Chapter 5: ARIMA Models

When successive observations show serial dependence, forecasting methods based on exponential smoothing may be inefficient.
Why?
They fail to take advantage of the serial dependence in the data effectively.

I. Linear models for Stationary Series

Consider the Linear Filter or Linear Operator; L:

$$y_t = L(x_t) = \sum_{i=-\infty}^{\infty} \psi_i x_{t-i} \quad , \quad t = \ldots, -1, 0, 1, \ldots$$

This is a process that converts the input x_t into an output y_t. Also, the conversion involves all values of the input in the form of a summation with different weights that are time-invariant. Finally if $\sum_{-\infty}^{\infty} |\psi_i| < \infty$, then this filter is also stable.

A. Stationarity

Recall that we are interested in "Weak Stationarity":

- Expected Value is not dependent on time
- ~~AC~~ Autocovariance at any lag is NOT a function of time, but of the given lag.

∴ A stationary series exhibits a "similar" statistical behaviour (distribution) in time.

B. <u>Stationary Time Series</u>

Suppose $\{x_t\}$ is a stationary series, with

$$E(x_t) = \mu_x \text{ , and}$$
$$Cov(x_t, x_{t+k}) = \gamma_x(k).$$

Then, if $y_t = L(x_t) = \sum_{i=-\infty}^{\infty} \psi_i x_{t-i}$ is a linear filter,

$$E(y_t) = \mu_y = \sum_{i=-\infty}^{\infty} \psi_i \mu_x \text{ , and}$$
$$Cov(y_t, y_{t+k}) = \gamma_y(k) = \sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} \psi_i \psi_j \gamma_x(i-j+k).$$

→ Consider the following <u>stable linear process</u> with white noise, ε_t :

$$y_t = \mu + \sum_{i=0}^{\infty} \psi_i \varepsilon_{t-i} \text{ , } E(\varepsilon_t) = 0 ; \gamma_\varepsilon(h) = \begin{cases} \sigma^2 ; h=0 \\ 0 ; h \neq 0 \end{cases}$$
$$= \mu + \psi_0 \varepsilon_t + \psi_1 \varepsilon_{t-1} + \psi_2 \varepsilon_{t-2} + \cdots$$

Then $E(y_t) = \mu$, and

$$\gamma_y(k) = \sigma^2 \sum_{i=0}^{\infty} \psi_i \psi_{i+k}.$$

<u>Infinite Moving Average</u> :

$$y_t = \mu + \psi_0 \varepsilon_t + \psi_1 \varepsilon_{t-1} + \psi_2 \varepsilon_{t-2} + \cdots$$
$$= \mu + \underbrace{\sum_{i=0}^{\infty} \psi_i B^i}_{\Psi(B)} \varepsilon_t = \mu + \Psi(B) \cdot \varepsilon_t .$$

This is the* WOLD DECOMPOSITION THEOREM.

The Wold decomp. theorem states that a stationary time series can be seen as the weighted ~~or~~ sum of the present and past "disturbances" that are random.

→ Note that it is critical that these disturbances are random ⇒ <u>Uncorrelated</u> random shocks that have <u>Constant</u> variance.

→ Also Uncorrelated $\neq$ Independent.

Independence $\Longrightarrow$ $f(X,Y) = f_X(X) \cdot f_Y(Y)$

⇒ Knowing X does not give any info. about Y.

Uncorrelated ⇒ $E(XY) = E(X) \cdot E(Y)$.

This happens <u>because</u> $Cov(X,Y) = E(XY) - E(X) \cdot E(Y) = 0$

→ Finally, <u>Independent</u> $\Longrightarrow$ <u>Uncorrelated</u>, but <u>Uncorrelated</u> $\not\Longrightarrow$ <u>Independence</u>.

For example, consider X and $Y = |X|$.

Clearly not independent, but can show that they are uncorrelated!

In this chapter, when using Wold's decomp. Theorem, we will only need "uncorrelated" random disturbances, not "independent" random errors.

II. Finite Order Moving Average, MA(q), processes.

MA(q) : $y_t = \mu + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q}$, where

$\{\varepsilon_t\}$ is white noise.

In terms of the backshift operator, B,

$$\begin{aligned} \text{MA}(q): y_t &= \mu + (\varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q}) \\ &= \mu + (1 - \theta_1 B - \cdots - \theta_q B^q)\varepsilon_t \\ &= \mu + \left(1 - \sum_{i=1}^{q} \theta_i B^i\right)\varepsilon_t \end{aligned}$$

$\therefore \boxed{y_t = \mu + \Theta(B)\,\varepsilon_t}$, where $\Theta(B) = 1 - \sum_{i=1}^{q} \theta_i B^i$

Also, (a) $E(y_t) = E(\mu + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q})$
$= \mu$.

(b) $\text{Var}(y_t) = \text{Var}(\mu + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q})$
$\Rightarrow \gamma_y(0) = \sigma^2(1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2)$.

(c) $\gamma_y(k) = \text{Cov}(y_t, y_{t+k})$
$= E\left[(\varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q})(\varepsilon_{t+k} - \theta_1 \varepsilon_{t+k-1} - \cdots - \theta_q \varepsilon_{t+k-q})\right]$

$$= \begin{cases} \sigma^2(-\theta_k + \theta_1 \theta_{k+1} + \cdots + \theta_{q-k}\theta_q) & ; k = 1, 2, \ldots, q \\ 0 & ; \text{o/w} \cdot (k > q) \end{cases}$$

③ $$\rho_y(k) = \frac{\gamma_y(k)}{\gamma_y(0)} = \begin{cases} \dfrac{-\theta_k + \theta_1\theta_{k+1} + \cdots + \theta_{q-k}\theta_q}{1+\theta_1^2 + \cdots + \theta_q^2} & ; k=1,2,\ldots,q \\ 0 & ; k>q. \end{cases}$$

Note : The ACF of an MA(q) model "cuts off" after lag 'q'. [Or becomes very small in absolute value after lag 'q']

A. First-Order Moving Average process, MA(1)

MA(1): $y_t = \mu + \varepsilon_t - \theta_1 \varepsilon_{t-1}$; $q = 1$.

$$\gamma_y(0) = \sigma^2 (1 + \theta_1^2)$$
$$\gamma_y(1) = -\theta_1 \sigma^2$$
$$\gamma_y(k) = 0 \;, \; k = 2, 3, \ldots$$
$$\rho_y(1) = \frac{-\theta_1}{1+\theta_1^2}$$
$$\rho_y(k) = 0 \;, \; k = 2, 3, \ldots$$

Note that the ACF for an MA(1) process cuts off after lag 1.

B. Second – Order Moving average process, MA(2).

$$y_t = \mu + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2}$$
$$= \mu + (1 - \theta_1 B - \theta_2 B^2)\varepsilon_t$$

$$\gamma_y(0) = \sigma^2(1 + \theta_1^2 + \theta_2^2)$$
$$\gamma_y(1) = \sigma^2(-\theta_1 + \theta_1\theta_2)$$
$$\gamma_y(2) = \sigma^2(-\theta_2)$$
$$\gamma_y(k) = 0 \quad , \; k = 3, 4, \cdots \text{ or } k > 2$$
$$\rho_y(1) = \frac{-\theta_1 + \theta_1\theta_2}{1 + \theta_1^2 + \theta_2^2}$$
$$\rho_y(2) = \frac{-\theta_2}{1 + \theta_1^2 + \theta_2^2}$$
$$\rho_y(k) = 0 \quad \text{for} \quad k > 2$$

Note that the ACF for an MA(2) process cuts off after lag 2.

III. Finite – Order ~~AR~~ Autoregressive Processes, AR(p)

Albeit powerful, the Wold decomposition Theorem requires us to estimate infinitely many weights, $\{\psi_i\}$

One interepretation of the finite order, MA(q) process is that at any given time, only a finite number of the infinitely many past disturbances "contribute" to the current value of the time series, and as the time window of the contributors moves in time, the oldest disturbance becomes obsolete for the next observation.

A. <u>First-Order Autoregressive process</u>, AR(1)

$$y_t = \mu + \sum_{i=0}^{\infty} \psi_i \varepsilon_{t-i}$$
$$= \mu + \sum_{i=0}^{\infty} \psi_i B^i \varepsilon_t$$
$$= \mu + \Psi(B)\, \varepsilon_t$$

One approach is to assume that the contributions of the disturbances that are way in the past should be small compared to the more recent disturbances.

Since the disturbances are i.i.d; we can assume a set of exponentially decaying weights for this purpose.

So, set $\psi_i = \phi^i$, where $|\phi| < 1$. Then

$$y_t = \mu + \varepsilon_t + \phi \varepsilon_{t-1} + \phi^2 \varepsilon_{t-2} + \cdots$$
$$= \mu + \sum_{i=0}^{\infty} \phi^i \varepsilon_{t-i}$$

Also, $y_{t-1} = \mu + \varepsilon_{t-1} + \phi\, \varepsilon_{t-2} + \phi^2 \varepsilon_{t-3} + \cdots$

$\Rightarrow \phi\, y_{t-1} = \phi\mu + \phi \varepsilon_{t-1} + \phi^2 \varepsilon_{t-2} + \phi^3 \varepsilon_{t-3} + \cdots$

$\therefore \phi\, y_{t-1} - \phi\mu = \phi \varepsilon_{t-1} + \phi^2 \varepsilon_{t-2} + \phi^3 \varepsilon_{t-3} + \cdots$

Substituting this in y_t, we have

$y_t = \mu + \varepsilon_t + \phi\, y_{t-1} - \phi\mu$

$$\Rightarrow \quad \boxed{y_t = \delta + \phi y_{t-1} + \varepsilon_t} \quad ; \quad \delta = (1-\phi)\mu .$$

This is an AR(1) process $\Rightarrow$ regressing y_t on y_{t-1}.
This a CAUSAL AR model.
Note also that this model is stationary if $|\phi| < 1$.
[If $|\phi| > 1$, the time series is explosive. Although there is a stationary solution for an AR(1) model when $|\phi|>1$ it results in a NON-CAUSAL model, where forecasting requires knowledge about the future!]

AR(1): $E(y_t) = \mu = \dfrac{\delta}{1-\phi}$

$$Var(y_t) = \gamma_y(0) = \sigma^2 \cdot \frac{1}{1-\phi^2} .$$

$$Cov(y_t, y_{t+k}) = \gamma_y(k) = \sigma^2 \cdot \phi^k \cdot \frac{1}{1-\phi^2} \quad ; \quad k = 0, 1, 2, \ldots$$

$$\rho(k) = \frac{\gamma_y(k)}{\gamma_y(0)} = \phi^k \quad ; \quad k = 0, 1, 2, \ldots$$

Exponential decay form.

B. <u>Second-Order Autoregressive Process, AR(2)</u>

$$y_t = \delta + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t .$$

$$\Rightarrow \quad y_t - \phi_1 y_{t-1} - \phi_2 y_{t-2} = \delta + \varepsilon_t$$

$$\Rightarrow \quad (1 - \phi_1 B - \phi_2 B^2) y_t = \delta + \varepsilon_t$$

$$\Rightarrow \quad \Phi(B) y_t = \delta + \varepsilon_t .$$

$$\Rightarrow \quad y_t = \underbrace{\Phi(B)^{-1}\delta}_{\mu} + \underbrace{\Phi(B)^{-1}}_{\Psi(B)} \cdot \varepsilon_t .$$

$$\Rightarrow \quad y_t = \mu + \Psi(B)\varepsilon_t$$

$$\Rightarrow \quad y_t = \mu + \sum_{i=0}^{\infty} \psi_i \varepsilon_{t-i}$$

$$\Rightarrow \quad y_t = \mu + \sum_{i=0}^{\infty} \psi_i B^i \varepsilon_t$$

$$\Rightarrow \quad \therefore \boxed{y_t = \mu + \sum_{i=0}^{\infty} \psi_i B^i \varepsilon_t} \text{, where}$$

IMA Representation of AR(2) model.

$$\mu = \Phi(B)^{-1}\delta \cancel{= (1-\phi_1 B - \phi_2 B^2)^{-1}\delta} \text{, and}$$

$$\Phi(B)^{-1} = \sum_{i=0}^{\infty} \psi_i B^i = \Psi(B)$$

IMP. Q: What makes this AR(2) model <u>stationary</u> and <u>causal</u>?

Solve the polynomial : $m^2 - \phi_1 m - \phi_2 = 0$

Roots of the equation are:

$$m_1, m_2 = \frac{\phi_1 \pm \sqrt{\phi_1^2 + 4\phi^2}}{2} \quad \left[\frac{-b \pm \sqrt{b^2 - 4ac}}{2a}\right]$$

If $|m_1|, |m_2| < 1$, then $\sum_{i=0}^{\infty} |\psi_i| < \infty$.

(Also, for complex roots, see p. 343 MJK)

$\therefore$ For an AR(2) to be stationary & have an IMA representatio[n]

$$\boxed{\begin{aligned} &\phi_1 + \phi_2 < 1, \\ &\phi_2 - \phi_1 < 1, \\ &|\phi_2| < 1 \end{aligned}}$$

$$E(y_t) = \mu = \frac{\delta}{1-\phi_1-\phi_2} \quad ; \quad 1-\phi_1-\phi_2 \neq 0$$

$$\begin{cases} Var(y_t) = \gamma_y(0) = \phi_1 \gamma_y(1) + \phi_2 \gamma_y(2) + \sigma^2 \\ Cov(y_t, y_{t+k}) = \gamma_y(k) = \phi_1 \gamma_y(k-1) + \phi_2 \gamma_y(k-2) \; ; \; k = 1, 2, \ldots \end{cases}$$

YULE - WALKER EQUATIONS.

Also, $\rho(k) = \phi_1 \rho(k-1) + \phi_2 \rho(k-2)$, $k = 1, 2, \ldots$

Case 1: If there are two real roots, the ACF is a mixture of two "exponential decay" terms.

Case 2: If there are two complex roots, the ACF has the form of a "damped sinusoid".

Case 3: If there is one real root, m_0 $[m_1 = m_2 = m_0]$, the ACF will exhibit an "exponential decay" pattern.

C. General Autoregressive process, AR(p)

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

This AR(p) is causal and stationary if the roots of the following polynomial are less than 1 in absolute value:

$$m^p - \phi_1 m^{p-1} - \phi_2 m^{p-2} - \cdots - \phi_p = 0 .$$

$$AR(p):\ E(y_t) = \mu = \frac{\delta}{1 - \phi_1 - \phi_2 - \cdots - \phi_p}.$$

$$Var(y_t) = \gamma_y(0) = \sum_{i=1}^{p} \phi_i \gamma(i) + \sigma^2$$

$$Cov(y_t, y_{t+k}) = \gamma_y(k) = \sum_{i=1}^{p} \phi_i \gamma(k-i) \ ; \ k = 1, 2, \cdots$$

Yule-Walker Eqn's: $$\rho(k) = \sum_{i=1}^{p} \phi_i \rho(k-i) \ ; \ k = 1, 2, \ldots$$

As before, depending on the roots of the AR(p) process, the ACF of an AR(p) process can be a mixture of exponential decay and damped sinusoidal patterns.

IV. <u>Partial Autocorrelation Function</u>, PACF

ACF useful for MA(q) processes, but not so for AR(p) processes, because of the structure of the ACF of AR processes (mixture of exp. decay, sinusoidal, etc.)

<u>Partial Correlation</u> = Correlation b/w two variables after adjusting for a common factor(s) that may be affecting them.

PACF between y_t and y_{t-k} is the autocorrelation b/w them after adjusting ~~for~~ for everything ~~in between them~~ pairs.;
i.e., $y_{t-1}, y_{t-2}, \ldots, y_{t-k+1}$.

Impt: Thus, for an AR(p) model, the PACF b/w y_t & y_{t-k} should be zero for all $k > p$.

∴ a.) Use ACF to detect order 'q' of MA(q).
b.) Use PACF to detect order 'p' of AR(p).

A Note on "Invertibility" of time series.

MA(1): Let $y_t = \mu + \varepsilon_t - \theta_1 \varepsilon_{t-1}$

WLOG, let $E(y_t) = \mu = 0$

Then $y_t = \varepsilon_t - \theta_1 \varepsilon_{t-1}$

$$\Rightarrow \quad \begin{aligned} \varepsilon_t &= y_t + \theta_1 \varepsilon_{t-1} \\ &= y_t + \theta_1 [y_{t-1} + \theta_1 \varepsilon_{t-2}] \\ &= y_t + \theta_1 y_{t-1} + \theta_1^2 \varepsilon_{t-2} \\ &\;\vdots \\ &= y_t + \theta_1 y_{t-1} + \theta_1^2 y_{t-2} + \cdots \\ &= \sum_{i=0}^{\infty} \theta_1^i y_{t-i} \end{aligned}$$

If $|\theta_1| < 1$, then $\{\varepsilon_t\}$ is a convergent series.
This is called an INVERTIBLE Process.

And, importantly,
An Invertible MA(q) process can be expressed as an infinite AR(p) process!

V. Mixed Autoregressive – Moving Average Processes, ARMA(p,q)

ARMA(p,q):
$$y_t = \delta + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \dots - \theta_q \varepsilon_{t-q}.$$

$$\Rightarrow y_t = \delta + \sum_{i=1}^{p} \phi_i y_{t-i} + \varepsilon_t - \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}.$$

Stationarity: If all roots of the following polynomial are less than 1 in absolute value, then the ARMA(p, q) process is stationary:
$$m^p - \phi_1 m^{p-1} - \phi_2 m^{p-2} - \dots - \phi_p = 0.$$

Invertibility: If all the roots of the following polynomial are less than 1 in absolute value, then the ARMA(p, q) process is invertible:
$$m^q - \theta_1 m^{q-1} - \theta_2 m^{q-2} - \dots - \theta_q = 0.$$
Then, the ARMA(p, q) has an infinite AR representation

VI. Nonstationary Processes

A time series, y_t, is "homogenous nonstationary" if it is not stationary, but its first difference, or higher order differences ~~are~~ produce a stationary series.

E.g.:
$$w_t = (y_t - y_{t-1}) = (1-B)\, y_t$$
OR
$$w_t = (1-B)^d\, y_t$$
If w_t is stationary, then y_t is homogenous non-stationary.

Let $d = 1$: $y_t = w_t + y_{t-1}$
$= w_t + w_{t-1} + y_{t-2}$
$= w_t + w_{t-1} + w_{t-2} + \dots + w_1 + y_0$

We will call an ARIMA(p, d, q) if its d'th difference, $(1-B)^d$ produces a stationary ARMA(p, q) process.

Example : ARIMA(0, 1, 1) : $(1-B)y_t = \delta + (1-\theta B)\varepsilon_t$.

VII. Building ARIMA models : See R Code.

VIII. Forecasting ARIMA processes :
Best forecast is the one that minimizes the MSE of forecast errors, i.e.;

$$E\left[\left(y_{T+\tau} - \hat{y}_{T+\tau}(T)\right)^2\right] = E\left[e_T(\tau)^2\right]$$

Consider an ARIMA(p, d, q) process at time $T+\tau$.

$$y_{T+\tau} = \delta + \sum_{i=1}^{p+d} \phi_i y_{T+\tau-i} + \varepsilon_{T+\tau} - \sum_{i=1}^{q} \theta_i \varepsilon_{T+\tau-i}$$

(Recall that an ARIMA(p, d, q) process can be represented as

$$\Phi(B)(1-B)^d y_t = \delta + \Theta(B)\varepsilon_t.$$)

→ This process has an IMA representation

$$y_{T+\tau} = \mu + \sum_{i=0}^{\infty} \psi_i \varepsilon_{T+\tau-i}$$

Forecast Error:

$$e_T(\tau) = y_{T+\tau} - \hat{y}_{T+\tau}(T)$$

$$= \sum_{i=0}^{\tau-1} \psi_i \, \varepsilon_{T+\tau-i}$$

Note that: $E[e_T(\tau)] = 0$

$$Var[e_T(\tau)] = \sigma^2(\tau), \quad \tau = 1, 2, \ldots$$

$\Rightarrow$ Variance of FE gets bigger with increasing forecast lead times τ.

The $100(1-\alpha)\%$ P.I. for $y_{T+\tau}$ is:

$$\hat{y}_{T+\tau}(T) \pm z_{\alpha/2} \cdot \sigma(\tau).$$

IX. Seasonal processes:

Data may exhibit strong seasonal/periodic pattern.

Consider a simple additive model:

$$y_t = S_t + N_t, \text{ where}$$

S_t = Deterministic component with periodicty s, and
N_t = Stochastic ~~pr~~ component that may be modeled as an ARMA process.

Also, note that

$$S_t = S_{t+s} = S_{t-s}$$

$$\Rightarrow \quad S_t - S_{t-s} = (1-B^s) S_t = 0.$$

Now, $\quad y_t = S_t + N_t$

$$\Rightarrow (1-B^s) y_t = (1-B^s) S_t + (1-B^s) N_t$$

$$\Rightarrow (1-B^s) y_t = (1-B^s) N_t$$

$$\Rightarrow \quad w_t = (1-B^s) N_t.$$

Since an ARMA process can be used to model N_t, we have

$$\Phi(B) w_t = (1-B^s)\Theta(B)\varepsilon_t, \text{ where}$$

ε_t is white noise.

A more general seasonal ARIMA model of order $(p,d,q) \times (P^*, D^*, Q)$ with period s is:

$$\Phi^*(B^s)\,\Phi(B)\,(1-B)^d\,(1-B^s)^D y_t = \delta + \Theta^*(B^s)\,\Theta(B)\,\varepsilon_t.$$

<u>Example</u>: ARIMA $(0,1,1) \times (0,1,1)$ model with $s=12$

$$(1-B)(1-B^{12}) y_t = (1 - \theta_1 B - \theta_1^* B^{12} + \theta_1 \theta_1^* B^{13})\varepsilon_t.$$

See R code for application of SARIMA model to clothing sales data.