A Tour of Time Series Analysis with R

James Balamuta, Stéphane Guerrier, Roberto Molinari and Haotian Xu

2016-09-09

Contents

Preface **5**

Contributing 5

Bibliographic Note 5

Rendering Mathematical Formulae 6

R Code Conventions 6

License 6

1 Introduction **7**

1.1 Time Series 7

1.2 Exploratory Data Analysis for Time Series 8

1.3 Basic Time Series Models 13

1.4 Composite Stochastic Processes 20

2 Autocorrelation and Stationarity **23**

2.1 The Autocorrelation and Autocovariance Functions 23

2.2 Stationarity 27

2.3 Estimation of Moments of Stationary Processes 32

A Proofs **51**

A.1 Proof of Theorem 1 51

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

```r
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

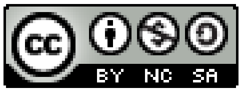


Figure 1: This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

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

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

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.

t = 1
t = 2
...
t = n − 2
t = n − 1
t = n

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:

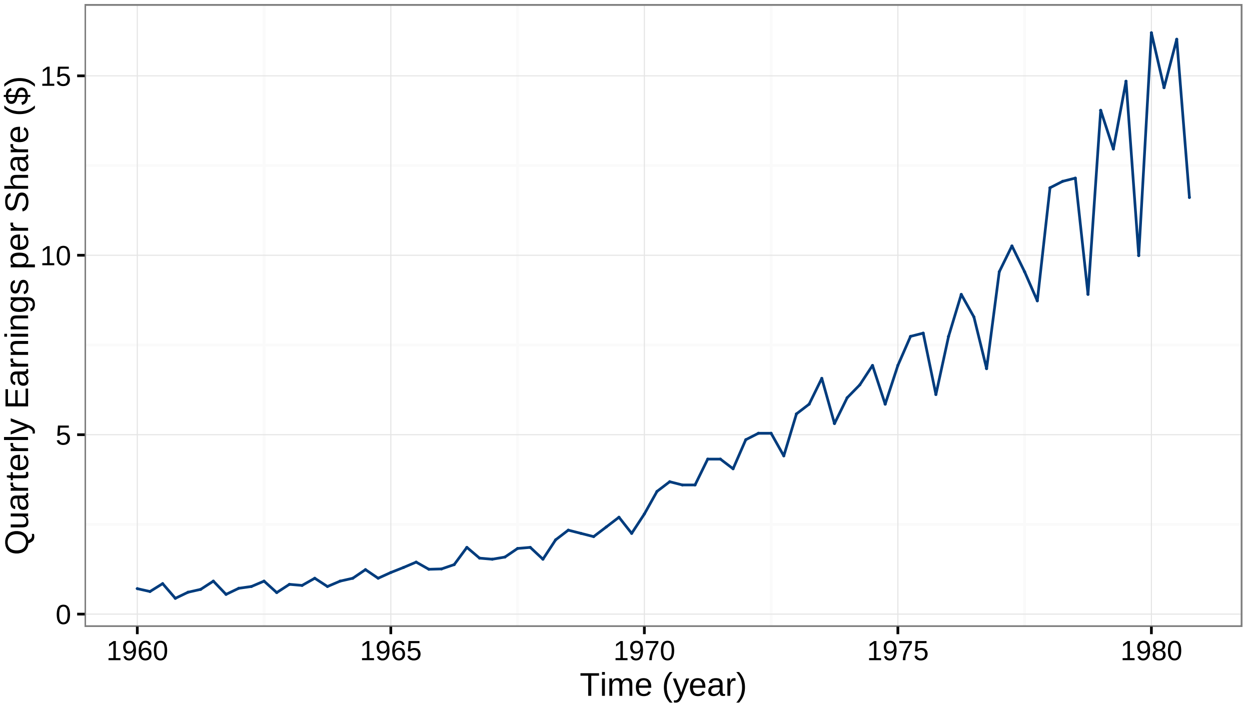
1. **Seasonal trends** (periodic): These are the cyclical patterns which repeat after a fixed/regular time period. This could be due to business cycles (e.g. bust/recession, recovery).

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: 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:

```r
# 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 ($)")
```



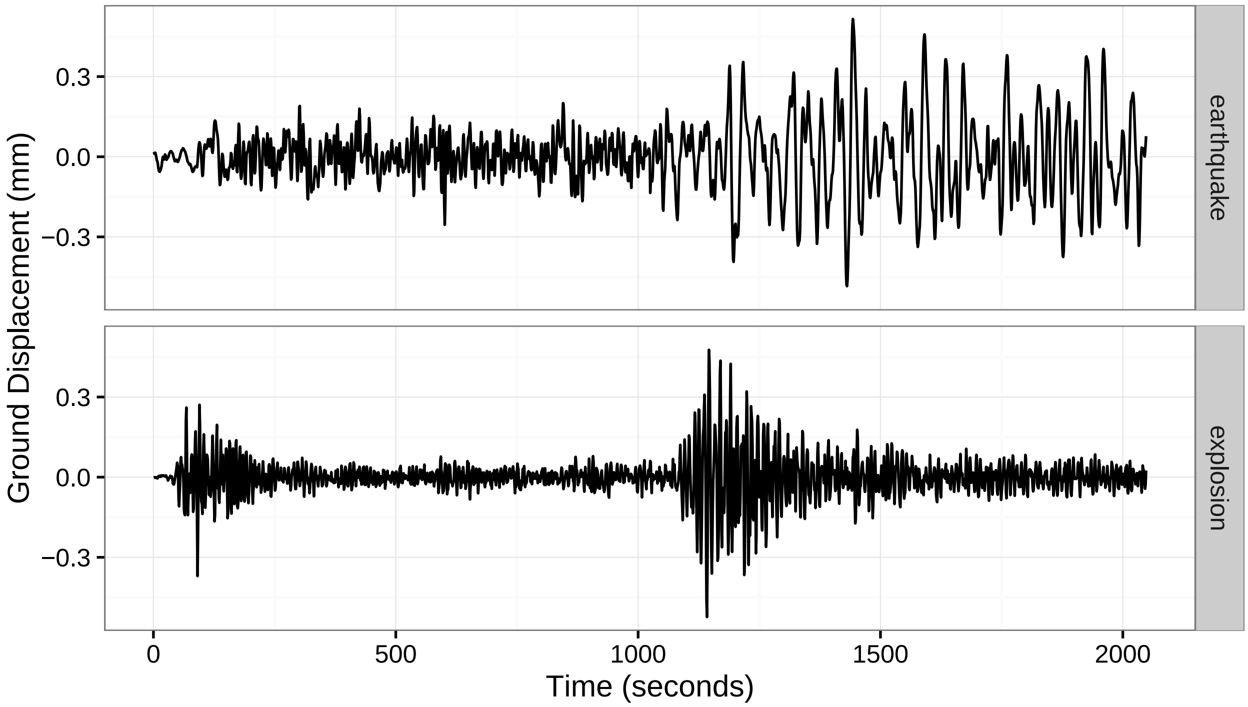
One trait that the graph makes evident is that the data contains a non-linear increasing trend as well as a yearly seasonal component. In addition, one can note that the *variability* of the data seems to increase with time. Being able to make such observations provides important information to select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

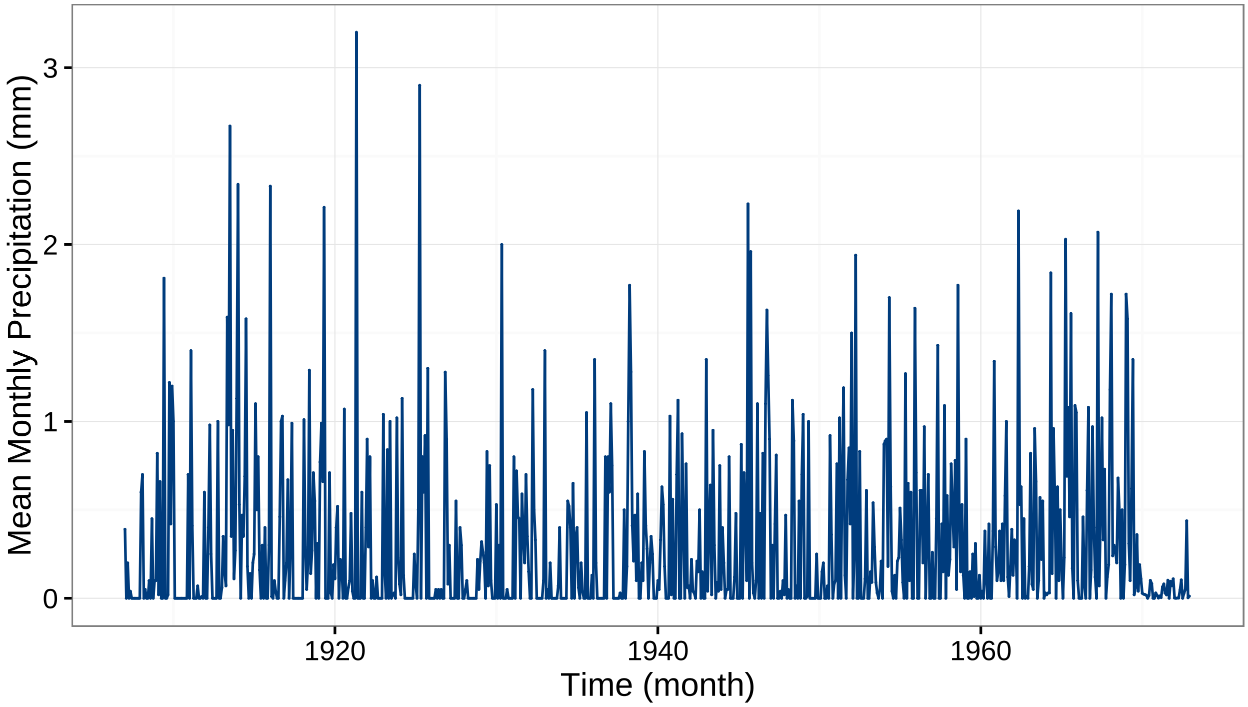


From the graph, it can be observed that the statistical properties of the time series appear to change over time. For instance, the variance of the time series shifts at around $t = 1150$ for both series. The shift in variance also opens “windows” where there appear to be distinct states. In the case 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: 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:

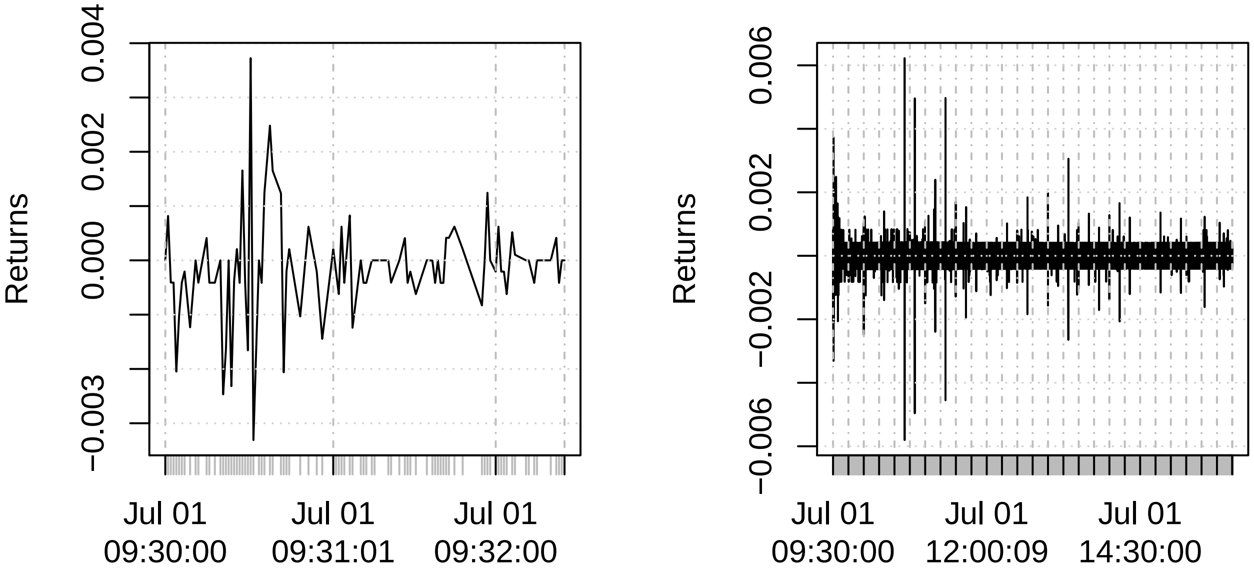
```
# Load data
hydro = read.csv("data/precipitation.csv", header=T, sep=";")

# 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: 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).

```
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")
```



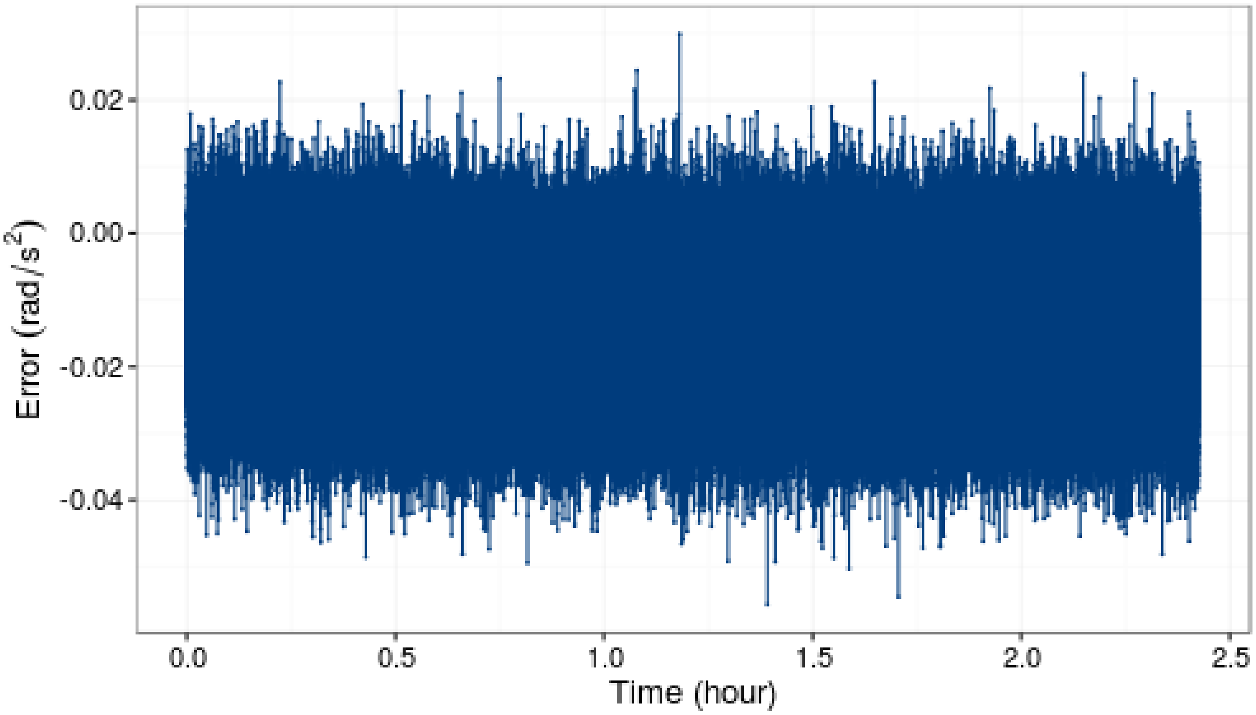
It can be observed on the left panel that observations are not equally spaced. Indeed, in high-frequency data the intervals between two points is typically not constant and, even worse, is a random variable. This implies that the time when a new observation will be available is in general unknown. On 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: 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:

```r
# 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 Ω_t as all the information avaiable up to time $t-1$, i.e.

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

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 $h \neq 0$.

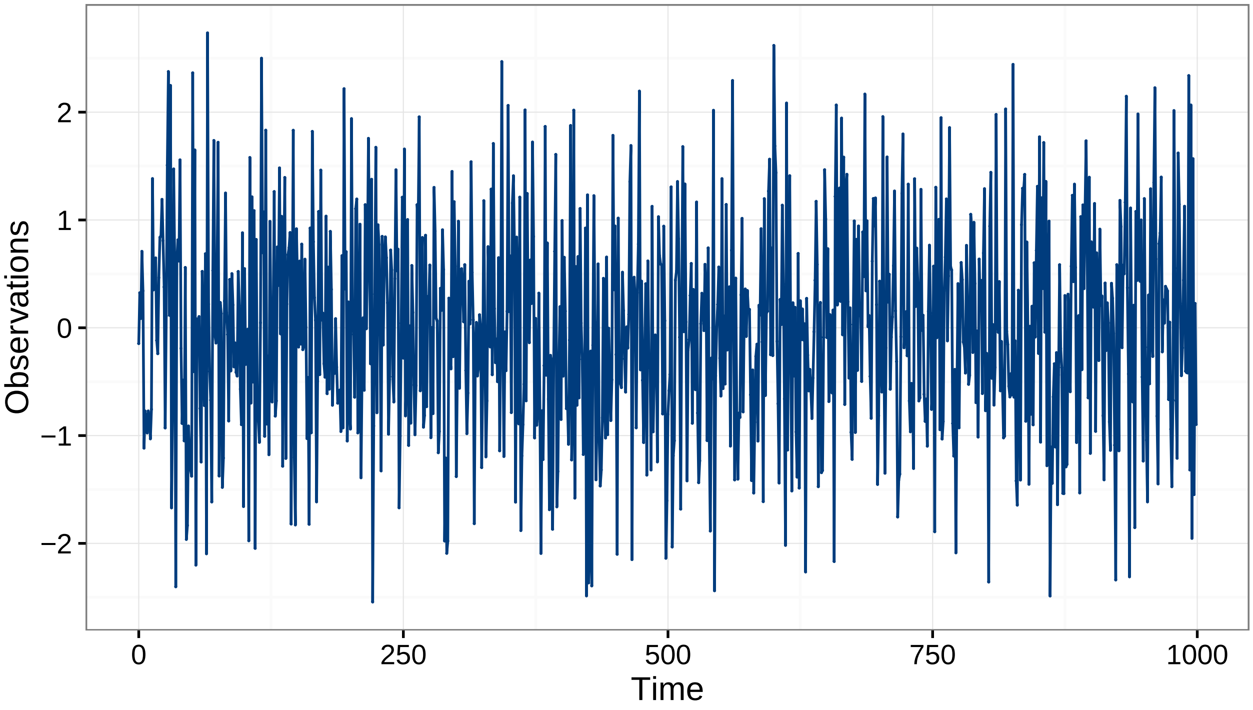
It is clear from these definitions that if a process is a strong white noise it is also a weak white noise. However, the converse is not true as shown in the following example:

Example: 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

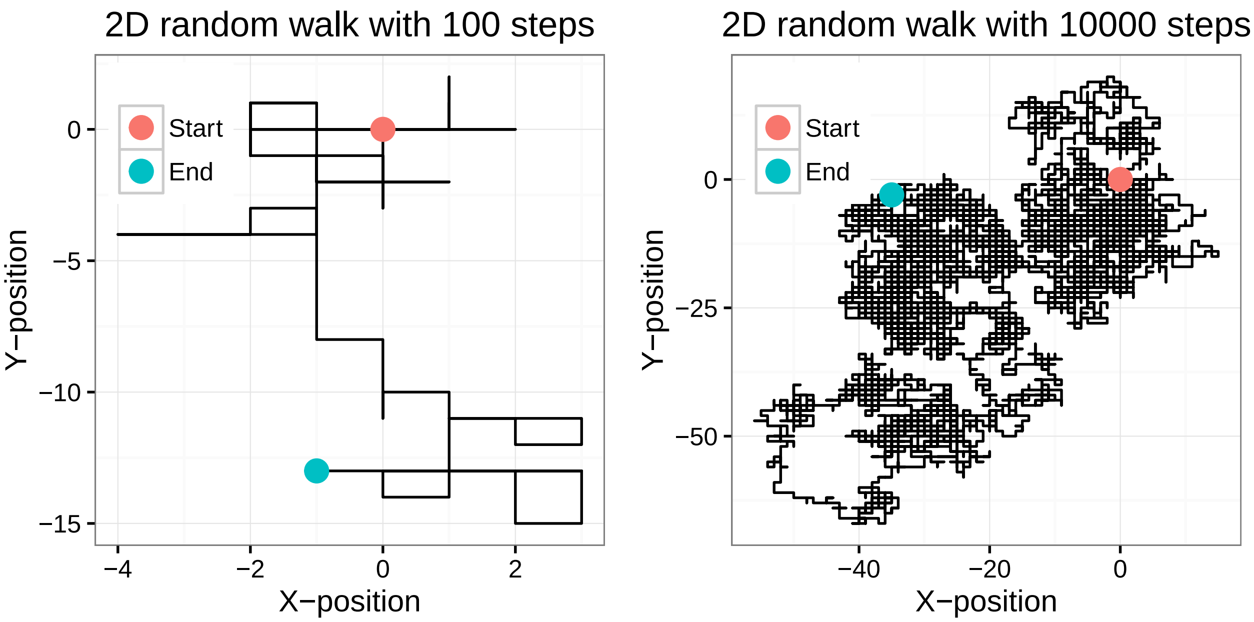
```
# This code simulates a gaussian white noise process
n = 1000                                    # process length
sigma2 = 1                                  # process variance
Xt = gen.gts(WN(sigma2 = sigma2), N = n)
plot(Xt)
```



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:

```r
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 class, we only consider one very specific form of random walk, namely the Gaussian random walk which can be defined as:

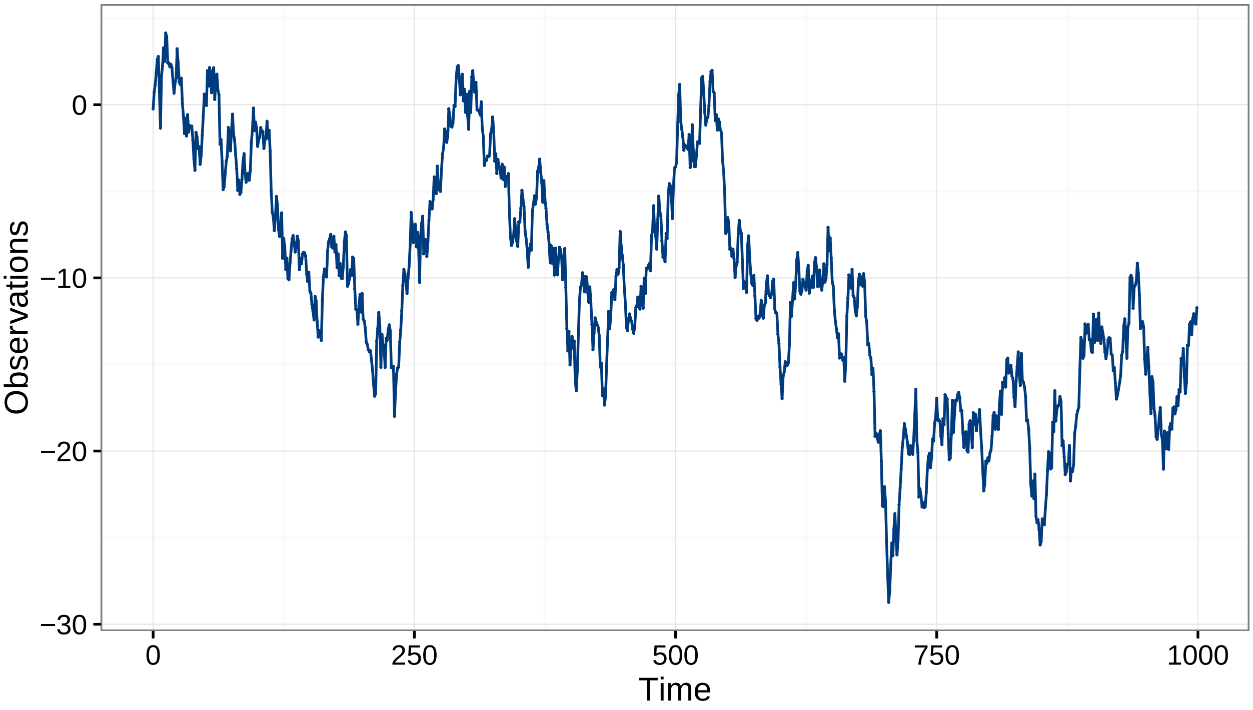
$$X_t = X_{t-1} + W_t,$$

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

```r
# 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: 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 μ, i.e.

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

which is of course equivalent to

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

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

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

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

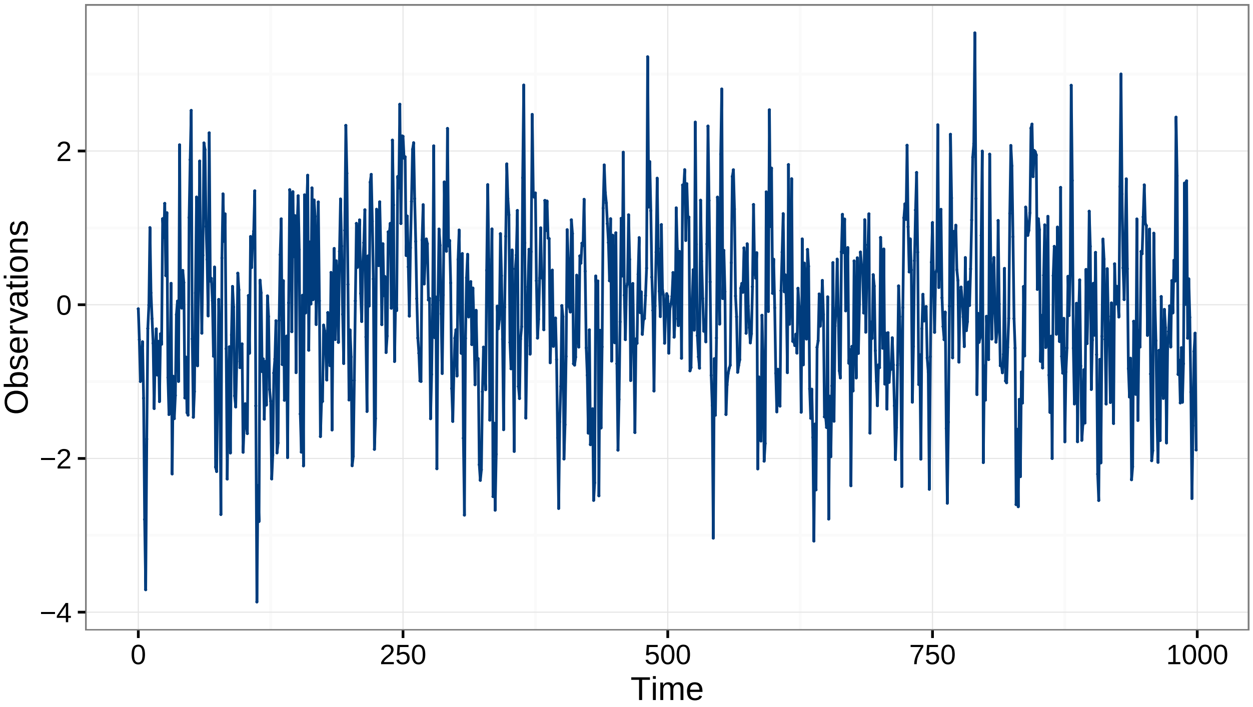
Under the assumption of infinite past (i.e. $t \in \mathbb{Z}$) and $|\phi| < 1$, we obtain

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

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)
```



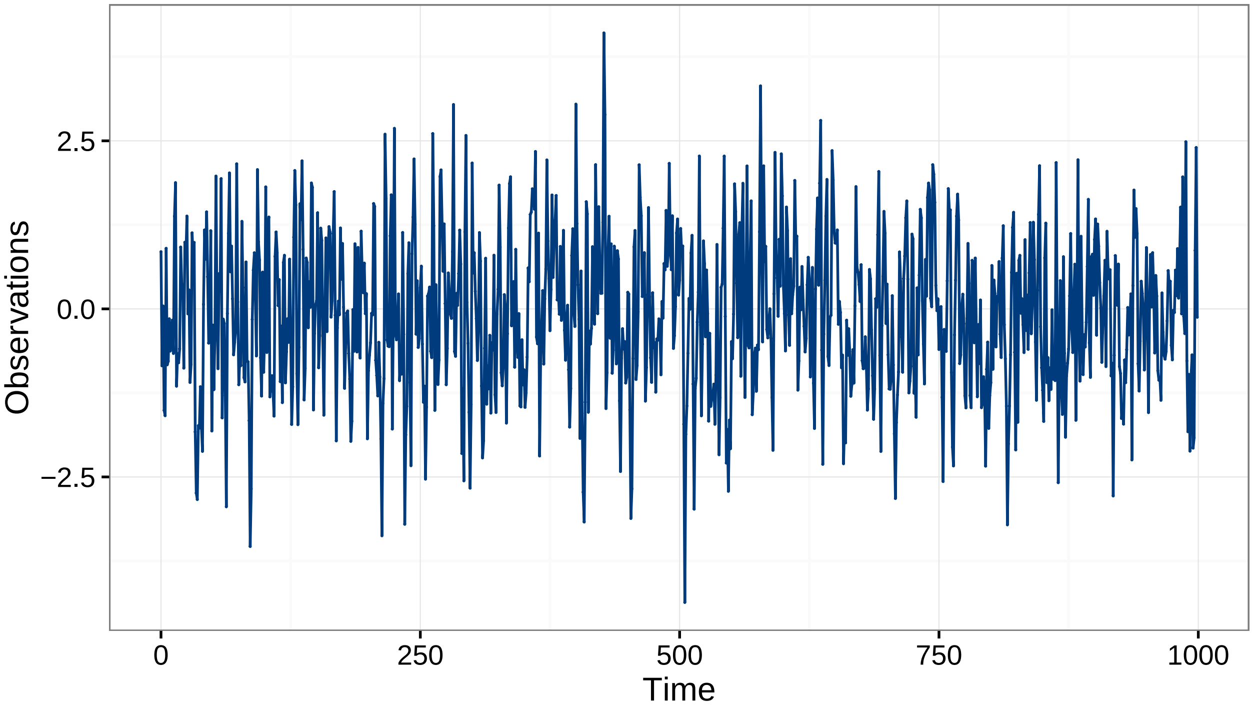
1.3.4 Moving Average Process of Order 1

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

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

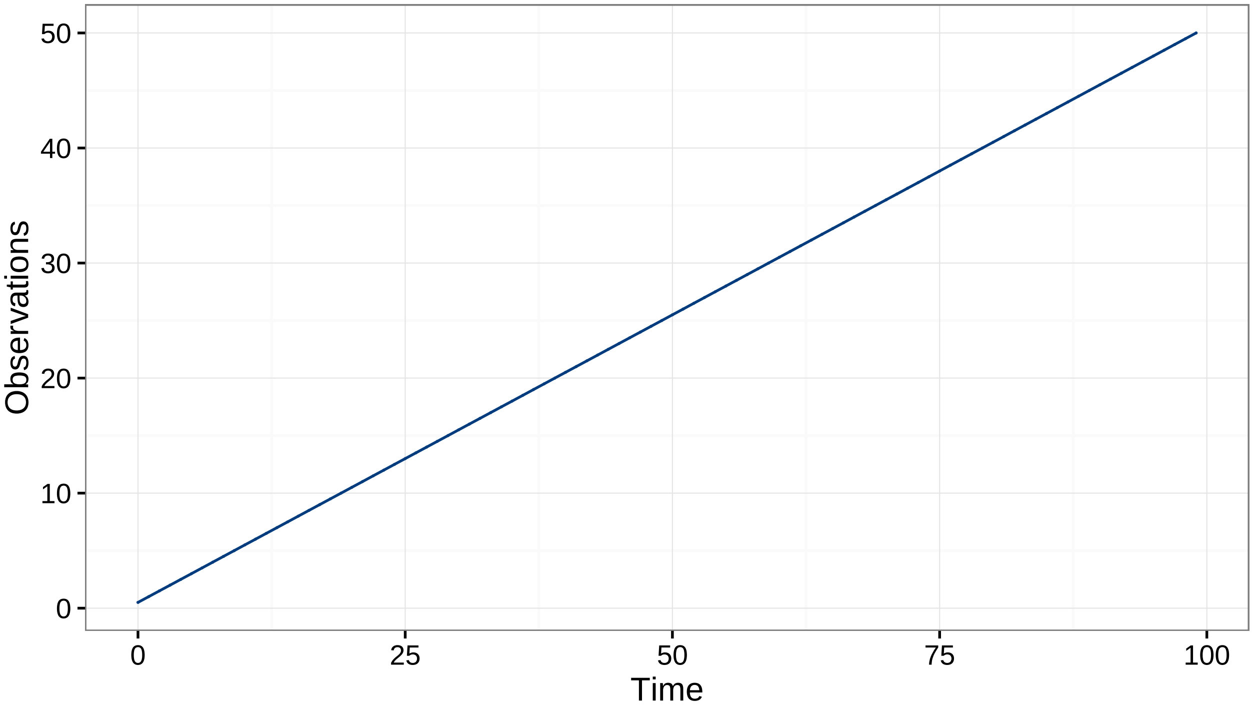
where ω is a constant and with the initial condition $X_0 = c$, an arbitrary constant (typically zero). This process can be expressed in a more familiar form as follows:

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

Therefore, a (linear) drift corresponds to a simple linear model with slope ω 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

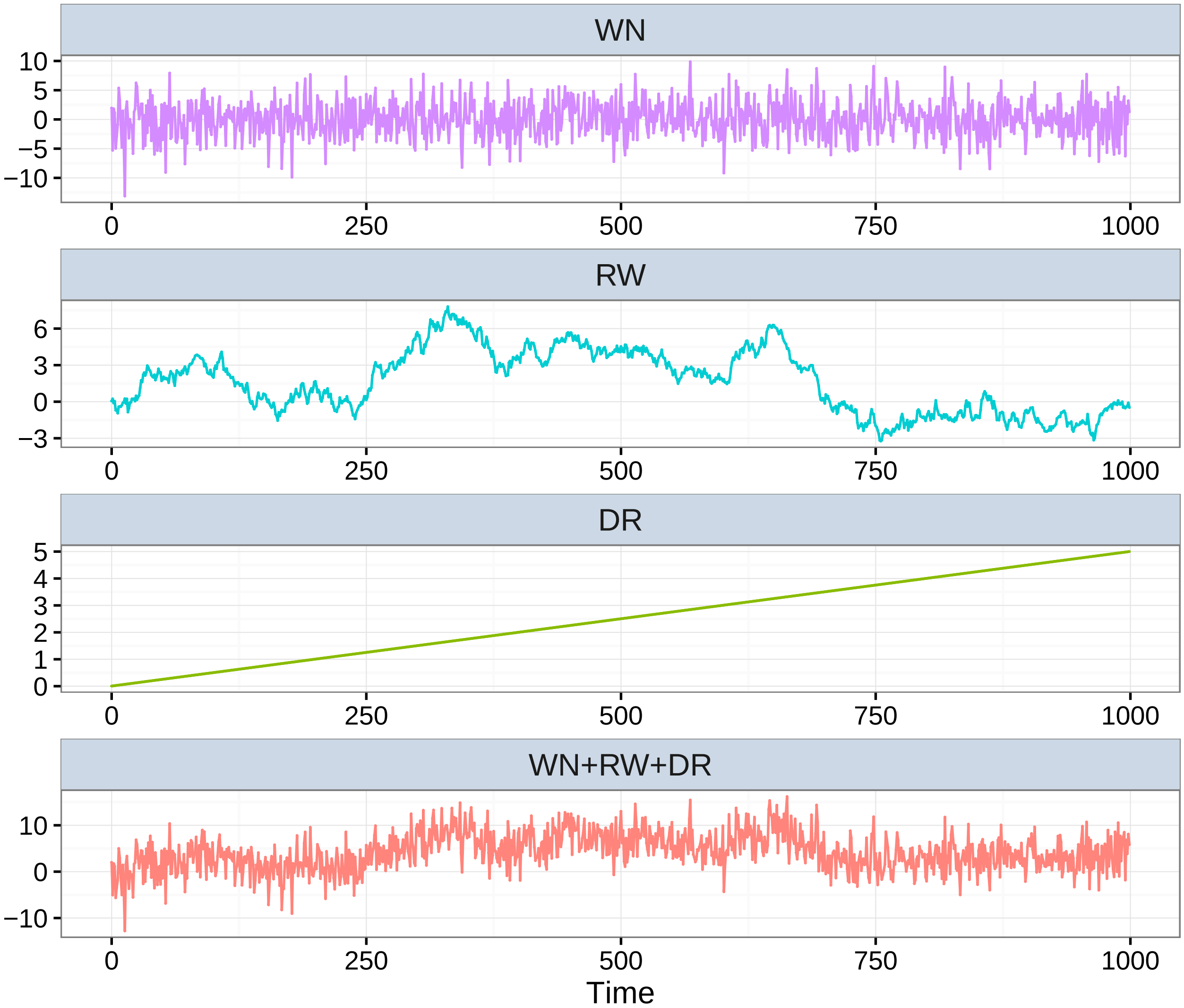
$$\begin{aligned} Y_t &= Y_{t-1} + W_t + \delta \\ X_t &= Y_t + Z_t, \end{aligned}$$

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:

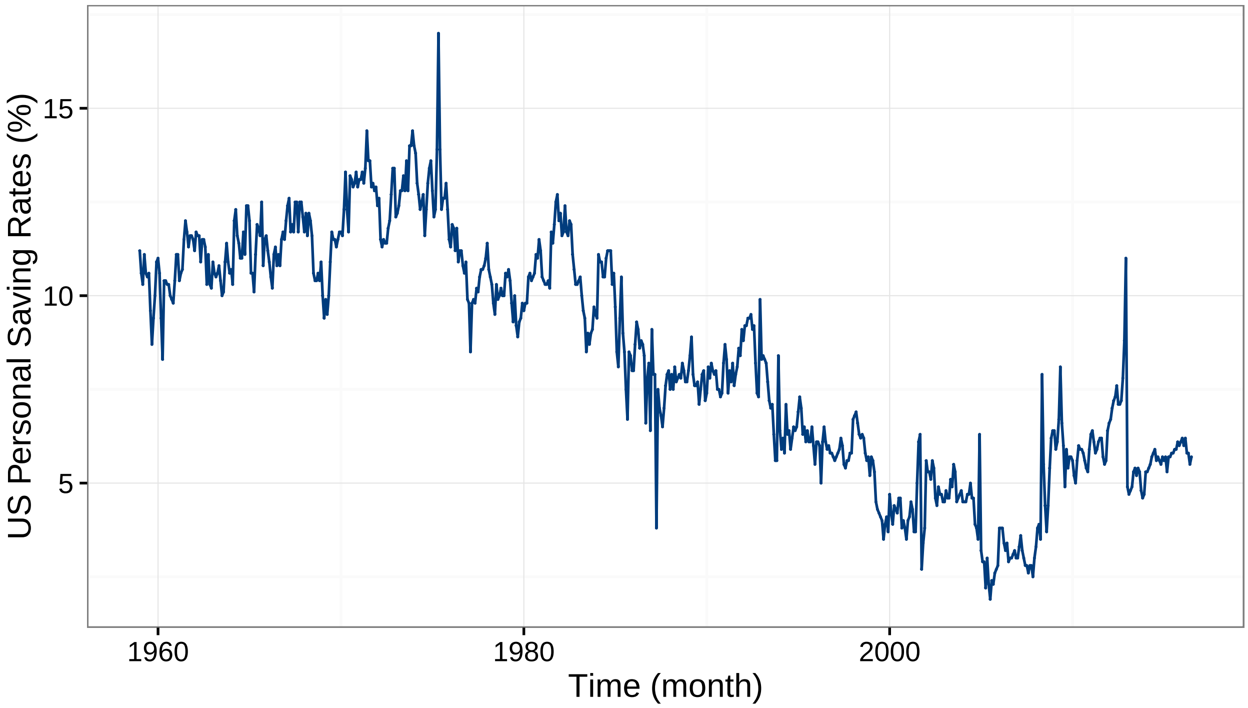
```
n = 1000                                # process length
delta = 0.005                           # delta parameter (drift)
sigma2 = 10                             # variance parameter (white noise)
gamma2 = 0.1                            # innovation variance (random walk)
model = WN(sigma2 = sigma2) + 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 Ω_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 $\text{corr}\left(X_t, X_{t+h}\right)$. 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 $\text{corr}\left(X_t, X_{t+h}\right)$ 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

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

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

2.1.1 Definitions

The *autocovariance function* of a series (X_t) is defined as

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

where the definition of covariance is given by:

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

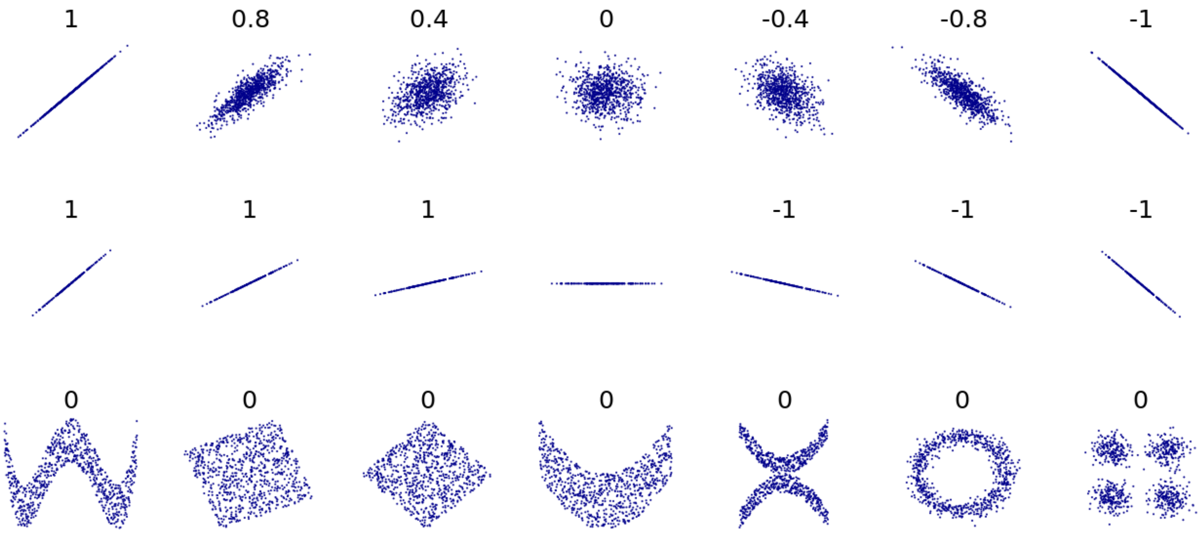


Figure 2.1: Different forms of dependence and their Pearson's r value

Similarly, the above expectations are defined to be:

$$\mathbb{E}\left[X_t\right] = \int_{-\infty}^{\infty} x \cdot f_t\left(x\right) dx,$$

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

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

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

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 $\gamma\left(t, t+h\right)$ instead of $\gamma_x\left(t, t+h\right)$. 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:

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

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

1. The autocovariance function is *symmetric*. That is, $\gamma\left(h\right) = \gamma\left(-h\right)$ since $\text{cov}\left(X_t, X_{t+h}\right) = \text{cov}\left(X_{t+h}, X_t\right)$.
2. The autocovariance function "contains" the variance of the process as $\text{var}\left(X_t\right) = \gamma\left(0\right)$.

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.

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

4. Just as any covariance, $\gamma\left(h\right)$ is "scale dependent" since $\gamma\left(h\right) \in \mathbb{R}$, or $-\infty \le \gamma\left(h\right) \le +\infty$. We therefore have:
   - if $|\gamma\left(h\right)|$ is "close" to zero, then X_t and X_{t+h} are "weakly" (linearly) dependent;
   - if $|\gamma\left(h\right)|$ 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. $\gamma\left(h\right)=0$ 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 $\gamma\left(h\right)=0$ 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

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

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

2.1.2 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.*

2.1.3 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

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

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

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

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

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

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

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

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, ρ_1 and ρ_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.

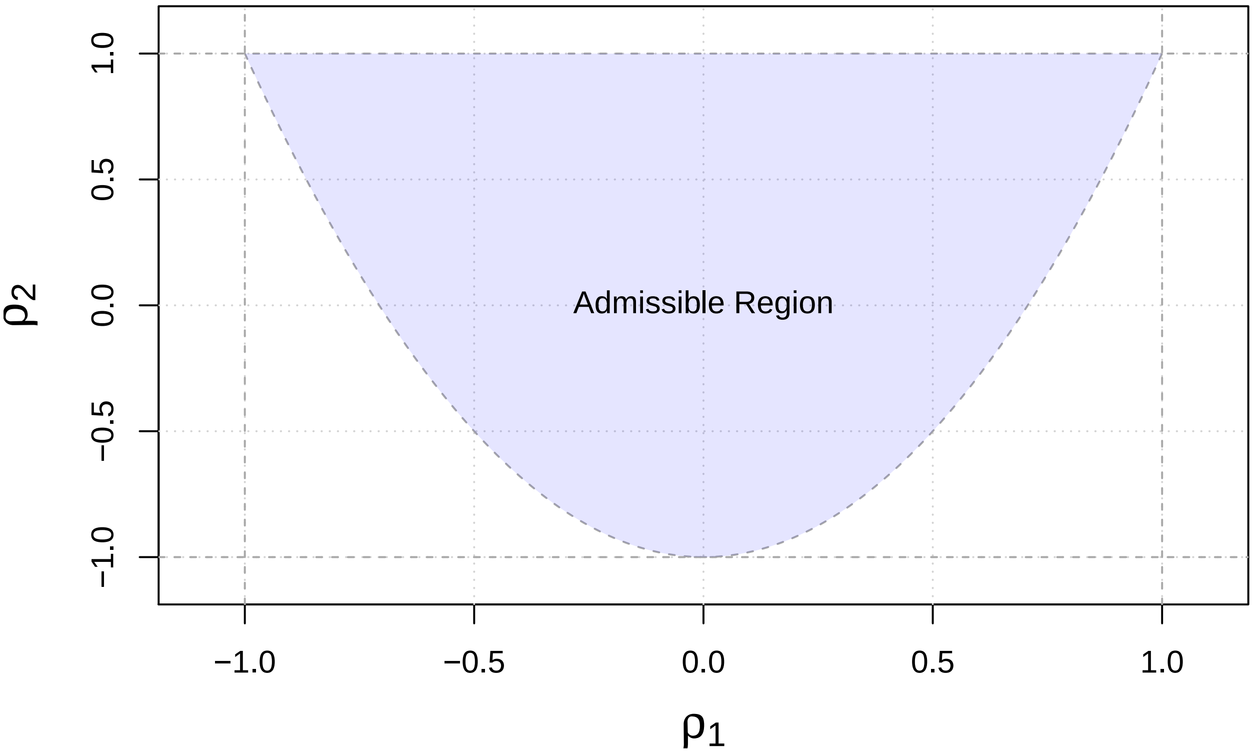


Figure 2.2: Admissible autocorrelation functions

2.2 Stationarity

2.2.1 Definitions

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

- 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.
- 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.

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*: an iid Cauchy process is strongly but not weakly stationary.
- Weak stationarity *does not imply* strong stationarity. *Example*: 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.

2.2.2 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 1.3.

Example: 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: 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:

```r
# 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: MA(1) Similarly to our previous examples, we attempt to verify the stationary properties for the MA(1) model defined in Section 1.3.4:

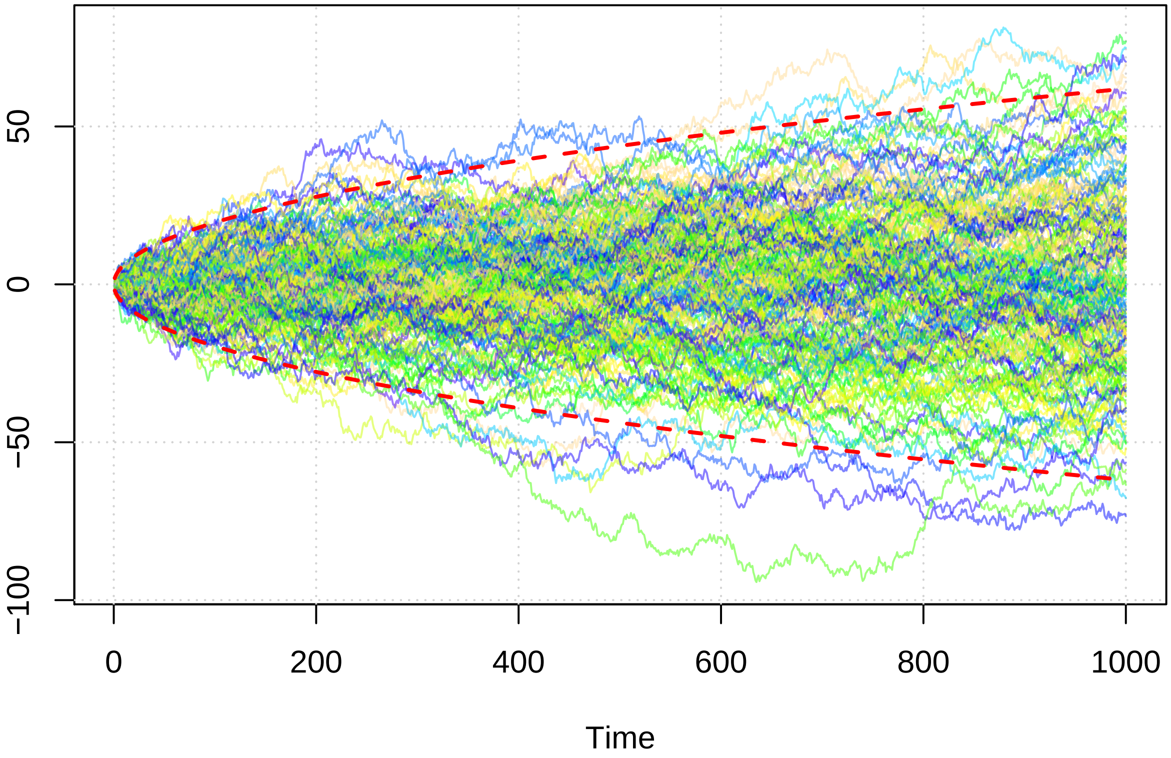


Figure 2.3: Two hundred simulated random walks.

1.
$$\mathbb{E}\left[X_t\right] = \mathbb{E}\left[\theta_1 W_{t-1} + W_t\right] = \theta_1 \mathbb{E}\left[W_{t-1}\right] + \mathbb{E}\left[W_t\right] = 0.$$

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

3. Regarding the autocovariance, we have

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

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

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

implying the following autocovariance function:

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

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

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

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

Example AR(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(∞). 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

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

   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

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

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

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

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

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

After fully simplifying, we obtain

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

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

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

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

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

The variance of such estimator can be derived as follows:

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

where $\boldsymbol{\Sigma}(k) \in \mathbb{R}^{n\times n}$ and its ith, 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

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

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: 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

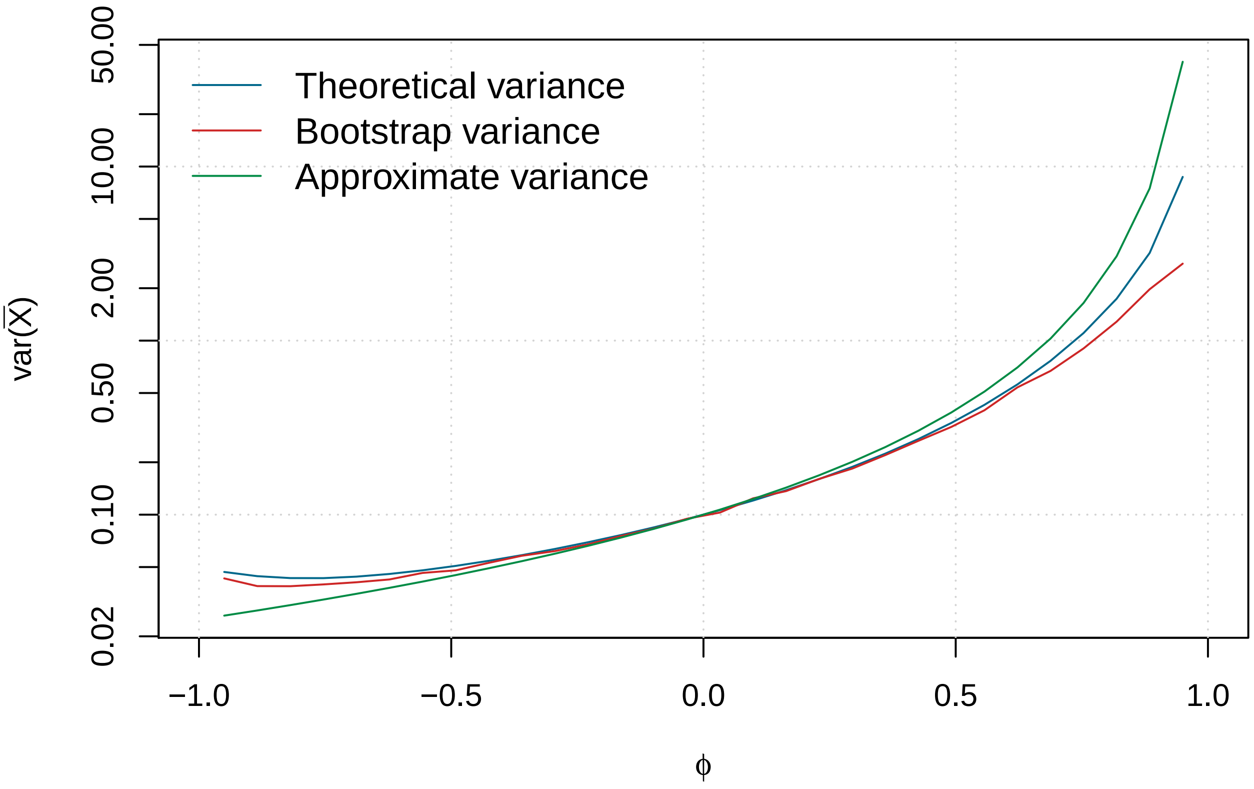
$$\hat{\sigma}_B^2 = \frac{1}{B-1}\sum_{i=1}^{B}\left(\bar{X}_i^* - \bar{X}^*\right)^2, \quad \text{where} \quad \bar{X}^* = \frac{1}{B}\sum_{i=1}^{B}\bar{X}_i^*,$$

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 ϕ going from -0.95 to 0.95:

```r
# 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 ϕ. 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 (2.2) while the asymptotic formula provides a reasonable approximate for ϕ 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:

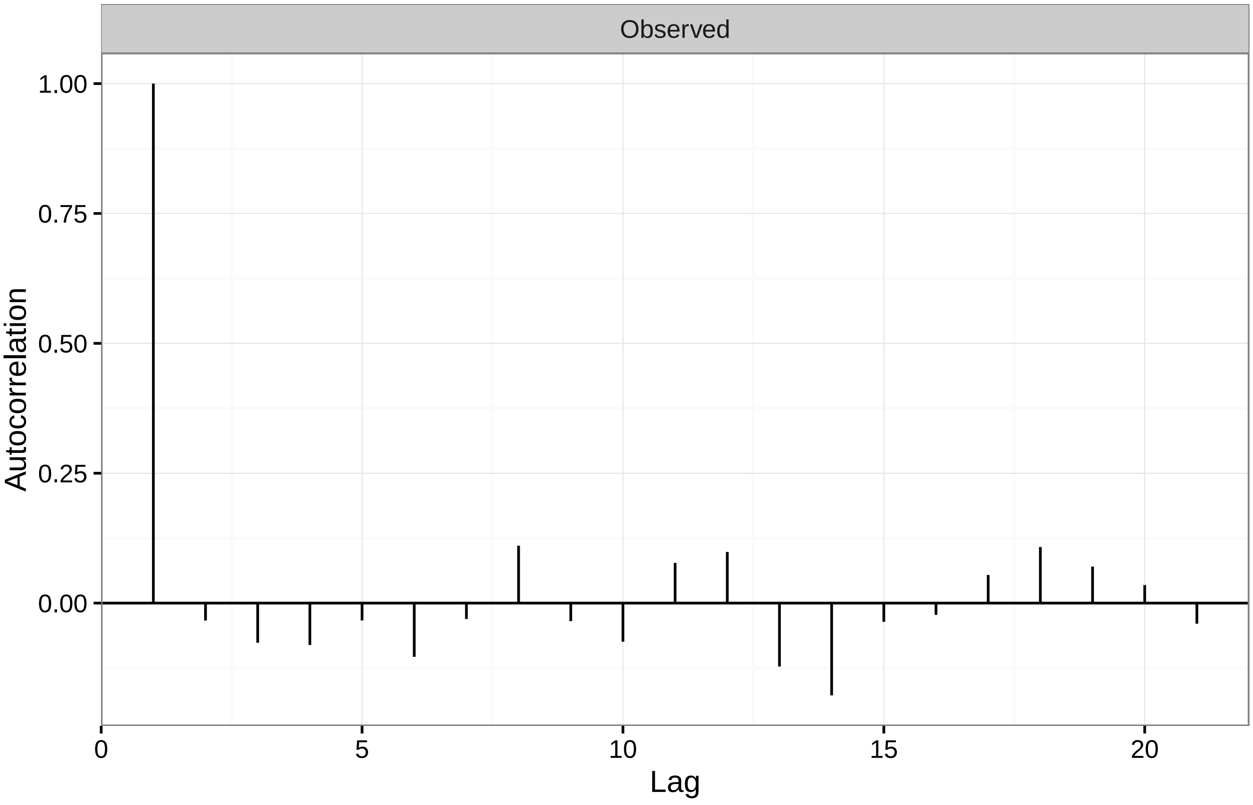
$$\hat{\gamma}(h) = \frac{1}{T}\sum_{t=1}^{T-h}\left(X_t - \bar{X}\right)\left(X_{t+h} - \bar{X}\right)$$

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:

```r
# 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)
```



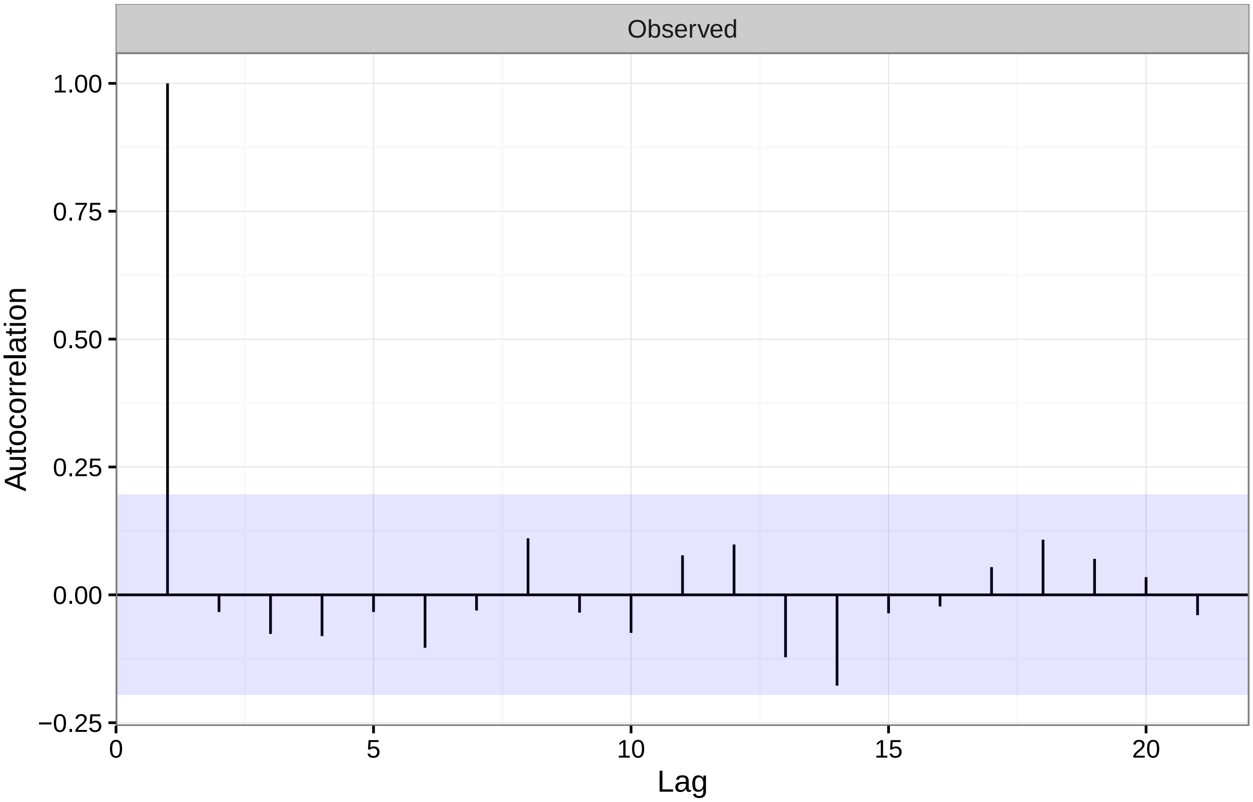
In this example, the true autocorrelation is equal to zero at any lag $h \neq 0$ but obviously the estimated autocorrelations are random variables and are not equal to their true values. It would therefore be useful to have some knowledge about the variability of the sample autocorrelations (under some 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: 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 **??**.

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: 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)
```

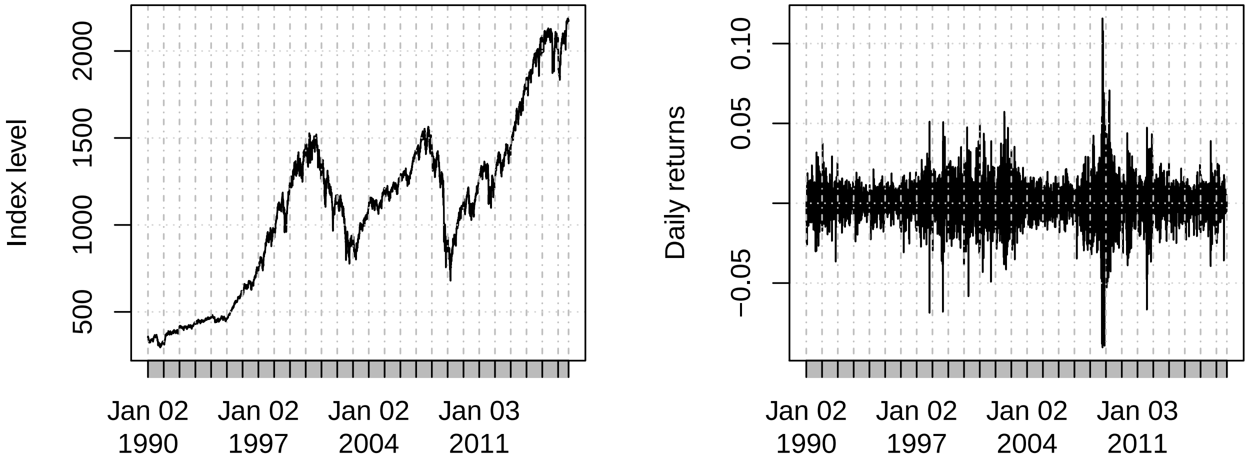
```
# Download S&P index
getSymbols("^GSPC", from="1990-01-01", to = Sys.Date())
```

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

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

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

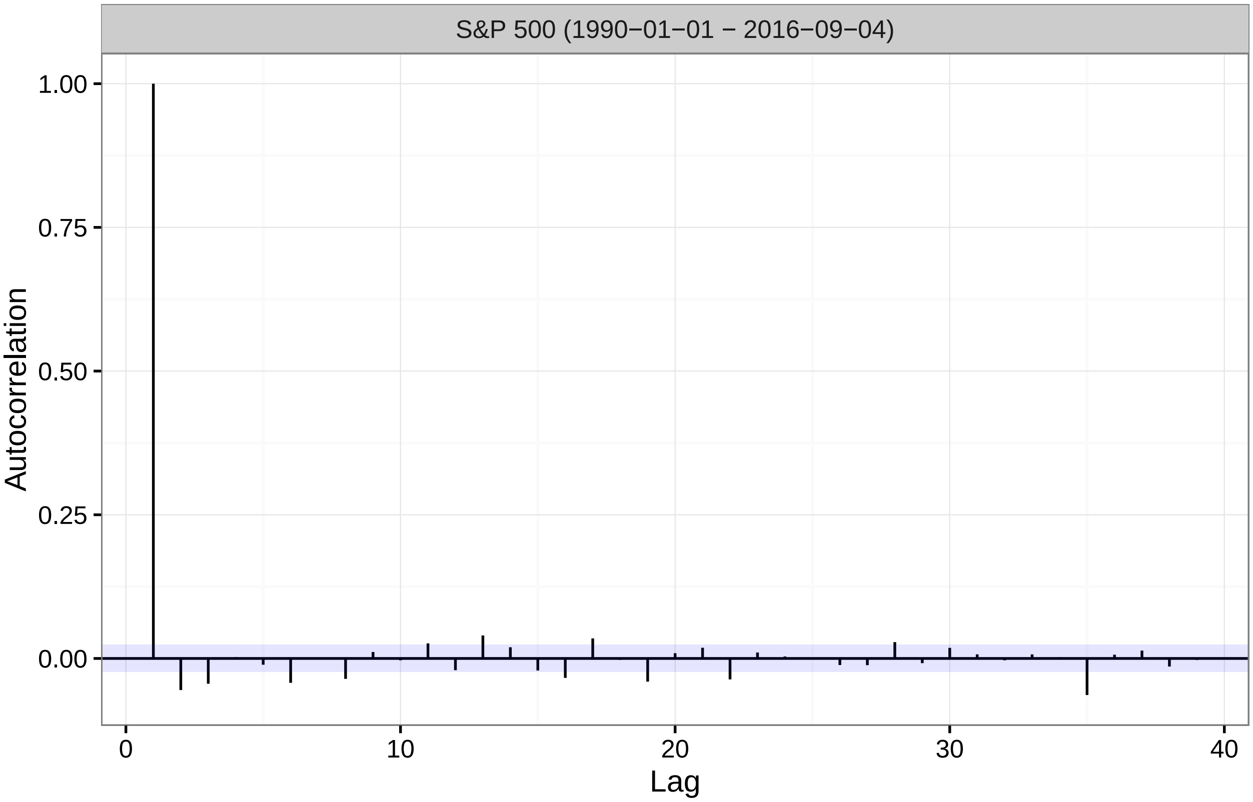
```
## Warning in plot.xts(GSPC, main = " ", ylab = "Index level"): only
## the univariate series will be plotted
```

```
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.

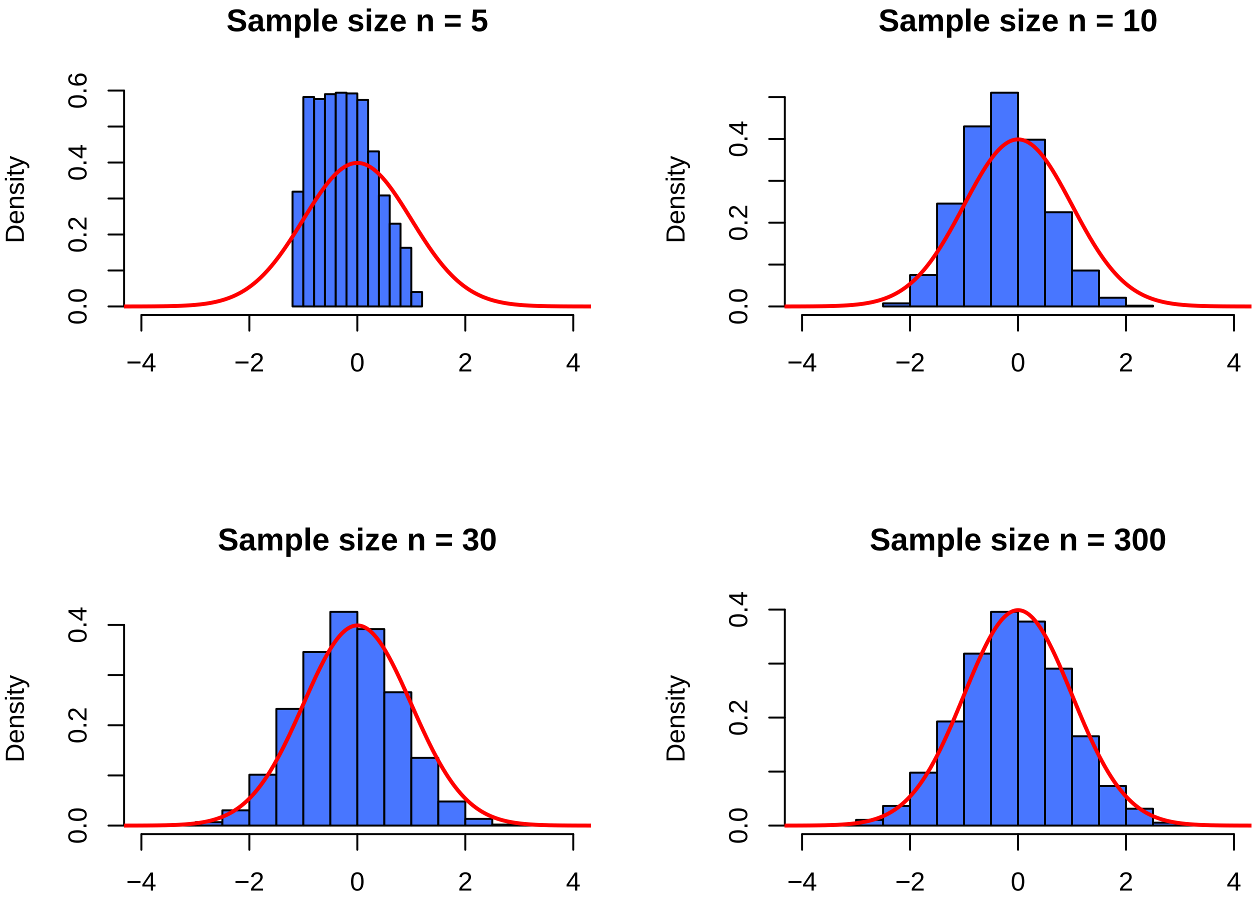
```
sp500 = na.omit(GSPC.ret)
names(sp500) = paste("S&P 500 (1990-01-01 - ",Sys.Date(),")", sep = "")
plot(ACF(sp500))
```



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

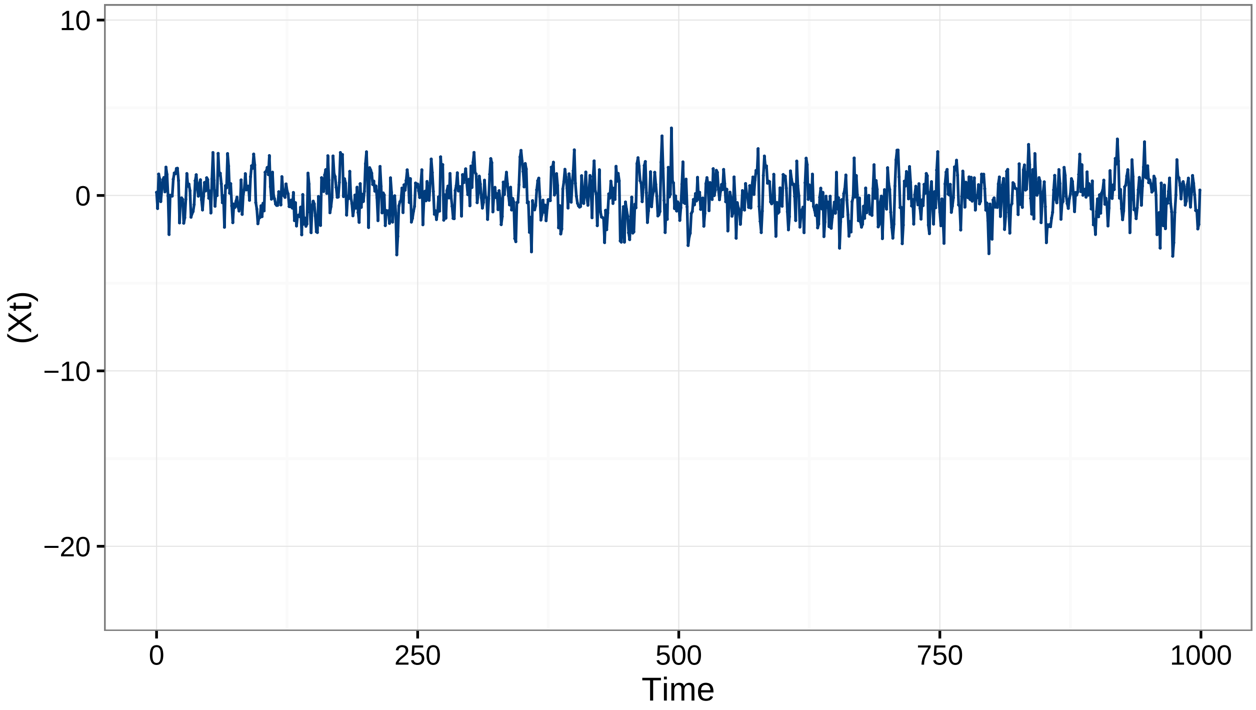
The data generating process delivers a theoretical autocorrelation (autocovariance) function that, as explained in the previous section, can then be estimated through the sample autocorrelation (autocovariance) functions. However, in practice, the sample is often issued from a data generating process 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.

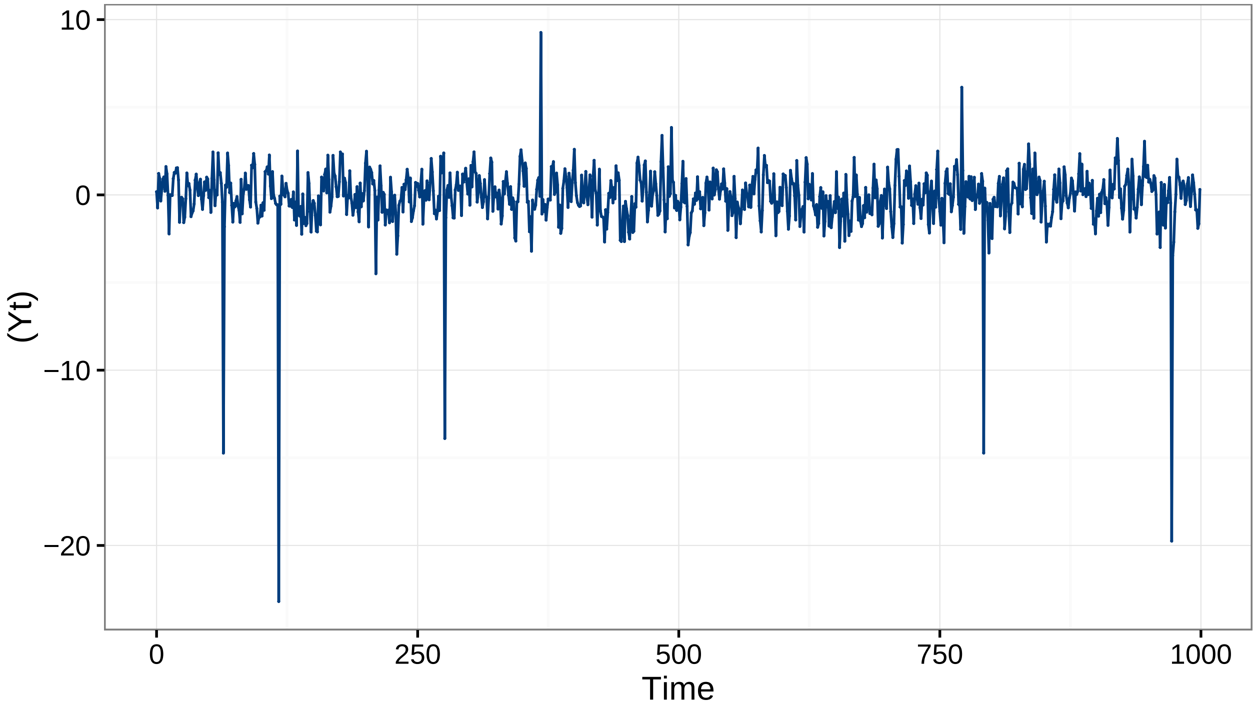
The fact that the sample can suffer from outliers implies that the standard estimation of the autocorrelation (autocovariance) functions through the sample functions could be highly biased. The standard estimators presented in the previous section are therefore not "robust" and can behave badly when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

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

when ε 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 $\alpha = 0.05$.

```r
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)
```

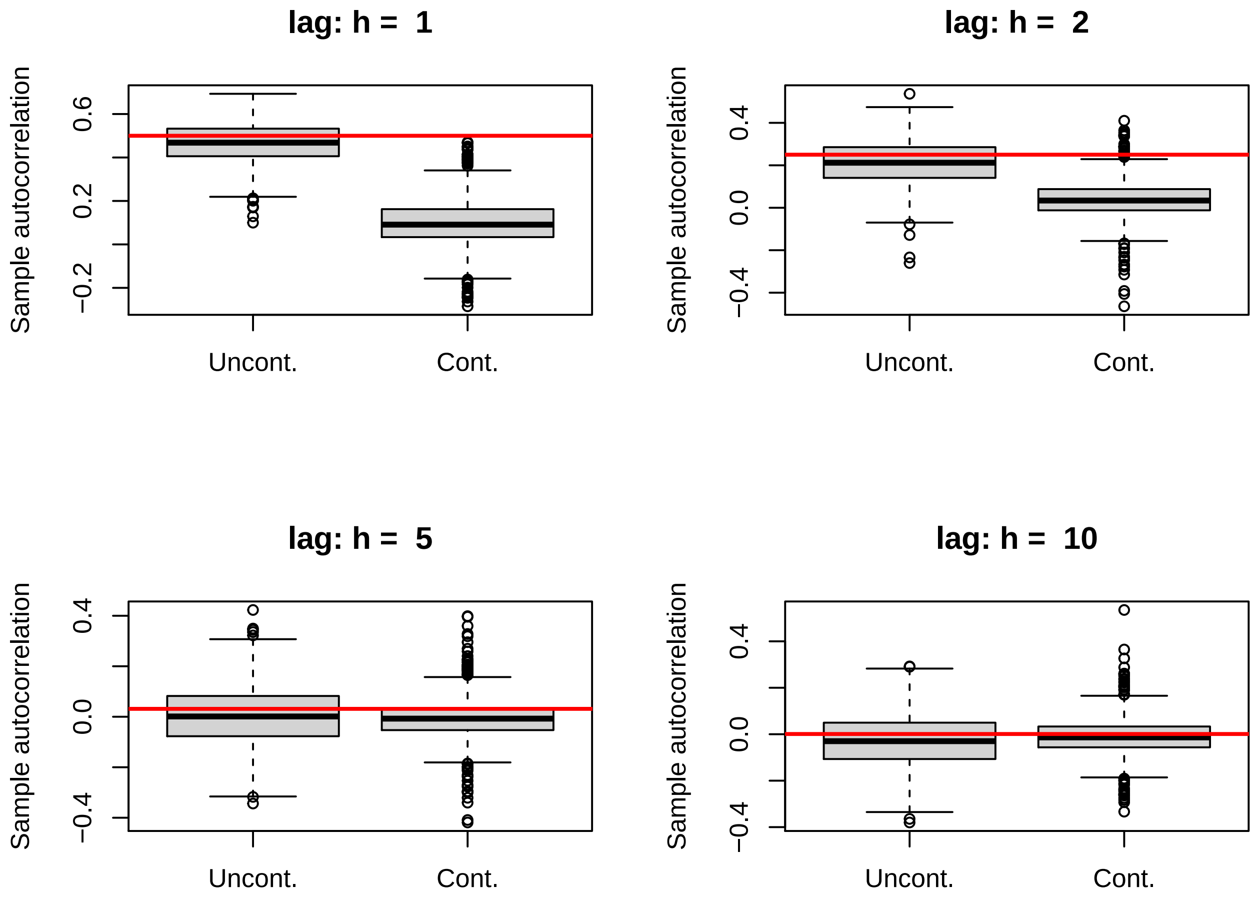




Next, we consider a simulated example to highlight how the performance of the "classical" autocorrelation can deteriorate if the sample is contaminated (i.e. what is the impact of using (Y_t) instead of (X_t), the "uncontaminated" process). In this simulation, we used the setting presented above and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

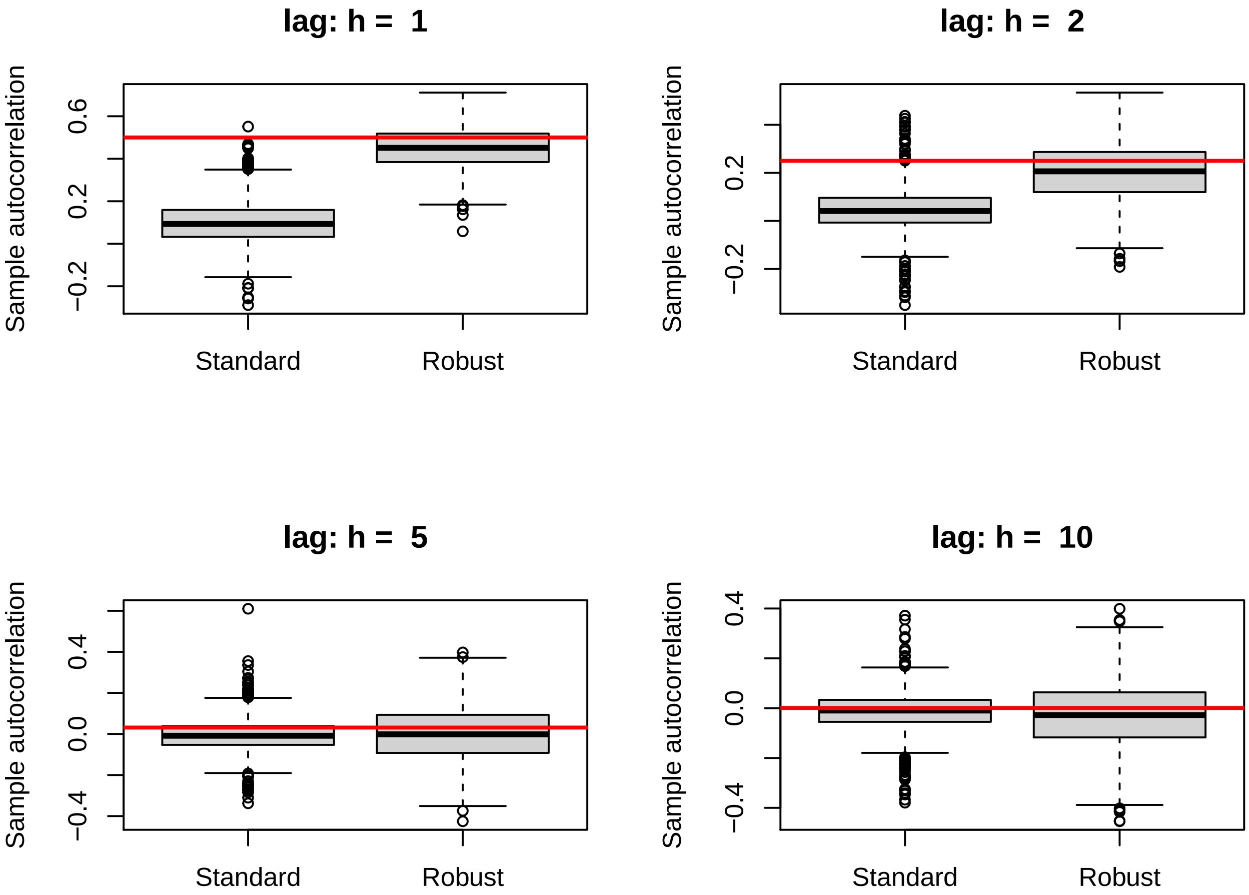


The boxplots in each figure show how the standard autocorrelation estimator is centered around the true value (red line) when the sample is not contaminated (left boxplot) while it is considerably biased when the sample is contaminated (right boxplot), especially at the smallest lags. Indeed, it can 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.

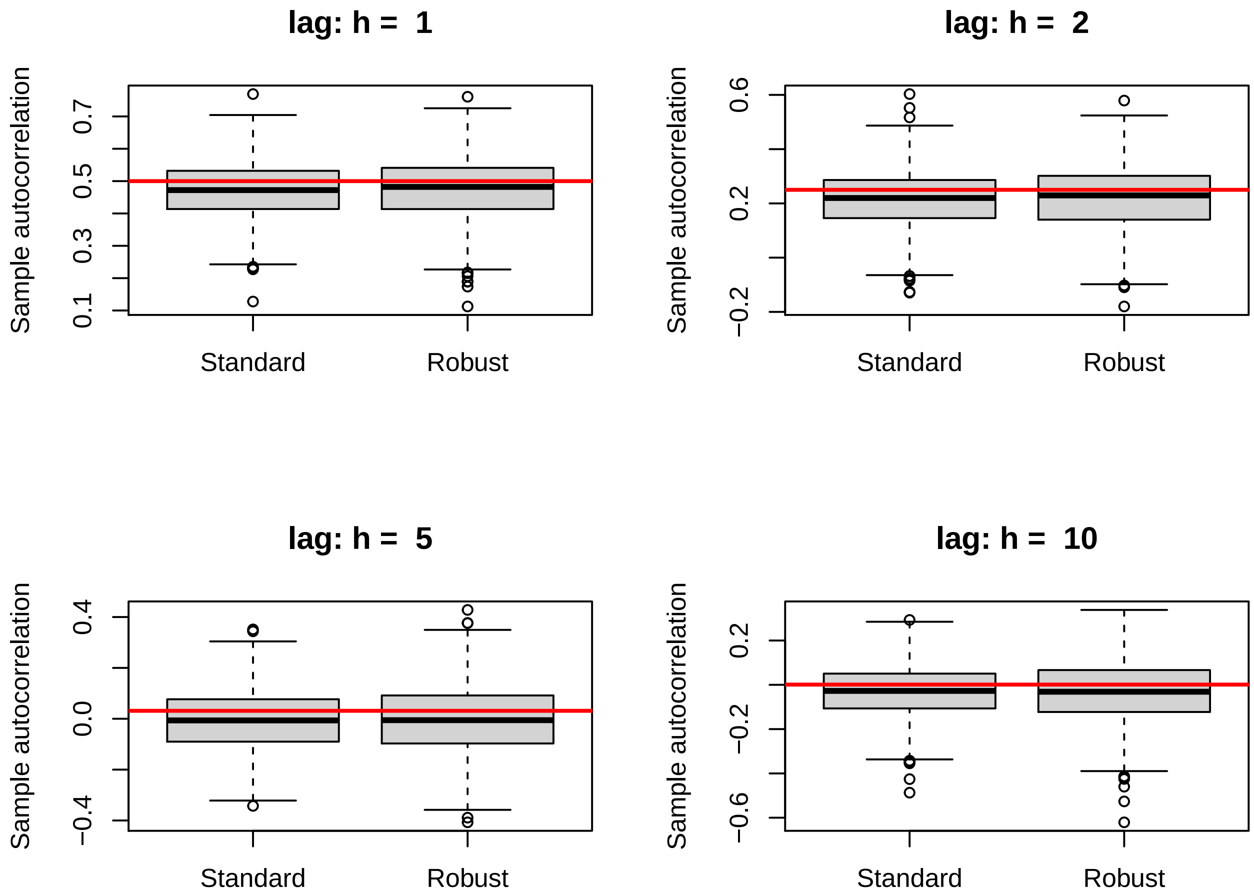
In order to limit this problematic, different robust estimators exist for time series problems allowing to reduce the impact of contamination on the estimation procedure. Among these estimators there are a few that estimate the autocorrelation (autocovariance) functions in a robust manner. One of these 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)
}
```



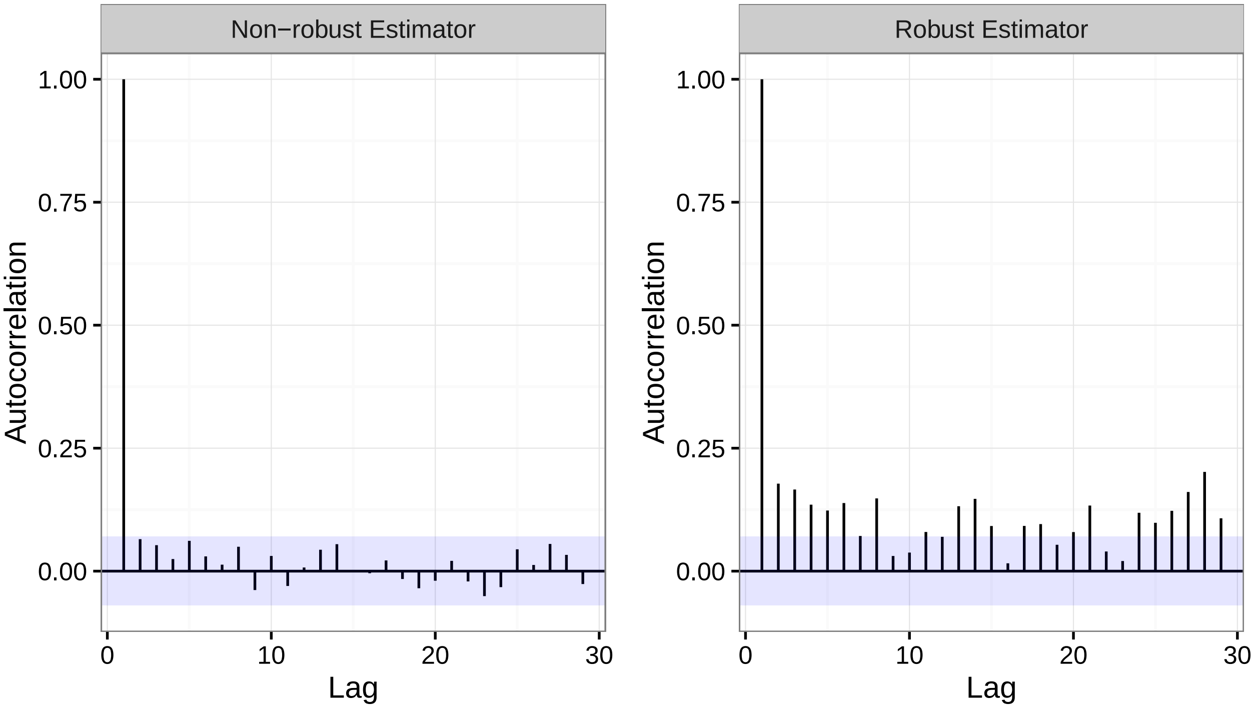
The robust estimator remains close to the true value represented by the red line in the boxplots as opposed to the standard estimator. It can also be observed that to reduce the bias induced by contamination in the sample, robust estimators pay a certain price in terms of efficiency as highlighted by 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 α 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:

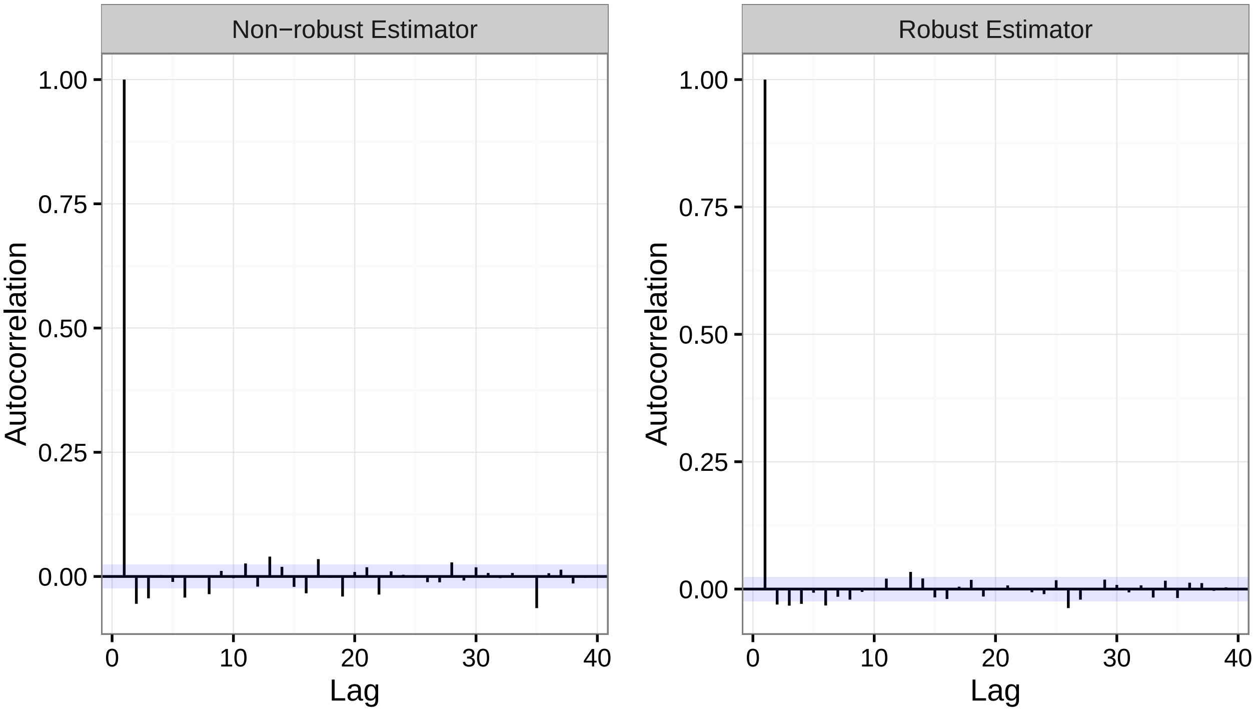
```r
# Load packages
library(gmwm)
library(gridExtra)
library(robcor)

# 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.

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 σ^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.

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

For this equation, it is clear that $\hat{\gamma}\left(0\right)$ and $\hat{\gamma}\left(h\right)$ (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 $[\hat{\gamma}\left(h\right) \quad \hat{\gamma}\left(h\right)]^T$. 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

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

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

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

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

Lemma A1: Let

$$X_t = \mu + \sum_{j=-\infty}^{\infty} \psi_j W_{t-j}, \tag{A.1}$$

where (W_t) is a strong white process with variance σ^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

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

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

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

Next, we consider each term of the above equation. For the first term, since $(X_t-\mu)^2 = \left(\sum_{j=-\infty}^{\infty}\psi_j W_{t-j}\right)^2$, and $\mathbb{E}[W_iW_j] \neq 0$ 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

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

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

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

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

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

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

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

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

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

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

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

only when $i = j$.

Therefore, we obtain

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

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,

$$\begin{aligned} \sqrt{n}\left\{\begin{bmatrix}\hat{\gamma}(0)\\ \hat{\gamma}(h)\end{bmatrix} - \begin{bmatrix}\sigma^2\\ 0\end{bmatrix}\right\} &= \sqrt{n}\left\{\begin{bmatrix}\tilde{\gamma}(0)\\ \tilde{\gamma}(h)\end{bmatrix} - \begin{bmatrix}\sigma^2\\ 0\end{bmatrix}\right\} + \underbrace{\sqrt{n}\left\{\begin{bmatrix}\hat{\gamma}(0)\\ \hat{\gamma}(h)\end{bmatrix} - \begin{bmatrix}\tilde{\gamma}(0)\\ \tilde{\gamma}(h)\end{bmatrix}\right\}}_{\overset{p}{\to}0} \\ &\overset{\mathcal{D}}{\to} \mathcal{N}\left(0, \begin{bmatrix}2\sigma^4 & 0\\ 0 & \sigma^4\end{bmatrix}\right). \end{aligned}$$

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 $h \neq 0$

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

where

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

Thus, we have

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

which concludes the proof the Theorem 1. ■

Bibliography