Autoregressive Moving Average (ARMA) Models and their Practical Applications

Massimo Guidolin

February 2018

1 Essential Concepts in Time Series Analysis

1.1 Time Series and Their Properties

Time series: a sequence of random variables $y_1, y_2, ..., y_T$, or a stochastic process $\{y_t\}_{t=1}^{T}$, of which we only observe the empirical realisations. It is continuous when the observations are recorded continuously over some time interval, discretely sampled when the observations are recorded at equally spaced time intervals.

Observed time series:$\{y_t\}_{t=1}^{T}$ is a selection (technically, a sub-sequence because limited to a finite sample) of the realized values of a family of random variables $\{y_t\}_{-\infty}^{+\infty}$defined on an appropriate probability space.

Time series model for the observations $\{y_t\}_{t=1}^{T}$: a specification of the joint distribution of the set of random variables of which the sampled data are a realization.

- A complete probabilistic time series model would be impractical (very large number of parameters), so we often specify only the first- and second-order moments of such a joint distribution, i.e. the mean, variances and covariances of $\{y_t\}$.

- If the joint distribution of the variables in the random sequence $\{y_t\}$ is multivariate normal, the second-order moments of the process are sufficient to give a complete statistical characterization of $\{y_t\}$.

- When the assumption above is not respected, if the stochastic process can be assumed

to be linear, its second-order characterization is still sufficient. Indeed, the theory of minimum mean squared error linear prediction depends only on the second-order moment properties of the process.

Linear process: A time series $\{y_t\}$ is said to be a linear process if it has the representation

$$y_t = \mu + \sum_{j=-\infty}^{\infty} \phi_j z_{t-j} \quad \forall t$$

where μ is a constant, $\{\phi_j\}$ is a sequence of constant coefficients where $\phi_0 = 1$ and $\sum_{j=-\infty}^{\infty} |\phi_j| < \infty$, and $\{z_t\}$ is a sequence of IID random variables with a defined distribution function. In particular, we assume that the distribution of z_t is continuous, with $E[z_t] = 0$ and $Var[z_t] = \sigma_z^2$. Noticeably, if $\sigma_z^2 \sum_{i=1}^{\infty} \phi_i^2 < \infty$, then y_t is weakly stationary.

1.2 Stationarity

- In order to use past realizations of a variable of interest to forecast its future values, it is necessary for the stochastic process to be stationary.

Strict stationarity: a process is strictly stationary if the joint distribution of the variable associated to any sub-sequence of times $t_1, t_2, ..., t_n$ is the same as the joint distribution of the sequence of all times $t_{1+k}, t_{2+k}, ..., t_{n+k}$ (where k is an arbitrary time shift). In other words, a strictly stationary time series $\{y_t\}$ has the following properties:

1. the random variables y_t are identically distributed;
2. the two random vectors $[y_t, y_{t+k}]'$ and $[y_1, y_{1+k}]'$ have the same joint distribution for any t and k.

- Strict stationarity requires that all the moments of the distribution are time invariant.

Weak stationarity: a stochastic process $\{y_t\}$ is weakly stationary (or, alternatively, covariance stationary) if it has time invariant first and second moments, i.e., if for any choice of $t = 1, 2, ..., \infty$, the following conditions hold:

1. $\mu_y \equiv E(y_t) \quad \text{with} \quad |\mu_y| < \infty$
2. $\sigma_y^2 \equiv E[(y_t - \mu_y)(y_t - \mu_y)] = E[(y_t - \mu_y)^2] < \infty$
3. $\gamma_h \equiv E[(y_t - \mu_y)(y_{t-h} - \mu_y)] \quad \forall h \quad \text{with} \quad |\gamma_h| < \infty$

where $h = ..., -3, -2, -1, 1, 2, 3, ...$

- Weak stationarity requires that the mean and the variance are time invariant.

- Weak stationarity implies that $Cov(y_t, y_{t-h}) = Cov(y_{t-j}, y_{t-j-h}) = \gamma_h$, that is it does not vary over time, but only depends on h.

Autocovariance function (ACVF): $\gamma_h \equiv Cov(y_t, y_{t-h})$ for $h = ..., -3, -2, -1, 1, 2, 3, ...$

Autocorrelation function (ACF) : $\rho_h = \gamma_h/\gamma_0$ where γ_0 is the variance.

- Autocorrelations convey more useful information than autocovariance coefficients do, as the latter depend on the units of measurement of y_t.

- The autocovariance and autocorrelation functions are important for the characterization and classification of time series, given that, for stationary processes, both $\gamma()$ and $\rho()$ should eventually decay to zero.

- Strict stationarity implies weak stationarity (provided that the first and the second moments exist) while the reverse is generally not true.

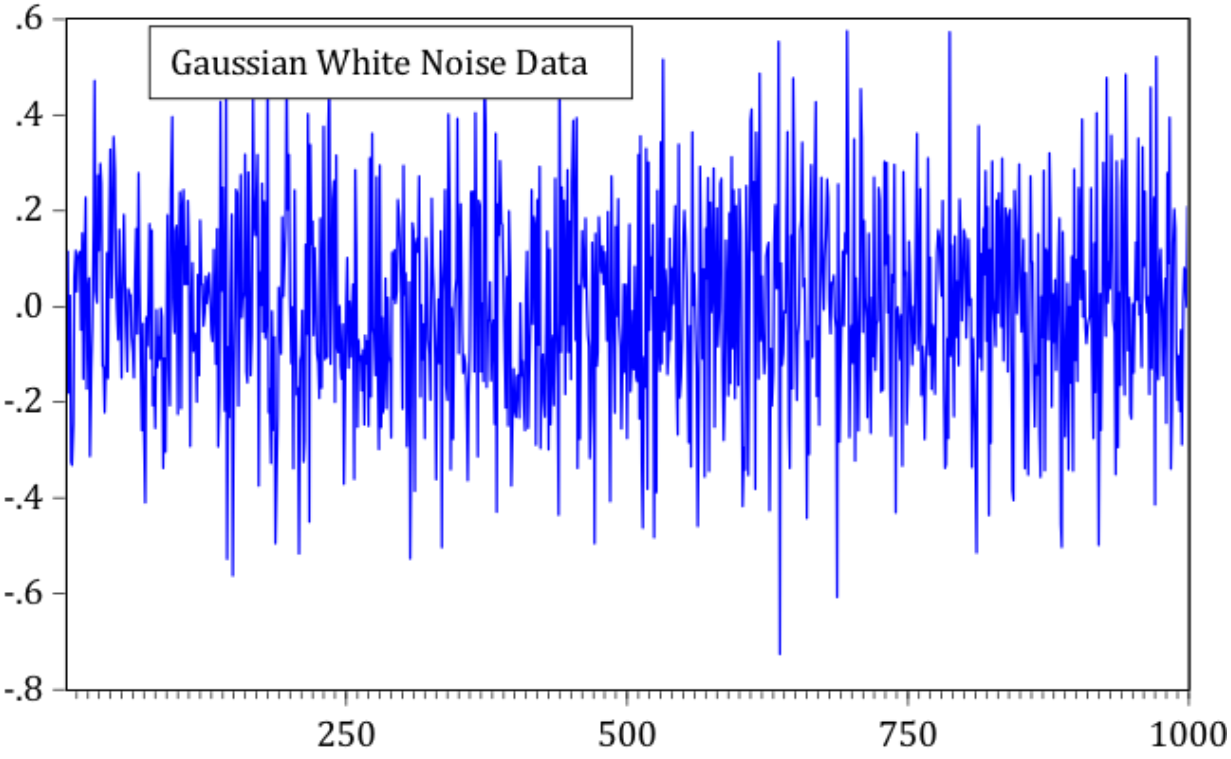

Figure 1: Plot of Data Simulated from a Gaussian white noise process

White Noise: a white noise (WN) process is a sequence of random variables $\{z_t\}$ with mean equal to zero, constant variance equal to σ^2 , and zero autocovariances (and autocorrelations) except at lag zero. If $\{z_t\}$ is normally distributed, we shall speak of a Gaussian white noise.

The figure shows an example of Gaussian white noise, from which 1,000 simulations have been generated with a standard deviation of 0.2. It represents pure noise in the sense that—apart from lucky patterns that may accidentally appear - there is no appreciable structure in the plotted data, either in terms of its levels or in terms of its tendency to randomly fluctuate.

1.3 Sample Autocorrelations and Sample Partial Autocorrelations

- Plotting the data is not always sufficient to determine whether the data generating process (DGP) is stationary or not.
- Since for a series that originates from a stationary process autocorrelations tend to die out quickly as the length h increases, they may be used to investigate the properties of the data.
- We only observe one of the possible realized and necessarily finite sequences of sample data from the process $\{y_t\}$, so we shall only be able to compute sample autocorrelations.

Sample autocorrelations: given a sample of T observations of the variable y_t,$y_1, y_2, ..., y_T$, the estimated or sample autocorrelation function $\hat{\rho}_h$ (where h is a positive integer) is computed as

$$\hat{\rho}_h = \frac{\sum_{t=h+1}^{T}(y_t - \hat{\mu})(y_{t-h} - \hat{\mu})}{\sum_{t=1}^{T}(y_t - \hat{\mu})^2} = \frac{\hat{\gamma}_h}{\hat{\gamma}_0}$$

where $\hat{\mu}$ is the sample mean computed as $\hat{\mu} = T^{-1}\sum_{t=1}^{T} y_t$
$\hat{\gamma}_h$ is the sample autocovariance
$\hat{\gamma}_0$ is the sample variance.

- Sample autocorrelations only provide estimates of the true and unobserved autocorrelations of $\{y_t\}$ and may contain large sample variation and substantial uncertainty.
- They measure the strength of the linear relationship between y_t and the lagged values of the series (the higher the autocorrelation coefficient, the stronger the linear predictive relationship between past and future values of the series).

- If $\{y_t\}$ is an IID process with finite variance, then for a large sample, the estimated autocorrelations will be asymptotically normally distributed as $N(0, 1/T)$.

- Tests of hypotheses on the autocorrelation coefficients can determine whether they are significantly different from zero, e.g. the 95% confidence interval given by $\hat{\rho}_h \pm 1.96\sqrt{T}$.

- If $\{y_t\}$ is a weakly stationary time series satisfying $y_t = \mu + \sum_{j=0}^{q} \phi_j z_{t-j}$ where $\phi_0 = 1$ and $\{z_j\}$ is a Gaussian white noise, then:

$$\hat{\rho}_h \overset{a}{\sim} N(0, T^{-1}(1 + 2\sum_{j=1}^{q} \rho_j^2))$$

- In finite samples $\hat{\rho}_h$ is a biased estimator of ρ_h , with a bias of the order of $1/T$, meaning that it can be large for small samples, but it becomes less and less relevant as the number of observations increases.

- Jointly test whether several M consecutive autocorrelation coefficients are equal to zero using:

  1. Box and Pierce test:
  $$Q(M) \equiv T \sum_{h=1}^{M} \hat{\rho}_h^2 \overset{a}{\sim} \chi_M^2$$
  under the null hypothesis $H_0 = \rho_1 = \rho_2 = ... = \rho_M = 0$
  Therefore, if $Q(M)$ exceeds the critical value from the χ^2 distribution with degrees of freedom at a probability $1 - \alpha$, then H_0 can be rejected and at least one ρ is significantly different from zero.

  2. Ljung and Box test, with better small sample properties than the previous one:
  $$Q^*(M) = T(T + 2) \sum_{h=1}^{M} \frac{\hat{\rho}_h^2}{T - h} \sim \chi_M^2$$

Sample partial autocorrelations: formally, the partial autocorrelation a_h is defined as

$$a_h = Corr(y_t, y_{t-h} | y_{t-1}, ..., y_{t-h+1})$$

Partial autocorrelation function (PACF): sequence $a_1, a_2, .., a_T$ as a function of $h = 1, 2, ...$ of partial autocorrelations.

- The sample estimate $\hat{a}_h$ of the partial autocorrelation at lag h is obtained as the ordinary least square estimator of ϕ_h in an autoregressive model $y_t = \phi_0 + \phi_1 y_{t-1} + ... + \phi_h y_{t-h} + \epsilon_t$.

- They measure the "added" predictive power of the h-th lag of the series y_{t-h}, when $y_{t-1}, ..., y_{t-h+1}$ are already accounted for in the predictive regression.

- ACFs and PACFs can only measure the degree of linear association between y_t and y_{t-k} for $k \geq 1$ and thus they may fail to detect important nonlinear dependence relationships in the data.

2 Moving Average and Autoregressive Processes

2.1 Finite Order Moving Average Processes

Moving average process (MA): a q-th order moving average, $MA(q)$, is a process that can be represented as

$$y_t = \mu + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + ... + \theta_q \epsilon_{t-q}$$

where the process of $\{\epsilon_t\}$ is an IID white noise with mean zero and constant variance equal to σ_ϵ^2 . In a compact form:

$$y_t = \mu + \epsilon_t + \sum_{j=1}^{q} \theta_j \epsilon_{t-j}$$

Properties:
Always stationary as they are finite, linear combination of white noise processes for which the first two moments are time-invariant. Therefore, MA(q) has constant mean, variance and autocovariances that may be different from zero up to lag q, but are equal to zero afterwards.

1. Mean: $E(y_t) = \mu$
2. Variance: $Var(y_t) = \gamma_0 = (1 + \theta_1^2 + \theta_2^2 + ... + \theta_q^2)\sigma^2$
3. Covariance: $\gamma_h = Cov(y_t, y_{t-h}) = (\theta_h + \theta_{h+1}\theta_1 + \theta_{h+2}\theta_2 + ... + \theta_q \theta_{q-h})\sigma^2$ for $h = 1, 2, ..., q$ and $\gamma_h = 0$ for $h > q$

2.2 Autoregressive Processes

Autoregressive process: a p-th order autoregressive process, denoted as $AR(p)$, is a process that can be represented by the p-th order stochastic difference equation

$$y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} + \epsilon_t$$

where the process of $\{\epsilon_t\}$ is an IID white noise with mean zero and constant variance σ_ϵ^2 . In a compact form:

$$y_t = \phi_0 + \sum_{j=1}^{p} \phi_j y_{t-j} + \epsilon_t$$

- $AR(p)$ models are simple univariate devices to capture the Markovian nature of financial and macroeconomic data, i.e., the fact that the series tends to be influenced at most by a finite number of past values of the same series.

Lag operator (L): it shifts the time index of a variables regularly sampled over time backward by one unit, e.g. $Ly_t = y_{t-1}$.

Difference operator (Δ): it expresses the difference between consecutive realizations of a time series, e.g $\Delta y_t = y_t - y_{t-1}$.

We can rewrite the AR model using the lag operator

$$(1 - \phi_1 L - \phi_2 L^2 - ... - \phi_p L^p) y_t = \phi_0 + \epsilon_t$$

In a compact form

$$\phi(L) y_t = \phi_0 + \epsilon_t$$

where $\phi(L)$ is the polynomial of order p, $(1 - \phi_1 L - \phi_2 L^2 - ... - \phi_p L^p)$.

(Reverse) Characteristic equation: the equation obtained by replacing in the polynomial $\phi(L)$ the lag operator L by a variable λ and setting it equal to zero, i.e., $\phi(\lambda) = 0$.

Root of the polynomial $\phi(L)$: any value of λ which satisfies the polynomial equation $\phi(\lambda) = 0$. It is a determinant of the behavior of the time series: if the absolute value of all the roots of the characteristic equations is higher than one the process is said to be stable.

- A stable process is always weakly stationary.

Properties:

1. Mean: $\mu = \frac{\phi_0}{1-\phi_1}$ in case of an $AR(1)$
   $\mu = \frac{\phi_0}{1-\phi_1-\phi_2-...-\phi_p}$ in case of an $AR(p)$

2. Variance: $Var[y_t] = \frac{\sigma_\epsilon^2}{1-\phi_1^2}$ in case of an $AR(1)$
   $Var[y_t] = \frac{\sigma_\epsilon^2}{1-\phi_1^2-\phi_2^2-...-\phi_p^2}$ in case of an $AR(p)$

Wold's decomposition: Every weakly stationary, purely non-deterministic, stochastic process $(y_t - \mu)$ can be written as an infinite, linear combination of a sequence of white noise components:

$$y_t - \mu = \sum_{i=0}^{\infty} \theta_j \epsilon_{t-i}$$

- Wold's theorem states that only an autoregressive process of order p with no constant and no other predetermined, fixed terms can be expressed as an infinite order moving average process, $MA(\infty)$.

- This result is useful for deriving the autocorrelation function of an autoregressive process.

- It can be used to check the stationary condition of the mean and the variance of an AR model.

Example:

1. Mean:
   Starting from an $AR(1)$ model

   $$y_t = \phi_0 + \phi_1 y_{t-1} + \epsilon_t$$

   Take the expectation of the model equation

   $$E(y_t) = \varphi_0 + \phi_1 E(y_{t-1})$$

   Under the stationarity condition, it must be that

   $$E(y_t) = E(y_{t-1}) = \mu$$

   and hence

   $$E(y_t) = \mu = \phi_0 + \phi_1 \mu$$

which solved for the unknown unconditional mean gives:

$$\mu = \frac{\phi_0}{1 - \phi_1}$$

and then $\phi_0 = (1 - \phi_1)\mu$.
For μ to be constant and to exist, it must be that $\phi_1 \neq 1$. Substitute ϕ_0 above in the initial model equation and obtain

$$y_t - \mu = \phi_1(y_{t-1} - \mu) + \epsilon_t$$

It must also be the case that

$$y_{t-1} - \mu = \phi_1(y_{t-2} - \mu) + \epsilon_{t-1}$$

and therefore

$$y_t - \mu = \phi_1(\phi_1(y_{t-2} - \mu) + \epsilon_{t-1}) + \epsilon_t$$

By a process of infinite backward substitution, we obtain

$$y_t - \mu = \epsilon_t + \phi_1\epsilon_{t-1} + \phi_1^2\epsilon_{t-2} + ... = \sum_{i=0}^{\infty} \phi_1^i \epsilon_{t-i}$$

Following similar algebraic steps, we can derive also the unconditional mean of an $AR(p)$ model

$$\mu = \frac{\phi_0}{1 - \phi_1 - \phi_2 - ... - \phi_p}$$

For μ to be constant and to exist, it must be that

$$|\phi_1 + \phi_2 + ... + \phi_p| < 1$$

2. Variance:
Starting from an $AR(1)$ model

$$E[(y_t - \mu)^2] = {\phi^2}_1 E[(y_{t-1} - \mu)^2] + E[{\epsilon_t}^2]$$

under the stationarity assumption $E[(y_t - \mu)^2] = E[(y_{t-1} - \mu)^2] = Var[y_t]$ it becomes

$$Var[y_t] = {\phi^2}_1 Var[y_t] + {\sigma_\epsilon}^2$$

and therefore

$$Var[y_t] = \frac{\sigma_\epsilon^2}{1 - \phi_1^2}$$

provided that ${\phi^2}_1 < 1$.
Therefore, putting all the conditions derived before together, for an $AR(1)$ model, weak stationarity requires $|\phi_1| < 1$.

In case of a general $AR(p)$ model the formula of the variance becomes

$$Var[y_t] = \frac{\sigma_\epsilon^2}{1 - \phi_1^2 - \phi_2^2 - ... - \phi_p^2}$$

In the general case, covariance stationarity can be checked using the Schur criterion.

Schur criterion: Construct two lower-triangular matrices, $\mathbf{A_1}$ and $\mathbf{A_2}$ of the form:

$$\mathbf{A_1} \equiv \begin{pmatrix} 1 & 0 & ... & 0 & 0 \\ -\phi_1 & 1 & & & 0 \\ -\phi_2 & -\phi_1 & & & \\ ... & -\phi_2 & & & \\ & ... & & & 0 \\ -\phi_{p-1} & -\phi_{p-2} & ... & -\phi_1 & 1 \end{pmatrix} \mathbf{A_2} \equiv \begin{pmatrix} -\phi_p & 0 & ... & 0 & 0 \\ -\phi_{p-1} & -\phi_p & & & 0 \\ -\phi_{p-2} & -\phi_{p-1} & & & \\ ... & -\phi_{p-2} & & & \\ & ... & & & 0 \\ -\phi_1 & -\phi_2 & ... & -\phi_{p-1} & -\phi_p \end{pmatrix}$$

The $AR(p)$ process is covariance stationary if and only if the matrix $\mathbf{S} = \mathbf{A_1}\mathbf{A_1'} - \mathbf{A_2}\mathbf{A_2'}$ is positive definite.

The autocovariances and autocorrelations of $AR(p)$ processes can be computed by solving a set of Yule-Walker equations.

Yule-Walker equations for an $AR(2)$ process where $\mu = 0$

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \epsilon_t$$

We multiply the previous equation by y_{t-s} with $s = 1, 2, ...$ and take the expectation of each resulting stochastic difference equation

$$E[y_t y_t] = \phi_1 E[y_{t-1} y_t] + \phi_2 [y_{t-2} y_t] + E[y_t \epsilon_t]$$
$$E[y_t y_{t-1}] = \phi_1 E[y_{t-1} y_{t-1}] + \phi_2 \mathrm{E}[y_{t-2} y_{t-1}] + E[y_{t-1} \epsilon_t]$$
$$E[y_t y_{t-2}] = \phi_1 E[y_{t-1} y_{t-2}] + \phi_2 \mathrm{E}[y_{t-2} y_{t-2}] + E[y_{t-2} \epsilon_t]$$
$$E[y_t y_{t-s}] = \phi_1 E[y_{t-1} y_{t-s}] + \phi_2 \mathrm{E}[y_{t-2} y_{t-s}] + E[y_{t-s} \epsilon_t]$$

By definition

$$E[y_t y_{t-s}] = E[y_{t-k} y_{t-k-s}] = \gamma_s$$

We know that

$$E[\epsilon_t y_t] = \sigma^2 \quad \text{and} \quad E[y_{t-s} \epsilon_t] = 0$$

Therefore

$$\gamma_0 = \phi_1 \gamma_1 + \phi_2 \gamma_2 + \sigma^2$$
$$\gamma_1 = \phi_1 \gamma_0 + \phi_2 \gamma_1$$
$$\gamma_s = \phi_1 \gamma_{s-1} + \phi_2 \gamma_{s-2} \quad s \geq 2$$

Divide γ_1 and γ_s by γ_0

$$\rho_1 = \phi_1 \rho_0 + \phi_2 \rho_1$$
$$\rho_s = \phi_1 \rho_{s-1} + \phi_2 \rho_{s-2} \quad s \geq 2$$

By construction $\rho_0 = 1$, then

$$\rho_1 = \frac{\phi_1}{1 - \phi_2}$$

and, consequently, we can solve by recursive substitution ρ_s for any $s \geq 2$.

- The autocorrelation function of an $AR(p)$ will decay geometrically to zero because the leading term will also take the form of powers of sums of the coefficients which need to be restricted to absolute sums that are less than one.

We can use the sample PACF function to identify the order of an $AR(p)$ model. From the definition of PACF

$$y_t = \phi_{0,1} + \phi_{1,1} y_{t-1} + e_{1t}$$
$$y_t = \phi_{0,2} + \phi_{1,2} y_{t-1} + \phi_{2,2} y_{t-2} + e_{2t}$$
$$y_t = \phi_{0,3} + \phi_{1,3} y_{t-1} + \phi_{2,3} y_{t-2} + \phi_{3,3} y_{t-3} + e_{3t}$$

These models are in the form of multiple linear regressions and can be estimated by simple least squares, so that $\hat{\phi}_{j,j}$ is the sample j-order PACF of y_t and should converge to zero for all orders $j > p$.

Invertibility: An invertible MA(q) model can be expressed as an $AR(\infty)$:

$$y_t = \sum_{i=0}^{\infty} \phi_i L^i y_{t-1} + u_t$$

A MA(q) is invertible when the magnitude of all the roots of the MA polynomial exceeds the one.

- The ACF of a MA model has the same shape of the PACF of an AR model, and the PACF of a MA model has the same shape of an AR model.

2.3 Autoregressive Moving Average Processes

ARMA process: A time series is said to follow an ARMA(p, q) if it satisfies

$$y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + ... + ...\theta_q \epsilon_{t-q} + \epsilon_t$$

where the process of $\{\epsilon_t\}$ is an IID white noise with mean zero and constant variance equal to σ_ϵ^2. In a compact form:

$$y_t = \phi_0 + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{i=0}^{q} \theta_i \epsilon_{t-i}$$

with $\theta_0 = 1$.
Using the lag operator

$$(1 - \sum_{i=1}^{p} \phi_i L^i) y_t = \phi_0 + \sum_{i=0}^{q} \theta_i \epsilon_{t-i}$$

with $\theta_0 = 1$, so that the solution for y_t is stable and equal to:

$$y_t = \frac{\phi_0 + \sum_{i=0}^{q} \theta_i \epsilon_{t-i}}{1 - \sum_{i=1}^{p} \phi_i L^i}$$

and will be covariance stationary if the roots of the polynomial $(1-\phi_1 L-\phi_2 L^2-...-\phi_p L^p)$ lie outside the unit circle.

- ARMA models allow to overcome the problem of AR and MA models with many parameters, as they are a compact-form combination of the two.

- ARMA models are rarely needed to fit the most commonly used data, but are highly relevant in volatility modeling.

Properties:

1. Mean: $\mu = \frac{\phi_0}{1-\phi_1}$ in case of an ARMA(1,1)
   $\mu = \frac{\phi_0}{1-\phi_1-\phi_2-...-\phi_p}$ in case of an ARMA(p,q)

   that can be derived, for instance when p=q=1, by taking the expectation of the ARMA process equation $E[y_t] = \phi_0 + \phi_1 E[y_{t-1}] + \theta_1 E[\epsilon_{t-1}] + E[\epsilon_t]$ and substituting $E[\epsilon_{t-1}] = E[\epsilon_t] = 0$.

2. ACF: $\rho_1 = \frac{\gamma_1}{\gamma_0} = \frac{(1+\theta_1\phi_1)(\theta_1+\phi_1)}{(1+\theta_1^2+2\theta_1\phi_1)}$
   $\rho_s = \frac{\gamma_s}{\gamma_0} = \phi_1\rho_{s-1}$for any $\quad s \geq 2$

   Derived by solving the Yule-Walker equations

$$E[y_t y_t] = \phi_1 E[y_{t-1} y_t] + \theta_1 E[\epsilon_{t-1} y_t] + E[\epsilon_t y_t]$$

$$E[y_t y_{t-1}] = \phi_1 E[y_{t-1} y_{t-1}] + \theta_1 E[\epsilon_{t-1} y_{t-1}] + E[\epsilon_t y_{t-1}]$$
$$E[y_t y_{t-2}] = \phi_1 E[y_{t-1} y_{t-2}] + \theta_1 E[\epsilon_{t-1} y_{t-2}] + E[\epsilon_t y_{t-2}]$$
$$E[y_t y_{t-s}] = \phi_1 E[y_{t-1} y_{t-s}] + \theta_1 E[\epsilon_{t-1} y_{t-s}] + E[\epsilon_t y_{t-s}]$$

   so that we find

$$\gamma_0 = \phi_1\gamma_1 + \theta_1(\phi_1+\theta_1)\sigma_\epsilon^2 + \sigma_\epsilon^2$$
$$\gamma_1 = \phi_1\gamma_0 + \theta_1\sigma_\epsilon^2$$
$$\gamma_2 = \phi_1\gamma_1$$
$$\gamma_s = \phi_1\gamma_{s-1}$$

   Plug γ_1 into γ_0

$$\gamma_0 = \phi_1(\phi_1\gamma_0 + \theta_1\sigma_\epsilon^2) + \theta_1(\phi_1+\theta_1)\sigma_\epsilon^2 + \sigma_\epsilon^2$$

   and thus

$$\gamma_0 = \frac{(1+\theta_1^2+2\theta_1\phi_1)\sigma_\epsilon^2}{1-\phi_1^2}$$

   Substitute the last equation into γ_1

$$\gamma_1 = \frac{(1 \pm \theta_1\phi_1)(\gamma_1 \pm \phi_1)}{(1-\phi_1^2)}\sigma_\epsilon^2$$

ρ_1 is than equal to

$$\rho_1 = \frac{\gamma_1}{\gamma_0} = \frac{(1+\theta_1\phi_1)(\theta_1+\phi_1)}{(1+\theta_1^2+2\theta_1\phi_1)}$$

$$\rho_s = \frac{\gamma_s}{\gamma_0} = \phi_1\rho_{s-1} \quad \text{for any} \quad s \geq 2$$

For a general ARMA(p,q) model, beginning with lag q the values of ρ_s will satisfy:

$$\rho_s = \frac{\gamma_s}{\gamma_0} = \phi_1\rho_{s-1} + \phi_2\rho_{s-2} + ... + \phi_p\rho_{s-p}$$

- After the first lag, the autocorrelation functions of an ARMA(1,1) will decline geometrically at a rate that depends on ϕ_1, meaning that after the first lag, it is the AR component of the process that determines the behavior of the ACF.

- After the qth lag, the ACF of an ARMA model is geometrically declining, similarly to the one of a pure AR(p) model.

- The initial p values for the series of interest can be treated as initial conditions that satisfy the Yule-Walker equations. For these initial p lags, the shape of the ACF is determined by the characteristic equation.

- Difference between an AR(p) and an ARMA(p,q) model: both have geometrically declining ACFs, but the former will have a PACF which cuts off to zero after p lags, while the latter will have a PACF which declines geometrically.

3. PACF: use the Yule-Walker equations and compute the partial autocorrelation coefficients from the autocorrelations as:

$$\phi_{1,1} = \rho_1$$

$$\phi_{2,2} = \frac{\rho_2 - \rho_1^2}{1-\rho_1^2}$$

$$\phi_{3,3} = \frac{\rho_s - \sum_{j=1}^{s-1}\phi_{s-1,j}\rho_{s-j}}{1-\sum_{j=1}^{s-1}\phi_{s-1,j}\rho_j}$$

where $\phi_{s,j} = \phi_{s-1,j} - \phi_{s,s}\phi_{s-1,s-j}$ with $j = 1, 2, ..., s-1$.

3 Selection and Estimation of AR, MA and ARMA models

3.1 The Selection of the Model and the Role of Information criteria

Box and Jenkins identification procedure:

1. Compute the sample ACF and PACF of the observed time series.
2. Compare the results with the theoretical shape of the ACF and PACF of the alternative models, summarized in the table.

	ACF	PACF
AR(p)	Decays towards zero.	Cuts off after lag p.
MA(q)	Cuts off after lag q.	Decays towards zero.
ARMA(p, q)	Decays towards zero starting at lag q.	Decays towards zero starting at lag p.

Figure 2: A summary of the characteristics of AR, MA and ARMA models

- The interpretation of the ACF and PACF can be difficult and sometimes even subjective.

Information criteria (IC): criteria that trade off the goodness of (in-sample) fit and the parsimony of the model and provide a (cardinal, even if specific to an estimation sample) summary measure.
They always include two terms to avoid the so-called overfitting problem:

1. a function of the sum of squared residual (SSR)
2. a penalty for the loss of degrees of freedom from the number of parameters that the model implies.

- The best performing model is the one that minimizes the information criteria.

Overfitting: a phenomenon deriving from the use of too many parameters, due to the fact that in this case noise, rather than the dependence structure in the data is fitted, thus reducing the forecasting power of the model.

Example: the sample residual variance as a measure of the goodness of fit favors larger models, as they tend provide a better in-sample fit: each estimated parameter provides additional flexibility in approximating the data set.

Common Information criteria (classified by increasing penalty):

1. Akaike IC: $AIC = ln(\hat{\sigma}^2) + \frac{2k}{T}$
   - It asymptotically overestimates the order/complexity of a model with positive probability.
2. Hannan-Quinn IC: $HQIC = ln(\hat{\sigma}^2) + \frac{2k}{T} ln(ln(T))$
3. Schwarz's Bayesian IC: $SBIC = ln(\hat{\sigma}^2) + \frac{2k}{T} ln(T)$
   - It is a consistent criterion: it determinates the true model asymptotically, meaning that, as the sample size approaches infinity, it will select the correct model order.

where T is the sample size, $\hat{\sigma}^2$ is the residual variance and k is the total number of parameters to be estimated.

- The three IC may select different models.

- No criteria will always outperform the others: the ICs should give supplementary guidance and not substitute the visual inspection of sample ACFs and PACFs.

3.2 Estimation methods

1. Estimation of an AR model, using the method of (conditional) ordinary least square (OLS):

   Starting from
   $$y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} + \epsilon_t$$
   where $\phi_0 = (1 - \phi_1 - \phi_2 - ... - \phi_p)\mu$, so that $\mu \equiv T^{-1} \sum_{t=1}^{T} y_t$.
   Assume that pre-sample values $y_{1-p}, ..., y_0$ are available in addition to the sample values $y_1, ..., y_T$, then

   (a) OLS estimator of $\varphi = [\phi_0, \phi_1, ..., \phi_p]'$:
   $$\hat{\varphi} = (\sum_{t=1}^{T} x_{t-1} x_{t-1}')^{-1} \sum_{t=1}^{T} x_{t-1} y_t$$
   where $x_{t-1} = [1, y_{t-1}, ..., y_{t-p}]'$.

(b) Fitted model:

$$y_t = \hat{\phi}_0 + \hat{\phi}_1 y_{t-1} + \hat{\phi}_2 y_{t-2} + ... + \hat{\phi}_p y_{t-p} + \hat{\epsilon}_t$$

(c) Associated residuals:

$$\hat{\epsilon}_t = y_t - \hat{y}_t$$

(d) Estimated variance of the residuals:

$$\hat{\sigma}^2_\epsilon = \frac{\sum_{t=p+1}^{T} \hat{\epsilon}^2_t}{T - 2p - 1}$$

- If the process is normally distributed, the OLS estimation is equivalent to the maximum likelihood estimation (MLE) conditional on initial values.

2. Estimation of an ARMA model, using the likelihood maximization procedure:
Define the probability distribution of the observed data.
The log-likelihood function $\phi(L)y_t = \theta(L)\epsilon_t$ has the form

$$lnL(\phi_1, ..., \phi_p, ..., \theta_1, ...\theta_q) = \sum_{t=1}^{T} \ell_t(\cdot)$$

where

$$\ell_t(\cdot) = -\frac{1}{2} ln 2\pi - \frac{1}{2} ln\sigma^2_\epsilon - \frac{(\theta(L)^{-1}\phi(L)y_t)^2}{2\sigma^2_\epsilon}$$

under the assumption that the conditional distributions of y_t are normally distributed.

- Under general conditions, the resulting estimators are consistent and have an asymptotic normal distribution.

- Iterative numerical algorithms are used to optimize the log-likelihood function, that requires start-up values for the parameters. Therefore, the estimates reported by standard software packages can strongly depend on the methods used to set such starting conditions.

3.3 Residual Diagnostics

- Given that if the model is correctly specified, the residuals shall not exhibit any predictable pattern, an analysis of the residuals of the estimated model can be used to check its adequacy.

Analysis of the residuals:

1. Plot the standardized residuals, i.e. $\hat{\epsilon}_t^s = (\hat{\epsilon}_t^s - \hat{\bar{\epsilon}})/\hat{\sigma}_\epsilon$, where $\hat{\bar{\epsilon}}$ is the predicted mean for time t and $\hat{\sigma}_\epsilon$ is the standard deviation of the residuals.

   - If the model is correctly specified, that is the residuals are normally distributed with zero mean and unit variance, then approximately 95% of the standardized residuals should fall in an interval of ± 2 around zero.

2. Plot the squared (standardized) residuals.

   - If the model is correctly specified, such a plot should not display any clusters, i.e., the tendency of high (low) squared residuals to be followed by other high (low) squared standardized residuals.

3. Jarque-Bera test:
   $$H_0 : \frac{\mu_3}{\sigma_3} = 0 \quad \text{and} \quad H_0 : \frac{\mu_4}{\sigma_4} - 3 = 0$$
   Because the normal distribution is symmetric, μ_3 should be zero and μ_4 should satisfy $\mu_4 = 3\sigma_\epsilon^4$.
   It can be seen as a joint test that λ_1 and λ_2 are zero and tested as
   $$H_0 : \lambda_3 = 0$$
   where $\lambda_3 = \lambda^2{}_1 + \lambda^2{}_2$, which is asymptotically distributed as a $\chi^2(2)$, thanks to the fact that
   $$\lambda_1 = \frac{1}{6T}\sum_{t=1}^{T}(\frac{\hat{\epsilon}_t^3}{\hat{\sigma}_\epsilon^3}) \text{ and } \quad \lambda_2 = \frac{1}{24T}\sum_{t=1}^{T}(\frac{\hat{\epsilon}_t^4}{\hat{\sigma}_\epsilon^4} - 3)$$
   are both N(0,1) distributed.

   - In case of small samples, the results of the Jarque-Bera test should be interpreted with caution, given that small sample properties of the sample moments may deviate considerably from their theoretical counterparts.

4. Compute the residuals sample autocorrelations and perform tests of hypotheses to assess whether there is any remaining serial dependence in them.

   - If the residuals are truly uncorrelated, we would expect 95% of the sample autocorrelations to fall inside the approximate (asymptotically valid) confidence interval $\pm 2/\sqrt{T}$ around zero.

5. Durbin-Watson (DW) test: it assesses the presence of first-order serial correlation in the residuals of an estimated ARMA model.

$$H_0 : \rho = 0 \quad \text{and} \quad H_1 : \rho > 0$$

$$DW = \frac{\sum_{t=2}^{T}(\hat{\epsilon}_t - \hat{\epsilon}_{t-1})^2}{\sum_{t=1}^{T}\hat{\epsilon}_t^2} \simeq 2(1 - \hat{\rho}_1)$$

where $\epsilon_t = \rho\epsilon_{t-1} + u_t$ and $\hat{\rho}_1$ is the one-lag ACF of $\{\hat{\epsilon}_t\}$.

- If there is no serial correlation ($\hat{\rho}_1$ is near zero), the $DW \simeq 2$; if there is a positive correlation $DW < 2$.

- The DW test does not follow a standard statistical distribution and thus we must compare it with two sets of critical values, d_U and d_L, that depend on the chosen level of significance, on the number of observations and of explanatory variables:

  (a) If $DW < d_L$ we reject H_0 in favor of H_1;

  (b) If $DW > d_U$ we fail to reject H_0;

  (c) If $d_L \leq DW \leq d_U$ the test is inconclusive.

Skewness: index of asymmetry based on μ_3

$$\hat{S} = \frac{1}{T}\sum_{t=1}^{T}\frac{\hat{\epsilon}_t^3}{\hat{\sigma}_\epsilon^3}$$

Excess kurtosis: index of tail thickness based on μ_4

$$\hat{K} = \frac{1}{T}\sum_{t=1}^{T}\frac{\hat{\epsilon}_t^4}{3\hat{\sigma}_\epsilon^4} - 3$$

4 Forecasting ARMA processes

4.1 Standard Principles of Forecasting

Forecasting: attempt to predict which value a random variable is most likely to take in the future.

In-sample forecast: forecast generated with reference to the same data that were used to estimate the parameters of the model.

Out-of-sample forecast: forecast that predict the value of observations that were not used initially to specify and estimate the model.

One-step-ahead forecast $\hat{y}_t(1)$: forecast of the value that a random variable y_t is likely to assume in the next period. It is computed at time t on the basis of the information then available.

Multi-step-ahead forecast $\hat{y}_t(h)$: forecast at time t that the same random variable is likely to assume at time $h > 1$, y_{t+h}.

Loss function: function that defines how concerned we are if our forecast were to be off relative to the realized value, by a certain amount.

Mean square forecast error (MSFE) associated to the forecast $\hat{y}_t(h)$:

$$MSFE[\hat{y}_t(h)] \equiv E[(y_{t+h} - \hat{y}_t(h))^2]$$

- Very convenient results are obtained if one assumes a quadratic loss function, that is a forecast $\hat{y}_t(h)$ that minimizes the MSFE.

- $MSFE[\hat{y}_t(h)]$ is minimized when $\hat{y}_t(h)$ is equal to $E[y_{t+h}|\Im_t]$, where $\Im_t$ is the information set available to the forecaster.

- The conditional mean of y_{t+h} given its past observations is the best estimator of $\hat{y}_t(h)$ in terms of MSFE.

  Proof:

  Starting from

  $$MSFE[\hat{y}_t(h)] \equiv E[(y_{t+h} - \hat{y}_t(h))^2]$$

  add and subtract $E[y_{t+h}|\Im_t]$

  $$MSFE[\hat{y}_t(h)] \equiv E[(y_{t+h} - E[y_{t+h}|\Im_t] + E[y_{t+h}|\Im_t] - \hat{y}_t(h))^2]$$

  Given that $y_{t+h} - E[y_{t+h}|\Im_t]$ only depends on $\epsilon_{t+1}, ..., \epsilon_{t+h}$ and $E[y_{t+h}|\Im_t] - \hat{y}_t(h)$ only depends on $y_t, y_{t-1}, y_{t-2}, ...$

  $$E[(y_{t+h} - E[y_{t+h}|\Im_t])(E[y_{t+h}|\Im_t] - \hat{y}_t(h))] = 0$$

Therefore, squaring the expression in brackets in the MSFE equation above we obtain

$$MSFE[\hat{y}_t(h)] = MSFE(E[y_{t+h}|\Im_t]) + E[(E[y_{t+h}|\Im_t] - \hat{y}_t(h))^2]$$

which is minimized when $\hat{y}_t(h) = E[y_{t+h}|\Im_t]$.

4.2 Forecasting an AR(p) Process

1. One-step-ahead forecast:

   Starting from the equation of an AR(p)

   $$y_{t+1} = \phi_0 + \phi_1 y_t + ... + \phi_p y_{t+1-p} + \epsilon_{t+1}$$

   Under the MSFE loss function

   $$\hat{y}_t(1) = E[y_{t+1}|y_t, y_{t-1}, y_{t-2}, ...] = \phi_0 + \sum_{i=1}^{p} \phi_i y_{t+1-i}$$

   and the associated forecast error is

   $$u_t(1) = y_{t+1} - \hat{y}_t(1) = \epsilon_{t+1}$$

   Therefore, the variance of the one-step-ahead forecast is

   $$Var[u_t(1)] = Var(\epsilon_{t+1}) = \sigma_\epsilon^2$$

   and the confidence interval for the point forecast y_{t+1}, when ϵ_t is normally distributed, is

   $$\hat{y}_t(1) \pm 1.96 \times \sigma_\epsilon$$

   - Forecasts are computed using estimated parameters of the model and not the true, unobserved ones, so they do not consider the uncertainty in the parameter estimates.

   - When the sample size used in the estimation of the model is large, the forecast based on estimated parameters should be close to the true one.

2. Multi-step ahead forecast:

   $$\hat{y}_t(h) = E[y_{t+h}|y_t, y_{t-1}, ...] = \phi_0 + \sum_{i=1}^{p} \phi_i \hat{y}_t(h-i)$$

   where $\hat{y}_t(i) = y_{t+i}$ if $i < 1$.
   The h-step-ahead forecast error is

   $$u_t(h) = y_{t+h} - \hat{y}_t(h)$$

Mean reversion property: for a stationary AR(p) model, $\hat{y}_t(h)$ converges to $E[y_t]$ as h approaches infinity.

- When the forecast horizon is long enough, the best prediction of an AR(p) process is its unconditional mean and the variance of the forecast error is equal to the unconditional variance of the process.