Contents

3 ARMA models

We introduce a class of models which constitute the standard "Box–Jenkins" toolbox for time series analysis and prediction since the 1960s. In the previous notes we have introduced (autoregressive) AR and (moving average) MA models, as two simple examples of time series models. Here we will see that these are instances of a larger class of models, called ARMA, which can be used to produce forecast along with standard errors for a variety of applications.

3.1 Linear processes

First we will define a larger class of processes, called "linear". Let $(W_t)_{t\in\mathbb{Z}}$ be uncorrelated random variables with zero mean and variance σ_W^2 (white noise). The time index is a positive or negative integer, i.e. $t \in \mathbb{Z}$. A process $(Y_t)_{t\in\mathbb{Z}}$ is linear if it can be represented as

$$\forall t \in \mathbb{Z} \quad Y_t = \sum_{j=-\infty}^{\infty} \psi_j W_{t-j}, \tag{3.1}$$

where $(\psi_t)_{t\in\mathbb{Z}}$ is a sequence of real numbers. Because of the infinite sum in the definition we might first wonder whether Y_t is well-defined, and under what conditions on $(\psi_t)_{t\in\mathbb{Z}}$.

At this point the sum over $j \in \mathbb{Z}$ might seem confusing: Y_t is defined in terms of all noise terms, past and future. We could restrict the sum to be over $j \in \mathbb{N}$, or equivalently, assume that the coefficients ψ_j for $j \leq -1$ are equal to zero. On the other hand the symmetry in (3.1) is appealing and relates to the intuitive notion of moving average.

We can always define variables $Z_t^{(n)} = \sum_{j=-n}^{n} \psi_j W_{t-j}$ for all $n \in \mathbb{N}$. The question is whether we can let n go to infinity: does it make sense to write $Y_t = \lim_{n\to\infty} Z_t^{(n)}$?

Here is a proof that this definition makes sense:

Proof. We have $|\sum_{j=-n}^{n} \psi_j W_{t-j}| \leq \sum_{j=-n}^{n} |\psi_j||W_{t-j}|$ by the triangle inequality. Furthermore,

$$\mathbb{E}[\lim_{n\to\infty} \sum_{j=-n}^{n} |\psi_j||W_{t-j}|] = \lim_{n\to\infty} \mathbb{E}[\sum_{j=-n}^{n} |\psi_j||W_{t-j}|]$$

by the monotone convergence theorem, since $\sum_{j=-n}^{n} |\psi_j||W_{t-j}|$ is increasing with n. Both sides of the equality could be infinite. On the right hand side we have

$$\lim_{n\to\infty} \mathbb{E}[\sum_{j=-n}^{n} |\psi_j||W_{t-j}| = \lim_{n\to\infty} \sum_{j=-n}^{n} |\psi_j| \mathbb{E}\left[|W_{t-j}|\right].$$

By Hölder's inequality, $\mathbb{E}\left[|W_t|\right] \leq \mathbb{E}[|W_t|^2]^{\frac{1}{2}} = \sigma_W$. So the right hand side is finite if $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$. We have just shown that, under that condition, $\lim_{n\to\infty} \sum_{j=-n}^{n} |\psi_j||W_{t-j}|$ is finite with probability one, and therefore $\sum_{j=-\infty}^{\infty} \psi_j W_{t-j}$ is finite with probability one. In other words if we assume $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$ then Y_t is well-defined.

■

The linear process is weak stationary. Indeed, the conditions on the mean, variance and autocovariance functions are satisfied.

1. The mean is constant in time and equal to zero:

$$\mathbb{E}\left[Y_t\right] = \mathbb{E}\left[\lim_{n\to\infty} \sum_{j=-n}^{n} \psi_j W_{t-j}\right] = \lim_{n\to\infty} \mathbb{E}\left[\sum_{j=-n}^{n} \psi_j W_{t-j}\right] = 0.$$

   We have switched limit and expectation above; this requires some technical conditions, e.g. assuming $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$ and using the dominated convergence theorem.

2. The autocovariance function $\gamma\left(h\right) = \mathbb{C}\mathrm{ov}(Y_t, Y_{t+h})$ satisfies for all $h \geq 0$

$$\gamma(h) = \sum_{j=-\infty}^{\infty} \psi_j \psi_{j+|h|} \sigma_W^2. \tag{3.2}$$

   which does not depend on t.

Proof. Let's check this: first, since $\mathbb{E}[Y_t] = 0$ then $\mathbb{C}\mathrm{ov}\,(Y_t, Y_{t+h}) = \mathbb{E}\,[Y_t Y_{t+h}]$. Next,

$$\begin{aligned}\mathbb{E}\,[Y_t Y_{t+h}] &= \mathbb{E}\left[\left(\sum_{j=-\infty}^{\infty} \psi_j W_{t+h-j}\right)\left(\sum_{k=-\infty}^{\infty} \psi_k W_{t-k}\right)\right] \\ &= \sum_{j=-\infty}^{\infty}\sum_{k=-\infty}^{\infty} \psi_j \psi_k \mathbb{E}\,[W_{t+h-j} W_{t-k}]\,, \quad \text{(swapping limit and expectation)}\end{aligned}$$

where $\mathbb{E}\,[W_{t+h-j}W_{t-k}] = 0$ if $h - j \neq -k$ and $\mathbb{E}\,[W_{t+h-j}W_{t-k}] = \sigma_W^2$ otherwise. The above expression simplifies to Eq. (3.2). The swap of limit and expectation is again justifiable using dominated convergence and the condition $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$, which also ensures that (3.2) is finite for all h. ■

An important result, called ***Wold representation theorem*** states that all weak stationary processes $(Y_t)_{t\in\mathbb{Z}}$ with zero mean can be represented as

$$Y_t = \sum_{j=0}^{\infty} d_j \varepsilon_{t-j} + \eta_t, \tag{3.3}$$

where (ε_t) constitute noise terms (uncorrelated variables with zero mean), and η_t is called a deterministic or predictable process: η_t is a random variable that is a deterministic function of all the "past" random variables $Y_{t-1}, Y_{t-2}, \ldots$, and η_t is uncorrelated with the noise terms ε_s for all s, t. Finally, a technical condition imposes $\sum_j d_j^2 < \infty$ to ensure the variance of Y_t is finite.

The result suggests that linear processes are flexible enough to represent any stationary process. However linear processes are defined through a infinite sequence of coefficients $(\phi_j)_{j\in\mathbb{Z}}$, and thus are not direct candidates for statistical models: we would not want to estimate infinitely many parameters from a finite observed series. ARMA models constitute a subclass of linear processes, parametrized by a user-chosen number of coefficients, that proves convenient as a statistical model for many time series.

3.2 Autoregressive models

Definition. Introduce B, the *backshift operator*, also denoted L for *lag operator*, such that $BY_t = LY_t = Y_{t-1}$, i.e. the operator shifts the index of the process by one unit backward (we can think of it as operating on the "notation"). It is an operator for it transforms the process

$\{\ldots, Y_{t-1}, Y_t, Y_{t+1}, \ldots\}$ into another process which at time t takes value Y_{t-1}:

$$\begin{array}{rl} & \{\ldots, Y_{t-1}, Y_t, Y_{t+1}, \ldots\} \\ B, L \downarrow & \\ & \{\ldots, Y_{t-2}, Y_{t-1}, Y_t, \ldots\} \end{array}$$

Likewise, define $B^2 = BB$ so that $B^2 Y_t = BY_{t-1} = Y_{t-2}$, and so forth, $B^p Y_t = Y_{t-p}$ for any $p \in \mathbb{N}$.

We define an autoregressive process (Y_t), written AR(p), as

$$Y_t = \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \ldots + \varphi_p Y_{t-p} + W_t, \tag{3.4}$$

where $p \in \mathbb{N}$ and $\varphi_1, \ldots, \varphi_p$ are the autoregressive coefficients in $\mathbb{R}$, with $\varphi_p \neq 0$. We can write (3.4) as

$$\varphi(B) Y_t = W_t, \tag{3.5}$$

where φ is called the *autoregressive polynomial* defined as $\varphi(B) = 1 - \varphi_1 B - \ldots - \varphi_p B^p$. We can change the AR($p$) model to $\varphi(B)(Y_t - \mu) = W_t$, for some $\mu \in \mathbb{R}$, and then we allow for a non-zero mean μ in the process.

Remark 3.1. The AR(p) process is an example of a Markov chain of order p, which means that, conditionally on $(Y_{t-1}, \ldots, Y_{t-p})$, Y_t is independent of all the variables before time $t - p$.

To quickly see the connection with linear processes, consider the case $p = 1$: $\varphi(B) = 1 - \varphi_1 B$, and unraveling the definition of Y_t recursively we get

$$Y_t = \varphi_1 Y_{t-1} + W_t = \varphi_1^2 Y_{t-2} + \varphi_1 W_{t-1} + W_t = \ldots = \varphi_1^{k+1} Y_{t-k-1} + \sum_{j=0}^{k} \varphi_1^j W_{t-j}. \tag{3.6}$$

Next suppose that $|\varphi_1| < 1$, so that $\varphi_1^k \to 0$ as k goes to infinity, and we find

$$Y_t = \sum_{j=0}^{\infty} \varphi_1^j W_{t-j}. \tag{3.7}$$

Thus we recognize a linear process with coefficients $\psi_j = 0$ for $j \leq -1$ and $\psi_j = \varphi_1^j$ for $j \geq 0$. The above motivates a restriction for AR(1) coefficients: we impose $|\varphi_1| < 1$. We will see below how this condition generalizes to AR(p) processes.

Initial distribution. We have not yet mentioned the initial distribution for (Y_t), i.e. the distribution of $Y_0, \ldots, Y_{p-1}$ from which to start the recursion in (3.4). A natural choice is to define

the initial distribution in such a way that the process is stationary. Let us consider again the case of AR(1) for simplicity, and let's find a distribution for Y_0 such that (Y_t) is stationary. First, we want $\mathbb{E}[Y_t]$ to be constant. Since $\mathbb{E}[Y_t] = \varphi_1 \mathbb{E}[Y_{t-1}] + \mathbb{E}[W_t] = \varphi_1 \mathbb{E}[Y_t]$, we need $\mathbb{E}[Y_t] = 0$ for all $t \in \mathbb{N}$. Secondly, for stationarity we require the covariance function $\gamma(t,s) = \mathbb{C}\mathrm{ov}(Y_t, Y_s)$ to depend only on the lag $|t-s|$, and not on the values of t and s. Considering the case $|t-s| = 0$, this implies that $\mathbb{V}[Y_t]$ is constant for all t; let's write $\upsilon = \mathbb{V}[Y_t]$. For the AR(1) process, we have $\mathbb{V}[Y_t] = \varphi_1^2 \mathbb{V}[Y_{t-1}] + \mathbb{V}[W_t]$, where we use the fact that the noise terms (W_t) are uncorrelated. In other words, $\upsilon = \varphi_1^2 \upsilon + \sigma_W^2$, i.e. $\upsilon = \sigma_W^2/(1-\varphi_1^2)$. With this choice of υ, if $\mathbb{V}[Y_0] = \upsilon$ then $\mathbb{V}[Y_t] = \upsilon$ for all t. Thus, to ensure stationarity we need to initialize Y_0 in such a way that $\mathbb{E}[Y_0] = 0$ and $\mathbb{V}[Y_0] = \upsilon$. Assuming that (W_t) is i.i.d. $\mathcal{N}(0, \sigma_W^2)$, we can set $Y_0 \sim \mathcal{N}(0, \upsilon)$ and the process (Y_t) is then stationary.

Another way to understand the initial condition, is to consider that the process started many periods ago at $t - \kappa$, so expression (3.6) rewrites

$$Y_t = \varphi_1 Y_{t-1} + W_t = \varphi_1^\kappa Y_{t-\kappa} + \sum_{j=0}^{\kappa-1} \varphi_1^j W_{t-j}$$

with moments

$$\begin{aligned}
\mathbb{E}[Y_t] &= \varphi_1^\kappa \mathbb{E}[Y_{t-\kappa}] \\
\mathbb{V}[Y_t] &= \varphi_1^{2\kappa} \mathbb{V}[Y_{t-\kappa}] + \sigma_W^2 \sum_{j=0}^{\kappa-1} \varphi_1^{2j} = \varphi_1^{2\kappa} \mathbb{V}[Y_{t-\kappa}] + \frac{1-\varphi_1^{2\kappa}}{1-\varphi_1^2} \sigma_W^2,
\end{aligned}$$

so we see that if the process started $\kappa \to \infty$ periods ago, *at the beginning of times*, then Y_t is normal and its expectation and variance are $Y_t \sim \mathcal{N}(0, \sigma_W^2/(1-\varphi_1^2))$. In other words we can also think that Y_0 is the output of an infinite history. In a heuristic way, stationarity is reached when the process originated many periods ago.

Remark 3.2. Alternatively it is possible to "condition" on the initial value of the series, i.e. in the case of AR(p) models to restrict the series being modeled to $Y_{p+1}, \ldots, Y_n$, and to consider the first p values as fixed. This way there is no need to specify an initial distribution. The disadvantage is that the series used to calibrate the model has p fewer values.

Condition on the coefficients of AR(p) models. Consider the autoregressive polynomial $\varphi(B) = 1 - \sum_{j=1}^p \varphi_j B^j$. We suspect that there might be conditions on the coefficients $(\varphi_j)_{j=1}^p$ for the AR(p) process to be stationary, similarly to the condition $|\varphi_1| < 1$ identified above. Let's

assume that we can find a sequence $(\psi_j)_{j\geq 0}$ such that

$$Y_t = \sum_{j=0}^{\infty} \psi_j W_{t-j}. \tag{3.8}$$

Equation (3.8) is called the *infinite moving average* (or MA(∞)) representation of the process. We recognize a particular case of a linear process as in Eq. (3.1). Then we know that under the assumption $\sum_{j\geq 0} |\psi_j| < \infty$ the process is stationary with autocovariance given by Eq. (3.2).

Now the question is to find some condition on the AR coefficients $(\varphi_j)_{j=1}^p$ so that the AR(p) process (Y_t) has an MA(∞) representation as in Eq. (3.8). Introduce the polynomial $\psi(B) = \sum_{j=0}^{\infty} \psi_j B^j$ such that $Y_t = \psi(B) W_t$; then we can rewrite Eq. (3.5) as

$$\begin{aligned} &\varphi(B)\,\psi(B)\,W_t = W_t \\ \Leftrightarrow \quad &(1 - \sum_{j=1}^{p} \varphi_j B^j)(\sum_{k=0}^{\infty} \psi_k B^k) W_t = (1 + 0B + 0B^2 + \ldots) W_t. \end{aligned} \tag{3.9}$$

Let's find the sequence $(\psi_j)_{j\geq 0}$ and conditions on $(\varphi_j)_{j=1}^p$ for the above equation to hold. If we develop the product on the left hand side, and collect the terms by powers of B, we find $\psi_0 B^0$, $(\psi_1 - \varphi_1\psi_0)B^1$, $(\psi_2 - \varphi_1\psi_1 - \varphi_2\psi_0)B^2$, etc. Equating to the coefficients on the right hand side ($1B^0$, $0B^1$, $0B^2$, etc), we obtain a recursive solution for $(\psi_j)_{j\geq 0}$: first $\psi_0 = 1$, and then

$$\begin{aligned} \psi_1 - \varphi_1\psi_0 = 0 \quad &\Rightarrow \psi_1 = \varphi_1 \\ \psi_2 - \varphi_1\psi_1 - \varphi_2\psi_0 = 0 \quad &\Rightarrow \psi_2 = \varphi_1^2 + \varphi_2, \end{aligned}$$

etc. The general equation for $k \geq p$ is given by $\psi_k - \sum_{j=1}^{p} \psi_{k-j}\varphi_j = 0$; it is called a homogeneous linear difference equation of order p. We can always solve these equations for finitely many steps recursively as above, but the resulting sequence $(\psi_j)_{j\geq 0}$ might not converge, or it might not satisfy a technical condition such as $\sum_{j=0}^{\infty} |\psi_j| < \infty$ and so the sum in Eq. (3.8) might be infinite. It is not clear from the above reasoning what conditions on φ would lead to $\sum_{j=0}^{\infty} |\psi_j| < \infty$.

We next present another approach to finding $(\psi_j)_{j\geq 0}$, which leads to direct conditions on the coefficients of the autoregressive polynomial. We view the task of finding $(\psi_j)_{j\geq 0}$ as equivalent to finding the multiplicative inverse of the polynomial $\varphi(B)$, as seen from Eq. (3.9). Since φ is a polynomial of order p, we might remember from complex analysis that there are p solutions to the equation $\varphi(z) = 0$ for $z \in \mathbb{C}$, where $\mathbb{C}$ is the complex plane. Denote these solutions, called *roots* of the polynomial, by $\lambda_1, \ldots, \lambda_p$. We can write the polynomial as $\varphi(z) = \prod_{j=1}^{p} (1 - \lambda_j^{-1} z)$. If the

roots are all distinct, then we could write, for some $(a_j)_{j=1}^p$,

$$\frac{1}{\varphi(z)} = \sum_{j=1}^{p} \frac{a_j}{1 - \lambda_j^{-1} z},$$

it is called a *partial fraction decomposition*; this can be extended to the case where some roots have a multiplicity larger than one, but we focus on the case of distinct roots for simplicity. Now we see that if $|\lambda_j^{-1}| < 1$ for all $j \in \{1, \dots, p\}$, we could write $(1 - \lambda_j^{-1} z)^{-1} = \sum_{k=0}^{\infty} (\lambda_j^{-k}) z^k$ provided $|z| \leq 1$ as well, using the standard series expansion $1/(1-x) = 1 + x + x^2 + x^3 + \dots$. This would finally lead to

$$\frac{1}{\varphi(z)} = \sum_{j=1}^{p} a_j \sum_{k=0}^{\infty} (\lambda_j^{-k}) z^k = \sum_{k=0}^{\infty} (\sum_{j=1}^{p} a_j \lambda_j^{-k}) z^k.$$

This leads to the correspondence: $\psi_k = \sum_{j=1}^{p} a_j \lambda_j^{-k}$. We assumed $|\lambda_j^{-1}| < 1$ for all $j \in \{1, \dots, p\}$, so we can directly check that $\sum_{k \geq 0} |\psi_k|$ is finite.

To summarize, we impose that the roots $(\lambda_j)_{j=1}^p$ of the autoregressive polynomial φ are such that $|\lambda_j^{-1}| < 1$ for all $j \in \{1, \dots, p\}$, so that we can define an MA(∞) representation as in Eq. (3.8). When $|\lambda_j|^{-1} < 1$, or equivalently $|\lambda_j| > 1$, we say that λ_j lies *outside the unit circle*. The existence of the MA(∞) representation ensures that $(Y_t)_{t \in \mathbb{Z}}$ is stationary with zero mean and covariance function given by Eq. (3.2).

3.3 Moving average models

The second fundamental class of time series models is called MA(q) models (for *moving average* of order q). We say that (Y_t) is a MA(q) process if

$$Y_t = W_t + \theta_1 W_{t-1} + \dots + \theta_q W_{t-q}, \tag{3.10}$$

where $\theta_q \neq 0$. We can define the *moving average polynomial* $\theta(B) = 1 + \theta_1 B + \dots + \theta_q B^q$, and write the MA(q) model as

$$Y_t = \theta(B) W_t. \tag{3.11}$$

Using Eq. (3.2), we can directly obtain the autocovariance function of MA(q) processes.

For any choice of coefficients $\theta_1, \dots, \theta_q$, the MA(q) process is stationary. However, by analogy with the autoregressive case, we will impose that the polynomial $\theta(B)$ is invertible (i.e. its roots

are all outside the unit circle). This allows the so-called AR(∞) representation

$$\sum_{j=0}^{\infty} \pi_j Y_{t-j} = W_t, \tag{3.12}$$

where the sequence $(\pi_j)_{j\geq 0}$ is obtained by inverting the polynomial $\theta(B)$. Therefore, under conditions on the roots of their polynomials, AR(p) processes have an MA(∞) representation and MA(q) processes have an AR(∞) representations. It will prove convenient to have both MA(∞) and AR(∞) representations, for prediction purposes as explained in Section 3.5.

Remark 3.3. The reason why we impose invertibility of the MA lag polynomial can be easily understood in the MA(1) case;

$$Y_t = W_t + \theta_1 W_{t-1} = (1 + \theta_1 B) W_t.$$

The root of the lag polynomial is θ_1^{-1} and

$$\begin{aligned}
\mathbb{V}[Y_t] &= \left(1 + \theta_1^2\right) \sigma_W^2, \\
\mathbb{C}\text{ov}[Y_t, Y_{t-k}] &= \theta_1 \sigma_W^2 \text{ if } k = 1 \text{ and } 0 \text{ if } k > 1, \\
\mathbb{C}\text{orr}[Y_t, Y_{t-k}] &= \theta_1 / \left(1 + \theta_1^2\right).
\end{aligned}$$

Notice that there exists a white noise process (V_t) with variance $\sigma_V^2 = \left(1 + \theta_1^2\right) / \left(1 + \theta_1^{-2}\right) \sigma_W^2$ such that Y_t^* defined as

$$Y_t^* = V_t + \theta_1^{-1} V_{t-1}$$

has zero expectation,

$$\begin{aligned}
\mathbb{V}[Y_t^*] &= \left(1 + \theta_1^{-2}\right) \sigma_V^2 = \mathbb{V}[Y_t], \\
\mathbb{C}\text{orr}\left[Y_t^*, Y_{t-k}^*\right] &= \theta_1^{-1} / \left(1 + \theta_1^{-2}\right) = \theta_1 / \left(1 + \theta_1^2\right) = \mathbb{C}\text{orr}[Y_t, Y_{t-k}]
\end{aligned}$$

hence $\mathbb{C}\text{ov}\left[Y_t^*, Y_{t-k}^*\right] = \mathbb{C}\text{orr}\left[Y_t^*, Y_{t-k}^*\right] \mathbb{V}[Y_t^*] = \mathbb{C}\text{ov}[Y_t, Y_{t-k}]$. The two processes (Y_t) and (Y_t^*) have the same moments and autocovariance structure. It follows that when we observe an MA(1) process, we cannot tell whether the lag polynomial has a root that is below or above unity since there are two different processes that are observationally equivalent. Since we cannot distinguish them and since in most cases it does not matter, we choose by convention the process with an invertible MA lag polynomial.

3.4 ARMA(p,q) models

Definition. We finally combine both AR(p) and MA(q) components into ARMA(p,q) models, defined as processes satisfying

$$Y_t - \varphi_1 Y_{t-1} - \ldots - \varphi_p Y_{t-p} = W_t + \theta_1 W_{t-1} + \ldots + \theta_q W_{t-q},$$

or, equivalently

$$\varphi(B) Y_t = \theta(B) W_t.$$

The integers p and q are called the orders of the AR and MA components of the model. We put the following restrictions on the parameters, based on the previous discussions.

- $\varphi_p \neq 0$ and $\theta_q \neq 0$; otherwise the process would be ARMA($\tilde{p}$,$\tilde{q}$) with smaller orders $\tilde{p}$,$\tilde{q}$.
- The polynomials $\varphi(B)$ and $\theta(B)$ are both invertible (i.e. their roots are all outside the unit circle), so that we have both MA(∞) and AR(∞) representations, of the form given by Eqs. (3.8) and (3.12).
- The polynomials $\varphi(B)$ and $\theta(B)$ have no common roots.

The last condition ensures that we could not simplify the ARMA(p,q) model into a simpler ARMA(p-1,q-1) model, by removing a common factor from both polynomials. For instance, consider the white noise process $Y_t = W_t$ for all $t \in \mathbb{Z}$. It also satisfies the equation $Y_t - \alpha Y_{t-1} = W_t - \alpha W_{t-1}$ for all $\alpha \neq 0$, so it would be in the class of ARMA(1,1) models if we did not add that third condition.

Remark 3.4. In `R`, the coefficients $(\psi_j)_{j \geq 1}$ of the MA(∞) representation can be calculated from the AR and MA coefficients using the function `ARMAtoMA`. By swapping the arguments of the function, we can use it to obtain the coefficients $(\pi_j)_{j \geq 1}$ of the AR(∞) representation.

Identification and estimation. The question of "identification" refers to the selection of the orders p and q based on data. There are no magic recipes for this important task. Traditionally the advice was to stare at ACF and PACF plots. Indeed it turns out that, if a series is generated from a pure AR(p) model (i.e. without an MA part), then the PACF cuts off after lag p, while the ACF "tails off" (i.e. smoothly goes to zero). On the other hand, if a series is generated from a pure MA(q) model, the ACF cuts off after lag q, while the PACF tails off. Thus we can identify the orders of pure AR and MA processes from these visual tools. Unfortunately this does not provide guidelines when both ACF and PACF tail off smoothly, and when there are no clear cut-offs e.g. due to a limited sample size.

With modern computing power, it is possible to simultaneously consider many choices of p and q, and to select a model based on a generic "model selection" criterion, such as Akaike's Information Criterion, a Bayesian Information Criterion, etc. These criteria involve the value of the log-likelihood function at its maximizer, and penalize model complexity through terms that depend on the number of parameters, e.g. $p+q+1$ for ARMA(p,q) models with an unknown variance σ_W^2. For example, the corrected AIC (AICc in Eq. (9.3.4) of Brockwell & Davis' book) is

$$\text{AICc} = -2\log p(y_1,\dots,y_n;\hat{\varphi},\hat{\theta},\hat{\sigma}_W^2) + \frac{2kn}{n-k-1},$$

where $k = p+q+1$, $\hat{\varphi},\hat{\theta},\hat{\sigma}_W^2$ are the maximum likelihood estimators, and n is the length of the time series.

Note that it might be not be necessary to select one model: for forecasting purposes one might prefer to aggregate multiple models.

The likelihood plays a key role in parameter estimation for ARMA models, since the most common approach is maximum likelihood estimation. How to compute the likelihood is not straightforward (give it a try!) and we will defer this point to the discussion of Kalman filters in the next chapter; see also Section 3.8 below.

3.5 Forecasting with ARMA

We assume here that we have identified parameters that satisfy the conditions of Section 3.4, and that we want to use to predict the future values of a time series. We will rely on the AR(∞) and MA(∞) representations, given in Eqs. (3.8) and (3.12). Assume for a moment that we have an infinitely long time series, (Y_t) where $t = -\infty,\dots,0,1,\dots,n$. Suppose that we want to predict Y_{n+h} for some $h \geq 1$. We have the two representations

$$Y_{n+h} = \sum_{j=0}^{\infty} \psi_j W_{n+h-j} \quad \text{and} \quad W_{n+h} = \sum_{j=0}^{\infty} \pi_j Y_{n+h-j},$$

where $\pi_0 = \psi_0 = 1$. Under the squared loss it is sensible to predict Y_{n+h} using the conditional expectation of Y_{n+h} given the available data $Y_{-\infty},\dots,Y_n$. We introduce the notation $Y_{n+h|n}$ for the conditional expectation $\mathbb{E}[Y_{n+h}|Y_{-\infty},\dots,Y_n]$, for all h. Taking the conditional expectation on

both sides of the AR(∞) representation, we obtain for all $h \geq 1$,

$$\begin{aligned}\mathbb{E}\left[W_{n+h}|Y_{-\infty},\dots,Y_n\right] &= \sum_{j=0}^{\infty}\pi_j\mathbb{E}\left[Y_{n+h-j}|Y_{-\infty},\dots,Y_n\right]\\ \Leftrightarrow \quad 0 &= \sum_{j=0}^{h-1}\pi_j Y_{n+h-j|n} + \sum_{j=h}^{\infty}\pi_j Y_{n+h-j}.\end{aligned}$$

We used the fact that W_{n+h} is independent of all $Y_{-\infty},\dots,Y_n$, so that its conditional expectation is zero; and the conditional expectation of Y_{n+h-j} is equal to Y_{n+h-j} itself if $n+h-j \leq n$. Therefore, we obtain the formula

$$Y_{n+h|n} = -\sum_{j=1}^{h-1}\pi_j Y_{n+h-j|n} - \sum_{j=h}^{\infty}\pi_j Y_{n+h-j}, \tag{3.13}$$

which can be used to obtain $Y_{n+h|n}$ recursively, starting from $Y_{n+1|n} = -\sum_{j=0}^{\infty}\pi_{1+j}Y_{n-j}$ when $h=1$. To assess the error made by these predictions, we take the conditional expectation in both sides of the MA(∞) representation to obtain

$$\begin{aligned}Y_{n+h|n} &= \sum_{j=0}^{\infty}\psi_j\mathbb{E}\left[W_{n+h-j}|Y_{-\infty},\dots,Y_n\right],\\ &= \sum_{j=0}^{h-1}\psi_j\mathbb{E}\left[W_{n+h-j}|Y_{-\infty},\dots,Y_n\right] + \sum_{j=h}^{\infty}\psi_j\mathbb{E}\left[W_{n+h-j}|Y_{-\infty},\dots,Y_n\right].\end{aligned}$$

In the first sum, $n+h-j > n$, thus W_{n+h-j} is independent of all $Y_{-\infty},\dots,Y_n$, and thus its mean is zero; if $n+h-j \leq n$, W_{n+h-j} is entirely determined by $Y_{-\infty},\dots,Y_n$ (as can be seen from the AR(∞) equation), so that its conditional expectation is W_{n+h-j} itself. We can therefore write

$$Y_{n+h|n} = \sum_{j=h}^{\infty}\psi_j W_{n+h-j}.$$

Therefore, the error made by these predictions satisfies

$$\mathbb{E}\left[\left(Y_{n+h} - \mathbb{E}\left[Y_{n+h}|Y_{-\infty},\dots,Y_n\right]\right)^2\right] = \mathbb{E}\left[\left(\sum_{j=0}^{\infty}\psi_j W_{n+h-j} - \sum_{j=h}^{\infty}\psi_j W_{n+h-j}\right)^2\right] = \sigma_W^2\sum_{j=0}^{h-1}\psi_j^2.$$

Note that the above formula for the prediction error goes to a constant (in h) as $h \to \infty$. Note also that we have finitely many observations in practice, so that we need to truncate the infinite sum in Eq. (3.13).

3.6 Seasonal ARMA

An important feature of many time series is seasonality. For instance, monthly measurements might exhibit similar patterns every 12 time steps. This can be accounted by ARMA models using enough coefficients. For instance, we could systematically consider an AR(12) model for monthly data, an AR(52) for weekly observations, an AR(365) for daily data, etc. However it seems a bit unfortunate to introduce so many parameters.

An approach to capture seasonality without having too many parameters is provided by seasonal ARMA models. These models are obtained using AR and MA polynomial as in standard ARMA models, but applied to a backshift operator of order s, where s is the seasonality. For example, instead of an AR(12) model, we consider a seasonal AR(1) model with seasonality 12:

$$(1 - \Phi_1 B^{12})Y_t = Y_t - \Phi_1 Y_{t-12} = W_t,$$

where Φ_1 is a coefficient. If we denote by $\Phi(B^s)$ a seasonal autoregressive polynomial, and by $\Theta(B^s)$ a seasonal moving average polynomial, then a purely seasonal ARMA model is

$$\Phi(B^s)Y_t = \Theta(B^s)W_t.$$

It is exactly an ARMA model with the backshift operator B replaced B^s. We can formulate the same conditions on Φ and Θ as we did on φ and θ, to ensure that these polynomials are invertible.

In practice, we typically mix seasonal and non-seasonal components. For example, to capture dependencies at lag 1 and at lag 4, one can consider the model $(1 - \varphi_1 B)(1 - \Phi_1 B^4)Y_t = W_t$, with both an AR(1) and a seasonal AR(1). This can be written

$$Y_t = \varphi_1 Y_{t-1} + \Phi_1 Y_{t-4} - \varphi_1 \Phi_1 Y_{t-5} + W_t.$$

Therefore it is akin to an AR(5) model with some constraints on the coefficients: $\varphi_2 = \varphi_3 = 0$, and $\varphi_5 = -\varphi_1 \varphi_4$.

3.7 ARIMA

Differencing towards stationarity. Time series in raw form are not always adequately modeled as stationary processes. However, in surprisingly many cases, differencing the data leads to series that look fairly stationary. Recall that ∇Y_t is defined as $Y_t - Y_{t-1}$, i.e. $\nabla = 1 - B$, and $\nabla^k = (1-B)^k$ for all $k \geq 1$. Figure XX shows raw series and differenced series, in various cases where the differenced series appear to be stationary, in the sense of looking stable over time and having ACF and PACF decaying to zero.

For random walk processes, differencing brings stationary. Indeed consider a random walk defined as $X_t = \delta + X_{t-1} + W_t$ for all $t \geq 1$ and $X_0 = 0$, where W_t is a zero-mean stationary process. Then (X_t) itself is not stationary, but $Y_t = \nabla X_t = \delta + W_t$ is stationary. Similarly consider (X_t) defined as a linear trend plus noise: $X_t = a + bt + W_t$ for all $t \geq 1$. Then (X_t) is not stationary but $Y_t = \nabla X_t$ is stationary.

Thus one might hope that differencing brings stationarity, and if differencing once is not enough, we can keep on differencing. However, over-differencing can be problematic. Indeed if (X_t) is a stationary process, then (∇X_t) is still stationary, but with a different autocovariance function. It can be checked that if (X_t) is uncorrelated noise, then (∇X_t) has non-trivial autocorrelations, in fact corresponding to an MA(1) process; differencing more induces dependencies with a longer range. It can even generate troubling issues, for $\nabla W_t = W_t - W_{t-1}$ is an MA(1) whose lag polynomial has a root equal to -1, i.e. the MA lag polynomial is not invertible into an AR(∞) although it is stationary. This ∇W_t is a very specific process with odd properties that we will not cover here (estimators may not be Gaussian, even asymptotically).

Consider a process that is modeled as an ARMA(p, q),

$$\phi(B) Y_t = \theta(B) W_t.$$

If Y_t is stationary then the roots of $\phi(\cdot)$ are such that $\phi(z) = 0 \Rightarrow |z| > 1$ but some processes can be modeled as having "unit roots", i.e. $\phi(1) = 0$. In this case, the polynomial ϕ can be factored as

$$\phi(z) = \prod_{j=1}^{p} \left(1 - \frac{1}{\lambda_j} z\right)$$

where λ_j are the roots of $\phi(\cdot)$. Assume, for instance that $\lambda_1 = \lambda_2 = 1$ (a unit root with order of multiplicity equal to 2) then

$$\phi(z) = (1-z)(1-z) \prod_{j=3}^{p} \left(1 - \frac{1}{\lambda_j} z\right) = \left[\prod_{j=3}^{p} \left(1 - \frac{1}{\lambda_j} z\right)\right] (1-z)^2$$

and, letting $\varphi(z) = \left[\prod_{j=3}^{p} \left(1 - \frac{1}{\lambda_j} z\right)\right]$,

$$\phi(B) Y_t = \left[\prod_{j=3}^{p} \left(1 - \frac{1}{\lambda_j} B\right)\right] (1-B)^2 Y_t = \varphi(B) \nabla^2 Y_t = \theta(B) W_t,$$

where if the roots of $\varphi(\cdot)$ are all greater than one in modulus, $\nabla^2 Y_t$ follows a stationary ARMA$(p-2, q)$.

Here Y_t is nonstationary (the roots of ϕ are not in the stationary region) but taking the second difference of Y_t transforms the nonstationary process into a stationary one. This is the notion of *unit roots*, also called stochastic trends, i.e. a specific class of nonstationary processes whose first, second or higher order differences are stationary. We call Y_t *integrated of order* d if (i) it is nonstationary, (ii) $\nabla^k Y_t$ is nonstationary for $k < d$ and (iii) $\nabla^d Y_t$ is stationary (i.e. integrated of order zero). We denote it as $Y_t \sim \mathsf{I}(d)$.

We say that a process (Y_t) is autoregressive integrated moving average, written ARIMA(p,d,q) if $\nabla^d Y_t$ follows an ARMA(p,q) model, i.e.

$$\varphi(B)\nabla^d Y_t = \theta(B)W_t. \tag{3.14}$$

ARIMA models with $d \geq 1$ are non-stationary, but in a very specific way; they can't be expected to model adequately all non-stationary processes. One can try a few choices of d and compare predictive performance, or a "model selection criterion" such as AIC, in order to choose the appropriate order. Figure 3.1 presents the series of annual log Consumer Price Index, its first difference (inflation) and second difference for a series of countries over 1960-2020. We observe series that are integrated of different orders.

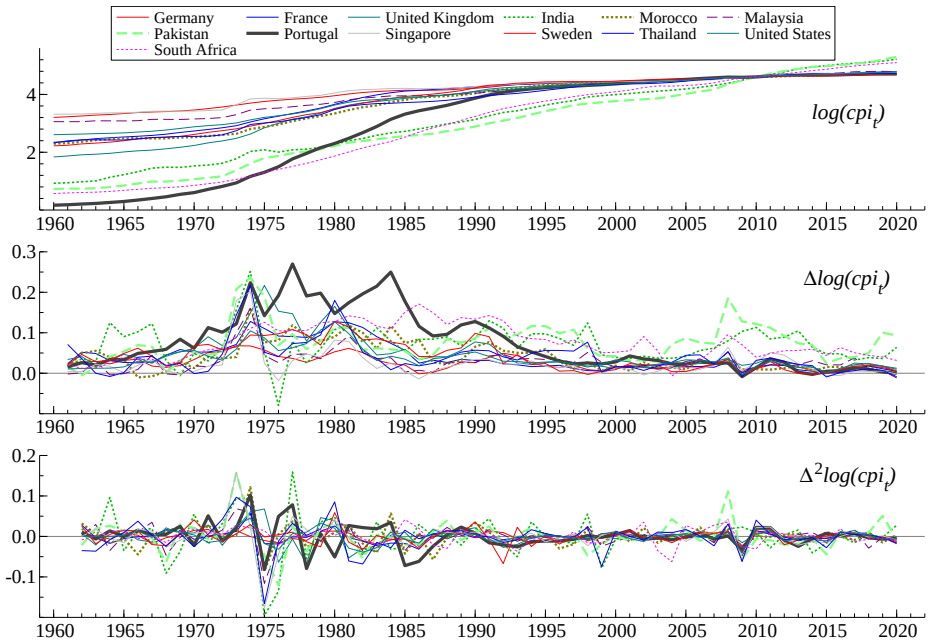

Figure 3.1: Logarithm of the annual Consumer Price Index and its first and second differences for a variety of countries. Source: World Bank.

Unit root tests: the Dickey-Fuller approach A question related to the order d in ARIMA models is testing for "unit roots" in the AR and MA polynomials of an ARMA model. Evidence of a root near one in the AR polynomial suggests that the series could be differenced before fitting an ARMA model. A root near one in the MA model might suggest that the series were over-differenced.

The main approach to testing for unit roots is called the (augmented) Dickey-Fuller test (from Said Dickey and Wayne Fuller who proposed it in 1979). Essentially, the principle is that if Y_t is non-stationary but ∇Y_t is stationary, the two cannot be linearly related in a regression such as

$$\nabla Y_t = \phi Y_{t-1} + X_t \tag{3.15}$$

where X_t is stationary. Indeed the distribution of Y_{t-1} changes over time but neither those of X_t or ∇Y_t do: it must be the case that in the regression above $\phi = 0$. As an example consider the AR(1) case: $Y_t = \rho Y_{t-1} + W_t$. Then

$$\nabla Y_t = (\rho - 1) Y_{t-1} + W_t$$

and we see that if $\rho = 1$ (Y_t is a random walk, i.e. nonstationary) then $\rho - 1 = 0$ and ∇Y_t and Y_{t-1} are not correlated. If $|\rho| < 1$, which corresponds to $Y_t \sim \mathsf{I}(0)$, then $\rho - 1 < 0$. The case $\rho - 1 > 0$ corresponds to an "explosive" root that is very rare in practice and can be ignored.

The principle of the Dickey–Fuller test is to regress ∇Y_t on Y_{t-1}

$$\nabla Y_t = \phi Y_{t-1} + W_t$$

and perform a one-sided t-test for the null hypothesis $\mathsf{H}_0 : \phi = 0$ vs the alternative $\mathsf{H}_1 : \phi < 0$. The *null hypothesis* of the Dickey–Fuller test is therefore that the process is nonstationary, and rejecting it leads to the alternative hypothesis that Y_t is stationary. Most unit-root tests conform with this null hypothesis, except the KPSS test (see below) for which the null is that of stationarity.

Because Y_{t-1} is nonstationary, the t-statistic has a nonstandard distribution and we cannot use the usual critical values or p-values, so we must resort to a specific distribution for this t-stat, the so-called **Dickey–Fuller distribution** which is tabulated in standard packages.

There are two additional issues related to unit-root testing:

1. The presence of deterministic terms has differing impacts under the null and alternative so this creates confusion and impacts inference. For instance, consider the regression $\nabla Y_t = \tau + \phi Y_{t-1} + W_t$, with $\tau > 0$. If $\phi \in (-2, 0)$, then $Y_t = \tau + (\phi + 1) Y_{t-1} + W_t$ is stationary around a nonzero mean $\mu = -\tau/\phi$. If $\phi = 0$ then $Y_1 = \tau + Y_0 + W_1$, $Y_2 = \tau + Y_1 + W_2 = 2\tau + Y_0 + W_1 + W_2$ and so on, $Y_t = \tau t + Y_0 + \sum_{j=1}^{t} W_j$ presents a linear deterministic trend τt. The fact that the

Table 1: Distribution of the F statistic for the test $(\tau, \delta, \phi) = (\tau, 0, 0)$ in $\nabla Y_t = \tau + \delta t + \phi Y_{t-1} + W_t$.

	Probability of a value less than							
Sample Size	.01	.025	.05	.10	.90	.95	.975	.99
25	.74	.90	1.08	1.33	5.91	7.24	8.65	10.61
20	.76	.93	1.11	1.37	5.61	6.73	7.81	9.31
100	.76	.94	1.12	1.38	5.47	6.49	7.44	8.73
250	.76	.94	1.13	1.39	5.39	6.34	7.25	8.43
500	.76	.94	1.13	1.39	5.36	6.30	7.20	8.34
∞	.77	.94	1.13	1.39	5.34	6.25	7.16	8.27

Source: Dickey & Fuller (1976), Table VI

role of τ depends on the value of ϕ generates difficulties in the test. The solution to the issue of constants and trends is either to follow a precise sequential testing algorithm (starting from the more general and verifying in turn each hypothesis precisely) or to resort to alternative tests. The latter solution is generally preferred since the Dickey–Fuller test does not perform necessarily well in finite samples. In the model

$$\begin{aligned} Y_t &= \tau + \delta t + \rho Y_{t-1} + W_t \\ \nabla Y_t &= \tau + \delta t + (\rho - 1) Y_{t-1} + W_t, \end{aligned}$$

depending on the values of (τ, δ, ρ) the following behaviors for Y_t result:

(τ, δ, ρ)	$\lvert\rho\rvert < 1$	$\lvert\rho\rvert = 1$
$\delta \neq 0$	stationary around a linear trend	integrated and exhibits a quadratic trend
$\tau \neq 0, \delta = 0$	stationary with a nonzero mean	integrated and exhibits a linear trend
$\tau = 0, \delta = 0$	stationary with zero mean	integrated without deterministic trend

It is often recommended to start with the more general model with nonzero (τ, δ). Test $\phi = \rho - 1 = 0$, then test for $\delta = 0$ using the F test for the joint hypothesis $(\delta, \phi) = (0, 0)$. This means computing the Fisher statistic

$$F = \frac{ESS_R - ESS_{UR}}{(N - k)\, q},$$

where N is the number of observations used in the regression ($T - 1$ since we use ΔY_t, not Y_t), k the number of estimated parameters in the unrestricted regression (i.e. estimating δ and ρ), q the number of restrictions (two here), ESS_R is the sum of squares of the modeled variables (i.e. $\sum_{i=1}^{T} \Delta \widehat{Y}_i^2$) under the null hypothesis $(\delta, \phi) = (0, 0)$ (i.e. excluding the corresponding

regressors from the regression) and ESS_{UR} the sum $\sum_{i=1}^{T} \Delta\widehat{Y}_i^2$ in the unrestricted regression. The critical values are given in Table 1. If the hypothesis is accepted then impose it and retest for $\phi = 0$ using the appropriate Dickey–Fuller distribution, and so on. Hopefully no contradiction appears.

2. In equation (3.15), we only specified that X_t is stationary, so it may not be white noise and could be generated by an ARMA(p,q). We must take into account this dependance or the estimator of α is impacted. The so-called **Augmented Dickey–Fuller** (ADF) consists in using the t-statistic obtained by fitting an AR(p) with the value of p chosen by AIC:

$$\begin{aligned} Y_t &= \tau + \delta t + \rho_1 Y_{t-1} + \ldots + \rho_p Y_{t-p} + W_t \Leftrightarrow \\ \Delta Y_t &= \tau + \delta t + \phi Y_{t-1} + \alpha_1 \Delta Y_{t-1} + \ldots + \alpha_{p-1} \Delta Y_{t-p+1} + W_t, \end{aligned} \tag{3.16}$$

i.e. $\rho(B) Y_t = \tau + \delta t + W_t$, with

$$\begin{aligned} \phi &= \rho_1 + \ldots + \rho_p - 1 = -\rho(1), \\ \alpha_i &= -(\rho_{i+1} + \ldots + \rho_p), \ i = 1, \ldots, p-1. \end{aligned}$$

A unit root value then implies a zero value of $\rho(1)$ which is identical to ϕ. Least squares regression can therefore directly be applied to obtain the unit root parameter ϕ. The standard t-statistic for testing for a zero value of ϕ can thus be used to test for a unit root in higher order AR models. The (asymptotic) distribution for the t-test for the null that $\phi = 0$ is the same as in the AR(1) model. Hence, we can use the same critical values as for the AR(1) model.

The ADF is a **parametric** test for the presence of a unit root since it relies on estimating the parameters of model for Y_t which is fitted to the data by specifying the lag order p and/or the presence of nonzero drift/linear trend. Expression (3.16) assumed W_t is white noise, this necessitates that p is not underestimated. The solution is to start from a large enough p so W_t passes all the diagnostic tests and then find p using the AIC.

Various tests for the presence of unit roots are available in `R` packages, such as the `tseries` package, and explained in most time series books. See also the lecture notes of Eric Zivot, available here.

Phillips-Perron (1988) test (PP) Peter Phillips and Pierre Perron suggested a *nonparametric* correction to the DF test that replaces the ADF. Here the autocorrelation of ΔY_t is not modeled. Instead a simple DF model ($p = 1$) is fitted to the data. The residuals are assumed to be

stationary (this is true if Y_t is either stationary or integrated of order 1) so they admit a Wold representation. The distribution of the t statistic for ϕ depends on the autocorrelation structure of the residuals. Phillips-Perron proposed a correction for t using the difference between the estimated sum of all covariances of the residuals (estimated *spectral density at frequency zero, or long-run variance*) and their estimated variance. This difference is zero if the residuals are *iid*.

The PP still needs to specify the deterministic terms in the regression. This test is robust to misspecification of the correlation structure of ΔY_t. As with nonparametric corrections, it is less precise when the ADF equation correctly represents the reality. Both PP and ADF are asymptotically equivalent.

Eliott-Rothenberg-Stock (1996) test (ERS) This test used the fact that in finite samples, values of ρ that are close to, yet strictly different from, unity are indistinguishable. Instead of differencing Y_t, ERS suggest quasi-differencing Y_t as $\widetilde{Y}_t = Y_t - \rho_T Y_{t-1}$, with $\rho_T = 1 - \frac{c}{T}$, and $c = 7$ or 13.5 respectively if only a constant, or a constant and a linear trend are estimated. The ERS statistic is

$$P_T = \frac{SSR\left(\rho_T\right) - \rho_T SSR\left(1\right)}{f_0}$$

where $SSR\left(x\right)$ is the sum of the squared residuals from the Dickey-Fuller equation using quasi-difference parameter x. f_0 is as usual computed under the null. This test is optimal against the point alternative ρ_T and is quite robust in general. This test can be combined with GLS detrending.

Kwiatkowski, Phillips, Schmidt, and Shin (1992) test (KPSS) This test is based on a *different null hypothesis:* that of stationarity (the others test for the presence of a unit root). Here the equation that is involved is

$$Y_t = X_t\beta + u_t$$

where X_t is an *exogenous* variable (a constant and a deterministic trend here, but it could be any other variable, you will see this in the context of cointegration testing). The LM statistic is then defined as

$$LM = \frac{\sum_t S_t^2}{T^2 f_0}$$

where $S_t = \sum_{i=1}^{t} \widehat{u}_t$ and f_0 is an estimate of the spectral density of u_t at frequency zero. Under the null of stationarity, KPSS tabulated the distribution of the LM test. Unfortunately this test is not very robust and tends to under reject stationarity.

Forecasting with ARIMA. Focusing on the case $d = 1$, suppose that we have $Z_t = \nabla Y_t$ and that we model Z_t through an ARMA(p,q) model. After having estimated the parameters, we can

perform predictions of Z_{n+h} for $h \geq 1$, as described in Section 3.5. Denote the predictions of future values Z_{n+h} given $Z_1, \dots, Z_n$ by $Z_{n+h|n}$. How can we obtain predictions for the original series (Y_t)?

We have by definition $Z_{n+1} = Y_{n+1} - Y_n$. If we predict Z_{n+1} with $Z_{n+1|n}$, then this leads to predicting Y_{n+1} with $Y_{n+1|n} = Y_n + Z_{n+1|n}$. Then, using the formula $Y_{n+2} = Y_{n+1} + Z_{n+2}$, we can predict Y_{n+2} with $Y_{n+2|} = Y_{n+1|n} + Z_{n+2|n}$. We can recursively compute predictions $Y_{n+h|n}$ from the forecast $Z_{n+h|n}$ for all $h \geq 1$.

What is the prediction error associated with these forecasts? The generic calculations are a bit tedious, but we can consider specific cases. First, consider the random walk with drift $X_t = \delta + X_{t-1} + W_t$ for all $t \geq 1$, with $X_0 = 0$. This can be seen as an ARIMA(0,1,0). The prediction of X_{n+1} using $X_1, \dots, X_n$ based on the conditional expectation yields

$$X_{n+1|n} = \mathbb{E}[X_{n+1}|X_1, \dots, X_n] = \delta + X_n,$$

and for all $h \geq 1$,

$$X_{n+h|n} = \mathbb{E}[X_{n+h}|X_1, \dots, X_n] = h\delta + X_n.$$

We consider the mean squared error $\mathbb{E}[(X_{n+h|n} - X_{n+h})^2]$. By writing X_{n+h} in terms of the past noise terms, as

$$X_{n+h} = h\delta + X_n + \sum_{j=n+1}^{n+h} W_j,$$

we can express the difference between X_{n+h} and its prediction as

$$X_{n+h} - X_{n+h|n} = \sum_{j=n+1}^{n+h} W_j.$$

We can now compute the expectation of the square of $X_{n+h} - X_{n+h|}$, and we find $h\sigma_W^2$. This error goes to infinity as $h \to \infty$. This is different from stationary ARMA models, for which the prediction error standard deviation converges to a constant as $h \to \infty$.

Next consider an ARIMA(0,1,1) model, which will draw a connection to exponential smoothing. We have $Y_t = Y_{t-1} + W_t - \theta_1 W_{t-1}$ for all $t \in \mathbb{Z}$. We assume $|\theta_1| < 1$, which allows to write $(1 - \theta_1 B)^{-1}$ as $\sum_{j=0}^{\infty} \theta_1^j B^j$, and we obtain

$$\left(\sum_{j=0}^{\infty} \theta_1^j B^j \right) (1 - B) Y_t = W_t$$

$$\text{or, equivalently, } \sum_{j=0}^{\infty} \theta_1^j B^j Y_t - \sum_{j=0}^{\infty} \theta_1^j B^{j+1} Y_t = W_t.$$

Note that the first sum is $Y_t + \sum_{j=1}^{\infty} \theta_1^j Y_{t-j}$, and the second sum is $\sum_{j=1}^{\infty} \theta_1^{j-1} Y_{t-j}$, so

$$Y_t = \sum_{j=1}^{\infty} (1-\theta_1)\theta_1^{j-1} Y_{t-j} + W_t.$$

From this expression, we can calculate the forecast of Y_{n+1} given $Y_{-\infty}, \ldots, Y_n$ using the conditional expectation:

$$\begin{aligned} Y_{n+1n} &= \mathbb{E}[Y_{n+1} | Y_{-\infty}, \ldots, Y_n] \\ &= \sum_{j=1}^{\infty} (1-\theta_1)\theta_1^{j-1} Y_{n+1-j} \end{aligned}$$

We recognize the method called "exponential smoothing" in Chapter 2! This time we see the method as a particular instance of ARIMA, thus in a probabilistic framework which suggests ways of estimating the parameter θ_1, and the prediction error can be defined meaningfully.

3.8 State space representation

We have not yet explained how to compute the likelihood associated with ARMA models. Here we describe how ARMA models can be represented as "state space models". In the next chapter we will see how to generically evaluate the likelihood of state space models using an algorithm called the Kalman filter.

Linear Gaussian state space models are of the form

$$\begin{aligned} X_t &= \Phi X_{t-1} + W_t, \\ Y_t &= A X_t + V_t, \end{aligned} \tag{3.17}$$

where W_t, V_t are independent multivariate Normal variables with mean zero and covariance matrices Σ_W and Σ_V respectively. The coefficients Φ and A are matrices. To be more precise, if X_t is of dimension d, Φ has to be $d \times d$, and W_t has to be d-dimensional. If Y_t is of dimension p, then A has to be of dimension $p \times d$ and V_t has to be of dimension p. The initial distribution of X_0 is given by a Normal distribution $\mathcal{N}(m_0, C_0)$, with mean m_0 and covariance matrix C_0. The parameters of the model are: $m_0, C_0, A, \Phi, \Sigma_W, \Sigma_V$.

We next show that ARMA models are linear Gaussian models as in Eq. (3.17). We will see other examples in the next chapter. Let us start with an AR(p) model: $Y_t = \varphi_1 Y_{t-1} + \ldots + \varphi_p Y_{t-p} + W_t$. This process can be represented as a vector-valued Markov chain with p components; consider the

state equation

$$X_t = \begin{pmatrix} Y_{t-p+1} \\ Y_{t-p+2} \\ \vdots \\ Y_{t-1} \\ Y_t \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \\ \varphi_p & \varphi_{p-1} & \varphi_{p-2} & \dots & \varphi_1 \end{pmatrix} \begin{pmatrix} Y_{t-p} \\ Y_{t-p+1} \\ \vdots \\ Y_{t-2} \\ Y_{t-1} \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} W_t.$$

Consider the equation row by row: the first $p-1$ rows mean "$Y_j = Y_j$" for $j \in \{t-p+1, \dots, t-1\}$, and the last row is the AR(p) model definition. We define the observation equation to be

$$Y_t = \begin{pmatrix} 0 & 0 & \dots & 0 & 1 \end{pmatrix} X_t + V_t,$$

where the noise term V_t is equal to zero. We have expressed an AR(p) model in the form of Eq. (3.17). Next consider an MA(q) model: $Y_t = W_t + \theta_1 W_{t-1} + \dots + \theta_q W_{t-q}$. We can define the observation equation as

$$Y_t = \begin{pmatrix} \theta_q & \theta_{q-1} & \dots & \theta_1 & 1 \end{pmatrix} X_t + V_t,$$

where again V_t is zero for all times t, and

$$X_t = \begin{pmatrix} W_{t-q} \\ W_{t-q+1} \\ \vdots \\ W_{t-1} \\ W_t \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \\ 0 & 0 & 0 & \dots & 0 \end{pmatrix} \begin{pmatrix} W_{t-q-1} \\ W_{t-q} \\ \vdots \\ W_{t-2} \\ W_{t-1} \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} W_t.$$

In this representation, the latent process X_t contains all the noise terms that are used in the definition of Y_t.

Now, let us consider an ARMA(p,q) model. Define $r = \max(p, q+1)$, and extend φ or θ with zeros. For instance, if $p = 2$ and $q = 2$, then $r = 3$, and define $\varphi_3 = 0$. Consider a latent process X_t made of r elements $(X_{t,1}, \dots, X_{t,r})$ such that

$$\begin{pmatrix} X_{t,1} \\ X_{t,2} \\ \vdots \\ X_{t,r-1} \\ X_{t,r} \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \\ \varphi_r & \varphi_{r-1} & \varphi_{r-2} & \dots & \varphi_1 \end{pmatrix} \begin{pmatrix} X_{t-1,1} \\ X_{t-1,2} \\ \vdots \\ X_{t-1,r-1} \\ X_{t-1,r} \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} W_t.$$

From the above $r-1$ first rows, we see that $X_{t,j} = X_{t-1,j+1}$ for $j \in \{1,\dots,r-1\}$. Therefore, $X_{t,r-1} = X_{t-1,r}$, $X_{t,r-2} = X_{t-1,r-1} = X_{t-2,r}$, etc. Then the last element $X_{t,r}$ satisfies $X_{t,r} = \varphi_1 X_{t-1,r} + \varphi_2 X_{t-1,r-1} \dots + \varphi_r X_{t-1,1} + W_t$ which can be rewritten $\varphi(B)X_{t,r} = W_t$. Define the observation equation as

$$Y_t = \begin{pmatrix} \theta_{r-1} & \theta_{r-2} & \dots & \theta_1 & 1 \end{pmatrix} X_t + V_t,$$

where V_t is zero for all times t. Then $Y_t = X_{t,r} + \theta_1 X_{t,r-1} + \dots + \theta_{r-1} X_{t,1}$, which can be rewritten $Y_t = \theta(B)X_{t,r}$. Therefore $\varphi(B)Y_t = \varphi(B)\theta(B)X_{t,r} = \theta(B)\varphi(B)X_{t,r} = \theta(B)W_t$, so that Y_t follows an ARMA(p,q) model.

3.9 Box–Jenkins methodology

We conclude this chapter by a quick overview of the so-called "Box–Jenkins" methodology, championed by George Box and Gwilym Jenkins in the early 1970s, and synonymous to "ARIMA modeling". That methodology is now considered classical, and somewhat superseded by the broader framework of state space models. However it is worth knowing about, arguably it is one of the first attempts at automatizing the process of time series forecasting in a statistically sound manner, and it has had many practical successes.

The idea superseding the Box-Jenkins methodology is that the Wold decomposition theorem implies that weakly stationary processes $\nabla^d Y_t$ admit an infinite MA(∞) representation:

$$\begin{aligned} \nabla^d Y_t &= \sum_{j=0}^{\infty} d_j \varepsilon_{t-j} + \eta_t = \left(\sum_{j=0}^{\infty} d_j B^j \right) \varepsilon_t + \eta_t, \\ &= d(B)\,\varepsilon_t + \eta_t \end{aligned}$$

where $d(B)$ denotes an infinite expansion. In practice, since we deal with finite samples of observations, there is no hope to estimate all the parameters d_j but we can make an approximation using a rational fraction of the sort

$$d(B) \approx \frac{\theta(B)}{\phi(B)} = \frac{1 + \theta_1 B + \dots \theta_q B^q}{1 - \varphi_1 B - \dots - \varphi_p B^p}$$

so, ignoring the "deterministic" η_t,

$$Y_t = \frac{\theta(B)}{\phi(B)} \varepsilon_t \quad \Leftrightarrow \quad \phi(B)\, Y_t = \theta(B)\, \varepsilon_t,$$

i.e., an ARMA(p,q) model is a good approximation to almost all weakly stationary processes. One of the only limitations is that the ACF of ARMA processes declines exponentially fast, i.e., at a rate

$\gamma(k) \sim \gamma_0 \alpha^k$ while some processes exhibit "long memory" that decays more slowly at a polynomial rate $\gamma(k) \sim \gamma_0 k^{-\lambda}$. The latter are typically called fractionally integrated processes and modeled as $\varphi(B)(1-B)^d Y_t = \theta(B) W_t$ with $d \in \mathbb{R}$ rather that $d \in \mathbb{N}$. They are labeled ARFIMA(p, d, q).

The algorithm is as follows:

1. The process starts with data preparation, transformation and visualization, in particular with the objective of selecting the orders p, d, q in the ARIMA model (3.14).
2. Once a model is selected, parameters are estimated e.g. by maximizing the likelihood using a numerical optimizer.
3. From the fitted values, the residuals $\hat{W}_t$ are constructed and one checks that they are compatible with the white noise assumption, otherwise the model can be changed and the process iterated.

The latter is part of "model checking", similar to the inspection of residuals in linear regressions (lack of heteroscedasticity, normality of the errors...). A popular model check in time series analysis is the Ljung–Box test on the residuals, which tests whether the autocorrelations are different from zero. The test statistic involves a weighted sum of sample autocorrelations computed from the residuals, up to a certain lag.