# A Tour of Time Series Analysis with R

*James Balamuta, Stéphane Guerrier, Roberto Molinari and Haotian Xu*

*2016-10-17*

# Contents

# List of Tables

# List of Figures

# Preface

This text is designed as an introduction to time series analysis and is used as a support document for the class STAT 429 (Time Series Analysis) given at the University of Illinois at Urbana-Champaign. It preferable to always access the text online rather than a printed to be sure you are using the latest version. The online version so affords additional features over the traditional PDF copy such as a scaling text, variety of font faces, and themed backgrounds. However, if you are in need of a local copy, a **pdf version** is also available.

This document is under active development and as a result is likely to contains many errors. As Montesquieu puts it:

> "*La nature semblait avoir sagement pourvu à ce que les sottises des hommes fussent passagères, et les livres les immortalisent.*"

## Contributing

If you notice any errors, we would be grateful if you would let us know. To let us know about the errors, there are two options available to you. The first and subsequently the fastest being if you are familiar with GitHub and know RMarkdown, then make a pull request and fix the issue yourself!. Note, in the online version, there is even an option to automatically start the pull request by clicking the edit button in the top-left corner of the text.

The second option, that will have a slightly slower resolution time is to send an email to `balamut2 AT illinois DOT edu` that includes: the error and a possible revision. Please put in the subject header: `[TTS]`.

## Bibliographic Note

This text is heavily inspired by the following three execellent references:

1. "*Time Series Analysis and Its Applications*", Third Edition, Robert H. Shumway & David S. Stoffer.
2. "*Time Series for Macroeconomics and Finance*", John H. Cochrane.
3. "*Cours de Séries Temporelles: Théorie et Applications*", Volume 1, Arthur Charpentier.

## Rendering Mathematical Formulae

Throughout the book, there will be mathematical symbols used to express the material. Depending on the version of the book, there are two different render engines.

- For the online version, the text uses MathJax to render mathematical notation for the web. In the event the formulae does not load for a specific chapter, first try to refresh the page. 9 times out of 10 the issue is related to the software library not loading quickly.
- For the pdf version, the text is built using the recommended AMS LaTeX symbolic packages. As a result, there should be no issue displaying equations.

An example of a mathematical rendering capabilities would be given as:

$$a^2 + b^2 = c^2$$

## R Code Conventions

The code used throughout the book will predominately be `R` code. To obtain a copy of `R`, go to the Comprehensive R Archive Network (CRAN) and download the appropriate installer for your operating system.

When `R` code is displayed it will be typeset using a `monospace` font with syntax highlighting enabled to ensure the differentiation of functions, variables, and so on. For example, the following adds 1 to 1

```
a = 1L + 1L
a
```

Each code segment may contain actual output from `R`. Such output will appear in grey font prefixed by `##`. For example, the output of the above code segment would look like so:

```
## [1] 2
```

Alongside the PDF download of the book, you should find the R code used within each chapter.

## License


Figure 1: This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

# Chapter 1

# Introduction

*Prévoir consiste à projeter dans l'avenir ce qu'on a perçu dans le passé.* Henri Bergson

After reading this chapter you will be able to:

- Describe what a *time series* is.
- Perform exploratory data analysis on time series data.
- Evaluate different characteristics of a time series.
- Classify basic time series models through equations and plots.
- Manipulate a time series equation using *backsubstitution*.

## 1.1 Time Series

Generally speaking a *time series* (or stochastic process) corresponds to set of "repeated" observations of the same variable such as price of a financial asset or temperature in a given location. In terms of notation a time series is often written as

$$(X_1, X_2, ..., X_n) \quad \text{or} \quad (X_t)_{t=1,...,n}\,.$$

The time index $t$ is contained within either the set of reals, $\mathbb{R}$, or integers, $\mathbb{Z}$. When $t \in \mathbb{R}$, the time series becomes a *continuous-time* stochastic process such a Brownian motion, a model used to represent the random movement of particles within a suspended liquid or gas, or an ElectroCardioGram (ECG) signal, which corresponds to the palpitations of the heart. However, within this text, we will limit ourselves to the cases where $t \in \mathbb{Z}$, better known as *discrete-time* processes. *Discrete-time* processes are where a variable is measured sequentially at fixed and equally spaced intervals in time akin to 1.1. This implies that we will have two assumptions:

1. $t$ is not random e.g. the time at which each observation is measured is known, and
2. the time between two consecutive observations is constant.

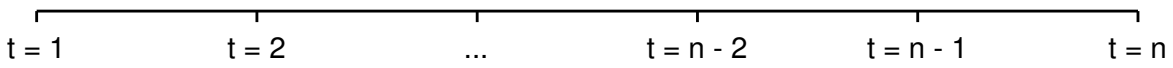



Figure 1.1: Discrete-time can be thought of as viewing a number line with equally spaced points.

Moreover, the term "time series" can also represent a probability model for a set of observations. For example, one of the fundamental probability models used in time series analysis is called a *white noise* process and is defined as

$$W_t \overset{iid}{\sim} N(0, \sigma^2).$$

This statement simply means that $(W_t)$ is normally distributed and independent over time. This model may appear to be dull but as we will see it is a crucial component to constructing more complex models. Unlike the white noise process, time series are typically *not* independent over time. Suppose that the temperature in Champaign is unusually low, then it is reasonable to assume that tomorrow's temperature will also be low. Indeed, such behavior would suggest the existence of a dependency over time. The time series methods we will discuss in this text consists of parametric models used to characterize (or at least approximate) the joint distribution of $(X_t)$. Often, time series models can be decomposed into two components, the first of which is what we call a *signal*, say $(Y_t)$, and the second component is a *noise*, say $(W_t)$, leading to the model

$$X_t = Y_t + W_t.$$

Typically, we have $E[Y_t] \neq 0$ while $E[W_t] = 0$ (although we may have $E[W_t|W_{t-1}, ..., W_1] \neq 0$). Such models impose some parametric structure which represents a convenient and flexible way of studying time series as well as a means to evalute *future* values of the series through forecasting. As we will see, predicting future values is one of the main aspects of time series analysis. However, making predictions is often a daunting task or as famously stated by Nils Bohr:

"*Prediction is very difficult, especially about the future.*"

There are plenty of examples of predictions that turned out to be completely erroneous. For example, three days before the 1929 crash, Irving Fisher, Professor of Economics at Yale University, famously predicted:

"*Stock prices have reached what looks like a permanently high plateau*".

Another example is given by Thomas Watson, president of IBM, who said in 1943:

"*I think there is a world market for maybe five computers.*"

## 1.2 Exploratory Data Analysis for Time Series

When dealing with relatively small time series (e.g. a few thousands), it is often useful to look at a graph of the original data. These graphs can be informative to "detect" some features of a time series such as trends and the presence of outliers.

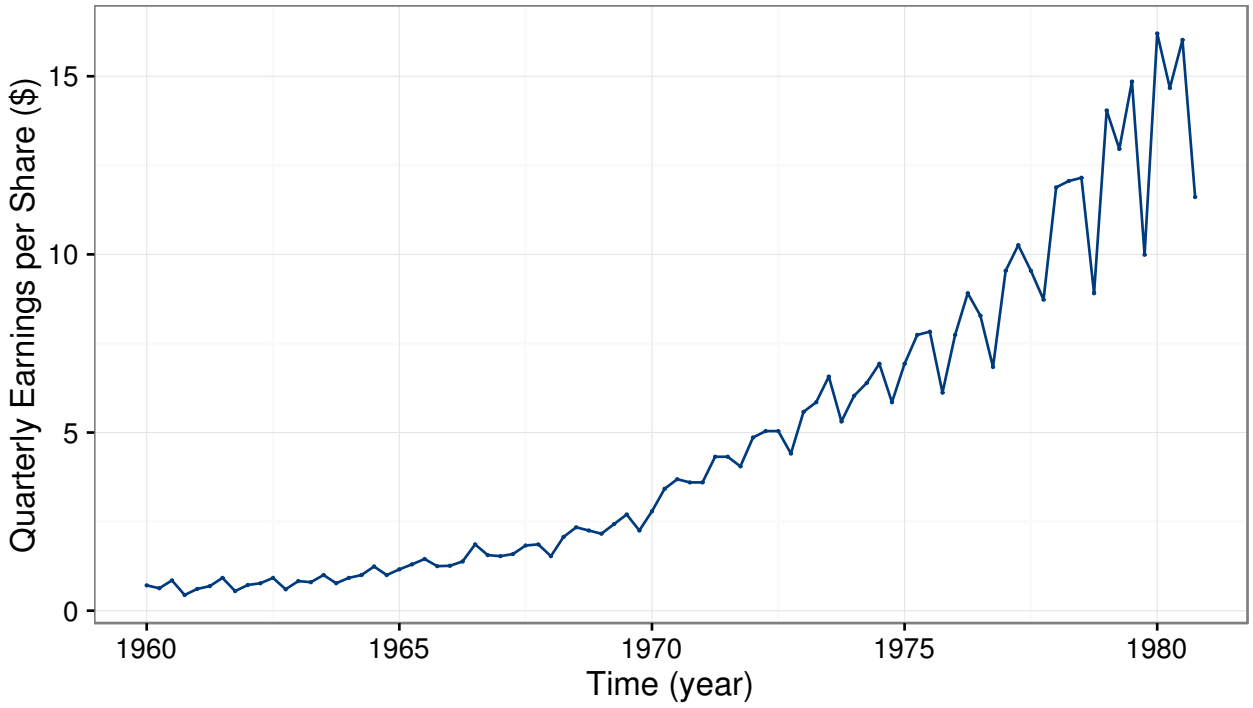
Indeed, a trend is typically assumed to be present in a time series when the data exhibit some form of long term increase or decrease or combination of increases or decreases. Such trends could be linear or non-linear and represent an important part of the "signal" of a model. Here are a few examples of non-linear trends:

1. **Seasonal trends** (periodic): These are the cyclical patterns which repeat after a fixed/regular time period. This could be due to business cycles (e.g. bust/recession, recovery).
2. **Non-seasonal trends** (periodic): These patterns cannot be associated to seasonal variation and can for example be due to an external variable such as, for example, the impact of economic indicators on stock returns. Note that such trends are often hard to detect based on a graphical analysis of the data.

3. **"Other" trends**: These trends have typically no regular patterns and are over a segment of time, known as a "window", that change the statistical properties of a time series. A common example of such trends is given by the vibrations observed before, during and after an earthquake.

**Example 1.** A traditional example of a time series is the quarterly earnings of the company Johnson and Johson. In the figure below, we present these earnings between 1960 and 1980

```
# Load data
data(jj, package = "astsa")

# Construct gts object
jj = gts(jj, start = 1960, freq = 4, name = 'Johnson and Johnson Quarterly Earnings',
         unit = "year")

# Plot time series
autoplot(jj) + ylab("Quarterly Earnings per Share ($)")
```



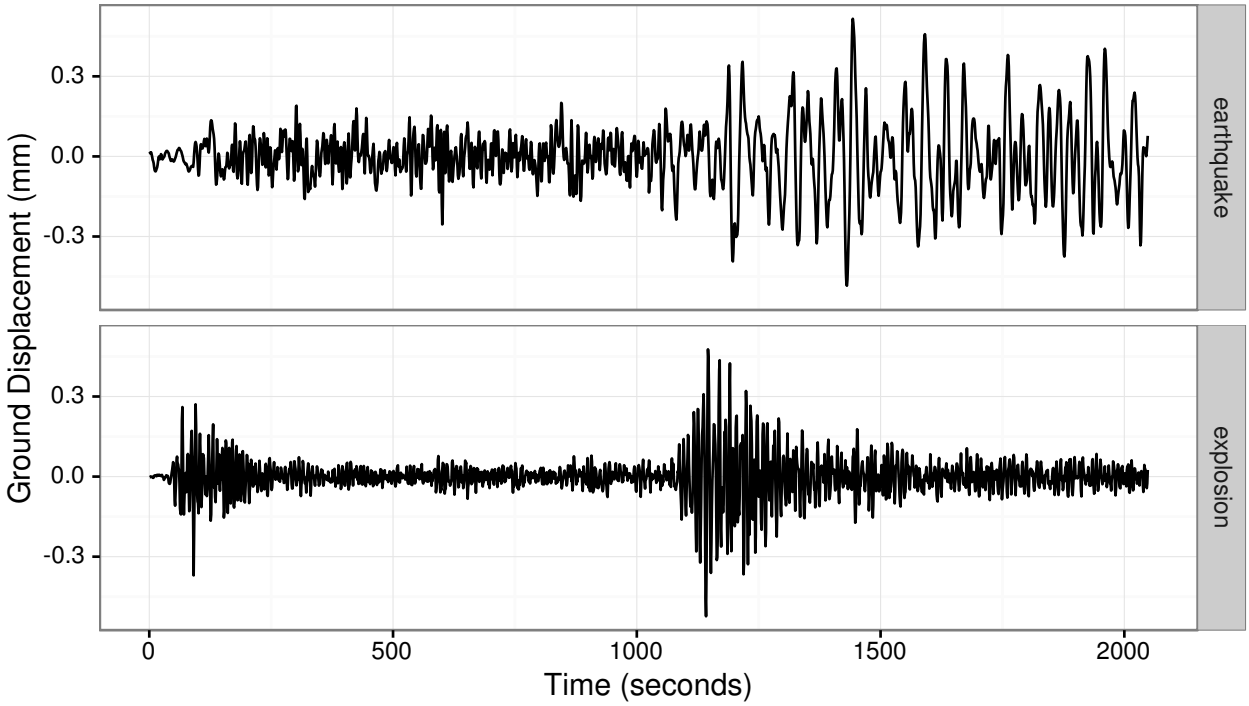

One trait that the graph makes evident is that the data contains a non-linear increasing trend as well as a yearly seasonal component. In addition, one can note that the *variability* of the data seems to increase with time. Being able to make such observations provides important information to select suitable models for the data.

Moreover, when observing "raw" time series data it is also interesting to evaluate if some of the following phenomena occur:

1. **Change in Mean:** Does the mean of the process shift over time?
2. **Change in Variance:** Does the variance of the process evolve with time?
3. **Change in State:** Does the time series appear to change between "states" having distinct statistical properties?
4. **Outliers** Does the time series contain some "extreme" observations? Note that this is typically difficult to assess visually.

**Example 2.** In the figure below, we present an example of displacement recorded during an earthquake as well as an explosion.

```
# Load data
data(EQ5, package="astsa")
data(EXP6, package="astsa")

EQ5.df = fortify(EQ5)
EQ5.df$type = "earthquake"
EXP6.df = fortify(EXP6)
EXP6.df$type = "explosion"
eq.df = rbind(EQ5.df, EXP6.df)

# Plot time series
ggplot(data = eq.df, aes(Index, Data)) + geom_line() + facet_grid( type ~ .) +
  ylab("Ground Displacement (mm)") + xlab("Time (seconds)") + theme_bw()
```

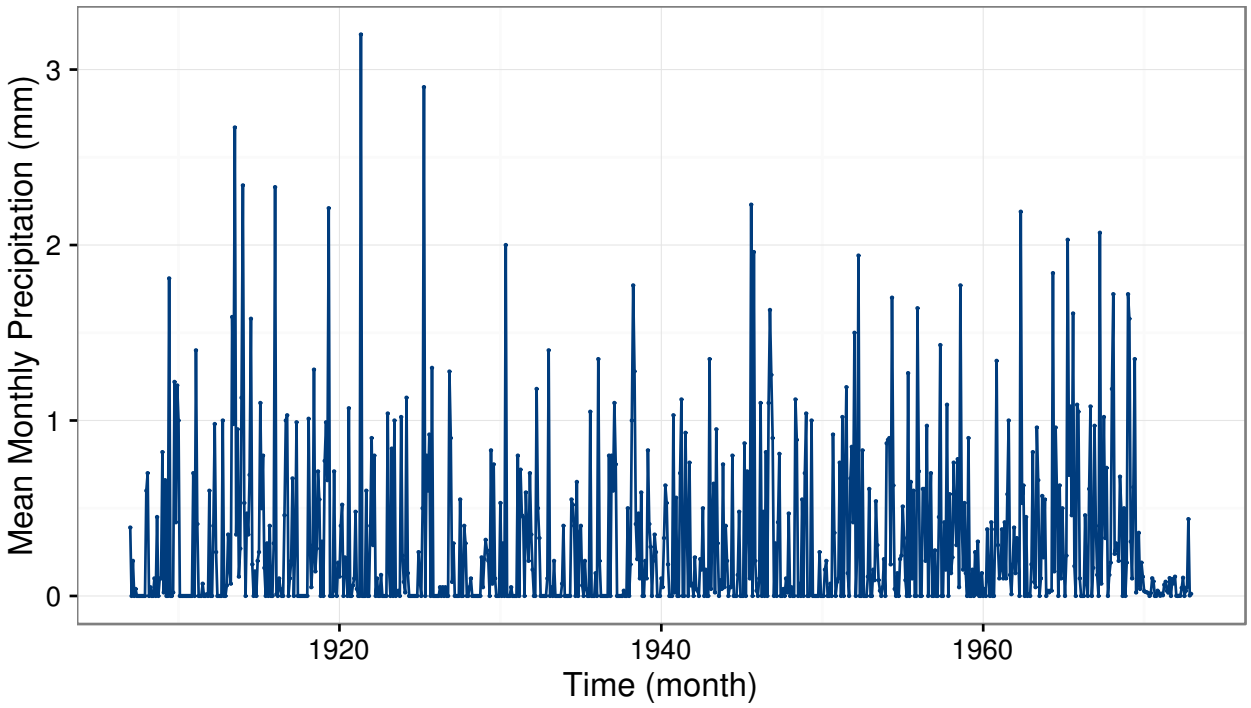



From the graph, it can be observed that the statistical properties of the time series appear to change over time. For instance, the variance of the time series shifts at around $t = 1150$ for both series. The shift in variance also opens "windows" where there appear to be distinct states. In the case of the explosion data, this is particularly relevant around $t = 50, \cdots, 250$ and then again from $t = 1200, \cdots, 1500$. Even within these windows, there are "spikes" that could be considered as outliers most notably around $t = 1200$ for explosion series.

Extreme observations or outliers are commonly observed in real time series data, this is illustrated in the following example.

**Example 3.** We consider here a data set coming from the domain of hydrology. The data concerns monthly precipitation (in mm) over a certain period of time (1907 to 1972) and is interesting for scientists in order to study water cycles. The data are presented in the graph below:

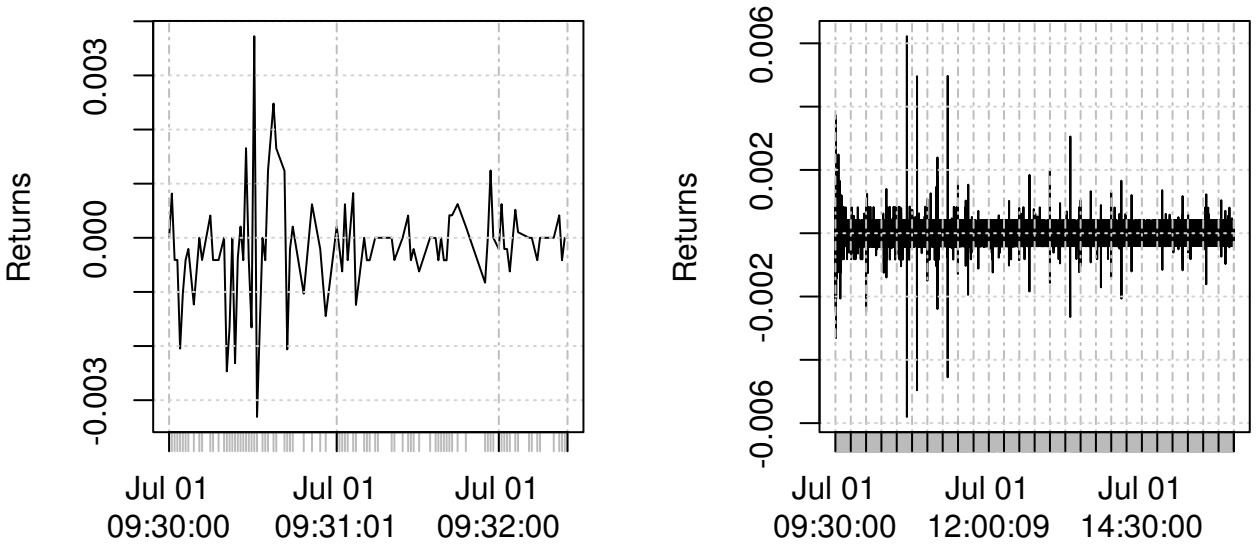
```r
# Load data
hydro = read.csv("data/precipitation.csv", header=T, sep=";")

# Construct gts object
hydro = gts(hydro[,2], start = 1907, freq = 12, name = 'Precipitation Data',
            unit = "month")

# Plot data
autoplot(hydro)  +  ylab("Mean Monthly Precipitation (mm)")
```




Next, we consider an example coming from high-frequency finance to illustrate the limitations our current framework.

**Example 4.** The figure below presents the returns or price innovations (i.e. informally speaking the changes in price from one observation to the other) for the Starbuck's stock on July 1, 2011 for about 150 seconds (left panel) and about 400 minutes (right panel).

```r
library(timeDate)

# Load "high-frequency" Starbucks returns for Jul 01 2011
data(sbux.xts, package = "highfrequency")

# Plot returns
par(mfrow = c(1,2))
plot(sbux.xts[1:89], main = " ", ylab = "Returns")
plot(sbux.xts, main = " ", ylab = "Returns")
```



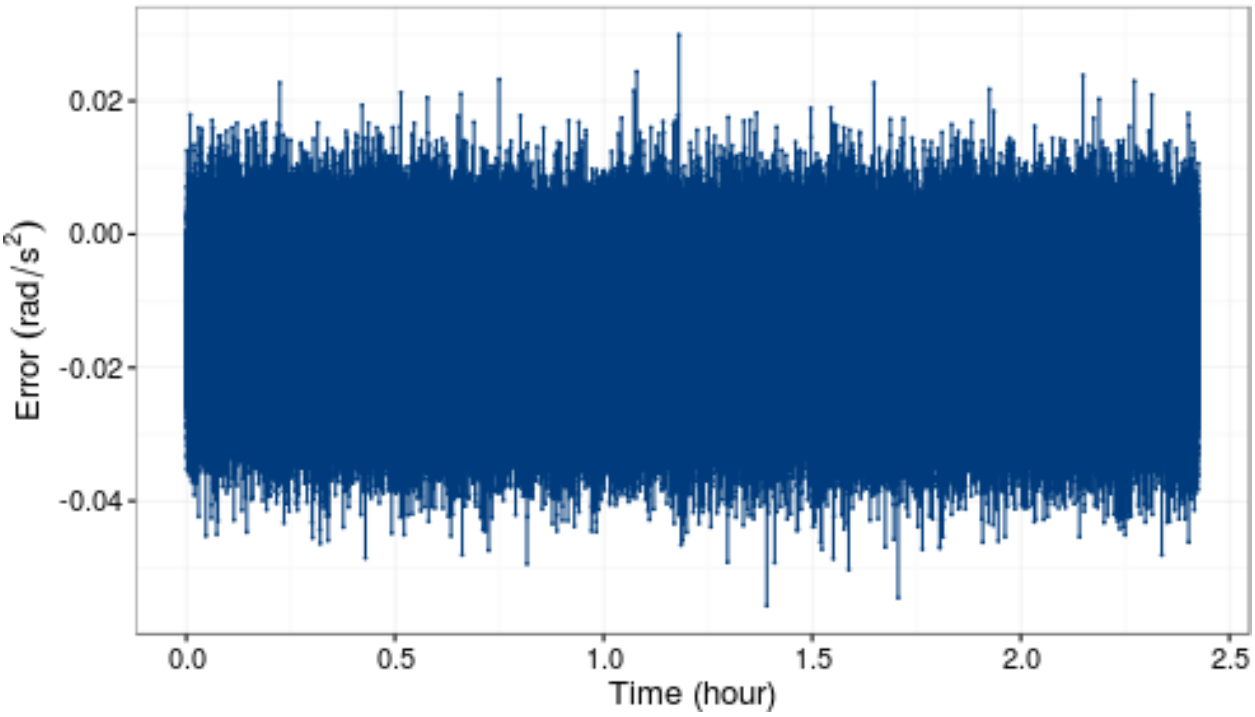

It can be observed on the left panel that observations are not equally spaced. Indeed, in high-frequency data the intervals between two points is typically not constant and, even worse, is a random variable. This implies that the time when a new observation will be available is in general unknown. On the right panel, one can observe that the variability of the data seems to change during the course of the trading day. Such a phenomenon is well known in the finance community since a lot of variation occurs at the start (and the end) of the day while the middle of the day is associated with small changes. Moreover, clear extreme observations can also be noted in this graph at around 11:00

Finally, let us consider the limitations of a direct graphical representation of a time series when the sample size is large. Indeed, due to visual limitations, a direct plotting of the data will probably result in an uninformative aggregation of points between which it is unable to distinguish anything. This is illustrated in the following example.

**Example 5.** We consider here the data coming from the calibration procedure of an Inertial Measurement Unit (IMU) which, in general terms, is used to enhance navigation precision or reconstruct three dimensional movements (see e.g. link). These sensors are used in a very wide range of applications such as robotics, virtual reality, vehicle stability control, human and animal motion capture and so forth (see e.g. link). The signals coming from these instruments are measured at high frequencies over a long time and are often characterized by linear trends and numerous underlying stochastic processes. If you have never heard

The code below retrieves some data from an IMU and plots it directly:

```
# Load packages
library(gmwm)
library(imudata)

# Load IMU data
data(imu6, package = "imudata")

# Construct gst object
Xt = gts(imu6[,1], name = "Gyroscope data", unit = "hour", freq = 100*60*60)

# Plot time series
autoplot(Xt) + ylab(expression(paste("Error ", (rad/s^2))))
```




Although a linear trend and other processes are present in this signal (time series), it is practically impossible to understand or guess anything from the plot.

## 1.3 Basic Time Series Models

In this section, we introduce some simple time series models. Before doing so it is useful to define $\Omega_t$ as all the information avaiable up to time $t-1$, i.e.

$$\Omega_t = (X_{t-1}, X_{t-2}, ..., X_0)\,.$$

As we will see this compact notation is quite useful.

### 1.3.1 White noise processes

The building block for most time series models is the Gaussian white noise process, which can be defined as

$$W_t \overset{iid}{\sim} N\left(0, \sigma_w^2\right).$$

This definition implies that:

1. $E[W_t|\Omega_t] = 0$ for all $t$,
2. $\text{cov}\left(W_t, W_{t-h}\right) = \mathbf{1}_{h=0}\,\sigma^2$ for all $t, h$.

Therefore, in this process there is an absence of temporal (or serial) dependence and is homoskedastic (i.e it has a constant variance). This definition can be generalized into two sorts of processes, the *weak* and *strong* white noise. The process $(W_t)$ is a weak white noise if

1. $E[W_t] = 0$ for all $t$,
2. $\text{var}\,(W_t) = \sigma^2$ for all $t$,
3. $\text{cov}\,(W_t, W_{t-h}) = 0$, for all $t$, and for all $h \neq 0$.

Note that this definition does not imply that $W_t$ and $W_{t-h}$ are independent (for $h \neq 0$) but simply uncorrelated. However, the notion of independence is used to define a *strong* white noise as

1. $E[W_t] = 0$ and $\text{var}(W_t) = \sigma^2 < \infty$, for all $t$,
2. $F(W_t) = F(W_{t-h})$, for all $t, h$ (where $F(W_t)$ denotes the distribution of $W_t$),
3. $W_t$ and $W_{t-h}$ are independent for all $t$ and for all $h \neq 0$.

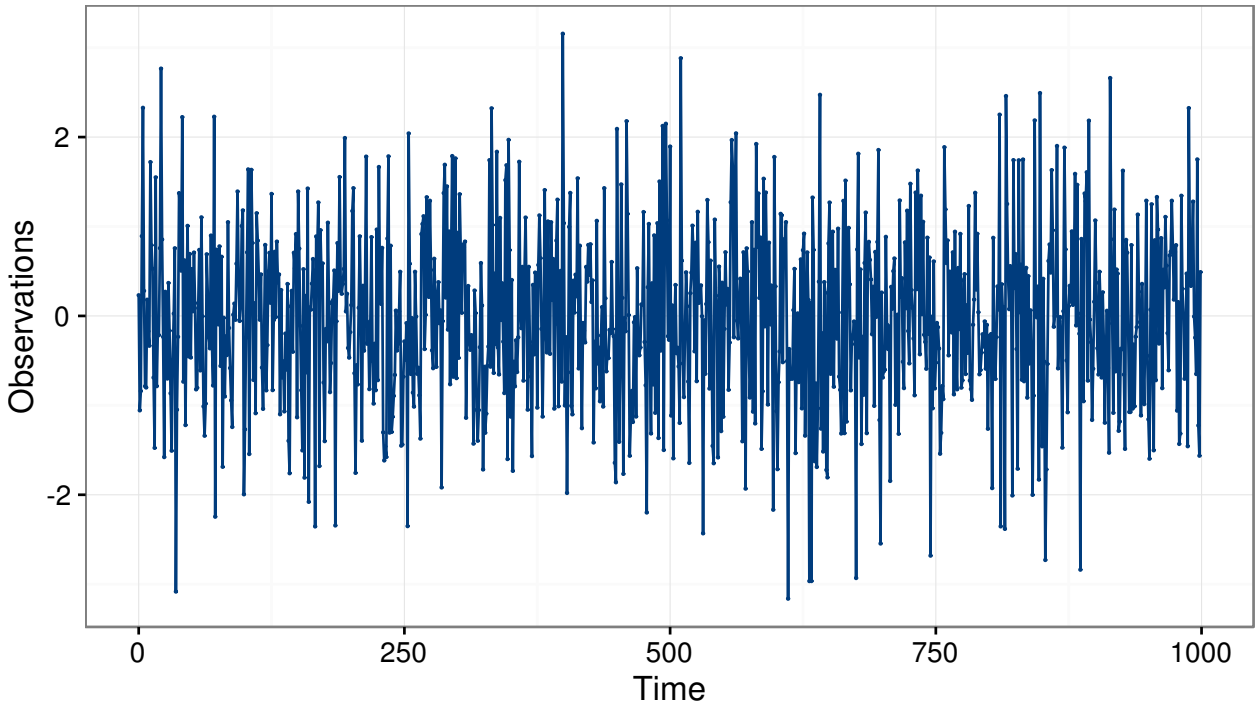
It is clear from these definitions that if a process is a strong white noise it is also a weak white noise. However, the converse is not true as shown in the following example:

**Example 6.** Let $Y_t \sim F_{t+2}$, where $F_{t+2}$ denotes a Student distribution with $t+2$ degrees of freedom. Assuming the sequence $(Y_1, \ldots, Y_n)$ to be independent, we let $X_t = \sqrt{\frac{t}{t+2}} Y_t$. Then, the process $(X_t)$ is obviously not a strong white noise as the distribution of $X_t$ changes with $t$. However, this process is a weak white noise since we have:

- $E[X_t] = \sqrt{\frac{t}{t+2}} E[Y_t] = 0$ for all $t$.
- $\text{var}(X_t) = \frac{t}{t+2} \text{var}(Y_t) = \frac{t}{t+2} \frac{t+2}{t} = 1$ for all $t$.
- $\text{cov}(X_t, X_{t+h}) = 0$ (by independence), for all $t$, and for all $h \neq 0$.

The code below presents an example of how to simulate a Gaussian white noise process

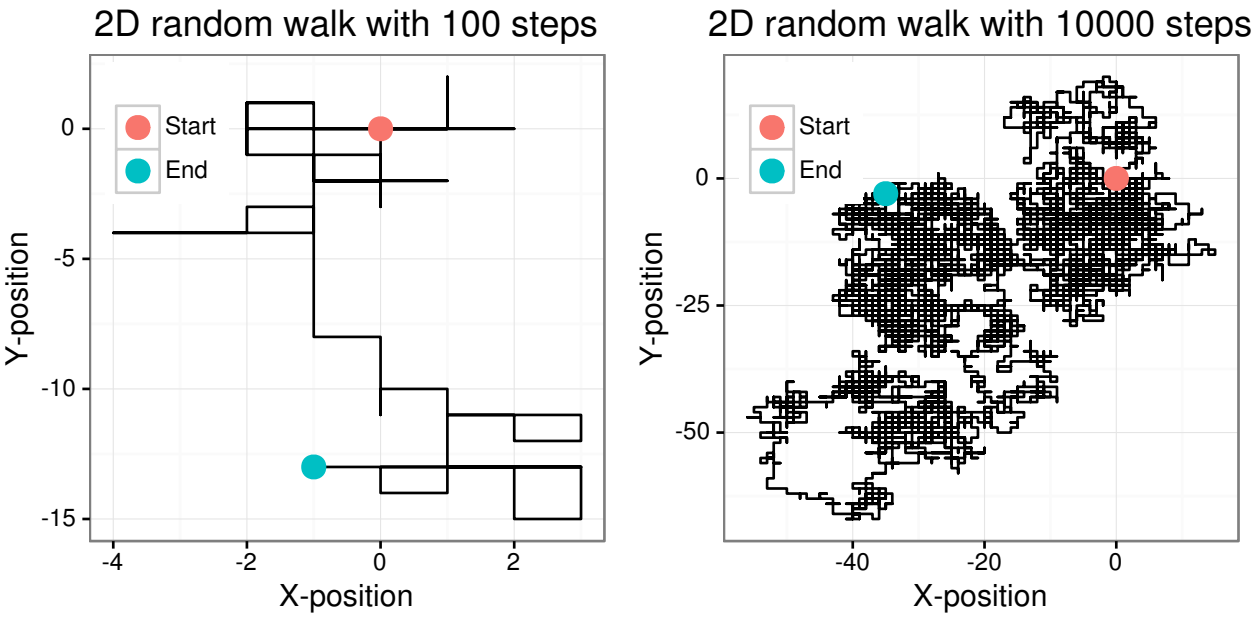
```
# This code simulates a gaussian white noise process
n = 1000                                  # process length
sigma2 = 1                                # process variance
Xt = gen.gts(WN(sigma2 = sigma2), N = n)
plot(Xt)
```



### 1.3.2 Random Walk Processes

The term *random walk* was first introduce by Karl Pearson in the early 19 hundreds. As for the white noise, there exist a large range of random walk processes. For example, one of the simplest forms of random walk can be explained as follows: suppose that you are walking on campus and your next step can either be to your left, your right, forward or backward (each with equal probability). Two realizations of such processes are represented below:

```
library("gridExtra")

# Function computes direction random walk moves
RW2dimension = function(steps = 100){
  # Initial matrix
  step_direction = matrix(0, steps+1, 2)

  # Start random walk
  for (i in seq(2, steps+1)){
    # Draw a random number from U(0,1)
    rn = runif(1)

    # Go right if rn \in [0,0.25)
    if (rn < 0.25) {step_direction[i,1] = 1}

    # Go left if rn \in [0.25,0.5)
    if (rn >= 0.25 && rn < 0.5) {step_direction[i,1] = -1}

    # Go forward if rn \in [0.5,0.75)
    if (rn >= 0.5 && rn < 0.75) {step_direction[i,2] = 1}

    # Go backward if rn \in [0.75,1]
    if (rn >= 0.75) {step_direction[i,2] = -1}
  }

  # Cumulative steps
  position = data.frame(x = cumsum(step_direction[, 1]), y = cumsum(step_direction[, 2]))

  # Mark start and stop locations
  start_stop = data.frame(x = c(0, position[steps+1, 1]), y = c(0, position[steps+1, 2]),
    type = factor(c("Start","End"), levels = c("Start","End")))

  # Plot results
  ggplot(mapping = aes(x = x, y = y)) + geom_path(data = position) +
    geom_point(data = start_stop, aes(color = type), size = 4) +
    theme_bw() + labs(x = "X-position", y = "Y-position",
    title = paste("2D random walk with", steps, "steps"),
    color = "") + theme(legend.position = c(0.15, 0.84))
}


# Plot 2D random walk with 10^2 and 10^5 steps
set.seed(5)
a = RW2dimension(steps = 10^2)
b = RW2dimension(steps = 10^4)
grid.arrange(a, b, nrow = 1)
```




Such processes inspired Karl Pearson's famous quote that

> "*the most likely place to find a drunken walker is somewhere near his starting point.*"

Empirical evidence of this phenomenon is not too hard to find on a Friday night in Champaign. In this text, we only consider one very specific form of random walk, namely the Gaussian random walk which can be defined as:

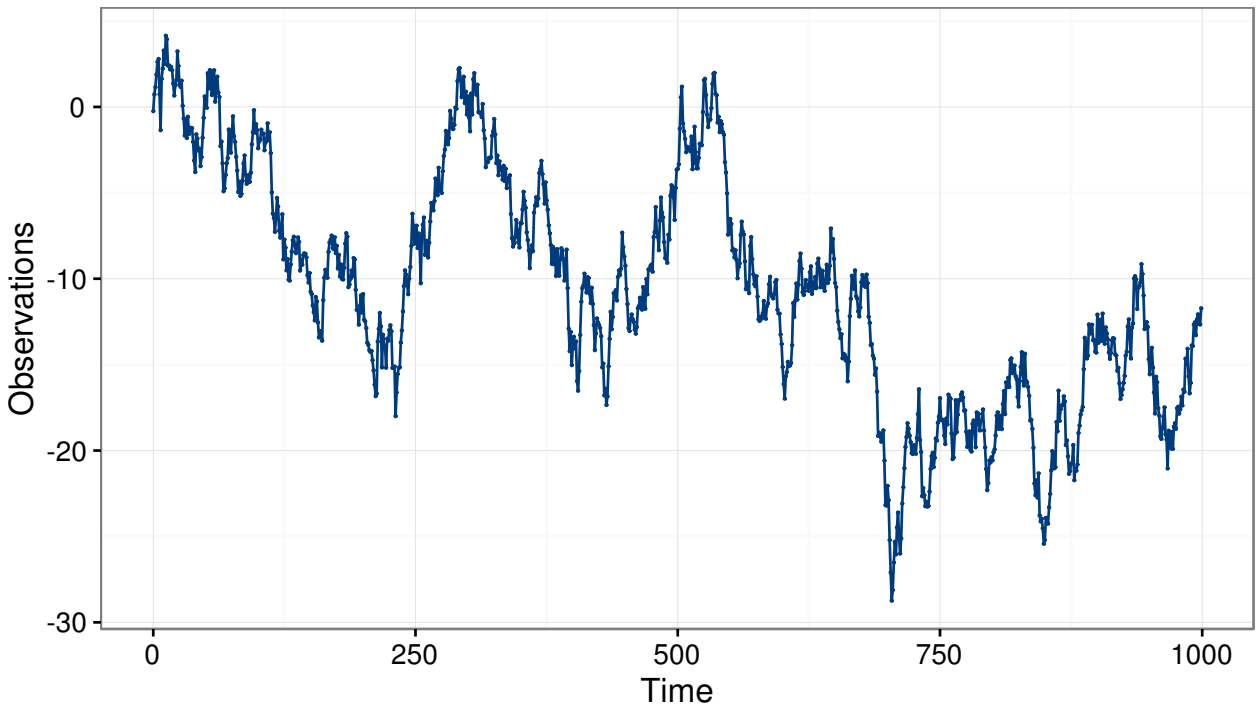
$$X_t = X_{t-1} + W_t,$$

where $W_t$ is a Gaussian white noise and with initial condition $X_0 = c$ (typically $c = 0$). This process can be expressed differently by *backsubstitution* as follows:

$$\begin{aligned} X_t &= X_{t-1} + W_t \\ &= (X_{t-2} + W_{t-1}) + W_t \\ &= \vdots \\ X_t &= \sum_{i=1}^{t} W_i + X_0 = \sum_{i=1}^{t} W_i + c \end{aligned}$$

The code below presents an example of how to simulate a such process

```
# This code simulates a gaussian random walk process
n = 1000                                # process length
gamma2 = 1                              # innovation variance
Xt = gen.gts(RW(gamma2 = gamma2), N = n)
plot(Xt)
```




### 1.3.3 Autoregressive Process of Order 1

An autoregressive process of order 1 or AR(1) is a generalization of both the white noise and random walk processes which are both themselves special cases of an AR(1). A (Gaussian) AR(1) process can be defined as

$$X_t = \phi X_{t-1} + W_t,$$

where $W_t$ is a Gaussian white noise. Clearly, an AR(1) with $\phi = 0$ is a Gaussian white noise and when $\phi = 1$ the process becomes a random walk.

**Remark 1.** We generally assume that an AR(1), as well as other time series models, have zero mean. The reason for this assumption is only to simplfy the notation but it is easy to consider an AR(1) process around an arbitrary mean $\mu$, i.e.

$$(X_t - \mu) = \phi\,(X_{t-1} - \mu) + W_t,$$

which is of course equivalent to

$$X_t = (1 - \phi)\,\mu + \phi X_{t-1} + W_t.$$

Thus, we will generally only work with zero mean processes since adding means is simple.

**Remark 2.** An AR(1) is in fact a linear combination of the past realisations of the white noise $W_t$. Indeed, we have

$$\begin{aligned} X_t &= \phi_t X_{t-1} + W_t = \phi\,(\phi X_{t-2} + W_{t-1}) + W_t \\ &= \phi^2 X_{t-2} + \phi W_{t-1} + W_t = \phi^t X_0 + \sum_{i=0}^{t-1} \phi^i W_{t-i}. \end{aligned}$$

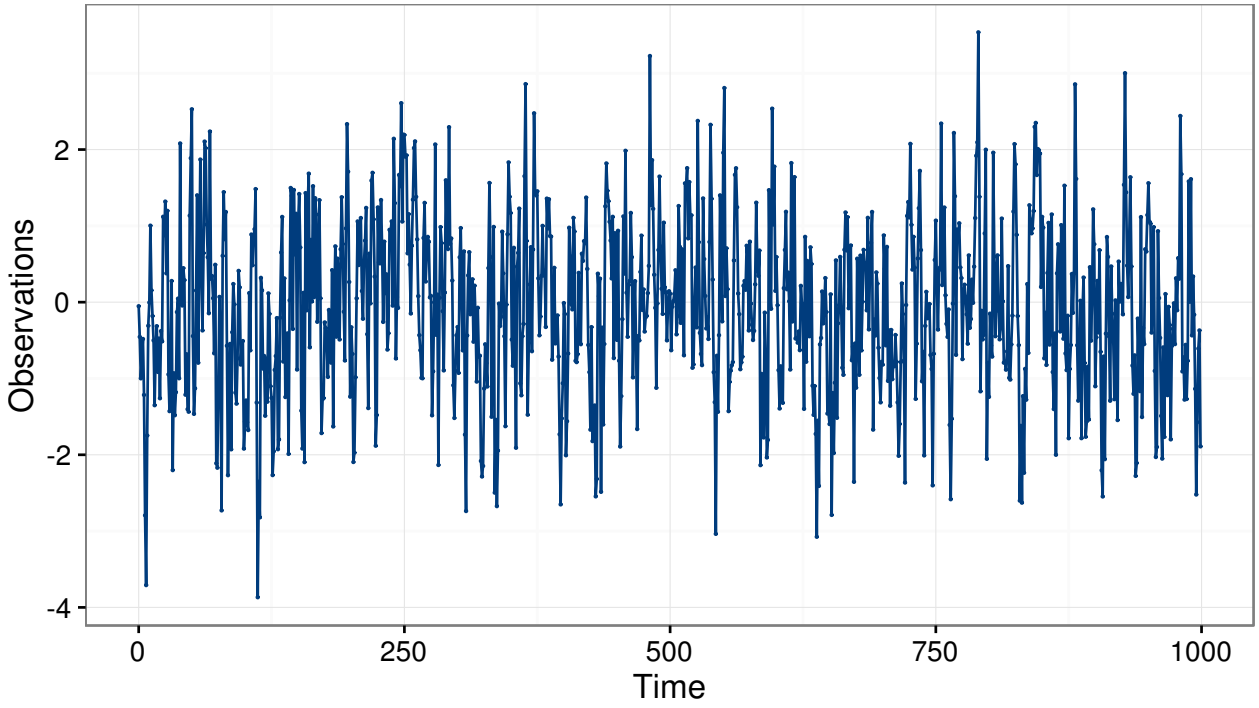
Under the assumption of infinite past (i.e. $t \in \mathbb{Z}$) and $|\phi| < 1$, we obtain

$$X_t = \sum_{i=0}^{\infty} \phi^i W_{t-i},$$

since $\lim_{i\to\infty} \phi^i X_{t-i} = 0$.

The code below presents an example of how an AR(1) can be simulated

```
# This code simulate a gaussian random walk process
n = 1000                                  # process length
phi = 0.5                                 # phi parameter
sigma2 = 1                                # innovation variance
Xt = gen.gts(AR1(phi = phi, sigma2 = sigma2), N = n)
plot(Xt)
```



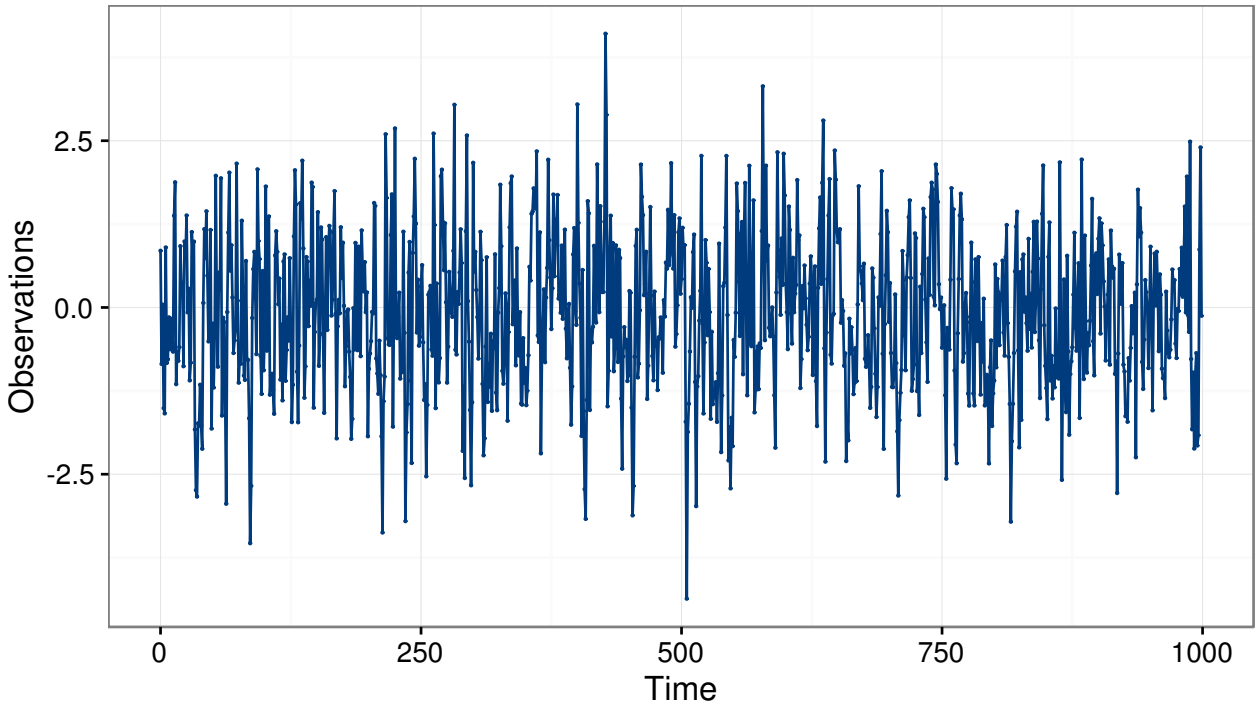

### 1.3.4 Moving Average Process of Order 1

As we have seen in the previous example, an AR(1) can be expressed as a linear combination of all past observations of $(W_t)$ while the next process, called a moving average process of order 1 or MA(1), is (in some sense) a "truncated" version of an AR(1). It is defined as

$$X_t = \theta W_{t-1} + W_t, \tag{1.1}$$

where (again) $W_t$ denotes a Gaussian white noise process. An example on how to generate an MA(1) is given below:

```
# This code simulates a gaussian white noise process
n = 1000                                  # process length
sigma2 = 1                                # innovation variance
theta = 0.5                               # theta parameter
Xt = gen.gts(MA1(theta = theta, sigma2 = sigma2), N = n)
plot(Xt)
```




### 1.3.5 Linear Drift

A linear drift is a very simple deterministic time series model which can be expressed as

$$X_t = X_{t-1} + \omega,$$

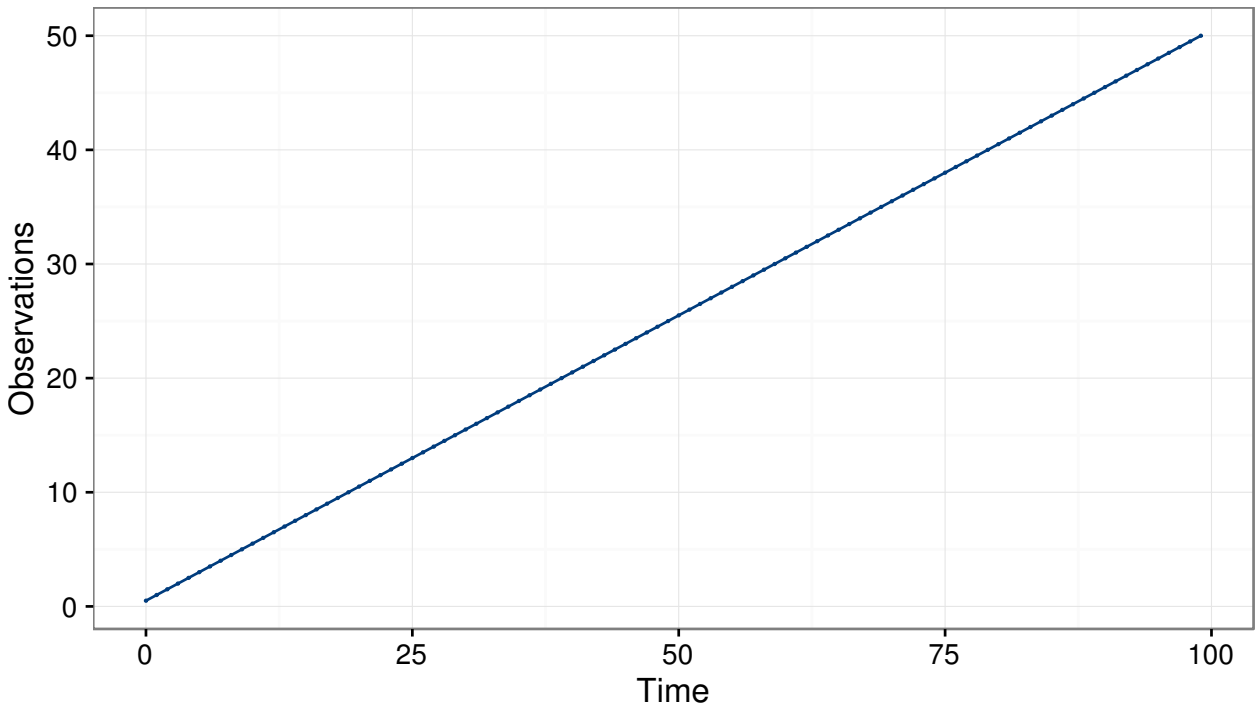
where $\omega$ is a constant and with the initial condition $X_0 = c$, an arbitrary constant (typically zero). This process can be expressed in a more familiar form as follows:

$$X_t = X_{t-1} + \omega = (X_{t-2} + \omega) + \omega = t\omega + c$$

Therefore, a (linear) drift corresponds to a simple linear model with slope $\omega$ and intercept $c$.

A drift can simply be generated using the code below:

```
# This code simulate a linear drift with 0 intercept
n = 100                                   # process length
omega = 0.5                               # slope parameter
Xt = gen.gts(DR(omega = omega), N = n)
plot(Xt)
```




## 1.4 Composite Stochastic Processes

A composite stochastic process can be defined as the sum of underlying (or latent) stochastic processes. In this text, we will use the term *latent time series* as a synomym for composite stochastic processes. A simple example of such a process is given by

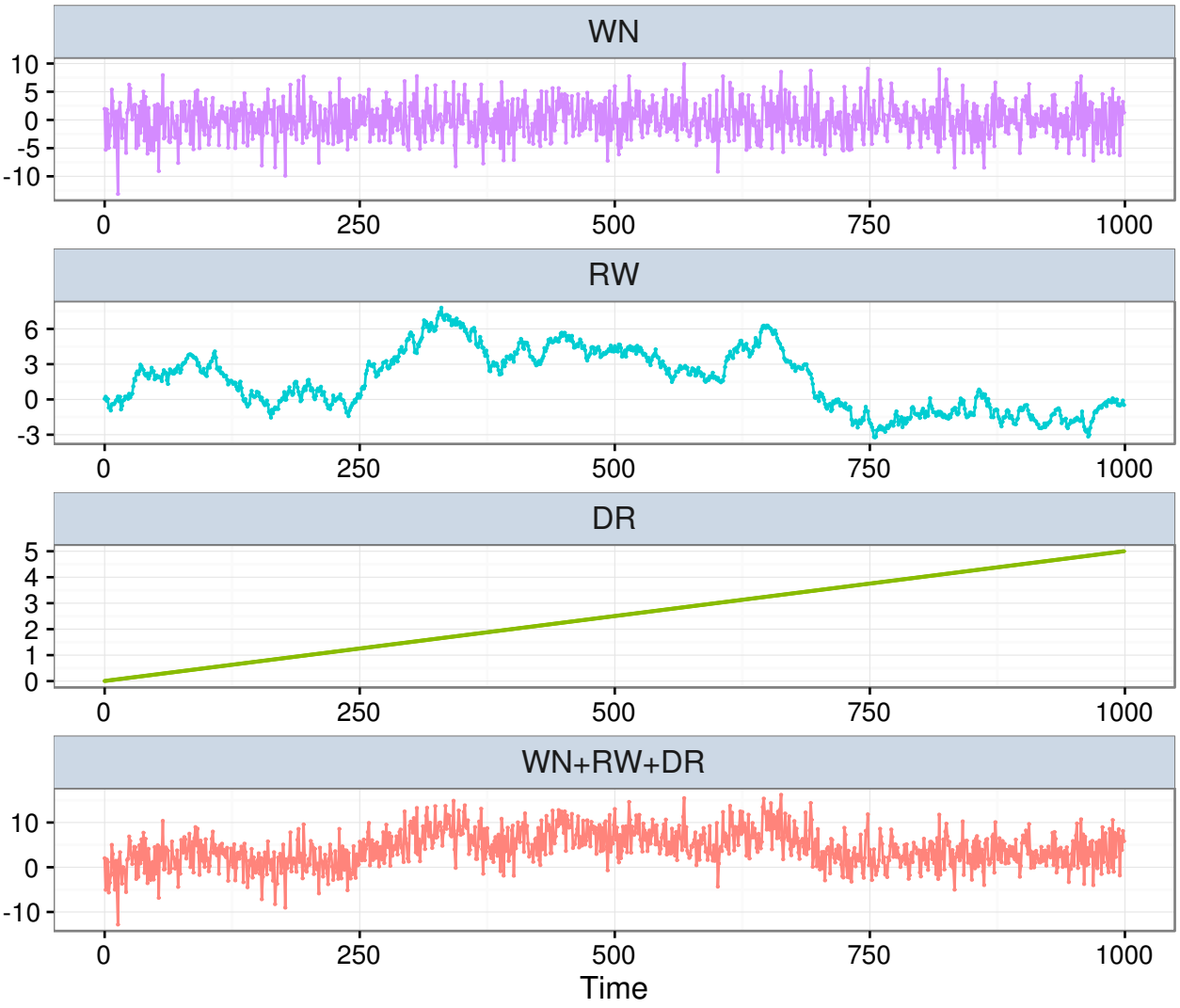
$$\begin{aligned} Y_t &= Y_{t-1} + W_t + \delta \\ X_t &= Y_t + Z_t, \end{aligned}$$

where $W_t$ and $Z_t$ are two independent Gaussian white noise processes. This model is often used as a first tool to approximate the number of individuals in the context ecological population dynamics. For example, suppose we want to study the population of Chamois in the Swiss Alps. Let $Y_t$ denote the "true" number of individuals in this population at time $t$. It is reasonable that $Y_t$ is (approximately) the population at the previous time $t-1$ (e.g the previous year) plus a random variation and a drift. This random variation is due to the natural randomness in ecological population dynamics and reflects the changes in the number of predators, in the aboundance of food or in the weather conditions. On the other hand, the drift is often of particular interest for ecologists as it can be used to determine the "long" term trends of the population (e.g. is the population increasing, stable or decreasing). Of course, $Y_t$ (the number of individauls) is typically unknown and we observe a noisy version of it, denoted as $X_t$. This process corresponds to the true population plus a measurement error since some individuals may not be observed while others may have been counted several times. Interestingly, this process can clearly be expressed as a *latent time series model* (or composite stochastic process) as follows:

$$\begin{aligned} R_t &= R_{t-1} + W_t \\ S_t &= \delta t \\ X_t &= R_t + S_t + Z_t, \end{aligned}$$

where $R_t$, $S_t$ and $Z_t$ denote, respectively, a random walk, a drift and a white noise. The code below can be used to simulate such data:

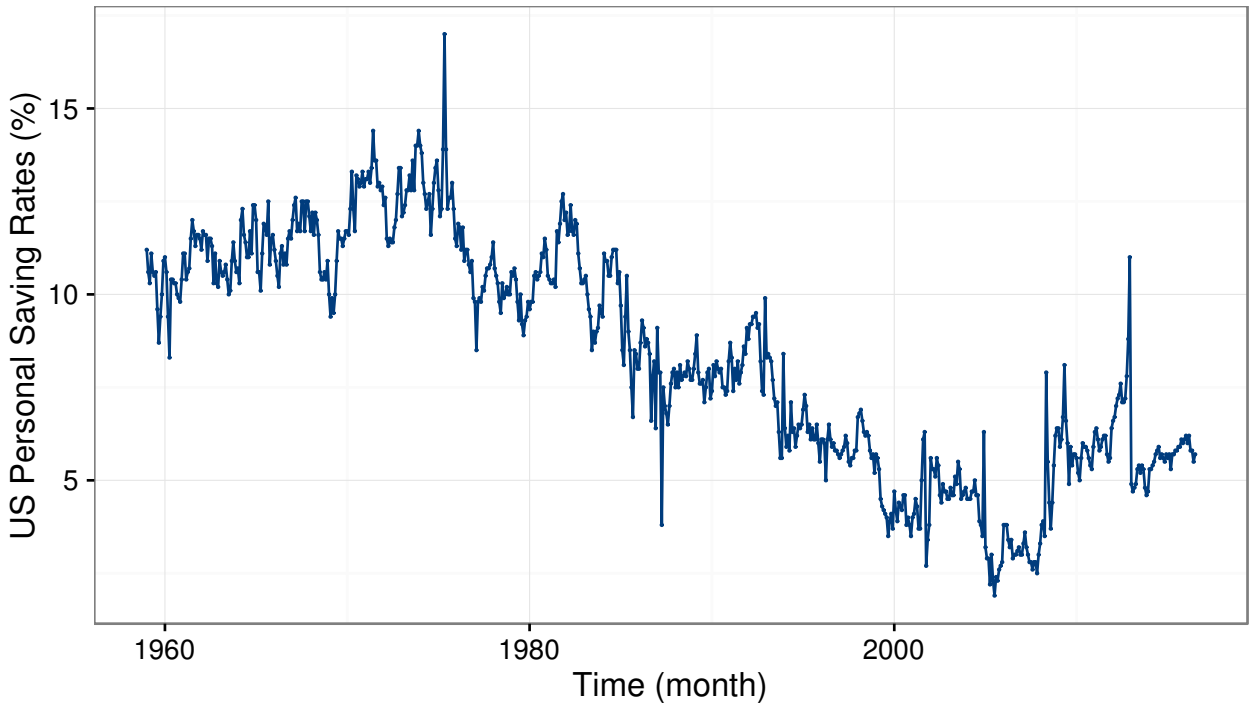
```
n = 1000                                    # process length
delta = 0.005                               # delta parameter (drift)
sigma2 = 10                                 # variance parameter (white noise)
gamma2 = 0.1                                # innovation variance (random walk)
model = WN(sigma2 = sigma2) + RW(gamma2 = gamma2) + DR(omega = delta)
Xt = gen.lts(model, N = n)
plot(Xt)
```




In the above graph, the first three plots represent the latent (unobserved) processes (i.e. white noise, random walk and drift) and the last one represents the sum of the three (i.e. $(X_t)$).

Let us consider a real example where these latent processes are useful to describe (and predict) the behavior of economic variables such as Personal Saving Rates (PSR). A process that is used for these settings is the "random-walk-plus-noise" model, meaning that the data can be explained by a random walk process in addition to which we observe some other process (e.g. a white noise model, an autoregressive model such as an AR(1), etc.). The PSR taken from the Federal Reserve of St. Louis from January 1, 1959, to May 1, 2015, is presented in the following plot:

```
# Saving Rates
data("savingrt", package="smacdata")

# Plot time series
autoplot(savingrt) + ylab("US Personal Saving Rates (%)")
```




It can be observed that the mean of the process seems to vary over time, suggesting that a random walk can indeed be considered as a possible model to explain this data. In addition, aside from some "spikes" and occasional sudden changes, the observations appear to gradually change from one time point to the other, suggesting that some other form of dependence between them could exist.

# Chapter 2

# Autocorrelation and Stationarity

> "*One of the first things taught in introductory statistics textbooks is that correlation is not causation. It is also one of the first things forgotten.*", Thomas Sowell

In this chapter we will discuss and formalize how knowledge about $X_{t-1}$ (or more generally about all the information from the past $\Omega_t$) can provide us with some information about the properties of $X_t$. In particular, we will consider the correlation (or covariance) of $(X_t)$ at different times such as $\text{corr}(X_t, X_{t+h})$. This "form" of correlation (covariance) is called the *autocorrelation* (*autocovariance*) and is a very useful tool in time series analysis. However, if we do not assume that a time series is characterized by a certain form of "stability", it would be rather difficult to estimate $\text{corr}(X_t, X_{t+h})$ as this quantity would depend on both $t$ and $h$ leading to more parameters to estimate than observations available. Therefore, the concept of *stationarity* is convenient in this context as it allows (among other things) to assume that

$$\text{corr}(X_t, X_{t+h}) = \text{corr}(X_{t+j}, X_{t+h+j}), \quad \text{for all } j,$$

implying that the autocorrelation (or autocovariance) is only a function of the lag between observations (rather than time itself). These two concepts (i.e. autocorrelation and stationarity) will be discussed in this chapter. Before moving on, it is helpful to remember that correlation (or autocorrelation) is only appropriate to measure a very specific kind of dependence, i.e. the linear dependence. There are many other forms of dependence as illustrated in the bottom panels of the graph below, which all have a (true) zero correlation:

Several other metrics have been introduced in the literature to assess the degree of "dependence" of two random variables however this goes beyond the material discussed in this chapter.

## 2.1 The Autocorrelation and Autocovariance Functions

**Definition 1.** *The autocovariance function of a series $(X_t)$ is defined as*

$$\gamma_x(t, t+h) = \text{cov}(X_t, X_{t+h}),$$

where the definition of covariance is given by:

$$\text{cov}(X_t, X_{t+h}) = \mathbb{E}[X_t X_{t+h}] - \mathbb{E}[X_t]\,\mathbb{E}[X_{t+h}].$$

Similarly, the above expectations are defined to be:

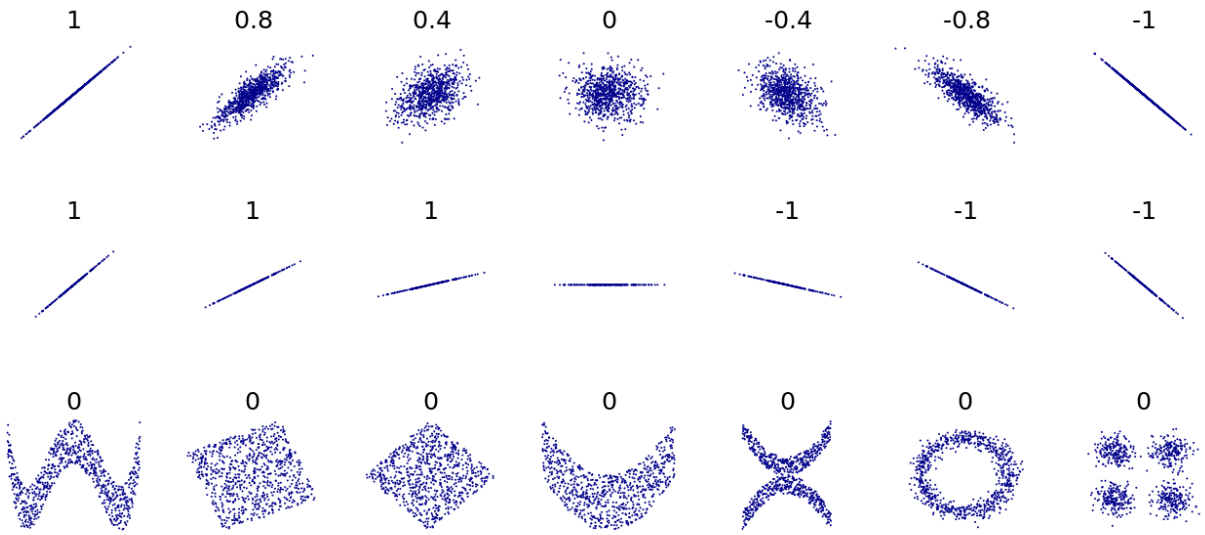



Figure 2.1: Different forms of dependence and their Pearson's r value

$$\mathbb{E}[X_t] = \int_{-\infty}^{\infty} x \cdot f_t(x)\, dx,$$
$$\mathbb{E}[X_t X_{t+h}] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1 x_2 \cdot f_{t,t+h}(x_1, x_2)\, dx_1 dx_2,$$

where $f_t(x)$ and $f_{t,t+h}(x_1, x_2)$ denote, respectively, the density of $X_t$ and the joint density of the pair $(X_t, X_{t+h})$. Since we generally consider stochastic processes with constant zero mean we often have

$$\gamma_x(t, t+h) = \mathbb{E}[X_t X_{t+h}].$$

In addition, we normally drop the subscript referring to the time series (i.e. $x$ in this case) if it is clear from the context which time series the autocovariance refers to. For example, we generally use $\gamma(t, t+h)$ instead of $\gamma_x(t, t+h)$. Moreover, the notation is even further simplified when the covariance of $X_t$ and $X_{t+h}$ is the same as that of $X_{t+j}$ and $X_{t+h+j}$ (for all $j$), i.e. the covariance depends only on the time between observations and not on the specific time $t$. This is an important property called *stationarity*, which will be discuss in the next section. In this case, we simply use to following notation:

$$\gamma(h) = \operatorname{cov}(X_t, X_{t+h}).$$

This notation will generally be used throughout the text and implicitly assume certain properties (i.e. stationarity) on the process $(X_t)$. Several remarks can be made on the autocovariance:

1. The autocovariance function is *symmetric*. That is, $\gamma(h) = \gamma(-h)$ since $\operatorname{cov}(X_t, X_{t+h}) = \operatorname{cov}(X_{t+h}, X_t)$.
2. The autocovariance function "contains" the variance of the process as $\operatorname{var}(X_t) = \gamma(0)$.
3. We have that $|\gamma(h)| \leq \gamma(0)$ for all $h$. The proof of this inequality is direct and follows from the Cauchy-Schwarz inequality, i.e.

$$\begin{aligned}(|\gamma(h)|)^2 = \gamma(h)^2 &= (\mathbb{E}[(X_t - \mathbb{E}[X_t])(X_{t+h} - \mathbb{E}[X_{t+h}])])^2 \\ &\leq \mathbb{E}\left[(X_t - \mathbb{E}[X_t])^2\right] \mathbb{E}\left[(X_{t+h} - \mathbb{E}[X_{t+h}])^2\right] = \gamma(0)^2.\end{aligned}$$

4. Just as any covariance, $\gamma(h)$ is "scale dependent" since $\gamma(h) \in \mathbb{R}$, or $-\infty \leq \gamma(h) \leq +\infty$. We therefore have:
   - if $|\gamma(h)|$ is "close" to zero, then $X_t$ and $X_{t+h}$ are "weakly" (linearly) dependent;
   - if $|\gamma(h)|$ is "far" from zero, then the two random variable present a "strong" (linear) dependence. However it is generally difficult to asses what "close" and "far" from zero means in this case.
5. $\gamma(h) = 0$ does not imply that $X_t$ and $X_{t+h}$ are independent but simply $X_t$ and $X_{t+h}$ are uncorrelated. The independence is only implied by $\gamma(h) = 0$ in the jointly Gaussian case.

As hinted in the introduction, an important related statistic is the correlation of $X_t$ with $X_{t+h}$ or *autocorrelation*, which is defined as

$$\rho(h) = \text{corr}(X_t, X_{t+h}) = \frac{\text{cov}(X_t, X_{t+h})}{\sigma_{X_t}\sigma_{X_{t+h}}} = \frac{\gamma(h)}{\gamma(0)}.$$

Similarly to $\gamma(h)$, it is important to note that the above notation implies that the autocorrelation function is only a function of the lag $h$ between observations. Thus, autocovariances and autocorrelations are one possible way to describe the joint distribution of a time series. Indeed, the correlation of $X_t$ with $X_{t+1}$ is an obvious measure of how *persistent* a time series is.

Remember that just as with any correlation:

1. $\rho(h)$ is "scale free" so it is much easier to interpret than $\gamma(h)$.
2. $|\rho(h)| \leq 1$ since $|\gamma(h)| \leq \gamma(0)$.
3. **Causation and correlation are two very different things!**

### 2.1.1 A Fundamental Representation

Autocovariances and autocorrelations also turn out to be very useful tools as they are one of the *fundamental representations* of time series. Indeed, if we consider a zero mean normally distributed process, it is clear that its joint distribution is fully characterized by the autocovariances $\mathbb{E}[X_t X_{t+h}]$ (since the joint probability density only depends of these covariances). Once we know the autocovariances we know *everything* there is to know about the process and therefore: *if two processes have the same autocovariance function, then they are the same process.*

### 2.1.2 Admissible Autocorrelation Functions

Since the autocorrelation is related to a fundamental representation of time series, it implies that one might be able to define a stochastic process by picking a set of autocorrelation values (assuming for example that $\text{var}(X_t) = 1$). However, it turns out that not every collection of numbers, say $\{\rho_1, \rho_2, ...\}$, can represent the autocorrelation of a process. Indeed, two conditions are required to ensure the validity of an autocorrelation sequence:

1. $\max_j \ |\rho_j| \leq 1$.
2. $\text{var}\left[\sum_{j=0}^{\infty} \alpha_j X_{t-j}\right] \geq 0$ for all $\{\alpha_0, \alpha_1, ...\}$.

The first condition is obvious and simply reflects the fact that $|\rho(h)| \leq 1$ but the second is far more difficult to verify. To further our understanding of the latter we let $\alpha_j = 0$ for $j > 1$, then condition 2 implies that

$$\text{var}[\alpha_0 X_t + \alpha_1 X_{t-1}] = \gamma_0 \begin{bmatrix} \alpha_0 & \alpha_1 \end{bmatrix} \begin{bmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \alpha_1 \end{bmatrix} \geq 0.$$

Thus, the matrix

$$\boldsymbol{A}_1 = \begin{bmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{bmatrix}$$

must be positive semi-definite. Taking the determinant we have

$$\det(\boldsymbol{A}_1) = 1 - \rho_1^2$$

implying that the condition $|\rho_1| \leq 1$ must be respected. Now, let $\alpha_j = 0$ for $j > 2$, then we must verify that:

$$\text{var}[\alpha_0 X_t + \alpha_1 X_{t-1} + \alpha_2 X_{t-2}] = \gamma_0 \begin{bmatrix} \alpha_0 & \alpha_1 & \alpha_2 \end{bmatrix} \begin{bmatrix} 1 & \rho_1 & \rho_2 \\ \rho_1 & 1 & \rho_1 \\ \rho_2 & \rho_1 & 1 \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \end{bmatrix} \geq 0.$$

Again, this implies that the matrix

$$\boldsymbol{A}_2 = \begin{bmatrix} 1 & \rho_1 & \rho_2 \\ \rho_1 & 1 & \rho_1 \\ \rho_2 & \rho_1 & 1 \end{bmatrix}$$

must be positive semi-definite and it is easy to verify that

$$\det(\boldsymbol{A}_2) = (1 - \rho_2)\left(-2\rho_1^2 + \rho_2 + 1\right).$$

Thus, this implies that

$$\begin{aligned} &-2\rho_1^2 + \rho_2 + 1 \geq 0 \Rightarrow 1 \geq \rho_2 \geq 2\rho_1^2 - 1 \\ &\Rightarrow 1 - \rho_1^2 \geq \rho_2 - \rho_1^2 \geq -(1 - \rho_1^2) \\ &\Rightarrow 1 \geq \frac{\rho_2 - \rho_1^2}{1 - \rho_1^2} \geq -1. \end{aligned}$$

Therefore, $\rho_1$ and $\rho_2$ must lie in a parabolic shaped region defined by the above inequalities as illustrated in Figure 2.2.

From our derivation it is clear that the restrictions on the autocorrelation are very complicated thereby justifying the need for other forms of fundamental representation which we will explore later in this text. Before moving on to the estimation of the autocorrelation and autocovariance functions, we must first discuss the stationarity of $(X_t)$, which will provide a convenient framework in which $\gamma(h)$ and $\rho(h)$ can be used (rather that $\gamma(t, t+h)$ for example) and (easily) estimated.

## 2.2 Stationarity

There are two kinds of stationarity that are commonly used. They are defined as follows:

**Definition 2.** *A process $(X_t)$ is strongly stationary or strictly stationary if the joint probability distribution of $(X_{t-h}, ..., X_t, ..., X_{t+h})$ is independent of $t$ for all $h$.*

**Definition 3.** *A process $(X_t)$ is weakly stationary, covariance stationary or second order stationary if $\mathbb{E}[X_t]$, $\mathbb{E}[X_t^2]$ are finite and $\mathbb{E}[X_t X_{t-h}]$ depends only on $h$ and not on $t$.*

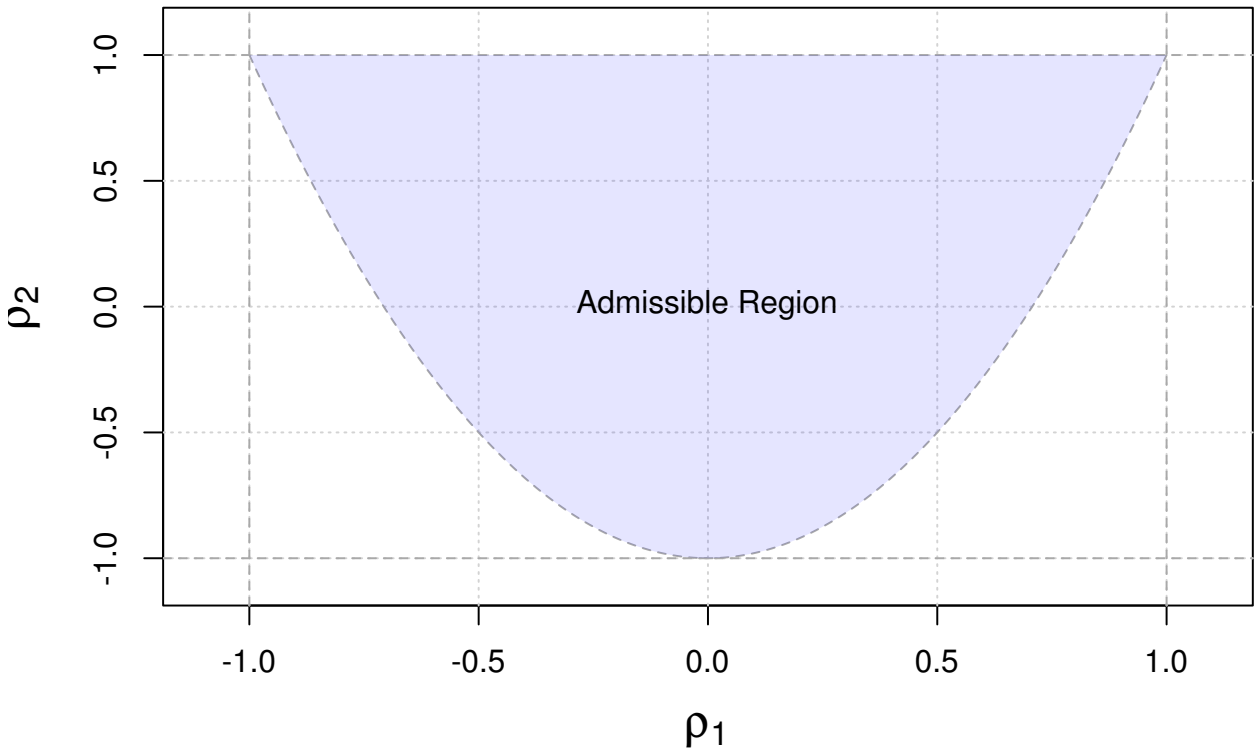



Figure 2.2: Admissible autocorrelation functions

These types of stationarity are *not equivalent* and the presence of one kind of stationarity does not imply the other. That is, a time series can be strongly stationary but not weakly stationary and vice versa. In some cases, a time series can be both strongly and weakly stationary and this is occurs, for example, in the (jointly) Gaussian case. Stationarity of $(X_t)$ matters because *it provides the framework in which averaging dependent data makes sense* thereby allowing to easily estimate quantities such as the autocorrelation function.

Several remarks and comments can be made on these definitions:

- As mentioned earlier, strong stationarity *does not imply* weak stationarity.

**Example 7.** an iid Cauchy process is strongly but not weakly stationary.

- Weak stationarity *does not imply* strong stationarity.

**Example 8.** Consider the following weak white noise process:

$$X_t = \begin{cases} U_t & \text{if } t \in \{2k : \ k \in \mathbb{Z}\}, \\ V_t & \text{if } t \in \{2k+1 : \ k \in \mathbb{Z}\}, \end{cases}$$

where $U_t \overset{iid}{\sim} N(1,1)$ and $V_t \overset{iid}{\sim} \mathcal{E}(1)$ is a weakly stationary process that is *not* strongly stationary.

- Strong stationarity combined with bounded values of $\mathbb{E}[X_t]$ and $\mathbb{E}[X_t^2]$ *implies* weak stationarity
- Weak stationarity combined with normality distributed processes *implies* strong stationarity.

### 2.2.1 Assessing Weak Stationarity of Time Series Models

It is important to understand how to verify if a postulated model is (weakly) stationary. In order to do so, we must ensure that our model satisfies the following three properties:

1. $\mathbb{E}[X_t] = \mu_t = \mu < \infty$,
2. $\text{var}[X_t] = \sigma_t^2 = \sigma^2 < \infty$,
3. $\text{cov}(X_t, X_{t+h}) = \gamma(h)$.

In the following examples we evaluate the stationarity of the processes introduced in Section **??**.

**Example 9.** [Gaussian White Noise] It is easy to verify that this process is stationary. Indeed, we have:

1. $\mathbb{E}[X_t] = 0$,
2. $\gamma(0) = \sigma^2 < \infty$,
3. $\gamma(h) = 0$ for $|h| > 0$.

**Example 10.** [Random Walk] To evaluate the stationarity of this process we first derive its properties:

1. We begin by calculating the expectation of the process

$$\mathbb{E}[X_t] = \mathbb{E}[X_{t-1} + W_t] = \mathbb{E}\left[\sum_{i=1}^{t} W_i + X_0\right] = \mathbb{E}\left[\sum_{i=1}^{t} W_i\right] + c = c.$$

   Observe that the mean obtained is constant since it depends only on the value of the first term in the sequence.

2. Next, after finding the mean to be constant, we calculate the variance to check stationarity:

$$\begin{aligned}\text{var}(X_t) &= \text{var}\left(\sum_{i=1}^{t} W_t + X_0\right) = \text{var}\left(\sum_{i=1}^{t} W_i\right) + \underbrace{\text{var}(X_0)}_{=0} \\ &= \sum_{i=1}^{t} \text{var}(W_i) = t\sigma_w^2,\end{aligned}$$

   where $\sigma_w^2 = \text{var}(W_t)$. Therefore, the variance depends on time $t$ contradicting our second property. Moreover, we have:

$$\lim_{t\to\infty} \text{var}(X_t) = \infty.$$

   This process is therefore not weakly stationary.

3. Regarding the autocovariance of a random walk we have:

$$\begin{aligned}\gamma(h) = \text{cov}(X_t, X_{t+h}) &= \text{cov}\left(\sum_{i=1}^{t} W_i, \sum_{j=1}^{t+h} W_j\right) = \text{cov}\left(\sum_{i=1}^{t} W_i, \sum_{j=1}^{t} W_j\right) \\ &= \min(t, t+h)\,\sigma_w^2 = (t + \min(0, h))\,\sigma_w^2,\end{aligned}$$

   which further illustrates the non-stationarity of this process.

Moreover, the autocorrelation of this process is given by

$$\rho(h) = \frac{t + \min(0, h)}{\sqrt{t}\sqrt{t+h}},$$

implying (for a fixed $h$) that

$$\lim_{t\to\infty} \rho(h) = 1.$$

In the following simulated example, we illustrate the non-stationary feature of such a process:

```
# In this example, we simulate a large number of random walks

# Number of simulated processes
B = 200

# Length of random walks
n = 1000

# Output matrix
out = matrix(NA,B,n)

# Set seed for reproducibility
set.seed(6182)

# Simulate Data
for (i in seq_len(B)){
  # Simulate random walk
  Xt = gen.gts(RW(gamma=1), N = n)

  # Store process
  out[i,] = Xt
}

# Plot random walks
plot(NA, xlim = c(1,n), ylim = range(out), xlab = "Time", ylab = " ")
grid()
color = sample(topo.colors(B, alpha = 0.5))
grid()
for (i in seq_len(B)){
  lines(out[i,], col = color[i])
}

# Add 95% confidence region
lines(1:n, 1.96*sqrt(1:n), col = 2, lwd = 2, lty = 2)
lines(1:n, -1.96*sqrt(1:n), col = 2, lwd = 2, lty = 2)
```

In the plot, two hundred simulated random walks are plotted along with the theoretical 95% confidence intervals (red-dashed lines). The relationship between time and variance can clearly be observed (i.e. the variance of the process increases with the time).

**Example 11.** [Moving Average of Order 1] Similarly to our previous examples, we attempt to verify the stationary properties for the MA(1) model defined in Section 1.3.4:

1. $$\mathbb{E}[X_t] = \mathbb{E}[\theta_1 W_{t-1} + W_t] = \theta_1 \mathbb{E}[W_{t-1}] + \mathbb{E}[W_t] = 0.$$
2. $$\text{var}(X_t) = \theta_1^2 \,\text{var}(W_{t-1}) + \text{var}(W_t) = \left(1+\theta^2\right)\sigma_w^2.$$

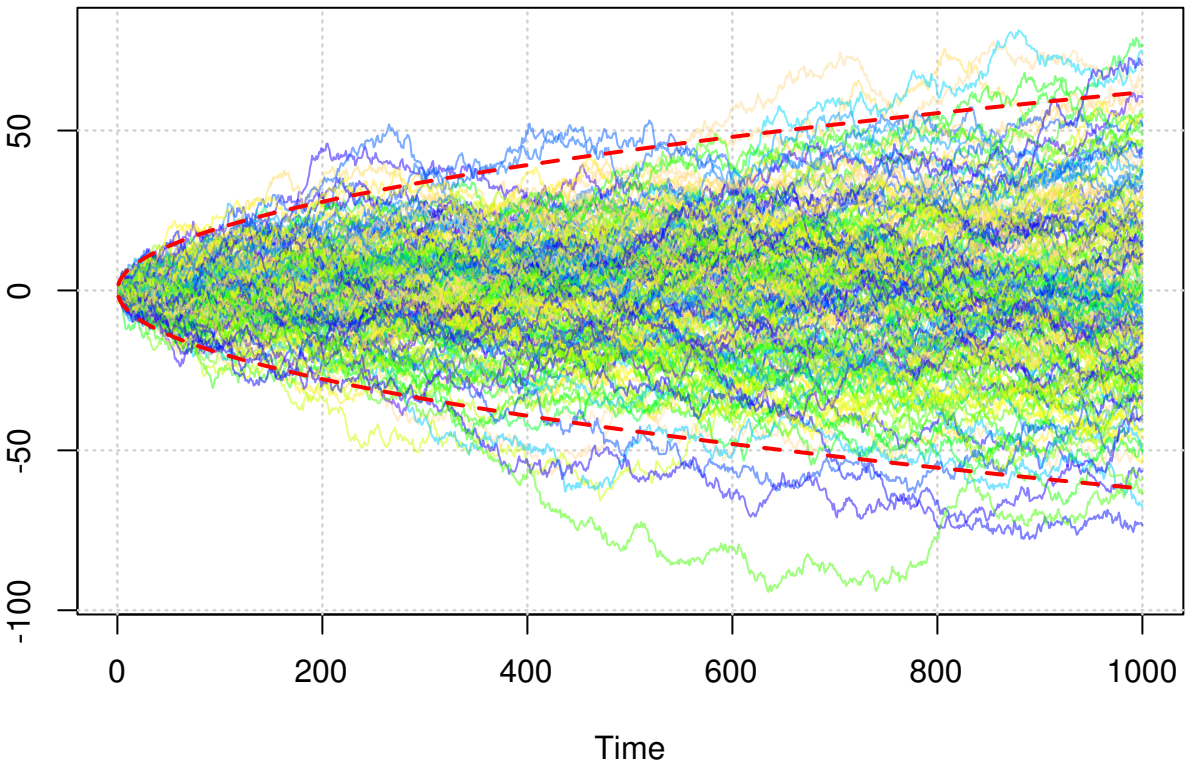



Figure 2.3: Two hundred simulated random walks.

3. Regarding the autocovariance, we have

$$\begin{aligned}\operatorname{cov}(X_t, X_{t+h}) &= \mathbb{E}\left[\left(X_t - \mathbb{E}\left[X_t\right]\right)\left(X_{t+h} - \mathbb{E}\left[X_{t+h}\right]\right)\right] = \mathbb{E}\left[X_t X_{t+h}\right] \\ &= \mathbb{E}\left[\left(\theta W_{t-1} + W_t\right)\left(\theta W_{t+h-1} + W_{t+h}\right)\right] \\ &= \mathbb{E}\left[\theta^2 W_{t-1} W_{t+h-1} + \theta W_t W_{t+h} + \theta W_{t-1} W_{t+h} + W_t W_{t+h}\right].\end{aligned}$$

It is easy to see that $\mathbb{E}\left[W_t W_{t+h}\right] = \mathbf{1}_{\{h=0\}}\sigma_w^2$ and therefore, we obtain

$$\operatorname{cov}(X_t, X_{t+h}) = \left(\theta^2 \mathbf{1}_{\{h=0\}} + \theta \mathbf{1}_{\{h=1\}} + \theta \mathbf{1}_{\{h=-1\}} + \mathbf{1}_{\{h=0\}}\right)\sigma_w^2$$

implying the following autocovariance function:

$$\gamma(h) = \begin{cases} \left(\theta^2+1\right)\sigma_w^2 & h = 0 \\ \theta\sigma_w^2 & |h| = 1 \\ 0 & |h| > 1 \end{cases}.$$

Therefore, an MA(1) process is weakly stationary since both the mean and variance are constant over time and its covariance function is only a function of the lag $h$. Finally, we can easily obtain the autocorrelation for this process, which is given by

$$\rho(h) = \begin{cases} 1 & h = 0 \\ \frac{\theta\sigma_w^2}{(\theta^2+1)\sigma_w^2} = \frac{\theta}{\theta^2+1} & |h| = 1 \\ 0 & |h| > 1 \end{cases}.$$

Interestingly, we can note that $|\rho(1)| \leq 0.5$.

**Example 12.** [Autoregressive of Order 1] As another example, we shall verify the stationary properties for the AR(1) model defined in Section 1.3.3.

Using the *backsubstitution* technique, we can rearrange an AR(1) process so that it is written in a more compact form, i.e.

$$\begin{aligned} X_t &= \phi X_{t-1} + W_t = \phi\left[\phi X_{t-2} + W_{t-1}\right] + W_t = \phi^2 X_{t-2} + \phi W_{t-1} + W_t \\ &\vdots \\ &= \phi^k X_{t-k} + \sum_{j=0}^{k-1} \phi^j W_{t-j}. \end{aligned}$$

By taking the limit in $k$ (which is perfectly valid as we assume $t \in \mathbb{Z}$) and assuming $|\phi| < 1$, we obtain

$$X_t = \lim_{k\to\infty} X_t = \sum_{j=0}^{\infty} \phi^j W_{t-j}$$

and therefore such process can be interpreted as a linear combination of the white noise $(W_t)$ and corresponds (as we will later on) to an MA($\infty$). In addition, the requirement $|\phi| < 1$ turns out to be extremely useful as the above formula is related to Geometric series which would diverge if $\phi \geq 1$. Indeed, remember that an infinite (converging) Geometric series is given by

$$\sum_{k=0}^{\infty} ar^k = \frac{a}{1-r}, \quad \text{if } |r| < 1.$$

With this setup, we demonstrate how crucial this property is by calculating each of the requirements of a stationary process.

1. First, we will check if the mean is stationary. In this case, we opt to use limits to derive the expectation

$$\begin{aligned} \mathbb{E}\left[X_t\right] &= \lim_{k\to\infty} \mathbb{E}\left[\phi^k X_{t-k} + \sum_{j=0}^{k-1} \phi^j W_{t-j}\right] \\ &= \lim_{k\to\infty} \underbrace{\phi^k \mathbb{E}[X_{t-k}]}_{=0} + \lim_{k\to\infty} \sum_{j=0}^{k-1} \phi^j \underbrace{\mathbb{E}\left[W_{t-j}\right]}_{=0} = 0. \end{aligned}$$

   As expected, the mean is zero and, hence, the first criteria for weak stationarity is satisfied.
2. Next, we opt to determine the variance of the process

$$\begin{aligned} \operatorname{var}\left(X_t\right) &= \lim_{k\to\infty} \operatorname{var}\left(\phi^k X_{t-k} + \sum_{j=0}^{k-1} \phi^j W_{t-j}\right) = \lim_{k\to\infty} \sum_{j=0}^{k-1} \phi^{2j} \operatorname{var}\left(W_{t-j}\right) \\ &= \lim_{k\to\infty} \sum_{j=0}^{k-1} \sigma_W^2 \, \phi^{2j} = \underbrace{\frac{\sigma_W^2}{1-\phi^2}}_{\text{Geom. Serie}}. \end{aligned}$$

   Once again, the above result only holds because we are able to use the geometric series convergence as a result of $|\phi| < 1$.

3. Finally, we consider the autocovariance of an AR(1). For $h > 0$, we have

$$\gamma(h) = \operatorname{cov}(X_t, X_{t+h}) = \phi \operatorname{cov}(X_t, X_{t+h-1}) = \phi\,\gamma(h-1).$$

Therefore, we using the symmetry of the autocovariance we have that

$$\gamma(h) = \phi^{|h|}\,\gamma(0).$$

Both the mean and variance do not depend on time in addition the autocovariance function can be viewed as function dependent on only lags and, thus, the AR(1) process is weakly stationary if $|\phi| < 1$. Lastly, we can obtain the autocorrelation for this process. Indeed, for $h > 0$, we have

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)} = \frac{\phi\gamma(h-1)}{\gamma(0)} = \phi\rho(h-1).$$

After fully simplifying, we obtain

$$\rho(h) = \phi^{|h|}.$$

Thus, the autocorrelation function for an AR(1) exhibits a *geometric decay*, meaning, the smaller the $|\phi|$, the faster the autocorrelation reaches zero. If $|\phi|$ is close to 1, then the decay rate is slower.

## 2.3 Estimation of Moments of Stationary Processes

In this section, we discuss how moments and related quantities of stationary process can be estimated. Informally speaking, the use of "averages" is meaningful for such processes suggesting that classical moments estimators can be employed. Indeed, suppose that one is interested in estimating $\alpha \equiv \mathbb{E}[m(X_t)]$, where $m(\cdot)$ is a known function of $X_t$. If $(X_t)$ is a strongly stationary process, we have

$$\alpha = \int m(x)\, f(x) dx$$

where $f(x)$ denotes the density of $(X_t)$, $\forall t$. Replacing $f(x)$ by $f_n(x)$, the empirical density, we obtain the following estimator

$$\hat{\alpha} = \frac{1}{n}\sum_{i=1}^{n} m(x_i).$$

In the next subsection, we examine how this simple idea can be used to estimate the mean, autocovariance and autocorrelation functions. Moreover, we discuss some of the properties of these estimators.

### 2.3.1 Estimation of the Mean Function

If a time series is stationary, the mean function is constant and a possible estimator of this quantity is, as discussed above, given by

$$\bar{X} = \frac{1}{n}\sum_{t=1}^{n} X_t.$$

Naturally, the $k$-th moment, say $\beta_k \equiv \mathbb{E}[X_t^k]$ can be estimated by

$$\hat{\beta}_k = \frac{1}{n}\sum_{t=1}^{n} X_t^k, \;\; k \in \{x \in \mathbb{N} : 0 < x < \infty\}.$$

The variance of such estimator can be derived as follows:

$$\begin{aligned}
\operatorname{var}\left(\hat{\beta}_k\right) &= \operatorname{var}\left(\frac{1}{n}\sum_{t=1}^{n} X_t^k\right) \\
&= \frac{1}{n^2}\operatorname{var}\left(\begin{bmatrix} 1 & \cdots & 1 \end{bmatrix}_{1\times n}\begin{bmatrix} X_1^k \\ \vdots \\ X_n^k \end{bmatrix}_{n\times 1}\right) \\
&= \frac{1}{n^2}\begin{bmatrix} 1 & \cdots & 1 \end{bmatrix}_{1\times n}\boldsymbol{\Sigma}(k)\begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}_{n\times 1},
\end{aligned} \tag{2.1}$$

where $\boldsymbol{\Sigma}(k) \in \mathbb{R}^{n\times n}$ and its $i$th, $j$-th element is given by

$$\left(\boldsymbol{\Sigma}(k)\right)_{i,j} = \operatorname{cov}\left(X_i^k, X_j^k\right).$$

In the case $k = 1$, ((2.1)) can easily be further simplified. Indeed, we have

$$\begin{aligned}
\operatorname{var}\left(\bar{X}\right) &= \operatorname{var}\left(\frac{1}{n}\sum_{t=1}^{n} X_t\right) \\
&= \frac{1}{n^2}\begin{bmatrix} 1 & \cdots & 1 \end{bmatrix}_{1\times n}\begin{bmatrix} \gamma(0) & \gamma(1) & \cdots & \gamma(n-1) \\ \gamma(1) & \gamma(0) & & \vdots \\ \vdots & & \ddots & \vdots \\ \gamma(n-1) & \cdots & \cdots & \gamma(0) \end{bmatrix}_{n\times n}\begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}_{n\times 1} \\
&= \frac{1}{n^2}\left(n\gamma(0) + 2(n-1)\gamma(1) + 2(n-2)\gamma(2) + \cdots + 2\gamma(n-1)\right) \\
&= \frac{1}{n}\sum_{h=-n}^{n}\left(1 - \frac{|h|}{n}\right)\gamma(h).
\end{aligned}$$

Obviously, when the $(X_t)$ is a white noise, the above formula reduces to the usual $\operatorname{var}\left(\bar{X}\right) = \sigma_w^2/n$. In the following example, we consider the case of an AR(1) process and discussed how $\operatorname{var}\left(\bar{X}\right)$ can be obtained or estimated.

**Example 13.** For an AR(1) we have $\gamma(h) = \phi^h\sigma_w^2\left(1-\phi^2\right)^{-1}$, therefore, we obtain (after some computations):

$$\operatorname{var}\left(\bar{X}\right) = \frac{\sigma_w^2\left(n - 2\phi - n\phi^2 + 2\phi^{n+1}\right)}{n^2\left(1-\phi^2\right)\left(1-\phi\right)^2}. \tag{2.2}$$

Unfortunately, deriving such an exact formula is often difficult when considering more complex models. However, asymptotic approximations are often employed to simplify the calculation. For example, in our case we have

$$\lim_{n\to\infty} n\operatorname{var}\left(\bar{X}\right) = \frac{\sigma_w^2}{\left(1-\phi\right)^2},$$

providing the following approximate formula:

$$\text{var}\left(\bar{X}\right) \approx \frac{\sigma_w^2}{n\left(1-\phi\right)^2}.$$

Alternatively, simulation methods can also be employed. For example, a possible strategy (i.e. parametric bootstrap) could be the following:

1. Simulate a new sample under the postulated model, i.e. $X_t^* \sim F_{\boldsymbol{\theta}}$ (*note:* if $\boldsymbol{\theta}$ is unknown it can be replace by $\hat{\boldsymbol{\theta}}$, a suitable estimator).
2. Compute the statistics of interest on the simulated sample $(X_t^*)$ (i.e. $\bar{X}^*$ in our example).
3. Repeat Steps 1 and 2 $B$ times where $B$ is sufficiently "large" (typically $100 \leq B \leq 10000$).
4. Compute the empirical variance of the statistics of interest based on the $B$ independent replications. In our example, we would have

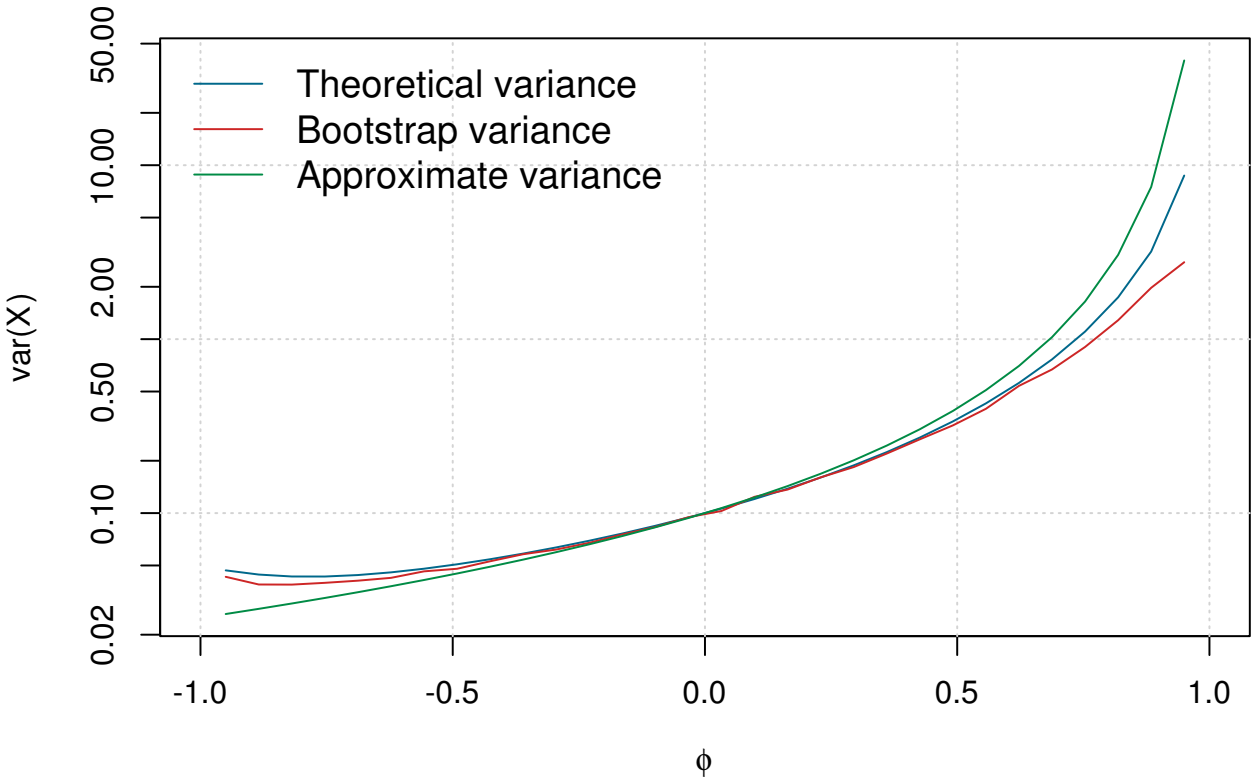
$$\hat{\sigma}_B^2 = \frac{1}{B-1}\sum_{i=1}^{B}\left(\bar{X}_i^* - \bar{X}^*\right)^2, \quad \text{where} \quad \bar{X}^* = \frac{1}{B}\sum_{i=1}^{B}\bar{X}_i^*,$$

and where $\bar{X}_i^*$ denotes the value of the mean estimated on the $i$-th simulated sample.

The figure below generated by the following code compares these three methods for $n = 10$, $B = 1000$, $\sigma^2 = 1$ and a grid of values for $\phi$ going from $-0.95$ to $0.95$:

```
# Define sample size
n = 10

# Number of Monte-Carlo replications
B = 5000

# Define grid of values for phi
phi = seq(from = 0.95, to = -0.95, length.out = 30)

# Define result matrix
result = matrix(NA,B,length(phi))

# Start simulation
for (i in seq_along(phi)){
  # Define model
  model = AR1(phi = phi[i], sigma2 = 1)

  # Monte-Carlo
  for (j in seq_len(B)){
    # Simulate AR(1)
    Xt = gen.gts(model, N = n)

    # Estimate Xbar
    result[j,i] = mean(Xt)
  }
}

# Estimate variance of Xbar
```

```
var.Xbar = apply(result,2,var)

# Compute theoretical variance
var.theo = (n - 2*phi - n*phi^2 + 2*phi^(n+1))/(n^2*(1-phi^2)*(1-phi)^2)

# Compute (approximate) variance
var.approx = 1/(n*(1-phi)^2)

# Compare variance estimations
plot(NA, xlim = c(-1,1), ylim = range(var.approx), log = "y",
     ylab = expression(paste("var(", bar(X), ")")),
     xlab= expression(phi), cex.lab = 1)
grid()
lines(phi,var.theo, col = "deepskyblue4")
lines(phi, var.Xbar, col = "firebrick3")
lines(phi,var.approx, col = "springgreen4")
legend("topleft",c("Theoretical variance","Bootstrap variance","Approximate variance"),
       col = c("deepskyblue4","firebrick3","springgreen4"), lty = 1,
       bty = "n",bg = "white", box.col = "white", cex = 1.2)
```




It can be observed that the variance of $\bar{X}$ typically increases with the $\phi$. As expected when $\phi = 0$ we have $\text{var}(\bar{X}) = 1/n$ as in this case the process is a white noise. Moreover, the bootstrap approach appears to approximate well the curve of (@ref(eq:chap2_exAR1)) while the asymptotic formula provides a reasonable approximate for $\phi$ being between -0.5 and 0.5. Naturally, the quality of this approximation would be far better for larger sample size (here we consider $n = 10$, which is a little "extreme").

### 2.3.2 Sample Autocovariance and Autocorrelation Functions

A natural estimator of the *autocovariance function* is given by:

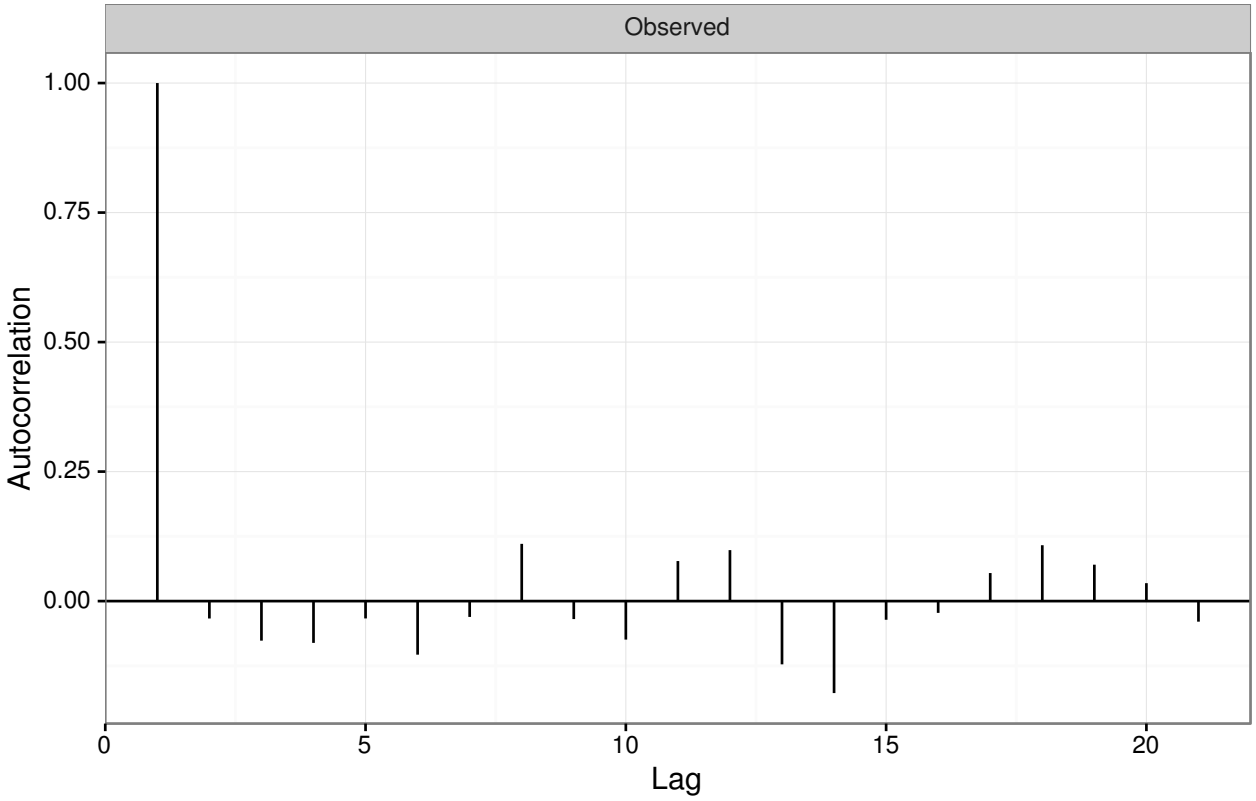
$$\hat{\gamma}(h) = \frac{1}{T}\sum_{t=1}^{T-h}\left(X_t - \bar{X}\right)\left(X_{t+h} - \bar{X}\right)$$

leading the following "plug-in" estimator of the *autocorrelation function*

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}.$$

A graphical representation of the autocorrelation function is often the first step for any time series analysis (again assuming the process to be stationary). Consider the following simulated example:

```
# Load package
library("gmwm")

# Set seed for reproducibility
set.seed(2241)

# Simulate 100 observation from a Gaussian white noise
Xt = gen.gts(WN(sigma2 = 1), N = 100)

# Compute autocorrelation
acf_Xt = ACF(Xt)

# Plot autocorrelation
plot(acf_Xt, show.ci = FALSE)
```



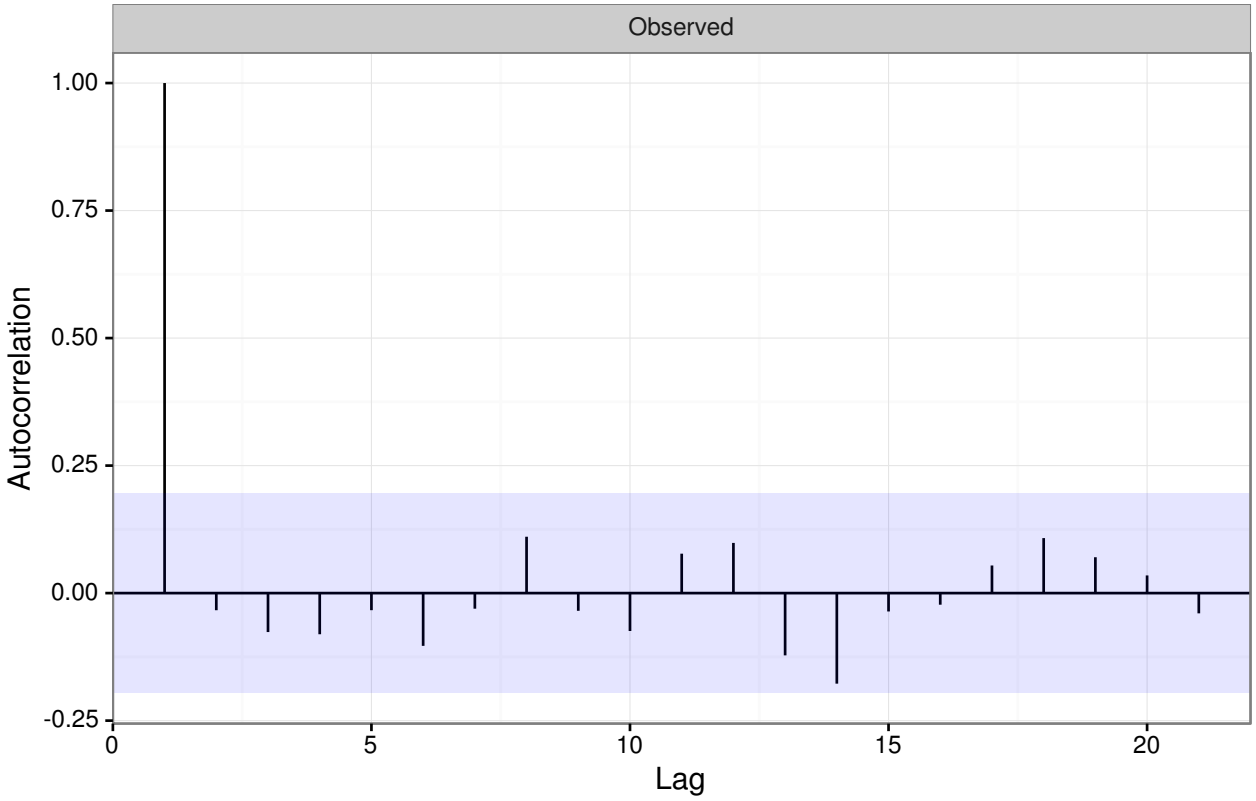

In this example, the true autocorrelation is equal to zero at any lag $h \neq 0$ but obviously the estimated autocorrelations are random variables and are not equal to their true values. It would therefore be useful to have some knowledge about the variability of the sample autocorrelations (under some conditions) to assess whether the data comes from a completely random series or presents some significant correlation at some lags. The following result provides an asymptotic solution to this problem:

**Theorem 1.** *If $X_t$ is a strong white noise with finite fourth moment, then $\hat{\rho}(h)$ is approximately normally distributed with mean 0 and variance $n^{-1}$ for all fixed $h$.*

The proof of this Theorem is given in Appendix A.

Using this result, we now have an approximate method to assess whether peaks in the sample autocorrelation are significant by determining whether the observed peak lies outside the interval $\pm 2/\sqrt{T}$ (i.e. an approximate 95% confidence interval). Returning to our previous example and adding confidence bands in the previous graph, we obtain:

```
# Plot autocorrelation with confidence bands
plot(acf_Xt)
```




It can now be observed that most peaks lie within the interval $\pm 2/\sqrt{T}$ suggesting that the true data generating process is uncorrelated.

**Example 14.** To illustrate how the autocorrelation function can be used to reveal some "features" of a time series we download the level of the Standard & Poor's 500 index, often abbreviated as the S&P 500. This financial index is based on the market capitalization of 500 large companies having common stock listed on the New York Stock Exchange or the NASDAQ Stock Market. The graph below shows the index level and daily returns from 1990.

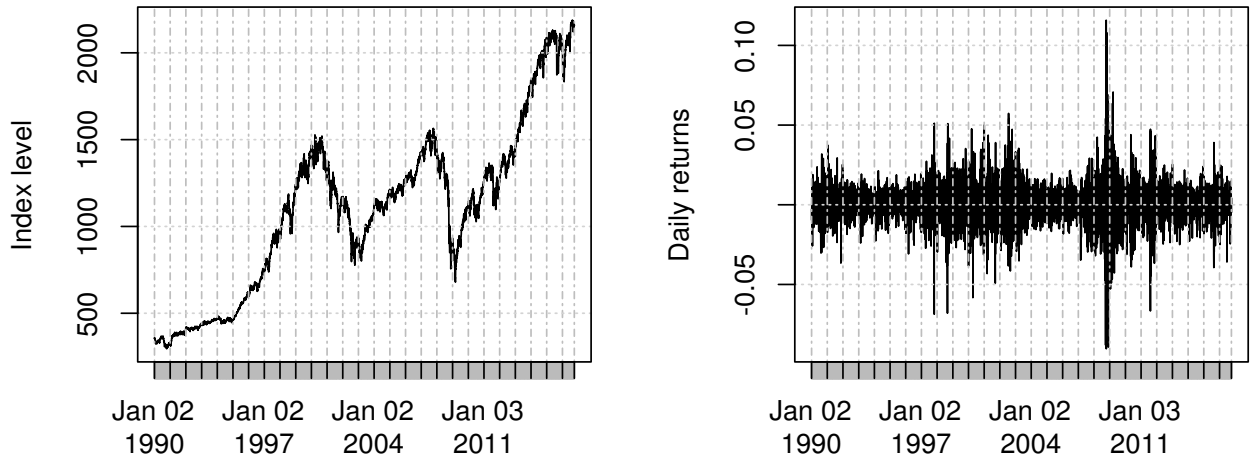
```
# Load package
library(quantmod)

# Download S&P index
getSymbols("^GSPC", from="1990-01-01", to = Sys.Date())
```

```
## [1] "GSPC"
```

```
# Compute returns
GSPC.ret = ClCl(GSPC)

# Plot index level and returns
par(mfrow = c(1,2))
plot(GSPC, main = " ", ylab = "Index level")
```

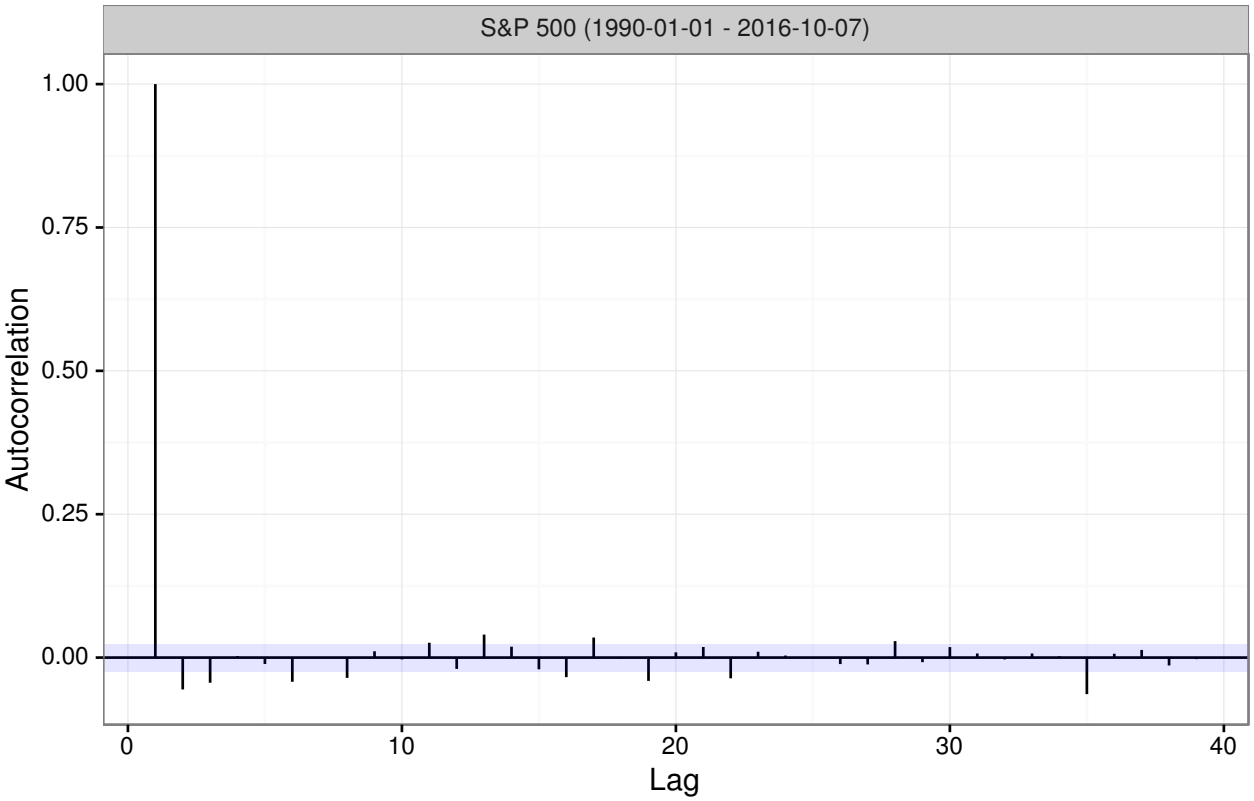
```
## Warning in plot.xts(GSPC, main = " ", ylab = "Index level"): only
## the univariate series will be plotted
```

```
plot(GSPC.ret, main = " ", ylab = "Daily returns")
```




From these graphs it is clear that the returns are not identically distributed as the variance seems to vary with time and clusters with either high or low volatility can be observed. These characteristic of financial time series is well known and in the Chapter 5, we will discuss how the variance of such process can be approximated. Nevertheless, we compute the empirical autocorrelation function of the S&P 500 return to evaluate the degree of "linear" dependence between observation. The graph below presents the empirical autocorrelation.

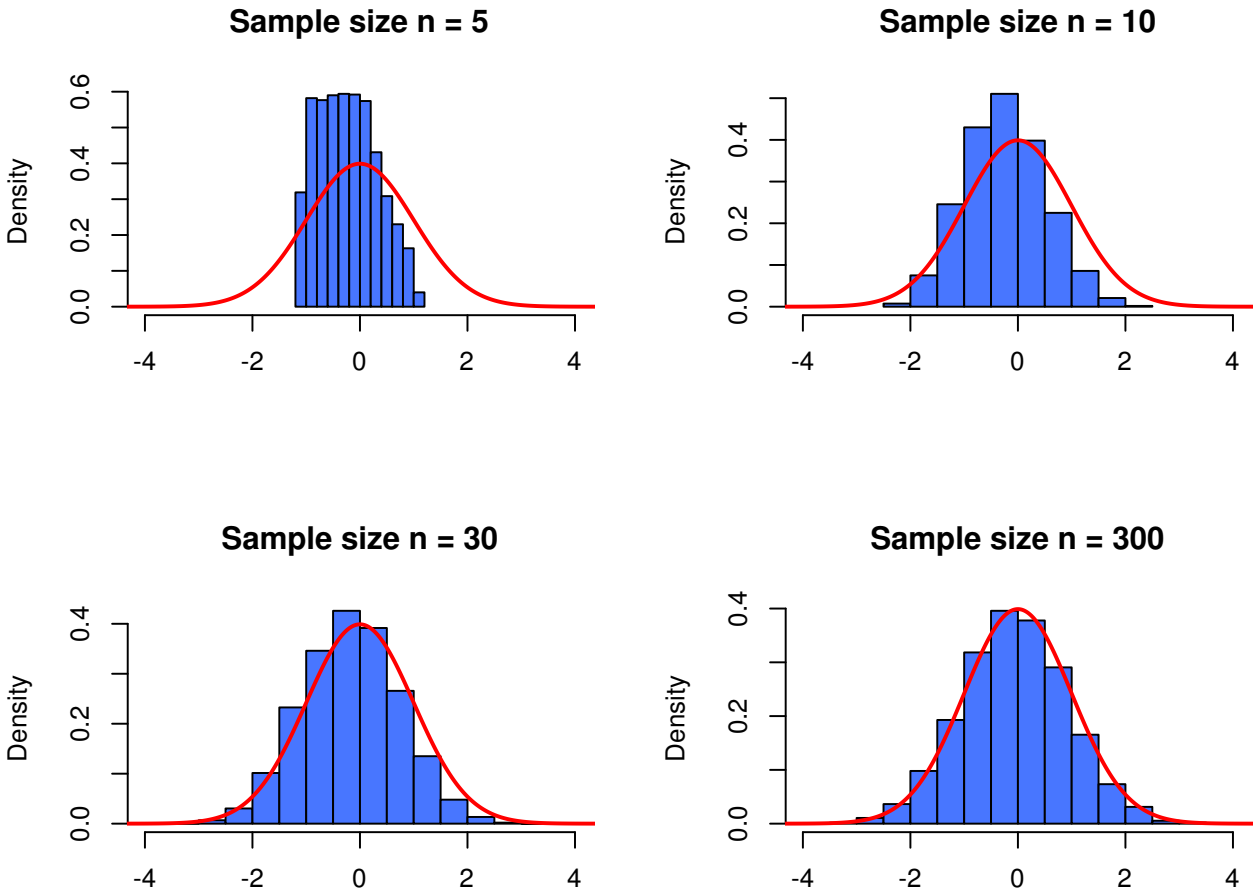
```
sp500 = na.omit(GSPC.ret)
names(sp500) = paste("S&P 500 (1990-01-01 - ",Sys.Date(),")", sep = "")
plot(ACF(sp500))
```



As expected, the autocorrelation is small but it might be reasonable to believe that this sequence is not purely uncorrelated.

Unfortunately, Theorem 1 is based on asymptotic argument and therefore the confidence bands constructed are also asymptotic and there are no "exact" tools that can be used in this case. To study the validity of this results when $n$ is "small" we performed a simulation. In the latter, we simulated processes following from a Gaussian white noise and examine the empirical distribution of $\hat{\rho}(3)$ with different sample sizes (i.e. $n$ is set to 5, 10, 30 and 300). Intuitively, the "quality" of of the approximation provided by Theorem should increase with the sample size $n$. The code below perform such simulation and compares the empirical distribution of $\sqrt{n}\hat{\rho}(3)$ with a normal distribution with mean 0 and variance 1, i.e. its asymptotic distribution, which is depicted using a red line.

```
# Number of Monte Carlo replications
B = 10000

# Define considered lag
h = 3

# Sample size considered
N = c(5,10,30,300)

# Initialisation
result = matrix(NA,B,length(N))

# Set seed
set.seed(1)

# Start Monte Carlo
for (i in seq_len(B)){
  for (j in seq_along(N)){
    # Simluate process
    Xt = rnorm(N[j])

    # Save autocorrelation at lag h
    result[i,j] = acf(Xt, plot = FALSE)$acf[h+1]
  }
}

# Plot results
par(mfrow = c(2,length(N)/2))
for (i in seq_along(N)){
  # Estimated empirical distribution
  hist(sqrt(N[i])*result[,i], col = "royalblue1",
       main = paste("Sample size n =",N[i]), probability = TRUE,
       xlim = c(-4,4), xlab = " ")

  # Asymptotic distribution
  xx = seq(from = -10, to = 10, length.out = 10^3)
  yy = dnorm(xx,0,1)
  lines(xx,yy, col = "red", lwd = 2)
}
```




As expected, it can clearly be observed that the asymptotic approximation is quite poor when $n = 5$ but as the sample size increases the approximation improves and is very close when, for example, $n = 300$. This simulation could suggest that Theorem 1 provides a relatively "close" approximation of the distribution of $\hat{\rho}(h)$.

### 2.3.3 Robustness Issues

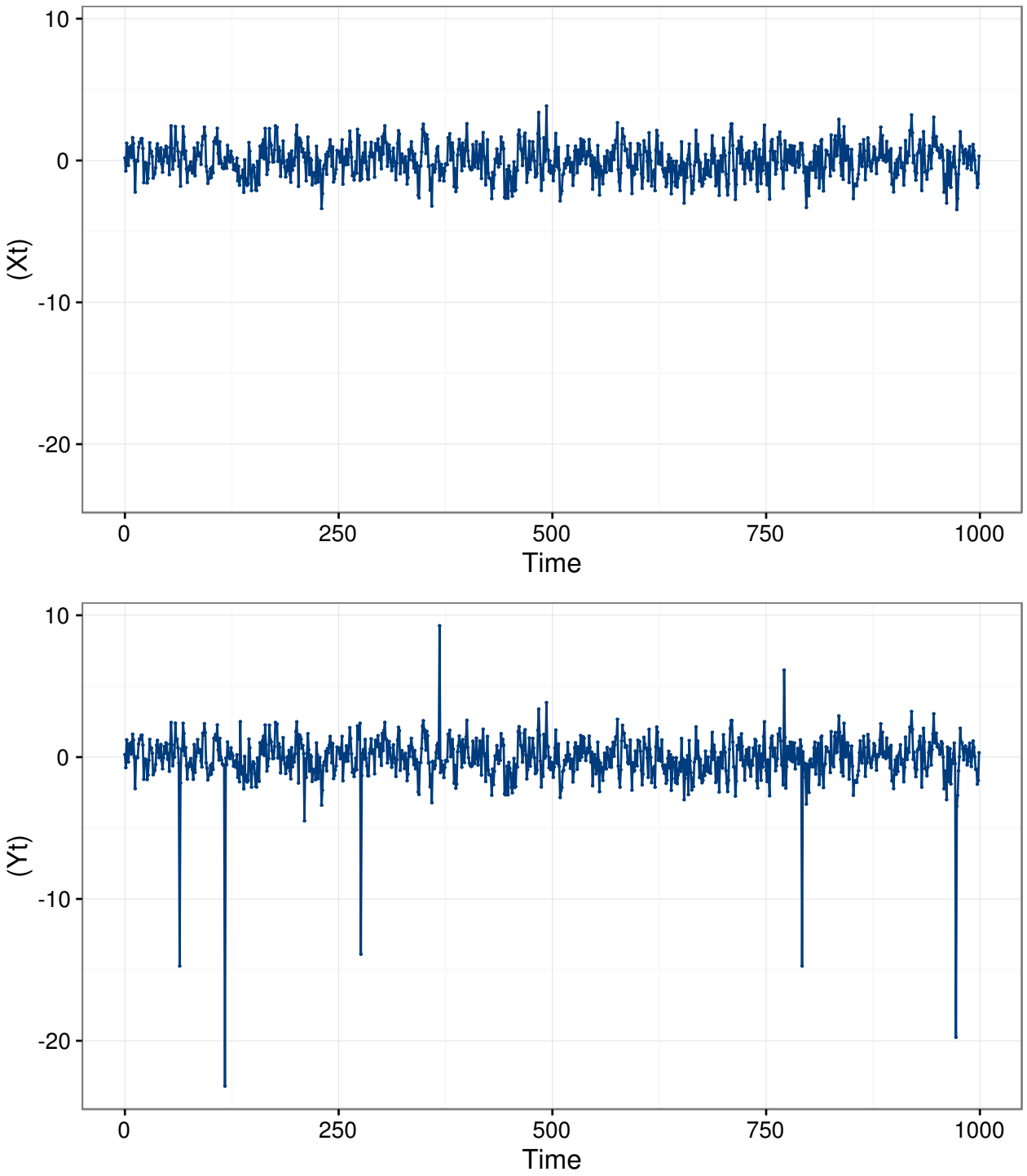
The data generating process delivers a theoretical autocorrelation (autocovariance) function that, as explained in the previous section, can then be estimated through the sample autocorrelation (autocovariance) functions. However, in practice, the sample is often issued from a data generating process that is "close" to the true one, meaning that the sample suffers from some form of small contamination. This contamination is typically represented by a small amount of extreme observations that are called "outliers" that come from a process that is different from the true data generating process.

The fact that the sample can suffer from outliers implies that the standard estimation of the autocorrelation (autocovariance) functions through the sample functions could be highly biased. The standard estimators presented in the previous section are therefore not "robust" and can behave badly when the sample suffers from contamination. To illustrate this limitation of classical estimator we consider the following two processes:

$$X_t = \phi X_{t-1} + W_t, \quad W_t \sim \mathcal{N}(0, \sigma_w^2),$$
$$Y_t = \begin{cases} X_t & \text{with probability } 1-\varepsilon \\ U_t & \text{with probability } \varepsilon \end{cases}, \quad U_t \sim \mathcal{N}(0, \sigma_u^2),$$

when $\varepsilon$ is "small" and $\sigma_u^2 \gg \sigma_w^2$, the process $(Y_t)$ can be interpreted as a "contaminated" version of $(X_t)$. The figure below represents one relalization of the processes $(X_t)$ and $(Y_t)$ using the following setting: $n = 100$, $\sigma_u^2 = 10$, $\phi = 0,5$, $\sigma_w^2 = 1$ as well as $\alpha = 0.05$.

```
library(gmwm)
library(gridExtra)

# Simulate Xt
set.seed(1)
model = AR1(phi = 0.5, sigma2 = 1)
Xt = gen.gts(model)

# Construct Yt
epsilon = 0.01
nb_outlier = rbinom(1,length(Xt),epsilon)
Yt = Xt
Yt[sample(1:length(Xt),nb_outlier)] = rnorm(nb_outlier,0,10)

# Add names
Xt = gts(Xt)
Yt = gts(Yt, name = paste("(",expression(Y[t]),")",sep = ""))

# Plot data
a = autoplot(Xt) + ylim(range(Yt)) + ylab("(Xt)")
b = autoplot(Yt) + ylab("(Yt)")
grid.arrange(a, b, nrow = 2)
```

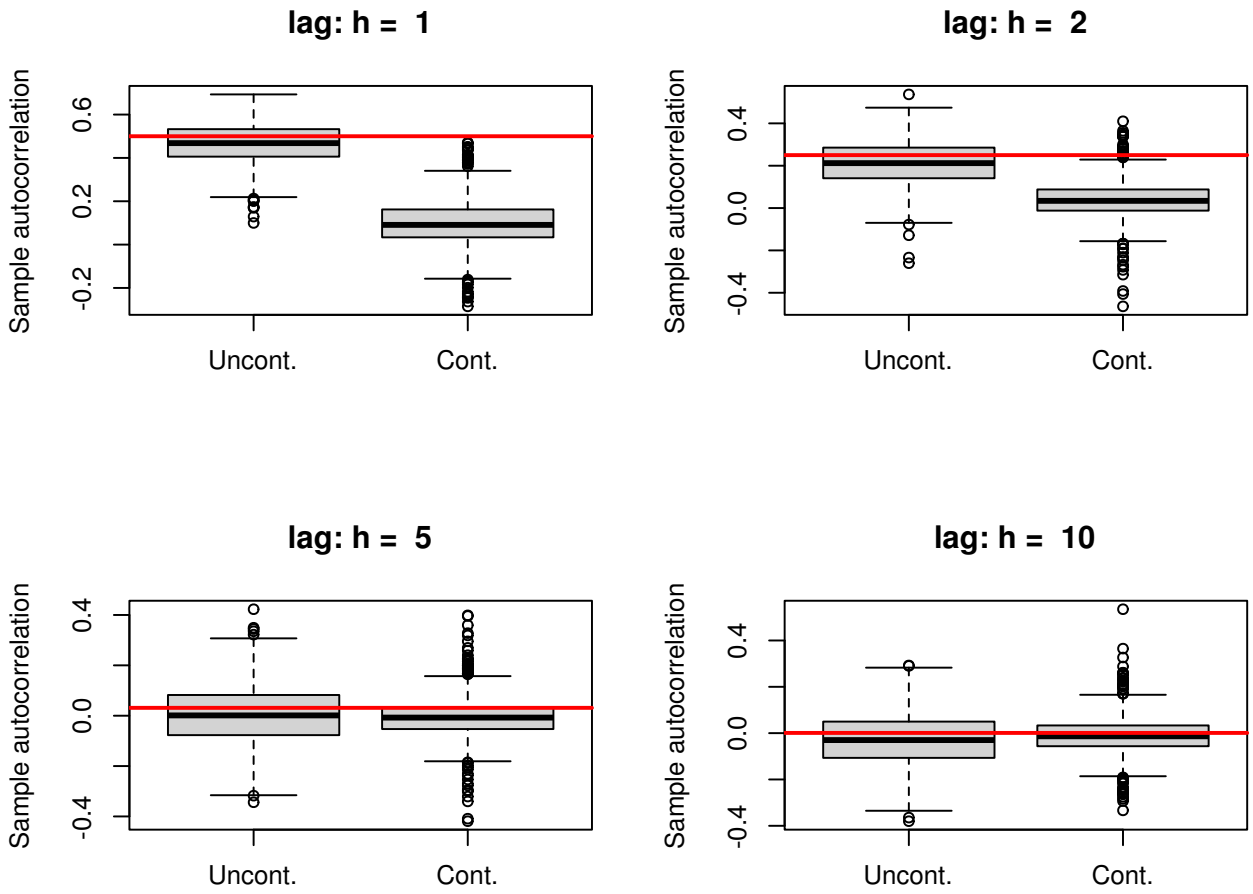



Next, we consider a simulated example to highlight how the performance of the "classical" autocorrelation can deteriorate if the sample is contaminated (i.e. what is the impact of using $(Y_t)$ instead of $(X_t)$, the "uncontaminated" process). In this simulation, we used the setting presented above and consider $B = 10^3$ bootstrap replications.

```
# Define sample size
n = 100
```

```
# Define proportion of "extreme" observation
alpha = 0.05

# Extreme observation are generated from N(0,sigma2.cont)
sigma2.cont = 10

# Number of Monte-Carlo replications
B = 1000

# Define model AR(1)
phi = 0.5
sigma2 = 1
model = AR1(phi = phi, sigma2 = sigma2)

# Initialization of result array
result = array(NA, c(B,2,20))

# Set seed for reproducibility
set.seed(3298)

# Start Monte-Carlo
for (i in seq_len(B)){
  # Simulate AR(1)
  Xt = gen.gts(model, N = n)

  # Create a copy of Xt
  Yt = Xt

  # Add a proportion alpha of extreme observations to Yt
  Yt[sample(1:n,round(alpha*n))] = rnorm(round(alpha*n), 0, sigma2.cont)

  # Compute ACF of Xt and Yt
  acf_Xt = ACF(Xt)
  acf_Yt = ACF(Yt)

  # Store ACFs
  result[i,1,] = acf_Xt[1:20]
  result[i,2,] = acf_Yt[1:20]
}


# Compare empirical distribution of ACF based on Xt and Yt

# Vector of lags considered (h <= 20)
lags = c(1,2,5,10) + 1

# Make graph
par(mfrow = c(2,2))

for (i in seq_along(lags)){
  boxplot(result[,1,lags[i]], result[,2,lags[i]], col = "lightgrey",
          names = c("Uncont.","Cont."), main = paste("lag: h = ", lags[i]-1),
          ylab = "Sample autocorrelation")
```

```
  abline(h = phi^(lags[i]-1), col = 2, lwd = 2)
}
```

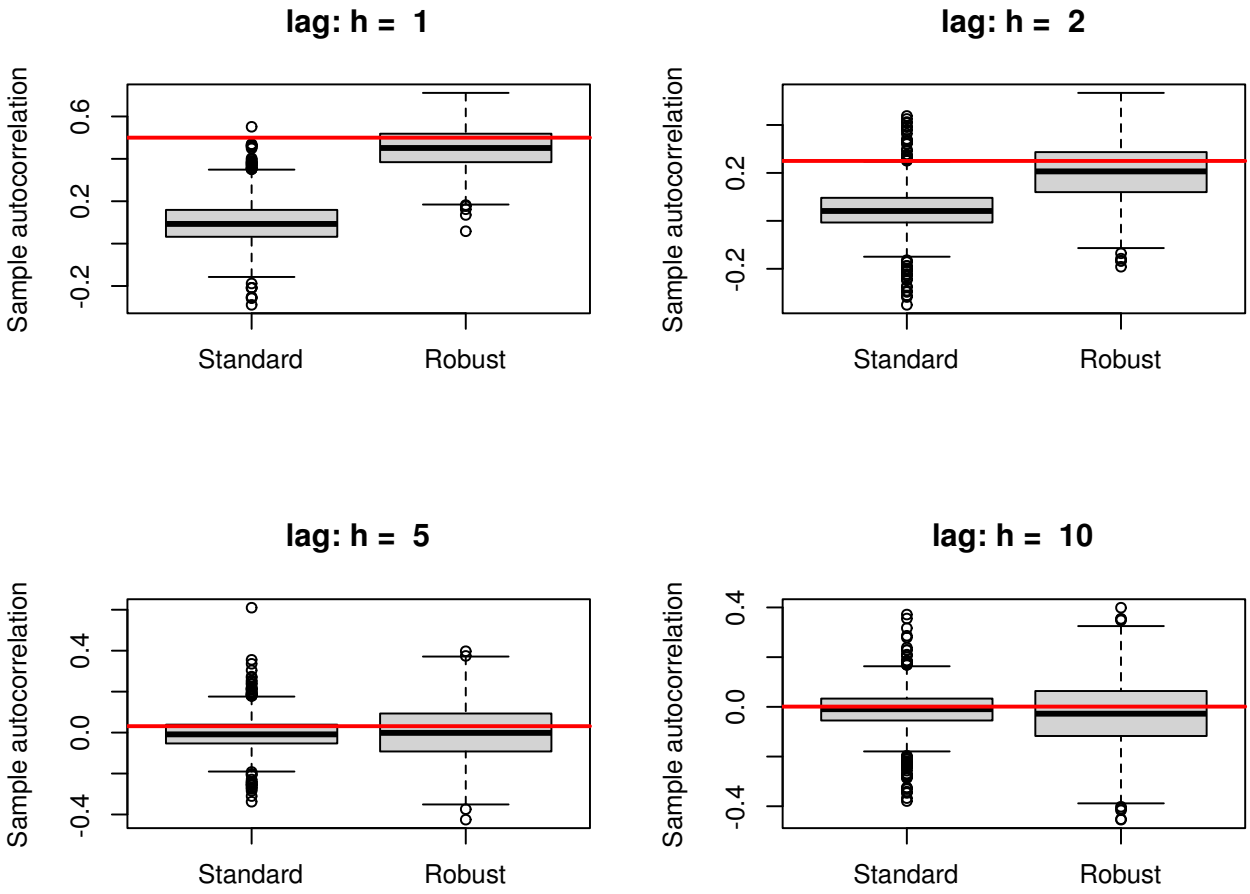



The boxplots in each figure show how the standard autocorrelation estimator is centered around the true value (red line) when the sample is not contaminated (left boxplot) while it is considerably biased when the sample is contaminated (right boxplot), especially at the smallest lags. Indeed, it can be seen how the boxplots under contamination are often close to zero indicating that it does not detect much dependence in the data although it should. This is a known result in robustness, more specifically that outliers in the data can break the dependence structure and make it more difficult for the latter to be detected.

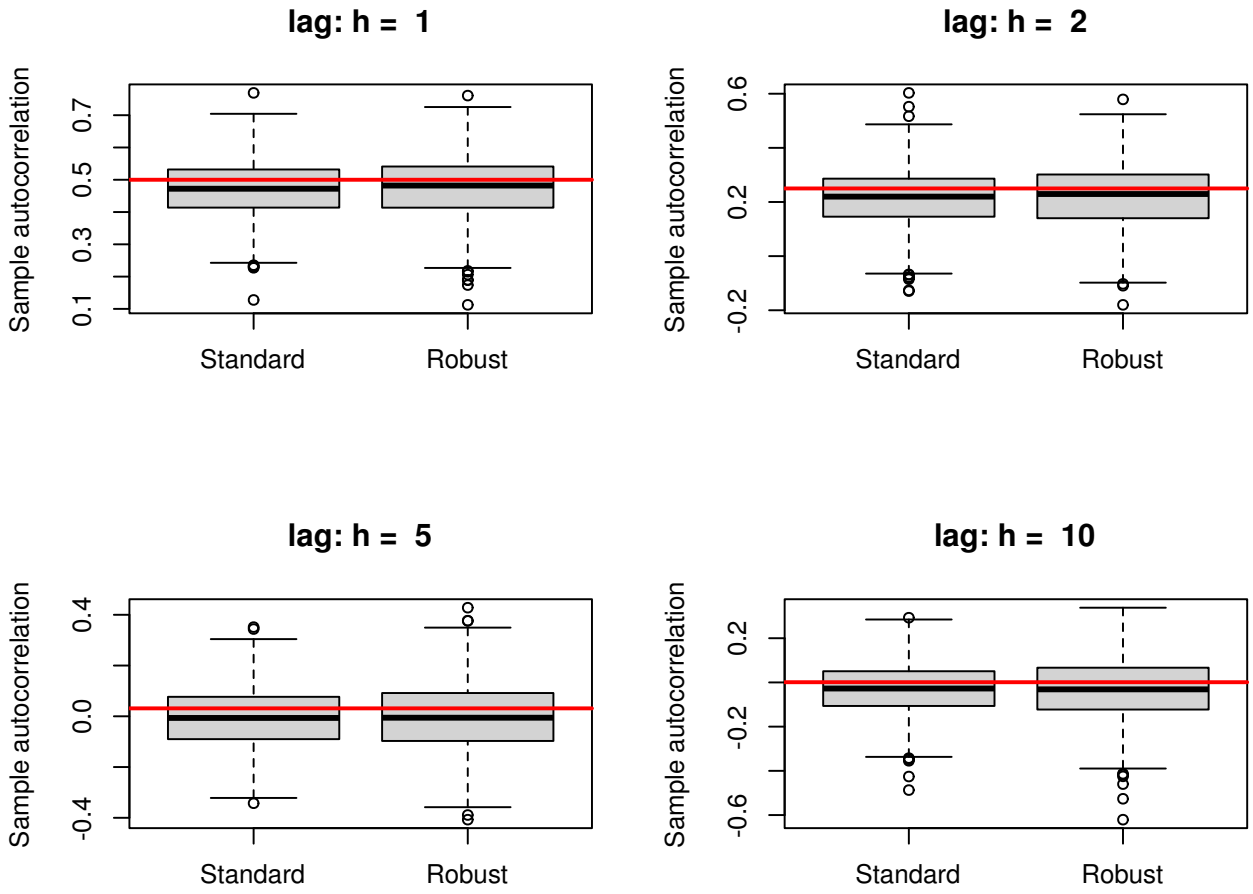
In order to limit this problematic, different robust estimators exist for time series problems allowing to reduce the impact of contamination on the estimation procedure. Among these estimators there are a few that estimate the autocorrelation (autocovariance) functions in a robust manner. One of these estimators is provided in the `robacf()` function in the "robcor" package and the following simulated example shows how it limits the bias due to contamination. Unlike in the previous simulation, we only consider in this example data issued from the contaminated model, i.e. $(Y_t)$, and compare the performance of two estimators (i.e. classical and robust autocorrelation estimators):

```
# Load packages
library("robcor")

# Define sample size
n = 100

# Define proportion of "extreme" observation
alpha = 0.05
```

```r
# Extreme observation are generated from N(0,sigma2.cont)
sigma2.cont = 10

# Number of Monte-Carlo replications
B = 1000

# Define model AR(1)
phi = 0.5
sigma2 = 1
model = AR1(phi = phi, sigma2 = sigma2)

# Initialization of result array
result = array(NA, c(B,2,20))

# Set seed for reproducibility
set.seed(5585)

# Start Monte-Carlo
for (i in seq_len(B)){
  # Simulate AR(1)
  Xt = gen.gts(model, N = n)

  # Add a proportion alpha of extreme observations to Yt
  Xt[sample(1:n,round(alpha*n))] = rnorm(round(alpha*n), 0, sigma2.cont)

  # Compute standard and robust ACF of Xt and Yt
  acf = ACF(Xt)
  rob_acf = robacf(Xt, plot=FALSE)$acf

  # Store ACFs
  result[i,1,] = acf[1:20]
  result[i,2,] = rob_acf[1:20]
}


# Compare empirical distribution of standard and robust ACF based on Xt

# Vector of lags considered (h <= 20)
lags = c(1,2,5,10) + 1

# Make graph
par(mfrow = c(2,2))

for (i in seq_along(lags)){
  boxplot(result[,1,lags[i]], result[,2,lags[i]], col = "lightgrey",
          names = c("Standard","Robust"), main = paste("lag: h = ", lags[i]-1),
          ylab = "Sample autocorrelation")
  abline(h = phi^(lags[i]-1), col = 2, lwd = 2)
}
```



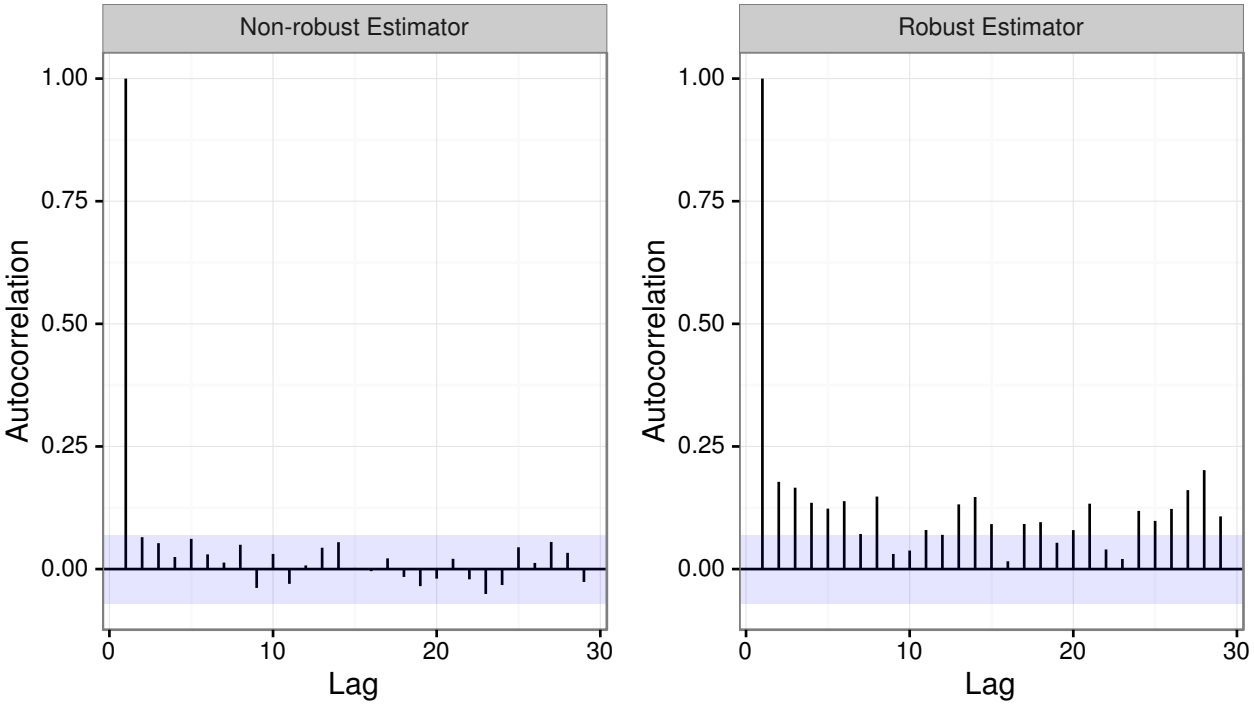

The robust estimator remains close to the true value represented by the red line in the boxplots as opposed to the standard estimator. It can also be observed that to reduce the bias induced by contamination in the sample, robust estimators pay a certain price in terms of efficiency as highlighted by the boxplots that show more variability compared to those of the standard estimator. To assess how much is "lost" by the robust estimator compared to the classical one in terms of efficiency, we consider one last simulation where we examine the performance of two estimators on data issued from the uncontaminated model, i.e. $(X_t)$. Therefore, the only difference between this simulation and the previous one is the value of $\alpha$ set to 0, the code shall thus be omitted and the results are depicted below:




It can be observed that both estimators provide extremely similar results although the robust estimator is slightly more variable.

Next, we consider the issue of robustness on the real data set coming from the domain of hydrology presented in Section 1.2. This data concerns monthly precipitation (in mm) over a certain period of time (1907 to 1972). Let us compare the standard and robust estimators of the autocorrelation functions:

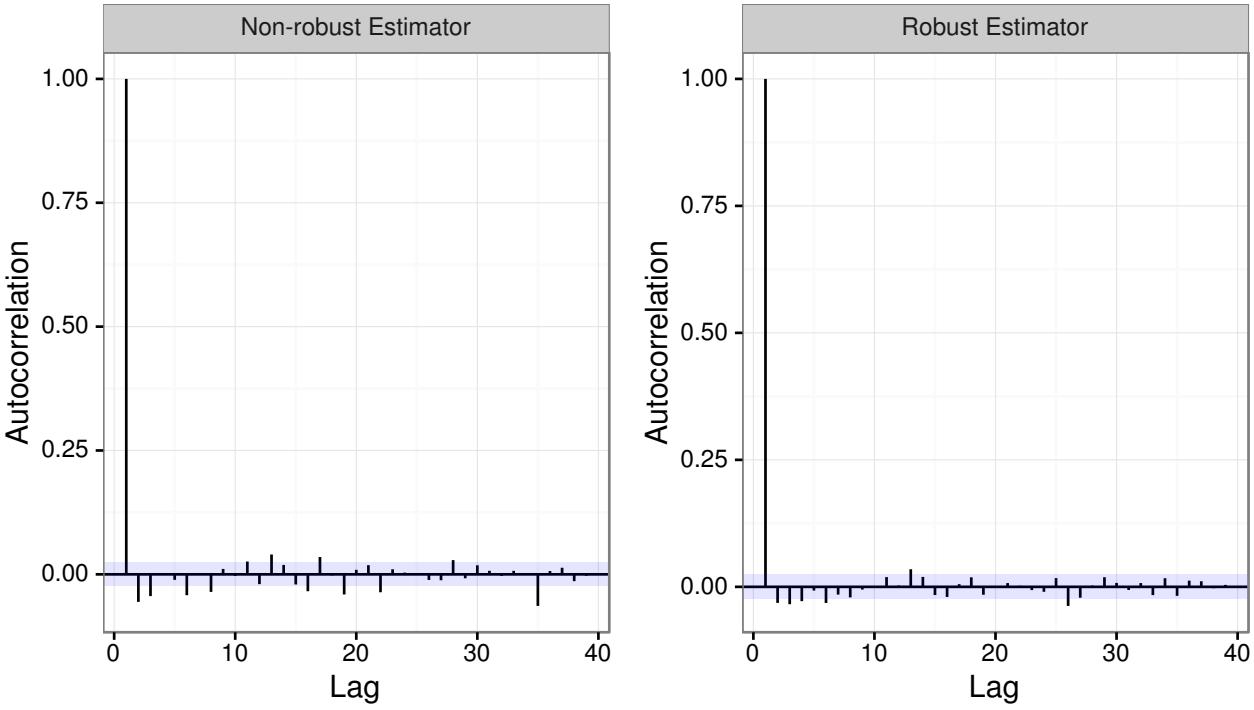
```
# Load packages
library(gmwm)
library(gridExtra)
library(robcor)

# Load data
data("hydro", package = "smacdata")

# Construct gts objects
hydro1 = gts(hydro, name = 'Non-robust Estimator')
hydro2 = gts(hydro, name = 'Robust Estimator')

# Plot data
a = plot(ACF(hydro1))
inter = ACF(hydro2)
inter[,,] = robacf(hydro2, plot=FALSE)$acf
b = plot(inter)
grid.arrange(a, b, nrow = 1)
```




It can be seen that, under certain assumptions (e.g. linear dependence), the standard estimator does not detect any significant autocorrelation between lags since the estimations all lie within the asymptotic confidence intervals. However, many of the robust estimations lie outside these confidence intervals at different lags indicating that there could be dependence within the data. If one were only to rely on the standard estimator in this case, there may be erroneous conclusions drawn on this data. Robustness issues therefore need to be considered for any time series analysis, not only when estimating the autocorrelation (autocovariance) functions.

Finally, we return to S&P 500 returns and compare the classical and robust autocorrelation estimators, which are presented in the figure below.

```
# Construct gts objects
sp500c = gts(sp500, name = 'Non-robust Estimator')
sp500r = gts(sp500, name = 'Robust Estimator')

# Plot data
a = plot(ACF(sp500c))
inter = ACF(sp500r)
inter[,,] = robacf(sp500r, plot=FALSE)$acf
b = plot(inter)
grid.arrange(a, b, nrow = 1)
```




It can be observed that both estimators are very similar. Nevertheless, some small discrepancies can be observed, in particular, the robust estimators seems to indicate an absence of linear dependence while a slightly different interpretation might be achieved with the classical estimator.

## 2.4 Joint Stationarity

The notion of joint stationarity implies that the time series under investigation is bivariate in nature, e.g. $(X_t)$ and $(Y_t)$ are two distinct time series with matching time points. In order to fulfill bivariate stationarity, both processes must be considered to be weakly stationary (constant mean and autocovariance depends on lag $h$) as well as a cross-covariance function depend only on lag $h$. The ideas discussed next can extended beyond the bivariate case to a general multivariate setting.

**Definition 4.** *The **cross-covariance** function of two jointly stationary processes $\{X_t\}$ and $\{Y_t\}$ is given by:*

$$\gamma_{XY}(t+h,t) = \operatorname{cov}(X_{t+h}, Y_t) = \mathbb{E}[(X_{t+h} - \mu_X)(Y_t - \mu_Y)] = \gamma_{XY}(h)$$

Unlike the symmetry found in the autocovariance function of a stationary process $\{X_t\}$, e.g. $\gamma_X(h) = \gamma_X(-h)$, the cross-covariance function is only equal when $\gamma_{XY}(h) = \gamma_{YX}(-h)$. Notice the switch in indices and negative lag change.

**Definition 5.** *The cross-correlation function for two jointly stationary time series $\{X_t\}$ and $\{Y_t\}$ can be expressed as:*

$$\rho_{XY}(t+h,t) = \frac{\gamma_{XY}(t+h,t)}{\sqrt{\gamma_X(0)}\sqrt{\gamma_Y(0)}} = \frac{\gamma_{XY}(h)}{\sigma_{X_{t+h}}\sigma_{Y_t}} = \rho_{XY}(h)$$

Due to the previously discussed symmetry regarding the cross-covariance function, note that the cross-correlation function also only has equality when $\rho_{XY}(h) = \rho_{YX}(-h)$. Thus, note that $\rho_{XY}(h) \neq \rho_{XY}(-h)$.

**Example 15.** Consider two time series processes given by:

$$X_t = W_t - W_{t-1}$$
$$Y_t = W_t + W_{t-1}$$

where $W_t \sim WN(0, \sigma_W^2)$.

First check to see if $\{X_t\}$ and $\{Y_t\}$ are weakly stationary.

The means of both time series are evident to be

$$E[X_t] = E[Y_t] = 0$$

The autocovariance for $\{X_t\}$ is given as:

$$\begin{aligned}
\gamma_X(h) &= \operatorname{cov}(X_{t+h}, X_t) \\
&= \operatorname{cov}(W_{t+h} - W_{t+h-1}, W_t - W_{t-1}) \\
&= 1_{\{h=0\}} 2\sigma^2 + 1_{\{h=-1\}} - \sigma^2 + 1_{\{h=1\}} - \sigma^2 \\
&= \begin{cases} 2\sigma_W^2, & \text{if} h = 0 \\ -\sigma_W^2, & \text{if} |h| = 1 \\ 0, & \text{if} |h| \geq 1 \end{cases}
\end{aligned}$$

Similarly, the autocovariance for $\{Y_t\}$ is calculated to be:

$$\begin{aligned}
\gamma_Y(h) &= \operatorname{cov}(Y_{t+h}, Y_t) \\
&= \operatorname{cov}(W_{t+h} + W_{t+h-1}, W_t + W_{t-1}) \\
&= 1_{\{h=0\}} 2\sigma^2 + 1_{\{h=-1\}} \sigma^2 + 1_{\{h=1\}} \sigma^2 \\
&= \begin{cases} 2\sigma_W^2, & \text{if} h = 0 \\ \sigma_W^2, & \text{if} |h| = 1 \\ 0, & \text{if} |h| \geq 1 \end{cases}
\end{aligned}$$

Next, the cross-covariance for $\{X_t\}$ and $\{Y_t\}$:

$$\begin{aligned}
\gamma_{XY}(h) &= \operatorname{cov}(X_{t+h}, Y_t) \\
&= \operatorname{cov}(W_{t+h} - W_{t+h-1}, W_t + W_{t-1}) \\
&= \begin{cases} 0, & \text{if } h = 0 \\ -\sigma_W^2, & \text{if } h = 1 \\ \sigma_W^2, & \text{if } h = -1 \\ 0, & \text{if } h \geq 2 \end{cases}
\end{aligned}$$

Therefore, based on obtain the weak stationarity for each process and obtain a cross-covariance function that only depends on the lag $h$, the process is joint stationary.

Furthermore, we also have the cross-correlation function as:

$$\rho_{XY}(X_{t+h}, Y_t) = \frac{\gamma_{XY}(h)}{\sigma_X \sigma_Y} = \frac{\gamma_{XY}(h)}{\sqrt{\gamma_x(0)}\sqrt{\gamma_y(0)}} = \frac{\gamma_{XY}(h)}{\gamma_X(0)} = \begin{cases} 0, & \text{if } h = 0 \\ -\frac{1}{2}, & \text{if } h = 1 \\ \frac{1}{2}, & \text{if } h = -1 \\ 0, & \text{if } h \geq 2 \end{cases}$$

### 2.4.1 Sample Cross-Covariance and Cross-Correlation Functions

A natural estimator of the *cross-covariance function* is given by:

$$\hat{\gamma}_{XY}(h) = \frac{1}{T}\sum_{t=1}^{T-h}\left(X_{t+h} - \bar{X}\right)\left(Y_t - \bar{Y}\right)$$

With this in mind, the "plug-in" estimator for the *cross-correlation function* follows:

$$\hat{\rho}_{XY}(h) = \frac{\hat{\gamma}_{XY}(h)}{\sqrt{\hat{\gamma}_X(0)}\sqrt{\hat{\gamma}_Y(0)}}$$

Both of the above estimators are again only symmetric under the above index and lag transformation.

## 2.5 Portmanteau test

In this section we give a brief introduction to Portmanteau tests used in time series analysis. In linear regression, we always need to do diagnostic test for residuals after building our model, to check whether our assumptions are satisfied. If there is no evidence to reject any of the assumptions, we can say that the linear model we built is adequate. Otherwise, the linear models are not adequate, some modifications or transformations need to be done either for the previous model or for the data. This rule also applies to time series modeling. In time series analysis, a wide variety of Portmanteau tests can be used to check the white noise residuals assumption. We will introduce two of them as follows, which are based on the ACF of residuals, in order to illustrate some of ideas of this kind of tests.

Dating back to 1970, Box and Pierce proposed the well-known Box-Pierce test statistic as the following form:

$$Q_{BP} = n\sum_{h=1}^{m}\hat{\rho}_h^2,$$

where the empirical autocorrelation of residuals at lag $h$ is defined as $\hat{\rho}_h = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}$. It is obvious that under alternative hypothesis, $\hat{\rho}_h$ would be deviate from 0, thus a large $Q_{BP}$ gives us the evidence to reject the null. And under null hypothesis that the residuals are white noise (or equivalently the time series are adequate), it can be shown that when $n \to \infty$, we have

$$Q_{BP} \xrightarrow{\mathcal{D}} \chi^2_{m-m'},$$

where $m'$ is the number of parameters in the time series model.

Then on 1978, Ljung and Box improved Box-Pierce test by standardizing each $\hat{\rho}_h$ by its asymptotic variance. The Ljung and Box test statistic is

$$Q_{LB} = n\sum_{h=1}^{m}\frac{n+2}{n-h}\hat{\rho}_h^2.$$

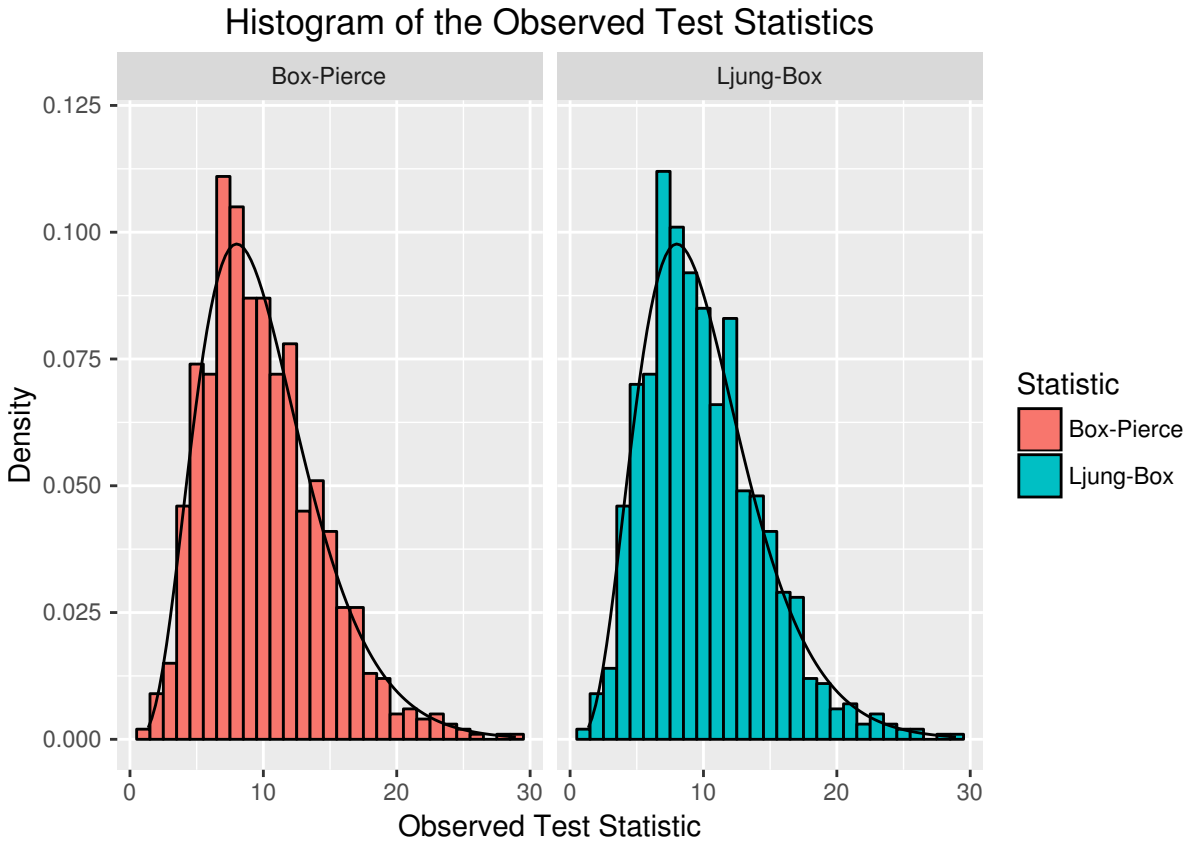
It can also be shown that $Q_{LB} \xrightarrow{\mathcal{D}} \chi^2_{m-m'}$ under the null. However, compared to $Q_{BP}$, the distribution of $Q_{BP}$ under the null is closer to $\chi^2_{m-m'}$, when $n$ is finite.

In the above two examples, the test statistic contains a user specified parameter $m$. And for different $m$, the power of the test would be different. Thus many work has been done to either select the optimal $m$, or propose a new test statistic without user specified parameters. Moreover, testing white noise can also be

done by checking PACF, or by checking the spectral density in the frequency domain. Therefore these lead to many different Portmanteau tests.

Take for an example the following use of a Portmanteau test to show the distribution of test statistics under the null:

```
library(gmwm)

# set seed
set.seed(1345)

# Specify models
model = WN(sigma2 = 1) # WN

B = 1000 # number of parametric bootstrap
BP.obs = rep(NA, B)
LB.obs = rep(NA, B)

for (j in seq_len(B)){
  x = gen.gts(model, N = 1000)
  BP.obs[j] = Box.test(x, lag = 10, "Box-Pierce", fitdf = 0)$statistic
  LB.obs[j] = Box.test(x, lag = 10, "Ljung-Box", fitdf = 0)$statistic
}

sim_results = data.frame(sim = c(BP.obs, LB.obs),
                         simtype = c(rep("Box-Pierce",B), rep("Ljung-Box",B)))

ggplot(data = sim_results, aes(x = sim)) +
  geom_histogram(aes(y = ..density.., fill = simtype),
                 binwidth = 1, color = "black") +
  stat_function(fun = dchisq, args = list(df = 10)) +
  facet_wrap( ~ simtype) + ylim(0, 0.12) +
  labs(fill = "Statistic", title = "Histogram of the Observed Test Statistics",
       y = "Density", x = "Observed Test Statistic")
```

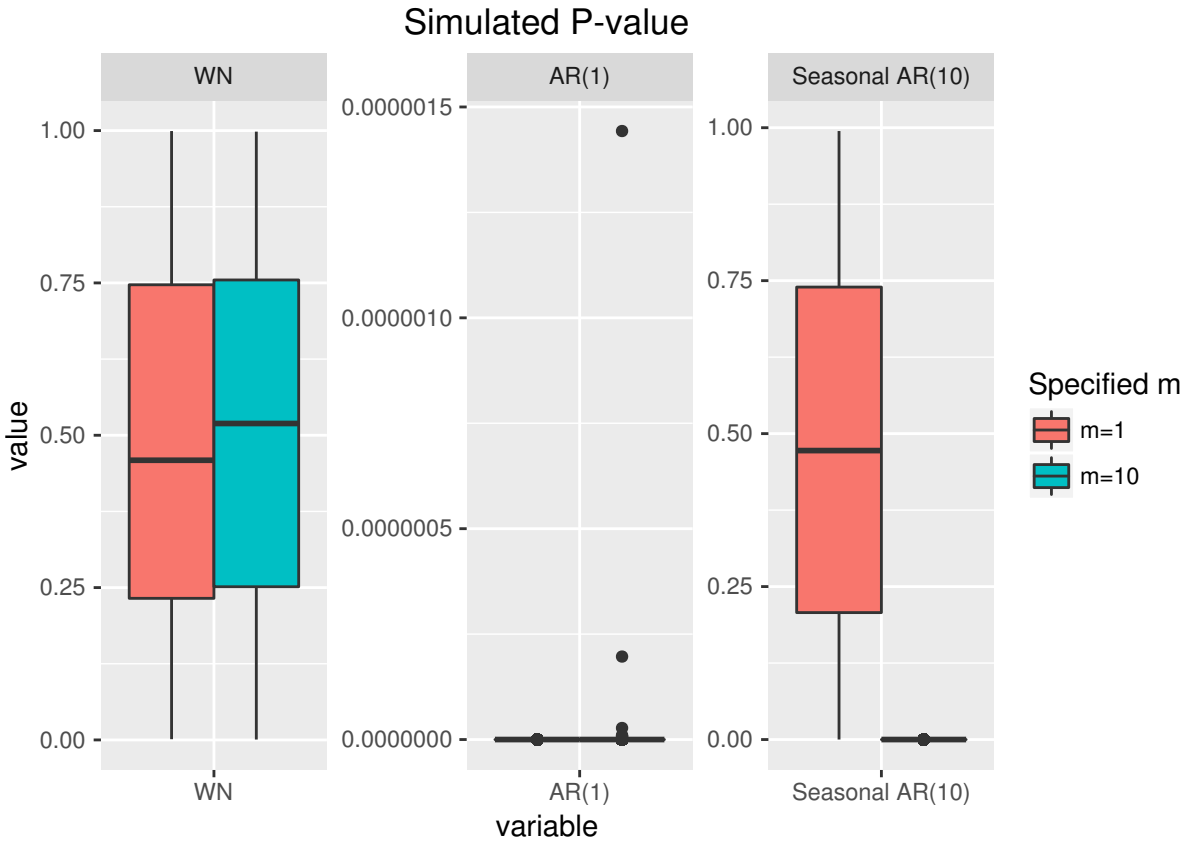



From the histogram, we can see that under the null the distribution of both BP and LB are close to Chi-square distribution, but LP is slightly better.

To show the distribution of P-values under different alternatives and show that the test depends on specified $m$.

```
# set seed
set.seed(1234)

# Specify models
model1 = WN(sigma2 = 1)                          # WN
model2 = AR(phi = 0.3, sigma2 = 1)               # AR(1)
model3 = AR(phi = c(rep(0,9), 0.3), sigma2 = 1) # seasonal AR(10)

B = 1000 # number of parametric bootstrap
LB.pvalue = matrix(NA, nrow = B, ncol = 6)

for (i in 1:3){
  for (j in seq_len(B)){
    x = gen.gts(get(paste0("model", i)), N = 1000)
    LB.pvalue[j,2*i-1] = Box.test(x, lag = 1, "Ljung-Box", fitdf = 0)$p.value
    LB.pvalue[j,2*i] = Box.test(x, lag = 10, "Ljung-Box", fitdf = 0)$p.value
  }
}

para_1 = data.frame(lag = 1, LB.pvalue[,c(1,3,5)])
para_2 = data.frame(lag = 2, LB.pvalue[,c(2,4,6)])
para = rbind(para_1, para_2)
```

```
colnames(para)[2:4] = c("WN", "AR(1)", "Seasonal AR(10)")

library("reshape2")
para.melt = melt(para, id.vars = "lag")

ggplot(data = para.melt, aes(x=variable, y=value)) +
  geom_boxplot(aes(fill=factor(lag))) + facet_wrap( ~ variable, scales="free") +
  ggtitle("Simulated P-value") +
  scale_fill_hue(name="Specified m", breaks = c(1,2) , labels = c("m=1", "m=10"))
```



# Chapter 3

# Autoregressive Moving Average Models

"*Essentially, all models are wrong, but some are useful*", George Box

In this chapter we introduce a class of time series models that is flexible and among the most commonly used to describe stationary time series. This class is represented by the AutoRegressive Moving Average (ARMA) models which combine and include the AutoRegressive (AR) and Moving Average (MA) models seen in the previous chapter. We first discuss AR and MA models in further detail before introducing the general ARMA class. The importance of the latter class of models resides in its flexibility as well as its capacity of describing (or closely approximating) the features of stationary time series. The autoregressive parts of these models describe how consecutive observations in time influence each other while the moving average parts capture some possible unobserved shocks thereby allowing to model different phenomena going from biology to finance.

To introduce and explain this class of models, this chapter is organized in the following manner. First of all we will discuss the class of linear processes of which the ARMA models are part of. We will then proceed to a detailed description of AR models in which we review their definition, explain their properties, introduce the main estimation methods for their parameters and highlight the diagnostic tools which can help understand if the estimated models appear to be appropriate or sufficient to well describe the observed time series. Once this is done, we will then use most of the results given for AR models to further describe and discuss MA models, for which we underline the property of invertibility, and finally the ARMA models. Indeed, the properties and estimation methods for the latter class are directly inherited from the discussions on AR and MA models.

## 3.1 Linear Operators and Processes

Within this section, the concepts of linear operators and processes will be discussed. Both concepts are critical to being able to model time series data. In fact, every time series that can be denoted as weakly stationary is either a linear process or has the ability to be after being transformed using a linear operation such as differencing deterministic components.

### 3.1.1 Linear Operators

Prior to introducing linear processes, discussion must first focus upon a very useful tool: linear operators. These operators play a fundamental role in easing the manipulation of time series equations. Principally,

the manipulations serve three fold: 1. converting into a linear process form, 2. differencing deterministic components, and 3. adjusting seasonal time windows.

**Definition 6** (Linear Operator)**.** *A linear operator, $L$, is an operator which satisfies the following two conditions:*

$$\begin{aligned} L(X_t + Y_t) &= LX_t + LX_t && \textit{(Preserves Addition)} \\ L(\lambda \cdot X_t) &= \lambda \cdot (LX_t) && \textit{(Preserves Scalar Multiplication)} \end{aligned}$$

Linear operators are defined to exists within a Hilbert space. As a result, these operations are mathematical very rigorous. Details on the rigor will not be invested indepth within this text and are left as an exercise better suited to Abstract Linear Algebra.

#### 3.1.1.1 Backshift (Lag) Operator

One of the principal uses of a linear operator within time series is the backshift operator. The backshift operator provides the ability to change the indices of the time series by one period e.g. $t \to t-1$.

**Definition 7** (Backshift Operator)**.** *The backshift operator is defined as:*

$$BX_t = X_{t-1}$$

*The backshift operator has the following properties:*

1. *Multiple backshifts yield:* $B^k X_t = B^{k-1} B X_t = B^{k-1} X_{t-1} = X_{t-k}$
2. *By raising to a negative power, the backshift operator moves observations forward:* $B^{-1} = X_{t+1}$
3. *Given a constant,* $c \in \mathbb{R}$*:* $Bc = c$
4. $B^0 = 1$

The notation for the backshift operator changes depending on the author's preferences. An alternative way to denote a backshift operation given by $B$ is to use the lag operator $L$. The same conventions as discussed above apply equally to the lag operator.

**Example 16.** [Backshifting with respect to seasonality] If the backshift operator is raised to a higher power, say $k = 12$, then this leads to allowing the exploration of seasonality. Specifically, in the case of monthly data, $B^{12} X_t = X_{t-12}$ would provide a way to look at the data that occurred one year prior.

#### 3.1.1.2 Differencing Operator

The differencing operator is helpful when trying to remove trend from the data. Many make an analogy between this operator and taking a first derivative of a function.

**Definition 8** (Differencing Operator)**.** *The* ***Differencing Operator*** *is defined as the gradient symbol applied to a time series:*

$$\nabla X_t = X_t - X_{t-1} = (1 - B) X_t$$

*The difference operator has the following properties:*

1. $\nabla^k X_t = \nabla^{k-1} (\nabla X_t) = (1 - B)^k X_t$
2. $\nabla^1 X_t = \nabla X_t$
3. $\nabla c = 0$

*4.* $\nabla^0 X_t = X_t$

**Example 17.** [Repetitive Differencing] To understand the repetitive nature of the difference operation, consider the case when the number of differences is $k = 2$:

$$\begin{aligned}\nabla^2 X_t &= \nabla\left(\nabla X_t\right) \\ &= \nabla\left(X_t - X_{t-1}\right) \\ &= \left(X_t - X_{t-1}\right) - \left(X_{t-1} - X_{t-2}\right) \\ &= X_t - 2X_{t-1} + X_{t-2}\end{aligned}$$

Notice that at the end of the process, we are able to transform the differencing operation to rely upon the backshift operator:

$$\begin{aligned}\nabla^2 X_t = X_t - 2X_{t-1} + X_{t-2} &= X_t - 2BX_t + B^2 X_t \\ &= \left(1 - B\right)\left(1 - B\right) X_t = \left(1 - B\right)^2 X_t\end{aligned}$$

#### 3.1.1.3 Seasonal Differencing Operator

As discussed previously, the backshift operator has a subtle way of shifting one time period. Trends that exists within time series sometimes span months or even years. Therefore, to remove such a deterministic component, the seasonality will have to be differenced out.

**Definition 9** (Seasonal Difference Operator)**.** *The seasonal difference operator can applied by using a seasonal backshift:*

$$\nabla_s X_t = X_t - X_{t-s} = \left(1 - B_s\right) X_t$$

*As this is a modification of the difference operator, the previously discussed properties associated with the differencing operation largely are the same excluding the differentiation. That is, the seasonal difference operator is generalized such that:*

$$\nabla_s^k X_t = \left(1 - B_s\right)^k X_t$$

### 3.1.2 Linear Processes

The models that that have been studied thus far have a common theme that unites them. Underlying each model, there is an input sequence of independent random errors that is mapped to a series of correlated weighted random variables. As only the output $\{X_t\}$ is observable, the type of process and subsequently weighting must be differentiated during the course of modeling.

Recall the Example 12 where *backsubstitution* was used to manipulate an AR(1) model into a condensed series form. From the condensed form, the ability to ascertain whether an $X_t$ existed such that the model had a solution was readily apparent. With the newfound knowledge of the *backshift* operator, the example is ripe to be revisited.

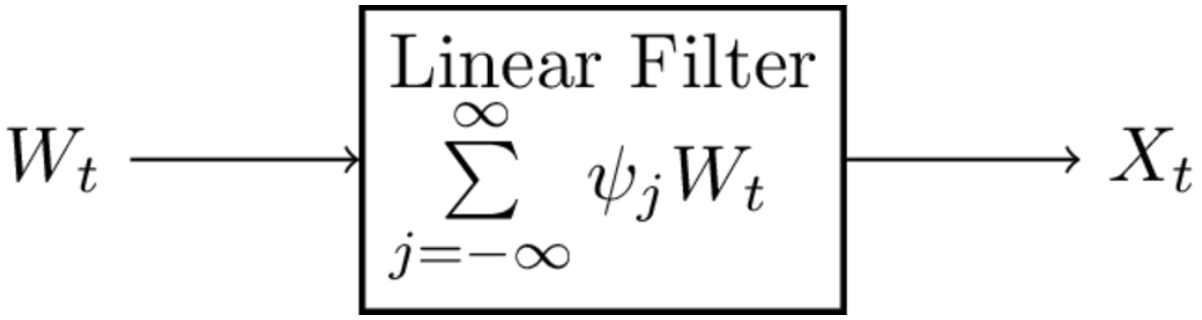



Figure 3.1: Processing a signal

$$\begin{aligned} X_t &= \phi X_{t-1} + W_t \\ X_t - \phi X_{t-1} &= W_t \\ (1 - \phi B) X_t &= W_t \\ X_t &= \frac{1}{1 - \phi B} W_t \\ &= \sum_{j=0}^{\infty} \phi^j B^j W_t = \sum_{j=0}^{\infty} \phi^j W_{t-j} \end{aligned}$$

Thus, just as before, there does exist a solution for an AR(1) when $|\phi| < 1$. However, there is a specific name for a process that is able to be written in the above form.

**Definition 10** (Linear Process)**.** *A time series, $\{X_t\}$, is defined to be a linear process if it can be expressed as a linear combination of white noise by:*

$$X_t = \mu + \sum_{j=-\infty}^{\infty} \psi_j W_{t-j}$$

*where $W_t \sim WN(0, \sigma^2)$ and $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$.*

Note, the later assumption is required to ensure that the series has a limit. Furthermore, the set of coefficients

$$\{\psi_j\}_{j=-\infty,\cdots,\infty}$$

can be viewed as linear filter. These coefficients do not have to be all equal nor symmetric as later examples will show.

**Theorem 2** (Stationarity of Linear Process)**.** *A linear process $\{X_t\}$ in the form of Definition 10 is stationary with:*

$$\begin{aligned} \mathbb{E}[X_t] &= \mu \\ \gamma(h) &= \sigma^2 \sum_{j=-\infty}^{\infty} \psi_j \psi_{h+j}. \end{aligned}$$

*Proof.* The expectation and autocovariance are able to be obtained immediately given the series converges under $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$. Therefore, we have for the expectation:

$$E[X_t] = E\left[\mu + \sum_{j=-\infty}^{\infty} \psi_j W_{t-j}\right] = \mu + \sum_{j=-\infty}^{\infty} \psi_j E[W_{t-j}] = \mu + 0 = \mu$$

In addition, the autocovariance is able to be obtained as:

$$\begin{aligned}
\operatorname{cov}(X_{t+h}, X_t) &= \operatorname{cov}\left(\mu + \sum_{j=-\infty}^{\infty} \psi_j W_{t+h-j}, \mu + \sum_{i=-\infty}^{\infty} \psi_i W_{t-i}\right) \\
&= \sum_{j=-\infty}^{\infty} \sum_{i=-\infty}^{\infty} \psi_j \psi_i \operatorname{cov}(W_{t+h-j}, W_{t-i}) = \sum_{j=-\infty}^{\infty} \psi_{j+h} \psi_j \operatorname{cov}(W_t, W_t) + \sum_{j=-\infty}^{\infty} \sum_{i=-\infty}^{\infty} \psi_j \psi_i \underbrace{\operatorname{cov}(W_{t+h-j}, W_{t-i})}_{=0} \\
&= \sigma_W^2 \sum_{j=-\infty}^{\infty} \psi_j \psi_i
\end{aligned}$$

Within the above derivation, the key is to realize that $\operatorname{cov}(W_{t+h-j}, W_{t-i}) = 0$ since $i \neq j + h$. □

**Example 18.** [Linear Processes under backshift operator] Linear processes are able to written in a more compact form by utilizing the backshift operator.

$$\begin{aligned}
X_t &= \mu + \psi(B) W_t \\
\psi(B) &= \sum_{j=0}^{\infty} \psi_j B^j
\end{aligned}$$

### 3.1.3 Examples of Linear Processes

Next, let us consider how to represent the previously discussed time series in Section 1.3 as a linear process. Generally speaking, the work required to transform the time series representation into a linear process rests solely upon the ability to find a set of $\psi_j$.

**Example 19.** [Linear Process of White Noise] The white noise process $\{X_t\}$, defined in 1.3.1, can be expressed as a linear process under:

$$\psi_j = \begin{cases} 1, & \text{if } j = 0 \\ 0, & \text{if } |j| \geq 1, \end{cases}$$

and $\mu = 0$.

Therefore, $X_t = W_t$, where $W_t \sim WN(0, \sigma_W^2)$.

**Example 20.** [Linear Process of Moving Average Order 1] Similarly, consider $\{X_t\}$ to be a MA(1) process, given by 1.3.4. The process can be expressed linearly under:

$$\psi_j = \begin{cases} 1, & \text{if } j = 0 \\ \theta, & \text{if } j = 1 \\ 0, & \text{if } j \geq 2, \end{cases}$$

and $\mu = 0$.

Thus, we have: $X_t = W_t + \theta W_{t-1}$.

**Example 21.** [Linear Process and Symmetric Moving Average] Consider a symmetric moving average given by:

$$X_t = \frac{1}{2q+1} \sum_{j=-q}^{q} W_{t+j}$$

Thus, $\{X_t\}$ is defined for $q+1 \leq t \leq n-q$. The above process would be a linear process since:

$$\psi_j = \begin{cases} \frac{1}{2q+1}, & \text{if } -q \leq j \leq q \\ 0, & \text{if } |j| > q. \end{cases}$$

and $\mu = 0$.

In practice, if $q = 1$, we would have:

$$X_t = \frac{1}{3}\left(W_{t-1} + W_t + W_{t+1}\right).$$

**Example 22.** [Linear Process of Autoregressive Process of Order 1] A more intensive application of a linear process is $\{X_t\}$ as an AR1 process, defined in 1.3.3. The intensity comes from the necessity to define the weights with respect to the time lag.

$$\psi_j = \begin{cases} \phi^j, & \text{if } j \geq 0 \\ 0, & \text{if } j < 0, \end{cases}$$

and $\mu = 0$.

Under the condition that $|\phi| < 1$ the process can be considered to be the traditional $X_t = \phi X_{t-1} + W_t$.

## 3.2 Autoregressive Models

The class of autoregressive models is based on the idea that previous values in the time series are needed to explain current values in the series. For this class of models, we assume that the $p$ previous observations are needed for this purpose and we therefore denote this class as AR($p$). In the previous chapter, the model we introduced was an AR(1) in which only the immediately previous observation is needed to explain the following one and therefore represents a particular model which is part of the more general class of AR(p) models.

**Definition 11** (Autoregressive Models of Order $p$)**.** *The AR(p) models can be formally represented as follows*

$$X_t = \phi_1 X_{t-1} + ... + \phi_p X_{t-p} + W_t,$$

*where $\phi_p \neq 0$ and $W_t$ is a (Gaussian) white noise process with variance $\sigma^2$.*

In general, we will assume that the expectation of the process $(X_t)$, as well as that of the following ones in this chapter, is zero. The reason for this simplification is that if $\mathbb{E}[X_t] = \mu$, we can define an AR process *around* $\mu$ as follows:

$$X_t - \mu = \sum_{i=1}^{p} (\phi_i X_{t-i} - \mu) + W_t,$$

which is equivalent to

$$X_t = \mu^\star + \sum_{i=1}^{p} \phi_i X_{t-i} + W_t,$$

where $\mu^\star = \mu(1 - \sum_{i=1}^{p} \phi_i)$. Therefore, to simplify the notation we will generally consider only zero mean processes, since adding means (as well as other deterministic trends) is easy.

A useful way of representing AR processes is through the backshift operator introduced in the previous section and is as follows

$$\begin{aligned} X_t &= \phi_1 X_{t-1} + ... + \phi_p y_{t-p} + w_t \\ &= \phi_1 B X_t + ... + \phi_p B^p X_t + W_t, \\ &= (\phi_1 B + ... + \phi_p B^p) X_t + W_t \end{aligned}$$

which finally yields

$$(1 - \phi_1 B - ... - \phi_p B^p) X_t = W_t,$$

or, in abbreviated form, can be expressed as

$$\phi(B) X_t = W_t.$$

We will see that $\phi(B)$ is important to establish the stationarity of these processes and is called the *autoregressive* operator. Moreover, this quantity is closely related to another important property of AR processes called *causality*. Before formally defining this new property, we consider the following example, which provides an intuitive illustration of its importance.

**Example 23.** [Non-causal AR(1)] Consider a classical AR(1) model with $|\phi| > 1$. Such a model could be expressed as

$$X_t = \phi^{-1} X_{t+1} - \phi^{-1} W_t = \phi^{-k} X_{t+k} - \sum_{i=1}^{k-1} \phi^{-i} W_{t+i}.$$

Since $|\phi| > 1$, we obtain

$$X_t = -\sum_{i=1}^{\infty} \phi^{-j} W_{t-j},$$

which is a linear process and therefore is stationary. Unfortunately, such a model is useless because we need the future to predict the future and such processes are called non-causal.

### 3.2.1 Properties of AR models

In this section we discuss the main properties of AR($p$) processes. We first consider the causality of these models which has already been introduced in the previous paragraphs and then discuss the autocorrelation (and partial autocorrelation) of AR($p$) models.

#### 3.2.1.1 Causality

In the previous section we already introduced the idea of causal time series model and therefore let us now introduce this concept in a more formal manner.

**Definition 12** (Causality of AR models)**.** *An AR(p) model is said to be causal, if the time series* $(X_t)$ *can be written as a one-sided linear process:*

$$X_t = \sum_{j=0}^{\infty} \psi_j W_{t-j} = \frac{1}{\phi(B)} W_t = \psi(B) W_t,$$

*where* $\phi(B) = \sum_{j=0}^{\infty} \phi_j B^j$, $\sum_{j=0}^{\infty} |\phi_j| < \infty$ *and* $\phi_0 = 1$.

As discussed earlier this condition implies that only the past values of the time series can explain the future values of it, and not vice versa. As discussed in the previous section, linear processes are (weakly) stationary processes and therefore causality directly implies (weak) stationarity. However, it might be difficult and not obvious to show the causality of AR($p$) processes by using the above definitions directly, thus the following property is particularly useful in practice.

**Theorem 3** (Causality of AR models)**.** *If an AR(p) model is causal if and only if* $\phi(z) \neq 0$ *for* $|z| \leq 1$. *Then, the coefficients of the one-sided linear process given in 12 can be obtained by solving*

$$\psi(z) = \frac{1}{\sum_{j=0}^{\infty} \phi_j z^j} = \frac{1}{\phi(z)}, \ |z| \leq 1.$$

The proof of this result is omitted but can (for example) be found in Appendix B of Shumway and Stoffer (2010). Moreover, it can be seen how there is no solution to the above equation if $\phi(z) = 0$ for $|z| \leq 1$ and therefore an AR($p$) is causal if and only if $\phi(z) \neq 0$ for $|z| \leq 1$. A condition for this to be respected is for the roots of $\phi(z) = 0$ to lie *outside the unit circle*.

Before further discussing the properties of AR($p$) models we consider the following two examples which discuss the causality (and stationarity) of AR(2) models.

**Example 24.** [Transforming an AR(2) into a linear process] Given the following AR(2) $X_t = 1.3X_{t-1} - 0.4X_{t-2} + W_t$ one could wonder if this process can be written as a one-sided linear process as in Definition 12. This can be done using the following approach:

*Step 1:* The AR operator of this model can be expressed as:

$$\phi(z) = 1 - 1.3z + 0.4z^2 = (1 - 0.5z)(1 - 0.8z),$$

and has roots $2 > 1$ and $1.25 > 1$. Thus, we should be able to covert it into linear process.

*Step 2:* From Theorem 3 we know that if AR process has all its roots lie outside the unit circle, we can write $X_t = \phi^{-1}(B) W_t$ and "inverse" the autoregressive operator $\phi(B)$ as follows:

$$\phi^{-1}(z) = \frac{1}{(1 - 0.5z)(1 - 0.8z)} = \frac{c_1}{(1 - 0.5z)} + \frac{c_2}{(1 - 0.8z)} = \frac{c_2(1 - 0.5z) + c_1(1 - 0.8z)}{(1 - 0.5z)(1 - 0.8z)},$$

since we can think of the above equation is valid for any $z$, we will solve the following system of equations to get $c_1$ and $c_2$,

$$\begin{cases} c_1 + c_2 & = 1 \\ -0.5c_2 - 0.8c_1 & = 0 \end{cases} \Longrightarrow \begin{cases} c_1 & = -\frac{5}{3} \\ c_2 & = \frac{8}{3}, \end{cases}$$

and, we obtain

$$\phi^{-1}(z) = \frac{-5}{3(1-0.5z)} + \frac{8}{3(1-0.8z)}.$$

*Step 3:* Using Geometric series, i.e. $a\sum_{k=0}^{\infty} r^k = \frac{a}{1-r}$, if $|r|<1$, we have

$$\begin{cases} \frac{-5}{3(1-0.5z)} & = \frac{-5}{3}\sum_{j=0}^{\infty} 0.5^j z^j, \text{ if } |z|<2 \\ \frac{8}{3(1-0.8z)} & = \frac{8}{3}\sum_{j=0}^{\infty} 0.8^j z^j, \text{ if } |z|<1.25. \end{cases}$$

This allows to expressed $\phi^{-1}(z)$ as

$$\phi^{-1}(z) = \sum_{j=0}^{\infty}\left[\frac{-5}{3}(0.5)^j + \frac{8}{3}(0.8)^j\right] z^j, \text{ if } |z|<1.25.$$

*Step 4:* Finally, we obtain

$$X_t = \phi^{-1}(B)W_t = \sum_{j=0}^{\infty}\left[\frac{-5}{3}(0.5)^j + \frac{8}{3}(0.8)^j\right] B^j W_t = \sum_{j=0}^{\infty}\left[\frac{-5}{3}(0.5)^j + \frac{8}{3}(0.8)^j\right] W_{t-j},$$

which verifies that the AR(2) is causal (and stationary).

**Example 25.** [Causal conditions for an AR(2) processes] We already know that an AR(1) is causal with the simple condition $|\phi_1|<1$. It could seem natural to believe that an AR(2) should be causal (implies stationary) with the conditon: $|\phi_i|<1$, $i=1,2$, however, this is not the case. Indeed, an AR(2) can be expressed as

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + W_t = \phi_1 B X_t + \phi_2 B^2 X_t + W_t,$$

corresponding to the following autoregressive operator:

$$\phi(z) = 1 - \phi_1 z - \phi_2 z^2.$$

Therefore, the process is causal when the roots of $\phi(z)$ lies outside of the unit circle. Letting $z_1$ and $z_2$ denote those roots, we impose the following constraints to ensure the causality of the model:

$$|z_1|>1, \quad \text{where} \quad z_1 = \frac{\phi_1 + \sqrt{\phi_1^2 + 4\phi_2}}{-2\phi_2},$$
$$|z_2|>1, \quad \text{where} \quad z_2 = \frac{\phi_1 - \sqrt{\phi_1^2 + 4\phi_2}}{-2\phi_2},$$

note that $z_1$ and $z_2$ can be complex values.

Thus we can represent $\phi_1$ and $\phi_2$ by $z_1$ and $z_2$,

$$\phi_1 = (z_1^{-1} + z_2^{-1}),$$
$$\phi_2 = -(z_1 z_2)^{-1}.$$

Moreover we have the following equivalent condition for causality:

$$\begin{cases} |z_1| & > 1 \\ |z_2| & > 1, \end{cases}$$

if and only if

$$\begin{cases} \phi_1 + \phi_2 & < 1 \\ \phi_2 - \phi_1 & < 1 \\ |\phi_2| & < 1. \end{cases}$$

We can show "*if*"

$\phi_1 + \phi_2 = \frac{1}{z_1} + \frac{1}{z_2} - \frac{1}{z_1 z_2} = \frac{1}{z_1}\left(1 - \frac{1}{z_2}\right) + \frac{1}{z_2} < 1 - \frac{1}{z_2} + \frac{1}{z_2} = 1$, (since $\left(1 - \frac{1}{z_2}\right) > 0$)

$\phi_2 - \phi_1 = -\frac{1}{z_1 z_2} - \frac{1}{z_1} - \frac{1}{z_2} = -\frac{1}{z_1}\left(\frac{1}{z_2} + 1\right) - \frac{1}{z_2} < \frac{1}{z_2} + 1 - \frac{1}{z_2} = 1$, (since $\left(\frac{1}{z_2} + 1\right) > 0$)

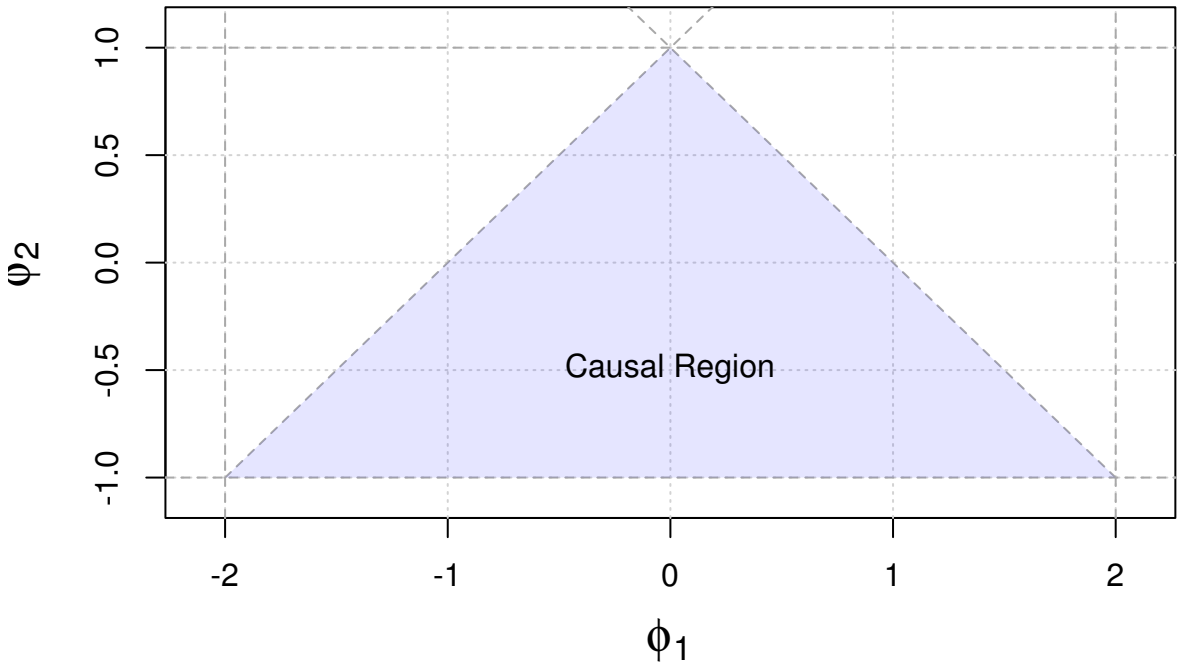
$|\phi_2| = \frac{1}{|z_1||z_2|} < 1.$

We can also show "*only if*"

Since $z_1 = \frac{\phi_1 + \sqrt{\phi_1^2 + 4\phi_2}}{-2\phi_2}$ and $\phi_2 - 1 < \phi_1 < 1 - \phi_2$, then $z_1^2 = \frac{\left(\phi_1 + \sqrt{\phi_1^2 + 4\phi_2}\right)^2}{4\phi_2^2} < \frac{\left((1-\phi_2) + \sqrt{(1-\phi_2)^2 + 4\phi_2}\right)^2}{4\phi_2^2} = \frac{4}{4\phi_2^2} \leq 1.$

Since $z_2 = \frac{\phi_1 - \sqrt{\phi_1^2 + 4\phi_2}}{-2\phi_2}$ and $\phi_2 - 1 < \phi_1 < 1 - \phi_2$, then $z_2^2 = \frac{\left(\phi_1 - \sqrt{\phi_1^2 + 4\phi_2}\right)^2}{4\phi_2^2} < \frac{\left((\phi_2 - 1) + \sqrt{(\phi_2 - 1)^2 + 4\phi_2}\right)^2}{4\phi_2^2} = \frac{4\phi_2^2}{4\phi_2^2} = 1.$

Finally, the causal region of an AR(2) is depicted on the figure below.




#### 3.2.1.2 Autocorrelation

In this section we discuss the autocorrelation of (causal) AR($p$) processes. Before considering the general case of an AR($p$), we revisit Example 24 and derive the ACF of the AR(2) model presented in this example.

**Example 26.** [Autocorrelation of an AR(2)] Considering the same model as in Example 24, i.e. $X_t = 1.3X_{t-1} - 0.4X_{t-2} + W_t$, we can derive the ACF using the following steps:

*Step 1:* Find the homogeneous difference equation with respect to the ACF $\rho(h)$, which in this case is given by:

$$\rho(h) - 1.3\rho(h-1) + 0.4\rho(h-2) = 0, \ h = 1, 2, ...$$

and the initial conditions are $\rho(0) = 1$ and $\rho(-1) = \frac{13}{14}$. Note that the above equation is an homogenous difference equation of order 2.

*Step 2:* Using the results of Example 24, we have:

$$\phi(z) = 1 - 1.3z + 0.4z^2 = (1 - 0.5z)(1 - 0.8z),$$

and the roots of this equation are given by $z_1 = 2 > 1$ and $z_2 = 1.25 > 1$. Moreover, since $z_1$ and $z_2$ are real and distinct, we obtain (since it corresponds to the solution of an homogenous difference equation of order 2):

$$\rho(h) = c_1 z_1^{-h} + c_2 z_2^{-h}.$$

*Step 3:* Solve $c_1$ and $c_2$ based on two initial conditions found in Step 1, i.e.

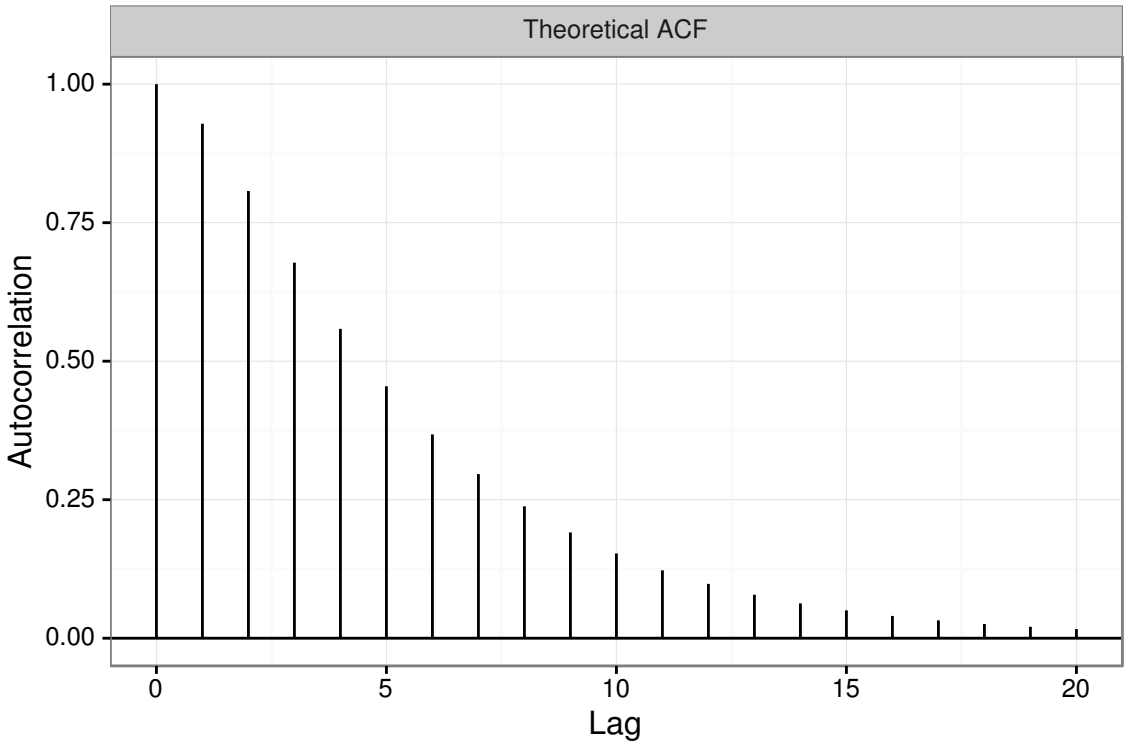
$$\begin{cases} \rho(0) = c_1 + c_2 = 1 \\ \rho(-1) = 2c_1 + 1.25c_2 = \frac{13}{14}, \end{cases},$$

implying that $c_1 = -\frac{3}{7}$ and $c_2 = \frac{10}{7}$. The ACF for this model is therefore given by

$$\rho(h) = -\frac{3}{7}2^{-h} + \frac{10}{7}\left(\frac{5}{4}\right)^{-h}.$$

The graph depicts the ACF of this process:

```
library(exts)
autoplot(theo_acf(AR(phi = c(1.3, -0.4))))
```



The method used in the previous is only applicable for AR(2) with roots that are distinct and real, which is true when $\phi_2 > -\phi^1/4$. In the case where $\phi_2 = -\phi^1/4$, the autoregressive operator has single root and $\rho(h)$ can be obtain by determining (using initial conditions) the constants $c_1$ and $c_2$ of the following expression:

$$\rho(h) = z_1^{-h}\left(c_1 + c_2 h\right).$$

When $\phi_2 > -\phi^1/4$ the roots are complex conjugate pairs and solution is given by:

$$\rho(h) = c_1|z_1|^{-h}\cos\left(h\theta + c_2\right),$$

where the constants depends on initial conditions, while $\theta = \arg(z_1)$.

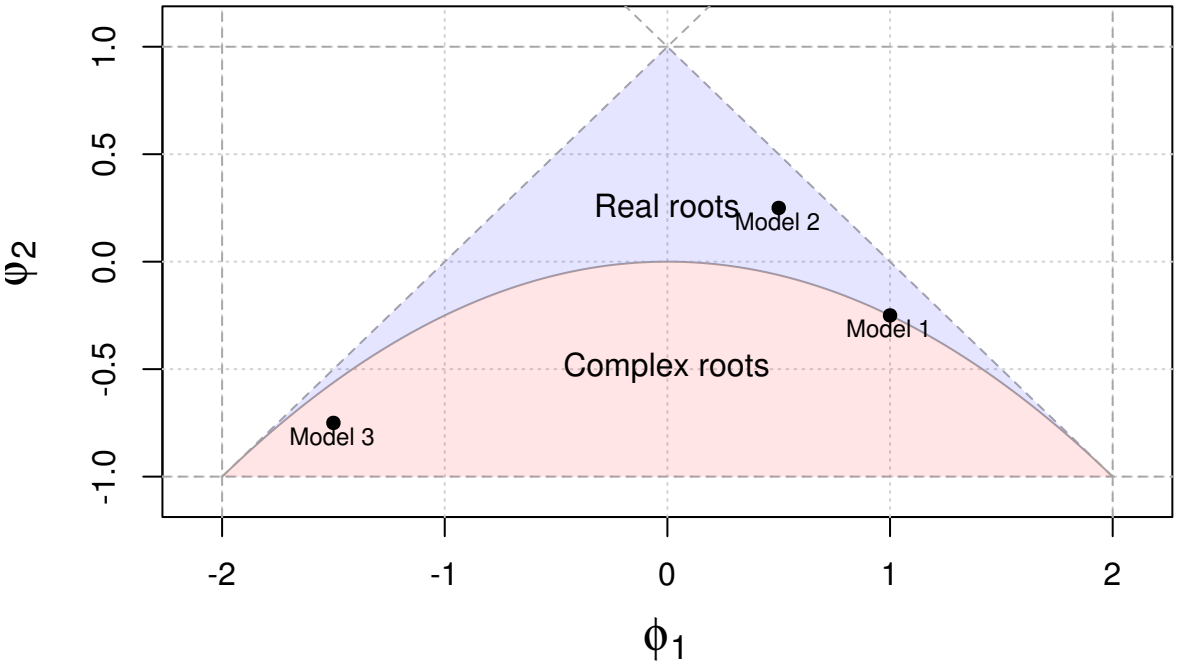
Therefore, we have that $\rho(h) \to 0$ exponential fast as $h \to \infty$ and when $\phi_2 > -\phi^1/4$ $\rho(h)$ goes to zero in an sinusoidal fashion. This behavior is illustrated in the next example.

**Example 27.** [Autocorrelation AR(2) Processes] Consider the following models:

$$\begin{aligned}
\text{Model 1}: &\quad X_t = X_{t-1} - 0.25X_{t-2} + W_t \\
\text{Model 2}: &\quad X_t = 0.5X_{t-1} + 0.25X_{t-2} + W_t \\
\text{Model 3}: &\quad X_t = -1.5X_{t-1} - 0.75X_{t-2} + W_t.
\end{aligned}$$

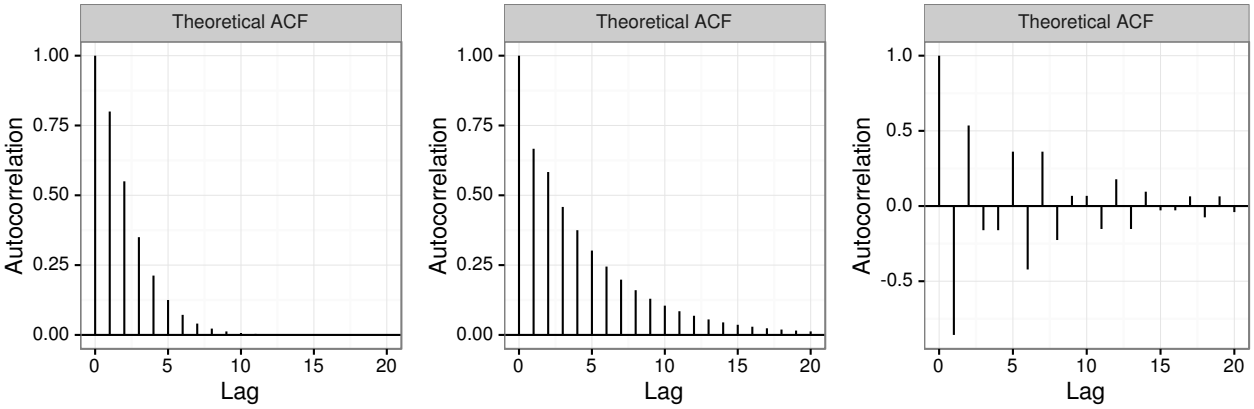
It is easy to verify the first one has real distrinct roots, the second a unique real root while the latter has complexe roots. This is illustrated in the figure which depicts the causal region of an AR(2) that has been separated between models with real and complexe roots.

Next we present an example to show how to derive the ACF for general causal AR(p) model. It is much more complicated than what we did for AR(2).




The ACF of these models are represented on the figure below. As expected the three models corresponds to an ACF that dampens exponentially fast but only Model 3 exhibits a sinusoidal features.

```
library(gridExtra)

# Define models
m1 = AR(phi = c(1, -0.25))
m2 = AR(phi = c(0.5,0.25))
m3 = AR(phi = c(-1.5, -0.75))

# Theoretical ACF
acf1 = theo_acf(m1)
acf2 = theo_acf(m2)
acf3 = theo_acf(m3)

# Plot ACFs
a1 = autoplot(acf1)
a2 = autoplot(acf2)
a3 = autoplot(acf3)
grid.arrange(a1, a2, a3, nrow = 1)
```




Next, we consider the ACF for a general causal AR($p$) model. Unfortunately, this is far more complicated than our previous example.

**Example 28.** [Autocorrelation of an AR($p$)] Recall that the AR($p$) models can be formally represented as follows

$$X_t = \phi_1 X_{t-1} + ... + \phi_p X_{t-p} + W_t,$$

where $W_t$ is a (Gaussian) white noise process with variance $\sigma^2$.

There are two ways to derive the ACF for general AR($p$) models. For the first one, since we assume our AR($p$) model is causal, we can write it as a one-sided linear process: $X_t = \sum_{j=0}^{\infty} \psi_j W_{t-j}$, then the autocovariance function $\gamma(h) = cov(X_{t+h}, X_t) = \sigma^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+h}$, for $h \geq 0$. The difficulty of this method is to solve the $\psi$-weights. Treating AR(p) as a special case of ARMA(p,q), we can let the MA polynomials $\theta(z) = 1$ and solve the $\psi$-weights by matching the coefficients in $\phi(z)\psi(z) = \theta(z)$.

For the second method, we can also use the procedure we used for AR(2). That is finding a homogeneous difference equation with respect to ACF $\rho(h)$, and solve it directly. To do so we need following steps.

*Step 1*: Find the homogeneous difference equation with repsect to the ACF $\rho(h)$.

Firstly verify whether our model is causal. If it is then we have,

$$\rho(h) - \phi_1 \rho(h-1) - \cdots - \phi_p \rho(h-p) = 0, \ h \geq p.$$

*Step 2*: Solve the roots of the associated AR polynomials.

The polynomials can be written as:

$$\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p.$$

Suppose the polynomials have $r$ distinct roots, and let $m_j$ denoted the number of replicates of $z_j$, such that $m_1 + m_2 + \cdots + m_r = p$.

Then the general solution of the homogeneous difference equation is

$$\rho(h) = z_1^{-h} P_1(h) + z_2^{-h} P_2(h) + \cdots + z_r^{-h} P_r(h), \ h \geq p,$$

where $P_j(h)$ is the polynomial in $h$ of degree $m_j - 1$.

*Step 3*: Solve every $P_j(h)$ based on $p$ given initial conditions on $\rho(h)$.

**Remark:** Since the AR($p$) model is causal, the roots we obtained in step 2 should be outside of the unit circle (i.e. $|z_i| > 1$, for $i = 1, \cdots, r$). Then the absolute value of the general solution

$$|\rho(h)| = \left| \frac{P_1(h)}{z_1^h} + \frac{P_2(h)}{z_2^h} + \cdots + \frac{P_r(h)}{z_r^h} \right| \leq \left| \frac{P_1(h)}{z_1^h} \right| + \left| \frac{P_2(h)}{z_2^h} \right| + \cdots + \left| \frac{P_r(h)}{z_r^h} \right| \leq \frac{r \left| P_r(h) \right|}{\min_{j=1,\ldots,r} |z_j|^h},$$

from the right hand side of the last inequality, we can find the rate of convergence would be dominated by $\frac{1}{\min_{j=1,\ldots,r} |z_j|^h}$. Thus the ACF $\rho(h)$ will goes to zero exponentially as $h \to \infty$.

EXAMPLE MISSING HERE + DISCUSSION ON LIMITS OF ACF

#### 3.2.1.3 Partial autocorrelation of AR models

**Definition 13** (Partial autocorrelation)**.** *For a stationary process, $X_t$, the Partial AutoCorrelation Function (PACF) can be denoted as $\phi_{hh}$, for $h = 1, 2, \ldots$, which are*

$$\phi_{11} = \text{corr}(X_{t+1}, X_t) = \rho(1),$$

*and*

$$\phi_{hh} = \text{corr}(X_{t+h} - \hat{X}_{t+h}, X_t - \hat{X}_t), \ h \geq 2.$$

**Remark** From the above definition we can think of the partial correlation $\phi_{hh}$ as the correlation between the residual of $X_{t+h}$ after removing its best linear predictor on the vector space spanned by $\{X_{t+1}, \ldots, X_{t+h-1}\}$ and the residual of $X_t$ after removing its best linear predictor on the same vector space. Similar to the linear regression, after projecting the residuals should be independent with the vector space spanned by $\{X_{t+1}, \ldots, X_{t+h-1}\}$.

We will discuss about the best linear predictor in the section Forecasting, here we just mention how to obtain the best linear predictor. Given a time series $X_1, X_2, \ldots, X_t$ with zero mean, the best linear predictor of $X_{t+h}$ can be written as $\hat{X}_{t+h} = \sum_{j=1}^{t} \alpha_j X_j$ such than it satisfies the prediction equations:

$$\mathbb{E}(X_{t+h} - \hat{X}_{t+h}) = 0,$$

and

$$\mathbb{E}[(X_{t+h} - \hat{X}_{t+h}) X_j] = 0, \text{ for } i = 1, \ldots, t.$$

According to the projection theorem, we can show that satisfying the prediction equations is equivalent to minimizing the mean square error $\mathbb{E}(X_{t+h} - \hat{X}_{t+h})$. Thus we have two equivalent to obtain the best linear predictor.

**Example 29.** [PACF of AR(1)] Consider a casual AR(1) model $X_t = \phi X_{t-1} + W_t$, We have,

$$\phi_{11} = \text{corr}(X_{t+1}, X_t) = \rho(1) = \phi,$$

and

$$\phi_{22} = \text{corr}(X_{t+2} - \hat{X}_{t+2}, X_t - \hat{X}_t),$$

Next, we find $\hat{X}_t$ and $\hat{X}_{t+2}$. Since $X_{t+1}$ is the only random between $X_t$ and $X_{t+2}$, $\hat{X}_t$ and $\hat{X}_{t+2}$ are the best linear predictors on the vector space spaned by $X_t$, we can obtain them by minimizing the MSE.

$$\mathbb{E}(X_{t+2} - \hat{X}_{t+2})^2 = \mathbb{E}(X_{t+2} - \beta_1 X_{t+1})^2 = \gamma(0) - 2\beta_1\gamma(1) + \beta_1^2\gamma(0),$$

Then by minimizing the MSE, we have $\beta_1 = \frac{\gamma(1)}{\gamma(0)} = \phi$.

Similarily, by minimizing

$$\mathbb{E}(X_t - \hat{X}_t)^2 = \mathbb{E}(X_t - \beta_2 X_{t+1})^2 = \gamma(0) - 2\beta_2\gamma(1) + \beta_2^2\gamma(0),$$

we have $\beta_2 = \frac{\gamma(1)}{\gamma(0)} = \phi$.

Or equivalently we can use the prediction equations. Thus we have

$$\mathbb{E}[(X_{t+2} - \hat{X}_{t+2})X_{t+1}] = \mathbb{E}[(X_{t+2}X_{t+1} - \beta_1 X_{t+1}^2)] = \gamma(1) - \beta_1\gamma(0) = 0,$$

and

$$\mathbb{E}[(X_t - \hat{X}_t)X_{t+1}] = \mathbb{E}[(X_t X_{t+1} - \beta_2 X_{t+1}^2)] = \gamma(1) - \beta_2\gamma(0) = 0,$$

Thus we can get the same solutions.

Therefore,

$$\phi_{22} = \text{corr}(X_{t+2} - \phi X_{t+1}, X_t - \phi X_{t+1}) = \text{corr}(W_{t+2}, X_t - \phi X_{t+1}) = 0,$$

note that the last equation is based on casuality.

**Example 30.** [PACF of AR(p)] In this example we would like to show that the PACF characterize the order of AR(p) models. That is when $h > p$, the PACF $\phi_{hh} = 0$. Suppose a causal AR(p) model, $X_{t+h} = \sum_{j=1}^{p} \phi_j X_{t+h-j} + W_{t+h}$, we want to calculate $\phi_{hh}$.

The best linear predictor of $X_{t+h}$ is

$$\hat{X}_{t+h} = \mathbb{E}\left[X_{t+h}|X_t, \ldots, X_{t+h-1}\right] = \mathbb{E}\left[\sum_{j=1}^{p} \phi_j X_{t+h-j} + W_{t+h}|X_t, \ldots, X_{t+h-1}\right] = \sum_{j=1}^{p} \mathbb{E}\left[\phi_j X_{t+h-j}|X_t, \ldots, X_{t+h-1}\right] = \sum_{j=1}^{p} \phi_j X$$

Thus when $h > p$,

$$\phi_{hh} = corr(X_{t+h} - \hat{X}_{t+h}, X_t - \hat{X}_t) = corr(W_{t+h}, X_t - \hat{X}_t) = 0.$$

### 3.2.2 Estimation of AR($p$) models

Given the above defined properties of the AR($p$) models, we will now discuss how these models can be estimated, more specifically how the $p+1$ parameters can be obtained from an observed time series. Indeed, a reliable estimation of these models is necessary in order to interpret and describe different natural phenomena and/or forecast possible future values of the time series.

A first approach builds upon the earlier definition of the AR(p) as being a linear process. Recall that

$$X_t = \sum_{j=1}^{p} \phi_j X_{t-j} + W_t \tag{3.1}$$

which delivers the following autocovariance function

$$\gamma(h) = cov(X_{t+h}, X_t) = cov(\sum_{j=1}^{p} \phi_j X_{t-j} + W_{t+h}, X_t) = \sum_{j=1}^{p} \phi_j \gamma(h-j),\ h \geq 0. \tag{3.2}$$

Rearranging the above expressions we obtain the following general equations

$$\gamma(h) - \sum_{j=1}^{p} \phi_j \gamma(h-j) = 0,\ h \geq 1 \tag{3.3}$$

and, recalling that $\gamma(h) = \gamma(-h)$,

$$\gamma(0) - \sum_{j=1}^{p} \phi_j \gamma(j) = \sigma_w^2. \tag{3.4}$$

We can now define the Yule-Walker equations.

**Definition 14** (Yule-Walker Equations)**.** *The Yule-Walker equations are given by*

$$\gamma(h) = \phi_1 \gamma(h-1) + ... + \phi_p \gamma(h-p),\ h = 1, ..., p \tag{3.5}$$

*and*

$$\sigma_w^2 = \gamma(0) - \phi_1 \gamma(1) - ... - \phi_p \gamma(p). \tag{3.6}$$

*which in matrix notation can be defined as follows*

$$\Gamma_p \phi = \gamma_p and \sigma_w^2 = \gamma(0) - \phi' \gamma_p \tag{3.7}$$

*where $\Gamma_p$ is the $p \times p$ matrix containing the autocovariances $\gamma(k-j), j, k = 1, ..., p$ while $\phi = (\phi_1, ..., \phi_p)'$ and $\gamma_p = (\gamma(1), ..., \gamma(p))'$ are $p \times 1$ vectors.*

Considering the Yule-Walker equations, it is possible to use a method of moments approach and simply replace the theoretical quantities given in the previous definition with their empirical (estimated) counterparts that we saw in the previous chapter. This gives us the following Yule-Walker estimators

$$\hat{\phi} = \hat{\Gamma}_p^{-1} \hat{\gamma}_p \text{and} \hat{\sigma}_w^2 = \hat{\gamma}(0) - \hat{\gamma}_p' \hat{\Gamma}_p^{-1} \hat{\gamma}_p \tag{3.8}$$

These estimators have the following asymptotic properties.

**Property: Consistency and Asymptotic Normality of Yule-Walker estimators** The Yule-Walker estimators for a causal AR(p) model have the following asymptotic properties:

$$\sqrt{T}(\hat{\phi} - \phi) \xrightarrow{\mathcal{D}} \mathcal{N}(\mathbf{0}, \sigma_w^2 \Gamma_p^{-1}) \text{and} \hat{\sigma}_w^2 \xrightarrow{\mathcal{P}} \sigma_w^2.$$

Therefore the Yule-Walker estimators have an asymptotically normal distribution and the estimator of the innovation variance is consistent. Moreover, these estimators are also optimal for AR(p) models, meaning that they are also efficient. However, there exists another method which allows to achieve this efficiency also for general ARMA models and this is the maximum likelihood method. Considering an AR(1) model

as an example, and assuming without loss of generality that the expectation is zero, we have the following representation of the AR(1) model

$$X_t = \phi X_{t-1} + W_t$$

where $|\phi| < 1$ and $W_t \overset{iid}{\sim} \mathcal{N}(0, \sigma_w^2)$. Supposing we have observations issued from this model $(x_t)_{t=1,...,T}$, then the likelihood function for this setting is given by

$$L(\phi, \sigma_w^2) = f(\phi, \sigma_w^2 | x_1, ..., x_T)$$

which, for an AR(1) model, can be rewritten as follows

$$L(\phi, \sigma_w^2) = f(x_1) f(x_2|x_1) \cdots f(x_T|x_{T-1}).$$

If we define $\Omega_t^p$ as the information contained in the previous $p$ observations to time $t$, the above can be generalized for an AR(p) model as follows

$$L(\phi, \sigma_w^2) = f(x_1, ..., x_p) f(x_{p+1}|\Omega_{p+1}^p) \cdots f(x_T|\Omega_{T-1}^p)$$

where $f(x_1, ..., x_p)$ is the joint probability distribution of the first $p$ observations. A discussion on how to find $f(x_1, ..., x_p)$ will be presented in the following paragraphs based on the approach to find $f(x_1)$ in the AR(1) likelihood. Going back to the latter, we know that $x_t|x_{t-1} \sim \mathcal{N}(\phi x_{t-1}, \sigma_w^2)$ and therefore we have that

$$f(x_t|x_{t-1}) = f_w(x_t - \phi x_{t-1})$$

where $f_w(\cdot)$ is the distribution of $w_t$. This rearranges the likelihood function as follows

$$L(\phi, \sigma_w^2) = f(x_1) \prod_{t=2}^{T} f_w(x_t - \phi x_{t-1})$$

where $f(x_1)$ can be found through the causal representation

$$x_1 = \sum_{j=0}^{\infty} \phi^j w_{1-j}$$

which implies that $x_1$ follows a normal distribution with zero expectation and a variance given by $\frac{\sigma_w^2}{(1-\phi^2)}$. Based on this, the likelihood function of an AR(1) finally becomes

$$L(\phi, \sigma_w^2) = (2\pi\sigma_w^2)^{-\frac{n}{2}} (1-\phi)^2 \exp\left(-\frac{S(\phi)}{2\sigma_w^2}\right)$$

with $S(\phi) = (1-\phi)^2 x_1^2 + \sum_{t=2}^{T} (x_t - \phi x_{t-1})^2$. Once the derivative of the logarithm of the likelihood is taken, the minimization of the negative of this function is usually done numerically. However, if we condition on the initial values, the AR(p) models are linear and, for example, we can then define the conditional likelihood of an AR(1) as

$$L(\phi, \sigma_w^2 | x_1) = (2\pi\sigma_w^2)^{-\frac{n-1}{2}} \exp\left(-\frac{S_c(\phi)}{2\sigma_w^2}\right)$$

where

$$S_c(\phi) = \sum_{t=2}^{T}(x_t - \phi x_{t-1})^2.$$

The latter is called the conditional sum of squares and $\phi$ can be estimated as a straightforward linear regression problem. Once an estimate $\hat{\phi}$ is obtained, this can be used to obtain the conditional maximum likelihood estimate of $\sigma_w^2$

$$\hat{\sigma}_w^2 = \frac{S_c(\hat{\phi})}{(n-1)}.$$

The estimation methods presented so far are standard ones for these kind of models. Nevertheless, if the data suffers from some form of contamination, these methods can become highly biased. For this reason, some robust estimators are available to limit this problematic if there are indeed outliers in the observed time series. A first solution is given by the estimator proposed in Kunsch (1984) who underlines that the MLE score function of an AR(p) is given by

$$\kappa(\theta|x_j, ...x_{j+p}) = \frac{\partial}{\partial\theta}(x_{j+p} - \sum_{k=1}^{p}\phi_k x_{j+p-k})^2$$

where $\theta$ is the parameter vector containing, in the case of an AR(1) model, the two parameters $\phi$ and $\sigma_w^2$ (i.e. $\theta = [\phi\ \sigma_w^2]$). This delivers the estimating equation

$$\sum_{j=1}^{n-p}\kappa(\hat{\theta}|x_j, ...x_{j+p}) = 0.$$

The score function $\kappa(\cdot)$ is clearly not bounded, in the sense that if we arbitrarily move a value of $(x_t)$ to infinity then the score function also goes to infinity thereby delivering a biased estimation procedure. To avoid that outlying observations bias the estimation excessively, a bounded score function can be used to deliver an M-estimator given by

$$\sum_{j=1}^{n-p}\psi(\hat{\theta}|x_j, ...x_{j+p}) = 0,$$

where $\psi(\cdot)$ is a function of bounded variation. When conditioning on the first $p$ observations, this problem can be brought back to a linear regression problem which can be applied in a robust manner using the robust regression tools available in R such as `rlm(...)` or `lmrob(...)`. However, another available tool in R which can applied directly without conditioning also for general ARMA models is the `gmwm(...)` function which, by specifying the option `robust = TRUE`. This function makes use of a quantity called the wavelet variance (denoted as $\nu$) which is estimated robustly and then used to retrieve the parameters $\theta$ of the time series model. The robust estimate is obtained by solving the following minimization problem

$$\hat{\theta} = \underset{\theta\in\Theta}{\operatorname{argmin}}(\hat{\nu} - \nu\theta)^T\Omega(\hat{\nu} - \nu\theta),$$

where $\hat{\nu}$ is the robustly estimated wavelet variance, $\nu\theta$ is the theoretical wavelet variance and $\Omega$ is positive definite weighting matrix. Below we show some simulation studies where we present the results of the above estimation procedures in absence and in presence of contamination in the data.

INSERT SIMULATIONS

For all the above methods, it would be necessary to understand how "precise" their estimates are. To do so we would need to obtain confidence intervals for these estimates and this can be done mainly in two manners:

- using the asymptotic distribution of the parameter estimates;
- using parametric bootstrap.

The first approach consists in using the asymptotic distribution of the estimators presented earlier to deliver approximations of the confidence intervals which get better as the length of the observed time series increases. Hence, for example, if we wanted to find a 95% confidence interval for the parameter $\phi$, we would use the quantiles of the normal distribution (given that all methods presented earlier present this asymptotic distribution). However, this approach can present some drawbacks, one of which is there behaviour when the parameters are close to the boundaries of the parameter space. Let us take the example of an AR(1) to illustrate this problematic and suppose that $\phi = 0.99 < 1$ and $\sigma^2_W$. It can be seen that the parameter $\phi$ respects the condition for stationarity but is very close to its boundary. The code below computes the confidence interval for $\phi$ using the asymptotic normal distribution.

```
## [1] 0.9840441 1.0068569
```

It can be seen how the confidence interval contains values that make the AR(1) non-stationary (i.e. values of $\phi$ larger than 1). For this purpose, the approach based on parametric bootstrap provides a viable solution. Indeed, parametric bootstrap takes the estimated parameter values and uses them in order to simulate from an AR(1) based on these parameter values. For each simulation the parameters are estimated again and saved. Finally, the empirical quantiles of the saved estimated parameter values provide a confidence interval which does not suffer from boundary problems. The code below gives an example of how this confidence interval is built based on the same estimation procedure but using parametric bootstrap (using 500 bootstrap replicates).

```
##      2.5%     97.5%
## 0.8764531 0.9989982
```

In this case, it can be observed that the confidence interval lies entirely within the boundaries of the parameter space.

# Appendix A

# Proofs

## A.1 Proof of Theorem 1

We let $X_t = W_t + \mu$, where $\mu < \infty$ and $(W_t)$ is a strong white noise process with variance $\sigma^2$ and finite fourth moment (i.e. $\mathbb{E}[W_t^4] < \infty$).

Next, we consider the sample autocovariance function computed on $(X_t)$, i.e.

$$\hat{\gamma}(h) = \frac{1}{n}\sum_{t=1}^{n-h}\left(X_t - \bar{X}\right)\left(X_{t+h} - \bar{X}\right).$$

For this equation, it is clear that $\hat{\gamma}(0)$ and $\hat{\gamma}(h)$ (with $h > 0$) are two statistics involving sums of different lengths. As we will see, this prevents us from using directly the multivariate central limit theorem on the vector $[\hat{\gamma}(h) \quad \hat{\gamma}(h)]^T$. However, the lag $h$ is fixed and therefore the difference in the number of elements of both sums is asymptotically negligible. Therefore, we define a new statistic

$$\tilde{\gamma}(h) = \frac{1}{n}\sum_{t=1}^{n}(X_t - \mu)(X_{t+h} - \mu),$$

which, as we will see, is easier to used and show that $\hat{\gamma}(h)$ and $\tilde{\gamma}(h)$ are asymptotically equivalent in the sense that:

$$n^{\frac{1}{2}}[\tilde{\gamma}(h) - \hat{\gamma}(h)] = o_p(1).$$

Therefore, assuming this results to be true, $\tilde{\gamma}(h)$ and $\hat{\gamma}(h)$ would have the same asymptotic distribution, it is sufficient to show the asymptotic distribution of $\tilde{\gamma}(h)$. So that before continuing the proof the Theorem 1 we first state and prove the following lemma:

**Lemma A1:** Let

$$X_t = \mu + \sum_{j=-\infty}^{\infty}\psi_j W_{t-j},$$

where $(W_t)$ is a strong white process with variance $\sigma^2$, and the coefficients satisfying $\sum|\psi_j| < \infty$. Then, we have

$$n^{\frac{1}{2}}[\tilde{\gamma}(h) - \hat{\gamma}(h)] = o_p(1).$$

*Proof:* By Markov inequality, we have

$$\mathbb{P}\left(|n^{\frac{1}{2}}[\tilde{\gamma}(h)-\hat{\gamma}(h)]|\geq\varepsilon\right)\leq\frac{\mathbb{E}|n^{\frac{1}{2}}[\tilde{\gamma}(h)-\hat{\gamma}(h)]|}{\varepsilon},$$

for any $\varepsilon>0$. Thus, it is enough to show that

$$\lim_{n\to\infty}\ \mathbb{E}\left[|n^{\frac{1}{2}}[\tilde{\gamma}(h)-\hat{\gamma}(h)]|\right]=0$$

to prove Lemma A1.By the definitions of $\tilde{\gamma}(h)$ and $\hat{\gamma}(h)$, we have

$$\begin{aligned}
n^{\frac{1}{2}}[\tilde{\gamma}(h)-\hat{\gamma}(h)] &= \frac{1}{\sqrt{n}}\sum_{t=n-h+1}^{n}(X_t-\mu)(X_{t+h}-\mu)\\
&\quad+\frac{1}{\sqrt{n}}\sum_{t=1}^{n-h}\left[(X_t-\mu)(X_{t+h}-\mu)-(X_t-\bar{X})(X_{t+h}-\bar{X})\right]\\
&=\frac{1}{\sqrt{n}}\sum_{t=n-h+1}^{n}(X_t-\mu)(X_{t+h}-\mu)+\frac{1}{\sqrt{n}}\sum_{t=1}^{n-h}\left[(\bar{X}-\mu)(X_t+X_{t+h}-\mu-\bar{X})\right]\\
&=\frac{1}{\sqrt{n}}\sum_{t=n-h+1}^{n}(X_t-\mu)(X_{t+h}-\mu)+\frac{1}{\sqrt{n}}(\bar{X}-\mu)\sum_{t=1}^{n-h}(X_t+X_{t+h}-\mu-\bar{X})\\
&=\frac{1}{\sqrt{n}}\sum_{t=n-h+1}^{n}(X_t-\mu)(X_{t+h}-\mu)+\frac{1}{\sqrt{n}}(\bar{X}-\mu)\left[\sum_{t=1+h}^{n-h}X_t-(n-h)\mu+h\bar{X}\right]\\
&=\frac{1}{\sqrt{n}}\sum_{t=n-h+1}^{n}(X_t-\mu)(X_{t+h}-\mu)+\frac{1}{\sqrt{n}}(\bar{X}-\mu)\left[\sum_{t=1+h}^{n-h}(X_t-\mu)-h(\mu-\bar{X})\right]\\
&=\frac{1}{\sqrt{n}}\sum_{t=n-h+1}^{n}(X_t-\mu)(X_{t+h}-\mu)+\frac{1}{\sqrt{n}}(\bar{X}-\mu)\sum_{t=1+h}^{n-h}(X_t-\mu)+\frac{h}{\sqrt{n}}(\bar{X}-\mu)^2,
\end{aligned}$$

where $\bar{X}=\frac{1}{n}\sum_{t=1}^{n}X_t=\mu+\frac{1}{n}\sum_{t=1}^{n}\sum_{j=-\infty}^{\infty}\psi_jW_{t-j}=\mu+\frac{1}{n}\sum_{j=-\infty}^{\infty}\sum_{t=1}^{n}\psi_jW_{t-j}$.
Then, we have

$$\begin{aligned}
\mathbb{E}\left[\left|n^{\frac{1}{2}}[\tilde{\gamma}(h)-\hat{\gamma}(h)]\right|\right] &\leq \frac{1}{\sqrt{n}}\sum_{t=n-h+1}^{n}\mathbb{E}\left[|(X_t-\mu)(X_{t+h}-\mu)|\right]\\
&\quad+\frac{1}{\sqrt{n}}\mathbb{E}\left[\left|(\bar{X}-\mu)\sum_{t=1+h}^{n-h}(X_t-\mu)\right|\right]+\frac{h}{\sqrt{n}}\mathbb{E}\left[(\bar{X}-\mu)^2\right].
\end{aligned}$$

Next, we consider each term of the above equation. For the first term, since $(X_t-\mu)^2=\left(\sum_{j=-\infty}^{\infty}\psi_jW_{t-j}\right)^2$, and $\mathbb{E}[W_iW_j]\neq0$ only if $i=j$. By Cauchy–Schwarz inequality we have

$$\mathbb{E}\left[|(X_t-\mu)(X_{t+h}-\mu)|\right]\leq\sqrt{\mathbb{E}\left[|(X_t-\mu)|^2\right]\mathbb{E}\left[|(X_{t+h}-\mu)|^2\right]}=\sigma^2\sum_{i=-\infty}^{\infty}\psi_i^2.$$

Then, we consider the third term, since it will be used in the second term

$$\mathbb{E}[(\bar{X}-\mu)^2]=\frac{1}{n^2}\sum_{t=1}^{n}\sum_{i=-\infty}^{\infty}\psi_i^2\mathbb{E}\left[W_{t-i}^2\right]=\frac{\sigma^2}{n}\sum_{i=-\infty}^{\infty}\psi_i^2.$$

Similarly, for the second term we have

$$\begin{aligned}
\mathbb{E}\left[\left|(\bar{X}-\mu)\sum_{t=1+h}^{n-h}(X_t-\mu)\right|\right] &\leq \sqrt{\mathbb{E}\left[|(\bar{X}-\mu)|^2\right]\mathbb{E}\left[|\sum_{t=1+h}^{n-h}(X_t-\mu)|^2\right]} \\
&= \sqrt{\mathbb{E}\left[(\bar{X}-\mu)^2\right]\mathbb{E}\left[\sum_{t=1+h}^{n-h}(X_t-\mu)^2+\sum_{t_1\neq t_2}(X_{t_1}-\mu)(X_{t_2}-\mu)\right]} \\
&\leq \sqrt{\frac{\sigma^2}{n}\sum_{i=-\infty}^{\infty}\psi_i^2\cdot(n-2h)\sigma^2\left(\sum_{j=-\infty}^{\infty}|\psi_j|\right)^2} \\
&\leq \sqrt{\frac{n-2h}{n}}\sigma^2\left(\sum_{i=-\infty}^{\infty}|\psi_i|\right)^2.
\end{aligned}$$

Combining the above results we obtain

$$\begin{aligned}
\mathbb{E}|n^{\frac{1}{2}}[\tilde{\gamma}(h)-\hat{\gamma}(h)]| &\leq \frac{1}{\sqrt{n}}h\sigma^2\sum_{i=-\infty}^{\infty}\psi_i^2+\sqrt{\frac{n-2h}{n^2}}\sigma^2\left(\sum_{i=-\infty}^{\infty}|\psi_i|\right)^2+\frac{h}{n\sqrt{n}}\sigma^2\sum_{i=-\infty}^{\infty}\psi_i^2 \\
&\leq \frac{1}{n\sqrt{n}}(nh+\sqrt{n-2h}+h)\sigma^2\left(\sum_{i=-\infty}^{\infty}|\psi_i|\right)^2,
\end{aligned}$$

By the taking the limit in $n$ we have

$$\lim_{n\to\infty}\mathbb{E}\left[|n^{\frac{1}{2}}[\tilde{\gamma}(h)-\hat{\gamma}(h)]|\right]\leq\sigma^2\left(\sum_{i=-\infty}^{\infty}|\psi_i|\right)^2\lim_{n\to\infty}\frac{nh+\sqrt{n-2h}+h}{n\sqrt{n}}=0.$$

We can therefore conclude that

$$\sqrt{n}[\tilde{\gamma}(h)-\hat{\gamma}(h)]=o_p(1),$$

which concludes the proof of Lemma A1. ■

Returning to the proof of Theorem 1, since the process $(Y_t)$, where $Y_t=(X_t-\mu)(X_{t+h}-\mu)$, is iid, we can apply multivariate central limit theorem to the vector $[\tilde{\gamma}(h)\quad\tilde{\gamma}(h)]^T$, and we obtain

$$\begin{aligned}
\sqrt{n}\left\{\begin{bmatrix}\tilde{\gamma}(0)\\ \tilde{\gamma}(h)\end{bmatrix}-\mathbb{E}\begin{bmatrix}\tilde{\gamma}(0)\\ \tilde{\gamma}(h)\end{bmatrix}\right\} &= \frac{1}{\sqrt{n}}\begin{bmatrix}\sum_{t=1}^{n}(X_t-\mu)^2-n\mathbb{E}[\tilde{\gamma}(0)]\\ \sum_{t=1}^{n}(X_t-\mu)(X_{t+h}-\mu)-n\mathbb{E}[\tilde{\gamma}(h)]\end{bmatrix} \\
&\xrightarrow{\mathcal{D}}\mathcal{N}\left(0,n\ \text{var}\left(\begin{bmatrix}\tilde{\gamma}(0)\\ \tilde{\gamma}(h)\end{bmatrix}\right)\right)
\end{aligned}$$

Moreover, by Cauchy–Schwarz inequality and since $\text{var}(X_t)=\sigma^2$, we have

$$\sum_{t=1}^{n}(X_t-\mu)(X_{t+h}-\mu)\leq\sqrt{\sum_{t=1}^{n}(X_t-\mu)^2\sum_{t=1}^{n}(X_{t+h}-\mu)^2}<\infty.$$

Therefore, by bounded convergence theorem and $(W_t)$ is iid, we have

$$\begin{aligned}\mathbb{E}[\tilde{\gamma}(h)]&=\frac{1}{n}\mathbb{E}\left[\sum_{t=1}^{n}(X_t-\mu)(X_{t+h}-\mu)\right]\\&=\frac{1}{n}\left[\sum_{t=1}^{n}\mathbb{E}(X_t-\mu)\,\mathbb{E}(X_{t+h}-\mu)\right]=\begin{cases}\sigma^2, & \text{for } h=0\\ 0, & \text{for } h\neq 0\end{cases}.\end{aligned}$$

Next, we consider the variance of $\tilde{\gamma}(h)$ when $h\neq 0$,

$$\begin{aligned}var[\tilde{\gamma}(h)]&=\frac{1}{n^2}\mathbb{E}\left\{\left[\sum_{t=1}^{n}(X_t-\mu)(X_{t+h}-\mu)\right]^2\right\}\\&=\frac{1}{n^2}\mathbb{E}\left\{\left[\sum_{i=1}^{n}(X_i-\mu)(X_{i+h}-\mu)\right]\left[\sum_{j=1}^{n}(X_j-\mu)(X_{j+h}-\mu)\right]\right\}\\&=\frac{1}{n^2}\mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i-\mu)(X_{i+h}-\mu)(X_j-\mu)(X_{j+h}-\mu)\right].\end{aligned}$$

Also by Cauchy–Schwarz inequality and the finite fourth moment assumption, we can use the bounded convergence theorem. Once again since $(W_t)$ is white noise process, we have

$$\mathbb{E}\left[(X_i-\mu)(X_{i+h}-\mu)(X_j-\mu)(X_{j+h}-\mu)\right]\neq 0$$

only when $i=j$.

Therefore, we obtain

$$\begin{aligned}var[\tilde{\gamma}(h)]&=\frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left[(X_i-\mu)^2(X_{i+h}-\mu)^2\right]\\&=\frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}(X_i-\mu)^2\,\mathbb{E}(X_{i+h}-\mu)^2=\frac{1}{n}\sigma^4.\end{aligned}$$

Similarly, for $h=0$, we have

$$var[\tilde{\gamma}(0)]=\frac{1}{n^2}\mathbb{E}\left\{\left[\sum_{t=1}^{n}(X_t-\mu)^2\right]^2\right\}-\frac{1}{n^2}\left[\mathbb{E}\sum_{t=1}^{n}(X_t-\mu)^2\right]^2=\frac{2}{n}\sigma^4.$$

Next, we consider the covariance between $\tilde{\gamma}(0)$ and $\tilde{\gamma}(h)$, for $h\neq 0$, and we obtain

$$\begin{aligned}cov[\tilde{\gamma}(0),\tilde{\gamma}(h)]&=\mathbb{E}[\tilde{\gamma}(0)\tilde{\gamma}(h)]-\mathbb{E}[\tilde{\gamma}(0)]\mathbb{E}[\tilde{\gamma}(h)]=\mathbb{E}[\tilde{\gamma}(0)\tilde{\gamma}(h)]\\&=\mathbb{E}\left[\left[\sum_{t=1}^{n}(X_t-\mu)^2\right]\left[\sum_{t=1}^{n}(X_t-\mu)(X_{t+h}-\mu)\right]\right]=0.\end{aligned}$$

Therefore by Slutsky's Theorem we have,

$$\sqrt{n}\left\{\begin{bmatrix}\hat{\gamma}(0)\\ \hat{\gamma}(h)\end{bmatrix}-\begin{bmatrix}\sigma^2\\ 0\end{bmatrix}\right\}=\sqrt{n}\left\{\begin{bmatrix}\tilde{\gamma}(0)\\ \tilde{\gamma}(h)\end{bmatrix}-\begin{bmatrix}\sigma^2\\ 0\end{bmatrix}\right\}+\underbrace{\sqrt{n}\left\{\begin{bmatrix}\hat{\gamma}(0)\\ \hat{\gamma}(h)\end{bmatrix}-\begin{bmatrix}\tilde{\gamma}(0)\\ \tilde{\gamma}(h)\end{bmatrix}\right\}}_{\xrightarrow{p}0}$$

$$\xrightarrow{\mathcal{D}}\mathcal{N}\left(0,\begin{bmatrix}2\sigma^4 & 0\\ 0 & \sigma^4\end{bmatrix}\right).$$

Next, we define the function $g\left(\begin{bmatrix}a\\ b\end{bmatrix}\right)=b/a$, where $a\neq 0$. For this function it is clear that

$$\nabla g\left(\begin{bmatrix}a\\ b\end{bmatrix}\right)=\begin{bmatrix}-\frac{b}{a^2}\\ \frac{1}{a}\end{bmatrix}^T,$$

and thus using the Delta method, we have for $h\neq 0$

$$\sqrt{n}\hat{\rho}(h)=\sqrt{n}\left\{g\left(\begin{bmatrix}\hat{\gamma}(0)\\ \hat{\gamma}(h)\end{bmatrix}\right)-\mu\right\}\xrightarrow{\mathcal{D}}\mathcal{N}\left(0,\sigma_r^2\right),$$

where

$$\begin{aligned}\mu&=g\left(\begin{bmatrix}\sigma^2 & 0\end{bmatrix}\right)=0,\\ \sigma_r^2&=\nabla g\left(\begin{bmatrix}\sigma^2\\ 0\end{bmatrix}\right)\begin{bmatrix}2\sigma^4 & 0\\ 0 & \sigma^4\end{bmatrix}\nabla g\left(\begin{bmatrix}\sigma^2\\ 0\end{bmatrix}\right)^T=\begin{bmatrix}0 & \sigma^{-2}\end{bmatrix}\begin{bmatrix}2\sigma^4 & 0\\ 0 & \sigma^4\end{bmatrix}\begin{bmatrix}0\\ \sigma^{-2}\end{bmatrix}=1.\end{aligned}$$

Thus, we have

$$\sqrt{n}\hat{\rho}(h)\xrightarrow{\mathcal{D}}\mathcal{N}(0,1),$$

which concludes the proof the Theorem 1. ■

# Bibliography

Shumway, R. and Stoffer, D. (2010). *Time Series Analysis and Its Applications: With R Examples*. Springer Texts in Statistics. Springer New York.