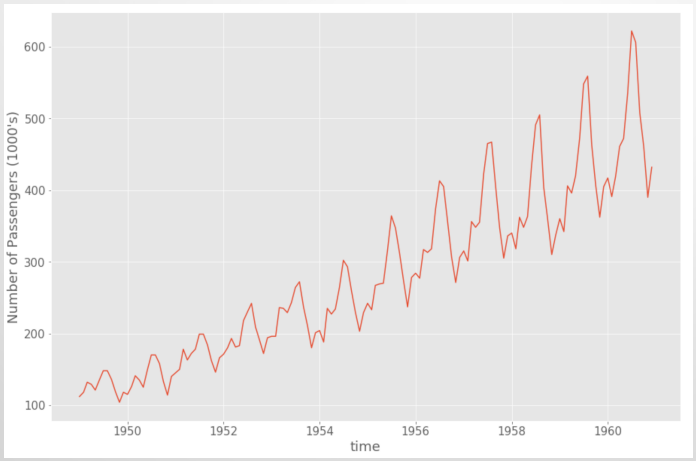
43

Train & test method

Walk – forward validation

44

Example: Air Passenger Data



Monthly totals of a US airline passengers, from 1949 to 1960

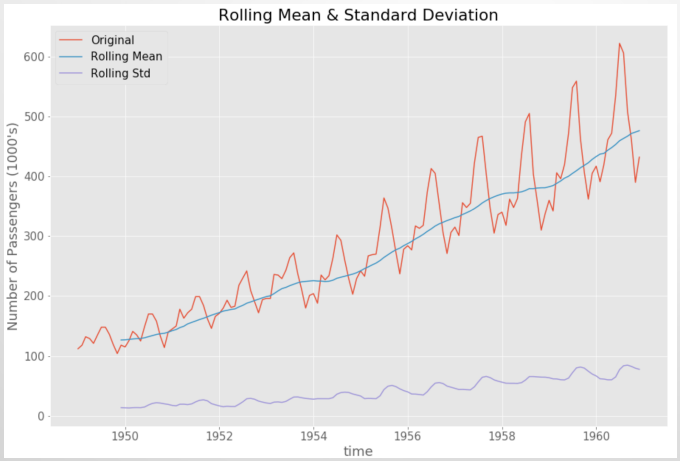
45

Model identification

- As with any data analysis, we should construct a time plot of the data, and inspect the graph for any anomalies
- The most important thing in this phase is to determine if the time series is **stationary** and if there is any **significant seasonality** that needs to be handled
- **Test Stationarity**
  - Recall the definition, if the mean or variance changes over time then it's non-stationary, thus an intuitive way is to **plot rolling statistics**
  - We can also make an **autocorrelation plot**, as non-stationary time series often shows very slow decay

46

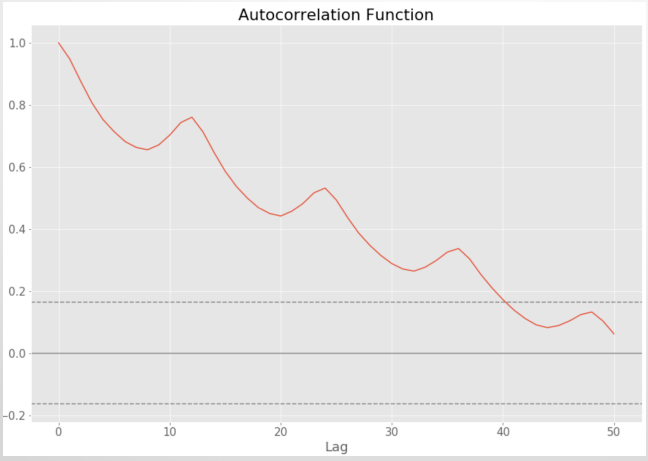
Stationarity Test: Rolling Statistics



Rolling statistics with sliding window of 12 months

47

Stationarity Test: ACF Plot



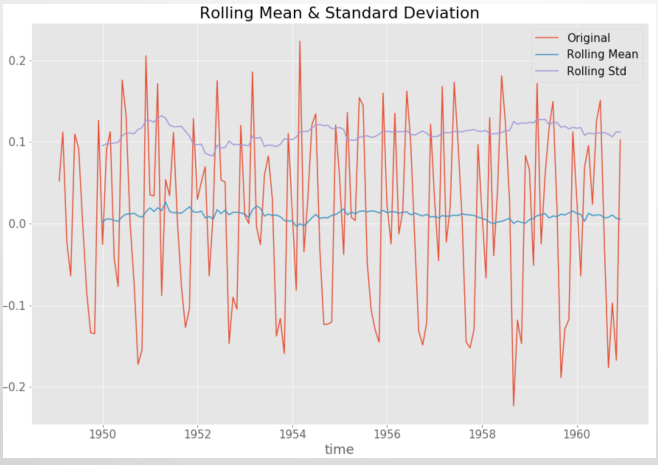
48

Stationarize Time Series

- As all previous methods show that the initial time series is non-stationary, it is necessary to perform **transformations** to make it stationary for ARMA modelling
  - Detrending
  - **Differencing**
  - Transformation: Applying arithmetic operations like log, square root, cube root, etc. to stationarize a time series
  - Aggregation: Taking average over a longer time period, like weekly/monthly
  - Smoothing: Removing rolling average from original time series
  - Decomposition: Modeling trend and seasonality explicitly and removing them from the time series

49

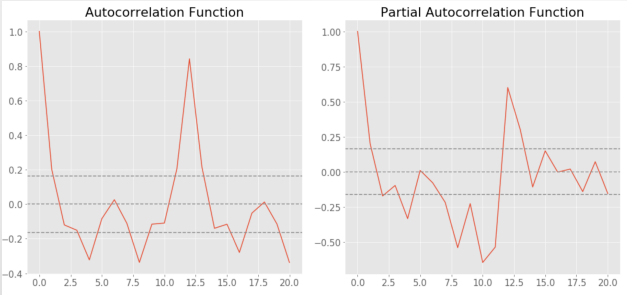
Stationarized Time Series



First order differencing over number of passengers

50

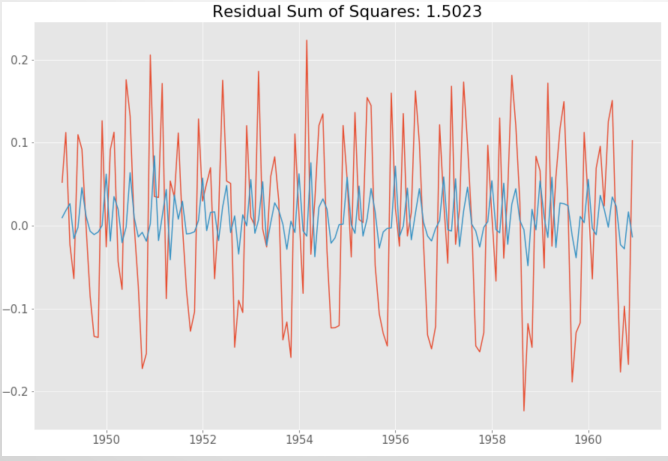
Choosing model specification



- Firstly we notice an obvious peak at $h = 12$, because for simplicity we didn't model the yearly cyclical effect
- It seems $p = 2$, $q = 2$ is a reasonable choice. Let's see three models, $AR(2)$, $MA(2)$ and $ARMA(2, 2)$

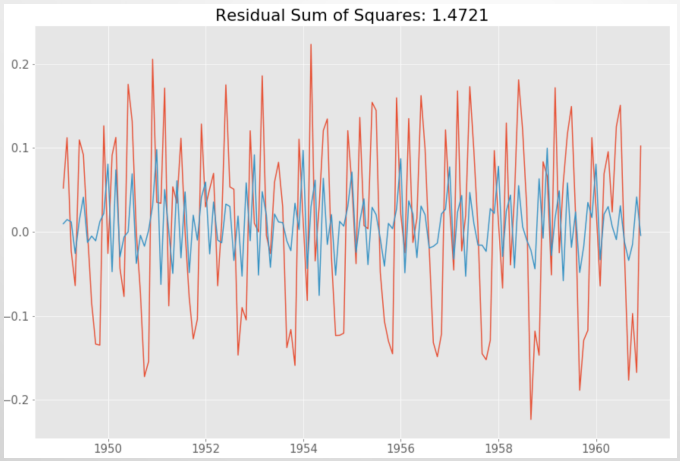
51

$AR(2)$: Predicted on Residuals



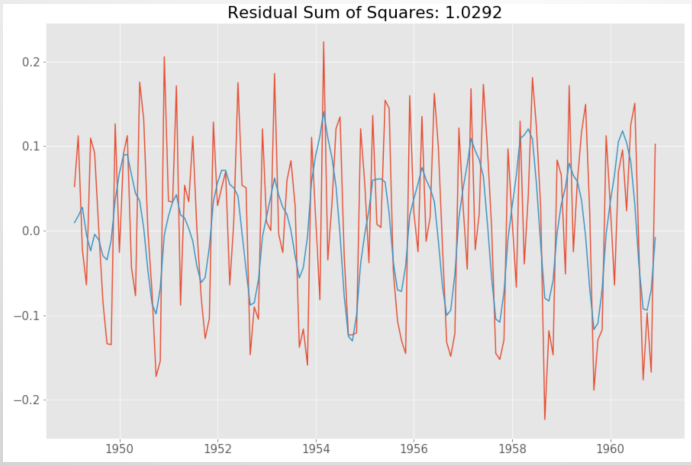
52

$MA(2)$: Predicted on Residuals



53

$ARMA(2)$: Predicted on Residuals

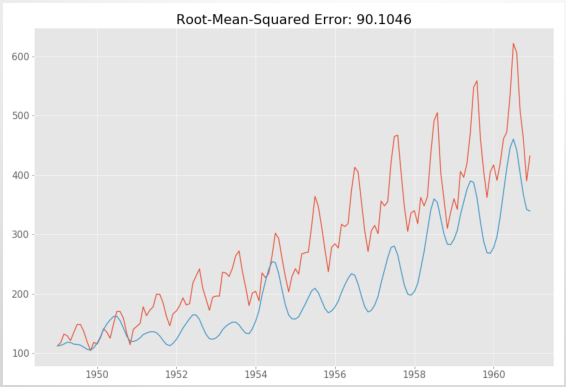


The $AR(2)$ and $MA(2)$ models have almost the same MSE (RSS)
Combined the $ARMA(2,2)$ performs significantly better

54

Forecasting

- The last step is to reverse the transformations we've done to get the prediction on original scale



55

References

- *Time Series: Autoregressive models AR, MA, ARMA, ARIMA*, Mingda Zhang, Univesity of Pittsburgh, 2018
- *Forecasting: principles and practice*, Hyndman, Rob J and Athanasopoulos, George, 2018
- *Time series analysis and its applications*, Shumway, Robert H and Stoffer, David S, 2017
- *A comprehensive beginner's guide to create a Time Series Forecast* (with Codes in Python), https://www.analyticsvidhya.com/blog/2016/02/time-series-forecasting-codes-python/
- *An Introductory Study on Time Series Modeling and Forecasting*, Ratnadip Adhikari, R. K. Agrawal, 2013
- Wikipedia articles for various concepts

56