Financial Time Series

Annika Lang & Andreas Petersson

version May 11, 2021

Preface to v2020

Compared to last year, some changes were made to the material. The Student's t-distribution was added to the introductory chapter. The section on the sample autocorrelation function and the Ljung–Box test was updated with a new example. The section on trend and seasonality was restructured with the addition of applications for all trend smoothing methods to real data, along with an explanation of how both trend and seasonal components can be estimated by linear least squares. New material on how to compute forecasts for data which have been differenced to make it stationary was added. Finally, the section on ARIMA processes was extended to SARIMA processes and some details were changed in the chapter on GARCH processes.

Göteborg, May 2020, Andreas Petersson

Preface to v2019

Some additional introductory material was added and some forecasting and estimation algorithms for linear time series were removed. Details were added for the theoretical results on linear times series. The chapter on GARCH processes was significantly expanded. Some clarifications were added in the chapter on non-linear time series.

Göteborg, May 2019, Andreas Petersson

Preface to v2018

Minor changes and corrections were done.

Göteborg, May 2018, Annika Lang & Andreas Petersson

Preface to v2017

The lecture notes are updated and references are given with respect to the new third edition of "Introduction to Time Series and Forecasting" by Brockwell and Davis [6]. Furthermore, new examples were added to support the understanding of the theory.

Göteborg, May 2017, Annika Lang & Andreas Petersson

Preface to v2016

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 [5, 4, 13, 22].

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 [17]. For a more extended but still easy introduction to probability theory in English than that given below the reader is referred for example to [20]. A nice text to extend a statistical point of view to a probabilistic one can be found in [18]. 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 $\mu : \mathcal{A} \to \mathbb{R}_+ \cup \{\infty\}$ 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 note that if X i non-negative, i.e., if $P(X < 0) = 0$, and if $\mathbb{E}(X) = 0$ then, as an easy consequence of the definition of the expectation, $X = 0$ *P-almost surely*. This means that $P(X = 0) = 1$. As we have already learned, $g \circ X$ 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 with finite variance, 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. The covariance is well-defined as a consequence of the *Cauchy–Schwarz equality*, which says that $|\,\mathbb{E}(XY)|^2 \leq \mathbb{E}(|X||Y|)^2 \leq \mathbb{E}(|X|^2)\,\mathbb{E}(|Y|^2)$.

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
$$\mathbb{E}(a_1 X_1 + a_2 X_2 | Y) = a_1 \, \mathbb{E}(X_1|Y) + a_2 \, \mathbb{E}(X_2|Y).$$

(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). In addition, it holds that

(vi) if the random variables $(X, Y_1, \ldots, Y_n)$ are all independent of the random variable Z, then
$$\mathbb{E}(X|Y_1, \ldots, Y_n, Z) = \mathbb{E}(X|Y_1, \ldots, Y_n).$$

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

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

for $x \in \mathbb{R}$. 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 (see the concept of a joint distribution function below).

Example 1.1.7 (Student's t-distribution)**.** A random variable X has a *Student's t-distribution with $\nu > 0$ degrees of freedom* (denoted by $X \sim t_\nu$) if it is continuously distributed with density given by

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

for $x \in \mathbb{R}$, where Γ is the gamma function. If $\nu > 1$, $\mathbb{E}(|X|) < \infty$ and $\mathbb{E}(X) = 0$. If $\nu > 2$, $\mathbb{E}(X^2) < \infty$ and $\mathsf{Var}(X) = \nu/(\nu-2)$. In order to be able to specify what the mean and variance of a t-distributed random variable is, we say that X has a *generalized Student's t-distribution with $\nu > 2$ degrees of freedom* with mean μ and variance σ^2 if

$$Y := \frac{(X-\mu)\sqrt{\nu}}{\sigma\sqrt{\nu-2}} \sim t_\nu.$$

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 [20]. 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 1.1.8 (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 $x \in \mathbb{R}$*

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

The concept of random variables is straightforward to extend to random vectors. A *random vector* in $\mathbb{R}^n$ is an n-dimensional vector $X = (X_1, X_2, \ldots, X_n)'$ whose components X_i, $i = 1, 2, \ldots, n$, are random variables. The cumulative distribution function or *joint distribution function* of X, F_X, is then defined by

$$F_X(x) = F_X(x_1, x_2, \ldots, x_n) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n)$$

where $x = (x_1, x_2, \ldots, x_n)'$. For continuously distributed random vectors, the density function of X is defined by the equation

$$F_X(x) = F_X(x_1, x_2, \ldots, x_n) = \int_{-\infty}^{x_n} \cdots \int_{-\infty}^{x_2} \int_{-\infty}^{x_1} f(y_1, \ldots, y_n) \mathrm{d}y_1 \mathrm{d}y_2 \ldots \mathrm{d}y_n$$

if there exists such a function $f : \mathbb{R}^n \to \mathbb{R}$. The concept of the mean extends to random vectors via

$$\mathbb{E}(X) = (\mathbb{E}(X_1), \mathbb{E}(X_2), \ldots, \mathbb{E}(X_n))' \tag{1.2}$$

and the notion of variance corresponds to the *covariance matrix* $\mathsf{Cov}(X) \in \mathbb{R}^{n\times n}$ defined by

$$\mathsf{Cov}(X) = \mathbb{E}((X - \mathbb{E}(X))(X - \mathbb{E}(X))'), \tag{1.3}$$

which can also be written as $\mathsf{Cov}(X) = (\mathsf{Cov}(X_i, X_j))_{i,j=1}^n$.

The most important example of a random vector will for us be a *Gaussian random vector*, i.e., a random vector that follows the multivariate normal distribution.

Example 1.1.9 (Multivariate normal distribution)**.** X has a *multivariate normal distribution* with mean $\mu = \mathbb{E}(X)$ and non-singular covariance matrix $\Sigma = \mathsf{Cov}(X)$, which we write $X \sim \mathcal{N}(\mu, \Sigma)$, if it has the density function

$$f_X(x) = \frac{1}{\sqrt{2\pi^n \det \Sigma}} \exp\left(-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right). \tag{1.4}$$

A useful fact is that a random vector $X = (X_1, X_2, \ldots, X_n)'$ has a multivariate normal distribution if and only if the random variable $\sum_{i=1}^n a_i X_i$ is normally distributed is normally distributed for any vector $a \in \mathbb{R}^n$. Note that, by Example 1.1.6, X is a vector of independent Gaussian random variables if and only if $\Sigma = I$, where I is the identity matrix.

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. The *level* of a hypothesis test is the probability that H_0 is rejected given that H_0 is true. The *power* of a hypothesis test is the probability that H_0 is rejected given that H_1 is true.

CHAPTER 2

Stationary time series and seasonality

This chapter is based on Brockwell and Davis' book "Introduction to Time Series and Forecasting" [6] as well as Grandell's lecture notes "Time series analysis" [13], which are based on [6] and [4]. d 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 $\mathbb{T}$, usually a subset of $\mathbb{R}$. 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, e.g., $\mathbb{T} = \{t_n, n \in \mathbb{N}\}$ 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 $(x_t, t = 1, \ldots, 42)$ for the Swedish clothing company H&M. Figure 2.1 shows these earnings in million SEK from January 2006 through April 2016.

Example 2.1.3. Another example of a type of time series, that will be important for us in the latter half of the course, are *log-returns*, or *continuously compounded returns* $(r_t, t \in \mathbb{Z})$. These are

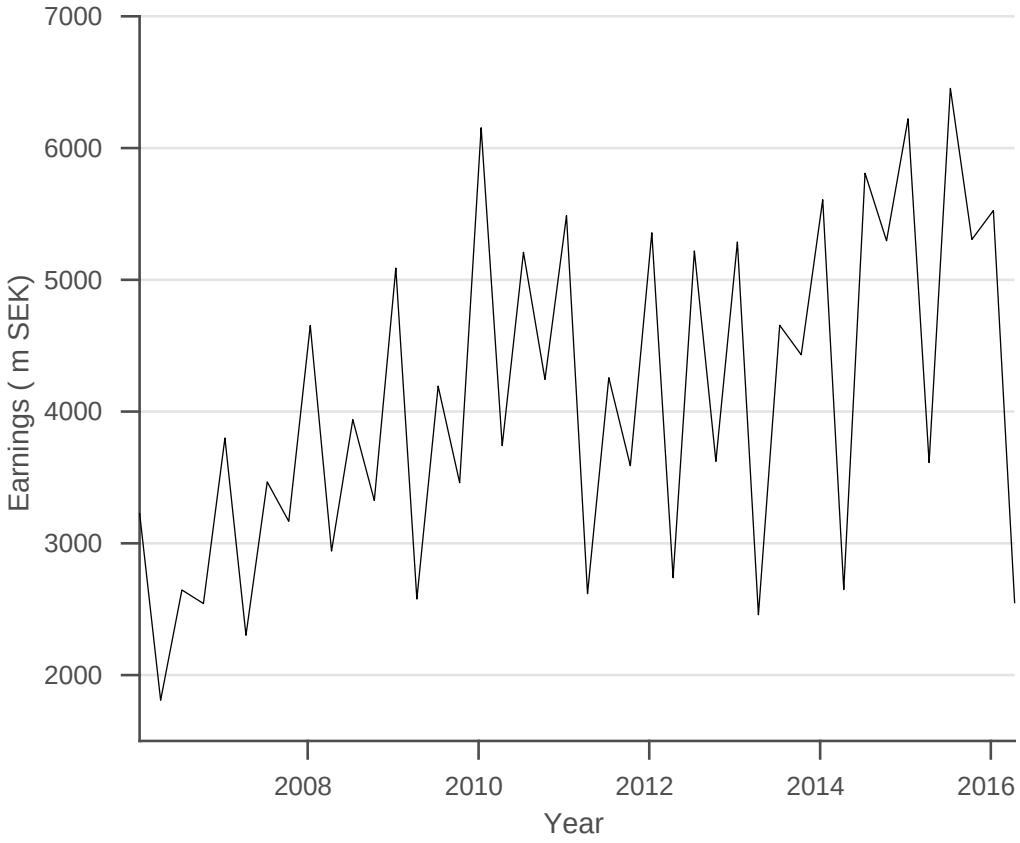


Figure 2.1: Quarterly earnings of H&M from January 2006 through April 2016.

defined by $r_t = \log(P_t) - \log(P_{t-1})$, where P_t is the price of some asset at time $t \in \mathbb{Z}$. In Figure 2.2, we plot daily log-returns $(r_t, t = 1, \dots 2584)$ of the S&P500 index from January 2006 to April 2016. At each point t, r_t is the logarithm of the percentage (divided by 100) that a stock portfolio allocated according to this index has increased or decreased from one month to the next.

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 and allow also for infinitely many observations, to allow for the arrival of new data. 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 $(X_n, n \in \mathbb{N})$ or $(X_n, n \in \mathbb{Z})$ if we want to account for unlimited historical data. 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.

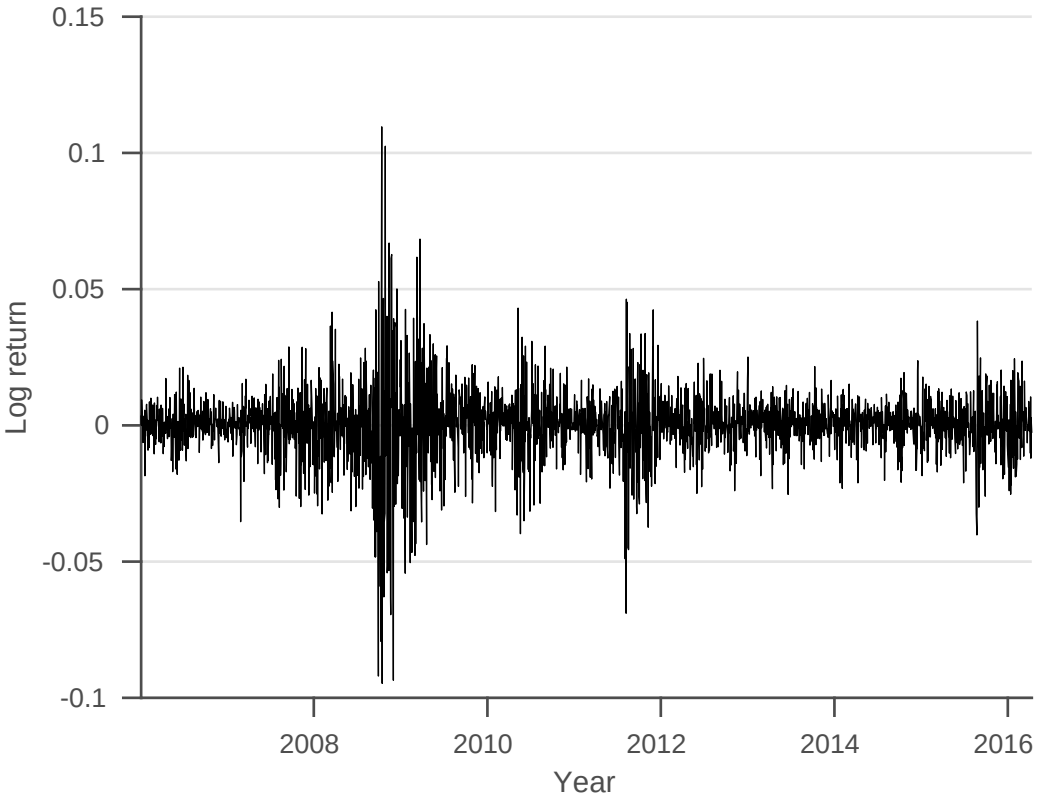


Figure 2.2: Daily log-returns of the S&P500 index from January 2006 to April 2016.

Definition 2.1.4. A stochastic process $X = (X_t, t \in \mathbb{T})$ is called *iid noise* with mean μ 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 $\mathbb{E}(X_t) = \mu$ and $\mathsf{Var}(X_t) = \sigma^2$ for all $t \in \mathbb{T}$. An iid noise is denoted by $X \sim \mathrm{IID}(\mu, \sigma^2)$.

Please note that iid noise is sometimes called *white noise* in the literature (e.g., in [22]). 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 2.1.5 (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 $(X_n, n \in \mathbb{Z})$ 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 2.1.6 (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 $(X_n, n \in \mathbb{Z})$, 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 2.1.5, 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 2.1.7. A time series X is said to be a *Gaussian time series* if all finite-dimensional distributions are normal, i.e., all finite-dimensional vectors are multivariate Gaussian distributed.

A useful fact of the multivariate normal distribution is that any linear combination of the components of a multivariate normal random vector is also normal. In our context this means that a time series $(X_t, t \in \mathbb{T})$ is Gaussian if and only if for any collection $\{t_1, \dots t_n\}$ of times, $n \in \mathbb{N}$, and any vector $a \in \mathbb{R}^n$, the sum $\sum_{i=1}^{n} a_i X_{t_i}$ is normally distributed.

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 $r, s, h \in \mathbb{Z}$.

Definition 2.2.3. A time series $X = (X_t, t \in \mathbb{Z})$ (for which $\mathsf{Var}(X_t) < \infty$ for all $t \in \mathbb{Z}$ is not necessarily true) is said to be *strictly stationary* if the random variables $(X_1, \dots, X_n)$ and $(X_{1+h}, \dots, X_{n+h})$ 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.

Example 2.2.4. Consider the time series $X := (X_t, t \in \mathbb{Z})$ given by

$$X_t := Y_t(Z_t + Z_{t-1})$$

where $Z := (Z_t, t \in \mathbb{Z})$ is $\mathrm{IID}(0, \sigma_Z^2)$ and $Y := (Y_t, t \in \mathbb{Z})$ is a stationary time series independent of Z. We use Definition 2.2.2 to show that X is a (weakly) stationary time series. First we check Condition (i), i.e., we show that μ_X is constant. We have for all $t \in \mathbb{Z}$

$$\mu_X(t) = \mathbb{E}(X_t) = \mathbb{E}(Y_t(Z_t + Z_{t-1})) = \mathbb{E}(Y_t)\,\mathbb{E}(Z_t + Z_{t-1}) = \mathbb{E}(Y_t)(\mathbb{E}(Z_t) + \mathbb{E}(Z_{t-1})) = \mathbb{E}(Y_t) \cdot 0 = 0$$

which is constant and therefore does not depend on t. Here we used the independence of Y and Z in the second equality and after that Z has zero mean. Using the same independence and that X has mean zero, we obtain further that

$$\begin{aligned}\mathsf{Var}(X_t) &= \mathbb{E}(X_t^2) = \mathbb{E}(Y_t^2(Z_t + Z_{t-1})^2) = \mathbb{E}(Y_t^2)\,\mathbb{E}((Z_t + Z_{t-1})^2) \\ &= \mathbb{E}(Y_t^2)(\mathbb{E}(Z_t^2) + 2\,\mathbb{E}(Z_t)\,\mathbb{E}(Z_{t-1}) + \mathbb{E}(Z_{t-1}^2)) = 2(\mathsf{Var}(Y_t) + \mathbb{E}(Y_t)^2)\sigma_Z^2 < +\infty,\end{aligned}$$

since the mean and variance of Y are finite due to its stationarity. It remains to show Condition (ii) of Definition 2.2.2. Using the definition and the independence of Y and Z, we first observe for all $r, s, h \in \mathbb{Z}$ that

$$\begin{aligned}\gamma_X(r+h, s+h) &= \mathbb{E}(X_{r+h}X_{s+h}) = \mathbb{E}(Y_{r+h}(Z_{r+h} + Z_{r+h-1})Y_{s+h}(Z_{s+h} + Z_{s+h-1})) \\ &= \mathbb{E}(Y_{r+h}Y_{s+h})\,\mathbb{E}(Z_{r+h}Z_{s+h} + Z_{r+h}Z_{s+h-1} + Z_{r+h-1}Z_{s+h} + Z_{r+h-1}Z_{s+h-1}).\end{aligned}$$

The stationarity of Y further implies that

$$\mathbb{E}(Y_{r+h}Y_{s+h}) = \gamma_Y(r+h, s+h) + \mu_Y^2 = \gamma_Y(r, s) + \mu_Y^2 = \mathbb{E}(Y_rY_s)$$

and similarly for Z that

$$\begin{aligned}&\mathbb{E}(Z_{r+h}Z_{s+h} + Z_{r+h}Z_{s+h-1} + Z_{r+h-1}Z_{s+h} + Z_{r+h-1}Z_{s+h-1}) \\ &\qquad = \mathbb{E}(Z_rZ_s + Z_rZ_{s-1} + Z_{r-1}Z_s + Z_{r-1}Z_{s-1}).\end{aligned}$$

Doing the same computations back in the other direction, we derive

$$\gamma_X(r+h, s+h) = \mathbb{E}(X_{r+h}X_{s+h}) = \mathbb{E}(X_rX_s) = \gamma_X(r, s),$$

which shows stationarity of X.

Now, we observe that Condition (ii) in Definition 2.2.2 is equivalent to $\gamma_X(r, s)$, with $r, s \in \mathbb{Z}$, being 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 2.2.5. 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 and, since the covariance is symmetric, γ_X is even, i.e., $\gamma_X(h) = \gamma_X(-h)$, $h \in \mathbb{Z}$. 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 2.2.6. 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. For a white noise to be iid noise on the other hand, the random variables must be independent and identically distributed. For example, a white noise that is Gaussian is necessarily iid, since random variables that are uncorrelated and jointly normal are independent.

Example 2.2.7. Let us continue with the time series X of Example 2.2.4 and compute its ACVF. From the computations in the previous example we obtain first that

$$\begin{aligned} \gamma_X(h) = \gamma_X(h,0) &= (\gamma_Y(h,0) + \mu_Y^2))(\gamma_Z(h,0) + \gamma_Z(h,-1) + \gamma_Z(h-1,0) + \gamma_Z(h-1,-1)) \\ &= (\gamma_Y(h) + \mu_Y^2)(\gamma_Z(h) + \gamma_Z(h+1) + \gamma_Z(h-1) + \gamma_Z(h)). \end{aligned}$$

Since $Z \sim \mathrm{IID}(0, \sigma_Z^2)$, this expression simplifies to

$$\gamma_X(h) = \begin{cases} 2(\gamma_Y(h) + \mu_Y^2)\sigma_Z^2 & \text{if } h = 0, \\ (\gamma_Y(h) + \mu_Y^2)\sigma_Z^2 & \text{if } |h| = 1, \\ 0 & \text{else.} \end{cases}$$

2.3 Sample mean and sample autocorrelation function

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 *estimates* (the estimators applied to data) are just numbers.

Definition 2.3.1. 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 $h = 0, \dots, n-1$. Furthermore the *sample autocorrelation function* $\hat{\rho}$ is given by

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

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

We remark that the definitions of the sample autocovariance and autocorrelation function can be extended to $h = -n, \dots, -1$ 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, which could be interpreted as a "stronger law of large numbers", since it is stronger than the weak law of large numbers.

Proposition 2.3.2. *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 covariance is bilinear, we observe that

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

Let us simplify the sum next. It holds that

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

Coming back to our original computation we obtain that

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

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

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

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 at least one, i.e., for all $n \in \mathbb{N}$

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

It can be shown that $\lim_{h\to\infty}\gamma_X(h) = 0$ suffices to show that $\lim_{n\to\infty}\mathsf{Var}(\bar{X}_n) = 0$ (without a rate), but the proof is more involved. We leave it to the interested reader to confirm this as a somewhat challenging *exercise*.

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

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

Even if the time series is not Gaussian, this is usually a good approximation.

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 [3], which says that one should take $n \geq 50$ and $h \leq n/4$. Moreover, it is possible to prove asymptotic results such as Proposition 2.3.2 also for $\gamma_X(h)$, but this requires stronger assumptions on the time series. The interested reader is referred to [4, Chapter 7].

In applications such as forecasting, the sample autocovariance function is often used in a matrix. 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 [6, 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 $\hat{\gamma}(0) > 0$, i.e., if there is a $t \in \mathbb{Z}$ s.t. $X_t \neq \bar{X}$. One can show that if the factor n^{-1} were replaced by $(n-h)^{-1}$ in the definition of $\hat{\gamma}$, the matrices could be singular even if $\hat{\gamma}(0) > 0$.

Observe that these functions can be defined for all observed time series, even if the time series is not stationary. In fact, the sample autocorrelation function can be used to detect deviations from stationarity, as we shall see.

Example 2.3.3. Let us revisit the data from Example 2.1.2, i.e., the set $(x_t, t = 1, \ldots, 42)$ 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 2.3. 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 2.5. We also note that the bounds $\pm 1.96/\sqrt{n}$ have been included in the figure, something that is explained below.

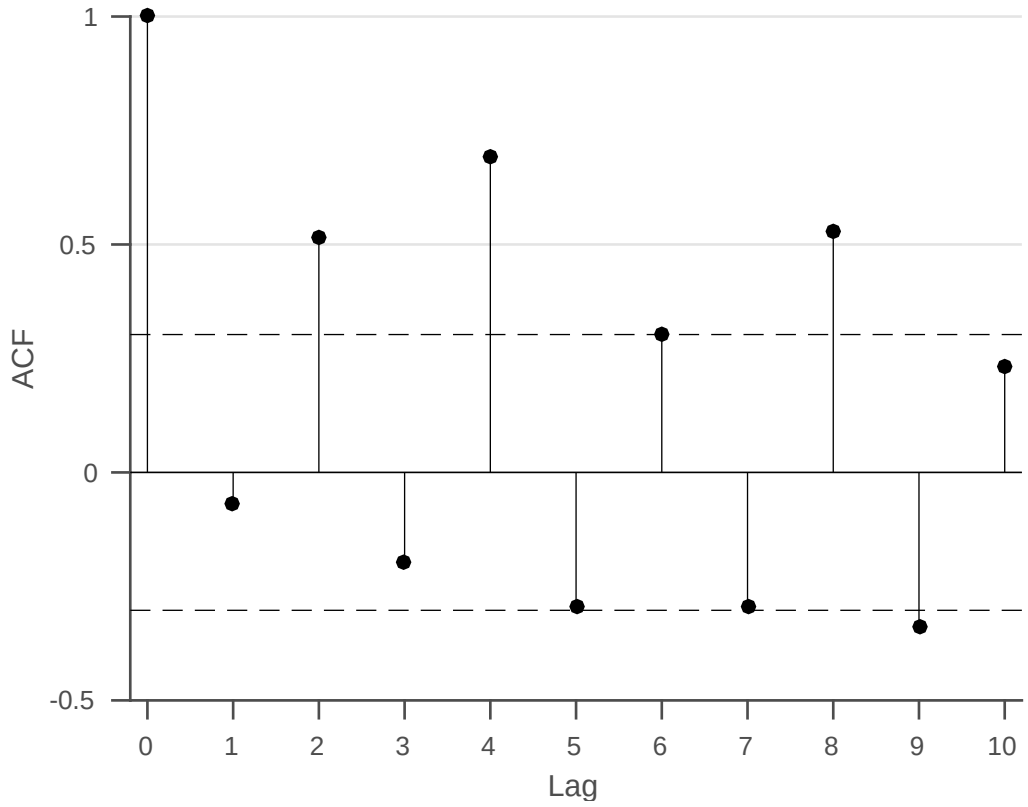


Figure 2.3: Sample autocorrelation function for the H&M data of Example 2.1.2.

The first application of the estimators introduced above is in the context of hypothesis testing. The first thing that we do when giving a data set is to check if there is any temporal structure to it, i.e., if we believe that it is something other than iid noise. If now $Y = (Y_1, \dots, Y_n)$ is a sequence of iid random variables with finite variance, then the sample autocorrelations $\hat{\rho}(h)$, $h = 1, 2, 3, \dots$ for sufficiently large n approximately iid and $\mathcal{N}(0, n^{-1})$ distributed. So if we are given a data set $(y_1, \dots, y_n)$, an informal hypothesis test is to check if 95% of the computed values of $\hat{\rho}$ should fall between the bounds $\pm 1.96/\sqrt{n}$ if the data set is a realization of Y. Otherwise we can reject the null hypothesis of Y being iid noise.

The same reasoning is behind the following hypothesis tests that we introduce to check if there is any temporal structure to the data we have at hand. More details on the concept of hypothesis testing can be found in Section 1.2 if the reader is not familiar with ideas.

If $(Y_1, \dots, Y_n)$ is a sequence of iid random variables with finite variance, then one can show that the test statistic

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

is approximately χ^2_h distributed, i.e., chi-squared distributed with h degrees of freedom. This forms the basis of the *Portmanteau test*.

Method 2.3.4 (Portmanteau test, Box–Pierce test)**.** We have

$$H_0 : Y \sim \mathrm{IID}(\mu, \sigma^2),$$
$$H_1 : Y \nsim \mathrm{IID}(\mu, \sigma^2).$$

and the test statistic

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

The null hypothesis is rejected at the α-level, with $\alpha \in (0,1)$, if $\lambda > \chi^2_{1-\alpha,h}$, where $\chi^2_{1-\alpha,h}$ denotes the $1-\alpha$-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 2.3.5 (Ljung–Box test)**.** This test is a modification of the Portmanteau test. Use instead the test statistic

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

which is asymptotically χ^2_h-distributed for iid random variables. Use the same rejection regions as in the Portmanteau test 2.3.4.

Example 2.3.6. Let us revisit the data from Example 2.1.2, i.e., the set $(x_t, t = 1, \ldots, 42)$ of quarterly earnings of H&M. We apply the Ljung–Box test (Method 2.3.5) with $h = 4$ and $\alpha = 0.05$ to this data set to see whether it is likely that the earnings are a realization of iid noise. 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 x being a realization of $X \sim \mathrm{IID}(\mu, \sigma^2)$ at the 5% level, something that is not surprising given Figure 2.1.

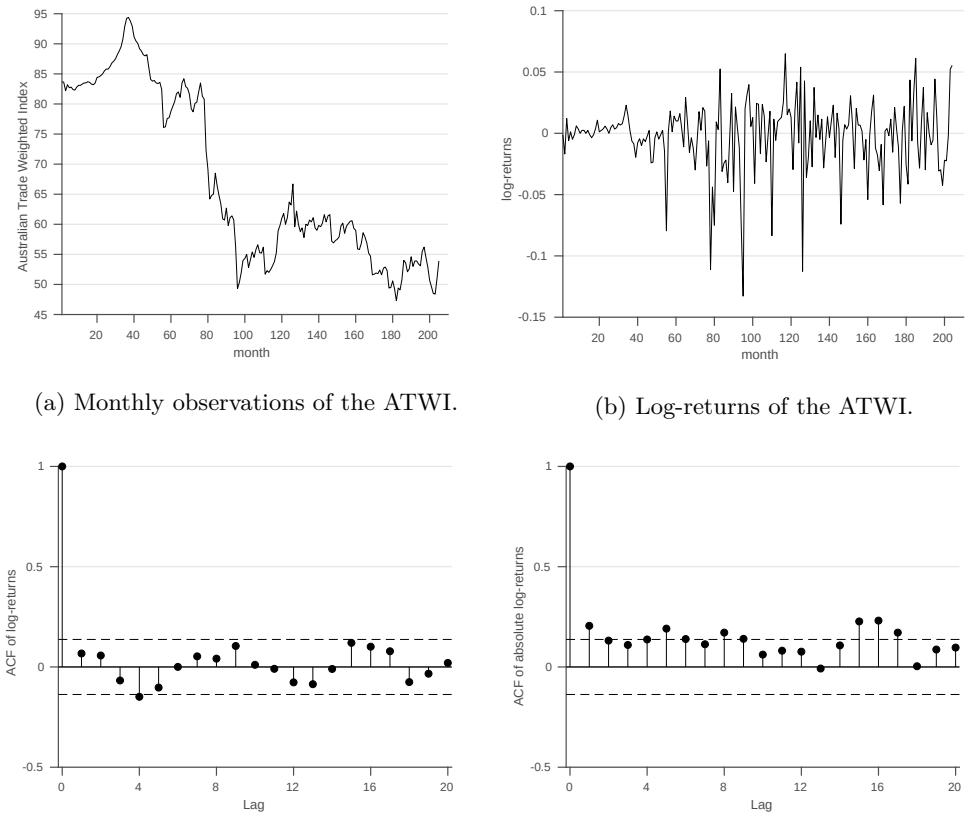
Some care has to be taken when applying these tests to financial time series. In many contexts, the tests have *low power*, i.e., there is a large probability that they fail to reject the null hypothesis when the alternative hypothesis is true. This is illustrated in the next example.

Example 2.3.7. In Figure 2.4a we plot $x = (x_t, t = 1, \ldots, 205)$, monthly observations of the Australian Trade Weighted Index (ATWI), a weighted sum of exchange rates between the Australian dollar and other currencies from January 1978 to January 1995. In Figure 2.4b we show the corresponding log-returns $y = (y_t, t = 1, \ldots, 204)$ with $y_t = \log(y_{t+1}) - \log(y_t)$ and in Figure 2.4c the sample autocorrelation function of these. We again apply the Ljung–Box test to y, now with $h = 20$ and $\alpha = 0.05$. The test statistic (2.1) is found to be 24.37 which is smaller than the critical value $\chi^2_{0.95,20} = 31.41$. From this result we have no reason to reject the null hypothesis that y is the realization of $Y \sim \mathrm{IID}(\mu, \sigma^2)$ for some μ, σ^2. However, if $(Y_t, t = 1, \ldots, 204)$ is a sequence of iid random variables, then so are $(|Y_t|)_{t=1}^{204}$. Therefore, if the power of the Ljung–Box test were big, we would expect H_0 to not be rejected if we applied it to the time series $(|y_t|, t = 1, \ldots, 204)$. In Figure 2.4d we plot the sample ACF of this series. If we now calculate (2.1) again, we find it to be 84.25, which means that we decidedly reject the claim that the data is iid at the 5% level.

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 [6, Section 1.6]. As the example above illustrates, it is usually a good idea to apply a couple of transformations to the data when testing for iid, so that we do not fail to reject the null hypothesis by mistake.

2.4 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 $(X_t, t = 1, \ldots, n)$. We will find best predictors in the sense of minimal mean squared errors. To that end, let us start with the necessary definitions.



(a) Monthly observations of the ATWI.

(b) Log-returns of the ATWI.

(c) Sample ACF of the log-returns.

(d) Sample ACF of absolute log-returns.

Figure 2.4: Observations and sample ACF of the ATWI data.

Definition 2.4.1. 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*. Here we only consider forecasting as minimization of the mean squared error, since its theoretical properties leads to easy-implementable algorithms.

Definition 2.4.2. Let $(X_t, t \in \mathbb{Z})$ be a time series with $\mathsf{Var}(X_t) < \infty$ for all $t \in \mathbb{Z}$ 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 $t \in \mathbb{Z}$, if it minimizes the mean squared error, i.e.,

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

where the minimum is taken over all measurable functions $g : \mathbb{R}^n \to \mathbb{R}$.

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

Proposition 2.4.3. *Let $(X_t, t \in \mathbb{Z})$ be a time series with $\mathsf{Var}(X_t) < \infty$ for all $t \in \mathbb{Z}$ 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 predictor of X_t for some $t \in \mathbb{Z}$ 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 (vi) 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 2.4.4. Let $(X_t, t \in \mathbb{Z})$ be a time series with $\mathsf{Var}(X_t) < \infty$ for all $t \in \mathbb{Z}$ 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 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 $t \in \mathbb{Z}$ if it minimizes the mean squared error, i.e.,

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

where the minimum is taken over all linear functions g of 1 and X^n, i.e., for all functions g such that there exist $a_0, \dots, a_n \in \mathbb{R}$ such that $g(X^n) := a_0 + a_1 X_{t_n} + a_2 X_{t_{n-1}} + \cdots + a_n X_{t_1}$.

Note that when $X^n := (X_1, \dots, X_n)$, the authors of [6] usually denote $b_{n+h}^l(X^n)$ by $P_n X_{n+h}$ and $b_{n+1}^l(X^n)$ by $\hat{X}_{n+1}$.

Let us now derive the coefficients $(a_i, i = 0, \dots, n)$ explicitly, 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 $a = (a_0, \dots, a_n)$, 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

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

as well as, for $j = 1, \dots, n$,

$$\frac{\partial S(a)}{\partial a_j} = 2\,\mathbb{E}(X_{t_{n+1-j}}(a_0 + a_1 X_{t_n} + \cdots + a_n X_{t_1} - X_t)). \tag{2.3}$$

By setting these equations equal to zero, we derive from (2.2) that

$$\mathbb{E}(X_t - (a_0 + a_1 X_{t_n} + \cdots + a_n X_{t_1})) = \mathbb{E}(X_t - b_t^l(X^n)) = 0 \tag{2.4}$$

and from (2.3)

$$\mathbb{E}\left((X_t - (a_0 + a_1 X_{t_n} + \cdots + a_n X_{t_1}))\, X_{t_{n+1-j}}\right) = \mathbb{E}\left(\left(X_t - b_t^l(X^n)\right) X_{t_{n+1-j}}\right) = 0 \tag{2.5}$$

for $j = 1, \dots, n$.

In fact, these two equations determine the predictor uniquely. To see this, let $(a_j^{(1)}, j = 0, \dots, n)$ and $(a_j^{(2)}, j = 0, \dots, n)$ 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 (2.4) and

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

for all $j = 1, \dots, n$ by (2.5), 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 2.4.5. *Let $(X_t, t \in \mathbb{Z})$ be a time series with $\mathsf{Var}(X_t) < \infty$ for all $t \in \mathbb{Z}$ 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 of X_t is given by*

$$b_t^l(X^n) = a_0 + a_1 X_{t_n} + a_2 X_{t_{n-1}} + \cdots + a_n X_{t_1},$$

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

(i) $\mathbb{E}(X_t - b_t^l(X^n)) = 0$,

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

An important property of the best linear predictor of a *Gaussian* time series is that it coincides with the conditional expectation (see, e.g., [4, Exercise 2.20]). This means due to Proposition 2.4.3 that under this assumption we obtain $b_t^l(X^n) = b_t(X^n) = \mathbb{E}(X_t|X^n)$.

If $(X_t, t \in \mathbb{Z})$ is stationary with mean μ and autocovariance function γ, then it is an easy *exercise* to check that (2.4) simplifies to

$$a_0 = \mu \left(1 - \sum_{i=1}^{n} a_i\right), \tag{2.6}$$

which implies, together with (2.5), that, for $j = 1, 2, \dots, n$,

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

The following corollary expresses these facts in matrix notation.

Corollary 2.4.6. *Let $X = (X_t, t \in \mathbb{Z})$ and X^n be as in Proposition 2.4.5 and assume in addition that X is stationary with mean μ and autocovariance function γ. Then the coefficients $(a_i, i = 0, \dots, n)$ of $b_t^l(X^n)$ are determined by the linear equations*

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

and

$$\Gamma_n(a_1, \dots, a_n)' = (\gamma(t - t_n), \dots, \gamma(t - t_1))' \tag{2.7}$$

with

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

Moreover,

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

Proof. The first two claims follows directly by the previous proposition. To compute the mean squared error 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 the first two claims 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 that

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

Therefore we conclude that

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

□

Remark 2.4.7. Note that (2.7) holds with the autocovariance function γ replaced by the autocorrelation function ρ, since this corresponds to dividing both sides of the equation with $\gamma(0)$.

Remark 2.4.8. In the case that $X^n := (X_1,\dots,X_n)$ the equations to derive the coefficients $(a_0,\dots,a_n)$ for the prediction of X_{n+h} (i.e., the coefficients in $b_{n+h}^l(X^n)$) simplify to

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

and

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

Observe that X^n just determines the length of the used history but not necessarily the absolute time. In other words, under stationarity, you will obtain the same coefficients $(a_0,a_1,\dots,a_n)$ in your best linear predictor $b_{t+n+h}^l((X_{t+1},\dots,X_{t+n}))$ for all $t\in\mathbb{Z}$.

Note that for a stationary time series, the coefficients $(a_1,\dots,a_n)$ in Corollary 2.4.6 only depend on the ACVF γ while a_0 only depends on $(a_1,\dots,a_n)$ and μ. This means that only a_0 will change if we consider different μ. Therefore, considering only zero-mean stationary time series is not a restriction in the context of forecasting.

To see an application of the theory, let us now treat two zero-mean stationary examples.

Example 2.4.9 (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. Assuming that Z_t is uncorrelated with X_{t-j} for $j>0$, something that follows from the inequality $|\phi_1|<1$ as we will see in Chapter 3, 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 $\mathbb{E}(X_t)=0$ and $\mathbb{E}(Z_tX_{t-1}) = \mathbb{E}(Z_t)\,\mathbb{E}(X_{t-1}) = 0$ by assumption. This implies that

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

and

$$\gamma_X(h) = \mathbb{E}(X_tX_{t+h}) = \mathbb{E}(X_t(Z_{t+h}+\phi_1X_{t+h-1})) = \phi_1\gamma_X(h-1)$$

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 $b_{n+1}^l(X^n)$ 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 Corollary 2.4.6 $\sum_{i=1}^n a_i X_{n+1-i}$, 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

$$b_{n+1}^l(X^n) = \phi_1 X_n$$

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

Example 2.4.10 (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 $\mathrm{WN}(0, \sigma^2)$ 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

$$\begin{aligned} \gamma_X(h) &= \mathsf{Cov}(X_t, X_{t+h}) = \mathbb{E}(X_t X_{t+h}) \\ &= \mathbb{E}(Z_t Z_{t+h}) + \theta_1 \mathbb{E}(Z_t Z_{t+h-1}) + \theta_1 \mathbb{E}(Z_{t-1} Z_{t+h}) + \theta_1^2 \mathbb{E}(Z_{t-1} Z_{t+h-1}) \end{aligned}$$

so that, since $Z \sim \mathrm{WN}(0, \sigma^2)$,

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

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

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

The best linear predictor $b_{n+1}^l(X^n)$ with respect to $X^n := (X_1, \ldots, X_n)$ is by Proposition 2.4.5 $\sum_{i=1}^n a_i X_{n+1-i}$. Note that $a_0 = 0$ due to the fact that the series has zero mean. We find the coefficients a_i with Proposition 2.4.5. 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. [6, 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 Corollary 2.4.6 to compute the mean squared error

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

which finishes this example.

Example 2.4.11 (AR(1) with missing value)**.** Let us consider the AR(1) model from Example 2.4.9 again. Assume that we have observed X_1 and X_3 but that we are missing X_2. Then the best linear predictor $b_2^l((X_1, X_3))$ of X_2 is $a_1X_3 + a_2X_1$ by Corollary 2.3.6, 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

$$b_2^l((X_1, X_3)) = \frac{\phi_1}{1 + \phi_1^2}(X_3 + X_1)$$

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

We have stated in Proposition 2.4.5 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., $b_{n+1}^l(X^n)$ is used to compute $b_{n+2}^l(X^{n+1})$ 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

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

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 2.4.5.

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 [6, Section 2.5.3]. 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 [4, Proposition 5.1.1].

Method 2.4.12 (Durbin–Levinson algorithm)**.** Compute the coefficients $a_{n1}, \dots, a_{nn}$ 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

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

Lemma 2.4.13. *There exist unique coefficients* $(\theta_{ij},\ 1 \le j \le i \le n)$ *such that the best linear predictors satisfy*

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

Proof. Observe first that we can change the order of summation to

$$\hat{X}_{n+1} = \sum_{j=1}^{n} \theta_{n(n+1-j)}(X_j - \hat{X}_j),$$

which is the formula that we are going to show. Furthermore, the best linear predictors satisfy with the notation of the Durbin–Levinson algorithm that

$$\hat{X}_{n+1} = \sum_{j=1}^{n} a_{n(n+1-j)} X_j.$$

Using these representations, we prove the claim by induction.

We first obtain for $n = 1$ that

$$\hat{X}_2 = a_{11} X_1 = a_{11}(X_1 - \hat{X}_1),$$

since $\hat{X}_1 = 0$ by definition. Setting $\theta_{11} := a_{11}$, we establish the base case.

Let us next assume that the claim holds for $j = 1, \dots, n-1$ and prove the claim for n. We observe with the definition of the best linear predictor that

$$\hat{X}_{n+1} = \sum_{j=1}^{n} a_{n(n+1-j)}(X_j - \hat{X}_j) + \sum_{j=1}^{n} a_{n(n+1-j)}\hat{X}_j$$

and derive for the second term that

$$\sum_{j=1}^{n} a_{n(n+1-j)}\hat{X}_j = \sum_{j=1}^{n} a_{n(n+1-j)} \sum_{i=1}^{j-1} \theta_{(j-1)(j-i)}(X_i - \hat{X}_i)$$

by the induction hypothesis. If we change the order of summation and change the roles of i and j in a second step, the last equation can be rewritten to

$$\begin{aligned}\sum_{j=1}^{n} a_{n(n+1-j)}\hat{X}_j &= \sum_{i=1}^{n-1}\Big(\sum_{j=i+1}^{n} a_{n(n+1-j)}\theta_{(j-1)(j-i)}\Big)(X_i - \hat{X}_i) \\ &= \sum_{j=1}^{n-1}\Big(\sum_{i=j+1}^{n} a_{n(n+1-i)}\theta_{(i-1)(i-j)}\Big)(X_j - \hat{X}_j).\end{aligned}$$

Plugging this result in the initial computation, we conclude that

$$\hat{X}_{n+1} = \sum_{j=1}^{n}\Big(a_{n(n+1-j)} + \sum_{i=j+1}^{n} a_{n(n+1-i)}\theta_{(i-1)(i-j)}\Big)(X_j - \hat{X}_j),$$

and therefore we set

$$\theta_{n(n+1-j)} := a_{n(n+1-j)} + \sum_{i=j+1}^{n} a_{n(n+1-i)}\theta_{(i-1)(i-j)}.$$

Uniqueness is a consequence of [6, Problem 2.20], which we leave for the reader to confirm. □

The innovations algorithm generates these coefficients and the mean squared errors $v_j = \mathrm{MSE}(\hat{X}_{j+1}, X_{j+1})$ recursively.

Method 2.4.14 (Innovations algorithm)**.** Compute the coefficients $\theta_{n1}, \dots, \theta_{nn}$ 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 $0 \le k < n$ and

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

Proof. The statement about v_0 is true by definition. Note next that by [6, Problem 2.20], we have for $i, j \in \{1, \dots, n+1\}$,

$$\mathbb{E}((X_i - \hat{X}_i)(X_j - \hat{X}_j)) = 0 \tag{2.8}$$

as long as $i \neq j$ (this is a fairly straightforward consequence of Proposition 2.4.5). Therefore, we obtain that for $0 \leq k < n$,

$$\begin{aligned}\mathbb{E}(\hat{X}_{n+1}(X_{k+1} - \hat{X}_{k+1})) &= \mathbb{E}\left(\sum_{j=1}^{n} \theta_{nj}(X_{n+1-j} - \hat{X}_{n+1-j})(X_{k+1} - \hat{X}_{k+1})\right) \\ &= \sum_{j=1}^{n} \theta_{nj}\,\mathbb{E}((X_{n+1-j} - \hat{X}_{n+1-j})(X_{k+1} - \hat{X}_{k+1})) = \theta_{n(n-k)}v_k\end{aligned} \tag{2.9}$$

where we expanded $\hat{X}_{n+1}$ using Lemma 2.4.13. By (2.8) we also have

$$\begin{aligned}\mathbb{E}(X_{n+1}(X_{k+1} - \hat{X}_{k+1})) &= \mathbb{E}((X_{n+1} - \hat{X}_{n+1})(X_{k+1} - \hat{X}_{k+1})) + \mathbb{E}(\hat{X}_{n+1}(X_{k+1} - \hat{X}_{k+1})) \\ &= \mathbb{E}(\hat{X}_{n+1}(X_{k+1} - \hat{X}_{k+1})) = \theta_{n(n-k)}v_k,\end{aligned} \tag{2.10}$$

using (2.9) in the last step. Therefore, by Lemma 2.4.13 again,

$$\begin{aligned}\theta_{n(n-k)} &= v_k^{-1}\,\mathbb{E}(X_{n+1}(X_{k+1} - \hat{X}_{k+1})) \\ &= v_k^{-1}\left(\mathbb{E}(X_{n+1}X_{k+1}) - \sum_{j=1}^{k} \theta_{kj}\,\mathbb{E}(X_{n+1}(X_{k+1-j} - \hat{X}_{k+1-j}))\right) \\ &= v_k^{-1}\left(\kappa(n+1, k+1) - \sum_{j=1}^{k} \theta_{kj}\theta_{n(n-(k-j))}v_{k-j}\right) \\ &= v_k^{-1}\left(\kappa(n+1, k+1) - \sum_{j=0}^{k-1} \theta_{k(k-j)}\theta_{n(n-j)}v_j\right)\end{aligned}$$

so we have derived the expression for $\theta_{n(n-k)}$. Note that we used (2.10) twice, once in the first equality and once in the third. To find the expression for v_n, we note that by (2.8) and Lemma 2.4.13,

$$\begin{aligned}\mathbb{E}(X_{n+1}\hat{X}_{n+1}) &= \mathbb{E}(\hat{X}_{n+1}^2) + \mathbb{E}((X_{n+1} - \hat{X}_{n+1})\hat{X}_{n+1}) \\ &= \mathbb{E}(\hat{X}_{n+1}^2) + \sum_{j=1}^{n} \theta_{nj}\,\mathbb{E}((X_{n+1} - \hat{X}_{n+1})(X_{n+1-j} - \hat{X}_{n+1-j})) = \mathbb{E}(\hat{X}_{n+1}^2)\end{aligned}$$

so that

$$\begin{aligned}v_n &= \mathbb{E}((X_{n+1} - \hat{X}_{n+1})^2) = \mathbb{E}(X_{n+1}^2) - 2\,\mathbb{E}(X_{n+1}\hat{X}_{n+1}) + \mathbb{E}(\hat{X}_{n+1}^2) = \mathbb{E}(X_{n+1}^2) - \mathbb{E}(\hat{X}_{n+1}^2) \\ &= \kappa(n+1, n+1) - \sum_{j=1}^{n} \theta_{nj}^2 v_{n-j} = \kappa(n+1, n+1) - \sum_{j=0}^{n-1} \theta_{n,n-j}^2 v_j,\end{aligned}$$

where we again used (2.8) in the fourth equality. □

Note that we solve for θ and v in the order $v_0, \theta_{11}, v_1, \theta_{22}, \theta_{21}, v_2, \theta_{33}, \theta_{32}, \theta_{31}, v_3, \ldots$.

2.5 Trend and seasonality

So far we have only discussed stationary time series. This allows us to estimate statistical properties and make forecasts but typically data requires transformations before it can be considered

stationary. In this chapter we will assume that the data is non-stationary in a particular way, namely that it is a realization of the stochastic process X which can be split into

$$X_t = m_t + s_t + Y_t. \tag{2.11}$$

This is called 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 with mean zero.

The trend component is the most important component for financial purposes – when we model returns on an investment as a time series we hope that we make money on average, and that therefore the time series has a positive trend. In terms of modeling stock returns the seasonal component is less relevant – seasonality in stock returns allow for arbitrage – but it is important in the modeling of business data such as retail sales figures.

Determining if a time series X contains a trend component and or a seasonal component can be a difficult task. Visual representations of the data in the form of plots of the raw data or the ACF can be useful. For example, a plot of the ATWI time series of Example 2.3.7 (Figure 2.4a) suggests a clear negative trend for this data. A time series with a trend component has a sample ACF that is decaying very slowly, see Figure 2.5a for the sample ACF of the ATWI data. Similarly, a seasonal component can be seen by the sample ACF displaying periodicity, see Example 2.1.2. A priori knowledge can also be used to decide on whether to include trend and seasonality or not. Another possibility is to let models with and without such components compete by forecasting known data and using the result to determine whether to include them or not. Finally spectral methods (see [6, Chapter 4]) can be used to detect seasonality, but this is outside the scope of this course. We will assume 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) or by inspection of the raw data plot and the sample ACF as described above. Typical periods are 24 hours per day, 7 days per week, 12 months per year, or 4 quarters per year.

Once it has been decided that the time series has a trend and/or a seasonal component, there are two main approaches to obtaining the stationary process Y from (2.11): *estimation* and *differencing*. We will cover the former approach first. In all methods below, 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.

2.5.1 Estimation of trend and seasonal components

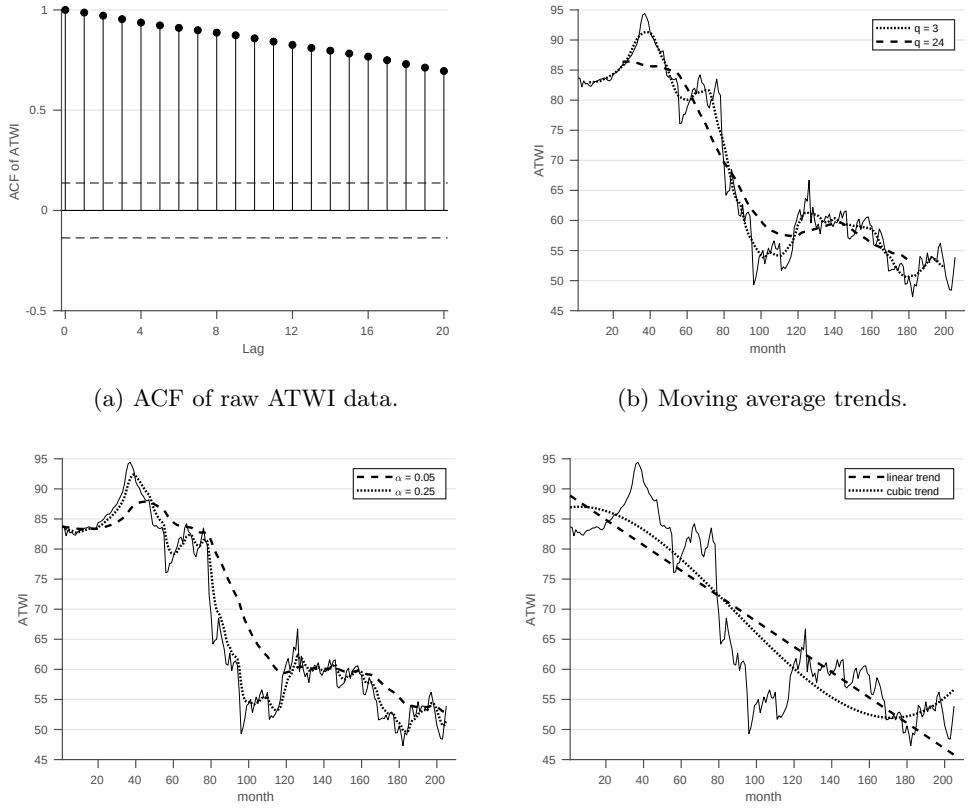
In what follows we first introduce two methods for trend estimation before giving two methods that estimate both trend and seasonality. In the first two, we assume that the seasonal component is zero, i.e., that $s_t = 0$ for all $t \in \mathbb{Z}$.

Method 2.5.1 (Estimation of trend by a 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

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



(a) ACF of raw ATWI data.

(b) Moving average trends.

(c) Exponential smoothing trends.

(d) Linear least squares trends.

Figure 2.5: Sample ACF and estimated trends of the ATWI data.

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. In Figure 2.5b we apply this method with $q = 3$ and $q = 24$ to the ATWI data.

Method 2.5.2 (Estimation of trend 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. In Figure 2.5c we apply this method with $\alpha = 0.05$ and $\alpha = 0.25$ to the ATWI data.

In the next two methods, we estimate trend and seasonality in parallel.

Method 2.5.3 (Estimation of trend and seasonality by linear least squares)**.** Assume that the trend m is given by the polynomial

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

for $t \in \mathbb{Z}$ and some $q \in \mathbb{N}$. Let the seasonal component s with known period d be given by the sinusoid

$$s_t := \sum_{k=0}^{p} b_k \cos(2\pi\lambda_k t/d) + c_k \sin(2\pi\lambda_k t/d)$$

for $t \in \mathbb{Z}$, $p \in \mathbb{N}$ and some *known* coefficients $(\lambda_j, j = 1, \ldots, p) \subset \mathbb{N}$. The unknown coefficients $(a_j, j = 1, \ldots, q)$, $(b_k, k = 1, \ldots, p)$ and $(k_j, k = 1, \ldots, p)$ are obtained by the *least square minimization*

$$\arg\min_{a_j, b_k, c_k} \sum_{t=1}^{n} (x_t - m_t - s_t)^2,$$

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

$$C := \begin{pmatrix} 1 & \cdots & 1^q & \cos(2\pi\lambda_1/d) & \cdots & \cos(2\pi\lambda_p/d) & \sin(2\pi\lambda_1/d) & \cdots & \sin(2\pi\lambda_p/d) \\ & & & & & \vdots & & & \\ 1 & \cdots & t^q & \cos(2\pi\lambda_1 t/d) & \cdots & \cos(2\pi\lambda_p t/d) & \sin(2\pi\lambda_1 t/d) & \cdots & \sin(2\pi\lambda_p t/d) \\ & & & & & \vdots & & & \\ 1 & \cdots & n^q & \cos(2\pi\lambda_1 n/d) & \cdots & \cos(2\pi\lambda_p n/d) & \sin(2\pi\lambda_1 n/d) & \cdots & \sin(2\pi\lambda_p n/d) \end{pmatrix}$$

has the property that $C'C$ is non-singular, then $\sum_{t=1}^{n}(x_t - m_t - s_t)^2$ has a unique minima given by $(\hat{a}_0, \hat{a}_1, \ldots, \hat{a}_q, \hat{b}_1, \ldots, \hat{b}_p, \hat{c}_1, \ldots, \hat{c}_p)' = (C'C)^{-1}C'X$, where $X = (x_1, x_2, \ldots, x_n)'$. In Figure 2.5d we apply this method to the ATWI data and fit a linear ($q = 1$) and a cubic trend ($q = 3$), assuming no seasonal component.

Method 2.5.4 (Estimation of trend and seasonality by moving averages)**.** Let us assume that (for simplicity) the size of the observed data n covers a multiple of the period d, i.e., $n/d \in \mathbb{N}$. 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. Note that we sum between $q+1$ and $n-q$ since we have estimated $m(t)$ at precisely these indices. 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 e.g., Method 2.5.3 to the deseasonalized series $(x_t - \hat{s}_t, t = 1, \ldots, n)$ to obtain a trend estimator in parametric form which is convenient for prediction and simulation.

Example 2.5.5. We apply Method 2.5.4 to the data of Example 2.1.2. As we noted in Example 2.3.3, 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==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 2.6.

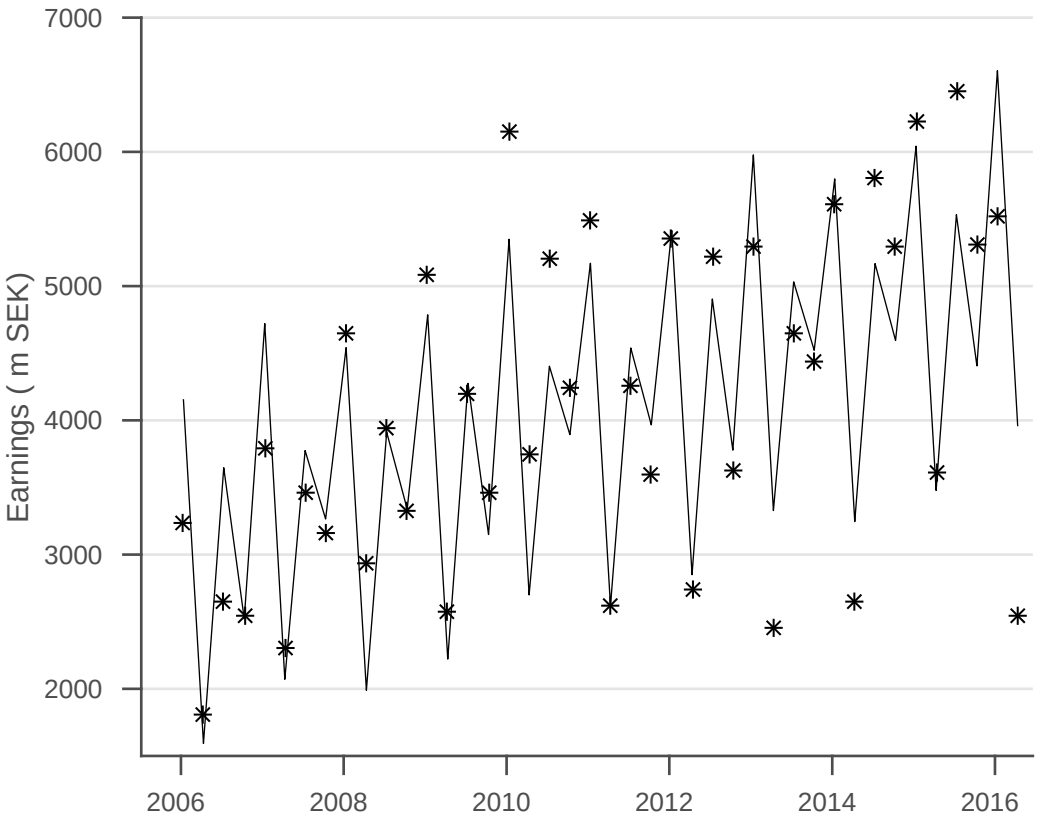


Figure 2.6: The H&M data of Example 2.1.2 (stars) with an estimated linear trend and seasonal component with period 4, $\hat{m}_t + \hat{s}_t$ (line).

2.5.2 Elimination of trend and seasonal components by differencing

The former section deals with explicit estimation of s and m. This is useful for descriptive purposes, to answer the question of what any trend or seasonal component of the time series under analysis looks like. If one is only interested in modeling (and forecasting) it may suffice to *eliminate them.* For this, the method of *differencing* is easy to implement and common in practice.

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.,

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

We apply powers of ∇ to X in order to eliminate trend components. Suppose that the trend m in (2.11) 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.) If $s = 0$ in (2.11), we therefore have 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 if $s = 0$, $\nabla^q X$ is a mean $q!\, a_q$, stationary process.

In order to deal with time series with a seasonal component s with period d, we introduce the *lag-d differencing operator* ∇_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 *exercise* that $\nabla_d Y$ is a stationary time series, which implies that $\nabla_d X$ is a stochastic process without seasonal component. The trend component $\nabla_d m$ is then eliminated with multiple applications of ∇ like above.

We end this section with a note on how to forecast a differenced time series. Suppose that we have eliminated m and s from $X = (X_t, t \in \mathbb{Z})$ by applying ∇^N and ∇_d^M for some powers N and M. Then, by what we have noted above and by expanding ∇^N and ∇_d^M into powers of B, we see that there are coefficients $b_1, \ldots, b_{N+Md}$ and a mean zero stationary process $\tilde{Y} = (\tilde{Y}_t, t \in \mathbb{Z})$ such that

$$\tilde{Y}_t = \nabla^N \nabla_d^M X_t = (1 - B)^N (1 - B^d)^M X_t = \sum_{k=0}^{N+Md} b_k B^k X_t = \sum_{k=0}^{N+Md} b_k X_{t-k}$$

for all $t \in \mathbb{Z}$, where $b_0 := 1$. Therefore, it holds in particular that

$$X_{n+h} = \tilde{Y}_{n+h} - \sum_{k=1}^{N+Md} b_k X_{n+h-k}.$$

for some $h, n \in N$. Suppose now (by relabeling the indices of your data set if need be) that X is observed at times $-N - Md + 1, -N - Md, \ldots, n$, $n \in \mathbb{N}$ so that $\tilde{Y}$ is observed at times $1, \ldots, n$ and let us write $X^{n+N+Md} := (X_{-N-Md+1}, \ldots, X_n)$ and $\tilde{Y}^n := (\tilde{Y}_1, \ldots, \tilde{Y}_n)$ for the sets of observed

values of the time series. If we assume that the initial observations $X_{-N-Md+1}, \ldots, X_0$ are all uncorrelated with $\tilde{Y}^n$, then it can be shown that

$$b^\ell_{n+h}(X^{n+N+Md}) = b^\ell_{n+h}(\tilde{Y}^n) - \sum_{k=1}^{N+Md} b_k b^\ell_{n+h-k}(X^{n+N+Md}). \tag{2.12}$$

Since $\tilde{Y}$ is stationary, $b^\ell_{n+h}(\tilde{Y}^n)$ can be computed using the methods of Section 2.4. This means that (2.12) defines a recursive formula that allows us to compute forecasts of X^{n+h} by first computing $b^\ell_{n+1}(X^{n+N+Md})$, then $b^\ell_{n+2}(X^{n+N+Md})$ and so on. Note that $b^\ell_{n+h-k}(X^{n+N+Md}) = X_{n+h-k}$ if $h \leq k$.

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

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 $Y = Z \sim \mathrm{WN}(0, \sigma^2)$, i.e., 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. We first observe that $\mathbb{E}(|Y_t|) \leq \mathbb{E}(|Y_t|^2)^{1/2} = \gamma_Y(0)^{1/2}$ by the Cauchy–Schwarz inequality and therefore we may (by Fubini's theorem) interchange sum and expectation

$$\mathbb{E}(|X_t|) \leq \sum_{j\in\mathbb{Z}} |\psi_j| \, \mathbb{E}(|Y_{t-j}|) \leq \left(\sum_{j\in\mathbb{Z}} |\psi_j| \right) \gamma_Y(0)^{1/2} < +\infty.$$

This implies (by, e.g., Chebyshev's inequality) that the series converges with probability one and that X 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. For details we refer to [6, Example C.1.1]. Therefore X is stationary.

If Y is mean zero white noise with variance σ^2, then the previous expression simplifies 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 $\gamma_Y(h+j-k) \neq 0$ if and only 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 zeros. 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

$$\phi(B) X_t = \theta(B) Z_t, \tag{3.2}$$

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*

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

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

Proof. We only prove that the condition on ϕ implies that a stationary solution exists. For uniqueness, and for the fact that no solution exists if $\phi(z) = 0$ for some $z \in \mathbb{C}$ with $|z| = 1$, the reader is referred to [4, Chapter 3] and [4, Problem 4.28]. If $\phi(z) \neq 0$ for all $z \in \mathbb{C}$ on the unit

circle, then, by a fact from complex analysis (see, e.g., [1, Chapter 5]), 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. Let us define the operator $\phi(B)^{-1}$ by

$$\phi(B)^{-1} Y_t := \sum_{j=-\infty}^{\infty} \chi_j B^j Y_t = \sum_{j=-\infty}^{\infty} \chi_j Y_{t-j}$$

for stationary time series $Y = (Y_t, t \in \mathbb{Z})$. Note that the sum on the right hand side exists and is stationary as a consequence of Proposition 3.1.2. We define a time series $X = (X_t, t \in \mathbb{Z})$ by setting, with $\theta_0 := 1$,

$$X_t := \phi(B)^{-1}\theta(B) Z_t = \phi(B)^{-1} \sum_{i=0}^{q} \theta_i Z_{t-i} = \sum_{j=-\infty}^{\infty} \sum_{i=0}^{q} \chi_j \theta_i Z_{t-i-j}. \tag{3.3}$$

By Proposition 3.1.2, this is a well-defined stationary time series, so it only remains to check that it fulfills (3.2).

To this end, we note that, with $\phi_0 := -1$, we have

$$1 = \phi(z)\phi(z)^{-1} = -\left(\sum_{k=0}^{p} \phi_k z^k\right)\left(\sum_{j=-\infty}^{\infty} \chi_j z^j\right) = -\sum_{j=-\infty}^{\infty} \sum_{k=0}^{p} \chi_j \phi_k z^{j+k}.$$

If we make the index change $m = j + k$ we therefore get

$$-1 = \sum_{m=-\infty}^{\infty} \left(\sum_{k=0}^{p} \chi_{m-k}\phi_i\right) z^m.$$

By equating coefficients of z^m, $m \in \mathbb{Z}$, we find that the coefficient in front of z^0 is 1 and the rest 0, i.e.,

$$\sum_{k=0}^{p} \chi_{-k}\phi_k = -1 \text{ and } \sum_{k=0}^{p} \chi_{m-k}\phi_k = 0 \text{ for all } m \neq 0.$$

If we apply this fact to $\phi(B)X_t$, after making the same index change $m = j + k$, we have

$$\begin{aligned}\phi(B)X_t &= -\sum_{k=0}^{p} \phi_k X_{t-k} = -\sum_{j=-\infty}^{\infty} \sum_{i=0}^{q} \sum_{k=0}^{p} \chi_j \theta_i \phi_k Z_{t-i-j-k} \\ &= -\sum_{m=-\infty}^{\infty} \sum_{i=0}^{q} \left(\sum_{k=0}^{p} \chi_{m-k}\phi_k\right) \theta_i Z_{t-i-m} = \sum_{i=0}^{q} \theta_i Z_{t-i} = \theta(B) Z_t,\end{aligned}$$

which completes the proof. □

An index change in the explicit formula (3.3) for X shows that

$$X_t = \phi(B)^{-1}\theta(B) Z_t = \sum_{k=-\infty}^{\infty} \left(\sum_{i=0}^{q} \chi_{k-i}\theta_i\right) Z_{t-k}.$$

With $\psi(z) := \phi(z)^{-1}\theta(z)$, we can write this as

$$X_t = \sum_{k=-\infty}^{\infty} \psi_k Z_{t-k},$$

where $\psi_k = (\sum_{i=0}^{q} \chi_{k-i}\theta_i)$ are the coefficients in the Laurent series expansion of $\psi(z)$.

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 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 $\psi_j = 0$ for $j < 0$, i.e., the $\psi(z)$ as defined in the proof of Proposition 3.2.4 has a power series expansion with only positive powers,

$$\psi(z) = \sum_{j=0}^{\infty} \psi_j z^j$$

and $\sum_{j=0}^{\infty} |\psi_j| < +\infty$, so that $|\psi(z)| < \infty$ for all $z \in \mathbb{C}$ with $|z| \leq 1$. Since

$$\psi(z) = \theta(z)/\phi(z)$$

and ϕ and θ have no common zeros, this means that $\phi(z) \neq 0$ for all $z \in \mathbb{C}$ with $|z| \leq 1$. Conversely, if $\phi(z) \neq 0$ for all $z \in \mathbb{C}$ with $|z| \leq 1$ then $\psi(z) = \theta(z)/\phi(z)$ can, by a fact from complex analysis, be expanded into a power series of only positive powers from which existence and causality follows. □

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 for a causal process, $\psi(z)\phi(z) = \theta(z)$, 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

$$\psi_j - \sum_{k=1}^{p} \phi_k \psi_{j-k} = \theta_j \tag{3.4}$$

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 (3.4), 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 $j < 0$.

Example 3.2.10. Let us again consider the time series X of Example 2.2.4, which was given by

$$X_t = Y_t(Z_t + Z_{t-1})$$

for all $t \in \mathbb{Z}$, where $Z \sim \text{IID}(0, \sigma_Z^2)$ and Y is a stationary time series with ACVF γ_Y. For simplicity we now assume that $\gamma_Y(h) = 2^{-|h|}$ and that $\mu_Y = 0$. Then, using the computations from Example 2.2.7, we derive

$$\gamma_X(h) = \begin{cases} 2\sigma_Z^2 & \text{if } h = 0, \\ \frac{\sigma_Z^2}{2} & \text{if } |h| = 1, \\ 0 & \text{else.} \end{cases}$$

This ACVF is zero when $|h| > 1$, exactly like the ACVF of the MA(1) process in Example 2.4.10. Therefore, we can represent X as an MA(1) process by

$$X_t = V_t + \theta_1 V_{t-1},$$

for some white noise process $V \sim \mathrm{WN}(0, \sigma_V^2)$. To find θ_1 and σ_V^2, we note from Example 2.4.10 that $\gamma_X(0) = (1+\theta_1^2)\sigma_V^2$ and that $\gamma_X(1) = \theta_1\sigma_V^2$. Therefore, by solving the system of equations

$$\begin{aligned} 2\sigma_Z^2 &= (1+\theta_1^2)\sigma_V^2, \\ \frac{\sigma_Z^2}{2} &= \theta_1\sigma_V^2 \end{aligned}$$

and picking one of the solutions, we have the representation

$$X_t = V_t + (2-\sqrt{3})V_{t-1}, \tag{3.5}$$

where $V \sim \mathrm{WN}(0, \frac{\sigma_Z^2}{2(2-\sqrt{3})})$. That the process V_t exists follows from [6, Proposition 2.1.1] and due to the representation (3.5) we directly see that the process X is causal by definition. To see if it is invertible, we must by Lemma 3.2.9 check if any of the roots of the polynomial $f(z) = 1+(2-\sqrt{3})z$ is inside of the unit circle. Since the only root is $-(2-\sqrt{3})^{-1}$ and since $|-(2-\sqrt{3})^{-1}| > 1$, the process X *is* invertible.

From now on, we will restrict our intention to causal and invertible models, since any non-causal and/or non-invertible ARMA process can be rewritten as a causal/invertible ARMA process with respect to another white noise series, see [4, page 127].

3.2.1 Autocovariance 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 (3.4).

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

Method 3.2.11. Proposition 3.1.2 implies with the above representation that

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

The coefficients can be computed using (3.4).

Method 3.2.12. 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

$$\gamma(k) - \sum_{j=1}^{p} \phi_j \gamma(k-j) = \sigma^2 \sum_{j=0}^{q-k} \theta_{k+j} \psi_j \tag{3.6}$$

for $0 \leq k \leq q$ and

$$\gamma(k) - \sum_{j=1}^{p} \phi_j \gamma(k-j) = 0 \tag{3.7}$$

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

Method 3.2.13. This method is a numerical version of Method 3.2.12. Use (3.6), and if $p > q$ also (3.7), to create a system of linear equations for $k = 0, \ldots, p$ and solve them numerically to find $\gamma(0), \ldots, \gamma(p)$. Afterwards, use the result to successively determine $\gamma(j)$ 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 3.2.14. Let X be an ARMA(p, q) process. The *partial autocorrelation function* α (*PACF* for short) of X is defined by

$$\begin{aligned} \alpha(0) &:= 1, \\ \alpha(h) &:= \phi_{hh} \end{aligned}$$

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

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

i.e., the coefficient a_h in

$$b_{h+1}^{l}((X_1, \ldots, X_h)) = \sum_{i=1}^{h} a_i X_{h+1-i}.$$

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

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

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

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

and for $h \geq 2$

$$\alpha(h) = \mathsf{Cor}(X_{t+h} - b^l_{t+h}(X^h), X_t - b^l_t(X^h))$$

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. This equivalence holds for more general time series, but not for all weakly stationary time series, which is why we choose to restrict the definition to ARMA processes.

Example 3.2.15. Let us again revisit the set of quarterly earnings of H&M from Example 2.1.2, after having removed the trend and seasonality in Example 2.5.5. We are interested in the correlation structure of the remaining data. We can compute its sample PACF $\hat{\alpha}$, which we do for $h = 0, 1, \ldots, 10$ in MATLAB with the following code, assuming that we have calculated the sample ACF beforehand:

```
% Input:
%   gamma double vec, acf in vector form
%   lags int, number of lags to evaluate

% Output:
%       pacf double vec, row vector of sample pacf values

pacf=ones(1,lags+1);
% Find phi_k,k
phi=zeros(lags,lags);
v=zeros(1,lags);
phi(1,1)=gamma(2)/gamma(1);
v(1)=gamma(1)*(1-phi(1,1)^2);
for k=2:lags
        phi(k,k)=(gamma(k+1)-gamma(k+1-(1:k-1))*phi(k-1,1:k-1)')
            /v(k-1);
        phi(k,1:k-1)=phi(k-1,1:k-1)-phi(k,k)*phi(k-1,k-1:-1:1);
        v(k)=v(k-1)*(1-phi(k,k)^2);
end
pacf(2:end)=diag(phi); % phi_k,k
```

The result is shown in Figure 3.1. We see that after removing any deterministic effects on the earnings, the earnings for a given quarter are most influenced by the earnings at the previous quarter (lag 1 in Figure 3.1) and of the previous quarter from the previous year (lag 5).

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 1.1.8 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.

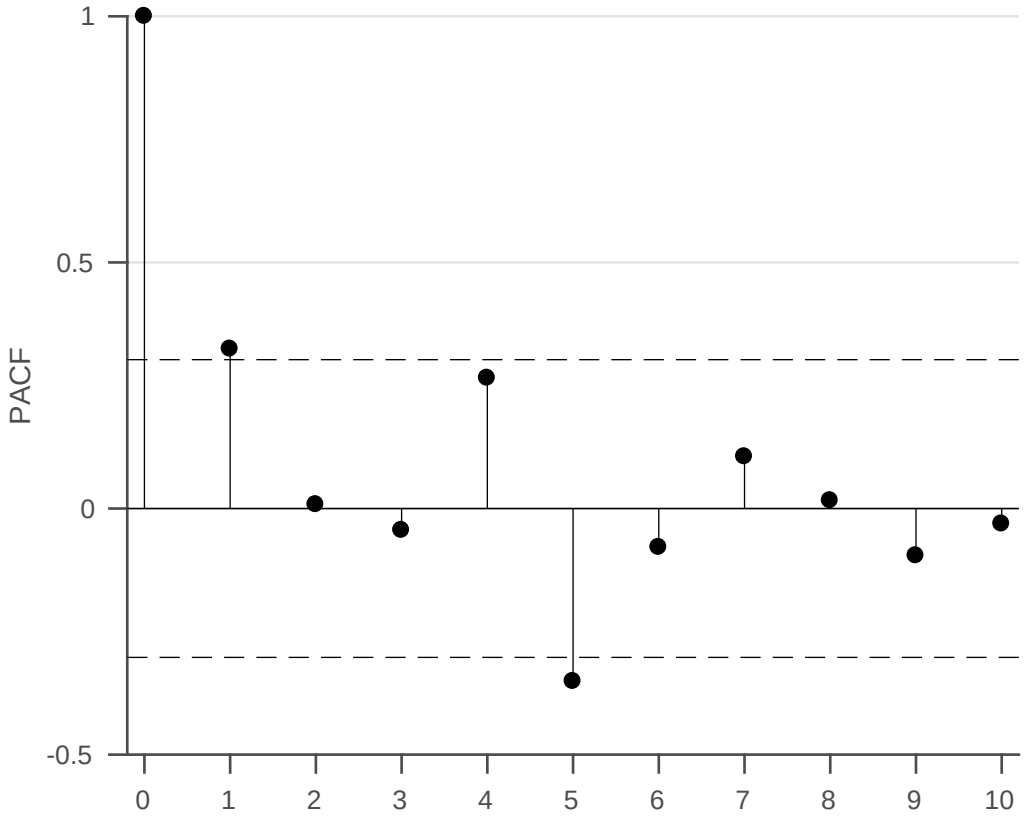


Figure 3.1: Sample partial autocorrelation function for the *deseasonlized* H&M data of Example 2.1.2 and Example 2.5.5.

Example 3.2.16. In Figure 3.2 we plot the ACF and PACF of the causal AR(3) process given by the equation

$$X_t - \frac{5}{4}X_{t-1} + \frac{1}{2}X_{t-2} - \frac{1}{16}X_{t-3} = Z_t$$

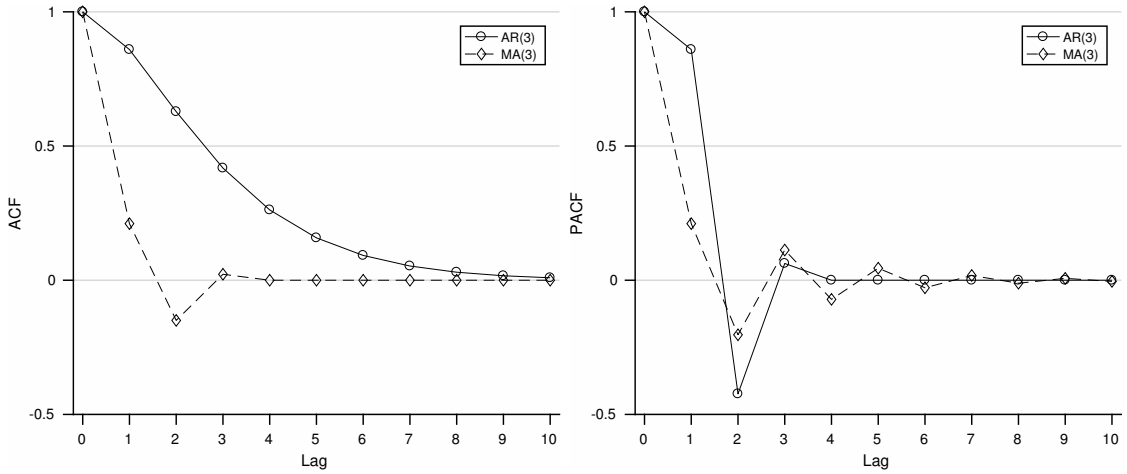
and of the invertible MA(3) process given by

$$X_t = Z_t + \frac{5}{4}Z_{t-1} - \frac{1}{2}Z_{t-2} + \frac{1}{16}Z_{t-3}.$$

We note that the PACF of the purely autoregressive process is zero for lags greater than 3, as expected. In the same way, the ACF of the MA(3) process is zero for lags greater than 3, something that can be seen as an easy consequence of Proposition 3.1.2.

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 cases. We will discuss the order selection in Section 3.2.3. Here we will give methods to determine the parameters $(\phi_j, j = 1, \ldots, p)$ and $(\theta_j, j = 1, \ldots, q)$. The main method is maximum likelihood, but this requires good initial values of $(\phi_j, j = 1, \ldots, p)$ and $(\theta_j, j = 1, \ldots, q)$. For this reason we first describe preliminary methods. We start with a computation method for purely autoregressive, AR(p), models.



(a) The ACF (b) The PACF

Figure 3.2: The ACFs and PACFs of an AR(3) and an MA(3) process.

So called *Yule–Walker estimation* can be derived from Method 3.2.12. We observe that the equations in Method 3.2.12 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 3.2.17 (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 $\hat{R}_p := (\hat{\rho}(i-j))_{i,j=1}^{p}$ denotes the sample autocorrelation matrix.

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 intervals: since $\hat{\phi}_i \sim \mathcal{N}(\phi_i, n^{-1}\sigma^2\Gamma_{p,ii}^{-1})$, $i = 1, \ldots, p$, the interval $\hat{\phi}_i \pm \Phi_{1-\alpha/2} n^{-1/2}\sigma(\Gamma_{p,ii}^{-1})^{1/2}$ will contain ϕ_i with approximate $1-\alpha$ probability. Here $\Phi_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution, $\Gamma_{p,ii}^{-1}$ is the i:th diagonal entry of Γ_p^{-1} and $n^{-1/2}\sigma(\Gamma_{p,ii}^{-1})^{1/2}$.

Furthermore we remark that 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. Instead, we have the following algorithm, which is a variant of a least square regression.

Method 3.2.18 (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 3.2.17. 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 maximum likelihood estimation in the context of ARMA(p,q) processes. For brevity, we will write $\phi := (\phi_1, \ldots, \phi_p)$ and $\theta := (\theta_1, \ldots, \theta_q)$ for the vectors of parameters that we are interested in. We assume that we are given a Gaussian ARMA(p,q) process (which follows if we assume that $Z \sim \mathrm{IID}\,\mathcal{N}(0, \sigma^2)$). Then for any fixed values ϕ, θ, and σ^2, since the innovations $X_1 - \hat{X}_1, \ldots, X_n - \hat{X}_n$ are uncorrelated (see the proof of Method 2.4.14, the innovations algorithm) and Gaussian, they are independent, where $\hat{X}_1 := 0$ and $\hat{X}_j := b_j^l((X_1, \ldots, X_{j-1}))$, $j \geq 2$. If we had observed the innovations instead of the data X_i, the definition of the likelihood would be as simple as a product of densities. However, as it turns out, the likelihood function for the data set $(X_1, \ldots, X_n)$ can be considerably simplified by rewriting it in terms of innovations.

Let now $\Gamma_n = \Gamma_n(\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q, \sigma^2)$ denote the covariance matrix of $\mathbf{X}_n := (X_1, \ldots, X_n)'$, i.e., let $(\Gamma_n)_{i,j} = \gamma_X(i-j)$, and assume that it is nonsingular. The *likelihood* $L(\Gamma_n)$ of $(X_1, \ldots, X_n)$ is given by

$$L(\Gamma_n) = f(\mathbf{X}_n|\Gamma_n) = (2\pi)^{-n/2}(\det \Gamma_n)^{-1/2} \exp\left(-2^{-1}\mathbf{X}_n'\Gamma_n^{-1}\mathbf{X}_n\right), \tag{3.8}$$

i.e., the density function of $\mathbf{X}_n = (X_1, \ldots, X_n)'$ but written as a function of Γ_n. We simplify this equation using the properties of the innovations noted above.

A consequence of Lemma 2.4.13 is that one can write $\mathbf{X}_n$ as

$$\mathbf{X}_n = C_n(\mathbf{X}_n - \hat{\mathbf{X}}_n)$$

where $\hat{\mathbf{X}}_n = (\hat{X}_1, \ldots, \hat{X}_n)'$ and C_n is a $n \times n$ lower triangular matrix with ones on the diagonal. Moreover, since $(\mathbf{X}_n - \hat{\mathbf{X}}_n)$ is a vector of independent random variables, it has the covariance matrix $D_n = \text{diag}(v_0, \ldots, v_{n-1})$ where, for $i = 0, \ldots, n$, $v_i = \text{MSE}(X_{i+1}, \hat{X}_{i+1}) = \mathbb{E}((X_{i+1} - \hat{X}_{i+1})^2)$. By linearity of the expectation operator and some matrix algebra, we have

$$\Gamma_n = \mathbb{E}(\mathbf{X}_n\mathbf{X}_n') = \mathbb{E}(C_n(\mathbf{X}_n - \hat{\mathbf{X}}_n)(\mathbf{X}_n - \hat{\mathbf{X}}_n)'C_n') = C_nD_nC_n'.$$

Using this observation, along with the fact that $C_n^{-1}\mathbf{X}_n = (\mathbf{X}_n - \hat{\mathbf{X}}_n)$, we get that,

$$\begin{aligned}\mathbf{X}_n'\Gamma_n^{-1}\mathbf{X}_n &= \mathbf{X}_n'(C_nD_nC_n')^{-1}\mathbf{X}_n = \mathbf{X}_n'(C_n')^{-1}D_n^{-1}(C_n)^{-1}\mathbf{X}_n = ((C_n)^{-1}\mathbf{X}_n)'D_n^{-1}(C_n)^{-1}\mathbf{X}_n \\ &= (\mathbf{X}_n - \hat{\mathbf{X}}_n)'D_n^{-1}(\mathbf{X}_n - \hat{\mathbf{X}}_n) = \sum_{j=1}^{n}(X_j - \hat{X}_j)^2/v_{j-1},\end{aligned}$$

since $D_n^{-1} = \text{diag}(v_0^{-1}, \ldots, v_{n-1}^{-1})$. We also have, since D_n is diagonal and C_n is lower triangular with ones on the diagonal,

$$\det(\Gamma_n) = \det(C_n)\det(D_n)\det(C_n) = v_0v_1 \cdots v_{n-1}.$$

Now we have a simple expression for $L(\Gamma_n)$. In order to estimate the white noise variance σ^2, it turns out that defining $r_j := v_j/\sigma^2 = E((X_{j+1} - \hat{X}_{j+1})^2)/\sigma^2$ for $j = 0, \ldots, n-1$ simplifies our algorithm. We leave it as an easy *exercise* to show that, as a consequence of Proposition 3.1.2 and Corollary 2.4.6, $\hat{X}_{j+1}$ and r_j do not depend on σ^2.

Plugging these equations into the likelihood, we have derived the *Gaussian likelihood for an* ARMA(p, q) *process*

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

which can be determined by the innovations algorithm 2.4.14. The maximum likelihood estimates are now obtained by minimizing $-\ln(L(\phi, \theta, \sigma^2))$ with respect to ϕ,θ and σ^2, which is equivalent to maximizing L with respect to these parameters.

Method 3.2.19 (Maximum likelihood estimators)**.** The maximum likelihood estimators of σ^2, ϕ, and θ are determined from the expressions

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

and

$$(\hat{\phi}, \hat{\theta}) = \underset{(\phi,\theta)}{\arg\min}\, \ell(\phi, \theta).$$

Here

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

where $\hat{X}_j$ and r_{j-1}^{-1} can be computed with Method 2.4.14 using the parameters ϕ, θ, (which shows that S is a function of ϕ and θ). ℓ is the function given by

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

The minimization of ℓ is typically done numerically. Initial values can 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 remember that $\hat{X}_j$ and r_{j-1}, $j = 1, \ldots, n$, do not depend on σ^2.)

Remark 3.2.20. For large sample sizes n, the maximum likelihood estimates $\hat{\beta} = (\hat{\phi}_1, \ldots, \hat{\phi}_p, \hat{\theta}_1, \ldots, \hat{\theta}_q)$ is approximately normally distributed with mean $\beta = (\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q)$ and covariance matrix $2H^{-1}(\beta)/n$, where $H = (\delta^2 \ell(\beta)/\delta\beta_i \delta\beta_j)_{i,j=1}^{p+q}$ is the Hessian matrix of the log-likelihood function of Method 3.2.19. This fact is used to compute confidence intervals for $\hat{\phi}_1, \ldots, \hat{\phi}_p, \hat{\theta}_1, \ldots, \hat{\theta}_q$, similar to what was done for the Yule–Walker estimates in Method 3.2.17.

Remark 3.2.21. A common alternative to assuming that $Z = (Z_t, t \in \mathbb{Z})$ is Gaussian white noise (and therefore IID) is to assume that it is $Z \sim \text{IID}(0, \sigma^2)$ and that Z_t follows a Student t-distribution for all times t. The maximum likelihood algorithm then takes on a more complicated form. However, in this case the Gaussian maximum likelihood estimates can still provide good approximate values of the parameters or at least good initial values.

Having fitted the model, it remains to check that the model was chosen adequately. The *rescaled residuals* are defined by

$$\hat{R}_t := (X_t - \hat{X}_t)/v_{t-1}^{1/2},$$

where $\hat{X}_t$ and v_{t-1} are computed using, e.g., Method 2.4.14 based on the maximum likelihood estimates $\hat{\phi}_1, \ldots, \hat{\phi}_p, \hat{\theta}_1, \ldots, \hat{\theta}_q$ and $\hat{\sigma}^2$. If the model describes the data accurately, they should behave like the standardized driving noise Z_t, i.e., looking like white noise with variance 1 and being approximately normally distributed if a Gaussian maximum likelihood estimation procedure was used. If instead a maximum likelihood estimation assuming a Student t-distribution for Z_t was used, $\hat{R}_t$ should follow a t-distribution instead and be approximately IID with variance 1, and so on. The motivation for this is the fact that $\mathbb{E}((\hat{R}_t - Z_t/\sigma)^2) \to 0$ as $t \to \infty$.

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.

Recalling Example 3.2.16, we note that 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" along with a slowly decaying partial autocorrelation function 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.

There are many ways to do order selection in a more systematic way. We introduce two methods based on information theory.

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 first introduce the AICC criterion, where AIC stands for *Akaike's Information Criterion* and the last C for *biased-Corrected*. The AICC is an estimate of the *Kullback–Leibler divergence* (a measure of how one probability distribution is different from another) of the estimated distribution from the exact distribution of the data. It is based on the assumption that $Z \sim \mathrm{IID}\,\mathcal{N}(0, \sigma^2)$, but it seems to be robust against moderate deviations from normality (such as when Z_t follows a t-distribution).

Method 3.2.22 (AICC criterion)**.** Choose p, q, ϕ_p, and θ_q to minimize

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

where $\phi_p = (\phi_1, \ldots, \phi_p)$ and $\theta_q = (\theta_1, \ldots, \theta_q)$.

To calculate this criterion in practice, we fit a large set of of models with different orders (p, q) to the data and choose the one that minimizes the negative log-likelihood, modified by the penalty factor $2n(p+q+1)/(n-p-q-2)$.

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 are for example obtained by the BIC (*Bayesian Information Criterion*), which also punishes the choice of large values for p and q and tries to avoid overfitting.

Method 3.2.23 (BIC criterion)**.** Choose p and q to minimize

$$(n-p-q) \ln\left(n\hat{\sigma}^2/(n-p-q)\right) + n\left(1 + \ln\sqrt{2\pi}\right) + (p+q) \ln\left(\left(\sum_{t=1}^{n} X_t^2 - n\hat{\sigma}^2\right)/(p+q)\right),$$

where $\hat{\sigma}^2$ denotes the maximum likelihood estimate of the white noise variance.

One negative aspect of the BIC is the efficiency of finding minimizers. While order selection by minimization of the AICC is asymptotically efficient for causal and invertible ARMA processes, this does not hold for the BIC. Efficiency in this context means that choosing the model that minimizes the AICC will asymptotically produce a model with minimum one-step ahead prediction errors, in a certain sense.

Under *constrained* maximum likelihood estimation, i.e., when specific coefficients are assumed to be zero in the maximum likelihood procedure, the value $p+q+1$ is replaced by m, the number of non-zero coefficients, in both the AICC and the BIC. For more details and references to the literature, the reader is referred to [6, Section 5.5].

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 [13].

3.2.4 Model building for ARMA processes

We take this opportunity to summarize what we have noted about building an ARMA time series model so far. If the reader have a data set that he or she wants to model with an ARMA model, one could proceed in the following way (see also [6, Chapter 5.3]).

- Remove trend and seasonality until you believe that the data can be modeled as a stationary time series, see Section 2.5.
- Identify the order of the ARMA model for the time series by either looking at the ACF/PACF as noted in sections 3.2.1 and 3.2.3 or (more systematically) by fitting (using maximum likelihood or Hannan–Rissanen estimation) successively higher order ARMA(p, q) to the data and choose a number of candidate models with small AICC and/or BIC values, as outlined in 3.2.3. Try to simplify the models by setting parameters to zero if they are not significant (i.e., if their confidence interval includes 0), and see if this reduces the AICC/BIC (the number $p + q$ in these criteria are then replaced by m, the number of non-zero coefficients).
- Estimate the final candidate models using maximum likelihood, see Section 3.2.19.
- Compute the residuals $\hat{R}_t$ for the different models and check that they are consistent with the specified distribution and temporal covariance structure for Z_t, see Section 3.2.19. If they are, the model is considered to be adequate for the data. The final model is chosen to be the one with residuals $\hat{R}_t$ most like Z_t. Another alternative is to reserve some *test data* in the form of the end of the time series at the start of this process and then compute forecasts for this test data using the candidate models. Then one chooses the final model as the one with the minimum forecast error.

3.2.5 Forecasting of ARMA processes

The innovations algorithm 2.4.14 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 $Z \sim \mathrm{WN}(0, \sigma^2)$, 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

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

where $m = \max\{p, q\}$ (cf. [2]).

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 2.4.14 to the process W we obtain

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

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 $j > q$, which follows from the fact that $\kappa(r, s) = 0$ for $r > m$ and $|r - s| > q$.

Note that $\hat{W}_{n+1}$ is the best linear one-step predictor of W_{n+1} in terms of $W^n := (W_1, \ldots, W_n)$, i.e.,

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

Furthermore one can show that

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

using the linearity of the prediction estimator and the fact that W can be expressed as a linear combination of X and vice versa. This implies that $\hat{W}_t - W_t = \sigma^{-1}(\hat{X}_t - X_t)$ for $t \geq 1$, 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

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

where we recall that the coefficients $(\theta_{nj}, n \in \mathbb{N}, j \leq \min\{n, m\})$ and the mean squared errors $\mathrm{MSE}(\hat{W}_{n+1}, W_{n+1}) = v_n$ are found recursively from the innovations algorithm 2.4.14. The best linear estimators can be computed recursively.

Let us next consider h-step predictors of an ARMA(p, q) process. For this, let us first note that, using Lemma 2.4.13, we can write

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

We now claim that $b^l_{n+h}(W^n)$ is obtained by dropping the first $h - 1$ terms of this sum. This follows by Proposition 2.4.5: $b^l_{n+h}(W^n)$ is the unique linear combination of $W_1, \ldots, W_n$ that fulfills $\mathbb{E}((W_{n+h} - b^l_{n+h}(W^n))W_k) = 0$ for $k = 1, \ldots, n$ and by linearity of the expectation, we see that for these k

$$\begin{aligned} &\mathbb{E}\left(\left(W_{n+h} - \sum_{j=h}^{n+h-1} \theta_{(n+h-1)j}(W_{n+h-j} - \hat{W}_{n+h-j})\right) W_k\right) \\ &= \mathbb{E}\left(\left(W_{n+h} - \hat{W}_{n+h} + \sum_{j=1}^{h-1} \theta_{(n+h-1)j}(W_{n+h-j} - \hat{W}_{n+h-j})\right) W_k\right) \\ &= \mathbb{E}\Big((W_{n+h} - \hat{W}_{n+h})W_k\Big) + \sum_{j=1}^{h-1} \theta_{(n+h-1)j}\,\mathbb{E}\Big((W_{n+h-j} - \hat{W}_{n+h-j})W_k\Big) = 0, \end{aligned}$$

where we have used Proposition 2.4.5 for each of the individual terms. Therefore

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

Similar to the case that $h = 1$, one can show that the h-step predictors $b^l_{n+h}(X^n)$ with $X^n := (X_1, \ldots, X_n)$ satisfy

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

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

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

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

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

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 2.4.

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 $h \ge 1$ 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. If Z is not a Gaussian time series, one can still construct an approximate prediction bounds in the case that $h = 1$ using the fact that for an invertible ARMA process,

$$\mathbb{E}((X_t - \hat{X}_t - Z_t)^2) \to 0$$

as $t \to \infty$. This implies that the distribution of $X_t - \hat{X}_t$ will be approximately equal to the distribution of Z_t for large t, so any confidence interval for Z_t can be used as a confidence interval for $X_t - \hat{X}_t$.

3.3 (S)ARIMA models and unit root tests

We have seen in Section 2.5 that one way of transforming a non-stationary time series into a stationary one is given by the differencing approach. That is, we apply $\nabla = 1 - B$ one or several times to delete the trend component and $\nabla_s = 1 - B^s$ to delete a seasonal component of period $s \in \mathbb{N}$ from the classical decomposition model (2.11). The class of SARIMA models is a way of mathematically formalizing this approach in the case that the differenced stationary time series is an ARMA(p, q) process.

Definition 3.3.1. Let X be a stochastic process and let $d, D \in \mathbb{N}_0$. Then X is a SARIMA$(p, d, q) \times (P, D, Q)_s$ *process* if the process Y defined by $Y_t := \nabla^d \nabla_s X_t$ is a causal ARMA process defined by

$$\phi(B)\Phi(B^s)Y_t = \theta(B)\Theta(B^s)Z_t,$$

where $\phi(z) = 1 - \phi_1 z - \ldots - \phi_p z^p$, $\Phi(z) = 1 - \Phi_1 z - \ldots - \Phi_P z^P$, $\theta(z) = 1 + \theta_1 z^1 + \ldots + \theta_q z^q$ and $\Theta(z) = 1 + \Theta_1 z^1 + \ldots + \Theta_Q z^Q$ for $p, q, P, Q \in \mathbb{N}_0$.

Here the abbreviation ARIMA stands for *seasonal autoregressive integrated moving average*. Note that X is in general a non-stationary time series, and neither the mean nor the covariance function are determined by the difference equations above. The process Y in this definition is a causal ARMA$(p + sP, q + sQ)$ process since the polynomials $\phi^*(\cdot) = \phi(\cdot)\Phi(\cdot)$ and $\theta^*(\cdot) = \theta(\cdot)\Theta(\cdot)$ are of degree $p + sP$ and $q + sQ$. The reason for including the operator $\Phi(B^s)$ and $\Theta(B^s)$ is to account for stochastic seasonal effects which can remain after removing the seasonal component in the classical decomposition model (2.11). It is often found that a high order ARMA process provides a good fit to the remaining data, provided that most of its coefficients ϕ_i and θ_j are zero except for at a few indices i, j that are multiples of s, or close to such multiples. One example of such an ARMA process is given by Y in the definition above. In modelling, an equivalent approach is to instead let Y be a standard ARMA process and manually specify which coefficients should be set to zero when estimating the model.

A SARIMA model without any seasonality is called an ARIMA model. These are particularly important in financial time series analysis, for example when judging whether an asset $(S_t, t \in \mathbb{Z})$ is best modeled by *log-returns* X, i.e., $X_t := \log(S_{t+1}) - \log(S_t)$, or *log-prices* X, i.e., $X_t := \log(S_t)$.

Definition 3.3.2. Let X be a stochastic process and d a nonnegative integer. Then X is an ARIMA(p, d, q) *process* if the process Y defined by $Y_t := (1 - B)^d X_t$ is a causal ARMA(p, q) process.

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 $Z \sim \text{WN}(0, \sigma^2)$ 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 $z = 1$. It is very difficult in practice 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.

An ARIMA model is an appropriate choice if the autocovariance function is very slowly decaying. 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. As noted above, 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 follows a model with roots on the unit circle is due to Dickey and Fuller (see [8]) and described in what follows. We restrict our attention to autoregressive processes and start with the case of an AR(1) model. Therefore assume that $(X_1, \ldots, X_n)$ are observations of the AR(1) model

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

where $Z \sim \text{WN}(0, \sigma^2)$, $|\phi_1| < 1$, 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 [12, 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 $\mathrm{AR}(p)$ 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 $\mathrm{AR}(p-1)$ 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 3.3.3 (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}, …, ∇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 for 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

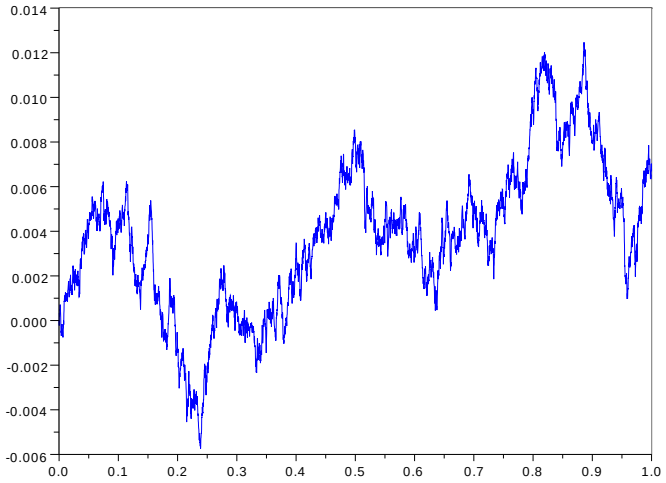


Figure 4.1: Path of a Brownian motion.

the explicit solution

$$P_t = P_0 \exp((\mu - 2^{-1}\sigma^2)t + \sigma B_t).$$

The corresponding log-returns $X_t := \log(P_{t+1}) - \log(P_t)$, evaluated at time points $t \in \mathbb{Z}$, then take the form

$$X_t = \log(P_{t+1}) - \log(P_t) = \mu - \frac{\sigma^2}{2} + \sigma(B_{t+1} - B_t).$$

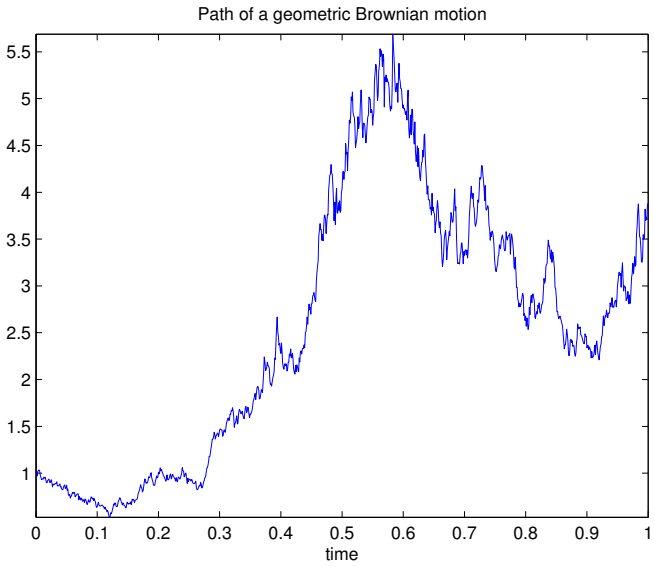


Figure 4.2: Path of a geometric Brownian motion with $P_0 = \mu = \sigma = 1$

If we assume that the mean has been removed before we model this time series, then, since the increments $B_{t+1} - B_t \sim \mathcal{N}(0,1)$ and are mutually independent, this means that $X_t = \sigma Z_t$, where $Z = (Z_t, t \in \mathbb{Z}) \sim \text{IID}\,\mathcal{N}(0,1)$. However, this simple model does not exhibit the properties we see in practice when we want X_t to model a time series of log-returns. Specifically, the problems are the following.

- Since Z is IID in the model, so is X which means that for two time points s and t, X_t and X_s should be independent. In practice, we only observe that they are uncorrelated, i.e., there is some dependence in time but it is not linear.
- The volatility σ is in the model constant over time, but often one see *volatility clustering*, i.e., periods of low and high variance tend to cluster together (see Figure 4.3).
- X_t should be Gaussian, but often one observes "fat tails" (i.e., large positive and negative jumps occur more frequently than one would expect from a normal distribution). This means that the *kurtosis* $\frac{\mathbb{E}(X_t^4)}{\mathbb{E}(X_t^2)^2}$ is greater than 3, which is the value under the Gaussian IID model.

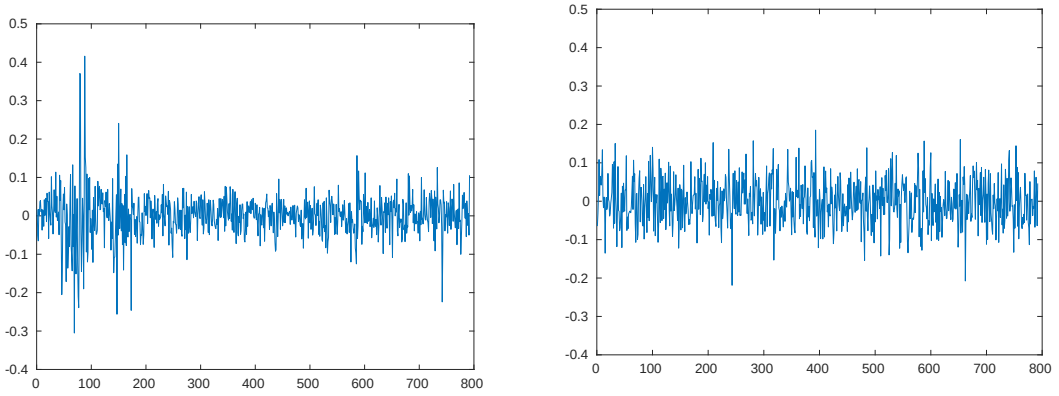
In this chapter we will look at models which attempt to solve at least some of these problems. Most importantly, we will replace the volatility σ in the equation $X_t = \sigma Z_t$ with a stochastic process $(\sigma_t, t \in \mathbb{Z})$.

Definition 4.0.1. A stochastic process $X = (X_t, t \in \mathbb{Z})$ is said to follow a *random variance model* if

$$X_t = \sigma_t Z_t \tag{4.1}$$

for all $t \in \mathbb{Z}$, where $Z = (Z_t, t \in \mathbb{Z})$ is IID$(0,1)$ and $\sigma = (\sigma_t, t \in \mathbb{Z})$ is an unspecified stochastic process called the *volatility*. If X_t can be written as a deterministic function of $(Z_s, s \leq t)$ for all $t \in \mathbb{Z}$, then X is said to be *causal*.

Since the random variables $(Z_s, s \leq t)$ are independent of the random variables $(Z_s, s > t)$, any process that is causal in this sense is independent of the future, if we by this mean any random variable that can be written as a function of $(Z_s, s > t)$. We note that there is no standard nomenclature for random variance models, we mainly use that of [6]. For example, in [22] a causal



(a) Returns of the S&P500 index.

(b) Observations from the model $X_t = \sigma Z_t$, where $Z \sim \text{IID}\,\mathcal{N}(0,1)$ and $\sigma \in \mathbb{R}$.

Figure 4.3: Comparison of 792 observations of simple monthly returns of the S&P500 index (left) and Gaussian white noise (right). Both time series have the same sample variance.

process is called *non-anticipative*. All models that we encounter in this chapter will be random variance models. It should be noted that modeling volatility is a difficult problem and no one model works well in all situations.

One option to estimate the volatility from data 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.

4.1 Definitions and properties

Let us start by introducing the necessary definitions and the theoretical background of the ARCH and GARCH models. The main idea behind both processes is to incorporate the possibility of volatility clustering in the realizations of the model. One possibility would be to let σ_t be a function of time, but then we could end up with the variance of X_t being non-constant, so that we would have a non-stationary model which would be hard to do estimate from data. Instead we allow for non-constant *conditional* variance, i.e.,

$$\mathsf{Var}(X_t | X_{t-1}, X_{t-2}, \ldots) = \mathbb{E}((X_t - \mathbb{E}(X_t))^2 | X_{t-1}, X_{t-2}, \ldots) \neq \text{ constant}$$

for $t \in \mathbb{Z}$. This concept is called *conditional heteroscedasticity*. This is accomplished by letting σ_t^2 in $X_t = \sigma_t Z_t$ be an ARMA-like process.

Note that the simple model $X_t = \sigma Z_t$ for $t \in \mathbb{Z}$, with $\sigma \in \mathbb{R}$, $Z \sim \text{IID}\,\mathcal{N}(0,1)$, does not exhibit conditional heteroscedasticity. This is because that since $Z \sim \text{IID}(0,1)$, $X_t = \sigma Z_t$ is independent of the random variables $(\sigma Z_s, s < t) = (X_s, s < t)$, so the conditional variance above coincides with the usual variance.

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 it satisfies the ARCH equations

$$X_t = \sigma_t Z_t,$$

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

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

$\alpha_0 > 0$, $\alpha_j \geq 0$ for $j = 1, \ldots, p$.

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

The requirements $\alpha_0 > 0$ and $\alpha_j \geq 0$, $j \geq 1$, guarantee that $\sigma_t > 0$.

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. If X is a causal ARCH(p) process, we find from the defining equation for σ^2 that σ_t^2 can be written as a deterministic function of $(Z_s, s \leq t-1)$ and is therefore independent of the random variables $(Z_s, s \geq t)$. We leave it as a simple exercise to show that this implies that $\mathbb{E}(X_t^2) = \mathbb{E}(\sigma_t^2)$. This, in combination with the fact that (by definition) X is causal, yields

$$\mathbb{E}(X_t^2) = \mathbb{E}(\sigma_t^2) = \alpha_0 + \sum_{j=1}^{p} \alpha_j \, \mathbb{E}(X_t^2) = \alpha_0 + \alpha(1) \, \mathbb{E}(X_t^2).$$

This means that we must have $1 - \alpha(1) > 0$, which in turn yields

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

If $\mathbb{E}(X_t^4) < \infty$, it can be shown that $(X_t^2, t \in \mathbb{Z})$ is an AR process, a fact that can be useful in identification of ARCH processes.

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 a stationary solution to the GARCH equations

$$X_t = \sigma_t Z_t,$$

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

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

with $\alpha_0 > 0$, $\alpha_j \geq 0$ for $j = 1, \ldots, p$, $\beta_i \geq 0$ for $i = 1, \ldots, q$.

Remark 4.1.3. Typical one chooses the distribution of Z_t as $Z_t \sim \mathcal{N}(0,1)$ or $\sqrt{\nu/(\nu-2)} Z_t \sim t_\nu$ for all $t \in \mathbb{Z}$, where the factor $\sqrt{\nu/(\nu-2)}$ in front of Z_t ensures that $\mathsf{Var}(Z_t) = 1$. In the latter case, Z is said to follow a *generalized* or *non-standardized* t-distribution, but we usually refer to it as a t-distribution when there is no risk of confusion.

In practice, it is commonly found that a GARCH(1, 1) process yields a sufficient model, without the need for higher orders. In the next proposition, we show a condition for the existence of this process.

Proposition 4.1.4 (Existence of a GARCH(1, 1) process)**.** *If $\alpha_1+\beta_1 < 1$, there exists a stationary solution $X = (X_t, t \in \mathbb{Z})$ to the* GARCH(1, 1) *equations that is given by the equation*

$$X_t = \sigma_t Z_t,$$

where $Z \sim \mathrm{IID}(0, 1)$ and

$$\sigma_t^2 = \alpha_0 \left(1 + \sum_{i=1}^{\infty} (\alpha_1 Z_{t-1}^2 + \beta_1)(\alpha_1 Z_{t-2}^2 + \beta_1) \cdots (\alpha_1 Z_{t-i}^2 + \beta_1)\right). \tag{4.2}$$

It is unique (P-a.s.), strictly stationary and causal. Conversely, if $\alpha_1 + \beta_1 \geq 1$, then there no non-zero stationary solution to the GARCH(1, 1) *equations for which σ_t can be written as a deterministic function of $(Z_s, s < t)$ for all $t \in \mathbb{Z}$.*

Proof. First, let us note that due to the monotone convergence theorem, and since $Z \sim \mathrm{IID}(0, 1)$,

$$\begin{aligned}\mathbb{E}(|\sigma_t^2|) = \mathbb{E}(\sigma_t^2) &= \alpha_0 \left(1 + \sum_{j=1}^{\infty} \mathbb{E}\left((\alpha_1 Z_{t-1}^2 + \beta_1)(\alpha_1 Z_{t-2}^2 + \beta_1) \cdots (\alpha_1 Z_{t-j}^2 + \beta_1)\right)\right) \\ &= \alpha_0 \left(1 + \sum_{j=1}^{\infty} (\alpha_1 + \beta_1)^j\right) = \frac{\alpha_0}{1 - \alpha_1 - \beta_1} < \infty,\end{aligned}$$

by assumption, where we used the geometric sum formula. This shows that σ^2 is a time series with finite mean, and hence that $P(\sigma_t^2 < \infty) = 1$. To see that X with σ_t^2 defined by (4.2) satisfies the GARCH equations, i.e., that

$$\sigma_t^2 = \alpha_0 + \alpha_1 X_{t-1}^2 + \beta_1 \sigma_{t-1}^2,$$

we apply $X_{t-1}^2 = \sigma_{t-1}^2 Z_{t-1}^2$ to the right hand side and then use (4.2) directly to see that

$$\begin{aligned}&\alpha_0 + \alpha_1 X_{t-1}^2 + \beta_1 \sigma_{t-1}^2 = \alpha_0 + \alpha_1 \sigma_{t-1}^2 Z_{t-1}^2 + \beta_1 \sigma_{t-1}^2 = \alpha_0 + (\alpha_1 Z_{t-1}^2 + \beta_1)\sigma_{t-1}^2 \\ &= \alpha_0 + \alpha_0 \left(\alpha_1 Z_{t-1}^2 + \beta_1 + \sum_{i=1}^{\infty} (\alpha_1 Z_{t-1}^2 + \beta_1)(\alpha_1 Z_{t-2}^2 + \beta_1)(\alpha_1 Z_{t-3}^2 + \beta_1) \cdots (\alpha_1 Z_{t-1-i}^2 + \beta_1)\right) \\ &= \alpha_0 \left(1 + \alpha_1 Z_{t-1}^2 + \beta_1 + \sum_{i=2}^{\infty} (\alpha_1 Z_{t-1}^2 + \beta_1)(\alpha_1 Z_{t-2}^2 + \beta_1) \cdots (\alpha_1 Z_{t-i}^2 + \beta_1)\right) \\ &= \alpha_0 \left(1 + \sum_{i=1}^{\infty} (\alpha_1 Z_{t-1}^2 + \beta_1)(\alpha_1 Z_{t-2}^2 + \beta_1) \cdots (\alpha_1 Z_{t-i}^2 + \beta_1)\right) = \sigma_t^2.\end{aligned}$$

For uniqueness and strict stationarity, we refer the reader to [11]. Clearly, X is causal by construction, and weak stationarity follows from strict stationarity and the fact that by the Cauchy–Schwarz inequality, $\mathbb{E}(X_t^2) \leq \mathbb{E}(\sigma_t^2) < \infty$.

We now show that if $\alpha_1 + \beta_1 \geq 1$, there is no stationary solution to the GARCH(1, 1) equations for which σ_t can be written as a deterministic function of $(Z_s, s < t)$. Suppose that there was such a solution, so that σ_t were independent of Z_t. This would imply that $\mathbb{E}(X_t) = \mathbb{E}(\sigma_t Z_t) = \mathbb{E}(\sigma_t)\,\mathbb{E}(Z_t) = 0$ and, similarly, that $\mathbb{E}(X_t^2) = \mathbb{E}(\sigma_t^2)$. Then, since X should be stationary and have zero mean, $\mathbb{E}(X_t^2) = \gamma_X(0) = \mathbb{E}(X_s^2)$ for all $s, t \in \mathbb{Z}$. Combining these facts yield

$$\mathbb{E}(X_t^2) = \mathbb{E}(\sigma_t^2) = \alpha_0 + \alpha_1\,\mathbb{E}(X_{t-1}^2) + \beta_1\,\mathbb{E}(\sigma_{t-1}^2) = \alpha_0 + \mathbb{E}(X_t^2)(\alpha_1 + \beta_1).$$

Since $\alpha_0 > 0$, this means that $\alpha_1 + \beta_1 < 1$ unless $\mathbb{E}(X_t^2) = 0$ (i.e., unless $X_t = 0$ P-a.s. for all $t \in \mathbb{Z}$). Therefore we get a contradiction, so no such solution can exist. □

Let us now return to the three properties of log-returns that we noted at the start of this chapter and see how they relate to the properties of GARCH(1, 1)-processes. First, log-returns are often uncorrelated in time but not independent. To see that this holds for causal GARCH(1, 1)-processes with $\alpha_1 + \beta_1 < 1$, note first that since σ_t only depends on $(Z_\tau, \tau < t)$ (see above), it is independent of Z_t and we get for $s < t$,

$$\mathsf{Cov}(X_t, X_s) = \mathbb{E}(Z_t Z_s \sigma_t \sigma_s) = \mathbb{E}(Z_t)\,\mathbb{E}(\sigma_t \sigma_s Z_s) = 0,$$

while X_t and X_s are clearly not independent in general. Next, we noted that log-returns display volatility clustering. That this also holds for GARCH(1, 1)-processes is obvious from the defining equation

$$\sigma_t^2 = \alpha_0 + \alpha_1 X_{t-1}^2 + \beta_1 \sigma_{t-1}^2.$$

Finally, log-returns typically have "fat tails", i.e. kurtosis greather than 3. For the GARCH(1, 1) model (cf. [22, Section 3.16]) the kurtosis is given by

$$\frac{\mathbb{E}(X_t^4)}{\mathbb{E}(X_t^2)^2} = \frac{\mu_4(1 - (\alpha_1 + \beta_1)^2)}{1 - \beta_1^2 - 2\alpha_1\beta_1 - \mu_4\alpha_1^2},$$

where $\mu_4 = \mathbb{E}(Z_t^4)/\,\mathbb{E}(Z_t^2)^2$ is the kurtosis of Z_t, $t \in \mathbb{Z}$. In the common case that $Z_t \sim \mathcal{N}(0, 1)$, $\mu_4 = 3$ and it is an easy exercise to deduce that in this case the kurtosis of the GARCH(1, 1) process is indeed greater than 3. Note that this result can also be used to assess uncertainty in volatility estimation.

For general GARCH(p, q) processes, the situation is more complicated. We no longer have a simple expression for σ_t^2, but if a similar condition to $\alpha_1 + \beta_1 < 1$ holds, we have the following result. The notation uses the fact that, similar to the ARCH process, we can rewrite the volatility equation to

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

where

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

Proposition 4.1.5 (Existence of a GARCH(p, q) process)**.** *If*

$$\alpha(1) + \beta(1) = \sum_{i=1}^{p} \alpha_i + \sum_{j=1}^{q} \beta_j < 1,$$

there exists a unique weakly and strictly stationary causal solution $X = (X_t, t \in \mathbb{Z})$ to the GARCH(p, q) *equations and a real-valued sequence $(\psi_j)_{j=0}^{\infty}$ such that $\sum_{j=0}^{\infty} |\psi_j| < \infty$ and σ^2 can be expressed by*

$$\sigma_t^2 = \psi_0 + \sum_{j=1}^{\infty} \psi_j X_{t-j}^2.$$

Conversely, if

$$\alpha(1) + \beta(1) = \sum_{i=1}^{p} \alpha_i + \sum_{j=1}^{q} \beta_j \geq 1,$$

then no stationary and causal solution to the GARCH(p, q) *equations for which σ_t can be written as a deterministic function of $(Z_s, s < t)$ for all $t \in \mathbb{Z}$ exists.*

For a proof, the reader is referred to [11].

Under the condition $\alpha(1) + \beta(1) < 1$ we get, similar to the ARCH case, $\mathbb{E}(X_t^2) = \mathbb{E}(\sigma_t^2)$, 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

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

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

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

with mean $\alpha_0(1-\alpha(1)-\beta(1))^{-1}$, i.e., a process that can be represented by

$$X_t^2 - \sum_{i=1}^{\max\{p,q\}} (\alpha_i + \beta_i) X_{t-i}^2 = \alpha_0 + \eta_t - \sum_{j=1}^{q} \beta_j \eta_{t-j}, \tag{4.3}$$

where $\alpha_i := 0$ for $i > p$, $\beta_i := 0$ for $i > q$ and $\eta_t := X_t^2 - \sigma_t^2$ can be shown to be white noise.

4.2 Parameter estimation and order selection

Let us introduce two methods in this section for estimation of parameters in ARCH and GARCH models. The first one is a direct maximum likelihood method, while the second one uses the ARMA representation to estimate the parameters α_i and β_i. For more methods the reader is referred to [22].

The most common method for parameter estimation in GARCH models is some variant of *conditional maximum likelihood estimation*, which we introduce below. We assume that $\alpha(1) + \beta(1) < 1$ so that a causal stationary solution exists and σ_t^2 only depends on $(Z_s, s < t)$. It is an approximate maximum likelihood method based on the observation that a joint probability density function can be written as a product of a conditional density function and a marginal density function, i.e., $f_{X,Y}(x,y) = f(x|y) f_Y(y)$. Generalizing this approach to a set of observations $(x_j, j = 1, \ldots, n)$ of a GARCH(p,q)-process, we get for the density function of $\mathbf{X}_n := (X_1, \ldots, X_n)'$

$$\begin{aligned} f_{\mathbf{X}_n}((x_1,\ldots,x_n)') &= f(x_n|x_{n-1}, x_{n-2}, \ldots, x_1) \cdots f(x_{p+1}|x_p, x_{p-1}, \ldots, x_1) f(x_p, x_{p-1}, \ldots, x_1) \\ &= f(x_p, x_{p-1}, \ldots, x_1) \prod_{t=p+1}^{n} f(x_t|x_{t-1}, x_{t-2}, \ldots, x_1) \end{aligned}$$

Now, since

$$X_t = \sigma_t Z_t,$$

and

$$\sigma_t^2 = \alpha_0 + \sum_{j=1}^{p} \alpha_j X_{t-j}^2 + \sum_{i=1}^{q} \beta_i \sigma_{t-i}^2, \tag{4.4}$$

the distribution of X_t, $t \geq p+1$, conditioned on $X_1, X_2, \ldots, X_{t-1}$ is known if the values $(\sigma_{p+1-j}^2, j = 1, 2, \ldots, q)$ are assumed to be known. Moreover, when a random variable Y belongs to the *location-scale family*, which includes the Gaussian distribution and the (generalized) t-distribution, and have zero mean, one can show that in terms of its density f_Y, $f_{\sigma Y}(x) = \frac{1}{\sigma} f_Y(\frac{x}{\sigma})$, $\sigma > 0$. This means that we can express the density of X_t in terms of Z_t if Z_t belongs to the location-scale family. Finally, the term $f(x_p, x_{p-1}, \ldots, x_1)$ in (4.2) is deemed too complicated to handle, so it is dropped and we end up with the method below.

Method 4.2.1 (Conditional maximum likelihood estimation)**.** The conditional maximum likelihood estimators $(\hat{\alpha}_0, \ldots, \hat{\alpha}_p, \hat{\beta}_1, \ldots, \hat{\beta}_q, \hat{\theta}_Z)$ are obtained as the values that maximize the *conditional likelihood function*

$$L(\alpha_0, \ldots, \alpha_p, \beta_1, \ldots, \beta_q, \theta_Z) = \prod_{t=p+1}^{n} \frac{1}{\sigma_t} f_Z\left(\frac{x_t}{\sigma_t}\right) \tag{4.5}$$

where f_Z is the density of the white noise Z and θ_Z is any other parameter Z depends on (such as the degrees of freedom if Z is t-distributed). The likelihood function is a function of the parameters $\alpha_0, \ldots, \alpha_p, \beta_1, \ldots, \beta_q$ via σ_t, which is computed recursively with (4.4), along with supposing for all $t \leq 0$ that

$$\sigma_t = \sqrt{\hat{\sigma}^2},$$
$$X_t = 0,$$

where $\hat{\sigma}^2$ is the sample variance of $\{x_1, x_2, \ldots, x_n\}$. Equivalently, one obtains the parameter estimates $(\hat{\alpha}_0, \ldots, \hat{\alpha}_p, \hat{\beta}_1, \ldots, \hat{\beta}_q, \hat{\theta}_Z)$ as the values that minimize $-\ln L(\alpha_0, \ldots, \alpha_p, \beta_1, \ldots, \beta_q, \theta_Z)$.

The *residuals* for this estimation are the numbers $(x_t/\hat{\sigma}_t, t = p+1, p+2, \ldots, n)$ where $\hat{\sigma}_t$ is the approximated forecasted conditional variance given $(Z_s, s < t)$. These are obtained by recursion using (4.4) with the true parameters replaced by the maximum likelihood estimates with the same initial conditions as in the algorithm above. They should approximately resemble the process Z in distribution. We leave it as an *exercise* to show that for $Z \sim \text{IID}\,\mathcal{N}(0,1)$,

$$-\ln L(\alpha_0, \ldots, \alpha_p, \beta_1, \ldots, \beta_q) = \frac{1}{2} \sum_{t=p+1}^{n} \left(\ln 2\pi + \ln \sigma_t^2 + \frac{X_t^2}{\sigma_t^2}\right).$$

The final method uses the ARMA representation of a GARCH process. It often provides 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 3.2.19 to estimate the parameters of the ARMA representation (4.3) 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

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

If this final method is used, order selection can be done using the same methods as for ARMA processes, using information criteria for example. This can be done for the conditional maximum likelihood method too, but then one should control for the fact that only a subset of observations are used in the likelihood function (4.5). For example, the AICC becomes

Method 4.2.3 (AICC criterion)**.** Choose $p, q, \alpha_0, \ldots, \alpha_p, \beta_1, \ldots, \beta_q$ and θ_Z to minimize

$$-2\frac{n}{n-p} \ln L(\alpha_0, \ldots, \alpha_p, \beta_1, \ldots, \beta_q, \theta_Z) + 2n\frac{m+1}{n-m-2},$$

where n is the sample size and m is the number of non-zero estimated parameters.

4.3 Model building and ARMA – GARCH processes

Sometimes a time series of returns on some asset may display some significant autocorrelations, which is to say it is unlikely that it is white noise. In this case, an ARMA – GARCH-model may be appropriate.

Definition 4.3.1. A time series X is said to be an ARMA – GARCH *process* if it is an ARMA(p, q) process driven by GARCH$(\tilde{p}, \tilde{q})$ noise, i.e., if it is stationary and

$$X_t - \sum_{j=1}^{p} \phi_j X_{t-j} = \sigma_t Z_t + \sum_{j=1}^{q} \theta_j \sigma_{t-j} Z_{t-j}, \tag{4.6}$$

where $Z \sim \text{IID}(0, 1)$, the polynomials $(1 - \sum_{j=1}^{p} \phi_j z^j)$ and $(1 + \sum_{j=1}^{q} \theta_j z^j)$ have no common zeros and

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

with $\alpha_0 > 0$, $\alpha_j \geq 0$ for $j = 1, \ldots, \tilde{p}$, $\beta_i \geq 0$ for $i = 1, \ldots, \tilde{q}$. Furthermore, it is called an ARMA – GARCH *process with mean* μ if $X - \mu$ is an ARMA – GARCH process.

Such a process exists under the same conditions as for a GARCH process, and it is causal if both the ARMA part and GARCH part are causal. Typically, the parameters $(\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q, \alpha_0, \alpha_1, \ldots, \alpha_{\tilde{p}}, \beta_1, \ldots, \beta_{\tilde{q}})$ are estimated using conditional maximum likelihood, as in the previous section. For example, if $Z \sim \text{IID}\,\mathcal{N}(0, 1)$ then, if $\phi(z) \neq 0$ and $\theta(z) \neq 0$ for all $z \in \mathbb{C}$ such that $|z| \leq 1$ and if also $\alpha(1) + \beta(1) < 1$, we have that conditioned on $(X_{t-j}, j \in \mathbb{N})$,

$$X_t \sim \mathcal{N}\left(\sum_{j=1}^{p} \phi_j X_{t-j} + \sum_{j=1}^{q} \theta_j \sigma_{t-j} Z_{t-j}, \alpha_0 + \sum_{j=1}^{\tilde{p}} \alpha_j \sigma_{t-j}^2 Z_{t-j}^2 + \sum_{i=1}^{\tilde{q}} \beta_i \sigma_{t-i}^2 \right),$$

for all $t \in \mathbb{Z}$, and X is said to be a *conditional mean/variance model*. The unknown observations of σ and Z are approximated by recursion with proper initialization, similar to the pure GARCH case, see [11] for further details.

We summarize the essential steps in a suggested model building procedure for ARMA – GARCH models below.

- Remove trend and seasonality until you believe that the data can be modeled as a stationary time series. For the time series of this chapter, this mainly involves removing the sample mean.
- If one determines that the conditional mean is non-zero, i.e., by performing statistical tests and inspecting the sample ACFs and PACFs, fit an ARMA process to the data as described in Chapter 3.
- Determine from the residuals of the ARMA fit if the white noise of the model should be modeled using a GARCH process, i.e., by inspecting the sample ACFs and PACFs of the squared residuals.
- Fit several GARCH-processes to the residuals and compare them using an information criterion.
- Estimate the final full model using conditional maximum likelihood.
- Compute the residuals $\hat{R}_t = (X_t - \hat{X}_t)/\sqrt{\hat{\sigma}_t}$ for this fit and check that they are consistent with the specified distribution and temporal covariance structure for Z_t. If they are, the model is considered to be adequate for the data.

4.4 Extensions

Even though GARCH models are much more flexible than the simple Gaussian model $X_t = \sigma Z_t$, $\sigma \in \mathbb{R}$, there are features of log-returns that GARCH models fail to capture. Extensions have been developed to account for these failures and we end this chapter by mentioning two of them. They are both random variance models, i.e., of the form $X = (X_t, t \in \mathbb{Z})$ with $X_t = \sigma_t Z_t$ for all $t \in \mathbb{Z}$, where $Z \sim \text{IID}(0,1)$ and $\sigma = (\sigma_t, t \in \mathbb{Z})$ is a stochastic process.

In a causal GARCH model, the value of the volatility σ_t depends only on the past values of the squares $(X_{t-j}^2, j \in \mathbb{N})$, which is to say that it does not depend on the *sign* of previous values. However, the volatility of log-returns typically responds differently to positive/negative values of $(X_{t-j}, j \in \mathbb{N})$. The market tends to behave more erratically after a big negative return. The following extension of a GARCH model takes this into account.

Definition 4.4.1. A stochastic process $X = (X_t, t \in \mathbb{Z})$ is called an EGARCH(p,q) *process* if it is stationary and satisfies the EGARCH equations

$$X_t = \sigma_t Z_t,$$

where $Z \sim \text{IID}(0,1)$ has a symmetric distribution, i.e., Z_t and $-Z_t$ have the same distribution, and

$$\ln(\sigma_t^2) = \alpha_0 + \sum_{j=1}^{p} \alpha_j g(Z_{t-j}) + \sum_{i=1}^{q} \beta_i \ln(\sigma_{t-i}^2)$$

with $g(x) = x + \lambda(|x| - \mathbb{E}(|Z_t|))$ and $\alpha_0, \alpha_1, \dots \alpha_p, \beta_1, \dots, \beta_q, \lambda$ being real numbers.

The parameter λ determines how $g(Z_t)$ reacts to different signs of Z_t. A major difference between GARCH and EGARCH processes is that the previous volatility values in the latter model has a *multiplicative* effect on the current volatility as opposed to an additive one, which can be seen from

$$\sigma_t^2 = \exp(\alpha_0) \cdot \prod_{j=1}^{p} \exp(\alpha_j g(Z_{t-j})) \cdot \prod_{i=1}^{q} \sigma_{t-i}^{2\beta_i}.$$

This equation also explains why we do not have any positivity constraints on the parameters of the EGARCH model since the exponential function is always positive.

Similar to the processes above, we can rewrite the volatility equations using polynomials of the backshift operator,

$$\ln(\sigma_t^2) = \alpha_0 + \alpha(B) g(Z_t) + \beta(B) \ln(\sigma_t^2),$$

where

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

Since Z_t and $-Z_t$ are assumed to the same distribution, it holds that $\mathbb{E}(Z_t|Z_t|) = \mathbb{E}(-Z_t|-Z_t|)$ which implies that $\mathbb{E}(Z_t|Z_t|) = -\mathbb{E}(Z_t|Z_t|)$, i.e., that $\mathbb{E}(Z_t|Z_t|) = 0$ for all $t \in \mathbb{Z}$. From that point on it is *an exercise* to show that $g(Z) = (g(Z_{t-1}), t \in \mathbb{Z}) \sim \text{WN}(0, 1 + \lambda^2 \,\mathsf{Var}(|Z_t|))$ and hence that $\ln(\sigma^2)/\alpha_{p'} := (\ln(\sigma_t^2)/\alpha_{p'}, t \in \mathbb{Z})$ is an ARMA$(q, p-p')$ (where p' is the first $j \in \mathbb{N}$ such that $\alpha_j \neq 0$) process with mean $\mu = \alpha_0/(\alpha_{p'}(1-\beta(1)))$.

If $1 - \beta(z) \neq 0$ for all $z \in \mathbb{C}$ such that $|z| = 1$, and $1 - \beta(z)$ and $\alpha(z)$ have no common zeros, then, one can show that $\ln(\sigma^2)$ defined by the equations above is a strictly stationary time series which implies that the same is true for σ^2. If $1 - \beta(z) \neq 0$ for all $z \in \mathbb{C}$ such that $|z| \leq 1$, then $\ln(\sigma^2)$ is also causal and the process X defined by the EGARCH equations is strictly stationary and causal, and therefore exists in the weakly stationary sense if it has finite variance. One can

show that this is true if $Z \sim \text{IID}\,\mathcal{N}(0,1)$. In fact, then all moments of the process exist, which indicates that it is hard for an EGARCH model to capture the "fat tails" of log-returns. Therefore, Z is often assumed to be non-normal. Given the distribution of Z, estimation for EGARCH models is accomplished via conditional maximum likelihood in the same way as for GARCH models.

The final model we consider tries to capture the fact that volatility can display "long memory". This means, for our purposes, that the sample ACF of the squares of the set of log-returns may decay very slowly. This is bad news for us if we want to model the log-returns as a GARCH process, since we have noted that, under suitable conditions, the squares of such a process follow an ARMA process

$$\phi(B)X_t^2 = \alpha_0 + \theta(B)\eta_t,$$

where $\phi(z) = 1 - \alpha(z) - \beta(z)$, $\theta(z) = 1 - \beta(z)$ and $\eta_t = X_t^2 - \sigma_t^2$ is white noise, and the ACF of an ARMA process decays rapidly. In the previous chapter we noted that ARIMA processes, on the other hand, have the property of a very slowly decaying sample ACF. So if we want this property to hold also for $X^2 = (X_t^2, t \in \mathbb{Z})$, we should require that $\phi(B) = \tilde{\phi}(B)(1 - B)$, where $\tilde{\phi}$ is some polynomial such that $\tilde{\phi}(z) \neq 0$ for all $z \in \mathbb{C}$ with $|z| \leq 1$, i.e., that ϕ has a simple root at 1. This, then, motivates the definition of the IGARCH process as a kind of integrated GARCH process.

Definition 4.4.2. A stochastic process $X = (X_t, t \in \mathbb{Z})$ is called an IGARCH(p, q) *process* if it is a *strictly* stationary solution to the GARCH equations

$$X_t = \sigma_t Z_t,$$

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

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

with $\alpha_1, \ldots, \alpha_p, \beta_1, \ldots, \beta_q$ being non-negative numbers, $\alpha_0 > 0$ and $\sum_{i=1}^{p} \alpha_i + \sum_{j=1}^{q} \beta_j = 1$.

It has been shown that a causal strictly stationary IGARCH(p, q) process exists if the distribution of Z_t has unbounded support and no atom at zero (true if $Z \sim \text{IID}\,\mathcal{N}(0,1)$), but then it has infinite variance, which motivates the use of strict instead of weak stationarity in the definition. Despite this fact, it appears to be useful in practice since GARCH models fitted to empirical data often display the unit-root property $\alpha(1) + \beta(1) \approx 1$.

CHAPTER 5

Nonlinear models and nonparametric methods

This chapter is mainly based on [22]. It is important to mention that white noise in [22] is called iid noise in [6] as well as here. The reader should be aware of this when looking for details in [22] and comparing it to the presented content of these lecture notes.

5.1 Introduction to nonlinear models

As has been seen in Chapter 3, a zero mean 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. Let us therefore 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

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

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

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

where $\mathcal{F}_{t-1}$ represents the the information available up to time $t-1$, generated by X, Z and possibly some other independent time series. If g is nonlinear, the model is called *nonlinear in mean*, while it is called *nonlinear in variance* if $h(\mathcal{F}_{t-1})$ is time variant. The models in Chapter 3 are linear. The models introduced in Chapter 4 are nonlinear in variance.

Let us in what follows introduce three examples of nonlinear models, the bilinear model and two switching models. For more examples the reader is referred to [22].

Example 5.1.1. 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, the other parameters are real-valued, and Z is a white noise. This model was introduced by Grander and Andersen [14] and has been widely investigated.

The next two models are switching models – a process which switches between two or more simple models (or *regimes*) depending on a state of an observed or unobserved condition. Such models are not uncommon in finance. In [22], the author considers a model for the monthly employment rate, which takes on two different regimes depending on which side of a given threshold the value of the rate is. This is motivated by regulators being more inclined to take action when the employment rate is low, thus changing the dynamics of the time series.

Example 5.1.2. A time series $X = (X_t, t \in \mathbb{Z})$ follows a *k-regime self-exciting threshold autoregressive (SETAR) model* if it satisfies

$$X_t = c_j + \sum_{i=1}^{p} \phi_{ji} X_{t-i} + Z_{jt} \text{ if } X_{t-d} \in [\gamma_{j-1}, \gamma_j),$$

where $d \in \mathbb{N}$ and the parameters $c_j, \phi_{ji}, \gamma_{j-1}, \gamma_j$ are real valued with $i = 1, \ldots, p$, $j = 1, \ldots, k$ and the series $Z_1, \ldots, Z_k$ with $Z_j = (Z_{jt}, t \in \mathbb{Z})$ are mutually independent IID$(0, \sigma_j^2)$ noises.

There are also threshold generalizations of ARMA and GARCH models which are formulated in the same way. They can have nice qualitative properties when it comes to describing the dynamics of data, but are numerically hard to estimate with potentially unreliable results. One possibility is to use *conditional least squares*, based on minimizing

$$S(\theta) := \sum_{t=p+1}^{n} \left(X_t - \mathbb{E}_\theta(X_t | \mathcal{F}_{t-1})\right)^2,$$

where the conditional expectation is computed with respect to $\theta := c_j, \phi_{ji}, \gamma_{j-1}, \gamma_j$.

The third 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.

Definition 5.1.3. 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

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

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 model below follows two different AR(p) models depending on what state a given Markov chain S is in. The chain S could describe the state of "the market" if X is a model for log-returns of a stock. For example, we could let $S_t = 1$ denote the belief that the market is in a "bear state" and $S_t = 2$ that the market is in a "bull state".

Example 5.1.4. A time series $X = (X_t, t \in \mathbb{Z})$ follows a *Markov switching autoregressive model* (MSA for short) with two states 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 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 time series $Z_1 = (Z_{1t}, t \in \mathbb{Z})$ and $Z_2 = (Z_{2t}, t \in \mathbb{Z})$ are $\text{IID}(0, \sigma^2)$ noises and independent of each other.

Markov switching models are hard to estimate since they depend on an *unobserved* state S. One common approach is to use Markov chain Monte Carlo (MCMC) in a Bayesian framework.

5.2 Nonlinearity tests

The basic idea behind various nonlinearity test is that the residuals of the fit of a given linear model should (more or less) fulfill the assumptions that one makes about the driving noise. The null hypothesis is typically some variant of $H_0 = \{\text{the linear model with the given assumption is adequate}\}$ and the alternative hypothesis vary. The majority of tests only control the risk of doing *type I* errors, so that typically $P(\text{make a type I error}) = P(\text{reject } H_0 | H_0 \text{ is true}) \leq 0.05$. Therefore, there is no formal problem with specifying different alternative hypotheses H_1 in a test. In fact, alternative hypotheses were not present in the original formulation of hypothesis tests. Thinking about null and alternative hypotheses must be made carefully in a time series context. Since we can formulate the null hypothesis as $H_0 = \{\text{the data is generated by the assumed (linear model)}\}$, H_0 will almost always be false since all models are a simplification of reality. But a model can still be useful even if H_0 is false. Moreover, it is not clear what to do if the test rejects H_0. If we have $H_1 = \neg H_0$ all we can do is give up. But typically, the tests are designed to have good power against a *specific* H_1 like $H_1 = \{\text{the driving noise follows a GARCH process}\}$, which gives a guideline about what to do if the test is rejected. The *power* of a test is defined as $P(\text{reject } H_0 | H_1 \text{ is true})$. Therefore, when applying hypothesis tests in this context one has to carefully think about:

- What are the statistical assumptions for the test?
- Does a rejection mean that my model is not useful for the purpose I am working with?
- What aspect of my data lead to the test rejecting the null hypothesis?
- Against what H_1 does my test have good power? What does that suggest that I do next?

In this section we discuss both, nonparametric and parametric statistics that have decent power against some of the models considered in Section 5.1.

5.2.1 Nonparametric tests

The following test has good power for testing ARMA models driven by $Z = \text{IID}\,\mathcal{N}(0, 1)$ versus ARMA models driven by a GARCH process as the noise term Z. It is based on the assumption that if the residuals are IID, the ACF of all powers of them should not be significantly different from zero.

Method 5.2.1 (Q-statistic of squared residuals)**.** This method by McLeod and Li applies the Ljung–Box statistics 2.3.5 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 can be formulated as

$$H_0 : \beta_1 = \cdots = \beta_m = 0,$$

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 $t = m+1, \ldots, n$ with innovations $(e_t, t = m+1, \ldots, n)$.

Method 5.2.2 (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.

$$X_t = \mu + \sum_{i=0}^{\infty} \psi_i Z_{t-i},$$

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 $\psi_0 := 1$ and $\psi_k := 0$ for $k < 0$, the centered third-order moment is defined as

$$c(u, v) := \mathbb{E}\left((X_t - \mu)(X_{t+u} - \mu)(X_{t+v} - \mu)\right) = \mathbb{E}(Z_t^3) \sum_{k=-\infty}^{\infty} \psi_k \psi_{k+u} \psi_{k+v}$$

for $u, v \in \mathbb{Z}$ and arbitrary $t \in \mathbb{Z}$ due to stationary. For frequencies w_1 and w_2 the Fourier transform of c 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 *bispectral 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. It can be shown that for a linear Gaussian time series, the white noise sequence $(Z_t, t \in \mathbb{Z})$ is Gaussian. Since the third moment of a standard normal random variable is zero, $\mathbb{E}(Z_t^3) = 0$ and the bispectrum is zero for all frequencies.

5.2.2 Parametric tests

To conclude the section about nonlinearity tests we introduce one parametric method and remark at the end how this can be extended. A parametric test assumes that the residuals are generated from a *specific* model.

Method 5.2.3 (RESET test)**.** Ramsey [19] proposes a specification test for linear least-squares regression analysis referred to as a *RESET test*, which is readily applicable to linear AR models and which stands for *Regression Equation Specification Error Test*. Therefore consider the linear AR(p) model

$$X_t = \phi_0 + \sum_{j=1}^{p} \phi_j X_{t-j} + Z_t.$$

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

$$\hat{X}_t := \hat{\phi}_0 + \sum_{j=1}^{p} \hat{\phi}_j X_{t-j},$$

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

$$\hat{Z}_t = \alpha_{10} + \sum_{j=1}^{p} \alpha_{1j} X_{t-j} + \sum_{i=1}^{s} \alpha_{2i} \hat{X}_t^{1+i} + V_t$$

for some $s \geq 1$ and innovations $(V_t, t = 1, \ldots, n)$ and compute the least-squares residuals

$$\hat{V}_t = \hat{Z}_t - \left(\hat{\alpha}_{10} + \sum_{j=1}^{p} \hat{\alpha}_{1j} X_{t-j} + \sum_{i=1}^{s} \hat{\alpha}_{2i} \hat{X}_t^{1+i} \right)$$

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.

5.3 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. We will discuss nonparametric methods in a simple autoregressive context where we have a time series $X = (X_t, t \in \mathbb{Z})$ that can be modeled by

$$X_t = m(X_{t-r}) + Z_t, \tag{5.1}$$

where $Z \sim IID(0, \sigma^2)$. However, the method we discuss do not directly use the property that X_t and X_{t-r} are observations from the same time series. Therefore, we generalize and assume that we are given two time series X and Y that are related by

$$Y_t = m(X_t) + Z_t, \tag{5.2}$$

where m is an arbitrary, smooth, but unknown function and $Z \sim IID(0, \sigma^2)$. Our goal is to estimate the nonlinear function m from the data. Assuming that Z and Y are independent, we can view this as attempting to estimate $\mathbb{E}(Y_t|X_t)$. To start, let $X = x$ be constant and independent of Z. 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

$$\hat{m}(x) := \sum_{t=1}^{n} w_t(x) y_t, \tag{5.3}$$

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 5.3.1 (Kernel regression)**.** This method determines the weights by a kernel, which is typically a probability density function denoted by $K : \mathbb{R} \to \mathbb{R}^+$ and which satisfies that it is nonnegative and

$$\int K(z)\, \mathrm{d}z = 1.$$

To increase the flexibility in distance measure, the kernel is often rescaled by the *bandwidth* $h > 0$ and becomes

$$K_h(x) = h^{-1} K(x h^{-1})$$

and

$$\int K_h(z)\, \mathrm{d}z = 1.$$

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 (5.3), 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 [16, 23]). Possible choices of the kernel include the *Gaussian kernel*

$$K_h(x) := (2\pi h^2)^{-1/2} \exp(-(2h^2)^{-1} x^2)$$

and the *Epanechnikov kernel* [9]

$$K_h(x) := 0.75\, h^{-1}(1 - (x/h)^2) I(|x/h| \leq 1),$$

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, for the Epanechikov kernel, $\hat{m}(x_t) \to y_t$ 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 [15] as well as Fan and Yao [10].

Method 5.3.2 (Bandwidth selection with MISE)**.** This method is based on the asymptotic expansion of the *mean integrated squared error* (MISE for short) for kernel smoothers

$$\mathrm{MISE} := \mathbb{E}\left(\int_{-\infty}^{\infty} (\hat{m}(x) - m(x))^2 \, \mathrm{d}x \right),$$

where m is the true function and $\hat{m}$ the estimator which depends on h. Under some regularity conditions on m, one can derive the optimal bandwidth that minimizes the MISE. This will typically depend on several unknown quantities that must be estimated from the data with preliminary smoothing, i.e., computing $\hat{m}$ with a reference bandwith selector. 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. One way of choosing s is to take the sample standard deviation of the data. This works well if the data is normally distributed but might else lead to oversmoothing. For details the reader is referred to [21, Section 3.4.2].

Method 5.3.3 (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 CV is minimized.

Having presented two methods to choose the bandwidth in kernel regression, we note that $\hat{m}(x)$ can equally be defined as the value $\hat{a}$ that minimizes

$$L(a) := \sum_{t=1}^{n} (y_t - a)^2 K_h(x - x_t).$$

This leads a more general method for the estimation of m in Equation (5.2).

Method 5.3.4 (Local linear regression method)**.** Denote the available observations by $((x_t, y_t), t = 1, \ldots, n)$. 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 5.3.2. 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

$$\hat{a} = \frac{\sum_{t=1}^{n} w_t(x) y_t}{\sum_{t=1}^{n} w_t(x)}$$

and

$$\hat{b} = \frac{\sum_{t=1}^{n} \tilde{w}_t(x) y_t}{\sum_{t=1}^{n} w_t(x)},$$

where

$$w_t(x) := K_h(x - x_t)(s_{n,2}(x) - (x - x_t)s_{n,1}(x)),$$
$$\tilde{w}_t(x) := K_h(x - x_t)((x - x_t)s_{n,0}(x) - s_{n,1}(x)),$$

and

$$s_{n,j}(x) := \sum_{t=1}^{n} K_h(x - x_t)(x - x_t)^j$$

for $j = 0, 1, 2$. We leave the derivation as an *exercise* 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 $m(x)$.

One can show that if m in Equation (5.2) is twice continuously differentiable, using local linear regression results in smaller mean squared prediction errors than kernel regression. Let us finally note that more complex autoregressive models like

$$X_t = m_1(X_{t-1}) + m_2(X_{t-2}) + \ldots + m_k(X_{t-k}) + Z_t,$$

where $(m_i)_{i=1}^{k}$ is a sequence of smooth functions, or

$$X_t = m(X_{t-1}, \ldots, X_{t-k}) + Z_t,$$

where $m : \mathbb{R}^k \to \mathbb{R}$ is a smooth function, can be handled by multivariate kernels, such as the multivariate Gaussian density $K_h \colon \mathbb{R}^k \to \mathbb{R}$

$$K_h(\mathbf{x}) = (2\pi h^2)^{-k/2} (\det \Sigma)^{-1/2} \exp(-(2h^2)^{-1} \mathbf{x}' \Sigma^{-1} \mathbf{x}).$$

5.4 Forecasting and evaluation

We have seen in Section 2.4 that forecasting of 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 bootstrapping to compute nonlinear forecasts.

Method 5.4.1 (Parametric bootstrap)**.** Given data $(X_1, X_2, \ldots, X_n)$, we want to forecast X_{n+h} for some $h > 0$. We denote this forecast by $X_n(h)$. The parametric bootstrap computes forecasts of $X_{n+1}, \ldots, X_{n+h}$ sequentially in the following way. For $i = 1, \ldots, h$ repeat:

(i) Generate a random sample of the driving noise at time $n + i$ according to the underlying model.

(ii) Compute $\tilde{X}_{n+i}$ using the generated sample, the model, the data, and the previous forecasts $X_n(1), \ldots, X_n(i-1)$.

(iii) Repeat the previous two steps K times to get K realizations $(\tilde{X}_{n+i}^{(k)}, k = 1, \ldots, K)$. A point forecast for X_{n+i} is then obtained via the sample average $X_n(i) = K^{-1} \sum_{k=1}^{K} \tilde{X}_{n+h}^{(k)}$.

This procedure is also known as *Monte Carlo simulation*. If the model is adequate, $X_n(h) \approx \mathbb{E}(X_{n+h}|X_n, X_{n-1}, \ldots, X_1)$. The samples that make up the average 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 an h-step ahead forecast. Therefore let us do the following: Given a data set $(X_1, X_2, \ldots, X_N)$, we subdivide it into two subsamples: $(X_1, X_2, \ldots, X_n)$, which we refer to as *training subsample* (or *estimation subsample*), and $(X_{n+h}, X_{n+h+2}, \ldots, X_N)$, referred to as the *test subsample* (or *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 test subsample. It is usually appropriate to use a *rolling forecasting procedure*, so that $(X_1, X_2, \ldots, X_n)$ is used to compute $X_n(h)$, $(X_1, X_2, X_3, \ldots, X_{n+1})$ is used to compute $X_{n+1}(h)$ and so on.

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. We denote by $m = N - n - h + 1$ the size of the test subsample. Stationarity (strict or weak) is implicitly assumed.

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 of X_{n+h} in the test 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 $m = N - n = m_{01} + m_{02} = m_{10} + m_{20}$ is the total number of h-step-ahead forecasts in the test 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. The entries in the rightmost column and bottom row are calculated by $m_{i0} = m_{i1} + m_{i2}$ and $m_{0i} = m_{1i} + m_{2i}$ respectively,

where $i \in \{1, 2\}$. 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 mild assumptions, χ^2 has an asymptotic χ^2 distribution with one degree of freedom. Therefore we say that our forecasts are better than random choice at level α if $\chi^2 \geq \chi^2_{1-\alpha}$, where $\chi^2_{1-\alpha}$ is the $1-\alpha$-percentile of the χ^2-distribution with 1 degree of freedom. For example, the forecasts are better than random choice at the 5% level if $\chi^2 \geq \chi^2_{0.95} = 3.841$, which is to be understood in the sense that if we were to predict upward and downward movements in the data using random choice, there would be less than 5% chance of observing $\chi^2 \geq 3.841$. For more details the reader is referred to the literature, especially to [7].

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 data points available in the test subsample. Note that the error computation is done between the data from the test subsample and the h-step-ahead forecasts computed from the model that was derived from the training subsample.

In applications 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.

The following measure is based on the observation that if X is a continuous random variable with cumulative distribution function F_X, then the distribution of the random variable $Y = F_X(X) \sim \mathcal{U}([0, 1])$. This follows by the calculation $P(Y \leq y) = P(F_X(X) \leq y) = P(X \leq F^{-1}(y)) = F(F^{-1}(y)) = y$ for $y \in [0, 1]$. The inverse exists since for continuous random variables X, F_X is strictly increasing.

The measure is different from the ones above in that it does not only measure the forecasting performance of the considered model but also the goodness of fit.

Method 5.4.4 (Distributional measure)**.** For each data point X_{n+h+j}, $j = 0, \ldots, m-1$, in the test subsample, compute the *empirical distribution function* $\hat{F}$

$$\hat{F}_j(y) = \frac{1}{K} \sum_{k=1}^{K} I(\tilde{X}^{(k)}_{n+h+j} \leq y)$$

out of the sample obtained by the parametric bootstrap method 5.4.1 for the h-step-ahead forecast $X_{n+h+j}(h)$. Use the test subsample to compute

$$u_{n+j}(h) := \hat{F}_j(X_{n+h+j})$$

for all $j = 0, \ldots, m-1$, where m denotes the total number of h-step-ahead forecasts in the test 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

$$D = \sup_{x \in [0,1]} \left| \frac{1}{m} \sum_{j=0}^{m-1} I(u_{n+j}(h) \leq x) - x \right|$$

can be used to test the sample with respect to the uniform distribution. The (asymptotic) distribution for this statistic is complicated but if the model is adequate the statistic D should be small. This fact can be used to choose between several models.

Bibliography

[1] Lars V. Ahlfors. *Complex analysis*. McGraw-Hill Book Co., New York, third edition, 1978. An introduction to the theory of analytic functions of one complex variable, International Series in Pure and Applied Mathematics.

[2] Craig F. Ansley. An algorithm for the exact likelihood of a mixed autoregressive-moving average process. *Biometrika*, 66:59–65, 1979.

[3] 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.

[4] Peter J. Brockwell and Richard A. Davis. *Time Series: Theory and Methods*. Springer Series in Statistics. Springer-Verlag, New York, second edition, 1991.

[5] Peter J. Brockwell and Richard A. Davis. *Introduction to Time Series and Forecasting*. Springer Texts in Statistics. New York, NY: Springer, 2nd edition, 2002.

[6] Peter J. Brockwell and Richard A. Davis. *Introduction to Time Series and Forecasting*. Springer Texts in Statistics. New York, NY: Springer, 3rd edition, 2016.

[7] 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.

[8] 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.

[9] V.A. Epanechnikov. Nonparametric estimates of a multivariate probability density. *Theory of Probability and Its Applications*, 14:153–158, 1969.

[10] Jianqing Fan and Qiwei Yao. *Nonlinear Time Series. Nonparametric and Parametric Methods*. New York, NY: Springer, 2003.

[11] Christian Francq and Jean-Michel Zakoian. *GARCH models : structure, statistical inference and finance applications.* Wiley-Blackwell, 2010.

[12] Wayne A. Fuller. *Introduction to Statistical Time Series*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, 1976.

[13] Jan Grandell. Time series analysis. Lecture notes, 2011.

[14] 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.

[15] Wolfgang Härdle. *Applied Nonparametric Regression*, volume 19 of *Econometric Society Monographs*. Cambridge: Cambridge University Press, 1991.

[16] Èlizbar A. Nadaraya. On estimating regression. *Theory and Probability Application*, 10:186–190, 1964.

[17] Jürgen Potthoff. Einführung in die Wahrscheinlichkeitstheorie. Lecture notes for an introductory course in probability theory.

[18] Michael A. Proschan and Pamela A. Shaw. *Essentials of Probability Theory for Statisticians*. Chapman & Hall/CRC Texts in Statistical Science Series. CRC Press, Boca Raton, FL, 2016.

[19] James B. Ramsey. Tests for specification errors in classical linear least-squares regression analysis. *J. R. Stat. Soc., Ser. B*, 31:350–371, 1969.

[20] Sheldon M. Ross. *A First Course in Probability*. Pearson, 9th edition, 2014.

[21] Bernard W. Silverman. *Density Estimation for Statistics and Data Analysis*. Monographs on Statistics and Applied Probability. London - New York: Chapman and Hall, 1986.

[22] Ruey S. Tsay. *Analysis of Financial Time Series*. Wiley Series in Probability and Statistics. Hoboken, NJ: John Wiley & Sons, 3rd edition, 2010.

[23] Geoffrey S. Watson. Smooth regression analysis. *Sankhyā, Ser. A*, 26:359–372, 1964.