Lecture Notes for Stat Øk 2

Thomas Mikosch

Contents

1. Introduction

In these lecture notes we consider different approaches to modeling *dependence* of discrete time processes. It is common to refer to these processes as *time series*. Time series are of interest to people in very different fields. For example, the meteorologist collects and studies temperature, air pressure and other data as a function of time and location. The financial engineer looks at share prices, exchange rates, interest rates, etc. The actuary considers individual claims in a portfolio. Government offices collect time series about employment, tax income, health status, suicide rates, number of prison cells, and many other features which may or may not be of interest for society. Some pedagogically or otherwise gifted colleagues collect time series about different study and social behavior of male and female students and draw wise conclusions from those numbers. Some of them also collect time series about the number of students over decades and publish those as scientific publication once a year. The astronomer counts sun-spots and measures the brightness of stars. The newspapers confront us every day with time series about records and developments in sports, finance, economy. These few examples show that a *time series analysis* is very much what we need in order to understand what these series represent, what their theoretical backbones are and how we can predict them.

The naive understanding of a time series is that at certain discrete instants of time t a random value X_t (or random vector $\mathbf{X}_t$) is observed. In mathematical language, we are given a family of random variables X_t, a so-called *discrete-time stochastic process*. The notion of *time* is clearly a mathematical one; it can also be understood as *space*, for example in crystallography, where the time t is not important, then t stands for the location, where X_t has been measured.

We may ask what we can expect from time series analysis. First of all we want to look at some theoretical *probabilistic* models which fit sufficiently large classes of real-life data. We will study the dependence structure of these models in terms of covariances and correlations of the X_t's. This is the so-called *time domain* of time series analysis. An equivalent approach is provided in the *frequency domain*, where one studies a time series as superposition of random sinusoids, i.e., deterministic trigonometric functions with a random amplitude. This helps one to determine the dominating periodicities in a time series.

We will study some of the important parametric time series models, including ARMA, FARIMA, GARCH, and stochastic volatility models. To fit a parametric model means to estimate its parameters from data. Thus statistical estimation techniques are of importance in time series analysis. We will discuss some of the important estimation techniques of these parametric models.

After fitting a model to data it can be of interest to forecast future (still unobserved) values of the time series by using the structure of the fitted model. This can be understood as a physical prediction of these values or as a study of the distribution of future values, their extremal behavior (e.g. maxima, records, exceedances), their periodicity, etc.

The literature on time series is vast and increases rapidly. A special *Journal of Time Series Analysis* exists and dozens of monographs have been written about the topic. Closest in spirit for our purposes are the textbooks Brockwell and Davis [8, 9]; other relevant sources will be cited in the corresponding sections.

Classical time series analysis mostly deals with the second order structure of a time series, i.e., with its correlation or covariance structure. This is sufficient if the underlying model for (X_t) is a Gaussian stochastic process since the dependence structure of a mean zero Gaussian process is determined by its second order structure. However, there are many time series of interest, where Gaussianity is not a reasonable assumption, such as for financial or insurance data, where one often has heavy-tailed marginal distributions. Then correlations and covariances describe the dependence in an insufficient way.

Over the past few years, various alternative measures of dependence have been developed. As a matter of fact, there is nothing like a unique quantitative description of dependence, as for

Gaussian X_t's via the covariances. For non-Gaussian sequences, one can only describe certain aspects of dependence in a suitable way. One of them is *extremal dependence*. We will briefly touch on how this kind of dependence can be described and estimated in a time series. In particular, we will have a look at the *extremogram* which is taylored for the extremes in a time series and motivated by the classical autocorrelation function.

2. Stationary processes

2.1. **Autocovariance and autocorrelation function, stationarity.** Time series will be modeled as a *stochastic process*.

Definition 2.1. (Stochastic process)
A stochastic process *is a family of real-valued random variables* $(X_t)_{t\in T}$ *defined on a probability space* $[\Omega, \mathcal{F}, P]$. *The functions* $(X.(\omega))_{\omega\in\Omega}$ *are called* realizations *or* trajectories *or* sample paths *of the process.*

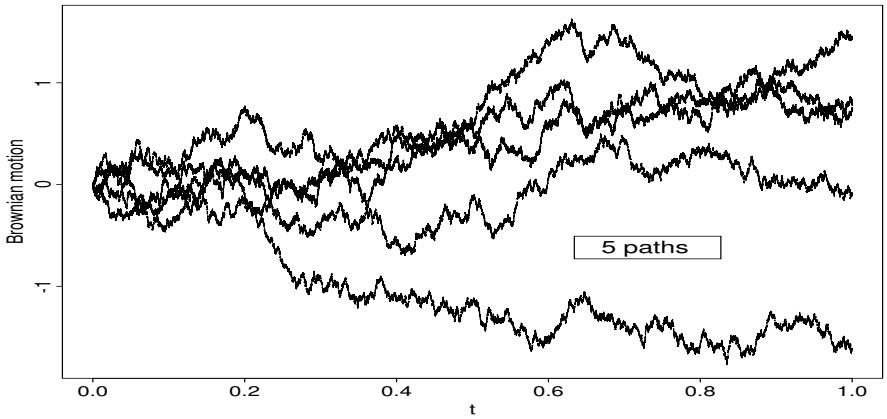

Figure 2.2. *Realizations of a stochastic process* (*Brownian motion*) *on* $[0, 1]$.

Thus a stochastic process is a function of two variables ω and t. It is common to refer to t as *time* even if it does not have this meaning. Throughout we will consider subsets $T \subset \mathbb{R}$. In most cases we will assume that $T = \mathbb{Z}$ or $T = \mathbb{N}$. The notion *time series* will be used in a twofold way: as a stochastic process with a discrete index set (mostly $T = \mathbb{Z}$) and as the observations on such a stochastic process. This means we use the name "time series" for both the process and its realization (or observation).

When looking at a time series we hope to see some sort of "regularity". In particular, when looking at different segments of the series we might expect to discover similar patterns or similar behavior. This can be made precise by introducing the notion of "stationarity". Before we can do that we need a fundamental tool:

Definition 2.3. (Autocovariance function)
Let $(X_t)_{t\in T}$ *be a process such that* $\mathrm{var}(X_t) < \infty$ *for all* $t \in T$. *The function*

$$\gamma_X(s,t) = \mathrm{cov}(X_s, X_t) = E[(X_s - EX_s)(X_t - EX_t)]\,, \quad s,t \in T\,,$$

is called the autocovariance function *of the process* (X_t).

Definition 2.4. (Stationary process)
The time series $(X_t)_{t\in\mathbb{Z}}$ *is said to be* stationary *if the following relations hold:*

(1) $E|X_t|^2 < \infty\,, t \in \mathbb{Z}$.
(2) $EX_t = m\,, t \in \mathbb{Z}$, *for a constant* m.
(3) $\gamma_X(s,t) = \gamma_X(s+h, t+h)$ *for all* $s,t,h \in \mathbb{Z}$.

In the literature stationarity is sometimes called *second order stationarity*, *covariance stationarity*, *stationarity in the wide sense*, *weak stationarity*.

If (X_t) is stationary we have

$$\gamma_X(s,t) = \gamma_X(0, t-s) = \mathrm{cov}(X_0, X_{t-s}) \,, \quad s,t \in \mathbb{Z} \,.$$

For this reason we redefine the autocovariance function as

$$\gamma_X(h) \equiv \gamma_X(0,h) \,, \quad h \in \mathbb{Z} \,.$$

We also introduce a normalized autocovariance function, the *autocorrelation function* of (X_t),

$$\rho_X(h) = \frac{\gamma_X(h)}{\gamma_X(0)} = \frac{\mathrm{cov}(X_h, X_0)}{\mathrm{var}(X_0)} = \mathrm{corr}(X_h, X_0) \,, \quad h \in \mathbb{Z} \,.$$

Autocovariances, autocorrelations and their sample versions are relevant for the study of the dependence structure and for building theoretical time series models. Whenever we work with these quantities we are in the *time domain* of time series analysis. Another way of looking at time series is the *frequency domain* where one studies the spectral properties of such series. We will come to this topic in Section 6.

2.2. **Examples of stationary processes.** We continue with some simple examples of stationary processes.

Example 2.5. (White noise process)
Let (X_t) be iid finite variance random variables. Then, obviously, $m = EX_t = EX_0$ and

$$\gamma_X(s,t) = \mathrm{cov}(X_s, X_t) = \mathrm{cov}(X_{s+h}, X_{t+h}) = \begin{cases} 0 & s \neq t \\ \mathrm{var}(X_0) & s = t \,. \end{cases}$$

Sequences of mean–zero, finite variance iid random variables are particular cases of *white noise processes*: a stochastic process $(X_t)_{t\in\mathbb{Z}}$ is *white noise* if

(1) $EX_t = 0$ for all $t \in \mathbb{Z}$,
(2) $E(X_t X_s) = 0$ for all $s \neq t$, i.e., X_t and X_s are uncorrelated,
(3) $EX_t^2 = \sigma^2$ for a finite constant σ^2, all $t \in \mathbb{Z}$.

Obviously, every white noise process is stationary.

The name "white noise" is used in a different way in general stochastic process theory (e.g. as a generalized derivative of Brownian motion). Property 2 of white noise is often called (*pairwise*) *orthogonality* of the process (X_t). Orthogonal functions are intensively studied in Fourier analysis.

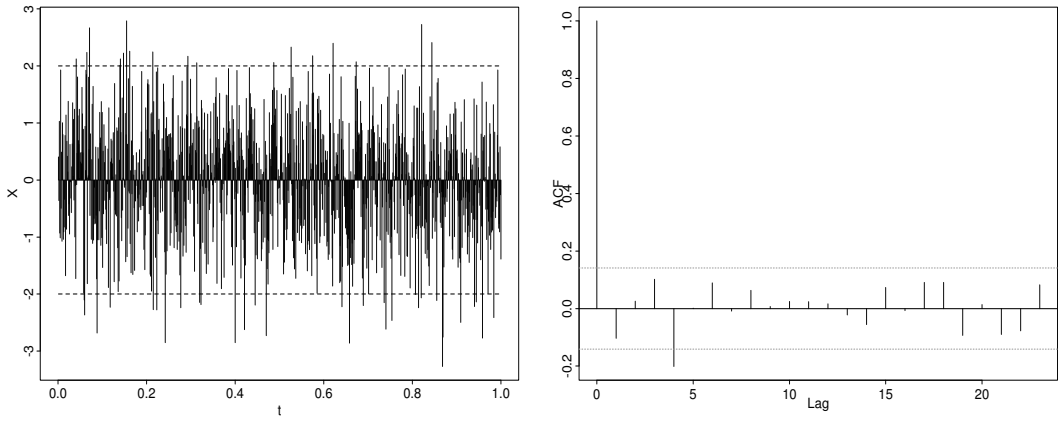

Figure 2.6. *IID* (*standard Gaussian*) *white noise* (left) *and its estimated autocorrelation function* (right). *The dotted vertical lines indicate* 95% *asymptotic confidence bands for the estimators.*

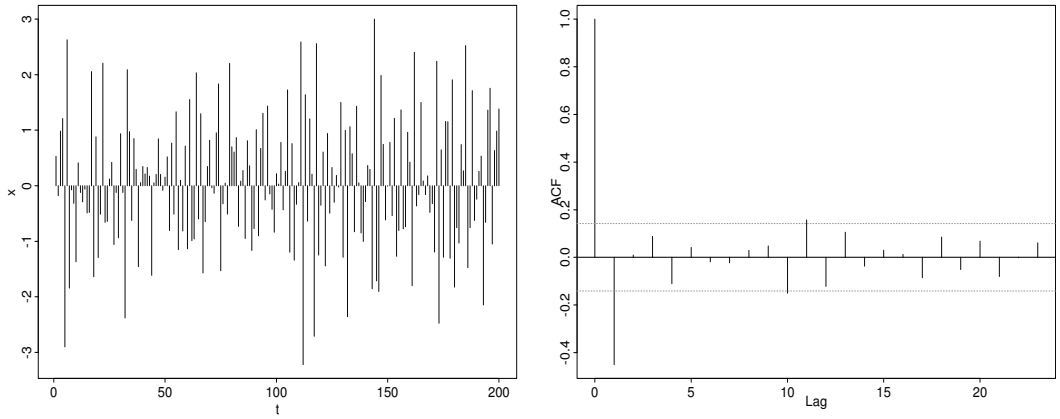

Figure 2.9. *Sample path of the* MA(1) *process* $X_t = Z_t - 0.5Z_{t-1}$ *for iid standard Gaussian white noise* (Z_t) *and its estimated autocorrelation function. The vertical lines indicate* 95% *asymptotic confidence bands for the case of an iid standard Gaussian white noise sequence.*

Exercise 2.7. Consider an iid sequence (W_i) of standard normal random variables and define the time series

$$X_1 = \frac{W_1 + W_2}{\sqrt{2}}, X_2 = \frac{W_1 - W_2}{\sqrt{2}}, X_3 = \frac{W_3 + W_4}{\sqrt{2}}, X_4 = \frac{W_3 - W_4}{\sqrt{2}}, \dots,$$
$$X_1 = \text{sign}(W_2)|W_1|, X_2 = \text{sign}(W_1)|W_2|, X_3 = \text{sign}(W_4)|W_3|, X_4 = \text{sign}(W_3)|W_4|, \dots,$$
$$X_1 = \text{sign}(W_1)|W_1|, X_2 = \text{sign}(W_2)|W_1|, X_3 = \text{sign}(W_3)|W_1|, \dots.$$

Which of these time series models is white noise and which of them does not consist of independent random variables?
Hint: For an iid sequence (Y_i) of symmetric random variables the sequences $(\text{sign}(Y_i))$ and $(|Y_i|)$ are independent.

Example 2.8. (Moving average process with white noise)
Let $(Z_t)_{t\in\mathbb{Z}}$ be white noise, $\theta_1, \dots, \theta_q$ be real numbers for some $q \geq 1$. The stochastic process $(X_t)_{t\in\mathbb{Z}}$ defined as

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

is called a *moving average of order* q (MA(q) *process*). It is a stationary process.

Example 2.11. (Sinusoid)
Let $X_t = A\cos(\theta t) + B\sin(\theta t)$, where A, B are random variables such that $EA = EB = E(AB) = 0$ and $EA^2 = EB^2 = 1$, $\theta \in [-\pi, \pi]$. We have $EX_t = 0$, $\text{var}(X_t) = 1$ and

$$\begin{aligned}\text{cov}(X_{t+h}, X_t) &= E[(A\cos(\theta(t+h)) + B\sin(\theta(t+h)))(A\cos(\theta t) + B\sin(\theta t))] \\ &= \cos(\theta t)\cos(\theta(t+h)) + \sin(\theta t)\sin(\theta(t+h)) \\ &= \cos(\theta h),\end{aligned}$$

which is independent of t and hence (X_t) is stationary.

The example of a sinusoid looks artificial. Indeed, the randomness in a sinusoid (X_t) comes in only through the random variables A and B. Suppose we choose A, B at time zero, the values X_t for $t > 0$ are completely determined by a deterministic recurrence equation. A sinusoid is an example of a *deterministic time series*.

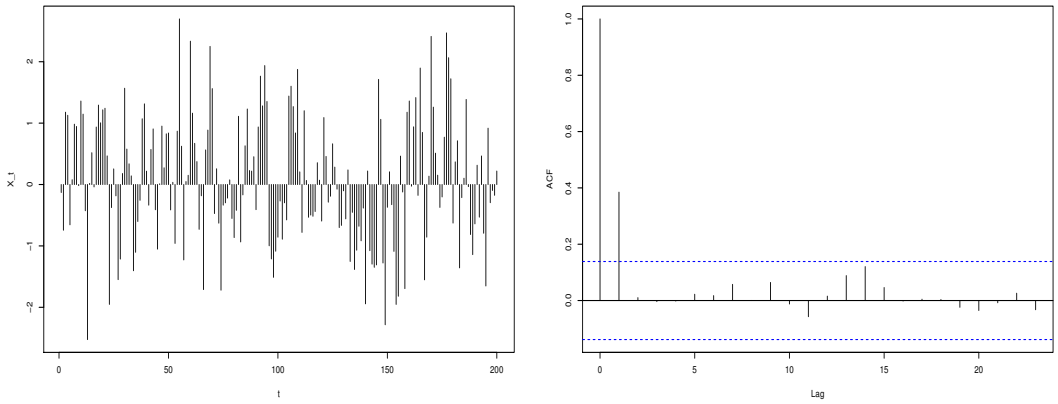

Figure 2.10. *Sample path of the* MA(1) *process* $X_t = Z_t + 0.5Z_{t-1}$ *for iid standard Gaussian white noise* (Z_t) *and its estimated autocorrelation function. The vertical lines indicate* 95% *asymptotic confidence bands for the case of an iid standard Gaussian white noise sequence.*

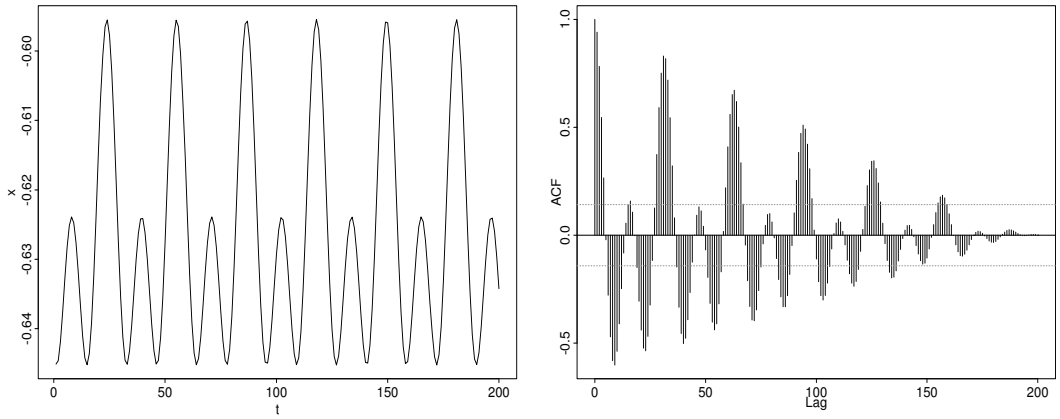

Figure 2.12. *A sample path of the process* $X_t = A\cos(0.4t) + B\sin(0.4t)$ *from Example 2.11* (left) *and estimated autocorrelation function of the process* $X_t = A\cos(0.4t) + B\sin(0.4t)$.

2.3. **Strict stationarity, Gaussian stationary processes.** Many processes we will consider are stationary in a much stronger sense.

Definition 2.13. (Strict stationarity)
The time series $(X_t)_{t\in\mathbb{Z}}$ *is said to be* strictly stationary *if for any* $h \in \mathbb{Z}$ *and* $t \geq 0$*, the random vectors* $(X_h, \ldots, X_{t+h})$ *and* $(X_0, \ldots, X_t)$ *have the same distribution.*

In particular, for a strictly stationary process (X_t) the distribution functions of $(X_0, \ldots, X_t)$ and $(X_h, \ldots, X_{t+h})$ coincide:

$$P(X_h \leq x_0, \ldots, X_{t+h} \leq x_t) = P(X_0 \leq x_0, \ldots, X_t \leq x_t)\,, \quad x_i \in \mathbb{R}\,, \quad i = 1, \ldots, t\,.$$

This is one way to check strict stationarity. Another one is to check whether the characteristic functions (or moment generating functions if they exist) of $(X_0, \ldots, X_t)$ and $(X_h, \ldots, X_{t+h})$ are the same:

$$E\mathrm{e}^{\,i(\lambda_0 X_h + \cdots + \lambda_t X_{t+h})} = E\mathrm{e}^{\,i(\lambda_0 X_0 + \cdots + \lambda_t X_t)}\,, \quad \lambda_i \in \mathbb{R}\,, \quad i = 1, \ldots, t\,.$$

A simple way of constructing a strictly stationary time series (Y_t) from a given strictly stationary time series (X_t) is the following.

Proposition 2.14. *Consider a deterministic real-valued function g acting on $\mathbb{R}^m$ for some $m \geq 1$ and a strictly stationary sequence $(X_t)_{t\in\mathbb{Z}}$. Then the time series*

$$Y_t = g(X_t, \ldots, X_{t-m+1}), \qquad t \in \mathbb{Z},$$

is strictly stationary.

Proof. In view of the strict stationarity of (X_t) it is easy to see the following identity of the distributions

$$\big((X_t, \ldots, X_{t-m+1})\big)_{t=1,\ldots,n} \stackrel{d}{=} \big((X_{t+h}, \ldots, X_{t+h-m+1})\big)_{t=1,\ldots,n}$$

for any $h \in \mathbb{Z}$ and $n \geq 1$. Hence

$$(Y_t)_{t=1,\ldots,n} = \big(g(X_t, \ldots, X_{t-m+1})\big)_{t=1,\ldots,n} \stackrel{d}{=} \big(g(X_{t+h}, \ldots, X_{t+h-m+1})\big)_{t=1,\ldots,n} = (Y_{t+h})_{t=1,\ldots,n}$$

But this is the defining property of strict stationarity. □

Example 2.15. Let (X_t) be strictly stationary. Consider the function

$$g(x_0, \ldots, x_q) = x_0 + \theta_1 x_1 + \cdots + \theta_q x_q,$$

for given real numbers $\theta_1, \ldots, \theta_q$. Then

$$Y_t = g(X_t, \ldots, X_{t-q}) = X_t + \theta_1 X_{t-1} + \cdots + X_{t-q}, \qquad t \in \mathbb{Z},$$

constitutes an MA(q) process driven by (X_t). The new process is again strictly stationary. Hence a moving average process of a strictly stationary process is strictly stationary. Of course, an iid sequence is strictly stationary, hence a moving average process of an iid sequence is strictly stationary.

Other simple time series models are given by functions g of one variable, e.g. $Y_t = |X_t|, Y_t = X_t^2$. Since returns (X_t) of speculative prices are often uncorrelated it is common to look at the autocorrelation functions of their absolute values and squares. The functions $g(x_1, x_2) = x_1 x_2$, $g(x_1, x_3), \ldots$ yields the strictly stationary time series $Y_t = X_t X_{t+1}$, $Y_t = X_t X_{t+2}, \ldots$, $t \in \mathbb{Z}$ which appear in the definitions of the sample autocovariance and autocorrelation functions.

Proposition 2.14 is often not sufficient for our purposes. For example, an autoregressive process of order 1 is given by the difference equation

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

where (Z_t) is iid white noise and $\phi \in (-1, 1)$. The unique solution to this equation is

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

The right-hand side is given by an infinite series which is a deterministic function $g(Z_t, Z_{t-1}, \ldots)$ involving the infinite sequence $(Z_j)_{j\leq t}$. It is not obvious that this infinite series is meaningful; we would have to check whether it defines a finite random variable X_t for every t.

Proposition 2.16. *Consider a deterministic real-valued function g acting on $\mathbb{R}^{\mathbb{Z}}$ for some strictly stationary sequence $(X_t)_{t\in\mathbb{Z}}$. Then the time series*

$$Y_t = g(X_t, X_{t-1}, \ldots), \qquad t \in \mathbb{Z}, \tag{2.1}$$

is strictly stationary provided the right-hand side is finite with probability 1 for all t.

The proof is analogous to the one of Proposition 2.14; you are encouraged to check the arguments. In what follows, we will encounter several examples of time series models which are of the form (2.1).

Remark 2.17. We also mention that much of the theory in these notes can be extended to the even more general models

$$Y_t = g(\dots, X_{t+1}, X_t, X_{t-1}, \dots), \qquad t \in \mathbb{Z},$$

i.e., a deterministic function is acting on the sequence $(X_s)_{s\in\mathbb{Z}}$, involving at time t the past and present $(X_s)_{s\le t}$ but also the future $(X_s)_{s>t}$ of the strictly stationary sequence $(X_s)_{s\in\mathbb{Z}}$. For example, a two-sided infinite moving process

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

with iid white noise (Z_t) is of this form with

$$g\big(\dots, x_{-1}, x_0, x_1, \dots\big) = \sum_{j=-\infty}^{\infty} \psi_j x_j .$$

However, such processes are often considered unnatural because they involve future values of a time series.

Remark 2.18. By strict stationarity, we have $X_t \stackrel{d}{=} X_0$ for any t. Therefore the right-hand side in (2.1) is finite a.s. for any t if and only if it is finite for $t = 0$.

In what follows, we study the relationship between stationarity and strict stationarity. For a stationary process we only require that $\mathrm{var}(X_t) < \infty$ for all t and both the expectation EX_t and the autocovariance function $\mathrm{cov}(X_t, X_s)$ be shift-invariant. Therefore the following statement is not surprising.

Proposition 2.19. *A finite-variance strictly stationary time series $(X_t)_{t\in\mathbb{Z}}$ is stationary.*

Proof. For a strictly stationary process (X_t) the distributions of X_h and X_0 coincide. Hence EX_t and $\mathrm{var}(X_t)$ do not depend on t. Moreover, strict stationarity implies that the pairs (X_t, X_{t+h}) have the same distribution as (X_0, X_h) for every t and $h \ge 0$. Thus $\mathrm{cov}(X_t, X_{t+h})$ does not depend on t. □

Remark 2.20. In general, strict stationarity of a time series (X_t) is a more restrictive property than stationarity since one needs knowledge about the distributions of all vectors $(X_t, \dots, X_s)$, $t \le s$. On the other hand, strict stationarity goes beyond stationarity in the sense that it includes time series whose marginal distributions have infinite variance. There is a strong belief based on statistical evidence that most real-life time series of interest have finite variance, in particular most return series of speculative prices. However, infinite variance time series are not uncommon for returns of electricity prices, for reinsurance claims and earthquake magnitudes; see, for example, Adler et al. [1], Embrechts et al. [13], and Garcia et al. [17].

The converse to Proposition 2.19 is in general not true. For example, white noise processes are in general not strictly stationary. However, there is the important exception of the Gaussian processes. Before we define them recall the notion of *multivariate Gaussian vector* and *multivariate Gaussian distribution*: the random vector $\mathbf{G} = (G_1, \dots, G_n)'$[1] is (non-degenerate) Gaussian if it has density

$$f_{\mathbf{G}}(\mathbf{y}) = \frac{1}{(2\pi)^{n/2}(\det \Sigma_{\mathbf{G}})^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{y}-\mu_{\mathbf{G}})'\Sigma_{\mathbf{G}}^{-1}(\mathbf{y}-\mu_{\mathbf{G}})\right\}, \quad \mathbf{y} \in \mathbb{R}^n . \tag{2.2}$$

Here $\mu_{\mathbf{G}} = E\mathbf{G} = (EG_1, \dots, EG_n)'$ and $\Sigma_{\mathbf{G}} = (\mathrm{cov}(G_i, G_j))_{i,j=1,\dots,n}$, where the latter is supposed to be non-singular. Alternatively, a multivariate Gaussian vector is uniquely determined by its

[1] For a vector $\mathbf{a}$ we write $\mathbf{a}'$ for its transpose.

characteristic function

$$E\mathrm{e}^{\,i\mathbf{u}'\mathbf{G}} = \exp\left\{ i\mathbf{u}'\mu_{\mathbf{G}} - \frac{1}{2}\mathbf{u}'\Sigma_{\mathbf{G}}\,\mathbf{u} \right\}, \quad \mathbf{u} \in \mathbb{R}^n . \tag{2.3}$$

The latter definition also allows for a singular covariance matrix Σ_G.

Example 2.21. (Bivariate normal distribution)
It has density

$$\begin{aligned} & f_{(G_1,G_2)}(y_1,y_2) \\ = \quad & \frac{1}{2\pi}\frac{1}{\sqrt{(1-\rho^2)\sigma_1^2\sigma_2^2}} \exp\left\{ -\frac{1}{2} \begin{pmatrix} y_1-\mu_1 \\ y_2-\mu_2 \end{pmatrix}' \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}^{-1} \begin{pmatrix} y_1-\mu_1 \\ y_2-\mu_2 \end{pmatrix} \right\}, \end{aligned}$$

where $\sigma_i^2 = \mathrm{var}(G_i)$, $\mu_i = EG_i$, $i = 1, 2$, and

$$\rho = \mathrm{corr}(G_1, G_2) = \frac{\mathrm{cov}(G_1, G_2)}{\sqrt{\mathrm{var}(G_1)\mathrm{var}(G_2)}} = \frac{\rho\sigma_1\sigma_2}{\sigma_1\sigma_2} \in [-1, 1] .$$

Definition 2.22. (Gaussian process)
The stochastic process $(X_t)_{t\in T}$ is called Gaussian *if its* finite-dimensional distributions*, i.e., the distributions of the vectors $(X_{t_1}, \ldots, X_{t_m})$ for any choice of $t_1, \ldots, t_m \in T$ and $m \geq 1$, are multivariate Gaussian; see* (2.2) *or* (2.3)*.*

A straightforward consequence of this definition is the following.

Corollary 2.23. *Consider a Gaussian time series $(X_t)_{t\in\mathbb{Z}}$.*

(1) *The finite-dimensional distributions of the process are completely determined by the mean value function $\mu(t) = EX_t$ and the autocovariance function $\gamma_X(s,t) = \mathrm{cov}(X_s, X_t)$ for $s, t \in \mathbb{Z}$.*
(2) *If $(X_t)_{t\in\mathbb{Z}}$ is stationary it is strictly stationary.*

Proof. The first claim is straightforward since the finite-dimensional distributions of (X_t) are determined only by the values of μ and γ_X; see (2.2) or (2.3).

If (X_t) is stationary the random vectors $(X_0, \ldots, X_t)$ and $(X_h, \ldots, X_{t+h})$, $h \in \mathbb{Z}$, have the same expectation vector and covariance matrix for every $t \geq 0$, i.e., they are invariant under shifts h. These vectors are Gaussian and hence, by the first part, the distributions of $(X_0, \ldots, X_t)$ and $(X_h, \ldots, X_{t+h})$, $h \in \mathbb{Z}$, are the same for all $t \geq 0$. This is another way of saying that (X_t) is strictly stationary. □

When dealing with real-life data, one has to decide by looking at the data whether they come from a stationary/strictly stationary model or one has to transform the data in one way or the other "to come closer" to the stationarity assumption. We refer to Section 2.6 for a discussion of suitable transformations. When looking at the data one can be fooled by the appearance of their irregular behavior to conclude that they do not come from a stationary model; see Figure 2.24 for an example.

2.4. **Ergodic time series.** An even more restrictive notion than strict stationarity is *ergodicity*. It is not as easily explained as stationarity or strict stationarity. In an intuitive sense a strictly stationary ergodic time series (X_t) satisfies the strong law of large numbers (called *ergodic theorem* in this context) for all "good" functions acting on it.

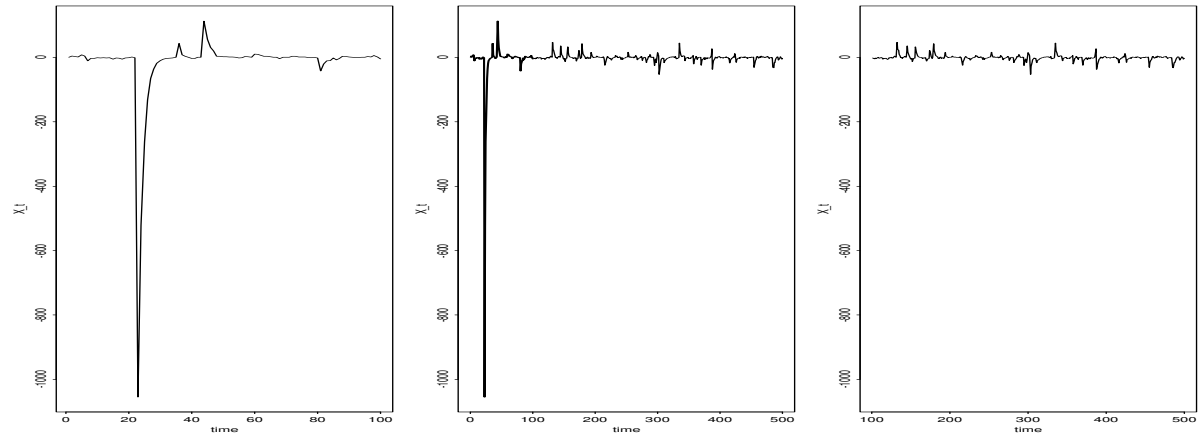

Figure 2.24. *We see a realization of* 500 *values of the strictly stationary* AR(1) *process* $X_t = 0.5X_{t-1} + Z_t$ *with iid Cauchy noise* (Z_t) (*middle*). *The left* (*right*) *graph visualizes the first* 100 (*last* 400) *values. If one looks at the whole series the first piece gives one the impression that the series is not stationary. When the time series in the right graph is analyzed the assumption of stationarity is plausible, while, when in possession of the whole series, a structural break seems more likely.*
The Cauchy distribution has infinite first moment. Therefore a very big value Z_t *may occur once in a while, making* X_t *and some of its successors very big themselves.*

To make this precise consider any real-valued (measurable) deterministic function f on the sequence space $\mathbb{R}^{\mathbb{Z}}$ and define a new time series via *shifts*:

$$\begin{aligned} Y_0 &= f(\ldots, X_{-1}, X_0\,, X_1, \ldots)\,, \\ Y_1 &= f(\ldots, X_0, X_1\,, X_2, \ldots)\,, \\ \cdots &= \cdots \\ Y_t &= f(\ldots, X_{t-1}, X_t\,, X_{t+1}, \ldots)\,, \qquad t \in \mathbb{Z}\,, \end{aligned}$$

and assume that

$$E|Y_0| = E[|f(\ldots, X_{-1}, X_0\,, X_1, \ldots)|] < \infty$$

An example of this kind of time series is the infinite moving average

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

for real-valued (ψ_j). If $E|Y_t| < \infty$ the right-hand series is finite a.s. Simpler examples are $Y_t = f(X_t)$ (e.g. $Y_t = X_t$), $Y_t = X_t + \theta_1 X_{t-1} + \cdots + \theta_q X_{t-q}$, for real values θ_j, $Y_t = X_t X_{t-1}$, $t \in \mathbb{Z}$. The latter time series only involve finitely many elements of (X_s).

Theorem 2.25. *A strictly stationary time series* $(X_t)_{t\in\mathbb{Z}}$ *is* ergodic *if and only if it satisfies the* ergodic theorem*: for a time series* (Y_t) *constructed from any real-valued function* f *by shifts (see above) with the property* $E|Y_0| < \infty$ *the strong law of large numbers holds:*

$$\frac{1}{n}\sum_{t=1}^{n} Y_t \overset{\text{a.s.}}{\to} EY_0\,. \tag{2.4}$$

Remark 2.26. There are many stationary processes (X_t) that satisfy the strong law of large numbers

$$\frac{1}{n}\sum_{t=1}^{n} X_t \stackrel{\text{a.s.}}{\to} EX_0\,.$$

(e.g. the sinusoid in Example 2.11) or even the strong law of large numbers for $(f(X_t))$ for various functions f. This does not mean that (X_t) is ergodic. Relation (2.4) is a much stronger requirement: the strong law of large numbers must hold for any time series constructed as a function of shifts of (X_t), not only for special functions.

Remark 2.27. Showing ergodicity of a concrete strictly stationary time series is not easy. However, one can show that an iid sequence (Z_t) is ergodic and therefore one can construct many interesting new ergodic time series from (Z_t) by using shifts.

Remark 2.28. Many parameter estimators in statistics and time series analysis have the form of a sample average of independent or dependent observations. In this context, the ergodic theorem (2.4) is utterly useful for proving strong consistency of these estimators.

A general reference to ergodic theory is Krengel [24]. The recent textbook of Samorodnitsky [31] is very accessible and gives a good introduction to stationarity, ergodicity, and related topics. The proof of Theorem 2.25 can be found in Section 2.1 in [31].

Not every strictly stationary time series is ergodic. A simple example is the time series $X_t = X$, $t \in \mathbb{Z}$. Assuming a finite expectation for X we have

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

This means that the "limit" of the sample mean is random, not the expected value of X as required by the strong law of large numbers.

In Proposition 2.1.6 of [31] one can find a more general result which characterizes a strictly stationary *non-ergodic* time series.

Proposition 2.29. *A strictly stationary process is non-ergodic if and only if there is a probability space supporting two strictly stationary processes (Y_n) and (Z_n) with different finite-dimensional distributions, and a Bernoulli(p) distributed random variable B with $0 < p < 1$ independent of them such that*

$$(X_n) = \begin{cases} (Y_n) & \text{with probability } p, \\ (Z_n) & \text{with probability } 1-p. \end{cases}$$

One can think of flipping a coin whose two outcomes 1 and 0 correspond to the Bernoulli random variable B and appear with probabilities p and $1-p$, respectively. If the outcome 1 appears you choose the sequence (Y_n), otherwise (Z_n). This means that a non-ergodic strictly stationary process is a *mixture of two strictly stationary processes.*

In Corollary 2.1.8 of [31] one finds the following very useful result.

Theorem 2.30. (Functions of shifts of ergodic sequences yield ergodic sequences)
Let (X_t) be a strictly stationary ergodic time series and f be a real-valued (measurable) function on $\mathbb{R}^{\mathbb{Z}}$. Then the sequence of shifts

$$Y_t = f(\ldots, X_{t-1}, X_t\,, X_{t+1}, \ldots)\,, \qquad t \in \mathbb{Z}\,, \tag{2.5}$$

constitutes a strictly stationary ergodic sequence provided the right-hand side is finite with probability 1.

Example 2.31. (MA(q) process) The simplest example of a strictly stationary ergodic sequence is an iid sequence (Z_t). Any sequence (Y_t) of the form (2.5) (replacing (X_t) by (Z_t)) defined on the shifts of (Z_t) is ergodic. In particular, any MA(q) process $X_t = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}$, $t \in \mathbb{Z}$, for real-valued θ_i, is ergodic, but also the infinite moving average

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

for real (ψ_j) if we can ensure that the right-hand side is finite a.s.

In what follows, we will work with various ergodic time series models which are obtained as functions acting on the shifts of an iid sequence (Z_t) and involving infinite subsets of (Z_t). Those models include the autoregressive processes and the GARCH process.

Example 2.32. (Sample mean and sums of products of lagged X_t's satisfy the strong law of large numbers)
We conclude from (2.4) that

$$\overline{X}_n = \frac{1}{n} \sum_{t=1}^{n} X_t \stackrel{\text{a.s.}}{\to} EX_0 \,,$$

if $E|X_1| < \infty$ and for any $h \geq 0$,

$$\frac{1}{n} \sum_{t=1}^{n} X_t \, X_{t+h} \stackrel{\text{a.s.}}{\to} E(X_0 \, X_h) \,,$$

if $E|X_0 X_h| < \infty$.

Exercise 2.33. Let (X_t) be a strictly stationary ergodic sequence with finite variance. Show that for every fixed $h \geq 0$, the sample autocovariances

$$\gamma_{n,X}(h) = \frac{1}{n} \sum_{t=1}^{n-h} (X_t - \overline{X}_n) \, (X_{t+h} - \overline{X}_n)$$

and the sample autocorrelations

$$\rho_{n,X}(h) = \frac{\gamma_{n,X}(h)}{\gamma_{n,X}(0)}$$

are consistent estimators of their deterministic counterparts:

$$\gamma_{n,X}(h) \stackrel{\text{a.s.}}{\to} \gamma_X(h) \quad \text{and} \quad \rho_{n,X}(h) \stackrel{\text{a.s.}}{\to} \rho_X(h) \,.$$

Exercise 2.34. Consider the daily S&P 500 return data from 1928 until 25 May 2000 from the website www.math.ku.dk/∼mikosch/Time. Plot the annual sample means and variances of the returns and their absolute values. A business year has about 250 days. Does this series make the impression of an ergodic time series?

2.5. **Mixing and strong mixing of strictly stationary processes.** Ergodicity of a strictly stationary time series (X_t) ensures that parameter estimators of the underlying model are consistent, i.e., converge (a.s. or in probability) to the parameter of interest. For the construction of confidence bands around these estimators one needs central limit theorems. In this case, one needs stronger assumptions on (X_t) than ergodicity. Such conditions run under the name *mixing* and constitute some kind of *asymptotic independence condition* in a time series. This can be seen from

the definition of the *mixing property* of a strictly stationary sequence: for any "good" Borel sets A, B in $\mathbb{R}^{\mathbb{Z}}$,

$$\begin{aligned} &P\big((\ldots, X_{-1}, X_0, X_1, \ldots) \in A\,, (\ldots, X_{n-1}, X_n, X_{n+1}, \ldots) \in B\big) \\ &\to \quad P\big((\ldots, X_{-1}, X_0, X_1, \ldots) \in A\big)\, P\big((\ldots, X_{-1}, X_0, X_1, \ldots) \in B\big)\,, \qquad n \to \infty\,. \end{aligned} \tag{2.6}$$

To make this notion more accessible we quote Theorem 2.2.7 from [31]:

Theorem 2.35. *A strictly stationary time series (X_t) is mixing if and only if for every $k \geq 1$,*

$$\big(X_1, \ldots, X_k, X_{n+1}, \ldots, X_{n+k}\big) \xrightarrow{d} \big(X_1, \ldots, X_k, X_1', \ldots, X_k'\big)\,, \qquad n \to \infty\,, \tag{2.7}$$

where (X_t') is an independent copy of (X_t).

In words: the further the two vectors $(X_1, \ldots, X_k)$ and $(X_{n+1}, \ldots, X_{n+k})$ are apart from each other in time the weaker is the dependence between them, and "in the limit" they are independent.

Relation (2.7) follows directly from (2.6). Indeed, choose

$$A = \cdots \times \mathbb{R} \times \mathbb{R} \times (-\infty, x_1] \times \cdots \times (-\infty, x_k] \times \mathbb{R} \times \mathbb{R} \times \cdots \tag{2.8}$$

$$B = \cdots \times \mathbb{R} \times \mathbb{R} \times (-\infty, x_1'] \times \cdots \times (-\infty, x_k'] \times \mathbb{R} \times \mathbb{R} \times \cdots\,, \tag{2.9}$$

for any real numbers x_i, x_i', $i = 1, \ldots, k$, where the subscript i indicates the ith coordinate. Then (2.6) turns into

$$\begin{aligned} &P\big(X_1 \leq x_1, \ldots, X_k \leq x_k, X_{n+1} \leq x_1', \ldots, X_{n+k} \leq x_k'\big) \\ &\to \quad P\big(X_1 \leq x_1, \ldots, X_k \leq x_k\big)\, P\big(X_1 \leq x_1', \ldots, X_k \leq x_k'\big)\,, \quad n \to \infty\,, \end{aligned} \tag{2.10}$$

provided we choose $(x_1, \ldots, x_k)$ and $(x_1', \ldots, x_k')$ as continuity points of the distribution function of $(X_1, \ldots, X_k)$. But this just means convergence in distribution of the left-hand side in (2.7) to a random vector whose first and last k components have the distribution of $(X_1, \ldots, X_k)$ and the first k and last k components are independent of each other.

The mixing property is not easily established for particular models. Exceptions are iid sequences and moving averages acting on them.

Example 2.36. For an iid sequence (Z_t) relation (2.10) (with (X_t) replaced by (Z_t)) holds with equality of the left- and right-hand sides for $n > k$. Now consider a time series of the form

$$X_t = g(Z_t, \ldots, Z_{t-m+1})\,, \qquad t \in \mathbb{Z}\,, \tag{2.11}$$

for some real-valued (measurable) function g and $m \geq 1$. The random vectors $(X_1, \ldots, X_k)$ and $(X_{n+1}, \ldots, X_{n+k})$ are independent if $n \geq k + m - 1$ and therefore (2.10) holds with equality for large n. The model (2.11) contains the moving averages $X_t = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_m Z_{t-m+1}$. One can show that infinite moving averages of an iid sequence (Z_t), i.e.,

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

are mixing as well; see Theorem 2.38. This class of processes plays a major role in these notes.

Remark 2.37. By construction of the time series (2.11), the infinite sequences $(\ldots, X_{-1}, X_0)$ and $(X_m, X_{m+1}, \ldots)$ are independent. More generally, by strict stationarity, $(\ldots, X_{s-1}, X_s)$ and $(X_t, X_{t+1}, \ldots)$, $s < t$, are independent if $t - s > m$. Such a strictly stationary time series is called *m-dependent*. m-dependent sequences are not necessarily of the form (2.11), i.e., not necessarily functions of independent random variables. Strictly stationary m-dependent sequences are mixing.

As ergodicity is inherited by time series which are a function acting on the shifts of an ergodic time series, so mixing is inherited from mixing processes acting on shifts. The following is Corollary 2.2.5 of Samorodnitsky [31].

Theorem 2.38. (Functions of shifts of mixing sequences yield ergodic sequences)
Let (X_t) be a strictly stationary mixing time series and f be a real-valued (measurable) function on $\mathbb{R}^{\mathbb{Z}}$. Then the sequence of shifts

$$Y_t = f(\dots, X_{t-1}, X_t, X_{t+1}, \dots), \qquad t \in \mathbb{Z},$$

constitutes a strictly stationary mixing sequence provided the right-hand side is finite with probability 1.

Remark 2.39. A deeper analysis of mixing sequences shows that they are always ergodic. There exist ergodic sequences which are not mixing; see [31] for examples.

Our goal was to ensure that a strictly stationary time series (X_t) satisfies the central limit theorem. Unfortunately, mixing is not enough for this; one also needs to control the rate of convergence in (2.6), i.e., how fast two pieces of a time series become independent when the distance in time between them increases. This was discovered by Murray Rosenblatt [30] in 1956.

We introduce the *strong mixing coefficients* or *mixing rate function* $(\alpha_n)_{n\geq 1}$:

$$\alpha_n = \sup_{C\in\mathcal{F}^X_{-\infty,0}, D\in\mathcal{F}^X_{n,\infty}} \big|P(C\cap D) - P(C)\,P(D)\big|, \tag{2.12}$$

where $\mathcal{F}^X_{k,l} = \sigma(X_k, \dots, X_l)$ is the Borel σ-field generated by the random variables $X_k, \dots, X_l$. So $\mathcal{F}^X_{-\infty,0}$ is the Borel σ-field generated by the "past" and "present" $(\dots, X_{-1}, X_0)$ (given that $t = 0$ is "now") and $\mathcal{F}^X_{n,\infty}$ is the corresponding σ-field representing the "future", starting at $t = n$.

The strictly stationary sequence (X_t) is *strongly mixing* if $\alpha_n \to 0$ as $n \to \infty$, i.e., the past/present and future of the time series become asymptotically independent in a uniform fashion at the rate (α_n).

Strong mixing is indeed more restrictive than mixing; cf. (2.6). This can be seen from the fact that

$$P(C\cap D) = P\big((\dots, X_{-1}, X_0) \in A, (X_n, X_{n+1}, \dots) \in B\big)$$

for suitable Borel sets $A, B \subset \mathbb{R}^{\mathbb{Z}}$. In particular, one can choose A, B as in (2.8), (2.9) and obtains (2.10) by stationarity and since $\alpha_n \to 0$.

Notice that

$$P(C\cap D) - P(C)\,P(D) = E[I_C I_D] - EI_C\,EI_D = \operatorname{cov}(I_C, I_D),$$

hence strong mixing is about the convergence rate of the indicator functions of the events C and D whose distance in time increases to infinity. Using approximations of bounded functions by linear combinations, one can show that

$$\sup_{\|f\|,\|g\|\leq 1} |\operatorname{cov}(f, g)| = \alpha_n, \tag{2.13}$$

where f is any (measurable) function acting on $\dots, X_{-1}, X_0$, g is any (measurable) function acting on $X_n, X_{n+1}, \dots$, and both have supremum norm bounded by 1; see Doukhan [11].

One of the main problems for proving a central limit theorem for the sample mean $\overline{X}_n$ and related estimators like the sample autovariances $\gamma_{n,X}(h)$ of a time series (X_t) is to control their

variances. First assume that (X_t) is only stationary and has mean μ. Then

$$\begin{aligned}\operatorname{var}(\overline{X}_n) &= E\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)\right)^2 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}E[(X_i-\mu)(X_j-\mu)] \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\operatorname{cov}(X_i,X_j) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\gamma_X(i-j) \\ &= \frac{1}{n}\sum_{|h|<n}\left(1-\frac{|h|}{n}\right)\gamma_X(h) \le \frac{1}{n}\sum_{|h|<n}|\gamma_X(h)| \to 0\,,\end{aligned}$$

provided $\gamma_X(h) \to 0$ as $|h| \to \infty$. This condition is satisfied for most reasonable (meaning: used in practice) time series models. By Markov's inequality we conclude that $\overline{X}_n \xrightarrow{P} \mu$. Here we did not need strict stationarity or ergodicity.

For the central limit theorem we need to ensure that $\operatorname{var}(\sqrt{n}\overline{X}_n)$ does not explode. Notice that

$$\begin{aligned}\operatorname{var}(\sqrt{n}\,\overline{X}_n) &= \sum_{|h|<n}\left(1-\frac{|h|}{n}\right)\gamma_X(h) \\ &= \sum_{|h|\le h_0}\gamma_X(h) + \sum_{n>|h|>h_0}\left(1-\frac{|h|}{n}\right)\gamma_X(h) + o(1) \\ &= I_{h_0} + J_{n,h_0} + o(1)\,,\end{aligned}$$

where h_0 is any fixed natural number and $o(1)$ denotes any function that converges to zero as $n \to \infty$. In view of the structure of I_{h_0} we guess that

$$\begin{aligned}\operatorname{var}(\sqrt{n}\,\overline{X}_n) &\to \sum_{|h|<\infty}\gamma_X(h) \\ &= \gamma_X(0) + 2\sum_{h=1}^{\infty}\gamma_X(h)\,, \qquad (2.14)\end{aligned}$$

where we used that $\gamma_X(h) = \gamma_X(-h)$ but we need to show that

$$\lim_{h_0\to\infty}\limsup_{n\to\infty}|J_{n,h_0}| = 0\,.$$

This is not straightforward and requires extra conditions on the rate of decay to zero of the autocovariance function $\gamma_X(h)$ as $h \to \infty$. For example, if γ_X is absolutely summable,

$$|J_{n,h_0}| \le 2\sum_{h>h_0}|\gamma_X(h)| \to 0\,, \qquad h_0 \to \infty\,. \qquad (2.15)$$

If we assume strict stationarity and strong mixing on (X_t) a famous result from the classical book by Ibragimov and Linnik [21] (Theorem 17.2.2) provides a clear relation between the mixing rate (α_n) and the covariances of functions in such a time series.

Theorem 2.40. *Consider a strictly stationary strongly mixing time series (X_t) with rate function (α_h). Let Y and Z be random variables which are (measurable) functions of $(\ldots, X_{t-1}, X_t)$ and $(X_{t+h}, X_{t+h+1}, \ldots)$, respectively, for some $h \ge 1$. Also assume that $E[|Y|^{2+\delta}] + E[|Z|^{2+\delta}] < \infty$ for some $\delta > 0$. Then there exists a constant $c > 0$ (depending only on the distributions of (Y, Z), not on s, t) such that*

$$|\operatorname{cov}(Y,Z)| \le c\,\alpha_h^{\delta/(2+\delta)}\,.$$

Of course, this result immediately applies to $\gamma_X(h) = \mathrm{cov}(X_0, X_h)$:

$$|\gamma_X(h)| \le c\,\alpha_h^{\delta/(2+\delta)}\,, \qquad h \ge 1\,.$$

Now assume that

$$\sum_{h=1}^{\infty} \alpha_h^{\delta/(2+\delta)} < \infty\,. \tag{2.16}$$

Then we may conclude from (2.14) and (2.15) the following result:

Corollary 2.41. *If the mixing rate (α_h) of the strictly stationary time series (X_t) satisfies the growth condition* (2.16) *then*

$$\mathrm{var}\big(\sqrt{n}\overline{X}_n\big) \to \sigma^2 = \gamma_X(0) + 2\sum_{h=1}^{\infty}\gamma_X(h)\,, \tag{2.17}$$

and the right-hand side is finite.

Now the following result is perhaps less mysterious. It was a benchmark in the limit theory for dependent sequences, proved by I.A. Ibragimov in 1962. A proof can be found in Ibragimov and Linnik [21].

Theorem 2.42. (Ibragimov's central limit theorem for strongly mixing sequences) *Let (X_t) be a strictly stationary time series whose strong mixing rate (α_h) satisfies* (2.16) *and $E[|X_0|^{2+\delta}] < \infty$ for some $\delta > 0$. If σ^2 in* (2.17) *is positive then the central limit theorem for the sample mean holds:*

$$P\Big(\frac{\sqrt{n}}{\sigma}\big(\overline{X}_n - EX_0\big) \le x\Big) \to \Phi(x)\,, \qquad n \to \infty\,, \quad \textit{for any real } x,$$

where Φ denotes the standard normal distribution function.

Remark 2.43. The value σ^2 is possibly zero. An example is the strictly stationary sequence $X_t = Z_t - Z_{t-1}$, $t \in \mathbb{Z}$, where (Z_t) is iid with finite variance. In this case, we conclude from (2.17) that $\sqrt{n}\,\overline{X}_n \xrightarrow{P} 0$ as $n \to \infty$. If we also assume $E[|Z_0|^{4+\delta}] < \infty$ for some $\delta > 0$ then the time series (X_t^2) satisfies all conditions of Theorem 2.42 and the central limit theorem applies to it with asymptotic variance $\mathrm{var}(X_0^2) > 0$.

Remark 2.44. The strength of a result like Theorem 2.42 is its generality. It immediately applies to any (measurable) function f acting on X_t, $f(X_t)$, like $|X_t|$, X_t^2, etc. Transformed time series inherit the strong mixing property from (X_t) with the same rate. Indeed, the σ-field $\mathcal{F}_{k,l}^{f(X)}$ generated by $f(X_k), \ldots, f(X_l)$ for $k \le l$ is a sub-σ-field of $\mathcal{F}_{k,l}^X$. Therefore the strong mixing rate $\alpha_h^{f(X)}$ of $(f(X_t))$ is bounded by α_h. This follows directly from the definition of α_h in (2.12). The same argument applies to any (measurable) transformation of the multivariate sequence $(X_t, \ldots, X_{t-k})$ for any fixed $k \ge 0$. In particular, it applies to the functions $f(X_t, X_{t+h}) = X_t X_{t+h}$ and hence the central limit theorem of Theorem 2.42 applies to the sample covariances $\gamma_{n,X}(h)$ with normalization $\sqrt{n}$.

Using the Cramér-Wold device, we can even prove a joint central limit theorem for a vector of sample covariances and correlations.

Corollary 2.45. *Assume the conditions of Theorem 2.42 for (X_t) and that $E[|X_0X_h|^{2+\delta}] < \infty$ for $h = 1, \ldots, m$, some $m \ge 1$ and $\delta > 0$.*

(1) *Then*

$$\sqrt{n}\big(\gamma_{n,X}(1) - \gamma_X(1), \ldots, \gamma_{n,X}(m) - \gamma_X(m)\big) \xrightarrow{d} \mathbf{Y}_m\,,$$

where $\mathbf{Y}_{0,m}$ is a mean-zero Gaussian vector with a complicated covariance structure.

(2) *If also* $E[|X_0|^{4+\delta}] < \infty$ *then we have*

$$\sqrt{n}\big(\gamma_{n,X}(0) - \gamma_X(0), \gamma_{n,X}(1) - \gamma_X(1), \ldots, \gamma_{n,X}(m) - \gamma_X(m)\big) \xrightarrow{d} \mathbf{Y}_{0,m}\,,$$

where[2] $\mathbf{Y}_{0,m} = \mathrm{vec}(Y_0, \mathbf{Y}_m)$ *is a mean-zero Gaussian vector with a complicated covariance structure, and*

$$\sqrt{n}\,\big(\rho_{n,X}(1) - \rho_X(1), \ldots, \rho_{n,X}(m) - \rho_X(m)\big) \xrightarrow{d} \frac{1}{\gamma_X(0)}\big(\mathbf{Y}_{0,m} - Y_0\,\mathbf{R}_X\big)\,.$$

Here

$$\mathbf{R}_X = (\rho_X(1), \ldots, \rho_X(m))\,.$$

How can one prove a central limit theorem for dependent strongly mixing data? There are various ways of doing this. One way is the *method of small and large blocks*. It is applicable to more general dependence concepts than strong mixing.[3]

The main idea of the proof comes from an iid sequence (X_t) with mean zero and variance 1. In this case, we can factorize the characteristic function of $\sqrt{n}\,\overline{X}_n$:

$$\begin{aligned}\phi_n(s) &= E\Big[\exp\Big(i\,s\frac{1}{\sqrt{n}}(X_1 + \cdots + X_n)\Big)\Big] \\ &= \Big(E\exp(isX_1/\sqrt{n})\Big)^n, \qquad s \in \mathbb{R}\,.\end{aligned}$$

Since X_1 has mean zero and variance one a Taylor expansion yields

$$E\exp(isX_1/\sqrt{n}) = 1 - \frac{s^2}{2n}(1 + o(1))\,.$$

Hence $\phi_n(s) \to \exp(-s^2/2)$. The limit is the characteristic function of a standard normal random variable. Convergence of the characteristic functions implies the convergence of the underlying distributions, hence the central limit theorem holds.

For a dependent sequence we try to proceed in a similar way. We assume without loss of generality that X_0 has mean zero. We intend to factorize the characteristic function $\phi_n(s)$ in some way. This is achieved in an asymptotic way. For this reason, we divide the sample $X_1, \ldots, X_n$ into blocks. For the sake of argument, assume that $n = m_n k_n$ for two integer sequences $k_n, m_n \to \infty$, where m_n is the block length and k_n the number of blocks. We will suppress the dependence of m_n and k_n on n:

$$X_1, \ldots, X_m, X_{m+1}, \ldots, X_{2m}, \ldots, X_{(k-1)m+1}, \ldots, X_{km}\,.$$

Write $S_i^{(n)}$ for the sums of the elements in block $i = 1, \ldots, k$. These sums are dependent but we hope that $S_i^{(n)}$ and $S_j^{(n)}$ become asymptotically independent if $|i - j|$ is sufficiently large. However, neighboring blocks are always dependent. For this reason, we chop the first $l = l_n$ elements off each block sum $S_i^{(n)}$:

$$S_i^{(n)} = \sum_{t=(i-1)m+1}^{(i-1)m+l} X_i + \sum_{t=(i-1)m+l+1}^{im} X_i = \underline{S}_i^{(n)} + \overline{S}_i^{(n)}\,.$$

If we assume that $l = l_n = o(m_n)$ (therefore the name small block - large block) and $l_n \to \infty$ slowly, using the strong mixing condition, we can get a bound for the variance (one can proceed as for the

[2]The vec notation just means that we stick the first component $Y_{0,m}$ to $\mathbf{Y}_m$.

[3]As a matter of fact, there exist many other mixing conditions. They can be found in books like Ibragimov and Linnik [21] and Doukhan [11] and focus on different aspects of the dependence structure.

verification of the finiteness of σ^2 above)

$$\mathrm{var}\Big((\underline{S}_1^{(n)} + \cdots + \underline{S}_k^{(n)})/\sqrt{n} \Big) \to 0\,.$$

This means that the normalized sum of all small block sums does not contribute to the distributional limit of the normalized sample mean.

Now it suffices to deal with the sum of the large blocks:

$$T_n = \overline{S}_1^{(n)} + \cdots + \overline{S}_k^{(n)}\,.$$

Each of the block sums $\overline{S}_1^{(n)}$ has the same distribution – by strict stationarity of (X_t). Moreover, indices of X_t's in neighboring blocks i and $i+1$ are at least l_n steps apart. Since $l_n \to \infty$ we hope that even neighboring block sums become asymptotically independent. We again use characteristic functions:

$$\begin{aligned}
\Delta_n &= E\Big[\exp\big(isT_n/\sqrt{n}\big)\Big] - \prod_{j=1}^{k} E\Big[\exp\Big(is\overline{S}_j^{(n)}/\sqrt{n}\Big)\Big] \\
&= E\Big[\exp\big(isT_n/\sqrt{n}\big)\Big] - \Big(E\Big[\exp\big(is\overline{S}_1^{(n)}/\sqrt{n}\big)\Big]\Big)^k \\
&= E\Big[\prod_{j=1}^{k}\exp\big(is\overline{S}_j^{(n)}/\sqrt{n}\big) - \Big(E\Big[\exp\big(is\overline{S}_1^{(n)}/\sqrt{n}\big)\Big]\Big)^k\Big] \\
&= \sum_{v=1}^{k} E\Big[\prod_{j=1}^{v-1}\exp\big(is\overline{S}_j^{(n)}/\sqrt{n}\big)\Big(\exp\big(is\overline{S}_v^{(n)}/\sqrt{n}\big) - E\Big[\exp\big(is\overline{S}_1^{(n)}/\sqrt{n}\big)\Big]\Big)\Big] \\
&\quad \times\Big(E\Big[\exp\big(is\overline{S}_1^{(n)}/\sqrt{n}\big)\Big]\Big)^{k-v} \\
&= \sum_{v=1}^{k} \mathrm{cov}\Big[\prod_{j=1}^{v-1}\exp\big(is\overline{S}_j^{(n)}/\sqrt{n}\big), \exp\big(is\overline{S}_v^{(n)}/\sqrt{n}\big)\Big]\Big(E\Big[\exp\big(is\overline{S}_1^{(n)}/\sqrt{n}\big)\Big]\Big)^{k-v}\,.
\end{aligned}$$

Here we used the elementary (telescoping sum) identity

$$\prod_{j=1}^{k} a_j - \prod_{j=1}^{k} b_j = \sum_{v=1}^{k}\prod_{j=1}^{v-1} a_j \times (a_v - b_v) \times \prod_{j=v+1}^{k} b_j\,.$$

Recall from (2.13) (the absolute value of a characteristic function is bounded by 1) that

$$\Big|\mathrm{cov}\Big[\prod_{j=1}^{v-1}\exp\big(is\overline{S}_j^{(n)}/\sqrt{n}\big), \exp\big(is\overline{S}_v^{(n)}/\sqrt{n}\big)\Big]\Big| \le \alpha_l\,.$$

Therefore

$$|\Delta_n| \le k\,\alpha_l$$

Under (2.16) it is possible to choose $k = k_n \to \infty$ and $l = l_n \to \infty$ such that the right-hand side converges to zero. We conclude that $|\Delta_n| \to 0$. The quantity

$$\Big(E\Big[\exp\big(is\overline{S}_1^{(n)}/\sqrt{n}\big)\Big]\Big)^k$$

in Δ_n is the characteristic function of

$$\frac{T_n'}{\sqrt{n}} = \frac{(\overline{S}_1^{(n)})' + \cdots + (\overline{S}_k^{(n)})'}{\sqrt{n}}$$

where $(\overline{S}_1^{(n)})', \ldots, (\overline{S}_k^{(n)})'$ are independent copies of $\overline{S}_1^{(n)}$. One can show (exactly in the same as we derived the limit σ^2 for $\mathrm{var}(\sqrt{n}\overline{X}_n)$) that

$$\mathrm{var}\Big(T_n'/\sqrt{n}\Big) = \frac{k}{n}\mathrm{var}\big(\overline{S}_1^{(n)}\big) = \mathrm{var}\big(\overline{S}_1^{(n)}/\sqrt{m}\big) \to \sigma^2\,.$$

Classical limit theory for sums of iid random variables whose distribution may change from n to $n+1$ (see Petrov [28]) yields that

$$P\big(T_n'/(\sigma\sqrt{n}) \le x\big) \to \Phi(x)\,, \qquad n \to \infty\,, \quad x \in \mathbb{R}\,.$$

Write

$$\frac{\sqrt{n}}{\sigma}\overline{X}_n = \frac{T_n}{\sigma\sqrt{n}} + R_n \stackrel{d}{=} \frac{T_n'}{\sigma\sqrt{n}} + o_P(1)\,,$$

where $o_P(1)$ is a sequence of random variables converging to zero in probability. Collecting all the arguments above, we have

$$\frac{\sqrt{n}}{\sigma}\overline{X}_n \stackrel{d}{\to} N(0,1)\,, \qquad n \to \infty\,,$$

as desired.

2.6. **Transformation to stationarity.** Classical time series analysis is about stationary and strictly stationary processes. This does not mean that there have not been made attempts to deviate from this assumption. Quite often it is supposed that a time series becomes stationary after some transformation or that the time series is "locally stationary". This naturally means that we must have some additional information about the structure of the time series. In what follows, we consider some ad hoc transformations which sometimes transform real-life data which are not believed to come from a stationary sequence into a "more stationary" form.

Notice that real-life data will hardly ever come from a stationary (in the wide or strict senses) sequence; if we assume that a suitable transformation of the data yields "something similar to a sample from a stationary sequence" this is nothing but a convenient model assumption, which cannot be verified. Indeed, the definition of stationarity involves the distribution of the infinite sequence (X_t) which is never available.

One can, however, assume a parametric model for a time series, estimate its parameters and afterwards check the goodness-of-fit of the model. This means that one particular class of stationary models can give a good fit to the data, but it is not a "proof" of the stationarity of the underlying data.

For real-life data, properties like ergodicity, mixing and strong mixing cannot be tested in principle. One way of convincing oneself of stationarity/ergodicity is to apply the same estimator to distinct pieces of the time series with equal length. For example, if one has daily data one can check the stability of the sample mean, sample variance, and other sample moments calculated on these pieces, e.g. the annual means, annual variance, etc. If these quantities vary heavily one may have doubts about stationarity/ergodicity.

Example 2.46. Financial time series such as share prices, exchange rates, composite stock indices are usually not modeled by stationary processes. Standard financial theory tells us that prices increase roughly exponentially through time, thus their expectation can certainly not be a constant; see for example the Consumer Price Index (CIP) available on the website of Statistics Denmark. The CIP corrects prices for inflation and allows one to compare prices from different time periods.

In the financial time series and econometrics literature the following two transformations of the original price time series (X_t) are proposed:

$$Y_t = \frac{X_t - X_{t-1}}{X_{t-1}}\,, \quad t \in \mathbb{Z}\,, \tag{2.18}$$

$$Y_t = \log(X_t) - \log(X_{t-1}) = \log\left(\frac{X_t}{X_{t-1}}\right)\,, \quad t \in \mathbb{Z}\,. \tag{2.19}$$

It is believed that these transformations provide stationarity. Notice that (2.18) just defines "daily returns" if we understand X_t as a price on a given day t, e.g. closing or high or low daily prices, as provided by agencies such as Reuters, Bloomberg, or Yahoo Finance. The expressions (2.18) and (2.19) essentially define the same quantities: by a Taylor expansion argument we see that

$$\log\left(\frac{X_t}{X_{t-1}}\right) = \log\left(1 + \frac{X_t - X_{t-1}}{X_{t-1}}\right) \approx \frac{X_t - X_{t-1}}{X_{t-1}}\,.$$

This approximation works quite well since the daily returns $(X_t - X_{t-1})/X_{t-1}$ are usually very small.

The definition of "returns" indicates that they measure the relative change of a price at equidistant instants of time, e.g. days. In contrast to the price, they are independent of the monetary unit, therefore they allow for a comparison of the performance of different prices on the same time scale.

Another argument for the transformations (2.18) and (2.19) is that they express the general belief that prices increase roughly exponentially through time. The popular Black-Scholes model for speculative prices assumes that

$$X_t = X_0 \exp\left(ct + \sigma B_t\right)\,, \qquad t \geq 0\,,$$

for a standard Brownian motion B and positive c, σ. Assuming independence between X_0 and $(B_t)_{t>0}$, we have

$$EX_t = EX_0 \exp\left((c + 0.5\sigma^2)t\right)\,,$$

i.e., the series increases exponentially on average. Given this model, (2.19) yields the log-returns $Y_t = c + \sigma(B_t - B_{t-1})$, $t = 1, 2, \ldots$. In view of the independent and stationary increments of Brownian motion this log-return sequence consists of iid Gaussian random variables with mean c and variance σ^2. We will see later that real-life return series are dependent and have rather heavy tails, not comparable with the normal distribution.

There exist various books on financial time series. One of the first ones in this context was Taylor [35]. Although it is not the most recent monograph, it still contains a wealth of interesting material on empirical features of financial time series and arguments for using returns instead of prices. By now, returns are the standard objects of financial time series analysis.

Exercise 2.47. Look at the USD/DEM and USD/FRF foreign exchange rates on the website www.math.ku.dk/∼mikosch/Time and calculate the return and log-return series (Y_t) and $\log(1+Y_t)$, respectively. Plot the differences $|Y_t - \log(1 + Y_t)|$ and determine their maximum. Calculate the sample autocorrelations of the resulting return time series and their absolute values. Choose the maximum number of lags as 10% of the sample size. (Function acf in R.)

In what follows, we consider more transformations which may lead to stationary time series. However, we must be aware that there is no unique rule for dealing with real–life time series data and any procedure of "making them stationary" is subjective. One also has to decide whether such transformations really provide a gain of information. Clearly, we can apply the whole existing theory for stationary processes to the new time series but it is not always possible to translate this into

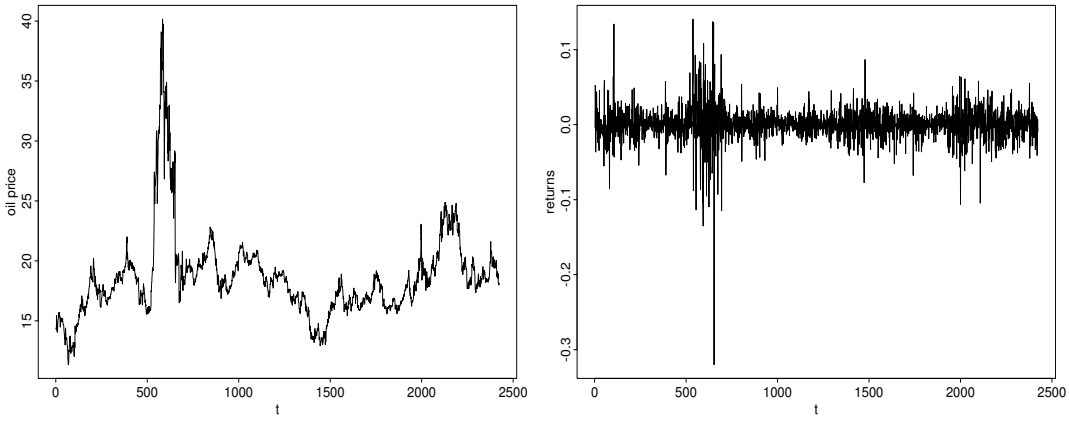

Figure 2.48. *Left: Daily prices of crude oil over a period of* 2423 *days. Right: The corresponding return series.*

information about the original series. For example, a theory about the maxima of the transformed time series does usually not give information about the maxima of the original time series.

Now suppose that we know that the original time series is given as

$$X_t = m_t + s_t + Y_t\ , \quad t = 1, \dots, n\ . \tag{2.20}$$

Here m_t is a slowly changing deterministic function, called *trend*, and s_t is another deterministic function with period d, say, i.e., $s_t = s_{t+d}$, called *seasonal component*. The process (Y_t) is supposed to be stationary. Thus the model (2.20) follows the classical pattern of "signal + noise" philosophy. For example, daily temperature data measured at the same place clearly have a seasonal component which is usually well known (e.g. by taking averages over annual data at the same day of the year) and can be subtracted from the observations (X_t) giving the stationary noise (Y_t). If there was a global warming one would also have to take into account a trend m_t which, however, is difficult to detect although meteorologists and climate researchers have done their best.

Presence of trend, absence of seasonality. We consider a submodel of (2.20): suppose there is no seasonality. Then we arrive at

$$X_t = m_t + Y_t\ , \quad t = 1, \dots, n\ .$$

Least squares estimation of m_t. Suppose that m_t is quadratic in t:

$$m_t = a_0 + a_1 t + a_2 t^2\ .$$

A natural way of estimating the a_i's is by least squares: minimize

$$\sum_{t=1}^{n} (X_t - m_t)^2$$

with respect to a_0, a_1, a_2 (by taking partial derivatives with respect to the a_i's, setting them equal to zero and solving the corresponding system of equations). Brockwell and Davis [8], p. 15, apply this procedure to US population data, 1790-1980, with estimated parameter values

$$\widehat{a}_0 = 2.097911 \times 10^{10},\ \widehat{a}_1 = -2.334962 \times 10^{7},\ \widehat{a}_2 = 6.498591 \times 10^{3}\ . \tag{2.21}$$

See Figure 2.49. Of course, instead of a quadratic polynomial we can choose other functions m_t and estimate their parameters by using least squares estimation. In general, this does not lead to explicit expressions for the estimators and we would depend on numerical approximations.

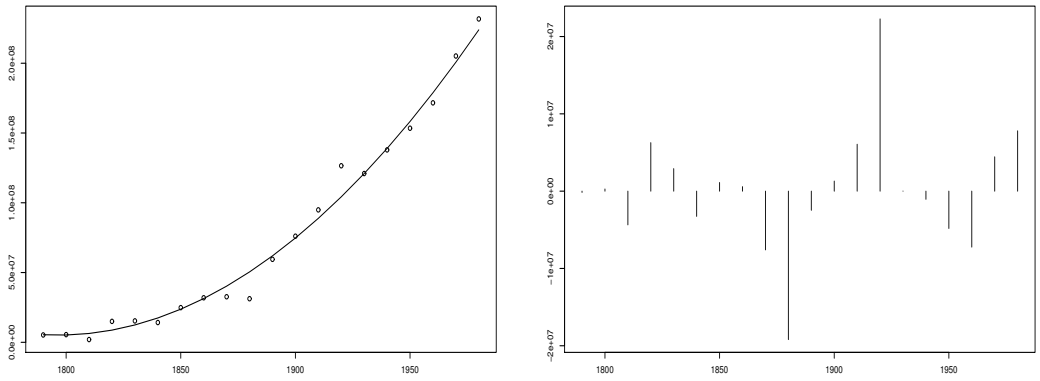

Figure 2.49. Left: *Population of the USA at ten year intervals,* 1790-1980 (*US Bureau of the Census*) *with parabola fitted by least squares; see* (2.21). Right: *The residuals* (*data minus fitted parabola*) *give one the impression of a stationary sequence.*

Smoothing via moving averages. Let q be a positive integer. Consider the moving averages

$$\frac{1}{2q+1}\sum_{|j|\le q} m_{t+j} + \frac{1}{2q+1}\sum_{|j|\le q} Y_{t+j} \approx m_t\,. \tag{2.22}$$

The latter relation will be satisfied if the noise (Y_t) is "small", e.g. if (Y_t) is ergodic with mean value zero then the strong law of large numbers applies, and m_t is "locally linear". For a sample $X_1, \dots, X_n$ one has to modify this procedure when $t \le q$ or $t > n - q$.

The smoothing procedure (2.22) automatically introduces some kind of dependence between the X_t's. This is easily seen if the Y_t's are iid: the outcome of (2.22) is a time series which is dependent over $2q + 1$ lags. Although a large q usually gives one a smoother approximation to m_t, it also introduces undesirable dependence effects in the time series, and therefore one should try to work with a q which is rather small.

Procedure (2.22) can be generalized in various ways by giving other weights than $1/(2q + 1)$ to each X_{t-j}. See for example the classical text by Kendall and Stuart [22] or Brockwell and Davis [8, 9]. As a matter of fact, the estimation of m_t is closely related to kernel curve estimation for which a vast literature exists; see for example Wand and Jones [36].

Differencing. Now we explain one of the most popular methods for transforming time series into a "stationary" regime. We start with a time series of the form

$$X_t = at + b + Y_t\,, \quad t \in \mathbb{Z}\,,$$

i.e., (X_t) has a linear trend. Let $BX_t = X_{t-1}$ denote the *backshift operator* and

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

be the *difference operator*. Notice that

$$\Delta X_t = a + \Delta Y_t\,.$$

Polynomial trends m_t can be treated in the same way: let

$$B^j(X_t) = X_{t-j}\,, \quad \Delta^0(X_t) = X_t\,, \quad \Delta^j(X_t) = \Delta(\Delta^{j-1}(X_t))\,.$$

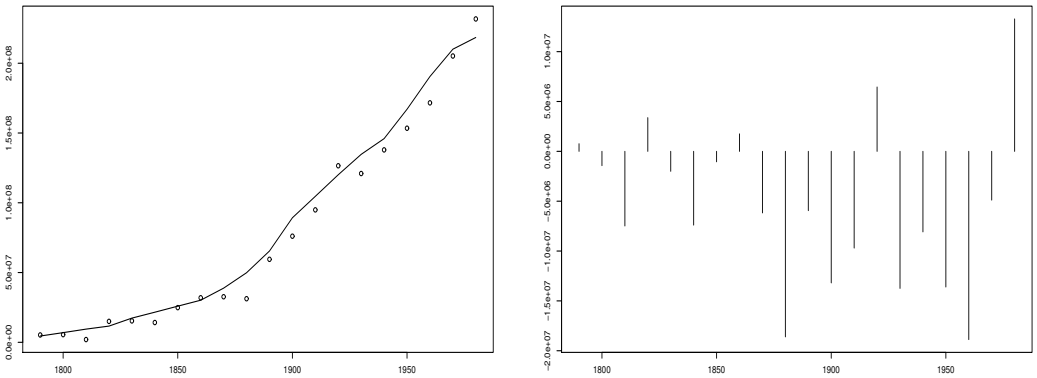

Figure 2.50. Left: *Population of the USA at ten year intervals,* 1790-1980 (*US Bureau of the Census*) *with estimated mean* (*solid line*) *obtained by smoothing the data by a moving average* (2.22) *with* $q = 2$. Right: *The residuals* (*data minus fitted mean curve*) *give one the impression of a stationary sequence.*

Polynomials in Δ and B can be manipulated in the same way as polynomials in real variables. For example,

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

Starting with the model

$$X_t = m_t + Y_t = \sum_{j=0}^{k} a_j t^j + Y_t\,, \quad t \in \mathbb{Z}\,,$$

one gets after k times differencing

$$\Delta^k(X_t) = k!\, a_k + \Delta^k(Y_t)\,,$$

which is stationary if (Y_t) is stationary; see Proposition 4.9.

Notice that the construction of log-returns in (2.19) is a differencing procedure as described above applied to time series of log-prices.

Presence of trend and seasonality. There exist many methods to deal with this problem; see e.g. Brockwell and Davis [8], Section 1.4. We restrict ourselves to one particular method. Since the seasonal component s_t has period d (e.g. $d = 12$ for monthly data, $d = 364$ for daily data, $d = 250$ for daily speculative price data, etc.) it is natural to apply a difference operator $\Delta_d = 1 - B^d$ to (X_t). This results in

$$\Delta_d(X_t) = m_t - m_{t-d} + Y_t - Y_{t-d}\,,$$

i.e., we get rid of the seasonal component. The new trend $m_t - m_{t-d}$ can be eliminated by an application of the methods described above. This (hopefully) leads to a stationary time series.

Differencing of time series is one of the standard procedures in time series analysis with the aim of "coming closer to a stationary regime". When dealing with real-life data it is not always possible to decide in a unique way how often one should difference the data or whether one should difference the data at all. As said before, it is a subjective decision as to which procedure one should use in order to transform the data to stationarity. Often additional information (such as in the case of the construction of returns from price data) helps one to apply an appropriate transformation. But in any case one should be clear about the following: stationarity is a convenient mathematical

assumption, which we will need in the sequel for building up a nice mathematical theory, but real-life data will hardly come from pure stationary models as described above. We can only hope to "approximate" real-life data by a stationary process via suitable transformations.

Example 2.51. To show what happens if we neglect trends or seasonalities in a time series we consider the two toy models

$$X_t = t + Y_t \qquad \text{and} \qquad X'_t = \cos t + Y_t\,, \tag{2.23}$$

where (Y_t) is an iid standard normal sequence. Of course, the data are independent but not stationary. If we consider the sample autocorrelation function $\rho_{n,X}$ of the time series (X_t) (see Section 3.2 for its definition and properties) we observe that $\rho_{n,X}(h)$ stays very close to one for a large number of lags. This follows from the fact that $\rho_{n,X}(h) \overset{\text{a.s.}}{\to} 1$ as $n \to \infty$. On the other hand, $\rho_{n,X'}$ has a strong seasonal component. This follows from the fact that $\rho_{n,X'}(h) \overset{\text{a.s.}}{\to} \cos h$ as $n \to \infty$. These effects are illustrated in Figure 2.51. If one sees sample autocorrelations functions like these (the values are hardly close to zero even at large lags) one should be suspicious about the stationarity assumption of the data.

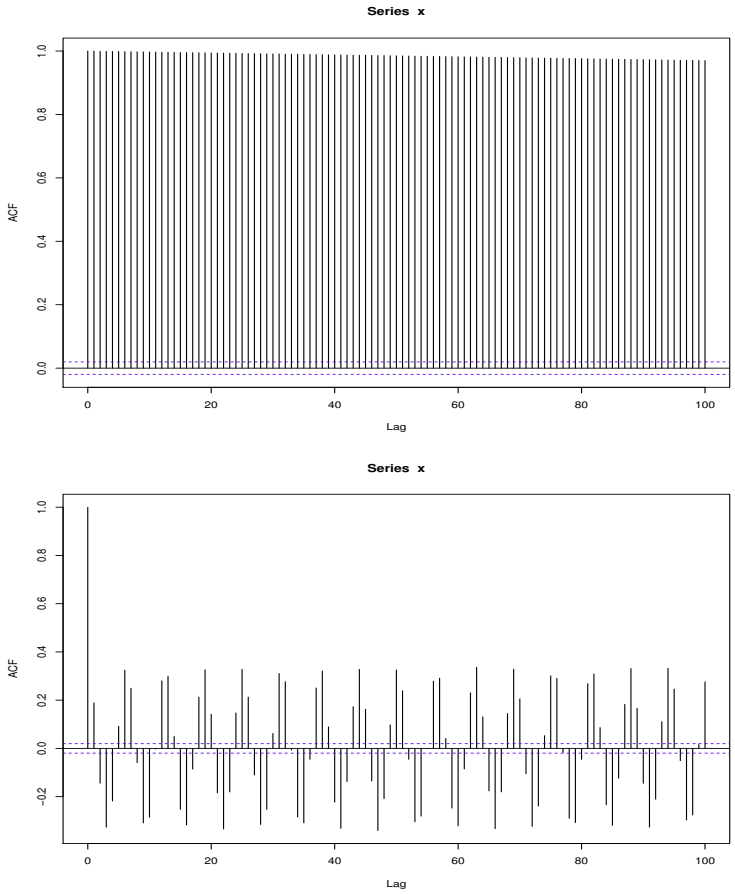

Figure 2.52. *Sample autocorrelation function of the non-stationary time series* (X_t) *(top) and* (X'_t) *(bottom) from* (2.23)*. The sample size is* $n = 10000$*.*

3. The autocovariance and the autocorrelation functions

3.1. **Some basic properties.** Throughout this section $(X_t)_{t\in\mathbb{Z}}$ is a real-valued stationary process. In Section 2 we have learned that the second order structure (covariances, correlations) of such a process gives one a first impression of dependencies in this sequence. In this section we want to study some of the basic properties of the *autocovariance function*

$$\gamma_X(h) = \mathrm{cov}(X_h, X_0)\,, \qquad h \in \mathbb{Z}\,.$$

These properties can immediately be translated to the *autocorrelation function*

$$\rho_X(h) = \gamma_X(h)/\gamma_X(0)\,, \qquad h \in \mathbb{Z}\,,$$

since it is the autocovariance function of the stationary process $(X_t/\sqrt{\mathrm{var}(X_0)})_{t\in\mathbb{Z}}$.

It is immediate that

$$\gamma_X(0) = \mathrm{var}(X_0) \geq 0$$

and, by the Cauchy-Schwarz inequality,

$$|\gamma_X(h)| = |\mathrm{cov}(X_h, X_0)| \leq (\mathrm{var}(X_0)\mathrm{var}(X_h))^{1/2} = \gamma_X(0)\,,$$

and therefore $|\rho_X(h)| \leq 1$. Moreover,

$$\gamma_X(-h) = \mathrm{cov}(X_{-h}, X_0) = \mathrm{cov}(X_0, X_h) = \gamma_X(h)\,.$$

Now let $a_1, \ldots, a_n \in \mathbb{R}$, $t_1, \ldots, t_n \in \mathbb{Z}$. Then

$$\begin{aligned}
\sum_{i,j=1}^{n} a_i\, a_j\, \gamma_X(t_i - t_j) &= \sum_{i,j=1}^{n} a_i\, a_j\, E[(X_{t_i} - EX_0)(X_{t_j} - EX_0)] \\
&= E\left[\sum_{i,j=1}^{n} a_i\, a_j\, (X_{t_i} - EX_0)(X_{t_j} - EX_0)\right] \\
&= E\left|\sum_{i=1}^{n} a_i\, (X_{t_i} - EX_0)\right|^2 \geq 0\,.
\end{aligned}$$

This property of the autocovariance function leads to the following definition:

Definition 3.1. (Non-negative definiteness)
A function $\gamma : \mathbb{Z} \to \mathbb{R}$ is called non-negative definite *if*

$$\sum_{i,j=1}^{n} a_i\, a_j\, \gamma(t_i - t_j) \geq 0 \tag{3.1}$$

for every n and every choice of $a_1, \ldots, a_n \in \mathbb{R}$, $t_1, \ldots, t_n \in \mathbb{Z}$.

Non-negative definiteness characterizes the autocovariance function of a stationary process:

Theorem 3.2. (Characterization of the autocovariance function)
A function $\gamma : \mathbb{Z} \to \mathbb{R}$ is the autocovariance function of a real-valued stationary time series if and only if it is even (i.e., $\gamma(h) = \gamma(-h)$) and non-negative definite. Moreover, for every even non-negative function γ on $\mathbb{Z}$ there exists a stationary Gaussian process with autocovariance function γ.

Proof. The first part of the proof was given above. The proof of the second part of this theorem is based on the so-called Kolmogorov consistency theorem which allows one to construct a stationary Gaussian process from its finite-dimensional distributions (which only depend on the expectation and covariance structures).

Kolmogorov's theorem (cf. Theorem 1.2.1 in Brockwell and Davis [8]) tells us that a collection of distribution functions $(F_{\mathbf{t}})_{\mathbf{t}\in\mathcal{T}}$, where $\mathcal{T} = \{\mathbf{t} = (t_1, \ldots, t_n) : t_i \in \mathbb{Z}, t_1 < \cdots < t_n, n = 1, 2, \ldots\}$, are

the distribution functions corresponding to the finite-dimensional distributions of some stochastic process $(X_t)_{t\in\mathbb{Z}}$ if and only if for any $n = 1, 2, \ldots$, $\mathbf{t} = (t_1, \ldots, t_n) \in \mathcal{T}$ and $1 \le i \le n$,

$$\lim_{u_i \to 0} \phi_{\mathbf{t}}(\mathbf{u}) = \phi_{\mathbf{t}(i)}(\mathbf{u}(i)) \tag{3.2}$$

where

$$\phi_{\mathbf{t}}(\mathbf{u}) = \int_{\mathbb{R}^n} \exp(i\mathbf{u}'\mathbf{x})\, F_{\mathbf{t}}(dx_1, \ldots, dx_n)\,, \quad \mathbf{u} \in \mathbb{R}^n\,,$$

is the characteristic function corresponding to $F_{\mathbf{t}}$. We know that $\phi_{\mathbf{t}}$ determines the distribution function $F_{\mathbf{t}}$ and vice versa. The characteristic function $\phi_{\mathbf{t}(i)}(\mathbf{u}(i))$ is obtained from $\phi_{\mathbf{t}}(\mathbf{u})$ by deleting the ith components of $\mathbf{t}$ and $\mathbf{u}$. Condition (3.2) is the so-called *consistency property* of $(F_{\mathbf{t}})_{\mathbf{t}\in\mathcal{T}}$; it ensures that marginal distribution functions of $F_{\mathbf{t}}$ should coincide with the specified lower-dimensional distribution functions.

Now assume that γ is even and non-negative definite. We will show that there exists a Gaussian mean-zero stationary process with γ as its autocovariance function. For given $\mathbf{t} \in \mathbb{Z}^n$ such that $t_1 < \cdots < t_n$ define

$$\Gamma_n = \big(\gamma(t_i - t_j)\big)_{i,j=1,\ldots,n}\,.$$

By virtue of non-negative definiteness of γ the matrix Γ_n is non-negative definite, i.e.,

$$\mathbf{u}'\Gamma_n\mathbf{u} \ge 0\,, \qquad \mathbf{u} \in \mathbb{R}^n\,.$$

Consequently,

$$\phi_{\mathbf{t}}(\mathbf{u}) = \exp\big(-\frac{1}{2}\mathbf{u}'\Gamma_n\mathbf{u}\big)\,, \qquad \mathbf{u} \in \mathbb{R}^n\,,$$

is a characteristic function corresponding to some distribution function $F_{\mathbf{t}}$. Indeed, it is the characteristic function of an n-dimensional Gaussian distribution with mean zero and covariance matrix Γ_n. Clearly, (3.2) is satisfied, hence the distribution functions $F_{\mathbf{t}}$ are consistent. By Kolmogorov's theorem, there exists a time series (X_t) with finite-dimensional distribution functions $F_{\mathbf{t}}$ and characteristic functions $\phi_{\mathbf{t}}$, $\mathbf{t} \in \mathcal{T}$. In particular, $\mathrm{cov}(X_i, X_j) = \gamma(i-j)$, as required for a stationary process. □

The verification of the non-negative definiteness of a given function is in general difficult. This property is mainly of theoretical interest.

For parameter estimation we will need the following result:

Proposition 3.3. *Let (X_t) be a stationary process. If its autovariance function satisfies $\gamma_X(0) > 0$ and $\gamma_X(h) \to 0$ as $h \to \infty$ then the inverse of the covariance matrix*

$$\Gamma_n = (\gamma_X(i-j))_{i,j=1,\ldots,n}$$

exists for every n.

Proof. Suppose that Γ_n is singular for some n. We may assume without loss of generality that $EX_0 = 0$. In view of Exercise 3.4 we know that there exists an integer $r \ge 1$ and real constants $a_1, \ldots, a_r$ such that Γ_r is non-singular and

$$X_{r+1} = \sum_{j=1}^{r} a_j X_j\,.$$

By stationarity we then also have

$$X_{r+h} = \sum_{j=1}^{r} a_j X_{j+h-1}, \qquad \text{for all } h \ge 1.$$

Hence for all $n \geq r+1$ there exist real vectors $\mathbf{a}^{(n)} = (a_1^{(n)}, \ldots, a_r^{(n)})'$ such that

$$X_n = (\mathbf{a}^{(n)})' \mathbf{X}_r \,, \qquad \text{where } \mathbf{X}_r = (X_1, \ldots, X_r)'$$

Calculating the variance in the latter relation, we obtain

$$\gamma_X(0) = (\mathbf{a}^{(n)})' \Gamma_r \mathbf{a}^{(n)} = (\mathbf{a}^{(n)})' O \Lambda O' \mathbf{a}^{(n)} \,,$$

where O is an $r \times r$ orthonormal matrix, i.e., $OO' = I_r$, where I_r is the r-dimensional identity matrix, and Λ is a diagonal matrix whose diagonal entries are the eigenvalues $\lambda_1 \leq \cdots \leq \lambda_r$ of Γ_r. Since this matrix is invertible, hence positive definite, its eigenvalues are positive, in particular $\lambda_1 > 0$. Hence

$$\gamma_X(0) \geq \lambda_1 (\mathbf{a}^{(n)})' O O' \mathbf{a}^{(n)} = \lambda_1 \sum_{i=1}^{r} (a_i^{(n)})^2,$$

and therefore each $a_i^{(n)}$ is a bounded function of n. On the other hand,

$$\gamma_X(0) = \operatorname{cov}(X_n, \sum_{i=1}^{r} a_i^{(n)} X_i) \leq \sum_{i=1}^{r} |a_i^{(n)}| \, |\gamma_X(n-i)| \,.$$

In view of this inequality and since $a_i^{(n)}$ are bounded we cannot have $\gamma_X(0) > 0$ and $\gamma_X(h) \to 0$ as $h \to \infty$ at the same time if Γ_n is singular. □

Exercise 3.4. Let $\mathbf{X} = (X_1, \ldots, X_n)'$ be a random vector with covariance matrix Σ. Show that Σ is singular (i.e., non-invertible) if and only if there exists a non-zero vector $\mathbf{b} \in \mathbb{R}^n$ such that $\operatorname{var}(\mathbf{b}'\mathbf{X}) = 0$. In particular, if $\mathbf{X}$ has mean zero $\mathbf{b}'\mathbf{X} = 0$.

3.2. **The sample autocovariance and autocorrelation functions.** Now assume we observed a sample $X_1, \ldots, X_n$ from a stationary time series (X_t). An important question is to estimate its autocorrelation and autocovariance functions from this sample. Natural sample estimators (method of moment estimators) are given by the *sample autocovariance function*

$$\gamma_{n,X}(h) = \begin{cases} \dfrac{1}{n} \displaystyle\sum_{j=1}^{n-|h|} (X_j - \overline{X}_n)(X_{j+|h|} - \overline{X}_n) & |h| \leq n-1 \,, \\ 0 & |h| \geq n \,, \end{cases}$$

and by the *sample autocorrelation function*

$$\rho_{n,X}(h) = \frac{\gamma_{n,X}(h)}{\gamma_{n,X}(0)} \,, \quad h \in \mathbb{Z} \,,$$

where

$$\overline{X}_n = \frac{1}{n} \sum_{j=1}^{n} X_j$$

denotes the sample mean. The quantities $\gamma_{n,X}(h)$ and $\rho_{n,X}(h)$ are the *sample autocovariance* and the *sample autocorrelation* at *lag* h, respectively.

Proposition 3.5. *The functions* $(\gamma_{n,X}(h))_{h\in\mathbb{Z}}$ *and* $(\rho_{n,X}(h))_{h\in\mathbb{Z}}$ *are even and non–negative definite for every realization of* (X_t).

Proof. As for the autocovariance and autocorrelation functions, it suffices to prove the statement for the sample autocovariance function since the sample autocorrelation function is a scaled version

of $\gamma_{n,X}$. Write

$$\begin{aligned} Y_i &= X_i - \overline{X}_n\,, \\ T_n &= \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 & Y_1 & Y_2 & \cdots & Y_{n-1} & Y_n \\ 0 & 0 & 0 & \cdots & Y_1 & Y_2 & Y_3 & \cdots & Y_n & 0 \\ \vdots & & & & & \vdots & & & & \vdots \\ 0 & Y_1 & Y_2 & \cdots & Y_{n-1} & Y_n & 0 & \cdots & 0 & 0 \end{pmatrix}, \\ \Gamma_{n,X} &= (\gamma_{n,X}(i-j))_{i,j=1,\ldots,n}\,. \end{aligned}$$

Then for any n-dimensional column vector $\mathbf{a}$,

$$\mathbf{a}'\,\Gamma_{n,X}\,\mathbf{a} = \mathbf{a}'\,(n^{-1}T_n\,T_n')\,\mathbf{a} = n^{-1}\,(\mathbf{a}'\,T_n)\,(\mathbf{a}'\,T_n)' \geq 0\,.$$

This means that the matrix $\Gamma_{n,X}$ is non-negative definite or, alternatively,

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

for any choice of a_i's, which implies (3.1) for $\gamma = \gamma_{n,X}$. Hence $\gamma_{n,X}$ is non-negative definite and we also have $\gamma_{n,X}(h) = \gamma_{n,X}(-h)$ by definition of these quantities. □

Remark 3.6. Recall from Proposition 3.3 that the covariance matrices $\Gamma_n = (\gamma_X(i-j))_{i,j=1,\ldots,n}$ of a stationary process (X_t) are invertible if $\gamma_X(0) > 0$ and $\gamma_X(h) \to 0$ as $h \to \infty$. We learn from Proposition 3.5 that the sample autocovariance function $\gamma_{n,X}$ is the autocovariance function of some process. Brockwell and Davis [8], Proposition 3.2.1, show that $\gamma_{n,X}$ can be interpreted as the autocovariance function of some MA$(n-1)$ process driven by white noise, i.e., $\gamma_{n,X}$ is the autocovariance function of some stationary process. Clearly, $\gamma_{n,X}(h) = 0$ for $h \geq n$. Thus, if we have $\gamma_{n,X}(0) > 0$ we may conclude that the sample covariance matrices $\Gamma_{h,X} = (\gamma_{n,X}(i-j))_{1\leq i,j\leq h}$, $h \geq 1$, are invertible.

Why do we choose the normalization n for $\gamma_{n,X}(h)$? For every realization of (X_t), both $\gamma_{n,X}$ and $\rho_{n,X}$ are the autocovariance/autocorrelation functions of a stationary process. It is this property which made us define the sample autocovariance $\gamma_{n,X}$ with normalization n instead of $n-h$, corresponding to the number of summands in the sum constituting $\gamma_{n,X}$. For large n and h small compared to n, the different normalizations do not matter. For small n, the normalization n leads to a substantial bias of $\gamma_{n,X}(h)$.

Consistency and asymptotic normality. Under general conditions, $\gamma_{n,X}(h)$ and $\rho_{n,X}(h)$ are consistent and asymptotically normal estimators of $\gamma_X(h)$ and $\rho_X(h)$, respectively. *Consistency* means that for every fixed $h \in \mathbb{Z}$,

$$\gamma_{n,X}(h) \stackrel{\text{a.s.}}{\to} \gamma_X(h) \quad \text{and} \quad \rho_{n,X}(h) \stackrel{\text{a.s.}}{\to} \rho_X(h)\,.$$

We concluded in Exercise 2.33 that this property holds if $\mathrm{var}(X_0) < \infty$ and (X_t) is a strictly stationary ergodic sequence.

Asymptotic normality of the sample autocovariances means that

$$\sqrt{n}\,(\gamma_{n,X}(h) - \gamma_X(h))_{h=1,\ldots,m} \stackrel{d}{\to} \mathbf{Y}_m \sim N(\mathbf{0}, \Sigma)\,, \tag{3.3}$$

where $N(\mathbf{0}, \Sigma)$ denotes the Gaussian distribution with mean $\mathbf{0}$ and covariance matrix Σ; see (2.2) for the Gaussian density. The limiting covariance matrix is rather complicated and therefore omitted; see Brockwell and Davis [8], Section 7.2, or (4.15) below for some particular cases. Relation (3.3) does not generally hold for stationary or ergodic processes (X_t) with a regular covariance matrix Σ. In addition to ergodicity one needs to ensure that (X_t) satisfies some additional structural

conditions, e.g. that (X_t) is a linear process (see p. 44), a martingale difference sequence, or that the *strong mixing condition* holds with a rate function converging to zero sufficiently fast. In the latter case, we formulated conditions for (3.3) and the corresponding results for the sample autocorrelation function in Corollary 2.45.

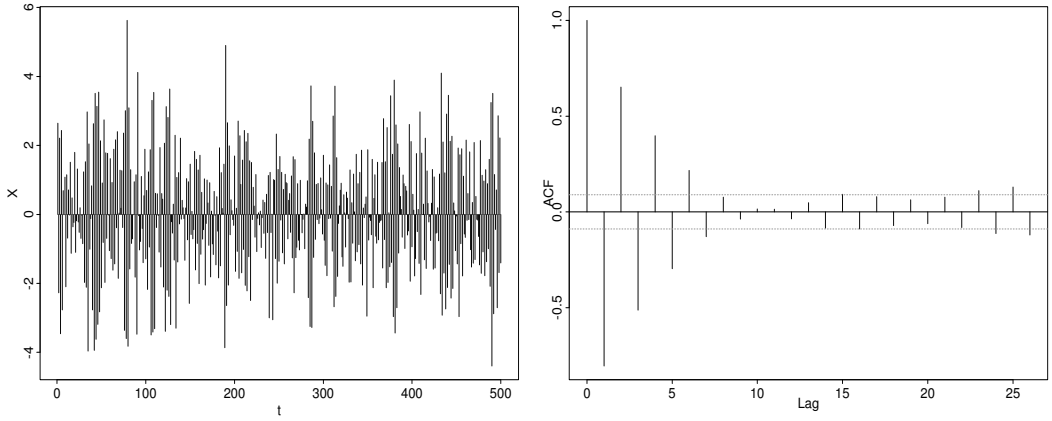

Figure 3.7. *One sample path of the* AR(1) *process* $X_t = -0.8X_{t-1} + Z_t$ *with iid Gaussian* (Z_t) *and its sample autocorrelation function.*

The asymptotic normality of the sample autocorrelations follows in a similar way by a continuous mapping argument. Indeed,

$$\sqrt{n}\,(\rho_{n,X}(h) - \rho_X(h)) = \frac{\sqrt{n}\,(\gamma_{n,X}(h) - \gamma_X(h))}{\gamma_{n,X}(0)} + \gamma_X(h)\,\frac{\sqrt{n}\,(\gamma_X(0) - \gamma_{n,X}(0))}{\gamma_{n,X}(0)\gamma_X(0)}\,. \tag{3.4}$$

If we assume consistency and joint asymptotic normality for $\gamma_{n,X}(h)$ and $\gamma_{n,X}(0)$, then

$$(\sqrt{n}\,(\gamma_{n,X}(h) - \gamma_X(h))\,, \sqrt{n}\,(\gamma_{n,X}(0) - \gamma_X(0))\,, \gamma_{n,X}(0)) \xrightarrow{d} (N_1, N_2, \gamma_X(0))\,,$$

where (N_1, N_2) are jointly Gaussian. This, (3.4) and the continuous mapping theorem yield

$$\begin{aligned} \sqrt{n}\,(\rho_{n,X}(h) - \rho_X(h)) &\xrightarrow{d} N_1\,[\gamma_X(0)]^{-1} - \gamma_X(h)N_2\,[\gamma_X(0)]^{-2} \\ &= [\gamma_X(0)]^{-1}[N_1 - \rho_X(h)\,N_2]\,. \end{aligned} \tag{3.5}$$

In particular, the limit distribution is normal. The standard deviation $\sigma(h)$ of the limiting Gaussian random variable can sometimes be calculated (if one assumes a particular time series model) or estimated from the data. It can be used to construct pointwise confidence bands for $\rho_{n,X}(h)$. For example, the event $\{\rho(h) \in [\rho_{n,X}(h) - 1.96\sigma(h)/\sqrt{n}, \rho_{n,X}(h) + 1.96\,\sigma(h)/\sqrt{n}]\}$ corresponds to an asymptotic 95% confidence band. In the case of linear processes with iid white noise (including moving average and autoregressive processes with iid white noise) $\sigma(h)$ is explicitly known; see (4.15) below.

What do the confidence bands in statistical software for $\rho_{n,X}$ mean? From these calculations we can see that the limiting normal distribution depends on the dependence structure, in particular, the covariance structure of (X_t). For the construction of confidence bands for the sample autocorrelations $\rho_{n,X}(h)$ one would therefore depend on estimates for these unknown parameters. Since this is often inconvenient or impossible, statistical software for the sample autocorrelation function does usually not show asymptotic confidence bands based on the central limit theorem (3.5) for $\rho_{n,X}(h)$ with the true (or estimated) asymptotic variance. Software such as R or S+ gives 95% asymptotic confidence for $\rho_{n,X}(h)$ assuming that (X_t) is an iid Gaussian sequence, for which $\rho_X(h) = 0$ for

all $h \neq 0$. Therefore we see in standard software for the sample autocorrelation function the two horizontal lines $\pm 1.96/\sqrt{n}$, indicating those confidence bands. Note in particular that these confidence bands do not depend on the lag h; for a dependent sequence (X_t) the width of the confidence bands usually differs for distinct h. All graphs in these notes which show sample autocorrelation functions also give the lines for the Gaussian iid case. Therefore we have to interpret these bands with caution: they are not the confidence bands we are interested in but they indicate how far the sample autocorrelation function of the data deviates from the autocorrelation structure of iid standard Gaussian white noise.

Exercise 3.8. Assume that (X_t) is an iid white noise sequence and that $E[|X|^{2+\delta}] < \infty$ for some $\delta > 0$.[4] Then $\gamma_X(h) = \rho_X(h) = 0$ for $h \neq 1$. Show that

$$\sqrt{n}\left(\gamma_{n,X}(h)\right)_{h=1,\ldots,m} \xrightarrow{d} \mathbf{Y}_m \sim N(\mathbf{0}, \sigma^4 I_m)\,, \qquad n \to \infty,$$

where $\sigma^2 = \mathrm{var}(X_0)$ and I_m is the m-dimensional identity matrix. Apply Ibragimov's central limit theorem to show that

$$\sqrt{n}\left(\rho_{n,X}(h)\right)_{h=1,\ldots,m} \xrightarrow{d} \frac{\mathbf{Y}_m}{\sigma^2} \sim N(\mathbf{0}, I_m)\,, \qquad n \to \infty.$$

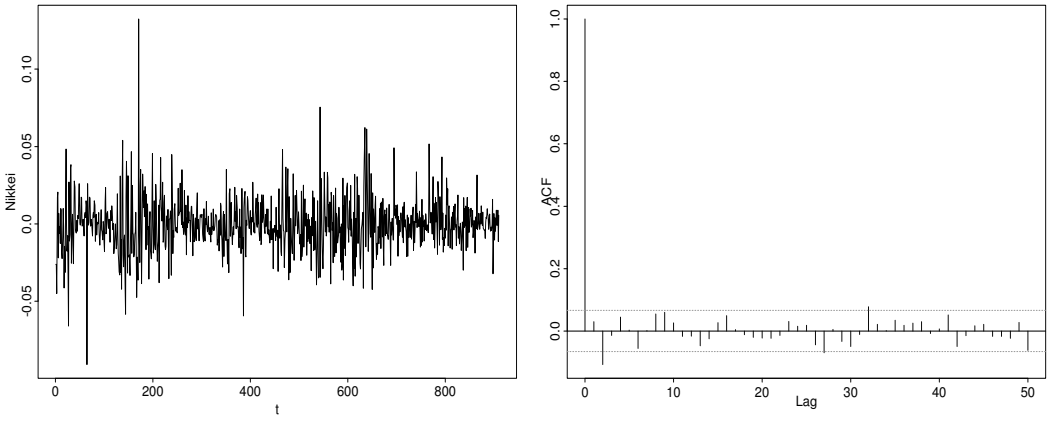

Figure 3.9. *The NIKKEI (Japanese composite stock index) daily closing log-returns over a period of 4 years and its sample autocorrelation function. The sample autocorrelation function and the confidence bands suggest that the data constitute white noise.*

Can the sample autocorrelations fool us? Since we do not know the autocorrelations, we depend on the values of sample autocorrelations as approximations. These approximations can be doubtful in various situations.

- The sample can be too small. Asymptotic theory (consistency, asymptotic normality) is then not applicable. The interpretation of the estimators $\rho_{n,X}(h)$ and their confidence bands can be meaningless. For an application of asymptotic results one should require sample sizes which exceed 100 by far.
- Even if the sample size n is large enough to apply the asymptotic theory, the sample autocorrelations at too large lags h are meaningless since $\gamma_{n,X}(h)$ contains only $n - h$ summands. In a classical monograph on time series analysis, Box and Jenkins [7], p. 33, suggest a rule of thumb: one should not use sample autocorrelations at lags $h > n/4$.

[4] One can actually show that the condition $\mathrm{var}(X_0) < \infty$ suffices; see Brockwell and Davis [8], Theorem 7.2.2.

- Asymptotic confidence bands can be very unreliable when one deals with time series which have heavy-tailed marginal distribution. This can be observed in financial time series analysis, where it is believed that returns often do not have sufficiently high moments. Then the confidence bands based on the central limit theorem as above are not applicable. As a matter of fact, in such situations, confidence bands can be larger than the autocorrelations to be estimated and therefore the sample autocorrelations can be meaningless; see Mikosch [25] for a discussion.
- One of the basic theoretical assumptions about the interpretation of the sample autocorrelation function is the validity of the ergodic theorem. This requires strict stationarity of the underlying time series. If one studies too long time series it is not unlikely that the dependence structure of the data changes when time goes by, i.e., the data can be "rather non-stationary". Then the interpretation of the sample autocorrelation function can become rather difficult. In Example 3.11 we illustrate how structural breaks can fool one if one interprets the sample autocorrelation function uncritically.

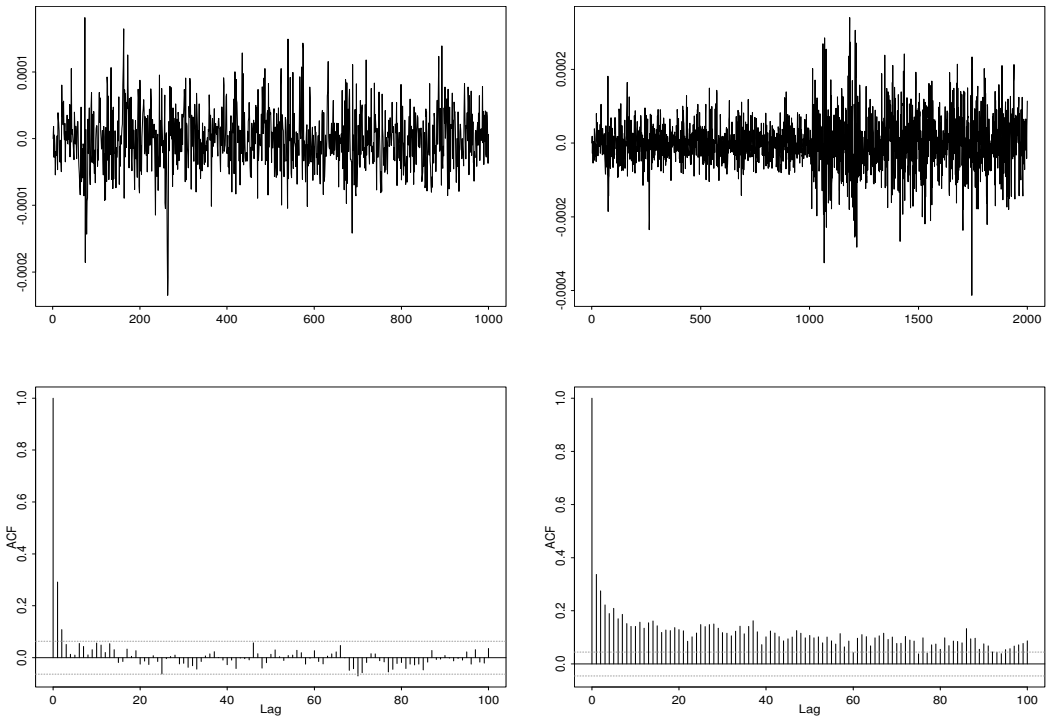

Figure 3.10. Left column: *A sample of size $n = 1000$ from a stationary ergodic model and the sample autocorrelation function of its absolute values. It decays to zero very quickly.* Right column: *Concatenation of the previous sample with another sample of size n. The second piece of the data comes from another stationary ergodic model, as can be seen by an eyeball inspection. The lower graph gives the sample autocorrelation function of the absolute values of the whole sample. One can see that this function decays to zero very slowly due to the term in $0.25|E|X_0^{(1)}| - E|X_0^{(2)}||^2$ in the limit of* (3.6)*. If one did not know where the data came from, one could interpret this effect as long-range dependence.*

Example 3.11. Consider a sample $X_1, \ldots, X_n$, where we assume that for some $p \in (0,1)$,

$$X_i^{(1)} = X_i\,, \quad i = 1, \ldots, [np]\,,$$

comes from a strictly stationary ergodic model with finite variance and expectation $EX_1^{(1)}$ and

$$X_i^{(2)} = X_i\,, \quad i = [np]+1, \ldots, n\,,$$

comes from another strictly stationary ergodic model with finite variance and expectation $EX_1^{(2)}$. Straightforward calculation and the ergodic theorem show that

$$\gamma_{n,X}(h) \stackrel{\text{a.s.}}{\to} p\,\gamma_{X^{(1)}}(h) + (1-p)\,\gamma_{X^{(2)}}(h) + p\,(1-p)\,|EX_0^{(1)} - EX_0^{(2)}|^2\,. \tag{3.6}$$

This simple calculation shows that, if there were a structural break at $t = [np]$ in an ergodic time series and the sample autocorrelation were calculated from this non-stationary sequence, the interpretation of $\gamma_{n,X}(h)$ as approximation to $\gamma_X(h)$ (which does not make sense) would be meaningless. In particular, if $EX_0^{(1)}$ and $EX_0^{(2)}$ are significantly different the sample autocorrelation function would not disappear and have a tendency to stay away from 0. If we observe this in a real-life time series we may doubt the stationarity assumption on our data and rather try to split the data into disjoint parts, where stationarity might be more appropriate.

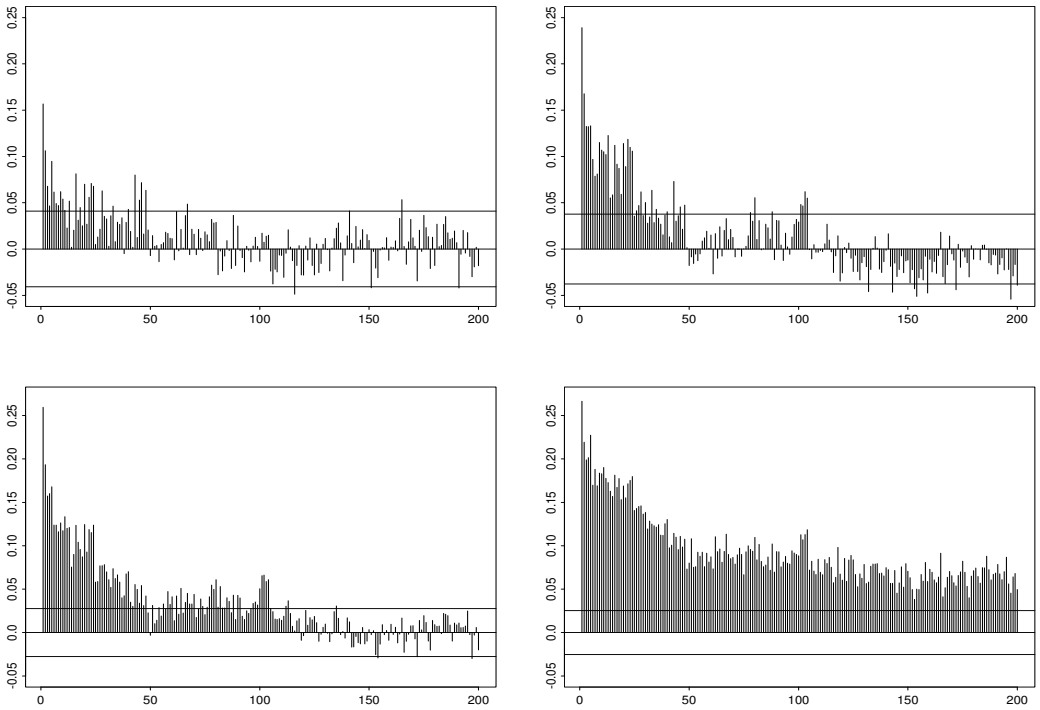

Figure 3.12. *The sample ACF for the absolute values of the daily log-returns of the first* 9 *and* 11 (top left and right), 20 *and* 24 *years* (bottom left and right) *of the major US composite stock index* S&P500 *data, starting in* 1953. *See Figure 5.1 for a visualization of the S&P series.*

For real-life data structural breaks will mostly not occur abruptly as in the previous example, but it may happen that small trends and seasonalities occur which cause the sample autocorrelation function to fool us in one way or the other. In particular, if we work with very long time series (thousands, millions,...) of data, which is not uncommon in financial time series analysis or in the analysis of teletraffic data, one may doubt the meaning of the sample autocorrelation function, since in these large data sets structural breaks or trends are likely to occur.

Over the last 30 years attempts have been made to explain the effect of non-vanishing sample autocorrelations for large lags by introducing the notion of *long-range dependence* or *long memory*. This means that a *stationary* time series (X_t) has a very slowly decaying autocorrelation function in the sense that

$$\sum_h |\rho_X(h)| = \infty \,. \tag{3.7}$$

This definition of long-range dependence is useless from a statistical point of view; it could never be checked by calculating the sample autocorrelations which vanish at lags $|h| \geq n$. Therefore sufficient conditions for (3.7) such as

$$|\rho_X(h)| = c\, h^{-d}\,(1 + o(1))\,, \quad h \to \infty\,,$$

for some $c > 0$ and $d \in (0, 1)$ were introduced which would allow one to estimate d from data by calculating $|\rho_{n,X}(h)|$ for a variety of lags h, given the sample size n is large as well. In this way, the slow decay of $|\rho_{n,X}(h)|$ can be explained in the framework of stationary time series. Given a sample, it is impossible to decide what causes the slow decay of $|\rho_{n,X}(h)|$, and therefore it is a matter of belief which theory one finds more appropriate for modeling the data at hand. Encyclopedic treatments of long-range dependence can be found in Doukhan et al. [12] and Samorodnitsky [31]. Samorodnitsky and Taqqu [32] and Brockwell and Davis [8] contain some chapters about long-range dependence and models for this phenomenon; see also p. 54 below for FARIMA long memory models.

4. ARMA processes

4.1. **Basic properties and examples.** In this section we consider a class of stationary processes which is most important in classical time series analysis: the *autoregressive moving average processes* (*ARMA processes*). Many stationary processes of interest are close to ARMA processes in the sense that their autocorrelation function can be approximated by the autocorrelation function of a suitable ARMA process. Their theory is well understood and their analysis is one of the main building blocks of all monographs on time series analysis.

Definition 4.1. (ARMA(p, q) process)
The time series $(X_t)_{t\in\mathbb{Z}}$ is said to be an ARMA(p, q) process *or* ARMA process of order (p, q) *if it is stationary and satisfies the* ARMA difference equations

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}\,, \quad t \in \mathbb{Z}\,, \tag{4.1}$$

for given real numbers $\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q$ and a white noise sequence (Z_t) with $0 < \mathrm{var}(Z_0) = \sigma^2$.

For our purposes, we will frequently assume that (Z_t) is a finite variance iid sequence. Also notice that the choice of the coefficient one in (4.1) in front of X_t and Z_t is some kind of a standardization.

Equation (4.1) can be rewritten using polynomials in the backshift operator B from p. 24. Introduce the polynomials

$$\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p \quad \text{and} \quad \theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q\,, \quad z \in \mathbb{C}\,,$$

and recall that for integers $d \geq 0$, $B^d X_t = X_{t-d}$. Then we can simply rewrite (4.1) by using the polynomials $\phi(B)$ and $\theta(B)$ in the backshift operator:

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

Example 4.2. (MA(q) process)
In Example 2.8 we learnt about the MA(q) process which is a special case of an ARMA process with $\phi(z) \equiv 1$. Then

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

There is actually nothing to solve in this difference equation and it is not difficult to see that it defines a stationary process: set $\theta_0 = 1$,

$$EX_t = 0 \quad \text{and} \quad \operatorname{cov}(X_{t+h}, X_t) = \begin{cases} \sigma^2 \sum_{j=0}^{q-|h|} \theta_j \theta_{j+|h|} & |h| \le q \,, \\ 0 & |h| > q \,. \end{cases}$$

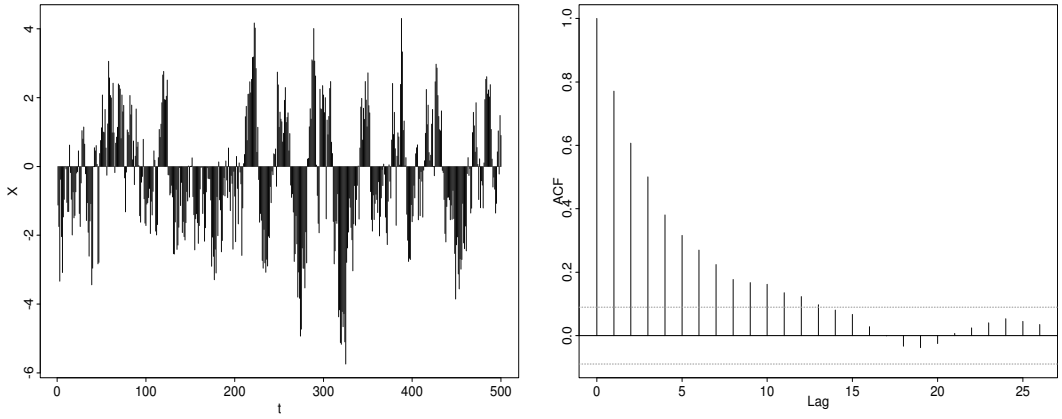

Figure 4.3. *One sample path of the* AR(1) *process* $X_t = 0.8 * X_{t-1} + Z_t$*. and its sample autocorrelation function.*

Example 4.4. (AR(p) process)
An autoregressive process of order p (AR(p) process) is given by the difference equations

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

or by the relation

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + Z_t \,, \quad t \in \mathbb{Z} \,.$$

In this case it is not obvious for which coefficients ϕ_j a stationary solution (X_t) exists. This can already be seen for the AR(1) process which is given by

$$X_t = \phi X_{t-1} + Z_t \,, \quad t \in \mathbb{Z} \,. \tag{4.2}$$

Iterating this equation we get

$$X_t = Z_t + \phi Z_{t-1} + \cdots + \phi^n Z_{t-n} + \phi^{n+1} X_{t-n-1} \,. \tag{4.3}$$

Suppose first that $|\phi| < 1$. Relation (4.3) suggests that the representation

$$X_t = \sum_{j=0}^{\infty} \phi^j Z_{t-j} \,, \quad t \in \mathbb{Z} \,, \tag{4.4}$$

holds. The random variable on the right-hand side exists as an a.s. limit since

$$E \left| \sum_{j=0}^{\infty} \phi^j Z_{t-j} \right| \le \sum_{j=0}^{\infty} |\phi|^j \, E|Z_{t-j}| \le \sup_j E|Z_j| \sum_{j=0}^{\infty} |\phi|^j \le \sigma \, (1 - |\phi|)^{-1} < \infty \,,$$

where we used Lyapunov's inequality $E|Z_j| \le (EZ_j^2)^{1/2} = \sigma$. Notice that (X_t) with representation (4.4) solves the AR(1) difference equation (4.2). Moreover, the series representation (4.4) indeed

yields a stationary time series:

$$EX_t = \sum_{j=0}^{\infty} \phi^j E Z_{t-j} = 0\,, \tag{4.5}$$

$$\begin{aligned} \operatorname{cov}(X_t, X_{t+h}) &= E\left(\sum_{j=0}^{\infty} \phi^j Z_{t-j} \sum_{k=0}^{\infty} \phi^k Z_{t+h-k}\right) \\ &= \lim_{n\to\infty} E\left(\sum_{j=0}^{n} \phi^j Z_{t-j} \sum_{k=0}^{n} \phi^k Z_{t+h-k}\right) \qquad (4.6) \\ &= \phi^{|h|}\sigma^2 \sum_{j=0}^{\infty} \phi^{2j} = \sigma^2\, \phi^{|h|}\,(1-|\phi|^2)^{-1}\,, \\ \operatorname{corr}(X_t, X_{t+h}) &= \phi^{|h|}\,. \end{aligned}$$

The justification of the interchange of expectation and limits is left as Exercise 4.5. We also leave it to show that the series representation (4.4) yields a (a.s.) *unique* stationary solution to the difference equations (4.2).

Now suppose that $|\phi| > 1$. Then we can write

$$\begin{aligned} X_t &= \phi X_{t-1} + Z_t\,, \\ \phi^{-1} X_t &= X_{t-1} + \phi^{-1} Z_t\,, \quad \text{and} \\ X_t &= -\phi^{-1} Z_{t+1} + \phi^{-1} X_{t+1}\,, \end{aligned}$$

and similar arguments as above show that

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

is the unique stationary solution of the AR(1) equation (4.2). However, this solution is usually considered as unnatural since it depends on the noise Z_{t+j} at future instants of time. For most practical applications, the condition $|\phi| < 1$ is assumed.

Exercise 4.5. i) Show that the interchange of expectation and infinite series in relations (4.5) and (4.6) is justified.
Hints: a) For (4.5) use a domination argument.
b) In order to prove (4.6) prove first that for X_t with series representation (4.4) and

$$X_t^{(n)} = \sum_{j=0}^{n} \phi^j Z_{t-j}\,,$$

$E(X_t - X_t^{(n)})^2 \to 0$ as $n \to \infty$. This means $X_t^{(n)} \overset{L^2}{\to} X_t$ as $n \to \infty$, where $\overset{L^2}{\to}$ denotes mean square convergence or convergence in the Hilbert space L^2 of mean-zero square integrable random variables. The latter space is equipped with the inner product $(X, Y) = E(XY)$ and norm $\|X\| = (EX^2)^{1/2}$. Since L^2 is a complete space, it suffices to show that for every Cauchy sequence $X_t^{(n)} - X_t^{(m)} = \sum_{j=m+1}^{n} \phi^j Z_{t-j} \overset{L^2}{\to} 0$ as $n, m \to \infty$. For the limit (4.6) also notice that the operation of inner product (X, Y) is a continuous one with respect to the distance $\|X - Y\|$ induced by the norm in L^2.

ii) Show that the stationary solution (X_t) with representation (4.4) is a.s. unique.
Hint: Assume there is another stationary solution $(\widetilde{X}_t)$ to (4.2). By iterating (4.2), show that $E|X_t - \widetilde{X}_t| = 0$, hence $X_t = \widetilde{X}_t$ a.s.
iii) Show that the AR(1) equation (4.2) does not have a stationary solution for $\phi = \pm 1$.

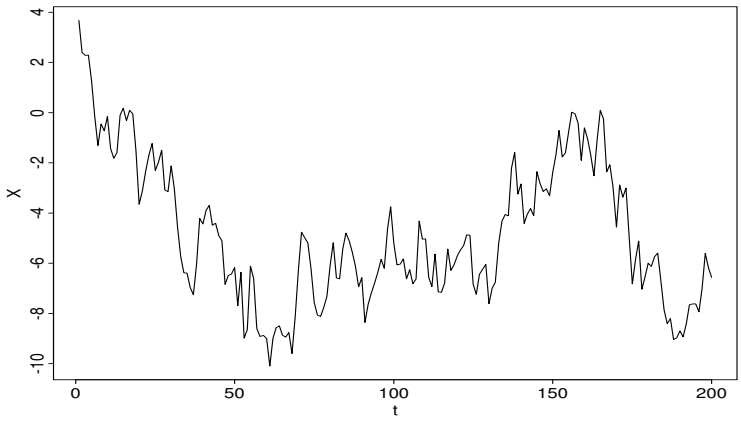

Figure 4.6. *One sample path of the non-stationary* AR(1) *process* $X_t = X_{t-1} + Z_t$ *for iid Gaussian* (Z_t).

As already mentioned, the AR(1) equations for $|\phi| > 1$ have stationary solutions which depend on the noise in the future. To exclude such phenomena the notion of *causality* is introduced:

Definition 4.7. (Causal ARMA process)
An ARMA(p, q) *process is said to be causal if it has representation*

$$X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j} \ , \quad t \in \mathbb{Z} \ , \tag{4.7}$$

for constants (ψ_j) *satisfying*

$$\sum_{j=0}^{\infty} |\psi_j| < \infty \ . \tag{4.8}$$

4.2. **Linear process representation.** The process (4.7) is a special *linear process*: (X_t) is linear if it has representation

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

given the so defined series make sense. All causal ARMA(p, q) processes have a linear series representation. The sequence (Z_t) is also called *noise* or *the innovations* of the linear process. The sequence of coefficients (ψ_j) is called a *linear filter*. In this sense, the linear filter (ψ_j) acts on the noise (Z_t) generating the time series (X_t).

The following is an auxiliary result.

Lemma 4.8. *Assume the random variables* (X_t) *satisfy the condition* $\sup_t E|X_t| < \infty$. *If* $\sum_{j=0}^{\infty} |\psi_j| < \infty$ *then the series*

$$\psi(B)X_t = \sum_{j=0}^{\infty} \psi_j B^j X_t = \sum_{j=0}^{\infty} \psi_j X_{t-j}$$

converges absolutely with probability 1.
If in addition $\sup_t EX_t^2 < \infty$ *then this series converges in mean square to the same limit.*

The statement about the a.s. convergence is a consequence of the fact that

$$E\left[\sum_{j=0}^{\infty} |\psi_j||X_{t-j}|\right] \le \sup_t E|X_t| \sum_{j=0}^{\infty} |\psi_j| < \infty \ .$$

The assertion about mean square convergence can be checked by the Cauchy convergence criterion in L^2 in the spirit of Exercise 4.5.

We continue with another auxiliary result:

Proposition 4.9. *Assume* (X_t) *is a stationary time series with autocovariance function* $(\gamma_X(h))_{h\in\mathbb{Z}}$. *If* $\sum_{j=0}^{\infty} |\psi_j| < \infty$ *then*

$$Y_t = \psi(B)X_t = \sum_{j=0}^{\infty} \psi_j X_{t-j}$$

converges for each t *with probability* 1 *and in mean square. Moreover, the process* (Y_t) *is stationary with autocovariance function*

$$\gamma_Y(h) = \sum_{j,k=0}^{\infty} \psi_j \, \psi_k \, \gamma_X(h-j+k) \ .$$

Proof. The convergence statement follows from Lemma 4.8 since $\sup_t EX_t^2 = EX_0^2 < \infty$. Next we check stationarity:

$$\begin{aligned}
EY_t &= EX_t \sum_{j=0}^{\infty} \psi_j = EX_0 \sum_{j=0}^{\infty} \psi_j \ , \\
E(Y_t Y_{t+h}) &= E\left[\sum_{j=0}^{\infty} \psi_j X_{t+h-j} \sum_{k=0}^{\infty} \psi_k X_{t-k}\right] \\
&= \sum_{j,k=0}^{\infty} \psi_j \psi_k E(X_{h-j} X_{-k}) \\
&= \sum_{j,k=0}^{\infty} \psi_j \psi_k [\gamma_X(h-j+k) + (EX_0)^2] \\
&= \sum_{j,k=0}^{\infty} \psi_j \psi_k \gamma_X(h-j+k) + \left(EX_0 \sum_{j=0}^{\infty} \psi_j\right)^2 \ .
\end{aligned}$$

As in the case of an AR(1) process we would have to justify the interchange of expectation and infinite series. This can be done in the spirit of Exercise 4.5. □

The next result is important since it tells us under which conditions an ARMA process is causal and what its representation as a linear process is.

Theorem 4.10. (Criterion for causality of an ARMA process)
Let (X_t) *be an* ARMA(p,q) *process such that* $\phi(z)$ *and* $\theta(z)$ *have no common zeros for all complex*

z. Then (X_t) is causal if and only if $\phi(z) \neq 0$, $z \in \mathbb{C}, |z| \leq 1$. The coefficients (ψ_j) in the linear process representation $X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}$ are then determined by the relation

$$\psi(z) = \sum_{j=0}^{\infty} \psi_j z^j = \frac{\theta(z)}{\phi(z)} , \quad |z| \leq 1 . \tag{4.9}$$

Remark 4.11. 1) Notice that determining (ψ_j) via (4.9) makes sense since the coefficients of a convergent power series are unique.
2) Assume that $\alpha(z) = \sum_{j=0}^{\infty} \alpha_j z^j$ and $\beta(z) = \sum_{j=0}^{\infty} \beta_j z^j$ are two power series such that $\sum_{j=0}^{\infty}(|\alpha_j| + |\beta_j|) < \infty$. Then $\psi(z) = \alpha(z)\beta(z)$ has again a power series representation $\psi(z) = \sum_{j=0}^{\infty} \psi_j z^j$ which converges for $|z| \leq 1$. It is also reasonable to define the backshift power series

$$\alpha(B)\beta(B)X_t = \psi(B)X_t .$$

Sketch of the proof. We restrict ourselves to the sufficiency part. Assume that $\phi(z) \neq 0$ for $|z| \leq 1$. By continuity of ϕ, $\phi(z) \neq 0$ for $|z| \leq 1 + \varepsilon$ for some $\varepsilon > 0$. Then we can divide by $\phi(z)$ and the function $1/\phi(z)$ has again a power series representation for $|z| \leq 1 + \varepsilon$:

$$1/\phi(z) = \sum_{j=0}^{\infty} \xi_j z^j \equiv \xi(z) , \quad |z| \leq 1 + \varepsilon . \tag{4.10}$$

This is a fact from complex function theory. It also follows that the coefficients (ξ_j) satisfy $|\xi_j| < a^j$ for some $a < 1$ and large j. Indeed, for the convergence of the series in (4.10) the condition $|\xi_j|(1 + \varepsilon)^j \to 0$ is necessary. Hence $\sum_{j=0}^{\infty} |\xi_j| < \infty$. Moreover, $\xi(z)\phi(z) \equiv 1$. An application of $\xi(B)$ to both sides of the ARMA equations $\phi(B)X_t = \theta(B)Z_t$ is justified by Proposition 4.9:

$$\xi(B)\phi(B)X_t = X_t = \xi(B)\theta(B)Z_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j} .$$

Remarks. 1) Since

$$\frac{1}{\phi(z)} = \sum_{j=0}^{\infty} \xi_j z^j , \quad |z| < 1 + \varepsilon,$$

converges and $\theta(z)$ is a finite polynomial in z,

$$\psi(z) = \frac{\theta(z)}{\phi(z)} = \sum_{j=0}^{\infty} \psi_j z^j , \quad |z| < 1 + \varepsilon ,$$

is also a convergent power series. By the same argument as in the proof above, it follows that

$$|\psi_j| < a^j \quad \text{for some} \quad 0 < a < 1 , \quad \text{large } j . \tag{4.11}$$

2) Since $X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}$ and (Z_t) is a white noise process, it follows immediately from Proposition 4.9 that (X_t) is stationary. Moreover, for a white noise process

$$\gamma_Z(h) = \begin{cases} 0 & |h| \geq 1 , \\ \sigma^2 & h = 0 . \end{cases}$$

Hence we obtain from Proposition 4.9 the formulae

$$\text{var}(X_0) = \sigma^2 \sum_{j=0}^{\infty} \psi_j^2 \,, \tag{4.12}$$

$$\gamma_X(h) = \sigma^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+|h|} \,, \quad h \in \mathbb{Z} \,. \tag{4.13}$$

$$\rho_X(h) = \frac{\sum_{j=0}^{\infty} \psi_j \psi_{j+|h|}}{\sum_{j=0}^{\infty} \psi_j^2} \,, \quad h \in \mathbb{Z} \,.$$

From Theorem 2.30 and the fact that an iid sequence (Z_t) is strictly stationary and ergodic we then conclude that a linear process (X_t) is strictly stationary and ergodic.

3) The linear process representation and the formulae (4.12) and (4.13) make it evident that the knowledge of the coefficients ψ_j is crucial for the understanding of the dependence structure of a concrete ARMA process. There exist many ways of calculating these coefficients. We refer to Brockwell and Davis [8], Section 3.3, for the general case and restrict ourselves to the calculation of one example.

Example 4.12. (Calculation of the coefficients of an ARMA(2,1) process)
Assume

$$(1 - B + \frac{1}{4}B^2)X_t = (1 + B)Z_t \,.$$

Then $\phi(z) = 1 - z + \frac{1}{4}z^2, \theta(z) = 1 + z$,

$$\psi(z) = \frac{\theta(z)}{\phi(z)} \,, \quad \phi(z)\psi(z) = \theta(z) \,.$$

Hence

$$(1 - z + \frac{1}{4}z^2)(\psi_0 + \psi_1 z + \psi_2 z^2 + \psi_3 z^3 + \cdots) = 1 + z \,.$$

Comparing the coefficients on the left-hand and right-hand sides, we obtain

$$\begin{aligned} 1 &= \psi_0 \,, \\ 1 &= -\psi_0 + \psi_1 \,, \quad \psi_1 = 2 \,, \\ 0 &= \frac{1}{4}\psi_0 - \psi_1 + \psi_2 \,, \quad \psi_2 = 1.75 \,, \\ \cdots &\quad \cdots \end{aligned}$$

By Remark 1 above, the coefficients ψ_j decrease exponentially fast, and so only the first few ψ_j's are relevant for determining the autocovariance function of the process.

There is still another concept for ARMA processes which is analogous to causality, the so-called *invertibility* of an ARMA process.

Definition 4.13. (Invertible ARMA process)
An $ARMA(p, q)$ is called invertible *if there exists a sequence of constants (π_j) such that $\sum_{j=0}^{\infty} |\pi_j| < \infty$ and*

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

The following result is similar to the characterization of causality.

Theorem 4.14. (Characterization of an invertible ARMA process)
Let (X_t) be an ARMA(p,q) *process such that $\phi(z)$ and $\theta(z)$ do not have common zeros for all complex z. Then (X_t) is invertible if and only if $\theta(z) \neq 0$, $z \in \mathbb{C}, |z| \leq 1$. The coefficients (π_j) in the linear process representation $Z_t = \sum_{j=0}^{\infty} \pi_j X_{t-j}$ are then determined by the relation*

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

The proof of this theorem is similar to the one for Theorem 4.10. We can merge Theorems 4.10 and 4.14 to the following

Corollary 4.15. (Characterization of a causal, invertible ARMA process)
Let (X_t) be an ARMA(p,q) *process such that $\phi(z)$ and $\theta(z)$ do not have common zeros for all complex z. Then (X_t) is causal and invertible if and only if $\phi(z)\theta(z) \neq 0$, $z \in \mathbb{C}, |z| \leq 1$.*

The notions of causality and invertibility are important when it comes to estimating the parameters of an ARMA process. Most of the estimation theory is based on causal and invertible ARMA processes. Invertibility is also relevant if one wants to judge the goodness-of-fit of an ARMA process: this property allows one to estimate the innovations (so-called *residuals*) and test whether they are close in some sense to white noise or to an iid sequence.

4.3. **Estimation of ARMA processes.** In this section we consider some estimation procedures for ARMA and, more generally, for stationary processes. Let (X_t) be a stationary process. Which quantities can be of interest to be estimated?

- The mean value $\mu = EX_t = EX_0$.
- The autocovariances $\gamma_X(h) = \mathrm{cov}(X_0, X_h)$.
- The autocorrelations $\rho_X(h) = \mathrm{corr}(X_0, X_h)$.
- The innovation variance $\sigma^2 = \mathrm{var}(Z_0)$.

For an ARMA process

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q} , \quad t \in \mathbb{Z} ,$$

it is of interest to estimate

- the parameters

$$\beta = (\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q)'$$

as well as

- the order (p,q).

The estimation of the order (p,q) is difficult; it is based on so-called *information criteria*. Among them the *Akaike* (*AIC*) and *Bayesian* (*BIC*) *criteria* are best known. Roughly speaking, these information criteria estimate the parameters of the ARMA process by adding a penalty term to the likelihood function, which shall avoid over-parameterization of the model; see Brockwell and Davis [8], Section 9.2, for an introduction to this topic and p. 50 below for an example.

Estimation of μ. Given a sample $X_1, \ldots, X_n$ from a stationary process, a natural estimator of μ is given by its sample mean $\overline{X}_n$. Obviously, $\overline{X}_n$ is an unbiased estimator of μ.

If (X_t) is strictly stationary and ergodic, in particular, if the noise (Z_t) is iid, the ergodic theorem yields $\overline{X}_n \overset{\text{a.s.}}{\to} \mu$, provided μ is defined. If, in addition, (X_t) satisfies a strong mixing condition, it is also asymptotically normally distributed in view of Ibragimov's central limit theorem. Strong mixing conditions are often difficult to verify for ARMA processes; see Doukhan [11]. Then it is easier to exploit the particular structure of the ARMA processes to derive asymptotic normality; see Chapter 7 in [8].

Example 4.16. ARMA processes with iid white noise (Z_t) are strongly mixing if Z_0 has a positive Lebesgue density. Of course, MA(m) processes are m-dependent, hence strongly mixing with vanishing rate function for large lags. Even for the simple stationary causal AR(1) process $X_t = \phi X_{t-1} + Z_t$, $t \in \mathbb{Z}$, strong mixing is not easily verified unless Z_0 has a density. In particular, an AR(1) process with Bernoulli noise (Z_t) is not strongly mixing. Nevertheless, the central limit theorem for $\overline{X}_n$ and the sample autocovariances and autocorrelations holds; Brockwell and Davis [8], Chapter 7.

If we give up the condition of ergodicity and only assume stationarity of (X_t), it is not difficult to see that, under general conditions, $\overline{X}_n$ is a consistent estimator of μ; see p. 17 for a proof:

Proposition 4.17. (Consistency of $\overline{X}_n$)
Let (X_t) be a stationary process. If the condition $\gamma_X(n) \to 0$ holds, then $\overline{X}_n \xrightarrow{P} \mu$ and $\overline{X}_n \xrightarrow{L^2} \mu$.

Example 4.18. Let $(X_t - \mu)$ be a causal mean-zero ARMA process driven by white noise (Z_t) with variance σ^2. We know from Section 4.1 that $X_t - \mu = \sum_{j=0}^{\infty} \psi_j Z_{t-j}$ and that $|\psi_j| \le K a^j$ for a positive constant K, a constant $a < 1$ and for all j; see (4.11). Hence we conclude by the Cauchy-Schwarz inequality that

$$\begin{aligned} |\gamma_X(h)| &= \sigma^2 \left| \sum_{j=0}^{\infty} \psi_j \psi_{j+|h|} \right| \le \sigma^2 \left(\sum_{j=0}^{\infty} \psi_j^2 \sum_{i=0}^{\infty} \psi_{i+|h|}^2 \right)^{1/2} \\ &\le K\sigma^2 \left(\sum_{j=0}^{\infty} a^{2j} \sum_{i=0}^{\infty} a^{2(i+|h|)} \right)^{1/2} \le const\ a^{|h|} \,. \end{aligned} \tag{4.14}$$

Hence $\gamma_X(n) \to 0$ and by Proposition 4.17, $\overline{X}_n \to \mu$ both in mean square and in probability.

For an ARMA process driven by iid noise (Z_t) with variance σ^2 one can show that $(\sqrt{n}(\overline{X}_n - \mu)) \xrightarrow{d} Y$ for a Gaussian $\mathrm{N}(0, \sum_h \gamma_X(h))$ random variable Y; see Theorem 7.1.2 in Brockwell and Davis [8]. Notice that

$$\begin{aligned} \mathrm{var}(\sqrt{n}\,\overline{X}_n) &\to \sum_{h=-\infty}^{\infty} \gamma_X(h) = \gamma_X(0) + 2\sum_{h=1}^{\infty} \gamma_X(h) \\ &= \sigma^2 (\sum_j \psi_j)^2 \,. \end{aligned}$$

The first identity was proved on p. 17 under the assumption that γ_X is absolutely summable. The second identity follows from the fact that $\gamma_X(h) = \sigma^2 \sum_j \psi_j \psi_{j+|h|}$, $h \in \mathbb{Z}$.

For ARMA and, more generally, linear processes $X_t = \sum_j \psi_j Z_{t-j}$, $t \in \mathbb{Z}$, with iid white noise limit theory for $\overline{X}_n$, $\gamma_{n,X}$, etc., can be derived by exploiting the linear structure of the process. We illustrate this for the sample mean of an MA(2) process:

$$\begin{aligned} \overline{X}_n &= \frac{1}{n} \sum_{t=1}^{n} \big(Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} \big) \\ &= \overline{Z}_n + \theta_1 \big(\overline{Z}_n - (Z_n - Z_0)/n \big) + \theta_2 \big(\overline{Z}_n - (Z_n + Z_{n-1} - Z_0 - Z_{-1})/n \big) \\ &= \overline{Z}_n \big(1 + \theta_1 + \theta_2 \big) - \frac{1}{n} \big(\theta_1 (Z_n - Z_0) + \theta_2 (Z_n + Z_{n+1} - Z_0 - Z_{-1}) \big) \,. \end{aligned}$$

We have $Z_i/n \xrightarrow{P} 0$ as $n \to \infty$ for any choice of i. Therefore and by the central limit theorem for $\overline{Z}_n$,

$$\sqrt{n}\overline{X}_n \xrightarrow{d} N\big(0, \sigma^2(1+\theta_1+\theta_2)^2\big)\,, \qquad n \to \infty\,.$$

Similarly, $Z_n/n \xrightarrow{\text{a.s.}} 0$ and $\overline{Z}_n \xrightarrow{\text{a.s.}} 0$ by the strong law of large numbers for an iid sequence. Therefore

$$\overline{X}_n \xrightarrow{\text{a.s.}} 0\,, \qquad n \to \infty\,.$$

Estimation of the autocorrelations. According to Section 3.2, natural estimators of the autocovariances are given by the sample autocovariances

$$\gamma_{n,X}(h) = \frac{1}{n}\sum_{t=1}^{n-|h|}(X_t - \overline{X}_n)(X_{t+|h|} - \overline{X}_n)\,, \quad |h| < n\,,$$

from which we can determine the sample autocorrelations

$$\rho_{n,X}(h) = \gamma_{n,X}(h)/\gamma_{n,X}(0)\,, \quad |h| < n\,.$$

If (X_t) is strictly stationary and ergodic, in particular, if (Z_t) is iid, the ergodic theorem yields consistency of $\gamma_{n,X}(h)$, hence of $\rho_{n,X}(h)$, provided $\text{var}(X_0) < \infty$; see Exercise 2.33. The asymptotic normality again depends on strong mixing conditions which are not easily verified. Using the particular *linear* structure of the ARMA process, one can avoid these conditions. We cite here a result which can be found in [8], Theorem 7.2.1.

Theorem 4.19. (Asymptotic normality of the sample autocorrelations)
Let (Z_t) be iid noise with $EZ_1 = 0, \sigma^2 = \text{var}(Z_1), EZ_1^4 < \infty$. Suppose (X_t) is a linear process with representation $X_t - \mu = \sum_{j=-\infty}^{\infty}\psi_j Z_{t-j}$ and $\sum_j |\psi_j| < \infty$. Then the relation

$$\sqrt{n}\,(\rho_{n,X}(h) - \rho_X(h))_{h=1,\dots,m} \xrightarrow{d} \mathbf{Y} = (Y_h)_{h=1,\dots,m}$$

holds, where $\mathbf{Y}$ is $\text{N}(\mathbf{0}, W)$ and the covariance matrix W is given by Bartlett's formula*:*

$$\begin{aligned} w_{ij} &= \sum_{k=1}^{\infty}[\rho_X(k+i) + \rho_X(k-i) - 2\rho_X(i)\rho_X(k)] \times \\ &\qquad \times[\rho_X(k+j) + \rho_X(k-j) - 2\rho_X(j)\rho_X(k)]\,. \end{aligned} \tag{4.15}$$

The assumptions of this theorem are satisfied for a causal ARMA process. Hence the sample autocorrelations of a causal ARMA process are consistent and asymptotically normally distributed estimators of the underlying autocorrelations.

The proof of this statement is given in Section 7.3 of [8]. It is based on the observation, derived from Ibragimov's central limit theorem, that

$$\frac{1}{\sigma^2\sqrt{n}}\left(\sum_{t=1}^{n-h} Z_t\, Z_{t+h}\right)_{h=1,\dots,m} \xrightarrow{d} (N_h)_{h=1,\dots,m}$$

for iid standard normal random variables (N_h), W is then the covariance matrix of the vector $(Y_h)_{h=1,\dots,m}$, where

$$Y_i = \sum_{k=1}^{\infty}\left[\rho_X(k+i) + \rho_X(k-i) - 2\rho_X(i)\rho_X(k)\right] N_k\,.$$

Bartlett's formula is particularly important for determining the variance of the considered estimators and for constructing confidence intervals for the sample autocorrelations. Notice that standard software does not give one confidence bands for the sample autocorrelations of a fitted ARMA processes but for an iid white noise sequence.

Example 4.20. (IID white noise)
For iid mean-zero Z_t's, Bartlett's formula yields

$$w_{ij} = \begin{cases} 1 & i = j \ , \\ 0 & \text{otherwise} \ . \end{cases}$$

This is due to the fact that $\rho_X(h) = 0$ for $h \neq 0$.

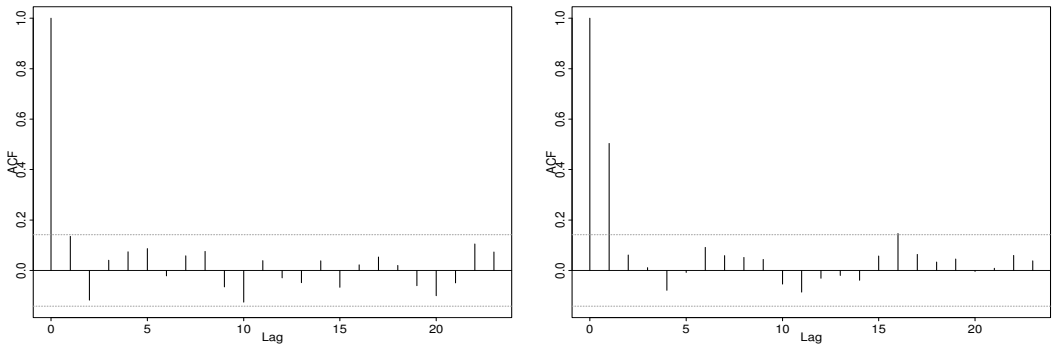

Figure 4.21. Left: *Sample autocorrelation function for iid white noise* ($n = 200$). *The lines parallel to the x-axis give* 95% *asymptotic confidence bands* $\pm 1.96/\sqrt{n}$. Right: *Sample autocorrelation function of the* MA(1) *process* $X_t = Z_t + 0.8Z_{t-1}$ *with sample size* $n = 200$. *The lines parallel to the x-axis give* 95% *asymptotic confidence bands* $\pm 1.96/\sqrt{n}$ *for iid white noise. This is standard in all statistical packages. The autocorrelation at lag* 1 *clearly indicates that we have an* MA(1) *process; the other autocorrelations are not significant.*

Example 4.22. (MA(q)-process)
Consider the MA(q)-process

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

driven by iid white noise. Bartlett's formula gives the asymptotic variance

$$w_{ii} = [1 + 2\rho_X^2(1) + \cdots + 2\rho_X^2(q)] \ , \quad i > q \ ,$$

which is the variance of the normal limit of the sequence $(\sqrt{n}(\rho_{n,X}(i) - \rho_X(i))) = (\sqrt{n}\rho_{n,X}(i))$ for $i > q$. This means in particular, that the confidence bands for lags $i > q$ must be wider than those for iid white noise shown by professional software.

Exercise 4.24. i) Calculate w_{ii} for an AR(1) process as well as the limit of w_{ii} as $i \to \infty$.
ii) Simulate a sample of size $n = 200$ from the AR(1) process $X_t = 0.8X_{t-1} + Z_t$ for iid standard normal white noise (Z_t) (use arima.sim in R). Draw the sample autocorrelation function for (X_t) with maximal lag $h = 25$ and asymptotic confidence bands for iid white noise (this is standard in the function acf in R). Then draw in the same graph (use the function lines in R) the 95% asymptotic confidence bands based on the calculations of i).

The Yule-Walker estimates. Now we consider one of the most important parameter estimates in time series analysis: the *Yule-Walker estimates*. Assume we have a causal AR(p) process:

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = Z_t \ , \quad t \in \mathbb{Z} \ , \tag{4.16}$$

driven by white noise with variance σ^2. By causality, we can write

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

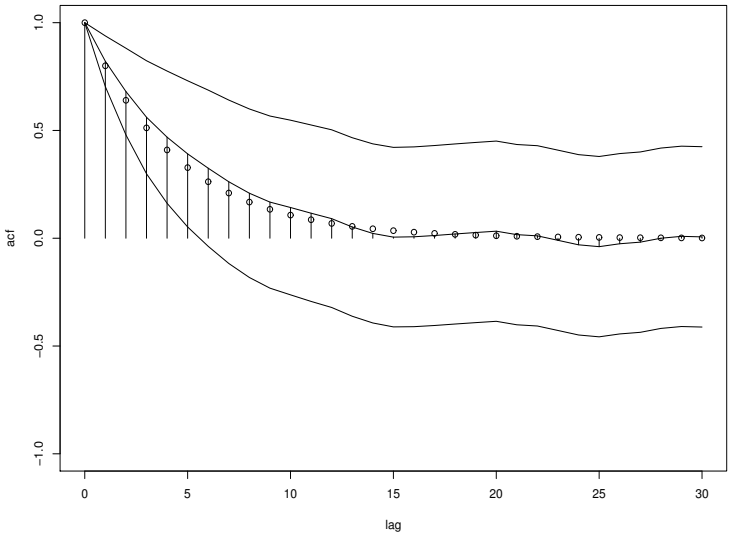

Figure 4.23. *The sample autocorrelation function of an* AR(1) *process* $X_t = 0.8X_{t-1} + Z_t$ *for iid Gaussian white noise. The asymptotic confidence band at lag* h *is derived from Bartlett's formula:* $\rho_{n,X}(h) \pm 1.96\sqrt{w_{hh}/n}$, $n = 1000$. *The dots correspond to the theoretical autocorrelations* $\rho_X(h) = 0.8^h$.

A consequence of this relation is that

$$E[Z_t X_{t-i}] = \begin{cases} \sigma^2 & i = 0\,, \\ 0 & i > 0\,. \end{cases} \tag{4.17}$$

Now multiply (4.16) by $X_{t-i}, i = 0, \ldots, p$, and then take expectations on both sides. Using (4.17), we conclude that

$$\begin{aligned} \sigma^2 &= \gamma_X(0) - \phi_1\gamma_X(1) - \cdots - \phi_p\gamma_X(p)\,, \\ 0 &= \gamma_X(1) - \phi_1\gamma_X(0) - \cdots - \phi_p\gamma_X(p-1)\,, \\ \vdots &= \vdots \\ 0 &= \gamma_X(p) - \phi_1\gamma_X(p-1) - \cdots - \phi_p\gamma_X(0)\,. \end{aligned}$$

Writing

$$\begin{aligned} \phi &= (\phi_1, \ldots, \phi_p)'\,, \\ \Gamma_p &= (\gamma_X(i-j))_{i,j=1,\ldots,p}\,, \\ \gamma_p &= (\gamma_X(1), \ldots, \gamma_X(p))'\,, \end{aligned}$$

we thus obtain the $p+1$ linear equations

$$\begin{aligned} \sigma^2 &= \gamma_X(0) - \phi'\gamma_p\,, \\ \Gamma_p\,\phi &= \gamma_p\,. \end{aligned}$$

We argued in Proposition 3.3 that Γ_p^{-1} exists since $\gamma_X(0) > 0$ and $\gamma_X(h) \to 0$ as $h \to \infty$ are satisfied for a stationary autoregressive process. Hence

$$\begin{aligned} \phi &= \Gamma_p^{-1} \gamma_p \,, \\ \sigma^2 &= \gamma_X(0) - \gamma_p' \Gamma_p^{-1} \gamma_p \,. \end{aligned}$$

Replacing γ_p by

$$\gamma_{n,p} = (\gamma_{n,X}(1), \ldots, \gamma_{n,X}(p))'$$

and Γ_p by

$$\hat{\Gamma}_{n,p} = (\gamma_{n,X}(i-j))_{i,j=1,\ldots,p}$$

we arrive at the *Yule-Walker equations* for $\widehat{\sigma}^2$ and $\widehat{\phi}$:

$$\begin{aligned} \widehat{\sigma}^2 &= \gamma_{n,X}(0) - \widehat{\phi}' \gamma_{n,p} \,, \\ \hat{\Gamma}_{n,p} \, \widehat{\phi} &= \gamma_{n,p} \,. \end{aligned}$$

Following again Proposition 3.3 and the discussion in Remark 3.6 (we have $\gamma_{n,X}(0) > 0$ for sufficiently large n), we may conclude that the inverse $\hat{\Gamma}_{n,p}^{-1}$ exists. Hence there exists a unique solution $\hat{\phi}$ to the Yule-Walker equations. Thus, writing

$$\begin{aligned} \rho_{n,p} &= \gamma_{n,p} / \gamma_{n,X}(0) \,, \\ \hat{R}_{n,p} &= \hat{\Gamma}_{n,p} / \gamma_{n,X}(0) \,, \end{aligned}$$

we obtain the equations

$$\begin{aligned} \widehat{\sigma}^2 &= \gamma_{n,X}(0) \, (1 - \rho_{n,p}' \hat{R}_{n,p}^{-1} \rho_{n,p}) \,, \\ \widehat{\phi} &= \hat{R}_{n,p}^{-1} \, \rho_{n,p} \,. \end{aligned}$$

The evaluation of the Yule-Walker equations can be done via the *Levinson-Durbin recursive algorithm* or the *innovations algorithm*. For details see Brockwell and Davis [8], Section 8.2, or Section 7.3 below.

Example 4.25. (Yule-Walker estimates for AR(1) process)
Let $X_t - \phi X_{t-1} = Z_t$ be causal, i.e., $|\phi| < 1$. The Yule-Walker equations are then given by

$$\begin{aligned} \widehat{\sigma}^2 &= \gamma_{n,X}(0) - \widehat{\phi} \, \gamma_{n,X}(1) \,, \\ 0 &= \gamma_{n,X}(1) - \widehat{\phi} \, \gamma_{n,X}(0) \,. \end{aligned}$$

Easy calculation yields

$$\begin{aligned} \widehat{\phi} &= \gamma_{n,X}(1) / \gamma_{n,X}(0) = \rho_{n,X}(1) \,, \\ \widehat{\sigma}^2 &= \gamma_{n,X}(0) \, [1 - \rho_{n,X}^2(1)] \,. \end{aligned} \tag{4.18}$$

Exercise 4.26. Prove the consistency of $\widehat{\phi}$ and $\widehat{\sigma}^2$ for an AR(1) model with iid white noise (Z_t) innovations. Also prove asymptotic normality for $\hat{\phi}$ and $\widehat{\sigma}^2$, assuming $E[Z_0^4] < \infty$. Hint: use Bartlett's central limit theorem; see Theorem 4.19.

From the construction of the Yule-Walker estimate it is clear that it is a method of moment estimate – the theoretical covariances are replaced by their sample counterparts.

For a causal AR(p) process driven by iid noise (Z_t) with variance σ^2,

$$\sqrt{n} \, (\widehat{\phi} - \phi) \xrightarrow{d} \mathbf{Y}_p$$

with a Gaussian $\mathrm{N}(\mathbf{0}, \sigma^2\Gamma_p^{-1})$ random vector $\mathbf{Y}_p$. This allows one to construct asymptotic confidence bands for $\widehat{\phi}$. Moreover, $\widehat{\sigma}^2 \xrightarrow{P} \sigma^2$; see Theorem 8.1.1 in [8].

Gaussian maximum likelihood. The Yule-Walker estimates are restricted to AR(p) processes. However, there exist also several estimation procedures for general causal, invertible ARMA processes:

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q} \; .$$

We write

$$\beta = (\phi_1, \ldots, \phi_p \, ; \theta_1, \ldots, \theta_q)'$$

for the corresponding parameter vector.

One estimation method is given by the *Gaussian maximum likelihood approach*: suppose for the moment that (Z_t) is iid mean-zero Gaussian with variance σ_0^2. Assume that the sample $X_1, \ldots, X_n$ comes from an ARMA(p, q) process with true parameter β_0 from the parameter space

$$\begin{aligned} C \;\; = \;\; & \{\beta \in \mathbb{R}^{p+q} : \phi(z)\theta(z) \neq 0 \; , |z| \leq 1 \text{ and } \phi(\cdot) \text{ and } \theta(\cdot) \\ & \text{do not have common zeros.}\} \end{aligned}$$

We assume $\beta_0 \in C$ which means that (X_t) is causal and invertible; see Corollary 4.15.

Example 4.27. Assume (X_t) is a causal AR(1) process. Then $\beta = \phi_1 \in C = (-1, 1)$. Assume (X_t) is an MA(q) process, $X_t = Z_t + \theta_1 Z_{t-1}$. This process is causal. To ensure invertibility we need that $\theta(z) = 1 + \theta_1 z$ does not have zeros for $|z| \leq 1$. We have $\theta(z) = 0$ if and only if $z = -1/\theta_1$. Thus (X_t) is invertible if and only if $|\theta_1| < 1$. Then $\beta = \theta_1 \in C = (-1, 1)$.

Let $\Gamma_n(\beta_0, \sigma_0^2)$ be the (non-singular) covariance matrix of $\mathbf{X} = (X_1, \ldots, X_n)'$. Then $\mathbf{X}$ has density

$$L(\beta_0, \sigma_0^2)(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}(\det \Gamma_n(\beta_0, \sigma_0^2))^{1/2}} \exp\left\{-\frac{1}{2}\, \mathbf{x}'\, \Gamma_n^{-1}(\beta_0, \sigma_0^2)\, \mathbf{x}\right\} \; , \quad \mathbf{x} \in \mathbb{R}^n \; ;$$

see (2.2) on p. 10. Taking logarithms and plugging in $\mathbf{X}$, we obtain,

$$-2 \log L(\beta_0, \sigma_0^2)(\mathbf{X}) = n \log(2\pi) + \log \det \Gamma_n(\beta_0, \sigma_0^2) + \mathbf{X}' \Gamma_n^{-1}(\beta_0, \sigma_0^2) \mathbf{X} \; .$$

Since we do not know β_0 and σ_0^2, a natural (maximum likelihood) argument suggests to minimize the quantity

$$\log \det \Gamma_n(\beta, \sigma^2) + \mathbf{X}' \Gamma_n^{-1}(\beta, \sigma^2) \mathbf{X}$$

with respect to $\beta \in C$ and $\sigma^2 > 0$. The so-defined estimator (β_n, σ_n^2) of (β_0, σ_0^2) is called the *Gaussian maximum likelihood estimator of* (β_0, σ_0^2). It is not given explicitly; for its calculation one depends on numerical optimization procedures.

For a sample $\mathbf{X}$ with a general (non-Gaussian) distribution we do in general not know the underlying density (if it exists at all), thus we cannot determine the likelihood function. However, we can take instead the Gaussian likelihood function, plug in $\mathbf{X}$ and minimize $-2 \log L(\beta, \sigma^2)(\mathbf{X})$ with respect to $\beta \in C$ and $\sigma^2 > 0$. This concept works for a surprisingly large class of time series; see Section 10.8 in Brockwell and Davis [8]:

Theorem 4.28. (Asymptotic normality of the Gaussian maximum likelihood estimator)
If (X_t) is a causal invertible ARMA process with true parameter $\beta_0 \in C$ driven by iid white noise (Z_t), then

$$\sqrt{n}\, (\beta_n - \beta_0) \xrightarrow{d} \mathbf{Y}_{p+q}$$

for a Gaussian $\mathrm{N}(0, W(\beta_0))$ *vector $\mathbf{Y}_{p+q}$, where the covariance matrix $W(\beta_0)$ can be expressed via the spectral density (to be defined in Section 6; the spectral density is a function which depends on the autocorrelation function of the* ARMA(p, q) *model corresponding to the parameter vector β_0) of the underlying process. Moreover, $\sigma_n^2 \xrightarrow{\text{a.s.}} \sigma_0^2$.*

In general, one depends on numerical methods for determining the estimators β_n and σ_n^2 from the likelihood equations. There is one simple case where we can determine the Gaussian maximum likelihood estimator explicitly.

Example 4.29. (Gaussian maximum likelihood for an AR(1) process) Consider the AR(1) process $X_t = \phi X_{t-1} + Z_t$ with an iid white noise (Z_t) with positive variance σ^2. This process constitutes a *Markov chain* $(X_n)_{n\geq 0}$ *with continuous state space*, i.e.,

$$P\big(X_{n+1} \in A \mid X_0, \ldots, X_n\big) = P\big(X_{n+1} \in A \mid X_n\big)\,,$$

for any choice of Borel sets $A \subset \mathbb{R}$, $n \geq 0$, and X_n may assume any values in $\mathbb{R}$. In particular, we have

$$P\big(X_{n+1} \leq y \mid X_n = x\big) = P\big(\phi x + Z_{n+1} \leq y \mid X_n = x\big) = P\big(Z_0 \leq y - \phi x\big)\,.$$

Here we used the independence between Z_{n+1} and X_n. These *transition probabilities* do not depend on n and therefore we deal with a *homogeneous Markov chain*. Now we write the joint density of $\mathbf{X}_n = (X_1, \ldots . X_n)'$ as follows (assuming that all densities are positive)

$$f_{\mathbf{X}_n}(x_1, \ldots, x_n) = \frac{f_{\mathbf{X}_n}(x_1, \ldots, x_n)}{f_{\mathbf{X}_{n-1}}(x_1, \ldots, x_{n-1})} \frac{f_{\mathbf{X}_{n-1}}(x_1, \ldots, x_{n-1})}{f_{\mathbf{X}_{n-2}}(x_1, \ldots, x_{n-2})} \cdots \frac{f_{\mathbf{X}_2}(x_1, x_2)}{f_{X_1}(x_1)} f_{X_1}(x_1)\,.$$

(4.19)

The ratios

$$\frac{f_{\mathbf{X}_k}(x_1, \ldots, x_k)}{f_{\mathbf{X}_{k-1}}(x_1, \ldots, x_{k-1})}, \quad k = 2, \ldots, n\,, \tag{4.20}$$

are the densities of the transition probabilities

$$P\big(X_k \leq y \mid X_1 = x_1, \ldots, X_{k-1} = x_{k-1}\big) = P\big(X_k \leq y \mid X_{k-1} = x_{k-1}\big)\,.$$

By the Markov property, the conditional densities in (4.20) reduce to

$$\frac{f_{X_{k-1}, X_k}(x_{k-1}, x_k)}{f_{X_{k-1}}(x_{k-1})}\,, \quad k = 2, \ldots, n\,.$$

and (4.19) turns into

$$\begin{aligned} f_{\mathbf{X}_n}(x_1, \ldots, x_n) &= \frac{f_{X_{n-1}, X_n}(x_{n-1}, x_n)}{f_{X_{n-1}}(x_{n-1})} \frac{f_{X_{n-2}, X_{n-1}}(x_{n-2}, x_{n-1})}{f_{X_{n-2}}(x_{n-2})} \cdots \frac{f_{X_1, X_2}(x_1, x_2)}{f_{X_1}(x_1)} f_{X_1}(x_1) \\ &= \frac{f_{X_0, X_1}(x_{n-1}, x_n)}{f_{X_0}(x_{n-1})} \frac{f_{X_0, X_1}(x_{n-2}, x_{n-1})}{f_{X_0}(x_{n-2})} \cdots \frac{f_{X_0, X_1}(x_1, x_2)}{f_{X_0}(x_1)} f_{X_0}(x_1)\,, \end{aligned}$$

where we used the strict stationarity of (X_t) in the last step. Now assume that (Z_t) is iid $N(0, \sigma^2)$ distributed. We observe that the transition probabilities $P(Z_0 \leq y - \phi x)$ have density

$$\frac{f_{X_0, X_1}(x_0, x_1)}{f_{X_0}(x_0)} = \frac{\exp(-(x_1 - \phi x_0)^2/(2\sigma^2))}{\sigma\sqrt{2\pi}}\,.$$

Hence the density of $\mathbf{X}_n$ turns into

$$f_{\mathbf{X}_n}(x_1, \ldots, x_n) = \frac{1}{\sigma^n(\sqrt{2\pi})^n} \exp\Big(-\sum_{t=2}^{n}(x_t - \phi x_{t-1})^2/(2\sigma^2)\Big) \exp(-x_1^2/(2\sigma^2))$$

Taking logarithms, we obtain

$$-2\,\log f_{\mathbf{X}_n}(x_1, \ldots, x_n) = n\,\log(\sigma^2) + n\log(2\pi) + \frac{1}{\sigma^2}\Big(\sum_{t=2}^{n}(x_t - \phi x_{t-1})^2 + x_1^2\Big)\,.$$

Plugging in the sample $\mathbf{X}_n$ on the right-hand side and minimizing the resulting function with respect to σ^2 and $\phi \in (-1, 1)$ we arrive at the Gaussian maximum likelihood estimators of the true parameter (ϕ_0, σ_0^2) underlying the sample $\mathbf{X}_n$

$$\hat{\phi} = \rho_{n,X}(1) \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{t=1}^{n}(X_t - \hat{\phi}X_{t-1})^2 \,.$$

(after some slight corrections). We notice that $\hat{\phi}$ is the Yule-Walker estimate of ϕ on p. (4.18).

As a matter of fact, the Yule-Walker and Gaussian maximum likelihood estimators for the parameters $\phi_1, \ldots, \phi_p$ of an AR(p) process are asymptotically equivalent in the sense that these estimators are asymptotically normal with the same covariance matrix; see Brockwell and Davis [8], Chapter 7.

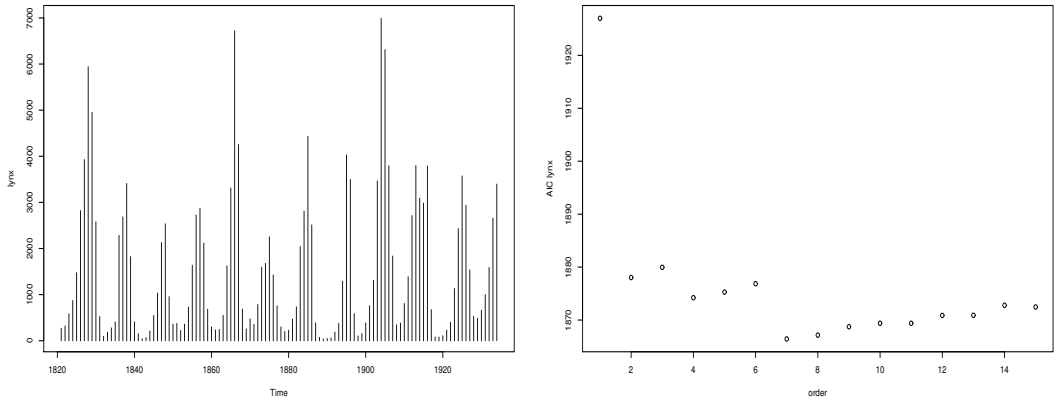

Figure 4.30. Left: *The Canadian lynx data is a famous time series, representing the annual counts of lynx trappings* 1821–1934. Right: *The AIC for an* AR(p) *model fitted to the lynx data. The order p is given on the x-axis. The minimum is achieved at the order* 7.

It was mentioned before that *information criteria* such as the AIC and BIC (see p. 42) help one to determine the order (p, q) of an ARMA process. For example, the AIC of a sample $\mathbf{X}$ from an ARMA(p, q) is determined by minimizing the quantity

$$\mathrm{AIC}(\beta) = -2\log L(\beta, \sigma^2)(\mathbf{X}) + 2\,(p + q + 1)\,,$$

where both β and σ^2 have to be replaced by estimators, depending on (p, q). This means that the likelihood function is supplemented with an additional penalty term: the larger $p + q$, the more unlikely is it that AIC(β) is minimal. It seems that the AIC is a rather arbitrary choice of penalizing the maximum likelihood procedure. This, however, is not correct; the choice of the form of the AIC and various other information criteria is based on some deep results on information theory which explain how one can make best use of the information contained in the data.

In statistical software such as R or S+ one finds the Yule-Walker and Gaussian maximum likelihood procedures for ARMA processes with the corresponding order determination via the AIC, estimation of the mean value μ, the estimation of the autocorrelation function and of the variance σ^2 of the noise.

Exercise 4.31. The Wölfer sunspot number series is a standard time series which is available in R (data(sunspots)). First transform the monthly data to annual data by taking annual averages.
i) Calculate the AIC for the sunspot numbers for an AR(p) model, $p = 1, \ldots, 20$, and plot the AIC in a graph against p.

ii) Fit an AR(p) model where p minimizes the AIC (functions arima or ar). Simulate a time series from this AR(p) model (arima.sim) with iid noise and the same sample size as the sunspot numbers. Plot the sunspot numbers and the simulated time series for comparison. Experiment with the distribution of the noise in order to get the size of the data right.
iii) Plot the sample autocorrelation functions of the sunspot numbers and the simulated time series and compare them.

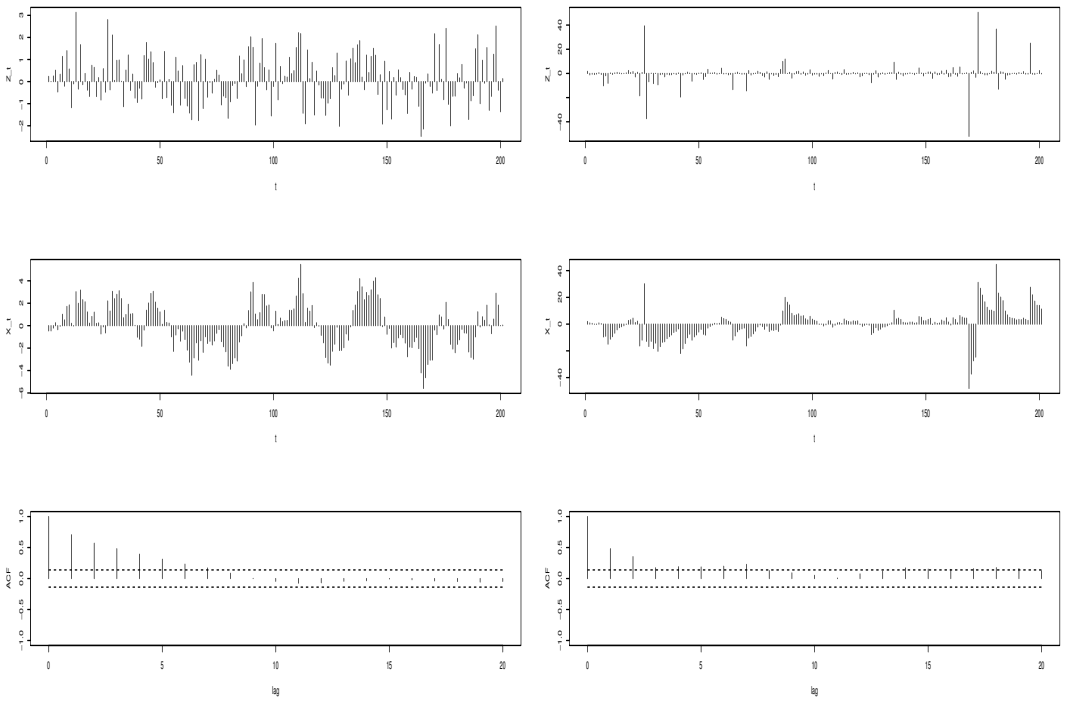

Figure 4.32. *IID standard normal noise* (top left) *and IID Cauchy noise* (top right) *and the corresponding* AR(1) *processes* $X_t = 0.8X_{t-1} + Z_t$ (middle graphs). *The bottom graphs are the corresponding sample autocorrelations functions. The latter functions are almost indistinguishable. Mind the differences in the order of magnitude of the values of the* X*- and* Z*-processes for the normal and Cauchy random variables: in the Cauchy case the extremes are much more pronounced due to* $E|Z_0| = E|X_0| = \infty$.

4.4. **Variations on ARMA models.** In this section we consider some time series models which are derived from ARMA processes in a suitable way.

Infinite variance ARMA processes. In the previous sections we considered ARMA processes driven by white noise which, by definition, consists of random variables with a finite variance. However, the condition of a finite variance is not necessary to define an ARMA(p, q) process. Under the standard assumptions that the polynomials $\phi(\cdot)$ and $\theta(\cdot)$ do not have common zeros and that $\phi(z)\theta(z) \neq 0$ for $|z| \leq 1$, the ARMA equations $\phi(B)X_t = \theta(B)Z_t$, $t \in \mathbb{Z}$, have a unique strictly stationary solution (X_t) if the noise (Z_t) consists of iid centered (i.e., $EZ_0 = 0$ if $E|Z_0| < \infty$ or Z_0 symmetric) random variables and $E|Z_0|^p < \infty$ for some $p > 0$. The solution has again a representation as a linear process $X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}$, $t \in \mathbb{Z}$, with the same coefficients (ψ_j) as in the finite variance case. For example, an AR(1) process with parameter $|\phi| < 1$ has representation $X_t = \sum_{j=0}^{\infty} \phi^j Z_{t-j}$.

Although the autocovariances and the autocorrelations of an infinite variance process do not make sense (they are defined via second moments!) the corresponding sample versions $\gamma_{n,X}(h)$ and

$\rho_{n,X}(h) = \gamma_{n,X}(h)/\gamma_{n,X}(0)$ are clearly defined. Moreover, under quite general conditions one can show that

$$\rho_{n,X}(h) \xrightarrow{P} \frac{\sum_{j=0}^{\infty} \psi_j \psi_{j+|h|}}{\sum_{j=0}^{\infty} \psi_j^2} , \quad |h| \geq 1 , \tag{4.21}$$

and one can derive the asymptotic distribution under further restrictions on the distribution of Z_0. These results are somewhat surprising since, in the finite variance case, the right-hand side of (4.21) is nothing but $\rho_X(h)$, which, in the infinite variance case, does not exist, but can still be interpreted as some kind of a "population correlation". Similar astonishing results can be shown for the classical parameter estimation procedures under an infinite variance condition. For example, the Yule-Walker and the Gaussian maximum likelihood estimators estimate the underlying parameters consistently.

One might ask how these results can be interpreted. The "practitioner" who rarely believes in an infinite variance model will take these results as "robustness of the classical estimators under outliers in the innovations". From a purely probabilistic point of view this is wrong since the notion "outlier" suggests that large values of Z_t are an accident, but they actually belong to the distribution of Z_t.

Infinite variance models have attracted the attention for a long time. Mandelbrot suggested in a series of papers around 1960 to model financial time series (stock returns, exchange rates, etc.) by infinite variance (in particular stable) processes; see Samorodnitsky and Taqqu [32] for an enjoyable reading about stable processes. The use of infinite variance models in finance has a history of controversial discussions; see Taylor [35] or Mikosch [25]. The existence of real-life time series with infinite variance is confirmed in reinsurance; see Embrechts et al. [13], Chapter 6, as well as in telecommunications; see Willinger et al. [37]. In this area it is believed that the ON/OFF processes of sources (computers) in Local Area Networks and the Internet can be well modeled by infinite variance time series. However, to describe the dependence structure of such data by an ARMA process is wishful thinking; see Figure 4.34.

Introductions to infinite variance ARMA processes can be found in Brockwell and Davis [8], Section 13.3, and Embrechts et al. [13], Chapter 7.

Exercise 4.33. i) Simulate $n = 200$ values of the AR(1) process $X_t = -0.8X_{t-1} + Z_t$ a) for iid standard Gaussian Z_t (this is standard in arima.sim) and b) for iid standard Cauchy Z_t (rcauchy). Make scatter plots for both time series, i.e., plot (X_t, X_{t+1}).
ii) Repeat the simulation of the series in a) and b) 500 times and calculate the Yule-Walker estimate (function ar) of $\phi = -0.8$ for each of the series. Calculate the mean and the standard deviation of these distinct series of 500 estimates of ϕ. Make a boxplot (function boxplot) comparison. Give an intuitive interpretation of the results.

ARIMA processes. ARIMA processes are models for non-stationary processes with a polynomial trend.

Definition 4.35. (ARIMA process)
(X_t) *is an* integrated ARMA process of order (p, d, q) (ARIMA(p, d, q) *process*) *for integers* $p, d, q \geq 0$ *if the* d *times differenced process* (X_t) *is a causal* ARMA(p, q) *process, i.e.,* $Y_t = (1-B)^d X_t, t \in \mathbb{Z}$, *is a causal* ARMA$(p, q)$ *process.*

The so-defined process is non-stationary for $d \geq 1$. This can be seen by an evaluation of $\mathrm{var}(X_t)$ which depends on t.

Example 4.36. (ARIMA$(1, 1, 0)$ process)
An ARIMA$(1, 1, 0)$ process is given by the equations

$$(1 - B)(1 - \phi B)X_t = Z_t , \quad t \in \mathbb{Z} ,$$

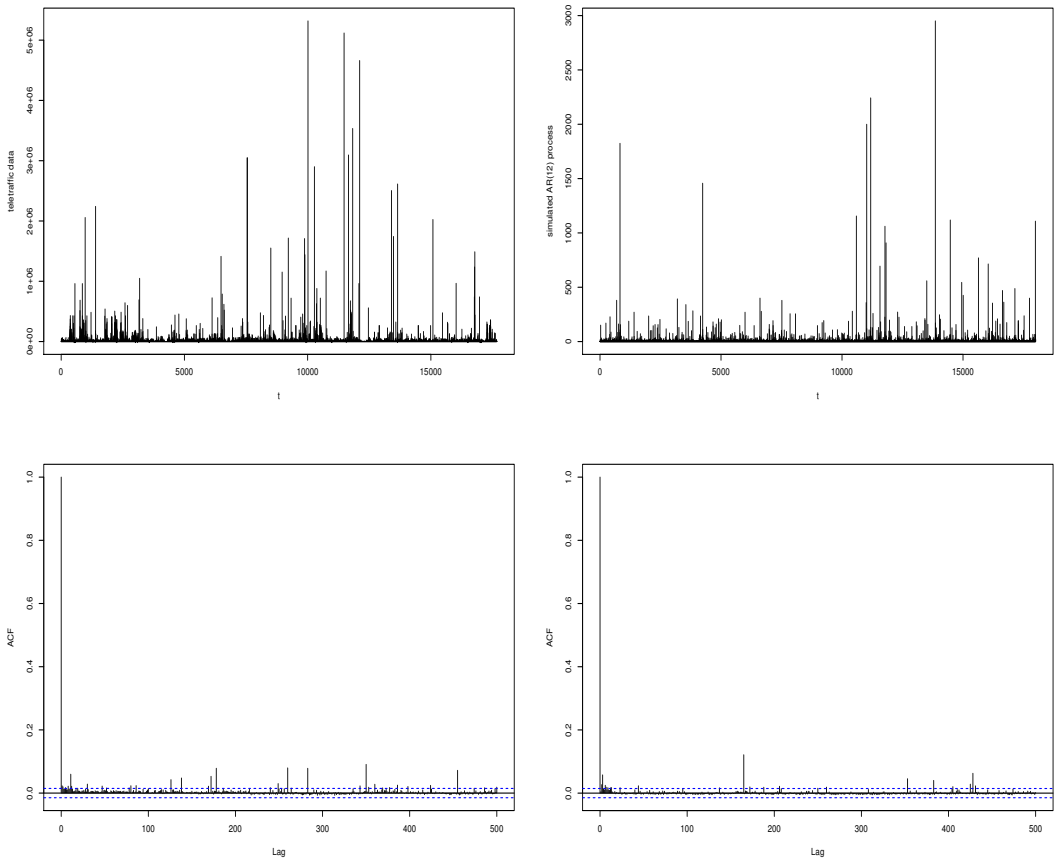

Figure 4.34. Left column: *A time series of* 18000 *lengths of ON-periods of a local area computer network and its sample autocorrelation function. The data are believed to come from an infinite variance model, explaining the extremes in this time series. The sample autocorrelations of this model are not easily modeled by an ARMA process; although most autocorrelations are negligible, the autocorrelations at some lags are different from zero and seem to indicate the existence of some intricate dependence structure.* Right column: *Simulation of teletraffic data based on the fit of the real-life data to an* AR(12) *model with infinite variance iid noise and the corresponding sample autocorrelation function. The order* 12 *was chosen by the AIC. All* 12 *parameters of the model are significantly smaller than* 0.1. *They cannot explain the occurrence of non-negligible autocorrelations at very high lags in the real-life time series. However, since the data are believed to have infinite variance, the interpretation of the sample autocorrelations is quite doubtful.*

where (Z_t) is a white noise process and $|\phi| < 1$. This means that

$$X_t = X_0 + \sum_{j=1}^{t} Y_j \ , \quad t \geq 1 \ ,$$

where (Y_t) is an AR(1) process:

$$Y_t = (1-B)X_t = X_t - X_{t-1} = \sum_{j=0}^{\infty} \phi^j Z_{t-j} \ , \quad t \in \mathbb{Z} \ .$$

In Section 2 we learnt that differencing is one of the standard methods in order to get rid of a polynomial trend in the time series. Thus one can add a polynomial trend to (X_t); after d times differencing this trend disappears.

Notice that the case $d = 0$ corresponds to the usual ARMA case. A general ARIMA(p, d, q) process can be described by the equations

$$\phi^*(B)X_t = (1-B)^d\phi(B)X_t = \theta(B)Z_t\,, \quad t \in \mathbb{Z}\,,$$

where $\phi(\cdot)$ and $\theta(\cdot)$ are the polynomials corresponding to a causal ARMA(p, q) process. The polynomial $\phi^*(z) = (1-z)^d\phi(z)$ has a zero at $z = 1$. This is in contrast to a causal stationary ARMA process where we know that $\phi(z) \neq 0$ for $|z| \leq 1$.

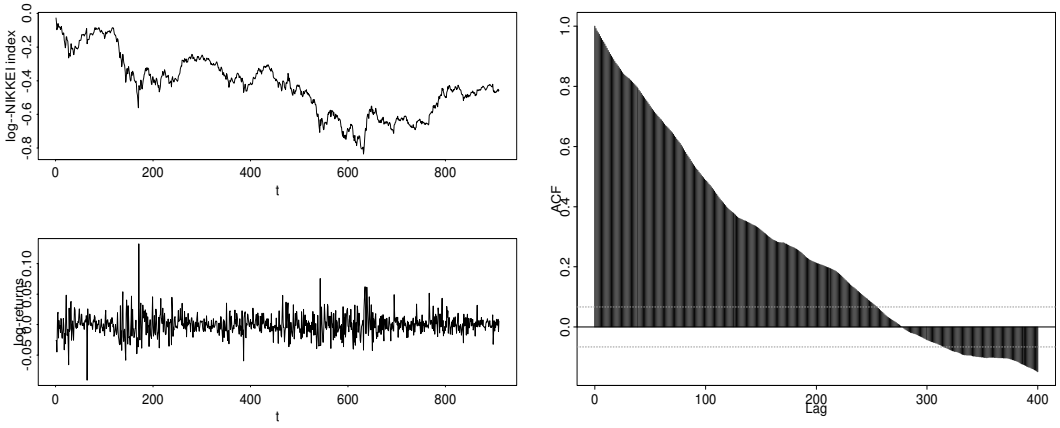

Figure 4.37. *The logarithm of the Japanese composite stock index NIKKEI gives the impression of a non-stationary time series* (top left)*. After differencing the time series "looks stationary"* (bottom left)*. These are the log-returns.* Right: *The sample autocorrelations of the logarithmic NIKKEI index. They decay very slowly indicating that the data come from a non-stationary model. An alternative approach would be to assume stationarity and long memory.*

Fractional ARIMA processes. We know that a causal ARMA process has a linear process representation $X_t = \sum_{j=0}^{\infty}\psi_j Z_{t-j}$, where (Z_t) is white noise; see p. 38. The coefficients ψ_j decrease to zero as $j \to \infty$ at an exponential rate; see (4.11). This implies in particular that $|\psi_j| < Ka^j$ for a positive constant K and an $0 < a < 1$, for all j, but also that $|\rho_X(h)| \leq Ka^j$; see (4.14). Hence

$$\sum_{h=0}^{\infty}|\rho_X(h)| < \infty\,. \tag{4.22}$$

This indicates that the dependence in an ARMA process is "weak"; if $|h|$ is large, the correlation between X_t and X_{t+h} dies out rather fast. Processes satisfying (4.22) are called *short memory processes* or *processes with short-range dependence.*

In contrast to these processes, there exist real-life time series whose sample autocorrelations can be shown to decay very slowly. If the corresponding *stationary* model does not satisfy (4.22), i.e., if $|(\rho_X(n))|$ is not summable, the process is said to have *long memory* or *long-range dependence*, i.e., there is very strong dependence in the data indeed. Examples of such time series have been observed in hydrology, economics and teletraffic; see for example the absolute values of the S&P500 log-returns in Figure 3.12 or the teletraffic data in Figure 4.34. Hurst [20] considered the time series of the annual flow of the river Nile at Ashwan over some hundred years of observations; see

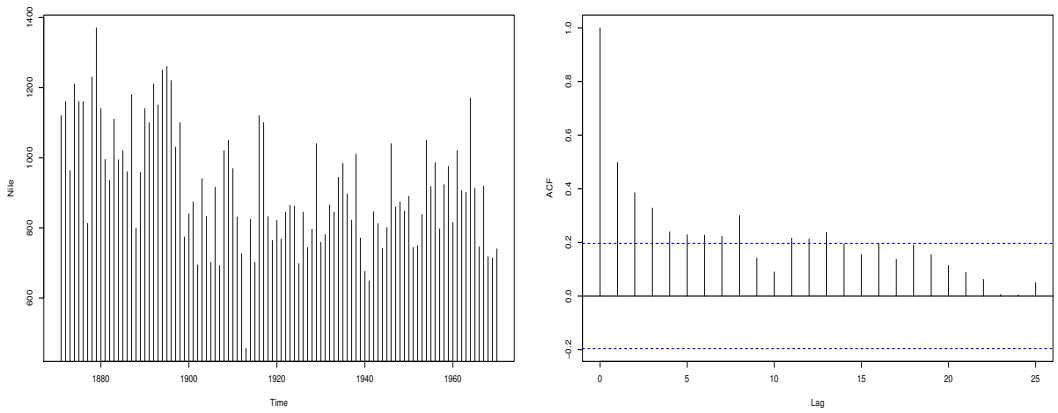

Figure 4.38. *The river Nile data* $1871 - 1970$ *present annual flows of the river Nile at Ashwan* (left). *The corresponding sample autocorrelations* (right).

Figure 4.38 for a piece of the data. He got empirical evidence that these water levels exhibit some kind of long-range dependence. Therefore one sometimes refers to long memory as *Hurst effect*.

The monographs Brockwell and Davis [8], Section 13.3, Samorodnitsky and Taqqu [32] and the book Doukhan et al. [12] are relevant references on long memory. As mentioned after p. 32, the phenomenon of long memory is a question of belief. It is possible to model the same kind of sample autocorrelation behavior — slow decay at large lags — by a non-stationary time series which is subject to changes of its structure, but also by a stationary time series with a slowly decaying autocorrelation function.

One of the standard models in this context was introduced in 1980 by Granger and Joyeux [19]; Granger received one of the Nobel Prizes for Economics in 2003.

Definition 4.39. (FARIMA process)
The process (X_t) *is a* fractional ARIMA/fractionally integrated ARMA process of order $(0, d, 0)$ (FARIMA$(0, d, 0)$ process) *for some* $d \in (-0.5, 0.5)$ *or* fractional noise process *if it is the stationary solution of the FARIMA equations*

$$(1 - B)^d X_t = Z_t \,, \quad t \in \mathbb{Z} \,, \tag{4.23}$$

where (Z_t) *is a white noise process.*

The defining equation (4.23) can be shown to admit a unique stationary solution. Relation (4.23) has to be interpreted as follows. Notice that

$$(1 - z)^d = \sum_{j=0}^{\infty} \pi_j z^j \,, \quad |z| < 1 \,,$$

is a binomial (Taylor) expansion with coefficients

$$\pi_j = \frac{\Gamma(j - d)}{\Gamma(j + 1)\Gamma(-d)} = \prod_{0<k\le j} \frac{k - 1 - d}{k} \,.$$

Plugging in the backshift operator we thus obtain

$$(1 - B)^d X_t = \sum_{j=0}^{\infty} \pi_j X_{t-j} \,.$$

Thus we get something like an autoregressive process of order ∞. Similarly to the ARMA processes of finite order, the equations (4.23) can be solved:

Theorem 4.40. (Linear process representation of FARIMA processes)
For $d \in (-0.5, 0.5)$ there exists a unique stationary solution of the equations (4.23)*. It can be written as*

$$X_t = (1-B)^{-d} Z_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j} \ ,$$

where (ψ_j) are the coefficients of the power series expansion of $(1-z)^{-d}$ around zero. In particular,

$$\psi_j = \frac{\Gamma(j+d)}{\Gamma(j+1)\Gamma(d)} = \prod_{0<k\le j} \frac{k-1+d}{k} \ .$$

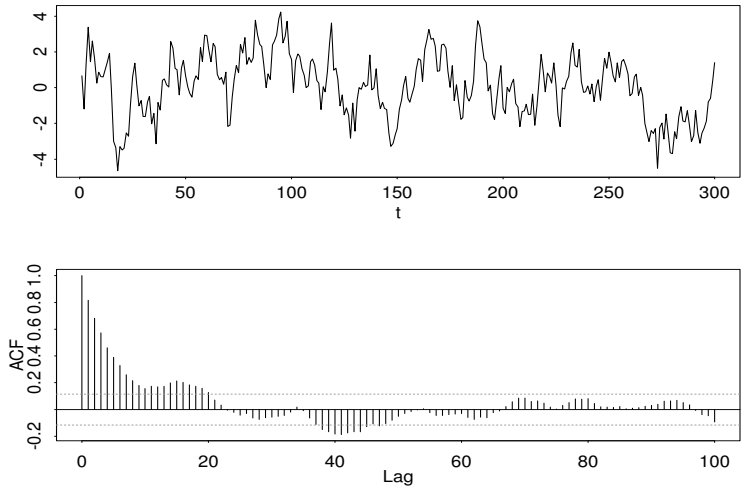

Figure 4.41. *Simulation of a* FARIMA$(0.8, 0.4, 0)$ *process driven by Gaussian white noise* (top) *and the corresponding sample autocorrelations* (bottom).

Using the properties of the gamma function (Stirling's formula), one can show that

$$\rho_X(h) \sim h^{2d-1} \frac{\Gamma(1-d)}{\Gamma(d)} \ , \quad h \to \infty \ ,$$

implying that these processes have long memory if $d \in (0, 0.5)$.

In a similar fashion one can define FARIMA(p, d, q) processes via the equations

$$\phi(B)(1-B)^d X_t = \theta(B) Z_t \ , \quad t \in \mathbb{Z} \, , \quad d \in (-0.5, 0.5) \ .$$

For $d \in (0, 0.5)$ these processes exhibit long-range dependence.

The parameters of FARIMA processes can be estimated in a similar fashion as for ARMA processes. The proofs of consistency and the derivation of the limit distributions are very technical. Long memory processes have non-standard limit behavior. The central limit theorem for the sample mean with normalization $\sqrt{n}$ fails in this case whereas it holds (under general assumptions) for ARMA processes. A central limit theorem can be proved with an appropriate normalization depending on the parameter d. A consequence would be that, under long-range dependence, we could not trust any standard statistical procedure (as implemented in all statistical software packages) any more.

5. ARCH and GARCH processes

In the econometrics literature, the ARCH processes (*autoregressive processes with conditional heteroscedasticity*) and their numerous modifications have attracted significant attention. One of the 2003 Bank of Sweden Prizes for Economics, better known under the name of Nobel Prize for Economics, was awarded to Robert Engle who introduced the ARCH model in the celebrated 1982 paper [14]. We also refer to the collection of papers on the theme "ARCH" edited by Engle [15].

ARCH processes were introduced to describe typical features of log-return data $X_t = \log S_t - \log S_{t-1}$ of share prices, foreign exchange rates, composite stock indices, etc., denoted by S_t. Among them are the following "stylized facts":

- Zero sample autocorrelations for (X_t) at almost all lags, with a possible exception at the first lag although $\rho_{n,X}(1)$ is always rather small (often about 0.1 in absolute value).
- Very slowly decaying sample autocorrelations of $(|X_t|)$ and (X_t^2). In this context, one often refers to *long memory in the volatility*.
- Occurrence of extremely large and small X_t's clustered at certain instants of time, caused by turbulences in the market due to financial crashes, political decisions, war, etc.

If we wanted to explain the dependence structure of such a model by an ARMA model with iid noise (Z_t), we would have to restrict our attention to models of the form $X_t = Z_t$ or moving average models of very low order. Indeed, for an MA(q) model the autocorrelations vanish at lag $q+1$ and therefore only MA-models would fit the autocorrelation structure of the data. On the other hand, for an MA(q) model with iid noise, X_t and X_{t+q+1} are independent, hence $|X_t|^r$ and $|X_{t+q+1}|^r$ are independent for any $r > 0$ and therefore $\rho_{|X|^r}(h) = 0$ for $|h| > q$. This means that the effect of non-vanishing autocorrelations of the $(|X_t|^r)$ processes for $r = 1, 2$ cannot be explained by an MA(q) model with iid noise (Z_t).

5.1. **The ARCH(1) model.** We start by considering the simplest element of the ARCH family, an ARCH(1) process: let (Z_t) be an iid noise sequence with $EZ_0 = 0$ and $\mathrm{var}(Z_0) = 1$. Then define

$$X_t = \sigma_t Z_t\,, \quad \sigma_t^2 = \alpha_0 + \alpha_1 X_{t-1}^2\,, \quad t \in \mathbb{Z}\,,$$

for some positive α_0, α_1. In contrast to the ARMA or linear processes, (Z_t) is referred to as multiplicative noise. The σ_t's are called *volatility*. This is another name for standard deviation.

Notice that, $EX_t = 0$ and, by independence of Z_{t+h} and $\sigma_{t+h}X_t$ (later we will show that σ_k is a function of $(Z_s)_{s\le k-1}$)

$$\gamma_X(h) = E(X_t X_{t+h}) = E(X_t\,\sigma_{t+h})\,EZ_{t+h} = 0\,, \quad h \ge 1\,,$$

hence $\rho_X(h) = 0$ for $h \neq 0$. This property captures the empirical fact that the sample autocorrelations of log-returns are negligible at almost all lags.

Notice that σ_t^2 is the conditional variance of X_t given the past $X_{t-1}, X_{t-2}, \ldots$:

$$E(X_t^2 \mid X_{t-1}, X_{t-2}, \ldots) = E(X_t^2 \mid X_{t-1}) = \sigma_t^2\,\mathrm{var}(Z_0) = \alpha_0 + \alpha_1 X_{t-1}^2\,.$$

Thus a large value of X_{t-1}^2 (yesterday's squared return) will substantially contribute to the conditional variance of X_t (today's return). It is common to assume strict stationarity for (X_t) and (σ_t^2). Then $\mathrm{var}(X_t)$ is not dependent on time, but the conditional variance of X_t is time-dependent and "gets adjusted to the recent history" of the time series. This is the reason why this kind of model is referred to as "conditionally heteroscedastic", i.e., its conditional variance changes over time.

Writing

$$B_t = \alpha_0\,, \quad A_t = \alpha_1 Z_{t-1}^2 \quad \text{and} \quad Y_t = \sigma_t^2\,,$$

we have

$$Y_t = A_t Y_{t-1} + B_t\,, \quad t \in \mathbb{Z}\,. \tag{5.1}$$

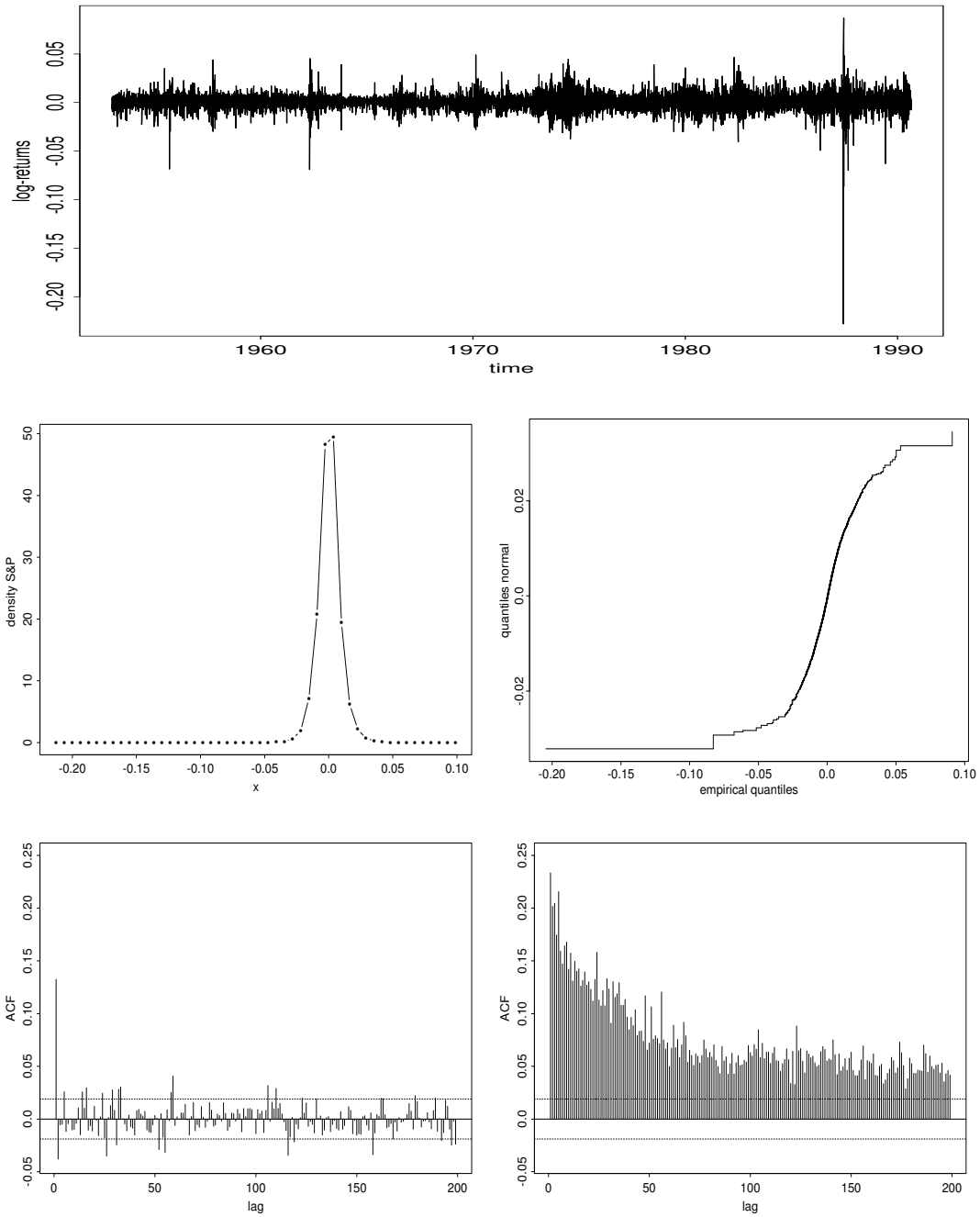

Figure 5.1. Top: *Plot of* 9558 S&P500 *daily log-returns from January* 2, 1953, *to December* 31, 1990. *The year marks indicate the beginning of the calendar year. The* S&P500 *is one of the major US composite stock indices. This time series is one of the warhorses of the financial time series community.* Middle, left: *Density plot of the* S&P500 *data. The limits on the* x*-axis indicate the range of the data.* Right: *QQ-plot of the* S&P500 *data against the normal distribution whose mean and variance are estimated from the* S&P500 *data. These graphs give a clear indication that the data are non-Gaussian and heavy-tailed.* Bottom: *Sample ACFs for the log-returns* (left) *and absolute log-returns* (right) *of the* S&P500.

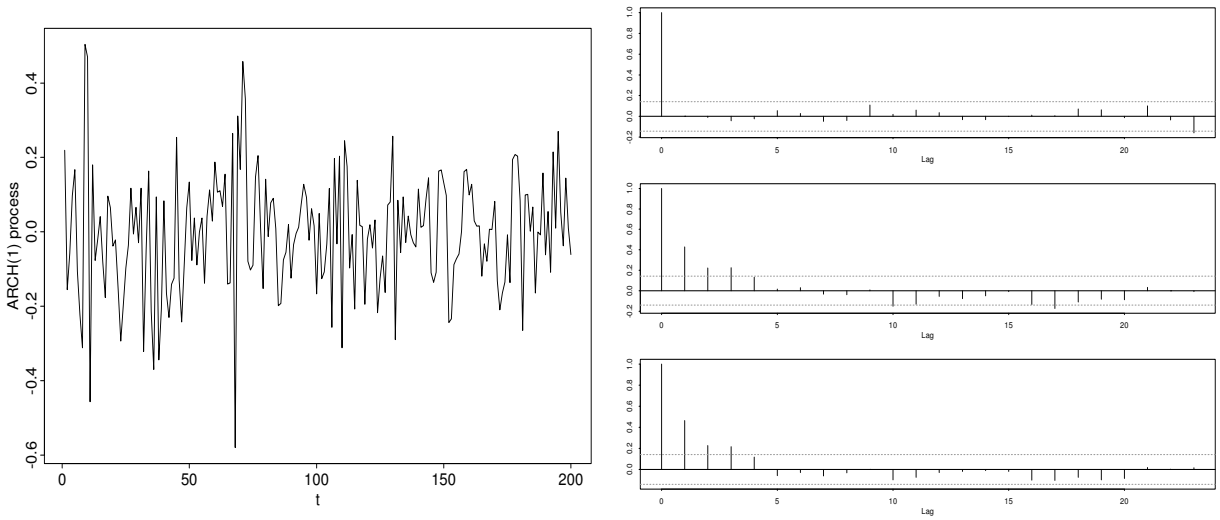

Figure 5.2. Left: *A realization of the* ARCH(1) *process* $X_t = (0.01 + 0.5X_{t-1}^2)^{1/2} Z_t$ *with iid standard Gaussian noise* (Z_t). Right: *The corresponding sample autocorrelation functions for* (X_t) (top), $(|X_t|)$ (middle) *and* (X_t^2). *The sample autocorrelations of* $\rho_{n,|X|}(h)$ *and* $\rho_{n,X^2}(h)$ *decay very quickly to zero. This is in agreement with the theory, i.e., the empirically observed "long memory" of* $(|X_t|)$ *and* (X_t^2) *cannot be captured by any ARCH or GARCH model.*

We refer to this equation as a *stochastic recurrence equation*. The sequence of pairs (A_t, B_t) is iid and (A_t, B_t) and Y_{t-1} are independent. Equation (5.1) can be interpreted as a *random coefficient autoregressive model*. The representation (5.1) helps one to find conditions for strict stationarity of (Y_t). Indeed, if $A_t = \varphi$ were a constant, (5.1) would describe an AR(1) process with parameter φ. For an AR(1) process with iid noise (B_t) we know that a unique stationary causal solution exists if and only if $|\varphi| < 1$; see Example 4.4.

We proceed as in the case of a causal AR(1) model. Iterating (5.1) r times , we obtain

$$Y_t = A_t \cdots A_{t-r} Y_{t-r-1} + \alpha_0 \sum_{i=t-r}^{t} A_t \cdots A_{i+1} \,.$$

Now, letting r go to infinity, we hope that the first term on the right-hand side will disappear and the second one will converge. Notice that

$$\sum_{i=-\infty}^{t} A_t \cdots A_{i+1} = 1 + \sum_{i=-\infty}^{t-1} \exp\left\{ (t-i) \left[\frac{1}{t-i} \left(\sum_{j=i+1}^{t} \log A_j \right) \right] \right\} \,. \tag{5.2}$$

For fixed t, the strong law of large numbers tells us that as $i \to -\infty$,

$$\frac{1}{t-i} \sum_{j=i+1}^{t-1} \log A_j \stackrel{\text{a.s.}}{\to} E \log A_1 \,,$$

provided that $E \log A_1$ is defined, finite or infinite. Hence, under the moment condition $-\infty \le E \log A_1 < 0$, the infinite series (5.2) converges a.s. for every fixed t. Then the sequence

$$\widetilde{Y}_t = \alpha_0 \sum_{i=-\infty}^{t} A_t \cdots A_{i+1} \,, \quad t \in \mathbb{Z} \,,$$

constitutes a strictly stationary solution to equation (5.1). If there is another strictly stationary solution $(\widehat{Y}_t)$ we have by iterating (5.1),

$$|\widetilde{Y}_t - \widehat{Y}_t| \;=\; A_t \cdots A_{t-r}\, |\widetilde{Y}_{t-r-1} - \widehat{Y}_{t-r-1}|\,, \tag{5.3}$$

and since $A_t \cdots A_{t-r}$ and $|\widetilde{Y}_{t-r-1} - \widehat{Y}_{t-r-1}|$ are independent, the weak law of large numbers and $E \log A_1 < 0$ imply that the right-hand side in (5.3) converges to zero in probability as $r \to \infty$. Therefore $\widetilde{Y}_t = \widehat{Y}_t$ for every t with probability 1.

More sophisticated arguments show that $E \log A_1 < 0$ is also necessary for the existence and uniqueness of a non-trivial strictly stationary solution of the stochastic recurrence equation $Y_t = A_t Y_{t-1} + B_t$, $t \in \mathbb{Z}$. Hence we proved that the squared volatility process (σ_t^2) has representation

$$\sigma_t^2 = \alpha_0 \sum_{j=-\infty}^{t} \prod_{k=j+1}^{t} \alpha_1 Z_{k-1}^2 = f(Z_{t-1}, Z_{t-2}, \ldots)\,, \quad t \in \mathbb{Z}\,,$$

Recalling the theory from Section 2, we immediately see that the right-hand side is a function acting on the shifts of an iid sequence. Hence (σ_t) is a strictly stationary ergodic and mixing time series. Moreover,

$$X_t = \sigma_t Z_t = \sqrt{f(Z_{t-1}, Z_{t-2}, \ldots)}\, Z_t = g(Z_t, Z_{t-1}, \ldots)$$

has the same properties.

Theorem 5.3. (Nelson [27], Bougerol and Picard [6]) *There exists an a.s. unique non-vanishing strictly stationary ergodic causal*[5] *solution of the* ARCH(1) *stochastic recurrence equation* (5.1) *if and only if* $\alpha_0 > 0$, $E \log(\alpha_1 Z_0^2) < 0$.

The assumption $\alpha_0 > 0$ is crucial; otherwise $X_t \equiv 0$ a.s. would be the solution to (5.1).

5.2. **The ARCH family, definition and relation with ARMA processes.** As a generalization of the ARCH(1) model, Engle [14] suggested the following simple model for the volatility σ_t:

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i\, X_{t-i}^2\,, \quad t \in \mathbb{Z}\,. \tag{5.4}$$

Here α_i are non-negative constants with $\alpha_p \alpha_0 > 0$ for $p \geq 1$. The model

$$X_t = \sigma_t\, Z_t\,, \quad (Z_t) \text{ iid}, \; EZ_0 = 0, \; \mathrm{var}(Z_0) = 1\,, \tag{5.5}$$

with the specification (5.4) for σ_t^2 is called an ARCH(p) process (*autoregressive conditionally heteroscedastic model of order* p).

The autoregressive structure can be seen by the following argument. Writing

$$\nu_t = X_t^2 - \sigma_t^2 = \sigma_t^2\,(Z_t^2 - 1)\,,$$

with the help of (5.4) one obtains

$$\varphi(B) X_t^2 = \alpha_0 + \nu_t\,, \quad t \in \mathbb{Z}\,, \tag{5.6}$$

where

$$\varphi(z) = 1 - \sum_{i=1}^{p} \alpha_i\, z^i\,,$$

and $BC_t = C_{t-1}$ is the backshift operator. If (Z_t) is an iid sequence with unit variance, finite 4th moment, and (X_t) is stationary with finite 4th moment, then (ν_t) constitutes a white noise sequence. Therefore (X_t^2) is an AR(p) process with noise sequence (ν_t). However, (ν_t) is not an iid sequence.

[5] This means it depends only on past and present values of the Z's.

Exercise 5.4. Verify that strict stationarity of (σ_t^2) with $E\sigma_0^4 < \infty$ and $EZ_0^4 < \infty$ imply that (ν_t) constitutes white noise.

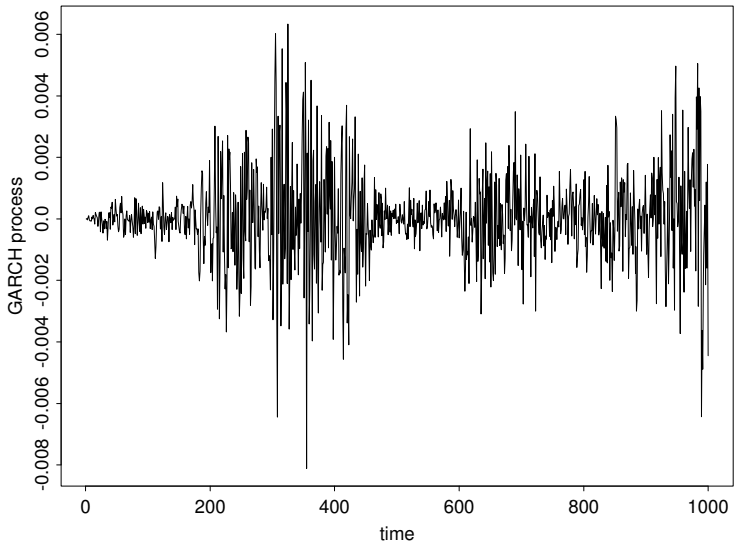

Figure 5.5. *A simulated path of the* GARCH(1,1) *time series* $X_t = (0.0001 + 0.1X_{t-1}^2 + 0.9\sigma_{t-1}^2)^{0.5}Z_t$, $t = 1, \ldots, 1000$, *for iid standard normal* (Z_t).

Since ARCH(p) processes do not fit log-returns very well unless one chooses the order p quite large (which is not desirable when the sample is small), various people have thought about improvements. Because (5.6) bears some resemblance with an autoregressive structure, it is natural to impose an ARMA structure on the squared returns:

$$\varphi(B)\, X_t^2 = \alpha_0 + \beta(B)\, \nu_t\,, \quad t \in \mathbb{Z}\,, \tag{5.7}$$

where $\varphi(B)$ and $\beta(B)$ are polynomials in the backshift operator B with coefficients φ_i, β_j. More precisely, let α_i, $i = 0, \ldots, p$, and β_j, $j = 1, \ldots, q$, be non-negative coefficients with $\alpha_p > 0$ if $p \geq 1$ and $\beta_q > 0$ if $q \geq 1$, then

$$\varphi(z) = 1 - \sum_{i=1}^{p} \alpha_i\, z^i - \sum_{j=1}^{q} \beta_j\, z^j \quad \text{and} \quad \beta(z) = 1 - \sum_{j=1}^{q} \beta_j\, z^j\,.$$

This construction leads to the GARCH(p, q) *process* (*generalized ARCH process of order* (p, q)) which was independently introduced by Bollerslev [5] and Taylor [35]. The latter process, with its ramifications and modifications, has become the model for returns which is used most frequently in applications. It is more conveniently written as the multiplicative model (5.5) with specification:

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i\, X_{t-i}^2 + \sum_{j=1}^{q} \beta_j\, \sigma_{t-j}^2\,, \quad t \in \mathbb{Z}\,. \tag{5.8}$$

Here α_i and β_j are non-negative constants. To understand the motivation behind the ARCH processes it pays to read some of the original articles of which some were mentioned; see Engle [15] for a good collection.

Exercise 5.6. (Integrated GARCH)
For real-life log-returns one often observes that the estimated GARCH parameters sum up to a value close to 1; see Figure 5.7:

$$\sum_{j=1}^{p} \widehat{\alpha}_j + \sum_{k=1}^{q} \widehat{\beta}_k \approx 1\,.$$

This observation led Engle and Bollerslev [16] to the introduction of the *integrated* GARCH(p,q) *process* (IGARCH(p,q)) by requiring

$$\sum_{j=1}^{p} \alpha_j + \sum_{k=1}^{q} \beta_k = 1\,.$$

A strictly stationary version of an IGARCH process has the undesirable and empirically not observed property that both σ_t and X_t have infinite variance. Verify this property by assuming that (X_t) and (σ_t) are both strictly stationary. Also show that σ_t and X_t have infinite variance if

$$\sum_{j=1}^{p} \alpha_j + \sum_{k=1}^{q} \beta_k > 1\,.$$

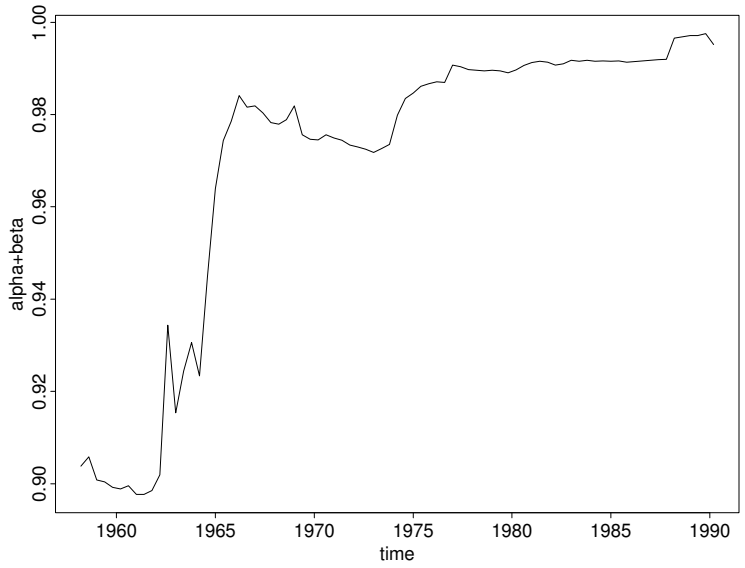

Figure 5.7. *The estimated values of $\alpha_1 + \beta_1$, using quasi-MLE, see Section 5.5, for an increasing sample of the* S&P500 *log-returns from Figure 5.1. An initial* GARCH(1,1) *model is fitted to the first* 1500 *observations* (6 *business years*). *Then $k * 100$, $k = 1, 2, \ldots$, data points are successively added to the sample and α_1 and β_1 are re-estimated on these samples. The labels on the time axis indicate the date of the latest observation used for the estimation procedure.*

5.3. **The** GARCH(1,1) **process.** The GARCH(1,1) process is most frequently used in applications to return series. Main reasons are that

(1) this simple model with three parameters $\alpha_0, \alpha_1, \beta_1$ and iid standard normal or iid student distributed (standardized to unit variance) innovations (Z_t) already gives a reasonable fit to real-life returns,

(2) in contrast to higher-order GARCH models, one can calculate certain distributional characteristics (moments, conditions for stationarity, tails,...) (almost) explicitly.

5.3.1. *Conditions for strict and weak stationarity.* The squared volatility σ_t^2 of a GARCH(1,1) process $X_t = \sigma_t Z_t$ satisfies the one-dimensional stochastic recurrence equation

$$\sigma_t^2 = \alpha_0 + \alpha_1 X_{t-1}^2 + \beta_1 \sigma_{t-1}^2 = \alpha_0 + (\alpha_1 Z_{t-1}^2 + \beta_1)\, \sigma_{t-1}^2 \,. \tag{5.9}$$

Writing $Y_t = \sigma_t^2$, $A_t = \alpha_1 Z_{t-1}^2 + \beta_1$, $B_t = \alpha_0$, it is not difficult to see that the equation $Y_t = A_t Y_{t-1} + B_t$, $t \in \mathbb{Z}$, in (5.9) has the solution

$$Y_t = \alpha_0 \sum_{i=-\infty}^{t} A_t \cdots A_{i+1} = \alpha_0 \sum_{i=-\infty}^{t} (\alpha_1 Z_{t-1}^2 + \beta_1) \cdots (\alpha_1 Z_i^2 + \beta_1) \,, \qquad t \in \mathbb{Z} \,, \tag{5.10}$$

provided $-\infty \le E \log A_1 = E \log(\alpha_1 Z_0^2 + \beta_1) < 0$. Indeed, one can follow the lines of the proof in the ARCH(1) case.

As for the ARCH(1) case we may conclude that a GARCH(1,1) process has the following structure:

$$X_t = g(Z_t, Z_{t-1}, \ldots) \,, \qquad t \in \mathbb{Z} \,,$$

for some function g acting on the shifts of the iid sequence (Z_t). Following the results in Section 2, we conclude that (X_t) is strictly stationary ergodic and mixing. It can also be shown to be strongly mixing with a mixing rate (α_h) which decays to zero exponentially fast if Z_0 has a Lebesgue density in some interval; see Doukhan [11].

Theorem 5.8. (Nelson [27], Bougerol and Picard [6]) *There exists an a.s. unique non-vanishing strictly stationary ergodic causal (i.e., depending only on past and present values of the Z's) solution of the equations defining a* GARCH(1,1) *process if and only if $\alpha_0 > 0$ and $E \log(\alpha_1 Z_1^2 + \beta_1) < 0$.*

In particular, the condition $E \log(\alpha_1 Z_1^2 + \beta_1) < 0$ is satisfied for $\alpha_1 + \beta_1 < 1$. This follows by an application of Jensen's inequality:

$$E \log(\alpha_1 Z_1^2 + \beta_1) \le \log(E(\alpha_1 Z_1^2 + \beta_1)) = \log(\alpha_1 + \beta_1) < 0 \,.$$

It follows from Exercise 5.6 hat the GARCH(1,1) process (X_t) has infinite variance if $\alpha_1 + \beta_1 \ge 1$. The case $\alpha_1 + \beta_1 < 1$ covers the finite variance case; see the arguments below. This case is sufficient for many practical purposes. Thus a GARCH(1,1) process is stationary if and only if $\alpha_1 + \beta_1 < 1$.

In the ARMA case the conditions for stationarity do not depend on the distribution of the innovations. This is different in the GARCH(1,1) case: the relation $E \log(\alpha_1 Z_1^2 + \beta_1) < 0$ (which is necessary for strict stationarity) involves the distribution of the noise Z_1; see Figure 5.9 for an illustration of the region where (X_t) is strictly stationary. Therefore one obtains different parameter regions for strict stationarity of a GARCH(1,1) process, depending on the noise distribution.

5.3.2. *Moments and tails.* The even integer moments of a GARCH(1,1) process can be calculated by exploiting the stochastic recurrence equation $Y_t = A_t Y_{t-1} + B_t$ with $Y_t = \sigma_t^2$, $B_t = \alpha_0$, $A_t = \alpha_1 Z_{t-1}^2 + \beta_1$. Indeed, we have

$$E[X_0^{2k}] = E[\sigma_0^{2k}]\, E[Z_0^{2k}] \,, \quad k = 1, 2, \ldots \,,$$

and

$$\begin{aligned} E[Y_0^k] &= E\Big[(A_1 Y_0 + \alpha_0)^k\Big] \\ &= \sum_{i=0}^{k} \binom{k}{i} E[(A_1 Y_0)^i] \alpha_0^{k-i} = \sum_{i=0}^{k} \binom{k}{i} E[A_1^i]\, E[Y_0^i] \alpha_0^{k-i} \,. \end{aligned}$$

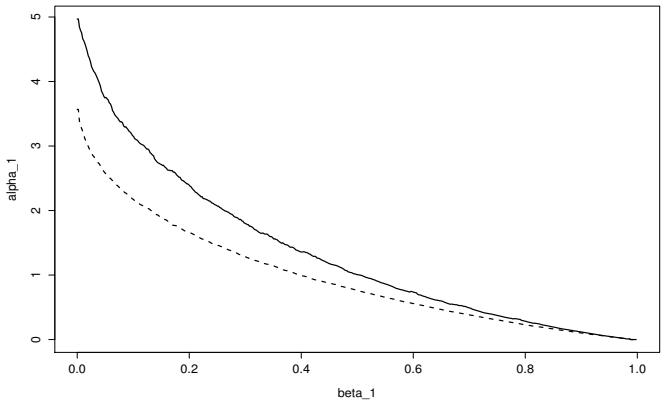

Figure 5.9. *The (α_1, β_1)-areas below the two curves guarantee the existence of a strictly stationary* GARCH(1,1) *process.* Solid line: *IID student noise with* 4 *degrees of freedom with variance* 1. Dotted line: *IID standard normal noise. The regions were determined by checking the condition* $E\log(\alpha_1 Z_1^2 + \beta_1) < 0$.

Hence

$$E[Y_0^k] = \frac{\sum_{i=0}^{k-1} \binom{k}{i} E[A_1^i]\, E[Y_0^i]\alpha_0^{k-i}}{1 - E[A_1^k]}$$

provided the denominator satisfies

$$1 - E[A_1^k] = 1 - E[(\alpha_1 Z_0^2 + \beta_1)^k] > 0\,.$$

This leads us to a recursive relation for the moments $E[Y_0^k]$ given we know α_0 and the moments $E[A_1^i] = E[(\alpha_1 Z_0^2 + \beta_1)^i]$. For example, we have

$$\begin{aligned} E[\sigma_0^2] &= \frac{\alpha_0}{1 - (\alpha_1 + \beta_1)}\,, \\ E[\sigma_0^4] &= \frac{\alpha_0^2 + 2\alpha_0 \dfrac{(\alpha_1 + \beta_1)\alpha_0}{1 - (\alpha_1 + \beta_1)}}{1 - E[(\alpha_1 Z_0^2 + \beta_1)^2]} \end{aligned}$$

The calculation of $E[\sigma_0^{2k}]$ is only possible if $E[(\alpha_1 Z_0^2 + \beta_1)^k] < 1$. However, if $P(\alpha_1 Z_0^2 + \beta_1 > 1) > 0$ we have $E[(\alpha_1 Z_0^2 + \beta_1)^k] \to \infty$ as $k \to \infty$. This means that certain moments of σ_t are infinite, hence certain moments of X_t are infinite. Since

$$E[|X_0|^p] = E[\sigma_0^p]\, E[|Z_0|^p]\,, \qquad p > 0\,,$$

the responsibility for these infinite moments is due to the distribution of σ_0 provided Z_0 has all moments finite. For example, if we assume that Z_0 has the standard normal distribution then we also have $E[\sigma_0^p] = \infty$ for some $p > 0$. When looking at the structure of σ_t^2 given by the infinite series in (5.10), it may be surprising that certain moments of σ_t^2 can be infinite although each summand in the infinite series representation has all moments finite. An explanation of this phenomenon is given by results of Kesten [23] and Goldie [18]. They deal with the general stochastic recurrence equation $Y_t = A_t Y_{t-1} + B_t$ for an iid sequence $\big((A_t, B_t)\big)_{t\in\mathbb{Z}}$, the process (σ_t^2) for a GARCH(1,1) process $X_t = \sigma_t Z_t$ being a particular example. These results explain that, under mild conditions on the distribution of Z_0 and if the equation

$$E[A_1^{\kappa/2}] = E[(\alpha_1 Z_0^2 + \beta_1)^{\kappa/2}] = 1$$

has a positive solution κ, this solution is unique and one has the following relation for the tails of σ_0 and X_0:

$$P(\sigma_0 > x) \sim c_0\, x^{-\kappa}\,, \quad \text{and} \quad P(\pm X_0 > x) \sim E[(Z_0)_\pm^\kappa]\, P(\sigma_0 > x)\,, \quad x \to \infty$$

for a positive constant c_0. Then, in particular, $E[|X_0|^\kappa] = \infty$.

Power-law tails are often observed for returns of financial time series, including stock indices, foreign exchange rates, stock prices. For an illustration, see Figures 5.10 and 5.11.

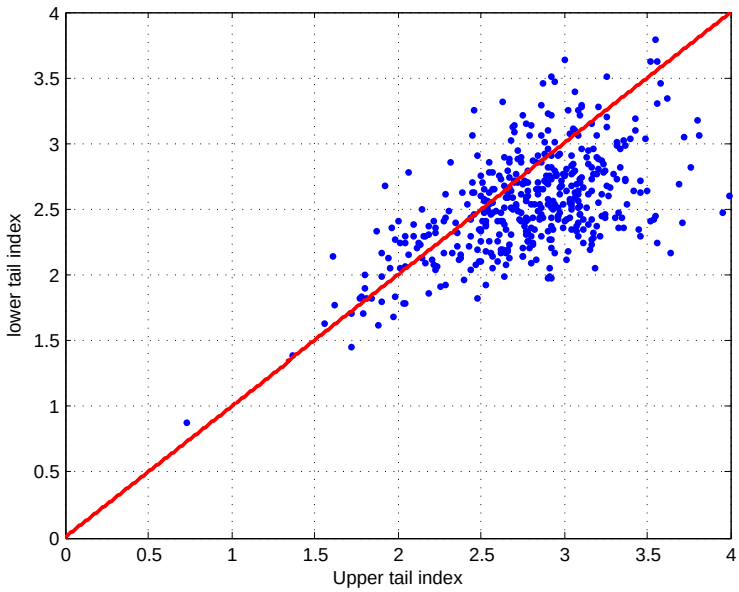

Figure 5.10. *Estimates for the upper and lower tail indices of the returns of the 500 components of the S&P 500 composite stock index. This means that we assume that there exist positive κ_{up} and κ_{low} such that for any of the 500 series, say (X_t), $P(X_t > x) \sim c_+ x^{-\kappa_{up}}$ and $P(X_t < -x) \sim c_- x^{-\kappa_{low}}$ as $x \to \infty$. The graph shows that the estimates of the tail indices are typically between 2 and 4, implying that the data would not have a finite 4th moment.*

Consider a GARCH(1,1) process (X_t) with power-law tails with index κ. Write

$$Y_t = (X_t)_+ = \max(X_t, 0) \qquad \text{and} \qquad M_n = \max(Y_1, \ldots, Y_n)\,.$$

Then $P(Y_t > x) = P(X_t > x) = c_+\, x^{-\kappa}(1 + o(1))$ as $x \to \infty$. Assume for the moment that (Y_t) is an iid sequence. Then

$$\begin{aligned} P\big((c_+ n)^{-1/\kappa} M_n \le x\big) &= \big[P\big((c_+ n)^{-1/\kappa} Y_1 \le x\big)\big]^n \\ &= \big[1 - P\big((c_+ n)^{-1/\kappa} Y_1 > x\big)\big]^n \\ &= \Big[1 - \frac{1}{n\, x^\kappa}(1 + o(1))\Big]^n \\ &\to \mathrm{e}^{-x^{-\kappa}} = \Phi_\kappa(x)\,,\ n \to \infty\,, \qquad x > 0\,. \end{aligned}$$

The distribution function Φ_κ belongs to the *Fréchet distribution* which is one of the three possible non-degenerate limit distributions for normalized and centered maxima of an iid sequence. Since a GARCH(1,1) process is a dependent sequence one needs to correct the limit distribution by a positive number:

$$P\big((c_+ n)^{-1/\kappa} M_n \le x\big) \to \Phi_\kappa^\theta(x)\,, \quad n \to \infty\,, \quad x > 0\,,$$

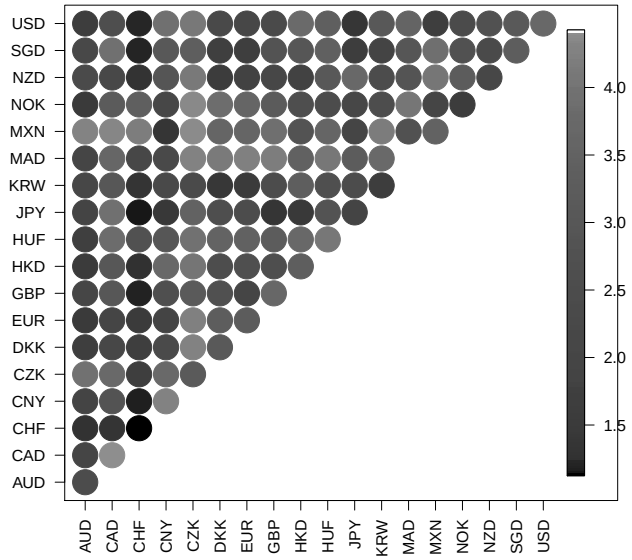

Figure 5.11. *Estimates for the lower tail indices of some foreign exchange rate returns. All indices are between 2 and 4.*

where $\theta \in (0,1)$ is the *extremal index* of the GARCH(1,1) sequence; see Embrechts et al. [13], Section 8.4 for details. The extremal index θ is a measure of the size of extremal clusters above high thresholds. Notice that Φ_κ^θ is a distribution in the scale family of Φ_κ.

5.4. **Why GARCH?.** The popularity of the GARCH model can be explained by various arguments.

- Its relation to ARMA processes suggests that the theory behind it might be closely related to ARMA process theory which is well studied, widely known and seemingly "easy".

This opinion is, however, wishful thinking. The difference to standard ARMA processes is due to the fact that the noise sequence (ν_t) in (5.7) depends on the X_t's themselves, so a complicated non-linear relationship of the X_t's builds up. For example, in order to show that a stationary version of (X_t^2) exists, one would have to iterate equation (5.7), hoping that X_t^2 becomes an explicit expression only of the sequence (ν_t) which expression one might take as the solution to the difference equations (5.7). For an iid noise sequence (ν_t) this recipe is known to work; see Brockwell and Davis [8], Chapter 3, who study conditions for the validity of this approach. However, the noise $\nu_t = X_t^2 - \sigma_t^2$ itself depends on the stationary sequence (X_t) to be constructed, and so one has basically gained not so much by this approach.

If one knows that (X_t) is a well defined strictly stationary process, the relation with ARMA processes can be useful. For example, one can derive formulae for the moments of X_t^2 by using the moments of an ARMA process in terms of the ARMA parameters and the moments of the underlying noise sequence (ν_t). Moreover, if this ARMA process is causal we also know that the autocovariance function $\gamma_{X^2}(h)$ of (X_t^2) decays exponentially fast as $|h| \to \infty$. This shows that (X_t^2) has exponentially short memory.

Conditions for the existence of a strictly stationary version of a GARCH process are not easy and difficult to verify if the order (p,q) is such that $p > 1$ or $q > 1$. They are based on multivariate versions of the stochastic recurrence equations of Section 5.1; see Mikosch [25] and Buraczewski et

al. [10] for a book treatment. Axceptions are the ARCH(1) and GARCH(1,1) processes for which necessary and sufficient conditions for the existence of a strictly stationary version of (X_t) in terms of α_0, α_1, β_1 and the distribution of Z_1 are known.

A second argument in favor of GARCH processes is the fact that, under mild conditions,

- the tails of a strictly stationary GARCH process have power law tails. In particular, certain moments of X_t are infinite. This is in agreement with the stylized facts for financial return: the heavy tails of X_t cause the occurrence of very large positive and very small negative values in observed return time series.

A third argument *for* the use of GARCH models is that,

- even for a GARCH(1,1) model with three parameters one often gets a reasonable fit to real-life financial data, provided that the sample has not been chosen from a too long period making the stationarity assumption questionable. Tests for the residuals of GARCH(1,1) models with estimated parameters $\alpha_0, \alpha_1, \beta_1$ give the impression that the residuals very much behave like an iid sequence.

Some evidence on this issue can be found in the paper of Mikosch and Stărică [26]; see Figures 5.12 and 5.13.

A fourth argument for the GARCH is the following:

- The GARCH model allows for a simple distributional forecast.

Indeed, the definition of the GARCH model tells us that the distribution of X_{t+1} given the past $X_t, X_{t-1}, \ldots$ is the conditional distribution of X_{t+1} given σ_t. For example, if we assume (Z_t) iid standard Gaussian, then X_{t+1} has the conditional $\mathrm{N}(0, \sigma_t^2)$ distribution. This distribution can be updated every day, depending on the observations $X_1, \ldots, X_t$, the parameter estimates $\widehat{\alpha}_i$, $\widehat{\beta}_j$ and the resulting calculated values $\widehat{\sigma}_1, \ldots, \widehat{\sigma}_t$ which are obtained by plugging the X_i's, $\widehat{\alpha}_i$, $\widehat{\beta}_j$ into the definition of σ_{t+1}^2 and by choosing some initial values for the $\widehat{\sigma}$-values. We refer to Figure 5.14 for an illustration of this simple forecast procedure.

A fifth, and perhaps the most powerful argument in favor of GARCH models, from an applied point of view, is the fact that

- the statistical estimation of the parameters of a GARCH process is rather uncomplicated; see Section 5.5.

This attractive property has led S+ to provide us with a module for the statistical inference and simulation of GARCH models, called S+FinMetrics.

5.5. **Gaussian quasi-maximum likelihood.** The estimation technique used most frequently in applications is a *Gaussian quasi-maximum likelihood* procedure which we want to explain briefly. Assume for the moment that the noise (Z_t) in an ARCH(p) model of a given order p is iid standard normal. Then X_t is Gaussian $N(0, \sigma_t^2)$ given the whole past $X_{t-1}, X_{t-2}, \ldots$, and a conditioning argument yields the density function $f_{X_p,\ldots,X_n}$ of $X_p, \ldots, X_n$ through the conditional Gaussian densities of the X_t's given $X_1 = x_1, \ldots, X_n = x_n$:

$$
\begin{aligned}
& f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) \\
&\quad = \; f_{X_n}(x_n \,|\, X_{n-1} = x_{n-1}, \ldots, X_1 = x_1)\, f_{X_{n-1}}(x_{n-1} \,|\, X_{n-2} = x_{n-2}, \ldots, X_1 = x_1) \cdots \\
&\qquad f_{X_{p+1}}(x_{p+1} \,|\, X_p = x_p, \ldots, X_1 = x_1)\, f_{X_1,\ldots,X_p}(x_1, \ldots, x_p) \\
&\quad = \; (2\pi)^{-(n-p)/2} \prod_{t=p+1}^{n} \sigma_t^{-1} \mathrm{e}^{-x_t^2/(2\sigma_t^2)}\, f_{X_1,\ldots,X_p}(x_1, \ldots, x_p)\,. \qquad (5.11)
\end{aligned}
$$

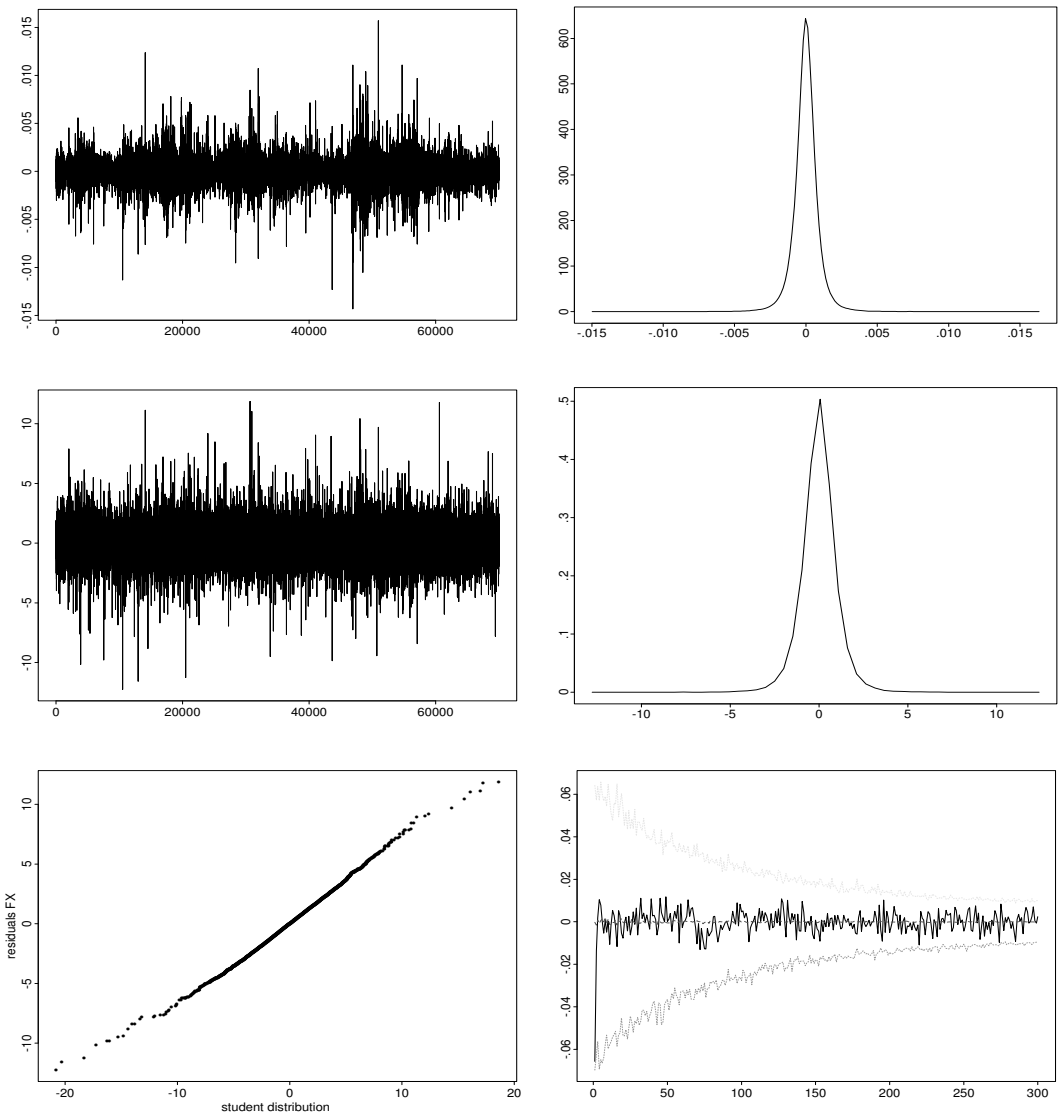

Figure 5.12. Top: $70,000$ *values of* 30 *minute foreign exchange JPY-USD log-returns between* 1992 *and* 1996 (left) *and their density* (right). Middle: *The residuals of the JPY-USD foreign exchange log-returns* (left) *after fitting a* GARCH$(1,1)$ *with parameters* $\alpha_0 = 10^{-7}$, $\alpha_1 = 0.11$ *and* $\beta_1 = 0.88$. *This means that one calculates the values* $\widehat{Z}_t = X_t/\widehat{\sigma}_t$, *where the* $\widehat{\sigma}_t^2$ *are calculated from the definition of a* GARCH(1,1) *process and the parameters* α_i *and* β_1 *are replaced by their estimators. The density of the residuals* (right). *The scale difference on the* x*-axis when compared with the foreign exchange density is due to the standardization* $\mathrm{var}(Z_0) = 1$. Bottom, left: *QQ-plot of the* GARCH(1,1) *residuals against the quantiles of a student distribution with* 4 *degrees of freedom. The residuals are nicely fitted by this distribution. Notice that this distribution has very heavy tails in the sense that its* 4*th moment is infinite, implying that the distribution of the returns must also be very heavy-tailed.* Bottom, right: *The sample ACF of the foreign exchange rate data with* 95% *asymptotic confidence bands for a fitted* GARCH(1,1) *process.*

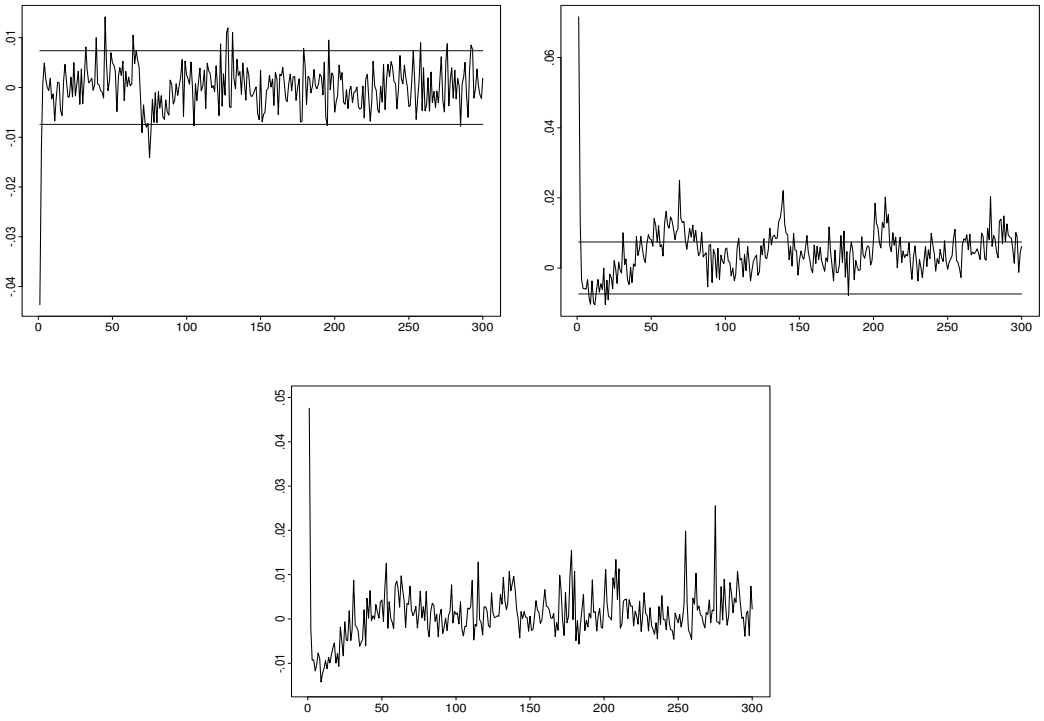

Figure 5.13. *Continuation of Figure 5.12. The sample ACFs of the residuals of the foreign exchange rate log-returns* (top, left)*, their absolute values* (top, right) *and squares* (bottom)*. The straight lines in the two upper graphs indicate the* $\pm 1.96/\sqrt{n}$ *asymptotic confidence bands for an iid sequence with finite second moment. In the lower graph we refrain from giving* $(1/\sqrt{n})$*-confidence bands because* Z_1 *possibly has an infinite 4th moment. Compare with the sample ACFs of the foreign exchange log–returns in Figure 5.12, in particular observe the differences in scale.*

Here we used the fact that the conditional density of a vector $(\mathbf{A}, \mathbf{B})$ given $\mathbf{B}$ can be expressed by the corresponding densities $f_{\mathbf{A},\mathbf{B}}$ and $f_{\mathbf{B}}$ by

$$f_{\mathbf{A}}(\mathbf{a} \mid \mathbf{B} = \mathbf{b}) = \frac{f_{\mathbf{A},\mathbf{B}}(\mathbf{a}, \mathbf{b})}{f_{\mathbf{B}}(\mathbf{b})}\,.$$

Ignoring the density $f_{X_1,\ldots,X_p}$ and replacing $t = p+1$ by $t = 1$ in (5.11), the "Gaussian log-likelihood" of $X_1, \ldots, X_n$ is given by

$$\begin{aligned} L_n(\alpha_0, \alpha_1, \ldots, \alpha_p, \beta_1, \ldots, \beta_q)(X_1, \ldots, X_n) &= L_n(\theta)(X_1, \ldots, X_n) \\ &= -\frac{1}{2n} \sum_{t=1}^{n} [2 \log \sigma_t + \sigma_t^{-2} X_t^2] \\ &= -\frac{1}{2n} \sum_{t=1}^{n} [2 \log \sigma_t(\theta) + \sigma_t^{-2}(\theta) X_t^2]\,, \end{aligned} \tag{5.12}$$

where $\theta = (\alpha_0, \alpha_1, \ldots, \alpha_p, \beta_1, \ldots, \beta_q)'$ is any parameter in a suitable parameter space and θ_0 is the true parameter of the GARCH model, where the observations $X_t = \sigma_t(\theta_0) Z_t$ come from. The quantity (5.12) is also formally defined for general GARCH(p, q) processes and it can be maximized as a function of the α_i's and β_j's involved. The resulting value in the parameter space is the *Gaussian quasi-maximum likelihood estimator* (MLE) of the parameters of a GARCH(p, q) process.

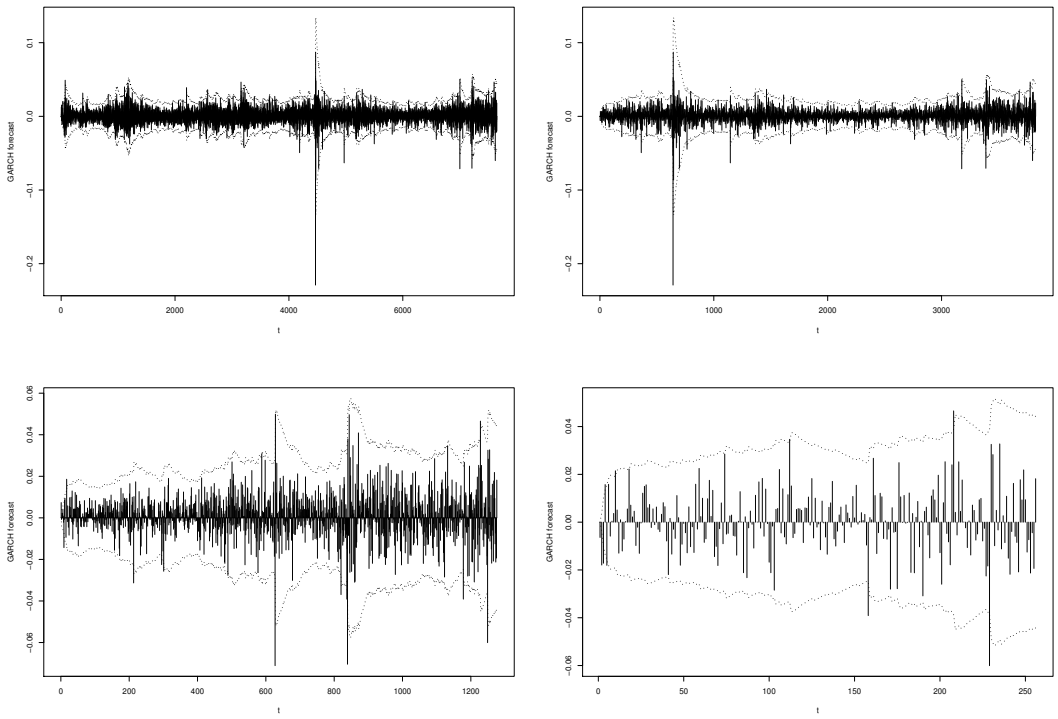

Figure 5.14. *One day* 95% *distributional forecasts of log-returns of the S&P500 composite stock index* (*from top left, top right, bottom left to bottom right:* 30*,* 15*,* 5*,* 1 *years of data*) *based on a* GARCH(1,1) *model with iid standard normal noise and parameters* $\alpha_0 = 10^{-6}$*,* $\alpha_1 = 0.07$*,* $\beta_1 = 0.96$*. The extreme values of the log-returns are not correctly captured by the model.*

There are obvious problems with this estimation procedure. For example, one might be surprised about the assumption of Gaussian noise (Z_t). Although this is not the most realistic assumption[6] theoretical work (see the references below) shows that asymptotic properties such as $\sqrt{n}$*-consistency* (i.e., consistency and asymptotic normality with $\sqrt{n}$-rate) of the Gaussian quasi-MLE remain valid for large classes of noise distributions. This observation is similar to other estimation procedures in time series analysis where one does not maximize the "true" maximum likelihood function of the underlying data but rather assumes Gaussianity of the data and maximizes a corresponding score function. This approach works for the Gaussian maximum likelihood of ARMA processes (see Brockwell and Davis [8], Section 10.8, and Section 4.3 in these notes) and in more general situations.

Attempts to replace the Gaussian densities in L_n by a "more realistic" density of the Z_t's (for example, a t-density) can lead to non-consistency of the MLE. Consistency of the estimators can be achieved if one knows the exact density underlying Z_t but when dealing with data one can never rely on this assumption. Even if one tries to estimate the parameters of the density of Z_t together with the GARCH parameters (for example, some professional software offers to estimate the degrees of freedom of t-distributed Z_t's from the data) the MLE based on these densities can lead to non-consistent estimators.[7]

The careful reader might also have observed that the derivation of the maximum likelihood function (5.11) is not directly applicable if the model deviates from an ARCH(p) process. Indeed,

[6]Empirical evidence indicates that the Z_t's are much better modeled by a t-distribution; see Figure 5.12 for some evidence.

[7]These facts I learned from Daniel Straumann.

that formula requires calculating the *unobservable* values σ_t, $t = 1, \ldots, n$, from the *observed* sample $X_1, \ldots, X_n$. A glance at the defining formula (5.8) convinces one that this is not possible in the general GARCH(p, q) case. Indeed, an iteration of (5.8) yields that one would have to know *all* values $X_{n-1}, \ldots, X_0, X_{-1}, \ldots$ for the calculation of $\sigma_1, \ldots, \sigma_n$. Alternatively, one needs to know finitely many values of the unobservable values $X_0, X_{-1}, \ldots$ and $\sigma_0, \sigma_{-1}, \ldots$. Therefore practitioners (and software packages) have to choose a finite number of such initial values in order to make the iteration for the σ's run. The choice of deterministic initial values implies that the calculated $\sigma_1, \ldots, \sigma_n$ cannot be considered as a realization of a stationary sequence. One may, however, hope that the dependence on the initial values disappears for large values of n in a way similar to a Markov chain with arbitrary initial value whose distribution becomes closer to the stationary distribution, and this hope can be justified by theoretical means; see Berkes et al. [3] and Straumann and Mikosch [34].

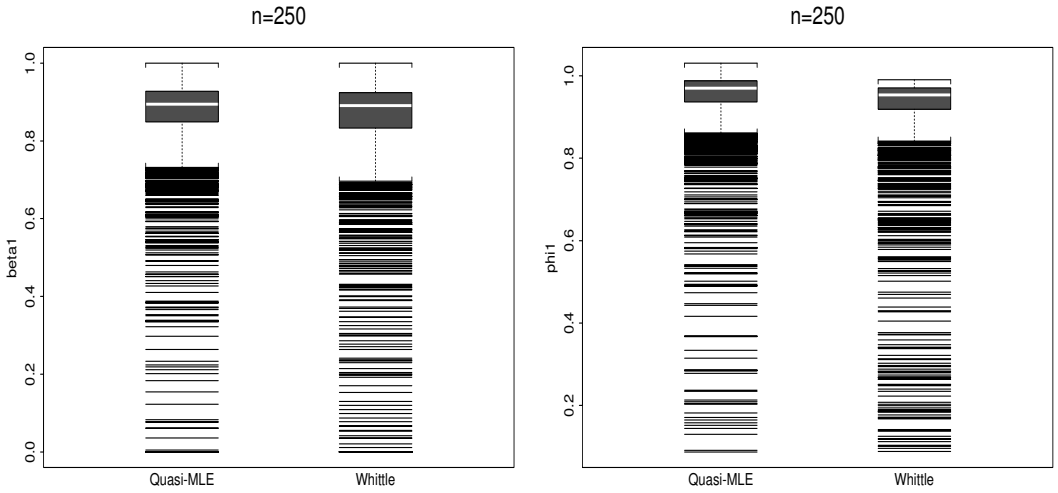

Figure 5.15. *A boxplot comparison of the distributions of the Gaussian quasi-MLE and the Whittle estimator, another important estimation technique for ARMA processes, for* β_1 (left) *and* $\varphi_1 = \alpha_1 + \beta_1$ (right) *in a* GARCH(1,1) *model with parameters* $\alpha_0 = 8.58 \times 10^{-6}$, $\alpha_1 = 0.072$, $\beta_1 = 0.92$. *The sample size is* $n = 250$. *The boxplots are based on* 1000 *independent repetitions of the parameter estimation procedures.*

Simulation results, see Figure 5.15, indicate that the Gaussian quasi-MLE does not work too well for small sample sizes of a couple of hundred values. On the other hand, it is not very realistic to fit a particular GARCH model to several years of daily log-returns — the data do not behave like a stationary process over such long periods of time. The accuracy of the estimation procedure based on one business year of data (250 days) is rather non-satisfactory. The poor behavior of the quasi-MLE for GARCH models seems to be due to the fact that the log-likelihood function L_n in (5.12) is rather flat in the parameter space and therefore it is difficult to find its maximum.

5.6. **Some ideas about the proof of the asymptotic normality of the Gaussian quasi-MLE.** An excellent reference to parameter estimation in GARCH models is Straumann [33]. There one also finds a proof of the consistency and asymptotic normality of the maximizer θ_n of the Gaussian likelihood function $L_n(\theta)$ in (5.12) for a suitable parameter space C such that the true

parameter θ_0 underlying the data $X_1, \ldots, X_n$ is an inner point of C. Recall the Gaussian log-likelihood of $X_1, \ldots, X_n$ given by

$$\begin{aligned} L_n(\theta) &= L_n(\theta)(X_1, \ldots, X_n) \\ &= -\frac{1}{2n} \sum_{t=1}^{n} [\log \sigma_t^2(\theta) + \sigma_t^{-2}(\theta) X_t^2] \,. \end{aligned}$$

For convenience, we write $h_t(\theta) = \sigma_t^2(\theta)$ and

$$\ell_t(\theta) = -\frac{1}{2} \Big[\log h_t(\theta) + \frac{X_t^2}{h_t(\theta)} \Big] \,.$$

Then its gradient (derivative with respect to all parameters α_i and β_j) is given by

$$\ell_t'(\theta) = -\frac{1}{2} \frac{h_t'(\theta)}{h_t(\theta)} \Big[1 - \frac{X_t^2}{h_t(\theta)} \Big]$$

Then we also have

$$L_n(\theta) = \frac{1}{n} \sum_{t=1}^{n} \ell_t(\theta) \,.$$

A first order Taylor expansion yields

$$L_n(\theta_n) = L_n(\theta_0) + L_n'(\theta_0) \, (\theta_n - \theta_0) + R_n^{(1)} \,, \tag{5.13}$$

where the remainder can be shown to satisfy $R_n^{(1)} \stackrel{\text{a.s.}}{\to} 0$ as $n \to \infty$. Notice that

$$h_t(\theta) = f_\theta(Z_{t-1}, Z_{t-2}, \ldots)$$

for some function f acting on the shifts of the iid sequence (Z_t) and therefore $(h_t(\theta))$ is an ergodic sequence of stochastic processes indexed by $\theta \in C$ and $\ell_t(\theta), \ell_t'(\theta)$ inherit ergodicity.

The ergodic theorem yields

$$\begin{aligned} L_n(\theta_0) &= \frac{1}{n} \sum_{t=1}^{n} \ell_t(\theta_0) \\ &\stackrel{\text{a.s.}}{\to} E\ell_0(\theta_0) = -\frac{1}{2} \big(E \log h_0(\theta_0) + E[X_0^2 / h(\theta_0)] \big) \\ &= -\frac{1}{2} \big(E \log h_0(\theta_0) + E[Z_0^2] \big) = -\frac{1}{2} \big(E \log h_0(\theta_0) + 1 \big) \\ L_n'(\theta_0) &= \frac{1}{n} \sum_{t=1}^{n} \ell_t'(\theta_0) \\ &\stackrel{\text{a.s.}}{\to} E\ell_0'(\theta_0) = -\frac{1}{2} E \Big[\frac{h_0'(\theta_0)}{h_0(\theta_0)} \Big(1 - \frac{X_0^2}{h_0(\theta_0)} \Big) \Big] \\ &= -\frac{1}{2} E \Big[\frac{h_0'(\theta_0)}{h_0(\theta_0)} \Big] \, (1 - E[Z_0^2]) \\ &= \mathbf{0} \,. \end{aligned}$$

Here we used the fact that h_0 is a function of $Z_{-1}, Z_{-2}, \ldots$, independent of Z_0^2 with $\operatorname{var}(Z_0) = 1$ and $X_0^2 = h_0(\theta_0) Z_0^2$. It follows from (5.13) that

$$L_n(\theta_n) - L_n(\theta_0) \stackrel{\text{a.s.}}{\to} 0 \,.$$

Using the properties of the pointwise limiting function

$$-\frac{1}{2} \big(E \log h_0(\theta) + 1 \big) \,, \qquad \theta \in C \,,$$

which has a unique maximum at $\theta = \theta_0 \in C$, one can show that $\theta_n \overset{\text{a.s.}}{\to} \theta_0$.

In a next step we consider the Taylor expansion

$$L_n'(\theta_n) = L_n'(\theta_0) + L_n''(\theta_0)\,(\theta_n - \theta_0) + R_n^{(2)}\,. \tag{5.14}$$

Notice that

$$L_n''(\theta_0) = \frac{1}{n}\sum_{t=1}^{n} \ell_t''(\theta_0)\,,$$

where the second derivative $\ell_t''(\theta_0)$ is a matrix and can be written as a function of $Z_t, Z_{t-1}, \ldots$, hence the ergodic theorem applies:

$$L_n''(\theta_0) \overset{\text{a.s.}}{\to} \mathbf{F}_0 = E[\ell_0''(\theta_0)]\,, \tag{5.15}$$

and the limiting matrix $\mathbf{F}_0$ is invertible. One can show that $\sqrt{n}\,R_n^{(2)} \overset{P}{\to} 0$ as $n \to \infty$. Now, since θ_n maximizes $L_n(\theta)$ in C and therefore $L_n'(\theta_n) = \mathbf{0}$, we have in view of (5.14)

$$\sqrt{n}\,L_n'(\theta_0) = \big(L_n''(\theta_0)\big)^{-1}\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\frac{h_t'(\theta_0)}{\sigma_t^2}\,(Z_t^2-1) + o_P(1) = \sqrt{n}\,\big(\theta_n - \theta_0\big) + o_P(1)\,, \quad n \to \infty\,,$$

where $o_P(1)$ converges to zero in probability. With (5.15) we also have

$$\sqrt{n}\,\big(\theta_n - \theta_0\big) = \mathbf{F}_0^{-1}\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\frac{h_t'(\theta_0)}{\sigma_t^2}\,(Z_t^2-1) + o_P(1)\,. \tag{5.16}$$

Now let $\mathcal{F}_t = \sigma(Z_t, Z_{t-1}, \ldots)$ be the σ-field generated by $Z_t, Z_{t-1}, \ldots,$. We notice that the summands on the right-hand side have the following properties:

- For any t,

$$\mathbf{Y}_t = \frac{h_t'(\theta_0)}{\sigma_t^2}\,(Z_t^2-1) \tag{5.17}$$

  is $\mathcal{F}_t$-measurable, i.e., it is a function of $Z_t, Z_{t-1}, \ldots$.

- $$E\Big[\frac{h_t'(\theta_0)}{\sigma_t^2}\,(Z_t^2-1) \mid \mathcal{F}_{t-1}\Big] = \frac{h_t'(\theta_0)}{\sigma_t^2}\,E[Z_t^2-1] = 0\,.$$

  since $h_t'(\theta_0)/\sigma_t^2$ only depends on $Z_{t-1}, Z_{t-2}, \ldots$ and Z_t is independent of it.

Now recall the definition of a *martingale difference sequence* $(\mathbf{Y}_t)_{t=0,1,2,\ldots}$ with values in $\mathbb{R}^d$ with respect to a filtration $(\mathcal{G}_t)$ of σ-fields $\mathcal{G}_t$, i.e., $\mathcal{G}_{t-1} \subset \mathcal{G}_t$ for any t.

- $E[|\mathbf{Y}_t|] < \infty$ for all t.
- $\mathbf{Y}_t$ is $\mathcal{G}_t$-measurable for all t.
- $E[\mathbf{Y}_t \mid \mathcal{G}_{t-1}] = 0$ a.s.

Therefore the sequence $(\mathbf{Y}_t)$ in (5.17) is a martingale difference sequence with respect to the filtration $(\mathcal{F}_t)$.

In Billingsley [4], Theorem 23.1, we find the following elegant central limit theorem for martingale difference sequences:

Theorem 5.16. (Billingsley's central limit theorem) *Let (ξ_t) be a strictly stationary ergodic real-valued martingale difference sequence with respect to a filtration $(\mathcal{H}_t)_{t=0,1,2,\ldots}$. Assume $E\xi_0 = 0$ and $\sigma^2 = \operatorname{var}(\xi_0) < \infty$. Then the central limit theorem holds*

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\xi_t \overset{d}{\to} N(0,\sigma^2)\,.$$

Our sequence $(\mathbf{Y}_t)$ in (5.17) is vector-valued. However, we may use the *Cramér-Wold device:* it suffices to show that for any column vector $\mathbf{c}$,

$$\frac{1}{\sqrt{n}} \sum_{t=1}^{n} \mathbf{c}'\mathbf{Y}_t \xrightarrow{d} c'\mathbf{Y}\,, \qquad n \to \infty\,,$$

where $\mathbf{Y}$ is Gaussian $N(\mathbf{0}, \mathbf{G_0})$ column vector with

$$\mathbf{G}_0 = E[Z^4 - 1]E\big[(h'(\theta_0))^T h'(\theta_0)/\sigma_0^4\big]\,. \tag{5.18}$$

Here we also used the fact that $\mathbf{c}'\mathbf{Y}_t$ is a martingale with respect to $(\mathcal{F}_t)$ for every fixed $\mathbf{c}$.

We combine (5.18) and (5.16) and obtain:

Theorem 5.17. *Under suitable conditions, the Gaussian quasi-maximum likelihood estimator θ_n of θ_0 in the GARCH(p,q) model is strongly consistent and asymptotically normal with $N\big(\mathbf{0}, \mathbf{F_0^{-1}G_0F_0^{-1}}\big)$ limit.*

For a detailed proof, see Straumann [33], Theorem 5.6.1.

6. Spectral analysis of time series

6.1. **An example.** In what follows, we will assume that the time series (X_t) is complex-valued. This assumption is for mathematical convenience only and has some tradition in the area of signal processing and electrical engineering. Since we are only interested in real-valued time series the corresponding results follow by considering the real part of the random variables. We will deal with the Fourier analysis of stationary processes. In Fourier analysis, and in applied areas such as electrical engineering, it is standard to work with the complex trigonometric function system $(\mathrm{e}^{\,ikt})_{t\in[-\pi,\pi],k\in\mathbb{Z}}$, and so we will follow this pattern.

We first define some standard notions:

Definition 6.1. (Complex-valued stationary process)
A complex-valued process $(X_t)_{t\in\mathbb{Z}}$ is stationary *if the following properties hold:*
1) $EX_t = m$ for a constant m, all t.
2) $E|X_t|^2 < \infty$ for all t.
3) $E(X_t\overline{X}_{t+h})$ does not depend on t.

Notice that $(X,Y) = E(X\overline{Y})$ defines an inner product in the space $L^2[\Omega,\mathcal{F},P]$ of square-integrable random variables on $[\Omega,\mathcal{F},P]$. Equipped with this inner product, $L^2[\Omega,\mathcal{F},P]$ is a Hilbert space. For a stationary process (X_t) we introduce the *autocovariance function*:

$$\gamma_X(h) = E[\overline{(X_t - EX_0)}(X_{t+h} - EX_0)]\,, \quad h \in \mathbb{Z}\,.$$

The following simple example of a complex-valued stationary process will turn out to be crucial for the understanding of the structure of a general stationary process.

Example 6.2. (Linear combination of sinusoids)
Let $(A(\lambda_j))_{j=1,\dots,n}$ be uncorrelated complex-valued random variables such that

$$EA(\lambda_j) = 0\,, \quad E|A(\lambda_j)|^2 = \sigma_j^2 > 0\,, \quad j = 1,\dots,n\,.$$

Here $-\pi < \lambda_1 < \cdots < \lambda_n \le \pi$ are the so-called *frequencies*. We consider the process

$$X_t = \sum_{j=1}^{n} \mathrm{e}^{\,it\lambda_j}\, A(\lambda_j)\,, \quad t \in \mathbb{Z}\,. \tag{6.1}$$

It is stationary:

$$EX_t = 0 \quad \text{and} \quad E(\overline{X}_t\, X_{t+h}) \;=\; \sum_{j=1}^{n} \sigma_j^2\, \mathrm{e}^{\,ih\lambda_j}\,.$$

Thus the autocovariance function of this process is given by

$$\gamma_X(h) = \sum_{j=1}^{n} \sigma_j^2 \, \mathrm{e}^{\,ih\lambda_j} \; . \tag{6.2}$$

This formula bears some resemblance with the expectation of a discrete random variable $\mathrm{e}^{\,ihX}$ with atoms at λ_j. However, (6.2) is in general not an expectation since, in general, $\sum_j \sigma_j^2 \neq 1$. Nevertheless, in analogy to the expectation of a discrete random variable we can write

$$\gamma_X(h) = \int_{(-\pi,\pi]} \mathrm{e}^{\,ih\lambda} \, dF_X(\lambda) \; , \tag{6.3}$$

as a Lebesgue–Stieltjes integral with respect to the *spectral distribution function*

$$F_X(\lambda) = \sum_{j:\, \lambda_j \leq \lambda} \sigma_j^2 \; , \quad \lambda \in [-\pi, \pi] \; .$$

It is, in general, not the distribution function of a probability distribution. Relation (6.3) is the *spectral representation of the autocovariance function* γ_X.

It is also possible to give a (random) spectral representation of the process (X_t). The stochastic process

$$Z_X(\lambda) = \sum_{j:\lambda_j \leq \lambda} A(\lambda_j) \; , \quad \lambda \in (-\pi, \pi] \; .$$

jumps at the λ_j's by the values $A(\lambda_j)$. Then we can interpret (6.1) as the integral representation

$$X_t = \int_{(-\pi,\pi]} \mathrm{e}^{\,it\lambda} dZ_X(\lambda) \; , \quad t \in \mathbb{Z} \; . \tag{6.4}$$

The integral above is *stochastic* which means that we integrate with respect to a random measure or with respect to a stochastic process Z_X very much in the spirit of a Lebesgue-Stieltjes integral. In contrast to a classical Lebesgue-Stieltjes integral, the weights at the λ_j's are random and may assume complex values.

Since the $A(\lambda_j)$'s are uncorrelated,

$$E[(Z_X(b) - Z_X(a))\overline{(Z_X(d) - Z_X(c))}] = 0 \quad \text{if } (a,b] \cap (c,d] = \emptyset.$$

Such a process is called a *process with orthogonal increments*. Indeed, in the Hilbert space $L^2[\Omega, \mathcal{F}, P]$ the inner product of the increments of the stochastic process Z_X on the disjoint sets $(a,b]$ and $(c,d]$ is zero, hence the increments are orthogonal in this space. Also notice that

$$E|Z_X(b) - Z_X(a)|^2 = F_X(b) - F_X(a) \, , \quad a < b \, , \tag{6.5}$$

which closely links the process Z_X and the spectral distribution function F_X.

The importance of this example is that it shows in a simple way the features which are typical for *any* stationary process. Indeed, every stationary process (X_t) can be shown to have a representation as a stochastic integral (6.4) with respect to a process Z_X with orthogonal increments which defines a spectral distribution function F_X by means of formula (6.5) and the autocovariance function γ_X then has representation (6.3).

In the following we will make the notion of the spectral distribution function more precise.

6.2. **The spectral representation of a stationary process.** Recall from Section 3 that the autocovariance function of a real-valued stationary process is *non-negative definite*. This notion can easily be extended to complex-valued stationary processes: the function $\gamma : \mathbb{Z} \to \mathbb{C}$ is said to be *non-negative definite* if

$$\sum_{i,j=1}^{n} a_i \overline{a}_j \gamma(i-j) \geq 0 \,,$$

for any choice of complex numbers $(a_i)_{i=1,\ldots,n}$ and $n \geq 1$. Moreover, a result analogous to the real-valued case holds: $\gamma : \mathbb{Z} \to \mathbb{C}$ is the autocovariance function of a complex-valued stationary process if and only if $\gamma(-h) = \overline{\gamma(h)}$ for all $h \in \mathbb{Z}$ and γ is non-negative definite. This fact will help us to better understand the following important theorem which relates autocovariance functions and distribution functions:

Theorem 6.3. (Herglotz's theorem)
The function $\gamma : \mathbb{Z} \to \mathbb{C}$ with $\gamma(-h) = \overline{\gamma(h)}$ is the autocovariance function of a stationary process if and only if there exists a right–continuous, non–decreasing, bounded function F on $[-\pi, \pi]$ such that $F(-\pi) = 0$ and

$$\gamma(h) = \int_{(-\pi,\pi]} \mathrm{e}^{\,ih\lambda} \, dF(\lambda) \,, \quad h \in \mathbb{Z} \,. \tag{6.6}$$

The function F satisfying (6.6) is unique. This follows in the same way as proving that there is a unique relationship between a probability distribution and a characteristic function, by applying the inversion formula for characteristic functions. However, notice that F is in general not a *probability distribution function* since $F(\pi) \neq 1$ is possible.

Proof. We only prove the sufficiency part. It suffices to prove that γ as defined in (6.6) is non-negative definite:

$$\begin{aligned} \sum_{r,s=1}^{n} a_r \overline{a}_s \gamma(r-s) &= \int_{(-\pi,\pi]} \sum_{r,s=1}^{n} a_r \overline{a}_s \mathrm{e}^{\,i\lambda(r-s)} dF(\lambda) \\ &= \int_{(-\pi,\pi]} \left| \sum_{r=1}^{n} a_r \mathrm{e}^{\,i\lambda r} \right|^2 dF(\lambda) \geq 0 \,. \end{aligned}$$

□

Herglotz's theorem motivates the following definition:

Definition 6.4. (Spectral distribution function of a stationary process)
Suppose that the stationary process (X_t) has an autocovariance function with representation

$$\gamma_X(h) = \int_{(-\pi,\pi]} \mathrm{e}^{\,ih\lambda} dF_X(\lambda) \,, \quad h \in \mathbb{Z} \,, \tag{6.7}$$

where F_X is the right-continuous, non-decreasing, bounded function F_X on $[-\pi, \pi]$ with $F_X(-\pi) = 0$ corresponding to γ_X in Herglotz's theorem. The function F_X is called the spectral distribution function *of the process (X_t), the corresponding measure the* spectral distribution*, and* (6.7) *is the* spectral representation of the autocovariance function γ_X*. Moreover, if F_X is absolutely continuous with respect to Lebesgue measure then the corresponding density function f_X, i.e.,*

$$F_X(\lambda) = \int_{(-\pi,\lambda]} f_X(x) \, dx, \quad \lambda \in (-\pi, \pi] \,,$$

is called the spectral density *of (X_t).*

Example 6.5. (Spectral density of white noise)
Let (X_t) be white noise with autocovariance function

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

Then we immediately see that the corresponding (unique) spectral distribution function is

$$F_X(\lambda) = \frac{\sigma^2}{2\pi} \int_{-\pi}^{\lambda} dx\,, \quad \lambda \in [-\pi, \pi]\,,$$

and the corresponding density is

$$f_X(\lambda) \equiv \frac{\sigma^2}{2\pi}\,, \quad \lambda \in [-\pi, \pi]\,.$$

Thus white noise is characterized by a constant spectral density.

The following very powerful theorem states that any stationary process can be understood as the superposition of a possibly infinite number of random sinusoids, i.e., trigonometric functions with random weights. Before we can formulate this result we still need another notion: the process $(Z_X(\lambda))_{\lambda\in[-\pi,\pi]}$ has *orthogonal increments* if the following conditions are satisfied:
- $EZ_X(\lambda) \equiv 0$.
- $E|Z_X(\lambda)|^2 < \infty$ for all λ.
- $E(Z_X(\lambda_4) - Z_X(\lambda_3))\overline{((Z_X(\lambda_2) - Z_X(\lambda_1))} = 0$ provided $(\lambda_1, \lambda_2] \cap (\lambda_3, \lambda_4] = \emptyset$.

Theorem 6.6. (Spectral representation of a stationary process)
Let (X_t) be a mean–zero stationary process. Then it has representation as a stochastic integral

$$X_t = \int_{(-\pi,\pi]} \mathrm{e}^{\,it\lambda}\, dZ_X(\lambda)\,, \quad t \in \mathbb{Z}\,,$$

where Z_X is a process with orthogonal increments such that

$$F_X(\lambda) = E|Z_X(\lambda) - Z_X(-\pi)|^2\,, \quad \lambda \in [-\pi, \pi]\,.$$

We do not intend to give a rigorous definition of this stochastic integral. It is similar to the definition of an Itô integral and allows for an approximation via discrete random sums:

$$X_t = \int_{(-\pi,\pi]} \mathrm{e}^{\,it\lambda}\, dZ_X(\lambda) \approx \sum_{j=1}^{n} \mathrm{e}^{\,it\lambda_{j-1}}\, (Z_X(\lambda_j) - Z_X(\lambda_{j-1})) \tag{6.8}$$

for $-\pi < \lambda_0 < \cdots < \lambda_n \leq \pi$. Notice that the $(Z_X(\lambda_j) - Z_X(\lambda_{j-1}))$ are uncorrelated. Thus we are in the framework of Example 6.2. The latter example gives the exact stochastic integral representation of the particular stochastic process considered there. It corresponds to a stationary process whose spectral distribution is discrete with a finite number of jumps. In the case of a general stationary process, the stochastic integral representation is achieved by letting the mesh of the partition of the λ_j's go to zero, i.e., the number of λ_j's becomes dense in $[-\pi, \pi]$. In contrast to a Lebesgue-Stieltjes integral, the limits are not defined in a pathwise sense, i.e., for a fixed sample path of Z_X, but the limit has to be taken in the space $L^2[\Omega, \mathcal{F}, P]$. We refer to Brockwell and Davis [8], Chapter 4, for an exact definition of the stochastic integral.

From (6.8) it is intuitively clear that the influence of the trigonometric function $\mathrm{e}^{\,it\lambda_{j-1}}$ on X_t is the bigger the larger the random coefficient $Z_X(\lambda_j) - Z_X(\lambda_{j-1})$. A measure for the order of magnitude of the latter is given by the quantity

$$E|Z_X(\lambda_j) - Z_X(\lambda_{j-1})|^2 = F_X(\lambda_j) - F_X(\lambda_{j-1})\,.$$

If a (sufficiently regular) spectral density f_X exists, the latter difference can be approximated by

$$f_X(\lambda_{j-1})\, (\lambda_j - \lambda_{j-1})\,.$$

Now assume that f_X has one significant peak at λ_{j-1}. In view of the discussion above we may expect that (X_t) is essentially determined by one term in its spectral representation:

$$X_t \approx \mathrm{e}^{\,it\lambda_{j-1}} \left(Z_X(\lambda_j) - Z_X(\lambda_{j-1})\right) , \quad t \in \mathbb{Z} .$$

This means that X_t is essentially determined by one trigonometric function $\mathrm{e}^{\,it\lambda_{j-1}}$ with random coefficient (amplitude) $Z_X(\lambda_j) - Z_X(\lambda_{j-1})$ such that $E|Z_X(\lambda_j) - Z_X(\lambda_{j-1})|^2 \approx f_X(\lambda_{j-1})\,(\lambda_j - \lambda_{j-1})$. Since $\mathrm{e}^{\,it\lambda_{j-1}} = \mathrm{e}^{\,it2\pi[\lambda_{j-1}/(2\pi)]}$ is periodic we may expect that the X_t's have a big value roughly once in $2\pi/\lambda_{j-1}$ units of time.

What has been said about the "largest peak" of the spectral density translates in a similar fashion to the second, third,... largest peak of the density, so that X_t can indeed be understood as a superposition of trigonometric functions with random amplitudes. Clearly, if the density does not have "clear peaks" this means that all trigonometric functions $\exp\{i\lambda t\}$ have roughly the same influence on X_t. In reality, one typically observes time series which are the superposition of many trigonometric functions with a few "leading" trigonometric functions.

Example 6.7. (Wölfer sunspot numbers)
The Wölfer sunspot numbers, see Figure 6.8, is a famous time series which can be found in any textbook on time series analysis as well as in S+ and R. The data are annual averages from 1749 until 1976 (228 years) of observed sunspots. It is obvious that there is roughly a cycle of ten years where the maximum number of spots is achieved. The estimated spectral density is given in the right graph of Figure 6.8. The plotting positions are the *Fourier frequencies* $2\pi j/228, j = 1, \ldots, 114$. A sharp peak can be observed at $\approx 2\pi 0.1$ which corresponds to a $1/0.1 = 10$–year cycle.

Next we consider a method to calculate the spectral density of a stationary process from the autocovariances $\gamma_X(h)$. We start with an auxiliary result:

Proposition 6.10. *Let $(K(n))$ be a sequence of real numbers which is absolutely summable, i.e.,*

$$\sum_{n=-\infty}^{\infty} |K(n)| < \infty . \tag{6.9}$$

Then

$$K(h) = \int_{-\pi}^{\pi} \mathrm{e}^{\,ihx} f(x)\, dx , \quad h \in \mathbb{Z} ,$$

where

$$f(\lambda) = \frac{1}{2\pi} \sum_{n=-\infty}^{\infty} \mathrm{e}^{\,-in\lambda} K(n) .$$

Proof. Notice that

$$\int_{-\pi}^{\pi} \mathrm{e}^{\,ihx} dx = \begin{cases} 2\pi & h = 0 , \\ 0 & h \neq 0 . \end{cases} \tag{6.10}$$

Having this in mind we conclude from (6.9) and a domination argument, ensuring the interchange of infinite series and integral, that

$$\begin{aligned} \int_{-\pi}^{\pi} \mathrm{e}^{\,ihx} f(x) dx &= \int_{-\pi}^{\pi} \mathrm{e}^{\,ihx} \frac{1}{2\pi} \left(\sum_{n=-\infty}^{\infty} \mathrm{e}^{\,-inx} K(n) \right) dx \\ &= \frac{1}{2\pi} \sum_{n=-\infty}^{\infty} K(n) \int_{-\pi}^{\pi} \mathrm{e}^{\,i(h-n)x} dx = K(h) . \end{aligned}$$

□

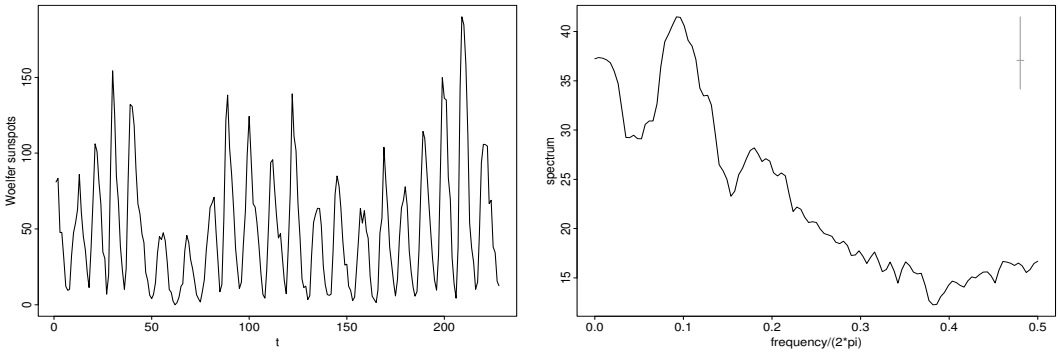

Figure 6.8. Left: *The Wölfer sunspot numbers. An eyeball inspection indicates that extremes of the time series have roughly a cycle of* 10 *years.* Right: *Estimated spectral log-density of the Wölfer sunspot numbers. The vertical line in the right upper corner is the width of a* 95% *asymptotic confidence band which applies uniformly at all frequencies. Note: In S+ and R not the frequencies* λ *are indicated on the x-axis, but the values* $\lambda/(2\pi)$ *for* $\lambda \in [0, \pi]$.

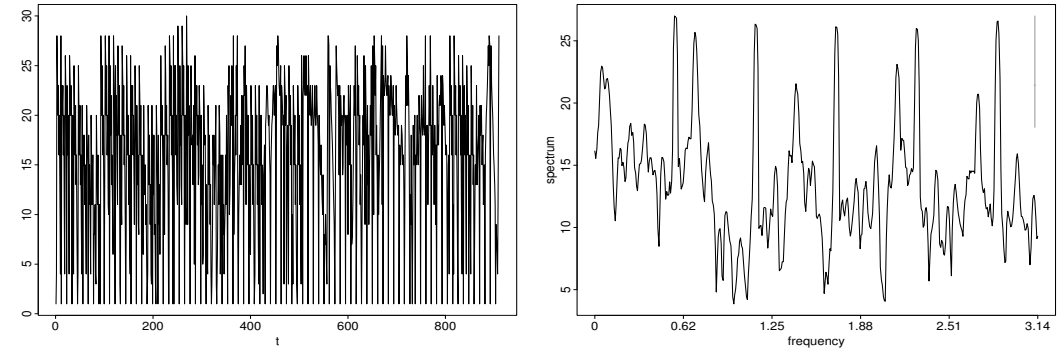

Figure 6.9. Left: *J.S. Bach's Suite No 4 für Violoncello. The notes roughly correspond to the logarithms of the real frequencies played by the instrument.* Right: *Estimated spectral log-density. Not completely surprising, the series contains various cycles with almost the same influence on the time series.*

Herglotz's theorem tells us that the autocovariance function of a stationary process can be calculated from the spectral density. The following result is a partial converse; it tells us that we can calculate the spectral density from an absolutely summable sequence of autocovariances.

Corollary 6.11. *Let* $\gamma : \mathbb{Z} \to \mathbb{C}$ *be such that*

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

The function γ *is the autocovariance function of a stationary process if and only if*

$$f(\lambda) = \frac{1}{2\pi} \sum_{n=-\infty}^{\infty} \mathrm{e}^{\,-in\lambda} \gamma(n) \geq 0 \, , \quad \lambda \in [-\pi, \pi] \, . \tag{6.11}$$

Then f *is the spectral density of the stationary process.*

Proof. We restrict ourselves to the sufficiency part: assume that f with representation (6.11) is non–negative. Since $(\gamma(n))$ is absolutely summable we can apply Proposition 6.10. Hence

$$\gamma(h) = \int_{-\pi}^{\pi} \mathrm{e}^{\,ihx} f(x)dx \;, \quad h \in \mathbb{Z} \;.$$

Notice that the latter can be written as

$$\gamma(h) = \int_{-\pi}^{\pi} \mathrm{e}^{\,ihx} dF(x) \;, \quad h \in \mathbb{Z} \;,$$

where

$$F(\lambda) = \int_{-\pi}^{\lambda} f(x)dx \;, \quad \lambda \in [-\pi, \pi] \;.$$

The latter is a spectral distribution function. Herglotz's theorem implies that γ is an autocovariance function. This proves the corollary. □

Exercise 6.12. Show that, under the conditions of Corollary 6.11 the spectral density f of a real-valued time series (X_t) satisfies $f(\lambda) = f(-\lambda)$. Conclude that it suffices to calculate/estimate the spectral density/distribution only on $[0, \pi]$. This is the reason why S+ and R give the estimated spectral density only on $[0, \pi]$.

6.3. **The spectral density of an ARMA process.** Recall the following fact from Proposition 4.9, adjusted to the case of a complex-valued stationary process: if (Y_t) is stationary with autocovariance function γ_Y and the coefficients (ψ_j) are absolutely summable then $X_t = \sum_{j=0}^{\infty} \psi_j Y_{t-j}, t \in \mathbb{Z}$, is again stationary with autocovariance function

$$\gamma_X(h) = \sum_{j,k=0}^{\infty} \psi_j \overline{\psi}_k \gamma_Y(h - j + k) \;. \tag{6.12}$$

A similar transformation result holds for the spectral distribution functions of the processes (X_t) and (Y_t):

Theorem 6.13. *Suppose (Y_t) is stationary and has spectral distribution function F_Y. Assume that the real coefficients (ψ_j) are absolutely summable. Then the linear process*

$$X_t = \sum_{j=0}^{\infty} \psi_j Y_{t-j} \;, \quad t \in \mathbb{Z} \;, \tag{6.13}$$

is stationary with spectral distribution function

$$F_X(\lambda) = \int_{(-\pi,\lambda]} \left| \sum_{j=0}^{\infty} \psi_j \mathrm{e}^{\,-ijx} \right|^2 dF_Y(x) \;.$$

The function $\psi(\mathrm{e}^{\,-i\lambda}) = \sum_{j=0}^{\infty} \psi_j \mathrm{e}^{\,-ij\lambda}$ is called the *transfer function of the linear filter* (ψ_j), and $|\psi(\mathrm{e}^{\,-i\lambda})|^2$ is the *power transfer function.* We will see later that the power transfer function is a crucial part of the spectral density of a linear process.

Proof. The stationarity of (X_t) follows from Proposition 4.9.

Using Herglotz's theorem, we obtain from (6.12) the following:

$$\begin{aligned}
\gamma_X(h) &= \sum_{j,k=0}^{\infty} \psi_j \overline{\psi}_k \gamma_Y(h-j+k) \\
&= \sum_{j,k=0}^{\infty} \psi_j \overline{\psi}_k \int_{(-\pi,\pi]} \mathrm{e}^{\,i(h-j+k)x} dF_Y(x) \\
&= \int_{(-\pi,\pi]} \left(\sum_{j=0}^{\infty} \psi_j \mathrm{e}^{\,-ijx} \right) \left(\sum_{k=0}^{\infty} \overline{\psi}_k \mathrm{e}^{\,ikx} \right) \mathrm{e}^{\,ihx} dF_Y(x) \\
&= \int_{(-\pi,\pi]} \mathrm{e}^{\,ihx} \left| \sum_{j=0}^{\infty} \psi_j \mathrm{e}^{\,-ijx} \right|^2 dF_Y(x) \,.
\end{aligned}$$

Here we needed the absolute summability of (ψ_j) for interchanging summation and integral. An application of Herglotz's theorem completes the proof. □

Now assume that (X_t) is a causal ARMA process satisfying the ARMA equations

$$\phi(B)X_t = \theta(B)Z_t \,, \quad t \in \mathbb{Z} \,,\ \phi(z) \neq 0 \,,\ |z| \leq 1 \,,$$

where (Z_t) is white noise with variance σ^2. We know that (X_t) has representation (6.13) as a linear process, where (ψ_j) is determined by the equation $\psi(z) = \theta(z)/\phi(z)$, $|z| \leq 1$. Now we may apply Theorem 6.13: recall that white noise has spectral density $f_Z \equiv \sigma^2/(2\pi)$; see Example 6.5. Hence a causal ARMA process has spectral distribution function

$$\begin{aligned}
F_X(\lambda) &= \int_{(-\pi,\lambda]} |\psi(\mathrm{e}^{\,-ix})|^2 dF_Z(x) = \int_{(-\pi,\lambda]} |\psi(\mathrm{e}^{\,-ix})|^2 \, \frac{\sigma^2}{2\pi} \, dx \\
&= \frac{\sigma^2}{2\pi} \int_{(-\pi,\lambda]} \frac{|\theta(\mathrm{e}^{\,-ix})|^2}{|\phi(\mathrm{e}^{\,-ix})|^2} \, dx \,.
\end{aligned}$$

An immediate consequence of the latter formula is

Theorem 6.14. (Spectral density of an ARMA process)
A causal ARMA process (X_t) has spectral density

$$f_X(\lambda) = \frac{\sigma^2}{2\pi} \, \frac{|\theta(\mathrm{e}^{\,-i\lambda})|^2}{|\phi(\mathrm{e}^{\,-i\lambda})|^2} \,, \quad \lambda \in [-\pi, \pi] \,.$$

Notice that we got two different representations of the spectral density of an ARMA process. On the one hand, we have the representation via the coefficients (ψ_j) in Theorem 6.14. On the other hand, we have from (6.11) representation

$$f_X(\lambda) \equiv \frac{1}{2\pi} \sum_{n=-\infty}^{\infty} \mathrm{e}^{\,-in\lambda} \gamma_X(n) \,.$$

Example 6.16. (Spectral density of an MA(1) process)
Consider the MA(1) process $X_t = Z_t + \theta Z_{t-1}$. It has spectral density

$$f_X(\lambda) = \frac{\sigma^2}{2\pi} \left| 1 + \theta \mathrm{e}^{\,-i\lambda} \right|^2 = \frac{\sigma^2}{2\pi} (1 + 2\theta \cos\lambda + \theta^2) \,.$$

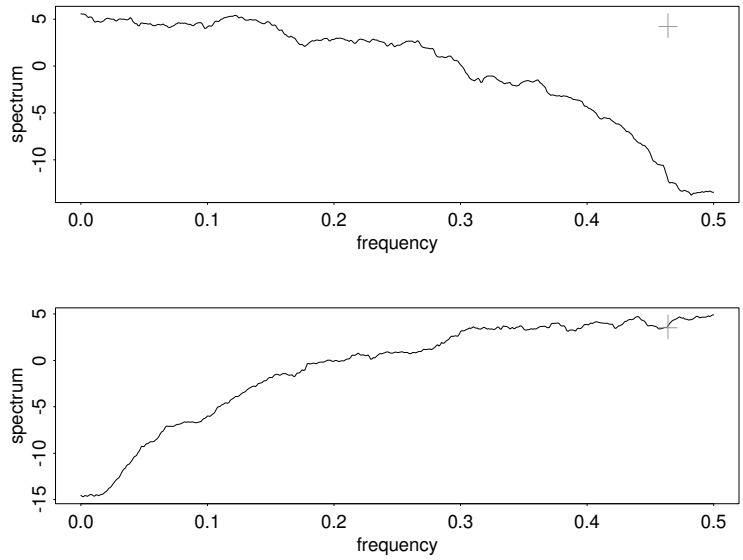

Figure 6.15. *Estimated spectral log-density of an* MA(1) *process with* $\theta = 0.8$ (top) *and* $\theta = -0.8$ (bottom). *The* λ*'s on the* x*-axis correspond to the frequency* $2\pi\lambda$.

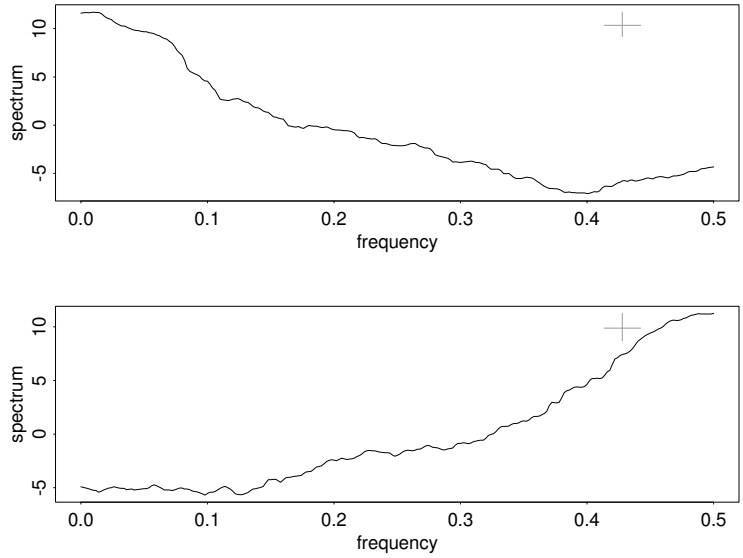

Figure 6.17. *Estimated spectral log-density of an* AR(1) *process with* $\phi = -0.8$ (top) *and* $\theta = 0.8$ (bottom). *The* λ*'s on the* x*-axis correspond to the frequency* $2\pi\lambda$.

Example 6.18. (Spectral density of an AR(1) process)
Consider the AR(1) process $X_t = \phi X_{t-1} + Z_t$ with $|\phi| < 1$. It has spectral density

$$f_X(\lambda) = \frac{\sigma^2}{2\pi}\left|1 - \phi \mathrm{e}^{-i\lambda}\right|^{-2} = \frac{\sigma^2}{2\pi}(1 - 2\phi\cos\lambda + \phi^2)^{-1} .$$

6.4. **Estimation of the spectral density.** In this section we study some statistical estimators of the spectral density. Recall from Corollary 6.11 that the spectral density of a stationary process

(X_t) has representation

$$f_X(\lambda) \equiv \frac{1}{2\pi} \sum_{n=-\infty}^{\infty} \mathrm{e}^{-in\lambda} \gamma_X(n)$$

provided the autocovariances $\gamma_X(h)$ are absolutely summable. Thus it is natural to replace the autocovariances $\gamma_X(h)$ by their sample versions and to get an estimator of f_X in this way. For simplicity, we will assume that (X_t) is already centered, i.e., $EX_t = 0$. In that case it makes sense to modify the sample autocovariances as follows:

$$\widetilde{\gamma}_{n,X}(h) = \begin{cases} n^{-1} \sum_{t=1}^{n-|h|} X_t X_{t+|h|} & \text{if } |h| < n \,, \\ 0 & \text{otherwise} \end{cases}$$

This means we consider the X_t's instead of the $X_t - \overline{X}_n$'s in the definition of the sample autocovariances. It will turn out soon that this is not really a restriction. Thus a natural estimator of $f_X(\lambda)$ is given by

$$I_{n,X}(\lambda) = \frac{1}{2\pi} \sum_{|h|<n} \mathrm{e}^{-ih\lambda} \widetilde{\gamma}_{n,X}(h) \,, \quad \lambda \in [-\pi, \pi] \,.$$

The so defined statistic is called the (raw) *periodogram* of the sample $X_1, \ldots, X_n$. Notice that we can write

$$\begin{aligned} I_{n,X}(\lambda) &= \frac{1}{2\pi} n^{-1} \sum_{t=1}^{n} \sum_{s=1}^{n} X_t X_s \mathrm{e}^{-i\lambda(t-s)} \\ &= \frac{1}{2\pi} \left| n^{-1/2} \sum_{t=1}^{n} \mathrm{e}^{-i\lambda t} X_t \right|^2 . \end{aligned}$$

The periodogram is usually evaluated at the *Fourier frequencies*

$$\lambda_j = 2\pi j/n \,, \quad 0 < j < n/2 \,.$$

Since $\sum_{t=1}^{n} \mathrm{e}^{i\lambda_j t} = 0$ for $\lambda_j \neq 0$, the periodogram ordinates $I_{n,X}(\lambda_j)$ of the sample $X_1, \ldots, X_n$ at the Fourier frequencies have the same value as the periodogram of the sequence $X_1 - m, \ldots, X_n - m$ for any constant m. This is another indication that the centering of the X_t's is not essential. In particular, we have at the Fourier frequencies λ_j,

$$\begin{aligned} I_{n,X}(\lambda_j) &= \frac{1}{2\pi} \left| n^{-1/2} \sum_{t=1}^{n} \mathrm{e}^{-i\lambda_j t} (X_t - \overline{X}_n) \right|^2 \\ &= \frac{1}{2\pi} \sum_{|h|<n} \mathrm{e}^{-ih\lambda_j} \gamma_{n,X}(h) \,, \end{aligned}$$

which is the direct sample analog of the spectral density.

In order to get an impression of the properties of the periodogram we consider some particular case of (X_t):

Example 6.20. (Periodogram of Gaussian white noise)
Suppose that (Z_t) is iid Gaussian white noise with variance σ^2. We consider the periodogram at the Fourier frequencies λ_j. Since

$$I_{n,Z}(\lambda_j) = \frac{1}{2\pi} \left| n^{-1/2} \sum_{t=1}^{n} \mathrm{e}^{-i\lambda_j t} Z_t \right|^2 ,$$

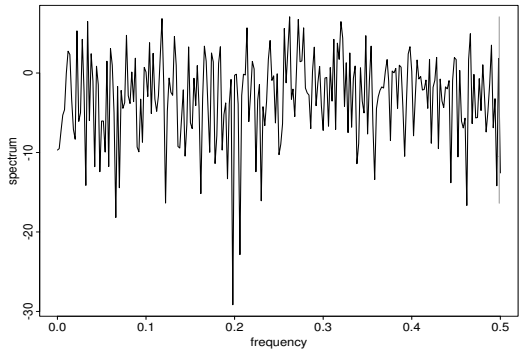

Figure 6.19. *The raw log-periodogram of iid Gaussian white noise at the Fourier frequencies. The x-axis has to be scaled by 2π in order to get the usual frequencies. The erratic behavior of the periodogram indicates that it is not a consistent estimator of the spectral density.*

its distributional properties are determined by the (complex–valued) Gaussian random variables

$$n^{-1/2} \sum_{t=1}^{n} \mathrm{e}^{\,-i\lambda_j t} Z_t \; . \tag{6.14}$$

For $1 \le j \le k < n/2$ we observe that

$$\begin{aligned} & E \left(n^{-1/2} \sum_{t=1}^{n} \mathrm{e}^{\,-i\lambda_j t} Z_t \right) \left(n^{-1/2} \sum_{s=1}^{n} \mathrm{e}^{\,i\lambda_k s} Z_s \right) \\ = \;& n^{-1} \sum_{t=1}^{n} E Z_1^2 \, \mathrm{e}^{\,-i(\lambda_j - \lambda_k)t} \\ = \;& \sigma^2 \begin{cases} 1 & j = k \; , \\ \dfrac{\mathrm{e}^{\,i(n+1)(\lambda_j - \lambda_k)} - 1}{\mathrm{e}^{\,i(\lambda_j - \lambda_k)} - 1} - 1 = \dfrac{\mathrm{e}^{\,i2\pi(j-k)/n} - 1}{\mathrm{e}^{\,i2\pi(j-k)/n} - 1} - 1 = 0 & j \ne k \; . \end{cases} \end{aligned}$$

Since the random variables (6.14) are mean–zero uncorrelated Gaussian and have the same second moment we may conclude that they must be independent and identically distributed. In particular, the sequence $(I_{n,Z}(\lambda_j))_{1 \le j < n/2}$ is iid. In order to determine the distribution of one $I_{n,Z}(\lambda_j)$ we observe that $n^{-1/2} \sum_{t=1}^{n} Z_t \cos(\lambda_j t)$ and $n^{-1/2} \sum_{t=1}^{n} Z_t \sin(\lambda_j t)$ have mean zero and are uncorrelated, hence they are independent Gaussian, each with variance $\sigma^2/2$. Moreover,

$$I_{n,Z}(\lambda_j) = \frac{1}{2\pi} \left(n^{-1/2} \sum_{t=1}^{n} Z_t \cos(\lambda_j t) \right)^2 + \frac{1}{2\pi} \left(n^{-1/2} \sum_{t=1}^{n} Z_t \sin(\lambda_j t) \right)^2 \; . \tag{6.15}$$

Hence $2\pi I_{n,Z}(\lambda_j)$ has the same distribution as the sum of the squares of two $\mathrm{N}(0, \sigma^2/2)$ random variables, i.e., a $\sigma^2 \chi^2_2/2$ or a $\sigma^2 \mathrm{Exp}(1)$ distribution. Thus we conclude that the periodogram ordinates of an iid standard Gaussian sequence (Z_t) at the Fourier frequencies are iid exponential with mean $\sigma^2/(2\pi)$. We conclude in particular that the periodogram is certainly not a consistent estimator of the (in this case constant) spectral density $\sigma^2/(2\pi)$.

The iid property of the periodogram at the Fourier frequencies has been used by R.A. Fisher to construct his g–test for Gaussian white noise. He proposed the test statistic

$$g_n = \max_{1\le j\le q} I(\lambda_j) \Big/ \left(q^{-1} \sum_{j=1}^{q} I(\lambda_j) \right)$$

with $q = [n/2]$. Its asymptotic distribution can be calculated under the null hypothesis that (Z_t) is iid Gaussian; see e.g. Brockwell and Davis (1991), p. 339. Notice that $q^{-1} \sum_{j=1}^{q} I(\lambda_j) \stackrel{\text{a.s.}}{\to} \sigma^2$.

Though the periodogram is not a consistent estimator of f_X it is not too far away from consistency:

Proposition 6.21. *Assume (X_t) is mean–zero stationary with an absolutely summable autocovariance function. Then, for $\lambda \in (0, \pi]$, $EI_{n,X}(\lambda) \to f_X(\lambda)$.*

Proof. Since $EX_t = 0$ we have

$$\begin{aligned} EI_{n,X}(\lambda) &= \frac{1}{2\pi} \sum_{|h|<n} \mathrm{e}^{-ih\lambda} E\widetilde{\gamma}_{n,X}(h) \\ &= \frac{1}{2\pi} \sum_{|h|<n} \mathrm{e}^{-ih\lambda} \left(n^{-1} \sum_{t=1}^{n-|h|} EX_t X_{t+|h|} \right) \\ &= \frac{1}{2\pi} \sum_{|h|<n} \left(1 - \frac{|h|}{n} \right) \mathrm{e}^{-ih\lambda} \gamma_X(h) \\ &\to \frac{1}{2\pi} \sum_{|h|<\infty} \mathrm{e}^{-ih\lambda} \gamma_X(h) = f_X(\lambda) \, . \end{aligned}$$

In the proof we used the absolute summability of $(\gamma_X(h))$ to show that the latter limit is well defined. □

Now we study the asymptotic distribution of the periodogram:

Theorem 6.22. (Asymptotic properties of the periodogram)
Assume that (X_t) is a linear process with absolutely summable coefficients (ψ_j) which is driven by iid white noise with variance σ^2. Let $0 < \omega_1 < \cdots < \omega_m < \pi$ be fixed frequencies. Then

$$(I_{n,X}(\omega_j))_{j=1,\ldots,m} \stackrel{d}{\to} \left(\frac{\sigma^2}{2\pi} \, |\psi(\mathrm{e}^{-i\omega_j})|^2 \, E_j \right)_{j=1,\ldots,m} = (f_X(\omega_j) E_j)_{j=1,\ldots,m}$$

for iid standard exponential random variables (E_j).

The limit distribution depends on the spectral density at a given frequency. In order to avoid this it is common to plot the logarithm of the estimated spectral density. Then the confidence bands become independent of the frequency.

Notice the similarities of the limit distribution for the general linear process and for the iid Gaussian white noise sequence from Example 6.20. The basic ideas of the proof are the following ones:
1) First one can show that

$$I_{n,X}(\lambda) = I_{n,Z}(\lambda) \, |\psi(\mathrm{e}^{-i\lambda})|^2 + o_P(1) \, .$$

Thus the limit distribution of $I_{n,X}(\lambda)$ is determined by the one of $I_{n,Z}(\lambda)$ for iid white noise.
2) One uses the decomposition

$$2\pi I_{n,Z}(\omega_j) \quad = \quad \left(n^{-1/2} \sum_{t=1}^{n} Z_t \cos(\omega_j t) \right)^2 + \left(n^{-1/2} \sum_{t=1}^{n} Z_t \sin(\omega_j t) \right)^2 .$$

Then one can show a two–dimensional CLT for

$$n^{-1/2} \left(\sum_{t=1}^{n} Z_t \cos(\omega_j t), \sum_{t=1}^{n} Z_t \sin(\omega_j t) \right)$$

using the Cramér–Wold device and the Lindeberg–Feller CLT. The limit consists of two iid $N(0, 0.5\sigma^2)$ random variables whose sum of squares gives an exponential distribution.

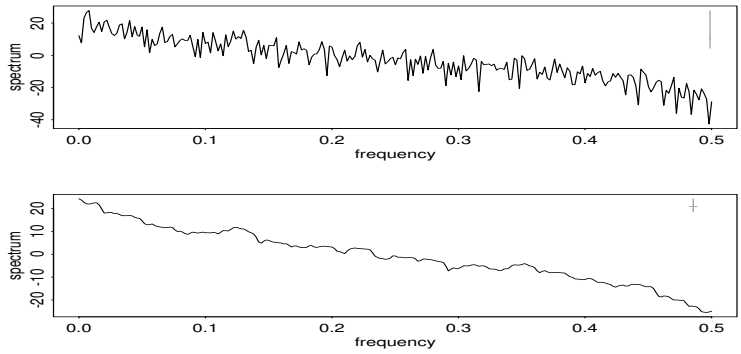

Figure 6.23. *Raw log-periodogram* (left) *and smoothed log-periodogram* (right) *for a time series from the* ARMA$(1,1)$ *model* $X_t - 0.9X_{t-1} = Z_t + 0.9Z_{t-1}$.

Although the periodogram is not a consistent estimator of $f_X(\lambda)$, it is not far away from consistency. Observe that

$$I_{n,X}(\omega_j) \approx f_X(\omega_j) E_j$$

for iid standard exponential random variables E_j. Note that $T_j = E_j - 1$ is mean–zero with very light (exponential) tails. The following idea can be made to work: assume that there are given non–negative weights $(W_n(k))_{|k| \le m}$ such that

$$W_n(k) = W_n(-k) \quad \text{for} \quad |k| \le m \,,$$

$$\sum_{|k| \le m} W_n(k) = 1 \quad \text{and} \quad \sum_{|k| \le m} W_n^2(k) \to 0 \,.$$

The integers $m = m_n$ have to be chosen such that $m_n \to \infty$ and $m_n/n \to 0$. We apply the following heuristic argument:

$$\sum_{|k| \le m} W_n(k) I_{n,X}(\lambda + 2\pi k/n) \approx \sum_{|k| \le m} W_n(k) f_X(\lambda + 2\pi k/n)\ E_k$$

for iid standard exponential E_k's. If f_X is continuous at λ then $f_X(\lambda + 2\pi k/n) \sim f_X(\lambda)$. Hence

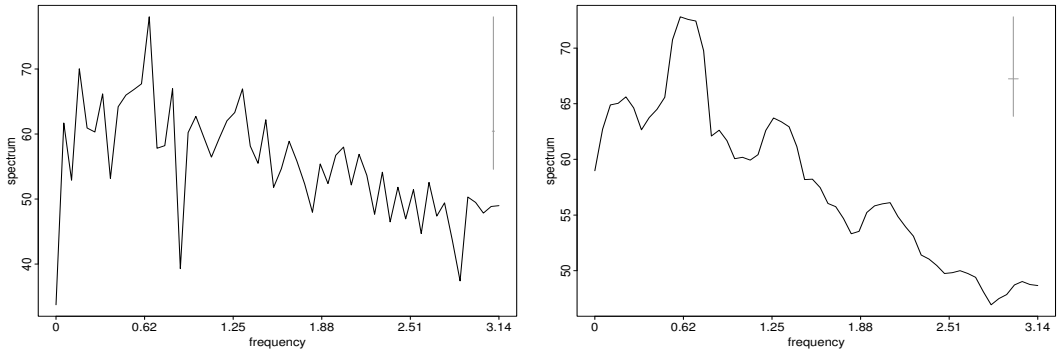

Figure 6.24. *The raw* (top) *and the smoothed* (bottom) *log-periodograms of the Canadian lynx data; see Figure 4.30. There is a clear peak about frequency* $2\pi * 0.1$ *indicating a cycle of about* 10 *years.*

$$\begin{aligned}\sum_{|k|\le m} W_n(k) I_{n,X}(\lambda + 2\pi k/n) &\approx f_X(\lambda) \sum_{|k|\le m} W_n(k)(1+T_k) \\ &= f_X(\lambda)\left(1 + \sum_{|k|\le m} W_n(k) T_k\right) \\ &= f_X(\lambda) + o_P(1)\,,\end{aligned}$$

since

$$\operatorname{var}\left(\sum_{|k|\le m} W_n(k) T_k\right) = \sum_{|k|\le m} W_n^2(k) \to 0\,.$$

The simplest weights we can choose are of the form $W_n(k) = 1/(2m+1)$. They are called the *Daniell weights*.

We have only indicated the rough idea of the use of smoothing techniques for estimating the spectral density. There exists indeed a whole industry for estimating spectral densities. We refer to Priestley [29] for an intensive discussion of various estimators of the spectral density including the discrete weighted averages discussed above, lag window estimators and kernel density estimators. A useful overview is also given in Brockwell and Davis [8], Section 10.4.

7. Prediction of time series

7.1. **The projection theorem in Hilbert space.** Recall the notion of a Hilbert space: it is a normed vector space with inner product product (x, y) and with a norm defined by $\|x\| = \sqrt{(x, x)}$. Moreover, it is complete, i.e., every Cauchy sequence in H has a limit in H.

The following theorem is the basis for the prediction of time series:

Theorem 7.1. (Projection theorem)
Assume that M is a closed subspace of the Hilbert space H and $x \in H$.
1) There exists a unique element $\widehat{x} \in H$ such that

$$\|x - \widehat{x}\| = \inf_{y \in M} \|x - y\| \,. \tag{7.1}$$

2) (7.1) *holds if and only if $\widehat{x} \in M$ and $x - \widehat{x}$ is orthogonal to M, i.e., for every $y \in M$, $(x - \widehat{x}, y) = 0$.*

Notice that the projection theorem is close to the Pythagorean theorem. The projection theorem also offers a way how to calculate $\widehat{x}$. Indeed, 2) above is equivalent to $(x - \widehat{x}, y) = 0$ for all elements $y \in M$ which span the subspace M. We will make intensive use of that property. The element $\widehat{x}$ is usually called the *orthogonal projection of x on M*.

In what follows, we will exploit special cases of the projection theorem.

7.2. **Linear prediction of time series.** We assume that (X_t) is a stationary process with mean μ. Given we observed $X_1, \ldots, X_n$, we want to predict X_{n+h} by a linear combination of these observations:

$$P_n X_{n+h} = a_0 + a_1 X_n + \cdots + a_n X_1. \tag{7.2}$$

We want to choose the weights (a_j) in such a way that the mean squared error $E[(X_{n+h} - P_n X_{n+h})^2]$ is minimized.

Proposition 7.2. (Best linear h-step prediction)
The minimum mean squared error linear h–step predictor $P_n X_{n+h}$ is determined by the equation

$$\Gamma_n \mathbf{a}_n = \gamma_n(h) \,,$$

and $a_0 = \mu\big(1 - \sum_{i=1}^n a_i\big)$, where

$$\mathbf{a}_n = (a_1, \ldots, a_n)' \,, \quad \Gamma_n = (\gamma_X(i-j))_{i,j=1,\ldots,n}$$

and

$$\gamma_n(h) = (\gamma_X(h), \ldots, \gamma_X(h+n-1))' \,.$$

The h-step predictor is then given by

$$P_n X_{n+h} = \mu + \sum_{i=1}^n a_i (X_{n+1-i} - \mu) \,.$$

Proof. We have to minimize the function

$$f(\mathbf{a}_n) = E\big[(X_{n+h} - (a_0 + a_1 X_n + \cdots + a_n X_1))^2\big] \,. \tag{7.3}$$

We know that for any square integrable random variable Y and any constant c we have $E[(Y-c)^2] \geq \mathrm{var}(Y)$ with equality if and only if $c = EY$. Hence we necessarily have the relation

$$E\big[X_{n+h} - (a_0 + a_1 X_n + \cdots + a_n X_1)\big] = 0 \,,$$

and

$$a_0 = \mu\big(1 - \sum_{i=1}^n a_i\big) \,.$$

From now on we may assume without loss of generality that (X_t) has mean zero and that $a_0 = 0$. A necessary condition for the minimization of $f(\mathbf{a}_n)$ is then

$$\frac{\partial f(\mathbf{a}_n)}{\partial a_i} = 2E\big[\big(X_{n+h} - \sum_{j=1}^{n} a_j X_{n+1-j}\big) X_{n+1-i}\big] = 0\,, \qquad i = 1, \dots, n\,.$$

In matrix form this system of equations is just $\Gamma_n \mathbf{a}_n = \gamma_n(h)$ as desired. □

Exercise 7.3. (a) Show that the prediction error is given by

$$E\big[(X_{n+h} - P_n X_{n+h})^2\big] = \gamma_X(0) - \mathbf{a}_n' \gamma_n(h)\,. \tag{7.4}$$

(b) Show that $P_n X_{n+h}$ is unique, i.e., if there are two solutions $\mathbf{a}_n^{(1)}$ and $\mathbf{a}_n^{(2)}$ to the system of equations $\Gamma_n \mathbf{a}_n = \gamma_n(h)$ then the random variable

$$Z = a_0^{(1)} - a_0^{(2)} + \sum_{j=1}^{n} (a_j^{(1)} - a_j^{(2)})\, X_{n+1-j}$$

is zero a.s.
(c) Show that

$$E[(X_{n+h} - P_n X_{n+h}) X_j] = 0\,, \qquad j = 1, \dots, n\,. \tag{7.5}$$

(d) Consider the Hilbert space of all linear combinations $a_0 + \sum_{i=1}^{n} a_i X_{n+1-i}$ equipped with the inner product $(X, Y) = \mathrm{cov}(X, Y)$. Show that the unique linear h-step predictor $P_n X_{n+h}$ can be derived from the projection theorem in Hilbert space by solving (7.5). (Note: one can always assume without loss of generality that $a_0 = 0$ and $\mu = 0$.)

Remark 7.4. If Γ_n is invertible we have $\mathbf{a}_n = \Gamma_n^{-1} \gamma_n$. In that case, $\mathbf{a}_n$ is unique. If Γ_n is singular then there exist infinitely many solutions $\mathbf{a}_n$. However, all of them represent the same predictor $P_n X_{n+h}$ since it is unique.

Example 7.5. (Linear 1-step prediction of AR processes)
From p. 45 recall the Yule-Walker estimator of a causal AR(p) process given by the equations

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = Z_t\,, \quad t \in \mathbb{Z}\,,$$

driven by white noise with variance σ^2. The Yule-Walker equations for the parameter vector $\phi = (\phi_1, \dots, \phi_p)'$ and of σ^2 were given by the $p + 1$ linear equations

$$\begin{aligned} \sigma^2 &= \gamma_X(0) - \phi' \gamma_p(1)\,, \\ \Gamma_p\, \phi &= \gamma_p(1)\,. \end{aligned}$$

If we identify the vector ϕ with $\mathbf{a}_n = (a_1, \dots, a_p)'$ we see that the second equation is nothing but the 1-step prediction equation. In other words, Yule-Walker estimation is based on the idea of minimum least squares prediction of X_{n+1} by linear combinations of $1, X_n, \dots, X_{n+1-p}$. In this sense, the Yule-Walker estimators are least squares estimators.

We also notice that the first equation for σ^2 is identical with the mean squares 1-step prediction error given in (7.4).

Example 7.6. (A deterministic time series)
We consider the stationary process (X_t) from Example 2.11: $X_t = A\cos(\omega t) + B\sin(\omega t)$, where A, B are random variables such that $EA = EB = E(AB) = 0$ and $E[A^2] = E[B^2] = 1$, $\omega \in (0, \pi)$. It has autocovariance function $\gamma_X(h) = \cos(\omega h)$. Notice that

$$X_{n+1} = (2\cos\omega) X_n - X_{n-1}\,, \qquad n \in \mathbb{Z}\,.$$

Since the linear prediction $P_n X_{n+1}$ of X_{n+1} based on $X_1, \dots, X_n$ is unique we have

$$P_n X_{n+1} = X_{n+1}\,, \qquad n \in \mathbb{Z}\,.$$

This means that X_{n+1} is completely predictable given its past. This property is shared by a whole class of time series, the *deterministic time series*: given their entire past, their future is completely determined by these values.

The *Wold decomposition* of a stationary time series shows that every stationary process (X_t) can be decomposed into two stationary processes

$$X_t = Y_t + A_t\,, \qquad t \in \mathbb{Z}\,,$$

where (Y_t) is a causal linear process

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

with respect to some white noise process (Z_t) and (A_t) is a *deterministic process* in the sense that, for any t, A_t can be represented as a (possibly infinite) linear combination of $(A_s)_{s\le t-1}$, i.e., it is completely predictable given its infinite past; see Brockwell and Davis [8], Section 5.7. Deterministic time series models are often considered as unnatural and most often excluded from the analyses.

The solution $\mathbf{a}_n$ to the prediction equation $\Gamma_n \mathbf{a}_n = \gamma_n(h)$ is uniquely determined if Γ_n is invertible, $\mathbf{a}_n = \Gamma_n^{-1}\gamma_n(h)$. We recall from Proposition 3.3 that Γ_n is invertible for every n if $\gamma_X(0) > 0$ and $\gamma_X(h) \to 0$ as $h \to \infty$.

Corollary 7.7. *If $\gamma_X(0) > 0$ and $\gamma_X(h) \to 0$ as $h \to \infty$ then the best linear h-step predictor $P_n X_{n+h}$ of X_{n+h} in terms of $X_1, \ldots, X_n$ is*

$$P_n X_{n+h} = \sum_{j=1}^{n} a_j X_{n+1-j}\,,$$

where $\mathbf{a}_n = \Gamma_n^{-1}\gamma_n(h)$ with prediction error

$$E\big[(P_n X_{n+h} - X_{n+h})^2\big] = \gamma_X(0) - \gamma_n(h)'\Gamma_n^{-1}\gamma_n(h)\,.$$

7.3. **The innovations algorithm.** The innovations algorithm is a recursive method for determining the linear prediction of a stationary process. We introduce the subspaces

$$M_n = \left\{ \sum_{j=1}^{n} a_j X_j \;\; : a_j \in \mathbb{R}\,, j = 1, \ldots, n \right\}$$

of the Hilbert space of square integrable random variables equipped with the inner product $(X, Y) = \mathrm{cov}(X, Y)$. Notice that we also have

$$M_n = \left\{ \sum_{j=1}^{n} a_j (X_j - P_{j-1}X_j) \;\; : a_j \in \mathbb{R}\,, j = 1, \ldots, n \right\}\,,$$

since $P_{j-1}X_j \in M_{j-1}$. The latter representation of M_n has the advantage that $\{X_1 - P_0 X_1, \ldots, X_n - P_{n-1}X_n\}$ are orthogonal since $X_i - P_{i-1}X_i \in M_{j-1}$ for $i < j$ and $X_j - P_{j-1}X_j$ is orthogonal to M_{j-1} by definition of the predictor $P_{j-1}X_j$.

We may conclude that

$$P_n X_{n+1} = \sum_{j=1}^{n} \phi_{nj} X_{n+1-j} = \sum_{j=1}^{n} \theta_{nj}(X_{n+1-j} - P_{n-j}X_{n+1-j})$$

for appropriate constants (θ_{nj}) which we will determine in what follows.

Proposition 7.8. (Innovations algorithm, Brockwell and Davis [8], Section 5.2)
Let (X_t) be mean–zero stationary and such that Γ_n^{-1} exists for all n. Then

$$P_n X_{n+1} = \begin{cases} 0 & n = 0 \,, \\ \sum_{j=1}^{n} \theta_{nj}(X_{n+1-j} - P_{n-j}X_{n+1-j}) & n \geq 1 \,, \end{cases}$$

where the coefficients (θ_{nj}) and the prediction errors (v_n) are determined recursively:

$$\begin{aligned} v_0 &= \gamma_X(0) \,, \\ \theta_{n,n-k} &= v_k^{-1}\left(\gamma_X(n-k) - \sum_{j=0}^{k-1} \theta_{k,k-j}\theta_{n,n-j}v_j\right) \,, \quad k = 0, \ldots, n-1 \,, \\ v_n &= \gamma_X(0) - \sum_{j=0}^{n-1} \theta_{n,n-j}^2 v_j \,. \end{aligned}$$

The innovations algorithm has to be applied recursively in the following order:

$$v_0; \theta_{11}, v_1; \theta_{22}, \theta_{21}, v_2; \theta_{33}, \theta_{32}, \theta_{31}, v_3; \ldots$$

Proof. As mentioned the random variables $X_j - P_{j-1}X_j$ are orthogonal. Thus, for $0 \leq k < n$,

$$\begin{aligned} (P_n X_{n+1}, X_{k+1} - P_k X_{k+1}) &= \left(\sum_{j=1}^{n} \theta_{nj}(X_{n+1-j} - P_{n-j}X_{n+1-j}), X_{k+1} - P_k X_{k+1}\right) \\ &= \theta_{n,n-k} E(X_{k+1} - P_k X_{k+1})^2 = \theta_{n,n-k} v_k \,. \end{aligned}$$

Adding and subtracting X_{n+1} on the left-hand side we obtain

$$v_k^{-1}(X_{n+1}, X_{k+1} - P_k X_{k+1}) = \theta_{n,n-k} \,. \tag{7.6}$$

Now, recalling the definition of $P_k X_{k+1}$, we obtain

$$\begin{aligned} \theta_{n,n-k} &= v_k^{-1}\left(\gamma_X(n-k) - \left(X_{n+1}, \sum_{j=0}^{k-1} \theta_{k,k-j}(X_{j+1} - P_j X_{j+1})\right)\right) \\ &= v_k^{-1}\left(\gamma_X(n-k) - \sum_{j=0}^{k-1} \theta_{k,k-j}(X_{n+1}, X_{j+1} - P_j X_{j+1})\right) \\ &= v_k^{-1}\left(\gamma_X(n-k) - \sum_{j=0}^{k-1} \theta_{k,k-j}\theta_{n,n-j}v_j\right) \,. \end{aligned}$$

In the last step we used (7.6). Thus we got the desired formula for $\theta_{n,n-k}$.
It remains to calculate the prediction error

$$\begin{aligned} v_n &= (X_{n+1} - P_n X_{n+1}, X_{n+1} - P_n X_{n+1}) \\ &= (X_{n+1}, X_{n+1}) - (P_n X_{n+1}, P_n X_{n+1}) \\ &= \gamma_X(0) - \left(\sum_{j=1}^{n} \theta_{nj}(X_{n+1-j} - P_{n-j} X_{n+1-j}), \sum_{j=1}^{n} \theta_{nj}(X_{n+1-j} - P_{n-j} X_{n+1-j}) \right) \\ &= \gamma_X(0) - \sum_{j=1}^{n} \theta_{nj}^2 v_{n-j} \\ &= \gamma_X(0) - \sum_{j=0}^{n-1} \theta_{n,n-j}^2 v_j \ . \end{aligned}$$

Example 7.9. (MA(1) process)
Let $X_t = \theta Z_{t-1} + Z_t$ be an MA(1) process driven by white noise (Z_t). Then $\gamma_X(0) = \sigma^2(1+\theta^2)$, $\gamma_X(1) = \theta\sigma^2$ and $\gamma_X(k) = 0$ for all lags $|k| > 1$. Straightforward calculation yields that

$$\begin{aligned} v_0 &= (1+\theta^2)\sigma^2 \ , \qquad v_n = [1 + \theta^2 - v_{n-1}\theta^2\sigma^2]\sigma^2 \ , \\ \theta_{nj} &= 0 \ , 2 \le j \le n \ , \quad \theta_{n1} = v_{n-1}^{-1}\theta\sigma^2 \ . \end{aligned}$$

We can also write

$$P_n X_{n+1} = \theta(X_n - P_{n-1}X_n)/r_{n-1} \ ,$$

where $r_n = v_n\sigma^2$. It can be shown that $r_n \xrightarrow{P} 1$. Thus $P_n X_{n+1} \sim \theta(X_n - P_{n-1}X_n)$.

An application of the innovations algorithm and some more sophistication yields the following results for an ARMA process.

Proposition 7.10. (Linear prediction of an ARMA process)
Suppose that (X_t) is a causal ARMA(p,q) *process satisfying the equations*

$$\phi(B)X_t = \theta(B)Z_t \, ,$$

and $m = \max(p,q)$. Then

$$P_n X_{n+1} = \begin{cases} \sum_{j=1}^{n} \theta_{nj}(X_{n+1-j} - P_{n-j}X_{n+1-j}) \ , & q \le n < m \ , \\ \\ \phi_1 X_n + \cdots + \phi_p X_{n+1-p} + \sum_{j=1}^{q} \theta_{nj}(X_{n+1-j} - P_{n-j}X_{n+1-j}) \ , & n \ge m \ , \end{cases}$$

and

$$v_n = \sigma^2 r_n \ .$$

Moreover, $r_n \xrightarrow{P} 1$ and $\theta_{nj} \xrightarrow{P} \theta_j$ if (X_t) is also invertible.

This proposition shows that the linear predictor $\widehat{X}_{n+1}$ is basically a "rewritten ARMA equation". It also shows that the prediction error is roughly of the order σ^2 and can therefore not be improved even for large n.

Example 7.11. For an AR(p) process the innovations algorithm yields

$$P_n X_{n+1} = \phi_1 X_n + \cdots + \phi_p X_{n+1-p} \ , \quad n \ge p \ .$$

For an MA(q) process we obtain

$$P_n X_{n+1} = \sum_{j=1}^{q} \theta_{nj}(X_{n+1-j} - P_{n-j}X_{n+1-j}) \,, \quad n \geq q \,.$$

Notice that for large n, the θ_{nj} could be replaced by the θ_j.
For a causal ARMA$(1,1)$ process we obtain

$$P_n X_{n+1} = \phi X_n + \theta_{n1}(X_n - P_{n-1}X_n) \,.$$

7.4. **Some comments on the general prediction problem.** The following lemma is the basis for general prediction in the space of square integrable random variables.

Lemma 7.12. (Best prediction of a random variable given some information) *Let X be a random variable with finite variance and W be a random element on the same probability space. Then $P_W X = E(X \mid W)$ minimizes the quadratic risk $E[(X-Y)^2]$ in the class of all square integrable random variables Y which are measurable functions of W.*

Proof. We have for any Y which is a function of W,

$$\begin{aligned} E[(X-Y)^2] &= E\big[\big((X - E(X \mid W)) + (E(X \mid W) - Y)\big)^2\big] \\ &= E\big[(X - E(X \mid W))^2\big] + E\big[(E(X \mid W) - Y)^2\big] + 2E\big[(X - E(X \mid W))(E(X \mid W) - Y)\big] \\ &= E\big[(X - E(X \mid W))^2\big] + E\big[(E(X \mid W) - Y)^2\big] \\ &\geq E\big[(X - E(X \mid W))^2\big] \end{aligned}$$

with equality if and only if $E(X \mid W) = Y$ a.s. Here we used the facts that both Y and $E(X \mid W)$ are functions of W, and that

$$\begin{aligned} E\big[(X - E(X \mid W))(E(X \mid W) - Y)\big] &= E\Big[E\big[(X - E(X \mid W))(E(X \mid W) - Y) \mid W\big]\Big] \\ &= E\Big[(E(X \mid W) - Y)\, E\big[(X - E(X \mid W)) \mid W\big]\Big] \\ &= E\Big[(E(X \mid W) - Y)\,\big(E(X \mid W) - E(X \mid W)\big)\Big] = 0 \,. \end{aligned}$$

□

Consider a time series (X_t). We may conclude that the quantities (abusing previous notation)

$$P_n X_{n+h} = E(X_{n+h} \mid X_1, \ldots, X_n) \,, \qquad n \geq 1 \,,$$

are the best predictors of X_{n+h} given the sample $X_1, \ldots, X_n$. In many situations, this h-step predictor is difficult to calculate even when considering ARMA processes. An exception is the AR process.

Example 7.13. (Best prediction of a causal AR(p) process given the past)
Consider the causal AR(p) process

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

where (Z_t) is iid white noise. Then for $n > p$,

$$\begin{aligned} E(X_{n+1} \mid X_1, \ldots, X_n) &= E(X_{n+1} \mid X_n, \ldots, X_{n-p+1}) \\ &= \phi_1 X_n + \cdots + \phi_p X_{n-p+1} \,, \\ E(X_{n+2} \mid X_1, \ldots, X_n) &= E(\phi_1 X_{n+1} + \cdots + \phi_p X_{n-p+2} \mid X_n, \ldots, X_{n-p+1}) \\ &= \phi_1\big(\phi_1 X_n + \cdots + \phi_p X_{n-p+1}\big) + \phi_2 X_n + \cdots + \phi_p X_{n-p+2} \,, \\ &\quad \ldots \,. \end{aligned}$$

This means that the best h-step predictor can be calculated as a linear combination of $X_1, \ldots, X_n$. In this case, the linear predictor and the best predictor coincide.

Example 7.14. (Best prediction in a stochastic volatility and GARCH model given the past) One of the popular models in financial time series analysis is the simple Gaussian stochastic volatility model given by

$$X_t = \sigma_t Z_t\,, \qquad t \in \mathbb{Z}\,,$$

where (Z_t) is iid white noise and (σ_t) is a strictly stationary positive volatility sequence independent of (Z_t). Often one assumes that the log-volatility is given by a linear process:

$$\log \sigma_t = \sum_{j=0}^{\infty} \psi_j \eta_{t-j}\,, \qquad t \in \mathbb{Z}\,,$$

where (η_t) is iid standard Gaussian and (ψ_j) a square summable sequence of real numbers. Assume that

$$Y_t = \log \sigma_t = \phi\, Y_{t-1} + \eta_t\,, \qquad t \in \mathbb{Z}\,,$$

for some $|\phi| < 1$. Then the best prediction of X_{n+1} given $X_1, \ldots, X_n$ is

$$E(X_{n+1} \mid X_1, \ldots, X_n) = E(\mathrm{e}^{\,\phi Y_n + \eta_{n+1}} Z_{n+1} \mid X_1, \ldots, X_n) = E(\sigma_n^{\phi} \mid X_1, \ldots, X_n) E(\mathrm{e}^{\,\eta_{n+1}} Z_{n+1}) = 0\,,$$

while

$$\begin{aligned} E(|X_{n+1}| \mid X_1, \ldots, X_n) &= E[|Z_{n+1}|] E[\mathrm{e}^{\,\eta_{n+1}}] E(\mathrm{e}^{\,\phi Y_n} \mid X_1, \ldots, X_n) \\ &= E[|Z_0|] E[\mathrm{e}^{\,\eta_0}] E\big[\sigma_n^{\phi} \mid X_1, \ldots, X_n\big]\,. \end{aligned}$$

The expression on the right-hand side cannot be calculated unless one has additional information about the (unobserved) volatility sequence (σ_t). Indeed,

$$\begin{aligned} E(|X_{n+1}| \mid \sigma_1, \ldots, \sigma_n) &= E[|Z_0|] E[\mathrm{e}^{\,\eta_0}] E(\sigma_n^{\phi} \mid \sigma_n) \\ &= E[|Z_0|] E[\mathrm{e}^{\,\eta_0}] \sigma_n^{\phi}\,. \end{aligned}$$

This is an example of a highly non-linear prediction of X_{n+1} given the past volatilities.

Consider an ARCH(p) model where for positive α_i,

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i}^2\,.$$

If $n > p$ we have

$$\begin{aligned}
E(X_{n+1}^2 \mid X_1, \ldots, X_n) &= E[Z_{n+1}^2] E(\sigma_{n+1}^2 \mid X_1, \ldots, X_n) \\
&= \sigma_{n+1}^2 \,, \\
E(X_{n+2}^2 \mid X_1, \ldots, X_n) &= E(\sigma_{n+2}^2 \mid X_1, \ldots, X_n) \\
&= \alpha_0 + \sum_{i=2}^{p} \alpha_i X_{n+2-i}^2 + \alpha_1 \, E(X_{n+1}^2 \mid X_1, \ldots, X_n) \\
&= \alpha_0 + \sum_{i=2}^{p} \alpha_i X_{n+2-i}^2 + \alpha_1 \, \sigma_{n+1}^2 \\
&= \alpha_0 + \sum_{i=2}^{p} \alpha_i X_{n+2-i}^2 + \alpha_1 \, \sigma_{n+1}^2 \,, \\
E(X_{n+3}^2 \mid X_1, \ldots, X_n) &= E(\sigma_{n+3}^2 \mid X_1, \ldots, X_n) \\
&= \alpha_0 + \sum_{i=3}^{p} \alpha_i X_{n+3-i}^2 + \alpha_1 \, E(X_{n+2}^2 \mid X_1, \ldots, X_n) \\
&\quad + \alpha_2 \, E(X_{n+1}^2 \mid X_1, \ldots, X_n) \\
&= \alpha_0 + \sum_{i=3}^{p} \alpha_i \, X_{n+3-i}^2 + \alpha_1 \big(\alpha_0 + \sum_{i=2}^{p} \alpha_i X_{n+2-i}^2 + \alpha_1 \sigma_{n+1}^2 \big) \\
&\quad + \alpha_2 \, \sigma_{n+1}^2 \,, \ldots \,.
\end{aligned}$$

In contrast to the stochastic volatility model, the predictions of X_{n+h}^2 and σ_{n+h}^2 are easy to calculate.

References

[1] Adler,R.J., Feldman, R.E. and Taqqu, M.S. (1998) *A Practical Guide to Heavy Tails.* Birkhauser, Boston.

[2] Andersen, T.G., Davis, R.A., Kreiss, J.-P. and Mikosch, T. (Eds.) (2009) *The Handbook of Financial Time Series.* Springer, Berlin.

[3] Berkes, I., Horváth, L. and Kokoszka, P. (2003) GARCH processes: structure and estimation. *Bernoulli* **9**, 201–227.

[4] Billingsley, P. (1968) *Convergence of Probability Measures.* Wiley, New York.

[5] Bollerslev, T. (1986) Generalized autoregressive conditional heteroskedasticity. *J. Econometrics* **31**, 307–327.

[6] Bougerol, P. and Picard, N. (1992) Stationarity of GARCH processes and of some nonnegative time series. *J. Econometrics* **52**, 115–127.

[7] Box, G.E.P. and Jenkins, G.M. (1976) *Time Series Analysis: Forecasting and Control.* Holden-Day, San Francisco.

[8] Brockwell, P. and Davis, R.A. (1991) *Time Series: Theory and Methods.* Springer, New York.

[9] Brockwell, P.J. and Davis, R.A. (2016) *Introduction to Time Series and Forecasting.* 3rd Edition. Springer, New York.

[10] Buraczewski, D., Damek, E. and Mikosch, T. (2016) *Stochastic Models with Power-Laws. The Equation* $X = AX + B$. Springer, New York.

[11] Doukhan, P. (1994) *Mixing. Properties and Examples.* Lecture Notes in Statistics **85**. Springer Verlag, New York.

[12] Doukhan, P., Oppenheim, G. and Taqqu, M.S. (Eds.) *Long Range Dependence.* Birkhäuser, Boston.

[13] Embrechts, P., Klüppelberg, C. and Mikosch, T. (1997) *Modelling Extremal Events for Insurance and Finance.* Springer, Berlin.

[14] Engle, R.F. (1982) Autoregressive conditional heteroscedastic models with estimates of the variance of United Kingdom inflation. *Econometrica* **50**, 987–1007.

[15] Engle, R.F. (Ed.) (1995) *ARCH Selected Readings.* Oxford University Press, Oxford (UK).

[16] Engle, R.F. and Bollerslev, T. (1986) Modelling the persistence of conditional variances. With comments and a reply by the authors. *Econometric Rev.* **5**, 1–87.

[17] Garcia, I., Klüppelberg, C. and Müller, G. (2011) Estimation of stable CARMA models with an application to electricity spot prices. *Stat. Model.* **11**, 447470.

[18] Goldie, C.M. (1991) Implicit renewal theory and tails of solutions of random equations. *Ann. Appl. Probab.* **1**, 126–166.

[19] Granger, C.W.J. and Joyeux, R. (1980) An introduction to long-memory time series and fractional differencing. *J, Time Series Anal.* **1** 15–30.

[20] Hurst, H.E. (1951) Long-term storage capacity of reservoirs. *Trans. Amer. Soc. Civil Engineers* **116**, 770–808.

[21] Ibragimov, I.A. and Linnik, Yu.V. (1971) *Independent and Stationary Sequences of Random Variables.* Wolters–Noordhoff, Groningen.

[22] Kendall, M.G. and Stuart, A. (1976) *The Advanced Theory of Statistics.* Vol. 3. Griffin, London.

[23] Kesten, H. (1973) Random difference equations and renewal theory for products of random matrices. *Acta Math.* **131**, 207–248.

[24] Krengel, U. (1985) *Ergodic Theorems.* De Gruyter, Berlin.

[25] Mikosch, T. (2003) Modeling dependence and tails of financial time series. In: Finkenstädt, B. and Rootén, H. (Eds.) *Extreme Values in Finance, Telecommunications, and the Environment,* pp. 185–286. Chapman and Hall, Boca Raton. A version of this paper is available under www.math.ku.dk/∼mikosch/Semstat.

[26] Mikosch, T. and Stărică, C. (2000) Limit theory for the sample autocorrelations and extremes of a GARCH(1,1) process. *Ann. Statist.* **28**, 1427–1451. An extended version is available under www.math.ku.dk/∼mikosch/preprint.

[27] Nelson, D.B. (1990) Stationarity and persistence in the GARCH$(1,1)$ model. *Econometric Theory* **6**, 318–334.

[28] Petrov, V.V. (1995) *Limit Theorems of Probability Theory.* Oxford University Press, Oxford.

[29] Priestley, M.B. (1981) *Spectral Analysis and Time Series, vols. I and II.* Academic Press, New York.

[30] Rosenblatt, M. (1956) A central limit theorem and a strong mixing condition. *Proc. National Acad. Sci.* **42**, 43–47.

[31] Samorodnitsky, G. (2016) *Stochastic Processes and Long Range Dependence.* Springer, New York.

[32] Samorodnitsky, G. and Taqqu, M.S. (1994) *Stable Non–Gaussian Random Processes. Stochastic Models with Infinite Variance.* Chapman and Hall, London.

[33] Straumann, D. (2004) *Estimation in Conditionally Heteroscedastic Time Series Models.* Springer. Berlin.

[34] Straumann, D. and Mikosch, T. (2006) Quasi-maximum likelihood estimation in heteroscedastic time series: a stochastic recurrence equation approach. *Ann. Statist.* **34**, 2449–2495

[35] Taylor, S.J. (1986) *Modelling Financial Time Series.* Wiley, Chichester, New York.

[36] Wand, M.P. and Jones, M.C. (1995) *Kernel Smoothing*. Chapman and Hall, London.
[37] Willinger, W., Taqqu, M.S., Leland, M. and Wilson. D. (1995) Self–similarity through high variability: statistical analysis of ethernet lan traffic at the source level. *Computer Communications Review*, 25:100–113, 1995. Proceedings of the ACM/SIGCOMM'95, Cambridge, MA.