Time Series: Autoregressive models
AR, MA, ARMA, ARIMA

Mingda Zhang

University of Pittsburgh
mzhang@cs.pitt.edu

October 23, 2018

Overview

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 in order to extract meaningful statistics and other characteristics of the data.

Categories and Terminologies

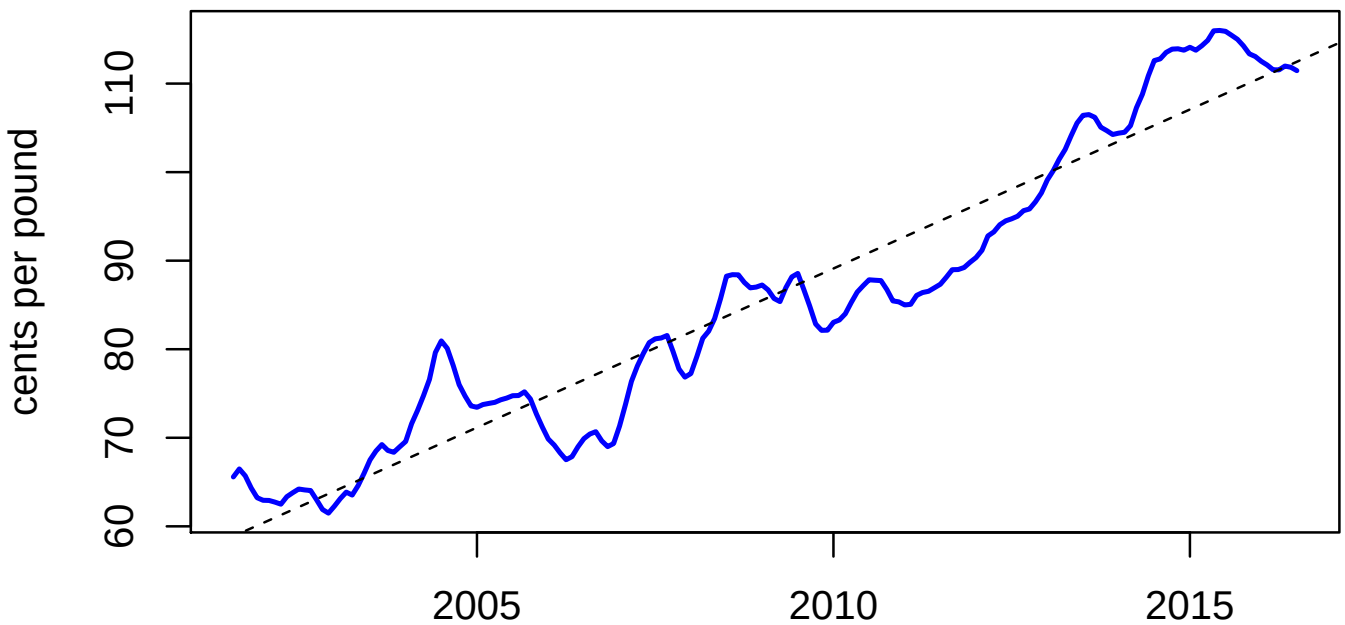
- **Time-domain** vs. **Frequency-domain**
  - Time-domain approach: how does what happened today affect what will happen tomorrow?
    These approaches view the investigation of lagged relationships as most important, e.g. autocorrelation analysis.
  - Frequency-domain approach: what is the economic cycle through periods of expansion and recession?
    These approaches view the investigation of cycles as most important, e.g. spectral analysis and wavelet analysis.
- This lecture will focus on time-domain approaches.

Categories and Terminologies (cont.)

- **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.
- This lecture will focus on univariate, linear, discrete time series.

Components of a Time Series

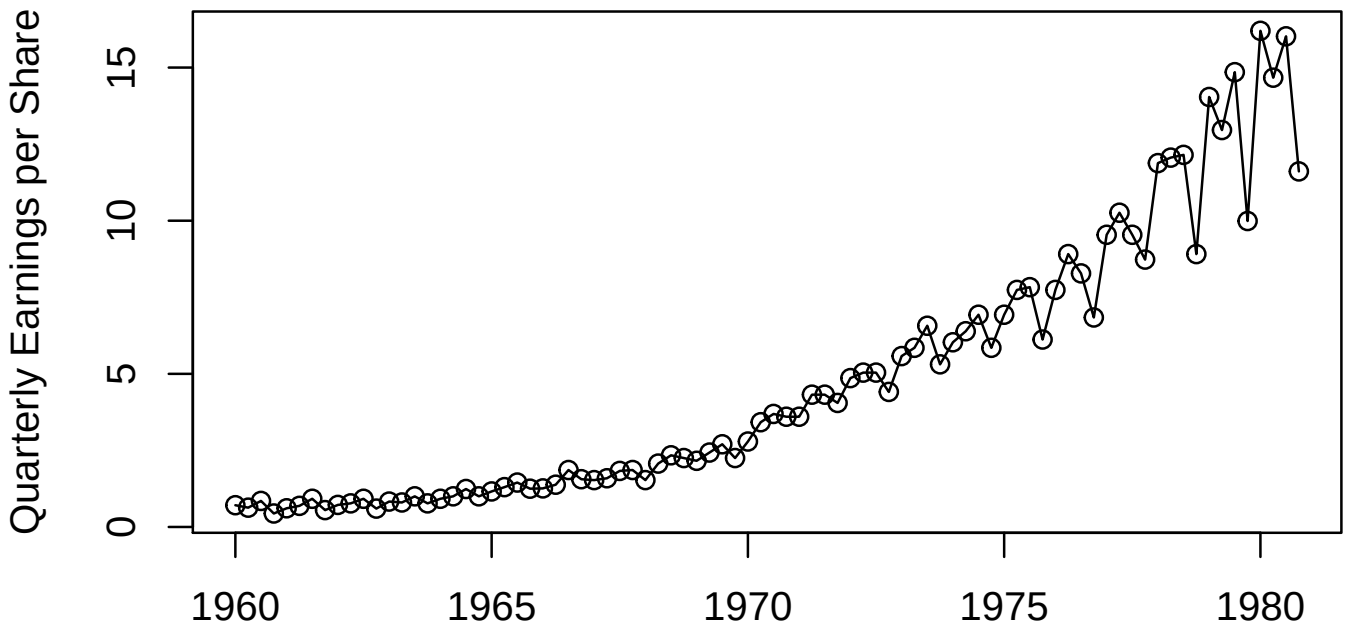
- In general, a time series is affected by four components, i.e. trend, seasonal,cyclical and irregular components.
  - **Trend**
    The general tendency of a time series to increase, decrease or stagnate over a long period of time.



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.

Components of a Time Series (cont.)

- In general, a time series is affected by four components, i.e. trend, seasonal,cyclical and irregular components.
  - **Seasonal variation**
    This component explains fluctuations within a year during the season, usually caused by climate and weather conditions, customs, traditional habits, etc.

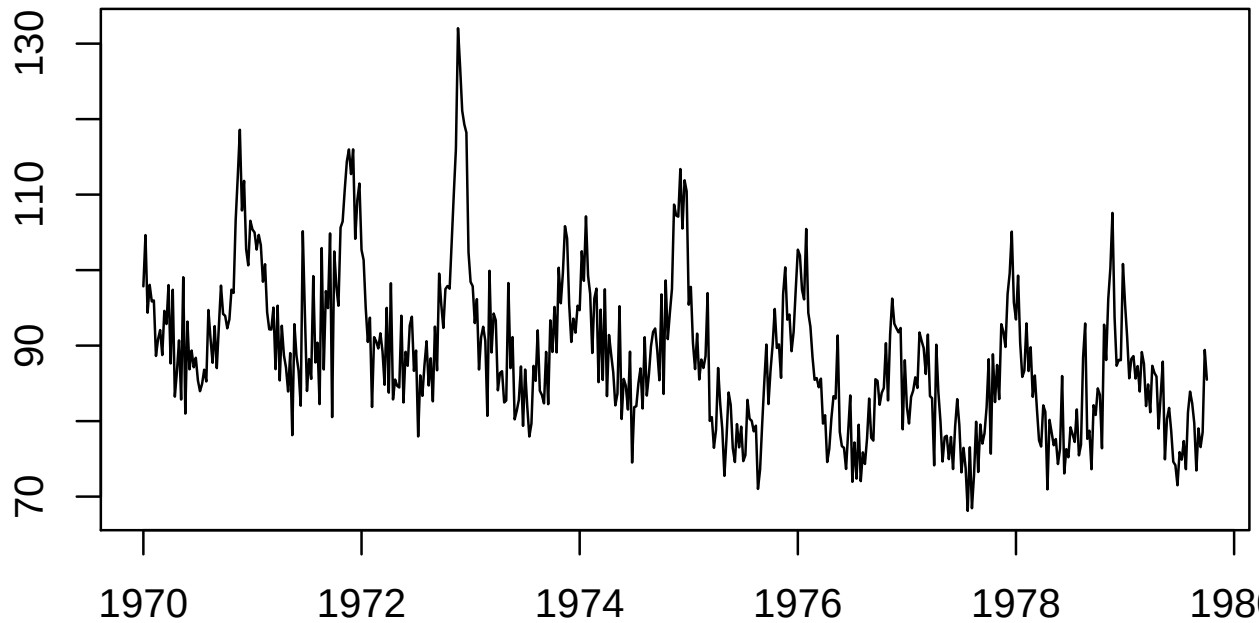


Johnson & Johnson quarterly earnings per share, 84 quarters, 1960-I to 1980-IV.

Components of a Time Series (cont.)

- In general, a time series is affected by four components, i.e. trend, seasonal,cyclical and irregular components.
  - **Cyclical variation**

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



Average weekly cardiovascular mortality in Los Angeles County. There are 508 six-day smoothed averages obtained by filtering daily values over the 10 year period 1970-1979.

Components of a Time Series (cont.)

- In general, a time series is affected by four components, i.e. trend, seasonal,cyclical and irregular components.
  - **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.

Combination of Four Components

- Considering the effects of these four components, two different types of models are generally used for a time series.
  - Additive Model

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

    Assumption: These four components are independent of each other.

  - Multiplicative Model

$$Y(t) = T(t) \times S(t) \times C(t) \times I(t)$$

    Assumption: These four components of a time series are not necessarily independent and they can affect one another.

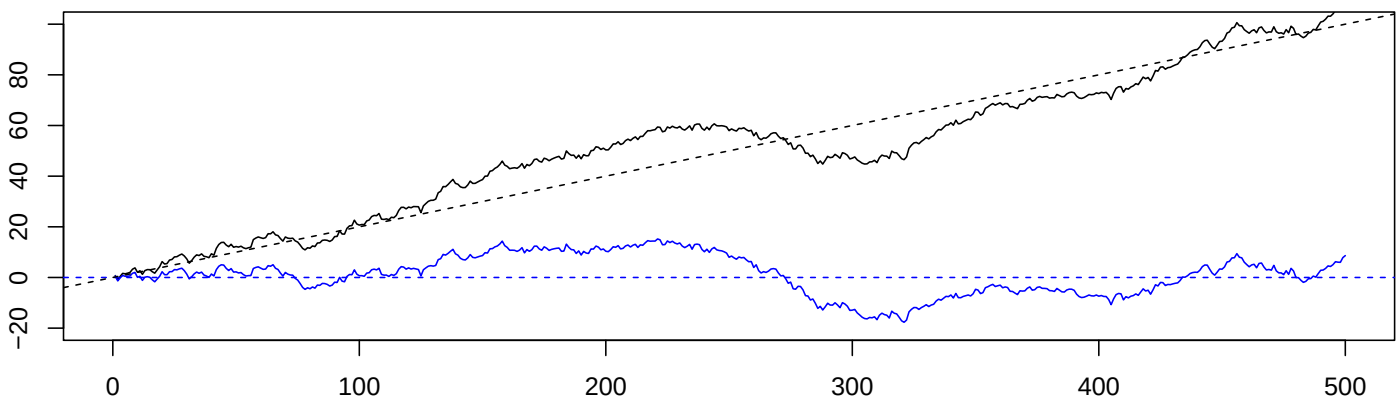
Time Series Example: White Noise

- **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_w^2, denoted as $w_t \sim wn(0, \sigma_w^2)$.
- **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_w^2), denoted as $w_t \sim$ iid $\mathcal{N}(0, \sigma_w^2)$.
- White noise time series is of great interest because if the stochastic behavior of all time series could be explained in terms of the white noise model, then classical statistical methods would suffice.

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.

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



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 will focus 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 $t = 0, \pm 1, \pm 2, \ldots$
  - Note that 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.

Measures of Dependence

- A complete description of a time series, observed as a collection of n random variables at arbitrary time points $t_1, t_2, \ldots, t_n$, for any positive integer n, is provided by the joint distribution function, evaluated as the probability that the values of the series are jointly less than the n constants, $c_1, c_2, \ldots, c_n$; i.e.,

$$F_{t_1,t_2,\ldots,t_n}(c_1, c_2, \ldots, c_n) = Pr(X_{t_1} \leq c_1, X_{t_2} \leq c_2, \ldots, X_{t_n} \leq c_n).$$

- Unfortunately, these multidimensional distribution functions cannot usually be written easily.
- Therefore some informative descriptive measures can be useful, such as mean function and more.

Measurement Functions

- **Mean function**
  - The mean function is defined as

$$\mu_t = \mu_{Xt} = E[X_t] = \int_{-\infty}^{\infty} x f_t(x) dx,$$

    provided it exists, where E denotes the usual expected value operator.
- Clearly for white noise series, $\mu_{w_t} = E[w_t] = 0$ for all t.
- For random walk with drift ($\delta \neq 0$),

$$\mu_{X_t} = E[X_t] = \delta t + \sum_{i=1}^{t} E[w_i] = \delta t$$

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 choppy series tend to have autocovariance functions that are nearly zero for large separations.

Autocorrelation for Time Series

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

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

  - According to Cauchy-Schwarz inequality

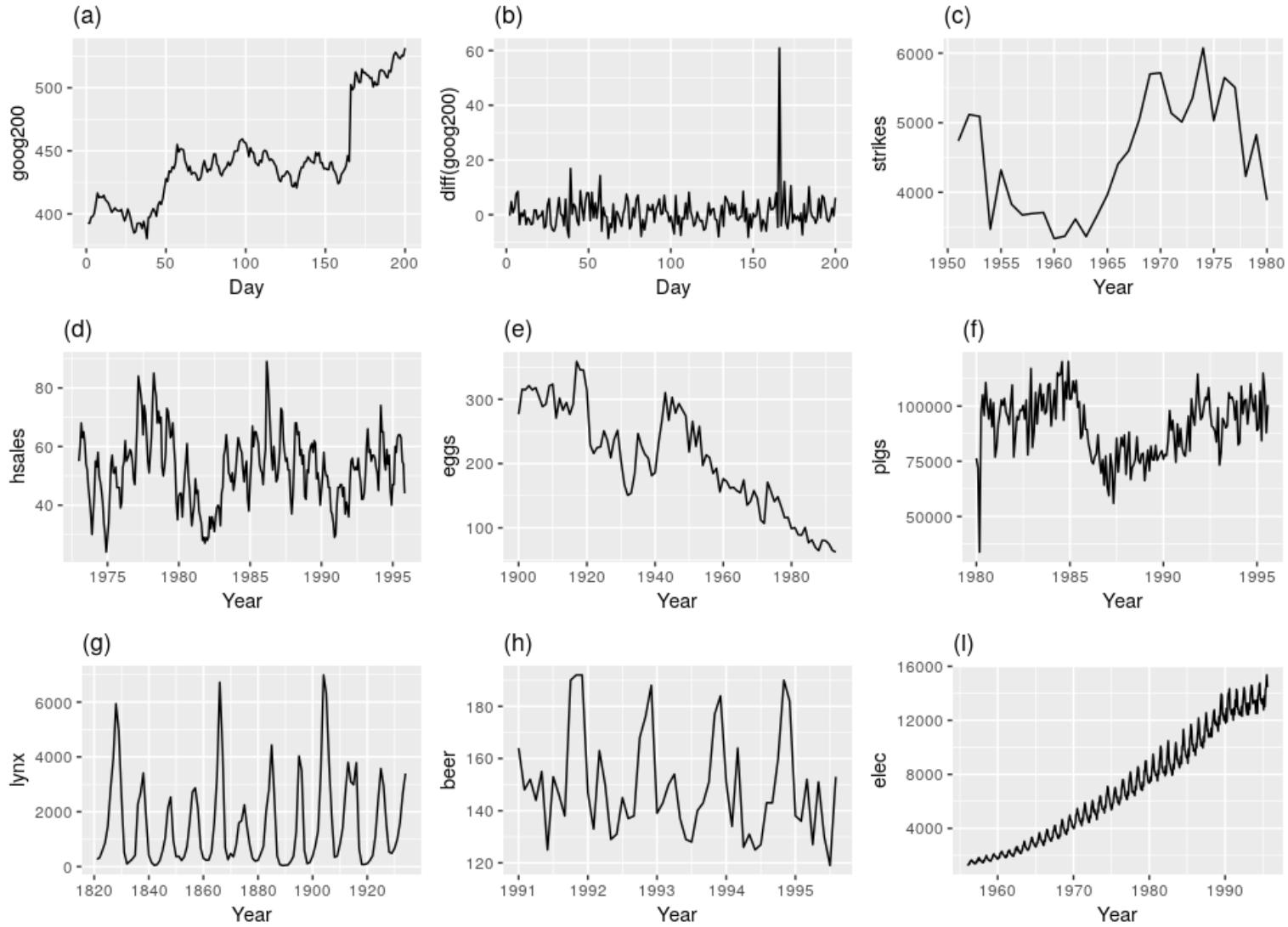
$$|\gamma(s,t)|^2 \leq \gamma(s,s)\gamma(t,t),$$

    it's easy to show that $-1 \leq \rho(s,t) \leq 1$.
- 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.

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 a little bit 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.

Which of these are stationary?



Strict Stationarity

- There are two types of stationarity, i.e. strictly stationary and weakly stationary.
- **Strict Stationarity**
  - The time series $\{X_t, t \in \mathbb{Z}\}$ is said to be strictly stationary if the joint distribution of $(X_{t_1}, X_{t_2}, \ldots, X_{t_k})$ is the same as that of $(X_{t_1+h}, X_{t_2+h}, \ldots, X_{t_k+h})$.
  - In other words, strict stationarity means that the joint distribution only depends on the "difference" h, not the time $(t_1, t_2, \ldots, t_k)$.
- However in most applications this stationary condition is too strong.

Weak Stationarity

- **Weak Stationarity**
  - The time series $\{X_t, t \in \mathbb{Z}\}$ is said to be weakly stationary if
    1. $E[X_t^2] < \infty, \quad \forall t \in \mathbb{Z}$;
    2. $E[X_t] = \mu, \quad \forall t \in \mathbb{Z}$;
    3. $\gamma_X(s,t) = \gamma_X(s+h, t+h), \quad \forall s, t, h \in \mathbb{Z}$.
  - In other words, a weakly stationary time series $\{X_t\}$ must have three features: finite variation, constant first moment, and that the second moment $\gamma_X(s,t)$ only depends on $|t-s|$ and not depends on s or t.
- Usually the term *stationary* means weakly stationary, and when people want to emphasize a process is stationary in the strict sense, they will use strictly stationary.

Remarks on Stationarity

- Strict stationarity does not assume finite variance thus strictly stationary does NOT necessarily imply weakly stationary.
  - Processes like i.i.d Cauchy is strictly stationary but not weakly stationary.
- A nonlinear function of a strictly stationary time series is still strictly stationary, but this is not true for weakly stationary.
- Weak stationarity usually does not imply strict stationarity as higher moments of the process may depend on time t.
- If time series $\{X_t\}$ is Gaussian (i.e. the distribution functions of $\{X_t\}$ are all multivariate Gaussian), then weakly stationary also implies strictly stationary. This is because a multivariate Gaussian distribution is fully characterized by its first two moments.

Autocorrelation for Stationary Time Series

- Recall that 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**

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

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 $h = 1, 2, \ldots$ is defined as

$$\phi_{11} = \text{corr}(X_{t+1}, X_t) = \rho_1$$
$$\phi_{hh} = \text{corr}(X_{t+h} - \hat{X}_{t+h}, X_t - \hat{X}_t), \quad h \geq 2$$

    where $\hat{X}_{t+h}$ and $\hat{X}_t$ is defined as:

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

  - If X_t is Gaussian, then ϕ_{hh} is actually conditional correlation

$$\phi_{hh} = \text{corr}(X_t, X_{t+h} | X_{t+1}, X_{t+2}, \ldots, X_{t+h-1})$$

ARIMA Models

- ARIMA is an acronym that stands for **A**uto-**R**egressive **I**ntegrated **M**oving **A**verage. Specifically,
  - **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 that **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.

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.
- In other words, ARIMA model can be configured to perform the function of an ARMA model, and even a simple AR, I, or MA model.

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, $w_t \sim wn(0, \sigma_w^2)$, and $\phi_1, \phi_2, \ldots, \phi_p$ ($\phi_p \neq 0$) are model parameters. The hyperparameter p represents the length of the "direct look back" in the series.

Backshift Operator

- Before we dive deeper into the AR process, we need some new notations to simplify the representations.
- **Backshift Operator**
  - The backshift operator is defined as

$$BX_t = X_{t-1}.$$

It can be extended, $B^2X_t = B(BX_t) = B(X_{t-1}) = X_{t-2}$, and so on. Thus,

$$B^kX_t = X_{t-k}$$

- We can also define an inverse operator (*forward-shift operator*) by enforcing $B^{-1}B = 1$, such that

$$X_t = B^{-1}BX_t = B^{-1}X_{t-1}.$$

Autoregressive Operator of AR Process

- Recall the definition for $AR(p)$ process:

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

By using the backshift operator we can rewrite it as:

$$X_t - \phi_1 X_{t-1} - \phi_2 X_{t-2} - \cdots - \phi_p X_{t-p} = w_t$$
$$(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p) X_t = w_t$$

- The autoregressive operator is defined as:

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p = 1 - \sum_{j=1}^{p} \phi_j B^j,$$

then the $AR(p)$ can be rewritten more concisely as:

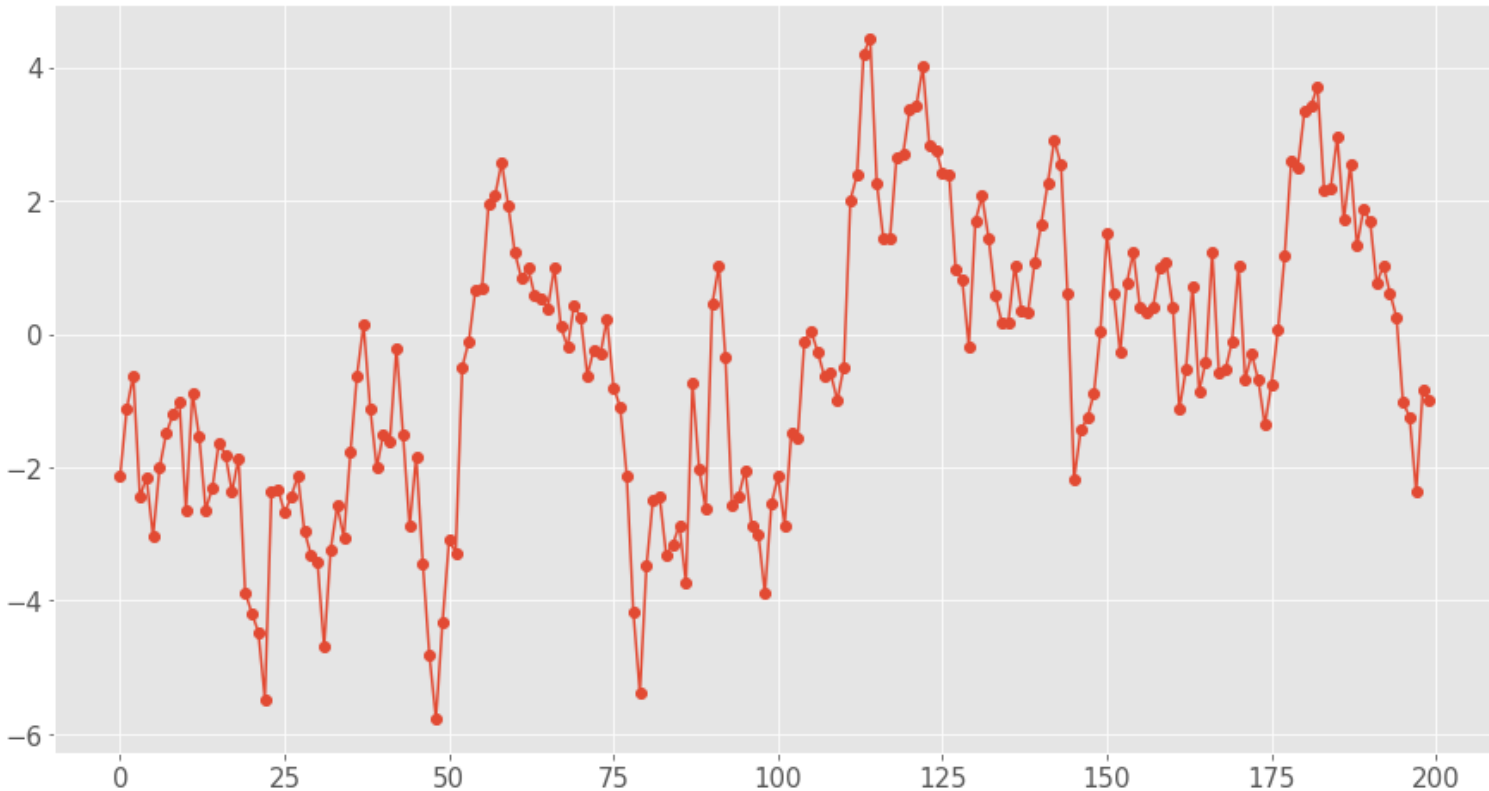
$$\boxed{\phi(B) X_t = w_t}$$

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 $|\phi_1|$ 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).

AR Examples: AR(1) Process

- Simulated $AR(1)$ Process $X_t = 0.9X_{t-1} + w_t$:

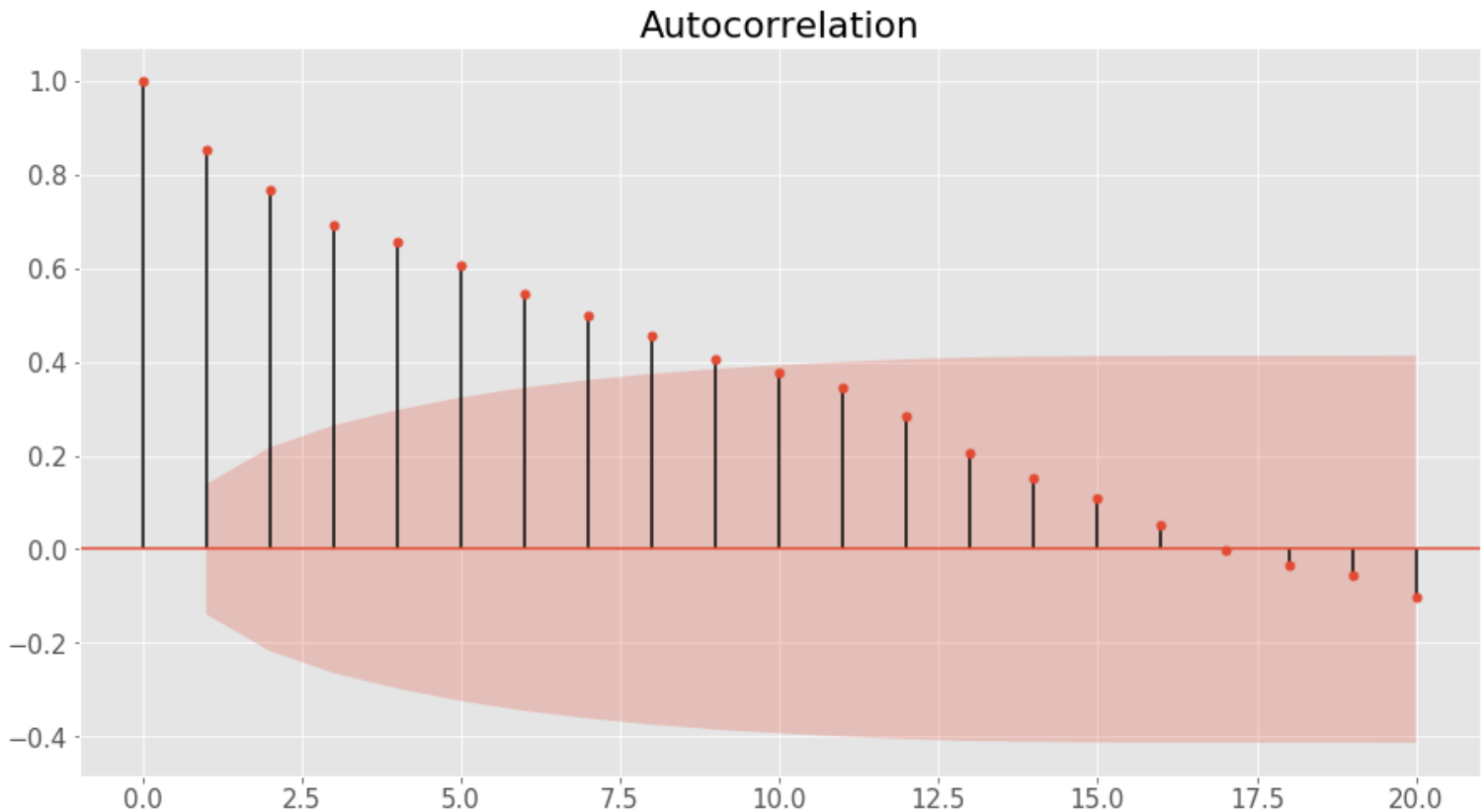


- **Mean** $E[X_t] = 0$
- **Variance** $\text{Var}(X_t) = \dfrac{\sigma_w^2}{(1 - \phi_1^2)}$

AR Examples: AR(1) Process

- **Autocorrelation Function (ACF)**

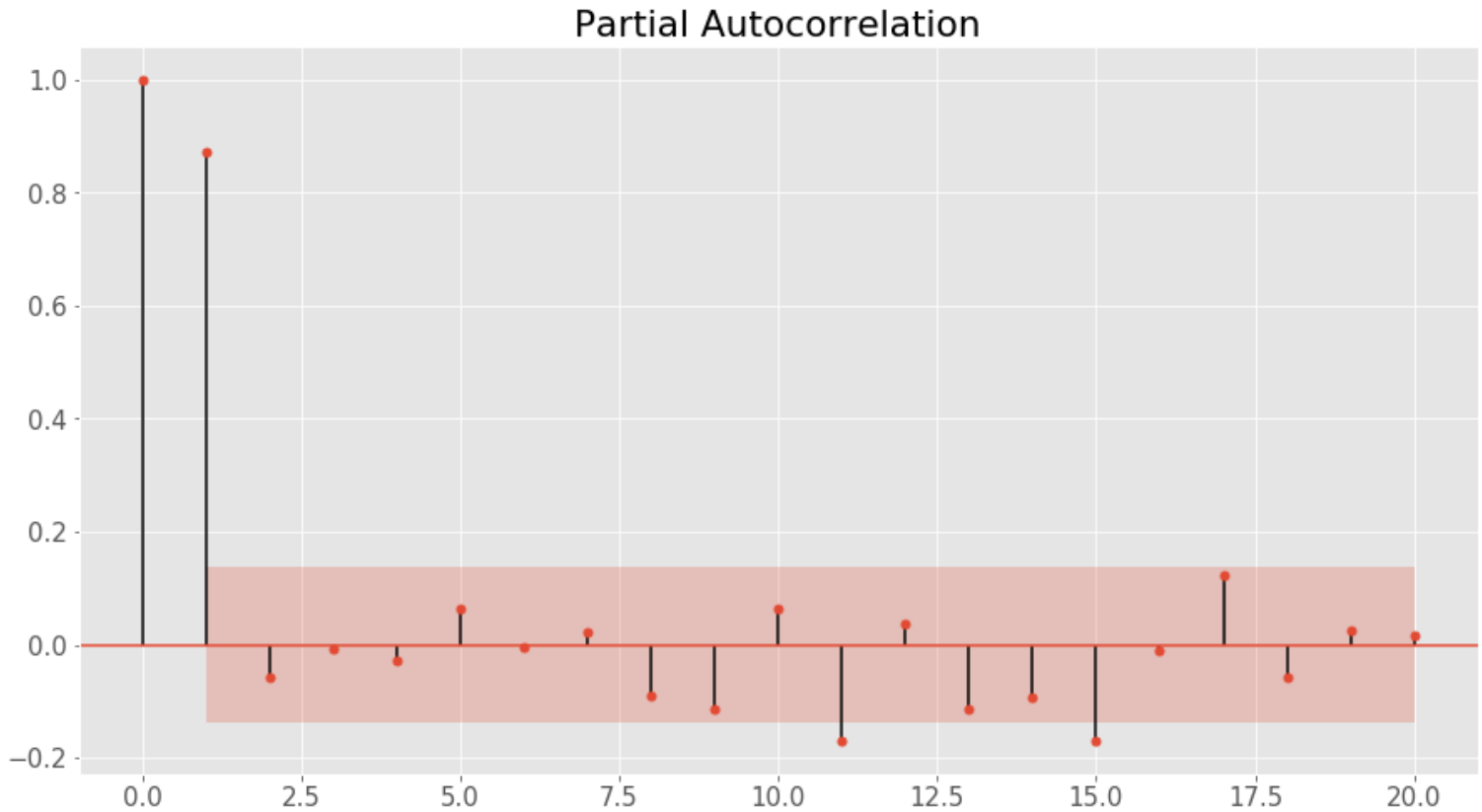
$$\rho_h = \phi_1^h$$



AR Examples: AR(1) Process

- **Partial Autocorrelation Function (PACF)**

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



AR Examples: AR(1) Process

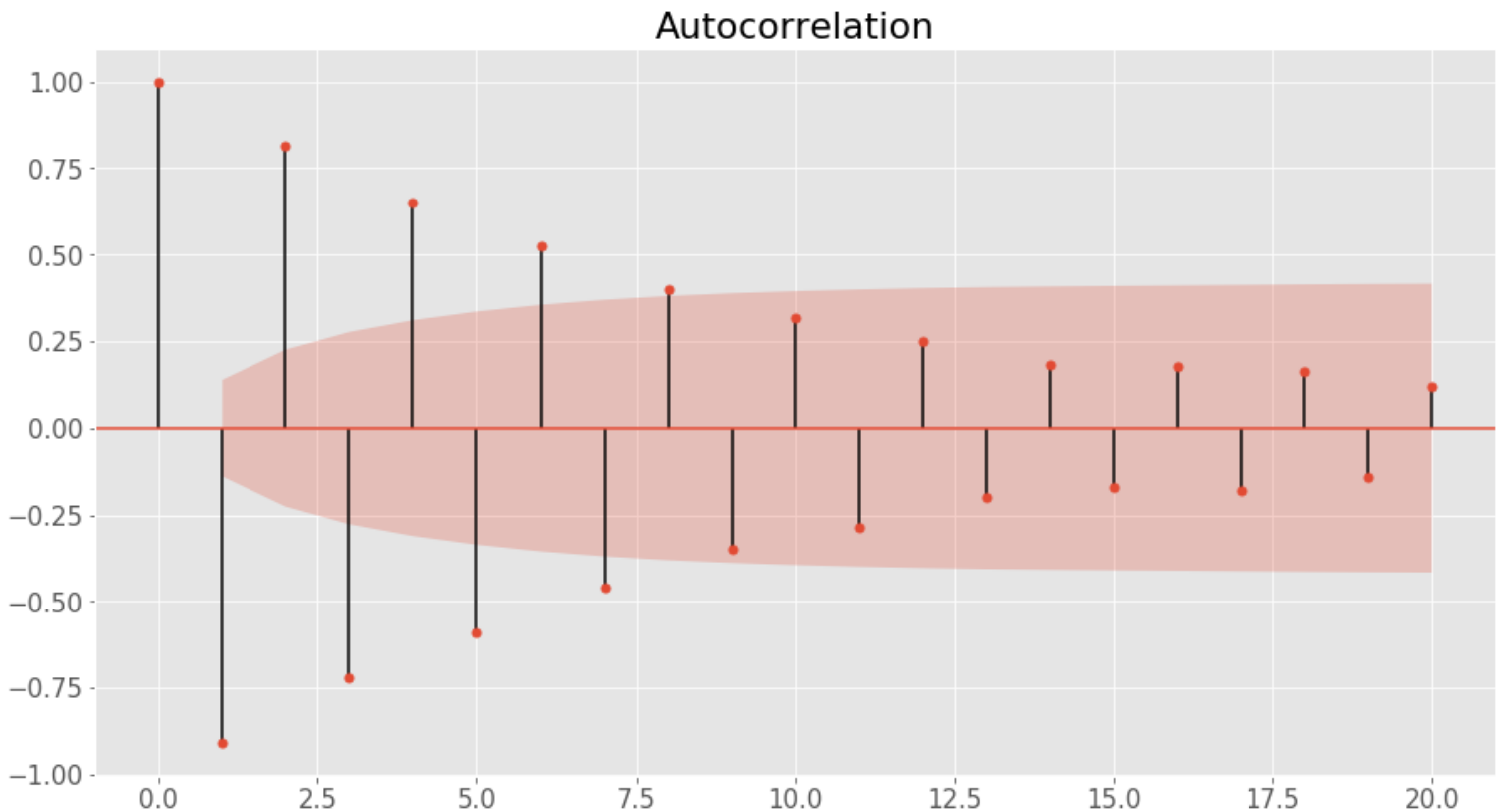
- Simulated $AR(1)$ Process $X_t = -0.9X_{t-1} + w_t$:

- **Mean** $E[X_t] = 0$
- **Variance** $\text{Var}(X_t) = \dfrac{\sigma_w^2}{(1 - \phi_1^2)}$

AR Examples: AR(1) Process

- **Autocorrelation Function (ACF)**

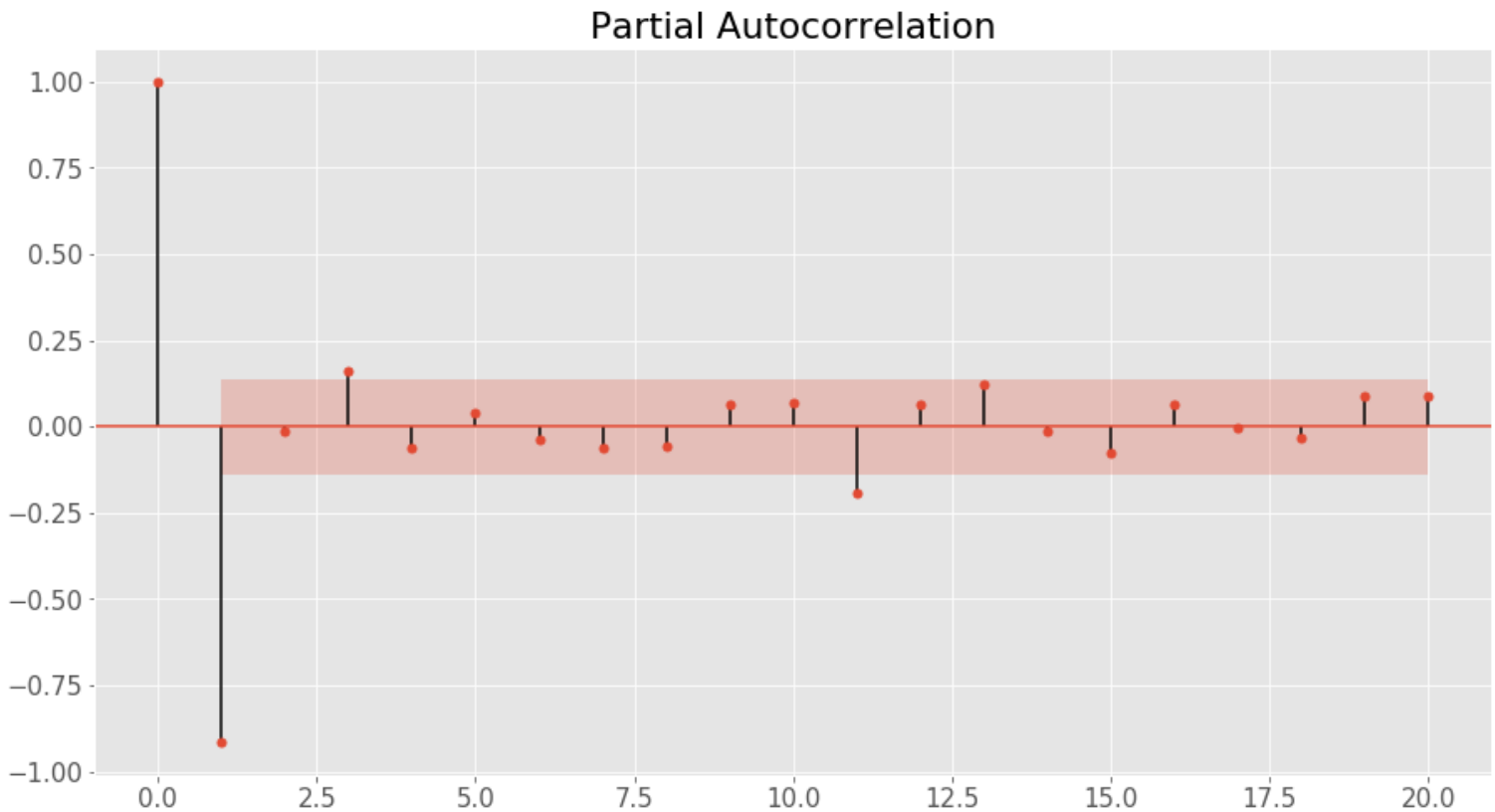
$$\rho_h = \phi_1^h$$



AR Examples: AR(1) Process

- **Partial Autocorrelation Function (PACF)**

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



Stationarity of AR(1)

- We can iteratively expand $AR(1)$ representation as:

$$\begin{aligned} X_t &= \phi_1 X_{t-1} + w_t \\ &= \phi_1(\phi_1 X_{t-2} + w_{t-1}) + w_t = \phi_1^2 X_{t-2} + \phi_1 w_{t-1} + w_t \\ &\vdots \\ &= \phi_1^k X_{t-k} + \sum_{j=0}^{k-1} \phi_1^j w_{t-j} \end{aligned}$$

- Note that if $|\phi_1| < 1$ and $\sup_t \mathrm{Var}(X_t) < \infty$, we have:

$$X_t = \sum_{j=0}^{\infty} \phi_1^j w_{t-j}$$

This representation is called the stationary solution.

AR Problem: Explosive AR Process

- We've seen $AR(1)$: $X_t = \phi_1 X_{t-1} + w_t$ while $|\phi_1| \leq 1$.
- What if $|\phi_1| > 1$? Intuitively the time series will "explode".
- However, technically it still can be stationary, because we expand the representation differently and get:

$$X_t = -\sum_{j=1}^{\infty} \phi_1^{-j} w_{t+j}$$

- But clearly this is not useful because we need the future (w_{t+j}) to predict now (X_t).
- We use the concept of *causality* to describe time series that is not only stationary but also NOT future-dependent.

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.

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)
  - Maximum Likelihood Estimation (MLE) estimator
  - Ordinary Least Squares (OLS) 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.

Moving Average Models (MA)

- The name might be misleading, but moving average models should 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.

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 $w_t \sim wn(0, \sigma_w^2)$, and $\theta_1, \theta_2, \ldots, \theta_q$ $(\theta_q \neq 0)$ 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.

Moving Average Operator

- **Moving Average Operator**
  - Equivalent to autoregressive operator, we define moving average operator as:

$$\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q,$$

  where B stands for backshift operator, thus $B(w_t) = w_{t-1}$.
- Therefore the moving average model can be rewritten as:

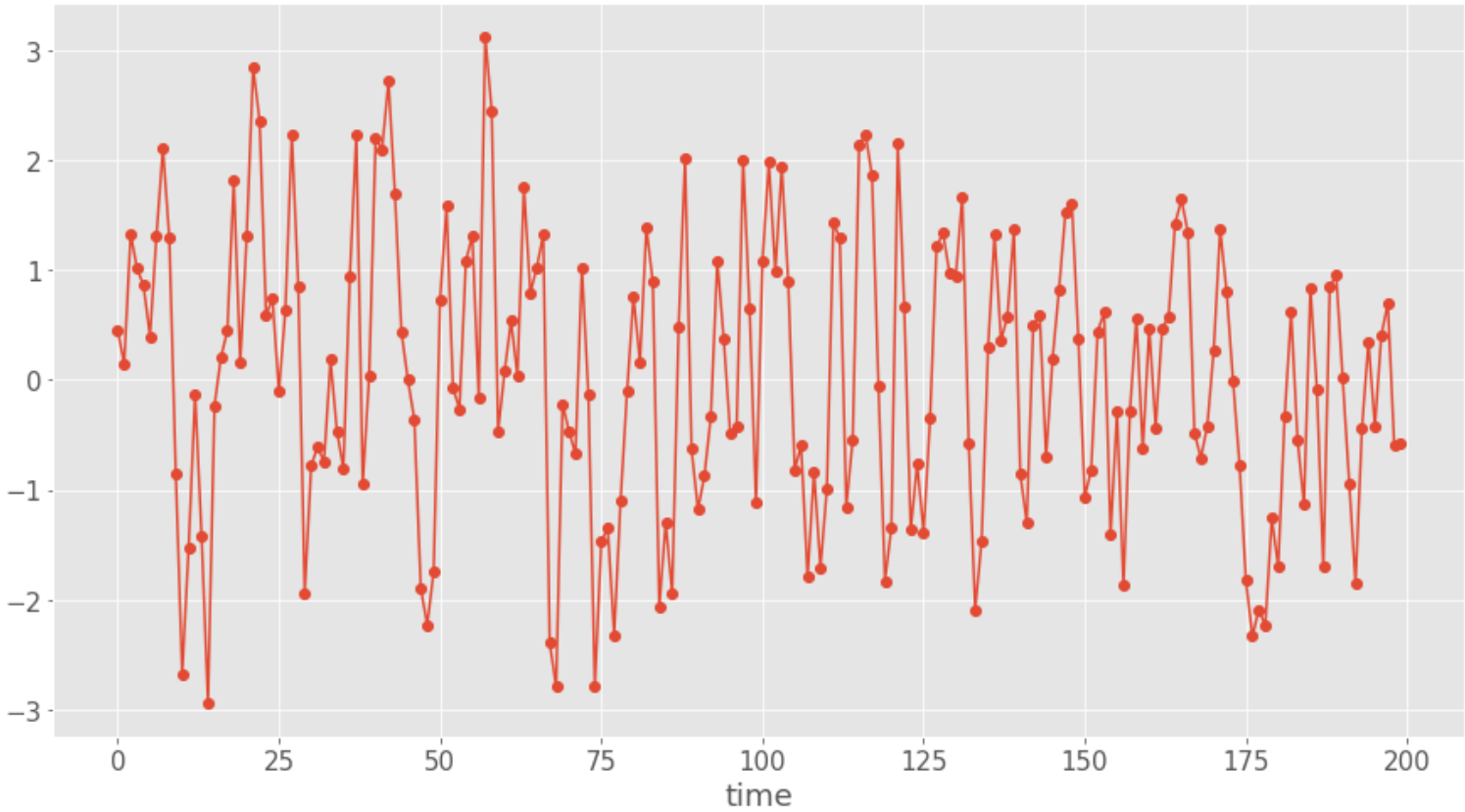
$$X_t = w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + \cdots + \theta_q w_{t-q}$$

$$X_t = (1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q) w_t$$

$$\boxed{X_t = \theta(B) w_t}$$

MA Examples: MA(1) Process

- Simulated $MA(1)$ Process $X_t = w_t + 0.8 \times w_{t-1}$:

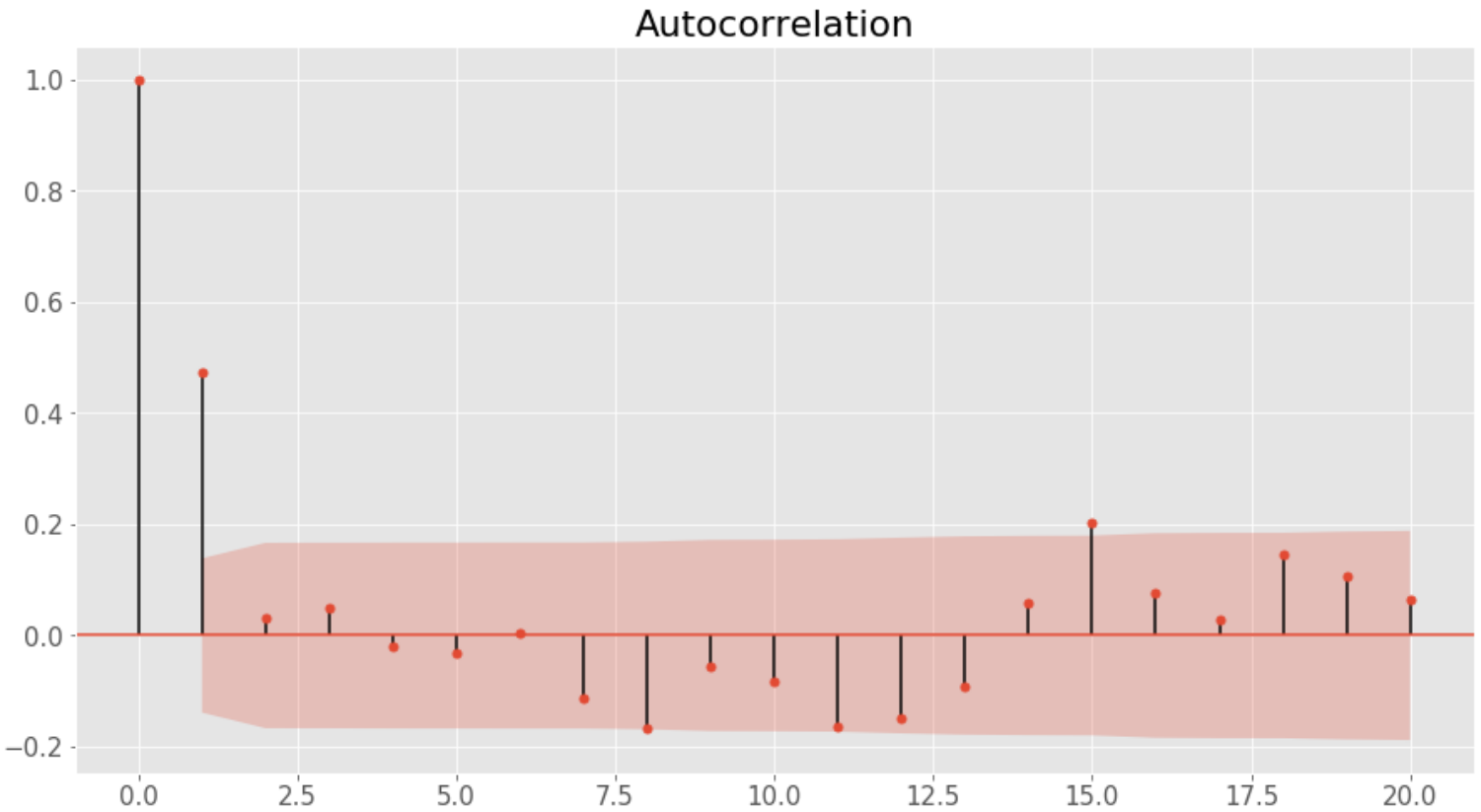


- **Mean** $E[X_t] = 0$
- **Variance** $\mathrm{Var}(X_t) = \sigma_w^2(1 + \theta_1^2)$

MA Examples: MA(1) Process

- **Autocorrelation Function (ACF)**

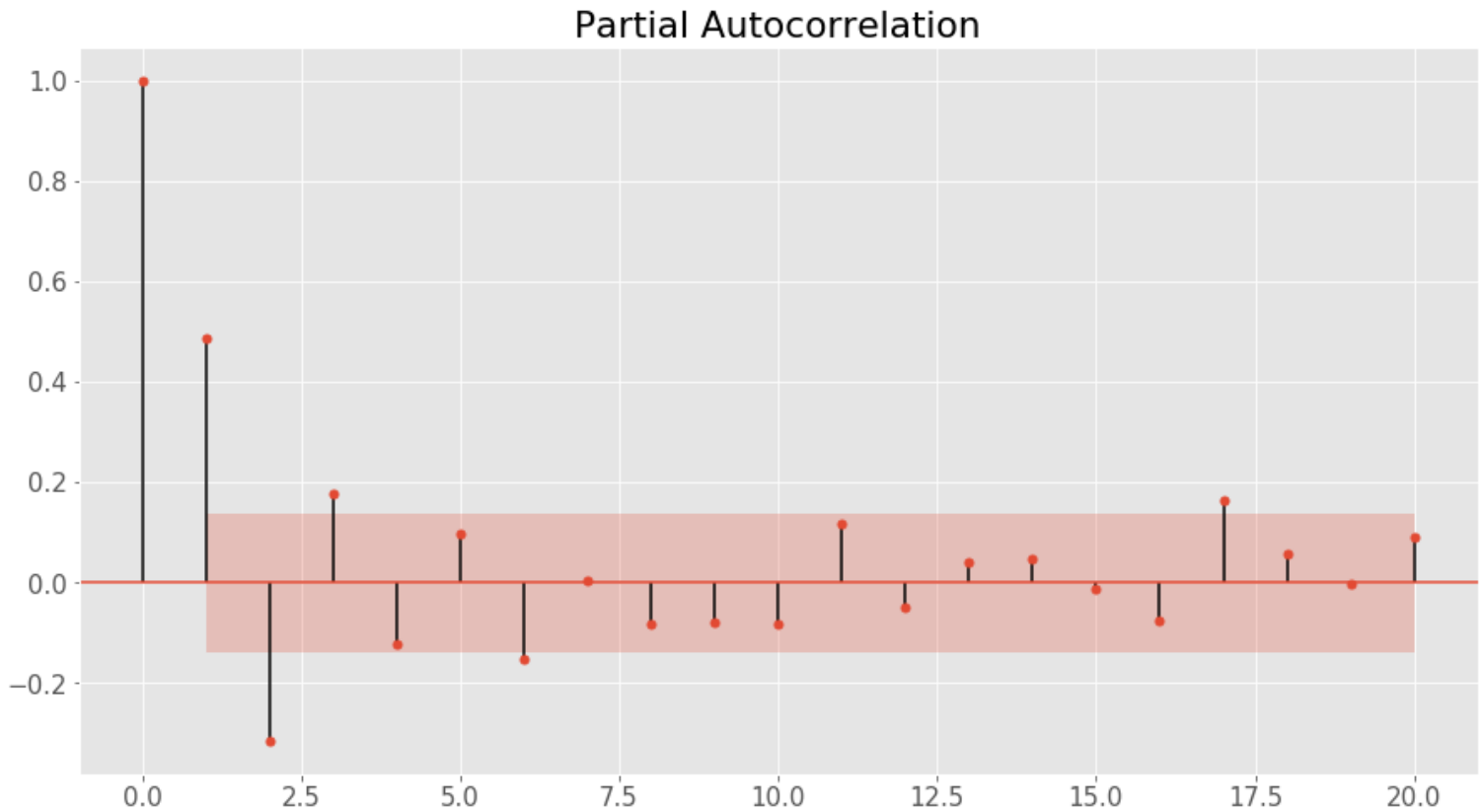
$$\rho_1 = \frac{\theta_1}{1+\theta_1^2} \qquad \rho_h = 0, \forall h \geq 2$$



MA Examples: MA(1) Process

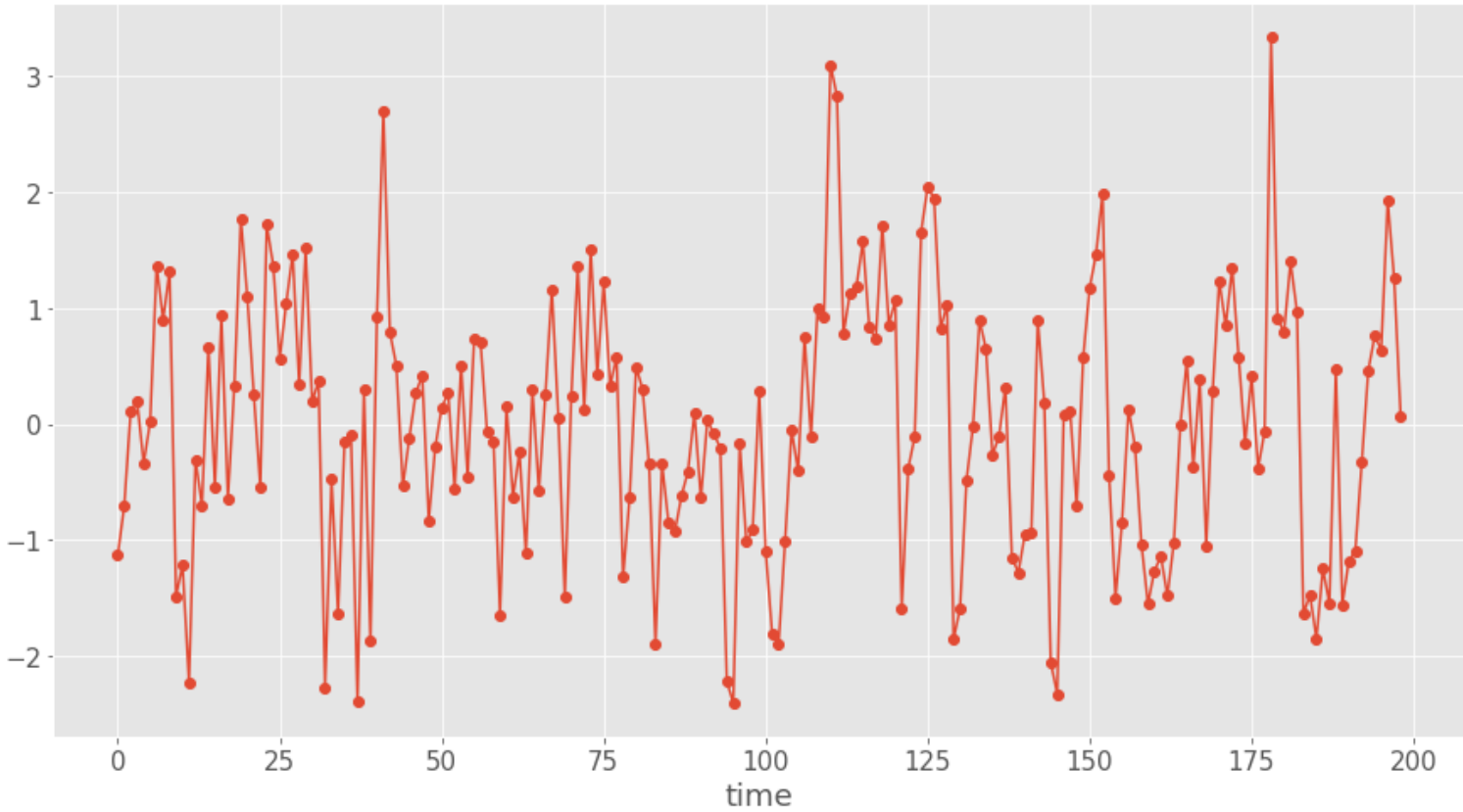
- **Partial Autocorrelation Function (PACF)**

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



MA Examples: MA(2) Process

- Simulated $MA(2)$ Process $X_t = w_t + 0.5 \times w_{t-1} + 0.3 \times w_{t-2}$:

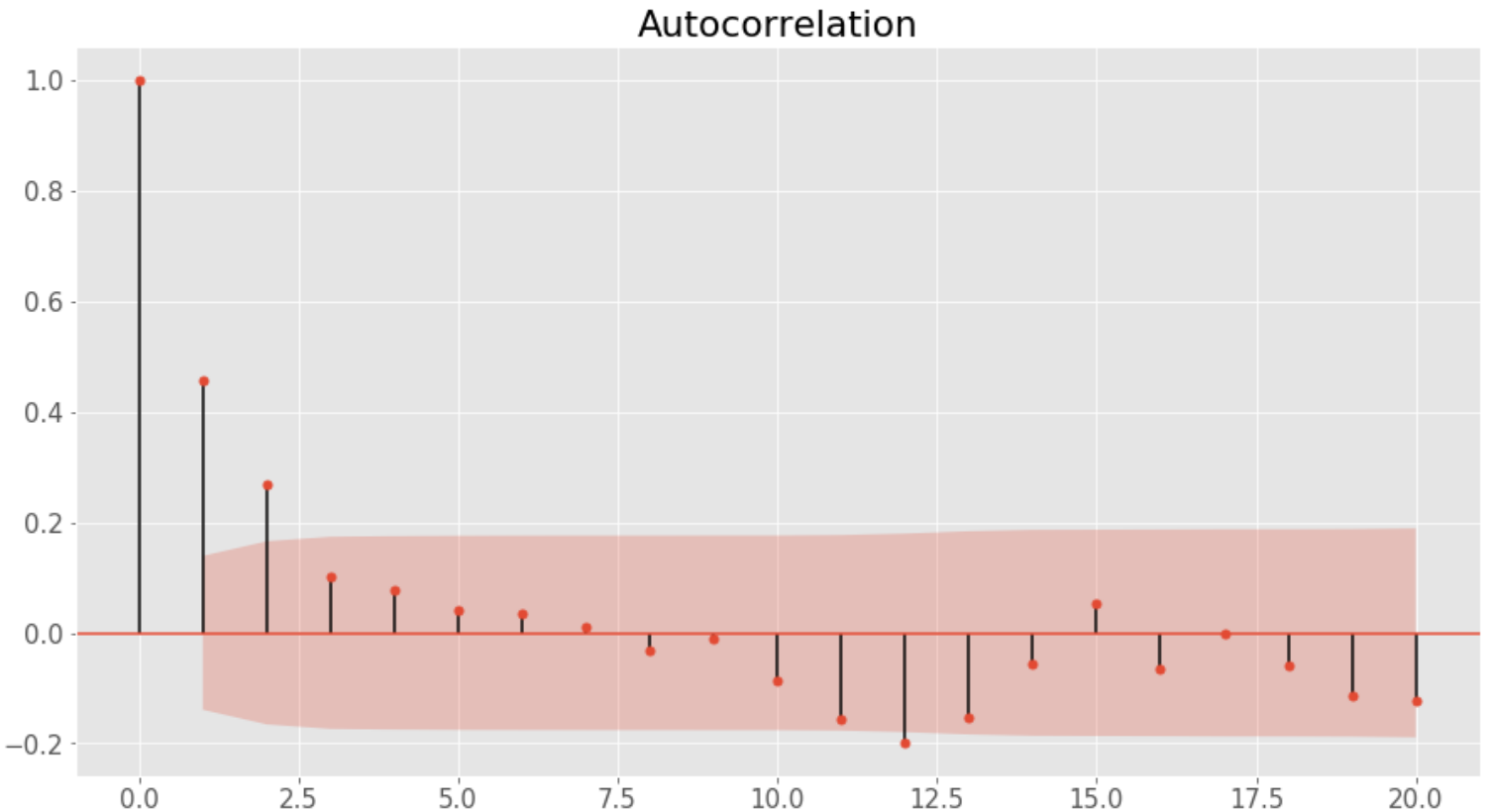


- **Mean** $E[X_t] = 0$
- **Variance** $\text{Var}(X_t) = \sigma_w^2(1 + \theta_1^2 + \theta_2^2)$

MA Examples: MA(2) Process

- **Autocorrelation Function (ACF)**

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



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 $\rho_h = 0$ for all lags $h > q$.
- In other words, 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.

MA Problem: Non-unique MA Process

- Consider the following two $MA(1)$ models:
  - $X_t = w_t + 0.2w_{t-1}, \qquad w_t \sim \text{iid } \mathcal{N}(0, 25),$
  - $Y_t = v_t + 5v_{t-1}, \qquad v_t \sim \text{iid } \mathcal{N}(0, 1)$
- Note that both of them have $\text{Var}(X_t) = 26$ and $\rho_1 = \frac{5}{26}$.
- In fact, these two $MA(1)$ processes are essentially the same. However, since we can only observe X_t (Y_t) but not noise terms w_t (v_t), we cannot distinguish them.
- Conventionally, we define the concept *invertibility* and always choose the invertible representation from multiple alternatives.
  - Simply speaking, for $MA(1)$ models the invertibility condition is $|\theta_1| < 1$.

Comparisons between AR and MA

- Recall that we have seen for $AR(1)$ process, if $|\phi_1| < 1$ and $\sup_t \mathrm{Var}(X_t) < \infty$,

$$X_t = \sum_{j=0}^{\infty} \phi_1^j w_{t-j}$$

- In fact, all causal $AR(p)$ processes can be represented as $MA(\infty)$; In other words, **infinite moving average processes are finite autoregressive processes**.
- All invertible $MA(q)$ processes can be represented as $AR(\infty)$. i.e. **finite moving average processes are infinite autoregressive processes**.

MA Models: Parameters Estimation

- A well-known fact is that 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, but we won't get the optimal estimators with Yule-Walker equations.
- 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.

ARMA Models

- Autoregressive and moving average models can be combined together to form ARMA models.
- **Definition**
  - A time series $\{x_t; t = 0, \pm 1, \pm 2, \dots\}$ 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 $\phi_p \neq 0$, $\theta_q \neq 0$, and $\sigma_w^2 > 0$, $w_t \sim wn(0, \sigma_w^2)$.
  - With the help of AR operator and MA operator we defined before, the model can be rewritten more concisely as:

$$\boxed{\phi(B)X_t = \theta(B)w_t}$$

ARMA Problems: Redundant Parameters

- You may have observed that if we multiply a same factor on both sides of the equation, it still holds.

$$\eta(B)\phi(B)X_t = \eta(B)\theta(B)w_t$$

- For example, consider a white noise process $X_t = w_t$ and $\eta(B) = (1 - 0.5B)$:

$$(1 - 0.5B)X_t = (1 - 0.5B)w_t$$
$$X_t = 0.5X_{t-1} - 0.5w_{t-1} + w_t$$

- Now it looks exactly like a $ARMA(1, 1)$ process!
- If we were unaware of parameter redundancy, we might claim the data are correlated when in fact they are not.

Choosing Model Specification

- Recall we have discussed that ACF and PACF can be used for determining ARIMA 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

- Other criterions can be used for choosing q and q too, such as AIC (Akaike Information Criterion), AICc (corrected AIC) and BIC (Bayesian Information Criterion).
- Note that the selection for p and q is not unique.

"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 and a zero-mean stationary series, i.e. $X_t = \mu_t + Y_t$.
- **Strategies**
  - Detrending: Subtracting with an estimate for trend and deal with residuals.

$$\hat{Y}_t = X_t - \hat{\mu}_t$$

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

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

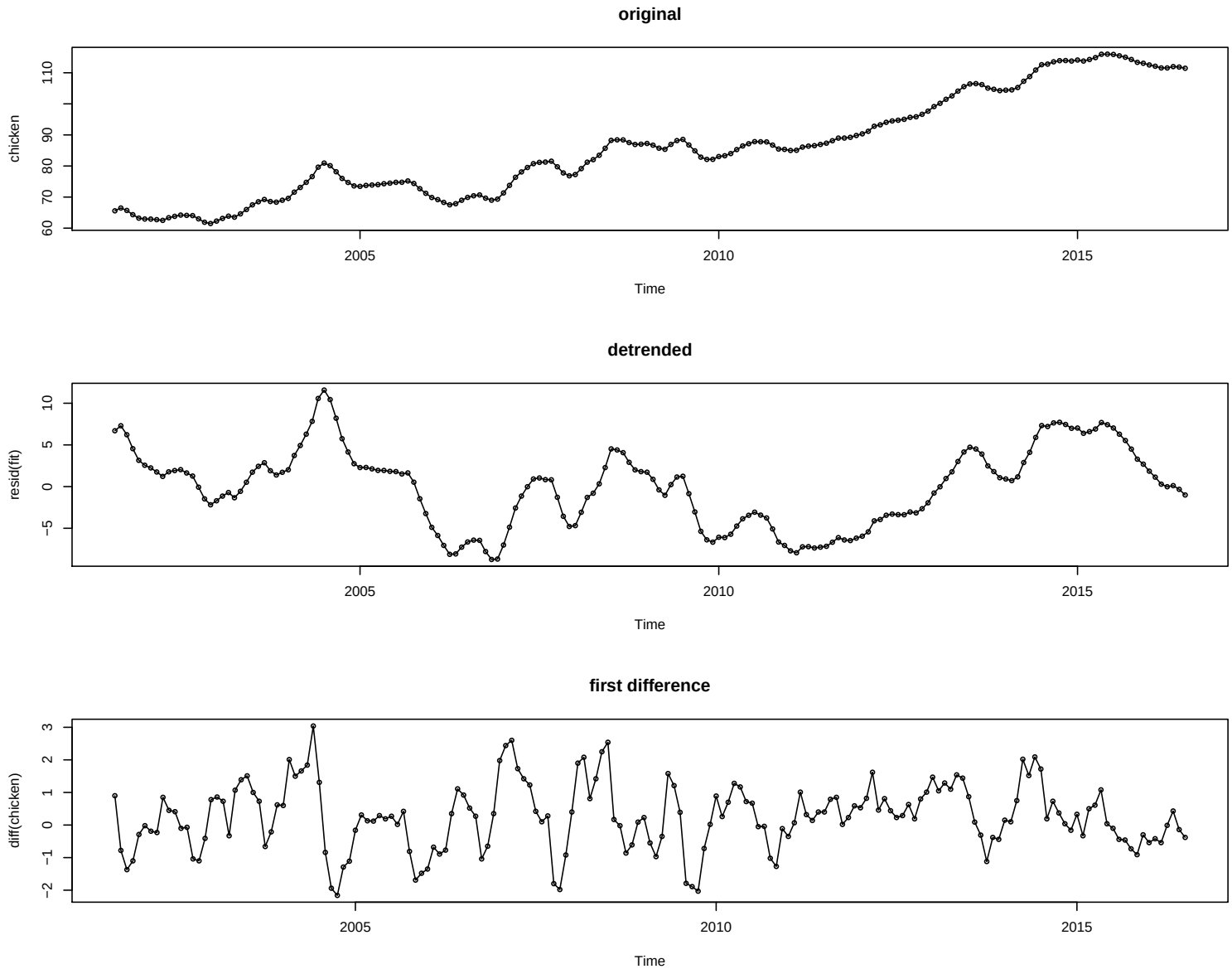
∇ is defined as the first difference and it can be extended to higher orders.

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.
- Recall the backshift operator $X_t = BX_{t-1}$:

$$\begin{aligned}
\nabla X_t &= X_t - X_{t-1} = X_t - BX_t = (1-B)X_t \\
\nabla^2 X_t &= \nabla(\nabla X_t) = \nabla(X_t - X_{t-1}) \\
&= (X_t - X_{t-1}) - (X_{t-1} - X_{t-2}) \\
&= X_t - 2X_{t-1} + X_{t-2} = X_t - 2BX_t + B^2 X_t \\
&= (1 - 2B + B^2)X_t = (1-B)^2 X_t
\end{aligned}$$

Detrending vs. Differencing



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 $ARMA(p, q)$.
  - In general, $ARIMA(p, d, q)$ model can be written as:

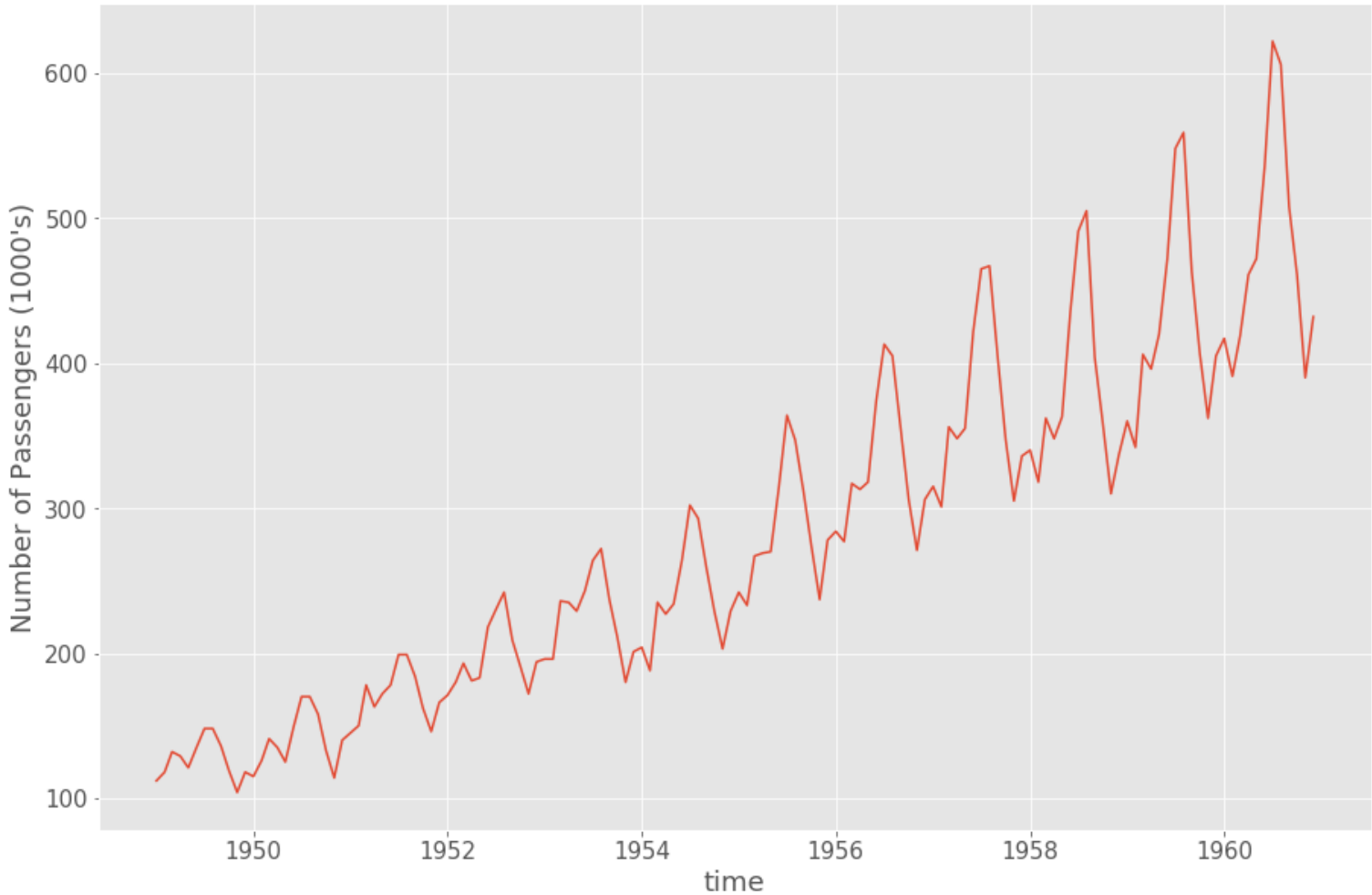
$$\phi(B)(1 - B)^d X_t = \theta(B) w_t$$

    .

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.
- Let's go through a concrete example together for this procedure.

Air Passenger Data

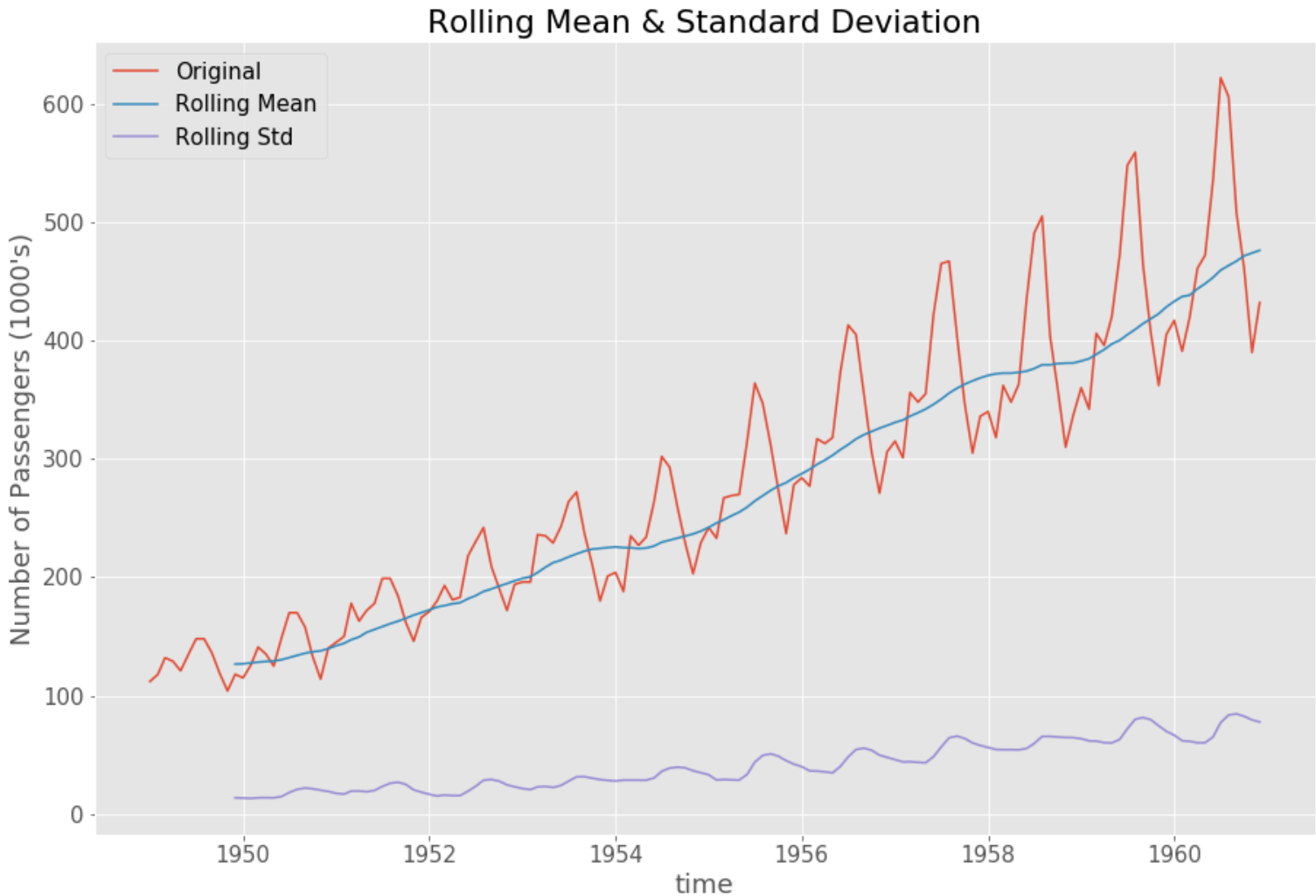


Monthly totals of a US airline passengers, from 1949 to 1960

Model Identification

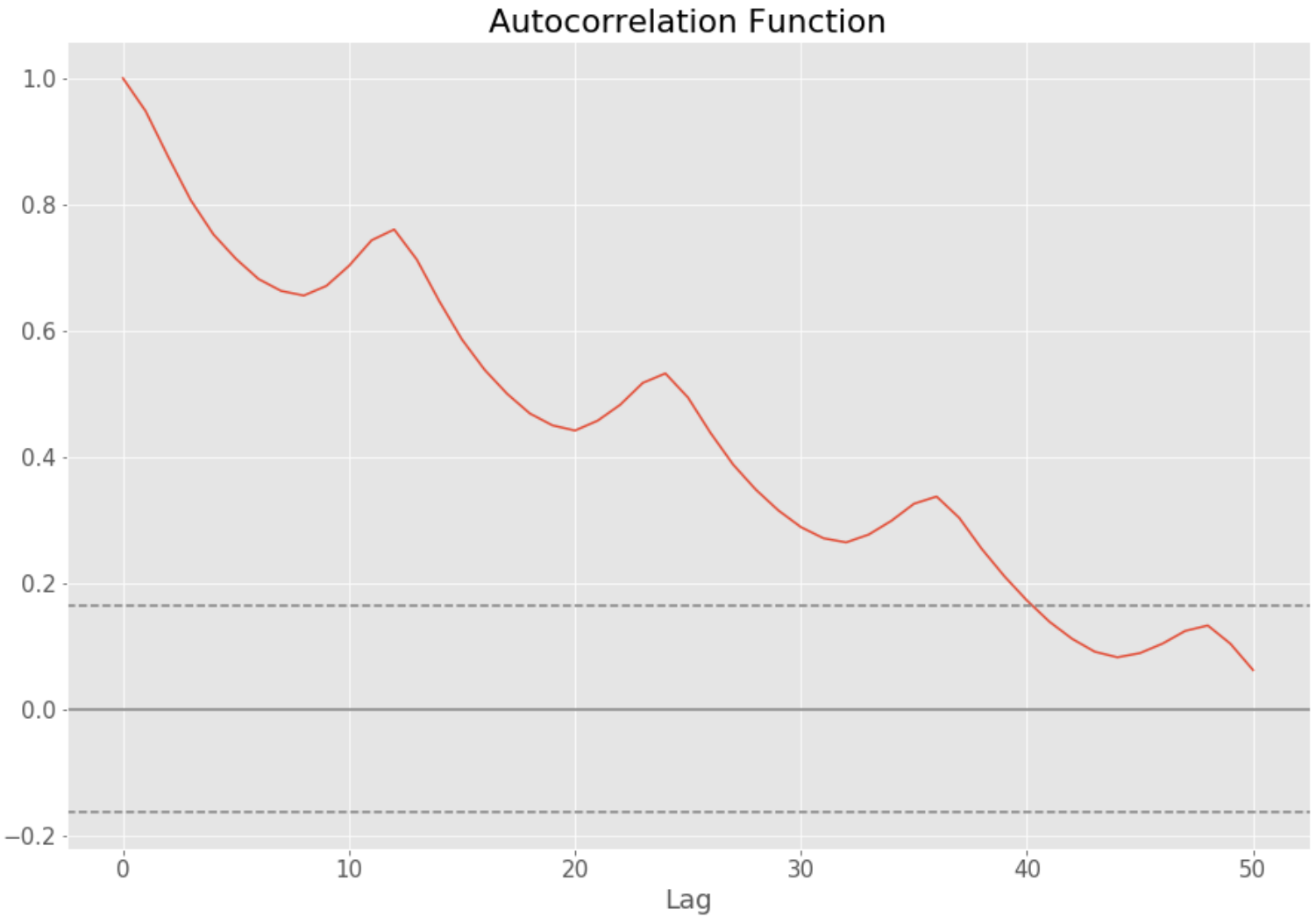
- As with any data analysis, we should construct a time plot of the data, and inspect the graph for any anomalies.
- The most important thing in this phase is to determine if the time series is stationary and if there is any significant seasonality that needs to be handled.
- **Test Stationarity**
  - 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.
  - A well-established statistical test called augmented Dickey-Fuller Test can help. The null hypothesis is *the time series is non-stationary*.

Stationarity Test: Rolling Statistics



Rolling statistics with sliding window of 12 months

Stationarity Test: ACF Plot



Autocorrelation with varying lags

Stationarity Test: ADF Test

- Results of Augmented Dickey-Fuller Test

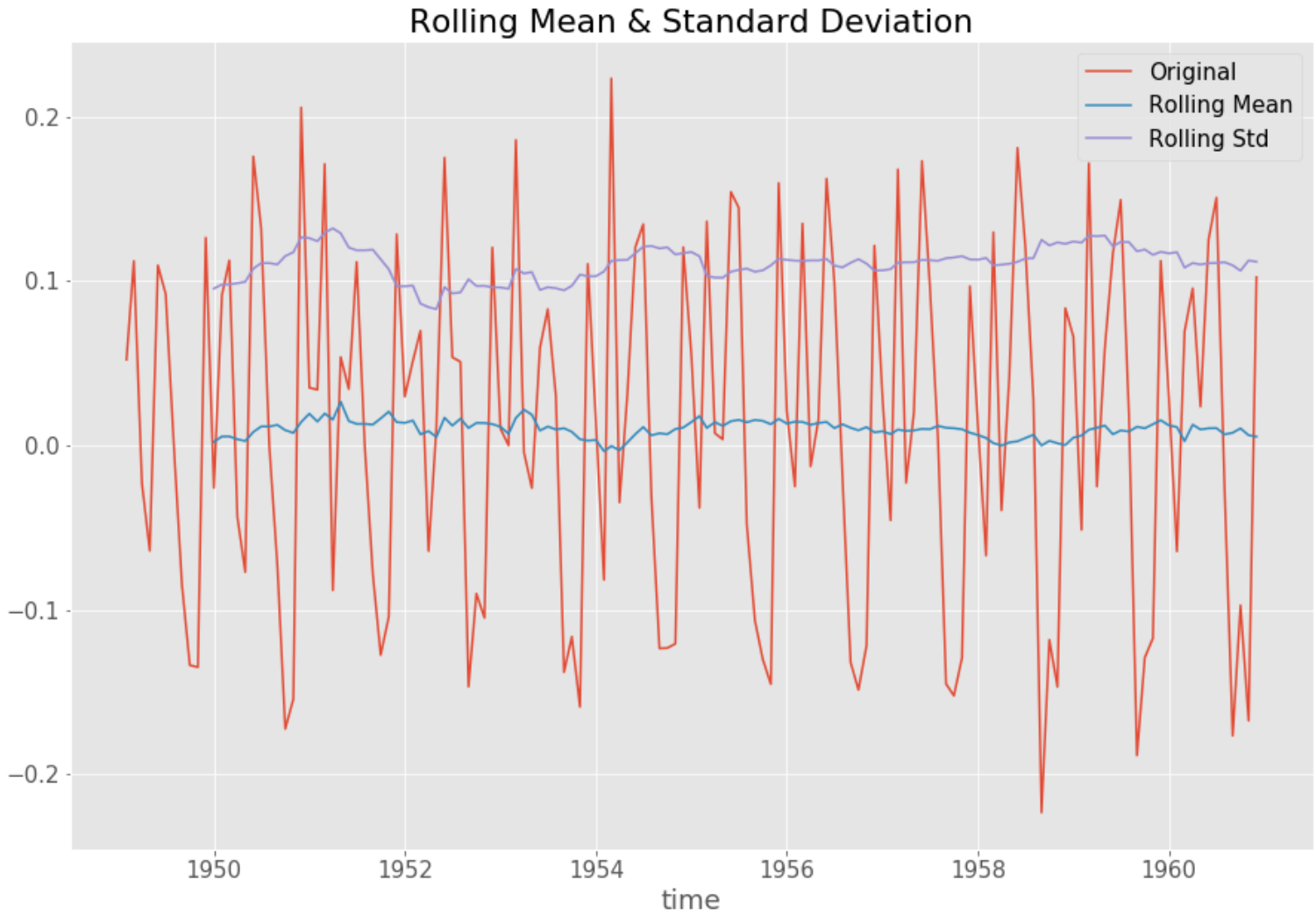
Item	Value
Test Statistic	0.815369
p-value	0.991880
#Lags Used	13.000000
Number of Observations Used	130.000000
Critical Value (1%)	-3.481682
Critical Value (5%)	-2.884042
Critical Value (10%)	-2.578770

- The test statistic is a negative number.
- The more negative it is, the stronger the rejection of the null hypothesis.

Stationarize Time Series

- As all previous methods show that the initial time series is non-stationary, it's necessary to perform transformations to make it stationary for ARMA modeling.
  - 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.

Stationarized Time Series: ACF Plot



First order differencing over logarithm of passengers

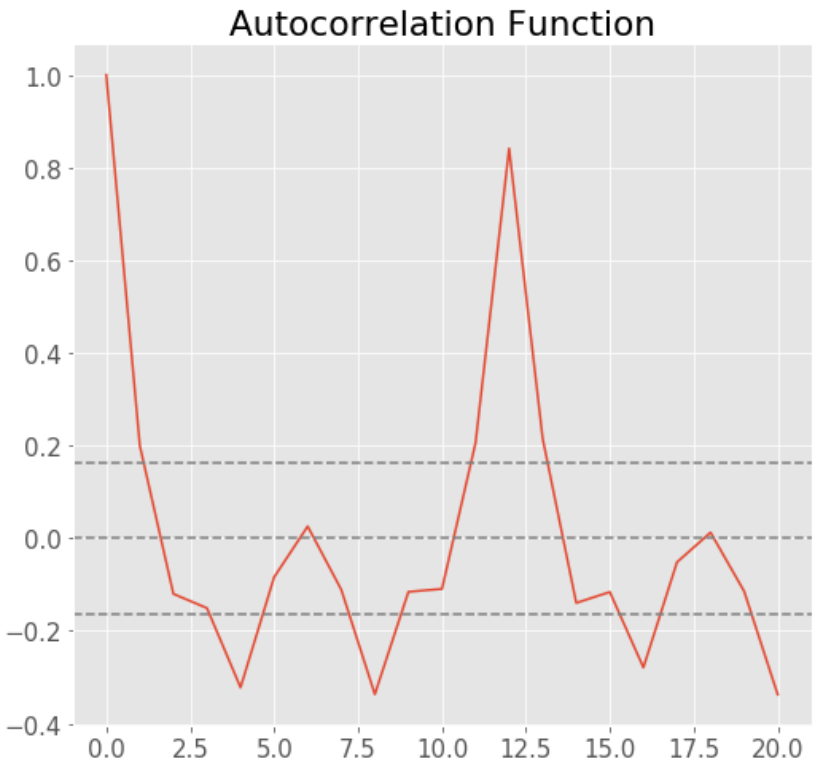
Stationarized Time Series: ADF Test

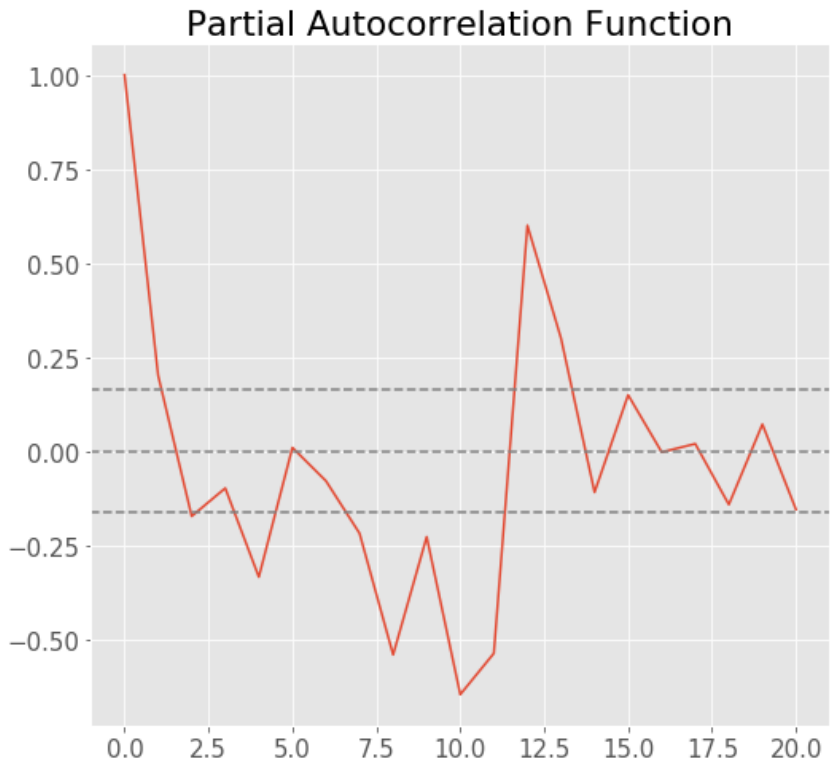
- Results of Augmented Dickey-Fuller Test

Item	Value
Test Statistic	-2.717131
p-value	0.071121
#Lags Used	14.000000
Number of Observations Used	128.000000
Critical Value (1%)	-3.482501
Critical Value (5%)	-2.884398
Critical Value (10%)	-2.578960

- From the ACF plot, we can see that the mean and std variations have much smaller variations with time.
- Also, the ADF test statistic is less than the 10% critical value, indicating the time series is stationary with 90% confidence.

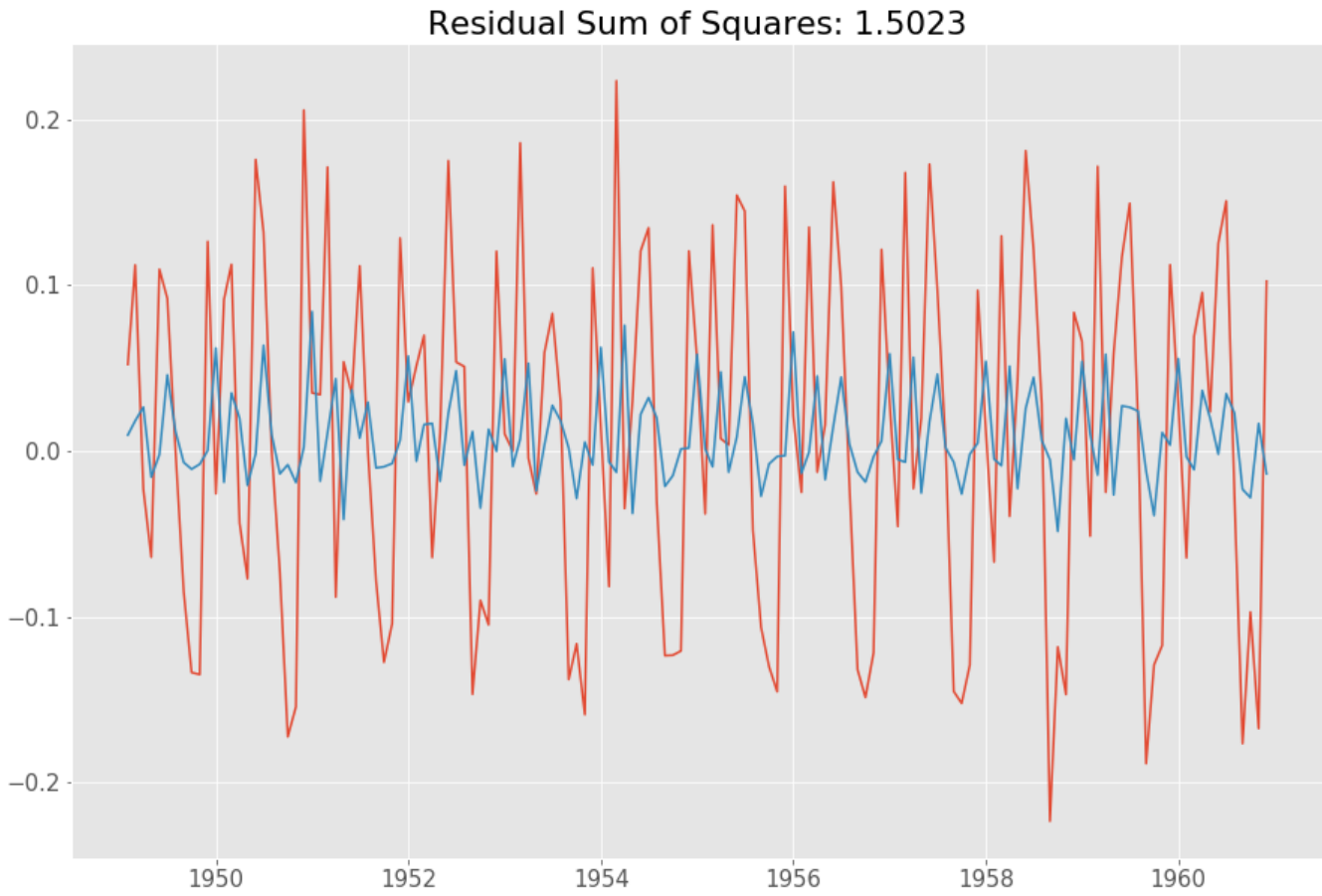
Choosing Model Specification





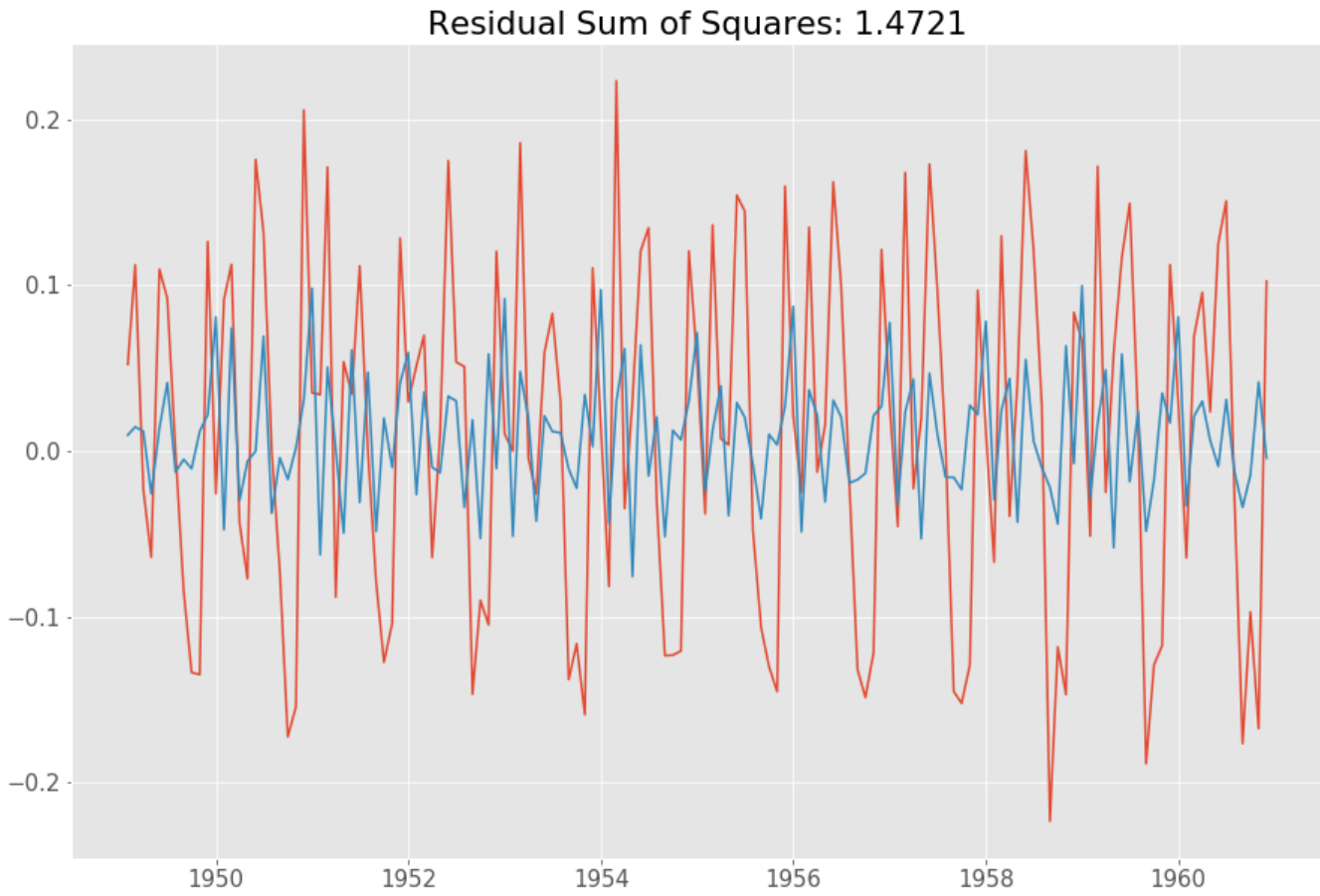
- Firstly we notice an obvious peak at $h = 12$, because for simplicity we didn't model the cyclical effect.
- It seems $p = 2$, $q = 2$ is a reasonable choice. Let's see three models, $AR(2)$, $MA(2)$ and $ARMA(2, 2)$.

$AR(2)$: Predicted on Residuals

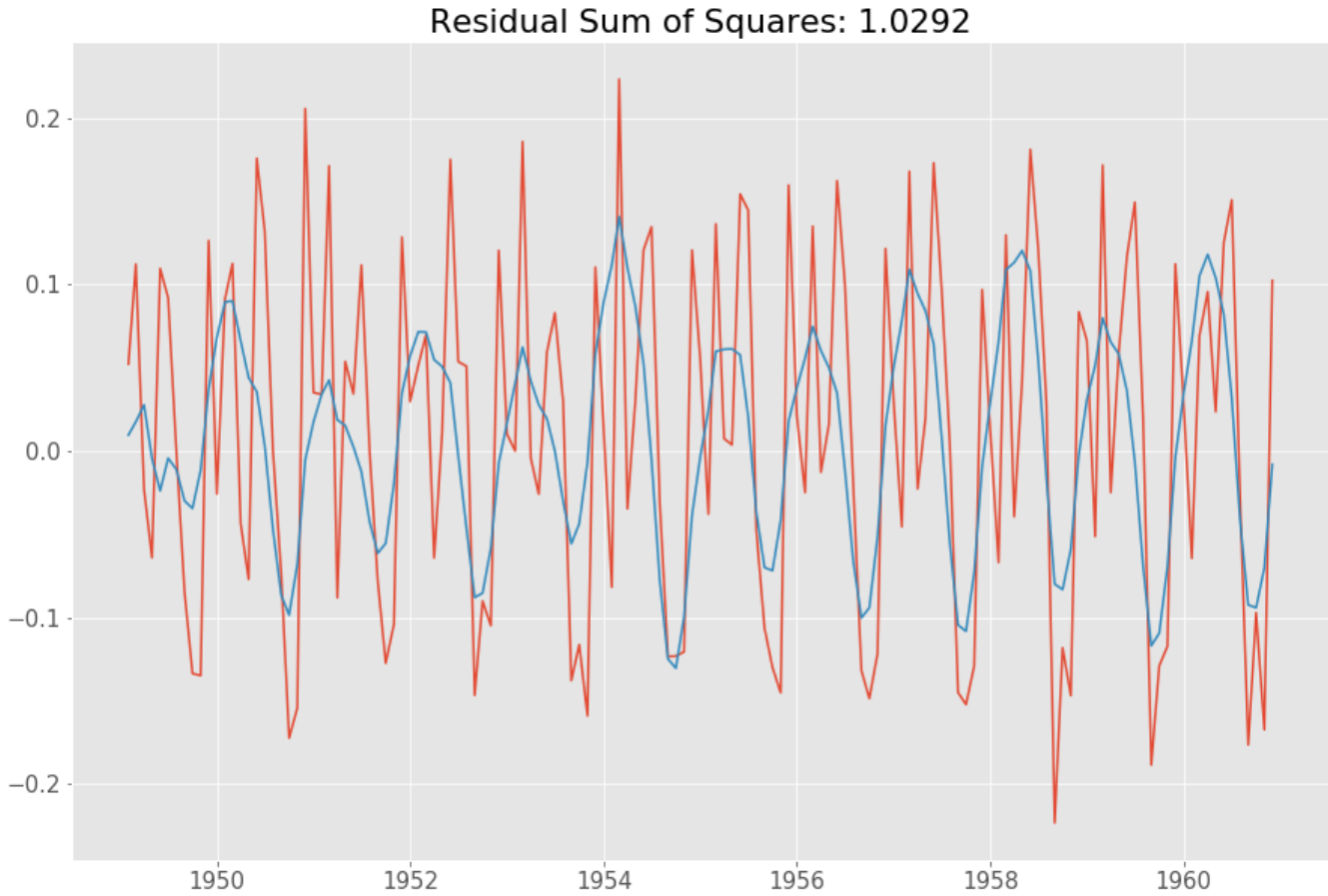


- RSS is a measure of the discrepancy between the data and the estimation model.
  - A small RSS indicates a tight fit of the model to the data.

$MA(2)$: Predicted on Residuals



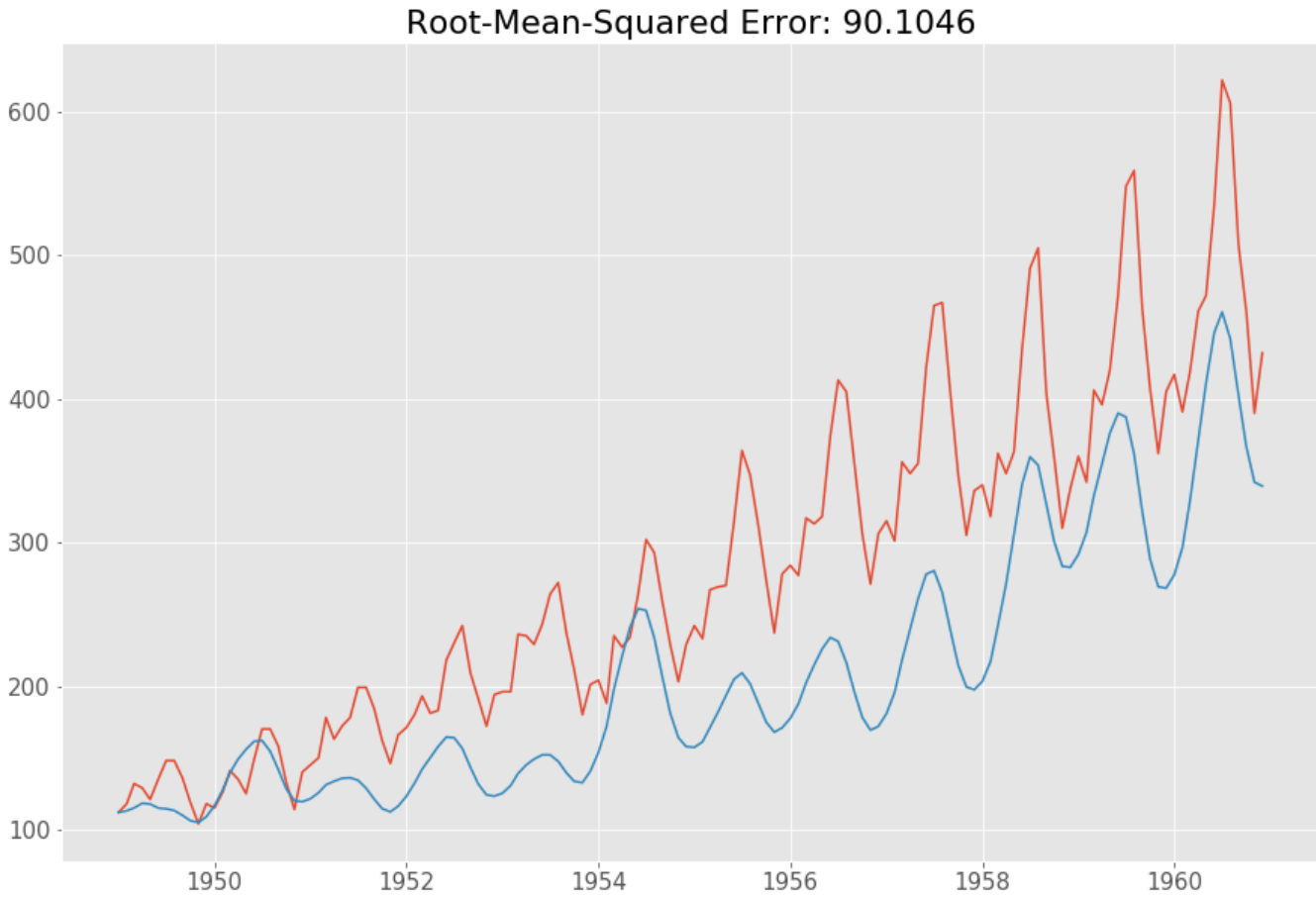
$ARMA(2,2)$: Predicted on Residuals



- Here we can see that the $AR(2)$ and $MA(2)$ models have almost the same RSS but combined is significantly better.

Forecasting

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



SARIMA: Seasonal ARIMA Models

- One problem in the previous model is the lack of seasonality, which can be addressed in a generalized version of ARIMA model called seasonal ARIMA.
- **Definition**
  - A seasonal ARIMA model is formed by including additional seasonal terms in the ARIMA models, denoted as

$$ARIMA\,(p,d,q)\,(P,D,Q)_m,$$

    i.e.

$$\phi(B^m)\,\phi(B)\,(1-B^m)^D\,(1-B)^d\,X_t = \theta(B^m)\,\theta(B)\,w_t$$

    where m represents the number of observations per year.
  - The seasonal part of the model consists of terms that are similar to the non-seasonal components, but involve backshifts of the seasonal period.

Off-the-shelf Packages

- R has `arima` function in standard package `stats`.
- Mathematica has a complete library of time series functions including ARMA.
- MATLAB includes functions such as `arima`.
- Python has a `statsmodels` module provides time series analysis including ARIMA. Another Python module called `pandas` provides dedicated class for time series objects.
- STATA includes the function `arima` for estimating ARIMA models.
- SAS has an econometric package, `ETS`, for ARIMA models.
- GNU Octave can estimate AR models using functions from extra packages.

References

- *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/
- *STAT510: Applied Time Series Analysis*, Penn State Online Courses
- An Introductory Study on Time Series Modeling and Forecasting, Ratnadip Adhikari, R. K. Agrawal, 2013.
- Wikipedia articles for various concepts.

Thanks for your attention!