ARMA and ARIMA time series

School of Statistics for Astrophysics
Variability and Time Series Analysis
Autrans 6-11 October 2019

Outline

Preliminaries

Autoregressive AR(p) and moving average stationary time series are frequently used in many fields of applications: econometrics, marketing, engeneering, biostatistics, astrophysics, geophysics, climatology,...
AR(p) models are built in situations where the variable at time t is dependent of his recent past: this will be written in the form:

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \varepsilon_t$$

where (ε_t) is a white noise.
MA(q) time series are useful whan Y_t is the result of shocks in the past; these shocks are represented by a white noise, and are possibly damped as they are far from t. This gives

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

In some situations we can think that the above both dependence forms are acting on (Y_t). We the have an ARMA(p,q) model. Clearly it can be written:

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}$$

Backward operator. Difference operator.
It is useful to define the backward operator ([1]) B:

$$BY_t = Y_{t-1}$$

When iterated we have $B^2Y_t = Y_{t-2}, \dots, B^kY_t = Y_{t-k}, \dots$ and for any scalar α, αB satisfy $(\alpha B)Y_t = \alpha(BY_t) = \alpha Y_{t-1}$ and so on for B^k. It follows that given a polynomial $P(B) = \alpha_0 + \alpha_1 B + \dots + \alpha_r B^r$

$$P(B)Y_t = \alpha_0 Y_t + \alpha_1 Y_{t-1} + \dots + \alpha_r Y_{t-r}$$

Polynomials in B are used along the same algebra as a polynomial of a real (or complex) variable. In particular when $|\alpha| < 1$ ([2])

$$(1-\alpha B)^{-1} = \frac{1}{1-\alpha B} = 1 + \alpha B + \alpha^2 B^2 + \dots = \sum_{j\geq 0} \alpha^j B^j$$

We use also the difference operator $\nabla = (I - B)$:

$$\nabla Y_t = (I - B)Y_t = Y_t - Y_{t-1}$$

Note that we can iterate : $\nabla^2 Y_t = (I-B)^2 Y_t = Y_t - 2Y_{t-1} + Y_{t-2}$ and so on.

[1] Some authors (often in econometrics) prefer use L as Lag

[2] Along the text $|\cdot|$ denotes the absolute value of a real as well as the modulus of a complex number

A linear causal process $Y_t = \varepsilon_t + \psi_1 \varepsilon_{t-1} + \psi_1 \varepsilon_{t-2} + \cdots$ can be written:

$$Y_t \quad = (1 + \psi_1 B + \psi_2 B^2 + \cdots)\varepsilon_t \quad = \psi(B)\varepsilon_t$$

In the same way an invertible process $Y_t = (\pi_1 Y_{t-1} + \pi_2 Y_{t-2} + \cdots) + \varepsilon_t$ will satisfy

$$\begin{aligned} Y_t - \pi_1 Y_{t-1} - \pi_2 Y_{t-2} - \cdots &= \varepsilon_t \\ (1 - \pi_1 B - \pi_2 B^2 - \cdots) Y_t &= \varepsilon_t \\ \pi(B) Y_t &= \varepsilon_t \end{aligned}$$

Autoregressive models

Definition

A mean zero time series (Y_t) is an **AR(p)** series when (Y_t) is stationary and satisfy:

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \varepsilon_t \qquad (1)$$

where $\phi_p \neq 0$ and $\varepsilon_t \sim wn(0, \sigma_\varepsilon^2)$.
When $\mathbb{E}(Y_t) = \mu \neq 0$, (Y_t) is **AR(p)** when (Y_t) is stationary and we have

$$Y_t - \mu = \phi_1(Y_{t-1} - \mu) + \phi_2(Y_{t-2} - \mu) + \cdots + \phi_p(Y_{t-p} - \mu) + \varepsilon_t$$

that is

$$Y_t = \alpha + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \varepsilon_t$$

with $\alpha = \mu(1 - \phi_1 - \phi_2 - \cdots - \phi_p)$.

- (1) can be rewritten $Y_t - \phi_1 Y_{t-1} - \phi_2 Y_{t-2} - \cdots - \phi_p Y_{t-p} = \varepsilon_t$, that is $\phi(B) Y_t = \varepsilon_t$. The operator $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$ is the **autoregressive operator**, and $\phi(z)$ is the **autoregressive polynomial** of the complex variable z.

- A mean zero AR(p) time series is causal iff the roots of $\phi(z) = 0$ lie outside of the unit circle. For example, let (Y_t) be an AR(1) process, $Y_t = \phi_1 Y_{t-1} + \varepsilon_t$, then $\phi(B) = 1 - \phi_1 B$. Hence (Y_t) is causal when $|\phi_1| < 1$.
- The causal representation $Y_t = \psi(B)\varepsilon_t$ is obtained along:

$$\begin{aligned} \phi(B) Y_t &= \varepsilon_t \\ \phi(B)\psi(B)\varepsilon_t &= \varepsilon_t \end{aligned}$$

  which leads to $\phi(B)\psi(B) = 1$. Developing the product on the left and identifying the coefficients in the left and right parts provides the ψ_i coefficients.
  When $(Y_t) \sim \mathrm{AR}(1)$ we easily get $\psi_j = \phi^j$, $j \geq 0$. Thus the causal representation is given by $Y_t = \varepsilon_t + \phi_1\varepsilon_{t-1} + \phi_1^2\varepsilon_{t-2} + \cdots + \phi_1^j\varepsilon_{t-j} + \cdots$.
- An AR(p) time series is clearly invertible with $\pi_i = \phi_i$ for $i \leq p$ and $\pi_i = 0$ for $i > p$.

The AR(1) time series

Let's consider the process $Y_t = \phi_1 Y_{t-1} + \varepsilon_t$, with $|\phi_1| < 1$ and $\varepsilon_t \sim wn(0, \sigma_\varepsilon^2)$. The process (Y_t) is causal, hence stationary, with $\psi_j = \phi_1^j$, i.e. $Y_t = \sum_{j\geq 0} \phi_1^j \varepsilon_{t-j}$. Elementary calculus provides the covariance and correlation. For $h \geq 0$ we have

$$\gamma(h) = \sigma_\varepsilon^2 \frac{\phi_1^h}{1-\phi_1^2} \quad \text{and} \quad \rho(h) = \phi_1^h$$

The partial correlation is given by:

$$\phi_{hh} = \begin{cases} \rho(1) = \phi_1 & \text{for} \quad h = 1 \\ 0 & \text{for} \quad h > 1 \end{cases}$$

The following R code gives 200 observations of two AR(1) series one with $\phi_1 = 0.85$ and another with $\phi_1 = -0.85$. See figure 1.

```
...................................................................
par(mfrow= c(2,1))
phi1<-  0.85; phi2<-  -0.85
set.seed(1001)
y1<- arima.sim(n=200,list(ar=phi1),sd=1.0)
plot(y1,type="line", main=bquote(Y[t]==.(phi1)*Y[t-1]+epsilon[t]))
set.seed(1001)
y2<- arima.sim(n=200,list(ar=phi2),sd=1.0)
plot(y2,type="line", main=bquote(Y[t]==.(phi2)*Y[t-1]+epsilon[t]))
...................................................................
```

The figure 2 gives the theoretical acf and pacf. It shows that the absolue value of the ACF tails off exponentialy and that the PACF cuts off after $h = 1$. We can check also that $\phi_{11} = \phi_1$.

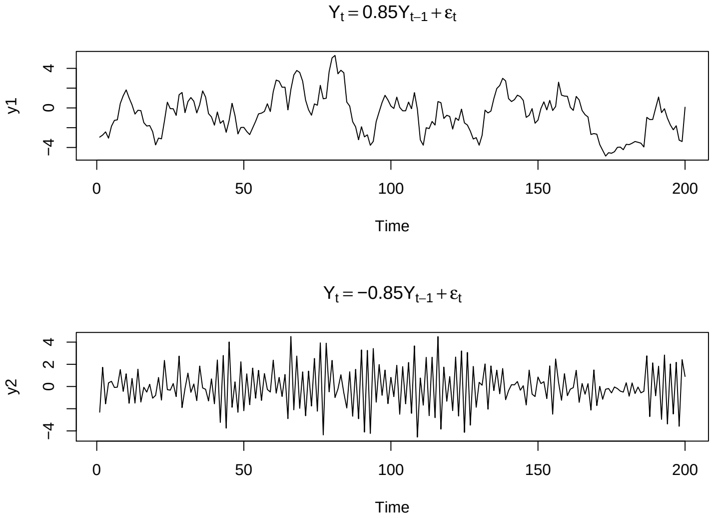


Figure 1: Samples of AR(1) series

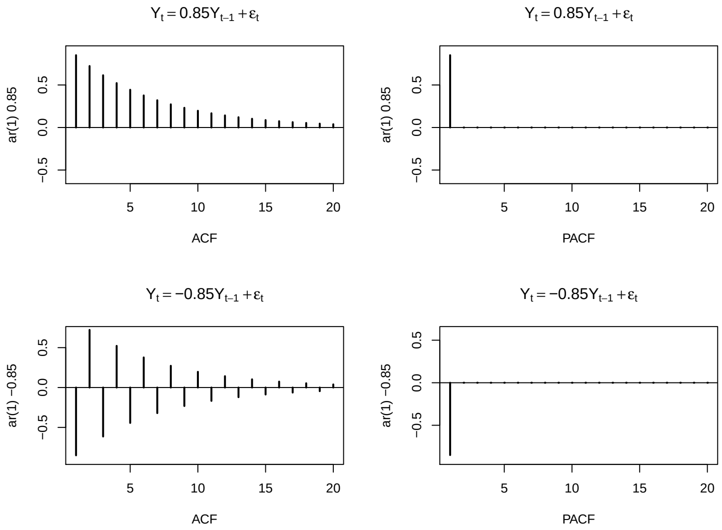


Figure 2: Theoretical acf and pacf of AR(1) series

The AR(2) time series

Let's consider the process $(Y_t) = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \varepsilon_t$ with $\varepsilon_t \sim wn(0, \sigma_\varepsilon^2)$.
To (Y_t) be causal the roots of $\phi(z) = 1 - \phi_1 z - \phi_2 z^2$ must satisfy $|z| > 1$. It can be shown that this is equivalent to:

$$\phi_1 + \phi_2 < 1, \quad \phi_2 - \phi_1 < 1, \quad \text{and} \quad |\phi_2| < 1.$$

See figure 3. For the causality to hold we need (ϕ_1, ϕ_2) to be inside the triangle. The parabolic curve is the limit between reals roots (upper part) and complex roots (under the curve).
The weigths ψ_j of the causal representation can be obtained by solving:

$$(1 - \phi_1 z - \phi_2 z^2)(\psi_0 + \psi_1 z + \psi_2 z^2 + \cdots) = 1$$

We get $\psi_0 = 1$, $\psi_1 - \phi_1\psi_0 = 0$, $\psi_2 - \phi_1\psi_1 - \phi_2\psi_0 = 0$, $\psi_3 - \phi_1\psi_2 - \phi_2\psi_1 - \phi_3 = 0$ and so on.
It follows

$$\psi_1 = \phi_1 \quad \psi_2 = \phi_1^2 + \phi_2.$$

And for $j \geq 2$ we get in fact

$$\psi_j = \phi_1 \psi_{j-1} + \phi_2 \psi_{j-2}. \tag{2}$$

The general solution of this equation can be expressed as exponential functions of the inverses of the roots of the autoregressive polynomial.

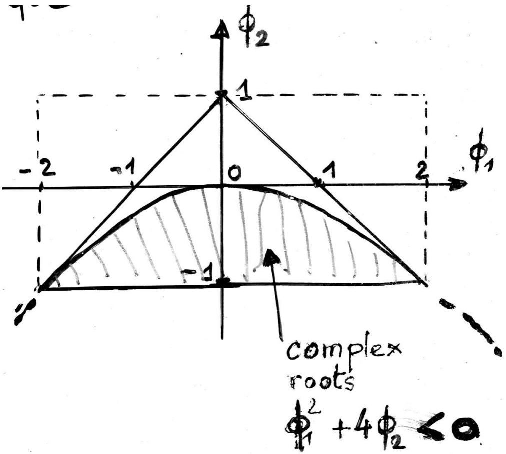


Figure 3: Causality of AR(2) series

Note that, given the numeric values of the autoregressive coefficients, R provides the values of ψ_j. As an example, for $Y_t = 1/6Y_{t-1} - 1/6Y_{t-2} + \varepsilon_t$, using ARMAtoMA(stats) we get the ten first ψ_j along :

```
........................................................................
ARMAtoMA(ar=c(1/6,-1/6),lag.max=10)
 plot(ARMAtoMA(ar=c(1/6,-1/6),lag.max=40) # to get a graph
[1]  0.16666667 -0.57222222 -0.19537037  0.31077160  0.16901749
[6] -0.15829338 -0.12779272  0.07367724  0.08895517 -0.02938048
........................................................................
```

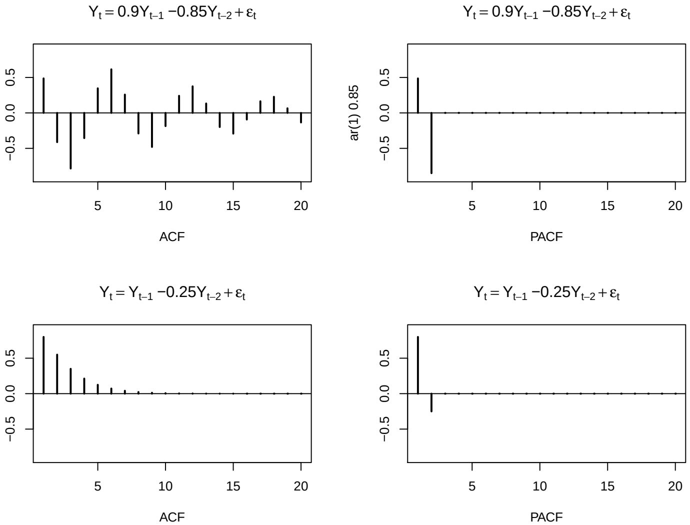
To get the autocovariance we write

$$\begin{aligned}\gamma(h) = \mathbb{E}(Y_t Y_{t+h}) &= \mathbb{E}[Y_t(\phi_1 Y_{t+h-1} + \phi_2 Y_{t+h-2} + \varepsilon_t)] \\ &= \phi_1 \gamma(h-1) + \phi_2 \gamma(h-2)\end{aligned}$$

which leads to:

$$\rho(h) = \phi_1 \rho(h-1) + \phi_2 \rho(h-2) \tag{3}$$

For $h = 1$ and $h = 2$, this gives $\rho(1) = \phi_1 + \phi_2 \rho(1)$ and $\rho(2) = \phi_1 \rho(1) + \phi_2$. The equation (3) is the same as (2) and therefore the solutions can be written as exponential functions of the inverses of the roots of the autoregressive polynomial.
The partial autocorrelation satisfies $\phi_{11} = \rho(1) = \phi_1/(1-\phi_2)$,
$\phi_{22} = (\rho(2) - \rho(1)^2)/(1 - \rho(1)^2)$ and $\phi_{hh} = 0$ for $h > 2$.
Figures 4 and 5 show the ACF and PACF of AR(2) series for different cases of autoregressive polynomials.


Yt = 0.9Yt−1 −0.85Yt−2 + εt
Yt = 0.9Yt−1 −0.85Yt−2 + εt
ACF
PACF
ar(1) 0.85
Yt = Yt−1 −0.25Yt−2 + εt
Yt = Yt−1 −0.25Yt−2 + εt
ACF
PACF

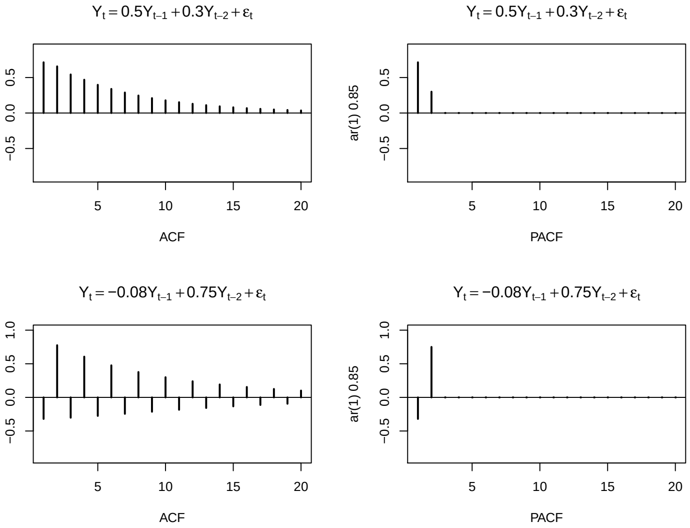


Figure 5: ACF and PACF of AR(2) series. Top: two real positive roots. Bottom two real roots with

ACF and PACF of a general AR(p) time series
The qualitative behavior of ACF and PACF we seen for AR(1) and AR(2) holds true in the case of general AR(p) models.
The equation (3) is generalized to

$$\rho(h) = \phi_1 \rho(h-1) + \phi_2 \rho(h-2) + \cdots + \phi_p \rho(h-p) \tag{4}$$

From the results of difference equations it follows that $\rho(h)$ tails off as a mixture of exponential decays (corresponding to real roots of the autoregressive polynomial) and/or damped sine wave corresponding to real roots.

Concerning the PACF, the main result is that it cuts off after p, $\phi_{hh} = 0$ for $h > p$.
This result is quite useful in the model identification step.

Moving average models

Definition

A time series (Y_t) is a MA(q) series when (Y_t) satisfies:

$$Y_t = \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \cdots + \theta_q\varepsilon_{t-q} \quad (5)$$

where $\theta_q \neq 0$ and $\varepsilon_t \sim wn(0, \sigma_\varepsilon^2)$.

- The operator $\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q$ is the **moving average operator**. We can write $Y_t = \theta(B)\varepsilon_t$. The polynomial $\theta(z)$ is the **moving average polynomial** of the complex variable z.
- A moving average time series is trivially causal ($\psi_0 = 1$, $\psi_j = \theta_j$ for $1 \leq j \leq q$ and $\psi_j = 0$ for $j > q$) and consequently stationary for any choice of the parameters $\theta_1, \theta_2, \ldots, \theta_q$.
- A MA(q) time series is invertible when the roots of $\theta(z) = 0$ lie outside of the unit circle. For uniqueness and also issues related to forecasting we assume henceforth that the MA(q) series we deal with are invertible. The invertible representation $\varepsilon_t = \pi(B)Y_t$, $\pi(B) = 1 - \pi_1 B - \pi_2 B^2 - \cdots$, will follow from $\varepsilon_t = \pi(B)Y_t = \pi(B)\theta(B)\varepsilon_t$ which leads to $\pi(B)\theta(B) = 1$. The π_i will be obtained by identification as seen for AR series.

The MA(1) time series

Consider $Y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1}$ with $|\theta_1| < 1$. (Y_t) is a zero mean process and we easily get:

$$\gamma(h) = \begin{cases} (1+\theta_1^2)\sigma_\varepsilon^2 & h = 0 \\ \theta_1 \sigma_\varepsilon^2 & h = 1 \\ 0 & h > 1 \end{cases}$$

from which it comes:

$$\rho(h) = \begin{cases} \theta_1/(1+\theta_1^2) & h = 1 \\ 0 & h > 1 \end{cases}$$

We note that $|\rho(1)| < 0.5$.

The partial autocorrelation is not so easy to get. It can be shown that:

$$\phi_{hh} = \frac{(-1)^{h+1}\theta_1^h}{1-\theta_1^{2(h+1)}} \quad \text{for} \quad h \geq 1$$

Hence $|\phi_{hh}|$ is exponentially decreasing, alternating in sign when $\theta_1 > 0$, being always negative when $\theta_1 < 0$, while $\rho(h)$ cuts off after $h = 1$.

Note also that, for any h, $|\phi_{hh}| < 0.5$.

Finally it is worthwhile to note that the behavior of the MA(1) is the dual of the AR(1): the acf cuts off after $h = 1$ and the pacf tails off. It is the reverse of AR(1). The figure 6 shows the ACF and PACF for $\theta_1 = 0.85$ and $\theta_1 = -0.85$

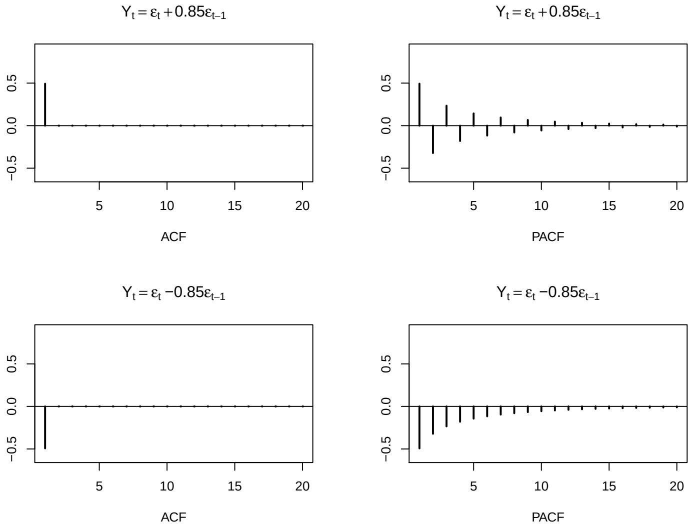


Figure 6: ACF and PACF of two MA(1) series. Top $\theta = 0.85$. Bottom $\theta = -0.85$

The MA(2) time series

Consider $Y_t = \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2}$. for invertibility the roots of the polynomial $\pi(z) = 1 + \theta_1 z + \theta_2 z^2$ must lie outside the unit circle. This leads to the conditions:

$$\theta_1 + \theta_2 > -1, \quad \theta_1 - \theta_2 < 1, \quad \text{and} \quad |\theta_2| < 1.$$

This looks like the conditions needed for the AR(2) series to be causal, but the triangle is mirrored about the axis $\phi_1 = 0$.

The autocovariance function is:

$$\gamma(h) = \begin{cases} (1+\theta_1^2+\theta_2^2)\sigma_\varepsilon^2 & h = 0 \\ \theta_1(1+\theta_2)\sigma_\varepsilon^2 & h = 1 \\ \theta_2\sigma^2 & h = 2 \end{cases}$$

and $\gamma(h) = 0$ for $h > 2$. It follows:

$$\rho(h) = \begin{cases} \frac{\theta_1(1+\theta_2)}{1+\theta_1^2+\theta_2^2} & h = 1 \\ \frac{\theta_2}{1+\theta_1^2+\theta_2^2} & h = 2 \\ 0 & h > 2 \end{cases}$$

The first partial autocorrelations are given by:

$$\begin{aligned}\phi_{11} &= \rho(1)\\ \phi_{22} &= \frac{\rho(2)-\rho(1)^2}{1-\rho(1)^2}\\ \phi_{33} &= \frac{\rho(1)^3-\rho(1)\rho(2)(2-\rho(2))}{1-\rho(2)^2-2\rho(1)^2(1-\rho(2))}\end{aligned}$$

The PACF will decrease at an exponential rate when the roots of the moving average polynomial are real, and will be damped sine wave when the roots are complex.

The general MA(q) process

For the general MA(q) we get the variance:

$$\gamma(0) = \sigma_\varepsilon^2 \sum_{j=0}^{q} \theta_j^2$$

where $\theta_0 = 1$, and the following autocovariances are given by

$$\gamma(h) = \begin{cases} \sigma_\varepsilon^2(\theta_h + \theta_1\theta_{h+1} + \cdots + \theta_{q-h}\theta_q), & h = 1, 2, \ldots, q \\ 0 & h > q \end{cases}$$

This leads to the autocorrelation function:

$$\rho(h) = \begin{cases} \frac{\theta_h + \theta_1\theta_{h+1} + \cdots + \theta_{q-h}\theta_q}{1+\theta_1^2+\cdots+\theta_q^2}, & h = 1, 2, \ldots, q \\ 0 & h > q \end{cases}$$

Thus the autocorrelation functions cuts off after lag q, which generalizes what was observed for MA(1) and MA(2).
Furthermore, depending on the roots of $\theta(z) = 1 + \theta_1 z + \theta_2 z^2 + \cdots + \theta_q z^q = 0$ the partial autocorrelation will tail off as a mixture of exponential decreasing functions (related to real roots) and/or damped sine waves associated with the complex roots.

The ARMA(p,q) process

Definition

A mean zero time series (Y_t) is an **ARMA(p,q)** series when (Y_t) is stationary and satisfy:

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q} \qquad (6)$$

where $\phi_p \neq 0$, $\theta_q \neq 0$ and $\varepsilon_t \sim wn(0, \sigma_\varepsilon^2)$.
When $\mathbb{E}(Y_t) = \mu \neq 0$, (Y_t) (6) is replaced by

$$\begin{aligned} Y_t - \mu = \phi_1(Y_{t-1} - \mu) &+ \phi_2(Y_{t-2} - \mu) + \cdots + \phi_p(Y_{t-p} - \mu) + \varepsilon_t + \theta_1 \varepsilon_{t-1} \\ &+ \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q} \end{aligned}$$

that is

$$Y_t = \alpha + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}$$

with $\alpha = \mu(1 - \phi_1 - \phi_2 - \cdots - \phi_p)$.

Remarks

1. (6) is written
$$\phi(B)Y_t = \theta(B)\varepsilon_t$$
where $\phi(B)$ and $\theta(B)$ are the autoregressive and moving average operators:
$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p \quad (7)$$
$$\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q \quad (8)$$
2. Henceforth an ARMA(p,q) series is always supposed to be causal and invertible. These conditions are satisfied iff the roots of $\theta(z) = 0$ are outside of the unit circle (invertibility), and the same for $\phi(z) = 0$ (causality). We assume also that $\phi(z) = 0$ and $\theta(z) = 0$ have no common roots.
3. p and q are called the orders of the autoregressive and moving average models respectively.

A consequence of remark 2 just above is that we can get the causal representation of an ARMA(p,q) (Y_t), $Y_t = \sum_{j\geq 0} \psi_j \varepsilon_j$, $\psi_0 = 1$ by

$$\sum_{j\geq 0} \psi_j z^j = \frac{\theta(z)}{\phi(z)}, \quad |z| \leq 1. \quad (9)$$

In the same way the autoregressive representation follows from

$$\sum_{j\geq 0} \pi_j z^j = \frac{\phi(z)}{\theta(z)}, \quad |z| \leq 1. \tag{10}$$

Note that the ψ_i terms as well as the π_i decrease exponentially.

The ACF of the ARMA(p,q) time series

$$\begin{aligned}
\gamma(h) &= \mathbb{C}\text{ov}\big(Y_t, Y_{t+h}\big) \\
&= \mathbb{C}\text{ov}\big(Y_t, \sum_{j=1}^{p} \phi_j Y_{t+h-j} + \sum_{j=0}^{q} \theta_j \varepsilon_{t+h-j}\big) \\
&= \sum_{j=1}^{p} \phi_j \gamma(h-j) + \mathbb{C}\text{ov}\big(Y_t, \sum_{j=0}^{q} \theta_j \varepsilon_{t+h-j}\big) \\
&= \sum_{j=1}^{p} \phi_j \gamma(h-j) + \sum_{j=0}^{q-h} \theta_j \psi_{h-j}
\end{aligned}$$

where we used the causal representation of Y_t.
It comes:

$$\gamma(h) = \phi_1 \gamma(h-1) + \phi_2 \gamma(h-2) + \cdots + \phi_p \gamma(h-p) \quad h \geq \max(p, q+1) \tag{11}$$

and the initial conditions are given by:

$$\gamma(h) = \sum_{j=1}^{p} \phi_j \gamma(h-j) + \sigma_\varepsilon^2 \sum_{j=0}^{q-h} \theta_j \psi_{h-j} \quad h < \max(p, q+1) \tag{12}$$

Dividing by $\gamma(0)$ yields similar equations for $\rho(h)$.

It is worthwhile to note:

1. The equation (11) is identical to the one for AR(p) models (4). That is the ACF of an ARMA model depend only on the AR part of the ARMA. Hence, as in AR models, tha ACF tails off as a mixture of exponential decays and/or damped sine waves. Note that the q first autocorrelations depend on both autoregressive and moving average parameters.
2. It is not so easy to give formulas for the partial autocorrelations. But the behavior will also be a mixture of exponential decays and/or damped sine waves depending on the roots of boths the polynomials $\phi(z)$ and $\theta(z)$.

The ARMA(1,1) model

The solutions of (9) and (10) are:

$$\pi_j = (-\theta_1)^{j-1}(\phi_1 + \theta_1), \quad j \geq 1, \quad \text{and} \quad \psi_j = \phi_1^{j-1}(\phi_1 + \theta_1), \quad j \geq 1 \qquad (13)$$

From (11) it comes

$$\gamma(h) = \phi_1 \gamma(h-1), \quad h = 2, 3, \ldots$$

Writing (12) for $h = 0$ and $h = 1$ and using (13) gives:

$$\gamma(0) = \phi_1 \gamma(1) + \sigma_\varepsilon^2(1 + \theta_1\phi_1 + \theta_1^2) \quad \text{and} \quad \gamma(1) = \phi_1\gamma(0) + \sigma_\varepsilon^2 \theta_1$$

This yields

$$\gamma(0) = \sigma_\varepsilon^2 \frac{(1 + 2\theta_1\phi_1 + \theta_1^2)}{1 - \phi_1^2} \quad \text{and} \quad \gamma(1) = \sigma_\varepsilon^2 \frac{(1 + \theta_1\phi_1)(\theta_1 + \phi_1)}{1 - \phi_1^2}$$

Finally the solution for $\gamma(h)$ is

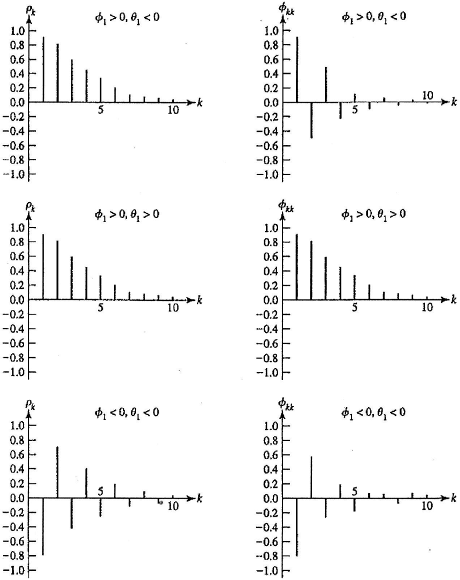
$$\gamma(h) = \sigma_\varepsilon^2 \frac{(1 + \theta_1\phi_1)(\theta_1 + \phi_1)}{1 - \phi_1^2} \phi_1^{h-1}, \quad h \geq 1$$

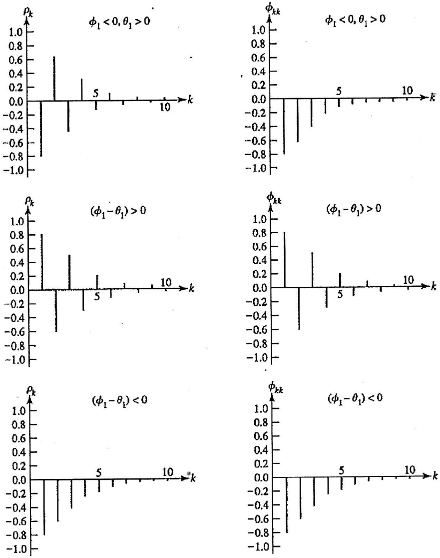
Dividing by $\gamma(0)$ gives the ACF:

$$\rho(h) = \frac{(1+\theta_1\phi_1)(\theta_1+\phi_1)}{1+2\theta_1\phi_1+\theta_1^2}\phi_1^{h-1}, \quad h \geq 1$$

Hence after the initial lags, as it could be anticipated, the ACF of an ARMA(1,1) is closed to the one of his AR(1) part but differ by the multiplicative constant and the lag in the exponential term..
Like the ACF, the PACF of an ARMA(1,1) tails off exponentially with a shape related to the signs and magnitudes of the parameters ϕ_1 and θ_1.

We show in figures (7) and (8) for different values of (ϕ_1, θ_1) the shape of ACF and PACF. It appears that often, examining both ACF and PACF allows to preclude AR or MA models.


ρ_k
$\phi_1 > 0, \theta_1 < 0$
ϕ_{kk}
$\phi_1 > 0, \theta_1 < 0$
ρ_k
$\phi_1 > 0, \theta_1 > 0$
ϕ_{kk}
$\phi_1 > 0, \theta_1 > 0$
ρ_k
$\phi_1 < 0, \theta_1 < 0$
ϕ_{kk}
$\phi_1 < 0, \theta_1 < 0$
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
−1.0
5
10
k


ρ_k
$\phi_1 < 0, \theta_1 > 0$
ϕ_{kk}
$\phi_1 < 0, \theta_1 > 0$
ρ_k
$(\phi_1 - \theta_1) > 0$
ϕ_{kk}
$(\phi_1 - \theta_1) > 0$
ρ_k
$(\phi_1 - \theta_1) < 0$
ϕ_{kk}
$(\phi_1 - \theta_1) < 0$
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
−1.0
5
10
k

ARIMA series

Definition

A time series (Y_t) is an ARIMA(p,d,q) series when $\nabla^d Y_t = (1-B)^d Y_t$ is stationary and is an ARMA(p,q) model, that is

$$\phi(B)\nabla^d Y_t = \theta(B)\varepsilon_t \tag{14}$$

where:

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p \tag{15}$$
$$\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q \tag{16}$$

The roots of the polynomials $\phi(z)$ and $\theta(z)$ are supposed to be outside the unit circle and $\phi(z)$ and $\theta(z)$ are supposed to have no common root.
(Y_t) is said to be an Autoregressive Integrated Moving Average time series.

To take into account that $\nabla^d Y_t$ could have a non zero mean (14) is given a more general form by adding a constant term θ_0:

$$\phi(B)\nabla^d Y_t = \theta_0 + \theta(B)\varepsilon_t \tag{17}$$

Examples

- ARIMA(0,1,0) and ARIMA(0,2,0)
  - **Random walk**

$$\begin{aligned} \nabla Y_t &= \varepsilon_t \\ Y_t - Y_{t-1} &= \varepsilon_t \end{aligned}$$

That is (Y_t) is a random walk $Y_t = Y_{t-1} + \varepsilon_t = Y_1 + \sum_{i=0}^{t-2} \varepsilon_{t-i},\ t > 1$.

  - **Random walk with drift**

$$\nabla Y_t = \theta_0 + \varepsilon_t$$

That is (Y_t) is a random walk with drift
$Y_t = Y_{t-1} + \theta_0 + \varepsilon_t = Y_1 + (t-1)\theta_0 + \sum_{i=0}^{t-2} \varepsilon_{t-i},\ t > 1$.

  - **Linear trend**

$$\nabla^2 Y_t = \varepsilon_t$$

That is $Y_t - 2Y_{t-1} + Y_{t-2} = \varepsilon_t$ which can be seen as a nonhomogeneous order 2 difference equation. The solution can be written:

$$Y_t = A_0 + A_1 t + \tilde{\varepsilon}_t$$

where A_0 and A_1 are random and depending on Y_1 and Y_2 and $\tilde{\varepsilon}_t$ is a particular solution of the complete equation.

- ARIMA(0,1,1), ARIMA(0,2,2), ARIMA(1,1,1)
  - ARIMA(0,1,1)
    $$\nabla Y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1}$$
    That is $Y_t = Y_{t-1} + \varepsilon_t + \theta_1 \varepsilon_{t-1}$
  - ARIMA(0,2,2)
    $$\nabla^2 Y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2}.$$
    That is $Y_t = 2Y_{t-1} - Y_{t-2} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2}$.
  - ARIMA(1,1,1)
    $$(1 - \phi_1 B)\nabla Y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1}$$
    That is $\nabla Y_t - \phi_1 \nabla Y_{t-1} + \varepsilon_t + \theta_1 \varepsilon_{t-1}$, and finally:
    $$Y_t = (1 + \phi_1) Y_{t-1} - \phi_1 Y_{t-2} + \varepsilon_t + \theta_1 \varepsilon_{t-1}$$

A main interest of ARIMA(p,d,q) is that they can model processes with a random polynomial with degree $d-1$ trend. The operator ∇^d filters the degree $d-1$ polynomials.
Let :

$$Y_t = A_0 + A_1 t + \cdots + A_{d-1} t^{d-1} + Z_t$$

with $A_0, A_1, \ldots, A_{d-1}$ random variables. Then

$$\nabla^d Y_t = \nabla^d Z_t$$

Estimating trend and seasonal components by moving average

Consider :

$$Y_t = Z_t + S_t + \varepsilon_t$$

where Z_t is a trend, S_t a seasonal component and ε_t a white noise with variance σ_ε^2.
S_t satisfy $S_{t+p_0} = S_t$, that is we have p_0 seasonality parameters to estimate :$S_1, \ldots, S_{p_0}$.
A practical method to estimate both the trend and the seasonality components is to use an moving average operator:

$$MY_t = \theta_{-m_1} Y_{t-m_1} + \theta_{-m_1+1} Y_{t-m_1+1} + \cdots + \theta_{-1} Y_{-1} + \theta_0 Y_0 + \theta_1 Y_1 + \cdots + \theta_{m_2} Y_{m_2},$$

M is denoted $M = (\theta_{-m_1}, \theta_{-m_1+1}, \cdots, \theta_{-1}, \theta_0, \theta_1, \cdots, \theta_{m_2})$ and $m_1 + m_2 + 1$ is the order of M.
M is chosen in such a way that

- Concerned seasonality is eliminated.
- Some adequate series are invariant: for instance polynomials of degree d.
- $\mathbb{V}\mathrm{ar}(\varepsilon_t^*) = \mathbb{V}\mathrm{ar}(M\varepsilon_t)$ is reduced as much as possible.

Note that $\mathbb{V}\mathrm{ar}(\varepsilon_t^*) = \sigma_\varepsilon^2 \sum_{i=-m_1}^{m_2} \theta_i^2$. Hence we have to choose M in such a way that the ratio $\mathbb{V}\mathrm{ar}(\varepsilon_t^*)/\sigma_\varepsilon^2 = \sum_{i=-m_1}^{m_2} \theta_i^2$ to be small as possible.
Further, to leave invariant a degree d polynomial we have to satisfy:

$$\sum \theta_i = 1, \sum i\theta_i = 0, \ldots, \sum i^d \theta_i = 0$$

For instance, consider the Spencer moving average with 15 terms. It is a symetric m.a. (i.e. $m_1 = m_2 = m$, and $\theta_{-i} = \theta_i,\ i = 1, \dots, m$) and $(\theta_{-7}, \dots, \theta_0)$ are given by

$$(\theta_{-7}, \dots, \theta_0) = \frac{1}{320}(-3, -6, -5, 3, 21, 46, 67, 74)$$

This m.a. leaves invariant degree 3 polynomials and delete the period 4 seasonality, i.e. the functions $\cos(2\pi \frac{j}{4} t)$ and $\sin(2\pi \frac{j}{4} t)$ for $j = 1, 2$.

Exercise. Check these properties by simulation (Hint: generate a polynomial and apply filter. Do the same for the periodic functions)

The principle of the estimation method, in its simplest form, consists in:

$$\begin{aligned} Y_t^* &= MY_t \\ &= Z_t^* + S_t^* + \varepsilon_t^* \\ &\approx Z_t + \varepsilon_t^* \end{aligned}$$

Then we define:

$$\widehat{S}_t = Y_t - Y_t^*.$$

For the season k, $k = 1, \ldots, p_0$; we calculate:

$$\widehat{S}_k = \frac{1}{n_k} \sum_{i=0}^{n_k-1} \widehat{S}_{k+ip_0}$$

Finally we correct each term to force their sum to be 0:

$$\widehat{\widehat{S}}_k = \widehat{S}_k - \frac{1}{p_0} \sum_{k=1}^{p_0} \widehat{S}_k$$

Let $k(t)$ the season of t, $Y_t - \widehat{\widehat{S}}_{k(t)}$ is referred to as the *seasonally adjusted series* (In french: serie CVS, corrigée des variations saisonnières).
Note that often the procedure is iterated, possibly with different m.a.

Seasonal ARIMA models

Methods presented in the Ch0 (linear regression) and in the foregoing paragraph assume that the seasonal component satisfy:

- The seasonal component is deterministic
- The seasonal component is independent of the other non-seasonal components (for instance a deterministic trend).

In practical situation the seasonal component may be stochastic and correlated with non-seasonal components.

Assume we have monthly data Y_t, where Y_t is related to Y_{t-1} but also to Y_{t-12}.That is we have a relationship month by month, i.e. inside a period, but also year by year, i.e. between periods.

If we don't know this relation year to year, we fit a non-seasonal ARIMA model:

$$\phi_p(B)(1-B)^d Y_t = \theta_q(B)u_t$$

Then u_t will not be a white noise, with a non zero aucorrelation for lags multiple of $s = 12$

$$\rho(js) = \frac{\mathbb{E}\left[(u_t - \mu_u)(u_{t+js} - \mu_u)\right]}{\sigma_u^2}$$

where the dynamic of u_t could be modelled by :

$$\Phi_P(B^s)(1-B^s)^D u_t = \Theta_Q(B^s)\varepsilon_t$$

with

$$\Phi_P(B) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \cdots - \Phi_p B^{Qs} \quad (18)$$

$$\Theta_Q(B) = 1 + \Theta_1 B^s + \Theta_2 B^{2s} + \cdots + \Theta_q B^{Qs} \quad (19)$$

This yields a general expression for the ARIMA seasonal model.

Definition

A time series (Y_t) is an ARIMA$(p,d,q) \times (P,D,Q)_s$ series when we can write

$$\Phi_P(B^s)\phi_p(B)(1-B)^d(1-B^s)^D Y_t = \theta_q(B)\Theta_Q(B^s)\varepsilon_t \quad (20)$$

where the polynomials Φ_P and Θ_Q, respectively ϕ_p and θ_q were defined previously and have roots which satisfy the usual constraints (no common roots and roots outside of the unit circle).

Exercise

- Given $(1-\Phi_1 B^{12})Y_t = \varepsilon_t$, that is an ARIMA$(0,0,0) \times (1,0,0)_{12}$ which can be seen also as particular AR(12), calculate the ACF of (Y_t).
- Do the same for the model ARIMA$(0,0,0) \times (0,0,1)_{12}$, that is $Y_t = (1+\Theta_1 B^1 2)\varepsilon$.

Example

Let Y_t following the model ARIMA$(0,1,1) \times (0,1,1)_{12}$ and set $Z_t = (1-B)(1-B^{12})Y_t$, that is

$$Z_t = (1+\theta_1 B)(1+\Theta_1 B^{12})\varepsilon_t$$

The ACF of Z_t satisfy:

$$\begin{aligned}
\rho_Z(1) &= \frac{\theta_1}{1+\theta_1^2} \\
\rho_Z(11) &= \frac{\theta_1\Theta_1}{(1+\theta_1^2)(1+\Theta_1^2)} \\
\rho_Z(12) &= \frac{\Theta_1}{(1+\Theta_1^2)} \\
\rho_Z(13) &= \rho_Z(11)
\end{aligned}$$

and $\rho(j) = 0$ for any other j. We have also $\gamma_Z(0) = (1+\theta_1^2)(1+\Theta_1^2)\sigma_\varepsilon^2$.

Forecasting

Several methods are available to forecast Y_{t+m}^t of Y_{t+m} when we observe Y_t and its past.

A) **Method by linear projection on** $\overline{sp}\{Y_s,\, 1 \leq s \leq t\}$

In ch1.4 we give a theoretical answer for $m = 1$ when we observe $Y_1, \dots, Y_t$:

$$Y_{t+1}^t = \phi_{t1} Y_t + \phi_{t2} Y_{t-1} + \cdots + \phi_{tt} Y_1$$

where $\boldsymbol{\phi}_t = (\phi_{t1}, \phi_{t2}, \cdots, \phi_{tt})'$ is given by

$$\boldsymbol{\phi}_t = \Gamma_t^{-1} \boldsymbol{\gamma}_t, \tag{21}$$

with $\boldsymbol{\gamma}_t = (\gamma(1), \dots, \gamma(t))'$, and $\Gamma_t = \Big(\gamma(i-j)\Big)_{i,j=1,\dots,t}$.

Due to Γ_t^{-1} this method may be time consuming. Even though the Levinson-Durbin method provide a more efficient algorithm, this forecasting method is used only when t is small.

B) **Methods based on the AR(∞) and MA(∞) possibly truncated representations.**

Let (Y_t) be a mean zero stationary time series causal and invertible, and $\varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2)$.

Linear representations $Y_t = \sum_{i=0}^{\infty} \psi_i \varepsilon_{t-i}$ and $Y_t = \sum_{i=1}^{\infty} \pi_i Y_{t-i} + \varepsilon_t$ are the basis for defining the forecast.

We can write:

$$\begin{aligned} Y_{t+m} &= \sum_{j=0}^{\infty} \psi_j \varepsilon_{t+m-j} \\ &= \sum_{j=0}^{m-1} \psi_j \varepsilon_{t+m-j} + \sum_{j=m}^{\infty} \psi_j \varepsilon_{t+m-j} \end{aligned} \tag{22}$$

Then the conditional expectation of Y_{t+m} given $(Y_{t-i}, i = 0, \ldots, \infty)$ is given by:

$$Y_{t+m}^t = \sum_{j=m}^{\infty} \psi_j \varepsilon_{t+m-j} = \sum_{j=0}^{\infty} \psi_{m+j} \varepsilon_{t-j} \tag{23}$$

Using the AR(∞) representation we get:

$$\begin{aligned} Y_{t+m}^t &= \sum_{i=1}^{m-1} \pi_i Y_{t+m-i}^t + \sum_{i=m}^{\infty} \pi_i Y_{t+m-i} \\ &= \pi_1 Y_{t+m-1}^t + \cdots + \pi_{m-1} Y_{t+1}^t + \sum_{i=m}^{\infty} \pi_i Y_{t+m-i} \end{aligned} \tag{24}$$

(24) allows to compute iteratively $Y_{t+1}^t, Y_{t+2}^t, \ldots, Y_{t+m}^t$ along

$$
\begin{aligned}
Y_{t+1}^{t} &= \sum_{i=0}^{\infty} \pi_{1+i} Y_{t-i} \\
Y_{t+2}^{t} &= \pi_1 Y_{t+1}^{t} + \sum_{i=0}^{\infty} \pi_{2+i} Y_{t-i} \\
Y_{t+3}^{t} &= \pi_1 Y_{t+2}^{t} + \pi_2 Y_{t+1}^{t} + \sum_{i=0}^{\infty} \pi_{3+i} Y_{t-i}
\end{aligned}
$$

and so on.

Clearly

$$\mathbb{E}(Y_{t+m}^{t}) = \mathbb{E}(Y_{t+m}) = 0$$

and from (22) and (23) we get

$$P_{t+m}^{t} = \mathbb{E}\left[(Y_{t+m}^{t} - Y_{t+m})^2\right] = \sigma_\varepsilon^2 \sum_{j=0}^{m-1} \psi_j^2.$$

Note also that the forecast errors $e_m = (Y_{t+m}^{t} - Y_{t+m})$ and $e_{m+k} = (Y_{t+m+k}^{t} - Y_{t+m+k})$ are correlated:

$$\mathbb{C}\mathrm{ov}(e_m, e_{m+k}) = \sigma_\varepsilon^2 \sum_{j=0}^{m-1} \psi_j \psi_{j+k}$$

When $\mathbb{E}(Y_t) = \mu$, Y_t is replaced by $Y_t - \mu$ and we get:

$$Y^t_{t+m} = \mu + \sum_{j=0}^{\infty} \psi_{m+j}\varepsilon_{t-j}$$

It follows that when $m \to \infty$:

$$\begin{aligned} Y^t_{t+m} &\to \mu \quad \text{and} \\ P^t_{t+m} &\to \sigma^2_\varepsilon \sum_{j=0}^{\infty} \psi_j^2 = \gamma(0) = \mathbb{V}\mathrm{ar}(Y_t) \end{aligned}$$

C) **Truncated conditional expectation**
As $Y_0, Y_{-1}, Y_{-2}, \ldots$ are not available we truncate the last left term of (24) to $\sum_{i=m}^{t+m-1} \pi_i Y_{t+m-i}$, and consider the forecast

$$\widetilde{Y}^t_{t+m} = \sum_{i=1}^{m-1} \pi_i \widetilde{Y}^t_{t+m-i} + \sum_{i=m}^{t+m-1} \pi_i Y_{t+m-i}$$

and we proceed iteratively as above.
The case of ARMA models
First recall that for an AR(p) model $Y_{t+1} = \phi_1 Y_t + \cdots + \phi_p Y_{t-p} + \varepsilon_t$, the predictor $\widetilde{Y}^t_{t+m} = Y^t_{t+m} = \phi_1 Y_{t+m-1} + \cdots + \phi_p Y_{t+m-p}$ is exact, and we can proceed iteratively:

$$\begin{aligned} Y^t_{t+1} &= \phi_1 Y_t + \cdots + \phi_p Y_{t-p} \\ Y^t_{t+2} &= \phi_1 Y^t_{t+1} + \phi_2 Y_t + \cdots + \phi_p Y_{t-p+2} \end{aligned}$$

and so on.

Finally, for a general ARMA(p,q) series, the truncated predictors for $m = 1, 2, \ldots$ takes the particular form:

$$\widetilde{Y}^t_{t+m} = \phi_1 \widetilde{Y}^t_{t+m-1} + \cdots + \phi_p \widetilde{Y}^t_{t+m-p} + \theta_1 \widetilde{\varepsilon}^t_{t+m-1} + \cdots + \theta_q \widetilde{\varepsilon}^t_{t+m-q}$$

where $\widetilde{Y}^t_s = Y_s$ for $1 \leq s \leq t$, and $\widetilde{Y}^t_s = 0$ for $s \leq 0$. The truncated prediction errors are given by $\widetilde{\varepsilon}^t_s = 0$ for $s \leq 0$ or $s \geq t$, and

$$\widetilde{\varepsilon}^t_s = \phi(B) \widetilde{Y}^t_s - \theta_1 \widetilde{\varepsilon}^t_{s-1} - \cdots - \theta_q \widetilde{\varepsilon}^t_{s-q}$$

for $1 \leq s \leq t$.

Estimation

Several methods are used to estimate the parameters of ARMA(p,q) models.

A) **Methods of moments.**
The principle is to define estimators by identifying empirical moments with their expressions in terms of the parameters.
In the particular case of AR(p) models, the method of moments through the Yule-Walker equations has proved to be efficient. When $Y_t \sim$AR(p), the YW equations are the difference equations we already encountered:

$$\begin{aligned} \gamma(h) &= \phi_1\gamma(h-1) + \cdots + \phi_p\gamma(h-p), \quad h = 1, 2, \ldots, p \\ \gamma(0) &= \phi_1\gamma(1) + \cdots + \phi_p\gamma(p) + \sigma_\varepsilon^2 \end{aligned}$$

In matrix notation these equations can be written $\phi = \Gamma_p^{-1}\gamma_p = R_p^{-1}\rho_p$ with notations introduced before (see (21)).
This leads to $\widehat{\phi} = \widehat{R}_p^{-1}\widehat{\rho}_p$ and $\widehat{\sigma}_\varepsilon^2 = \widehat{\gamma}(0)[1 - \widehat{\rho}'_p\widehat{R}_p^{-1}\widehat{\rho}_p]$
Then $\widehat{\rho}_p = (\widehat{\rho}(1), \ldots, \widehat{\rho}(p))$ is provided by the empirical ACF, and yields $\widehat{R}_p = (\widehat{\rho}(i-j))$.
The variance $\gamma(0)$ is estimated by the empirical estimate of $\mathbb{V}\mathrm{ar}(Y_t)$. Finally we get $\widehat{\phi}$ and $\widehat{\sigma}_\varepsilon^2$ by plugging these empirical estimates in the formulas just above.

This method called Yule-Walker methods is efficient for AR(p) models . It's not the same for MA(q) or general ARMA(p,q).

B) **Likelihood and least squares methods**

For general gaussian ARMA(p,q) models, methods are based on the likelihood function, unconditional or conditional on the initial observations. Least squares methods are also proposed which used a part of the expression of the likelihood.

We assume that Y_t ~ARMA(p,q) and that we observe $y_1, y_2, \ldots, y_n$.

We note $\beta = (\mu, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q)'$. Thus the likelihood can be written:

$$L(\beta, \sigma_\varepsilon^2) = \prod_{t=1}^{n} f(y_t | \, y_{t-1}, \ldots, y_1)$$

and recall that the distribution of Y_t conditional on $(Y_{t-1}, \ldots, Y_1)$ is $\mathcal{N}(Y_{t-1}^t, P_{t-1}^t)$.

We know that $\gamma(0) = \sigma_\varepsilon^2 \sum_{j=0}^{\infty} \psi_j^2$ and $P_t^{t-1} = \sigma_\varepsilon^2 \Big(\sum_{j=0}^{\infty} \psi_j^2 \Big) \Big[\prod_{j=1}^{t-1} (1 - \phi_{jj}^2) \Big] \equiv \sigma_\varepsilon^2 r_t$.

The likelihood is given by:

$$L(\beta, \sigma_\varepsilon^2) = (2\pi\sigma_\varepsilon^2)^{-n/2} \Big[\prod_{t=1}^{n} r_t(\beta) \Big]^{-1/2} exp \Big[- \frac{S(\beta)}{2\sigma_\varepsilon^2} \Big] \qquad (25)$$

where

$$S(\beta) = \sum_{t=1}^{n} \frac{(y_t - y_t^{t-1})^2}{r_t(\beta)} \qquad (26)$$

The estimates $\widehat{\beta}$ and $\widehat{\sigma}_\varepsilon^2$ are obtained by maximizing $L(\beta, \sigma_\varepsilon^2)$. As emphasized, S and r_t are only function of β. It follows:

$$\widehat{\sigma}_\varepsilon^2 = S(\widehat{\beta})/n.$$

Plugging $\sigma_\varepsilon^2 = n^{-1} S(\beta)$ in (25) and taking the log leads to minimize

$$n \log(S(\beta)) + \sum_{t=1}^{n} \log(r_t(\beta)).$$

this is done via iterative procedures.

The *unconditional least squares* method get $\widehat{\beta}$ by minimizing $S(\beta)$ given by (26). Then $\widehat{\sigma}_\varepsilon^2 = S(\widehat{\beta})/(n - (p + q + 1))$
The *conditional least squares* method minimizes also $S(\beta)$ given by (26) but conditionaly on the initial observations.

As set in the following result, under some conditions, the mle, the unconditional least squares and the conditional least squares methods lead to optimal estimators.

Asymptotic optimality for the maximum likelihood, unconditional least squares, conditional least squares estimators

Under some conditions, for invertible and causal time series, when initialized by the method of moments, the maximum likelihood, the unconditional least squares, and the conditional least squares, all result in optimal estimators of σ_ε^2 and β in the sense: $\widehat{\sigma}_\varepsilon^2$ is consistent and the asymptotic normal distribution has the best variance. We have:

$$\sqrt{n}(\widehat{\beta} - \beta) \to^d \mathcal{N}\left(0, \sigma_\varepsilon^2 \Gamma_{p,q}^{-1}\right). \tag{27}$$

The $\Gamma_{p,q}$ matrix is the information matrix whose can be written

$$\Gamma_{p,q} = \begin{pmatrix} \Gamma_{\phi,\phi} & \Gamma_{\phi,\theta} \\ \Gamma_{\theta,\phi} & \Gamma_{\theta,\theta} \end{pmatrix}$$

where $\Gamma_{\phi,\phi} = (\gamma(i-j)), \ldots i, j = 1, \ldots, p$ of an AR(p) process $\phi(B)X_t = \varepsilon_t$, and $\Gamma_{\theta,\theta}$ is the same for an AR(q) process $\theta(B)Y_t = \varepsilon_t$. The matrix $\Gamma_{\phi,\theta} = (\gamma_{X,Y}(i-j))$ where $\gamma_{X,Y}$ is the cross-covariance between the two AR processes $\phi(B)X_t = \varepsilon_t$ and $\theta(B)Y_t = \varepsilon_t$. Finally $\Gamma_{\theta,\phi} = \Gamma'_{\phi,\theta}$.

Building ARIMA models strategy

Given a dataset, experience shows that some steps are to be taken into account to build ARIMA model(s) for this dataset. Note that often several models may be adequatly fitted to the data and the fit could be approximatively the same.
Some authors consider first an identification phase and then a process of diagnostic to evaluate and possibly modify the tentatively selected model.

- **A) Identification**
  A first step is to plot the time series. This allows often to detect trend, seasonality, outliers, heteroscedasticity, non-stationarity...This possibly gives informations as the necessity of apply a transformation and/or differencing the data.

  a) Preliminary transformation.
  Often transformations, such as $\log(x)$, $1/x$, $\sqrt{x}$, x^2 are used to stabilize the variance. It is worthwhile to note that it happens that normality is also improved by these transformations. When differenting seems to be carried out, transformations must be applied *before* differenting.
  The log transformation is often useful when data take on low values as well as some high values: the log-transform will spread low values and stretch high values.

A very often used power-transformation, called Box-Cox procedure, can be applied to find the best power-transformation to make y more homoscedastic, more gaussian or more symmetric. The family of power transformations is given by :

$$y^{(\lambda)} = \begin{cases} \frac{y^{\lambda}-1}{\lambda} & \text{if } \lambda \neq 0, \\ \ln(y) & \text{if } \lambda = 0. \end{cases}$$

In R time series library BoxCox{forecast} provides the $\widehat{\lambda}$ optimal, the transformed series $y_t^{(\widehat{\lambda})}$ and the inverse procedure.
The principle is the following. For a λ value, an AR approximation is fitted and the residuals variance is calculated. This procedure is performed for a grid of λ values, and $\widehat{\lambda}$ is choosen as the λ which minimizes the residuals variance.

b) Identifying d, and p and q.
We compute and plot ACF and PACF to determine whether we need to differentiate. An essential issue is to detect a trend that would imply differentiation. Some cases are clear: when the ACF is decreasing very slowly and PACF cuts off after lag 1 (sometimes 2, more rarely 3), this is typical of a trend dominating the signal. Then we need to differentiate (see example *** below). We get $Y_t^{(1)} = \nabla Y_t = (1-B)Y_t$. The DF (Dickey-Fuller) test or the ADF (Augmented Dickey-Fuller) version can be worked out. The procedure is iterated to see whether $Y_t^{(1)}$ needs also to be differentiate. Note that both under-differentiation and over-differentiation are to be avoided.

A first determination of p and q rests on the behavior of the ACF and PACF several times observed.

- When $(Y_t) \sim$AR(p) then the ACF tails off exponentially or as a damped sinus, and the PACF cuts off after lag p
- When $(Y_t) \sim$MA(q) then the ACF cuts off after lag q and the PACF tails off exponentially or as a damped sinus.
- When $(Y_t) \sim$ARMA(p,q) then the ACF tails off after lag $q-p$ and the PACF tails off after lag $p-q$.

Note that, in practice, due to samples variations these features could be not so clear.

d) Deterministic trend θ_0 when $d > 0$.
In the general model $\phi(B)(1-B)^d Y_t = \theta_0 + \theta(B)\varepsilon_t$, θ_0 is a deterministic trend mean of the differentiated series. θ_0 can be put in the model and its significance tested after the fitting of the model. Another way is to consider $Z_t = (1-B)^d Y_t$ and to test that Z_t is a mean zero process. This can be done by an approximate t-test based on $\overline{Z}_t / S_{\overline{Z}}$. The standard error $S_{\overline{Z}}$ is usually approximated by:

$$S_{\overline{Z}} = \left[\frac{\widehat{\gamma}_Z(0)}{n}(1 + 2\widehat{\rho}_Z(1) + 2\widehat{\rho}_Z(2) + \cdots + 2\widehat{\rho}_Z(k)\right]$$

where $\widehat{\gamma}_Z(0)$ is the sample variance of (Z_t) and $\widehat{\rho}_Z(1), \widehat{\rho}_Z(2), \ldots, \widehat{\rho}_Z(k)$ are the first k significant terms of the sample ACF of (Z_t).

- **B) Checking adequacy of the tentative model**
  Given the data have been possibly transformed and possibly differentiated , and a first choice for p, d an q have been done, we investigate the adequacy of this tentative model. The main tool is to examine the residuals.
  Recall that the residuals/innovations $e_t = y_t - y_t^{t-1}$, which can be seen as estimates of ε_t, when normalized by the standard error $(P_t^{t-1})^{1/2}$ should be approximatly:
  - uncorrelated under the usual assumption that $\varepsilon_t \sim WN(0, \sigma_\varepsilon^2)$
  - independent when $\varepsilon_t \sim IID(0, \sigma_\varepsilon^2)$
  - independent $\mathcal{N}(0,1)$ when $\varepsilon_t \sim IID\mathcal{N}(0, \sigma_\varepsilon^2)$

For checking these properties it is recommanded :
-to plot the residuals to detect visually any departure from these properties
-to check normality visually by drawing histograms and Q-Q plots
-use Box-Pierce and Ljung and Box tests to detect any inapropriate ACF behavior (the $\widehat{\rho}_e(i)$ should be approximatively independent and $\mathcal{N}(0, \sqrt{1/n})$.
- check that the ACF and PACF don't show any evidence for an AR, or MA, or ARMA structure, which could lead to put in question the tentative model and try a new one taking into account the residuals structure.