Financial Time Series

Annika Lang

2014/15 LP4
draft version June 3, 2015

Preface

These lecture notes were written in parallel to the lecture "Financial Time Series" (TMS087/MSA410) held by the author at Chalmers University of Technology and University of Gothenburg in Spring 2015. They are based on [4, 3, 14, 21].

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, in May 2015
Annika Lang

Contents

1 Stationary time series and seasonality

This chapter is based on Brockwell and Davis' book "Introduction to Time Series and Forecasting" [4] as well as Grandell's lecture notes "Time series analysis" [14], which are based on [4] and [3].

We start with an introduction to times series in general before focusing on stationary time series. We discuss especially the testing and forecasting of stationary times series before we finally introduce methods to remove trend an seasonal components from observed data. For proofs the reader is currently referred to the literature introduced at the beginning of this chapter.

1.1 Introduction to time series

Let from here on $(\Omega, \mathcal{A}, P)$ denote a probability space. In what follows we are concerned with special types of stochastic processes and the corresponding observations that are used to do model fitting, statistics, and forecasting.

Definition 1.1.1. A *time series* is a real-valued series of observations $(x_t, t \in \mathbb{T})$ with respect to an index set $\mathbb{T} \subset \mathbb{R}$.

Definition 1.1.2. 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.

Remark 1.1.3. If no confusion arises, we shall use the term *time series* to mean both the data and the stochastic process which it is a realization of.

The index set $\mathbb{T}$ can in principle be all $\mathbb{R}$ or a subinterval, but in reality we are just able to observe the stochastic process at finitely many times. If $\mathbb{T} = [0, T]$, $T \leq +\infty$, then $(X_t, t \in \mathbb{T})$ is called a *stochastic process in continuous time*. Nevertheless, we will 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. The underlying sequence of random variables is then called a *stochastic process in discrete time*. Therefore 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 $n \in \mathbb{N}$.

In this context a complete probabilistic time series model for the sequence of random variables $(X_n, n \in \mathbb{N})$ specifies the joint distributions of the random vectors $(X_{i_1}, \ldots, X_{i_m})$

for all subsets $\{i_1, \ldots, i_m\}$ of $\mathbb{N}$, $m \in \mathbb{N}$, or equivalently all probabilities

$$P(X_{i_1} \leq y_1, \ldots, X_{i_m} \leq y_m)$$

for all $y_j \in \mathbb{R}$, $j = 1, \ldots, m$, and $m \in \mathbb{N}$.

Let us consider a special time series which leads to easy computations. Please note that the following iid noise is sometimes called white noise (e.g., in [21]), while we will introduce white noise later in these notes as noise with weaker assumptions on the process.

Definition 1.1.4. 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 $X \sim \mathrm{IID}(0, \sigma^2)$.

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

Example 1.1.5 (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 $(X_n, n \in \mathbb{N})$ is a sequence of iid random variables characterized by

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

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

$$\mathbb{E}(X_1) = \int_\Omega X_1(\omega)\, dP(\omega) = 0,$$

and variance 1, i.e.,

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

Example 1.1.6 (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 $(X_n, n \in \mathbb{N})$, 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 1.1.5, the corresponding random walk is called a *simple symmetric random walk*.

1.2 Stationarity and autocorrelation

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 1.2.1. 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 1.2.2. 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.

Condition (ii) in Definition 1.2.2 implies that γ_X is actually a function of the distance 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 1.2.3. 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 $h, t \in \mathbb{Z}$.

Definition 1.2.4. 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. The conclusion in the other direction is just true if the process is centered and uncorrelated implies independence of the random variables. This means that a white noise is an iid noise if $\mu = 0$ and the random variables are independent, e.g., if they are normally distributed.

Definition 1.2.5. A function $\kappa : \mathbb{Z} \to \mathbb{R}$ is said to be *nonnegative definite* or *positive-semi-definite* if for any $n \in \mathbb{N}$

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

for all $(a_1, \ldots, a_n) \in \mathbb{R}^n$ and $(t_1, \ldots, t_n) \in \mathbb{Z}^n$.

Theorem 1.2.6. *A real-valued even function defined on $\mathbb{Z}$ is nonnegative definite if and only if it is the autocovariance function of a stationary time series.*

Definition 1.2.7. A time series X is said to be a *Gaussian time series* if all finite-dimensional distributions are normal.

As mentioned before, observe that a stationary Gaussian time series is also strictly stationary since the normal distribution is determined by its mean and covariance.

1.2.1 Sample mean and sample autocorrelation function

In practice time series analysis is concerned with fitting available data to theoretical models that can then be used for predictions. Since mean and covariance are in general not known, estimators have to be used. In what follows we introduce the concept of sample mean, sample autocovariance function, and sample autocorrelation function as well as their properties.

Definition 1.2.8. 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 $h = -n, \ldots, n$. Furthermore the *sample autocorrelation function* $\hat{\rho}$ is given by

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

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

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.

The following result shows the mean square convergence of the sample mean.

Proposition 1.2.9. *Let X be a stationary time series with mean μ and autocovariance γ_X. Then*

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

if $\lim_{n\to\infty} \gamma(n) = 0$. *Furthermore*

$$\lim_{n\to\infty} n\, \mathbb{E}((\bar{X}_n - \mu)^2) = \sum_{|h|<\infty} \gamma(h)$$

if $\sum_{h=-\infty}^{\infty} |\gamma(h)| < +\infty$.

Furthermore it is clear that for a Gaussian time series X it holds that the sample mean is Gaussian. More specifically, it is clear by the properties of Gaussian random variables 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 denote by

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

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

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

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

It is evident that it is impossible to give reasonable estimates for $\gamma_X(h)$ and $\rho_X(h)$ for $h \geq n$, and even for h near to n the results are not reliable due to few samples. A useful guide can be found in [2], which says that one should take $n \geq 50$ and $h \leq n/4$.

Observe that this function 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.

1.2.2 Testing for autocorrelation

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. We assume in what follows that the reader is familiar with the basics of statistics, especially with hypothesis testing.

Method 1.2.10 (Normality)**.** If $(Y_1, \ldots, Y_n)$ is a sequence of iid random variables with finite variance, then one can show that the sample autocorrelation is 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

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

Method 1.2.11 (Portmanteau test)**.** If $(Y_1, \ldots, Y_n)$ 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

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

and the test statistic

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

Method 1.2.12 (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.

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

1.2.3 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 focus on finding the linear combination of $(X_t, t = 1, \dots, n)$ that forecasts X_{n+h} with minimum mean squared error, i.e.,

$$P_n X_{n+h} = a_0 + a_1 X_n + \cdots + a_n X_1$$

for real coefficients $(a_i, i = 0, \dots, n)$, where $P_n X_{n+h}$ denotes the best linear predictor, such that

$$\mathbb{E}((X_{n+h} - (a_0 + a_1 X_n + \cdots + a_n X_1))^2)$$

is minimized. It can be shown that $P_n X_{n+h}$ exists and is unique. For details the reader is referred to [4, Section 2.5]. We collect the properties of the best prediction in the following proposition. Let us from now on denote the autocovariance function by γ and omit X if there is no reason for confusion.

Proposition 1.2.13. *Let $P_n X_{n+h}$ be the best linear predictor of the stationary time series X with known mean μ and autocovariance function γ. Then it has the following properties:*

(i) $P_n X_{n+h} = \mu + \sum_{i=1}^{n} a_i (X_{n+1-i} - \mu)$*, where the coefficients $(a_1, \dots, a_n)$ are determined by the unique solution of*

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

(ii) $\mathbb{E}((X_{n+h} - P_n X_{n+h})^2) = \gamma(0) - \sum_{i=1}^{n} a_i \gamma(h-1+i)$*;*

(iii) $\mathbb{E}(X_{n+h} - P_n X_{n+h}) = 0$*, i.e., $P_n X_{n+h}$ is unbiased;*

(iv) $\mathbb{E}((X_{n+h} - P_n X_{n+h}) X_j) = 0$ *for all* $j = 1, \dots, n$.

It follows from the proposition that the solution can be computed and is given by

$$a_0 = \mathbb{E}(X_0)\left(1 - \sum_{i=1}^{n} a_i\right)$$

and

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

if the autocovariance matrix is nonsingular and can therefore be inverted.

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

Example 1.2.14 (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 WN$(0, \sigma^2)$ process and $|\phi_1| < 1$. This model will be called an AR(1) model in the framework of Chapter 2. Then we first compute the autocovariance function which is given by

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

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

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

and

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

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

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

The best linear predictor $P_n X_{n+1}$ is then by Proposition 1.2.13 $\sum_{i=1}^n a_i X_{n+1-i}$, 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

$$P_n X_{n+1} = \phi_1 X_n$$

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

Example 1.2.15 (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 2.

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

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

and

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

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

Due to the properties of the white noise, we have that $P_n X_{n+h} = 0$ for all $h > 1$. So all we have to compute is $P_n X_{n+1}$, which satisfies

$$\begin{pmatrix} \gamma_X(0) & \gamma_X(1) \\ \gamma_X(1) & \gamma_X(0) \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} \gamma_X(1) \\ \gamma_X(2) \end{pmatrix}.$$

In other words we have to solve the system of linear equations

$$\begin{aligned} \sigma^2(1 + \theta_1^2)a_1 + \theta_1 \sigma^2 a_2 &= \theta_1 \sigma^2, \\ \theta_1 \sigma^2 a_1 + \sigma^2(1 + \theta_1^2)a_2 &= 0. \end{aligned}$$

As solution we obtain

$$a_1 = \frac{\theta_1(1 + \theta_1^2)}{1 + \theta_1^2 + \theta_1^4} \quad \text{and} \quad a_2 = -\frac{\theta_1^2}{1 + \theta_1^2 + \theta_1^4}.$$

This leads to the best linear predictor

$$P_n X_{n+1} = a_1 X_n + a_2 X_{n-1} = \frac{\theta_1(1 + \theta_1^2)}{1 + \theta_1^2 + \theta_1^4} X_n - \frac{\theta_1^2}{1 + \theta_1^2 + \theta_1^4} X_{n-1}$$

and finishes the example.

More generally we introduce the prediction operator $P(\cdot|W)$ in what follows. In this context we have seen before how to compute the best linear predictor $P(Y|W)$ of Y in terms of $W = (W_n, \ldots, W_1)$ and 1, where $W = (X_n, \ldots, X_1)$ so far. The function $P(\cdot|W)$ which converts Y into $P(Y|W)$ is called the *prediction operator*. Instead of using $W = (X_n, \ldots, X_1)$, one can use other input, e.g., to estimate a missing value in a time series. Prediction operators have useful properties which are collected in the following proposition.

Proposition 1.2.16. *Let the random vector $W = (W_n, \ldots, W_1)$ be given and set $\Gamma := \mathsf{Cov}(W, W)$. Let U and V be two random variables with finite second moment and $\beta, \alpha_1, \ldots, \alpha_n \in \mathbb{R}$. Then the prediction operator $P(\cdot|W)$ has the following properties:*

(i) $P(U|W) = \mathbb{E}(U) + a(W' - \mathbb{E}(W'))$, *where* $\Gamma a' = \mathsf{Cov}(U, W)'$;

(ii) $\mathbb{E}((U - P(U|W))W) = 0$ *and* $\mathbb{E}(U - P(U|W)) = 0$, *i.e., the prediction operator is orthogonal and unbiased;*

(iii) *the mean squared error is given by* $\mathbb{E}((U - P(U|W))^2) = \mathsf{Var}(U) - a\,\mathsf{Cov}(U, W)$;

(iv) $P(\alpha_1 U + \alpha_2 V + \beta|W) = \alpha_1 P(U|W) + \alpha_2 P(V|W) + \beta$, *i.e., the prediction operator is linear;*

(v) $P(\sum_{i=1}^n \alpha_i W_i + \beta|W) = \sum_{i=1}^n \alpha_i W_i + \beta$;

(vi) $P(U|W) = \mathbb{E}(U)$ *if* $\mathsf{Cov}(U, W) = 0$;

(vii) $P(U|W) = P(P(U|W, Y)|W)$, *if* $Y = (Y_1, \ldots, Y_n)$ *is such that all components of* $\mathbb{E}(Y'Y)$ *are finite.*

Example 1.2.17 (AR(1) with missing value)**.** Let us consider the AR(1) model from Example 1.2.14 again. Assume that we have observed X_1 and X_3 but that we are missing X_2. Then the best predictor $P(X_2|X_3, X_1)$ of $X_@$ is $a_1 X_3 + a_2 X_1$ by Proposition 1.2.16, 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

$$P(X_2|X_3, X_1) = \frac{\phi_1}{1 + \phi_1^2}(X_3 + X_1)$$

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

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

We have stated in Proposition 1.2.13 and Proposition 1.2.16 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., $P_n X_{n+1}$ is used to compute $P_{n+1} X_{n+2}$ in a cheaper way.

To make the algorithm readable we have to adapt the notation for changing sizes of matrices. In what follows let the best linear one-step estimator be given by

$$P_n X_{n+1} = \sum_{i=1}^n \phi_{ni} X_{n+1-i},$$

where $\phi_{ni} := a_i$ 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 1.2.13.

One way to compute the estimators more efficiently all in parallel is the Durbin–Levinson algorithm, which is presented next.

Method 1.2.18 (Durbin–Levinson algorithm)**.** Compute the coefficients $\phi_{n1}, \ldots, \phi_{nn}$ recursively from the equations

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

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

and

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

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

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

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

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

$$\hat{X}_n := \begin{cases} 0 & \text{for } n = 1, \\ P_{n-1} X_n & \text{for } n > 1, \end{cases}$$

and the mean squared errors

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

One can show that the best linear predictors satisfy

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

for some coefficients θ_{ij}, $i, j = 1, \ldots, n$. The innovations algorithm generates these coefficients and the mean squared errors $v_j = \mathbb{E}((X_{j+1} - \hat{X}_{j+1})^2)$ recursively.

Method 1.2.19 (Innovations algorithm)**.** Compute the coefficients $\theta_{n1}, \ldots, \theta_{nn}$ recursively from the equations

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

and

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

for $0 \leq k < n$ and

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

We finish this section with a result on the decomposition of a stationary time series which is called the *Wold decomposition*. It states how a nondeterministic time series can be decomposed into a deterministic one and a white noise sum. To make this more precise, we need to define what we mean by a deterministic time series and talk about an infinite past.

To allow for linear predictors that are not forced to depend on $(X_n, \dots, X_1)$, let us denote by $P_{m,n}X_{n+h}$ the best linear predictor for the observations of $X_m, \dots, X_0, X_1, \dots, X_n$ for $m < 0$ and $n > 0$. For large $|m|$, we can approximate the mean square limit which we denote by

$$\tilde{P}_n X_{n+h} := \lim_{m \to -\infty} P_{m,n} X_{n+h}$$

and which we call the *prediction operator based on the infinite past*. We collect some properties of $\tilde{P}_n$ in the following lemma.

Lemma 1.2.20. *Let U and V be random variables with finite second moment and $a, b, c \in \mathbb{R}$. The prediction operator $\tilde{P}_n$ based on the infinite past satisfies the following properties:*

(i) the error is orthogonal to the past, i.e., $\mathbb{E}((U - \tilde{P}_n(U))X_j) = 0$, $j \leq n$;

(ii) $\tilde{P}_n(aU + bV + c) = a\tilde{P}_n(U) + b\tilde{P}_n(V) + c$, i.e., it is linear;

(iii) $\tilde{P}_n(U) = U$ if U is a limit of linear combinations of $(X_j, j \leq n)$;

(iv) $\tilde{P}_n(U) = \mathbb{E}(U)$ if $\mathsf{Cov}(U, X_j) = 0$ for all $j \leq n$.

Furthermore we have to define what we mean by a deterministic and nondeterministic time series.

Definition 1.2.21. A stationary time series $X = (X_t, t \in \mathbb{Z})$ is called *deterministic* if $X_n - \tilde{P}_{n-1}X_n = 0$ for all $n \in \mathbb{Z}$. Otherwise it is called *nondeterministic*.

This enables us finally to state the Wold decomposition theorem.

Theorem 1.2.22 (Wold decomposition)**.** *If X is a nondeterministic stationary time series, then*

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

for all $t \in \mathbb{Z}$, where

(i) $\psi_0 = 1$ and $\sum_{j=0}^{\infty} \psi_j^2 < +\infty$,

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

(iii) $\mathsf{Cov}(Z_s, V_t) = 0$ for all $s, t \in \mathbb{Z}$,

(iv) $Z_t = \tilde{P}_t Z_t$ for all $t \in \mathbb{Z}$,

(v) $V_t = \tilde{P}_s V_t$ for all $s, t \in \mathbb{Z}$,

(vi) $V = (V_t, t \in \mathbb{Z})$ is deterministic.

One derives from the theorem that Z, V, and $(\psi_j, j \in \mathbb{N}_0)$ are unique and can be written explicitly as

$$Z_t = X_t - \tilde{P}_{t-1}X_t$$

and

$$\psi_j = \frac{\mathbb{E}(X_t Z_{t-j})}{\mathbb{E}(Z_t^2)},$$

which can be plugged into

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

For details the reader is referred to [3, p.188].

1.3 Trend and seasonality

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

$$X_t = m_t + s_t + Y_t, \tag{1.1}$$

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

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

1.3.1 Trend in absence of seasonality

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

$$X_t = m_t + Y_t,$$

where we assume without loss of generality that $\mathbb{E}(Y_t) = 0$ for all $t \in \mathbb{Z}$.

In what follows we first introduce methods for trend estimation before we give a method that does trend elimination directly without estimation.

Method 1.3.1 (Estimation by smoothing with a finite moving average filter)**.** Consider for $q \in \mathbb{N}$ and $q + 1 \leq t \leq n - q$ the *two-sided moving average*

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

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

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

if we assume that q is sufficiently small that $(m_s, s = t-q, \ldots, t+q)$ is approximately linear and that the average of 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 does not lead to estimates for all $t = 1, \ldots, n$ depending on the choice of q.

Method 1.3.2 (Estimation by exponential smoothing)**.** For any fixed $\alpha \in [0,1]$ define the *one-sided moving averages* $(\hat{m}_t, t = 1, \ldots, n)$ by the recursion

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

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

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

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

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

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

Method 1.3.3 (Estimation by polynomial fitting)**.** Assume that the trend m is given by the polynomial

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

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

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

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

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

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

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

Method 1.3.4 (Elimination by differencing)**.** Define the *difference operator* ∇ by

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

for $t \geq 2$, where B denotes the *ward 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 have

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

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

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

Assume that the trend m is given by the polynomial

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

Then one shows in an *exercise* that

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

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

Coming back to the stochastic process X, one obtains that

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

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

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

1.3.2 Trend and seasonality in parallel

Let us come back to the classical decomposition model (1.1). We recall that

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

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

Method 1.3.5 ("S1": estimation by moving averages)**.** Let us assume that we are given observations $(x_t, t = 1, \dots, n)$ of the stochastic process X. We start with the estimation of the trend by applying a moving average filter that eliminates the seasonal component and dampens the noise. For an even period $d := 2q$, we set

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

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

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

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

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

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

Finally, we reestimate the trend by applying one of the methods from Section 1.3.1 to the deseasonalized series $(x_t - \hat{s}_t, t = 1, \dots, n)$. The reestimation of the trend is done to have a parametric form for the trend that can be used for prediction and simulation.

Method 1.3.6 ("S2": elimination by differencing)**.** The method of differencing that was introduced in Method 1.3.4 for time series without seasonality can be adapted to the general classical decomposition model by introducing the *lag-d differencing operator* ∇_d for a period d. It is defined by

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

Applying this operator to the model, we obtain that

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

due to the periodicity of s. Therefore $\nabla_d X$ is a stochastic process without seasonal component and the methods introduced in Section 1.3.1 can be applied to estimate the trend component.

Once we have removed the trend and the seasonal components from the time series, we have to test the remaining sequence for stationarity with the methods from Section 1.2.2. This helps us to check if the assumed model assumptions, e.g., on the period d, are reasonable.

2 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.

2.1 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 remark that we use stochastic process and time series from now on as synonyms.

Definition 2.1.1. A stochastic process X is called a *linear process* if it has the representation

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

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

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

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

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

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

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

where we define the operator

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

This is used in the following proposition to characterize the properties of X.

Proposition 2.1.2. *Let Y be a mean-zero, stationary time series with 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*

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

is stationary with mean zero and autocovariance function γ_X given by

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

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

In the special case that X is a linear process with variance σ^2 of the corresponding white noise, it holds that the autocovariance function γ_X is given by

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

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

2.2 ARMA models

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

Definition 2.2.1. 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 2.2.2. 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 2.2.3. A time series X is an ARMA(p,q) *process* if X is stationary and if for all $t \in \mathbb{Z}$

$$X_t - \sum_{j=1}^{p} \phi_j X_{t-j} = Z_t + \sum_{j=1}^{q} \theta_j Z_{t-j}, \tag{2.1}$$

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

To simplify the notation, we set

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

and

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

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

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

where we recall that B denotes the backward shift operator.

Proposition 2.2.4 (Existence and uniqueness)**.** *A stationary solution X of Equation* (2.1) *exists and is the unique stationary solution if and only if*

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

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

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

Definition 2.2.5. An ARMA(p,q) process X is *causal* or a *causal function of Z* if there exists a sequence of constants $(\psi_j, j \in \mathbb{N}_0)$ 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 process.

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

Lemma 2.2.6. *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* $|z| \leq 1$.

Together with Proposition 2.2.4, the lemma implies the following corollary.

Corollary 2.2.7. *A causal* ARMA(p,q) *process has a unique stationary solution.*

Definition 2.2.8. An ARMA(p,q) process X is *invertible* if there exists a sequence of constants $(\pi_j, j \in \mathbb{N}_0)$ 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 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 frequently checked.

Lemma 2.2.9. *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* $|z| \leq 1$.

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 $j < 0$.

2.2.1 Autocorrelation and partial autocorrelation function

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

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

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

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

as well as

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

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

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

where the coefficients ψ_j are determined by the expansion

$$\frac{\theta(z)}{\phi(z)} = \sum_{j=0}^{\infty} \psi_j z^j$$

for $|z| \leq 1$.

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

Method 2.2.10. Proposition 2.1.2 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 $(\psi_j, j \in \mathbb{N}_0)$ are computed by the methods introduced in Section 2.2.2.

Method 2.2.11. If we multiply each side of the equations

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

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

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

for $0 \leq k < m$ and

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

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

Method 2.2.12. This method is a numerical version of Method 2.2.11. Use the equations there for the first $p+1$ equations and solve them numerically to get $\gamma(0), \ldots, \gamma(p)$. Afterwards, use the result and determine successively $\gamma(j)$ for $j > p$.

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

Definition 2.2.13. 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

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

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

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

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

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

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.

2.2.2 Parameter estimation

Let us assume in this section that the order parameters p and q of an ARMA(p,q) model are known, which is not true in most realistic cases. We will discuss the order selection in Section 2.2.3. Here we will give methods to determine the parameters $(\phi_j, j = 1, \ldots, p)$ and $(\theta_j, j = 1, \ldots, q)$.

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

First we introduce the Yule–Walker estimation, which can be derived from Method 2.2.11. We observe that the equations in Method 2.2.11 simplify for an AR(p) model to

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

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

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

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 on 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 2.2.14 (Yule–Walker estimation)**.** Compute estimators $\hat{\sigma}^2$ and $(\hat{\phi}_j, j = 1, \dots, p)$ from the equations

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

and

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

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

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

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

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

of the preceding i observations, which are called the *forward prediction errors*. Similarly for $0 \leq i < n$ let $(v_i(t), t = i+1, \ldots, n)$ be the differences between x_{n+1-t} and its best linear estimate in terms of the subsequent i observations, which are referred to *backward prediction errors*. In an *exercise* one shows that the forward and backward prediction errors satisfy the recursions

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

The transformation of these recursions leads to the following algorithm, which computes estimators for σ^2 and ϕ_{ii}. The remaining $(\phi_{ij}^{(B)}, j < i)$ can be obtained by the Durbin–Levinson algorithm 1.2.18, where (B) indicates that the estimators are computed using Burg's algorithm.

Method 2.2.15 (Burg's algorithm)**.**

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

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

Similarly to the application of the Durbin–Levinson algorithm 1.2.18 to fit AR models, we can use the innovations algorithm 1.2.19 to fit MA models

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

of orders $m \in \mathbb{N}$, where $Z \sim \mathrm{WN}(0, \hat{v}_m)$.

Method 2.2.16 (Innovations algorithm)**.** Apply the innovations algorithm 1.2.19 with the sample autocovariance function instead of the autocovariance function to obtain the coefficients of the *fitted innovations* MA(m) *model*

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

where $Z \sim \mathrm{WN}(0, \hat{v}_m)$.

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

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

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

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

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

with $\theta_0 := 1$ and $\theta_j := 0$ for $j > q$. To estimate $(\psi_j, j = 1, \ldots, p+q)$ we can use the innovation estimates $\hat{\theta}_{m1}, \ldots, \hat{\theta}_{m(p+q)}$ whose large-sample behavior is (under smoothness assumptions, for details see [4, Remark 1]) that for any positive integer k the joint distribution function of

$$\sqrt{n}(\hat{\theta}_{m1} - \theta_1, \hat{\theta}_{m2} - \theta_2, \ldots, \hat{\theta}_{mk} - \theta_k)$$

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

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

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

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

Method 2.2.17 (Hannan–Rissanen algorithm)**.**

Step 1 Fit a high-order AR(m) model (with $m > \max\{p, q\}$) to the data using the Yule–Walker estimates from Method 2.2.14. For estimated coefficients $(\hat{\phi}_{m1}, \ldots, \hat{\phi}_{mm})$, compute the estimated residuals from the equations

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

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

Step 2 Estimate the vector of parameters $\beta = (\phi_j, \theta_i, j = 1, \ldots, p, i = 1, \ldots, q)$ 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*

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

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 estimate of the white noise variance is

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

Let us assume for the next method that we are given a Gaussian ARMA(p, q) process (or at least act as if). Then for any fixed values $(\phi_1, \dots, \phi_p)$, $(\theta_1, \dots, \theta_q)$, and σ^2, the random variables $X_1 - \hat{X}_1, \dots, X_n - \hat{X}_n$ are independent and normally distributed, where $\hat{X}_j := \mathbb{E}(X_j | X_1, \dots, X_{j-1}) = P_{j-1}X_j$, $j \geq 2$. Let Γ_n denote the covariance matrix of $(X_1, \dots, X_n)$ and assume that it is nonsingular. The likelihood of $(X_1, \dots, X_n)$ is

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

Assuming that the process is Gaussian, which also makes kind of sense for other processes if large sample sizes are used, one derives the *Gaussian likelihood* for an ARMA process

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

where

$$r_j := \frac{\mathsf{Var}(X_j - \hat{X}_j)}{\sigma^2},$$

which can be determined by the innovations algorithm 1.2.19.

Method 2.2.18 (Maximum likelihood estimators)**.** The maximum likelihood estimators of σ^2, ϕ, and θ are determined from the expression

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

where

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

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

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

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

The derivation of the equations is left as an *exercise* to the reader.

Method 2.2.19 (Least squares estimation of mixed models)**.** Minimize S instead of ℓ in Method 2.2.18 to obtain the least squares estimates $\tilde{\phi}$ and $\tilde{\theta}$. The least squares estimate of σ^2 is

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

Having fitted the model, it remains to check that the model was chosen adequately. If this is the case, the residuals should behave like white noise.

2.2.3 Order selection

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

In general a selection for an AR(p) or MA(q) model may be made using autocorrelation and partial autocorrelation functions. Typically, an autocorrelation function with "q peaks and then zero" indicates a MA(q) model, while a slowly decaying autocorrelation function and a partial autocorrelation function with "p peaks and then zero" indicates an AR(p) model. After parameter estimation, which can be done by the procedures from Section 2.2.2, the model should be checked if the obtained residuals behave like white noise by the methods introduced in Section 1.2.2.

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

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

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

Method 2.2.20 (AICC criterion)**.** Choose p, q, ϕ_p, and θ_q to minimize

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

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

In general one may say that order selection is a difficult problem and many criteria have been proposed. Rissanen's minimum description length (MDL) criterion seems to be rather much used according to [14].

2.2.4 Forecasting of ARMA processes

The innovations algorithm 1.2.19 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

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

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

The autocovariance function γ_X of X can easily be computed using any method of Section 2.2.1. The autocovariances $\kappa(i,j) = \mathbb{E}(W_i W_j)$, $i, j \geq 1$, are then found from

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

Applying the innovations algorithm 1.2.19 to the process W we obtain

$$\hat{W}_{n+1} = \sum_{j=1}^{n}\theta_{nj}(W_{n+1-j} - \hat{W}_{n+1-j})$$

for $q \leq n < m$ and

$$\hat{W}_{n+1} = \sum_{j=1}^{q}\theta_{nj}(W_{n+1-j} - \hat{W}_{n+1-j})$$

for $n \geq m$, where the coefficients $(\theta_{nj}, n \in \mathbb{N}, j \leq \min\{n, m\})$ and the mean squared errors $\mathbb{E}((W_{n+1} - \hat{W}_{n+1})^2)$ are found recursively from the innovations algorithm 1.2.19 with κ defined above. The notable feature of the predictors $(\hat{W}_{n+1}, n \in \mathbb{N})$ is the vanishing of θ_{nj} when both $n \geq m$ and $j > q$.

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

$$\hat{W}_{n+1} = P_n W_{n+1}.$$

Furthermore we obtain that

$$\hat{W}_t = \sigma^{-1} \hat{X}_t$$

for $t = 1, \ldots, m$ and

$$\hat{W}_t = \sigma^{-1}(\hat{X}_t - \sum_{j=1}^{p} \phi_j X_{t-j})$$

for $t > m$ due to the linearity of P_n. So we obtain as best linear estimator for X_{n+1}

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

with mean squared error

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

where the coefficients $(\theta_{nj}, n \in \mathbb{N}, j \leq \min\{n, m\})$ and the mean squared errors $\mathbb{E}((W_{n+1} - \hat{W}_{n+1})^2)$ are found recursively from the innovations algorithm 1.2.19. The best linear estimators can be computed recursively.

Let us next consider h-step predictors of an ARMA(p, q) process. Therefore we recall from Section 1.2.3 that

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

We conclude with the properties of W that the h-step predictors $P_n X_{n+h}$ satisfy

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

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

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

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

The mean squared error of $P_n X_{n+h}$ is computed from the formula

$$\sigma_n^2(h) := \mathbb{E}((X_{n+h} - P_n X_{n+h})^2) = \sum_{j=0}^{h-1} \left(\sum_{r=0}^{j} \chi_r \theta_{(n+h-r-1)(j-r)} \right)^2 v_{n+h-j-1},$$

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

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

for $j \in \mathbb{N}$, and the $v_{n+h-j-1}$ denote the mean squared errors of the one-step predictors as introduced in Section 1.2.3.

Assuming—as usual—that the ARMA(p, q) process defined by

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

where $Z \sim \mathrm{WN}(0, \sigma^2)$, is causal and invertible, we have the representations

$$X_{n+h} = \sum_{j=0}^{\infty} \psi_j Z_{n+h-j}$$

and

$$Z_{n+h} = X_{n+h} + \sum_{j=1}^{\infty} \pi_j X_{n+h-j},$$

where the sequences $(\psi_j, j \in \mathbb{N}_0)$ and $(\pi_j, j \in \mathbb{N})$ are uniquely determined as discussed previously. Let $\tilde{P}_n$ denote the best linear approximation as in Section 1.2.3, then the application of $\tilde{P}_n$ to each side of the above equations yields

$$\tilde{P}_n X_{n+h} = \sum_{j=h}^{\infty} \psi_j Z_{n+h-j}$$

and

$$\tilde{P}_n X_{n+h} = -\sum_{j=1}^{\infty} \pi_j \tilde{P}_n X_{n+h-j}.$$

For $h = 1$ this yields a formula that enables us to compute the predictors successively from one-step predictors. One obtains the h-step prediction error

$$X_{n+h} - \tilde{P}_n X_{n+h} = \sum_{j=0}^{h-1} \psi_j Z_{n+h-j},$$

from which one sees that the mean squared error is

$$\tilde{\sigma}^2(h) = \sigma^2 \sum_{j=0}^{h-1} \psi_j^2.$$

The predictors obtained in this way have the form

$$\tilde{P}_n X_{n+h} = \sum_{j=0}^{\infty} c_j X_{n-j}.$$

Due to just finitely many observations, the series has to be truncated in practice after n terms. The resulting predictor is a useful approximation to $P_n X_{n+h}$ if n is large and the coefficients $(c_j, j \in \mathbb{N}_0)$ converge to zero. Further one can show that $\tilde{\sigma}^2(h)$ is an easily calculated approximation to $\sigma_n^2(h)$ for large n.

Finally in this section we remark that in the special case that the ARMA process is driven by Gaussian white noise, i.e., $Z \sim \mathrm{IID}\mathcal{N}(0, \sigma^2)$, for each $h \geq 1$ the prediction error $X_{n+h} - P_n X_{n+h}$ is normally distributed with mean zero and variance $\sigma_n^2(h)$. 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.

2.3 ARIMA models

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

Definition 2.3.1. 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.

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

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

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

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

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

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

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

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

with d roots on the unit circle. A more systematic approach is due to Dickey and Fuller (see [7]) and described in what follows.

Assume that $(x_1,\ldots,x_n)$ are observations from the AR(1) model

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

where $Z\sim \mathrm{WN}(0,\sigma^2)$, $|\phi_1|<1$, and $\mu:=\mathbb{E}(X_t)$. Since the normal hypothesis test fails, 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}. The estimated standard error of $\hat{\phi}_1^*$ is

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

where

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

and $\bar{X}_n$ is 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 $H_0 : \phi_1 = 1$ 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 the 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 as an *exercise* to the reader.

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

Method 2.3.2 (Dickey–Fuller test)**.** Estimate ϕ_1^* as the coefficient of X_{t-1} in the OLS regression of ∇X_t onto 1, X_{t-1}, ∇X_{t-1}, $\ldots$, ∇X_{t-1+p}. For large n the t-ratio

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

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

Testing the moving-average polynomial for unit roots is equivalent to testing that the time series has been overdifferenced.

3 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 the explicit solution

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

Nevertheless, this does not seem to be the case in realistic models. One option to measure the volatility is the *realized volatility*, which is computed by

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

for observed data $(x_j, j = 1, \ldots, n)$, 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.

3.1 Definitions and properties

Definition 3.1.1. A stochastic process $X = (X_t, t \in \mathbb{Z})$ is called an ARCH(p) *process* if it is stationary and if

$$X_t = \sigma_t Z_t,$$

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

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

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

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

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

Consider now an ARCH(p) process and the polynomial

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

Thus we can rewrite the equation of the volatility σ_t to

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

where we recall that B denotes the backward shift operator introduced in Chapter 1. Due to stationarity and the fact that $\mathbb{E}(X_t^2) = \mathbb{E}(\sigma_t^2)$, which can be shown in an easy *exercise*, it holds that

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

This implies that

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

It can be shown that $(X_t^2, t \in \mathbb{Z})$ is an AR process (see, e.g. [14]).

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

$$X_t = \sigma_t Z_t,$$

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

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

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

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

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

where

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

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

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

which implies that

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

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

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

and

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

as well as mean $\alpha_0(1 - \alpha(1) - \beta(1))^{-1}$ (see, e.g. [14]). The ARMA model can be represented by

$$X_t^2 = \alpha_0 + \sum_{i=1}^{\max\{p,q\}} (\alpha_i + \beta_i) X_{t-i}^2 + \eta_t - \sum_{j=1}^{q} \beta_j \eta_{t-j}, \tag{3.1}$$

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

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

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

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

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

We close this section by remarking that uncertainty in volatility estimation is an important issue that is often overlooked. To assess the variability of an estimated volatility, it is necessary to consider the kurtosis of a volatility model (cf. [21, Section 3.16]). For the GARCH$(1,1)$ model it is given by

$$\frac{3(1 - (\alpha + \beta)^2)}{1 - \beta^2 - 2\alpha\beta - 3\alpha^2} > 3,$$

where the lower bound of three can be easily computed.

3.2 Estimation

Let us introduce two methods in this section to do estimation for ARCH and GARCH models. While the first one tests given data for ARCH effects, the second uses the ARMA representation to estimate the parameters α_i and β_i. For more methods the reader is referred to [21].

Method 3.2.1 (Test for ARCH effects)**.** Test the null hypothesis $H_0 : \alpha_1 = \cdots = \alpha_p = 0$.

Set

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

where

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

and

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

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

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

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

Then the test statistic for ARCH effects is

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

which is asymptotically χ_p^2 distributed.

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

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

which is maximized numerically.

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

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

Method 3.2.2 (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 2.2.18 to estimate the parameters of the ARMA representation (3.1) for $(x_j^2, j = 1, \ldots, n)$, denoted by $\hat{\phi}_i$ and $\hat{\theta}_i$. Obtain the parameter estimates of the GARCH coefficients by setting

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

3.3 Extensions

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

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

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

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

4 Nonlinear models

This section is mainly based on [21]. It is important to mention that white noise in [21] is called iid noise in [4] as well as here. The reader should be aware of this when looking for details in [21] and comparing it to the presented content of these lecture notes.

As seen in Chapter 2, a centered linear model can be expressed by

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

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

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

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

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

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

and the conditional variance σ_t^2 by

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

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

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

where $\epsilon_t = Z_t/\sigma_t$ is a standardized shock (or innovation). Furthermore assume for simplicity that Z is iid noise, i.e., $\epsilon \sim \mathrm{IID}(0, 1)$. If g is nonlinear, the model is called *nonlinear in mean*, while it is called *nonlinear in variance* if h is time variant. The models in Chapter 2 are linear. One can show that those introduced in Chapter 3 are nonlinear in variance.

4.1 Nonlinear models

In this section we introduce the bilinear and the Markov switching model as two examples of nonlinear models. For more examples the reader is for the moment is referred to [21].

4.1.1 Bilinear model

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, ad s are nonnegative integers and Z is a white noise. This model was introduced by Grander and Andersen [15] and has been widely investigated.

4.1.2 Markov switching model

Definition 4.1.1. A stochastic process X is a *Markov process* if its conditional distribution function satisfies

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

for arbitrary $h > t$. If X is a discrete-time stochastic process, 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*.

Definition 4.1.2. A time series X follows a *Markov switching autoregressive model* (MSA for short) if it satisfies

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

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

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

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

4.2 Nonparametric methods for model fitting

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

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

$$Y_t = m(X_t) + Z_t, \tag{4.1}$$

where m is an arbitrary, smooth, but unknown function and Z is iid noise. Our goal is to estimate the nonlinear function m from the data. Let for the beginning $X = x$ be constant and $(y_t, t = 1, \ldots, n)$ be a series of independent observations. Then the problem simplifies to

$$y_t = m(x) + Z_t$$

and taking the sample average yields

$$n^{-1} \sum_{t=1}^{n} y_t = m(x) + n^{-1} \sum_{t=1}^{n} Z_t.$$

By the properties of the iid noise and the law of large numbers, the averaged noise converges to zero for large n. Therefore

$$\bar{y} := n^{-1} \sum_{t=1}^{n} y_t$$

is a consistent estimator for $m(x)$, i.e., $\bar{y} \approx m(x)$.

As long as m is sufficiently smooth and $X_t \approx x$ still almost constant, the method continues to work fine. In other cases one possibility is to use a weighted average of y instead of the simple one, which we denote by

$$\hat{m}(x) := n^{-1} \sum_{t=1}^{n} w_t(x) y_t, \tag{4.2}$$

where the weights $w_t(x)$ are larger for those y_t with x_t close to x and smaller for those far away.

We introduce two methods to determine the weights in what follows.

Method 4.2.1 (Kernel regression)**.** This method determines the weights by a kernel, which is typically a probability density function denoted by K and which satisfies that it is nonnegative and

$$\int K(z)\, dz = 1.$$

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

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

and

$$\int K_h(z)\,dz = 1.$$

Define the weight function by

$$w_t(x) := \frac{K_h(x - x_t)}{\sum_{s=1}^{n} K_h(x - x_s)}.$$

Plugging this into Equation (4.2), the *Nadaraya–Watson kernel estimator*

$$\hat{m}(x) = \sum_{t=1}^{n} w_t(x) y_t = \frac{\sum_{t=1}^{n} K_h(x - x_t) y_t}{\sum_{t=1}^{n} K_h(x - x_t)}$$

is obtained (see [19, 22]). Possible choices of the kernel include the *Gaussian kernel*

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

and the *Epanechnikov kernel* [8]

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

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 $\hat{m}(x_t) \to y_t$ 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 [16] as well as Fan and Yao [9].

Method 4.2.2 (Bandwidth selection with MISE)**.** This method is a plug-in method, which is based on the asymptotic expansion of the *mean integrated squared error* (MISE for short) for kernel smoothers

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

where m is the true function and $\hat{m}$ the estimator which depends on h. Under some regularity conditions, one derives the optimal bandwidth by minimization of the MISE, which typically depends on several unknown quantities that must be estimated from the data with some preliminary smoothing. In practice the choice of preliminary smoothing can become a problem. A normal reference bandwidth selector is given by Fan and Yao by

$$\hat{h}_{\mathrm{opt}} = \begin{cases} 1.06\, s\, n^{-1/5} & \text{for the Gaussian kernel,} \\ 2.34\, s\, n^{-1/5} & \text{for the Epanechnikov kernel,} \end{cases}$$

where s is the sample standard error of the independent variable, which is assumed to be stationary.

Method 4.2.3 (Bandwidth selection with cross validation)**.** The *leave-one-out cross validation* starts with omitting one observation (x_j, y_j). The remaining $n-1$ data points are used to obtain the following smoother at x_j:

$$\hat{m}_{h,j}(x_j) := (n-1)^{-1} \sum_{t \neq j} w_t(x_j) y_t,$$

which is an estimate of y_j where the weights $w_t(x_j)$ sum to $n-1$. Afterwards the same is performed for all remaining $n-1$ observations and

$$\mathrm{CV}(h) := n^{-1} \sum_{j=1}^{n} (y_j - \hat{m}_{h,j}(x_j))^2 W(x_j)$$

is defined, where W is a nonnegative weight function satisfying $\sum_{j=1}^{n} W(x_j) = n$ that can be used to down-weight the boundary points if necessary. This might be the case since points at the boundary have often fewer neighboring observations. The function CV is called the *cross-validation function* because it validates the ability of the smoother to predict y. The bandwidth h is chosen such that it minimizes CV.

Having presented two methods to choose the bandwidth in kernel regression, we continue with another method to estimate m in Equation (4.1).

Method 4.2.4 (Local linear regression method)**.** Assume that m in Equation (4.1) is twice continuously differentiable at some given point x in the support of m. Denote the available observations by $((y_t, x_t), t = 1, \ldots, n)$. The *local linear regression method* to nonparametric regression is to find a and b that minimize

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

where K_h is a kernel with bandwidth h as in Method 4.2.2. Denote the minimum of a by $\hat{a}$, which is the estimate of $m(x)$, while the minimum of b denoted by $\hat{b}$ can be used as an estimate of $m'(x)$. The least-squares problem has a closed-form solution, which is given by

$$\hat{a} = \frac{\sum_{t=1}^{n} w_t y_t}{\sum_{t=1}^{n} w_t},$$

where

$$w_t := K_h(x - x_t)(s_{n,2} - (x - x_t) s_{n,1})$$

and

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

for $j = 1, 2$. We leave the derivation to the interested reader.

In practice, to avoid that the denominator becomes zero,

$$\hat{m}(x) := \frac{\sum_{t=1}^{n} w_t y_t}{\sum_{t=1}^{n} w_t + n^{-2}}$$

is used as an estimate for $m(x)$.

4.3 Nonlinearity tests

In this section we discuss both, nonparametric and parametric statistics that have decent power against the models considered in Section 4.1.

4.3.1 Nonparametric tests

Under the null hypothesis of linearity, residuals of a properly specified linear model should be uncorrelated. Let us here consider the stronger assumption that they are independent, which holds automatically true for Gaussian noise. Any violation of independence in the residuals indicates inadequacy of the entertained model, including the linearity assumption. This is the basic idea behind various nonlinearity tests. Here we introduce two methods for the moment. For more the reader is referred to the literature.

Method 4.3.1 (Q-statistic of squared residuals)**.** This method by McLeod and Li applies the Ljung–Box statistics 1.2.12 to the squared residuals of an ARMA(p, q) model to check for model inadequacy. The test statistic is

$$Q(m) := n(n+2) \sum_{i=1}^{m} \frac{\hat{\rho}_i^2(Z_t^2)}{n-i},$$

where n is the number of observations, m is a properly chosen number of autocorrelations used in the test, $(Z_t, t = 1, \ldots, n)$ denotes the residual series, and $\hat{\rho}_i(Z_t^2)$ is the lag i autocorrelation function of Z_t^2. If the entertained linear model is adequate, $Q(m)$ is asymptotically χ^2_{m-p-q}-distributed.

The null hypothesis of the test is

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

where the parameter β_i is the coefficient of Z_{t-i}^2 in the linear regression

$$Z_t^2 = \beta_0 + \sum_{i=1}^{m} \beta_i Z_{t-i}^2 + e_t$$

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

Method 4.3.2 (Bispectral test)**.** This test can be used to test for linearity and Gaussianity. It depends on the result that a properly normalized bispectrum of a linear time series is constant over all frequencies and that the constant is equal to zero under normality. Here, the *bispectrum* of a time series is the Fourier transform of its third-order moments, but let us treat this in detail in what follows.

For a stationary time series

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

where μ is a constant, $Z \sim \mathrm{IID}(0, \sigma^2)$ and $(\psi_j, j \in \mathbb{Z})$ is a sequence of real numbers with $\psi_0 = 1$, the third-order moment is defined as

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

for $u, v \in \mathbb{Z}$, where we set $\psi_0 := 1$ and $\psi_i := 0$ for $k < 0$. For frequencies w_1 and w_2 the Fourier transform is then given by

$$b_3(w_1, w_2) := \frac{\mathbb{E}(Z_t^3)}{4\pi^2} \Gamma(-(w_1 + w_2)) \Gamma(w_1) \Gamma(w_2),$$

where Γ is defined by

$$\Gamma(w) := \sum_{u=0}^{\infty} \psi_u \exp(-iwu)$$

and $i = \sqrt{-1}$. Since the spectral density of X is given by

$$p(w) = \frac{\sigma^2}{2\pi} |\Gamma(w)|^2,$$

one obtains that the *bispectrum*

$$b(w_1, w_2) := \frac{|b_3(w_1, w_2)|^2}{p(w_1) p(w_2) p(w_1 + w_2)}$$

is constant for all (w_1, w_2). The bispectrum test estimates b over a suitably chosen grid of points and applies a test statistic similar to *Hotelling's T^2 statistic* to check the constancy. Since for a linear Gaussian series $\mathbb{E}(Z_t^3) = 0$, the bispectrum is zero for all frequencies.

4.3.2 Parametric tests

To conclude the section about nonlinearity tests we introduce one parametric method and remark at the end how this can be extended.

Method 4.3.3 (RESET test)**.** Ramsey [20] proposes a specification test for linear least-squares regression analysis referred to as a *RESET test*, which is readily applicable to linear AR models. Therefore consider the linear AR(p) model

$$X_t = (1, X_{t-1}, \ldots, X_{t-p})(\phi_0, \phi_1, \ldots, \phi_p)' + Z_t.$$

The first step of the RESET test is to obtain the least-squares estimate $\hat{\phi}$ and compute the fit

$$\hat{X}_t := (1, X_{t-1}, \ldots, X_{t-p})(\hat{\phi}_0, \hat{\phi}_1, \ldots, \hat{\phi}_p)',$$

the residual $\hat{Z}_t := X_t - \hat{X}_t$, and the sum of squared residuals

$$\mathrm{SSR}_0 := \sum_{t=p+1}^{n} \hat{Z}_t^2,$$

where n is as usual the sample size.

In the second step, consider the linear regression

$$\hat{Z}_t = (1, X_{t-1}, \ldots, X_{t-p})(\alpha_{10}, \ldots, \alpha_{1p})' + (\hat{X}_t^2, \ldots, \hat{X}_t^{s+1})(\alpha_{21}, \ldots, \alpha_{2s})' + V_t$$

for some $s \geq 1$ and compute the least-squares residuals

$$\hat{V}_t = \hat{Z}_t - (1, X_{t-1}, \ldots, X_{t-p})(\hat{\alpha}_{10}, \ldots, \hat{\alpha}_{1p})' - (\hat{X}_t^2, \ldots, \hat{X}_t^{s+1})(\hat{\alpha}_{21}, \ldots, \hat{\alpha}_{2s})'$$

and the sum of squared residuals

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

of the regression. The idea of the RESET test is that if the linear AR(p) model is adequate, then all α_{1i} and α_{2j} should be zero. This can be tested by using the F statistic given by

$$F := \frac{(\mathrm{SSR}_0 - \mathrm{SSR}_1)(n - p - g)}{\mathrm{SSR}_1 g},$$

where $g := s + p + 1$, which under linearity and normality assumption, has an F distribution with degrees of freedom g and $n - p - g$.

We remark that there exist several improvements of the RESET test. We here mention only the modification of the second step of the RESET test by Keenan and a different choice of the regressor by Tsay. For details the reader is referred to the literature.

4.4 Forecasting

We have seen in Section 1.2.3 that forecasting for linear time series can be done with closed-form formulas. This does not hold for most nonlinear models when the forecast horizon is greater than 1. In what follows we introduce parametric bootstraps to compute nonlinear forecasts.

Method 4.4.1 (Parametric bootstrap)**.** Given X_n, we want to forecast X_{n+h} for some $h > 0$. The parametric bootstrap computes realizations $X_{n+1}, \ldots, X_{n+h}$ sequentially in the following way. For $i = 1, \ldots, h$ repeat:

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

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

Repeat this whole procedure M times to get M realizations of X_{n+h}. Compute the sample average over the realizations to obtain a point forecast for X_{n+h} which we denote by $X_n(h)$.

The realizations could also be used to obtain an empirical distribution function which might be of use in the following methods when forecasting evaluation is done.

In what follows we introduce different methods to evaluate the performance of a forecast. Therefore let us do the following: Given a data set, we subdivide it into two subsamples which we refer to as *estimation subsample* and *forecasting subsample*. We will use the first one to build a nonlinear model. We derive the performance then by comparing the obtained forecasts computed by the model with the data of the forecasting subsample. In what follows three measures are used to get an idea of the performance which are commonly used in the literature. Nevertheless, we should mention that there exists no widely accepted measure to compare models.

Method 4.4.2 (Directional measure)**.** A typical measure for the evaluation of the performance of forecasts is to use a 2×2 contingency table that summarizes the number of "hits" and "misses" of the model in predicting ups and downs up to x_{n+h} in the forecasting subsample. Let the table be given by

Actual	Predicted		
	Up	Down	
Up	m_{11}	m_{12}	m_{10}
Down	m_{21}	m_{22}	m_{20}
	m_{01}	m_{02}	m

where m is the total number of h-step-ahead forecasts in the forecasting subsample, m_{11} is the number of "hits" in the predicting upward movements, m_{21} is the number of

"misses" in predicting downward movements of the market, and so on. It is clear that larger values in m_{11} and m_{22} indicate better forecasts. The test statistic

$$\chi^2 := \sum_{i,j=1}^{2} \frac{(m_{ij} - m_{i0}m_{0j}/m)^2}{m_{i0}m_{0j}/m}$$

can be used to evaluate the performance of the model, where a large χ^2 signifies that the model outperforms the chance of random choice. Under some assumptions, χ^2 has an asymptotic χ^2 distribution with one degree of freedom. For more details the reader is referred to the literature, especially to [6].

Method 4.4.3 (Magnitude measure)**.** Three statistics that are commonly used to measure performance of point forecasts are

- the *mean squared error*

$$\mathrm{MSE}(h) := m^{-1} \sum_{j=0}^{m-1} (X_{n+h+j} - X_{n+j}(h))^2,$$

- the *mean absolute deviation*

$$\mathrm{MAD}(h) := m^{-1} \sum_{j=0}^{m-1} |X_{n+h+j} - X_{n+j}(h))|,$$

- the *mean absolute percentage error*

$$\mathrm{MAPE}(h) := m^{-1} \sum_{j=0}^{m-1} \left| \frac{X_{n+j}(h)}{X_{n+h+j}} - 1 \right|,$$

where m is the number of h-step-ahead forecasts available in the forecasting subsample. The error computation is done between the data from the forecasting subsample and the h-step-ahead forecasts computed from the model that was derived from the estimation subsample.

In application one often chooses one of the above measures and then the model with the smallest magnitude on that measure. This is regarded as the best h-step-ahead forecasting model. Be aware that it might happen that different models are chosen for different forecast horizons h. For limitations in model comparison of the different measures, the reader is referred to the literature.

Method 4.4.4 (Distributional measure)**.** Compute the empirical distribution function $\hat{F}$ out of the sample obtained by the parametric bootstrap method 4.4.1 for the desired h-step-ahead forecast according to the model obtained from the estimation subsample. Use the forecasting subsample to compute

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

for all $j = 0, \ldots, m-1$, where m denotes the total number of h-step-ahead forecasts in the forecasting subsample. If the model is adequate, then $(u_{n+j}(h), j = 0, \ldots, m-1)$ will behave like a random sample from the uniform distribution on $[0, 1]$. For sufficiently large m, the Kolmogorov–Smirnov statistic can be used to test the sample with respect to the uniform distribution.

The method can be used for both, model checking and forecasting comparison.

5 Extreme value theory

Let us briefly discuss in this chapter how to treat extreme price movements in financial markets which are rare but important. We apply extreme value theory to give a quantification of risk. There exist three types of risk: credit risk, operational risk, and market risk. In what follows we discuss *value at risk*, which is mainly concerned with market risk but which can also be applied to credit risk. Afterwards we give a short introduction to extreme value theory before combining both section. We follow closely [21] in this chapter, but we refer the reader to [11] for a more detailed and mathematical introduction to risk management including value at risk.

5.1 Value at Risk

Assume that we are interested in the risk of a financial position at time t for the next h periods. Let $\Delta V(h)$ be the change in value of the underlying assets of the financial position from time t to $t+h$. Then the associated *loss function* $L(h)$ is a positive or negative function of $\Delta V(h)$ depending on the position being short or long. Let the *cumulative distribution function* (CDF for short) of $L(h)$ be denoted by F_h, i.e.,

$$F_h(x) := P(L(h) \leq x),$$

which is a nondecreasing, right-continuous (i.e., càdlàg) function.

Definition 5.1.1. The *value at risk* V@R (also abbreviated by VaR) of a financial position with associated loss function L over the time horizon h with tail probability $p = 1-\alpha$ is defined by

$$\text{V@R}_\alpha := \inf\{x \in \mathbb{R}, P(L(h) > x) \leq 1-\alpha\} = \inf\{x \in \mathbb{R}, F_h(x) \geq \alpha\}$$

We remark that V@R_α is the α-quantile of the cumulative distribution function of the loss function, sometimes also referred to as the upper pth quantile because p is the upper tail probability of the loss distribution.

Typical values for α are 0.95 and 0.99, although 0.999 might be required for stress testing.

Let us introduce in what follows different approaches to value at risk calculations.

5.1.1 RiskMetrics

The RiskMetrics methodology was developed by J. P. Morgan to value at risk calculation. Let us consider a simple form of it in what follows.

Assume that the continuously compounded daily return of a portfolio follows a conditional normal distribution. Denote the daily log returns by $(r_t, t \in \mathbb{Z})$ and the information available at time $t-1$, i.e., the filtration of the underlying stochastic processes, by $\mathcal{F}_{t-1}$. RiskMetrics assumes that the conditional distribution of r_t given $\mathcal{F}_{t-1}$ is $\mathcal{N}(\mu_t, \sigma_t^2)$, where μ_t is the conditional mean and σ_t^2 the conditional variance of r_t given $\mathcal{F}_{t-1}$. Furthermore it is assumed that these two quantities evolve according to the simple model that $\mu_t = 0$ for all t and

$$\sigma_t^2 = \beta\sigma_{t-1}^2 + (1-\beta)r_{t-1}^2$$

for all t and some fixed $\beta \in (0,1)$, typically $\beta = 0.94$. We remark that the method assumes that the logarithm of the daily price $p_t = \ln P_t$ of the portfolio satisfies the difference equation $p_t - p_{t-1} = a_t$, where $a_t = \sigma_t\epsilon_t$ follows an IGARCH$(1,1)$ model without drift.

One shows in an easy *exercise* that the log return from time $t+1$ to time $t+k$ (inclusive) is

$$r_t[k] = r_{t+1} + \cdots + r_{t+k}.$$

Under the assumed model, $\mathbb{E}(r_t[k]|\mathcal{F}_t)$ is normally distributed with mean zero and variance

$$\sigma_t^2[k] := \mathsf{Var}(r_t[k]|\mathcal{F}_t) = \sum_{i=1}^{k} \mathsf{Var}(a_{t+i}|\mathcal{F}_t),$$

where

$$\mathsf{Var}(a_{t+i}|\mathcal{F}_t) = \mathbb{E}(\sigma_{t+i}^2|\mathcal{F}_t)$$

can be obtained in practice recursively using forecasting methods for IGARCH models. One shows further that

$$\mathsf{Var}(r_{t+i}|\mathcal{F}_t) = \sigma_{t+1}^2$$

for all $i = 1, \ldots, k$ and hence

$$\sigma_t^2[k] = k\sigma_{t+1}^2,$$

which implies that $\mathbb{E}(r_t[k]|\mathcal{F}_t) \sim \mathcal{N}(0, \sigma_{t+1}^2)$.

RiskMetrics uses this result to calculate value at risk for the log return. Under the normality assumption, one obtains for the upper 5% quantile that

$$\mathrm{V@R}_{0.05} = 1.65\sigma_{t+1}$$

for the next trading day and for the next k trading days

$$\mathrm{V@R}_{0.05}[k] = 1.65\sqrt{k}\sigma_{t+1}.$$

This scaling by $\sqrt{k}$ is referred to as the *square root of time role* in value at risk calculation under RiskMetrics.

The advantage of RiskMetrics is its simplicity and that it makes risk more transparent. Due to heavy tails that occur frequently in practice, the normality assumption of RiskMetrics leads to underestimation of value at risk.

5.1.2 Quantile estimation

This section introduces a nonparametric approach to value at risk calculation, i.e., it makes no specific assumption on the distribution.

In what follows we have to assume that the distribution of return in the prediction period is the same as that in the sample period. Then one can use the empirical quantile of the return to calculate value at risk.

Let $r_1, \ldots, r_n$ be the returns of a portfolio in the sample period. The *order statistic* is the values sorted in increasing order, which we denote by

$$r_{(1)} \leq r_{(2)} \leq \cdots \leq r_{(n)}.$$

Then $r_{(1)}$ is the sample minimum and $r_{(n)}$ the sample maximum.

Assuming that the returns are iid random variables with continuous distribution, probability density function f and cumulative distribution function F, the following theorem (see, e.g., [5, Appendix 2]) applies:

Theorem 5.1.2. *Let $(r_1, \ldots, r_n)$ be a sample of a sequence of iid random variables with cumulative distribution function F and density f, $\ell = np$ for some fixed $p \in (0,1)$, and $x_p = F^{-1}(p)$. Assume that $f(x_p) \neq 0$. Then the order statistic of $r_{(\ell)}$ satisfies asymptotically*

$$r_{(\ell)} \sim \mathcal{N}(x_p, n^{-1}p(1-p)f(x_p)^{-2}).$$

This theorem gives an opportunity to estimate the quantile x_p by using $r_{(\ell)}$.

In practice it might happen that the probability of interest p does not satisfy that $np \in \mathbb{N}$. Then simple interpolation can be used to obtain estimates. More specifically, let ℓ_1 and ℓ_2 be the two neighboring positive integers of np. Set

$$p_i := n^{-1}\ell_i$$

for $i = 1, 2$, which yields that $r_{(\ell_i)}$ is a consistent estimate of x_{p_i}. Then x_p can be estimated by

$$\hat{x}_p := \frac{p_2 - p}{p_2 - p_1} r_{(\ell_1)} + \frac{p - p_1}{p_2 - p_1} r_{(\ell_2)}.$$

Advantages of the empirical quantile method to value at risk calculation include its simplicity and that no specific distributional assumption is used. The drawbacks of the

method include the assumptions that the distribution of the returns remains unchanged from the sample to the prediction period. This implies that the predicted loss cannot be greater than the historical one, which is not good in practice. When the tail probability is small, the empirical quantile is not an efficient estimate of the theoretical quantile. Furthermore no explanatory variables are included in the estimation procedure, which are discussed in the following section. In real application the estimate of value at risk obtained by the empirical quantile method can just serve as a lower bound for the actual value at risk.

5.1.3 Quantile regression

In practice one often has explanatory variables available that are important to the problem, e.g., the action taken by central banks on interest rates. It therefore makes sense to include this information into $(\mathcal{F}_t, t \in \mathbb{Z})$. The estimation of the quantiles using $\mathbb{E}(r_{t+1}|\mathcal{F}_t)$ under this filtration is referred to as *regression quantiles* in the literature.

For that let us consider the linear regression

$$r_t = \beta' x_t + a_t,$$

where β is a k-dimensional vector of parameters and x_t a vector of predictors that are elements of $\mathcal{F}_{t-1}$. The distribution of $\mathbb{E}(r_t|\mathcal{F}_{t-1})$ is a translation of the distribution of a_t since $\beta' x_t$ is known under $\mathcal{F}_{t-1}$, i.e., it is obvious by construction that it is $\mathcal{F}_t$-measurable. Koenker and Bassett suggest in [18] estimating the conditional quantile $\mathbb{E}(x_p|\mathcal{F}_{t-1})$ of r_t given $\mathcal{F}_{t-1}$ as

$$\hat{x}_p|\mathcal{F}_{t-1} \equiv \inf\{\beta_0' x | R_p(\beta_0) = \min\},$$

where $R_p(\beta_0) = \min$ means that β_0 is obtained by

$$\beta_0 = \arg\min_{\beta} \sum_{t=1}^{n} w_p(r_t - \beta' x_t)$$

and w_p is defined by

$$w_p(z) := \begin{cases} pz & \text{if } z \geq 0, \\ (p-1)z & \text{if } z < 0. \end{cases}$$

5.2 Extreme value theory

In this section let us consider a series of n returns $(r_1, \ldots, r_n)$ with order statistic $(r_{(1)}, \ldots, r_{(n)})$, where we recall that $r_{(1)} = \min\{r_j, j = 1, \ldots, n\}$ and $r_{(n)} = \max\{r_j, j = 1, \ldots, n\}$. Following the literature and using the loss function in value at risk calculation, it will be focused on properties of the maximum return $r_{(n)}$, while those of the minimum can be easily obtained by a sign change $r_{(1)} = -\max\{-r_j, j = 1, \ldots, n\}$.

5.2.1 Introduction to extreme value theory

For a review of extreme value theory let us assume that the returns are serially independent with common cumulative distribution function F in what follows. Furthermore let the range of the returns be $[l, u]$, where $l = -\infty$ and $u = +\infty$ for log returns. In an easy *exercise*, one shows that the cumulative distribution function of $r_{(n)}$ denoted by $F_{n,n}$ is given by

$$F_{n,n}(x) = (F(x))^n.$$

In practice F as well as $F_{n,n}$ are unknown. It is clear from the obtained formula that $F_{n,n}$ degenerates for $n \to +\infty$. More specifically

$$F_{n,n}(x) \to \begin{cases} 0 & \text{if } x < u, \\ 1 & \text{if } x \geq u, \end{cases}$$

i.e., the cumulative distribution function has no practical value.

Extreme value theory is concerned with finding two sequences $(\alpha_n, n \in \mathbb{N})$ and $(\beta_n, n \in \mathbb{N})$ with $\alpha_n > 0$ for all $n \in \mathbb{N}$ such that the distribution of

$$r_{(n*)} := \frac{r_{(n)} - \beta_n}{\alpha_n}$$

converges to a random variable with nondegenerate distribution as $n \to +\infty$.

The sequence $(\beta_n, n \in \mathbb{N})$ is a location series and $(\alpha_n, n \in \mathbb{N})$ is a series of scaling factors. Under the independence assumption, the limiting distribution of $r_{(n*)}$ is given by

$$F_*(x) = \begin{cases} \exp\left(-(1+\xi x)^{-1/\xi}\right) & \text{if } \xi \neq 0, \\ \exp\left(-\exp(-x)\right) & \text{if } \xi = 0, \end{cases}$$

for $x < -1/\xi$ if $\xi < 0$ and for $x > -1/\xi$ if $\xi > 0$. The case $\xi = 0$ is taken as the limit when $\xi \to 0$. The parameter ξ is referred to as the *shape parameter* that governs the tail behavior of the limiting distribution. Furthermore the parameter $\alpha := \xi^{-1}$ is called the *tail index* of the distribution.

The limiting distribution is the *generalized extreme value distribution* of Jenkinson [17] for the maximum that encompasses the three types of limiting distributions of Gnedenko [13]

- Type I: $\xi = 0$, the *Gumbel* family with cumulative distribution function

$$F_*(x) = \exp(-\exp(-x)), \quad x \in (-\infty, \infty);$$

- Type II: $\xi > 0$, the *Fréchet* family with cumulative distribution function

$$F_*(x) = \begin{cases} \exp\left(-(1+\xi x)^{-1/\xi}\right) & \text{if } x > -1/\xi, \\ 0 & \text{else;} \end{cases}$$

- Type III: $\xi < 0$, the *Weibull* family with cumulative distribution function

$$F_*(x) = \begin{cases} \exp\left(-(1+\xi x)^{-1/\xi}\right) & \text{if } x < -1/\xi, \\ 1 & \text{else.} \end{cases}$$

Gnedenko gave necessary and sufficient conditions for the cumulative distribution function F of the returns to be associated with one of the three types of limiting distribution. The right tail of the distribution declines exponentially for the Gumbel family, by a power function for the Fréchet family, and is finite for the Weibull family.

For risk management, we are mainly interested in the Fréchet family, which includes stable and Student-t distributions. The Gumbel family consists of thin-tailed distributions such as normal and lognormal distributions.

The density function f_* of F_* can be obtained easily by differentiation

$$f_*(x) = \begin{cases} (1+\xi x)^{-1/\xi-1} \exp\left(-(1+\xi x)^{-1/\xi}\right) & \text{if } \xi \neq 0, \\ \exp\left(-x - \exp(-x)\right) & \text{if } \xi = 0, \end{cases} \tag{5.1}$$

where $x \in (-\infty, \infty)$ for $\xi = 0$, $x < -\xi^{-1}$ for $\xi < 0$, and $x > -\xi^{-1}$ for $\xi > 0$.

Extreme value theory has two important implications. First the tail behavior of the cumulative distribution function F or the return series and not the specific distribution determines the limiting distribution of the (normalized) maximum. The location series and series of scaling factors may however depend on F. Secondly, Feller shows in [10, p. 279] that the tail index ξ does not depend on the time interval of the observations, which becomes handy in value at risk calculations.

5.2.2 Empirical estimation

The goal of this section is to give an estimation method for the parameters ξ, β_n, and α_n. Since for a given sample, one extreme value is not sufficient to do an extreme value theory, other approaches have to be found. Therefore divide your sample of T returns $(r_t, t = 1, \ldots, T)$ into g subsamples of n observations, such that $T = ng$ (which is assumed to match for simplicity). Let $r_{n,i}$ denote the maximum of the ith subsample. When n is sufficiently large

$$x_{n,i} := \frac{r_{n,i} - \beta_n}{\alpha_n}$$

should follow an extreme value distribution and the collection of subsample maxima $(r_{n,i}, i = 1, \ldots, g)$ can then be regarded as a sample of g observations from that extreme value distribution.

Let us here just introduce one parametric method to estimate the parameters of the extreme value distribution. Another parametric method would be a regression method. For this as well as a nonparametric approach, the reader is referred to [21, Sec. 7.5.2].

Method 5.2.1 (Maximum likelihood method)**.** Assuming that the sample $(r_{n,i}, i = 1, \ldots, g)$ follows a generalized extreme value distribution such that the density of $x_{n,i}$ is given by (5.1), then the density of $r_{n,i}$ is by a simple transformation

$$f(r_{n,i}) = \begin{cases} \alpha_n^{-1}(1+\alpha_n^{-1}\xi_n(r_{n,i}-\beta_n))^{-(1+\xi_n)/\xi_n} \exp(-(1+\alpha_n^{-1}\xi_n(r_{n,i}-\beta_n))^{-1/\xi_n}) & \text{if } \xi_n \neq 0, \\ \alpha_n^{-1}\exp\left(-\alpha_n^{-1}(r_{n,i}-\beta_n) - \exp(-\alpha_n^{-1}(r_{n,i}-\beta_n))\right) & \text{if } \xi_n = 0, \end{cases}$$

where it is understood that $1+\alpha_n^{-1}\xi_n(r_{n,i}-\beta_n) > 0$ if $\xi_n \neq 0$. Here ξ_n is ξ in the previous notation, which indicates that it depends on the choice of n. Under the independence assumption, the likelihood function is

$$\ell(r_{n,1}, \ldots, r_{n,g}|\xi_n, \alpha_n, \beta_n) = \prod_{i=1}^{g} f(r_{n,i}).$$

Nonlinear estimation procedures can then be used to obtain maximum likelihood estimates of ξ_n, β_n, and α_n.

5.3 Extreme value approach to value at risk

Assuming the framework of Section 5.2, we obtain estimates of the location, scale, and shape parameters for the extreme value distribution of a sample of returns of size T. This can be used to compute quantiles of the extreme value distribution. Then the $r_n^* = (1-p^*)$th quantile for a small upper tail probability p^* can be computed by

$$r_n^* = \begin{cases} \beta_n - \frac{\alpha_n}{\xi_n}\left(1-(-\ln(1-p^*))^{-\xi_n}\right) & \text{if } \xi_n \neq 0, \\ \beta_n - \alpha_n \ln(-\ln(1-p^*)) & \text{if } \xi_n = 0, \end{cases}$$

where the case of $\xi_n \neq 0$ is of major interest in financial applications. The quantile r_n^* is the value at risk based on the extreme value theory for the subperiod maximum. Finally the value at risk for the original data has to be computed and retrieved from that. In summary one has to do the following to compute the value at risk for the original log returns:

Method 5.3.1. Denote by $(r_t, t = 1, \ldots, T)$ the observed log returns.

(i) Select the length of the subperiod n and obtain subperiod maxima $(r_{n,i}, i = 1, \ldots, g)$, where $T = gn$.

(ii) Obtain the maximum likelihood estimates of β_n, α_n, and ξ_n.

(iii) Check the adequacy of the fitted extreme value model, which is omitted at that stage in the lecture notes but can be looked up in [21].

(iv) If the extreme value model is adequate, compute

$$\mathrm{V@R}_p = \begin{cases} \beta_n - \frac{\alpha_n}{\xi_n}\left(1 - (-n\ln(1-p))^{-\xi_n}\right) & \text{if } \xi_n \neq 0, \\ \beta_n - \alpha_n \ln(-n\ln(1-p)) & \text{if } \xi_n = 0, \end{cases}$$

where p is a given small upper tail probability.

Bibliography

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

[2] George E. P. Box and Gwilym M. Jenkins. *Time Series Analysis: Forecasting and Control*. Holden-Day, San Francisco, Calif.-Düsseldorf-Johannesburg, revised edition, 1976. Holden-Day Series in Time Series Analysis.

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

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

[5] David R. Cox and David V. HInkley. *Theoretical Statistics*. London: Chapman & Hall Ltd., 1974.

[6] 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.

[7] 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.

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

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

[10] William Feller. *An Introduction to Probability Theory and Its Applications. Volume II*. Wiley Series in Probability and Mathematical Statistics. New York etc.: John Wiley and Sons, 1971.

[11] Hans Föllmer and Alexander Schied. *Stochastic Finance: An Introduction in Discrete Time*. Berlin: de Gruyter, 3rd revised and extended ed. edition, 2011.

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

[13] Boris V. Gnedenko. Sur la distribution limite du terme maximum d'une série aléatoire. *Annals of Mathematics*, 44:423–453, 1943.

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

[15] 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.

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

[17] A.F. Jenkinson. The frequency distribution of the annual maximum (or minimum) of meteorological elements. *Quarterly Journal of the Royal Meteorological Society*, 81:158–171, 1955.

[18] Roger Koenker and Gilbert Bassett. Regression quantiles. *Econometrica*, 46:33–50, 1978.

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

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

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

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