Trading Algorithms

TIME SERIES ANALYSIS

Lecturer: Reza Entezari-Maleki

entezari@iust.ac.ir

School of Computer Engineering

Iran University of Science and Technology

Outlines

- Introduction to time series
- Stationary and non-stationary time series
- Unit roots and ADF test
- Non-stationary time series data
- Autocorrelation and partial autocorrelation (ACF and PACF)
- Autoregressive (AR) and moving average (MA) models
- ARMA, ARIMA, and SARIMA models
- ARCH and GARCH models
- Vector autoregressive (VAR) models
- Granger causality

Introduction to time series

- In mathematics, a time series is a series of data points indexed (or listed or graphed) in time order.

- The observations (data points) each occur at some time t, where t belongs to the set of allowed times, T.

$$\{x_t\}, t \in T$$

- Most commonly, a time series is a sequence taken at **successive equally spaced points** in time. Thus it is a sequence of discrete-time data.

- In traditional machine learning, a dataset is a **collection** of observations. It makes predictions based on **unseen data**, and it predicts the future with all the previous observations taken into consideration.

- In a time series, the dataset is different. A time-series adds a definite order of dependence between observations.

Introduction to time series …

- Suppose we want to show the number of ice cream cones sold during a year in two different ways: (1) **based on the temperature**, and (2) **based on the previous sales**.

- Our concern in this problem is to predict how many ice cream cones will be sold tomorrow.

- Both methods help us to predict this metric, but the first approach is a non-time series way to represent the problem and the second one is a time series way.

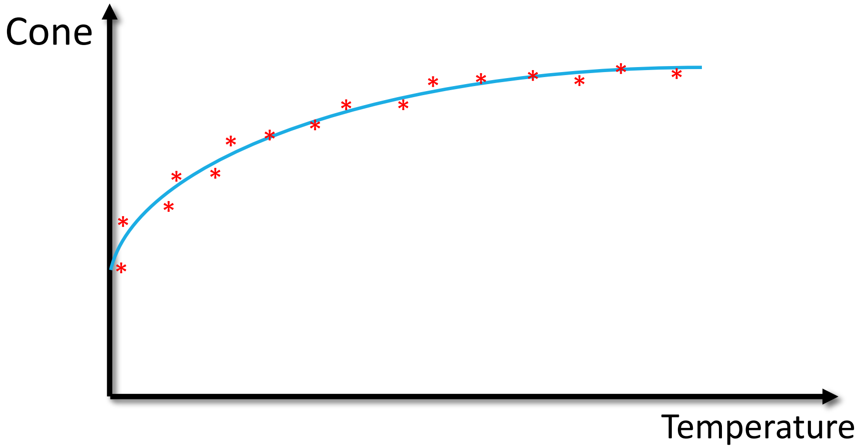
- In the first approach we should predict the tomorrow's temperature to be able to predict the number of tomorrow's ice cream cones.

- In the second approach, the number of ice creams in **day t'+1 is predicted using the number of ice creams sold from t to t'.**

Introduction to time series …

- In the first way, the colder days the less ice cream was sold and the hotter days the more ice cream was sold.

- We show the sales using red strikes, and use **interpolation** to estimate new data points based on the range of a discrete set of known data points (red strikes).

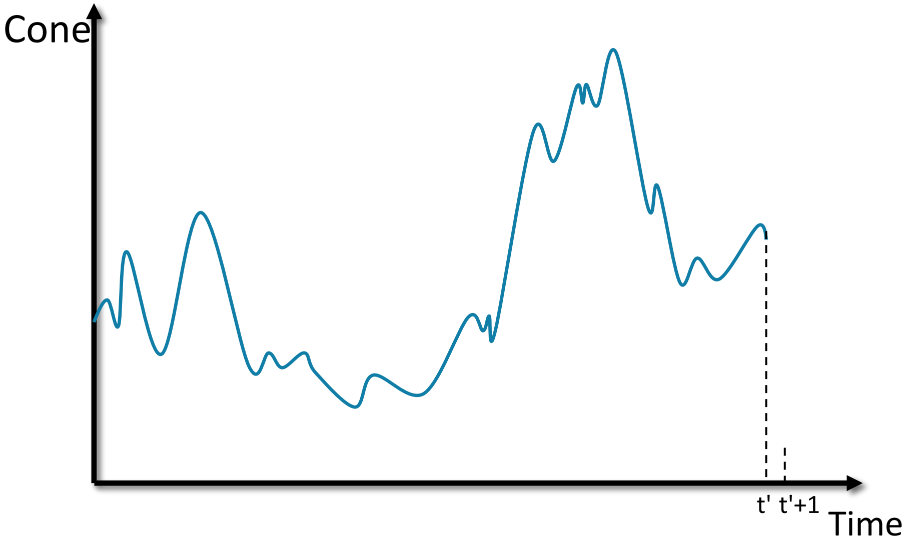


Introduction to time series …

- In the second way, the temperature is not longer the variable, instead the lag version of cones is our variable.

- In this case, we do **extrapolation** to predict a point outside the range.

- Time series are always going to be extrapolation problem.



Introduction to time series …

- Extrapolation is more difficult compared to the interpolation because the error is propagated and gets higher when we wish to estimate a point far from the known points.
- In other words, **the uncertainty of day t'+2** not only depends on the **uncertainty of days t to t'** but also the uncertainty of day t'+1 whish is recently predicted.
- Therefore, the error (confidence interval or prediction interval) is growing when we get further from the known range.
- However, time series can act as a powerful prediction tools, especially for short prediction periods, say for prediction tomorrow's ice cream cones.

Stationary and non-stationary time series

- A **stationary time series** is one **whose properties do not depend on the time** at which the series is observed.

- When a time series is stationary, it means that **certain attributes of the data do not change over time**.

- **It does not mean that the series does not change over time, just that the way it changes does not itself change over time.**

- However, some time series are non-stationary, whereby values and associations between and among variables do vary with time.

- **In finance, many processes are non-stationary, and so must be handled appropriate.**

Stationary and non-stationary time series …

- **Time series** with trends, or with seasonality, are **not stationary,** the trend and seasonality will affect the value of the time series at different times.

- In contrast to the non-stationary process that has a variable mean and variance that does not remain near, or returns to a long-run mean over time, **the stationary process reverts around a constant long-term mean and has a constant variance** independent of time.

- In **non-stationary time series**, data points often have **means and/or variances that change over time**.

- Non-stationary behaviors can be trends, cycles, random walks, or combinations of the three.

Stationary and non-stationary time series …

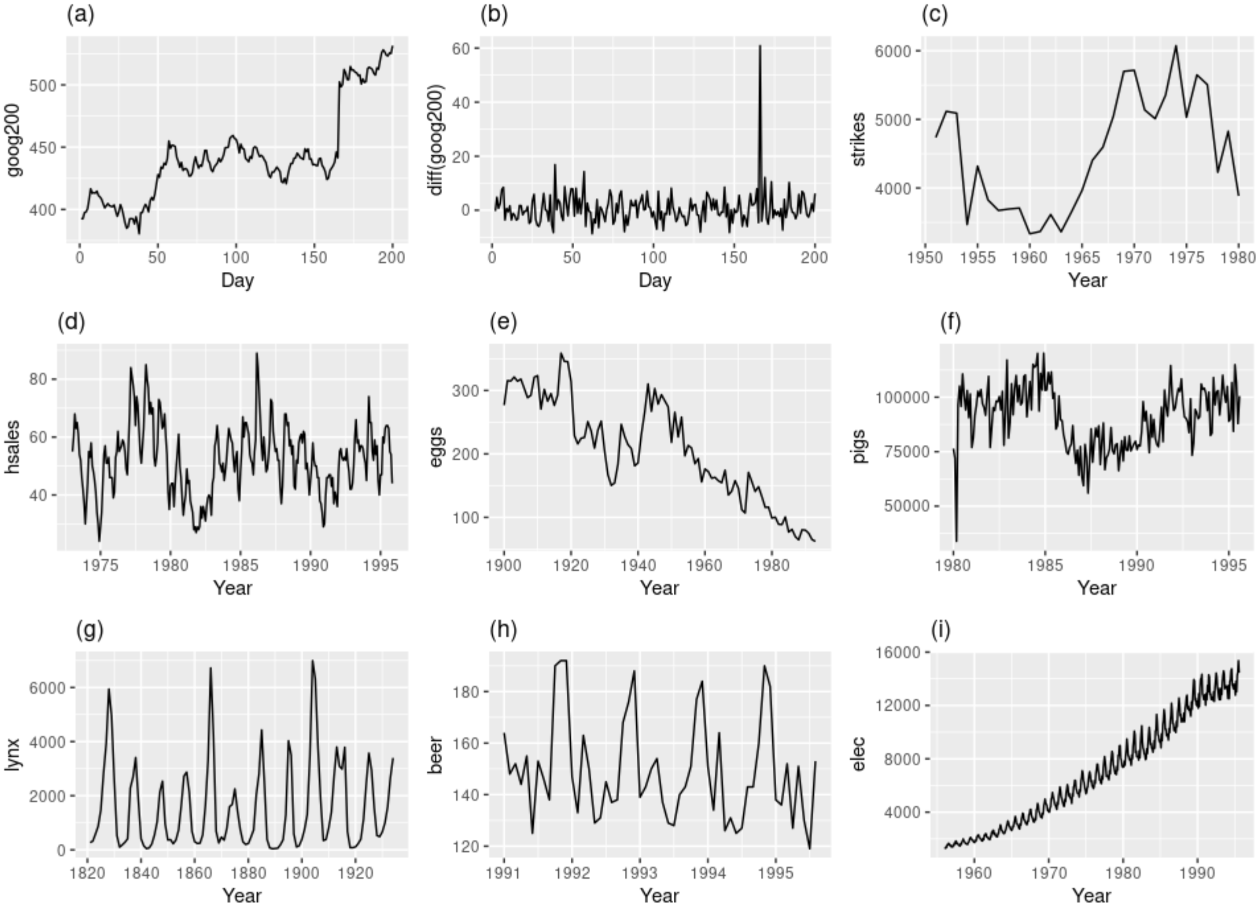
- **In general, if a time series has variable mean or variance, or it has the seasonality component, it cannot be considered as a stationary time series.**

- It should be noted that seasonality component of time series is different from the cyclicity component:
  - **Seasonality** refers to **predictable changes** that occur over a **one-year** period in a business or economy based on the seasons including **calendar or commercial seasons**.
  - The **cyclical** component of a time series refers to **fluctuations around the trend**, when data exhibit rises and falls that are **not of fixed period**.

- Usually seasonal cycles are observed within one calendar year, while cyclical effects, span time periods shorter or longer than one calendar year.

- If the fluctuations are not of a fixed frequency then they are cyclic (in the long-term, the timing of these cycles is not predictable); if the frequency is unchanging and associated with some aspect of the calendar, then the pattern is seasonal.

Stationary and non-stationary time series …



Unit roots

- A time series has stationarity if a shift in time doesn’t cause a change in the shape of the distribution; **unit roots are one cause for non-stationarity**. However, we can remove them.

- A unit root is a unit of measurement to determine **how much stationarity a time series model has**.

- Also called a unit root process, we determine the stochasticity of the model using statistical hypothesis testing.

- There is a myriad of ways to check for presence of a unit root process, one of the well-known methods is Augmented Dickey Fuller test.

Unit roots …

A simple time series example:

$$\boldsymbol{y_t = \rho y_{t-1} + \varepsilon_t}$$

$$= \rho(\rho y_{t-2} + \varepsilon_{t-1}) + \varepsilon_t = \rho^2 y_{t-2} + \rho\varepsilon_{t-1} + \varepsilon_t$$

$$= \rho^2(\rho y_{t-3} + \varepsilon_{t-2}) + \rho\varepsilon_{t-1} + \varepsilon_t = \rho^3 y_{t-3} + \rho^2\varepsilon_{t-2} + \rho\varepsilon_{t-1} + \varepsilon_t$$

$$\ldots$$

$$= \rho^{t-1}\left(\rho y_{t-t} + \varepsilon_{t-(t-1)}\right) + \rho^{t-2}\varepsilon_{t-(t-2)} + \cdots + \rho^2\varepsilon_{t-2} + \rho\varepsilon_{t-1} + \varepsilon_t$$

$$\boldsymbol{= \rho^t y_0 + \sum_{k=0}^{t-1} \rho^k \varepsilon_{t-k}}$$

$$\boldsymbol{Var(y_t)} = \sigma^2\left[\rho^0 + \rho^2 + \rho^4 + \cdots + \rho^{2(t-1)}\right]$$

$$\varepsilon_t \in N(0, \sigma^2)$$

$$\boldsymbol{E(y_t)} = \rho E(y_{t-1}) = \rho^2 E(y_{t-2}) = \cdots = \rho^t y_0$$

Unit roots …

- We study three different cases ofρ; $|\rho| < 1$,$|\rho| > 1$, and $|\rho| = 1$.
- First case: $|\boldsymbol{\rho}| < \mathbf{1}$

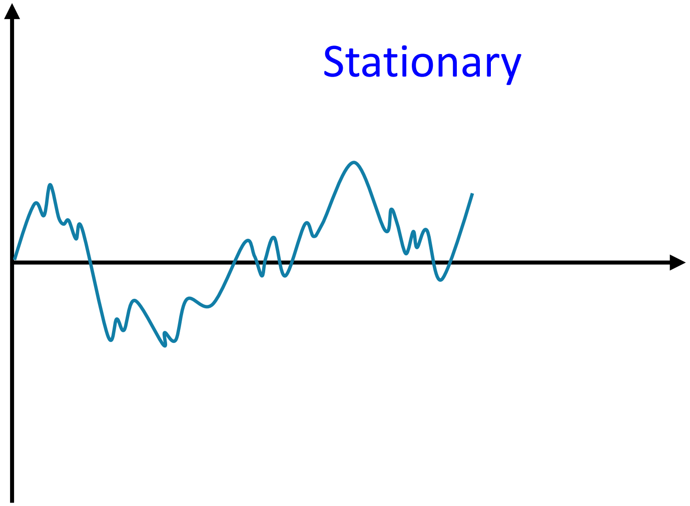
$$\boldsymbol{E(y_t)} = \rho^t y_0$$

$$\boldsymbol{Var(y_t)} = \sigma^2\left[\rho^0 + \rho^2 + \rho^4 + \cdots + \rho^{2(t-1)}\right]$$

$$\xRightarrow{|\rho|<1 \ \ and \ \ t \to \infty}$$

$$\boldsymbol{E(y_t)} = 0$$

$$\boldsymbol{Var(y_t)} = \frac{\sigma^2}{1-\rho^2}$$


Stationary

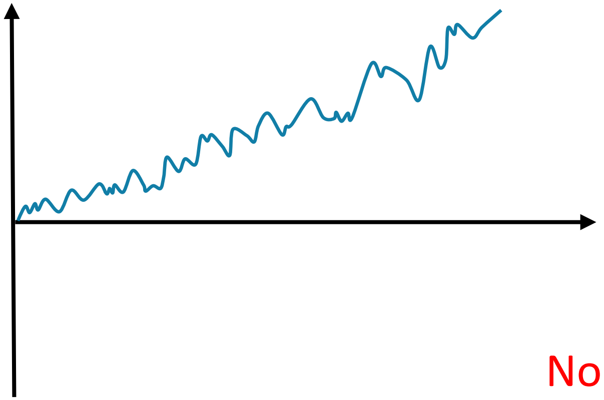
Unit roots …

- Second case: $|\boldsymbol{\rho}| > \mathbf{1}$

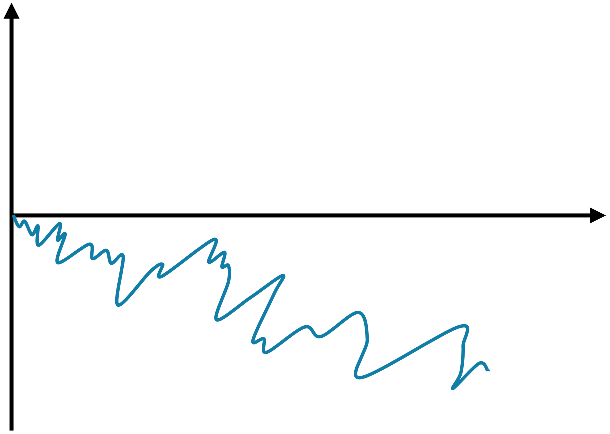
$$\boldsymbol{E(y_t)} = \rho^t y_0$$

$$\boldsymbol{Var(y_t)} = \sigma^2\left[\rho^0 + \rho^2 + \rho^4 + \cdots + \rho^{2(t-1)}\right]$$

$$\xRightarrow{|\rho|>1 \quad and \quad t \to \infty} \quad \boldsymbol{E(y_t)} = \pm\infty$$



Non-stationary



Unit roots ...

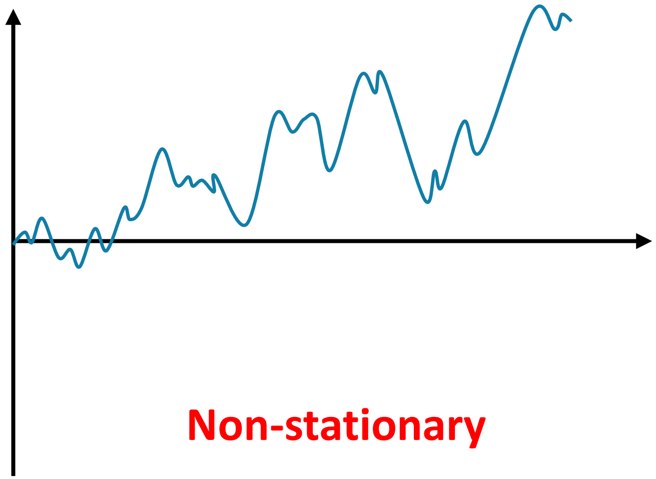
- Third case: $|\boldsymbol{\rho}| = \mathbf{1}$

$$\boldsymbol{E(a_t)} = \rho^t y_0$$

$$\boldsymbol{Var(a_t)} = \sigma^2\left[\rho^0 + \rho^2 + \rho^4 + \cdots + \rho^{2(t-1)}\right]$$

$$\xRightarrow{|\rho|=1 \quad and \quad t\to\infty}$$

$$\boldsymbol{E(y_t)} = y_0$$
$$\boldsymbol{Var(y_t)} = t\sigma^2$$



Non-stationary

Unit roots …

- Solution in the third case $|\boldsymbol{\rho}| = \mathbf{1}$: **first difference**

$$\boldsymbol{y_t = \rho y_{t-1} + \varepsilon_t}$$

Let $$\boldsymbol{\Delta y_t = y_t - y_{t-1}}$$

$$\boldsymbol{\Delta y_t = \varepsilon_t}$$

$$\boldsymbol{E(\Delta y_t)} = 0$$
$$\boldsymbol{Var(\Delta y_t)} = \sigma^2$$

$$\boldsymbol{\varepsilon_t \in N(0, \sigma^2)}$$

Dickey Fuller and Augmented Dickey Fuller tests

- When we make a model for forecasting purposes in time series analysis, we require a stationary time series for better prediction.

- So the first step to work on modeling is to make a time series stationary.

- Testing for stationarity is a frequently used activity in **autoregressive modeling**.

- There are various tests for this purpose, and Augmented Dickey-Fuller (ADF) is one of the most famous ones.

- The Dickey-Fuller test is a **statistical hypothesis test** that measures the amount of stochasticity in a time series model. The Dickey-Fuller test is based on linear regression.

Dickey Fuller and Augmented Dickey Fuller tests …

- The Dickey-Fuller test actually creates a t-statistic that is compared to predetermined critical values.

- Being below that critical statistic allows us to reject the null hypothesis and accept the alternative.

- If we are above this test statistic we fail to reject the null hypothesis.

- There are three main versions of the test:

$$\Delta y_t = \delta y_{t-1} + \varepsilon_t$$

$$\Delta y_t = \alpha + \delta y_{t-1} + \varepsilon_t$$

$$\Delta y_t = \alpha + \beta t + \delta y_{t-1} + \varepsilon_t$$

Dickey Fuller and Augmented Dickey Fuller tests …

- DF test

$$y_t = \alpha + \rho y_{t-1} + \varepsilon_t$$

H_0: $\rho = 1$ ← Unit root exists (non-stationary)

H_1: $\rho < 1$ ← No unit root (stationary)

$$y_t - y_{t-1} = \alpha + (\rho - 1) y_{t-1} + \varepsilon_t$$

$$\Delta y_t = \alpha + \delta y_{t-1} + \varepsilon_t$$

H_0: $\delta = 0$ ← Unit root exists (non-stationary)

H_1: $\delta < 0$ ← No unit root (stationary)

Dickey Fuller and Augmented Dickey Fuller tests …

- The value of the test statistic is $t_{\hat{\delta}} = \frac{\hat{\delta}}{SE(\hat{\delta})}$ which should be compared with the Dickey-Fuller Distribution critical values.

$$t_{\hat{\delta}} = \frac{\hat{\delta}}{SE(\hat{\delta})}$$

$t_{\hat{\delta}} < DF_{critical}$ ← Reject H_0

$t_{\hat{\delta}} > DF_{critical}$ ← Don’t reject H_0

Dickey Fuller and Augmented Dickey Fuller tests …

- ADF test

$$y_t = \alpha + \sum_{i=1}^{P} \rho_i y_{t-i} + \varepsilon_t$$

$$\Delta y_t = \alpha + \delta y_{t-1} + \sum_{i=1}^{P} \beta_i \Delta y_{t-1} + \varepsilon_t$$

$$t_{\hat{\delta}} = \frac{\hat{\delta}}{SE(\hat{\delta})}$$

$$t_{\hat{\beta}} = \frac{\hat{\beta}}{SE(\hat{\beta})}$$

Python programming example

ADF test

Non-stationary time series data

- Data points are often non-stationary or have means, variances, and covariances that change over time.
- Before we get to the point of transformation for the non-stationary financial time series data, we should **distinguish between the different types of non-stationary processes**.
- This will provide us with a better understanding of the processes and allow us to apply the correct transformation.
- Examples of non-stationary processes are random walk with or without a drift (a slow steady change) and deterministic trends (trends that are constant, positive, or negative, independent of time for the whole life of the series).

Non-stationary time series data …

- **Random walk theory**
  - suggests that changes in stock prices have the same distribution and are independent of each other.
  - infers that the past movement or trend of a stock price or market cannot be used to predict its future movement.
  - believes it's impossible to outperform the market without assuming additional risk.
  - considers technical analysis undependable because it results in chartists only buying or selling a security after a move has occurred.
  - considers fundamental analysis undependable due to the often-poor quality of information collected and its ability to be misinterpreted.
  - claims that investment advisors add little or no value to an investor’s portfolio.

Non-stationary time series data …

- Pure Random Walk ($y_t = y_{t-1} + \varepsilon_t$)
  - Random walk predicts that the value at time t will be equal to the last period value plus a stochastic component that is a white noise, which means ε_t is independent and identically distributed with mean 0 and variance σ^2.
  - Random walk can also be named a process integrated of some order, a process with a unit root or a process with a stochastic trend. **It is a non-mean-reverting process that can move away from the mean either in a positive or negative direction**.
  - Another characteristic of a random walk is that the variance evolves over time and goes to infinity as time goes to infinity; therefore, **a random walk cannot be predicted**.

Non-stationary time series data …

- Random Walk with Drift ($y_t = \alpha + y_{t-1} + \varepsilon_t$)
  - If the random walk model predicts that the value at time t will equal the last period's value plus a constant, or drift (α), and a white noise term (ε_t), then the process is random walk with a drift.
  - **It also does not revert to a long-run mean and has variance dependent on time.**
- Deterministic Trend ($y_t = \alpha + \beta t + \varepsilon_t$)
  - Often a random walk with a drift is confused for a deterministic trend.
  - Both include a drift and a white noise component, but the value at time t in the case of a random walk is regressed on the last period's value (y_{t-1}), while in the case of a deterministic trend it is regressed on a time trend (βt).

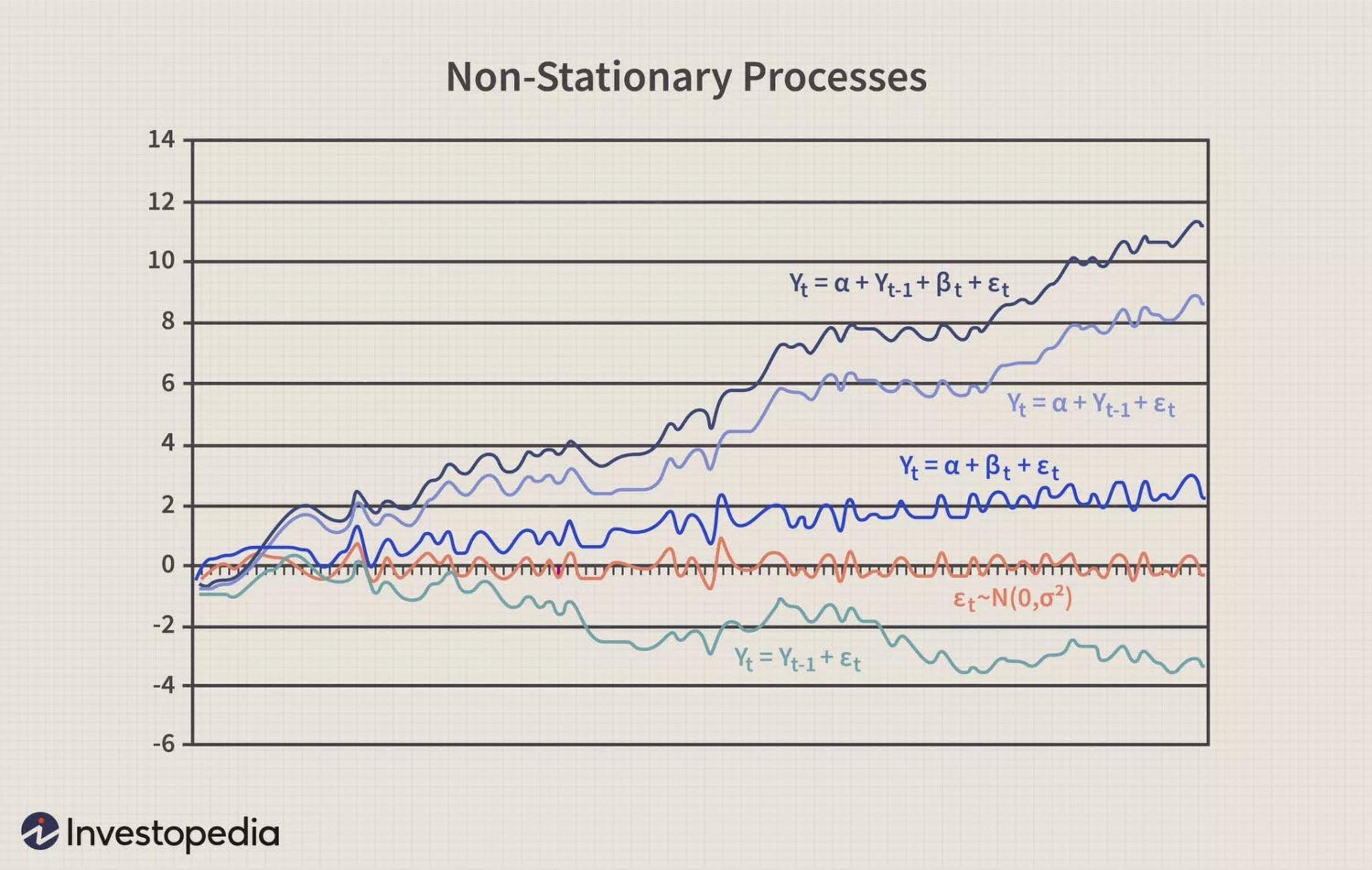
Non-stationary time series data …

- A non-stationary process with a deterministic trend has a mean that grows around a fixed trend, which is constant and independent of time.

➢Random Walk with Drift and Deterministic Trend ($y_t = \alpha + y_{t-1} + \beta t + \varepsilon_t$)

- Another example is a non-stationary process that combines a random walk with a drift component (α) and a deterministic trend (βt).
- It specifies the value at time t by the last period's value, a drift, a trend, and a stochastic component.

Non-stationary time series data …



Making a time series stationary

- There are different ways to check if a time series is stationary or not. For example,
  - in the fist step, we visually investigate a time series and check its trend and volatility.
  - in the second step, we can run local and global tests on the time series to check the trend, volatility and seasonality components.
  - in the third step and as a more formal way, we can apply ADF tests on a time series.
- Now, the question is that what will be the next step if we realize a time series is not stationary.
- One of the popular methods for making a non-stationary time series stationary is **differencing**.
- Differencing can help stabilize the mean of a time series by removing changes in the level of a time series, and therefore eliminating (or reducing) trend and seasonality.

Making a time series stationary …

- A random walk with or without a drift can be transformed to a stationary process by **differencing** (subtracting y_{t-1} from y_t, taking the difference y_t - y_{t-1}) correspondingly to $y_t - y_{t-1} = \varepsilon_t$ or $y_t - y_{t-1} = \alpha + \varepsilon_t$ and then the process becomes difference-stationary.

- The disadvantage of differencing is that the process loses one observation each time the difference is taken.

- A non-stationary process with a deterministic trend becomes stationary after removing the trend, or **detrending**.

- For example, $y_t = \alpha + \beta t + \varepsilon_t$ is transformed into a stationary process by subtracting the trend βt: $y_t - \beta t = \alpha + \varepsilon_t$.

- No observation is lost when **detrending** is used to transform a non-stationary process to a stationary one.

Making a time series stationary …

- We can also subtract y_{t-1} from y_t in a non-stationary process with a deterministic trend, as below

$$y_t = \alpha + \beta t + \varepsilon_t$$

$$\Delta y_t = y_t - y_{t-1} = \alpha + \beta t + \varepsilon_t - \alpha - \beta(t-1) - \varepsilon_{t-1}$$

$$\Delta y_t = \beta + \varepsilon_t - \varepsilon_{t-1}$$

- Since $\varepsilon_t \sim N(0, \sigma^2)$

$$E(\Delta y_t) = \beta$$

$$Var(\Delta y_t) = 2\sigma^2$$

Python programming example

Making a time series stationary

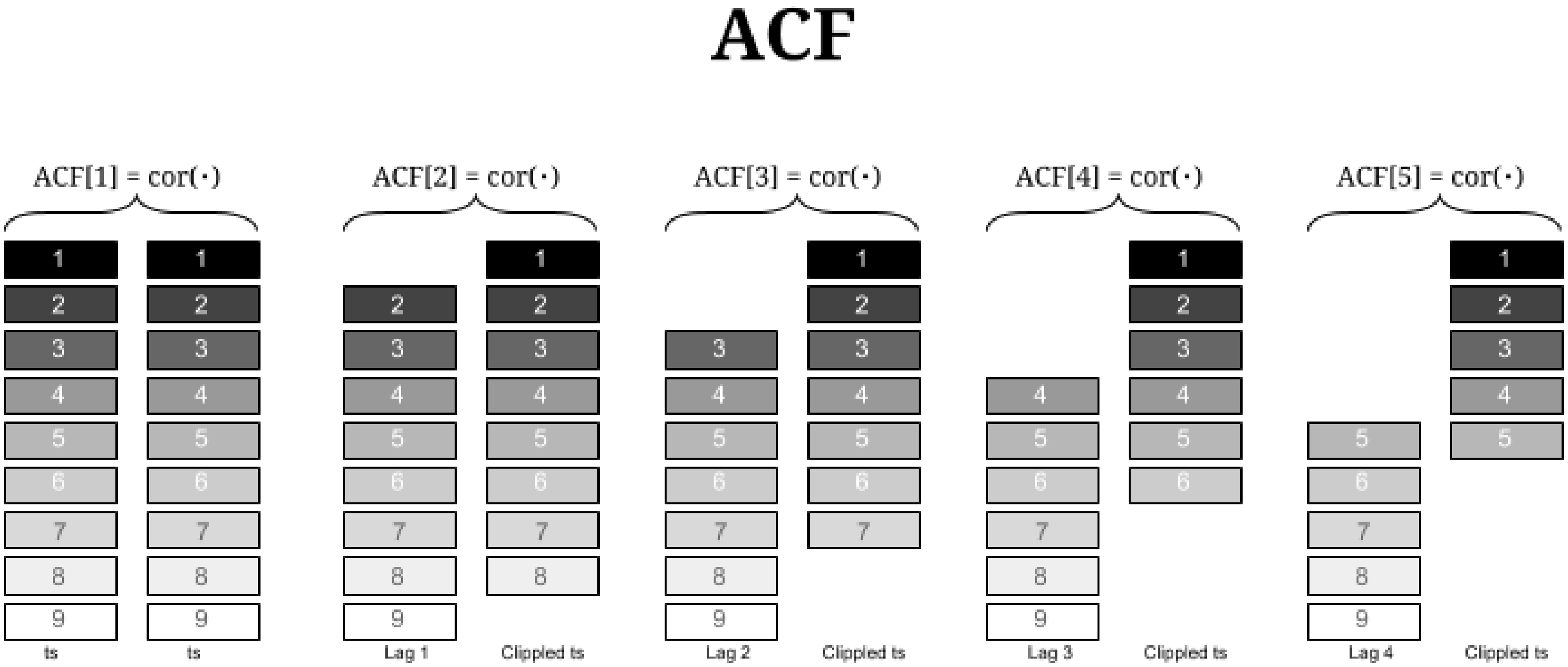
Autocorrelation and partial autocorrelation

- Autocorrelation is the correlation between a time series with a lagged version of itself.
- The **autocorrelation analysis** helps detect patterns and check for randomness.
- It's especially important when you intend to use an autoregressive–moving-average (ARMA) model for forecasting because it **helps to determine its parameters**.
- The analysis involves looking at the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots.

Autocorrelation and partial autocorrelation …

- **ACF** is an (complete) auto-correlation function which gives us values of auto-correlation of any series with its lagged values.
- In other words: we have a time series, and basically make multiple copies of it, understanding that each copy is going to be offset by one entry from the prior copy, because the initial data contains *t* data points, while the previous time series length (which excludes the last data point) is only *t−1*.
- Each copy is correlated to the original, keeping in mind that **we need identical lengths**, and to this end, we'll have to keep on **clipping the tail end of the initial data series** to make them comparable.
- The ACF plot describes how well the present value of the series is related with its past values.

Autocorrelation and partial autocorrelation …



Autocorrelation and partial autocorrelation …

- The correlation between two variables y_1 and y_2 is defined as:

$$r = \frac{E[(y_1 - \mu_1)(y_2 - \mu_2)]}{\sigma_1 \sigma_2} = \frac{Cov(y_1, y_2)}{\sigma_1 \sigma_2}$$

- Where E is the expectation operator, μ_1 and μ_2 are the means respectively for y_1 and y_2 and σ_1, σ_2 are their standard deviations.

- In the context of a single variable, i.e. auto-correlation, y_1 is the original series and y_2 is a lagged version of it.

- Upon the above definition, sample autocorrelations of order $k = 0,1,2, \ldots$ can be obtained by computing the following expression with the observed series $y_t, t = 1,2 \ldots, n$, where $\bar{y}$ is the sample mean of data.

$$r(k) = \frac{\frac{1}{n-k} \sum_{t=k+1}^{n} (y_t - \bar{y})(y_{t-k} - \bar{y})}{\sqrt{\frac{1}{n} \sum_{t=1}^{n} (y_t - \bar{y})^2} \sqrt{\frac{1}{n-k} \sum_{t=k+1}^{n} (y_{t-k} - \bar{y})^2}}$$

Autocorrelation and partial autocorrelation …

- The partial autocorrelation function (PACF) gives the partial correlation of a stationary time series with its own lagged values, regressed the values of the time series at all shorter lags.

- For example, the partial autocorrelation of order 2 measures the effect (linear dependence) of y_{t-2} on y_t after removing the effect of y_{t-1} on both y_t and y_{t-2}.

- One way to compute the sample PACF is solving the following system for each order k, (γ_k is PACF of order k).

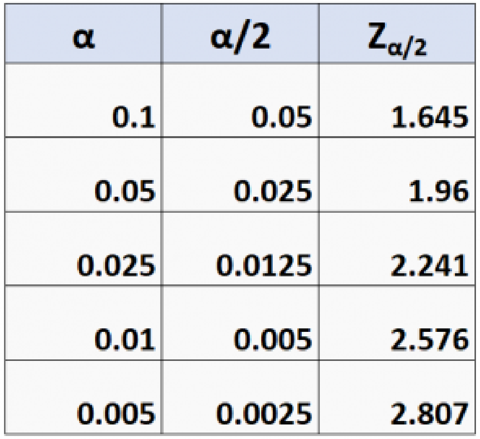
$$\begin{pmatrix} r(0) & r(1) & \dots & r(k-1) \\ r(1) & r(0) & \dots & r(k-2) \\ \vdots & \vdots & \ddots & \vdots \\ r(k-1) & r(k-2) & \dots & r(0) \end{pmatrix} \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_k \end{pmatrix} = \begin{pmatrix} r(1) \\ r(2) \\ \vdots \\ r(k) \end{pmatrix}$$

Autocorrelation and partial autocorrelation …

- We plot these two values (ACF and PACF) along with the confidence band.

- Confidence bands can be computed as the value of the sample autocorrelations $\pm\frac{z_{\frac{\alpha}{2}}}{\sqrt{n}}$, where $z_{\frac{\alpha}{2}}$ is referring to the z critical value from the z table that corresponds to $\frac{\alpha}{2}$, e.g. 1.96 for 95% confidence bands.

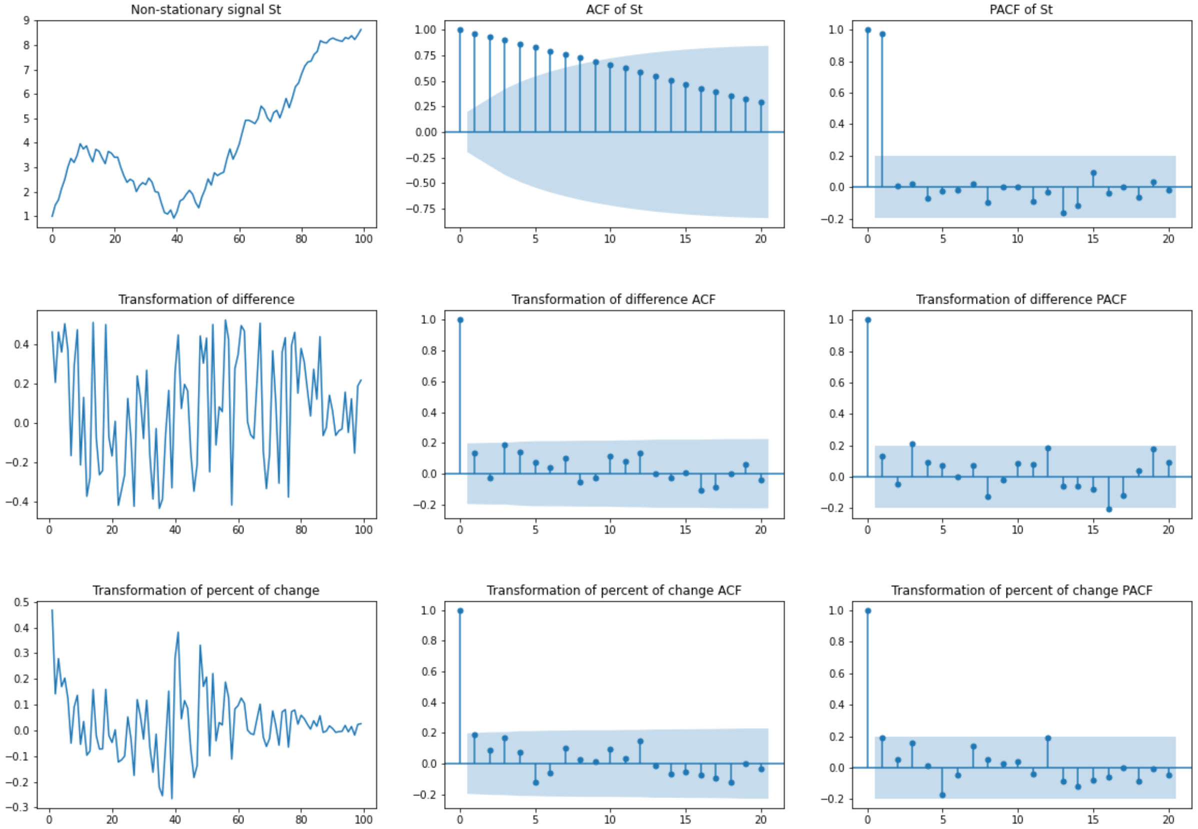
- Sometimes confidence bands that increase as the order increases are used. In this cases the bands can be defined as $\pm\frac{z_{\frac{\alpha}{2}}\times\sqrt{(1+2\sum_{i=1}^{k} r(i)^2)}}{\sqrt{n}}$



α	α/2	$Z_{\alpha/2}$
0.1	0.05	1.645
0.05	0.025	1.96
0.025	0.0125	2.241
0.01	0.005	2.576
0.005	0.0025	2.807

Autocorrelation and partial autocorrelation …

- Some sample plots of ACF and PACF



Python programming example

ACF and PACF

Autoregressive (AR) models

- A statistical model is autoregressive if it predicts future values based on its past values.
- For example, an autoregressive model might seek to predict a stock's future prices based on its past performance.
- Thus, an autoregressive model of **order p** can be written as

$$y_t = \alpha + \rho_1 y_{t-1} + \rho_2 y_{t-2} + \cdots + \rho_p y_{t-p} + \varepsilon_t$$

- This is similar to a multiple regression but with lagged values of y_t as predictors.
- We refer to this as an AR(p) model, an autoregressive model of order p.
- Autoregressive models are remarkably flexible at handling a wide range of different time series patterns.

Autoregressive (AR) models …

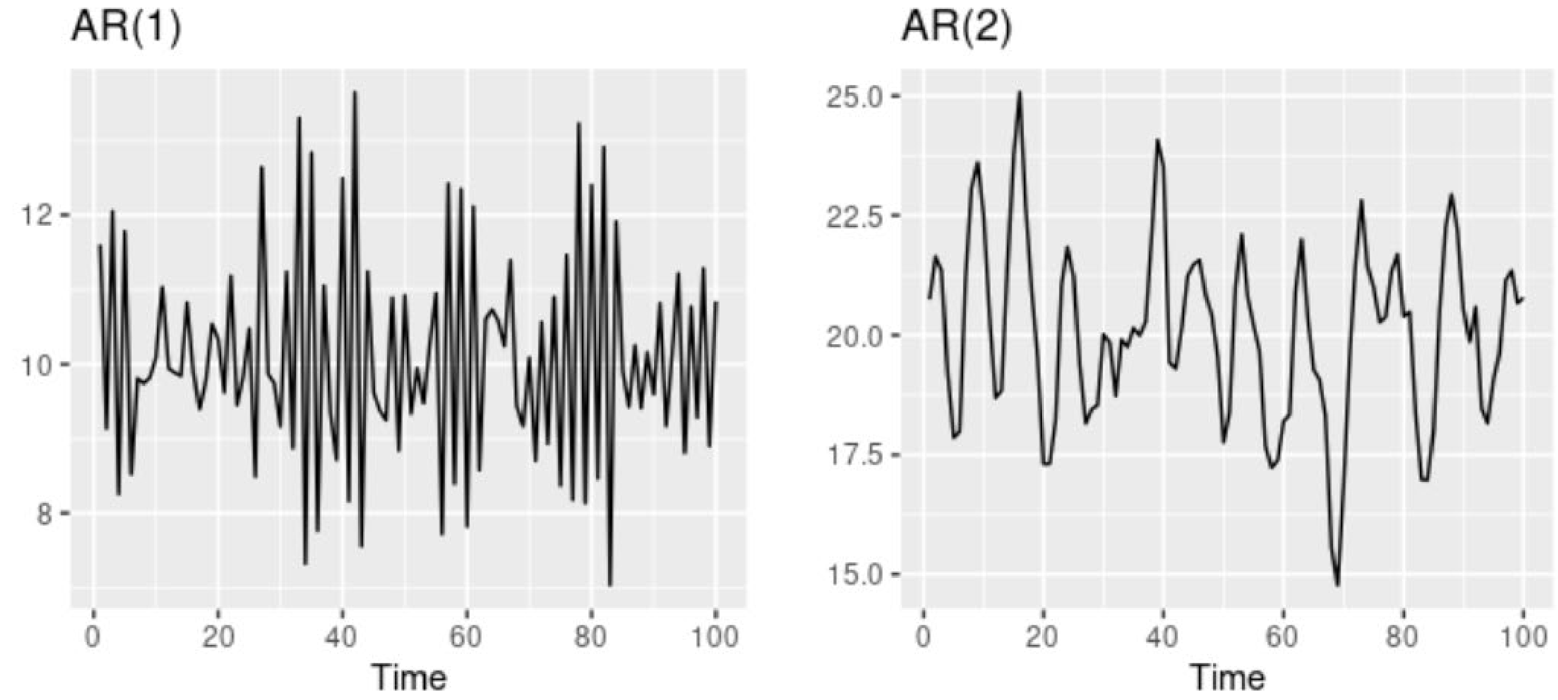


Figure 8.5: Two examples of data from autoregressive models with different parameters. Left: AR(1) with $y_t = 18 - 0.8y_{t-1} + \varepsilon_t$. Right: AR(2) with $y_t = 8 + 1.3y_{t-1} - 0.7y_{t-2} + \varepsilon_t$. In both cases, ε_t is normally distributed white noise with mean zero and variance one.

Autoregressive (AR) models …

- For an **AR(1)** model:

$$\boldsymbol{y_t = \alpha + \rho_1 y_{t-1} + \varepsilon_t}$$

  - ○ when $\rho_1 = 0$, y_t is equivalent to white noise
  - ○ when $\rho_1 = 1$ and $\alpha = 0$, y_t is equivalent to a random walk (pure random walk)
  - ○ when $\rho_1 = 1$ and $\alpha \neq 0$, y_t is equivalent to a random walk with drift
  - ○ when $|\rho_1| < 1$, y_t, tends to oscillate around the mean.

- These concepts and techniques are used by technical analysts to forecast security prices.

- However, since autoregressive models base their predictions only on past information, they **implicitly assume** that the fundamental forces that influenced the past prices will not change over time.

Moving average (MA) models

- Rather than using past values of the forecast variable in a regression, a **moving average model** uses past forecast errors in a regression-like model.

$$y_t = \alpha + \varepsilon_t + \varphi_1 \varepsilon_{t-1} + \varphi_2 \varepsilon_{t-2} + \cdots + \varphi_q \varepsilon_{t-q}$$

- We refer to this as an MA(q) model, a moving average model of order q.
- Of course, **we do not observe the values of ε_t, so it is not really a regression in the usual sense**.
- Notice that each value of y_t can be thought of as a weighted moving average of the past few forecast errors.
- **However, moving average models should not be confused with the moving average smoothing.**
- A moving average model is used for forecasting future values, while moving average smoothing is used for estimating the trend-cycle of past values.

Moving average (MA) models …

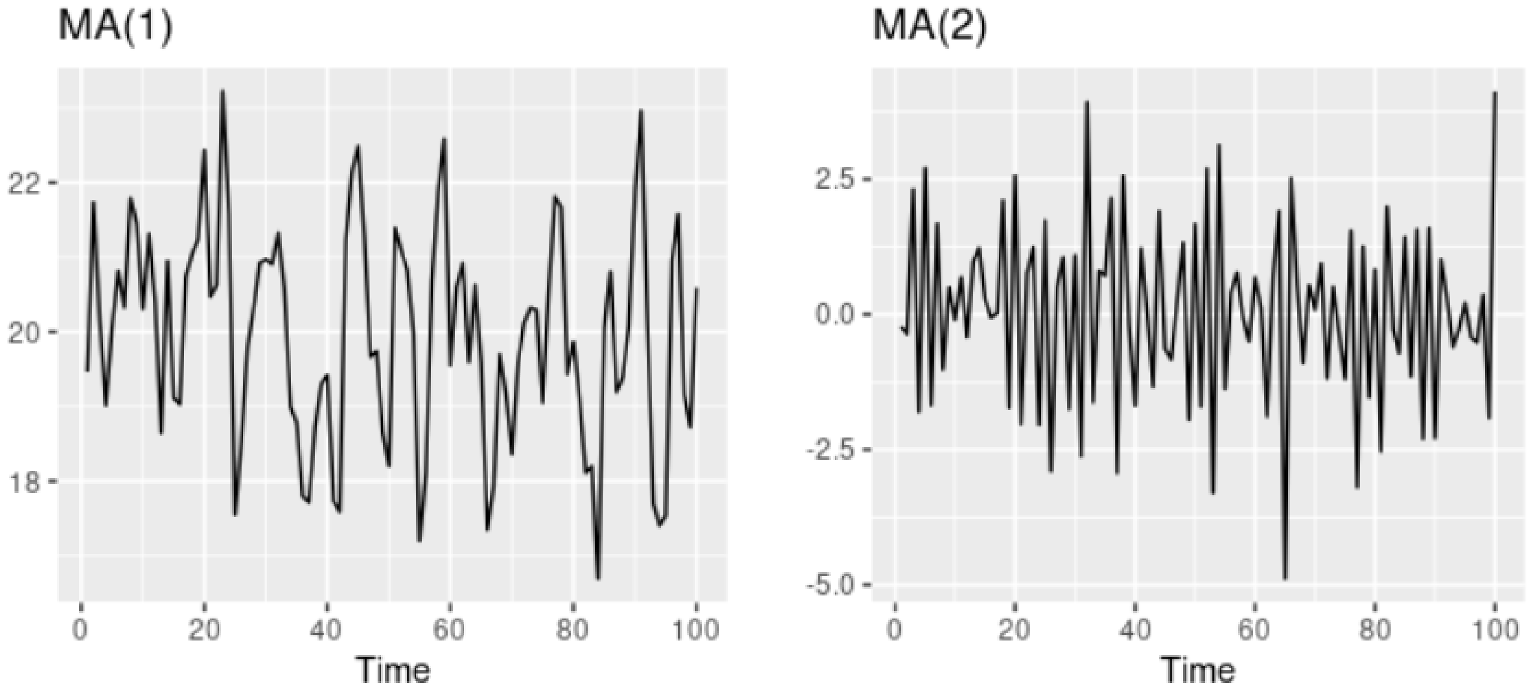


Figure 8.6: Two examples of data from moving average models with different parameters. Left: MA(1) with $y_t = 20 + \varepsilon_t + 0.8\varepsilon_{t-1}$. Right: MA(2) with $y_t = \varepsilon_t - \varepsilon_{t-1} + 0.8\varepsilon_{t-2}$. In both cases, ε_t is normally distributed white noise with mean zero and variance one.

Interpretation of ACF and PACF plots for Identifying AR and MA

- In time series analysis, autocorrelation function (ACF) and the partial autocorrelation function (PACF) plots are essential in providing the model's orders such as ***p*** for **AR** and ***q*** for **MA** to select the best model for forecasting.

- The basic guideline for interpreting the ACF and PACF plots is to look for tail off pattern in either ACF or PACF.

- Though ACF and PACF do not directly dictate the order of the ARMA model, the plots can facilitate finding the orders of AR and MA.

Interpretation of ACF and PACF plots for Identifying AR and MA …

- We can select the order p for AR(p) model based on significant **spikes from the PACF plot**.

- **One more indication** of the AR process is that the **ACF plot decays more slowly**.

- **Every bar outside the blue boundary** of the PACF plot (**confidence bands**) tell us the order of the AR model.

- In contrast to the AR model, we can select the order q for model MA(q) from ACF if this plot has a sharp cut-off after lag q.

- **One more indication** of the MA process is that the PACF plot decays more slowly.

Python programming example

AR model

Python programming example

MA model

ARMA model

- In the statistical analysis of time series, autoregressive moving average (ARMA) models provide a parsimonious description of a (weakly) stationary stochastic process in terms of **two polynomials**, one for the **autoregression (AR)** and the second for the **moving average (MA)**.

- The notation *ARMA(p, q)* refers to the model with *p* autoregressive terms and *q* moving-average terms. This model contains the AR(p) and MA(q) models together as the below.

$$y_t = \varepsilon_t + \sum_{i=1}^{p} \rho_i y_{t-i} + \sum_{i=1}^{q} \varphi_i \varepsilon_{t-i}$$

Python programming example

ARMA model

Lag (backshift) operator

- The lag operator L is a useful notational device when working with time series lags.

$$Ly_t = y_{t-1}$$

- Some references use **B** for backshift instead of **L** for lag.

- Two application of L to y_t shifts the data back two periods.

$$L(Ly_t) = L^2 y_t = y_{t-2}$$

- For monthly data, if we wish to consider the same month last year, the notation is $L^{12} y_t = y_{t-12}$.

- The lag operator is convenient for describing the process of differencing. A first difference can be written as

$$\Delta y_t = y_t - y_{t-1} = y_t - Ly_t = (1 - L)y_t$$

Lag (backshift) operator …

- So, the **first difference** is represented by $(\mathbf{1}-\boldsymbol{L})$.
- Similarly, if second-order differences have to be computed, then:

$$\Delta^2 y_t = \Delta(\Delta y_t) = \Delta(y_t - y_{t-1}) = \Delta y_t - \Delta y_{t-1}$$
$$(y_t - y_{t-1}) - (y_{t-1} - y_{t-2}) = y_t - 2y_{t-1} + y_{t-2}$$
$$= (1 - 2L + L^2)y_t = (1-L)^2 y_t$$

- In general, a **d***th*-order difference can be written as

$$\boldsymbol{(1-L)^d y_t}$$

- Backshift notation is particularly useful when combining differences, as the operator can be treated using ordinary algebraic rules. **In particular, terms involving L can be multiplied together**.

Invertibility of time series

- The fundamental idea in **invertibility** is the connection between MA and AR models.

- AR and MA are two different models derived for different purposes but when we look at them we realize that there is a fundamental connection between these models.

- First, we show that how the MA(1) model is really one AR(∞) model.

- Second, we will show that how the AR(1) model is really one MA(∞) model.

- We remember that AR(p) model is $y_t = \alpha + \rho_1 y_{t-1} + \rho_2 y_{t-2} + \cdots + \rho_p y_{t-p} + \varepsilon_t$ and MA(q) model is represented by $y_t = \alpha + \varphi_1 \varepsilon_{t-1} + \varphi_2 \varepsilon_{t-2} + \cdots + \varphi_q \varepsilon_{t-q} + \varepsilon_t$.

Invertibility of time series …

- $\mathbf{MA}(\mathbf{1}) \Leftrightarrow \mathbf{AR}(\infty)$

MA(1): $\boldsymbol{y_t = -\varphi\varepsilon_{t-1} + \varepsilon_t}$

$$y_t = (1 - \varphi L)\varepsilon_t$$

$$\frac{y_t}{(1-\varphi L)} = \varepsilon_t, \quad |\varphi|<1$$

$$(1 + \varphi L + \varphi^2 L^2 + \cdots)y_t = \varepsilon_t$$

$$y_t + \varphi y_{t-1} + \varphi^2 y_{t-2} + \cdots = \varepsilon_t$$

AR(∞): $\boldsymbol{y_t = -\varphi y_{t-1} - \varphi^2 y_{t-2} - \cdots + \varepsilon_t}$

$$\begin{tikzcd}
y_t & \varepsilon_t \arrow[l] \\
\varepsilon_{t-1} \arrow[u] & y_{t-1} \arrow[l] \\
\varepsilon_{t-2} \arrow[u] & y_{t-2} \arrow[l] \\
\varepsilon_{t-3} \arrow[u] & y_{t-3} \arrow[l] \\
\vdots \arrow[u] & \cdots \arrow[l] \\
{} \arrow[u] &
\end{tikzcd}$$

Invertibility of time series …

- $\mathbf{AR(1)} \Leftrightarrow \mathbf{MA(\infty)}$

AR(1): $\boldsymbol{y_t = \rho y_{t-1} + \varepsilon_t}$

$$\varepsilon_t = (1 - \rho L) y_t$$

$$\frac{\varepsilon_t}{(1-\rho L)} = y_t, \quad |\rho|<1$$

$$(1 + \rho L + \rho^2 L^2 + \cdots)\varepsilon_t = y_t$$

$$\varepsilon_t + \rho \varepsilon_{t-1} + \rho^2 \varepsilon_{t-2} + \cdots = y_t$$

MA(∞): $\boldsymbol{y_t = -\rho \varepsilon_{t-1} - \rho^2 \varepsilon_{t-2} - \cdots + \varepsilon_t}$

$$\begin{tikzcd}
y_t & \varepsilon_t \arrow[l] \\
y_{t-1} \arrow[u] & \varepsilon_{t-1} \arrow[l] \\
y_{t-2} \arrow[u] & \varepsilon_{t-2} \arrow[l] \\
y_{t-3} \arrow[u] & \varepsilon_{t-3} \arrow[l] \\
\vdots \arrow[u] & \cdots \arrow[l] \\
{} \arrow[u] &
\end{tikzcd}$$

ARIMA model

- A popular and widely used statistical method for time series forecasting is the **ARIMA model**.
- It is a class of models that captures a suite of different standard temporal structures in time series data.
- This acronym is descriptive, capturing the key aspects of the model itself. Briefly, they are:
  - **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** (e.g. subtracting an observation from an observation at the previous time step) 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**.

ARIMA model …

- Each of these components are explicitly specified in the model as a parameter.
- 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.
- The parameters of the ARIMA model are defined as follows:
  - 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.

$$\left(1-\sum_{i=1}^{p}\rho_i L^i\right)(1-L)^d y_t=\left(1+\sum_{i=1}^{q}\varphi_i L^i\right)\varepsilon_t$$

SARIMA model

- So far, we have restricted our attention to non-seasonal data and non-seasonal ARIMA models.
- However, ARIMA models are also capable of modelling a wide range of seasonal data.
- A seasonal ARIMA model is formed by including additional seasonal terms in the ARIMA models we have seen so far. It is written as follows:

$$\text{ARIMA} \qquad \underbrace{(p,d,q)}_{\substack{\uparrow \\ \text{Non-seasonal part} \\ \text{of the model}}} \quad \underbrace{(P,D,Q)_m}_{\substack{\uparrow \\ \text{Seasonal part of} \\ \text{of the model}}}$$

- where m = number of observations per year. **We use uppercase notation for the seasonal part of the model**, and lowercase notation for the non-seasonal parts of the model.

SARIMA model ...

- The seasonal part of the model consists of terms that are similar to the non-seasonal components of the model, but involve backshifts of the seasonal period.

- Generally, an $SARIMA(p,d,q)(P,D,Q)_m$ model is defined as

$$\left(1-\sum_{i=1}^{p}\rho_i L^i\right)\left(1-\sum_{i=1}^{P}\mathrm{P}_i L^{m.i}\right)(1-L)^d(1-L^m)^D y_t$$

$$=\left(1+\sum_{i=1}^{q}\varphi_i L^i\right)\left(1+\sum_{i=1}^{Q}\Phi_i L^{m.i}\right)\varepsilon_t$$

- For example, an $SARIMA(1,1,1)(1,1,1)_4$ model is for **quarterly data** $(\boldsymbol{m}=\mathbf{4})$ and can be written as

$$(1-\rho_1 L)(1-\mathrm{P}_1 L^4)(1-L)(1-L^4)y_t=(1+\varphi_1 L)(1+\Phi_1 L^4)\varepsilon_t$$

Python programming example

SARIMA model

ARCH model

- Autoregressive models can be developed for univariate time series data that is stationary (AR), has a trend (ARIMA), and has a seasonal component (SARIMA).
- One aspect of a univariate time series that these autoregressive models do not model is **a change in the variance over time**.
- Classically, a time series with modest changes in variance can sometimes be adjusted using a **power transform**, such as by taking the Log.
- There are some time series where the variance changes consistently over time. In the context of a time series in the financial domain, this would be called increasing and decreasing volatility.
- **In time series where the variance is increasing in a systematic way, such as an increasing trend, this property of the series is called heteroskedasticity**. It's a fancy word from statistics that means changing or unequal variance across the series.

ARCH model …

- If the **change in variance can be correlated over time**, then it can be modeled using an **autoregressive process**, such as **ARCH**.

- **Autoregressive Conditional Heteroskedasticity (ARCH)** is a method that explicitly models the change in variance over time in a time series.

- Specifically, an ARCH method models the variance at a time step as a function of the residual errors from a mean process (e.g. a zero mean).

- A **lag parameter** must be specified to define the number of **prior residual errors** to include in the model.

- A generally accepted notation for an ARCH model is to specify the ARCH() function with the p parameter *ARCH(p)*; for example, ARCH(1) would be a first order ARCH model.

ARCH model …

- The approach expects the series is stationary, other than the change in variance, meaning it does not have a trend or seasonal component.

- **An ARCH model is used to predict the variance at future time steps.**

- In practice, this can be used to model the expected variance on the residuals after another autoregressive model has been used, such as an ARMA or similar.

- An ARCH model could be used for any series that has periods of increased or decreased variance. This might, for example, be a property of residuals (ϵ_t) after an ARIMA model has been fit to the data.

ARCH model …

- Suppose that we are modeling the variance of a series y_t. The ARCH(1) model for the variance of model y_t is that conditional on y_{t-1}, the variance at time t is

$$Var(y_t|y_{t-1}) = \sigma_t^2 = \alpha_0 + \alpha_1 y_{t-1}^2$$

- We impose the **constraints $\boldsymbol{\alpha_0, \alpha_1 \geq 0}$ to avoid negative variance**.

- The variance at time *t* is connected to the value of the series at time *t-1*.

- A relatively large value of y_{t-1}^2 gives a relatively large value of the variance at time.

- If we assume that the series has mean = 0 (this can always be done by centering), the ARCH model could be written as

ARCH model …

$$y_t = \sigma_t \varepsilon_t$$

with $\sigma_t = \sqrt{\alpha_0 + \alpha_1 y_{t-1}^2}$

and $\varepsilon_t \sim N(\mu = 0, \sigma^2 = 1)$

- In ARCH (1) model of y_t, we can see that y_t^2 has a AR(1) model.
- An ARCH(m) process is one for which the variance at time t is conditional on observations at the previous m times, and the relationship is:

$$Var(y_t | y_{t-1}, \dots, y_{t-m}) = \sigma_t^2 = \alpha_0 + \alpha_1 y_{t-1}^2 + \dots + \alpha_m y_{t-m}^2$$

- With certain constraints imposed on the coefficients, the y_t^2 series will theoretically be AR(m).

GARCH model

- **Generalized Autoregressive Conditional Heteroskedasticity (GARCH)** is an extension of the ARCH model that incorporates a moving average component together with the autoregressive component.
- Specifically, the model includes lag variance terms (e.g. the observations if modeling the white noise residual errors of another process), together with lag residual errors from a mean process.
- The introduction of a moving average component allows the model to both model the conditional change in variance over time as well as changes in the time-dependent variance.
- As such, the model introduces a new parameter q that describes the number of lag variance terms:
  - p: The number of lag residual errors to include in the GARCH model.
  - q: The number of lag variances to include in the GARCH model.

GARCH model …

- A generally accepted notation for a GARCH model is to specify the GARCH function with the p and q parameters GARCH(p, q); for example GARCH(1, 1) would be a first order GARCH model and can be represented as:

$$\boldsymbol{\sigma_t^2 = \alpha_0 + \alpha_1 y_{t-1}^2 + \beta_1 \sigma_{t-1}^2}$$

- In the GARCH notation, the first subscript refers to the order of the y^2 terms on the right side, and the second subscript refers to the order of the σ^2 terms.
- The configuration for an ARCH model is best understood in the context of ACF and PACF plots of the variance of the time series.
- This can be achieved by subtracting the mean from each observation in the series and squaring the result, or just squaring the observation if you’re already working with residuals from another model.

GARCH model …

- The best identification tool may be a time series plot of the series.

- It can be fruitful to look at the ACF and PACF of both y_t and y_t^2. For instance, if y_t appears to be white noise and y_t^2 appears to be AR(1), then an ARCH(1) model for the variance is suggested.

- If the PACF of the y_t^2 suggests AR(m), then ARCH(m) may work.

- GARCH models may be suggested by an ARMA type look to the ACF and PACF of y_t^2.

- **In practice, things won't always fall into place as nicely as they did for the simulated example, you might have to experiment with various ARCH and GARCH structures after spotting the need in the time series plot of the series**.

Python programming example

GARCH model (synthetic sample)

Python programming example

GARCH model (stock forecasting)

Vector autoregressive (VAR) models

- The vector autoregressive (VAR) model is a workhouse multivariate time series model that relates current observations of a variable with past observations of itself and past observations of other variables in the system.
- VAR models differ from univariate autoregressive models because they allow feedback to occur between the variables in the model.
- As an example, suppose that we measure three different time series variables, denoted by $y_{t,1}$, $y_{t,2}$, and $y_{t,3}$. The vector autoregressive model of order 1, denoted as VAR(1), is as follows:

$$y_{t,1} = \alpha_1 + \rho_{11} y_{t-1,1} + \rho_{12} y_{t-1,2} + \rho_{13} y_{t-1,3} + \varepsilon_{t,1}$$
$$y_{t,2} = \alpha_2 + \rho_{21} y_{t-1,1} + \rho_{22} y_{t-1,2} + \rho_{23} y_{t-1,3} + \varepsilon_{t,2}$$
$$y_{t,3} = \alpha_3 + \rho_{31} y_{t-1,1} + \rho_{32} y_{t-1,2} + \rho_{33} y_{t-1,3} + \varepsilon_{t,3}$$

Vector autoregressive (VAR) models …

- In general, we can define VAR(1) model with *n* variables using matrix representation as below.

$$Y_t = \mathrm{A} + \mathrm{P}Y_{t-1} + E_t$$

- where $Y_t = \begin{bmatrix} y_{t,1} \\ \vdots \\ y_{t,n} \end{bmatrix}$, $\mathrm{A} = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix}$, $\mathrm{P} = \begin{bmatrix} \rho_{11} & \cdots & \rho_{1n} \\ \rho_{21} & \cdots & \rho_{2n} \\ \vdots & \vdots & \vdots \\ \rho_{n1} & \cdots & \rho_{nn} \end{bmatrix}$, $Y_{t-1} = \begin{bmatrix} y_{t-1,1} \\ \vdots \\ y_{t-1,n} \end{bmatrix}$, and

$E_t = \begin{bmatrix} \varepsilon_{t,1} \\ \vdots \\ \varepsilon_{t,n} \end{bmatrix}$

- In general, for a VAR(p) model, the first p lags of each variable in the system would be used as regression predictors for each variable.

Python programming example

VAR model

Granger causality

- **Granger causality** is a statistical concept of causality that is based on prediction.

- According to Granger causality, if a signal X1 *Granger-causes* or *G-causes* a signal X2, then past values of X1 should contain information that helps predict X2 above and beyond the information contained in past values of X2 alone.

- Its mathematical formulation is **based on linear regression** modeling of stochastic processes.

- More complex extensions to nonlinear cases exist, however these extensions are often more difficult to apply in practice.

- *Granger causality* (or *G-causality*) was developed in 1960s and has been widely used in economics since the 1960s. However it is only within the last few years that applications in **neuroscience** have become popular.

Granger causality …

- The basic Granger causality definition is quite simple:
  - Suppose that we have **three terms, $\boldsymbol{X_t}$, $\boldsymbol{Y_t}$, and $\boldsymbol{W_t}$**, and that we first attempt to forecast X_{t+1} using past terms of X_t and W_t.
  - We then try to forecast X_{t+1} using past terms of X_t, Y_t, and W_t. If the second forecast is found to be more successful, according to standard cost functions, then **the past of $\boldsymbol{Y}$ appears to contain information helping in forecasting $\boldsymbol{X_{t+1}}$ that is not in past $\boldsymbol{X_t}$ or $\boldsymbol{W_t}$**.
  - Thus, Y_t would **Granger cause** X_{t+1} if
    - Y_t occurs before X_{t+1}, and
    - it contains information useful in forecasting X_{t+1} that is not found in a group of other appropriate variables.

Granger causality …

- As a simple example, suppose we want to predict the value of X_t using an **AR(3)** model, like the model below

$$X_t = \alpha + \rho_1 X_{t-1} + \rho_3 X_{t-3} + \varepsilon_t$$

- Now, we want to increase the accuracy of the model above by introducing past values of Y_t. Suppose the new model is

$$X_t = \alpha + \rho_1 X_{t-1} + \rho_3 X_{t-3} + \varphi_3 Y_{t-3} + \varphi_5 Y_{t-5} + \varepsilon_t$$

- Now we should apply the **t-test** on each of the lags i of Y_t to see if each of these lags has any contribution to the result by itself or not.

- After that, we should apply the **F-test** to see which combinations of the lagged values of Y_t that passed the t-test is a most suitable one and can be help to accurately predict X_t.

- If **none of the lags/combinations could pass both the t-test and F-test**, then we say that $\boldsymbol{Y_t}$ **does not Granger causes** $\boldsymbol{X_t}$.

Python programming example

Granger causality

Python programming example

Full example (undo transformation)

Python programming example

Full example (model selection)