Financial Time Series

Annika Lang

version April 15, 2016

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 [5]. For a more extended but still easy introduction to probability theory in English than that given below the reader is referred for example to [6]. We start with the very basic concept of a probability space.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Then with

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

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

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

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

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

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

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

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

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

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

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

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

and for discrete random variables to

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

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

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

while for discrete ones we compute

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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 $X_1, \dots, X_n$ be independent random variables and $g_1, \dots, g_n$ 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.,

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

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).

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

Example 1.1.4 (Bernoulli distribution)**.** The *Bernoulli distribution* is a discrete distribution that takes values in $\{0, 1\}$ and that models a coin flipping experiment. It is characterized by the parameter $p \in (0, 1)$. A Bernoulli distributed random variable X has the distribution

$$P(X = 1) := p, \qquad P(X = 0) := 1 - p.$$

In an easy computation one obtains that

$$\mathbb{E}(X) = p, \qquad \mathsf{Var}(X) = p(1 - p).$$

Example 1.1.5 (Uniform distribution)**.** A random variable X is *uniformly distributed* on the interval $[a, b]$ denoted by $X \sim \mathcal{U}([a, b])$ if it is continuous with density given by

$$f_X(x) := \begin{cases} (b-a)^{-1} & \text{if } x \in [a, b], \\ 0 & \text{else.} \end{cases}$$

It is an easy *exercise* to compute that

$$\mathbb{E}(X) = \frac{a+b}{2}, \qquad \mathsf{Var}(X) = \frac{(b-a)^2}{12}.$$

A useful observation especially for simulations is that if $X \sim \mathcal{U}([0, 1])$, then

$$a + (b - a)X \sim \mathcal{U}([a, b])$$

for real numbers $a < b$.

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

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

The cumulative distribution function of this distribution is usually denoted by

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

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

A central property of the normal distribution is the simple but remarkable fact that the sample or empirical mean $\bar{X}_n$ of a large number of random variables of any distribution will be approximately normally distributed under some simple conditions. This is not proven in these lecture notes but merely stated below. For a proof the reader is referred for example to [6]. The reader should be aware that there exist many versions of this theorem with different assumptions on the underlying random variables. The following is one of the most common with the strongest assumptions.

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

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

1.2 Review on mathematical statistics

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

CHAPTER 2

Stationary time series and seasonality

This chapter is based on Brockwell and Davis' book "Introduction to Time Series and Forecasting" [3] as well as Grandell's lecture notes "Time series analysis" [4], which are based on [3] and [2].

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

2.1 Introduction to time series

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

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

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

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

and for finite observations and models $(x_1, \ldots, x_n)$ and $(X_1, \ldots, X_n)$, 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

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

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

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

Definition 2.1.2. A stochastic process $X = (X_t, t \in \mathbb{T})$ is called *iid noise* with mean zero and variance σ^2 if the sequence of random variables $(X_t, t \in \mathbb{T})$ is independent and identically distributed (abbreviated by *iid*) with $\mathbb{E}(X_t) = 0$ and $\mathsf{Var}(X_t) = \sigma^2$ for all $t \in \mathbb{T}$. An iid noise is denoted by $X \sim \mathrm{IID}(0, \sigma^2)$.

Please note that iid noise is sometimes called *white noise* in the literature (e.g., in [7]). 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 2.1.3 (Binary process)**.** A simple stochastic process and an example of an iid noise is the *binary process* which describes the flipping of a fair coin. In this case $(X_n, n \in \mathbb{N})$ is a sequence of iid random variables characterized by

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

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

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

and variance 1, i.e.,

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

Example 2.1.4 (Random walk)**.** A *random walk* $(S_n, n \in \mathbb{N}_0)$ is obtained by the cumulative summing of iid random variables, i.e., for a given iid noise $(X_n, n \in \mathbb{N})$, it is defined by $S_0 := 0$ and for $n \in \mathbb{N}$ by

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

If the sequence of random variables is given by the binary process in Example 2.1.3, 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 2.1.5. A time series X is said to be a *Gaussian time series* if all finite-dimensional distributions are normal.

2.2 Characterization of stationary time series

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

which makes them symmetric functions around zero.

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

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

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

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

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

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

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

Let us simplify the sum next. It holds that

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

Coming back to our original computation we obtain that

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

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

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

which finishes the proof. □

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

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

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

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

Let us next have a look at the estimation of the autocovariance and autocorrelation function. First of all, it is evident that it is impossible to give reasonable estimates for $\gamma_X(h)$ and $\rho_X(h)$ for $h \geq n$, and even for h near to n the results are not reliable due to few samples. A useful guide can be found in [1], which says that one should take $n \geq 50$ and $h \leq n/4$.

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

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

the k-dimensional *sample covariance matrix*. It is nonnegative definite, which is shown in [3, Section 2.4.2]. The same holds true for the sample autocorrelation matrix $\hat{R}_k$ defined by

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

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

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

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

Method 2.2.7 (Normality)**.** If $(Y_1, \dots, Y_n)$ is a sequence of iid random variables with finite variance, then the sample autocorrelation is for sufficiently large n by the Central Limit Theorem 1.1.7 approximately $\mathcal{N}(0, n^{-1})$ distributed. Hence 95% should fall between the bounds $\pm 1.96/\sqrt{n}$. Use this for hypothesis testing at lag h with

$$\begin{aligned} H_0 &: \rho_Y(h) = 0, \\ H_1 &: \rho_Y(h) \neq 0, \end{aligned}$$

and the test statistic

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

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

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

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

Method 2.2.8 (Portmanteau test, Box–Pierce test)**.** If $(Y_1, \dots, Y_n)$ is a sequence of iid random variables with finite variance, then one can show that $n \sum_{i=1}^{h} \hat{\rho}(i)^2$ is approximately χ^2_h distributed, i.e., chi-squared distributed with h degrees of freedom. Use this for hypothesis testing with

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

and the test statistic

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

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

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

Method 2.2.9 (Ljung–Box test)**.** This test is a modification of the Portmanteau test. Use instead the test statistic

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

which is asymptotically χ^2_h-distributed for iid random variables. Use the same rejection regions as in the Portmanteau test 2.2.8.

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 [3, Section 1.6].

2.3 Forecasting stationary time series

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Putting these results together we have just shown that

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

Therefore we have transformed our minimization problem to

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

Due to the positivity of squares, it is clear that

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

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

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

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

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

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

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

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

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

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

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

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

$$\begin{aligned}\frac{\partial S(a)}{\partial a_j} &= 2a_j\, \mathbb{E}(X^2_{t_{n+1-j}}) \\ &\quad + 2\,\mathbb{E}(X_{t_{n+1-j}}(a_0 + a_1 X_{t_n} + \cdots + a_{j-1} X_{t_{n+2-j}} + a_{j+1} X_{t_{n-j}} + \cdots + a_n X_{t_1} - X_t))\end{aligned} \tag{2.1}$$

as well as

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

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

$$a_0 = \mu \left(1 - \sum_{i=1}^{n} a_i\right). \tag{2.3}$$

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

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

which simplifies with (2.3) to

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

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

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

where

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

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

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

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

Then

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

by (2.2) and

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

for all $j = 1, \ldots, n$ by (2.1), which implies that

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

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

In conclusion we have shown the following proposition:

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

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

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

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

and

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

with

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

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

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

and

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

Bibliography

[1] 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.

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

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

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

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

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

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