LECTURE NOTES
FINANCIAL TIME SERIES
WI3411TU[1]
Version 1.12

F.H. van der Meulen

[1]Please contact me at f.h.vandermeulen@tudelft.nl to point out mistakes.

Contents

Preface v

Chapter 1. Introduction 1
1. Returns 5

Chapter 2. Stationary time-series 7
1. Recap covariance and correlation 7
2. Stationarity 8
3. Some examples of time-series models 9
4. Best linear prediction for stationary time-series 15
5. The partial autocorrelation function 18
6. Exercises from old exams 19

Chapter 3. From data to time-series model 21
1. Transforming time-series to stationarity 21
2. Estimating the autocorrelation function 23
3. Estimating the partial autocorrelation function 27
4. Asymptotics for the PACF 27
5. Some simulations in R for the PACF 28
6. Some methods to check the noise sequence 29
7. Checking for normality using a probability plot 31
8. Exercises from old exams 32

Chapter 4. ARMA modeling 35
1. Moving Average models 35
2. Autoregressive models 38
3. Residual analysis and diagnostic plots 42
4. An example 42
5. ARMA models 47
6. Exercises from old exams 49

Chapter 5. Conditional expectation and best prediction 51
1. Conditional expectation in the discrete case 51
2. Conditional expectation in the general case 52
3. Best Predictor 53
4. Exercises from old exams 56

Chapter 6. GARCH modeling 59

1. Introduction 59
2. ARCH(1)-model 64
3. Extending the model: GARCH models 75
4. Combining ARMA and GARCH models 76
5. Exercises from old exams 78

Chapter 7. Parameter estimation 87
1. Maximum likelihood estimation 87
2. Least squares estimation 88

Chapter 8. Appendix 91
1. Setting up your R working environment 91
2. List of R commands 92
3. The multivariate normal distribution 93

Chapter 9. Hints to exercise solutions 95

Bibliography 101

Preface

These are lecture notes for the course "financial time-series analysis" (WI-3411TU) at Delft University of Technology. There are quite some good books on time-series analysis available. At an introductory level there are the books by Brockwell and Davis [2002] and Shumway and Stoffer [2011]. Somewhat more directed to financial time-series is the book Tsay [2010]. A more advanced book is Durbin and Koopman [2012], which uses the state-space modelling approach to time-series analysis. The latter is a very powerful approach, including Kalman filtering methods and multivariate time-series analysis.

Exercises are scattered throughout the text. Many are taken from Brockwell and Davis [2002]. At the end of most chapters there are exercises from old assignments and exams. These are representative for the kind of questions that you can expect at the exam. Throughout the text I added `R`-commands and corresponding output to help you carrying out time-series analysis yourself on real data.

Delft, August 2021
Frank van der Meulen

CHAPTER 1

Introduction

Time-series analysis is concerned with summarising and drawing conclusions from data that are obtained sequentially. In this course, we will assume the data are obtained at times $t = 0, 1, 2, \ldots$. The value that is measured at time t will be denoted by x_t and the whole sequence of measurements will be denoted by $\{x_t\}$. If we explicitly wish to stress that the data are obtained at times $t = 0, 1, \ldots, n$, then we write $\{x_t\}_{t=0}^n$. When $x_t \in \mathbb{R}^d$ with $d > 1$ we speak of a multivariate time series $\{x_t\}$. Here, we will only be concerned with univariate time-series, for which $d = 1$.

DEFINITION 1.1. A *time-series* is a collection of observations $x_t \in \mathbb{R}^d$, recorded at times $t = 0, 1, \ldots, n$.

Strictly speaking, the nomenclature "time-series" is a bit awkward, as it suggests a series, which in mathematics is like $\sum_{i \geq 0} a_i$, whereas we actually mean time-sequence. The terminology is however well-established and henceforth we'll stick to that. The value x_t can for example be the value of a stock at day t (assuming that $t = 0$ corresponds to the first date at which we record its value). Other examples are interest rates, temperature, humidity, voltage, sea-level etc. Clearly, time-series analysis has its roots in many fields of applications, explaining the wide variety of tools that have been developed. Here, we are specifically interested in financial time-series. Those of who had a course in signal processing for example, will discover overlap with the theory covered in this course. Needless, to say, these lecture notes only give an introduction to some key concepts of time-series analysis, the goal being to understand the application of ARMA- and GARCH-models to financial time-series. The exposition is at many places non rigorous (meaning that the particular statement is essentially correct, but the mathematical formulation can be sharpened), due to time-constraints on the course and foremost the assumed background knowledge.

Important objectives of time-series analysis include:

- providing a compact description / summary of the data;
- identifying global trends and or seasonal components;
- predicting future values of the series;
- separation (or filtering) of noise from signal.

All of these in some sense require the identification of structure in the data. Structure can be simple, if for example it is seen that $t \mapsto x_t$ is linearly

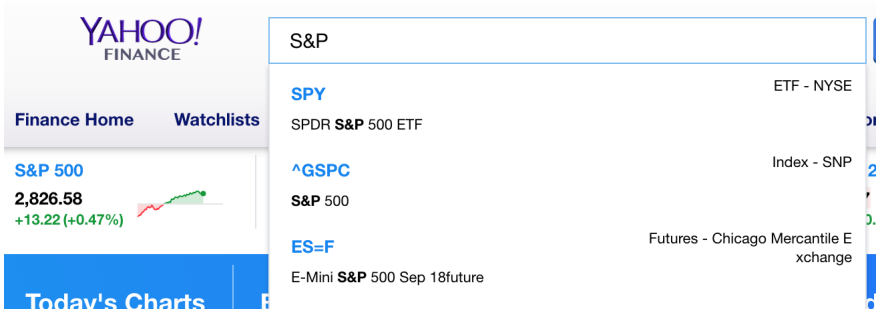

FIGURE 1. Identifying the symbol for S&P 500.

increasing. For financial time-series however, this structure is often much more hidden. Let's illustrate some of the problems using an example.
We consider prices of the S&P index over the period starting at January 1, 2007 and ending at June 19, 2018. Here, the time-unit is a day, and time $t = 0$ corresponds to January 1, 2007. These data can be downloaded into `R` directly from `finance.yahoo.com`. To do this, first of all, you need to set up your `R` environment by installing a couple of packages. This is explained in Section 1 of the appendix. Next, run the following code to get the data:

```
getSymbols("^gspc",src='yahoo')
```

This makes an object containing the data, called `GSPC`. The first argument of the function `getSymbols` is a string that is the identifier of the data on yahoo. This can be extracted from the website by simply searching for S&P, as in the screenshot in Figure 1.
The data can easily be subsetted, for example if you only wish to have the data from June 2007 up to January 12 2008, you get these from

```
GSPC['2007-06::2008-01-12']
```

A first step in data-analysis always consists of numerically and graphically summarising the data. So let's have a closer look at the data we downloaded: using `head` we display the first 6 lines of the downloaded data.

```
> head(GSPC)
           GSPC.Open GSPC.High GSPC.Low GSPC.Close GSPC.Volume
    GSPC.Adjusted
2007-01-03  1418.03   1429.42  1407.86    1416.60 3429160000
    1416.60
2007-01-04  1416.60   1421.84  1408.43    1418.34 3004460000
    1418.34
2007-01-05  1418.34   1418.34  1405.75    1409.71 2919400000
    1409.71
```

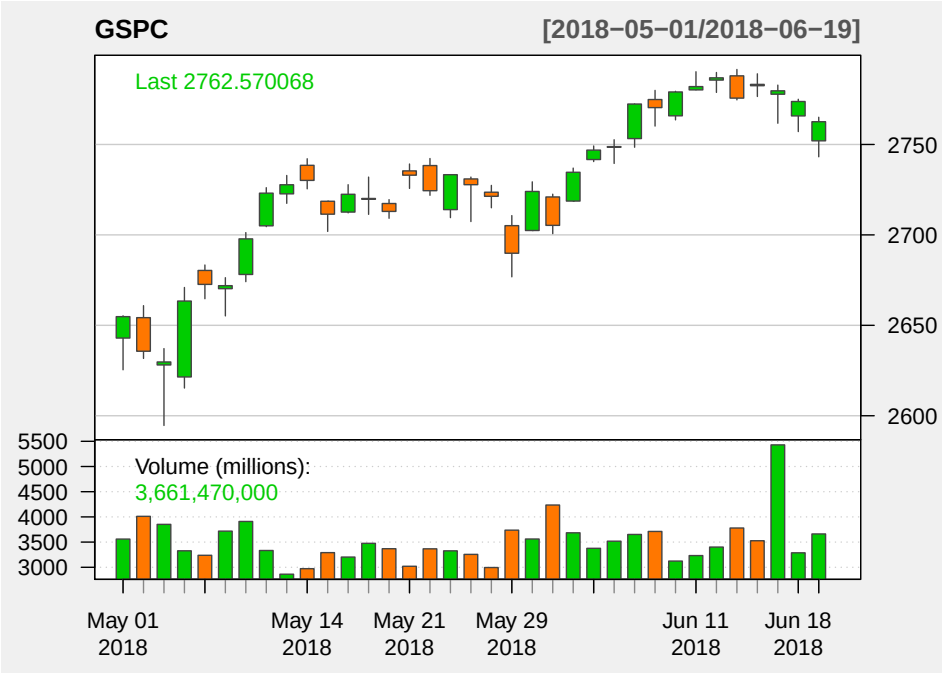

FIGURE 2. OHLC-chart for the S&P index over the past two months (today = June 19, 2018).

```
2007-01-08  1409.26  1414.98 1403.97   1412.84 2763340000
    1412.84
2007-01-09  1412.84  1415.61 1405.42   1412.11 3038380000
    1412.11
2007-01-10  1408.70  1415.99 1405.32   1414.85 2764660000
    1414.85
```

So we have downloaded the opening and closing prices, the highest- and lowest value during the day, the volume, and the adjusted closing prices. These are sometimes conveniently summarised using an OHLC (Open-High-Low-Closing) chart. For the data of the past 2 months, this chart can be obtained using the command.

```
chartSeries(GSPC,subset='last 2 months',theme=chartTheme('white'))
```

The resulting figure is depicted in Figure 2. In the bottom panel the trading volumes are depicted. On the top panel, for each day there is a "box including line segment". The latter gives the lowest- and highest value during that day, together with the opening-and closing prices. To identify the opening and closing prices, note that

- if the box is green, then the closing price was higher than the opening price;

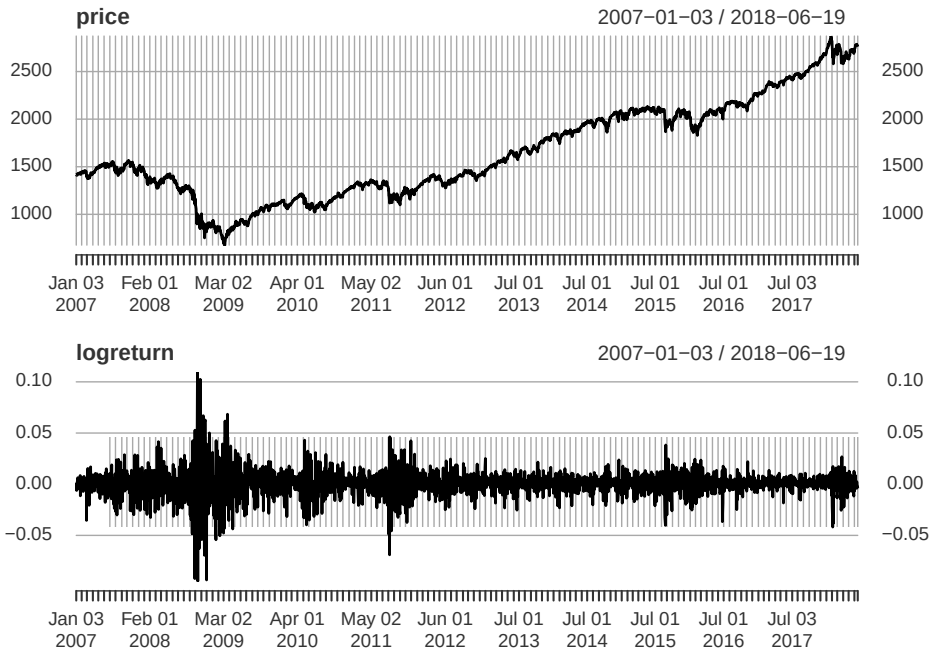

FIGURE 3. Top: adjusted closing prices of S&P index from January 1, 2007 until June 19. Bottom: corresponding log-returns.

- if on the other hand the opening price was higher than the closing price, the box is in orange.

In the top panel of Figure 3 the adjusted closing are shown of all data (not just the data of the past two months). There are clear trends in the time-series, but these are not easily captured by assuming a low order polynomial for example. The bottom panel shows log-returns, which we will define shortly. This figure was obtained with the code.

```
price <- Cl(GSPC) # get closing price
logreturn <- diff(log(price)) # convert the prices into returns

par(mfrow=c(2,1)) # set up graphics device to have two figures on 2
    rows and 1 column
plot(price)
plot(logreturn)
```

Hence, it is not too hard to download financial time-series and obtain some useful visualisations.

Exercise 1.1 Try this out yourself for some other financial indices. The identifier you can find on the yahoo site. For example Apple Inc. on NasdaqGS has `AAPL` as identifier.

1. Returns

Instead of the raw time-series x_t, one often considers derived quantities, which are summarised in the following definition.

DEFINITION 1.2. Suppose x_t is the value (price) of a financial quantity at time t. The *simple gross return* is defined by x_t/x_{t-1}. The *simple return* r_t is defined by

$$r_t = \frac{x_t - x_{t-1}}{x_{t-1}},$$

this is the percentage profit (or loss). The *log-return* is define by

$$\ell_t = \log \frac{x_t}{x_{t-1}}.$$

If $r_t \approx 0$, then $\ell_t \approx r_t$ as can be seen from

$$\ell_t = \log\left(1 + \frac{x_t - x_{t-1}}{x_{t-1}}\right) = \log(1 + r_t) \approx r_t.$$

Here, I used that for y close to zero $\log(1 + y) \approx y$, which follows from a first order-Taylor expansion.

Exercise 1.2 Plot the simple returns versus the log-returns for the `GSPC` data (consider adjusted closing prices).

Exercise 1.3 Suppose $x_t = e^{\alpha t}$, for some $\alpha \in \mathbb{R}$. Compute ℓ_t.

The most commonly applied transformation of financial data is that to log-returns. It is fair to ask the reason for transforming. What do we potentially gain from this? For that we need to know what a stationary time-series is. Intuitively, it is a time-series that has no obvious trends and remains in some sense the same over time. From the bottom panel in figure 3 it can be seen that the log-returns indeed do not show particular trends and fluctuate around zero. To make precise what we exactly mean by stationarity we need to view the data as a realisation of a stochastic process. In the next chapter we introduce the necessary mathematical notions for this.

CHAPTER 2

Stationary time-series

1. Recap covariance and correlation

Covariance and correlation measure dependence between two random variables X_1 and X_2. To be well defined, we assume that $\mathrm{E}\left[X_1^2\right] < \infty$ and $\mathrm{E}\left[X_2^2\right] < \infty$. In words: X_1 and X_2 are assumed to be square integrable random variables. This is a mathematical assumption to get the concept of covariance well defined.

DEFINITION 2.1. The *covariance* of X_1 and X_2 is defined by

$$\mathrm{Cov}(X_1, X_2) = \mathrm{E}\left[(X_1 - \mathrm{E}[X_1])(X_2 - \mathrm{E}[X_2])\right] = \mathrm{E}[X_1 X_2] - \mathrm{E}[X_1]\,\mathrm{E}[X_2]\,.$$

Clearly, $\mathrm{Cov}(X_1, X_2) = \mathrm{Cov}(X_2, X_1)$. If X_1 and X_2 are independent, then $\mathrm{Cov}(X_1, X_2) = 0$. The opposite direction is essentially only true for the bivariate normal distribution. For a definition of this distribution we refer to Section 3 in the appendix. We have the following linearity property of covariance: if X, Y and Z are (square integrable) random variables and a, b and c real numbers, then

$$\mathrm{Cov}(aX + bY + c, Z) = a\mathrm{Cov}(X, Z) + b\mathrm{Cov}(Y, Z)\,. \tag{1}$$

We will use this property over and over again, so it is a good idea to memorise this formula!

DEFINITION 2.2. The *correlation-coefficient* $\rho(X_1, X_2)$ of the random variables X_1 and X_2 is defined by

$$\rho(X_1, X_2) = \frac{\mathrm{Cov}(X_1, X_2)}{\sqrt{\mathrm{Var}(X_1)\,\mathrm{Var}(X_2)}}.$$

(and 0 if $\mathrm{Var}(X_1) = 0$ or $\mathrm{Var}(X_2) = 0$).

This is essentially a normalisation of covariance, as it can be proved that $-1 \leq \rho(X_1, X_2) \leq 1$ (this is a consequence of the Cauchy-Schwartz inequality). We have

$$\begin{aligned} \rho(X_1, X_2) = 1 \quad &\Longleftrightarrow \quad \mathrm{P}(X_2 = aX_1 + b) = 1 \quad \text{with } a > 0 \\ \rho(X_1, X_2) = -1 \quad &\Longleftrightarrow \quad \mathrm{P}(X_2 = aX_1 + b) = 1 \quad \text{with } a < 0 \end{aligned}$$

from which we infer that correlation measures *linear* dependence.

Exercise 2.1 Give an example of a random variable X for which $\mathrm{E}[X]$ is infinite.

2. Stationarity

An important objective of time-series analysis is prediction. We may wish to make statements like "with probability 0.9 the price of stock A will not drop below value z over the next month". This is far from easy! Just reading the statement, it is clear that we need at least

(1) a probability model for the time-series data (else we cannot make probability statements);
(2) an assumption within the model that says that some properties remain the same over time (else we cannot do prediction).

As usual in probability theory, we view the observed data $\{x_t\}_{t=0}^n$ as a realisation of random variables $\{X_t\}_{t=0}^n$. The statistical model then consists of the specification of the probability distribution of $\{X_t\}_{t=0}^n$. Hence, this is the joint distribution of the random vector $(X_0, X_1, \ldots, X_n)$. Somewhat confusingly, both $\{x_t\}$ and $\{X_t\}$ are usually referred to as time-series. However, it is clear that $\{x_t\}$ simply consists of real numbers, whereas $\{X_t\}$ is a sequence of random variables!

DEFINITION 2.3. A time-series $\{X_t\}$ is called *strictly stationary* if for any set of times $t_1, \ldots, t_k$ the distribution of $(X_{t_1+h}, \ldots, X_{t_k+h})$ does not depend on h.

This is not a very handy definition and therefore we will consider a weaker concept of stationarity.

DEFINITION 2.4. Let $\{X_t\}$ be a time series with $\mathrm{E}\left[X_t^2\right] < \infty$. $\{X_t\}$ is called *stationary*, if

(1) $\mathrm{E}[X_t] = \mu$ is independent of t
(2) $\mathrm{Cov}(X_{t+h}, X_t)$ is independent of t for each h

Loosely speaking a stationary process is one whose statistical properties do not change over time. By property 2, it follows that

$$\mathrm{Var}(X_t) = \mathrm{Cov}(X_t, X_t)$$

is constant. Hence, for a stationary time-series both the expectation and variance are constant over time.

DEFINITION 2.5. Let $\{X_t\}$ be a stationary time series with mean μ.
The *autocovariance function* of $\{X_t\}$ at lag h is

$$\gamma_X(h) = \mathrm{Cov}(X_{t+h}, X_t)\,.$$

The *autocorrelation function* of $\{X_t\}$ at lag h is

$$\rho_X(h) = \frac{\gamma_X(h)}{\gamma_X(0)} = \rho\,(X_{t+h}, X_t)\,.$$

Note that both functions indeed do not depend on t, as we assume $\{X_t\}$ is stationary. It is customary to only consider nonnegative values of h because

$$\gamma_X(h) = \gamma_X(-h) \qquad \rho_X(h) = \rho_X(-h).$$

That is, both $h\gamma_X(h)$ and $h \mapsto \rho_X(h)$ are even functions. The value of h is often referred to as the *lag* and then $\rho_X(h)$ is called the *lag h-correlation*. In the next section we give some examples of time-series for which we will compute the autocorrelation function.

DEFINITION 2.6. A real-valued function f defined on the integers is *nonnegative definite* if

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i f(i-j) a_j \geq 0$$

for all positive integers n and vectors $\boldsymbol{a} = \begin{bmatrix} a_1 & \dots & a_n \end{bmatrix}' \in \mathbb{R}^n$.

THEOREM 2.7. A real-valued function defined on the integers is the autocovariance function of a stationary time-series if and only if it is even and nonnegative definite.

PROOF. Let $\boldsymbol{X}_n = \begin{bmatrix} X_1 & \cdots X_n \end{bmatrix}'$. Then $\boldsymbol{a}'\boldsymbol{X}_n$ is a random variable and hence $\operatorname{Var}(\boldsymbol{a}'\boldsymbol{X}_n) \geq 0$. Define the $n \times n$ matrix Γ by $\Gamma_{i,j} = \operatorname{Cov}(X_i, X_j)$. By stationarity $\Gamma_{i,j} = \gamma(i-j)$. Now

$$\begin{aligned}\operatorname{Var}(\boldsymbol{a}'\boldsymbol{X}_n) &= \operatorname{Cov}\left(\sum_{i=1}^{n} a_i X_i, \sum_{j=1}^{n} a_j X_j\right) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \gamma(i-j).\end{aligned}$$

This proves that the autocovariance function of a stationary time-series is nonnegative definite. The proof for the reverse statement is more difficult and omitted. □

The proof also reveals that

$$\operatorname{Var}(\boldsymbol{a}'\boldsymbol{X}_n) = \boldsymbol{a}'\Gamma\boldsymbol{a}, \tag{2}$$

a property that will turn out to be useful later on.

3. Some examples of time-series models

EXAMPLE 2.8. Suppose $\{X_t\}$ is a sequence of independent and identically distributed random variables. In that case we write $X_t \sim \text{IID}$. If we assume $\mathrm{E}[X_t] = 0$ and $\operatorname{Var}(X_t) = \sigma^2$, then we write $X_t \sim \text{IID}(0, \sigma^2)$. Clearly

$$\operatorname{Cov}(X_s, X_t) = \begin{cases} \sigma^2 & \text{if} \quad s = t \\ 0 & \text{otherwise} \end{cases}.$$

Hence $\{X_t\}$ is a stationary time-series (put shortly "$\{X_t\}$ is stationary") with

$$\gamma_X(h) = \begin{cases} \sigma^2 & \text{if} \quad h = 0 \\ 0 & \text{if} \quad h \neq 0 \end{cases}$$

and

$$\rho_X(h) = \begin{cases} 1 & \text{if} \quad h = 0 \\ 0 & \text{if} \quad h \neq 0 \end{cases}.$$

For a IID-sequence, knowing X_t gives us no information about X_{t+h} (for all $h \neq 0$).

DEFINITION 2.9. A sequence $\{X_t\}$ of uncorrelated random variables, where $\mathrm{E}[X_t] = 0$ and $\mathrm{Var}(X_t) = \sigma^2$ (for each t) is called a *white-noise sequence*. We write $X_t \sim \mathrm{WN}(0, \sigma^2)$.

The autocovariance- and autocorrelation function of a white-noise sequence is exactly the same as for an IID-sequence. If $X_t \sim \mathrm{IID}(0, \sigma^2)$ then $X_t \sim \mathrm{WN}(0, \sigma^2)$. However, the converse does not hold. Here is a counterexample.

EXAMPLE 2.10. Start with $\{Z_t\} \sim \mathrm{IID}(0, 1)$, where $Z_t \sim N(0, 1)$. Now define

$$X_t = \begin{cases} Z_t & \text{if } t \text{ is even} \\ (Z_{t-1}^2 - 1)/\sqrt{2} & \text{if } t \text{ is odd} \end{cases}.$$

So $X_0 = Z_0$, $X_1 = (Z_0^2 - 1)/\sqrt{2}$, $X_2 = Z_2$, $X_3 = (Z_2^2 - 1)/\sqrt{2}$, etc. Note that for $\{X_t\}$ we only use $Z_0, Z_2, Z_4, \ldots$. We show that $X_t \sim \mathrm{WN}$. We have $\mathrm{E}[X_t] = 0$ for all t.
Next, we compute $\gamma_X(h) = \mathrm{Cov}(X_t, X_{t+h})$. As $\{Z_t\} \sim \mathrm{IID}(0, 1)$, we have that $\gamma_X(h) = 0$ for all $h \geq 2$. Now consider $h = 1$. Two cases can occur

- Adjacent X_t and X_{t+1} share the same random variable Z_t. This is for example the case for X_0 and X_1 (take $t = 0$). Then

  $$\mathrm{Cov}(X_0, X_1) = \mathrm{E}[X_0 X_1] - \mathrm{E}[X_0]\,\mathrm{E}[X_1] = \mathrm{E}[X_0 X_1],$$

  where the final equality holds since $\mathrm{E}[X_0] = \mathrm{E}[Z_0] = 0$. Now

  $$\mathrm{E}[X_0 X_1] = (1/\sqrt{2})\mathrm{E}\left[Z_0^3 - Z_0\right] = 0.$$

  Hence X_0 and X_1 are uncorrelated.
- Adjacent X_t and X_{t+1} do not share the same random variable Z_t. Take for example $t = 1$. As Z_0 and Z_2 are independent, so are X_1 and X_2. Hence we immediately get $\mathrm{Cov}(X_1, X_2) = 0$.

Hence $\gamma_X(1) = 0$. Finally, we check that $\mathrm{Var}(X_t)$ is constant. Obviously, if t is even we have $\mathrm{Var}(X_t) = 1$. Now if t is odd, then

$$\begin{aligned} \mathrm{Var}(X_t) &= \mathrm{Var}\Big((Z_{t-1}^2 - 1)/\sqrt{2}\Big) = \frac{1}{2}\mathrm{Var}\big(Z_{t-1}^2\big) \\ &= \frac{1}{2}\left(\mathrm{E}\left[Z_{t-1}^4\right] - (\mathrm{E}\left[Z_{t-1}^2\right])^2\right) = \frac{1}{2}(3 - 1^2) = 1. \end{aligned}$$

Here, we used that $\mathrm{E}\left[Z^4\right] = 3$ if $Z \sim N(0, 1)$. Hence, $\mathrm{Var}(X_t) = 1$ for all t.

This example may seem a bit artificial. However, important financial time-series models such as ARCH-models turn out to be white-noise, but certainly not IID.

Exercise 2.2 Check that $\mathrm{E}[X_t] = 0$ for all t in the preceding example

Next, we give an example of a non-stationary time-series.

EXAMPLE 2.11. Suppose $\{Z_t\} \sim \mathrm{IID}(0, \sigma^2)$. Set

$$X_t = \sum_{i=1}^{t} Z_i.$$

The sequence $\{X_t\}$ is called a *random walk*. This time-series is not stationary, since for all $h \geq 0$

$$\begin{aligned}\mathrm{Cov}(X_{t+h}, X_t) &= \mathrm{Cov}(X_t + Z_{t+1} + \cdots Z_{t+h}, X_t) \\ &= \mathrm{Cov}(X_t, X_t) + \mathrm{Cov}(Z_{t+1}, X_t) + \cdots \mathrm{Cov}(Z_{t+h}, X_t) \\ &= \mathrm{Cov}(X_t, X_t) = \mathrm{Var}(X_t) = t\sigma^2.\end{aligned}$$

We now come to somewhat more interesting time-series models. These usually take a white-noise sequence as input.

DEFINITION 2.12. The time-series $\{X_t\}$ is a *first order moving average process* if it satisfies the relation

$$X_t = Z_t + \theta Z_{t-1}, \qquad t = 0, \pm 1, \pm 2, \ldots,$$

where $\{Z_t\} \sim \mathrm{WN}(0, \sigma^2)$ and $\theta \in \mathbb{R}$. We write $\{X_t\} \sim \mathrm{MA}(1)$.

It is easy to see that $\mathrm{E}[X_t] = 0$. We have

$$\mathrm{Cov}(X_t, X_{t+h}) = \begin{cases} \sigma^2(1+\theta^2) & \text{if } h = 0 \\ \sigma^2\theta & \text{if } h = \pm 1 \\ 0 & \text{if } |h| \geq 2. \end{cases} \tag{3}$$

Therefore,

$$\rho_X(h) = \begin{cases} 1 & \text{if } h = 0 \\ \dfrac{\theta}{1+\theta^2} & \text{if } h = \pm 1 \\ 0 & \text{if } |h| \geq 2. \end{cases}$$

Note that ρ_X is the same for θ and $1/\theta$. In Figure 1 realisations of MA(1) processes with $\theta = 0.5$ and $\theta = -0.5$ are shown.

Exercise 2.3 Verify the form of the autocovariance function in (3).

(1) If $h = 0$, note that

$$\mathrm{Var}(X_t) = \mathrm{Var}(Z_t + \theta Z_{t-1}) = \mathrm{Var}(Z_t) + 2\theta\mathrm{Cov}(Z_t, Z_{t-1}) + \mathrm{Var}(\theta Z_{t-1}).$$

Now complete the derivation.

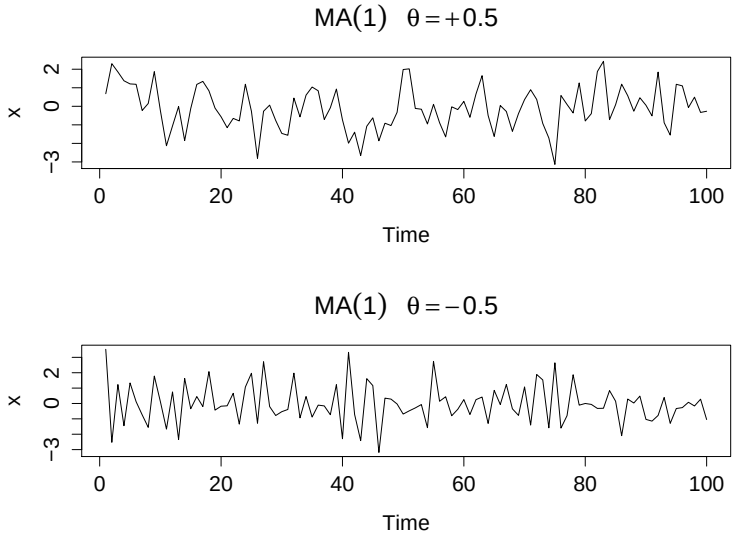

FIGURE 1. Realisations of MA(1) processes with $\theta = 0.5$ and $\theta = -0.5$.

(2) If $h = 1$, note that

$$\mathrm{Cov}(X_t, X_{t+1}) = \mathrm{Cov}(Z_t + \theta Z_{t-1}, Z_{t+1} + \theta Z_t)\,.$$

Next, use equation (1) and that $\{Z_t\} \sim \mathrm{WN}$.

DEFINITION 2.13. A time series $\{X_t\}$ follows a *first order autoregressive process*, if there exists a white noise process $\{Z_t\}$ and a coefficient $\phi \in \mathbb{R}$ such that

$$X_t = \phi X_{t-1} + Z_t \tag{4}$$

We write $\{X_t\} \sim \mathrm{AR}(1)$.

Once X_0 and $\{Z_t\}$ are given, all other values of X_t ($t \neq 0$) can be computed from the defining relation. Usually, we are interested in a (the?) stationary solution to equation (4). For that, one may pose the following questions.

- Does such a process exist?
- Is it uniquely defined?
- How should the process be initialised if we only consider $t \geq 0$?

We will get back to these questions later in more detail. For now, we note that we can repeatedly use the defining relation of a AR(1)-process to obtain

$$\begin{aligned} X_t &= \phi X_{t-1} + Z_t \\ &= \phi(\phi X_{t-2} + Z_{t-1}) + Z_t \\ &= \phi^2 X_{t-2} + \phi Z_{t-1} + Z_t = \cdots \\ &= \phi^k X_{t-k} + \sum_{j=0}^{k-1} \phi^j Z_{t-j} \end{aligned}$$

If we assume $|\phi| < 1$, then it is tempting to state that this expression converges to

$$\sum_{j=0}^{\infty} \phi^j Z_{t-j}.$$

Note however that is not a statement about convergence of real numbers, but of random variables. Convergence of a sequence of random variables can be made precise in different ways and that topic is beyond the scope of this course. Here, it would mean that

$$\mathrm{E}\left[\Big(\sum_{j=J}^{\infty} \phi^j Z_{t-j} \Big)^2 \right] \to 0 \quad \text{as} \quad J \to \infty.$$

Now from the representation $X_t = \sum_{j=0}^{\infty} \phi^j Z_{t-j}$ we get

$$\mathrm{E}[X_t] = \sum_{j=0}^{\infty} \phi^j \mathrm{E}[Z_{t-j}] = 0,$$

assuming that expectation and summation can be interchanged [1]. We can derive the auto-covariance function as follows: take $h \geq 0$, then

$$\begin{aligned} \gamma_X(h) = \mathrm{Cov}(X_t, X_{t+h}) &= \mathrm{Cov}\left(\sum_{j=0}^{\infty} \phi^j Z_{t-j}, \sum_{k=0}^{\infty} \phi^k Z_{t+h-k} \right) \\ &= \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} \phi^{j+k} \mathrm{Cov}(Z_{t-j}, Z_{t+h-k}) \end{aligned}$$

For fixed j, the covariance term is only nonzero if $t - j = t + h - k$. That is, if and only if $k = h + j$. Hence

$$\gamma_X(h) = \sum_{j=0}^{\infty} \phi^{h+2j} \mathrm{Var}(Z_{t-j}) = \sigma^2 \phi^h \sum_{j=0}^{\infty} (\phi^2)^j = \sigma^2 \frac{\phi^h}{1-\phi^2}.$$

[1] This is not always allowed, but we skip all mathematical difficulties here. If you want to learn probability theory you need to learn measure theory.

So $\gamma_X(h) = \phi^h \gamma_X(0)$ and therefore

$$\rho_X(h) = \phi^h.$$

To obtain a stationary process, we need to think about the initialisation X_0. As $\mathrm{E}[X_t] = 0$ we should have $\mathrm{E}[X_0] = 0$. For the variance, from the defining relation we get

$$\mathrm{Var}(X_t) = \mathrm{Var}(\phi X_{t-1} + Z_t) = \phi^2 \mathrm{Var}(X_{t-1}) + \sigma^2.$$

Here we used that X_{t-1} only depends on $Z_{t-1}, Z_{t-2}, \ldots$. Hence, if $\mathrm{Var}(X_t)$ is constant, say equal to c, then

$$c = \phi^2 c + \sigma^2$$

which implies

$$\mathrm{Var}(X_0) = c = \frac{\sigma^2}{1 - \phi^2}.$$

Exercise 2.4 (Brockwell and Davis [2002], exercise 1.4.) Let $\{Z_t\} \sim$ IID with the $N\left(0, \sigma^2\right)$-distribution. Suppose $a, b, c \in \mathbb{R}$. Which of the following processes are stationary? For each stationary process, find its mean and covariance function.

(1) $X_t = a + bZ_t + cZ_{t-2}$.
(2) $X_t = Z_1 \cos(ct) + Z_2 \sin(ct)$.
(3) $X_t = a + bZ_0$.
(4) $X_t = Z_t Z_{t-1}$.

Exercise 2.5 (Brockwell and Davis [2002], exercise 1.5.) Let $X_t = Z_t + \theta Z_{t-2}$, where $\{Z_t\} \sim \mathrm{WN}(0, 1)$.

(1) Find the autocovariance and autocorrelation function when $\theta = 0.8$.
(2) Compute the variance of the sample mean of $(X_1 + X_2 + X_3 + X_4)/4$ when $\theta = 0.8$.
(3) Repeat (2) when $\theta = -0.8$ and compare the results.

Exercise 2.6 (Brockwell and Davis [2002], exercise 1.6.) Suppose $\{X_t\}$ and $\{Y_t\}$ are stationary time series with autocovariance functions γ_X and γ_Y respectively. Compute the autocovariance function of $\{Z_t\}$, where $Z_t = X_t + Y_t$, assuming that $\mathrm{Cov}(X_s, Y_t) = 0$ for all s and t.

Exercise 2.7 (Brockwell and Davis [2002], exercise 2.2.) Show that the process

$$X_t = A \cos(\omega t) + B \sin(\omega t), \qquad t = 0, \pm 1, \ldots$$

(where A and B are uncorrelated random variables with mean 0 and variance 1 and ω is a fixed frequency in $[0, \pi]$), is stationary and find its mean and autocovariance function. Deduce that the function $\kappa(h) = \cos(\omega h)$, $h = 0, \pm 1, \ldots$ is nonnegative definite.

Exercise 2.8 (Brockwell and Davis [2002], exercise 2.3.)

(1) Find the autocovariance function of $X_t = Z_t + 0.3Z_{t-1} - 0.4Z_{t-2}$, where $\{Z_t\} \sim \mathrm{WN}(0, 1)$.
(2) Find the autocovariance function of $Y_t = U_t - 1.2U_{t-1} - 1.6U_{t-2}$, where $\{U_t\} \sim \mathrm{WN}(0, 1/4)$. Compare the results with that found under (1).

4. Best linear prediction for stationary time-series

Prediction is an important objective in time-series analysis. In case we impose a particular structure on the predictor (linearity), this turns out to be tractable for stationary time-series.

4.1. A simple example. We will start with a simple example. Suppose we have a mean-zero stationary time series $\{X_t\}$ and we wish to predict X_{t+1} using only the observation X_t. We assume a linear predictor, that is, we assume that the predictor P_tX_{t+1} is of the form

$$P_tX_{t+1} = a_0 + a_1X_t.$$

We define (a_0, a_1) to minimise the *mean squared prediction error*

$$F(a_0, a_1) = \mathrm{E}\left[(a_0 + a_1X_t - X_{t+1})^2\right].$$

Computing partial derivatives gives

$$\frac{\partial F(a_0, a_1)}{\partial a_0} = 2\mathrm{E}\left[a_0 + a_1X_t - X_{t+1}\right]$$
$$\frac{\partial F(a_0, a_1)}{\partial a_1} = 2\mathrm{E}\left[X_t(a_0 + a_1X_t - X_{t+1})\right].$$

Setting the first equation equal to zero yields $a_0 = 0$. Plugging this into the second equation, and equating to zero here as well gives

$$a_1\mathrm{E}\left[X_t^2\right] = \mathrm{E}\left[X_tX_{t+1}\right].$$

Hence

$$a_1 = \frac{\mathrm{E}\left[X_tX_{t+1}\right]}{\mathrm{E}\left[X_t^2\right]} = \rho_X(1).$$

The Jacobian matrix of F is given by

$$2\begin{bmatrix} 1 & \mathrm{E}[X_t] \\ \mathrm{E}[X_t] & \mathrm{E}\left[X_t^2\right] \end{bmatrix}$$

and its determinant equals $2\mathrm{Var}(X_t)$. Hence, $\mathrm{Var}(X_t) > 0$ implies that the mapping $(a_0, a_1) \mapsto F(a_0, a_1)$ is strictly convex and has a unique minimiser given by $(a_0, a_1) = (0, \rho_X(1))$. We conclude that the best linear predictor for X_{t+1}, using X_t, is given by

$$P_tX_{t+1} = \rho_X(1)X_t.$$

The prediction error is given by

$$\mathrm{E}\left[(P_tX_{t+1} - X_{t+1})^2\right] = \left(1 - \rho_X(1)^2\right)\sigma^2,$$

where $\sigma^2 = \mathrm{Var}(X_t)$.

Exercise 2.9 Verify this.

Hence, the stronger the lag-1 autocorrelation, the smaller the prediction error. Intuitively this really makes sense! If $\mathrm{E}[X_t] = \mu \neq 0$, we can immediately see that the best linear predictor is given by

$$P_t X_{t+1} = \mu + \rho_X(1)(X_t - \mu)$$

(apply the previous derivation to the time-series $X_t - \mu$).

4.2. The general case. More generally, we are interested in predicting X_{t+h} $(h \geq 1)$, based on $X_1, \ldots, X_t$. We will assume that $\{X_t\}$ is a mean zero stationary time series with autocovariance function γ.

DEFINITION 2.14. The *best linear predictor* of X_{t+h}, based on $X_1, \ldots, X_t$ is defined to be the random variable

$$P_{1:t} X_{t+h} = a_0 + a_1 X_t + \cdots + a_t X_1$$

that minimises the mean squared prediction error

$$(a_0, a_1, \ldots, a_n) \mapsto \mathrm{E}\left[(P_{1:t} X_{t+h} - X_{t+h})^2\right].$$

To find $a_0, a_1, \ldots, a_t$, we proceed by equating all partial derivatives to zero:

$$\frac{\partial}{\partial a_j} \mathrm{E}[X_{t+h} - a_0 - a_1 X_t - \cdots - a_t X_1]^2 = 0, \quad \text{for } j = 0, 1, \ldots, t$$

This gives

$$\mathrm{E}\left[X_{t+h} - a_0 - \sum_{i=1}^{t} a_i X_{t+1-i}\right] = 0$$

$$\mathrm{E}\left[\left(X_{t+h} - \sum_{i=1}^{t} a_i X_{t+1-i}\right) X_{t+1-j}\right] = 0, \quad j = 1, 2, \ldots, t$$

Note that the second equation reads as

$$\mathrm{E}[(\text{Error}) \times (\text{Predictor Variable})] = 0$$

which helps in recalling the formula. Using that $\mathrm{E}[X_t] = 0$ we find that $a_0 = 0$ from the first equation. Using

$$\mathrm{E}[X_s X_t] = \mathrm{Cov}(X_s, X_t) = \gamma(s - t),$$

the second equation yields that $\mathbf{a}_t = (a_1, \ldots, a_t)$ satisfies

$$\underbrace{\begin{bmatrix} \gamma(0) & \gamma(1) & \cdots & \gamma(t-1) \\ \gamma(1) & \gamma(0) & \cdots & \gamma(t-2) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma(t-1) & \gamma(t-2) & \cdots & \gamma(0) \end{bmatrix}}_{\Gamma_t} \underbrace{\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_t \end{bmatrix}}_{\boldsymbol{a}_t} = \underbrace{\begin{bmatrix} \gamma(h) \\ \gamma(h+1) \\ \vdots \\ \gamma(h+t-1) \end{bmatrix}}_{\boldsymbol{\gamma}_t(h)} \tag{5}$$

Therefore, if Γ_t is non-singular, then

$$\mathbf{a}_t = \Gamma_t^{-1}\boldsymbol{\gamma}_t(h). \tag{6}$$

This gives

$$P_{1:t}X_{t+h} = \Gamma_t^{-1}\boldsymbol{\gamma}_t(h)\boldsymbol{X}_t,$$

with

$$\boldsymbol{X}_t = \begin{bmatrix} X_t & X_{t-1} & \cdots X_1 \end{bmatrix}'. \tag{7}$$

The mean squared predictor error is given by

$$\begin{aligned}
\mathrm{E}\left[(X_{t+h} - P_{1:t}X_{t+h})^2\right] &= \mathrm{E}\left[\left(X_{t+h} - \boldsymbol{a}_t'\boldsymbol{X}_t\right)^2\right] \\
&= \mathrm{E}\left[X_{t+h}^2\right] - 2\mathrm{E}\left[X_{t+h}(a_1X_t + \ldots a_tX_1)\right] + \mathrm{E}\left[\left(\boldsymbol{a}_t'\boldsymbol{X}_t\right)^2\right] \\
&= \sigma^2 - 2(a_1\gamma(h) + \ldots + a_t\gamma(t+h-1) + \mathrm{Var}\left(\boldsymbol{a}_t'\boldsymbol{X}_t\right) \\
&= \sigma^2 - 2\boldsymbol{a}_t'\boldsymbol{\gamma}_t(h) + \boldsymbol{a}_t'\Gamma_t\boldsymbol{a}_t,
\end{aligned}$$

where we used (2) as the final equality. Using (6) we get

$$\mathrm{E}\left[(X_{t+h} - P_{1:t}X_{t+h})^2\right] = \sigma^2 - \boldsymbol{\gamma}_t(h)'\Gamma_t^{-1}\boldsymbol{\gamma}_t(h) \leq \sigma^2$$

since Γ is nonnegative definite.

Let's summarise what we have derived.

LEMMA 2.15. If Γ_t is nonsingular, then

$$P_{1:t}X_{t+h} = \Gamma_t^{-1}\boldsymbol{\gamma}_t(h)\boldsymbol{X}_t,$$

where Γ_t and $\boldsymbol{\gamma}_t(h)$ are defined in (5) and $\boldsymbol{X}_t$ is defined in (7). The mean squared prediction error satisfies

$$\mathrm{E}\left[(X_{t+h} - P_{1:t}X_{t+h})^2\right] = \sigma^2 - \boldsymbol{\gamma}_t(h)'\Gamma_t^{-1}\boldsymbol{\gamma}_t(h).$$

EXAMPLE 2.16. Assume $\{X_t\} \sim \mathrm{AR}(1)$, i.e. $|\phi| < 1$, $\{Z_t\} \sim \mathrm{WN}(0, \sigma^2)$ and

$$X_t = \phi X_{t-1} + Z_t, \qquad t = 0, \pm 1, \pm 2, \ldots,$$

Suppose we are interested in the one-step ahead best linear predictor $P_{1:t}X_{t+1}$. Because $\gamma(h) = \phi^{|h|}$, the preceding derivation gives that the coefficients $a_1, \ldots, a_t$ must be a solution of

$$\begin{bmatrix} 1 & \phi & \phi^2 & \cdots & \phi^{t-1} \\ \phi & 1 & \phi & \cdots & \phi^{t-2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \phi^{t-1} & \phi^{t-2} & \phi^{t-3} & \cdots & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_t \end{bmatrix} = \begin{bmatrix} \phi \\ \phi^2 \\ \vdots \\ \phi^t \end{bmatrix}$$

Clearly $\mathbf{a}_t = (\phi, 0, \ldots, 0)$ is a solution, so $P_{1:t}X_{t+1} = \phi X_t$.

For an AR(1) time-series this example shows that the best linear predictor is easily found. In general this is not the case.

Exercise 2.10 Consider $\{X_t\} \sim \mathrm{MA}(1)$. Find $P_{1:3}X_4$ by solving the system in (5).

5. The partial autocorrelation function

Suppose $\{X_t\}$ is a mean zero stationary time-series.

DEFINITION 2.17. The *partial auto-correlation* at lag $k \geq 1$ is defined as the coefficient $\alpha(k)$ of X_1 in the best linear predictor $P_{1:k}X_{k+1}$. The mapping $\alpha\colon \{1, 2, \ldots\} \to \mathbb{R}$ is called the partial autocorrelation function (PACF).

That is, if we take $h = 1$ and $t = k$ in (5), then a_{kk} is uniquely defined by

$$\begin{bmatrix} \gamma(0) & \gamma(1) & \cdots & \gamma(k-1) \\ \gamma(1) & \gamma(0) & \cdots & \gamma(k-2) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma(k-1) & \gamma(k-2) & \cdots & \gamma(0) \end{bmatrix} \begin{bmatrix} a_{1k} \\ a_{2k} \\ \vdots \\ a_{kk} \end{bmatrix} = \begin{bmatrix} \gamma(1) \\ \gamma(2) \\ \vdots \\ \gamma(k) \end{bmatrix}$$

and $\alpha(k) = a_{kk}$. This definition is admittedly complicated. However, it turns out to be very useful. From the results in subsection 4.1 we get that always

$$\alpha(1) = \rho_X(1).$$

Now let's go to $k = 2$ and consider $P_{1:2}X_3$. This is of the form $a_1X_2 + a_2X_1$. As $\alpha(2) = a_2$ it gives the contribution of X_1 to the best linear predictor 2 lags ahead. For $k = 3$ we consider $P_{1:3}X_4$. This is of the form $a_1X_3 + a_2X_2 + a_3X_1$. As $\alpha(3) = a_3$ it gives the contribution of X_1 to the best linear predictor 3 lags ahead. Hence, we see that $\alpha(k)$ gives the contribution of the best linear predictor k lags ahead.

LEMMA 2.18. If $\{X_t\} \sim \text{AR}(1)$, then $\alpha(k) = 0$ for all $k \geq 2$.

PROOF. This follows immediately from example 2.16. □

While very simple, this lemma will turn out to be very useful in identifying AR(1) time-series.

EXAMPLE 2.19. Consider a causal AR(p) process

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} Z_t, \qquad \{Z_t\} \sim \text{WN}(0, \sigma^2).$$

Then it is immediately clear that

$$\alpha(h) = \begin{cases} \phi_p & \text{if } h = p \\ 0 & \text{if } h > p \end{cases}$$

Values of $\alpha(h)$ for $1 \leq h < p$ can be computed separately. This result is useful in case we know $\{X_t\}$ is an AR-process, but its order is unknown. If we would know the partial autocorrelation, then the order is the determined by exactly that lag for which all higher order lag partial autocorrelations vanish.

6. Exercises from old exams

Exercise 2.11 Let $\{Z_t\}$ be a sequence of independent normally distributed random variables. Assume $\mathrm{E}[Z_1] = 0$ and $\mathrm{Var}(Z_1) = \sigma^2$. Let a, b and c be constants.

(1) Suppose $X_t = a + bZ_t + cZ_{t-2}$, $t = 0, 1, \ldots$. Compute the autocovariance function for lags 1 and 2.
(2) Suppose $X_t = Z_0 \cos(\pi t)$, $t = 0, 1, \ldots$. Compute the mean and autocovariance function to conclude that this process is stationary.

Exercise 2.12 Suppose $\{X_t\}$ is a mean-zero stationary time series. Show that the best linear predictor for X_{t+1}, based on only X_{t-1} is given by $\rho_X(2)X_{t-1}$. Here $\rho_X(h) = \rho(X_t, X_{t+h})$.

Exercise 2.13 Simulate 3 realisations of length 10000 from the "2-regime AR(1)" model

$$X_t = \begin{cases} -1.5X_{t-1} + Z_t & \text{if } X_{t-1} < 0 \\ \alpha X_{t-1} + Z_t & \text{if } X_{t-1} \geq 0 \end{cases},$$

with $\{Z_t\} \sim$ IID $N(0, 1)$ and $\alpha > 0$. Take $X_0 = 0$ and both $\alpha = 0.5$ and $\alpha = 1$. Next consider $\alpha = 0.5$, $\alpha = 1$ and $\alpha = 1.5$: for which values of α does it appear that $\{X_t\}$ is stationary?

CHAPTER 3

From data to time-series model

1. Transforming time-series to stationarity

We have seen in the previous chapter how to do prediction for stationary time-series. For an observed time-series (so the data), stationarity is an assumption that we impose. We can never be absolutely sure that it applies, but the use of a statistical model is a simplification in itself anyway. As a stationary time-series has constant expectation and variance, trends indicate nonstationarity. Many time-series encountered in practise appear indeed to be nonstationary. Fortunately, we can transform the data such that the stationarity assumption is more reasonable for the transformed data. The general approach is as follows:

- The observations $\{x_t\}$ are assumed to be a realisation of a time series $\{X_t\}$.
- We plot the series and inspect its graph with the *classical decomposition* in mind, i.e.

$$X_t = m_t + s_t + Y_t$$

  where
  - m_t is a slowly changing *trend component*;
  - s_t is a periodic function that reflects a *seasonal component*;
  - Y_t is a random noise term that is assumed to be *stationary*.
- We estimate and extract the non-random components m_t and s_t.
- We find find a suitable probabilistic model for Y_t (which we often choose from a collection of tractable time-series models).

There are various ways to estimate the trend component m_t. One option is to assume a *polynomial trend*

$$m_t = a_0 + a_1 t + \cdots + a_k t^k$$

for some $k \geq 1$ and where $a_0, a_1, \ldots, a_k \in \mathbb{R}$.
Then we can find estimates $\widehat{a}_0, \widehat{a}_1, \ldots, \widehat{a}_k$ through fitting the observations $\{x_t\}$ by means of least squares, i.e., by minimising

$$\sum_{t=1}^{n} (x_t - a_0 - a_1 t - \cdots - a_k t^k)^2$$

over $a_0, a_1, \ldots, a_k$. The resulting trend estimate is then given by $\widehat{m}_t = \widehat{a}_0 + \widehat{a}_1 t + \cdots + \widehat{a}_k t^k$ Subsequently, the de-trended time series is given by

$$Y_t = X_t - \widehat{m}_t = X_t - \widehat{a}_0 - \widehat{a}_1 t - \cdots - \widehat{a}_k t^k.$$

Another option for removing trends is *differencing*.

DEFINITION 3.1. Define the *backward shift operator* B by

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

and the *differencing operator* by

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

Polynomial trends can be removed by considering repeated differencing of the original time series. To see this, first assume

$$m_t = c_0 + c_1 t.$$

Then is is easy to verify that

$$\nabla m_t = c_1$$

from which we conclude that a linear drift can be removed by differencing once. If we difference the time-series twice, we get

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

Exercise 3.1 Show that if X_t is a quadratic function of t, twice differencing removes the time dependence.

Exercise 3.2 Suppose $\{X_t\}$ is a monthly time-series with yearly seasonal pattern, i.e. $X_t = X_{t-12} + Z_t$, where $\{Z_t\}$ is white-noise. Explain why $\nabla^{12} X_t$ removes the seasonality.

REMARK 3.2. If $X_t = a_0 + a_1 t + Y_t$, with $\{Y_t\} \sim$ WN, the trend can be removed by either polynomial fitting or differencing. Note however that these two methods are really different.

For financial time-series, it is often assumed that the logreturns are stationary. With the just defined notation we can write

$$\ell_t = \nabla \log X_t.$$

Hence, there is a simple transformation first. Next, one would like to know the dependence structure of ℓ_t. For that, we need to estimate the autocorrelation function.

2. Estimating the autocorrelation function

Once we believe our time-series data can be modeled as a realisation of a stationary time-series, we are interested in finding a suitable time-series model. For that, one needs to learn some flexible classes of time-series models capturing various characteristics in the autocorrelation and partial autocorrelation functions. To identify which class "fits" the data best, it is useful to estimate the mean and autocorrelation function from the data.

For a stationary time-series, we can estimate the mean $\mu = \mathrm{E}[X_t]$ by the sample mean estimate

$$\bar{x} = \frac{1}{n}\sum_{t=1}^{n} x_t.$$

For $h \geq 0$, we can estimate the covariance $\mathrm{Cov}(X_{t+h}, X_t) = \mathrm{E}[(X_{t+h} - \mu)(X_t - \mu)]$ by the sample covariance of the datasets

$$\begin{array}{lcccc} \text{dataset 1:} & x_{1+h} & x_{2+h} & \cdots & x_n \\ \text{dataset 2:} & x_1 & x_2 & \cdots & x_{n-h} \end{array}$$

This leads to the following definition.

Definition 3.3. The *sample autocovariance* at lag h is defined by

$$\widehat{\gamma}(h) = \frac{1}{n}\sum_{t=1}^{n-h}(x_{t+h} - \bar{x})(x_t - \bar{x}), \qquad \text{for } h \geq 0.$$

We note that estimates of sample autocovariances for higher lags are less reliable than those of smaller lags, as the averaging takes place over less terms.

Remark 3.4. Dividing by n instead of $n - |h|$ is to ensure that sample covariance matrix

$$\widehat{\Gamma}_n := [\widehat{\gamma}(i-j)]_{i,j=1}^{n} = \begin{pmatrix} \widehat{\gamma}(0) & \widehat{\gamma}(1) & \cdots & \widehat{\gamma}(n-1) \\ \widehat{\gamma}(1) & \widehat{\gamma}(0) & \cdots & \widehat{\gamma}(n-2) \\ \vdots & \vdots & \ddots & \vdots \\ \widehat{\gamma}(n-1) & \widehat{\gamma}(n-2) & \cdots & \widehat{\gamma}(0) \end{pmatrix}$$

is nonnegative definite. This makes sense, as we know that the true covariance matrix Γ always has this property.

As $\rho_X(h) = \gamma_X(h)/\gamma_X(0)$, the following definition should not come as a surprise.

Definition 3.5. The *sample autocorrelation* at lag h is defined by

$$\widehat{\rho}(h) = \frac{\widehat{\gamma}(h)}{\widehat{\gamma}(0)} = \frac{\displaystyle\sum_{t=1}^{n-h}(x_{t+h} - \bar{x})(x_t - \bar{x})}{\displaystyle\sum_{t=1}^{n-h}(x_t - \bar{x})^2}, \qquad \text{for } h \geq 0.$$

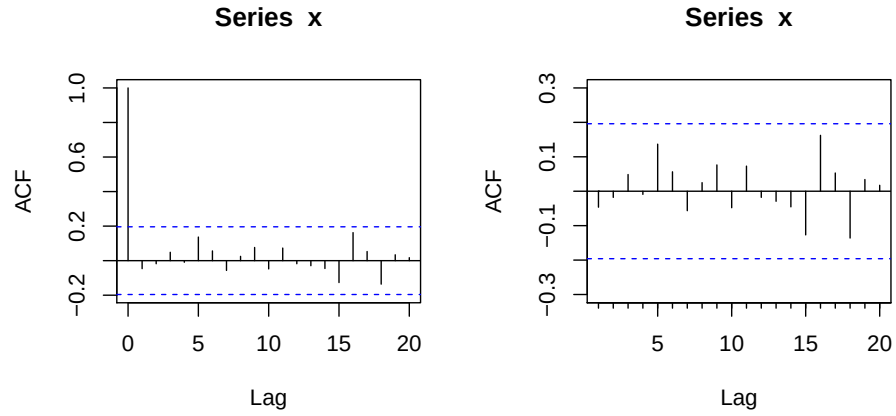

FIGURE 1. ACF-plots for 100 independent draws from the standard Normal distribution. Left: lag 0 included. Right: lag 0 omitted.

2.1. Some simulations in R for the ACF. We sample 100 independent draws from the standard Normal distribution and compute the (estimated) autocorrelation function.

```
x <- rnorm(100)
pdf('acf-ind.pdf',width=7,height=3.5)
par(mfrow=c(1,2))
acf(x)
Acf(x) # acf without lag 0
dev.off()
```

In the code, the way to export your the figure to a pdf figure is also included. The resulting figure is displayed in Figure 1. The two figures are the same, however the left one includes lag 0, while the right one does not. We know that all autocorrelations beyond lag 0 are equal to zero for an IID-sequence. However, since we only have limited information (a sample), many lags are estimated to be close to zero, but not exactly equal to zero. The dashed-blue lines in the figures will be explained later. Now we can also draw data from an AR(1) or MA(1) time-series model. Figures 2 and 3 show ACF-plots for a sample of size 100, with the true values of the ACF superimposed in red. It can be seen that estimates tend to be slightly more off for higher lags. The code for producing these figures is below (the commands for saving the figure to a pdf have been left out).

```
y <- arima.sim(list(ar=0.8),n=100)
acf(y,ylim=c(-0.4,1.2),main='AR(0.8)')
points(0:25,ARMAacf(ar=0.8,lag=25),type="b",col='red')
```

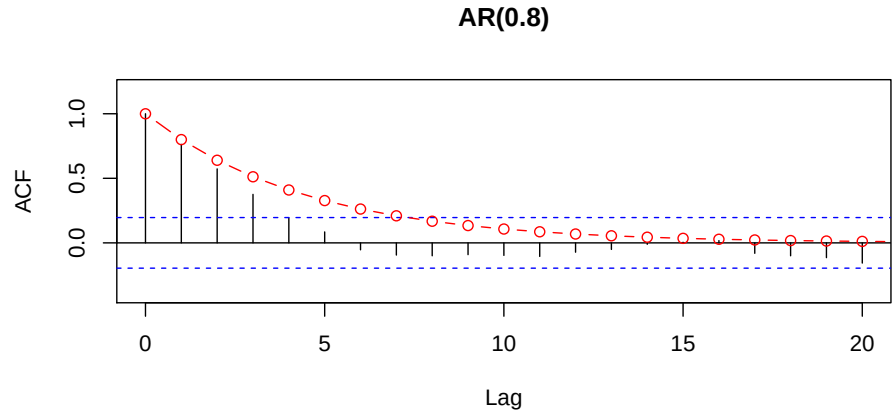

FIGURE 2. ACF-plots for 100 independent draws from a AR(1) time series with $\phi = 0.8$. The true values of the ACF are superimposed in red.

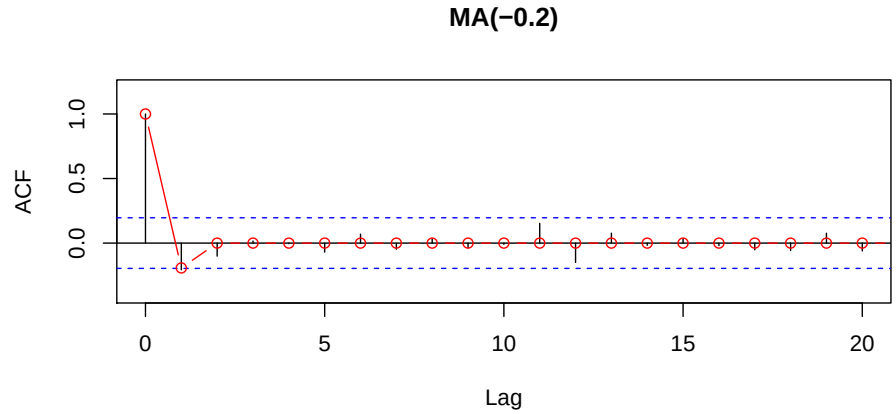

FIGURE 3. ACF-plots for 100 independent draws from a MA(1) time series with $\theta = -0.2$. The true values of the ACF are superimposed in red.

```
z <- arima.sim(list(ma=-0.2),n=100)
acf(z,ylim=c(-0.4,1.2),main='MA(-0.2)')
points(0:25,ARMAacf(ma=-0.2,lag=25),type="b",col='red')
```

2.2. Asymptotics for the ACF. Asymptotics refers to the large sample limit $n \to \infty$. Intuitively, when $n \to \infty$ one expects $\rho_n(h)$ to converge to

$\rho_X(h)$. Here we have written $\rho_n(h)$ instead of $\rho(h)$ to emphasise the dependence of the estimator on n. It requires some effort to make this statement precise. First of all, we need to make the notion of convergence precise.

DEFINITION 3.6. For a sequence of random variables $\{X_n\}_{n=1}^{\infty}$ we say that X_n *converges in distribution* to the random variable X if

$$\mathrm{P}(X_n \leq x) \to \mathrm{P}(X \leq x) \qquad (n \to \infty)$$

for all points where $x \mapsto \mathrm{P}(X \leq x)$ is continuous. We write $X_n \rightsquigarrow X$.

This definition can be extended in a straightforward way to convergence in distribution of random vectors.
The following theorem shows that the rescaled and centred vector of autocorrelations converges to a multivariate normal distribution.

THEOREM 3.7. Assume

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

for an IID-sequence $\{Z_t\}$ with mean zero and $\mathrm{E}\left[Z_t^4\right] < \infty$ and numbers ψ_j with $\sum_j |\psi_j| < \infty$. Then

$$\sqrt{n}\left(\begin{bmatrix}\hat{\rho}_n(1)\\ \cdots \\ \hat{\rho}_n(h)\end{bmatrix} - \begin{bmatrix}\rho_X(1)\\ \cdots \\ \rho_X(h)\end{bmatrix}\right) \rightsquigarrow N_h(0, W),$$

where W has (g, h) element

$$W_{g,h} = \sum_k \big(\rho_X(k+g)\rho_X(k+h) + \rho_X(k-g)\rho_X(k+h) + 2\rho_X(g)\rho_X(h)\rho_X(k)^2 \\ -2\rho_X(g)\rho_X(k)\rho_X(k+h) - 2\rho_X(h)\rho_X(k)\rho_X(k+g)\big).$$

The expression for W is known as *Bartlett's formula*.

COROLLARY 3.8. If $X_t = \mu + Z_t$, where $\{Z_t\} \sim$ IID, then $W = I$. Therefore, $\hat{\rho}_n(1), \ldots, \hat{\rho}_n(h)$ are approximately independent normal random variables with mean zero and variance $1/n$.

From this corollary it follows that for a IID-time series, for n large and $h \geq 1$, $|\rho_n(h)| \leq 2/\sqrt{n}$ with approximately probability 0.95 (actually the value 1.96 is more precise here than 2). Now look at the dotted blue lines in figures 1, 2 and 3: these are exactly at height $\pm 2/\sqrt{n}$ (which is 0.2 as $n = 100$). Hence, if data are IID, then we would expect 95% of the (estimated) autocorrelations to be within the two blue lines. We should also expect that 5% are not!
Another application of this result is the *Ljung-Box*-test. Here, we test the null hypothesis

$$H_0 \colon \rho_X(1) = \ldots, \rho_X(k) = 0,$$

for some fixed value of k. Under H_0, the sum $\sum_{h=1}^{k} (\sqrt{n}\hat{\rho}_n(h))^2$ is approximately (in the large sample limit) distributed as the sum of k independent

squared normal random variables. This distribution is well-known from probability theory and is the $\chi^2(k)$-distribution (the "chi-square distribution with k degrees of freedom"). The Ljung-Box test statistic (as implemented in `R`) is actually slightly different, it is defined by

$$T_{LB} = \sum_{h=1}^{k} \frac{n+2}{n-h} n\hat{\rho}_n(h)^2, \tag{8}$$

the factor $(n+2)/(n-h)$ being included so that the chi-square approximation is better for small sample sizes.

3. Estimating the partial autocorrelation function

In section 5 we defined the PACF at lag k to be equal to the solution of a_{kk} in the system of equations

$$\begin{bmatrix} \gamma(0) & \gamma(1) & \cdots & \gamma(k-1) \\ \gamma(1) & \gamma(0) & \cdots & \gamma(k-2) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma(k-1) & \gamma(k-2) & \cdots & \gamma(0) \end{bmatrix} \begin{bmatrix} a_{1k} \\ a_{2k} \\ \vdots \\ a_{kk} \end{bmatrix} = \begin{bmatrix} \gamma(1) \\ \gamma(2) \\ \vdots \\ \gamma(k) \end{bmatrix}.$$

Plugging in estimators $\gamma_n(h)$ for $\gamma(h)$ yields the *sample auto correlation* at lag h, $\hat{\alpha}(k)$ to be defined as the solution of $\hat{a}_{kk}$ to the system of equations

$$\begin{bmatrix} \hat{\gamma}(0) & \gamma(1) & \cdots & \hat{\gamma}(k-1) \\ \hat{\gamma}(1) & \gamma(0) & \cdots & \hat{\gamma}(k-2) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\gamma}(k-1) & \hat{\gamma}(k-2) & \cdots & \hat{\gamma}(0) \end{bmatrix} \begin{bmatrix} \hat{a}_{1k} \\ \hat{a}_{2k} \\ \vdots \\ \hat{a}_{kk} \end{bmatrix} = \begin{bmatrix} \hat{\gamma}(1) \\ \hat{\gamma}(2) \\ \vdots \\ \hat{\gamma}(k) \end{bmatrix}.$$

4. Asymptotics for the PACF

Just as for the ACF, for the PACF it can be shown that $\sqrt{n}(\hat{\alpha}_n(k) - \alpha(k)) \rightsquigarrow N(0, V)$ for some $V > 0$. That is, the sample PACF is asymptotically normally distributed. We omit the expression for V (which is complicated). The following theorem can be skipped now, but will be used in the sequel.

THEOREM 3.9. Let $\{X_t\}$ be a causal AR(p) process with driving noise sequence $\{Z_t\} \sim \text{IID}(0, \sigma^2)$. Then for $h > p$

$$\sqrt{n}\widehat{\alpha}(h) \rightsquigarrow N(0,1) \quad \text{as} \quad n \to \infty.$$

This theorem can be applied for order selection of AR processes. If

$$\begin{cases} |\widehat{\alpha}(h)| > 1.96/\sqrt{n} & \text{for } 0 \le h \le p \\ |\widehat{\alpha}(h)| \le 1.96/\sqrt{n} & \text{for } h > p. \end{cases}$$

then this suggests that an AR(p) process may be appropriate.

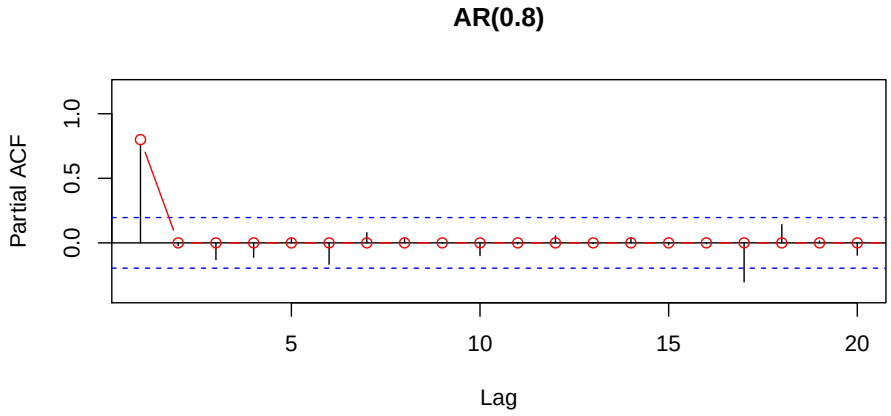

FIGURE 4. PACF-plots for 100 independent draws from a AR(1) time series with $\phi = 0.8$. The true values of the PACF are superimposed in red.

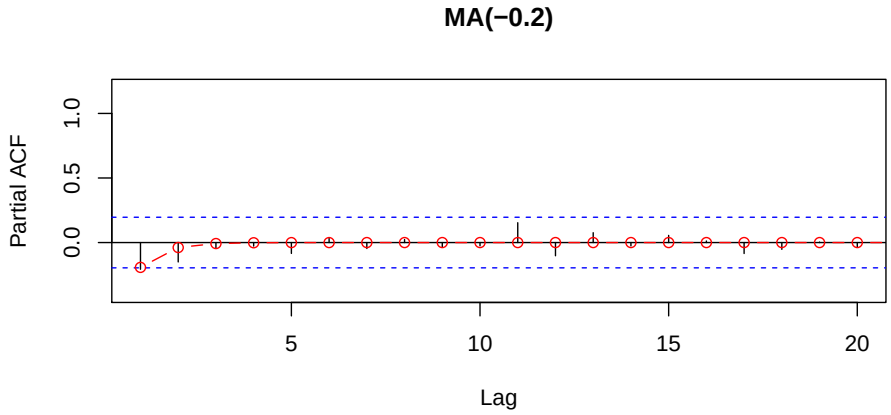

FIGURE 5. PACF-plots for 100 independent draws from a MA(1) time series with $\theta = -0.2$. The true values of the PACF are superimposed in red.

5. Some simulations in R for the PACF

Again we consider data from an AR(1) or MA(1) time-series model. Figures 4 and 5 show PACF-plots for a sample of size 100, with the true values of the PACF superimposed in red. It can be seen that estimates tend to be slightly more off for higher lags. The code for producing these figures is below.

```
pacf(y,ylim=c(-0.4,1.2),main='AR(0.8)')
```

```
points(1:25,ARMAacf(ar=0.8,lag=25,pacf=TRUE),type="b",col='red')

pacf(z,ylim=c(-0.4,1.2),main='MA(-0.2)')
points(1:25,ARMAacf(ma=-0.2,lag=25,pacf=TRUE),type="b",col='red')
```

Note that the AR(1) process has exponentially decreasing ACF, while the PACF equals zero after lag 1. For the MA(1) process it is exactly the other way around: it appears that the PACF is exponentially decreasing, while the ACF is zero after lag 1. We will get back to this in more generality.

6. Some methods to check the noise sequence

Once the data have been transformed to stationarity, one tries to capture the dependence (correlation) structure in the data by fitting a time-series model to the transformed data. Obviously, it is of interest fo assess whether the transformed data can be considered a realisation of an IID-sequence. For if that is the case, it does not make sense to do any further modelling on the dependence structure. From the asymptotics in the preceding section we summarise three ways for assessing that the data $x_1, \ldots, x_n$ can be considered to be a realisation of a IID-sequence

(1) Visually assess the ACF-plot: about 95% of the estimated autocorrelations should be in between the blue curves at $\pm 1.96/\sqrt{n}$.
(2) Compute the value of the Ljung-Box statistic as given in (8) for a fixed (chosen) value of k. If it exceeds the α-upper quantile of the $\chi^2(k)$-distribution, this implies rejection of the null hypothesis. Formally, this test can only be used to detect violations from the null-hypothesis, that states that the first k autocorrelations are all equal to zero. In Brockwell and Davis [2002], the Ljung-Box test is called the "portmanteau-test".
(3) Apply the turning-point-test. This test is based on the following idea: if $\{Z_t\} \sim$ IID then three successive values are equally likely to occur in any of the six possible orders

Under the hypothesis that $\{Z_t\} \sim$ IID, the probability of a turning point equals $2/3$. If we let T_n denote the number of turning points in a series of n points, then $\mathrm{E}[T_n] = \frac{2}{3}(n-2)$. Now it can be shown that under the null-hypothesis

$$U_n = \frac{T_n - 2(n-2)/3}{\sqrt{(16n-29)/90}} \rightsquigarrow N(0,1).$$

Hence, the null-hypothesis is rejected at significance level 0.05 if $|U_n| > 1.96$.

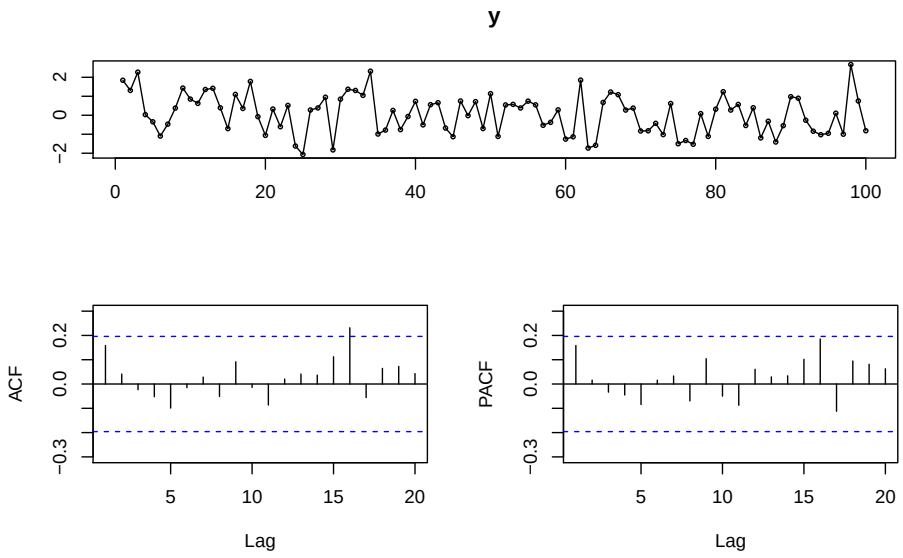

FIGURE 6. Time-series plot, ACF-plot and PACF-plot for a realisation of $\{Y_t\}$.

To illustrate these methods, we simulate three datasets from different time-series models:

(1) $\{X_t\} \sim$ IID $N(0,1)$;
(2) $\{Y_t\} \sim$ AR(1) with $\phi = 0.2$ and driving noise sequence $\{Z_t\} \sim$ IID $N(0,1)$;
(3) $\{Y_t\} \sim$ AR(1) with $\phi = 0.9$ and driving noise sequence $\{Z_t\} \sim$ IID $N(0,1)$;

Using sample size 100 this can easily be done:

```
x <- rnorm(100)
y <- arima.sim(list(ar=0.2),100)
z <- arima.sim(list(ar=0.9),100)
```

Using

```
tsdisplay(y)
```

(and similarly for `x` and `z`) one figure can be obtained in which the data, ACF and PACF are shown (note the ACF for lag 0 is omitted, as it is trivially known to be equal to 1). See Figure 6. To each of the time-series, we apply Ljung-Box tests and the turning point test. For `y` the code is [1]

[1] Unfortunately, the turning-point-test is not available in `R`, but you can source the script `tpt.R`.

```
Box.test(y,lag=5,type='Ljung-Box')
Box.test(y,lag=10,type='Ljung-Box')
Box.test(y,lag=20,type='Ljung-Box')
turning.point.test(y)
```

We summarise the p-values obtained (rounded to two digits):

Model	LB($k=5$)	LB($k=10$)	LB ($k=20$)	Turning point test
$\{X_t\}$	0.75	0.24	0.49	0.52
$\{Y_t\}$	0.52	0.85	0.68	0.87
$\{Z_t\}$	$< 2.2 \cdot 10^{-16}$	$< 2.2 \cdot 10^{-16}$	$< 2.2 \cdot 10^{-16}$	$3.70 \cdot 10^{-6}$

Given the small sample size, the moderate deviation of $\{Y_t\}$ from the IID-case remains undetected. In view of Figure 6 this is not too surprising.

Exercise 3.3 Repeat this analysis with larger sample sizes. Check whether $\{Y_t\}$ is indeed "flagged" to be different from IID.

7. Checking for normality using a probability plot

It is a common assumption that the marginal distribution of a white-noise series $\{Z_t\}$ is Gaussian. So can we check whether data $z_1, \ldots, z_n$ are realisations of the $N(\mu, \sigma^2)$-distribution? There are many statistical tests for this purpose, but there is also a powerful graphical way to assess normality. To explain this, let Φ denotes the cumulative distribution function of the standard Normal distribution. So

$$\Phi(x) = \mathrm{P}(U \leq x)\,, \qquad U \sim N(0,1), \qquad x \in \mathbb{R}.$$

Assume F is the cumulative distribution function of the $N(\mu, \sigma^2)$-distribution. Then

$$F^{-1}(x) = \mu + \sigma \Phi^{-1}(x).$$

Exercise 3.4 Prove this. Hint: if $U \sim N(0,1)$, then $\mu + \sigma U \sim N(\mu, \sigma^2)$.

If we order the data to $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$, then, assuming the data are realisations from the $N(\mu, \sigma^2)$-distribution,

$$F(x_{(i)}) \approx i/n.$$

Applying F^{-1} on both sides gives

$$x_{(i)} \approx F^{-1}(i/n) = \mu + \sigma \Phi^{-1}(i/n).$$

So then we can check whether the points

$$\big(\Phi^{-1}(i/n), x_{(i)}\big) \qquad 1 \leq i \leq n$$

are on a straight line. As the definition of $\Phi^{-1}(x)$ is problematic for $x = 1$, instead the points

$$\big(\Phi^{-1}(i/(n+1)), x_{(i)}\big) \qquad 1 \leq i \leq n$$

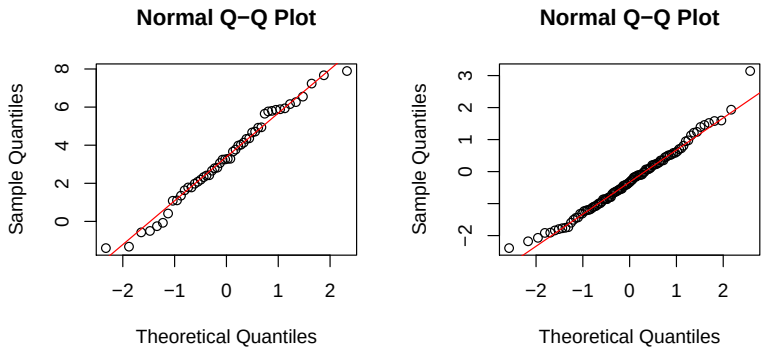

FIGURE 7. Normal probability plots. Left: sample size 50 from the $N(3,4)$-distribution. Right: sample size 100 from the $N(0,1)$-distribution.

are plotted (changing n to $n+1$ should not make a big difference for sufficiently large sample size n). The resulting plot is called a *(normal) probability plot*.
In Figure 7 we show two probability plots for data sampled from a Normal distribution. Indeed the points appear to be on a straight line. The code for making this figure is as follows:

```
x <- rnorm(50,3,2); y <- rnorm(100)
par(mfrow=c(1,2))
qqnorm(x);qqline(x,col='red')
qqnorm(y);qqline(y,col='red')
```

Exercise 3.5 Sample data from distributions different than the Normal distribution, such as the exponential distribution and make normal probability plots. Sampling from the exponential distribution can be done with the command `rexp` in `R`.

8. Exercises from old exams

Exercise 3.6
Suppose $\{W_t\} \sim \text{WN}(0, \sigma^2)$. Define the time-series $\{X_t\}$ by

$$X_t = W_t + \theta W_{t-3}$$

with $\theta \in \mathbb{R}$.

(1) Simulate two realisations of $\{X_t\}$ of length 250, where the sequence $\{W_t\}$ consists of independent random variables with the $N(0, 0.5)$-distribution and $\theta = -0.4$. Make a figure with both realisations.
(2) Compute $\text{Cov}(X_t, X_{t+h})$ for all values of h and verify that $\{X_t\}$ is stationary.

(3) Compute the autocorrelation function of $\{X_t\}$.
(4) Generate figures of the autocorrelation function for the simulated time-series in subquestion (a) and verify whether this is in agreement with your answer to exercise (c).
(5) Suppose that the time-series $\{Y_t\}$ is defined by $Y_t = \nabla X_t$. Show that $\gamma_Y(h) = 2\gamma_X(h) - \gamma_X(h+1) - \gamma_X(h-1)$.
(6) Take one of the realisations of exercise (a) and compute $y_t = \nabla x_t$ for this realisation. Now generate an autocorrelation plot of $\{y_t\}$. Using exercise (d), explain why the lag-one autocorrelation is approximately equal to -0.5.
(7) Compute the standard deviation of $\{X_t\}$. Do the empirical standard-deviations of the two realisations simulated under (a) agree with your result?

Exercise 3.7 In this exercise we are going to analyse the adjusted closing prices of Apple Inc. from January 1, 2010 until august 1, 2017. Download the data from Yahoo using the commands (you need internet access)

```
getSymbols("AAPL",src='yahoo')
```

The closing data over the required time period can be extracted and plotted as follows:

```
Cl.AAPL <- Cl(AAPL['2010-01::2017-08-01']) plot(Cl.AAPL,col='blue',main='closing values')
```

Finally the data can be converted to type `numeric` by issuing the command

```
x <- as.numeric(Cl.AAPL)
```

This changes the internal format of the time-series to enable further calculations in `R`.

(1) Is this time-series stationary? Explain your answer.
(2) Compute the logreturns and make a time-series plot of these.
(3) Make an autocorrelation plot of the logreturns and execute the Ljung-Box test with lag equal to 5 (use significance level 0.001). Add a lag-plot using the function `lag1.plot` from the `astsa` library. Is it reasonable to assume the logreturns are white-noise?
(4) Investigate whether the logreturns are normally distributed.
(5) Consider the time-series $y_t = x_t^2$, where x_t is the logreturn at time t. Make an autocorrelation plot of $\{y_t\}$. Can we conclude that $\{x_t\}$ is a realisation of an IID-sequence?

Exercise 3.8 Suppose the random variable X is uniformly distributed on $[0, 1]$. Let $Y = e^{-X}$ and $Z = e^X$.

(1) Compute $\mathrm{Cov}(Y, Z)$.
(2) Verify your answer using Monte-Carlo simulation.
(3) Mathematically prove that Y and Z are not independent.

CHAPTER 4

ARMA modeling

1. Moving Average models

Many important stationary processes can be written as a linear process.

DEFINITION 4.1. The time series $\{X_t\}$ is a *linear process* if it has the representation

$$X_t = \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j},$$

for all t, where $\{Z_t\} \sim \text{WN}(0, \sigma^2)$ and $\{\psi_j\}$ is a sequence of constants with $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$.

- If $\psi_j = 0$ for $j < 0$, then the process is called *causal*.
- For causal linear processes we write

$$X_t = \mathbf{\Psi}(B) Z_t,$$

with

$$\mathbf{\Psi}(B) = \psi_0 + \psi_1 B + \psi_2 B^2 + \dots.$$

Here, recall the definition of the backshift-operator

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

In definition 2.12 we defined the first-order moving average process. A straightforward generalisation of this process is given in the following definition.

DEFINITION 4.2. A time series $\{X_t\}$ is a *moving-average process of order q*, if

$$X_t = Z_t + \theta_1 Z_{t-1} + \dots + \theta_q Z_{t-q}$$

where $\{Z_t\} \sim \text{WN}(0, \sigma^2)$, $\theta_1, \dots, \theta_q \in \mathbb{R}$ and $\theta_q \neq 0$.

This is a causal linear process with

$$\psi_j = \begin{cases} \theta_j & \text{if} \quad 0 \leq j \leq q \\ 0 & \text{if} \quad j > q \end{cases},$$

where $\theta_0 \stackrel{\text{def}}{=} 1$. We write

$$X_t = \mathbf{\Theta}(B) Z_t,$$

with $\mathbf{\Theta}(B)$ the *moving average operator*

$$\mathbf{\Theta}(B) = 1 + \theta_1 B + \theta_2 B^2 + \dots + \theta_q B^q.$$

THEOREM 4.3. If $\{X_t\} \sim \text{MA}(q)$, then it is stationary and

$$\gamma(h) = \begin{cases} \sigma^2 \sum_{j=0}^{q-h} \theta_j \theta_{j+h} & |h| \leq q \\ 0 & |h| > q \end{cases}.$$

Moreover, if $\{X_t\}$ is a stationary time series with mean zero, for which $\gamma(h) = 0$ for $|h| > q$, then it can be represented as a $\text{MA}(q)$ process.

PROOF. We only give a proof of the first statement. For that, define $\theta_0 = 1$ and consider

$$\begin{aligned} \text{Cov}(X_t, X_{t+h}) &= \text{Cov}\left(\sum_{j=0}^{q} \theta_j Z_{t-j}, \sum_{k=0}^{q} \theta_k Z_{t+h-k} \right) \\ &= \sum_{j=0}^{q} \sum_{k=0}^{q} \theta_j \theta_k \text{Cov}(Z_{t-j}, Z_{t+h-k}) \end{aligned}$$

For fixed $j \in \{0, \ldots, q\}$ the covariance term is only nonzero if $k = h + j$. Hence

$$\text{Cov}(X_t, X_{t+h}) = \sum_{j=0}^{q} \theta_j \theta_{h+j} \sigma^2.$$

Finally note the terms with $j \in \{q - h + 1, \ldots, q\}$ are all equal to zero. □

A realisation of the process can easily be obtained using the commands

```
n <- 150
x <- arima.sim(list(ma=c(0.6,-0.3)),n)
```

In Figure 1 the simulated time-series is plotted together with its ACF. There is a slight difficulty in defining $\text{MA}(q)$ processes. To explain the issue, consider the following two $\text{MA}(1)$ time series $\{X_t\}$ and $\{Y_t\}$:

$$\begin{aligned} X_t &= Z_t + \frac{1}{5} Z_{t-1} \qquad \{Z_t\} \sim \text{IID } N(0, 25) \\ Y_t &= Z_t + 5 Z_{t-1} \qquad \{Z_t\} \sim \text{IID } N(0, 1). \end{aligned} \tag{9}$$

Note that $X_t | Z_t, Z_{t-1} \sim N(0, 26)$ and $Y_t | Z_t, Z_{t-1} \sim N(0, 26)$. Hence, simulating $\{X_t\}$ or $\{Y_t\}$ both yield a $\text{MA}(1)$-process with the same (joint) distribution. However, note that the parameter θ_1 is equal to $1/5$ and 5 in case of $\{X_t\}$ and $\{Y_t\}$ respectively. Let's see how this indeterminacy is dealt with within R. We simulate the data using `arima.sim`, where we explicitly specify $\{Z_t\}$ via the `rand.gen`-argument.

```
set.seed(10) # set random number generator
x <- arima.sim(list(ma=1/5),1000,rand.gen=function(n,...)
    rnorm(n,sd=5))
set.seed(10) # reset random number generator
```

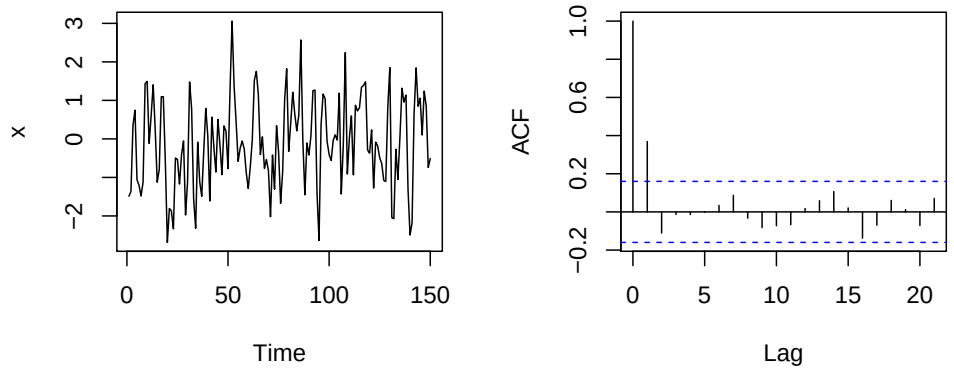

FIGURE 1. A realisation of length 150 of the process $X_t = Z_t + 0.6Z_{t-1} - 0.3Z_{t-2}$, where $\{Z_t\} \sim \text{IID } N(0,1)$.

```
y <- arima.sim(list(ma=5),1000,rand.gen=function(n,...)
    rnorm(n,sd=1))
```

Exercise 4.1 Make a plot with `x` and `y` overlaid. Explain what you see.

Now we fit a MA(1) model to both `x` and `y`. This makes sense, right? After all, the data were generated from this model.

```
sarima(x,0,0,1)
sarima(y,0,0,1)
```

Don't bother two much about the syntax here: just remember that MA(q) models can be fitted using the `sarima`-command. The first argument is a vector that contains the data, the 2nd and 3th argument should be zero, the 4th argument should be q. By "fitted" we mean that the parameters are estimated. By default, it is assumed that

$$X_t = Z_t + \theta Z_{t-1}, \quad \text{where} \quad \{Z_t\} \sim \text{IID } N(0, \sigma^2).$$

Hence, there are two parameters to be estimated: θ and σ^2. We don't go into details about the estimation method here, but discuss this later.
Here is part of the output with `x` as data:

```
Coefficients:
         ma1   xmean
      0.2425  0.0740
s.e.  0.0300  0.1946
```

For `y` we obtain similarly:

```
Coefficients:
         ma1   xmean
      0.2426  0.0693
s.e.  0.0300  0.1945
```

Note that the estimates are about the same. How come? The point is that we will always be concerned with the invertible solution of the MA-equation.

THEOREM 4.4. Let $\{X_t\} \sim \text{MA}(q)$, i.e. $X_t = \Theta(B)Z_t$ for a white-noise sequence $\{Z_t\}$. Then

$$Z_t = \sum_{j=0}^{\infty} \pi_j X_{t-j}$$

for a sequence $\{\pi_j\}$ such that $\sum_{j=0}^{\infty} |\pi_j| < \infty$ if and only if Θ has no roots on or inside the unit circle. In that case we speak of a *invertible* solution.

You can see why it is called invertible: the roles of $\{X_t\}$ and $\{Z_t\}$ are inverted.

Exercise 4.2 Show that $\{X_t\}$ as defined in (9) is invertible, while $\{Y_t\}$ is not.

To get some further intuition: from $X_t = Z_t + \theta Z_{t-1}$, we get

$$Z_t = -\theta Z_{t-1} + X_t.$$

Now we can iteratively apply this recursion to obtain

$$Z_t = \sum_{j=0}^{\infty} (-\theta)^j X_{t-j}.$$

Without being mathematically precise, it is conceivable that the right-hand-side is only summable if $|\theta| < 1$.

Exercise 4.3 Show that this reasoning is in agreement with Theorem 4.4.

2. Autoregressive models

DEFINITION 4.5. A time series $\{X_t\}$ follows an AR(p) model (autoregression model of order p), if there exists a white noise process $\{Z_t\}$ and coefficients $\phi_1, \ldots, \phi_p$ with $\phi_p \neq 0$ such that

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + Z_t$$

We write

$$\mathbf{\Phi}(B)X_t = Z_t,$$

with $\mathbf{\Phi}$ the autoregressive operator

$$\mathbf{\Phi}(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p.$$

Once we specify $\{Z_t\}$ and X_0, all other values X_t ($t \neq 0$) follow from the defining equation. So there exist many AR(p)-processes!. Usually, we impose

the additional assumption that the process is *stationary*. Then one natural question is when a stationary version exists.
To get a feeling for this, consider the AR(1)-model:

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

- If $|\phi| < 1$, then
$$X_t = \sum_{j=0}^{\infty} \phi^j Z_{t-j}.$$
satisfies the equation.
- On the other hand, if $|\phi| > 1$ then we can do the recursion in the other direction:
$$\begin{aligned} X_t &= \phi^{-1} X_{t+1} - \phi^{-1} Z_{t+1} \\ &= \phi^{-1} \left(\phi^{-1} X_{t+2} - \phi^{-1} Z_{t+2} \right) - \phi^{-1} Z_{t+1} \\ &= \dots \end{aligned}$$
So by repeatedly doing this recursion, it is conceivable that
$$X_t = -\sum_{j=1}^{\infty} \phi^{-j} Z_{t+j}$$
is also a stationary solution.

Hence we reach the preliminary conclusion that the stationary solution is given by

$$X_t = \begin{cases} \sum_{j=0}^{\infty} \phi^j Z_{t-j} & \text{if} \quad |\phi| < 1 \\ -\sum_{j=1}^{\infty} \phi^{-j} Z_{t+j} & \text{if} \quad |\phi| > 1 \end{cases}.$$

It is clear that no stationary solution exists when $|\phi| = 1$ (why?). Put differently, a stationary solution exists, when the equation $|\Phi(z)| = 0$ has no solutions with $|z| = 1$. In theorem 4.6 we see how this generalises to AR(p)-processes.
The solution for $|\phi| > 1$ is not useful for prediction, since it depends on the future: it is not a causal linear process. Hence, only for $|\phi| < 1$ there exists a (unique) causal solution to the auto-regressive defining equation.

THEOREM 4.6. An AR(p) process $\{X_t\}$ is *causal* (or a causal function of the process $\{Z_t\}$), i.e.,

$$X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j} \qquad \text{with} \quad \sum_{j=0}^{\infty} |\psi_j| < \infty,$$

if and only if the roots of the equation

$$\mathbf{\Phi}(z) = 1 - \phi_1 z - \dots - \phi_p z^p = 0$$

are outside the unit circle (so a root z always satisfies $|z| > 1$).

Now we see that we have two representations for a causal stationary autoregressive process:

$$\begin{aligned} \text{definition AR}(p) \qquad & \boldsymbol{\Phi}(B)X_t = Z_t \\ \text{causal linear process} \qquad & X_t = \boldsymbol{\Psi}(B)Z_t. \end{aligned}$$

Combining these, we get

$$\boldsymbol{\Phi}(B)\boldsymbol{\Psi}(B) = 1.$$

Hence, the coefficients $\{\psi_j\}$ satisfy

$$\left(1 - \phi_1 z - \phi_2 z^2 - \phi_p z^p\right)\left(1 + \psi_1 z + \psi_2 z^2 + \ldots\right) = 1$$

and we can solve for $\{\psi_j\}$ by reorganising coefficients. As an example, if $\{X_t\} \sim \text{AR}(1)$, then

$$\boldsymbol{\Psi}(\boldsymbol{z}) = \frac{1}{\boldsymbol{\Phi}(z)} = \frac{1}{1-\phi z} = \sum_{j=0}^{\infty} \phi^j z^j = \sum_{j=0}^{\infty} \psi_j z^j.$$

Hence we get

$$\psi_j = \phi^j.$$

This shows explicitly that a causal stationary AR(1) model can be represented as an infinite MA-model. So if we have a time series with non-vanishing autocorrelations over large lags, an AR model can provide a more parsimonious model (i.e. a model that uses less parameters) than a MA model.

EXAMPLE 4.7. In Figure 2 we show realisations of a AR(2)-model for various choices of its parameters ϕ_1 and ϕ_2. The corresponding ACF plots are in Figure 3.
From the figures it can be seen that a wide variety of dependence structures can be modelled, depending on the values of ϕ_1 and ϕ_2.

Estimation of AR(p) Models in `R` in R most easily done with the function

```
sarima(x,p,0,0,no.constant=TRUE)
```

Just the fit can be obtained by

```
sarima(x,p,0,0,no.constant=TRUE)$fit
```

This fits the model

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + Z_t,$$

assuming $\{Z_t\} \sim \text{IID}\, N(0, \sigma^2)$. For this model it is assumed that $\text{E}[X_t] = 0$. There is also a function `arima` for fitting autoregressive time series in `R`, but the advantage of `sarima` is that certain "diagnostic plots" are generated automatically (we explain later what this means).

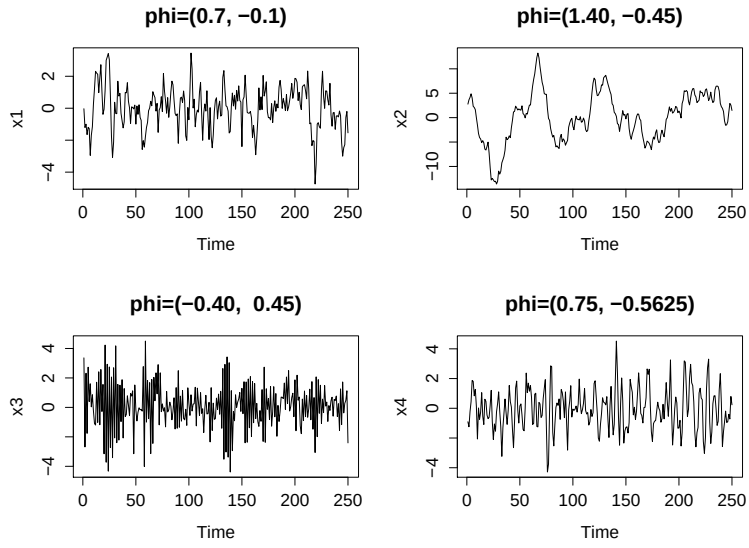

FIGURE 2. Realisations of a AR(2)-model for various choices of its parameters ϕ_1 and ϕ_2.

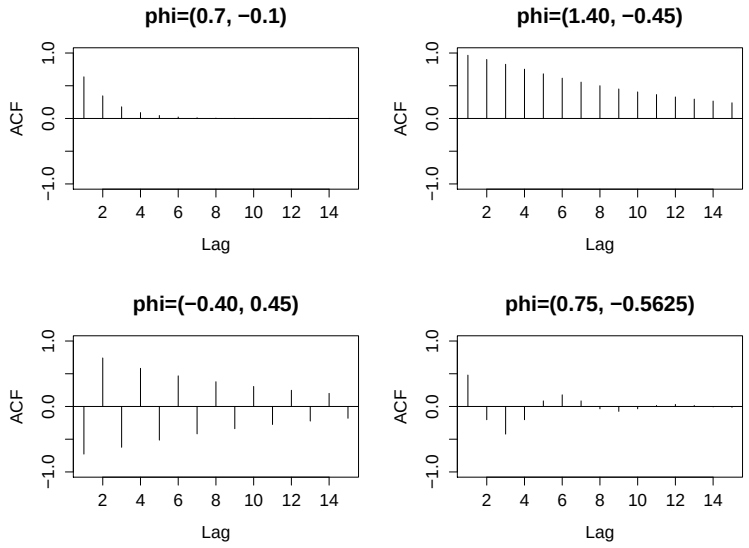

FIGURE 3. ACF plots corresponding to the time-series in Figure 2.

If you believe the time-series may have nonzero mean, then the parameters of the model

$$(X_t - \mu) = \phi_1(X_{t-1} - \mu) + \phi_2(X_{t-2} - \mu) + \ldots + \phi_p(X_{t-p} - \mu) + Z_t,$$

assuming $\{Z_t\} \sim$ IID $N(0, \sigma^2)$, can be estimated using

```
sarima(x,p,0,0)$fit
```

As we have $\mathrm{E}[X_t] = \mu$, the estimate for μ is called `xmean` in the session output.
Computing the ACF of a AR(p)-model is more involved than for a MA(q) model. Exercise 8 reveals how to derive the ACF in case $p = 2$.

3. Residual analysis and diagnostic plots

For parameter estimation, it is typically assumed that the white noise sequence $\{Z_t\}$ is IID $N(0, \sigma^2)$. How do we know that this assumption is satisfied? The answer is that we never know for sure, but this applies to virtually all statistical models. We can however check whether the assumption is *reasonable* for the given data.
Suppose we have fitted a AR(p)-model to a particular dataset. The estimates are denoted by $\hat{\phi}_1, \ldots, \hat{\phi}_p, \hat{\sigma}$. The *residuals* are defined by

$$\widehat{Z}_t = X_t - \widehat{\phi}_1 X_{t-1} - \widehat{\phi}_2 X_{t-2} - \cdots - \widehat{\phi}_p X_{t-p}.$$

The should "resemble" $\{Z_t\}$. The *standardised residuals* are defined by

$$e_t = \frac{\widehat{Z}_t}{\hat{\sigma}}$$

Now *residual analysis* means that you check whether either the residuals or standardised residuals are like a realisation of a IID sequence with the Normal distribution. Clearly, the techniques discussed in sections 6 and 7 in Chapter 3 can be used for this purpose.
Whereas, for autoregressive models the way to define residuals is fairly straightforward, this is not the case in general. As an example, for a moving average model, their definition is somewhat more involved. One way to go is to exploit invertibility and to write $Z_t = \sum_{j=0}^{\infty} \pi_j X_{t-j}$. Now the coefficients $\{\pi_j\}$ are related to the coefficients $\theta_1, \ldots, \theta_q$ in the polynomial $\boldsymbol{\Theta}$ that defines the moving average process. Hence, using this relation, estimated values $\hat{\theta}_1, \ldots, \hat{\theta}_q$ can be used to obtain estimated coefficients $\{\hat{\pi}\}_j$. Then the residuals are given by $\hat{Z}_t = \sum_{j=0}^{\infty} \hat{\pi}_j X_{t-j}$. In practice, this summation needs to be truncated to $j = t$, as only observations $x_0, \ldots, x_n$ are available.

4. An example

Suppose we are given the time series data in Figure 4. Actually, these were simulated as follows:

```
n <- 200
x <- arima.sim(list(ar=c(0.6,-0.3)),n,sd=0.3)
```

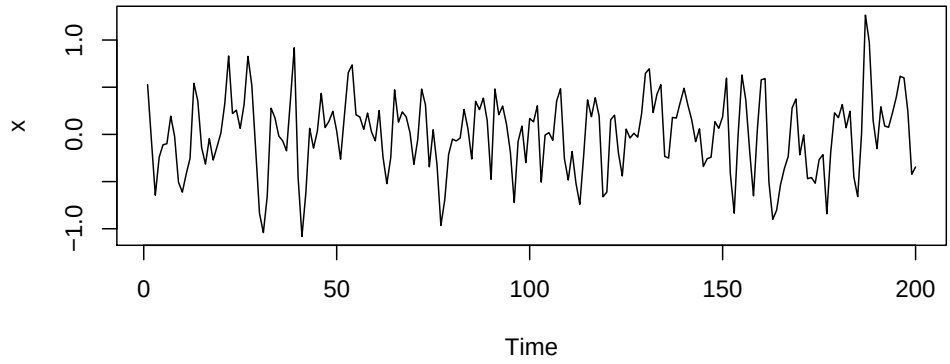

FIGURE 4. A realisation of length 200 of a AR(2)-model with $\phi_1 = 0.6$ and $\phi_2 = -0.3$.

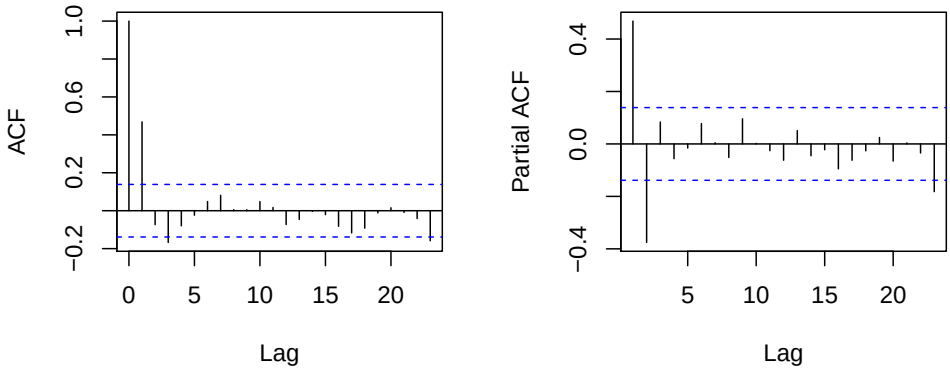

FIGURE 5. A realisation of length 200 of a AR(2)-model with $\phi_1 = 0.6$ and $\phi_2 = -0.3$.

In Figure 5 we show the ACF and PACF for this dataset. Based on the ACF, a MA(3) model appears reasonable. Based on the PACF on the other hand, a AR(2)-model appears reasonable (note that this matches with the way the data were generated). For illustration purpose, we fit a AR(p) model for $p \in \{1, 2, 3\}$. For fitting the AR(1) model, we use the code

```
sarima(x,1,0,0,no.constant=T)
```

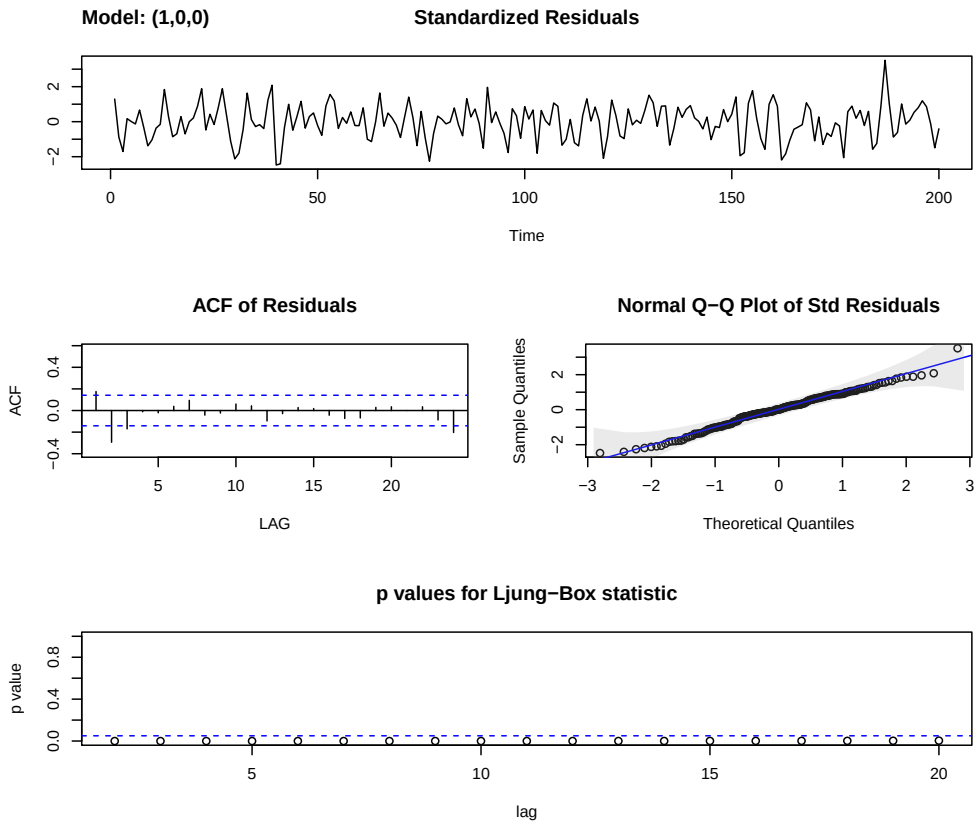

FIGURE 6. Diagnostic plots for AR(1)-fit.

(it is easy to see how to adjust the code to fit an AR(p)-model. In Figure 6 the diagnostic plots (constructed after fitting the model) are shown. Clearly, the ACF of the residuals is significant at lag 2. Moreover, from the Ljung-Box statistics it is clear that the residuals are not IID. We conclude that the AR(1) model is not appropriate for modelling the data. In figures 7 and 8 we show diagnostic plots for AR(2) and AR(3) models respectively. Both models seem fine: there is no reason to reject either the fitted AR(2) or AR(3) model. Then which should we prefer? This is known as the *model selection* problem. We prefer a model with least parameters. That is, the model that is most *parsimonious*.

For model comparison, there are many methods available. Here we restrict attention to two methods, for simplicity we focus on model order selection for AR(p)-models (for other statistical models, such as MA(q) models, similar methods exist).

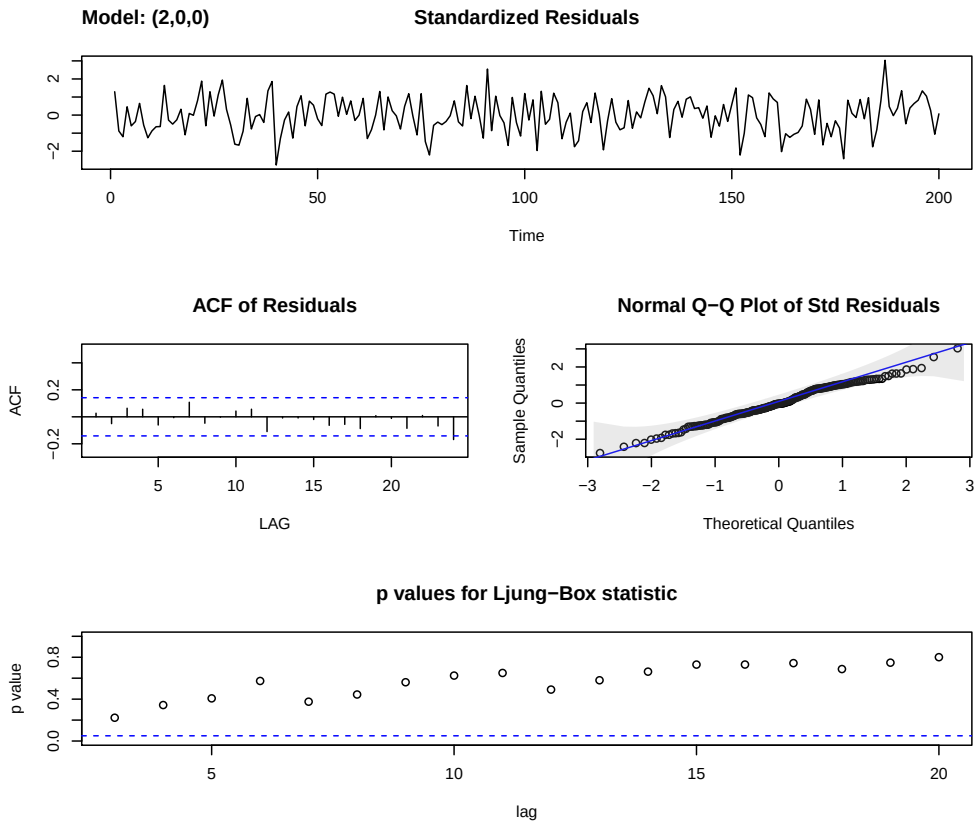

FIGURE 7. Diagnostic plots for AR(2)-fit.

(1) Compare the $\{\psi_j\}$-weights in the models, where we use the representation $X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}$ (recall that we assume the time-series to be stationary and causal).
(2) Take the model (that is, the value of p) that minimises the *AIC* (Akaike Information Criterion)

$$\mathrm{AIC} = -2 \log L(\phi_1, \ldots, \phi_p, \sigma) + 2(p+1)$$

or its corrected version

$$\mathrm{AICC} = -2 \log L(\phi_1, \ldots, \phi_p, \sigma) + \frac{2(p+1)n}{n-p-2}$$

Here L denotes the likelihood function, which can be interpreted as the probability of the data under the assumed model (the likelihood is defined more precisely in Section 1 in Chapter 7).

The idea behind AIC and AICC is that we balance the number of parameters in the model and the minimum value of the negative likelihood. Clearly, the more parameters, the lower value of the

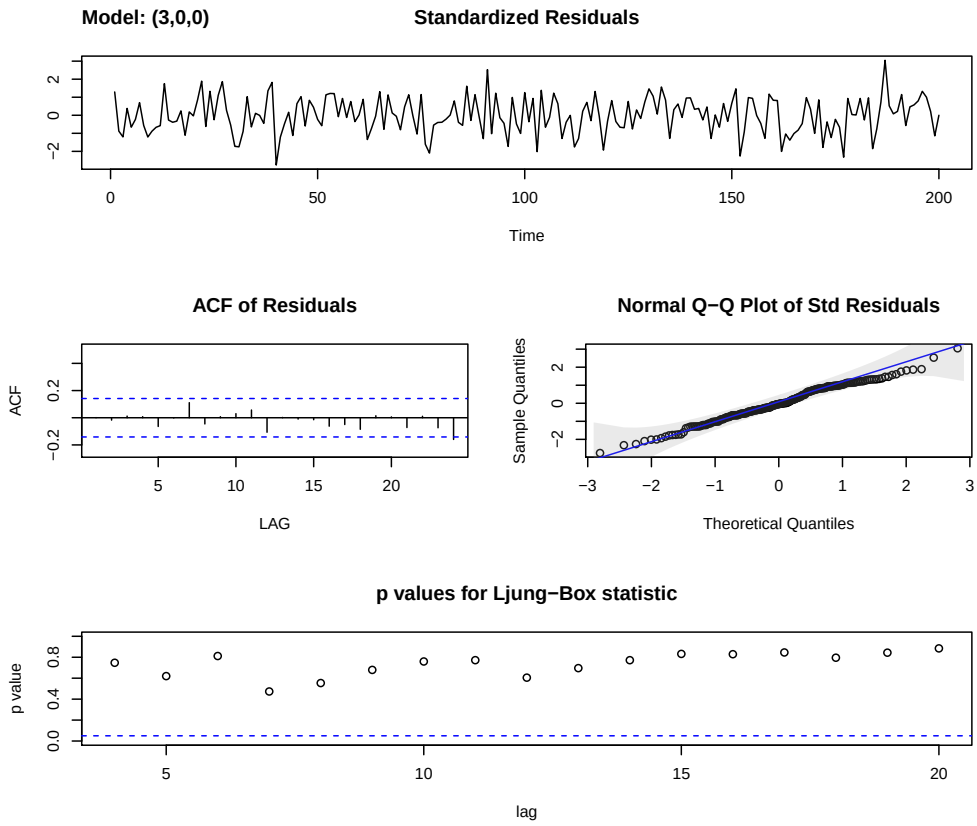

FIGURE 8. Diagnostic plots for AR(3)-fit.

negative likelihood that can be attained. The justification of AIC is outside the scope of this course. AICC is thought to give improved accuracy over AIC (this is a sloppy statement, as it requires to specify what we mean by accuracy in this context).

The ψ_j weights and AIC can be obtained with the following code:

```
out2 <- sarima(x,2,0,0,no.constant=T)
out2_fit <- out2$fit
ARMAtoMA(ar=out2_fit$coef,lag.max=10)
out2_fit$aic
```

This results in

Model	AIC	ψ_1	ψ_2	ψ_3	ψ_4
AR(2)	133.55	0.65	0.04	-0.22	-0.16
AR(3)	134.27	0.68	0.03	-0.19	-0.09

It can be seen that the differences between the AR(2) and AR(3) fits are small, but based on AIC we would prefer the AR(2)-model.
Finally, for completeness, we also fitted a MA(3) model:

```
Coefficients:
         ma1     ma2      ma3
      0.6744  0.0285  -0.1614
s.e.  0.0713  0.0912   0.0746

sigma^2 estimated as 0.1099: log likelihood = -63.23, aic = 134.46
```

Note that the estimates are in agreement with our earlier findings.

5. ARMA models

ARMA models are time-series models that include MA and AR models as special cases.

DEFINITION 4.8. The time series $\{X_t\}$ is an *ARMA(p,q) process* if there exist polynomials

$$\mathbf{\Theta}(z) = 1 + \theta_1 z + \cdots + \theta_q z^q$$

and

$$\mathbf{\Phi}(z) = 1 - \phi_1 z - \cdots - \phi_p z^p$$

and a white noise process $\{Z_t\}$ such that

$$\Phi(B)X_t = \Theta(B)Z_t. \tag{10}$$

We implicitly assume that $\theta_q \neq 0$ and $\phi_p \neq 0$. Equation (10) specifies a relation between $\{X_t\}$ and $\{Z_t\}$. Note that if $\theta_1 = \cdots = \theta_q = 0$, then $\{X_t\} \sim \text{AR}(p)$ and that if $\phi_1 = \cdots = \phi_p = 0$, then $\{X_t\} \sim \text{MA}(q)$.
As a MA process is not guaranteed to be invertible and an AR process is not guaranteed to be causal and stationary, we may ask whether a causal stationary invertible solution to the ARMA model equation exists.
Another problem that may occur in defining a ARMA process is *parameter redundancy*. This is easily explained by an example.

EXAMPLE 4.9. Suppose $\{Z_t\} \sim \text{WN}(0, \sigma^2)$ and set $X_t = Z_t$. Then obviously $\{X_t\}$ is a s a white noise process. Now if we multiply the equation $X_t = Z_t$, and subtract this equation after first multiplying it with 0.5 we get the equation

$$X_t - 0.5X_{t-1} = Z_t - 0.5Z_{t-1}.$$

One may be tempted to identify this process as an ARMA$(1,1)$ process, while clearly $\{X_t\}$ is white-noise. The problem is caused by *parameter redundancy* which can be identified by noting that $\mathbf{\Theta}(z)$ and $\mathbf{\Phi}(z)$ have common factors.

THEOREM 4.10. A *causal and stationary* ARMA(p,q) process $\{X_t\}$ exists if and only if all roots of $\mathbf{\Phi}$ are outside the unit circle. In that case we can write

$$X_t = \mathbf{\Psi}(B)Z_t \qquad \text{with} \quad \mathbf{\Psi}(B) = \frac{\mathbf{\Theta}(B)}{\mathbf{\Phi}(B)}$$

The solution is *unique* if Θ and Φ have no common factors.

THEOREM 4.11. An ARMA(p,q) process $\{X_t\}$ is invertible i.e.,

$$Z_t = \sum_{j=0}^{\infty} \pi_j X_{t-j} \qquad \text{with} \sum_{j=0}^{\infty} |\pi_j| < \infty,$$

if and only if all roots of $\mathbf{\Theta}$ are outside the unit circle.

Fitting with `R` returns parameters such that the time series is invertible and causal.

EXAMPLE 4.12. Consider the process

$$X_t = 0.4X_{t-1} + 0.45X_{t-2} + Z_{t-1} + 0.25Z_{t-2} + Z_t.$$

We check for parameter redundancy, causality and invertibility.
For this, first write the defining relation as $\mathbf{\Phi}(B)X_t = \mathbf{\Theta}(B)Z_t$ with with

$$\mathbf{\Phi}(B) = 1 - 0.4B - 0.45B^2 = (1 + 0.5B)(1 - 0.9B)$$

and

$$\mathbf{\Theta}(B) = 1 + B + 0.25B^2 = (1 + 0.5B)^2$$

Clearly, there is parameter redundancy: the common term $1 + 0.5B$ can be cancelled. After this, we see that we are left with an ARMA$(1,1)$-process with

$$\mathbf{\Phi}(B) = 1 - 0.9B \qquad \text{and} \quad \mathbf{\Theta}(B) = (1 + 0.5B)$$

Now

- $\mathbf{\Phi}(B) = 0 \Rightarrow B = 10/9$, so the process is causal;
- $\mathbf{\Theta}(B) = 0 \Rightarrow B = -2$, so process is invertible.

To find the linear process representation, recall that

$$\mathbf{\Phi}(z)\mathbf{\Psi}(z) = \mathbf{\Theta}(z)$$

so that

$$(1 - 0.9z)(\psi_0 + \psi_1 z + \psi_2 z^2 + \ldots) = (1 + 0.5z).$$

Now matching of coefficients gives $\psi_0 = 1$, $\psi_1 = 1.4$ and

$$\psi_j = 0.9\psi_{j-1} = 1.4(0.9)^{j-1} \qquad j > 1.$$

We conclude that $\{X_t\}$ can be represented as

$$X_t = Z_t + 1.4\sum_{j=1}^{\infty} (0.9)^{j-1} Z_{t-j}.$$

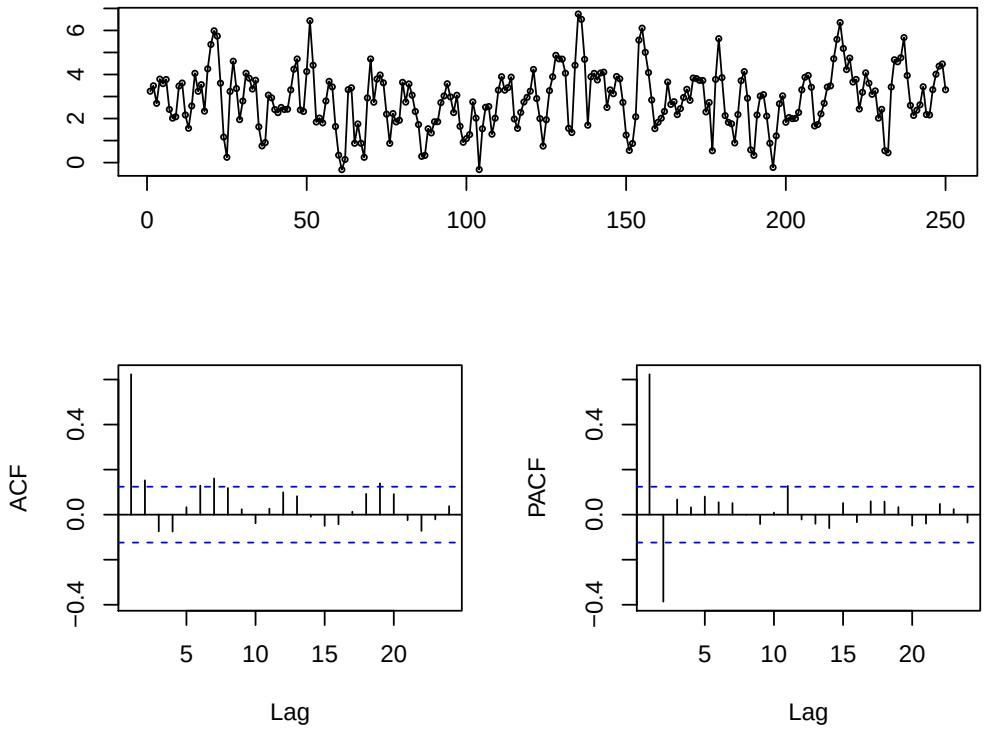

FIGURE 9. Data for Exercise 4.

6. Exercises from old exams

Exercise 4.4 See Figure 9. Give 2 reasonable ARMA-processes for modelling these data. Motivate your choice.

Exercise 4.5 Suppose $\{X_t\}$ is a stationary mean-zero causal AR(1) process. Hence

$$X_t = \phi X_{t-1} + Z_t \qquad \text{where} \qquad \{Z_t\} \sim \mathrm{WN}(0, \sigma_z^2),$$

where $|\phi| < 1$. Define

$$Y_t = X_t + W_t$$

where $W_t \sim \mathrm{WN}(0, \sigma_w^2)$. Assume $\mathrm{E}[W_s Z_t] = 0$ for all s and t.

(1) Define $U_t := Y_t - \phi Y_{t-1}$. Compute $\mathrm{Cov}(U_t, U_{t+h})$ for all $h \geq 1$.
(2) It is claimed that $Y_t \sim \mathrm{ARMA}(p, q)$. Is this true? If so, what are p and q?

Exercise 4.6 Suppose $Z_t \sim \mathrm{WN}(0, \sigma^2)$ and $\psi > 0$. Verify for which values of ψ the AR(2)-equation

$$X_t = \psi X_{t-2} + Z_t$$

admits a causal, stationary and invertible solution.

Exercise 4.7
Suppose $\{Z_t\} \sim \mathrm{WN}(0, 1)$ and consider the time-series $\{X_t\}$ defined by

$$X_t = Z_t + \theta Z_{t-1}.$$

(1) For which values of θ is $\{X_t\}$ a causal time-series?
(2) For which values of θ is $\{X_t\}$ an invertible time-series?
(3) Show that $|\rho_X(1)| \leq 1/2$ for all $\theta \in \mathbb{R}$. For which values of θ is $|\rho_X(1)|$ maximal?
(4) Simulate a realisation of length 300 of the time-series $\{X_t\}$. Take $\theta = 3$ and assume $\{Z_t\}$ is a sequence of independent random variables, each with the standard normal distribution Next estimate the parameters of the model using these simulated data and the `R`-command `sarima`. Explain the estimates. Are these correct?

Exercise 4.8 Suppose $\{Z_t\} \sim \mathrm{WN}$ and that the time-series $\{X_t\}$ is defined by the recursive relation

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + Z_t. \tag{11}$$

You may assume $\{X_t\}$ is stationary and $\mathrm{E}[X_t] = 0$.

(1) Give a criterion for ϕ_1 and ϕ_2 such that the time-series is stationary and causal.
(2) Take an integer $h > 0$. Show that the ACF of $\{X_t\}$, $h \mapsto \rho(h)$ satisfies
$$\rho(h) = \phi_1 \rho(h-1) + \phi_2 \rho(h-2).$$
Hint: multiply both sides in equation (11) with X_{t-h} and take the expectation on both sides. Note that Z_t is independent of X_{t-h} for $h = 1, 2, \ldots$.
(3) Show that $\rho(1) = \phi_1/(1 - \phi_2)$.

Exercise 4.9 Suppose $\{W_t\} \sim \mathrm{WN}(0, \sigma^2)$. Define the causal stationary time series $\{X_t\}$ by

$$X_t = \theta X_{t-2} + W_t, \qquad X_0 = 0$$

with $|\theta| < 1$. You can assume $E[X_t] = 0$ for all t. Denote the ACF of the time series $\{X_t\}$, evaluated at lag h, by $\rho(h)$.

(1) Show that $\rho(h) = \theta \rho(h-2)$.
(2) Show that $\rho(1) = 0$.
(3) Express $\rho(4)$ in terms of θ.

CHAPTER 5

Conditional expectation and best prediction

Conditional expectation is a very useful concept in probability theory. Unfortunately, its mathematical definition requires knowledge of *measure theory* (actually, this theory is necessary to do any sort of "advanced" probability theory, thereby leaving the world of throwing dice, coins, etc.). The difficulty resides in the fact that when Y has a "continuous" distribution, defining

$$\mathrm{E}[X \mid Y = y]$$

is problematic since $\mathrm{P}(Y = y) = 0$ for all y.

1. Conditional expectation in the discrete case

In case the joint distribution of (X, Y) is "discrete", all is elementary.

DEFINITION 5.1 (Conditional probability: discrete case.). The *conditional probability mass function* of Y given $X = x$ is defined by

$$p_{Y|X}(y \mid x) = \mathrm{P}(Y = y \mid X = x)$$

whenever $\mathrm{P}(X = x) > 0$.

From the definition it follows that

$$p_{Y|X}(y \mid x) = \frac{\mathrm{P}(X = x, Y = y)}{\mathrm{P}(X = x)}.$$

If X and Y are independent, then

$$p_{Y|X}(y \mid x) = \mathrm{P}(Y = y)\,.$$

EXAMPLE 5.2. Suppose we throw a dice twice. Let S and M denote the sum and maximum of the outcomes respectively. The joint distribution of (S, M) is given in Table 1. The values of $a \mapsto \mathrm{P}(S = a \mid M = 3)$ can be computed from the column with numbers in italics. Applying the definition of conditional probability we get

a	4	5	6
$\mathrm{P}(S = a \mid M = 3)$	$\frac{2}{5}$	$\frac{2}{5}$	$\frac{1}{5}$

DEFINITION 5.3 (Conditional expectation: discrete case.). The *conditional expectation* of Y given X is denoted by $\mathrm{E}[Y \mid X = x]$ and defined as

$$\mathrm{E}[Y \mid X = x] = \sum_y y\, p_{Y|X}(y \mid x).$$

$$\mathrm{P}(S = a, M = b)$$

	b					
a	1	2	3	4	5	6
2	1/36	0	*0*	0	0	0
3	0	2/36	*0*	0	0	0
4	0	1/36	*2/36*	0	0	0
5	0	0	*2/36*	2/36	0	0
6	0	0	*1/36*	2/36	2/36	0
7	0	0	*0*	2/36	2/36	2/36
8	0	0	*0*	1/36	2/36	2/36
9	0	0	*0*	0	2/36	2/36
10	0	0	*0*	0	1/36	2/36
11	0	0	*0*	0	0	2/36
12	0	0	*0*	0	0	1/36

TABLE 1. Joint distribution of (S, M).

The following properties can be verified by direct computation:

(1) If X and Y are independent, then

$$\mathrm{E}[Y \mid X = x] = \mathrm{E}[Y]\,.$$

(2) Expectation of conditional expectation.

$$\mathrm{E}[Y] = \sum_x \mathrm{E}[Y \mid X = x]\,\mathrm{P}(X = x)\,.$$

That is,

$$\mathrm{E}[Y] = \mathrm{E}[\mathrm{E}[Y \mid X]]\,.$$

2. Conditional expectation in the general case

As alluded to in the introductory alinea of this chapter, a precise mathematical definition of conditional expectation is outside the scope of this course. However, the intuitive way to think about it is just as in Section 2.3 of Shreve [2004]. In this book, the expectation of the random variable Y conditional on all information at time n is denoted by $\mathbb{E}_n$. Here, we will use slightly different notation. Denote the information set at time t by $\mathcal{F}_t$. Intuitively, this information consists of all knowledge on $X_1, \ldots, X_t$. The conditional expectation of Y, conditional on $\mathcal{F}_t$, is then denoted by $\mathrm{E}[Y \mid \mathcal{F}_t]$. In Theorem 2.3.2 of Shreve [2004] five fundamental properties of conditional expectation are stated. We will need the first 4 of those, which are given by

(1) **Linearity.** For all constants c_1 and c_2, we have

$$\mathrm{E}[c_1X + c_2Y \mid \mathcal{F}_t] = c_1\mathrm{E}[X \mid \mathcal{F}_t] + c_2\mathrm{E}[Y \mid \mathcal{F}_t].$$

(2) **Taking out what is known.** If X depends on $\mathcal{F}_t$, then

$$\mathrm{E}[XY \mid \mathcal{F}_t] = X\mathrm{E}[Y \mid \mathcal{F}_t].$$

(3) **Iterated conditioning.** If $0 \leq s \leq t$

$$\mathrm{E}[\mathrm{E}[X \mid \mathcal{F}_t] \mid \mathcal{F}_s] = \mathrm{E}[X \mid \mathcal{F}_s].$$

(4) **Independence.** If X does not depend on $\mathcal{F}_t$, then

$$\mathrm{E}[X \mid \mathcal{F}_t] = \mathrm{E}[X].$$

In fact, the second property can be generalised to: if X is independent of $\mathcal{F}_t$, then for "nicely behaved" functions f and g

$$\mathrm{E}[f(X)\, g(Z) \mid \mathcal{F}_t] = f(X)\mathrm{E}[g(Z) \mid \mathcal{F}_t].$$

A particular case of iterated conditioning is that

$$\mathrm{E}[X] = \mathrm{E}[\mathrm{E}[X \mid \mathcal{F}_t]]. \tag{12}$$

3. Best Predictor

In Section 4 of chapter 2 we discussed the best linear predictor. In some cases the best predictor can be calculated in closed form, thereby lifting the linearity assumption.

Definition 5.4. Let $h \geq 1$. The *best predictor* for X_{t+h} is defined by

$$P_tX_{t+h} = \mathrm{E}[X_{t+h} \mid \mathcal{F}_t].$$

The *h-step ahead prediction error* is defined by

$$e_t(h) = X_{t+h} - P_tX_{t+h}.$$

This definition is motivated by the following result

Lemma 5.5. Suppose that $X_1, X_2, \ldots$ is a sequence of random variables with $\mathrm{E}\left[X_i^2\right] < \infty$ and $\mathrm{E}[X_i] = \mu$. The random variable $f(X_1, \ldots, X_t)$ that minimises

$$\mathrm{E}\left[(X_{t+h} - f(X_1, \ldots, X_t))^2\right]$$

is given by

$$f(X_1, \ldots, X_t) = \mathrm{E}[X_{t+h} \mid \mathcal{F}_t].$$

To see this, go through the steps of the following exercise (taken from Brockwell and Davis [2002], chapter 1, exercise 1.1).

Exercise 5.1 Let X and Y be two random variables. Assume $\mathrm{E}[Y] = \mu$ and $\mathrm{E}\left[Y^2\right] < \infty$.

(a) Show that the constant c that minimises $\mathrm{E}\left[(Y-c)^2\right]$ is given by $c = \mu$.

(b) Deduce that the random variable $f(X)$ that minimises $\mathrm{E}\left[(Y - f(X))^2 \mid X\right]$ is $f(X) = \mathrm{E}[Y \mid X]$.
Hint: use part (a) with $c = f(X)$ *and* $\mu = \mathrm{E}[Y \mid X]$.

(c) Deduce that the random variable $f(X)$ that minimises $\mathrm{E}\left[(Y - f(X))^2\right]$ is also $f(X) = \mathrm{E}[Y \mid X]$.
Hint: equation (12).

EXAMPLE 5.6. If $\{X_t\} \sim \mathrm{AR}(p)$, then

$$\begin{aligned} X_{t+1} &= \phi_1 X_t + \phi_2 X_{t-1} + \ldots, \phi_p X_{t+1-p} + Z_{t+1} \\ &= \sum_{i=1}^{p} \phi_i X_{t+1-i} + Z_{t+1}. \end{aligned}$$

Hence

$$P_t X_{t+1} = \mathrm{E}[X_{t+1} \mid \mathcal{F}_t] = \sum_{i=1}^{p} \phi_i X_{t+1-i}.$$

and

$$e_t(1) = X_{t+1} - P_t X_{t+1} = Z_{t+1}.$$

If we assume $\{Z_t\} \sim N(0, \sigma^2)$, then a 95% *prediction interval* is given by

$$P_t X_{t+1} - 1.96\sigma \leq X_{t+1} \leq P_t X_{t+1} + 1.96\sigma$$

This follows from $e_t(1)/\sigma \sim N(0,1)$. In practice, σ is estimated from the data (with e.g. maximum likelihood), which induces extra uncertainty (so wider prediction intervals). This is usually neglected.

EXAMPLE 5.7. If $\{X_t\} \sim \mathrm{AR}(p)$, then

$$X_{t+2} = \phi_1 X_{t+1} + \phi_2 X_t + \ldots, \phi_p X_{t+2-p} + Z_{t+2}$$

and the best 2-step ahead predictor is given by

$$P_t X_{t+2} = \mathrm{E}[X_{t+2} \mid \mathcal{F}_t] = \phi_1 P_t X_{t+1} + \sum_{i=2}^{p} \phi_i X_{t+2-i}$$

The two step ahead forecast error is given by

$$\begin{aligned} e_t(2) &= X_{t+2} - P_t X_{t+2} \\ &= \phi_1 (X_{t+1} - P_t X_{t+1}) + Z_{t+2} \\ &= \phi_1 Z_{t+1} + Z_{t+2}. \end{aligned}$$

Note that $\mathrm{Var}(e_t(2)) \geq \mathrm{Var}(e_t(1))$.

The preceding examples can be extended to general h-step ahead forecasts.

EXAMPLE 5.8. If $\{X_t\} \sim \text{MA}(1)$, then

$$X_{t+1} = Z_{t+1} + \theta_1 Z_t.$$

Hence

$$P_t X_{t+1} = \text{E}[X_{t+1} \mid \mathcal{F}_t] = \theta_1 Z_t.$$

and the one step ahead forecast error is given by

$$e_t(1) = X_{t+1} - P_t X_{t+1} = Z_{t+1}$$

Since we always consider *invertible* MA-processes, we have

$$Z_t = \sum_{j=0}^{\infty} \pi_j X_{t-j}$$

EXAMPLE 5.9. If $\{X_t\} \sim \text{MA}(1)$, then

$$X_{t+2} = Z_{t+2} + \theta_1 Z_{t+1}.$$

Hence

$$P_t X_{t+2} = \text{E}[X_{t+2} \mid \mathcal{F}_t] = 0.$$

and

$$e_t(2) = X_{t+2} - P_t X_{t+2} = Z_{t+2} + \theta_1 Z_{t+1}$$

In fact, for $\ell \geq 2$:

$$\begin{aligned} P_t X_{t+\ell} &= \text{E}[X_{t+\ell} \mid \mathcal{F}_t] \\ &= \text{E}[Z_{t+\ell} + \theta_1 Z_{t+\ell-1} \mid \mathcal{F}_t] = 0. \end{aligned}$$

The *ℓ step ahead forecast error* is given by

$$e_t(\ell) = X_{t+\ell} - P_t X_{t+\ell} = Z_{t+\ell} + \theta_1 Z_{t+\ell-1}.$$

3.1. Prediction for ARMA models in R. Once an ARMA model has been fitted, it is quite easy to obtain forecasts in R. In fact, it can be done at once. Let's generate data from a AR(1) model, exhibiting positive correlation:

```
x <- arima.sim(list(ar=0.8),50)
```

A plot of the data with 10 steps ahead predictions is obtained from

```
sarima.for(x, 10, 1, 0, 0)
```

Here, indeed a AR(1)-model is assumed. If the data are in `x` and we wish to predict n steps ahead assuming a ARMA(p, q)-model, then the correct syntax is `sarima.for(x,n,p,0,q)`. In Figure 1 we show the resulting plot.

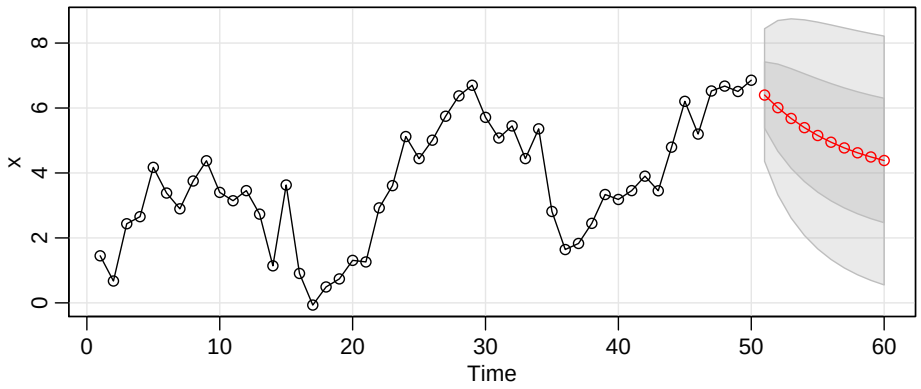

FIGURE 1. Ten-step ahead forecast, assuming the data can be modelled by a AR(1) time-series model. The shaded areas are determined by the best predictor ± 1 and ± 2 its standard deviation.

4. Exercises from old exams

Exercise 5.2 On brightspace you can find the files `x1.txt` en `x2.txt` Take the final number of your student-number.
If it is equal to 1, 2, 3, 4 or 5, then take the data from the file `x1.txt`.
If it is equal to 6, 7, 8, 9 or 0, then take the data from the file `x2.txt`.

Read the data. You can use the R-command `scan` or `read.table` for this purpose. Save your dataset into a vector `x1` or `x2`, depending on whether you are working with `x1.txt` or `x2.txt`.

(1) Plot the time series and make figures of the ACF and PACF.
(2) Would you model your data with a AR(p)-model or a MA(q)-model? Specify and explain which choice for either p or q you think is appropriate.
(3) Fit the model chosen at (b) and verify whether the residuals can be assumed to be white-noise. Do you think the residuals can be modeled with the normal distribution?
(4)
- In case you are working with dataset `x1`: choose an ARMA$(1,1)$-model.
- In case you are working with dataset `x2`: choose an ARMA$(2,2)$-model.

Fit the model and verify whether the residuals can be considered white noise.
(5) A causal ARMA process $\{X_t\}$ can be written as $X_t = \sum_{j\geq 0} \psi_j Z_{t-j}$. Compare the first 10 $\{\psi_j\}$-weights of the fitted models under (c) and (d) using a figure in which $j \mapsto \psi_j$ is depicted. Do you think the two fitted models are completely different?

Hint: you can use the command `ARMAtoMA`.

(6) Use the command

```
fit <- auto.arima(x1,d=0,trace=TRUE,ic='aic')
```

or

```
fit <- auto.arima(x2,d=0,trace=TRUE,ic='aic')
```

depending on which dataset you are analysing. You need the `forecast` library for this command. It searches for a good fitting ARMA model using AIC ("Akaike's Information Criterion"). Using this criterion, the model with the smallest value of AIC is to be preferred. Which model is optimal in terms of AIC for your dataset? Write down the equation(s) for your fitted model.

To help you interpreting `R`*'s output: the function* `auto.arima` *returns the best fitting ARIMA-model. An* $ARIMA(p,0,q)$ *model is the same as an* $ARMA(p,q)$ *model.*

(7) Plot the time-series together with the bestpredictor for the next 20 time steps ahead (using the model derived in part (f)).

Exercise 5.3 Consider a causal invertible ARMA$(1,1)$-process:

$$X_t = \phi X_{t-1} + Z_t + \theta Z_{t-1}$$

with $\{Z_t\} \sim \text{IID}(0,1)$. We can write $Z_t = \sum_{j\geq 0} \pi_j X_{t-j}$ for summable weights $\{\pi_j\}$.

(1) Show that $P_n X_{n+1}$, the one-step-ahead forecast, can be written as $\sum_{j\geq 0} \beta_j X_{n-j}$ where

$$\beta_j = \begin{cases} \phi + \theta\pi_0 & \text{if } j = 0 \\ \theta\pi_j & \text{if } j > 0 \end{cases}.$$

(2) Show that $e_n(1)$, the one-step-ahead forecast error, has variance equal to 1.

CHAPTER 6

GARCH modeling

1. Introduction

GARCH is an abbreviation to *Generalised Autoregressive Conditionally Heteroscedastic*. It constitutes a class of time-series models that was proposed by the economist Robert Engle in the eighties and extended by Tim Bollerslev a couple of years later. These models, of which many variations exist, are very popular for volatility modelling in finance. The starting point is the log-return of a financial derivative or index. The log-return at time t will be denoted by X_t.
We will write $\mathcal{F}_t$ for the information flow up till time t. This is sometimes written as

$$\mathcal{F}_t = \sigma(X_1, \ldots, X_t),$$

where σ refers to the notion of a σ-algebra. Understanding this requires measure theory and is outside the scope of these lecture notes. Whereas ARMA modelling is used for specifying $\mathrm{E}[X_t \mid \mathcal{F}_{t-1}]$, GARCH modelling is used for specifying $\mathrm{Var}(X_t \mid \mathcal{F}_{t-1})$. Usually, the *volatility* at time t, denoted by σ_t, is defined as the conditional standard-deviation of the log-return at time t, i.e.

$$\sigma_t = \sqrt{\mathrm{Var}(X_t \mid \mathcal{F}_{t-1})}.$$

It has been observed empirically that many time-series data share common properties, these are known as *stylised facts (or features) of financial returns*. These include:

- *Volatility clustering*: volatility may be high for certain subsequent times and low in other periods.
- *Non-normality* and *fat (heavy) tails*: the marginal distribution of the returns is not normal and decays slowly to zero in the tails (much slower than the normal distribution).
- *Nonlinear dependence*: while the ACF of the returns indicates white noise, the returns are dependent (recall that correlation only measures *linear* dependence).

Kurtosis is a measure of the thickness/fatness/heaviness of the tails of a probability density.

DEFINITION 6.1. For a random variable Y with expectation μ its *kurtosis* is defined by

$$\kappa_Y = \frac{\mathrm{E}\left[(Y-\mu)^4\right]}{(\mathrm{E}\left[(Y-\mu)^2\right])^2}.$$

The *excess kurtosis* of Y is defined by $\kappa_Y - 3$.

The latter definition is motivated by the fact that $\kappa_Y = 3$ if $Y \sim N(0,1)$. The excess kurtosis for a number of well known distributions, along with their density, is depicted in Figure 1.
A natural estimator for the kurtosis is given by the sample kurtosis.

DEFINITION 6.2. The *sample kurtosis* of a sample $Y_1, Y_2, \ldots, Y_n$ is defined by

$$\hat{\kappa}_n = \frac{\frac{1}{n}\sum_{i=1}^n (Y_i - \bar{Y}_n)^4}{\left(\frac{1}{n}\sum_{i=1}^n (Y_i - \bar{Y}_n)^2\right)^2}.$$

Whereas, the normality assumption often simplifies calculations, it may lead to gross underestimation of risk. This has been widely recognised:

> ... as you well know, the biggest problems we now have with the whole evolution of risk is the fat-tail-problem, which is really creating very large conceptual difficulties. Because we all know, the assumption of normality enables us to drop off the uge amount of complexity in our equations... Because once you start putting in non-normality assumptions, which is unfortunately what characterises the real world, then these issues become extremely difficult.
>
> Alan Greenspan (1997).

1.1. Illustrations of stylised facts. Here, we show the stylised facts for a number of time-series.

1.1.1. *Apple data.* First we analyse the log-returns of daily closing prices of AAPL. We take all data from 2010 up till August 2012. In Figure 2 the raw data and log-returns are plotted. Next, in Figure 3, ACF and PACF plots are shown, together with a histogram and normal QQ-plot of the data. These figures can be made using the following code

```
# Make an object containing the data, called AAPL
getSymbols("AAPL",scr='yahoo')

# Extract closing prices of AAPL, from 2012 onwards
cl_AAPL <- Cl(AAPL['2012::'])

# Compute log-returns
lr_AAPL <- diff(log(cl_AAPL))

plot(cl_AAPL,col='blue')
```

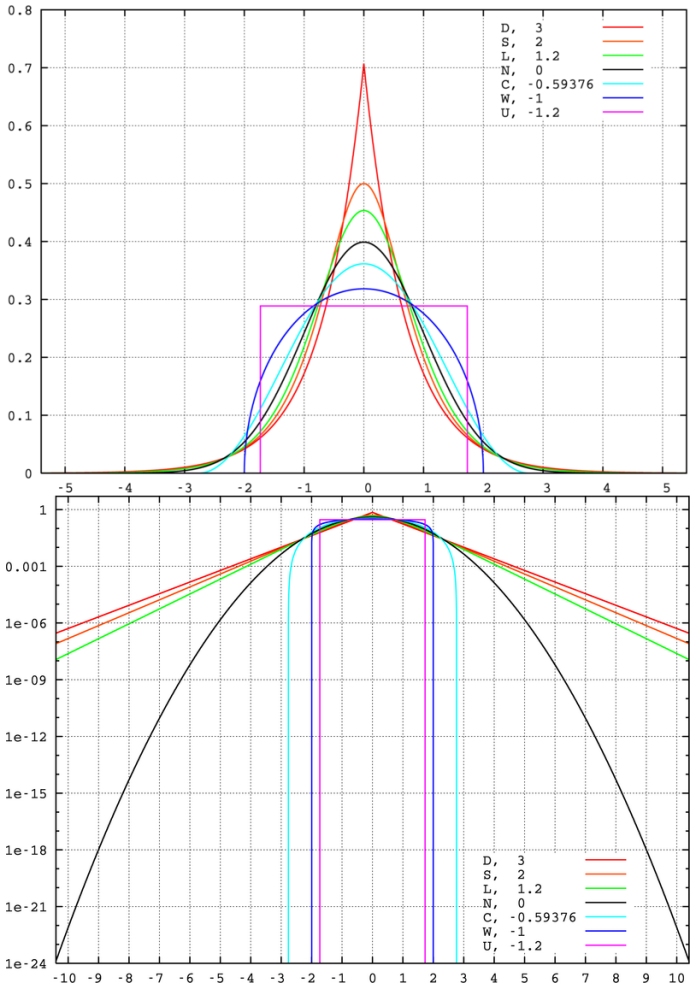

FIGURE 1. Excess kurtosis for a number of well known distributions. The lower figure uses a logarithmic scale (figures taken from Wikipedia).

```
plot(lr_AAPL,col='blue')

tsdisplay(lr_AAPL)
par(mfrow=c(1,2))
hist(lr_AAPL,breaks='FD',prob=TRUE,xlab="",ylab="",main="")
y <- as.numeric(lr_AAPL)
qqnorm(y); qqline(y,col='red')
```

To check for stylised features we carried out a Ljung-Box test on both the logreturns and squared logreturns.

```
> Box.test(y,lag=10,type='Ljung-Box')
```

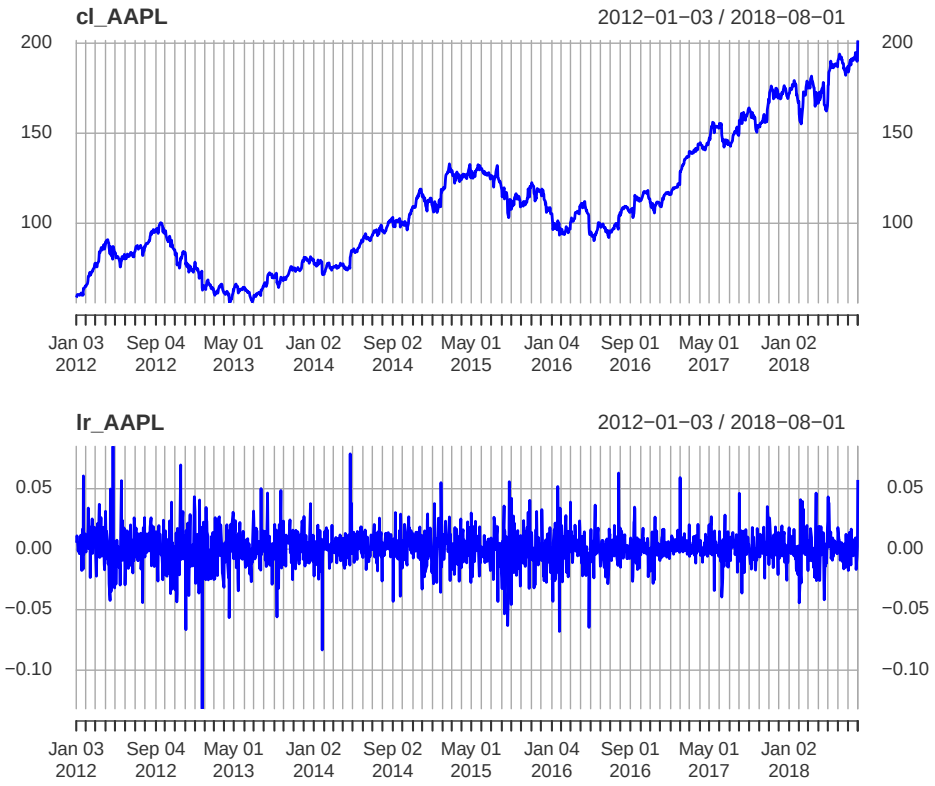

FIGURE 2. AAPL data. Upper panel: raw data. Lower panel: logreturns.

```
   Box-Ljung test

data:  y
X-squared = 18.709, df = 10, p-value = 0.04411

> Box.test(y^2,lag=10,type='Ljung-Box')

   Box-Ljung test

data:  y^2
X-squared = 28.458, df = 10, p-value = 0.001524
```

Depending on the significance level chose, the white noise assumption on the logreturns is rejected. For the squared log-returns, it is rejected at level 0.01.
By using the code

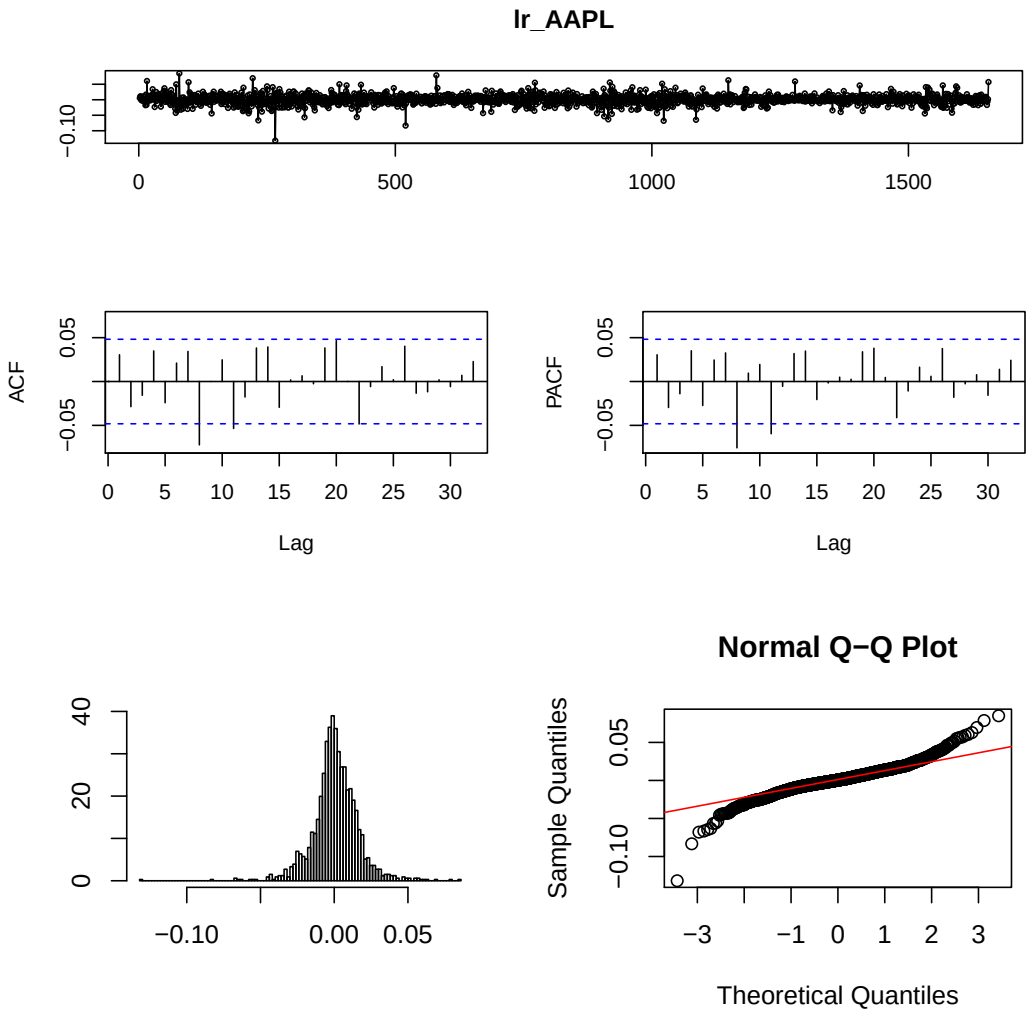

FIGURE 3. Analysis of the log-returns for AAPL.

```
kurtosis(y,na.rm=TRUE) # excess kurtosis
```

we can compute that the excess kurtosis of the data is 6.0, confirming heavy tails.

1.1.2. *Vodafone data.* We repeat the analysis for Vodafone Group Public Limited Company - NasdaqGS with symbol "VOD" over the years 2007 up till 2010. Figures 4 and 5 show the corresponding figures. For these data the volatility clustering effect is clearly visible. Again we notice heavy tails, the excess kurtosis being 5.34.

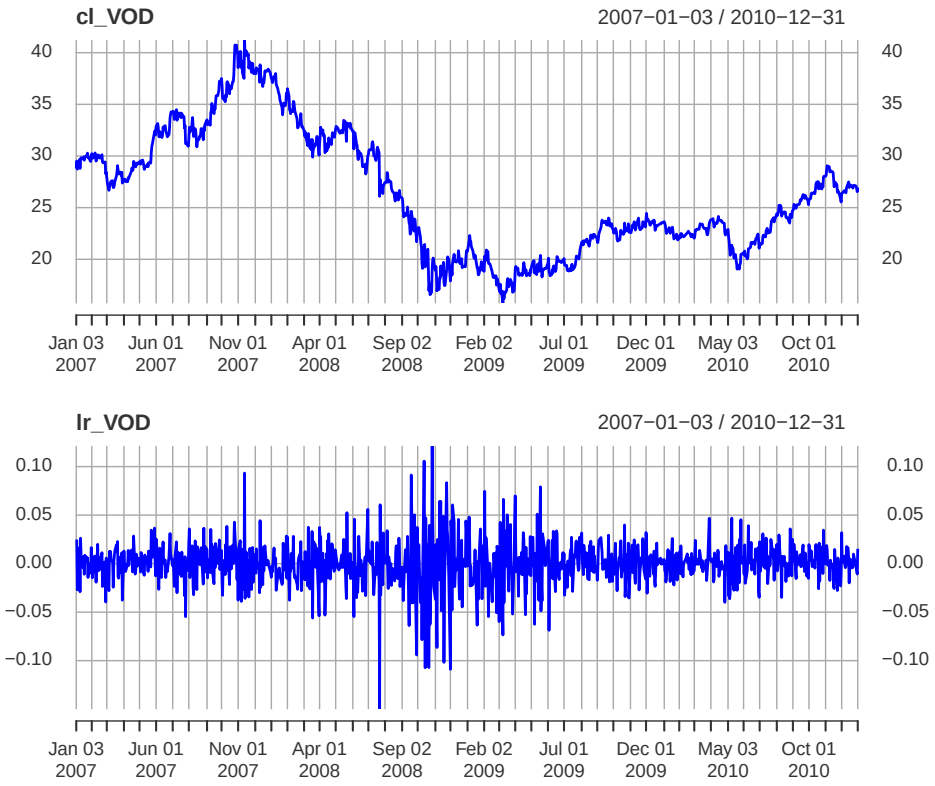

FIGURE 4. VOD data. Upper panel: raw data. Lower panel: logreturns.

Exercise 6.1 Try this analysis yourself on Koninklijke Philips Electronics NV (PHG) - NYSE.

2. ARCH(1)-model

We first study the most basic GARCH-model in detail.

DEFINITION 6.3. Let $\{Z_t\} \sim \text{IID}(0,1)$. Suppose $\omega, \alpha \geq 0$. Then $\{X_t\} \sim \text{ARCH}(1)$ if $X_t = \sigma_t Z_t$ where

$$\sigma_t^2 = \omega + \alpha\ X_{t-1}^2.$$

Please note that the sequence $\{Z_t\}$ is assumed to be IID, which is a stronger assumption than WN.

LEMMA 6.4 (Conditional properties). If $\{X_t\} \sim \text{ARCH}(1)$, then

$$\text{E}[X_t \mid \mathcal{F}_{t-1}] = 0 \qquad \text{Var}(X_t \mid \mathcal{F}_{t-1}) = \sigma_t^2$$

PROOF. First note that

$$\text{E}[X_t \mid \mathcal{F}_{t-1}] = \text{E}[\sigma_t Z_t \mid \mathcal{F}_{t-1}] = \sigma_t \text{E}[Z_t \mid \mathcal{F}_{t-1}] = \sigma_t \text{E}[Z_t] = 0.$$

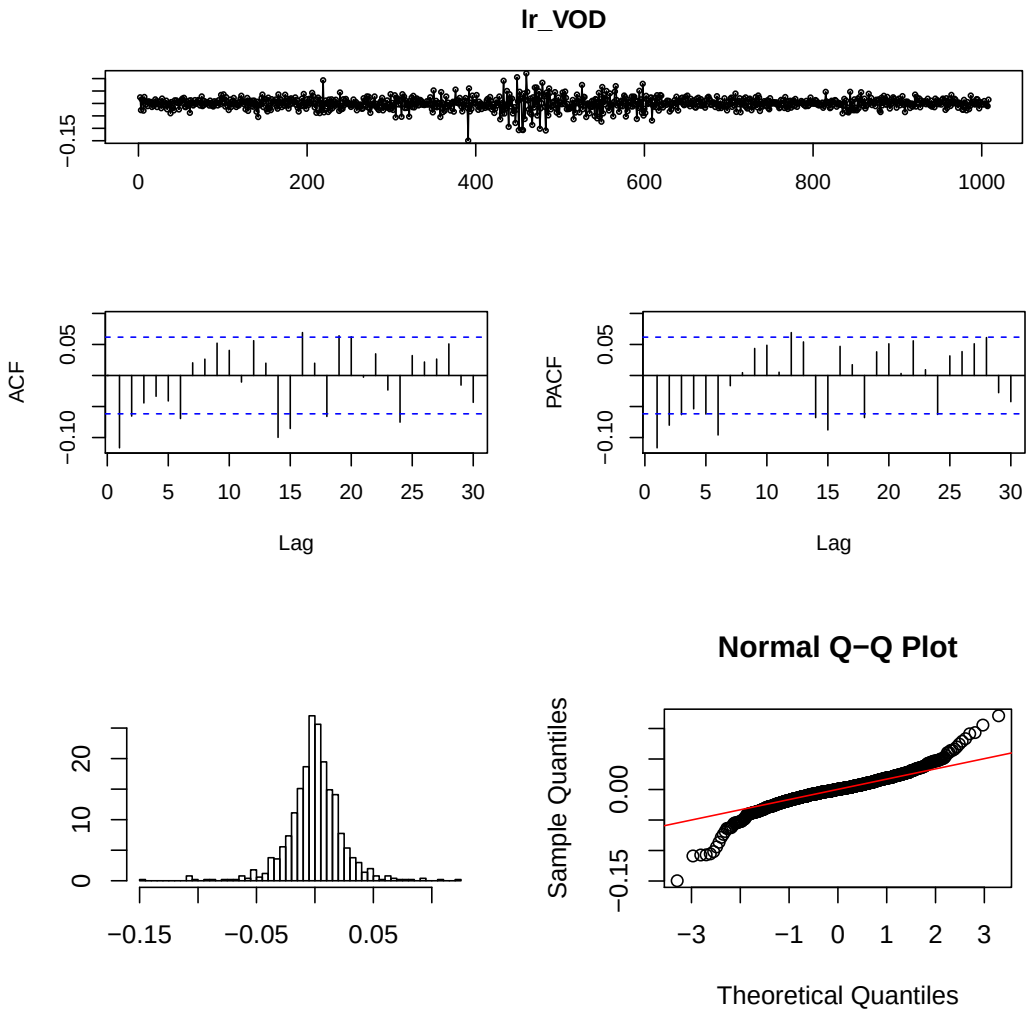

FIGURE 5. Analysis of the log-returns for VOD.

Along the same lines of proof it follows that

$$\mathrm{Var}(X_t \mid \mathcal{F}_{t-1}) = \mathrm{E}\left[X_t^2 \mid \mathcal{F}_{t-1}\right] = \sigma_t^2 \mathrm{E}\left[Z_t^2\right] = \sigma_t^2.$$

□

Hence

$$\mathrm{Var}(X_t \mid \mathcal{F}_{t-1}) = \mathrm{Var}(X_t \mid X_{t-1}) = \sigma_t^2 = \omega + \alpha\ X_{t-1}^2.$$

So indeed, the conditional variance is modelled. In particular, if we assume $\{Z_t\} \sim N(0,1)$, then

$$X_t \mid X_{t-1} \sim N\left(0, \omega + \alpha\ X_{t-1}^2\right).$$

The conditional distribution has nonconstant variance, which explain the naming "Conditionally Heteroscedastic". The previous display reveals that an ARCH(1) model exhibits volatility clustering: large values of X_{t-1}^2 are likely to be following by large values of X_t.
The following lemma shows that an ARCH(1) model is a particular white-noise model.

LEMMA 6.5 (Unconditional properties). If $\{X_t\} \sim$ ARCH(1), then

(1) $\mathrm{E}[X_t] = 0$;
(2) For $h > 0$, $\mathrm{Cov}(X_t, X_{t+h}) = 0$;
(3) $\mathrm{Var}(X_t) = \omega + \alpha \mathrm{Var}(X_{t-1})$.

So, with proper initialization, $\{X_t\}$ is a white-noise series.

PROOF. (1)

$$\mathrm{E}[X_t] = \mathrm{E}\,\mathrm{E}[X_t \mid \mathcal{F}_{t-1}] = 0$$

(2) For $h > 0$,

$$\begin{aligned}\mathrm{E}[X_t X_{t+h}] &= \mathrm{E}\,\mathrm{E}[X_t X_{t+h} \mid \mathcal{F}_{t+h-1}] \\ &= \mathrm{E}\left[X_t \underbrace{\mathrm{E}[X_{t+h} \mid \mathcal{F}_{t+h-1}]}_{=0}\right] = 0\end{aligned}$$

(3)

$$\mathrm{E}\left[X_t^2\right] = \mathrm{E}\,\mathrm{E}\left[X_t^2 \mid \mathcal{F}_{t-1}\right] = \mathrm{E}\left[\sigma_t^2\right] = \omega + \alpha \mathrm{E}\left[X_{t-1}^2\right]$$

□

The following lemma implies that if $\{X_t\} \sim$ ARCH(1), then the partial ACF of $\{X_t^2\}$ is zero after lag 1.

LEMMA 6.6. If $\{X_t\} \sim$ ARCH(1), then $X_t^2 \sim$ AR(1).

PROOF. Note that

$$X_t^2 - (\omega + \alpha X_{t-1}^2) = V_t,$$

where

$$V_t = \sigma_t^2 (Z_t^2 - 1).$$

Hence

$$X_t^2 = \omega + \alpha X_{t-1}^2 + V_t.$$

It remains to prove that $\{V_t\}$ is a white noise sequence. To this end, note that $\{V_t\}$ has mean zero since

$$\mathrm{E}[V_t] = \mathrm{E}\,\mathrm{E}\left[\sigma_t^2 (Z_t^2 - 1) \mid \mathcal{F}_{t-1}\right] = \mathrm{E}\left[\sigma_t^2\right] \mathrm{E}\left[Z_t^2 - 1\right] = 0\,.$$

The remainder of the proof is not part of the exam. For $h > 0$,

$$\begin{aligned}
\mathrm{E}\,[V_t V_{t+h}] &= \mathrm{E}\,\mathrm{E}\,[V_t V_{t+h} \mid \mathcal{F}_{t+h-1}] \\
&= \mathrm{E}\,[V_t \mathrm{E}\,[V_{t+h} \mid \mathcal{F}_{t+h-1}]] \\
&= \mathrm{E}\,\left[V_t \mathrm{E}\,\left[\sigma^2_{t+h}(Z^2_{t+h} - 1) \mid \mathcal{F}_{t+h-1}\right]\right] \\
&= \mathrm{E}\,\left[V_t \sigma^2_{t+h} \mathrm{E}\,\left[(Z^2_{t+h} - 1) \mid \mathcal{F}_{t+h-1}\right]\right] \\
&= \mathrm{E}\,\left[V_t \sigma^2_{t+h}\right] \mathrm{E}\,\left[Z^2_{t+h} - 1\right] = 0.
\end{aligned}$$

This shows that $\mathrm{Cov}(V_t, V_{t+h}) = 0$ for $h > 0$. □

As a consequence, there exists a causal stationary process

$$X_t^2 = \omega + \alpha X_{t-1}^2 + V_t$$

if $\alpha \in [0, 1)$.

Exercise 6.2 Show that for a stationary ARCH(1)-process

$$\mathrm{Var}(X_t) = \frac{\omega}{1 - \alpha}.$$

We summarise the main properties of a ARCH(1) process

- A causal stationary process exists if $\omega \geq 0$ and $\alpha \in [0, 1)$.
- The conditional mean is zero, whereas
$$\mathrm{Var}(X_t \mid \mathcal{F}_{t-1}) = \sigma_t^2 = \omega + \alpha X_{t-1}^2.$$
- The squared process $\{X_t^2\}$ is AR(1) with nonzero mean and non-Gaussian noise sequence
$$X_t^2 = \omega + \alpha X_{t-1}^2 + V_t.$$

2.1. Kurtosis of stationary ARCH(1). The kurtosis of a ARCH(1)-process depends on the kurtosis of the IID-sequence $\{Z_t\}$.

LEMMA 6.7. For a stationary ARCH(1) process $\{X_t\}$ with noise sequence $\{Z_t\}$ we have

$$\kappa_X = \frac{1 - \alpha^2}{1 - \tau\alpha^2} \kappa_Z,$$

provided that $\alpha^2 < 1/\tau$. Here $\tau = \mathrm{E}\,\left[Z_t^4\right]$.

If $\{Z_t\} \sim N(0, 1)$, then $\tau = 3$ and we need to impose the restriction $\alpha < 1/\sqrt{3}$ for X_t to have finite kurtosis. As a consequence of this lemma

$$\kappa_X > \kappa_Z$$

which reveals that the distribution of X_t has heavier tails than that of Z_t. Hence, the stylised feature "heavy tails" is indeed accounted for by the model. Furthermore, the higher the kurtosis of Z_t, the higher that of X_t. By default it is assumed that $Z_t \sim N(0, 1)$, but in practise for some datasets an IID sequence $\{Z_t\}$ with heavier tails may be desirable. The following choices can account for this:

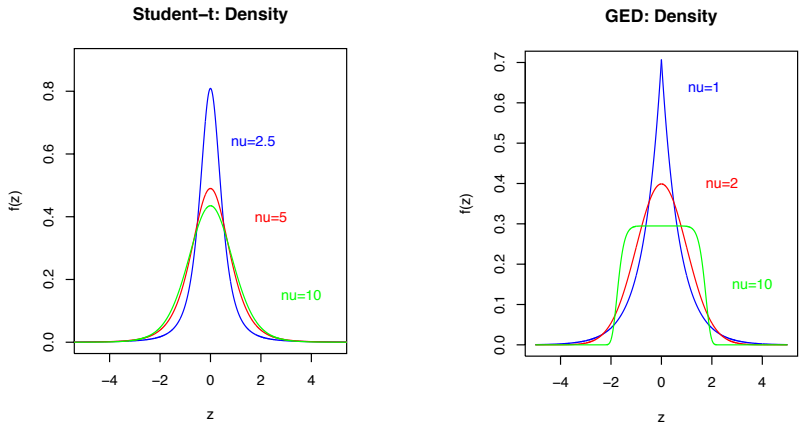

FIGURE 6. Left: density plot of the standardised Student-t distribution for various values of ν. Right: density plot of the Generalised Error distribution for various values of ν.

- the Standardised Student-t (standardised such that $\mathrm{Var}(Z_t) = 1$):

$$Z_t \sim \sqrt{\frac{\nu-2}{\nu}} t_\nu, \qquad \nu > 2$$

where t_ν denotes a Student-t with ν degrees of freedom. Its density is given by

$$f(x) = \frac{\Gamma\left((\nu+1)/2\right)}{\Gamma(\nu/2)\sqrt{(\nu-2)\pi}} \left(1 + \frac{x^2}{\nu-2}\right)^{-(\nu+1)/2}.$$

- The Generalized Error Distribution (GED) with density

$$f(x) = \frac{\nu \exp\left(-\frac{1}{2}|x/\lambda|^\nu\right)}{\lambda 2^{1+1/\nu}\Gamma(1/\nu)}, \qquad 0 < \nu \leq \infty, \qquad \lambda = \sqrt{\frac{\Gamma(1/\nu)}{2^{2\nu}\Gamma(3/\nu)}}.$$

The value $\nu = 2$ corresponds to $N(0,1)$ and when $0 < \nu < 2$ the distribution has heavier tails than the normal.

In R, the parameter ν is referred to as the `shape` parameter. A plot of the densities of these distributions for various values of ν is given in Figure 6

2.2. Simulating and fitting an ARCH model in R. Simulating a realisation of a ARCH(1) in R is easy. By default the noise sequence $\{Z_t\}$ is taken to be IID Normal with mean zero and variance 1.

```
spec <- garchSpec(model = list(omega=5, alpha = 0.4, beta = 0))
x <- garchSim(spec, n = 500)
```

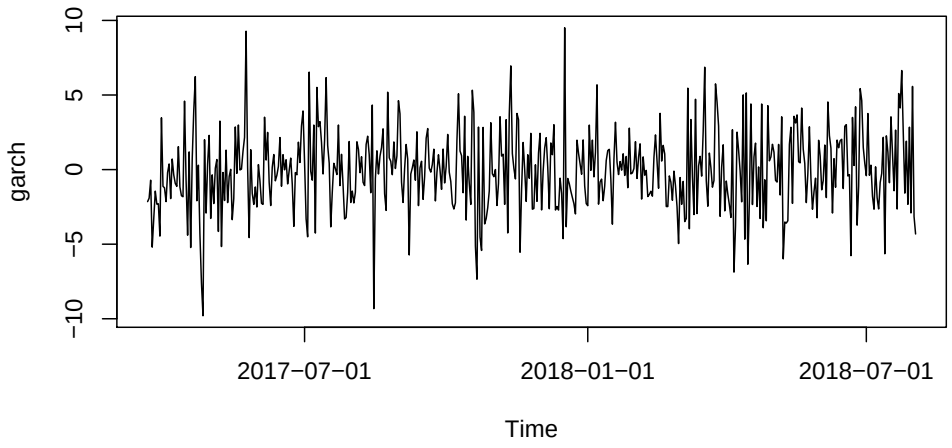

FIGURE 7. Realisation of length 500 of a ARCH(1)-process, with $\omega = 5$ and $\alpha = 0.4$. The sequence $\{Z_t\}$ has the standard normal as marginal distribution.

Figure 7 shows a time-series plot of the generated data. In Figure 8 the ACF and PACF of both x and x^2 are shown. These are as expected (check this yourself).
With the command `garchFit` the following model is fitted:

$$X_t = \mu + \sigma_t Z_t \qquad \sigma_t^2 = \omega + \alpha X_{t-1}^2$$

Hence note that by default an intercept μ is included in the model. As an illustration, we fit an ARCH(1) model to the data that we just generated and visualised in Figure 7.

```
gfit <- garchFit(~garch(1,0),data=x)
```

The syntax becomes a bit more clear when we have discussed what a GARCH-process is (it turns out that ARCH(1) is the same as GARCH$(1, 0)$).
From the object `gfit` useful information on the fitted model can be extracted:

- `summary(gfit)`: as the name suggests, this gives a summary of the fitted model, including parameter estimates

  $$\hat{\mu},\ \hat{\omega},\ \hat{\alpha}$$

  and the estimates of the standard deviations of the corresponding estimators.
- `plot(gfit)` allows to make various plots, including residual plots.

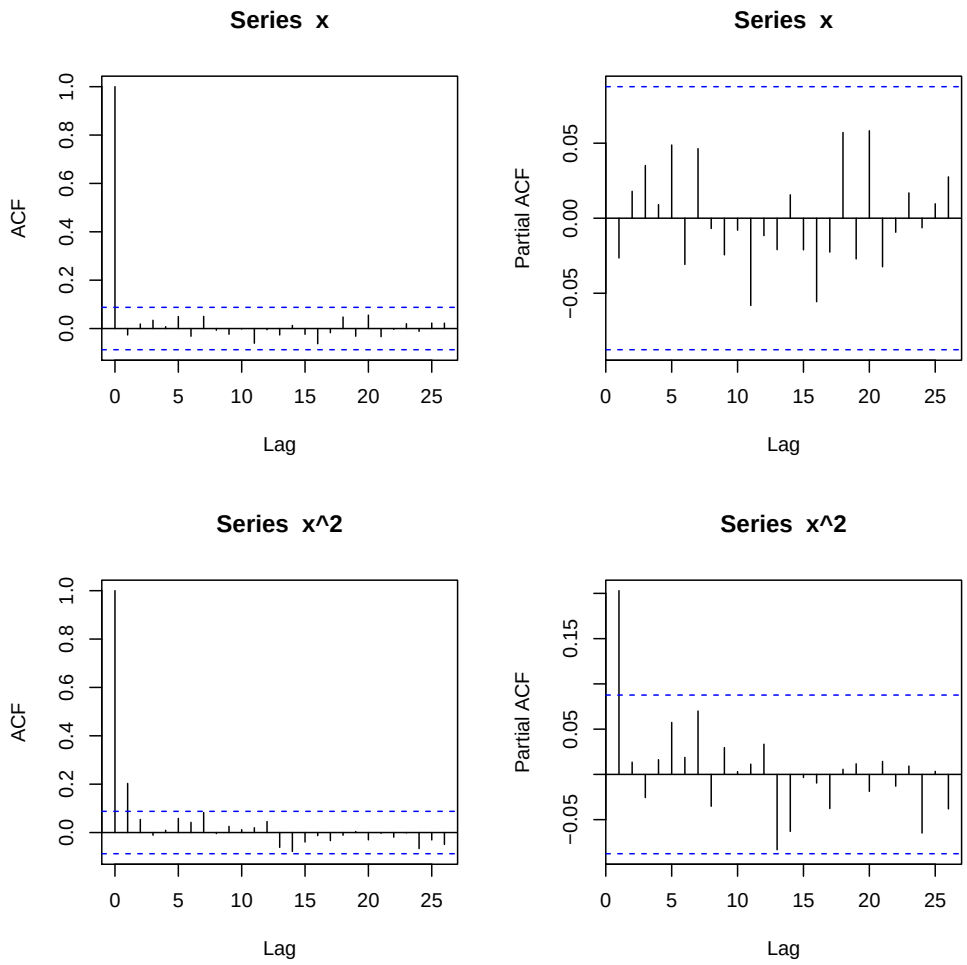

FIGURE 8. ACF, PACF plots for both x and x^2, where x is the realisation in Figure 7.

- Specific quantities that can be extracted include:
  - information criteria such as AIC:

    `gfit@fit$ics`

  - residuals $\widehat{Z}_t$:

    `gfit@residuals` or `residuals(gfit)`

  - standardised residuals

    `gfit@residuals/gfit@sigma.t`

– volatility: $\widehat{\sigma}_t$:

`gfit@sigma.t` or `volatility(gfit)`

– conditional variance: $\widehat{\sigma}_t^2$:

`gfit@h.t`

For a GARCH model (so in particular a ARCH(1) model), the residuals and standardised residuals are defined by

$$X_t - \hat{\mu}$$

and

$$\hat{Z}_t := \frac{X_t - \hat{\mu}}{\hat{\sigma}_t}$$

respectively. Here $\hat{\sigma}_t$ is the (predicted) volatility, which is given by

$$\hat{\sigma}_t = \sqrt{\hat{\omega} + \hat{\alpha} X_{t-1}^2}.$$

Standardised residuals are useful for model checking: the intuition being that if the model is adequate, then the standardised residuals should be like the IID$(0,1)$ sequence $\{Z_t\}$ (which is by default the standard normal distribution). In particular, model checking should at least include:

- a check that $\{\hat{Z}_t\}$ behaves like white-noise using the ACF, possibly augmented with a Ljung-Box test;
- a check on the validity of the distributional assumption on $\{Z_t\}$ by making a QQ-plot to $\{\hat{Z}_t\}$;
- a check that $\{\hat{Z}_t^2\}$ is of AR(1) type, by making a PACF plot for $\{\hat{Z}_t^2\}$.

We apply the `summary` command on the fitted model.

```
> summary(gfit)

Title:
 GARCH Modelling

Call:
 garchFit(formula = ~garch(1, 0), data = x)

Mean and Variance Equation:
 data ~ garch(1, 0)

Conditional Distribution:
 norm

Coefficient(s):
      mu     omega    alpha1
-0.051593  4.565640  0.407811

Std. Errors:
```

```
 based on Hessian

Error Analysis:
       Estimate Std. Error t value Pr(>|t|)
mu     -0.05159    0.10204  -0.506    0.613
omega   4.56564    0.45032  10.139  < 2e-16 ***
alpha1  0.40781    0.08932   4.566 4.98e-06 ***
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 0.1 1

Standardised Residuals Tests:
                               Statistic p-Value
 Jarque-Bera Test R    Chi^2  2.980174 0.2253531
 Shapiro-Wilk Test R   W      0.995511 0.1609105
 Ljung-Box Test   R    Q(10)  4.907731 0.8972558
 Ljung-Box Test   R    Q(15)  7.172082 0.9526974
 Ljung-Box Test   R    Q(20)  13.85067 0.8379856
 Ljung-Box Test   R^2  Q(10)  9.390504 0.4954871
 Ljung-Box Test   R^2  Q(15)  16.02895 0.3801301
 Ljung-Box Test   R^2  Q(20)  18.23344 0.5720335
 LM Arch Test     R    TR^2   10.91223 0.5364554

Information Criterion Statistics:
     AIC      BIC      SIC     HQIC
4.750271 4.775559 4.750200 4.760194
```

(Some redundant output was not included here.) Besides parameter estimates, various tests are performed on the standardised residuals and their squared values. We have discussed the Ljung-Box test: if the model is adequate, then all p-values of these tests are expected to be uniformly distributed on $[0, 1]$. The "Jarque-Bera" and "Shapiro-Wilk" tests are tests for normality of the standardised residuals: we have not discussed such tests. Similarly, we have not discussed the "LM Arch Test".
All kinds of plots can easily be made:

```
> plot(gfit)

Make a plot selection (or 0 to exit):

 1:  Time Series
 2:  Conditional SD
 3:  Series with 2 Conditional SD Superimposed
 4:  ACF of Observations
 5:  ACF of Squared Observations
 6:  Cross Correlation
 7:  Residuals
 8:  Conditional SDs
 9:  Standardized Residuals
```

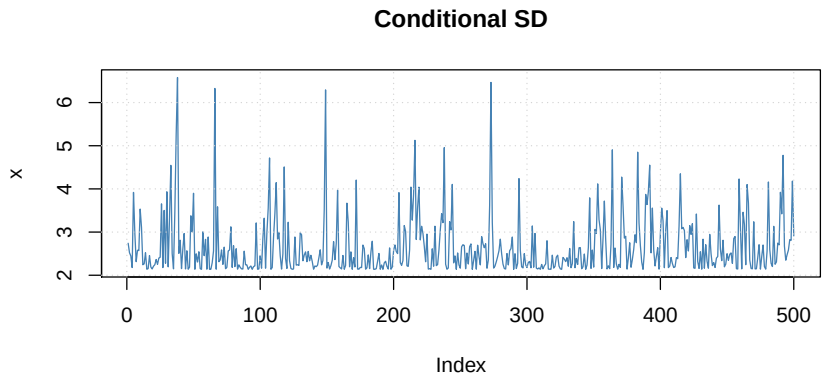

FIGURE 9. Plot of $\hat{\sigma}_t$, based on the ARCH(1) fit for the data x in Figure 7.

```
10:   ACF of Standardized Residuals
11:   ACF of Squared Standardized Residuals
12:   Cross Correlation between r^2 and r
13:   QQ-Plot of Standardized Residuals
```

Now the user can choose which plot needs to be generated. Alternatively, if you already know which plot you wish to generate, you can use the command

```
plot(gfit, which=2) # plot of conditional SD
```

The resulting plot is shown in Figure 9.

2.3. Best prediction in a ARCH(1) model. Suppose $\{X_t\} \sim$ ARCH(1) with $\{Z_t\} \sim N(0,1)$, then

$$X_{t+1} \mid \mathcal{F}_t \sim N\left(0, \sigma_{t+1}^2 = \omega + \alpha\ X_t^2\right).$$

This is the *one-step ahead forecast*. This says we simply forecast zero, but the prediction bands depend on the previous value X_t.

LEMMA 6.8 (Forecasting h-steps ahead). For $h \geq 1$

$$\mathrm{E}\left[X_{t+h} \mid \mathcal{F}_t\right] = 0 \qquad \mathrm{Var}(X_{t+h} \mid \mathcal{F}_t) = \mathrm{E}\left[X_{t+h}^2 \mid \mathcal{F}_t\right]$$

and

$$\mathrm{E}\left[X_{t+h}^2 \mid \mathcal{F}_t\right] = \omega\left(1 + \alpha + \ldots + \alpha^{h-2}\right) + \alpha^{h-1}\sigma_{t+1}^2$$

PROOF. We first give the proof for the case $h = 2$.

$$\begin{aligned} X_{t+2}^2 &= \sigma_{t+2}^2 Z_{t+2}^2 \\ &= \left(\omega + \alpha X_{t+1}^2\right) Z_{t+2}^2 \\ &= \left(\omega + \alpha\sigma_{t+1}^2 Z_{t+1}^2\right) Z_{t+2}^2 \\ &= \omega Z_{t+2}^2 + \alpha\sigma_{t+1}^2 Z_{t+1}^2 Z_{t+2}^2 \end{aligned}$$

Hence,

$$\mathrm{E}\left[X_{t+2}^2 \mid \mathcal{F}_t\right] = \omega + \alpha\sigma_{t+1}^2.$$

Next, consider the case $h = 3$

$$\begin{aligned} X_{t+3}^2 &= \sigma_{t+3}^2 Z_{t+3}^2 \\ &= \left(\omega + \alpha X_{t+2}^2\right) Z_{t+3}^2 \\ &= \left(\omega + \alpha\left[\omega Z_{t+2}^2 + \alpha\sigma_{t+1}^2 Z_{t+1}^2 Z_{t+2}^2\right]\right) Z_{t+3}^2 \\ &= \omega Z_{t+3}^2 + \alpha\omega Z_{t+2}^2 Z_{t+3}^2 + \alpha^2\sigma_{t+1}^2 Z_{t+1}^2 Z_{t+2}^2 Z_{t+3}^2 \end{aligned}$$

Hence,

$$\mathrm{E}\left[X_{t+3}^2 \mid \mathcal{F}_t\right] = \omega + \alpha\omega + \alpha^2\sigma_{t+1}^2.$$

In general

$$\mathrm{E}\left[X_{t+h}^2 \mid \mathcal{F}_t\right] = \omega + \alpha\mathrm{E}\left[X_{t+h-1}^2 \mid \mathcal{F}_t\right].$$

So if we define

$$v(h) := \mathrm{Var}(X_{t+h} \mid \mathcal{F}_t),$$

then

$$\begin{aligned} v(h) &= \omega + \alpha v(h-1), \qquad h \geq 2 \\ v(1) &= \sigma_{t+1}^2 \end{aligned}$$

from which the result follows. □

Note that

$$\mathrm{E}\left[X_{t+h}^2 \mid \mathcal{F}_t\right] \to \omega/(1-\alpha) \qquad (h \to \infty),$$

which is the variance of the stationary ARCH(1)-process.

2.4. Discussion: weaknesses of ARCH models. While ARCH models do account for a number of stylised features, they also have some weaknesses:

- The ARCH model assumes that positive and negative shocks have the same effect on volatility.
- The ARCH model provides a way to describe conditional variance. It does not give indications what causes such behavior.
- The ARCH model is restrictive, i.e., constraints, such as $3\alpha^2 < 1$, limit the ability of ARCH models with conditional Gaussian innovations to capture excess kurtosis.

These weaknesses have led to the introduction of more advanced models.

3. Extending the model: GARCH models

The ARCH(1) model can be extended to the ARCH(m) in a straightforward way.

DEFINITION 6.9. Let $\{Z_t\} \sim \text{IID}(0,1)$. Assume $\omega, \alpha_1, \ldots, \alpha_m \geq 0$. Then $\{X_t\} \sim \text{ARCH}(m)$ if $X_t = \sigma_t Z_t$ where

$$\sigma_t^2 = \omega + \alpha_1 X_{t-1}^2 + \ldots + \alpha_m X_{t-m}^2$$

Without proof we state that if $\{X_t\} \sim \text{ARCH}(m)$ then

- $\{X_t\} \sim \text{WN}$;
- $\{X_t^2\} \sim \text{AR}(m)$. Hence the PACF of $\{X_t^2\}$ can be used for determining the order m.

It is observed empirically that often a high order high order ARCH-model is needed. For that reason, GARCH models were introduced to provide a more parsimonious model.

DEFINITION 6.10. Let $\{Z_t\} \sim \text{IID}(0,1)$. Assume $\omega, \alpha, \beta \geq 0$. Then $\{X_t\} \sim \text{GARCH}(1,1)$ if $X_t = \sigma_t Z_t$ with

$$\sigma_t^2 = \omega + \alpha X_{t-1}^2 + \beta \sigma_{t-1}^2.$$

It is now also clear how to define a GARCH(p,q) model.
In the following lemmas we state some properties of the GARCH(1,1) model without proof.

LEMMA 6.11. Conditional properties

$$\text{E}[X_t \mid \mathcal{F}_{t-1}] = 0 \qquad \text{Var}(X_t \mid \mathcal{F}_{t-1}) = \sigma_t^2$$

LEMMA 6.12. Unconditional properties

(1) $\text{E}[X_t] = 0$;
(2) For $h > 0$, $\text{Cov}(X_t, X_{t+h}) = 0$;
(3) Suppose $\alpha + \beta < 1$. With proper initialization, $\{X_t\}$ is a *white-noise series* with

$$\text{Var}(X_t) = \frac{\omega}{1 - \alpha - \beta}.$$

LEMMA 6.13. If $X_t \sim \text{GARCH}(1,1)$, then $\{X_t^2\}$ admits a non-Gaussian ARMA(1,1) representation:

$$X_t^2 = \omega + (\alpha + \beta) X_{t-1}^2 + V_t - \beta V_{t-1}$$

where $\{V_t\}$ is non-Gaussian white noise.

PROOF. From

$$X_t = \sigma_t Z_t$$

we get

$$X_t^2 - \sigma_t^2 = \sigma_t^2 Z_t^2 - \sigma_t^2 = V_t, \tag{13}$$

where $V_t = \sigma_t^2(Z_t^2 - 1)$. Hence

$$\beta(X_{t-1}^2 - \sigma_{t-1}^2) = \beta V_{t-1} \tag{14}$$

Now subtracting equation (14) from equation (13) gives

$$X_t^2 - \beta X_{t-1}^2 - (\sigma_t^2 - \beta\sigma_{t-1}^2) = V_t - \beta V_{t-1}.$$

The representation now follows since by the equation for σ_t^2 gives

$$\sigma_t^2 - \beta\sigma_{t-1}^2 = \omega + \alpha X_{t-1}^2.$$

The proof that $\{V_t\} \sim \text{WN}$ follows similarly as in Lemma 6.6. □

Further properties and extensions of GARCH$(1,1)$ Many variations of GARCH exist: APARCH, GARCH in mean, I-GARCH, ... A "readable" paper on the use of GARCH models in econometrics is Engle [2001].

4. Combining ARMA and GARCH models

ARMA models can be combined with GARCH models:

- the ARMA-part models $\mathrm{E}[X_t \mid \mathcal{F}_{t-1}]$;
- the GARCH-part models $\mathrm{Var}(X_t \mid \mathcal{F}_{t-1})$.

As an example, the model

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

with

$$X_t = \sigma_t Z_t \qquad \sigma_t^2 = \omega + \alpha X_{t-1}^2 + \beta\sigma_{t-1}^2$$

specifies a AR(1) + GARCH$(1,1)$ model for $\{Y_t\}$.

EXAMPLE 6.14. We analyse the closing prices of Koninklijke Philips Electronics NV (PHG) - NYSE over het period 2011-01-01 up till 2013-10-28. To obtain the data:

```
getSymbols('PHG')
price <- Cl(PHG['2011-01-01::2013-10-28'])
lr <- diff(log(price))[-1] # remove first element, as it is NA
```

A time-series plot of the log-returns is in Figure 10. Based on the ACF of the data, shown in top-left panel of Figure 11, we fit a MA(1)-model to the data.

```
acf(lr)
# fit MA(1) model
ma_fit <- sarima(lr,0,0,1,de=F,no.con=T)$fit
summary(ma_fit)
```

To verify if the fitted model is correct, we first check for normality of the residuals. This turns out to be problematic (the figure of the normal probability plot is not included here). Even more interesting, in Figure 11 we show the ACF of the residuals (which looks fine) together with the ACF and

PACF of the squared residuals. The latter two figures indicate an ARCH-effect: the squared residuals are dependent. Based on the PACF of the squared residuals we decide to fit a MA(1) + ARCH(3)-model.

```
g_fit <- garchFit(~ arma(0,1) + garch(3,0),data=lr,
                  include.mean=FALSE,trace=F)
```

The following diagnostic plots are most interesting

```
plot(g_fit,which=11) # ACF of Squared Standardized Residuals
plot(g_fit,which=13) # QQ-Plot of Standardized Residuals
```

The resulting plots, shown in Figure 12, reveal that the ARCH-effect is properly taken into account, but that the distributional problem of the IID-sequence $\{Z_t\}$ has not yet been fixed. Apparently, we need the noise sequence to have heavier tails. To accomplish this, we refit the model, but now with $\{Z_t\}$ assumed to have the scaled t-distribution.

```
g_fit2 <- garchFit(~ arma(0,1) + garch(3,0),data=lr,
                  include.mean=FALSE,trace=F,cond.dist="std")
```

Now we can again make diagnostic plots for the fitted model: see Figure 13. Indeed, the distributional problem has been resolved! The fitted model can be extracted from:

```
> summary(g_fit2)

Title:
 GARCH Modelling

Call:
 garchFit(formula = ~arma(0, 1) + garch(3, 0), data = lr, cond.dist
     = "std",
    include.mean = FALSE, trace = F)

Conditional Distribution:
 std

Coefficient(s):
       ma1       omega      alpha1      alpha2      alpha3
           shape
-0.06144578  0.00024584  0.04115813  0.24114731  0.27866369
    4.50115104
```

(Some of the output has been left out.) If you only wish to get the coefficients, the command

```
coef(g_fit2)
```

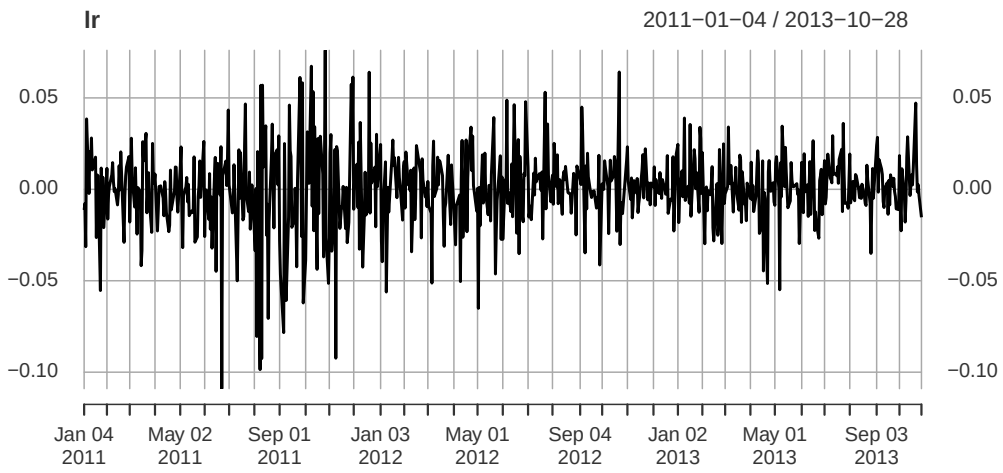

FIGURE 10. Log-returns for closing prices of Koninklijke Philips Electronics NV (PHG) - NYSE over het period 2011-01-01 up till 2013-10-28.

will also do. Hence, If $\{Y_t\}$ denotes the PHG log-returns at time t, then the model postulates that

$$\begin{aligned} Y_t &= X_t - 0.06X_{t-1} \\ X_t &= \sigma_t Z_t \\ \sigma_t^2 &= 2.5 \cdot 10^{-4} + 0.04X_{t-1}^2 + 0.24X_{t-2}^2 + 0.27X_{t-3}^2 \\ \{Z_t\} &\sim \text{IID} \quad \text{with} \quad Z_t \sim \text{scaled } t_\nu \text{ with } \nu = 4.5. \end{aligned}$$

Note that the parameter ν of the scaled t-distribution is referred to as the `shape` parameter in `R`.
With the fitted model we can forecast. Suppose we wish to forecast 50 days ahead

```
N <- 50 # number of time steps ahead to predict
predict(g_fit2, n.ahead=N, plot=TRUE)
```

The predictions are shown in Figure 14.

5. Exercises from old exams

Exercise 6.3
For each of the given time series models for $\{X_t\}$ state whether it concerns an AR(p), MA(q), ARMA(p, q), ARCH(p), GARCH(p, q) model or combined model. Provide the values for p and q. In this exercise you can neglect issues such as "parameter redundancy", stationarity, invertibility and causality.

(1) $X_t - 2X_{t-2} - Z_t = Z_{t-2}$, where $\{Z_t\} \sim \text{WN}(0, 1)$.

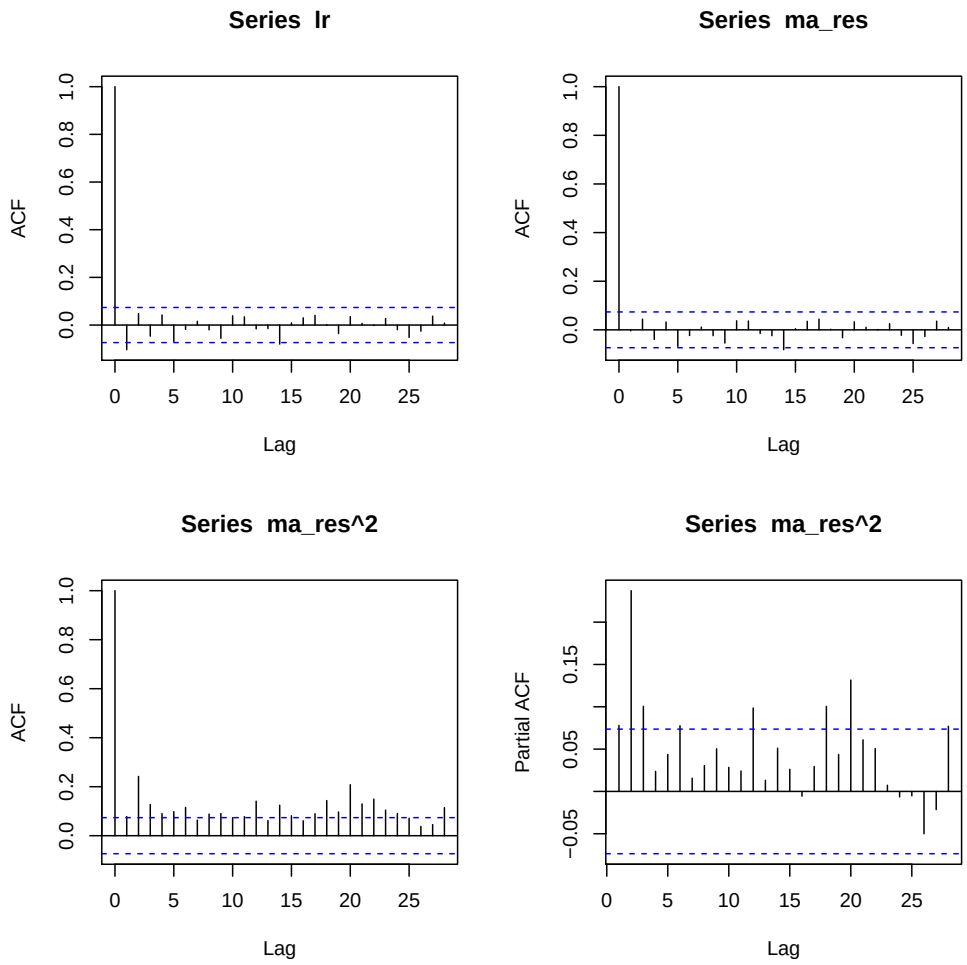

FIGURE 11. Top-left: ACF of the log-returns. Top-right: ACF of the residuals of the fitted MA(1)-model. Bottom-left: ACF of the squared residuals of the fitted MA(1)-model. Bottom-right: PACF of the squared residuals of the fitted MA(1)-model.

(2) $X_t = Z_t + 0.2Z_{t-1}$, where $Z_t = \sqrt{0.2 + 0.05Z_{t-1}^2}Y_t$ and $\{Y_t\} \sim \text{IID}(0,1)$.

Exercise 6.4 Suppose we have data saved in the vector **x**. Consider the following R-output from fitting a time-series model:

```
Call: garchFit(formula =  garch(2, 0), data = x)
Mean and Variance Equation: data   garch(2, 0) <environment: 0x10d8e1d48>
[data = x]
```

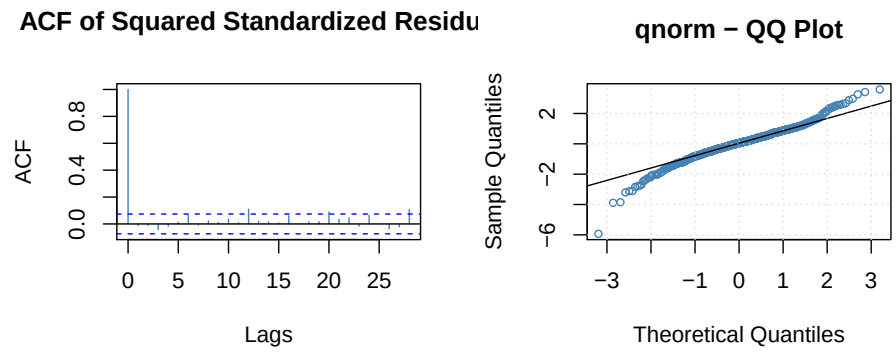

FIGURE 12. Diagnostic plot for the MA(1) + ARCH(3) model, where $\{Z_t\}$ has the normal distribution as marginal distribution. Left: ACF of the squared standardised residuals. Right: normal probability plot of the standardised residuals.

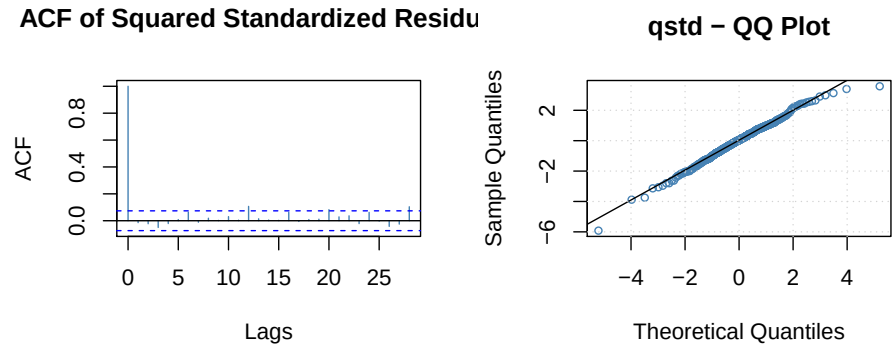

FIGURE 13. Diagnostic plot for the MA(1) + ARCH(3) model, where $\{Z_t\}$ has the scaled t-distribution as marginal distribution. Left: ACF of the squared standardised residuals. Right: normal probability plot of the standardised residuals.

```
Conditional Distribution: norm
Coefficient(s): mu omega alpha1 alpha2 -0.010232 4.877309 0.486111
0.075708
Std. Errors: based on Hessian
Error Analysis: Estimate Std. Error t value Pr(>|t|) mu -0.01023
0.07661 -0.134 0.8937 omega 4.87731 0.37785 12.908 < 2e-16 *** alpha1
```

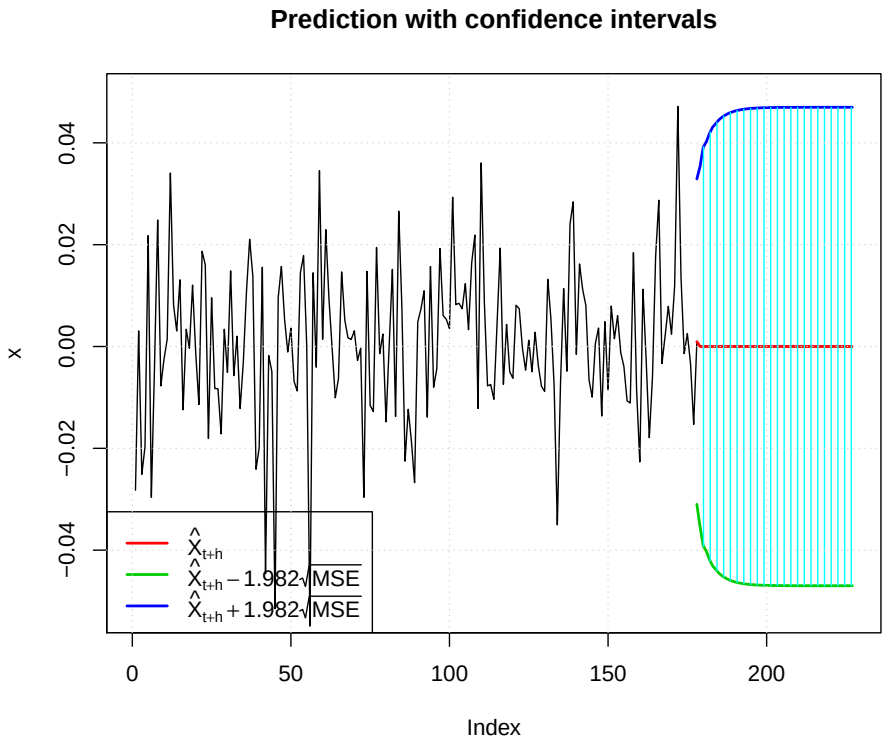

FIGURE 14. Forecasts obtained from the fitted MA(1) + ARCH(3) model, where $\{Z_t\}$ has the scaled t-distribution as marginal distribution.

```
0.48611 0.06078 7.998 1.33e-15 *** alpha2 0.07571 0.02959 2.559
0.0105 * ---
```

(1) Write down the equations for the fitted model. Values of fitted parameters can be rounded to two decimals.
(2) Consider figures 15 and 16. Is there evidence of departures from the assumed model? If so, what are discrepancies?

Exercise 6.5 Go to `finance.yahoo.com`, and search for 'coke' to find that the index on the NYSE has name 'KO'.

(1) Download the data into `R` (you can use the `quantmod` library) for this and consider the data between 2008-01-01 and 2017-08-01 (meaning the first of august in 2017). Make an Open-High-Low-Close (OHLC) chart of the data till one year back from today.
(2) Fit a MA(q)-model without intercept to the logreturns and save the residuals in a vector. Call this vector `z`. Can we consider the residuals to be white-noise?

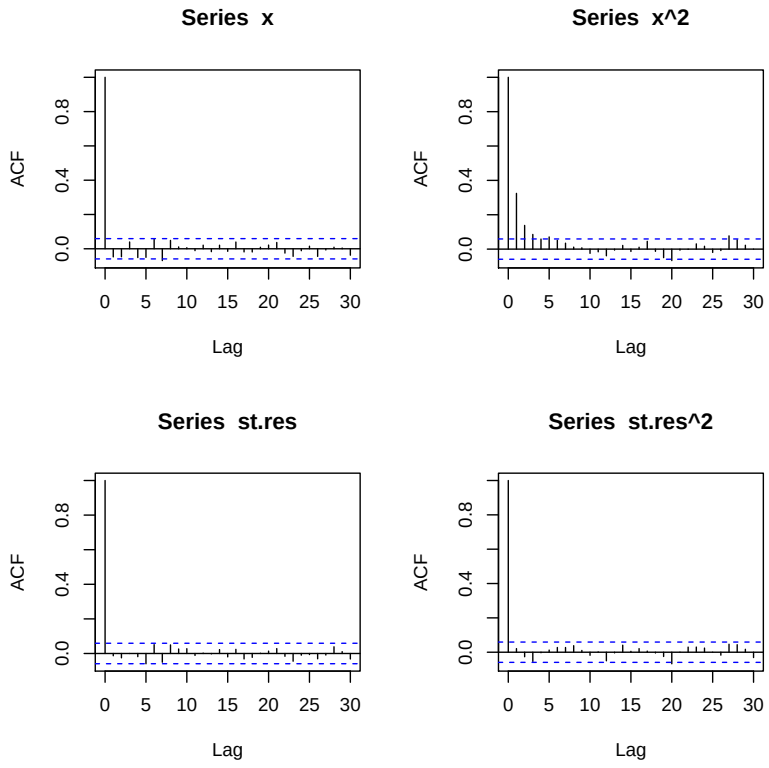

FIGURE 15.

(3) Investigate the time-series `z` for ARCH-effects using the autocorrelation function of the time-series $\mathtt{z}^2$. Do you find any ARCH-effects in `z`?
(4) Verify whether the marginal distribution of `z` can be assumed to have the Normal distribution using a probability-plot.
(5) Fit a GARCH$(1,1)$ model for `z` using the command

```
fit <- garchFit( garch(1,1),z,trace=F)
```

Give the equations of the fitted model (include *all* estimated parameters).
(6) Check whether the model assumption of the assumed GARCH model are fulfilled using the standardised residuals.
(7) We can also fit the MA(q) model jointly with GARCH$(1,1)$ white noise :

```
garchFit( arma(0,q)+ garch(1,1), data=y, trace=FALSE,cond.dist='std')
```

Here, we specify that the IID sequence appearing in the definition of the GARCH$(1,1)$-model has the student-t distribution. Fit

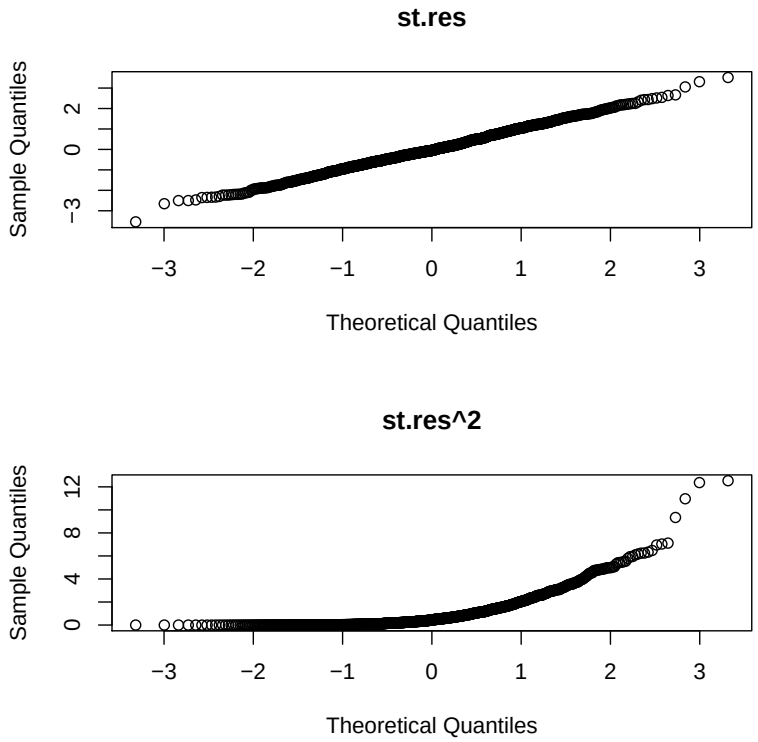

FIGURE 16.

this model and give the equations of the fitted model (include *all* estimated parameters).

Exercise 6.6 Suppose $\{Z_t\} \sim$ IID $N(0,1)$ and that $\{A_t\}$ and $\{Y_t\}$ are stationary time-series defined by

$$A_t = Z_t\sqrt{7 + 0.5A_{t-1}^2 + 0.05A_{t-2}^2}$$

and

$$Y_t = 2 + \frac{2}{3}Y_{t-1} + A_t$$

(1) Give the ACF for $\{A_t\}$; hence the mapping $h \mapsto \rho_A(h)$ (no need for simulation here) *Hint: what type of time-series is* $\{A_t\}$*?*
(2) Give the ACF of the process $\{Y_t\}$; Hence the function $h \mapsto \rho_Y(h)$.
(3) Does the distribution of $\{A_t\}$ have lighter or heavier tails compared to those of the Normal distribution?
(4) Suppose we simulate a realisation of $\{A_t\}$. What do you expect to see when you make a figure of the partial autocorrelation function of the time-series $\{A_t^2\}$?

Exercise 6.7 Consider the R-output below, obtained by the `summary`-command: Use this output to answer the following questions.

(1) Write down the model equations of the fitted model.
(2) Are the standardised residuals normally distributed?
(3) Can the standardised residuals be considered white-noise?
(4) Explain whether the fitted model is reasonable.

```
Mean and Variance Equation:
 data ~ arma(2, 0) + garch(1, 0)
<environment: 0x10d985178>
 [data = x]

Conditional Distribution:
 norm

Coefficient(s):
     ar1       ar2     omega    alpha1
 0.74111  -0.22913   0.13059   0.50846

Std. Errors:
 based on Hessian

Error Analysis:
        Estimate  Std. Error  t value Pr(>|t|)
ar1      0.74111     0.04371   16.955  < 2e-16 ***
ar2     -0.22913     0.03748   -6.114 9.71e-10 ***
omega    0.13059     0.01205   10.834  < 2e-16 ***
alpha1   0.50846     0.08460    6.010 1.85e-09 ***
---

Log Likelihood:
 -354.245    normalized:  -0.6440818

Standardised Residuals Tests:
                                Statistic p-Value
 Jarque-Bera Test   R    Chi^2  0.3903963 0.8226716
 Shapiro-Wilk Test  R    W      0.9978911 0.7328051
 Ljung-Box Test     R    Q(10)  14.34987  0.1576228
 Ljung-Box Test     R    Q(15)  18.65636  0.2297306
 Ljung-Box Test     R    Q(20)  21.02628  0.395587
 Ljung-Box Test     R^2  Q(10)  45.78526  1.568576e-06
 Ljung-Box Test     R^2  Q(15)  52.95579  3.921785e-06
 Ljung-Box Test     R^2  Q(20)  54.4559   4.949604e-05
 LM Arch Test       R    TR^2   56.92809  8.144013e-08
```

Exercise 6.8 Suppose $\{Z_t\} \sim \text{IID} \quad N(0,1)$. Let $\{A_t\}$ and $\{Y_t\}$ be stationary processes defined by

$$A_t = Z_t \sqrt{7 + 0.5A_{t-1}^2}$$

and

$$Y_t = 2 + \frac{2}{3} Y_{t-1} + A_t$$

(1) Give the ACF of the process $\{A_t\}$; the mapping $h \mapsto \rho_A(h)$.
(2) After which lag is the PACF of the process $\{Y_t\}$ zero?

Exercise 6.9 Consider the time series model

$$X_t = 0.3X_{t-1} + Z_t - 0.3Z_{t-1},$$

where $\{Z_t\}$ is a white noise sequence with mean zero and variance equal to 2. State whether this concerns an AR(p), MA(q), ARMA(p, q), ARCH(p), GARCH(p, q) model, WN or combined model. Provide the values for p and q.

CHAPTER 7

Parameter estimation

In this chapter we review two methods for estimating parameters in time-series models. In fact, these are methods that can be applied much more generally, for a whole range of statistical models. We discuss maximum likelihood estimation and least squares estimation. These are certainly not the only methods for parameter estimation: method of moments estimation, Bayesian estimation and spectral estimation are examples of methods which can perform equally well. Then one may ask: how do you measure performance. i.e. when is an estimator good, or even optimal? There is no universal agreement on the way to make this mathematically precise, but a whole lot of theory has been developed. This is really the field of *mathematical statistics*. Here, we simply discuss two methods, which usually (but certainly not always) lead to reasonable parameter estimates.

1. Maximum likelihood estimation

Maximum likelihood estimation is a general estimation method within statistics. First a few remarks on notation: we denote the density of the random variable Y, conditional on X by $f_{Y|X}$. When evaluated in (x, y), this becomes $f_{Y|X}(y \mid x)$. Given observations $X_1, \ldots, X_n$ with the joint density $(x_1, \ldots, x_n) \mapsto f_{X_1,\ldots,X_n}(x_1, \ldots, x_n; \theta)$ that depends on the parameter θ, the *likelihood function* is the function

$$\theta \mapsto f_{X_1,\ldots,X_n}(X_1, \ldots, X_n; \theta).$$

The *maximum likelihood estimator* for θ, if it exists, is the value of θ that maximises the likelihood function. If $X_1, \ldots, X_n$ are independent, then the likelihood factorises into a product of n terms. For time-series models, the likelihood may be of a more complicated form. A useful decomposition of the likelihood is the following

$$\begin{aligned} f_{X_1,\ldots,X_n}(X_1, \ldots, X_n; \theta) = f_{X_1}(X_1; \theta) f_{X_2|X_1}(X_2 \mid X_1; \theta) \\ \ldots f_{X_n|X_{n-1},\ldots,X_1}(X_n \mid X_{n-1}, \ldots, X_1; \theta). \end{aligned}$$

In practise, often an adjusted form of the likelihood is used, which is then referred to as a *pseudo-likelihood*. This is best illustrated by an example.

EXAMPLE 7.1. Suppose $\{X_t\} \sim \text{AR}(p)$ (causal, stationary) with noise sequence $\{Z_t\}$ being IID with marginal $N(0, \sigma^2)$-distribution. Then the pseudo-likelihood is defined by

$$(\sigma, \phi_i, \ldots, \phi_p) \mapsto \prod_{t=p+1}^{n} f_Z(\mathbf{\Phi}(B)X_t; \sigma),$$

where

$$\mathbf{\Phi}(B)X_t = X_t - \phi_1 X_{t-1} - \ldots - \phi_p X_{t-p}.$$

Here, the pseudo-likelihood is obtained from the likelihood by simply dropping some intractable terms. Intuitively, when the sample size n is large, this should not affect the performance of the estimation procedure.

Exercise 7.1 Take $p = 1$ in the preceding example and derive expressions for the maximum likelihood estimator for (σ, ϕ_1).

EXAMPLE 7.2. Suppose $\{X_t\} \sim \text{ARCH}(1)$, with $\{Z_t\}$ a IID sequence with marginal $N(0, 1)$-distribution. Since $X_t \mid X_{t-1} \sim N(0, \omega + \alpha X_{t-1}^2)$ the pseudo likelihood can be defined as the mapping

$$(\omega, \alpha) \mapsto \prod_{t=2}^{n} \frac{1}{\sqrt{2\pi(\omega + \alpha X_{t-1}^2)}} \exp\left(-\frac{x_t^2}{2(\omega + \alpha X_{t-1}^2)}\right).$$

This function can be numerically optimised; a closed-form expression not being available.

In a wide variety of cases it can be proved that maximum likelihood estimators have certain optimality properties asymptotically. We do not go into details.

2. Least squares estimation

Suppose the time series $\{X_t\}$ has mean zero: $\text{E}[X_t] = 0$. We denote the parameter (vector) in the model by θ. As an example, in a AR(2)-model we have $\theta = (\phi_1, \phi_2, \sigma^2)$, where σ^2 is the variance of the noise sequence. Least squares estimation is based on comparing a prediction of X_t (which depends on θ), with X_t itself. Suppose we are given observations $X_1, \ldots, X_n$. Denote $P_{t-1}X_t$ the predictor of X_t based on $X_1, \ldots, X_{t-1}$. As it depends on θ, we will write $P_{t-1}X_t(\theta)$.

DEFINITION 7.3. A *least squares estimator* is any θ for which the mapping

$$\theta \mapsto \sum_{t=1}^{n} (X_t - P_{t-1}X_t(\theta))^2$$

is minimal.

The underlying idea is that the "true" θ should give "best predictions". We haven't specified the predictor itself yet. Two possibilities are

- the best (nonlinear) predictor $P_{t-1}X_t(\theta) = \text{E}[X_t \mid X_{t-1}, \ldots, X_1]$.

- the best (linear) predictor $P_{t-1}X_t(\theta) = \beta_1 X_{t-1} + \ldots \beta_{t-1}X_1$. As we now from section 4 that the best linear predictor depends on $h \mapsto \gamma_X(h;\theta)$, the least squares estimator will also depend on the data via the autocorrelations.

If the best predictor is tractable, then using it is preferable over the best linear predictor (which is more restricted in its form). However, in many models this is simply not the case. The quantity $X_t - P_{t-1}X_t(\theta)$ is known as the *innovation*, and then least squares estimation can be reformulated as minimising the sum of the squared innovations.
So far, we have assumed that $\mathrm{E}[X_t] = 0$. If this is not the case, then we can either plug-in the estimator $\bar{X}_n$ for μ, or apply the preceding discussion to the time-series $X_t - \mu$ (which as mean zero) and add μ as an extra parameter to θ.

EXAMPLE 7.4. If $\{X_t\} \sim \mathrm{AR}(p)$, then the best predictor and best linear predictor coincide:

$$P_{t-1}X_t(\theta) = \phi_1 X_{t-1} + \ldots, \phi_p X_{t-p}.$$

Exercise 7.2 Derive an expression for the least squares estimator when $p = 1$.

EXAMPLE 7.5. Suppose $\{X_t\} \sim \mathrm{ARCH}(p)$. As $\{X_t\}$ is a white-noise sequence, we have $P_{t-1}X_t = 0$. The least squares criterion is then given by $\sum_{t=1}^n X_t^2$, which does not depend on the parameter. Clearly, in this case least squares estimation does not make sense. Sometimes, instead *weighted least squares* estimation is proposed, where θ is determined by minimising

$$\theta \mapsto \sum_{t=1}^{n} \frac{(X_t - P_{t-1}X_t(\theta))^2}{v_{t-1}(\theta)},$$

where $v_{t-1}(\theta) = \mathrm{E}_\theta\left[(X_t - P_{t-1}X_t(\theta))^2\right]$ is the (squared) prediction error. Unfortunately, also this adjusted criterion does not make much sense for ARCH-models.

CHAPTER 8

Appendix

1. Setting up your R working environment

First install R, then install Rstudio. Open Rstudio. We will make use of certain libraries specifically designed for time-series analysis in R. You need to install these on your computer. In R-studio go to Packages - install packages. Choose the packages: forecast, tseries, astsa, fGarch and quantmod. You only need to do this once.
Next, each time you open Rstudio, run the following few lines of code:

```
library(forecast)
library(tseries)
library(astsa)
library(fGarch)
library(quantmod)
```

2. List of R commands

The following list of commands will be handed out at the exam.

mean	average
sd	standard deviation
var	variance
diff	differencing of a vector
acf	autocorrelation function
Acf	autocorrelation function (lag 0 omitted)
pacf	partial autocorrelation function
tsdisplay	plot time-series, ACF and PACF
Box.test	Box-Pierce or Ljung-Box test
kurtosis	excess kurtosis
qqnorm and qqline	make normal probability plot
ARMAacf	theoretical ACF for an ARMA process
arima.sim	simulate from a ARMA model
sarima	fit ARMA-model (use "$fit" for the fitted object)
sarima.for	forecast, assuming a ARMA-model
garchSpec	specify GARCH model.
garchSim	simulate from GARCH-model
garchFit	fit GARCH-model
predict	forecasts for fitted GARCH-model
residuals(gfit)	residuals of fitted GARCH-model in gfit
residuals(gfit, standardize=T)	standardised residuals of fitted GARCH-model in gfit
volatility(gfit)	volatility of fitted GARCH-model in gfit

An example of reading data from finance.yahoo.com

```
# read the data
getSymbols("PHG",src='yahoo')
# subset
PHG['2007-06::2008-01-12']
# OHLC chart
chartSeries(PHG,subset='last 2 months',theme=chartTheme('white'))
```

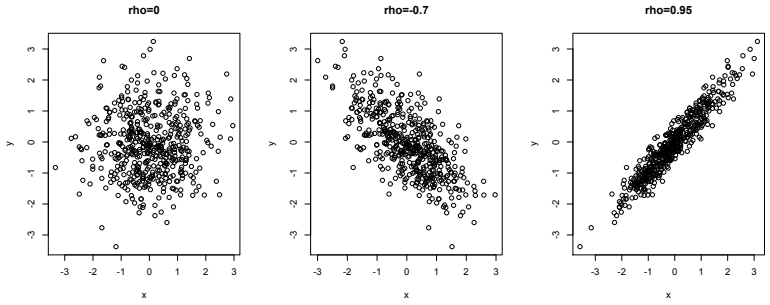

FIGURE 1. 500 realisations of the pair (X_1, X_2) (as defined in (15)) for various values of ρ. Left: $\rho = 0$, middle: $\rho = 0.7$, right: $\rho = 0.95$.

3. The multivariate normal distribution

DEFINITION 8.1. Suppose $Z_1, \ldots, Z_k$ are independent $N(0,1)$-distributed random variables. Define $Z = \begin{bmatrix} Z_1 & \cdots & Z_k \end{bmatrix}'$. A k-dimensional random vector X has the *multivariate normal distribution* with mean vector μ and covariance matrix Σ if X has the same probability distribution as the vector $\mu + LZ$, for a $k \times k$ matrix L with $\Sigma = LL'$ and k-dimensional vector μ. The density of X is then given by

$$f_X(x) = (2\pi)^{-k/2}(\det\Sigma)^{-1/2} \exp\left(-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right).$$

If Σ is diagonal, then the density factorises into a product of k densities of the univariate normal distribution.

EXAMPLE 8.2. Let's consider a particular construction in dimension 2 (leading to the *bivariate* normal distribution. Start with independent random variables $Z_1 \sim N(0,1)$ and $Z_2 \sim N(0,1)$. Take $\rho \in (-1,1)$ and put

$$\begin{aligned} X_1 &= \sqrt{1-\rho^2}Z_1 + \rho\, Z_2 \\ X_2 &= \qquad\qquad\quad Z_2 \end{aligned} \tag{15}$$

Then we can easily verify that $X_2 \sim N(0,1)$ and $X_1 \sim N(0,1)$; In Figure 1 we simulated 500 realisations of the pair (X_1, X_2) for various values of ρ. This figure illustrates that ρ measures *linear* dependence. In this example it can be proved that in fact $\rho(X_1, X_2) = \rho$. From (15) it is clear that X_1 and X_2 are independent if and only if $\rho = 0$. Hence, in this specific example: $\rho(X_1, X_2) = 0$ implies that X_1 and X_2 are independent.
In general, the bivariate normal distribution has 5 parameters: $\mu_1, \mu_2, \sigma_1, \sigma_2$ and ρ. Sampling from this distribution can be done by sampling Z_1 and Z_2 independently from the standard normal distribution and setting

$$\begin{aligned} X_1 &= \mu_1 + \sigma_1(\sqrt{1-\rho^2}Z_1 + \rho\, Z_2) \\ X_2 &= \mu_2 + \sigma_2 \qquad\qquad\quad Z_2 \end{aligned}$$

CHAPTER 9

Hints to exercise solutions

Solution 1.3

$$\ell_t = \log \frac{x_t}{x_{t-1}} = \log e^{\alpha(t+1)-\alpha t} = \alpha.$$

Solution 2.1 Hint: think of the Pareto distribution.

Solution 2.2 If t is even, then this is trivial. If t is odd, then

$$\mathrm{E}[X_t] = 2^{-1/2}(\mathrm{E}\left[Z_{t-1}^2\right] - 1) = 0$$

since for the standard normal distribution its second moment equals 1.

Solution 2.3 If $h = 0$, using that the $\{Z_t\} \sim \mathrm{WN}$ we get

$$\mathrm{Var}(X_t) = (1 + \theta^2)\sigma^2.$$

For $h = 1$, using that the covariance-operator is half-linear, the expression is equal to

$$\mathrm{Cov}(Z_t, Z_{t+1}) + \theta\mathrm{Var}(Z_t) + \theta\mathrm{Cov}(Z_{t-1}, Z_{t+1}) + \theta^2\mathrm{Cov}(Z_{t-1}, Z_t)\,.$$

Now only the second term does not vanish and it equals $\theta\sigma^2$. Proving that all autocorrelations at lags $h \geq 2$ goes similarly.

Solution 2.4 Here, only the solution is given, not the derivation:

(1) $\mu_X(t) = a$ and

$$\gamma_X(t) = \begin{cases} (b^2 + c^2)\sigma^2 & \text{if } h = 0 \\ bc\sigma^2 & \text{if } |h| = 2 \\ 0 & \text{otherwise} \end{cases}$$

The process is (weakly) stationary.

(2) $\mu_X(t) = 0$ and

$$\gamma_X(t) = \sigma^2 \cos(ch)$$

using that $\cos(\alpha - \beta) = \cos\alpha\cos\beta + \sin\alpha\sin\beta$. The process is (weakly) stationary.

(3) $\mu_X(t) = a$ and

$$\gamma_X(t) = \sigma^2 b^2.$$

The process is (weakly) stationary.

(4) $\mu_X(t) = 0$ and

$$\gamma_X(t) = \begin{cases} \sigma^2 & \text{if } h = 0 \\ 0 & \text{otherwise} \end{cases}.$$

The process is (weakly) stationary.

Solution 2.5

(1)

$$\gamma_X(h) = \begin{cases} 1+\theta^2 & \text{if } h = 0 \\ \theta & \text{if } |h| = 2 \end{cases} = \begin{cases} 1.64 \\ 0.8 \end{cases}$$

Hence

$$\rho_X(h) = \frac{\gamma_X(h)}{\gamma_X(0)} = \begin{cases} 1 & \text{if } h = 0 \\ 0.8/1.64 \approx 0.49 & \text{if } |h| = 2 \end{cases}.$$

(2) We have

$$\mathrm{Var}((X_1 + X_2 + X_3 + X_4)/4) =$$

$$\frac{1}{16}\left(\mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \mathrm{Var}(X_3) + \mathrm{Var}(X_4) + 2\mathrm{Cov}(X_1, X_3) + 2\mathrm{Cov}(X_2, X_4)\right).$$

This is seen to equal $(\gamma_X(0) + \gamma_X(2))/4 = (1.64 + 0.8)/4 = 0.61$.

(3) In this case $(\gamma_X(0) + \gamma_X(2))/4 = (1.64 - 0.8)/4 = 0.21$. Considerably smaller due to the negative covariance at lag 2.

Solution 2.10 Note that the best linear predictor is given by

$$P_{1:t}X_{t+h} = a_0 + a_1 X_t + \ldots + a_t X_1.$$

Now if $\mathrm{E}[X_t] = 0$ then $a_0 = 0$. We need to compute $P_{1:3}X_4$, so $t = 3$ and $h = 1$. Equation (5) implies that

$$\begin{bmatrix} \gamma(0) & \gamma(1) & \gamma(2) \\ \gamma(1) & \gamma(0) & \gamma(1) \\ \gamma(2) & \gamma(1) & \gamma(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} \gamma(1) \\ \gamma(2) \\ \gamma(3) \end{bmatrix}.$$

Now divide both sides by $\gamma(0)$, then each $\gamma(h)$ gets replaced with $\rho(h)$. Furthermore, use the fact that for a MA(1)-process $\rho(2) = \rho(3) = 0$. Denote $\rho(1)$ by η. Then the system reduces to

$$\begin{bmatrix} 1 & \eta & 0 \\ \eta & 1 & \eta \\ 0 & \eta & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} \eta \\ 0 \\ 0 \end{bmatrix}.$$

Note that if $X_t = Z_t + \theta Z_{t-1}$, then η can be expressed in terms of θ. The remainder is just simple linear algebra. The third equation gives $a_3 = -\eta a_2$. Furthermore we find $a_2 = \frac{\eta^2}{2\eta^2 - 1}$ and $a_1 = \eta(1 - a_2)$.

Solution 2.13

```
simforward <- function(alpha,N=10000)
{
  z <- rnorm(N)
  x <- rep(0,N)
  for (i in 2:N)
  {
    beta <- ifelse(x[i-1]<0,-1.5,alpha)
    x[i] <- beta * x[i-1] + z[i]
  }
  x
}

par(mfrow=c(3,3))
alphas = c(0.5, 1, 1.5) # vector of alphas we want to consider
for (j in 1:3)
{
  for (b in 1:3) # simulate 3 times for each alpha
  {
    x = simforward(alphas[j],100)
    # I take 100 here, as otherwise the time series with alpha=1.5
        explodes to infinity
    plot.ts(x,main=as.character(alphas[j]))
  }
}
```

Solution 3.2 The model for monthly data with a yearly seasonal pattern is given by $X_t = X_{t-12} + Z_t$, where $\{Z_t\} \sim$ WN. Hence $Y_t = \nabla^{12} X_t = X_t - X_{t-12} = W_t$. As a numerical illustration,

```
n <- 60
SD <- 0.25
x <- sin(seq(0,2*pi,l=12))
for (i in 13:n) x[i] <- x[i-12] + rnorm(1,sd=SD)
y <- diff(x,lag=12)

par(mfrow=c(2,2))
plot.ts(x,main='Seasonal time-series'); points(x)
abline(v=seq(0,120,by=12),col='red')
acf(x,main='Seasonal time-series')

plot.ts(y,main='De-seasonalised time-series');points(y)
abline(v=seq(0,120,by=12),col='red')
acf(y,main='De-seasonalised time-series')
```

This gives Figure 1.

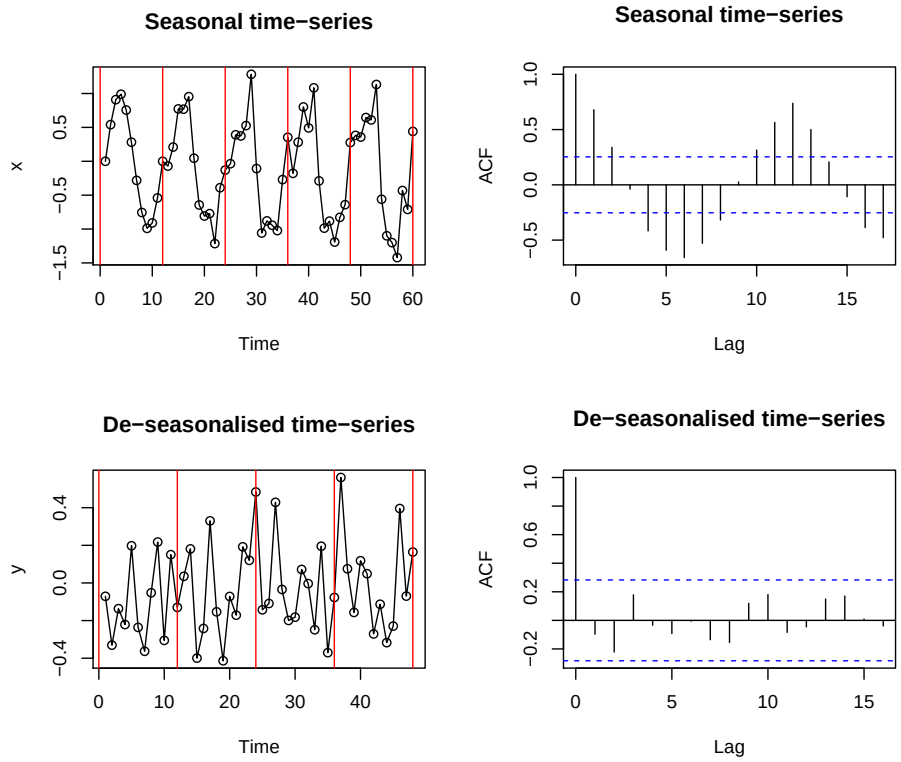

FIGURE 1. Top: seasonal pattern with ACF. Bottom: de-seasonalised pattern with ACF.

Solution 3.6

```
th <- -0.4
N <- 250
y1 <- arima.sim(n=N, list(ma = c(0,0,th)))
y2 <- arima.sim(n=N, list(ma = c(0,0,th)))
par(mfrow=c(1,1))
plot.ts(y1)
lines(y2,col='red',ylab="MA(3)(time)")

par(mfrow=c(1,2))
acf(y1)
acf(y2)

acf(diff(y1))

sd(y1)
sd(y2)
# should be
```

```
sqrt(1+th*th)
```

Solution 3.8

(1)

$$\begin{aligned}\mathrm{Cov}(Y,Z) &= \mathrm{E}[YZ] - \mathrm{E}[Y]\,\mathrm{E}[Z] = 1 - \left(\int_0^1 e^{-x}dx\right)\left(\int_0^1 e^{x}dx\right) \\ &= 1 - (1-e^{-1})(e-1) = 3 - e - e^{-1}\end{aligned}$$

(2)

```
x <- runif(10000)
cov(exp(-x),exp(x)) # gives -0.08518148
3-exp(1)-exp(-1) # true value -0.08616127
```

(3) If they would be independent, then they should be uncorrelated, which is not the case.

Solution 4.2 The equations are

$$\begin{aligned} X_t &= Z_t + \frac{1}{5}Z_{t-1} \qquad \{Z_t\} \sim \text{IID } N(0,25) \\ Y_t &= Z_t + 5Z_{t-1} \qquad \{Z_t\} \sim \text{IID } N(0,1). \end{aligned}$$

Hence for $\{X_t\}$ we have $\Theta(B) = 1 + \frac{1}{5}B$ while for $\{Y_t\}$ we have $\Theta(B) = 1 + 5B$. Now just verify the location of solutions to $|\Theta(B)| = 0$.

Solution 4.4 On the one hand the ACF plot shows that a MA(2)-process is reasonable, on the other hand the PACF plots show that a AR(2)-process is reasonable. The evidence for the latter being appropriate seems somewhat stronger.

Solution 4.6 This process is of AR(2)-type and therefore invertible (put otherwise, in ARMA-representation, $\Theta(B) \equiv 1$). Write $\Phi(B)X_t = Z_t$ with $\Phi(B) = 1 - \psi B^2$. Then $|\Phi(B)| = 0$ gives $B = \pm 1/\sqrt{\psi}$. Hence $|B| > 1$ iff $|\psi| < 1$. For these values a stationary, causal version of the process exists.

Solution 4.9

(1) Assume $h > 1$. Multiply both sides of the equation by X_{t-h} to get

$$\mathrm{Cov}(X_t, X_{t-2}) = \theta \mathrm{Cov}(X_{t-2}, X_{t-h}) + \mathrm{Cov}(W_t, X_{t-h})\,.$$

By causality, X_{t-h} can only dependent on $W_{t-h}, W_{t-h-1}, \ldots$ (as a linear combination). As $\{W_t\}$ is white-noise, the rightmost term must be zero. We conclude that

$$\gamma_X(h) = \theta\gamma_X(h-2).$$

The answer follows by dividing both sides by $\gamma_X(0)$.

(2) Take $h = 1$ and use that $\rho_X(h) = \rho_X(-h)$. Then $(1-\theta)\rho_X(1) = 0$. As the process is stationary, we cannot have $\theta = 1$. Hence $\rho_X(1) = 0$.
(3)
$$\rho_X(4) = \theta\rho_X(2) = \theta\theta\rho_X(0) = \theta^2.$$

Solution 6.6

(1) $\{A_t\} \sim$ ARCH(2), hence white noise. Therefore, $\rho_A(h) = 1$ if $h = 0$, else it equals zero.
(2) We first rewrite the model equation to
$$Y_t - 6 = \frac{2}{3}(Y_{t-1} - 6) + A_t.$$
So $U_t := Y_t - 6$ is of AR(1)-type. Hence $\rho_U(h) = (2/3)^h$. Clearly, as $\{U_t\}$ and $\{Y_t\}$ only differ in an additive constant we have $\rho_U(h) = \rho_Y(h)$.
(3) Heavier.
(4) as $\{A_t\} \sim$ ARCH(2), we know that $\{A_t^2\} \sim AR(2)$, so its PACF will be zero after lag 2.

Solution 6.7

(1) The fitted model is of AR(2) + ARCH(1) type. Hence
$$X_t = 0.74X_{t-1} - 0.23X_{t-2} + Z_t$$
with
$$Z_t = \sigma_t W_t, \qquad \sigma_t^2 = 0.13 + 0.51Z_{t-1}^2.$$
Here $\{W_t\}$ is IID with the standard Normal distribution.
(2) Both the Jarque-Bera and Shapiro-Wilk test give no indication to reject normality (Note: in 2019 I only discussed normal-probability plots for verifying normality, hence I would state the question differently).
(3) Yes, at least, the Ljung-Box-tests for `R` give no reason to doubt this assumption.
(4) There is still significant autocorrelation in the squared residuals (look at Ljung-Box-tests), so the fit is not reasonable (yet).

Solution 6.8

(1) Note that $\{A_t\} \sim$ ARCH(1), so $\rho_A(h) = 1$ if $h = 0$ and else 0.
(2) $\{Y_t\} \sim$ AR(1) and hence the PACf is zero after lag 1.

Solution 6.9 This can be written as $\Phi(B)X_t = \Theta(B)Z_t$, with
$$\Phi(B) = 1 - 0.3B \qquad \Theta(B) = 1 - 0.3B.$$
Hence, $X_t = Z_t$, so $\{X_t\}$ is white-noise.

Bibliography

Peter J. Brockwell and Richard A. Davis. *Introduction to time series and forecasting*. Springer Texts in Statistics. Springer-Verlag, New York, second edition, 2002. ISBN 0-387-95351-5. doi: 10.1007/b97391. URL https://doi.org/10.1007/b97391. With 1 CD-ROM (Windows).

J. Durbin and S. J. Koopman. *Time series analysis by state space methods*, volume 38 of *Oxford Statistical Science Series*. Oxford University Press, Oxford, second edition, 2012. ISBN 978-0-19-964117-8. doi: 10.1093/acprof:oso/9780199641178.001.0001. URL https://doi.org/10.1093/acprof:oso/9780199641178.001.0001.

Robert Engle. Garch 101: The use of arch/garch models in applied econometrics. *Journal of Economic Perspectives*, 15(4):157–168, 2001.

Steven E. Shreve. *Stochastic calculus for finance. I.* Springer Finance. Springer-Verlag, New York, 2004. ISBN 0-387-40100-8. The binomial asset pricing model.

Robert H. Shumway and David S. Stoffer. *Time series analysis and its applications*. Springer Texts in Statistics. Springer, New York, third edition, 2011. ISBN 978-1-4419-7864-6. doi: 10.1007/978-1-4419-7865-3. URL https://doi.org/10.1007/978-1-4419-7865-3. With R examples.

Ruey S. Tsay. *Analysis of financial time series*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, third edition, 2010. ISBN 978-0-470-41435-4. doi: 10.1002/9780470644560. URL https://doi.org/10.1002/9780470644560.