Temporal Information Processing

Student

Máster en Ciencia de Datos
Universidad Autónoma de Madrid

tu-web.es

Contents

Part I.

Introduction

1. Time Series

A time series, also known as discrete time signal, is a sequence of observations taken periodically in time. We can use time series to perform many tasks such as predictions of future values, behaviour analysis or information extraction. Examples of time series are audio signals, industrial instrument measures or diary finantial activity.

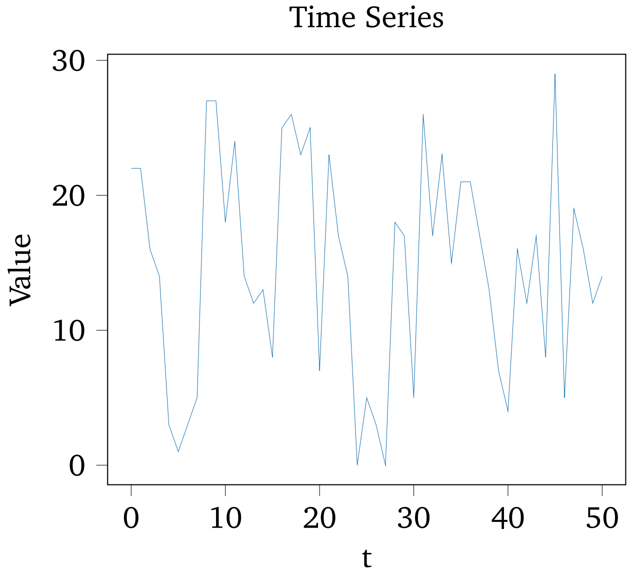


Figure 1: Example of a random time series.

A system can be determined comparing the input and the output. We call the system a filter if it is linear and time invariant. Considering the dynamic system as a black box, we can estimate the transference function or the impulse response to taht filter.

We can also consider **multivariate** time series, where some values of the time series have an influence on the other values in different or the same time instant. We can **classify** the time series in two wide types:

- Determinist: based in dynamic systems, they exploit the phisics of the generation algorithm of the time series.
- Stochastic: where the series are realizations of a stochastic process, which can be modelated.

In this subject, we will focus on stochastic models.

1.1. Stochastic Models

We can make three big considerations on the stochastic models.

- Stationary models.

  Definition 1.1. Let $\{X_t\}$ be a stochastic process and let $F_X\left(x_{t_1+\tau}, \ldots, x_{t_n+\tau}\right)$ represent the CDF of the **unconditional** joint distribution of $\{X_t\}$ at times $t_1+\tau, \ldots, t_n+\tau$. Then $\{X_t\}$ is strictly stationary if

  $$F_X\left(x_{t_1+\tau}, \ldots, x_{t_n+\tau}\right) = F_X\left(X_{t_1}, \ldots, x_{t_n}\right)$$

  However, we will use the case of **weak stationarity**, where we assume that the expectation of the stochastic process and the covariance at times $t, t+\tau$ are constant.

  Example 1.1. AR, MA, ARMA

- Non stationary models, where we do not make the assumption that the average of the process is constant in time and that there is seasonality

  Example 1.2. ARIMA, SARIMA

- Influenced by exogenous(extern) variables. In this cases, the exogenous variable affects the model, but the model does not affect this variable.

  Example 1.3. SARIMAX

Let us introduce some **notation** for the following explanations

Definition 1.2. Let z_t be the value of the time series at instant t.

- The **backward shift** operator is $z_{t-m} = B^m z_t$
- The **forward shift** operator is $z_{t+m} = F^m z_t = B^{-m} z_t$
- The difference or discrete gradient operator is $\nabla z_t = z_t - z_{t-1} = (1 - B) z_t$

Recall that, having a time series we can consider its **Z-transform**, that converts the discrete-time signal into a complex frequency-domain representation. In the Z-transform representation, the previously introduced notation is:

- The backward shift is $z_{t-m} = B^m z_t = Z^{-m} z_t$
- The forward shift is $z_{t+m} = B^{-m} z_t = Z^m z_t$
- The difference or discrete gradient is $\nabla z_t = (1 - Z^{-1}) z_t$

2. Linear filter based models

The stochastic models we use are based on time series z_t in which sucessive values are highly dependent. In these cases, we can see that the time series is generated from a series of independependent "shocks".

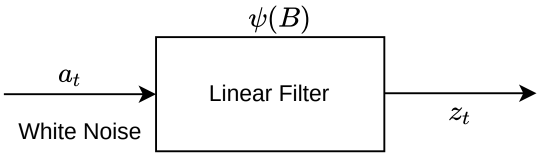
Definition 2.1. Let $a_t \sim \mathcal{N}\left(0, \sigma_a^2\right)$ be *white noise* (where each *shock* is related to a_t) which is not observed. Consider a linear filter that transforms the unobserved a_t to a observed time series z_t. We say that a **linear filter model is**

$$z_t = \mu + a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} + \cdots = \mu + \psi(B) a_t, \tag{1}$$

where

$$\psi(B) = 1 + \psi_1 B + \psi_2 B^2 + \ldots$$

is called the **transfer function** of the filter.



As we can see, we are expressing the filter in terms of a infinite sum of the coefficients ψ_i. If there are finite coefficients of the sum is *absolutely summable*, that is: $\sum_{j=0}^{\infty} \left|\psi_j\right| < \infty$ or the vector of coefficients has finite ℓ^1 norm, we say that the filter is **stable** and the process z_t is **stationary**.

In the case where the ℓ^1 norm is not finite, our filter are non-stable and produce non-stationary series.

2.1. Autoregressive Models (AR)

Let us firstly consider the simplest case of linear filter. An **autoregressive model** is a linear filter where the current value of the process $\tilde{z}_t$ is expressed as a finite sum of the previous values and a random shock a_t.

Definition 2.2. Let us denote the values of a process af equally spaced times $t, t-1, \ldots$ by $z_t, z_{t-1}, \ldots$. Consider that the values are centered, that is $\tilde{z}_t = z_t - \mu$. Then, the **autoregressive (AR) process** of **order p** is

$$\tilde{z}_t = \phi_1 \tilde{z}_{t-1} + \phi_2 \tilde{z}_{t-2} + \cdots + \phi_p \tilde{z}_{t-p} + a_t \tag{2}$$

Note that it is called autoregressive since, if you consider $\tilde{z}_{i-k}$ for $k = 1, \ldots, p$ as points, you are doing a *linear regression* over the past values.

Now, if we define the **autoregressive operator of order p** using the backward shift operator B as:

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p,$$

we can economically write the autoregressive model in (2) as

$$\phi(B)\tilde{z}_t = a_t \tag{3}$$

In practice, this model has $p+2$ unknown parameters $\mu, \phi_1, \ldots, \phi_p, \sigma_a^2$ which have to be estimated from the data.

Proposition 2.1. The autoregressive model is a particula case of a linear filter

Proof. Although we will not be estrictly formal in this proof, we will give an intuition on the iterative process that has to be done.

Consider the term $\tilde{z}_{t-1}$, let us eliminate it. Recall that

$$\tilde{z}_{t-1} = \phi_1 \tilde{z}_{t-2} + \cdots + \phi_p \tilde{z}_{t-p-1} + a_{t-1}.$$

We can substitute this term in the expression of the AR model given in Equation (2). The same can be done for $\tilde{z}_{t-2}$ and so on, to yield eventually an infinite series in the a terms.

□

In the case where $p = 1$, we have the AR process $\tilde{z}_t = \phi \tilde{z}_{t-1} + a_t$. After m sucessive substitutions of $\tilde{z}_{t-j} = \phi \tilde{z}_{t-j-1} + a_{t-j}$, with $j = 1, \ldots, m$, we obtain

$$\tilde{z}_t = \phi^{m+1} \tilde{z}_{t-m-1} + a_t + \phi a_{t-1} + \phi^2 a_{t-2} + \cdots + \phi^m a_{t-m}$$

Now, if we take the limit $m \to \infty$ this leads to the *convergent inifinite series representation* $\tilde{z}_t = \sum_{j=0}^{\infty} \phi^j a_{t-j}$, with $\psi_j = \phi^j, j \geq 1$, provided that $|\phi| < 1$. In the general AR case,

$$\phi(B)\tilde{z}_t = a_t$$

is equivalent to

$$\tilde{z}_t = \phi^{-1}(B) a_t = \psi(B) a_t, \qquad \psi(B) = \phi^{-1}(B) = \sum_{j=0}^{\infty} \psi_j B^j.$$

AR processes can be stationary or nonstationary. From the definition, it is clear that for a AR process to be stationary, the coefficients ϕ must be such that the weights $\psi_1, \psi_2, \ldots$ in $\psi(B) = \phi^{-1}(B)$ form a convergent series. A **necessary requirement** for stationarity is that the autoregressive operator $\phi(B)$, considered a polynomial in B of degree p, must have all roots greater than 1 in absolute value.

2.2. Application: Linear Prediction Coefficients in Speech Coding

Let us now set in the case of the **Speech Coding** topic. It is considered that a speech sample can be approximated as a linear combination of the past samples, which is how an AR model behaves. We have to find the coefficients that best suit our problem, using for instance the mean squared error prediction. We use the obtained **Linear Prediction Coefficients (LPCs)** to represent the signal frame.

Using this technique, we would be **reducing** the signal size significantly. However, since we are only approximating the signal, we would be most probably losing information. Two examples of codification of the audio signals are:

- MP3: which produces a different audio signal, involving loss of information
- FLAC: where the output is almost equal to the input, no loss of information

Signals are digitalized using a coding system.

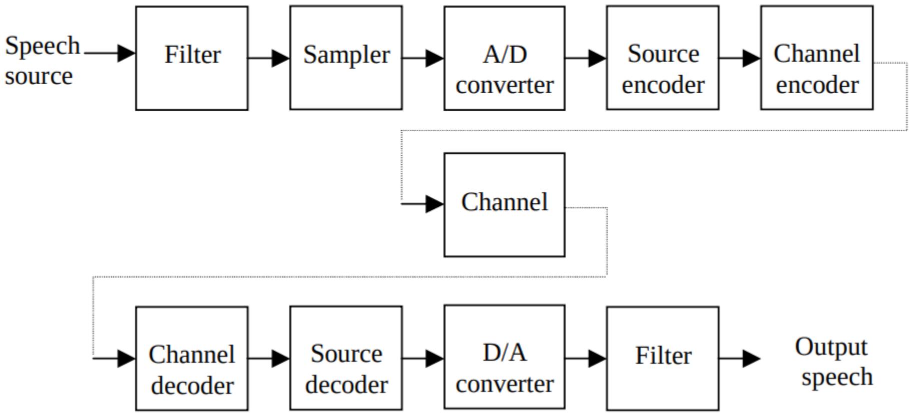


Figure 2: Block diagram of a speech coding system.

The filter eliminates aliasing and the sampler makes the continuous to discrete time conversion.

Example 2.1. In this example, we present the digital CD audio signal and why we would like to reduce its size without losing information. This signal has the following properties:

1. Sample rate $\Omega_s = 44.1kHz$
2. Bits per sample: 16
3. 2 channels (although sometimes 3 are used)

With this properties, the input bit rate is

$$R = \Omega_s \cdot \text{Bits/sample} \cdot \text{Channels} = 44.1 * 10^3 * 16 * 2 = 14112000 \frac{\text{bits}}{s} = 1.41 \frac{\text{Mb}}{s}$$

Which implies that, in a single minute we would need

$$60s;*;1,4112\frac{\text{Mb}}{s};*;\frac{1\,\text{byte}}{8\,\text{bits}} = 10.09\,\text{MB},$$

which is a high size for a single minute audio.

Example 2.2. In this example, we will present the input bit rate for the speech digital signal. Its common properties are:

1. Sample rate $\Omega_s = 8;\text{kHz}$
2. Bits per sample: 16
3. 1 channel

With this properties, the input bit rate is 128 Kb per second.

As a quick note, remember that **to quantify** a continuous time series is to assign it discrete amplitude values. When we do this, we are introducing a **cuantification error**,

$$\text{error}(t) = z_{\text{quantified}}(t) - z_{\text{original}}(t)$$

In each $t \in \mathbb{R}$, this error can be positive or negative. We can consider that the error is additive.

Let us consider a simplified version of the speech production model. Consider that there is a source and a filter, such as in the next figure:

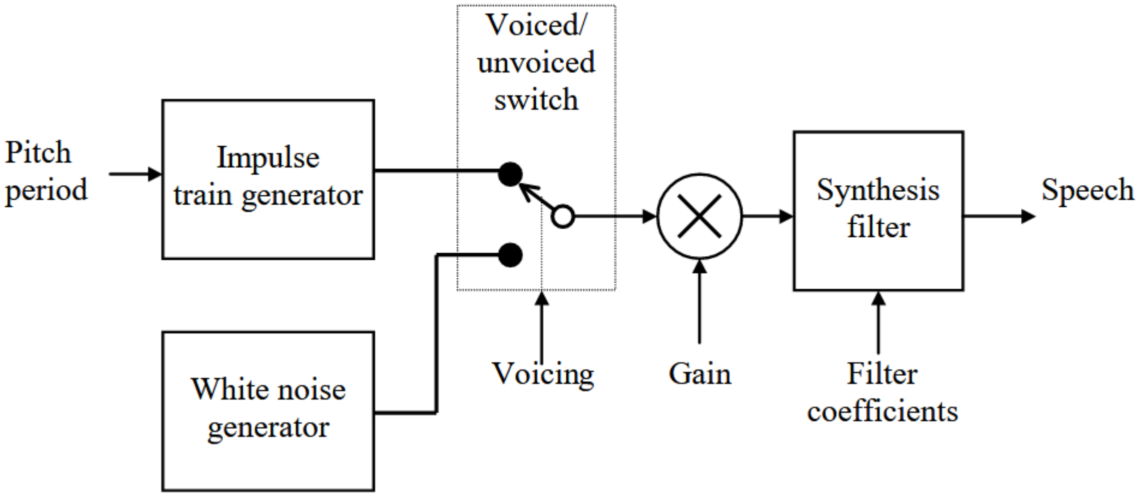


Figure 3: LPC model of speech production.

Then,

- We assume that we can separate the voice in non overlapping frames that are short enough to keep the model parameters constant.

- Then, we estimate the model parameters for each frame . These parameters are: voicing, gain (energy level of the frame), filter coefficients (response of the synthesis filter), pitch period (time length between consecutive excitation impulses)

Using the AR model to compute the synthesis filter coefficients

We can use the autoregressive model to predict a speech sample. Let $x[n]$ be the discrete signal. The prediction aims to find the coefficients a_k, $k = 1, \dots, p$ such that we can compute that sample using the previous samples

$$x[n] \approx \sum_{k=1}^{p} a_k\, x[n-k] \implies x[n] = \sum_{k=1}^{p} a_k\, x[n-k] + e[n] = \sum_{k=1}^{p} a_k\, x[n-k] + Ge'[n],$$

where $e[n]$ is the error at time step n, $e'[n]$ is the theoretical excitation and G is the gain. We **minimize the mean squared error**(also called mean energy) of the prediction error $e[n]$ in order to fit this model.

3. ARMA, (S)ARIMA(X) and Multivariate series

In this section we will generalize the Autoregressive model, adding complexity (and thus more generalization capability) to it.

3.1. MA and ARMA

Firstly, we bring back the general expression of a linear filter given in Equation (1):

$$\tilde{z}_t = a_t + \sum_{j=1}^{\infty} \psi_j a_{t-j}.$$

We can consider a special case when only the first $q \in \mathbb{N}$ are nonzero.

Definition 3.1. A **Moving Average (MA)** process of order q is a linear filter where only the first q terms are nonzero

$$\tilde{z}_t = a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q}. \tag{4}$$

As it can be appreciated, we now use the symbols $-\theta_1, \dots, -\theta_q$ for the finite set of *weights*. What we are doing is to *smooth* the white noise a_t.

Recall that, as we did before, we can express the moving average model as

$$\tilde{z}_t = (1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q) a_t = \theta(B) a_t.$$

The Moving Average models are **not addecuate** when the series has *autocorrelation* (that is, relation to its past values). Also, real-life time series are “more than white noise” to smooth.

A solution for this disadvantages could be combining the MA model with the AR model linearly:

> **Definition 3.2.** The **Autoregressive-Moving Average (ARMA)** process of order p, q is defined as a linear combination of both models
>
> $$\tilde{z}_t = \sum_{i=1}^{p} \phi_i \tilde{z}_{t-i} + a_t - \sum_{j=1}^{q} \theta_j a_{t-j},$$
>
> or
>
> $$\phi(B)\tilde{z}_t = \theta(B)a_t.$$

Using an ARMA model, we are not only capturing the relationships between a point $\tilde{z}_t$ and its previous ones (AR), but also smoothing the influence of the white noise (MA).

A great **disadvantage** of the ARMA models is that they are **always stationary**, so we cannot model non-stationary time series.

The following proposition presents a very interesting results on ARMA models:

> **Proposition 3.1.** An ARMA process is stationary if all the roots of $\phi(B) = 0$ have module greater than one, and it is (explosive) non stationary if the roots have module lesser than one.

The left-to-mention case where the roots of $\phi(B) = 0$ lie **on** the unit circle is very interesting. Nonseasonal series are often well represented by models in which one or more roots are unitary.

4. Non estationarity

Many time series appearing in real life have non stationary behaviour. Thus, we have to obtain new models to be able to make prediction of future values.

However, we can decompose these time series and treat them separately.

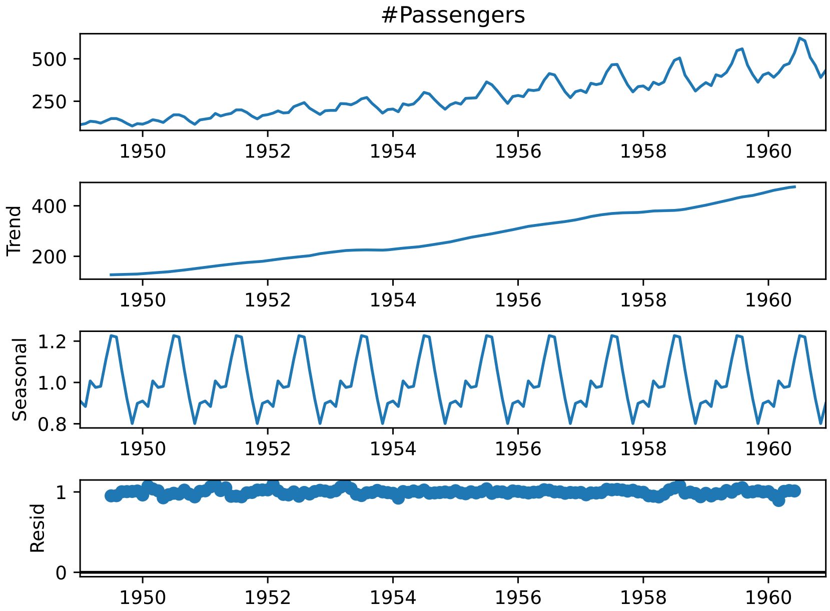


Figure 4: Multiplicative decomposition of a time series.

Tipically, we use the following components (all of them at time t):

1. T_t, the **trend** component, reflecting the long-term progression of the series. It exists when there is a persistent increasing or decreasing direction on the data. It is not neccesarily linear.
2. S_t, the **seasonal** component, reflecting the seasonal variation. A seasonal pattern exists when a time series is influenced by seasonal factors. Seasonality occurs over a fixed known period of time.
3. R_t, the **residual** component, describing randomness or irregular influences

Ocassionally, an additional cyclical component C_t is considered, but we will not do consider that case. With the considered components, using an **additive** model (used when the variations around the trend do not vary with the level of the time series), we can think of the time series as

$$z_t = T_t + S_t + R_t.$$

Using a **multiplicative** model (used when the trend is proportional to the level of the time series), our time series can be written as:

$$z_t = T_t \times S_t \times R_t.$$

Having the different components of a time series, we can look at the properties of each of the individual components and study them separately. Some properties that will help us in the creation of new models are:

1. The residual R_t is usually stationary, so we can use an already known model.

2. The trend T_t is usually a smooth function, which we can derivate (one or multiple times) in orden to *make it dissapear*
3. The stationality S_t has a periodic component.

4.1. ARIMA

Firstly, we will deal with the trend T_t. As we have said, we can make it *dissapear* by differenciating it.

Firstly, it can be shown that if d roots of the generalized autoregressive operator $\varphi(B)$ are unitary, then this operator can be written as

$$\varphi(B) = \phi(B)(1-B)^d,$$

where $\phi(B)$ is a stationary autoregressive operator. Thus, a model that can represent homogeneous nonstationary behaviour has the form:

$$\varphi(B)z_t = \phi(B)(1-B)^d z_t = \theta(B)a_t.$$

Now, if we name

$$w_t = (1-B)^d z_t = \nabla^d z_t,$$

we can rewrite the previous equation as

$$\phi(B)w_t = \theta(B)a_t,$$

and we are representing homogeneous nonstationary behaviour using the $d-$th diference of the process and calling it to be stationary. In practise, d is not usually greater than 2.

We can now use this reasoning to give a formal definition of the ARIMA process.

Definition 4.1. The **Autoregressive integrated moving average process (ARIMA)**of order p,d,q is defined by:

$$w_t = \sum_{i=1}^{p} \phi_i w_{t-i} + a_t - \sum_{j=1}^{q} \theta_j a_{t-j},$$

where $w_t = \nabla^d z_t$.

Remark 4.1. If we replace w_t by $z_t - \mu$, in the $d = 0$ case, the model includes the *estationary mixed model*, the AR and the MA models.

The following explanation gives an intuition of why is the model called *integrated* (although it probably should be called *summed*):

Let us find the **inverse relation** for $w_t = (1-B)^d z_t = \nabla^d z_t$. Consider

$$S = \nabla^{-1} = (1-B)^{-1} = \sum_{i=0}^{\infty} B^i.$$

Then, we can consider this *inverse relation* expressed as:

$$z_t = S^d w_t = \sum_{j=0}^{\infty} w_{t-j}.$$

Hence, it can be said that ARIMA may be generated by *summing* (or integrating) the stationary ARMA process w_t, d times.

To sum up, ARIMA has **three steps**

1. Derivate d times to *remove* trend:

$$w_t = (1-B)^d z_t = \nabla^d z_t.$$

2. Apply stationary model ARMA to w_t:

$$w_t = \sum_{i=1}^{p} \phi_i w_{t-i} + a_t - \sum_{j=1}^{q} \theta_j a_{t-j}.$$

3. Predict *reversing* the derivation

$$S^d w_t = \sum_{j=0}^{\infty} w_{t-j}$$

4.1.1. SARIMA

Having eliminated the Trend component, we would now like to estimate the **stationality** of the time series, assuming a multiplicative or additive relation. We perform the following extension of ARIMA.

> **Definition 4.2.** Consider a non stationary model, which **stationality** component has period S. Apply ARIMA to obtain a model that has S temporal units. Then, we obtain the **Seasonality ARIMA (SARIMA)**, of order S:
>
> $$\Phi(B^S)\nabla_S^D s_t = \Theta(B^S)\alpha_t.$$

It is common to assume that the stationality is *multiplicative*, obtaining the model *SARIMA* $(p,d,q)\times(P,D,Q)$. This stational component modlules the width/module of the rest of the components of the time series.

4.2. Exogenous variables

There is an especial case in which we consider variables outside our model that directly affect our time series. To model this case, we use AR and MA models, and add extra coefficients to the already known equations. Let us define one of the models:

Definition 4.3. Consider an ARMA model. Adding exogenous variables to it result in the **ARMAX** model, which has the following expression:

$$z_t = \sum_{i=1}^{p} \phi_i \tilde{z}_{t-i} + a_t - \sum_{j=1}^{q} \theta_i a_{t-j} - \sum_{k=1}^{r} \beta_k e_{t-k}.$$

The same way, **ARX,MAX,ARIMAX,SARIMAX** can be defined.

5. Model selection and fitting. Box-Jenkins method.

Now that we know all these fantastic models, we would like to apply them in time series analysis. We would like to choose which of the model **best fits** a time series data.

Firstly, we will introduce two defintions of functions that will be used to determine the best *hyperparameters* of our models. Since our purpose is to look at the *already-computed* functions, the following definitions will only give an intuitive idea and will not be very formal.

Definition 5.1. The **autocorrelation function ACF** is the correlation between a signal and a delayed copy of the signal. It is useful for finding repeating patterns (periodicity) or missing fundamental frequency.

Note. Unit root processes, trend-stationary processes, AR and MA processes are specific forms of processes with autocorrelation.

Definition 5.2. The **partial autocorrelation function (PACF)** gives the partial correlation of a stationary time series with its own lagged values, *regressed the values of the time series at all shorter lags*. That is, eliminates the influence of other lag values.

With these two definitions, we proceed to describe the Box-Jenkins Method. We execute the following steps:

1. Model class postulation. In this step, we select a family of models that we postulate our model will be in. As an example, we can consider *linear model based filters*.
2. Model identification: There are variations of what it should be done in this step. We will consider the following:
   a) Identifying estacionarity:
      i. Trend: It can be detected by differenciating the time series and checking that the autocorrelation function is transformed to a constant function 0 in $t = 0$ and 0 in the rest of times

   ii. Seasonality: Using the autocorrelation, we must find if there is a peak in the AFC in a specific time.

b) Eliminating *estationality and trend*

c) Determining the type of stationary model that we will use, looking at the ACF signal. As a **guide**, Table 1 can be used.

ACF	Model
Exponential decay as $\|lag\|$ increases	AR
Positive/negative decay as $\|lag\|$ increases, or smoothed sinusoid	AR
A few peaks in shiftings (lags) that are not $t = 0$	MA
Decay as $\|lag\|$ increases with an initial lag_0	ARMA
All values equal to zero and peak in $t = 0$	White Noise
High values in fixed intervals	Estationality component
No decay to zero	Non stationary time series

Table 1: Guide to select model using ACF.

Having selected our model, we have to find the optimal hyperparameters p and q that best fit our data. Usually, we will use:

- p such that the last *important peak* in the PACF has p shifting.
- q such that the last *important peak* in the ACF has q shifting.

There exist other theoretical estrategies that can be followed, but we will not explain them in this notes.

Lastly, it has to be mentioned that there are different measures that quantify the *goodness* of the fitted model, such as the **Akaike Information criterion** or the **Bayesian Information criterion**.

Note. The exogenous-variable models may have some variations in the Box-Jenkins method.