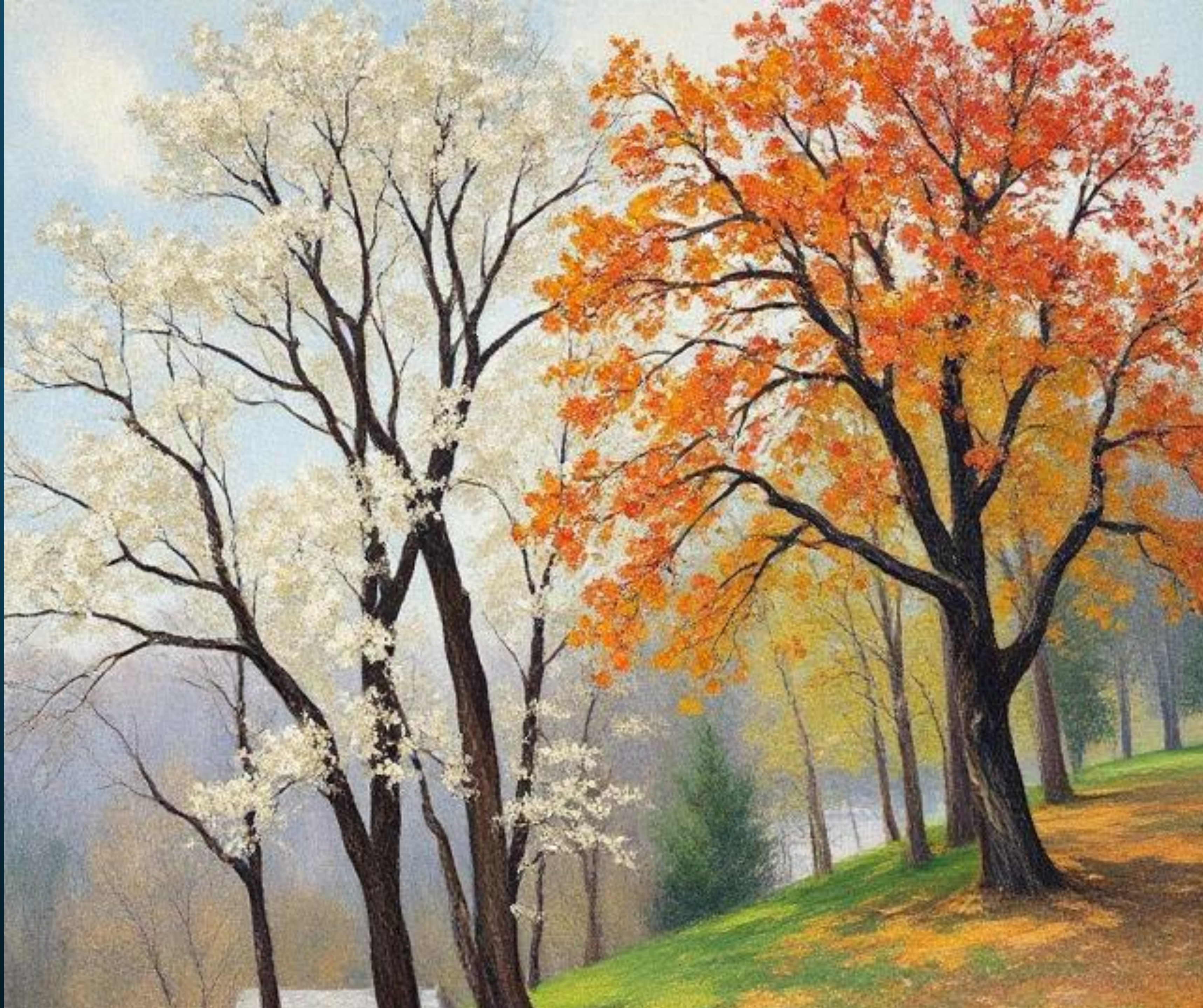
Seasonal Decomposition and SARIMA



LECTURE OUTLINE

Last time we studied the following topics:

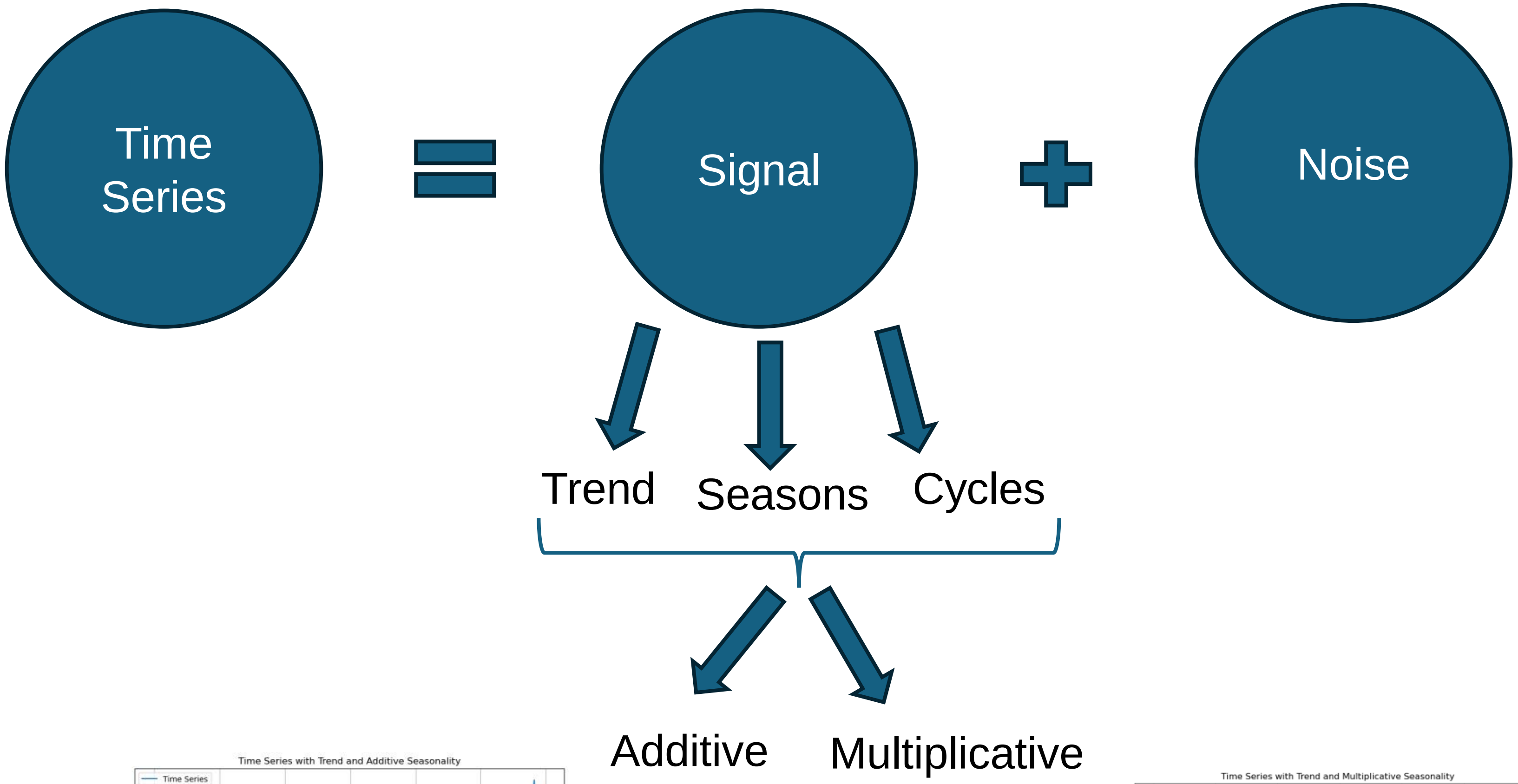
- Vector AutoRegression
- Error Correction Models
- ARIMAX

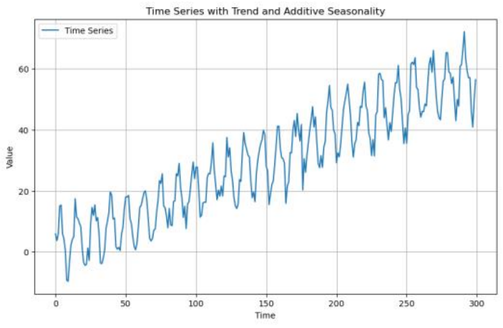
In this lecture we will study:

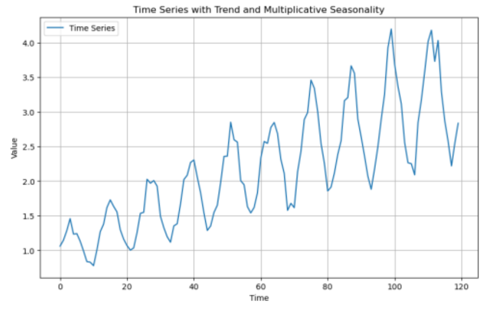
- Seasonal Decomposition
- SARIMA
- Recap

Seasonal Decomposition

Time Series Decomposition



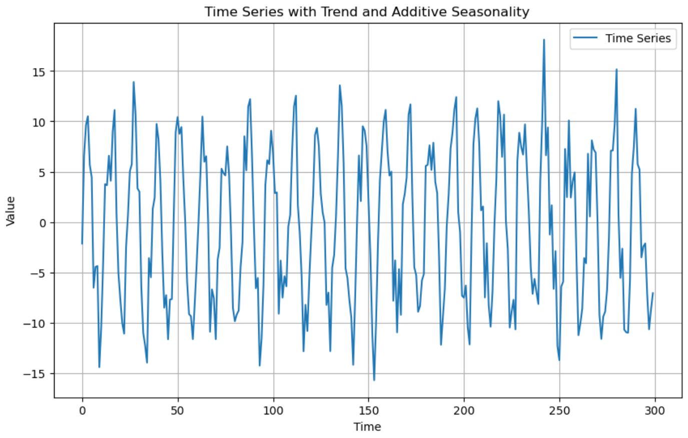


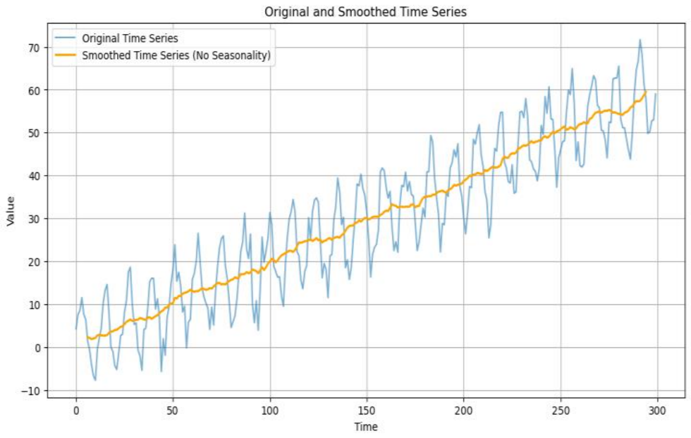


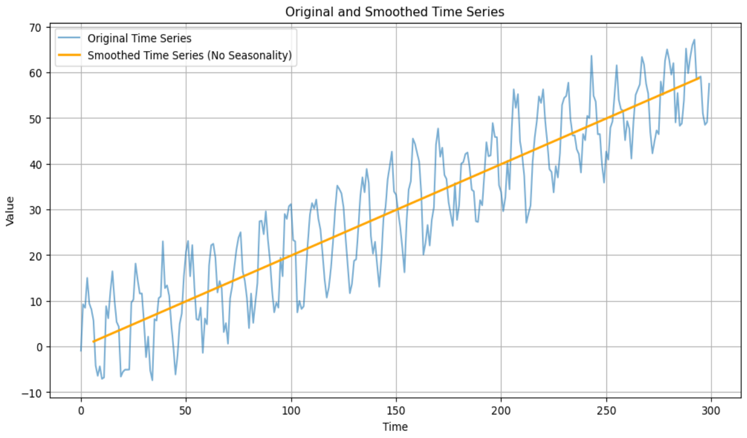
Seasons amplitude is growing

Time Series Decomposition

It is possible to decompose a time series into its components







Seasonality only

Differencing

Trend only

Moving average

Trend only with no noise

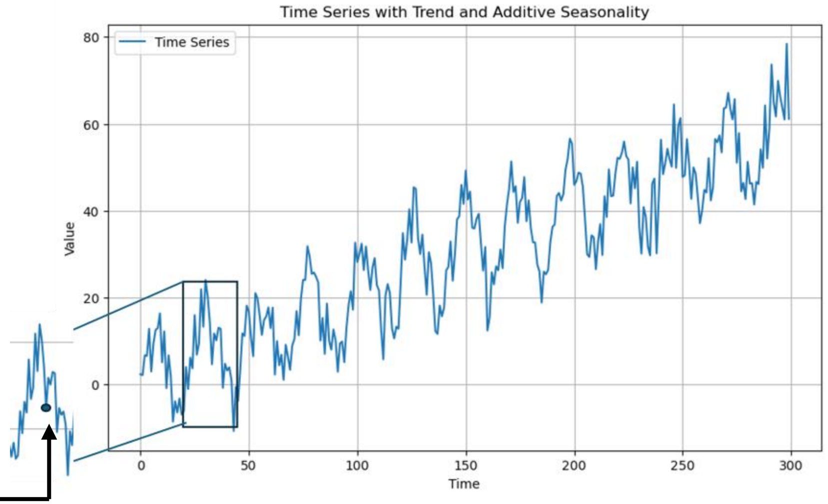
Linear regression

LOESS principle

- Moving average to erase seasonality is ok but better methods exist

→ **LOESS:** Locally Estimated Scatterplot Smoothing

→ Also called Localized Regression



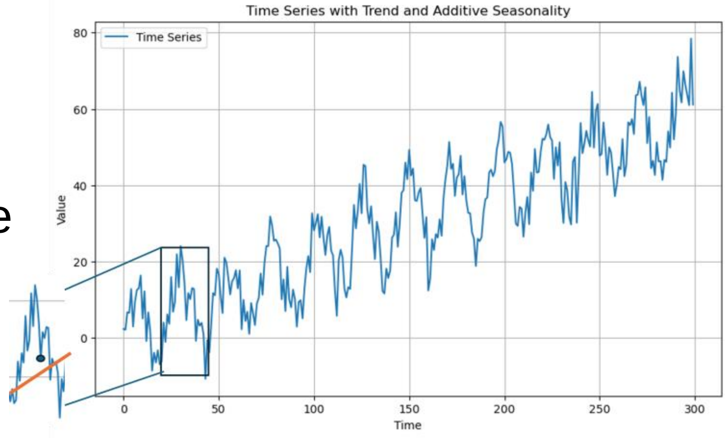
1. LOESS zooms on a small window of the data

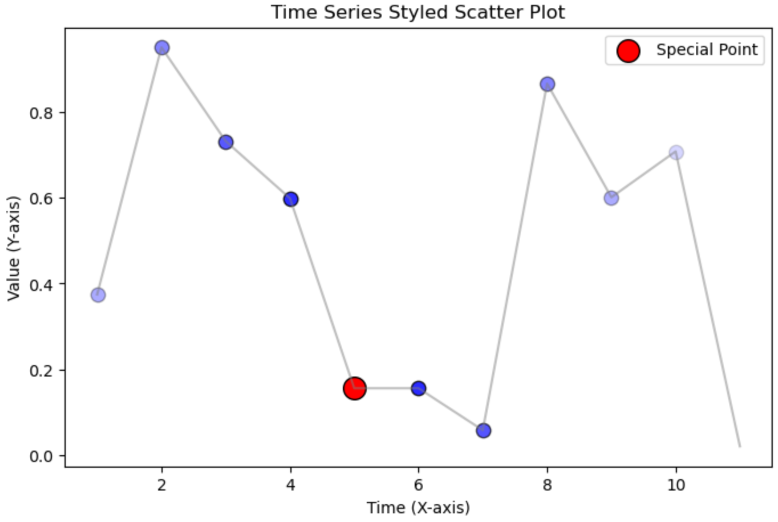
2. Then we highlight one point and take a local window of observations around it

3. Weight observations depending on their distance to the point of interest

4. Fit a weighted linear regression

5. Keep the predicted value for the observation of interest

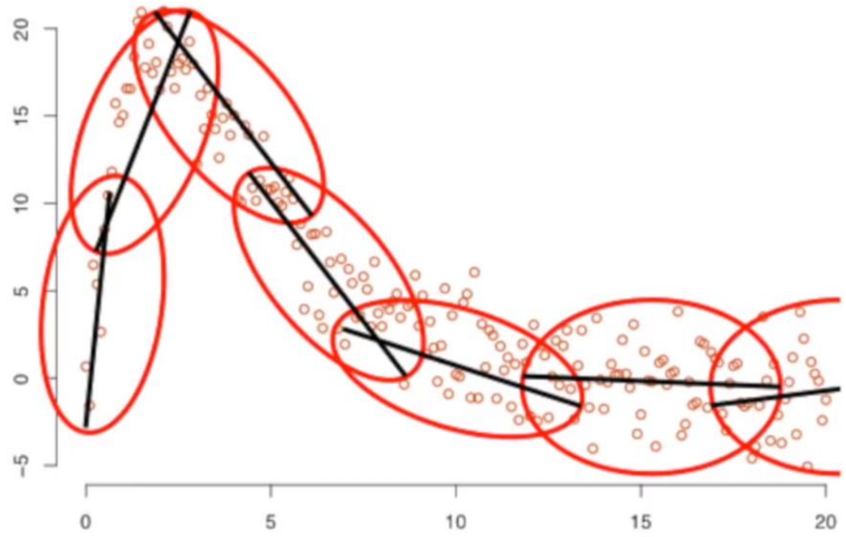




LOESS principle

- We follow the previous procedure for every point
- **Result:** Seasonality is taken away smoothly

Unlike global models (e.g., linear regression), LOESS does not assume a fixed functional form for the data. Instead, it builds a flexible, data-driven model.

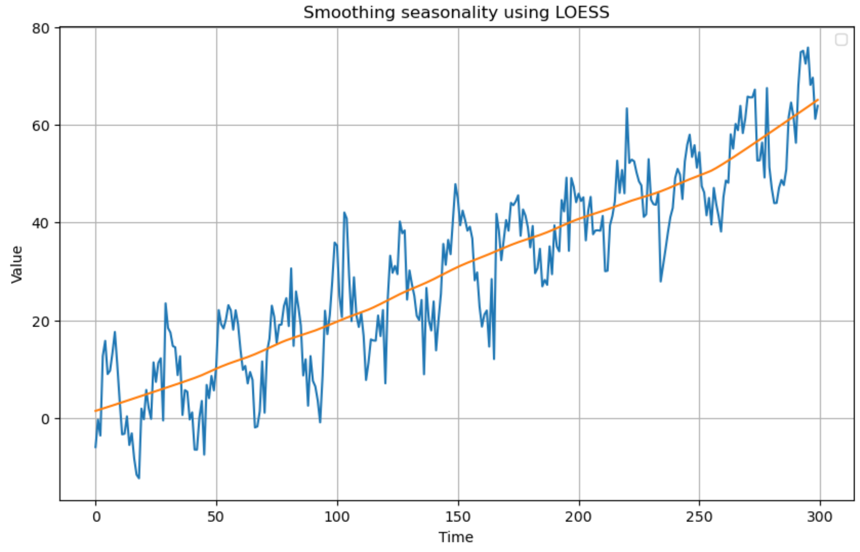


Some details:

- The size of this neighborhood is controlled by a parameter called the span or bandwidth. The span is typically defined as a proportion of the total data points
- The weights are computed using a kernel function, such as the tricube weight function:

$$w(x) = \begin{cases} (1 - |d|^3)^3 & \text{for } |d| < 1, \\ 0 & \text{otherwise,} \end{cases}$$

$$\text{where } d = \frac{|x - x_i|}{\text{max distance in neighborhood}}$$

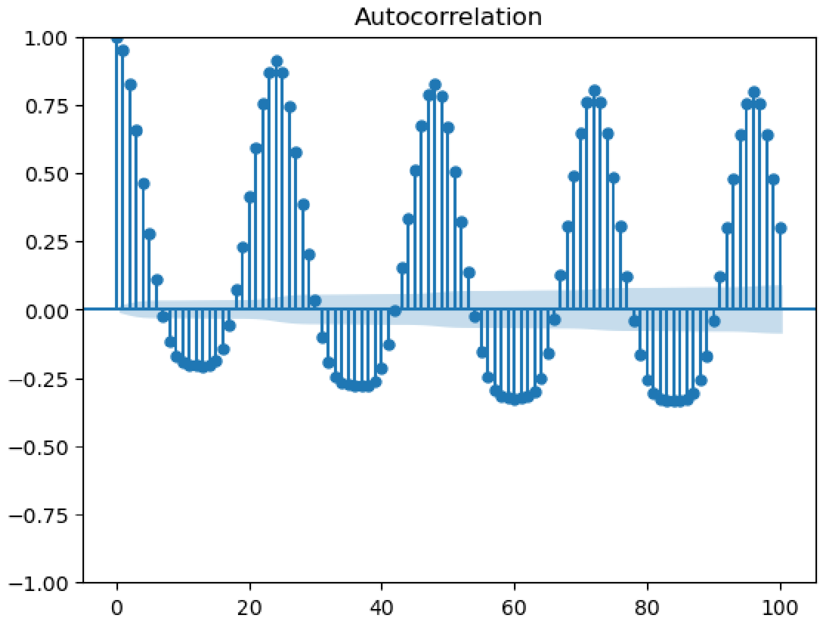


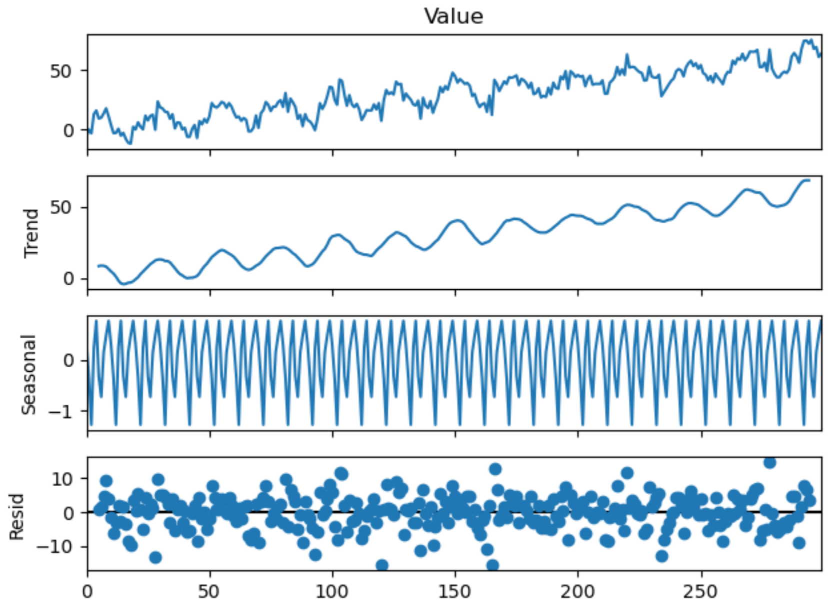
Limits:

- Computationally expensive
- Boundary effects
- Choice of parameters

Visualizing seasonality

- You can use the ACF
- Statsmodel library offers a component decomposition module





SARIMA

How to forecast with seasonality

- We recall that ARIMA has 3 parameters (p,d,q)

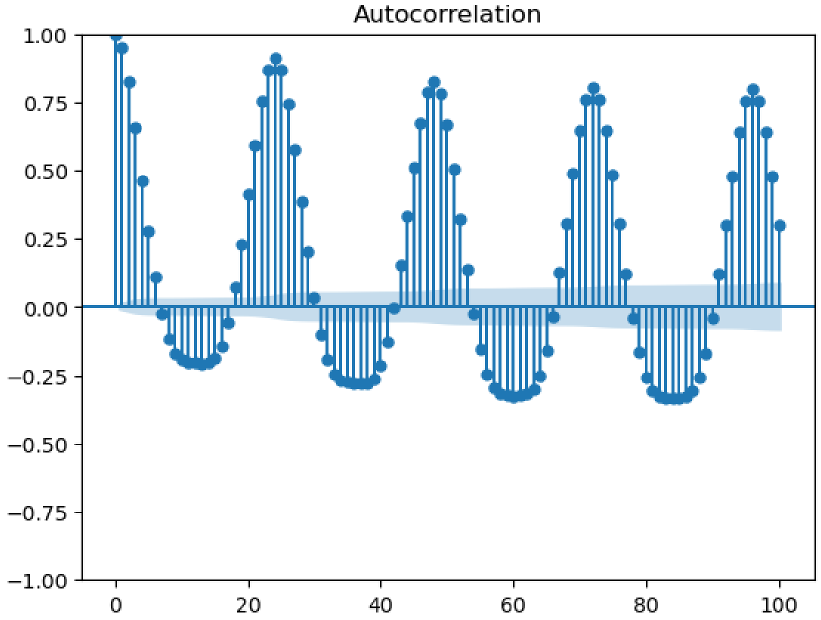
- SARIMA adds 4 more parameters

→ . It is written as **SARIMA(p,d,q)(P,D,Q)m**

- (p,d,q) are exwactly the same as for ARIMA models
- m parameter: number of periods that it takes for a seasonality to repeat

→ In this example m=12

- The (P, D, Q) parameters are the analogs of (p, d, q).



D parameter

- Consider a SARIMA(1,1,1)(1,1,1)12 model and recall the backshift operator: $L^k y_t = y_{t-k}$
- We have d = 1, meaning: $y_t - y_{t-1} = y_t - L^1 y_t = (1 - L^1) y_t$
- The D parameter says that we will also backshift the series by D × m.

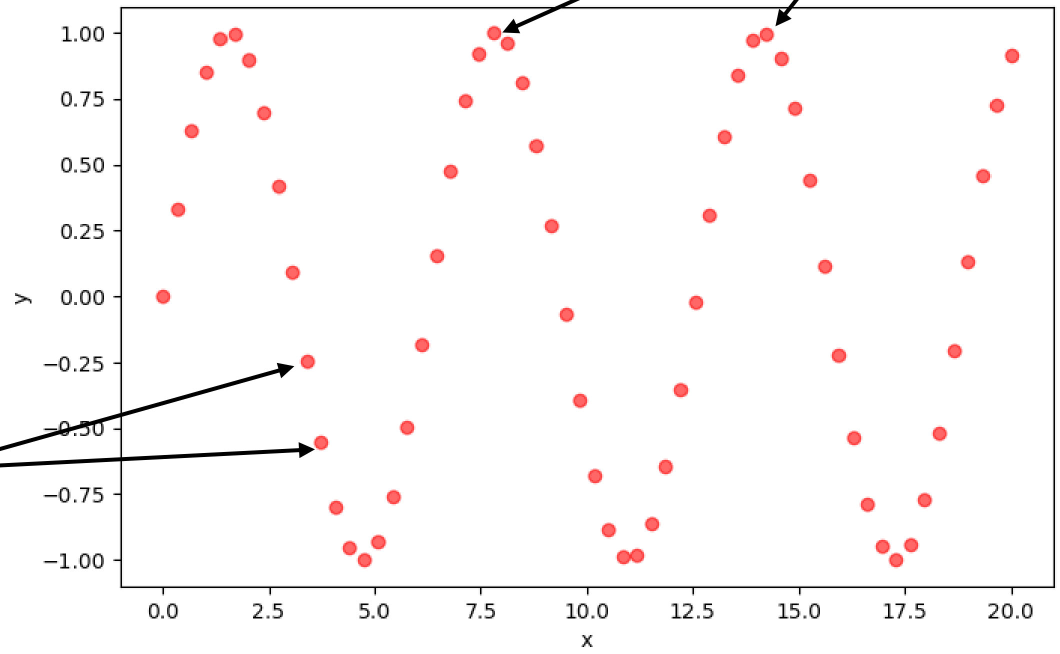
$$(1 - L^{12}) y_t$$

This difference is important too

Intuition: We difference d times to handle stationarity but then given a period of m, we difference D times between y_t and y_{t-m} to handle seasonality
→ The immediate variation is interesting but the variation between a given point and the same point in the last period is also interesting

→ When we write SARIMA(1,1,1)(1,1,1)12, there is an implicit 1, such as SARIMA(1, 1, 1)$_1$(1, 1, 1)12.

Immediate difference



P parameter

- For an AR(p) process we have (ignoring the intercept):

$$y_t = (\phi_1 \cdot L^1 + \phi_2 \cdot L^2 + \cdots + \phi_p \cdot L^p) y_t + \epsilon_t$$

$$(1 - \phi_1 \cdot L^1 - \phi_2 \cdot L^2 - \cdots - \phi_p \cdot L^p) y_t = \epsilon_t$$

- The P parameter has the same logic, but we multiply the power of the backshift parameter by the value m:

$$y_t = (\Phi_1 \cdot L^{1 \times m} + \Phi_2 \cdot L^{2 \times m} + \cdots + \Phi_P \cdot L^{P \times m}) y_t + \epsilon_t$$

$$(1 - \Phi_1 \cdot L^{1 \times m} - \Phi_2 \cdot L^{2 \times m} - \cdots - \Phi_P \cdot L^{P \times m}) y_t = \epsilon_t$$

Note that $\phi \neq \Phi$ and $p \neq P$

- So with P = 1 and m = 12, we have: $(1 - \Phi \cdot L^{12}) y_t = \epsilon_t$

Q parameter

We recall that in an MA model we have: $$y_t = \epsilon_t + \theta_1 \epsilon_{t-1} + \ldots \theta_q \epsilon_{t-q}$$

$$y_t = \epsilon_t + \theta_1 L^1 \epsilon_t + \cdots + \theta_q \cdot L^q \epsilon_t$$

$$y_t = (1 + \theta_1 \cdot L^1 + \cdots + \theta_q \cdot L^q) \epsilon_t$$

The Q parameter bears the same logic, we just multiply again the power of L by m.

$$y_t = (1 + \Theta_1 \cdot L^{1 \times m} + \cdots + \Theta_Q \cdot L^{Q \times m}) \epsilon_t$$

All parameters

- Expressing ARIMA(p,d,q) uusing the backshift operator, we get:

$$(1-\phi_1 \cdot L^1 - \phi_2 \cdot L^2 - \cdots - \phi_p \cdot L^p)(1-L^1)^d y_t = (1+\theta_1 \cdot L^1 + \cdots + \theta_q \cdot L^q)\epsilon_t$$

- The SARIMA model keeps the same logic but adds the terms with the powers multiplicated by m.

$$(1-\phi_1 \cdot L^1 - \phi_2 \cdot L^2 - \cdots - \phi_p \cdot L^p)(1-\Phi_1 \cdot L^{1\times m} - \Phi_2 \cdot L^{2\times m} - \cdots - \Phi_P \cdot L^{P\times m})(1-L^1)^d(1-L^m)^D y_t$$
$$= (1+\theta_1 \cdot L^1 + \cdots + \theta_q \cdot L^q)(1+\Theta_1 \cdot L^{1\times m} + \cdots + \Theta_Q \cdot L^{Q\times m})\epsilon_t$$

- If we consider again our SARIMA(1, 1, 1)(1, 1, 1)12 example, it can be written as:

$$(1-\phi_1 \cdot L^1)(1-\Phi_1 \cdot L^{12})(1-L^1)(1-L^{12})y_t = (1+\theta_1 \cdot L^1)(1+\Theta_1 \cdot L^{12})\epsilon_t$$

ARIMA parameters

Seasonality parameters

Additional Examples

Intuition: ARIMA just differenciates the time series by the period length

Consider the SARIMA(1,0,0)(0,1,1)12: $(1-\phi_1 \cdot L^1)(1-L^{12})y_t = (1+\Theta_1 \cdot L^{12})\epsilon_t$

Let's expand this formula: $(1-\phi_1 \cdot L^1 - L^{12} + \phi_1 \cdot L^{13})y_t = (1+\Theta_1 \cdot L^{12})\epsilon_t$

$$y_t - \phi_1 \cdot y_{t-1} - y_{t-12} + \phi_1 \cdot y_{t-13} = \epsilon_t + \Theta_1 \cdot \epsilon_{t-12}$$

$$y_t - y_{t-12} = \phi_1 \cdot y_{t-1} - \phi_1 \cdot y_{t-13} + \epsilon_t + \Theta_1 \cdot \epsilon_{t-12}$$

Let $z_t = y_t - y_{t-12}$

$$z_t = \phi_1 \cdot z_{t-1} + \Theta_1 \cdot \epsilon_{t-12} + \epsilon_t$$

Consider SARIMA(1,0,1)(2,1,0)12:

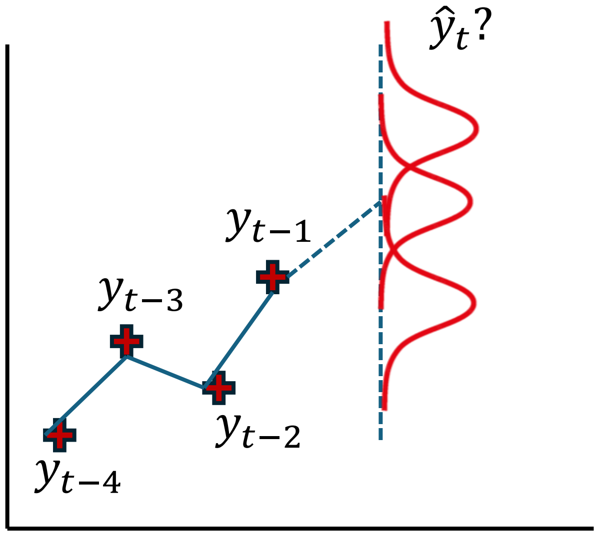
- start by defining your new series: $z_t = y_t - y_{t-12}$
- Then:

$$z_t = \phi_0 + \phi_1 \cdot z_{t-1} + \theta_1 \epsilon_{t-1} + \Phi_1 \cdot z_{t-12} + \Phi_2 \cdot z_{t-24} + \epsilon_t$$

Final note: It is possible to extend the SARIMA model to create a SARIMAX version

ARCH:AutoRegressive Conditional Heteroskedasticity

Reminder on MLE



$$\mathcal{L} = P(obs|\theta_0, \mathcal{M})$$

Normal Distribution

y_t

What is this exactly? It is the parameters of the model M, **its mean and variance**

$$y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-1} + \cdots + \phi_p y_{t-p} + \epsilon_t$$

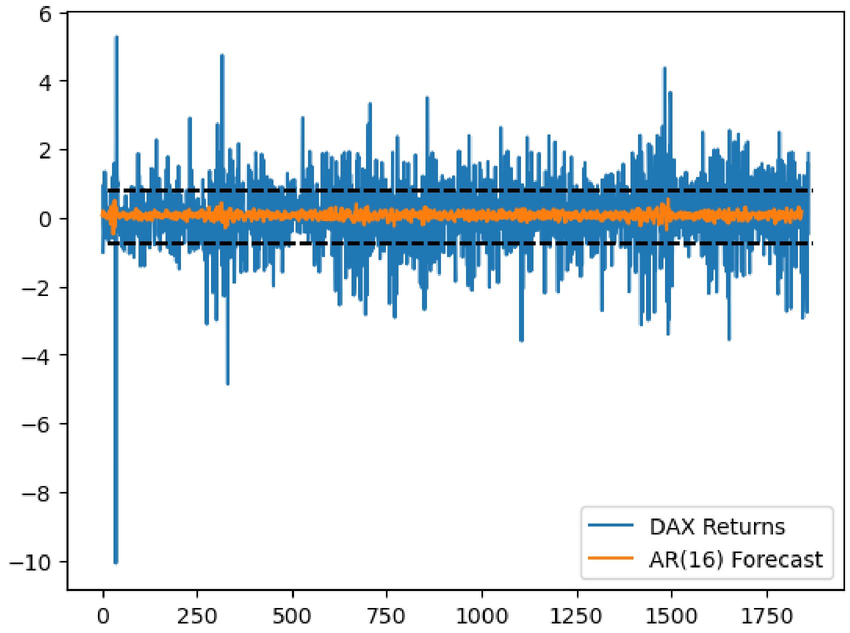
$$p(y|x) = \mathcal{N}(y; \hat{y}(x), \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\hat{y}(x) - y)^2}{2\sigma^2}\right)$$

$\hat{y}_t$ is the mean of a Normal distribution conditionned by the previous values

→ **Bonus:** MLE allows to estimate the variance of the distribution → you get uncertainty over your prediction for free

→ But this uncertainty is fixed

→ It would be great to predict a flexible variance

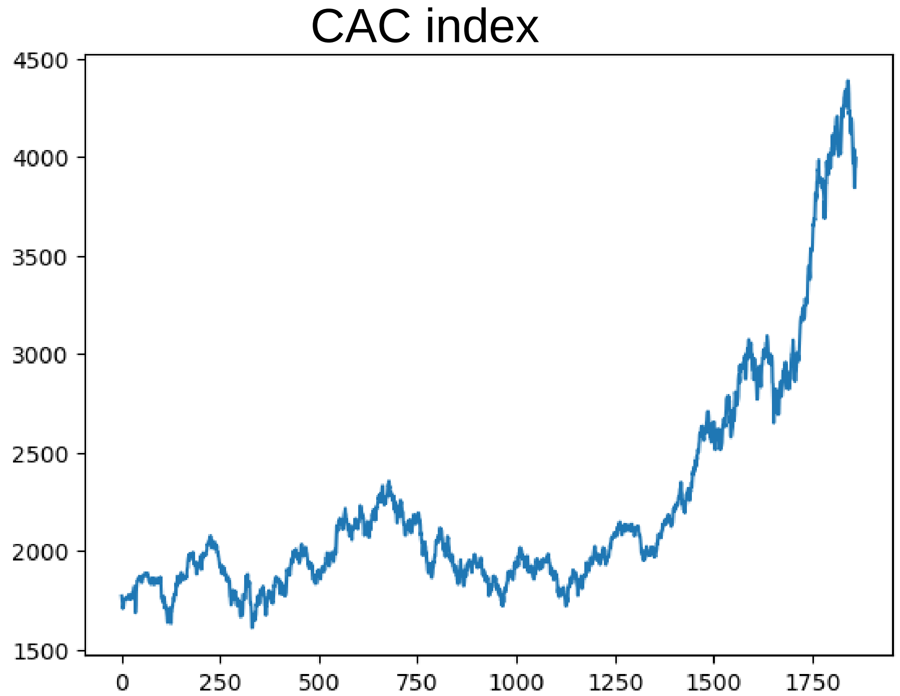


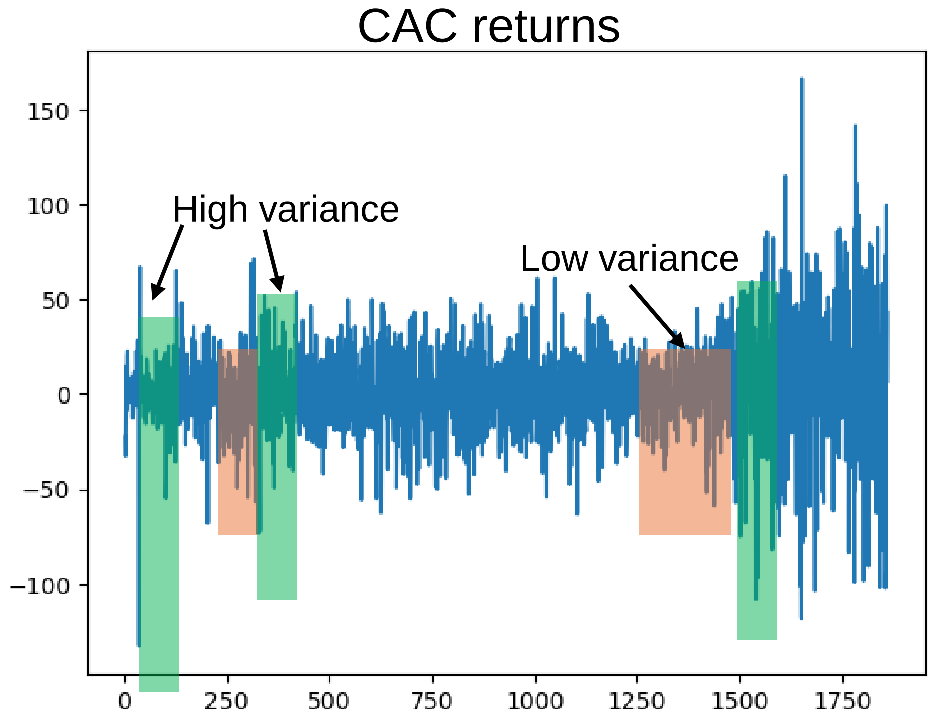
Conditional Variance

- The variance we computed before using MLE is said unconditional, it is the variance accross the entire dataset. A kind of mean of the variance.

- ARCH models compute the conditional variance, a variance conditioned on the previous values of the variance; hence the name AR **conditional Heteroskedasticity**

- Empirically, periods of large volatility are followed by periods of low volatility





- Estimating the volatility i.e the conditional variance, it can help take more informed decisions, mitigate risk, price options, or balance a portfolio of assets

Conditional mean but fixed variance

AR process: $y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-1} + \cdots + \phi_p y_{t-p} + \epsilon_t$

The prediction is given by: $\hat{y}_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-1} + \cdots + \phi_p y_{t-p} = \mu(y_t|y_{t-i})$

So actually:

$$y_t = \hat{y}_t + \epsilon_t = \mu(y_t|y_{t-i}) + \epsilon_t$$

In AR process, the real value is considered to be the sum of a conditional mean and a fixed error or variance.
→ It could be could be cool it this variance was conditional too

How to model volatility

- ARIMA models condition predictions on the past values of the time series.

→ For the variance or volatility we can do the same. We take an auto-regressive approach.

> The defining feature of ARCH models is that the variance of ϵ_t, conditional on past observations, follows a specific autoregressive structure

- ϵ_t is the error made by your model (ARIMA for instance) and we consider that its variance depends on the variance of the errors we made before.

$$\mathrm{var}(\epsilon_t) = \sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i \sigma_{t-i}^2$$

But how do we get σ_{t-i}^2 ?

- This is the variance of the errors we made yesterday but we cannot observe it, all we can observe is the error that we actually made yesterday.
- The best proxy we have for volatility at time t − 1 is the realized squared error ϵ_{t-1}^2 since the true variance is unobserved.

⚠ We approximate the variance of yesterday with the square ot the error we made yesterday.

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i \epsilon_{t-i}^2$$

ARCH process

1. fit an ARIMA (or another time series model) to capture the conditional mean μ

$$y_t = \mu(y_t | y_{t-i}) + \epsilon_t$$

2. Once you obtain the residuals ˆϵt from the ARIMA model, you check if they exhibit conditional heteroskedasticity (i.e., time-varying volatility). If so, you model their variance using an ARCH process:

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i \epsilon_{t-i}^2$$

where σ_t^2 gives the time-dependent variance of the residuals.

3. The point forecast for the next observation is given by the ARIMA model. The uncertainty (variance) around that prediction is given by the ARCH model. Now every forecast is taken from N ($\hat{\mu}_{t+1}$, $\hat{\sigma}_{t+1}$). Therefore you can sample for each timestep different possible forecasts:

$$y_{t+1} = \hat{\mu}_{t+1} + \hat{\sigma}_{t+1} z_{t+1}, \quad z_{t+1} \sim N(0,1)$$

4. You can compute an interval forecast (e.g., 95% confidence interval):

$$(\hat{\mu}_{t+1} - 1.96\hat{\sigma}_{t+1}, \hat{\mu}_{t+1} + 1.96\hat{\sigma}_{t+1})$$

Example

Suppose you are forecasting stock returns.

1. Fit an ARIMA model to predict expected returns $\hat{\mu}_{t+1}$.

2. Fit an ARCH model to capture the time-varying volatility $\hat{\sigma}^2_{t+1}$.

3. our final return forecast follows

$$R_{t+1} \sim N(\hat{\mu}_{t+1}, \hat{\sigma}^2_{t+1})$$

4. If $\hat{\mu}_{t+1}$ = 0.002 (0.2% expected return) and $\hat{\sigma}^2_{t+1}$ = 0.01 (1% volatility), your 95% confidence interval is:

$$(0.002 - 1.96(0.01), 0.002 + 1.96(0.01)) = (-0.0186, 0.0226)$$

→ This means the next return is expected to be 0.2%, but with a range of possible values between -1.86% and 2.26%.

How to fit ARCH using MLE

Recall that the likelihood function, assuming a Normal distribution is:

$$\mathcal{L}(\theta) = \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{(x_t - \mu)^2}{2\sigma^2}\right)$$

$$= \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{(\hat{y}_t - y_t)^2}{2\sigma^2}\right)$$

Replace the mean by the prediction

$$= \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{\epsilon_t^2}{2\sigma^2}\right)$$

Prediction – real value = residual

$$= \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{\epsilon_t^2}{2\hat{\sigma}_t^2}\right)$$

Replace the variance by your estimate

$$\log \mathcal{L}(\theta) = \sum_{t=1}^{T} \left(-\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\hat{\sigma}_t^2) - \frac{\epsilon_t^2}{2\hat{\sigma}_t^2}\right)$$

When to use ARCH

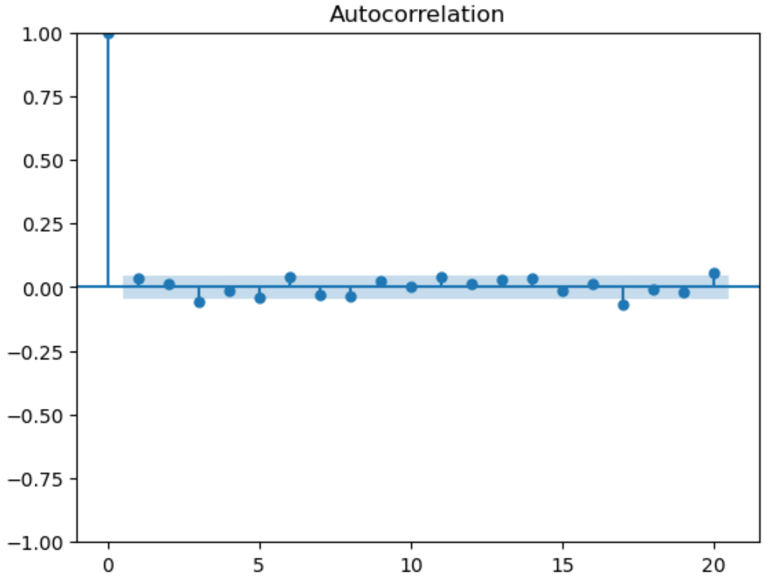
We said that after fitting your ARIMA model, you need to check for the heteroscedasticity of the residuals.
How to do that?
→ Use the ACF and PACF of the residuals. If you see significant spikes that are decreasing overtime, this means that there is persistence in the variance and you should model it using ARCH.

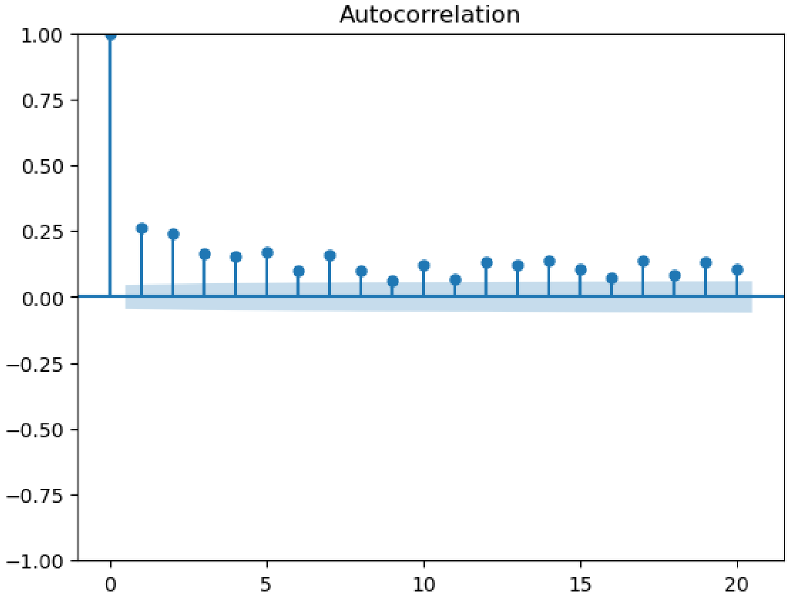
- you can also test if you should use ARCH even **before** getting residuals from your ARIMA model.
  - The best proxy for the unobserved volatility is the squared observed error.
  - Error is just the difference between the real value and the predicted one.
  - With a naïve prediction $\hat{y}_t = \hat{y}_{t+1}$, error is yesterday's value minus today's value = return

  → you can consider the squared return as a proxy for the volatility

  → Plot squared return ACF and if you see significant spikes that decrease over time, then you should use ARCH.

Return ACF





Squared Return ACF

How to evaluate an ARCH model (1/2)

- your forecast value is the output from ARIMA that you can compare to the real value using MSE. But there is no MSE to compute for the variance.

Alternatives:

- Likelihood
- AIC/BIC
- Coverage:
  - construct 95% prediction intervals: $(\mu_t - 1.96\hat{\sigma}_t, \mu_t + 1.96\sigma_t)$
  - Then we check how often the actual value y_t falls inside:

$$\text{Coverage} = \frac{\text{Number of times } y_t \text{ is in the interval}}{T}$$

  - Ideally, for a 95 prediction interval, we should have close to 95% coverage.
  - If coverage is too low, your model is underestimating risk
  - If coverage is too high, your model is overestimating risk.

If your model consistently predicts very high variance, the coverage may indeed be 100% but this doesn't necessarily mean the model is good. In fact, it could indicate that the model is overestimating the uncertainty

How to evaluate an ARCH model (2/2)

- Standardized residuals normality test

$$\hat{z}_t = \frac{y_t - \hat{\mu}_t}{\hat{\sigma}_t}$$

If $\hat{\sigma}_t^2$ is well-calibrated, then $\hat{z}_t \sim$ N (0, 1).

→ **Usual tests:** Jarque-Bera, Ljung-Box

- ARCH assumes that after accounting for time-varying volatility, the residuals should be i.i.d. and follow a standard normal distribution N(0,1). If standardized residuals are still heteroskedastic, the model is misspecified.

- If the model underestimates volatility, then $\hat{z}_t$ will be larger than expected (too many extreme values)
- If the model overestimates volatility, then $\hat{z}_t$ will be too small (not enough dispersion)

- The ideal case: $\hat{z}_t \sim \mathcal{N}(0, 1)$

Lags and Limitations

How many lags include in an ARCH model?

→ Unfortunately, there is no principled way to figure out beforehand. You have to try many different lags, compute their AIC/BIC and choose the parameter that gives the best result.

Limitations:

- Financial time series often exhibit long memory in volatility (i.e.,past volatility influences current volatility for a long time). ARCH models require a large number of lags to capture this, making them inefficient.

- ARCH models assume that positive and negative shocks have the same effect on volatility, meaning that a large positive return increases future volatility just as much as a large negative return. In reality, bad news (negative returns) often increases volatility more than good news.

GARCH: Generalized ARCH

Bootstrap

- ARCH considers that the best proxy for the last variance is the residual

→ **Better idea:** why not use our last variance forecast as an approximation for the variance?

- ARCH Uses only past squared residuals.: $\sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i \epsilon_{t-i}^2$

- GARCH model uses both past squared residuals and past variance estimates.:

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i \epsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j \hat{\sigma}_{t-j}^2$$

- ARCH assumes that volatility is driven only by past squared shocks. GARCH assumes that volatility is also persistent, meaning that today's variance depends on both past shocks and past variance estimates

Intuition and Advantages

- GARCH can be seen as an ARMA model for variance, where:
  - The ARCH part is like an AR(q) model for ϵ_t^2
  - The GARCH part adds an MA(p) component by including past variances

- **Why is GARCH More Efficient than ARCH?**
  - **Captures volatility persistence better:** Since past variance σ^2 is included, it smooths volatility over time and avoids the need for a large q in ARCH.
  - **Fewer parameters**: Instead of requiring a high-order ARCH(q) model, a GARCH(1,1) model can often capture long-term volatility behavior efficiently.

WRAP UP

Components of a Time Series

Long-Term Trend: A general upward or downward movement.

Seasonal Patterns: Regular fluctuations within a specific period.

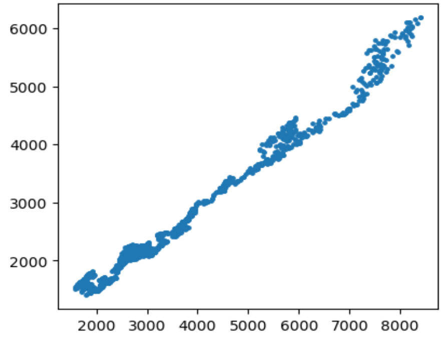
Cyclic Movements: Less predictable short-term cycles.

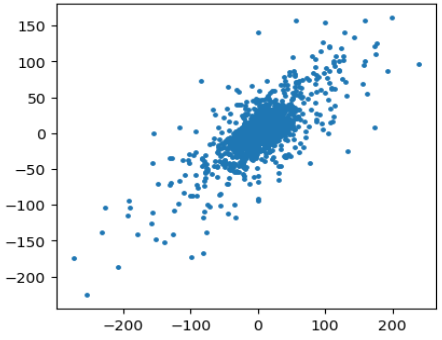
Random Fluctuations: Unpredictable noise or variations.

The importance of correlation

- In TS forecasting correlation is key, without it there would be no insight, contrary to classical ML.
- At the same time it can be misleading

Original Data





Differenced Data

Useless Correlation

Imputation methods and lookahead

Forward Fill: Carrying forward the last known value prior to the missing one.

Backward Fill: Propagating values backward (avoid lookaheads).

Moving Average: Using a rolling mean or median. Similar to a forward fill, however, you are using input from multiple recent times in the past.

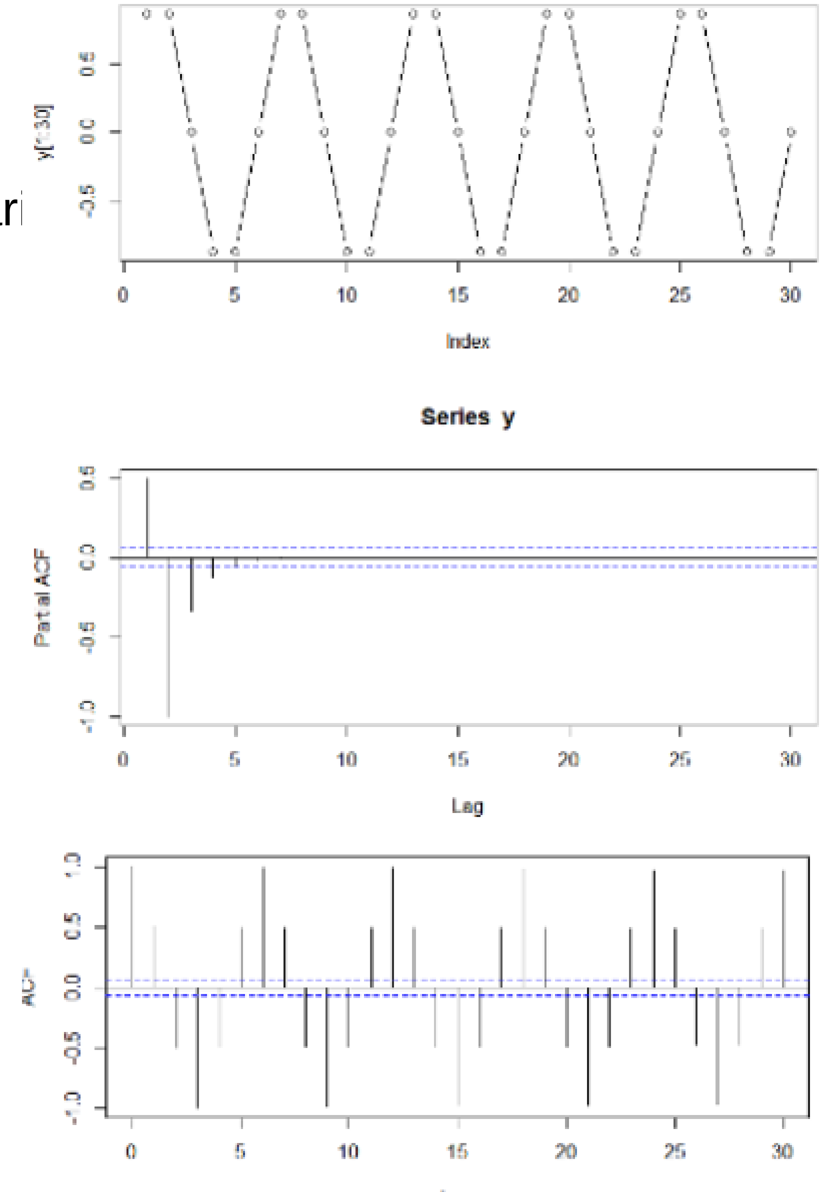
Interpolation: Estimating missing values using neighboring data points.

A note on lookahead:

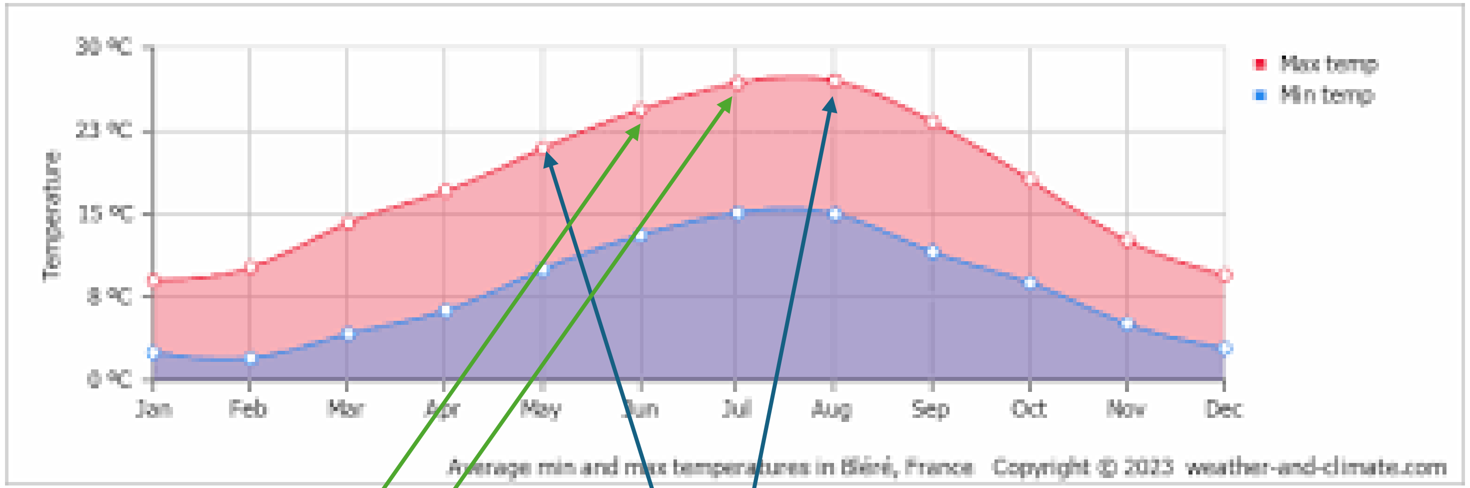
- **Definition:** Using future information to influence model behavior → You are not supposed to have such future information
- **Consequences**: Inaccurate predictions and biased models.
- **Prevention**: Use only past data for training and evaluation.

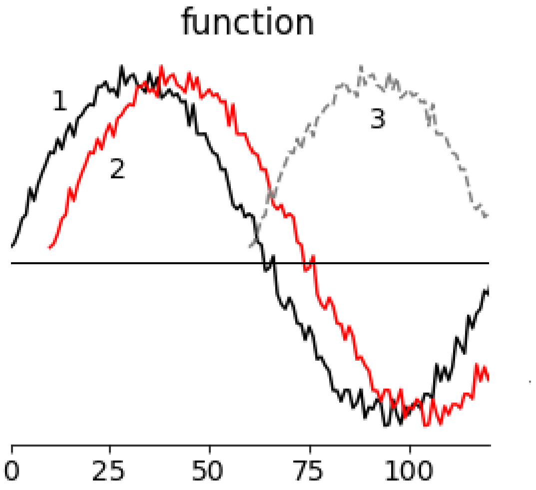
Trends, seasonality and stationarity

- How to remove trend and seasonality them and why?
- Transformations
- Definition of stationarity, difference between weak and strong stationari
- How to detect non stationarity?



Autocorrelation





Autocorrelation compares these two series

Partial autoorrelation compares these two values without the effect of these

- ACF, PACF, when to use them and why?
- How to compute them, especially PACF
- Their properties

Feature selection

- Granger causality test
- Entropy and Information Gain

MLE

$$\mathcal{L} = \Pi_{i=1}^{N} P(x_i|\theta_0, \mathcal{M})$$

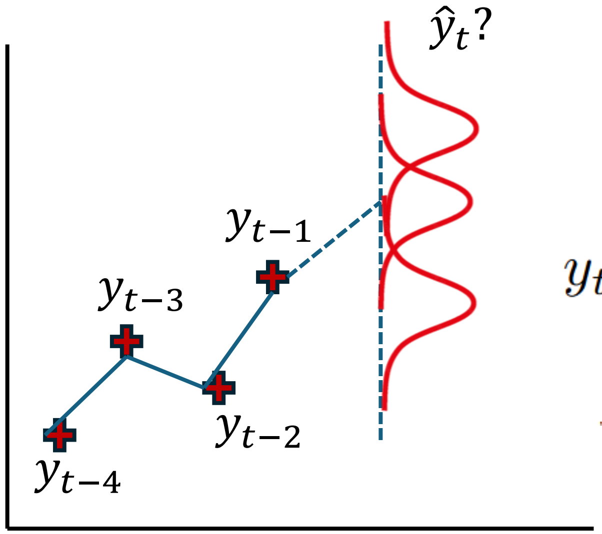
$$\theta^* = \arg\max_{\theta} P(obs|\theta, \mathcal{M})$$

$$\theta^* = \arg\max_{\theta} \sum_{i}^{m} \log P(x_i|\theta, \mathcal{M}$$

$$\mathcal{L} = P(obs|\theta_0, \mathcal{M})$$ Normal Distribution

y_t

What is this exactly? It is the parameters of the model M, **its mean and variance**



$$y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-1} + \cdots + \phi_p y_{t-p} + \epsilon_t$$

$$p(y|x) = \mathcal{N}(y; \hat{y}(x), \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\hat{y}(x) - y)^2}{2\sigma^2}\right)$$

$\hat{y}_t$ is the mean of a Normal distribution conditionned by the previous values

→ Given that we use an AR(p), the mean is modeled as $y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-1} + \cdots + \phi_p y_{t-p}$.

→ The mean itself is parametered by the coefficients ϕ_i !

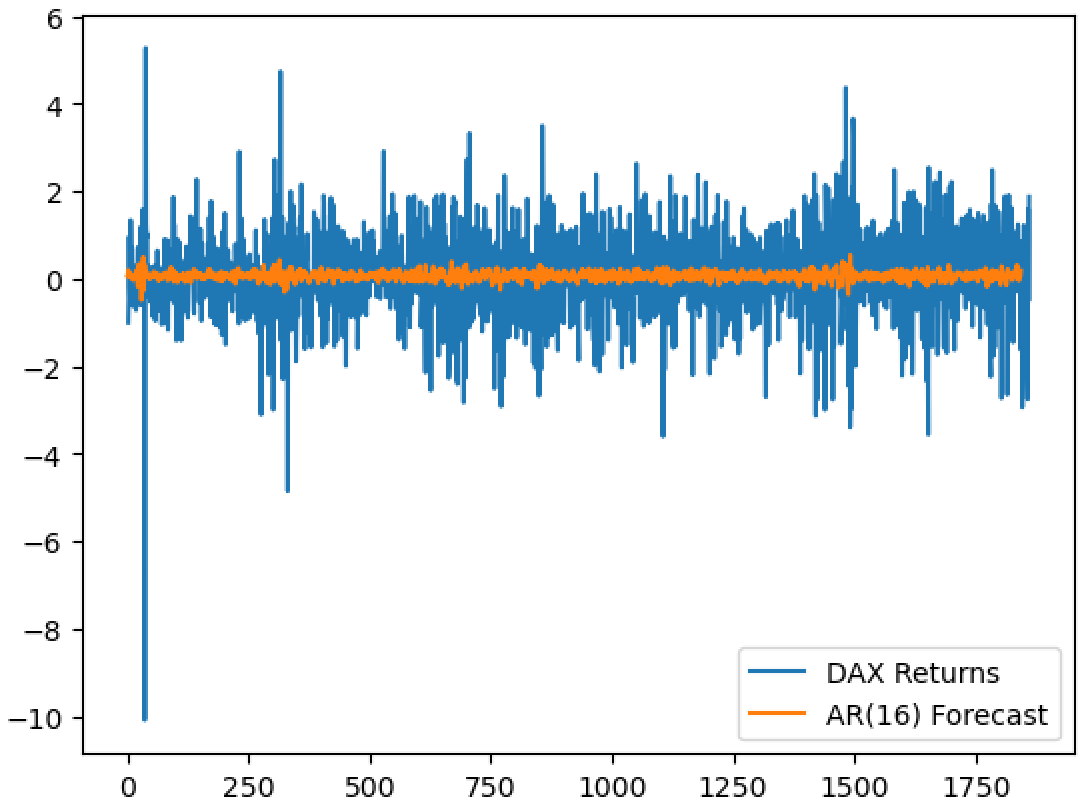
→ Finding the mean is equivalent to finding the parameters ϕ_i

→ Once the AR(p) model is fitted, the coefficients ϕ_i do not change but the y_{t-i} change, so the mean changes. Every time you are estimating a new distribution with a different mean. $\mathbb{E}[y_t|y_{t-1}, \ldots, y_{t-p}] = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p}$

→ **Bonus:** MLE allows to estimate the variance of the distribution → you get uncertainty over your prediction for free

AR models

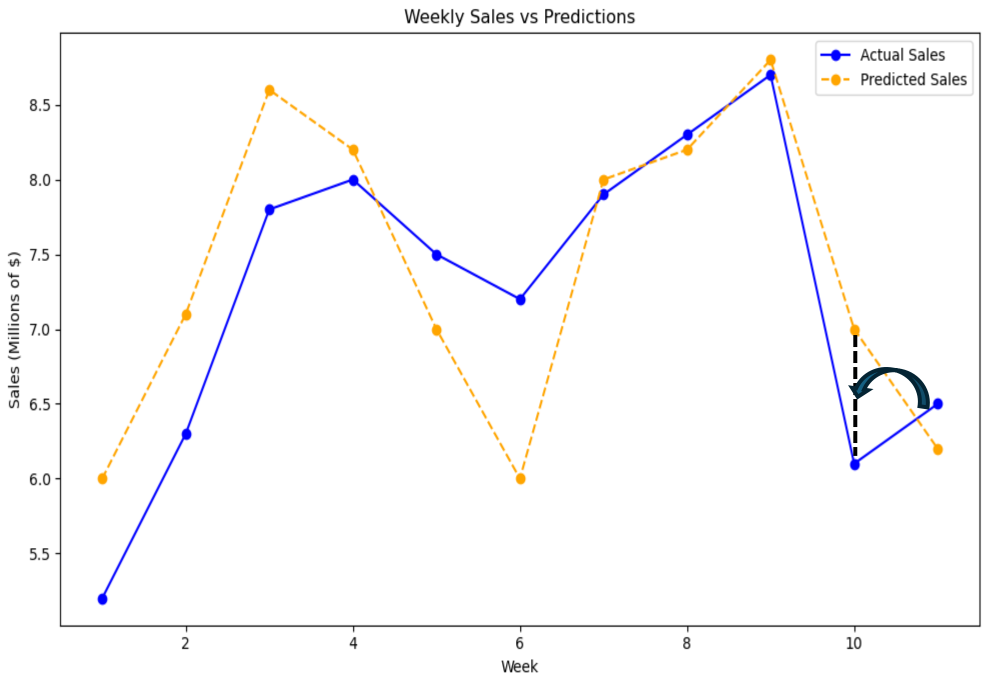
- When to use them
- Stationarity criteria, and how to test it
- How to choose the right number of lags
- Their pitfalls: regression to the mean, structural break, short term, problem with n-step forecasts



MA models

- The principle
- When to use them
- Stationarity criteria, and how to test it
- How to choose the right number of lags
- Their pitfalls: short memory models, problem with n-step forecasts

Intuition: The error from yesterday affects the current value. Some unknown predicted shock that shifted you off of where you expected to be is what's actually affecting the current time point.



How to evaluate your model

- We have mentionned a bunch of « accuracy » metrics but this is far from enough
- AIC/BIC to balance performance and complexity
- Residuals diagnostics to assess if you can do better

ARMA models

Number of autoregressive terms

ARMA(p,q) $\longleftarrow$ $$\boxed{y_t = \phi_0 + \sum_{i}^{p} \phi_i y_{t-i} + \epsilon_t + \sum_{j}^{q} \theta_j \epsilon_{t-j}}$$

Number of error terms

- How to choose ARMA order
- Backshift operator and how to use it to simplify an ARMA model
- How to check stationarity
- Limitations
- ARIMA

IRF

Steps to Analyze Dynamic Response:

1. Apply a one-time shock (e.g., set $e_t = 1$, $e_{t-1} = 0$, $e_{t-2} = 0, \ldots$).

2. Observe how the series y_t evolves using the model equations.

- What does equilibrium mean?

- IRF for AR, MA and ARMA and how does it inform us about the characteristics of AR and MA

VAR

- What it is used for
- What kind of information or insight can you gain from it
- Formulation
- VAR vs VARMA
- How to choose number of lags

FEVD

- Formula
- Sources of shocks in VAR
- How is it presented?
- FEVD vs IRF

Cointegration

- Spurious regression/ correlation and their causes
- R^2
- Why differencing does not work for long term
- Definition of cointegration
- How to verify cointegration

Error Correction Model

- Formulation
- What is it used for
- The rôle of the adjustment variable
- The 3 cases

ARIMAX

- Principle
- N-step forecasting and exogeneous variables
- Differencing for exogeneous variables
- Multicollinearity and how to handle it
- Cross correlation