Lecture Notes

STA 137
Applied Time Series Analysis

Alexander Aue
University of California, Davis

Spring 2010

Contents

Chapter 1

The Basic Concepts of Time Series Analysis

The first chapter explains the basic notions and highlights some of the objectives of time series analysis. In Section 1.1 we give several important examples, discuss their characteristic features and deduce a general approach to the data analysis. In Section 1.2, stationary processes are identified as a reasonably broad class of random variables which are able to capture the main features extracted from the examples. Finally, we discuss how to treat deterministic trends and seasonal components in Sections 1.3 and 1.4, and assess the residuals in Section 1.5. Section 1.6 concludes.

1.1 Introduction and Examples

The first definition clarifies the notion *time series analysis*.

Definition 1.1.1 (Time Series) *Let $T \neq \emptyset$ be an index set, conveniently being thought of as "time". A family $(X_t)_{t\in T}$ of random variables (random functions) is called a stochastic process. A realization of $(X_t)_{t\in T}$ is called a time series. We will use the notation $(x_t)_{t\in T}$ in the discourse.*

The most common choices for the index set T include the integers $\mathbb{Z} = \{0, \pm 1, \pm 2, \ldots\}$, the positive integers $\mathbb{N} = \{1, 2, \ldots\}$, the nonnegative integers $\mathbb{N}_0 = \{0, 1, 2, \ldots\}$, the real numbers $\mathbb{R} = (-\infty, \infty)$ and the positive halfline $\mathbb{R}_+ = [0, \infty)$. In this class, we are mainly concerned with the first three cases which are subsumed under the notion *discrete time series analysis*.

Oftentimes the stochastic process $(X_t)_{t\in T}$ is itself referred to as a time series, in the sense that a realization is identified with the probabilistic mechanism. The objective of time series analysis is to gain knowledge of this underlying random phenomenon through examining one (and typically only one) realization.

We start with a number of well known examples emphasizing the multitude of possible applications of time series analysis in various scientific fields.

Example 1.1.1 (Wölfer's sunspot numbers) In Figure 1.1, the number of sunspots (that is, dark spots observed on the surface of the sun) observed annually are plotted

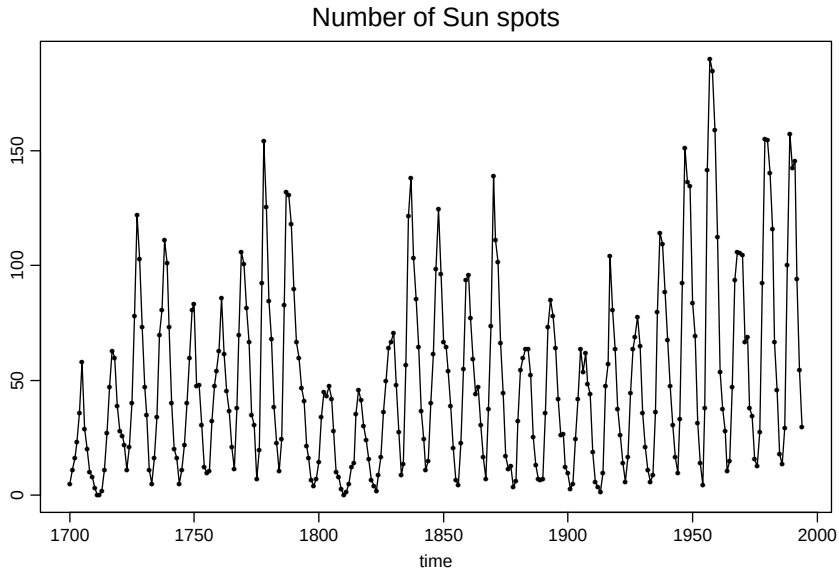

Figure 1.1: Wölfer's sunspot numbers from 1700 to 1994.

against time. The horizontal axis labels time in years, while the vertical axis represents the observed values x_t of the random variable

$$X_t = \#\text{ of sunspots at time } t, \qquad t = 1700, \ldots, 1994.$$

The figure is called a *time series plot*. It is a useful device for a preliminary analysis. Sunspot numbers are used to explain magnetic oscillations on the sun surface.

To reproduce a version of the time series plot in Figure 1.1 using the free software package R[1], download the file `sunspots.dat` from the course webpage and type the following commands:

```
> spots = read.table("sunspots.dat")
> spots = ts(spots, start=1700, frequency=1)
> plot(spots, xlab="time", ylab="", main="Number of Sun spots")
```

In the first line, the file `sunspots.dat` is read into the object `spots`, which is then in the second line transformed into a time series object using the function `ts()`. Using `start` sets the starting value for the x-axis to a prespecified number, while `frequency` presets the number of observations for one unit of time. (Here: one annual observation.) Finally, `plot` is the standard plotting command in R, where `xlab` and `ylab` determine the labels for the x-axis and y-axis, respectively, and `main` gives the headline.

Example 1.1.2 (Canadian lynx data) The time series plot in Figure 1.2 comes from a biological data set. It contains the annual returns of lynx at auction in London by the Hudson Bay Company from 1821–1934 (on a $\log_{10}$ scale). Here, we have observations of the stochastic process

$$X_t = \log_{10}(\text{number of lynx trapped at time } 1820 + t), \qquad t = 1, \ldots, 114.$$

[1]Downloads are available at `http://cran.r-project.org`.

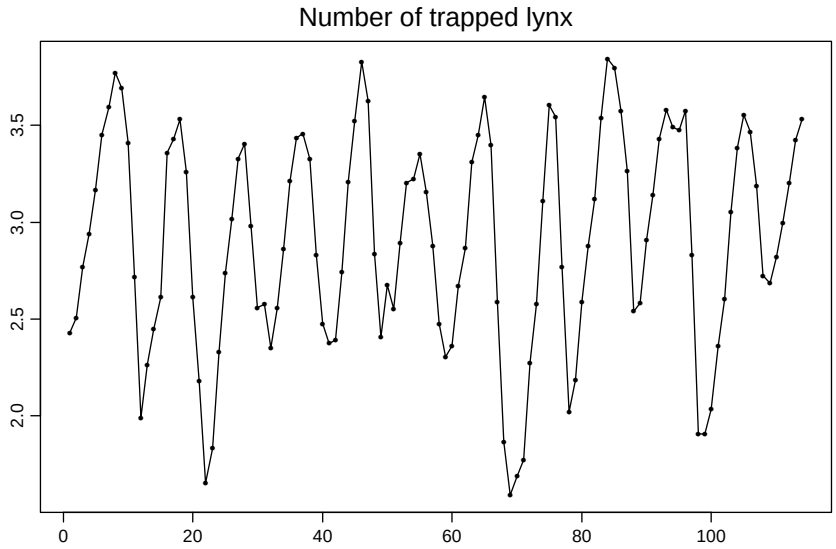

Figure 1.2: Number of lynx trapped in the MacKenzie River district between 1821 and 1934.

The data is used as an estimate for the number of lynx trapped along the MacKenzie River in Canada. This estimate is often used as a proxy for the true population size of the lynx. A similar time series plot could be obtained for the snowshoe rabbit, the primary food source of the Canadian lynx, hinting at an intricate predator-prey relationship.

Assuming that the data is stored in the file `lynx.dat`, the corresponding R commands leading to the time series plot in Figure 1.2 are

```
> lynx = read.table("lynx.dat")
> lynx = ts(log10(lynx), start=1821, frequency=1)
> plot(lynx, xlab="", ylab="", main="Number of trapped lynx")
```

Example 1.1.3 (Treasury bills) Another important field of application for time series analysis lies in the area of finance. To hedge the risks of portfolios, investors commonly use short-term risk-free interest rates such as the yields of three-month, six-month, and twelve-month Treasury bills plotted in Figure 1.3. The (multivariate) data displayed consists of 2,386 weekly observations from July 17, 1959, to December 31, 1999. Here,

$$X_t = (X_{t,1}, X_{t,2}, X_{t,3}), \qquad t = 1, \ldots, 2386,$$

where $X_{t,1}$, $X_{t,2}$ and $X_{t,3}$ denote the three-month, six-month, and twelve-month yields at time t, respectively. It can be seen from the graph that all three Treasury bills are moving very similarly over time, implying a high correlation between the components of X_t.

To produce the three-variate time series plot in Figure 1.3, you can use the R code

```
> bills03 = read.table("bills03.dat");
> bills06 = read.table("bills06.dat");
> bills12 = read.table("bills12.dat");
```

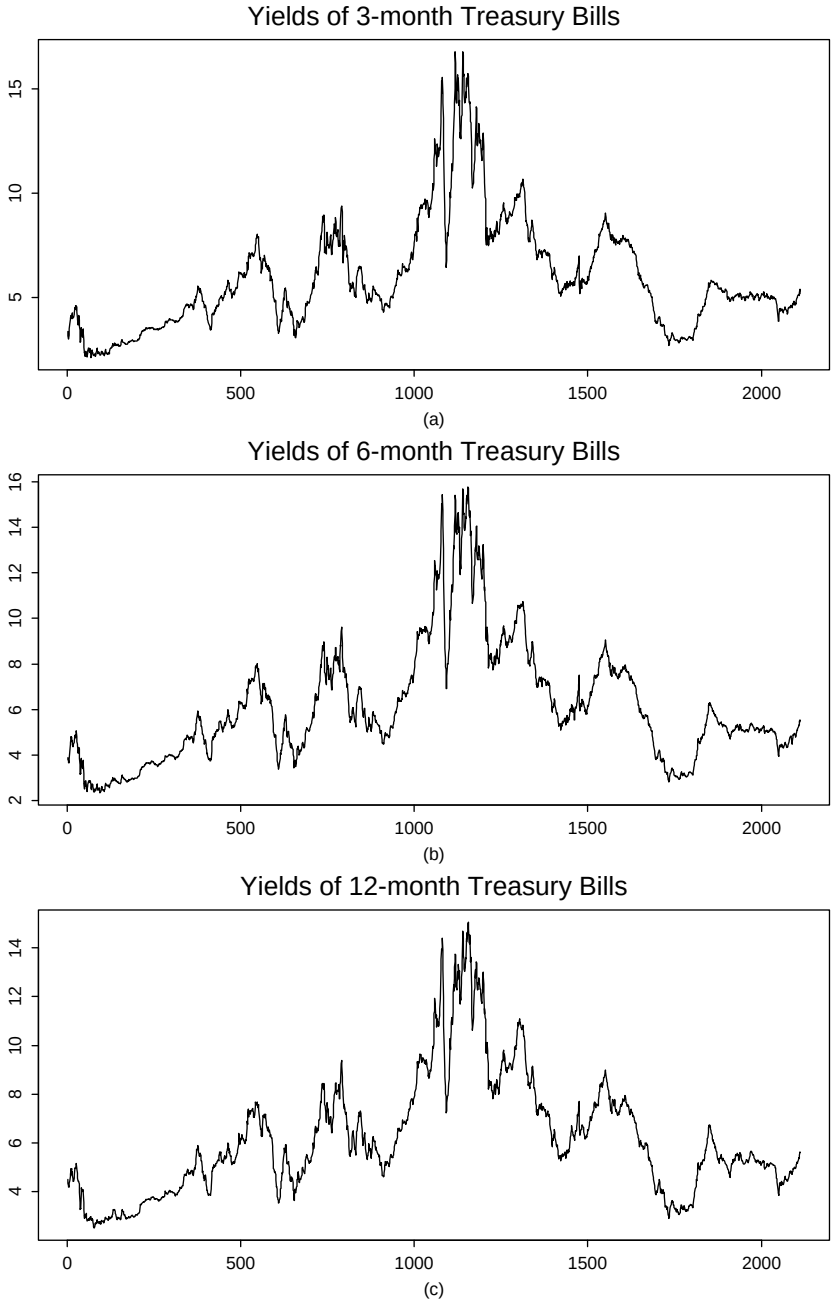

Figure 1.3: Yields of Treasury bills from July 17, 1959, to December 31, 1999.

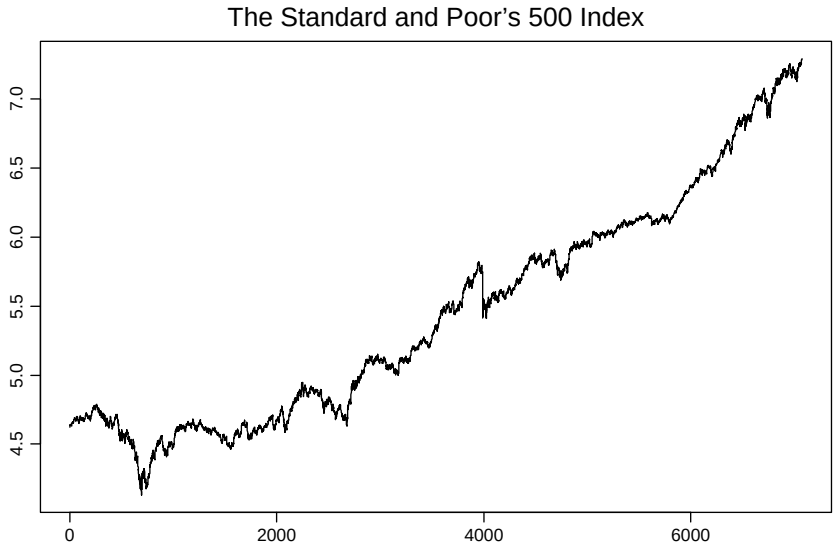

Figure 1.4: S&P 500 from January 3, 1972, to December 31, 1999.

```
> par(mfrow=c(3,1))
> plot.ts(bills03, xlab="(a)", ylab="",
          main="Yields of 3-month Treasury Bills")
> plot.ts(bills06, xlab="(b)", ylab="",
          main="Yields of 6-month Treasury Bills")
> plot.ts(bills12, xlab="(c)", ylab="",
          main="Yields of 12-month Treasury Bills")
```

It is again assumed that the data can be found in the corresponding files `bills03.dat`, `bills06.dat` and `bills12.dat`. The command line `par(mfrow=c(3,1))` is used to set up the graphics. It enables you to save three different plots in the same file.

Example 1.1.4 (S&P 500) The Standard and Poor's 500 index (S&P 500) is a value-weighted index based on the prices of 500 stocks that account for approximately 70% of the U.S. equity market capitalization. It is a leading economic indicator and is also used to hedge market portfolios. Figure 1.4 contains the 7,076 daily S&P 500 closing prices from January 3, 1972, to December 31, 1999, on a natural logarithm scale. We are consequently looking at the time series plot of the process

$$X_t = \ln(\text{closing price of S\&P 500 at time } t), \qquad t = 1, \ldots, 7076.$$

Note that the logarithm transform has been applied to make the returns directly comparable to the percentage of investment return. The time series plot can be reproduced in R using the file `sp500.dat`.

There are countless other examples from all areas of science. To develop a theory capable of handling broad applications, the statistician needs to rely on a mathematical framework that can explain phenomena such as

- trends (apparent in Example 1.1.4);

- seasonal or cyclical effects (apparent in Examples 1.1.1 and 1.1.2);
- random fluctuations (all Examples);
- dependence (all Examples?).

The classical approach taken in time series analysis is to postulate that the stochastic process $(X_t)_{t\in T}$ under investigation can be divided into deterministic trend and seasonal components plus a centered random component, giving rise to the model

$$X_t = m_t + s_t + Y_t, \qquad t \in T, \tag{1.1.1}$$

where $(m_t)_{t\in T}$ denotes the trend function ("mean component"), $(s_t)_{t\in T}$ the seasonal effects and $(Y_t)_{t\in T}$ a (zero mean) stochastic process. After an appropriate model has been chosen, the statistician may aim at

- estimating the model parameters for a better understanding of the time series;
- forecasting future values, for example, to develop investing strategies;
- checking the goodness of fit to the data to confirm that the chosen model is indeed appropriate.

We shall deal in detail with estimation procedures and forecasting techniques in later chapters of these notes. The rest of this chapter will be devoted to introducing the classes of strictly and weakly stationary stochastic processes (in Section 1.2) and to providing tools to eliminate trends and seasonal components from a given time series (in Sections 1.3 and 1.4), while some goodness of fit tests will be presented in Section 1.5.

1.2 Stationary Time Series

Fitting solely independent and identically distributed random variables to data is too narrow a concept. While, on one hand, they allow for a somewhat nice and easy mathematical treatment, their use is, on the other hand, often hard to justify in applications. Our goal is therefore to introduce a concept that keeps some of the desirable properties of independent and identically distributed random variables ("regularity"), but that also considerably enlarges the class of stochastic processes to choose from by allowing dependence as well as varying distributions. Dependence between two random variables X and Y is usually measured in terms of the *covariance function*

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

and the *correlation function*

$$\mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}.$$

With these notations, we can now introduce the classes of strictly and weakly dependent stochastic processes.

Definition 1.2.1 (Strict Stationarity) *A stochastic process $(X_t)_{t\in T}$ is called strictly stationary if, for all $t_1, \ldots, t_n \in T$ and h such that $t_1 + h, \ldots, t_n + h \in T$, it holds that*

$$(X_{t_1}, \ldots, X_{t_n}) \stackrel{\mathcal{D}}{=} (X_{t_1+h}, \ldots, X_{t_n+h}).$$

That is, the so-called finite-dimensional distributions of the process are invariant under time shifts. Here $=^{\mathcal{D}}$ indicates equality in distribution.

The definition in terms of the finite-dimensional distribution can be reformulated equivalently in terms of the cumulative joint distribution function equalities

$$P(X_{t_1} \leq x_1, \ldots, X_{t_n} \leq x_n) = P(X_{t_1+h} \leq x_1, \ldots, X_{t_n+h} \leq x_n)$$

holding true for all $x_1, \ldots, x_n \in \mathbb{R}$, $t_1, \ldots, t_n \in T$ and h such that $t_1 + h, \ldots, t_n + h \in T$. This can be quite difficult to check for a given time series, especially if the generating mechanism of a time series is far from simple, since too many model parameters have to be estimated from the available data rendering concise statistical statements impossible. A possible exception is provided by the case of independent and identically distributed random variables.

To get around these difficulties, a time series analyst will commonly only specify the first- and second-order moments of the joint distributions. Doing so then leads to the notion of weak stationarity.

Definition 1.2.2 (Weak Stationarity) *A stochastic process $(X_t)_{t\in T}$ is called weakly stationary if*

- *the second moments are finite: $E[X_t^2] < \infty$ for all $t \in T$;*
- *the means are constant: $E[X_t] = m$ for all $t \in T$;*
- *the covariance of X_t and X_{t+h} depends on h only:*

$$\gamma(h) = \gamma_X(h) = \mathrm{Cov}(X_t, X_{t+h}), \qquad h \in T \text{ such that } t + h \in T,$$

is independent of $t \in T$ and is called the autocovariance function (ACVF). Moreover,

$$\rho(h) = \rho_X(h) = \frac{\gamma(h)}{\gamma(0)}, \qquad h \in T,$$

is called the autocorrelation function (ACF).

Remark 1.2.1 *If $(X_t)_{t\in T}$ is a strictly stationary stochastic process with finite second moments, then it is also weakly stationary. The converse is not necessarily true. If $(X_t)_{t\in T}$, however, is weakly stationary and Gaussian, then it is also strictly stationary. Recall that a stochastic process is called Gaussian if, for any $t_1, \ldots, t_n \in T$, the random vector $(X_{t_1}, \ldots, X_{t_n})$ is multivariate normally distributed.*

This section is concluded with examples of stationary and nonstationary stochastic processes.

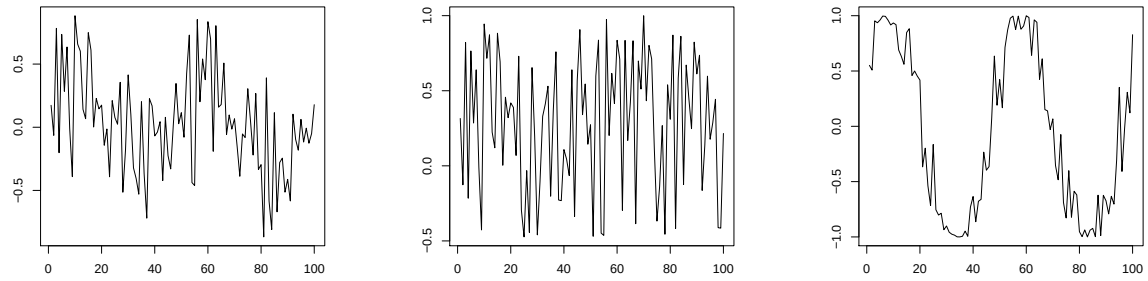

Figure 1.5: 100 simulated values of the cyclical time series (left panel), the stochastic amplitude (middle panel), and the sine part (right panel).

Example 1.2.1 (White Noise) Let $(Z_t)_{t\in\mathbb{Z}}$ be a sequence of real-valued, pairwise uncorrelated random variables with $E[Z_t] = 0$ and $0 < \mathrm{Var}(Z_t) = \sigma^2 < \infty$ for all $t \in \mathbb{Z}$. Then $(Z_t)_{t\in Z}$ is called *white noise*, which shall be abbreviated by $(Z_t)_{t\in\mathbb{Z}} \sim \mathrm{WN}(0, \sigma^2)$. It defines a centered, weakly stationary process with ACVF and ACF given by

$$\gamma(h) = \begin{cases} \sigma^2, & h = 0, \\ 0, & h \neq 0, \end{cases} \qquad \text{and} \qquad \rho(h) = \begin{cases} 1, & h = 0, \\ 0, & h \neq 0, \end{cases}$$

respectively. If the $(Z_t)_{t\in\mathbb{Z}}$ are moreover independent and identically distributed, they are called *iid noise*, shortly $(Z_t)_{t\in\mathbb{Z}} \sim \mathrm{IID}(0, \sigma^2)$. The left panel of Figure 1.6 displays 1000 observations of an iid noise sequence $(Z_t)_{t\in\mathbb{Z}}$ based on standard normal random variables. The corresponding R commands to produce the plot are

```
> z = rnorm(1000, 0, 1)
> plot.ts(z, xlab="", ylab="", main="")
```

The command `rnorm` simulates here 1000 normal random variables with mean 0 and variance 1. There are various built-in random variable generators in R such as the functions `runif(n,a,b)` and `rbinom(n,m,p)` which simulate the n values of a uniform distribution on the interval (a, b) and a binomial distribution with repetition parameter m and success probability p, respectively.

Example 1.2.2 (Cyclical Time Series) Let A and B be uncorrelated random variables with zero mean and variances $\mathrm{Var}(A) = \mathrm{Var}(B) = \sigma^2$, and let $\lambda \in \mathbb{R}$ be a frequency parameter. Define

$$X_t = A\cos(\lambda t) + B\sin(\lambda t), \qquad t \in \mathbb{R}.$$

The resulting stochastic process $(X_t)_{t\in\mathbb{R}}$ is then weakly stationary. Since $\sin(\lambda t + \varphi) = \sin(\varphi)\cos(\lambda t) + \cos(\varphi)\sin(\lambda t)$, the process can be represented as

$$X_t = R\sin(\lambda t + \varphi), \qquad t \in \mathbb{R},$$

so that R is the stochastic amplitude and $\varphi \in [-\pi, \pi]$ the stochastic phase of a *sinusoid*. Easy computations show that we must have $A = R\sin(\varphi)$ and $B = R\cos(\varphi)$. In the left panel of Figure 1.5, 100 observed values of a series $(X_t)_{t\in\mathbb{Z}}$ have been displayed. Therein,

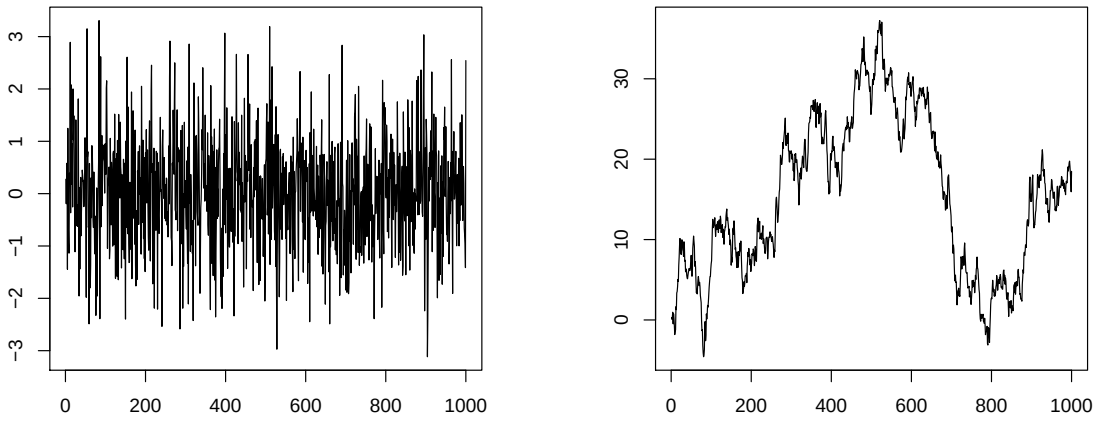

Figure 1.6: 1000 simulated values of iid $\mathcal{N}(0,1)$ noise (left panel) and a random walk with iid $\mathcal{N}(0,1)$ innovations (right panel).

we have used $\lambda = \pi/25$, while R and φ are random variables uniformly distributed on the interval $(-.5, 1)$ and $(0, 1)$, respectively. The middle panel shows the realization of R, the right panel the realization of $\sin(\lambda t + \varphi)$. Using cyclical time series bears great advantages when seasonal effects, such as annually recurrent phenomena, have to be modeled. You can apply the following R commands:

```
> t = 1:100; R = runif(100,-.5,1); phi = runif(100,0,1); lambda = pi/25
> cyc = R*sin(lambda*t+phi)
> plot.ts(cyc, xlab="", ylab="")
```

This produces the left panel of Figure 1.5. The middle and right panels follow in a similar fashion.

Example 1.2.3 (Random Walk) Let $(Z_t)_{t\in\mathbb{N}} \sim \mathrm{WN}(0, \sigma^2)$. Let $S_0 = 0$ and

$$S_t = Z_1 + \ldots + Z_t, \qquad t \in \mathbb{N}.$$

The resulting stochastic process $(S_t)_{t\in\mathbb{N}_0}$ is called a *random walk* and is the most important nonstationary time series. Indeed, it holds here that, for $h > 0$,

$$\mathrm{Cov}(S_t, S_{t+h}) = \mathrm{Cov}(S_t, S_t + R_{t,h}) = t\sigma^2,$$

where $R_{t,h} = Z_{t+1} + \ldots + Z_{t+h}$, and the ACVF obviously depends on t. In R, you may construct a random walk, for example, with the following simple command that utilizes the 1000 normal observations stored in the array `z` of Example 1.2.1.

```
> rw = cumsum(z)
```

The function `cumsum` takes as input an array and returns as output an array of the same length that contains as its jth entry the sum of the first j input entries. The resulting time series plot is shown in the right panel of Figure 1.6.

In Chapter 3 below, we shall discuss in detail so-called autoregressive moving average processes which have become a central building block in time series analysis. They are constructed from white noise sequences by an application of a set of stochastic difference equations similar to the ones defining the random walk $(S_t)_{t\in\mathbb{N}_0}$ of Example 1.2.3.

In general, however, the true parameters of a stationary stochastic process $(X_t)_{t\in T}$ are unknown to the statistician. Therefore, they have to be estimated from a realization $x_1, \ldots, x_n$. We shall mainly work with the following set of estimators. The *sample mean* of $x_1, \ldots, x_n$ is defined as

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

The *sample autocovariance function (sample ACVF)* is given by

$$\hat{\gamma}(h) = \frac{1}{n}\sum_{t=1}^{n-h}(x_{t+h} - \bar{x})(x_t - \bar{x}), \qquad h = 0, 1, \ldots, n-1. \tag{1.2.1}$$

Finally, the *sample autocorrelation function (sample ACF)* is

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}, \qquad h = 0, 1, \ldots, n-1.$$

Example 1.2.4 Let $(Z_t)_{t\in\mathbb{Z}}$ be a sequence of independent standard normally distributed random variables (see the left panel of Figure 1.6 for a typical realization of size $n = 1{,}000$). Then, clearly, $\gamma(0) = \rho(0) = 1$ and $\gamma(h) = \rho(h) = 0$ whenever $h \neq 0$. Table 1.1 gives the corresponding estimated values $\hat{\gamma}(h)$ and $\hat{\rho}(h)$ for $h = 0, 1, \ldots, 5$. The estimated values

h	0	1	2	3	4	5
$\hat{\gamma}(h)$	1.069632	0.072996	−0.000046	−0.000119	0.024282	0.0013409
$\hat{\rho}(h)$	1.000000	0.068244	−0.000043	−0.000111	0.022700	0.0012529

Table 1.1: Estimated ACVF and ACF for selected values of h.

are all very close to the true ones, indicating that the estimators work reasonably well for $n = 1{,}000$. Indeed it can be shown that they are asymptotically unbiased and consistent. Moreover, the sample autocorrelations $\hat{\rho}(h)$ are approximately normal with zero mean and variance $1/1000$. See also Theorem 1.2.1 below. In R, you may use the function `acf` to compute the sample ACF.

Theorem 1.2.1 *Let $(Z_t)_{t\in\mathbb{Z}} \sim \mathrm{WN}(0, \sigma^2)$ and let $h \neq 0$. Under a general set of conditions, it holds that the sample ACF at lag h, $\hat{\rho}(h)$, is for large n approximately normally distributed with zero mean and variance $1/n$.*

Theorem 1.2.1 and Example 1.2.4 suggest a first method to assess whether or not a given data set can be modeled conveniently by a white noise sequence: for a white noise sequence, approximately 95% of the sample ACFs should be within the the confidence interval $\pm 2/\sqrt{n}$. Using the data files on the course webpage, you can compute with R the corresponding sample ACFs to check for whiteness of the underlying time series. We will come back to properties of the sample ACF in Chapter 2.

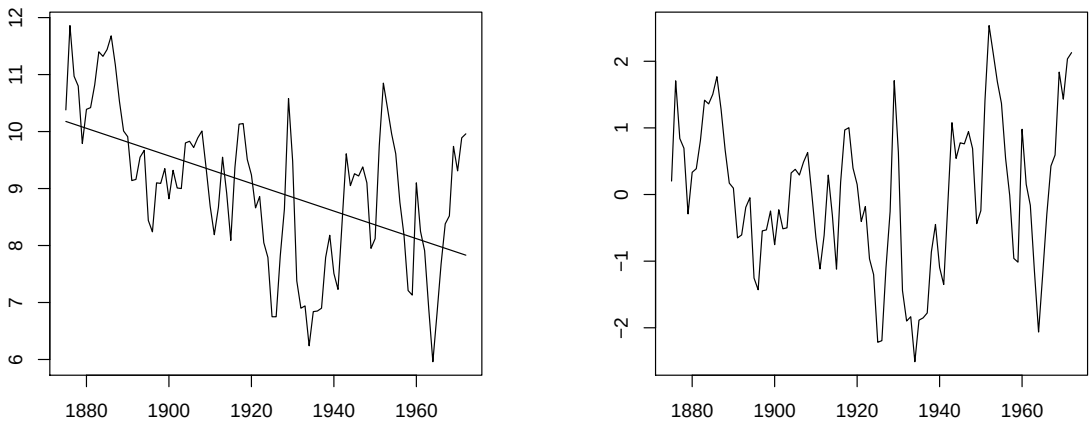

Figure 1.7: Annual water levels of Lake Huron (left panel) and the residual plot obtained from fitting a linear trend to the data (right panel).

1.3 Eliminating Trend Components

In this section we develop three different methods to estimate the trend of a time series model. We assume that it makes sense to postulate the model (1.1.1) with $s_t = 0$ for all $t \in T$, that is,

$$X_t = m_t + Y_t, \qquad t \in T, \tag{1.3.1}$$

where (without loss of generality) $E[Y_t] = 0$. In particular, we will discuss three different methods, (1) the least squares estimation of m_t, (2) smoothing by means of moving averages and (3) differencing.

Method 1 (Least squares estimation) It is often useful to assume that a trend component can be modeled appropriately by a polynomial,

$$m_t = b_0 + b_1 t + \ldots + b_p t^p, \qquad p \in \mathbb{N}_0.$$

In this case, the unknown parameters $b_0, \ldots, b_p$ can be estimated by the least squares method. Combined, they yield the estimated polynomial trend

$$\hat{m}_t = \hat{b}_0 + \hat{b}_1 t + \ldots + \hat{b}_p t^p, \qquad t \in T,$$

where $\hat{b}_0, \ldots, \hat{b}_p$ denote the corresponding least squares estimates. Note that we do not estimate the order p. It has to be selected by the statistician—for example, by inspecting the time series plot. The residuals $\hat{Y}_t$ can be obtained as

$$\hat{Y}_t = X_t - \hat{m}_t = X_t - \hat{b}_0 - \hat{b}_1 t - \ldots - \hat{b}_p t^p, \qquad t \in T.$$

How to assess the goodness of fit of the fitted trend will be subject of Section 1.5 below.

Example 1.3.1 (Level of Lake Huron) The left panel of Figure 1.7 contains the time series of the annual average water levels in feet (reduced by 570) of Lake Huron from 1875 to 1972. We are dealing with a realization of the process

$$X_t = (\text{Average water level of Lake Huron in the year } 1874 + t) - 570, \qquad t = 1, \ldots, 98.$$

There seems to be a linear decline in the water level and it is therefore reasonable to fit a polynomial of order one to the data. Evaluating the least squares estimators provides us with the values

$$\hat{b}_0 = 10.202 \qquad \text{and} \qquad \hat{b}_1 = -0.0242$$

for the intercept and the slope, respectively. The resulting observed residuals $\hat{y}_t = \hat{Y}_t(\omega)$ are plotted against time in the right panel of Figure 1.7. There is no apparent trend left in the data. On the other hand, the plot does not strongly support the stationarity of the residuals. Additionally, there is evidence of dependence in the data.

To reproduce the analysis in R, assume that the data is stored in the file `lake.dat`. Then use the following commands.

```
> lake = read.table("lake.dat")
> lake = ts(lake, start=1875)
> t = 1:length(lake)
> lsfit = lm(lake~t)
> plot(t, lake, xlab="", ylab="", main="")
> lines(lsfit$fit)
```

The function `lm` fits a linear model or regression line to the Lake Huron data. To plot both the original data set and the fitted regression line into the same graph, you can first plot the water levels and then use the `lines` function to superimpose the fit. The residuals corresponding to the linear model fit can be accessed with the command `lsfit$resid`.

Method 2 (Smoothing with Moving Averages) Let $(X_t)_{t\in\mathbb{Z}}$ be a stochastic process following model (1.3.1). Choose $q \in \mathbb{N}_0$ and define the *two-sided moving average*

$$W_t = \frac{1}{2q+1} \sum_{j=-q}^{q} X_{t+j}, \qquad t \in \mathbb{Z}. \tag{1.3.2}$$

The random variables W_t can be utilized to estimate the trend component m_t in the following way. First note that

$$W_t = \frac{1}{2q+1} \sum_{j=-q}^{q} m_{t+j} + \frac{1}{2q+1} \sum_{j=-q}^{q} Y_{t+j} \approx m_t,$$

assuming that the trend is locally approximately linear and that the average of the Y_t over the interval $[t-q, t+q]$ is close to zero. Therefore, m_t can be estimated by

$$\hat{m}_t = W_t, \qquad t = q+1, \ldots, n-q.$$

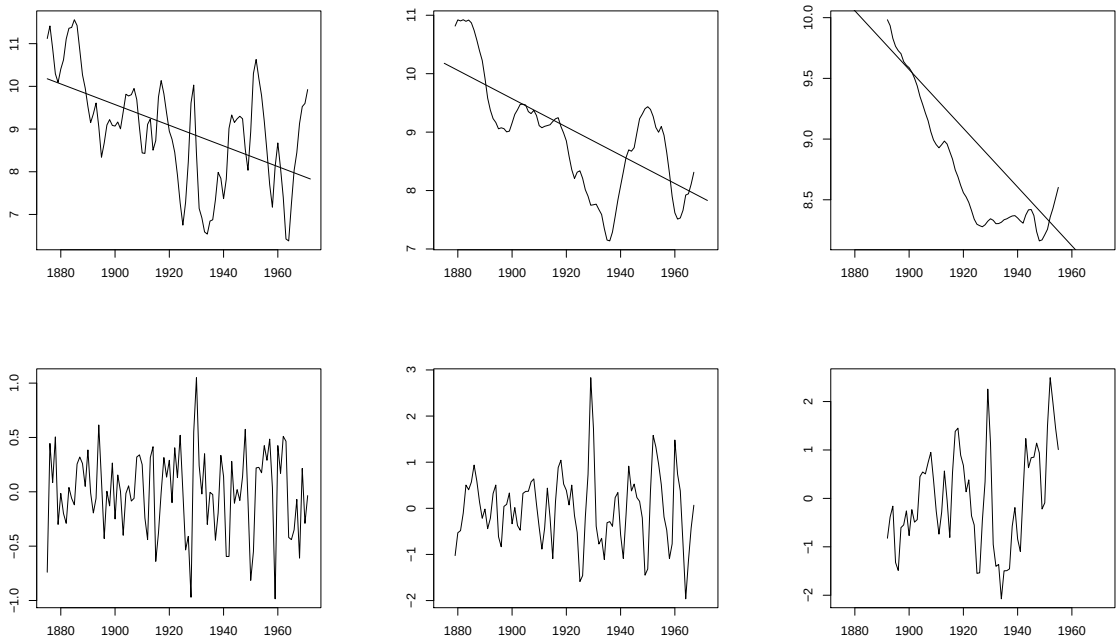

Figure 1.8: The two-sided moving average filters W_t for the Lake Huron data (upper panel) and their residuals (lower panel) with bandwidth $q = 2$ (left), $q = 10$ (middle) and $q = 35$ (right).

Notice that there is no possibility of estimating the first q and last $n - q$ drift terms due to the two-sided nature of the moving averages. In contrast, one can also define *one-sided moving averages* by letting

$$\hat{m}_1 = X_1, \qquad \hat{m}_t = aX_t + (1-a)\hat{m}_{t-1}, \quad t = 2, \ldots, n.$$

Figure 1.8 contains estimators $\hat{m}_t$ based on the two-sided moving averages for the Lake Huron data of Example 1.3.1 for selected choices of q (upper panel) and the corresponding estimated residuals (lower panel).

The moving average filters for this example can be produced in R in the following way:

```
> t = 1:length(lake)
> ma2 = filter(lake, sides=2, rep(1,5)/5)
> ma10 = filter(lake, sides=2, rep(1,21)/21)
> ma35 = filter(lake, sides=2, rep(1,71)/71)
> plot(t, ma2, xlab="", ylab="")
> lines(ma10); lines(ma35)
```

Therein, `sides` determines if a one- or two-sided filter is going to be used. The phrase `rep(1,5)` creates a vector of length 5 with each entry being equal to 1.

More general versions of the moving average smoothers can be obtained in the following way. Observe that in the case of the two-sided version W_t each variable $X_{t-q}, \ldots, X_{t+q}$ obtains a "weight" $a_j = (2q+1)^{-1}$. The sum of all weights thus equals one. The same is true for the one-sided moving averages with weights a and $1 - a$. Generally, one can

hence define a smoother by letting

$$\hat{m}_t = \sum_{j=-q}^{q} a_j X_{t+j}, \qquad t = q+1, \ldots, n-q, \tag{1.3.3}$$

where $a_{-q} + \ldots + a_q = 1$. These general moving averages (two-sided and one-sided) are commonly referred to as *linear filters*. There are countless choices for the weights. The one here, $a_j = (2q+1)^{-1}$, has the advantage that linear trends pass undistorted. In the next example, we introduce a filter which passes cubic trends without distortion.

Example 1.3.2 (Spencer's 15-point moving average) Suppose that the filter in display (1.3.3) is defined by weights satisfying $a_j = 0$ if $|j| > 7$, $a_j = a_{-j}$ and

$$(a_0, a_1, \ldots, a_7) = \frac{1}{320}(74, 67, 46, 21, 3, -5, -6, -3).$$

Then, the corresponding filters passes cubic trends $m_t = b_0 + b_1 t + b_2 t^2 + b_3 t^3$ undistorted. To see this, observe that

$$\sum_{j=-7}^{7} a_j = 1 \qquad \text{and} \qquad \sum_{j=-7}^{7} j^r a_j = 0, \qquad r = 1, 2, 3.$$

Now apply Proposition 1.3.1 below to arrive at the conclusion. Assuming that the observations are in `data`, you may use the R commands

```
> a =c(-3, -6, -5, 3, 21, 46, 67, 74, 67, 46, 21, 3, -5, -6, -3)/320
> s15 = filter(data, sides=2, a)
```

to apply Spencer's 15-point moving average filter. This example also explains how to specify a general tailor-made filter for a given data set.

Proposition 1.3.1 *A linear filter (1.3.3) passes a polynomial of degree p if and only if*

$$\sum_{j} a_j = 1 \qquad \textit{and} \qquad \sum_{j} j^r a_j = 0, \qquad r = 1, \ldots, p.$$

Proof. It suffices to show that $\sum_j a_j (t+j)^r = t^r$ for $r = 0, \ldots, p$. Using the binomial theorem, we can write

$$\begin{aligned} \sum_j a_j (t+j)^r &= \sum_j a_j \sum_{k=0}^{r} \binom{r}{k} t^k j^{r-k} \\ &= \sum_{k=0}^{r} \binom{r}{k} t^k \left(\sum_j a_j j^{r-k} \right) \\ &= t^r \end{aligned}$$

for any $r = 0, \ldots, p$ if and only if the above conditions hold. This completes the proof. □

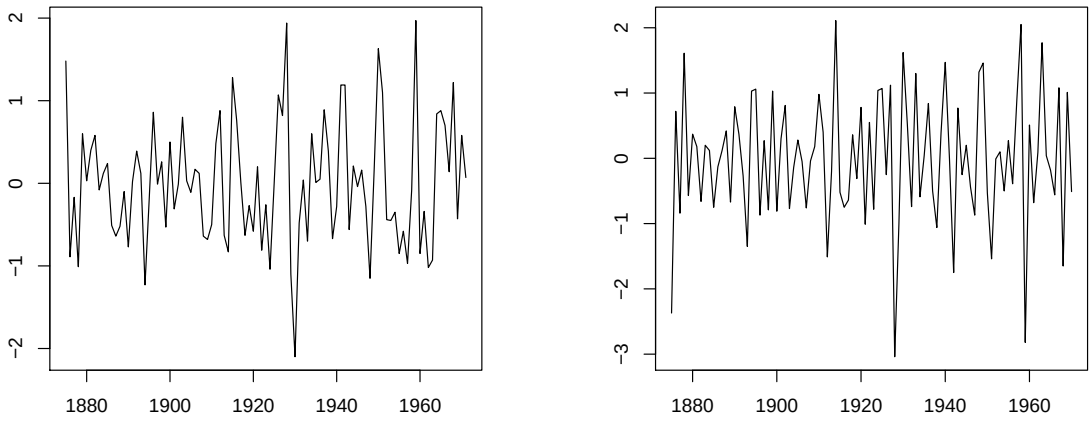

Figure 1.9: Time series plots of the observed sequences (∇x_t) in the left panel and $(\nabla^2 x_t)$ in the right panel of the differenced Lake Huron data described in Example 1.3.1.

Method 3 (Differencing) A third possibility to remove drift terms from a given time series is differencing. To this end, we introduce the *difference operator* ∇ as

$$\nabla X_t = X_t - X_{t-1} = (1 - B)X_t, \qquad t \in T,$$

where B denotes the backshift operator $BX_t = X_{t-1}$. Repeated application of ∇ is defined in the intuitive way:

$$\nabla^2 X_t = \nabla(\nabla X_t) = \nabla(X_t - X_{t-1}) = X_t - 2X_{t-1} + X_{t-2}$$

and, recursively, the representations follow also for higher powers of ∇. Suppose that you are applying the difference operator to a linear trend $m_t = b_0 + b_1 t$, then you obtain

$$\nabla m_t = m_t - m_{t-1} = b_0 + b_1 t - b_0 - b_1(t-1) = b_1$$

which is a constant. Inductively, this leads to the conclusion that for a polynomial drift of degree p, namely $m_t = \sum_{j=0}^{p} b_j t^j$, we have that $\nabla^p m_t = p! b_p$ and thus constant. Applying this technique to a stochastic process of the form (1.3.1) with a polynomial drift m_t, yields then

$$\nabla^p X_t = p! b_p + \nabla^p Y_t, \qquad t \in T.$$

This is a stationary process with mean $p! b_p$. The plots in Figure 1.9 contain the first and second differences for the Lake Huron data. In R, they may be obtained from the commands

```
> d1 = diff(lake)
> d2 = diff(d1)
> par(mfrow=c(1,2))
> plot.ts(d1, xlab="", ylab="")
> plot.ts(d2, xlab="", ylab="")
```

The next example shows that the difference operator can also be applied to a random walk to create stationary data.

Example 1.3.3 *Let $(S_t)_{t\in\mathbb{N}_0}$ be the random walk of Example 1.2.3. If we apply the difference operator ∇ to this stochastic process, we obtain*

$$\nabla S_t = S_t - S_{t-1} = Z_t, \qquad t \in \mathbb{N}.$$

In other words, ∇ does nothing else but recover the original white noise sequence that was used to build the random walk.

1.4 Eliminating Trend and Seasonal Components

Let us go back to the classical decomposition (1.1.1),

$$X_t = m_t + s_t + Y_t, \qquad t \in T,$$

with $E[Y_t] = 0$. In this section, we shall discuss three methods that aim at estimating both the trend and seasonal components in the data. As additional requirement on $(s_t)_{t\in T}$, we assume that

$$s_{t+d} = s_t, \qquad \sum_{j=1}^{d} s_j = 0,$$

where d denotes the period of the seasonal component. (If we are dealing with yearly data sampled monthly, then obviously $d = 12$.) It is convenient to relabel the observations $x_1, \ldots, x_n$ in terms of the seasonal period d as

$$x_{j,k} = x_{k+d(j-1)}.$$

In the case of yearly data, observation $x_{j,k}$ thus represents the data point observed for the kth month of the jth year. For convenience we shall always refer to the data in this fashion even if the actual period is something other than 12.

Method 1 (Small trend method) If the changes in the drift term appear to be small, then it is reasonable to assume that the drift in year j, say, m_j is constant. As a natural estimator we can therefore apply

$$\hat{m}_j = \frac{1}{d}\sum_{k=1}^{d} x_{j,k}.$$

To estimate the seasonality in the data, one can in a second step utilize the quantities

$$\hat{s}_k = \frac{1}{N}\sum_{j=1}^{N}(x_{j,k} - \hat{m}_j),$$

where N is determined by the equation $n = Nd$, provided that data has been collected over N *full* cycles. Direct calculations show that these estimators possess the property $\hat{s}_1 + \ldots + \hat{s}_d = 0$ (as in the case of the true seasonal components s_t). To further assess the quality of the fit, one needs to analyze the observed residuals

$$\hat{y}_{j,k} = x_{j,k} - \hat{m}_j - \hat{s}_k.$$

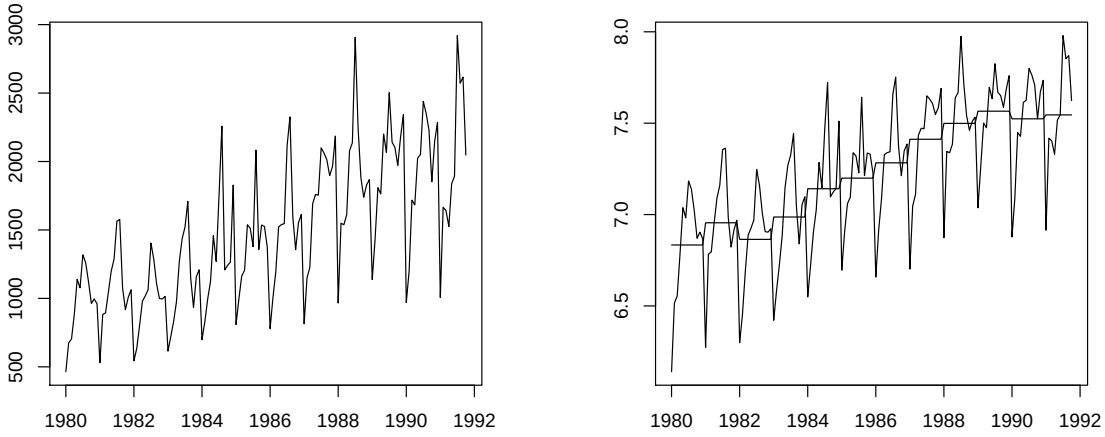

Figure 1.10: Time series plots of the red wine sales in Australia from January 1980 to October 1991 (left) and its log transformation with yearly mean estimates (right).

Note that due to the relabeling of the observations and the assumption of a slowly changing trend, the drift component is solely described by the "annual" subscript j, while the seasonal component only contains the "monthly" subscript k.

Example 1.4.1 (Australian wine sales) The left panel of Figure 1.10 shows the monthly sales of red wine (in kiloliters) in Australia from January 1980 to October 1991. Since there is an apparent increase in the fluctuations over time, the right panel of the same figure shows the natural logarithm transform of the data. There is clear evidence of both trend and seasonality. In the following, we will continue to work with the log transformed data. Using the small trend method as described above, we first estimate the annual means, which are already incorporated in the right time series plot of Figure 1.10. Note that there are only ten months of data available for the year 1991, so that the estimation has to be adjusted accordingly. The detrended data is shown in the left panel of Figure 1.11. The middle plot in the same figure shows the estimated seasonal component, while the right panel displays the residuals. Even though the assumption of small changes in the drift is somewhat questionable, the residuals appear to look quite nice. They indicate that there is dependence in the data (see Section 1.5 below for more on this subject).

Method 2 (Moving average estimation) This method is to be preferred over the first one whenever the underlying trend component is not constant. Three steps are to be applied to the data.

1st Step: Trend estimation. At first, we focus on the removal of the trend component with the linear filters discussed in the previous section. If the period d is odd, then we can directly use $\hat{m}_t = W_t$ as in (1.3.2) with q specified by the equation $d = 2q + 1$. If the period $d = 2q$ is even, then we slightly modify W_t and use

$$\hat{m}_t = \frac{1}{d}(.5x_{t-q} + x_{t-q+1} + \ldots + x_{t+q-1} + .5x_{t+q}), \qquad t = q+1, \ldots, n-q.$$

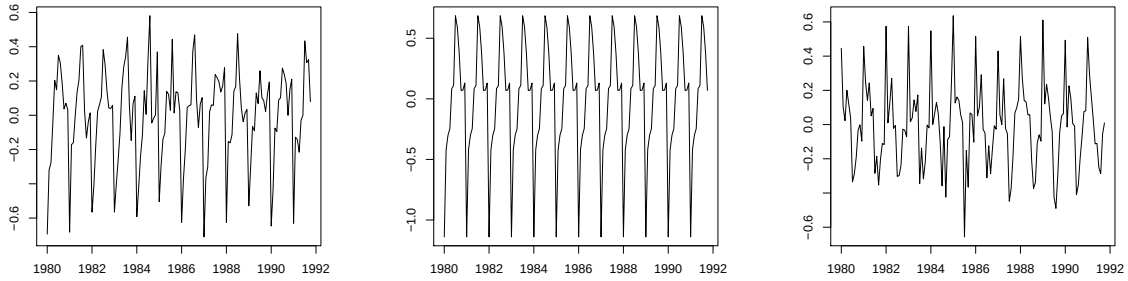

Figure 1.11: The detrended log series (left), the estimated seasonal component (center) and the corresponding residuals series (right) of the Australian red wine sales data.

2nd Step: Seasonality estimation. To estimate the seasonal component, let

$$\mu_k = \frac{1}{N-1}\sum_{j=2}^{N}(x_{k+d(j-1)} - \hat{m}_{k+d(j-1)}), \qquad k = 1, \ldots, q,$$

$$\mu_k = \frac{1}{N-1}\sum_{j=1}^{N-1}(x_{k+d(j-1)} - \hat{m}_{k+d(j-1)}), \qquad k = q+1, \ldots, d.$$

Define now

$$\hat{s}_k = \mu_k - \frac{1}{d}\sum_{\ell=1}^{d}\mu_\ell, \qquad k = 1, \ldots, d,$$

and set $\hat{s}_k = \hat{s}_{k-d}$ whenever $k > d$. This will provide us with *deseasonalized data* which can be examined further. In the final step, any remaining trend can be removed from the data.

3rd Step: Trend Reestimation. Apply any of the methods from Section 1.3.

Method 3 (Differencing at lag d) Introducing the *lag-d difference operator* ∇_d, defined by letting

$$\nabla_d X_t = X_t - X_{t-d} = (1 - B^d)X_t, \qquad t = d+1, \ldots, n,$$

and assuming model (1.1.1), we arrive at the transformed random variables

$$\nabla_d X_t = m_t - m_{t-d} + Y_t - Y_{t-d}, \qquad t = d+1, \ldots, n.$$

Note that the seasonality is removed, since $s_t = s_{t-d}$. The remaining noise variables $Y_t - Y_{t-d}$ are stationary and have zero mean. The new trend component $m_t - m_{t-d}$ can be eliminated using any of the methods developed in Section 1.3.

Example 1.4.2 (Australian wine sales) We revisit the Australian red wine sales data of Example 1.4.1 and apply the differencing techniques just established. The left plot of Figure 1.12 shows the the data after an application of the operator ∇_{12}. If we decide to estimate the remaining trend in the data with the differencing method from Section 1.3, we arrive at the residual plot given in the right panel of Figure 1.12. Note that the order of application does not change the residuals, that is, $\nabla\nabla_{12}x_t = \nabla_{12}\nabla x_t$. The middle panel of Figure 1.12 displays the differenced data which still contains the seasonal component.

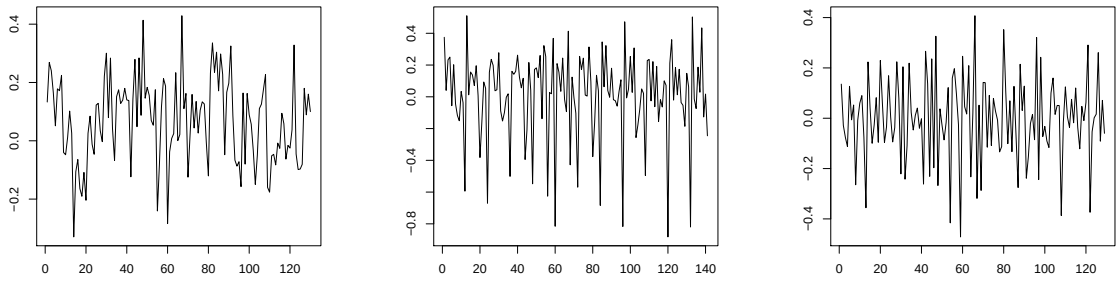

Figure 1.12: The differenced observed series $\nabla_{12}x_t$ (left), ∇x_t (middle) and $\nabla\nabla_{12}x_t = \nabla_{12}\nabla x_t$ (right) for the Australian red wine sales data.

1.5 Assessing the Residuals

In this subsection, we introduce several goodness-of-fit tests to further analyze the residuals obtained after the elimination of trend and seasonal components. The main objective is to determine whether or not these residuals can be regarded as obtained from a sequence of independent, identically distributed random variables or if there is dependence in the data. Throughout we denote by $Y_1, \ldots, Y_n$ the residuals and by $y_1, \ldots, y_n$ a typical realization.

Method 1 (The sample ACF) We have seen in Example 1.2.4 that, for $j \neq 0$, the estimators $\hat{\rho}(j)$ of the ACF $\rho(j)$ are asymptotically independent and normally distributed with mean zero and variance n^{-1}, provided the underlying residuals are independent and identically distributed with a finite variance. Therefore, plotting the sample ACF for a certain number of lags, say h, we expect that approximately 95% of these values are within the bounds $\pm 1.96/\sqrt{n}$. The R function `acf` helps you to perform this analysis. (See Theorem 1.2.1.)

Method 2 (The Portmanteau test) The Portmanteau test is based on the test statistic

$$Q = n \sum_{j=1}^{h} \hat{\rho}^2(j).$$

Using the fact that the variables $\sqrt{n}\hat{\rho}(j)$ are asymptotically standard normal, it becomes apparent that Q itself can be approximated with a chi-squared distribution possessing h degrees of freedom. We now reject the hypothesis of independent and identically distributed residuals at the level α if $Q > \chi^2_{1-\alpha}(h)$, where $\chi^2_{1-\alpha}(h)$ is the $1-\alpha$ quantile of the chi-squared distribution with h degrees of freedom. Several refinements of the original Portmanteau test have been established in the literature. We refer here only to the papers Ljung and Box (1978), and McLeod and Li (1983) for further information on this topic.

Method 3 (The rank test) This test is very useful for finding linear trends. Denote by

$$\Pi = \#\{(i,j) : Y_i > Y_j,\, i > j,\, i = 2, \ldots, n\}$$

the random number of pairs (i,j) satisfying the conditions $Y_i > Y_j$ and $i > j$. Clearly, there are $\binom{n}{2} = \frac{1}{2}n(n-1)$ pairs (i,j) such that $i > j$. If $Y_1, \ldots, Y_n$ are independent and

identically distributed, then $P(Y_i > Y_j) = 1/2$ (assuming a continuous distribution). Now it follows that $\mu_\Pi = E[\Pi] = \frac{1}{4}n(n-1)$ and, similarly, $\sigma_\Pi^2 = \mathrm{Var}(\Pi) = \frac{1}{72}n(n-1)(2n+5)$. Moreover, for large enough sample sizes n, Π has an approximate normal distribution with mean μ_Π and variance σ_Π^2. Consequently, one would reject the hypothesis of independent, identically distributed data at the level α if

$$P = \frac{|\Pi - \mu_\Pi|}{\sigma_\Pi} > z_{1-\alpha/2},$$

where $z_{1-\alpha/2}$ denotes the $1-\alpha/2$ quantile of the standard normal distribution.

Method 4 (Tests for normality) If there is evidence that the data are generated by Gaussian random variables, one can create the *qq plot* to check for normality. It is based on a visual inspection of the data. To this end, denote by $Y_{(1)} < \ldots < Y_{(n)}$ the order statistics of the residuals $Y_1, \ldots, Y_n$ which are normally distributed with expected value μ and variance σ^2. It holds that

$$E[Y_{(j)}] = \mu + \sigma E[X_{(j)}], \tag{1.5.1}$$

where $X_{(1)} < \ldots < X_{(n)}$ are the order statistics of a standard normal distribution. The qq plot is defined as the graph of the pairs $(E[X_{(1)}], Y_{(1)}), \ldots, (E[X_{(n)}], Y_{(n)})$. According to display (1.5.1), the resulting graph will be approximately linear with the squared correlation R^2 of the points being close to 1. The assumption of normality will thus be rejected if R^2 is "too" small. It is common to approximate $E[X_{(j)}] \approx \Phi_j = \Phi^{-1}((j-.5)/n)$ (Φ being the distribution function of the standard normal distribution) and the previous statement is made precise by letting

$$R^2 = \frac{\left[\sum_{j=1}^n (Y_{(j)} - \bar{Y})\Phi_j\right]^2}{\sum_{j=1}^n (Y_{(j)} - \bar{Y})^2 \sum_{j=1}^n \Phi_j^2},$$

where $\bar{Y} = \frac{1}{n}(Y_1 + \ldots + Y_n)$. The critical values for R^2 are tabulated and can be found, for example in Shapiro and Francia (1972). The corresponding R function is `qqnorm`.

1.6 Summary

In this chapter, we have introduced the classical decomposition (1.1.1) of a time series into a drift component, a seasonal component and a sequence of residuals. We have provided methods to estimate the drift and the seasonality. Moreover, we have identified the class of stationary processes as a reasonably broad class of random variables. We have introduced several ways to check whether or not the resulting residuals can be considered to be independent, identically distributed. In Chapter 3, we will discuss in depth the class of autoregressive moving average (ARMA) processes, a parametric class of random variables that are at the center of linear time series analysis because they are able to capture a wide range of dependence structures and allow for a thorough mathematical treatment. Before, we are dealing with the properties of the sample mean, sample ACVF and ACF in the next chapter.

Chapter 2

The Estimation of Mean and Covariances

In this brief second chapter, we will collect some results concerning asymptotic properties of the sample mean and the sample ACVF. Throughout, we denote by $(X_t)_{t\in\mathbb{Z}}$ a weakly stationary stochastic process with mean μ and ACVF γ. In Section 1.2 we have seen that such a process is completely characterized by these two quantities. We have estimated μ by computing the sample mean $\bar{x}$, and γ by $\hat{\gamma}$ defined in (1.2.1). In the following, we shall discuss the properties of these estimators in more detail.

2.1 Estimation of the Mean

Assume that we have to find an appropriate guess for the unknown mean μ of some weakly stationary stochastic process $(X_t)_{t\in\mathbb{Z}}$. The sample mean $\bar{x}$, easily computed as the average of n observations $x_1, \ldots, x_n$ of the process, has been identified as suitable in Section 1.2. To investigate its theoretical properties, we need to analyze the random variable associated with it, that is,

$$\bar{X}_n = \frac{1}{n}(X_1 + \ldots + X_n).$$

Two facts can be quickly established.

- $\bar{X}_n$ is an *unbiased* estimator for μ, since

$$E[\bar{X}_n] = E\left[\frac{1}{n}\sum_{t=1}^{n} X_t\right] = \frac{1}{n}\sum_{t=1}^{n} E[X_t] = \frac{1}{n}n\mu = \mu.$$

  This means that "on average", we estimate the true but unknown μ. Notice that there is no difference in the computations between the standard case of independent and identically distributed random variables and the more general weakly stationary process considered here.

- If $\gamma(n) \to 0$ as $n \to \infty$, then $\bar{X}_n$ is a *consistent* estimator for μ, since

$$\begin{aligned}\mathrm{Var}(\bar{X}_n) &= \mathrm{Cov}\left(\frac{1}{n}\sum_{i=1}^{n} X_i, \frac{1}{n}\sum_{j=1}^{n} X_j\right) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{Cov}(X_i, X_j) \\ &= \frac{1}{n^2}\sum_{i-j=-n}^{n}(n-|i-j|)\gamma(i-j) = \frac{1}{n}\sum_{h=-n}^{n}\left(1-\frac{|h|}{n}\right)\gamma(h).\end{aligned}$$

  Now, the quantity on the right-hand side converges to zero as $n \to \infty$ because $\gamma(n) \to 0$ as $n \to \infty$ by assumption. The first equality sign in the latter equation array follows from the fact that $\mathrm{Var}(X) = \mathrm{Cov}(X, X)$ for any random variable X, the second equality sign uses that the covariance function is linear in both arguments. For the third equality, you can use that $\mathrm{Cov}(X_i, X_j) = \gamma(i-j)$ and that each $\gamma(i-j)$ appears exactly $n - |i-j|$ times in the double summation. Finally, the right-hand side is obtained by replacing $i-j$ with h and pulling one n^{-1} inside the summation.

  In the standard case of independent and identically distributed random variables $\mathrm{Var}(\bar{X}) = \sigma^2$. The condition $\gamma(n) \to 0$ is automatically satisfied. However, in the general case of weakly stationary processes, it cannot be omitted.

More can be proved using an appropriate set of assumptions. We only collect the results as a theorem without giving the proofs.

Theorem 2.1.1 *Let $(X_t)_{t\in\mathbb{Z}}$ be a weakly stationary stochastic process with mean μ and ACVF γ. Then, the following statements hold true as $n \to \infty$.*
(a) If $\sum_{h=-\infty}^{\infty}|\gamma(h)| < \infty$, then

$$n\mathrm{Var}(\bar{X}_n) \to \sum_{h=-\infty}^{\infty}\gamma(h) = \tau^2;$$

(b) If the process is "close to Gaussianity", then

$$\sqrt{n}(\bar{X}_n - \mu) \sim AN(0, \tau_n^2), \qquad \tau_n^2 = \sum_{h=-n}^{n}\left(1-\frac{|h|}{n}\right)\gamma(h).$$

Here, $\sim AN(0, \tau_n^2)$ stands for approximately normally distributed with mean zero and variance τ_n^2.

Theorem 2.1.1 can be utilized to construct confidence intervals for the unknown mean parameter μ. To do so, we must, however, estimate the unknown variance parameter τ_n. For a large class of stochastic processes, it holds that τ_n^2 converges to τ^2 as $n \to \infty$. Therefore, we can use τ^2 as an approximation for τ_n^2. Moreover, τ^2 can be estimated by

$$\hat{\tau}_n^2 = \sum_{h=-\sqrt{n}}^{\sqrt{n}}\left(1-\frac{|h|}{n}\right)\hat{\gamma}(h),$$

where $\hat{\gamma}(h)$ denotes the ACVF estimator defined in (1.2.1). An approximate 95% confidence interval for μ can now be constructed as

$$\left(\bar{X}_n - 1.96\frac{\hat{\tau}_n}{\sqrt{n}}, \bar{X}_n + 1.96\frac{\hat{\tau}_n}{\sqrt{n}}\right).$$

Example 2.1.1 (Autoregressive Processes) Let $(X_t)_{t\in\mathbb{Z}}$ be given by the equations

$$X_t - \mu = \phi(X_{t-1} - \mu) + Z_t, \qquad t \in \mathbb{Z}, \tag{2.1.1}$$

where $(Z_t)_{t\in\mathbb{Z}} \sim \mathrm{WN}(0, \sigma^2)$ and $|\phi| < 1$. We will see in Chapter 3 that $(X_t)_{t\in\mathbb{Z}}$ defines a weakly stationary process. Utilizing the stochastic difference equations (2.1.1), we can determine both mean and autocovariances. It holds that $E[X_t] = \phi E[X_{t-1}] + \mu(1-\phi)$. Since, by stationarity, $E[X_{t-1}]$ can be substituted with $E[X_t]$, we finally obtain that

$$E[X_t] = \mu, \qquad t \in \mathbb{Z}.$$

In the following we shall work with the process $(Y_t)_{t\in\mathbb{Z}}$ given by letting $Y_t = X_t - \mu$. Clearly, $E[Y_t] = 0$. From the definition, it follows also that the covariances of $(X_t)_{t\in\mathbb{Z}}$ and $(Y_t)_{t\in\mathbb{Z}}$ coincide. So let us first compute the second moment of Y_t^2: We have

$$E[Y_t^2] = E\big[(\phi Y_{t-1} + Z_t)^2\big] = \phi^2 E[Y_{t-1}^2] + \sigma^2$$

and consequently, since $E[Y_{t-1}^2] = E[Y_t^2]$ by weak stationarity of $(Y_t)_{t\in\mathbb{Z}}$,

$$E[Y_t^2] = \frac{\sigma^2}{1-\phi^2}, \qquad t \in \mathbb{Z}.$$

It becomes apparent from the latter equation, why the condition $|\phi| < 1$ was needed in display (2.1.1). In the next step, we compute the autocovariance function. For $h > 0$, it holds that

$$\gamma(h) = E[Y_{t+h}Y_t] = E\big[(\phi Y_{t+h-1} + Z_{t+h})Y_t\big] = \phi E[Y_{t+h-1}Y_t] = \phi\gamma(h-1) = \phi^h\gamma(0)$$

after h iterations. But since $\gamma(0) = E[Y_t^2]$, we obtain by symmetry of the ACVF that

$$\gamma(h) = \frac{\sigma^2\phi^{|h|}}{1-\phi^2}, \qquad h \in \mathbb{Z}.$$

After these theoretical considerations, we can now construct a 95% confidence interval for the mean parameter μ. To check if Theorem 2.1.1 is applicable here, we need to check if the autocovariances are absolutely summable:

$$\begin{aligned}\tau^2 &= \sum_{h=-\infty}^{\infty}\gamma(h) = \frac{\sigma^2}{1-\phi^2}\left(1 + 2\sum_{h=1}^{\infty}\phi^h\right) = \frac{\sigma^2}{1-\phi^2}\left(1 + \frac{2}{1-\phi} - 2\right)\\ &= \frac{\sigma^2}{1-\phi^2}\frac{1}{1-\phi}(1+\phi) = \frac{\sigma^2}{(1-\phi)^2} < \infty.\end{aligned}$$

Therefore, a 95% confidence interval for μ which is based on the observed values $x_1, \ldots, x_n$ is given by

$$\left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}(1-\phi)}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}(1-\phi)}\right).$$

Therein, the parameters σ and ϕ have to be replaced with appropriate estimators. These will be introduced in Chapter 3 below.

2.2 Estimation of the Autocovariance Function

In this section, we deal with the estimation of the ACVF and ACF at lag h. Recall from equation (1.2.1) that we can use the estimator

$$\hat{\gamma}(h) = \frac{1}{n} \sum_{t=1}^{n-|h|} (X_{t+|h|} - \bar{X}_n)(X_t - \bar{X}_n), \qquad h = 0, \pm 1, \ldots, \pm(n-1),$$

as a proxy for the unknown $\gamma(h)$. As estimator for the ACF $\rho(h)$, we have identified

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}, \qquad h = 0, \pm 1, \ldots, \pm(n-1).$$

We quickly collect some of the theoretical properties of $\hat{\rho}(h)$. They are not as obvious to derive as in the case of the sample mean, and we skip all proofs. Note also that similar statements hold for $\hat{\gamma}(h)$ as well.

- The estimator $\hat{\rho}(h)$ is generally biased, that is, $E[\hat{\rho}(h)] \neq \rho(h)$. It holds, however, under non-restrictive assumptions that

  $$E[\hat{\rho}(h)] \to \rho(h) \qquad (n \to \infty).$$

  This property is called *asymptotic unbiasedness*.

- The estimator $\hat{\rho}(h)$ is consistent for $\rho(h)$ under an appropriate set of assumptions, that is, $\mathrm{Var}(\hat{\rho}(h) - \rho(h)) \to 0$ as $n \to \infty$.

We have already established in Section 1.5 how the sample ACF $\hat{\rho}$ can be used to test if residuals consist of white noise variables. For more general statistical inference, we need to know the sampling distribution of $\hat{\rho}$. Since the estimation of $\rho(h)$ is based on only a few observations for h close to the sample size n, estimates tend to be unreliable. As a rule of thumb, given by Box and Jenkins (1976), n should at least be 50 and h less than or equal to $n/4$.

Theorem 2.2.1 *For $m \geq 1$, let $\boldsymbol{\rho}_m = (\rho(1), \ldots, \rho(m))^T$ and $\hat{\boldsymbol{\rho}}_m = (\hat{\rho}(1), \ldots, \hat{\rho}(m))^T$, where ^{T} denotes the transpose of a vector. Under a set of suitable assumptions, it holds that*

$$\sqrt{n}(\hat{\boldsymbol{\rho}}_m - \boldsymbol{\rho}_m) \sim AN(\mathbf{0}, \Sigma) \qquad (n \to \infty),$$

where $\sim AN(0, \Sigma)$ stands for approximately normally distributed with mean vector $\mathbf{0}$ and covariance matrix $\Sigma = (\sigma_{ij})$ given by Bartlett's formula

$$\sigma_{ij} = \sum_{k=1}^{\infty} \big[\rho(k+i) + \rho(k-i) - 2\rho(i)\rho(k)\big]\big[\rho(k+j) + \rho(k-j) - 2\rho(j)\rho(k)\big].$$

The section is concluded with two examples. The first one recollects the results already known for independent, identically distributed random variables, the second deals with the autoregressive process of Example 2.1.1.

Example 2.2.1 Let $(X_t)_{t\in\mathbb{Z}} \sim \text{IID}(0, \sigma^2)$. Then, $\rho(0) = 1$ and $\rho(h) = 0$ for all $h \neq 0$. The covariance matrix Σ is therefore given by

$$\sigma_{ij} = 1 \quad \text{if } i = j \qquad \textit{and} \qquad \sigma_{ij} = 0 \quad \text{if } i \neq j.$$

This means that Σ is a diagonal matrix. In view of Theorem 2.2.1 it holds thus that the estimators $\hat{\rho}(1), \ldots, \hat{\rho}(k)$ are approximately independent and identically distributed normal random variables with mean 0 and variance $1/n$. This was the basis for Methods 1 and 2 in Section 1.6 (see also Theorem 1.2.1).

Example 2.2.2 Let us reconsider the autoregressive process $(X_t)_{t\in\mathbb{Z}}$ from Example 2.1.1 with $\mu = 0$. Dividing $\gamma(h)$ by $\gamma(0)$ yields that

$$\rho(h) = \phi^{|h|}, \qquad h \in \mathbb{Z}.$$

We can now compute the diagonal entries of Σ as

$$\begin{aligned}
\sigma_{ii} &= \sum_{k=1}^{\infty} \big[\rho(k+i) + \rho(k-i) - 2\rho(i)\rho(k)\big]^2 \\
&= \sum_{k=1}^{i} \phi^{2i}(\phi^{-k} - \phi^k)^2 + \sum_{k=i+1}^{\infty} \phi^{2k}(\phi^{-i} - \phi^i)^2 \\
&= (1-\phi^{2i})(1+\phi^2)(1-\phi^2)^{-1} - 2i\phi^{2i}.
\end{aligned}$$

Chapter 3

ARMA Processes

3.1 Introduction

In this chapter we discuss *autoregressive moving average* processes, which play a crucial role in specifying time series models for applications. They are defined as the solutions of *stochastic difference equations* with constant coefficients and therefore possess a linear structure.

Definition 3.1.1 (ARMA processes) *(a) A weakly stationary process $(X_t)_{t\in\mathbb{Z}}$ is called an autoregressive moving average time series of order (p,q), abbreviated by ARMA(p,q), if it satisfies the difference equations*

$$X_t = \phi_1 X_{t-1} + \ldots + \phi_p X_{t-p} + Z_t + \theta_1 Z_{t-1} + \ldots + \theta_q Z_{t-q}, \qquad t \in \mathbb{Z}, \tag{3.1.1}$$

where $\phi_1, \ldots, \phi_p$ and $\theta_1, \ldots, \theta_q$ are real constants, $\phi_p \neq 0 \neq \theta_q$, and $(Z_t)_{t\in\mathbb{Z}} \sim \mathrm{WN}(0, \sigma^2)$.

(b) A weakly stationary stochastic process $(X_t)_{t\in\mathbb{Z}}$ is called an ARMA(p,q) time series with mean μ if the process $(X_t - \mu)_{t\in\mathbb{Z}}$ satisfies the equation system (3.1.1).

A more concise representation of (3.1.1) can be obtained with the use of the backshift operator B. To this end, we define the *autoregressive polynomial* and the *moving average polynomial* by

$$\phi(z) = 1 - \phi_1 z - \phi_2 z^2 - \ldots - \phi_p z^p, \qquad z \in \mathbb{C},$$

and

$$\theta(z) = 1 + \theta_1 z + \theta_2 z^2 + \ldots + \theta_q z^q, \qquad z \in \mathbb{C},$$

respectively, where $\mathbb{C}$ denotes the set of complex numbers. Inserting the backshift operator into these polynomials, the equations in (3.1.1) become

$$\phi(B)X_t = \theta(B)Z_t, \qquad t \in \mathbb{Z}. \tag{3.1.2}$$

Example 3.1.1 Figure 3.1 displays realizations of three different autoregressive moving average time series based on independent, standard normally distributed $(Z_t)_{t\in\mathbb{Z}}$. The left panel is an ARMA(2,2) process with parameter specifications $\phi_1 = .2$, $\phi_2 = -.3$, $\theta_1 = -.5$ and $\theta_2 = .3$. The middle plot is obtained from an ARMA(1,4) process with parameters $\phi_1 = .3$, $\theta_1 = -.2$, $\theta_2 = -.3$, $\theta_3 = .5$, and $\theta_4 = .2$, while the right plot is from

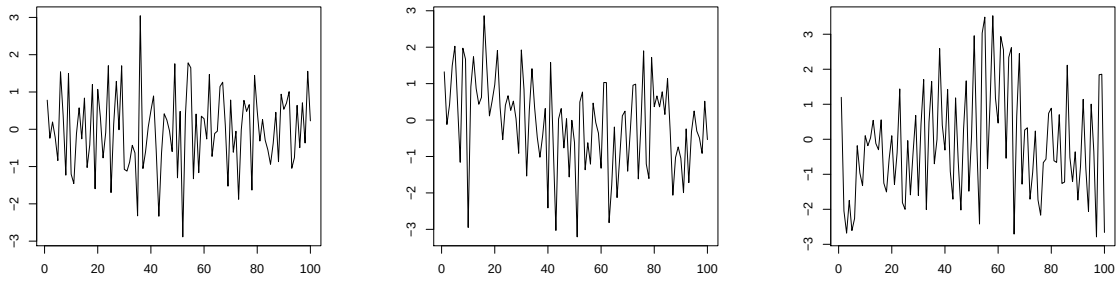

Figure 3.1: Realizations of three autoregressive moving average processes.

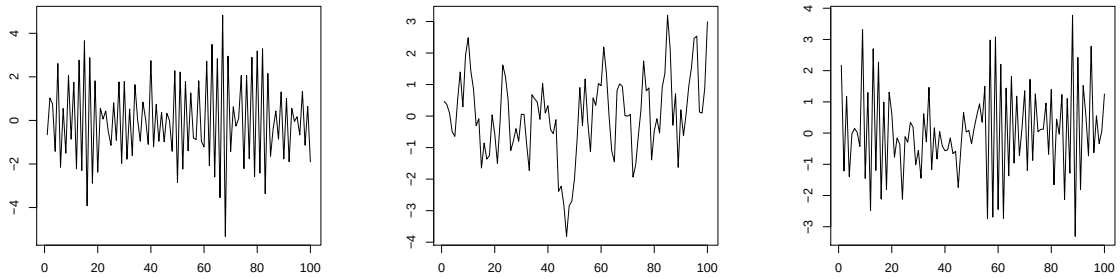

Figure 3.2: Realizations of three autoregressive processes.

an ARMA(4,1) with parameters $\phi_1 = -.2$, $\phi_2 = -.3$, $\phi_3 = .5$ and $\phi_4 = .2$ and $\theta_1 = .6$. The plots indicate that ARMA models can provide a flexible tool for modeling diverse residual sequences. We shall find out in the next section that all three realizations here come from (strictly) stationary processes. Similar time series plots can be produced in R using the commands

```
> arima22 =
     arima.sim(list(order=c(2,0,2), ar=c(.2,-.3), ma=c(-.5,.3)), n=100)
> arima14 =
     arima.sim(list(order=c(1,0,4), ar=.3, ma=c(-.2,-.3,.5,.2)), n=100)
> arima41 =
     arima.sim(list(order=c(4,0,1), ar=c(-.2,-.3,.5,.2), ma=.6), n=100)
```

Some special cases which we cover in the following two examples have particular relevance in time series analysis.

Example 3.1.2 (AR processes) If the moving average polynomial in (3.1.2) is equal to one, that is, if $\theta(z) \equiv 1$, then the resulting $(X_t)_{t\in\mathbb{Z}}$ is referred to as *autoregressive process of order p, AR(p)*. These time series interpret the value of the current variable X_t as a linear combination of p previous variables $X_{t-1}, \ldots, X_{t-p}$ plus an additional distortion by the white noise Z_t. Figure 3.2 displays two AR(1) processes with respective parameters $\phi_1 = -.9$ (left) and $\phi_1 = .8$ (middle) as well as an AR(2) process with parameters $\phi_1 = -.5$ and $\phi_2 = .3$. The corresponding R commands are

```
> ar1neg = arima.sim(list(order=c(1,0,0), ar=-.9), n=100)
```

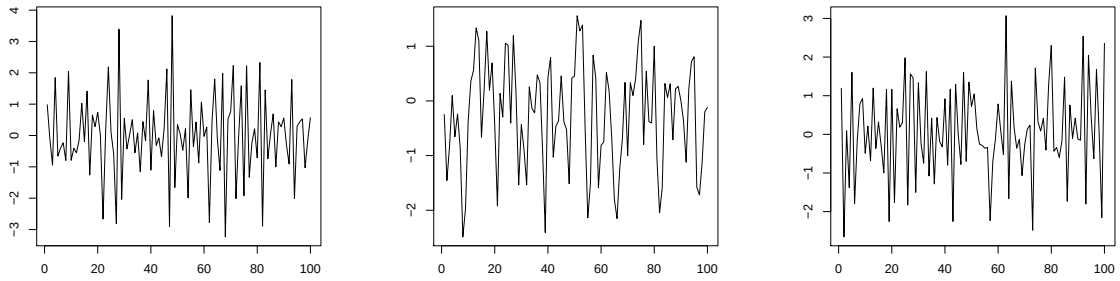

Figure 3.3: Realizations of three moving average processes.

```
> ar1pos = arima.sim(list(order=c(1,0,0), ar=.8), n=100)
> ar2 = arima.sim(list(order=c(2,0,0), ar=c(-.5,.3)), n=100)
```

Example 3.1.3 (MA processes) If the autoregressive polynomial in (3.1.2) is equal to one, that is, if $\phi(z) \equiv 1$, then the resulting $(X_t)_{t\in\mathbb{Z}}$ is referred to as *moving average process of order q, MA(q)*. Here the present variable X_t is obtained as superposition of q white noise terms $Z_t, \ldots, Z_{t-q}$. Figure 3.3 shows two MA(1) processes with respective parameters $\theta_1 = .5$ (left) and $\theta_1 = -.8$ (middle). The right plot is observed from an MA(2) process with parameters $\theta_1 = -.5$ and $\theta_2 = .3$. In R you may use

```
> ma1pos = arima.sim(list(order=c(0,0,1), ma=.5), n=100)
> ma1neg = arima.sim(list(order=c(0,0,1), ma=-.8), n=100)
> ma2 = arima.sim(list(order=c(0,0,2), ma=c(-.5,.3)), n=100)
```

For the analysis upcoming in the next chapters, we now introduce moving average processes of infinite order ($q = \infty$). They are an important tool for determining stationary solutions to the difference equations (3.1.1).

Definition 3.1.2 (Linear processes) *A stochastic process $(X_t)_{t\in\mathbb{Z}}$ is called linear process or MA(∞) time series if there is a sequence $(\psi_j)_{j\in\mathbb{N}_0}$ with $\sum_{j=0}^{\infty} |\psi_j| < \infty$ such that*

$$X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}, \qquad t \in \mathbb{Z}, \tag{3.1.3}$$

where $(Z_t)_{t\in\mathbb{Z}} \sim \mathrm{WN}(0, \sigma^2)$.

Moving average time series of any order q are special cases of linear processes. Just pick $\psi_j = \theta_j$ for $j = 1, \ldots, q$ and set $\psi_j = 0$ if $j > q$. It is common to introduce the power series

$$\psi(z) = \sum_{j=0}^{\infty} \psi_j z^j, \qquad z \in \mathbb{C},$$

to express a linear process in terms of the backshift operator. We can now rewrite display (3.1.3) in the form

$$X_t = \psi(B) Z_t, \qquad t \in \mathbb{Z}.$$

With the definitions of this section at hand, we shall investigate properties of ARMA processes such as stationarity and invertibility in the next section. We close the current section giving meaning to the notation $X_t = \psi(B)Z_t$. Note that we are possibly dealing with an *infinite* sum of random variables.

For completeness and later use, we derive in the following example the mean and ACVF of a linear process.

Example 3.1.4 (Mean and ACVF of a linear process) *Let $(X_t)_{t\in\mathbb{Z}}$ be a linear process according to Definition 3.1.2. Then, it holds that*

$$E[X_t] = E\left[\sum_{j=0}^{\infty} \psi_j Z_{t-j}\right] = \sum_{j=0}^{\infty} \psi_j E[Z_{t-j}] = 0, \qquad t \in \mathbb{Z}.$$

Next observe also that

$$\begin{aligned}\gamma(h) &= \mathrm{Cov}(X_{t+h}, X_t) \\ &= E\left[\sum_{j=0}^{\infty} \psi_j Z_{t+h-j} \sum_{k=0}^{\infty} \psi_k Z_{t-k}\right] \\ &= \sigma^2 \sum_{k=0}^{\infty} \psi_{k+h}\psi_k < \infty\end{aligned}$$

by assumption on the sequence $(\psi_j)_{j\in\mathbb{N}_0}$.

3.2 Causality and Invertibility

While a moving average process of order q will always be stationary without conditions on the coefficients $\theta_1, \ldots, \theta_q$, some deeper thoughts are required in the case of AR(p) and ARMA(p, q) processes. For simplicity, we start by investigating the autoregressive process of order one, which is given by the equations $X_t = \phi X_{t-1} + Z_t$ (writing $\phi = \phi_1$). Repeated iterations yield that

$$X_t = \phi X_{t-1} + Z_t = \phi^2 X_{t-2} + Z_t + \phi Z_{t-1} = \ldots = \phi^N X_{t-N} + \sum_{j=0}^{N-1} \phi^j Z_{t-j}.$$

Letting $N \to \infty$, it could now be shown that, with probability one,

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

is the weakly stationary solution to the AR(1) equations, provided that $|\phi| < 1$. These calculations would indicate moreover, that an autoregressive process of order one can be represented as linear process with coefficients $\psi_j = \phi^j$.

Example 3.2.1 (Mean and ACVF of an AR(1) process) Since we have identified an autoregressive process of order one as an example of a linear process, we can easily determine its expected value as

$$E[X_t] = \sum_{j=0}^{\infty} \phi^j E[Z_{t-j}] = 0, \qquad t \in \mathbb{Z}.$$

For the ACVF, we obtain that

$$\begin{aligned}\gamma(h) &= \mathrm{Cov}(X_{t+h}, X_t) \\ &= E\left[\sum_{j=0}^{\infty} \phi^j Z_{t+h-j} \sum_{k=0}^{\infty} \phi^k Z_{t-k}\right] \\ &= \sigma^2 \sum_{k=0}^{\infty} \phi^{k+h}\phi^k = \sigma^2\phi^h \sum_{k=0}^{\infty} \phi^{2k} = \frac{\sigma^2\phi^h}{1-\phi^2},\end{aligned}$$

where $h \geq 0$. This determines the ACVF for all h using that $\gamma(-h) = \gamma(h)$. It is also immediate that the ACF satisfies $\rho(h) = \phi^h$. See also Example 3.1.1 for comparison.

Example 3.2.2 (Nonstationary AR(1) processes) In Example 1.2.3 we have introduced the random walk as a nonstationary time series. It can also be viewed as a nonstationary AR(1) process with parameter $\phi = 1$. In general, autoregressive processes of order one with coefficients $|\phi| > 1$ are called *explosive* for they do not admit a weakly stationary solution that could be expressed as a linear process. However, one may proceed as follows. Rewrite the defining equations of an AR(1) process as

$$X_t = -\phi^{-1} Z_{t+1} + \phi^{-1} X_{t+1}, \qquad t \in \mathbb{Z}.$$

Apply now the same iterations as before to arrive at

$$X_t = \phi^{-N} X_{t+N} - \sum_{j=1}^{N} \phi^{-j} Z_{t+j}, \qquad t \in \mathbb{Z}.$$

Note that in the weakly stationary case, the present observation has been described in terms of past innovations. The representation in the last equation however contains only future observations with time lags larger than the present time t. From a statistical point of view this does not make much sense, even though by identical arguments as above we may obtain

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

as the weakly stationary solution in the explosive case.

The result of the previous example leads to the notion of causality which means that the process $(X_t)_{t\in\mathbb{Z}}$ has a representation in terms of the white noise $(Z_s)_{s\leq t}$ and that is hence independent of the future as given by $(Z_s)_{s>t}$. We give the definition for the general ARMA case.

Definition 3.2.1 (Causality) *An ARMA(p, q) process given by (3.1.1) is causal if there is a sequence* $(\psi_j)_{j\in\mathbb{N}_0}$ *such that* $\sum_{j=0}^{\infty}|\psi_j|<\infty$ *and*

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

Causality means that an ARMA time series can be represented as a linear process. We have seen earlier in this section how an AR(1) process whose coefficient satisfies the condition $|\phi|<1$ can be converted into a linear process. We have also seen that this is impossible if $|\phi|>1$. The conditions on the autoregressive parameter ϕ can be restated in terms of the corresponding autoregressive polynomial $\phi(z)=1-\phi z$ as follows. It holds that

$$|\phi|<1 \quad \text{if and only if} \quad \phi(z)\neq 0 \ \text{ for all } |z|\leq 1,$$
$$|\phi|>1 \quad \text{if and only if} \quad \phi(z)\neq 0 \ \text{ for all } |z|\geq 1.$$

It turns out that the characterization in terms of the zeroes of the autoregressive polynomials carries over from the AR(1) case to the general ARMA(p, q) case. Moreover, the ψ-weights of the resulting linear process have an easy representation in terms of the polynomials $\phi(z)$ and $\theta(z)$.The result is summarized in the next theorem.

Theorem 3.2.1 *Let* $(X_t)_{t\in\mathbb{Z}}$ *be an ARMA(p, q) process such that the polynomials* $\phi(z)$ *and* $\theta(z)$ *have no common zeroes. Then* $(X_t)_{t\in\mathbb{Z}}$ *is causal if and only if* $\phi(z)\neq 0$ *for all* $z\in\mathbb{C}$ *with* $|z|\leq 1$. *The coefficients* $(\psi_j)_{j\in\mathbb{N}_0}$ *are determined by the power series expansion*

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

A concept closely related to causality is *invertibility*. We motivate this notion with the following example that studies properties of a moving average time series of order 1.

Example 3.2.3 Let $(X_t)_{t\in\mathbb{N}}$ be an MA(1) process with parameter $\theta=\theta_1$. It is an easy exercise to compute the ACVF and the ACF as

$$\gamma(h)=\begin{cases}(1+\theta^2)\sigma^2, & h=0,\\ \theta\sigma^2, & h=1\\ 0 & h>1,\end{cases}\qquad \rho(h)=\begin{cases}1 & h=0.\\ \theta(1+\theta^2)^{-1}, & h=1.\\ 0 & h>1.\end{cases}$$

These results lead to the conclusion that $\rho(h)$ does not change if the parameter θ is replaced with θ^{-1}. Moreover, there exist pairs (θ,σ^2) that lead to the same ACVF, for example $(5, 1)$ and $(1/5, 25)$. Consequently, we arrive at the fact that the two MA(1) models

$$X_t=Z_t+\frac{1}{5}Z_{t-1},\qquad t\in\mathbb{Z},\qquad (Z_t)_{t\in\mathbb{Z}}\sim \text{iid}\,\mathcal{N}(0,25),$$

and

$$X_t=\tilde{Z}_t+5\tilde{Z}_{t-1},\qquad t\in\mathbb{Z},\qquad (\tilde{Z})_{t\in\mathbb{Z}}\sim \text{iid}\,\mathcal{N}(0,1),$$

are indistinguishable because we only observe X_t but not the noise variables Z_t and $\tilde{Z}_t$.

For convenience, the statistician will pick the model which satisfies the invertibility criterion which is to be defined next. It specifies that the noise sequence can be represented as a linear process in the observations.

Definition 3.2.2 (Invertibility) *An ARMA(p, q) process given by (3.1.1) is invertible if there is a sequence* $(\pi_j)_{j\in\mathbb{N}_0}$ *such that* $\sum_{j=0}^{\infty}|\pi_j| < \infty$ *and*

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

Theorem 3.2.2 *Let* $(X_t)_{t\in\mathbb{Z}}$ *be an ARMA(p, q) process such that the polynomials* $\phi(z)$ *and* $\theta(z)$ *have no common zeroes. Then* $(X_t)_{t\in\mathbb{Z}}$ *is invertible if and only if* $\theta(z) \neq 0$ *for all* $z \in \mathbb{C}$ *with* $|z| \leq 1$. *The coefficients* $(\pi_j)_{j\in\mathbb{N}_0}$ *are determined by the power series expansion*

$$\pi(z) = \sum_{j=0}^{\infty} \pi_j z^j = \frac{\phi(z)}{\theta(z)}, \qquad |z| \leq 1.$$

From now on we assume that all ARMA sequences specified in the sequel are causal and invertible ones if not explicitly stated otherwise. The final example of this section highlights the usefulness of the theory we have established. It deals with parameter redundancy and the calculation of the causality and invertibility sequences $(\psi_j)_{j\in\mathbb{N}_0}$ and $(\pi_j)_{j\in\mathbb{N}_0}$.

Example 3.2.4 (Parameter redundancy) Consider the ARMA equations

$$X_t = .4X_{t-1} + .21X_{t-2} + Z_t + .6Z_{t-1} + .09Z_{t-2},$$

which seem to generate an ARMA(2,2) sequence. However, the autoregressive and moving average polynomials have a common zero:

$$\tilde{\phi}(z) = 1 - .4z - .21z^2 = (1 - .7z)(1 + .3z),$$
$$\tilde{\theta}(z) = 1 + .6z + .09z^2 = (1 + .3z)^2.$$

Therefore, we can reset the ARMA equations to a sequence of order (1,1) and obtain

$$X_t = .7X_{t-1} + Z_t + .3Z_{t-1}.$$

Now, the corresponding polynomials have no common roots. Note that the roots of $\phi(z) = 1 - .7z$ and $\theta(z) = 1 + .3z$ are $10/7 > 1$ and $-10/3 < -1$, respectively. Thus Theorems 3.2.1 and 3.2.2 imply that causal and invertible solutions exist. In the following, we are going to calculate the corresponding coefficients in the expansions

$$X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j} \qquad \text{and} \qquad Z_t = \sum_{j=0}^{\infty} \pi_j X_{t-j}, \qquad t \in \mathbb{Z}.$$

We start with the causality sequence $(\psi_j)_{j\in\mathbb{N}_0}$. Writing, for $|z| \leq 1$,

$$\sum_{j=0}^{\infty} \psi_j z^j = \psi(z) = \frac{\theta(z)}{\phi(z)} = \frac{1+.3z}{1-.7z} = (1+.3z)\sum_{j=0}^{\infty}(.7z)^j,$$

it can be obtained from a comparison of coefficients that

$$\psi_0 = 1 \qquad \text{and} \qquad \psi_j = (.7+.3)(.7)^{j-1} = (.7)^{j-1}, \qquad j \in \mathbb{N}.$$

Similarly one computes the invertibility coefficients $(\pi_j)_{j\in\mathbb{N}_0}$ from the equation

$$\sum_{j=0}^{\infty} \pi_j z^j = \pi(z) = \frac{\phi(z)}{\theta(z)} = \frac{1-.7z}{1+.3z} = (1-.7z)\sum_{j=0}^{\infty}(-.3z)^j$$

($|z| \leq 1$) as

$$\pi_0 = 1 \qquad \text{and} \qquad \pi_j = (-1)^j(.3+.7)(.3)^{j-1} = (-1)^j(.3)^{j-1}.$$

Together, the previous calculations yield to the explicit representations

$$X_t = Z_t + \sum_{j=1}^{\infty}(.7)^{j-1}Z_{t-j} \qquad \text{and} \qquad Z_t = X_t + \sum_{j=1}^{\infty}(-1)^j(.3)^{j-1}X_{t-j}.$$

In the remainder of this section, we provide a general way to determine the weights $(\psi_j)_{j\geq 1}$ for a causal ARMA(p,q) process given by $\phi(B)X_t = \theta(B)Z_t$, where $\phi(z) \neq 0$ for all $z \in \mathbb{C}$ such that $|z| \leq 1$. Since $\psi(z) = \theta(z)/\phi(z)$ for these z, the weight ψ_j can be computed by matching the corresponding coefficients in the equation $\psi(z)\phi(z) = \theta(z)$, that is,

$$(\psi_0 + \psi_1 z + \psi_2 z^2 + \ldots)(1 - \phi_1 z - \ldots - \phi_p z^p) = 1 + \theta_1 z + \ldots + \theta_q z^q.$$

Recursively solving for $\psi_0, \psi_1, \psi_2, \ldots$ gives

$$\begin{aligned} \psi_0 &= 1, \\ \psi_1 - \phi_1\psi_0 &= \theta_1, \\ \psi_2 - \phi_1\psi_1 - \phi_2\psi_0 &= \theta_2, \end{aligned}$$

and so on as long as $j < \max\{p, q+1\}$. The general solution can be stated as

$$\psi_j - \sum_{k=1}^{j}\phi_k\psi_{j-k} = \theta_j, \qquad 0 \leq j < \max\{p, q+1\}, \tag{3.2.1}$$

$$\psi_j - \sum_{k=1}^{p}\phi_k\psi_{j-k} = 0, \qquad j \geq \max\{p, q+1\}, \tag{3.2.2}$$

if we define $\phi_j = 0$ if $j > p$ and $\theta_j = 0$ if $j > q$. To obtain the coefficients ψ_j one therefore has to solve the homogeneous linear difference equation (3.2.2) subject to the initial

1.0000000000	0.7000000000	0.4900000000	0.3430000000	0.2401000000
0.1680700000	0.1176490000	0.0823543000	0.0576480100	0.0403536070
0.0282475249	0.0197732674	0.0138412872	0.0096889010	0.0067822307
0.0047475615	0.0033232931	0.0023263051	0.0016284136	0.0011398895
0.0007979227	0.0005585459	0.0003909821	0.0002736875	0.0001915812

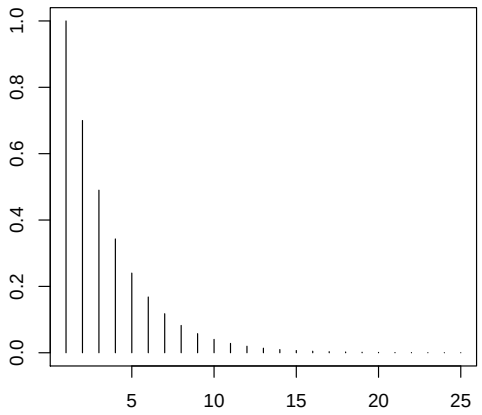

Figure 3.4: The R output for the ARMA(1,1) process of Example 3.2.4.

conditions specified by (3.2.1). For more on this subject, see Section 3.6 of Brockwell and Davis (1991) and Section 3.3 of Shumway and Stoffer (2006).

In R, these computations can be performed using the command `ARMAtoMA`. For example, you can use the commands

```
> ARMAtoMA(ar=.7,ma=.3,25)
> plot(ARMAtoMA(ar=.7,ma=.3,25))
```

which will produce the output displayed in Figure 3.4. The plot shows nicely the exponential decay of the ψ-weights which is typical for ARMA processes. The table shows row-wise the weights $\psi_0, \ldots, \psi_{24}$. This is enabled by the choice of `25` in the argument of the function `ARMAtoMA`.

3.3 The PACF of a causal ARMA Process

In this section, we introduce the *partial autocorrelation function (PACF)* to further assess the dependence structure of stationary processes in general and causal ARMA processes in particular. To start with, let us compute the ACVF of an moving average process of order q.

Example 3.3.1 (The ACVF of an MA(q) process) Let $(X_t)_{t\in\mathbb{Z}}$ be an MA(q) process specified by the polynomial $\theta(z) = 1 + \theta_1 z + \ldots + \theta_q z^q$. Then, letting $\theta_0 = 1$, it holds that

$$E[X_t] = \sum_{j=0}^{q} \theta_j E[Z_{t-j}] = 0.$$

To compute the ACVF, suppose that $h \geq 0$ and write

$$\begin{aligned}
\gamma(h) = \mathrm{Cov}(X_{t+h}, X_t) &= E[X_{t+h}X_t] \\
&= E\left[\left(\sum_{j=0}^{q} \theta_j Z_{t+h-j}\right)\left(\sum_{k=0}^{q} \theta_k Z_{t-k}\right)\right] \\
&= \sum_{j=0}^{q}\sum_{k=0}^{q} \theta_j\theta_k E[Z_{t+h-j}Z_{t-k}] \\
&= \begin{cases} \sigma^2 \sum_{k=0}^{q-h} \theta_{k+h}\theta_k, & 0 \leq h \leq q. \\ 0, & h > q. \end{cases}
\end{aligned}$$

The result here is a generalization of the MA(1) case, which was treated in Example 3.2.3. It is also a special case of the linear process in Example 3.1.4. The structure of the ACVF for MA processes indicates a possible strategy to determine in practice the unknown order q: plot the the sample ACF and select as order q the largest lag such that $\rho(h)$ is significantly different from zero.

While the sample ACF can potentially reveal the true order of an MA process, the same is not true anymore in the case of AR processes. Even for the AR(1) time series it has been shown in Example 3.2.1 that its ACF $\rho(h) = \phi^{|h|}$ is nonzero for all lags. As further motivation, however, we discuss the following example.

Example 3.3.2 Let $(X_t)_{t\in\mathbb{Z}}$ be a causal AR(1) process with parameter $|\phi| < 1$. It holds that

$$\gamma(2) = \mathrm{Cov}(X_2, X_0) = \mathrm{Cov}(\phi^2 X_0 + \phi Z_1 + Z_2, X_0) = \phi^2\gamma(0) \neq 0.$$

To break the linear dependence between X_0 and X_2, subtract ϕX_1 from both variables. Calcualting the resulting covariance yields

$$\mathrm{Cov}(X_2 - \phi X_1, X_0 - \phi X_1) = \mathrm{Cov}(Z_2, X_0 - \phi X_1) = 0,$$

since, due to the causality of this AR(1) process, $X_0 - \phi X_1$ is a function of $Z_1, Z_0, Z_{-1}, \ldots$ and therefore uncorrelated with $X_2 - \phi X_1 = Z_2$.

The previous example motivates the following general definition.

Definition 3.3.1 (Partial autocorrelation function) *Let $(X_t)_{t\in\mathbb{Z}}$ be a weakly stationary stochastic process with zero mean. Then, we call the sequence $(\phi_{hh})_{h\in\mathbb{N}}$ given by*

$$\phi_{11} = \rho(1) = \mathrm{Corr}(X_1, X_0),$$
$$\phi_{hh} = \mathrm{Corr}(X_h - X_h^{h-1}, X_0 - X_0^{h-1}), \qquad h \geq 2,$$

the partial autocorrelation function (PACF) of $(X_t)_{t\in\mathbb{Z}}$. Therein,

$$\begin{aligned} X_h^{h-1} &= \textit{regression of } X_h \textit{ on } (X_{h-1}, \ldots, X_1) \\ &= \beta_1 X_{h-1} + \beta_2 X_{h-2} + \ldots + \beta_{h-1} X_1 \end{aligned}$$

$$\begin{aligned} X_0^{h-1} &= \textit{regression of } X_0 \textit{ on } (X_1, \ldots, X_{h-1}) \\ &= \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_{h-1} X_{h-1}. \end{aligned}$$

Notice that there is no intercept coefficient β_0 in the regression parameters, since it is assumed that $E[X_t] = 0$. We demonstrate how to calculate the regression parameters in the case of an AR(1) process.

Example 3.3.3 (PACF of an AR(1) process) If $(X_t)_{t\in\mathbb{Z}}$ is a causal AR(1) process, then we have that $\phi_{11} = \rho(1) = \phi$. To calculate ϕ_{22}, we first calculate $X_2^1 = \beta X_1$, that is β. This coefficient is determined by minimizing the mean-squared error between X_2 and βX_1:

$$E[X_2 - \beta X_1]^2 = \gamma(0) - 2\beta\gamma(1) + \beta^2\gamma(0)$$

which is minimized by $\beta = \rho(1) = \phi$. (This follows easily by taking the derivative and setting it to zero.) Therefore $X_2^1 = \phi X_1$. Similarly, one computes $X_0^1 = \phi X_1$ and it follows from Example 3.3.2 that $\phi_{22} = 0$. Indeed all lags $h \geq 2$ of the PACF are zero.

More generally, let us briefly consider a causal AR(p) process given by $\phi(B)X_t = Z_t$ with $\phi(z) = 1 - \phi_1 z - \ldots - \phi_p z^p$. Then, for $h > p$,

$$X_h^{h-1} = \sum_{j=1}^{p} \phi_j X_{h-j}$$

and consequently

$$\phi_{hh} = \mathrm{Corr}(X_h - X_h^{h-1}, X_0 - X_0^{h-1}) = \mathrm{Corr}(Z_h, X_0 - X_0^{h-1}) = 0$$

if $h > p$ by causality (the same argument used in Example 3.3.2 applies here as well). Observe, however, that ϕ_{hh} is not necessarily zero if $h \leq p$. The forgoing suggests that the sample version of the PACF can be utilized to identify the order of an autoregressive process from data: use as p the largest lag h such that ϕ_{hh} is significantly different from zero.

On the other hand, for an invertible MA(q) process, we can write $Z_t = \pi(B)X_t$ or, equivalently,

$$X_t = -\sum_{j=1}^{\infty} \pi_j X_{t-j} + Z_t$$

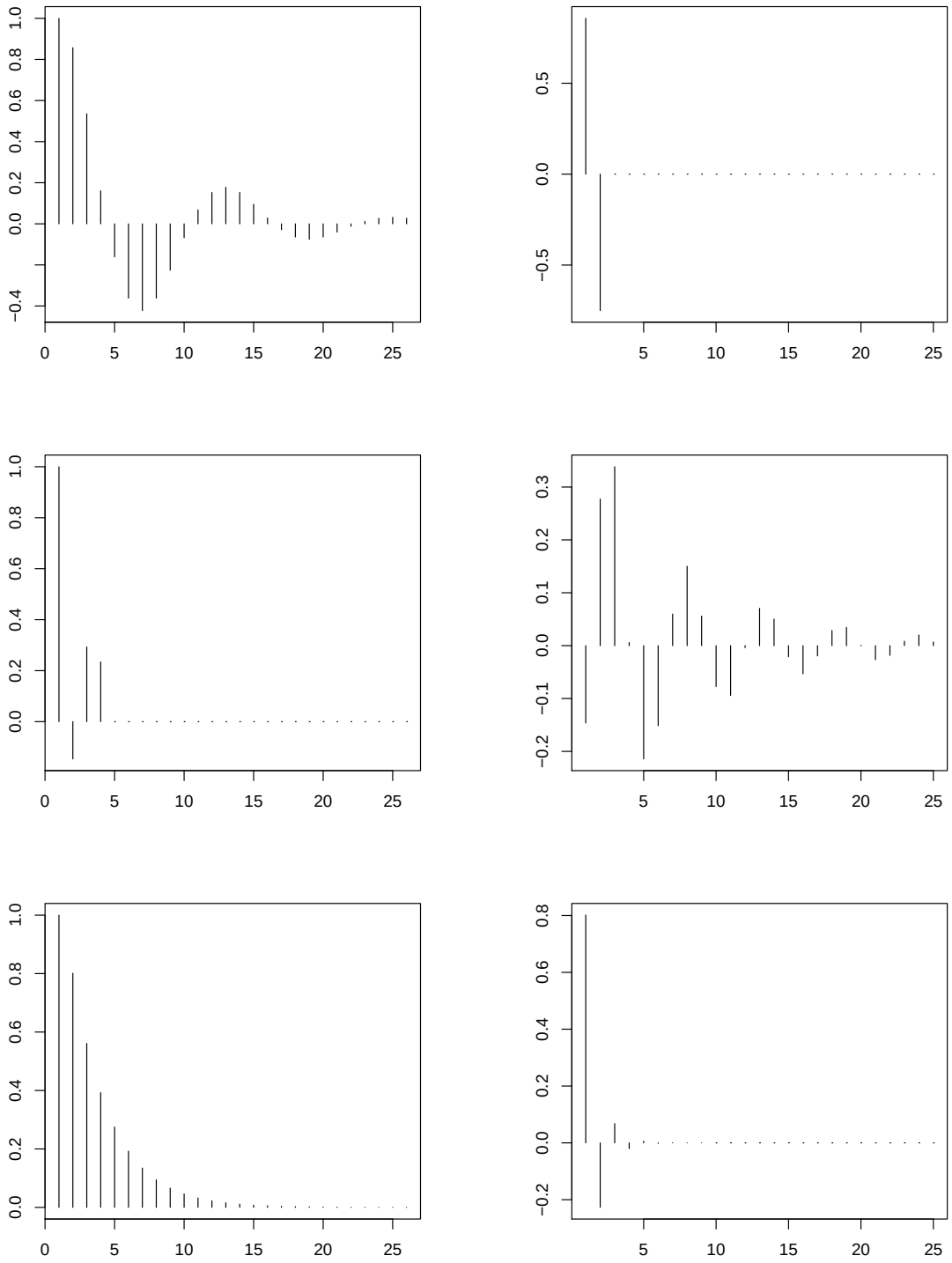

Figure 3.5: The ACFs and PACFs of an AR(2) process (upper panel), and MA(3) process (middle panel) and and ARMA(1,1) process (lower panel).

	AR(p)	MA(q)	ARMA(p, q)
ACF	tails off	cuts off after lag q	tails off
PACF	cuts off after lag p	tails off	tails off

Table 3.1: The behavior of ACF and PACF for AR, MA, and ARMA processes.

which shows that the PACF of an MA(q) process will be nonzero for all lags, since for a "perfect" regression one would have to use all past variables $(X_s)_{s<t}$ instead of only the quantity X_t^{t-1} given in Definition 3.3.1.

In summary, the PACF reverses the behavior of the ACVF for autoregressive and moving average processes. While the latter have an ACVF that vanishes after lag q and a PACF that is nonzero (though decaying) for all lags, AR processes have an ACVF that is nonzero (though decaying) for all lags but a PACF that vanishes after lag p.

ACVF (ACF) and PACF hence provide useful tools in assessing the dependence of given ARMA processes. If the estimated ACVF (the estimated PACF) is essentially zero after some time lag, then the underlying time series can be conveniently modeled with an MA (AR) process—and no general ARMA sequence has to be fitted. These conclusions are summarized in Table 3.1.

Example 3.3.4 Figure 3.5 collects the ACFs and PACFs of three ARMA processes. The upper panel is taken from the AR(2) process with parameters $\phi_1 = 1.5$ and $\phi_2 = -.75$. It can be seen that the ACF tails off and displays cyclical behavior (note that the corresponding autoregressive polynomial has complex roots). The PACF, however, cuts off after lag 2. Thus, inspecting ACF and PACF, we would correctly specify the order of the AR process.

The middle panel shows the ACF and PACF of the MA(3) process given by the parameters $\theta_1 = 1.5$, $\theta_2 = -.75$ and $\theta_3 = 3$. The plots confirm that $q = 3$ because the ACF cuts off after lag 3 and the PACF tails off.

Finally, the lower panel displays the ACF and PACF of the ARMA(1,1) process of Example 3.2.4. Here, the assessment is much harder. While the ACF tails off as predicted (see Table 3.1), the PACF basically cuts off after lag 4 or 5. This could lead to the wrong conclusion that the underlying process is actually an AR process of order 4 or 5. (The reason for this behavior lies in the fact that the dependence in this particular ARMA(1,1) process can be well approximated by that of an AR(4) or AR(5) time series.)

To reproduce the graphs in R, you can use the commands

```
> ar2.acf = ARMAacf(ar=c(1.5,-.75), ma=0, 25)
> ar2.pacf = ARMAacf(ar=c(1.5,-.75), ma=0, 25, pacf=T)
```

for the AR(2) process. The other two cases follow from straightforward adaptations of this code.

Example 3.3.5 (Recruitment Series) The data considered in this example consists of 453 months of observed recruitment (number of new fish) in a certain part of the Pacific

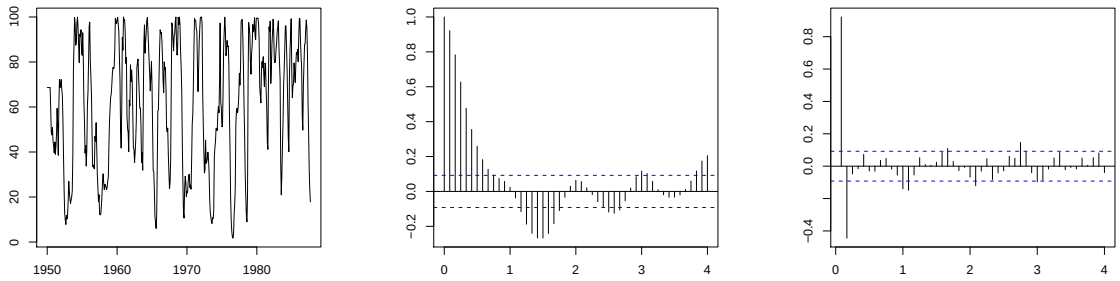

Figure 3.6: The recruitment series of Example 3.3.5 (left), its sample ACF (middle) and sample PACF (right).

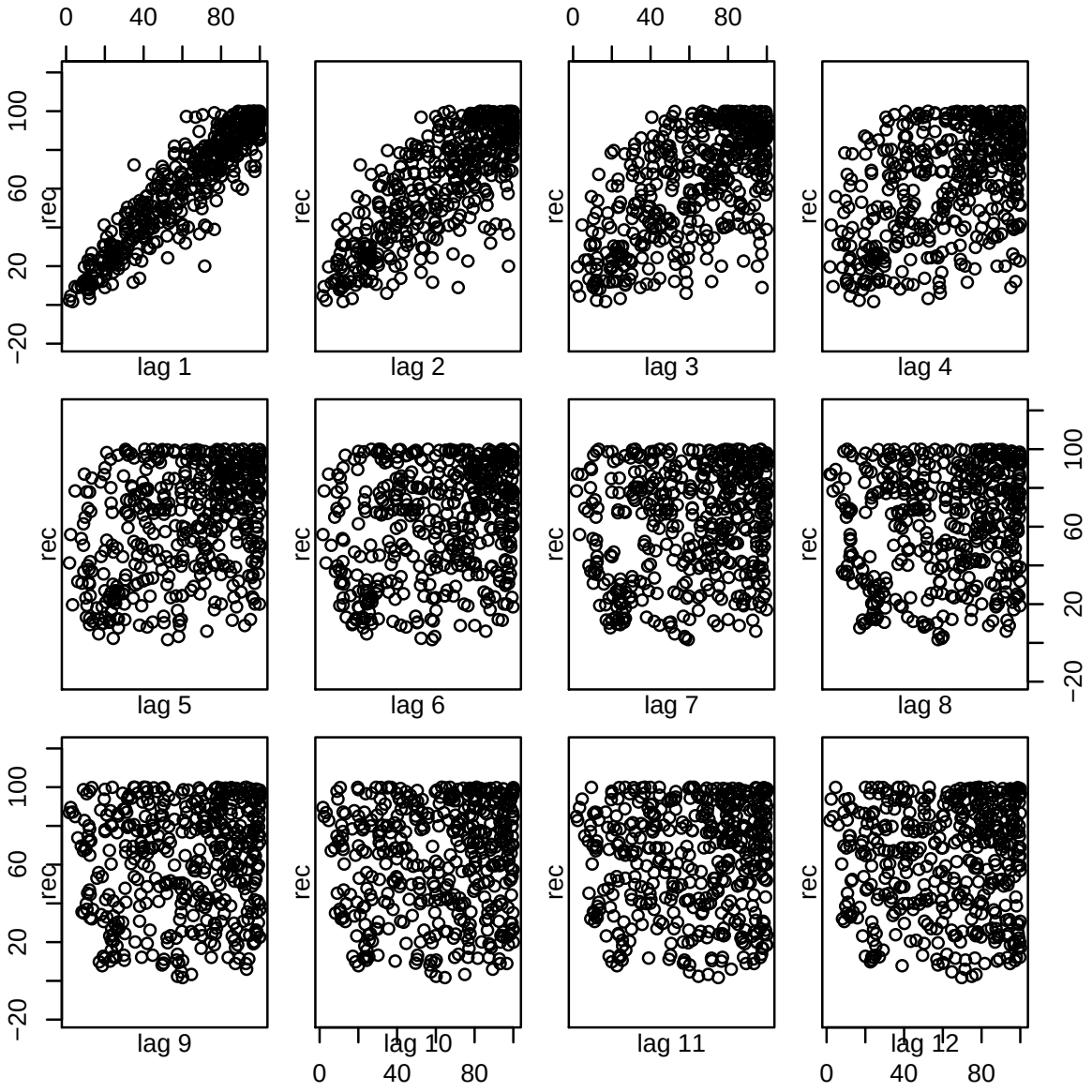

Figure 3.7: Scatterplot matrix relating current recruitment to past recruitment for the lags $h = 1, \ldots, 12$.

Ocean collected over the years 1950–1987. The corresponding time series plot is given in the left panel of Figure 3.6. The corresponding ACF and PACF displayed in the middle and right panel of the same figure recommend fitting an AR process of order $p = 2$ to the recruitment data. Assuming that the data is in `rec`, the R code to reproduce Figure 3.6 is

```
> rec = ts(rec, start=1950, frequency=12)
> plot(rec, xlab="", ylab="")
> acf(rec, lag=48)
> pacf(rec, lag=48)
```

This assertion is also consistent with the scatterplots that relate current recruitment to past recruitment at several time lags, namely $h = 1, \ldots, 12$. For lag 1 and 2, there seems to be a strong linear relationship, while this is not the case anymore for $h \geq 3$. The corresponding R commands are

```
> lag.plot(rec, lags=12, layout=c(3,4), diag=F)
```

Denote by X_t the recruitment at time t. To estimate the AR(2) parameters, you can run a regression on the observed data triplets included in the set $\{(x_t, x_{t-1}, x_{t-2})\colon j = 3, \ldots, 453\}$ to fit a model of the form

$$X_t = \phi_0 + \phi_1 X_{t-1} + \phi_2 X_{t-2} + Z_t, \qquad t = 3, \ldots, 453,$$

where $(Z_t) \sim \mathrm{WN}(0, \sigma^2)$. This task can be performed in R as follows.

```
> fit.rec = ar.ols(rec, aic=F, order.max=2, demean=F, intercept=T)
```

Now you can access the estimates typing `fit.rec` and the corresponding standard errors with `fit.rec$asy.se`. You will then obtain the parameter estimates $\hat{\phi}_0 = 6.737(1.111)$, $\hat{\phi}_1 = 1.3541(.042)$, $\hat{\phi}_2 = -.4632(.0412)$ and $\hat{\sigma}^2 = 89.72$. The standard errors are given in brackets.

3.4 Forecasting

Suppose that we have observed the variables $X_1, \ldots, X_n$ of a weakly stationary time series $(X_t)_{t\in\mathbb{Z}}$ and that our goal is to predict or forecast the future values of $X_{n+1}, X_{n+2}, \ldots$ based on this information. We shall focus here on so-called *one-step best linear predictors (BLP)*. These are, by definition, linear combinations

$$\hat{X}_{n+1} = \phi_{n0} + \phi_{n1} X_n + \ldots + \phi_{nn} X_1 \tag{3.4.1}$$

of the observed variables $X_1, \ldots, X_n$ that minimize the mean-squared error

$$E\left[\{X_{n+1} - g(X_1, \ldots, X_n)\}^2\right]$$

for functions g of $X_1, \ldots, X_n$. Straightforward generalizations yield definitions for the m-step best linear predictors $\hat{X}_{n+m}$ of X_{n+m} for arbitrary $m \in \mathbb{N}$ in the same fashion. Using Hilbert space theory, one can prove the following theorem which will be the starting point for our considerations.

Theorem 3.4.1 (Best linear prediction) *Let $(X_t)_{t\in\mathbb{Z}}$ be a weakly stationary stochastic process of which we observe $X_1, \ldots, X_n$. Then, the one-step BLP $\hat{X}_{n+1}$ of X_{n+1} is determined by the equations*

$$E\left[(X_{n+1} - \hat{X}_{n+1})X_{n+1-j}\right] = 0$$

for all $j = 1, \ldots, n+1$, where $X_0 = 1$.

The equations specified in Theorem 3.4.1 can be used to calculate the coefficients $\phi_{n0}, \ldots, \phi_{nn}$ in (3.4.1). We can focus on mean zero processes $(X_t)_{t\in\mathbb{Z}}$ and thus set $\phi_{n0} = 0$ as the following calculations show. Assume that $E[X_t] = \mu$ for all $t \in \mathbb{Z}$. Then, Theorem 3.4.1 gives that $E[\hat{X}_{n+1}] = E[X_{n+1}] = \mu$ (using the equation with $j = n+1$). Consequently, it holds that

$$\mu = E[\hat{X}_{n+1}] = E\left[\phi_{n0} + \sum_{\ell=1}^{n} \phi_{n\ell} X_{n+1-\ell}\right] = \phi_{n0} + \sum_{\ell=1}^{n} \phi_{n\ell}\mu.$$

Using now that $\phi_{n0} = \mu(1 - \phi_{n1} - \ldots - \phi_{nn})$, equation (3.4.1) can be rewritten as

$$\hat{Y}_{n+1} = \phi_{n1}Y_n + \ldots + \phi_{nn}Y_1,$$

where $\hat{Y}_{n+1} = \hat{X}_{n+1} - \mu$ has mean zero.

With the ACVF γ of $(X_t)_{t\in\mathbb{Z}}$, the equations in Theorem 3.4.1 can be expressed as

$$\sum_{\ell=1}^{n} \phi_{n\ell}\gamma(j-\ell) = \gamma(j), \qquad j = 1, \ldots, n. \tag{3.4.2}$$

Note that due to the convention $\phi_{n0} = 0$, the last equation in Theorem 3.4.1 (for which $j = n+1$) is omitted. More conveniently, this is restated in matrix notation. To this end, let $\boldsymbol{\Gamma}_n = (\gamma(j-\ell))_{j,\ell=1,\ldots,n}$, $\boldsymbol{\phi}_n = (\phi_{n1}, \ldots, \phi_{nn})^T$ and $\boldsymbol{\gamma}_n = (\gamma(1), \ldots, \gamma(n))^T$, where ^{T} denotes the transpose. With these notations, (3.4.2) becomes

$$\boldsymbol{\Gamma}_n\boldsymbol{\phi}_n = \boldsymbol{\gamma}_n \qquad \Longleftrightarrow \qquad \boldsymbol{\phi}_n = \boldsymbol{\Gamma}_n^{-1}\boldsymbol{\gamma}_n, \tag{3.4.3}$$

provided that $\boldsymbol{\Gamma}_n$ is nonsingular. The determination of the coefficients $\phi_{n\ell}$ has thus been reduced to solving a linear equation system and depends only on second-order properties of $(X_t)_{t\in\mathbb{Z}}$ which are given by the ACVF γ.

Let $\boldsymbol{X}_n = (X_n, X_{n-1}, \ldots, X_1)^T$. Then, $\hat{X}_{n+1} = \boldsymbol{\phi}_n^T\boldsymbol{X}_n$. To assess the quality of the prediction, one computes the mean-squared error with the help of (3.4.3) as follows:

$$\begin{aligned}
P_{n+1} &= E\left[(X_{n+1} - \hat{X}_{n+1})^2\right] \\
&= E\left[(X_{n+1} - \boldsymbol{\phi}_n^T\boldsymbol{X}_n)^2\right] \\
&= E\left[(X_{n+1} - \boldsymbol{\gamma}_n^T\boldsymbol{\Gamma}_n^{-1}\boldsymbol{X}_n)^2\right] \\
&= E\left[X_{n+1}^2 - 2\boldsymbol{\gamma}_n^T\boldsymbol{\Gamma}_n^{-1}\boldsymbol{X}_n X_{n+1} + \boldsymbol{\gamma}_n^T\boldsymbol{\Gamma}_n^{-1}\boldsymbol{X}_n\boldsymbol{X}_n^T\boldsymbol{\Gamma}_n^{-1}\boldsymbol{\gamma}_n\right] \\
&= \gamma(0) - 2\boldsymbol{\gamma}_n^T\boldsymbol{\Gamma}_n^{-1}\boldsymbol{\gamma}_n + \boldsymbol{\gamma}_n^T\boldsymbol{\Gamma}_n^{-1}\boldsymbol{\Gamma}_n\boldsymbol{\Gamma}_n^{-1}\boldsymbol{\gamma}_n \\
&= \gamma(0) - \boldsymbol{\gamma}_n^T\boldsymbol{\Gamma}_n^{-1}\boldsymbol{\gamma}_n.
\end{aligned} \tag{3.4.4}$$

As an initial example, we explain the prediction procedure for an autoregressive process of order 2.

Example 3.4.1 (Prediction of an AR(2) Process) Let $(X_t)_{t\in\mathbb{Z}}$ be the causal AR(2) process $X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + Z_t$. Suppose that you have only an observation of X_1 to forecast the value of X_2. In this simplified case, the single prediction equation (3.4.2) is

$$\phi_{11}\gamma(0) = \gamma(1),$$

so that $\phi_{11} = \rho(1)$ and $\hat{X}_{1+1} = \rho(1)X_1$. In the next step, assume that we have observed values of X_1 and X_2 at hand to forecast the value of X_3. Then, one similarly obtains from (3.4.2) that the predictor can be computed from

$$\begin{aligned}\hat{X}_{2+1} &= \phi_{21}X_2 + \phi_{22}X_1 = \boldsymbol{\phi}_2^T \boldsymbol{X}_2 = (\boldsymbol{\Gamma}_2^{-1}\boldsymbol{\gamma}_2)^T \boldsymbol{X}_2 \\ &= (\gamma(1), \gamma(2)) \begin{pmatrix} \gamma(0) & \gamma(1) \\ \gamma(1) & \gamma(0) \end{pmatrix}^{-1} \begin{pmatrix} X_2 \\ X_1 \end{pmatrix}.\end{aligned}$$

However, applying the arguments leading to the definition of the PACF in Section 3.3, one finds that

$$\begin{aligned}E\left[\{X_3 - (\phi_1 X_2 + \phi_2 X_1)\}X_1\right] &= E[Z_3 X_1] = 0, \\ E\left[\{X_3 - (\phi_1 X_2 + \phi_2 X_1)\}X_2\right] &= E[Z_3 X_2] = 0.\end{aligned}$$

Hence, $\hat{X}_{2+1} = \phi_1 X_2 + \phi_2 X_1$ and even $\hat{X}_{n+1} = \phi_1 X_n + \phi_2 X_{n-1}$ for all $n \geq 2$, exploiting the particular autoregressive structure. Since similar results can be proved for general causal AR(p) processes, the one-step predictors have the form

$$\hat{X}_{n+1} = \phi_1 X_n + \ldots + \phi_p X_{n-p+1}$$

whenever the number of observed variables n is at least p.

The major drawback of this approach is immediately apparent from the previous example: For larger sample sizes n, the prediction procedure requires the calculation of the inverse matrix $\boldsymbol{\Gamma}_n^{-1}$ which is computationally expensive. In the remainder of this section, we introduce two recursive prediction methods that bypass the inversion altogether. They are known as *Durbin-Levinson algorithm* and *innovations algorithm*. Finally, we deal with predictors based on the *infinite past* which are, in several cases, easily applicable for the class of causal and invertible ARMA processes.

Method 1 (The Durbin-Levinson algorithm) If $(X_t)_{t\in\mathbb{Z}}$ is a zero mean weakly stationary process with ACVF γ such that $\gamma(0) > 0$ and $\gamma(h) \to 0$ as $h \to \infty$, then the coefficients $\phi_{n\ell}$ in (3.4.2) and the mean squared errors P_n in (3.4.4) satisfy the recursions

$$\phi_{11} = \frac{\gamma(1)}{\gamma(0)}, \qquad P_0 = \gamma(0),$$

and, for $n \geq 1$,

$$\phi_{nn} = \frac{1}{P_{n-1}}\left(\gamma(n) - \sum_{\ell=1}^{n-1} \phi_{n-1,\ell}\gamma(n-\ell)\right),$$

$$\begin{pmatrix} \phi_{n1} \\ \vdots \\ \phi_{n,n-1} \end{pmatrix} = \begin{pmatrix} \phi_{n-1,1} \\ \vdots \\ \phi_{n-1,n-1} \end{pmatrix} - \phi_{nn} \begin{pmatrix} \phi_{n-1,n-1} \\ \vdots \\ \phi_{n-1,1} \end{pmatrix}$$

and

$$P_n = P_{n-1}(1 - \phi_{nn}^2).$$

It can be shown that under the assumptions made on the process $(X_t)_{t\in\mathbb{Z}}$, it holds indeed that ϕ_{nn} is equal to the value of the PACF of $(X_t)_{t\in\mathbb{Z}}$ at lag n. The result is formulated as Corollary 5.2.1 in Brockwell and Davis (1991). We highlight this fact in an example.

Example 3.4.2 (The PACF of an AR(2) process) Let $(X_t)_{t\in\mathbb{Z}}$ be a causal AR(2) process. Then, $\rho(1) = \phi_1/(1-\phi_2)$ and all other values can be computed recursively from

$$\rho(h) - \phi_1\rho(h-1) - \phi_2\rho(h-2) = 0, \qquad h \geq 2.$$

Note that the ACVF γ satisfies a difference equation with the same coefficients, which is easily seen by multiplying the latter equation with $\gamma(0)$. Applying the Durbin-Levinson algorithm gives first that

$$\phi_{11} = \frac{\gamma(1)}{\gamma(0)} = \rho(1) \qquad \text{and} \qquad P_1 = P_0(1-\phi_{11}^2) = \gamma(0)(1-\rho(1)^2).$$

Ignoring the recursion for the error terms P_n in the following, the next $\phi_{n\ell}$ values are obtained as

$$\begin{aligned} \phi_{22} &= \frac{1}{P_1}\left[\gamma(2) - \phi_{11}\gamma(1)\right] = \frac{1}{1-\rho(1)^2}\left[\rho(2) - \rho(1)^2\right] \\ &= \frac{\phi_1^2(1-\phi_2)^{-1} + \phi_2 - [\phi_1(1-\phi_2)^{-1}]^2}{1 - [\phi_1(1-\phi_2)^{-1}]^2} = \phi_2, \end{aligned}$$

$$\phi_{21} = \phi_{11} - \phi_{22}\phi_{11} = \rho(1)(1-\phi_2) = \phi_1,$$

$$\phi_{33} = \frac{1}{P_2}\left[\gamma(3) - \phi_{21}\gamma(2) - \phi_{22}\gamma(1)\right] = \frac{1}{P_2}\left[\gamma(3) - \phi_1\gamma(2) - \phi_2\gamma(2)\right] = 0.$$

Now, referring to the remarks after Example 3.3.3, no further computations are necessary to determine the PACF because $\phi_{nn} = 0$ for all $n > p = 2$.

Method 2 (The innovations algorithm) In contrast to the Durbin-Levinson algorithm, this method can also be applied to nonstationary processes. It should thus, in general, be preferred over Method 1. The innovations algorithm gets its name from the fact that one directly uses the form of the prediction equations in Theorem 3.4.1 which are stated in terms of the *innovations* $(X_{t+1} - \hat{X}_{t+1})_{t\in\mathbb{Z}}$. Observe that the sequence consists of uncorrelated random variables.

The one-step predictors $\hat{X}_{n+1}$ can be calculated from the recursions

$$\hat{X}_{0+1} = 0, \qquad P_1 = \gamma(0)$$

and, for $n \geq 1$,

$$\hat{X}_{n+1} = \sum_{\ell=1}^{n} \theta_{n\ell}(X_{n+1-\ell} - \hat{X}_{n+1-\ell})$$

$$P_{n+1} = \gamma(0) - \sum_{\ell=0}^{n-1} \theta_{n,n-\ell}^2 P_{\ell+1},$$

where the coefficients are obtained from the equations

$$\theta_{n,n-\ell} = \frac{1}{P_{\ell+1}} \left[\gamma(n-\ell) - \sum_{i=0}^{\ell-1} \theta_{\ell,\ell-i}\theta_{n,n-i}P_{i+1} \right], \qquad \ell = 0, 1, \ldots, n-1.$$

As example we show how the innovations algorithm is applied to a moving average time series of order 1.

Example 3.4.3 (Prediction of an MA(1) Process) Let $(X_t)_{t\in\mathbb{Z}}$ be the MA(1) process $X_t = Z_t + \theta Z_{t-1}$. Note that

$$\gamma(0) = (1+\theta^2)\sigma^2, \qquad \gamma(1) = \theta\sigma^2 \qquad \text{and} \qquad \gamma(h) = 0 \quad (h \geq 2).$$

Using the innovations algorithm, we can compute the one-step predictor from the values

$$\theta_{n1} = \frac{\theta\sigma^2}{P_n}, \qquad \theta_{n\ell} = 0 \quad (\ell = 2, \ldots, n-1),$$

and

$$P_1 = (1+\theta^2)\sigma^2,$$

$$P_{n+1} = (1 + \theta^2 - \theta\theta_{n1})\sigma^2$$

as

$$\hat{X}_{n+1} = \frac{\theta\sigma^2}{P_n}(X_n - \hat{X}_n).$$

Method 3 (Prediction based on the infinite past) Suppose that we are analyzing a causal and invertible ARMA(p, q) process. Assume further that we have the (unrealistic) ability to store the history of the process and that we can thus access all past variables $(X_t)_{t\leq n}$. Define now

$$\tilde{X}_{n+m} = E[X_{n+m}|X_n, X_{n-1}, \ldots],$$

as the m-step ahead predictor based on the infinite past. It can be shown that, for large sample sizes n, the difference between the values of $\hat{X}_{n+m}$ and $\tilde{X}_{n+m}$ vanishes at an exponential rate. Exploiting causality and invertibility of the ARMA process, one can transform the predictor $\tilde{X}_{n+m}$ so that it is in a computationally more feasible form. To do so, note that by causality

$$\tilde{X}_{n+m} = E[X_{n+m}|X_n, X_{n-1}, \ldots]$$

$$= E\left[\sum_{j=0}^{\infty} \psi_j Z_{n+m-j} \Big| X_n, X_{n-1}, \ldots\right]$$

$$= \sum_{j=m}^{\infty} \psi_j Z_{n+m-j} \tag{3.4.5}$$

because $E[Z_t|X_n, X_{n-1}, \ldots]$ equals zero if $t > n$ and equals Z_t if $t \leq n$ (due to invertibility!). The representation in (3.4.5) can be used to compute the mean squared prediction error $\tilde{P}_{n+m}$. Using causality, we obtain that

$$\tilde{P}_{n+m} = E[(X_{n+m} - \tilde{X}_{n+m})^2] = E\left[\left(\sum_{j=0}^{m-1} \psi_j Z_{n+m-j}\right)^2\right] = \sigma^2 \sum_{j=0}^{m-1} \psi_j^2. \tag{3.4.6}$$

On the other hand, (3.4.5) does not allow to directly calculate the forecasts because $\tilde{X}_{n+m}$ is given in terms of the noise variables Z_{n+m-j}. Instead we will use invertibility. Observe first that

$$E[X_{n+m-j}|X_n, X_{n-1}, \ldots] = \begin{cases} \tilde{X}_{n+m-j}, & j < m. \\ X_{n+m-j}, & j \geq m. \end{cases}$$

By invertibility (the "0 =" part follows again from causality),

$$0 = E[Z_{n+m}|X_n, X_{n-1}, \ldots]$$

$$= E\left[\sum_{j=0}^{\infty} \pi_j X_{n+m-j} \Big| X_n, X_{n-1}, \ldots\right]$$

$$= \sum_{j=0}^{\infty} \pi_j E[X_{n+m-j}|X_n, X_{n-1}, \ldots].$$

Combining the previous two statements, we arrive at

$$\tilde{X}_{n+m} = -\sum_{j=1}^{m-1} \pi_j \tilde{X}_{n+m-j} - \sum_{j=m}^{\infty} \pi_j X_{n+m-j}. \tag{3.4.7}$$

The equations can now be solved recursively for $m = 1, 2, \ldots$ Note, however, that for any $m \geq 1$ the sequence $(X_{n+m+t} - \tilde{X}_{n+m+t})_{t\in\mathbb{Z}}$ does not consist of uncorrelated random variables. In fact, if $h \in \mathbb{N}_0$, it holds that

$$E[(X_{n+m} - \tilde{X}_{n+m})(X_{n+m+h} - \tilde{X}_{n+m+h})]$$

$$= E\left[\sum_{j=0}^{m-1} \psi_j Z_{n+m-j} \sum_{i=0}^{m+h-1} \psi_i Z_{n+m+h-i}\right]$$

$$= \sigma^2 \sum_{j=0}^{m-1} \psi_j \psi_{j+h}.$$

Finally, for practical purposes we need to truncate the given forecast for large n. This is accomplished by setting

$$\sum_{j=n+m}^{\infty} \pi_j X_{n+m-j} = 0.$$

The resulting equations (see (3.4.7) for comparison) yield recursively the truncated m-step predictors X^*_{n+m}:

$$X^*_{n+m} = -\sum_{j=1}^{m-1} \pi_j X^*_{n+m-j} - \sum_{j=m}^{n+m-1} \pi_j X_{n+m-j}. \tag{3.4.8}$$

3.5 Parameter Estimation

Let $(X_t)_{t\in\mathbb{Z}}$ be a causal and invertible ARMA(p,q) process with known orders p and q, possibly with mean μ. We are in this section concerned with estimation procedures for the unknown parameter vector

$$\boldsymbol{\beta} = (\mu, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q, \sigma^2)^T. \tag{3.5.1}$$

To simplify the estimation procedure, we assume that we can work with data that has been adjusted by subtraction of the mean and we can restrict the discussion to zero mean ARMA models.

In the following, we shall introduce three methods of estimation. The method of moments works best in case of pure AR processes, while it does not lead to optimal estimation procedures for general ARMA processes. For the latter, more efficient estimators are provided by the maximum likelihood and least squares methods which will be discussed subsequently.

Method 1 (Method of Moments) Since this method is only efficient in their case, we restrict the presentation here to AR(p) processes

$$X_t = \phi_1 X_{t-1} + \ldots + \phi_p X_{t-p} + Z_t, \qquad t \in \mathbb{Z},$$

where $(Z_t)_{t\in\mathbb{Z}} \sim \mathrm{WN}(0,\sigma^2)$. The parameter vector $\boldsymbol{\beta}$ consequently reduces to $(\boldsymbol{\phi}, \sigma^2)^T$ with $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)^T$ and can be estimated using the *Yule-Walker equations*

$$\boldsymbol{\Gamma}_p \boldsymbol{\phi} = \boldsymbol{\gamma}_p \qquad \text{and} \qquad \sigma^2 = \gamma(0) - \boldsymbol{\phi}^T \boldsymbol{\gamma}_p,$$

where $\boldsymbol{\Gamma}_p = (\boldsymbol{\gamma}(k-j))_{k,j=1,\ldots,p}$ and $\boldsymbol{\gamma}_p = (\gamma(1), \ldots, \gamma(p))^T$. Observe that the equations are obtained by the same arguments applied to derive the Durbin-Levinson algorithm in the previous section. The method of moments suggests to replace every quantity in the Yule-Walker equations with their estimated counterparts, which yields the *Yule-Walker estimators*

$$\widehat{\boldsymbol{\phi}} = \hat{\boldsymbol{\Gamma}}_p^{-1} \hat{\boldsymbol{\gamma}}_p = \hat{\boldsymbol{R}}_p^{-1} \hat{\boldsymbol{\rho}}_p \tag{3.5.2}$$

$$\hat{\sigma}^2 = \hat{\gamma}(0) - \hat{\boldsymbol{\gamma}}_p^T \hat{\boldsymbol{\Gamma}}_p^{-1} \hat{\boldsymbol{\gamma}}_p = \hat{\gamma}(0) \left[1 - \hat{\boldsymbol{\rho}}_p^T \hat{\boldsymbol{R}}_p^{-1} \hat{\boldsymbol{\rho}}_p\right]. \tag{3.5.3}$$

Therein, $\hat{\boldsymbol{R}}_p = \hat{\gamma}(0)^{-1}\hat{\boldsymbol{\Gamma}}_p$ and $\hat{\boldsymbol{\rho}}_p = \hat{\gamma}(0)^{-1}\hat{\boldsymbol{\gamma}}_p$ with $\hat{\gamma}(h)$ defined as in (1.2.1). Using $\hat{\gamma}(h)$ as estimator for the ACVF at lag h, we implicitly obtain a dependence on the sample size n. This dependence is suppressed in the notation used here. The following theorem contains the limit behavior of the Yule-Walker estimators as n tends to infinity.

Theorem 3.5.1 *If $(X_t)_{t\in\mathbb{Z}}$ is a causal AR(p) process, then*

$$\sqrt{n}(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}) \xrightarrow{\mathcal{D}} N(\mathbf{0}, \sigma^2\boldsymbol{\Gamma}_p^{-1}) \qquad \text{and} \qquad \hat{\sigma}^2 \xrightarrow{P} \sigma^2$$

as $n \to \infty$, where $\to^P$ indicates convergence in probability.

A proof of this result is given in Section 8.10 of Brockwell and Davis (1991). Since equations (3.5.2) and (3.5.3) have the same structure as the corresponding equations (3.4.3) and (3.4.4), the Durbin-Levinson algorithm can be used to solve recursively for the estimators $\widehat{\boldsymbol{\phi}}_h = (\widehat{\phi}_{h1}, \ldots, \widehat{\phi}_{hh})$. Moreover, since ϕ_{hh} is equal to the value of the PACF of $(X_t)_{t\in\mathbb{Z}}$ at lag h, the estimator $\widehat{\phi}_{hh}$ can be used as its proxy. Since we already know that, in the case of AR(p) processes, $\phi_{hh} = 0$ if $h > p$, Theorem 3.5.1 implies immediately the following corollary.

Corollary 3.5.1 *If $(X_t)_{t\in\mathbb{Z}}$ is a causal AR(p) process, then*

$$\sqrt{n}\widehat{\phi}_{hh} \xrightarrow{\mathcal{D}} Z \qquad (n \to \infty)$$

for all $h > p$, where Z stands for a standard normal random variable.

Example 3.5.1 (Yule-Walker estimates for AR(2) processes) Suppose that we have observed $n = 144$ values of the autoregressive process $X_t = 1.5X_{t-1} - .75X_{t-2} + Z_t$, where $(Z_t)_{t\in\mathbb{Z}}$ is a sequence of independent standard normal variates. Assume further that $\hat{\gamma}(0) = 8.434$, $\hat{\rho}(1) = 0.834$ and $\hat{\rho}(2) = 0.476$ have been calculated from the data. The Yule-Walker estimators for the parameters are then given by

$$\widehat{\boldsymbol{\phi}} = \begin{pmatrix} \widehat{\phi}_1 \\ \widehat{\phi}_2 \end{pmatrix} = \begin{pmatrix} 1.000 & 0.834 \\ 0.834 & 1.000 \end{pmatrix}^{-1} \begin{pmatrix} 0.834 \\ 0.476 \end{pmatrix} = \begin{pmatrix} 1.439 \\ -0.725 \end{pmatrix}$$

and

$$\hat{\sigma}^2 = 8.434\left[1 - (0.834, 0.476)\begin{pmatrix} 1.439 \\ -0.725 \end{pmatrix}\right] = 1.215.$$

To construct asymptotic confidence intervals using Theorem 3.5.1, the unknown limiting covariance matrix $\sigma^2\boldsymbol{\Gamma}_p^{-1}$ needs to be estimated. This can be done using the estimator

$$\frac{\hat{\sigma}^2\hat{\boldsymbol{\Gamma}}_p^{-1}}{n} = \frac{1}{144}\frac{1.215}{8.434}\begin{pmatrix} 1.000 & 0.834 \\ 0.834 & 1.000 \end{pmatrix}^{-1} = \begin{pmatrix} 0.057^2 & -0.003 \\ -0.003 & 0.057^2 \end{pmatrix}.$$

Then, the $1-\alpha$ level confidence interval for the parameters ϕ_1 and ϕ_2 are computed as

$$1.439 \pm 0.057 z_{1-\alpha/2} \qquad \text{and} \qquad -0.725 \pm 0.057 z_{1-\alpha/2},$$

respectively, where $z_{1-\alpha/2}$ is the corresponding normal quantile.

Example 3.5.2 (Recruitment Series) Let us reconsider the recruitment series of Example 3.3.5. There, we have first established an AR(2) model as appropriate for the data and then estimated the model parameters using an ordinary least squares approach. Here, we will instead estimate the coefficients with the Yule-Walker procedure. The R command is

```
> rec.yw = ar.yw(rec, order=2)
```

The mean estimate can be obtained from `rec.yw$x.mean` as $\hat{\mu} = 62.26$, while the autoregressive parameter estimates and their standard errors are accessed with `rec.yw$ar` and `sqrt(rec.yw$asy.var.coef)` as $\hat{\phi}_1 = 1.3316(.0422)$ and $\hat{\phi}_2 = -.4445(.0422)$. Finally, the variance estimate is obtained from `rec.yw$var.pred` as $\hat{\sigma}^2 = 94.7991$. All values are close to their counterparts in Example 3.3.5.

Example 3.5.3 Consider the invertible MA(1) process $X_t = Z_t + \theta Z_{t-1}$, where $|\theta| < 1$. Using invertibility, each X_t has an infinite autoregressive representation

$$X_t = \sum_{j=1}^{\infty} (-\theta)^j X_{t-j} + Z_t$$

that is nonlinear in the unknown parameter θ to be estimated. The method of moments is here based on solving

$$\hat{\rho}(1) = \frac{\hat{\gamma}(1)}{\hat{\gamma}(0)} = \frac{\hat{\theta}}{1 + \hat{\theta}^2}.$$

for $\hat{\theta}$. The foregoing quadratic equation has the two solutions

$$\hat{\theta} = \frac{1 \pm \sqrt{1 - 4\hat{\rho}(1)^2}}{2\hat{\rho}(1)},$$

of which we pick the invertible one. Note moreover, that $|\hat{\rho}(1)|$ is not necessarily less or equal to $1/2$ which is required for the existence of real solutions. (The theoretical value $|\rho(1)|$, however, is always less than $1/2$ for any MA(1) process, as an easy computation shows). Hence, θ can not always be estimated from given data samples.

Method 2 (Maximum Likelihood Estimation) The innovations algorithm of the previous section applied to a causal ARMA(p, q) process $(X_t)_{t\in\mathbb{Z}}$ gives

$$\hat{X}_{i+1} = \sum_{j=1}^{i} \theta_{ij}(X_{i+1-j} - \hat{X}_{i+1-j}), \qquad 1 \le i < \max\{p, q\},$$

$$\hat{X}_{i+1} = \sum_{j=1}^{p} \phi_j X_{i+1-j} + \sum_{j=1}^{q} \theta_{ij}(X_{i+1-j} - \hat{X}_{i+1-j}), \qquad i \ge \max\{p, q\},$$

with prediction error

$$P_{i+1} = \sigma^2 R_{i+1}.$$

In the last expression, σ^2 has been factored out due to reasons that will become apparent from the form of the likelihood function to be discussed below. Recall that the sequence $(X_{i+1} - \hat{X}_{i+1})_{i\in\mathbb{Z}}$ consists of uncorrelated random variables if the parameters are known. Assuming normality for the errors, we moreover obtain even independence. This can be exploited to define the *Gaussian maximum likelihood estimation (MLE) procedure*. Throughout, it is assumed that $(X_t)_{t\in\mathbb{Z}}$ has zero mean ($\mu = 0$). We collect the parameters of interest in the vectors $\boldsymbol{\beta} = (\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma^2)^T$ and $\boldsymbol{\beta}' = (\boldsymbol{\phi}, \boldsymbol{\theta})^T$, where $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)^T$ and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)^T$. Assume finally that we have observed the variables $X_1, \ldots, X_n$. Then, the Gaussian likelihood function for the innovations is

$$L(\boldsymbol{\beta}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \left(\prod_{i=1}^{n} R_i^{1/2} \right) \exp\left(-\frac{1}{2\sigma^2} \sum_{j=1}^{n} \frac{(X_j - \hat{X}_j)^2}{R_j} \right). \tag{3.5.4}$$

Taking the partial derivative of $\ln L(\boldsymbol{\beta})$ with respect to the variable σ^2 reveals that the MLE for σ^2 can be calculated from

$$\hat{\sigma}^2 = \frac{S(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}})}{n}, \qquad S(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}}) = \sum_{j=1}^{n} \frac{(X_j - \hat{X}_j)^2}{R_j}.$$

Therein, $\hat{\boldsymbol{\phi}}$ and $\hat{\boldsymbol{\theta}}$ denote the MLEs of $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ obtained from minimizing the *profile likelihood* or *reduced likelihood*

$$\ell(\boldsymbol{\phi}, \boldsymbol{\theta}) = \ln\left(\frac{S(\boldsymbol{\phi}, \boldsymbol{\theta})}{n} \right) + \frac{1}{n} \sum_{j=1}^{n} \ln(R_j).$$

Observe that the profile likelihood $\ell(\boldsymbol{\phi}, \boldsymbol{\theta})$ can be computed using the innovations algorithm. The speed of these computations depends heavily on the quality of initial estimates. These are often provided by the non-optimal Yule-Walker procedure. For numerical methods, such as the *Newton-Raphson* and *scoring algorithms*, see Section 3.6 in Shumway and Stoffer (2006).

The limit distribution of the MLE procedure is given as the following theorem. Its proof can be found in Section 8.8 of Brockwell and Davis (1991).

Theorem 3.5.2 *Let $(X_t)_{t\in\mathbb{Z}}$ be a causal and invertible ARMA(p, q) process defined with an iid sequence $(Z_t)_{t\in\mathbb{Z}}$ satisfying $E[Z_t] = 0$ and $E[Z_t^2] = \sigma^2$. Consider the MLE $\hat{\boldsymbol{\beta}}'$ of $\boldsymbol{\beta}'$ that is initialized with the moment estimators of Method 1. Then,*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}' - \boldsymbol{\beta}') \xrightarrow{\mathcal{D}} N(0, \sigma^2 \boldsymbol{\Gamma}_{p,q}^{-1}) \qquad (n \to \infty).$$

The result is optimal. The covariance matrix $\boldsymbol{\Gamma}_{p,q}$ is in block form and can be evaluated in terms of covariances of various autoregressive processes.

Example 3.5.4 (Recruitment Series) The MLE estimation procedure for the recruitment series can be applied in R as follows:

```
> rec.mle = ar.mle(rec, order=2)
```

The mean estimate can be obtained from `rec.mle$x.mean` as $\hat{\mu} = 62.26$, while the autoregressive parameter estimates and their standard errors are accessed with `rec.mle$ar` and `sqrt(rec.mle$asy.var.coef)` as $\hat{\phi}_1 = 1.3513(.0410)$ and $\hat{\phi}_2 = -.4099(.0410)$. Finally, the variance estimate is obtained from `rec.yw$var.pred` as $\hat{\sigma}^2 = 89.3360$. All values are very close to their counterparts in Example 3.3.5.

Method 3 (Least Squares Estimation) An alternative to the method of moments and the MLE is provided by the least squares estimation (LSE). For causal and invertible ARMA(p, q) processes, it is based on minimizing the weighted sum of squares

$$S(\boldsymbol{\phi}, \boldsymbol{\theta}) = \sum_{j=1}^{n} \frac{(X_j - \hat{X}_j)^2}{R_j} \tag{3.5.5}$$

with respect to $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$, respectively. Assuming that $\tilde{\boldsymbol{\phi}}$ and $\tilde{\boldsymbol{\theta}}$ denote these LSEs, the LSE for σ^2 is computed as

$$\tilde{\sigma}^2 = \frac{S(\tilde{\boldsymbol{\phi}}, \tilde{\boldsymbol{\theta}})}{n - p - q}.$$

The least squares procedure has the same asymptotics as the MLE.

Theorem 3.5.3 *The result of Theorem 3.5.2 holds also if $\hat{\boldsymbol{\beta}}'$ is replaced with $\tilde{\boldsymbol{\beta}}'$.*

Example 3.5.5 (Recruitment Series) The least squares estimation has already been discussed in Example 3.3.5, including the R commands.

3.6 Model Selection

In this section, a rough guide for going about the data analysis will be provided. It consists of several parts, most of which have been discussed previously. The main focus is on the order selection of p and q in the case that these parameters are in fact unknown.

Step 1. Plot the data and check whether or not the variability remains reasonably stable throughout the observation period. If that is not the case, use preliminary transformations to stabilize the variance. One popular class is given by the *Box-Cox transformations* (Box and Cox, 1964)

$$f_\lambda(U_t) = \begin{cases} \lambda^{-1}(U_t^\lambda - 1), & U_t \geq 0,\ \lambda > 0. \\ \ln U_t & U_t > 0,\ \lambda = 0. \end{cases}$$

In practice f_0 or $f_{1/2}$ are often adequate choices. (Recall, for instance, the Australian wine sales data of Example 1.4.1.)

Step 2. Remove, if present, trend and seasonal components from the data. Chapter 1 introduced a number of tools to do so, based on the classical decomposition of a time series

$$X_t = m_t + s_t + Z_t$$

into a trend, a seasonality and a residual component. Note that differencing works also without the specific representation in the last display. If the data appears stationary,

move on to the next step. Else apply, for example, another set of difference operations.

Step 3. Suppose now that Steps 1 and 2 have provided us with observations that are well described by a stationary sequence $(X_t)_{t\in\mathbb{Z}}$. The goal is then to find the most appropriate ARMA(p,q) model to describe the process. In the unlikely case that p and q can be assumed known, utilize the estimation procedures of Section 3.5 directly. Otherwise, choose them according to one of the following criteria.

(a) The standard criterion that is typically implemented in software packages is a modification of *Akaike's information criterion*, see Akaike (1969), which was given by Hurvich and Tsai (1989). In this paper, it is suggested that the ARMA model parameters be chosen that they minimize the objective function

$$\mathrm{AIC}_C(\boldsymbol{\phi},\boldsymbol{\theta},p,q) = -2\ln L(\boldsymbol{\phi},\boldsymbol{\theta},S(\boldsymbol{\phi},\boldsymbol{\theta})/n) + \frac{2(p+q+1)n}{n-p-q-2}. \tag{3.6.1}$$

Here, $L(\boldsymbol{\phi},\boldsymbol{\theta},\sigma^2)$ denotes the Gaussian likelihood defined in (3.5.4) and $S(\boldsymbol{\phi},\boldsymbol{\theta})$ is the weighted sum of squares in (3.5.5). It can be seen from the definition that the AIC_C does not attempt to minimize the log-likelihood function directly. The introduction of the penalty term on the right-hand side of (3.6.1) reduces the risk of overfitting.

(b) For pure autoregressive processes, Akaike (1969) introduced a criterion that is based on a minimization of the *final prediction error*. Here, the order p is chosen as the minimizer of the objective function

$$\mathrm{FPE} = \hat{\sigma}^2\frac{n+p}{n-p},$$

where $\hat{\sigma}^2$ denotes the MLE of the unknown noise variance σ^2. For more on this topic and other procedures that help fit a model, we refer here to Section 9.3 of Brockwell and Davis (1991).

Step 4. The last step in the analysis is concerned with *diagnostic checking* by applying the goodness of fit tests of Section 1.5.

3.7 Summary

The class of autoregressive moving average processes has been introduced to model stationary stochastic processes. We have examined theoretical properties such as causality and invertibility, which depend on the zeroes of the autoregressive and moving average polynomials, respectively.

We have learned how the causal representation of an ARMA process can be utilized to compute its covariance function which contains all information about the dependence structure.

Assuming known parameter values, several forecasting procedures have been discussed. The Durbin-Levinson algorithm works well for pure AR processes, while the innovations algorithm is particularly useful for pure MA processes. Predictions using an infinite past work well for causal and invertible ARMA processes. For practical purposes, however, a truncated version is more relevant.

Since the exact parameter values are in general unknown, we have introduced various estimation procedures. The Yule-Walker procedure is only optimal in the AR case but

provides useful initial estimates that can be used for the numerical derivation of maximum likelihood or least squares estimates.

Finally, we have provided a framework that may potentially be useful when facing the problem of analyzing a data set in practice.

Chapter 4

Spectral Analysis

4.1 Introduction

Many of the time series discussed in the previous chapters displayed strong periodic components: The sunspot numbers of Example 1.1.1, the number of trapped lynx of Example 1.1.2 and the Australian wine sales data of Example 1.4.1. Often, there is an obvious choice for the period d of this cyclical part such as an annual pattern in the wine sales. Given d, we could then proceed by removing the seasonal effects as in Section 1.4. In the first two examples it is, however, somewhat harder to determine the precise value of d. In this chapter, we discuss therefore a general method to deal with the periodic components of a time series. To complicate matters, it is usually the case that several cyclical patterns are simultaneously present in a time series. As an example recall the southern oscillation index (SOI) data which exhibits both an annual pattern and a so-called El Niño pattern.

The sine and cosine functions are the prototypes of periodic functions. We are going to utilize them here to describe cyclical behavior in time series. Before doing so, we define a *cycle* as one complete period of a sine or cosine function over a time interval of length 2π. We also define the *frequency*

$$\omega = \frac{1}{d}$$

as the number of cycles per observation, where d denotes the period of a time series (that is, the number of observations in a cycle). For monthly observations with an annual period, we have obviously $d = 12$ and hence $\omega = 1/12 = .083$ cycles per observation. Now we can reconsider the process

$$X_t = R\sin(2\pi\omega t + \varphi)$$

as introduced in Example 1.2.2, using the convention $\lambda = 2\pi\omega$. To include randomness in this process, we choose the amplitude R and the phase φ to be random variables. An equivalent representation of this process is given by

$$X_t = A\cos(2\pi\omega t) + B\sin(2\pi\omega t),$$

with $A = R\sin(\varphi)$ and $B = R\cos(\varphi)$ usually being independent standard normal variates. Then, $R^2 = A^2 + B^2$ is a χ-squared random variable with 2 degrees of freedom and

$\varphi = \tan^{-1}(B/A)$ is uniformly distributed on $(-\pi, \pi]$. Moreover, R and φ are independent. Choosing now the value of ω we can describe one particular periodicity. To accommodate more than one, it seems natural to consider mixtures of these periodic series with multiple frequencies and amplitudes:

$$X_t = \sum_{j=1}^{m} \big[A_j \cos(2\pi\omega_j t) + B_j \sin(2\pi\omega_j t)\big], \qquad t \in \mathbb{Z},$$

where $A_1, \ldots, A_m$ and $B_1, \ldots, B_m$ are independent random variables with zero mean and variances $\sigma_1^2, \ldots, \sigma_m^2$, and $\omega_1, \ldots, \omega_m$ are *distinct* frequencies. Generalizing the solution to one of our homework problems, we find that $(X_t)_{t\in\mathbb{Z}}$ is a weakly stationary process with lag-h ACVF

$$\gamma(h) = \sum_{j=1}^{m} \sigma_j^2 \cos(2\pi\omega_j h), \qquad h \in \mathbb{Z}.$$

The latter result yields in particular that $\gamma(0) = \sigma_1^2 + \ldots + \sigma_m^2$. The variance of X_t is consequently the sum of the component variances.

Example 4.1.1 Let $m = 2$ and choose $A_1 = B_1 = 1$, $A_2 = B_2 = 4$ to be constant as well as $\omega_1 = 1/12$ and $\omega_2 = 1/6$. This means that

$$X_t = X_t^{(1)} + X_t^{(2)} = \big[\cos(2\pi t/12) + \sin(2\pi t/12)\big] + \big[4\cos(2\pi t/6) + 4\sin(2\pi t/6)\big]$$

is the sum of two periodic components of which one exhibits an annual cycle and the other a cycle of six months. For all processes involved, realizations of $n = 48$ observations (4 years of data) are displayed in Figure 4.1. Also shown is a fourth time series plot which contains the X_t distorted by standard normal independent noise, $\tilde{X}_t$. The corresponding R code is

```
> t = 1:48
> x1 = cos(2*pi*t/12)+sin(2*pi*t/12)
> x2 = 4*cos(2*pi*t/6)+4*sin(2*pi*t/6)
> x = x1+x2
> tildex = x+rnorm(48)
```

Note that the squared amplitude of $X_t^{(1)}$ is $1^2 + 1^2 = 2$. The maximum and minimum values of $X_t^{(1)}$ are therefore $\pm\sqrt{2}$. Similarly, we obtain $\pm\sqrt{32}$ for the second component.

For a statistician it is now important to develop tools to recover the periodicities from the data. The branch of statistics concerned with this problem is called spectral analyis. The standard method in this area is based on the *periodogram* which we are introducing now. Suppose for the moment that we know the frequency parameter $\omega_1 = 1/12$ in Example 4.1.1. To obtain estimates of A_1 and B_1, one could try to run a regression using the explanatory variables $Y_{t,1} = \cos(2\pi t/12)$ or $Y_{t,2} = \sin(2\pi t/12)$ to compute the least squares estimators

$$\hat{A}_1 = \frac{\sum_{t=1}^n X_t Y_{t,1}}{\sum_{t=1}^n Y_{t,1}^2} = \frac{2}{n}\sum_{t=1}^{n} X_t \cos(2\pi t/12),$$

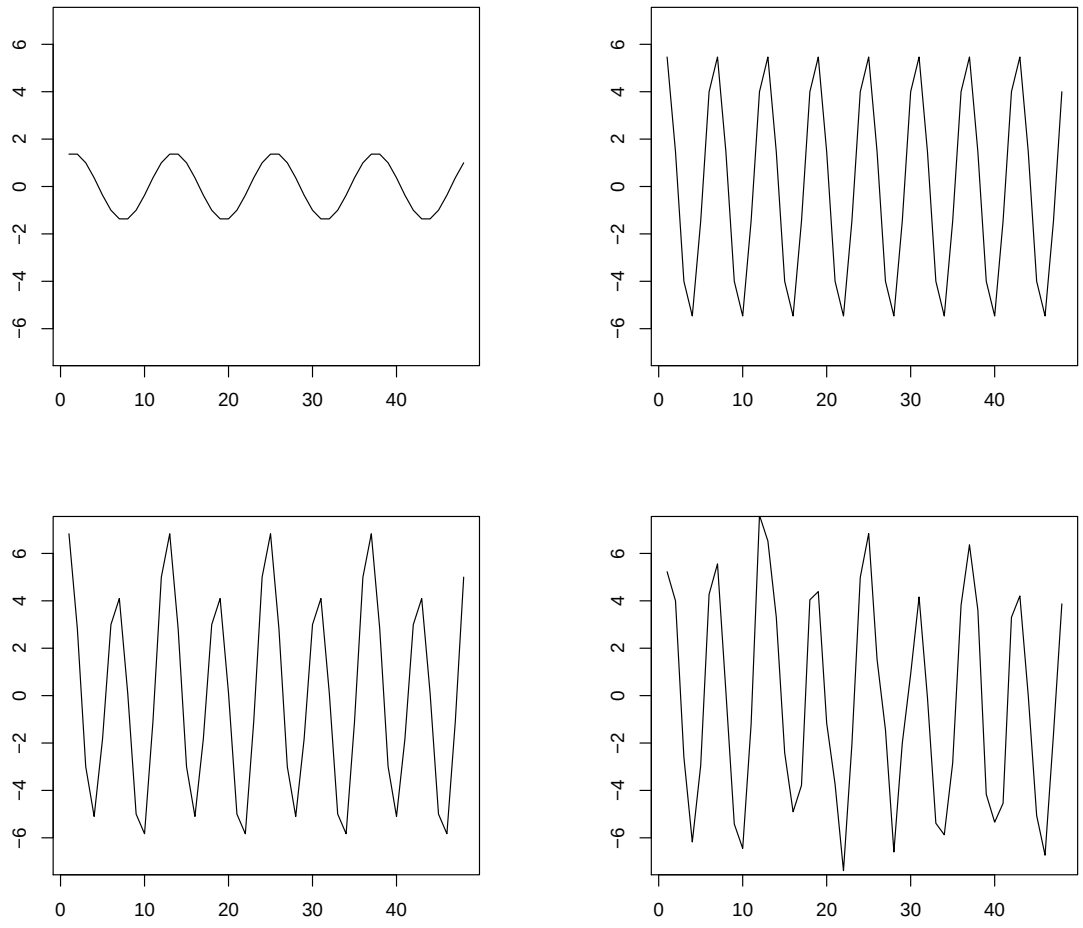

Figure 4.1: Time series plots of $(X_t^{(1)})$, $(X_t^{(2)})$, (X_t) and $(\tilde{X}_t)$.

$$\hat{B}_1 = \frac{\sum_{t=1}^n X_t Y_{t,2}}{\sum_{t=1}^n Y_{t,2}^2} = \frac{2}{n}\sum_{t=1}^n X_t \sin(2\pi t/12).$$

Since, in general, the frequencies involved will not be known to the statistician prior to the data analysis, the foregoing suggests to pick a number of potential ω's, say j/n for $j = 1, \ldots, n/2$ and to run a long regression of the form

$$X_t = \sum_{j=0}^{n/2} \big[A_j \cos(2\pi jt/n) + B_j \sin(2\pi jt/n)\big]. \tag{4.1.1}$$

This leads to least squares estimates $\hat{A}_j$ and $\hat{B}_j$ of which the "significant" ones should be selected. Note that the regression in (4.1.1) is a perfect one because there are as many unknowns as variables! Note also that

$$P(j/n) = \hat{A}_j^2 + \hat{B}_j^2$$

is essentially (up to a normalization) an estimator for the correlation between the time series X_t and the corresponding sum of the periodic cosine and sine functions at frequency j/n. The collection of all $P(j/n)$, $j = 1, \ldots, n/2$, is called the *scaled periodogram*. It can be computed quickly via an algorithm known as the fast Fourier transform (FFT) which in turn is based on the discrete Fourier transform (DFT)

$$d(j/n) = \frac{1}{\sqrt{n}}\sum_{t=1}^n X_t \exp(-2\pi ijt/n).$$

(For apparent reasons, the frequencies j/n are called the Fourier or fundamental frequencies.) Since $\exp(-ix) = \cos(x) - i\sin(x)$ and $|z|^2 = z\bar{z} = (a+ib)(a-ib) = a^2 + b^2$ for any complex number $z = a + ib$, it follows that

$$I(j/n) = |d(j/n)|^2 = \frac{1}{n}\left(\sum_{t=1}^n X_t \cos(2\pi jt/n)\right)^2 + \frac{1}{n}\left(\sum_{t=1}^n X_t \sin(2\pi jt/n)\right)^2.$$

We refer to $I(j/n)$ as the *periodogram*. It also follows immediately that the periodogram and the scaled periodogram are related via the identity $4I(j/n) = nP(j/n)$.

Example 4.1.2 Using the expressions and notations of Example 4.1.1, we can compute the periodogram and the scaled periodogram in R as follows:

```
> t = 1:48
> I = abs(fft(x)/sqrt(48))^2
> P = 4*I/48
> f = 0:24/48
> plot(f, P[1:25], type="l")
> abline(v=1/12)
> abline(v=1/6)
```

The corresponding (scaled) periodogram for $(\tilde{X}_t)$ can be obtained in a similar fashion.

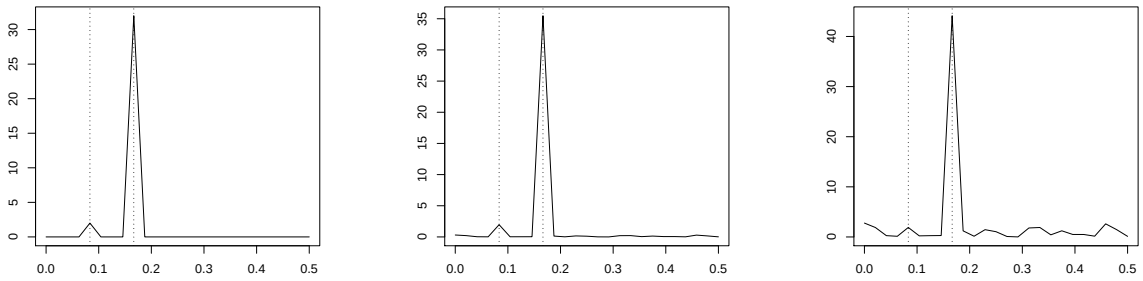

Figure 4.2: The scaled periodograms of (X_t), $(\tilde{X}_t^{(1)})$ and $(\tilde{X}_t^{(2)})$.

The scaled periodograms are shown in the left and middle panel of Figure 4.2. The right panel displays the scaled periodogram of another version of $(\tilde{X}_t)$ in which the standard normal noise has been replaced with normal noise with variance 9. From these plots it can be seen that the six months periodicity is clearly visible in the graphs (see the dashed vertical lines at $x = 1/6$). The less pronounced annual cycle (vertical line at $x = 1/12$) is still visible in the first two scaled periodograms but is lost if the noise variance is increased as in the right plot. Note, however, that the y-scale is different for all three plots.

In the ideal situation that we observe the periodic component without additional contamination by noise, we can furthermore see the variance decomposition from above. We have shown in the lines preceding Example 4.1.1 that $\gamma(0) = \sigma_1^2 + \sigma_2^2$, where in this example $\sigma_1^2 = 2$ and $\sigma_2^2 = 32$. These values are readily read from the scaled periodogram in the left panel of Figure 4.2. The contamination with noise alters these values.

In the next section, we establish that the time domain approach (based on properties of the ACVF, that is, regression on past values of the time series) we have discussed so far and the frequency domain approach (using a periodic function approach via fundamental frequencies, that is, regression on sine and cosine functions) are equivalent. We discuss in some detail the spectral density (the population counterpart of the periodogram) and properties of the periodogram itself.

4.2 The spectral density and the periodogram

The fundamental technical result which is at the core of spectral analysis states that any (weakly) stationary time series can be viewed (approximately) as a random superposition of sine and cosine functions varying at various frequencies. In other words, the regression in (4.1.1) is approximately true for all weakly stationary time series. In Chapters 1–3, we have seen how the characteristics of a stationary stochastic process can be described in terms of its ACVF $\gamma(h)$. It is our first goal to introduce the quantity corresponding to $\gamma(h)$ in the frequency domain.

Definition 4.2.1 (Spectral Density) *If the ACVF $\gamma(h)$ of a stationary time series*

$(X_t)_{t\in\mathbb{Z}}$ satisfies the condition

$$\sum_{h=-\infty}^{\infty} |\gamma(h)| < \infty,$$

then there exists a function f defined on $(-1/2, 1/2]$ such that

$$\gamma(h) = \int_{-1/2}^{1/2} \exp(2\pi i\omega h) f(\omega) d\omega, \qquad h \in \mathbb{Z},$$

and

$$f(\omega) = \sum_{h=-\infty}^{\infty} \gamma(h) \exp(-2\pi i\omega h), \qquad \omega \in (-1/2, 1/2].$$

The function f is called the spectral density of the process $(X_t)_{t\in\mathbb{Z}}$.

Definition 4.2.1 (which contains a theorem part as well) establishes that each weakly stationary process can be equivalently described in terms of its ACVF or its spectral density. It also provides the formulas to compute one from the other. Time series analysis can consequently be performed either in the time domain (using $\gamma(h)$) or in the frequency domain (using $f(\omega)$). Which approach is the more suitable one cannot be decided in a general fashion but has to be reevaluated for every application of interest.

In the following, we collect several basic properties of the spectral density and evaluate f for several important examples. That the spectral density is analogous to a probability density function is established in the next proposition.

Proposition 4.2.1 *If $f(\omega)$ is the spectral density of a weakly stationary process $(X_t)_{t\in\mathbb{Z}}$, then the following statements hold:*

(a) $f(\omega) \geq 0$ for all ω. This follows from the positive definiteness of $\gamma(h)$;

(b) $f(\omega) = f(-\omega)$ and $f(\omega + 1) = f(\omega)$;

(c) The variance of $(X_t)_{t\in\mathbb{Z}}$ is given by

$$\gamma(0) = \int_{-1/2}^{1/2} f(\omega) d\omega.$$

Part (c) of the proposition tells us that the variance of a weakly stationary process is equal to the integrated spectral density over all frequencies. We will come back to this property below, when we will discuss a spectral analysis of variance (spectral ANOVA). Before, we turn to three examples.

Example 4.2.1 (White Noise) If $(Z_t)_{t\in\mathbb{Z}} \sim \mathrm{WN}(0, \sigma^2)$, then its ACVF is nonzero only for $h = 0$, in which case we have $\gamma_Z(h) = \sigma^2$. Plugging this result into the defining equation in Definition 4.2.1 yields that

$$f_Z(\omega) = \gamma_Z(0) \exp(-2\pi i\omega 0) = \sigma^2.$$

The spectral density of a white noise sequence is therefore constant for all $\omega \in (-1/2, 1/2]$, which means that every frequency ω contributes equally to the overall spectrum. This explains the term "white" noise (in analogy to "white" light).

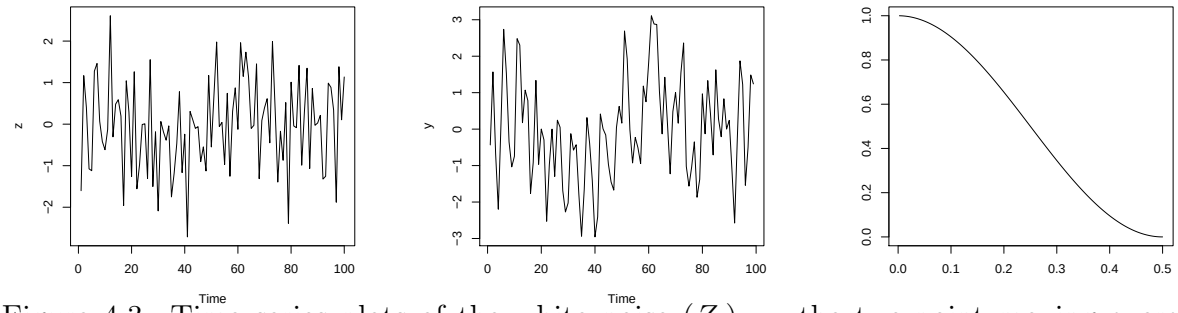

Figure 4.3: Time series plots of the white noise $(Z_t)_{t\in\mathbb{Z}}$, the two-point moving average $(X_t)_{t\in\mathbb{Z}}$ (left and middle) and the spectral density of $(X_t)_{t\in\mathbb{Z}}$ (right).

Example 4.2.2 (Moving Average) Let $(Z_t)_{t\in\mathbb{Z}} \sim \mathrm{WN}(0,\sigma^2)$ and define the time series $(X_t)_{t\in\mathbb{Z}}$ by

$$X_t = \tfrac{1}{2}\left(Z_t + Z_{t-1}\right), \qquad t \in \mathbb{Z}.$$

It is an easy exercise to show that

$$\gamma_X(h) = \frac{\sigma^2}{4}\left(2 - |h|\right), \qquad h = 0, \pm 1,$$

and that $\gamma_X = 0$ otherwise. Therefore,

$$\begin{aligned} f_X(\omega) &= \sum_{h=-1}^{1} \gamma_X(h)\exp(2\pi i\omega h) \\ &= \frac{\sigma^2}{4}\left[\exp(-2\pi i\omega(-1)) + 2\exp(-2\pi i\omega 0) + \exp(-2\pi i\omega 1)\right] \\ &= \frac{\sigma^2}{2}\left[1 + \cos(2\pi\omega)\right] \end{aligned}$$

using that $\exp(ix) = \cos(x) + i\sin(x)$, $\cos(x) = \cos(-x)$ and $\sin(x) = -\sin(-x)$. It can be seen from the two time series plots in Figure 4.3 that the application of the two-sided moving average to the white noise sequence smoothes the sample path. This is due to an attenuation of the higher frequencies which is visible in the form of the spectral density in the right panel of Figure 4.3. All plots have been obtained using Gaussian white noise with $\sigma^2 = 1$.

Example 4.2.3 (AR(2) Process) Let $(X_t)_{t\in\mathbb{Z}}$ be an AR(2) process which can be written in the form

$$Z_t = X_1 - \phi_1 X_{t-1} - \phi_2 X_{t-2}, \qquad t \in \mathbb{Z}.$$

In this representation, we can see that the ACVF γ_Z of the white noise sequence can be obtained as

$$\gamma_Z(h) = E\left[\left(X_t - \phi_1 X_{t-1} - \phi_2 X_{t-2}\right)\left(X_{t+h} - \phi_1 X_{t+h-1} - \phi_2 X_{t+h-2}\right)\right]$$

$$= (1+\phi_1^2+\phi_2^2)\gamma_X(h) + (\phi_1\phi_2-\phi_1)[\gamma_X(h+1)+\gamma_X(h-1)]$$
$$- \phi_2[\gamma_X(h+2)+\gamma_X(h-2)]$$

Now we know from Definition 4.2.1 that

$$\gamma_X(h) = \int_{-1/2}^{1/2} \exp(2\pi i\omega h) f_X(\omega) d\omega \qquad \text{and} \qquad \gamma_Z(h) = \int_{-1/2}^{1/2} \exp(2\pi i\omega h) f_Z(\omega) d\omega,$$

where $f_X(\omega)$ and $f_Z(\omega)$ denote the respective spectral densities. We find consequently that

$$\gamma_Z(h) = \int_{-1/2}^{1/2} \exp(2\pi i\omega h) f_Z(\omega) d\omega$$
$$= (1+\phi_1^2+\phi_2^2)\gamma_X(h) + (\phi_1\phi_2-\phi_1)[\gamma_X(h+1)-\gamma_X(h-1)] - \phi_2[\gamma_X(h+2)-\gamma_X(h-2)]$$
$$= \int_{-1/2}^{1/2} \big[(1+\phi_1^2+\phi_2^2) + (\phi_1\phi_2-\phi_1)(\exp(2\pi i\omega)+\exp(-2\pi i\omega))$$
$$-\phi_2(\exp(4\pi i\omega)+\exp(-4\pi i\omega))\big] \exp(2\pi i\omega h) f_X(\omega) d\omega$$
$$= \int_{-1/2}^{1/2} \big[(1+\phi_1^2+\phi_2^2) + 2(\phi_1\phi_2-\phi_1)\cos(2\pi\omega) - 2\phi_2\cos(4\pi\omega)\big] \exp(2\pi i\omega h) f_X(\omega) d\omega.$$

The foregoing implies together with $f_Z(\omega) = \sigma^2$ that

$$\sigma^2 = \big[(1+\phi_1^2+\phi_2^2) + 2(\phi_1\phi_2-\phi_1)\cos(2\pi\omega) - 2\phi_2\cos(4\pi\omega)\big] f_X(\omega).$$

Hence, the spectral density of an AR(2) process has the form

$$f_X(\omega) = \sigma^2 \left[(1+\phi_1^2+\phi_2^2) + 2(\phi_1\phi_2-\phi_1)\cos(2\pi\omega) - 2\phi_2\cos(4\pi\omega)\right]^{-1}.$$

In Figure 4.4 you can see the time series plot of an AR(2) process with parameters $\phi_1 = 1.35$, $\phi_2 = -.41$ and $\sigma^2 = 89.34$. These values are very similar to the ones obtained for the recruitment series in Section 3.5. The same figure also shows the corresponding spectral density using the formula we just derived.

With the contents of this Section, we have so far established the spectral density $f(\omega)$ as a population quantity describing the impact of the various periodic components. Next, we shall verify that the periodogram $I(\omega_j)$ introduced in Section 4.1 is the sample counterpart of the spectral density.

Proposition 4.2.2 *Let $\omega_j = j/n$ denote the Fourier frequencies. If $I(\omega_j) = |d(\omega_j)|^2$ is the periodogram based on observations $X_1, \ldots, X_n$ of a weakly stationary process $(X_t)_{t\in\mathbb{Z}}$, then*

$$I(\omega_j) = \sum_{h=-n+1}^{n-1} \hat{\gamma}_n(h) \exp(-2\pi i\omega_j h), \qquad j \neq 0.$$

If $j = 0$, then $I(\omega_0) = I(0) = n\bar{X}_n^2$.

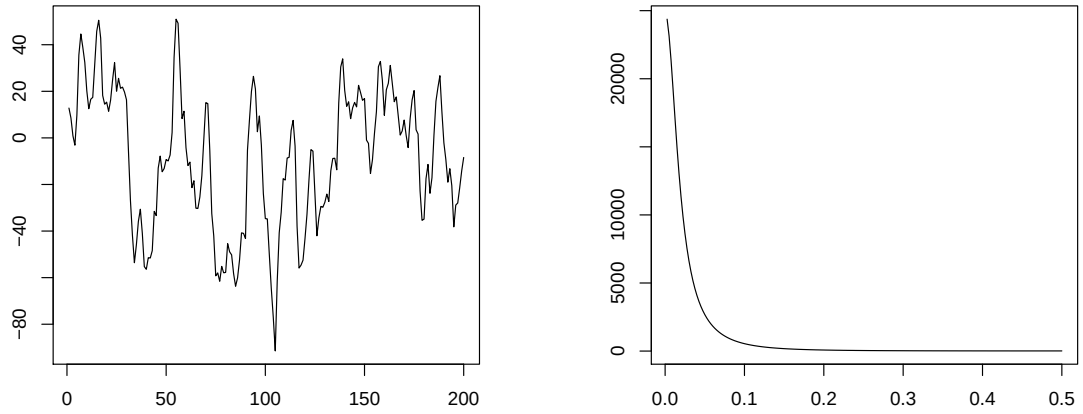

Figure 4.4: Time series plot and spectral density of the AR(2) process in Example 4.2.3.

Proof. Let first $j \neq 0$. Using that $\sum_{t=1}^{n} \exp(-2\pi i \omega_j t) = 0$, we can write

$$\begin{aligned} I(\omega_j) &= \frac{1}{n} \sum_{t=1}^{n} \sum_{s=1}^{n} (X_t - \bar{X}_n)(X_s - \bar{X}_n) \exp(-2\pi i \omega_j (t-s)) \\ &= \frac{1}{n} \sum_{h=-n+1}^{n-1} \sum_{t=1}^{n-|h|} (X_{t+|h|} - \bar{X}_n)(X_t - \bar{X}_n) \exp(-2\pi i \omega_j h) \\ &= \sum_{h=-n+1}^{n-1} \hat{\gamma}_n(h) \exp(-2\pi i \omega_j h), \end{aligned}$$

which proves the first claim of the proposition. If $j = 0$, we have with $\cos(0) = 1$ and $\sin(0) = 0$ that $I(0) = n\bar{X}_n^2$. This completes the proof. □

More can be said about the periodogram. In fact, one can interpret spectral analysis as a spectral analysis of variance (ANOVA). To see this, let first

$$d_c(\omega_j) = \mathrm{Re}(d(\omega_j)) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} X_t \cos(2\pi\omega_j t),$$

$$d_s(\omega_j) = \mathrm{Im}(d(\omega_j)) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} X_t \sin(2\pi\omega_j t).$$

Then, $I(\omega_j) = d_c^2(\omega_j) + d_s^2(\omega_j)$. Let us now go back to the introductory example and study the process

$$X_t = A_0 + \sum_{j=1}^{m} \big[A_j \cos(2\pi\omega_j t) + B_j \sin(2\pi\omega_j t) \big],$$

where $m = (n-1)/2$ and n odd. Suppose we have observed $X_1, \ldots, X_n$. Then, using regression techniques as before, we can see that $A_0 = \bar{X}_n$ and

$$A_j = \frac{2}{n}\sum_{t=1}^{n} X_t \cos(2\pi\omega_j t) = \frac{2}{\sqrt{n}} d_c(\omega_j),$$

$$B_j = \frac{2}{n}\sum_{t=1}^{n} X_t \sin(2\pi\omega_j t) = \frac{2}{\sqrt{n}} d_s(\omega_j).$$

Therefore,

$$\sum_{t=1}^{n}(X_t - \bar{X}_n)^2 = 2\sum_{j=1}^{m}\left[d_c^2(\omega_j) + d_s^2(\omega_j^2)\right] = 2\sum_{j=1}^{m} I(\omega_j)$$

and we obtain the following ANOVA table. If the underlying stochastic process exhibits

Source	df	SS	MS
ω_1	2	$2I(\omega_1)$	$I(\omega_1)$
ω_2	2	$2I(\omega_2)$	$I(\omega_2)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
ω_m	2	$2I(\omega_m)$	$I(\omega_m)$
Total	$n-1$	$\sum_{t=1}^{n}(X_t - \bar{X}_n)^2$	

a strong periodic pattern at a certain frequency, then the periodogram will most likely pick these up.

Example 4.2.4 Let us consider the $n = 5$ data points $X_1 = 2$, $X_2 = 4$, $X_3 = 6$, $X_4 = 4$ and $X_5 = 2$, which display a cyclical but nonsinosoidal pattern. This suggests that $\omega = 1/5$ is significant and $\omega = 2/5$ is not. In R, you can produce the spectral ANOVA as follows.

```
> x = c(2,4,6,4,2), t=1:5
> cos1 = cos(2*pi*t*1/5)
> sin1 = sin(2*pi*t*1/5)
> cos2 = cos(2*pi*t*2/5)
> sin2 = sin(2*pi*t*2/5)
```

This generates the data and the independent cosine and sine variables. Now we can run a regression and check the ANOVA output.

```
> reg = lm(x~cos1+sin1+cos2+sin2)
> anova(reg)
```

This leads to the following output. According to our previous reasoning (check the previous table!), the periodogram at frequency $\omega_1 = 1/5$ is given as the sum of the `cos1` and `sin1` coefficients, that is, $I(1/5) = (d_c(1/5) + d_s(1/5))/2 = (7.1777 + 3.7889)/2 = 5.4833$. Similarly, $I(2/5) = (d_c(2/5) + d_s(2/5))/2 = (0.0223 + 0.2111)/2 = 0.1167$. Note, however,

```
Response:   x
           Df  Sum Sq  Mean Sq  F  value Pr(>F)
cos1        1  7.1777   7.1777
cos2        1  0.0223   0.0223
sin1        1  3.7889   3.7889
sin2        1  0.2111   0.2111
Residuals   0  0.0000
```

that the mean squared error is computed differently in R. We can compare these values with the periodogram:

```
> abs(fft(x))^2/5
  [1]  64.8000000  5.4832816  0.1167184  0.1167184  5.4832816
```

The first value here is $I(0) = n\bar{X}_n^2 = 5 * (18/5)^2 = 64.8$. The second and third value are $I(1/5)$ and $I(2/5)$, respectively, while $I(3/5) = I(2/5)$ and $I(4/5) = I(1/5)$ complete the list.

In the next section, we will discuss some large sample properties of the periodogram to get a better understanding of spectral analysis.

4.3 Large sample properties

Let $(X_t)_{t\in\mathbb{Z}}$ be a weakly stationary time series with mean μ, absolutely summable ACVF $\gamma(h)$ and spectral density $f(\omega)$. Proceeding as in the proof of Proposition 4.2.2, we obtain

$$I(\omega_j) = \frac{1}{n}\sum_{h=-n+1}^{n-1}\sum_{t=1}^{n-|h|}(X_{t+|h|}-\mu)(X_t-\mu)\exp(-2\pi i\omega_j h),$$

provided $\omega_j \neq 0$. Using this representation, the limiting behavior of the periodogram can be established.

Proposition 4.3.1 *Let $I(\cdot)$ be the periodogram based on observations $X_1,\ldots,X_n$ of a weakly stationary process $(X_t)_{t\in\mathbb{Z}}$, then, for any $\omega \neq 0$,*

$$E[I(\omega_{j:n})] \to f(\omega) \qquad (n\to\infty),$$

where $\omega_{j:n} = j_n/n$ with $(j_n)_{n\in\mathbb{N}}$ chosen such that $\omega_{j:n}\to\omega$ as $n\to\infty$. If $\omega = 0$, then

$$E[I(0)] - n\mu^2 \to f(0) \qquad (n\to\infty).$$

Proof. There are two limits involved in the computations of the periodogram mean. First, we take the limit as $n\to\infty$. This, however, requires secondly that for each n we have to work with a different set of Fourier frequencies. To adjust for this, we have introduced the notation $\omega_{j:n}$. If $\omega_j \neq 0$ is a Fourier frequency (n fixed!), then

$$E[I(\omega_j)] = \sum_{h=-n+1}^{n-1}\left(\frac{n-|h|}{n}\right)\gamma(h)\exp(-2\pi i\omega_j h).$$

Therefore ($n \to \infty$!),

$$E[I(\omega_{j:n})] \to \sum_{h=-\infty}^{\infty} \gamma(h) \exp(-2\pi i \omega h) = f(\omega),$$

thus proving the first claim. The second follows from $I(0) = n\bar{X}_n^2$ (see Proposition 4.2.2), so that $E[I(0)] - n\mu^2 = n(E[\bar{X}_n^2] - \mu^2) = n\text{Var}(\bar{X}_n) \to f(0)$ as $n \to \infty$ as in Chapter 2. The proof is complete. □

Proposition 4.3.1 shows that the periodogram $I(\omega)$ is asymptotically unbiased for $f(\omega)$. It is, however, inconsistent. This is implied by the following proposition which is given without proof and it not surprising considering that each value $I(\omega_j)$ is the sum of squares of only two random variables irrespective of the sample size.

Proposition 4.3.2 *If $(X_t)_{t\in\mathbb{Z}}$ is a (causal or noncausal) weakly stationary time series such that*

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

with $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$ and $(Z_t)_{t\in\mathbb{Z}} \sim WN(0, \sigma^2)$, then

$$\left(\frac{2I(\omega_{1:n})}{f(\omega_1)}, \ldots, \frac{2I(\omega_{m:n})}{f(\omega_m)} \right) \xrightarrow{\mathcal{D}} (\xi_1, \ldots, \xi_m),$$

where $\omega_1, \ldots, \omega_m$ are m distinct frequencies with $\omega_{j:n} \to \omega_j$ and $f(\omega_j) > 0$. The variables $\xi_1, \ldots, \xi_m$ are independent, identical chi-squared distributed with two degrees of freedom.

The result of this proposition can be used to construct confidence intervals for the value of the spectral density at frequency ω. To this end, denote by $\chi_2^2(\alpha)$ the lower tail probability of the chi-squared variable ξ_j, that is,

$$P(\xi_j \leq \chi_2^2(\alpha)) = \alpha.$$

Then, we get from Proposition 4.3.2 that an approximate confidence interval with level $1 - \alpha$ is given by

$$\frac{2I(\omega_{j:n})}{\chi_2^2(1-\alpha/2)} \leq f(\omega) \leq \frac{2I(\omega_{j:n})}{\chi_2^2(\alpha/2)}.$$

Proposition 4.3.2 also suggests that confidence intervals can be derived simultaneously for several frequency components. Before we compute confidence intervals for the dominant frequency of the recruitment data we return for a moment to the computation of the FFT which is the basis for the periodogram usage. To ensure a quick computation time, highly composite integers n' have to be used. To achieve this in general, the length of time series is adjusted by padding the original but detrended data by adding zeroes. In R, spectral analysis is performed with the function `spec.pgram`. To find out which n' is used for your particular data, type `nextn(length(x))`, assuming that your series is in `x`.

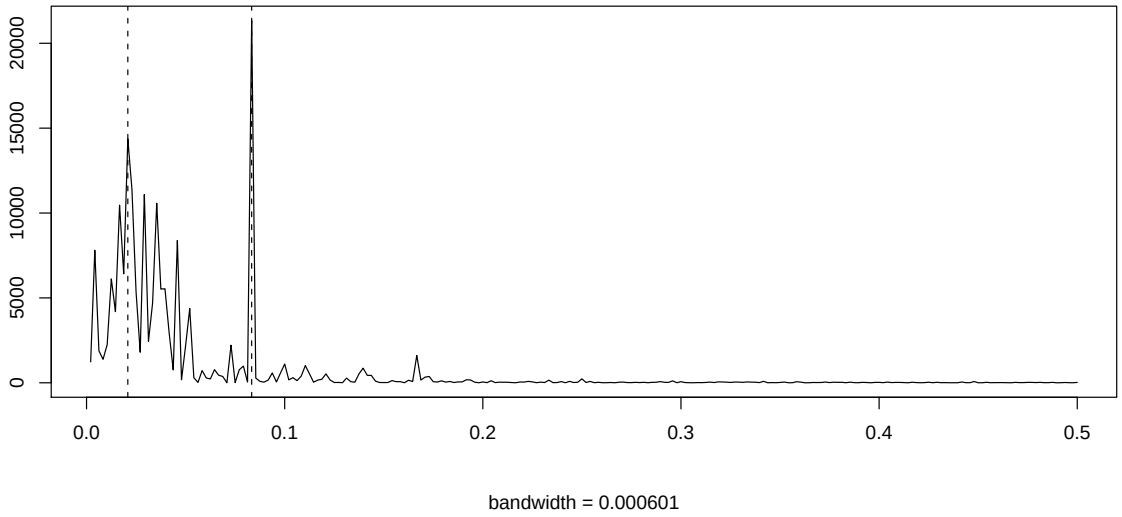

Figure 4.5: Periodogram of the recruitment data discussed in Example 4.3.1.

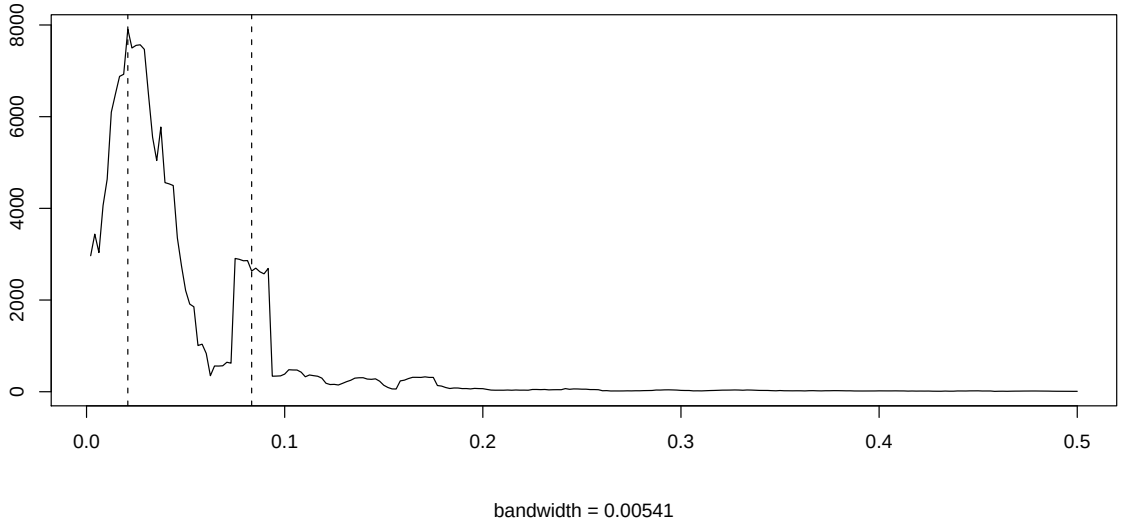

Figure 4.6: Averaged periodogram of the recruitment data discussed in Example 4.3.1.

Example 4.3.1 Figure 4.5 displays the periodogram of the recruitment data which has been discussed in Example 3.3.5. It shows a strong annual frequency component at $\omega = 1/12$ as well as several spikes in the neighborhood of the El Niño frequency $\omega = 1/48$. Higher frequency components with $\omega > .3$ are virtually absent. Even though we have fitted an AR(2) model to this data in Chapter 3 and forecasted future values based on this fit, we see that the periodogram here does not validate this fit as the spectral density of an AR(2) process (as computed in Example 4.2.3) is qualitatively different. In R, you can use the following commands (`nextn(length(rec))` provides you with $n' = 480$ here if the recruitment data is stored in `rec` as before).

```
> rec.pgram = spec.pgram(rec, taper=0, log="no")
> abline(v=1/12, lty=2)
> abline(v=1/48, lty=2)
```

The function `spec.pgram` allows you to fine-tune the spectral analysis. For our purposes, we always use the specifications given above for the raw periodogram (`taper` allows you, for example, to exclusively look at a particular frequency band, `log` allows you to plot the log-periodogram and is the R standard).

To compute the confidence intervals for the two dominating frequencies 1/12 and 1/48, you can use the following R code, noting that $1/12 = 40/480$ and $1/48 = 10/480$.

```
> rec.pgram$spec[40]
  [1]  21332.94
> rec.pgram$spec[10]
  [1]  14368.42
> u = qchisq(.025, 2), l = qchisq(.975, 2)
> 2*rec.pgram$spec[40]/l
> 2*rec.pgram$spec[40]/u
> 2*rec.pgram$spec[10]/l
> 2*rec.pgram$spec[10]/u
```

Using the numerical values of this analysis, we obtain the following confidence intervals at the level $\alpha = .9$:

$$f(1/12) \in (5783.041, 842606.2) \qquad \text{and} \qquad f(1/48) \in (3895.065, 567522.5).$$

These are much too wide and alternatives to the raw periodogram are needed. These are provided, for example, by a smoothing approach which uses an averaging procedure over a band of neighboring frequencies. This can be done as follows.

```
> k = kernel("daniell",4)
> rec.ave = spec.pgram(rec, k, taper=0, log="no")
> abline(v=1/12, lty=2)
> abline(v=1/48, lty=2)
> rec.ave$bandwidth
  [1]  0.005412659
```

The resulting smoothed periodogram is shown in Figure 4.6. It is less noisy, as is expected from taking averages. More precisely, we have taken here a two-sided Daniell filter with

$m = 4$ which uses $L = 2m + 1$ neighboring frequencies

$$\omega_k = \omega_j + \frac{k}{n}, \qquad k = -m, \ldots, m,$$

to compute the periodogram at $\omega_j = j/n$. The resulting plot in Figure 4.6 shows, on the other hand, that the sharp annual peak has been flattened considerably. The bandwidth reported in R can be computed as $b = L/(\sqrt{12}n)$. To compute confidence intervals one has to adjust the previously derived formula. This is done by taking changing the degrees of freedom from 2 to $df = 2Ln/n'$ (if the zeroes where appended) and leads to

$$\frac{df}{\chi^2_{df}(1-\alpha/2)} \sum_{k=-m}^{m} f\Big(\omega_j + \frac{k}{n}\Big) \le f(\omega) \le \frac{df}{\chi^2_{df}(\alpha/2)} \sum_{k=-m}^{m} f\Big(\omega_j + \frac{k}{n}\Big)$$

for $\omega \approx \omega_j$. For the recruitment data we can use the R code

```
> df = ceiling(rec.ave$df)
> u=qchisq(.025,df), l = qchisq(.975,df)
> df*rec.ave$spec[40]/l
> df*rec.ave$spec[40]/u
> df*rec.ave$spec[10]/l
> df*rec.ave$spec[10]/u
```

to get the confidence intervals

$$f(1/12) \in (1482.427, 5916.823) \qquad \text{and} \qquad f(1/48) \in (4452.583, 17771.64).$$

The compromise between the noisy raw periodogram and further smoothing as described here (with $L = 9$) reverses the magnitude of the 1/12 annual frequency and the 1/48 El Niño component. This is due to the fact that the annual peak is a very sharp one, with neighboring frequencies being basically zero. For the 1/48 component, there are is a whole band of neighboring frequency which also contribute (the El Niño phenomenon is irregular and does only on average appear every four years). Moreover, the annual cycle is now distributed over a whole range. One way around this issue is provided by the use of other kernels such as the modified Daniell kernel given in R as `kernel("modified.daniell", c(3,3))`. This leads to the spectral density in Figure 4.7.

4.4 Linear filtering

A linear filter uses specified coefficients $(\psi_s)_{s\in\mathbb{Z}}$, called the impulse response function, to transform a weakly stationary input series $(X_t)_{t\in\mathbb{Z}}$ into an output series $(Y_t)_{t\in\mathbb{Z}}$ via

$$Y_t = \sum_{s=-\infty}^{\infty} \psi_s X_{t-s}, \qquad t \in \mathbb{Z},$$

where $\sum_{s=-\infty}^{\infty} |\psi_s| < \infty$. Then, the frequency response function

$$\Psi(\omega) = \sum_{s=-\infty}^{\infty} \psi_s \exp(-2\pi i \omega s)$$

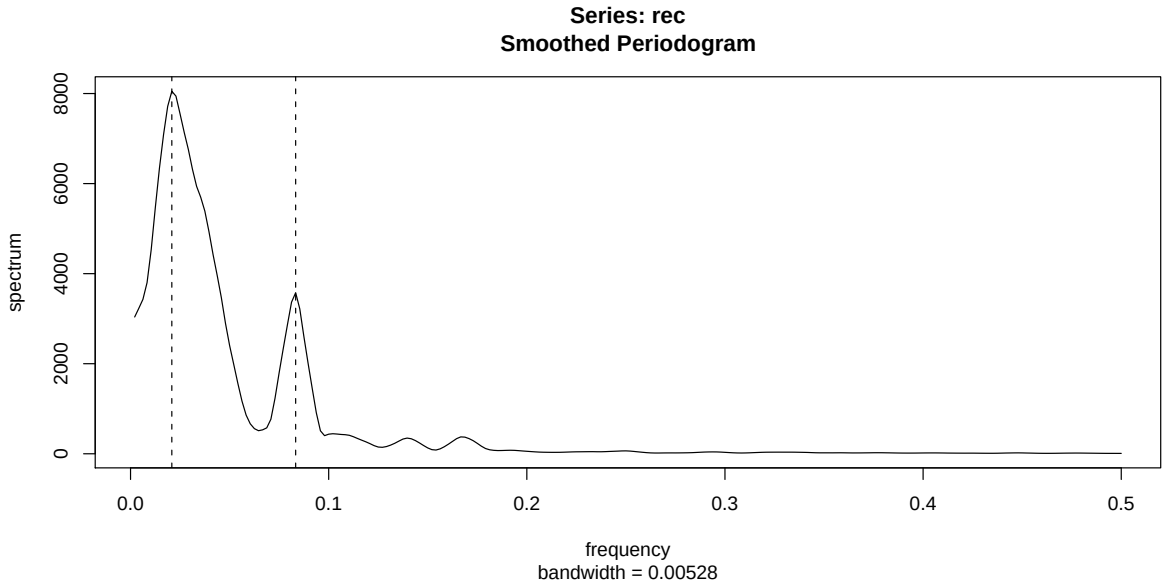

Figure 4.7: The modified Daniell periodogram of the recruitment data discussed in Example 4.3.1.

is well defined. Note that the two-point moving average of Example 4.2.2 and the differenced sequence ∇X_t are examples of linear filters. On the other hand, we can identify *any* causal ARMA process as a linear filter applied to a white noise sequence. Implicitly we have used this concept already to compute the spectral densities in Exampels 4.2.2 and 4.2.3. To investigate this in further detail, let $\gamma_X(h)$ and $\gamma_Y(h)$ denote the ACVF of the input process $(X_t)_{t\in\mathbb{Z}}$ and the output process $(Y_t)_{t\in\mathbb{Z}}$, respectively, and denote by $f_X(\omega)$ and $f_Y(\omega)$ the corresponding spectral densities. The following is the main result in this section.

Theorem 4.4.1 *Under the assumptions made in this section, we have that* $f_Y(\omega) = |\Psi(\omega)|^2 f_X(\omega)$.

Proof. We have that

$$\begin{aligned}
\gamma_Y(h) &= E\big[(Y_{t+h}-\mu_Y)(Y_t-\mu_Y)\big] \\
&= \sum_{r=-\infty}^{\infty}\sum_{s=-\infty}^{\infty}\psi_r\psi_s\gamma(h-r+s) \\
&= \sum_{r=-\infty}^{\infty}\sum_{s=-\infty}^{\infty}\psi_r\psi_s\int_{-1/2}^{1/2}\exp(2\pi i\omega(h-r+s))f_X(\omega)d\omega \\
&= \int_{-1/2}^{1/2}\Big(\sum_{r=-\infty}^{\infty}\psi_r\exp(-2\pi i\omega r)\Big)\Big(\sum_{s=-\infty}^{\infty}\psi_s\exp(2\pi i\omega s)\Big)\exp(2\pi i\omega h)f_X(\omega)d\omega
\end{aligned}$$

$$= \int_{-1/2}^{1/2} \exp(2\pi i\omega h)|\Psi(\omega)|^2 f_X(\omega)d\omega.$$

Now we can identify $f_Y(\omega) = |\Psi(\omega)|^2 f_X(\omega)$, which is the assertion of the theorem. □

Theorem 4.4.1 suggests a way to compute the spectral density of a causal ARMA process. To this end, let $(Y_t)_{t\in\mathbb{Z}}$ be such a causal ARMA(p, q) process satisfying $Y_t = \psi(B)Z_t$, where $(Z_t)_{t\in\mathbb{Z}} \sim \text{WN}(0, \sigma^2)$ and

$$\psi(z) = \frac{\theta(z)}{\phi(z)} = \sum_{s=0}^{\infty} \psi_s z^s, \qquad |z| \leq 1.$$

with $\theta(z)$ and $\phi(z)$ being the moving average and autoregressive polynomial, respectively. Note that the $(\psi_s)_{s\in\mathbb{N}_0}$ can be viewed as a special impulse response function.

Corollary 4.4.1 *If $(Y_t)_{t\in\mathbb{Z}}$ be a causal ARMA(p, q) process. Then, its spectral density is given by*

$$f_Y(\omega) = \sigma^2 \frac{|\theta(e^{-2\pi i\omega})|^2}{|\phi(e^{-2\pi i\omega})|^2}.$$

Proof. We apply Theorem 4.4.1 with input sequence $(Z_t)_{t\in\mathbb{Z}}$. Then $f_Z(\omega) = \sigma^2$, and moreover the frequency response function is

$$\Psi(\omega) = \sum_{s=0}^{\infty} \psi_s \exp(-2\pi i\omega s) = \psi(e^{-2\pi i\omega}) = \frac{\theta(e^{-2\pi i\omega})}{\phi(e^{-2\pi i\omega})}.$$

Since $f_Y(\omega) = |\Psi(\omega)|^2 f_X(\omega)$, the proof is complete. □

Corollary 4.4.1 gives an easy approach to define parametric spectral density estimates for causal ARMA(p, q) processes by simply replacing the population quantities by appropriate sample counterparts. This gives the spectral density estimator

$$\hat{f}(\omega) = \hat{\sigma}_n^2 \frac{|\hat{\theta}(e^{-2\pi i\omega})|^2}{|\hat{\phi}(e^{-2\pi i\omega})|^2}.$$

Now any of the estimation techniques discussed in Section 3.5 may be applied when computing $\hat{f}(\omega)$.

4.5 Summary

In this chapter we have introduced the basic methods of frequency domain time series analysis. These are based on a regression of the given data on cosine and sine functions varying at the Fourier frequencies. On the population side, we have identified the spectral densities as the frequency domain counterparts of absolutely summable autocovariance functions. These are obtained from one another by the application of (inverse) Fourier transforms. On the sample side, the periodogram has been shown to be an estimator for the unknown spectral density. Since it is an incosistent estimator, various techniques have been discussed to overcome this fact. Finally, we have introduced linear filters which can, for example, be used to compute spectral densities of causal ARMA processes and to derive parametric spectral density estimators other than the periodogram.

Bibliography

[1] Akaike, H. (1969). Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics* **21**, 243–247.

[2] Box, G.E.P., and Cox, D.R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B* **26**, 211–252.

[3] Brillinger, D.R. (1975). *Time Series: Data Analysis and Theory.* Holt, Rinehart & Winston, New York.

[4] Brockwell, P.J., and Davis, R.A. (1991). *Time Series: Theory and Methods (2nd ed.).* Springer-Verlag, New York.

[5] Brockwell, P.J., and Davis, R.A. (2002). *An Introduction to Time Series and Forecasting (2nd ed.).* Springer-Verlag, New York.

[6] Fan, J., and Yao, Q. (2003). *Nonlinear Time Series.* Springer-Verlag, New York.

[7] Hurvich, C.M., and Tsai, C.L. (1989). Regression and time series model selection in small samples. *Biometrika* **76**, 297–307.

[8] Ljung, G.M., and Box, G.E.P. (1978). On a measure of lack of fit in time series models. *Biometrika* **65**, 297–303.

[9] McLeod, A.I., and Li, W.K. (1983). Diagnostic checking ARMA time series models using squared-residual autocorrelations. *Journal of Time Series Analysis* **4**, 269–273.

[10] Peña, D., Tiao, G.C., and Tsay, R.S. (eds.) (2001). *A Course in Time Series Analysis.* John Wiley & Sons, New York.

[11] Shapiro, S.S., and Francia, R.S. (1972). An approximate analysis of variance test for normality. *Journal of the American Statistical Association* **67**, 215–216.

[12] Shumway, R.H., and Stoffer, D.A. (2006). *Time Series Analysis and its Applications (2nd ed.).* Springer-Verlag, New York.

[13] Tong, H. (1990). *Non-linear Time Series: A Dynamical System Approach.* Clarendon Press, Oxford.