Financial Time Series

Annika Lang & Andreas Petersson

version May 27, 2016

Preface

These lecture notes provide additional material to the lecture "Financial Time Series" (TMS087/MSA410) at Chalmers University of Technology and University of Gothenburg. The course is the fourth in a series of four financial courses (Options and Mathematics, Financial Risk, Stochastic Calculus) and assumes the student to be familiar with the mathematical content of the lectures as well as with basic statistical methods. It is to be seen as an early master level course. The intention of the course is to provide a mathematical framework for financial time series rather than to introduce as many available methods as possible. The student should understand the mathematical background of easy linear models, should be able to implement these from scratch, and should learn to transfer the knowledge to more advanced models.

First simulation examples have been added this year, more are planned in future years. Furthermore an introductory chapter to probability theory was included and the last chapter about extreme value theory was omitted. Since the notes are still under construction, we are thankful for any comments that help to improve them.

Göteborg, May 2016, Annika Lang & Andreas Petersson

Preface to v2015

These lecture notes were written in parallel to the lecture "Financial Time Series" (TMS087/MSA410) held by the author at Chalmers University of Technology and University of Gothenburg in Spring 2015. They are based on [4, 3, 10, 17].

The lecture notes are no more than a first draft, where examples, especially financial applications, as well as graphs and plots are still missing. They are first just exclusively handed to the students of the class.

Please help to improve the notes for future students and send any typos, problems, and remarks to the author (annika.lang@chalmers.se).

Göteborg, May 2015, Annika Lang

Contents

CHAPTER 1

Prerequisites

The intention of this chapter is to provide all readers with the necessary prerequisites in probability theory, mathematical statistics, and financial mathematics. In the lecture it is assumed that this content is already known and it is the personal responsibility of every student to be familiar with the introduced definitions, notations, and results included in this chapter.

1.1 Review on probability theory

The attempt of this section is to give an introduction to probability theory that is as short as possible but provides the reader with all basics that are required throughout the lecture. The presentation of results is highly inspired by [14]. For a more extended but still easy introduction to probability theory in English than that given below the reader is referred for example to [16]. We start with the very basic concept of a probability space.

Let Ω be a nonempty set. A system $\mathcal{A}$ of subsets $A \subseteq \Omega$ is called a *σ-algebra on* Ω if $\Omega \in \mathcal{A}$, it is closed under complements, i.e., $A \in \mathcal{A}$ implies $A^c = \Omega \backslash A \in \mathcal{A}$, and it is closed under countable unions, i.e., for all sequences $(A_n, n \in \mathbb{N})$, $A_n \in \mathcal{A}$ for all $n \in \mathbb{N}$, it holds that $\bigcup_{n\in\mathbb{N}} A_n \in \mathcal{A}$. The pair $(\Omega, \mathcal{A})$ is called a *measurable space* and elements of $\mathcal{A}$ are called *measurable sets*. A subset $\mathcal{G} \subset \mathcal{A}$ is a *sub-σ-algebra* of the σ-algebra $\mathcal{A}$ if $\mathcal{G}$ is a σ-algebra itself.

There exist many different σ-algebras. The simplest (and most boring) σ-algebra just consists of the empty set $\emptyset$ and Ω. It is an easy *exercise* to show that this is actually a σ-algebra. More interesting and frequently used σ-algebras include the power set $\mathcal{P}(\Omega)$ of Ω, which is the set of all subsets of Ω, the σ-algebra generated by a subset $\mathcal{E}$ of the power set, which is the smallest σ-algebra that contains $\mathcal{E}$, and the *Borel σ-algebra* over $\Omega = \mathbb{R}$, which is the σ-algebra generated by all half-open intervals of $\mathbb{R}$. This latter σ-algebra is denoted by $\mathcal{B}(\mathbb{R})$.

To "measure sizes" on a measurable space $(\Omega, \mathcal{A})$, let $\mu : \mathcal{A} \to \mathbb{R}_+$ be a mapping that satisfies $\mu(\emptyset) = 0$ as well as being *σ-additive*, i.e., for all sequences $(A_n, n \in \mathbb{N})$ of pairwise disjoint sets being elements of $\mathcal{A}$, it holds that

$$\mu\Big(\biguplus_{n\in\mathbb{N}} A_n\Big) = \sum_{n\in\mathbb{N}} \mu(A_n).$$

Then μ is called a *measure* on $(\Omega, \mathcal{A})$ and the triple $(\Omega, \mathcal{A}, \mu)$ is called a *measure space*. If furthermore $\mu(\Omega) = 1$, μ is called a *probability measure* and usually denoted by $P : \mathcal{A} \to [0, 1]$. The triple $(\Omega, \mathcal{A}, P)$ is then called a *probability space*.

A well-known measure is the *Lebesgue measure* λ which is defined on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ by

$$\lambda([a, b)) := b - a$$

for all half-open intervals $[a, b) \subset \mathbb{R}$.

Next, let $f : \Omega \to \mathbb{R}$ be a function and set for $B \in \mathcal{B}(\mathbb{R})$

$$f^{-1}(B) := \{\omega \in \Omega, f(\omega) \in B\}.$$

If $f^{-1}(B) \in \mathcal{A}$ for all $B \in \mathcal{B}(\mathbb{R})$, f is called *measurable*. When $\Omega = \mathbb{R}$ in this definition, $\mathcal{A}$ is taken to be $\mathcal{B}(\mathbb{R})$. The $\sigma-$algebra $\sigma(f)$ is generated by $\{f^{-1}(B), B \in \mathcal{B}(\mathbb{R})\} \subseteq \mathcal{P}(\Omega)$ and it is the smallest σ-algebra on Ω with respect to which f is measurable. It is an easy *exercise* to show that $\{f^{-1}(B), B \in \mathcal{B}(\mathbb{R})\}$ is a σ-algebra so that in fact $\sigma(f) = \{f^{-1}(B) : B \in \mathcal{B}(\mathbb{R})\}$.

In the following lemma, it is shown that measurability is preserved under the composition of measurable functions.

Lemma 1.1.1. *Let $g : \Omega \to \mathbb{R}$ and $f : \mathbb{R} \to \mathbb{R}$ be measurable functions, then $f \circ g : \Omega \to \mathbb{R}$ is measurable.*

Proof. Observe that for any $B \in \mathcal{B}(\mathbb{R})$

$$(f \circ g)^{-1}(B) = \{\omega \in \Omega, f(g(\omega)) \in B\} = \{\omega \in \Omega, g(\omega) \in f^{-1}(B)\} = g^{-1}(f^{-1}(B)).$$

Since $f^{-1}(B) \in \mathcal{B}(\mathbb{R})$ due to the measurability of f, $g^{-1}(f^{-1}(B)) \in \mathcal{A}$ by the measurability of g and the claim is proven. □

In the context of a probability space $(\Omega, \mathcal{A}, P)$, a measurable mapping $X : \Omega \to \mathbb{R}$ is called a (real-valued) *random variable* and the lemma implies that for any measurable function $f : \mathbb{R} \to \mathbb{R}$ the function $f \circ X$ is also a random variable.

Let X be a random variable and consider for $B \in \mathcal{B}(\mathbb{R})$

$$P_X(B) := P(X^{-1}(B)) = P(\{\omega \in \Omega, X(\omega) \in B\}) = P(X \in B),$$

where we use all notations as synonyms. Then it can be shown that $P_X : \mathcal{B}(\mathbb{R}) \to [0, 1]$ is a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ called the *image measure of P under X*. It is also called the *distribution of X*. The *cumulative distribution function* $F_X : \mathbb{R} \to [0, 1]$ is then defined by

$$F_X(x) := P_X((-\infty, x]) = P(X \leq x), \qquad x \in \mathbb{R}.$$

To omit the introduction of Lebesgue integration in what follows, we have to distinguish between continuous and discrete random variables and use Riemann integration and summation rules to define expectations of random variables.

A random variable X is called *discretely distributed* if it takes values in a countable subset of $\mathbb{R}$ with probability 1, i.e., there exists a real-valued (and possibly finite—but we use the infinite notation for simplicity) sequence $(x_i, i \in \mathbb{N})$ with $x_i \neq x_j$ for all $i, j \in \mathbb{N}$ such that

$$P(X = x_i) = p_X(x_i) > 0$$

for all $i \in \mathbb{N}$ and

$$P(X = x_i, i \in \mathbb{N}) = P\Big(\biguplus_{i \in \mathbb{N}} \{X = x_i\}\Big) = \sum_{i \in \mathbb{N}} p_X(x_i) = 1.$$

Then with

$$\epsilon_x(A) := \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{else} \end{cases}$$

for all $A \in \mathcal{B}(\mathbb{R})$ one obtains with the properties of a probability measure that the distribution of X can be expressed by

$$P_X(A) = \sum_{i=1}^{\infty} p_X(x_i)\epsilon_{x_i}(A).$$

While the cumulative distribution function of a discrete random variable is a stepfunction, a random variable X is called *continuously distributed* if its cumulative distribution function F_X is continuous. In what follows let us take the stronger assumption that F_X is differentiable with derivative f_X. Then it holds that

$$F_X(x) = \int_{-\infty}^{x} f_X(x)\,\mathrm{d}x$$

and f_X is called the *density* of X. This implies that for all intervals $(a, b]$ we are able to compute the probability that X is in $(a, b]$ by

$$P(X \in (a,b]) = P_X((a,b]) = P_X((-\infty,b]) - P_X((-\infty,a]) = F_X(b) - F_X(a) = \int_a^b f_X(x)\,\mathrm{d}x.$$

We should remark that not all random variables follow either a continuous or a discrete distribution but that there exist mixtures of both.

An important quantity of interest is the "average" or "mean" of a random variable. What can we expect to be its value when observing it? Put into a mathematical framework, the average is described by the *expectation* of a random variable which is formally (or if Lebesgue integration is known and X is integrable with respect to P) given by the integration of the random variable with respect to the probability measure P

$$\mathbb{E}(X) := \int_\Omega X(\omega)\,\mathrm{d}P(\omega).$$

By the transformation theorem this rather abstract expression can be simplified for continuous random variables to

$$\mathbb{E}(X) = \int_{\mathbb{R}} x f_X(x)\,\mathrm{d}x$$

and for discrete random variables to

$$\mathbb{E}(X) = \sum_{i=1}^{\infty} p_X(x_i)x_i.$$

We have already learned that $g \circ X$ is a random variable if g is measurable. Frequently we will compute expectations of more general expressions than $\mathbb{E}(X)$ which are of the form $g(X) = g \circ X$. Therefore we include the computing rules for expectations of these random variables for the convenience of the reader. For continuous random variables we obtain

$$\mathbb{E}(g(X)) = \int_{\mathbb{R}} g(x) f_X(x)\,\mathrm{d}x,$$

while for discrete ones we compute

$$\mathbb{E}(g(X)) = \sum_{i=1}^{\infty} p_X(x_i) g(x_i).$$

An important inequality that relates probabilities and expectations is *Chebyshev's inequality* which states that for any $\alpha > 0$ and any strictly increasing function g on $\mathbb{R}^+$ it holds that

$$P(|X| > \alpha) \leq g(\alpha)^{-1}\,\mathbb{E}(g(|X|)). \tag{1.1}$$

If g is the identity, the inequality is known as *Markov's inequality*. Another frequently used function is $g(x) = x^2$ which has the desired properties.

The definition of the expectation enables us to define the *variance* of a random variable X, which is given by

$$\mathsf{Var}(X) := \mathbb{E}((X - \mathbb{E}(X))^2).$$

In an easy *exercise* one shows that the variance is equal to

$$\mathsf{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2.$$

From this expression it is clear that a finite variance requires besides a finite expectation that the *second moment* exists, i.e., $\mathbb{E}(X^2) < +\infty$.

If X and Y are two random variables, then a "generalization" of the variance is the so-called *covariance* of X and Y which is defined by

$$\mathsf{Cov}(X, Y) := \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))).$$

It can be scaled to a quantity taking values in $[-1, 1]$ by

$$\mathsf{Cor}(X, Y) := \frac{\mathsf{Cov}(X, Y)}{\sqrt{\mathsf{Var}(X)\,\mathsf{Var}(Y)}}$$

and is called the *correlation* of X and Y.

If $\mathsf{Cov}(X, Y) = 0$ and therefore also $\mathsf{Cor}(X, Y) = 0$ (under the assumption of the non-trivial case that neither $\mathsf{Var}(X)$ nor $\mathsf{Var}(Y)$ is equal to zero), X and Y are said to be *uncorrelated* or *orthogonal* (in the sense of $L^2(\Omega; \mathbb{R})$).

While the expectation is *linear*, i.e., for random variables X, Y and constants $\alpha, \beta \in \mathbb{R}$

$$\mathbb{E}(\alpha X + \beta Y) = \alpha\,\mathbb{E}(X) + \beta\,\mathbb{E}(Y),$$

which is due to the linearity of the integral and of sums, respectively, this does not hold for the variance and covariance. Nevertheless, under the assumption of uncorrelated random variables we obtain the following formula for the variance of sums of random variables.

Theorem 1.1.2 (Bienaymé)**.** *Let $X_1, \ldots, X_n$ be pairwise uncorrelated random variables and $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$, then*

$$\mathsf{Var}\left(\sum_{i=1}^{n} \alpha_i X_i\right) = \sum_{i=1}^{n} \alpha_i^2\,\mathsf{Var}(X_i).$$

Proof. Let us divide the proofs into two steps. We first observe that for $\alpha \in \mathbb{R}$ and a random variable X is holds that

$$\mathsf{Var}(\alpha X) = \mathbb{E}((\alpha X - \mathbb{E}(\alpha X))^2) = \alpha^2\,\mathbb{E}((X - \mathbb{E}(X))^2) = \alpha^2\,\mathsf{Var}(X).$$

Therefore it is sufficient to prove the claim for $\alpha_1 = \cdots = \alpha_n = 1$. Furthermore we can assume without loss of generality that $\mathbb{E}(X_1) = \cdots = \mathbb{E}(X_n) = 0$. We compute

$$\mathsf{Var}\left(\sum_{i=1}^{n} X_i\right) = \mathbb{E}((X_1 + \cdots + X_n)^2) = \sum_{i=1}^{n} \mathbb{E}(X_i^2) + \sum_{i \neq j} \mathbb{E}(X_i X_j) = \sum_{i=1}^{n} \mathsf{Var}(X_i) + \sum_{i \neq j} \mathsf{Cov}(X_i, X_j).$$

Since the random variables are pairwise uncorrelated, it holds that $\mathsf{Cov}(X_i, X_j) = 0$ for all $i \neq j$ by definition and the claim follows. □

Let us next consider a stronger assumption on sequences of random variables than was assumed in the theorem of Bienaymé. Therefore let $(X_n, n \in \mathbb{N})$ be a sequence of random variables. The sequence is called *independent* if for all $n \in \mathbb{N}$, all positive integers $k_1 < \cdots < k_n$, and all choices $x_{k_1}, \ldots, x_{k_n} \in \mathbb{R}$ it holds that

$$P(X_{k_1} < x_{k_1}, \ldots, X_{k_n} < x_{k_n}) = \prod_{i=1}^{n} P(X_{k_i} < x_{k_i}) = P(X_{k_1} < x_{k_1}) \cdots P(X_{k_n} < x_{k_n}).$$

One can prove that this definition is actually sufficient for independence and implies the "usual" condition that for all $B_{k_1}, \ldots, B_{k_n} \in \mathcal{B}(\mathbb{R})$

$$P(X_{k_1} \in B_{k_1}, \ldots, X_{k_n} \in B_{k_n}) = \prod_{i=1}^{n} P(X_{k_i} \in B_{k_i}).$$

In order to show that the independence of random variables is stronger than the requirement that they are uncorrelated, we need the following result first.

Theorem 1.1.3. *Let $X_1, \ldots, X_n$ be independent random variables and $g_1, \ldots, g_n$ measurable functions such that*

$$\mathbb{E}(g_1(X_1) \cdots g_n(X_n)) < +\infty$$

exists, then

$$\mathbb{E}(g_1(X_1) \cdots g_n(X_n)) = \mathbb{E}(g_1(X_1)) \cdots \mathbb{E}(g_n(X_n)).$$

We remark for the interested reader that the theorem is proven by the observation that the product measure of the random variables is equal to the product of the image measures, i.e.,

$$P_{X_1,\ldots,X_n} = P_{X_1} \otimes \cdots \otimes P_{X_n}$$

and Fubini's theorem.

Coming back to the comparison of independent and uncorrelated random variables, let us set for two independent random variables X and Y

$$g_1(X) := g_2(X) := X - \mathbb{E}(X),$$

which is a measurable function under the assumption that $\mathbb{E}(X) < +\infty$ and $\mathbb{E}(Y) < +\infty$. Then the theorem implies that

$$\mathsf{Cov}(X, Y) = \mathbb{E}(g_1(X)g_2(Y)) = \mathbb{E}(g_1(X))\,\mathbb{E}(g_2(Y)) = (\mathbb{E}(X) - \mathbb{E}(X))(\mathbb{E}(Y) - \mathbb{E}(Y)) = 0,$$

i.e., we have shown that the independence of two random variables implies that they are uncorrelated. Nevertheless, the reader should be aware that uncorrelated random variables are usually not independent.

Product measures were already mentioned in the remark on the proof of Theorem 1.1.3 but were not discussed so far. The *product measure* $P_{X,Y}$ of two random variables X and Y is defined by the completion of

$$P_{X,Y}(A \times B) := P(X \in A, Y \in B), \qquad A, B \in \mathcal{B}(\mathbb{R}).$$

The *conditional probability* of X given Y is defined by

$$P(X \in A | Y \in B) := \frac{P(X \in A, Y \in B)}{P(Y \in B)}$$

for $A, B \in \mathcal{B}(\mathbb{R})$ with $P(Y \in B) \neq 0$, which leads for continuously distributed random variables with joint density $f_{X,Y}$ to the *conditional density* given by

$$f(x|y) = \begin{cases} \frac{f_{X,Y}(x,y)}{f_Y(y)} & \text{if } f_Y(y) \neq 0, \\ 0 & \text{else} \end{cases}$$

for $x, y \in \mathbb{R}$. Here f_Y is the (marginal) density of Y which can be derived by

$$f_Y(y) = \int_{\mathbb{R}} f_{X,Y}(x, y)\,\mathrm{d}x.$$

For discrete random variables we obtain

$$p(x|y) = \begin{cases} \frac{p_{X,Y}(x,y)}{p_Y(y)} & \text{if } p_Y(y) \neq 0, \\ 0 & \text{else,} \end{cases}$$

where the weights are given by $p_{X,Y}(x, y) = P(X = x, Y = y)$ and the (marginal) weights $p_Y(y)$ can be computed by

$$p_Y(y) = \sum_{i=1}^{\infty} p_{X,Y}(x_i, y) = P(Y = y),$$

where $(x_i, i \in \mathbb{N})$ denotes the values in $\mathbb{R}$ with strictly positive probability.

In what follows next, we use this concept to define conditional expectations. The reader should be aware that we are doing this introduction for a very specific case. Usually conditional expectations are considered in the more general setting with respect to σ-algebras instead of random variables. The experienced reader will observe quite easily that using the σ-algebra $\sigma(Y)$ generated by the random variable Y instead of Y leads to the same conditional expectations as those introduced in what follows.

Let X and Y be two random variables and assume that X is integrable or positive. Then, by the theorem of Radon–Nikodym, there exists a P-almost surely unique random variable Z with the properties that there exists a measurable function $g : \mathbb{R} \to \mathbb{R}$ such that

$$Z(\omega) = g(Y(\omega))$$

for all $\omega \in \Omega$ and for all $B \in \mathcal{B}(\mathbb{R})$

$$\int_{\{Y \in B\}} Z(\omega)\,\mathrm{d}P(\omega) = \int_{\{Y \in B\}} X(\omega)\,\mathrm{d}P(\omega).$$

The random variable Z is called the *conditional expectation* of X given Y and denoted by $\mathbb{E}(X|Y)$. Observe that in contrast to $\mathbb{E}(X)$, the conditional expectation $\mathbb{E}(X|Y)$ is a random variable which could be interpreted as the best approximation of X given just Y. In this context *P-almost surely* means that for all random variables Z' that also satisfy the two properties it holds that $P(Z = Z') = 1$.

For practical purposes and a more specific and concrete form of the conditional expectation we add that the abstract condition of integration with respect to the probability measure implies for continuous random variables that the conditional expectation is given by

$$\mathbb{E}(X|Y) = \int_{\mathbb{R}} x f(x|Y)\,\mathrm{d}x.$$

For discrete random variables one obtains that

$$\mathbb{E}(X|Y) = \sum_{i=1}^{\infty} p(x_i|Y)x_i,$$

where one should be aware that the result is a random variable which could be characterized by computing $\mathbb{E}(X|Y = y_j)$ for all $y_j \in \mathbb{R}$, $j \in \mathbb{N}$, with $P(Y = y_j) > 0$.

In what follows we give a selection of properties of the conditional expectation, where the reader is referred to the literature for the proofs or derives the results in easy computations. The conditional expectation has the following properties:

(i) The conditional expectation is linear, i.e., for $a_1, a_2 \in \mathbb{R}$ and random variables X_1, X_2, and Y it holds that
$$\mathbb{E}(a_1X_1 + a_2X_2|Y) = a_1\,\mathbb{E}(X_1|Y) + a_2\,\mathbb{E}(X_2|Y).$$

(ii) The expectation of the conditional expectation $\mathbb{E}(X|Y)$ is equal to the expectation of the random variable X, i.e.,
$$\mathbb{E}(\mathbb{E}(X|Y)) = \mathbb{E}(X).$$

(iii) If X is independent of Y, the conditional expectation satisfies
$$\mathbb{E}(X|Y) = \mathbb{E}(X),$$
i.e., the best approximation of X given Y is the expectation of X.

(iv) For every constant $a \in \mathbb{R}$ it holds that
$$\mathbb{E}(a|Y) = a.$$

(v) For any measurable function $g : \mathbb{R} \to \mathbb{R}$ and random variables X and Y it holds that
$$\mathbb{E}(g(Y)X|Y) = g(Y)\,\mathbb{E}(X|Y).$$

We will need conditional expectations given a whole family of random variables in the lecture to obtain the best forecast using the past observations of a time series. Therefore we have to generalize the conditional expectation to $\mathbb{E}(X|Y_1, \ldots, Y_n)$ for random variables X and $Y_1, \ldots, Y_n$. This is easily be done by finding a measurable function $g : \mathbb{R}^n \to \mathbb{R}$ such that $Z = g(Y_1, \ldots, Y_n)$. All presented results stay the same under this generalization (and instead of $\sigma(Y)$ one considers $\sigma(Y_1, \ldots, Y_n)$ to consider it in the "usual approach" of conditional expectations).

We continue this very short introduction to probability theory with a collection of examples of frequently used distributions.

Example 1.1.4 (Bernoulli distribution)**.** The *Bernoulli distribution* is a discrete distribution that takes values in $\{0, 1\}$ and that models a coin flipping experiment. It is characterized by the parameter $p \in (0, 1)$. A Bernoulli distributed random variable X has the distribution
$$P(X = 1) := p, \qquad P(X = 0) := 1 - p.$$
In an easy computation one obtains that
$$\mathbb{E}(X) = p, \qquad \mathsf{Var}(X) = p(1 - p).$$

Example 1.1.5 (Uniform distribution)**.** A random variable X is *uniformly distributed* on the interval $[a, b]$ denoted by $X \sim \mathcal{U}([a, b])$ if it is continuous with density given by
$$f_X(x) := \begin{cases} (b-a)^{-1} & \text{if } x \in [a, b], \\ 0 & \text{else.} \end{cases}$$
It is an easy *exercise* to compute that
$$\mathbb{E}(X) = \frac{a+b}{2}, \qquad \mathsf{Var}(X) = \frac{(b-a)^2}{12}.$$
A useful observation especially for simulations is that if $X \sim \mathcal{U}([0, 1])$, then
$$a + (b - a)X \sim \mathcal{U}([a, b])$$
for real numbers $a < b$.

Example 1.1.6 (Normal distribution)**.** One of the most famous and most frequently used distributions is the normal distribution. A random variable X is *normally distributed* or *Gaussian* with expectation μ and variance σ^2 denoted by $X \sim \mathcal{N}(\mu, \sigma^2)$ if it is continuously distributed with density given by

$$f_X(x) := \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

The cumulative distribution function of this distribution is usually denoted by

$$\Phi(x) := \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{x} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) \mathrm{d}y.$$

The expectation and the variance fully characterize the distribution and a family of normally distributed random variables is independent if it is jointly normally distributed and uncorrelated.

A central property of the normal distribution is the simple but remarkable fact that the sample or empirical mean $\bar{X}_n$ of a large number of random variables of any distribution will be approximately normally distributed under some simple conditions. This is not proven in these lecture notes but merely stated below. For a proof the reader is referred for example to [16]. The reader should be aware that there exist many versions of this theorem with different assumptions on the underlying random variables. The following is one of the most common with the strongest assumptions.

Theorem 1.1.7 (Central Limit Theorem (CLT))**.** *Let $(X_n, n \in \mathbb{N})$ be a sequence of independent and identically distributed random variables, each having finite mean μ and finite non-zero variance σ^2 and let $\bar{X}_n = n^{-1}\sum_{i=1}^{n} X_i$. Then the distribution of the standardized sample mean tends to the standard normal distribution, i.e. for all $x \in \mathbb{R}$*

$$P\left(\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \leq x\right) \to \Phi(x).$$

1.2 Review on mathematical statistics

A *hypothesis* is a statement about a parameter. We have two complementary hypotheses in a hypothesis testing problem which are called the *null hypothesis* H_0 and the *alternative hypothesis* H_1. Finally a *hypothesis testing procedure* or *hypothesis test* is a rule that specifies for which sample values the decision is made to accept H_0 as true and for which H_0 is rejected and H_1 is accepted as true.

CHAPTER 2

Stationary time series and seasonality

This chapter is based on Brockwell and Davis' book "Introduction to Time Series and Forecasting" [4] as well as Grandell's lecture notes "Time series analysis" [10], which are based on [4] and [3].

We start with an introduction to times series in general before focusing on stationary time series. We discuss especially the testing and forecasting of stationary times series. Finally we introduce methods to remove trend and seasonal components from observed data in order to obtain a stationary time series. Let us from here on in all of what follows consider random variables with respect to a fixed probability space $(\Omega, \mathcal{A}, P)$.

2.1 Introduction to time series

The goal of this section is to set up a mathematical framework that describes the behavior of observed data which might come from the stock market but many other sources in engineering, ecology, and finance can be treated in a similar way. We consider special types of stochastic processes which we are observing and trying to estimate, fit, and forecast. Therefore we first recall that a *stochastic process* $X := (X_t, t \in \mathbb{T})$ is a collection of random variables with respect to an index set $\mathbb{T}$. In the context of these lecture notes let $\mathbb{T} \subset \mathbb{R}$. We call X a *stochastic process in continuous time* if $\mathbb{T}$ is a (possibly unbounded) interval while it is called a *stochastic process in discrete time* if $\mathbb{T}$ is countable, i.e., $\mathbb{T} = \{t_n, n \in \mathbb{N}\}$ with $t_n \in \mathbb{R}$ for all $n \in \mathbb{N}$. While a stochastic process is the mathematical construction of some random behavior over time, we are interested in the observation of this process, e.g., of the evolution of a stock price. This will be done in the following framework:

Definition 2.1.1. A *time series* is a real-valued sequence of observations $(x_t, t \in \mathbb{T})$ with respect to an index set $\mathbb{T} \subset \mathbb{R}$. A *time series model* for the observed data $(x_t, t \in \mathbb{T})$ is a specification of the joint distributions (or possibly only the means and covariances) of a sequence of random variables $(X_t, t \in \mathbb{T})$ of which $(x_t, t \in \mathbb{T})$ postulates to be a realization.

The definition implies that a time series model is a stochastic process, but it might happen that we do not know all of its properties explicitly but just some specific quantities like the expectation or the covariances. We remark that we use the term *time series* to mean both the data and the underlying stochastic process if there is no danger of confusion.

Example 2.1.2. An example of a time series in the sense of a realization of a stochastic process is the set of quarterly earnings $(x_t)_{t=1}^{42}$ for the Swedish clothing company H&M. Figure 2.1.2 shows these earnings in million SEK from January 2006 through April 2016.

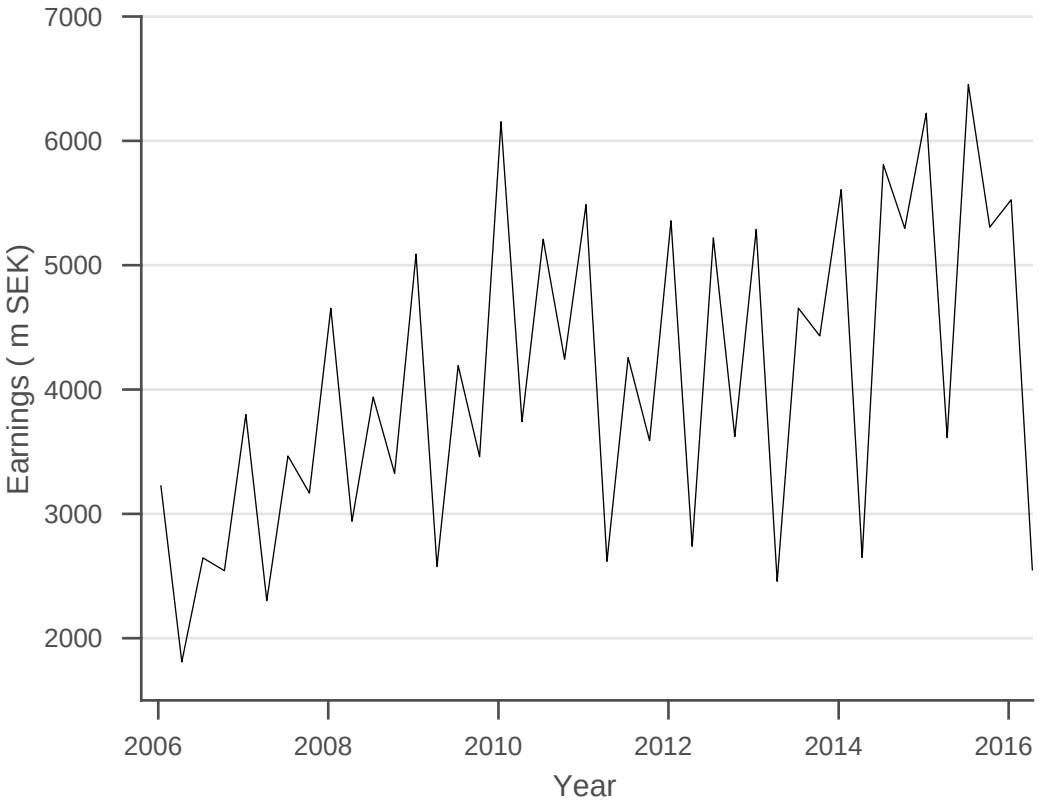

Figure 2.1: Quarterly earnings of H&M from January 2006 through April 2016.

Let us observe that in reality we are just able to observe the stochastic process at finitely many times. Therefore, we focus in these lecture notes on *discrete-time* time series, i.e., $\mathbb{T} = \{t_n, n \in \mathbb{N}\}$, and allow also for infinitely many observations, because the number of observations is not necessarily bounded (from the beginning). Let us assume from now on that $\mathbb{T}$ is a discrete set $\{t_n, n \in \mathbb{N}\}$ and let us abbreviate $(X_{t_n}, n \in \mathbb{N})$ by $(X_n, n \in \mathbb{N})$. Equivalently we write $(x_n, n \in \mathbb{N})$ and for finite observations and models $(x_1, \dots, x_n)$ and $(X_1, \dots, X_n)$, resp., for some finite and fixed $n \in \mathbb{N}$.

For a discrete time series, the specification of the joint distributions in Definition 2.1.1 simplifies to the knowledge of all probabilities

$$P_{X_{i_1},\dots,X_{i_m}}((-\infty, y_1], \dots, (-\infty, y_m]) = P(X_{i_1} \leq y_1, \dots, X_{i_m} \leq y_m)$$

for all finite random vectors $(X_{i_1}, \dots, X_{i_m})$ of any $\{i_1, \dots, i_m\} \subset \mathbb{N}$ with finite $m \in \mathbb{N}$ and all $y_j \in \mathbb{R}$, $j = 1, \dots, m$.

Although we claimed that the characterization of the joint distribution of a discrete time series is already simpler, it is still not convenient and in general not easy to derive results in this framework. To keep the technicalities in these lecture notes as low as possible, we will therefore introduce in what follows so-called iid noise.

Definition 2.1.3. A stochastic process $X = (X_t, t \in \mathbb{T})$ is called *iid noise* with mean zero and variance σ^2 if the sequence of random variables $(X_t, t \in \mathbb{T})$ is independent and identically distributed (abbreviated by *iid*) with $\mathbb{E}(X_t) = 0$ and $\mathsf{Var}(X_t) = \sigma^2$ for all $t \in \mathbb{T}$. An iid noise is denoted by $X \sim \mathrm{IID}(0, \sigma^2)$.

Please note that iid noise is sometimes called *white noise* in the literature (e.g., in [17]). We will use the terminology *white noise* for a more general process that satisfies weaker assumptions than iid noise.

In what follows we treat two simple examples of time series models.

Example 2.1.4 (Binary process)**.** A simple stochastic process and an example of an iid noise is the *binary process* which describes the flipping of a fair coin. In this case $(X_n, n \in \mathbb{N})$ is a sequence of iid random variables characterized by

$$P(X_1 = 1) = P(X_1 = -1) = \frac{1}{2}.$$

It is easy to see that it has mean zero, i.e.,

$$\mathbb{E}(X_1) = (-1) \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} = 0,$$

and variance 1, i.e.,

$$\mathsf{Var}(X_1) = \mathbb{E}((X_1 - \mathbb{E}(X_1))^2) = \mathbb{E}(X_1^2) = (-1)^2 \cdot \frac{1}{2} + 1^2 \cdot \frac{1}{2} = 1.$$

Example 2.1.5 (Random walk)**.** A *random walk* $(S_n, n \in \mathbb{N}_0)$ is obtained by the cumulative summing of iid random variables, i.e., for a given iid noise $(X_n, n \in \mathbb{N})$, it is defined by $S_0 := 0$ and for $n \in \mathbb{N}$ by

$$S_n := \sum_{i=1}^{n} X_i = S_{n-1} + X_n.$$

If the sequence of random variables is given by the binary process in Example 2.1.4, the corresponding random walk is called a *simple symmetric random walk*.

We finish this section by introducing the important example of a Gaussian time series.

Definition 2.1.6. A time series X is said to be a *Gaussian time series* if all finite-dimensional distributions are normal.

2.2 Characterization of stationary time series

Having seen time series models in general in the previous section, let us focus on the specific class of stationary times series and its properties in what follows.

Definition 2.2.1. Let $X = (X_t, t \in \mathbb{T})$ be a stochastic process with $\mathsf{Var}(X_t) < +\infty$ for all $t \in \mathbb{T}$. The *mean function* $\mu_X : \mathbb{T} \to \mathbb{R}$ of X is given by

$$\mu_X(t) := \mathbb{E}(X_t)$$

for all $t \in \mathbb{T}$ and the *covariance function* $\gamma_X : \mathbb{T} \times \mathbb{T} \to \mathbb{R}$ is defined by

$$\gamma_X(r, s) := \mathsf{Cov}(X_r, X_s) = \mathbb{E}\left((X_r - \mu_X(r))(X_s - \mu_X(s))\right)$$

for all $r, s \in \mathbb{T}$.

In order to avoid problems with the index set of the stochastic process especially when summing indices, let us consider for simplicity $\mathbb{T} = \mathbb{Z}$ in what follows, where we allow for negative times keeping in mind historical data.

Definition 2.2.2. Let $X = (X_t, t \in \mathbb{Z})$ be a time series with $\mathsf{Var}(X_t) < +\infty$ for all $t \in \mathbb{Z}$. The time series X is called *(weakly) stationary* if

(i) there exists $\mu \in \mathbb{R}$ such that $\mu_X(t) = \mu$ for all $t \in \mathbb{Z}$ and

(ii) $\gamma_X(r, s) = \gamma_X(r + h, s + h)$ for all $r, s, h \in \mathbb{Z}$.

Further, a time series X is said to be *strictly stationary* if the random variables $(X_1, \ldots, X_n)$ and $(X_{1+h}, \ldots, X_{n+h})$ have the same joint distributions for all $h \in \mathbb{Z}$ and $n \in \mathbb{N}$.

It is an easy *exercise* that a strictly stationary time series with finite variance is also weakly stationary. Whenever *stationary* is used in what follows, we shall mean weak stationarity. Furthermore observe that the converse is just true in very special cases. The typical example is that a weakly stationary Gaussian time series is also strictly stationary since the normal distribution is completely determined by its mean and covariance.

Furthermore, we observe that Condition (ii) in Definition 2.2.2 implies that $\gamma_X(r, s)$ with $r, s \in \mathbb{Z}$ is actually a function of the distance $|r - s|$ and therefore it is convenient and sufficient to write

$$\gamma_X(h) := \gamma_X(h, 0) = \gamma_X(s + h, s)$$

for $h, s \in \mathbb{Z}$ for stationary time series. In this context h is called the *lag*.

Definition 2.2.3. Let X be a stationary time series. The *autocovariance function (ACVF)* $\gamma_X : \mathbb{Z} \to \mathbb{R}$ of X is defined by

$$\gamma_X(h) := \mathsf{Cov}(X_{t+h}, X_t)$$

for $h \in \mathbb{Z}$ and $t \in \mathbb{Z}$. The *autocorrelation function (ACF)* $\rho_X : \mathbb{Z} \to [-1, 1]$ of X is defined by

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

for $h \in \mathbb{Z}$.

Note that γ_X is well-defined due to the stationarity of X. Furthermore we observe that ρ_X is given by the correlations of the time series. It is straightforward to see that

$$\rho_X(h) = \mathsf{Cor}(X_{t+h}, X_t) = \frac{\mathsf{Cov}(X_{t+h}, X_t)}{\sqrt{\mathsf{Var}(X_{t+h})\,\mathsf{Var}(X_t)}}$$

for all $h, t \in \mathbb{Z}$.

Let us introduce next the already announced generalization of iid noise.

Definition 2.2.4. A stochastic process $X = (X_t, t \in \mathbb{Z})$ is called a *white noise* with mean μ and variance σ^2 if it is a stationary process with $\mathbb{E}(X_t) = \mu$, $t \in \mathbb{Z}$, and for $h \in \mathbb{Z}$

$$\gamma_X(h) = \begin{cases} \sigma^2 & \text{if } h = 0, \\ 0 & \text{else.} \end{cases}$$

If X is a white noise it is denoted by $X \sim \mathrm{WN}(\mu, \sigma^2)$.

In other words a white noise is a sequence of uncorrelated random variables with constant mean and variance. It is clear from the definition that an iid noise is a white noise with mean 0 and variance σ^2. For a white noise to be iid noise on the other hand, it must be centered (i.e. $\mu = 0$) and the random variables must be independent and identically distributed. For example, a centered

white noise that is Gaussian is necessarily iid, since random variables that are uncorrelated and jointly normal are independent.

The mean and the covariance function as well as the autocovariance and the autocorrelation function of a time series are theoretical properties of the time series model. In practice we observe data and they are unknown. We assume a certain model that our data follows and try to estimate the parameters such as the four mentioned functions. In what follows we introduce *estimators* for the quantities of interest which is indicated by adding *sample* to the names. Observe that the introduced estimators are random variables while the quantities of interest themselves are deterministic.

Definition 2.2.5. Let $X = (X_t, t \in \mathbb{N})$ be a time series. The *sample mean* $\bar{X}_n$ of X is given by

$$\bar{X}_n := n^{-1} \sum_{t=1}^{n} X_t.$$

The *sample autocovariance function* $\hat{\gamma}$ is defined by

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

for $h = 0, \ldots, n-1$. Furthermore the *sample autocorrelation function* $\hat{\rho}$ is given by

$$\hat{\rho}(h) := \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}$$

for $h = 0, \ldots, n-1$.

We remark that the definitions of the sample autocovariance and autocorrelation function can be extended to $h = -n, \ldots, -1$ by setting for $h < 0$

$$\hat{\gamma}(h) := \hat{\gamma}(|h|),$$

which makes them symmetric functions around zero.

In an *exercise* one shows the well-known facts that $\bar{X}_n$ is an unbiased estimator for the mean if X is stationary, i.e., $\mathbb{E}(\bar{X}_n) = \mathbb{E}(X_1)$, while $\hat{\gamma}$ and $\hat{\rho}$ are not. We observe that the sample autocovariance and autocorrelation functions even stay biased if the factor n^{-1} is replaced by $(n-h)^{-1}$. Nevertheless, for large sample sizes they will nearly be unbiased.

Furthermore, we observe the convergence of the sample mean to the mean in the sense of the mean squared error in the following proposition.

Proposition 2.2.6. *Let X be a stationary time series with mean μ and autocovariance γ_X. Then*

$$\lim_{n\to\infty} \mathsf{Var}(\bar{X}_n) = \lim_{n\to\infty} \mathbb{E}\left((\bar{X}_n - \mu)^2\right) = 0$$

if $\sum_{|h|<\infty} |\gamma_X(|h|)| < +\infty$.

Proof. Let $n \in \mathbb{N}$ be fixed. Since the sample mean is an unbiased estimator of μ, we observe that

$$\mathsf{Var}(\bar{X}_n) = \frac{1}{n^2} \sum_{i,j=1}^{n} \mathbb{E}\left((X_i - \mu)(X_j - \mu)\right) = \frac{1}{n^2} \sum_{i,j=1}^{n} \mathsf{Cov}(X_i, X_j) = \frac{1}{n^2} \sum_{i,j=1}^{n} \gamma_X(|i-j|).$$

Let us simplify the sum next. It holds that

$$\sum_{i,j=1}^{n} \gamma_X(|i-j|) = 2 \sum_{i=1}^{n} \sum_{j=1}^{i-1} \gamma_X(i-j) + \sum_{i=1}^{n} \gamma_X(0) = 2 \sum_{h=1}^{n} (n-h)\gamma_X(h) + n\gamma_X(0).$$

Coming back to our original computation we obtain that

$$\mathsf{Var}(\bar{X}_n) = \frac{2}{n}\sum_{h=1}^{n}\left(1-\frac{h}{n}\right)\gamma_X(h) + \frac{1}{n}\gamma_X(0) = \frac{1}{n}\sum_{|h|<n}\left(1-\frac{|h|}{n}\right)\gamma_X(|h|) \leq \frac{1}{n}\sum_{|h|<n}|\gamma_X(|h|)|.$$

The assumption that $C := \sum_{|h|<\infty}|\gamma_X(|h|)| < +\infty$ yields that

$$\lim_{n\to\infty}\mathsf{Var}(\bar{X}_n) \leq \lim_{n\to\infty}\frac{C}{n} = 0,$$

which finishes the proof. □

From the last line of the proof we obtain especially that the rate of convergence of the mean squared error in the size of the sample is one, i.e., for all $n \in \mathbb{N}$

$$\mathbb{E}\left((\bar{X}_n - \mu)^2\right) \leq C \cdot \frac{1}{n}.$$

We remark that the sample mean of a Gaussian time series X is Gaussian since sums of Gaussian random variables are Gaussian. More specifically, one computes in an *exercise* that

$$n^{1/2}(\bar{X}_n - \mu) \sim \mathcal{N}\left(0, \sum_{|h|<n}(1-n^{-1}|h|)\gamma(h)\right).$$

Let us next have a look at the estimation of the autocovariance and autocorrelation function. First of all, it is evident that it is impossible to give reasonable estimates for $\gamma_X(h)$ and $\rho_X(h)$ for $h \geq n$, and even for h near to n the results are not reliable due to few samples. A useful guide can be found in [2], which says that one should take $n \geq 50$ and $h \leq n/4$.

For a compact notation and an efficient representation in a computer we denote by

$$\hat{\Gamma}_k := \begin{pmatrix} \hat{\gamma}(0) & \hat{\gamma}(1) & \cdots & \hat{\gamma}(k-1) \\ \hat{\gamma}(1) & \hat{\gamma}(0) & \cdots & \hat{\gamma}(k-2) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\gamma}(k-1) & \hat{\gamma}(k-2) & \cdots & \hat{\gamma}(0) \end{pmatrix}$$

the k-dimensional *sample covariance matrix*. It is nonnegative definite, which is shown in [4, Section 2.4.2]. The same holds true for the sample autocorrelation matrix $\hat{R}_k$ defined by

$$\hat{R}_k := \hat{\gamma}(0)^{-1}\hat{\Gamma}_k.$$

The matrices are nonsingular if $\hat{\gamma}(0) > 0$.

Observe that these functions can be defined for all observed time series. It has to be treated in what follows if this makes sense, i.e., if it is likely that the underlying stochastic process is stationary.

Example 2.2.7. Let us revisit the data from Example 2.1.2, i.e., the set $(x_t)_{t=1}^{42}$ of quarterly earnings of H&M. The sample mean is found to be $\bar{x} \approx 4007$. For the sample autocorrelation function, we compute $\hat{\rho}(h)$ for $h = 0, 1, \ldots, 10$ in MATLAB with the following code:

```
n=length(data);
mx=mean(data);
lags=10;
gamma=zeros(1,lags+1);
for h=0:lags
    gamma(h+1)=(data(1+h:end)-mx)*(data(1:end-h)-mx)'/n;
end
acf=gamma/gamma(1);
```

The result is shown in Figure 2.2.7. We see that $\hat{\rho}(h)$ displays a periodic behaviour - the peaks at lag h are similar to the peaks at lag $h-4$. This is often evidence of the presence of a so-called *seasonal component*, something that will be discussed in Section 2.4. We also note that the bounds $\pm 1.96/\sqrt{n}$ have been included in the figure, something that is explained in Method 2.2.8 below.

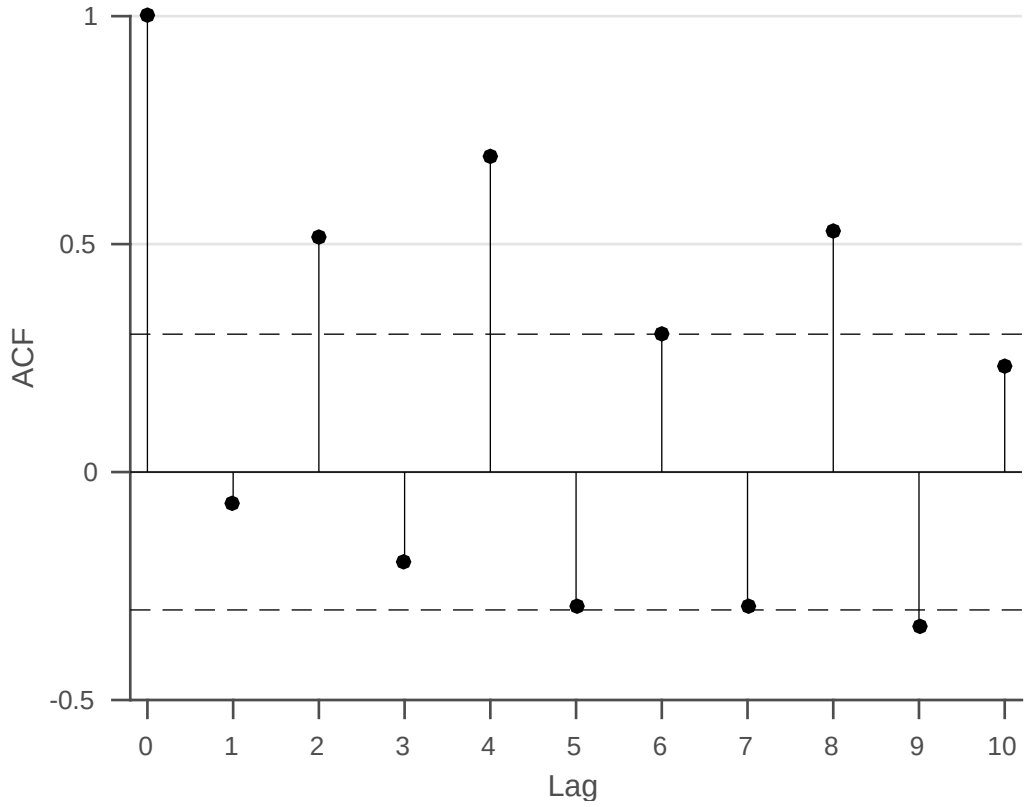

Figure 2.2: Sample autocorrelation function for the H&M data of Example 2.1.2.

In what follows we introduce several methods to test for stationarity or more specifically for independence of the observed data by using the properties of the (sample) autocorrelation. If we obtain that the observations are not iid random variables, we have to choose time series models that are more complicated than just the generation of iid random variables, where the knowledge of the distribution is sufficient. More details on the concept of hypothesis testing can be found in Section 1.2 if the reader is not familiar with these basic ideas of statistics.

Method 2.2.8 (Normality)**.** If $(Y_1, \ldots, Y_n)$ is a sequence of iid random variables with finite variance, then the sample autocorrelation is for sufficiently large n by the Central Limit Theorem 1.1.7 approximately $\mathcal{N}(0, n^{-1})$ distributed. Hence 95% should fall between the bounds $\pm 1.96/\sqrt{n}$. Use this for hypothesis testing at lag h with

$$H_0 : \rho_Y(h) = 0,$$
$$H_1 : \rho_Y(h) \neq 0,$$

and the test statistic

$$\lambda := \hat{\rho}(h) \left(\left(1 + 2 \sum_{i=1}^{h-1} \hat{\rho}(i)^2 \right) / n \right)^{-1/2}.$$

If $(Y_1, \dots, Y_n)$ is additionally Gaussian with $\rho_Y(j) = 0$ for $j > h$, the test statistic λ is asymptotically standard normally distributed. Hence H_0 is rejected if $|\lambda| > Z_{\alpha/2}$, where $Z_{\alpha/2}$ is the $100(1-\alpha/2)$th percentile of the normal distribution, i.e., $Z_{\alpha/2}$ is chosen such that

$$\int_{-\infty}^{Z_{\alpha/2}} \frac{1}{\sqrt{2\pi}} \exp\left(-\tfrac{x^2}{2}\right) \mathrm{d}x = 1 - \alpha/2.$$

Standard values for α include 0.05, 0.01, and 0.005, i.e., 5%, 1%, and 0.5%.

Method 2.2.9 (Portmanteau test, Box–Pierce test)**.** If $(Y_1, \dots, Y_n)$ is a sequence of iid random variables with finite variance, then one can show that $n \sum_{i=1}^{h} \hat{\rho}(i)^2$ is approximately χ^2_h distributed, i.e., chi-squared distributed with h degrees of freedom. Use this for hypothesis testing with

$$\begin{aligned} H_0 &: \rho_Y(1) = \cdots = \rho_Y(h) = 0, \\ H_1 &: \exists\, \rho_Y(i) \neq 0, i = 1, \dots, h, \end{aligned}$$

and the test statistic

$$\lambda := n \sum_{i=1}^{h} \hat{\rho}(i)^2.$$

The null hypothesis is rejected if $\lambda > \chi^2_{1-\alpha,h}$, where $\chi^2_{1-\alpha,h}$ denotes the α-quantile of the χ^2 distribution with h degrees of freedom.

This classical test, which originates to Box and Pierce in 1970, has been modified by Ljung and Box in 1978 and it has been shown that it performs better especially also for small sample sizes (of less than 100 elements). In what follows the modified test statistic is given.

Method 2.2.10 (Ljung–Box test)**.** This test is a modification of the Portmanteau test. Use instead the test statistic

$$n(n+2) \sum_{i=1}^{h} \frac{\hat{\rho}(i)^2}{n-i}, \tag{2.1}$$

which is asymptotically χ^2_h-distributed for iid random variables. Use the same rejection regions as in the Portmanteau test 2.2.9.

Example 2.2.11. Let us revisit the data from Example 2.1.2, i.e., the set $(x_t)_{t=1}^{42}$ of quarterly earnings of H&M. We apply the Ljung–Box test (Method 2.2.10) with $h = 4$ and $\alpha = 0.05$ to this data set to see whether it is likely that the earnings are a realization of white noise, or more specifically whether the autocorrelation function satisfies that $\rho_X(i) = 0$ for $i = 1, 2, 3, 4$. Recall that the sample autocorrelation function was computed in Example 2.1.2. The test statistic (2.1) is found to be 38.87 which is much greater than the critical value $\chi^2_{0.95,4} = 9.49$. Therefore we *reject* the null hypothesis of no autocorrelation for lags $i = 1, 2, 3, 4$ at the 5% level, something that is not surprising given Figure 2.1.2.

More tests like the turning point test, the difference sign test, and the rank test are available but not treated in these lecture notes. For those the reader is referred to [4, Section 1.6].

2.3 Forecasting stationary time series

The goal of forecasting a stationary time series with known mean μ and autocovariance function γ is to predict $(X_{n+h}, h > 0)$ in terms of $(X_t, t = 1, \dots, n)$. We will find best predictors in the sense of minimal mean squared errors. To that end, let us start with the necessary definitions.

Definition 2.3.1. Let X and Y be random variables and let Y be an approximation of X. The *mean squared error* of Y is defined by

$$\mathrm{MSE}(Y, X) := \mathbb{E}((Y - X)^2).$$

Note that the mean squared error is one (very popular) way to measure the error of a prediction and that one could think of many other "measures". This choice of error measure influences essentially the following analysis and definition of *best*. We will nevertheless restrict the forecasting to optimization with respect to the mean squared error since it is the usual choice and widely used.

Definition 2.3.2. Let $(X_t, t \in \mathbb{Z})$ be a time series and $X^n := (X_{t_1}, \ldots, X_{t_n})$ a collection of random variables of the time series at n different times. Then the function of X^n denoted by $b_t(X^n)$ is called a *best predictor* of X_t for some $t \neq t_j$, $j = 1, \ldots, n$, if it minimizes the mean squared error, i.e.,

$$b_t(X^n) := \operatorname*{arg\,min}_{g(X^n)} \mathrm{MSE}(g(X^n), X_t) = \operatorname*{arg\,min}_{g(X^n)} \mathbb{E}((g(X^n) - X_t)^2),$$

i.e., $g : \mathbb{R}^n \to \mathbb{R}$.

In the following proposition we show that a best predictor exists theoretically and that it is unique.

Proposition 2.3.3. *Let $(X_t, t \in \mathbb{Z})$ be a time series and $X^n := (X_{t_1}, \ldots, X_{t_n})$ a collection of random variables of the time series at n different times. Then the best predictor of X_t for some $t \neq t_j$, $j = 1, \ldots, n$, is the conditional expectation of X_t given X^n, i.e.,*

$$b_t(X^n) = \mathbb{E}(X_t | X^n).$$

Proof. To prove that the conditional expectation is a best predictor of X_t, let us first observe that

$$\begin{aligned}
\mathbb{E}((g(X^n) - X_t)^2) &= \mathbb{E}((g(X^n) - \mathbb{E}(X_t|X^n) + \mathbb{E}(X_t|X^n) - X_t)^2) \\
&= \mathbb{E}((g(X^n) - \mathbb{E}(X_t|X^n))^2) + \mathbb{E}(\mathbb{E}(X_t|X^n) - X_t)^2) \\
&\quad + 2\,\mathbb{E}((g(X^n) - \mathbb{E}(X_t|X^n))(\mathbb{E}(X_t|X^n) - X_t)).
\end{aligned}$$

We show next that the last term is equal to zero. To do this, we use the properties of the conditional expectation. We obtain by Property (ii) and since $g(X^n)$ and $\mathbb{E}(X_t|X^n)$ are both measurable functions of X^n with Property (v) that

$$\begin{aligned}
\mathbb{E}((g(X^n) - \mathbb{E}(X_t|X^n))(\mathbb{E}(X_t|X^n) - X_t)) &= \mathbb{E}(\mathbb{E}((g(X^n) - \mathbb{E}(X_t|X^n))(\mathbb{E}(X_t|X^n) - X_t)|X^n)) \\
&= \mathbb{E}((g(X^n) - \mathbb{E}(X_t|X^n))\,\mathbb{E}(\mathbb{E}(X_t|X^n) - X_t|X^n)).
\end{aligned}$$

Next the linearity of the conditional expectation Property (i) implies together with the measurability of $\mathbb{E}(X_t|X^n)$ that

$$\mathbb{E}(\mathbb{E}(X_t|X^n) - X_t|X^n) = \mathbb{E}(\mathbb{E}(X_t|X^n)|X^n) - \mathbb{E}(X_t|X^n) = \mathbb{E}(X_t|X^n) - \mathbb{E}(X_t|X^n) = 0.$$

Putting these results together we have just shown that

$$\mathbb{E}((g(X^n) - \mathbb{E}(X_t|X^n))(\mathbb{E}(X_t|X^n) - X_t)) = 0.$$

Therefore we have transformed our minimization problem to

$$\begin{aligned}
\min_{g(X^n)} \mathbb{E}((g(X^n) - X_t)^2) &= \min_{g(X^n)} \left(\mathbb{E}((g(X^n) - \mathbb{E}(X_t|X^n))^2) + \mathbb{E}(\mathbb{E}(X_t|X^n) - X_t)^2)\right) \\
&= \mathbb{E}(\mathbb{E}(X_t|X^n) - X_t)^2) + \min_{g(X^n)} \mathbb{E}((g(X^n) - \mathbb{E}(X_t|X^n))^2).
\end{aligned}$$

Due to the positivity of squares, it is clear that

$$\min_{g(X^n)} \mathbb{E}((g(X^n) - X_t)^2) \geq \mathbb{E}(\mathbb{E}(X_t|X^n) - X_t)^2).$$

By choosing $g(X^n) := \mathbb{E}(X_t|X^n)$ we therefore obtain a minimum, which finishes the proof for the existence of a minimum.

Uniqueness (in P-a.s. sense) follows since the minimum in the previous computation is just attained if

$$\mathbb{E}((g(X^n) - \mathbb{E}(X_t|X^n))^2) = 0,$$

i.e., if $g(X^n) = \mathbb{E}(X_t|X^n)$ in mean square and therefore also P-almost surely by Chebyshev's inequality (1.1). □

We have just seen that the conditional expectation is the best predictor with respect to the mean squared error. It remains to see how we compute its value in practice if we are given a finite set of observations, e.g., of an asset, that we want to use to predict future values as accurate as possible. Since the conditional expectation is not necessarily linear and computable in closed form, we restrict ourselves next to linear predictors.

Definition 2.3.4. Let $(X_t, t \in \mathbb{Z})$ be a time series and $X^n := (X_{t_1}, \ldots, X_{t_n})$ a collection of random variables of the time series at n different times. Then the linear function of 1 and X^n denoted by $b_t^l(X^n)$ is called a *best linear predictor* of X_t for some $t \neq t_j$, $j = 1, \ldots, n$, if it minimizes the mean squared error, i.e.,

$$b_t^l(X^n) := \operatorname*{arg\,min}_{g(X^n)} \mathrm{MSE}(g(X^n), X_t) = \operatorname*{arg\,min}_{g(X^n)} \mathbb{E}((g(X^n) - X_t)^2),$$

where g is a linear function of 1 and X^n, i.e., there exist $a_0, \ldots, a_n$ such that $g(X^n) := a_0 + a_1 X_{t_n} + a_2 X_{t_{n-1}} + \cdots + a_n X_{t_1}$.

Let us now derive the coefficients $(a_i, i = 0, \ldots, n)$ explicitly for a stationary time series with mean μ and autocovariance function γ, which automatically also shows the existence of the minimum. From calculus we know that we obtain an extremum of a (sufficiently smooth) function by differentiation. Therefore set

$$S(a) := \mathbb{E}((a_0 + a_1 X_{t_n} + \cdots + a_n X_{t_1} - X_t)^2)$$

with $a = (a_0, \ldots, a_n)$, which is a positive and quadratic function in terms of the coefficients and bounded from below by zero. Therefore at least one minimum exists. To find it explicitly, we compute for $j = 1, \ldots, n$

$$\frac{\partial S(a)}{\partial a_j} = 2\,\mathbb{E}(X_{t_{n+1-j}}(a_0 + a_1 X_{t_n} + \cdots + a_n X_{t_1} - X_t)) \tag{2.2}$$

as well as

$$\frac{\partial S(a)}{\partial a_0} = 2a_0 + 2\,\mathbb{E}(a_1 X_{t_n} + \cdots + a_n X_{t_1} - X_t). \tag{2.3}$$

By setting the last equation equal to zero and using the stationarity of the time series, we derive that

$$a_0 = \mu \left(1 - \sum_{i=1}^{n} a_i\right). \tag{2.4}$$

For $j = 1, \ldots, n$, we obtain with the definition of the autocovariance function and by setting the derivatives equal to zero that

$$a_j(\gamma(0)+\mu^2)+a_0\mu+a_1(\gamma(t_n-t_{n+1-j})+\mu^2)+\cdots+a_n(\gamma(t_1-t_{n+1-j})+\mu^2)-(\gamma(t-t_{n+1-j})+\mu^2) = 0,$$

which simplifies with (2.4) to

$$a_j\gamma(0) + a_1\gamma(t_n - t_{n+1-j}) + \cdots + a_n\gamma(t_1 - t_{n+1-j}) = \gamma(t - t_{n+1-j}).$$

Combining all $j = 1, \ldots, n$, we can rewrite the last equation in matrix vector notation as

$$\Gamma_n(a_1, \ldots, a_n)' = (\gamma(t - t_n), \ldots, \gamma(t - t_1))',$$

where

$$\Gamma_n = (\gamma(t_{n+1-j} - t_{n+1-i}))_{i,j=1}^n.$$

The solution of this system of equations is a minimum since S is a quadratic function bounded from below by zero.

To show uniqueness let $(a_j^{(1)}, j = 0, \ldots, n)$ and $(a_j^{(2)}, j = 0, \ldots, n)$ be two different solutions and denote by Z the difference between the two resulting predictors, i.e.,

$$Z := a_0^{(1)} - a_0^{(2)} + (a_1^{(1)} - a_1^{(2)})X_{t_n} + (a_2^{(1)} - a_2^{(2)})X_{t_{n-1}} + \cdots + (a_n^{(1)} - a_n^{(2)})X_{t_1}.$$

Then

$$\mathbb{E}(Z) = 0 + \mathbb{E}(X_t) - (0 + \mathbb{E}(X_t)) = 0$$

by (2.3) and

$$\mathbb{E}(ZX_{t_{n+1-j}}) = 0$$

for all $j = 1, \ldots, n$ by (2.2), which implies that

$$\begin{aligned}\mathbb{E}(Z^2) &= \mathbb{E}(Z(a_0^{(1)} - a_0^{(2)} + (a_1^{(1)} - a_1^{(2)})X_{t_n} + (a_2^{(1)} - a_2^{(2)})X_{t_{n-1}} + \cdots + (a_n^{(1)} - a_n^{(2)})X_{t_1})) \\ &= (a_0^{(1)} - a_0^{(2)})\,\mathbb{E}(Z) \\ &= 0.\end{aligned}$$

Therefore the mean squared error of the difference is zero and the predictors are (P-almost surely) the same.

In conclusion we have shown the following proposition:

Proposition 2.3.5. *Let $(X_t, t \in \mathbb{Z})$ be a stationary time series with mean μ and autocovariance function γ and $X^n := (X_{t_1}, \ldots, X_{t_n})$ a collection of random variables of the time series at n different times. Then the best linear predictor of X_t is given by*

$$b_t^l(X^n) = a_0 + a_1X_{t_n} + a_2X_{t_{n-1}} + \cdots + a_nX_{t_1},$$

where the coefficients $(a_i, i = 0, \ldots, n)$ are determined by the linear equations

$$a_0 = \mu\left(1 - \sum_{i=1}^n a_i\right)$$

and

$$\Gamma_n(a_1, \ldots, a_n)' = (\gamma(t - t_n), \ldots, \gamma(t - t_1))'$$

with

$$\Gamma_n = (\gamma(t_{n+1-j} - t_{n+1-i}))_{i,j=1}^n.$$

Let us finally remark that in the case that $X^n := (X_1, \ldots, X_n)$ the equations to derive the coefficients $(a_0, \ldots, a_n)$ for the prediction of X_{n+h} simplify to

$$a_0 = \mu\left(1 - \sum_{i=1}^n a_i\right)$$

and

$$(\gamma(i - j))_{i,j=1}^n(a_1, \ldots, a_n)' = (\gamma(h), \ldots, \gamma(h + n - 1))'.$$

In the following proposition we collect some properties of the best linear predictor.

Proposition 2.3.6. *Let $(X_t, t \in \mathbb{Z})$ be a stationary time series with mean μ and autocovariance function γ and $X^n := (X_{t_1}, \dots, X_{t_n})$ a collection of random variables of the time series at n different times. Then the best linear predictor $b_t^l(X^n)$ satisfies:*

(i) $\mathbb{E}(X_t - b_t^l(X^n)) = 0$, *i.e., it is unbiased,*

(ii) $\mathrm{MSE}(b_t^l(X^n), X_t) = \mathbb{E}((b_t^l(X^n) - X_t)^2) = \gamma(0) - (a_1, \dots, a_n)(\gamma(t-t_n), \dots, \gamma(t-t_1))'$,

(iii) $\mathbb{E}(X_{t_j}(X_t - b_t^l(X^n))) = 0$ *for all* $j = 1, \dots, n$.

Proof. The first claim follows by construction from (2.3). To compute the mean squared error in the second claim we observe that

$$\begin{aligned}\mathrm{MSE}(b_t^l(X^n), X_t) &= 2a_0\,\mathbb{E}(a_0 + a_1 X_{t_n} + \cdots + a_n X_{t_1} - X_t) - a_0^2 \\ &\quad + \sum_{i=1}^n a_i \sum_{j=1}^n \mathbb{E}(X_{t_{n+1-i}} X_{t_{n+1-j}}) a_j - 2\sum_{i=1}^n a_i\,\mathbb{E}(X_{t_{n+1-i}} X_t) + \mathbb{E}(X_t^2) \\ &= 0 - a_0^2 + (a_1, \dots, a_n)(\Gamma_n(a_1, \dots, a_n)' - 2(\gamma(t-t_n), \dots, \gamma(t-t_1))') \\ &\quad + \mu^2 \sum_{i,j=1}^n a_i a_j - 2\mu^2 \sum_{i=1}^n a_i + \gamma(0) + \mu^2.\end{aligned}$$

Due to Proposition 2.3.5 we derive that

$$(a_1, \dots, a_n)(\Gamma_n(a_1, \dots, a_n)' - 2(\gamma(t-t_n), \dots, \gamma(t-t_1))') = -(a_1, \dots, a_n)(\gamma(t-t_n), \dots, \gamma(t-t_1))'.$$

Additionally we obtain with the same proposition—more specifically with (2.4)—that

$$\begin{aligned}\mu^2 \sum_{i,j=1}^n a_i a_j - 2\mu^2 \sum_{i=1}^n a_i + \mu^2 &= \mu^2 \left(\sum_{i=1}^n a_i \Big(\sum_{j=1}^n a_j - 1 \Big) + 1 - \sum_{i=1}^n a_i \right) = a_0 \mu \left(- \sum_{i=1}^n a_i + 1 \right) \\ &= a_0^2.\end{aligned}$$

Therefore we conclude that

$$\mathrm{MSE}(b_t^l(X^n), X_t) = \gamma(0) - (a_1, \dots, a_n)(\gamma(t-t_n), \dots, \gamma(t-t_1))',$$

which proves the second claim. The third claim follows by construction from (2.2). □

To see an application of the theory, let us treat two examples in what follows.

Example 2.3.7 (AR(1))**.** Let us assume that the stationary time series model is given by

$$X_t - \phi_1 X_{t-1} = Z_t,$$

where $(Z_t, t \in \mathbb{Z})$ is a $\mathrm{WN}(0, \sigma^2)$ process and $|\phi_1| < 1$. This model will be called an AR(1) model in the framework of Chapter 3. Then we first compute the autocovariance function which is given by

$$\gamma_X(0) = \mathbb{E}(X_t^2) = \mathbb{E}((Z_t + \phi_1 X_{t-1})^2) = \sigma^2 + \phi_1^2 \gamma_X(0),$$

since Z is a white noise and therefore Z_t and X_{t-1} are uncorrelated. This implies that

$$\gamma_X(0) = \frac{\sigma^2}{1 - \phi_1^2}$$

and

$$\gamma_X(h) = \mathbb{E}(X_t X_{t+h}) = \mathbb{E}(X_t(Z_{t+h} + \phi_1 X_{t+h-1})) = \phi_1 \gamma_X(h-1)$$

with $h > 0$ for the same reasons as before. Solving the recursion leads to

$$\gamma_X(h) = \frac{\sigma^2 \phi_1^{|h|}}{1 - \phi_1^2}.$$

The best linear predictor $b^l_{n+1}(X^n)$ with respect to $X^n := (X_1, \ldots, X_n)$ is then (noting that the process has zero mean which implies $a_0 = 0$) by Proposition 2.3.5 $\sum_{i=1}^n a_i X_{n+1-i}$, where the coefficients a_i are determined by the solution of the system of linear equations

$$\begin{pmatrix} 1 & \phi_1 & \phi_1^2 & \cdots & \phi_1^{n-1} \\ \phi_1 & 1 & \phi_1 & \cdots & \phi_1^{n-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi_1^{n-1} & \phi_1^{n-2} & \phi_1^{n-3} & \cdots & 1 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} \phi_1 \\ \phi_1^2 \\ \vdots \\ \phi_1^n \end{pmatrix}.$$

It is clear that $a_1 = \phi_1$ and $a_i = 0$, $i = 2, \ldots, n$, solves the system of linear equations and therefore that the best linear predictor of X_{n+1} is

$$b^l_{n+1}(X^n) = \phi_1 X_n$$

with mean squared error σ^2, which should be computed in an *exercise*.

Example 2.3.8 (MA(1))**.** Let us assume that the stationary time series model is given by

$$X_t = Z_t + \theta_1 Z_{t-1},$$

where $(Z_t, t \in \mathbb{Z})$ is a WN$(0, \sigma^2)$ process. This model will be called a MA(1) model in the framework of Chapter 3.

It is clear that the mean of the series is $\mu = 0$. Furthermore we get that

$$\gamma_X(0) = (1 + \theta_1^2)\sigma^2$$

and

$$\gamma_X(1) = \theta_1 \sigma^2$$

as well as $\gamma_X(h) = 0$ for all $|h| > 1$ by easy computations.

The best linear predictor $b^l_{n+1}(X^n)$ with respect to $X^n := (X_1, \ldots, X_n)$ is again by Proposition 2.3.5 $\sum_{i=1}^n a_i X_{n+1-i}$. Note that $a_0 = 0$ due to the fact that the series has zero mean. We find the coefficients a_i with Proposition 2.3.6(iii). The equation

$$\mathbb{E}[(X_{n+1} - \sum_{i=1}^n a_i X_{n+1-i}) X_n] = 0$$

simplifies by use of γ_X to

$$\theta_1 = (1 + \theta_1^2)a_1 + \theta_1 a_2.$$

Similarly for $2 \leq j \leq n-1$ we obtain

$$0 = \theta_1 a_{j-1} + (1 + \theta_1^2)a_j + \theta_1 a_{j+1}$$

and also

$$0 = \theta_1 a_{n-1} + (1 + \theta_1^2)a_n.$$

It is clear that these difference equations uniquely defines the coefficients $a_1, \ldots, a_n$ and they can in fact for $1 \leq j \leq n-1$ be shown (cf. [4, Problem 3.12]) to be

$$a_j = \frac{1 + \theta_1^2 + \ldots + \theta_1^{2(n-j)}}{(-\theta_1)^{(n-j)}} a_n$$

with

$$a_n = \frac{-(-\theta_1)^n}{1+\theta_1^2+\ldots+\theta_1^{2n}}.$$

Given these coefficients, we can use Proposition 2.3.6(ii) to compute the mean squared error

$$\begin{aligned}\mathbb{E}((b^l_{n+1}(X^n)-X_{n+1})^2) &= \gamma(0)-(a_1,\ldots,a_n)(\gamma_X(1),\ldots,\gamma_X(n))'\\ &= (1+\theta_1^2)\sigma^2-a_1\theta_1\sigma^2,\end{aligned}$$

which finishes this example.

Example 2.3.9 (AR(1) with missing value)**.** Let us consider the AR(1) model from Example 2.3.7 again. Assume that we have observed X_1 and X_3 but that we are missing X_2. Then the best linear predictor $b^l_2((X_1,X_3))$ of X_2 is $a_1X_3+a_2X_1$ by Proposition 2.3.5, where the coefficients a_1 and a_2 solve the system of linear equations

$$\begin{pmatrix}1 & \phi_1^2\\ \phi_1^2 & 1\end{pmatrix}\begin{pmatrix}a_1\\ a_2\end{pmatrix}=\begin{pmatrix}\phi_1\\ \phi_1\end{pmatrix}.$$

An easy computation shows that

$$a_1=a_2=\frac{\phi_1}{1+\phi_1^2}$$

is a solution and therefore that the best linear predictor is

$$b^l_2((X_1,X_3))=\frac{\phi_1}{1+\phi_1^2}(X_3+X_1)$$

with mean squared error $\sigma^2/(1+\phi_1^2)$.

We remark that due to the linearity properties of the prediction operator, it is sufficient to consider just stationary time series with mean zero.

We have stated in Proposition 2.3.5 that there exists a unique solution that is the best linear predictor, but this involves solving a system of n linear equations. For large n this might be difficult and especially time consuming. To save computational time, we will introduce two algorithms in what follows that use a recursive approach, i.e., $b^l_{n+1}(X^n)$ is used to compute $b^l_{n+2}(X^{n+1})$ in a cheaper way, where $X^n:=(X_1,\ldots,X_n)$ for all $n\in\mathbb{N}$. To turn the algorithms into something readable we adapt the notation for changing sizes of matrices. In what follows let the best linear one-step estimator be given by

$$b^l_{n+1}(X^n)=\sum_{i=1}^n a_{ni}X_{n+1-i},$$

where $a_{ni}:=a_i$ in terms of the previously used notation, i.e., the coefficients are determined by the solution of the system of linear equations given in Proposition 2.3.5.

One way to compute the estimators more efficiently all at once is the Durbin–Levinson algorithm, which is introduced next. For a proof the reader is referred to [4, Section 2.5.1]. For this algorithm to work, one needs that the coefficients a_{ni} are uniquely determined at each time n, i.e., that $\Gamma_n=(\gamma(i-j))_{i,j=1}^n$ is non-singular for every n. A sufficient condition for this is the requirement that $\gamma(0)>0$ and $\gamma(n)\to 0$ as $n\to\infty$, which holds for many time series used in applications. For a proof the reader is referred to [3, Proposition 5.1.1].

Method 2.3.10 (Durbin–Levinson algorithm)**.** Compute the coefficients $a_{n1},\ldots,a_{nn}$ recursively from the equations

$$a_{nn}:=\left(\gamma(n)-\sum_{i=1}^{n-1}a_{(n-1)i}\gamma(n-i)\right)v_{n-1}^{-1},$$

$$\begin{pmatrix} a_{n1} \\ \vdots \\ a_{n(n-1)} \end{pmatrix} := \begin{pmatrix} a_{(n-1)1} \\ \vdots \\ a_{(n-1)(n-1)} \end{pmatrix} - a_{nn} \begin{pmatrix} a_{(n-1)(n-1)} \\ \vdots \\ a_{(n-1)1} \end{pmatrix},$$

and

$$v_n := v_{n-1}(1 - a_{nn}^2),$$

where $a_{11} = \gamma(1)/\gamma(0)$ and $v_0 := \gamma(0)$.

Observe that v_n computes the mean squared error of $b_{n+1}^l(X^n)$, which can be seen from the proof of the algorithm.

A second algorithm is the so-called *innovations algorithm*, which can be applied to all time series with finite second moments, i.e., stationarity is not a requirement. Therefore let us consider the more general framework that $(X_t, t \in \mathbb{Z})$ is a time series with mean zero, $\mathbb{E}(X_t^2) < +\infty$ for all $t \in \mathbb{Z}$, and covariance

$$\mathsf{Cov}(X_i, X_j) = \mathbb{E}(X_i X_j) = \kappa(i, j).$$

For convenience let us use the following notation for the best linear one-step predictors

$$\hat{X}_n := \begin{cases} 0 & \text{for } n = 1, \\ b_n^l(X^{n-1}) & \text{for } n > 1, \end{cases}$$

and the mean squared errors

$$v_n := \mathrm{MSE}(\hat{X}_{n+1}, X_{n+1}) = \mathbb{E}((\hat{X}_{n+1} - X_{n+1})^2).$$

One can show that the best linear predictors satisfy

$$\hat{X}_{n+1} = \begin{cases} 0 & \text{for } n = 0, \\ \sum_{j=1}^n \theta_{nj}(X_{n+1-j} - \hat{X}_{n+1-j}) & \text{for } n \geq 1 \end{cases}$$

for some coefficients θ_{ij}, $i, j = 1, \ldots, n$. The innovations algorithm generates these coefficients and the mean squared errors $v_j = \mathrm{MSE}(\hat{X}_{j+1}, X_{j+1})$ recursively. For a proof of the method the reader is referred to [3, Proposition 5.2.2].

Method 2.3.11 (Innovations algorithm)**.** Compute the coefficients $\theta_{n1}, \ldots, \theta_{nn}$ recursively from the equations

$$v_0 := \kappa(1, 1)$$

and

$$\theta_{n(n-k)} := v_k^{-1} \left(\kappa(n+1, k+1) - \sum_{j=0}^{k-1} \theta_{k(k-j)} \theta_{n(n-j)} v_j \right)$$

for $0 \leq k < n$ and

$$v_n := \kappa(n+1, n+1) - \sum_{j=0}^{n-1} \theta_{n(n-j)}^2 v_j.$$

2.4 Trend and seasonality

One possible treatment of data is to assume that the data set is a realization of the stochastic process X that can be split into

$$X_t = m_t + s_t + Y_t, \tag{2.5}$$

i.e., it follows the *classical decomposition model*. Here $m : \mathbb{Z} \to \mathbb{R}$ is a slowly changing function called the *trend component*, $s : \mathbb{Z} \to \mathbb{R}$ is a function with known period d referred to as the *seasonal component*, i.e., $s_{t+d} = s_t$ and $\sum_{j=1}^d s_j = 0$, and $Y = (Y_t, t \in \mathbb{Z})$ is a stationary time series.

The aim of this section is to estimate and extract the deterministic functions m and s such that the remaining stochastic process Y becomes hopefully a stationary time series.

2.4.1 Trend in absence of seasonality

Assume in this section that the stochastic process $X = (X_t, t \in \mathbb{Z})$ is given by

$$X_t = m_t + Y_t.$$

Without loss of generality let $\mathbb{E}(Y_t) = 0$ for all $t \in \mathbb{Z}$.

In what follows we first introduce a collection of methods for trend estimation before we give a method that does trend elimination directly without estimation. It should be mentioned that seasonal components are hard to find in financial data and that it is therefore reasonable to treat models without seasonal component but an overall trend. In all of the methods let $X^n := (X_1, \ldots, X_n)$, $n \in \mathbb{N}$, denote the finite number of random variables from the time series with available observed data.

Method 2.4.1 (Estimation by smoothing with a finite moving average filter)**.** Let $q \in \mathbb{N}$ with $2q < n$ be fixed. Compute for all $t = q+1, \ldots, n-q$ the *two-sided moving average*

$$W_t := (2q+1)^{-1} \sum_{j=-q}^{q} X_{t-j}$$

of X. Then by the definition of X, it holds that

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

if we assume that q is sufficiently small such that $(m_s, s = t-q, \ldots, t+q)$ is approximately linear and that the average of the error terms $(Y_s, s = t-q, \ldots, t+q)$ is close to zero. The details for the validity of the assumptions are left to the reader as an *exercise*.

The moving average therefore leads to the estimator

$$\hat{m}_t := (2q+1)^{-1} \sum_{j=-q}^{q} X_{t-j}$$

for $q+1 \leq t \leq n-q$. Observe that this method does not lead to estimates of m_t for all $t = 1, \ldots, n$ but only at the "inner" time points, where the definition of "inner" depends on the choice of q.

Method 2.4.2 (Estimation by exponential smoothing)**.** For any fixed $\alpha \in [0,1]$ define the *one-sided moving averages* $(\hat{m}_t, t = 1, \ldots, n)$ by the recursion

$$\hat{m}_t := \alpha X_t + (1-\alpha)\hat{m}_{t-1}$$

for $t = 2, \ldots, n$ and

$$\hat{m}_1 := X_1.$$

The method is referred to as *exponential smoothing* since the recursion implies for $t \geq 2$ that

$$\hat{m}_t = \sum_{j=0}^{t-2} \alpha(1-\alpha)^j X_{t-j} + (1-\alpha)^{t-1} X_1,$$

which is a weighted moving average of X with exponentially decreasing weights.

Method 2.4.3 (Estimation by polynomial fitting)**.** Assume that the trend m is given by the polynomial

$$m_t := a_0 + a_1 t + a_2 t^2,$$

then the coefficients a_0, a_1, and a_2 are obtained by the *least square minimization*

$$\min_{a_0,a_1,a_2} \sum_{t=1}^{n} (x_t - m_t)^2,$$

where $(x_t, t = 1, \ldots, n)$ is the series of observed data.

Similarly one could also use higher-order polynomials for the estimation of the trend, i.e.,

$$m_t := \sum_{j=0}^{q} a_j t^j$$

for some $q \in \mathbb{N}$ with $q < n$ and do a least square minimization of that.

Method 2.4.4 (Elimination by differencing)**.** Define the *difference operator* ∇ by

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

for $t \geq 2$, where B denotes the *backward shift operator* given by

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

Powers of B are defined by

$$B^j X_t = B^{j-1} B X_t = B^{j-1} X_{t-1} = \cdots = X_{t-j}$$

for $j < t$. Similarly we obtain

$$\nabla^j X_t = \nabla \nabla^{j-1} X_t$$

for $j < t$, e.g.,

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

Assume that the trend m is given by the polynomial

$$m_t := \sum_{j=0}^{q} a_j t^j$$

for some $q < n$. Then one shows in an *exercise* that

$$\nabla^q m_t = q!\, a_q.$$

(*Hint: Start with* $q = 1$ *and* ∇m_t.)

Coming back to the stochastic process X, one obtains that

$$\nabla^q X_t = q!\, a_q + \nabla^q Y_t.$$

Since Y is assumed to be a stationary process with mean zero, one can show in an *exercise* that the same holds for $\nabla^q Y$. This implies that $\nabla^q X$ is a mean $q!\, a_q$, stationary process, which is left to the reader as an *exercise*.

In reality it is often sufficient to consider $q = 1$ or $q = 2$.

2.4.2 Trend and seasonality in parallel

Let us come back to the classical decomposition model (2.5). We recall that

$$X_t = m_t + s_t + Y_t,$$

where $\mathbb{E}(Y_t) = 0$ for all $t \in \mathbb{Z}$, $s_{t+d} = s_t$, and $\sum_{j=1}^{d} s_j = 0$. Let us assume for simplicity that the size of the observed data n covers a multiple of the period d, i.e., $n/d \in \mathbb{N}$. Typical periods are 24 hours per day, 7 days per week, 12 months per year, or 4 quarters per year.

In what follows we introduce two methods to remove seasonal components in connection with the estimation of the trend. While the first methods "S1" is in analogy with Method 2.4.1, the second one called "S2" uses differencing which was already exploited for the trend in Method 2.4.4.

Method 2.4.5 ("S1": estimation by moving averages)**.** Let us assume that we are given observations $(x_t, t = 1, \ldots, n)$ of the stochastic process X. We start with the estimation of the trend by applying a moving average filter that eliminates the seasonal component and dampens the noise. For an even period $d := 2q$ we set

$$\hat{m}_t := d^{-1}(2^{-1}x_{t-q} + x_{t-q+1} + \cdots + x_{t+q-1} + 2^{-1}x_{t+q})$$

for $q < t \leq n - q$. Similarly we set for an odd period $d := 2q + 1$

$$\hat{m}_t := d^{-1} \sum_{j=-q}^{q} x_{t-j}.$$

To estimate the seasonal component, we average over the trend eliminated series elements with the same seasonal component, i.e., we set for $k = 1, \ldots, d$ and $q < k + jd \leq n - q$

$$w_k := |\{j \in \mathbb{N}_0, q < k + jd \leq n - q\}|^{-1} \sum_{q<k+jd\leq n-q} (x_{k+jd} - \hat{m}_{k+jd}),$$

where $|A|$ denotes the size of a finite set A, i.e., its number of elements. To satisfy the condition of the model that $\sum_{j=1}^{d} s_j = 0$, we have to modify the w_k's to obtain a valid seasonal component $\hat{s}$ by setting the components

$$\hat{s}_k := w_k - d^{-1} \sum_{j=1}^{d} w_j$$

for $k = 1, \ldots, d$ and $\hat{s}_k := \hat{s}_{k-d}$ for $k > d$.

Finally, we reestimate the trend by applying one of the methods from Section 2.4.1 to the deseasonalized series $(x_t - \hat{s}_t, t = 1, \ldots, n)$. The reestimation of the trend is only done to obtain a trend estimator in parametric form which is convenient for prediction and simulation.

Example 2.4.6. We apply Method 2.4.5 to the data of Example 2.1.2. As we noted in Example 2.2.7, there is evidence of a seasonal component with period $d = 4$. We also choose to fit a linear trend to the data. The method is applied with the following MATLAB code:

```
% Input:
%   data double vec, data.
%   time double vec, time indices
%   period int, length of period
%
% Output:
%    season double vec, row vector of season, length=length(DATA
    )
```

```
%    trend double vec, row vector of coefficients of trend,
    length=2
n=length(data);
%% Preestimation of trend
q=floor(period/2);
preltrend=zeros(1,n-2*q);
if(q==floor(period/2)) % Even period
    for i=1:n-2*q
        preltrend(i)=sum([data(i)/2,data(i+1:i+2*q-1),data(i+2*q
            )/2])/2/q;
    end
else % Odd period
    for i=1:n-2*q
        preltrend=mean(data(i:i+2*q));
    end
end

%% Estimation of seasonality
season=zeros(1,period);
for i=1:period
    if(i<=q)
        season(i)=mean(data(i+period:period:n-q)-preltrend(i+
            period-q:period:end));
    else
        season(i)=mean(data(i:period:n-q)-preltrend(i-q:period:
            end));
    end
end
season=season-mean(season); % Sum to zero
season=repmat(season,1,floor(n/period));
season(n-mod(n,period)+1:n)=season(1:mod(n,period)); % Extend
    season to full data set

%% Reestimation of linear trend by least square
deseasondata=data-season;
X = ones(n,2);
X(:,2) = time;
trend = X\deseasondata';
```

The estimated trend and seasonality $\hat{m}_t + \hat{s}_t$ are shown along with the original data x_t in Figure 2.4.6.

Method 2.4.7 ("S2": elimination by differencing)**.** The method of differencing that was introduced in Method 2.4.4 for time series without seasonal component can be adapted to the general classical decomposition model by introducing the *lag-d differencing operator* ∇_d for a period d. It is defined by

$$\nabla_d X_t := X_t - X_{t-d} = (1 - B^d)X_t.$$

Applying this operator to the model, we obtain that

$$\nabla_d X_t = m_t - m_{t-d} + s_t - s_{t-d} + Y_t - Y_{t-d} = \nabla_d m_t + \nabla_d Y_t$$

due to the periodicity of s. One shows in an easy *exercise* that $\nabla_d Y$ is a stationary time series, which implies that $\nabla_d X$ is a stochastic process without seasonal component. The methods introduced in Section 2.4.1 can be applied to estimate or eliminate the differenced trend component $\nabla_d m$.

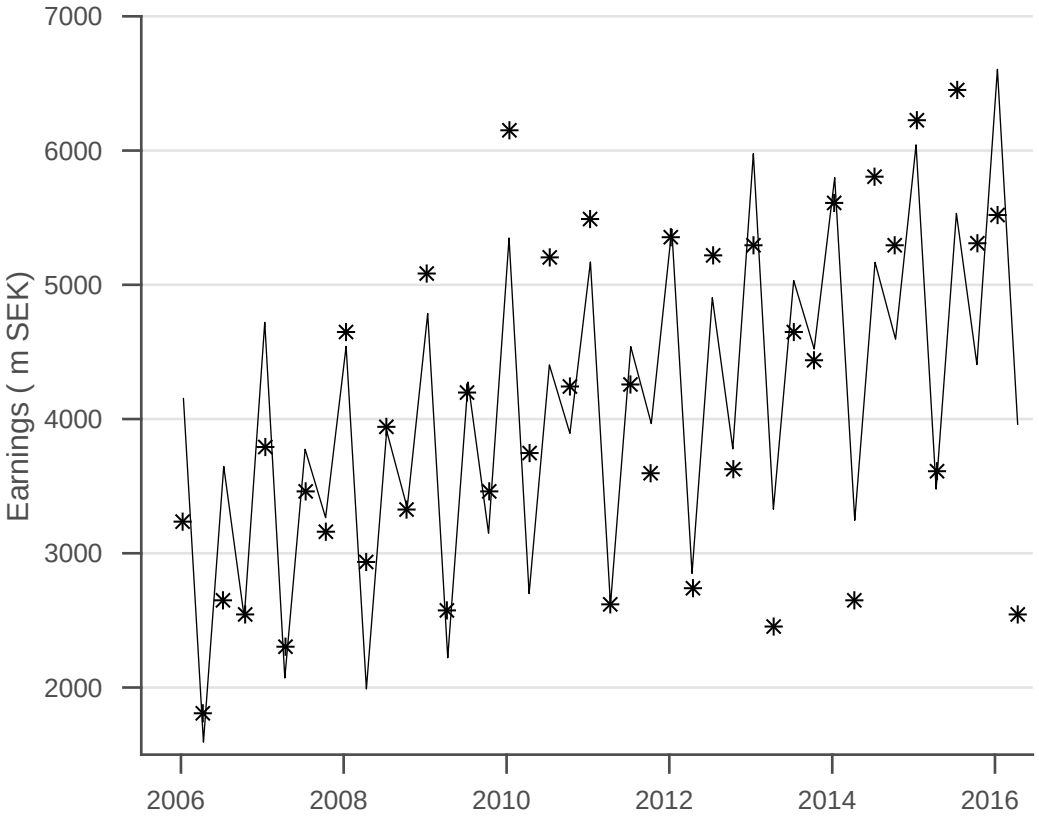

Figure 2.3: The H&M data of Example 2.1.2 (stars) with an estimated linear trend and seasonal component with period 4, $\hat{m}_t + \hat{s}_t$ (line).

Once we have removed the trend and the seasonal components from the time series, we have to test the remaining sequence for stationarity. We could start with the methods given in Section 2.2. If we observe that the remaining data is iid, we are completely done. If we do not succeed, we should test for different stationary models, which will be introduced in the next chapter.

Overall one should remark that we assumed that the period d is known. In practice we have to guess d, which could for example be done by choosing "reasonable" periods (for example, a seasonal component with period 12 for monthly sales by a clothing company), by looking at the plotted graph, or by testing the remaining series for stationarity for different d.

CHAPTER 3

Linear time series models

In this chapter we consider linear time series models, where we focus mainly on ARMA models. For those we discuss parameter estimation, order selection as well as forecasting methods. The chapter is finished with an extension of the ARMA model to ARIMA models.

3.1 Linear processes

Before we introduce the specific class of ARMA models, let us consider the more general class of linear processes and its properties. Let us assume that $X = (X_t, t \in \mathbb{Z})$ is a stochastic process in discrete time in what follows. We recall that we use stochastic process and time series as synonyms.

Definition 3.1.1. A stochastic process X is called a *linear process* if it has the representation

$$X_t = \sum_{j \in \mathbb{Z}} \psi_j Z_{t-j}$$

for all $t \in \mathbb{Z}$, where $Z \sim \mathrm{WN}(0, \sigma^2)$ and $(\psi_j, j \in \mathbb{Z})$ is a sequence of real numbers with $\sum_{j \in \mathbb{Z}} |\psi_j| < +\infty$.

A linear process is called a *moving average* or $\mathrm{MA}(\infty)$ *process* if $\psi_j = 0$ for all negative j, i.e., if X has the representation

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

We remark that the summability condition $\sum_{j=-\infty}^{\infty} |\psi_j| < +\infty$ ensures that the infinite sum converges with probability one and in mean square, which is left as an *exercise* to the interested reader.

We can rewrite the series in terms of the previously introduced backward shift operator B by

$$X_t = \psi(B) Z_t,$$

where we define the operator $\psi(B)$ by

$$\psi(B) := \sum_{j \in \mathbb{Z}} \psi_j B^j.$$

This is used in the following proposition to characterize the properties of $\psi(B)Y$, where Y is assumed to be stationary.

Proposition 3.1.2. *Let Y be a stationary time series with mean zero and autocovariance function γ_Y and let $(\psi_j, j \in \mathbb{Z})$ be a real-valued sequence such that $\sum_{j\in\mathbb{Z}} |\psi_j| < +\infty$. Then the time series X defined by*

$$X_t := \psi(B)Y_t$$

for all $t \in \mathbb{Z}$ is stationary with mean zero and autocovariance function γ_X given by

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

for all $h \in \mathbb{Z}$.

In the special case that X is a linear process, it holds that the autocovariance function γ_X is given by

$$\gamma_X(h) = \sum_{j\in\mathbb{Z}} \psi_j\psi_{h+j}\sigma^2$$

for all $h \in \mathbb{Z}$, where σ^2 is the variance of the underlying white noise sequence.

Proof. By a limit argument one shows similarly to the proof for linear time series that the sum converges and the expression is well-defined.

Since $\sum_{j\in\mathbb{Z}} |\psi_j| < +\infty$ and Y is stationary with finite variance, we are allowed to interchange the expectation and the sum to obtain for all $t \in \mathbb{Z}$

$$\mathbb{E}(X_t) = \mathbb{E}(\psi(B)Y_t) = \mathbb{E}\left(\sum_{j\in\mathbb{Z}} \psi_j Y_{t-j}\right) = \sum_{j\in\mathbb{Z}} \psi_j\,\mathbb{E}(Y_{t-j}) = 0,$$

where we used in the last step that Y is stationary with mean zero.

With a similar argument we are allowed to compute the covariance for $t, h \in \mathbb{Z}$

$$\begin{aligned}\mathsf{Cov}(X_t, X_{t+h}) &= \mathbb{E}\left(\sum_{j,k\in\mathbb{Z}} \psi_j Y_{t-j}\psi_k Y_{t+h-k}\right) = \sum_{j,k\in\mathbb{Z}} \psi_j\psi_k\,\mathbb{E}(Y_{t-j}Y_{t+h-k}) \\ &= \sum_{j,k\in\mathbb{Z}} \psi_j\psi_k\gamma_Y((t+h-k)-(t-j)) = \sum_{j,k\in\mathbb{Z}} \psi_j\psi_k\gamma_Y(h+j-k),\end{aligned}$$

which does not depend on t. One shows that X has finite variance by the mean square convergence of the series, which can be found in [4, Example C.1.1]. Therefore X is stationary.

If X is a linear process, Y is white noise with variance σ^2, which simplifies the previous expression to

$$\gamma_X(h) = \sum_{j,k\in\mathbb{Z}} \psi_j\psi_k\gamma_Y(h+j-k) = \sum_{j\in\mathbb{Z}} \psi_j\psi_{h+j}\sigma^2,$$

since $\gamma_Y(h+j-k)$ is just unequal to zero if $k = h+j$. This finishes the proof. □

3.2 ARMA models

An important class of linear processes is the one given by ARMA models. To understand the notation and the background, we first define autoregressive and moving average processes.

Definition 3.2.1. A time series X is called an *autoregressive process of order p* or AR(p) *process* if X is stationary and if for all $t \in \mathbb{Z}$

$$X_t - \sum_{j=1}^{p} \phi_j X_{t-j} = Z_t,$$

where $Z \sim \mathrm{WN}(0, \sigma^2)$.

Definition 3.2.2. A time series X is called a *moving average process of order q* or MA(q) *process* if X is stationary and if for all $t \in \mathbb{Z}$

$$X_t = Z_t + \sum_{j=1}^{q} \theta_j Z_{t-j},$$

where $Z \sim \mathrm{WN}(0, \sigma^2)$.

If we combine AR(p) and MA(q) processes, we end up with the following generalization to an ARMA(p, q) process.

Definition 3.2.3. A time series X is an ARMA(p, q) *process* if X is stationary and if for all $t \in \mathbb{Z}$

$$X_t - \sum_{j=1}^{p} \phi_j X_{t-j} = Z_t + \sum_{j=1}^{q} \theta_j Z_{t-j}, \tag{3.1}$$

where $Z \sim \mathrm{WN}(0, \sigma^2)$ and the polynomials $(1 - \sum_{j=1}^{p} \phi_j z^j)$ and $(1 + \sum_{j=1}^{q} \theta_j z^j)$ have no common factors. Further a time series X is called an ARMA(p, q) *process with mean μ* if $X - \mu$ is an ARMA(p, q) process.

To simplify the notation, we set

$$\phi(z) := 1 - \sum_{j=1}^{p} \phi_j z^j$$

and

$$\theta(z) := 1 + \sum_{j=1}^{q} \theta_j z^j.$$

Then the recursive form of the ARMA(p, q) process can be rewritten as

$$\phi(B) X_t = \theta(B) Z_t,$$

where we recall that B denotes the backward shift operator.

Proposition 3.2.4 (Existence and uniqueness)**.** *A stationary solution X of Equation* (3.1) *exists and is the unique stationary solution if and only if*

$$1 - \sum_{j=1}^{p} \phi_j z^j \neq 0$$

for all $z \in \mathbb{C}$ with $|z| = 1$.

Proof. The proof uses concepts of complex analysis. To get an idea let us look at the recursive form

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

Then one would usually—B being not an operator—solve the equation if $\phi(B)$ is invertible by multiplying from both sides with $\phi(B)^{-1}$, which yields

$$\phi(B)^{-1}\phi(B)X_t = \phi(B)^{-1}\theta(B)Z_t$$

and therefore

$$X_t = \phi(B)^{-1}\theta(B)Z_t.$$

Like that we obtain a solution to the recursion. The existence of $\phi(B)^{-1}$ is justified by the fact from complex analysis that if $\phi(z) \neq 0$ for all $z \in \mathbb{C}$ on the unit circle, then there exists $\delta > 0$ such that for all $1 - \delta < |z| < 1 + \delta$

$$\phi(z)^{-1} = \sum_{j=-\infty}^{\infty} \chi_j z^j$$

with $\sum_{j=-\infty}^{\infty} |\chi_j| < +\infty$, i.e., it has a Laurent series expansion. Setting $\psi(z) := \chi(z)\theta(z) = \sum_{j=-\infty}^{\infty} \psi_j z^j$, we obtain the solution

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

which is unique due to the uniqueness of the inverse. The converse can be proven by observing that the existence of a solution and the invertibility of $\phi(B)$ are coupled by an "if and only if" condition. □

In what follows two important properties and their equivalent characterizations are introduced that allow us to regard the ARMA process either as an infinite dimensional autoregressive or an infinite dimensional moving average process.

Definition 3.2.5. An ARMA(p, q) process X is *causal* or a *causal function of* Z if there exists a real-valued sequence $(\psi_j, j \in \mathbb{N}_0)$ such that $\sum_{j=0}^{\infty} |\psi_j| < +\infty$ and

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

for all $t \in \mathbb{Z}$, i.e., if X is a moving average/MA(∞) process.

The following lemma enables us to check for causality in practice.

Lemma 3.2.6. *An* ARMA(p, q) *process* X *is causal if and only if*

$$1 - \sum_{j=1}^{p} \phi_j z^j \neq 0$$

for all $z \in \mathbb{C}$ *with* $|z| \leq 1$.

Proof. We have seen in the proof of Proposition 3.2.4 that the unique solution is given by

$$X_t = \psi(B)Z_t = \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}.$$

Therefore we can conclude that X is causal if and only if

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

which holds under the assumption that ϕ and θ have no common factors if and only if χ as defined in the proof of Proposition 3.2.4 has a power series expansion

$$\chi(z) = \phi(z)^{-1} = \sum_{j=0}^{\infty} \chi_j z^j.$$

The existence of the power series expansion is equivalent to the condition that $\phi(z) \neq 0$ for all $z \in \mathbb{C}$ on the (closed) unit disc, which is another fact from complex analysis. □

Together with Proposition 3.2.4, the lemma implies the following corollary as an immediate consequence.

Corollary 3.2.7. *A causal* ARMA(p, q) *process has a unique stationary solution.*

Since $\psi(z)\phi(z) = \chi(z)\theta(z)\phi(z) = \theta(z)$, or equivalently

$$(\psi_0 + \psi_1 z + \cdots)(1 - \phi_1 z - \cdots - \phi_p z^p) = (1 + \theta_1 z + \cdots + \theta_q z^q),$$

one can explicitly find the coefficients in the power series expansion of $\psi(z)$ by equating coefficients of z^j, $j \in \mathbb{N}_0$, in this expression. This means that the sequence $(\psi_j, j \in \mathbb{N}_0)$, can be determined from the equations

$$\psi_j - \sum_{k=1}^{p} \phi_k \psi_{j-k} = \theta_j \tag{3.2}$$

for all $j \in \mathbb{N}_0$, where $\theta_0 := 1$, $\theta_j := 0$ for $j > q$, and $\psi_j := 0$ for $j < 0$.

A second important property of an ARMA(p, q) process is invertibility which in some sense interchanges the roles of X and Z.

Definition 3.2.8. An ARMA(p, q) process X is *invertible* if there exists a real-valued sequence $(\pi_j, j \in \mathbb{N}_0)$ such that $\sum_{j=0}^{\infty} |\pi_j| < +\infty$ and

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

for all $t \in \mathbb{Z}$, i.e., if X is an autoregressive/AR(∞) process.

A similar lemma and characterization as for causal processes also holds for invertible processes that helps for practical purposes since it can be checked relatively easy.

Lemma 3.2.9. *An* ARMA(p, q) *process X is invertible if and only if*

$$1 + \sum_{j=1}^{q} \theta_j z^j \neq 0$$

for all $z \in \mathbb{C}$ with $|z| \leq 1$.

Proof. Looking at the structure of the problem, it is clear that the method of proof of Lemma 3.2.6 can be used by interchanging the roles of X and Z, i.e., the invertibility is equivalent to the existence of the inverse of θ with power series expansion

$$\theta(z)^{-1} = \sum_{j=0}^{\infty} \pi_j z^j,$$

which—again as consequence of complex analysis—holds if and only if $\theta(z) \neq 0$ for all $z \in \mathbb{C}$ on the (closed) unit disc. □

In the same way as in (3.2), the sequence $(\pi_j, j \in \mathbb{N}_0)$ is determined by the equations

$$\pi_j + \sum_{k=1}^{q} \theta_k \pi_{j-k} = -\phi_j$$

for $j \in \mathbb{N}_0$, where we set $\phi_0 := -1$, $\phi_j := 0$ for $j > p$, and $\pi_j := 0$ for $j < 0$.

3.2.1 Autocorrelation and partial autocorrelation function

Let us consider autocovariance, autocorrelation, and partial autocorrelation functions as well as their computation in this section. We start with the calculation of the autocovariance function. Therefore we recall that an ARMA(p, q) process is given by

$$\phi(B)X_t = \theta(B)Z_t,$$

where $Z \sim \mathrm{WN}(0, \sigma^2)$ and

$$\phi(z) := 1 - \sum_{i=1}^{p} \phi_i z^i$$

as well as

$$\theta(z) := 1 + \sum_{j=1}^{q} \theta_j z^j.$$

Let us assume that the process is causal, then by definition there exists a real-valued sequence $(\psi_j, j \in \mathbb{N}_0)$ such that

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

where the coefficients ψ_j are determined by Equation (3.2).

In what follows we introduce three methods to compute the autocovariance function of an ARMA(p, q) process.

Method 3.2.10. Proposition 3.1.2 implies with the above representation that

$$\gamma(h) = \mathbb{E}(X_{t+h}X_t) = \sigma^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+|h|}.$$

If the coefficients $(\psi_j, j \in \mathbb{N}_0)$ are unknown, estimators can be computed with the methods introduced in the next Section 3.2.2.

Method 3.2.11. If we multiply each side of the equations

$$X_t - \sum_{j=1}^{p} \phi_j X_{t-j} = Z_t + \sum_{j=1}^{q} \theta_j Z_{t-j}$$

by X_{t-k} for $k \in \mathbb{N}_0$ and take expectations on each side, noting that

$$X_{t-k} = \sum_{j=0}^{\infty} \psi_j Z_{t-k-j},$$

we obtain

$$\gamma(k) - \sum_{j=1}^{p} \phi_j \gamma(k-j) = \sigma^2 \sum_{j=0}^{\infty} \theta_{k+j} \psi_j$$

for $0 \leq k < m$ and

$$\gamma(k) - \sum_{j=1}^{p} \phi_j \gamma(k-j) = 0$$

for $k \geq m$, where $m = \max\{p, q+1\}$, $\psi_j := 0$ for $j < 0$, $\theta_0 := 1$, and $\theta_j := 0$ for $j \notin \{0, \ldots, q\}$. These equations can sometimes be solved explicitly.

Method 3.2.12. This method is a numerical version of Method 3.2.11. Use the equations there for $k = 0, \ldots, p$ and solve them numerically to approximate $\gamma(0), \ldots, \gamma(p)$. Afterwards, use the result and determine successively $\gamma(j)$ for $j > p$.

Another important function for the estimation and fitting of models is the partial autocorrelation function. We will first define the function before we give the definition of the sample partial autocorrelation function that can be computed from observed data.

Definition 3.2.13. Let X be an ARMA(p, q) process. The *partial autocorrelation function* α (*PACF* for short) of X is defined by

$$\alpha(0) := 1,$$
$$\alpha(h) := \phi_{hh}$$

for $h \geq 1$, where ϕ_{hh} is the last component of

$$\phi_h = \left((\gamma(i-j))_{i,j=1}^{h}\right)^{-1} (\gamma(1), \gamma(2), \ldots, \gamma(h))'.$$

For any series of observations $(x_1, \ldots, x_n)$ with $x_i \neq x_j$ for some i and j, the *sample partial autocorrelation function* $\hat{\alpha}$ is given by

$$\hat{\alpha}(0) := 1,$$
$$\hat{\alpha}(h) := \hat{\phi}_{hh}$$

for $h \geq 1$, where $\hat{\phi}_{hh}$ is similarly the last component of

$$\hat{\phi}_h = \left((\hat{\gamma}(i-j))_{i,j=1}^{h}\right)^{-1} (\hat{\gamma}(1), \hat{\gamma}(2), \ldots, \hat{\gamma}(h))'.$$

It can be shown that, when X has mean zero,

$$\alpha(1) = \mathsf{Cor}(X_{t+1}, X_t) = \rho(1)$$

and for $h \geq 2$

$$\alpha(h) = \mathsf{Cor}(X_{t+h} - b_{t+h}^{l}(X^h), X_t - b_t^{l}(X^h))$$

where $X^h := (X_{t+1}, \ldots, X_{t+h-1})$, which means that the partial autocorrelation $\alpha(h)$ can be thought of as the correlation between X_t and X_{t+h} *when adjusting for the intervening observations* $X_{t+1}, \ldots, X_{t+h-1}$, hence the name of the function.

Furthermore it can be shown that the partial autocorrelation function of a causal AR(p) process is zero for lags greater than p. Since algebraic computations of the partial autocorrelation function are in general quite complicated, one should prefer numerical computations in many cases.

One method to choose an appropriate AR(p) model is to look at the sample partial autocorrelation function $\hat{\alpha}$. If $\hat{\alpha}(h)$ is significantly different from zero for $h = 0, \ldots, p$ and negligible for $h > p$, an AR(p) model might be a good choice for the observed data. For sample size n one defines "negligible" according to the Central Limit Theorem 1.1.7 that around 95% of the sample partial autocorrelation function values beyond lag p should fall within the bounds $\pm 1.96/\sqrt{n}$, which is justified by the fact that the sample partial autocorrelation function values at lags greater than p are approximately independent $\mathcal{N}(0, 1/n)$ distributed random variables.

3.2.2 Parameter estimation

Let us assume in this section that the order parameters p and q of an ARMA(p,q) model are known, which is not true in most realistic cases. We will discuss the order selection in Section 3.2.3. Here we will give methods to determine the parameters $(\phi_j, j = 1,\dots,p)$ and $(\theta_j, j = 1,\dots,q)$.

We start with computation methods for purely autoregressive models. To do parameter estimation for an AR(p) model, we introduce two methods.

First we introduce the Yule–Walker estimation, which can be derived from Method 3.2.11. We observe that the equations in Method 3.2.11 simplify for an AR(p) model to

$$\gamma(k) - \sum_{j=1}^{p} \phi_j \gamma(k-j) = \begin{cases} 0 & k \in \{1,\dots,p\}, \\ \sigma^2 & k = 0, \end{cases}$$

which are called the *Yule–Walker equations*. These equations can be rewritten as

$$\sum_{j=1}^{p} \phi_j \gamma(k-j) = \begin{cases} \gamma(k) & k \in \{1,\dots,p\}, \\ \gamma(0) - \sigma^2 & k = 0, \end{cases}$$

which leads to the linear system

$$(\gamma(i-j))_{i,j=1}^{p} (\phi_1,\dots,\phi_p)' = (\gamma(1),\dots,\gamma(p))'$$

and to

$$(\phi_1,\dots,\phi_p) \cdot (\gamma(1),\dots,\gamma(p))' = \gamma(0) - \sigma^2.$$

Often the Yule–Walker equations are used to determine γ from σ^2 and $(\phi_j, j = 1,\dots,p)$. For estimation we do it the other way around by using the sample autocovariance function $\hat{\gamma}$ from the made observations to get estimates of σ^2 and $(\phi_j, j = 1,\dots,p)$. Due to better properties of the sample autocorrelation function $\hat{\rho}$ compared to $\hat{\gamma}$, we transform the equations by dividing them by $\hat{\gamma}(0)$ and obtain the following method.

Method 3.2.14 (Yule–Walker estimation)**.** Compute estimators $\hat{\sigma}^2$ and $(\hat{\phi}_j, j = 1,\dots,p)$ from the equations

$$(\hat{\phi}_1,\dots,\hat{\phi}_p)' = \hat{R}_p^{-1} (\hat{\rho}(1),\dots,\hat{\rho}(p))',$$

and

$$\hat{\sigma}^2 = \hat{\gamma}(0) \left(1 - (\hat{\rho}(1),\dots,\hat{\rho}(p)) \hat{R}_p^{-1} (\hat{\rho}(1),\dots,\hat{\rho}(p))'\right),$$

where $\hat{R}_p$ denotes the sample autocorrelation matrix.

We observe that for large sample sizes n the vector $(\hat{\phi}_1,\dots,\hat{\phi}_p)$ is approximately normally distributed with mean $(\phi_1,\dots,\phi_p)$ and covariance matrix $n^{-1}\sigma^2\Gamma_p^{-1}$, where $\Gamma_p := (\gamma(i-j))_{i,j=1}^{p}$. This knowledge can be used to compute confidence regions.

Furthermore we remark that the the Yule–Walker estimates are special cases of moment estimators. The analogous procedure for ARMA(p,q) models with $q > 0$ is easily formulated, but the corresponding equations are nonlinear in the unknown coefficients. This might lead to nonexistence and nonuniqueness of solutions.

The second method that we introduce for AR(p) models is *Burg's algorithm*. We start by introducing the necessary notation and quantities. Therefore let $(x_i, i = 1,\dots,n)$ be n observations of a stationary times series X with mean zero. For $0 \le i < n$ let $(u_i(t), t = i+1,\dots,n)$ be the differences between $x_{n+1+i-t}$ and its best linear estimate in terms of the preceding i observations, which are called the *forward prediction errors*. Similarly for $0 \le i < n$ let $(v_i(t), t = i+1,\dots,n)$

be the differences between x_{n+1-t} and its best linear estimate in terms of the subsequent i observations, which are referred to as *backward prediction errors*. In an *exercise* one shows that the forward and backward prediction errors satisfy the recursions

$$\begin{aligned} u_0(t) &= v_0(t) = x_{n+1-t}, \\ u_i(t) &= u_{i-1}(t-1) - \phi_{ii} v_{i-1}(t), \\ v_i(t) &= v_{i-1}(t) - \phi_{ii} u_{i-1}(t-1). \end{aligned}$$

The transformation of these recursions leads to the following algorithm, which computes estimates for σ^2 and ϕ_{ii}. The remaining $(\phi_{ij}^{(\mathrm{B})}, j < i)$ can be obtained by replacing ϕ_{ii} with $\phi_{ii}^{(\mathrm{B})}$ in the Durbin–Levinson algorithm 2.3.10, where (B) indicates that the estimates are computed using Burg's algorithm.

Method 3.2.15 (Burg's algorithm)**.**

$$\begin{aligned} d(1) &:= \sum_{t=2}^{n} \left(u_0^2(t-1) + v_0^2(t) \right), \\ \phi_{ii}^{(\mathrm{B})} &:= \frac{2}{d(i)} \sum_{t=i+1}^{n} v_{i-1}(t) u_{i-1}(t-1), \\ d(i+1) &:= \left(1 - \phi_{ii}^{(\mathrm{B})2} \right) d(i) - v_i^2(i+1) - u_i^2(n), \\ \sigma_i^{(\mathrm{B})2} &:= \frac{\left(1 - \phi_{ii}^{(\mathrm{B})2} \right) d(i)}{2(n-i)}. \end{aligned}$$

For MA models and ARMA models we introduce the following two methods.

Similarly to the application of the Durbin–Levinson algorithm 2.3.10 to fit AR models, we can use the innovations algorithm 2.3.11 to fit MA models

$$X_t = Z_t + \sum_{j=1}^{m} \hat{\theta}_{mj} Z_{t-j}$$

of given orders $m \in \mathbb{N}$, where $Z \sim \mathrm{WN}(0, \hat{v}_m)$ with estimated white noise variance $\hat{v}_m$.

Method 3.2.16 (Innovations algorithm)**.** Apply the innovations algorithm 2.3.11 with the sample autocovariance function instead of the autocovariance function to obtain the coefficients of the *fitted innovations* $\mathrm{MA}(m)$ *model*

$$X_t = Z_t + \sum_{j=1}^{m} \hat{\theta}_{mj} Z_{t-j},$$

as well as the estimated white noise variance $\hat{v}_m$ of $Z \sim \mathrm{WN}(0, \hat{v}_m)$.

We remark that the obtained estimators are just consistent for invertible $\mathrm{MA}(q)$ processes with $Z \sim \mathrm{IID}(0, \sigma^2)$ and $\mathbb{E}(Z_t^4) < +\infty$ for all $t \in \mathbb{Z}$.

In the case of an $\mathrm{ARMA}(p, q)$ model with $p > 0$ and $q > 0$, we observe that the assumption of causality ensures that

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

where the coefficients $(\psi_j, j \in \mathbb{N}_0)$ satisfy for $j \in \mathbb{N}_0$ that

$$\psi_j = \theta_j + \sum_{i=1}^{\min\{j,p\}} \phi_i \psi_{j-i} \tag{3.3}$$

with $\theta_0 := 1$ and $\theta_j := 0$ for $j > q$ by (3.2). To estimate $(\psi_j, j = 1, \ldots, p+q)$ we can use the innovation estimates $\hat{\theta}_{m1}, \ldots, \hat{\theta}_{m(p+q)}$. One then estimates $(\phi_i, i = 1, \ldots, p)$ by solving the system of equations given by (3.3) for $q < j \leq p+q$. Having done that, the coefficients $(\theta_j, j = 1, \ldots, q)$ are then determined from the q remaining equations given by (3.3) for $j \leq q$.

The large-sample behavior of $\hat{\theta}_{m1}, \ldots, \hat{\theta}_{m(p+q)}$ is as follows (under smoothness assumptions, for details see [4, Remark 5.1.3 1]): For any positive integer k the joint distribution function of

$$\sqrt{n}(\hat{\theta}_{m1} - \psi_1, \hat{\theta}_{m2} - \psi_2, \ldots, \hat{\theta}_{mk} - \psi_k)$$

converges for $n \to +\infty$ to a multivariate normal distribution with mean zero and covariance matrix $A = (a_{ij})_{i,j=1}^k$, where

$$a_{ij} := \sum_{r=1}^{\min\{i,j\}} \theta_{i-r}\theta_{j-r}.$$

This result enables us to find approximate large-sample confidence intervals for the moving average coefficients. Moreover, the estimator $\hat{v}_m$ is consistent for σ^2.

The following algorithm is a variant of a least square regression.

Method 3.2.17 (Hannan–Rissanen algorithm)**.**

(i) Fit a high-order AR(m) model (with $m > \max\{p, q\}$) to the data using the Yule–Walker estimates from Method 3.2.14. For estimated coefficients $(\hat{\phi}_{m1}, \ldots, \hat{\phi}_{mm})$, compute the estimated residuals $\hat{Z}_t$ from the equations

$$\hat{Z}_t = X_t - \sum_{j=1}^{m} \hat{\phi}_{mj} X_{t-j}$$

for $t = m+1, \ldots, n$.

(ii) Estimate the vector of parameters $\beta := (\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q)$ by a least squares linear regression of X_t onto $(X_{t-1}, \ldots, X_{t-p}, \hat{Z}_{t-1}, \ldots, \hat{Z}_{t-q})$, $t = m+1+q, \ldots, n$, i.e., minimize the sum of squares

$$S(\beta) = \sum_{t=m+1+q}^{n} \left(X_t - \sum_{j=1}^{p} \phi_j X_{t-j} - \sum_{i=1}^{q} \theta_i \hat{Z}_{t-i} \right)^2$$

with respect to β. This gives the *Hannan–Rissanen estimator*

$$\hat{\beta} = (Z'Z)^{-1} Z'(X_{m+1+q}, \ldots, X_n)',$$

where

$$Z = \begin{pmatrix} X_{m+q} & X_{m+q-1} & \cdots & X_{m+q+1-p} & \hat{Z}_{m+q} & \hat{Z}_{m+q-1} & \cdots & \hat{Z}_{m+1} \\ X_{m+q+1} & X_{m+q} & \cdots & X_{m+q+2-p} & \hat{Z}_{m+q+1} & \hat{Z}_{m+q} & \cdots & \hat{Z}_{m+2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ X_{n-1} & X_{n-2} & \cdots & X_{n-p} & \hat{Z}_{n-1} & \hat{Z}_{n-2} & \cdots & \hat{Z}_{n-q} \end{pmatrix}$$

(If $p = 0$, Z contains only the last q columns.) The Hannan–Rissanen estimator of the white noise variance is

$$\hat{\sigma}^{(\mathrm{HR})2} = \frac{S(\hat{\beta})}{n - m - q}.$$

Let us finally introduce the frequently used and well-known concept of likelihood which we use for maximum likelihood estimation as the last but not least method in this section. Therefore we assume that we are given a Gaussian ARMA(p,q) process (or at least act as if). Then for any fixed values $(\phi_1,\dots,\phi_p)$, $(\theta_1,\dots,\theta_q)$, and σ^2, the random variables $X_1-\hat{X}_1,\dots,X_n-\hat{X}_n$ are independent and normally distributed, where $\hat{X}_1:=0$ and $\hat{X}_j:=\mathbb{E}(X_j|X_1,\dots,X_{j-1})$, $j\geq 2$. Let Γ_n denote the covariance matrix of $(X_1,\dots,X_n)$ and assume that it is nonsingular. The *likelihood* $L(\Gamma_n)$ of $(X_1,\dots,X_n)$ is given by

$$L(\Gamma_n)=(2\pi)^{-n/2}(\det\Gamma_n)^{-1/2}\exp\left(-2^{-1}(X_1,\dots,X_n)\Gamma_n^{-1}(X_1,\dots,X_n)'\right).$$

Keeping the assumption that the process is Gaussian—which also kind of makes sense for other processes if large sample sizes are used—one derives in our case of interest the *Gaussian likelihood* for an ARMA(p,q) process

$$L(\phi,\theta,\sigma^2)=(2\pi\sigma^2)^{-n/2}\left(\prod_{j=1}^{n}r_{j-1}\right)^{-1/2}\exp\left(-(2\sigma^2)^{-1}\sum_{j=1}^{n}r_{j-1}^{-1}(X_j-\hat{X}_j)^2\right),$$

where

$$r_j:=\frac{\mathsf{Var}(X_{j+1}-\hat{X}_{j+1})}{\sigma^2},$$

which can be determined by the innovations algorithm 2.3.11.

Method 3.2.18 (Maximum likelihood estimators)**.** The maximum likelihood estimators of σ^2, ϕ, and θ are determined from the expression

$$\hat{\sigma}^2=n^{-1}S(\hat{\phi},\hat{\theta}),$$

where

$$S(\hat{\phi},\hat{\theta})=\sum_{j=1}^{n}r_{j-1}^{-1}(X_j-\hat{X}_j)^2,$$

and $\hat{\phi}$ and $\hat{\theta}$ are the values of ϕ and θ that minimize

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

Do the minimization of ℓ numerically. Initial values can be computed by the methods introduced previously in this section.

The derivation of the equations is left as an *exercise* to the reader. (*Hint:* Differentiate $\ln L(\phi,\theta,\sigma^2)$ with respect to σ^2 and observe that $\hat{X}_j$ and r_{j-1}, $j=1,\dots,n$, are independent of σ^2.)

Method 3.2.19 (Least squares estimation of mixed models)**.** Minimize S instead of ℓ in Method 3.2.18 to obtain the least squares estimates $\tilde{\phi}$ and $\tilde{\theta}$. The least squares estimate of σ^2 is

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

Having fitted the model, it remains to check that the model was chosen adequately. If this is the case, the residuals should behave like white noise and be approximately normally distributed if a maximum likelihood estimation procedure was used. If the residuals do not appear to be normally distributed, one can assume that the innovations follow a Student t-distribution instead and see if the residuals fit better after a similar maximum likelihood estimation procedure.

3.2.3 Order selection

Assume in this section that our data is already transformed, e.g., trend and seasonal components are removed, such that the remaining series can potentially be fitted by a zero-mean ARMA(p, q) model. In this section we treat the problem to choose appropriate values for p and q.

In general a selection for an AR(p) or MA(q) model may be made using autocorrelation and partial autocorrelation functions. Typically, an autocorrelation function with "q peaks and then zero" indicates a MA(q) model. This is clear since the autocovariance function of a MA(q) process is by Proposition 3.1.2 given by

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

On the other hand side a slowly decaying autocorrelation function and a partial autocorrelation function with "p peaks and then zero" indicates an AR(p) model, which we discussed in Section 3.2.1. After parameter estimation, which can be done with the procedures introduced in Section 3.2.2, the model should be checked if the obtained residuals behave like white noise by the methods introduced in Section 2.2.

To find an ARMA(p, q) model systematically, we just introduce the following method, although a lot more could be written about that.

It is always possible to fit an ARMA(p, q) model with (too) large p and q, which is not an advantage from a forecasting point of view. In general it results in a small estimated white noise variance, but for forecasting the mean squared error of the forecast will additionally depend on the errors arising from the parameter estimation. Therefore we introduce a "penalty factor" to discourage the fitting of models with too many parameters.

We just introduce the AICC criterion, where AIC stands for *Akaike's Information Criterion* and the last C for *biased-Corrected*.

Method 3.2.20 (AICC criterion)**.** Choose p, q, ϕ_p, and θ_q to minimize

$$-2 \ln L(\phi_p, \theta_q, S(\phi_p, \theta_q)/n) + 2n \frac{p+q+1}{n-p-q-2}$$

One problem with the AICC criterion that we remark is that the estimators for p and q are not consistent, i.e., it does not hold that they converge almost surely to p and q. Consistent estimators include, e.g., those obtained by the BIC.

In general one may say that order selection is a difficult problem and many criteria have been proposed. Rissanen's minimum description length (MDL) criterion seems to be rather much used according to [10].

3.2.4 Forecasting of ARMA processes

The innovations algorithm 2.3.11 provides us with a recursive method for forecasting second-order zero-mean processes that are not necessarily stationary. For the causal ARMA process

$$\phi(B) X_t = \theta(B) Z_t,$$

where $Z \sim \text{WN}(0, \sigma^2)$, it is possible to simplify the application drastically. The idea is to apply the algorithm to the transformed process $W = (W_t, t \in \mathbb{N})$ defined by

$$W_t := \begin{cases} \sigma^{-1} X_t & t = 1, \ldots, m, \\ \sigma^{-1} \phi(B) X_t & t > m, \end{cases}$$

where $m = \max\{p, q\}$ (cf. [1]).

The autocovariance function γ_X of X can easily be computed using any method of Section 3.2.1. The autocovariances $\kappa(i, j) := \mathbb{E}(W_i W_j)$ for $i, j \geq 1$, are then found from

$$\kappa(i,j) = \begin{cases} \sigma^{-2}\gamma_X(i-j) & \max(i,j) \leq m, \\ \sigma^{-2}\left(\gamma_X(i-j) - \sum_{r=1}^{p} \phi_r \gamma_X(r - |i-j|)\right) & \min\{i,j\} \leq m < \max\{i,j\} \leq 2m, \\ \sum_{r=0}^{q} \theta_r \theta_{r+|i-j|} & \min\{i,j\} > m, \\ 0 & \text{otherwise.} \end{cases}$$

Applying the innovations algorithm 2.3.11 to the process W we obtain

$$\hat{W}_{n+1} = \begin{cases} \sum_{j=1}^{n} \theta_{nj}(W_{n+1-j} - \hat{W}_{n+1-j}) & q \leq n < m, \\ \sum_{j=1}^{q} \theta_{nj}(W_{n+1-j} - \hat{W}_{n+1-j}) & n \geq m, \end{cases}$$

where the coefficients $(\theta_{nj}, n \in \mathbb{N}, j \leq \min\{n, m\})$ and the mean squared errors $\mathrm{MSE}(\hat{W}_{n+1}, W_{n+1})$ are found recursively with κ as defined above. The notable feature of the predictors $(\hat{W}_{n+1}, n \in \mathbb{N})$ is the vanishing of θ_{nj} when both $n \geq m$ and $j > q$.

One derives that the predictor $\hat{W}_{n+1}$ is the best linear one-step predictor of W_{n+1}, i.e.,

$$\hat{W}_{n+1} = b^l_{n+1}(W^n)$$

with $W^n := (W_1, \ldots, W_n)$. Furthermore we obtain that

$$\hat{W}_t = \begin{cases} \sigma^{-1}\hat{X}_t & t = 1, \ldots, m, \\ \sigma^{-1}\Big(\hat{X}_t - \sum_{j=1}^{p} \phi_j X_{t-j}\Big) & t > m \end{cases}$$

due to the linearity of the prediction estimator. So we obtain as best linear estimator for X_{n+1} with respect to $X^n := (X_1, \ldots, X_n)$

$$b^l_{n+1}(X^n) = \hat{X}_{n+1} = \begin{cases} \sum_{j=1}^{n} \theta_{nj}(X_{n+1-j} - \hat{X}_{n+1-j}) & 1 \leq n < m, \\ \sum_{j=1}^{p} \phi_j X_{n+1-j} + \sum_{j=1}^{q} \theta_{nj}(X_{n+1-j} - \hat{X}_{n+1-j}) & n \geq m, \end{cases}$$

with mean squared error

$$\mathrm{MSE}(\hat{X}_{n+1}, X_{n+1}) = \mathbb{E}((\hat{X}_{n+1} - X_{n+1})^2) = \sigma^2\,\mathbb{E}((\hat{W}_{n+1} - W_{n+1})^2) = \sigma^2\,\mathrm{MSE}(\hat{W}_{n+1}, W_{n+1}),$$

where we recall that the coefficients $(\theta_{nj}, n \in \mathbb{N}, j \leq \min\{n, m\})$ and the mean squared errors $\mathrm{MSE}(\hat{W}_{n+1}, W_{n+1})$ are found recursively from the innovations algorithm 2.3.11. The best linear estimators can be computed recursively.

Let us next consider h-step predictors of an ARMA(p, q) process. One can show similarly to the results of Section 2.3 (cf. [3, Section 5.2]) that for $W^n := (W_1, \ldots, W_n)$ the best linear h-step predictor is given by

$$b^l_{n+h}(W^n) = \sum_{j=h}^{n+h-1} \theta_{(n+h-1)j}(W_{n+h-j} - \hat{W}_{n+h-j}) = \sigma^{-2} \sum_{j=h}^{n+h-1} \theta_{(n+h-1)j}(X_{n+h-j} - \hat{X}_{n+h-j}).$$

We conclude with the properties of W that the h-step predictors $b^l_{n+h}(X^n)$ with $X^n := (X_1, \ldots, X_n)$ satisfy

$$b^l_{n+h}(X^n) = \begin{cases} \sum_{j=h}^{n+h-1} \theta_{(n+h-1)j}(X_{n+h-j} - \hat{X}_{n+h-j}) & 1 \leq h \leq m - n, \\ \sum_{i=1}^{p} \phi_i b^l_{n+h-i}(X^n) + \sum_{j=h}^{n+h-1} \theta_{(n+h-1)j}(X_{n+h-j} - \hat{X}_{n+h-j}) & h > m - n. \end{cases}$$

If, as is almost always the case in practice, $n > m := \max\{p, q\}$, then for all $h \geq 1$

$$b^l_{n+h}(X^n) = \sum_{i=1}^{p} \phi_i \, b^l_{n+h-i}(X^n) + \sum_{j=h}^{n+h-1} \theta_{(n+h-1)j}(X_{n+h-j} - \hat{X}_{n+h-j}).$$

Once the predictors $\hat{X}_1, \ldots, \hat{X}_n$ have been computed, it is a straightforward calculation (with fixed n) to determine the best linear predictors $b^l_{n+h}(X^n)$ recursively.

The mean squared error of $b^l_{n+h}(X^n)$ is computed from the formula

$$\mathrm{MSE}(b^l_{n+h}(X^n), X_{n+h}) = \sum_{j=0}^{h-1} \left(\sum_{r=0}^{j} \chi_r \theta_{(n+h-r-1)(j-r)} \right)^2 v_{n+h-j-1},$$

where the coefficients χ_j are computed recursively from the equations $\chi_0 := 1$ and

$$\chi_j = \sum_{k=1}^{\min\{p,j\}} \phi_k \phi_{j-k}$$

for $j \in \mathbb{N}$, and the coefficients $(v_{n+h-j-1}, j = 0, \ldots, h-1)$ denote the mean squared errors of the one-step predictors as introduced in the methods in Section 2.3.

Finally in this section we remark that in the special case that the ARMA process is driven by Gaussian white noise, i.e., $Z \sim \mathrm{IID}\,\mathcal{N}(0, \sigma^2)$, for each $h \geq 1$ the prediction error $b^l_{n+h}(X^n) - X_{n+h}$ is normally distributed with mean zero and variance $\mathrm{MSE}(b^l_{n+h}(X^n), X_{n+h})$. This allows to compute confidence intervals. These bounds are called $(1-\alpha)$ *prediction bounds* for X_{n+h} if the $(1-\alpha/2)$ quantile of the standard normal distribution is used.

3.3 ARIMA models

In this section we focus on a nonstationary time series model, which can be considered if the observations do not seem to follow a stationary model. The class of ARIMA processes that we introduce here is a generalization of ARMA processes. It is the class of processes that reduce to ARMA processes when differenced finitely many times. More precisely we define it in the following way.

Definition 3.3.1. Let X be a stochastic process and d a nonnegative integer. Then X is an $\mathrm{ARIMA}(p, d, q)$ *process* if the process Y defined by $Y_t := (1-B)^d X_t$ is a causal $\mathrm{ARMA}(p, q)$ process.

Here the abbreviation ARIMA stands for *autoregressive integrated moving average*. Stated in another way this definition states that X satisfies a difference equation of the form

$$\phi^*(B)X_t := \phi(B)(1-B)^d X_t = \theta(B)Z_t,$$

where $Z \sim \mathrm{WN}(0, \sigma^2)$ and ϕ and θ are polynomials of degree p and q, respectively. Furthermore $\phi(z) \neq 0$ for $|z| \leq 1$, while the polynomial ϕ^* has a zero of degree d at $z = 1$. The process is stationary if and only if $d = 0$, in which case it reduces to an $\mathrm{ARMA}(p, q)$ process. Furthermore for $d \geq 1$ neither the mean nor the covariance function are determined by the above difference equation.

Observe that if $d \geq 1$, we can add an arbitrary polynomial trend of degree $(d-1)$ to X without violating the difference equation. ARIMA models are therefore useful for representing data with trend.

An ARIMA model is an appropriate choice if the autocovariance function is slowly decaying. Nevertheless, in practice it is very difficult to distinguish between an ARIMA$(p, 1, q)$ process and an ARMA$(p + 1, q)$ process with a root of $\phi(z) = 0$ near the unit circle.

In what follows we treat unit roots to determine an appropriate model.

To treat and find ARIMA models one applies the difference operator $(1 - B)$ to the observed data until the sample autocorrelation function is no longer slowly decaying with values near 1 at small lags but rapidly decreasing. The differenced time series can then be modeled by a low-order ARMA(p, q) process. The resulting ARIMA(p, d, q) model for the original data has then an autoregressive polynomial

$$\phi^*(z) = \left(1 - \sum_{j=1}^{p} \phi_j z^j\right)(1 - z)^d$$

with d roots on the unit circle.

A more systematic approach to decide if the observed data follow a model with roots on the unit circle is due to Dickey and Fuller (see [6]) and described in what follows. Let us start with the case of an AR(1) model before we extend the problem to the general case. Therefore assume that $(x_1, \ldots, x_n)$ are observations of $(X_1, \ldots, X_n)$ of the AR(1) model

$$X_t - \mu = \phi_1(X_{t-1} - \mu) + Z_t,$$

where $Z \sim \mathrm{WN}(0, \sigma^2)$, $|\phi_1| < 1$, and $\mu := \mathbb{E}(X_t)$. For large sample sizes n the maximum likelihood estimator of ϕ_1 is approximately normally distributed with mean ϕ_1 and variance $(1-\phi_1^2)/n$, which does not hold if $\phi_1 = 1$ and is therefore of no use for testing $H_0 : \phi_1 = 1$ versus $H_1 : \phi_1 < 1$. To construct a hypothesis test, we rewrite the model as

$$\nabla X_t = X_t - X_{t-1} = \phi_0^* + \phi_1^* X_{t-1} + Z_t,$$

where $Z \sim \mathrm{WN}(0, \sigma^2)$, $\phi_0^* := \mu(1 - \phi_1)$, and $\phi_1^* := \phi_1 - 1$. Let $\hat{\phi}_1^*$ be the ordinary least squares (OLS for short) estimator of ϕ_1^* found by regressing ∇X_t on 1 and X_{t-1}, i.e.,

$$(\hat{\phi}_0^*, \hat{\phi}_1^*) = \operatorname*{arg\,min}_{(\phi_0^*, \phi_1^*)} \sum_{t=2}^{n} (\nabla X_t - \phi_0^* - \phi_1^* X_{t-1})^2.$$

Then the estimated standard error of $\hat{\phi}_1^*$ can be shown to be

$$\widehat{\mathrm{SE}}(\hat{\phi}_1^*) = S \left(\sum_{t=2}^{n} (X_{t-1} - \bar{X}_{n-1})^2 \right)^{-1/2},$$

where

$$S^2 := (n-3)^{-1} \sum_{t=2}^{n} (\nabla X_t - \hat{\phi}_0^* - \hat{\phi}_1^* X_{t-1})^2$$

and we recall that $\bar{X}_{n-1}$ denotes the sample mean of $(X_1, \ldots, X_{n-1})$. Dickey and Fuller derived the limit distribution for $n \to +\infty$ of the t-ratio

$$\hat{\tau}_\mu := \frac{\hat{\phi}_1^*}{\widehat{\mathrm{SE}}(\hat{\phi}_1^*)}$$

under the unit root assumption $\hat{\phi}_1^* = 0$, from which a test of the null hypothesis $H_0 : \phi_1 = 1$ versus $H_1 : \phi_1 < 1$ and confidence regions can be constructed. The 0.01, 0.05, and 0.10 quantiles of the limit distribution of $\hat{\tau}_\mu$ are -3.43, -2.86, and -2.57, respectively, which can be found in [9, Table 8.5.2]. The augmented Dickey–Fuller test then rejects the null hypothesis of a unit root at level 0.05 if $\hat{\tau}_\mu < -2.86$.

Note that the cutoff value for this test statistic is much smaller than the standard cutoff value of -1.645 obtained from the normal approximation to the t-distribution, so that the unit root hypothesis is less likely to be rejected using the correct limit distribution.

The above procedure can be extended to the case where X follows an AR(p) model with mean μ given by

$$X_t - \mu = \sum_{j=1}^{p} \phi_j (X_{t-j} - \mu) + Z_t,$$

where $Z \sim \mathrm{WN}(0, \sigma^2)$. Similarly, the model can be rewritten as

$$\nabla X_t = \phi_0^* + \phi_1^* X_{t-1} + \sum_{j=2}^{p} \phi_j^* \nabla X_{t+1-j} + Z_t,$$

where

$$\begin{aligned}
\phi_0^* &:= \mu \left(1 - \sum_{i=1}^{p} \phi_i \right), \\
\phi_1^* &:= \sum_{i=1}^{p} \phi_i - 1, \\
\phi_j^* &:= -\sum_{i=j}^{p} \phi_i
\end{aligned}$$

for $j = 2, \ldots, p$, which is left to the reader as an *exercise*.

If the autoregressive polynomial has a unit root at 1, then $\phi_1^* = 0$ and the differenced series is an AR($p-1$) process. Consequently, we can do a similar procedure as in the AR(1) case, which can be applied recursively and which is summarized in the following method.

Method 3.3.2 (Dickey–Fuller test)**.** Estimate ϕ_1^* as the coefficient of X_{t-1} in the OLS regression of ∇X_t onto 1, X_{t-1}, ∇X_{t-1}, $\ldots$, ∇X_{t-1+p}. For large n the t-ratio

$$\hat{\tau}_\mu := \frac{\hat{\phi}_1^*}{\widehat{\mathrm{SE}}(\hat{\phi}_1^*)},$$

where $\widehat{\mathrm{SE}}(\hat{\phi}_1^*)$ is the estimated standard error of $\hat{\phi}_1^*$, has the same limit distribution as the AR(1) process with 0.01, 0.05, and 0.10 quantiles -3.43, -2.86, and -2.57, respectively. Test the null hypothesis $H_0 : \phi_1^* = 0$ and reject according to the chosen level. If a root is detected, repeat the procedure with the differenced process until rejection to determine d.

CHAPTER 4

ARCH and GARCH processes

In this section we introduce processes that are used to model volatility.

In the famous Black-Scholes framework, volatility is assumed to be constant over time to obtain the well-known equations. There, it is assumed that the price follows a geometric Brownian motion, i.e., it is the solution to the stochastic differential equation

$$dP_t = \mu P_t \, dt + \sigma P_t \, dB_t$$

with initial condition P_0 driven by a *Brownian motion* $B = (B_t, t \in \mathbb{R}_+)$, also known as *Wiener process*. The volatility σ is assumed to be a constant and the stochastic differential equation has

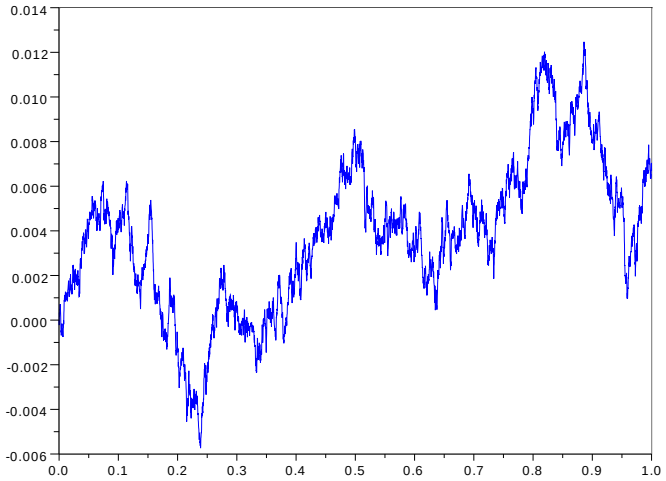

Figure 4.1: Path of a Brownian motion.

the explicit solution

$$P_t = P_0 \exp(\mu t + \sigma B_t).$$

Nevertheless, this does not seem to be the case in practice if one tests observed data. One option to measure the volatility is the *realized volatility*, which is computed by

$$\hat{\sigma}_t^2 := (\tau - 1)^{-1} \sum_{j=t-\tau}^{t} (x_j - \bar{x}_t)^2$$

for observed data $(x_1, \ldots, x_n)$, fixed $\tau < n$, and $\tau < t \leq n$, where

$$\bar{x}_t := \tau^{-1} \sum_{j=t-\tau}^{t} x_j.$$

The time frame for τ depends on the availability of data. If intra-day data is available, the time frame may be one day. For daily data it is typically 30 days.

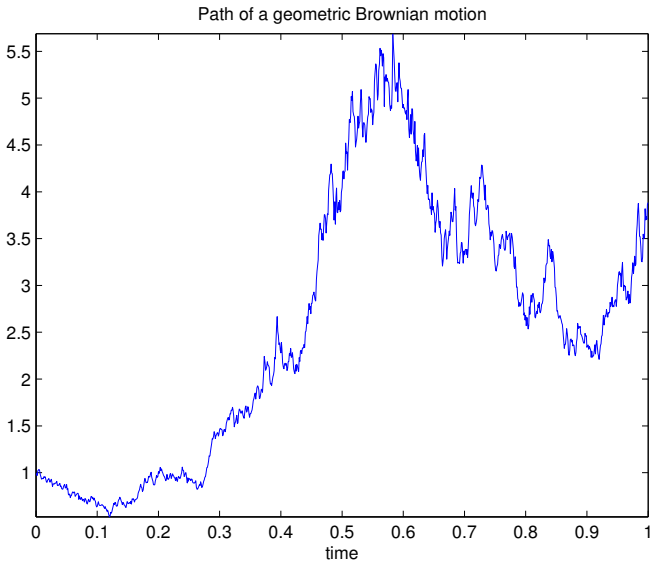

Figure 4.2: Path of a geometric Brownian motion with $P_0 = \mu = \sigma = 1$

The goal of this chapter is to introduce more general volatility models that allow for time dependent volatilities.

4.1 Definitions and properties

Let us start by introducing the necessary definitions and the theoretical background of ARCH and GARCH processes.

Definition 4.1.1. A stochastic process $X = (X_t, t \in \mathbb{Z})$ is called an ARCH(p) *process* if it is stationary and if

$$X_t = \sigma_t Z_t,$$

where $Z \sim \text{IID}\,\mathcal{N}(0,1)$,

$$\sigma_t^2 = \alpha_0 + \sum_{j=1}^{p} \alpha_j X_{t-j}^2,$$

$\alpha_0 > 0$, $\alpha_j \geq 0$ for $j = 1, \ldots, p$, and if Z_t and $(X_{t-j}, j \in \mathbb{N})$ are independent for all t.

Here the abbreviation ARCH stands for *autoregressive conditional heteroscedasticity*.

The requirements $\alpha_0 > 0$ and $\alpha_j \geq 0$, $j \geq 1$, guarantee that $\sigma_t > 0$. It is, however, not at all easy to find conditions on α_0 and α_j which ascertain that there really exists an ARCH(p) process.

Consider now an ARCH(p) process and the polynomial

$$\alpha(z) := \alpha_1 z + \cdots + \alpha_p z^p.$$

Thus we can rewrite the equation of the volatility σ_t to

$$\sigma_t^2 = \alpha_0 + \alpha(B)X_t^2,$$

where we recall that B denotes the backward shift operator introduced in Chapter 2. Due to stationarity and the fact that $\mathbb{E}(X_t^2) = \mathbb{E}(\sigma_t^2)$, which can be shown in an easy *exercise*, it holds that

$$\mathbb{E}(X_t^2) = \alpha_0 + \alpha(1)\,\mathbb{E}(X_t^2).$$

This implies that

$$\mathbb{E}(X_t^2) = \frac{\alpha_0}{1-\alpha(1)}.$$

It can be shown that $(X_t^2, t \in \mathbb{Z})$ is an AR process (see, e.g., [10]).

Since the order p of an ARCH process has to be rather large to be fitted to the observed data in practice, we now consider a generalization of ARCH processes, the so-called GARCH processes. This is one of many extensions of ARCH processes and certainly the most important one, where GARCH means *generalized* ARCH.

Definition 4.1.2. A stochastic process $X = (X_t, t \in \mathbb{Z})$ is called a GARCH(p, q) *process* if it is stationary and if

$$X_t = \sigma_t Z_t,$$

where $Z \sim \text{IID}\,\mathcal{N}(0,1)$,

$$\sigma_t^2 = \alpha_0 + \sum_{j=1}^{p} \alpha_j X_{t-j}^2 + \sum_{i=1}^{q} \beta_i \sigma_{t-i}^2,$$

with $\alpha_0 > 0$, $\alpha_j \geq 0$ for $j = 1, \ldots, p$, $\beta_i \geq 0$ for $i = 1, \ldots, q$, and if Z_t and $(X_{t-j}, j \in \mathbb{N})$ are independent for all t.

In the literature one finds that the GARCH$(1,1)$ process is often regarded to be a reasonably realistic model. Nevertheless, let us perform the following computations for GARCH(p, q) processes. Similarly to the ARCH process we rewrite the volatility equation to

$$\sigma_t^2 = \alpha_0 + \alpha(B)X_t^2 + \beta(B)\sigma_t^2,$$

where

$$\alpha(z) := \alpha_1 z + \cdots + \alpha_p z^p,$$
$$\beta(z) := \beta_1 z + \cdots + \beta_q z^q.$$

Again since $\mathbb{E}(X_t^2) = \mathbb{E}(\sigma_t^2)$ and due to stationarity we get

$$\mathbb{E}(X_t^2) = \alpha_0 + (\alpha(1) + \beta(1))\,\mathbb{E}(X_t^2),$$

which implies that

$$\mathbb{E}(X_t^2) = \frac{\alpha_0}{1-\alpha(1)-\beta(1)}.$$

Under the assumption that $\mathbb{E}(\sigma_t^4) < +\infty$ one can derive that $(X_t^2, t \in \mathbb{Z})$ is an ARMA$(\max\{p,q\}, q)$ process with generating polynomials

$$\phi(z) = 1 - \alpha(z) - \beta(z)$$

and

$$\theta(z) = 1 - \beta(z)$$

as well as mean $\alpha_0(1-\alpha(1)-\beta(1))^{-1}$ (see, e.g. [10]). The ARMA model can be represented by

$$X_t^2 = \alpha_0 + \sum_{i=1}^{\max\{p,q\}} (\alpha_i + \beta_i) X_{t-i}^2 + \eta_t - \sum_{j=1}^{q} \beta_j \eta_{t-j}, \tag{4.1}$$

where $\eta_t := X_t^2 - \sigma_t^2$. The interested reader checks in an *exercise* that $(\eta_t, t \in \mathbb{Z})$ is a martingale difference series, i.e., $\mathbb{E}(\eta_t) = 0$ and $\mathsf{Cov}(\eta_t, \eta_{t-j}) = 0$ for $j \neq 0$. However, the series is in general not an iid noise.

Since—as mentioned before—it is believed that a GARCH$(1,1)$ model is (often) sufficient, we state some properties of this specific choice of parameters.

The GARCH$(1,1)$ model is (weakly) stationary with $\mathsf{Cov}(X_t, X_s) = 0$ for $t \neq s$ if and only if $\alpha_1 + \beta_1 < 1$. Furthermore the $2m$-th moments of X exist if and only if

$$\sum_{j=0}^{m} \binom{m}{j} a_j \alpha_1^j \beta_1^{m-j} < 1,$$

where $a_0 := 1$ and $a_j := \prod_{i=1}^{j}(2i-1)$ for $j \geq 1$.

We close this section by remarking that uncertainty in volatility estimation is an important issue that is often overlooked. To assess the variability of an estimated volatility, it is necessary to consider the kurtosis of a volatility model (cf. [17, Section 3.16]). For the GARCH$(1,1)$ model it is given by

$$\frac{3(1-(\alpha_1+\beta_1)^2)}{1-\beta_1^2-2\alpha_1\beta_1-3\alpha_1^2} > 3,$$

where the lower bound of three can be easily computed.

4.2 Estimation

Let us introduce two methods in this section to do estimation for ARCH and GARCH models. While the first one tests given data for ARCH effects, the second one uses the ARMA representation to estimate the parameters α_i and β_i. For more methods the reader is referred to [17].

Method 4.2.1 (Test for ARCH effects)**.** Test the null hypothesis $H_0 : \alpha_1 = \cdots = \alpha_p = 0$.

Set

$$\mathrm{SSR}_0 := \sum_{t=p+1}^{n} (X_t^2 - \overline{X_n^2})^2,$$

where

$$\overline{X_n^2} := n^{-1} \sum_{t=1}^{n} X_t^2,$$

and

$$\mathrm{SSR}_1 := \sum_{t=p+1}^{n} \hat{e}_t^2,$$

where $\hat{e}_t$ is the residual from the least squares estimation of the regression model

$$X_t^2 = \alpha_0 + \sum_{j=1}^{p} \alpha_j X_{t-j}^2 + e_t$$

for $t = p+1, \ldots, n$.

Then the test statistic for ARCH effects is

$$\frac{(\mathrm{SSR}_0 - \mathrm{SSR}_1)(n-2p-1)}{\mathrm{SSR}_1 p},$$

which is asymptotically χ^2_p distributed.

Parameter estimation is often done using maximum likelihood estimation. For Gaussian noise, the log-likelihood function is given by

$$-2^{-1} \sum_{t=p+1}^{n} \left(\ln \sigma_t^2 + \sigma_t^{-2} X_t^2 \right),$$

which is maximized numerically.

Other noise distributions such as Student-t distribution or a generalized error distribution are also possible.

The following method uses the ARMA representation of a GARCH process. It provides often good approximations in practice but the statistical properties have not been investigated rigorously so far.

Method 4.2.2 (Two-pass estimation of GARCH)**.** Assume that a zero-mean set of observations $(x_j, j = 1, \ldots, n)$ is given. Use the maximum likelihood method 3.2.18 to estimate the parameters of the ARMA representation (4.1) for $(x_j^2, j = 1, \ldots, n)$, denoted by $\hat{\phi}_i$ and $\hat{\theta}_i$. Obtain the parameter estimates of the GARCH coefficients by setting

$$\hat{\beta}_i := \hat{\theta}_i \quad \text{and} \quad \hat{\alpha}_i := \hat{\phi}_i - \hat{\theta}_i.$$

4.3 Extensions

The ARCH and GARCH model both do not allow for asymmetries. Furthermore they have problems to treat extreme events. In the literature many modifications of the GARCH model have been proposed to overcome the problems of the model. We here just mention the *exponential* GARCH model, which is abbreviated by EGARCH and given by the formula

$$\ln(\sigma_t^2) = \alpha_0 + \sum_{i=1}^{p} \alpha_i \frac{|X_{t-i}| + \gamma_i X_{t-i}}{\sigma_{t-i}} + \sum_{j=1}^{q} \beta_j \ln(\sigma_{t-j}^2),$$

where the parameter γ_i signifies the leverage effect of X_{t-i} or accounts for skewness. In contrast to the GARCH model it allows for asymmetric effects.

Nevertheless, it should be mentioned at that point that the modeling of volatility is a difficult problem and it seems that no final satisfactory solution has been found so far.

CHAPTER 5

Nonlinear models

This section is mainly based on [17]. It is important to mention that white noise in [17] is called iid noise in [4] as well as here. The reader should be aware of this when looking for details in [17] and comparing it to the presented content of these lecture notes.

As seen in Chapter 3, a centered linear model can be expressed by

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

where $Z \sim \mathrm{WN}(0, \sigma^2)$ and $(\psi_j, j \in \mathbb{Z})$ is a sequence of real numbers.

This model might not always be sufficient for observed data. In this chapter we discuss more general models, how to test them and how to do forecasting, which becomes a lot more involved in this case than for linear models. Therefore let us consider the more general form of a time series model

$$X_t = f(Z_s, s \le t),$$

where f is some not necessary linear function. If we denote by $\mathcal{F}_t$ the sigma algebra generated by $(X_s, s \le t)$ and $(Z_s, s \le t)$, i.e., $(\mathcal{F}_t, t \in \mathbb{Z})$ is the *filtration* generated by X and Z, the conditional mean μ_t of X_t given $\mathcal{F}_{t-1}$ is given by

$$\mu_t = \mathbb{E}(X_t|\mathcal{F}_{t-1}) =: g(\mathcal{F}_{t-1})$$

and the conditional variance σ_t^2 by

$$\sigma_t^2 = \mathsf{Var}(X_t|\mathcal{F}_{t-1}) =: h(\mathcal{F}_{t-1}),$$

where g and h are well-defined functions and h is additionally positive. Let us restrict in what follows our class of nonlinear models to those which can be written as

$$X_t = g(\mathcal{F}_{t-1}) + \sqrt{h(\mathcal{F}_{t-1})}\epsilon_t,$$

where $\epsilon_t = Z_t/\sigma_t$ is a standardized shock (or innovation). Furthermore assume for simplicity that Z is iid noise, i.e., $\epsilon \sim \mathrm{IID}(0, 1)$. If g is nonlinear, the model is called *nonlinear in mean*, while it is called *nonlinear in variance* if h is time variant. The models in Chapter 3 are linear. One can show that those introduced in Chapter 4 are nonlinear in variance.

5.1 Nonlinear models

In this section we introduce the bilinear and the Markov switching model as two examples of nonlinear models. For more examples the reader is referred to [17].

The basic idea of *bilinear models* is to extend linear models, which can be seen as first-order Taylor expansion of nonlinear models, by the second-order Taylor terms. This leads to

$$X_t = c + \sum_{i=1}^{p} \phi_i X_{t-i} - \sum_{j=1}^{q} \theta_j Z_{t-j} + \sum_{i=1}^{m} \sum_{j=1}^{s} \beta_{ij} X_{t-i} Z_{t-j} + Z_t,$$

where p, q, m, and s are nonnegative integers and Z is a white noise. This model was introduced by Grander and Andersen [11] and has been widely investigated.

The second nonlinear model that we want to mention here is the so called Markov switching autoregressive model. In order to introduce it, we first have to give the definition of a Markov process, which is one of the important concepts in stochastic analysis.

Definition 5.1.1. A stochastic process $X = (X_t, t \in \mathbb{T})$ on some index set $\mathbb{T}$ is a *Markov process* if its conditional distribution function satisfies

$$P(X_h | X_s, s \leq t) = P(X_h | X_s)$$

for arbitrary $h > t$. If X is a discrete-time stochastic process, i.e., $\mathbb{T} = \mathbb{N}$ or $\mathbb{Z}$, then the property becomes

$$P(X_h | X_t, X_{t-1}, \ldots) = P(X_h | X_t)$$

for arbitrary $h > t$ and the process is also known as (first-order) *Markov chain*.

The following model uses the concept of Markov processes or Markov chains.

Definition 5.1.2. A time series $X = (X_t, t \in \mathbb{Z})$ follows a *Markov switching autoregressive model* (MSA for short) if it satisfies

$$X_t = \begin{cases} c_1 + \sum_{i=1}^{p} \phi_{1i} X_{t-i} + Z_{1t} & \text{if } S_t = 1, \\ c_2 + \sum_{i=1}^{p} \phi_{2i} X_{t-i} + Z_{2t} & \text{if } S_t = 2, \end{cases}$$

where S assumes values in $\{1, 2\}$ and is a first-order Markov chain with transition probabilities

$$P(S_t = 2 | S_{t-1} = 1) = w_1,$$
$$P(S_t = 1 | S_{t-1} = 2) = w_2$$

with $w_1, w_2 \in [0, 1]$. The innovational series $Z_1 = (Z_{1t}, t \in \mathbb{Z})$ and $Z_2 = (Z_{2t}, t \in \mathbb{Z})$ are $\text{IID}(0, \sigma^2)$ noise for finite σ^2 and independent of each other.

5.2 Nonparametric methods for model fitting

Nonparametric methods are highly data dependent and can easily result in overfitting. They are used if there is no sufficient knowledge about the nonlinear structure between random variables. The essence of nonparametric methods is smoothing. To get an idea of the problem, let us look into the following:

Assume that we are given two time series X and Y that are related by

$$Y_t = m(X_t) + Z_t, \tag{5.1}$$

where m is an arbitrary, smooth, but unknown function and Z is iid noise. Our goal is to estimate the nonlinear function m from the data. Let for the beginning $X = x$ be constant and $(y_t, t = 1, \ldots, n)$ be a series of independent observations. Then the problem simplifies to

$$y_t = m(x) + Z_t$$

and taking the sample average yields

$$n^{-1}\sum_{t=1}^{n} y_t = m(x) + n^{-1}\sum_{t=1}^{n} Z_t.$$

By the properties of the iid noise and the law of large numbers, the averaged noise converges to zero for large n. Therefore

$$\bar{y} := n^{-1}\sum_{t=1}^{n} y_t$$

is a consistent estimator for $m(x)$, i.e., $\bar{y} \approx m(x)$.

As long as m is sufficiently smooth and $X_t \approx x$ still almost constant, the method continues to work fine. In other cases one possibility is to use a weighted average of y instead of the simple one, which we denote by

$$\hat{m}(x) := \sum_{t=1}^{n} w_t(x) y_t, \tag{5.2}$$

where the weights $w_t(x)$ are larger for those y_t with x_t close to x and smaller for those far away. The weights sum up to one.

We introduce two methods to determine the weights in what follows.

Method 5.2.1 (Kernel regression)**.** This method determines the weights by a kernel, which is typically a probability density function denoted by K and which satisfies that it is nonnegative and

$$\int K(z)\,dz = 1.$$

To increase the flexibility in distance measure, the kernel is often rescaled by the *bandwidth* $h > 0$ and becomes

$$K_h(x) = h^{-1}K(xh^{-1})$$

and

$$\int K_h(z)\,dz = 1.$$

Define the weight function by

$$w_t(x) := \frac{K_h(x - x_t)}{\sum_{s=1}^{n} K_h(x - x_s)}.$$

Plugging this into Equation (5.2), the *Nadaraya–Watson kernel estimator*

$$\hat{m}(x) = \sum_{t=1}^{n} w_t(x) y_t = \frac{\sum_{t=1}^{n} K_h(x - x_t) y_t}{\sum_{t=1}^{n} K_h(x - x_t)}$$

is obtained (see [13, 18]). Possible choices of the kernel include the *Gaussian kernel*

$$K_h(x) := h^{-1}(2\pi)^{-1/2}\exp(-2^{-1}(x/h)^2)$$

and the *Epanechnikov kernel* [7]

$$K_h(x) := 0.75\,h^{-1}(1 - (x/h)^2)I(|x/h| \le 1),$$

where I denotes the indicator function, i.e., $I(A) = 1$ if A holds and $I(A) = 0$ else.

To understand the role of the bandwidth h one observes that $\hat{m}(x_t) \to y_t$ for $h \to 0$ and $\hat{m}(x_t) \to \bar{y}$ for $h \to +\infty$. Therefore one could regard h as the parameter that chooses the size of the neighborhood that is used for smoothing. In general bandwidth selection is a well-known problem in kernel regression. In what follows we introduce two methods to determine a "good" choice for h. For an overview to bandwidth selection, the reader is referred to Härdle [12] as well as Fan and Yao [8].

Method 5.2.2 (Bandwidth selection with MISE)**.** This method is a plug-in method, which is based on the asymptotic expansion of the *mean integrated squared error* (MISE for short) for kernel smoothers

$$\mathrm{MISE} := \mathbb{E}\left(\int_{-\infty}^{\infty} (\hat{m}(x) - m(x))^2\, dx\right),$$

where m is the true function and $\hat{m}$ the estimator which depends on h. Under some regularity conditions, one derives the optimal bandwidth by minimization of the MISE, which typically depends on several unknown quantities that must be estimated from the data with some preliminary smoothing. In practice the choice of preliminary smoothing can become a problem. A normal reference bandwidth selector is given by Fan and Yao by

$$\hat{h}_{\mathrm{opt}} = \begin{cases} 1.06\, s\, n^{-1/5} & \text{for the Gaussian kernel,} \\ 2.34\, s\, n^{-1/5} & \text{for the Epanechnikov kernel,} \end{cases}$$

where s is the sample standard error of the independent variable, which is assumed to be stationary.

Method 5.2.3 (Bandwidth selection with cross validation)**.** The *leave-one-out cross validation* starts with omitting one observation (x_j, y_j). The remaining $n-1$ data points are used to obtain the following smoother at x_j:

$$\hat{m}_{h,j}(x_j) := \sum_{t \neq j} w_t(x_j) y_t,$$

which is an estimate of y_j where the weights $w_t(x_j)$ sum to 1. Afterwards the same is performed for all remaining $n-1$ observations and

$$\mathrm{CV}(h) := \sum_{j=1}^{n} (y_j - \hat{m}_{h,j}(x_j))^2 W(x_j)$$

is defined, where W is a nonnegative weight function satisfying $\sum_{j=1}^{n} W(x_j) = 1$ that can be used to down-weight the boundary points if necessary. This might be the case since points at the boundary have often fewer neighboring observations. The function CV is called the *cross-validation function* because it validates the ability of the smoother to predict y. The bandwidth h is chosen such that it minimizes CV.

Having presented two methods to choose the bandwidth in kernel regression, we continue with another method to estimate m in Equation (5.1).

Method 5.2.4 (Local linear regression method)**.** Assume that m in Equation (5.1) is twice continuously differentiable at some given point x in the support of m. Denote the available observations by $((y_t, x_t), t = 1, \ldots, n)$. The *local linear regression method* to nonparametric regression is to find a and b that minimize

$$L(a, b) := \sum_{t=1}^{n} (y_t - a - b(x - x_t))^2 K_h(x - x_t),$$

where K_h is a kernel with bandwidth h as in Method 5.2.2. Denote the minimum of a by $\hat{a}$, which is the estimate of $m(x)$, while the minimum of b denoted by $\hat{b}$ can be used as an estimate of $m'(x)$. The least-squares problem has a closed-form solution, which is given by

$$\hat{a} = \frac{\sum_{t=1}^{n} w_t(x) y_t}{\sum_{t=1}^{n} w_t(x)},$$

where

$$w_t(x) := K_h(x - x_t)(s_{n,2}(x) - (x - x_t) s_{n,1}(x))$$

and

$$s_{n,j}(x) := \sum_{t=1}^{n} K_h(x - x_t)(x - x_t)^j$$

for $j = 1, 2$. We leave the derivation to the interested reader.

In practice, to avoid that the denominator becomes zero,

$$\hat{m}(x) := \frac{\sum_{t=1}^{n} w_t(x) y_t}{\sum_{t=1}^{n} w_t(x) + n^{-2}}$$

is used as an estimate for $m(x)$.

5.3 Nonlinearity tests

In this section we discuss both, nonparametric and parametric statistics that have decent power against the models considered in Section 5.1.

5.3.1 Nonparametric tests

Under the null hypothesis of linearity, residuals of a properly specified linear model should be uncorrelated. Let us here consider the stronger assumption that they are independent, which holds automatically true for Gaussian noise. Any violation of independence in the residuals indicates inadequacy of the entertained model, including the linearity assumption. This is the basic idea behind various nonlinearity tests. Here we introduce two methods for the moment. For more the reader is referred to the literature.

Method 5.3.1 (Q-statistic of squared residuals)**.** This method by McLeod and Li applies the Ljung–Box statistics 2.2.10 to the squared residuals of an ARMA(p, q) model to check for model inadequacy. The test statistic is

$$Q(m) := n(n + 2) \sum_{i=1}^{m} \frac{\hat{\rho}_i^2(Z_t^2)}{n - i},$$

where n is the number of observations, m is a properly chosen number of autocorrelations used in the test, $(Z_t, t = 1, \ldots, n)$ denotes the residual series, and $\hat{\rho}_i(Z_t^2)$ is the lag i autocorrelation function of Z_t^2. If the entertained linear model is adequate, $Q(m)$ is asymptotically χ^2_{m-p-q}-distributed.

The null hypothesis of the test is

$$H_0 : \beta_1 = \cdots = \beta_m = 0,$$

where the parameter β_i is the coefficient of Z_{t-i}^2 in the linear regression

$$Z_t^2 = \beta_0 + \sum_{i=1}^{m} \beta_i Z_{t-i}^2 + e_t$$

for $t = m + 1, \ldots, n$.

Method 5.3.2 (Bispectral test)**.** This test can be used to test for linearity and Gaussianity. It depends on the result that a properly normalized bispectrum of a linear time series is constant over all frequencies and that the constant is equal to zero under normality. Here, the *bispectrum* of a time series is the Fourier transform of its third-order moments, but let us treat this in detail in what follows.

For a stationary time series

$$X_t = \mu + \sum_{i=0}^{\infty} \psi_i Z_{t-i},$$

where μ is a constant, $Z \sim \mathrm{IID}(0, \sigma^2)$ and $(\psi_j, j \in \mathbb{Z})$ is a sequence of real numbers with $\psi_0 = 1$, the third-order moment is defined as

$$c(u, v) := \mathbb{E}(Z_t^3) \sum_{k=-\infty}^{\infty} \psi_k \psi_{k+u} \psi_{k+v}$$

for $u, v \in \mathbb{Z}$, where we set $\psi_0 := 1$ and $\psi_i := 0$ for $k < 0$. For frequencies w_1 and w_2 the Fourier transform is then given by

$$b_3(w_1, w_2) := \frac{\mathbb{E}(Z_t^3)}{4\pi^2} \Gamma(-(w_1 + w_2)) \Gamma(w_1) \Gamma(w_2),$$

where Γ is defined by

$$\Gamma(w) := \sum_{u=0}^{\infty} \psi_u \exp(-iwu)$$

and $i = \sqrt{-1}$. Since the spectral density of X is given by

$$p(w) = \frac{\sigma^2}{2\pi} |\Gamma(w)|^2,$$

one obtains that the *bispectrum*

$$b(w_1, w_2) := \frac{|b_3(w_1, w_2)|^2}{p(w_1) p(w_2) p(w_1 + w_2)}$$

is constant for all (w_1, w_2). The bispectrum test estimates b over a suitably chosen grid of points and applies a test statistic similar to *Hotelling's T^2 statistic* to check the constancy. Since for a linear Gaussian series $\mathbb{E}(Z_t^3) = 0$, the bispectrum is zero for all frequencies.

5.3.2 Parametric tests

To conclude the section about nonlinearity tests we introduce one parametric method and remark at the end how this can be extended.

Method 5.3.3 (RESET test)**.** Ramsey [15] proposes a specification test for linear least-squares regression analysis referred to as a *RESET test*, which is readily applicable to linear AR models. Therefore consider the linear AR(p) model

$$X_t = (1, X_{t-1}, \ldots, X_{t-p})(\phi_0, \phi_1, \ldots, \phi_p)' + Z_t.$$

The first step of the RESET test is to obtain the least-squares estimate $(\hat{\phi}_0, \hat{\phi}_1, \ldots, \hat{\phi}_p)$ and compute the fit

$$\hat{X}_t := (1, X_{t-1}, \ldots, X_{t-p})(\hat{\phi}_0, \hat{\phi}_1, \ldots, \hat{\phi}_p)',$$

the residuals $\hat{Z}_t := X_t - \hat{X}_t$, and the sum of squared residuals

$$\mathrm{SSR}_0 := \sum_{t=p+1}^{n} \hat{Z}_t^2,$$

where n is as usual the sample size.

In the second step, consider the linear regression

$$\hat{Z}_t = (1, X_{t-1}, \ldots, X_{t-p})(\alpha_{10}, \ldots, \alpha_{1p})' + (\hat{X}_t^2, \ldots, \hat{X}_t^{s+1})(\alpha_{21}, \ldots, \alpha_{2s})' + V_t$$

for some $s \geq 1$ and compute the least-squares residuals

$$\hat{V}_t = \hat{Z}_t - (1, X_{t-1}, \ldots, X_{t-p})(\hat{\alpha}_{10}, \ldots, \hat{\alpha}_{1p})' - (\hat{X}_t^2, \ldots, \hat{X}_t^{s+1})(\hat{\alpha}_{21}, \ldots, \hat{\alpha}_{2s})'$$

and the sum of squared residuals

$$\mathrm{SSR}_1 := \sum_{t=p+1}^{n} \hat{V}_t^2$$

of the regression. The idea of the RESET test is that if the linear AR(p) model is adequate, then all α_{1i} and α_{2j} should be zero. This can be tested by using the F statistic given by

$$F := \frac{(\mathrm{SSR}_0 - \mathrm{SSR}_1)(n-p-g)}{\mathrm{SSR}_1 g},$$

where $g := s + p + 1$, which under linearity and normality assumption, has an F distribution with degrees of freedom g and $n - p - g$.

We remark that there exist several improvements of the RESET test. We here mention only the modification of the second step of the RESET test by Keenan and a different choice of the regressor by Tsay. For details the reader is referred to the literature.

5.4 Forecasting

We have seen in Section 2.3 that forecasting for linear time series can be done with closed-form formulas. This does not hold for most nonlinear models when the forecast horizon is greater than 1. In what follows we introduce parametric bootstraps to compute nonlinear forecasts.

Method 5.4.1 (Parametric bootstrap)**.** Given X_n, we want to forecast X_{n+h} for some $h > 0$. The parametric bootstrap computes realizations $X_{n+1}, \ldots, X_{n+h}$ sequentially in the following way. For $i = 1, \ldots, h$ repeat:

(i) Generate a random sample of the innovation at time $n+i$ according to the underlying model.

(ii) Compute X_{n+i} using the generated sample, the model, the data, and the previous forecasts $X_n, \ldots, X_{n+i-1}$.

Repeat this whole procedure M times to get M realizations of X_{n+h}. Compute the sample average over the realizations to obtain a point forecast for X_{n+h} which we denote by $X_n(h)$. The latter is also known as *Monte Carlo simulation.*

The realizations could also be used to obtain an empirical distribution function which might be of use in the following methods when forecasting evaluation is done.

In what follows we introduce different methods to evaluate the performance of a forecast. Therefore let us do the following: Given a data set, we subdivide it into two subsamples which we refer to as *estimation subsample* and *forecasting subsample.* We will use the first one to build a nonlinear model. We derive the performance then by comparing the obtained forecasts computed by the model with the data of the forecasting subsample. In what follows three measures are used to get an idea of the performance which are commonly used in the literature. Nevertheless, we should mention that there exists no widely accepted measure to compare models.

Method 5.4.2 (Directional measure)**.** A typical measure for the evaluation of the performance of forecasts is to use a 2×2 contingency table that summarizes the number of "hits" and "misses" of the model in predicting ups and downs up to x_{n+h} in the forecasting subsample. Let the table be given by

Actual	Predicted		
	Up	Down	
Up	m_{11}	m_{12}	m_{10}
Down	m_{21}	m_{22}	m_{20}
	m_{01}	m_{02}	m

where m is the total number of h-step-ahead forecasts in the forecasting subsample, m_{11} is the number of "hits" in predicting upward movements, m_{21} is the number of "misses" in predicting downward movements of the market, and so on. It is clear that larger values in m_{11} and m_{22} indicate better forecasts. The test statistic

$$\chi^2 := \sum_{i,j=1}^{2} \frac{(m_{ij} - m_{i0}m_{0j}/m)^2}{m_{i0}m_{0j}/m}$$

can be used to evaluate the performance of the model, where a large χ^2 signifies that the model outperforms the chance of random choice. Under some assumptions, χ^2 has an asymptotic χ^2 distribution with one degree of freedom. For more details the reader is referred to the literature, especially to [5].

Method 5.4.3 (Magnitude measure)**.** Three statistics that are commonly used to measure performance of point forecasts are

- the *mean squared error*

$$\mathrm{MSE}(h) := m^{-1} \sum_{j=0}^{m-1} (X_{n+h+j} - X_{n+j}(h))^2,$$

- the *mean absolute deviation*

$$\mathrm{MAD}(h) := m^{-1} \sum_{j=0}^{m-1} |X_{n+h+j} - X_{n+j}(h))|,$$

- the *mean absolute percentage error*

$$\mathrm{MAPE}(h) := m^{-1} \sum_{j=0}^{m-1} \left| \frac{X_{n+j}(h)}{X_{n+h+j}} - 1 \right|,$$

where m is the number of h-step-ahead forecasts available in the forecasting subsample. The error computation is done between the data from the forecasting subsample and the h-step-ahead forecasts computed from the model that was derived from the estimation subsample.

In application one often chooses one of the above measures and then the model with the smallest magnitude on that measure. This is regarded as the best h-step-ahead forecasting model. Be aware that it might happen that different models are chosen for different forecast horizons h. For limitations in model comparison of the different measures, the reader is referred to the literature.

Method 5.4.4 (Distributional measure)**.** Compute the empirical distribution function $\hat{F}$ out of the sample obtained by the parametric bootstrap method 5.4.1 for the desired h-step-ahead forecast according to the model obtained from the estimation subsample. Use the forecasting subsample to compute

$$u_{n+j}(h) := \hat{F}(x_{n+h+j})$$

for all $j = 0, \ldots, m-1$, where m denotes the total number of h-step-ahead forecasts in the forecasting subsample. If the model is adequate, then $(u_{n+j}(h), j = 0, \ldots, m-1)$ will behave like a random sample from the uniform distribution on $[0, 1]$. For sufficiently large m, the Kolmogorov–Smirnov statistic can be used to test the sample with respect to the uniform distribution.

The method can be used for both, model checking and forecasting comparison.

At the end of this chapter it should be mentioned that in contrast to linear models, where we proved most of the results and methods or referred for the proofs to the literature, prediction with nonlinear models is a lot more heuristically. More complicated models lead to more complicated theory—if available. This implies that extra care has to be taken into account when applying these methods in practice possibly without completely understanding them. Consequences of that (as one reason) could be seen in the financial crisis.

Bibliography

[1] Craig F. Ansley. An algorithm for the exact likelihood of a mixed autoregressive-moving average process. *Biometrika*, 66:59–65, 1979.

[2] George E. P. Box and Gwilym M. Jenkins. *Time Series Analysis: Forecasting and Control*. Holden-Day, San Francisco, Calif.-Düsseldorf-Johannesburg, revised edition, 1976. Holden-Day Series in Time Series Analysis.

[3] Peter J. Brockwell and Richard A. Davis. *Time Series: Theory and Methods*. Springer Series in Statistics. Springer-Verlag, New York, second edition, 1991.

[4] Peter J. Brockwell and Richard A. Davis. *Introduction to Time Series and Forecasting*. Springer Texts in Statistics. New York, NY: Springer, 2nd edition, 2002.

[5] Christian M. Dahl and Svend Hylleberg. Specifying nonlinear econometric models by flexible regression models and relative forecast performance. Working paper, Department of Economics, University of Aarhus, Denmark, 1999.

[6] David A. Dickey and Wayne A. Fuller. Distribution of the estimators for autoregressive time series with a unit root. *J. Am. Stat. Assoc.*, 74:427–431, 1979.

[7] V.A. Epanechnikov. Nonparametric estimates of a multivariate probability density. *Theory of Probability and Its Applications*, 14:153–158, 1969.

[8] Jianqing Fan and Qiwei Yao. *Nonlinear Time Series. Nonparametric and Parametric Methods*. New York, NY: Springer, 2003.

[9] Wayne A. Fuller. *Introduction to Statistical Time Series*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, 1976.

[10] Jan Grandell. Time series analysis. Lecture notes, 2011.

[11] Clive William John Granger and Allan Paul Andersen. An introduction to bilinear time series models. Angewandte Statistik und Ökonometrie. Heft 8. Göttingen: Vandenhoeck & Ruprecht, 1978.

[12] Wolfgang Härdle. *Applied Nonparametric Regression*, volume 19 of *Econometric Society Monographs*. Cambridge: Cambridge University Press, 1991.

[13] Èlizbar A. Nadaraya. On estimating regression. *Theory and Probability Application*, 10:186–190, 1964.

[14] Jürgen Potthoff. Einführung in die Wahrscheinlichkeitstheorie. Lecture notes for an introductory course in probability theory.

[15] James B. Ramsey. Tests for specification errors in classical linear least-squares regression analysis. *J. R. Stat. Soc., Ser. B*, 31:350–371, 1969.

[16] Sheldon M. Ross. *A First Course in Probability.* Pearson, 9th edition, 2014.

[17] Ruey S. Tsay. *Analysis of Financial Time Series.* Wiley Series in Probability and Statistics. Hoboken, NJ: John Wiley & Sons, 3rd edition, 2010.

[18] Geoffrey S. Watson. Smooth regression analysis. *Sankhyā, Ser. A*, 26:359–372, 1964.