Time series prediction

Prakarsh Yadav
11755 - Machine Learning for Signal Processing

Contents:

- Introduction to time series and forecasting
- Stationarity and Wold Representation
- Autoregressive and Moving Average processes
- Autoregressive moving average processes and forecasting
- Non-stationary processes

Contents:

What is a time series?

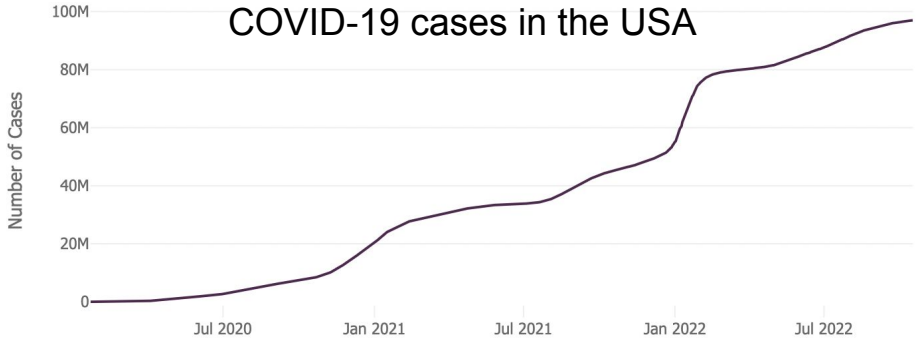
Time series: A collection of data which is indexed in time order

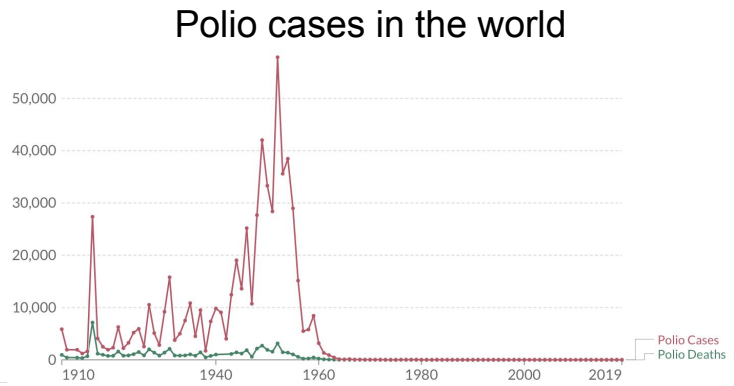
Any measurement we make at timed intervals can be considered as a time series

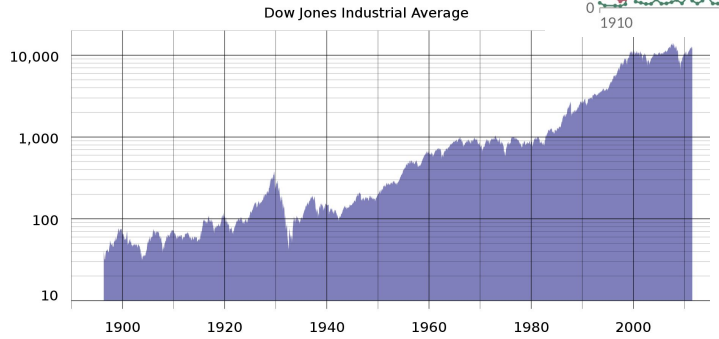
But given a time series, What can we say about it?

- Increasing?
- Decreasing?

  Can we forecast?







Time series and forecasting

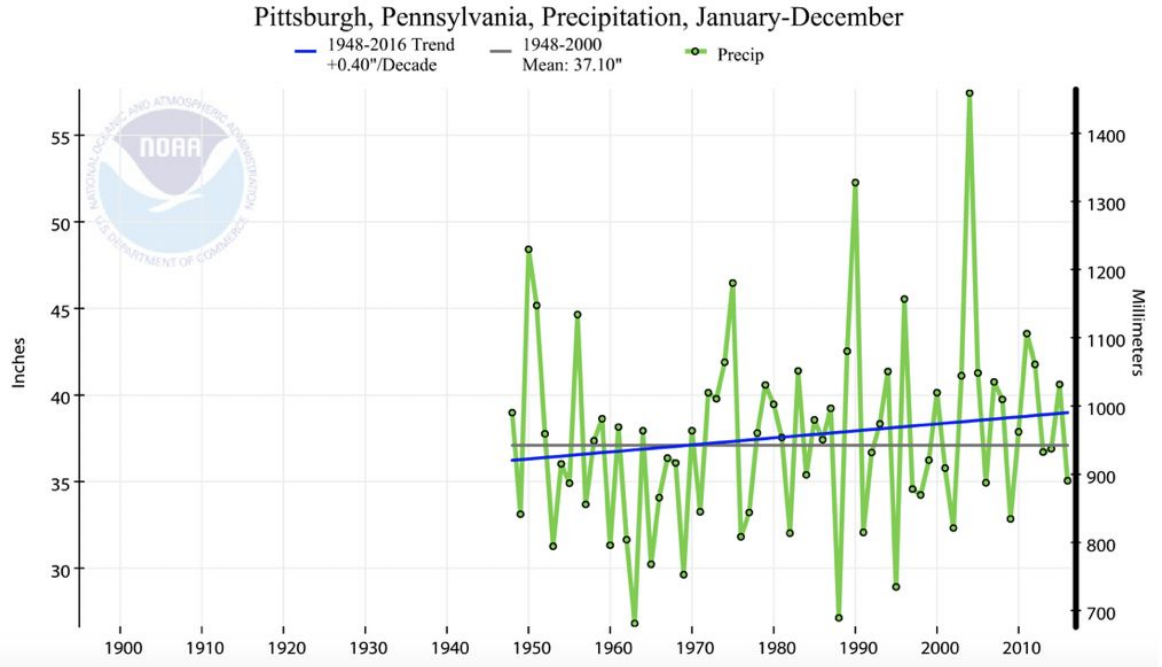
A time series can comprise of any measurements:

- Daily high/low temperatures
- Stock prices
- Weekly precipitation

This leads us to …

Given a time series, can we predict the most likely next measurement?

- **Forecasting!**



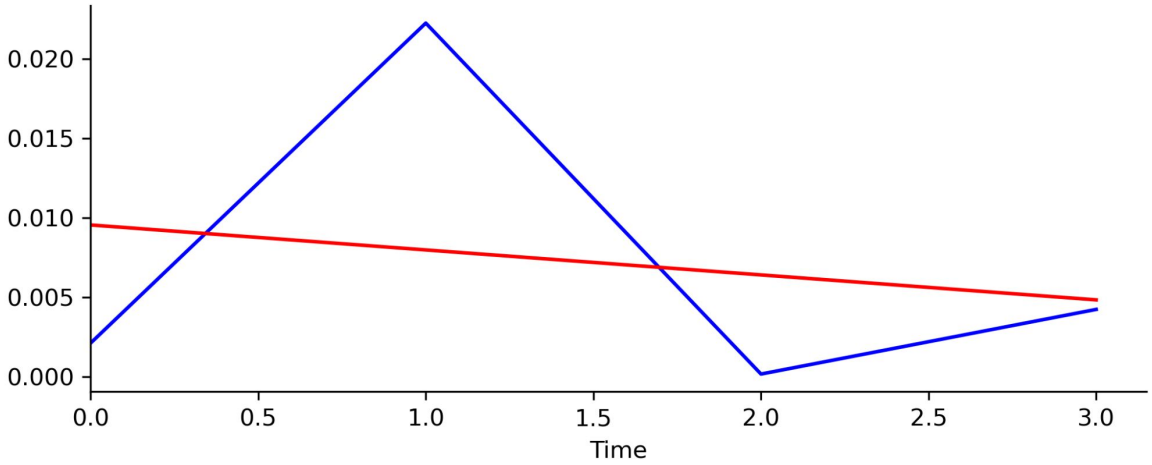
Forecasting

Say we just started recording a time series and we assume it has a linear relationship with time

We got 4 measurements, and fit a line to it

We are just getting started!

Slope: -0.00156



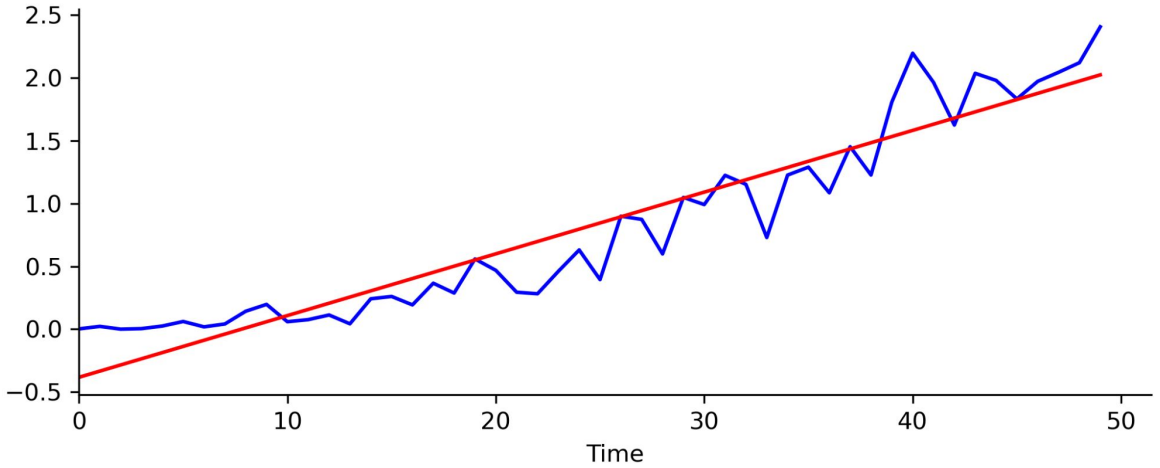
Forecasting

We are confident in our linear model of the time series

We got ambitious and collected next 50 measurements

The slope changed but still not a bad fit!

Slope: 0.04912



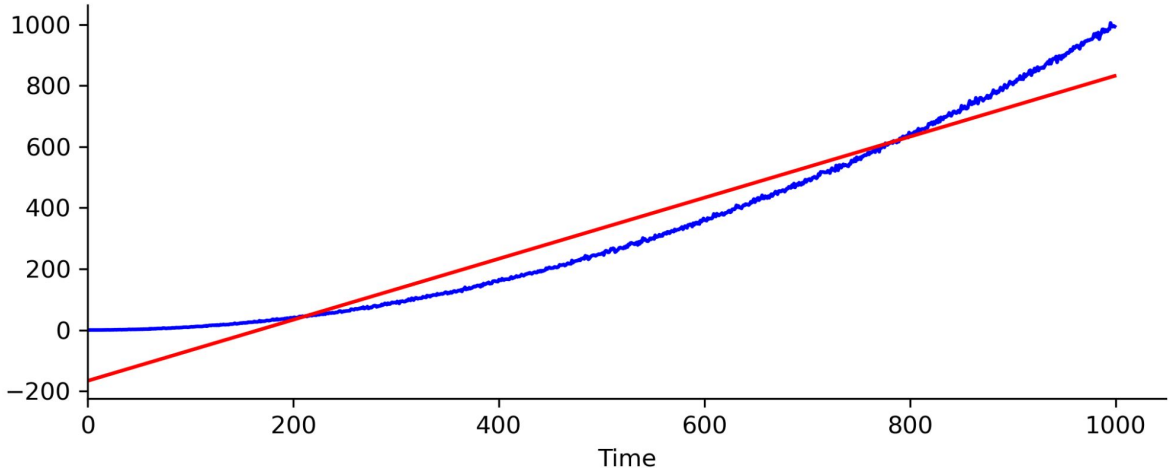
Forecasting

Like diligent grad students we get started on data collection!

We collect 1000 measurements!

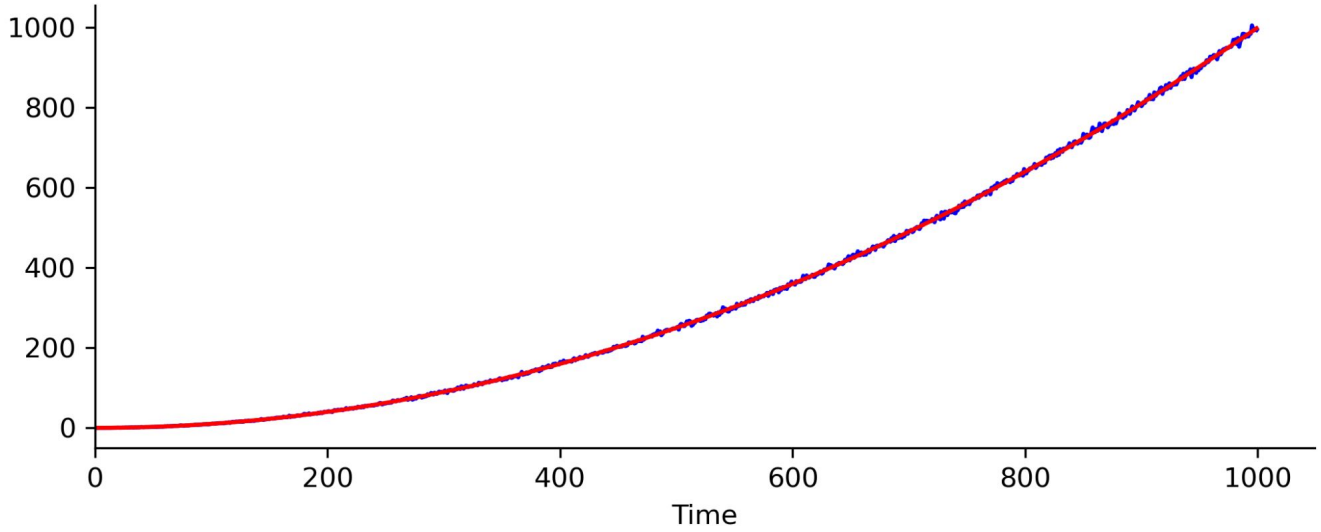
Maybe linear is not a valid assumption, what if we assume it quadratic?

Slope: 0.9992



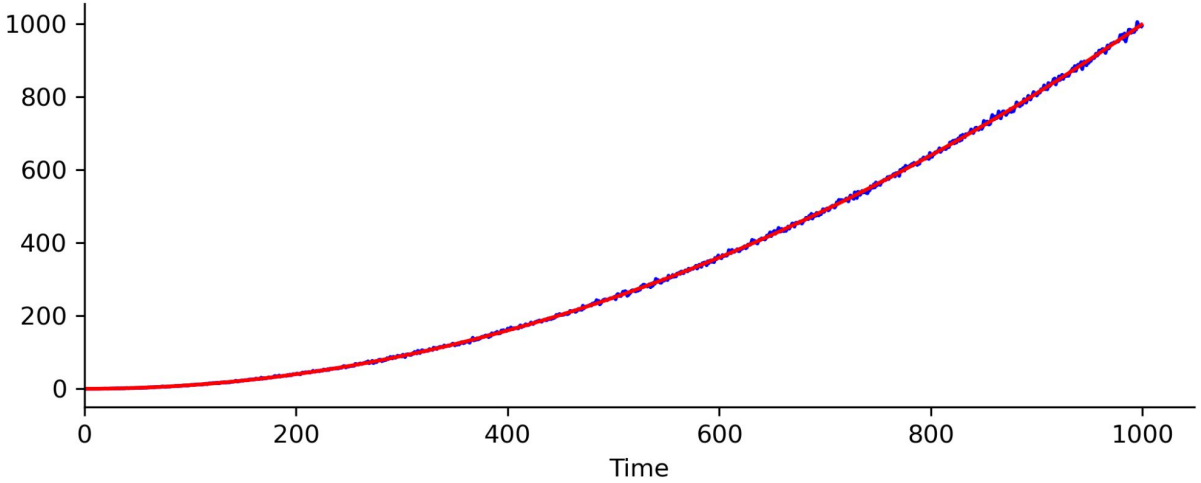
Forecasting

Quadratic assumption is looking good now and we are even happier!



Forecasting

Quadratic assumption is looking good now and we are even happier!

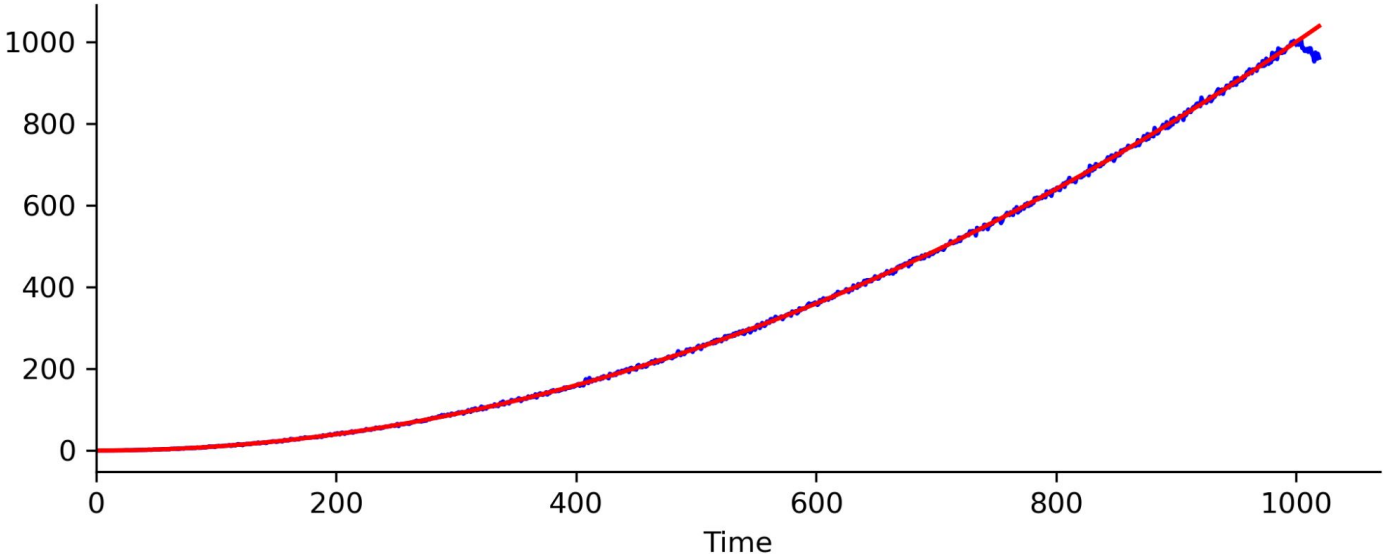


Now we are getting tired of data collection.

With this understanding, can we forecast the next measurements?

Forecasting

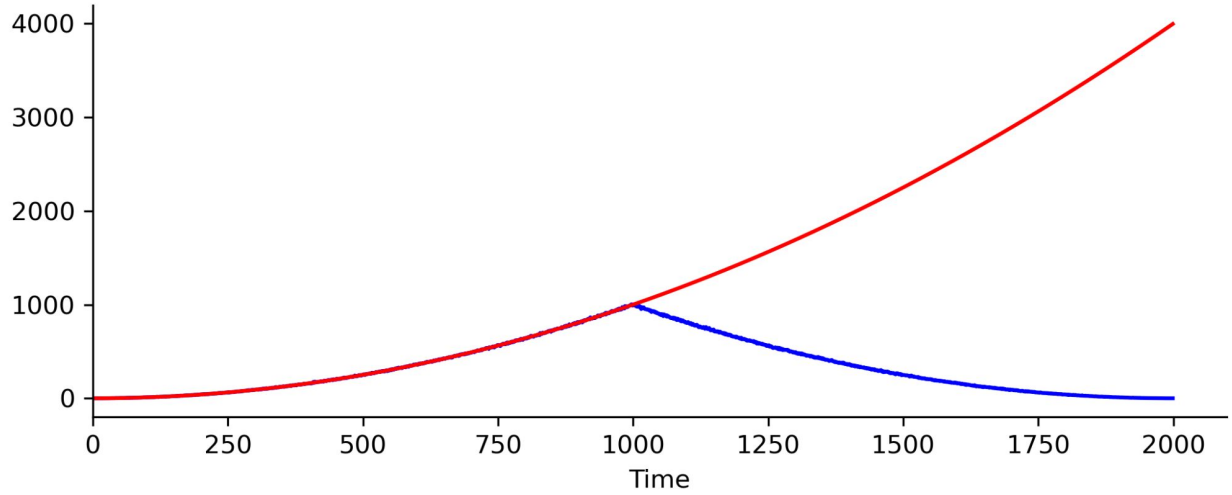
We predict the next 20 values but as a validation we also collect measurements



It seems like we are maybe getting an outlier, but we have faith in our model and well continue with it

Forecasting

We get ambitious and predict the next 1000 values

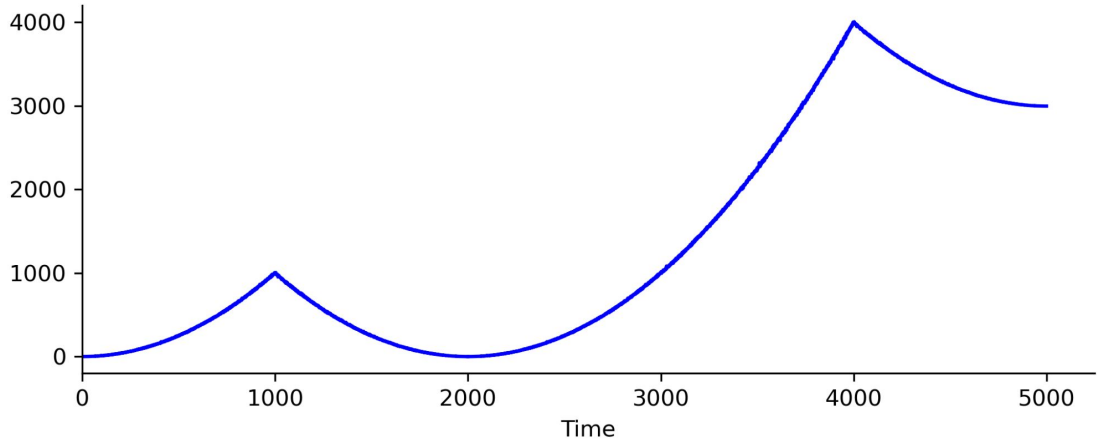


Looks like we made an oopsie! We did not expect this to have happened.

Forecasting

Now we dial down out predictions and **measure a lot more** more before forecasting

So the time series is doing its own thing and neither linear nor quadratic models would work!



In forecasting, our assumptions regarding the time series are critical!

We will now discuss how to make valid assumptions and where they will hold true!

Contents:

Wouldn’t it be nice if the properties of time series do not change?

Assume: The time series we are measuring is being sampled from a distribution whose parameters do not change with time.

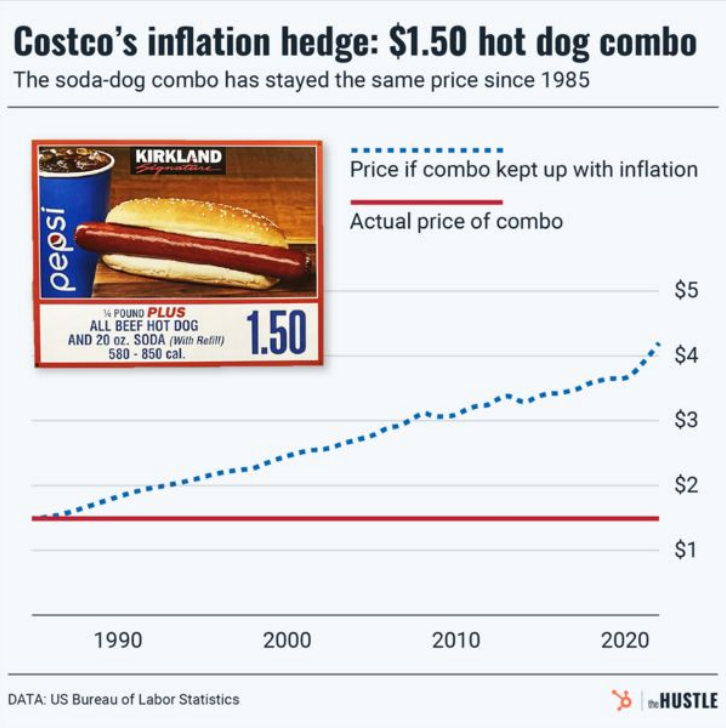
Then, the process can be considered to be “**stationary**” in time

Parameters such as **mean** and **variance** are not a function of time

Assuming that the time series is stationary confers some advantages:

- We do not violate any statistical analysis assumptions
- Law of large number and central limit theorem still apply
- Easier to model and forecast such processes

Consider Costco hot dog pricing, very likely it is going to be $1.5 for tomorrow



Strictly stationary time series

Let $\{X_t\}$ be a time series, and Fx is the cumulative distribution function of $\{X_t\}$

The $\{X_t\}$ is strictly stationary if

$$\mathbf{Fx}(x_{t1}, \ldots, x_{tn}) = \mathbf{Fx}(x_{t1+\tau}, \ldots, x_{tn+\tau})$$

For all τ, t_1, $t_n \in \mathbb{R}$ and for all $n \in \mathbb{N}$

Then the **mean** and **variance** of the time series will not change with time!

Examples,

Constant function; $X_t = Y$, for all t; $X_t = \cos(t+Y)$, for all $t \in \mathbb{R}$

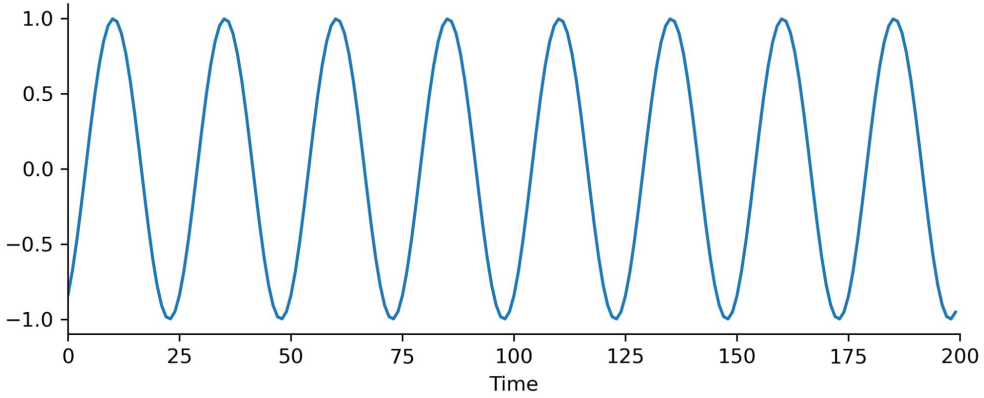
Is it useful? (a question for later)

- How far from reality is this? (Very much so!)

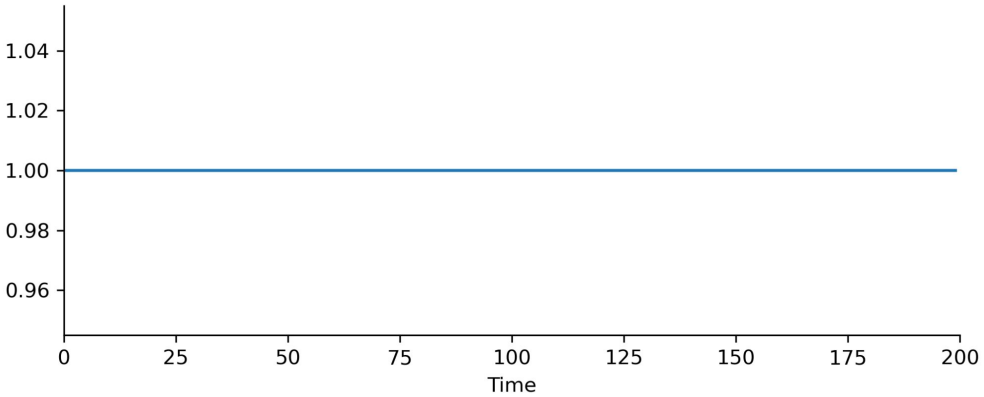
Strictly stationary time series

Examples:

$X_t = \cos(t+10)$



$X_t = Y$; $Y = 1$



Very restrictive view of the world!

N^{th} order stationary time series

Let $\{X_t\}$ be a time series, and Fx is the cumulative distribution function of $\{X_t\}$

The $\{X_t\}$ is **N^{th} order stationary** if

$$\mathbf{Fx(x_{t1}, \ldots, x_{tn}) = Fx(x_{t1+\tau}, \ldots, x_{tn+\tau})}$$

For all τ, t_1, $t_n \in \mathbb{R}$ and for all **$n \in \{1, \ldots, N\}$**

Then the **mean** and **variance** of the time series will not change **within the time window**!

Another view: **the time series is stationary within the period we observe**

Note: some definitions in literature refer to PDF instead of CDF

Going further: Covariance stationary processes

A weaker constraint on stationarity requirement

A time series $\{X_t\}$ is Covariance stationary if

$$\mathbf{E(X_t) = \mu}$$

$$\mathbf{Var(X_t) = \sigma^2_X}$$

$$\mathbf{Cov(X_t, X_{t+\tau}) = \gamma(\tau)}$$

Note: All are **independent of t**

Advantages?

- We assume covariance only depends on τ and is not a constant
- Model more complex time series

Autocorrelation function (ACF)

Autocorrelation Function (ACF) of a time series is defined as

$$\mathbf{R_{t,t+\tau} = Cov(X_t, X_{t+\tau}) \left[\sqrt{Var(X_t)}\sqrt{Var(X_{t+\tau})}\right]^{-1}}$$

Useful properties:

- $\mathbf{R_{t,t} = Var(X_t)}$
- For covariance stationary processes, $\mathbf{R_{t,t+\tau} = R_{t-\tau,\, t}}$, i.e. **ACF is a function of τ only**
- For covariance stationary processes, $\mathbf{Cov(X_t, X_{t+\tau}) = Cov(X_{t-\tau}, X_t)}$

Poll 1: which of the following time series is stationary?

- A coin toss of 400 trials
- A biased coin toss (always heads) with 400 trials
- Temperature at 4 PM for Shadyside
- Random walk

https://tinyurl.com/mlsp23-1024-01

Poll 1: which of the following time series is stationary?

- A coin toss of 400 trials
- A biased coin toss (always heads) with 400 trials
- Temperature at 4 PM for Shadyside
- Random walk

Why stationarity?

- Stationary time series are convenient to work with
  - Easier parameter discovery and optimization
  - Existing literature and methodologies that readily translates
- A reasonable compromise between what is (real data) and what we can work with (our model)
- An interesting perspective,
  - The more informed and less constrained stationary model we have, the less deviation it will have from the real series
  - Example, in case of stock prices,
    - **Case 1**: The future price is the running mean of previous price and plus stochastic noise
    - **Case 2**: The future price is multivariate (depends on previous prices, profits, geopolitical scenario, etc.) and what is still unknown is stochastic noise
  - Are either of the models perfectly accurate? **NO!**
  - Is one better than the other? **YES!**

Wold Representation theorem

Any zero-mean covariance stationary time series $\{X_t\}$ can be decomposed as the sum of two time series, one deterministic and one stochastic

$$\mathbf{X_t = V_t + S_t}$$

Here $\{V_t\}$ a **linear deterministic process**,

- A linear combination of previous values of V_t, with constant coefficients

S_t is an **infinite moving average** process of error terms (stochastic)

- $\mathbf{S_t = \sum^{\infty}_{i=0} \psi_i \eta_{t-i}}$
- ψ is the moving average weights, and $\mathbf{\{\eta_t\}}$ is the linear white noise

Named after Herman Wold

Wold Representation theorem: Properties

$$X_t = V_t + S_t$$

$$S_t = \sum_{i=0}^{\infty} \psi_i \eta_{t-i}$$

- Weights are stable, $\sum_{i=0}^{\infty} \psi_i^2 < \infty$ (square summable)
- Conventionally, $\psi_0 = 1$
- White noise $\{\eta_t\} \sim$ **iid** $N(0, \sigma^2)$ has, $E(\eta_t) = 0$, $E(\eta_t^2) = \sigma^2$, $E(\eta_t \eta_s) = 0$,
- White noise $\{\eta_t\}$ is uncorrelated with $\{V_t\}$, $E(\eta_t V_s) = 0$

Wold Representation theorem

Why is this representation useful?

- We just modelled a time series, $\mathbf{X}_t$, as a sum of two linear variables!
  - We can estimate the variables!
- It forms the basis of moving average (MA) and autoregressive (AR) models to explain time series (discussed later)
- It is a MA(∞) and AR(∞) representation

Challenge:

- It needs infinite number of parameters to represent the time series
  - But they decay rapidly in practice

Contents:

- Introduction to time series and forecasting
- Stationarity and Wold Representation
- Autoregressive and Moving Average processes
- Autoregressive moving average processes and forecasting
- Non-stationary processes

A slight detour to Lag operator

The lag operator L() shifts a time series back by one time increment

- Basically, a short hand to manipulate time series data

For $\{X_t\}$, $L(X_t) = X_{t-1}$

We can also have different orders of lag operator,

Or recursive application $L(L(X_t) = L^2(X_t)$

$L^0(X_t) = X_t$; $L^1(X_t) = X_{t-1}$; $L^2(X_t) = X_{t-2}$; …; $L^n(X_t) = X_{t-n}$;

Inverse also exists,

$L^{-n}(X_t) = X_{t+n}$

Wold Representation theorem with Lag operator

Defining some notation and setup

$$X_t = V_t + S_t$$

$$X_t = \sum_{i=0}^{\infty} \psi_i \eta_{t-i} + V_t$$

$$X_t = \sum_{i=0}^{\infty} \psi_i \mathbf{L^i(\eta_t)} + V_t$$

$$X_t = \psi(L)\eta_t + V_t,$$

where $\psi(\mathbf{L}) = \sum_{i=0}^{\infty} \psi_i \mathbf{L^i}$

Wold Representation theorem with Lag operator

Now, assume $\psi(L)$ is invertible

$$\psi^{-1}(L) = \sum^{\infty}_{i=0} \psi^{*}_{i} L^{i}, \text{ s.t.}$$

$$\psi^{-1}(L)\, \psi(L) = I = L^{0}$$

Also, we assume $V_t = 0$, i.e. $X_t = X_t - V_t$

$$X_t = \psi(L)\eta_t$$

$$\psi^{-1}(L)X_t = \psi^{-1}(L)\psi(L)\eta_t$$

Then, if $\psi^{-1}(L)$ exists, $\{X_t\}$ is also invertible, and can be represented as

$$X_t = \left(\sum^{\infty}_{i=0} \psi^{*}_{i} X_{t-i}\right) + \eta_t$$

Note: This is an **autoregressive representation** of X_t

What if the time series has persistence?

The data future measurements are dependent on the previous measurement, s.t. they can be described as a function of previous value

$$\text{Recall, } X_t = (\sum_{i=0}^{\infty} \psi^*_i X_{t-i}) + \eta_t$$

$$\mathbf{X_t = \psi X_{t-1} + \eta_t,}$$

$$\text{where } \eta_t \sim \text{iid } N(0,\sigma^2)$$

This is an autoregressive process of first order AR(1)

- Recall Wold representation is AR(∞)
- Think of Markov processes and their orders!

What if the time series has persistence on p values?

The data future measurements are dependent on the previous measurements, s.t. they can be described as a function of previous values

$$\text{Recall, } X_t = (\sum^{\infty}_{i=0} \psi^*_i X_{t-i}) + \eta_t$$

$$\mathbf{X_t = \psi_1 X_{t-1} + \psi_2 X_{t-2} + \ldots + \psi_p X_{t-p} + \eta_t,}$$

$$\text{where } \eta_t \sim \text{iid } N(0,\sigma^2)$$

This is an autoregressive process of order p, AR(p)

- It is a linear combination of previous values (**explanatory variables**)

AR(1) process representation

Solving AR with recursion:

$$X_t = \psi X_{t-1} + \eta_t,$$

$$X_t = \psi(\psi X_{t-2} + \eta_{t-1}) + \eta_t,$$

$$X_t = \psi^{j+1} X_{t-(j+1)} + \psi^j \eta_{t-j} + \ldots + \psi^2 \eta_{t-2} + \psi \eta_{t-1} + \eta_t,$$

Note:

- if $|\psi| < 1$, then $\psi^{j+1} X_{t-(j+1)} \to 0$, for large enough j (important relation with MA(∞))
- if $|\psi| > 1$, then $\psi^{j+1} X_{t-(j+1)} \to \infty$, for large enough j (not summable, hence non-stationary)

AR(1) Properties $X_t = \psi X_{t-1} + \eta_t$,

Given an AR(1), $X_t = \psi^{j+1} X_{t-(j+1)} + \psi^j \eta_{t-j} + \ldots + \psi^2 \eta_{t-2} + \psi \eta_{t-1} + \eta_t$,

$E[X_t] = \mu$

Now, if data is mean centered, $X_t = X_t - \mu$

$Var[X_t] = E[X_t - E[X_t]]^2 = E[\eta_t + \psi \eta_{t-1} + \psi^2 \eta_{t-2} + \ldots + \psi^j \eta_{t-j}]^2$

$Var[X_t] = Var[\eta_t] + \psi^2 Var[\eta_t] + \psi^4 Var[\eta_t] + \ldots = (1 + \psi^2 + \psi^4 + \ldots)\sigma^2$

$Var[X_t] = \sigma^2 (1 - \psi^2)^{-1}$, as, $\eta_t \sim$ iid $N(0, \sigma^2)$

AR(1) Properties $X_t = \psi X_{t-1} + \eta_t$,

Given an AR(1), $X_t = \psi^{j+1} X_{t-(j+1)} + \psi^j \eta_{t-j} + \ldots + \psi^2 \eta_{t-2} + \psi \eta_{t-1} + \eta_t$,

$Cov[X_t, X_{t-1}] = E[(X_t - E[X_t])\,(X_{t-1} - E[X_{t-1}])]$; $\quad X_t = X_t - \mu$

$Cov[X_t, X_{t-1}] = E[(\eta_t + \psi\eta_{t-1} + \psi^2\eta_{t-2} + \ldots)\;(\eta_{t-1} + \psi\eta_{t-2} + \psi^2\eta_{t-3} + \ldots)]$

$Cov[X_t, X_{t-1}] = (\psi^1 + \psi^3 + \psi^5 + \ldots)\sigma^2 = \psi(1 + \psi^2 + \psi^4 + \ldots)\sigma^2$

$Cov[X_t, X_{t-1}] = \psi\sigma^2(1 - \psi^2)^{-1} = \psi Var[X_t]$

AR(2) process representation

Given a AR(2), $X_t = \psi_1 X_{t-1} + \psi_2 X_{t-2} + \eta_t$,

We define $$Z_t = \begin{bmatrix} X_t \\ X_{t-1} \end{bmatrix}, v_t = \begin{bmatrix} \eta_t \\ 0 \end{bmatrix}, \Gamma = \begin{bmatrix} \psi_1 & \psi_2 \\ 1 & 0 \end{bmatrix}$$

Such that, $$Z_t = \Gamma Z_{t-1} + v_t$$

$$\begin{bmatrix} X_t \\ X_{t-1} \end{bmatrix} = \begin{bmatrix} \psi_1 & \psi_2 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} X_{t-1} \\ X_{t-2} \end{bmatrix} + \begin{bmatrix} \eta_t \\ 0 \end{bmatrix}$$

This is matrix representation of AR(2)

Also, referred to as Yule-Walker equations

AR(2) process stationarity test

To find roots of the equation $\begin{bmatrix} X_t \\ X_{t-1} \end{bmatrix} = \begin{bmatrix} \psi_1 & \psi_2 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} X_{t-1} \\ X_{t-2} \end{bmatrix} + \begin{bmatrix} \eta_t \\ 0 \end{bmatrix}$

We can compute eigenvalues (m_1,m_2) of $\Gamma = \begin{bmatrix} \psi_1 & \psi_2 \\ 1 & 0 \end{bmatrix}$

Which satisfy the characteristic equation $\mathbf{x^2 - \psi_1 x - \psi_2 = 0}$

Then, $$m_1, m_2 = \frac{(\psi_1 \pm \sqrt{\psi_1^2 + 4\psi_2})}{2}$$

AR(2) process stationarity test

For stationarity eigenvalues have less than 1 absolute value,

i.e. $|m_1|,|m_2| < 1$

$$m_1, m_2 = \frac{(\psi_1 \pm \sqrt{\psi_1^2 + 4\psi_2})}{2}$$

Which will be the case if,

$$\psi_1 + \psi_2 < 1$$

$$-\psi_1 + \psi_2 < 1$$

$$\psi_2 > -1$$

AR(2) process stationarity test: an alternate view

Write the characteristic equation as $\mathbf{1 - x^2 - \psi_1 x - \psi_2 = 0}$, Hamilton (1994)

In this representation, the roots of this equation must lie **outside the unit circle,** for the AR process to be stationary

The roots are the inverse of ψ_1 and ψ_2,

so their magnitude is >1 to ensure $|\psi_1|$ and $|\psi_2| < 1$

Hamilton, J. D.: Time Series Analysis, Princeton University Press (1994).

What if the error terms (η) varies with time?

Then we have a Moving Average (MA) process

$$\mathbf{X_t = \mu + \eta_t + \theta\eta_{t-1}}, \text{ MA(1)}$$

Here, **μ** is a constant, $\boldsymbol{\eta_t}$ and $\boldsymbol{\eta_{t-1}}$, are iid $N(0,\sigma^2)$

Additionally,

$E[X_t] = E[\mu + \eta_t + \theta\eta_{t-1}] = E[\mu] + E[\eta_t] + E[\theta\eta_{t-1}] = \boldsymbol{\mu}$

$Var[X_t] = E[X_t - E[X_t]]^2 = E[(\mu + \eta_t + \theta\eta_{t-1}) - \mu]^2 = E[\eta_t]^2 + E[\theta\eta_{t-1}]^2 + 2\theta E[\eta_t\eta_{t-1}]$

$Var[X_t] = \sigma^2 + \theta^2\sigma^2 + 0$

$Var[X_t] = (1+\theta^2)\sigma^2$

Note: Both quantities are independent of t

MA Covariance

We have an MA(1) process, $\mathbf{X_t = \mu + \eta_t + \theta\eta_{t-1}}$,

$Cov[X_t, X_{t-1}] = E[(X_t - E[X_t])(X_{t-1} - E[X_{t-1}])] = E[(\eta_t + \theta\eta_{t-1})(\eta_{t-1} + \theta\eta_{t-2})]$

$Cov[X_t, X_{t-1}] = E[\eta_t\eta_{t-1}] + \theta E[\eta_{t-1}^2] + \theta E[\eta_t\eta_{t-2}] + \theta^2 E[\eta_{t-1}\eta_{t-2}] = 0 + \theta\sigma^2 + 0 + 0$

$\mathbf{Cov[X_t, X_{t-1}] = \theta\sigma^2}$

Generally, for MA(q) process, $\mathbf{X_t = \mu + \theta_0\eta_t + \theta_1\eta_{t-1} + \ldots + \theta_q\eta_{t-q}}$,

$Cov[X_t, X_{t-q}] = E[(X_t - E[X_t])(X_{t-q} - E[X_{t-q-1}])] = E[(\eta_t + \theta\eta_{t-1})(\eta_{t-q} + \theta\eta_{t-q-1})]$

$\mathbf{Cov[X_t, X_{t-q}] = 0}$, for $q>1$

Note: With constant $E[X_t]$, $Var[X_t]$ and $Cov[X_t, X_{t-q}]$ independent of t,

MA processes are covariance stationary!

Call back to Wold Representation Theorem

$$X_t = V_t + S_t$$

Here $\{V_t\}$ a **linear deterministic process**,

- A linear combination of of previous values of V_t, with constant coefficients

S_t is an **infinite moving average** process of error terms (stochastic)

- $S_t = \sum^{\infty}_{i=0} \psi_i \eta_{t-i}$

Now we know what AR(p) and MA(q) processes look like,

We can appreciate that WRT represents time series as a **linear combination of MA(∞)** and **AR(∞) processes**

Poll 2: Mark true statements

- Lag operator requires that the time series is modeled by linear variables
- For Wold Representation Theorem, moving average weights are constrained
- Stationary time series can be modeled through the lag operator
- For AR processes, if $|\psi|$ >1, then they it is stationary

https://tinyurl.com/mlsp23-1024-02

Poll 2: Mark true statements

- Lag operator requires that the time series is modeled by linear variables
- For Wold Representation Theorem, moving average weights are constrained
- Stationary time series can be modeled through the lag operator
- For AR processes, if $|\psi|$ >1, then they it is stationary

Contents:

- Introduction to time series and forecasting
- Stationarity and Wold Representation
- Autoregressive and Moving Average processes
- Autoregressive moving average processes and forecasting
- Non-stationary processes

Auto Regressive Moving Average (ARMA) processes

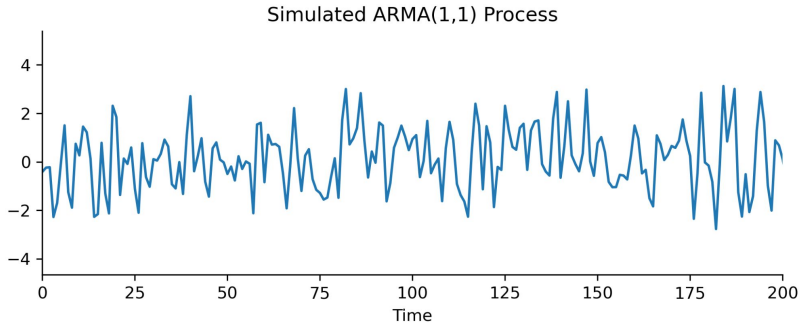
A linear combination of AR and MA processes

Simplest ARMA,

ARMA(1,1), $\mathbf{X_t = \psi X_{t-1} + \eta_t + \theta \eta_{t-1}}$,

As lag operator, $\mathbf{\psi(L) X_t = \theta(L) \eta_t}$,

Where, $\mathbf{\psi(L) = 1 - \sum^{p}_{i=0} \psi_i L^i}$, and $\mathbf{\theta(L) = 1 - \sum^{q}_{i=0} \theta_i L^i}$,



Note: With these definitions we can construct ARMA(2,1) or generally ARMA(p,q)

ARMA (p,q)

A process is ARMA(p,q) if it is **autoregressive with order p** and **moving average with order q**

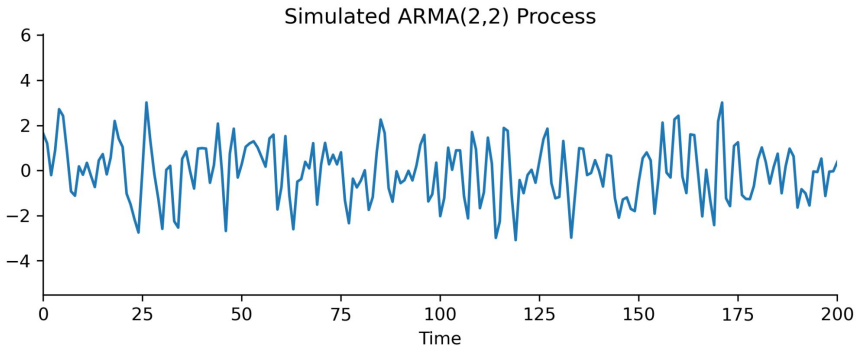
ARMA(p,q), $\mathbf{X_t = \psi_1 X_{t-1} + \psi_2 X_{t-2} + \ldots + \psi_p X_{t-p} + \eta_t + \theta_1 \eta_{t-1} + \theta_2 \eta_{t-2} + \ldots + \theta_q \eta_{t-q}}$,

As lag operator, $\mathbf{\psi(L) X_t = \theta(L) \eta_t}$,

Where, $\mathbf{\psi(L) = 1 - \sum^{p}_{i=0} \psi_i L^i}$, and $\mathbf{\theta(L) = 1 + \sum^{q}_{i=0} \theta_i L^i}$,

Wold decomposition of ARMA(p,q)

$\mathbf{X_t = [\psi(L)]^{-1} \theta(L) \eta_t}$,



Recall ARMA(1,1) process

$$(1 - \psi(L))X_t = (1 + \theta(L))\eta_t,$$

We can approximate the lag operator inverse as,

$$X_t = [(1 - \psi(L))]^{-1}(1 + \theta(L))\eta_t$$

$$X_t = [(1 - \psi(L))]^{-1}\eta_t + [(1 - \psi(L))]^{-1}\theta_1\eta_{t-1}$$

Expressing this as a Geometric Progression,

$$X_t = \sum_{j=0}^{\infty}(\psi L)^j\eta_t + \theta\sum_{j=0}^{\infty}(\psi L)^j\eta_{t-1}$$

$$X_t = \eta_t + \sum_{j=1}^{\infty}\psi^j\eta_{t-j} + \theta\sum_{j=1}^{\infty}\psi^{j-1}\eta_{t-j}$$

$$X_t = \eta_t + \sum_{j=1}^{\infty}(\psi^j + \theta\psi^{j-1})\eta_{t-j}$$

if $|\psi| < 1$, weights are summable and **Var[X_t] and Cov[X_t ,X_{t-1}] are finite**

AR model parameter estimation

Assume AR(1), with μ mean; $\mathbf{X_t = \mu + \psi X_{t-1} + \eta_t}$, alternatively, $\mathbf{\eta_t = X_t - \mu - \psi X_{t-1}}$

where, $\mathbf{\eta_t \sim N(0,\sigma^2)}$,

MLE formulation is: Find θ*, such that L(θ|X) is maximized

$$L(\theta|X) = \prod_{i=1}^{N} p(x_i|\theta)$$

Alternatively we can minimize the negative log likelihood function (NLL)

$$NLL = -\log(L(\theta|X))$$

AR model parameter estimation

For $\mathbf{\eta_t = X_t - \mu - \psi X_{t-1}}$ and $\mathbf{\eta_t \sim N(0,\sigma^2)}$,

We can estimate $p(X_1,...,X_t|X_0,\theta)$ as,

$p(X_1,...,X_t|X_0,\theta) = (2\pi\sigma^2)^{-T/2} \exp(-1/(2\sigma^2) * \sum^T_{t=1} \eta_t^2$

$p(X_1,...,X_t|X_0,\theta) = (2\pi\sigma^2)^{-T/2} \exp(-1/(2\sigma^2) * \sum^T_{t=1} (X_t - \mu - \psi X_{t-1})^2$

NLL is

$\mathbf{-\log(L(\theta|X)) = \tfrac{1}{2}(T*\log \sigma^2) + 1/(2\sigma^2)*\sum^T_{t=1} (X_t - \mu - \psi X_{t-1})^2 + const.}$

Note: This is quadratic in X and can be minimized by…

AR model parameter estimation

$-\log(L(\theta|X)) = \frac{1}{2}(T*\log \sigma^2) + 1/(2\sigma^2)*\sum_{t=1}^{T} (X_t - \mu - \psi X_{t-1})^2 + \text{const.}$

Note: This is quadratic in X and can be minimized by…

$$\begin{pmatrix} \psi' \\ \mu' \end{pmatrix} = \begin{pmatrix} T & \sum_{t=0}^{T-1} X_t \\ \sum_{t=0}^{T-1} X_t & \sum_{t=0}^{T-1} X_t^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{t=0}^{T-1} X_{t+1} \\ \sum_{t=0}^{T-1} X_t X_{t+1} \end{pmatrix}$$

$$\sigma'^2 = \frac{1}{T} \sum_{t=1}^{T} (X_t - \mu' - \psi' X_{t-1})^2$$

AR model parameter estimation

Rewriting,

$$\begin{pmatrix}\psi' \\ \mu'\end{pmatrix} = \begin{pmatrix} T & \sum_{t=0}^{T-1} X_t \\ \sum_{t=0}^{T-1} X_t & \sum_{t=0}^{T-1} X_t^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{t=0}^{T-1} X_{t+1} \\ \sum_{t=0}^{T-1} X_t X_{t+1} \end{pmatrix}$$

as,

$$\psi' = \frac{\sum_{t=0}^{T-1}(X_t - X')(X_{t-1}X'_+)}{\sum_{t=0}^{T-1}(X_t - X')^2}$$

$$\mu' = X'_+ + \psi' X'$$

where,

$$X' = \frac{1}{T}\sum_{t=0}^{T-1} X_t$$

$$X'_+ = \frac{1}{T}\sum_{t=0}^{T-1} X_{t+1}$$

MA model parameter estimation

Assume MA(1), with μ mean; $\mathbf{X_t = \mu + \eta_t + \theta\eta_{t-1}}$,, alternatively,$\mathbf{\eta_t = X_t - \mu - \theta\eta_{t-1}}$

The PDF of $\mathbf{X_t}$, is $p(X_t|X_{t-1},...,X_1, \eta_0=0, \theta) = \mathbf{(2\pi\sigma^2)^{-1/2} \exp(-\eta_t^2 / (2\sigma^2))}$

NLL is

$-\log(L(\theta|X,\eta_0=0)) = ½(T*\log \sigma^2) + 1/(2\sigma^2)*\sum^T_{t=1} \eta_t^2 + \text{const.}$

Forecasting with ARMA(p,q)

Problem statement: Given $\mathbf{X_{1:t} = X_1, \ldots X_t}$, what is the best estimate ($\mathbf{X^*_{t+1}}$) of $\mathbf{X_{t+1}}$?

Equivalent to, minimizing Mean Squared Error (MSE)

$$\min E((\mathbf{X^*_{t+1}} - \mathbf{X_{t+1}})^2)$$

To **prove the minimization of MSE is indeed forecasting,**

Say Z is any random variable measurable with respect to the information

set generated by $\mathbf{X_{1:t}}$, then

$$E((X_{t+1} - Z)^2) = E((X_{t+1} - E(X_{t+1}) + E(X_{t+1}) - Z)^2)$$

$$E((X_{t+1} - Z)^2) = E((X_{t+1} - E(X_{t+1}))^2) + E((E(X_{t+1}) - Z)^2) + 2E((X_{t+1} - E(X_{t+1}))(E(X_{t+1}) - Z))$$

Forecasting with ARMA(p,q)

$E((X_{t+1}- Z)^2) = E((X_{t+1}- E(X_{t+1}))^2) + E((E(X_{t+1}) - Z)^2) + 2\mathbf{E((X_{t+1}- E(X_{t+1}))(E(X_{t+1}) - Z))}$

Here, $E((X_{t+1}- E(X_{t+1}))(E(X_{t+1}) - Z)) = E((X_{t+1}- E(X_{t+1}))(E(X_{t+1}) - Z))$

$= (E(X_{t+1}) - E(E(X_{t+1})))(E(X_{t+1}) - Z))$

$= 0*(E(X_{t+1}) - Z))$, as $E(E(X_{t+1})) = E(X_{t+1})$

then, $E((X_{t+1}- Z)^2) = E((X_{t+1}- E(X_{t+1}))^2) + E((E(X_{t+1}) - Z)^2)$

which is **minimum at, $Z = E(X_{t+1}) = E(X^*_{t+1})$**

Poll 3: All ARMA processes can be decomposed according to Wold representation

- True

- False

https://tinyurl.com/mlsp23-1024-03

Poll 3: All ARMA processes can be decomposed according to Wold representation

- True

- False

Contents:

- Introduction to time series and forecasting
- Stationarity and Wold Representation
- Autoregressive and Moving Average processes
- Autoregressive moving average processes and forecasting
- Non-stationary processes

How do we tell if a process is stationary?

We conduct statistical test: (Null Hypothesis)

Given AR(1), $X_t = \psi X_{t-1} + \eta_t$,

We assume, $\mathbf{H_0: \psi = 1}$ (unit root, non-stationarity) OR $\mathbf{H_1: |\psi| < 1}$ (stationarity)

This is, **Autoregressive Unit Root Test**

Additional Stationarity tests

Unit Root test (H_0: non-stationarity)

- Dickey-Fuller (DF) test - Dickey and Fuller (1979)
- Augmented Dickey-Fuller (ADF) test - Said and Dickey (1984)
- Unit Root (PP) test - Phillips and Perron (1988)
- Efficient Unit (ERS) Root Test - Elliot, Rothenberg, and Stock (2001)

Stationarity test (H_0: stationarity)

- KPSS test - Kwiatkowski, Phillips, Schmidt, and Shin (1992)

What if a process is non stationary?

Consider the example of Random Walk, $\mathbf{X_t = X_{t-1} + \eta_t}$,

where $\eta_t \sim N(0,\sigma^2)$

$Var[X_t] = Var[X_{t-1} + \eta_t] = Var[X_{t-1}] + \sigma^2 = Var[X_{t-2} + \eta_t] + \sigma^2 = Var[X_{t-3} + \eta_t] + 2\sigma^2$

$\mathbf{Var[X_t] = Var[X_0] + t\sigma^2}$

Why is this not stationary?

Violations of stationarity

A time series $\{X_t\}$ is Covariance stationary if

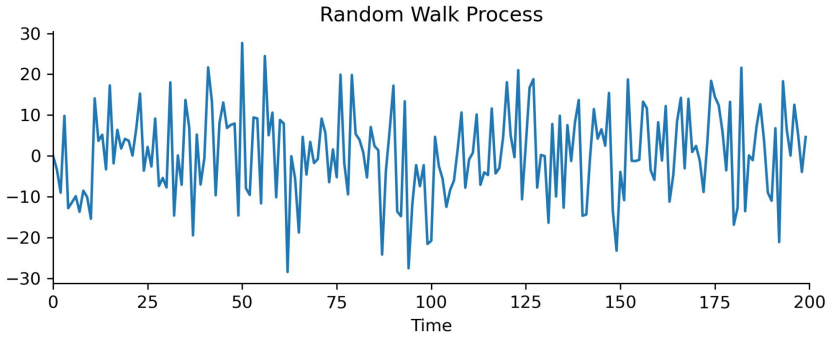
$$E(X_t) = \mu$$

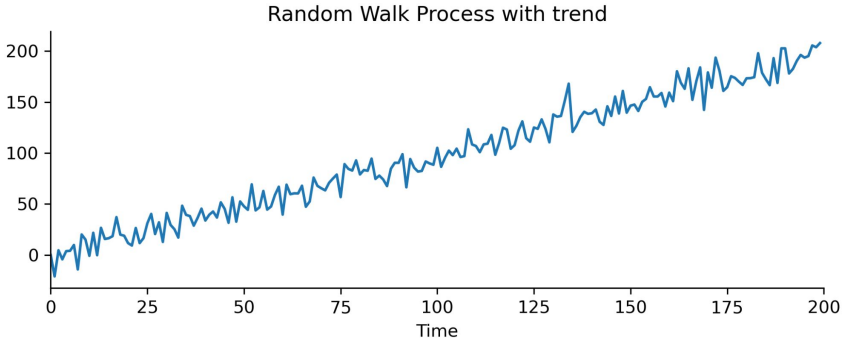
$$Var(X_t) = \sigma^2_X$$

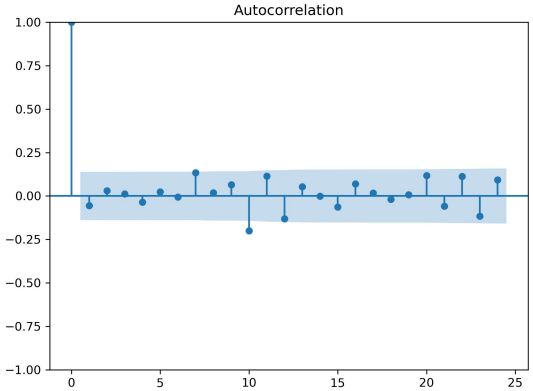
$$Cov(X_t, X_{t+\tau}) = \gamma(\tau)$$

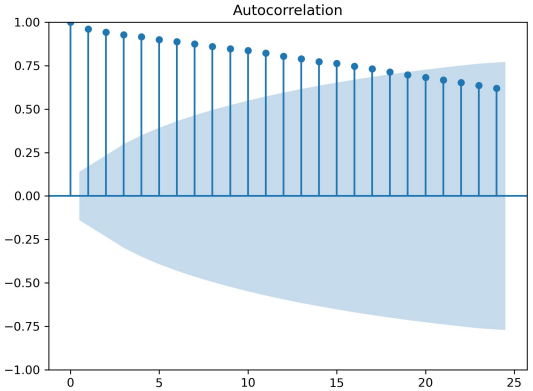
If any of these are a function of time, then process may be non-stationary!

Examples of non-stationary processes



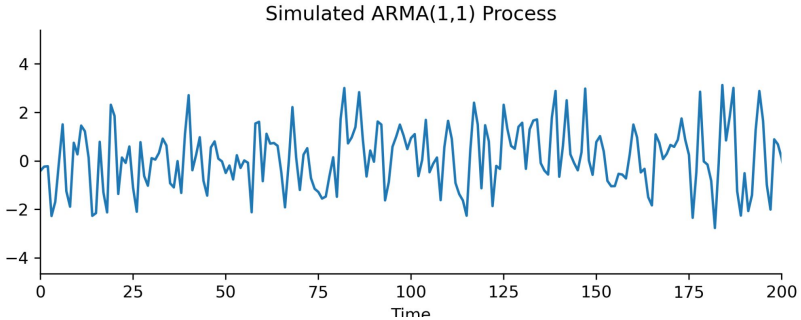


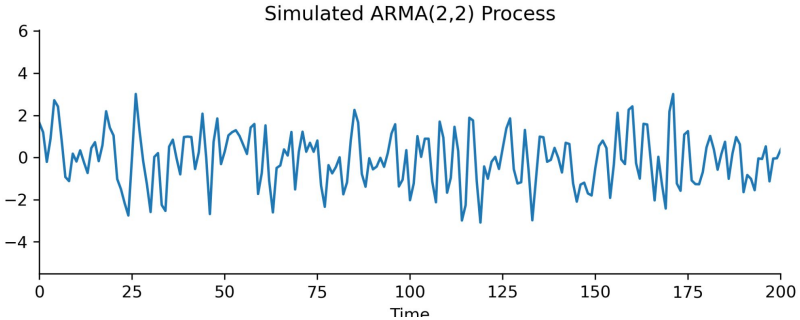


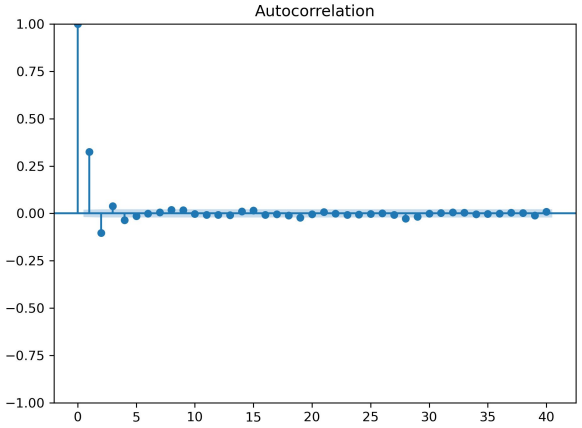


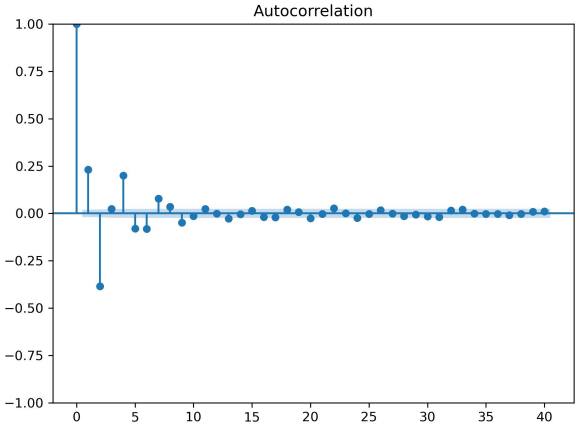
Examples of stationary processes

Autocorrelation Function (ACF), $R_{t,t+\tau} = Cov(X_t, X_{t+\tau}) \, [\sqrt{Var(X_t)}\sqrt{Var(X_{t+\tau})}]^{-1}$









Autocorrelation function (ACF)

Autocorrelation Function (ACF) of a time series is defined as

$$\mathbf{R_{t,t+\tau} = Cov(X_t, X_{t+\tau}) \left[\sqrt{Var(X_t)}\sqrt{Var(X_{t+\tau})}\right]^{-1}}$$

Useful properties:

- $\mathbf{R_{t,t} = Var(X_t)}$
- For covariance stationary processes, $\mathbf{R_{t,t+\tau} = R_{t-\tau,\, t}}$, i.e. **ACF is a function of τ only**
- For covariance stationary processes, $\mathbf{Cov(X_t, X_{t+\tau}) = Cov(X_{t-\tau}, X_t)}$

How do we deal with non stationary processes?

We can remove the non stationary trend behavior from the data! (Box-Jenkins method)

How do we do this?

By differencing method,

Assume,

- If the process **$\{X_t\}$ has a linear trend** in time, then by transformation we can obtain a process **$\{\Delta X_t\}$ that has no trend.**
- If the process **$\{X_t\}$ has a quadratic trend in time,** then by transformation we can obtain a second order process **$\{\Delta^2 X_t\}$ that has no trend.**

Differencing operators

For $X_t = \psi X_{t-1} + \eta_t$,

The first order difference operator is,

$\Delta \mathbf{X_t = X_t - X_{t-1}}$, alternatively, $\Delta \mathbf{X_t = (1- L)X_t}$, where L is lag operator

The second order difference operator is,

$\Delta(\Delta \mathbf{X_t}) = \Delta X_t - \Delta X_{t-1}$, $\Delta^2 X_t = (1- L)\Delta X_t = (1- L)(1- L)X_t$

$\mathbf{\Delta^2 X_t = (1- L)^2 X_t}$

Generally, the i^{th} order difference operator is

$\mathbf{\Delta^i X_t = (1- L)^i X_t}$

Trend removal by differencing

A trend is a time dependent variation in the time series $\{X_t\}$

i.e. $X_t = TD_t + \eta_t$, where $TD_t = a + bt$ (deterministic linear trend)

also, $\eta_t \sim AR(1)$, i.e. $\eta_t = \psi\eta_{t-1} - \varepsilon_t$, where $|\psi| < 1$ and ε_t is $WN(0,\sigma^2)$

Then moments of X_t are , $\mathbf{E[X_t]} = E[TD_t] + E[\eta_t] = a + bt$

and $\mathbf{Var[X_t]} = Var[\eta_t] = \sigma^2/(1-\psi)$

$\Delta\mathbf{X_t} = \mathbf{X_t - X_{t-1}}$, where $X_t = TD_t + \eta_t$,

$\Delta\mathbf{X_t} = TD_t + \eta_t - TD_{t-1} + \eta_{t-1} = a + bt - a - b(t-1) + \Delta\eta_t = b + \Delta\eta_t$

$\Delta\mathbf{X_t} = b + (\eta_t - \eta_{t-1}) = b + (1-L)\eta_t = b + (1-L)(1-\psi L)\varepsilon_{t-1}$,

$\Delta\mathbf{X_t = b + (1-L)(1-\psi L)\varepsilon_{t-1}}$

Incorporating i^{th} order difference and ARMA: ARIMA

The time series $\{X_t\}$ follows an ARIMA(p, q, d) model ("Integrated ARMA"), if

- $\{\Delta^d X_t\}$ is **stationary** (and non-stationary for lower-order differencing)
- In d order differencing it follows an **ARMA(p, q) model**

Practical Challenges:

- **Determining the order of differencing** required to remove time trends (deterministic or stochastic).
- **Estimating the unknown parameters** of an ARIMA(p, q, d) model.
- **Model Selection**: choosing among alternative models with different (p, q, d) specifications

Poll 4: Any time series can be made stationary with the differencing method

- True
- False

https://tinyurl.com/mlsp23-1024-04

Poll 4: Any time series can be made stationary with the differencing method

- True
- False

Limitations of Differencing: trend removal in Random Walk

$\mathbf{X_t = X_{t-1} + \eta_t}$, where η_t is $WN(0,\sigma^2)$ (also called a **Pure Integrated Process I(1)**)

then, $\Delta\mathbf{X_t = (1-L)X_t + \eta_t}$

Given X_0, we can rewrite this as, $\mathbf{X_t = X_0 + TS_t}$, where $TS_t = \sum_{\mathbf{j=0}}^{t}\eta_t$

then, TS_t is a **Stochastic Trend** process,

$\mathbf{TS_t = TS_{t-1} + \eta_t}$, where η_t is $WN(0,\sigma^2)$

Note:

- As a consequence of η_t, the Stochastic trend processes are not perfectly predictable (non-stationary)
- Differencing operator cannot remove trend associated with η_t

Application of Time series forecasting in context of 11755 projects

- The data is a time series (exchange rate, AQi)
- We can model the time series with ARMA, ARIMA model (for trend removal)

Considerations:

- Parameter estimation can be carried out by **MLE**
- The ARMA series can be represented as matrices through **Yule-Walker equations**
- The prediction will not be accurate
  - We have partial information and decomposition is not accurate
- The discussed variations of ARMA and ARIMA are for **the univariate case**, explore multivariate ARMA or **Vectorized ARMA (VARMA)**
- Careful determination of the deterministic and stochastic variable decomposition of time series
  - **Do not model stochastic components as deterministic**

Additional sources for time series forecasting

- MIT OCW: Topics in mathematics with applications in finance (18.S096) by Peter Kempthorne
- Kevin Kotzé's notes on time series prediction