Model Identification and Data Analysis

Niccoló Didoni

February 2022

Contents

Part I

Models

Chapter 1

Basic concepts

The real world can be seen as a collection of systems that evolves and periodically produce an output.

1.1 Time series

A system S produces an output (also quantity) of interest $y(t)$ generated by some mechanism according to what happens in the real world. Each quantity of interest $y(t)$ is observable. A representation of a system S is shown in Figure 1.1.

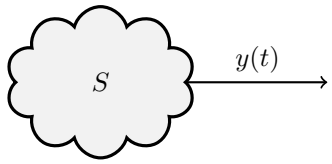


Figure 1.1: A system S and its output of interest $y(t)$.

A collection of consecutive values of $y(t)$ (i.e. for $t = 1, ..., n$) is a time series

$$\{y(1), y(2), ..., y(n)\}$$

It's important to underline that a time series denotes

- The mechanism that generates the output.
- The observations $y(1), ..., y(t)$.

Let us consider some examples to clear things up.

- An audio signal at a given frequency is a time series, in fact at every instant t we observe the output signal $y(t)$ of a system S (e.g. a speaker).
- Rain precipitation data collected every day is a time series, in fact every day we observe how much is raining.
- The data about the concentration of pollution in the air observed every day.
- The concentration of hormones in the blood monitored every hour.

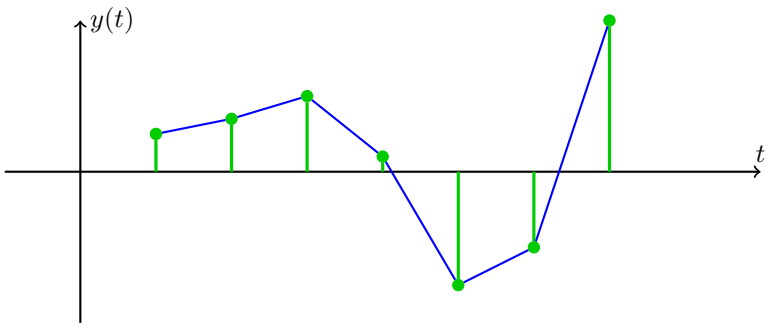


Figure 1.2: A graph representing the output $y(t)$ a system.

1.1.1 Digital signals

We will always consider systems whose data is a digital signal, thus the output is always observed at discrete intervals. Namely if we define T as the length of the interval between two observation, we can write a time series as

$$y(iT) : i \in \{0, ..., n\}$$

This means that the data can be plotted on a time-output plane as in Figure 1.2. In the graph

- The actual output is represented in red.
- The observations are represented in green.
- The observations are connected with blue lines. Notice that these lines do not represent the actual output but are used for clarity.

The value of the interval T depends on the type of system. If we give a look at the examples we did before, we notice that T is

- A number that depends on the frequency in the audio sampling example.
- A day in the rain and pollution example.
- A minute in the hormones example.

1.2 Input/Output systems

The output of many systems in the real world usually depends on some input $u(t)$, that is an observable source of variation. Such systems are called **Input/Output systems** (or I/O systems for short) and are represented as in Figure 1.3 As for output-only systems we can observe both inputs $(u(1), ..., u(n))$ and outputs $(y(1), ..., y(n))$. Here's some examples of I/O systems.

- An audio communication channel is an I/O system, in fact it receives an audio signal as input and sends such data to the receiver as output.
- A sensor that takes as input the rain data of a city and produces the pollution levels of the air is an I/O system because it generates an output (the pollution levels) given some input data (the rain data).

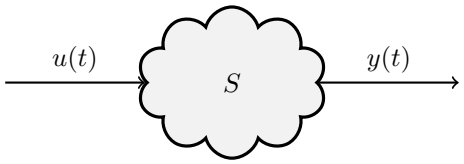


Figure 1.3: An I/O system S, its output $y(t)$ and input $u(t)$ of interest.

1.3 Models

Since we want to study the output $y(t)$ of a system, we need to build models for an I/O system and the time series it produces. The main problem is that in an I/O system, the output depends on many variable of the real world, thus multiple concurrent variations can modify the output data and the time series. Such variations can be unknown or too many to consider, thus we have to find some models that suitably describe I/O systems with its many variations.

1.3.1 Stochastic models

A good approach for describing I/O systems is stochastic models. This model uses a stochastic input $e(t)$ and the discrete, dynamic I/O system to describe the output data. A representation of a stochastic model is shown in Figure 1.4. The output stochastic model can also be seen as the sum of the outputs of two models

- A model that takes as input the stochastic input $e(t)$.
- A model that takes $u(t)$ as input.

Basically, what we are saying is that the output of a system can be seen as the overlap (i.e. superposition) of the effects of a stochastic behaviour and the actual behaviour of the system. A representation of such model is shown in Figure 1.5.

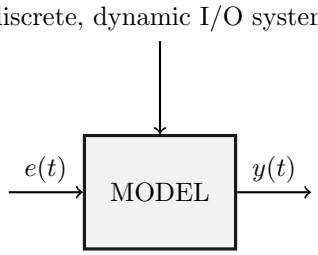


Figure 1.4: A stochastic model.

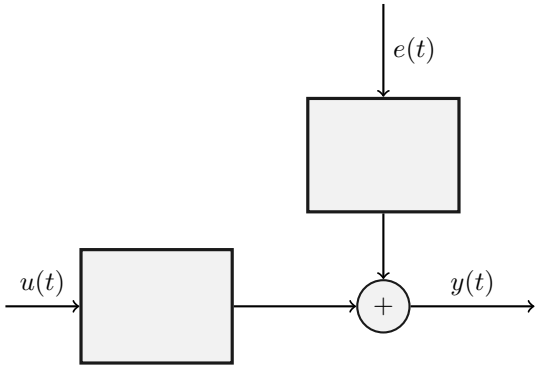


Figure 1.5: The output of an I/O system as the sum of the outputs of two models.

Chapter 2

Stochastic processes

2.1 Stochastic processes

The observations $y(1), \ldots, y(n)$ of an I/O system described with a stochastic model, are interpreted as realisations of a stochastic process. In particular,

> **Definition 1** (Stochastic process)**.** *A stochastic process is a countable infinite sequence of random variables all defined on the same probabilistic space (i.e. with the same law of distribution).*

From this definition we understand that we need a collection of random variables

$$v(1, s), v(2, s), \ldots, v(t, s)$$

where

- t (i.e. the first argument) is the **time index** and represents the time instant at which the variable has been measured. Notice that the time index can also be negative (usually to consider events in the past).
- s is the **outcome in a probability space** and considers that many different unknown sources of variation concur to produce the system's output. Namely s is a function that returns random values (with a specific distribution of probability) used to generate the values of $v(t, s)$.

Notice that sometimes, specially for long equations, we are going to drop the outcome s from the notation.

2.1.1 Realisations

A realisation of a stochastic process for a particular value of s (e.g. $s = \bar{s}$) is represented as $y(t, \bar{s})$. Basically we are defining a probability distribution $\bar{s}$ and we are using it to generate the values of the time sequence $y(1), \ldots, y(t)$. This means that when representing a time series in a I/O system described with a stochastic model, we have to draw all the possible realisations $y(t, \bar{s}_i)$ for each value $\bar{s}_i$ of s. A representation for two values of s is shown in Figure 2.1.

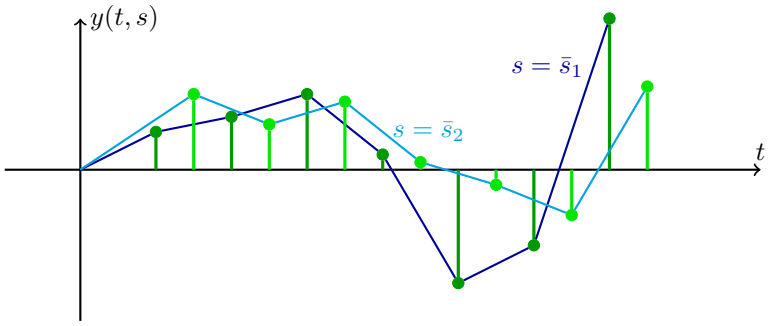


Figure 2.1: A graph representing the output $y(t,s)$ with two possible values of s.

2.2 Wide-sense descriptions

Say variables $v(t,s)$ are the random outcome of a system S and a stochastic process $\mathbb{P}$. These variables can be used to give a precise description of a stochastic process SP called **strong description**. The strong description of a stochastic process is the most complete way to describe a SP and allows to compute the probability value of v in a range $\mathbb{P}\{a < v(t,s) < b\}$. Strong descriptions are very hard to obtain in practice, thus we have to use a partial description called **wide-sense description**. A wide-sense description is based on partial **indicators** that can approximate the values of the probability distribution of the realisation of a stochastic process. Basically, we are representing some characteristics of the realisations instead of all the realisations.

2.2.1 Indicators

Mean

The first indicator we are going to use to describe a stochastic process is the mean $m_v(t)$ (i.e. the average or the expected value) of the realisations. The function $m_v(t)$ returns the sequence of means of $v(t,s)$ for every time instant t. Namely, for each t we consider all possible assignment $s = s_i$ compute do the average between the realisations $v(t,s_i)$. In other words, $m_v(t)$ is the sequence of values, for each t, of central values around which realisations take value. In formulas

$$m_v(t) = \mathbb{E}\big[v(t,s)\big] = \int_S v(t,s)\ \mathbb{P}\{ds\} \tag{2.1}$$

Function $m_v(t)$ is represented in Figure 2.2. The blue line represents the points around which the realisations take value.

Variance

Mean alone isn't enough to describe a stochastic process, in fact two processes may have a very different behaviour but same mean. Consider for instance the stochastic processes in Figure 2.3. The mean is the same for both processes but the realisations have a different amplitude. We need variance to differentiate processes in Figure 2.3.

Definition 2 (Variance)**.** *Variance $V_v(t)$ is the quantification of dispersion around the mean and is computed as the expected value of the squared difference between the realisations of $v(t,s)$ and*

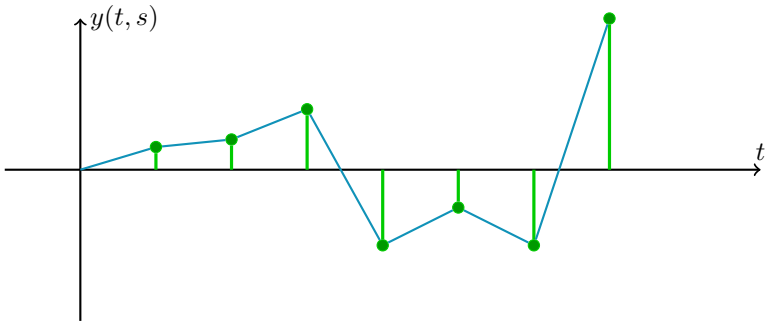
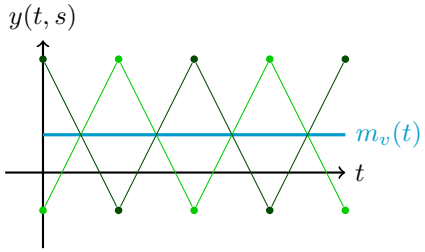
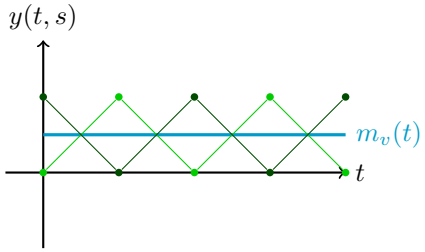


Figure 2.2: A graph representing the mean $m_v(t)$ of a stochastic process.



(a) Mean and realisation for system S_1.



(b) Mean and realisation for system S_2.

Figure 2.3: Two system with same mean but very different behaviours.

the mean $m_v(t)$.

$$V_v(t) = \mathbb{E}\Big[\big(v(t,s) - m_v(t)\big)^2\Big] = \int_S \big(v(t,s) - m_v(t)\big)^2 \, \mathbb{P}\{ds\} \tag{2.2}$$

Basically the variance averages, for each time instance t, the distances between the realisation for each $s = s_i$ and the mean $m_v(t)$.

Covariance

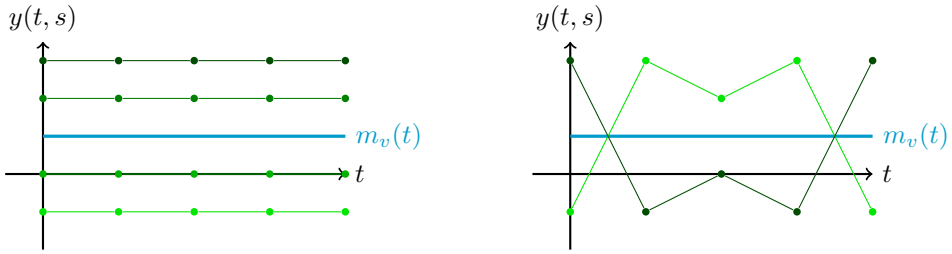
Even average isn't enough to sufficiently describe the realisations of a stochastic process. Consider the realisations in Figure 2.4. Both systems have the same mean and variance but are rather different. To distinguish these behaviours we have to introduce the covariance.

Definition 3 (Covariance)**.** *Covariance* $\gamma_v(t,\tau)$ *is the indication of* ***correlations*** *of random variables defining the stochastic process at different time instances.*

Covariance is computed as the expected value of the product between the deviations from the mean at time t and $t-\tau$. The deviation is the distance between v and m_v.

$$\gamma_v(t,\tau) = \mathbb{E}\Big[\big(v(t,s) - m_v(t)\big)\big(v(t-\tau,s) - m_v(t-\tau)\big)\Big] \tag{2.3}$$

$$= \int_S \big(v(t,s) - m_v(t)\big)\big(v(t-\tau,s) - m_v(t-\tau) \, \mathbb{P}\{ds\} \tag{2.4}$$



(a) Variance and realisation for system S_1. (b) Variance and realisation for system S_2.

Figure 2.4: Two system with same variance but very different behaviours.

The covariance can give us an hint on how a stochastic process behaves, in particular

- If $\gamma_v(t,\tau) > 0$ then the realisations of v have a tendency at being of the same of the deviation from t to $t-\tau$.
- If $\gamma_v(t,\tau) > 0$ then the realisations of v have a tendency at changing sign.

Alternative form The covariance can be expressed as

$$\gamma(t, t_2 = t-\tau) = \mathbb{E}\Big[\big(v(t,s) - m_v(t)\big)\big(v(t_2,s) - m_v(t_2)\big)\Big] \tag{2.5}$$

This form is equivalent to the former one since we have replaced $t-\tau$ with $t_2 = t-\tau$.

Variance and covariance The covariance can be used to obtain the covariance, in fact if we try to compute the covariance for $\tau = 0$ we obtain

$$\gamma(t,0) = \mathbb{E}\Big[\big(v(t,s) - m_v(t)\big)\big(v(t-\tau,s) - m_v(t-\tau)\big)\Big] \tag{2.6}$$

$$= \mathbb{E}\Big[\big(v(t,s) - m_v(t)\big)\big(v(t-0,s) - m_v(t-0)\big)\Big] \tag{2.7}$$

$$= \mathbb{E}\Big[\big(v(t,s) - m_v(t)\big)^2\Big] = V_v(t) \tag{2.8}$$

This means that the variance is an indicator of the dispersion around the mean at every time index t.

2.2.2 Weak characterisation

Thanks to mean, variance and covariance we can define a weak characterisation of a stochastic process.

Definition 4 (Weak characterisation)**.** *A weak characterisation of a stochastic process is given by*

- *Its* ***mean*** $m_v(t)$.
- *Its* ***covariance*** $\gamma_v(t,\tau)$.

The weak characterisation is just an approximation of the real behaviour of a stochastic process, but it's enough for our purposes.

Equivalence

All processes with the same mean $m_v(t)$ and covariance $\gamma_v(t,\tau)$ are equivalent.

2.3 Stationary stochastic processes

Among all stochastic processes, stationary stochastic processes are very important because of their properties.

Definition 5 (Stationary stochastic process)**.** *A stochastic process is said to be stationary if*

- *The mean $m_v(t)$ is constant for each time instant t.*

$$m_v(t) = m_v \ \forall t$$

- *The covariance $\gamma_v(t,\tau)$ doesn't depend on time but only on the time lag τ.*

$$\gamma(t,\tau) = \gamma(\tau) \ \forall t$$

This means that any two intervals of length τ have the same correlation (i.e. covariance).

In a nutshell, the indicators of a SSP are time invariant. From the definition of Stationary Stochastic Process SSP we can entail that the variance (i.e. the magnitude of dispersion around the mean) is constant

$$V_v(t) = \gamma(t,0) \tag{2.9}$$
$$= \gamma(0) \ \forall t \tag{2.10}$$

SSP are important because time invariance is typical of many situations and systems. Even non stationary processes can be written as the sum of a stationary process and a non stationary one.

$$v(t,s) = v_{stat}(t,s) + v_{non}(t,s)$$

Furthermore, SSP are easy to study.

Properties

Given the importance of stationary stochastic processes, it's useful to study their properties; in particular we are going to focus on the properties of the covariance $\gamma_v(t,\tau)$

- The covariance for $\tau = 0$ (i.e. the variance) is always non-negative. This property comes from the fact that $\big(v(t) - m_v\big)^2$ is always positive (because it's a square), thus the integral (i.e. the expected value) is also positive.

$$\gamma_v(t,0) = \mathbb{E}\big[\big(v(t) - m_v\big)^2\big] \geq 0$$

- The absolute value of the covariance for a generic τ is always smaller than the variance

$$|\gamma_v(t,\tau)| \leq \gamma_v(t,0)$$

- The covariance for negative values of τ is the same as the covariance for positive values of τ

$$\gamma_v(t, -\tau) = \gamma_v(t, \tau)$$

2.3.1 White noise

A very important stationary stochastic process is white noise. White noise is a sequence of uncorrelated random variables with the same mean and variance. More formally

Definition 6 (White noise)**.** *A stationary stochastic process $e(t, s)$ is called white noise if*

- *The mean is constant and equal to μ.*

$$\mathbb{E}\big[e(t, s)\big] = \mu \quad \forall t$$

- *The variance is constant and equal to λ^2.*

$$\mathbb{E}\Big[\big(e(t, s) - \mu\big)^2\Big] = \gamma_e(0) = \lambda^2 \quad \forall t$$

- *The covariance is null for each non-null time lag τ.*

$$\gamma_e(\tau) = \mathbb{E}\Big[\big(e(t, s) - \mu\big)\big(e(t - \tau, s) - \mu\big)\Big] = 0 \quad \forall \tau \neq 0$$

and we write

$$e(t, s) \sim WN(\mu, \lambda^2)$$

The third property of white noise is the one that characterise it (the others are directly inherited from the definition of stationary stochastic process); for this reason it's called **whiteness property**. In particular, the whiteness property states that there is no correlation between the realisations at different time instances (i.e. at time instances t and $t - \tau$ with $\tau \neq 0$) and highlight the unpredictability of the realisations. To sum things up, the covariance of white noise can be written as

$$\gamma_e(\tau) = \begin{cases} \lambda^2 & \tau = 0 \\ 0 & \tau \neq 0 \end{cases}$$

and can be represented as in Figure 2.5.

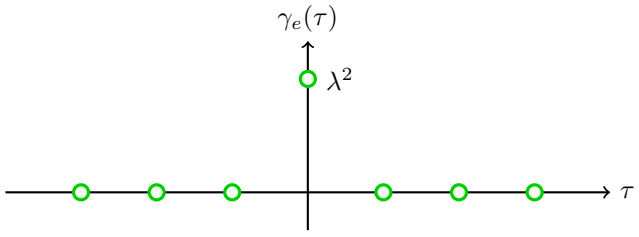


Figure 2.5: The covariance of white noise

2.3.2 Moving average process

Another important stationary stochastic process is the moving average $MA(n)$ where n is a parameter that represents the order of the process. A MA process is the linear combination of white noise. Namely, it takes as input a white noise $e(t)$ and linearly combines its previous values. Initially we are going to consider a zero-mean white noise $e(t,s) \sim WN(0,\lambda^2)$ (thus a zero-mean moving average). In formulas we can write

$$y(t,s) = c_0 \cdot e(t,s) + c_1 \cdot e(t-1,s) + c_2 \cdot e(t-2,s) + \cdots + c_n \cdot e(t-n,s) \tag{2.11}$$

$$= \sum_i^n c_i \cdot e(t-i,s) \tag{2.12}$$

where c_i are real parameters.

The order n of the moving average process defines the length of a windows that selects only the last n previous values of the input $e(t)$.

Stationary stochastic process

Let's demonstrate that a moving average process is a stationary stochastic process.

Constant mean Let's start by the first property. We have to check that the mean of $y(t,s)$ is constant in time t

$$m_y(t) = \mathbb{E}\big[y(t,s)\big] \tag{2.13}$$

$$= \mathbb{E}\big[\sum_i^n c_i \cdot e(t-i,s)\big] \tag{2.14}$$

The parameters c_i can be taken out of the expected value because they are constant (linearity of the expected value)

$$m_y(t) = \mathbb{E}\big[\sum_i^n c_i e(t-i,s)\big] \tag{2.15}$$

$$= \sum_i^n c_i \cdot \mathbb{E}\big[e(t-i,s)\big] \tag{2.16}$$

Since $e(t,s)$ is white noise, the expected value of e is μ (because $e \sim WN(\mu,\lambda^2)$), we obtain

$$m_y(t) = \mathbb{E}\big[\sum_i^n c_i e(t-i,s)\big] \tag{2.17}$$

$$= \sum_i^n c_i \cdot \mathbb{E}\big[e(t-i,s)\big] \tag{2.18}$$

$$= \sum_i^n c_i \cdot \mu \tag{2.19}$$

If we consider a zero-mean white noise, then the mean on the moving average process is 0. In both cases (zero-mean or non zero-mean) the mean $m_y(t)$ is constant and time independent (because it's the linear combination of a constant μ).

Covariance for zero time lag To demonstrate that the covariance doesn't depend on time t but only on the time lag τ, we can start by writing down the definition of covariance for a moving average process. To make things more concise we are going to drop the s from the notation of variables y (i.e. $y(t,s)$ becomes $y(t)$)

$$\gamma_y(t,\tau) = \mathbb{E}\Big[\big(y(t) - m_y(t)\big) \cdot \big(y(t-\tau) - m_y(t-\tau)\big)\Big] \tag{2.20}$$

Since we have already proven that $m_y(t)$ is 0 for zero-mean moving average processes (i.e. MA's with zero-mean white noise) we can rewrite the covariance as

$$\gamma_y(t,\tau) = \mathbb{E}\Big[\big(y(t) - 0\big) \cdot \big(y(t-\tau) - 0\big)\Big] \tag{2.21}$$

$$= \mathbb{E}\Big[y(t) \cdot y(t-\tau)\Big] \tag{2.22}$$

To prove that $\gamma_y(t,\tau)$ doesn't depend on time we can start from $\tau = 0$

$$\gamma_y(t,\tau=0) = \mathbb{E}\Big[y(t) \cdot y(t-0)\Big] \tag{2.23}$$

$$= \mathbb{E}\Big[y(t) \cdot y(t)\Big] \tag{2.24}$$

$$= \mathbb{E}\Big[y(t)^2\Big] \tag{2.25}$$

$$= \mathbb{E}\Big[\big(\sum_i^n c_i \cdot e(t-i)\big)^2\Big] \tag{2.26}$$

At this point we can expand the square to obtain

$$\gamma_y(t,0) = \mathbb{E}\Big[\sum_i^n c_i^2 \cdot e(t-i)^2 + \sum_{(i,j)} 2c_i \cdot c_j \cdot e(t-i) \cdot e(t-j)\Big] \tag{2.27}$$

Now we can use the linearity of the expected value to obtain

$$\gamma_y(t,0) = \sum_i^n c_i^2 \cdot \mathbb{E}\Big[e(t-i)^2\Big] + \sum_{(i,j)} 2c_i \cdot c_j \cdot \mathbb{E}\Big[e(t-i) \cdot e(t-j)\Big] \tag{2.28}$$

We are almost done, let's analyse the expected values $\mathbb{E}\big[e(t-i)^2\big]$. These are the definitions of covariance of white noise $e(t-i)$, in fact

$$\gamma_e(t,\tau=0) = \mathbb{E}\big[(e(t) - m_e(t))(e(t-\tau) - m_e(t-\tau))\big] \tag{2.29}$$

$$= \mathbb{E}\big[(e(t) - m_e(t))(e(t) - m_e(t))\big] \tag{2.30}$$

$$= \mathbb{E}\big[(e(t) - m_e(t))^2\big] \tag{2.31}$$

$$= \mathbb{E}\big[(e(t) - 0)^2\big] \quad (e(t) \sim WN(0,\lambda^2)) \tag{2.32}$$

$$= \mathbb{E}\big[e(t)^2\big] \tag{2.33}$$

The covariance of e for $\tau = 0$ is the variance, thus $\mathbb{E}[e(t)^2] = \lambda^2$ for each t because e is a stationary process (thus the covariance is independent from time for Definition 6 of white noise). Finally we can analyse the expected value of the cross products $\mathbb{E}\big[e(t-i)e(t-j)\big]$. Every term is a product between e at different time instants. For the whiteness property, realisations at different time instances are uncorrelated, thus $\mathbb{E}\big[e(t-i)e(t-j)\big]$ is zero. To sum things up, we can write the covariance for $\tau = 0$ as

$$\gamma_e(t, \tau = 0) = \sum_i^n c_i^2 \cdot \lambda^2 + \sum_{(i,j)} 2c_i \cdot c_j \cdot 0 \tag{2.34}$$

thus $\gamma_y(t, \tau = 0)$ constant and time independent.

Covariance for 1 time lag Proving that $\gamma_y(t, \tau = 0)$ is constant isn't enough to demonstrate that y is stationary. In particular we should prove that it is time independent for every time lag τ. Let's continue our demonstration proving that y is time independent for $\tau = 1$. Let's start by writing the covariance for $\tau = 1$ and remembering that we are dealing with an input white noise (i.e. $e(t) \sim WN(0, \lambda^2)$ with mean 0 and variance λ^2.

$$\gamma_e(t, 1) = \mathbb{E}\Big[\big(y(t) - m_y(t)\big) \cdot \big(y(t-1) - m_y(t-1)\big)\Big] \tag{2.35}$$

$$= \mathbb{E}\Big[y(t) \cdot y(t-1)\Big] \tag{2.36}$$

In the second equation we have used the fact that y has null mean for zero-mean inputs. Now we can expand $y(t)$ and $y(t-1)$

$$\gamma_e(t, 1) = \mathbb{E}\Big[\big(\sum_i^n c_i e(t-i)\big) \cdot \big(\sum_i^n c_i e(t-1-i)\big)\Big] \tag{2.37}$$

If we compute the product between the two factors we obtain some terms that refer to the same time instant (i.e. $c_i c_{i+1} e(t-i-1)^2$) and other that consider different time instants. The latter terms are always null, in fact e is a white noise and by definition realisations at different time instants are uncorrelated (whiteness property). The covariance becomes

$$\gamma_e(t, 1) = \sum_i^n c_i c_{i+1} \cdot \mathbb{E}\big[e(t-i-1)\big] \tag{2.38}$$

The remaining terms contain the variance of the white noise $V_e(t) = \gamma_e(t, 0) = \mathbb{E}[e(t)^2]$ that we have already demonstrated is time-independent and equal to λ^2. To wrap things up, we can write the covariance as

$$\gamma_e(t, 1) = \sum_i^n c_i c_{i+1} \cdot \lambda^2 \tag{2.39}$$

Clearly, $\gamma_e(t, 1)$ is constant and time independent. If we repeat the same process for $\tau = 2, ..., n$ we obtain the same results (the covariance hasn't the same value for each τ but it's constant for a certain τ), thus we can conclude that a moving average process is stationary. Notice that the covariance for $\tau > n$ is always null, in fact every realisation happens at different time instances (i.e. we don't have a term $e(t_1)e(t_2)$), thus for the white property of e, such realisations are uncorrelated.

General formula of the covariance Demonstrating that γ is time independent allowed us to write a general definition for the covariance of a moving average process.

$$\gamma_y(t,\tau) = \begin{cases} (c_0^2 + c_1^2 + \cdots + c_n^2)\lambda^2 & \tau = 0 \\ (c_0c_1 + c_1c_2 + \cdots + c_{n-1}c_n)\lambda^2 & \tau = \pm 1 \\ \vdots & \\ c_0c_n\lambda^2 & \tau = \pm n \\ 0 & |\tau| > n \end{cases} \tag{2.40}$$

This result (i.e., a MA process is stationary) has been obtained thanks to the stationarity of the white noise. In other words, the stationarity of white noise has induced the stationarity of the moving average process.

2.3.3 Moving average processes of order infinite

Among all moving average processes, $MA(\infty)$ have great relevance, in fact all stationary processes of interests are stationary processes of order infinite. Basically, the class of moving averages of order infinite cover almost all the class of SSP.

As for the general case, we will start from a zero-mean white noise $e(t) \sim WN(0,\lambda)$ (the generalisation to non zero-mean is easy), hence the moving average is

$$MA(\infty) = \sum_{i=0}^{+\infty} c_i e(t-i) \tag{2.41}$$

It's important to highlight that $MA(\infty)$ is a series of random variables, not numbers, because $e(t)$ depends on s (remember that $e(t)$ is actually $e(t,s)$). We have to put particular attention on this because the series is the limit

$$\lim_{n\to\infty} \sum_{i=0}^{+n} c_i e(t-i)$$

of random variables (i.e. of functions), thus we have to properly define the concept of limit. Also notice that convergence to a proper random variable is not always guaranteed. In particular a moving average process $MA(\infty)$ is well-defined (i.e. it converges) for every time instance t if and only if

$$\sum_{i=0}^{+\infty} c_i^2 < \infty \tag{2.42}$$

A well-defined moving average process $MA(\infty)$ is stationary. The process for proving that $MA(\infty)$ is stationary is the same as the one used for $MA(n)$ and relies on property 2.42 to apply the linearity of the expected value (the expected value is an integral and the series is a limit, thus we need the convergence property to swap integral and limit). From the demonstration of stationarity we obtain that the covariance of $MA(\infty)$ is

$$\gamma_y(t,\tau) = \left(\sum_{j=0}^{\infty} c_j c_{j+\tau} \right) \cdot \lambda^2 \tag{2.43}$$

As expected γ_y doesn't depend on time.

Complexity of moving average processes of infinite order

Moving average processes of order infinite cover almost all stationary stochastic processes. The only processes that can't described with an MA are completely predictable processes, that are of little interest because by observing y at one time instant we can predict y at every successive time instant.

2.3.4 Auto Regressive stochastic processes

The generality of MA processes of order infinite comes at a cost, in fact, we have to define an infinite number of parameters and work with series (i.e. we have to study the convergence of the series). At the same time $MA(n)$ is too limited, thus we have to introduce some processes that are a trade off between these two extremes. Auto Regressive (AR) stochastic processes allow to describe a large number of interesting stationary stochastic processes fed with white noise (or in general with a SSP) $e(t)$.

Let us consider a system that takes as input a zero-mean white noise $e(t) \sim WN(0, \lambda^2)$ (the same will be true for non zero-mean white noises and the generalisation is easy).

Definition 7 (Auto Regressive process)**.** *A process $y(t)$ is an auto regressive stochastic process if*

- *It's **stationary**.*
- *It satisfies the **recursive equation***

$$y(t) = a_1 y(t-1) + a_2 y(t-2) + \cdots + a_m y(t-m) + e(t) = \sum_{i=1}^{m} a_i y(t-i) + e(t) \tag{2.44}$$

where

  - *m is the **order** of the process.*
  - *$a_1, \ldots, a_m$ are the **coefficients** of the process.*

and we write

$$y(t) \sim AR(m)$$

Let's analyse a little more in detail the recursive equation. It can be divided in two main components

- $\sum_{i=1}^{m} a_i y(t-i)$ is the regression over the past values of the process.
- $e(t)$ is the term that considers the new input (i.e. the innovation).

Steady state solution

Now that we have defined what an AR process is, we should find a solution $y(t)$ of the recursive equation. In particular Equation 2.44 has multiple solutions, but we have to find the one for which $y(t)$ is stationary. Such solution is called **steady state solution**.

Definition 8 (Steady state solution)**.** *The steady state solution $y(t)$ of an $AR(n)$ process is the solution that satisfies the recursive equation $y(t) = \sum_{i=1}^{m} a_i y(t-i) + e(t)$ and for which the output $y(t)$ is stationary.*

To obtain the steady state solution of an Auto Regressive process, given its recursive Equation 2.44, we have to

1. Set the values of y at time t_0 to 0. This means creating a vector $Y(t_0)$ of all zeros.
$$Y(t_o) = \big[y(t_0 - 1), y(t_0 - 2), \ldots, y(t_0 - m)\big] = \underline{0}$$
2. Compute $y_{t_0}(t)$ expanding the recursive terms and using $Y(t_0)$ to stop recursion.
3. Compute the limit of $y_{t_0}(t)$ (as obtained at the previous point) for $t_0 \to -\infty$
$$y(t) = \lim_{t_0 \to -\infty} y_{t_0}(t)$$

The process $y(t)$ obtained satisfies Equation 2.44 and is a steady state solution. Furthermore, $y(t)$ is also a moving average process of order infinite, in fact it's a linear combination of the previous values of the input $e(t)$.

Well defined auto regressive processes

After finding a solution $y(t)$ to the recursive equation and showing that it's a steady state solution, we should also try to prove that $y(t)$ is a well-defined moving average process of order infinite. This can be achieved remembering that the sum of the squared coefficients c_i should be smaller than infinite
$$\sum_i c_i^2 < +\infty$$
Notice that the coefficients of $y(t)$ are functions of the parameters a_i of Equation 2.44 (i.e. $c_i = f(a_1, \ldots, a_m)$).

2.3.5 Auto Regressive Moving Average processes

Auto Regressive Moving Average (ARMA) processes add a moving average to an AR process. More specifically,

Definition 9 (Auto Regressive Moving Average)**.** *A process $y(t)$ is an Auto Regressive Moving Average process if it satisfies the recursive equation*
$$y(t) = \sum_{i=1}^{m} a_i y(t-i) + \sum_{i=0}^{n} c_i e(t-i) \tag{2.45}$$
and we write
$$y(t) \sim ARMA(m, n)$$

Steady state solution

As for AR processes, $y(t)$ is the steady state solution, obtained with the same process of AR, to the recursive equation. ARMA processes are moving average processes of order infinite, but in this case the coefficients are complicated functions of a_i and c_i. This makes hard proving that an ARMA process is well-defined, in fact we have to demonstrate
$$\sum_i^{+\infty} f(a_1, \ldots, a_m, c_1, \ldots, c_n)^2 < +\infty \tag{2.46}$$

2.4 Operational representation of an ARMA process

To demonstrate that an ARMA process is well-defined we can write the process using a different notation. In particular we have to introduce two operators

- The **backward shift operator** z^{-1}. This operator allows to write a signal $y(t-1,s)$ as $z^{-1}[y(t,s)]$
$$z^{-1}[y(t,s)] = y(t-1,s)$$
- The **forward shift operator** z^1. As for the backward shift operator we can write
$$z^1[y(t,s)] = y(t+1,s)$$

The z operators are linear, hence we can write

$$z^{-1}\big(ax(t) + by(t)\big) = z^{-1}\big(ax(t)\big) + z^{-1}\big(by(t)\big)$$

where x and y are stationary stochastic processes. The z operators can also be recursively applied to obtain

$$\begin{aligned} z^{-1}\big(z^{-1}\big(z^{-1}\big(y(t)\big)\big)\big) &= z^{-1}\big(z^{-1}\big(y(t-1)\big)\big) \\ &= z^{-1}\big(y(t-2)\big) \\ &= y(t-3) \\ &= z^{-3}\big(y(t)\big) \end{aligned}$$

The operators z and z^{-1} can be linearly combined to obtain arbitrary complex equations.

2.4.1 Transfer function

An ARMA's equation can be rewritten thanks to the z operator. In particular we can write the general form of an ARMA process

$$y(t) = \sum_{i=1}^{m} a_i y(t-1) + \sum_{i=0}^{n} c_i e(t-i)$$

as

$$y(t) = \sum_{i=1}^{m} a_i z^{-i} y(t) + \sum_{i=0}^{n} c_i z^{-i} e(t)$$

Now we can collect $y(t)$ and bring all the terms that contain $y(t)$ on the left of the equation to obtain

$$y(t) - \left(\sum_{i=1}^{m} a_i z^{-i}\right) y(t) = \sum_{i=0}^{n} c_i z^{-i} e(t)$$

$$\left(1 - \left(\sum_{i=1}^{m} a_i z^{-i}\right)\right) y(t) = \sum_{i=0}^{n} c_i z^{-i} e(t)$$

Finally we can isolate $y(t)$ on the left side and obtain

$$y(t) = \frac{\sum_{i=0}^{n} c_i z^{-i}}{1 - \left(\sum_{i=1}^{m} a_i z^{-i}\right)} e(t) \tag{2.47}$$

From equation 2.47 we notice that the steady state solution of an ARMA process defines an operator that takes a white noise $e(t)$ as input and returns a moving average $MA(\infty)$ as output, using the recursive equation. This operator (i.e. the function that maps $e(t)$ in $y(t)$) is called **transfer function** or **digital filter**

$$W(z) = \frac{\sum_{i=0}^{n} c_i z^{-i}}{1 - \left(\sum_{i=1}^{m} a_i z^{-i}\right)}$$

Formally a transfer function $W(z)$ is a function of the polynomials of the operations in the process's recursive function. In other words the transfer function $W(z)$ returns the steady-state solution of the recursive equation associated to the ARMA process. A transfer function $W(z)$ can also be written as

$$W(z) = \frac{C(z)}{A(z)}$$

where

- $C(z) = c_0 + \sum_{i=1}^{n} c_i z^{-1}$
- $A(z) = 1 - \left(\sum_{i=1}^{m} a_i z^{-1}\right)$

$W(z)$ can be graphically represented as in Figure 2.6.

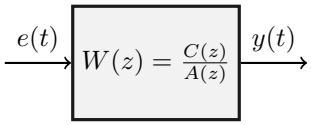


Figure 2.6: A transfer function $W(z)$.

2.4.2 Extension to general input SSPs

In general, the steady state solution of a ARMA process that takes as input a general stationary stochastic process $v(t)$ can be written as

$$y(t) = \frac{C(z)}{A(z)} v(t) = W(z)v(t) \tag{2.48}$$

In other words $y(t)$ is the steady state solution of the recursive equation

$$A(z)y(t) = C(z)v(t)$$

2.4.3 Composition of transfer functions

Transfer functions can be combined

- In **parallel**.
- In **series**.

Series of transfer functions

Let us consider a system as in Figure 2.7 where two systems S_1 and S_2 are in parallel, i.e., the output $x(t)$ of S_1 is used as input of S_2. The output of S_1 can be written as

$$x(t) = W_1(z)u(t)$$

and the output of S_2 can be written as

$$y(t) = W_2(z)x(t)$$

If we replace the value of $x(t)$ from the first equation we obtain

$$y(t) = W_2(z)W_1(z)u(t) = \big[W_2(z)W_1(z)\big]u(t)$$

The result we have just obtained is summed up in the following theorem.

Theorem 1 (Series of systems)**.** *Given the output of the series of two systems S_1 and S_2*

$$\begin{aligned} y(t) &= \big[W_1(z)W_2(z)\big]u(t) \\ &= \left[\frac{C_1(z)C_2(z)}{A_1(z)A_2(z)}\right]u(t) \end{aligned}$$

$y(t)$ is the steady state output of W_1W_2 fed by the stochastic process $u(t)$, thus it's the solution of the recursive equation

$$\big[A_1(z)A_2(z)\big]y(t) = \big[C_1(z)C_2(z)\big]u(t)$$

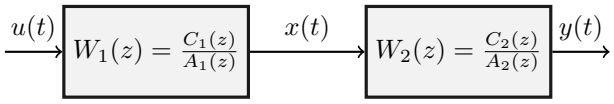


Figure 2.7: A series of systems.

Parallel of transfer functions

Let us consider a system as in Figure 2.8 where systems S_1 and S_2 are fed with the same input signal $u(t)$ and their output signal is summed to obtain $y(t)$. The equations of the outputs of systems S_1 and S_2 are

$$y_1(t) = W_1(z)u(t)$$

and

$$j_2(t) = W_2(z)u(t)$$

These two signals can be summed to obtain the output $y(t)$

$$\begin{aligned} y(t) &= y_1(t) + j_2(t) \\ &= W_1(z)u(t) + W_2(z)u(t) \\ &= \big[W_1(z) + W_2(z)\big]u(t) \end{aligned}$$

The result we have just obtained is summed up by the following theorem.

Theorem 2 (Parallel of transfer functions)**.** *Given the output of a parallel of systems,*

$$\begin{aligned} y(t) &= \big[W_1(z) + W_2(z)\big]u(t) \\ &= \left[\frac{C_1(z)A_2(z) + C_2(z)A_1(z)}{A_1(z)A_2(z)}\right]u(t) \end{aligned}$$

$y(t)$ is the steady state output of the recursive equation associated to the formal sum $W_1(z)+W_2(z)$ when fed with $u(t)$ and $y(t)$ is the solution of the recursive equation

$$\big[A_1(z)A_2(z)\big]y(t) = \big[A_2(z)C_1(z) + A_1(z)C_2(z)\big]u(t)$$

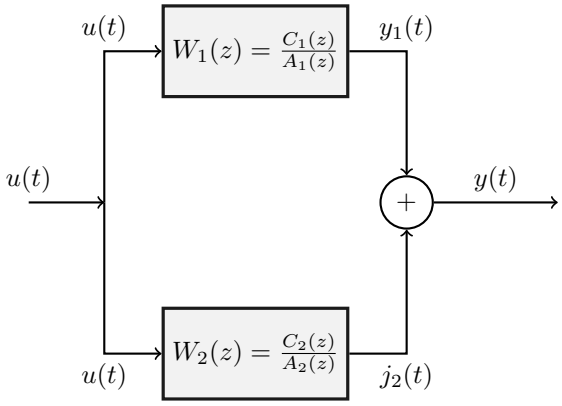


Figure 2.8: A parallel of two systems.

2.4.4 Changing sign of the polynomial shift operator

Given a transfer function

$$y(t) = W(z)u(t)$$

where $W(z)$ is a rational function of polynomial shift operators z, it's always possible to go from positive power operators to negative ones (and vice versa). Practically, if we expand the transfer function $W(z)$ we obtain

$$y(t) = \left[\frac{\sum_{i=0}^{n} c_i z^{-i}}{\sum_{i=0}^{m} a_i z^{-i}}\right]u(t)$$

Now we can take the maximum between n and m (say m for the sake of example) and multiply numerator and denominator for z^m to obtain

$$\begin{aligned} y(t) &= \left[\frac{\sum_{i=0}^{n} c_i z^{-i}}{\sum_{i=0}^{m} a_i z^{-i}}\frac{z^m}{z^m}\right]u(t) \\ &= \left[\frac{\sum_{i=0}^{n} c_i z^{-i+m}}{\sum_{i=0}^{m} a_i z^{-i+m}}\right]u(t) \end{aligned}$$

Notice that, multiplying $W(z)$ for $\frac{z^m}{z^m}$ is like putting in series two systems with transfer functions $W(z)$ and $\frac{z^m}{z^m}$.

2.4.5 Poles and zeros

Let us consider a transfer function $W(z) = \frac{C(z)}{A(z)}$. We can interpret it as a function of variables z, where z is a complex value variable (even if we know it actually is an operator). Thanks to this description we can import the concept of poles and zeros to the digital filter $W(z)$, in particular

- The **poles** of $W(z)$ are all the values of $z \in \mathbb{C}$ such that $W^{-1}(z) = 0$. Put it differently, the poles of $W(z)$ are the values of z that make the denominator $A(z)$ equal to 0 (i.e., the roots of $A(z)$).
- The **zeros** of $W(z)$ are all the values of $z \in \mathbb{C}$ such that $W(z) = 0$. Put it differently, the zeros of $W(z)$ are the values of z that make the nominator $C(z)$ equal to 0 (i.e., the roots of $C(z)$).

It's important to underline that, **to analyse zeros and poles,** $W(z)$ **has to be in positive form**, i.e., all exponents of z have to be positive (and we know that we can always go from the negative form to the positive form).

2.4.6 Auto Regressive Moving Average processes with exogenous input

Auto Regressive Moving Average with exogenous input are a further generalisation of ARMA processes.

Definition 10 (Auto Regressive Moving Average with eXogenous input)**.** *An Auto Regressive Moving Average with eXogenous input (ARMAX) process*

$$ARMAX(m, n, p, k)$$

is a process $y(t)$, generated by a white noise input $e(t)$ and by a measurable exogenous input $u(t)$.

$$\begin{aligned} y(t) &= a_1 y(t-1) + \cdots + a_m y(t-m) \\ &\quad + c_0 e(t) + c_1 e(t-1) + \cdots + c_n e(t-n) \\ &\quad + b_0 u(t-k) + b_1 u(t-k-1) + b_p u(t-k-p) \\ &= \sum_{i=1}^{m} a_i y(t-i) + \sum_{i=0}^{n} c_i e(t-i) + \sum_{i=0}^{p} b_i u(t-k-i) \end{aligned}$$

where

- $\sum_{i=1}^{m} a_i y(t-i)$ *is the Auto Regressive $AR(m)$ part.*
- $\sum_{i=0}^{n} c_i y(t-i)$ *is the Moving Average $MA(n)$ part.*
- $\sum_{i=0}^{p} b_i u(t-k-i)$ *is the exogenous input $X(p,k)$ part.*

Given an Auto Regressive Moving Average process with exogenous input $ARMAX(m, n, p, k)$, we say that

- m is the order of the AR process.
- n is the order of the MA process.
- p is the order of the X process.

- k is the delay of the X process.

Note that ARMAX and ARMA models are time-invariant and linear. They also are very general, and can be used to describe many processes of interest (clearly, a suitable selection of the model orders is required).

An ARMAX process is also the most general representation of the processes seen so far, in fact if we put some parameters to zero, we can obtain all the analysed processes (e.g., for $c_i = b_i = 0$ we obtain an AR process).

Extension of the ARMAX process

An ARMAX process can be further generalised considering a non linear combination of $y(t)$, $e(t)$ and $u(t)$. In such cases we can write

$$\begin{aligned} y(t) = f(&a_1 y(t-1), \dots, a_m y(t-m), \\ &c_0 e(t) + c_1 e(t-1), \dots, c_n e(t-n), \\ &b_0 u(t-k), b_1 u(t-k-1), b_p u(t-k-p)) \end{aligned}$$

where $f(\cdot)$ is a function (usually non linear, parametric and with some good approximation properties). These processes are called N-ARMAX.

2.4.7 Asymptotic stability

Given a digital filter $W(z)$, we say that

Theorem 3 (Asymptotic stability)**.** *A digital filer $W(z)$ is asymptotically stable if and only if all its poles have absolute value smaller than 1.*

Since z are complex variables, we can easily see that a digital filter is asymptotically stable if and only if it's poles are inside a circle of unitary radius in the complex plane. If we consider zeros, we obtain that

Theorem 4 (Minimum phase filter)**.** *A digital filer $W(z)$ is a minimum phase filter if and only if all its zeros have absolute value smaller than 1 (i.e., are inside the unitary circumference).*

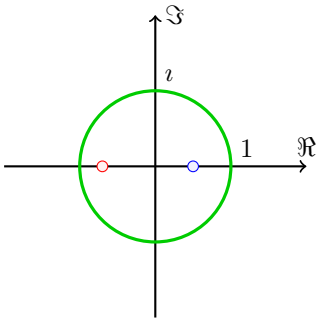


Figure 2.9: A circumference of unitary radius in the complex plane with a pole (red) and a zero (blue) on the real plane.

2.4.8 Well defined ARMA processes

Thanks to the operative representation of a process with its transfer function, we can finally say when an ARMA process is well defined.

Theorem 5 (Well defined ARMA process)**.** *Let*

- $W(z)$ *be a rational transfer function.*
- $v(t)$ *be an input stochastic process.*
- $y(t)$ *be the steady-state output of digital filter* $W(z)$ *fed with* $v(t)$*.*

The steady-state output $y(t)$ *is stationary, well defined (i.e., the limit in the definition of steady-state output converges) and converges if and only if*

- $v(t)$ *is a* ***stationary*** *stochastic process.*
- $W(z)$ *is* ***asymptotically stable****.*

In other words, the steady state output of an asymptotically stable digital filter $W(z)$ fed by a stationary stochastic process $v(t)$ is stationary as well.

Asymptotic convergence

Let us now consider an ARMA process $y(t)$ obtained as the output of an asymptotically stable digital filer $F(z)$ fed by an input stationary stochastic process $y(t)$

$$y(t) = F(z)v(t)$$

If we remember when we started studying well defined outputs, we had to initialise the value of the output as $y(t_0) = 0$ and use this value as starting point. If we consider the solution obtained starting from $y(t_0) = \bar{\gamma}$ instead of $y(t_0) = 0$, then $\tilde{y}$ is the solution of the recursive equation associated with $F(z)$. In general $\tilde{y}(t)$ is not a stationary stochastic process and it's different from $y(t)$ (as obtained starting from $y(t_0) = 0$). In other words, the asymptotic stability is related to initialisation in finite time. That being said, if $F(z)$ is asymptotically stable, then $\tilde{y}(t)$ will converge to $y(t)$ with $t \to \infty$ exponentially fast, for every finite starting value $y(t_0) = \bar{\gamma}$ we choose.

$$\tilde{y}(t) \to_{t\to\infty} y(t) \quad \forall t_0 \text{ finite} \quad \forall \bar{\gamma}$$
$$\tilde{y}(t,s) \to_{t\to\infty} y(t,s) \quad \forall s$$

In other words, if $F(z)$ is asymptotically stable, we start from a different value but we get close to the process $y(t)$, initiated with 0, exponentially fast. This result is summed up by the following theorem.

Theorem 6 (Gamma asymptotic stability)**.** *There is just one stationary output* $y(t)$ *which corresponds to the steady-state solution of* $A(z)y(z) = C(z)v(t)$*. However, if* $F(z)$ *is asymptotically stable, then all possible outputs* $\bar{y}(t)$ *obtained for different initialization* $y(t_0) = \bar{\gamma}$ *of the digital filter* $F(z)$ *tends asymptotically (as* $t \to \infty$*) to the solution, i.e. to the stationary output* $y(t)$*.*

Chapter 3

Weak characterisation of an ARMA process

Now that we know AR and ARMA processes, we would like to compute their mean and covariance function to have their weak characterisation.

3.1 AR processes

Let's start from Auto Regressive processes. In particular, we can start analysing an $AR(1) = ARMA(1,0)$ process

$$y(t) = ay(t-1) + e(t)$$

To better handle the process, let us switch to its operational representation

$$\begin{aligned} y(t) &= ay(t-1) + e(t) \\ y(t) &= az^{-1}y(t) + e(t) \\ y(t) - az^{-1}y(t) &= e(t) \\ (1 - az^{-1})y(t) &= e(t) \\ y(t) &= \frac{1}{1 - az^{-1}}e(t) \end{aligned}$$

Basically, we have found out that the transfer function $W(z)$ is

$$W(z) = \frac{1}{1 - az^{-1}}$$

The transfer function is in negative form, however we would like to have it in positive form to analyse it's poles and zeros and check its stability. To achieve this goal we can multiply it for $\frac{z}{z}$ and obtain

$$\begin{aligned} W(z) &= \frac{1}{1 - az^{-1}} \\ &= \frac{1}{1 - az^{-1}} \cdot \frac{z}{z} \\ &= \frac{z}{z - a} \end{aligned}$$

Now that the transfer function is in positive form, we can see that it has

- One zero in $z = 0$.
- One pole in $z = a$.

Since we want to ensure asymptotic stability, the pole has to be in the unitary circumference, hence we have to impose

$$|a| \leq 1$$

If this condition holds, $y(t)$ is stationary, well-defined and satisfies the recursive equation. Notice that, since $e(t)$ is a white noise, it's a SSP, so we have to check only that $|a| \leq 1$. Since $y(t)$ is well defined, it is a stationary stochastic process, hence we know that

- $m_y(t) = \mathbb{E}[y(t)] = m_y$ where m_y is constant.
- $\gamma_y(t, \tau) = \gamma_y(\tau)$, hence covariance is time independent.

However, we would also like to compute the values of m_y and $\gamma_y(\tau)$.

3.1.1 From an AR to a MA process

Since we know how to compute the average and the covariance of a Moving Average process, we can try to write an AR process as a MA process and than compute its characterisation using the formulas we have used for MA processes. In particular, can write an $AR(1)$ process as a $MA(\infty)$ process. Starting from

$$y(t) = \frac{1}{1 - az^{-1}} e(t)$$

we can infinitely expand $\frac{1}{1-az^{-1}}$ diving the numerator by the numerator to obtain

$$\begin{aligned} y(t) &= \frac{1}{1 - az^{-1}} e(t) \\ &= \left(1 + \frac{az^{-1}}{1 - az^{-1}}\right) e(t) \\ &= \left(1 + az^{-1} + \frac{a^2 z^{-2}}{1 - az^{-1}}\right) e(t) \\ &= \left(1 + az^{-1} + a^2 z^{-2} + \frac{a^3 z^{-3}}{1 - az^{-1}}\right) e(t) \end{aligned}$$

If we continue infinitely, we obtain

$$y(t) = (1 + az^{-1} + a^2 z^{-2} + a^3 z^{-3} + \cdots) e(t) = \left(1 + \sum_{i=1}^{\infty} a^i z^{-i}\right) e(t)$$

which is the definition of a $MA(\infty)$ process. A $MA(\infty)$ is well defined and stationary if $\sum_{i=0}^{\infty} |a|^i < \infty$. Since this is a geometric series, the necessary and sufficient condition is that $|a| < 1$, for $MA(\infty)$ to be stationary. This process can be repeated with whatever process $AR(n)$, simply executing the division of the reminder obtained at the previous iteration.

3.1.2 Alternative method

Mean

Using the $MA(\infty)$ representation might be a little hard, hence we can try and find a new method to compute mean and covariance. Let us start from the time-domain description of $y(t)$.

$$y(t) = ay(t-1) + e(t)$$

Now we can apply the expected value operator on both sides to obtain

$$\mathbb{E}[y(t)] = \mathbb{E}[ay(t-1) + e(t)]$$

Thanks to the linearity of the expected value, we can write

$$\mathbb{E}[y(t)] = a\mathbb{E}[y(t-1)] + \mathbb{E}[e(t)]$$

At this point we can remember that y is stationary, hence $m_y(t) = \mathbb{E}[y(t)]$ is time independent, hence $\mathbb{E}[y(t)] = \mathbb{E}[y(t-1)] = m_y$. Replacing this result we get

$$m_y = am_y + \mathbb{E}[e(t)]$$

Finally, we know that $e(t) \sim WN(0, \lambda^2)$, hence $\mathbb{E}[e(t)] = 0$, which leads us to

$$m_y = am_y + 0$$

which means that $m_y = 0$, as expected (because the mean of a $MA(\infty)$ is 0).

Covariance

To compute the covariance, let us start by the variance, which is the covariance for $\tau = 0$. As before, we can start from the time-domain representation.

$$\gamma_y(\tau = 0) = \mathbb{E}\big[(y(t-\tau) - m_y)^2\big]$$

Since we have already found out that the mean is 0, we can immediately rewrite the expression for the variance as

$$\gamma_y(\tau = 0) = \mathbb{E}\big[(y(t))^2\big]$$

Now we can replace the definition of $y(t)$ to obtain

$$\gamma_y(\tau = 0) = \mathbb{E}\big[(ay(t-1) + e(t))^2\big]$$

At this point, we can expand the square and apply the linearity of the expected value to get to

$$\begin{aligned}\gamma_y(\tau = 0) &= \mathbb{E}\big[(ay(t-1) + e(t))^2\big] \\ &= \mathbb{E}\big[a^2y^2(t-1) + e^2(t) + 2ay(t-1)e(t)\big] \\ &= a^2\mathbb{E}\big[y^2(t-1)\big] + \mathbb{E}\big[e^2(t)\big] + 2a\mathbb{E}\big[y(t-1)e(t)\big]\end{aligned}$$

Since y is stationary, the covariance is time independent, hence $\gamma_y(t-1, \tau = 0) = \gamma_y(t, \tau = 0) = \gamma_y(0)$. Moreover, $\mathbb{E}[e^2(t)]$ is the covariance of $e(t) \sim WN(0, \lambda^2)$, so $\mathbb{E}[e^2(t)] = \lambda^2$

$$\gamma_y(\tau = 0) = a^2\gamma_y(\tau = 0) + \lambda^2 + 2a\mathbb{E}\big[y(t-1)e(t)\big]$$

At this point we only have to understand what is the value of $\mathbb{E}[y(t-1)e(t)]$. We can start replacing $y(t-1) = ay(t-2) + e(t-1)$ to obtain

$$\begin{aligned}\mathbb{E}[y(t-1)e(t)] &= \mathbb{E}\big[(ay(t-2) + e(t-1)) \cdot e(t)\big] \\ &= \mathbb{E}\big[ay(t-2)e(t) + e(t-1)e(t)\big] \\ &= a\mathbb{E}[y(t-2)e(t)] + \mathbb{E}[e(t-1)e(t)\big]\end{aligned}$$

However, the whiteness property tells us that the white noise at different time instants in uncorrelated, hence $\mathbb{E}[e(t)e(t-1)] = 0$. If we keep replacing $y(t)$, we always obtain uncorrelated white noises ($\sum_i a^i \mathbb{E}[e(t-1-i)e(t)]$), hence we can say that $\mathbb{E}[y(t-1)e(t)] = 0$. If we replace this result in the equation of the covariance we get

$$\gamma_y(0) = a^2\gamma_y(0) + \lambda^2$$

We can now solve this equation to obtain the value of $\gamma_y(0)$

$$\begin{aligned}\gamma_y(0) &= a^2\gamma_y(0) + \lambda^2 \\ \gamma_y(0) - a^2\gamma_y(0) &= \lambda^2 \\ \big(1 - a^2\big)\gamma_y(0) &= \lambda^2 \\ \gamma_y(0) &= \frac{\lambda^2}{1-a^2}\end{aligned}$$

The same reasoning can be applied for $\tau = 1$ and (skipping the explanation of each passage, which is similar to the case $\tau = 0$) we obtain

$$\begin{aligned}\gamma_y(1) &= \mathbb{E}[y(t)y(t-1)] \\ &= \mathbb{E}\big[(ay(t-1) + e(t))y(t-1)\big] \\ &= \mathbb{E}\big[ay(t-1)y(t-1) + e(t)y(t-1)\big] \\ &= a\mathbb{E}\big[y^2(t-1)\big] + \mathbb{E}\big[e(t)y(t-1)\big] \\ &= a\gamma_y(0) + 0 \\ &= a\gamma_y(0) \\ &= \frac{a \cdot \lambda^2}{1-a^2}\end{aligned}$$

If we finally consider $\tau = 2$ we get

$$\begin{aligned}\gamma_y(2) &= \mathbb{E}[y(t)y(t-2)] \\ &= \mathbb{E}\big[(ay(t-1) + e(t))y(t-2)\big] \\ &= \mathbb{E}\big[ay(t-1)y(t-2) + e(t)y(t-2)\big] \\ &= a\mathbb{E}\big[y(t-1)y(t-2)\big] + \mathbb{E}\big[e(t)y(t-2)\big] \\ &= a\mathbb{E}\big[y(t-1)y(t-2)\big] + 0 \\ &= a\gamma_y(1) + 0\end{aligned}$$

Note that, to prove that $\mathbb{E}\big[e(t)y(t-2)\big] = 0$ we can replace $y(t-2)$ with its $MA(\infty)$ representation. Finally we can see a pattern and if we consider the general case (i.e., $\tau = \bar{\tau}$), we always obtain

$$\gamma_y(\bar{\tau}) = a\mathbb{E}\big[y(t-1)y(t-\bar{\tau})\big] + \mathbb{E}\big[e(t)y(t-\bar{\tau})\big]$$

The second element is always 0, for the same reason seen for $\tau = 0$ and the first element can always be written as $\gamma_y(\bar{\tau} - 1)$, hence we can derive a general formula for the covariance

$$\gamma_y(\tau) = \frac{a^{|\tau|}}{1-a^2}\lambda^2$$

The equations $\gamma_y(\bar{\tau}) = a\gamma_y(\bar{\tau} - 1)$ that allowed us to compute the general formula for the variance are called **Yule-Walker equations**.

3.2 ARMA processes

After analysing AR processes, let us switch to the more general case of ARMA processes. In particular, let us consider an ARMA process

$$y(t) = \sum_{i=1}^{m} a_i y(t-i) + e(t) + \sum_{i=0}^{n} c_i e(t-1)$$

where $e(t) \sim WN(0, \lambda^2)$.

3.2.1 Mean

To compute the mean of an ARMA process, let us apply the definition using the time-domain form of the process

$$\begin{aligned}\mathbb{E}[y(t)] &= \mathbb{E}\bigg[\sum_{i=1}^{m} a_i y(t-i) + \sum_{i=0}^{n} c_i e(t-1)\bigg] \\ &= \sum_{i=1}^{m} a_i \mathbb{E}\Big[y(t-i)\Big] + \sum_{i=0}^{n} c_i \mathbb{E}\Big[e(t-i)\Big]\end{aligned}$$

The mean $\mathbb{E}[e(t-i)]$ is 0 for all values of i because $e(t) \sim WN(0, \lambda^2)$ and $\mathbb{E}\Big[y(t-i)\Big]$ is constant and equal to m_y because y is stationary. Put this all together and we obtain

$$\begin{aligned}\mathbb{E}[y(t)] = m_y &= \sum_{i=1}^{m} a_i m_y + \sum_{i=0}^{n} c_i 0 \\ &= \sum_{i=1}^{m} a_i m_y\end{aligned}$$

The solution to this equation is trivially $m_y = 0$, hence we obtain

$$\mathbb{E}[y(t)] = 0$$

3.2.2 Covariance

Variance

As always, let's start by computing the covariance for $\tau = 0$. The story is always the same, let's replace the time-domain representation of $y(t)$

$$\begin{aligned}
\mathbb{E}\big[(y(t))^2\big] &= \mathbb{E}\left[\left(\sum_{i=1}^{m} a_i y(t-i) + \sum_{i=0}^{n} c_i e(t-i)\right)^2\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{m} a_i^2 y^2(t-i) + \sum_{i=0}^{n} c_i^2 e(t-i)^2 + \sum_{(i,j):i\neq j} 2a_i c_j y(t-i)e(t-j)\right] \\
&= \sum_{i=1}^{m} a_i^2 \mathbb{E}[y^2(t-i)] \\
&+ \sum_{i=0}^{n} c_i^2 \mathbb{E}[e(t-i)^2] \\
&+ \sum_{(i,j)\in(m\times m):i\neq j} 2a_i a_j \mathbb{E}[y(t-i)y(t-j)] \\
&+ \sum_{(i,j)\in(n\times n):i\neq j} 2a_i c_j \mathbb{E}[y(t-i)e(t-j)] \\
&+ \sum_{(i,j)\in(m\times n):i\neq j} 2c_i c_j \mathbb{E}[e(t-i)e(t-j)]
\end{aligned}$$

The last term can be immediately ruled out since it's the covariance of a white noise, hence we can rewrite the variance as

$$\begin{aligned}
\mathbb{E}\big[(y(t))^2\big] &= \sum_{i=1}^{m} a_i^2 \mathbb{E}[y^2(t-i)] \\
&+ \sum_{i=0}^{n} c_i^2 \mathbb{E}[e(t-i)^2] \\
&+ \sum_{(i,j):i\neq j} 2a_i a_j \mathbb{E}[y(t-i)y(t-j)] \\
&+ \sum_{(i,j):i\neq j} 2a_i c_j \mathbb{E}[y(t-i)e(t-j)]
\end{aligned}$$

Since $y(t)$ is stationary, its variance is time invariant and equal to $\gamma_y(0)$, hence we can replace $\mathbb{E}[y^2(t-i)]$ with $\gamma_y(0)$. Since we are already replacing $\mathbb{E}[y^2(t-i)] = \gamma_y(0)$, let us also replace $\mathbb{E}\big[(y(t))^2\big] = \gamma_y(0)$

$$\begin{aligned}
\gamma_y(0) = &\sum_{i=1}^{m} a_i^2 \gamma_y(0) + \sum_{i=0}^{n} c_i^2 \mathbb{E}[e(t-i)^2] \\
&+ \sum_{(i,j):i\neq j} 2a_i a_j \mathbb{E}[y(t-i)y(t-j)] + \sum_{(i,j):i\neq j} 2a_i c_j \mathbb{E}[y(t-i)e(t-j)]
\end{aligned}$$

Let us now focus on the terms $2a_i a_j \mathbb{E}[y(t-i)y(t-j)]$. For each value of $i \in [0, m]$, we can write

$$\sum_{j=1}^{m-i} 2a_i a_{i+j} \mathbb{E}[y(t-i)y(t-i-j)] = \sum_{j=1}^{m-i} 2a_i a_{i+j} \gamma_y(j)$$

hence we can rewrite the variance as

$$\begin{aligned}\mathbb{E}\big[(y(t))^2\big] &= \sum_{i=1}^{m} a_i^2 \gamma_y(0) + \sum_{i=0}^{n} c_i^2 \mathbb{E}[e(t-i)^2] \\ &+ \sum_{i=1} \sum_{j=1}^{m-i} 2a_i a_{i+j} \gamma_y(j) \\ &+ \sum_{(i,j):i\neq j} 2a_i c_j \mathbb{E}[y(t-i)e(t-j)]\end{aligned}$$

Another thing we can see is that $\mathbb{E}[e(t-i)^2]$ is the variance of $e(t) \sim WN(0, \lambda)$, hence all the terms $\mathbb{E}[e(t-i)^2]$ can be replaced with λ^2 to obtain

$$\begin{aligned}\mathbb{E}\big[(y(t))^2\big] &= \sum_{i=1}^{m} a_i^2 \gamma_y(0) + \sum_{i=0}^{n} c_i^2 \lambda^2 \\ &+ \sum_{i=1} \sum_{j=1}^{m-i} 2a_i a_{i+j} \gamma_y(j) \\ &+ \sum_{(i,j):i\neq j} 2a_i c_j \mathbb{E}[y(t-i)e(t-j)]\end{aligned}$$

Finally, using the $MA(\infty)$ representation of an AR process, we can cancel out the terms $\mathbb{E}[y(t-i)e(t-j)]$ since they are all zero. The final form of the variance is

$$\mathbb{E}\big[(y(t))^2\big] = \sum_{i=1}^{m} a_i^2 \gamma_y(0) + \sum_{i=0}^{n} c_i^2 \lambda^2 + \sum_{i=1} \sum_{j=1}^{m-i} 2a_i a_{i+j} \gamma_y(j)$$

Proceeding this way we get to a set of equations, called **Yule-Walker equations**

$$\begin{cases} \gamma_y(0) \\ \gamma_y(1) \\ \vdots \\ \gamma_y(m) \end{cases}$$

from which we can compute $\gamma_y(m)$, $\gamma_y(m+1)$, and so on.

3.3 Processes with non null mean

Until now, we have considered white noises with null mean. Let's try to compute mean and covariance for processes in which $e(t) \sim WN(\mu, \lambda^2)$.

3.3.1 Mean

As always, let's start from the definition of mean.

$$\begin{aligned}\mathbb{E}[y(t)] &= \mathbb{E}\Big[\sum_{i=1}^{m} a_i y(t-i) + \sum_{i=0}^{n} c_i e(t-1)\Big] \\ &= \sum_{i=1}^{m} a_i \mathbb{E}\big[y(t-i)\big] + \sum_{i=0}^{n} c_i \mathbb{E}\big[e(t-i)\big]\end{aligned}$$

Differently from the zero-mean case, we have to replace $\mathbb{E}\big[e(t-i)\big]$ with μ while $\mathbb{E}\big[y(t-i)\big]$ can still be replaced with m_y since $y(t)$ is stationary.

$$m_y = \sum_{i=1}^{m} a_i m_y + \sum_{i=0}^{n} c_i \mu$$

Now we can obtain m_y as follows

$$\begin{aligned} m_y &= \sum_{i=1}^{m} a_i m_y + \sum_{i=0}^{n} c_i \mu \\ m_y - \sum_{i=1}^{m} a_i m_y &= \sum_{i=0}^{n} c_i \mu \\ (1 - \sum_{i=1}^{m} a_i) m_y &= \sum_{i=0}^{n} c_i \mu \\ m_y &= \frac{\sum_{i=0}^{n} c_i \mu}{1 - \sum_{i=1}^{m} a_i} \end{aligned}$$

This means that, the mean for a general non-null process is

$$m_y = \frac{\sum_{i=0}^{n} c_i}{1 - \sum_{i=1}^{m} a_i} \cdot \mu$$

3.3.2 Covariance

Now it's time to compute the covariance. Let's start by writing the definition of covariance $\gamma_y(\tau)$.

$$\gamma_y(\tau) = \mathbb{E}\big[(y(t) - m_y)(y(t-\tau) - m_y)\big]$$

Before going on with the actual computations, let us also highlight that

$$\begin{aligned}\lambda^2 &= \mathbb{E}[(e(t) - \mu)^2] \\ &= \mathbb{E}[e(t)^2 + \mu^2 - 2\mu e(t)] \\ &= \mathbb{E}[e(t)^2] + \mu^2 - 2\mu\mathbb{E}[e(t)] \\ &= \mathbb{E}[e(t)^2] + \mu^2 - 2\mu^2 \\ &= \mathbb{E}[e(t)^2] - \mu^2\end{aligned}$$

This means that

$$\mathbb{E}[e(t)^2] = \lambda^2 + \mu^2$$

Moreover

$$\begin{aligned}
0 &= \mathbb{E}[(e(t) - \mu)(e(t-\tau) - \mu)] \\
&= \mathbb{E}[e(t)e(t-\tau) - \mu e(t) - \mu e(t-\tau) + \mu^2] \\
&= \mathbb{E}[e(t)e(t-\tau)] - \mathbb{E}[\mu e(t)] - \mathbb{E}[\mu e(t-\tau)] + \mathbb{E}[\mu^2] \\
&= \mathbb{E}[e(t)e(t-\tau)] - \mu\mathbb{E}[e(t)] - \mu\mathbb{E}[e(t-\tau)] + \mu^2 \\
&= \mathbb{E}[e(t)e(t-\tau)] = -\mu^2
\end{aligned}$$

then

$$\mathbb{E}[e(t)e(t-\tau)] = \mu^2$$

Knowing these information, we can compute the covariance directly, however it's much easier to use unbiased processes.

3.3.3 Unbiased processes

Now that we have all the ingredients ready, we can introduce an easy way to handle zero-mean processes. In particular, given two processes with non null-mean $e(t)$ and $y(t)$ we can define two new processes $\tilde{e}(t)$ and $\tilde{y}(t)$, called **unbiased processes**, that have 0 mean.

$$\tilde{y}(t) = y(t) - m_y$$

and

$$\tilde{e}(t) = e(t) - m_e$$

An important property of the unbiased processes is that they have null mean

$$\begin{aligned}
m_{\tilde{y}} &= \mathbb{E}[\tilde{y}(t)] \\
&= \mathbb{E}[y(t) - m_y] \\
&= \mathbb{E}[y(t)] - m_y \\
&= m_y - m_y = 0
\end{aligned}$$

Schematically,

$$\begin{aligned}
m_{\tilde{y}} &= \mathbb{E}[\tilde{y}(t)] = 0 \\
m_{\tilde{e}} &= \mathbb{E}[\tilde{e}(t)] = 0
\end{aligned}$$

As for now, we know that $\tilde{y}(t)$ has 0 mean, however, we would like to know what type of process it is. Let's understand it by replacing $y(t)$ with its definition (remember it's an ARMA process) and let's see where it leads us.

$$\begin{aligned}
\tilde{y}(t) &= y(t) - m_y \\
&= \sum_{i=1}^{m} a_i y(t-i) + \sum_{i=0}^{n} c_i e(t-i) - m_y
\end{aligned}$$

Now we can rewrite

- $e(t-i) = (\tilde{e}(t-i) + m_e)$
- $y(t-i) = (\tilde{y}(t-i) + m_y)$

and obtain

$$\begin{aligned}
\tilde{y}(t) &= \sum_{i=1}^{m} a_i(\tilde{y}(t-i) + m_y) + \sum_{i=0}^{n} c_i(\tilde{e}(t-i) + m_e) - m_y \\
&= \sum_{i=1}^{m} a_i\tilde{y}(t-i) + \sum_{i=1}^{m} a_i m_y + \sum_{i=0}^{n} c_i\tilde{e}(t-i) + \sum_{i=0}^{n} c_i m_e - m_y \\
&= \sum_{i=1}^{m} a_i\tilde{y}(t-i) + \left(\sum_{i=1}^{m} a_i - 1\right)m_y + \sum_{i=0}^{n} c_i\tilde{e}(t-i) + \sum_{i=0}^{n} c_i m_e \\
&= \sum_{i=1}^{m} a_i\tilde{y}(t-i) - \left(1 - \sum_{i=1}^{m} a_i\right)m_y + \sum_{i=0}^{n} c_i\tilde{e}(t-i) + \sum_{i=0}^{n} c_i m_e
\end{aligned}$$

Now we can remember that, for an ARMA process,

$$m_y = \frac{\sum_{i=0}^{n} c_i}{1 - \sum_{i=1}^{m} a_i} \cdot \mu$$

hence

$$\left(1 - \sum_{i=1}^{m} a_i\right)m_y = \sum_{i=0}^{n} c_i \cdot \mu$$

The term $\left(1 - \sum_{i=1}^{m} a_i\right)m_y$ can be replaced in the expression of $\tilde{y}(t)$ to derive (also remembering that $m_e = \mu$)

$$\begin{aligned}
\tilde{y}(t) &= \sum_{i=1}^{m} a_i\tilde{y}(t-i) - \sum_{i=0}^{n} c_i \cdot \mu + \sum_{i=0}^{n} c_i\tilde{e}(t-i) + \sum_{i=0}^{n} c_i m_e \\
&= \sum_{i=1}^{m} a_i\tilde{y}(t-i) + \sum_{i=0}^{n} c_i\tilde{e}(t-i)
\end{aligned}$$

It's easy to see that $\tilde{y}(t)$ has the same shape of an ARMA process, hence $\tilde{y}(t)$ is an ARMA process, i.e., the steady state output of a transfer function $W(z) = \frac{C(z)}{A(z)}$ fed by a zero-mean white noise $\tilde{e}(t) \sim WN(0, \lambda^2)$.

Finally, we can demonstrate that the covariance of $y(t)$ is the same as the one for a zero-mean process $\tilde{y}(t)$, in fact

$$\begin{aligned}
\gamma_y(\tau) &= \mathbb{E}[(y(t) - m_y)(y(t-\tau) - m_y)] \\
&= \mathbb{E}[\tilde{y}(t)\tilde{y}(t-\tau)] \\
&= \gamma_{\tilde{y}}(\tau)
\end{aligned}$$

and we have demonstrated that

$$\gamma_y(\tau) = \gamma_{\tilde{y}}(\tau)$$

Alternative interpretation

To understand why we can use unbiased processes, let us use a graphical approach. First things first, starting from $y(t) = W(z)e(t)$, let's try to draw it after replacing $e(t) = \tilde{e}(t)+\mu$ (from $\tilde{e}(t) = e(t)-\mu$).

$$y(t) = W(z)(\tilde{e}(t) + \mu)$$

The result is shown in Figure 3.1. Now, thanks to the linearity of the transfer function, we can write

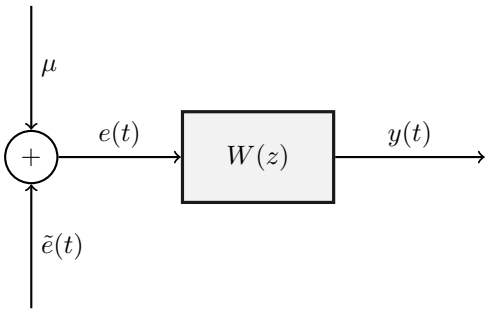


Figure 3.1: A visual representation of the equation $y(t) = W(z)(\tilde{e}(t) + \mu)$.

$y(t)$ as

$$y(t) = \tilde{e}(t)W(z) + \mu W(z)$$

Notice that, $\tilde{e}(t)W(z) = \tilde{y}(t)$. This new configuration is shown in Figure 3.2.

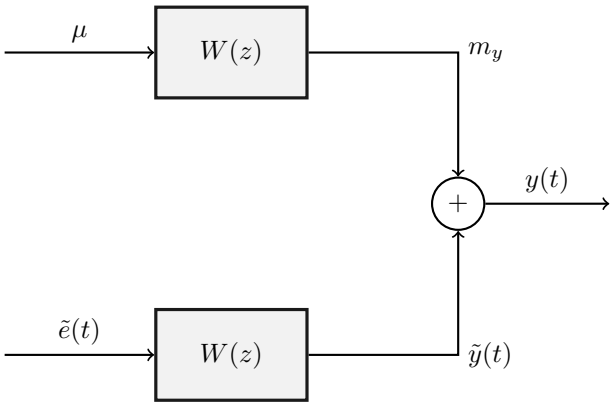


Figure 3.2: A visual representation of the equation $y(t) = W(z)\tilde{e}(t) + W(z)\mu$.

3.4 Gain theorem

Let us consider a non-zero mean process

$$y(t) = \frac{C(z)}{A(z)}e(t)$$

where $e(t) \sim WN(\mu, \lambda^2)$. The equation can be rewritten as

$$A(z)y(t) = C(z)e(t)$$

If we consider $A(z)$ and $C(z)$ as functions of z we can consider them in $z = 1$ and apply the expected operator to both sides to obtain

$$\mathbb{E}\big[A(1)y(t)\big] = \mathbb{E}[C(1)e(t)]$$

Since $A(1)$ and $C(1)$ are constant value, we can take them outside the expected value. Moreover, $\mathbb{E}[e(t)] = \mu$, hence we get

$$\begin{aligned} A(1)\mathbb{E}\big[y(t)\big] &= C(1)\mathbb{E}[e(t)] \\ &= C(1)\mu \end{aligned}$$

Finally we can divide both sides for $A(1)$ to obtain

$$\mathbb{E}\big[y(t)\big] = \frac{C(1)}{A(1)}\mu$$

This is an important result, which is called **gain theorem**.

Theorem 7 (Gain theorem)**.** *Given a non-zero mean stationary process $y(t)$ which is the steady-state output of a transfer function fed by $e(t) \sim WN(\mu, \lambda^2)$, its mean m_y can be computed as*

$$\mathbb{E}[y(t)] = m_y = \frac{C(z)}{A(z)}\bigg|_{z=1} \cdot \mu = \frac{C(1)}{A(1)} \cdot \mu$$

An important consequence of this theorem is that, given a stochastic process $y(t)$ generated as the output of a digital filter $W(z)$ fed by a stationary stochastic process, if

- $F(z)$ is asymptotically stable (i.e., $y(t)$ is a stationary stochastic process
- $E[v(t)] = m_v = 0$.

then $E[y(t)] = m_y = 0$.

Chapter 4

Frequency domain analysis

4.1 Spectral density

Let us consider a stationary stochastic process $y(t)$ with

- Mean $\mathbb{E}[y(t)] = m_y$.
- Covariance $\mathbb{E}[(y(t) - m_y)(y(t-\tau) - m_y)] = \gamma_y(\tau)$.

To analyse $y(t)$ in the frequency domain, we have to introduce the concept of spectral density (or spectrum).

Definition 11 (Spectral density)**.** *The spectral density of a stationary stochastic process $y(t)$, also called the spectrum of $y(t)$ is the discrete Fourier transform of the covariance function $\gamma_y(\tau)$*

$$\Gamma_y(\omega) = \sum_{\tau=-\infty}^{\infty} \gamma_y(\tau) \cdot e^{-j\omega\tau} = \mathcal{F}\{\gamma_y(\tau)\}$$

In the definition of spectrum, we recognise the following components

- ω is the frequency.
- Γ_y is the spectrum of $y(t)$.
- $\gamma_y(\tau)$ is a discrete time signal.

Notice that, the spectrum isn't the Fourier transform of the process $y(t)$ itself but the transform of the covariance function $\gamma_y(\tau)$.

4.1.1 Properties

Some important properties of a process's spectrum are

1. $\Gamma_y(\omega)$ is a **real function** of real values ω.

$$\Gamma_y(\omega) \in \mathbb{R} \quad \forall\omega \tag{4.1}$$

2. $\Gamma_y(\omega)$ is a **positive function**.
$$\Gamma_y(\omega) \geq 0 \quad \forall\omega \tag{4.2}$$
3. $\Gamma_y(\omega)$ is an **even function**.
$$\Gamma_y(\omega) = \Gamma_y(-\omega) \quad \forall\omega \tag{4.3}$$
4. $\Gamma_y(\omega)$ is a **periodic** function with period 2π.
$$\Gamma_y(\omega) = \Gamma_y(\omega + 2k\pi) \quad \forall\omega, k = 0, \pm1, \pm2, \dots \tag{4.4}$$

As a consequence of property 4.4 we can draw the spectrum only between $[-\pi, \pi]$, exploiting the spectrum periodicity. Moreover, thanks to property 4.3, we can further reduce the interval and draw the spectrum only in $[0, \infty)$. This means that, $\omega = \pi$ is the largest frequency for sinusoidal discrete time signals. This also means that, since
$$\omega = \frac{2\pi}{T}$$
then minimum period T is
$$T = \frac{2\pi}{\pi} = 2$$

Spectrum of a sum of stationary stochastic processes

Say $\xi(t)$ and $e(t)$ are two stationary stochastic processes not correlated one to the other. If a process $y(t)$ is the sum of $\xi(t)$ and $e(t)$,
$$y(t) = W(z)\xi(t) + T'(z)e(t)$$
then its spectrum can be written as the sum of the spectra of $\xi(t)$ and $e(t)$.
$$\Gamma_y(\omega) = \Gamma_e(\omega) + \Gamma_\xi(\omega)$$

4.1.2 Frequency response

Given a stochastic process
$$y(t) = W(z)v(t) = \frac{C(z)}{A(z)}v(t)$$
where $W(z)$ is asymptotically stable, we can compute its spectrum using Definition 11, however it's a little complex. Luckily, we can introduce a new method to compute $\Gamma_y(\omega)$.

Theorem 8 (Frequency response)**.** *The spectrum $\Gamma_y(\omega)$ of a stochastic process $y(t) = W(z)v(t)$ is the product of the spectrum of the input $v(t)$ and the square of the transfer function evaluated in $e^{j\omega}$*
$$\Gamma_y(\omega) = \left|W(e^{j\omega})\right|^2 \cdot \Gamma_v(\omega)$$

Notice that, we are considering $W(z)$ as a function of the complex variable z.

4.1.3 Euler representation of the exponential

An important formula to remember and that is going to be very useful is the following
$$e^{-j\omega} + e^{j\omega} = \cos(\omega) - j\sin(\omega) + \cos(\omega) + j\sin(\omega) = 2\cos(\omega)$$

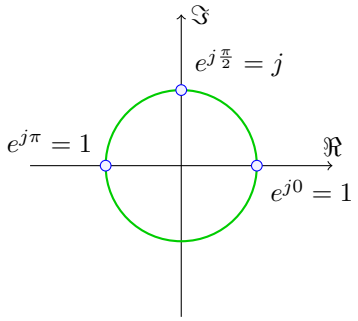


Figure 4.1: The unitary circumference in the complex plane and the points where it's interesting to compute the value of the spectrum.

4.1.4 Spectrum of the white noise

Now that we know enough about the spectrum of a process, we can try to compute the spectral density of a white noise $e(t) \sim WN(\mu, \lambda^2)$. Remember that, the covariance of the white noise is zero for all τ except for $\tau = 0$. Applying the definition of spectrum we get

$$\begin{aligned}\Gamma_e(\omega) &= \sum_{\tau=-\infty}^{\infty} \gamma_e(\tau) \cdot e^{-j\omega\tau} \\ &= \lambda^2 e^{-j\omega 0} \sum_{\tau \neq 0} 0 \cdot e^{-j\omega\tau} \\ &= \lambda^2\end{aligned}$$

Hence, the spectrum of a white noise with variance λ^2 is constantly λ^2. A graphical representation is shown in Figure 4.2.

Now that we know how the shape of the white noise's spectrum, we can use Theorem 8 to compute the spectrum of a process that is output of a transfer function $W(z)$ fed by a white noise

$$\Gamma_y(\omega) = \left|W(e^{j\omega})\right|^2 \cdot \lambda^2$$

4.1.5 Anti-transformation

Given the spectrum $\Gamma_y(\omega) = \mathcal{F}\{\gamma_y(\tau)\}$ of a process $y(t)$, we can always obtain the covariance function using the anti-transformation

Definition 12 (Anti-transformation)**.** *The anti-transformation $\gamma_y(\tau)$ of a spectrum $\Gamma_y(\omega)$ is the integral between $-\pi$ and π of the product between the spectrum and $e^{j\omega\tau}$*

$$\begin{aligned}\gamma_y(\tau) &= \mathcal{F}^{-1}\{\Gamma_y(\omega)\} \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \Gamma_y(\omega) e^{j\omega\tau}\, d\omega\end{aligned}$$

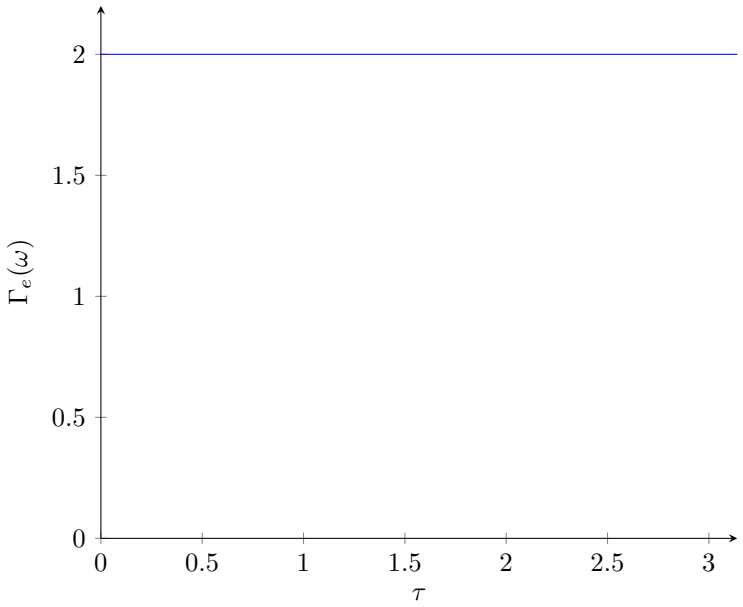


Figure 4.2: The spectrum of a white noise $e(t) \sim WN(\mu, \lambda^2)$.

From this definition, we can compute the variance ($\tau = 0$) of a spectrum $\Gamma_y(\omega)$ as

$$\begin{aligned}\gamma_y(0) = V_y &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \Gamma_y(\omega) e^{j\omega 0}\, d\omega \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \Gamma_y(\omega)\, d\omega\end{aligned}$$

This means that, the variance of a process is the area under the spectrum of that process.

The relationship between the spectrum and the covariance is bi-univocal, hence the notion of spectrum doesn't add anything to the weak description of a process, hence the description $\langle m_y, \gamma_y(\tau)\rangle$ is equivalent to $\langle m_y, \Gamma_y(\omega)\rangle$. The spectrum only gives a different perspective and is useful to understand some properties.

4.1.6 Kinchine-Wiener theorem

Say $y(t)$ is a stationary stochastic process and that we filter it with an ideal pass-band filter $F(z)$ (as in Figure 4.3) to obtain a process $\tilde{y}(t)$

$$\tilde{y}(t) = F(z)y(t)$$

The pass-band filter, lets only the frequencies between $\bar{\omega}$ and $\bar{\omega} + \delta$ pass. In this conditions, the following theorem holds.

Theorem 9 (Wiener Kinchine)**.** *The spectrum $\Gamma_y(\bar{\omega})$ of the input process $y(t)$ is the limit, for*

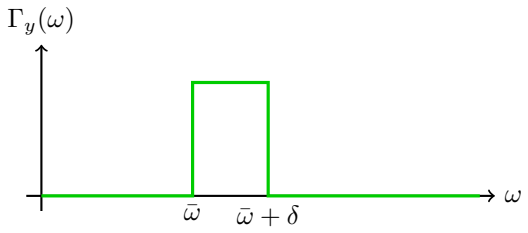


Figure 4.3: A pass-band filter.

$\delta \to 0$ of the variance of the filtered output $\tilde{y}(t)$.

$$\Gamma_y(\bar{\omega}) = \lim_{\delta \to 0} \gamma_{\tilde{y}(0)}$$

4.2 Characterisations of a process

From what we have seen until now, a process can be described (at least) in four different ways

1. The **time domain** representation uses the equations in the time domain.

$$y(t) = \sum_{i=1}^{m} a_i y(t-i) + \sum_{i=0}^{n} c_i e(t-i)$$
$$e(t) \sim WN(\mu, \lambda^2)$$

2. The **operational** representation uses the transfer functions.

$$y(t) = \frac{C(z)}{A(z)} e(t)$$
$$e(t) \sim WN(\mu, \lambda^2)$$

3. The **probabilistic** characterisation uses mean and covariance.

$$m_y$$
$$\gamma_y(\tau)$$

4. The **frequency-domain** characterisation uses mean and the spectrum.

$$m_y$$
$$\Gamma_y(\omega)$$

The first and second characterisations are equivalent because the latter is obtained from the former using the time-shift operator. Moreover, the third and fourth characterisations are equivalent because of the Fourier transform and anti-transform. From representations 1 and 2 we can obtain 3 and 4 since $\gamma_y(\tau)$ is computed from the transfer function $W(z)$. What we haven't analysed yet, is how to get from characterisations 3 and 4 to 1 and 2.

To achieve this goal, let's start from the operational representation

$$y(t) = \frac{C(z)}{A(z)} e(t)$$

Using Theorem 8 we can write

$$\begin{aligned}\Gamma_y(\omega) &= |W(z)|^2 \Gamma_e(\omega) \\ &= \frac{|C(z)|^2}{|A(z)|^2} \lambda^2\end{aligned}$$

From this description, we can say that $y(t)$ has a rational spectral density because it is a rational function of the variable $e^{j\omega}$. The reverse is also true thanks to the following theorem

Theorem 10. *Let $y(t)$ be a stationary stochastic process with rational spectral density. Then, there exists a white noise process $\xi(t)$ with suitable mean and variance and a rational transfer function $W(z)$ such that*

$$y(t) = W(z)\xi(t)$$

hence $y(t)$ is an ARMA process.

However, this representation isn't unique, in fact multiple ARMA processes (e.g., $W(z)\xi(t)$ and $\tilde{W}(z)\tilde{\xi}(t)$) can generate the same process $y(t)$.

Among all possible ARMA models we can choose, one is very important. Let us consider the operational representation

$$y(t) = W(z) = \frac{C(z)}{A(z)} \xi(t)$$

Now let us consider a zero $z = q$ of $W(z)$, with module greater than 1 ($|q| > 1$). The transfer function can be then rewritten as

$$W(z) = W_1(z) \cdot (z - q)$$

If we replace the new formulation of $W(z)$ in the ARMA model, we obtain

$$y(t) = W_1(z)(z - q)\xi(t)$$

We can all agree that we can multiply $W_1(z)$ for $\frac{z-\frac{1}{q}}{z-\frac{1}{q}}$ and obtain

$$\begin{aligned}y(t) &= W_1(z)(z - q)\frac{z - \frac{1}{q}}{z - \frac{1}{q}} \xi(t) \\ &= W_1(z)\left(z - \frac{1}{q}\right)\left[\frac{z - q}{z - \frac{1}{q}} \xi(t)\right] \\ &= \tilde{W}(z)\tilde{\xi}(t)\end{aligned}$$

where

- $\tilde{W}(z) = W_1(z)\left(z - \frac{1}{q}\right)$
- $\tilde{\xi}(t) = \frac{z-q}{z-\frac{1}{q}} \xi(t)$

The new transfer function $\tilde{W}(z)$ is rational (numerator and denominator are polynomials) since it's the product between a rational transfer function $W_1(z)$ and a polynomial factor. Now we should check if $\tilde{\xi}(t)$ is a white noise. To do so, we can compute the spectral density of $\tilde{\xi}(t)$, considering $\frac{z-q}{z-\frac{1}{q}}$ as the transfer function and $\xi(t)$ as the input signal. Applying Theorem 8 we get

$$\begin{aligned}
\Gamma_{\tilde{\xi}}(\omega) &= \left|\frac{z-q}{z-\frac{1}{q}}\right|_{z=e^{j\omega}} \Gamma_{\xi}(\omega) = \frac{|e^{j\omega}-q|^2}{\left|e^{j\omega}-\frac{1}{q}\right|^2}\lambda^2 \\
&= \frac{(e^{j\omega}-q)(e^{-j\omega}-q)}{\left(e^{j\omega}-\frac{1}{q}\right)\left(e^{j\omega}-\frac{1}{q}\right)}\lambda^2 \\
&= \frac{1+q^2-q(e^{j\omega}+e^{-j\omega})}{1+\frac{1}{q^2}-\frac{1}{q}(e^{j\omega}+e^{-j\omega})}\lambda^2 \\
&= \frac{1+q^2-2q\cos(\omega)}{1+\frac{1}{q^2}-\frac{1}{q}\cos(\omega)}\lambda^2 \\
&= \frac{q^2\left(\frac{1}{q^2}+1-\frac{2}{q}\cos(\omega)\right)}{1+\frac{1}{q^2}-\frac{1}{q}\cos(\omega)}\lambda^2 \\
&= q^2\lambda^2
\end{aligned}$$

The spectral density is constant for all values of γ, hence $\tilde{\xi}(t)$ is a white noise. Now that we know that $\tilde{\xi}(t)$ is a white noise of variance $q^2\lambda^2$, let us compute it's mean so that we can fully characterise it. In particular, to compute the mean we can use the Gain Theorem 7.

$$\begin{aligned}
\mathbb{E}[\tilde{\xi}(t)] &= \left|\frac{z-q}{z-\frac{1}{q}}\right|_{z=1} \mathbb{E}[\xi(t)] \\
&= \frac{1-q}{1-\frac{1}{q}}\xi(t) \\
&= \frac{q\left(\frac{1}{q}-1\right)}{1-\frac{1}{q}}\xi(t) \\
&= q\mu
\end{aligned}$$

Hence, $\tilde{\xi}(t) = \frac{1-q}{1-\frac{1}{q}}\xi(t)$ is a white noise of mean $q\mu$ and variance $q^2\lambda^2$

$$\tilde{\xi}(t) \sim WN(q\mu, q^2\lambda^2)$$

and

$$y(t) = \tilde{W}(z)\tilde{\xi}(t) = W_1(z)\left(z-\frac{1}{q}\right)\left[\frac{z-q}{z-\frac{1}{q}}\xi(t)\right]$$

is a new representation of $y(t) = W(z)y(t)$ (and $y(t)$ is the same process in both cases, just written in a different way).

Apart from the one examined (and three others), there are no other sources of ambiguity in defining an ARMA process. This is expressed by the following theorem.

Theorem 11 (Spectral factorisation)**.** *Let $y(t)$ be a stationary stochastic process with rational spectral density. Then, there exists an unique white noise process $\xi(t)$ with suitable mean and variance and an unique rational transfer function $W(z)$ such that*

$$y(t) = W(z)\xi(t) = \frac{C(z)}{A(z)}\xi(t)$$

and

1. *$C(z)$ and $A(z)$ are **monic** (i.e. the coefficients of the maximum degree terms of $C(z)$ and $A(z)$ are equal to 1).*
2. *$C(z)$ and $A(z)$ have **null relative degree**.*
3. *$C(z)$ and $A(z)$ are **coprime** (i.e. they have no common factors).*
4. *The absolute value of the poles and the zeroes of $W(z)$ is less than or equal to 1 (i.e. poles and zeroes are inside the unit circle).*

*When all the four conditions above are satisfied, $y(t) = W(z)\xi(t)$ is a **canonical representation** of $y(t)$.*

It's important to underline that **every ARMA process admits a canonical representation**. Until now, we have seen that a digital filter, depending on the module of the poles and zeros of its transfer function, can have different properties. A summary of such properties is shown in Table 4.1.

Module	Property
poles: $\lvert\cdot\rvert < 1$	Asymptotically stable
zeros: $\lvert\cdot\rvert < 1$	Minimum phase filter
poles: $\lvert\cdot\rvert \leq 1$ zeros: $\lvert\cdot\rvert \leq 1$	Canonical representation

Table 4.1: Properties of a digital filter $W(z) = \frac{C(z)}{A(z)}$ depending on the module of its poles and zeros.

Part II

Linear optimal prediction

Chapter 5

k-step prediction

Let us consider an ARMA process

$$y(t) = \frac{C(z)}{A(z)} e(t)$$

with $e(t) \sim WN(0, \lambda^2)$ and $W(z) = \frac{C(z)}{A(z)}$ asymptotically stable. Before going on, let us also assume that

1. $y(t) = W(z)e(t)$ is a **canonical representation** of the process $y(t)$ (as defined by Theorem 11).
2. Every zero of the transfer function $W(z)$ is strictly inside the unitary circle.

   $$|z_i| < 1 \quad \forall \text{ zeros } z_i$$

   Notice that this assumption is a further constraint of the first one, in fact $y(t)$ is in canonical form if all it's poles are smaller or equal to 1, however, with this assumption, we are ruling out the case in which the poles are equal to one. The reason for this will be clear later on.

Under these assumptions we can try and solve the k-step prediction problem. The goal is, given n observations of process $y(t)$ from $t - n - 1$ to t, to predict the future value of process y at time $t + k$ (i.e., $y(t + k)$).

$$y(t-n-1), y(t-n), \ldots, y(t-1), y(t) \xmapsto{predict} y(t+k)$$

The function that, given the values up to time t, returns the value of the process at time $t + k$, is

$$y(t + k|n)$$

and is called **predictor**. Notice that, the information up to time t is an infinite series of observed values and $y(t + k|t)$ is a function of the values $y(t - i)$ from $i = 0$ to $i = \infty$

$$y(t + k|t) = f\big(y(t), y(t-1), y(t-2), \ldots\big)$$

That being said, there exists many functions $f(\cdot)$ that take as input the values $y(t - i)$, however we would like to have most accurate one, i.e., the one that returns a value as close as possible to the actual value at time $t + k$. In other words, we want to **minimise the error that we do when we predict** $y(t + k|t)$.

To make things clear before introducing the concept of error, let us call

- $\hat{y}(t+k|t)$ the value predicted by the predictor.
- $y(t+k)$ the actual value we want to approximate.

Basically we want to use to build a predictor that generates a value $\hat{y}(t+k|t)$ as close as possible to the real value $y(t+k)$. Always remember that the values $y(t)$ depend on the random variable s, hence also $y(t+k)$ and the values till time t are values that depend on the random variable s. In other words,

- The value at $t+k$ can be written as $y(t+k,s)$.
- The values $y(t-i)$ until time t can be written as $y(t-i,s)$.

Hence the predictor can be written as

$$\hat{y}(t+k|t) = f(y(t,s), y(t-1,s), y(t-2,s), \dots) = y(t+k|t,s)$$

This means that the predictor $y(t+k|t,s)$ is stochastic itself and it depends on the realization of the stochastic process $y(t)$ we measured.

5.1 Measuring error

As we said before, among all possible predictors, we want to find the one with the minimum error, i.e., the one whose predictions are closer to the actual values. To understand how good a prediction is, we have to introduce the concept of error and a way to measure it.

5.1.1 Mean square prediction error

A widely used error for its simplicity is the mean square prediction error.

Definition 13 (Mean square prediction error)**.** *The mean square prediction error is the squared distance between the true value $y(t+k)$ and the prediction $y(t+k|t)$*

$$\mathbb{E}\big[\big(y(t+k) - y(t+k|t)\big)^2\big]$$

and

$$\varepsilon(t+k|t) = y(t+k) - y(t+k|t)$$

*is called **prediction error**.*

5.2 Optimal linear predictor

Now that we know how to measure the error (i.e., using the mean square prediction error) we can define the optimal linear predictor $\hat{y}(t+k|t)$

Definition 14 (Optimal linear predictor)**.** *The optimal linear predictor $\hat{y}(t+k|t)$ is the one with the optimal coefficients α_i^o that minimise the mean square error.*

$$\min_{\alpha_i} \mathbb{E}\big[\big(y(t+k) - \hat{y}(t+k|t)\big)^2\big] = \min_{\alpha_i} \mathbb{E}\left[\left(y(t+k) - \sum_{i=0}^{\infty} \alpha_i y(t-i)\right)^2\right]$$

5.2.1 Linear predictors

The number of predictors we can choose is too big, hence we should limit ourselves to linear predictors.

Definition 15 (Linear predictor)**.** *A linear predictor is a predictor $\hat{y}(t+k|t)$ that linearly combines the process's values until time t. In other words, $\hat{y}(t+k|t)$ is a linear combination (of parameters α) of the previous values of y*

$$\hat{y}(t+k|t) = \sum_{i=0}^{\infty} \alpha_i y(t-i)$$

The parameters α_i of the predictor should guarantee that $\hat{y}(t+k|t)$ is well defined, hence we should ensure that

$$\sum_{i=0}^{\infty} \alpha_i^2 < +\infty$$

Notice that, being a process, we can write the linear predictor $\hat{y}(t+k|t)$ in its operational form

$$\begin{aligned}
\hat{y}(t+k|t) &= \sum_{i=0}^{\infty} \alpha_i y(t-i) \\
&= \sum_{i=0}^{\infty} \alpha_i z^{-i} y(t) \\
&= \left(\sum_{i=0}^{\infty} \alpha_i z^{-i} \right) y(t) \\
&= F_\alpha(z) y(t)
\end{aligned}$$

This also means that the predictor $\hat{y}(t+k|t)$ is the steady-state solution of the linear filter $F_\alpha(z)$ fed with the input signal $y(t)$.

5.2.2 Optimal predictor from the noise

Before minimising the mean square prediction error, we should solve an easier problem. Let's start by saying that $y(t)$ is the steady-state solution of the recursive equation $y(t) = W(z)v(t)$ and can be written as a Moving Average process of infinite order (to see how, check the chapter on the weak characterisation of ARMA processes)

$$y(t) = \sum_{i=0}^{\infty} w_i e(t-i)$$

where w_i is a function of the coefficients of $C(z)$ and $A(z)$.

This representation of $y(t)$ can be used in the linear predictor. In particular, if we replace it in the equation of the predictor, we obtain

$$\begin{aligned}
\hat{y}(t+k|t) &= \sum_{i=0}^{\infty} \alpha_i y(t-i) \\
&= \sum_{i=0}^{\infty} \alpha_i \sum_{j=0}^{\infty} w_j e(t-j-i)
\end{aligned}$$

In the expression above, we can collect all terms with same $e(t)$ to obtain the following expression

$$\hat{y}(t+k|t) = \sum_{i=0}^{\infty} \beta_i e(t-i)$$

This means that, every linear predictor based on the previous observations of the process can be written as a linear predictor based on the previous observations (i.e., measurements) of the noise $e(t)$ in input up to time t. This means that, instead of finding the parameters α_i that minimise the predictor, we can find the optimal values of β_i (i.e., those that minimise the linear predictor written as a function of the noise). Basically, we want to find the optimal noise-based linear predictor, defined as

Definition 16 (Optimal noise-based linear predictor)**.** *The optimal noise-based linear predictor is the linear predictor*

$$\hat{y}(t+k|t) = \sum_{i=0}^{\infty} \beta_i^o e(t-i)$$

whose optimal parameters β_i^o minimise the mean square prediction error.

$$\min_{\{\beta_i\}} \mathbb{E}\big[\big(y(t+k) - \hat{y}(t+k|t)\big)^2\big] = \min_{\{\beta_i\}} \mathbb{E}\big[\big(y(t+k) - \sum_{i=0}^{\infty} \beta_i^o e(t-i)\big)^2\big]$$

To minimise the error $\varepsilon(t+k|t)$, we can try and write $y(t+k)$ in a different way. In particular, we can notice that, it can be written as a $MA(\infty)$, too. In particular, we can write

$$\begin{aligned} y(t+k) &= \sum_{i=0}^{\infty} w_i e(t+k-i) \\ &= \sum_{i=0}^{k-1} w_i e(t+k-i) + \sum_{i=k}^{\infty} w_i e(t+k-i) \end{aligned}$$

Finally, we can replace $i = j + k$ in the second addend to obtain

$$y(t+k) = \sum_{i=0}^{k-1} w_i e(t+k-i) + \sum_{j=0}^{\infty} w_{j+k} e(t-j)$$

Thanks to this representation, we can write write the mean square error as

$$\mathbb{E}\big[\big(y(t+k) - \hat{y}(t+k|t)\big)^2\big] = \mathbb{E}\Big[\Big(\sum_{i=0}^{k-1} w_i e(t+k-i) + \sum_{j=0}^{\infty} w_{j+k} e(t-j) - \sum_{i=0}^{\infty} \beta_i^o e(t-i)\Big)^2\Big]$$

Now we can expand the square and directly apply the linearity of the expected value to obtain

$$\begin{aligned} \mathbb{E}\big[\big(y(t+k) - \hat{y}(t+k|t)\big)^2\big] &= \mathbb{E}\Big[\Big(\sum_{i=0}^{k-1} w_i e(t+k-i)\Big)^2\Big] \\ &\quad + \mathbb{E}\Big[\Big(\sum_{j=0}^{\infty} w_{j+k} e(t-j) - \sum_{i=0}^{\infty} \beta_i^o e(t-i)\Big)^2\Big] \\ &\quad + 2\mathbb{E}\Big[\Big(\sum_{i=0}^{k-1} w_i e(t+k-i)\Big)\Big(\sum_{j=0}^{\infty} w_{j+k} e(t-j) - \sum_{i=0}^{\infty} \beta_i^o e(t-i)\Big)\Big] \end{aligned}$$

Since we have zero-mean white noises all around, we can try and reason on whether they are correlated or not. In particular, we can put our focus on the last added. The first term in the added contains values of $e(t)$ that go from $e(t+k)$ to $e(t+1)$, hence all values at time instances in the future. On the other hand, in the second term, $e(t)$ is evaluated only at time instances in the past. This means that, multiplying the first and second term, we can't obtain values of the noise at the same time instance, hence all values are uncorrelated and 0 for the whiteness properties. Long story short, we can eliminate this added. The new form of the mean square prediction error is therefor

$$\mathbb{E}\big[\big(y(t+k)-\hat{y}(t+k|t)\big)^2\big] = \mathbb{E}\Big[\Big(\sum_{i=0}^{k-1} w_i e(t+k-i)\Big)^2\Big] + \mathbb{E}\Big[\Big(\sum_{j=0}^{\infty} w_{j+k} e(t-j) - \sum_{i=0}^{\infty} \beta_i^o e(t-i)\Big)^2\Big]$$

The first element can't be controlled, since it doesn't contains the parameters β_i. On the other hand, the second addend contains the parameters β_i, hence we can try and minimise it. If we notice that it's always positive (because of the square), we understand that the minimum value it can take is 0, hence we want to find the values of β_i for which

$$\sum_{j=0}^{\infty} w_{j+k} e(t-j) - \sum_{i=0}^{\infty} \beta_i^o e(t-i) = 0$$

Moving the second addend to the right-hand side we obtain

$$\sum_{j=0}^{\infty} w_{j+k} e(t-j) = \sum_{i=0}^{\infty} \beta_i^o e(t-i)$$

which means that the values of β_i that minimise the error are

$$\beta_i^o = w_{j+k} \quad \forall i = 0, 1, 2, \dots$$

If we replace the newly found optimal values of β_i^o in the expression of the noise-based linear predictor we obtain

$$\hat{y}(t+k|t) = \sum_{i=0}^{\infty} w_{i+k} e(t-i)$$

Practical computation of the linear predictor

Now that we know what is the shape of the optimal noise-based linear predictor, we would like to understand how to compute the values of w_{i+k}. To achieve this goal we have to come back to the operational representation of the process $y(t)$ we are trying to predict. In particular $y(t+k)$ can be written as

$$y(t+k) = W(z)e(t+k) = \frac{C(z)}{A(z)} e(t+k)$$

To compute the coefficients w_{i+k} we have to do the k-step division between $C(z)$ and $A(z)$. In particular, we want to write $\frac{C(z)}{A(z)}$ as

$$\frac{C(z)}{A(z)} = E(z) + \frac{z^{-k} F(z)}{A(z)}$$

where

- $z^{-k}F(z)$ is the reminder of the division between $C(z)$ and $A(z)$ after k steps.
- $E(z)$ is the integer division between $C(z)$ and $A(z)$.

We can also explicitly write the values of the quotient and the reminder as

- $E(z) = \sum_{i=0}^{k-1} w_i z^{-i}$
- $z^{-k}\frac{F(z)}{A(z)} = \sum_{i=k}^{\infty} w_i z^{-i} = \sum_{i=0}^{\infty} w_{k+i} z^{-k-i}$

Since we are going to need it in a few moments, we can derive $\frac{F(z)}{A(z)}$ as

$$\begin{aligned} z^{-k}\frac{F(z)}{A(z)} &= \sum_{i=0}^{\infty} w_{k+i} z^{-k-i} \\ z^k z^{-k}\frac{F(z)}{A(z)} &= \sum_{i=0}^{\infty} w_{k+i} z^{-k-i} z^k \\ \frac{F(z)}{A(z)} &= \sum_{i=0}^{\infty} w_{k+i} z^{-k-i+k} \\ \frac{F(z)}{A(z)} &= \sum_{i=0}^{\infty} w_{k+i} z^{-i} \end{aligned}$$

If we replace the new formulation of $W(z)$ in the recursive equation we obtain

$$\begin{aligned} y(t+k) &= \left(E(z) + z^{-k}\frac{F(z)}{A(z)}\right)e(t+k) \\ &= E(z)e(t+k) + z^{-k}\frac{F(z)}{A(z)}e(t+k) \\ &= E(z)e(t+k) + \frac{F(z)}{A(z)}e(t) \\ &= \sum_{i=0}^{k-1} w_i z^{-i} e(t+k) + \sum_{i=0}^{\infty} w_{k+i} z^{-i} e(t) \\ &= \sum_{i=0}^{k-1} w_i e(t+k-i) + \sum_{i=0}^{\infty} w_{k+i} e(t-i) \end{aligned}$$

As we can see, $y(t+k)$ can be divided in two parts:

- $E(z)e(t+k)$ is uncorrelated with the past and can't be predicted since we have only the values up to $e(t)$. Note that, since we are talking about white noise, these values are uncorrelated to the value at time t.
- $\frac{F(z)}{A(z)}e(t)$ depends only on the current time instant and is predictable depending on the information at time t.

Since the first part is not correlated to the current time instant and can't be predicted we can forget about it and obtain the optimal predictor from the noise.

Definition 17 (Optimal predictor from the noise)**.** *The optimal predictor from the noise for $y(t)$ is*

$$\hat{y}(t+k|t) = \frac{F(z)}{A(z)}e(t)$$

5.2.3 Optimal predictor from the output

Unfortunately, the past data related to the white noise isn't available because the white noise is something we use to model the complexity of the input, but we can't actually measure it. This means that the optimal predictor from the noise is a good theoretical result, however it can't be used in practice. To obtain a result practically useful we have to reconstruct the noise $e(t)$ from the output $y(t)$. This means that if we can write $e(t)$ as a function of $y(t)$, then we can replace it in the predictor from noise and obtain a predictor from the output.

$$\hat{y}(t+k|t)\big(e(t)\big) \xmapsto{e(t)=f(y(t))} \hat{y}(t+k|t)\big(f(y(t))\big)$$

If we want to write $e(t)$ as a function of $y(t)$, we can invert the recursive equation

$$y(t) = \frac{C(z)}{A(z)}e(t)$$

We know that $y(t)$ is well defined since $\frac{C(z)}{A(z)}$ is asymptotically stable (Theorem 5). If invert the transfer function $W(z)$ to obtain

$$e(t) = W^{-1}y(t) = \frac{A(z)}{C(z)}y(t)$$

we would also try to verify if $e(t)$ is well-defined. To prove this, according to Theorem 5, we have to prove that

- $y(t)$ is well defined.
- $W^{-1}(z) = \frac{A(z)}{C(z)}$ is asymptotically stable.

To prove this two properties, we have to remember the assumptions we did when starting analysing linear prediction. In particular, we assumed that

1. $y(t) = W(z)e(t)$ is a **canonical representation** of the process $y(t)$ (as defined by Theorem 11).
2. Every zero of the transfer function $W(z)$ is strictly inside the unitary circle.

$$|z_i| < 1 \quad \forall \text{ zeros } z_i$$

The first assumption can be used to prove that $y(t)$ is well-defined, in fact if $y(t)$ is in its canonical representation, then it's also well-defined. On the other hand, the second assumption can be used to prove that $W^{-1}(z)$ is asymptotically stable, in fact $W^{-1}(z)$ is asymptotically stable if and only if all its poles are smaller than 1. The assumption imposes that the zeros of $W(z)$, i.e., the roots of $C(z)$ have to be smaller than 1. However, $C(z)$ is at the denominator in $W^{-1}(z)$, hence the poles of $W^{-1}(z)$ are smaller than 1. Basically, we have just shown that $W^{-1}(z)$ is asymptotically stable.

Since $W^{-1}(z)$ is asymptotically stable and $y(t)$ is well-defined, we can write

$$e(t) = W^{-1}(z)y(t) = \frac{A(z)}{C(z)}y(t)$$

The process $e(t)$ can also be written as $MA(\infty)$ as

$$e(t) = \sum_{i=0}^{\infty} \breve{w}_i y(t-i)$$

Having obtained a representation of the white noise as a function of the output, we can replace it in the optimal predictor from the noise to obtain the optimal predictor from the output

$$\begin{aligned}
\hat{y}(t+k|t) &= \frac{F(z)}{A(z)}e(t) \\
&= \sum_{i=0}^{\infty} w_{k+i} z^{-i} e(t) \\
&= \sum_{i=0}^{\infty} w_{k+i} e(t-i) \\
&= \sum_{i=0}^{\infty} w_{k+i} \Big(\sum_{j=0}^{\infty} \breve{w}_j y(t-i-j) \Big)
\end{aligned}$$

Alternatively, we can replace directly $e(t) = \frac{A(z)}{C(z)}y(t)$, to derive

$$\begin{aligned}
\hat{y}(t+k|t) &= \frac{F(z)}{A(z)}e(t) \\
&= \frac{F(z)}{A(z)}\frac{A(z)}{C(z)}y(t) \\
&= \frac{F(z)}{C(z)}y(t)
\end{aligned}$$

It's important to remember that to obtain the predictor from the output we have to start from the canonical representation of $y(t)$. To sum things up, the optimal predictor from the output is

Definition 18 (Optimal linear predictor from the output)**.** *The optimal linear predictor from the output of a process $y(t) = W(z)e(t)$ in canonical form is*

$$\hat{y}(t+k|t) = \frac{F(z)}{C(z)}y(t) = \sum_{i=0}^{\infty} w_{k+i} \Big(\sum_{j=0}^{\infty} \breve{w}_j y(t-i-j) \Big)$$

Also notice that, since $y(t)$ is stationary and $\frac{F(z)}{C(z)}$ is asymptotically stable, then also the predictor, for Theorem 5 is a stationary stochastic process and $\hat{y}(t+k|t)$ is the steady-state output of a digital filter $\frac{F(z)}{C(z)}$ fed by $y(t)$. The fact that $\hat{y}(t+k|t)$ is stationary means that it's statistical properties (i.e., mean and covariance) don't depend on time, hence the predictor from the output can also be written as

$$\hat{y}(t|t-k) = \frac{F(z)}{C(z)}y(t-k)$$

The prediction $\hat{y}(t+k|t)$ is computed thanks to $y(t), \ldots, y(t-\infty)$, however, in practice we don't have all the values of the sequence. In other words, we only have a finite sequence that starts from when we start observing the process $y(1)$ and terminates at time t

$$y(t), \cdots, y(1)$$

In this conditions, we can compute the output $\tilde{\hat{y}}(t+k|t)$ of $\frac{F(z)}{C(z)}$ with an arbitrary initialisation at time $t = 1$ of the process (e.g., all zeros, like we did for MA processes), fed with $y(t)$. The output $\tilde{\hat{y}}(t+k|t)$ is sub-optimal, however it converges exponentially fast to the the optimal predictor $\hat{y}(t+k|t)$ and if t is large enough (which is typically true since t is a running time and the data set always increases) we can neglect the approximation.

$$\tilde{\hat{y}}(t+k|t) \approx_{t \gg 1} \hat{y}(t+k|t)$$

5.2.4 Optimal prediction error

Let $y(t) = W(z)e(t)$ be a canonical representation of an ARMA process, with $e(t) \sim WN(0, \lambda^2)$. If $\hat{y}(t+k|t)$ is an optimal linear predictor, we can write its error as

$$\varepsilon(t+k|t) = y(t+k) - \hat{y}(t+k|t)$$

To evaluate this error, we can replace the input process and the predictor with their $MA(\infty)$ representation. Doing so leads us to

$$\begin{aligned}
\varepsilon(t+k|t) &= y(t+k) - \hat{y}(t+k|t) \\
&= E(z)e(t+k) + \frac{F(z)}{A(z)}e(t) - \frac{F(z)}{C(z)}e(t) \\
&= E(z)e(t+k)
\end{aligned}$$

This result makes sense because since we choose the values of β_i to put to 0 the controllable part of the process to predict, the error we do is the part we can't control. If we expand the term $E(z)$ we obtain

$$\begin{aligned}
\varepsilon(t+k|t) &= E(z)e(t+k) \\
&= \sum_{i=0}^{k+1} w_i z^{-i} e(t+k) \\
&= \sum_{i=0}^{k+1} w_i e(t+k-i)
\end{aligned}$$

This means that the error $\varepsilon(t+k|t)$ of an optimal linear predictor is a $MA(k-1)$ process. Being a Moving Average process, we can easily compute its variance

$$V_\varepsilon = \sum_{i=0}^{k-1} c_i^2 \lambda^2$$

If we evaluate the variance for increasing values of k we notice that it increases.

$$\begin{aligned}
V_\varepsilon(k=1) &= (c_0^2 + c_1^2)\lambda^2 \\
V_\varepsilon(k=2) &= (c_0^2 + c_1^2 + c_2^2)\lambda^2 \\
V_\varepsilon(k=3) &= (c_0^2 + c_1^2 + c_2^2 + c_3^2)\lambda^2
\end{aligned}$$

This makes sense because we are trying to predict a value in the distant future $t+k$ without all the realisation between t and $t+k$.

Moreover, for $k \to \infty$, the variance goes to $\sum_{i=0}^{\infty} w_i^2 \lambda^2$ and since w_i are the coefficients of the $MA(\infty)$ representation of $y(t+k)$, then the variance of the error tends to the variance of the process to predict

$$V[\varepsilon(t+k|t)] \to_{k\to\infty} V[y(t+k)]$$

This happens because, as $k \to \infty$, the information on $y(t+k)$ carried by the previous values $y(t), y(t-1), \dots$ vanishes and the sole possible prediction for $y(t+k)$ is given by its mean value

$$\hat{y}(t+\infty) = \mathbb{E}[y(t+\infty)] = 0$$

If we use this value to compute the variance of ε we get

$$\begin{aligned} V[\varepsilon(t+\infty|t)] &= \mathbb{E}\big[\big(y(t+\infty) - \mathbb{E}[y(t+\infty)]\big)^2\big] \\ &= \mathbb{E}\big[\big(y(t+\infty) - 0)\big)^2\big] \\ &= V\big[y(t+\infty)\big] \\ &= V_y \end{aligned}$$

5.3 Prediction of non-zero mean processes

Up until now, we have considered only zero mean processes. Let us now consider a process

$$y(t) = W(z)e(t) = \frac{C(z)}{A(z)}e(t)$$

in canonical form and with

$$e(t) \sim WN(\mu, \lambda^2)$$

Before going on, let us immediately write the expected value of $y(t)$ using the Gain Theorem (7)

$$\begin{aligned} \mathbb{E}[y(t)] = \bar{y} &= W(1) \cdot \mathbb{E}[e(t)] \\ &= W(1) \cdot \mu \end{aligned}$$

The optimal predictor $\hat{y}(t+k|t)$ relies on the fact that the white noise is zero mean, however now the noise has mean μ. Luckily, we can use unbiased processes to go from non-zero mean processes to zero mean processes. Let us define the unbiased processes for $y(t)$

$$\begin{aligned} \tilde{y}(t) &= y(t) - \mathbb{E}[y(t)] \\ &= y(t) - \bar{y} \\ &= y(t) - W(1)\mu \end{aligned}$$

and for $e(t)$

$$\begin{aligned} \tilde{e}(t) &= e(t) - \mathbb{E}[e(t)] \\ &= e(t) - \mu \end{aligned}$$

Thanks to these processes we can write

$$\tilde{y}(t) = W(z)\tilde{e}(t) = \frac{C(z)}{A(z)}\tilde{e}(t)$$

Since $\tilde{y}(t)$ is zero-mean, we can apply to it the theory seen so far. In particular we can predict the values of $\tilde{y}(t+k|t)$ using the optimal linear predictor from the output

$$\widehat{\tilde{y}(t+k|k)} = \frac{F(z)}{C(z)}\tilde{y}(t)$$

Notice that this predictor uses the measurements $\tilde{y}(t), \tilde{y}(t-1), \dots$, which are equivalent to the measurements $y(t), y(t-1), \dots$ because

$$\tilde{y}(t-i) = y(t-i) - m_y$$

The linear predictor $\hat{\tilde{y}}(t+k|k)$ predicts the values of $\tilde{y}$, however we would like to predict the values of $y(t)$. In other words we would like to build the predictor $\hat{y}(t+k|t)$. To achieve this goal, we can write $y(t)$ from $\tilde{y}(t) = y(t) - m_y$

$$y(t) = \tilde{y}(t) + m_y$$

and remember that the mean is the same for all time instants t. This means that

$$y(t+k) = \tilde{y}(t+k) + m_y$$

Starting from this observation, we can build the optimal predictor as

$$\begin{aligned}
\hat{y}(t+k|t) &= \widehat{\tilde{y}(t+k|t)} + m_y \\
&= \hat{\tilde{y}}(t+k|t) + m_y \\
&= \frac{F(z)}{C(z)}\tilde{y}(t) + m_y \\
&= \frac{F(z)}{C(z)}\big(y(t) - m_y\big) + m_y \\
&= \frac{F(z)}{C(z)}y(t) - \frac{F(z)}{C(z)}m_y + m_y \\
&= \frac{F(z)}{C(z)}y(t) + \left(1 - \frac{F(z)}{C(z)}\right)m_y \\
&= \frac{F(z)}{C(z)}y(t) + \left(1 - \frac{F(1)}{C(1)}\right)m_y
\end{aligned}$$

The last step is obtained applying the Gain Theorem.

5.3.1 Prediction of an ARMAX process

Applying the same reasoning seen for ARMA processes, we can compute the optimal linear predictor of an ARMAX process fed by a zero-mean white noise

$$y(t) = \frac{B(z)}{A(z)}u(t-d) + \frac{C(z)}{A(z)}e(t)$$

as

$$\hat{y}(t+k|t) = \frac{B(z)E(z)}{C(z)}u(t+k-d) + \frac{F(z)}{C(z)}y(t)$$

Part III

Model Identification

Chapter 6

Model Identification

6.1 Introduction

Up until now, we have described and analysed stochastic processes $y(t)$ and used them to predict future values. However, we haven't still said where does the model come from. Basically, we have always written $y(t) = W(z)e(t)$, taking the transfer function $W(z)$ for granted. The problem of obtaining a model for $y(t)$ is called **model identification**.

The idea of model identification is to obtain a model from the data (i.e., the experiments on the real system), which is our primary source of knowledge. Better said, we have a system S (as in Figure 6.1) and we want to describe it as an ARMAX process. The system's input and outputs can be observed and measured to obtain a sequence of measurements

$$y(1), \ldots, y(k) = \big\{y(t)\big\}$$

and

$$u(1), \ldots, u(k) = \big\{u(t)\big\}$$

Model identification provides automatic ways to map the observations $\big\{y(t)\big\}$ and $\big\{u(t)\big\}$ into the equations needed for the ARMAX representation (as in Figure 6.2) of the system S.

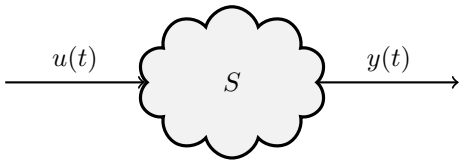


Figure 6.1: An I/O system S, its output $y(t)$ and input $u(t)$ of interest.

6.2 Parametric methods

An ARMAX model can be described by

- Its statistical properties m_y, $\gamma_y(\tau)$, $\Gamma_y(\omega)$. This description is said to be **non-parametric**.

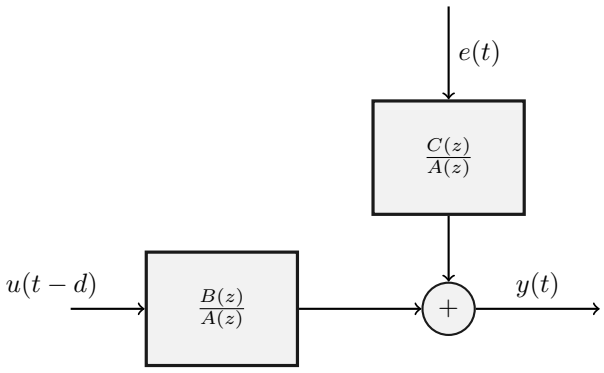


Figure 6.2: An ARMAX model of S obtained from the observations of S's inputs and outputs.

- The coefficients of $A(z)$, $B(z)$ and $C(z)$. This characterisation is said to be **parametric**.

Both characterisations can be derived from the input data, however we will focus on the parametric one. The identification of a model using the parametric method can be divided in three phases

1. **Experiment design and data collection**.
2. **Selection of a parametric model class**

   $$\mathcal{M}(\vartheta) = \{M(\vartheta), \vartheta \in \Theta\}$$

   where ϑ is a vector of parameters and each vector $\vartheta \in \Theta$ corresponds to a different model.
3. **Choice of the identification criterion** $J_N(\vartheta) \geq 0$. The identification criterion measures the performance of the model corresponding to ϑ in describing the collected data. The best model is the one that minimises $J_N(\vartheta)$.
4. **Minimization of** $J_N(\vartheta)$ **with respect to** ϑ.
5. **Model validation**. Once the optimal model has been obtained, we verify whether this model is actually a good one. If it is not, the identification process must be repeated.

6.2.1 Experiment design and data collection

In this phase we have to collect the values of $y(t)$ and $u(t)$. The main problems we can encounter when measuring the inputs and the outputs are

- The number N of samples we should collect.
- How we should design the input $u(t)$.

6.2.2 Choice of the parametric model class

Given a system S, many different models and model classes can describe it. After collecting data, we should define which class of models we want to use to describe our system. A model class $\mathcal{M}(\theta)$

is a set of parametric models $M(\theta)$ of parameter θ. In our case we want to use ARMAX processes to model the system of interest, hence the models we want to use are

$$\begin{aligned} M(\vartheta) : y(t) &= \frac{B(z)}{A(z)}u(t-d) + \frac{C(z)}{A(z)}e(t) \\ &= \frac{\sum_{i=0}^{p} b_i z^{-i}}{\sum_{i=0}^{m} a_i z^{-i}}u(t-d) + \frac{\sum_{i=0}^{n} c_i z^{-i}}{\sum_{i=0}^{m} a_i z^{-i}}e(t) \end{aligned}$$

where $e(t) \sim WN(0, \lambda^2)$. Since we want to obtain the parameters a_i, b_i and c_i of the model above, the parameters vector is

$$\vartheta = \begin{bmatrix} a_0, \dots, a_m, b_0, \dots, b_p, c_0, \dots, c_n \end{bmatrix}^T$$

Notice that the variance λ^2 of the white noise should be a parameter of the model, too, since we don't observe the white noise and we con only estimate it. However, it has a marginal impact on the model, hence we can, for now leave it apart. Now that we know the shape of the shape of the parameters vector, we can rewrite the models $M(\vartheta)$ as

$$M(\vartheta) : y(t) = \frac{B(z,\vartheta)}{A(z,\vartheta)}u(t-d) + \frac{C(z,\vartheta)}{A(z,\vartheta)}e(t)$$

This representation describes a family of ARMAX models whose coefficients $a_i(\vartheta)$, $b_i(\vartheta)$, $c_i(\vartheta)$ of the polynomials A, B and C depend on the parameter vector ϑ.

To perform identification, we will rely on the theory of prediction, hence, we have to assume that for every $\vartheta \in \Theta$, the stochastic part of $M(\vartheta)$ (i.e., the part depending on the white noise $e(t) \sim WN(0, \lambda^2)$) is canonical and has no zeroes on the unit circle. This is not a big issue, in fact

- Given a non-canonical representation we can always obtain its canonical representation.
- Zeroes on the unit circle are not usually required to model the behavior of a given system and the behavior of models with zeroes on the unit circle can be approximate by means of models with zeroes close to the unit circle.

System order

Until now, the parameters vector θ contains only the parameters of $A(z)$, $B(z)$ and $C(z)$ and the variance of $e(t)$, however it doesn't contains the orders n, m, p of the system and the time delay d of the input. This means that, for now we are considering them as fixed values that can even be considered null to use a different model (e.g., we can put $b_i = 0$ to obtain an ARMA model).

Black box identification

When we designed the model class to use for identifying a model for S, we didn't add any a-priori information about the system and we simply described it as a general ARMAX process (i.e., we have defined the general class of model but not a specific shape). This approach is called **black-box approach** because no a-priori information about the system is available and the parametrisation is completely free. The black-box approach is counterposed to the grey-box approach. In this case, some information about the the system is known. For instance, a grey-box approach could define a parameters vector

$$\vartheta = \begin{bmatrix} k \\ \tau \end{bmatrix}$$

and a model

$$y(t) = \frac{k + k^{\tau} z^{-1}}{1 - \tau^2 z^{-2}} u(t-d) + \frac{k^3 k^4 z^{-1}}{1 - \tau^2 z^{-2}} e(t)$$

As we can see, we have added some information to the model, which is less generic.

6.2.3 Choice of the identification criterion

After defining the class of models $M(\vartheta)$ we want to use, we have to find a criterion $J(\vartheta)$ to measure how good a parameter vector ϑ is, i.e., how good ϑ is at describing the collected data $\{u(t)\}$ and $\{y(t)\}$.

To measure how good can ϑ describe the collected data we can notice that the sequence of measurements $\{u(t)\} \cup \{y(t)\}$ is a sequence of known values while the model class $M(\vartheta)$ is a stochastic model (it depends on s, even if it's not written). To understand why this observation is useful, we can write the predictor from the output of $M(\theta)$

$$\hat{y}(t|t-k) = \frac{B(z,\vartheta)E(z,\vartheta)}{C(z,\vartheta)} u(t-d) + \frac{F(z,\vartheta)}{C(z,\vartheta)} y(t-k)$$

If we consider $k = 1$, we can fed the optimal predictor $\hat{y}(t|t-1)$ with the values of $y(t-1)$ and $u(t-d)$, which are part of the sequence of observed values, and compare the prediction with the actual value of the output $y(t)$, which is also part of the known collected data. A visual representation of the identification process that we have just described is shown in Figure 6.3.

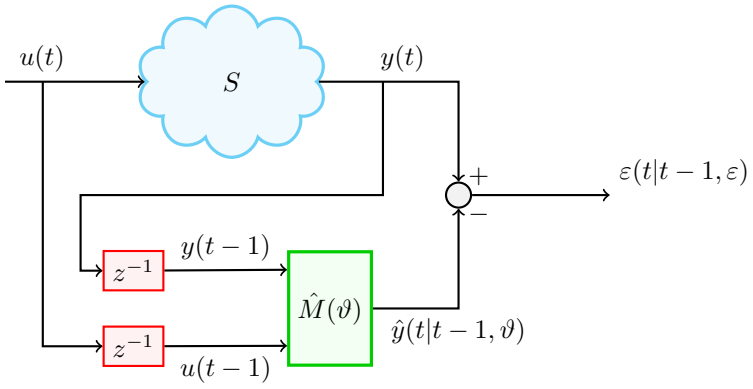


Figure 6.3: A visual representation of the process used to compute the error from the observed data and the predictor from the output.

An important thing to highlight is that the predictor doesn't contain the variance of the white noise. That's the reason why we decided to exclude λ^2 from the parameters vector ϑ.

Prediction Error Minimisation

Given the qualitative description above, we can obtain the Prediction Error Minimisation (PEM) identification criterion.

Definition 19 (Prediction Error Minimisation (PEM) criterion)**.** *The identification criterion $J_N(\vartheta)$ for a ARMAX model class $\mathcal{M}(\vartheta)$ is the empirical variance of the prediction error $\varepsilon(t|t-$*

$1, \vartheta)$

$$J_N(\vartheta) = \frac{1}{N}\sum_{i=1}^{N}\Big(y(i) - \hat{y}(i|i-1,\vartheta)\Big)^2$$
$$= \frac{1}{N}\sum_{i=1}^{N}\big(\varepsilon(i|i-1,\vartheta)\big)^2$$

Now that we have defined the criterion to evaluate ϑ, we can use it to compute the best parameter vector ϑ. In particular, the vector ϑ that better describes the data input data is the which minimum error (i.e., which predictions are closer to the actual data observed). More formally,

Definition 20 (Optimal PEM parameter vector)**.** *The optimal $\hat{\vartheta}_N$ is the value of $\vartheta \in \Theta$ that minimises the empirical variance of the prediction error.*

$$\hat{\vartheta}_N = arg\min_{\vartheta\in\Theta} J_N(\vartheta)$$
$$= arg\min_{\vartheta\in\Theta} \frac{1}{N}\sum_{i=1}^{N}\big(\varepsilon(i|i-1,\vartheta)\big)^2$$

Noise variance

Now that we have computed the optimal value for ε, we can reveal that the optimal value of the white noise's variance is

$$\hat{\lambda}_N^2 = J_N(\hat{\vartheta}_N) = \frac{1}{N}\sum_{i=1}^{N}\varepsilon(i,\hat{\vartheta}_N)^2$$

6.2.4 Minimisation of the identification criterion

Now that we know that we have to minimise the identification criterion $J_N(\vartheta)$, we should understand how to practically do it. In particular,

- The identification criterion of AR and ARX processes is a quadratic function of ϑ. Since the criterion is a quadratic function, we can explicitly compute the minimum, which is also unique.
- The identification criterion of $ARMA$, $ARMAX$ and AR processes is a non quadratic function of ϑ. Since the criterion is a non-quadratic function, we can't compute explicitly the minimum and we should use iterative methods that start from an hypothesis and keep improving it. Such methods guarantee that we reach a minimum, however we can't tell if it's the global minimum or just a local one.

This makes a big difference when we optimise (i.e., minimise) the identification criterion $J_N(\vartheta)$.

AR and ARX processes

Let's start by analysing the how to compute the minimum of the identification criterion $J_N(\vartheta)$ for an AR or ARX process. First things first, let's write the recursive equation for a generic ARX process

$$y(t) = \frac{B(z)}{A(z)}u(t-d) + \frac{1}{A(z)}e(t)$$

Since the polynomial $C(z)$ is not there in an ARX process, we can rewrite the parameters vector as

$$\vartheta = \left[a_0, \dots, a_m, b_0, \dots, b_p\right]^T$$

which is a vector of $n_\vartheta = p + m$ dimensions. The recursive equation can be rewritten in the form we initially used to introduce AR process to obtain

$$\begin{aligned} A(z)y(t) &= B(z)u(t-d) + e(t) \\ y(t) &= \sum_{i=1}^{m} a_i z^{-i} y(t) \sum_{i=0}^{p} b_i z^{-i} u(t-d) + e(t) \\ &= \sum_{i=1}^{m} a_i y(t-i) \sum_{i=0}^{p} b_i u(t-d-i) + e(t) \end{aligned}$$

If we write the variables $y(t-i)$ and $u(t-d-i)$ in vector form, we obtain a vector (of dimension $n_\phi = m + p$)

$$\phi(t) = \left[y(t-1), \dots, y(t-m), u(t-d), \dots, u(t-d-p)\right]^T$$

This means that $y(t)$ can be rewritten in vector form as

$$y(t) = \vartheta^T \phi(t) + e(t)$$

Now we can notice that $\phi(t)$ depends only on the quantities observable at time $t-1$, hence it's fully predictable, while $e(t)$, being a white noise, is fully unpredictable, hence we can forget about it. What we obtain is

$$y(t) = \vartheta^T \phi(t) = \phi(t)^T \vartheta$$

Since $y(t)$ is fully predictable at time $t-1$, we can write

$$M(\vartheta) : \hat{y}(t|t-i, \vartheta) = \vartheta^T \phi(t)$$

Now that we have found the shape of the predictor, we can replace it in the identification criterion to obtain

$$\begin{aligned} J_N(\vartheta) &= \frac{1}{N} \sum_{i=1}^{N} \left(y(i) - \hat{y}(i|i-1, \vartheta) \right)^2 \\ &= \frac{1}{N} \sum_{i=1}^{N} \left(y(i) - \vartheta^T \phi(t) \right)^2 \end{aligned}$$

Since $\vartheta^T \phi(t)$ is linear in ϑ, then, if we square it, we obtain a quadratic function that can be minimised computing putting the first derivative to 0

$$\frac{dJ_N(\vartheta)}{d\vartheta} = 0 \tag{6.1}$$

and imposing that the second derivative is not negative

$$\frac{d^2 J_N(\vartheta)}{d\vartheta}^2 \geq 0 \tag{6.2}$$

Since ϑ is a vector, the derivative of $J_N(\vartheta)$ is also a vector of partial derivatives for all elements of ϑ.

$$\frac{dJ_N(\vartheta)}{d\vartheta} = \begin{bmatrix} \frac{\partial J_N(\vartheta)}{\partial \vartheta_1} \\ \vdots \\ \frac{\partial J_N(\vartheta)}{\partial \vartheta_{n_\vartheta}} \end{bmatrix}$$

First derivative Let us now compute this vector.

$$\begin{aligned}
\frac{dJ_N(\vartheta)}{d\vartheta} &= \frac{d}{d\vartheta}\frac{1}{N}\sum_{i=1}^{N}\left(y(i)-\phi(t)^T\vartheta\right)^2 \\
&= \frac{1}{N}\sum_{i=1}^{N}\frac{d}{d\vartheta}\left(y(i)-\phi(t)^T\vartheta\right)^2 \\
&= \frac{1}{N}\sum_{i=1}^{N}2(y(i)-\phi(t)^T\vartheta)\frac{d}{d\vartheta}\left(y(i)-\phi(t)^T\vartheta\right) \\
&= \frac{1}{N}\sum_{i=1}^{N}2(y(i)-\phi(t)^T\vartheta)\frac{d}{d\vartheta}(-\phi(t)) \\
&= -\frac{2}{N}\sum_{i=1}^{N}\phi(t)(y(i)-\phi(t)^T\vartheta)
\end{aligned}$$

Notice that the derivative of $y(i)-\phi(t)^T\vartheta$ is $-\phi(t)$ because ϑ is linear and $\frac{d}{d\vartheta_i}\vartheta_i\phi_i(t)=\phi_i$. If we put the derivative to 0 we obtain

$$\begin{aligned}
\frac{dJ_N(\vartheta)}{d\vartheta} &= 0 \\
-\frac{2}{N}\sum_{t=1}^{N}\phi(t)(y(t)-\phi(t)^T\vartheta) &= 0 \\
\sum_{t=1}^{N}\phi(t)(y(t)-\phi(t)^T\vartheta) &= 0 \\
\sum_{t=1}^{N}\phi(t)y(t)-\sum_{t=1}^{N}\phi(t)\phi(t)^T\vartheta &= 0 \\
\sum_{t=1}^{N}\phi(t)y(t) &= \sum_{t=1}^{N}\left(\phi(t)\phi(t)^T\right)\vartheta
\end{aligned}$$

The equations in the system of n_ϑ equations in n_ϑ unknowns

$$\sum_{t=1}^{N}\left(\phi(t)\phi(t)^T\right)\vartheta=\sum_{t=1}^{N}\phi(t)y(t) \tag{6.3}$$

are called Least Squares (LS) normal equations. If the matrix $\sum_{t=1}^{N}\left(\phi(t)\phi(t)^T\right)$ is **non singular** (otherwise we couldn't invert it), the system admits one solution, which is the optimal value of the vector ϑ. Let us analyse a little more in depth the shape of the Least Squares normal equations. Let's start by recalling what are the components of the equations.

- ϑ is a column vector of n_ϑ elements.
- $y(t)$ is a scalar.
- $\phi(t)$ is a column vector of dimensions n_ϑ.

Let's now analyse the right hand-side of the equation. If we multiply $\phi(t)$ for $y(t)$ we get another vector of dimensions n_ϑ. Moreover, since the sum is applied to each element of the vector $\phi(t)$, the whole term $\sum_{t=1}^{N} \phi(t)y(t)$ is a column vector of dimensions n_ϑ.

Focusing on the left hand-side, we have that $\phi(t)\phi(t)^T$ is the product between a $n_\vartheta \times 1$ matrix and a $1 \times n_\vartheta$ matrix, hence we get a $n_\vartheta \times n_\vartheta$ matrix.

$$\begin{bmatrix} \sum_{t=1}^{N} y(t-1)y(t-1) & \dots & \sum_{t=1}^{N} y(t-1)u(t-q) \\ \vdots & \ddots & \vdots \\ \sum_{t=1}^{N} y(t-m)y(t-1) & \dots & \sum_{t=1}^{N} y(t-m)u(t-q) \\ \sum_{t=1}^{N} u(t-d)y(t-1) & \dots & \sum_{t=1}^{N} u(t-d)u(t-q) \\ \vdots & \ddots & \vdots \\ \sum_{t=1}^{N} u(t-q)y(t-1) & \dots & \sum_{t=1}^{N} u(t-q)u(t-q) \end{bmatrix} \begin{bmatrix} a_1 \\ \vdots \\ a_m \\ b_1 \\ \vdots \\ v_p \end{bmatrix} = \begin{bmatrix} \sum_{\tau=1}^{N} \phi(\tau-1)y(\tau) \\ \vdots \\ \sum_{\tau=1}^{N} \phi(\tau-m)y(\tau) \\ \sum_{\tau=1}^{N} u(\tau-d)y(\tau) \\ \vdots \\ \sum_{\tau=1}^{N} u(\tau-q)y(\tau) \end{bmatrix} \tag{6.4}$$

Second derivative Finally, we have to check if the second derivative is semi-definite positive. Let's start by computing the second derivative of $J_N(\vartheta)$

$$\begin{aligned} \frac{d^2 J_N(\vartheta)}{d\vartheta^2} &= \frac{dJ'_N(\vartheta)}{d\vartheta} \\ &= \frac{2}{N} \sum_{t=1}^{N} \phi(t)\phi(t) \end{aligned}$$

To check if the matrix $\frac{2}{N} \sum_{t=1}^{N} \phi(t)\phi(t)$ is semi-definite positive we have to verify that

$$x^T \cdot \frac{2}{N} \sum_{t=1}^{N} \phi(t)\phi(t) \cdot x$$

is always not negative for all non null vectors x. If we take the vectors x and x^T inside the summation we get

$$\frac{2}{N} \sum_{t=1}^{N} x^T \cdot \phi(t)\phi(t) \cdot x$$

Now we can remember that $x^T y = y^T x$, hence we obtain

$$\frac{2}{N} \sum_{t=1}^{N} x^T \cdot \phi(t) \cdot x^T \cdot \phi(t)$$

$$\frac{2}{N} \sum_{t=1}^{N} \left(x^T \cdot \phi(t) \right)^2$$

which means that the point obtained from the Least Squares (LS) normal equations is always a minimum. Given this result, if the matrix $\sum_{t=1}^{N} \left(\phi(t)\phi(t)^T \right)$ is not singular, we get an infinite number of points which are however equivalent for prediction purposes. This situation happens if

- The data record was not representative enough of the underlying physical phenomenon.
- The chosen model class was too complex and there are equivalent models for describing the same phenomenon.

ARMAX processes

In a general ARMAX process, the optimal predictor has $C(z)$ at the denominator, hence the function to minimise is not quadratic and might have multiple local minima. To solve this problem we have to use an iterative approach

1. The algorithm is initialized with an initial estimate (typically, randomly chosen) of the optimal parameter vector ϑ^1.
2. The hypothesis is refined using a function $f(\vartheta)$ to obtain a new more precise estimate $\vartheta^{n+1} = f(\vartheta^n)$.
3. The sequence of estimates should converge to the optimal parameter vector $\hat{\vartheta}_N$.

This process guarantees to reach a minimal which could be local. To solve this problem we can repeat the algorithm multiple times starting from different initialisation ϑ^1 and choose the minimum among the values of $\hat{\vartheta}_N$ we have obtained in different iterations. Even this techniques doesn't guarantee to reach a global optimum, but it can help. Moreover, the more times we repeat the algorithm, more chances we have to find the global optimum.

6.2.5 Model validation

The last part of the model identification phase is called model validation and allows to measure the performance of the model and understand if the model we have chosen is useful or not. If the model isn't performing good enough, we have to go back to the previous step and change the model class. In particular, we can choose to change

- The **model class** $\mathcal{M}(\vartheta)$.
- The **model orders** n, m and p. Notice that, changing the model order also means selecting a specific model (e.g., with $p = 0$ we are selecting ARMA models).
- The **delay** d.

For now, let us assume $m = n = p \neq 0$, to find some methods to evaluate the performance of the model. This means that writing $ARMAX(m)$ is equivalent to writing $ARMAX(m, n = m, p = m, d = m)$.

Evaluation form the indicator

The first idea one could have to compute the performance of a model $M(\hat{\vartheta})$ could be to use the indicator $J_N(\vartheta)$. The idea is to consider different values of m and take the model with the lowest indicator (i.e., the one whose predictor has the lowest error with respect to the data observed). The algorithm is shown in 1. The problem with this algorithm is that it uses the error computed on the same data we used for the predictor. This means that we select the model that fits better the observed data but not the actual system we want to analyse. This behaviour is called overfitting and

Algorithm 1 Model evaluation from the indicator J_N.

$\hat{\vartheta}_N = \infty$
$J_N = \infty$
for $m = 1$ to $\hat{m}$ **do**
  $\mathcal{M}^m = ARMAX(m)$
  $\vartheta_N^m = arg\min_{\vartheta \in \theta} \frac{1}{N} \sum_i \varepsilon(i, \vartheta)$
  if $J_N(\vartheta_N^m) \leq J_N$ **then**
    $J_N \leftarrow J_N(\vartheta_N^m)$
    $\hat{\vartheta}_N \leftarrow \vartheta_N^m$
  end if
end forreturn $\hat{\vartheta}_N$

we say that a model has high variance if it precisely resembles the behaviour of the data, but can't generalise and capture the behaviour of the system. Practically, the algorithm chooses the highest order possible because, with an higher order, the linear predictor is a function that fluctuates a lot and it can, in the extreme case, pass for all observed data points.

Cross-validation

The problem with Algorithm 1 is that it's evaluated against the same data it's built on. To solve this problem we can split the observed data, that contains data with indices from 1 to N, in two sets

- The **training data**, with indices from 1 to M. The training data is used to compute the best parameter vector $\hat{\vartheta}_N$.
- The **validation data**, with indices from $M + 1$ to N. The validation data is used to evaluate the parameter vector obtained from the training data.

The cross validation algorithm is shown in Figure 2. This approach has a lower variance with

Algorithm 2 Cross-validation.

$\hat{\vartheta} = \infty$
$J = \infty$
for $m = 1$ to $\hat{m}$ **do**
  $\mathcal{M}^m = ARMAX(m)$
  $\vartheta_T^m = arg\min_{\vartheta \in \theta} \frac{1}{M} \sum_i^M \varepsilon(i, \vartheta)$ ▷ Computed using the training set
  $J_V(\vartheta_T^m) = \frac{1}{N-M} \sum_{i=M+1}^{N} \left(y(i) - \hat{y}(i|i-1, \vartheta_T^m)\right)^2$
  if $J_V(\vartheta_T^m) \leq J$ **then**
    $J \leftarrow J_V(\vartheta_T^m)$
    $\hat{\vartheta} \leftarrow \vartheta_T^m$
  end if
end forreturn $\hat{\vartheta}$

respect to the previous one (i.e., it doesn't suffer from overfitting), however, since some data is used for validation, the model itself is trained on less data, hence we could select a model that can't really

capture all the characteristics of the system. This behaviour is called **underfitting** and is generated by an high bias. Deciding how to split the data set is therefore very important.

Direct model penalisation

A way to balance variance and bias is to use verification algorithms that automatically penalise models with a lot of parameters (i.e., with a large m). The idea of such algorithms is to use the whole data-set to train the model (i.e., to compute $\hat{\vartheta}_N$) and then use a penalised version of $J_N(\vartheta)$ to compute the best model. The general version of the algorithm is shown in Algorithm 3. Some of

Algorithm 3 Model penalisation algorithm.

$\hat{\vartheta}_N = \infty$
$V = \infty$
for $m = 1$ to $\hat{m}$ **do**
  $\mathcal{M}^m = ARMAX(m)$
  $\hat{\vartheta}_N^m = arg\min_{\vartheta\in\theta} \frac{1}{N}\sum_i \varepsilon(i,\vartheta)$
  $V(m) = V(m, J_N(\hat{\vartheta}_N^m))$
  if $V(m) \leq V$ **then**
    $V \leftarrow V(m)$
    $\hat{\vartheta}_N \leftarrow \hat{\vartheta}_N^m$
  end if
end forreturn $\hat{\vartheta}_N$

the most used functions $V(m)$ are

- **Final Prediction Error** FPE.

$$FPE\big(m, J_N(\hat{\vartheta}_N^m)\big) = \frac{N+m}{N-m} \cdot J_N(\hat{\vartheta}_N^m)$$

  It's clear that, the bigger the value of m, the bigger is the numerator and the smaller the denominator, hence the value of V gets bigger.

- **Akaikes Identification Criterion** AIC.

$$AIC\big(m, J_N(\hat{\vartheta}_N^m)\big) = 2\frac{m}{N}\ln\left(J_N(\hat{\vartheta}_N^m)\right)$$

  Thanks to the m at the numerator, the value of AIC increases proportionally to m, hence models with a big number of parameters are penalised.

- **Minimum Description Length** MDL.

$$MDL\big(m, J_N(\hat{\vartheta}_N^m)\big) = \ln(N)\frac{m}{N} + \ln\left(J_N(\hat{\vartheta}_N^m)\right)$$

It's possible to demonstrate that the FPE and AIC functions are equivalent, in fact, if we apply the natural logarithm to the FPE function we obtain

$$\begin{aligned}\ln(FPE) &= \ln\left(\frac{N+m}{N-m}\right) + \ln\left(J_N(\hat{\vartheta}_N^m)\right) \\ &= \ln\left(\frac{N(1+\frac{m}{N})}{N(1-\frac{m}{N})} \cdot J_N(\hat{\vartheta}_N^m)\right) \\ &= \ln\left(\frac{1+\frac{m}{N}}{1-\frac{m}{N}} \cdot J_N(\hat{\vartheta}_N^m)\right) \\ &= \ln\left(1+\frac{m}{N}\right) - \ln\left(1-\frac{m}{N}\right) + \ln\left(J_N(\hat{\vartheta}_N^m)\right)\end{aligned}$$

If we assume that the number of samples is much bigger than the number of parameters of the model, i.e. $N \gg m$, then $\frac{m}{N} \approx 0$, hence we can approximate the logarithm as

$$\ln\left(1+\frac{m}{N}\right) \approx \frac{m}{N}$$

If we replace this result in the equation above we obtain

$$\begin{aligned}\ln(FPE) &= \ln\left(1+\frac{m}{N}\right) - \ln\left(1-\frac{m}{N}\right) + \ln\left(J_N(\hat{\vartheta}_N^m)\right) \\ &\approx \frac{m}{N} - \left(-\frac{m}{N}\right) + \ln\left(J_N(\hat{\vartheta}_N^m)\right) \\ &\approx 2\frac{m}{N} + \ln\left(J_N(\hat{\vartheta}_N^m)\right)\end{aligned}$$

This means that, the function AIC can be approximated with the logarithm of the FPE function and the both function lead to the same result. Also notice that, applying the logarithm to the FPE function isn't a problem since we are interesting in computing the minimum, hence if we apply a function to the function to minimise what changes is the value of the minimum but not the point where it is.

Moreover, we can also notice that the MDL function is very similar to the AIC function and the only difference is that in the latter function, the term $\ln(N)$ is replaced with a 2. Since N is usually big, MDL penalises the models more than AIC (and FPE).

6.3 Asymptotic analysis of PEM identification

Model identification allows to find a model $M(\hat{\vartheta}_N)$ for a system S. It would be interesting to find out if $M(\hat{\vartheta}_N)$ is in fact a good one. Finding an answer to this problem for a finite number of samples N is hard, however we can try and do it for $N \to \infty$. Before starting, let us formally define the system S we are going to analyse. The system S has an input

$$u(t) = F(z)r(t) + S(z)e(t)$$

and an output

$$y(t) = G(z)u(t) + H(z)e(t)$$

Both $u(t)$ and $y(t)$ are stationary stochastic processes, $e(t)$ and $r(t)$ are white noises with zero-mean and variance λ^2 and σ^2, respectively.

The measured data sequence corresponds to a particular realization of input/output signals of S. Basically, the sequences we have collected are evaluated in $s = \bar{s}$

$$\{y(1, \bar{s}), \dots, y(N, \bar{s})\}$$
$$\{u(1, \bar{s}), \dots, u(N, \bar{s})\}$$

This means that also the predictions $\{\hat{y}(t|t-1, \vartheta, \bar{s})\}$, the errors $\{\varepsilon(t, \vartheta, \bar{s})\}$ and the parameters $\{\vartheta(\bar{s})\}$ depend on the random variable $\bar{s}$. If we collect data different sets of data with different realisations of s (i.e., we collect a data set for $\bar{s}$, one for $\bar{\bar{s}}$ and so on) we obtain indicators J_N with very different behaviours. However, if we consider $N \to \infty$ we notice that all the J_N shrink to a single asymptotic curve. Practically, this means that, if N is large enough, our model doesn't depend on the realisation of s. This result is summed up in the following theorem

Theorem 12 (Empirical variance convergence)**.** *The performance index $J(\vartheta, s)$, i.e., the empirical variance of the prediction error of a linear predictor, tends to the actual variance of the error, when the number of samples goes to infinity ($N \to \infty$).*

$$J(\vartheta, s) \xrightarrow[N\to\infty]{} \bar{J}(\vartheta) = \mathbb{E}[\varepsilon(t, \varepsilon)^2]$$

Moreover, by letting

$$\Delta = \{\vartheta^* : J(\vartheta^*) \leq J(\vartheta), \forall \vartheta\}$$

be the set of global minimum points of $J(\vartheta)$, we have, that

$$\hat{\vartheta}_N(s) \xrightarrow[N\to\infty]{} \Delta$$

In other words, $\bar{J}(\vartheta)$ is the asymptotic counterpart of $J(\vartheta_N)$. As a corollary, we have that

Theorem 13 (Indicator convergence)**.** *If $\Delta = \{\theta^*\}$, i.e., the indicator $\hat{J}_N(\varepsilon)$ has a unique minimum point, then the parameter vector converges to that point*

$$\hat{\vartheta}_N(s) \xrightarrow[N\to\infty]{} \vartheta^*$$

This means that, the distance between the optimal parameter vector $\vartheta_N(s)$ and Δ tends to 0, as N goes to infinity.

The theorem says that the result of PEM identification is the same, independently of the measured realizations of process, as long as N is large enough. This means that, instead of studying the quality of the identified models for a finite N, we can easily evaluate their asymptotic quality, i.e. the quality of $M(\vartheta^*)$. If the model is asymptotically good, then we can infer that $M(\hat{\vartheta}_N)$ is also good, as long as N is large enough.

PEM approximation capabilities

Let's assume that the system S we are trying to model belongs to the class of models $\mathcal{M}(\vartheta)$ that we choose. This means that it exists a parameter vector ϑ^0 for which the model $M(\vartheta^0)$ behaves just like the system does (the model has the same structure of the system).

$$\exists \vartheta^0 \in \theta : M(\vartheta^0) = S$$

If vector ϑ^0 belongs to the set Δ of minimum points of $\bar{J}(\vartheta)$, then we can say, thanks to Theorem 12 that our identification method is able to retrieve asymptotically the true parameterisation of the data generating system S. To demonstrate that $\vartheta^0 \in \Delta$ we can start from remembering that the error $\varepsilon(t, \vartheta)$ is

$$\varepsilon(t, \vartheta) = y(t) - \hat{y}(t|t-1, \vartheta)$$

If we add and subtract the predictor $\hat{y}(t|t-1, \vartheta^0)$ in the right-hand side, we obtain

$$\varepsilon(t, \vartheta) = y(t) - \hat{y}(t|t-1, \vartheta^0) + \hat{y}(t|t-1, \vartheta^0) - \hat{y}(t|t-1, \vartheta)$$

In the new equation we see that the first two addends are actually the error of the predictor that uses the true parameter vector ϑ^0, hence the optimal prediction error, which is equal to the white noise $e(t)$, since it is the only non controllable part.

$$y(t) - \hat{y}(t|t-1, \vartheta^0) = \varepsilon(t, \vartheta^0) = e(t)$$

This means that we can rewrite $\varepsilon(t, \vartheta)$ as

$$\varepsilon(t, \vartheta) = e(t) + \hat{y}(t|t-1, \vartheta^0) - \hat{y}(t|t-1, \vartheta)$$

If we compute the variance of the prediction error $\varepsilon(t, \vartheta)$, which is $\bar{J}_N(\vartheta)$, we get

$$\begin{aligned}
\mathbb{E}\Big[\varepsilon(t, \vartheta)^2\Big] = \bar{J}_N(\vartheta) &= \mathbb{E}\Big[\big(e(t) + \hat{y}(t|t-1, \vartheta^0) - \hat{y}(t|t-1, \vartheta)\big)^2\Big] \\
&= \mathbb{E}\Big[\big(e(t)\big)^2\Big] + \mathbb{E}\Big[\big(\hat{y}(t|t-1, \vartheta^0) - \hat{y}(t|t-1, \vartheta)\big)^2\Big] \\
&\quad + 2\mathbb{E}\Big[e(t)\big(\hat{y}(t|t-1, \vartheta^0) - \hat{y}(t|t-1, \vartheta)\big)\Big]
\end{aligned}$$

Since the predictions $\hat{y}$ are uncorrelated with the white noise, the last addend is 0. Moreover, the first addend is the variance of the white noise (it has zero mean), hence we get

$$\mathbb{E}\Big[\varepsilon(t, \vartheta)^2\Big] = \bar{J}_N(\vartheta) = \lambda^2 + \mathbb{E}\Big[\big(\hat{y}(t|t-1, \vartheta^0) - \hat{y}(t|t-1, \vartheta)\big)^2\Big]$$

Finally we can see that the second addend is always positive (because it's squared) and in particular it is zero when $\vartheta = \vartheta^0$. This means that

$$\bar{J}_N(\vartheta) \geq \lambda^2 \quad \forall \theta \in \Theta$$

Moreover, since $J_N(\vartheta^0) = \lambda^2$ we obtain

$$\bar{J}_N(\vartheta) \geq J_N(\vartheta^0) \quad \forall \theta \in \Theta$$

This means that, if a system $S \in \mathcal{M}(\vartheta)$, then PEM identification guarantees that the identified model tends asymptotically to the true model for S.

Definitions

Theorems and principles