Lecture Notes

Time Series Analysis

Alexander Lindner
Ulm University

Winter semester 2025/26

Contents

Bibliography

[BD1] Brockwell, P.J. and Davis, R.A. (1990): Time series: theory and methods. 2nd edition, Springer.

[BD2] Brockwell, P.J. and Davis, R.A. (2016): Introduction to time series and forecasting. 3rd edition, Springer.

[Fu] Fuller, W.A. (1996): Introduction to statistical time series. 2nd edition. Wiley.

[Ha] Hamilton, J.D. (1994): Time series analysis. Princeton Univ. Press.

[KN] Kreiß, J.-P. and Neuhaus, G. (2006): Einführung in die Zeitreihenanalyse, Springer.

Foreword

Dear students,

these are the lecture notes for the course on Time Series Analysis.

For whom is this course?

This course is (mainly) for master students in mathematics, business mathematics, mathematical biometry, finance, data science or CSE. Other students are welcome, and if they are allowed to take the exam by their Study and Examination Regulations, they may also take the exam.

Time of lectures

The course takes place in the first half of the semester, i.e. from October 14th, 2025, to December 4, 2025. Lecturing hours are on Tuesdays, 10:15 – 11:45 in E60 (Helmholtzstraße 18) and on Thursdays, 08:30 – 10:00 in E20 (Helmholtzstraße 18). The first lecture is on Thursday, October 14th.

Continuation of this lecture

In the second half of the semester, i.e. from December 9, 2005, to February 12, 2026, will be the course 'Advanced topics in time series analysis'. This is basically a continuation of the course on 'Time series analysis', but is formally independent from it. The Advanced Topics will require knowledge of 'Time Series Analysis', but are examined in a different exam. It is possible to take 'Time Series Analysis' also without taking 'Advanced topics in time series analysis'.

Exercises

Exercise sheets will be handed out every week in the first half of the semester in Moodle. You do not have to hand them in and they are no prerequisites for the exam. The exercises are organised by Sebastian Aichmann. More information regarding the exercises will be posted separately.

Exam

We will be having an oral exam. It can be taken in the second half of the semester, or in the semester break in March/April 2026.

Literature

The lecture notes are based on the book [BD1] by Brockwell and Davis. This is really an excellent book and I will mostly follow it.

As further literature, I recommend the books by

- Brockwell, P.J. and Davis, R.A. [BD2]: This book covers more the practical aspects of time series analysis than the other book by Brockwell and Davis.
- Fuller, W.A. [Fu]: A very good book, which however concentrates more on the statistical aspects of time series.
- Hamilton, J.D. [Ha]: Time series analysis. Princeton Univ. Press. (Very good book, but the book is a big one)
- Kreiß, J.-P. and Neuhaus, G. [KN]: This book is also excellent in my opinion and close to what I will be doing in this lecture. Unfortunately, so far it is only available in German, although the authors are working on an English version.

Alexander Lindner.

Chapter 1

Introduction

This is an introductory chapter. We address the question what a time series is and present some examples. There are two possible definitions of a time-series. One is that it is a collection of data in time, the other is that it is a certain stochastic process which is a model for these data. From the practical point of view, the first definition is the relevant one, from the theoretical point of view, the second. Let us first give the practical definition.

Definition 1.1. A *time series* is a sequence of observations $x_1, \dots, x_T$, or $(x_n)_{n \in \mathbb{N}}$, or $(x_n)_{n \in \mathbb{Z}}$, recorded at specific time points $1, 2, \dots$ When the observations are one-dimensional, we speak of *univariate time series*, when they are multi-dimensional, we speak of *multivariate time series.*

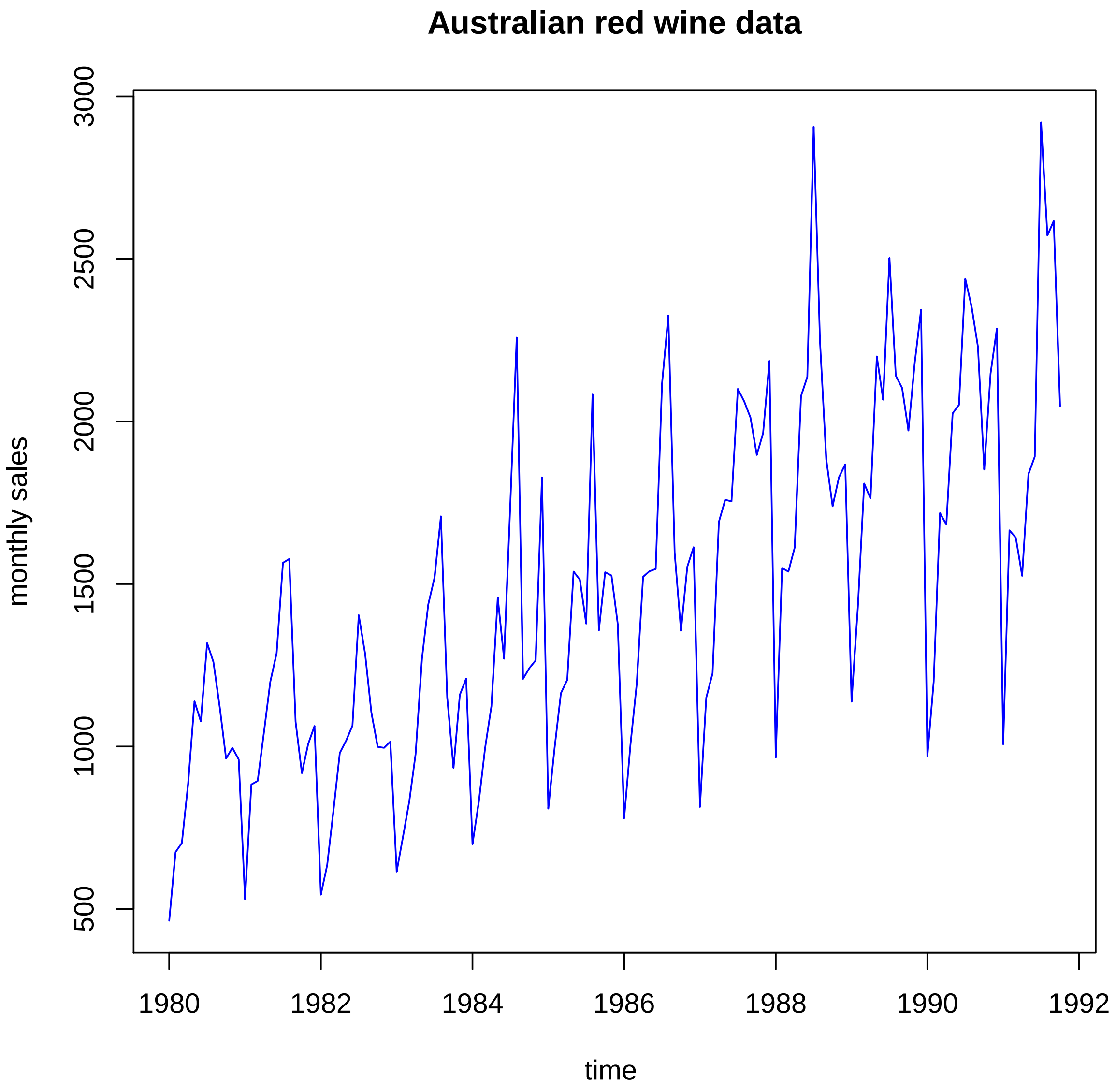
Example 1.2. Denote by x_t the monthly sales of Australian red wine per month in kilolitres, taken from ITSM (a programme incorporated in the book [BD2] by Brockwell and Davis), starting from January 1980 ($t = 1$) over February 1980 ($t = 2$) to October 1991 ($t = 142$). The graph is displayed in Figure 1.1. When looking at the data, we see the following features:
(i) There seems to be an upward trend, i.e. on average people consume more read wine when time moves on.
(ii) There is a seasonal pattern, namely most red wine is drunk in July, while not so much in January. This is because red wine is more often drunk in winter than in summer, and in Australia, January is summer while July is winter.
(iii) There is an increase in variability, meaning that the fluctuations become larger and larger with increasing time.

So what does one usually do first when one has a time series and is interested in its structure?
The first thing is to plot the data. Then one can often already see if

- there is a trend over time, i.e. if the data increase or decrease in time, and if so, how they do that

- there is a seasonal pattern, or some cyclic pattern

Figure 1.1: Monthly sales of Australian read wine per months, Jan 1980 – Oct. 1991, as discussed in Example 1.2

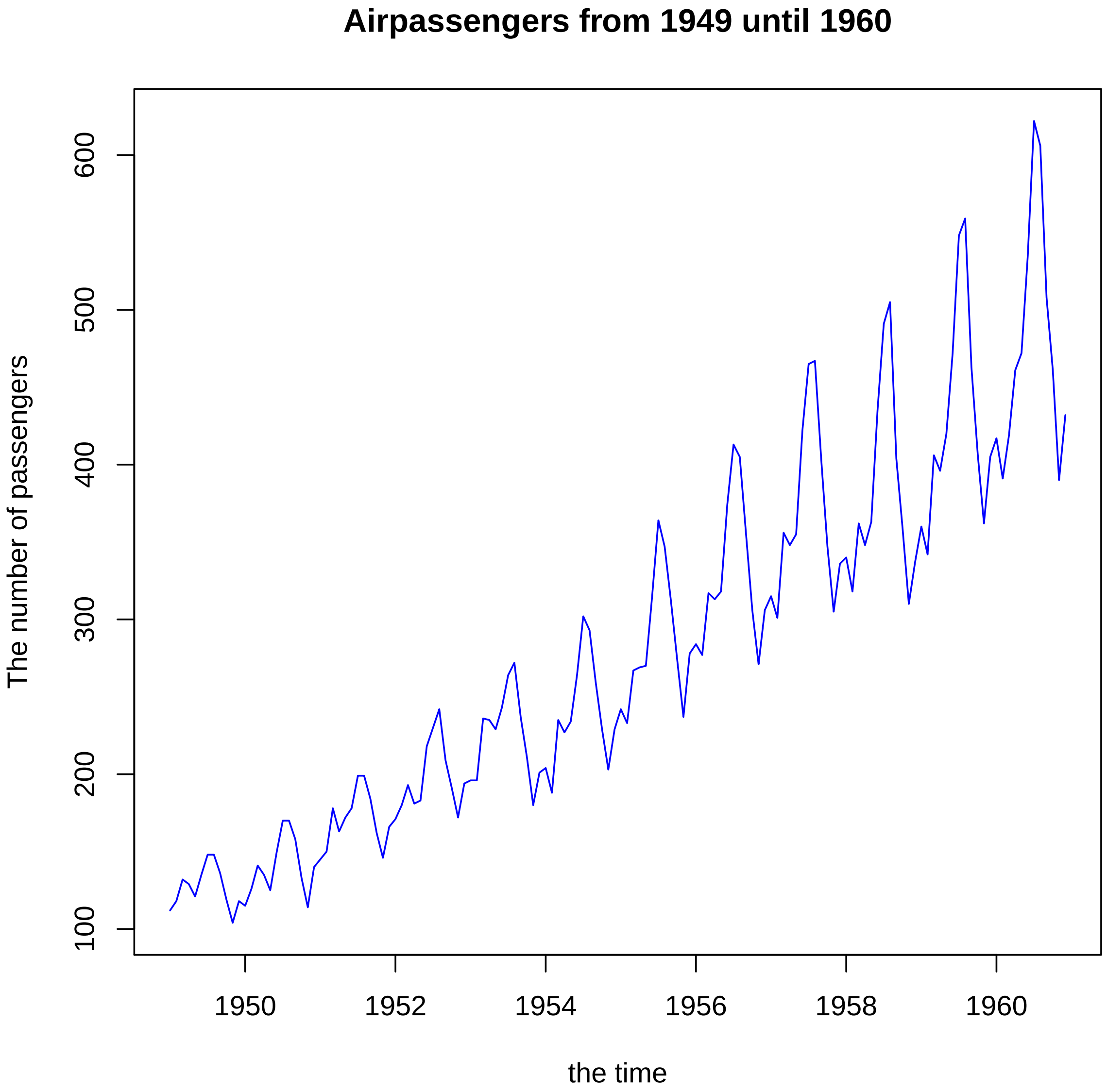


- the variability is constant over time or varies
- there are other systematic features present in the data.

In the third chapter we will be concerned with identifying trend or seasonality. In this chapter, we look at some specific examples.

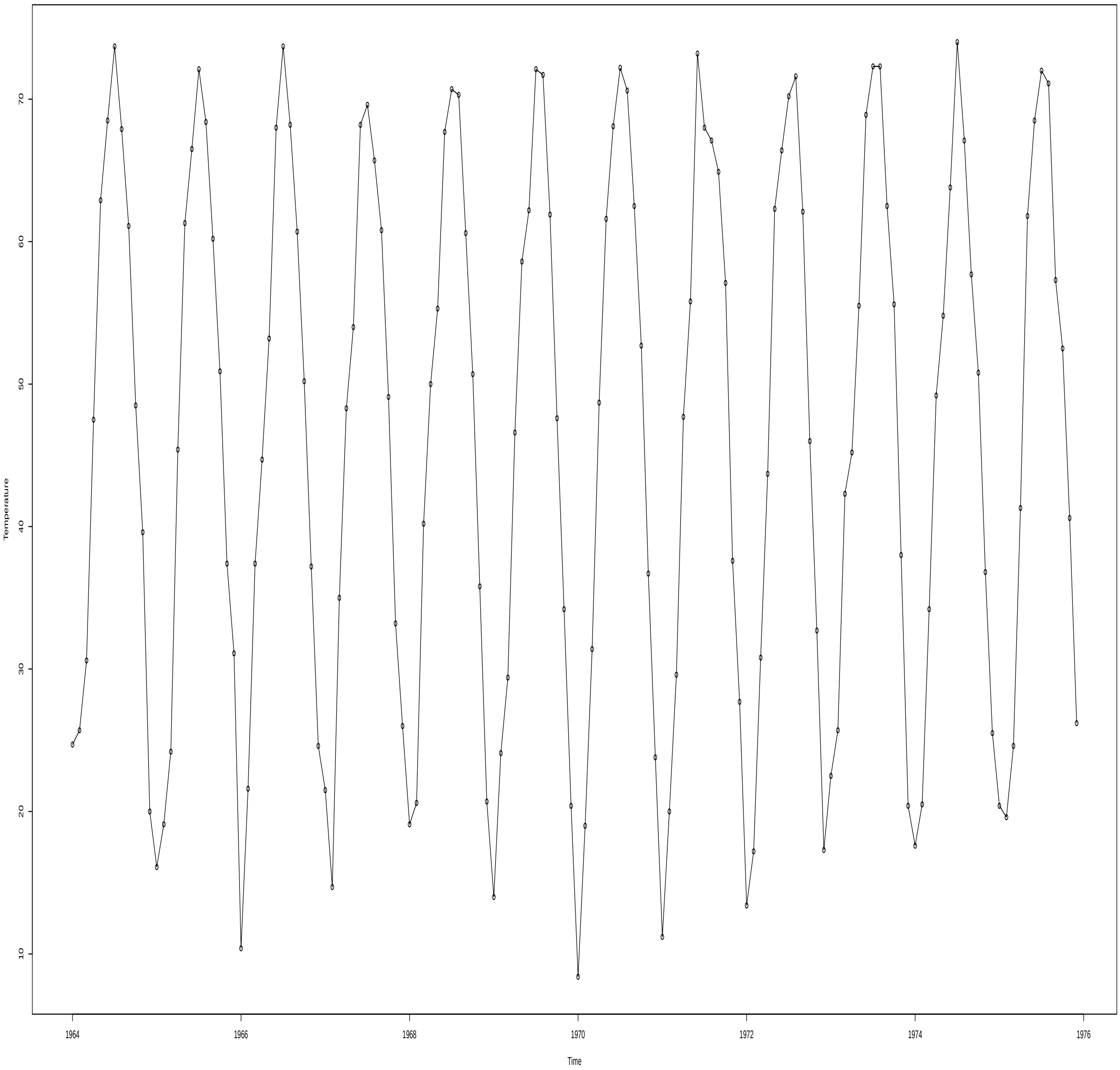
Example 1.3. Figure 1.2 shows the monthly total airline passenger numbers (in thousands), from January 1949 – December 1960, taken from ITSM. They seem to exhibit a linear upwards trend and a seasonal component, as well as increase in variability.

Figure 1.2: International airline passenger data (in thousands) from Janaury 1949 – December 1960.



Example 1.4. Figure 1.3 shows the average monthly temperature (in Fahrenheit) at the place Dubuque (which is in Iowa, USA), in the time period 1964 – 1976. The data are taken from the R-library TSA, data(tempdub). The data seem to exhibit a seasonal component, but no trend.

Figure 1.3: Monthly temperature at Dubuque in Iowa from 1964 – 1976

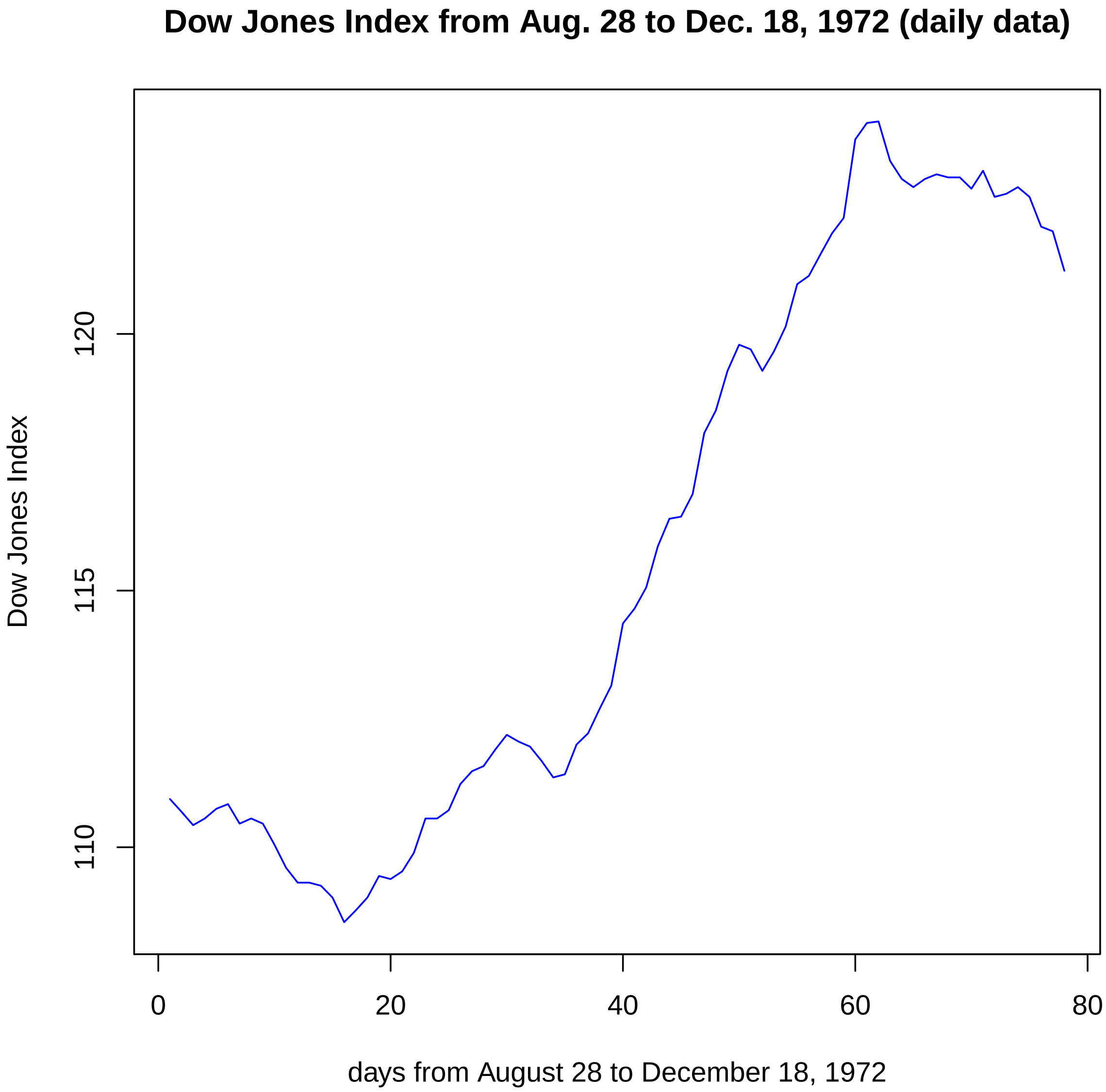


Example 1.5. Figure 1.4 shows the Dow Jones Utilities Index from Aug. 28 – Dec. 18, 1972 (daily data), taken from ITSM. Looking at it there seems to be an upwards trend, but no seasonal component. It is a typical financial time series. A financial time series is often much better analysed using the differenced series or the log returns.

Example 1.6. Figure 1.5 displays the daily log returns (multiplied by 100) based on closing prices $P(t)$ of 7 stock indices. The log return is defined as

$$\log \frac{P(t)}{P(t-1)} = \log P(t) - \log P(t-1);$$

Figure 1.4: Dow Jones Utilities Index from Aug. 28 – Dec. 18, 1972 (daily)



here, log denotes the natural logarithm. The first return is for July 2, 1997, the last is for April 9, 1999. The indices are: Australian All-ordinaries, Dow-Jones Industrial, Hang Seng, JSI (Indonesia), KLSE (Malaysia), Nikkei 225, KOSTI (South Korea), and the data are taken from ITSM. No trend or seasonal pattern visible. We will hardly analyse multivariate data in this course, but of course there may be connections between the single time series present here.

Example 1.7. Further examples of time series include e.g.

- Monthly unemployment rate in Germany,
- annual bottom water level of the river nile,
- Canadian lynx data: number of captured lynx from 1821 to 1934 at the MacKenzie River,
- stock prices, rates of exchange, log returns
- sunspot number (average number per year),
- accident data,
- diseases, clinical trials, etc.

So, we have real data given. But how do probability and statistics come in? The idea is that one models a phenomenon as a stochastic process, and then views the given time series as one realisation of this stochastic process. Let us recall what a stochastic process (with a time domain specified below) is.

Definition 1.8. A real valued or $\mathbb{R}^d$-valued *stochastic process* with index set $T = \mathbb{N}$, $T = \mathbb{N}_0$ or $T = \mathbb{Z}$ is a sequence $(X_t)_{t\in T}$ of random variables defined on a probability space $(\Omega, \mathcal{F}, P)$. We call $(X_t)_{t\in T}$ a *time series.* The functions $(T \ni t \mapsto X_t(\omega))$ for $\omega \in \Omega$ are called *paths* or *realizations* of the time series. In reality, one usually observes only one path.

So this is the second and theoretical definition of a time series, namely simply as a stochastic process $(X_t)_{t\in T}$. As said, for most of our analysis in these notes we will work with the second definition, with the understanding that the real data time series is a realisation $x_t = X_t(\omega)$ for all t and some fixed $\omega \in \Omega$, so basically we see one path of the stochastic process.

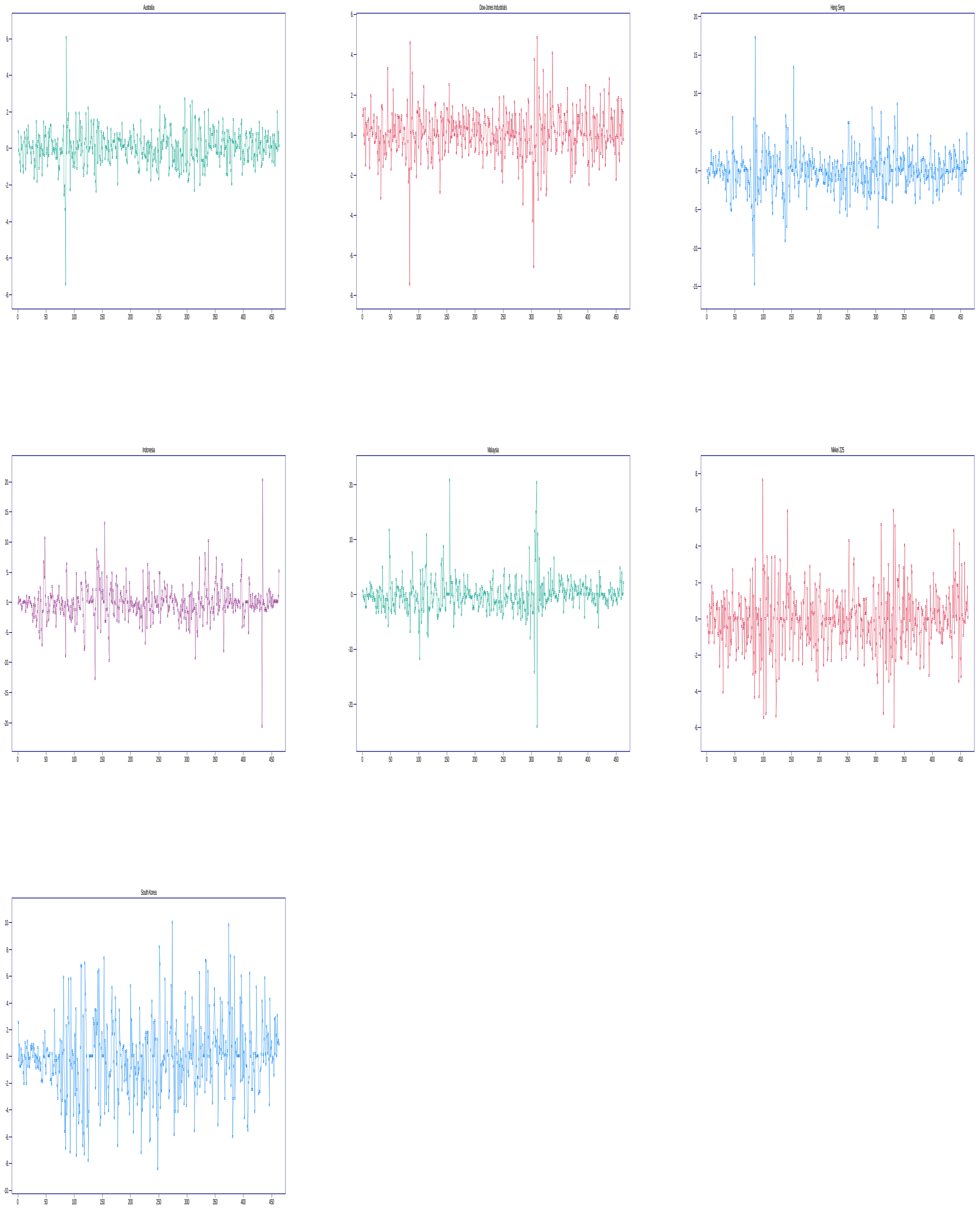
The interesting feature in time series analysis is that the stochastic process $(X_t)_{t\in T}$ does not come from an i.i.d. sequence, but that there are dependencies between different X_t.

Task 1.9. What are the objectives of time series analysis?
(i) The first thing is that one finds a model type for the observed data (x_t), which is usually done using a stochastic process, and one assumes that one has exactly one realisation of that.
(ii) Then one is interested in identifying the “obvious” patterns, like trend, seasonality and other deterministic quantities. Although we will not work much with those, this is actually one of the most important things and often has the most influence. We shall treat this shortly in the third chapter.
(iii) Having estimated trend, seasonality and other deterministic quantities (call their sum y_t), one subtracts these from the data and is left with the “residuals”$z_t = x_t - y_t$. In those no obvious structure should be present any more.
(iv) Then fit a stochastic model to the remaining residuals (z_t). For that, one needs to know a variety of model classes.

(v) Having done this, one should check the model for goodness of fit. This means that it may well be that one has fitted a model class that is not good at all. If so, one has to start again and fit another model class.
(vi) Supposing the model is good, one can use this to forecast (i.e. predict) future values. This is of course the issue one is most interested in, to predict.
(vii) If the model is good, one can also use it to test certain hypotheses.
These are the tasks we try to carry out (at least partially).

In the next chapter we shall treat the notion of stationarity. This is the property we usually impose on the residuals.

Figure 1.5: Daily log returns of 7 indices as described in Example 1.6



Chapter 2

Stationary time series and the autocorrelation function

In this chapter we will learn what a stationary time series is. The idea is that one usually estimates or eliminates trend and seasonality in data first, and then tries to model the residuals with a stationary model. So what does stationarity mean? There are two notions of stationarity, namely strict stationarity and weak (or second order) stationarity.

Suppose throughout that $T = \mathbb{N}$, $T = \mathbb{N}_0$ or $T = \mathbb{Z}$. For simplicity, we assume that our time series $(X_t)_{t \in T}$ is real valued (or complex valued, later), and we think of a time series as a stochastic process.

Definition 2.1. A time series $(X_t)_{t\in T}$ is said to be *strictly stationary*, if for all $t_1, \dots, t_n \in T$, $k \in \mathbb{N}$ holds that

$$(X_{t_1}, X_{t_2}, \dots, X_{t_n}) \overset{\mathrm{d}}{=} (X_{t_1+k}, X_{t_2+k}, \dots, X_{t_n+k}),$$

i.e. the finite dimensional distributions are shift invariant.

So not only do I have the same distribution at every time point, but also the distribution of (X_1, X_2) is the same as the distribution of (X_5, X_6), so the dependency between X_1 and X_2 is the same as the dependency between X_5 and X_6. In some sense, one is in an equilibrium, and this is called strict stationarity.

Before we define weak stationarity, let us define the mean function and the covariance function of a time series.

Definition 2.2. Let $X = (X_t)_{t\in T}$ be a time series with $\mathbb{E}X_t^2 < \infty$ for all $t \in T$. Then

$$\mu_X(t) := \mathbb{E}X_t, \quad t \in T,$$

is called the *mean function* and

$$\gamma_X(r, s) := \mathrm{Cov}\,(X_r, X_s) = \mathbb{E}[(X_r - \mu_X(r))(X_s - \mu_X(s))], \quad r, s \in T,$$

the *covariance function* of X.

The notion of weak stationarity is indeed weaker (at least if we have finite variance):

Definition 2.3. A real valued time series $(X_t)_{t\in T}$ is said to be *weakly stationary* or *second order stationary* or simply *stationary*, if

i) $\mathbb{E}X_t^2 < \infty$ for all $t \in T$,

ii) $\mathbb{E}X_t = \mathbb{E}X_{t'}$ for all $t, t' \in T$ (, i.e. μ_X is constant),

iii) $\gamma_X(t+h,t) = \gamma_X(t'+h,t')$ for all $t,t' \in T$, $h \in \mathbb{N}_0$, i.e. $\gamma_X(r,s)$ depends only on $r-s$.

Then $\mu_X := \mathbb{E}X_t$ is called the *mean of X*,

$$\gamma_X(h) := \gamma_X(t+h,t) \quad h \in \mathbb{N}_0 \text{ (if } T = \mathbb{N}_0, T = \mathbb{N}) \quad \text{or} \quad h \in \mathbb{Z} \quad \text{if} \quad T = \mathbb{Z},$$

is called *autocovariance of X at lag h*. The function/sequence $(\gamma_X(h))_{h\in\mathbb{N}_0/\mathbb{Z}}$ is called the *autocovariance function of X* (ACVF).
The *autocorrelation of X at lag h* is defined by (if $\gamma_X(0) \neq 0$)

$$\varrho_X(h) := \frac{\gamma_X(h)}{\gamma_X(0)} = \operatorname{Corr}(X_{t+h}, X_t),\ h \in \mathbb{N}_0/\mathbb{Z},$$

and $(\varrho_X(h))_{h\in\mathbb{N}_0,\mathbb{Z}}$ is called *autocorrelation function of X* (ACF).

So while strict stationarity is in terms of the finite dimensional distributions, weak stationarity is only defined in terms of the mean and covariances. We have:

Proposition 2.4. Let $(X_t)_{t\in T}$ be strictly stationary.

a) For all $t, t' \in T$ holds true that $X_t \overset{\mathrm{d}}{=} X_{t'}$.

b) If $\mathbb{E}X_t^2 < \infty$ for all $t \in T$ then $(X_t)_{t\in T}$ is weakly stationary.

Proof. a) clear, b) exercise/ clear. □

The easiest example of a strictly stationary sequence is i.i.d. noise:

Example 2.5. [IID-Noise]
If $X = (X_t)_{t\in T}$ is independent and identically distributed (i.i.d.), then X is called *i.i.d. noise* or *i.i.d. white noise.* Denote by ρ the distribution of X_1. By the i.i.d. property, the distribution of $(X_{t_1}, \ldots, X_{t_n})$ for $t_1 < \ldots < t_n$ is then given by $\rho^{\otimes n}$, the n'fold product measure of ρ with itself. From this we see immediately that $(X_t)_{t\in T}$ is strictly stationary. If $\mathbb{E}X_t = 0$ and $\mathbb{E}X_t^2 = \sigma^2 < \infty$, we write $(X_t)_{t\in T} \sim \operatorname{IID}(0,\sigma^2)$.

i.i.d. noise is the most basic building block for time series when thinking of strict stationarity. When thinking of weak stationarity, the most basic building block is white noise:

Example 2.6. [White Noise]
If $\mathbb{E}X_t^2 < \infty$, $\mathbb{E}X_t = 0, \operatorname{Var} X_t = \mathbb{E}X_t^2 = \sigma^2 \in (0,\infty)$ for all $t \in T$ and $\operatorname{Cov}(X_t, X_{t'}) = 0$ for all $t \neq t'$, then $(X_t)_{t\in T}$ is called *white noise*, written as

$$(X_t)_{t\in T} \sim WN(0,\sigma^2).$$

This time series is weakly stationary and γ_X is given by

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

There are also examples of time series that are not stationary:

Example 2.7. [Random Walk]
Let $(Z_t)_{t\in\mathbb{N}}$ be $IID(0,\sigma^2)$ with $\sigma > 0$ and define $S_t := \sum\limits_{i=1}^{t} Z_i$, $t \in \mathbb{N}_0$. It follows that $\mathbb{E}S_t = 0$, $\operatorname{Var} S_t = \operatorname{Var} \sum\limits_{i=1}^{t} Z_i = t\sigma^2 < \infty$,

$$\gamma_S(t+h,t) = \operatorname{Cov}(S_{t+h}, S_t) = \operatorname{Cov}(S_t + X_{t+1} + \cdots + X_{t+h}, S_t) = \operatorname{Cov}(S_t, S_t) = t\sigma^2,$$

so $(S_t)_{t\in\mathbb{N}_0}$ is not weakly stationary, hence neither strictly stationary.

The easiest example of a weakly stationary process built from white noise is the moving average process:

Example 2.8. [$MA(q)$-Process]
Let $(Z_t)_{t\in\mathbb{Z}} \sim WN(0,\sigma^2)$ and

$$X_t := Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q},\ t \in \mathbb{Z},$$

with $\theta_1, \ldots, \theta_q \in \mathbb{R}$.
$(X_t)_{t\in\mathbb{Z}}$ is called *$MA(q)$-process* or *moving-average process of order q*. It holds true that $\mathbb{E}X_t^2 < \infty$, $\mathbb{E}X_t = 0$, and with $\theta_0 := 1$,

$$\gamma_X(h) = \operatorname{Cov}(\sum_{i=0}^{q} \theta_i Z_{t+h-i}, \sum_{j=0}^{q} \theta_j Z_{t-j}) = \sum_{i,j=0}^{q} \theta_i\theta_j \operatorname{Cov}(Z_{t+h-i}, Z_{t-j}) = \sum_{j=0}^{q-|h|} \theta_j\theta_{|h|+j}\sigma^2,$$

(see this first for $h \geq 0$; for $h < 0$ use $\gamma_X(-h) = \gamma_X(h)$), so a $MA(q)$-process is weakly stationary).

Remark 2.9. *The converse of Proposition 2.4 a) does not hold true, i.e. there exists a weakly stationary time series which is not strictly stationary.*

Proof. Exercise. □

Let us recall the definition of the normal distribution:

Definition 2.10. Let $\sigma \geq 0$ and $\mu \in \mathbb{R}$. A probability measure ρ on $(\mathbb{R}, \mathcal{B}_1)$ is a *normal distribution with mean μ and variance σ^2*, if

- either $\sigma = 0$ and $\rho = \delta_\mu$, the Dirac measure at μ, or
- $\sigma \in (0, \infty)$ and ρ has probability density
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),$$
i.e.
$$\rho(A) = \int_A \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx$$
for all $A \in \mathcal{B}_1$.

We write $\rho = N(\mu, \sigma^2)$. A random variable $X : \Omega \to \mathbb{R}$ is called *normally distributed* or *Gaussian distributed* or a *Gaussian random variable*, if its distribution is $N(\mu, \sigma^2)$ for some $\sigma \in [0, \infty)$ and $\mu \in \mathbb{R}$. When $\mu = 0$ and $\sigma = 1$ we speak of $N(0, 1)$ as the *standard normal distribution.*

Remark 2.11. If $X \overset{d}{=} N(\mu, \sigma^2)$, then X has indeed expectation μ and variance σ^2. Also, if $\sigma^2 = 0$, then X is equal to μ almost surely.

Recall that the characteristic function $\varphi_X : \mathbb{R}^d \to \mathbb{C}$ of a random vector $X : \Omega \to \mathbb{R}^d$ is given by
$$\varphi_X(u) := \mathbb{E}e^{i\langle u,X\rangle} = \mathbb{E}e^{iu'X}, \quad u \in \mathbb{R}^d,$$
where $\langle \cdot, \cdot \rangle$ denotes the Euclidian product in $\mathbb{R}^d$. We write a' to denote the transpose of a vector or matrix a. Usually, vectors in $\mathbb{R}^d$ will be column vectors. The characteristic function uniquely describes the distribution of a random vector. Normal random variables can be characterised as follows:

Proposition 2.12. *Let $X : \Omega \to \mathbb{R}$ be a random variable and $\mu \in \mathbb{R}$, $\sigma \geq 0$. Then X is normally distributed with mean μ and variance σ^2 if and only if the characteristic function φ_X of X is given by*
$$\varphi_X(u) = e^{iu\mu - \sigma^2 u^2/2} \quad \forall\, u \in \mathbb{R}.$$

The proof can be found in any standard text on probability.

The definition of a multivariate normal random variable can be deduced from the one-dimensional setting.

Definition 2.13. A random vector $X = (X_1, \ldots, X_d)' : \Omega \to \mathbb{R}^d$ is *Gaussian* or *normally distributed*, if for each $a = (a_1, \ldots, a_d)' \in \mathbb{R}^d$ the linear combination $a'X = \sum_{j=1}^d a_j X_j$ is one-dimensional Gaussian distributed.

This may be a bit different from the definition you know of multivariate normal distributions. The next theorem gives the relation with the familiar concept.

Theorem 2.14. *(a) If a random vector $X = (X_1, \dots, X_d)'$ is Gaussian, then each of its components has finite variance and the distribution of X is uniquely determined by the mean $\mu := \mathbb{E}(X)$ and the covariance matrix*

$$\Sigma := \operatorname{Cov}(X) = (\operatorname{Cov}(X_i, X_j))_{i,j=1,\dots,d} \in \mathbb{R}^{d \times d}.$$

We write $X \sim N(\mu, \Sigma)$ or $X \overset{d}{=} N(\mu, \Sigma)$ in that case.
(b) To every $\mu \in \mathbb{R}^d$ and every symmetric positive semidefinite matrix $\Sigma \in \mathbb{R}^{d \times d}$ (i.e. $\Sigma' = \Sigma$ and $x'\Sigma x \geq 0$ for all $x \in \mathbb{R}^d$) there exists a random vector X with the distribution $N(\mu, \Sigma)$.
(c) A random vector $X : \Omega \to \mathbb{R}^d$ is $N(\mu, \Sigma)$-distributed (where $\mu \in \mathbb{R}$ and $\Sigma \in \mathbb{R}^{d \times d}$), if and only if the characteristic function of X is given by

$$\varphi_X(u) = \mathrm{e}^{\mathrm{i}\langle u, \mu \rangle - u'\Sigma u/2} = \mathrm{e}^{\mathrm{i}\mu' u - u'\Sigma u/2} \quad \forall\, u \in \mathbb{R}^d.$$

(d) If $X \sim N(\mu, \Sigma)$ and the matrix Σ is invertible, then X has a probability density which is given by

$$f_X(x) = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} \exp\left(-\frac{1}{2}(x - \mu)'\Sigma^{-1}(x - \mu) \right), \quad x \in \mathbb{R}^d.$$

(e) For a random vector X and $\mu \in \mathbb{R}^d$ and a positive semidefinite matrix $\Sigma \in \mathbb{R}^{d \times d}$ we have $Y \sim N(\mu, \Sigma)$ if and only if $a'Y \sim N(a'\mu, a'\Sigma a)$ for all $a \in \mathbb{R}^d$.

A proof can be found in the book by Brockwell and Davis [BD1], Section 1.6. Observe that the expectation of the vector X is defined componentwise, i.e. $\mathbb{E}(X_1, \dots, X_d)' = (\mathbb{E}X_1, \dots, \mathbb{E}X_d)'$.

Knowing the multivariate normal distribution, we can define what a Gaussian time series is.

Definition 2.15. A time series $(X_t)_{t \in \mathbb{Z}}$ is said to be a *Gaussian time series*, if $(X_{t_1}, \dots, X_{t_n})'$ is Gaussian for all $n \in \mathbb{N}$ and $t_1, \dots, t_n \in \mathbb{Z}$, i.e. all finite-dimensional distributions are normally distributed.

For Gaussian time series, the concepts of weak and strict stationarity coincide:

Proposition 2.16. *Let $X = (X_t)_{t \in \mathbb{Z}}$ be a Gaussian time series. Then the following statements are equivalent:*

(i) X is weakly stationary,

(ii) X is strictly stationary.

Proof. That "$(ii) \implies (i)$" follows from Proposition 2.4 (b), since $\mathbb{E}X_t^2 < \infty$ for all t. Let us prove that "$(i) \implies (ii)$". For all $n \in \mathbb{N}$ and $t_1, \dots, t_n \in \mathbb{Z}$ it holds true that

$$\begin{aligned}
(X_{t_1+h}, \dots, X_{t_n+h})' &\overset{d}{=} N((\mathbb{E}X_{t_1+h}, \dots, \mathbb{E}X_{t_n+h})', (\operatorname{Cov}(X_{t_i+h}, X_{t_j+h}))_{i,j=1,\dots,n}) \\
&\overset{(i)}{=} N((\mu, \dots, \mu)', (\gamma_X(t_i + h - t_j - h))_{i,j=1,\dots,n}) \\
&= N((\mu, \dots, \mu)', (\gamma_X(t_i - t_j))_{i,j=1,\dots,n}) \\
&\overset{(i)}{=} N((\mathbb{E}X_{t_1}, \dots, \mathbb{E}X_{t_n})', (\operatorname{Cov}(X_{t_i}, X_{t_j}))_{i,j=1,\dots,n}) \\
&\overset{d}{=} (X_{t_1}, \dots, X_{t_n})'.
\end{aligned}$$

□

Proposition 2.16 is the reason why one often considers only weakly stationary time series. The idea is that one often has (approximately) a Gaussian time series as they arise in reality by virtue of the central limit theorem, and for Gaussian time series then weak and strict stationarity are the same concept.

When we have empirical data, which we assume to be a realisation of a weakly stationary time series, we would like to estimate the mean and the autocovariance function. This is usually done with the following estimators:

Definition 2.17. *Let $x_1, \dots, x_n$ be a realisation of an $\mathbb{R}$-valued time series (regardless if is weakly stationary or not). Then the* empirical mean *of $x_1, \dots, x_n$ is defined by*

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

The empirical autocovariance function $\widehat{\gamma}$ *is defined by*

$$\widehat{\gamma}(h) := \frac{1}{n} \sum_{t=1}^{n-h} (x_{t+h} - \overline{x})(x_t - \overline{x}), \quad h = 0, 1, \dots, n-1 \tag{2.1}$$

and

$$\widehat{\gamma}(h) := \widehat{\gamma}(-h), \quad h = -n+1, \dots, -1. \tag{2.2}$$

Provided $\widehat{\gamma}(0) \neq 0$ (this is the case if $x_1, \dots, x_n$ are not all equal), the empirical autocorrelation function *is given by*

$$\widehat{\varrho}(h) := \frac{\widehat{\gamma}(h)}{\widehat{\gamma}(0)}, \quad -n < h < n. \tag{2.3}$$

In later chapters we shall discuss the asymptotic behaviour of the empirical mean and the empirical autocovariance function. In many cases of interest they are strongly consistent and asymptotically normal estimators for the corresponding mean and autocovariance function of a strictly and weakly stationary process.

Sometimes it is necessary to work with complex-valued random variables and time series. If $X = Y + \mathrm{i}Z$, where Y, Z are real-valued, then set $\overline{X} = Y - \mathrm{i}Z$ to be the complex conjugate of X.

We define the covariance of two complex random variables in such a way that it is linear in the first component. More precisely, we have:

Definition 2.18. *a) For two complex-valued random variables X, Y on the same probability space with $\mathbb{E}|X|^2 < \infty$, $\mathbb{E}|Y|^2 < \infty$, we define*

$$\mathrm{Cov}\,(X, Y) := \mathbb{E}(X\overline{Y}) - (\mathbb{E}X)(\mathbb{E}\overline{Y}) = \mathbb{E}\big[(X - \mathbb{E}(X))\overline{(Y - \mathbb{E}(Y))}\big],$$

which is called the covariance *of X and Y. When $X = Y$ we call* $\mathrm{Var}\,(X) = \mathrm{Cov}\,(X, X)$ *the* variance *of X. Observe that* $\mathrm{Var}\,(X) \in \mathbb{R}$ *and even* $\mathrm{Var}\,(X) \geq 0$, *while* $\mathrm{Cov}\,(X, Y) \in \mathbb{C}$.

b) A complex-valued time series $(X_t)_{t\in\mathbb{Z}}$ is said to be weakly stationary, *if* $\mathbb{E}X_t = \mathbb{E}X_{t'}$ *for all $t, t' \in \mathbb{Z}$, $\mathbb{E}|X_t|^2 < \infty$ for all $t \in \mathbb{Z}$ and*

$$\gamma_X(r, s) := \mathrm{Cov}\,(X_r, X_s)$$

depends only on the difference $r - s$. Then

$$\gamma_X(r - s) := \gamma_X(r, s)$$

is called the autocovariance function at lag $r - s$.
If additionally $\gamma_X(0) \neq 0$, then the autocorrelation function at lag h *is defined by*

$$\rho_X(h) := \frac{\gamma_X(h)}{\gamma_X(0)}.$$

c) A time series $(X_t)_{t\in\mathbb{Z}}$ is strictly stationary, *if all its finite dimensional distributions are shift-invariant, i.e. if*

$$(X_{t_1}, X_{t_2}, \ldots, X_{t_n}) \overset{d}{=} (X_{t_1+k}, X_{t_2+k}, \ldots, X_{t_n+k}),$$

for all $n \in \mathbb{N}$ and all $t_1, \ldots, t_n \in \mathbb{Z}$ and all $k \in \mathbb{Z}$.

Remark 2.19. (a) If $(X_t)_{t\in\mathbb{Z}}$ is real-valued and weakly stationary, then

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

For a complex-valued weakly stationary time series we have

$$\gamma_X(h) = \overline{\gamma_X(-h)}.$$

This follows immediately from the fact that $\mathrm{Cov}\,(Y, X) = \overline{\mathrm{Cov}\,(X, Y)}$.
(b) If $x_1, \ldots, x_n \in \mathbb{C}$ are realisations of a complex valued time series, then the empirical mean of $x_1, \ldots, x_n$ is defined by $\overline{x} := \frac{1}{n}\sum_{t=1}^n x_t$, the empirical autocovariance function by (2.1) for $h = 0, \ldots, n-1$ and by $\widehat{\gamma}(h) := \overline{\widehat{\gamma}(-h)}$ for $h = -n+1, \ldots, -1$, and the empirical autocorrelation function again by (2.3).

Convention 2.20. *When we speak of* stationary *time series in this lecture, we shall always mean weakly stationary time.*

Chapter 3

Data cleansing from trends and seasonal effects

In this chapter we present some methods of how to estimate trend and season or how to eliminate them. Before speaking of these quantities, one should have a model behind.

3.1 Decomposition of time series

We have seen that we should first identify trend and seasonality and other deterministic quantities. In order to speak about these, we should first have a model behind. This is done in the classical (additive) decomposition model:

Definition 3.1. The *classical (additive) decomposition model* describes a time series (x_t) as a sum

$$x_t = m_t + s_t + y_t,$$

where

- (m_t) denotes the *trend component*, which is a slowly changing function
- (s_t) denotes the *seasonal component*, which is a function with known period d
- (y_t) denotes a *random noise component*, which is often hoped to be a realisation of a stationary time series. The random noise component is responsible for the fluctuations of a time series.

Sometimes, the trend component is further decomposed into a *cyclic component* corresponding e.g. to an economic cycle, and a *pure trend component*, which might be increasing or decreasing. We shall however not do this but assume the classical additive decomposition model.

There are also many other models possible, one could e.g. also think of *multiplicative decomposition model* given as

$$x_t = m_t \cdot s_t \cdot y_t$$

When all components are positive, taking the logarithm leads to

$$\log x_t = \log m_t + \log s_t + \log y_t,$$

which is again additive.

We shall not touch on these other models, but as said, throughout we shall assume the classical additive model as underlying.

3.2 Estimation of the trend in the absence of seasonality

In this section we present some methods to estimate the trend when there is no seasonal component present. In the next section we shall then be concerned with the estimation of the seasonality when no trend is present, and in the last section with the estimation of both trend and season simultaneously.

Assumption 3.2. The model for the time series (x_t) is given by

$$x_t = m_t + y_t, \quad t = 1, \dots, T,$$

where $y_1, \dots, y_T$ is a "well-behaved" noise term, in particular it has expectation 0. The trend is denoted by (m_t) and is a "slowly changing" deterministic function.

We shall consider the following methods for estimating / eliminating the trend of a given time series:

(a) Linear regression

(b) Polynomial regression

(c) Moving average smoothing

(d) Exponential smoothing

(e) Differencing

(f) Transformation of the data

3.2.1 Linear regression

Most of you know linear regression from statistics courses. We assume here that an (affine-)linear trend is present. More precisely, the additional model assumption throughout Subsection 3.2.1 is

Assumption 3.3. The trend (m_t) is affine linear, i.e. of the form

$$m_t = \alpha t + \beta, \quad t = 1, \ldots, T,$$

for some $\alpha, \beta \in \mathbb{R}$.

Definition 3.4. Under the model Assumptions 3.2 and 3.3, the *linear regression* chooses those α and β in $\mathbb{R}$ which fit best in a least squares sense, i.e. that satisfy

$$(\alpha, \beta) := \text{argmin}_{(a,b)\in\mathbb{R}^2} \sum_{t=1}^{T} (x_t - at - b)^2.$$

These α and β can be explicitly given as done in the following theorem:

Theorem 3.5. *Given data $x_t = m_t + y_t$, $t = 1, \ldots, T$, as above (with $T \geq 2$), define the empirical mean*

$$\overline{x}_T = \frac{1}{T} \sum_{t=1}^{T} x_t$$

and the time average

$$\overline{t}_T := \frac{1}{T} \sum_{t=1}^{T} t = \frac{1}{T} \frac{T(T+1)}{2} = \frac{T+1}{2}.$$

Then the optimal solution (α, β) in the linear regression framework above is given by

$$\alpha := \frac{\sum_{t=1}^{T} (t - \overline{t}_T) x_t}{\sum_{t=1}^{T} (t - \overline{t}_T)^2}$$

and

$$\beta := \overline{x}_T - \alpha \overline{t}_T.$$

You should know that result from an introductory statistics course. I give the proof only for completeness.

Proof. Write

$$g(a, b) := \sum_{t=1}^{T} (x_t - at - b)^2, \quad a, b \in \mathbb{R}.$$

For minimizing over (a, b) we need to get the partial derivatives and set them equal to 0. So

$$0 = \frac{\partial g}{\partial a}(a, b) = 2 \sum_{t=1}^{T} (x_t - at - b) \cdot (-t), \tag{3.1}$$

$$0 = \frac{\partial g}{\partial b}(a, b) = -2 \sum_{t=1}^{T} (x_t - at - b). \tag{3.2}$$

Rearranging (3.2) leads to

$$\overline{x}_T = b + a\frac{1}{T}\sum_{t=1}^{T} t = b + a\overline{t}_T, \tag{3.3}$$

and multiplying this by $\overline{t}_T$ gives

$$\overline{x}_T \cdot \overline{t}_T = b\overline{t}_T + a(\overline{t}_T)^2. \tag{3.4}$$

Rearranging (3.1) we obtain

$$\frac{1}{T}\sum_{t=1}^{T} x_t t = b\overline{t}_T + a\frac{1}{T}\sum_{t=1}^{T} t^2,$$

and substracting (3.4) from that we obtain

$$\begin{aligned} a\left(\frac{1}{T}\sum_{t=1}^{T} t^2 - (\overline{t}_T)^2\right) &= \frac{1}{T}\sum_{t=1}^{T} x_t t - \overline{x}_T\overline{t}_T \\ &= \frac{1}{T}\sum_{t=1}^{T}(x_t t - x_t\overline{t}_T) = \frac{1}{T}\sum_{t=1}^{T}(t - \overline{t}_T)x_t. \end{aligned} \tag{3.5}$$

Observe that

$$\begin{aligned} \frac{1}{T}\sum_{t=1}^{T} t^2 - (\overline{t}_T)^2 &= \frac{1}{T}\sum_{t=1}^{T}(t - \overline{t}_T + \overline{t}_T)^2 - (\overline{t}_T)^2 \\ &= \frac{1}{T}\sum_{t=1}^{T}(t - \overline{t}_T)^2 + \frac{2}{T}\overline{t}_T \underbrace{\sum_{t=1}^{T}(t - \overline{t}_T)}_{=0} + \frac{1}{T}T(\overline{t}_T)^2 - (\overline{t}_T)^2 \\ &= \frac{1}{T}\sum_{t=1}^{T}(t - \overline{t}_T)^2. \end{aligned}$$

Plugging this into (3.4) we get

$$\alpha = \alpha_{\text{opt}} = \frac{\sum_{t=1}^{T}(t - \overline{t}_T)x_t}{\sum_{t=1}^{T}(t - \overline{t}_T)^2},$$

and plugging this again into (3.3) we get

$$\beta = \beta_{\text{opt}} = \overline{x}_T - \alpha\overline{t}_T.$$

Formally, we should still prove that there is indeed the global minimum (not only a local one, or not a maximum), but we leave that out. It is clear from the setting that a global minimum must exist, and there the partial derivatives must be 0. As we found only one point with partial derivatives being 0, this must be the global minimum. □

Example 3.6. Recall the Dow Jones Utility Index data from Example 1.5 and Figure 1.4. Applying the linear regression as described above gives for the trend

$$m_t = 100 + 0.23t.$$

Figure 3.1 shows the data together with the plotted trend. Figure 3.2 shows the residuals, i.e. $y_t = x_t - m_t$.
A simple model for the Dow Jones utility index could hence be

$$x_t = 100 + 0.23t + y_t,$$

and one could try to forecast $x_{90} = 100 + 0.23 \cdot 90 = 120.7$.
However, looking at the fit it does not really seem to be a good fit, there might still be room to find more structure in (y_t), or make a more general regression.

3.2.2 Polynomial or more general regression

Again we assume Assumption 3.2, but Assumption 3.3 is replaced by the following:

Assumption 3.7. (a) The trend (m_t) obeys a polynomial regression, i.e. there are $\beta_0, \beta_1, \dots, \beta_{p-1} \in \mathbb{R}$ such that

$$m_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \ldots + \beta_{p-1} t^{p-1}, \quad t = 1, \dots, T.$$

Or, more generally:
(b) The trend (m_t) obeys a regression of the form

$$m_t = \beta_1 f_1(t) + \beta_2 f_2(t) + \ldots + \beta_p f_p(t), \quad t = 1, \dots, T,$$

where $f_1 : \{1, 2, \dots, T\} \to \mathbb{R}, \dots, f_p : \{1, 2, \dots, T\} \to \mathbb{R}$ are given functions.

Definition 3.8. Under the model Assumptions 3.2 and 3.7 (a), the *polynomial regression* chooses the $\beta_0, \dots, \beta_{p-1}$ which fit best in a least squares sense, i.e.

$$(\beta_0, \dots, \beta_{p-1}) := \operatorname{argmin}_{(b_0, \dots, b_{p-1}) \in \mathbb{R}^p} \sum_{t=1}^{T} (x_t - b_0 - b_1 t - \ldots - b_{p-1} t^{p-1})^2.$$

Under the model Assumptions 3.2 and 3.7 (b), the *general regression* chooses the $\beta_1, \dots, \beta_p$ which fit best in a least squares sense, i.e.

$$(\beta_1, \dots, \beta_p) := \operatorname{argmin}_{(b_1, \dots, b_p) \in \mathbb{R}^p} \sum_{t=1}^{T} (x_t - b_1 f_1(t) - \ldots - b_p f_p(t))^2.$$

The polynomial regression is obviously a special case of the general regression by choosing $f_j(t) = t^{j-1}$. The linear regression is the special case of polynomial regression with $p = 2$.

The solution of the general regression problem is given in the next theorem. Recall that the rank of a matrix denotes the maximal number of linearly independent columns, equivalently the maximal number of linearly independent rows.

Theorem 3.9. *In the framework of general regression, suppose that $T \geq p$ and that* $\operatorname{rank}(A) = p$, *where*

$$A := \begin{pmatrix} f_1(1) & f_2(1) & \dots & f_p(1) \\ \vdots & \vdots & & \vdots \\ f_1(T) & f_2(T) & \dots & f_p(T) \end{pmatrix} \in \mathbb{R}^{T\times p}$$

is the design matrix*. Then $A'A \in \mathbb{R}^{p\times p}$ is invertible (here: A' denotes the transpose of A) and the optimal solution to the regression problem is given by*

$$\begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} = (A'A)^{-1}A' \begin{pmatrix} x_1 \\ \vdots \\ x_T \end{pmatrix}.$$

Again, the result should be known from a standard statistics course, but since I have it ready anyway I can also include it in these notes.

Proof. Write

$$g(b_1, \dots, b_p) = \sum_{t=1}^{T} (x_t - b_1 f_1(t) - \dots - b_p f_p(t))^2.$$

When the minimum is attained, the partial derivatives must be 0, so

$$0 = \frac{\partial g}{\partial b_j}(b_1, \dots, b_p) = 2\sum_{t=1}^{T} \Big(x_t - b_1 f_1(t) - \dots - b_p f_p(t)\Big)(-f_j(t)) \quad \forall\, j = 1, \dots, p.$$

This is equivalent to

$$(f_j(1), \dots, f_j(T)) \cdot \begin{pmatrix} f_1(1) & \dots & f_p(1) \\ \vdots & & \vdots \\ f_1(T) & \dots & f_p(T) \end{pmatrix} \cdot \begin{pmatrix} b_1 \\ \vdots \\ b_p \end{pmatrix} = (f_j(1), \dots, f_j(T)) \cdot \begin{pmatrix} x_1 \\ \vdots \\ x_T \end{pmatrix}$$

for all $j \in 1, \dots, p$; to see this, simply multiply the matrix equation out and multiply it by 2; then one gets the above equation.
The p equations above can again be rewritten as a single matrix equation of the form

$$\begin{pmatrix} f_1(1) & \dots & f_1(T) \\ \vdots & & \vdots \\ f_p(1) & \dots & f_p(T) \end{pmatrix} \begin{pmatrix} f_1(1) & \dots & f_p(1) \\ \vdots & & \vdots \\ f_1(T) & \dots & f_p(T) \end{pmatrix} \cdot \begin{pmatrix} b_1 \\ \vdots \\ b_p \end{pmatrix} = \begin{pmatrix} f_1(1) & \dots & f_1(T) \\ \vdots & & \vdots \\ f_p(1) & \dots & f_p(T) \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_T \end{pmatrix}.$$

With the definition of the design matrix this can be rewritten as

$$A'A \begin{pmatrix} b_1 \\ \vdots \\ b_p \end{pmatrix} = A' \begin{pmatrix} x_1 \\ \vdots \\ x_T \end{pmatrix}.$$

Hence, if $A'A$ is invertible, we get the desired form for the least squares estimator in the general regression. That we have indeed a global and unique minimum is checked by an elementary curve discussion.
So we only have to verify that $A'A$ is invertible. Suppose not. Then there is $c \in \mathbb{R}^p \setminus \{0\}$ such that $A'Ac = 0$. Multiplying from the left by c' we obtain (with $\langle \cdot, \cdot \rangle$ the standard Euclidian product and $|\cdot|$ the norm in $\mathbb{R}^T$)

$$0 = c'A'Ac = \langle Ac, Ac \rangle = |Ac|^2.$$

Hence we conclude that $Ac = 0$. But since $\text{rank}(A) = p \leq T$, it is known from linear algebra that A is injective. This is a contradiction to the existence of $c \in \mathbb{R}^p \setminus \{0\}$ with $Ac = 0$. Hence $A'A$ must be invertible. □

Applying this to polynomial regression we obtain:

Corollary 3.10. *In the framework of polynomial regression, suppose that $T \geq p$, and let*

$$A := \begin{pmatrix} 1 & 1 & \dots & 1 \\ 1 & 2 & \dots & 2^{p-1} \\ \vdots & \vdots & & \vdots \\ 1 & T & \dots & T^{p-1} \end{pmatrix}.$$

Then $A'A$ is invertible and the optimal solution to the polynomial regression problem is given by

$$\begin{pmatrix} \beta_0 \\ \vdots \\ \beta_{p-1} \end{pmatrix} = (A'A)^{-1}A' \begin{pmatrix} x_1 \\ \vdots \\ x_T \end{pmatrix}.$$

Proof. The submatrix

$$B := \begin{pmatrix} 1 & 1 & \dots & 1 \\ 1 & 2 & \dots & 2^{p-1} \\ \vdots & \vdots & & \vdots \\ 1 & p & \dots & p^{p-1} \end{pmatrix} \in \mathbb{R}^{p \times p}$$

is a Vandermond matrix and hence has determinant

$$\det B = \prod_{1 \leq i < j \leq p} (i - j) \neq 0;$$

if you do not know this fact from linear algebra, it can be found e.g. on Wikepedia. Hence $\text{rank}(B) = p$ and hence also $\text{rank}(A) = p$. □

Example 3.11. Figure 3.3 shows the development of the population of the US from 1790 to 1990, where a time step corresponds to 10 years (t=1 means 1790, t=2 means 1800, etc.) The data and graph are taken from ITSM and are displayed in green. A quadratic regression seems reasonable, and indeed gives a good fit:

$$m_t = 0.70 * 10^7 - 0.22 * 10^7 t + 0.65 * 10^6 t^2$$

The regression line is also plotted in Figure 3.3 and indeed looks good. Using this model as forecasts, we get
Forecast for 2000: $273.2 \cdot 10^6$: The true value was $285.1 * 10^6$.
Forecast for 2010: $298.95 \cdot 10^6$: The true value was $309.33 * 10^6$.
Observe that the relative error for 2000 is

$$\frac{285.1 - 273.2}{285.1} = 4.1\%,$$

which is not so bad for quite a crude estimate that does not take into account any other circumstances that happen in the world.
The residuals of the data (i.e. $x_t - m_t$) are plotted in Figure 3.4. There, one does not see any more a typical pattern. As a time series analyst one is happy then and can try to fit stochastic models to the residuals.

Remark 3.12. It is also possible to use other regression functions like

- exponential, so a fit of the form $a + be^{\lambda t}$ with unknown a, b (λ known or assumed)
- harmonic regression of the form

  $$a_0 + \sum_{j=1}^{k} (a_j \cos(\lambda_j t) + b_j \sin(\lambda_j t)),$$

  where $a_0, a_1, \dots, a_k, b_1, \dots, b_k$ are unknown parameters and $\lambda_1, \dots, \lambda_k$ are fixed frequences. This is mainly interesting for estimating seasonal effects which will be done later.
- Many other specifications of regression types are possible.

3.2.3 Moving average smoothing

The idea behind this method is that the trend behaves approximately linearly over an interval of a (small) length, and that the residuals sum approximately up to 0 over these small intervals. The latter is justified because we assume that the residuals have expectation zero, so an average of them should be approximately close to their expectation, i.e. zero. Of course, the larger the interval, the better the approximation to zero, hence one has to balance which length to take.

To make this precise, we have the following:

Method 3.13. Let $q \in \mathbb{N}_0$ and $(x_t)_{t=1,\dots,T}$ be a time series. Define

$$W_t := \frac{1}{2q+1} \sum_{j=-q}^{q} x_{t+j}.$$

So this is an average with the q neighbouring elements to the right and the q neighbouring elements to the left of x_t. Later this will be called a two-sided moving average process, which is why the method is called *moving average smoothing*. Using the assumption that m_t is approximately affine linear, i.e. of the form

$$m_t \approx a + ct,$$

and that

$$\frac{1}{2q+1}\sum_{j=-q}^{q} y_{t+j} \approx 0,$$

we obtain

$$W_t = \frac{1}{2q+1} \underbrace{\sum_{j=-q}^{q} m_{t+j}}_{\approx \sum_{j=-q}^{q}(m_t+cj)=(2q+1)m_t} + \underbrace{\frac{1}{2q+1}\sum_{j=-q}^{q} y_{t+j}}_{\approx 0} \approx m_t$$

Hence W_t seems to be a good estimate for the trend!
The formula above is not defined for $t \le q$ or $t \ge T-q$, but by formally setting $x_t := x_1$ for $t \le 0$ and $x_t := x_T$ for $t \ge T-q$ we can use the above formula for all $t \in \{1, \dots, T\}$.

Remark 3.14. One can also use other weights than uniform weights, e.g.

$$W_t = \frac{1}{\sum_{j=-q}^{q} a_j} \sum_{j=-q}^{q} a_j x_{t-j},$$

provided $(\sum_{j=-q}^{q} a_j)^{-1} \sum_{j=-q}^{q} a_j m_{t-j} \approx m_t$ [See exerc. for conditions for that] and $\sum_{j=-q}^{q} a_j y_{t-j} \approx 0$.
In particular, when $q = 7$, $a_j = a_{-j}$ for $j \in \{0, \dots, 7\}$ and

$$(a_0, a_1, \dots, a_7) := (74, 67, 46, 21, 3, -5, -6, -3)$$

we get the *Spencer 15-point moving average filter*, which lets trends of polynomial degree 3 pass without distortion, i.e.

$$\frac{1}{\sum_{j=-7}^{7} a_j} \sum_{j=-7}^{7} a_j m_{t-q} = m_t$$

as long as m_t is a polynomial of degree ≤ 3 (Exercise).
One speaks of the numbers $\left(\frac{1}{\sum_{k=-q}^{q} a_k} a_j\right)_{j=-q,\dots,q}$ as *low pass filters*, since they remove the rapid fluctuations (or high frequency) from the data (since $\sum_{j=-q}^{q} y_{t-q} \approx 0$) and leave the slowly changing trend estimate W_t.

Example 3.15. Applying the moving average fit of the Dow Jones Utility Index (cf. Example 1.5, Figure 1.4) when choosing $q = 5$ and uniformly distributed weights gives the graph displayed in Figure 3.5. Also this seems to be a reasonable fit.

3.2.4 Exponential smoothing

The idea behind exponential smoothing is that the trend is a convex combination of the current observation and the last (estimated) trend. Recall that a convex combination of two numbers u, v is given by $\alpha u + (1-\alpha)v$ for some $\alpha \in [0,1]$. Letting α vary through $[0,1]$ one gets all numbers between u and v. But here we fix α. The precise method is:

Method 3.16. Let $\alpha \in [0,1]$ and $(x_t)_{t=1,\dots,T}$ be a time series. Define recursively

$$m_1 := x_1, \quad m_t := \alpha x_t + (1-\alpha)m_{t-1}.$$

(So we use the same α for all recursions).
For $t \geq 2$ this gives

$$m_t = \sum_{j=0}^{t-2} \alpha(1-\alpha)^j x_{t-j} + (1-\alpha)^{t-1} x_t,$$

which is a weighted moving average of $x_t, \dots, x_1$ with exponentially decreasing weights (except for the last one). This explains the name *exponential smoothing*.

3.2.5 Differencing

Rather than estimating a trend one can try to *eliminate* it. This can be done by differencing. Let

$$\nabla x_t := x_t - x_{t-1}.$$

(∇ is spoken "nabla". In analysis, often the gradient is denoted by ∇, and since $x_t - x_{t-1} = \frac{x_t - x_{t-1}}{t-(t-1)}$ is a difference quotient, we have the analogy to the derivative).
This gives a new series $z_t = \nabla x_t$, $t = 2, \dots, T$, and we can try to find a model for (z_t). If we have a model for (z_t), we also get one for (x_t) by setting

$$x_t = x_1 + \sum_{j=1}^{t} z_j.$$

The differencing operator eliminates linear trends to constant trends, and constant trends to no trends. More precisely, if $m_t = b + at$, then $\nabla m_t = a$. The proof is simple.

Lemma 3.17. *Suppose $m_t = b_0 + b_1 t + \ldots + b_{p-1} t^{q-1}$ for $t \in \{1, \dots, T\}$. Then ∇m_t is a polynomial of degree $\leq q-2$. Hence, applying the ∇-operator q times, the trend will be eliminated.*

Proof. Exercise □

Definition 3.18. Write

$$\begin{aligned} \nabla^2 x_t &= \nabla\nabla x_t = \nabla(x_t - x_{t-1}) \\ &= (x_t - x_{t-1}) - (x_{t-1} - x_{t-2}) = x_t - 2x_{t-1} + x_{t-2}, \\ \nabla^3 x_t &= \nabla\nabla\nabla x_t, \\ \dots \end{aligned}$$

Example 3.19. The Dow Jones Utility Index (x_t) of Example 1.5 can be seen in Figure 1.4. The corresponding differenced sequence (∇x_t) can be seen in Figure 3.6. One immediately sees that deterministic components in there are largely eliminated, although there is still a small pattern visible.

3.2.6 Transformation of the data

Sometimes it is helpful to apply a deterministic function to the data and then find a model for the transformed data. I.e. let $f : \mathbb{R} \to \mathbb{R}$ be a function and consider $v_t := f(x_t)$, $t = 1, \ldots, T$, and then try to estimate the mean (and season) of (v_t) rather than that of (x_t).

This is often applied using the logarithmic transform, i.e. $v_t = \log x_t$, and then often combined with differencing (v_t), giving the log returns $v_t - v_{t-1} = \log x_t - \log x_{t-1} = \log(x_t/x_{t-1})$. We have already come across the log-returns in Example 1.6.

Figure 3.1: The linear regression method applied to the Dow Jones Utility Index as described in Example 3.6. We see the index data together with the regression line

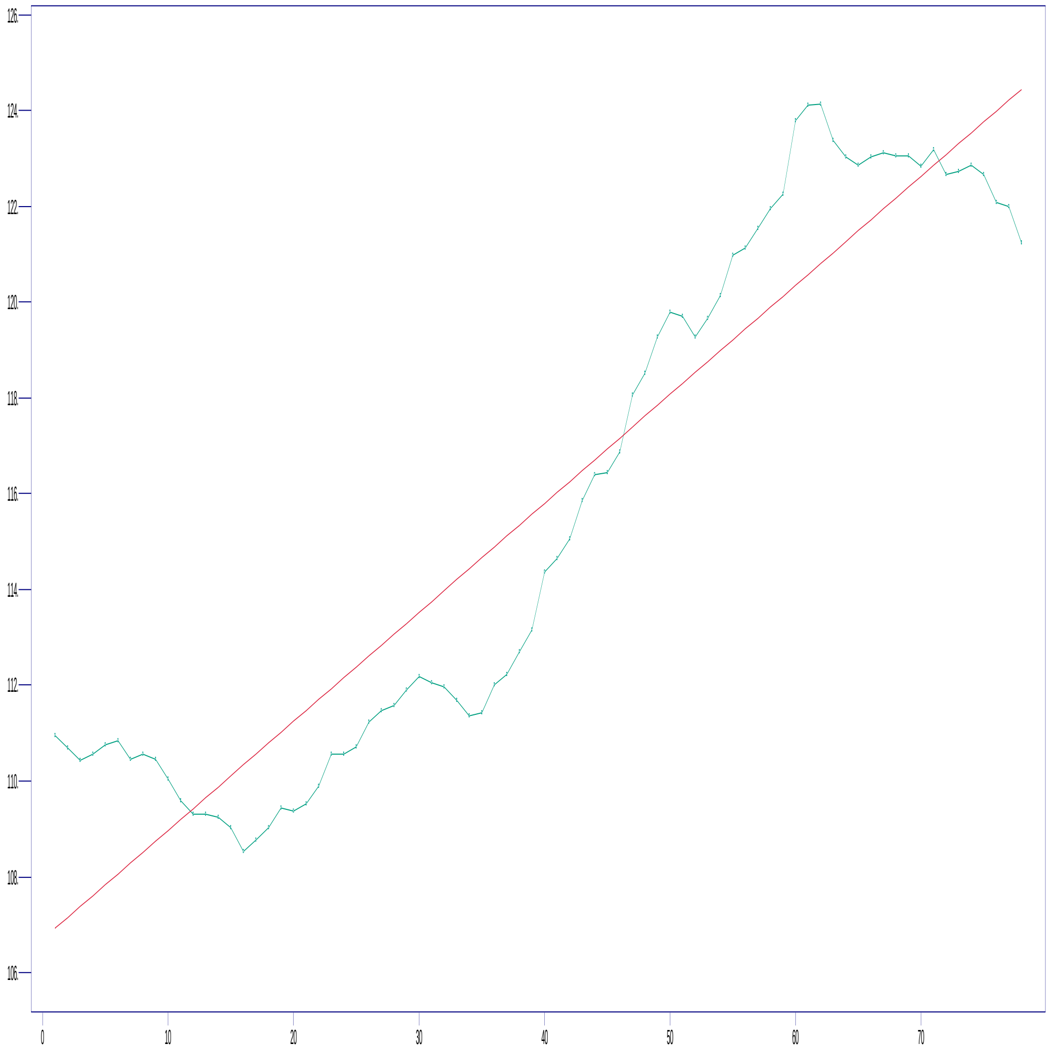


Figure 3.2: The residuals of the linear regression method applied to the Dow Jones Utility Index, arising from Figure 3.1, cf. Example 3.6

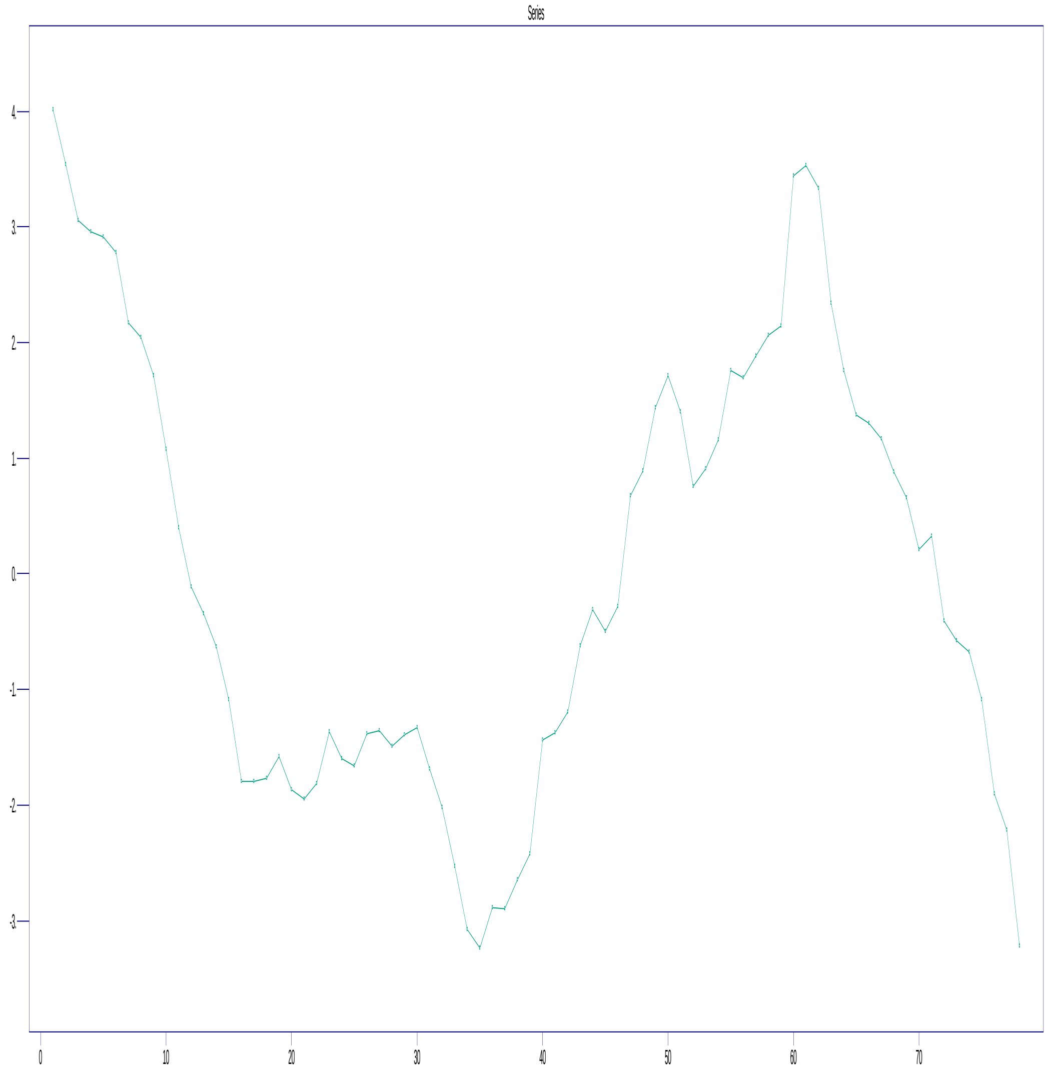


Figure 3.3: A quadratic regression applied to the US population data from Example 3.11. In green we have the data, in red the quadratic regression fit.

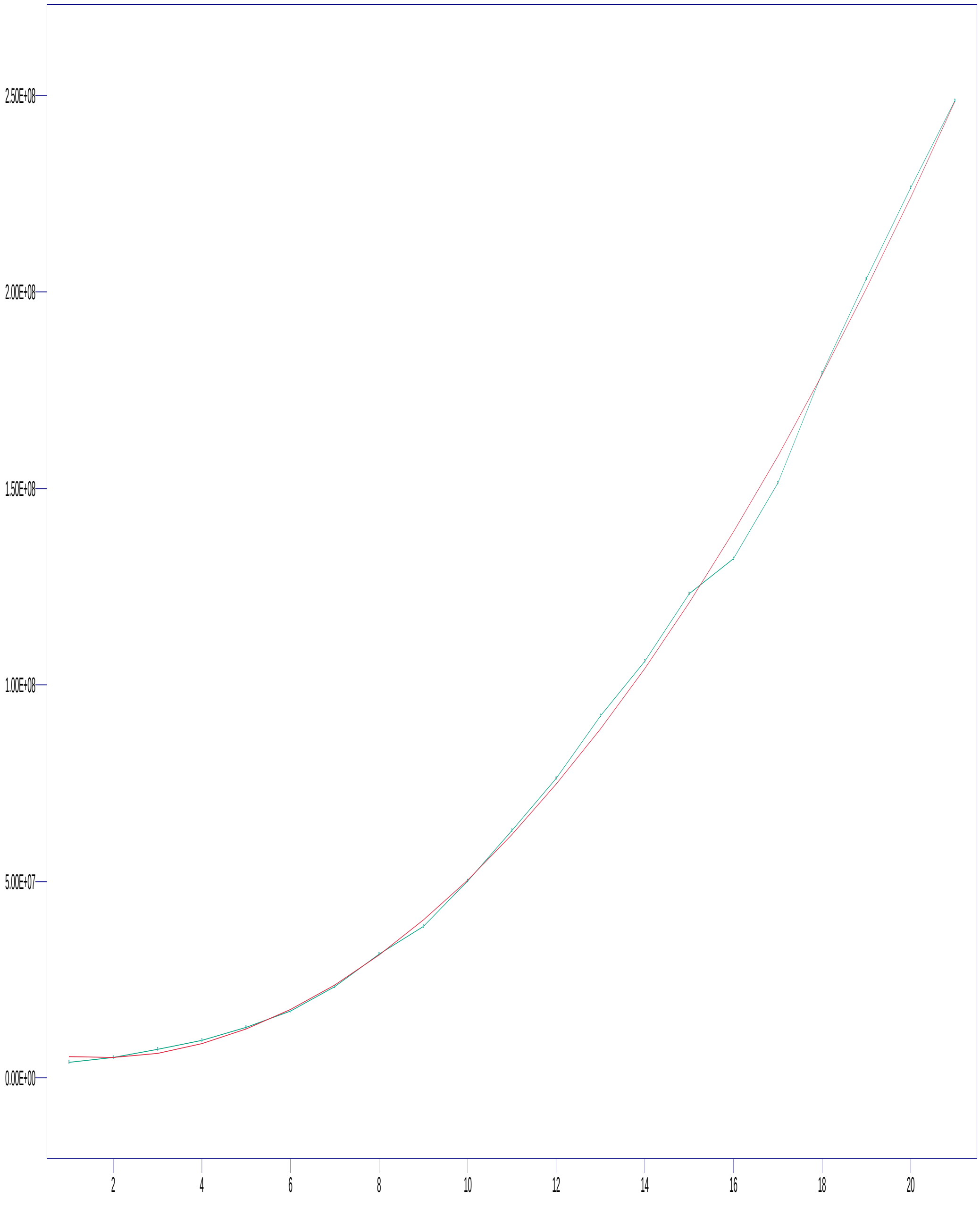


Figure 3.4: The residuals after the quadratic regression of Example 3.11 (cf. also Figure 3.3) have been done.

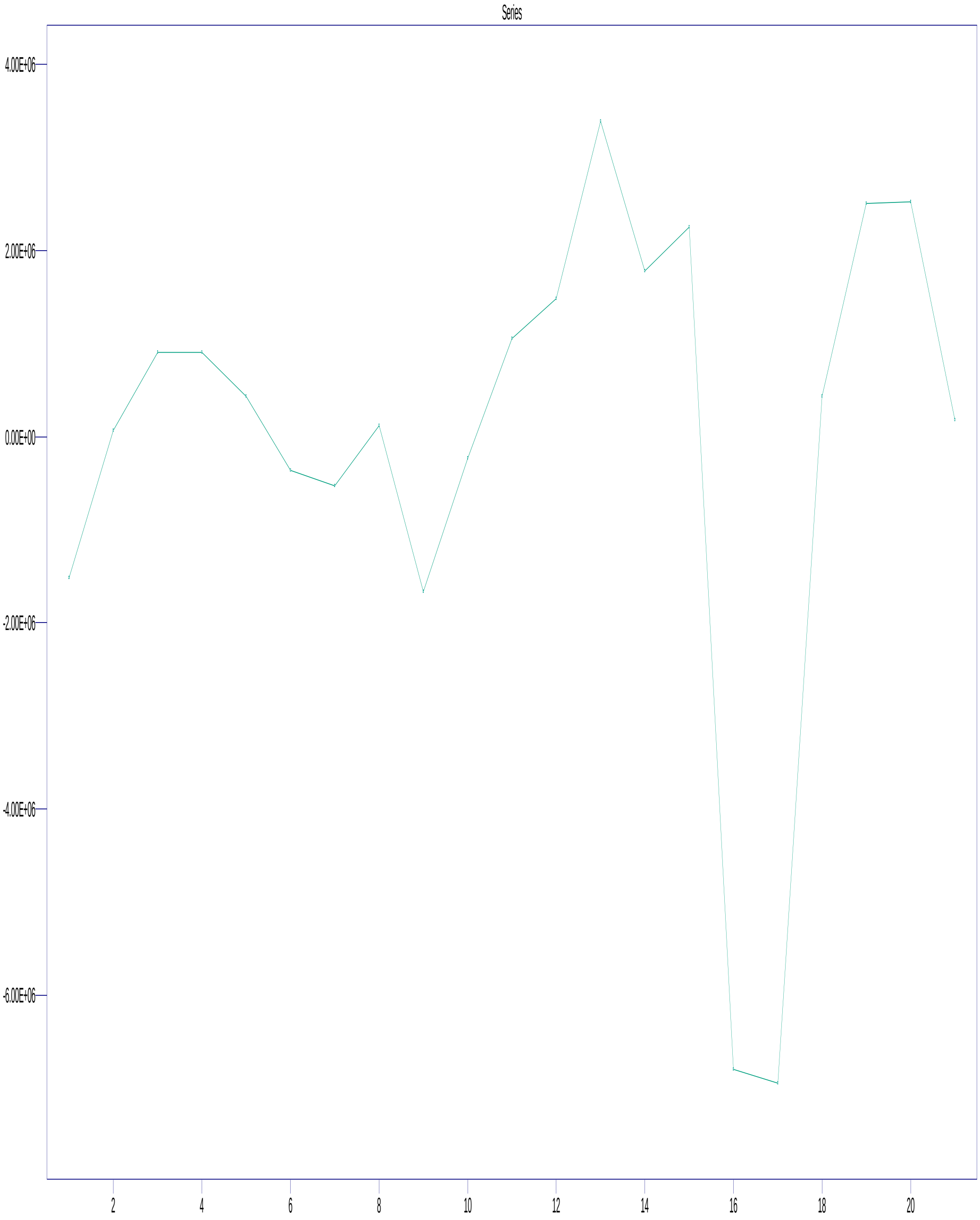


Figure 3.5: The moving average fit to the Dow Jones Utility Index (cf. Example 1.5, Figure 1.4) when choosing $q = 5$ and uniformly distributed weights. The data points are in green, the fitted line in red.

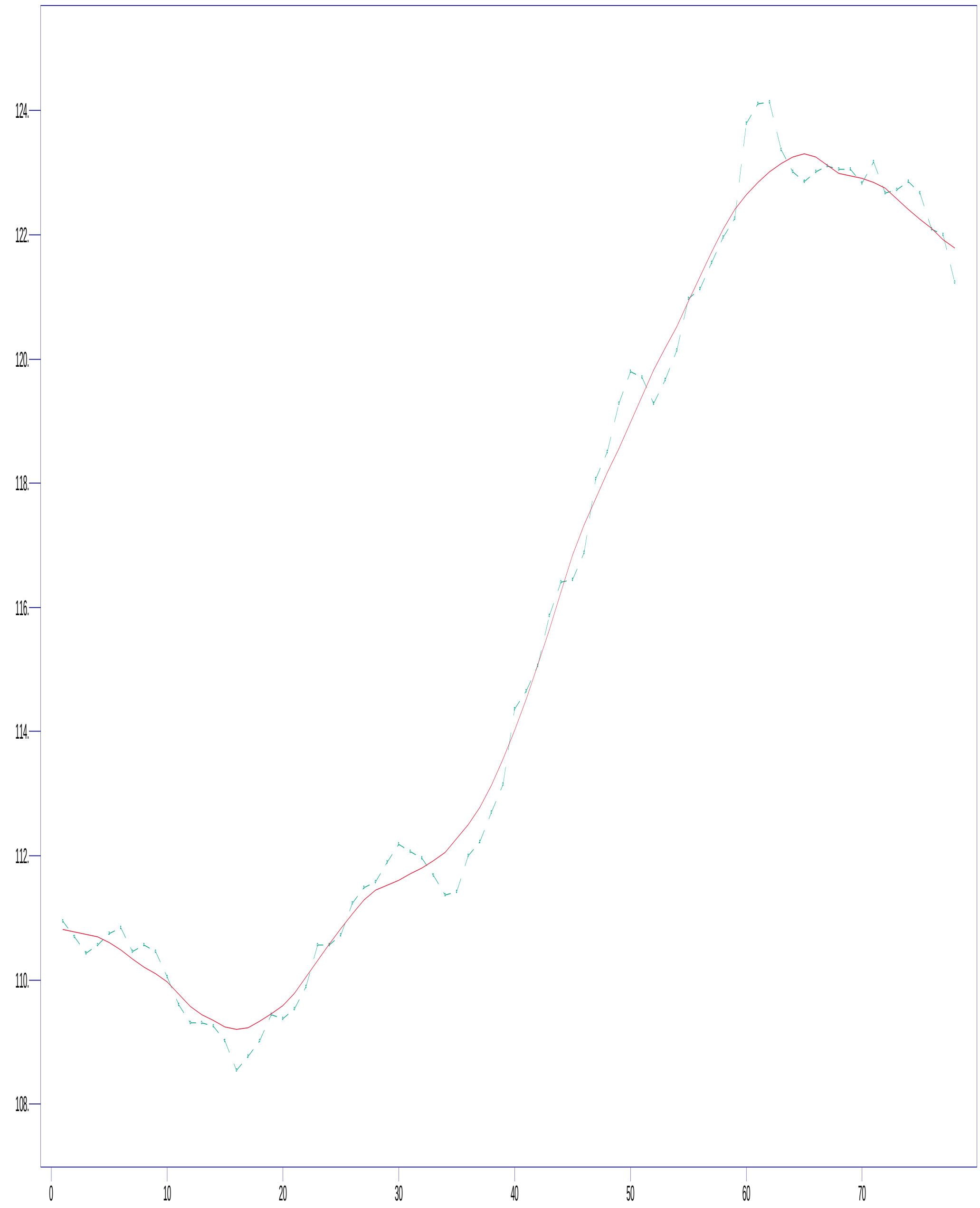
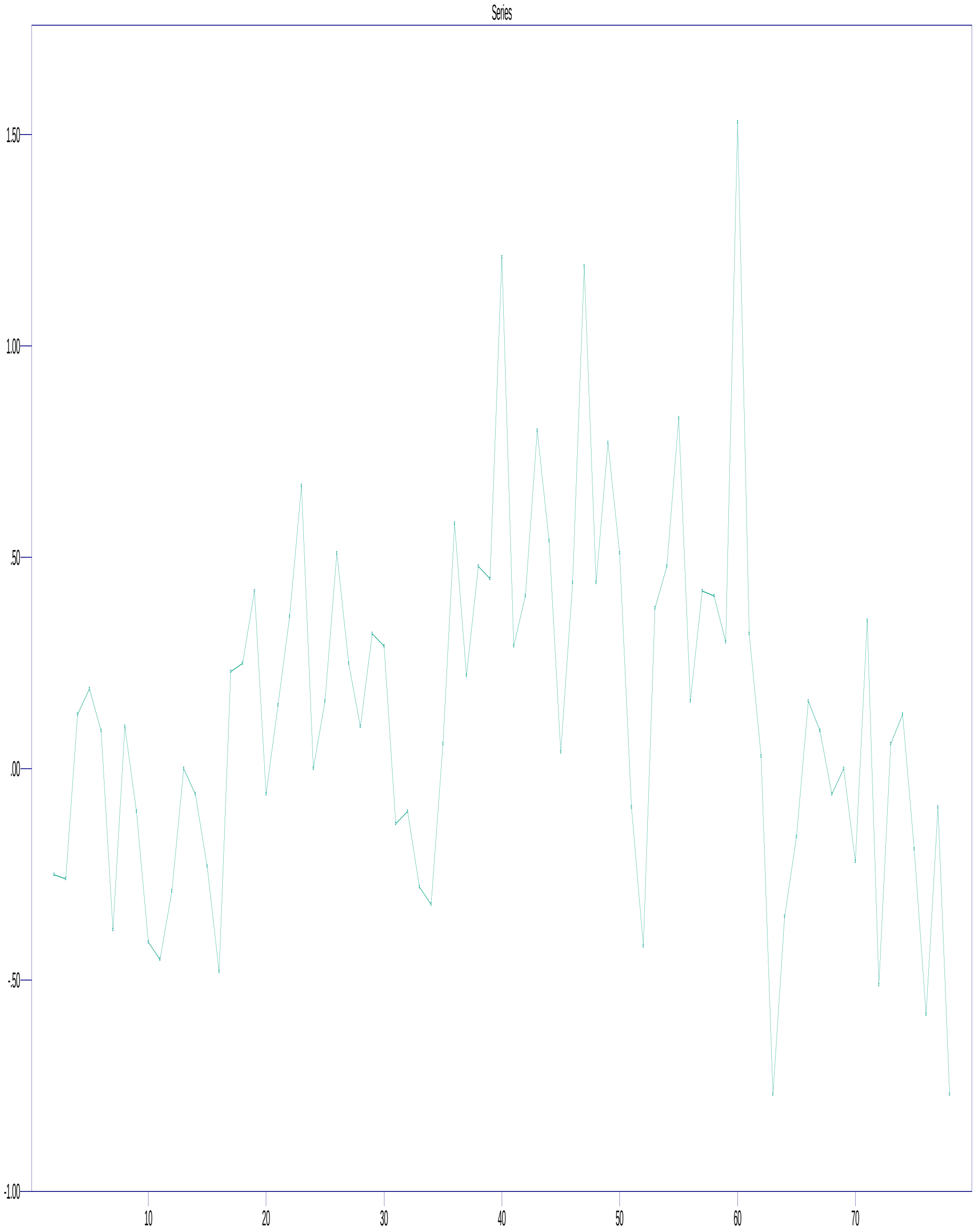


Figure 3.6: The (once) differenced sequence obtained from the Dow Jones Utility Index of Example 1.5 that was shown in Figure 1.4.



3.3 Estimating and elimination of seasonality in the absence of trend

In this section we consider the problem of the estimation of a seasonal component when no trend component is present. The precise model assumptions (replacing Assumption 3.2) for this section are:

Assumption 3.20. The model for the time series (x_t) is given by

$$x_t = s_t + y_t, \quad t = 1, \dots, T$$

where $y_1, \dots, y_T$ is a well behaved noise term, and s_t a *seasonal component*. By the latter we mean in particular that (s_t) is *d-periodic*, i.e.

$$s_{t+d} = s_t.$$

Further, assume that

$$\sum_{t=1}^{d} s_t = 0.$$

The last assumption is justified because we could also consider

$$s_t' := s_t - \frac{1}{d}\sum_{j=1}^{d} s_j$$

which is d-periodic and sums up to 0. The constant term $\sum_{j=1}^{d} s_t$ could be considered as a trend and packed there (but here, trend assumed to be absent). However, when considering trend and season simultaneously, we simply pack this constant to the trend.

To estimate the season, we can

- take the average of $x_t, x_{t+d}, x_{t+2d}, \dots$,
- or do harmonic regression,
- or eliminate the season by differencing.

3.3.1 Average method

Basically, one takes the average of $x_t, x_{t+d}, \dots$ and then substracts their own sum. More precisely, one has (recall that $\lfloor x \rfloor$ for $x \in \mathbb{R}$ denotes the largest number $n \in \mathbb{Z}$ such that $n \leq x$; so $\lfloor x \rfloor$ denotes the *floor function* or also called *Gauß bracket*):

Method 3.21. Given a time series $x_1, \dots, x_T$ and a fixed period d (where $d << T$), define

$$W_k := \frac{1}{\lfloor T/d \rfloor} \sum_{j=0}^{\lfloor T/d-1 \rfloor} x_{k+jd}, \quad k = 1, \dots, d.$$

This is a good candidate for the seasonal component, but does not necessarily sum up to 0. Hence denote

$$s_k := W_k - \frac{1}{d} \sum_{j=1}^{d} W_j, \quad k = 1, \dots, d$$

and

$$s_{k+jd} := s_k, \quad k = 1, \dots, d.$$

This is a good estimator for the seasonal component (it differs from W_k only by a constant) but is done such that it sums up to 0.

Definition 3.22. For an estimator $\widehat{s}_t$ of the seasonal component s_t, $d_t := x_t - \widehat{s}_t (\approx y_t)$ are called the *deseasonalised data.*

Usually, the s_t defined in Method 3.21 is just an estimator of the season, as one does not really know the true seasonal component. After all, all these are just models for the reality.

Example 3.23. The average method applied to the Dubuque temperature data of Example 1.3 (plotted in Figure 1.2), when the mean 46.26597 was substracted, gives

$$\begin{aligned} &s_1 = -29.66, s_2 = -25.62, s_3 = -13.79, s_4 = 0.26, \\ &s_5 = 11.82, s_6 = 21.23, s_7 = 25.45, s_8 = 23.07, \\ &s_9 = 14.76, s_{10} = 4.71, s_{11} = -9.62, s_{12} = -22.62. \end{aligned}$$

Figure 3.7 shows that Dubuque data (line) together with the estimated seasonal components plus mean (dots). Figure 3.8 shows the residuals after subtracting mean and seasonal component (the deseasonalised data when also the mean was subtracted).

3.3.2 Harmonic regression

This method is motivated by the fact that the seasonal component is supposed to be d-periodic and to fluctuate around 0.
Known functions which do that are $t \mapsto \sin(t\frac{j2\pi}{d})$ and $t \mapsto \cos(t\frac{j2\pi}{d})$ for $j \in \mathbb{N}$ (obviously

Figure 3.7: The Dubuque temperature data of Example 1.3 together with the estimated season (plus the mean of the data) of Example 3.23.

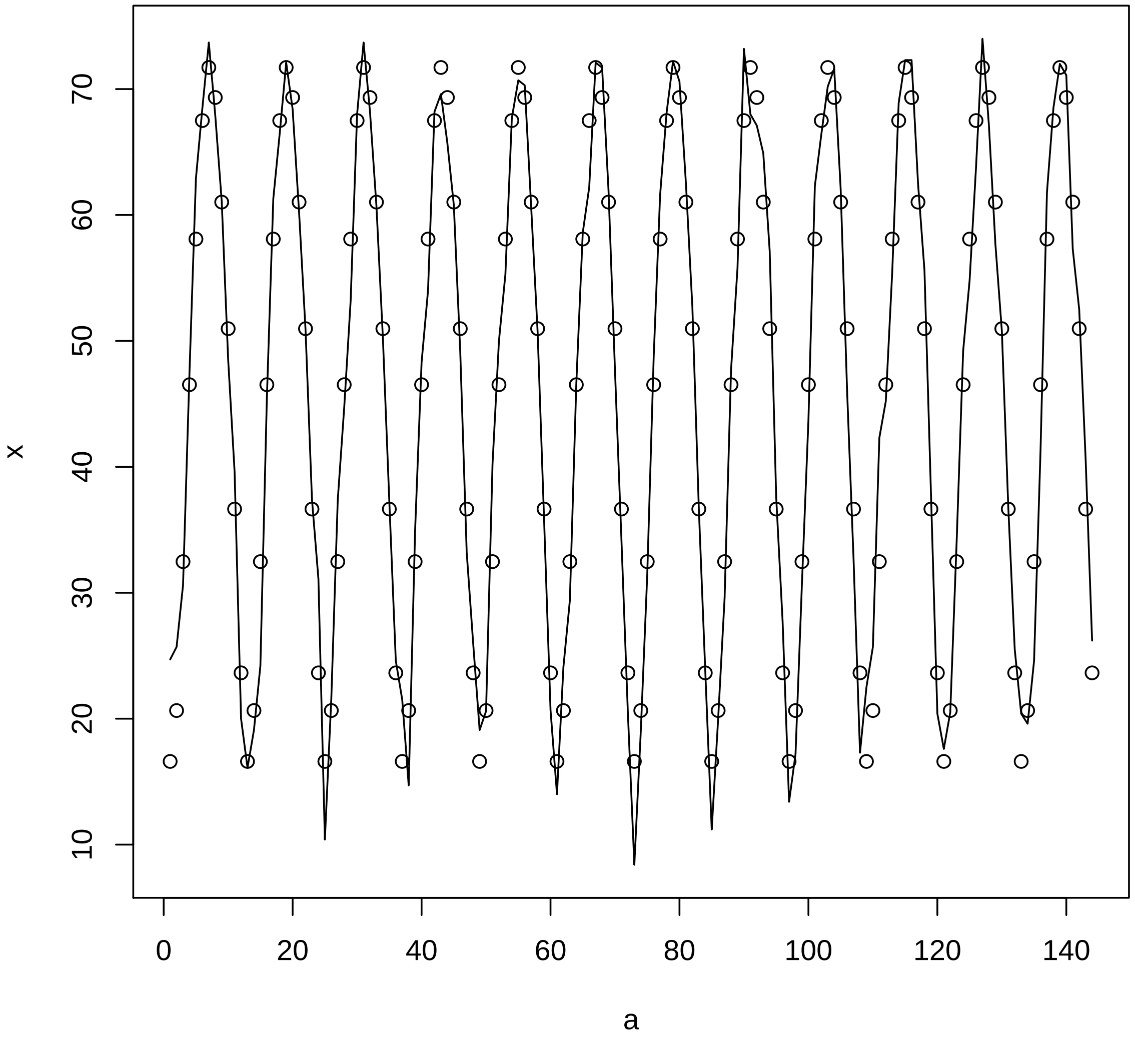
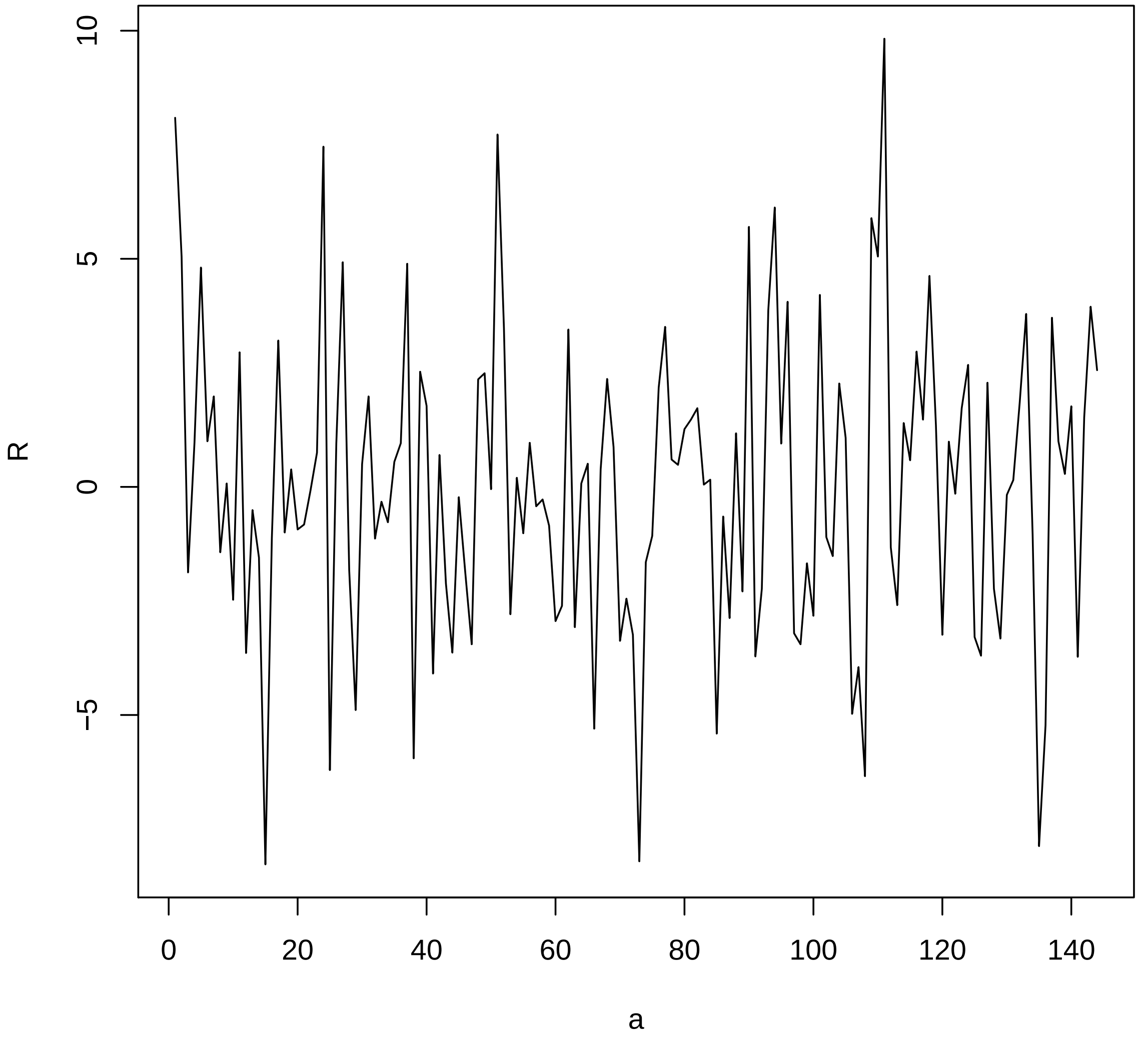


Figure 3.8: The residuals of the Dubuque temperature data after the mean and the seaonal component were subtracted, cf. Example 3.23.



not the only ones, but at least we know some functions that do that).
Hence can try to make a harmonic regression of the form

$$x_t = \sum_{j=1}^{k} \left(a_j \cos(t\frac{2\pi j}{d}) + b_j \sin(t\frac{2\pi j}{d}) \right) + y_t$$

for given k and take the regression function as value for the seasonal component. This is a particular case of a general regression (cf. Section 3.2.2).

3.3.3 Season elimination by differencing of higher order

As for the trend, rather than estimating the season we can also try to eliminate it by differencing with respect to lag d. By that we mean:

Definition 3.24. For a time series $(x_t)_{t=1,\dots,T}$ and some $d \in \mathbb{N}$ ($d << T$) define

$$\nabla_d x_t := x_t - x_{t-d},$$

which is the *difference operator of lag d*.

Remark 3.25. (a) In general, $\nabla_d x_t \neq \nabla^d x_t$ (when $d \geq 2$). This should be done as an exercise.
(b) Since

$$\nabla_d s_t = s_t - s_{t-d} = 0$$

by periodicity of s, the application of ∇_d eliminates the seasonal component from the data.

The air passenger data of Example 1.3, after having differenced them with lag 12, is shown in Figure 3.9.

3.4 Estimating trend and season simultaneously

In this section we assume that we have both trend and season simultaneously, i.e. a model of the form

$$x_t = m_t + s_t + y_t$$

as in Definition 3.1. As before, we assume that s is d-periodic and that $\sum_{t=1}^{d} s_t = 0$. We use the following method.

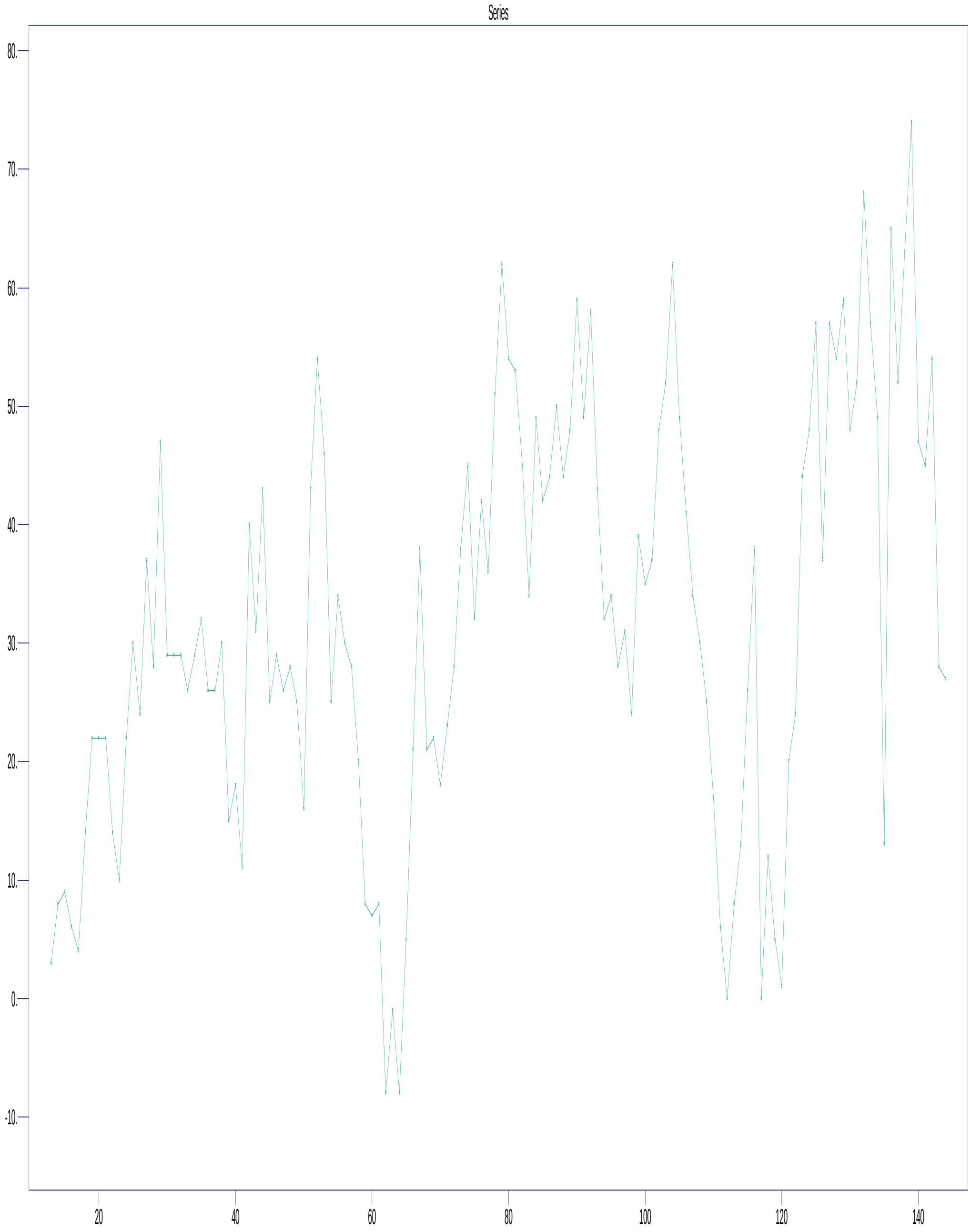
Method 3.26.

- Assume the model

$$x_t = m_t + s_t + y_t,$$

where m_t is the trend and s_t the seasonal component with period d, $s_1 + \ldots + s_d = 0$

Figure 3.9: ∇_{12} applied to the monthly total airline passenger data from Example 1.3, the undifferenced data which were displayed in Figure 1.2.



- We first do a preliminary estimation of the trend: If $d = 2q + 1$ with $q \in \mathbb{N}_0$, take

  $$\widehat{m}_{t,prelim} := \frac{1}{2q+1} \sum_{j=-q}^{q} x_{t+j}.$$

  If $d = 2q$ with $q \in \mathbb{N}$, take

  $$\widehat{m}_{t,prelim} := \frac{1}{2q} \left(\frac{1}{2} x_{t-q} + x_{t-q+1} + \ldots + x_{t+q-1} + \frac{1}{2} x_{t+q} \right).$$

- For estimating the seasonal component, denote

  $$W_k := \frac{1}{\lfloor T/d \rfloor} \sum_{j=0}^{\lfloor T/d \rfloor - 1} (x_{k+jd} - \widehat{m}_{k+jd,prelim}), \quad k = 1, \ldots, d.$$

- Since not necessarily $\sum_{k=1}^{d} W_k = 0$, denote

  $$\widehat{s}_k := W_k - \frac{1}{d} \sum_{j=1}^{d} W_j, \quad k = 1, \ldots, d,$$

  and

  $$\widehat{s}_{k+jd} := \widehat{s}_k$$

  Then $\widehat{s}_t$ is the seasonal component

- Denote

  $$d_t := x_t - \widehat{s}_t, \quad t = 1, \ldots, T,$$

  which are the deseasonalised data.

- Estimate the trend $\widehat{m}_t$ of (d_t) through the previous methods and get residuals $y_t = x_t - \widehat{s}_t - \widehat{m}_t$.

Observe that the preliminary trend $\widehat{m}_{t,prelim}$ is the average value *within one complete season.* By subtracting it, one basically makes the different cycles "similar", then one can apply the previous averaging method to estimate the season. After having deseasonalised, one can get the true trend.

Chapter 4

Properties of the autocovariance function

In this chapter we have a deeper look at properties of the autocovariance function of stationary time series. Can we even characterise when a function $\gamma : \mathbb{Z} \to \mathbb{R}$ is the autocovariance function of some stationary time series?

The answer is yes. For the proof of the corresponding theorem, we first recall the consistency theorem of Kolmogorov, which is usually proved in a course on Stochastic Processes. Alternatively, you can find a proof in the book of P. Billingsley, Probability and Measure, Section 36.

Theorem 4.1. [Consistency theorem of Kolmogorov]
Let $T \subset \mathbb{R}$, $d \in \mathbb{N}$ and for all $\vec{t} = (t_1, \dots, t_n)'$ with $t_1 < \cdots < t_n$, $t_i \in T$, let $\nu_{\vec{t}}$ be a distribution on $(\mathbb{R}^d)^n$. If the consistency condition

$$\begin{aligned} &\nu_{(t_1,\dots,t_n)}(B_1 \times \cdots \times B_{i-1} \times \mathbb{R}^d \times B_{i+1} \times \cdots \times B_n) \\ =&\nu_{(t_1,\dots,t_{i-1},t_{i+1},\dots,t_n)}(B_1 \times \cdots \times B_{i-1} \times B_{i+1} \times \cdots \times B_n) \end{aligned} \tag{4.1}$$

holds for all $\vec{t} = (t_1, \dots, t_n)'$ and $B_1, \dots, B_n \in \mathcal{B}(\mathbb{R}^d)$, then there exists a stochastic process $(X_t)_{t \in T}$ such that

$$P_{(X_{t_1},\dots,X_{t_n})} = \nu_{(t_1,\dots,t_n)} \ \forall t_1 < \cdots < t_n,$$

i.e. whose system of finite dimensional distributions is given by the $\nu_{\vec{t}}$.

The characterisation we will obtain below in Theorem 4.4 will be that a function $\kappa : \mathbb{Z} \to \mathbb{R}$ is the autocovariance function of a stationary real valued time series if and only if it is even and positive semidefinite. Let us first give the corresponding definitions:

Definition 4.2. *Let $\kappa : \mathbb{Z} \to \mathbb{R}$ be a function.*

a) κ is said to be even, *if $\kappa(h) = \kappa(-h)$ for all $h \in \mathbb{Z}$.*

b) κ is called positive semidefinite*, if*

$$\sum_{i,j=1}^{n} a_i \kappa(t_i - t_j) a_j \geq 0 \; \forall n \in \mathbb{N}, \; \forall a_1, \dots, a_n \in \mathbb{R}, \; \forall t_1, \dots, t_n \in \mathbb{Z}. \tag{4.2}$$

The notion of positive semidefiniteness appears strange at first sight, bu the following remark explains it better:

Remark 4.3. Denoting

$$\kappa_{\vec{t}} := (\kappa(t_i - t_j))_{i,j=1,\dots,n} \in \mathbb{R}^{n \times n}$$

for $\vec{t} = (t_1, \dots, t_n)' \in \mathbb{Z}^n$, then (4.2) is equivalent to

$$a' \kappa_{\vec{t}} a \geq 0 \quad \forall \; n \in \mathbb{N}, \; a \in \mathbb{R}^n, \; \vec{t} \in \mathbb{Z}^n.$$

Hence κ is even and positive semidefinite if and only if $\kappa_{\vec{t}}$ is positive semidefinite for all $n \in \mathbb{N}$ and all $\vec{t} \in \mathbb{Z}^n$ (speaking of positive semidefiniteness of the matrix $\kappa_{\vec{t}}$ needs symmetry (or at least being hermitian) of the matrix).

We can now achieve the desired characterisation:

Theorem 4.4. [Characterisation of an ACVF of a weakly stationary time series]
Let $\gamma : \mathbb{Z} \to \mathbb{R}$ be a function. Then the following statements are equivalent:

i) γ is an ACVF of a real-valued weakly stationary time series,

ii) γ is an ACVF of a weakly stationary Gaussian time series,

iii) γ is even and positive semidefinite.

Proof.
"*ii)* $\Longrightarrow$ *i)*" clear.

"*i)* $\Longrightarrow$ *iii)*" Let $X = (X_t)_{t \in \mathbb{Z}}$ be weakly stationary with ACVF γ. Then

$$\gamma(-h) = \operatorname{Cov}(X_{-h}, X_0) = \operatorname{Cov}(X_0, X_h) = \gamma(h), \quad h \in \mathbb{Z},$$

so γ is even.
Let $\vec{t} = (t_1, \dots, t_n)^T \in \mathbb{Z}^n$ and $\Gamma_{\vec{t}} = (\gamma(t_i - t_j))_{i,j=1,\dots,n}$, so $\Gamma_{\vec{t}} = \operatorname{Cov}((X_{t_1}, \dots, X_{t_n})')$. It follows that $\Gamma_{\vec{t}}$ is positive semidefinite, since every covariance matrix is positive semidefinite. Hence γ is positive semidefinite by Remark 4.3.

"*iii)* $\Longrightarrow$ *ii)*" Let γ be even and positive semidefinite. For every vector $\vec{t} = (t_1, \dots, t_n)' \in \mathbb{Z}^n$ with $t_1 < \cdots < t_n$ let $\nu_{\vec{t}}$ be a normal distribution with mean $\vec{0}$ and covariance matrix $\Gamma_{\vec{t}} = (\gamma(t_i - t_j))_{i,j=1,\dots,n}$. This distribution exists by Theorem 2.14 (b), since $\Gamma_{\vec{t}}$ is positive semidefinite.

Let $\varphi_{\vec{t}}(a) = \widehat{\nu_{\vec{t}}}(a) = \int\limits_{\mathbb{R}^n} e^{ia'x}\nu_{\vec{t}}(dx)$ be the Fourier transform of $\nu_{\vec{t}}$ (i.e. the characteristic function) at $a \in \mathbb{R}^n$, then

$$\varphi_{\vec{t}}(a) = \exp(-\frac{1}{2}a'\,\Gamma_{\vec{t}}\,a) \quad \forall a = (a_1, \dots, a_n) \in \mathbb{R}^n.$$

It follows that

$$\varphi_{\vec{t}}(a_1, \dots, a_{i-1}, 0, a_{i+1}, \dots, a_n) = \varphi_{(t_1,\dots,t_{i-1},t_{i+1},\dots,t_n)}(a_1, \dots, a_{i-1}, a_{i+1}, \dots, a_n). \tag{4.3}$$

Let $q_i : \mathbb{R}^n \to \mathbb{R}^{n-1}$, $(x_1, \dots, x_n) \mapsto (x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n)$. Denote by $q_i(\nu_{\vec{t}})$ the image measure of $\nu_{\vec{t}}$ under q_i. Using the transformation rule for integration with respect to image measures according to which

$$\int_{\mathbb{R}^{n-1}} f(z)\; q_i(\nu_{\vec{t}})(\mathrm{d}z) = \int_{\mathbb{R}^n} f(q_i(x))\, \nu_{\vec{t}}(\mathrm{d}x), \quad x \in \mathbb{R}^n,\, z \in \mathbb{R}^{n-1},$$

for positive (and measurable) or integrable $f : \mathbb{R}^{n-1} \to \mathbb{C}$, we calculate the Fourier transform of $q_i(\nu_{\vec{t}})$ as

$$\begin{aligned}
&\widehat{q_i(\nu_{\vec{t}})}(a_1, \dots, a_{i-1}, a_{i+1}, \dots, a_n) \\
&= \int\limits_{\mathbb{R}^{n-1}} e^{i(a_1x_1+\cdots+a_{i-1}x_{i-1}+a_{i+1}x_{i+1}+\cdots+a_nx_n)}\; q_i(\nu_{\vec{t}})\,(\mathrm{d}(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n)) \\
&= \int\limits_{\mathbb{R}^n} e^{i\langle(a_1,\dots,a_{i-1},a_{i+1},\dots,a_n)',q_i(x)\rangle}\nu_{\vec{t}}(\mathrm{d}x) \\
&= \int\limits_{\mathbb{R}^n} e^{i\langle(a_1,\dots,a_{i-1},0,a_{i+1},\dots,a_n)',x\rangle}\nu_{\vec{t}}(\mathrm{d}x) = \varphi_{\vec{t}}(a_1, \dots, a_{i-1}, 0, a_{i+1}, \dots, a_n).
\end{aligned}$$

Since the Fourier transformation determines the distribution, (4.3) implies that $q_i(\nu_{\vec{t}}) = \nu_{(t_1,\dots,t_{i-1},t_{i+1},\dots,t_n)}$. Kolmogorov's Consistency Theorem (Theorem 4.1) now implies that there exists a stochastic process $X = (X_t)_{t\in\mathbb{Z}}$ such that

$$(X_{t_1}, \dots, X_{t_n})' \overset{d}{=} N(\vec{0}, \Gamma_{\vec{t}}) \quad \forall\, n \in \mathbb{N} \quad \forall t_1 < \cdots < t_n$$

and since $\Gamma_{\vec{t}+h\vec{e}_n} = \Gamma_{\vec{t}}$ with $\vec{e}_n = (1, \dots, 1)' \in \mathbb{Z}^n$, X is strictly stationary and a Gaussian process. Especially, it holds true that $(X_i, X_j)' \sim N\left(\vec{0}, \begin{pmatrix} \gamma(0) & \gamma(i-j) \\ \gamma(i-j) & \gamma(0) \end{pmatrix}\right)$, so we have $\mathrm{Cov}\,(X_i, X_j) = \gamma(i-j)$. □

Remark 4.5. If $\gamma : \mathbb{Z} \to \mathbb{R}$ is an ACVF, then $\gamma(0) \geq 0$ and $|\gamma(h)| \leq \gamma(0)\ \forall h \in \mathbb{Z}$.

Proof. $\gamma(0) = \mathrm{Var}\, X_0 \geq 0$ and by the Cauchy-Schwarz inequality,

$$|\gamma(h)| = |\mathrm{Cov}\,(X_h, X_0)| \overset{\text{C.S.}}{\leq} \sqrt{\mathrm{Var}\, X_h}\sqrt{\mathrm{Var}\, X_0} = \gamma(0).$$

□

Remark 4.6. Sometimes it is hard to say if a function γ is positive semidefinite and it is easier to construct an example. The standard practice is:

i) If you think γ is not an ACVF, then check if γ fulfills the properties of Remark 4.5, or if γ is not even, or if γ is not positive semidefinite.

ii) If you think γ is an ACVF, then try to construct an explicit time series with this ACVF.

Example 4.7. The function $\kappa : \mathbb{Z} \to \mathbb{R}$, $h \mapsto \cos(wh)$ is for every $w \in \mathbb{R}$ an ACVF.

Proof. Let A, B be two uncorrelated random variables with $\mathbb{E}A = \mathbb{E}B = 0$ and $\operatorname{Var} A = \operatorname{Var} B = 1$. Define

$$X_t := A\cos(wt) + B\sin(wt).$$

It holds that $\mathbb{E}X_t = 0$ and

$$\begin{aligned}\operatorname{Cov}(X_{t+h}, X_t) &= \operatorname{Cov}(A\cos(w(t+h)) + B\sin(w(t+h)), A\cos(wt) + B\sin(wt)) \\ &= \operatorname{Var} A\cos(w(t+h))\cos(wt) + \operatorname{Var} B\sin(w(t+h))\sin(wt) \\ &= \cos(w(t+h) - wt) = \cos(wh) = \kappa(h).\end{aligned}$$

We conclude that X_t is weakly stationary with ACVF κ. □

When we have a realisation of a weakly stationary time series we will usually try to estimate the autocovariance function by the empirical autocovariances. Will this then also give a positive semidefinite empirical autocovariance function? Let us first give the definition:

Definition 4.8. *Let $x_1, \dots, x_n$ be realisations of a real-valued time series and*

$$\widehat{\gamma}(h) = \begin{cases} \frac{1}{n}\sum\limits_{t=1}^{n-|h|}(x_{t+|h|} - \overline{x})(x_t - \overline{x}), & h \in \{-n+1, \dots, n-1\} \\ 0, & h \in \mathbb{Z} : |h| \geq n \end{cases}$$

the empirical autocovariance function, as defined in Definition 2.17 (where we extend the definition by setting $\widehat{\gamma}(h) = 0$ for $|h| \geq n$). For $k \in \mathbb{N}$ we define the matrix

$$\widehat{\Gamma}_k := \begin{pmatrix} \widehat{\gamma}(0) & \widehat{\gamma}(1) & \dots & \widehat{\gamma}(k-1) \\ \widehat{\gamma}(1) & \widehat{\gamma}(0) & \dots & \widehat{\gamma}(k-2) \\ \vdots & & \ddots & \vdots \\ \widehat{\gamma}(k-1) & \widehat{\gamma}(k-2) & \dots & \widehat{\gamma}(0) \end{pmatrix} = (\widehat{\gamma}(i-j))_{i,j=1,\dots,k} \in \mathbb{R}^{k\times k}.$$

$\widehat{\Gamma}_k$ *is called* empirical autocovariance matrix *(of this realisation). Provided $\widehat{\gamma}(0) \neq 0$, the matrix $\widehat{R}_k := \frac{1}{\widehat{\gamma}(0)}\widehat{\Gamma}_k$ is called* empirical autocorrelation matrix.

Proposition 4.9. *The empirical autocovariance matrix $\widehat{\Gamma}_k$ of the realisations $x_1, \dots, x_n$ of a real valued time series is for all $k \in \mathbb{N}$ positive semidefinite.*

Proof. If $\widehat{\Gamma}_m$ is positive semidefinite, then $\widehat{\Gamma}_k$ is pos. semidefinite for $k < m$ as the upper left corner submatrix of $\widehat{\Gamma}_m$. So choose $k \geq n$. Set

$$Y_i = \begin{cases} x_i - \overline{x}_n, & i = 1, \dots, n \\ 0, & i = n+1, \dots k \end{cases}$$

and define the $(k \times 2k)$-matrix

$$T := \begin{pmatrix} 0 & \dots & 0 & 0 & Y_1 & Y_2 & \dots & Y_k \\ 0 & \dots & 0 & Y_1 & Y_2 & \dots & Y_k & 0 \\ \vdots & & \iddots & & \vdots & \iddots & & \vdots \\ 0 & Y_1 & Y_2 & \dots & Y_k & 0 & \dots & 0 \end{pmatrix}.$$

Then it holds that $\widehat{\Gamma}_k = \frac{1}{n} TT'$. If $a \in \mathbb{R}^k$, then

$$a^T \widehat{\Gamma}_k a = n^{-1}(a'T)(T'a) \geq 0,$$

and $\widehat{\Gamma}_k$ is clearly symmetric. □

We shall now come to a characterisation similar to Theorem 4.4 for complex valued time series. We need some definitions:

Definition 4.10. (a) A complex matrix $A \in \mathbb{C}^{n \times n}$ is called *hermitian*, if $\overline{A}' = A$. It is called *positive semidefinite*, if additionally

$$\overline{a}' A a \geq 0 \quad \forall\, a \in \mathbb{C}^n.$$

(b) A function $\kappa : \mathbb{Z} \to \mathbb{C}$ is called *hermitian*, if $\kappa(h) = \overline{\kappa(-h)}$ for all $h \in \mathbb{Z}$. It is called *positive semidefinite*, if

$$\sum_{j,k=1}^{n} a_j \kappa(t_j - t_k) \overline{a_k} \geq 0 \quad \forall\, n \in \mathbb{N}, \quad \forall\, a_1, \dots, a_n \in \mathbb{C}, \quad \forall\, t_1, \dots, t_n \in \mathbb{Z}.$$

Observe that if $A \in \mathbb{R}^{n \times n}$ is such that $a'Aa \geq 0$ for all $a \in \mathbb{R}^n$, then A does not need to be symmetric. An example is given by the matrix $A = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \in \mathbb{R}^{2 \times 2}$, for which $a'Aa = 0$ for all $a \in \mathbb{R}^2$. This is the reason why by definition a real matrix $A \in \mathbb{R}^{n \times n}$ is called *positive semidefinite* if it is symmetric *and* satisfies $a'Aa \geq 0$ for all $a \in \mathbb{R}^n$, so symmetry is a part of positive semidefiniteness. For complex matrices, it is a bit different. We also required a positive semidefinite matrix to be hermitian by definition, but as it turns out, this is not necessary, it is already implied by $\overline{a}'Aa \in \mathbb{R}$ for all $a \in \mathbb{C}^n$. More precisely, we have:

Lemma 4.11. *Let $A = B + \mathrm{i}C \in \mathbb{C}^{n \times n}$ with $B, C \in \mathbb{R}^{n \times n}$ such that $\overline{a}'Aa \in \mathbb{R}$ for all $a \in \mathbb{C}^n$. Then A must be hermitian, i.e. $\overline{A}' = A$, i.e. $B' = B$ and $C' = -C$.*

Proof. We have $\overline{a}'(B + \mathrm{i}C)a \in \mathbb{R}$ for all $a \in \mathbb{C}^n$, hence also $a'(B + \mathrm{i}C)a \in \mathbb{R}$ for all $a \in \mathbb{R}^n$. This implies $a'Ca = 0$ for all $a \in \mathbb{R}^n$. Taking the transpose of this equation implies $a'C'a = 0$ for all $a \in \mathbb{R}^n$, and adding both equations up shows $a'(C + C')a = 0$ for all $a \in \mathbb{R}^n$. Since $C + C'$ is symmetric, it can be diagonalised, and the above equation says that the only eigenvalue of $C + C'$ is 0, so that $C + C' = 0$. Hence $C' = -C$.

To see that B is symmetric, write $a = d + \mathrm{i}f$ with $d, f \in \mathbb{R}^n$ for $a \in \mathbb{C}^n$. Then by assumption we have

$$(d' - \mathrm{i}f')(B + \mathrm{i}C)(d + \mathrm{i}f) \in \mathbb{R} \quad \forall\, d, f \in \mathbb{R}^n.$$

Hence the imaginary part of the left-hand side must be zero, and calculating this imaginary part we obtain

$$d'Bf + d'Cd - f'Bd + f'Cf = 0 \quad \forall\, d, f \in \mathbb{R}^n.$$

Now write $B = (b_{jk})_{j,k=1,\ldots,n}$ and $C = (c_{jk})_{j,k=1,\ldots,n}$ and take $d = e_j$ and $f = e_k$, the j'th and k'th unit vector in $\mathbb{R}^n$. Then $d'Bf = b_{jk}$, $f'Bd = b_{kj}$, $d'Cd = c_{jj}$ and $f'Cf = c_{kk}$. But $c_{jj} = c_{kk} = 0$ as a consequence of $C = -C'$. Hence the above equation gives

$$b_{jk} - b_{kj} = 0 \quad \forall\, j, k \in \{1, \ldots, n\},$$

showing that $B = B'$. □

We can now give the analogue to Theorem 4.4:

Theorem 4.12. *Let $\gamma : \mathbb{Z} \to \mathbb{C}$ be a function. Then the following statements are equivalent:*

(i) γ is the ACVF of a complex-valued weakly stationary time series.

(ii) γ is hermitian and positive semidefinite.

(iii) γ is positive semidefinite.

Proof. "(i) $\Longrightarrow$ *(ii)*": That γ is hermitian was already observe in Remark 2.19 (a). To see that it is positive semidefinite, let $\vec{t} = (t_1, \ldots, t_n)' \in \mathbb{Z}^n$ and $a_1, \ldots, a_n \in \mathbb{C}$. Then

$$\begin{aligned} 0 \;\; &\leq \;\; \mathrm{Var}\left(\sum_{j=1}^n a_j X_{t_j}\right) = \mathrm{Cov}\left(\sum_{j=1}^n a_j X_{t_j}, \sum_{k=1}^n a_k X_{t_k}\right) \\ &= \sum_{j,k=1}^n a_j \mathrm{Cov}\,(X_{t_j}, X_{t_k}) \overline{a}_k = \sum_{j,k=1}^n a_j \overline{a}_k \gamma(t_j - t_k). \end{aligned}$$

"(ii) $\Longrightarrow$ (iii)" is clear.

"(iii) $\Longrightarrow$ (i)": Let $\kappa : \mathbb{Z} \to \mathbb{C}$ be positive semidefinite. For $n \in \mathbb{N}$ and $\vec{t} = (t_1, \ldots, t_n)' \in \mathbb{Z}^n$ we have by assumption

$$\sum_{j,k=1}^n a_j \kappa(t_j - t_k) \overline{a_k} \geq 0 \quad \forall\, n \in \mathbb{N} \quad \forall\, a_1, \ldots, a_n \in \mathbb{C}. \tag{4.4}$$

Denoting

$$a = (a_1, \dots, a_n)' \in \mathbb{C}^n \quad \text{and} \quad A_{\vec{t}} := (\kappa(t_j - t_k))_{j,k=1,\dots,n},$$

Equation (4.4) can be rewritten as $a' A_{\vec{t}} \overline{a} \geq 0$ for all $a \in \mathbb{C}^n$, and replacing a by $\overline{a}$ it can be rewritten as

$$\overline{a}' A_{\vec{t}} a \geq 0 \quad \forall\, n \in \mathbb{N} \quad \forall\, a \in \mathbb{C}^n. \tag{4.5}$$

From Lemma 4.11 we see that $A_{\vec{t}}$ is hermitian and positive semidefinite. Write

$$A_{\vec{t}} = B_{\vec{t}} + \mathrm{i} C_{\vec{t}} \quad \text{with} \quad B_{\vec{t}}, C_{\vec{t}} \in \mathbb{R}^{n \times n},$$

where $C'_{\vec{t}} = -C_{\vec{t}}$ and $B'_{\vec{t}} = B_{\vec{t}}$. Writing $a = d + \mathrm{i}f$ with $d, f \in \mathbb{R}^n$, similar to the proof of Lemma 4.11, we have from (4.5)

$$\begin{aligned} 0 &\leq (d' - \mathrm{i}f')(B_{\vec{t}} + \mathrm{i}C_{\vec{t}})(d + \mathrm{i}f) \\ &= \big(d'B_{\vec{t}}d - d'C_{\vec{t}}f + f'B_{\vec{t}}f + f'C_{\vec{t}}d\big) + \mathrm{i}\underbrace{\big(d'B_{\vec{t}}f + d'C_{\vec{t}}d - f'B_{\vec{t}}d + f'C_{\vec{t}}f\big)} = 0. \end{aligned} \tag{4.6}$$

Define the real valued $2n \times 2n$-matrix

$$\Lambda_{\vec{t}} := \begin{pmatrix} B_{\vec{t}} & -C_{\vec{t}} \\ C_{\vec{t}} & B_{\vec{t}} \end{pmatrix} \in \mathbb{R}^{2n \times 2n}.$$

Then $\Lambda_{\vec{t}}$ is symmetric since $B_{\vec{t}} = B'_{\vec{t}}$ and $C'_{\vec{t}} = -C_{\vec{t}}$, and for each vector $\begin{pmatrix} d \\ f \end{pmatrix} \in \mathbb{R}^{2n}$ we have from (4.6)

$$(d', f')\Lambda_{\vec{t}} \begin{pmatrix} d \\ f \end{pmatrix} = d'B_{\vec{t}}d - d'C_{\vec{t}}f + f'C_{\vec{t}}d + f'B_{\vec{t}}f \geq 0.$$

Hence the matrix $\Lambda_{\vec{t}} \in \mathbb{R}^{2n \times 2n}$ is symmetric and positive semidefinite, hence a covariance matrix. For each $\vec{t} \in \mathbb{Z}^n$ let $\nu_{\vec{t}}$ be a normal distribution in $\mathbb{R}^{2n}$ with mean $\vec{0}$ and covariance matrix $\Lambda_{\vec{t}}$. Denote by

$$\varphi_{\vec{t}}(b, c) := \int_{\mathbb{R}^{2n}} \mathrm{e}^{\mathrm{i}(b'x + c'y)}\, \nu_{\vec{t}}(\mathrm{d}(x, y)),$$

with $x, y \in \mathbb{R}^n$ and $b, c \in \mathbb{R}^n$, the Fourier transform of $\nu_{\vec{t}}$ at $(b', c')'$. Then

$$\varphi_{\vec{t}}(b, c) = \exp\left(-\frac{1}{2}(b', c')\Lambda_{\vec{t}}(b', c')'\right) \quad \forall\, (b', c')' \in \mathbb{R}^{2n}. \tag{4.7}$$

For fixed $i \in 1, \dots, n$ let $q_i : \mathbb{R}^n \to \mathbb{R}^{n-1}$, $(x_1, \dots, x_n)' \mapsto (x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n)'$ and $Q_i : \mathbb{R}^{2n} \to \mathbb{R}^{2(n-1)}$, $(x', y')' \mapsto ((q_i(x))', (q_i(y))')$. Writing $b = (b_1, \dots, b_n)'$ and $c = (c_1, \dots, c_n)'$ we have

$$\begin{aligned} &\widehat{Q_i(\nu_{\vec{t}})}(b_1, \dots, b_{i-1}, b_{i+1}, \dots, b_n, c_1, \dots, c_{i-1}, c_{i+1}, \dots, c_n) \\ &\quad = \int_{\mathbb{R}^{2(n-1)}} \mathrm{e}^{\mathrm{i}(c_1x_1 + \dots + c_{i-1}x_{i-1} + c_{i+1}x_{i+1} + \dots + c_nx_n + d_1y_1 + \dots + d_{i-1}y_{i-1} + d_{i+1}y_{i+1} + \dots + d_ny_n)} \\ &\qquad\qquad Q_i(\nu_{\vec{t}})(\mathrm{d}(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n, y_1, \dots, y_{i-1}, y_{i+1}, \dots, y_n)) \\ &\quad = \int_{\mathbb{R}^{2n}} \mathrm{e}^{\mathrm{i}(\langle q_i(b), q_i(x)\rangle + \langle q_i(c), q_i(y)\rangle)}\, \nu_{\vec{t}}(\mathrm{d}(x, y)) \\ &\quad = \varphi_{\vec{t}}(b_1, \dots, b_{i-1}, 0, b_{i+1}, \dots, b_n, c_1, \dots, c_{i-1}, 0, c_{i+1}, \dots, c_n). \end{aligned} \tag{4.8}$$

Denote

$$\widetilde{b} = (b_1, \ldots, b_{i-1}, 0, b_{i+1}, \ldots, b_n)' \quad \text{and} \quad \widetilde{c} = (c_1, \ldots, c_{i-1}, 0, c_{i+1}, \ldots, c_n)'.$$

Then

$$\begin{aligned}
&\widehat{Q_i(\nu_{\vec{t}})}(q_i(b), q_i(c)) \\
&\overset{(4.8)}{=} \quad \varphi_{\vec{t}}(\widetilde{b}, \widetilde{c}) \\
&\overset{(4.7)}{=} \quad \exp\left[-\frac{1}{2}(\widetilde{b}', \widetilde{c}')\begin{pmatrix} B_{\vec{t}} & -C_{\vec{t}} \\ C_{\vec{t}} & B_{\vec{t}} \end{pmatrix}\begin{pmatrix} \widetilde{b} \\ \widetilde{c} \end{pmatrix}\right] \\
&= \quad \exp\left[-\frac{1}{2}\left[\widetilde{b}' B_{\vec{t}} \widetilde{b} - \widetilde{b}' C_{\vec{t}} \widetilde{c} + \widetilde{c}' C_{\vec{t}} \widetilde{b} - \widetilde{c}' B_{\vec{t}} \widetilde{c}\right]\right] \\
&= \quad \exp\left[-\frac{1}{2}\left[q_i(b)' B_{q_i(\vec{t})}\, q_i(b) - q_i(b)' C_{q_i(\vec{t})}\, q_i(c) + q_i(c)' C_{q_i(\vec{t})}\, q_i(b) - q_i(c)' B_{q_i(\vec{t})}\, q_i(c)\right]\right] \\
&= \quad \exp\left[-\frac{1}{2}(q_i(b)', q_i(c)')\begin{pmatrix} B_{q_i(\vec{t})} & -C_{q_i(\vec{t})} \\ C_{q_i(\vec{t})} & B_{q_i(\vec{t})} \end{pmatrix}\begin{pmatrix} q_i(b) \\ q_i(c) \end{pmatrix}\right] \\
&\overset{(4.7)}{=} \quad \varphi_{q_i(\vec{t})}(q_i(b), q_i(c)).
\end{aligned}$$

This shows $\widehat{Q_i(\nu_{\vec{t}})} = \widehat{\nu_{q_i(\vec{t})}}$ and hence $Q_i(\nu_{\vec{t}}) = \nu_{q_i(\vec{t})} = \nu_{(t_1,\ldots,t_{i-1},t_{i+1},\ldots,t_n)'}$. Since Q_i can be seen as the projection of $(\mathbb{R}^2)^n$ onto $(\mathbb{R}^2)^{n-1}$, Kolmogorov's consistency theorem (Theorem 4.1) ensures the existence of an $\mathbb{R}^2$-valued stochastic process $(Y', Z')' = ((Y_n, Z_n)')_{n\in\mathbb{Z}}$ such that

$$(Y_{t_1}, \ldots, Y_{t_n}, Z_{t_1}, \ldots, Z_{t_n})' \overset{d}{=} N(\vec{0}, \Lambda_{\vec{t}}) \quad \forall\, n \in \mathbb{N} \quad \forall\, t_1 < \ldots < t_n.$$

Since $\Lambda_{\vec{t}+h\vec{e}_n} = \Lambda_{\vec{t}}$ for $h \in \mathbb{Z}$ with $\vec{e}_n = (1, \ldots, 1)' \in \mathbb{R}^n$, we see that $(Y', Z')'$ is strictly stationary, in the sense that its finite dimensional distributions are shift invariant. Now set

$$X_n = Y_n - \mathrm{i}Z_n, \quad n \in \mathbb{Z}.$$

Then $X = (X_n)_{n\in\mathbb{Z}}$ is $\mathbb{C}$-valued and strictly stationary (by strict stationarity of $(Y', Z')'$, easily checked), and since X has finite second moments, it is also weakly stationary. Further,

$$\mathbb{E}(X_n) = \mathbb{E}(Y_n) - \mathrm{i}\mathbb{E}(Z_n) = 0.$$

Since $(Y_h, Y_0, Z_h, Z_0)' \overset{d}{=} N(\vec{0}, \Lambda_{(h,0)'}) \in \mathbb{R}^{4\times 4}$ we have

$$\Lambda_{(h,0)'} = \begin{pmatrix} \operatorname{Cov}(Y_h, Y_h) & \operatorname{Cov}(Y_h, Y_0) & \operatorname{Cov}(Y_h, Z_h) & \operatorname{Cov}(Y_h, Z_0) \\ \operatorname{Cov}(Y_0, Y_h) & \operatorname{Cov}(Y_0, Y_0) & \operatorname{Cov}(Y_0, Z_h) & \operatorname{Cov}(Y_0, Z_0) \\ \operatorname{Cov}(Z_h, Y_h) & \operatorname{Cov}(Z_h, Y_0) & \operatorname{Cov}(Z_h, Z_h) & \operatorname{Cov}(Z_h, Z_0) \\ \operatorname{Cov}(Z_0, Y_h) & \operatorname{Cov}(Z_0, Y_0) & \operatorname{Cov}(Z_0, Z_h) & \operatorname{Cov}(Z_0, Z_0) \end{pmatrix}.$$

On the other hand, by definition, $A_{(h,0)'} = \begin{pmatrix} \kappa(0) & \kappa(-h) \\ \kappa(h) & \kappa(0) \end{pmatrix} \in \mathbb{R}^{2\times 2}$ and hence

$$\Lambda_{(h,0)'} = \begin{pmatrix} \Re A_{(h,0)'} & -\Im A_{(h,0)'} \\ \Im A_{(h,0)'} & \Re A_{(h,0)'} \end{pmatrix} = \begin{pmatrix} \Re\kappa(0) & \Re\kappa(-h) & -\Im\kappa(0) & -\Im\kappa(-h) \\ \Re\kappa(h) & \Re\kappa(0) & -\Im\kappa(h) & -\Im\kappa(0) \\ \Im\kappa(0) & \Im\kappa(-h) & \Re\kappa(0) & \Re\kappa(-h) \\ \Im\kappa(h) & \Im\kappa(0) & \Re\kappa(h) & \Re\kappa(0) \end{pmatrix}.$$

Equating both expressions for $\Lambda_{(h,0)'}$ we obtain

$$\begin{aligned}
\gamma_X(h) &= \operatorname{Cov}(X_h, X_0) = \operatorname{Cov}(Y_h - \mathrm{i}Z_h, Y_0 - \mathrm{i}Z_0) \\
&= \operatorname{Cov}(Y_h, Y_0) + \mathrm{i}\operatorname{Cov}(Y_h, Z_0) - \mathrm{i}\operatorname{Cov}(Z_h, Y_0) + \operatorname{Cov}(Z_h, Z_0) \\
&= \operatorname{Cov}(Y_0, Y_h) + \mathrm{i}\operatorname{Cov}(Z_0, Y_h) - \mathrm{i}\operatorname{Cov}(Y_0, Z_h) + \operatorname{Cov}(Z_0, Z_h) \\
&= \Re\kappa(h) + \mathrm{i}\Im\kappa(h) + \mathrm{i}\Im\kappa(h) + \Re\kappa(h) \\
&= 2\kappa(h).
\end{aligned}$$

This shows that 2κ is the ACVF of the weakly stationary complex time series X. Then κ is the ACVF of the weakly stationary time series $(\frac{1}{\sqrt{2}}X_t)_{t\in\mathbb{Z}}$. □

Remark 4.13. As in Remark 4.5 we can show that the ACVF $\gamma : \mathbb{Z} \to \mathbb{C}$ of a weakly stationary time series satisfies $\gamma(0) \geq 0$ and $|\gamma(h)| \leq \gamma(0)$ for all $h \in \mathbb{Z}$.

Remark 4.14. If $x_1, \ldots, x_n$ is a realisation of a complex-valued stationary time series, we can define the *empirical autocovariance function* $\widehat{\gamma}(h)$ for $h = -n+1, \ldots, n-1$ as in Remark 2.19 and $\widehat{\gamma}(h) = 0$ for $|h| \geq n$. Then the empirical *autocovariance matrix* is defined by

$$\widehat{\Gamma}_k = (\widehat{\gamma}(i-j))_{i,j=1,\ldots,k} \in \mathbb{C}^{k\times k}.$$

Provided $\widehat{\gamma}(0) \neq 0$, the matrix $\widehat{R}_k := \frac{1}{\widehat{\gamma}(0)}\widehat{\Gamma}_k$ is called the *empirical autocorrelation matrix*.

Chapter 5

Linear filters and two-sided moving average processes of infinite order

In this chapter we shall learn how to construct new stationary time series from given ones. This will be done by applying so called linear filters.

Let $(Y_n)_{n\in\mathbb{N}}$ be a sequence of $\mathbb{C}$-valued random variables and Y a $\mathbb{C}$-valued random variable, all defined on the same probability space $(\Omega, \mathcal{F}, P)$. Recall that (Y_n) *converges in* L^p to Y as $n \to \infty$, where $p \in [1, \infty)$, if $\mathbb{E}|Y_n|^p < \infty$ for all n, $\mathbb{E}|Y|^p < \infty$ and

$$\lim_{n\to\infty} \mathbb{E}|Y_n - Y|^p = 0;$$

it *converges in probability* to Y, if

$$\lim_{n\to\infty} P(|Y_n - Y| > \varepsilon) = 0 \quad \forall\, \varepsilon > 0;$$

it converges *almost surely to* Y if the set $\{\omega \in \Omega : \lim_{n\to\infty} Y_n(\omega) = Y(\omega)\}$ has probability 1. The limit random variable Y is in all cases unique almost surely, and both almost sure convergence as well as L^p-convergence imply convergence in probability to the same limit, but other implications do not hold without extra assumptions. The L^p*-norm* of a random variable $Y : \Omega \to \mathbb{C}$ is defined by

$$\|Y\|_p = (\mathbb{E}|Y|^p)^{1/p}$$

for $p \in [1, \infty)$ and it is known that $L^p(\Omega, \mathcal{F}, P; \mathbb{C})$, the space of (equivalence classes of) p-integrable complex random variables (p-integrable means $\|Y\|_p < \infty$), is complete, i.e. that every Cauchy sequence in L^p connverges in L^p.

Given a sequence $(X_n)_{n\in\mathbb{N}}$ of complex random random variables, we say that the series $\sum_{n=1}^{\infty} X_n$ converges almost surely (resp. in L^p or in probability), if the sequence $Y_n = \sum_{j=1}^{n} X_j$, $n \in \mathbb{N}$, of partial sums converges almost surely (resp. in L^p or in probability) to some random variable Y, and we say that it *converges almost surely absolutely*, if

$$P\left(\left\{\omega \in \Omega : \sum_{n=1}^{\infty} |X_n(\omega)| < \infty\right\}\right) = 1.$$

Since absolute convergence of a series implies its convergence, this implies that $\sum_{n=1}^{\infty} X_n$ also converges almost surely. Finally, given a double series $\sum_{n=-\infty}^{\infty} X_n$, where the X_n are random variables, we say that it converges almost surely (resp. almost surely absolutely, or in L^p or in probability), if both the sums $\sum_{n=0}^{\infty} X_n$ and $\sum_{n=1}^{\infty} X_{-n}$ converge almost surely (almost surely absolutely, or in L^p or in probability, respectively). It is clear that a double series $\sum_{n=-\infty}^{\infty} X_n$ converges almost surely absolutely if and only if

$$P\left(\left\{\omega \in \Omega : \sum_{n=-\infty}^{\infty} |X_n(\omega)| < \infty\right\}\right) = 1.$$

A handy criterion for almost sure absolute convergence of an infinite series or a double series is the following:

Lemma 5.1. *Let $(X_n)_{n\in\mathbb{Z}}$ be a sequence of complex valued random variables on a probability space $(\Omega, \mathcal{F}, P)$ and let $p \in [1, \infty)$. Assume that $\sum_{n=-\infty}^{\infty} \|X_n\|_p < \infty$. Then $\sum_{n=-\infty}^{\infty} X_n$ converges in L^p and also almost surely absolutely. It also converges in $L^{p'}$ (to the same limit) for every $p' \in [1, p]$ and*

$$\left\| \sum_{n=-\infty}^{\infty} X_n \right\|_{p'} \leq \sum_{n=-\infty}^{\infty} \|X_n\|_{p'} \leq \sum_{n=-\infty}^{\infty} \|X_n\|_p < \infty. \tag{5.1}$$

Proof. For $1 \leq p' < p$, an application of Hölder's inequality (applied with p/p' and $(1 - p'/p)^{-1}$) we have for a random variable Y that

$$\|Y\|_{p'}^{p'} = \mathbb{E}(|Y|^{p'} \cdot 1) \leq \left(\mathbb{E}(|Y|^{p'})^{p/p'}\right)^{p'/p} (\mathbb{E}(1^{(1-p'/p)^{-1}}))^{1-p'/p} = \|Y\|_p^{p'},$$

so that $\|Y\|_{p'} \leq \|Y\|_p$ whenever $1 \leq p' \leq p < \infty$, and by assumption we see that also $\sum_{n=-\infty}^{\infty} \|X_n\|_{p'} \leq \sum_{n=-\infty}^{\infty} \|X_n\|_p < \infty$. Using Minkowski's inequality we obtain for $m, n \in \mathbb{N}$, $n \geq m$, and $1 \leq p' \leq p < \infty$ that

$$\left\| \sum_{j=m+1}^{n} X_j \right\|_{p'} \leq \sum_{j=m+1}^{n} \|X_j\|_{p'},$$

and since $\sum_{j=0}^{\infty} \|X_j\|_{p'} < \infty$, this is a Cauchy sequence in $L^{p'}$. It follows that $\sum_{j=0}^{\infty} X_j$ converges in $L^{p'}$, and similarly for $\sum_{j=1}^{\infty} X_{-j}$. Also, by Minkowski's inequality, $\| \sum_{j=0}^{n} X_j\|_{p'} \leq \sum_{j=0}^{n} \|X_j\|_{p'}$ for every $n \in \mathbb{N}$, so that taking the limit gives $\| \sum_{j=0}^{\infty} X_j\|_{p'} \leq \sum_{j=0}^{\infty} \|X_j\|_{p'}$ and similarly for $\sum_{j=-\infty}^{-1} X_j$ so that we obtain (5.1).

To see that the sum converges almost surely absolutely, define the numerical random variable $Y : \Omega \to [0, \infty]$ by $Y = \sum_{n=-\infty}^{\infty} |X_n|$. Then by monotone convergence,

$$\mathbb{E}Y = \mathbb{E} \sum_{n=-\infty}^{\infty} |X_n| = \sum_{n=-\infty}^{\infty} \mathbb{E}|X_n| = \sum_{n=-\infty}^{\infty} \|X_n\|_1 < \infty$$

(using $p' := 1$). But a numerical random variable with finite expectation must be almost surely finite, i.e. $P(Y < \infty) = 1$. This shows almost sure absolute convergence of $\sum_{n=-\infty}^{\infty} X_n$. □

As an immediate corollary we obtain:

Corollary 5.2. *Let $p \in [1,\infty)$ and $(X_t)_{t\in\mathbb{Z}}$ be a sequence of complex-valued random variables with*

$$\sup_{t\in\mathbb{Z}} \mathbb{E}|X_t|^p < \infty.$$

Let $(\psi_j)_{j\in\mathbb{Z}}$ be an absolutely summable sequence of complex numbers, i.e. such that

$$\sum_{j\in\mathbb{Z}} |\psi_j| < \infty.$$

Then for each $t \in \mathbb{Z}$, the series $\sum_{j=-\infty}^{\infty} \psi_j X_{t-j}$ converges almost surely (a.s.) absolutely and in $L^{p'}$ for every $p' \in [1,p]$.

Proof. This follows immediately from the previous lemma (and its proof) by observing that $\|\psi_j X_{t-j}\|_{p'} = |\psi_j| \, \|X_{t-j}\|_{p'} \leq |\psi_j| \|X_{t-j}\|_p$ for $p' \in [1,p]$. □

The previous corollary allows us to obtain stationary sequences $(Y_t)_{t\in\mathbb{Z}}$ from a given stationary time series $(X_t)_{t\in\mathbb{Z}}$ by defining $Y_t = \sum_{j=-\infty}^{\infty} \psi_j X_{t-j}$, provided $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$. More precisely, we have

Theorem 5.3. *Let $(X_t)_{t\in\mathbb{Z}}$ be a $\mathbb{C}$-valued time series and $(\psi_j)_{j\in\mathbb{Z}}$ be a sequence in $\mathbb{C}$ with $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$.*
(a) Suppose that $(X_t)_{t\in\mathbb{Z}}$ is weakly stationary with ACVF γ_X. Then for every $t \in \mathbb{Z}$, the sum

$$Y_t := \sum_{j=-\infty}^{\infty} \psi_j X_{t-j}$$

converges a.s. absolutely, as well as in L^1 and in L^2. Further, the sequence $(Y_t)_{t\in\mathbb{Z}}$ is weakly stationary with ACVF γ_Y given by

$$\gamma_Y(h) = \sum_{j,k=-\infty}^{\infty} \psi_j \overline{\psi_k} \gamma_X(h-j+k), \quad h \in \mathbb{Z},$$

(the double sum converging absolutely) and with mean $\mathbb{E}X_0 \sum_{j=-\infty}^{\infty} \psi_j$.
(b) Suppose that $(X_t)_{t\in\mathbb{Z}}$ is strictly stationary and that $Y_t := \sum_{j=-\infty}^{\infty} \psi_j X_{t-j}$ converges in probability for each $t \in \mathbb{Z}$ (satisfied for example under the assumptions of (a)). Then $Y = (Y_t)_{t\in\mathbb{Z}}$ is strictly stationary.

Proof. (a) Since X is weakly stationary, and

$$\mathbb{E}|X_t|^2 = \mathrm{Var} X_t + |\mathbb{E}X_t|^2$$

and since $\mathrm{Var}(X_t)$ and $\mathbb{E}X_t$ do not depend on t (by weak stationarity), it follows that $\sup_{t\in\mathbb{Z}} \mathbb{E}|X_t|^2 < \infty$. Hence by Corollary 5.2, the sum defining Y_t converges almost surely absolutely, as well as in L^1 and in L^2 (to the same limit). Then also $\sum_{j=-n}^{n} \psi_j X_{t-j}$ converges in L^1 and in L^2 to Y_t as $n \to \infty$. Using the L^1-convergence we obtain

$$\mathbb{E}X_0 \sum_{j=-n}^{n} \psi_j = \mathbb{E} \sum_{j=-n}^{n} \psi_j X_{t-j} \to \mathbb{E}Y_t,\ n \to \infty,$$

so that Y_t has expectation $\mathbb{E}X_0 \sum_{j=-\infty}^{\infty} \psi_j$.

We already know that $\mathbb{E}|Y_t|^2 < \infty$ for each t, so we only have to show that the autocovariance function of Y depends only on its lag. For that, fix $h, t \in \mathbb{Z}$ and define for each $n \in \mathbb{Z}$

$$A_n := \sum_{j=-n}^{n} \psi_j X_{t+h-j} \quad \text{and} \quad B_n := \sum_{j=-n}^{n} \overline{\psi_j}\, \overline{X_{t-j}}.$$

Then A_n converges in L^2 to Y_{t+h} and B_n converges in L^2 to $\overline{Y}_t$ as $n \to \infty$. We hence conclude from the Cauchy-Schwarz inequality that

$$\begin{aligned}
&\mathbb{E}|Y_{t+h}\overline{Y_t} - A_n B_n| \\
&\quad \leq\ \mathbb{E}|(Y_{t+h}\overline{Y_t} - A_n\overline{Y_t}) + (A_n\overline{Y_t} - A_n B_n)| \\
&\quad \leq\ \underbrace{\sqrt{\mathbb{E}|Y_{t+h} - A_n|^2}}_{\to 0} \sqrt{\mathbb{E}|\overline{Y_t}|^2} + \underbrace{\sqrt{\mathbb{E}|A_n|^2}}_{bounded} \underbrace{\sqrt{\mathbb{E}|\overline{Y_t} - B_n|^2}}_{\to 0} \to 0, \quad n \to \infty,
\end{aligned}$$

so

$$\begin{aligned}
\mathbb{E}(Y_{t+h}\overline{Y_t}) = \lim_{n\to\infty} \mathbb{E}(A_n B_n) \ &= \ \lim_{n\to\infty} \mathbb{E} \sum_{j,k=-n}^{n} \psi_j \overline{\psi_k} X_{t+h-j} \overline{X_{t-k}} \\
&= \ \lim_{n\to\infty} \sum_{j,k=-n}^{n} \psi_j \overline{\psi_k} \big(\gamma_X(h-j+k) + |\mathbb{E}X_0|^2\big), \qquad (5.2)
\end{aligned}$$

which implies that $\mathbb{E}(Y_{t+h}\overline{Y_t})$ does not depend on t, and since also $\mathbb{E}Y_t$ does not depend on t, we conclude that Cov (Y_{t+h}, Y_t) does not depend on t so that $(Y_t)_{t\in\mathbb{Z}}$ is weakly stationary. Observe that

$$\sum_{j,k=-\infty}^{\infty} |\psi_j \overline{\psi_k} \gamma_X(h-j+k)| \leq \underbrace{\sup_{i\in\mathbb{Z}} |\gamma_X(i)|}_{\leq \gamma_X(0) < \infty} \sum_{j=-\infty}^{\infty} |\psi_j| \sum_{k=-\infty}^{\infty} |\psi_k| < \infty$$

by assumption, and similarly $\sum_{j,k=-\infty}^{\infty} \left|\psi_j \overline{\psi_k} |\mathbb{E}X_0|^2\right| < \infty$. From (5.2) and $\mathbb{E}Y_t = \mathbb{E}X_0 \sum_{j=-\infty}^{\infty} \psi_j$ we then obtain

$$\gamma_Y(h) \ = \ \mathbb{E}(Y_{t+h}\overline{Y_t}) - \mathbb{E}Y_{t+h}\mathbb{E}\overline{Y_t} = \sum_{i,k=-\infty}^{\infty} \psi_j \overline{\psi_k} \gamma_X(h-j+k).$$

(b) Let $t_1 < \cdots < t_n \in \mathbb{Z}$ and $h \in \mathbb{Z}$. Define for all $m \in \mathbb{N}$ the $\mathbb{C}^{(2m+1)n}$-valued random vectors

$$\begin{aligned}
U_m \ &:= \ (X_{t_1+h-m}, \ldots, X_{t_1+h+m}, \ldots, X_{t_n+h-m}, \ldots, X_{t_n+h+m})' \\
\text{and} \quad V_m \ &:= \ (X_{t_1-m}, \ldots, X_{t_1+m}, \ldots, X_{t_n-m}, \ldots, X_{t_n+m})'.
\end{aligned}$$

Then $U_m \overset{d}{=} V_m$ for all $m \in \mathbb{Z}$ since X is strictly stationary. Now define the mapping

$$\begin{aligned} \Psi_m : \mathbb{C}^{(2m+1)n} &\to \mathbb{C}^n, \\ (z_{-m,1}, \ldots, z_{m,1}, \ldots, z_{-m,n}, \ldots, z_{m,n})' &\mapsto \left(\sum_{j=-m}^{m} \psi_j z_{-j,1}, \ldots, \sum_{j=-m}^{m} \psi_j z_{-j,n} \right)'. \end{aligned}$$

Then

$$\begin{aligned} \left(\sum_{j=-m}^{m} \psi_j X_{t_1+h-j}, \ldots, \sum_{j=-m}^{m} \psi_j X_{t_n+h-j} \right)' &= \Psi_m(U_m) \\ &\overset{d}{=} \Psi_m(V_m) \\ &= \left(\sum_{j=-m}^{m} \psi_j X_{t_1-j}, \ldots, \sum_{j=-m}^{m} \psi_j X_{t_n-j} \right)'. \end{aligned}$$

But the left-hand side converges in probability to $(Y_{t_1+h}, \ldots, Y_{t_n+h})'$ and the right-hand side in probability to $(Y_{t_1}, \ldots, Y_{t_n})'$ as $m \to \infty$. Since left-hand side and right-hand side have the same distribution, so do their probability limits, showing that $(Y_{t_1+h}, \ldots, Y_{t_n+h})' \overset{d}{=} (Y_{t_1}, \ldots, Y_{t_n})'$, so that Y is strictly stationary. □

Applying the previous theorem to the case when $X \sim WN(0, \sigma^2)$ we obtain:

Corollary 5.4. *Let $(Z_t)_{t\in\mathbb{Z}} \sim WN(0, \sigma^2)$ and $(\psi_j)_{j\in\mathbb{Z}}$ a sequence in $\mathbb{C}$ with $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$. Then $(Y_t = \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j})_{t\in\mathbb{Z}}$ is weakly stationary with ACVF*

$$\gamma_Y(h) = \sigma^2 \sum_{j=-\infty}^{\infty} \psi_j \overline{\psi_{j-h}}.$$

Proof. This is clear from Theorem 5.3 since $\gamma_X(0) = \sigma^2$ and $\gamma_X(h) = 0$ for $h \neq 0$. □

The processes arising in Corollary 5.4 have special names:

Definition 5.5. Let $(Z_t)_{t\in\mathbb{Z}} \sim WN(0, \sigma^2)$ and $(\psi_j)_{j\in\mathbb{Z}}$ a sequence in $\mathbb{C}$ with $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$ and define $Y_t = \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}$ for $t \in \mathbb{Z}$. Then $(Y_t)_{t\in\mathbb{Z}}$ is called a *two-sided moving average process of order infinity with coefficients ψ_j*. If additionally $\psi_j = 0$ for all $j < 0$, then $Y_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}$ is called a one-sided *$MA(\infty)$-process*, or *one sided moving average process of infinite order.*

The process of building a moving average process from a white noise process is often called 'applying a filter'. The precise definition is as follows:

Definition 5.6. An absolutely summable sequence $(\psi_j)_{j\in\mathbb{Z}}$ of complex numbers is called a *linear filter* (more precisely, one should say an *l^1-linear filter*, since one can still relax the condition of absolute summability a bit, but we shall not do this here). If

$$Y_t = \sum_{j=-\infty}^{\infty} \psi_j X_{t-j}, \quad t \in \mathbb{Z},$$

for two time series $X = (X_t)_{t\in\mathbb{Z}}$ and $Y = (Y_t)_{t\in\mathbb{Z}}$, we say that the time series Y *arises from X through application of the linear filter* $(\psi_j)_{j\in\mathbb{Z}}$.

So a moving average process (two sided, one-sided, of infinite or of finite order) arises from a white noise through application of a linear filter. For example, if $Z = (Z_t)_{t\in\mathbb{Z}} \sim WN(0, \sigma^2)$, then the MA($q$)-process $Y_t = Z_t + \theta_1 Z_{t-1} + \ldots + \theta_q Z_{t-q}$ arises from Z by application of the linear filter $(\psi_j)_{j\in\mathbb{Z}}$ with $\psi_0 := 1$, $\psi_j := \theta_j$ for $j \in \{1, \ldots, q\}$ and $\psi_j := 0$ otherwise.

It will be convenient in many aspects to describe linear filters by their associated filter functions:

Definition 5.7. (a) Denote by S^1 the unit circle in the complex plane, i.e.

$$S^1 := \{z \in \mathbb{C} : |z| = 1\}.$$

(b) Let $(\psi_j)_{j\in\mathbb{Z}}$ be a linear filter. Then the *associated filter function* ψ is defined by

$$\psi : S^1 \to \mathbb{C}, \quad z \mapsto \sum_{j=-\infty}^{\infty} \psi_j z^j.$$

Remark 5.8. (a) Since $\sum_{j=-\infty}^{\infty} |\psi_j z^j| = \sum_{j=-\infty}^{\infty} |\psi_j| < \infty$ for all $z \in S^1$, the sum defining $\psi(z)$ converges absolutely, so that ψ is indeed a function into the complex plane.
(b) Identifying the unit circle S^1 with $[0, 2\pi]$ via $[0, 2\pi] \ni s \mapsto \mathrm{e}^{\mathrm{i}s} \in S^1$ (well, actually we should identify it formally with $[0, 2\pi)$, but $\mathrm{e}^{\mathrm{i}2\pi} = \mathrm{e}^{\mathrm{i}0} = 1$, so we are working with 2π-periodic functions), and setting $z = \mathrm{e}^{\mathrm{i}s}$, we have

$$\psi(z) = \psi(\mathrm{e}^{\mathrm{i}s}) = \sum_{n=-\infty}^{\infty} \psi_n \left(\mathrm{e}^{\mathrm{i}s}\right)^n = \sum_{n=-\infty}^{\infty} \psi_n \mathrm{e}^{ins} =: \widetilde{\psi}(s).$$

By the Weierstraß convergence theorem, the convergence of the right-hand side is uniformly and hence $[0, 2\pi] \ni s \mapsto \widetilde{\psi}(s) := \sum_{n=-\infty}^{\infty} \psi_n \mathrm{e}^{ins}$ defines a continuous function on $[0, 2\pi]$ with $\widetilde{\psi}(2\pi) = \widetilde{\psi}(0)$. Hence also $\psi : S^1 \to \mathbb{C}$ must be continuous. So every filter function must necessarily be continuous.
(c) The set of all filter functions is the set of all continuous functions with absolutely summable Fourier coefficients; in Fourier analysis it is called the *Wiener algebra.*

One immediate question is whether a filter function determines uniquely its linear filter, or in the language of Fourier analysis, are the Fourier coefficients unique? Indeed, they are:

Theorem 5.9. *Let $(\psi_j)_{j\in\mathbb{Z}}$ and $(\varphi_j)_{j\in\mathbb{Z}}$ be two linear filters. Assume that their associated filter functions are equal, i.e. that*

$$\psi(z) = \varphi(z) \quad \forall\, z \in S^1.$$

Then $\psi_j = \varphi_j$ for all $j \in \mathbb{Z}$.

Proof. Let $\widetilde{\psi}(s)$ be defined as in Remark 5.8 (b). Since

$$\int_0^{2\pi} \mathrm{e}^{\mathrm{i}hs}\,\mathrm{d}s = \int_0^{2\pi} \cos(hs)\,\mathrm{d}s + \mathrm{i}\int_0^{2\pi} \sin(hs)\,\mathrm{d}s = \begin{cases} 0, & h \in \mathbb{Z} \setminus \{0\}, \\ 2\pi, & h = 0, \end{cases}$$

we conclude from Lebesgue's dominated convergence theorem that for $m \in \mathbb{Z}$

$$\int_0^{2\pi} \widetilde{\psi}(s)\mathrm{e}^{-\mathrm{i}ms}\,\mathrm{d}s = \int_0^{2\pi} \sum_{n=-\infty}^{\infty} \psi_n \mathrm{e}^{\mathrm{i}ns}\,\mathrm{e}^{-\mathrm{i}ms}\,\mathrm{d}s = \sum_{n=-\infty}^{\infty} \psi_n \int_0^{2\pi} \mathrm{e}^{\mathrm{i}(n-m)s}\,\mathrm{d}s = 2\pi\psi_m.$$

Since $\widetilde{\psi} = \widetilde{\varphi}$ we obtain $\psi_m = \varphi_m$ for all $m \in \mathbb{Z}$. □

Next, we introduce the notion of the backshift operator.

Definition 5.10. For a time series $X = (X_t)_{t\in\mathbb{Z}}$ on a probability space $(\Omega, \mathcal{F}, P)$, let

$$BX_t := X_{t-1},\ t \in \mathbb{Z},$$

i.e. $(BX_t)(\omega) := X_{t-1}(\omega)$ for all $\omega \in \Omega$. We call B the *backshift operator* (also: *lag operator*).

Actually, it would be more precise to write $(BX)_t = X_{t-1}$, in which case B would be seen as an operator acting on the class of two sided complex sequences $z = (z_n)_{n\in\mathbb{Z}}$ mapping $(z_n)_{n\in\mathbb{Z}}$ to $(z_{n-1})_{n\in\mathbb{Z}}$. Then the t'th entry $(Bz)_t$ of Bz is given by z_{t-1}. It is however custom to write immediately $BX_t = X_{t-1}$. It is clear that applying this operator k times should transform X_t to X_{t-k}, and the inverse B^{-1} should transform X_t to X_{t+1}. Hence the following definition makes sense:

Definition 5.11. Let $X = (X_t)_{t\in\mathbb{Z}}$ be a complex valued time-series. Then define for all $k, t \in \mathbb{Z}$

$$B^k X_t = X_{t-k}.$$

Especially, we have $BBX_t = B^2X_t = X_{t-2}$, $BB^{-1}X_t = X_t, \ldots$ The operator B^{-1} is also called the *forwardshift operator.*

Now let $(\psi_j)_{j\in\mathbb{Z}}$ be a linear filter and $X = (X_t)_{t\in\mathbb{Z}}$ a time series. Assume that $Y_t := \sum_{j=-\infty}^{\infty} \psi_j X_{t-j}$ converges in probability for each $t \in \mathbb{Z}$ (satisfied e.g. if X is weakly stationary) so that $Y = (Y_t)_{t\in\mathbb{Z}}$ arises from X through application of the linear filter $(\psi_j)_{j\in\mathbb{Z}}$. Then

$$Y_t = \sum_{j=-\infty}^{\infty} \psi_j X_{t-j} = \sum_{j=-\infty}^{\infty} \psi_j B^j X_t.$$

On the other hand, recall the filter function associated with $(\psi_j)_{j\in\mathbb{Z}}$ was given by $\psi(z) = \sum_{j=-\infty}^{\infty} \psi_j z^j$. Thus, *formally* substituting B for z should lead to

$$Y_t = \sum_{j=-\infty}^{\infty} \psi_j B^j X_t = \left(\sum_{j=-\infty}^{\infty} \psi_j z^j \right)_{|z=B} X_t = \psi(z)_{|z=B} X_t = \psi(B)X_t.$$

This was a formal deviation of what $\psi(B)X_t$ should be, and we take it now as a definition:

Definition 5.12. Let $(\psi_j)_{j\in\mathbb{Z}}$ be a linear filter with associated filter function $\psi : S^1 \to \mathbb{C}$. Let $X = (X_t)_{t\in\mathbb{Z}}$ be a time series such that $Y_t := \sum_{j=-\infty}^{\infty} \psi_j X_{t-j}$ converges in probability for every $t \in \mathbb{Z}$ (satisfied for example if X is weakly stationary), so that the time series $Y = (Y_t)_{t\in\mathbb{Z}}$ arises from X through application of the linear filter $(\psi_j)_{j\in\mathbb{Z}}$. Then we define

$$\psi(B)X_t := Y_t = \sum_{j=-\infty}^{\infty} \psi_j X_{t-j}, \quad t \in \mathbb{Z}.$$

Remark 5.13. The definition of $\psi(B)$ is unambiguous, since if $(\varphi_j)_{j\in\mathbb{Z}}$ is another linear filter with the same filter function $\varphi(z) = \psi(z)$, then $\varphi_j = \psi_j$ for all $j \in \mathbb{Z}$ by Theorem 5.9.

So far, these have only been definitions. But the use of filter functions and the back shift operator has some nice features when applying two filters consecutively:

Theorem 5.14. *Let $(\psi_j)_{j\in\mathbb{Z}}$ and $(\varphi_j)_{j\in\mathbb{Z}}$ be two linear filters with filter functions $\psi(z) = \sum_{j=-\infty}^{\infty} \psi_j z^j$ and $\varphi(z) = \sum_{j=-\infty}^{\infty} \varphi_j z^j$. For $m \in \mathbb{Z}$ let*

$$\theta_m := \sum_{j+k=m} \psi_j \varphi_k = \sum_{j\in\mathbb{Z}} \psi_j \varphi_{m-j} = \sum_{k\in\mathbb{Z}} \varphi_k \psi_{m-k}.$$

Then $\sum_{m=-\infty}^{\infty} |\theta_m| < \infty$, hence $(\theta_m)_{m\in\mathbb{Z}}$ is a linear filter. Define the filter function

$$\theta : S^1 \to \mathbb{C}, \quad z \mapsto \theta(z) := \sum_{m=-\infty}^{\infty} \theta_m z^m.$$

Then

$$\psi(z)\varphi(z) \;=\; \sum_{j\in\mathbb{Z}} \psi_j z^j \sum_{k\in\mathbb{Z}} \varphi_k z^k = \sum_{m\in\mathbb{Z}} \left(\sum_{j+k=m} \psi_j \varphi_k \right) z^m = \sum_{m\in\mathbb{Z}} \theta_m z^m = \theta(z).$$

In addition, if $X = (X_t)_{t\in\mathbb{Z}}$ is a weakly stationary time series, then

$$\psi(B)\varphi(B)X_t = \varphi(B)\psi(B)X_t = \theta(B)X_t.$$

Hence, applying two linear filters corresponds to applying one linear filter, where the filter function of this is the product of the two original filter functions.

Proof. The assertions regarding the product of the filter function are elementary analysis using the Cauchy product of infinite sums. The assertion regarding the application to X follows similarly by rearranging the double sums. For example, setting $Y_t := \psi(B)X_t$,

$$\varphi(B)\psi(B)X_t = \varphi(B)Y_t = \sum_{k\in\mathbb{Z}} \varphi_k Y_{t-k} = \sum_{k\in\mathbb{Z}} \varphi_k \sum_{j\in\mathbb{Z}} \psi_j X_{t-k-j} = \sum_{k\in\mathbb{Z}} \sum_{j\in\mathbb{Z}} \varphi_k \psi_j X_{t-k-j}; \tag{5.3}$$

here, the double series converges almost surely absolutely, since the expectation of the numerical random variable $W = \sum_{k\in\mathbb{Z}} \sum_{j\in\mathbb{Z}} |\varphi_k \psi_k Z_{t-k-j}| : \Omega \to [0,\infty]$ is given by

$$\mathbb{E}(W) = \sum_{k\in\mathbb{Z}} \sum_{j\in\mathbb{Z}} |\varphi_k|\,|\psi_j| \underbrace{\mathbb{E}|X_{t-k-j}|}_{\leq \|X_{t-k-j}\|_2 = \|X_0\|_2 < \infty} \leq (\mathbb{E}|X_0|^2)^{1/2} \sum_{k=-\infty}^{\infty} |\varphi_k| \sum_{j=-\infty}^{\infty} |\psi_j| < \infty,$$

so that W is finite almost surely. Using the substitution $k + j = m$, the right-hand side of (5.3) can be rearranged to

$$\sum_{m\in\mathbb{Z}} \sum_{k,j\in\mathbb{Z}:k+j=m} \varphi_k \psi_j X_{t-m} = \sum_{m\in\mathbb{Z}} \theta_m X_{t-m}$$

by the Cauchy product for infinite sums, since the double sum converges absolutely with probability one. So we have proved that

$$\varphi(B)\psi(B)X_t = \theta(B)X_t = (\varphi\psi)(B)X_t.$$

That $\psi(B)\varphi(B)X_t = \theta(B)X_t$ follows similarly. □

Example 5.15. Let $Z = (Z_t)$ be weakly stationary, and let

$$X_t := \sum_{j=0}^{q} \psi_j Z_{t-j}, \quad Y_t := \sum_{k=0}^{p} \varphi_k X_{t-k}.$$

To represent Y in terms of Z, define the filter functions

$$\psi(z) = \sum_{j=0}^{q} \psi_j z^j, \quad \varphi(z) := \sum_{j=0}^{p} \varphi_j z^j,$$

that are polynomials. Calculate their product $\theta(z) := \varphi(z)\psi(z) = \sum_{m=0}^{p+q} \theta_m z^m$. Then $Y_t = \sum_{m=0}^{p+q} \theta_m Z_{t-m}$.

As an application of the previous results, assume that we would like to solve an equation of the form $\varphi(B)X_t = \theta(B)Z_t$, where $Z = (Z_t)_{t\in\mathbb{Z}}$ is white noise, and φ and θ are given filter functions, and we are looking for a weakly stationary solution $X = (X_t)_{t\in\mathbb{Z}}$ of this equation. Assuming that such a solution (if it exists) can be represented as a linear filter $(\psi_j)_{j\in\mathbb{Z}}$ applied to Z, i.e. $X_t = \psi(B)Z_t$, it must fulfill

$$(\varphi\psi)(B)Z_t = \varphi(B)\psi(B)Z_t = \varphi(B)X_t = \theta(B)Z_t \quad \forall\, t \in \mathbb{Z}.$$

Denoting the filter to which the filter function $\varphi\psi$ is associated to by $(\xi_j)_{j\in\mathbb{Z}}$, so that $\sum_{j=-\infty}^{\infty} \xi_j z^j = \varphi(z)\psi(z)$, we have

$$\sum_{j=-\infty}^{\infty} \xi_j Z_{t-j} = \sum_{j=-\infty}^{\infty} \theta_j Z_{t-j} \quad \forall\, t \in \mathbb{Z}. \tag{5.4}$$

Multiplying the above equation by $\overline{Z_0}$, we have from the L^2-convergence of the sum that

$$\mathbb{E}\left(\overline{Z_0}\sum_{j=-\infty}^{\infty}\xi_j Z_{t-j}\right) = \sum_{j=-\infty}^{\infty}\xi_j \mathbb{E}(\overline{Z_0}Z_{t-j}) = \xi_t\sigma^2,$$

since $Z \sim WN(0,\sigma^2)$. The same holds for the right hand side of (5.4) multiplied by $\overline{Z_0}$, leading to

$$\sigma^2\xi_t = \sigma^2\theta_t \quad \forall\, t \in \mathbb{Z}.$$

Since $\sigma^2 > 0$, this gives $\xi_j = \theta_j$ for all $j \in \mathbb{Z}$ and hence $\varphi(z)\psi(z) = \theta(z)$ for all $z \in S^1$. Let us summarise these findings in the following corollary:

Corollary 5.16. *Let $Z = (Z_t)_{t\in\mathbb{Z}} \sim WN(0,\sigma^2)$ with $\sigma^2 > 0$, and let $(\varphi_j)_{j\in\mathbb{Z}}$ and $(\theta_j)_{j\in\mathbb{Z}}$ be two linear filters. If there exists a filter $\psi = (\psi_j)_{j\in\mathbb{Z}}$ such that the process $X = (X_t)_{t\in\mathbb{Z}}$ defined by $\psi(B)Z_t$ satisfies the equation*

$$\varphi(B)X_t = \theta(B)Z_t, \quad t \in \mathbb{Z},$$

then the filter functions are related by

$$\varphi(z)\psi(z) = \theta(z) \quad \forall\, z \in S^1.$$

Corollary 5.16 will be important when treating solutions of ARMA equations in the next chapter.

Chapter 6

ARMA processes

In this chapter we treat the important class of ARMA processes. They can be regarded as the most fundamental class of linear time series, since they are easy to describe on the one hand, and on the other hand still cover a wide range of processes. Using spectral theory (what we shall not do in this course) one can actually show that the class of ARMA processes is dense in a certain sense in the set of all stationary processes, meaning that every stationary process can be approximated arbitrarily well (in a certain sense) by ARMA processes. This underlines the importance of ARMA proceses.

6.1 Definition, AR(1) and ARMA(1,1) process

Let us start with the definition of an ARMA process:

Definition 6.1. Let $p, q \in \mathbb{N}_0$.
(a) A time series $(X_t)_{t\in\mathbb{Z}}$ is called an *autoregressive moving average process or orders p and q*, in short an *ARMA(p,q)-process*, if $(X_t)_{t\in\mathbb{Z}}$ is weakly stationary and there exist coefficients $\varphi_1, \ldots, \varphi_p \in \mathbb{C}$, $\theta_1, \ldots, \theta_q \in \mathbb{C}$ and a complex-valued white noise $(Z_t)_{t\in\mathbb{Z}}$ such that

$$X_t - \varphi_1 X_{t-1} - \ldots - \varphi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \ldots + \theta_q Z_{t-q}. \tag{6.1}$$

$\theta_1, \ldots, \theta_q$ are called *moving-average-coefficients* and $\varphi_1, \ldots, \varphi_p$ *autoregressive coefficients.* Sometimes we say that $(X_t)_{t\in\mathbb{Z}}$ is an ARMA process with respect to the white noise $(Z_t)_{t\in\mathbb{Z}}$. Equation (6.1) is called an *ARMA(p,q)*-equation or simply an *ARMA* equation.
(b) A time series $(X_t)_{t\in\mathbb{Z}}$ is called *ARMA(p,q) process with mean* $\mu \in \mathbb{C}$, if $(X_t - \mu)_{t\in\mathbb{Z}}$ is an ARMA(p,q) process.

Remark 6.2. (a) The definition in (b) needs some clarification. Is this definition unambiguous, i.e. does an ARMA process with mean μ really have expectation μ? This is equivalent to the fact that every ARMA process has mean 0. Is this so? To investigate this question, let $(X_t)_{t\in\mathbb{Z}}$ be an ARMA(p,q) process and $1 - \varphi_1 - \ldots - \varphi_p \neq 0$, then $\mathbb{E}X_t = 0$, as we conclude from (6.1) that

$$(1-\varphi_1-\dots-\varphi_p)\mathbb{E}X_t = \mathbb{E}Z_t(1+\theta_1+\ldots+\theta_q) = 0 \implies \mathbb{E}X_t = 0.$$

So if X is an ARMA process and satisfies $1-\varphi_1-\ldots-\varphi_p \neq 0$, then an ARMA process has mean 0 and an ARMA process with mean μ indeed has expectation μ. This does not need to be true any longer if $1-\varphi_1-\ldots-\varphi_p = 0$. To see this, let $Z=(Z_t)_{t\in\mathbb{Z}} \sim WN(0,\sigma^2)$ and define $X_t := Z_t + 1$. Then

$$X_t - X_{t-1} = Z_t - Z_{t-1},$$

showing that $X = (X_t)_{t\in\mathbb{Z}}$ is an ARMA process, although it has expectation 1. Since $(X_t-\mu)_{t\in\mathbb{Z}}$ is an ARMA process for every μ, by definition we can say that X is an ARMA process with mean μ for any given μ, although the expectation of X is well defined and is equal to 1. So, the case when $1-\varphi_1-\ldots-\varphi_p = 0$ is the only case when the definition in (b) is not good. This will not cause a problem to us, since in most cases we will assume anyway that $1-\varphi_1-\ldots-\varphi_p \neq 0$.
(b) In the example above we have also seen that there is no unique weakly stationary solution of the ARMA(1,1) equation $X_t - X_{t-1} = Z_t - Z_{t-1}$. The deeper reason for that is that the autoregressive polynomial and moving average polynomial (to be defined in Definition 6.3) share a common root on the unit circle S^1. We shall not go into details but often exclude such cases by assuming that the moving average polynomial and autoregressive polynomial do not share a common zero (at least not on S^1).

With the aid of the backshift operator defined in Definitions 5.10, 5.11 and 5.12 we can rewrite the ARMA equation in a nicer way.

Definition 6.3. Consider the ARMA(p,q)-equation (6.1). We define for the *MA-coefficents (moving average coefficients)* $\theta_1,\ldots,\theta_q$ the *moving average polynomial* of the ARMA equation (6.1) by

$$\theta(z) := 1+\theta_1 z+\ldots+\theta_q z^q,\ z\in\mathbb{C}.$$

For the *AR coefficients (autoregressive coefficients)* $\varphi_1,\ldots,\varphi_p$ the polynomial

$$\varphi(z) := 1-\varphi_1 z-\ldots-\varphi_p z^p,\ z\in\mathbb{C}$$

is called the *autoregressive polynomial* of the ARMA equation (6.1), and both θ and φ are called *characteristic polynomials.* If we use the backshift operator B, then (6.1) can be rewritten as

$$\varphi(B)X_t = \theta(B)Z_t,\ t\in\mathbb{Z}. \tag{6.2}$$

An ARMA$(0,q)$-process is a weakly stationary process $(X_t)_{t\in\mathbb{Z}}$ of the form

$$X_t = Z_t+\theta_1 Z_{t-1}+\ldots+\theta_q Z_{t-q} = \theta(B)Z_t,$$

where $(Z_t)_{t\in\mathbb{Z}} \sim WN(0,\sigma^2)$. This is nothing else then the MA(q)-process we defined in Example 2.8. So an ARMA$(0,q)$-process is nothing else than an MA(q)-process. Similarly, when the other parameter is zero, i.e. when $p\in\mathbb{N}$ but $q=0$, we will speak of an autoregressive process or AR-process:

Definition 6.4. An ARMA$(p, 0)$ process is also called an *AR(p) process*, or an *autoregressive process of order p.*

An important question will be when the ARMA(p, q)-equations actually have a weakly stationary solution. Let us treat this first in the simple case of the AR(1) process.

Theorem 6.5 (The AR(1) process)**.** *Let $(Z_t)_{t\in\mathbb{Z}} \sim WN(0, \sigma^2)$ with $\sigma^2 \neq 0$ and $\varphi_1 \in \mathbb{C}$. Consider the AR(1) equation*

$$X_t - \varphi_1 X_{t-1} = Z_t,\ t \in \mathbb{Z}. \tag{6.3}$$

Then (6.3) *admits a weakly stationary solution if and only if $|\varphi_1| \neq 1$. If $|\varphi_1| \neq 1$, then the weakly stationary solution is unique and when $|\varphi_1| < 1$ is given by*

$$X_t := \sum_{k=0}^{\infty} \varphi_1^k Z_{t-k}, \quad t \in \mathbb{Z},$$

(the sum converging almost surely absolutely and in L^2), while if $|\varphi_1| > 1$ it is given by

$$X_t = -\sum_{j=1}^{\infty} \varphi_1^{-j} Z_{t+j}, \quad t \in \mathbb{Z}$$

(the sum converging almost surely absolutely and in L^2). When $|\varphi_1| < 1$, the ACVF of X is given by

$$\gamma_X(h) = \begin{cases} \frac{\sigma^2}{1-|\varphi_1|^2}\varphi_1^h, & h \in \mathbb{N}_0, \\ \frac{\sigma^2}{1-|\varphi_1|^2}\overline{\varphi_1}^{|h|}, & -h \in \mathbb{N}, \end{cases}$$

while for $|\varphi_1| > 1$ it is given by

$$\gamma_X(h) = \begin{cases} \frac{\sigma^2}{|\varphi_1|^2-1}\overline{\varphi_1}^{-h}, & h \in \mathbb{N}_0, \\ \frac{\sigma^2}{|\varphi_1|^2-1}\varphi_1{}^h, & -h \in \mathbb{N}. \end{cases}$$

Proof. (a) Consider first the case that $|\varphi_1| < 1$. To show the existence and form of the weakly stationary solution, define X_t by $X_t := \sum_{k=0}^{\infty} \varphi_1^k Z_{t-k}$, $t \in \mathbb{Z}$. Since $\sum_{j=0}^{\infty} |\varphi_1|^j = \frac{1}{1-|\varphi_1|} < \infty$ as a consequence of $|\varphi_1| < 1$, the sequence $(\varphi_1^j)_{j\in\mathbb{N}_0}$ is indeed a linear filter (define the filter coefficients as 0 for negative indices). Then X is weakly stationary by Corollary 5.4 with ACVF given for $h \in \mathbb{N}_0$ by

$$\gamma_X(h) = \sigma^2 \sum_{j=h}^{\infty} \varphi_1^j \overline{\varphi_1^{j-h}} = \sigma^2 \varphi_1^h \sum_{j=0}^{\infty} |\varphi_1|^{2j} = \frac{\sigma^2}{1-|\varphi_1|^2}\varphi_1^h,$$

and for $h < 0$ we use $\gamma_X(h) = \overline{\gamma_X(-h)}$. To see that X thus defined satisfies indeed the AR(1) equation observe that

$$\begin{aligned} X_t - \varphi_1 X_{t-1} &= \sum_{k=0}^{\infty} \varphi_1^k Z_{t-k} - \varphi_1 \sum_{k=0}^{\infty} \varphi_1^k Z_{t-1-k} \\ &= Z_{t-0} + \sum_{k=1}^{\infty} (\varphi_1^k - \varphi_1^k) Z_{t-k} = Z_t, \end{aligned}$$

so that X is indeed a solution of (6.3). To see the uniqueness of the solution, suppose that $X = (X_t)_{t\in\mathbb{Z}}$ is some weakly stationary solution of (6.3), so

$$X_t = Z_t + \varphi_1 X_{t-1} \quad \forall\, t \in \mathbb{Z}.$$

Iterating this again and again, we obtain

$$\begin{aligned} X_t &= Z_t + \varphi_1 X_{t-1} \\ &= Z_t + \varphi_1(Z_{t-1} + \varphi_1 X_{t-2}) \\ &= Z_t + \varphi_1 Z_{t-1} + \varphi_1^2 X_{t-2} \\ &= Z_t + \varphi_1 Z_t + \varphi_1^2(Z_{t-2} + \varphi_1 X_{t-3}) \\ &= Z_t + \varphi_1 Z_{t-1} + \varphi_1^2 Z_{t-2} + \varphi_1^3 X_{t-3} \\ &= \ldots \\ &= \sum_{k=0}^{N} \varphi_1^k Z_{t-k} + \varphi_1^{N+1} X_{t-N-1} \qquad (6.4) \end{aligned}$$

for any $N \in \mathbb{N}$. Letting $N \to \infty$ we see that $\sum_{k=0}^{N} \varphi_1^k Z_{t-k}$ converges (in mean square and almost surely) to $\sum_{k=0}^{\infty} \varphi_1^k Z_{t-k}$, and that $\varphi_1^{N+1} X_{t-N-1}$ converges in mean square to 0, since

$$\mathbb{E}\left|\varphi_1^{N+1} X_{t-N-1}\right|^2 = \underbrace{|\varphi_1|^{2(N+1)}}_{\to 0 \text{ since } |\varphi_1|<1} \cdot \underbrace{\mathbb{E}|X_{t-N-1}|^2.}_{\text{bounded since } X \text{ weakly stat.}}$$

It follows that $X_t = \sum_{k=0}^{\infty} \varphi_1^k Z_{t-k}$, in particular there is only one solution.

(b) When $|\varphi_1| > 1$, we can rewrite (6.3) as

$$X_{t-1} = -\frac{1}{\varphi_1} Z_t + \frac{1}{\varphi_1} X_t.$$

Iterating this we obtain

$$\begin{aligned} X_t &= -\frac{1}{\varphi_1} Z_{t+1} + \frac{1}{\varphi_1} X_{t+1} \\ &= -\frac{1}{\varphi_1} Z_{t+1} - \frac{1}{\varphi_1^2} Z_{t+2} + \frac{1}{\varphi_1^2} X_{t+2} \\ &= \ldots \\ &= -\sum_{k=1}^{N} \varphi_1^{-k} Z_{t+k} + \frac{1}{\varphi_1^N} X_{t+N}. \end{aligned}$$

Letting $N \to \infty$ this converges to $\sum_{k=1}^{\infty} \varphi_1^{-k} Z_{t+k}$, giving uniqueness (observe that $|\varphi_1^{-1}| < 1$) and a candidate for the solution. Now define $X_t := \sum_{k=1}^{\infty} \varphi_1^{-k} Z_{t+k}$. This is weakly stationary by Corollary 5.4, and similarly as in (a) it is checked that it defines indeed a solution of (6.3). Again by Corollary 5.4, its ACVF is given for $h \in \mathbb{N}_0$ by

$$\begin{aligned} \gamma_X(h) &= \sigma^2 \sum_{j<0, j-h<0} \varphi_1^j \overline{\varphi_1}^{j-h} = \sigma^2 \overline{\varphi_1}^{-h} \sum_{j<0} |\varphi_1|^{2j} = \sigma^2 \overline{\varphi_1}^{-h} |\varphi_1|^{-2} \sum_{j=0}^{\infty} \frac{1}{|\varphi_1|^{2j}} \\ &= \frac{\sigma^2 |\varphi_1|^{-2}}{1 - |\varphi_1|^2} \overline{\varphi_1}^{-h} = \frac{\sigma^2}{|\varphi_1|^2 - 1} \overline{\varphi_1}^{-h} \end{aligned}$$

and for $h \in -\mathbb{N}_0$ we use $\gamma_X(h) = \overline{\gamma_X(-h)}$.

(c) When $|\varphi_1| = 1$, assume that $X = (X_t)_{t\in\mathbb{Z}}$ is a weakly stationary solution of (6.3). We then obtain from (6.4) that

$$X_t - \varphi_1^{N+1} X_{t-N-1} = \sum_{k=0}^{N} \varphi_1^k Z_{t-k}.$$

Taking the variance of this equation we obtain

$$\mathrm{Var}\,\left(X_t - \varphi^{N+1} X_{t-N-1}\right) = \mathrm{Var}\,\left(\sum_{k=0}^{N} \varphi_1^k Z_{t-k}\right) = \sum_{k=0}^{N} \underbrace{|\varphi_1|^{2k}}_{=1} \cdot \underbrace{\mathrm{Var}\,(Z_{t-k})}_{=\sigma^2} \to \infty, \quad N \to \infty,$$

where we used that the variance of an uncorrelated sum is the sum of the variances. But now if $(X_t)_{t\in\mathbb{Z}}$ is weakly stationary, then

$$\mathrm{Var}\,(X_t - \varphi_1^{N+1} X_{t-N-1}) = \mathrm{Var}\,(X_t) + \mathrm{Var}\,(\varphi_1^{N+1} X_{t-N-1}) - 2\Re\mathrm{Cov}\,(X_t, \varphi_1^{N+1} X_{t-N-1})$$

(here $\Re z$ denotes the real part of a complex number z) which by stationarity of X, the Cauchy-Schwarz inequality and the fact that $|\varphi_1| = 1$ is bounded in N, hence a contradiction. We see that no weakly stationary solution can exist when $|\varphi_1| = 1$. □

Remark 6.6. (a) The solution $X_t = -\sum_{k=1}^{\infty} \varphi_1^{-k} Z_{t+k}$ when $|\varphi_1| > 1$ is somewhat unnatural since it depends on future values of the noise. Much more natural is the solution $X_t = \sum_{k=0}^{\infty} \varphi_1^k Z_{t-k}$ when $|\varphi_1| < 1$, which depends only on past values of the noise. A solution which depends only on past values of the noise is called a *causal solution.*
(b) Using the characteristic polynomial $\varphi(z) = 1 - \varphi_1 z$, the condition $|\varphi_1| \neq 1$ is equivalent to $\varphi(z) \neq 0$ for all $z \in S^1$. Further, the condition $|\varphi_1| < 1$ is equivalent to $\varphi(z) \neq 0$ for all $\{z \in \mathbb{C} : |z| \leq 1\}$, i.e. to the fact that φ does not have zeroes in the closed unit ball. We shall extend these conditions later to general ARMA equations.

Before we treat general ARMA(p, q) processes we first have a look at the special case of the ARMA(1,1) process.

Theorem 6.7 (The ARMA(1,1) process). *Let $Z = (Z_t)_{t\in\mathbb{Z}} \sim WN(0, \sigma^2)$ with $\sigma^2 > 0$ and let $\varphi_1, \theta_1 \in \mathbb{C}$. Consider the ARMA(1,1) equation*

$$X_t - \varphi_1 X_{t-1} = Z_t + \theta_1 Z_{t-1}. \tag{6.5}$$

(a) If $|\varphi_1| \neq 1$, there exists a unique stationary solution X to the ARMA(1,1) equation (6.5)*, which for $|\varphi_1| < 1$ is given by $X_t = Z_t + (\varphi_1 + \theta_1) \sum_{j=1}^{\infty} \varphi_1^{j-1} Z_{t-j}$, while if $|\varphi_1| > 1$ it is given by $X_t = -\frac{\theta_1}{\varphi_1} Z_t - \sum_{j=1}^{\infty} (\varphi_1 + \theta_1) \varphi_1^{-j-1} Z_{t+j}$.*
(b) Suppose that $|\varphi_1| = 1$. Then the ARMA(1,1) equation (6.5) *admits a stationary solution if and only if $\varphi_1 = -\theta_1$, in which case one such solution is given by $X_t = Z_t$. If the underlying probability space $(\Omega, \mathcal{F}, P)$ on which $(Z_t)_{t\in\mathbb{Z}}$ is defined is rich enough to support a Bernoulli distributed random variable U with parameter* 1/2 *that is independent of $(Z_t)_{t\in\mathbb{Z}}$, then also $X_t := Z_t + \varphi_1^t (2U - 1)$ is a stationary solution of* (6.5)*, in particular, the stationary solution is not unique.*

Proof. Much of the proof can be done in a similar way as the proof of Theorem 6.5, but we give another one using the characteristic polynomials, which paves the way for the more general case later. The characteristic polynomials are given by

$$\varphi(z) = 1 - \varphi_1 z \quad \text{and} \quad \theta(z) = 1 - \theta_1 z,$$

so that the ARMA(1,1) equation reads

$$\varphi(B)X_t = \theta(B)Z_t.$$

(a) (i) Let $|\varphi_1| < 1$. To show uniqueness, let $X = (X_t)_{t\in\mathbb{Z}}$ be a stationary solution. We would like to solve this equation and apply the filter function $S^1 \ni z \mapsto \frac{1}{\varphi(z)}$ to it. But is that actually a filter function, i.e. does there exist a liner filter $(\psi_j)_{j\in\mathbb{Z}}$ such that $\frac{1}{\varphi(z)} = \sum_{j=0}^{\infty} \psi_j z^j$ for all $z \in S^1$? The answer here is yes, and by using the geometric series we obtain that

$$\frac{1}{\varphi(z)} = \frac{1}{1-\varphi_1 z} = \sum_{j=0}^{\infty} \varphi_1^j z^j,$$

which converges absolutely for $z \in \mathbb{C}$ with $|z| < 1/|\varphi_1|$, in particular for $z \in S^1$. So now if X is a stationary solution of $\varphi(B)X_t = \theta(B)Z_t$, we can apply the linear filter described by the filter function $S^1 \ni z \mapsto \frac{1}{\varphi(z)}$ on both sides of the equation and obtain from Theorem 5.14

$$X_t = B^0 X_t = \left(\frac{1}{\varphi}\varphi\right)(B)\, X_t = \frac{1}{\varphi}(B)(\underbrace{\varphi(B)X_t}_{\theta(B)Z_t}) = \frac{\theta}{\varphi}(B)Z_t.$$

The right hand side of this equation is uniquely determined from the characteristic polynomials and Z, so that the solution is unique if existent.

To see the existence, we define $X_t := \frac{\theta}{\varphi}(B)Z_t$. We have already seen that $S^1 \ni z \mapsto \frac{1}{\varphi(z)}$ is a filter function, and since $S^1 \ni z \mapsto \theta(z)$ is one and since the product of two filter functions is again a filter function by Theorem 5.14, so is $S^1 \ni z \mapsto \frac{\theta(z)}{\varphi(z)}$. This shows that $X = (X_t)_{t\in\mathbb{Z}}$ is weakly stationary, and from Theorem 5.14 we obtain that

$$\varphi(B)X_t = \varphi(B)\frac{\theta}{\varphi}(B)Z_t = \frac{\varphi\theta}{\varphi}(B)Z_t = \theta(B)Z_t,$$

so that X solves the ARMA(1,1) equation. Finally, observe that for $z \in S^1$

$$\frac{\theta(z)}{\varphi(z)} = (1 + \theta_1 z)\sum_{j=0}^{\infty} \varphi_1^j z^j = z^0 + \sum_{j=1}^{\infty} (\varphi_1 + \theta_1)\varphi_1^{j-1} z^j.$$

This shows that $X_t = \frac{\theta}{\varphi}(B)Z_t = Z_t + \sum_{j=1}^{\infty}(\varphi_1 + \theta_1)\varphi_1^{j-1} Z_{t-j}$.

(ii) Now let $|\varphi_1| > 1$. We would like again to use the same trick, but we must assure ourselves that $S^1 \mapsto \frac{1}{\varphi(z)} = \frac{1}{1-\varphi_1 z}$ is a filter function. Unfortunately, $|\varphi_1 z| > 1$ now for $z \in S^1$, but we can use the trick of the geometric series nevertheless via

$$\frac{1}{\varphi(z)} = \frac{1}{1-\varphi_1 z} = \frac{-1}{\varphi_1 z} \cdot \frac{1}{1 - \frac{1}{\varphi_1 z}} = \frac{-1}{\varphi_1 z}\sum_{j=0}^{\infty}\left(\frac{1}{\varphi_1 z}\right)^j = -\sum_{j=1}^{\infty} \varphi_1^{-j} z^{-j}.$$

This converges for $|\varphi_1 z| > 1$, in particular for $z \in S^1$, and so $\frac{1}{\varphi_1(z)}$ is a filter function also in this case. The uniqueness and existence then follows as in (a), with the unique stationary solution being given by $X_t = \frac{\theta}{\varphi}(B)Z_t$. For the specific form of X_t, we only have to observe that

$$\begin{aligned}\frac{\theta(z)}{\varphi(z)} &= -(1+\theta_1 z)\sum_{j=1}^{\infty}\varphi_1^{-j}z^{-j} = -\sum_{j=1}^{\infty}\varphi_1^{-j}z^{-j} - \sum_{j=0}^{\infty}\theta_1\varphi_1^{-j-1}z^{-j} \\ &= -\frac{\theta_1}{\varphi_1}z^0 - (\varphi_1+\theta_1)\sum_{j=1}^{\infty}\varphi_1^{-j-1}z^{-j},\end{aligned}$$

so that

$$X_t = -\frac{\theta_1}{\varphi_1}Z_t - \sum_{j=1}^{\infty}(\varphi_1+\theta_1)\varphi_1^{-j-1}Z_{t+j}.$$

(b) Now let $|\varphi_1| = 1$. Suppose that there exists a weakly stationary solution $X = (X_t)_{t\in\mathbb{Z}}$ of Equation (6.5). Denote $W_t := Z_t + \theta_1 Z_{t-1}$, so that $X_t - \varphi_1 X_{t-1} = W_t$ for each $N \in \mathbb{N}$ and $t \in \mathbb{Z}$. Iterating as in Equation (6.4) we obtain

$$\begin{aligned}X_t &= \varphi_1^{N+1}X_{t-N-1} + \sum_{k=0}^{N}\varphi_1^k \underbrace{W_{t-k}}_{=Z_{t-k}+\theta_1 Z_{t-k-1}} \\ &= \varphi_1^{N+1}X_{t-N-1} + \sum_{k=0}^{N}\varphi_1^k Z_{t-k} + \sum_{k=1}^{N+1}\varphi_1^{k-1}\theta_1 Z_{t-k} \\ &= \varphi_1^{N+1}X_{t-N-1} + \varphi_1^N\theta_1 Z_{t-N-1} + \sum_{k=1}^{N}\varphi_1^{k-1}(\varphi_1+\theta_1)Z_{t-k}.\end{aligned}$$

Rearranging yields

$$X_t - \varphi_1^{N+1}X_{t-N-1} - \varphi_1^N\theta_1 Z_{t-N-1} = (\varphi_1+\theta_1)\sum_{k=1}^{N-1}\varphi_1^k Z_{t-k}. \tag{6.6}$$

As in the proof of part (c) of Theorem 6.5 we see that the variance of the right-hand side of Equation (6.6) is given by $|\varphi_1+\theta_1|^2\sigma^2\sum_{k=1}^{N-1}|\varphi_1|^{2k} = |\varphi_1+\theta_1|^2\sigma^2(N-1)$, which is unbounded in N unless $\varphi_1+\theta_1 = 0$. On the other hand, the variance of the left hand side of (6.6) is bounded in N by an argument similar to the proof of part (c) of Theorem 6.5. Altogether we see that a necessary condition for a weakly stationary solution to exist when $|\varphi_1| = 1$ is that $\varphi_1+\theta_1 = 0$, i.e. that $\varphi_1 = -\theta_1$.

Now assume that $|\varphi_1| = 1$ and that $\varphi_1 = -\theta_1$, so that we are considering the ARMA(1,1) equation $X_t - \varphi_1 X_{t-1} = Z_t - \varphi_1 Z_{t-1}$. It is clear that $X_t := Z_t$ is a stationary solution of this equation. To see the non-uniqueness, suppose that there exists a $b(1,1/2)$-distributed random variable U (i.e. $P(U=1) = P(U=0) = 1/2$) such that U is independent of $(Z_t)_{t\in\mathbb{Z}}$. Define $V_t := \varphi_1^t(2U-1)$. Then $V_t - \varphi_1 V_{t-1} = 0$, $\mathbb{E}|V_t|^2 < \infty$,

$$\mathbb{E}V_t = \varphi_1^t\mathbb{E}(2U-1) = \varphi_1^t(1\cdot\underbrace{P(U=1)}_{=1/2}+(-1)\underbrace{P(U=0)}_{=1/2}) = 0$$

for all $t \in \mathbb{Z}$ and

$$\operatorname{Cov}(V_{t+h}, V_t) = \operatorname{Cov}(\varphi_1^{t+h}(2U-1), \varphi_1^t(2U-1) = \varphi_1^{t+h}\overline{\varphi_1}^t \operatorname{Var}(2U-1) = \varphi_1^h \operatorname{Var}(2U-1)$$

for all $h, t \in \mathbb{Z}$, so that $V = (V_t)_{t\in\mathbb{Z}}$ is weakly stationary. Since $\operatorname{Cov}(V_t, Z_s) = 0$ for all $t, s \in \mathbb{Z}$ by independence of U and Z, we see that also $(Z_t + \varphi_1^t(2U-1))_{t\in\mathbb{Z}}$ is weakly stationary and obviously a solution of (6.5). □

Remark 6.8. (a) In the setting of Theorem 6.7, consider the characteristic polynomials $\varphi(z) = 1 - \varphi_1 z$ and $\theta(z) = 1 + \theta_1 z$. The unique zero of φ is at $z = 1/\varphi_1$, the unique zero of θ is at $z = -1/\theta_1$. The condition $|\varphi_1| \neq 1$ can then be expressed by saying that φ has no zero on the unit circle, the condition $|\varphi_1| < 1$ by the fact that all zeroes (i.e. *the* zero) of φ lie outside the closed unit ball $\{z \in \mathbb{C} : |z| \leq 1\}$, and the condition $\varphi_1 \neq -\theta_1$ when $|\varphi_1| = 1$ means in particular that φ and θ have no common zero on S^1. In this language we shall think when generalising Theorem 6.7 to ARMA equations of higher order.
(b) As in Remark 6.6 (a), when $|\varphi_1| < 1$, equivalently when all zeroes of φ lie outside the closed unit ball in $\mathbb{C}$, then the unique stationary solution depends only on the past values of the noise, which will be called a causal solution later on. When $|\varphi_1| > 1$ and $\varphi_1 \neq -\theta_1$, the stationary solution $X_t = -\frac{\theta_1}{\varphi_1} Z_t - \sum_{j=1}^\infty (\varphi_1 + \theta_1)\varphi_1^{j-1} Z_{t+j}$ depends on future values of the noise, which is again strange. Also here, the case $\varphi_1 = -\theta_1$ needs some extra attention, which is why later we shall often assume that φ and θ do not share common zeroes.

In the proof of Theorem 6.7 we showed in particular that $S^1 \ni z \mapsto \frac{1}{1-\varphi_1 z}$ is a filter function when $|\varphi_1| \neq 0$. As we shall need this also in the general case, we state our findings in a separate lemma:

Lemma 6.9. *Let $\varphi_1 \in \mathbb{C}$ with $|\varphi_1| \neq 1$. Then $S^1 \ni z \mapsto \frac{1}{1-\varphi_1 z}$ is a filter function. More precisely, if $|\varphi_1| < 1$, then we have*

$$\frac{1}{1-\varphi_1 z} = \sum_{j=0}^\infty \varphi_1^j z^j \quad \forall\, z \in \mathbb{C} : |z| < 1/|\varphi_1|,$$

with the sum being absolutely convergent for each such z, so that $S^1 \ni z \mapsto \frac{1}{1-\varphi_1 z}$ is the filter function associated with the filter $(\psi_j)_{j\in\mathbb{Z}}$ defined by $\psi_j = \varphi_1^j$ for $j \in \mathbb{N}_0$ and $\psi_j = 0$ for $-j \in \mathbb{N}$. In the case $|\varphi_1| > 1$, we have

$$\frac{1}{1-\varphi_1 z} = -\sum_{j=1}^\infty \varphi_1^{-j} z^{-j} \quad \forall\, z \in \mathbb{C} : |z| > 1/|\varphi_1|,$$

with the sum being absolutely convergent for each such z, so that $S^1 \ni z \mapsto \frac{1}{1-\varphi_1 z}$ is the filter function associated with the filter $(\psi_j)_{j\in\mathbb{Z}}$ defined by $\psi_j = -\varphi_1^j$ for $-j \in \mathbb{N}$ and $\psi_j = 0$ for $j \in \mathbb{N}_0$.

Proof. This was proved in part (a) of the proof of Theorem 6.7. □

6.2 Stationary solutions, causal and invertible ARMA processes

We can now improve on the method we developed in the proof of Theorem 6.7 to establish the existence of stationary solutions of general ARMA(p, q)-equations. The theorem is:

Theorem 6.10 (Stationary solutions of the ARMA(p, q) equation)**.**
Let $(Z_t)_{t\in\mathbb{Z}} \sim WN(0, \sigma^2)$ a white noise process with $\sigma^2 > 0$. Consider the ARMA(p, q)-equation (6.1) *and let $\varphi(z)$ and $\theta(z)$ the corresponding characteristic polynomials. Suppose that $\varphi(z)$ and $\theta(z)$ do not have a common zero on the unit circle, i.e. there is no $w \in S^1$ for which $\varphi(w) = 0$ and $\theta(w) = 0$. Then a stationary solution $(X_t)_{t\in\mathbb{Z}}$ to the ARMA(p, q)-equation* (6.1)*, i.e. to the equation* (6.2)

$$\varphi(B)X_t = \theta(B)Z_t, \quad t \in \mathbb{Z},$$

exists if and only if $\varphi(z) \neq 0$ for all $z \in S^1$, i.e. if φ does not have a zero on the unit circle. If this condition is satisfied, then the stationary solution is unique and given by

$$X_t = \frac{\theta}{\varphi}(B)Z_t, \quad t \in \mathbb{Z},$$

where $S^1 \ni z \mapsto \frac{\theta(z)}{\varphi(z)}$ is indeed a filter function.

Theorem 6.10 tells us that $\frac{\theta(z)}{\varphi(z)}$ is a filter function if φ has no zero on the unit circle, but we would also like to know how to calculate the filter coefficients in a convenient way. This is the contents of the next result. We state it first and then prove Theorem 6.10 and Proposition 6.11 together.

Proposition 6.11. *Let $(Z_t)_{t\in\mathbb{Z}} \sim WN(0, \sigma^2)$ be a white noise process with $\sigma^2 > 0$. Consider the ARMA(p, q)-equation* (6.1) *and let $\varphi(z)$ and $\theta(z)$ the corresponding characteristic polynomials. Suppose that $\varphi(z) \neq 0$ for all $z \in S^1$ and assume that $\varphi_p \neq 0$ (so that φ is a polynomial of degree p). Denote by $z_1, \ldots, z_p \in \mathbb{C}$ the p zeroes of the autoregressive polynomial $\varphi(z)$, counted with multiplicity. Assume that the zeroes are ordered in such a way that*

$$|z_1| \leq |z_2| \leq \ldots \leq |z_{k_0}| < 1 < |z_{k_0+1}| \leq \ldots \leq |z_p|,$$

where $k_0 \in \{0, \ldots, p\}$ (with $k_0 = 0$ meaning that all zeroes have modulus greater than 1, and $k_0 = p$ meaning that all zeroes have modulus less than 1). Then

$$\frac{\theta(z)}{\varphi(z)} = \theta(z)\,(-1)^{k_0} \prod_{k=1}^{k_0} \left(\sum_{j=1}^{\infty} z_k^j z^{-j} \right) \prod_{k=k_0+1}^{p} \left(\sum_{j=0}^{\infty} z_k^{-j} z^j \right) \quad \forall\, z \in S^1, \tag{6.7}$$

and by multiplying the right-hand side out as a Cauchy product $\sum_{j=-\infty}^{\infty} \psi_j z^j$ as in Theorem 5.14 we get the filter $(\psi_j)_{j\in\mathbb{Z}}$ with associated filter function $S^1 \ni z \mapsto \frac{\theta(z)}{\varphi(z)}$. With this filter, the unique stationary solution of (6.1) *is given by*

$$X_t = \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}, \quad t \in \mathbb{Z}.$$

Proof of Theorem 6.10 and Proposition 6.11. (a) Let us first prove Proposition 6.11 and the direct half of Theorem 6.10, i.e. assume that $\varphi(z) \neq 0$ for all $z \in S^1$. By reducing the order if necessary, we may assume without loss of generality that $\varphi_p \neq 0$ (if $\varphi(z) \equiv 1$ the assertion is trivial). Write

$$\varphi(z) = 1 - \varphi_1 z - \ldots - \varphi_p z^p = -\varphi_p(z - z_1) \cdots (z - z_p), \quad \forall\, z \in \mathbb{C},$$

where $z_1, \ldots, z_p$ are the zeroes of φ, counted with multiplicity, ordered as in the statement of Proposition 6.11. Since $1 = \varphi(0) = -\varphi_p \prod_{k=1}^p (-z_k)$ we must have $z_1, \ldots, z_p \neq 0$, hence we can also write

$$\varphi(z) = -\varphi_p \prod_{k=1}^p (-z_k) \prod_{k=1}^p (1 - z/z_k) = \prod_{k=1}^p (1 - z/z_k) \quad \forall\, z \in \mathbb{C}.$$

Denoting $A_k(z) := 1 - z/z_k$ for $z \in \mathbb{C}$ and $k \in \{1, \ldots, p\}$, the ARMA(p,q)-equation (6.2) can hence be written as

$$(1 - z_1^{-1}B) \ldots (1 - z_p^{-1}B) X_t = A_1(B) \ldots A_p(B) X_t = \theta(B) Z_t, \quad t \in \mathbb{Z}. \qquad (6.8)$$

Since $|z_k| \neq 1$ for all $k \in \{1, \ldots, p\}$, the functions $S^1 \ni z \mapsto \frac{1}{A_k(z)} = \frac{1}{1 - z_k^{-1} z}$ are again filter functions by Lemma 6.9. Hence, if $(X_t)_{t\in\mathbb{Z}}$ is a solution of (6.2), equivalently of (6.8), then applying consecutively $\frac{1}{A_1}(B), \ldots, \frac{1}{A_p}(B)$ to (6.8) gives

$$X_t = \frac{1}{A_p}(B) \ldots \frac{1}{A_1}(B)\, \theta(B) Z_t = \frac{\theta}{A_p \cdots A_1}(B) Z_t = \frac{\theta}{\varphi}(B) Z_t$$

by Theorem 5.14, giving uniqueness and the form of the solution. Conversely, to show the existence of a stationary solution, observe again that $S^1 \ni z \mapsto \frac{1}{A_k(z)}$ are filter functions for each $k \in \{1, \ldots, p\}$, hence so is $\frac{\theta(z)}{\varphi(z)} = \frac{\theta(z)}{A_p(z) \cdots A_1(z)}$. Denoting $X_t := \frac{\theta}{\varphi}(B)\, Z_t$ then gives a stationary process, and

$$\varphi(B) \underbrace{X_t}_{= \frac{\theta}{\varphi}(B) Z_t} = A_1(B) \cdots A_p(B) \frac{1}{A_p}(B) \cdots \frac{1}{A_1}(B) \theta(B) Z_t = \theta(B) Z_t$$

by Theorem 5.14, so that $(X_t)_{t\in\mathbb{Z}}$ is indeed a solution of (6.2).

To obtain (6.7), it is enough to observe from Lemma 6.9 that for $k \in \{1, \ldots, k_0\}$ (so that $1/|z_k| > 1$),

$$\frac{1}{A_k(z)} = \frac{1}{1 - z/z_k} = -\sum_{j=1}^\infty z_k^j z^{-j} \quad \forall\, z \in \mathbb{C} : |z| > |z_k|,$$

and for $k \in \{k_0 + 1, \ldots, p\}$ (so that $1/|z_k| < 1$),

$$\frac{1}{A_k(z)} = \frac{1}{(1 - z/z_k)} = \sum_{j=0}^\infty z_k^{-j} z^j \quad \forall\, z \in \mathbb{C} : |z| < |z_k|.$$

Equation (6.7) then follows again from Theorem 5.14.

(b) Let us now show that $\varphi(z) \neq 0$ for all $z \in S^1$ is necessary for a weakly stationary solution to exist, provided that φ and θ do not share common zeroes on the unit circle. Suppose there is some zero $\lambda \in S^1$ of φ such that $\theta(\lambda) \neq 0$ but nevertheless that there exists a stationary solution $(X_t)_{t\in\mathbb{Z}}$ of (6.2). Write

$$\varphi(z) = (1 - \lambda^{-1}z)\widetilde{\varphi}(z), \quad z \in \mathbb{Z},$$

with a polynomial $\widetilde{\varphi}$ of degree $\leq p-1$ and define

$$U_t := \widetilde{\varphi}(B)X_t, \quad W_t := \theta(B)Z_t, \quad t \in \mathbb{Z}.$$

Since $(X_t)_{t\in\mathbb{Z}}$ is weakly stationary, so is $(U_t)_{t\in\mathbb{Z}}$ and we have

$$U_t - \lambda^{-1}U_{t-1} = (1-\lambda^{-1}B)U_t = (1-\lambda^{-1})\widetilde{\varphi}(B)X_t = \varphi(B)X_t \overset{(6.2)}{=} \theta(B)Z_t = W_t, \quad t \in \mathbb{Z}.$$

Iterating as in Equation (6.4) we obtain

$$\begin{aligned} U_t &= \lambda^{-N-1}U_{t-N-1} + \sum_{k=0}^{N} \lambda^{-k} W_{t-k} \\ &= \lambda^{-N-1}U_{t-N-1} + \sum_{k=0}^{N} \lambda^{-k}(Z_{t-k} + \theta_1 Z_{t-k-1} + \ldots + \theta_q Z_{t-k-q}) \end{aligned} \tag{6.9}$$

for every $N \in \mathbb{N}$. For $N \geq q$ we can reorder the last sum according to the Z_{t-k}, distinguish whether $k \in \{0, \ldots, q-1\}$, $k \in \{q, \ldots, N\}$ or $k \in \{N+1, \ldots, N+q\}$ and see that there exist $a_0, \ldots, a_{q-1}, b_1, \ldots, b_q \in \mathbb{C}$ (not depending on N) such that (choose $t = 0$)

$$\begin{aligned} &\sum_{k=0}^{N} \lambda^{-k}(Z_{-k} + \theta_1 Z_{-k-1} + \ldots + \theta_q Z_{k-q}) \\ &\quad = a_0 Z_0 + \ldots + a_{q-1}Z_{-(q-1)} + \lambda^{-N}(b_1 Z_{-N-1} + \ldots + b_N Z_{-N-q}) \\ &\qquad + \sum_{k=q}^{N} \lambda^{-k}(1 + \theta_1\lambda + \ldots + \theta_q\lambda^q)Z_{-k} \quad \forall\, N \geq q. \end{aligned}$$

Inserting this into (6.9) we have

$$\begin{aligned} &U_0 - \lambda^{-N-1}U_{-N-1} - a_0 Z_0 - \ldots - a_{q-1}Z_{-(q-1)} - \lambda^{-N}(b_1 Z_{-N-1} + \ldots + b_N Z_{-N-q}) \\ = &\sum_{k=q}^{N} \lambda^{-k}(1 + \theta_1\lambda + \ldots + \theta_q\lambda^q)Z_{-k} \quad \forall\, N \geq q. \end{aligned}$$

Since $(U_t)_{t\in\mathbb{Z}}$ and $(Z_t)_{t\in\mathbb{Z}}$ are weakly stationary and since $|\lambda| = 1$, it is easily seen that the variance of the left-hand side of this equation is bounded as $N \to \infty$. On the other hand, the variance of the right-hand side is given by

$$\sum_{k=q}^{N} \underbrace{|\lambda|^{-2k}}_{1} \underbrace{|1 + \theta_1\lambda + \ldots + \theta_q\lambda^q}_{=\theta(\lambda)}|^2 \underbrace{\mathrm{Var}\,(Z_{-k})}_{=\sigma^2} = (N-q+1)|\theta(\lambda)|^2\sigma^2,$$

which is unbounded in N since $\theta(\lambda) \neq 0$ and $\sigma^2 > 0$. This is a contradiction, so that we see that a weakly stationary solution to (6.2) cannot exist if φ has zeroes on S^1 (that are not cancelled out by θ). □

Remark 6.12. Suppose that φ has no zeroes on S^1. Then the representation (6.7) actually does not only hold for $z \in S^1$, but for all $z \in \mathbb{C}$ such that

$$|z_{k_0}| < |z| < |z_{k_0+1}|$$

as seen in the proof (supposing that $k_0 \in \{1, \dots, p-1\}$). If $k_0 = 0$, then it is easily seen that the representation (6.7) even holds for all $z \in \mathbb{C} : |z| < |z_1|$ (observe that then $|z_1| > 1$), and if $k_0 = p$, then it holds for all $z \in \mathbb{C}$ such that $|z| > |z_p|$ (observe that then $|z_p| < 1$). In complex analysis such representations are called *Laurent expansions*, and they converge on an annulus. When $k_0 = 0$ the Laurent expansion degenerates to a power series expansion of the form $\sum_{j=0}^{\infty} \psi_j z^j$, valid for all $z \in \mathbb{C}$ with $|z| < |z_1|$.

Remark 6.13. (a) If φ and θ share some roots on the unit circle, then it is still possible that a stationary solution exists. To see this, suppose that there exist polynomials $\widetilde{\varphi}(z), \widetilde{\theta}(z)$ and $Q(z)$ such that

$$\varphi(z) = \widetilde{\varphi}(z)\, Q(z), \quad \theta(z) = \widetilde{\theta}(z)\, Q(z) \quad \forall\, z \in \mathbb{C}$$

and $\widetilde{\varphi}(z)$ does not have zeroes on the unit circle (i.e. all zeroes of φ on S^1 are also zeroes of θ of at least the same multiplicity). By Theorem 6.10 there exists a stationary solution $(X_t)_{t\in\mathbb{Z}}$ of the ARMA equation $\widetilde{\varphi}(B)X_t = \widetilde{\theta}(B)Z_t$, given by $X_t = \frac{\widetilde{\theta}}{\widetilde{\varphi}}(B)Z_t$. Applying $Q(B)$ to both sides of the ARMA equation gives

$$\varphi(B)X_t = Q(B)(\widetilde{\varphi}(B)X_t) = Q(B)(\widetilde{\theta}(B)Z_t) = \theta(B)Z_t,$$

so that $(X_t)_{t\in\mathbb{Z}}$ also solves (6.2). Since $\frac{\theta(z)}{\varphi(z)} = \frac{\widetilde{\theta}(z)}{\widetilde{\varphi}(z)}$ for all $z \in S^1$ (with a suitable continuous interpretation of the left-hand side for $\varphi(z) = 0$) the stationary solution $(X_t)_{t\in\mathbb{Z}}$ can also be written as $\frac{\theta}{\varphi}(B)Z_t$, as in Theorem 6.10. Observe however that the weakly stationary solution will in general no longer be unique if Q has zeroes on the unit circle. This can be seen as in Theorem 6.7 (c).
(b) With more effort one can show that the converse direction of the above remark also holds, i.e. that the ARMA(p, q)-equation (6.2) has a weakly stationary solution if and only if all zeroes of φ on S^1 are also zeroes of θ of at least the same multiplicity. This can be done using arguments from spectral theory or by elementary methods. We omit the details.

Solutions which depend on the future of the noise are strange. The concept of causal solutions avoids this:

Definition 6.14. Let $Z = (Z_t)_{t\in\mathbb{Z}}$ white noise. An ARMA(p, q) process $(X_t)_{t\in\mathbb{Z}}$ with $\varphi(B)X_t = \theta(B)Z_t$, $t \in \mathbb{Z}$, is said to be *causal* (more precisely *causal with respect to the noise* $(Z_t)_{t\in\mathbb{Z}}$), if there exist complex constants $(\psi_j)_{j\in\mathbb{N}_0}$ such that $\sum_{j=0}^{\infty} |\psi_j| < \infty$ and

$$X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j},\ t \in \mathbb{Z}.$$

So X_t only depends on the past of Z, and this in a linear way.

Please observe that causality does not only depend on (X_t), it depends on the interaction between (X_t) and (Z_t). It is well possible that a process (X_t) satisfies e.g. the AR(1) equations $X_t - 1/2X_{t-1} = Z_t$ and $X_t - 2X_{t-1} = W_t$, where both $(Z_t)_{t\in\mathbb{Z}}$ and $(W_t)_{t\in\mathbb{Z}}$ are white noise sequences (cf. exercises). By Theorem 6.5, $(X_t)_{t\in\mathbb{Z}}$ is causal with respect to $(Z_t)_{t\in\mathbb{Z}}$, but not causal with respect to $(W_t)_{t\in\mathbb{Z}}$. In most cases, the underlying white noise sequence will be given and hence we often only say that $(X_t)_{t\in\mathbb{Z}}$ is causal.

Before we characterise causality of ARMA processes, let us give the following lemma:

Lemma 6.15 (Uniqueness of filter coefficients of moving average processes)**.**
Let $Z = (Z_t)_{t\in\mathbb{Z}} \sim WN(0, \sigma^2)$ with $\sigma^2 > 0$. Let $(\psi_j)_{j\in\mathbb{Z}}$ and $(\xi_j)_{j\in\mathbb{Z}}$ be two linear filters such that $\psi(B)Z_t = \xi(B)Z_t$ for all $t \in \mathbb{Z}$, i.e. such that

$$\sum_{j=-\infty}^{\infty} \psi_j Z_{t-j} = \sum_{j=-\infty}^{\infty} \xi_j Z_{t-j} \quad \forall\, t \in \mathbb{Z}.$$

Then $\psi_j = \xi_j$ for all $j \in \mathbb{Z}$.

Proof. Multiplying both sides of $\sum_{j=-\infty}^{\infty} \psi_j Z_{t-j} = \sum_{j=-\infty}^{\infty} \xi_j Z_{t-j}$ by $\overline{Z_0}$ and taking expectations gives

$$\mathbb{E} \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}\overline{Z_0} = \mathbb{E} \sum_{j=-\infty}^{\infty} \xi_j Z_{t-j}\overline{Z}_0.$$

Since $\sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}$ converges in L^2, the sum $\sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}\overline{Z}_0$ converges in L^1 (similarly for the right-hand side) so that we can interchange expectation and summation and obtain

$$\sum_{j=-\infty}^{\infty} \psi_j \mathbb{E}(Z_{t-j}\overline{Z}_0) = \sum_{j=-\infty}^{\infty} \xi_j \mathbb{E}(Z_{t-j}\overline{Z}_0).$$

Since $Z \sim WN(0, \sigma^2)$ we have $\mathbb{E}(Z_{t-j}\overline{Z}_0) = 0$ if $t \neq j$ and $\mathbb{E}(Z_{t-j}\overline{Z}_0) = \sigma^2$ if $j = t$. Hence $\psi_t\sigma^2 = \xi_t\sigma^2$ for all $t \in \mathbb{Z}$, so that $\psi_t = \xi_t$ for all $t \in \mathbb{Z}$. □

We can now characterise when an ARMA(p, q) process is causal.

Theorem 6.16 (Causal ARMA processes)**.**
Let $Z = (Z_t)_{t\in\mathbb{Z}}$ be a white noise process with variance $\sigma^2 > 0$. Consider the ARMA(p, q)-equation (6.1) *and let $\varphi(z)$ and $\theta(z)$ be the characteristic polynomials.*
(a) Suppose that

$$\varphi(z) \neq 0 \quad \forall\, z \in \mathbb{C} : |z| \leq 1, \tag{6.10}$$

i.e. that φ does not have zeroes on or inside the unit circle S^1. Then the unique stationary solution $(X_t)_{t\in\mathbb{Z}}$ of (6.1) *is causal.*
(b) Conversely, if additionally φ and θ do not share common zeroes in $\{z \in \mathbb{C} : |z| \leq 1\}$, then Equation (6.10) *is also necessary for a causal solution of* (6.1) *to exist.*

Proof. (a) If $\varphi(z) \neq 0$ for all $z \in \mathbb{C}$ such that $|z| \leq 1$, then the zeroes of φ lie outside the unit circle, i.e. $k_0 = 0$ in the notion of Proposition 6.11. By (6.7), we have

$$\frac{\theta(z)}{\varphi(z)} = \theta(z) \prod_{k=1}^{p} \left(\sum_{j=0}^{\infty} z_k^{-j} z^j \right)$$

and multiplying this out it is clear (by the Cauchy product) that this has a representation of the form $\sum_{j=0}^{\infty} \psi_j z^j$ for $z \in S^1$. Then $X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}$ by Proposition 6.11, so that the unique stationary solution X is causal.

(b) Now let $X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}$ be a causal solution of (6.1). Then $\psi(z) := \sum_{j=0}^{\infty} \psi_j z^j$ converges for all $z \in \mathbb{C}$ with $|z| \leq 1$, in particular for $z \in S^1$. As in Theorem 5.14 we can multiply $\varphi(z)$ and $\psi(z)$ for all $|z| \leq 1$ (not only for $|z| = 1$) and find by the usual Cauchy product that

$$\varphi(z)\psi(z) = \sum_{j \in \mathbb{Z}} \xi_j z^j =: \xi(z) \quad \forall\, |z| \leq 1$$

with $\sum_{j \in \mathbb{Z}} |\xi_j| < \infty$. On the other hand, since $X_t = \psi(B) Z_t$ we obtain

$$(\varphi\psi)(B) Z_t = \varphi(B)\psi(B) Z_t = \varphi(B) X_t \overset{(6.2)}{=} \theta(B) Z_t \quad \forall\, t \in \mathbb{Z}.$$

Lemma 6.15 now implies that $\varphi(z)\psi(z) = \theta(z)$ for all $z \in S^1$ (since the underlying filters are the same). But since $\theta(z)$ and $\xi(z)$ agree for $z \in S^1$ we must have $\xi_j = \theta_j$ for all $j \in \mathbb{Z}$ by Theorem 5.9 (with $\theta_0 := 1$ and $\theta_j := 0$ for $j < 0$ or $j > q$). But this then implies that

$$\xi(z) = \theta(z) \quad \forall\, z \in \mathbb{C} : |z| \leq 1$$

(i.e. not only for $z \in S^1$). Hence we obtain

$$\varphi(z)\psi(z) = \theta(z) \quad \forall\, z \in \mathbb{C} : |z| \leq 1.$$

In particular, a zero z_0 of φ with $|z_0| \leq 1$ must also be a zero of θ. Hence $\varphi(z) \neq 0$ for all $|z| \leq 1$ since we assumed that there were no common zeroes of φ and θ on or inside the unit circle. This finishes the proof. □

Remark 6.17. As in Remark 6.13, common zeroes of φ and θ on or inside the unit circle can be factored out. It is possible to show that (6.1) admits a causal solution if and only if all zeroes $z \in \mathbb{C}$ of φ with $|z| \leq 1$ are also zeroes of θ with at least the same multiplicity. We omit the details.

A concept that is related to causality is that of invertibility. However, here one does not try to represent X_t in a linear way of the past noise, but rather to represent the noise in a linear way in terms of the past observations X_t. The reason why one is interested in such a concept is that in practice one usually does not observe the noise, but the time series X, and from the invertibility condition one can recover the noise. Having (in practise an estimator for) the noise, one can check if it is really white noise and see if the model assumptions were correct.

Let us now give the precise definition of invertibility:

Definition 6.18. Let $Z = (Z_t)_{t\in\mathbb{Z}} \sim WN(0, \sigma^2)$ and consider an ARMA process $(X_t)_{t\in\mathbb{Z}}$ satisfying $\varphi(B)X_t = \theta(B)Z_t$. Then $(X_t)_{t\in\mathbb{Z}}$ is said to be *invertible* (more precisely *invertible with respect to* $(Z_t)_{t\in\mathbb{Z}}$), if there exist complex coefficients $(\pi_j)_{j\in\mathbb{N}_0}$ with $\sum_{j=0}^{\infty} |\pi_j| < \infty$ such that

$$Z_t = \sum_{j=0}^{\infty} \pi_j X_{t-j}, \quad \forall\, t \in \mathbb{Z}.$$

As for causality, also invertibility is a concept that depends on the interplay of both X and Z and not only on X. Invertibility can be characterised in a similar fashion as causality, however this time in terms of the zeroes of θ rather than φ:

Theorem 6.19 (Invertible ARMA processes)**.**
Let $Z = (Z_t)_{t\in\mathbb{Z}} \sim WN(0, \sigma^2)$ with $\sigma^2 > 0$ and $X = (X_t)_{t\in\mathbb{Z}}$ be an ARMA(p, q) process satisfying (6.1) *with characteristic polynomials φ and θ.*
(a) If

$$\theta(z) \neq 0 \quad \forall\, z \in \mathbb{C} : |z| \leq 1, \tag{6.11}$$

i.e. if θ does not have zeroes on or inside the unit circle S^1, then X is invertible (with respect to Z). The invertible representation of Z is then given by

$$Z_t = \frac{\varphi}{\theta}(B)X_t = \sum_{j=0}^{\infty} \pi_j X_{t-j}, \quad t \in \mathbb{Z},$$

where

$$\frac{\varphi(z)}{\theta(z)} = \sum_{j=0}^{\infty} \pi_j z^j \quad \forall\, z \in S^1.$$

(b) Conversely, if additionally φ and θ do not share common zeroes in $\{z \in \mathbb{C} : |z| \leq 1\}$, then Equation (6.11) *is also necessary for X to be invertible.*

Proof. (a) Suppose that $\theta(z) \neq 0$ for all $|z| \leq 1$. Without loss of generality assume that $\theta_q \neq 0$ (otherwise reduce the order, and $q = 0$ is trival). Denote by $z_1, \ldots, z_q$ the zeroes of θ (counted with multiplicity), so that

$$\theta(z) = \theta_q(z - z_1) \ldots (z - z_q) = \prod_{k=1}^{q} (1 - z/z_k),$$

similar to the proof of Proposition 6.11. By assumption, $1/|z_k| < 1$ for all $k \in \{1, \ldots, p\}$, so that by Lemma 6.9 we have

$$\frac{1}{\theta(z)} = \prod_{k=1}^{q} \sum_{j=0}^{\infty} z_k^{-j} z^j \quad \forall\, z \in S^1.$$

From this representation we see that $\frac{\varphi(z)}{\theta(z)} = \sum_{j=0}^{\infty} \pi_j z^j$ for all $z \in S^1$ where $\sum_{j=0}^{\infty} |\pi_j| < \infty$. Applying now $\frac{1}{\theta}(B)$ to (6.2) gives

$$Z_t = \frac{\theta}{\theta}(B)Z_t = \frac{1}{\theta}(B)\underbrace{\theta(B)Z_t}_{=\varphi(B)X_t} = \frac{\varphi}{\theta}(B)X_t = \sum_{j=0}^{\infty} \pi_j X_{t-j}.$$

(b) Now assume that θ and φ do not have common zeroes on or inside the unit circle and that X is invertible, so that $Z_t = \sum_{j=0}^{\infty} \pi_j X_{t-j}$ for all $t \in \mathbb{Z}$ with $\sum_{j=0}^{\infty} |\pi_j| < \infty$. Denote $\pi(z) := \sum_{j=0}^{\infty} \pi_j z^j$ for all $z \in \mathbb{C}$ with $|z| \leq 1$. From (6.2) we then obtain

$$\varphi(B)Z_t = \varphi(B)\pi(B)X_t = \pi(B)\underbrace{\varphi(B)X_t}_{=\theta(B)Z_t} = (\pi\theta)(B)Z_t,$$

so that $\varphi(z) = \pi(z)\theta(z)$ for all $z \in S^1$ by Lemma 6.15. An argument similar to the proof of Theorem 6.16 then shows that even

$$\varphi(z) = \pi(z)\theta(z) \quad \forall\, z \in \mathbb{C} : |z| \leq 1.$$

Hence a zero z_0 of $\mathbb{C}$ with $|z_0| \leq 1$ must also be a zero of φ, which was excluded by assumption. It follows that $\theta(z) \neq 0$ for all $|z| \leq 1$. □

6.3 Homogenous linear difference equations with constant coefficients

In Section 6.4 we will show that the autocovariance function γ of a causal ARMA(p, q) process with autoregressive polynomial φ satisfies

$$\gamma(k) - \varphi_1\gamma(k-1) - \ldots - \varphi_p\gamma(k-p) = 0 \quad \forall\, k \geq q+1.$$

This is a homogeneous linear difference equation for γ which can also be written as

$$(\varphi(B)\gamma)(k) = 0 \quad \forall\, k \geq q+1,$$

if B denotes the backshift operator. Similarly, we will encounter that the filter coefficients ψ_j in the MA(∞)-representation of a causal ARMA(p, q) process satisfy the equation

$$\psi_j - \varphi_1\psi_{j-1} - \ldots - \varphi_p\psi_{j-k} = 0 \quad \forall\, j \geq \max(p, q+1),$$

or with the use of the backshift operator,

$$\varphi(B)\psi_j = 0 \quad \forall\, j \geq \max(p, q+1).$$

In this section we shall show that there is a closed form solution to equations of this form. As an introductory example, we start with the Fibonacci numbers.

Example 6.20 (Fibonacci numbers)**.** Consider the sequence $(a_n)_{n\in\mathbb{N}_0}$ defined by

$$a_0 := 0, \quad a_1 := 1, \quad a_n := a_{n-1} + a_{n-2} \quad \forall\, n \geq 2.$$

These are called the *Fibonacci numbers.* Surely, in Analysis 1 you have shown using induction that a closed form solution to this recurrence equation is given by

$$a_n = \frac{1}{\sqrt{5}}\left(\frac{1+\sqrt{5}}{2}\right)^n - \frac{1}{\sqrt{5}}\left(\frac{1-\sqrt{5}}{2}\right)^n \quad \forall\, n \in \mathbb{N}_0,$$

Proving that it is true by induction is one thing, but how on earth does one come to the formula? This is what we shall see now. For the moment, let us get rid of the initial conditions $a_0 = 0$ and $a_1 = 1$ and only look at the recursion $a_n = a_{n-1} + a_{n-2}$. Denote by V the set of all complex valued sequences $(x_n)_{n\in\mathbb{N}_0}$ such that

$$x_n = x_{n-1} + x_{n-2} \quad \forall\, n \geq 2. \tag{6.12}$$

If $(x_n)_{n\in\mathbb{N}_0}$ and $(y_n)_{n\in\mathbb{N}_0}$ are in V and $\lambda \in \mathbb{C}$, then

$$\lambda x_n + y_n = \lambda x_{n-1} + y_{n-1} + \lambda x_{n-2} + y_{n-2},$$

so that $(\lambda x_n + y_n)_{n\in\mathbb{N}_0}$ is also in V. This shows that V is a (complex) vector space (it is also non-empty since the zero sequence $x_n = 0$ is in V). So, the set of all sequences satisfying this recurrence equation is a vector space. So it should have a dimension (finite or infinite) and also a basis. Can we find such a basis and obtain the dimension?
Observe that a sequence in V is determined by its initial conditions x_0 and x_1 and then by the recursion (6.12). This suggests that there is a linear isomorphism between $\mathbb{C}^2$ and V. To be precisely, let

$$\Psi : \mathbb{C}^2 \to V, \quad (x_0, x_1)' \mapsto (x_n)_{n\in\mathbb{N}_0}, \quad \text{where } x_n \text{ is determined from (6.12) for } n \geq 2.$$

It is easily checked that Ψ is linear, i.e. $(\Psi(\lambda x_0 + y_0, \lambda x_1 + y_1)) = \lambda\Psi(x_0, x_1) + \Psi(y_0, y_1)$. Also, Ψ is clearly injective. It is also surjective, since a sequence $(x_n)_{n\in\mathbb{N}_0} \in V$ is obviously determined from (x_0, x_1), and then $(x_n)_{n\in\mathbb{N}_0} = \Psi(x_0, x_1)$. So, Ψ is a bijection from $\mathbb{C}^2$ to V. Since $\mathbb{C}^2$ is 2-dimensional, so must be V. A basis is obtained from $\Psi(0, 1)$ and $\Psi(1, 0)$, since $(1, 0), (0, 1)$ forms a basis of $\mathbb{C}^2$, but this is not a good one for our purpose. Let us look if we can find two nicer linearly independent solutions in V. For that, we look at particularly simply solutions. The following sequence seems to be particularly simply.

$$\text{Ansatz: } b_n = q^n, \quad n \in \mathbb{N}_0, \quad \text{for some } q \in \mathbb{C}.$$

For which q will this sequence $b = (b_n)_{n\in\mathbb{N}_0}$ be in V? Well, by definition, $b \in V$ if and only if $b_n = b_{n-1} + b_{n-2}$ for all $n \geq 2$, i.e. if and only if

$$q^n = q^{n-1} + q^{n-2} \quad \forall\, q \geq 2.$$

It is clear that $q = 0$ does not satisfy this for $n = 2$ (since $0 = 0^2 \neq 0^1 + 0^0 = 0 + 1 = 1$), so we may divide the above equation by q^{n-2}. This then leads to the equation

$$q^2 = q + 1.$$

So, the sequence $(q^n)_{n\in\mathbb{N}_0}$ is in V if and only if $q^2 = q+1$, i.e. $q^2-q-1 = 0$. This quadratic equation can be solved giving

$$q_{1,2} = \frac{1 \pm \sqrt{(-1)^2 - 4\cdot(-1)}}{2} = \frac{1 \pm \sqrt{5}}{2}.$$

Denote

$$c_n = q_1^n = \left(\frac{1+\sqrt{5}}{2}\right)^n \quad \text{and} \quad d_n = q_2^n = \left(\frac{1-\sqrt{5}}{2}\right)^n \quad \forall\, n \in \mathbb{N}_0.$$

Then we have just seen that $c = (c_n)_{n\in\mathbb{N}_0}$ and $d = (d_n)_{n\in\mathbb{N}_0}$ are in V. On the other hand, it is easily seen that they are linearly independent, because $\lambda_1 c + \lambda_2 d = 0$ implies

$$\lambda_1 c_0 + \lambda_2 d_0 = 0 \quad \text{and} \quad \lambda_1 c_1 + \lambda_2 d_1 = 0.$$

Since $c_0 = d_0 = 1$ this implies $\lambda_1 = -\lambda_2$, hence $c_1 = d_1$ if $\lambda_1 \neq 0$ or $\lambda_2 \neq 0$. But $c_1 \neq d_1$, so that we conclude that $\lambda_1 = \lambda_2 = 0$, showing that c and d are linearly independent. Since $\dim V = 2$ we conclude that c, d forms a basis of V. It follows that a sequence $x = (x_n)_{n\in\mathbb{N}_0}$ is in V if and only if there are $\lambda_1, \lambda_2 \in \mathbb{C}$ such that $x_n = \lambda_1 c_n + \lambda_2 d_n$ for all $n \in \mathbb{N}_0$.
So now we have described all solutions in V. Coming back to the Fibonacci numbers $(a_0, a_1, \ldots)$, we know that they are in V. Hence there are λ_1, λ_2 such that $a = \lambda_1 c + \lambda_2 d$, in particular we must have

$$0 = a_0 = \lambda_1 \underbrace{c_0}_{=1} + \lambda_2 \underbrace{d_0}_{=1} = \lambda_1 + \lambda_2$$

and

$$1 = a_1 = \lambda_1 \underbrace{c_1}_{(1+\sqrt{5})/2} + \underbrace{\lambda_2}_{=-\lambda_1 \text{ from above}} \underbrace{d_1}_{(1-\sqrt{5})/2} = \lambda_1 \left(\frac{1+\sqrt{5}}{2} - \frac{1-\sqrt{5}}{2} \right).$$

This shows $1 = \sqrt{5}\lambda_1$ and hence $\lambda_1 = 1/\sqrt{5} = -\lambda_2$. Hence

$$a_n = \lambda_1 c_n + \lambda_2 d_n = \frac{1}{\sqrt{5}} \left(\frac{1+\sqrt{5}}{2} \right)^n - \frac{1}{\sqrt{5}} \left(\frac{1-\sqrt{5}}{2} \right)^n \quad \forall\, n \in \mathbb{N}_0,$$

which is what we wanted. We could have done the same thing also considering only real sequences and the real vector space V, but it is often advantageous to work with complex vector spaces. □

We will generalise the previous example now. Let us first give a formal definition:

Definition 6.21. Let $T = \mathbb{Z}$, $T = \mathbb{N}_0$ or $T = \{t_0, t_0 + 1, t_0 + 2, \ldots\} = t_0 + \mathbb{N}_0$ for some $t_0 \in \mathbb{Z}$. Let $\varphi_1, \ldots, \varphi_p \in \mathbb{C}$ with $\varphi_p \neq 0$. Then an equation of the form

$$x_n = \varphi_1 x_{n-1} + \ldots + \varphi_p x_{n-p} \quad \forall\, n \in \mathbb{Z} : n - p \in T \tag{6.13}$$

is called a *homogeneous linear difference equation with constant coefficients.* A sequence $(x_n)_{n\in T}$ of complex numbers is called a solution of this equation, if it satisfies (6.13). The polynomial

$$1 - \varphi_1 z - \ldots - \varphi_p z^p, \quad z \in \mathbb{C},$$

is called the *characteristic polynomial* of this difference equation.

We will mostly be interested in the case when $T = \mathbb{N}_0$, but the theory can be made also for the general index sets T given in Definition 6.21. The name *homogeneous* refers to the fact that we can write (6.13) as

$$x_n - \varphi_1 x_{n-1} - \ldots - \varphi_p x_{n-p} = 0, \quad \forall\, n \in \mathbb{Z} : n - p \in T$$

i.e. with zero on the right hand side. For a given sequence $(b_n)_{n\in T}$, an equation of the form

$$x_n - \varphi_1 x_{n-1} - \ldots - \varphi_p x_{n-p} = b_n \quad \forall\, n \in \mathbb{Z} : n - p \in T,$$

would be called *inhomogeneous* if at least one of the b_n with $n - p \in T$ is different from zero. We shall only be interested in homogeneous equations.

We are concerned with finding all solutions of (6.13). To start with, as in Example 6.20, the set of all solutions of (6.13) forms a p-dimensional complex vector space:

Lemma 6.22. *In the setting of Definition 6.21, denote by V the set of all complex valued solutions of* (6.13)*. Then V is a p-dimensional complex vector space, and $\Psi : \mathbb{C}^p \to V$, defined by*

$$\Psi(x_{t_0}, \ldots, x_{t_0+p-1}) \mapsto (x_n)_{n\in t_0+\mathbb{N}_0},$$

where x_n for $n \geq t_0 + p$ is given recursively by (6.13)*, is a vector space isomorphism when $T = t_0 + \mathbb{N}_0$. If $T = \mathbb{Z}$, we can take $t_0 \in \mathbb{Z}$ arbitrary, and obtain a vector space isomorphism $\Psi : \mathbb{C}^n \to V$ by*

$$\Psi(x_{t_0}, \ldots, x_{t_0+p-1}) \mapsto (x_n)_{n\in\mathbb{Z}},$$

where x_n for $n \geq t_0 + p$ is given by (6.13)*, and for $n < t_0$ it is given recursively by*

$$x_n = \frac{1}{\varphi_p}\left(x_{n+p} - \varphi_1 x_{n+p-1} - \ldots - \varphi_{p-1} x_{n+1}\right).$$

Proof. That V is a vector space is readily checked, as well as that Ψ such defined is a vector space isomorphism. Observe that when $T = \mathbb{Z}$, the recursion to the left is simple a rewriting of (6.13). □

Now we would like to find solutions of (6.13). As in Example 6.20, we can make again an ansatz of the form $x_n = \xi^n$ for some $\xi \in \mathbb{C} \setminus \{0\}$ (when $\xi = 0$, we do not get an interesting solution, in particular none which is linearly independent from other ones). Plugging this into (6.13) we obtain

$$\xi^n - \varphi_1 \xi^{n-1} - \ldots - \varphi_p \xi^{n-p} = 0,$$

and dividing by $\xi^n \neq 0$ leads to

$$1 - \varphi_1 \xi^{-1} - \ldots - \varphi_p \xi^{-p} = 0.$$

With the aid of the characteristic polynomial $\varphi(z) = 1 - \varphi_1 z - \ldots - \varphi_p z^p$ this can be rewritten as $\varphi(\xi^{-1}) = 0$, so that ξ^{-1} must be a zero of φ in order for $(\xi^n)_{n\in T}$ to be in V. If the characteristic polynomial φ has p pairwise distinct roots $\xi_1^{-1}, \ldots, \xi_p^{-1}$, then it is possible to show that $(\xi_j^n)_{n\in T}$, $j = 1, \ldots, p$, forms a basis of V and we are done in the description of V (we shall do that later in a more general context). But what happens if the characteristic polynomial has multiple roots? E.g., the recurrence equation $x_n = x_{n-1} - \frac{1}{4} x_{n-2}$ has characteristic polynomial $\varphi(z) = 1 - z + z^2/4 = (1 - z/2)^2$, so that φ has a double zero at 2 and the above ansatz only gives us $(2^{-n})_{n\in T}$ as a solution, but not a second one. It turns out that then $(n2^{-n})_{n\in T}$ is a second, linearly independent solution, but we shall do this now in more generality.

For that, it is convenient to work again with the backshift operator. For a sequence $(x_n)_{n\in T}$ we write $Bx_n = x_{n-1}$ when $n-1 \in T$, and

$$\varphi(B)x_n = x_n - \varphi_1 x_{n-1} - \ldots - \varphi_p x_{n-p}, \quad n-p \in T.$$

As for linear filters, this is like inserting $z = B$ in the characteristic polynomial $\varphi(z)$ and then apply it to the sequence (x_n). Now if $\widetilde{\varphi}(z) = 1 - \widetilde{\varphi}_1 z - \ldots - \widetilde{\varphi}_{\widetilde{p}} z^p$ is the characteristic polynomial of another homogeneous difference equation of order $\widetilde{p}$, then

$$(\widetilde{\varphi}\varphi)(B)x_n = \widetilde{\varphi}(B)\left(\varphi(B)x\right)_n \quad \forall\, n \in \mathbb{Z} : n - p - \widetilde{p} \in T,$$

i.e. $(\widetilde{\varphi}\varphi)(B) = \widetilde{\varphi}(B) \circ \varphi(B)$; this follows as in the proof of Equation (5.3) in Theorem 5.14. But this equips us with a method to obtain more solutions. We can factorise the polynomial φ. Denote by $\xi_1^{-1}, \ldots, \xi_v^{-1}$ the pairwise different zeroes of φ and denote by r_u, $u = 1, \ldots, v$, their multiplicities. Then $\sum_{u=1}^{v} r_u = p$ and we can write

$$\varphi(z) = 1 - \varphi_1 z - \ldots - \varphi_p z^p = \prod_{u=1}^{v} (1 - \xi_u z)^{r_u} \quad \forall\, z \in \mathbb{C}.$$

Hence,

$$\varphi(B) = \prod_{u=1}^{v} (1 - \xi_u B)^{r_u}.$$

If we find a sequence $(x_n)_{n\in T}$ that satisfies $(1 - \xi_u B)^{r_u} x_n = 0$ for $n - r_u \in T$ for some fixed u, then

$$\varphi(B)x_n = \left(\prod_{j\neq u} (1 - \xi_j B)^{r_j}\right) \underbrace{(1 - \xi_u B)^{r_u} x_n}_{=0} = 0 \quad \forall\, n - p \in T, \tag{6.14}$$

so that this sequence satisfies (6.13). So let us first have a look at solutions of the equation $(1 - \xi_u B)^{r_u} x_n = 0$.

Lemma 6.23. *Let $\xi \in \mathbb{C} \setminus \{0\}$ and $a_0, \ldots, a_s \in \mathbb{C}$. Let*

$$x_n = (a_0 + a_1 n + \ldots + a_s n^s)\xi^n$$

for all $n \in T$. Then there are constants $b_0, \ldots, b_{s-1} \in \mathbb{C}$ such that

$$(1 - \xi B)x_n = \begin{cases} (b_0 + b_1 n + \ldots + b_{s-1} n^{s-1})\xi^n \quad \forall\, n \in \mathbb{Z} : n - 1 \in T, & \text{if} \quad s \geq 1, \\ 0 \quad \forall\, n \in \mathbb{Z} : n - 1 \in T & \text{if} \quad s = 0. \end{cases}$$

In particular,

$$(1 - \xi B)^{s+1} x_n = 0 \quad \forall\, n \in \mathbb{Z} : n - s - 1 \in T.$$

Proof. If $s = 0$ we have $x_n = a_0\xi^n$ and hence $(1 - \xi B)x_n = x_n - \xi x_{n-1} = 0$. If $s \geq 1$ we have

$$\begin{aligned}(1 - \xi B)x_n &= x_n - \xi x_{n-1} \\ &= (a_0 + a_1 n + \ldots + a_s n^s)\xi^n - \xi(a_0 + a_1(n-1) + \ldots + a_s(n-1)^s)\xi^{n-1} \\ &= \xi^n \left(c_0 + c_1 n + \ldots + c_{s-1}s^{n-1} + a_s\left(n^s - (n-1)^s\right)\right)\end{aligned}$$

for some $c_0, \ldots, c_{s-1} \in \mathbb{C}$. But also $n^s - (n-1)^s$ is a polynomial in the variable n of degree $s-1$, hence there are $d_1, \ldots, d_{s-1} \in \mathbb{C}$ such that $a_s(n^s - (n-1)^s) = d_0 + d_1 n + \ldots + d_{s-1}n^{s-1}$. The first claim then follows with $b_j = c_j + d_j$, $j \in \{0, \ldots, s-1\}$. Hence application of $(1 - \xi B)$ to x_n results in a polynomial in n of degree less than the previous one. Applying this $s + 1$ times we obtain $(1 - \xi B)^{s+1}x_n = 0$ for (x_n) of the given form. □

Corollary 6.24. *Let $\xi \in \mathbb{C}\backslash\{0\}$. Then the s sequences $x^{(j)} = (x_n^{(j)})_{n \in T}$, $j \in \{0, \ldots, s-1\}$, defined by*

$$x_n^{(j)} := n^j \xi^n, \quad n \in T,$$

are linearly independent and form a basis of the space of all sequences $(x_n)_{n \in T}$ satisfying the homogeneous difference equation $(1 - \xi B)^s x_n = 0$ for all $n \in \mathbb{Z}$ with $n - s \in T$. A sequence $x = (x_n)_{n \in T}$ is a solution of $(1 - \xi B)^s x_n = 0$ (for all $n - s \in T$) if and only if there are constants $a_0, \ldots, a_{s-1} \in \mathbb{C}$ such that $x_n = \sum_{j=0}^{s-1} a_j n^j \xi^n$ for all $n \in T$.

Proof. By Lemma 6.23, a sequence of the form $x_n = \sum_{j=0}^{s-1} a_j n^j \xi^n$ is a solution of $(1 - \xi B)^s x_n = 0$. Since the sequences $x^{(0)}, \ldots, x^{(s-1)}$ are of this form, they are also solutions. To see that they are linearly independent, assume that there are $\lambda_0, \ldots, \lambda_{s-1} \in \mathbb{C}$ such that

$$0 = \sum_{j=0}^{s-1} \lambda_j x_n^{(j)} = \sum_{j=0}^{s-1} \lambda_j n^j \xi^n \quad \forall\, n \in T.$$

Deviding this equation by $n^{s-1}\xi^n$ and letting $n \to \infty$ shows $\lambda_{s-1} = 0$. Next, we divide by $n^{s-2}\xi^n$ and obtain $\lambda_{s-2} = 0$. Continuing in this way we derive that $\lambda_{s-1} = \ldots = \lambda_0 = 0$, giving linear independence of $x^{(0)}, \ldots, x^{(s-1)}$. Since the space of solutions of $(1-\xi B)^s x_n = 0$ is s-dimensional, $x^{(0)}, \ldots, x^{(s-1)}$ must be a basis of this space. Since a sequence $(x_n)_{n \in T}$ is in the span of $x^{(0)}, \ldots, x^{(s-1)}$ if and only if it is of the form $\sum_{j=0}^{s-1} a_j n^j \xi^n$ for all $n \in T$ with some $a_0, \ldots, a_{s-1} \in \mathbb{C}$ we get the form of the space of solutions. □

Before we give the complete solution of the homogeneous linear difference equation (6.13), let us recall the following fact from linear algebra:

Lemma 6.25. *Let $x_1, \ldots, x_n \in \mathbb{C}$. Then the* Vandermonde matrix

$$A := \begin{pmatrix} 1 & 1 & 1 & \ldots & 1 \\ x_1 & x_2 & x_3 & \ldots & x_n \\ x_1^2 & x_2^2 & x_3^2 & \ldots & x_n^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_1^{n-1} & x_2^{n-1} & x_3^{n-1} & \ldots & x_n^{n-1} \end{pmatrix} \in \mathbb{C}^{n \times n}$$

has determinant

$$\det(A) = \prod_{1\leq i<j\leq n} (x_j - x_i).$$

This determinant is also called a Vandermonde determinant*.*

Proof. Exercises. □

Equipped with our previous findings we can now give the complete solution to (6.13).

Theorem 6.26. *[General solution of homogeneous difference equation with constant coefficients]*
Let $T = \mathbb{Z}$, $T = \mathbb{N}_0$ or $T = \{t_0, t_0 + 1, t_0 + 2, \ldots\} = t_0 + \mathbb{N}_0$ for some $t_0 \in \mathbb{Z}$. Let $\varphi_1, \ldots, \varphi_p \in \mathbb{C}$ with $\varphi_p \neq 0$. Denote by V the p-dimensional complex vector space of solutions $(x_n)_{n\in T}$ of (6.13), i.e. of

$$x_n = \varphi_1 x_{n-1} + \ldots + \varphi_p x_{n-p} \quad \forall\, n \in \mathbb{Z} : n - p \in T.$$

Denote by $\xi_1^{-1}, \ldots, \xi_v^{-1}$ the pairwise different zeroes of the characteristic polynomial $\varphi(z) = 1 - \varphi_1 z - \ldots - \varphi_p z^p$ and by r_u, $u = 1, \ldots, v$, their multiplicities. Then a sequence $(x_n)_{n\in T}$ is in V if and only if there are constants $\alpha_{u,j} \in \mathbb{C}$, $u \in \{1, \ldots, v\}$ and $j \in \{0, \ldots, r_u - 1\}$ such that

$$x_n = \sum_{u=1}^{v} \sum_{j=0}^{r_u-1} \alpha_{u,j} n^j \xi_u^n \quad \forall\, n \in T.$$

The p sequences $x^{(u,j)} = (x_n^{(u,j)})_{n\in T}$ with $u \in \{1, \ldots, v\}$ and $j \in \{0, \ldots, r_u - 1\}$ defined by

$$x_n^{(u,j)} := n^j \xi_u^n, \quad n \in T,$$

form a basis of V.

Proof. We know from Corollary 6.24 that each of the sequences $x^{(u,j)}$ satisfies the difference equation $(1-\xi_u B)^{r_u-1} x_n^{(u,j)} = 0$, hence $x^{(u,j)}$ also satisfies $\varphi(B) x_n^{(u,j)} = 0$ for all n with $n - p \in T$ by (6.14). It hence suffices to show that the $x^{(u,j)}$, $u \in \{1, \ldots, v\}$, $j \in \{0, \ldots, r_u - 1\}$ are linearly independent (then they must form a basis of V since the dimension of V is p). To see the linear independence, let the ξ_u be ordered such that

$$|\xi_1| \geq |\xi_2| \geq \ldots \geq |\xi_v|,$$

and if $|\xi_j| = |\xi_i|$ for some i, j, choose the ordering in such a way that additionally $r_i \geq r_j$ when $i \leq j$. Let $\lambda_{u,j} \in \mathbb{C}$ (for $u \in \{1, \ldots, v\}$ and $j \in \{0, \ldots, r_i - 1\}$) such that

$$\sum_{u=1}^{v} \sum_{j=0}^{r_u-1} \lambda_{u,j} x^{(u,j)} = 0,$$

i.e. that

$$\sum_{u=1}^{v} \sum_{j=0}^{r_u-1} \lambda_{u,j} n^j \xi_u^n = 0 \quad \forall\, n \in T. \tag{6.15}$$

We have to show that $\lambda_{u,j} = 0$ for all u and j. We do this recursively and show first that the $\lambda_{1,j}$-coefficients are zero. We distinguish between two cases.

(i) Case 1: $|\xi_1| > |\xi_2|$.
In that case we divide Equation (6.15) by $n^{r_1-1}\xi_1^n$ and let $n \to \infty$, resulting in $\lambda_{1,r_1-1} = 0$. Next, we divide (6.15) by $n^{r_1-2}\xi_1^n$ and let $n \to \infty$ to obtain $\lambda_{1,r_1-2} = 0$. Continuing in this way we obtain $\lambda_{1,0} = \ldots = \lambda_{1,r_1-1} = 0$.

(ii) Case 2: $|\xi_1| = |\xi_2| = \ldots = |\xi_m|$ for some $m \in \{2, \ldots, v\}$.
Then $r_1 \geq r_2 \geq \ldots \geq r_m$ by our assumed ordering. Denote by $k \in \{1, \ldots, m\}$ the largest integer such that $r_1 = \ldots = r_k$ (hence $r_1 = \ldots = r_k > r_{k+1} \geq \ldots \geq r_m$ if $k < m$). Dividing Equation (6.15) by $n^{r_1-1}|\xi_1|^n$ gives

$$\sum_{u=1}^{k} \lambda_{u,r_1-1} \left(\frac{\xi_u}{|\xi_u|}\right)^n + \sum_{u=1}^{k}\sum_{j=0}^{r_1-2} \lambda_{u,j} n^{j-(r_1-1)} \left(\frac{\xi_u}{|\xi_u|}\right)^n + \sum_{u=k+1}^{v}\sum_{j=0}^{r_u-1} \lambda_{u,j} n^{j-(r_1-1)} \left(\frac{\xi_u}{|\xi_1|}\right)^n = 0.$$

The second and third term tend to zero as $n \to \infty$ (since $|\xi_u| < |\xi_1|$ for $u > k$), hence we can write

$$\sum_{u=1}^{k} \lambda_{u,r_1-1} \left(\frac{\xi_u}{|\xi_u|}\right)^n = g_n \quad \forall\, n \in T$$

where $(g_n)_{n \in T}$ is a sequence with $\lim_{n\to\infty} g_n = 0$. Now let $\theta_u \in [0, 2\pi)$ such that $\mathrm{e}^{\mathrm{i}\theta_u} = \frac{\xi_u}{|\xi_u|}$ for $u \in \{1, \ldots, k\}$. Then the above equation can be written as

$$\sum_{u=1}^{k} \lambda_{u,r_1-1} \mathrm{e}^{in\theta_u} = g_n \quad \forall\, n \in T. \tag{6.16}$$

For each $n \in T$ denote

$$A_n := \begin{pmatrix} \mathrm{e}^{in\theta_1} & \ldots & \mathrm{e}^{in\theta_k} \\ \mathrm{e}^{\mathrm{i}(n+1)\theta_1} & \ldots & \mathrm{e}^{\mathrm{i}(n+1)\theta_k} \\ \vdots & \ddots & \vdots \\ \mathrm{e}^{\mathrm{i}(n+k-1)\theta_1} & \ldots & \mathrm{e}^{\mathrm{i}(n+k-1)\theta_k} \end{pmatrix} \in \mathbb{C}^{k\times k}, \quad \lambda := \begin{pmatrix} \lambda_{1,r_1-1} \\ \vdots \\ \lambda_{k,r_1-1} \end{pmatrix} \in \mathbb{C}^k$$

and

$$G_n := \begin{pmatrix} g_n \\ \vdots \\ g_{n+k-1} \end{pmatrix} \in \mathbb{C}^k.$$

Then $G_n \to 0$ as $n \to \infty$ and (6.16) implies

$$A_n\lambda = G_n \quad \forall\, n \in T.$$

Multiplying the u'th column of A_n by $\mathrm{e}^{-\mathrm{i}n\theta_u}$ and observing the rules for the calculation of the determinant we obtain

$$\det(A_n) = \mathrm{e}^{\mathrm{i}(\theta_1+\ldots+\theta_k)n} \det \begin{pmatrix} 1 & \ldots & 1 \\ \mathrm{e}^{\mathrm{i}\theta_1} & \ldots & \mathrm{e}^{\mathrm{i}\theta_k} \\ \vdots & \ddots & \vdots \\ \mathrm{e}^{\mathrm{i}(k-1)\theta_1} & \ldots & \mathrm{e}^{\mathrm{i}(k-1)\theta_k} \end{pmatrix} = \mathrm{e}^{\mathrm{i}(\theta_1+\ldots+\theta_k)n} \det(A_0).$$

But A_0 is a Vandermonde matrix, hence $\det(A_0) = \prod_{1\le u<w\le k}(\mathrm{e}^{\mathrm{i}\theta_w} - \mathrm{e}^{\mathrm{i}\theta_u}) \neq 0$ by Lemma 6.25. It follows that A_n is invertible. Denote by $a_u^{(n)} \in \mathbb{C}^k$ the u'th column of A_n. By Cramèr's rule (linear algebra!), the solution $\lambda = (\lambda_{1,r_1-1}, \ldots, \lambda_{k,r_1-1})'$ of $A_n\lambda = G_n$ is given by

$$\lambda_{u,r_1-1} = \frac{\det(a_1^{(n)}, \ldots, a_{u-1}^{(n)}, G_n, a_{u+1}^{(n)}, \ldots, a_k^{(n)})}{\det A_n}, \quad u \in \{1, \ldots, k\}.$$

But $G_n \to 0$ as $n \to \infty$, while the components of A_n remain bounded as $n \to \infty$. Since $|\det A_n| = |\det A_0|$ the right hand side of the above equation converges to 0 as $n \to \infty$, resulting in $\lambda_{u,r_1-1} = 0$ for all $u \in \{1, \ldots, k\}$. Repeating the above procedure we find that $\lambda_{u,j} = 0$ for all $u \in \{1, \ldots, m\}$ and $j \in \{1, \ldots, r_u - 1\}$, finishing Case 2.

With the above procedure we have in particular shown that $\lambda_{1,j} = 0$ for all $j \in \{0, \ldots, r_1 - 1\}$, hence we can eliminate $x^{(1,j)}$ for $j = 0, \ldots, r_1 - 1$. Repeating the same steps, we can eliminate $x^{(2,j)}$ by showing that $\lambda_{2,j} = 0$ for all $j \in \{0, \ldots, r_2 - 1\}$ and so on. Finally, we arrive at $\lambda_{u,j} = 0$ for all $u \in \{1, \ldots, v\}$ and $j \in \{0, \ldots, r_u - 1\}$ showing linear independence of $x^{(u,j)}$, $u \in \{1, \ldots, v\}$, $j \in \{0, \ldots, v - 1\}$. □

Remark 6.27. What we did in this section is in close analogy to homogeneous linear differential equations with constant coefficients. Also there one sets up the characteristic polynomial, considers their roots and then has exponentials that involve the roots multiplied by a polynomial of degree less than the multiplicity of the root. Students who are familiar with ordinary differential equations will have recognised the similarities.

Remark 6.28. What happens if all coefficients $\varphi_1, \ldots, \varphi_p$ are real valued? Then one might be interested in real valued solutions $(x_n)_{n\in T}$ and not in complex ones. But it may nevertheless happen that the roots of the characteristic polynomial are complex. As an example, consider the difference equation $x_n = 2x_{n-1} - 2x_{n-2}$. Its characteristic polynomial is given by $\varphi(z) = 1 - 2z + 2z^2$, hence the zeroes of φ are given by

$$\xi_{1,2}^{-1} = \frac{2 \pm \sqrt{(-2)^2 - 4\cdot 2}}{2\cdot 2} = \frac{1 \pm \mathrm{i}}{2}.$$

By Theorem 6.26, the general (complex) solution is given by

$$x_n = \alpha_1 \left(\frac{1+\mathrm{i}}{2}\right)^{-n} + \alpha_2 \left(\frac{1-\mathrm{i}}{2}\right)^{-n}, \quad n \in T,$$

and that is not easily seen to be real valued. The trick now is to observe that with each solution $(x_n)_{n\in T}$ of (6.13) also $(\Re x_n)_{n\in T}$ and $(\Im x_n)_{n\in T}$ are solutions of (6.13) (since the φ_i are real valued), but those are real valued. We have

$$((1+\mathrm{i})/2)^{-n} = 2^{n/2}\left(\mathrm{e}^{\mathrm{i}\pi/4}\right)^n = 2^{n/2}\mathrm{e}^{\mathrm{i}\pi n/4},$$

hence

$$\Re((1+\mathrm{i})/2)^{-n} = 2^{n/2}\cos(\pi n/4), \quad \Im((1+\mathrm{i})/2)^{-n} = 2^{n/2}\sin(\pi n/4).$$

Similarly,

$$\Re((1-\mathrm{i})/2)^{-n} = 2^{n/2}\cos(-\pi n/4), \quad \Im((1+\mathrm{i})/2)^{-n} = 2^{n/2}\sin(-\pi n/4).$$

Observe that considering the second solution does not give new linearly independent solutions when considering the real and the imaginary part, the reason is that the ξ_1^{-1} and ξ_2^{-1} are complex conjugates. One can now easily show that one gets all complex solutions (x_n) as

$$x_n = \beta_1 2^{n/2} \cos(\pi n/4) + \beta_2 2^{n/2} \sin(\pi n/4), \quad n \in T$$

with $\beta_1, \beta_2 \in \mathbb{C}$ and all real valued solutions with $\beta_1, \beta_2 \in \mathbb{R}$. Similar arguments can be made for general homogeneous linear difference equations with constant real coefficients, the key observation being that the complex zeroes with non-zero imaginary part occur then in pairs of complex conjugate zeroes. We omit the details.

6.4 Three methods to compute the autocovariance function of a causal ARMA process

In this section we will provide three methods to compute the autocovariance function of a causal ARMA(p, q) process.

6.4.1 First method: calculation of the MA(∞) representation

Let $(Z_t)_{t\in\mathbb{Z}} \sim WN(0, \sigma^2)$ and $(X_t)_{t\in\mathbb{Z}}$ be a causal ARMA(p, q) process satisfying (6.1) with characteristic polynomials $\varphi(z)$ and $\theta(z)$. Suppose that φ and θ have no common zeroes on S^1 (so that $\varphi(z) \neq 0$ for all $z \in S^1$). By Theorem 6.10 and causality, we have

$$\psi(z) = \sum_{j=0}^{\infty} \psi_j z^j = \frac{\theta(z)}{\varphi(z)} \quad \forall\, z \in S^1$$

for some absolutely summable coefficients $(\psi_j)_{j\in\mathbb{N}_0}$. Rewriting this gives $\varphi(z)\psi(z) = \theta(z)$ for all $z \in S^1$, i.e.

$$(1 - \varphi_1 z - \ldots - \varphi_p z^p) \sum_{j=0}^{\infty} \psi_j z^j = 1 + \theta_1 z + \ldots + \theta_q z^q \quad \forall\, z \in S^1.$$

Set $\theta_0 := 1$, $\theta_j := 0$ for $j > q$ and $\varphi_j = 0$ for $j > p$, then we have (by multiplying it out and comparing the coefficients which is valid by Theorem 5.9)

$$\psi_j - \sum_{k=1}^{j} \varphi_k \psi_{j-k} = \theta_j, \quad 0 \leq j < \max(p, q+1) \tag{6.17}$$

and

$$\psi_j = \sum_{k=1}^{p} \varphi_k \psi_{j-k} \text{ for } j \geq \max(p, q+1). \tag{6.18}$$

From Theorem 6.26 we know that the general solution of (6.18) is (provided $\varphi_p \neq 0$; otherwise reduce the order)

$$\psi_j = \sum_{u=1}^{v} \sum_{m=0}^{r_u-1} \alpha_{u,m} j^m \xi_u^j, \quad j \geq \max(p, q+1) - p, \tag{6.19}$$

where ξ_u^{-1}, $u = 1, \dots, v$, are the distinct zeroes of $\varphi(z)$ and r_u their multiplicities. The ψ_j, $0 \leq j < \max(p, q+1)$, are recursively defined from (6.17). The coefficients $\alpha_{u,m}$ are then obtained from solving (6.19) in $\alpha_{u,m}$ for $j = \max(p, q+1) - p, \dots, \max(p, q+1) - 1$ since the corresponding ψ_j are known from (6.17). Then the ACVF is given by Theorem 5.3 (one can obtain explicit expressions from this by some calculations).

6.4.2 Second method: finding a recursion for the autocovariance function

Let $(Z_t)_{t\in\mathbb{Z}} \sim WN(0, \sigma^2)$ and $(X_t)_{t\in\mathbb{Z}}$ be a causal ARMA(p, q) process with expectation 0 satisfying (6.1) with characteristic polynomials $\varphi(z)$ and $\theta(z)$. Hence we have $\varphi(B)X_t = \theta(B)Z_t$, i.e.

$$X_t - \varphi_1 X_{t-1} - \ldots - \varphi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \ldots + \theta_q Z_{t-q}.$$

Multiplying this equation by $\overline{X}_{t-k}$ gives

$$(X_t - \varphi_1 X_{t-1} - \ldots - \varphi_p X_{t-p})\,\overline{X}_{t-k} = (Z_t + \theta_1 Z_{t-1} + \ldots + \theta_q Z_{t-q})\overline{X}_{t-k} \quad \forall\, t, k \in \mathbb{Z}.$$

Since the process is causal, there are absolutely summable coefficients $(\psi_j)_{j\in\mathbb{N}_0}$ such that $X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}$, hence also $\overline{X}_{t-k} = \sum_{j=0}^{\infty} \overline{\psi}_j \overline{Z}_{t-k-j}$. Plugging this into the right-hand side of the above equation gives

$$(X_t - \varphi_1 X_{t-1} - \ldots - \varphi_p X_{t-p})\overline{X}_{t-k} = (Z_t + \theta_1 Z_{t-1} + \ldots + \theta_q Z_{t-q}) \sum_{j=0}^{\infty} \overline{\psi}_j \overline{Z}_{t-k-j}.$$

Taking now the expectation of this equation results in (with $\theta_0 := 1$)

$$\gamma(k) - \varphi_1\gamma(k-1) - \ldots - \varphi_p\gamma(k-p) = \sigma^2 \sum_{j=k}^{q} \theta_j \overline{\psi}_{j-k}, \quad 0 \leq k \leq q, \tag{6.20}$$

and

$$\gamma(k) - \varphi_1\gamma(k-1) - \ldots - \varphi_p\gamma(k-p) = 0, \quad k \geq q+1. \tag{6.21}$$

Equation (6.20) actually is true not only for $0 \leq k \leq q$, but also for $0 \leq k \leq \max(p-1, q)$ (for if $p - 1 > q$ and $q < k \leq p - 1$ then $\sum_{j=k}^{q} \theta_j \overline{\psi}_{j-k} = 0$ as an empty sum). Hence we can also write

$$\gamma(k) - \varphi_1\gamma(k-1) - \ldots - \varphi_p\gamma(k-p) = \sigma^2 \sum_{j=k}^{q} \theta_j \overline{\psi}_{j-k}, \quad 0 \leq k \leq \max(p-1, q), \tag{6.22}$$

and

$$\gamma(k) - \varphi_1\gamma(k-1) - \ldots - \varphi_p\gamma(k-p) = 0, \quad k \geq \max(p, q+1). \tag{6.23}$$

The values of $\psi_0, \ldots, \psi_q$ in the boundary conditions (6.22) can be obtained recursively from (6.16), and one can show that the system of $\max(p, q+1)$ equations given by (6.22) in conjunction with $\gamma(-k) = \overline{\gamma(k)}$ has a unique solution (we shall not do this here; you can also calculate the starting values $\gamma(0), \ldots, \gamma(\max(p-1, q))$ with the third method to come in Section 6.4.3. For (6.23) we can then use the closed form solution

$$\gamma(h) = \sum_{u=1}^{v}\sum_{j=0}^{r_i-1} \beta_{u,j} h^j \xi_u^h, \ h \geq \max(p, q+1) - p,$$

of Theorem 6.26, ξ_u^{-1} are the pairwise different zeroes of $\varphi(z)$ and r_u their multiplicities. The coefficients $\beta_{u,j}$ have to be so determined so that the initial values are correct. The most important message from this method is that the ACVF γ satisfies itself a homogeneous difference equation with constant coefficients given by (6.23), and that we can give a general form of the solution.

6.4.3 Third method: the autocovariance generating function

Now we will learn a very elegant method which works more generally for stationary time series with absolutely summable autocovariances.

Definition 6.29. Let $X = (X_t)_{t\in\mathbb{Z}}$ be a stationary time series with autocovariance function γ. Assume that $\sum_{h=-\infty}^{\infty} |\gamma(h)| < \infty$, i.e. that $(\gamma(h))_{h\in\mathbb{Z}}$ is a linear filter. Then the associated filter function $G = G_X : S^1 \to \mathbb{C}$ of $(\gamma(h))_{h\in\mathbb{Z}}$, given by

$$G(z) = G_X(z) = \sum_{h=-\infty}^{\infty} \gamma(h) z^h$$

is called the *autocovariance generating function of X.*

Remark 6.30. By Theorem 5.9, the autocovariance generating function G of X determines the autocovariance function uniquely. Even more, the proof of Theorem 5.9 shows that

$$\gamma(h) = \frac{1}{2\pi}\int_0^{2\pi} G(\mathrm{e}^{\mathrm{i}s})\mathrm{e}^{-\mathrm{i}hs}\,\mathrm{d}s \quad \forall\, h \in \mathbb{Z}.$$

In here lies the significance of the autocovariance generating function. If we can calculate G explicitly, we also have an expression for the autocovariance function.

For infinite moving average processes the autocovariances are absolutely summable and we can express the autocovariance generating function in terms of the filter ψ.

Theorem 6.31. *Let $Z = (Z_t)_{t\in\mathbb{Z}} \sim WN(0, \sigma^2)$ and $(\psi_j)_{j\in\mathbb{Z}}$ a linear filter with associated filter function ψ. Consider the stationary process $X = (X_t)_{t\in\mathbb{Z}}$ defined by $X_t = \psi(B)Z_t$. Then the autocovariance function of X is absolutely summable and the autocovariance generating function of X is given by*

$$G_X(z) = \sigma^2|\psi(z)|^2 \quad \forall\, z \in S^1.$$

Proof. From Corollary 5.4 we have

$$\gamma(h) = \sigma^2 \sum_{j=-\infty}^{\infty} \psi_j \overline{\psi}_{j-h} \quad \forall\, h \in \mathbb{Z}.$$

This implies

$$\sum_{h\in\mathbb{Z}} |\gamma(h)| \le \sigma^2 \sum_{h\in\mathbb{Z}} \sum_{j\in\mathbb{Z}} |\psi_j \overline{\psi_{j-h}}| = \sigma^2 \sum_{j\in\mathbb{Z}} \sum_{m\in\mathbb{Z}} |\psi_j \overline{\psi_m}| = \sigma^2 \sum_{j\in\mathbb{Z}} |\psi_j| \sum_{m\in\mathbb{Z}} |\psi_m| < \infty.$$

Now let $z \in S^1$. Then

$$\begin{aligned} G_X(z) &= \sum_{h\in\mathbb{Z}} \gamma(h) z^h \\ &= \sigma^2 \sum_{h\in\mathbb{Z}} \sum_{j\in\mathbb{Z}} \psi_j \overline{\psi_{j-h}} z^h \\ &= \sigma^2 \sum_{j\in\mathbb{Z}} \psi_j z^j \sum_{h\in\mathbb{Z}} \overline{\psi_{j-h}} z^{h-j} \\ &\overset{m=j-h}{=} \sigma^2 \sum_{j\in\mathbb{Z}} \psi_j z^j \overline{\sum_{m\in\mathbb{Z}} \psi_m \overline{z}^{-m}} \\ &= \sigma^2 \psi(z) \overline{\psi(\overline{z}^{-1})}; \end{aligned}$$

the interchange of the summations is allowed by the previous calculation which implied absolute summability of the autocovariance. But since $z \in S^1$ we have $\overline{z}z = 1$, hence $\overline{z}^{-1} = z$, so that we obtain

$$G_X(z) = \sigma^2 \psi(z) \overline{\psi(z)} = \sigma^2 |\psi(z)|^2$$

for all $z \in S^1$. □

Corollary 6.32. *Let $Z = (Z_t)_{t\in\mathbb{Z}} \sim WN(0, \sigma^2)$ and $X = (X_t)_{t\in\mathbb{Z}}$ an $ARMA(p, q)$ process satisfying* (6.1) *with characteristic polynomials $\varphi(z)$ and $\theta(z)$ such that $\varphi(z) \neq 0$ for all $z \in S^1$. Then the autocovariance generating function G_X is given*

$$G_X(z) = \sigma^2 \frac{|\theta(z)|^2}{|\varphi(z)|^2} \quad \forall\, z \in S^1.$$

Further, for each $h \in \mathbb{Z}$ the autocovariance function $\gamma_X(h)$ at lag h is given by

$$\gamma_X(h) = \frac{\sigma^2}{2\pi} \int_0^{2\pi} \left| \frac{\theta(\mathrm{e}^{\mathrm{i}s})}{\varphi(\mathrm{e}^{\mathrm{i}s})} \right|^2 \mathrm{e}^{-\mathrm{i}hs} \, \mathrm{d}s.$$

Proof. By Theorem 6.10, X has a two-sided moving average representation with filter function $\psi(z) = \frac{\theta(z)}{\varphi(z)}$. The claim then follows immediately from Theorem 6.31 and Remark 6.30. $\square$

While Corollary 6.32 gives a nice formula, it does not easily provide asymptotics of $\gamma_X(h)$ as $h \to \infty$. For the qualitative behaviour including such asymptotics, the method developed in Section 6.4.2 has advantages.

Remark 6.33. If X is a weakly stationary sequence with absolutely summable autocovariance function γ, then the function $f : [-\pi, \pi] \to \mathbb{C}$ defined by

$$f(s) = \frac{1}{2\pi} \sum_{h \in \mathbb{Z}} \gamma(h) \mathrm{e}^{-\mathrm{i}hs} = \frac{1}{2\pi} G_X(\mathrm{e}^{-\mathrm{i}s})$$

is also called the *spectral density* of the process X. The spectral density of an ARMA process is hence given by $s \mapsto \frac{\sigma^2}{2\pi} \left| \frac{\theta(\mathrm{e}^{-is})}{\varphi(\mathrm{e}^{-\mathrm{i}s})} \right|^2$.

Chapter 7

Linear prediction

We now come to the important problem of how to predict a time series (we use the words "predict" and "forecast" with the same meaning). In practice, you have data, say $x_1, \ldots, x_n$, and would like to predict into the future, say you want to have a rough idea of x_{n+1} or of x_{n+10}. How would one do that?

Well, the structure in the previous data $x_1, \ldots, x_n$ should play a role, the prediction should be different when the data come from an AR(1) process than when they come from i.i.d. observations. So as a first step, one should fit some model to the data, e.g. an ARMA model. How this is to be done will be treated later. Usually, such a fitting proceedure is done using the best ARMA model that satisfies certain constraints. But then it could be that no ARMA model at all fits the data well. Hence in a second step one has to check if the models fits the data well (I am not sure if we do this later, it is called diagnostic checking). But now, suppose that one has fitted a model and one is reasonably convinced that it is a good fit. And the idea now is to find within this model a forecasting technique. I.e. if $(X_t)_{t\in\mathbb{Z}}$ is the found model (e.g. an ARMA(1,1) process), find functions $f_n : \mathbb{R}^n \to \mathbb{R}$ and $g_n : \mathbb{R}^n \to \mathbb{R}$ such that $f_n(X_1, \ldots, X_n)$ is a reasonable forecast for X_{n+1}, and $g_n(X_1, \ldots, X_n)$ is a reasonable forecast for X_{n+10}. This is purely probabilistic without data (this chapter). The new forecasts (or prediction) are then random variables, say $f_n(X_1, \ldots, X_n)$ or $g_n(X_1, \ldots, X_n)$. But I do not want a random variable as a forecast, but a number. Hence I insert the observations $x_1, \ldots, x_n$ into these functions, and say that the data forecast is $f_n(x_1, \ldots, x_n)$ or $g_n(x_1, \ldots, x_n)$, respectively. Let us summarise this again:

Method 7.1. Given data $x_1, \ldots, x_n$, which are a realisation of a stationary time series, I would like to forecast x_{n+1}, or x_{n+10}. This is done using the following steps:

- Step 1: Fit a model to the data, for example an ARMA model (later).
- Step 2: Check if the model fits well to the data (possibly later).
- Step 3: Within the model, find a forecasting technique. I.e. if $(X_t)_{t\in\mathbb{Z}}$ is the found model (e.g. an ARMA(1,1) process), find functions $f_n : \mathbb{R}^n \to \mathbb{R}$ and $g_n : \mathbb{R}^n \to \mathbb{R}$ such that $f_n(X_1, \ldots, X_n)$ is a reasonable forecast for X_{n+1}, and $g_n(X_1, \ldots, X_n)$

is a reasonable forecast for X_{n+10}. This is purely probabilistic without data (this chapter).

- Step 4: The forecasts of the data are then $f_n(x_1, \dots, x_n)$ and $g_n(x_1, \dots, x_n)$, respectively.

Let us look at a first example to illustrate this idea:

Example 7.2. Suppose data $x_1, \dots, x_n$ are given, and there is good reason to think that the data come from a causal AR(2) series (by looking at the ACF and the PACF (the partial autocorrelation function, to be defined in this chapter). So we believe that the underlying model is

$$X_t = \varphi_1 X_{t-1} + \varphi_2 X_{t-2} + Z_t, \quad t \in \mathbb{Z}.$$

Now estimate the parameters φ_1 and φ_2 (later). It seems plausible that for an AR(2) model the forecast of X_{n+1} based on $X_1, \dots, X_n$ should be

$$f_n(X_1, \dots, X_n) = \varphi_1 X_n + \varphi_2 X_{n-1}$$

(true if the model is causal, this chapter). What a good forecast for X_{n+10} is, is not so clear (but also this chapter). Now the one-step ahead forecast of the data is

$$f_n(x_1, \dots, x_n) = \varphi_1 x_n + \varphi_2 x_{n-1}$$

(with the estimated φ_1 and φ_2, provided it is causal).

So the contents in this chapter will be to do forecasting within stationary models, and we assume in this chapter that the model is known, so e.g. that we have a given ARMA(1,1) process with known parameters and known variance of the noise. We will mainly do linear forecasting, what "linear" means will become clear later. Observe again that instead of forecasting also the name prediction is used.

Now the question of this chapter is: Given a model $(X_t)_{t\in\mathbb{Z}}$ for a stationary time series, what is a good definition of the forecast of X_{n+h} given $X_1, \dots, X_n$? The answer is via orthogonal projections in Hilbert spaces. Let us first collect some facts on Hilbert spaces:

7.1 Hilbert spaces

You hopefully have encountered Hilbert spaces in your earlier studies. We collect however the main definitions and facts needed for us. We will only treat complex Hilbert spaces.

Definition 7.3. (a) A *(complex) inner product space* is a pair $(H, \langle\cdot,\cdot\rangle)$, where H is a (complex) vector space and $\langle\cdot,\cdot\rangle_H = \langle\cdot,\cdot\rangle : H \times H \to \mathbb{C}$ an *inner product* (also called *scalar product*), i.e. it holds that

i) $\langle X, Y\rangle = \overline{\langle Y, X\rangle}\ \forall X, Y \in H$,

ii) $\langle \alpha X + \beta Y, Z\rangle = \alpha \langle X, Z\rangle + \beta \langle Y, Z\rangle\ \forall X, Y, Z \in H,\ \alpha, \beta \in \mathbb{C}$,

iii) $\langle X, X\rangle \geq 0 \ \forall X \in H$,

iv) $\langle X, X\rangle = 0 \iff X = 0$.

An inner product space is also called a *(complex) pre-Hilbert space.*

(b) For every inner product space the *associated norm* $\|\cdot\| = \|\cdot\|_H : H \to [0, \infty)$ is defined by

$$\|X\| := ||X||_H := \sqrt{\langle X, X\rangle}.$$

It is easily checked that $\|\cdot\|_H$ satisfies indeed the properties of a norm, i.e. that

(i) $\|X\|_H \in [0, \infty)$ for all $X \in H$,

(ii) $\|\lambda X\|_H = |\lambda| \, \|X\|_H$ for all $X \in H$ and $\lambda \in \mathbb{C}$,

(iii) $\|X + Y\|_H \leq \|X\|_H + \|Y\|_H$ for all $X, Y \in H$ (triangle inequality),

(iv) $\|X\|_H = 0$ if and only if $X = 0$.

(Well, the triangular inequality is not so trivial, it uses the Cauchy-Schwarz inequality which we will show below in Theorem 7.5 (a) for a general inner product space).

(c) A sequence $(X_n)_{n\in\mathbb{N}}$ in H is said to *converge (in H)* if there exists some $X \in H$ such that $||X_n - X||_H \to 0$ as $n \to \infty$, i.e. such that to every $\varepsilon > 0$ there exists some $N_\varepsilon \in \mathbb{N}$ such that $\|X_n - X\|_H \leq \varepsilon$ for all $n \geq N_\varepsilon$. We then write $\lim_{n\to\infty} X_n = X$ or $X_n \to X$ as $n \to \infty$. It is easily checked that the limit is unique (provided it exists).

(d) A sequence $(X_n)_{n\in\mathbb{N}}$ in H is a *Cauchy sequence*, if for all $\varepsilon > 0$ there exists some $N_\varepsilon \in \mathbb{N}$ such that $||X_n - X_m||_H \leq \varepsilon$ for all $n, m \geq N_\varepsilon$. By the triangle inequality, every convergent sequence is also a Cauchy sequence.

(e) A *(complex) Hilbert space*, abbreviated as *HS*, is an inner product space $(H, \langle\cdot,\cdot\rangle)$ such that the associated norm $\|\cdot\|_H$ is *complete*, i.e. such that every Cauchy sequence converges in H.

Usually we will simply speak of a Hilbert space H if the underlying inner product $\langle\cdot,\cdot\rangle$ is clear. The following examples are the most important ones in our context:

Example 7.4. a) $(\mathbb{C}^n, \langle\cdot,\cdot\rangle_2)$ is a Hilbert space with $\langle x, y\rangle_2 := \sum_{i=1}^n x_i \overline{y_i}$ (known from calculus).

b) If $(\Omega, \mathcal{F}, P)$ is a probability space , then $\mathcal{L}^2(\Omega, \mathcal{F}, P)$ is the set of all random variables $X : \Omega \to \mathbb{C}$ with $\mathbb{E}\,|X|^2 = \int_\Omega |X|^2 \, dP < \infty$, and $L^2 = L^2(\Omega, \mathcal{F}, P)$ is the set of the equivalence classes given by the identification of P-a.s. equal random variables. If one sets for $X, Y \in L^2$

$$\langle X, Y\rangle := \mathbb{E}\,(X\overline{Y}),$$

then $(L^2, \langle\cdot,\cdot\rangle)$ is a Hilbert space (Proof: Measure theory). It holds

$$\mathrm{Cov}\,(X, Y) = \langle X - \mathbb{E}(X), Y - \mathbb{E}(Y)\rangle\,.$$

Next we give some elementary properties of Hilbert spaces (actually, an inner product space is enough):

Theorem 7.5. *Let $(H, \langle, \cdot, \cdot\rangle)$ be a Hilbert space (or only an inner product space) with norm $||\cdot||$. Then the following are true:*

(a) The Cauchy-Schwarz inequality is valid, i.e.

$$|\langle X, Y\rangle| \leq \sqrt{|\langle X, X\rangle|}\sqrt{|\langle Y, Y\rangle|} = \|X\|_H \, \|Y\|_H \quad \forall\, X, Y \in H.$$

(b) Let $X_n, Y_n \in H$ ($n \in \mathbb{N}$), $X, Y \in H$ and assume that $X_n \to X$, $Y_n \to Y$ (in $(H, ||\cdot||)$) as $n \to \infty$. Then

$$||X_n|| \to ||X|| \quad \text{and} \quad \langle X_n, Y_n\rangle \to \langle X, Y\rangle \quad \text{as } n \to \infty.$$

(c) The parallelogram law holds, i.e. for all $X, Y \in H$ we have

$$||X+Y||^2 + ||X-Y||^2 = 2||X||^2 + 2||Y||^2.$$

Proof. (a) Let $X, Y \in H$ and $\alpha \in \mathbb{C}$. Then

$$0 \leq \langle X - \alpha Y, X - \alpha Y\rangle = \langle X, X\rangle - \alpha\langle Y, X\rangle - \overline{\alpha}\langle X, Y\rangle + |\alpha|^2\langle Y, Y\rangle.$$

Choose $\theta \in [0, 2\pi)$ and $b \in [0, \infty)$ such that $\langle Y, X\rangle = b\mathrm{e}^{\mathrm{i}\theta}$, and denote $\alpha_t := \mathrm{e}^{-\mathrm{i}\theta}t$ for general $t \in \mathbb{R}$. Then the inequality above gives

$$\begin{aligned} 0 &\leq \langle X, X\rangle - \mathrm{e}^{-\mathrm{i}\theta}tb\mathrm{e}^{\mathrm{i}\theta} - \mathrm{e}^{\mathrm{i}\theta}tb\mathrm{e}^{-\mathrm{i}\theta} + t^2\langle Y, Y\rangle \\ &= \langle X, X\rangle - 2bt + t^2\langle Y, Y\rangle =: q(t). \end{aligned}$$

The right hand side $q(t)$ is a quadratic polynomial in the real variable t, and the left hand side is non-negative. Assume for the moment that $|\langle X, Y\rangle| > \langle X, X\rangle\,\langle Y, Y\rangle$ and $Y \neq 0$ (hence $\langle Y, Y\rangle > 0$). Then, since $|\langle X, Y\rangle| = b$, also $4b^2 > 4\langle X, X\rangle\langle Y, Y\rangle$, showing that the quadratic polynomial $q(t)$ has the two different real roots

$$t_{1,2} := \frac{2b \pm \sqrt{4b^2 - 4\langle X, X\rangle\langle Y, Y\rangle}}{2\langle Y, Y\rangle},$$

and between t_1 and t_2 the function q must be negative, contradicting the above equation. Hence we have $|\langle X, Y\rangle| \leq \|X\|_H\|Y\|_H$ if $Y \neq 0$. If $Y = 0$, then

$$\langle X, Y\rangle = \langle X, 2Y\rangle = 2\langle X, Y\rangle$$

showing that $\langle X, Y\rangle = 0$, so that Cauchy-Schwarz inequality holds also in this case.

(b) Observe first that if $X_n \to X$, then $\|X_n\| \leq \|X_n - X\| + \|X\| \leq \varepsilon + \|X\|$ for large enough n, so that $(\|X_n\|)_{n\in\mathbb{N}}$ must be bounded. Similarly, $(\|Y_n\|)_{n\in\mathbb{N}}$ must be bounded. Now we write

$$\begin{aligned} \langle X_n, Y_n\rangle - \langle X, Y\rangle &= \langle X_n, Y_n\rangle - \langle X_n, Y\rangle + \langle Y_n, Y\rangle - \langle X, Y\rangle \\ &= \langle X_n, Y_n - Y\rangle + \langle X_n - X, Y\rangle. \end{aligned}$$

The Cauchy-Schwarz inequality then implies

$$|\langle X_n, Y_n\rangle - \langle X, Y\rangle| \leq \|X_n\| \, \|Y_n - Y\| + \|X_n - X\| \, \|Y\|.$$

But $\|X_n - X\| \to 0$ and $\|Y_n - Y\| \to 0$ as $n \to \infty$ by assumption, and $\|X_n\|$ remains bounded as seen above. This shows that $\langle X_n, Y_n\rangle \to \langle X, Y\rangle$ as $n \to \infty$. Choosing $Y = X$ and taking the square root we also obtain $\|X_n\| \to \|X\|$ as $n \to \infty$.

(c) This follows easily from

$$\begin{aligned} \|X+Y\|^2 + \|X-Y\|^2 &= \langle X+Y, X+Y\rangle + \langle X-Y, X-Y\rangle \\ &= \langle X, X\rangle + \langle X, Y\rangle + \langle Y, X\rangle + \langle Y, Y\rangle \\ &\quad + \langle X, X\rangle - \langle X, Y\rangle - \langle Y, X\rangle + \langle Y, Y\rangle \\ &= 2\|X\|^2 + 2\|Y\|^2. \end{aligned}$$

□

We now come to the concept of closed sets and the orthogonal complement:

Definition 7.6. Let $(H, \langle\cdot,\cdot\rangle)$ be a Hilbert space (or only an inner product space).
(a) A subset $M \subset H$ is called *closed*, if from $x_n \in M$, $x \in H$ and $||x_n - x|| \to 0$ $(n \to \infty)$ follows that $x \in M$.
(b) The *orthogonal complement* $M^\perp$ of a subset M of H is defined by

$$M^\perp := \{x \in H : \langle x, y\rangle = 0 \; \forall y \in M\},$$

i.e. the set of the elements in H that are perpendicular to all elements of M (we call two elements $x, y \in H$ *perpendicular* or *orthogonal*, if $\langle x, y\rangle = 0$). We also write

$$x \perp M$$

to indicate that $\langle x, y\rangle = 0$ for all $y \in M$, i.e. for $x \in M^\perp$ and say that x is *perpendicular* or *orthogonal to M*.

In finite dimensional Hilbert spaces or even finite dimensional inner product spaces, every linear subspace (i.e. subvector space) is automatically closed. This is no longer true in general Hilbert spaces with infinite dimension (exercises). However, the orthogonal complement of a set will always be a closed linear subspace:

Lemma 7.7. *Let $(H, \langle\cdot,\cdot\rangle)$ be a Hilbert space (or only an inner product space). Then for every subset $M \subset H$ of H the set $M^\perp$ is a closed, linear subspace of H.*

Proof. To show that it is a linear subspace, let $x_1, x_2 \in M^\perp$, $\alpha \in \mathbb{C}$. Then

$$\langle \alpha x_1 + x_2, y\rangle = \alpha \langle x_1, y\rangle + \langle x_2, y\rangle = 0 \quad \forall y \in M,$$

hence $\alpha x_1 + x_2 \in M^\perp$, so that $M^\perp$ is a linear subspace.

Now let $x_n \in M^\perp$, $x \in H$ and assume that $||x_n - x|| \to 0$ as $n \to \infty$. Then for all $y \in M$ it holds that

$$\langle x, y\rangle \overset{\text{Thm. 7.5 (b)}}{=} \lim_{n\to\infty} \langle x_n, y\rangle = 0,$$

hence $x \in M^\perp$ so that $M^\perp$ is closed. □

We now come to a fundamental theorem, namely the orthogonal projection theorem. This only holds for Hilbert spaces but not for general inner product spaces.

Theorem 7.8 (Orthogonal projection theorem)**.** *Let $M \subset H$ be a closed linear subspace of the Hilbert space H and let $x \in H$. Then the following hold true:*

(a) There exists a unique element $\widehat{x} \in M$ such that $||x - \widehat{x}|| = \inf_{y \in M} ||x - y||$.

(b) The element $\widehat{x}$ is the unique element in M such that $x - \widehat{x} \in M^\perp$, i.e. for $z \in H$ we have $z = \widehat{x}$ if and only if $z \in M$ and $x - z \in M^\perp$.

The element $\widehat{x}$ is called orthogonal projection of x onto M.

Proof. (a) To show the existence, let $d := \inf_{y \in M} ||x - y||^2$. By definition of the infimum this implies the existence of a sequence $(y_n) \subset M$ with $||y_n - x||^2 \to d$ as $n \to \infty$. Since $(y_m + y_n)/2 \in M$ by the subspace property, we conclude from the parallelogram law (Theorem 7.5 (c)) that

$$\begin{aligned} 0 \leq ||y_m - y_n||^2 &= -4||(y_m + y_n)/2 - x||^2 + 2||y_n - x||^2 + 2||y_m - x||^2 \\ &\leq -4d + 2(||y_n - x||^2 + ||y_m - x||^2) \to 0 \text{ for } n, m \to \infty. \end{aligned}$$

This shows that $(y_n)_{n \in \mathbb{N}}$ is a Cauchy sequence, and by completeness (here we need that it is a Hilbert space!) there exists an $\widehat{x} \in H$ with $||y_n - \widehat{x}|| \to 0$. Since M is closed, $\widehat{x} \in M$. From Theorem 7.5 (b) we then have: $||x - \widehat{x}||^2 = \lim_{n \to \infty} ||x - y_n||^2 = d$, showing the existence of the element with the required property.

To show uniqueness, let $\widehat{y} \in M$ be a further element such that $||x - \widehat{y}||^2 = ||x - \widehat{x}||^2 = d$ and $\widehat{x}$ as before. Then by the parallelogram law,

$$\begin{aligned} 0 \leq ||\widehat{x} - \widehat{y}||^2 &= -4||(\widehat{x} + \widehat{y})/2 - x||^2 + 2(||\widehat{x} - x||^2 + ||\widehat{y} - x||^2) \\ &\leq -4d + 2(d + d) = 0 \implies \widehat{y} = \widehat{x}. \end{aligned}$$

(b) For one direction, let $\widehat{y} \in M$ with $x - \widehat{y} \in M^\perp$. Then it holds that for all $y \in M$

$$\begin{aligned} ||x - y||^2 &= \langle x - \widehat{y} + \widehat{y} - y, x - \widehat{y} + \widehat{y} - y \rangle \\ &= \langle x - \widehat{y}, x - \widehat{y} \rangle + \langle \widehat{y} - y, \widehat{y} - y \rangle \geq ||x - \widehat{y}||^2, \end{aligned}$$

so that $\widehat{y}$ has the properties of $\widehat{x}$ from (a), so that $\widehat{x} = \widehat{y}$ by the uniqueness.

For the converse direction, let $\widehat{x}$ be the element constructed in (a). Then $\widehat{x} \in M$ and we only have to show that $x - \widehat{x} \in M^\perp$. Assume for the moment that $x - \widehat{x} \notin M^\perp$. Then there exists $y \in M$ with $\langle x - \widehat{x}, y \rangle =: a \neq 0$ (in particular $y \neq 0$). Define $\widetilde{x} = \widehat{x} + ay/||y||^2$. Then

$$\begin{aligned} ||x - \widetilde{x}||^2 &= \langle x - \widehat{x} + \widehat{x} - \widetilde{x}, x - \widehat{x} + \widehat{x} - \widetilde{x} \rangle \\ &= ||x - \widehat{x}||^2 + \frac{|a|^2}{||y||^2} + 2\,\Re\,\langle x - \widehat{x}, \widehat{x} - \widetilde{x} \rangle \\ &= ||x - \widehat{x}||^2 - \frac{|a|^2}{||y||^2} < ||x - \widehat{x}||^2, \end{aligned}$$

which is a contradiction to the definition of $\widehat{x}$ since $\widetilde{x} \in M$. □

Given a closed linear subspace M of a Hilbert space, we can associate to every $x \in H$ its orthogonal projection onto M. This defines a mapping:

Definition 7.9. Let H be a Hilbert space and M be a closed subspace. Denote for every $x \in H$ by $P_M x = P_M(x) := \widehat{x} \in M$ the unique element from Theorem 7.8 with

$$\|x - P_M(x)\| = \inf_{y \in M} \|x - y\|.$$

Then $P_M : H \to M$ is a mapping and called *orthogonal projection onto* M, or sometimes simply the *orthoprojection onto* M.

In the following we collect some properties of P_M:

Proposition 7.10. *Let H be a Hilbert space and M be a closed subspace. Then the orthogonal projection P_M onto M has the following properties:*

i) P_M is linear, i.e. $P_M(\alpha x + \beta y) = \alpha P_M(x) + \beta P_M(y)$ $\forall x, y \in H$ and $\alpha, \beta \in \mathbb{C}$.

ii) $||x||^2 = ||P_M x||^2 + ||(\mathrm{Id} - P_M)x||^2$ $\forall x \in H$, where $\mathrm{Id} : H \to H$ denotes the identity on H given by $\mathrm{Id}(x) = x$ for all $x \in H$.

iii) If $x_n \to x$ in H as $n \to \infty$, then $P_M x_n \to P_M x$ in H as $n \to \infty$, i.e. P_M is continuous.

iv) For $x \in H$ we have $x \in M$ if and only if $P_M x = x$.

v) For $x \in H$ we have $x \in M^\perp$ if and only if $P_M x = 0$.

Proof. (i) $\alpha P_M x + \beta P_M y \in M$ is clear and for every $z \in M$ it holds true that

$$\langle \alpha x + \beta y - \alpha P_M x - \beta P_M y, z \rangle = \alpha \langle x - P_M x, z \rangle + \beta \langle y - P_M y, z \rangle = 0$$

Theorem 7.8 (b) then implies that $\alpha P_M x + \beta P_M y = P_M(\alpha x + \beta y)$.

(ii) We have $x = P_M x + (\mathrm{Id} - P_M)x$. Since $(\mathrm{Id} - P_M)x = x - \widehat{x} \in M^\perp$ and $P_M x \in M$ we have $\langle P_M x, x - P_M x \rangle = 0$. The claim then follows since

$$\|x\|^2 = \big\langle P_M x + (\mathrm{Id} - P_M x), P_M x + (\mathrm{Id} - P_M x) \big\rangle = \|P_M x\|^2 + \|(\mathrm{Id} - P_M)x\|^2 + 2\Re\langle P_M x, (\mathrm{Id} - P_M)x \rangle.$$

(iii) From (ii), we have $||P_M(x - x_n)||^2 \leq ||x - x_n||^2$ and the latter tends to 0 by assumption as $n \to \infty$.

(iv) This is clear by the definition of $P_M x = \widehat{x}$.

(v) Since $\widehat{x} - x \in M^\perp$ and $M^\perp$ is a subspace, it holds that $\widehat{x} \in M^\perp \iff x \in M^\perp$. Since $\widehat{x} \in M$ and $M \cap M^\perp = \{0\}$, the assertion follows. □

7.2 Best prediction and best linear prediction

We will soon define the best predictor and the best linear predictor as an orthogonal projection in $L^2(P)$ onto a suitable closed subspace. Let us first describe the subspace needed for the best linear prediction.

Definition 7.11. Let $(x_t)_{t\in T}$ be a family of vectors in a Hilbert space H. Denote by $\text{span}\{x_t : t \in T\}$ the set of all finite linear combinations (the span of $\{x_t, t \in T\}$) of the elements $\{x_t, t \in T\}$, and by

$$\overline{\text{span}}\{x_t, t \in T\}$$

the closure of the span (i.e. the intersection of all closed supersets of $\text{span}\{x_t, t \in T\}$).

Remark 7.12. (a) It is easily seen that $\overline{\text{span}}\{x_t, t \in T\}$ is a closed subspace of H. It is the smallest closed subspace which contains all elements x_t with $t \in T$.
(b) If T is finite, then $\overline{\text{span}}\{x_t, t \in T\} = \text{span}\{x_t, t \in T\}$, since it is easily seen that finite dimensional subspaces are always closed (exercise).
(c) If T is not finite, then it can happen that $\text{span}\{x_t : t \in T\}$ is not closed and hence that $\overline{\text{span}}\{x_t, t \in T\} \supsetneq \text{span}\{x_t, t \in T\}$ (exercise).

We can now define the best linear predictor and the best predictor:

Definition 7.13. Let $(X_t)_{t\in T}$ be a family in $L^2(\Omega, \mathcal{F}, P)$ and $X \in L^2(\Omega, \mathcal{F}, P)$.

(a) Denote by

$$\widetilde{M} := L^2(\Omega, \sigma(X_t : t \in T), P)$$

the space of all (equivalence classes) of square integrable random variables that are $\sigma(X_t : t \in T)$ measurable. It is a closed subspace of $L^2(\Omega, \mathcal{F}, P)$ (since L^2-limits of $\sigma(X_t : t \in T)$-measurable random variables are again $\sigma(X_t : t \in T)$-measurable). Then

$$P_{\widetilde{M}}X$$

is said to be the *(best) prediction of X given $(X_t)_{t\in T}$* or *(best) forecast of X given $(X_t)_{t\in T}$*.

(b) Denote

$$M := \overline{\text{span}}\{X_t, t \in T\}.$$

Then M is a closed subspace of $L^2(\Omega, \mathcal{F}, P)$, and $P_M X$ is called the *best linear prediction of X given $(X_t)_{t\in T}$* or *best linear forecast of X given $(X_t)_{t\in T}$*.

(c) The quantities $||X - P_{\widetilde{M}}X||^2$ or $||X - P_M X||^2$ are called *squared prediction error.*

Interpretation 7.14. (a) The intuition behind this definition is that a predictor should be a function of what has been already observed. The most general setting is that the predictor can be expressed as $f(X_t : t \in T)$ for some Borel measurable function f (defined on a suitable space). Those who know the factorisation lemma (e.g. from Financial Mathematics I) know that this is equivalent to saying that the predictor should be

$\sigma(X_t : t \in T)$-measurable. So the best predictor should be $\sigma(X_t : t \in T)$-measurable. But how do we say it is better than other elements? The fact that we require it to be in $L^2(\Omega, \sigma(X_t : t \in T), P)$ and to be the orthogonal projection onto the space $\widetilde{M}$ can be expressed as

$$\mathbb{E}|X - P_{\widetilde{M}}X|^2 = \inf_{Y \in \widetilde{M}} \mathbb{E}|X - Y|^2,$$

i.e. $P_{\widetilde{M}}$ is the unique element in $\widetilde{M}$ that minimises the squared prediction error, i.e. the L^2-difference to X. This makes $P_{\widetilde{M}}X$ a natural choice for the definition of a best predictor.
(b) We have explained why we consider $P_{\widetilde{M}}X$ as the best predictor. But why consider $P_M X$, the best linear predictor? Well, there one only considers elements as possible predictors that can be written as linear combinations of the $(X_t)_{t \in T}$ or that can be approximated by such linear combinations (ensured by taking the closure). The reason why we use the orthogonal projection is as for $P_{\widetilde{M}}X$, the best linear predictor minimises the mean squared prediction error among all possible candidates in M.

Remark 7.15. (a) In the setting of Definition 7.13, we have

$$P_{\widetilde{M}}X = \mathbb{E}(X | X_t : t \in T),$$

the conditional expectation of X given $(X_t)_{t \in T}$. To see this, observe that $P_{\widetilde{M}}X$ is $\sigma(X_t : t \in T)$-measurable. Now let $A \in \sigma(X_t : t \in T)$. Then $\mathbf{1}_A$ (the indicator function of A) is in $\widetilde{M}$, hence

$$\mathbb{E}(P_{\widetilde{M}}X \cdot \mathbf{1}_A) = \langle P_{\widetilde{M}}X, \mathbf{1}_A \rangle = \langle X, \mathbf{1}_A \rangle + \langle P_{\widetilde{M}}X - X, \mathbf{1}_A \rangle = \mathbb{E}(X \mathbf{1}_A)$$

since $\langle P_{\widetilde{M}}X - X, \mathbf{1}_A \rangle = 0$ by the definition of the orthogonal projection. But this shows that $P_{\widetilde{M}}X$ satisfies the defining properties of the conditional expectation.
(b) We have $M \subset \widetilde{M}$ (observe that $\operatorname{span}(X_t : t \in T) \subset \widetilde{M}$ and that $\widetilde{M}$ is closed), so $P_M X \in \widetilde{M}$ and $\|P_M X - X\|^2 \geq \|P_{\widetilde{M}}X - X\|^2$. This shows that $P_{\widetilde{M}}$ is in general better than $P_M X$.

The following example shows that the best predictor can be strictly better than the best linear predictor:

Example 7.16. Let X and Z be independent $N(0,1)$-distributed random variables and $Y := X^2 + Z$. Then the best prediction of Y given $\{X, 1\}$ is $\mathbb{E}(Y|X, 1) = X^2$, the squared prediction error is $\mathbb{E}|Y - X^2|^2 = \mathbb{E}|Z|^2 = 1$. On the other hand, the best linear predictor has the form $P_M Y = aX + b$ for some $a, b \in \mathbb{C}$, since it must be in $\overline{\operatorname{span}}\{X, 1\}$. Since $Y - P_M Y \in \{1, X\}^\perp$ it holds that

(i) $\langle aX + b, X \rangle = \langle P_M Y, X \rangle = \langle Y, X \rangle = \mathbb{E}(YX) = \mathbb{E}(X^3 + XZ) = 0$, and

(ii) $\langle aX + b, 1 \rangle = \langle P_M Y, 1 \rangle = \langle Y, 1 \rangle = \mathbb{E}Y = 1$.

From (ii) we obtain $\mathbb{E}(aX + b) = 1$, hence $b = 1$. From (i) we then obtain $a\mathbb{E}|X|^2 + b\mathbb{E}X = 0$, hence $a = 0$. We conclude that $P_M Y = 1$. The squared prediction error is

$$\begin{aligned} ||Y - P_M Y||^2 &= \mathbb{E}|X^2 - 1 + Z|^2 = \operatorname{Var}(X^2 + Z) = \operatorname{Var} X^2 + \operatorname{Var} Z \\ &= \mathbb{E}X^4 - (\mathbb{E}X^2)^2 + 1 = 3 - 1 + 1 = 3. \end{aligned}$$

This shows that the best predictor is strictly better than the best linear predictor here.

So we have seen that the conditional expectation is a better prediction then the best linear. Nevertheless, one often use the linear prediction because:

- it is easier to calculate
- it only depends on the first and second moment of the random variables (as we shall see later)
- for multivariate normal distribution the linear prediction coincides with the conditional expectation, as we shall see now:

Proposition 7.17. *Let $(X, X_1, \dots, X_n)'$ be multivariate normally distributed. Then the best predictor of X given $X_1, \dots, X_n$ is equal to the best linear predictor of X given $1, X_1, \dots, X_n$, i.e.*

$$P_{\overline{\text{span}}\{1,X_1,\dots,X_n\}}X = \mathbb{E}(X|X_1, \dots, X_n).$$

Proof. Let $\widehat{X} = P_{\overline{\text{span}}\{1,X_1,\dots,X_n\}}(X)$. Then $\widehat{X} - X \in (\overline{\text{span}}\{1, X_1, \dots, X_n\})^{\perp}$, i.e.

$$0 = \left\langle \hat{X} - X, 1 \right\rangle = \mathbb{E}\hat{X} - \mathbb{E}X \quad \text{and}$$
$$0 = \left\langle \hat{X} - X, X_i \right\rangle = \mathbb{E}((\hat{X} - X)X_i) \qquad \forall\, i = 1, \dots, n.$$

But this implies that

$$\text{Cov}\,(\widehat{X} - X, X_i) = \mathbb{E}((\widehat{X} - X)X_i) - \mathbb{E}(\widehat{X} - X)\mathbb{E}X_i = 0 \quad \forall\, i = 1, \dots, n.$$

As $(\widehat{X} - X, X_1, \dots, X_n)'$ is multivariate normally distributed, this implies that $\widehat{X} - X$ is independent of $(X_1, \dots, X_n)'$. Hence if $W \in L^2(\Omega, \sigma(X_1, \dots, X_n), P)$, then $\widehat{X} - X$ is independent of W. This implies that $\mathbb{E}((\widehat{X} - X)W) = \mathbb{E}(\widehat{X} - X)\mathbb{E}W = 0$, hence $\mathbb{E}(\widehat{X}W) = \mathbb{E}(XW)$. Choosing $W = \mathbf{1}_A$ with $A \in \sigma(X_1, \dots, X_n)$ then shows that $\widehat{X} = \mathbb{E}(X|X_1, \dots, X_n)$ since $\widehat{X} \in L^2(\Omega, \sigma(X_1, \dots, X_n), P)$. □

Proposition 7.17 is the main reason why the best linear prediction plays such a prominent role, because in practice one often thinks of having a Gaussian time series and prediction for that then reduces to linear prediction. This is similar to the fact that weak stationarity plays such a crucial role, because for Gaussian time series weak and strict stationarity coincide.

In most cases we will have time series with mean zero. How is the best linear prediction affected if we do not have mean zero time series? This is explained in the following result.

Proposition 7.18. *Let $X, X_1, \dots, X_n \in L^2$ with $\mathbb{E}X = \mathbb{E}X_1 = \ldots = \mathbb{E}X_n = \mu$ (same mean) and define $Y = X - \mu$, $Y_i = X_i - \mu$. Then it holds true that*

$$P_{\overline{\text{span}}\{1,X_1,\dots,X_n\}}X = \mu + P_{\overline{\text{span}}\{Y_1,\dots,Y_n\}}Y. \tag{7.1}$$

Hence w.l.o.g. we can restrict our attention to random variables with mean zero, i.e. to Y_t. Especially, it holds that

$$P_{\overline{\text{span}}\{1,Y_1,\dots,Y_n\}}Y = P_{\overline{\text{span}}\{Y_1,\dots,Y_n\}}Y. \tag{7.2}$$

Proof. Let

$$\widehat{Y} = P_{\overline{\text{span}}\{1,Y_1,\dots,Y_n\}}Y = b + a_1Y_1 + \dots a_nY_n \tag{7.3}$$

for some $b, a_1, \dots, a_n \in \mathbb{C}$; it must be of this form because

$$\overline{\text{span}}\{1, Y_1, \dots, Y_n\} = \text{span}\{1, Y_1, \dots, Y_n\} = \{b + a_1Y_1 + \dots + a_nY_n : b, a_1, \dots, a_n \in \mathbb{C}\}.$$

Since $Y - \widehat{Y}$ is in the orthogonal complement of $\overline{\text{span}}\{1, Y_1, \dots, Y_n\}$ we have in particular

$$0 = \left\langle Y - \widehat{Y}, 1 \right\rangle = \mathbb{E}(Y - \widehat{Y}) \implies b \overset{(7.3)}{=} \mathbb{E}\widehat{Y} = \mathbb{E}Y = 0.$$

But (7.3) then implies that $\widehat{Y} \in \overline{\text{span}}\{Y_1, \dots, Y_n\}$, and since

$$Y - \widehat{Y} \perp \overline{\text{span}}\{1, Y_1, \dots, Y_n\} \supset \overline{\text{span}}\{Y_1, \dots, Y_n\}$$

we see that

$$\widehat{Y} = P_{\overline{\text{span}}\{Y_1,\dots,Y_n\}}Y,$$

so that (7.2) holds true. Equation (7.1) now follows by the linearity of $P_{\overline{\text{span}}\{\dots\}}$ and observing that $\overline{\text{span}}\{1, X_1, \dots, X_n\} = \overline{\text{span}}\{1, Y_1, \dots, Y_n\}$. □

The linear prediction of a random variable based on $X_1, \dots, X_n$ will get a special name:

Definition 7.19. Let $X_1, \dots, X_n \in L^2(\Omega, \mathcal{F}, P)$ with $\mathbb{E}X_1 = \dots = \mathbb{E}X_n = 0$ (e.g. coming from a mean zero stationary time series). Then we set

$$P_n := P_{\overline{\text{span}}\{X_1,\dots,X_n\}}, \quad (n \in \mathbb{N}),$$

the orthogonal projection from $L^2(\Omega, \mathcal{F}, P)$ onto $\overline{\text{span}}\{X_1, \dots, X_n\}$. If $X_{n+h} \in L^2(\Omega, \mathcal{F}, P)$ with $\mathbb{E}X_{n+h} = 0$, $h \in \mathbb{N}$, then

$$\widehat{X}_{n+h}^{(h)} := P_nX_{n+h}$$

is called the *h-step predictor (forecast) of* X_{n+h} (given $X_1, \dots, X_n$). We set

$$\widehat{X}_{n+1} := P_nX_{n+1}$$

for the 1-*step predictor*. The squared error

$$v_n^{(h)} := ||\widehat{X}_{n+h}^{(h)} - X_{n+h}||^2$$

is called *squared $h-$step prediction error.*

Let us calculate the 1-step predictor of a causal AR(p) process:

Proposition 7.20. *Let $(Z_t)_{t\in\mathbb{Z}} \sim WN(0, \sigma^2)$ and let $(X_t)_{t\in\mathbb{Z}}$ be a causal AR(p) process satisfying*

$$X_t - \varphi_1X_{t-1} - \dots - \varphi_pX_{t-p} = Z_t, \quad t \in \mathbb{Z}.$$

Then for each $n \geq p$ the 1-step predictor $\widehat{X}_{n+1}$ of X_{n+1} given $X_1, \dots, X_n$ is given by

$$\widehat{X}_{n+1} = \varphi_1X_n + \dots + \varphi_pX_{n-p+1}.$$

The squared 1*-step prediction error is σ^2.*

Proof. Define $Y := \varphi_1 X_n + \ldots + \varphi_p X_{n-p+1}$ (then clearly $Y \in L^2$). To show that $Y = \widehat{X}_{n+1}$ we have to show that Y satisfies the conditions of Theorem 7.8 (b), i.e. that

(i) $Y \in \overline{\operatorname{span}}\{X_1, \ldots, X_n\}$ and that

(ii) $X_{n+1} - Y \perp \overline{\operatorname{span}}\{X_1, \ldots, X_n\}$.

Condition (i) is clear from the definition of Y since $n \geq p$. For the proof of (ii) observe that

$$X_{n+1} - Y = X_{n+1} - \varphi_1 X_n - \ldots - \varphi_p X_{n-p+1} = Z_{n+1}$$

by the AR(p) equation. But since $(X_t)_{t \in \mathbb{Z}}$ is causal, for each $k \in \{1, \ldots, n\}$ we can write $X_k = \sum_{j=0}^{\infty} \psi_j Z_{k-j}$ with absolutely summable coefficients $(\psi_j)_{j \in \mathbb{N}_0}$. This shows that for $k \in \{1, \ldots, n\}$ we have

$$\langle X_k, X_{n+1} - Y \rangle = \langle X_k, Z_{n+1} \rangle = \langle \sum_{j=0}^{\infty} \psi_j Z_{k-j}, Z_{n+1} \rangle = \sum_{j=0}^{\infty} \psi_j \underbrace{\langle Z_{k-j}, Z_{n+1} \rangle}_{=0 \text{ since } n+1 \neq k-j} = 0,$$

hence for an arbitrary element $a_1 X_1 + \ldots + a_n X_n \in \overline{\operatorname{span}}\{X_1, \ldots, X_n\}$ we get

$$\langle a_1 X_1 + \ldots + a_n X_n, X_{n+1} - Y \rangle = 0,$$

so that also condition (ii) is satisfied. Since the orthogonal projection is uniquely determined by these properties we see that $\widehat{X}_{n+1} = Y$. Since $\widehat{X}_{n+1} - X_{n+1} = Z_{n+1}$ as seen we further have $v_n(1) = \mathbb{E}|Z_{n+1}|^2 = \sigma^2$ for the squared 1-step prediction error. □

Remark 7.21. The causality assumption in Proposition 7.20 is crucial and was also used in the proof. Consider e.g. the non-causal AR(1) process

$$X_t - 2X_{t-1} = Z_t$$

with $(Z_t)_{t \in \mathbb{Z}} \sim WN(0, \sigma^2)$. As shown in the exercises, there exists some white noise sequence $(W_n)_{n \in \mathbb{Z}} \sim WN(0, \sigma^2/4)$ such that

$$X_t - \frac{1}{2} X_{t-1} = W_t, \quad t \in \mathbb{Z}.$$

Proposition 7.20 then implies that for $n \geq 1$ the one-step predictor of X_{n+1} given $X_1, \ldots, X_n$ is given by $\frac{1}{2} X_n$ with squared 1-step prediction error $\sigma^2/4$. In particular, the 1-step predictor is **not** equal to $2X_n$.

The determination of the h-step predictor of a weakly stationary time series can be reduced to solving a system of linear equations that involve only the autocovariances. More precisely, we have:

Theorem 7.22 (Solution of the prediction problem; prediction equations).
Let $(X_t)_{t \in \mathbb{Z}}$ be a (complex-valued) stationary time series with mean value 0 *and ACVF γ.*

Let $h, n \in \mathbb{N}$. Then a random variable Y is the h-step predictor of X_{n+h} given $\{X_1, \dots, X_n\}$ if and only if there are constants $\varphi_{n,1}^{(h)}, \dots, \varphi_{n,n}^{(h)} \in \mathbb{C}$ such that

$$Y = \varphi_{n,1}^{(h)} X_n + \ldots + \varphi_{n,n}^{(h)} X_1 = \sum_{i=1}^{n} \varphi_{n,i}^{(h)} X_{n+1-i} \tag{7.4}$$

and the prediction equations

$$\Gamma_n \vec{\varphi}_n^{(h)} = \vec{\gamma}_n^{(h)} \tag{7.5}$$

are satisfied, where

$$\begin{aligned} \Gamma_n &:= (\gamma(i-j))_{i,j=1,\dots,n} \in \mathbb{C}^{n\times n}, \\ \vec{\gamma}_n^{(h)} &:= (\gamma(h), \gamma(h+1), \dots, \gamma(h+n-1))' \in \mathbb{C}^n, \quad \textit{and} \\ \vec{\varphi}_n^{(h)} &:= (\varphi_{n,1}^{(h)}, \dots, \varphi_{n,n}^{(h)})' \in \mathbb{C}^n. \end{aligned}$$

If Γ_n is invertible, then (7.5) has a unique solution. If not then there exist infinitely many solutions $\vec{\varphi}_n^{(h)}$ of (7.5), but they all lead to the same random variable Y in (7.4) which is the h-step predictor $\widehat{X}_{n+h}^{(h)} = Y$. Finally, if Γ_n is invertible, then the squared h-step prediction error is given by

$$v_n^{(h)} = \gamma(0) - \overline{(\vec{\gamma}_n^{(h)})'}\, \Gamma_n^{-1}\, \vec{\gamma}_n^{(h)}.$$

Proof. Being the orthogonal projection of X_{n+h} onto $\overline{\operatorname{span}}\{X_1, \dots, X_n\} = \operatorname{span}\{X_1, \dots, X_n\}$, we know that the h-step predictor $\widehat{X}_{n+h}^{(h)}$ is unique and characterised by $\widehat{X}_{n+h}^{(h)} \in \operatorname{span}\{X_1, \dots, X_n\}$ and $X_{n+h} - \widehat{X}_{n+h}^{(h)} \perp \operatorname{span}\{X_1, \dots, X_n\}$. For a random variable Y this means that $Y = \widehat{X}_{n+h}^{(h)}$ if and only if there are coefficients $\varphi_{n,1}^{(h)}, \dots, \varphi_{n,n}^{(h)} \in \mathbb{C}$ such that (7.4) holds (meaning that $Y \in \operatorname{span}\{X_1, \dots, X_n\}$) and such that $X_{n+h} - Y \perp \operatorname{span}\{X_1, \dots, X_n\}$. But $X_{n+h} - Y \perp \operatorname{span}\{X_1, \dots, X_n\}$ is equivalent to $\langle X_{n+h} - Y, X_k \rangle = 0$ for all $k \in \{1, \dots, n\}$, since every element of $\operatorname{span}\{X_1, \dots, X_n\}$ can be written as a linear combination $\sum_{j=1}^n a_k X_k$ and using the bilinearity of the inner product. So Y with representation (7.4) is equal to $\widehat{X}_{n+h}^{(h)}$ if and only if

$$\langle X_{n+h} - Y, X_{n+1-j} \rangle = 0 \quad \forall\, j = 1, \dots, n.$$

Inserting (7.4) for Y, the latter is equivalent to

$$\left\langle \sum_{i=1}^{n} \varphi_{n,i}^{(h)} X_{n+1-i}, X_{n+1-j} \right\rangle = \langle X_{n+h}, X_{n+1-j} \rangle \qquad \forall j = 1, \dots, n,$$

i.e. to

$$\sum_{i=1}^{n} \varphi_{n,i}^{(h)} \gamma(j-i) = \gamma(h-1+j) \qquad \forall j = 1, \dots, n.$$

This is a system of n linear equations, which in matrix form can be rewritten as (7.5). This shows that $Y = \widehat{X}_{n+h}^{(h)}$ if and only if Y has representation (7.4) such that the prediction equations (7.5) are satisfied.

It is clear that (7.5) has a unique solution if Γ_n is invertible. If Γ_n is not invertible, then the h-step predictor exists nevertheless and is unique, and by the previous characterisation it follows that (7.5) must have a solution and that all its solutions must via (7.4) lead to the same Y. That there are infinitely many solutions of (7.5) if Γ_n is not invertible is clear from linear algebra.

Finally, suppose again that Γ_n is invertible. Since $\gamma(i-j) = \overline{\gamma(j-i)}$ we have $\Gamma_n = \overline{\Gamma_n'} \in \mathbb{C}^{n\times n}$. Also, $\vec{\varphi}_n^{(h)} = \Gamma_n^{-1}\vec{\gamma}_n^{(h)} \in \mathbb{C}^n$. For the squared h-step prediction error we obtain

$$\begin{aligned}
v_n^{(h)} &= \|X_{n+h} - \widehat{X}_{n+h}^{(h)}\|^2 \\
&\overset{\text{Prop. 7.10}}{=} \|X_{n+h}\|^2 - \|\widehat{X}_{n+h}^{(h)}\|^2 \\
&= \gamma(0) - \left\langle \sum_{i=1}^n \varphi_{n,i}^{(h)} X_{n+1-i}, \sum_{j=1}^n \varphi_{n,j}^{(h)} X_{n+1-j} \right\rangle \\
&= \gamma(0) - \sum_{i,j=1,\dots,n} \varphi_{n,i}^{(h)}\, \overline{\varphi_{n,j}^{(h)}}\, \gamma(j-i) \\
&= \gamma(0) - \overline{(\vec{\varphi}_n^{(h)})'}\, \Gamma_n\, \vec{\varphi}_n^{(h)} \\
&= \gamma(0) - \overline{(\vec{\gamma}_n^{(h)})'}\overline{(\Gamma_n^{-1})'}\, \Gamma_n\, \Gamma_n^{-1}\vec{\gamma}_n^{(h)} \\
&= \gamma(0) - \overline{(\vec{\gamma}_n^{(h)})'}\, \Gamma_n^{-1}\, \vec{\gamma}_n^{(h)}.
\end{aligned}$$

□

One may wonder why the coefficients in (7.4) are ordered in a reverse way, i.e. why write $\sum_{i=1}^n \varphi_{n,1}^{(h)} X_{n+1-i}$ in (7.4) and not define the coefficients to satisfy $\sum_{i=1}^n \varphi_{n,i}^{(h)} X_i$? The reason is that the coefficients refering to the recent past (X_n, X_{n-1}, etc.) should have a higher significance and hence they get the lower indices. Further, as we have seen in Proposition 7.20, this corresponds in a nice way to the ordering of the coefficients when considering causal AR(p)-processes.

As a corollary of the proof of Theorem 7.22 we find:

Corollary 7.23. *Let $(X_t)_{t\in\mathbb{Z}}$ be stationary with mean zero, let $n, h \in \mathbb{N}$ and let Y be a random variable of the form*

$$Y = \sum_{i=1}^n c_i X_{n+1-i}$$

such that

$$\operatorname{Cov}(X_{n+h} - Y, X_i) = 0 \quad \forall\, i = 1, \dots, n.$$

Then

$$\widehat{X}_{n+h}^{(h)} = P_n X_{n+h} = Y.$$

Proof. We have $M \in \operatorname{span}\{X_1, \dots, X_n\}$ and $\langle X_{n+h} - Y, X_k\rangle = 0$ for all $k \in \{1, \dots, n\}$ by assumption. The proof of Theorem 7.22 then shows that $Y = \widehat{X}_{n+h}^{(h)}$. □

Let us have a look at some examples.

Example 7.24. Consider a causal AR(1) process

$$X_t = \varphi X_{t-1} + Z_t, \quad (Z_t) \sim WN(0, \sigma^2), \quad |\varphi| < 1.$$

We have $\mu = 0$ and $\gamma(k) = \frac{\sigma^2}{1-|\varphi|^2}\varphi^k$ for $k \in \mathbb{N}_0$ (Theorem 6.5). Let $n, h \in \mathbb{N}$. The prediction equation (7.5) for $\widehat{X}^{(h)}_{n+h} = P_{n+h}X_n$ reads (after dividing by $\frac{\sigma^2}{1-|\varphi|^2}$)

$$\begin{pmatrix} 1 & \overline{\varphi} & \overline{\varphi}^2 & \dots & \overline{\varphi}^{n-1} \\ \varphi & 1 & \overline{\varphi} & \dots & \overline{\varphi}^{n-2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \varphi^{n-1} & \varphi^{n-2} & \varphi^{n-3} & \dots & 1 \end{pmatrix} \begin{pmatrix} \varphi^{(h)}_{n,1} \\ \varphi^{(h)}_{n,2} \\ \vdots \\ \varphi^{(h)}_{n,n} \end{pmatrix} = \begin{pmatrix} \varphi^h \\ \varphi^{h+1} \\ \dots \\ \varphi^{n+h-1} \end{pmatrix}$$

We see that

$$(\varphi^{(h)}_{n,1}, \varphi^{(h)}_{n,2}, \dots, \varphi^{(h)}_{n,n})' = (\varphi^h, 0, \dots, 0)'$$

solves this equation. Hence

$$\widehat{X}^{(h)}_{n+h} = P_n X_{n+h} = \varphi^h X_n.$$

Example 7.25. Consider a causal AR(2) process

$$X_t = \varphi_1 X_{t-1} + \varphi_2 X_{t-2} + Z_t, \quad (Z_t) \sim WN(0, \sigma^2).$$

Then

$$\begin{aligned} X_{n+3} &= Z_{n+3} + \varphi_1 X_{n+2} + \varphi_2 X_{n+1} \\ &= Z_{n+3} + \varphi_1(Z_{n+2} + \varphi_1 X_{n+1} + \varphi_2 X_n) + \varphi_2 X_{n+1} \\ &= Z_{n+3} + \varphi_1 Z_{n+2} + \varphi_1\varphi_2 X_n + (\varphi_1^2 + \varphi_2)(Z_{n+1} + \varphi_1 X_n + \varphi_2 X_{n-1}) \\ &= \underbrace{Z_{n+3} + \varphi_1 Z_{n+2} + (\varphi_1^1 + \varphi_2)Z_{n+1}}_{=:W} + \underbrace{(\varphi_1^2 + 2\varphi_1\varphi_2)X_n + (\varphi_1^2 + \varphi_2)\varphi_2 X_{n-1}}_{=:Y}. \end{aligned}$$

By causality, $\operatorname{Cov}(X_{n+3} - Y, X_i) = \operatorname{Cov}(W, X_i) = 0$ for $i = 1, \dots, n$. Hence, for $n \geq 2$, $\widehat{X}^{(3)}_{n+3} = P_n X_{n+3} = Y$ by Corollary 7.23, i.e.

$$P_n X_{n+3} = (\varphi_1^2 + 2\varphi_1\varphi_2)X_n + (\varphi_1^2 + \varphi_2)\varphi_2 X_{n-1}.$$

For ARMA and already for Moving Average processes prediction is much more difficult. Let us exemplify this for the MA(1) process:

Example 7.26. Consider the MA(1) process

$$X_t = Z_t + \theta Z_{t-1}, \quad (Z_t) \sim WN(0, 1)$$

with $\theta \in \mathbb{R}$. From Example 2.8 we know

$$\gamma(h) = \begin{cases} 1 + \theta^2, & h = 0, \\ \theta, & h = \pm 1, \\ 0, & |h| \geq 2. \end{cases}$$

Consider the 1-step prediction $\widehat{X}_4$ of X_4 based on X_1, X_2, X_3. Hence the predication equation (7.5) reads

$$\begin{pmatrix} 1+\theta^2 & \theta & 0 \\ \theta & 1+\theta^2 & \theta \\ 0 & \theta & 1+\theta^2 \end{pmatrix} \begin{pmatrix} \varphi_{1,3}^{(1)} \\ \varphi_{2,3}^{(1)} \\ \varphi_{3,3}^{(1)} \end{pmatrix} = \begin{pmatrix} \theta \\ 0 \\ 0 \end{pmatrix}$$

The Gauss algorithm leads to the equivalent system

$$\begin{pmatrix} \frac{(1+\theta^2)(1+\theta^4)}{1+\theta^2+\theta^4} & 0 & 0 \\ \frac{\theta(1+\theta^2)}{1+\theta^2+\theta^4} & 1 & 0 \\ 0 & \frac{\theta}{1+\theta^2} & 1 \end{pmatrix} \begin{pmatrix} \varphi_{1,3}^{(1)} \\ \varphi_{2,3}^{(1)} \\ \varphi_{3,3}^{(1)} \end{pmatrix} = \begin{pmatrix} \theta \\ 0 \\ 0 \end{pmatrix}$$

$$\Longrightarrow \varphi_{1,3}^{(1)} = \frac{\theta(1+\theta^2+\theta^4)}{(1+\theta^2)(1+\theta^4)},$$

$$\varphi_{2,3}^{(1)} = -\frac{\theta(1+\theta^2)}{1+\theta^2+\theta^4}\varphi_{1,3}^{(h)} = \frac{-\theta^2}{1+\theta^4}, \quad \varphi_{3,3}^{(1)} = -\frac{\theta}{1+\theta^2}\varphi_{2,3}^{(1)} = \frac{\theta^3}{(1+\theta^2)(1+\theta^4)}$$

This implies

$$\widehat{X}_4 = \frac{\theta^3}{(1+\theta^2)(1+\theta^4)}X_1 - \frac{\theta^2}{1+\theta^4}X_2 + \frac{\theta(1+\theta^2+\theta^4)}{(1+\theta^2)(1+\theta^4)}X_3.$$

The determination of the best linear predictor for the MA(1) process was really very complicated and did not have at all a nice form. Hence one often uses approximate forecasts for ARMA processes, by approximating ARMA processes by AR processes as explained in the following remark:

Remark 7.27. Consider the ARMA(p, q) process

$$X_t - \varphi_1 X_{t-1} - \ldots - \varphi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \ldots + \theta_q Z_{t-q}, \tag{7.6}$$

and assume that the process is causal and invertible.

From invertibility, there exist coefficients (ξ_j) such that

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

they can be obtained from

$$\theta(z) \sum_{j=0}^{\infty} \xi_j z^j = \phi(z).$$

If n is large in comparison to p and q, choose $m > p, q$ reasonably large, but $m < n - q$ and approximate this by

$$Z_t \approx \sum_{j=0}^{m} \xi_j X_{t-j} \quad \text{for all } t. \tag{7.7}$$

Inserting (7.7) into (7.6) gives

$$X_{n+1} \approx Z_{n+1} + \sum_{i=1}^{p} \varphi_i X_{n+1-i} + \sum_{k=1}^{q} \theta_k \sum_{j=0}^{m} \xi_j X_{n+1-k-j},$$

i.e. we have approximated the ARMA(p,q) model by an AR$(m+q)$-model. Assuming this AR$(m+q)$-model is causal,

$$\sum_{i=1}^{p} \varphi_i X_{n+1-i} + \sum_{k=1}^{q} \theta_k \sum_{j=0}^{m} \xi_j X_{n+1-k-j}$$

should be a good approximate forecast for X_{n+1} by Proposition 7.20.

The prediction equation (7.5) is particularly nice to solve if the matrix Γ_n is invertible. The following result gives sufficient conditions for this to happen.

Theorem 7.28. *Let $(X_t)_{t\in\mathbb{Z}}$ be a (complex-valued) stationary time series with ACVF γ. If $\gamma(0) > 0$ and $\lim_{h\to\infty} \gamma(h) = 0$, then $\Gamma_n = (\gamma(i-j))_{i,j=1,\dots,n}$ is invertible for every $n \in \mathbb{N}$.*

Proof. By subtracting the mean we can assume without loss of generality that $\mathbb{E}X_t = 0$ for all t. Since Γ_n is hermitian, we can diagonalise it and see that

$$\begin{aligned}\Gamma_n \text{ singular } &\iff \exists a = (a_1, \dots, a_n)' \in \mathbb{C}^n \setminus \{0\} : \ \overline{a'}\,\Gamma_n\, a = 0 \\ &\iff \exists a = (a_1, \dots, a_n)' \in \mathbb{C}^n \setminus \{0\} : \ \sum_{i,j=1}^{n} \overline{a_i} a_j \gamma(i-j) = 0 \\ &\iff \exists a = (a_1, \dots, a_n)' \in \mathbb{C}^n \setminus \{0\} : \operatorname{Var}\left(\sum_{i=1}^{n} \overline{a_i} X_i\right) = 0.\end{aligned}$$

Assume that $r := \inf\{m \in \mathbb{N} : \Gamma_m \text{ singular}\} - 1 < \infty$. Then $r \geq 1$ (since $\gamma(0) > 0$) and $\exists a_1, \dots, a_{r+1} \in \mathbb{C}$, not all 0, such that

$$\operatorname{Var}\left(\sum_{i=1}^{r+1} \overline{a_i} X_i\right) = 0.$$

As Γ_r is non-singular (i.e. invertible) we must have $a_{r+1} \neq 0$ so there exist $b_1, \dots, b_r \in \mathbb{C}$ such that

$$X_{r+1} = \sum_{j=1}^{r} b_j X_j \qquad \text{a.s.}$$

Since $(X_t)_{t\in\mathbb{Z}}$ is stationary, it follows that $\forall h \geq 1$

$$X_{r+h} = \sum_{j=1}^{r} b_j X_{j+h-1} \qquad \text{a.s.}$$

(since $\mathrm{Var}\,(X_{r+h} - \sum_{j=1}^r b_j X_{j+h-1}) = \mathrm{Var}\,(X_{r+1} - \sum_{j=1}^r b_j X_j) = 0$).

$$\Longrightarrow\ \forall n \geq r+1\ \exists b_1^{(n)}, \dots, b_r^{(n)} \in \mathbb{C} : X_n = \sum_{j=1}^r b_j^{(n)} X_j = (\vec{b}^{(n)})' \vec{X},$$

where $\vec{X} := (X_1, \dots, X_r)' \in \mathbb{C}^r$ and $\vec{b}^{(n)} = (b_1^{(n)}, \dots, b_r^{(n)})' \in \mathbb{C}^r$. We obtain

$$\gamma(0) = \mathbb{E}(X_n \overline{X_n}) = (\vec{b}^{(n)})' \mathbb{E}(\vec{X}\,\overline{\vec{X}'}\,)\,\overline{\vec{b}^{(n)}} = (\vec{b}^{(n)})'\,\Gamma_r\,\overline{\vec{b}^{(n)}}.$$

Next observe that Γ_r is hermitian, hence there exists a unitary matrix U (i.e. $U^{-1} = \overline{U'}$) and a diagonal matrix $\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_r) \in \mathbb{R}^{r \times r}$ (real valued!) such that

$$\Gamma_r = U\,\Lambda\,\overline{U'}.$$

Further, since Γ_r is positive semidefinite (see the proof of the implication "(i) $\Longrightarrow$ (ii)" in Theorem 4.12) we have $\lambda_1, \dots, \lambda_r \geq 0$ and since the matrix Γ_r is invertible we even have $\lambda_1, \dots, \lambda_r > 0$. Using this we can write

$$\gamma(0) = (\vec{b}^{(n)})'\,U\,\Lambda\,\overline{U'}\,\overline{\vec{b}^{(n)}}.$$

Since

$$c'\,\Lambda\,\overline{c} = \sum_{i=1}^n |c_i|^2 \lambda_i \geq |c|^2 \min\{\lambda_1, \dots, \lambda_r\}$$

for all $c = (c_1, \dots, c_r)' \in \mathbb{C}^r$ we obtain with $c := \overline{U'}\,\overline{(\vec{b}^{(n)})}$ that

$$\begin{aligned} \gamma(0) &\geq \min\{\lambda_1, \dots, \lambda_r\}\,|\overline{U'}\,\overline{(\vec{b}^{(n)})'}|^2 \\ &= \min\{\lambda_1, \dots, \lambda_r\}\,|(\vec{b}^{(n)})|^2 = \underbrace{\min\{\lambda_1, \dots, \lambda_r\}}_{>0} \sum_{j=1}^r |b_j^{(n)}|^2. \end{aligned}$$

Hence $(b_j^{(n)})_{n\in\mathbb{N}}$ is bounded for each $j \in \{1, \dots, r\}$, and it follows that

$$\gamma(0) = \mathrm{Cov}\,(X_n, \sum_{j=1}^r b_j^{(n)} X_j) \leq \sum_{j=1}^r \underbrace{|b_j^{(n)}|}_{\text{bounded}}\ \underbrace{|\gamma(n-j)|}_{\to 0} \to 0$$

as $n \to \infty$. This is a contradiction to $\gamma(0) > 0$. □

7.3 Recursive calculation of one-step predictors

Often one would like to do many forecasts. Why that? Well, suppose I have fitted a model, say an ARMA(1,1) process. One way to see if it is a good model is to check how it performs in terms of prediction. E.g., if I have 1000 observations $x_1, \dots, x_{1000}$, then I could predict x_{501} based on $x_1, \dots, x_{500}$, or x_{502} based on $x_1, \dots, x_{501}$ and then check how good my predictor was, i.e. compare the predicted value for x_{501} with the truly observed x_{501}, and so on. For doing such a thing, I have to predict very often, and also recursively.

In principle, the prediction problem can of course be solved by solving (7.5)

$$\Gamma_n \vec{\varphi}_n^{(h)} = \gamma_n(h).$$

However, (7.5) is a matrix equation, inverting Γ_n for large n is computationally time consuming and not clever, and solving the system can also be messy, as seen for the MA(1) process. Hence it would be good to have a recursive scheme for predictors, which avoids matrix inversion and solving linear systems. One of these algorithms is the Durbin–Levinson algorithm which we now present, another the innovations algorithm.

We write

$$\widehat{X}_{n+1} = \widehat{X}_{n+1}^{(1)} = P_n X_{n+1}$$

for the 1-step predictor. For simplicity, we restrict to real valued time series although it would not take much extra effort to formulate it also for complex-valued time series.

Theorem 7.29 (Durbin-Levinson-Algorithm).
Let $(X_t)_{t\in\mathbb{Z}}$ be a stationary real-valued time series with mean 0 and ACVF γ, with $\gamma(0) > 0$ and $\lim_{h\to\infty} \gamma(h) = 0$. Denote $\widehat{X}_1 := 0$,

$$\widehat{X}_{n+1} = \phi_{n1} X_n + \ldots + \phi_{nn} X_1,\ n \geq 1$$

the one-step-predictor (the coefficients are unique by Theorems 7.22 and 7.28) and $v_n = \mathbb{E}|X_{n+1} - \hat{X}_{n+1}|^2$, the squared one-step prediction error, for $n \in \mathbb{N}_0$. Then

$$\phi_{11} = \gamma(1)/\gamma(0), \quad v_0 = \gamma(0), \tag{7.8}$$

$$\phi_{nn} = \left(\gamma(n) - \sum_{j=1}^{n-1} \phi_{n-1,j}\gamma(n-j)\right) v_{n-1}^{-1}, \quad n \geq 2, \tag{7.9}$$

$$\begin{pmatrix} \phi_{n1} \\ \vdots \\ \phi_{n,n-1} \end{pmatrix} = \begin{pmatrix} \phi_{n-1,1} \\ \vdots \\ \phi_{n-1,n-1} \end{pmatrix} - \phi_{nn} \begin{pmatrix} \phi_{n-1,n-1} \\ \vdots \\ \phi_{n-1,1} \end{pmatrix}, \quad n \geq 2 \tag{7.10}$$

and

$$v_n = v_{n-1}(1 - \phi_{nn}^2) > 0, \quad n \geq 1. \tag{7.11}$$

To illustrate this algorithm, one uses the equations as follows:

- (7.8) $\mapsto \phi_{11}, v_0$
- (7.11) $\mapsto v_1$

- (7.9) $\mapsto \phi_{22}$
- (7.10) $\mapsto \phi_{21}$
- (7.11) $\mapsto v_2$
- (7.9) $\mapsto \phi_{33}$
- (7.10) $\mapsto \phi_{31}, \phi_{32}, \dots$

Proof. As Γ_n is not singular $\forall n \in \mathbb{N}$ by Theorem 7.29, we have $\operatorname{Var}(X_{n+1} - \sum\limits_{j=1}^{n} a_j X_j) > 0$ for all $a_1, \dots, a_n \in \mathbb{R}$, so

$$v_n > 0 \quad \forall n \in \mathbb{N}_0.$$

Let

$$\mathcal{H}_n = \overline{\text{span}}\{X_1, \dots, X_n\}, \quad \mathcal{K}_1 = \overline{\text{span}}\{X_2, \dots, X_n\}$$

and

$$\mathcal{K}_2 = \overline{\text{span}}\{X_1 - P_{\mathcal{K}_1} X_1\} = \{\lambda(X_1 - P_{\mathcal{K}_1} X_1) : \lambda \in \mathbb{R}\}$$

(we can restrict here to real vector spaces since the time series is real valued). Since $X_1 - P_{\mathcal{K}_1} X_1 \perp \mathcal{K}_1$, $\mathcal{K}_1$ and $\mathcal{K}_2$ are orthogonal subspaces of $\mathcal{H}_n$ and it holds that

$$\mathcal{H}_n = \mathcal{K}_1 \oplus \mathcal{K}_2$$

(a direct sum of orthogonal subspaces, meaning that $\langle x_1, x_2 \rangle = 0$ for all $x_1 \in \mathcal{K}_1$ and $x_2 \in \mathcal{K}_2$ and that for each $x \in \mathcal{H}_n$ there are unique $x_1 \in \mathcal{K}_1$ and $x_2 \in \mathcal{K}_2$ such that $x = x_1 + x_2$). For $Y \in L^2(\Omega, \mathcal{F}, P)$ it hence holds that

$$P_{\mathcal{H}_n} Y = P_{\mathcal{K}_1} Y + P_{\mathcal{K}_2} Y,$$

which implies that

$$\hat{X}_{n+1} = P_{\mathcal{K}_1} X_{n+1} + P_{\mathcal{K}_2} X_{n+1} = P_{\mathcal{K}_1} X_{n+1} + a(X_1 - P_{\mathcal{K}_1} X_1) \tag{7.12}$$

with

$$a = \frac{\langle \hat{X}_{n+1}, X_1 - P_{\mathcal{K}_1} X_1 \rangle}{||X_1 - P_{\mathcal{K}_1} X_1||^2} = \frac{\langle X_{n+1}, X_1 - P_{\mathcal{K}_1} X_1 \rangle}{||X_1 - P_{\mathcal{K}_1} X_1||^2} \tag{7.13}$$

(related to simple facts from linear algebra on how to develop a vector with respect to an orthonormal basis). Since (X_t) is stationary and real-valued, $(X_1, \dots, X_n)'$ and $(X_n, X_{n-1}, \dots, X_1)'$ and $(X_2, \dots X_{n+1})'$ have all the same covariance matrix, hence we get

$$P_{\mathcal{K}_1} X_{n+1} = \sum_{j=1}^{n-1} \phi_{n-1,j} X_{n+1-j}, \tag{7.14}$$

$$P_{\mathcal{K}_1}X_1 = \sum_{j=1}^{n-1} \phi_{n-1,j} X_{j+1} \tag{7.15}$$

and

$$||X_1 - P_{\mathcal{K}_1}X_1||^2 = ||X_{n+1} - P_{\mathcal{K}_1}X_{n+1}||^2 = ||X_n - \hat{X}_n||^2 = v_{n-1}. \tag{7.16}$$

$(7.12), (7.14), (7.15) \implies$

$$\hat{X}_{n+1} = aX_1 + \sum_{j=1}^{n-1}(\phi_{n-1,j} - a\phi_{n-1,n-j})X_{n+1-j}. \tag{7.17}$$

With (7.13) and (7.15) and (7.16) we obtain

$$\begin{aligned} a &= \left(\langle X_{n+1}, X_1 \rangle - \sum_{j=1}^{n-1} \phi_{n-1,j} \langle X_{n+1}, X_{j+1} \rangle \right) v_{n-1}^{-1} \\ &= \left(\gamma(n) - \sum_{j=1}^{n-1} \phi_{n-1,j} \gamma(n-j) \right) v_{n-1}^{-1}. \end{aligned}$$

As Γ_n is not singular the ϕ_{nj} are unique and comparison of coefficients in (7.17) and $\hat{X}_{n+1} = \sum_{j=1}^{n} \phi_{nj} X_{n+1-j}$ gives us

$$\phi_{nn} = a$$

and $\phi_{n,j} = \phi_{n-1,j} - a\phi_{n-1,n-j}$, which implies (7.9) and (7.10). To see (7.11), observe that

$$\begin{aligned} v_n = ||X_{n+1} - \hat{X}_{n+1}||^2 &= ||X_{n+1} - P_{\mathcal{K}_1}X_{n+1} - P_{\mathcal{K}_2}X_{n+1}||^2 \\ &= ||X_{n+1} - P_{\mathcal{K}_1}X_{n+1}||^2 + ||P_{\mathcal{K}_2}X_{n+1}||^2 - 2\langle X_{n+1} - P_{\mathcal{K}_1}X_{n+1}, P_{\mathcal{K}_2}X_{n+1} \rangle \\ &= v_{n-1} + a^2 v_{n-1} - 2a \langle X_{n+1}, X_1 - P_{\mathcal{K}_1}X_1 \rangle \\ &= v_{n-1} + a^2 v_{n-1} - 2a^2 v_{n-1} = (1 - \phi_{nn}^2) v_{n-1}. \end{aligned}$$

□

For further use in the next section when we introduce the partial autocorrelation we note the following corollary:

Corollary 7.30. *Under the assumptions and notations of Theorem* 7.29 *we have*

$$\phi_{nn} = \mathrm{corr}(X_{n+1} - P_{\overline{\mathrm{span}}\{X_2,\dots,X_n\}}X_{n+1}, X_1 - P_{\overline{\mathrm{span}}\{X_2,\dots,X_n\}}X_1)$$

(where corr *denotes the correlation).*

Proof. With the notations in the proof of Theorem 7.29 we have $P_{\mathcal{K}_1}X_{n+1} \perp (X_1 - P_{\mathcal{K}_1}X_1)$ and it follows that

$$\begin{aligned} \phi_{nn} = a &= \frac{\langle X_{n+1}, X_1 - P_{\mathcal{K}_1}X_1 \rangle}{||X_1 - P_{\mathcal{K}_1}X_1||^2} \\ &= \frac{\langle X_{n+1} - P_{\mathcal{K}_1}X_{n+1}, X_1 - P_{\mathcal{K}_1}X_1 \rangle}{||X_1 - P_{\mathcal{K}_1}X_1||^2} \\ &= \mathrm{corr}(X_{n+1} - P_{\mathcal{K}_1}X_{n+1}, X_1 - P_{\mathcal{K}_1}X_1). \end{aligned}$$

□

Let us see the Durbin-Levinson algorithm in action when forecasting an MA(1) process:

Example 7.31. As in Example 7.26, consider again the MA(1) process

$$X_t = Z_t + \theta Z_{t-1}, \quad (Z_t) \sim WN(0,1)$$

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

$$\gamma(0) = 1 + \theta^2, \quad \gamma(1) = \theta, \quad \gamma(h) = 0 \text{ for } \ h \geq 2$$

by Example 2.8. Let us follow the steps of the Durbin–Levinson algorithm:
Step 1:

$$v_0 = \gamma(0) = 1 + \theta^2, \quad \phi_{1,1} = \frac{\gamma(1)}{\gamma(0)} = \frac{\theta}{1+\theta^2}$$

Step 2:

$$v_1 = v_0(1 - \phi_{1,1}^2) = (1+\theta^2)\left(1 - \frac{\theta^2}{(1+\theta^2)^2}\right) = \frac{1+\theta^2+\theta^4}{1+\theta^2}$$

Step 3:

$$\phi_{2,2} = \left(\underbrace{\gamma(2)}_{=0} - \phi_{2-1,1}\underbrace{\gamma(2-1)}_{=\theta}\right) v_1^{-1} = -\frac{\theta}{1+\theta^2}\theta\frac{1+\theta^2}{1+\theta^2+\theta^4} = \frac{-\theta^2}{1+\theta^2+\theta^4}$$

Step 4:

$$\begin{aligned}\phi_{2,1} &= \phi_{2-1,1} - \phi_{2,2}\phi_{2-1,2-1} = \left(1 + \frac{\theta^2}{1+\theta^2+\theta^4}\right)\frac{\theta}{1+\theta^2} \\ &= \frac{(1+2\theta^2+\theta^4)\theta}{(1+\theta^2+\theta^4)(1+\theta^2)} = \frac{(1+\theta^2)\theta}{1+\theta^2+\theta^4}\end{aligned}$$

Back to Step 2:

$$\begin{aligned}v_2 &= v_1(1-\phi_{2,2}^2) = \frac{1+\theta^2+\theta^4}{1+\theta^2}\left(1 - \frac{\theta^4}{(1+\theta^2+\theta^4)^2}\right) \\ &= \frac{(1+\theta^2+\theta^4)^2 - \theta^4}{(1+\theta^2)(1+\theta^2+\theta^4)} = \frac{1+2\theta^2+2\theta^4+2\theta^6+\theta^8}{(1+\theta^2)(1+\theta^2+\theta^4)} \\ &= \frac{1+\theta^2+\theta^4+\theta^6}{1+\theta^2+\theta^4}\end{aligned}$$

Step 3:

$$\begin{aligned}\phi_{3,3} &= \left(\underbrace{\gamma(3)}_{=0} - \phi_{3-1,1}\underbrace{\gamma(3-1)}_{=0} - \phi_{3-1,2}\underbrace{\gamma(3-2)}_{=\theta}\right) v_2^{-1} \\ &= -\phi_{2,2}\theta v_2^{-1} = \frac{\theta^2}{1+\theta^2+\theta^4}\theta\frac{1+\theta^2+\theta^4}{1+\theta^2+\theta^4+\theta^6} \\ &= \frac{\theta^3}{(1+\theta^2)(1+\theta^4)}\end{aligned}$$

Step 4:

$$\begin{pmatrix} \phi_{3,1} \\ \phi_{3,2} \end{pmatrix} = \begin{pmatrix} \phi_{2,1} \\ \phi_{2,2} \end{pmatrix} - \phi_{3,3} \begin{pmatrix} \phi_{2,2} \\ \phi_{2,1} \end{pmatrix}$$

$$\begin{aligned} \phi_{3,1} &= \phi_{2,1} - \phi_{3,3}\phi_{2,2} = \frac{(1+\theta^2)\theta}{1+\theta^2+\theta^4} + \frac{\theta^3}{(1+\theta^2)(1+\theta^4)} \frac{\theta^2}{1+\theta^2+\theta^4} \\ &= \ldots = \frac{\theta(1+\theta^2+\theta^4)}{(1+\theta^2)(1+\theta^4)} \end{aligned}$$

$$\begin{aligned} \phi_{3,2} &= \phi_{2,2} - \phi_{3,3}\phi_{2,1} = \frac{-\theta^2}{1+\theta^2+\theta^4} - \frac{\theta^3}{(1+\theta^2)(1+\theta^4)} \frac{(1+\theta^2)\theta}{1+\theta^2+\theta^4} \\ &= \ldots = \frac{-\theta^2}{1+\theta^4} \end{aligned}$$

Hence

$$P_1 X_2 = \phi_{1,1} X_1 = \frac{\theta}{1+\theta^2} X_1$$

$$P_2 X_3 = \phi_{2,1} X_2 + \phi_{2,2} X_1 = \frac{(1+\theta^2)\theta}{1+\theta^2+\theta^4} X_2 - \frac{\theta^2}{1+\theta^2+\theta^4} X_1$$

$$\begin{aligned} P_3 X_4 &= \phi_{3,1} X_3 + \phi_{3,2} X_2 + \phi_{3,3} X_3 \\ &= \frac{\theta(1+\theta^2+\theta^4)}{(1+\theta^2)(1+\theta^4)} X_1 - \frac{\theta^2}{1+\theta^4} X_2 + \frac{\theta^3}{(1+\theta^2)(1+\theta^4)} X_1 \end{aligned}$$

in accordance with our previous calculation of Example 7.26.

Although the Durbin–Levinson algorithm performs much better than solving systems of linear equations, for ARMA processes with a non-trivial moving average part it is still a bit cumbersome. A good alternative is presented by the so called innovations algorithm. Denoting the one-step predictors by $\widehat{X}_{n+1}$, the idea behind the innovation algorithm is that $\{X_1, X_2 - \widehat{X}_2, X_3 - \widehat{X}_3, \ldots, X_n - \widehat{X}_n\}$ forms an orthogonol system that spans $\operatorname{span}\{X_1, \ldots, X_n\}$, and hence (defining $\widehat{X}_1 := 0$) there must be coefficients $\theta_{n,j} = \theta_{nj}$ such that

$$\widehat{X}_{n+1} = \sum_{j=1}^{n} \theta_{nj}(X_{n+1-j} - \widehat{X}_{n+1-j}) = \theta_{n,1}(X_n - \widehat{X}_n) + \theta_{n,2}(X_{n-1} - \widehat{X}_{n-1}) + \ldots + \theta_{n,n}(X_1 - \widehat{X}_1).$$

The quantities $X_j - \widehat{X}_j$ are called the *innovations*, which is where the name of the innovation algorithm comes from. And the nice feature is then that also the coefficients $\theta_{n,j}$ allow an easy pattern for the recursive determination, and this even if the time series is not necessarily stationary.

Let us first prove the statement regarding the orthogonality and the span of the innovations.

Lemma 7.32. *Let $X_1, \ldots, X_n$ be mean zero random variables with finite variance and define $\widehat{X}_1 := 0$ and denote by $\widehat{X}_j$ the best linear one-step predictor of X_j given $X_1, \ldots, X_{j-1}$ (for $j = 2, \ldots, n$). Then $\{X_j - \widehat{X}_j : j = 1, \ldots, n\}$ forms an orthogonal system (i.e. $\langle X_j - \widehat{X}_j, X_k - \widehat{X}_k \rangle = 0$ for $j \neq k$) and*

$$\operatorname{span}\{X_1 - \widehat{X}_1, \ldots, X_n - \widehat{X}_n\} = \operatorname{span}\{X_1, \ldots, X_n\} = \overline{\operatorname{span}}\{X_1, \ldots, X_n\}.$$

If additionally the matrix $(\mathbb{E}(X_i \overline{X_j}))_{i,j=1,\ldots,n}$ is invertible, then $X_1 - \widehat{X}_1, \ldots, X_n - \widehat{X}_n$ are linearly independent, so that for each $Y \in \operatorname{span}\{X_1, \ldots, X_n\}$ there exist unique coefficients $\theta_{n1}, \ldots, \theta_{nn} \in \mathbb{C}$ such that $Y = \sum_{j=1}^{n} \theta_{nj}(X_{n+1-j} - \widehat{X}_{n+1-j})$.

Proof. We know that $\widehat{X}_j \in \operatorname{span}\{X_1, \ldots, X_{j-1}\} \subset \operatorname{span}\{X_1, \ldots, X_j\}$ for each j so that $X_j - \widehat{X}_j \in \operatorname{span}\{X_1, \ldots, X_j\}$ for each j. Since also $\widehat{X}_{j+1} - X_{j+1} \perp \operatorname{span}\{X_1, \ldots, X_j\}$ we obtain that $\{X_j - \widehat{X}_j : j = 1, \ldots, n\}$ forms an orthogonal system.

Denote $\mathcal{H}_n := \operatorname{span}\{X_1, \ldots, X_n\}$. The assertion on the span is proved by induction on n. It is trivially true for $n = 1$ (since then $\widehat{X}_1 = 0$), so now suppose that it is true for some $n \in \mathbb{N}$. We already know that $\widehat{X}_{n+1} \in \operatorname{span}\{X_1, \ldots, X_n\} = \mathcal{H}_n$, so

$$\begin{aligned} & \operatorname{span}\{X_{n+1} - \widehat{X}_{n+1}, X_n - \widehat{X}_n, \ldots, X_1 - \widehat{X}_1\} \\ \overset{\text{induction hypothesis}}{=} & \operatorname{span}\{X_{n+1} - \widehat{X}_{n+1}, \mathcal{H}_n\} \\ \overset{\widehat{X}_{n+1} \in \mathcal{H}_n}{=} & \operatorname{span}\{X_{n+1}, \mathcal{H}_n\} \\ = & \mathcal{H}_{n+1}. \end{aligned}$$

To see the linear independence, we can argue as in the proof of Theorem 7.28 to see that invertibility of $(\mathbb{E}(X_i \overline{X_j}))_{i,j=1,\ldots,n}$ implies $\overline{a}'(\mathbb{E}(X_i \overline{X_j}))_{i,j=1,\ldots,n} a \neq 0$ for all $a \in \mathbb{C}^n \setminus \{0\}$, which is equivalent to $\operatorname{Var}\left(\sum_{i=1}^{n} \overline{a}_i X_i\right) > 0$ whenever $(a_1, \ldots, a_n)' \neq (0, \ldots, 0)'$. Hence a non-trivial linear combination of the X_i has non-zero variance, in particular cannot be equal to 0 almost surely. This implies linear independence of $X_1, \ldots, X_n$, hence the dimension of $\mathcal{H}_n$ is n, and since $X_1 - \widehat{X}_1, \ldots, X_n - \widehat{X}_n$ span $\mathcal{H}_n$, they must be linearly independent, too. Hence the coefficients θ_{nj} must be unique. □

You might recall from linear algebra that coefficients with respect to orthonormal bases are often much easier to calculate than with respect to other bases. Well, here the system is only orthogonal and not orthonormal, nevertheless the determination of the coefficients turns out easy. The innovations algorithm now computes the coefficients in the innovation representation. We restrict to real valued sequences for simplicity.

Theorem 7.33 (The innovations algorithm).
Let $(X_t)_{t \in \mathbb{Z}}$ be a real valued time series with mean zero and finite variances (not necessarily stationary), and denote

$$\kappa(i, j) := \mathbb{E}(X_i X_j) \quad \forall i, j \in \mathbb{Z}$$

(so that $\kappa(i, j) = \gamma(i - j)$ if X is additionally stationary). Define $\widehat{X}_1 := 0$ and denote the best linear one-step predictor of X_{n+1} given $X_1, \ldots, X_n$ by $\widehat{X}_{n+1}$ for each $n \in \mathbb{N}_0$. Define

$$v_n := \mathbb{E}(X_{n+1} - \widehat{X}_{n+1})^2, \quad n \in \mathbb{N}_0,$$

the squared one-step prediction error. Assume that the matrix $(\kappa(i,j))_{i,j=1,\ldots,n}$ *is invertible for each* $n \in \mathbb{N}$ *(if the time series is stationary, this means that* $\Gamma_n = (\gamma(i-j))_{i,j=1,\ldots,n}$ *is invertible for every* $n \in \mathbb{N}$*, a sufficient condition for which is given in Theorem 7.28). Then for every* $n \in \mathbb{N}$ *there are unique coefficients* $\theta_{nj} \in \mathbb{R}$*,* $j = 1, \ldots, n$*, such that*

$$\widehat{X}_{n+1} = \sum_{j=1}^{n} \theta_{nj}(X_{n+1-j} - \widehat{X}_{n+1-j}), \tag{7.18}$$

and the coefficients and squared one-step prediction errors are given recursively by

$$v_0 = \kappa(1,1), \tag{7.19}$$

$$\theta_{n,n-k} = v_k^{-1}\left(\kappa(n+1,k+1) - \sum_{j=0}^{k-1} \theta_{k,k-j}\theta_{n,n-j}v_j\right), \quad k = 0, 1, \ldots, n-1, \tag{7.20}$$

$$v_n = \kappa(n+1,n+1) - \sum_{j=0}^{n-1} \theta_{n,n-j}^2 v_j. \tag{7.21}$$

Under the conditions of Theorem 7.33, the coefficients and squared predictions errors are solved recursively in the order

$$v_0; \theta_{11}; v_1; \theta_{22}, \theta_{21}; v_2; \theta_{33}, \theta_{32}, \theta_{31}; v_3, \ldots$$

Proof. Denote $\mathcal{H}_n := \operatorname{span}\{X_1, \ldots, X_n\}$. We know from Lemma 7.32 that $\{X_1 - \widehat{X}_1, \ldots, X_n - \widehat{X}_n\}$ forms a basis of $\mathcal{H}_n$ with orthogonal elements. Hence for each $n \in \mathbb{N}$, there are unique coefficients $\theta_{nj} \in \mathbb{R}, j = 1, \ldots, n$ such that (7.18) holds (that the θ_{nj} are real and not complex can be seen by repeating the proof of Lemma 7.32 for real spans which all works since the X_j are real valued). Further, since $X_1, \ldots, X_{n+1}$ are linearly independent by the proof of Lemma 7.32, we have $X_{n+1} \notin \mathcal{H}_n$ and hence $\widehat{X}_{n+1} \neq X_{n+1}$, so that $v_n > 0$ for all $n \in \mathbb{N}_0$. Now take in (7.18) the inner product with $X_{k+1} - \widehat{X}_{k+1}$ for $k \in \{0, 1, \ldots, n-1\}$, then by orthogonality we obtain

$$\left\langle \widehat{X}_{n+1}, X_{k+1} - \widehat{X}_{k+1} \right\rangle = \sum_{j=1}^{n} \theta_{nj} \underbrace{\left\langle X_{n+1-j} - \widehat{X}_{n+1-j}, X_{k+1} - \widehat{X}_{k+1} \right\rangle}_{=v_k \text{ if } n+1-j=k+1;\text{otherwise } =0} = \theta_{n,n-k}v_k.$$

Using again the orthogonality, we have $\langle X_{n+1} - \widehat{X}_{n+1}, X_{k+1} - \widehat{X}_{k+1}\rangle = 0$ for $k \leq n-1$, hence

$$\left\langle X_{n+1}, X_{k+1} - \widehat{X}_{k+1} \right\rangle = \theta_{n,n-k}\, v_k.$$

Dividing by $v_k \neq 0$ we obtain

$$\theta_{n,n-k} = v_k^{-1}\left\langle X_{n+1}, X_{k+1} - \widehat{X}_{k+1} \right\rangle, \quad k = 0, \ldots, n-1. \tag{7.22}$$

By (7.18) with n replaced by k, we have

$$\widehat{X}_{k+1} = \sum_{j=1}^{k} \theta_{k,j}(X_{k+1-j} - \widehat{X}_{k+1-j}) = \sum_{j=0}^{k-1} \theta_{k,k-j}(X_{j+1} - \widehat{X}_{j+1})$$

for $k \geq 0$ (for $k = 0$ we have $\widehat{X}_1 = 0$ which coincides with the above formula as the empty sum is zero by definition). Plugging this into (7.22) we obtain for $k = 0, \dots, n-1$

$$\begin{aligned} \theta_{n,n-k} &= v_k^{-1} \left(\langle X_{n+1}, X_{k+1} \rangle - \left\langle X_{n+1}, \sum_{j=0}^{k-1} \theta_{k,k-j} (X_{j+1} - \widehat{X}_{j+1}) \right\rangle \right) \\ &= v_k^{-1} \left(\kappa(n+1, k+1) - \sum_{j=0}^{k-1} \theta_{k,k-j} \langle X_{n+1}, X_{j+1} - \widehat{X}_{j+1} \rangle \right). \qquad (7.23) \end{aligned}$$

But by (7.22), $\langle X_{n+1}, X_{j+1} - \widehat{X}_{j+1} \rangle = \theta_{n,n-j} v_j$, hence we can rewrite (7.23) to

$$\theta_{n,n-k} = v_k^{-1} \left(\kappa(n+1, k+1) - \sum_{j=0}^{k-1} \theta_{k,k-j} \theta_{n,n-j} v_j \right), \quad k = 0, 1, \dots, n-1,$$

which is (7.20). Equation (7.21) follows from

$$v_n = \|X_{n+1} - \widehat{X}_{n+1}\|^2 \overset{\text{Prop. 7.10 (b)}}{=} \|X_{n+1}\|^2 - \|\widehat{X}_{n+1}\|^2 = \kappa(n+1, n+1) - \|\widehat{X}_{n+1}\|^2$$

and the fact that

$$\|\widehat{X}_{n+1}\|^2 \overset{(7.18)}{=} \left\langle \sum_{j=1}^{n} \theta_{nj} (X_{n+1-j} - \widehat{X}_{n+1-j}), \sum_{j=1}^{n} \theta_{nj} (X_{n+1-j} - \widehat{X}_{n+1-j}) \right\rangle = \sum_{k=0}^{n-1} \theta_{n,n-k}^2 v_k$$

by orthogonality of $\{X_1 - \widehat{X}_1, \dots, X_n - \widehat{X}_n\}$. □

Example 7.34. Consider again the MA(1) process $X_t = Z_t + \beta Z_{t-1}$, $t \in \mathbb{Z}$, where $Z = (Z_t)_{t \in \mathbb{Z}} \sim WN(0, \sigma^2)$ and $\beta \in \mathbb{R}$ (we use β rather than θ so that no confusion with the innovation coefficients arises). We have

$$\kappa(i,j) = \begin{cases} 0, & |i-j| > 1, \\ \beta\sigma^2, & |i-j| = 1, \\ \sigma^2(1+\beta^2), & i = j \end{cases}$$

from Example 2.8. With the notations of the innovations algorithm, we have

$$\begin{aligned} v_0 &= \kappa(1,1) = \sigma^2(1+\beta^2), \\ \theta_{1,1} &= v_0^{-1} \kappa(2,1) = \frac{\beta}{1+\beta^2}, \\ v_1 &= \kappa(2,2) - \theta_{1,1}^2 v_0 = (1+\beta^2)\sigma^2 - \left(\frac{\beta}{1+\beta^2} \right)^2 \sigma^2(1+\beta^2) = \sigma^2 \frac{1+\beta^2+\beta^4}{1+\beta^2}, \end{aligned}$$

and for $n \geq 2$ we obtain

$$(7.20) \Longrightarrow \theta_{n,n} = 0 \text{ (use } k=0) \overset{(7.20)}{\Longrightarrow} \theta_{n,n-1} = 0 \text{ (use } k=1) \Longrightarrow \dots \Longrightarrow \theta_{n,2} = 0$$

$$\overset{(7.20)}{\Longrightarrow} \theta_{n,1} = v_{n-1}^{-1}\,\kappa(n+1,n) = v_{n-1}^{-1}\beta\sigma^2$$

$$\overset{(7.21)}{\Longrightarrow} v_n = \kappa(n+1,n+1) - \theta_{n,1}^2 v_{n-1} = \sigma^2(1+\beta^2) - v_{n-1}^{-1}\beta^2\sigma^4.$$

The important message is that $\theta_{n,j} = 0$ when $2 \le j \le n$ so that for $n \ge 2$

$$\widehat{X}_{n+1} = \theta_{n,1}(X_n - \widehat{X}_n) = \beta(X_n - \widehat{X}_n)\frac{\sigma^2}{v_{n-1}}.$$

Defining

$$r_n := v_n/\sigma^2, \quad n \in \mathbb{N}_0$$

we can write

$$\widehat{X}_{n+1} = \beta(X_n - \widehat{X}_n)/r_{n-1} \quad \text{for } n \ge 2,$$

where

$$r_0 = 1 + \beta^2 \quad \text{and} \quad r_{n+1} = 1 + \beta^2 - \beta^2/r_n \quad \forall\, n \ge 1.$$

Remark 7.35. Suppose we have a realisation $x_1, \ldots, x_T$ of a real-valued time series satisfying the assumptions of Theorem 7.33. We then have

$$\widehat{x}_1 := 0, \quad \widehat{x}_2 = \theta_{1,1}(x_1 - \widehat{x}_1), \quad \widehat{x}_3 = \theta_{2,1}(x_2 - \widehat{x}_2) + \theta_{2,2}(x_1 - \widehat{x}_1), \quad \ldots.$$

This follows from the fact that if we have a function $f_n : \mathbb{R}^n \to \mathbb{R}$ such that $\widehat{X}_{n+1} = f_n(X_1, \ldots, X_n)$, then $\widehat{X}_{n+1}(\omega) = f_n(X_1(\omega), \ldots, X_n(\omega))$ for all $\omega \in \Omega$ (the sample space), regardless what the function f_n does outside the range of $(X_1, \ldots, X_n)$. Then if $x_1 = X_1(\omega_0), \ldots, x_n = X_n(\omega_0)$ for some realisation (i.e. one fixed ω_0, then the predictor is $\widehat{X}_{n+1}(\omega_0) = f_n(x_1, \ldots, x_n)$ (cf. Method 7.1). The innovations algorithm provides a function, where f_n is defined in terms of $f_{n-1}, \ldots, f_1$, but nevertheless it is a function and hence we must get the predictor of the realisation as stated. Observe that by the uniqueness of the predictor any two functions f_n and g_n with $\widehat{X}_{n+1} = f_n(X_1, \ldots, X_n) = g_n(X_1, \ldots, X_n)$ must agree on the range of $(X_1, \ldots, X_n)$ (apart from what happens on a null-set, which is negligible), the definition of the prediction of a realisation as in Method 7.1 is actually well-defined. Hence, the innovation algorithm gives the same result as would the Durbin-Levinson algorithm do or even a method by solving the linear system of the equations.

Remark 7.36. Given an ARMA(p, q) process, it is possible to transform the process in such a way such that the innovations algorithm becomes particularly easy to implement. The interested reader is referred to Brockwell and Davis [BD1, Section 5.3].

The representation of $\widehat{X}_{n+1}$ in terms of the innovations also opens the way for the determination of h-step predictors. We have:

Theorem 7.37. *With the notations and with the assumptions of Theorem 7.33, for $h \in \mathbb{N}$, the (linear) h-step predictor $P_nX_{n+h} = \widehat{X}_{n+h}^{(h)}$ of X_{n+h} given $X_1, \ldots, X_n$ is given by*

$$P_nX_{n+h} = \sum_{j=h}^{n+h-1} \theta_{n+h-1,j}(X_{n+h-j} - \widehat{X}_{n+h-j}),$$

where the coefficients θ_{nj} are determined as in the innovation algorithm. Moreover, the squared h-step prediction error is given by

$$v_n^{(h)} = \kappa(n+h, n+h) - \sum_{j=h}^{n+h-1} \theta_{n+h-1,j}^2 v_{n+h-j-1}.$$

Proof. Denote by P_m the orthogonal projection onto $\operatorname{span}\{X_1, \dots, X_m\}$ for $m \in \mathbb{N}$. Let $k \geq m$. We claim first that

$$P_m Y = P_m(P_k Y) \tag{7.24}$$

for all random variables Y with finite variance. To see this, observe that $P_m(P_k Y) \in \operatorname{span}\{X_1, \dots, X_m\}$ by definition of P_m. Further,

$$Y - P_m(P_k Y) = (Y - P_k(Y)) + (P_k(Y) - P_m(P_k Y)).$$

But $Y - P_k(Y)$ is orthogonal to $\operatorname{span}\{X_1, \dots, X_k\}$ and hence to $\operatorname{span}\{X_1, \dots, X_m\}$ since $k \geq m$. Also, $P_k Y - P_m(P_k Y)$ is orthogonal to $\operatorname{span}\{X_1, \dots, X_m\}$. Hence $Y - P_m(P_k Y)$ is orthogonal to $\operatorname{span}\{X_1, \dots, X_m\}$, and since $P_m(P_k Y) \in \operatorname{span}\{X_1, \dots, X_m\}$ the orthogonal projection theorem shows that $P_m(P_k Y) = P_m Y$.

Let us apply (7.24) now to our setting. We have for $h \geq 2$

$$\begin{aligned} P_n X_{n+h} &\overset{(7.24)}{=} P_n P_{n+h-1} X_{n+h} \\ &= P_n \widehat{X}_{n+h} \\ &\overset{(7.18)}{=} P_n \left(\sum_{j=1}^{n+h-1} \theta_{n+h-1,j}(X_{n+h-j} - \widehat{X}_{n+h-j}) \right) \\ &= \sum_{j=1}^{n+h-1} \theta_{n+h-1,j} P_n (X_{n+h-j} - \widehat{X}_{n+h-j}) \\ &= \sum_{j=h}^{n+h-1} \theta_{n+h-1,j}(X_{n+h-j} - \widehat{X}_{n+h-j}), \end{aligned}$$

since

$$X_{n+h-j} - \widehat{X}_{n+h-j} \begin{cases} \in \operatorname{span}\{X_1, \dots, X_n\}, & j \geq h, \\ \perp \operatorname{span}\{X_1, \dots, X_n\}, & j = 1, \dots, h-1 \end{cases}$$

(use Proposition 7.10 (iv),(v)). This shows

$$\|P_n X_{n+h}\|^2 = \sum_{j=h}^{n+h-1} \theta_{n+h-1,j}^2 \, v_{n+h-j-1}$$

by the orthogonality of the representation and with

$$v_n^{(h)} = \mathbb{E}(X_{n+h} - P_n X_{n+h})^2 \overset{\text{Prop. 7.10 (ii)}}{=} \|X_{n+h}\|^2 - \|P_n X_{n+h}\|^2$$

we obtain the desired formula for $v_n^{(h)}$. □

We end this section with a remark concerning the so called prediction bounds:

Remark 7.38. We have viewed the prediction problem in practice as something where you get a number, i.e. you have observations $x_1, \dots, x_n$ and you would like to have a number $\widehat{x}_{n+h}$ which should be close to what you expect for x_{n+h}. What would statements mean like 'The probability that a loss of at least 1000 Euros occurs is greater than 5 %', or in terms of weather forecasts statements like 'The probability that it rains tomorrow is 30 %'? Even if we have not done anything with respect to weather forecasts, the idea behind both statements is the same: one assumes a statistical model and is not only interested in getting a number (like I predict 300 ml of rain per square meter), but a probability distribution based on the knowledge of today. So actually I would be interested in the probability distribution of X_{n+h} given $X_1, \dots, X_n$ (this depends on my model, which of course may be wrong; just think of weather forecasts, and another person might use a different model and hence get other forecasts). Since I can calculate $\widehat{X}_{n+h}^{(h)} = P_n X_{n+h}$, this is equivalent to saying that I am interested in the distribution of

$$\Delta_n^{(h)} = X_{n+h} - P_n X_{n+h}$$

(given $X_1, \dots, X_n$). Now assume that the time series $(X_t)_{t \in \mathbb{Z}}$ is a Gaussian time series with zero mean. Then also $\Delta_n^{(h)}$ will be normally distributed with mean 0 and variance $\mathbb{E}(\Delta_n^{(h)})^2 = v_n^{(h)}$, which can be calculated using Theorem 7.37. The conditional distribution of X_{n+h} given $X_1, \dots, X_n$ is then a normal distribution with mean $P_n X_{n+h}$ and variance $v_n^{(h)}$. Now if $\Phi_{1-\alpha/2}$ denotes the $(1-\alpha/2)$-quantile of the standard normal distribution (so that $\int_{-\infty}^{\Phi_{1-\alpha/2}} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-x^2/2}\, \mathrm{d}x = 1 - \alpha/2$), then (conditional on $X_1, \dots, X_n$), X_{n+h} lies with probability $1-\alpha$ between the bounds $P_n X_{n+h} - \Phi_{1-\alpha/2}\sqrt{v_n^{(h)}}$ and $P_n X_{n+h} + \Phi_{1-\alpha/2}\sqrt{v_n^{(h)}}$. These bounds are then called the $(1-\alpha)$-*prediction bounds for* X_{n+h}.

7.4 The partial autocorrelation function and how to detect visually AR or MA processes?

To start with consider a moving average process $X_t = Z_t + \theta_1 Z_{t-1} + \dots + \theta_q Z_{t-q}$ with $\theta_1, \dots, \theta_q \in \mathbb{R}$ with $\theta_q \neq 0$ (so the true order is q), where $Z = (Z_t)_{t \in \mathbb{Z}} \sim WN(0, \sigma^2)$ with $\sigma > 0$. From Example 2.8 we know that the autocorrelation function $\rho_X(h) = \gamma_X(h)/\gamma_X(0)$ is given by

$$\rho_X(h) = \frac{\sum_{j=0}^{q-|h|} \theta_j \theta_{|h|+j}}{\sum_{j=0}^{q} \theta_j^2},$$

so in particular

$$\rho_X(h) = \begin{cases} 0, & \text{if} \quad h \geq q+1, \\ \frac{\theta_q}{\sum_{j=0}^{q} \theta_j^2} \neq 0, & \text{if} \quad h = q. \end{cases}$$

So the autocorrelation function is zero for lags greater than q and non-zero at lag q. Suppose now that you have a sample $x_1, \dots, x_n$ of a real valued stationary time series and you see that the sample autocorrelation function $\widehat{\rho}(h)$ is significantly different from zero for $h = q$ and looks close to zero (with possible a minor number of outliers) for $h > q$. Assuming that the sample autocorrelation function gives a good estimator of the true autocorrelation function (we will establish later conditions under which this is true), it is reasonable to assume that the data might come from an MA(q)-process. This is indeed justified, by a theoretical result which states that if $X = (X_t)_{t\in\mathbb{Z}}$ is a real valued zero-mean stationary process with autocovariance function γ such that $\gamma(h) = 0$ for $h > q$ and $\gamma(q) = 0$, then there must be a white noise sequence $(Z_t)_{t\in\mathbb{Z}}$ which respect to which X is an MA(q)-process. We shall not prove that result, although it would be feasible for us with the methods we know. A proof of that result can be found in Brockwell and Davis [BD1, Proposition 3.2.1]. So if the sample autocorrelation functions of a time series are significantly different from 0 for $h = q$ and look zero for higher lags, then this is a strong indication that the data might come from an MA(q)-process.

Is there a similar thing for autoregressive processes? The answer is yes, and it is the partial autocorrelation function. Its definition is a bit cumbersome, as it comes out of the proof of the Durbin–Levinson algorithm. We give it only for real-valued time series:

Definition 7.39. Let $X = (X_t)_{t\in\mathbb{Z}}$ be a stationary real-valued time series. Then the partial autocorrelation function $\alpha : \mathbb{N} \to \mathbb{R}$ is defined by

$$\alpha(k) = \begin{cases} \rho(1) = \operatorname{corr}(X_2, X_1) \text{ if } k = 1 \\ \operatorname{corr}(X_{k+1} - P_{\overline{\operatorname{span}}\{1,X_2,\dots,X_k\}}X_{k+1}, X_1 - P_{\overline{\operatorname{span}}\{1,X_2,\dots,X_k\}}X_1),\ k \geq 2. \end{cases}$$

(often one also defines $\alpha(0) := 1$).

This is a bit strange, but from the definition you can view $\alpha(n)$ as the correlation between X_1 and X_{n+1}, where the information 'between', i.e. $(X_2, \dots, X_n)$, is additionally taken into account, or as the correlation of two residua, after X_{n+1} and X_1 are regressed on the observations $X_2, \dots, X_n$. This is not yet terrible helpful either. How do I calculate the partial autocorrelation function. Fortunately, we have already solved this issue in Corollary 7.30:

Proposition 7.40. *Let $X = (X_t)_{t\in\mathbb{Z}}$ be a stationary real-valued time series with mean 0 and ACVF γ, with $\gamma(0) > 0$ and $\lim_{h\to\infty} \gamma(h) = 0$. For $n \geq 1$ denote the best linear one-step predictor of X_{n+1} by $\widehat{X}_{n+1}$ with representation*

$$\widehat{X}_{n+1} = \phi_{n1}X_n + \ldots + \phi_{nn}X_1,$$

where the coefficients $\phi_{n1}, \dots, \phi_{nn} \in \mathbb{R}$ are unique. Then $\alpha(n) = \phi_{nn}$, i.e. the partial autocorrelation at lag n is equal to the coefficient of X_1 in the representation of the 1-step predictor of X_{n+1}.

Proof. For $n \geq 2$ this is clear from Corollary 7.30, and for $n = 1$ it follows from the definition of $\alpha(1)$ and (7.8). □

As a corollary, we obtain:

Corollary 7.41. *Let $X = (X_t)_{t\in\mathbb{Z}}$ be a real valued stationary causal AR(p) process of the form $X_t - \varphi_1 X_{t-1} - \ldots - \varphi_p X_{t-p} = Z_t$ $\varphi_p \neq 0$. Then $\alpha(p) = \varphi_p \neq 0$ and $\alpha(h) = 0$ for $h > p$.*

Proof. This is immediate from Propositions 7.20 and 7.40. □

Remark 7.42. As for the relation of the autocovariance to moving average processes, there is also a converse to Corollary 7.41. It can be shown that a real valued stationary time series with mean zero is an AR(p) process with highest coefficient different from zero, if and only if the partial autocorrelation function is different from zero at p and zero for lags greater than p. See F.L. Ramsey, Characterization of the partial autocorrelation function, *The Annals of Statistics*, Vol. 2 No. 6, 1974, pp. 1296–1301. It is also possible to define other objects which similarly characterize general ARMA(p, q) processes for given orders. See W.A. Woodward and H.L. Gray, On the relationship between the S array and the Box-Jenkins methods of ARMA model identification, *Journal of the American Statistical Association*, Vol. 76 No. 375, 1981, pp. 579–587.

Proposition 7.40 allows the definition of an empirical partial autocorrrelation function: observe that by (7.5), $\Gamma_n \vec{\varphi}_n^{(1)} = \vec{\gamma}_n^{(1)}$, so that

$$\varphi_{nn} = e_n' \Gamma_n^{-1} \vec{\gamma}_n^{(1)},$$

where $e_n = (0, \ldots, 0, 1)' \in \mathbb{R}^n$ denotes the n'th unit vector in $\mathbb{R}^n$ (provided Γ_n is invertible). If we replace Γ_n and $\widehat{\gamma}_n^{(h)}$ by their corresponding empirical versions, we hence arrive at the empirical autocorrelation function:

Definition 7.43. Let $x_1, \ldots, x_n$ be a realisations of a stationary real-valued time series and define the empirical autocovariances $\widehat{\gamma}(h)$ and empirical autocovariance matrix $\widehat{\Gamma}_k = (\widehat{\gamma}(i-j))_{i,j=1,\ldots,k} \in \mathbb{R}^{k\times k}$ as in Definition 4.8. Suppose that $\widehat{\Gamma}_k$ is invertible. Then

$$\widehat{\varphi}_{kk} := e_k' \widehat{\Gamma}_k^{-1} (\widehat{\gamma}(1), \ldots, \widehat{\gamma}(k))'$$

is called the *empirical partial autocorrelation function (of this realisation) at lag $k < n$* or *sample partial autocorrelation function at lag k.* (For $k = 0$ one defines the empirical partial autocorrelation function as being equal to 1).

We shall see later (when doing the Yule-Walker estimators) that the condition that $\widehat{\Gamma}_k$ is invertible for some $k < n$ is equivalent to $\widehat{\gamma}(0) \neq 0$, i.e. to the fact that not all observations $x_1, \ldots, x_n$ are the same, so that the empirical autocorrelation function can actually be defined in every reasonable scenario. Assuming that the empirical partial autocorrelation gives a good estimate of the true partial autocorrelation, empirical autocorrelations that are significantly different from 0 at lag p and close to zero at lags greater than p give a strong indication that the underlying model might be an AR(p) process. Statistical software packages often plot both the empirical autocorrelation function and the empirical partial autocorrelation function.

Let us give some examples:

Example 7.44. Figure 7.1 shows a simulated MA(3) process $X_t = Z_t + 2Z_{t-1} - 3Z_{t-2} - 4Z_{t-3}$ with i.i.d. $N(0,1)$ noise (Z_t) of length 500. The sample ACF (left) and sample PACF (right) are shown in Figure 7.2. The sample ACF actually hints at an MA(2) process, but an MA(3) process is correct here. It is difficult to read something off from the sample PACF, but the ACF is already clear enough here.

Example 7.45. The (model) ACF (left) and PACF (right) of the AR(2) process

$$X_t + 1.4X_{t-1} + 0.45X_{t-2} = Z_t$$

are displayed in Figure 7.3.
The ACF (left) and PACF (right) of the AR(2) process

$$X_t - 1.4X_{t-1} + 0.45X_{t-2} = Z_t$$

are displayed in Figure 7.4.
Observe that these are the model ACF and PACF, not the sample ACF/PACF of a realisation. Here not too much can be read off from the ACF, but the PACF clearly reflects the findings of Corollary 7.41.

Example 7.46. The model ACF (left) and PACF (right) of the ARMA(2,1) process

$$X_t - \frac{3}{4}X_{t-1} + \frac{1}{8}X_{t-2} = Z_t + 2Z_{t-1}$$

is displayed in Figure 7.5. Here, neither the ACF or PACF becomes zero, so if confronted with such a sample ACF/PACF one would try neither an MA or an AR process, but possibly a mixture of them, i.e. an ARMA process.

Example 7.47. Consider the MA(1) process

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

Figure 7.6 shows the model ACF (left) and model PACF (right) of this process. For a simulation of 5000 data points, the sample ACF (left) and sample PACF (right) of the simulated process are shown in Figure 7.7. They approximate the model ACF/PACF quite well.

Figure 7.1: Simulated MA(3) process: $X_t = Z_t + 2Z_{t-1} - 3Z_{t-2} - 4Z_{t-3}$ of Example 7.44

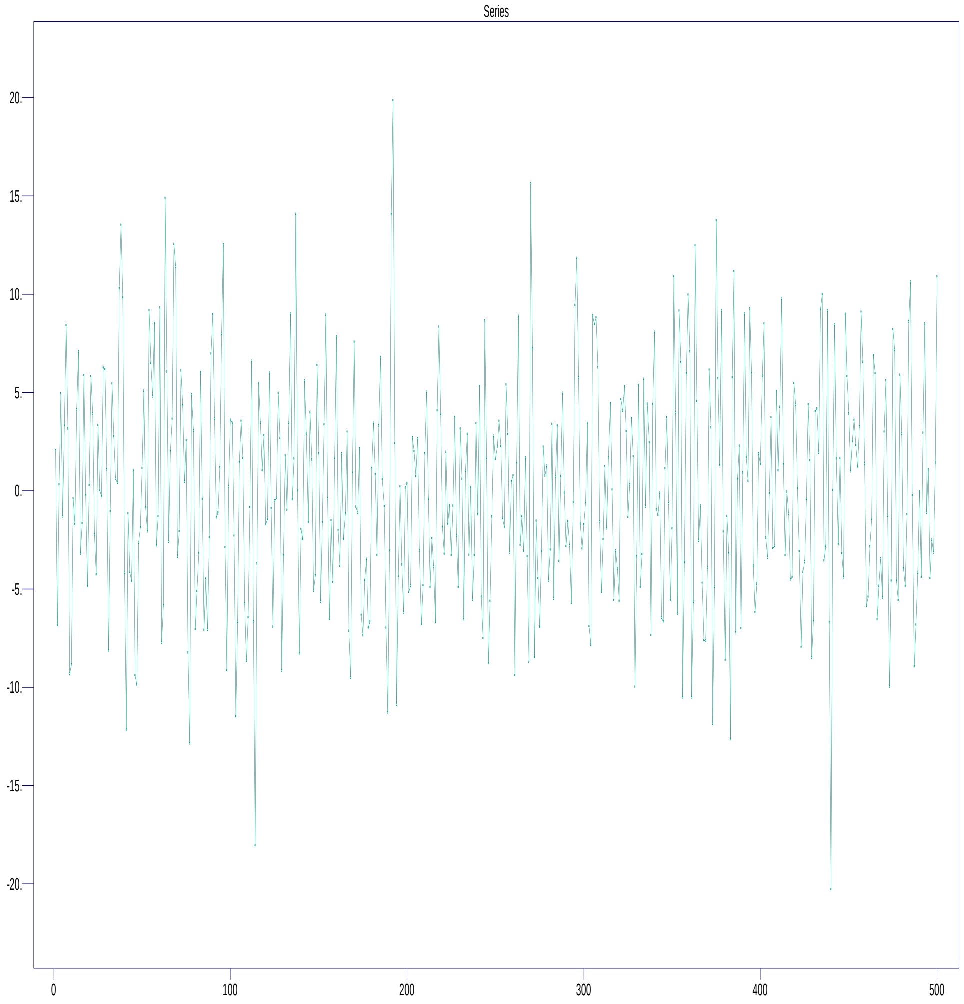


Figure 7.2: Sample ACF and sample PACF of the realisation of the MA(3) process shown in Figure 7.1

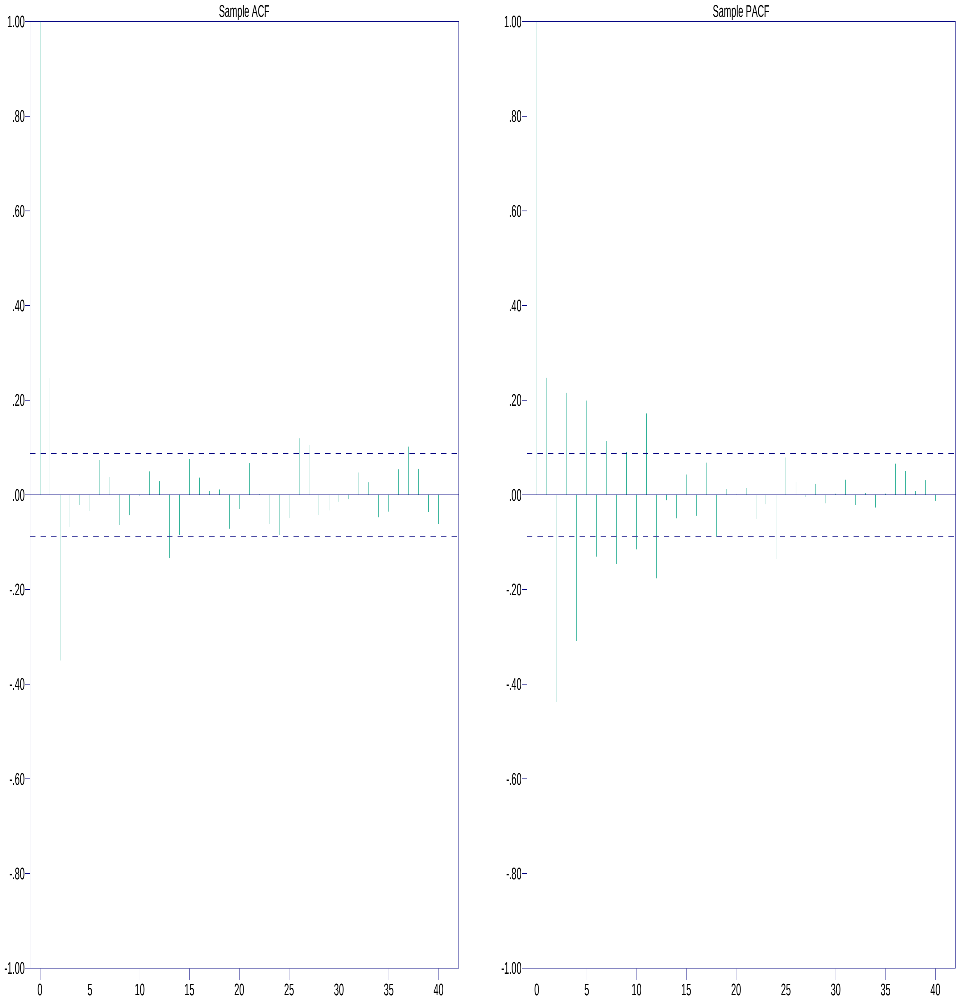


Figure 7.3: Model ACF (left) and model PACF (right) of the AR(2) process: $X_t + 1.4X_{t-1} + 0.45X_{t-2} = Z_t$

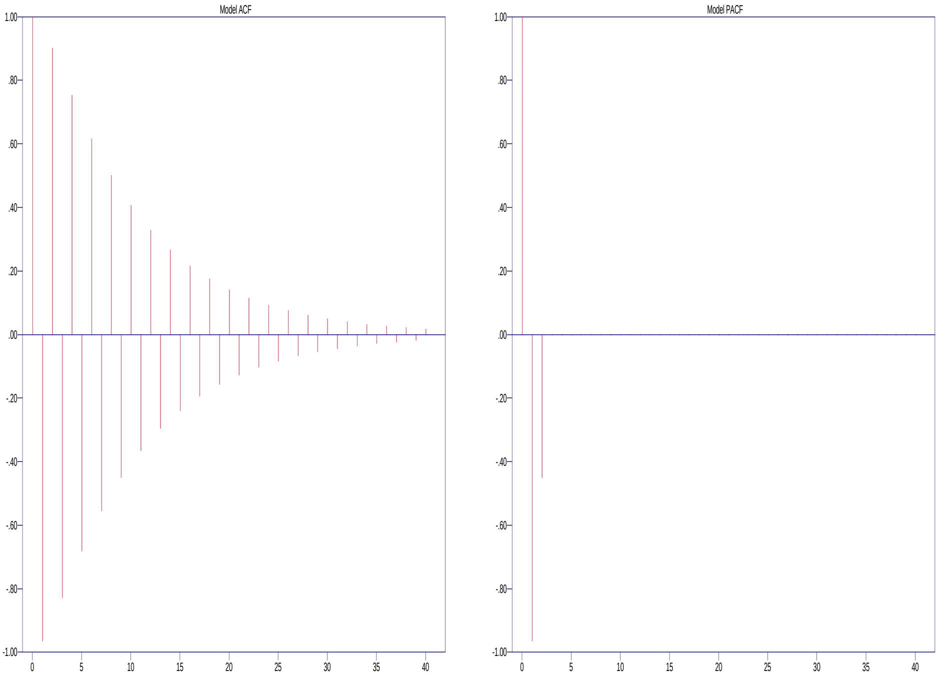


Figure 7.4: Model ACF (left) and model PACF (right) of the AR(2) process: $X_t - 1.4X_{t-1} + 0.45X_{t-2} = Z_t$

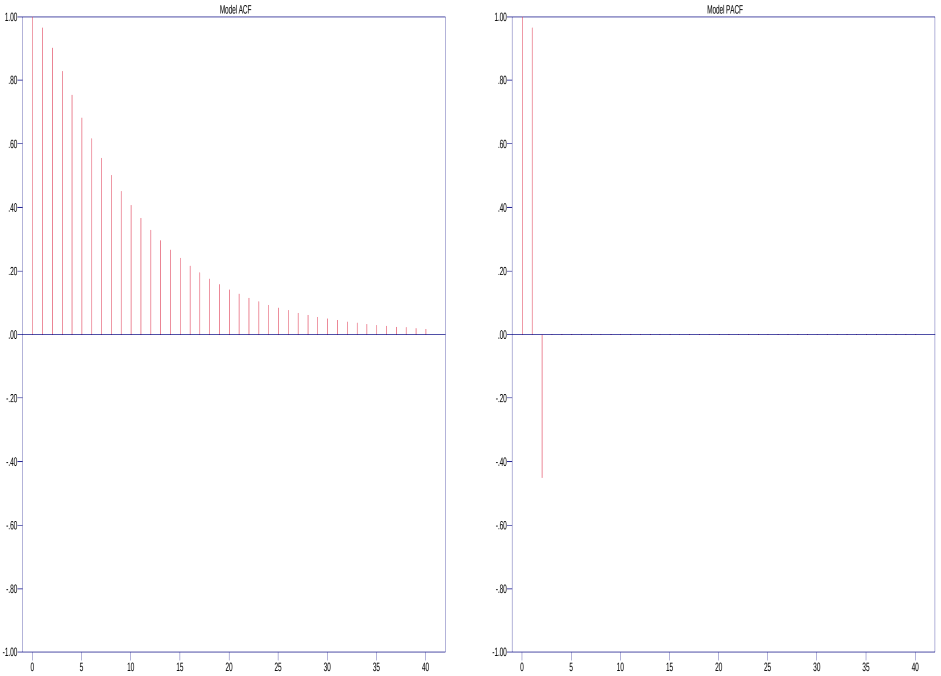


Figure 7.5: Model ACF (left) and model PACF (right) of the ARMA(2,1) process $X_t - \frac{3}{4}X_{t-1} + \frac{1}{8}X_{t-2} = Z_t + 2Z_{t-1}$

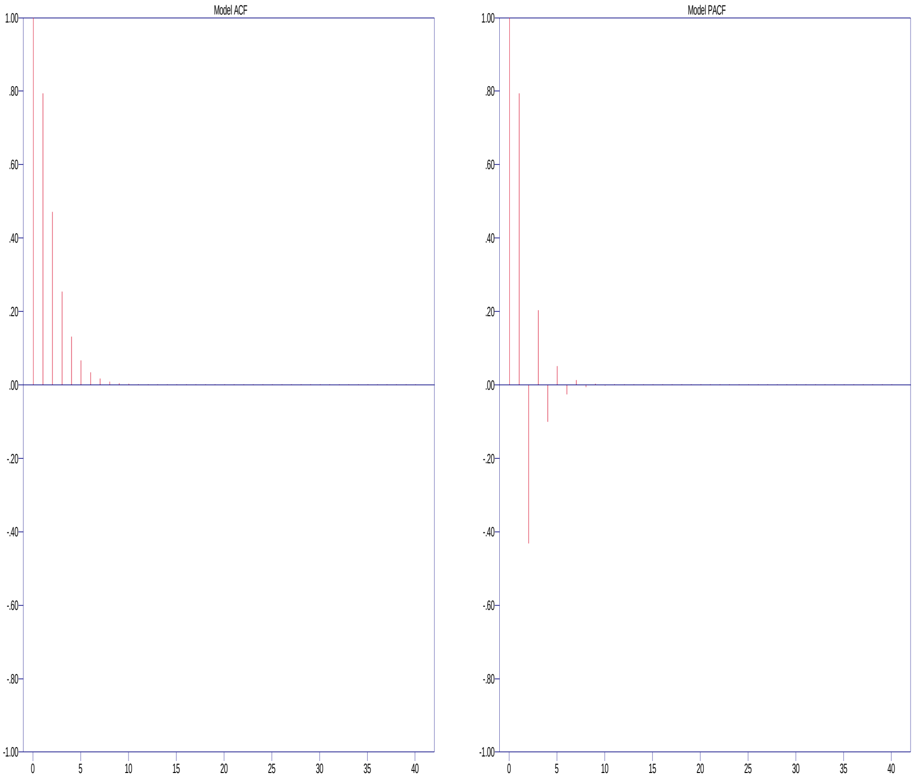


Figure 7.6: Model ACF (left) and model PACF of the MA(1) process $X_t = Z_t + 0.5Z_{t-1}$

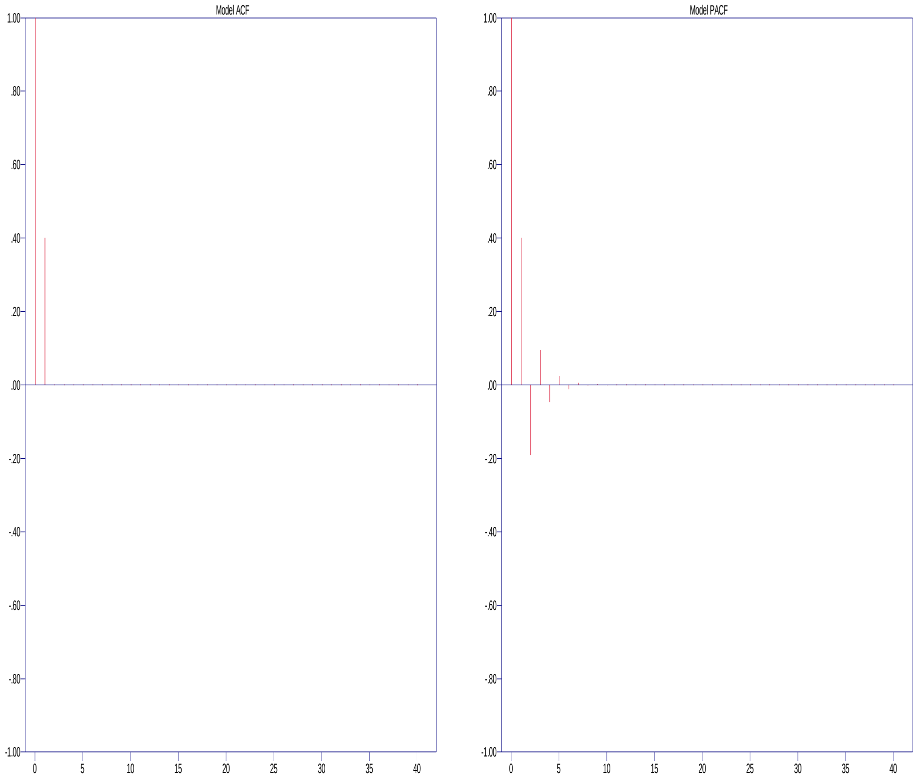
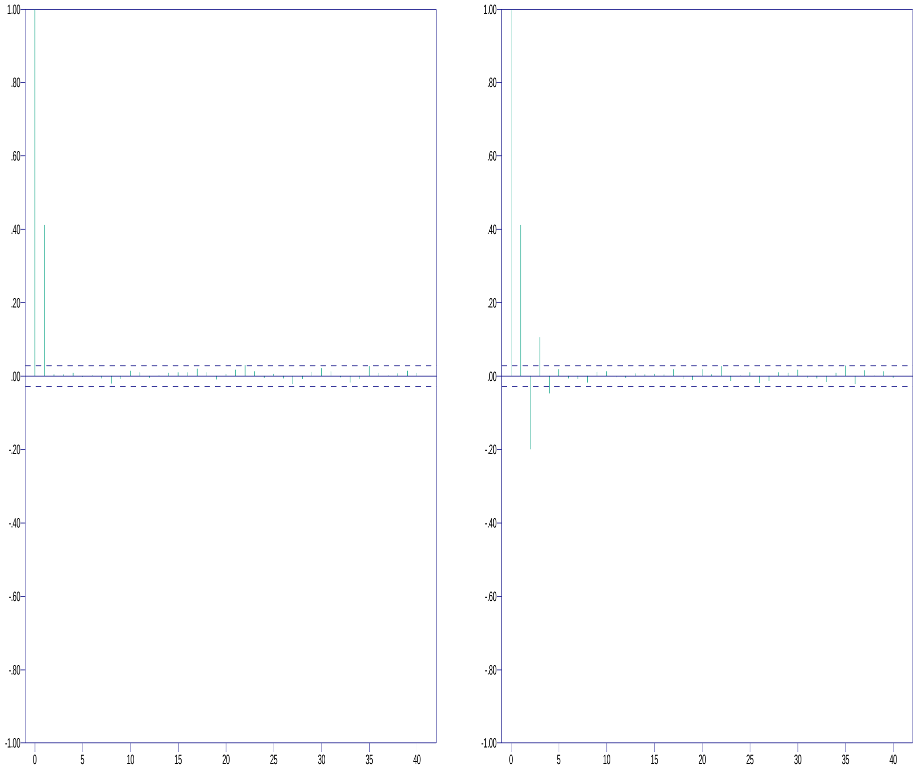


Figure 7.7: Sample ACF (left) and model PACF of a simulation of length 5000 of the MA(1) process $X_t = Z_t + 0.5Z_{t-1}$



Chapter 8

Estimation of the mean value

In this chapter we start with statistics. We would like to estimate various things in a stationary time series, e.g. the mean, the autocovariance function or if we suppose that the data come from an ARMA process, the parameters. In elementary stchastics courses and in large parts of statistics courses you probably have dealt with estimation for i.i.d. data. E.g., given i.i.d. data $(X_n)_{n\in\mathbb{N}}$ with finite expectation, it is known that the sample mean $\overline{X}_n := \frac{1}{n}\sum_{i=1}^n X_i$ converges almost surely to the true mean $\mu = \mathbb{E}(X_1)$ by the strong law of large numbers, termed *strong consistency* of the estimator. Further, if additionally the X_i have finite variance σ^2, then the central limit theorem assures that

$$\sqrt{n}(\overline{X}_n - \mathbb{E}(X_1)) \xrightarrow{d} N(0,\sigma^2) \quad \text{as} \quad n \to \infty,$$

(where $\xrightarrow{d}$ denotes convergence in distribution), so that we have the approximate distribution of $\overline{X}_n$ and can hence construct approximate confidence intervals. All this is based on the i.i.d. assumption of $(X_n)_{n\in\mathbb{N}}$. In this chapter we shall treat the asymptotic behaviour of the sample mean of a stationary time series. The data are then no longer i.i.d., but have dependencies within them. Nevertheless, we shall manage to get some limit results for stationary time series under additional assumptions.

Throughout this chapter, let $(X_t)_{t\in\mathbb{Z}}$ be a **real-valued stationary process** with expectation

$$\mu := \mathbb{E}X_t.$$

A natural estimator for μ is then the sample mean (or empirical mean), given by

$$\overline{X}_n := \frac{1}{n}\sum_{i=1}^n X_i,\ n \in \mathbb{N}.$$

Before we start proving L^2-consistency of the sample mean under wide assumptions, we recall the Césaro limit, which is just an elementary lemma from calculus:

Lemma 8.1 (Césaro limit)**.** *Let $(c_n)_{n\in\mathbb{N}}$ be a sequence of real numbers that converges to some $c \in \mathbb{R}$ as $n \to \infty$. Then*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n c_i = c.$$

Proof. Since (c_n) converges to c, for every $\varepsilon > 0$ there exists some $N(\varepsilon) \in \mathbb{N}$ such that $|c_n - c| < \varepsilon$ for all $n \in \mathbb{N}$. Then

$$\left| c - \frac{1}{n}\sum_{i=1}^{n} c_i \right| \le \frac{1}{n}\sum_{i=1}^{n} |c_i - c| \le \frac{1}{n}\sum_{i=1}^{N_\varepsilon - 1} |c_i - c| + \frac{1}{n}\sum_{i=N_\varepsilon}^{n} \varepsilon$$

for all $n \ge N_\varepsilon$, so that $\limsup_{n\to\infty} |c - \frac{1}{n}\sum_{i=1}^{n} c_i| \le 0 + \varepsilon$, and since this is true for all ε we get the claim. □

Remark 8.2. The limit $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} c_i$ is called a *Césaro limit*. Lemma 8.1 states that if a real sequence converges, then the Césaro limit exists, too, with the same limit. The converse is not true. A simple example is given by the sequence $c_n = (-1)^n$.

The following result establishes L^2-consistency of the sample mean of a stationary time series:

Theorem 8.3. *Let $(X_t)_{t\in\mathbb{Z}}$ be a real-valued stationary process with $\mathbb{E}X_t = \mu$ and autocovariance function γ. Then it holds that:*

a) If $\lim_{n\to\infty} \gamma(n) = 0$, then $\mathrm{Var}\,(\overline{X}_n) = \mathbb{E}(\overline{X}_n - \mu)^2 \to 0$ as $n \to \infty$, i.e.

$$\overline{X}_n \overset{L^2}{\to} \mu \quad \text{as} \quad n \to \infty, \tag{8.1}$$

meaning that $\overline{X}_n$ is an L^2-consistent estimator of μ. In particular,

$$\overline{X}_n \overset{P}{\to} \mu \quad \text{as} \quad n \to \infty \tag{8.2}$$

(where $\overset{P}{\to}$ denotes convergence in probability), meaning that $\overline{X}_n$ is a consistent estimator of μ.

b) If $\sum_{h\in\mathbb{Z}} |\gamma(h)| < \infty$, then $\lim_{n\to\infty}(n\mathrm{Var}\,(\overline{X}_n)) = \sum_{h\in\mathbb{Z}} \gamma(h)$.

Proof. First observe that

$$\begin{aligned} n\mathrm{Var}\,(\overline{X}_n) &= n\mathrm{Var}\,(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n}\sum_{i,j=1}^{n} \mathrm{Cov}\,(X_i, X_j) \\ &= \frac{1}{n}\sum_{|h|<n} (n - |h|)\gamma(h) \le \sum_{|h|\le n} |\gamma(h)|. \end{aligned} \tag{8.3}$$

(a) If $\lim_{n\to\infty} \gamma(h) = 0$, then it follows by considering the Césaro-limit (Lemma 8.1) that

$$\lim_{n\to\infty} \frac{1}{n}\sum_{|h|\le n} |\gamma(h)| = 2\lim_{n\to\infty} \frac{1}{n}\sum_{h=1}^{n} |\gamma(h)| = 2\lim_{h\to\infty} \gamma(h) = 0,$$

so that $\mathrm{Var}\,\overline{X}_n \to 0$ as $n \to \infty$ by (8.3), giving the claim (observe that $\mathbb{E}(\overline{X}_n) = \mu$).

(b) If $\sum_{h\in\mathbb{Z}}|\gamma(h)| < \infty$, then it follows from (8.3) and Lebesgue's dominated convergence theorem that

$$\lim_{n\to\infty} n\mathrm{Var}\,(\overline{X}_n) = \lim_{n\to\infty} \sum_{|h|<n} \underbrace{(1-\frac{|h|}{n})\gamma(h)}_{|\cdot|\le|\gamma(h)|,\ \to\gamma(h)} = \sum_{h\in\mathbb{Z}} \gamma(h).$$

□

Remark 8.4. Let $T_n = T_n(X_1,\dots,X_n)$ be a sequence of estimators for a parameter μ. If $T_n \overset{L^2}{\to} \mu$ as $n\to\infty$ for every μ in the parameter space, then the estimator T_n is called *L^2-consistent*, if T_n converges almost surely to μ as $n\to\infty$ for every μ in the parameter space, then T_n is called *strongly consistent*, and if T_n converges in probability to μ as $n\to\infty$ for all μ in the parameter space, then it is called *consistent*. Since almost sure convergence and L^2-convergence imply convergence in probability, both strong consistency and L^2-consistency imply consistency. Theorem 8.3 (a) then shows that the sample mean is an L^2-consistent and hence consistent estimator of the mean μ, and Theorem 8.3 (b) shows how quickly $\overline{X}_n$ converges in L^2 to μ.

Example 8.5. Let $(Z_t)_{t\in\mathbb{Z}} \sim WN(0,\sigma^2)$ (real valued), $\psi_j\in\mathbb{R}$ with $\sum_{j\in\mathbb{Z}}|\psi_j| < \infty$, $\mu\in\mathbb{R}$ and consider the (possibly) two-sided moving average process of (possibly) infinite order

$$X_t := \mu + \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}, \quad t\in\mathbb{Z}.$$

Then $\gamma(h) = \sigma^2 \sum\limits_{j=-\infty}^{\infty} \psi_j\psi_{j-h}$ by Corollary 5.4 and

$$\sum_{h=-\infty}^{\infty} |\gamma(h)| \le \sigma^2 \sum_{h=-\infty}^{\infty}\sum_{j=-\infty}^{\infty} |\psi_j\psi_{j-h}| = \sigma^2 \left(\sum_{j=-\infty}^{\infty} |\psi_j|\right)^2 < \infty,$$

$$\sum_{h=-\infty}^{\infty} \gamma(h) = \sigma^2 \sum_{h=-\infty}^{\infty}\sum_{j=-\infty}^{\infty} \psi_j\psi_{j-h} = \sigma^2 \left(\sum_{j=-\infty}^{\infty} \psi_j\right)^2.$$

Theorem 8.3 implies that $\overline{X}_n$ is an L^2-consistent estimator of μ with

$$\lim_{n\to\infty} n\mathrm{Var}\,(\overline{X}_n) = \sigma^2 \left(\sum_{j=-\infty}^{\infty} \psi_j\right)^2,$$

so that $\mathrm{Var}\,(\overline{X}_n) \sim \frac{\sigma^2}{n}(\sum_{j\in\mathbb{Z}}\psi_j)^2$ as $n\to\infty$. Here, the symbol $a_n \sim b_n$ as $n\to\infty$ for two sequences $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$ means that a_n/b_n converges to 1 as $n\to\infty$. If the noise $(Z_t)_{t\in\mathbb{Z}}$ is additionally i.i.d., then one can also show that the sample mean is a strongly consistent estimator for μ, see e.g. M. Taniguchi and Y. Kakizawa, Asmptotic Theory of Statistical Inference for Time Series, Springer, Theorems 1.3.4 and 1.3.5). We shall not persue this further, neither shall we need it.

Our main goal in this chapter is to derive a central limit theorem for the sample mean of a linear process $X_t = \mu + \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}$ (as in Example 8.5) if the noise is additionally i.i.d. We shall envoke the notation of asymptotic normality:

Definition 8.6. Let $\sigma \geq 0$, $(\mu_n)_{n\in\mathbb{N}}$ be a sequence of real numbers and $(v_n)_{n\in\mathbb{N}}$ be a sequence of strictly positive real numbers that converges to 0 as $n \to \infty$. Let $Y = (Y_n)_{n\in\mathbb{N}}$ be a sequence of of real-valued random variables. We write

$$Y_n \sim AN(\mu_n; v_n, \sigma^2), \quad n \to \infty,$$

if

$$\frac{1}{\sqrt{v_n}}(Y_n - \mu_n) \xrightarrow{d} N(0, \sigma^2) \quad \text{as} \quad n \to \infty,$$

and say that $(Y_n)_{n\in\mathbb{N}}$ is *asymptotically normal with mean μ_n, rate $1/\sqrt{v_n}$ and limit variance σ^2.* If additionally $\sigma > 0$, then we also say that (Y_n) is asymptotically normal with 'mean μ_n' and 'standard deviation $\sqrt{v_n}\sigma$ '.

Remark 8.7. (a) By definition, a sequence $(Y_n)_{n\in\mathbb{N}}$ of random variables *converges in distribution* to a random variable Y, if and only if the distribution P_{Y_n} of Y_n converges weakly to the distribution P_Y of Y, i.e. if

$$\lim_{n\to\infty} \int_{\mathbb{R}} f(x)\, P_{Y_n}(\mathrm{d}x) = \int_{\mathbb{R}} f(x)\, P_Y(\mathrm{d}x)$$

for all bounded and continuous functions $f : \mathbb{R} \to \mathbb{R}$. This is equivalent to saying that

$$\lim_{n\to\infty} \mathbb{E} f(Y_n) = \mathbb{E} f(Y)$$

for all bounded and continuous functions $f : \mathbb{R} \to \mathbb{R}$. By Lévy's continuity theorem, this is further equivalent to saying that

$$\lim_{n\to\infty} \varphi_{Y_n}(t) = \varphi_Y(t) \quad \forall\, t \in \mathbb{R},$$

where $\varphi_Y(t) = \mathbb{E}\mathrm{e}^{\mathrm{i}Yt}$ denotes the characteristic function of Y at t. We write $Y_n \xrightarrow{d} Y$ or also $Y_n \xrightarrow{d} P_Y$ to indicate convergence in distribution of Y_n to Y.
(b) If $\sigma > 0$ in the situation of Definition 8.6, denote $\sigma_n := \sqrt{v_n}\sigma$. Then $Y_n \sim AN(\mu_n; v_n, \sigma^2)$ if and only if

$$\frac{1}{\sigma_n}(Y_n - \mu_n) \xrightarrow{d} N(0, 1) \quad \text{as} \quad n \to \infty.$$

So in this case, the quantities v_n and σ^2 enter only in terms of the product $\sigma_n = v_n\sigma$, and it is custom to write then $Y_n \sim AN(\mu_n; \sigma_n^2)$. The reason why I chose the notation with three parameters is that when $\sigma = 0$ we cannot recover the rate v_n from $v_n\sigma = 0$. To say that $Y_n \sim AN(\mu_n; v_n, 0)$ then means that

$$\frac{1}{\sqrt{v_n}}(Y_n - \mu_n) \xrightarrow{d} N(0, 0) = \delta_0 \quad (n \to \infty),$$

where δ_0 is the Dirac measure at 0, so that $\frac{1}{\sqrt{v_n}}(Y_n - \mu_n)$ converges in distribution to 0. Since the rate v_n is important here, this cannot be reflected by just saying $Y_n \sim AN(\mu_n; v_n 0) = AN(\mu_n; 0)$.
(c) The sequences μ_n and $\sigma_n = \sqrt{v_n}\sigma$ do not have to reflect the true mean and standard deviation of Y_n, i.e. it can be that $\mu_n \neq \mathbb{E}(Y_n)$ and $\sigma_n \neq \sqrt{\mathrm{Var}\,(Y_n)}$. Indeed, the sequences μ_n and σ_n do not even have to be unique, since if $(\mu_n, \sigma_n)_{n\in\mathbb{N}}$ is a sequence such that $(Y_n - \mu_n)/\sigma_n \overset{d}{\to} N(0,1)$ as $n \to \infty$ (where $\sigma_n > 0$ and $\sigma_n \to 0$), and if σ_n' is another sequence such that $\sigma_n'/\sigma_n \to 1$ as $n \to \infty$, then also $(Y_n - \mu_n)/\sigma_n' \overset{d}{\to} N(0,1)$ by Slutsky's lemma. Further, if (μ_n') is a sequence such that $(\mu_n' - \mu_n)/\sigma_n \to 0$ (e.g. $\mu_n' := \mu_n + \sigma_n/n$), then also $(Y_n - \mu_n')/\sigma_n \overset{d}{\to} N(0,1)$ by Slutsky's lemma. It can even happen that $(Y_n - \mu_n)/\sigma_n \overset{d}{\to} N(0,1)$ for some sequences (μ_n) and (σ_n) if the Y_n have infinite variance (hence no standard deviation exists; see e.g. Kallenberg, Foundations of Modern Probability, Theorem 5.17 for an example which arises as a sum of i.i.d. random variables without finite second moment).

Before we can prove asymptotic normality of the sample mean, we need the following useful result which is sometimes called a generalised version of Slutsky's lemma. Slutsky's classical lemma states that if two sequences (X_n) and (Y_n) defined on the same probability space convergence in distribution to X and Y, respectively, where Y is constant, then the sum $X_n + Y_n$ and the product $X_n Y_n$ converge in distribution to $X + Y$ and $X_n Y_n$, respectively.

The result which is sometimes called generalised lemma of Slutsky is concerned with double series of random variables and reads as follows. We formulate it immediately for convergence in distribution of random vectors (the corresponding definitions for convergence in distribution of Remark 8.7 (a) carry over word by word to the multivariate setting):

Theorem 8.8. *Let $(\Omega, \mathcal{F}, P)$ be a probability space and $(X_n)_{n\in\mathbb{N}}$ and $(Y_{nj})_{j\in\mathbb{N}, n\in\mathbb{N}}$ be (double) sequences of $\mathbb{R}^k$-valued random vectors on $(\Omega, \mathcal{F}, P)$ such that*

i) for all $j \in \mathbb{N}$: $Y_{nj} \overset{d}{\to} Y_j$ as $n \to \infty$ for some random vector Y_j,

ii) $Y_j \overset{d}{\to} Y$ as $j \to \infty$ for some random vector Y and

iii) $\lim_{j\to\infty} \limsup_{n\to\infty} P(|X_n - Y_{nj}| > \varepsilon) = 0\ \forall \varepsilon > 0$.

Then $X_n \overset{d}{\to} Y$ as $n \to \infty$.

Proof. For the characteristic functions $\varphi_{X_n}(t)$ of X_n at $t \in \mathbb{R}^k$ and $\varphi_Y(t)$ of Y at t we have (here, $\langle \cdot, \cdot \rangle$ denotes the standard inner product in $\mathbb{R}^k$)

$$\begin{aligned}
&|\varphi_{X_n}(t) - \varphi_Y(t)| = |\mathbb{E}e^{i\langle t, X_n\rangle} - \mathbb{E}e^{i\langle t, Y\rangle}| \\
\leq &|\mathbb{E}e^{i\langle t, X_n\rangle} - \mathbb{E}e^{i\langle t, Y_{nj}\rangle}| + |\mathbb{E}e^{i\langle t, Y_{nj}\rangle} - \mathbb{E}e^{i\langle t, Y_j\rangle}| + |\mathbb{E}e^{i\langle t, Y_j\rangle} - \mathbb{E}e^{i\langle t, Y\rangle}| \\
=: &I_{n,j} + II_{n,j} + III_j, \quad \text{say.}
\end{aligned}$$

Then for any $\varepsilon > 0$ we have

$$\begin{aligned} I_{n,j} &\le \mathbb{E}|e^{i\langle t,X_n\rangle} - e^{i\langle t,Y_{nj}\rangle}| \\ &= \mathbb{E}|e^{i\langle t,X_n\rangle}(1 - e^{i\langle t,Y_{nj}-X_n\rangle})| \\ &= \mathbb{E}(|1 - e^{i\langle t,Y_{nj}-X_n\rangle}|\mathbf{1}_{\{|Y_{nj}-X_n|>\varepsilon\}}) + \mathbb{E}(|1 - e^{i\langle t,Y_{nj}-X_n\rangle}|\mathbf{1}_{\{|Y_{nj}-X_n|\le\varepsilon\}}) \\ &:= A_{n,j,\varepsilon} + B_{n,j,\varepsilon}, \quad \text{say.} \end{aligned}$$

Since $A_{n,j,\varepsilon} \le 2P(|Y_{nj} - X_n| > \varepsilon)$ we have $\lim_{j\to\infty} \limsup_{n\to\infty} A_{n,j\varepsilon} = 0$ by *iii)*.
Now choose for given $t \in \mathbb{R}^k$ and $\delta > 0$ an $\varepsilon > 0$ such that

$$|1 - e^{i\langle t,y-x\rangle}| \le \delta \quad \forall\, |y - x| \ \le \varepsilon.$$

Then $\limsup_{j\to\infty} \limsup_{n\to\infty} B_{n,j,\varepsilon} \le \delta$ so that $\limsup_{j\to\infty} \limsup_{n\to\infty} I_{n,j} \le \delta$ for every $\delta > 0$, and letting $\delta \downarrow 0$ we obtain

$$\lim_{j\to\infty} \limsup_{n\to\infty}(I_{n,j}) = 0.$$

Furthermore $\lim_{n\to\infty}(II_{n,j}) = 0$ by *i)* and $\lim_{j\to\infty}(III_j) = 0$ by *ii)*.
This implies

$$\lim_{n\to\infty} |\varphi_{X_n}(t) - \varphi_Y(t)| \le \lim_{j\to\infty} \limsup_{n\to\infty}(I_{j,n} + II_{j,n} + III_{j,n}) = 0,$$

so that $\varphi_{X_n}(t)$ converges to $\varphi_Y(t)$ as $n \to \infty$ for each fixed t, hence showing the required convergence in distribution by Lévy's continuity theorem. □

Before we come to the desired limit theorem for the linear processes $X_t = \mu + \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}$ with i.i.d. innovations, we first treat so called m-dependent sequences.

Definition 8.9. Let $m \in \mathbb{N}$. A sequence $(X_t)_{t\in\mathbb{Z}}$ is called *m-dependent*, if for all $t \in \mathbb{Z}$ the sequences $(X_j)_{j\le t}$ and $(X_j)_{j\ge t+m+1}$ are independent.

Maybe the notion m-dependent is a bit misleading and a better notion would be $(m+1)$-independent, because the difference of the two neighbouring endpoints is $(t+m+1)-t = m+1$ and those are in particular independent, but the standard notion is to call these sequences m-dependent and we stick to it.

Example 8.10. An MA(m)-process with i.i.d. noise is $m-$dependent. This is clear.

Now we come to the central limit theorem for m-dependent strictly stationary sequences with finite variance. It is due to Hoeffing and Robbins (1948):

Theorem 8.11 (Central Limit Theorem for m-dependent strictly stationary sequences)**.** *Let $(X_t)_{t\in\mathbb{Z}}$ be a strictly stationary real-valued m-dependent sequence of random variables with finite variance and $\mathbb{E}X_t = 0$. Denote by γ the autocovariance function of X (since X is strictly stationary with finite variance, it is also weakly stationary). Denote*

$$v_m := \sum_{|h|\le m} \gamma(h).$$

Then

i) $\lim_{n\to\infty} n\mathrm{Var}\,(\overline{X}_n) = v_m$ *and*

ii) $\overline{X}_n \sim AN(0; \frac{1}{n}, v_m)$ *as* $n \to \infty$.

Proof. (i) Since $(X_t)_{t\in\mathbb{Z}}$ is m-dependent, we have $\gamma(h) = 0$ for $|h| > m$. From (8.3) we then have

$$n\mathrm{Var}\,\overline{X}_n = \sum_{|h|<n} \left(1 - \frac{|h|}{n}\right)\gamma(h) \overset{n\geq m}{=} \sum_{|h|\leq m} \left(1 - \frac{|h|}{n}\right)\gamma(h) \to v_m = \sum_{|h|\leq m} \gamma(h) \quad \text{as} \quad n \to \infty.$$

(ii) We will apply Theorem 8.8. For that, let $k \in \mathbb{N}$ such that $k > 2m$. For $r = \lfloor \frac{n}{k} \rfloor$ (the largest integer smaller or equal than n/k) define

$$Y_{nk} := \frac{1}{\sqrt{n}} \left(\underbrace{(X_1 + \ldots + X_{k-m})}_{=:Z_1} + \underbrace{(X_{k+1} + \ldots + X_{2k-m})}_{=:Z_2} + \ldots + \underbrace{(X_{(r-1)k+1} + \ldots + X_{rk-m})}_{=:Z_r} \right).$$

Since $(X_t)_{t\in\mathbb{Z}}$ is m-dependent and strictly stationary, $\sqrt{n}Y_{nk} = Z_1 + \ldots + Z_r$ is the sum of r i.i.d. random variables. Every summand Z_j has mean 0 and variance

$$R_{k-m} = \mathrm{Var}\,(X_1 + \ldots + X_{k-m}) = \sum_{|j|\leq m} (k - m - |j|)\gamma(j).$$

Then the classical central limit theorem applies and we obtain when $R_{k-m} > 0$ that

$$\frac{\sqrt{n}Y_{nk}}{\sqrt{rR_{k-m}}} = \frac{Z_1 + \ldots + Z_r}{\sqrt{\mathrm{Var}\,(Z_1 + \ldots + Z_r)}} \xrightarrow{d} N(0,1) \quad \text{as} \quad n \to \infty,$$

so that

$$Y_{nk} = \sqrt{\frac{r}{n}R_{k-m}} \frac{\sqrt{n}Y_{nk}}{\sqrt{rR_{k-m}}} \xrightarrow{d} N(0, \frac{1}{k}R_{k-m}) \quad \text{as} \, n \to \infty$$

by Slutsky's lemma since $\lim_{n\to\infty} r/n = 1/k$. Similarly, if $R_{k-m} = 0$ we have $X_1 + \ldots + X_{k-m} = 0$ almost surely (having mean and variance 0), hence $Z_1 = 0$ and hence $Z_2 = \ldots = Z_r = 0$, so that $Y_{nk} = 0$ and hence $Y_{nk} \xrightarrow{d} N(0,0) = N(0, \frac{1}{k}R_{k-m})$ as $n \to \infty$.

Now let Y_k be an $N(0, \frac{1}{k}R_{k-m})$-distributed random variable (regardless if $R_{k-m} \neq 0$ or not). Then we have just seen that $Y_{nk} \xrightarrow{d} Y_k$ as $n \to \infty$, so that condition *i)* of Theorem 8.8 is satisfied.

Next, let Y be an $N(0, v_m)$-distributed random variable. Since

$$\frac{1}{k}R_{k-m} = \sum_{|j|\leq m} \left(1 - \frac{m + |j|}{k}\right)\gamma(j) \to v_m \quad \text{as} \quad k \to \infty,$$

it follows that $Y_k \xrightarrow{d} Y \overset{d}{=} N(0, v_m)$ as $k \to \infty$, since a sequence of normal distributions converges to a normal distribution if and only if the mean and the standard deviation converge to the corresponding limit. We see that condition *ii)* of Theorem 8.8 is satisfied.

Finally, to obtain condition *iii)* of Theorem 8.8, write

$$\sqrt{n}\overline{X}_n - Y_{nk} = \frac{1}{\sqrt{n}}\sum_{j=1}^{r-1}(X_{jk-m+1} + X_{jk-m+2} + \ldots + X_{jk}) + \frac{1}{\sqrt{n}}(X_{rk-m+1} + \ldots + X_n)$$

Then

$$\operatorname{Var}(\sqrt{n}\overline{X}_n - Y_{nk}) = \frac{1}{n}((r-1)R_m + R_{h(n)}),$$

where

$$R_m = \operatorname{Var}(X_1 + \ldots + X_m),\ R_{h(n)} = \operatorname{Var}(X_1 + \ldots + X_{h(n)}),\ h(n) := n - k\lfloor\frac{n}{k}\rfloor + m.$$

Observe that R_m is independent of n and that $0 \le h(n) \le k + m$. This implies that the sequence $(R_{h(n)})_{n\in\mathbb{N}}$ is bounded (for fixed k). This gives

$$\limsup_{n\to\infty}\operatorname{Var}(\sqrt{n}\overline{X}_n - Y_{nk}) \le \limsup_{n\to\infty}\frac{1}{n}(\frac{n}{k}R_m + R_{h(n)}) = \frac{1}{k}R_m.$$

An application of Chebyshev's inequality then gives

$$\lim_{k\to\infty}\limsup_{n\to\infty}P(|\sqrt{n}\overline{X}_n - Y_{nk}| > \varepsilon) \le \lim_{k\to\infty}\frac{R_m/k}{\varepsilon^2} = 0 \quad \forall\,\varepsilon > 0,$$

so that condition *iii)* of Theorem 8.8 is satisfied. Hence all conditions of Theorem 8.8 are satisfied and we obtain $\sqrt{n}\overline{X}_n \xrightarrow{d} N(0, v_m)$ as $n \to \infty$. □

Example 8.12. Let X be an MA(q) process with i.i.d. $(0,\sigma^2)$-noise and real coefficients $(\theta_j)_{j=1,\ldots,q}$. Then, with $\theta_0 := 1$, $v_q = \sigma^2\left(\sum_{j=0}^{q}\theta_j\right)^2$ by Example 8.5, so

$$\overline{X}_n \sim AN\left(0; \frac{1}{n}, \sigma^2\left(\sum_{j=0}^{q}\theta_j\right)^2\right) \quad \text{as} \quad n \to \infty.$$

We now come to the main result of this chapter, which gives asymptotic normality of the sample mean for linear processes (with i.i.d. noise):

Theorem 8.13 (Asymptotic normality of $\overline{X}_n$)**.**
Let $\mu \in \mathbb{R}$, $(\psi_j)_{j\in\mathbb{Z}}$ an absolutely summable sequence of real numbers and $Z = (Z_t)_{t\in\mathbb{Z}}$ a real valued i.i.d. noise with mean zero and finite variance σ^2. Define $X = (X_t)_{t\in\mathbb{Z}}$ by

$$X_t = \mu + \sum_{j=-\infty}^{\infty}\psi_j Z_{t-j}, \quad t \in \mathbb{Z}.$$

Denote the autocovariance function of X by γ. Then the sample mean $\overline{X}_n$ is asymptotically normal, more precisely

$$\overline{X}_n \sim AN\left(\mu; \frac{1}{n}, \sigma^2\left(\sum_{j=-\infty}^{\infty}\psi_j\right)^2\right) = AN\left(\mu; \frac{1}{n}, \sum_{h=-\infty}^{\infty}\gamma(h)\right) \quad \textit{as} \quad n \to \infty.$$

Proof. Observe first that $X = (X_t)_{t\in\mathbb{Z}}$ is weakly and strictly stationary by Theorem 5.3. Now for each $m \in \mathbb{N}$ define with a truncated sum

$$X_{t,m} := \mu + \sum_{j=-m}^{m} \psi_j Z_{t-j}, \quad t \in \mathbb{Z}, \quad \text{and}$$

$$Y_{nm} := \overline{X}_{n,m} = \frac{1}{n}\sum_{t=1}^{n} X_{t,m}, \quad n \in \mathbb{N}.$$

Then $(X_{t,m})_{t\in\mathbb{Z}}$ is $2m+1$ dependent, strictly stationary and has finite variance. If γ_m denotes its autocovariance function, then

$$\sum_{|h|\le 2m+1} \gamma_m(h) = \sum_{h\in\mathbb{Z}} \gamma_m(h) = \sigma^2 \left(\sum_{j=-m}^{m} \psi_j\right)^2$$

by Example 8.5. Now let Y_m be an $N(0, \sigma^2(\sum_{j=-m}^{m}\psi_j)^2)$-distributed random variable. Theorem 8.11 then implies that

$$\sqrt{n}(Y_{nm} - \mu) \xrightarrow{d} Y_m \quad \text{as} \quad n \to \infty.$$

Further, $\sigma^2(\sum_{j=-m}^{m}\psi_j)^2 \to \sigma^2(\sum_{j=-\infty}^{\infty}\psi_j)^2$ as $m \to \infty$, hence

$$Y_m \xrightarrow{d} N(0, \sigma^2(\sum_{j=-\infty}^{\infty} \psi_j)^2), \quad \text{as} \quad m \to \infty.$$

Finally, since also $(X_t - X_{t,m})_{t\in\mathbb{Z}} = (\sum_{|j|>m}\psi_j Z_{t-j})_{t\in\mathbb{Z}}$ is a linear process, we obtain from Example 8.5 that

$$\mathrm{Var}\,(\sqrt{n}(\overline{X}_n - Y_{nm})) = n\mathrm{Var}\,(\frac{1}{n}\sum_{t=1}^{n}\sum_{|j|>m}\psi_j Z_{t-j}) \to \sigma^2 \left(\sum_{|j|>m}\psi_j\right)^2, \quad n \to \infty.$$

An application of Chebyshev's inequality then shows

$$\lim_{m\to\infty} \limsup_{n\to\infty} P(|\sqrt{n}(\overline{X}_n - Y_{nm})| > \varepsilon) = 0 \quad \forall\, \varepsilon > 0.$$

Hence all conditions of Theorem 8.8 are satisfied and we conclude that $\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2\left(\sum \psi_j\right)^2)$, which is the claim (compare also Example 8.5 for the second form of the description of the limit). □

Remark 8.14. Theorems 8.11 and 8.13 give us approximate confidence intervals for the estimator $\overline{X}_n$. They imply that $\overline{X}_n - \mu$ is approximately distributed as $N(0, v/n)$ with $v = \sum_{j\in\mathbb{Z}}\gamma(h)$. Hence, if $\Phi_{1-\alpha/2}$ denotes the $(1-\alpha/2)$-quantile of the standard normal distribution, then μ is contained in the random interval $[\overline{X}_n - \sqrt{v/n}\Phi_{1-\alpha/2}, \overline{X}_n + \sqrt{v/n}\Phi_{1-\alpha/2}]$ with approximate probability $1-\alpha$.

Remark 8.15. If the process $X = (X_t)_{t\in\mathbb{Z}}$ is a stationary Gaussian process with mean μ (regardless if it is m-dependent or a linear process), then also $\overline{X}_n$ is Gaussian with mean μ and variance given by

$$\operatorname{Var} \overline{X}_n = \frac{1}{n} \sum_{|h|<n} \left(1 - \frac{|h|}{n}\right) \gamma(h)$$

(cf. (8.3)). Hence $\sqrt{n}(\overline{X}_n - \mu) \overset{d}{=} N(0, \sum_{|j|<n} \left(1 - \frac{|h|}{n}\right) \gamma(h))$ in this case. This allows again to construct confidence intervals, this time however exact confidence intervals and not only approximate ones.

Chapter 9

Estimation of the autocovariance function

Having estimated the mean of a stationary time series, we can now attack a more difficult task, namely the estimation of the autocovariance and of the autocorrelation function.

Throughout this chapter, we will consider a real-valued stationary time series $(X_t)_{t\in\mathbb{Z}}$ with autocovariance function γ. We assume that we have observations $X_1,\dots,X_n$ (it would be better to write $x_1,\dots,x_n$, but as here we are interested in the probabilistic properties of the estimators, we take capitel letters). Denoting the sample mean by

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

the empirical autocovariance function at lag $h \in \{0, 1\dots, n-1\}$ is given by

$$\widehat{\gamma}(h) := \frac{1}{n}\sum_{t=1}^{n-h}(X_t - \overline{X}_n)(X_{t+h} - \overline{X}_n).$$

It seems to be a natural estimator of $\gamma(h)$. If $\widehat{\gamma}(0) \neq 0$, then the empirical autocorrelation

$$\widehat{\rho}(h) := \frac{\widehat{\gamma}(h)}{\widehat{\gamma}(0)}, \quad 0 \le h \le n-1,$$

is a natural estimator of $\rho(h) = \frac{\gamma(h)}{\gamma(0)}$.

Observe that $\widehat{\rho}(h)$ is a function of the bivariate vector $(\widehat{\gamma}(0), \widehat{\gamma}(h))'$. In order to get asymptotic results of $\widehat{\rho}(h)$ we need multidimensional distributions. As a first result we recall (and prove) the Cramér-Wold device, which allows to reduce convergence in distribution of random vectors to convergence in distribution of random variables arising from linear combinations of the components. More precisely, we have

Theorem 9.1 (Cramér–Wold device).
Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of $\mathbb{R}^k$-valued random vectors and X an $\mathbb{R}^k$-valued random vector. Then the following statements are equivalent:

(i) $X_n \xrightarrow{d} X$ *as* $n \to \infty$.

(ii) $\lambda' X_n \xrightarrow{d} \lambda' X$ *as* $n \to \infty$ *for all* $\lambda \in \mathbb{R}^k$.

Proof. Denote by φ_{X_n} and φ_X the characteristic functions of X_n and X, respectively, and similarly the characteristic functions of $\lambda' X_n$ and $\lambda' X$. Then we have from Levy's continuity theorem

$$\begin{aligned}
X_n \xrightarrow{d} X \quad (n \to \infty) \quad &\Longleftrightarrow \quad \lim_{n\to\infty} \varphi_{X_n}(\lambda) = \varphi_X(\lambda) \quad \forall\, \lambda \in \mathbb{R}^k \\
&\Longleftrightarrow \quad \lim_{n\to\infty} \mathbb{E}\mathrm{e}^{\mathrm{i}\lambda' X_n} = \mathbb{E}\mathrm{e}^{\mathrm{i}\lambda' X} \quad \forall\, \lambda \in \mathbb{R}^k \\
&\Longleftrightarrow \quad \lim_{n\to\infty} \varphi_{\lambda' X_n}(1) = \varphi_{\lambda' X}(1) \quad \forall\, \lambda \in \mathbb{R}^k \\
&\Longleftarrow \quad \lim_{n\to\infty} \varphi_{\lambda' X_n}(u) = \varphi_{\lambda' X}(u) \quad \forall\, u \in \mathbb{R} \quad \forall\, \lambda \in \mathbb{R}^k \\
&\Longleftrightarrow \quad \lambda' X_n \xrightarrow{d} \lambda' X \quad (n \to \infty) \quad \forall\, \lambda \in \mathbb{R}^k.
\end{aligned}$$

This shows that (ii) implies (i). The converse direction is easier, because if $X_n \xrightarrow{d} X$ as $n \to \infty$, then $f(X_n) \xrightarrow{d} f(X)$ as $n \to \infty$ for every continuous function $f : \mathbb{R}^k \to \mathbb{R}^m$, and since $\mathbb{R}^k \to \mathbb{R}$, $x \mapsto \lambda' x$ is a continuous function for every fixed $\lambda \in \mathbb{R}^k$, we see that (i) implies (ii). □

Remark 9.2. Recall that by definition a random vector $X : \Omega \to \mathbb{R}^k$ is normally distributed, if and only if $\lambda' X$ is normally distributed $\forall \lambda \in \mathbb{R}^k$. Further, for a given vector $\mu \in \mathbb{R}^k$ and a covariance matrix $\Sigma \in \mathbb{R}^{k\times k}$ we have $X \overset{d}{=} N(\mu, \Sigma)$ if and only if $\lambda' X \overset{d}{=} N(\lambda'\mu, \lambda'\Sigma\lambda)$ for all $\lambda \in \mathbb{R}^k$.

Let us generalise Definition 8.6 to the multivariate setting.

Definition 9.3. Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of $\mathbb{R}^k$-valued random vectors, $(\mu_n)_{n\in\mathbb{N}}$ a sequence of vectors in $\mathbb{R}^k$ and $\Sigma \in \mathbb{R}^{k\times k}$ a covariance matrix (i.e. it is a symmetric and non-negative definite matrix). Let $(v_n)_{n\in\mathbb{N}}$ be a sequence of real numbers with $v_n > 0$ and $\lim_{n\to\infty} v_n = 0$. Then $(X_n)_{n\in\mathbb{N}}$ is called *asymptotically normal with mean* μ_n*, rate* $1/\sqrt{v_n}$ *and limit covariance matrix* Σ*,* in symbols

$$X_n \sim AN(\mu_n; v_n, \Sigma), \quad n \to \infty,$$

if

$$\frac{1}{\sqrt{v_n}}(X_n - \mu_n) \xrightarrow{d} N(0, \Sigma) \quad \text{as} \quad n \to \infty.$$

For $k = 1$ this coincides with Definition 8.6. With the aid of the Cramér-Wold device, multivariate asymptotic normality can can be reduced to univariate asymptotic normality. More precisely, we have:

Corollary 9.4. *Let* $(X_n)_{n\in\mathbb{N}}$ *be a sequence of* $\mathbb{R}^k$*-valued random vectors,* $(\mu_n)_{n\in\mathbb{N}}$ *a sequence in* $\mathbb{R}^k$ *and* $\Sigma \in \mathbb{R}^{k\times k}$ *a covariance matrix. Let* $(v_n)_{n\in\mathbb{N}}$ *be a sequence of real numbers with* $v_n > 0$ *and* $\lim_{n\to\infty} v_n = 0$*. Then the following statements are equivalent:*

(i) $X_n \sim AN(\mu_n; v_n, \Sigma)$ *as* $n \to \infty$.

(ii) $\lambda' X_n \sim AN(\lambda'\mu_n; v_n, \lambda'\Sigma\lambda)$ *as* $n \to \infty$ *for all* $\lambda \in \mathbb{R}^k$.

Proof. This is immediate from Theorem 9.1 and Remark 9.2. □

Remark 9.5. Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of $\mathbb{R}^k$-valued random vectors and X an $\mathbb{R}^k$-valued random vector. We defined convergence in distribution of random vectors in terms of $\int_{\mathbb{R}^k} f(x)\,\mathrm{d}P_{X_n}(x) \to \int_{\mathbb{R}^k} f(x)\,\mathrm{d}P_X(x)$ for every bounded and continuous function $f : \mathbb{R}^k \to \mathbb{R}$, and mentioned that it is equivalent to pointwise convergence of the characteristic functions. There are however also other characterisations of convergence in distribution. The *Portmanteau theorem* states that the following are equivalent:

(i) $X_n \xrightarrow{d} X$ as $n \to \infty$.

(ii) $\limsup_{n\to\infty} P(X_n \in F) \leq P(X \in F)$ for all closed subsets F of $\mathbb{R}^k$.

(iii) $\liminf_{n\to\infty} P(X_n \in G) \geq P(X \in G)$ for all open subsets G of $\mathbb{R}^k$.

(iv) $\lim_{n\to\infty} P(X_n \in A) = P(X \in A)$ for all Borel sets $A \subset \mathbb{R}^k$ with $P(X \in \partial A) = 0$, where ∂A denotes the topological boundary of A.

I assume that you have come across this theorem in your probability course. If not, a proof can be found e.g. in Billingsley, Probability and Measure, Theorem 29.1, or Klenke, Probability Theory, Theorem 13.16.

We know from probability theory that if X_n converges in distribution to X and if $f : \mathbb{R}^k \to \mathbb{R}^m$ is continuous, then $f(X_n)$ converges in distribution to $f(X)$. But what now if X_n is asymptotically normal, say $X_n \sim AN(\theta; v_n, \Sigma)$ for a fixed vector θ, so that

$$\frac{1}{\sqrt{v_n}}(X_n - \theta) \xrightarrow{d} N(0,\Sigma) \quad \text{as } n \to \infty?$$

If f is a continuous function, then it follows that

$$f\left(\frac{1}{\sqrt{v_n}}(X_n - \theta)\right) \xrightarrow{d} f(Z), \quad \text{where} \quad Z \overset{d}{=} N(0,\Sigma),$$

but this is maybe not what we want. We are more interested in the question whether $f(X_n)$ is also asymptotically normal, i.e. if there are a vector $\mu \in \mathbb{R}^m$, a covariance matrix $W \in \mathbb{R}^{m\times m}$ and a sequence $(w_n)_{n\in\mathbb{N}}$ of strictly positive real numbers tending to zero such that

$$\frac{1}{\sqrt{w_n}}(f(X_n) - \mu) \xrightarrow{d} N(0,W) \quad \text{as} \quad n \to \infty.$$

The next theorem gives sufficient conditions under which this holds. It is called the *Delta method* because derivatives are involved. Recall that a function $f : \mathbb{R}^m \to \mathbb{R}^k$ is *(totally)* *differentiable* at a point $x_0 \in \mathbb{R}^k$ if there exists a matrix $A \in \mathbb{R}^{m \times k}$ and a function $h : \mathbb{R}^k \to \mathbb{R}^m$ such that

$$f(x) = f(x_0) + A \cdot (x - x_0) + h(x) \quad \forall\, x \in \mathbb{R}$$

and

$$\lim_{x \to x_0} \frac{1}{|x - x_0|} h(x) = 0.$$

The matrix A then coincides with the Jacobi matrix of f at x_0 and is denoted by $Df(x_0)$. The Delta-method now reads as follows:

Theorem 9.6 (Delta-method)**.**
Let $\underline{X}_n = (X_n^{(1)}, \ldots, X_n^{(k)})'$*,* $n \in \mathbb{N}$*, be a sequence of* $\mathbb{R}^k$*-valued random vectors and let* $(v_n)_{n\in\mathbb{N}}$ *be a sequence of strictly positive real numbers converging to 0 as* $n \to \infty$*. Furthermore, suppose that*

$$\underline{X}_n \sim AN(\underline{\theta}; v_n, \Sigma), \quad \textit{as} \quad n \to \infty,$$

for a vector $\underline{\theta} \in \mathbb{R}^k$ *and a covariance matrix* $\Sigma \in \mathbb{R}^{k \times k}$*.*
Let $f : \mathbb{R}^k \to \mathbb{R}^m$ *be a function, which is totally differentiable in* $\underline{\theta} \in \mathbb{R}^k$ *with Jacobi matrix* $D = \left(\frac{\partial f_i}{\partial x_j}(\underline{\theta})\right)_{i=1,\ldots,m;j=1,\ldots,k} \in \mathbb{R}^{m \times k}$*. Then*

$$f(\underline{X}_n) \sim AN(f(\underline{\theta}); v_n, D\Sigma D') \quad \textit{as} \quad n \to \infty.$$

Proof. Since f is totally differentiable in $\theta \in \mathbb{R}^k$ we can write

$$f(\underline{x}) = f(\underline{\theta}) + D \cdot (\underline{x} - \underline{\theta}) + R(\underline{x} - \underline{\theta}) \quad \forall\, \underline{x} \in \mathbb{R}^k,$$

where $R : \mathbb{R}^k \to \mathbb{R}^m$ is a function with $\lim_{z \to 0} \frac{R(z)}{||z||} = 0$. Inserting $\underline{X}_n$ for $\underline{x}$ gives

$$f(\underline{X}_n) = f(\underline{\theta}) + D \cdot (\underline{X}_n - \underline{\theta}) + R(\underline{X}_n - \underline{\theta})$$

and hence

$$\frac{1}{\sqrt{v_n}}(f(\underline{X}_n) - f(\underline{\theta})) = D \cdot \frac{1}{\sqrt{v_n}}(\underline{X}_n - \underline{\theta}) + \frac{1}{\sqrt{v_n}} R(\underline{X}_n - \underline{\theta}).$$

Since by assumption, $\frac{1}{\sqrt{v_n}}(\underline{X}_n - \underline{\theta}) \xrightarrow{d} \underline{Z}$ as $n \to \infty$ for some $N(0, \Sigma)$-distributed random variable $\underline{Z}$, and since the mapping $\mathbb{R}^k \ni \underline{x} \mapsto D \cdot x \in \mathbb{R}^m$ is continuous, we get

$$D \frac{1}{\sqrt{v_n}}(\underline{X}_n - \underline{\theta}) \xrightarrow{d} D\underline{Z} \overset{d}{=} N(0, D\Sigma D') \quad \text{as} \quad n \to \infty.$$

By Slutsky's lemma it is hence enough to show that

$$\frac{1}{\sqrt{v_n}} R(\underline{X}_n - \underline{\theta}) \xrightarrow{P} 0 \quad (n \to \infty).$$

We write $|R(z)| = \varphi(z)|z|$, where $\varphi : \mathbb{R}^k \to [0,\infty)$ is a function with $\varphi(0) = 0$ and $\lim_{z\to 0}\varphi(z) = 0$ (observe that necessarily $R(0) = 0$). Then

$$\frac{1}{\sqrt{v_n}} R(\underline{X}_n - \underline{\theta}) = \frac{1}{\sqrt{v_n}}(\underline{X}_n - \underline{\theta})\,\varphi(\underline{X}_n - \underline{\theta}).$$

But $\frac{1}{\sqrt{v_n}}(\underline{X}_n - \underline{\theta})$ converges in distribution to $\underline{Z}$ by assumption, and since the mapping $\mathbb{R}^k \mapsto \mathbb{R}$, $\underline{x} \mapsto |\underline{x}|$ is continuous, we see that $\left|\frac{1}{\sqrt{v_n}}(\underline{X}_n - \underline{\theta})\right|$ converges in distribution to $|\underline{Z}|$ as $n\to\infty$. Again by Slutsky's lemma it is hence enough to show that $\varphi(\underline{X}_n - \underline{\theta}) \xrightarrow{P} 0$ $(n \to \infty)$. Since $\lim_{z\to 0}\varphi(z) = 0$ it is sufficient to show that $\underline{X}_n - \underline{\theta} \xrightarrow{P} 0$ (by the subsequence criterion for stochastic convergence; namely a sequence (Y_n) of random vectors converges in probability to a random vector Y if and only if for every subsequence (Y_{n_k}) of (Y_n) there exists a further subsequence $(Y_{n_{k_l}})$ of (Y_{n_k}) that converges almost surely to Y; then use that φ is continuous at 0). To see that $\underline{X}_n - \underline{\theta}$ converges in probability to 0, let $\varepsilon > 0$. Then

$$\begin{aligned}\limsup_{n\to\infty} P(|\underline{X}_n - \underline{\theta}| \geq \varepsilon) &= \limsup_{n\to\infty} P(\frac{1}{\sqrt{v_n}}|\underline{X}_n - \underline{\theta}| \geq \frac{\varepsilon}{\sqrt{v_n}}) \\ &\leq \limsup_{n\to\infty} P(\frac{1}{\sqrt{v_n}}|\underline{X}_n - \underline{\theta}| \geq r)\ \forall r > 0\end{aligned}$$

since $\lim_{n\to\infty}\varepsilon/\sqrt{v_n} = \infty$. But $\frac{1}{\sqrt{v_n}}|\underline{X}_n - \underline{\theta}|$ converges in distribution to $\underline{Z}$, and the set $\{\underline{z} \in \mathbb{R}^k : |\underline{z}| \geq r\}$ is closed, so that by Portmanteau's theorem (cf. Remark 9.5),

$$\limsup_{n\to\infty} P(\frac{1}{\sqrt{v_n}}|\underline{X}_n - \underline{\theta}| \geq r) \leq P(|\underline{Z}| \geq r) \quad \forall\, r > 0.$$

Hence

$$\limsup_{n\to\infty} P(|\underline{X}_n - \underline{\theta}| \geq \varepsilon) \leq P(|\underline{Z}| \geq r) \quad \forall\, r > 0$$

and letting $r\to\infty$ it follows that $\limsup_{n\to\infty} P(|\underline{X}_n - \underline{\theta}| \geq \varepsilon) = 0$, so we have $\underline{X}_n \xrightarrow{P} \underline{\theta}$, finishing the proof. □

The actual goal of this chapter is to derive the asymptotics for $(\widehat{\rho}(1), \dots, \widehat{\rho}(h))$ as $n\to\infty$ under suitable conditions. For that we need some lemmas and a central limit theorem for $(\widehat{\gamma}(1), \dots, \widehat{\gamma}(h))$ and $(\gamma^*(1), \dots, \gamma^*(h))$, where $\gamma^*(j)$ is defined in the next lemma.

Lemma 9.7. *Let $X_t = \sum_{j\in\mathbb{Z}} \psi_j Z_{t-j}$, $t \in \mathbb{Z}$, where $(Z_t) \sim iid(0,\sigma^2)$ (real-valued with $\sigma^2 > 0$) with*

$$\eta := \frac{\mathbb{E}Z_t^4}{\sigma^4} < \infty,$$

and $(\psi_j)_{j\in\mathbb{Z}}$ be real with $\sum_{j\in\mathbb{Z}}|\psi_j| < \infty$. Let γ be the ACVF of $(X_t)_{t\in\mathbb{Z}}$ and define

$$\gamma^*(h) := \frac{1}{n}\sum_{t=1}^{n} X_t X_{t+h},\ h \in \mathbb{N}_0.$$

Then it holds for every $p, q \geq 0$:

$$\begin{aligned}
&\lim_{n\to\infty} n\mathrm{Cov}\left(\gamma^*(p), \gamma^*(q)\right)\\
=&(\eta-3)\gamma(p)\gamma(q) + \sum_{k=-\infty}^{\infty} (\gamma(k)\gamma(k-p+q) + \gamma(k+q)\gamma(k-p)).
\end{aligned}$$

Proof. First observe that

$$\mathbb{E}\gamma^*(h) = \mathbb{E}(X_t X_{t+h}) = \gamma(h).$$

Next, since the (Z_t) are i.i.d.$(0, \sigma^2)$, we have

$$\mathbb{E}(Z_s Z_t Z_u Z_v) = \begin{cases} \eta\sigma^4 & , s=t=u=v, \\ \sigma^4 & , s=t\neq u=v, \\ 0 & , s\neq t, s\neq u, s\neq v \end{cases}$$

and similar quantities for the other cases (e.g. we get 0 when $t \neq s$, $t \neq s$, $t \neq u$, and we get σ^4 when $s = u \neq t = v$). Plugging this into the form of X_t we obtain (a small explanation follows below)

$$\begin{aligned}
&\mathbb{E}(X_t X_{t+p} X_{t+h+p} X_{t+h+p+q})\\
=&\sum_i \sum_j \sum_k \sum_l \psi_i \psi_{j+p} \psi_{k+h+p} \psi_{l+h+p+q} \mathbb{E}(Z_{t-i} Z_{t-j} Z_{t-k} Z_{t-l})\\
=&(\eta-3)\sigma^4 \sum_i \psi_i \psi_{i+p} \psi_{i+h+p} \psi_{i+h+p+q} && (i=j=k=l)\\
&+ \underbrace{\sigma^2 \sum_i \psi_i \psi_{i+p}}_{\gamma(p)} \underbrace{\sigma^2 \sum_k \psi_{k+h+p} \psi_{k+h+p+q}}_{\gamma(q)} && (i=j, k=l)\\
&+ \underbrace{\sigma^2 \sum_i \psi_i \psi_{i+h+p}}_{\gamma(h+p)} \underbrace{\sigma^2 \sum_j \psi_{j+p} \psi_{j+h+p+q}}_{\gamma(h+q)} && (i=k, j=l)\\
&+ \underbrace{\sigma^2 \sum_i \psi_i \psi_{i+h+p+q}}_{\gamma(h+p+q)} \underbrace{\sigma^2 \sum_j \psi_{j+p} \psi_{j+h+p}}_{\gamma(h)} && (i=l, j=k).
\end{aligned}$$

Why did we write $\eta - 3$ in the first line when $i = j = k = l$ and not just η? The reason is that in the lines below, e.g. the second line when $i = j$ and $k = l$, we wrote down $\mathbb{E}(Z_{t-i}Z_{t-j})\,\mathbb{E}(Z_{t-k}Z_{t-l})$ not only when $i = j \neq k = l$, but also when $i = j = k = l$. But the latter should not have been counted, so it has to be subtracted once. The same is true for the next two lines, which is why we have to substract this three times, resulting in the '-3' in the first line, which is why we have there $\eta - 3$ and not just η.

Continuing with our calculation, we obtain

$$\begin{aligned}\mathbb{E}(\gamma^*(p)\gamma^*(q)) &= \frac{1}{n^2}\mathbb{E}\left(\sum_{s=1}^{n}\sum_{t=1}^{n} X_t X_{t+p} X_s X_{s+q}\right) \\ &= \frac{1}{n^2}\sum_{s=1}^{n}\sum_{t=1}^{n}\Big(\gamma(p)\gamma(q) + \gamma(s-t)\gamma(s-t-p+q) \\ &\qquad + \gamma(s-t+q)\gamma(s-t-p) \\ &\qquad + (\eta-3)\sigma^4\sum_i \psi_i\psi_{i+p}\psi_{i+s-t}\psi_{i+s-t+q}\Big).\end{aligned}$$

This gives

$$\begin{aligned}\operatorname{Cov}\left(\gamma^*(p), \gamma^*(q)\right) &= \mathbb{E}(\gamma^*(p)\gamma^*(q)) - \gamma(p)\gamma(q) \\ &= \frac{1}{n^2}\sum_{s=1}^{n}\sum_{t=1}^{n}\Big(\gamma(s-t)\gamma(s-t-p+q) + \gamma(s-t+q)\gamma(s-t-p) \\ &\qquad + (\eta-3)\sigma^4\sum_i \psi_i\psi_{i+p}\psi_{i+s-t}\psi_{i+s-t+q}\Big) \\ &\overset{k=s-t}{=} \frac{1}{n^2}\sum_{|k|<n}(n-|k|)T_k\,,\end{aligned}$$

where

$$T_k := \gamma(k)\gamma(k-p+q) + \gamma(k+q)\gamma(k-p) + (\eta-3)\sigma^4\sum_i \psi_i\psi_{i+p}\psi_{i+k}\psi_{i+k+q}.$$

Since $(\psi_j)_{j\in\mathbb{Z}}$ is absolutely summable (and hence also $(\gamma(k))_{k\in\mathbb{Z}}$) it follows that $(T_k)_{k\in\mathbb{Z}}$ is absolutely summable. Lebesgue's dominated convergence theorem applied to sums (i.e. the counting measure) then gives

$$\begin{aligned}\lim_{n\to\infty} n\operatorname{Cov}\left(\gamma^*(p)\gamma^*(q)\right) &= \lim_{n\to\infty}\sum_{|k|<n}\left(1-\frac{|k|}{n}\right)T_k \\ &= \sum_{k=-\infty}^{\infty} T_k \\ &= (\eta-3)\underbrace{\left(\sigma^2\sum_i \psi_i\psi_{i+p}\right)}_{\gamma(p)}\underbrace{\left(\sigma^2\sum_k \psi_{i+k}\psi_{i+k+q}\right)}_{\gamma(q)} \\ &\quad + \sum_{k=-\infty}^{\infty}\left(\gamma(k)\gamma(k-p+q) + \gamma(k+q)\gamma(k-p)\right).\end{aligned}$$

This is the claim. □

The next step is to prove a limit result for γ^* when the sum $\sum_{j\in\mathbb{Z}}\psi_j Z_{t-j}$ defining X_t is actually a finite sum, i.e. when all but finitely many of the ψ_j are zero.

Lemma 9.8. *Let $m \in \mathbb{N}$ and $X_t = \sum\limits_{j=-m}^{m} \psi_j Z_{t-j}$, $t \in \mathbb{Z}$, where $\psi_j \in \mathbb{R}$, $(Z_t)_{t\in\mathbb{Z}} \sim iid(0, \sigma^2)$ real-valued with $\sigma^2 > 0$ and $\mathbb{E}Z_t^4 < \infty$, define*

$$\eta := \frac{\mathbb{E}Z_t^4}{\sigma^4}$$

and let γ be the ACVF of (X_t). Let $\gamma^*(j)$ be given as above. Fix $h \in \mathbb{N}_0$. Then

$$\begin{pmatrix} \gamma^*(0) \\ \vdots \\ \gamma^*(h) \end{pmatrix} \sim AN\left(\begin{pmatrix} \gamma(0) \\ \vdots \\ \gamma(h) \end{pmatrix}; \frac{1}{n}, V \right) \qquad (n \to \infty), \tag{9.1}$$

where $V \in \mathbb{R}^{(h+1)\times(h+1)}$ is given by

$$V = ((\eta - 3)\gamma(p)\gamma(q) + \sum_{k\in\mathbb{Z}} (\gamma(k)\gamma(k-p+q) + \gamma(k+q)\gamma(k-p))_{p,q=0,\dots,h}. \tag{9.2}$$

The matrix V is symmetric and positive semi-definite, hence a covariance matrix. The element $V_{0,0}$ is strictly positive if

$$\begin{cases} Z_0^2 \text{ is not almost surely constant and } \psi_j \neq 0 \text{ for at least one } j \in \mathbb{Z}, \\ \text{or } \gamma(j) \neq 0 \text{ for at least one } j \neq 0. \end{cases} \tag{9.3}$$

Proof. Let us first show that $V \in \mathbb{R}^{(h+1)\times(h+1)}$ is a covariance matrix. Write $V = (V_{pq})_{p,q=0,\dots,h}$. Then

$$V_{pq} = \lim_{n\to\infty} n\mathrm{Cov}\,(\gamma^*(p), \gamma^*(q))$$

by Lemma 9.7, from which we see immediately that $V_{pq} = V_{qp}$ so that V is symmetric. Further, for each $\lambda = (\lambda_0, \dots, \lambda_h)' \in \mathbb{R}^{h+1}$ we have

$$\lambda' V \lambda = \Big(\lim_{n\to\infty} n\lambda' (\mathrm{Cov}\,(\gamma^*(p), \gamma^*(q))_{p,q=0,\dots,h} \lambda \Big) = \lim_{n\to\infty} n\mathrm{Var}\left(\sum_{p=0}^{h} \lambda_p \gamma^*(p) \right) \geq 0, \tag{9.4}$$

since a variance is always non-negative. This shows that V is positive semidefinite, hence a covariance matrix.

For the proof of (9.1), the idea is to define a $(2m+h)$-dependent sequence of random vectors, apply Theorem 8.11 to the linear combinations of the random vectors, and then apply the Cramér-Wold device. The details are as follows. For each $t \in \mathbb{Z}$, define the $\mathbb{R}^{h+1}$ valued random vector Y_t by

$$Y_t := (X_t X_t, X_t X_{t+1}, X_t X_{t+2}, \dots, X_t X_{t+h})'.$$

Then

$$\frac{1}{n} \sum_{t=1}^{n} Y_t = \frac{1}{n} \begin{pmatrix} \sum_{t=1}^{n} X_t X_t \\ \sum_{t=1}^{n} X_t X_{t+1} \\ \vdots \\ \sum_{t=1}^{n} X_t X_{t+h} \end{pmatrix} = \begin{pmatrix} \gamma^*(0) \\ \gamma^*(1) \\ \vdots \\ \gamma^*(h) \end{pmatrix}. \tag{9.5}$$

Equation (9.1) is hence equivalent to

$$\sqrt{n}\left(\frac{1}{n}\sum_{t=1}^{n} Y_t - \begin{pmatrix}\gamma(0)\\ \vdots \\ \gamma(h)\end{pmatrix}\right) \xrightarrow{d} N(0,V) \quad \text{as} \quad n\to\infty.$$

By the Cramér-Wold device (Theorem 9.1), or directly by Corollary 9.4, this is equivalent to

$$\sqrt{n}\left(\frac{1}{n}\sum_{t=1}^{n} \lambda'\left(Y_t - \begin{pmatrix}\gamma(0)\\ \vdots \\ \gamma(h)\end{pmatrix}\right)\right) \xrightarrow{d} N(0,\lambda' V\lambda) \quad (n\to\infty) \quad \forall\, \lambda\in\mathbb{R}^{h+1}. \tag{9.6}$$

Observe that for each $\lambda\in\mathbb{R}^{h+1}$, the sequence $(\lambda' Y_t)_{t\in\mathbb{Z}}$ is strictly stationary and $(2m+h)$-dependent; this is a simple consequence of the i.i.d. property of the noise and the definition of X_t. Further, $\mathbb{E}Y_t = (\gamma(0),\ldots,\gamma(h))'$ so that $(\lambda'(Y_t-(\gamma(0),\ldots,\gamma(h))')_{t\in\mathbb{Z}}$ is strictly stationary, $2m+h$-dependent and has expectation 0. Hence, if $\gamma_{\lambda' Y}$ denotes the autocovariance function of $(\lambda'(Y_t-(\gamma(0),\ldots,\gamma(h))'))_{t\in\mathbb{Z}}$, then

$$\sqrt{n}\left(\frac{1}{n}\sum_{t=1}^{n} \lambda'\left(Y_t - \begin{pmatrix}\gamma(0)\\ \vdots \\ \gamma(h)\end{pmatrix}\right)\right) \xrightarrow{d} N\left(0,u_{2m+h}\right) \quad (n\to\infty)$$

by Theorem 8.11 (b), where

$$u_{2m+h} = \lim_{n\to\infty} n\mathrm{Var}\,(n^{-1}\sum_{t=1}^{n}\lambda' Y_t)$$

by Theorem 8.11 (a). But

$$u_{2m+h} \overset{(9.5)}{=} \lim_{n\to\infty} n\mathrm{Var}\,(\lambda'(\gamma^*(0),\ldots,\gamma^*(h))') \overset{(9.4)}{=} \lambda' V\lambda,$$

so that (9.6) and hence (9.1) follow.

To see that $V_{00}>0$ if (9.3) is satisfied, observe that

$$V_{00} = (\eta-3)(\gamma(0))^2 + 2\sum_{k\in\mathbb{Z}}(\gamma(k))^2 = (\eta-1)(\gamma(0))^2 + 2\sum_{k\in\mathbb{Z}\setminus\{0\}}(\gamma(k))^2.$$

Since

$$\mathbb{E}Z_t^4 = \mathbb{E}(Z_t^2)^2 \begin{cases} \geq (\mathbb{E}Z_t^2)^2 = \sigma^4, & \text{always,}\\ > (\mathbb{E}Z_t^2)^2 = \sigma^4, & \text{if } Z_t^2 \text{ is not constant,}\end{cases}$$

by Jensen's inequality, we see that $\eta\geq 1$ and that $\eta>1$ unless Z_t^2 is constant. This shows $V_{00}>0$ (the assumption $\psi_j\neq 0$ for at least one j guarantees that $\gamma(0)>0$). □

Remark 9.9. That $V_{00} = 0$ when Z_t^2 is constant and $\psi_j = 0$ for $j \neq 0$ is no surprise. For then $Z_t^2 = \sigma^2$ almost surely, hence $X_t^2 = \psi_0^2\sigma^2$ almost surely, so that $\gamma^*(0) = \psi_0^2\sigma^2$ almost surely. An example when Z_t^2 is constant is given by $P(Z_t = \sigma) = P(Z_t = -\sigma) = 1/2$ (this is actually the only possibility to construct such an example when $\mathbb{E}(Z_t) = 0$).

Similar to the proof of Theorem 8.13, it is now possible to relax the condition of having a finite sum by applying Theorem 8.8:

Lemma 9.10. *Lemma* 9.8 *holds also for processes of the form*

$$X_t = \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j},\ t \in \mathbb{Z},\ (Z_t) \sim iid(0, \sigma^2) \text{ real valued}$$

if $\sum_{j\in\mathbb{Z}} |\psi_j| < \infty$, $\sigma^2 > 0$ *and* $\mathbb{E}Z_t^4 = \eta\sigma^4 < \infty$.

Proof. We omit the proof. It can be found in Proposition 7.3.3 in the book of Brockwell and Davis [BD1]. □

The next step is to replace $\gamma^*(j) = \frac{1}{n}\sum_{j=1}^{n} X_t X_{t+h}$ by $\widehat{\gamma}(j) = \frac{1}{n}\sum_{j=1}^{n-h}(X_t - \overline{X}_n)(X_{t+h} - \overline{X}_n)$. This is based on algebraic calculations:

Lemma 9.11. *Let* $X_t = \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}$, $t \in \mathbb{Z}$, *where* $(Z_t) \sim iid(0, \sigma^2)$ *(real-valued) with* (ψ_j) *real-valued,* $\sum_{j\in\mathbb{Z}} |\psi_j| < \infty$, $\mathbb{E}Z_t^4 = \eta\sigma^4 < \infty$, $\sigma^2 > 0$ *and* γ *be the ACVF of* $(X_t)_{t\in\mathbb{Z}}$. *For* $h \in \mathbb{N}_0$ *let* $V \in \mathbb{R}^{(h+1)\times(h+1)}$ *be defined as in* (9.2). *Then*

$$\begin{pmatrix} \hat{\gamma}(0) \\ \vdots \\ \hat{\gamma}(h) \end{pmatrix} \sim AN\left(\begin{pmatrix} \gamma(0) \\ \vdots \\ \gamma(h) \end{pmatrix}; \frac{1}{n}, V\right) \quad \text{as} \quad n \to \infty.$$

The element $V_{0,0}$ *is strictly positive if* (9.3) *is satisfied.*

Proof. We omit the proof. See Brockwell and Davis [BD1], Propositon 7.3.4. □

Remark 9.12. If $X_t = \mu + \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}$, $t \in \mathbb{Z}$, $(Z_t) \sim iid(0, \sigma^2)$ (real valued) with conditions like above and ACVF γ, then

$$\begin{pmatrix} \hat{\gamma}(0) \\ \vdots \\ \hat{\gamma}(h) \end{pmatrix} \sim AN\left(\begin{pmatrix} \gamma(0) \\ \vdots \\ \gamma(h) \end{pmatrix}; \frac{1}{n}, V\right) \quad (n \to \infty),$$

where V is the covariance matrix (9.2) and

$$\hat{\gamma}(p) = \frac{1}{n}\sum_{j=1}^{n-p}(X_j - \overline{X}_n)(X_{j+h} - \overline{X}_n).$$

This follows by applying Lemma 9.11 to $(X_t - \mu)_{t\in\mathbb{Z}}$.

From Lemma 9.11 and Remark 9.12 we get consistency of the estimator $\widehat{\gamma}(j)$:

Corollary 9.13. *Under the conditions of Lemma 9.11, let $X_t = \mu + \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}$. Then $\widehat{\gamma}(j)$ is a consistent estimator of $\gamma(j)$ for each $j \in \mathbb{N}_0$, i.e. $\widehat{\gamma}(j) \xrightarrow{P} \gamma(j)$ for each $j \in \mathbb{N}_0$ (observe that $\widehat{\gamma}(j)$ depends on the sample size n, although we suppressed it in the notation). If additionally $\gamma(0) > 0$ (i.e. there is $k \in \mathbb{Z}$ with $\psi_k \neq 0$), then $\widehat{\rho}(j)$ is a consistent estimator of $\rho(j)$ for each $j \in \mathbb{N}_0$, i.e.*

$$\widehat{\rho}(j) = \frac{\widehat{\gamma}(j)}{\widehat{\gamma}(0)} \xrightarrow{P} \frac{\gamma(j)}{\gamma(0)} = \rho(j), \quad (n \to \infty).$$

Proof. From Lemma 9.11 and Remark 9.12 we have $\sqrt{n}(\widehat{\gamma}(j) - \gamma(j)) \xrightarrow{d} U$ for some normally distributed random variable U. Multiplying this with $1/\sqrt{n}$ we get $\widehat{\gamma}(j) - \gamma(j) \xrightarrow{d} 0 \cdot U = 0$ by Slutsky's lemma, and since convergence in distribution to a constant implies convergence in probability, we get $\widehat{\gamma}(j) \xrightarrow{P} \gamma(j)$ for each $j \in \mathbb{N}_0$. The assertion regarding the empirical autocorrelation function is then clear. □

We now come to the main result of this chapter, namely a central limit theorem for the sample autocorrelation function. Observe that we do not need V_{00} to be strictly positive to apply the Delta-method, but we need $\gamma(0)$ to be different from zero (hence it is also strictly positive).

Theorem 9.14. *Let $(X_t)_{t\in\mathbb{Z}}$ be the stationary process*

$$X_t = \mu + \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}, \quad t \in \mathbb{Z},$$

where $(Z_t)_{t\in\mathbb{Z}}$ is a real-valued i.i.d. sequence with mean zero and variance $\sigma^2 \in (0, \infty)$ and $\mathbb{E}Z_t^4 < \infty$, and where $\mu \in \mathbb{R}$ and $\psi_j \in \mathbb{R}$ with $0 < \sum_{j=-\infty}^{\infty} |\psi_j| < \infty$. Let

$$\begin{aligned} \underline{\hat{\rho}}(h) &= (\hat{\rho}(1), \dots, \hat{\rho}(h))', \\ \underline{\rho}(h) &= (\rho(1), \dots, \rho(h))' \end{aligned}$$

and $W = (w_{ij})_{i,j=1,\dots,h} \in \mathbb{R}^{h\times h}$ with

$$\begin{aligned} w_{ij} = \sum_{k=-\infty}^{\infty} \Big(&\rho(k+i)\rho(k+j) + \rho(k-i)\rho(k+j) + 2\rho(i)\rho(j)\rho^2(k) \\ &- 2\rho(i)\rho(k)\rho(k+j) - 2\rho(j)\rho(k)\rho(k+i) \Big), \; i,j = 1,\dots,h. \end{aligned} \tag{9.7}$$

Then

$$\underline{\widehat{\rho}}(h) \sim AN(\underline{\rho}(h); \frac{1}{n}, W) \quad as \quad n \to \infty.$$

Proof. Since $\sum_{j\in\mathbb{Z}} |\psi_j| > 0$, i.e. since at least one of the ψ_j is different from 0, we have $\gamma(0) > 0$. Let $g : \mathbb{R}^{h+1} \to \mathbb{R}^h$ be defined by

$$g((x_0, x_1, \ldots, x_h)') := \left(\frac{x_1}{x_0}, \frac{x_2}{x_0}, \ldots, \frac{x_h}{x_0}\right)' \quad \text{when} \quad x_0 \neq 0$$

(how it is defined when $x_0 = 0$ is irrelevant). Let V be the covariance matrix (9.2) and $D = \left(\frac{\partial g_i}{\partial x_j}((\gamma(0), \ldots, \gamma(h))')\right)_{i=1,\ldots,h,\, j=1,\ldots,h+1}$. Then a simple calculation shows that

$$D = \frac{1}{\gamma(0)} \begin{pmatrix} -\rho(1) & 1 & 0 & \cdots & 0 \\ -\rho(2) & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & & \ddots & \\ -\rho(h) & 0 & & & 1 \end{pmatrix}.$$

The Delta-method (Theorem 9.6) together with Lemma 9.11 and Remark 9.12 then imply that $\underline{\widehat{\rho}}(h) \sim AN(\underline{\rho}(h), n^{-1}DVD')$ as $n \to \infty$.
Write $V = (v_{ij})_{i,j=0,\ldots,h}$ and $DVD^T = (w_{ij})_{i,j=1,\ldots,h}$. Then

$$\begin{aligned} w_{ij} &= \frac{1}{(\gamma(0))^2}\left[v_{ij} - \rho(i)v_{0j} - \rho(j)v_{i0} + \rho(i)\rho(j)v_{00}\right] \\ &\overset{(9.2)}{=} \sum_{k=-\infty}^{\infty} \Big[\rho(k)\rho(k-i+j) + \rho(k-i)\rho(k+j) + 2\rho(i)\rho(j)\rho^2(k) \\ &\qquad - 2\rho(i)\rho(k)\rho(k+j) - 2\rho(k)\rho(j)\rho(k-i)\Big]. \end{aligned}$$

As $\sum_k \rho(k)\rho(k-i+j) = \sum_k \rho(k+i)\rho(k+j)$ and $\sum_k \rho(j)\rho(k)\rho(k-i) = \sum_k \rho(j)\rho(k+i)\rho(k)$, Equation (9.7) follows. □

Remark 9.15. Equation (9.7) is called *Bartlett's Formula.* By simple calculations we conclude that

$$w_{ij} = \sum_{k=1}^{\infty} \left(\rho(k+i) + \rho(k-i) - 2\rho(i)\rho(k)\right)\left(\rho(k+j) + \rho(k-j) - 2\rho(j)\rho(k)\right). \quad (9.8)$$

The formula (9.8) is a little bit simpler to calculate.

Remark 9.16. It is surprising that the quantity η no longer appears in Barlett's formula, although it appeared in the limit variance of the sample autocovariance. So coming from the sample autocovariance to the sample autocorrelation, the quantity η dropped out. Hence it is natural to wonder if the assumption that $\mathbb{E}Z_t^4 < \infty$ is essential for Theorem 9.14 to hold. The answer is that if one assumes not only that $\sum_{j\in\mathbb{Z}} |\psi_j| < \infty$, but additionally also that $\sum_{j\in\mathbb{Z}} \psi_j^2|j| < \infty$, then the requirement that $\mathbb{E}Z_t^4 < \infty$ in Theorem 9.14 can be dropped. See Brockwell and Davis [BD1], Theorem 7.2.2. there.

Example 9.17. (a) Let $X_t = Z_t + \theta_1 Z_{t-1} + \ldots + \theta_q Z_{t-q}$, $Z_t \sim iid(0, \sigma^2)$ (and real-valued) with $\sigma^2 > 0$, $\mathbb{E}Z_t^4 < \infty$, and $\theta_1, \ldots, \theta_q \in \mathbb{R}$. Then (9.8) implies

$$w_{ii} = (1 + 2\rho^2(1) + 2\rho^2(2) + \ldots + 2\rho^2(q)) \quad \forall\, i > q,$$

so we have $n^{1/2}(\widehat{\rho}(i) - \underbrace{\rho(i)}_{=0}) \xrightarrow{d} N(0, w_{ii})$ for $n \to \infty$ $(i > q)$.

(b) When $X_t = Z_t \sim i.i.d.(0, \sigma^2)$ (real valued) with $\mathbb{E}Z_t^4 < \infty$ and $\sigma^2 > 0$, then $w_{ii} = 1$ for $i \geq 1$ by (9.8), so that $\widehat{\rho}(i)$ is approximately $N(0, 1/n)$-distributed.

Remark 9.18. (a) The results of asymptotical normality gives us the possibility to construct confidence intervals. We know that $\sqrt{n}(\widehat{\rho}(h) - \rho(h))$ is approximately $N(0, w_{hh})$-distributed. Hence, if $\Phi_{1-\alpha/2}$ denotes the $(1 - \alpha/2)$-quantile of the standard normal distribution, then the interval $[\widehat{\rho}(h) - \sqrt{w_{hh}/n}\Phi_{1-\alpha/2}, \widehat{\rho}(h) + \sqrt{w_{hh}/n}\Phi_{1-\alpha/2}]$ contains the true autocorrelation $\rho(h)$ with approximate probability $1 - \alpha$.
(b) Statistics programs often plot the lines $\pm 1.96 n^{-1/2}$ when plotting sample autocorrelation functions. This is due to the fact that the 0.975-quantile of the standard normal distribution is 1.96. If data come from i.i.d. noise, then $\rho(i) = 0$ for $i \geq 1$ and $[-1.96/\sqrt{n}, 1.96/\sqrt{n}]$ is a 95 % approximate confidence interval for $\rho(i)$. If one has a plot of the empirical autocorrelation function and not more than 1 out of 20 of the sample autocorrelations for lags greater or equal than 1 lie outside these bounds, then this is a good indication that the data come from white noise. Sometimes, statistics programs also plot other confidence bounds, which makes sense since Barlett's formula indicates that the asymptotic variance depends on the model. One should always check which bounds are plotted there.

Chapter 10

Yule-Walker estimator for causal AR(p) processes

Now we come to the important issue of estimating the parameters of ARMA(p,q)-processes. In this chapter, we treat a very simply estimator for causal AR(p)-processes, the so called Yule–Walker estimator. To some extent this is a moment estimator. It is possible to express the coefficients of a causal AR(p)-process through the autocovariance function, which itself can be estimated with the sample autocovariance function, and then one can plug these estimators in. The details are given in this chapter.

Throughout this chapter, let $(Z_t)_{t\in\mathbb{Z}} \sim WN(0,\sigma^2)$ (real-valued) with $\sigma^2 > 0$ and $(X_t)_{t\in\mathbb{Z}}$ be a real and causal AR(p)-process with representation

$$X_t - \varphi_1 X_{t-1} - \ldots - \varphi_p X_{t-p} = Z_t, \quad (Z_t) \sim WN(0,\sigma^2). \tag{10.1}$$

Wanted: An estimator for $\vec{\varphi} = (\varphi_1, \ldots, \varphi_p)'$ and for σ^2.

The so called 'Yule-Walker equations' are derived as follows: since (X_t) is causal, there are complex coefficients $(\psi_k)_{k\in\mathbb{N}_0}$ with $\sum_{k=0}^\infty |\psi_k| < \infty$ and

$$X_t = \sum_{k=0}^{\infty} \psi_k Z_{t-k}.$$

Since the coefficients $\varphi_1, \ldots, \varphi_p$ are real valued, so are the coefficients ψ_k. This follows from Equations (6.17) and (6.18) (by solving them iteratively for $j = 0$, $j = 1$, $j = 2, \ldots$) Multiplying (10.1) by X_{t-j} and inserting the above expression for X_{t-j} on the right-hand side gives

$$X_t X_{t-j} - \varphi_1 X_{t-1} X_{t-j} - \ldots - \varphi_p X_{t-p} X_{t-j} = Z_t \sum_{k=0}^{\infty} \psi_k Z_{t-j-k}, \quad j = 0, \ldots, p, \quad t \in \mathbb{Z}.$$

Taking the expectation in the above equation gives for the ACVF γ of X (observe that $\psi_0 = 1$ by (6.17))

$$\gamma(j) - \varphi_1 \gamma(j-1) - \ldots - \varphi_p \gamma(j-p) = \begin{cases} \psi_0 \sigma^2 = \sigma^2, & j = 0, \\ 0, & j = 1, \ldots, p. \end{cases}$$

Actually, we had these equations derived before, namely in Equations (6.20) and (6.21). Denoting

$$\Gamma_p := (\gamma(i-j))_{i,j=1,\dots,p} \in \mathbb{R}^{p\times p} \quad \text{and} \quad \vec{\gamma}_p := (\gamma(1),\dots,\gamma(p))' \in \mathbb{R}^p$$

the above equations read

$$\Gamma_p \cdot \vec{\varphi} = \vec{\gamma}_p$$

(for $j = 1,\dots,p$) and for $j = 0$

$$\sigma^2 = \gamma(0) - \vec{\varphi}' \cdot \vec{\gamma}_p.$$

Definition 10.1. For the causal AR(p)-process given by (10.1) with $\Gamma_p = (\gamma(i-j))_{i,j=1,\dots,p}$ and $\vec{\gamma}_p := (\gamma(1),\dots,\gamma(p))'$ the equations

$$\Gamma_p \cdot \vec{\varphi} = \vec{\gamma}_p \tag{10.2}$$

and

$$\sigma^2 = \gamma(0) - \vec{\varphi}' \cdot \vec{\gamma}_p \tag{10.3}$$

are called *Yule–Walker equations.*

Remark 10.2. The Yule-Walker equation (10.2) is nothing else than the prediction equation (7.5) for the one-step predictor of X_{p+1} given $X_1,\dots,X_p$ (i.e. we have $h = 1$). However, one should observe that à priori there are different meanings associated to the φ_i. In the Yule-Walker equations, these are the coefficients of the AR(p)-process, while in (7.5) they are the coefficients of the predictor. A posteriori however, it turns out that these coefficients are the same, which is due to Proposition 7.20. So we can say that the Yule-Walker equations (at least (10.2)) are nothing else than the prediction equations (for the one-step predictor of X_{p+1}), and that they are also nothing else than the recurrence equations (6.20) and (6.21) for the autocovariance function (at least for lags $k = 0,\dots,p$). The significance of having the equations again is that now our viewpoint is different. Before we were assuming that we knew the coefficients and were interested in calculating the autocovariances and predictors. Now we think that we have the autocovariances, or at least estimators for them, and want to obtain knowledge regarding the coefficients from that.

Now assume that we have a realisation $X_1,\dots,X_n$ of a causal AR(p)-process. Then we can write down the empirical autocovariances. So substituting Γ_p by the empirical autocovariance matrix

$$\widehat{\Gamma}_p := (\widehat{\gamma}(i-j))_{i,j=1,\dots,p}$$

and $\vec{\gamma}_p$ by its empirical counterpart

$$\widehat{\vec{\gamma}}_p := (\widehat{\gamma}(1),\dots,\widehat{\gamma}(p))',$$

where

$$\widehat{\gamma}(-h) := \widehat{\gamma}(h) = \frac{1}{n}\sum_{i=1}^{n-h}(X_i - \overline{X}_n)(X_{i+h} - \overline{X}_n), \quad h = 0,\dots,n-1,$$

the corresponding equations

$$\widehat{\Gamma}_p \cdot \widehat{\vec{\varphi}} = \widehat{\vec{\gamma}}_p \tag{10.4}$$

and

$$\widehat{\sigma^2} = \widehat{\gamma}(0) - \widehat{\vec{\varphi}}' \cdot \widehat{\vec{\gamma}}_p \tag{10.5}$$

(if existent) and their solution are called *Yule-Walker estimators* $\widehat{\vec{\varphi}}$ and $\widehat{\sigma^2}$ of $\vec{\varphi}$ and σ^2.

Let us give the Yule–Walker estimator for the particular case of an AR(1) process:

Example 10.3. *For a causal AR(1) process, the Yule–Walker estimators are given by*

$$\widehat{\varphi}_1 = \widehat{\rho}(1) \quad \textit{and} \quad \widehat{\sigma}^2 = \widehat{\gamma}(0) - \widehat{\rho}(1)\widehat{\gamma}(1).$$

Proof. This is immediate from Equations (10.4) and (10.5) which then read

$$\widehat{\gamma}(0)\widehat{\varphi}_1 = \widehat{\gamma}(1) \quad \text{and} \quad \widehat{\sigma}^2 = \widehat{\gamma}(0) - \widehat{\varphi}_1\widehat{\gamma}(1).$$

□

When considering higher orders, an immediate question now is the following:

Question: When does Equation (10.4) have a unique solution?

The answer is 'always', well, always unless all observations are the same:

Lemma 10.4. *Suppose that* $\widehat{\gamma}(0) > 0$ *(i.e. not all observations are the same). Then* $\widehat{\Gamma}_p$ *is strictly positive definitive and hence invertible for all* $p \in \mathbb{N}$.

Proof. Given observations $X_1, \dots, X_n$, the empirical autocovariance matrix $\widehat{\Gamma}_p$ is positive semidefinite for all $p \in \mathbb{N}$ by Proposition 4.9. Using Remark 4.3, the function $\widehat{\gamma} : \mathbb{Z} \to \mathbb{R}$ is even and positive semidefinite (setting $\widehat{\gamma}(h) := 0$ for $|h| \geq n$).

Using Theorem 4.4, it follows that there exists a real-valued weakly stationary process $(Y_t)_{t\in\mathbb{Z}}$ whose theoretical autocovariance function γ_Y coincides with the empirical autocovariance function $\widehat{\gamma}$, i.e. such that

$$\gamma_Y(h) = \widehat{\gamma}(h) \quad \forall\, h \in \mathbb{Z}.$$

Since $\widehat{\gamma}(h) = 0$ for $h \geq n$, we have $\lim_{h\to\infty} \gamma_Y(h) = 0$. Since also $\gamma_Y(0) = \widehat{\gamma}(0) > 0$, we obtain from Theorem 7.28 that $\widehat{\Gamma}_p = \Gamma_{Y,p}$ is invertible for every $p \in \mathbb{N}$, hence also strictly positive definite. □

This is good news, unless all observations are equal, we can always write down the Yule-Walker estimator. With the estimator $\widehat{\vec{\varphi}} = (\widehat{\varphi}_1, \dots, \widehat{\varphi}_p)'$ and $\widehat{\sigma^2}$. We can then write down an estimated AR(p)-model

$$Y_t - \widehat{\varphi}_1 Y_{t-1} - \ldots - \widehat{\varphi}_p Y_{t-p} = W_t, \quad t \in \mathbb{Z},$$

where $(W_t)_{t\in\mathbb{Z}} \sim WN(0, \widehat{\sigma^2})$. Although $\sigma^2 \geq 0$, it is not immediately clear that also the estimator $\widehat{\sigma^2}$ is strictly positive. Similarly, the model (10.1) is assumed to be causal, but will the estimated model also be causal? And how does the theoretical autocovariance function of the estimated model compare to the observed empirical autocovariance? The answer is provided in the following theorem:

Theorem 10.5. *Let $\widehat{\gamma}(0) > 0$ and denote by $\widehat{\varphi}$ and $\widehat{\sigma^2}$ the Yule-Walker estimators given by* (10.4) *and* (10.5)*. Then*

$$\widehat{\sigma^2} > 0$$

and

$$1 - \widehat{\varphi}_1 z - \ldots - \widehat{\varphi}_p z^p \neq 0 \quad \forall\, z \in \mathbb{C} : |z| \leq 1. \tag{10.6}$$

The fitted AR(p) model

$$Y_t - \widehat{\varphi}_1 Y_{t-1} - \ldots - \widehat{\varphi}_p Y_{t-p} = W_t, \quad t \in \mathbb{Z}, \tag{10.7}$$

where $(W_t)_{t\in\mathbb{Z}} \sim WN(0, \widehat{\sigma^2})$, is causal, and its autocovariance function γ_Y satisfies

$$\gamma_Y(h) = \widehat{\gamma}(h) \quad \forall\, h \in \{-p, \ldots, p\}.$$

The proof of Theorem 10.5 needs a lemma from linear algebra, which we state now but we will only indicate how to prove it: It reads as follows:

Lemma 10.6. *Let $\psi_1, \ldots, \psi_p \in \mathbb{R}$ such that $1 - \psi_1 z - \ldots - \psi_p z^p \neq 0$ for all $z \in \mathbb{C}$ with $|z| \leq 1$. Let $b \in \mathbb{R}$. Then the linear system of $2p+1$ equations*

$$\begin{aligned} a_h - \psi_1 a_{h-1} - \ldots - \psi_p a_{h-p} &= \begin{cases} b, & h = 0, \\ 0, & h = 1, \ldots, p, \end{cases} \\ a_h &= a_{-h}, \quad h = 1, \ldots, p \end{aligned}$$

has exactly one solution $(a_{-p}, \ldots, a_p)' \in \mathbb{R}^{2p+1}$.

Sketch of proof. This is not so easy to prove as it looks like, and we refer to Hamilton, Time Series Analysis, p. 59 and Exercise 10.1 there. It uses Kronecker products and one can show that $(a_0, a_1, \ldots, a_{p-1})'$ is given by the first p elements of the first column of the $(p^2 \times p^2)$-matrix $b[I_{p^2} - (F \otimes F)]^{-1}$, where $A \otimes B$ denotes the Kronecker product of two matrices (cf. the book of Hamilton for the definition of that product), where I_{p^2} is the $(p^2 \times p^2)$-identity matrix and F is the companion matrix

$$F := \begin{pmatrix} \psi_1 & \psi_2 & \psi_3 & \ldots & \psi_{p-1} & \psi_p \\ 1 & 0 & 0 & \ldots & 0 & 0 \\ 0 & 1 & 0 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ldots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 1 & 0 \end{pmatrix}.$$

The causality assumption implies that the above matrix $I_{p^2} - F \otimes F$ is indeed invertible. □

Let us now prove Theorem 10.5:

Proof of Theorem 10.5. (i) Let us first prove (10.6): From the proof of Lemma 10.4 we see that $\widehat{\gamma} : \mathbb{Z} \to \mathbb{R}$ is positive semidefinite (and even), hence the autocovariance function of some real valued stationary time series $(U_t)_{t\in\mathbb{Z}}$ with mean zero and $\gamma_U = \widehat{\gamma}$ (cf. Theorem 4.4).

Claim 1: $\widehat{\varphi}_1 U_p + \ldots + \widehat{\varphi}_p U_1$ is the orthogonal projection of U_{p+1} onto $\operatorname{span}\{U_1, \ldots, U_p\}$. This follows from the fact that $\widehat{\varphi}_1 U_p + \ldots + \widehat{\varphi}_p U_1 \in \operatorname{span}\{U_1, \ldots, U_p\}$ and that

$$\langle U_{p+1} - \sum_{k=1}^{p} \widehat{\varphi}_k U_{p+1-k}, U_j \rangle = \gamma_U(p+1-j) - \sum_{k=1}^{p} \widehat{\varphi}_k \gamma_U(p+1-k-j) \overset{(10.4)}{=} 0 \quad \forall\, j = 1, \ldots, p.$$

Now denote

$$\widehat{\Phi}(z) := 1 - \widehat{\varphi}_1 z - \ldots - \widehat{\varphi}_p z^p, \quad z \in \mathbb{C}.$$

Let $a \in \mathbb{C}$ such that $1/a$ is a zero of $\widehat{\Phi}$ (obviously, 0 cannot be a zero of this polynomial). Then we can write

$$\widehat{\Phi}(z) = (1 - az)\xi(z), \quad \text{where} \quad \xi(z) = 1 - \sum_{k=1}^{p-1} \xi_k z^k$$

for some $\xi_1, \ldots, \xi_{p-1} \in \mathbb{C}$. Denote

$$V_t := \xi(B) U_t = U_t - \xi_1 U_{t-1} - \ldots - \xi_{p-1} U_{t-(p-1)}, \quad t \in \mathbb{Z},$$

which gives a weakly stationary process. We have $\|V_t\|^2 = \mathbb{E}(V_t^2) > 0$: for if not, then $V_t = 0$, hence $\widehat{\Phi}(B) U_t = (1 - aB)\xi(B) U_t = (1 - aB) V_t = 0$ implying $U_p = \widehat{\varphi}_1 U_{p-1} + \ldots + \widehat{\varphi}_p U_0$ so that $\Gamma_{U,p+1}$ is not invertible. But since $\Gamma_{U,p+1} = \widehat{\Gamma}_{p+1}$ is invertible by Lemma 10.4, we see that $\|V_t\| > 0$. Denote

$$\rho := \operatorname{Corr}(V_{p+1}, V_p) = \frac{\langle V_{p+1}, V_p \rangle}{\|V_p\|^2}.$$

Claim 2: $\rho = a$.
To see that, denote

$$\widetilde{\Phi}(z) := (1 - \rho z)\xi(z) = 1 - \sum_{k=1}^{p} b_k z^k, \quad z \in \mathbb{C}.$$

Then

$$\mathbb{E} \left| U_{p+1} - \sum_{k=1}^{p} b_k U_{p+1-k} \right|^2 = \mathbb{E} \left| \widetilde{\Phi}(B) U_{p+1} \right|^2 = \mathbb{E} \left| (1 - \rho B) V_{p+1} \right|^2 = \mathbb{E} |V_{p+1} - \rho V_p|^2. \quad (10.8)$$

Similarly we obtain for $\widehat{\Phi}(z)$:

$$\mathbb{E} \left| U_{p+1} - \sum_{k=1}^{p} \widehat{\varphi}_k U_{p+1-k} \right|^2 = \mathbb{E} |V_{p+1} - a V_p|^2. \quad (10.9)$$

From the minimisation property of the projection and Claim 1, we get

$$\mathbb{E} \left| U_{p+1} - \sum_{k=1}^{p} \widehat{\varphi}_k U_{p+1-k} \right|^2 \leq \mathbb{E} \left| U_{p+1} - \sum_{k=1}^{p} b_k U_{p+1-k} \right|^2,$$

which using (10.8) and (10.9) can be rewritten as

$$\|V_{p+1} - aV_p\|^2 \leq \|V_{p+1} - \rho V_p\|^2.$$

On the other hand, the projection of V_{p+1} onto $\operatorname{span}\{V_p\}$ is given by

$$P_{\operatorname{span}\{V_p\}} V_{p+1} = \frac{\langle V_{p+1}, V_p\rangle}{\|V_p\|^2} V_p = \rho V_p,$$

so that by the minimisation property of the projection we obtain $aV_p = \rho V_p$, i.e. $\rho = a$. This finishes the proof of Claim 2.

Since $|\rho| \leq 1$ as a correlation, we obtain also $|a| \leq 1$ from Claim 2. Supposing for the moment that $|\rho| = |a| = 1$, we have

$$\|V_{p+1} - \rho V_p\|^2 = \|V_{p+1}\|^2 + |\rho|^2\|V_p\|^2 - 2\Re(\rho\langle V_p, V_{p+1}\rangle) = 0,$$

hence $V_{p+1} = \rho V_p$. But this implies

$$\widetilde{\Phi}(B)U_{p+1} = (1-\rho B)\xi(B)U_{p+1} = (1-\rho B)V_{p+1} = (1-\rho B)\rho V_p = (1-\rho B)\rho\xi(B)U_p,$$

hence there exist $c_0, \ldots, c_p \in \mathbb{C}$ such that $U_{p+1} = \sum_{k=0}^{p} c_k U_k$, which again leads to $\widehat{\Gamma}_{p+2} = \Gamma_{U,p+2}$ being singular, contradicting Lemma 10.4. Hence we have $|a| = |\rho| < 1$ so that all zeroes of $\widehat{\Phi}$ lie outside the unit circle, showing (10.6).

(ii) Let us show that $\widehat{\sigma^2} > 0$ and that the autocovariances of the fitted model conincide with the sample autocovariances for $h \in \{-p, \ldots, p\}$. To this end, for the moment let $w^2 > 0$ be given and $(W_t)_{t\in\mathbb{Z}} \sim WN(0, w^2)$. With this white noise, let $(Y_t)_{t\in\mathbb{Z}}$ be a process that satisfies (10.7). Then $(Y_t)_{t\in\mathbb{Z}}$ is a causal AR(p)-process by (10.6). From (10.2) and (10.3) we hence obtain

$$\gamma_Y(h) - \widehat{\varphi}_1\gamma_Y(h-1) - \ldots - \widehat{\varphi}_p\gamma_Y(h-p) = \begin{cases} w^2, & h = 0, \\ 0, & h = 1, \ldots, p. \end{cases}$$

On the other hand, from (10.4) and (10.5) we have

$$\widehat{\gamma}(h) - \widehat{\varphi}_1\widehat{\gamma}(h-1) - \ldots - \widehat{\varphi}_p\widehat{\gamma}(h-p) = \begin{cases} \widehat{\sigma^2}, & h = 0, \\ 0, & h = 1, \ldots, p. \end{cases}$$

But the linear system

$$a_h - \widehat{\varphi}_1 a_{h-1} - \ldots - \widehat{\varphi}_p a_{h-p} = \begin{cases} b, & h = 0, \\ 0, & h = 1, \ldots, p \end{cases}$$

with $a_j = a_{-j}$ has for given $b \in \mathbb{R}$ exactly one solution $(a_0, \ldots, a_p)' \in \mathbb{R}^{p+1}$ by Lemma 10.6. The uniqueness (and form) of the solution implies

$$\widehat{\gamma}(h) = \frac{\widehat{\sigma^2}}{w^2}\gamma_Y(h) \quad \forall\, h = 0, \ldots, p.$$

Since $\gamma_Y(0) > 0$ and $\widehat{\gamma}(0) > 0$ we obtain $\widehat{\sigma^2} > 0$. Choosing $w^2 = \widehat{\sigma^2}$ gives $\widehat{\gamma}(h) = \gamma_Y(h)$ for all $h = 0, \ldots, p$. □

We now come to the asymptotic normality of the Yule-Walker estimator:

Theorem 10.7. *Let $(X_t)_{t\in\mathbb{Z}}$ be the causal AR(p)-process given by* (10.1)*, where $(Z_t)_{t\in\mathbb{Z}}$ is i.i.d. $(0,\sigma^2)$-noise with finite fourth moment (real valued and $\sigma^2 > 0$). Then there exists a matrix $R \in \mathbb{R}^{(p+1)\times(p+1)}$ such that the Yule-Walker estimator $(\widehat{\vec{\varphi}}, \widehat{\sigma^2})$ satisfies*

$$n^{1/2}\left(\begin{pmatrix} \widehat{\vec{\varphi}} \\ \widehat{\sigma^2} \end{pmatrix} - \begin{pmatrix} \vec{\varphi} \\ \sigma^2 \end{pmatrix}\right) \xrightarrow{d} N(0,R) \quad \text{as} \quad n\to\infty. \tag{10.10}$$

The asymptotic covariance matrix of $\widehat{\vec{\varphi}}$ only is $\sigma^2\Gamma_p^{-1}$, i.e.

$$n^{1/2}\left(\widehat{\vec{\varphi}} - \vec{\varphi}\right) \xrightarrow{d} N(0,\sigma^2\Gamma_p^{-1}) \quad \text{as} \quad n\to\infty. \tag{10.11}$$

In particular,

$$\begin{aligned} \widehat{\vec{\varphi}} &\xrightarrow{P} \vec{\varphi}, \quad n\to\infty, \\ \widehat{\sigma^2} &\xrightarrow{P} \sigma^2, \quad n\to\infty, \end{aligned}$$

i.e. the Yule-Walker estimators are consistent.

Proof. Define $\Psi : \mathbb{R}^{p+1} \to \mathbb{R}^{p\times p}\times\mathbb{R}^p\times\mathbb{R}$ by

$$\begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_p \end{pmatrix} \mapsto \left(\begin{pmatrix} x_0 & x_1 & x_2 & \dots & x_{p-1} \\ x_1 & x_0 & x_1 & \ddots & \\ x_2 & x_1 & x_0 & \ddots & \\ & & \ddots & \ddots & \\ x_{p-1} & x_{p-2} & x_{p-3} & \dots & x_0 \end{pmatrix}, \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix}, x_0 \right).$$

Then Ψ is totally differentiable. Denote

$$G := \{A \in \mathbb{R}^{p\times p} : \det A \neq 0\} \times \mathbb{R}^p \times \mathbb{R}$$

and consider

$$f : G \to \mathbb{R}^p \times \mathbb{R}, \quad (A, z, x_0) \mapsto \left(A^{-1}z,\ x_0 - z'(A^{-1})'z\right).$$

Since by Cramér's rule,

$$A^{-1} = \frac{1}{\det A}\,\mathrm{ad}(A),$$

where

$$\mathrm{ad}(A) = \left((-1)^{i+j}\det A_{ij}\right)'_{i,j=1,\dots,p}$$

and $A_{ij} \in \mathbb{R}^{(p-1)\times(p-1)}$ is obtained from A by deleting the i'th row and j'th column, we see that f is totally differentiable on G. From that we conclude that

$$f\circ\Psi : \mathbb{R}^{p+1} \supset \Psi^{-1}(G) \to \mathbb{R}^p\times\mathbb{R}$$

is totally differentiable on $\Psi^{-1}(G)$. Since

$$\begin{pmatrix} \widehat{\gamma}(0) \\ \widehat{\gamma}(1) \\ \vdots \\ \widehat{\gamma}(p) \end{pmatrix} \in \Psi^{-1}(G)$$

by Lemma 10.4, and since

$$\begin{pmatrix} \widehat{\gamma}(0) \\ \widehat{\gamma}(1) \\ \vdots \\ \widehat{\gamma}(p) \end{pmatrix} \sim AN\left(\begin{pmatrix} \gamma(0) \\ \gamma(1) \\ \vdots \\ \gamma(p) \end{pmatrix} ; n^{-1}, V \right) \quad (n \to \infty)$$

by Lemma 9.11, the Delta-method (Theorem 9.6) gives

$$\begin{pmatrix} \widehat{\vec{\varphi}} \\ \widehat{\sigma^2} \end{pmatrix} = f \circ \Psi \left(\begin{pmatrix} \widehat{\gamma}(0) \\ \widehat{\gamma}(1) \\ \vdots \\ \widehat{\gamma}(p) \end{pmatrix} \right) \sim AN\left(\begin{pmatrix} \vec{\varphi} \\ \sigma^2 \end{pmatrix} ; n^{-1}, D(f \circ \Psi)_{|(\vec{\varphi}',\sigma^2)'} V \left(D(f \circ \Psi)_{|(\vec{\varphi}',\sigma^2)'} \right)' \right).$$

With

$$R := D(f \circ \Psi)_{|(\vec{\varphi}',\sigma^2)'} V \left(D(f \circ \Psi)_{|(\vec{\varphi}',\sigma^2)'} \right)'$$

we get (10.10) and hence the consistency (as in the proof of Corollary 9.13). It also follows that there is $R_1 \in \mathbb{R}^{p \times p}$ such that

$$n^{1/2} \left(\widehat{\vec{\varphi}} - \vec{\varphi} \right) \xrightarrow{d} N(0, R_1), \quad n \to \infty.$$

The calculation that $R_1 = \sigma^2 \Gamma_p^{-1}$ is messy. To get it, one can either use Theorem 8.1.1 in Brockwell and Davis who derive it by comparison with the least-squares estimator, or one does hard algebra and gets the form of R from that. □

With the aid of Theorem 10.7 we can derive approximate confidence intervals for the coefficients:

Corollary 10.8. *For $\alpha \in (0,1)$ denote the $(1-\alpha/2)$-quantile of the standard normal distribution by $\Phi_{1-\alpha/2}$, i.e. the unique number $\Phi_{1-\alpha/2} > 0$ such that*

$$\int_{-\infty}^{\Phi_{1-\alpha/2}} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \, dx = 1 - \alpha/2,$$

equivalently that

$$\int_{\Phi_{1-\alpha/2}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \, dx = \alpha/2.$$

Consider the Yule-Walker estimator of a causal AR(p)-process based on observations $X_1, \dots, X_n$. Denote by $\widehat{v}_{jj}$ the j'th diagonal element of $\widehat{\Gamma}_p^{-1}$. Then the approximate confidence interval

$$\left[\widehat{\varphi}_j - \Phi_{1-\alpha/2}\sqrt{\widehat{\sigma}^2\widehat{v}_{jj}/n}, \widehat{\varphi}_j + \Phi_{1-\alpha/2}\sqrt{\widehat{\sigma}^2\widehat{v}_{jj}/n}\right]$$

contains the true parameter φ_j with approximate probability $1-\alpha$. In particular, the random interval

$$\left[\widehat{\varphi}_j - 1.96\sqrt{\widehat{\sigma}^2\widehat{v}_{jj}/n}, \widehat{\varphi}_j + 1.96\sqrt{\widehat{\sigma}^2\widehat{v}_{jj}/n}\right]$$

contains the true parameter φ_j with approximate probability 95 %, i.e. in about 19 of 20 cases.

Proof. We know from Theorem 10.7 that $\sqrt{n}(\widehat{\varphi}_j - \varphi_j) \xrightarrow{d} N(0, \sigma^2 v_{jj})$ as $n \to \infty$, hence

$$\sqrt{n}\frac{\widehat{\varphi}_j - \varphi_j}{\sqrt{\sigma^2 v_{jj}}} \xrightarrow{d} N(0,1), \quad n \to \infty$$

(that $v_{jj} > 0$ follows from the invertibility of Γ_p and the fact that also Γ_p^{-1} is positive definite). Since $\widehat{\sigma}^2 \xrightarrow{P} \sigma^2$ and $\widehat{v}_{jj} \xrightarrow{P} v_{jj}$ as a consequence of $\widehat{\Gamma}_p \xrightarrow{P} \Gamma_p$ as $n \to \infty$, we see that also

$$\sqrt{n}\frac{\widehat{\varphi}_j - \varphi_j}{\sqrt{\widehat{\sigma^2 v_{jj}}}} \xrightarrow{d} N(0,1), \quad n \to \infty,$$

by Slutsky's lemma. Hence we conclude that

$$\sqrt{\frac{n}{\widehat{\sigma}^2\widehat{v}_{jj}}}(\widehat{\varphi}_j - \varphi_j)$$

is approximately $N(0,1)$–distributed. Since the probability that the standard normal distribution is greater than $\Phi_{1-\alpha/2}$ is $\alpha/2$, and the probability that it is less than $-\Phi_{1-\alpha/2}$ is also $\alpha/2$ by symmetry, we have

$$P\left(\sqrt{\frac{n}{\widehat{\sigma}^2\widehat{v}_{jj}}}(\widehat{\varphi}_j - \varphi_j) > \Phi_{1-\alpha/2}\right) \approx \frac{\alpha}{2} \quad \text{and} \quad P\left(\sqrt{\frac{n}{\widehat{\sigma}^2\widehat{v}_{jj}}}(\widehat{\varphi}_j - \varphi_j) < -\Phi_{1-\alpha/2}\right) \approx \frac{\alpha}{2}.$$

But this shows

$$P\left(\varphi_j > \widehat{\varphi}_j + \Phi_{1-\alpha/2}\sqrt{\frac{\widehat{\sigma}^2\widehat{v}_{jj}}{n}}\right) \approx \frac{\alpha}{2} \quad \text{and} \quad P\left(\varphi_j < \widehat{\varphi}_j - \Phi_{1-\alpha/2}\sqrt{\frac{\widehat{\sigma}^2\widehat{v}_{jj}}{n}}\right) \approx \frac{\alpha}{2}$$

so that

$$P\left(\varphi_j \in \left[\widehat{\varphi}_j - \Phi_{1-\alpha/2}\sqrt{\widehat{\sigma}^2\widehat{v}_{jj}/n}, \widehat{\varphi}_j + \Phi_{1-\alpha/2}\sqrt{\widehat{\sigma}^2\widehat{v}_{jj}/n}\right]\right) \approx 1 - 2\frac{\alpha}{2} = 1 - \alpha,$$

which is the desired approximate confidence interval. The second assertion follows immediately from the first, since it is known that the 97.5%-quantile of the standard normal distribution is 1.96. □

Let us try the Yule–Walker estimator for a real data set:

Example 10.9. Recall the Dow Jones utility index from August 28 - December 18, 1972, as explained in Example 1.5. It had a clear trend and hence was not stationary (cf. Figure 1.4), but after having differenced it once, we got the picture displayed in Figure 3.6, and one might think of fitting a stationary model to it. In Figure 10.1 we display the sample autocorrelation function on the left and the sample partial autocorrelation function on the right. The sample autocorrelation function is significantly different from zero for lag 1 (and probably also 2), the sample partial autocorrelation function is significantly different from zero only for lag 1. Since the partial autocorrelation function of an AR(1) process is zero for lags greater than 1, and its partial autocorrelation function decays exponentially, looking at Figure 10.1 suggests to try an AR(1) process. Using the Yule-Walker estimator from Example 10.3 this leads to (after subtracting the sample mean 0.1336 first)

$$\widehat{\varphi}_1 = 0.421879, \quad \widehat{\sigma}^2 = 0.145669,$$

i.e. the fitted model to the mean corrected times series $Y_t = X_t - 0.1336$ is

$$Y_t = 0.4219 Y_{t-1} + Z_t, \quad (Z_t) \sim WN(0, 0.1457),$$

and the one for (X_t) is

$$X_t - 0.1336 - 0.4219(X_{t-1} - 0.1336) = Z_t, \quad (Z_t) \sim WN(0, 0.1447).$$

Since we have 77 observations, and since $\widehat{\gamma}(0) = 0.17992$, we have $\widehat{v}_{11} = \widehat{\gamma}(0)^{-1}$, hence we get the approximate 95 % confidence bound for φ_1

$$0.4219 \pm (1.96)\sqrt{\frac{0.1479}{0.17992 \cdot 77}} = (0.2194, 0.6244).$$

Now assume that we think that an AR(2) process is the correct choice and let us fit an AR(2) process to it using the Yule–Walker estimators (by hand or using ITSM or R). This leads to the estimated mean corrected model

$$\widehat{\varphi}_1 = 0.3739, \quad \widehat{\varphi}_2 = 0.1138, \quad \widehat{\sigma}^2 = 0.143646,$$

i.e.

$$X_t - 0.1336 = 0.3739(X_{t-1} - 0.1336) + 0.1138(X_{t-2} - 0.1336) = Z_t,$$
$$(Z_t) \sim WN(0, 0.143646).$$

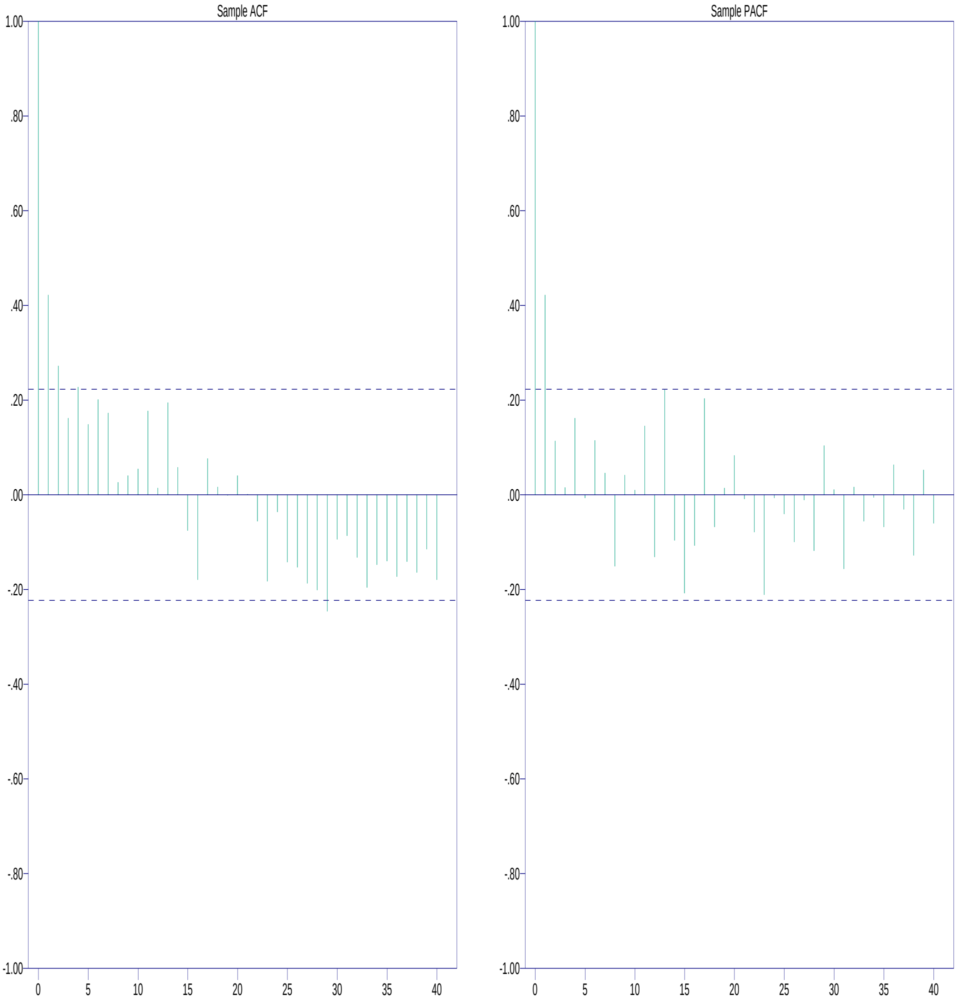
The computer programme ITSM (which comes with the book [BD2] by Brockwell and Davis) provides the confidence bounds in that it calculates the ratio

$$u_j := \frac{\widehat{\varphi}_j}{1.96\sqrt{\widehat{\sigma}^2 \widehat{v}_{jj}/n}},$$

which it calls the “Ratio of the AR coeff. to 1.96$*$ (standard error)”. The approximate 95 % confidence interval for φ_j is then $[\widehat{\varphi}_j - \widehat{\varphi}_j/u_j, \widehat{\varphi}_j + \widehat{\varphi}_j/u_j]$.

For the fitted AR(2) model we have $u_1 = 1.684793$, $u_2 = 0.512738$, hence the 95 % confidence intervals are $[0.1520, 0.5958]$ for φ_1, and $[-0.1079, 0.3353]$ for φ_2; since this second confidence interval contains 0, the value of φ_2 could very well be 0, in accordance with our first intuition to use an AR(1) process.

Figure 10.1: Sample autocorrelation function of the once differenced Dow Jones Utility Index displayed in Figure 3.6



Since the Yule–Walker equations are quite related to the prediction equations, it should be also possible to use some of the prediction machinery to obtain the Yule–Walker equations. In particular, it is possible to obtain the coefficients for a fitted AR(n) model recursively, by using the Durbin–Levinson algorithm. This is in particular useful if one is not sure about the order and fits first various autoregressive models of different order. The precise statement is as follows:

Proposition 10.10. [Recursive calculation of estimated coefficients]
The Yule–Walker estimators can be calculated recursively by applying the Durbin–Levinson algorithm to the sample autocovariance function $\widehat{\gamma}(h)$. More precisely, suppose that we have observations $X_1, \dots, X_T$ and denote the Yule–Walker estimators by $\widehat{\varphi}_{n,1}, \dots, \widehat{\varphi}_{n,n}$ and $\widehat{\sigma}_n^2$ when fitting an AR(n)-model of the form

$$X_t - \widehat{\varphi}_{n,1}X_{t-1} - \ldots - \widehat{\varphi}_{n,n}X_{t-n} = Z_t^{(n)}, \quad (Z_t^{(n)}) \sim WN(0, \widehat{\sigma}_n^2)$$

to the mean corrected data. Then the coefficients can be obtained iteratively from

$$\begin{aligned}
\widehat{\sigma}_0^2 &= \widehat{\gamma}(0), \quad \widehat{\varphi}_{1,1} = \widehat{\gamma}(1)/\widehat{\gamma}(0), \quad \widehat{\sigma}_n^2 = \widehat{\sigma}_{n-1}^2\left[1 - \widehat{\varphi}_{n,n}^2\right], \\
\widehat{\varphi}_{n,n} &= \left[\widehat{\gamma}(n) - \sum_{j=1}^{n-1}\widehat{\varphi}_{n-1,j}\widehat{\gamma}(n-j)\right]\widehat{\sigma}_{n-1}^{-2}, \\
\begin{pmatrix} \widehat{\varphi}_{n,1} \\ \vdots \\ \widehat{\varphi}_{n,n-1} \end{pmatrix} &= \begin{pmatrix} \widehat{\varphi}_{n-1,1} \\ \vdots \\ \widehat{\varphi}_{n-1,n-1} \end{pmatrix} - \widehat{\varphi}_{n,n}\begin{pmatrix} \widehat{\varphi}_{n-1,n-1} \\ \vdots \\ \widehat{\varphi}_{n-1,1} \end{pmatrix}.
\end{aligned}$$

Proof. By Theorem 10.5, the fitted model is causal and has the same theoretical autocovariance function γ_Y as $\widehat{\gamma}$ for lags up to $n \le T$, i.e. $\gamma_Y(h) = \widehat{\gamma}(h)$ for all $h \in \{0, 1, \dots, n\}$. Hence we can calculate the one-step prediction coefficients in the fitted AR(n)-model from the Durbin-Levinson algorithm, using the empirical autocovariance function (normally one would use γ_Y, but as this agrees with $\widehat{\gamma}$, we use these data). But by Proposition 7.20, the coefficients of the predictor in the fitted model are exactly the autoregressive coefficients for the fitted AR(n) process, hence they coincide with the Yule–Walker estimators. Furthermore, the white noise variance of the fitted model is

$$\mathbb{E}\left([X_{n+1} - \widehat{\varphi}_1X_n - \ldots - \widehat{\varphi}_nX_1]^2\right) = \mathbb{E}\left([X_{n+1} - P_nX_{n+1}]^2\right)$$

which is denoted by v_n in the Durbin-Levinson algorithm. This gives the claim. □

This is the end of Unit 13 (for Monday, July 19). There will still be a last Unit 14 for Monday, July 26, but Unit 14 will no longer be relevant for the exam. So the relevant topics for the exam end here. Regarding the exam itself and making appointments, we will make an announcement in moodle how to do this.

Chapter 11

Further estimators and order selection

In this final chapter we present the least squares estimator and the quasi-maximum likelihood estimator. We then move on to order selection criteria, i.e. a method to estimate the orders p and q of an ARMA(p, q) process. In this chapter, we shall not provide detailed proofs as we did so far, but mainly summarise the methods and give references for the proofs.

11.1 The least squares estimator

Let us first introduce the least squares estimator for causal AR(p)-processes.

For that, let $(Z_t)_{t\in\mathbb{Z}} \sim WN(0, \sigma^2)$ (real-valued) with $\sigma^2 > 0$ and $(X_t)_{t\in\mathbb{Z}}$ be a real and causal AR(p)-process with representation

$$X_t - \varphi_1 X_{t-1} - \ldots - \varphi_p X_{t-p} = Z_t, \quad (Z_t) \sim WN(0, \sigma^2) \tag{11.1}$$

(as in (10.1)). Define

$$\vec{\varphi} := (\varphi_1, \ldots, \varphi_p)' \in \mathbb{R}^p,$$

$$\vec{Y} := \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} \in \mathbb{R}^n, \quad \vec{Z} = \begin{pmatrix} Z_1 \\ \vdots \\ Z_n \end{pmatrix} \in \mathbb{R}^n, \quad \text{and} \quad X = \begin{pmatrix} X_0 & X_{-1} & \ldots & X_{1-p} \\ X_1 & X_0 & \ldots & X_{2-p} \\ \vdots & \vdots & & \vdots \\ X_{n-1} & X_{n-2} & \ldots & X_{n-p} \end{pmatrix} \in \mathbb{R}^{n\times p}.$$

Then (11.1) for $t = 1, \ldots, n$ is equivalent to

$$\vec{Y} = X\vec{\varphi} + \vec{Z}. \tag{11.2}$$

This can be seen as a linear regression model. Let $\vec{\varphi}^*$ be the least squares estimator of (11.2), i.e.

$$\vec{\varphi}^* = \text{argmin}_{a\in\mathbb{R}^p} |\vec{Y} - Xa|^2. \tag{11.3}$$

Denote

$$\mathcal{L} := X(\mathbb{R}^p) = \{Xa : a \in \mathbb{R}^p\}.$$

Then $\mathcal{L}$ is a (finite dimensional and hence closed) linear subspace of $\mathbb{R}^n$, and (11.3) means that $X\vec{\varphi}^*$ is the orthogonal projection of $\vec{Y}$ onto $\mathcal{L}$.

From Theorem 7.8 we conclude that $X\vec{\varphi}^*$ is uniquely determined, and that $\vec{\varphi}^*$ exists (but it may not be unique). Since it is an orthogonal projection, we have

$$\begin{aligned} \vec{Y} - X\vec{\varphi}^* &\perp Xz \quad \forall\, z \in \mathbb{R}^p, \quad \text{i.e.} \\ (\vec{Y} - X\vec{\varphi}^*)'Xz &= 0 \quad \forall\, z \in \mathbb{R}^p. \\ \Longrightarrow (\vec{Y} - X\vec{\varphi}^*)'X &= 0 \\ \Longrightarrow X'(\vec{Y} - X\vec{\varphi}^*) &= 0 \\ \Longrightarrow X'\vec{Y} &= X'X\vec{\varphi}^*. \end{aligned}$$

In the case that $X'X$ is invertible, we obtain

$$\vec{\varphi}^* = (X'X)^{-1}X'\vec{Y}. \tag{11.4}$$

Observe that the estimator $\vec{\varphi}^*$ requires also the observations $X_{1-p}, \ldots, X_{-1}, X_0$ and not only $X_1, \ldots, X_n$, but this is not an obstacle since we can easily shift everything by p (and assume that we have $n + p$ observations).

Observe that the (i, j)-element of $n^{-1}X'X$ is equal to $n^{-1}\sum_{k=1}^{n-i} X_k X_{k+|i-j|}$ and the i'th component of $n^{-1}X'\vec{Y}$ is equal to $n^{-1}\sum_{k=1-i}^{n-i} X_k X_{k+i}$. Hence, if γ^* as defined in the previous chapter is a consistent estimator (by Corollary 9.13 this would e.g. be the case for i.i.d. $(0, \sigma^2)$ noise with finite fourth moment, but much less is needed), then

$$n^{-1}X'X \xrightarrow{P} \Gamma_p \quad \text{and} \quad n^{-1}X'\vec{Y} \xrightarrow{P} \vec{\gamma}_p$$

as $n \to \infty$. Since Γ_p is invertible, so will be $X'X$. One can now show:

Theorem 11.1. *In addition to the above assumptions, assume that $(Z_t)_{t\in\mathbb{Z}}$ is i.i.d.$(0, \sigma^2)$ (not necessarily with finite fourth moment). Then*

$$n^{1/2}(\vec{\varphi}^* - \vec{\varphi}) \xrightarrow{d} N(0, \sigma^2\Gamma_p^{-1}) \quad \text{as} \quad n \to \infty$$

and (for the Yule-Walker estimator $\widehat{\vec{\varphi}}$)

$$n^{1/2}(\widehat{\vec{\varphi}} - \vec{\varphi}) \xrightarrow{d} N(0, \sigma^2\Gamma_p^{-1}) \quad \text{as} \quad n \to \infty.$$

Proof. For the first assertion, see Prop. 8.10.1 in Brockwell and Davis [BD1]. The second assertion is Theorem 8.1.1 in Brockwell and Davis [BD1], and reduces the result for the Yule-Walker estimator to that of the least squares estimator. □

11.2 The (quasi–)maximum-likelihood estimator

If something is estimated in statistics, moment estimators (like the Yule–Walker estimator) are often easy to obtain, but in most cases the estimator of choice is a maximum–likelihood estimator, i.e. an estimator that maximises the likelihood of a given observation. So we shall consider maximum–likelihood estimators also here. The problem is that without knowing the distribution of the noise once cannot give the likelihood of the observations either, but one can do so by assuming that the noise is Gaussian. One then maximises this Gaussian likelihood irrelevant if it is the true likelihood or not. One hence speaks of a quasi-maximum-likelihood estimator, since it may not be the true likelihood.

In this section we will be working in the following setting: let $x_1, \ldots, x_T$ be realisations of a causal and invertible ARMA(p,q) process

$$X_t - \varphi_1 X_{t-1} - \ldots - \varphi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \ldots + \theta_q Z_{t-q}, \quad (Z_t) \sim WN(0, \sigma^2),$$

were we assume that they are not all equal. The goal is to estimate $\phi = (\varphi_1, \ldots, \varphi_p)'$, $\theta = (\theta_1, \ldots, \theta_q)'$ and σ^2 from the data $x_1, \ldots, x_T$, equivalently, estimate Δ, where

$$\Delta := (\varphi_1, \ldots, \varphi_p, \theta_1, \ldots, \theta_q, \sigma^2)'.$$

As before we assume that the orders p and q are known.

In order to write down the likelihood we need some notation:

Notation 11.2. For the given realisations $x_1, \ldots, x_T$ of a causal and invertible ARMA(p,q) process, denote

$$\mathbf{x}_T := (x_1, \ldots, x_T)',$$

the vector of observations. Further, for any allowed parameter vector $\Delta = (\varphi_1, \ldots, \varphi_p, \theta_1, \ldots, \theta_q, \sigma^2)'$ that leads to a causal and invertible ARMA(p,q)-process $X = X(\Delta) = (X_t(\Delta))_{t\in\mathbb{Z}}$, denote the *model autocovariance function* by γ_Δ and then define the covariance matrix

$$\Gamma_{T,\Delta} := (\operatorname{Cov}(X_i(\Delta), X_j(\Delta)))_{i,j=1,\ldots,T} = (\gamma_\Delta(i-j))_{i,j=1,\ldots,T}.$$

Further, denote

$$\mathbf{X}_T := (X_1, \ldots, X_T)'$$

so that $\Gamma_{T,\Delta}$ is the covariance matrix of the vector $\mathbf{X}_T$. The idea is that $\mathbf{x}_T$ is then a realisation of the random vector $\mathbf{X}_T$.

Recall from Theorem 2.14 (d) that if $Y = (Y_1, \ldots, Y_T)'$ is a T-dimensional normally (i.e. Gaussian) distributed random vector with expectation $M = \mathbb{E}(Y) \in \mathbb{R}^T$ and covariance matrix $A = \operatorname{Cov}(Y) = (\operatorname{Cov}(Y_i, Y_j))_{i,j=1,\ldots,T} \in \mathbb{R}^{T\times T}$ such that A is invertible, then Y has a density that is given by

$$f_Y(\mathbf{y}) = \frac{1}{\sqrt{(2\pi)^T \det A}} \exp\left(-\frac{1}{2} (\mathbf{y} - M)' A^{-1} (\mathbf{y} - M) \right).$$

Assuming that the innovations $(Z_t)_{t\in\mathbb{Z}}$ in our ARMA(p,q)-process are Gaussian, we can immediately write down the probability density at the observed data $\mathbf{x}_T$, which is nothing else than the likelihood at Δ.

Lemma 11.3. *Let $X = (X_t)_{t\in\mathbb{Z}}$ be a causal and invertible ARMA(p,q) process with parameter vector $\Delta = (\varphi_1,\dots,\varphi_p,\theta_1,\dots,\theta_q,\sigma^2)' \in \mathbb{R}^{p+q+1}$, and assume that the noise $(Z_t)_{t\in\mathbb{Z}}$ is i.i.d. Gaussian $N(0,\sigma^2)$. Then given the observation vector $\mathbf{x}_T$ as defined before, the probability density function of the vector $\mathbf{X}_T$ at $\mathbf{x}_T$ is*

$$f_\Delta(\mathbf{x}_T) = \frac{1}{\sqrt{(2\pi)^T \det \Gamma_{T,\Delta}}} \exp\left(-\frac{1}{2}\mathbf{x}_T'\Gamma_{T,\Delta}^{-1}\mathbf{x}_T\right).$$

Proof. Using the representation of the stationary ARMA process $X_t = \sum_{j\in\mathbb{Z}}\psi_j Z_{t-j}$ with absolutely summable $(\psi_j)_{j\in\mathbb{Z}}$, it follows that with Z_t being normally distributed for all t that also $\mathbf{X}_T = (X_1,\dots,X_T)'$ is normally distributed. Since $\mathbb{E}(X_1) = 0$ and $\mathrm{Cov}\,(\mathbf{X}_T) = \Gamma_{T,\Delta}$ we immediately see the form of the probability density function ($\Gamma_{T,\Delta}$ is invertible by Theorem 7.28). □

The likelihood function is nothing else than the probability density function when the roles of $\mathbf{x}_T$ and f_Δ are reversed. This leads to:

Definition 11.4. In the previous setting, if $x_1,\dots,x_T$ are realisations of a Gaussian ARMA(p,q) process with parameter vector Δ, then

$$L(\Delta|\mathbf{x}_T) := f_\Delta(x_1,\dots,x_T) = \frac{1}{\sqrt{(2\pi)^T \det \Gamma_{T,\Delta}}} \exp\left(-\frac{1}{2}\mathbf{x}_T'\Gamma_{T,\Delta}^{-1}\mathbf{x}_T\right) \qquad (11.5)$$

is called the *likelihood of the parameter vector Δ given the observation* $\mathbf{x}_T = (x_1,\dots,x_T)'$. For a given observation $\mathbf{x}_T$, a parameter vector $\widehat{\Delta} \in \mathbb{R}^{p+q+1}$ is called a *Maximum-Likelihood estimator* of Δ, if

$$L(\widehat{\Delta}|\mathbf{x}_T) \geq L(\delta|\mathbf{x}_T)$$

for all allowed parameter vectors $\delta \in \mathbb{R}^{p+q+1}$, i.e. if

$$\widehat{\Delta} = \mathrm{argmax}_\delta L(\delta|\mathbf{x}_T),$$

and if $\widehat{\Delta}$ is itself allowed (usually meaning that it leads to a causal and invertible ARMA process).

Remark 11.5. Theoretically, everything is said with Definition 11.4. The likelihood is written down (assuming that the noise is Gaussian) and now we “only” have to maximise this, which nowadays with ready computer programs should be no problem anymore. Well, this is not quite correct. To determine the likelihood as written down we have to calculate $\Gamma_{T,\delta}^{-1}$ and $\det\Gamma_{T,\delta}$ for every allowed parameter vector δ to get $L(\delta,\mathbf{x}_T)$, which is computationally very expensive. In particular, the matrix inversion takes a lot of effort. Also, it would be good if we do not have to maximise over all $\delta \in \mathbb{R}^{p\times q+1}$, but only over those δ that we believe already to lie close to the true parameter. Here, the Yule–Walker estimator helps and can be used as a preliminary estimator, and then one maximises only in a neighbourhood of the Yule–Walker estimator. However, one problem remains, namely we need to simplify the calculation of $\mathbf{x}'_T \Gamma_{T,\delta}^{-1} \mathbf{x}_T$ and of $\det \Gamma_{T,\delta}$ that appear in the likelihood. How this can be done is the contents of the next theorem.

Theorem 11.6. *Let $(X_t)_{t\in\mathbb{Z}}$ be a causal and invertible Gaussian ARMA(p,q) process with parameter vector Δ. Define*

$$\widehat{X}_1 = P_0 X_1 := 0, \quad v_0 = v_0(\Delta) := \mathbb{E}[(X_1 - \widehat{X}_1)^2] = \gamma_\Delta(0),$$

and for $n \in \mathbb{N}$

$$\widehat{X}_{n+1} = P_n X_{n+1}, \quad v_n = v_n(\Delta) = \mathbb{E}[(X_{n+1} - P_n X_{n+1})^2],$$

the one-step ahead predictor based on $X_1, \dots, X_n$ and its mean squared error, respectively (all depend on the autocovariance function and hence on the underlying parameter vector Δ, also $P_n X_{n+1}$!). *Then*

$$\det \Gamma_{T,\Delta} = v_0(\Delta) v_1(\Delta) \cdots v_{T-1}(\Delta)$$

and

$$\mathbf{X}'_T \Gamma_{T,\Delta}^{-1} \mathbf{X}_T = \sum_{j=1}^{T} (X_j - \widehat{X}_j)^2 / v_{j-1}(\Delta).$$

Further, if

$$P_{j-1} X_j = \widehat{X}_j = \sum_{i=1}^{j-1} \phi_{j-1,i}(\Delta)\, X_{j-i},$$

(the coefficients depend on Δ!), let

$$\widehat{x}_j(\Delta) := \sum_{i=1}^{j-1} \phi_{j-1,i}(\Delta)\, x_{j-i}.$$

Then

$$\mathbf{x}'_T \Gamma_{T,\Delta}^{-1} \mathbf{x}_T = \sum_{j=1}^{T} (x_j - \widehat{x}_j(\Delta))^2 / v_{j-1}(\Delta).$$

Hence, the likelihood of the process at $\mathbf{x}_T$ can be written as

$$L(\Delta|\mathbf{x}_T) = (2\pi)^{-T/2} (v_0(\Delta) \cdots v_{T-1}(\Delta))^{-1/2} \exp\left\{ -\frac{1}{2} \sum_{j=1}^{T} (x_j - \widehat{x}_j(\Delta))^2 / v_{j-1}(\Delta) \right\}. \tag{11.6}$$

Proof. Brockwell and Davis [BD1], Time Series and Methods, Section 8.6 □

Remark 11.7. (a) The expression in (11.6) is much easier to compute for each Δ than inverting $\Gamma_{T,\Delta}$. It can be computed using the Durbin-Levinson-algorithm or the innovations algorithm.

(b) Theorem 11.6 implies that the dependence of σ^2 can be separated from the other parameters $(\phi', \theta') = (\varphi_1, \dots, \varphi_p, \theta_1, \dots, \theta_q)'$: from the prediction equation, it is easy to see that the variance of the noise does not enter when calculating the coefficients of the predictors, so that

$$\phi_{j,i}(\phi, \theta, \sigma^2) = \phi_{j,i}(\phi, \theta, 1)$$

(here, $\phi_{j,i}(\phi, \theta, \sigma^2)$ denote the coefficients for the one-step predictor $P_j X_{j+1} = \phi_{j,1} X_j + \ldots + \phi_{j,j} X_1$ in the model with parameter vector (ϕ, θ, σ^2)). This implies

$$\widehat{x}_j(\phi, \theta, \sigma^2) = \widehat{x}_j(\phi, \theta, 1)$$

Similarly, one sees

$$\frac{v_j(\phi, \theta, \sigma^2)}{\sigma^2} = v_j(\phi, \theta, 1).$$

Now define

$$r_j = r_j(\phi, \theta) := \frac{v_j(\phi, \theta, \sigma^2)}{\sigma^2} = v_j(\phi, \theta, 1).$$

Then

$$L(\phi, \theta, \sigma^2 | \mathbf{x}_T) = (2\pi\sigma^2)^{-n/2} (r_0 \cdots r_{T-1})^{-1/2} \exp\left\{ -\frac{1}{2} \sigma^{-2} \sum_{j=1}^{T} (x_j - \widehat{x}_j(\phi, \theta, 1))^2 / r_{j-1} \right\}.$$

Taking partial derivatives with respect to σ^2, φ_j and θ_j (which we usually do when we are interested in finding the maximum), it can be shown (cf. Brockwell and Davis [BD1], Section 8.7) that the maximum likelihood estimator $(\widehat{\phi}, \widehat{\theta}, \widehat{\sigma}^2)$ satisfies

$$\widehat{\sigma}^2 = T^{-1} S(\widehat{\phi}, \widehat{\theta}), \tag{11.7}$$

where

$$S(\phi, \theta) := \sum_{j=1}^{T} (x_j - \widehat{x}_j(\phi, \theta))^2 / r_{j-1}(\phi, \theta)$$

and $(\widehat{\phi}, \widehat{\theta})$ *minimises* the *reduced likelihood*

$$l(\phi, \theta) := \log(T^{-1} S(\phi, \theta)) + T^{-1} \sum_{j=1}^{T} \log r_{j-1}(\phi, \theta), \tag{11.8}$$

i.e.

$$(\widehat{\phi}, \widehat{\theta}) = \text{argmin}_{(\phi,\theta)} l(\phi, \theta). \tag{11.9}$$

So this is simpler, we only have to minimise the reduced likelihood that is a function in $(p+q)$-variables rather than $p+q+1$ and then we get the estimator for the variance gratis from (11.7). This is the way one usually obtains the maximum–likelihood–estimator.

So far we assumed that the noise was Gaussian, which allowed us to write down the likelihood and then maximise it. In reality in most cases we will not know that the noise is Gaussian and hence we will not be able to derive the likelihood function. However, we can still write down the expression on the right-hand side of (11.6) and then maximise it, or alternatively write down the reduced likelihood of (11.8) and then minimise it and determine $\widehat{\sigma}^2$ via (11.7). We then speak of a Quasi-Maximum-Likelihood estimator, as it does not maximise the true likelihood, but the likelihood after pretending the process to be Gaussian. Let us write this down again explicitly in a definition:

Definition 11.8. For a causal and invertible ARMA(p, q) process with parameters $\Delta = (\phi, \theta, \sigma^2)$, the quantity $L(\Delta|\mathbf{x}_T)$ defined by (11.6) is called the *Gaussian likelihood of the parameter* Δ (given the observations $\mathbf{x}_T$). The estimator $(\widehat{\phi}, \widehat{\theta}, \widehat{\sigma}^2)$ defined by

$$(\widehat{\phi}, \widehat{\theta}, \widehat{\sigma}^2) = \text{argmax}_{(\phi,\theta,\sigma^2)} L((\phi, \theta, \sigma^2)|\mathbf{x}_T)$$

(over all allowed parameter vectors and such that the estimated parameter is itself allowed), equivalently that satisfies (11.7) – (11.9), is called the *Quasi-Maximum-Likelihood estimator*. It does not maximise the likelihood, but the Gaussian likelihood (i.e. pretending the data were Gaussian).

Remark 11.9. (a) Maximum-Likelihood estimators and also Quasi-Maximum-Likelihood estimators often have very good statistical properties. This is also here the case. It can be shown that if the noise is i.i.d. with mean zero and variance σ^2, and if the characteristic polynomials have no common zeroes, then the Quasi-Maximum-Likelihood estimator will be asymptotically normal, i.e.

$$\widehat{\varphi}_j \sim AN(\varphi_j; T^{-1}, \sigma^2 a_{jj}), \quad \widehat{\theta}_j \sim AN(\theta_j; T^{-1}, \sigma^2 b_{jj}) \quad \text{as} \quad T \to \infty,$$

for suitable a_{jj} and b_{jj}. For the precise determination of a_{jj} and b_{jj}, see Brockwell and Davis [BD1], Section 8.8.
(b) It can be further shown that for causal AR(p) processes, the asymptotic variance is the same as for the Yule-Walker-estimator, i.e. $\sigma^2 a_{jj} = v_{jj}$ in that case.
(c) For the ARMA(1,1) process, one derives (under the given assumptions)

$$a_{11} = \frac{(1-\varphi_1^2)(1+\varphi_1\theta_1)^2}{(\varphi_1+\theta_1)^2}, \quad b_{11} = \frac{(1-\theta_1^2)(1-\varphi_1\theta_1)^2}{(\varphi_1+\theta_1)^2},$$

cf. Brockwell and Davis [BD1], Example 8.8.3.

Example 11.10. For the differenced and mean corrected Dow Jones data from Example 1.5 (cf. also Example 10.9) we obtain when fitting an AR(1) process for the maximum likelihood estimator

$$\widehat{\varphi}_{1,ML} = 0.4471, \quad \widehat{\sigma}^2_{ML} = 0.145509, \quad -2\log L(\widehat{\varphi}_{1,ML}, \widehat{\sigma}^2_{ML}) = 70.320889,$$

while for the Yule-Walker estimator we have

$$\widehat{\varphi}_{1,YW} = 0.4219,\ \widehat{\sigma}^2_{YW} = 0.145669,\ -2\log L_{YW}(\widehat{\varphi}_{1,YW}, \widehat{\sigma}^2_{YW}) = 70.378414.$$

So we see that there actually is not much difference between the Yule–Walker estimator and the Maximum-Likelihood estimator in this example.

11.3 Order selection

So far we have assumed that we knew the true orders p and q of our underlying ARMA process. But is this realistic? If we do not know the parameters, why should we know the orders? So we should also estimate the orders of the ARMA process. How can we do this?

- Clearly, for various fixed orders we should first fit the "best" model, and then compare the best models among them.
- The higher the order, the better the fit, so we must balance between too high orders and a weaker fit.
- Too high orders should be penalised.
- There are various criteria, we cover the AIC, the AICC and the BIC criterion.
- The likelihood seems to be a good measure for goodness of fit. The (quasi) maximum likelihood estimator maximises $L(\Delta|\mathbf{x}_T)$ among all Δ, equivalently, minimises $-\log L(\Delta|\mathbf{x}_T)$ among all Δ.
- So we should minimise $-\log L(\Delta)$ plus some penalty term.

We next give some of the various criteria that are in use. They all have in common that they penalise the (Gaussian) likelihood.

Definition 11.11. Let $x_1, \dots, x_T$ be realisations (not all equal) of a causal and invertible ARMA process with unknown orders. For fixed p, q, let $(\widehat{\phi}_{p,q}, \widehat{\theta}_{p,q}, \widehat{\sigma}^2_{p,q})$ be the maximum likelihood estimator for the given data when an ARMA(p,q) model is fitted. Denote by

$$L_{p,q} := L(\widehat{\phi}_{p,q}, \widehat{\theta}_{p,q}, \widehat{\sigma}^2_{p,q})$$

the corresponding (Gaussian) likelihood of the maximum-likelihood-estimator. Define

$$\begin{aligned}
\mathrm{AIC}(p,q) &:= -2\log L_{p,q} + 2(p+q+1),\\
\mathrm{AICC}(p,q) &:= -2\log L_{p,q} + \frac{2(p+q+1)T}{T-p-q-2},\\
\mathrm{BIC}(p,q) &:= (T-p-q)\log\left(\frac{T\widehat{\sigma}^2_{p,q}}{T-p-q}\right) + T(1+\log\sqrt{2\pi}) + (p+q)\log\frac{\sum_{t=1}^T x_t^2 - T\widehat{\sigma}^2_{p,q}}{p+q}.
\end{aligned}$$

Then the *AIC-criterion (Akaike information criterion)* chooses the orders p and q that minimise AIC(p,q). The *AICC-criterion (corrected Akaike information criterion)* chooses the orders p and q that minimise AICC(p,q), and the *BIC-criterion (Bayesian information criterion)* chooses the orders p and q that minimise BIC(p,q).

Due to our lack of time, we will not try to understand where these correction terms come from (but be assured that they do not fall from heaven but have some deeper meaning), we confine ourselves to know that they serve as order selection criteria when minimising the corresponding term over p and q. The criteria try to counterbalance the influence of too many parameters by penalising the negative loglikelihood. We simply give some facts known from practise:

Remark 11.12. (a) The AIC criterion tends to overfit p, i.e. to choose slightly too large values of p, which is counteracted by the AICC criterion.
(b) The BIC criterion is strongly consistent in the sense that if $x_1, \ldots, x_T$ are observations of an ARMA(p, q) model, and if $\widehat{p}$ and $\widehat{q}$ are the estimated orders by the BIC-criterion, then $\widehat{p} \to p$ and $\widehat{q} \to q$ almost surely as $T \to \infty$.
(c) The AIC and AICC-criteria are not strongly consistent, but have other desirable properties, in that they are "efficient" when it comes to forecasting (see Brockwell and Davis [BD1], Section 9.3, for more information).

Example 11.13. Let us return to the differenced Dow Jones utility index of Example 1.5, cf. also Examples 10.9 and 11.10. Using ITSM (the computer programme that comes with the book by Brockwell and Davis [BD2], but you could equally well use R), we find

$$\begin{aligned} -2\log L_{1,0} &= 70.320889, \quad \mathrm{AICC}(1,0) = 74.483051, \quad \mathrm{BIC}(1,0) = 73.993244, \\ -2\log L_{2,0} &= 69.272882, \quad \mathrm{AICC}(2,0) = 75.601650, \quad \mathrm{BIC}(2,0) = 75.552865, \\ -2\log L_{1,1} &= 68.803250, \quad \mathrm{AICC}(1,1) = 75.13132017, \quad \mathrm{BIC}(1,1) = 75.106867. \end{aligned}$$

So, among AR(1), AR(2) and ARMA(1,1) the AR(1) model is chosen by both AICC and BIC, although the ARMA(1,1) has the highest likelihood (i.e. lowest negative log-likelihood). Using the option "autofit" in ITSM, the programme also chooses the AR(1) model among all ARMA(p, q) models with $p, q \leq 5$ and gives the corresponding maximum-likelihood-estimator (using the AICC-criterion). Observe that when using the AIC-criterion, also the AR(1) model would be chosen.

Example 11.14. Figure 11.1 shows the lake level of the Lake Huron in feet (reduced by 570 feet) during the years 1875 – 1972, taken from ITSM. The sample autocorrelation (left) and sample partial autocorrelation function (right) are plotted in Figure 11.2. We see that the sample autocorrelation function decays exponentially fast and that only the sample partial autocorrelation functions at lags 1 (and 2) play a role. This suggests to try an AR(2), ARMA(1,1) or AR(1) or ARMA(2,1) model. The programme ITSM gives for the AR(2)-model

$$-2\log L_{2,0} = 207.296, \quad \mathrm{AICC}(2,0) = 213.551, \quad \mathrm{BIC}(2,0) = 217.619,$$

and the estimated mean corrected AR(2) model

$$X_t = 1.055 X_{t-1} - 0.2608 X_{t-2} + Z_t, \quad (Z_t) \sim WN(0, 0.478866).$$

For the AR(1) model it gives

$$-2\log L_{1,0} = 213.265, \quad \mathrm{AICC}(1,0) = 217.391, \quad \mathrm{BIC}(1,0) = 218.503,$$

so both AICC as well as BIC are higher for the AR(1) model than for the AR(2) model, so an AR(2) model should be clearly prefered to an AR(1) model. For the ARMA(1,1) model we obtain from ITSM

$$-2\log L_{1,1} = 206.512, \quad \mathrm{AICC}(1,1) = 212.768, \quad \mathrm{BIC}(1,1) = 216.859$$

Figure 11.1: Lake level of Lake Huron in foot during years 1875–1972, cf. Example 11.14.

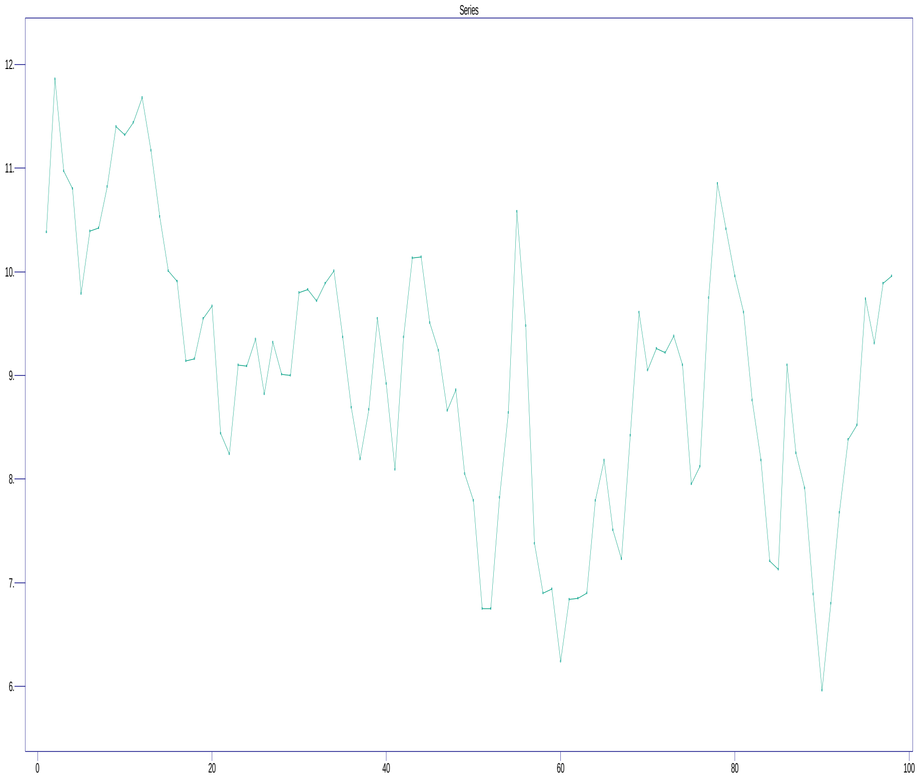
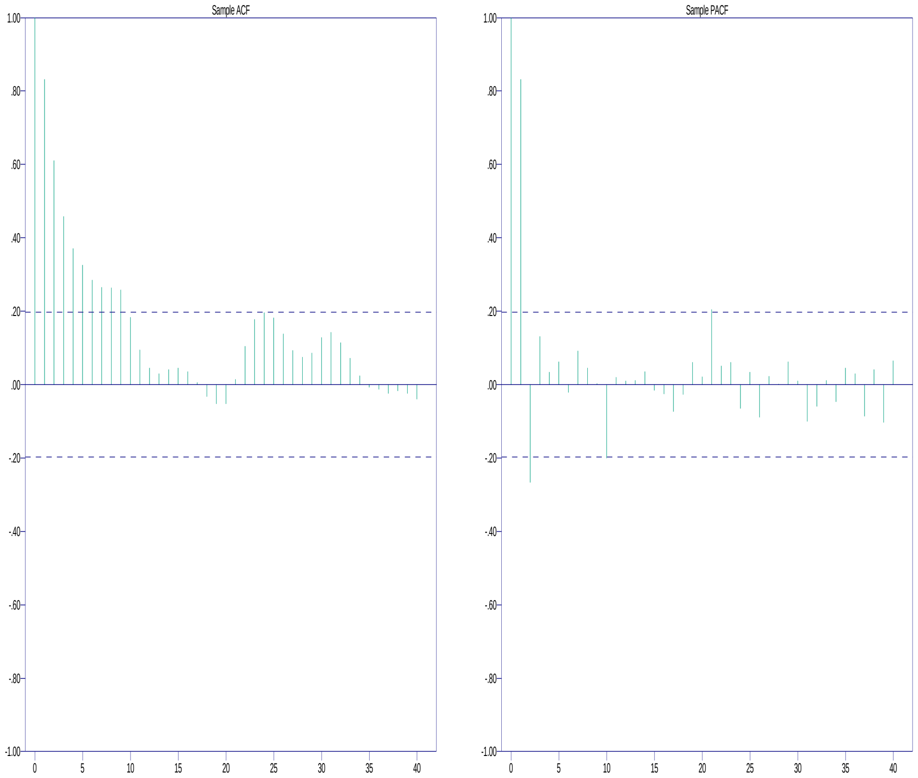


Figure 11.2: Sample autocorrelation function (left) and sample partial autocorrelation function (right) of the Lake Huron data from Figure 11.1, see also Example 11.14.



and the estimated mean corrected model as

$$X_t = 0.7431 X_{t-1} + Z_t + 0.3230 Z_{t-1}, \quad (Z_t) \sim WN(0, 0.475058)$$

So here the AICC and BIC values are even better than the corresponding ones for the AR(2) model, although there is not much difference to the AR(2) model. However, both AICC as well as BIC choose the ARMA(1,1) model, and even if one chooses among all ARMA(p,q)-models with $p, q \leq 10$ (using autofit of ITSM), then the ARMA(1,1) model is chosen. So the model choice should be

$$X_t = 0.7431 X_{t-1} + Z_t + 0.3230 Z_{t-1}, \quad (Z_t) \sim WN(0, 0.475058).$$

It would now be reasonable to do also diagnostic checking, i.e. to look at the residua after the best model has been fitted and check whether they resemble i.i.d. data or at least white noise, but we have done quite enough in this lecture. The interested reader is referred to Brockwell and Davis [BD1], Section 9.4.