ARIMA: Intro

to use non-stationary data → differencing

- $d \times 1$: remove temporal dependence on prev value
- $d \times 2$: non-linear " " removed

seasonality / trend

ARIMA (p, d, q)

d → Integration component

How to select optimal parameters p, d, q?

Autocorrelation

how much of current related to prev

+ve : large values ⟺ large values at lag

-ve : " " ⟺ small " " "

$$\rho_k = \frac{Cov(Y_{t-k}, Y_t)}{Var(Y_t)} = \frac{\sum_{t=k+1}^{n} (Y_{t-k} - \bar{Y})(Y_t - \bar{Y})}{\sum_{t=k+1}^{n} (Y_t - \bar{Y})^2}$$

Auto Correlation Function (ACF)

plot of autocors at diff. lags k → correlogram

Partial Auto Cor. Function (PACF)

remove influence from intermediate autocors b/w Y_{t-k} and Y_t

plot at lag 2: diff. b/w actual corr and expected corr due to propagation from lag 1

$$\rho_{pk}\big|_{k=2} = \frac{Cov(Y_{t-2}, Y_t \mid Y_{t-1})}{\sqrt{Var(Y_t \mid Y_{t-1})\, Var(Y_{t-2} \mid Y_{t-1})}}$$

Estimating p, q

$H_0: \rho_k = 0$ $\quad H_A: \rho_k \neq 0$

$H_0: \rho_{pk} = 0$ $\quad H_A: \rho_{pk} \neq 0$

Reject H_0 when $\rho_k > \frac{1.96}{\sqrt{n}}$, $\rho_{pk} > \frac{1.96}{\sqrt{n}}$

p → PACF

$\rho_{pk} > \frac{1.96}{\sqrt{n}}$ for first p values then cuts off to 0.

ρ_k decreases exponentially

AR(p)

q → ACF

$\rho_k > \frac{1.96}{\sqrt{n}}$ $\quad q$

ρ_{pk}

MA(q)

Statistical Analysis of Stationarity

ACF slow decay to zero → non stationary

cut off to zero → stationary

Dickey-Fuller Test

For a AR(1) process: $Y_{t+1} = \beta Y_t + \varepsilon_{t+1}$

$H_0: \beta = 1$ → non-stationary, presence of Unit Root

$H_1: \beta < 1$ → stationary

Alternatively,

$Y_{t+1} - Y_t = \beta Y_t - Y_t + \varepsilon_{t+1} \implies \Delta Y_t = Y_t(\beta - 1) + \varepsilon_{t+1}$

$$\boxed{\begin{aligned} &\Rightarrow \Delta Y_t = \psi Y_t + \varepsilon_{t+1} \\ &H_0: \psi = 0 \\ &H_1: \psi < 0 \end{aligned}}$$

Test statistic:

calculated like t-statistic, is NOT t-distrn

$$\boxed{DF \text{ statistic} = \frac{\psi}{SE(\psi)}}$$

Augmented Dickey-Fuller Test

DF assumes $\varepsilon_{t+1} \in$ white noise. If not: we add p lagged terms to soak up corr in errors

$$\boxed{\Delta Y_t = \psi Y_t + \sum_{i=1}^{p} \alpha_i Y_{t-i} + \varepsilon_{t+1}}$$

Everything else remains the same

Converting Non-Stationary To Stationary

1. identify cause → trend/seasonality
2. if trend: trend stationarity

   removing trend leaves stationary series

   $Y_t = \mu_t + \varepsilon_t$ → deterministic mean trend

   → De-trend = DG − Trend

   → Rescaling removes trend also. Eg: log scale
3. if not trend: difference stationarity

   mean trend is stochastic in nature → differencing

   $\nabla Y_t = Y_t - Y_{t-1}$] linear gone

   $\nabla^2 Y_t = \nabla Y_t - \nabla Y_{t-1}$

   $= [Y_t - Y_{t-1}] - [Y_{t-1} - Y_{t-2}]$

   $= Y_t - 2Y_{t-1} + Y_{t-2}$

   mostly everything gone $d \leq 2$

Random Walk

first differenced series = white noise

$Y_t - Y_{t-1} = \varepsilon_t$

$Y_t = Y_{t-1} + \varepsilon_t$

Naive → forecast same as prev value with some unpredictable change

Non-stationary, sudden changes in direction

with drift:

$Y_t - Y_{t-1} = \varepsilon_t + c$ $\quad$ c: avg of changes between consecutive observations

$Y_t = Y_{t-1} + \varepsilon_t + c$

$c > 0$: Y_t increases → drift upwards

$c < 0$: Y_t decreases → drift downwards

ARIMA Model Building

1. identify model from ACF, PACF
2. estimate parameters

   AIC (Akaike Info Criteria) $= -2(LL) + 2k$ → $p+q$

   BIC (Bayesian " ") $= -2(LL) + k\ln(n)$

   BIC: higher penalty than AIC for every additional var added

   LL: log likelihood

   n: no of observations
3. Validate

   Ljung-Box test for goodness of fit

NOTE:

non-stationary → stationary

⟹ undifferencing at the end

$d=1$: Actual = ARMA + Y_{t-1}

$d=2$: Actual = ARMA + $2Y_{t-1} - Y_{t-2}$

- +ve ρ_k at lag 1 → AR (borderline underdifferenced)
- -ve ρ_k at lag 1 → MA (mildly overdifferenced)
- MA(1) : fine tune effect of nonseasonal difference
- AR(1) : compensate for lack of " "

SARIMA: Seasonal . . .

$$Y_t = c + \sum_{n=1}^{p} a_n Y_{t-n} + \sum_{n=1}^{q} \theta_n \varepsilon_{t-n} + \sum_{n=1}^{P} \phi_n Y_{t-sn} + \sum_{n=1}^{Q} \eta_n \varepsilon_{t-sn} + \varepsilon_t$$

P: seasonal autoregressive terms

D: " differences

Q: " moving avg terms

Seasonal differencing

$D = 1$

$\nabla_s Y_t = Y_t - Y_{t-s}$

$\boxed{\text{regular first then seasonal}}$

- $P, D, Q \leq 1$
- only one of P or Q can be 1 at a time
- $d + D \leq 2$

$d = 1, D = 1$

1. $\nabla Y_t = Y_t - Y_{t-1}$
2. $\nabla_s Y_t = (Y_t - Y_{t-s}) - (Y_{t-1} - Y_{t-1-s})$

X: Exogenous variables

Externals that affect input and therefore forecasting.

Easier to just add them in model.

ARIMAX: → autoregression + regression

$$y_t = c + \sum_{n=1}^{p} a_n y_{t-n} + \sum_{n=1}^{q} \theta_n \varepsilon_{t-n} + \sum_{u=1}^{[x]} \beta_n x_{n_t} + \varepsilon_t$$

SARIMAX:

$$Y_t = c + \sum_{n=1}^{p} a_n Y_{t-n} + \sum_{n=1}^{q} \theta_n \varepsilon_{t-n} + \sum_{n=1}^{P} \phi_n Y_{t-sn} + \sum_{n=1}^{Q} \eta_n \varepsilon_{t-sn} + \sum_{n=1}^{n} \beta_n x_{n_t} + \varepsilon_t$$