DATA MINING 2
Time Series - Stationarity and Forecasting

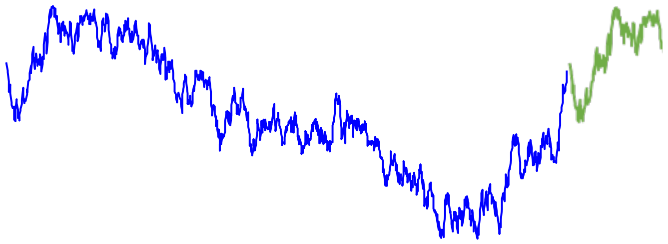
Riccardo Guidotti

a.a. 2019/2020

UNIVERSITÀ DI PISA

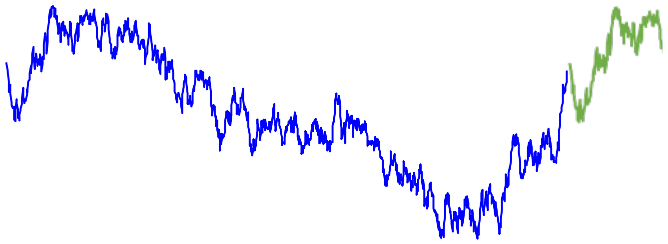
Time Series Forecasting (Prediction)

- Main difference between forecasting and classification: forecasting is about predicting a future state/value, rather than a current one.
- Applications:
  - Temperature, Humidity, CO2 Emissions
  - Epidemics
  - Pricing, Sales Volumes, Stocks
  - Forewarning of Natural Disasters (flooding, hurricane, snowstorm),
  - Electricity Consumption/Demands
- Techniques:
  - Statistical Methods,
  - Machine Learning Classifiers
  - Deep Neural Networks



Forecasting vs Regression

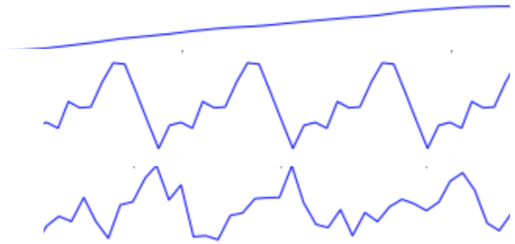
- Forecasting is **time dependent:** the basic assumption of a linear regression model that the observations are independent does not hold.
- Along with an increasing or decreasing **trend**, most TS have some form of **seasonality** trends, i.e. variations specific to a particular time frame.



Time Series Characteristics

Time Series Components

- A given TS consists of three systematic components including level, trend, seasonality, and one non-systematic component called noise.
  - **Level**: The average value in the series.
  - **Trend**: The increasing or decreasing value in the series.
  - **Seasonality**: The repeating short-term cycle in the series.
  - **Noise**: The random variation in the series.



- A **systematic** component have consistency or recurrence and can be described and modeled.
- A **Non-Systematic** component cannot be directly modeled.

Combining Time Series Components

- A TS is an aggregate or combination of these four components.
- All series have a level and noise. The trend and seasonality components are optional.

- **Additive Model**: y(t) = Level + Trend + Seasonality + Noise
  - Changes over time are consistently made by the same amount
  - A linear trend is a straight line.
  - A linear seasonality has the same frequency (width of cycles) and amplitude (height of cycles).

- **Multiplicative Model**: y(t) = Level * Trend * Seasonality * Noise
  - A multiplicative model is nonlinear, such as quadratic or exponential. Changes increase or decrease over time.
  - A nonlinear trend is a curved line.
  - A non-linear seasonality has an increasing/decreasing frequency and/or amplitude over time.

Time Series Models

- A TS model specifies the **joint distribution function** of the sequence $x_1, x_2, \ldots, x_n$ of n random variables as the probability that the values of the series are jointly less than n constants $c_1, c_2, \ldots, c_n$.
  - $F(c_1, c_2, \ldots, c_n) = P(x_1 \leq c_1, x_2 \leq c_2, \ldots, x_n \leq c_n)$

- Although the joint distribution function describes the data completely, it is an unwieldy tool for analyzing TS data

Time Series Descriptive Measures

- Another informative marginal descriptive measure is the **mean function** $\mu_t = E(x)$ where E denotes the expected value operator.

- The lack of independence between two subsequent values x_s at time *i* and x_t at time *i+k* can be assessed numerically, as in classical statistics, using the notions of **covariance** and **correlation**.
- Assuming the variance of a TS *x* is finite, we have the following definitions.

Time Series Descriptive Measures

- The **autocovariance function (AF)** is defined as
  - $\gamma_x(s,t) = cov(x_s, x_t) = E[(x_s - \mu_s)(x_t - \mu_t)]$

- It measures the linear dependence between lagged TS starting at two different time points on the same TS.
- Very smooth TS exhibit AF that stay large even when the *t* and *s* are far apart, whereas choppy TS tend to have AF that are nearly zero for large separations.
- If $\gamma_x(s,t)$ = 0 are not linearly related,
- For *s* = *t*, the AF reduces to the (assumed finite) variance, because
  - $\gamma_x(t,t) = var(x_t) = E[(x_t - \mu_t)^2]$

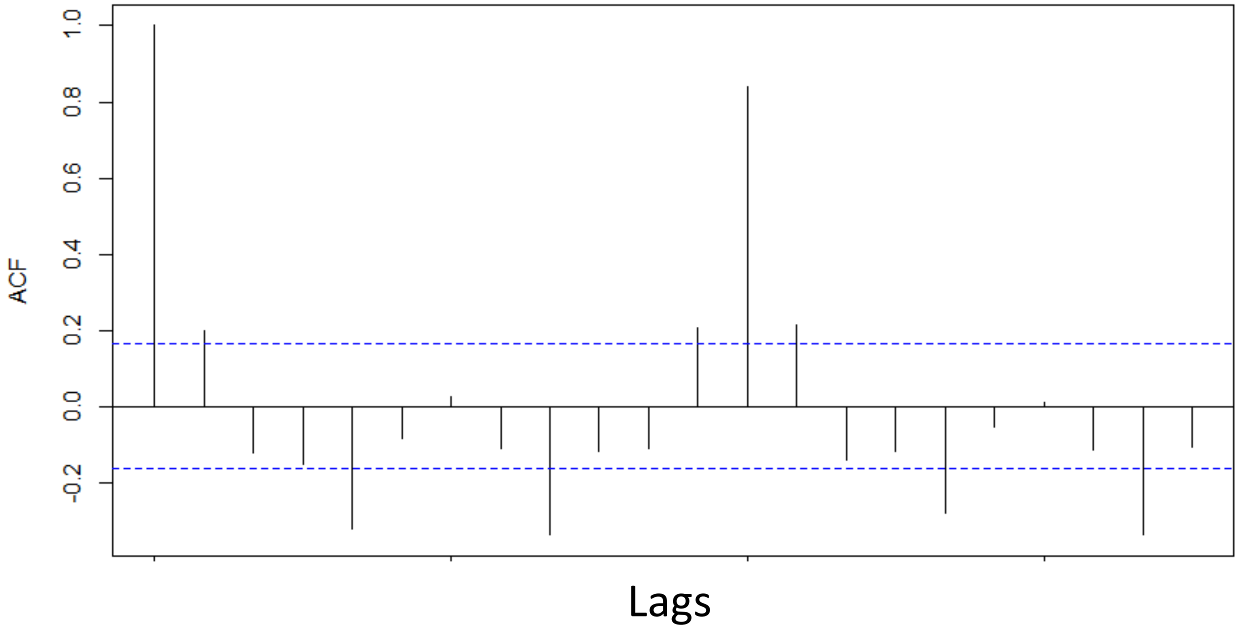
Time Series Descriptive Measures

$$r_k = \frac{\sum_{t=k+1}^{T} (y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}^{T} (y_t - \bar{y})^2},$$

- The **autocorrelation function (ACF)** is defined as
  - $\rho(s,t) = \frac{\gamma(s,t)}{\sqrt{\gamma(s,s)\gamma(t,t)}} \in [-1,1]$
- It measures the linear predictability of the series at time *t*, say x_t, using only the values from time *s*, x_s
- Hence, we have a rough measure of the ability to forecast the series at time *t* from the value at time *s*.
- ACF measures the linear relationship between lagged values of a TS.
- There are several autocorrelation coefficients, corresponding to each lag *k* = 1, 2, 3, …

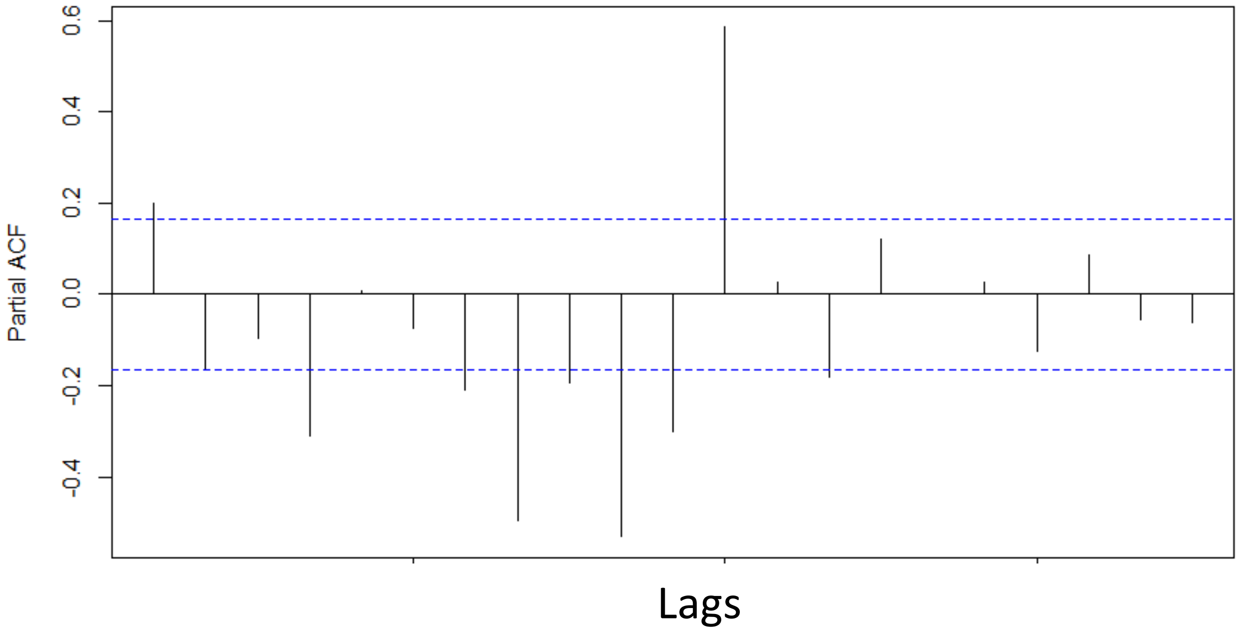
ACF plot

- The ACF plot shows the total correlation between different lag functions by calculating the correlation for TS with observations with previous time steps, called lags.
- Thus we calculate the ACF for x_t with x_{t+1} x_{t+2}, *etc.*



PACF plot

- A partial autocorrelation is a summary of the relationship between an observation in a TS with observations at prior time steps with the relationships of intervening observations *removed*.
- The partial autocorrelation at lag *k* is the correlation that results after removing the effect of any correlations due to the terms at shorter lags.



ACF and PACF Summary

- **Autocorrelation Function (ACF)**: It is a measure of the correlation between the TS with a lagged version of itself.
  - For instance at lag *k=5*, ACF would compare TS at time instant $t_1 \ldots t_n$ with TS at instant t_1-5, …, t_2-5 (t_1-5 and t_n being end points).
- **Partial Autocorrelation Function (PACF):** This measures the correlation between the TS with a lagged version of itself but after eliminating the variations already explained by the intervening comparisons.
  - Eg at lag *k=5*, would compare the correlation but remove the effects already explained by lags 1 to 4.

White Noise

- The differenced series is the *change* between consecutive observations in the original series, and can be written as $x'_t = x_t - x_{t-1}$
- Time series that show no autocorrelation are called **white noise**.
- In other words it is made of random values with a given mean and standard deviation but not autocorrelation.
- When the differenced series is white noise, i.e. $\varepsilon_t = x_t - x_{t-1}$, where ε_t denotes white noise, than $x_t = x_{t-1} + \varepsilon_t$ is a **random walk model**

Random Walk

- Random walk models are widely used for non-stationary data, e.g. financial and economic data.

- Random walks typically have:
  - long periods of apparent trends up or down
  - sudden and unpredictable changes in direction.

- The forecasts from a random walk model are equal to the last observation, as future movements are unpredictable, and are equally likely to be up or down.

Stationarity

Stationary Time Series

- A **strictly stationary** TS is one for which the probabilistic behavior of every collection of values $\{x_1, x_2, \ldots, x_n\}$ is identical to that of the time shifted set $\{x_{1+h}, x_{2+h}, \ldots, x_{n+h}\}$
  - $P(x_1 \leq c_1, x_2 \leq c_2, \ldots, x_n \leq c_n) = P(x_{1+h} \leq c_1, x_{2+h} \leq c_2, \ldots, x_{n+h} \leq c_n)$
  - for all n =1,2,…, all time points 1, 2, …, n, all numbers $c_1, c_2, \ldots, c_n$, all time shifts h.

- If a TS is strictly stationary, then all of the distribution functions for subsets of variables must agree with their counterparts in the shifted set for all values of the shift parameter h.
- In other words, shifting the time axis does not affect the distribution.

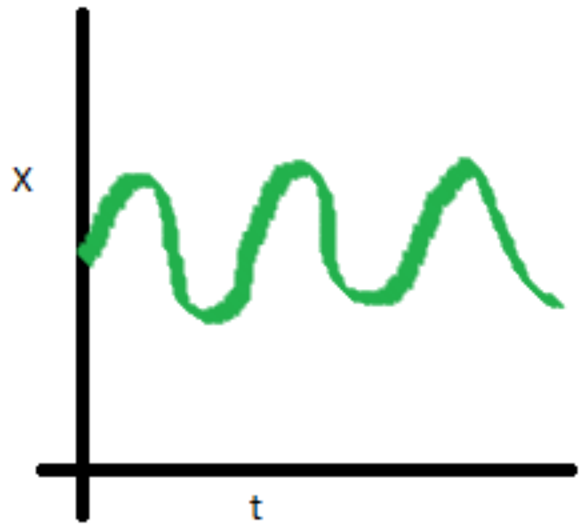
Stationary Time Series

- A **weakly stationary** TS, x_t, is a finite variance process such that
  - the mean value function, μ_t is constant and does not depend on time $\mu_t = \mu$
  - the autocovariance function, $\gamma(s, t)$ depends on *s* and *t* only through their difference $|s\text{-}t|$.
- We will use the term stationary to mean weakly stationary.

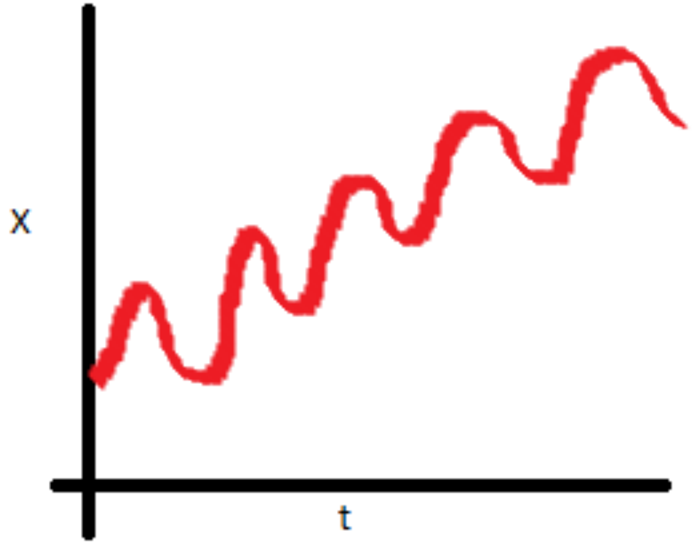
- A TS with a certain trend or with a certain seasonlity is not stationary.
- In practice, there are three basic criterion for a TS to be stationary

Stationary Time Series

- The mean of the series should not be a function of time, rather should be a constant.



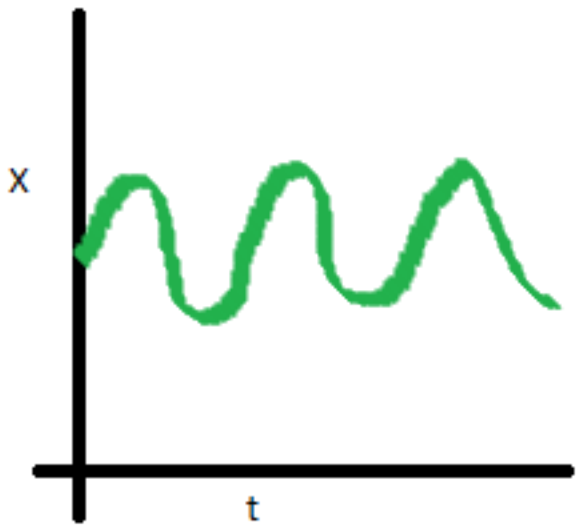
Stationary series



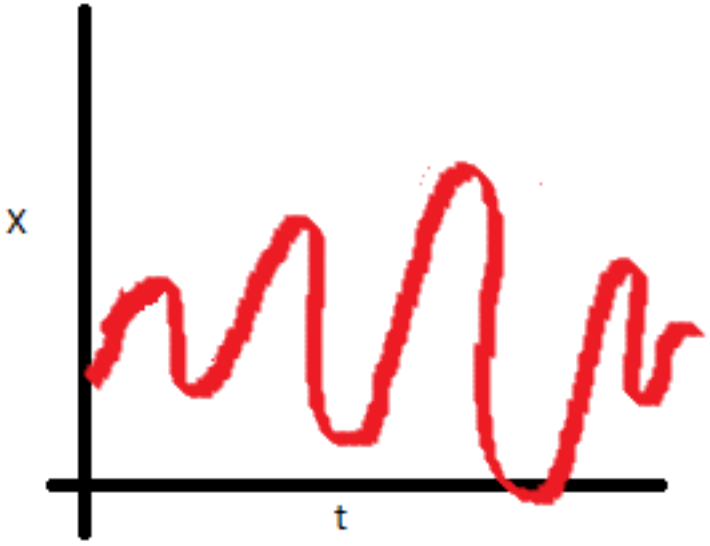
Non-Stationary series

Stationary Time Series

- The variance of the series should not a be a function of time, rather should be a constant.



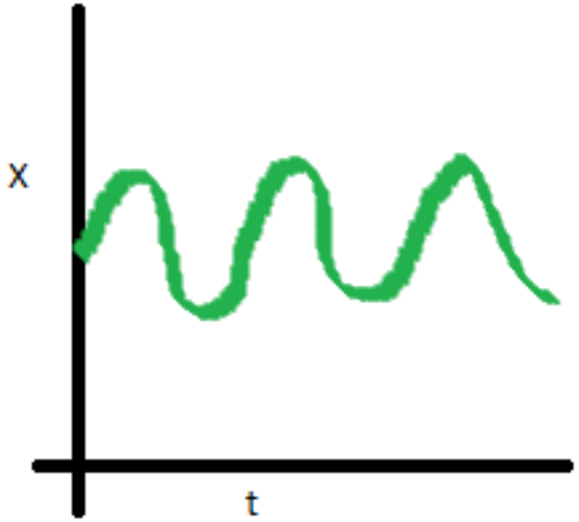
Stationary series



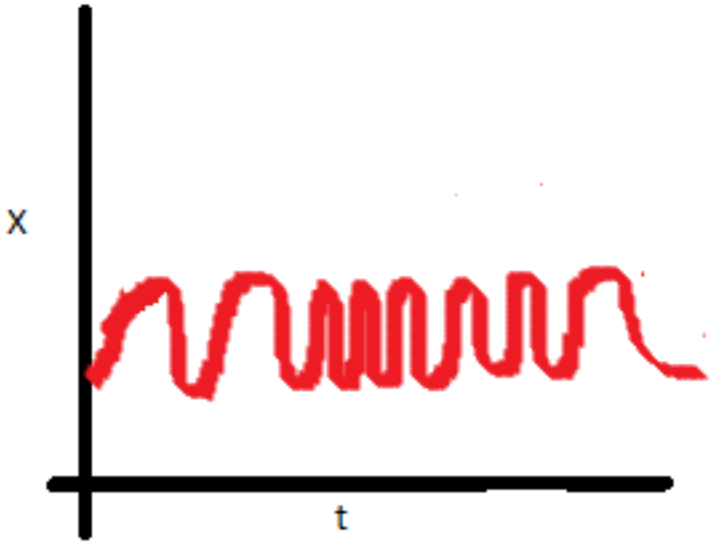
Non-Stationary series

Stationary Time Series

- The covariance of the i-th point and the (i+k)-th point should not be a function of time.



Stationary series



Non-Stationary series

Dickey Fuller Test of Stationarity

- The test results comprise of a Test Statistic and some Critical Values for different confidence levels.
- If the Test Statistic is less than the Critical Value, we can reject the null hypothesis and say that the series is stationary.
- The Dickey–Fuller test tests the null hypothesis that a unit root is present in an autoregressive model.
- A unit root is a feature of some stochastic processes (such as random walks) that can cause problems in statistical inference involving TS models.

```
Results of Dickey-Fuller Test:
Test Statistic                   0.815369
p-value                          0.991880
#Lags Used                      13.000000
Number of Observations Used    130.000000
Critical Value (5%)             -2.884042
Critical Value (1%)             -3.481682
Critical Value (10%)            -2.578770
dtype: float64
```

Dickey Fuller Test of Stationarity

```
Results of Dickey-Fuller Test:
Test Statistic                   0.815369
p-value                          0.991880
#Lags Used                      13.000000
Number of Observations Used    130.000000
Critical Value (5%)             -2.884042
Critical Value (1%)             -3.481682
Critical Value (10%)            -2.578770
dtype: float64
```

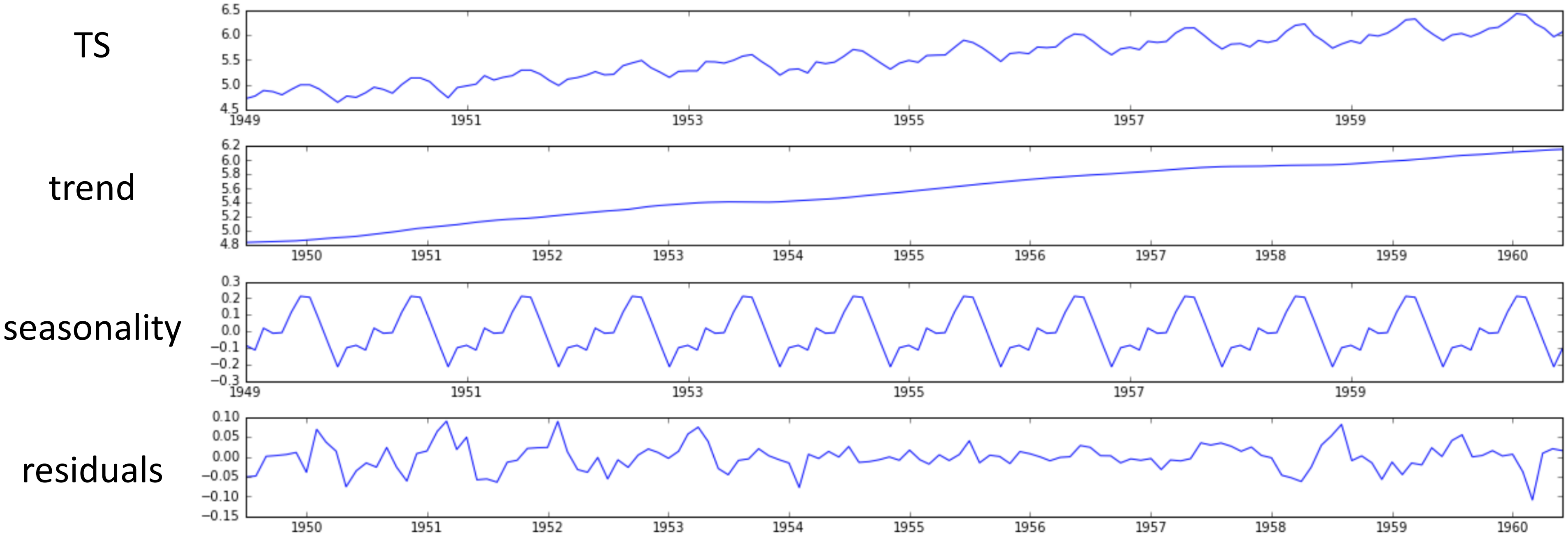
- First we build an autoregressive model
  - $y_t = \alpha y_{t-1} + u_t$
  - where y_t is the TS, t the time index, α a coefficient, and u_t the error term.
  - a unit root is present if $\alpha = 1$.
- We rewrite it as
  - $\Delta y_t = (\alpha - 1)y_{t-1} + u_t = \delta y_{t-1} + u_t$
  - where $\Delta y_t = y_t - y_{t-1}$ is the difference operator
  - a unit root is present if $\delta = 0$
- Then a test on α is run to understand if it is lower or equal than 1 (i.e., $\delta = 0$).
- If the null hypothesis is accepted then a trend exists.
- Since the test is done over the residual term Δy_t rather than raw data, it is not possible to use standard t-distribution to provide critical values.
- Therefore, this test has a specific distribution known as the Dickey–Fuller table.

Why Do I Care About Stationarity?

- If your TS is not stationary, you cannot build a TS predictive model.
- In cases where the stationary criterion are violated, the first requisite becomes to stationaries the TS.
- There are multiple ways of bringing stationarity by removing trend and/or seasonality.
- Some of them are Detrending, Differencing, Decomposition, etc.

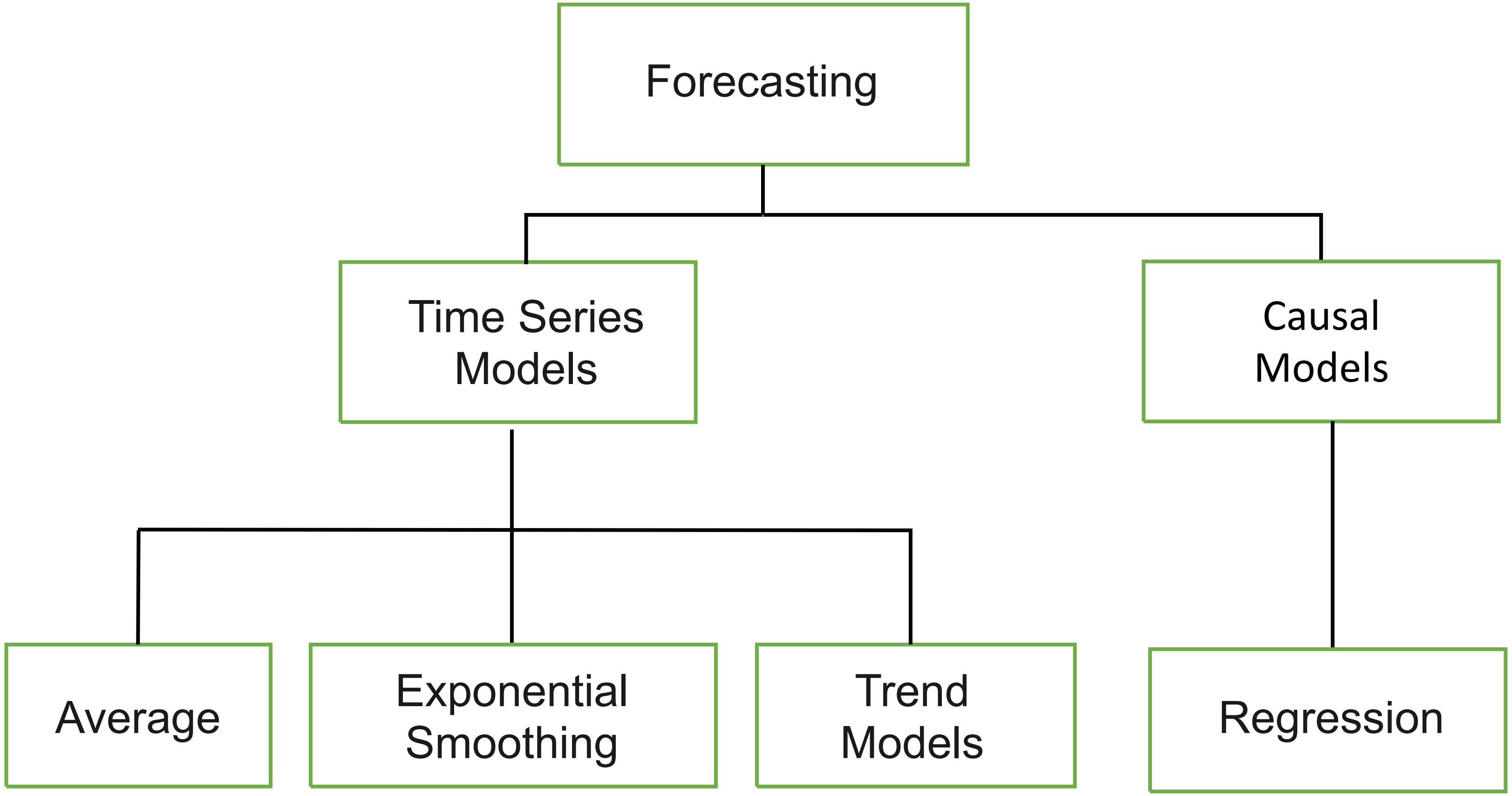
Eliminating Trend and Seasonality

- **Differencing**: we take the difference of the observation at a particular instant with that at the previous instant.
- **Detrending:** we simply remove the trend component from the TS.
- **Decomposing**: *trend* and *seasonality* are modeled separately and the remaining part of the TS, i.e., the *residual*, is returned.



Time Series Forecasting

It's Difficult to Make Predictions, Especially About the Future



ES and ARIMA models are the two most widely used approaches to time series forecasting, and provide complementary approaches to the problem.

Evaluating Forecast Accuracy

- A forecast “error” is the difference between an observed value and its forecast. An “error” is not a mistake, is the unpredictable part.

$$e_{T+h} = y_{T+h} - \hat{y}_{T+h|T}$$

- Forecast errors are different from residuals:
  - Residuals are calculated on the training set while forecast errors are calculated on the test set.
  - Residuals are based on one-step forecasts while forecast errors can involve multi-step forecasts.
- We can measure forecast accuracy by summarizing the forecast errors in different ways.

Scale-Dependent Errors

- Cannot be used to make comparisons between TS that involve different units.
- The two most commonly used scale-dependent measures are based on the absolute errors or squared errors:

$$\text{Mean absolute error: MAE} = \text{mean}(|e_t|),$$
$$\text{Root mean squared error: RMSE} = \sqrt{\text{mean}(e_t^2)}.$$

Percentage Errors

- Percentage errors are unit-free, and so are frequently used to compare forecast performances between data sets.
- The percentage error is given by

$$p_t = 100e_t/y_t$$

- The most commonly used measure is:

$$\text{Mean absolute percentage error: MAPE} = \text{mean}(|p_t|)$$

- Total and Median Absolute Percentage Error (TAPE, MedianApe) are also used.

Evaluation Measures from Regression

- **Coefficient of determination** R^2
  - is the proportion of the variance in the dependent variable that is predictable from the independent variable(s)

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

hat means predicted

$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\epsilon_i^2$

- **Mean Squared/Absolute Error** MSE/MAE
  - a risk metric corresponding to the expected value of the squared (quadratic)/absolute error or loss

$$\text{MSE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}}\sum_{i=0}^{n_{\text{samples}}-1}(y_i - \hat{y}_i)^2 \quad \text{MAE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}}\sum_{i=0}^{n_{\text{samples}}-1}|y_i - \hat{y}_i|$$

Simple Forecasting Methods

Simple Forecasting Methods

- **Average Method**: the forecasts of all future values are equal to the average (or “mean”) of the historical data.

$$\hat{y}_{T+h|T} = \bar{y} = (y_1 + \cdots + y_T)/T.$$

- **Naïve Method**: the forecasts of all future values are equal to the last value of the historical data.

$$\hat{y}_{T+h|T} = y_T.$$

- **Drift Method**: increase/decrease last value w.r.t. the amount of change over time (*drift*) as the average change in the historical data.

$$\hat{y}_{T+h|T} = y_T + \frac{h}{T-1}\sum_{t=2}^{T}(y_t - y_{t-1}) = y_T + h\left(\frac{y_T - y_1}{T-1}\right)$$

Exponential Smoothing

Simple Exponential Smoothing (SES)

- Is suitable for data with no clear trend or seasonal pattern.
- SES is in between the average and naive method.
- SES attaches larger weights to more recent observations than to observations from the distant past, while smallest weights are associated with the oldest observations
- Forecasts are calculated using weighted averages, where the weights decrease exponentially as observations come from further in the past.

$$\hat{y}_{T+1|T} = \alpha y_T + \alpha(1-\alpha) y_{T-1} + \alpha(1-\alpha)^2 y_{T-2} + \cdots$$

- $0 \leq \alpha \leq 1$ is the smoothing parameter

SES – Formalization in Components

- For SES the only component used is the level.
- Component form representations of SES comprise a forecast equation and a smoothing equation for each of the components in the method.

$$
\begin{aligned}
\text{Forecast equation} \qquad & \hat{y}_{t+h|t} = \ell_t \\
\text{Smoothing equation} \qquad & \ell_t = \alpha y_t + (1 - \alpha)\ell_{t-1}
\end{aligned}
$$

- where l_t is the level of the TS at time t

Holt's Linear Trend Method

- Holt extended simple exponential smoothing to allow the forecasting of data with a trend.

$$\begin{aligned} \text{Forecast equation} \quad & \hat{y}_{t+h|t} = \ell_t + hb_t \\ \text{Level equation} \quad & \ell_t = \alpha y_t + (1-\alpha)(\ell_{t-1} + b_{t-1}) \\ \text{Trend equation} \quad & b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1} \end{aligned}$$

- where l_t is the level of the TS at time t, b_t estimates the trend of TS, $0 \leq \alpha \leq 1$ is the smoothing parameter for the level and $0 \leq \beta^* \leq 1$ is the smoothing parameter for the trend.

Holt-Winters' Seasonal Method

- Holt (1957) and Winters (1960) extended Holt's method to capture seasonality.
- m denotes the frequency of the seasonality, i.e., the number of seasons in a reference period, while $0 \leq \gamma \leq 1 - \alpha$ is the smoothing parameter for the seasonality.
- The additive method is preferred when the seasonal variations are constant through the TS
- The multiplicative method is preferred when the seasonal variations are changing proportional to the level of the TS.

Holt-Winters' Seasonal Method

- Additive

$$\hat{y}_{t+h|t} = \ell_t + hb_t + s_{t+h-m(k+1)}$$
$$\ell_t = \alpha(y_t - s_{t-m}) + (1-\alpha)(\ell_{t-1} + b_{t-1})$$
$$b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1}$$
$$s_t = \gamma(y_t - \ell_{t-1} - b_{t-1}) + (1-\gamma)s_{t-m},$$

- Multiplicative

$$\hat{y}_{t+h|t} = (\ell_t + hb_t)s_{t+h-m(k+1)}$$
$$\ell_t = \alpha\frac{y_t}{s_{t-m}} + (1-\alpha)(\ell_{t-1} + b_{t-1})$$
$$b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1}$$
$$s_t = \gamma\frac{y_t}{(\ell_{t-1} + b_{t-1})} + (1-\gamma)s_{t-m}$$

k is the integer part of (h−1)/m, which ensures that the estimates of the seasonal indices come from the final period of the sample.

More on Exponential Smoothing

- ES methods are not restricted to those we have presented.

Trend	Seasonal N	Seasonal A	Seasonal M
N	$\hat{y}_{t+h\|t} = \ell_t$ $\ell_t = \alpha y_t + (1-\alpha)\ell_{t-1}$	$\hat{y}_{t+h\|t} = \ell_t + s_{t+h-m(k+1)}$ $\ell_t = \alpha(y_t - s_{t-m}) + (1-\alpha)\ell_{t-1}$ $s_t = \gamma(y_t - \ell_{t-1}) + (1-\gamma)s_{t-m}$	$\hat{y}_{t+h\|t} = \ell_t s_{t+h-m(k+1)}$ $\ell_t = \alpha(y_t/s_{t-m}) + (1-\alpha)\ell_{t-1}$ $s_t = \gamma(y_t/\ell_{t-1}) + (1-\gamma)s_{t-m}$
A	$\hat{y}_{t+h\|t} = \ell_t + hb_t$ $\ell_t = \alpha y_t + (1-\alpha)(\ell_{t-1} + b_{t-1})$ $b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1}$	$\hat{y}_{t+h\|t} = \ell_t + hb_t + s_{t+h-m(k+1)}$ $\ell_t = \alpha(y_t - s_{t-m}) + (1-\alpha)(\ell_{t-1} + b_{t-1})$ $b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1}$ $s_t = \gamma(y_t - \ell_{t-1} - b_{t-1}) + (1-\gamma)s_{t-m}$	$\hat{y}_{t+h\|t} = (\ell_t + hb_t)s_{t+h-m(k+1)}$ $\ell_t = \alpha(y_t/s_{t-m}) + (1-\alpha)(\ell_{t-1} + b_{t-1})$ $b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1}$ $s_t = \gamma(y_t/(\ell_{t-1} + b_{t-1})) + (1-\gamma)s_{t-m}$
A_d	$\hat{y}_{t+h\|t} = \ell_t + \phi_h b_t$ $\ell_t = \alpha y_t + (1-\alpha)(\ell_{t-1} + \phi b_{t-1})$ $b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)\phi b_{t-1}$	$\hat{y}_{t+h\|t} = \ell_t + \phi_h b_t + s_{t+h-m(k+1)}$ $\ell_t = \alpha(y_t - s_{t-m}) + (1-\alpha)(\ell_{t-1} + \phi b_{t-1})$ $b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)\phi b_{t-1}$ $s_t = \gamma(y_t - \ell_{t-1} - \phi b_{t-1}) + (1-\gamma)s_{t-m}$	$\hat{y}_{t+h\|t} = (\ell_t + \phi_h b_t)s_{t+h-m(k+1)}$ $\ell_t = \alpha(y_t/s_{t-m}) + (1-\alpha)(\ell_{t-1} + \phi b_{t-1})$ $b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)\phi b_{t-1}$ $s_t = \gamma(y_t/(\ell_{t-1} + \phi b_{t-1})) + (1-\gamma)s_{t-m}$

ARIMA Models

Auto-Regressive Integrated Moving Averages

- The ARIMA forecasting for a stationary time series is a linear equation (like a linear regression).
- While *ES* are based on a *description of the trend and seasonality*, *ARIMA* models aim to describe the *autocorrelations* in the data.
- Before we introduce ARIMA models, we recall the concept of stationarity and the technique of differencing TS.

Stationarity (again)

- A stationary TS is one whose properties do not depend on the time at which the series is observed.
- TS with trends, or with seasonality, are not stationary: the trend and seasonality affect the value of the TS at different times.
- A white noise series is stationary: it does not matter when you observe it, it looks much the same at any point in time.

Differencing (again)

- Differencing: compute the differences between consecutive observations.
- It is a possible transformation to make a non-stationary TS stationary.
- Indeed, it can help stabilize the mean of a TS by removing changes in the level, and thus eliminating (or reducing) trend and seasonality.
- In addition, transformations such as logarithms can help to stabilize the variance of a time series.

Autoregressive Models

- In multiple *regression* model, we predict the variable of interest using a linear combination of predictors.
- In an autoregression model, we forecast the variable of interest using a linear combination of past values of the variable.
- The term *autoregression* indicates that it is a regression of the variable against itself.
- An autoregressive model of order *p* can be written as

white noise

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t$$

- This is as an **AR(p) model** of order *p* (p = lag in the past)

Autoregressive Models

- We normally restrict AR models to stationary data, in which case some constraints on the values of the parameters are required.
- For AR(1): $-1 \leq \phi_1 \leq 1$
- For AR(2): $-1 \leq \phi_2 \leq 1,\ \phi_1 + \phi_2 < 1,\ \phi_2 - \phi_1 < 1$
- When *p>2* the restriction are much more complicated.

Moving Average Models

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

white noise → ε_t

$$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}$$

- This is as a **MA(q) model** of order *q* (q = lag in the past).
- MA models should not be confused with the moving average smoothing.
- It is possible to write any stationary AR(p) as MA(∞)

Moving Average Models

- It is possible to write any stationary AR(p) as MA(∞)
- The reverse result holds if we impose some constraints on the MA parameters.
- Then the MA model is called **invertible.**
- The invertibility constraints for other models are similar to the stationarity constraints.
- For MA(1): $-1 \leq \theta_1 \leq 1$
- For MA(2): $-1 \leq \theta_2 \leq 1, \theta_1 + \theta_2 > -1, \theta_1 - \theta_2 < 1$
- When p>2 the restriction are much more complicated.

ARIMA Models (Non-Seasonal)

- If we combine differencing with an AR model and a MA model, we obtain a non-seasonal ARIMA model. ARIMA is an acronym for AutoRegressive Integrated Moving Average (“integration” is the reverse of differencing).

$$y'_t = c + \phi_1 y'_{t-1} + \cdots + \phi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t$$

- where y'_t is the differenced series.
- We call this model **ARIMA(p,d,q) model**, where *p* is the order of the autoregressive part, *d* is the degree of first differencing involved, *q* is the order of the moving average part

ARIMA Models (Non-Seasonal)

- The same stationarity and invertibility conditions that are used for AR and MA models also apply to an ARIMA model.
- Special cases of ARIMA models

White noise	ARIMA(0,0,0)
Random walk	ARIMA(0,1,0) with no constant
Random walk with drift	ARIMA(0,1,0) with a constant
Autoregression	ARIMA(p,0,0)
Moving average	ARIMA(0,0,q)

- ARIMA(p,0,q) is also called ARMA(p,q)

ACF and PACF plots (again)

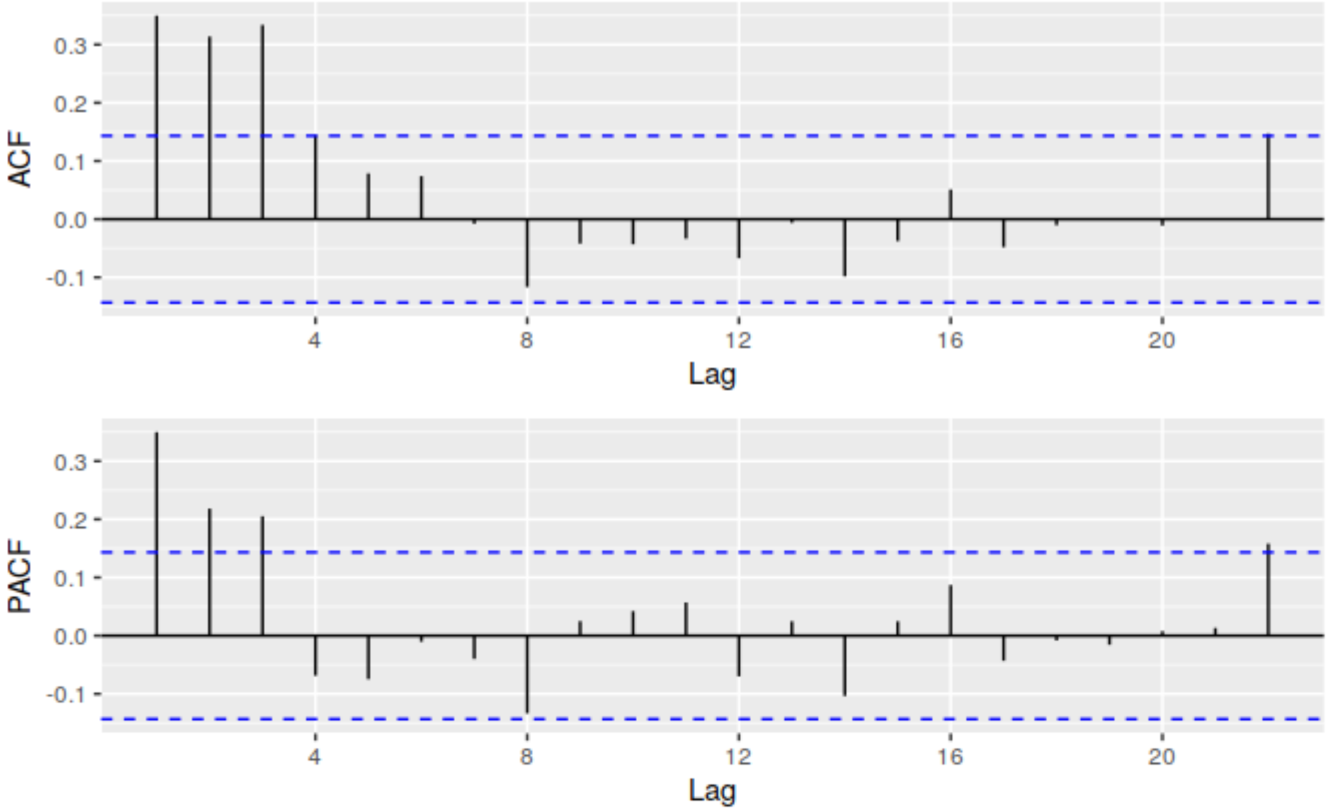
- It is sometimes possible to use the ACF plot, and the closely related PACF plot, to determine appropriate values for p and q.
- **ACF** plot shows the autocorrelations which measure the relationship between y_t and y_{t-k} for different values of k.
- **PACF** measure the relationship between y_t and y_{t-k} after removing the effects of lags $1,2,3,\ldots,k-1$.
- If the TS are from an ARIMA(p,d,0) or ARIMA(0,d,q), then the ACF and PACF plots can be helpful in determining the value of p or q.
- If p and q are both positive, then the plots do not help in finding suitable values of p and q.

ACF and PACF plots (again)

- The TS may follow an ARIMA(p,d,0) model if the ACF and PACF plots of the differenced TS show the following patterns:
  - the ACF is exponentially decaying or sinusoidal;
  - there is a significant spike at lag p in the PACF, but none beyond lag p.
- The data may follow an ARIMA(0,d,q) model if the ACF and PACF plots of the differenced TS show the following patterns:
  - the PACF is exponentially decaying or sinusoidal;
  - there is a significant spike at lag q in the ACF, but none beyond lag q.

ACF and PACF plots - Example

- There are three spikes in the ACF, followed by an almost significant spike at lag 4. In the PACF, there are three significant spikes, and then no significant spikes.
- The pattern in the first three spikes is what we would expect from an ARIMA(3,0,0), as the PACF tends to decrease.
- So in this case, the ACF and PACF lead us to think an ARIMA(3,0,0) model might be appropriate.



ARIMA – Parameters Estimation

- Once the model order has been identified (i.e., the values of *p,d,q*), we need to estimate the parameters $c, \phi_1, \dots, \phi_p, \theta_1, \dots, \theta_p$.
- *Maximum Likelihood Estimation* (MLE) can be used to find the values for these parameters.
- For ARIMA models, MLE is similar to the *least squares* estimates that would be obtained by minimizing

$$\sum_{t=1}^{T} \varepsilon_t^2$$

- Once the parameters are estimated they are placed in the equation and used to make the prediction of $y_{t+1}, y_{t+2}, \dots, y_{t+n}$

Determining the order of an ARIMA model

- Akaike's Information Criterion (AIC) $$\text{AIC} = -2\log(L) + 2(p+q+k+1)$$

- Bayesian Information Criterion (BIC) $$\text{BIC} = \text{AIC} + [\log(T) - 2](p+q+k+1)$$

- k=1 if c=0, k=0 otherwise
- Good models are obtained by minimizing the AIC, or BIC
- We highlight that AIC, or BIC are not good guides to selecting the appropriate *d*, but only for selecting *p* and *q*.
- This is because the differencing changes the data on which the likelihood is computed, making the AIC values between models with different orders of differencing not comparable.

Modelling Procedure

1. Visualize the time series
2. Stationarize the series
3. Plot ACF/PACF charts and find optimal parameters
4. Build the ARIMA model
5. Make Predictions

Seasonal ARIMA Models

- ARIMA models can also model 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.

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

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

Advanced Forecasting Methods

Advanced Forecasting Methods

- Machine Learning models in form of (auto-)regressors can be used for time series forecasting.
- Decision Tree Regressors
- (Deep) Neural Networks Regressors
  - Convolutional Neural Networks
  - Recurrent Neural Networks
- Ensemble Regressors
  - Bagging
  - Bootstrapping
  - Random Forest Regressors

References

- Forecasting: Principles and Practic. Rob J Hyndman and George Athanasaopoulus. (https://otexts.com/fpp2/)
- Time Series Analysis and Its Applications. Robert H. Shumway and David S. Stoffer. 4th edition.(https://www.stat.pitt.edu/stoffer/tsa4/tsa4.pdf)
- Mining Time Series Data. Chotirat Ann Ratanamahatana et al. 2010. (https://www.researchgate.net/publication/227001229_Mining_Time_Series_Data)
- Dynamic Programming Algorithm Optimization for Spoken Word Recognition. Hiroaki Sakode et al. 1978.
- Experiencing SAX: a Novel Symbolic Representation of Time Series. Jessica Line et al. 2009
- Compression-based data mining of sequential data. Eamonn Keogh et al. 2007.