Timeseries-ARIMA

Goal → univariate time series forecasting problems

Presented by: Sarit Maitra

Alliance University, Bengaluru, India

Fundamentals

Time Series Nomenclature

- Time Series Analysis → time series forecasting.
- t, t+1, t-1
- Level - baseline value for the series if it were a straight line.
- Trend - increasing or decreasing behavior of the series over time.
- Seasonality - repeating patterns or cycles of behavior over time.
- Noise - variability in the observations that cannot be explained by the model.

All time series have a level, most have noise, trend and seasonality optional.

Describe vs Predict

- Objective of time series analysis is to develop mathematical models
- Descriptive modeling - to determine seasonal patterns, trends, relation to external factors, etc.
- Predict - uses the information in a time series to predict / forecast future values

Forecasting

When forecasting, it is important to understand our goal.

- How much data do you have available and are you able to gather it all together?
- What is the time horizon of predictions that is required? Short, medium or long term?
- Can forecasts be updated frequently over time or must they be made once and remain static?
- At what temporal frequency are forecasts required?

Examples

- Forecasting the agri yield by state each year
- Forecasting the closing price of a stock each day.
- Forecasting sales in units sold each day for a store
- Forecasting the traffic each day.
- Forecasting unemployment for a state each quarter

Data preparation

Explore time series data

- Smoothing technique reduce the random variation in the observations.
- Log transformation - effective to smoothen exponential variance
- Moving average → $MA_{20}(t) = \frac{1}{20} \sum_{i=t-19}^{t} obs(i)$

Temporal structure

White noise

Model Diagnostic

- Time series is white noise if the variables are IID with a mean of zero
- If time series is white noise, then, by definition, it is random
- Errors from forecast model should ideally be white noise → $y(t) = signal(t)\ +\ noise(t)$

White noise identification

- line plot - features like a changing mean, variance, or obvious relationship between lagged variables.
- summary statistics - mean and variance of the whole series against the mean and variance of meaningful contiguous blocks of values in the series (e.g. days, months, or years).
- autocorrelation plot - correlation between lagged variables

Decompose time series

- Additive model $\rightarrow y(t) = Level + Trend + Seasonality + Noise$
- Multiplicative model $\rightarrow y(t) = Level \times Trend \times Seasonality \times Noise$

Trends

- Deterministic Trends → These are trends that consistently increase or decrease.
- Stochastic Trends → These are trends that increase and decrease inconsistently
- Identify Trend
- Remove trend

Seasonality

Types of seasonality

- Daily
- Weekly
- Monthly
- Yearly.

Simple way to correct for a seasonal component is to use differencing.

Stationary

- Linear model assumes that underlying data are a realization of a stationary process.
- So, our first step to check whether there is any evidence of a trend or seasonal effects and, if there is, remove them.
  - Look at plot
  - review summary statistics
  - Statistical Tests

ADF test

- ADF test statistical test called a unit root test .
- It determines how strongly a time series is defined by a trend.
- Null Hypothesis (H0): Fail to reject → time series has a unit root, meaning it is non-stationary.
- Alternate Hypothesis (H1): The null hypothesis is rejected → time series does not have a unit root.
  - $p > 0.05$: Fail to reject the null hypothesis (H0).
  - $p \leq 0.05$: Reject the null hypothesis (H0).

https://en.wikipedia.org/wiki/Augmented_Dickey%E2%80%93Fuller_test

Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American statistical association*, *74*(366a), 427-431.

Model evaluation

Performance Measures

- Forecast Error ($expected\ value\ -\ predicted\ value$)
- Mean Forecast Error / bias ($mean(forecast\ error)$)
- Mean Absolute Error $(mean(abs(forecast\ error))$
- Mean Squared Error ($mean(forecast\ error^2)$)
- Root Mean Squared Error ($\sqrt{mean\ squared\ error}$)

Residuals diagnostic

- Residual Line Plot → we expect the plot to be random around the value of 0.
- Residual Summary Statistics → we are interested in the mean value of the residual errors.
  - ≈ 0 suggests no bias in the forecasts, whereas positive and negative values suggest a positive or negative bias.
- Residual Histogram and Density Plots → we expect the forecast errors to be normally distributed around a zero mean.
- Residual Q-Q Plot → check the normality of the distribution of residual errors.
- Residual Autocorrelation Plot → we would not expect any correlation between the residuals.

Persistence model / Naïve Forecast

- Establishing a baseline is essential on any time series forecasting problem.
- A baseline in performance gives an idea of how well all other models will actually perform on the problem.
- Simplest forecast :
  - previous time step ⇒ next time step (y_{t+1}=y_t)
  - Forecast = Value from the same season in the previous cycle ($\hat{y}_{t+1}$=y_{t+1-m}, m=seasonality period.)

Box-Jenkins Method

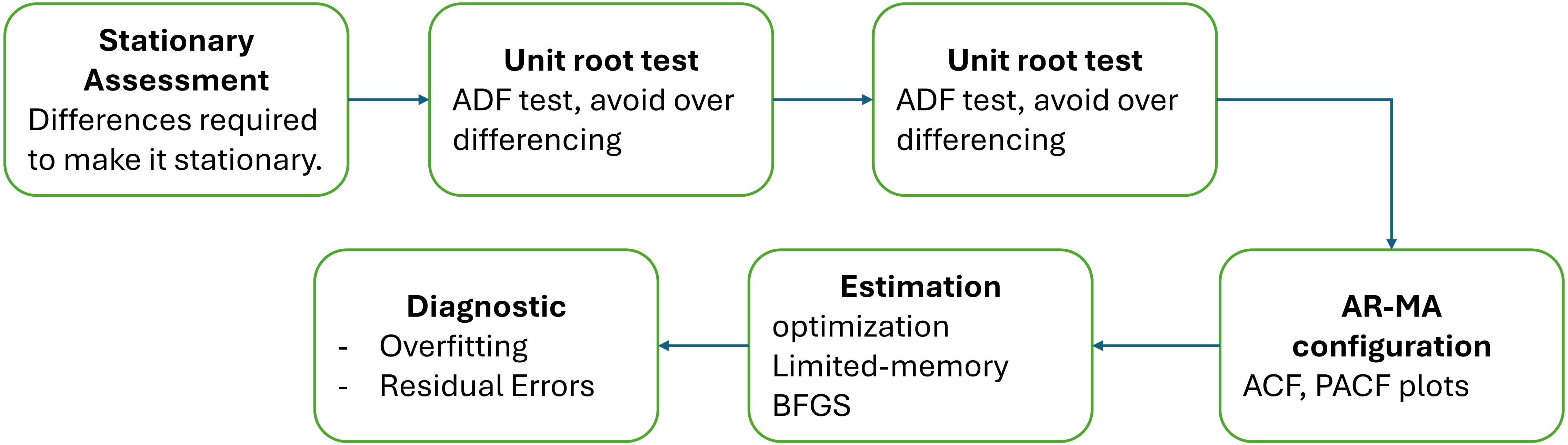
George Box and Gwilym Jenkins (1970), *Time Series Analysis: Forecasting and Control*

- Assumption that the process that generated the time series can be approximated using:
  - ARMA model if it is stationary
  - ARIMA model if it is non-stationary.
- ARIMA process (AR: Autoregression, I: Integrated and MA: Moving Average):
  - AR (p) → dependent relationship with the lagged observations.
  - I (d) → differencing to make series stationary
  - MA (q) → dependency between an observation and residual errors from a moving average model applied to lagged observations

ARIMA

- Each components are explicitly specified in the model as a parameter.
- Standard notation is used of ARIMA(p, d, q) where the parameters are substituted with integer values.
  - p → number of lag included in the model (lag order)
  - d → number of times observations are differenced (degree of differencing)
  - q → size of the moving average window (order of moving average)

Parameter identification



AR model

- Observations from previous time steps as input to predict the value at the next time step:

linear regression model→ $\hat{y} = \beta_0 + \beta_1 X_1 + \varepsilon$

AR (2) → $X_{(t\,+\,1)} = c + \phi_1 X_t + \phi_2 X_{t-1}$

standard AR(p) → $X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \phi_p X_{t-p} + \varepsilon_t$

- ε_t → Shock at time t ~ white noise ⇒ IID with E(ε_t) : stochastic disturbance term interpreted as random shocks.

AR Model summary

AR Model → Current value depends on past values of the series.

$$X_t = c + \sum_{i=1}^{p} \phi_i X_{t-i} + \varepsilon_t$$

predict next time step →

$$X_{t+1} = c + \sum_{i=1}^{p} \phi_i X_{t+1-i} + \varepsilon_{t+1}$$

MA model

- X_t depends only on the lagged forecast errors.

$$X_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} \ldots + \theta_q \varepsilon_{t-q}$$

- error terms → Past shocks directly influencing X_t
- We cannot observe ε_t directly in practice, they are latent, and must be estimated during model fitting
- Unlike AR models, past values of X do not appear on the right-hand side.
- Captures short-term dependencies or sudden shocks in the data.

MA model summary

- MA Model → Current value depends on past unobserved error terms (shocks).

$$X_t = \mu + \varepsilon_t + \sum_{j=1}^{q} \theta_j \, \varepsilon_{t-j}$$

- AR model → ε_t represents random shocks that affect X_t, not explicitly modeled over time (only today's shock enters directly).
- MA model → the same type of shocks ε_t used, their lagged effects are explicitly modeled, past shocks continue to influence future values for q periods.

AR / MA summary

- AR → Current value depends on past values of the series.
- MA → Current value depends on past unobserved errors.
- ε_t in AR → instantaneous shocks to X_t
- ε_t in MA → Shock whose lagged effects are modeled directly.

AR+MA Model

$$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} \ldots + \theta_q \varepsilon_{t-q}$$

- AR Model → for persistent trends (e.g., stock prices).
- MA Model → for shock-driven series (e.g., volatile returns).
- ARMA → Combines both.

Stationarity

- stationary → statistical properties do not change over time → ensure patterns are stable for reliable forecasts.
- Non stationary →
  - Trend: Long-term increase/decrease in mean (e.g., GDP growth).
  - Seasonality: Regular fluctuations (e.g., monthly sales peaks).
  - Structural Breaks: Sudden changes in behavior (e.g., post-pandemic)
- Stationarity test →
  - ADF Test : H_0 = non-stationary.
  - KPSS Test: H_0 = stationary.

Make series stationary

- Differencing:
  - $\Delta X_t = X_t - X_{t-1}$(removes trend)
  - $\Delta_{12} X_t = X_t - X_{t-12}$
- Transformations → Logarithm → Stabilizes variance.
- Decomposition → separate trend, seasonality, and residuals