6 MONOGRAPH SERIES: JAN MIELNICZUK - ANALYSIS OF TIME SERIES: THEORY

6

MONOGRAPH SERIES

JAN MIELNICZUK

ANALYSIS OF TIME SERIES: THEORY

INSTITUTE OF COMPUTER SCIENCE
POLISH ACADEMY OF SCIENCES

MONOGRAPH SERIES
INFORMATION TECHNOLOGIES: RESEARCH
AND THEIR INTERDISCIPLINARY APPLICATIONS

6

JAN MIELNICZUK

ANALYSIS OF TIME SERIES: THEORY

INSTITUTE OF COMPUTER SCIENCE
POLISH ACADEMY OF SCIENCES

Warsaw, 2015

Publication issued as a part of the project:
"Information technologies: research and their interdisciplinary applications",
Objective 4.1 of Human Capital Operational Programme.
Agreement number UDA-POKL.04.01.01-00-051/10-01.

Publication is co-financed by European Union from resources of European Social Fund.

Project leader: Institute of Computer Science, Polish Academy of Sciences

Project partners: System Research Institute, Polish Academy of Sciences, Nałęcz
Institute of Biocybernetics and Biomedical Engineering, Polish Academy of Sciences

Editors-in-chief: Olgierd Hryniewicz
Jan Mielniczuk
Wojciech Penczek
Jacek Waniewski

Reviewer: Piotr Pokarowski

Jan Mielniczuk
Institute of Computer Science, Polish Academy of Sciences
miel@ipipan.waw.pl
http://www.ipipan.waw.pl/staff/j.mielniczuk

Publication is distributed free of charge

ISBN 978-83-63159-16-0
e-ISBN 978-83-63159-17-7

Layout: Piotr Borkowski
Cover design: Waldemar Słonina

Contents

Preface

The monograph discusses mathematical foundations and recent theoretical developments in analysis of time series that is random temporal phenomena. Practical applications of the discussed methodology are mostly left out and intended for the second part of this book.

My motivations to write this monograph were twofold. I believe that a thorough knowledge of probabilistic and statistical modelling tools of time series analysis are necessary for understanding of its major current developments and thus also for an active research in this field. Secondly, theoretical aspects of time series analysis are relatively advanced technically and this frequently makes potential young researchers shy away from its topics. I tried to make this initiation process a little bit easier while avoiding an obvious trap of a short-cut, i.e. getting completely lost.

I attempted to make this book self-contained and assume only the knowledge of probability and statistics on intermediate level.

The monograph can be broadly divided into two parts: the one devoted to modelling and prediction of time series the second one to estimation and inference for time series; the division is clearly seen from the list of contents. I discuss several subjects which are intensively researched nowadays such as modelling and inference for long-range dependent data and novel methods of dependence quantification. There are however, many important fields of active research which are left undiscussed such as e.g. state-space modelling, multivariate time series (cf. e.g. Lütkepohl (2007)), various aspects of nonstationarity (cf. e.g. Hurd and Miamee (2007)) and nonlinear and nonparametric time series modelling (Fan and Yao (2003)).

I have drawn on many sources writing this monograph. The most important ones are Brockwell and Davis (1991), Pourahmadi (2001), Shumway and Stoffer (2006) and Ibragimow and Linnik (1971). Other books which can be consulted include e.g. Hamilton (1994) (especially for econometric applications) and Cryer and Chan (2008). Lindsay (2006) shows a different perspective on time series analysis based on point processes. Lahiri (2003) discusses resampling methods for dependent data and Taniguchi and Kakizawa (2000) advanced asymptotics of time series. McQuarrie and Tsai (1998) focuses on model selection and Makridakis et al. (1998) on forecasting. Applications of the time series analysis and specific models useful in financial engineering are discussed in Taylor (2005), Tsay (2010) and Ruppert (2011).

It is also necessary to mention here an excellent introductory book Chatfield

(2003) and a never aging masterpiece Hannan (1967).
Chosen parts of sections 1 and 3-9 may serve as a basis of MSc or PhD one-semester course on time series analysis.
Piotr Pokarowski and Hubert Szymanowski were among the first readers of the original version and they contributed much to its improvement. I thank them both.

1

Introduction to time series

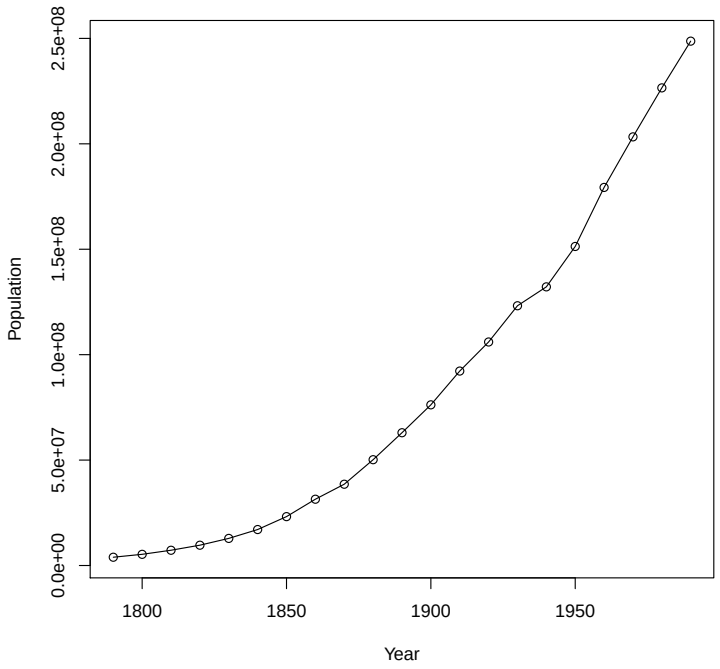
In this section we define time series concept, discuss linear subspaces related to it and its covariance structure. Weakly and strictly stationary time series are introduced together with several examples such as linear processes which will play important role further on.

1.1 Time series: basics

In the book we will discuss analysis and modelling of measurements of some recurring phenomenon taken at consecutive time points. These time points can be minutes, days, months, quarters, etc. Any set of measurements (x_t) indexed by time will be modeled by a sequence of random variables $(X_t)_{t\in T}$, where T is a countable subset of integers $\mathbb{Z}$ signifying consecutive time points. Such stochastic process $(X_t)_{t\in T}$ is called time series. For the most part of the book we consider the situation when the observations are quantitative and we assume that $X_t \in \mathbb{R}$. However, we will also briefly address the case when outcomes are qualitative, typical example being when they correspond to the consecutive words, or letters, in a book or a piece of human utterance. Usually, the index set T will be a set of natural numbers $\mathbb{N} = \{1, 2, \ldots\}$ or integers $\mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$. The second type of indexing is frequently adopted in modelling and corresponds to (sometimes implicit) assumption that the observations are recorded from infinite past on. Note that taking $T = \mathbb{N}$ or $T = \mathbb{Z}$ is adequate in situations when time points are equidistant. This may be sometimes problematic as e.g. when values of a certain financial index are recorded on consecutive working days. Then imposing stationarity structure on the underlying model, what we will frequently do here, will obviously ignore weekend effect.
Additional assumption which will (almost always) be adapted is that X_t are random variables defined on a probability space $(\Omega, \mathcal{F}, P)$ such that they are square integrable i.e. $X_t \in \mathcal{L}^2(\Omega, \mathcal{F}, P), t \in T$. This will enable us to treat elements of time series as elements of Hilbert space $\mathcal{L}^2(\Omega, \mathcal{F}, P)$.
Below we give an example of time series and its representation using statistical package R.

```
library(MASS)
USpop <- ts(data=scan("USPOP.DATA"), start=1790, end=1990,
frequency=0.1)
# option frequency- no. of obs per time unit, in this case unit
```

```
 #=1 year,frequency=0.1 means 1 observation  every 10 years
ts.plot(USpop, gpars=list(xlab="Year", ylab="Population",
type="o"))
```



1.2 Linear subspaces related to (X_t)

Random variable random variable X_t of time series $(X_t)_{t\in T}$ will be considered, unless it is explicitly stated otherwise, as an element of $\mathcal{L}^2(\Omega, \mathcal{F}, P)$, that is the space of square integrable *real* random variables on $(\Omega, \mathcal{F}, P)$ with a scalar product

$$< X, Y >= EXY = \int X(\omega)Y(\omega)dP(\omega).$$

In general, for complex random variables, we define $< X, Y >= EX\bar{Y} = \int X(\omega)\bar{Y}(\omega)dP(\omega)$, where $\bar{Y}$ stands for a complex conjugate of Y. When $X \in \mathcal{L}^2(\Omega, \mathcal{F}, P)$, we let

$$||X||^2 =< X, X > .$$

Note, however, that sometimes notation for $(X_t)_{t\in T}$ will be shortened to X. We will interchangingly use EX^2 and $||X||^2$ to denote squared norm of a random

variable $X \in \mathcal{L}^2$. Consistently with the geometry of $\mathcal{L}^2$ we will use the notion $X \perp Y$ when $< X, Y >= 0$.
Denote by $sp(X_t)_{t\in T}$ a linear span of random variables $\{X_t, t \in T\}$. We define two basic linear subspaces of $\mathcal{L}^2(\Omega, \mathcal{F}, P)$ related to time series $(X_t)_{t\in T}$. The first relates to the entire time series and the second to its history up to time t (moment t being included).

$$\begin{aligned} H(X) &= \overline{sp}(X_t, t \in T)(\text{closure in } \mathcal{L}^2) \\ &= \overline{\{a_1 X_{t_1} + a_2 X_{t_2} + \cdots + a_n X_{t_n}, \quad t_1, \ldots t_n \in T, \ a_1, \ldots, a_n \in \mathbb{R}, n \in \mathbb{N}\}} \end{aligned}$$

and

$$H_t = H_t(X) = \overline{sp}(X_s, s \leqslant t). \tag{1.1}$$

Obviously, we have

$$H_t(X) \subseteq H_{t+1}(X) \subseteq \cdots \subseteq H(X).$$

Moreover, for $T = \mathbb{Z}$, we define remote past subspace $H_{-\infty}$ as

$$H_{-\infty} = H_{-\infty}(X) = \bigcap_{t\in\mathbb{Z}} H_t(X). \tag{1.2}$$

Note that $H_{-\infty}$ is a closed subspace and as 0 (random variable equal 0 almost everywhere) belongs to any H_t it also belongs to $H_{-\infty}$. We show below that $H_{-\infty}$ may contain more elements.

Example 1.2.1 *Let $X_t = \varepsilon + \varepsilon_t$, $t \in \mathbb{Z}$, where ε_t are mean zero uncorrelated random variables such that $E\varepsilon_t^2 < \infty$ and $E\varepsilon^2 < \infty$. We check that $\varepsilon \in H_{-\infty}(X)$. Indeed, consider*

$$w_t = \frac{X_t + X_{t-1} + \cdots + X_{t-|t|+1}}{|t|} = \frac{\varepsilon_{t-|t|+1} + \cdots + \varepsilon_t}{|t|} + \varepsilon =: S_t + \varepsilon \quad \in H_t.$$

Note that as $||S_t||^2 = ES_t^2 = E\varepsilon_1^2/|t| \to 0$ when $t \to -\infty$, thus $w_t \to \varepsilon$ in $\mathcal{L}^2$ and this together with $w_t \in H_t$ yields $\varepsilon = \lim_{t\to-\infty} w_t \in H_{-\infty}$. Indeed, suppose that it belongs to its complement: $\varepsilon \in H_{-\infty}^c = \bigcup_{t\in\mathbb{Z}} H_t^c$, then it belongs to $H_{t_0}^c$ for a certain t_0. As the last set is open there is an open neighbourhood U of $\lim_{t\to-\infty} w_t$, contained in it. However, this is impossible as $U \subseteq H_{t_0}^c \subseteq H_t^c$ for $t < t_0$ contradicts $\varepsilon = \lim_{t\to-\infty} w_t, w_t \in H_t$ as any open neighborhood of ε contains elements of H_t for sufficiently large $|t|$.
Note also that it follows from the given reasoning that if (X_t) satisfies $(X_t + X_{t-1} + \cdots + X_{t-|t|+1})/|t| \to w$ when $t \to -\infty$ then $w \in H_{-\infty}(X)$.

1.3 Weakly and strictly stationary processes

We define a basic second order characteristic of time series.

Definition 1 *An autocovariance function of a (real-valued) time series* $(X_t)_{t\in T}$ *is defined as*

$$\gamma_X(s,t) = \mathrm{Cov}(X_s, X_t) = E((X_s - EX_s)(X_t - EX_t)) = < X_s - EX_s,\ X_t - EX_t > \tag{1.3}$$

From the Schwarz inequality follows that value $\gamma_X(s,t)$ is finite if $X_t, X_s \in \mathcal{L}^2$ (what we assume). For complex-valued time series the corresponding definition is

$$\gamma_X(s,t) = E((X_s - EX_s)(\overline{X_t - EX_t})). \tag{1.4}$$

Note that $\gamma_X(t,t) = \mathrm{Var}(X_t)$. We define now weakly stationary time series.

Definition 2 *Time series* $(X_t)_{t\in\mathbb{Z}}$ *is weakly stationary if*

(i) $EX_t = m, \quad t \in \mathbb{Z}$;
(ii) $\gamma_X(s,t) = \gamma_X(s+r, t+r), \quad$ for any $r, s, t \in \mathbb{Z}$.

It is sometimes assumed in the definition of weak stationarity that $\mathrm{Var}(X_t) < \infty, \quad t \in \mathbb{Z}$. Note however that since $X_t \in \mathcal{L}^2(\Omega, \mathcal{F}, P)$, we have $\mathrm{Var}(X_t) \leqslant EX_t^2 < \infty$ and this assumption is redundant. It follows from the second condition that $\mathrm{Var}(X_t)$ is constant and does not depend on t. We will call such process shortly WS (weakly stationary) time series.
For processes with $T = \mathbb{N}$ weak stationarity is defined completely analogously.
For weakly stationary process we have that $\gamma_X(s,t) = \gamma_X(s-t, 0))$ and thus $\gamma_X(s,t)$ is a function $s-t$ only. In such case we define autocovariance function $\gamma_X(h) : \mathbb{Z} \longrightarrow \mathbb{R}$ as a function of *one* variable

$$\gamma_X(h) := \gamma_X(h,0) = \gamma_X(s,t), \quad \text{for} \quad s - t = h. \tag{1.5}$$

ACVF is frequently used shorthand for autocovariance function. We list basic properties of $\gamma_X(h)$:

(i) $\mathrm{Var}(X_t) = \gamma(0),\ t \in \mathbb{Z}$;
(ii) $|\gamma(h)| \leqslant \gamma(0),\ h \in \mathbb{Z}$;
(iii) $\gamma(h) = \gamma(-h) \quad (\gamma(h) = \overline{\gamma(-h)}$ for complex-valued time series);
(iv) $\gamma_X(\cdot)$ non-negative definite, i.e. for any $a_1, \ldots, a_n \in \mathbb{R}, t_1, \ldots, t_n \in \mathbb{R}$,

$$\sum_{1\leqslant i,j\leqslant n} a_i a_j \gamma_X(t_i - t_j) \geqslant 0. \tag{1.6}$$

Proof. Note that (ii) follows from

$$|\mathrm{Cov}(X_{t+h}, X_t)| \leqslant (\mathrm{Var}X_{t+h})^{1/2}(\mathrm{Var}X_t)^{1/2} = \gamma(0)^{1/2}\gamma(0)^{1/2} = \gamma(0).$$

In order to prove (1.6) let $\mathbf{w} = (X_{t_1} - EX_{t_1}, \ldots, X_{t_n} - EX_{t_n})'$ and $\mathbf{a} = (a_1, \ldots, a_n)'$. Then expression in (1.6) equals

$$\mathrm{Var}(\mathbf{a}'\mathbf{w}) = \sum_{i,j} a_i a_j \gamma_X(t_i - t_j) \geqslant 0.$$

Observe that (iv) asserts that autocovariance matrix $\mathbf{\Gamma}_n = \big(\gamma_X(t_i - t_j)\big)_{i,j\leqslant n}$ is non-negative definite. Note that $\mathbf{\Gamma}_n$ is a Toeplitz matrix. Dropping the assumption of weak stationarity in a general case we have that $\big(\gamma_X(t_i, t_j)\big)_{i,j\leqslant n}$ is non-negative definite. The property that matrix $\mathbf{\Gamma}_n$ is non-negative definite for any $n \in \mathbb{N}$ characterizes autocovariance functions i.e. the following result is true.

Theorem 1.3.1 *Let $\gamma(h) : \mathbb{Z} \to \mathbb{R}$ be that for any n $\mathbf{\Gamma}_n = (\gamma(i-j))_{i,j\leqslant n}$ is non-negative definite. Then there exists time series $(X_t)_{t\in\mathbb{Z}}$ such that $\gamma(h)$ is its autocovariance function.*

Proof. As $\mathbf{\Gamma}_n$ is non-negative definite then multivariate n-dimensional normal distribution $\mathrm{N}(0, \mathbf{\Gamma}_n)$ exists. It is easy to see that if we specify finite dimensional distribution in n consecutive points as $\mathrm{N}(0, \mathbf{\Gamma}_n)$, then the family of these distributions is consistent i.e. it satisfies assumptions of Kolmogorov's existence theorem. The conclusion then follows from Kolmogorov's theorem.
Another characterization of the autocovariance is Herglotz's theorem discussed later. Now we define a stronger property of time series than weak stationarity.

Definition 3 *Time series $(X_t)_{t\in\mathbb{Z}}$ (not necessarily belonging to $\mathcal{L}^2$) is strictly stationary if for any $t_1, t_2, \ldots, t_k, h \in \mathbb{Z}$*

$$(X_{t_1}, X_{t_2}, \ldots, X_{t_k}) \overset{\mathcal{D}}{\sim} (X_{t_1+h}, X_{t_2+h}, \ldots, X_{t_k+h}), \tag{1.7}$$

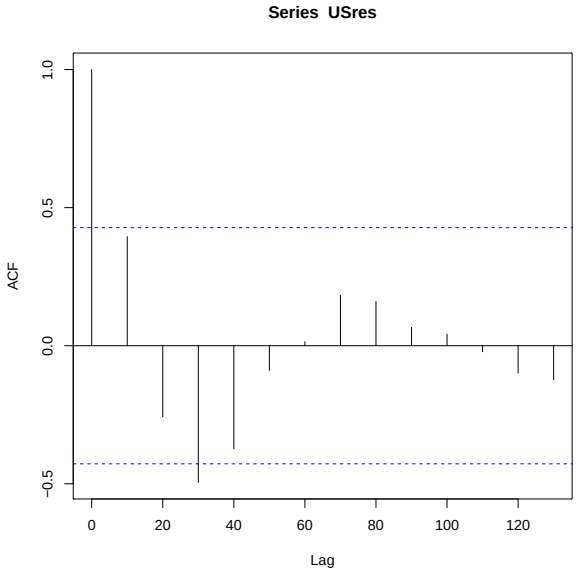
where $\overset{\mathcal{D}}{\sim}$ denotes equality of distributions.

Heuristically, time series is strictly stationary when its distributional properties do not depend on the moment from which the process is observed. Indeed, if the process is observed not from the moment 0 but from the moment h, then time moment t_i with respect to new origin corresponds to moment $t_i + h$ with respect to the old one. If we stipulate that the distributions of observations at time points $t_1, \ldots, t_k$ with respect to the new origin are the same as the those of observations taken at time points $t_1, \ldots, t_k$ with respect to the old one then this entails (1.7). Observe that in particular this means that distributions of any two variables X_s and X_t are equal. We will call such process shortly SS (strictly stationary) time series.
As equality of two univariate distributions implies equality of their means and equality of two bivariate distributions ensures equality of their covariances (if means and covariances exist) we have that that strictly stationary time series such that its elements are square integrable is weakly stationary. Thus, for all practical purposes, strict stationarity entails weak stationarity.
In the case of strictly stationary time series observations $X_1, \ldots, X_n$ have the same distribution and its parameters can be estimated based on one trajectory $X_1(\omega), \ldots, X_n(\omega)$ when n is large. Technical difficulty here which does not occur for i.i.d. sequences is that the observations are possibly dependent and usually

for the sake of inference additional properties such as ergodicity or more specific structure of time series has to be imposed. In the case of weakly stationary time series one can estimate its invariant characteristics such as mean, variance and covariance and their parameters such as spectral distribution.
In order to appreciate the difference between weak and strict stationarity consider a sequence of independent variables $(X_t)_{t\in\mathbb{Z}}$ such that $X_{2t} \overset{\mathcal{D}}{\sim} N(0,2)$ and $X_{2t+1} \overset{\mathcal{D}}{\sim} \chi^2_1 - 1$, where $\chi^2_1 - 1$ denotes centred χ^2 distribution with one degree of freedom. Such time series is weakly stationary as its mean is 0 and its covariance function $\gamma_X(s,t) = 2I\{s = t\}$ depends only on $s-t$ but it is obviously not strictly stationary as marginal distribution for even and odd indices differ.
Frequently a scale invariant version of autocovariance function is used. Namely, let $(X_t)_{t\in\mathbb{Z}}$ be WS time series such that $\gamma_X(0) > 0$ i.e. variables X_t are not equal to the mean of the process. An autocorrelation function (abbreviated to ACF) is defined as

$$\rho_X(h) = \rho(X_{t+h}, X_t) = \frac{\gamma_X(h)}{\{\gamma_X(0)\gamma_X(0)\}^{1/2}} = \frac{\gamma_X(h)}{\gamma_X(0)}. \tag{1.8}$$

It follows easily form the properties of the autocovariance function that $\rho_X(h) = \rho_X(-h)$ and $|\rho_X(h)| \leqslant \rho_X(0) = 1$. The plot below shows empirical autocorrelation function discussed further for residuals of the quadratic fit for `uspop.dat`. Confidence intervals depicted there correspond to white noise; their form will be derived in Chapter 7. Note that the value of empirical ACF for $h = 0$, similarily to $\rho(0)$, is also always 1.



Example 1.3.2 *We stress that two different stationary time series may have the same autocorrelation function. Namely, let Y_t be arbitrary zero mean WS time series and consider two WS processes defined as*

$$X_t = Y_t - \phi Y_{t-1}$$
$$W_t = Y_t - \phi^{-1} Y_{t-1},$$

where ϕ is an arbitrary constant different from 1. ACFs for both processes coincide. Indeed,

$$\gamma_X(h) = \gamma_Y(h) - \phi\gamma_Y(h-1) - \phi\gamma_Y(h+1) + \phi^2\gamma_Y(h)$$
$$\gamma_W(h) = \gamma_Y(h) - \phi^{-1}\gamma_Y(h-1) - \phi^{-1}\gamma_Y(h+1) + \phi^{-2}\gamma_Y(h).$$

and as one easily checks that as $\gamma_W(0) = \gamma_X(0)/\phi^2$, $\rho_X(h) = \rho_W(h)$. Obviously for $\phi \neq 1$ we have that $X_t \neq W_t$.

1.3.1 Main examples

We discuss now the main classes of weakly and strongly stationary processes which be frequently used in this book. A building block of many models of time series is a white noise process.

Example 1.3.3 *(white noise) Example above (1.8) is a special case of the situation when $(X_t)_{t\in\mathbb{Z}}$ is a sequence of uncorrelated (i.e. $\mathrm{Cov}(X_s, X_t) = 0$ for $r \neq s$) with mean m and variance $\sigma^2 > 0$. This is obviously weakly stationary time series as $\gamma_X(r,s) = \sigma^2 \cdot \delta_{rs}$, where $\delta_{rs} = I\{r \neq s\}$ and is strictly stationary if variables are independent. When $m = 0$ $(X_t)_{t\in\mathbb{Z}}$ is called (weak) white noise in the case of uncorrelated random variables and strong white noise when they are independent. The white noise will be abbreviated to WN$(0, \sigma^2)$ and denoted either by (ε_t) or (Z_t).*

We note that in the literature notation $\mathrm{WN}(0, \sigma^2)$ may refer to either a weak or a strong white noise. Here, unless stated otherwise, it will always refer to a *weak* white noise.
Linear process discussed below is a versatile tool used to model stationary time series. It is defined as a discrete convolution of white noise process.

Example 1.3.4 *(linear process) Let $(\varepsilon_t)_{t\in\mathbb{Z}}$ be WN$(0, \sigma^2)$ and $\psi_j \in \ell^2$, meaning that $\sum_{j=-\infty}^{\infty} \psi_j^2 < \infty$, be a fixed doubly-infinite sequence. Linear process is defined by*

$$X_t = \sum_{j=-\infty}^{\infty} \psi_j \varepsilon_{t-j}, \qquad t \in \mathbb{Z}. \tag{1.9}$$

It can be easily shown that the right hand side belongs to $\mathcal{L}^2$ for each t; to this end it is enough to convince oneself that its partial sums form a Cauchy sequence and use the fact that every Cauchy sequence is convergent in $\mathcal{L}^2$. Moreover, one can check that $EX_t = 0$. Also, continuity of scalar product $< \cdot, \cdot >$ implies that

$$< X_{t+k}, X_t >= \lim_{n,m\to\infty} < \sum_{i=-m}^{m} \psi_i \varepsilon_{t+k-i}, \sum_{j=-n}^{n} \psi_j \varepsilon_{t-j} >$$

$$= \lim_{n,m\to\infty} \sum_{i=-m}^{m} \sum_{j=-n}^{n} \psi_i \psi_j < \varepsilon_{t+k-i}, \varepsilon_{t-j} >= \sigma^2 \sum_{i=-\infty}^{\infty} \psi_i \psi_{i-k} = \sigma^2 \sum_{i=-\infty}^{\infty} \psi_i \psi_{i+k},$$

where we used the observation that $< \varepsilon_{t+k-i}, \varepsilon_{t-j} >\neq 0$ only for $j = i - k$ and the equality

$$\sum_{i=-\infty}^{\infty} \psi_i \psi_{i-k} = \sum_{i=-\infty}^{\infty} \psi_i \psi_{i+k}.$$

Thus $\gamma_X(s,t)$ depends on $s - t$ only. Whence $(X_t)_{t\in\mathbb{Z}}$ is WS time series. When (ε_t) is strong $WN(0,\sigma^2)$ then X_t is SS time series.

When $\psi_j = 0$ for $j < 0$ in (1.9) then linear process becomes one-sided linear process or moving average of infinite order denoted by MA(∞)

$$X_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}, \qquad t \in \mathbb{Z}. \tag{1.10}$$

In this case X_t depends on ε_s for $s \leqslant t$ only and moreover $X_t \in H_t(\varepsilon)$. Frequently used case of one-sided linear process is obtained when $\psi_j = 0$ for $j > q$. Then

$$X_t = \sum_{i=0}^{q} \psi_i \varepsilon_{t-i}.$$

In traditional notation used for ARMA processes coefficients ψ_i are usually named θ_i and with $c_0 = 1$ one has

$$X_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}. \tag{1.11}$$

Such a process is called the moving average of order q and denoted by MA(q). Observe that if (ε_t) is a strong $WN(0,\sigma^2)$, then $MA(q)$ is q-dependent that is for $|h| > q$ variables X_t i X_{t+h} are independent. In particular, when $q = 0$ MA(q) reduces to white noise process described in (i). Linear process is a very versatile model of time series which is frequently used in modelling and theoretical considerations. Main reason for this is that by choosing appropriate sequence (ψ_j) we can approximate given dependence structure. One of the possible generalizations of the linear process is Volterra series . Namely, suppose that for certain functions $g_k(u_1, \ldots, u_k)$ called Volterra kernels the following expression is well defined

$$\sum_{k=1}^{\infty} \sum_{u_1,\ldots,u_k=0}^{\infty} g_k(u_1,\ldots,u_k)\varepsilon_{t-u_1}\cdots\varepsilon_{t-u_k} \tag{1.12}$$

where (ε_t) is a strong white noise. Note that when all functions g_k for $k > 1$ are 0, expression (1.12) yields a linear process. In general conditions on $g_k(\cdot)$ for which (1.12) is meaningful can be quite restrictive.

Example 1.3.5 *(nonlinear autoregression) Very intuitive dynamics of time series is given by linear autoregression AR(p) of order $p \in \mathbb{N}$, namely*

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + \varepsilon_t, \tag{1.13}$$

where (ε_t) is WN$(0, \sigma^2)$. From the formal point of view (1.13) is nothing else then regression equation with X_t as a dependent variable and lagged variables $X_{t-1}, \ldots, X_{t-p}$ as the predictors. Stationary solutions to (1.13) will be discussed in Chapter 4. Of course, linear dependence of X_t on $X_{t-1}, \ldots, X_{t-p}$ is a serious restriction of the model and generalizations of the linear autoregressive structure were sought.

One of the possibilities is nonparametric autoregressive conditionally heteroscedastic model (NARCH) which stipulates that

$$X_t = f(X_{t-1}, \ldots, X_{t-p}) + \sigma(X_{t-1}, \ldots, X_{t-p})\varepsilon_t,$$

where f and σ are some functions parametric form of which is not known or does not exist. Although flexible, this model is very hard to estimate for large p due to curse of dimensionality: enormously large samples are needed to discern characteristic features of both unknown functions. Thus we fall between Scylla of model misspecification and Charybdis of unidentifiability. Much more modest and reasonable model than the last one is additive autoregression (AAR)

$$X_t = f_1(X_{t-1}) + \cdots + f_p(X_{t-p}) + \varepsilon_t, \tag{1.14}$$

the autoregressive 'brother' of an additive model, which avoids 'curse of dimensionality'. Functions $f_1, \ldots, f_p$ can be estimated using backpropagation algorithm, see e.g. Wood (2006).
Another model being nonlinear generalization of AR(1) time series is defined as

$$X_t = R(X_{t-1}, \varepsilon_t), \tag{1.15}$$

where R is a measurable function. Conditions under which (1.15) has a strictly stationary solution has been studied in Diaconis and Freedman (1999). They have shown that the following conditions are sufficient: there exists $\alpha > 0$ and x_0 such that

$$E(\log L_\varepsilon) < 0 \quad \text{and} \quad L_\varepsilon + |R(x_0, \varepsilon_0)| \in \mathcal{L}^\alpha, \tag{1.16}$$

where

$$L_\varepsilon = \sup_{x \neq x'} \frac{|R(x,\varepsilon) - R(x',\varepsilon)|}{|x - x'|}.$$

The first part of condition (1.16) asserts that on average a Lipschitz constant of the function $R(\cdot, \varepsilon)$ is strictly smaller than 1. Note that, intuitively, when stationary solution exists it is of the form $X_t = g(\ldots, \varepsilon_{t-1}, \varepsilon_t)$. Measures of dependence for such processes will be studied in Chapter 2.

1.4 Problems

1. Prove that:
(i) a linear process defined in (1.9) is well defined i.e. infinite sum defining X_t is convergent in $\mathcal{L}^2$ (to this end check that $X_t^n = \sum_{i=-n}^{n} \psi_j \varepsilon_{t-j}$ is a Cauchy sequence in $\mathcal{L}^2$) ;
(ii) $EX_t = 0$.
2. Define harmonic process

$$X_t = \sum_{j=-\infty}^{\infty} \psi_j e^{i\lambda_j t} \varepsilon_t,$$

where $(\psi_j) \in \ell^2$, $\psi_j \in \mathbb{C}$, $\lambda_j \in \mathbb{R}$, and (ε_t) is $\mathrm{WN}(0, \sigma^2)$. Justify the following statements:
(i) X_t is well defined and $EX_t = 0$;
(ii) Check that (X_t) is weakly stationary by calculating $\gamma_X(s,t)$.
3. Let (Y_t) be defined as

$$Y_t = m + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_{12} \varepsilon_{t-12},$$

where ε_t is $\mathrm{WN}(0, \sigma^2)$. Check that it is weakly stationary and find the autocovariance function of this process.
4. Show that for a linear process defined in (1.9) $\gamma(h) \to 0$ when $h \to \infty$. Produce an example of a weakly stationary process which does not have this property.
5.5. Let (X_t) be mean-zero stationary process. Find $\bar{A}, \bar{B}$ such that $(\bar{A}, \bar{B}) = \mathrm{argmin}_{A,B} F_h(A,B)$, where for $h \in \mathbb{Z}$

$$F_h(A,B) = E(X_{t+h} - A - BX_t)^2,$$

by two methods:
(i) analytically, by finding the stationary point of criterion function $F_h(A,B)$;
(ii) geometrically, by interpreting $A + BX_t$ as the perpendicular projection of X_{t+h} onto subspace $sp(1, X_t)$.
6. Let X, Z be independent standard normal $N(0,1)$ random variables and let $Y = X^{2k} + Z$, where $k \in \mathbb{N}$. Find the optimal linear predictor of Y based on X and $E(Y|X)$ (i.e. the best predictor among all square integrable functions of X) and the respective prediction errors. Interpret the result.

7. Let (X_t) be a mean 0 Gaussian process. Check that $E(X_{t+h}|X_t) = \rho(h)X_t$, that is

$$\rho(h)X_t = \inf_g E(X_{t+h} - g(X_t))^2,$$

where infimum is taken over all functions g such that $g(X_t)$ is square integrable. Compare the result with the result of problem 6.

8. Prove that if $(X, X_1, \ldots, X_n)$ has a joint multivariate normal distribution with mean 0, then optimal predictor of X based on $X_1, \ldots, X_n$ namely $E(X|X_1, \ldots, X_n)$ is linear, that is

$$E(X|X_1, \ldots, X_n) = a_1X_1 + \ldots + a_nX_n$$

for certain $a_1, \ldots, a_n \in \mathbb{R}$.

Hint. Consider $a_1X_1 + \ldots + a_nX_n$ which is the orthogonal projection of X on $sp(X_1, \ldots, X_n)$ and note that $< X - \sum_{i=1}^n a_iX_i, X_k = 0 >$ for $k = 1, \ldots, n$, thus $X - \sum_{i=1}^n a_iX_i$ and X_k are independent.

9. Consider an iid sequence (X_i) with finite second moment and let $S_n := \sum_{i=1}^n X_i$. (i) Prove that $\mathrm{Cov}(S_k, S_l) = \min(k, l)\mathrm{Var}(X_1)$. (ii) Prove that S_n is not weakly stationary unless $X_i \equiv 0$.

2

Quantification of dependence for time series

We discuss here main standard measures of dependence such as moments, cumulants and mixing coefficients as well as a concept of ergodicity. Moreover, predictive and functional measures are introduced as well as measures stemming from information theory. We show how such measures intervene in statements of the Central Limit Theorem for dependent data. For more extensive discussion of relevant subjects we refer to Rosenblatt (1985), Doukhan and Louhichi (1999) and Ibragimow and Linnik (1971).

2.1 Moments and cumulants

Autocovariance of two random variables is (unscaled) measure of their linear dependence. Below we define cumulants which generalize concept of covariance to many random variables and shortly describe their relation to moments. Let $\mathbf{X} = (X_1, \ldots, X_k)'$ and

$$m_{\mathbf{X}}^{(\nu_1,\cdots,\nu_k)} = E(X_1^{\nu_1} \cdots X_k^{\nu_k}).$$

Consider expansion of characteristic function $\varphi_{\mathbf{X}}$ of random vector $\mathbf{X}$

$$\begin{aligned}\varphi_{\mathbf{X}}(t_1, \ldots, t_k) = E \exp\{i\mathbf{t}'\mathbf{X}\} = \sum_{\nu_1+\cdots+\nu_k \leqslant n} \frac{i^{\nu_1+\nu_2+\cdots+\nu_k}}{\nu_1! \cdots \nu_k!} m_{\mathbf{X}}^{(\nu_1,\ldots,\nu_k)} t_1^{\nu_1} \cdots t_k^{\nu_k} \\ + \ o(|\mathbf{t}|^n) = \sum_{|\boldsymbol{\nu}| \leqslant n} \frac{i^{\boldsymbol{\nu}}}{\boldsymbol{\nu}!} m_{\mathbf{X}}^{(\nu)} \mathbf{t}^{\boldsymbol{\nu}} \ + \ o(|\mathbf{t}|^n),\end{aligned}$$

We consider parallel expansion of $\log \varphi_{\mathbf{X}}(t_1, \ldots, t_k)$

$$\begin{aligned}\log \varphi_{\mathbf{X}}(t_1, \ldots, t_k) = \sum_{\nu_1+\cdots+\nu_k \leqslant n} c_{\mathbf{X}}^{(\nu_1,\ldots,\nu_k)} \frac{i^{\nu_1+\nu_2+\cdots+\nu_k}}{\nu_1! \ldots \nu_k!} t_1^{\nu_1} \ldots t_k^{\nu_k} + \ o(|\mathbf{t}|^n) = \\ = \sum_{|\boldsymbol{\nu}| \leqslant n} \frac{i^{\boldsymbol{\nu}}}{\boldsymbol{\nu}!} c_{\mathbf{X}}^{(\nu)} \mathbf{t}^{\boldsymbol{\nu}} \ + \ o(|\mathbf{t}|^n),\end{aligned}$$

where

$$\mathbf{t}^{\boldsymbol{\nu}} = t_1^{\nu_1} \ldots t_k^{\nu_k}, \quad \boldsymbol{\nu} = \nu_1! \ldots \nu_k!, \quad |\boldsymbol{\nu}| = \nu_1 + \cdots + \nu_k.$$

Definition 4 *Cumulant of* $X_1^{\nu_1}, \dots, X_k^{\nu_k}$ *is defined as* $c_{\mathbf{X}}^{(\nu_1,\dots,\nu_k)}$ *. In particular cumulant of* $X_1, \dots, X_k$ *is*

$$\operatorname{cum}(X_1, \dots, X_k) = c_{\mathbf{X}}^{(1,\dots,1)}$$

that is coefficient related to $t_1 \cdot t_2 \cdots t_k$ *in expansion of* $\log \varphi(\mathbf{t})$ *when the term* i^k *is omitted.*

Now we consider correspondence between cumulants and moments. We can write

$$\varphi_{\mathbf{X}}(\mathbf{t}) = \exp(\log \varphi_{\mathbf{X}}(\mathbf{t})) = \sum_{q=0}^{\infty} \frac{1}{q!} \big\{ \log \varphi_{\mathbf{X}}(\mathbf{t}) \big\}^q$$

and comparing coefficients on both sides we have

$$m_{\mathbf{X}}^{(\boldsymbol{\nu})} = \sum_{q \geqslant 0} \sum_{\boldsymbol{\lambda}^{(1)} + \dots + \boldsymbol{\lambda}^{(q)} = \boldsymbol{\nu}} \frac{1}{q!} \frac{\boldsymbol{\nu}!}{\boldsymbol{\lambda}^{(1)}! \dots \boldsymbol{\lambda}^{(q)}!} \prod_{p=1}^{q} c_{\mathbf{X}}^{(\boldsymbol{\lambda}^{(p)})}$$

where the sum is over tuples $(\boldsymbol{\lambda}^{(1)}, \dots, \boldsymbol{\lambda}^{(q)})$ of *ordered* multi-indices. This means that e.g. tuples $(\boldsymbol{\lambda}^{(1)}, \dots, \boldsymbol{\lambda}^{(q)})$ and $(\boldsymbol{\lambda}^{(q)}, \dots, \boldsymbol{\lambda}^{(1)})$ both contribute to the sum above.
In particular we have

$$E(X_1 \cdots X_k) = \sum_{q \geqslant 1} \sum_{\substack{\text{dfferent partitions } \nu_1 \cdots \nu_q \\ \nu_1 \cup \dots \cup \nu_q = \{1,2,\dots,k\}}} D_{\boldsymbol{\nu}_1} \cdots D_{\boldsymbol{\nu}_q},$$

where $D_{\boldsymbol{\nu}_s} = \operatorname{cum}(X_{\alpha_1}, \dots, X_{\alpha_m})$ and $\{\alpha_1, \dots, \alpha_m\} = \boldsymbol{\nu}_s$. This is obviously due to the fact that number of *ordered* partitions of $\{\nu_1, \dots, \nu_q\}$ equals $q!$ times the number of different partitions.
Expanding analogously $\log E e^{i\mathbf{t}'\mathbf{X}}$, using formula,

$$\log x = \sum_{q \geqslant 1} \frac{(-1)^{q-1}}{q} x^q, \quad |x| < 1$$

we get

$$c_{\mathbf{X}}^{(\boldsymbol{\nu})} = \sum_{q \geqslant 1} \sum_{\substack{\boldsymbol{\lambda}^{(1)} + \dots + \boldsymbol{\lambda}^{(q)} = \boldsymbol{\nu} \\ \text{ordered indices}}} \frac{(-1)^{q-1}}{q} \frac{\boldsymbol{\nu}!}{\boldsymbol{\lambda}^{(1)}! \dots \boldsymbol{\lambda}^{(q)}!} \prod_{i=1}^{q} m_{\mathbf{X}}^{(\boldsymbol{\lambda}^{(i)})}$$

In particular for $\boldsymbol{\nu} = (1, \dots, 1)$ we obtain

$$\operatorname{cum}(X_1, \dots, X_k) = \sum_q \sum (-1)^{q-1} (q-1)! E\big(\prod_{i \in \nu_1} X_i \big) \cdots E\big(\prod_{i \in \nu_q} X_i \big) \tag{2.1}$$

where the second sum ranges over *different* partitions of $\{1, \dots, k\}$.

It follows from (2.1) in particular that $\mathrm{cum}(X_1) = EX_1$ and $\mathrm{cum}(X_1, X_2) = EX_1X_2 - EX_1EX_2 = \mathrm{Cov}(X_1, X_2)$, thus $\mathrm{cum}(X_1, \ldots, X_k)$ can be considered as generalization of covariance. Below we state three most useful properties of cumulants.

Proposition 2.1.1 *(i) If for non-void subset $I \subseteq \{1, \ldots, k\}$ coordinates $\mathbf{X}_I$ are independent of $\mathbf{X}_{I^c}$ then $\mathrm{cum}(X_1, \ldots, X_k) = 0$.*
(ii) Cumulants of order $k > 2$ of multivariate normal distribution are equal 0.
(iii) $\mathrm{cum}(\alpha_1 X_1 + \beta_1 Y_1, X_2, \ldots, X_k) =$

$$= \alpha_1 \mathrm{cum}(X_1, X_2, \ldots, X_k) + \beta_1 \mathrm{cum}(Y_1, X_2, \ldots, X_k).$$

Proof. In order to check (i) observe that if $J = I^c$ we have

$$\log Ee^{i\mathbf{t}'\mathbf{X}} = \log Ee^{i(\mathbf{t}_I'\mathbf{X}_I + \mathbf{t}_J'\mathbf{X}_J)} = \log Ee^{i\mathbf{t}_I'\mathbf{X}_I} + \log Ee^{i\mathbf{t}_J'\mathbf{X}_J}$$

and as the right-hand side does not contain in its expansion the term $t_1 \cdot \ldots \cdot t_k$ this proves (i). Analogously, the characteristic function of multivariate $N(\mathbf{m}, \boldsymbol{\Sigma})$ distribution equals $\varphi(\mathbf{t}) = \exp(\mathbf{t}'\mathbf{m} - \mathbf{t}'\boldsymbol{\Sigma}\mathbf{t})$. And the only terms in the expansion of $\log \varphi(\mathbf{t}) = \mathbf{t}'m - \mathbf{t}'\boldsymbol{\Sigma}\mathbf{t}$ correspond to linear terms t_i and quadratic terms $t_i \cdot t_j$. Part (iii) is obvious.
Part (i) is very useful. Note that the analogous property for moments is *not* satisfied.

Theorem 2.1.2 *(Diagram formula)*
Consider random variables X_{ij}, $i = 1, \ldots, I$, $j = 1, \ldots, J_i$ arranged in an array:

$$\begin{matrix} (1,1) & \ldots & (1,j_1) \\ \vdots & \vdots & \vdots \\ (I,1) & \ldots & (I,j_I) \end{matrix}$$

and

$$Y_i = \prod_{k=1}^{j_i} X_{ik}, \quad i = 1, \ldots, I.$$

Then the following diagram formula holds

$$\mathrm{cum}(Y_1, \ldots, Y_I) = \sum\nolimits^{*} \mathrm{cum}(X_{i_j}, i_j \in \nu_1) \times \cdots \times \mathrm{cum}(X_{i_j}, i_j \in \nu_p)$$

where the sum Σ^* is over all indecomposable partitions of the array i.e. such partitions that sum of elements of its proper sub-partition does not contain whole rows of diagram.

2.2 Ergodicity and mixing

Consider general probability space $(\Omega, \mathcal{A}, P)$ and a measurable transform $T : (\Omega, \mathcal{A}, P) \to (\Omega, \mathcal{A}, P)$. We call T measure preserving transform if induced

measure PT^{-1} coincides with P i.e. for any $A \in \mathcal{A}$ we have $P(\{\omega \in A\}) = PT^{-1}(\{\omega \in A\}) = P(\{\omega : T(\omega) \in A\})$. We define a family $\mathcal{J}$ of sets in $\mathcal{A}$ which are T-invariant i.e. such that $T^{-1}(A) = A$. It is easy to check that $\mathcal{J}$ is σ-algebra which will be called the σ-algebra of T-invariant sets.

Definition 5 *Measure preserving T is called ergodic on $(\Omega, \mathcal{A}, P)$ if for any $A \in \mathcal{J}$ we have $P(A) = P(A)^2$ i.e. $P(A)$ equals 1 or 0. Moreover T is called mixing if for any $A, B \in \mathcal{A}$ we have*

$$P(A \cap T^{-n}(B)) \to P(A)P(B), \qquad n \to \infty.$$

We will now define the analogous properties for time series $(X_t)_{t\in\mathbb{Z}}$. To this end observe that if we define $\mathbf{X}(\omega) = (\ldots, X_{k-1}(\omega), X_k(\omega), X_k(\omega)\ldots)$, $\mathbf{X}$ is a measurable transform of $(\Omega, \mathcal{A})$ to $(R^{\mathbb{Z}}, \mathcal{B}(R^{\mathbb{Z}}))$, where $\mathcal{B}(R^{\mathbb{Z}})$ denotes σ-algebra of Borel sets in $R^{\mathbb{Z}}$. Denote by $P\mathbf{X}^{-1}$, as usual, measure P induced by $\mathbf{X}$. Define now the *right* shift U as

$$(U(\mathbf{x}))_k = x_{k-1},$$

where $\mathbf{x} = (\ldots, x_{k-1}, x_k, x_{k+1}\ldots)$. We call U right shift because if we move an (imaginary) register to the right, value at position k is taken over by the value x_{k-1} which resided at position $k-1$ before. Observe that when $(X_t)_{t\in\mathbb{Z}}$ is weakly stationary than U preserves measure $P\mathbf{X}^{-1}$. Indeed, consider a cylinder C in $\mathcal{B}$ of the form $C = \{\mathbf{x} : x_{t_1} \in A_1, \ldots, x_{t_k} \in A_k\}$, where $A_1, \ldots, A_k$ are Borel sets in $\mathbb{R}$. Then

$$\begin{aligned} &P\mathbf{X}^{-1}(U^{-1}(C)) = P\mathbf{X}^{-1}(\{\mathbf{x} : (U\mathbf{x})_{t_1} \in A_1, \ldots, (U\mathbf{x})_{t_k} \in A_k\}) \\ &= P\mathbf{X}^{-1}(\{\mathbf{x} : x_{t_1-1} \in A_1, \ldots x_{t_k-1} \in A_k\}) \\ &= P(\{\omega : X_{t_1-1} \in A_1, \ldots X_{t_k-1} \in A_k\}) \\ &= P(\{\omega : X_{t_1} \in A_1, \ldots X_{t_k} \in A_k\}) = P\mathbf{X}^{-1}(C), \end{aligned}$$

where the penultimate equality follows from weak stationarity. It is easily seen that this property extends from cylinders to any sets in $\mathcal{B}(R^{\mathbb{Z}})$.

Definition 6 *We call weakly stationary $(X_t)_{t\in\mathbb{Z}}$ ergodic process (respectively, mixing process) if right shift U is ergodic (respectively, mixing) with respect to $P\mathbf{X}^{-1}$.*

Thus it means, in the case of ergodicity, that for any $A \in \mathcal{J}$, where $\mathcal{J}$ are invariant sets for U we have $P(\{\omega : \mathbf{X}(\omega) \in A\}) = 0$ or 1. Analogously, for mixing, we have that

$$P(\{\omega : \mathbf{X}(\omega) \in A \cap U^{-n}B\}) \to P(A)P(B). \tag{2.2}$$

Note that condition (2.2) written for two cylinders $A = \{\mathbf{x} : x_{t_1} \in A_1, \ldots, x_{t_k} \in A_k\}$ and $B = \{\mathbf{x} : x_{w_1} \in B_1, \ldots, x_{w_l} \in B_l\}$ yields

$$\begin{aligned} &P(\{\omega : \mathbf{X}(\omega) \in A \cap U^{-n}B\}) \\ &= P(\{X_{t_1} \in A_1, \ldots, X_{t_k} \in A_k, X_{w_1-n} \in B_1, \ldots, X_{w_l-n} \in B_l\}) \end{aligned}$$

$$\to P(\{X_{t_1} \in A_1, \ldots, X_{t_k} \in A_k)P(X_{w_1} \in B_1, \ldots, X_{w_l} \in B_l\}) \tag{2.3}$$

when $n \to \infty$. Note that the above convergence yields an intuitive interpretation of mixing as a sort of independence condition of the past from the present. Interpretation of ergodicity in similar vein will be given at the end of the section. The term ergodic was coined by G. Birkhoff from Greek words 'ergon' meaning 'work' and 'hodos'-'path'.
We have the following simple properties of ergodic and mixing time series.

Proposition 2.2.1 *(i) If weakly stationary $(X_t)_{t\in Z}$ is mixing it is also ergodic;*
(ii) If $(\varepsilon_t)_{t\in\mathbb{Z}}$ is strong white noise then $(\varepsilon_t)_{t\in\mathbb{Z}}$ is mixing;
(iii) Let $h : R^{\mathbb{N}} \to R$ be measurable and $X_t = h(\ldots, Y_{t-1}, Y_t)$ be one-sided moving function, where Y_t is mixing (ergodic). Then (X_t) is mixing (ergodic).

Proof. In order to prove (i) observe that if $A \in \mathcal{J}$ it follows from (2.2) applied to $A = B$ that $P\mathbf{X}^{-1}(A \cap U^{-n}A) = P\mathbf{X}^{-1}(A) = [P\mathbf{X}^{-1}(A)]^2$ and thus ergodicity condition is satisfied.
To see (ii) consider two cylinders $A = \{\mathbf{x} : x_{t_1} \in A_1, \ldots, x_{t_k} \in A_k\}$ and $B = \{\mathbf{x} : x_{w_1} \in B_1, \ldots, x_{w_l} \in B_l\}$ and note that (2.3) is satisfied for independent (ε_t) with equality replacing convergence provided n is such that $\max(w_k) - n < \min(t_i)$. The property above extends from cylinders to arbitrary measurable A and B in a standard way. Thus, strong white noise is mixing and thus ergodic.
(iii) We prove (iii) for mixing case. Let $\mathbf{Y}_t = (\ldots, Y_t)$ and $\mathbf{x}_t = (\ldots, x_t)$ for $(x_t) \in R^{\mathbb{Z}}$ and consider $B_1, B_2 \in R^{\mathbb{Z}}$. Note that

$$\mathbf{X}^{-1}(B_i) = \{\mathbf{X} \in B_i\} = \{(\ldots, h(\mathbf{Y}_{k-1}), h(\mathbf{Y}_k), h(\mathbf{Y}_{k+1}), \ldots) \in B_i\} = \mathbf{Y}^{-1}(A_i),$$

for $i = 1, 2$, where

$$A_i = \{\mathbf{x} : (\ldots, h(\mathbf{x}_{k-1}), h(\mathbf{x}_k), h(\mathbf{x}_{k+1}), \ldots) \in B_i\}.$$

Moreover note that $\mathbf{X}^{-1}(U^{-n}(B_i)) = \mathbf{Y}^{-1}(A_i^{(n)})$, where

$$\begin{aligned}
&A_i^{(n)} = \{\mathbf{x} \in R^{\mathbb{Z}} : (\ldots, h(\mathbf{x}_{k-1}), h(\mathbf{x}_k), h(\mathbf{x}_{k+1}), \ldots) \in U^{-n}(B_i)\} = \\
&\{\mathbf{x} \in R^{\mathbb{Z}} : U^n(\ldots, h(\mathbf{x}_{k-1}), h(\mathbf{x}_k), h(\mathbf{x}_{k+1}), \ldots) \in B_i\} = \\
&\{\mathbf{x} \in R^{\mathbb{Z}} : (\ldots, h(\mathbf{x}_{k-1-n}), h(\mathbf{x}_{k-n}), h(\mathbf{x}_{k+1-n}), \ldots) \in B_i\} = \\
&\{\mathbf{x} \in R^{\mathbb{Z}} : (\ldots, h(U^n(\mathbf{x}_{k-1})), h(U^n(\mathbf{x}_k)), h(U^n(\mathbf{x}_{k+1})), \ldots) \in B_i\} = \\
&\{\mathbf{y} \in R^{\mathbb{Z}} : U^n(\mathbf{y}) \in A_i\} = U^{-n}(A_i).
\end{aligned}$$

Thus using the equality above and the fact that (Y_t) is mixing

$$\begin{aligned}
&P\mathbf{X}^{-1}(B_1 \cap U^{-n}(B_2)) = P(\mathbf{X}^{-1}(B_1) \cap \mathbf{X}^{-1}(U^{-n}(B_2)) = P\mathbf{Y}^{-1}(A_1 \cap A_2^{(n)}) = \\
&P\mathbf{Y}^{-1}(A_1 \cap U^{-n}(A_2)) \to P\mathbf{Y}^{-1}(A_1)P\mathbf{Y}^{-1}(A_2) = P\mathbf{X}^{-1}(B_1)P\mathbf{X}^{-1}(B_2).
\end{aligned}$$

Note in particular that if $X_t = \sum_{i=0}^{\infty} a_i\varepsilon_{t-i}$ is one-sided moving average when (ε_t) is strong white noise, then using (i) and (iii) (X_t) is mixing and thus ergodic.

It also follows, what can be also proved directly, that ergodicity and mixing are preserved for $(g(X_t))$ where g is any measurable transform. Observe that events of the form $\{\liminf_{t\to\infty} X_t \in A\}$, $\{\limsup_{t\to\infty} X_t \in A\}$ are invariant with respect to the right shift and thus if (X_t) is ergodic, they have probability 0 or 1.
We will now state ergodic theorem proved by G. Birkhoff in 1932 from which in particular it follows that for ergodic time series its sample means converge to expected value of the marginal distribution. To this end note that if T is a measure-preserving on $(\Omega, \mathcal{A}, P)$ and X is square-integrable real-valued random variable on this space it follows that

$$X_t = X \circ T^t \qquad t = 1, 2, \ldots \tag{2.4}$$

is a stationary sequence. On the other hand if $\mathbf{Y} = (Y_t)_{t\in\mathbb{N}}$ is stationary time series it can be proved that there exist probability space and X and T as above defined on it that $\mathbf{Y} = \mathbf{X}$ in distribution, where coordinates of $\mathbf{X}$ are $X_t = X \circ T^t$. This has very important consequences for studying asymptotic properties of estimates as convergence results for $\mathbf{X}$ can be carried over for $\mathbf{Y}$. Thus, as it is done in ergodic theorem, it is enough to study asymptotic properties of stationary sequences given by (2.4). Observe also that if X_0 is random variable defined on $(R^{\mathbb{Z}}, \mathcal{B}(R^{\mathbb{Z}}))$ given by $X_0(\mathbf{x}) = x_0$ and we define $T = U^{-1}$ where U is the right shift, then X_t defined in (2.4) is given by $X_t(\mathbf{x}) = X_0 \circ U^{-t} = x_t$. This is called canonical representation of a stationary time series.
We define $S_t = t^{-1} \sum_{j=1}^{t} X_j$.

Theorem 2.2.2 *(Ergodic theorem) For X_t given in (2.4) we have*

$$\frac{S_t}{t} \to E(X_0|\mathcal{J}) \qquad a.s. \tag{2.5}$$

and

$$E\Big|\frac{S_t}{t} - E(X_0|\mathcal{J})\Big| \to 0. \tag{2.6}$$

In particular, for ergodic (X_t) $E(X_0|\mathcal{J})$ can be replaced by EX_0. As for any measurable A, $I\{X_t \in A\}$ is ergodic when (X_t) is ergodic, we have that

$$n^{-1} \sum_{t=1}^{n} I\{X_t \in A\} \to P(\{X \in A\}) = P\mathbf{X}^{-1}(A) \tag{2.7}$$

almost surely, that is frequency of visits to A converges to $P\mathbf{X}^{-1}(A)$ a.s. regardless of ω.
In the case of the one-sided linear process $X_t = \sum_{i=0}^{\infty} \psi_i Y_{t-i}$ with $(\psi_i) \in \ell^1$ and ergodic (Y_t) we have almost surely

$$n^{-1} \sum_{t=1}^{n} X_t = n^{-1} \sum_{k=1}^{n} \sum_{i=0}^{\infty} \psi_i Y_{t-i} \to EY_1 \sum_{i=0}^{\infty} \psi_i. \tag{2.8}$$

Ergodic theorem is frequently informally stated as follows: for ergodic time series averaging over trajectory and over sample space gives the same result in the sense

that for large sample sizes the sample mean is close to the expected value of the marginal distribution.
Ergodic theorem also yields nice equivalent definition of ergodicity.

Corollary 2.2.3 *A measure-preserving T on $(\Omega, \mathcal{A}, P)$ is ergodic if and only if*

$$\frac{1}{t}\sum_{j=1}^{t} P(A_1 \cap T^{-i}A_2) \to P(A_1)P(A_2) \qquad A_1, A_2 \in \mathcal{A}.$$

Comparing the last convergence with the definition of mixing we immediately see why mixing is a stronger concept than ergodicity. It is due to simple analytic fact that Cesàro means of convergent sequences are convergent to the same limit.

2.3 Mixing conditions

In this section we define several mixing conditions and mixing coefficients which quantify dependence between events which occur at lags at least n. The difference between them is that they quantify in a slightly different ways the departure from independence. The basic idea of mixing conditions is to ensure that the distant past and the distant future become asymptotically independent. One should realize that since we measure departures from independence, by imposing conditions on mixing coefficients we can only arrive at the same phenomena which hold for iid sequences. The name 'mixing' may be confusing as we have already defined mixing (in the ergodic sense). We will show that the weakest of the mixings introduced below, α-mixing implies mixing in the ergodic sense. Thus strong of large numbers holds under mixing. However, central limit theorems for mixing sequences impose unverifiable conditions on mixing coefficients. Let us law note that even checking mixing, apart from special cases such as Gaussian sequences or Markov series (for which more specific methods to establish asymptotic results exist) is practically impossible. We refer to Bradley (2005) for extensive discussion of topics below and the bibliography.
Let $(X_t)_{t\in\mathbb{Z}}$ be a *strictly* stationary time series and let $\mathcal{F}_i^j$ for $i \leqslant j$ be a sigma-algebra generated by variables $X_i, \ldots, X_j$, $\mathcal{L}^2(\mathcal{F}_i^j)$ a space of square integrable rvs measurable with respect to $\mathcal{F}_i^j$. We define the following mixing coefficients

$$\alpha(n) = \sup_{A\in\mathcal{F}_{-\infty}^0, B\in\mathcal{F}_n^\infty} |P(A\cap B) - P(A)P(B)|,$$

$$\beta(n) = \sup_{A_1,\ldots,A_I,B_1,\ldots,B_J} \frac{1}{2}\sum_{i=1}^{I}\sum_{j=1}^{J} |P(A_i\cap B_j) - P(A_i)P(B_j)|,$$

where the supremum is taken over all finite partitions $A_1, \ldots, A_I$ and $B_1, \ldots, B_J$ of Ω such that $A_i \in \mathcal{F}_{-\infty}^0$ $i = 1, \ldots, I$, $B_j \in B \in \mathcal{F}_n^\infty$, $j = 1, \ldots, J$.

$$\phi(n) = \sup_{A\in\mathcal{F}^0_{-\infty}, B\in\mathcal{F}^\infty_n, P(A)>0} |P(B) - P(B|A)|,$$

$$\rho(n) = \sup_{X\in\mathcal{L}^2(\mathcal{F}^0_{-\infty}), Y\in\mathcal{L}^2(\mathcal{F}^\infty_n)} |\mathrm{cor}(X,Y)|.$$

Coefficient $\rho(n)$ is called the maximal correlation between $\mathcal{F}^0_{-\infty}$ and $\mathcal{F}^\infty_n$. As $\mathcal{F}^j_i$ intuitively contains all information about the process between time points i and j the coefficients defined above quantify dependence between events described in terms of knowledge gathered on the process in intervals $(-\infty, 0]$ on $[n, \infty)$. Observe that as we assumed strict stationarity of time series the reference point 0 is irrelevant and all it counts is the length of the gap n between those intervals. Note that all mixing coefficients are monotonically nonincreasing and contained between 0 and 1. If any of them is equal to 0 this is equivalent to independence of the respective sigma-algebras.

We say that $(X_t)_{t\in\mathbb{Z}}$ is α-mixing if $\alpha(n) \to 0$ when $n \to \infty$ with the analogous wording applied to other coefficients. We remark that α-mixing is sometimes called , which is somewhat misleading, as we shall see, it is the weakest concept considered among the four introduced. Also, β-mixing sequences are also called absolutely regular.

We now discuss the interplay between the mixing coefficients. Note that for $A \in \mathcal{F}^0_{-\infty}, B \in \mathcal{F}^\infty_n$ we have

$$\begin{aligned} 4|P(A\cap B) - P(A)P(B)| &\leqslant \frac{|P(A\cap B) - P(A)P(B)|}{[(P(A)(1-P(A))(P(B)(1-P(B))]^{1/2}} \\ &= |\mathrm{cor}(I_A, I_B)| \end{aligned}$$

and this proves left hand side of the inequality

$$\alpha(n) \leqslant \frac{1}{4}\rho(n) \leqslant \frac{1}{2}\phi^{1/2}(n). \tag{2.9}$$

The right-hand side is proved in Coqburn (1960). From the above inequalities it follows that ϕ-mixing implies ρ-mixing and ρ-mixing implies α-mixing in its turn. Also, absolute regularity implies α-mixing, thus α-mixing is the weakest property among the four introduced. However, the weakest concept of α-mixing still implies mixing in ergodic sense defined above which follows in a straightforward way from the respective definitions. Moreover, ϕ-mixing implies β-mixing in view of the inequality $\beta(n) \leqslant \phi(n)$. It is also known that ρ-mixing and absolute regularity are incompatible in the sense that neither of them implies the other. For Gaussian time series ρ-mixing is equivalent to α-mixing.

There are some conceptual drawbacks of the weakest and the strongest of these measures. Observe that the reason $\alpha(n)$ is small may be due to the fact that the difference $|P(A\cap B) - P(A)P(B)|$ is small which is the effect we would like to quantify or to more trivial reason that simply one of the probabilities $P(A)$ or $P(B)$ is small, which we would like to disregard. This was the main reason of studying the remaining coefficients in particular ϕ mixing coefficient based on

quantity $|P(A \cap B) - P(A)P(B)|/P(A)$. However, ϕ-mixing is sometimes too strong a property as for Gaussian sequences it implies independence of X_t and X_{t+k} for sufficiently large k. (cf. Ibragimow and Rozanov (1978)). The advantage of the mixing coefficients is that if we consider transformed time series $f(X_t)$, its mixing coefficients are not larger than those for the original sequence. Moreover, it is easily that e.g. if $(X_t)_{t\in\mathbb{Z}}$ is α-mixing then $(f(X_t, X_{t-1}, \ldots, X_{t-p}))_{t\in\mathbb{Z}}$ for fixed p also has this property. Somewhat disappointingly $AR(1)$ process with innovations having Bernoulli distribution and $\phi = 1/2$ is not α-mixing (cf. Andrews (1984)). We consider also some special cases:
(i) if $(X_t)_{t\in\mathbb{Z}}$ is strictly stationary Markov sequence than it follows that we can replace in the definition of mixing coefficients the pair of sigma-algebras $(\mathcal{F}^0_{-\infty}, \mathcal{F}^\infty_n)$ by the pair $(\mathcal{F}^0_0, \mathcal{F}^n_n)$ i.e. sigma-algebras generated by a single random variables. Moreover, it is known that α-mixing is implied by convergence to 0 of the integral $\int |f_{0,n}(x,y) - f_0(x)f_n(y)|\,dxdy$, where the integrand is the absolute value of the difference between joint density of (X_0, X_n) and the product of marginal densities. It is also known that for a stationary Markov chain with a finite state space, mixing in ergodic sense is equivalent to α-mixing and is also equivalent to the fact that the chain is irreducible and aperiodic. Remarkably, condition $\phi(n) < 1$ implies ϕ-mixing with exponential rate of decay of the sequence $(\phi(n))_{n\in\mathbb{N}}$(we refer to Bradley (2005), p. 119 for the discussion).
(ii) For stationary Gaussian sequences, which are thoroughly treated in Ibragimow and Rozanov (1978), we have already mentioned that ϕ-mixing implies independence for sufficiently lagged observations. Moreover $\rho(n)$, the maximal correlation coefficient, reduces in this case to correlation $\mathrm{Cor}(X_0, X_n)$. Note that this in particular implies that a stationary Gaussian time series such that $\gamma(n) \to 0$ is ergodic (as it is α-mixing and thus mixing). An interesting inequality in the Gausssian case was proved by Kolmogorov and Rozanov, who have shown that (Kolmogorov and Rozanov (1960))

$$\rho(n) \leqslant 2\pi\alpha(n).$$

Thus in the stationary Gaussian case α-mixing implies ρ-mixing.

2.4 Another look at quantification of dependence

In this section we discuss another approach to quantify dependence for important subclass of strictly stationary processes defined in (2.15).

2.4.1 Method of projections

We discuss first a general method which avails itself of representation of a partial sum of strictly stationary sequence as an infinite sum of *uncorrelated* random

variables, which is called here method of projections. Its variants appear in Hannan (1979) and Ho and Hsing (1980) and it is a main tool to study the properties of predictive and functional measures of dependence introduced below. It can be also used to study nonparametric estimators of functional parameters such as marginal density or regression function for dependent data. The framework is quite general and can be shortly summarized as follows. Consider a strictly stationary square integrable sequence $(V_t)_{t\in\mathbb{Z}}$ such that V_t is measurable with respect to σ-algebra $\mathcal{F}_t$, where $(\mathcal{F}_t)_{t\in\mathbb{Z}}$ is increasing sequence of σ-algebras such that $\cap_{i=-\infty}^{t}\mathcal{F}_t$ is trivial. Then using $E(V_t|\mathcal{F}_{-j}) \to E(V_t|\cap_{i=-\infty}^{t}\mathcal{F}_t) = E(V_t)$ we obtain

$$V_t - EV_t = \sum_{k=-\infty}^{t} E(V_t|\mathcal{F}_k) - E(V_t|\mathcal{F}_{k-1}). \tag{2.10}$$

Note that projection of any square integrable U on $\mathcal{F}_k$ can be represented as a sum of a projection on $\mathcal{F}_{k-1}$ and a projection on its orthogonal complement

$$E(U|\mathcal{F}_k) = E(U|\mathcal{F}_{k-1}) + E(U|\mathcal{F}_k) - E(U|\mathcal{F}_{k-1}) := E(U|\mathcal{F}_{k-1}) + \mathcal{P}_k U$$

and thus it follows from (2.10) that

$$\sum_{t=1}^{n}(V_t - EV_t) = \sum_{t=1}^{n}\sum_{k=-\infty}^{t}\mathcal{P}_k V_t. \tag{2.11}$$

Provided that $\sum_t\sum_k ||\mathcal{P}_k V_t|| < \infty$ we can regroup terms in (2.11) and obtain

$$\sum_{t=1}^{n}(V_t - EV_t) = \sum_{k=-\infty}^{n}\mathcal{P}_k\sum_{t=1}^{n}V_t := \sum_{k=-\infty}^{n}U_{n,k}, \tag{2.12}$$

where $U_{n,k}$ and $U_{n,k'}$ are uncorrelated for $k \neq k'$. Using stationarity and orthogonality of components we obtain

$$||\sum_{t=1}^{n}(V_t - EV_t)||^2 = \sum_{k=-\infty}^{n}||U_{n,k}||^2 \leqslant \sum_{k=-\infty}^{n}\Big(\sum_{t=\max\{1,k\}}||P_1 V_{t-k+1}||\Big)^2. \tag{2.13}$$

In particular, if $||\mathcal{P}_1 V_t|| \leqslant \theta_t$ for $t \geqslant 1$ the bound in (2.13) is not larger than

$$\begin{aligned}&\sum_{k=-\infty}^{0}\Big(\theta_{t-k+1}\Big)^2 + \sum_{k=1}^{n}\Big(\theta_{t-k+1}\Big)^2 \leqslant \sum_{k=1}^{\infty}\Big(\sum_{t=1}^{n}\theta_{t+k}\Big)^2 + n\Theta_n^2\\ &= \sum_{k=1}^{\infty}(\Theta_{n+k} - \Theta_k)^2 =: \Xi_n^2,\end{aligned} \tag{2.14}$$

where $\Theta_n = \sum_{i=1}^{n}\theta_i$. Note that if (θ_i) are summable than $\Xi_n^2 = O(n)$ as $\sum_{k=1}^{\infty}(\Theta_{n+k} - \Theta_k)^2 = n(\sum_{i=1}^{n}\theta_i)^2$.

We note that Hannan (1979) used representation (2.11) in the case when $V_t = G(\varepsilon_t)$ and (ε_t) is stationary ergodic Markov chain. He proved that provided that $\sigma^2 = \sum_{t=0}^{\infty}||\mathcal{P}_0 V_t||^2 < \infty$ then $n^{-1/2}\sum_{t=1}^{n}V_t$ converges in distribution to $N(0,\sigma^2)$. We will discuss the genralisation of this result in the next section.

2.4.2 Predictive and functional measures of dependence

There are many other measures of dependence for time series we mention in particular weak dependence coefficients introduced in Dedecker et al. (2007) and projective measures of dependence Gordin (1969). Here we discuss a functional and predictive measure introduced by Wu (2005) which satisfy two natural criteria: their definition applies to a broad class of time series and they can be effectively calculated for interesting subclasses of this class. The considered class consists of strictly stationary processes defined by

$$X_t = g(\ldots, \varepsilon_{t-1}, \varepsilon_t), \quad t \in \mathbb{Z}, \tag{2.15}$$

where $(\varepsilon)_{t\in\mathbb{Z}}$ is a strong WN$(0, \sigma^2)$ and g is a measurable function. Such processes may be called subordinated Bernoulli shifts. Obviously, $(X_t)_{t\in\mathbb{Z}}$ is strictly stationary and causal in the sense that X_t depends only on innovations ε_s up to the moment t. This is a broad class of processes containing one-sided moving averages and their transforms, Volterra processes and many nonlinear processes e.g. processes of the form $X_t = R(X_{t-1}, \varepsilon_t)$ which admit stationary version. From the reason which will become apparent shortly right hand side of (2.15) is sometimes called nonlinear Wold representation.
Let $\xi_j = (\ldots, \varepsilon_{j-1}, \varepsilon_j)$ and for $j \geqslant 0$ consider its coupled version ξ_j^* with ε_0 replaced by its independent copy ε_0' i.e. random variable having the same distribution as ε_0 and independent of (ε_t). Moreover, define $g_n(\xi_0) = E(X_n|\xi_0) = E(g(\ldots, \varepsilon_{n-1}, \varepsilon_n)|\xi_0)$ and consider $g_n(\xi_0^*) = E(g(\xi_n^*)|\ldots, \varepsilon_{-1}, \varepsilon_0')$. For $p \geqslant 1$ and $n \geqslant 0$ functional measures of dependence are defined as

$$\delta_p(n) = ||g(\xi_n) - g(\xi_n^*)||_p$$

and predictive measures of dependence as

$$\omega_p(n) = ||g_n(\xi_0) - g_n(\xi_0^*)||_p.$$

Intuitively, $\delta_p(n)$ measures the influence of the change of an innovation at time zero on the value of the series at time n whereas $\omega_p(n)$ measures the impact of the respective change on the optimal nonlinear prediction of X_n based on innovations up to time 0. Also, introduce the general functional measure of an impact of the change at positions $I \subset \mathbb{Z}$, namely $\delta_p(I, n) = ||g(\xi_n) - g(\xi_{n,I})||_p$, where $\xi_{n,I}$ is a coupled version of ξ_n with innovations positioned at $i \in I$ replaced by their independent copies. Let $\mathcal{P}_k Z = E(Z|\xi_k) - E(Z|\xi_{k-1})$ be the projection operator for $Z \in \mathcal{L}^p$. The following result holds (Wu (2005))

Theorem 2.4.1 *(i) For $n \geqslant 0$ and $p \geqslant 1$ we have that $\omega_p(n) \leqslant \delta_p(n)$.*
(ii) For $n \geqslant 0$

$$||\mathcal{P}_0 X_n||_p \leqslant \omega_p(n) \leqslant 2||\mathcal{P}_0 X_n||_p.$$

(iii) Let $I \subset \mathbb{Z}$. We have

$$\delta_2^2(I, n) \leqslant 16 \sum_{i\in I} \delta_2^2(n - i).$$

Note that the last inequality shows that the strength of the change at positions in I can be bounded in element-wise manner by the strengths of dependence of X_n on individual ε_i for $i \in I$.
Proof. (i) We have for $n \geqslant 0$, in the view of $\xi_n^* = (\xi_{-1}, \varepsilon_0', \varepsilon_1, \ldots, \varepsilon_n)$

$$\begin{aligned} &E[g(\xi_n) - g(\xi_n^*)|\xi_{-1}, \varepsilon_0, \varepsilon_0'] = E[g(\xi_n)|\xi_{-1}, \varepsilon_0'] - E[g(\xi_n^*)|\xi_{-1}, \varepsilon_0] \\ &= g_n(\xi_0) - g_n(\xi_0^*). \end{aligned} \tag{2.16}$$

This in view of Jensen's inequality yields

$$(g_n(\xi_0) - g_n(\xi_0^*))^2 \leqslant E[(g(\xi_n) - g(\xi_n^*))^2|\xi_{-1}, \varepsilon_0, \varepsilon_0']$$

and integrating with respect to $\xi_{-1}, \varepsilon_0, \varepsilon_0'$ yields (i).
(ii) Observe that since $E(g(\xi_n)|\xi_{-1}) = E(g_n(\xi_0)|\xi_{-1})$ we have $E(g_n(\xi_0)|\xi_{-1}) = E(g_n(\xi_0^*)|\xi_0)$, and thus

$$\begin{aligned} &||\mathcal{P}_0 X_n||_p = ||E[g_n(\xi_0) - g_n(\xi_0^*)|\xi_0]||_p \leqslant ||g_n(\xi_0) - g_n(\xi_0^*)||_p \\ &\leqslant ||g_n(\xi_0) - E[g_n(\xi_0)|\xi_{-1}]||_p + ||E[g_n(\xi_0)|\xi_{-1}] - g_n(\xi_0^*)||_p \\ &= 2||\mathcal{P}_0 X_n||_p \end{aligned} \tag{2.17}$$

which ends the proof of (ii). We refer to Wu (2005) for the proof of (iii).
We say that the process (X_t) is p-stable if $\Omega_p := \sum_{n=0}^{\infty} \omega_p(n)$ is finite.

Example 2.4.2 *(i) Consider the one-sided linear process (1.10). Clearly* $\omega_p(n) = \delta_p(n) = 2|c_n|||\varepsilon_0 - \varepsilon_0'||_p$ *and thus 2-stability is equivalent to absolute convergence of* $\sum |\psi_i|$*. Moreover, under mild conditions on* K *it can be proved (cf. Wu (2005)) that for* $Y_t = K(X_t)$ *we have* $||\mathcal{P}_0 Y_t|| = O(|c_t|)$ *thus 2-stability also hold for linear subordinated process.*
(ii) Consider process $X_t = R(X_{t-1}, \varepsilon_t)$ *discussed in (1.3.3). It can be proved that under (1.16) the process has geometric moment contraction property i.e. for some* $p > 0$ *and* $r \in (0, 1)$

$$||X_n - g(\xi_{n,I})||_p = O(r^n),$$

where $I = \{\ldots, , -1, 0\}$*. Note that* $||X_n - g(\xi_{n,I})||_p$ *is another dependence index which measures the influence of the past* $\ldots, \varepsilon_{-1}, \varepsilon_0$ *on* X_n*. As by stationarity* $||g(\xi_n^*) - g(\xi_{n,I})||_p = ||g(\xi_{n+1}) - g(\xi_{n+1,I})||_p$ *the moment contraction property implies that*

$$\begin{aligned} \delta_p(n) &= ||g(\xi_n^*) - X_n)||_p \leqslant ||g(\xi_n^*) - g(\xi_{n,I})||_p + ||g(\xi_{n,I}) - X_n||_p \\ &= O(r^{n+1}) + O(r^n) = O(r^n). \end{aligned}$$

Thus in this case the series is p-stable. Conversely, if $\delta_p(n) = O(r^n)$ *for some* $p > 1$ *then Theorem 2.4.1 (i) implies that moment contraction property holds. We provide one representative example of the usefulness of this approach. The following result holds (cf Wu (2005)).*

Theorem 2.4.3 *Assume that* $\Omega_2 < \infty$. *Then*

$$\{S_{[nt]}/\sqrt{n}, 0 \leqslant t \leqslant 1\} \xrightarrow{\mathcal{D}} \{\sigma B(t), 0 \leqslant t \leqslant 1\}$$

when $n \to \infty$ *in* $\mathcal{D}[0,1]$, *where* $B(\cdot)$ *is the standard Brownian motion and* $\sigma = ||\sum_{i=0}^{\infty} \mathcal{P}_0 X_i|| \leqslant \Omega_2$. $\mathcal{D}[0,1]$ *denotes the space of right-continuous functions which have left-hand limits on* $[0,1]$ *with Skorochod topology (see Billingsley (1968)).*

2.5 Central Limit Theorems for dependent sequences

We now discuss limit theorems for mixing processes and linear processes. First note that as mixing in any sense considered by us implies ergodicity it follows from the ergodic theorem that $(X_1 + \dots X_n)/n \to EX_1$ almost surely provided $E|X_1|$ is finite i.e. strong law of large numbers holds. In order to state CLT for mixing sequences we first present some relevant inequalities. Let X, Y be two random variables and $\sigma(X), \sigma(X)$ sigma-algebras generated by them. Moreover, let $\alpha = \sup_{A \in \sigma(X), B \in \sigma(Y)} |P(A \cap B) - P(A)P(B)|$ and analogue fo ϕ-mixing coefficient defined analogously.

Lemma 2.5.1 *(i) If* $|X_1| \leqslant C_1$ *and* $|Y| \leqslant C_2$ *we have*

$$|\mathrm{Cov}(X,Y)| \leqslant 4\alpha C_1 C_2$$

(ii) If $E(|X|^p + |Y|^q) < \infty$ *for some* $p, q \geqslant 1$ *and* $r = (1 - 1/p - 1/q)^{-1} > 0$ *then*

$$|\mathrm{Cov}(X,Y)| \leqslant 8\alpha^{1/r} (E(|X|^p)^{1/p} E(|Y|^q)^{1/q}.$$

(iii) If $p, q > 1$ *are such that* $1/p + 1/q = 1$ *then*

$$|\mathrm{Cov}(X,Y)| \leqslant 2\phi^{1/p} (E(|X|^p)^{1/p} E(|Y|^q)^{1/q}.$$

Note that in particular it follows from (i) that if $(X_t)_{t \in \mathbb{Z}}$ is a bounded time series such that $\sum \alpha(i)$ is finite then $\sum_{-\infty}^{\infty} \gamma(h)$ is finite.
We also state an interesting general moment inequality proved in Doukhan and Louhichi (1999) which yields a bound for qth moment of a sum $S_n = X_1 + \cdots X_n$, where X_i are zero mean variables. Let

$$M_{r,q} = \sup |\mathrm{Cov}(X_{t_1} \cdots X_{t_p}, X_{t_{p+1}} \cdots X_{t_q})|,$$

where the supremum is taken over all $t_1, \dots, t_q$ such that $1 \leqslant p < q$ and $t_{p+1} - t_p = r$.

Theorem 2.5.2 *Assume that for some* $q \geqslant 2$ $M_{r,q} = O(r^{q/2})$ *when* $r \to \infty$. *Then for some* $C > 0$ *we have*

$$|E(S_n^q)| \leqslant C n^{q/2}.$$

The inequalities in Lemma 2.5.1 are used to prove a CLT below. The main device here is modified small-block and large-block argument originally due to Bernstein which we discuss later in connection with estimation of the mean.

Theorem 2.5.3 *Assume that $(X_t)_{t\in\mathbb{Z}}$ is strictly stationary zero mean process such that $\sigma^2 = \sum_{-\infty}^{\infty} \gamma(i)$ is positive and let $S_n = X_1 + \dots X_n$. Then*

$$S_n/\sqrt{n} \xrightarrow{\mathcal{D}} N(0, \sigma^2)$$

when $n \to \infty$ provided any of the three conditions is satisfied:
(i) $|X_t|$ is bounded almost surely and $\sum \alpha(i)$ is finite;
(ii) $E|X_t|^\delta < \infty$ and $\sum_{j\geqslant 1} \alpha(j)^{1-2/\delta} < \infty$ for some $\delta > 2$.
(iii) $\sum_{j\geqslant 1} \phi^{1/2}(j) < \infty$.

Note that that the Lemma 2.5.1 implies that in all cases (i)-(iii) σ^2 is finite. Moreover, note that in view of (2.9) condition in (iii) implies that $\sum_{j\geqslant 1} \alpha(j)$ is finite, however stronger condition on α-mixing coefficients is imposed in (ii).
There are many unresolved problems about properties of mixing sequences. The most known is Ibragimov's conjecture, now more then fifty years old, stating that ϕ-mixing sequence $(X_n)_{n\in\mathbb{N}}$ with the second moment of the marginal distribution finite and such that $\mathrm{Var}(X_1 + \cdots + X_n) \to \infty$ satisfies a CLT. The result is true when finiteness of the absolute moment of order $2 + \delta$ is imposed (cf. Theorem 18.5.1 in Ibragimow and Linnik (1971)).
A remarkable theorem of Ibragimov and Linnik which avoids imposing any mixing conditions at the expense of assumption that the process is linear will be discussed in Section 7.
The last result concerns the CLT for sums of martingale differences in a triangular array and is due to Brown (1971). Let $(X_{n,t}, t = 1, \dots, k_n)$ be an array of random variables on $(\Omega, \mathcal{F}, P)$ and let $\mathcal{F}_{n,t}$ for $0 \leqslant t \leqslant k_n$ be a sequence of sub σ-algebras of $\mathcal{F}$ such that $X_{n,t}$ is $\mathcal{F}_{n,t}$-measurable and $\mathcal{F}_{n,t-1} \subseteq \mathcal{F}_{n,t}$. Each row of $(X_{n,t}, t = 1, \dots, k_n)$ consists of martingale differences i.e. $E(X_{n,t}|\mathcal{F}_{n,t-1}) = 0$. Let $S_n = \sum_{t=1}^{k_n} X_{n,t}$.

Theorem 2.5.4 *Let $(X_{n,t}, \mathcal{F}_{n,t}, t = 1, \dots, k_n)$ be a martingale difference array such that*

$$\sum_{t=1}^{n} E(X_{t,n}^2|\mathcal{F}_{n,t-1}) \to 1$$

in probability and for any $\varepsilon > 0$

$$\sum_{t=1}^{n} E(X_{t,n}^2 I\{|X_{t,n}| > \varepsilon\}) \to 0.$$

Then $S_n \xrightarrow{\mathcal{D}} N(0,1)$ when $n \to \infty$.

2.6 Measures of dependence: information theoretic approach

We give a brief overview of the main concepts of information theory and their application to measuring dependence of time series, for a more exhaustive treatment and some aspects of an interplay of information theory with statistics we refer to Dębowski (2013) and Cover and Thomas (2006). Let X be a random variable with possible qualitative values in dictionary $\mathcal{D} = \{x_1, \ldots, x_k\}$ and denote $P(x) = P(X = x)$. To fix ideas we treat the case of finite dictionary only, although most of the properties below are true for countably infinite $\mathcal{D}$. The approach plays an important role in modelling and quantification of dependence for time series having qualitative values e.g. in Natural Language Modelling. Recall the following definitions.

Definition 7 *Entropy of random variable X is*

$$H(X) = E(-\log P(X)) = -\sum_{i=1}^{k} P(x_i) \log P(x_i). \tag{2.18}$$

The above definition has its analogue for continuous random variables X with density f for which differential entropy is defined as

$$H(X) = -\int f(x) \log f(x)\, dx. \tag{2.19}$$

Entropy (2.18) depends only on the probability mass function of X and frequently is defined for probability distributions without reference to a specific X. It is, along with the Gini index, the most popular measure of scatter for qualitatively distributed variables. In particular it is used in Classification and Regression Trees methodology for measuring variability of the class index in a node of a tree. This interpretation is confirmed by its properties as it satisfies $H(X) \geqslant 0$ and is equal 0 only for a single-valued X. Moreover, $H(X) \leqslant H(U) = \log k$, where U is uniformly distributed $P(U = x_i) = 1/k$ and thus is the most scattered among $\mathcal{D}$-valued random variables.
The definition above can be extended to entropy $H(X_1, X_2)$ of the pair of random variables having joint mass function $P(x_{1i}, x_{2j})$ for $i = 1, \ldots, k, j = 1, \ldots, l$ and in general to $H(X_1, X_2, \ldots, X_n)$.
One of the possible measures of dependence between X and Y is an entropy of conditional mass function $P(X = \cdot|Y = y)$ averaged with respect to the distribution of Y called conditional entropy and denoted by $H(X|Y)$. Note that

$$\begin{aligned} H(X|Y) &= \sum_{y:P(y)>0} P(Y = y) H(X|Y = y) \\ &= -\sum_{x,y:P(X=x,Y=y)>0} P(Y = y) P(X = x|Y = y) \log P(X = x|Y = y) \end{aligned}$$

$$= - \sum_{x,y:P(X=x,Y=y)>0} P(X=x,Y=y) \log \frac{P(X=x,Y=y)}{P(Y=y)}$$
$$= H(X,Y) - H(X), \tag{2.20}$$

which underlines usefulness of $H(X|Y)$ as a measure of dependence: it yields the incremental increase of entropy if we append X with Y. Equality (2.20) is generalised to so-called chain rule

$$H(X_1,\ldots,X_n) = H(X_1) + \sum_{i=2}^{n} H(X_i|X_1,\ldots,X_{i-1}). \tag{2.21}$$

We also define a mutual information $I(X;Y)$ as

$$I(X;Y) = E\big(\log \frac{P(X,Y)}{P(X)P(Y)}\big)$$

and note that it is Kullback-Leibler $D(p||q)$ divergence between bivariate mass function $p = P(x,y)$ and the product of its marginals $q = P(x)P(y)$. It follows from the properties of the latter that $I(X;Y) \geqslant 0$ and equals 0 only in the case when p and q coincide i.e. X and Y are independent. As $H(X,Y) + I(X;Y) = H(X) + H(Y)$ and all quantities are non-negative we obtain that $H(X|Y) \leqslant H(X)$.
Moreover, we define conditional mutual information of X and Y given Z

$$I(X;Y|Z) = E\big(\log \frac{P(X,Y|Z)}{P(X|Z)P(Y|Z)}\big),$$

as an averaged Kullback-Leibler between conditional distribution $P(X,Y|Z)$ and the product of its marginals. Nonnegativity of $I(X;Y|Z)$ is obvious as well as the property that it equals 0 only in the case when X and Y are conditionally independent given Z.
We now turn to defining some dependence measures for qualitatively-valued stationary time series $(X_t)_{t\in\mathbb{Z}}$. Stationarity is meant here in the strict sense, as a definition of weak stationarity is not applicable to qualitatively-valued random variables. Denote by X_k^l for $k \leqslant l$ a block of observations from time k to l and define block entropy $H(n) := H(X_{1+k}^{k+n})$, $n \geqslant 1, k \in \mathbb{Z}$ which due to stationarity depends only on the size of the block. We let also $H(0) := 0$. Dependence of (X_t) can be e.g. gauged by a differenced block entropy

$$\Delta H(n) = H(X_n|X_1^{n-1}) = H(X_1,\ldots,X_n) - H(X_1,\ldots,X_{n-1}). \tag{2.22}$$

Other possible measure is a block mutual information defined as

$$E(n) = I(X_{-n+1}^0; X_1^n). \tag{2.23}$$

We list below two important properties of the above measures.

Theorem 2.6.1 *(i) For a stationary process the following two limits exists and coincide*

$$\lim_{n\to\infty} \frac{H(n)}{n} = \lim_{n\to\infty} \Delta H(n) = h \geqslant 0.$$

Limit h is called the entropy rate.
(ii) We also have that

$$\lim_{n\to\infty} E(n) = \lim_{n\to\infty} H(n) - n\Delta H(n) = \lim_{n\to\infty} H(n) - nh = E.$$

E is called excess entropy (cf Crutchfield and Feldman (2003)).

Proof. We prove part (i) and we refer to Dębowski (2013), Theorem 4.9 for the proof of (ii). Observe that since conditioning does not increase entropy i.e. $H(X|Y) \leqslant H(X)$ we have

$$H(X_{n+1}|X_1,\ldots,X_n) \leqslant H(X_{n+1}|X_2,\ldots,X_n) = H(X_n|X_1,\ldots,X_{n-1}),$$

where the last equality follows from stationarity. Thus sequence $\Delta H(n)$ is non-increasing and whence its limit h exists. Since in view of chain rule (2.21) $H(n)/n = H(1)/n + \sum_{i=2}^{n} \Delta H(i)/n$ and as Cesàro means of $(\Delta(n))_n$ converge to h, (i) follows.
It follows from the proof that we have that $h \leqslant H(1)$, moreover, the equality $h = H(1)$ implies $H(X_n) = H(X_n|X_1,\ldots,X_{n-1})$ and thus X_n is independent of $X_1,\ldots,X_{n-1}$ for any n what implies that X_is are independent. Moreover, for the stationary Markov chain we have that $h = \lim_{n\to\infty} H(X_n|X_1,\ldots,X_{n-1}) = H(X_n|X_{n-1}) = H(X_2|X_1)$.
Note that by the chain rule for multivariate mass function $p(X_0,\ldots X_{n-1}) = \prod_{i=0}^{n-1} P(X_i|X_0^{i-1})$ we have that

$$-\frac{1}{n}\log P(X_0,\ldots,X_{n-1}) = -\frac{1}{n}\sum_{i=0}^{n-1} \log p(X_i|X_0^{i-1})$$

and for ergodic sequences it can be proved that the right hand side converges to

$$\lim_{n\to\infty} E(-\log(X_n|X_0^{n-1}) = h.$$

This is Shannon-McMillan-Breiman equipartition theorem, which states that (see e.g. Cover and Thomas (2006))

Theorem 2.6.2 *For a finite-valued strictly stationary ergodic sequence $(X_n)_{n\in\mathbb{N}}$ we have when $n \to \infty$*

$$-\frac{1}{n}\log P(X_0,\ldots,X_{n-1}) \to h$$

almost surely.

This means that asymptotically probability mass $p(x_1, \ldots, x_n)$ is approximately equi-distributed over 2^{nh} points with probability 2^{-nh} assigned to each of them. This has profound consequences for coding theory meaning that for sample paths of ergodic sequence we need on average 2^{nh} bits to code its subsequence of length n. This also indicates that upper bounds on h are valuable since they give us some idea on the optimal rate of compression.

2.7 Problems

1. Show that for any square-integrable real X and Y we have

$$\mathrm{Cov}(X, Y) = \int_{\mathbb{R}} \int_{\mathbb{R}} (P(X \geqslant s, Y \geqslant t) - P(X \geqslant s)P(Y \geqslant t))\, ds\, dt.$$

Prove it first for nonnegative random variables noting that

$$XY = \int_0^\infty \int_0^\infty I\{X \geqslant s\} I\{Y \geqslant t\}\, ds\, dt$$

and integrating both sides with respect to P. For arbitrary X and Y use decomposition $X = X^+ - X^-$, where $X^+ = \max(X, 0)$ and $X^- = \max(-X, 0)$ are nonnegative random variables together with $P(X > -t) = 1 - P(X^- \geqslant t)$ for $t > 0$.
2. Prove that when (Y_t) is a strong $WN(0, \sigma^2)$ then in (2.8) we have that $S_n/n \to 0$ a.s. under the weaker condition that $(\psi_i) \in \ell^2$.
3. Using (2.7) and Lebesgue dominated convergence theorem prove Corollary 2.2.3.
4. Prove that β-mixing coefficient satisfies (i) $\beta(n) \leqslant 1$ (ii) $\beta(n) \leqslant \phi(n)$.
5. (i) Using Lemma 2.5.1 prove that Theorem 2.5.2 holds true provided (X_t) is a strictly stationary sequence such that $E|X_t|^\delta < \infty$ for some $\delta > q$ and $\alpha(n) = O(n^{-\frac{\delta q}{2(\delta - q)}})$. (ii) Similarly, prove that the conclusion of (i) holds true provided X_1 is bounded and $\alpha(n) = O(n^{-q/2})$
6. Prove Lemma 2.5.1 (i). Hint. Use the conclusion of problem 1.
7. Show that for a stationary Markov chain with stationary distribution $(\mu_1, \ldots, \mu_k)$ and transition matrix (p_{ij}) we have that the entropy rate h satisfies

$$h = -\sum_{i=1}^{k} \sum_{j=1}^{k} \mu_i p_{ij} \log p_{ij}.$$

3

Optimal linear prediction

In this chapter we discuss a linear prediction problem, state prediction equations and introduce two algorithms by which prediction coefficients can be calculated, namely the Durbin-Levinson and the innovations algorithm. Prediction problem is frequently referred to as forecasting. Let $(X_t)_{t\in\mathbb{Z}}$ be a weakly stationary time series with the mean m and covariance function $\gamma(\cdot)$. We assume that m and covariance function $\gamma(\cdot)$ are known and moreover the covariance function is such that $\mathbf{\Gamma}_n = (\gamma(i-j))_{1\leqslant i,j\leqslant n}$ is nondegenerate i.e. $\mathbf{\Gamma}_n^{-1}$ exists for a given $n \in \mathbb{N}$.

3.1 The Yule - Walker equations

We look for the best linear predictor of X_{n+h} based on affine combination of $X_1, \ldots, X_n$ i.e. solution to the problem of finding

$$\arg \min_{a_0,a_1,\ldots,a_n} S(a_0, a_1, \ldots, a_n),$$

where

$$S(a_0, a_1, \ldots, a_n) = \| X_{n+h} - a_0 - \sum_{i=1}^{n} a_i X_{n+1-i} \|^2 = E\big(X_{n+h} - a_0 - \sum_{i=1}^{n} a_i X_{n+1-i}\big)^2.$$

It is known as the problem of h steps ahead (h-step) linear prediction. Before we develop an equation determining the coefficients of the best linear prediction let us note two facts. First, the condition we impose that $\mathbf{\Gamma}_n^{-1}$ exists is equivalent to the fact that $1, X_1, \ldots, X_n$ are not linearly dependent in the sense that there does not exist non-zero vector $(b_0, b_1 \ldots, b_n)$ such that $b_0 + \sum_{i=1}^{n} b_i X_i = 0$ with probability 1(Problem 3.1). This is a natural condition here as otherwise we can choose a proper subset of variables $\{X_{i_1}, \ldots, X_{i_k}\}$ which spans the same space i.e. $sp(1, X_{i_1}, \ldots, X_{i_k}) = sp(1, X_1, \ldots, X_n)$ and for which corresponding covariance matrix is non-degenerate. Secondly, it is easy to establish that the optimal predictor $P_n X_{n+h}$ is necessarily such that $X_{n+h} - P_n X_{n+h}$ is perpendicular to $sp(1, X_1, \ldots, X_n)$ as otherwise denoting by $X_\perp$ the vector having this property (existence of which we are about to establish) we will have

$$||X_{n+h} - P_n X_{n+h}||^2 = ||X_{n+h} - X_\perp||^2 + ||X_\perp - P_n X_{n+h}||^2$$

as $X_{n+h} - X_\perp$ is perpendicular to $X_\perp - P_n X_{n+h}$. This contradicts the fact that $P_n X_{n+h}$ is minimizer of $S(a_0, a_1, \ldots, a_n)$.
The geometric solution of the problem is shown below. We also note that our task is equivalent to problem of theoretical linear regression for which X_{n+h} is the response, $1, X_1, \ldots, X_n$ are random predictors and criterion function is given by $S(a_0, a_1, \ldots, a_n)$. Note that we want to take into account the fact that predictors are dependent random variables.

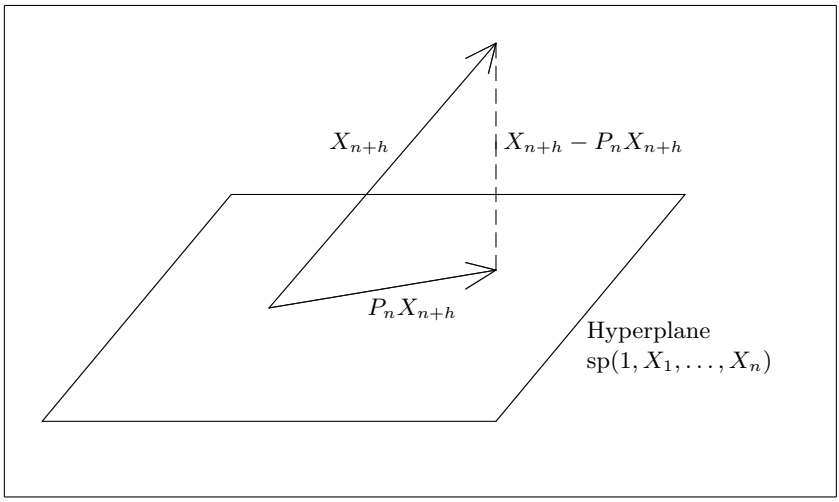


Fig. 3.1. $P_n X_{n+h}$ as a perpendicular projection of X_{n+h}

We already know that $X_{n+h} - P_n X_{n+h}$, that is residual of X_{n+h} after projecting it onto hyperplane $sp(1, X_1, \ldots, X_n)$, has to be perpendicular (in $\mathcal{L}^2$ space) to this hyperplane, or, equivalently, has to be perpendicular to its generators $1, X_1, \ldots, X_n$. We thus obtain

$$X_{n+h} - P_n X_{n+h} \perp 1 \tag{3.1}$$

and

$$X_{n+h} - P_n X_{n+h} \perp X_j, \quad j = 1, \ldots, n. \tag{3.2}$$

Property (3.1) is equivalent to

$$< 1, X_{n+h} - P_n X_{n+h} >= E\big(1(X_{n+h} - a_0 - \sum_{i=1}^{n} a_i X_{n+1-i})\big) = 0. \tag{3.3}$$

whereas (3.2) is equivalent to

$$E\big(X_j(X_{n+h} - a_0 - \sum_{i=1}^{n} a_i X_{n+1-i})\big) = 0, \quad j = 1, \ldots, n. \tag{3.4}$$

We obtain from (3.3) that ($m = EX_t$)

$$a_0 = m(1 - \sum_{i=1}^{n} a_i) \tag{3.5}$$

and substituting it into (3.4) we have

$$E\big(X_j(X_{n+h} - m - \sum_{i=1}^{n} a_i(X_{n+1-i} - m))\big) = 0, \tag{3.6}$$

or equivalently

$$\mathrm{Cov}(X_{n+h}, X_j) = \sum_{i=1}^{n} a_i \mathrm{Cov}(X_{n+1-i}, X_j). \tag{3.7}$$

Changing the index $j := n + 1 - j, \quad j = 1, \ldots, n$ we see that (3.7) is equivalent to

$$\gamma(h + j - 1) = \sum_{i=1}^{n} a_i \gamma(i - j), \quad j = 1, \ldots, n. \tag{3.8}$$

Recall that $\mathbf{\Gamma}_n = \big(\gamma(i-j)\big)_{i,j=1}^{n}$ and moreover define

$$\boldsymbol{\gamma}_n(h) = (\gamma(h), \gamma(h+1), \ldots, \gamma(h+n-1))'$$

Note that $\boldsymbol{\gamma}_n(h)$ is vector of covariances of X_{n+h} with $X_n, \ldots, X_1$. Let $\mathbf{a}_n = (a_1, a_2, \ldots, a_n)'$. Then (3.8) written in the vector form is

$$\mathbf{\Gamma}_n \mathbf{a}_n = \boldsymbol{\gamma}_n(h).$$

As $\mathbf{\Gamma}_n^{-1}$ exists, then $\mathbf{a}_n$ satisfying this equation is unique and

$$\mathbf{a}_n = \mathbf{\Gamma}_n^{-1} \boldsymbol{\gamma}_n(h). \tag{3.9}$$

Equations (3.5) and (3.9) are known as the Yule-Walker equations. Inverting $\mathbf{\Gamma}_n$ in (3.9) can be avoided by using Durbin-Levinson or innovations algorithm discussed below. Note that for $n = 1$ we have from (3.9) that $a_1 = \gamma(h)/\gamma(0) = \rho(h)$ and thus using (3.5) we obtain

$$P_1 X_{1+h} = a_0 + a_1 X_1 = m + \rho(h)(X_1 - m). \tag{3.10}$$

We derive now the formula for mean squared error of h–step prediction.

$$\begin{aligned} \sigma_{n,h}^2 &= \| X_{n+h} - P_n X_{n+h} \|^2 = \| X_{n+h} - m - \sum_{i=1}^{n} a_i(X_{n+1-i} - m) \|^2 = \\ &= \gamma(0) - 2\mathbf{a}_n' \boldsymbol{\gamma}_n(h) + \mathbf{a}_n' \mathbf{\Gamma}_n \mathbf{a}_n = \gamma(0) - \boldsymbol{\gamma}_n'(h) \mathbf{\Gamma}_n^{-1} \boldsymbol{\gamma}_n(h) \end{aligned} \tag{3.11}$$

using (3.5) for the second equality and (3.9) for the last one.
Equation (3.11) is a frequently used expression for mean squared error of h–step prediction. We give one of possible alternative equalities for one step prediction error below. For $h = 1$ let $\sigma_n^2 := \sigma_{n,1}^2$ and $\boldsymbol{\gamma}_n := \boldsymbol{\gamma}_n(1)$.

Proposition 3.1.1 *Assume that $\boldsymbol{\Gamma}_n$ is invertible. Then*

$$\sigma_n^2 = |\boldsymbol{\Gamma}_{n+1}| \ / \ |\boldsymbol{\Gamma}_n|, \tag{3.12}$$

where $|\boldsymbol{\Gamma}| = \det \boldsymbol{\Gamma}$.

Proof. Proof of (i) follows from the following property of determinants

$$\det \begin{pmatrix} A & B \\ C & D \end{pmatrix} = |A| \ |D - CA^{-1}B| \ = \ |D| \ |A - BD^{-1}C| \tag{3.13}$$

and the observation that

$$\boldsymbol{\Gamma}_{n+1} = \begin{pmatrix} \gamma(0) & \boldsymbol{\gamma}_n' \\ \boldsymbol{\gamma}_n & \boldsymbol{\Gamma}_n \end{pmatrix}$$

From (3.13) we have

$$|\boldsymbol{\Gamma}_{n+1}| = \big(\gamma(0) - \boldsymbol{\gamma}_n' \boldsymbol{\Gamma}_n^{-1} \boldsymbol{\gamma}_n\big) |\boldsymbol{\Gamma}_n|$$

and since $\gamma(0) - \boldsymbol{\gamma}_n' \boldsymbol{\Gamma}_n^{-1} \boldsymbol{\gamma}_n = \sigma_n^2$ equality (3.12) follows.

Remark 3.1.2 *(i) Equality (3.5) implies that*

$$P_n X_{n+h} = a_0 + \sum_{i=1}^{n} a_i X_{n+1-i} = m + \sum_{i=1}^{n} a_i (X_{n+1-i} - m). \tag{3.14}$$

It easily follows from (3.14) by contradiction that optimal linear predictor for mean m time series (X_t) is obtained by adding m to optimal linear predictor for the centred time series $(X_t - m)$.
(ii) The Yule-Walker equations imply that for a weakly stationary time series coefficients of prediction of X_{n+h} based on $1, X_n, X_{n-1}, \ldots, X_1$ are the same as for prediction of X_{t+h} based on $1, X_t, X_{t-1}, \ldots, X_{t-n+1}$ for any $t \in T$.
(iii)

$$\sigma_n^2 \leqslant \sigma_{n-1}^2 \leqslant \cdots \leqslant \sigma_0^2 = \mathrm{Var}(X_{n+1}) = \gamma(0).$$

(iv) Assume that (X_t) is mean-zero weakly stationary time series. Then (Problem 5.2)

$$\sigma_n^2 \to \sigma^2 = ||X_t - P_{H_{t-1}} X_t||^2 \quad \text{when } n \to \infty.$$

3.2 The Durbin–Levinson algorithm

We discuss now the Durbin–Levinson algorithm which finds solution to the Yule-Walker equations. Another method which also avoids inverting matrix $\mathbf{\Gamma}_n$ based on finding its modified Cholesky decomposition is discussed in section 3.3. We consider the case of $h = 1$ and assume as before that covariance matrix $\mathbf{\Gamma}_n$ of $(X_1, \ldots, X_n)$ is positive definite. We recall that this means that any nonzero linear combination of $X_1, \ldots, X_n$ is not constant. Then $\mathbf{\Gamma}_n$ is invertible and coefficients of a projection of X_{n+1} on $sp(1, X_1, \ldots, X_n)$ are uniquely defined. Traditionally, vector $(a_1, \ldots, a_n)'$ is denoted by $(\varphi_{n1}, \ldots, \varphi_{nn})'$ and we have (compare (3.14))

$$P_n X_{n+1} = m + \varphi_{n1}(X_n - m) + \varphi_{n2}(X_{n-1} - m) + \cdots + \varphi_{nn}(X_1 - m). \quad (3.15)$$

It is worthwhile to stress that the first index n in φ_{ni} corresponds to the number of observations we base our projection on. We usually have that $\varphi_{n+1,i}$ differs from φ_{ni}.

Definition 8 *Coefficient φ_{nn} corresponding to X_1 in representation (3.15) of $P_n X_{n+1}$ is called partial autocorrelation coefficient (PACF) or the Schur, Verblunsky coefficient of order n.*

Thus partial correlation coefficient of order n corresponds to the observation X_1 which is the furthest away time-wise from predicted observation X_{n+1}. intuitively, we would like to decide whether it is worthwhile to add an additional predictor X_1 to $X_n, \ldots, X_2$ based on absolute value or significance of φ_{nn}. Its equivalent definition will follow from the analysis of Durbin-Levinson algorithm. When $\mathbf{\Gamma}_n$ is not invertible, we let $\varphi_{nn} = 1$.

For the rest of this section we assume that $m = 0$. No generality is lost because of this in view of (3.14). Note that prediction equations (3.15) for the first $n+1$ observations can be written as

$$\mathbf{X}_{n+1} - \hat{\mathbf{X}}_{n+1} = \mathbf{\Phi}_n \mathbf{X}_{n+1}, \quad (3.16)$$

where $\mathbf{X}_{n+1} = (X_1, \ldots, X_{n+1})'$, $\hat{\mathbf{X}}_{n+1} = (\hat{X}_1, \ldots, \hat{X}_{n+1})'$ and $\mathbf{\Phi}_n$ is $(n+1) \times (n+1)$ lower triangular matrix defined as

$$\begin{pmatrix} 1 & 0 & \ldots & 0 & 0 \\ -\phi_{22} & 1 & \ldots & 0 & 0 \\ \vdots & & & \vdots & \vdots \\ -\phi_{n,n} & -\phi_{n,n-1} & \ldots & -\phi_{n,1} & 1 \end{pmatrix} \quad (3.17)$$

The Durbin–Levinson algorithm Let $\sigma_0^2 = \gamma(0)$ and assume that $\mathbf{\Gamma}_n$ is invertible. Coefficients φ_{ni} and σ_n^2 satisfy the following recursive equations

$$\varphi_{nn} = \left\{\gamma(n) - \sum_{j=1}^{n-1} \varphi_{n-1,j}\gamma(n-j)\right\} \sigma_{n-1}^{-2} \tag{3.18}$$

$$\begin{pmatrix} \varphi_{n,1} \\ \vdots \\ \varphi_{n,n-1} \end{pmatrix} = \begin{pmatrix} \varphi_{n-1,1} \\ \vdots \\ \varphi_{n-1,n-1} \end{pmatrix} - \varphi_{nn} \begin{pmatrix} \varphi_{n-1,n-1} \\ \vdots \\ \varphi_{n-1,1} \end{pmatrix} \tag{3.19}$$

$$\sigma_n^2 = (1-\varphi_{nn}^2)\sigma_{n-1}^2 = \cdots = \gamma(0)\prod_{i=1}^{n}(1-\phi_{ii}^2). \tag{3.20}$$

Thus having computed $\varphi_{n-1,1}, \ldots, \varphi_{n-1,n-1}$ we compute σ_{n-1}^2 (from (3.20)), then $\varphi_{n,n}$ from (3.18) and then finally $\varphi_{n,i},\ i = 1, \ldots, n-1$ from (3.19). Note that as $\mathbf{\Gamma}_n$ is invertible it implies that X_n does not belong to $sp(1, X_1, \ldots, X_{n-1})$ and thus $\sigma_{n-1}^2 > 0$. Then it follows from (3.20) that $\sigma_n^2 > 0$ is equivalent to $\varphi_{nn}^2 < 1$.

Proof. We give a proof of the algorithm based on the orthogonalization of subspaces $sp\{X_2, \ldots, X_n\}$ and $sp(X_1)$. Let $\mathcal{K}_1 = sp\{X_2, \ldots, X_n\}$ and $\mathcal{K}_2 = sp\{X_1 - P_{\mathcal{K}_1}X_1\}$. Thus $\mathcal{K}_2$ is univariate linear space such that $\mathcal{K}_2 \perp \mathcal{K}_1$ and moreover $\mathcal{X}_n == sp\{X_1, \ldots, X_n\}$ is a direct sum of $\mathcal{K}_1$ and $\mathcal{K}_2$

$$\mathcal{X}_n = \mathcal{K}_1 \oplus \mathcal{K}_2,$$

which means that $\mathcal{K}_1$ and $\mathcal{K}_2$ are orthogonal and any element of $x \in \mathcal{X}_n$ can be uniquely written as $x = x_1 + x_2$, where $x_i \in \mathcal{K}_i\,, i = 1, 2$. We have

$$\widehat{X}_{n+1} = P_{\mathcal{X}_n}X_{n+1} = P_{\mathcal{K}_1}X_{n+1} + P_{\mathcal{K}_2}X_{n+1} = P_{\mathcal{K}_1}X_{n+1} + a(X_1 - P_{\mathcal{K}_1}X_1) \tag{3.21}$$

$$a =< X_{n+1}, X_1 - P_{\mathcal{K}_1}X_1 > / \parallel X_1 - P_{\mathcal{K}_1}X_1 \parallel^2 . \tag{3.22}$$

As we noted that invertibility of $\mathbf{\Gamma}_n$ implies that $\sigma_{n-1}^2 > 0$ thus a is well defined. Note that from weak stationarity of (X_t) it follows that

$$\Sigma_{(X_1,\ldots,X_n)} = \Sigma_{(X_n,X_{n-1},\ldots,X_1)} = \Sigma_{(X_2,\ldots,X_{n+1})} = \mathbf{\Gamma}_n$$

and pertaining vectors $\gamma_n(1)$ coincide. The crucial step in the proof is now to note that the Yule-Walker equations yield that projection coefficients of X_1 on $X_2, \ldots, X_n$ are the same as projection coefficients of X_{n+1} on $X_n, \ldots, X_2$. Thus

$$P_{\mathcal{K}_1}X_1 = \sum_{j=1}^{n-1} \varphi_{n-1,j}X_{j+1} \tag{3.23}$$

$$P_{\mathcal{K}_1}X_{n+1} = \sum_{j=1}^{n-1} \varphi_{n-1,j}X_{n+1-j} \tag{3.24}$$

Thus (3.21), (3.23) and (3.24) imply that

$$\widehat{X}_{n+1} = aX_1 - aP_{\mathcal{K}_1}X_1 + P_{\mathcal{K}_1}X_{n+1} = aX_1 + \sum_{j=1}^{n-1}(\varphi_{n-1,j} - a\varphi_{n-1,n-j})X_{n+1-j} \tag{3.25}$$

where for the last equality we substitute $j := n - j$ in (3.23). Now, using (3.22) and (3.23) we have

$$a = \frac{< X_{n+1}, X_1 > - \sum_{j=1}^{n-1} \varphi_{n-1,j} < X_{n+1}, X_{j+1} >}{\sigma^2_{n-1}}, \tag{3.26}$$

where we use the observation that $\| X_1 - P_{\mathcal{K}_1}X_1 \|^2 = \| X_n - \widehat{X}_n \|^2 = \sigma^2_{n-1}$. As $\widehat{X}_{n+1} = \sum_{j=1}^{n} \varphi_{nj} X_{n+1-j}$ and this decomposition is unique we have from (3.25):

$$a = \varphi_{nn}, \quad \varphi_{n,j} = \varphi_{n-1,j} - a\varphi_{n-1,n-j},$$

which in view of (3.26) proves (3.18) and (3.19).
We have to show $\sigma^2_n = \sigma^2_{n-1}(1 - \varphi^2_{nn})$. Indeed,

$$\begin{aligned}
\sigma^2_n &= ||X_{n+1} - P_{\mathcal{X}_n}X_{n+1}||^2 = \| X_{n+1} - P_{\mathcal{K}_1}X_{n+1} - P_{\mathcal{K}_2}X_{n+1} \|^2 \\
&= \| X_{n+1} - P_{\mathcal{K}_1}X_{n+1} \|^2 + \| P_{\mathcal{K}_2}X_{n+1} \|^2 \\
&- 2 < X_{n+1} - P_{\mathcal{K}_1}X_{n+1}, P_{\mathcal{K}_2}X_{n+1} > \\
&= \| X_{n+1} - P_{\mathcal{K}_1}X_{n+1} \|^2 + \| P_{\mathcal{K}_2}X_{n+1} \|^2 \\
&- 2 < X_{n+1} - P_{\mathcal{X}_n}X_{n+1} + P_{\mathcal{K}_2}X_{n+1}, P_{\mathcal{K}_2}X_{n+1} > \\
&= \| X_{n+1} - P_{\mathcal{K}_1}X_{n+1} \|^2 + \| P_{\mathcal{K}_2}X_{n+1} \|^2 - 2||P_{\mathcal{K}_2}X_{n+1}||^2
\end{aligned}$$

and the ultimate equality follows from the fact that $X_{n+1} - P_{\mathcal{X}_n}X_{n+1}$ is perpendicular to $\mathcal{K}_2$. As $P_{\mathcal{K}_2}X_{n+1} = a(X_1 - P_{\mathcal{K}_1}X_1)$ the last expression equals

$$\sigma^2_{n-1} + a^2\sigma^2_{n-1} - 2a^2\sigma^2_{n-1} = \sigma^2_{n-1}(1 - a^2) = \sigma^2_{n-1}(1 - \varphi^2_{nn})$$

which ends the proof.

Remark 3.2.1 *We note that formula (3.20) is valid even when* $\mathbf{\Gamma}_n$ *is not invertible (or, equivalently,* $\sigma^2_{n-1} > 0$ *does not hold) as in this case we have that* $\sigma^2_{n-1} = 0$ *obviously implies* $\sigma^2_n = 0$.

Corollary 3.2.2 *For* (X_t) *WS time series such that* $\mathbf{\Gamma}_n$ *is invertible for some* $n \in \mathbb{N}$ *we have*

$$\begin{aligned}
\varphi_{nn} &= < X_{n+1}, X_1 - P_{\mathcal{K}_1}X_1 > / \| X_1 - P_{\mathcal{K}_1}X_1 \|^2 = \\
&= < X_{n+1} - P_{\mathcal{K}_1}X_{n+1}, X_1 - P_{\mathcal{K}_1}X_1 > / \| X_1 - P_{\mathcal{K}_1}X_1 \|^2 = \\
&= \rho(X_{n+1} - P_{\mathcal{K}_1}X_{n+1}, X_1 - P_{\mathcal{K}_1}X_1).
\end{aligned} \tag{3.27}$$

Right hand side of (3.27) is usually denoted by $\alpha(n)$ and is defined provided that $\mathbf{\Gamma}_n$ is invertible. If we additionally define $\alpha(n) = 1$ when $\mathbf{\Gamma}_n$ is non-invertible we obtain equality $\varphi_{nn} = \alpha(n)$ which is always true as the same convention has been adopted for ϕ_{nn}. For WS time series having arbitrary mean m we obtain equality of φ_{nn} and $\alpha(n)$ when slightly modifying definition of $\alpha(n)$

$$\alpha(n) = \rho(X_{n+1} - P_{sp(1,X_2,\ldots,X_n)}X_{n+1}, X_1 - P_{sp(1,X_2,\ldots,X_n)}X_1). \tag{3.28}$$

Graphical representation corresponding to the construction of $\alpha(n)$ is given below.

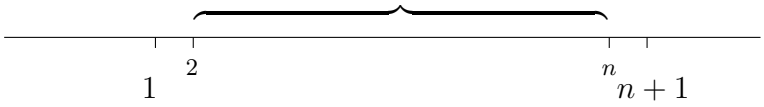


Fig. 3.2. Linear influence of $X_2, \ldots, X_n$ is removed from X_1 and X_{n+1}

From (3.27) we obtained an equivalent definition of ϕ_{nn} as a correlation coefficient between X_1 i X_{n+1} after removing linear influence of predictors $X_2, \ldots, X_n$. Note that importance of ϕ_{nn} is clearly seen when we consider prediction of X_{n+1} based on $X_n, \ldots, X_2$ and we want to decide whether it is worthwile to include X_1 as the predictor. If ϕ_{nn} is not negligible, enlarging set of predictors is justified. We note that a more general construction to construction of $\alpha(n)$ is used in regression analysis when in order to extract an influence of a particular predictor, say x_1, on response y, residuals res_y of y and residuals res_{x_1} of x_1 are computed and then res_y is considered as a response and res_{x_1} as a predictor in linear regression (partial regression plot).

Example 3.2.3 *Let X_t be a weakly stationary time series defined as*

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

where Z_t - $WN(0, \sigma^2)$ and $\theta \in \mathbb{R}$. It is a moving average of order 1 (MA(1)). We calculate partial autocorrelation coefficients $\alpha(1)$ i $\alpha(2)$ for this process. Obviously, $\alpha(1) = \rho(1) = \frac{\theta}{1+\theta^2}$. In order to calculate $\alpha(2)$ we calculate first $P_{sp(X_2)}X_3$.

$$P_{sp(X_2)}X_3 = \rho(1)X_2 = \frac{\theta}{1+\theta^2}X_2 = P_{sp(X_2)}X_1,$$

where the last equality (used also in greater generality in the proof of the Durbin-Levinson algorithm) holds due to stationarity of the process. Thus

$$\alpha(2) = \rho(X_3 - \frac{\theta}{1+\theta^2}X_2, X_1 - \frac{\theta}{1+\theta^2}X_2)$$

$$= \frac{-\frac{2\theta^2}{1+\theta^2} + \frac{\theta^2}{1+\theta^2}}{\mathrm{Var}(X_3 - \frac{\theta}{1+\theta^2}X_2)} = \frac{-\frac{\theta^2}{1+\theta^2}}{\mathrm{Var}(X_3 - \frac{\theta}{1+\theta^2}X_2)}.$$

We now compute $\mathrm{Var}(X_3 - \frac{\theta}{1+\theta^2}X_2)$

$$\mathrm{Var}(X_3 - \frac{\theta}{1+\theta^2}X_2) = 1 + \theta^2 + \frac{\theta^2}{1+\theta^2} - \frac{2\theta^2}{1+\theta^2} = \frac{1+\theta^2+\theta^4}{1+\theta^2}.$$

We thus obtain

$$\alpha(2) = \frac{-\theta^2}{1+\theta^2+\theta^4}.$$

Formula for $\alpha(n)$ *for arbitrary* n *is given in Problem 3.4.*

Let us also note that the following property of partial correlation holds:

$$\text{If } \ |\varphi_{nn}| = 1 \text{ then } \phi_{n+i,n+i} = 1, \quad \text{for } i \geqslant 1. \tag{3.29}$$

To see this assume for simplicity that m=0 and note that it is enough to consider the case when $\mathbf{\Gamma}_n$ is invertible. Then (3.27) holds and it follows from the Schwarz inequality that there exists a, such that $\mathrm{sign}(a) = \alpha(n)$ and

$$a(X_1 - \sum_{i=1}^{n-1} a_i X_{1+i}) = (X_{n+1} - \sum_{i=1}^{n-1} a_i X_{n+1-i}).$$

Thus X_{n+1} is a linear combination $X_1, \ldots, X_n$ and $\mathbf{\Gamma}_{n+i}$ is non-invertible for $i \geqslant 1$. It follows from the adopted convention that $\phi_{n+i,n+i} = \alpha(n+i) = 1$.
Let

$$S = \{(s_k)_0^\infty : \ |s_k| \leqslant 1 \ \forall k \text{ and if } |s_k| = 1 \Longrightarrow s_{k+1} = 1\}.$$

The following result holds, the second part of which has been already proved.

Theorem 3.2.4 *For any* $(s_k) \in S$ *there exists WS time series such that its partial autocorrelation function satisfies* $\alpha(n) = s_n$ *and for any WS time series its partial autocorrelation function belongs to* S.

Note that in view of Theorem 3.2.4 partial correlation coefficients provide *unconstrained* parametrization of the second order structure of the weakly stationary process in contrast to the autocovariance function. In the later case non-negative definiteness of the 'candidate' autocovariance has to be checked, whereas any sequence $(\alpha(n))_n$ such that $|\alpha(n)| < 1$ is valid sequence of PACFs. Thus PACFs should have a strong appeal to statisticians. Alas, mostly from computational reasons, it is not the case. We also remark that it follows e.g. from the Durbin-Levinson algorithm that if $\gamma(\cdot)$ is positive definite then we have $|\alpha(n)| < 1$ for any n.
The interesting question in the study of dependence of time series are conditions

which imply that $\alpha(n) \to 0$ when $n \to \infty$. This is answered by Rakhmanov's theorem which asserts that this happens when the density of the absolutely continuous part of the spectral measure is positive on the set of the full measure 2π.

3.2.1 Gaussian sequences

We briefly discuss the connections between autocorrelation and partial autocorrelation functions with information theoretic measures in the case when stationary sequence is Gaussian. Let $H(n)$ denote the entropy of the block X_1^n in (2.19). We note that all equalities stated in Section 2.6 Chapter 2 but the inequality $H(X) \geqslant 0$ valid only for discrete distributions, are also valid in the continuous case. It is proved in Cover and Thomas (2006), Theorem 4.8.1 that

$$H(n) = \frac{n}{2}\log(2\pi e) + \frac{1}{2}\log|\mathbf{\Gamma}_n|. \tag{3.30}$$

From this it follows that

Theorem 3.2.5 *For a stationary Gaussian $(X_t)_{t\in\mathbb{Z}}$ we have*

$$\begin{aligned} I(X_1; X_n) &= -\frac{1}{2}\log[1-\rho(n-1)^2] \\ I(X_1; X_n|X_2^n) &= -\frac{1}{2}\log[1-\alpha(n-1)^2] \end{aligned} \tag{3.31}$$

Proof. Note first that for a Gaussian pair (X, Y) we have

$$\begin{aligned} I(X;Y) &= H(X) + H(Y) - H(X,Y) \\ &= \frac{1}{2}(\log(2\pi e) + \log \mathrm{Var}(X) + -\log(2\pi e) + \log \mathrm{Var}(Y) \\ &- 2\log 2\pi e - \log[\mathrm{Var}(X)\mathrm{Var}(Y) - \mathrm{Cov}(X,Y)^2]) \\ &= -\frac{1}{2}\log[1-\rho(X,Y)^2]. \end{aligned} \tag{3.32}$$

This proves the first statement. In order to prove the second one note first that we have $H(f(X)|X) = 0$ and $H(X|Y) = H(X)$ if X and Y are independent. Thus we have

$$\begin{aligned} &I(X_1; X_n|X_2^{n-1}) = H(X_1|X_2^{n-1}) + H(X_n|X_2^{n-1}) - H(X_1, X_n|X_2^{n-1}) \\ &= H(X_1 - P_{sp(X_2,\ldots,X_n)}X_1) + H(X_n - P_{sp(X_2,\ldots,X_n)}X_n) \\ &- H(X_1 - P_{sp(X_2,\ldots,X_n)}X_1, X_n - P_{sp(X_2,\ldots,X_n)}X_n) \\ &= I(X_1 - P_{sp(X_2,\ldots,X_n)}X_1, X_n - P_{sp(X_2,\ldots,X_n)}X_n) \end{aligned} \tag{3.33}$$

and the second statement follows from the first one and the definition of the partial correlation. Note that the connection between $I(X_1; X_n|X_2^{n-1})$ and $\alpha(n)$ is clearly seen from the last equality in (3.33).

We also note that Gaussian vector $(Y_1, \ldots, Y_n)$ has the largest entropy among all n-dimensional vectors with the same covariance function $\mathbf{\Gamma}_n$. Indeed, let $(X_1, \ldots, X_n)$ be any such vector and denote by f_Y and f_X the respective densities. Then

$$\begin{aligned} &\int f_X(x_1, \ldots, x_n) \log f_Y(x_1, \ldots, x_n)\, dx_1 \ldots dx_n \\ &= \int f_Y(x_1, \ldots, x_n) \log f_Y(x_1, \ldots, x_n)\, dx_1 \ldots dx_n, \end{aligned}$$

both sides are quadratic forms of the same covariance matrix. Since

$$\int f_X(x_1, \ldots, x_n) \log \Big(\frac{f_Y(x_1, \ldots, x_n)}{f_X(x_1, \ldots, x_n)} \Big) dx_1 \ldots dx_n \leqslant 0$$

as $D(f_X || f_Y) \geqslant 0$, the both inequalities yield $H_n(X_1, \ldots, X_n) \leqslant H_n(Y_1, \ldots, Y_n)$. This is sometimes called maximum entropy principle.

3.3 The innovations algorithm

We show now how to compute coefficients of so called innovations representation of the projection of X_{n+1} on $\mathrm{sp}(X_1, \ldots, X_n)$,

$$\hat{X}_{n+1} = \sum_{j=1}^{n} \theta_{nj} (X_{n+1-j} - \hat{X}_{n+1-j}) = \sum_{j=0}^{n-1} \theta_{n,n-j} (X_{j+1} - \hat{X}_{j+1}) \tag{3.34}$$

for $n \geqslant 1$ and $\hat{X}_{n+1} = 0$ for $n = 0$. As before we consider zero-mean WS time series and assume that $\mathbf{\Gamma}_n$ is invertible. $\hat{X}_{j+1}$ denotes the orthogonal projection of X_{j+1} on $sp(X_1, \ldots, X_j)$. X_{n+1} can be represented as in (3.34) due to the fact that linear space $sp(X_1, \ldots, X_n)$ equals $sp(X_1 - \hat{X}_1, \ldots, X_n - \hat{X}_n)$. This equality in fact yields representation of $\mathrm{sp}(X_1, \ldots, X_n)$ as a direct sum of univariate orthogonal subspaces generated by $X_1 - \hat{X}_1, \ldots, X_n - \hat{X}_n$. Note that a corresponding matrix form of (3.34) is

$$\mathbf{X}_{n+1} = \mathbf{\Theta}_{n+1} (\mathbf{X}_{n+1} - \hat{\mathbf{X}}_{n+1}), \tag{3.35}$$

where $\mathbf{\Theta}_{n+1}$ is a lower triangular matrix defined as

$$\begin{pmatrix} 1 & 0 & \ldots & 0 & 0 \\ \theta_{11} & 1 & \ldots & 0 & 0 \\ \vdots & & & \vdots & \vdots \\ \theta_{nn} & \theta_{n,n-1} & \ldots & \theta_{n-1,1} & 1 \end{pmatrix} \tag{3.36}$$

Obviously, in view of (3.16) we have $\mathbf{\Theta}_{n+1} \mathbf{\Phi}_{n+1} = \mathbf{I}$. It follows from (3.35) that

$$\mathbf{\Gamma}_{n+1} = \mathbf{\Theta}_{n+1} \mathbf{D}_{n+1} \mathbf{\Theta}'_{n+1}, \tag{3.37}$$

where $\mathbf{D}_{n+1} = \mathrm{diag}(\sigma_0^2, \ldots, \sigma_n^2)$. Thus innovation algorith discussed below yields modified Cholesky decomposition (3.37).
We also note that innovation parameters in (3.34) are uniquely defined provided $\mathbf{\Gamma}_{n+1}$ is invertible. Representation (3.34) has a substantial advantage over usual representation $\hat{X}_{n+1} = \sum_{j=1}^{n} \phi_{nj} X_{n+1-j}$, namely the summands of the former are uncorrelated:

$$X_{n+1-j} - \hat{X}_{n+1-j} \perp X_{n+1-k} - \hat{X}_{n+1-k}$$

for $k \neq j$. We assume that $\mathbf{\Gamma}_n$ is invertible which implies that $\sigma_i^2 > 0$ for $i = 1, \ldots, n-1$. As we recall coefficients corresponding to projection onto an orthogonal set have very simple form

$$\theta_{nj} = \frac{< X_{n+1}, X_{n+1-j} - \hat{X}_{n+1-j} >}{\sigma_{n-j}^2}. \tag{3.38}$$

Thus $\theta_{n,j}$ is simply projection coefficient of X_{n+1} onto $X_{n+1-j} - \hat{X}_{n+1-j}$.
Coefficients θ_{nj} may be recursively computed based on previous coefficients θ_{kl} such that either $k < n$ or $k = n, l > j$ and prediction error σ_i^2 for $i \leqslant n-1$. Next, σ_n^2 is computed.
Indeed, applying (3.38) for $j = n-k$ and representation (3.34) with k in lieu of n we obtain

$$\begin{aligned}
\theta_{n,n-k} &= \frac{< X_{n+1}, X_{k+1} - \hat{X}_{k+1} >}{\sigma_k^2} \\
&= \frac{\gamma(n-k) - \sum_{j=1}^{k} \theta_{kj} < X_{n+1}, X_{k+1-j} - \hat{X}_{k+1-j} >}{\sigma_k^2} \\
&= \frac{\gamma(n-k) - \sum_{j=0}^{k-1} \theta_{k,k-j} < X_{n+1}, X_{j+1} - \hat{X}_{j+1} >}{\sigma_k^2}.
\end{aligned}$$

As it follows from (3.38) that $< X_{n+1}, X_{j+1} - \hat{X}_{j+1} >= \theta_{n,n-j}\sigma_j^2$, we obtain

$$\theta_{n,n-k} = \frac{\gamma(n-k) - \sum_{j=0}^{k-1} \theta_{k,k-j}\theta_{n,n-j}\sigma_j^2}{\sigma_k^2}, \tag{3.39}$$

for $k = 0, 1, \ldots, n-1$. Knowing values of $\theta_{n,n-k}$ we can compute σ_n^2:

$$\sigma_n^2 = ||X_{n+1} - \hat{X}_{n+1}||^2 = \gamma(0) - ||\hat{X}_{n+1}||^2 = \gamma(0) - \sum_{k=0}^{n-1} \theta_{n,n-k}^2 \sigma_k^2. \tag{3.40}$$

Innovation algorithms is based on equalities (3.39) and (3.40). Order of computing coefficients is as follows: $\sigma_0^2, \theta_{11}, \sigma_1^2, \theta_{22}, \theta_{21}, \sigma_2^2, \theta_{33}, \theta_{32}, \theta_{31}$ and so on. As we shall see later the innovation algorithm is instrumental for estimation of parameters of Gaussian ARMA(p, q) process.

We note that stationarity was actually not needed in the development of the algorithm which remains valid for any 0 mean time series when the term $\gamma(n-k)$ is replaced by $\gamma(n,k)$. Then equation (3.39) is generalized to

$$\theta_{n,n-k} = \frac{\gamma(n+1,k+1) - \sum_{j=0}^{k-1} \theta_{k,k-j}\theta_{n,n-j}\sigma_j^2}{\sigma_k^2} \tag{3.41}$$

and equation (3.40) to

$$\sigma_n^2 = \gamma(n+1,n+1) - \sum_{k=0}^{n-1} \theta_{n,n-k}^2 \sigma_k^2. \tag{3.42}$$

Moreover, if $\gamma(n-j) = 0$ for all $j \leqslant k$ (or, in general case $\gamma(n+1,j+1) = 0$ for all $j \leqslant k$) then we have $\theta_{n,n-j} = 0$ for such j. Indeed, it follows from the proof that $\sigma_j^2\theta_{n,n-j} =< X_{n+1}, X_{j+1} - \hat{X}_{j+1} >$ and the observation follows from orthogonality of X_{n+1} to X_{j+1} and to $\hat{X}_{j+1}$ and the fact that $\sigma_j^2 > 0$.
We finally note that as $X_{n+1} = \hat{X}_{n+1} + (X_{n+1} - \hat{X}_{n+1})$, (3.34) may be re-expressed in the form

$$X_{n+1} = \sum_{j=0}^{n} \theta_{nj}(X_{n+1-j} - \hat{X}_{n+1-j}),$$

where $\theta_{n,0} = 1$.

Example 3.3.1 *We apply the innovation algorithm to MA(1) process of the form* $X_t = \varepsilon_t + \theta\varepsilon_{t-1}$. *As* $\gamma(k) = 0$ *for* $k \geqslant 2$ *it follows from the remark below (3.42) and (3.39) that for* $n \geqslant 2$, *coefficient* $\theta_{nn} = \theta_{n,n-1} = \ldots = \theta_{n2} = 0$ *and* $\theta_{n1} = \sigma_{n-1}^{-2}\theta\sigma^2$. *Moreover,* $\sigma_0^2 = (1+\theta^2)\sigma^2$ *and*

$$\sigma_n^2 = \gamma(0) - \theta_{n1}^2\sigma_{n-1}^2 = (1 + \theta^2 - \sigma_{n-1}^{-2}\theta^2\sigma^2)\sigma^2,$$

thus letting $r_n = \sigma_n^2/\sigma^2$ *we can write this as* $r_n = 1+\theta^2-\theta^2/r_{n-1}$. *As* $r_0 = 1+\theta^2$ *we observe that* (r_i) *does not depend on* σ^2. *This is a special case of a general property true for causal ARMA processes proved in Chapter 4 which we will use later.*

Now for h-step ahead prediction we use easily verifiable equalities

$$P_nX_{n+h} = P_nP_{n+h-1}X_{n+h} = P_n\Big(\sum_{j=1}^{n+h-1} \theta_{n+h-1,j}(X_{n+h-j} - \hat{X}_{n+h-j})\Big) \tag{3.43}$$

and since $(X_{n+h-j} - \hat{X}_{n+h-j})$ is orthogonal to X_k for $j < h$ and $k \leqslant n$ and for $j \geqslant h$ it belongs to $sp(X_1,\ldots,X_n)$, we obtain

$$P_nX_{n+h} = \sum_{j=h}^{n+h-1} \theta_{n+h-1,j}(X_{n+h-j} - \hat{X}_{n+h-j}). \tag{3.44}$$

3.4 Problems

1. Show that condition that $\mathbf{\Gamma}_n = (\gamma(i-j))_{1\leqslant i,j\leqslant n}$ is non-degenerate is equivalent to the following statement: there does not exist non-zero vector $\mathbf{a} = (a_1, \ldots, a_n)'$ such that $\sum_{i=1}^n a_i X_i$ is constant almost everywhere i.e. $1, X_1, \ldots, X_n$ are not linearly dependent. Hint: note that the fact that $\mathbf{\Gamma}_n$ is degenerate is equivalent to existence of eigenvector $\mathbf{a} \neq 0$ corresponding to zero eigenvalue and thus $\mathbf{a}'\mathbf{\Gamma}_n\mathbf{a} = 0$.
2. State and prove Durbin-Levinson's algorithm for h-step prediction.
3. Prove that in general for a WS time series we have

$$\alpha(2) = \frac{\rho(2) - \rho(1)^2}{1 - \rho(1)^2}$$

and check that it is consistent with the calculations of the Example 3.2.3.
4. Prove that partial correlation coefficient $\alpha(n)$ for MA(1) time series equals

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

Hint . Compute $\mathbf{\Gamma}_n$ and use the Yule-Walker equations.
5. Autocovariance function of weakly stationary time series equals $\gamma(0) = 2$, $\gamma(\pm 1) = -1$ and zero for the remaining lags. Compute $\alpha(2)$ and then $\alpha(n)$.
6. Prove that prediction error $\sigma^2 = ||X_t - P_{H_{t-1}}X_t||^2 = \gamma(0)\prod_{i=0}^{\infty}(1 - \alpha^2(i))$.
7. Show that weakly stationary mean-zero time series is PND (i.e. $H_{-\infty} = 0$) if and only if for any t $P_{H_{t-s}}X_t \to 0$ when $s \to \infty$. Hint. Note that $H_{-\infty} = \bigcap H_t$ implies $P_{H_{t-s}}X_t \to P_{H_{-\infty}}X_t$. On the other hand from $P_{H_{-\infty}}X_t = 0$ for any t it follows that $P_{H_{-\infty}}Y = 0$ for $Y \in H_{-\infty}$.
8. Prove that if $\phi_1, \ldots, \phi_p$ are prediction coefficients of X_{p+1} on $sp\{X_1, \ldots, X_p\}$ for mean- zero weakly stationary process, then $\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p \neq 0$ for z such that $|z| \leqslant 1$. 9. Show that prediction error for h-step prediction in (3.44) equals

$$E(X_{n+h} - P_n X_{n+h})^2 = \gamma(0) - \sum_{j=h}^{n+h-1} \theta_{n+h-1}^2 \sigma_{n+h-j-1}^2.$$

10. Express (3.9) in terms of autocorrelation matrix $\Delta_n = \gamma(0)^{-1}\Gamma_n$ and autocorrelation vector ρ_n.

4

ARMA(p, q) processes

We define now a basic class of linear processes introduced by Box and Jenkins (see Box et al. (2008) for a thorough treatment of ARMA processes).

4.1 Definitions and examples

Definition 9 *$(X_t)_{t\in\mathbb{Z}}$ is real-valued mean zero ARMA(p, q) time series if the following two conditions are satisfied:*

(i) $(X_t)_{t\in\mathbb{Z}}$ is weakly stationary time series;
(ii) $(X_t)_{t\in\mathbb{Z}}$ satisfies a structural equation

$$X_t - \varphi_1 X_{t-1} - \cdots - \varphi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q} \tag{4.1}$$

for certain $\varphi_1, \ldots, \varphi_p, \theta_1, \ldots, \theta_q \in \mathbb{R}$, where $(Z_t)_{t\in\mathbb{Z}}$ is white noise WN$(0, \sigma^2)$ with $\sigma^2 > 0$ and $p, q \in \mathbb{N} \cup \{0\}$. $(X_t)_{t\in\mathbb{Z}}$ ARMA(p, q) with the mean μ, if $X_t - \mu$ is zero-mean ARMA(p, q). In this case the structural equation is

$$X_t - m - \varphi_1(X_{t-1} - \mu) - \cdots - \varphi_p(X_{t-p} - \mu) = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}, \tag{4.2}$$

or equivalently, letting $\varphi_0 = \mu(1 + \phi_1 + \ldots + \phi_p)$

$$X_t - \phi_0 - \varphi_1 X_{t-1} - \cdots - \varphi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}. \tag{4.3}$$

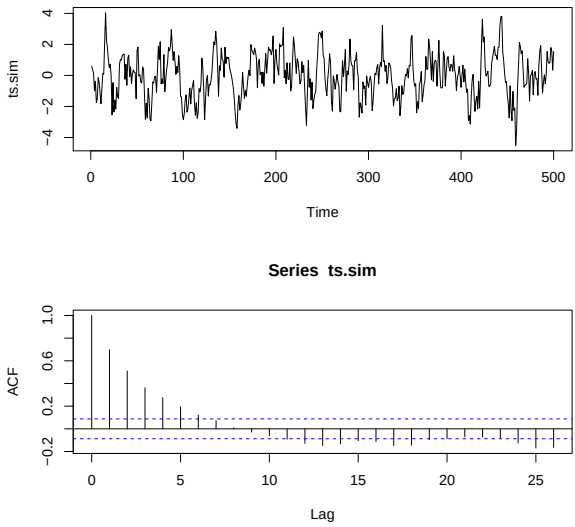
Note that we do not know that a time series satisfying (4.1) exists and whether, if it satisfies (4.1), it is weakly stationary. The existence of ARMA(p, q) time series will be discussed below. We consider two special cases of the definition above.

1. Autoregressive process of order p. Let $q = 0$ and consider the equation for process ARMA$(p, 0)$ denoted by AR(p):

$$X_t - \varphi_1 X_{t-1} - \cdots - \varphi_p X_{t-p} = Z_t. \tag{4.4}$$

Thus X_t is a linear combination of $X_{t-1}, \ldots, X_{t-p}$ with added noise. This is a usual regression equation with p predictors. In order to underline the fact that the predictors are given by previous values of the process, weakly stationary time series satisfying (4.4) is called autoregressive process of order p.

Example 4.1.1 *We will show below that AR(1) time series for* $|\phi_1| < 1$ *exists. The plot below shows simulated sample path* ($n = 500$) *for AR(1) time series with* $\phi_1 = 0.7$ *and related empirical autocorrelation coefficients. Note that for the smaller lags, say* $h \leqslant 5$*, for which autocorrelations are outside confidence band they appear to decay exponentially.*

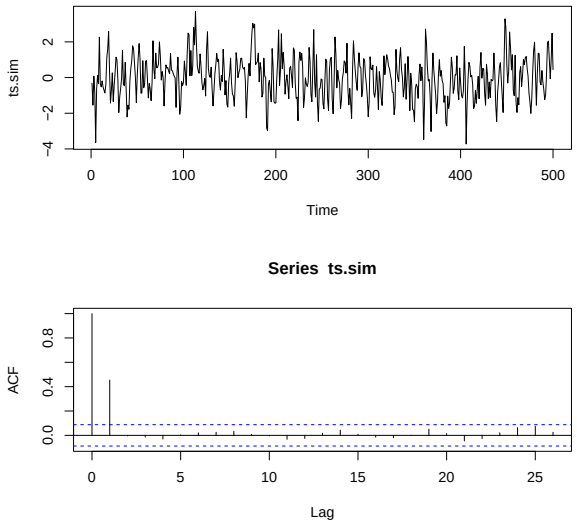


2. Moving average process of order q. Analogously, we let $p = 0$ and consider ARMA$(0, q)$ time series denoted by MA(q). The structural equation is

$$X_t = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}. \tag{4.5}$$

In this case we know from equation (4.5) that the MA(q) exists. Indeed, if we consider time series defined as the right hand side of (4.5) then such time series is weakly stationary as it is a special case of linear process MA(∞).

Example 4.1.2 *We also generated sample path* ($n = 500$) *for MA(1) time series with* $\theta = 0.7$ *and empirical partial correlation coefficients. Note that the plot of empirical autocorrelations looks qualitatively different from the analogous plot for the sample path of AR(1) process.*



Before we discuss existence of ARMA(p, q) process we introduce two useful properties of such time series.

Definition 10 *ARMA(p, q) time series $(X_t)_{t\in\mathbb{Z}}$ is causal with respect to white noise $(\varepsilon_t)_{t\in\mathbb{Z}}$ if for a certain $(\psi_t) \in \ell^1$*

$$X_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}. \tag{4.6}$$

Note that the righthand side of (4.6) is well defined even under weaker assumption that $(\psi_t) \in \ell^2$. This follows from Example 1.3.3 (ii). However, a little bit stronger assumption will enable us more convenient statement of Theorem 4.2.2 below. Note that in the case when we assume only that $(\psi_j) \in \ell^2$, $(X_t)_{t\in\mathbb{Z}}$ is one-sided linear process with innovations $(\varepsilon_t)_{t\in\mathbb{Z}}$. When (4.6) is satisfied, note that X_s is uncorrelated with ε_t for $s < t$, that is

$$\mathrm{Cov}(\varepsilon_t, X_s) = 0, \quad t > s.$$

Moreover, then (X_t) belongs to a closure of linear space $sp\{\varepsilon_s,\ s \leqslant t\}$, thus $H_t(X) \subseteq H_t(\varepsilon)$. Concept of invertibility of (X_t) naturally arises when we interchange the roles of X_t and (ε_t). Namely

Definition 11 *ARMA(p, q) time series $(X_t)_{t\in\mathbb{Z}}$ is invertible with respect to $(\varepsilon_t)_{t\in\mathbb{Z}}$ if for a certain $(\pi_i)_0^\infty \in \ell^1$*

$$\varepsilon_t = \sum_{i=0}^{\infty} \pi_i X_{t-i} \tag{4.7}$$

We can write (4.7) equivalently as

$$\pi_0 X_t = \varepsilon_t - \sum_{j=1}^{\infty} \pi_j X_{t-j}.$$

which is autoregressive equation of infinite order (AR(∞) representation).
We show in Lemma 4.2.1 that when $(\pi_i)_0^\infty \in \ell^1$ right hand side of (4.7) is well defined. Note that in this case $H_t(\varepsilon) \subseteq H_t(X)$. If (X_t) is at the same time causal and invertible with respect to white noise (ε_t), then we have $H_t(\varepsilon) \subseteq H_t(X)$ and $H_t(X) \subseteq H_t(\varepsilon)$ and thus equality $H_t(\varepsilon) = H_t(X)$ holds. As variables (ε_t) are uncorrelated it is easy to write down in such situation projection of X_{t+1} on $H_t(X)$ in terms of (ε_s) for $s \leqslant t$. Namely $\hat{X}_{t+1} = \sum_{i=1}^{\infty} \psi_i \varepsilon_{t-i}$, where $\psi_i = \sigma^{-2} < X_t, \varepsilon_{t-i} >$.

Example 4.1.3 *(i) Assume momentarily that AR(1) with parameter $\phi = \phi_1$ such that $|\phi| < 1$ exists and try to derive its expansion in terms of ε_t. Namely, for any $k \in N$*

$$\begin{aligned} X_t &= \phi X_{t-1} + \varepsilon_t = \phi(\phi X_{t-2} + \varepsilon_{t-1}) + \varepsilon_t = \phi^2 X_{t-2} + \phi\varepsilon_{t-1} + \varepsilon_t = \cdots \\ &= \sum_{i=0}^{k} \phi^i \varepsilon_{t-i} + \phi^{k+1} X_{t-k-1}. \end{aligned}$$

Note that as X_t is stationary then it follows that $||\phi^{k+1} X_{t-k-1}|| \leqslant |\phi|^{k+1} ||X_1|| \to 0$, since $|\phi| < 1$. Thus if AR(1) process with mean 0 and parameter $|\phi| < 1$ exists, it has the following causal representation with respect to white noise (ε_t)

$$X_t = \sum_{i=0}^{\infty} \phi^i \varepsilon_{t-i}.$$

Obviously, coefficients $(\phi^i)_{i=0}^\infty$ are absolutely summable. Note that if we define process (X_t) by the right hand side of the above equation it will satisfy the structural equation of AR(1) time series. Indeed

$$X_t = \sum_{i=0}^{\infty} \phi^i Z_{t-i} = \sum_{i=0}^{\infty} \phi^{i+1} \varepsilon_{t-(i+1)} + \varepsilon_t = \phi \sum_{i=0}^{\infty} \phi^i \varepsilon_{t-1-i} + \varepsilon_t = \phi X_{t-1} + \varepsilon_t.$$

Thus we proved existence of AR(1) process by guessing its causal representation and checking that this representation indeed yields AR(1) time series. Actually, this is a general procedure in the case of causal ARMA processes (see proof of Theorem 4.2.2).
(ii) Consider now MA(1) process with $|\theta| < 1$. Then we have

$$\varepsilon_t = X_t - \theta\varepsilon_{t-1} = X_t - \theta(X_{t-1} - \theta\varepsilon_{t-2})$$

$$= X_t - \theta X_{t-1} + \theta^2(X_{t-2} - \theta\varepsilon_{t-3}) = \cdots = \sum_{j=0}^{\infty}(-1)^j\theta^j X_{t-j}, \qquad (4.8)$$

where the last equality follows from the fact that $\theta^i\varepsilon_{t-i-1} \to 0$ *when* $i \to \infty$. *Thus we directly established that MA(1) process with* $|\theta| < 1$ *is invertible.*

Causality and invertibility defined above for ARMA processes may be defined in a completely analogous for any weakly stationary time series. More generally still, we can define these concepts with respect to an arbitrary weakly stationary time series Z_t. In order to ensure that the right hand sides of analogues of (4.6) and (4.7) are well defined in both cases we have to assume that coefficients $(\psi)_{i=0}^{\infty}$ and $(\pi)_{i=0}^{\infty}$ belong to ℓ^1 (are absolutely summable). This follows from Lemma 4.2.1 in the next section.

4.2 Causal and invertible ARMA processes

Below we will deal (in reversed order) with the problem of existence of ARMA(p, q) process, its causality and invertibility.
We define first backward shift operator B. For a given $(X_t)_{t\in\mathbb{Z}}$ we define

$$BX_t := X_{t-1}, \quad t \in \mathbb{Z} \qquad (4.9)$$

Thus j-times superposition of B is

$$B^j X_t = B \circ \cdots \circ BX_t = X_{t-j},$$

where in particular $B^0 X_t = \mathrm{Id}(X_t) = X_t$.
We define two complex polynomials related to ARMA structural equation

$$\varphi(z) = 1 - \varphi_1 z - \cdots - \varphi_p z^p$$

and

$$\theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q.$$

Related operators $\varphi(B)$ and $\theta(B)$ arise after formally plugging B as argument $\varphi(\cdot)$ and $\theta(\cdot)$ and interpreting B^j as j times superposition of B as above. This yields

$$\varphi(B) = Id - \varphi_1 B - \cdots - \varphi_p B^p$$

and analogously

$$\theta(B) = Id + \theta_1 B + \cdots + \theta_q B^q.$$

Note that due to sign convention in the structural equation, signs of coefficients in $\varphi(B)$ and $\theta(B)$ differ. Now structural equation (4.1) can be succinctly written as

$$\varphi(B)X_t = \theta(B)Z_t \tag{4.10}$$

corresponding to

$$\varphi(B)X_t = (B^0 - \varphi_1 B - \cdots - \varphi_p B^p)X_t = X_t - \varphi_1 X_{t-1} - \cdots - \varphi_p X_{t-p}$$

and

$$\theta(B)Z_t = (B^0 + \theta_1 B + \cdots + \theta_q B^q)Z_t = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}.$$

Now we address the question of existence and causal/invertible representation of ARMA time series. The following heuristics is helpful. Divide formally both sides of (4.10) by $\varphi(B)$ and consider process

$$\frac{\theta(B)}{\varphi(B)}\varepsilon_t$$

where $\theta(B)/\varphi(B)$ is interpreted as follows. We consider expansion of rational function $\theta(x)/\varphi(x)$ into infinite series and plug in shift operator B to obtain the definition of the process above. Theorems below specify conditions under which this process is correctly defined (what answers the question of existence of ARMA process) and is one-sided moving average (what solves the problem of its causal representation). We start with the later problem, which is simpler. First we prove a preliminary useful lemma.

Lemma 4.2.1 *Let $(X_t)_{t\in T}$ be arbitrary time series such that $\sup_t E|X_t| < \infty$ and $(\psi_j)_{j=-\infty}^{j=\infty}$ is such that $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$.*
(i) Then double-sided linear process

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

is well defined almost surely.
(ii) If additionally (X_t) is weakly stationary, then $(Y_t)_{t\in T}$ belongs to $\mathcal{L}^2$, is weakly stationary and

$$\gamma_Y(h) = \sum_{j,k} \psi_j \psi_k \gamma_X(h - j + k). \tag{4.11}$$

We note that we proved in Chapter 1 that when $(X_t)_{t\in T}$ is white noise than (ii) is satisfied under less stringent condition on (ψ_j), namely that $(\psi_j) \in \ell^2$.
Proof of (i). As for any random variable W we have that $E|W| < \infty$ implies $|W| < \infty$ almost surely, we show that

$$E\left(\sum_{j=-\infty}^{\infty} |\psi_j| \, |X_{t-j}|\right) < \infty.$$

Indeed, using the Lebesgue monotone convergence theorem for the first equality below we have

$$E\big(\sum_{j=-\infty}^{\infty}|\psi_j|\,|X_{t-j}|\big)=\lim_{n\to\infty}E\big(\sum_{j=-n}^{n}|\psi_j|\,|X_{t-j}|\big)$$
$$\leqslant\limsup_{n\to\infty}E|X_t|\sum_{j=-n}^{n}|\psi_j|<\infty.$$

(ii) Let $S_{t,n}=\sum_{|j|\leqslant n}\psi_jX_{t-j}$. We check the Cauchy condition for $(S_{t,n})_n$ in $\mathcal{L}^2$. The Schwarz inequality implies

$$|\sum_{m<|j|\leqslant n}\psi_jX_{t-j}\,|^2=\sum_{m<|j|,|k|\leqslant n}|\psi_j\psi_k|E|X_{t-j}X_{t-k}|\leqslant$$
$$\leqslant E|X_t|^2\big(\sum_{m<|j|\leqslant n}|\psi_j|\big)^2\longrightarrow 0,\quad\text{where } m,n\to\infty.$$

Obviously as limit of $S_{t,n}$ in $\mathcal{L}^2$ which is established above is also limit in probability, thus it is determined almost everywhere and it coincides with the limit in (i).
The form of covariance function of (Y_t) follows from covariance of $S_{t,n}=\sum_{|j|\leqslant n}\psi_jX_{t-j}$ and continuity of scalar product in w $\mathcal{L}^2$ (Problem 4.6 , see also Example 1.3.3).

Theorem 4.2.2 *Assume that $(X_t)_{t\in\mathbb{Z}}$ is ARMA(p,q) time series such that polynomials $\varphi(\cdot)$ i $\theta(\cdot)$ do not have common roots in $\mathbb{C}$. Then (X_t) is causal if and only if $\varphi(z)$ does not have zeros in the closed unit disc i.e. $\varphi(z)\neq 0$ for $z\in\mathbb{C}:\ |z|\leqslant 1$.*

Proof. We first prove that the condition $\varphi(z)\neq 0$ for $z\in\mathbb{C}:\ |z|\leqslant 1$ is sufficient. Let $\xi(z)=\varphi(z)^{-1}$ and note that it is analytic in the closed unit disc thus can be analytically extended to its open neighborhood that is there exists $\varepsilon>0$ such that

$$\xi(z)=\sum_{j=0}^{\infty}\xi_jz^j,\quad\text{for } |z|\leqslant 1+\varepsilon.$$

It is easy to check now that $(\xi_j)_{j=0}^{\infty}$ is absolutely summable by taking $z=1+\varepsilon/2$ for which $\xi(z)$ is convergent and noting that $|\xi_j|(1+\varepsilon/2)^j\ \to 0$ implies $\sum_{j=0}^{\infty}|\xi_j|<\infty$. It is also easy to check that processes $\varphi(B)X_t$ and $\theta(B)Z_t$ are weakly stationary. Thus in view of Lemma 4.2.1 application of $\xi(B)$ to any of them yields weakly stationary time series. Applying $\xi(B)$ to both sides of the structural equation we obtain

$$\xi(B)\varphi(B)X_t=\xi(B)\theta(B)Z_t.$$

Moreover note that $\xi(B)\varphi(B) = \mathrm{Id} \ \ (= B^0)$ that is $\xi(B)\varphi(B)X_t = X_t$. Thus the equality above yields causal representation of X_t as it is easy to see that coefficients of $\xi(z)\theta(z)$ are absolutely summable and whence square summable. In order to prove that the assumption $\varphi(z) \neq 0$ for $z \in \mathbb{C}$ such that $|z| \leqslant 1$ is necessary, suppose that

$$X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j} \ \text{ for some } \ (\psi_j), \ \ \sum_{j=0}^{\infty} |\psi_j| < \infty.$$

Let

$$\eta(z) := \varphi(z)\psi(z) =: \sum_{j=0}^{\infty} \eta_j z^j, \quad |z| \leqslant 1$$

Note that $\varphi(B)X_t = \theta(B)Z_t$, the fact that $\varphi(B)\psi(B) = \psi(B)\varphi(B)$ and $X_t = \psi(B)Z_t$ imply

$$\sum_{j=0}^{q} \theta_j Z_{t-j} = \varphi(B)X_t = \varphi(B)\psi(B)Z_t = \sum_{j=0}^{\infty} \eta_j Z_{t-j}.$$

Calculating scalar products of both sides with $Z_{t-k}, k \geqslant 0$ we obtain $\eta_k = \theta_k, \quad$ for $k = 0, 1, \ldots, q$ and $\eta_k = 0, \quad$ for $k > q$. This implies that $\theta(z) = \eta(z) = \varphi(z)\psi(z)$ and as $|\psi(z)| < \infty$ for $|z| \leqslant 1$ and $\theta(\cdot)$ i $\varphi(\cdot)$ do not have common roots, then $\varphi(z) \neq 0$ for $|z| \leqslant 1$. Note that to ensure the property $|\psi(z)| < \infty$ for $|z| \leqslant 1$ we have to assume that $(\psi_i) \in \ell^1$.

It follows from sufficiency proof that ARMA(p,q) time series such that $\varphi(\cdot)$ does not have roots in the unit disc is causal without assuming that $\phi(z)$ i $\theta(z)$ do not have common roots. Moreover, it follows from equation $\phi(z)\psi(z) = 1$ that $\psi_0 = 1$.
Using analogous reasoning as in the proof of the previous theorem we obtain equivalent condition for invertibility provided $\phi(z)$ i $\theta(z)$ do not have common roots.

Theorem 4.2.3 *Assume that $(X_t)_{t\in\mathbb{Z}}$ is ARMA(p,q) time series such that polynomials $\varphi(\cdot)$ i $\theta(\cdot)$ do not have common zeros in $\mathbb{C}$. Then (X_t) is invertible if and only if $\theta(z)$ does not have zeros in the closed unit disk i.e. $\theta(z) \neq 0$ for $z \in \mathbb{C} : |z| \leqslant 1$.*

As for the causality we also have that to obtain invertible representation one does not need to assume that $\varphi(\cdot)$ i $\theta(\cdot)$ do not have common zeros in $\mathbb{C}$, i.e. any ARMA(p,q) time series such that $\theta(\cdot)$ does not have roots in unit disk is invertible. Moreover, $\pi(z)\theta(z) = 1$ implies that $\pi_0 = 1$.
Now we address the question of existence of stationary solutions of structural equation i.e. the problem of existence of ARMA(p,q) process. Note that stationary solution does not always exist, for example it is easy to check that the solution of $X_t - X_{t-1} = Z_t$, where (Z_t) is WN$(0, \sigma^2)$ is not be stationary as it

implies that $X_t = X_0 + \sum_{i=1}^{t} Z_i$ and $\mathrm{Var}(X_t)$ is not constant. Indeed, it follows from the Schwarz inequality that

$$\mathrm{Var}(X_t) \geqslant \mathrm{Var}(X_0) + t\sigma^2 - 2(\mathrm{Var}(X_0)t\sigma^2)^{1/2} > \mathrm{Var}(X_0),$$

where the last inequality holds for large t. Note that in this case $\varphi(z) = 1 - z$ has root $z_0 = 1$ on the unit circle. In the next result we state sufficiency condition for existence and uniqueness of weakly stationary solution for the structural equation. We do *not* assume that the polynomials $\phi(\cdot)$ and $\theta(\cdot)$ do not have common roots.

Theorem 4.2.4 *If* $\varphi(z) \neq 0$ *for* $|z| = 1$ *then* $\varphi(B)X_t = \theta(B)Z_t$ *has the unique weakly stationary solution and it has the form*

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

where (ψ_j) *are given by the expansion*

$$\sum_{j=-\infty}^{\infty} \psi_j z^j = \frac{\theta(z)}{\varphi(z)}$$

and $\sum |\psi_j| < \infty$.

Proof. We prove first that stationary solution exists. As $\varphi(z) \neq 0$ for $|z| = 1$, thus $\psi(z) = \theta(z)/\varphi(z)$ is analytic on the circle $\{z : |z| = 1\}$ and can be analytically extended to some its open neighbourhood where it has Laurent expansion. Thus for a certain $\delta > 0$ we have

$$\frac{\theta(z)}{\varphi(z)} = \sum_{j=-\infty}^{\infty} \psi_j z^j, \qquad 1 - \delta < |z| < 1 + \delta.$$

Reasoning as in the proof of Theorem 4.2.2 we conclude that $\psi_j(1 + \delta/2)^j \to 0$ when $j \to \infty$ and $\psi_j(1 - \delta/2)^j \to 0$ when $j \to -\infty$ imply that $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$. In view of Lemma 4.2.1

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

exists and is weakly stationary. Application of $\phi(B)$ to both sides of above equality yields

$$\varphi(B)X_t = \varphi(B)\psi(B)X_t = \theta(B)X_t,$$

where the last equality follows from the definition of $\theta(\cdot)$. Thus $(X_t)_{t\in\mathbb{Z}}$ is weakly stationary solution to the structural equation of ARMA(p, q) process.
In order to prove uniqueness consider an arbitrary solution $(X_t)_{t\in\mathbb{Z}}$ to the structural equation. The first part of the proof gives

$$\xi(z) = \frac{1}{\varphi(z)} = \sum_{j=-\infty}^{\infty} \xi_j z^j, \quad 1-\delta < |z| < 1+\delta$$

Applying $\xi(B)$ to both sides of the structural equation

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

But $\xi(B)\theta(B) = \psi(B)$ defined in the first part of the proof and the last equality shows that (X_t) has to coincide with the previously obtained solution.
We stress that for the existence of the stationary version proved in the last result it is important that (X_t) starts at minus infinity. It follows e.g. from Example 4.1.3 that AR(1) process starting at 0 is not necessarily stationary even when $|\phi| < 1$.
We also note that in Example 6.2.8 we will show that in the case when $\phi(\cdot)$ and $\theta(\cdot)$ do not have zeros with $|z| = 1$ it is possible to find causal and invertible ARMA process having the same covariance structure as the original process.
We briefly discuss the case when autoregressive and moving average polynomials have common roots. The uniqueness of the solution depends then on whether φ has roots belonging to the unit circle or not. We consider first the case when φ does not have roots on the unit circle. In view of Theorem 4.2.4 if $\varphi(z) \neq 0$ for $|z| = 1$ the solution to problem pertaining to (θ, φ) exists and is unique. Let $\varphi = \eta\tilde{\varphi}$ and $\theta = \eta\tilde{\theta}$, where η is the greatest common divisor of φ i θ, which means that $\tilde{\varphi}$ i $\tilde{\theta}$ do not have common roots. Then it follows from the construction in the last proof that the solution for the pair (θ, φ) coincides with the unique solution for the pair $(\tilde{\theta}, \tilde{\varphi})$ and has representation

$$\frac{\eta\tilde{\theta}(B)}{\eta\tilde{\varphi}(B)} Z_t = \frac{\tilde{\theta}(B)}{\tilde{\varphi}(B)} Z_t.$$

The above equality is due to the fact that $\eta(z) \neq 0$ for $|z| = 1$. The outcome is different when the greatest common divisor η has a root z_0 on the unit circle. For any stationary solution X_t consider a process $W_t = Z z_0^{-t}$, where Z is an arbitrary zero-mean random variable in $\mathcal{L}^2$. Note that such process is weakly stationary due to the fact that $|z_0| = 1$. Note also that we have $(B - z_0 B^0)W_t = 0$ and since z_0 is a root of η and thus $\eta(z) = (z - z_0)t(z)$ it holds that $\eta(B)W_t = 0$. Since η is a common divisor of φ and θ we have $\varphi(B)W_t = 0$ and $\theta(B)W_t = 0$. It follows that $X_t + W_t$ is also stationary solution to structural equation and we do not have uniqueness of the ARMA process satisfying structural equation corresponding to (φ, θ) in this case.

Example 4.2.5 *We consider anew AR(1) time series*

$$X_t = \varphi X_{t-1} + \varepsilon_t$$

where ε_t is WN$(0, \sigma^2)$ in two cases (i) $|\varphi| < 1$ and (ii) $|\varphi| > 1$. Note that for $\varphi = 1$ we obtain the random walk. In case (i) we have shown that there exists

unique solution (X_t) which is causal process AR(1). The causal solution can be also easily obtained from the proof of the Theorem 4.2.3, namely

$$X_t = \frac{1}{1-\varphi B}\varepsilon_t = \sum_{i=0}^{\infty} \varphi^i Z_{t-i}.$$

In case (ii) weakly stationary process AR(1) exists in view of Theorem 4.2.4 but it is not causal. We find its representation as a linear process. To this end note

$$X_t = \frac{X_{t+1}}{\varphi} - \frac{\varepsilon_{t+1}}{\varphi} = \cdots = \frac{X_{t+k}}{\varphi^k} - \sum_{j=1}^{k-1} \frac{\varepsilon_{t+j}}{\varphi^j}.$$

As (X_t) is stationary, the first term on the right hand side in the above equality tends to 0 in $\mathcal{L}^2$ and we obtain the representation

$$X_t = -\sum_{j=1}^{\infty} \frac{\varepsilon_{t+j}}{\varphi^j}.$$

We calculate covariance function of this process. For $h > 0$ we have

$$\gamma(h) = \mathrm{Cov}\Big(-\sum_{j=1}^{\infty} \frac{\varepsilon_{t+j}}{\varphi^j}, -\sum_{j=1}^{\infty} \frac{\varepsilon_{t+h+j}}{\varphi^j}\Big) = \sigma^2 \sum_{j=1}^{\infty} \frac{1}{\varphi^{j+h}} \frac{1}{\varphi^j} = \frac{\sigma^2}{\varphi^2 - 1}\varphi^{-h} \quad (4.12)$$

and thus $\rho(h) = \varphi^{-|h|}$. Note that since correlation structure is exactly the same as for AR(1) process with autoregressive coefficient φ^{-1}, the Yule-Walker equations yield that prediction of X_{t+1} based on $X_1, \ldots, X_t$ equals $\hat{X}_{t+1} = \varphi^{-1}X_t = -\sum_{i=1}^{\infty} \varphi^{-i-1}\varepsilon_{t+i}$. We check that indeed $\varphi^{-1}X_t$ is perpendicular to $X_{t+1} - \hat{X}_{t+1} = \varphi^{-2}\varepsilon_{t+1} + \sum_{i=2}^{\infty}(\varphi^{-(i+1)} - \varphi^{-(i-1)})\varepsilon_{t+i}$

$$\mathrm{Cov}(X_{t+1} - X_{t+1}, \hat{X}_{t+1}) = -\frac{1}{\varphi^4} - \sum_{i=2}^{\infty} \Big(\frac{1}{\varphi^{2i+2}} - \frac{1}{\varphi^{2i}}\Big) = 0. \quad (4.13)$$

Example 4.2.6 *We consider now ARMA(1,1) with mean μ*

$$X_t - \mu = \varphi(X_{t-1} - \mu) + \varepsilon_t + \theta\varepsilon_{t-1}, \quad (4.14)$$

where $|\varphi| < 1$. We know that such process is causal with respect to white noise ε_t and we compute now its causal representation. Structural equation may be written as $(1-\varphi B)(X_t - \mu) = (1+\theta B)\varepsilon_t$. Whence

$$\begin{aligned} X_t - \mu &= \frac{(1+\theta B)}{(1-\varphi B)}\varepsilon_t = (1+\theta B)\sum_{i=0}^{\infty} \varphi^i B^i \varepsilon_t = \sum_{i=0}^{\infty} \varphi^i B^i \varepsilon_t + \theta \sum_{i=0}^{\infty} \varphi^i B^{i+1}\varepsilon_t \\ &= \varepsilon_t + \varphi \sum_{i=1}^{\infty} \varphi^{i-1} B^i \varepsilon_t + \theta \sum_{i=1}^{\infty} \varphi^{i-1} B^i \varepsilon_t = \varepsilon_t + (\varphi+\theta)\sum_{i=1}^{\infty} \varphi^{i-1} B^i \varepsilon_t. \end{aligned}$$

Thus in this case

$$\psi(z) = 1 + (\varphi + \theta)\sum_{i=1}^{\infty} \varphi^{i-1} z^i. \tag{4.15}$$

Using causal representation we will compute covariance function of (X_t)*. Causality implies that* $E((X_{t-j}-\mu)\varepsilon_{t-\tau}) = 0$ *for* $j > \tau$*. Thus for* $h \geqslant 2$ *terms* $\varepsilon_t + \theta\varepsilon_{t-1}$ *appearing in structural equation (4.14) of* $X_t - \mu$ *are uncorrelated with* X_{t-h} *and we have*

$$\gamma(h) = E(X_t - \mu)(X_{t-h} - \mu) = \varphi\gamma(h-1)$$

and thus it follows that $\gamma(h) = \varphi^{h-1}\gamma(1)$.
We compute now $\gamma(0)$ *and* $\gamma(1)$.

$$\begin{aligned}\gamma(0) = E(X_t - \mu)^2 &= E(X_t - \mu)(\varphi(X_{t-1} - \mu) + \varepsilon_t + \theta\varepsilon_{t-1}) \\ &= \varphi\gamma(1) + E(\varepsilon_t + (\varphi + \theta)\sum_{i=1}^{\infty} \varphi^{i-1} B^i \varepsilon_t)(\varepsilon_t + \theta\varepsilon_{t-1}) \\ &= \varphi\gamma(1) + \sigma^2 + (\varphi + \theta)\theta\sigma^2.\end{aligned}$$

Analogously

$$\gamma(1) = E(\varphi(X_{t-1} - \mu) + \varepsilon_t + \theta\varepsilon_{t-1})(X_{t-1} - \mu) = \varphi\gamma(0) + \theta\sigma^2,$$

where in the last equality we used $X_{t-1} - \mu = \varphi(X_{t-2} - \mu) + \varepsilon_{t-1} + \theta\varepsilon_{t-2}$*. Solving two last equations we have*

$$\gamma(0) = \frac{1 + 2\varphi\theta + \theta^2}{1 - \varphi^2}\sigma^2 \qquad \gamma(1) = \frac{(1 + \varphi\theta)(\varphi + \theta)}{1 - \varphi^2}\sigma^2.$$

Note that $\gamma(h)$ *can be directly calculated using (4.15) and derivations in Example 1.3.3(ii). Namely, for* $h > 0$

$$\begin{aligned}\gamma(h) = \sigma^2 \sum_{i=0}^{\infty} \psi_i \psi_{i+h} &= \sigma^2((\varphi + \theta)\varphi^{h-1} + \sum_{i=1}^{\infty} (\varphi + \theta)^2 \varphi^{2(i-1)+h}) \\ &= \sigma^2 \varphi^{h-1}(\varphi + \theta + \varphi \sum_{i=0}^{\infty} (\varphi + \theta)^2 \varphi^{2j}) \\ &= \sigma^2 \phi^{h-1}(\varphi + \theta + \frac{\varphi(\varphi + \theta)^2}{1 - \varphi^2})\end{aligned} \tag{4.16}$$

which coincides with previous calculations.
Covariance function of process ARMA(1,1) for $h \geqslant 1$ *has the same power law decay as AR(1) process but in contrast to it* $\gamma(1) \neq \varphi\gamma(0)$*, if* $\theta \neq 0$.

4.2.1 Covariance function for a causal ARMA(p, q) time series

We show that, as the previous example suggests, causality of ARMA(p, q) time series is an useful tool for calculating its covariance function. Namely, suppose that process (X_t) with structural equation $\varphi(B)X_t = \theta(B)Z_t$ satisfies assumptions of Theorem 4.2.2 and has causal representation $X_t = \sum_{i=0}^{\infty} \psi_j Z_{t-j}$. We calculate scalar products of both sides of (4.1) with

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

and using the above representation of X_{t-k} we have

$$\gamma(k) - \varphi_1\gamma(k-1) - \cdots - \varphi_p\gamma(k-p) = \sigma^2 \sum_{k \leqslant j \leqslant q} \theta_j \psi_{j-k}$$

for $0 \leqslant k \leqslant q$ and

$$\gamma(k) - \varphi_1\gamma(k-1) - \cdots - \varphi_p\gamma(k-p) = 0$$

for $k > q$.
It is known that for $h \geqslant \max(p, q+1) - p$ solution to the equations above has the form

$$\gamma(h) = \sum_{i=1}^{l} \sum_{j=0}^{r_i - 1} \beta_{ij} h^j \xi_i^{-h},$$

where ξ_i for $i = 1, \ldots, l$ are all different roots of $\varphi(z) = 0$, r_i their respective multiplicities and β_{ij} certain constants. As $|\xi_i| > 1$ (causality!), autocovariance function tends to 0 with $h \to \infty$, and its rate of decay is determined by a root (or roots) with the smallest distance to the unit circle. We note that from the last equation it follows that the autocovariance function of causal ARMA(p, q) process satisfies $|\gamma(h)| \leqslant a^{-h}$ for a certain $a > 1$ when $h \to \infty$.

For ARMA(p, q) process with only few non-zero coefficients of $\varphi(\cdot)$ i $\theta(\cdot)$ covariance function can be often calculated explicitely as in the case of ARMA(1,1). We consider one more example

Example 4.2.7 *Let (X_t) be ARMA(12,1) process defined as $X_t = \varphi X_{t-12} + Z_t + \theta Z_{t-1}$, $t \in \mathbb{Z}$, where $|\varphi| < 1$. From the results of this section it follows that X_t is causal with respect to Z_t. In order to compute its covariance function note that*

$$\begin{aligned}\gamma(0) &= < X_t, X_t > = < X_t, \varphi X_{t-12} + Z_t + \theta Z_{t-1} > \\ &= \varphi\gamma(12) + < \varphi X_{t-12} + Z_t + \theta Z_{t-1}, Z_t + \theta Z_{t-1} > = \varphi\gamma(12) + \sigma^2 + \theta^2\sigma^2,\end{aligned}$$

as from the causality of X_t it follows that $< X_{t-12}, Z_{t-i} > = 0$ for $i \leqslant 11$. In an analogous way we check that $\gamma(12) = \varphi\gamma(0)$. Thus

$$\gamma(0) = \frac{1+\theta^2}{1-\varphi^2}\sigma^2.$$

From the equality $\gamma(h) =< \varphi X_{t-12} + Z_t + \theta Z_{t-1}, X_{t-h} >$ *we have that*

$$\gamma(1) = \varphi\gamma(11) + \theta\sigma^2$$

and for $10 \geqslant h \geqslant 2$, $\gamma(h) = \varphi\gamma(12-h) = \varphi^2\gamma(h) = 0$. *Moreover,* $\gamma(12h+i) = \varphi\gamma(12(h-1)+i)$ *for* $h > 0$ *and* $|i| < 12$. *Consequently*

$$\gamma(12h) = \varphi^{|h|}\frac{1+\theta^2}{1-\varphi^2}\sigma^2,$$

$$\rho(12h \pm 1) = \varphi^{|h|}\frac{\theta}{1-\varphi^2}\sigma^2.$$

For the remaining values of lag h $\gamma(h)$ *equals 0.*
Note that for the process $X_t = \varphi X_{t-12} + \varepsilon_t$ *with* $|\varphi| < 1$ *correlation function equals* $\rho(12h) = \varphi^{|h|}$ *for lags* $12h$ *and 0 otherwise. In the above example, due to occurrence of MA(1) term in the structural equation, nonzero correlations for lags* $12h \pm 1$ *appear.*

The following two results provide important characterizations of AR(p) i MA(q) processes, respectively.

Theorem 4.2.8 *(i) If* (X_t) *is causal AR*(p) *time series then* $\alpha(n) = 0$ *for* $n > p$. *(ii) If* $\alpha(n) = 0$ *for* $n > p$ *and* (X_t) *is nondeterministic, that is* $||X_t - \hat{X}_t|| > 0$, *then* (X_t) *is AR*(p).

Proof. As (X_t) is causal, $\varepsilon_t \perp \sum_{i=1}^{p}\varphi_i X_{t-i}$ and thus it follows that for $n \geqslant p$ optimal linear predictor $\hat{X}_{t,n} = \sum_{i=1}^{p}\varphi_i X_{t-i}$ and whence for $n > p$ $\alpha(n) = \varphi_{nn} = 0$. Conversely, $\hat{X}_{t,n} \to \hat{X}_t$ (cf. Problem 5.2(ii)) and thus it follows that for $n \geqslant p$ $\hat{X}_{t,n} = \hat{X}_t$ and as $\varepsilon_t = X_t - \hat{X}_t$ is a nondegenerate white noise (we will check it in the proof of Wold's theorem next section) the conclusion follows.

Theorem 4.2.9 *Assume that* (X_t) *is nondeterministic. Then* $\gamma(h) = 0$ *for* $|h| > q$ *is equivalent to* (X_t) *is* MA(q) *for a certain white noise* (Z_t) *with nonzero variance.*

Proof. Let $Z_t := X_t - P_{H_{t-1}}X_t$. Again, we use the fact that (Z_t) is WN$(0,\sigma^2)$, where $\sigma^2 =\| X_t - H_{t-1}X_t \|^2$ is positive due to assumptions. As summands in the decomposition $X_t = P_{H_{t-1}}X_t + Z_t$ are uncorrelated, we have for any $q \in N$ that

$$H_t(X) = H_{t-1}(X) \oplus sp(Z_t) = \cdots = H_{t-q}(X) \oplus sp(Z_t, Z_{t-1}, \ldots, Z_{t-q+1})$$

and

$$H_{t-1}(X) = H_{t-q-1}(X) \oplus sp(Z_{t-1}, Z_{t-1}, \ldots, Z_{t-q}).$$

Then

$$X_t = Z_t + P_{H_{t-1}} X_t = Z_t + P_{H_{t-q-1}} X_t + P_{sp(Z_{t-1}, Z_{t-1}, \ldots, Z_{t-q})} X_t.$$

As X_t is uncorrelated with X_{t-i} for $i > q$, then it follows from the Yule-Walker equations that $P_{H_{t-1-q}} X_t = 0$ and the conclusion follows from $P_{sp(Z_{t-1}, Z_{t-1}, \ldots, Z_{t-q})} X_t = \sum_{i=1}^{q} \theta_i Z_{t-i}$ for certain $\theta_1, \ldots, \theta_q$.
From the last result we have the following conclusion.

Corollary 4.2.10 *X_t i Y_t are two mean zero weakly stationary time series with the same covariance function. If X_t is ARMA(p, q) , then Y_t is ARMA(p, q) (not necessarily with the same coefficients).*

Proof. Let $\varphi_1, \ldots, \varphi_p$ be autoregressive coefficients of the process X_t and consider the following process

$$\tilde{Y}_t := Y_t - \sum_{i=1}^{p} \varphi_i Y_{t-i}.$$

Then as $\tilde{X}_t := X_t - \sum_{i=1}^{p} \varphi_i X_{t-i} = \varepsilon_t + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j}$, covariance function of $\tilde{X}_t$ vanishes for lags $h > q$ and in view of assumptions covariance function of $\tilde{Y}_t$ has the same property. But then from the previous result it follows that $\tilde{Y}_t$ is MA(q) and the conclusion follows.

4.2.2 Prediction for causal ARMA(p, q) time series

We discuss now a useful fact that for *causal* ARMA(p, q) prediction of X_{n+1} for $n \geqslant \max(p, q)$ requires only q last innovations and p preceding values of the process instead of all n innovations. As a by-product we establish an important property which we will later use that predictor $\widehat{X}_{n+1}$ does not depend on the variance of errors σ^2.

Theorem 4.2.11 *Let $(X_t)_{t \in \mathbb{Z}}$ be causal process ARMA(p, q), $m = \max(p, q)$, $\widehat{X}_{n+1} = P_{\{X_1, \ldots, X_n\}} X_{n+1}$. Then*
(i)

$$\left.\begin{array}{ll} \widehat{X}_{n+1} = \sum_{i=1}^{n} \theta_{ni}(X_{n+1-i} - \widehat{X}_{n+1-i}), & \text{for } 1 \leqslant n < m \\ \quad \text{(the usual innovation representation)} & \\ \widehat{X}_{n+1} = \varphi_1 X_n + \cdots + \varphi_p X_{n+1-p} + & \\ \quad + \sum_{j=1}^{q} \theta_{nj}(X_{n+1-j} - \widehat{X}_{n+1-j}), & \text{for } n \geqslant m \end{array}\right\} \tag{4.17}$$

(ii) $\widehat{X}_{n+1}$ and $\| X_{n+1} - \widehat{X}_{n+1} \|^2 / \sigma^2$ do not depend on σ^2.

Proof. We consider an auxiliary process W_t defined for $1 \leqslant t \leqslant m$ as $W_t = \sigma^{-1} X_t$, and for $t > m$ as

$$W_t = \sigma^{-1}(X_t - \varphi_1 X_{t-1} - \cdots - \varphi_p X_{t-p}).$$

Note that (W_t) does not depend on σ^2. This follows from the observation that for $t > m$ we have

$$W_t = \frac{Z_t}{\sigma} + \theta_1 \frac{Z_{t-1}}{\sigma} + \cdots + \theta_q \frac{Z_{t-q}}{\sigma},$$

where $(Z_i/\sigma)_{t\in\mathbb{Z}}$ does not depend on σ whereas for $t \leqslant m$ in view of causality we have $X_t/\sigma = \sum_{i=0}^{\infty} \psi_i Z_{t-i}/\sigma$, where ψ_i, obtained as coefficients in the expansion of $\theta(z)/\varphi(z)$, also do not depend on σ.
We let $\theta_j = 0$ for $j > q$. W_t is not weakly stationary but its autocovariance function is easily computable. We have

$$\gamma(i,j) = \begin{cases} \sigma^{-2}\gamma_X(i-j) & \text{for } 1 \leqslant i,j \leqslant m \\ \sigma^{-2}\big(\ \gamma_X(i-j) \ - \sum\limits_{k=1}^{p} \varphi_r \gamma_X(k-|i-j|)\ \big) & \\ & \text{for } \min(i,j) \leqslant m < \max(i,j) \leqslant 2m \\ \sum\limits_{r=0}^{q} \theta_r \theta_{r+|i-j|} & \text{for } \min(i,j) > m \\ 0 & \text{otherwise} \end{cases}$$

Note that $m \geqslant p$ is used to calculate autocovariance under the second condition and $m \geqslant q$ is used in the case of the fourth condition.
Innovation representation of $\{W_t\}$ is

$$\widehat{W}_{n+1} = \sum_{j=1}^{n} \theta_{nj}(W_{n+1-j} - \widehat{W}_{n+1-j}), \quad 1 \leqslant n < m$$

$$\widehat{W}_{n+1} = \sum_{j=1}^{q} \theta_{nj}(W_{n+1-j} - \widehat{W}_{n+1-j}), \quad n \geqslant m.$$

For $n \geqslant m$ we have only q summands in the last equation as we observed while discussing innovation representation that $\gamma(n,j) = 0$ for $n \geqslant m$ and j: $|n-j| > q$ yields $\theta_{nj} = 0$.
Moreover, we note that $H_n = sp\{X_1,\ldots,X_n\} = sp\{W_1,\ldots,W_n\}$. This is obvious for $n \leqslant m$ and is easily obtained inductively for larger n. Then the projection of W_{n+1} on $sp\{W_1,\ldots,W_n\}$ coincides with its projection on $sp\{X_1,\ldots,X_n\}$. Projecting both sides on the last subspace in view of the definition of W_t we get
$\widehat{W}_t = \sigma^{-1}\widehat{X}_t, \qquad t = 1,2,\ldots,m$
$\widehat{W}_t = \sigma^{-1}[\widehat{X}_t - \varphi_1 X_{t-1} - \cdots - \varphi_p X_{t-p}], \qquad t > m$
and it follows that

$$X_{n+1} - \widehat{X}_{n+1} = \sigma(W_{n+1} - \widehat{W}_{n+1}), \quad n \geqslant 0.$$

This establishes (i). (ii) follows directly from (4.17), as θ_{ni} do not depend on σ^2 as the covariance function of (W_t) does not depend on it.

4.3 Problems

1. Let (X_t) be ARMA(12,12) time series of the form $X_t = \varphi X_{t-12} + Z_t + \theta Z_{t-12}$ for $|\varphi|, |\theta| < 1$. Compute its covariance function.
2. X_t is ARMA(2,1) time series given by $X_t - 0.06X_{t-1} + 0.09X_{t-2} = Z_t + Z_{t-1}$. Find:
(i) its causal representation;
(ii) linear prediction of X_2 based on $H_0(Z)$.
3. Consider ARMA(1,1) time series with mean μ

$$X_t - \mu = a(X_{t-1} - \mu) + Z_t + bZ_{t-1}$$

with $|b| < 1$. Find invertible representation of this process.
4. Check that for causal AR(2) time series we have for $h \geqslant 1$ $\gamma(h) = \varphi_1\gamma(h-1) + \varphi_2\gamma(h-2)$.
5. Show that invertibility is a stronger property than $Z_t \in H_t(X)$ i.e. it may happen that $Z_t = \lim_{n\to\infty} \sum_{j=0}^{n} a_{jn} X_{t-j}$, however, it is not true that $Z_t = \sum_{j=0}^{\infty} a_j X_{t-j}$ for some square integrable sequence (a_j). Hint. Consider non-invertible MA(2) process $X_t = Z_t - Z_{t-1}$.
6. Prove equality (4.11).
7. Let $(X_t)_{t\in\mathbb{Z}}$ be a zero-mean AR(2) such that (ε_t is$WN(0, \sigma^2)$)

$$X_t - 0,8X_{t-1} + 0,15X_{t-2} = \varepsilon_t.$$

(i) Check that (X_t) is causal.
(ii) Using representation of autoregressive polynomial $\varphi(z) = \prod_{i=1}^{p}(1 - a_i^{-1}z)$, where a_i are roots of $\varphi(z)$ find the causal representation of (X_t). What is the prediction error of one step prediction of X_t based on $\mathcal{H}_{t-1}$? Justify your answer.
8. Prove Theorem 4.2.3.

5

Representation of nondeterministic processes: the Wold theorem

The Wold theorem states important representation of weakly stationary deterministic processes (5.3) which is theoretical underpinning of using one sided moving average as a modelling tool. Unfortunately, for a given WS time series (X_t) it is usually hard to construct pertaining innovations Z_t which are building blocks of this representation as are based on projections of X_t on its past.

5.1 Deterministic and nondeterministic processes

Throughout this chapter $(X_t)_{t\in\mathbb{Z}}$ will denote a weakly stationary process. Recall that $H_t(X) := \overline{sp}(X_s, s \leqslant t)$. We consider projection of X_{t+1} on $H_t = H_t(X)$

$$\widehat{X}_{t+1} = P_{H_t} X_{t+1}.$$

Definition 12 *We call $(X_t)_{t\in\mathbb{Z}}$ deterministic process if*

$$X_{t+1} = \widehat{X}_{t+1} \quad \text{for any } t \in \mathbb{Z}. \tag{5.1}$$

Obviously, the above equality should hold with probability 1, or equivalently $H_{t+1}(X) \subseteq H_t(X)$.
It follows from weak stationarity that if the condition above is satisfied for some t then it is satisfied for all t. It is almost trivial to note that in view of weak stationarity (5.1) is equivalent to any of the two conditions below:
(i) For any t we have $||X_{t+1} - P_{H_t} X_{t+1}||^2 = E|X_{t+1} - P_{H_t} X_{t+1}|^2 = 0$
or
(ii) $X_{t+1} \in H_{-\infty}(X) = \bigcap_t H_t(X)$.
When weakly stationary process does not satisfy (5.1) we call it nondeterministic. Note that if the series (X_t) is nondeterministic we have that $\mathbf{\Gamma}_n > 0$ for $n \in \mathbb{N}$ or, putting it differently, for any $t_1, \ldots, t_n$ $X_{t_1}, \ldots, X_{t_n}$ are linearly independent. Deterministic process are not very interesting objects as far as randomness is concerned since they describe series for which a future value is wholly represented by its past. However, it turns out that this type of determinism is sufficient to describe deterministic part of the process, see (5.3).

Example 5.1.1 *If we consider $X_t = A\cos\omega t + B\sin\omega t$, where A, B are uncorrelated mean zero random variables with variance σ^2, $\omega \in \mathbb{R}$ is fixed frequency,*

then the randomness of X_t is due only to randomness of random variables A and B which do not depend on t.
It is easy to check that $X_{t+1} = 2\cos\omega\ X_t - X_{t-1}$ for any $t \in \mathbb{Z}$ thus (X_t) is deterministic since $X_{t+1} \in sp(X_{t-1}, X_t) \subseteq H_t(X)$.

We first give a less obvious characterization of non-deterministic processes in terms od their partial autocorrelation. Observe first that equality (3.20) and remark 3.2.1 below the proof of the Durbin-Levinson algorithm imply that

$$\sigma^2 = \gamma(0)\prod_{i=1}^{\infty}(1-\varphi_{ii}^2) = \gamma(0)\prod_{i=1}^{\infty}(1-\alpha^2(i)). \tag{5.2}$$

Another useful expression for the innovation variance σ^2 in terms of the logarithm of the spectral density is given by Kolmogorov-Szegő formula (cf Theorem 6.20). It follows from (5.2) that

Theorem 5.1.2 *Let (X_t) be a weakly stationary process. Then it is nondeterministic if and only if*

$$\sum_{i=1}^{\infty}\alpha^2(i) < \infty.$$

Proof. We consider two cases. Firstly, we can have that $\alpha^2(i) = 1$ for $i \geqslant i_0$. Then process is deterministic (cf proof of (3.29)) and at the same time $\sum_{i=1}^{\infty}\alpha^2(i)$ is infinite. In the second case, we have that $\alpha^2(i) < 1$ for all i. Then due to (5.2) the process is nondeterministic if and only if $\prod_{i=1}^{\infty}(1-\alpha^2(i))$ is nonzero. But this is equivalent to $(\alpha(i))_{i\in\mathbb{N}} \in \ell^2$ as $\alpha^2(i) < 1$ for all i .

5.2 The Wold theorem

Let $\sigma^2 = E|X_{t+1} - P_{H_t}X_{t+1}|^2$ which is strictly positive for a non-deterministic process. Note that $E|X_{t+1} - P_{H_t}X_{t+1}|^2$ does not depend on t as for any element $W_t = \sum_{i_k\geqslant 1} a_{i_k}X_{t+1-i_k}$ approximating X_{t+1} we have $E|X_{t+1} - W_t|^2 = E|X_{s+1} - W_s|^2$, where $W_s = \sum_{i_k\geqslant 1} a_{i_k}X_{s+1-i_k}$.

Theorem 5.2.1 *(Wold). Let $(X_t)_{t\in\mathbb{Z}}$ be a weakly stationary nondeterministic time series, $EX_t = 0$. Then X_t has representation*

$$X_t = \sum_{j=0}^{\infty}\psi_j Z_{t-j} + V_t, \tag{5.3}$$

where
(i) (Z_t) *is* $WN(0,\sigma^2)$;
(ii) $\sum_{j=0}^{\infty}\psi_j^2 < \infty, \quad \psi_0 = 1$;
(iii) $Z_t \in H_t(X)$;

(iv) $E(Z_t V_s) = 0$ *for all* $t, s \in \mathbb{Z}$;
(v) $V_t \in H_{-\infty}(X)$;
(vi) V_t *is deterministic* : $V_t \in H_{-\infty}(V)$;
(vii) Representation (5.3) is unique when (Z_t) *and* (V_t) *satisfy conditions (i)-(vi).*

Note that (5.3) is an representation of X_t as the sum of two terms: *one-sided* linear process and a deterministic one. Variables (Z_j) are innovations of the linear part.
Proof of (i) – (vi).
(i) Consider the decomposition $X_t = P_{H_{t-1}} X_t + X_t - P_{H_{t-1}} X_t$ and define Z_t as a term in this decomposition perpendicular to $H_{t-1}(X)$, namely

$$Z_t := X_t - P_{H_{t-1}} X_t. \tag{5.4}$$

Moreover, let $\sigma^2 = ||Z_t||^2$ and

$$\psi_j := \frac{< X_t, Z_{t-j} >}{\parallel Z_{t-j} \parallel^2} = \frac{< X_t, Z_{t-j} >}{\sigma^2},$$

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

It will follow from (i) and (ii) that V_t is well defined. Note that (5.4) implies that $Z_t \in H_t$ and $Z_t \perp H_{t-1}$ and thus $Z_t \in H_{t-1}^{\perp} \subseteq H_{t-2}^{\perp} \subseteq \dots$. In particular $Z_t \in H_s^{\perp}$ for $s < t$ and thus $E(Z_t Z_s) = 0$. We omit the proof that $EZ_t = 0$ and that weak stationarity implies that $\sigma^2 = ||X_{t+1} - P_{H_t} X_{t+1}||^2$ does not depend on t. These properties together mean that Z_t is weakly stationary white noise $\mathrm{WN}(0, \sigma^2)$.
(ii) Consider the projection of X_t on $\overline{sp}(Z_s, s \leqslant t)$ and call it $\widehat{X}_t$. As $(Z_s)_{s \leqslant t}$ form an orthogonal base in $\overline{sp}(Z_s, s \leqslant t)$, we have

$$\widehat{X}_t = \sum_{j=0}^{\infty} \frac{< X_t, Z_{t-j} >}{\parallel Z_{t-j} \parallel^2} Z_{t-j} = \sum_{j=0}^{\infty} \psi_j Z_{t-j},$$

where the form of the coefficients easily follows from the fact that (Z_i) are orthogonal. The fact that $\sum \psi_j^2 < \infty$ follows now from properties of projection and orthogonality of (Z_t):

$$\infty > \parallel X_t \parallel^2 \geqslant \parallel \widehat{X}_t \parallel^2 = \sigma^2 \sum_{j=0}^{\infty} \psi_j^2,$$

Moreover, note that

$$\psi_0 = < X_t, Z_t > / \sigma^2 = < X_t, X_t - P_{H_{t-1}} X_t > / \sigma^2 = \parallel X_t - P_{H_{t-1}} X_t \parallel^2 / \sigma^2 = 1.$$

Proof of (iii) follows from the definition of Z_t. In order to prove (iv) note that definition of V_t implies that $V_t \perp Z_s, \quad s \leqslant t$ as $\widehat{X}_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}$ is the projection of X_t on $\overline{sp}(Z_s, s \leqslant t)$. On the other hand for $s > t$ we have $Z_s \in H_{s-1}^{\perp} \subseteq H_t^{\perp}$ and $V_t \in H_t$, whence $Z_s \perp V_t$.
Proof of (v). We have that $V_t \in H_t = H_{t-1} \oplus sp(Z_t)$ but $V_t \perp Z_t$ thus $V_t \in H_{t-1} = H_{t-2} \oplus sp(Z_{t-1})$. Whence $V_t \in H_{t-2}$ and proceeding in an analogous way we obtain that $V_t \in H_{t-j}$ for any $j \in \mathbb{N}$ which is equivalent to

$$V_t \in \bigcap_{j=0}^{\infty} H_{t-j} = H_{-\infty}(X).$$

Proof of (vi). It follows from the definition of V_t and Z_t that

$$H_t(X) = \overline{sp}\{Z_j, j \leqslant t\} \oplus \overline{sp}\{V_j, j \leqslant t\} \tag{5.5}$$

We consider any $Y \in H_{-\infty}(X)$ and we show that $Y \in \oplus sp(V_j,\ j \leqslant t)$. As we have that $Y \in H_s \cap H_{s-1}$ but $H_s = H_{s-1} \oplus sp(Z_s)$ it follows that $Y \perp Z_s$ and in view of (5.5) it means that $Y \in \oplus \overline{sp}(V_j,\ j \leqslant t)$. This together with (5.5) yields $\overline{sp}(V_j,\ j \leqslant t) = H_{-\infty}(X)$. As this holds for any t it implies $= H_{-\infty}(V) = \overline{sp}(V_j,\ j \leqslant t)$.
We prove now the uniqueness of representation (5.3). Let (Z_t) i (V_t) be any processes satisfying (5.3) and conditions (i)-(vi). In this case representation (5.3) entails $H_{t-1} \subseteq \overline{sp}\{Z_j, j \leqslant t-1\} \oplus \overline{sp}\{V_j, j \leqslant t-1\}$, and thus conditions (i) and (iv) imply that Z_t is orthogonal to H_{t-1}, as Z_t is orthogonal to Z_j and to V_j for $j \leqslant t-1$. Projecting both sides of (5.3) on $H_{t-1}(X)$ we thus obtain $P_{H_{t-1}} X_t = \sum_{i=1}^{\infty} \psi_i Z_{t-i} + V_t$ (as Z_i for $i \leqslant t-1$ and V_t belong to $H_{t-1}(X)$), whence we have $Z_t = X_t - P_{H_{t-1}} X_t$ i.e. (5.4). Form of ψ_i follows from computing scalar product of both sides of (5.3) with Z_j and using (i). Finally, if (5.3) holds than V_t has to be defined as in the proof of the theorem.
Note that it follows from the proof how the projections of X_t on $H_{t-1}(X)$ and on $H_{t-1}(Z)$ are related. Namely we have

$$P_{H_{t-1}(X)} X_t = \sum_{i=1}^{\infty} \psi_i Z_{t-i} + V_t = P_{H_{t-1}(Z)} X_t + V_t.$$

Remark 5.2.2 *(i) Alternative proof of the Wold representation is obtained by proving that for nondeterministic process (X_t) it follows from $H_t(X) = H_s(X) \oplus sp(Z_{s+1}, \ldots, Z_t)$ for $s < t$ that (cf. problem 5.1)*

$$H_t(X) = H_t(Z) \oplus H_{-\infty}(X). \tag{5.6}$$

Indeed, then (5.6) implies

$$X_t = P_{H_t(X)}(X_t) = P_{H_t(Z)}(X_t) + P_{H_{-\infty}(X)}(X_t) =: U_t + V_t.$$

As $(Z_s)_{s \leqslant t}$ is an orthogonal base in $H_t(Z)$, we have $U_t = \sum_{j=0}^{\infty} \psi_i Z_{t-i}$ for a certain sequence $(\psi_i) \in \ell^2$.

(ii) Note that the proof of the Wold theorem yields also Wold decomposition of the finite order, namely for any $n \in \mathbb{N}$ we have

$$X_t = \sum_{k=0}^{n} \psi_k Z_{t-k} + e_{t,n},$$

where $e_{t,n} \in H_{t-n-1}$.

Remark 5.2.3 *Consider shortly the case when (X_t) is stationary, zero mean, nondeterministic and Gaussian. Then it follows that the projection on the whole past $\hat{X}_t$ is Gaussian as the variances σ_n^2 of (Gaussian) approximands of $\hat{X}_t$ in $\mathcal{H}_{t-1}$ converge to a certain σ^2 and thus $\hat{X}_t$ has $N(0, \sigma^2)$ distribution. Moreover, $(X_t, \hat{X}_t)$ is jointly Gaussian. Then it also follows from the proof of the Wold theorem that (Z_t) and (V_t) are Gaussian processes. If (X_t) is q-dependent for some $q \in \mathbb{N}$ we have $V_t \equiv 0$ and the formula for ψ_i implies that $\psi_i = 0$ for $i > q$, thus (X_t) is MA(q) series. On the other hand it is easy to check that if X_t and $X_{t-k}, k > p$ are independent given $X_{t-1}, \ldots, X_{t-p}$ then (X_t) is AR(p) causal series. namely this implies by conditioning that*

$$E([X_t - E(X_t|X_{t-1}, \ldots, X_{t-p})]X_{t-k}) = 0$$

for $k > 0$ and the conclusion follows from gaussianity and Problem 1.9.

We define now another property related to nondeterminism.

Definition 13 *Zero mean time series ws $(X_t)_{t\in\mathbb{Z}}$ is purely non-deterministic (PND) or linearly regular if $H_{-\infty} = \{0\}$. For an arbitrary mean WS time series (X_t) is PND if $X_t - EX_t$ is PND.*

It follows that, as the name suggests, that pure nondeterminism is a stronger property than nondeterminism, that is any time series which is purely nondeterministic is nondeterministic. Indeed, if $(X_t)_{t\in\mathbb{Z}}$ is deterministic this entails $H_t(X) \subseteq H_{t-1}(X)$ for any t and thus $H(X) \subseteq H_{-\infty}(X)$. In particular $H_{-\infty}(X) \neq \{0\}$. We have

Proposition 5.2.4 *White noise $(\varepsilon_t)_{t\in\mathbb{Z}}$ is PND.*

Proof. We want to prove that $Y \in H_{-\infty}(\varepsilon)$ implies that $Y = 0$ almost everywhere. We have that $Y \in H_{t-1}(\varepsilon) \subset H_t(\varepsilon)$, but for any $t \in \mathbb{Z}$ $\varepsilon_t \perp H_{t-1}(\varepsilon)$ and thus $< Y, \varepsilon_t >= 0$. However, as $Y \in H(\varepsilon)$ and $(\varepsilon_t)_{t\in\mathbb{Z}}$ is a maximal orthonormal subset in $H(\varepsilon)$ we have that

$$Y = \sum_{s=-\infty}^{\infty} c_s \varepsilon_s, \quad (c_s) \in \ell^2$$

But $< Y, \varepsilon_s >= c_s\sigma^2 = 0$ implies $c_s = 0$ and thus $Y = 0$ almost everywhere. Note that $(\varepsilon_t)_{t\in\mathbb{Z}}$ is maximal orthonormal subset in $H(\varepsilon)$ as for any putative $a \in H(\varepsilon)$ such that a is perpendicular to any ε_t it would imply that a is perpendicular to $\overline{sp(\varepsilon_t)} = H(\varepsilon)$, a contradiction.

In view of the Wold theorem any purely nondeterministic process satisfies

$$X_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}, \qquad \sum |\psi_j|^2 < \infty, \tag{5.7}$$

In other words, any PND process is causal (or has MA(∞) representation) with respect to a certain white noise, where (ε_t) is WN$(0, \sigma^2)$. Conversely, if $V_t \equiv 0$ in the Wold representation then it follows from Proposition 5.2.4 that (X_t) is PND. Indeed, in this case we have $H_{-\infty}(X) \subseteq H_{-\infty}(\varepsilon) = \{0\}$. Thus we obtain

Proposition 5.2.5 *Weakly stationary process $(X_t)_{t\in\mathbb{Z}}$ is causal with respect to a certain white noise if and only if it is PND.*

Representation (5.7) is also frequently called MA(∞) representation of PND process (X_t) and coefficients $(\psi_j)_{j=0}^{\infty}$ its MA parameters.

5.3 Prediction based on infinite past

Corollary 5.3.1 *Consider Wold decomposition (5.3) of weakly stationary nondeterministic process $(X_t)_{t\in\mathbb{Z}}$ and let $\widehat{X}_l$ be l-step prediction of X_l based on $X_0, X_{-1}, X_{-2}, \ldots$. Then*

$$\widehat{X}_l = \sum_{j=l}^{\infty} \psi_j Z_{l-j} + V_l.$$

and

$$\mathrm{Var}(X_l - \widehat{X}_l) = \sigma^2 \sum_{j=0}^{l-1} \psi_j^2$$

Proof. Rewriting Wold decomposition of X_l we have

$$X_l = \sum_{j=0}^{l-1} \psi_j Z_{l-j} + \sum_{j=l}^{\infty} \psi_j Z_{l-j} + V_l$$

From the construction of Z_t it follows that the first term on the right hand side is perpendicular to $H_0(X)$ whereas

$$\sum_{j=l}^{\infty} \psi_j Z_{l-j} \ \in H_0(X) \ \text{ and } \ V_l \in H_{-\infty}(X) \subseteq H_0(X),$$

and thus $\sum_{j=l}^{\infty} \psi_j Z_{l-j} + V_l$ is a projection of X_l on $H_0(X)$.

Remark 5.3.2 *(i) From Corollary 5.3.1 it follows that if (X_t) is PND then*

$$P_{H_{t-n}} X_t \xrightarrow{\mathcal{L}^2} 0, \quad \textit{when} \ \ n \to \infty,$$

since

$$\| P_{H_{t-n}} X_t \|^2 = \| \sum_{k=n}^{\infty} \psi_k Z_{t-k} \|^2 = \sigma^2 \sum_{k=n}^{\infty} |\psi_k|^2 \to 0.$$

(ii) If (X_t) is PND then

$$\gamma(k) = \sigma^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+k}, \quad k = 1, 2, \ldots, \quad \psi_0 = 1 \tag{5.8}$$

Note that (5.8) can be written in a matrix form

$$\begin{pmatrix} \gamma(0) & \gamma(1) & \ldots \\ \gamma(1) & \gamma(0) & \ldots \\ \ldots & \ldots & \ldots \\ \vdots & \vdots & \vdots \end{pmatrix} = \sigma^2 \begin{pmatrix} 1 & \psi_1 & \psi_2 & \ldots \\ 0 & 1 & \psi_1 & \ldots \\ \ldots & \ldots & \ldots & \ldots \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix} \begin{pmatrix} 1 & 0 & \ldots & \ldots \\ \psi_1 & 1 & \ldots & \ldots \\ \psi_2 & \psi_1 & \ldots & \ldots \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix}$$

or, equivalently, $\mathbf{\Gamma} = \sigma^2 \mathbf{\Psi} \mathbf{\Psi}'$, where $\mathbf{\Gamma} = (\gamma(i-j))_{i,j \geqslant 1}$ is the infinite covariance matrix and $\mathbf{\Psi} = (\psi_{i-j})_{i,j \geqslant 1}$, where $\psi_0 = 1$ and $\psi_i = 0$ for $i < 0$. This is obviously equivalent to Cholesky decomposition of Γ. Predictions above are constructed based on the Wold decomposition of a PND process, in particular in the development we used the property that $H_t(Z) = H_t(X)$ (cf. (5.6)). This is not necessarily true for any linear decomposition of the process. In particular, for the linear process X_t of the form

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

for $(\psi_i) \in \ell^2$, $\psi_0 \neq 0$ it is not necessarily true that $H_t(Z) = H_t(X)$ if X_t is not invertible and we know only that $H_t(X) \subseteq H_t(Z)$. Then for l-step prognosis we have

$$E(X_l - \hat{X}_l)^2 = E(X_l - P_{H_0(X)} X_l)^2 \geqslant E(X_l - P_{H_0(Z)} X_l)^2 = \sigma^2 \sum_{i=1}^{l-1} \psi_i^2. \tag{5.9}$$

However, in the case of ARMA(p, q) the following statement is valid.

Proposition 5.3.3 *Let (X_t) be causal and invertible ARMA(p, q) time series generated by white noise process with non-zero variance. The (X_t) is purely non-deterministic and its Wold decomposition coincides with causal representation of (X_t).*

Proof. Proof of the proposition follows from the observation that $Z_t = X_t - P_{H_{t-1}} X_t$, where Z_t is the white noise appearing in the definition of (X_t). Namely, as causality and invertibility implies that $H_{t-1}(X) = H_{t-1}(Z)$, we obtain by projecting both sides of causal representation of X_t that $P_{H_{t-1}} X_t = \sum_{i=1}^{\infty} \psi_i Z_{t-i}$,

and since $\psi_0 = 1$, $X_t - P_{H_{t-1}}X_t = Z_t$ we see that innovations defined in the proof of Wold theorem coincide with innovation Z_t. Similarly, equality $H_{t-1}(X) = H_{t-1}(Z)$ and(5.6) imply that $H_{-\infty}(X) = \{0\}$ and thus (X_t) is PND.
Actually, a more general statement holds true (see e.g. Doob (1953)), namely that equivalent condition for equality (5.9) to hold (or, equivalently that $H_t(X) = H_t(Z)$) is that $\psi(z) = \sum_{j=0}^{\infty} \psi_j z^j$ does not have roots for $|z| < 1$ provided $(\psi_j) \in \ell_1$.
Thus in view of the previous considerations we have

Corollary 5.3.4 *Let (X_t) be a causal and invertible ARMA(p, q) time series such that $\varphi(B)X_t = \theta(B)Z_t$ and Z_t is $WN(0, \sigma^2)$. Then*

$$P_{H_n(X)}X_{n+h} = -\sum_{j=1}^{\infty} \pi_j P_{H_n(X)}X_{n+h-j} \tag{5.10}$$

$$P_{H_n(X)}X_{n+h} = \sum_{j=h}^{\infty} \psi_j Z_{n+h-j}, \tag{5.11}$$

where (π_j) are coefficients of expansion of $\varphi(z)/\theta(z)$ and (ψ_j) are coefficients of expansion of $\theta(z)/\varphi(z)$.
Equality (5.11) implies that

$$E(X_{n+h} - P_{H_n(X)}X_{n+h})^2 = \sigma^2 \sum_{j=0}^{h-1} \psi_j^2.$$

We have proved that in the case of causal and invertible ARMA time series its Wold decomposition coincides with its causal representation and this is equivalent to (5.11). Moreover, using invertibility and $\pi_0 = 1$ we have

$$Z_{n+h} = \sum_{j=0}^{\infty} \pi_j X_{n+h-j} = X_{n+h} + \sum_{j=1}^{\infty} \pi_j X_{n+h-j}.$$

As $Z_{n+h} \perp H_n(Z) \supseteq H_n(X)$ for $h > 0$ in view of causality this implies $P_{H_n(X)}Z_{n+h} = 0$.
Projecting both sides of the equality above on $H_n(X)$ yields (5.10).

The Wiener–Kolmogorov formula. Let $\psi(z) = \theta(z)/\varphi(z)$. Then (5.11) can be stated in the following way, known as the Wiener–Kolmogorov formula.
Namely,

$$P_{H_n}X_{n+h} = \Big(\frac{\psi(B)}{B^h}\Big)_+ \frac{1}{\psi(B)}X_n, \tag{5.12}$$

where

$$\frac{\psi(B)}{B^h} = B^{-h} + \psi_1 B^{-h+1} + \cdots + \psi_h B^0 + \psi_{h+1} B^1 + \ldots$$

and $(\cdot)_+$ denotes a truncation operator defined as

$$(\sum_{i=-\infty}^{\infty} \psi_i B^i X_t)_+ := \sum_{i \geqslant 0} \psi_i B^i X_t.$$

Proof of (5.12). Note that

$$P_{H_n} X_{n+h} = \sum_{j=h}^{\infty} \psi_j Z_{n+h-j} = \Big(\frac{\psi(B)}{B^h}\Big)_+ Z_n.$$

At the same time

$$Z_n = \frac{\varphi(B)}{\theta(B)} X_n = \frac{1}{\psi(B)} X_n,$$

and the conclusion follows.

Example 5.3.5 *Let X_t be MA(1) time series $X_t = (1+\theta B)Z_t, \quad |\theta| < 1$ The Wiener-Kolmogorov formula implies that*

$$P_t X_{t+h} = \Big(\frac{1+\theta B}{B^h}\Big)_+ \quad \frac{1}{1+\theta B} X_t$$

For $h = 1$

$$\Big(\frac{1+\theta B}{B}\Big)_+ = \theta$$

and

$$P_t X_{t+1} = \frac{\theta}{1+\theta B} X_t = \theta X_t - \theta^2 X_{t-1} + \theta^3 X_{t-2} - \ldots$$

which also follows directly from (4.8).

We also provide an elegant formula for prediction error of X_t when prediction is based on the infinite past.

Proposition 5.3.6 *Assume that $\mathbf{\Gamma}_n$ is invertible. Then If $\sigma^2 = ||X_t - P_{H_{t-1}} X_t||^2 > 0$ (non-deterministic process), then $\sigma_n^2 > 0$ for any n and*

$$\sigma^2 = \exp(\lim_{n \to \infty} \frac{1}{n} \log |\mathbf{\Gamma}_n|). \tag{5.13}$$

The proof follows from (3.12) by noting that $\log \sigma_i^2 = \log(|\mathbf{\Gamma}_{i+1}|) - \log(|\mathbf{\Gamma}_i|)$ and $\sum_{i=1}^{n-1} \log \sigma_i^2 = \log(|\mathbf{\Gamma}_n|) - \log(|\Gamma_0|)$.

5.4 Predictive and autoregressive representations

We noted already that $\hat{X}_t = P_{H_{t-1}} X_t$ does not necessarily need to have representation as an infinite sum $\sum_{i>0} \varphi_i X_{t-i}$ (cf Problem 4.5). We now consider conditions under which it has such representation. The following elegant result can be proved.

Theorem 5.4.1 *Let X_t be zero mean weakly stationary non-deterministic process. Then for any $t \in \mathbb{N}$, $\hat{X}_t$ may be uniquely represented as*

$$\hat{X}_t = \sum_{i=1}^{n} \varphi_i X_{t-i} + Z_{t,n}, \tag{5.14}$$

where $Z_{t,n} \in H_{t-n-1}$.

Proof. We will use the useful fact that if $Y \in \bar{sp}(\mathcal{E}, X)$, where $\mathcal{E}$ is a closed subset of $\mathcal{L}^2$ and $X \notin \mathcal{E}$ then $Y = \gamma X + \varepsilon$, where γ is nonzero constant, $\varepsilon \in \mathcal{E}$, moreover this decomposition is unique (Problem 5.3). As $\hat{X}_t \in \bar{sp}(X_{t-1}, H_{t-2})$ and $X_{t-1} \notin H_{t-2}$ as (X_t) is non-deterministic, then $\hat{X}_t = \varphi_1 X_{t-1} + Z_{t,1}$, where $Z_{t,1} \in H_{t-2}$. Applying the same reasoning to $Z_{t,1}$ and continuing further in the same fashion we obtain the conclusion.
Note that it immediately follows from (5.14) that letting $\varepsilon_t = X_t - \hat{X}_t$ we have

$$X_t = \sum_{i=1}^{n} \varphi_i X_{t-i} + \varepsilon_t + Z_{t,n}.$$

The above decomposition is called predictive representation of (X_t).
It follows from the proof that the intuition concerning φ_i is as follows: φ_1 is coefficient corresponding to X_{t-1} in decomposition of $\hat{X}_t$ into linear combination of X_{t-1} and element of H_{t-2} (equal to $Z_{t,1}$). Further, φ_2 is coefficient corresponding to X_{t-2} in decomposition of $\hat{X}_t - \varphi_1 X_{t-1}$ into linear combination of X_{t-2} and an element from H_{t-3} and so on. Note that it is *not* decomposition into orthogonal components.
Series $\sum_{i=1}^{\infty} \varphi_i X_{t-i}$ does not need to converge. If it does, we say that (X_t) has an autoregressive (ARR) property, since then

$$X_t = \sum_{i=1}^{\infty} \varphi_i X_{t-i} + W_t + \varepsilon_t,$$

where $W_t \in H_{-\infty}(X)$ and $\varepsilon_t = X_t - \hat{X}_t$ is white noise. It follows from the fact that in this case $Z_{t,n}$ converges and it is easily shown that the limit has to belong to $H_{-\infty}$. Obviously, if the process is PND then $W_t = 0$ and we have that

$$X_t = \sum_{i=1}^{\infty} \varphi_i X_{t-i} + \varepsilon_t,$$

where the series converges in $\mathcal{L}^2$. This is AR(∞) representation. Coefficients (φ_i) are called autoregressive coefficients. Note that the decomposition solves the problem of representing $\hat{X}_t$ by infinite series. It also follows that the equivalent condition to ARR property is

$$\sum_{k=1}^{\infty}\sum_{l=1}^{\infty}\gamma(k-l)<\infty$$

(note that the above double sum is nonnegative due to the fact that $\gamma(\cdot)$ in nonnegatively definite) and sufficient condition is that (φ_i) are absolutely summable. AR(∞) representation is also useful because of the following reason. Let

$$\hat{X}_{t,n}=P_{sp(X_{t-1},\ldots,X_{t-n})}X_t.$$

Frequently in practice prediction of X_{n+h} based on n observations $X_1,\ldots,X_n$ is replaced by truncated prediction based on the whole past $H_n(X)=\overline{sp}\{X_j : j\leqslant n\}$ (truncation consisting of the first n terms of the sum) as the computation of the latter is frequently easier and, moreover, for n large both predictions are approximately equal. This is indeed the case as it may be proved that (cf. Problem 5.2 (ii)) that with no additional assumptions we have $\hat{X}_{t,n}\to\hat{X}_t$ when $n\to\infty$. As, using notation of Chapter 2, we have $\hat{X}_{t,n}=\sum_{i=1}^{n}\varphi_{n,i}X_{t-i}$ the convergence suggests that $\varphi_{n,i}$ should approximate φ_i when $n\to\infty$ and $\hat{X}_t$ is given by $\hat{X}_t=\sum_{i=1}^{\infty}\varphi_i X_{t-i}$.
It turns out that although convergence of coefficients $\varphi_{n,i}\to\varphi_i$ for any i holds for nondeterministic process (cf. Theorem 7.14 in Pourahmadi (2001)), however, norm convergence of the vectors $(\varphi_{n,1},\varphi_{n,2},\ldots,\varphi_{n,n})$ to $(\varphi_1,\varphi_2,\ldots)$ is equivalent to PND property. Representation of $\hat{X}_t=\sum_{i=1}^{\infty}\varphi_i X_{t-i}$ may not hold because of two reasons: series $(\sum_{i=1}^{n}\varphi_i X_{t-i})_n$ may not converge or $\hat{X}_t$ is not equal to it limit. It is interesting to note the connection between autoregressive and moving average coefficents.

Theorem 5.4.2 *Let $(X_t)_{t\in\mathbb{Z}}$ be nondeterministic stationary process with autoregressive and moving average coefficients (φ_i) and ψ_i respectively. Then*

$$\psi_l=\sum_{k=1}^{l-1}\psi_k\varphi_{l-k}. \tag{5.15}$$

Proof. Observe that we proved

$$\hat{X}_t-\sum_{k=1}^{n}\varphi_k X_{t-k}\in H_{t-n-1}.$$

Now, representing $\hat{X}_t$ as $\hat{X}_t=\sum_{l-1}^{n}\psi_l\varepsilon_{t-l}+r$, where $r\in H_{t-n-1}$ and $X_{t-k}=\sum_{i=0}^{n-1}\psi_i\varepsilon_{t-k-i}+r_{t-k,n-1}$, where $r_{t-k,n-1}\in H_{t-n-k}$, we have by equating coefficients corresponding to ε_{t-l} that

$$\hat{X}_t - \sum_{k=1}^{n} \varphi_k X_{t-k} = \sum_{l=1}^{n} (\psi_l - \sum_{k=1}^{l-1} \psi_k \varphi_{l-k}) \varepsilon_{t-l} + R_{n,t},$$

where $R_{n,t} \in H_{t-n-1}$. Note that the first term on the right-hand side is orthogonal to H_{t-n-1} and thus it equals 0, as the left-hand side belongs to H_{t-n-1}. Due to orthogonality of ε_{t-l} for different t this is only possible when (5.15) holds. As the last two results we state Baxter's and Debowski's inequality.

Theorem 5.4.3 *(Baxter (1960)) If $(X_t)_{t\in\mathbb{Z}}$ has continuous and positive spectral density on $(-\pi, \pi]$ then for certain $n \in \mathbb{N}$ and $c > 0$ we have that*

$$\sum_{k=1}^{n} |\varphi_{kn} - \varphi_k| \leqslant c \sum_{k=n+1}^{\infty} |\varphi_k|$$

holds when $n \geqslant N$.

Debowski's inequality is general in that it holds for any weakly stationary process.

Theorem 5.4.4 *(Dębowski (2007)) Let $\varphi_{nj} = 0$ for $j > n$ Then for $m > n$ we have*

$$\sum_{j=1}^{m} |\varphi_{mj} - \varphi_{nj}| \leqslant \prod_{k=1}^{m} (1 + |\alpha(k)|) - \prod_{k=1}^{n} (1 + |\alpha(k)|)$$

5.5 Problems

1. Let (X_t) be nondeterministic process and Z_t innovations defined in Wold decomposition.(i) Check that for $s < t$

$$H_t(X) = H_s(X) \oplus H_t(Z_{s+1}, \ldots, Z_t).$$

(ii) Prove that

$$H_t(X) = H_t(Z_t) \oplus H_{-\infty}(X).$$

2. (X_t) is weakly stationary zero mean process, $\hat{X}_{t,n} = P_{sp(X_{t-1},\ldots,X_{t-n})} X_t$ and $\hat{X}_t = P_{H_{t-1}(X)} X_t$. Prove that
(i) $\sigma_n^2 = ||X_t - \hat{X}_{t,n}||^2 \to \sigma^2 = ||X_t - \hat{X}_t||^2$ when $n \to \infty$
(ii) $\hat{X}_{t,n} \to \hat{X}_t$ when $n \to \infty$. Hint: $\sigma_n^2 - \sigma^2 = ||\hat{X}_t - \hat{X}_{t,n}||^2$.
3. Prove that if $Y \in \bar{sp}(\mathcal{E}, X)$, where $\mathcal{E}$ is a closed subset of $\mathcal{L}^2$ and $X \notin \mathcal{E}$ then $Y = \gamma X + \varepsilon$, where γ is non-zero constant and $\varepsilon \in \mathcal{E}$ and this decomposition is unique.
4. Let (X_t) is weakly stationary zero mean process. Prove that $\sigma_{n-1}^2 > 0$ is equivalent to $\Gamma_n = (\gamma(i-j)_{1\leqslant i,j\leqslant n})$ is positive definite.
5. Complete the missing details of the proof of (5.13).

6. Let $X_t = Z_t + \theta Z_{t-1}$ be MA(1) process with $|\theta| < 1$ having representation $X_{t+1} = Z_{t+1} - \sum_{j=1}^{\infty}(-\theta)^j Z_{t+1-j}$ and $\tilde{X}_{t+1} = -\sum_{j=1}^{t}(-\theta)^j Z_{t+1-j}$ is approximation of the projection of X_{t+1} on subspace $sp(Z_1, \ldots, Z_t)$. Prove that $||X_{t+1} - \tilde{X}_{t+1}||^2 = (1 + \theta^{2t+2})\sigma^2$.
7. Prove that $\varphi_{kn} \to \varphi_i$ when $n \to \infty$ for weakly stationary nondeterministic series.
8. Let $(X_t)_{t\in\mathbb{Z}}$ be PND series having ARR property. Show that

$$\sum_{k=1}^{n} |\varphi_{kn} - \varphi_k|^2 \leqslant \frac{4||\sum_{k=n+1}^{\infty} a_k X_{n-k}||^2}{\lambda_{min,n}^2},$$

where $\lambda_{min,n}$ is the minimal eigenvalue of the covariance matrix $\mathbf{\Gamma}_n$.
9. Complete the proof of the statement in the Remark 5.2.3 that if (X_t) is zero mean nondeterministic Gaussian series such that X_t and $X_{t-k}\,, k > p$ are independent given $X_{t-1}, \ldots, X_{t-p}$ then (X_t) is causal AR(p) series.

6

Spectral distribution functions and densities

In this chapter we consider weakly stationary time series with values in the complex domain $\mathbb{C}$. We discuss spectral representation of a covariance function and the weakly stationary time series. Moreover, we prove the Kolmogorov-Szegő theorem which relates error of prediction based on the whole past to spectral density.

6.1 Herglotz's theorem

Analogously to the proof of Theorem 1.1 (iii) we obtain that autocovariance function with values in the complex domain $\gamma(h) = E(X_{t+h} - EX_{t+h})(\overline{X_t - EX_t})$ is non-negative definite. It turns out that this is characterization of autocovariance functions.

Theorem 6.1.1 *Let $\gamma(\cdot) : \mathbb{Z} \longrightarrow \mathbb{C}$. Then the property that $\gamma(\cdot)$ is non-negative definite i.e. for any $a_1, \ldots, a_n \in \mathbb{C}$ and $n \in \mathbb{N}$*

$$\sum_{i,j=1}^{n} a_i \overline{a}_j \gamma(i-j) \;\geqslant\; 0$$

is equivalent to being an autocovariance function of a certain weakly stationary process. Note that the above condition implies in particular that the quadratic form is real-valued.
In this chapter we will use the following characterization of autocovariance due to Herglotz, also frequently attributed to Wiener or Khintchine.

Theorem 6.1.2 *$\gamma(\cdot) : \mathbb{Z} \longrightarrow \mathbb{C}$ is non-negative definite if and only if there exists a function $F : [-\pi, \pi] \longrightarrow \mathbb{R}^+$, which is right continuous, non-decreasing, bounded, $F(-\pi) = 0$ and such that*

$$\gamma(h) = \int_{-\pi}^{\pi} e^{ih\lambda} dF(\lambda). \tag{6.1}$$

F is uniquely defined for any $\lambda \in [-\pi, \pi]$, in particular values of possible jumps of F are uniquely specified.

Definition 14 *Function F which satisfies conditions of Theorem 6.1.2 and equality (6.1) is the called spectral distribution function. If*

$$F(\lambda) = \int_{-\pi}^{\lambda} f(s)ds \tag{6.2}$$

for a certain integrable $f \geqslant 0$, then f is called the spectral density.

If the spectral density exists then

$$\gamma(h) = \int_{-\pi}^{\pi} e^{ih\lambda} f(\lambda)\, d\lambda. \tag{6.3}$$

If spectral density exists then it is uniquely defined up to a set of Lebesgue measure 0 (m-almost everywhere). Obviously, spectral density may exist only in the case when F is continuous on $[-\pi, \pi]$. From the uniqueness in Herglotz's theorem it follows that, if a nonnegative function satisfies (6.3), then the integral of this function defined in (6.2) is the spectral distribution function pertaining to $\gamma(\cdot)$.

Proof of Theorem 6.1.2. Assume that condition (6.1) is satisfied. Then

$$\begin{aligned}\sum_{r,s=1}^{n} a_r \overline{a}_s \gamma(r-s) &= \int_{-\pi}^{\pi} \sum_{r,s=1}^{n} a_r \overline{a}_s \exp\left(i\lambda(r-s)\right) dF(\lambda) \\ &= \int_{-\pi}^{\pi} |\sum_{r=1}^{n} a_r \exp\left(i\lambda r\right)|^2\, dF(\lambda) \geqslant 0\end{aligned}$$

and thus in a view of Theorem 6.1.1 it is an autocovariance function.

Consider now arbitrary autocovariance function $\gamma(\cdot)$ and define for $N \in \mathbb{N}$

$$f_N(\lambda) = \frac{1}{2\pi N} \sum_{r,s=1}^{N} e^{i(s-r)\lambda} \gamma(r-s) e^{is\lambda} = \frac{1}{2\pi N} \sum_{|m|<N} (N - |m|) e^{-im\lambda} \gamma(m) \geqslant 0,$$

for $\lambda \in (-\pi, \pi]$, where the last inequality follows from the assumptions. The second equality follows from the fact that for $m : |m| < N$ there are exactly $N - |m|$ pairs (r, s) such that $1 \leqslant r, s \leqslant N$ and $r - s = m$. Let

$$F_N(\lambda) = \int_{-\pi}^{\lambda} f_N(\nu) d\nu.$$

Note that

$$\int_{-\pi}^{\pi} e^{ih\lambda}\, d\lambda = 2\pi I\{h = 0\}.$$

Thus for any $h \in \mathbb{Z}$ we then have

$$\int_{-\pi}^{\pi} e^{ih\lambda} dF_N(\lambda) = \frac{1}{2\pi} \sum_{|k|<N} \Big(1 - \frac{|k|}{N}\Big) \gamma(k) \int_{-\pi}^{\pi} e^{i(h-k)\lambda}\, d\lambda =$$

$$= \Big(1 - \frac{|h|}{N}\Big)\gamma(h) I\{|h| < N\}. \tag{6.4}$$

In particular for $h = 0$ it follows that $\int_{-\pi}^{\pi} dF_N(\lambda) = \gamma(0)$ for all N, which means that measures pertaining to $\{F_N\}_1^\infty$ are uniformly bounded. Moreover, they are concentrated on a compact set $[-\pi, \pi]$. Thus in view of Helly's theorem, sequence $\{F_N\}_1^\infty$ is tight (cf Billingsley (1968), chapter 6) and whence it follows that there exists sequence $\{N_k\}$ such that F_{N_k} is weakly convergent to a certain distribution function F. This means that for any g such that $g(-\pi) = g(\pi)$ we have

$$\int_{-\pi}^{\pi} g(\lambda) dF_{N_k}(\lambda) \to \int_{-\pi}^{\pi} g(\lambda) dF(\lambda).$$

In particular for $g(\lambda) = e^{ih\lambda}$ in conjunction with (6.4) it follows that

$$\gamma(h) = \int_{-\pi}^{\pi} e^{ih\lambda}\, dF(\lambda)$$

which is the needed representation of $\gamma(h)$.

Remark 6.1.3 *(i) The above proof suggests that if a spectral density exists it should satisfy*

$$f(\lambda) = \frac{1}{2\pi} \sum_{m \in \mathbb{Z}} e^{-im\lambda} \gamma(m).$$

Below we prove that this is indeed true for $\gamma(\cdot) \in \ell^1$.
(ii) In Theorem 6.1.2 it is sometimes assumed that γ is a Hermitian function i.e. it satisfies $\gamma(h) = \overline{\gamma(-h)}$. However, this follows from the assumption that $\gamma(\cdot)$ is non-negative definite. Indeed, Herglotz's theorem asserts that solely non-negative definiteness is required for function $\gamma(\cdot)$ to have a spectral representation (6.1). It sufffices to note that the right hand side of (6.1) is a Hermitian function.
(iii) Let m_F denotes a measure induced by F on $[-\pi, \pi]$. Then it can be decomposed as

$$m_F = m^F_{<<} + m^F_{\perp},$$

where $m^F_{<<}$ is absolutely continuous with respect to Lebesgue measure m on $[-\pi, \pi]$ and $m^F_{\perp}$ is singular with respect to this measure. If the singular part $m^F_{\perp}$ equals 0, then a spectral density exists and equals $f = dm^F_{<<}/dm$. It also turns out that we can characterize nondterminism and PND property in terms of this decomposition. Namely, it turns out that nondeterminism is equivalent to $\int \log f^a(\lambda)\, d\lambda > -\infty$, where f^a is the density of $m^F_{<<}$ (Szegő's theorem), if we additionally impose the condition that $m^F_{\perp} \equiv 0$ then the process is PND.
(iv) Note that $\gamma(0) = \int_{-\pi}^{\pi} dF(\lambda)$ thus Herglotz's theorem can be also stated for autocorrelatation function $\rho(h) = \gamma(h)/\gamma(0)$

$$\rho(h) = \int_{-\pi}^{\pi} e^{ih\lambda} dF^n(\lambda),$$

where $F^n(\lambda) = F(\lambda)/\gamma(0)$ is now proper cumulative distribution function called a normalized cumulative spectral distribution function.

Directly from the definition of spectral distribution function F it follows that if F_X is spectral distribution function of time series (X_t) then spectral distribution function of the process (aX_t+b) equals a^2F_X. Moreover, for the simplest possible time series $X_t = X$, where X is a given random variable, its spectral distribution takes two values only with jump at 0: $F(\lambda) = \sigma^2 I\{\lambda \geqslant 0\}$. Another example in the similar vein is given below.

Example 6.1.4 *Assume that $X_t = A\cos\omega t + B\sin\omega t$, where ω is given frequency, A, B are uncorrelated random variables and such that their mean is 0 and they have a common variance σ^2. Then $\gamma(h) = EA^2\cos\omega t\cos\omega(t+h) + EB^2\sin\omega t\sin\omega(t+h) = \sigma^2 cos(\omega(t+h) - \omega t) = \sigma^2\cos\omega h$*
Thus considering more general time series

$$X_t = \sum_{j=1}^{k} \big(A_j\cos\omega_j t + B_j\sin\omega_j t\big)$$

where $A_1, \ldots, A_k, B_1, \ldots, B_k$ are zero mean and uncorrelated, moreover $\mathrm{Var}A_j = \mathrm{Var}B_j = \sigma_j^2$, we have

$$\gamma(h) = \sum_{j=1}^{k} \sigma_j^2\cos\omega_j h. \tag{6.5}$$

Observe also that

$$F(\lambda) = \sum_{i=1}^{k} \frac{\sigma_i^2}{2}\big(I\{-\omega_i \leqslant \lambda\} + I\{\omega_i \leqslant \lambda\}\big).$$

Indeed,

$$\int_{-\pi}^{\pi} e^{ih\lambda}dF(\lambda) = \int_{-\pi}^{\pi} (\cos h\lambda + i\sin h\lambda)dF(\lambda) =$$

$$= \sum_{j=1}^{k} \big(\frac{\sigma_j^2}{2}\{\cos(-\omega_j h) + i\sin(-\omega_j h)\} + \frac{\sigma_j^2}{2}\{\cos\omega_j h + i\sin\omega_j h\}\big) = \gamma(h)$$

From the example above it follows that a spectral cumulative distribution does not cumulate probabilities as in the case of probability distribution function but

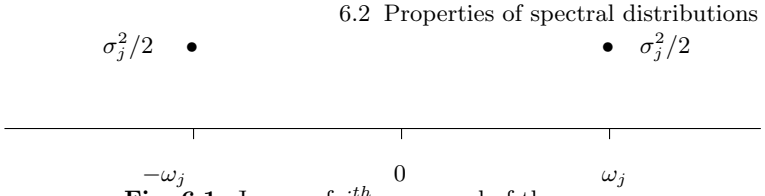


Fig. 6.1. Jumps of j^{th} summand of the process

variances of the components of the process. In the case when, informally speaking, time series is composed of a certain, possible infinite, number of sinusoidal waves of random amplitude with certain frequencies, contribution of a specific frequency to the cumulative distribution is determined by the variance of the corresponding amplitude. Also the following intuition can be gained from the example. If a distribution function has a large jump at ω or a spectral density having local maximum at ω, then the underlying process has significant periodic component with frequency ω and period $T = 2\pi/\omega$. Moreover, for ω such that $|\omega| \approx 2\pi$ we have that the period T is small and we can expect quick oscillations of the corresponding component. The opposite is true for small $|\omega|$.

6.2 Properties of spectral distributions

If $(X_t)_{t\in\mathbb{Z}}$ is real valued process, then spectral distribution function F satisfies $F(\pi^-) - F(\lambda) = F(-\lambda^-)$ for $0 \leqslant \lambda \leqslant \pi$ and the related measure μ_F determined by it on $(-\pi, \pi)$ is symmetric in the sense that for $A \subset [0, \pi)$ $\mu_F(A) = \mu_F(-A)$ and a corresponding spectral density (provided it exists) is symmetric. In such a case (6.3) reduces to the following equality

$$\gamma(h) = 2\int_0^\pi \cos(h\lambda) f(\lambda)\, d\lambda.$$

Moreover, note that as $\gamma(0) = \mathrm{Var}(X_t)$, then we have

$$\gamma(0) = \mathrm{Var}(X_t) = \int_{-\pi}^{\pi} f(\lambda) d\lambda.$$

We discuss now several results on existence and recovering of the spectral density from the related covariance function.

Theorem 6.2.1 *Assume that covariance function $\gamma(\cdot)$ is square summable and define*

$$f(\lambda) = \frac{1}{2\pi}\sum_{h=-\infty}^{\infty} e^{-ih\lambda}\gamma(h) = \frac{1}{2\pi}\big\{\gamma(0) + 2\sum_{h=1}^{\infty}\cos h\lambda\ \gamma(h)\big\}, \qquad (6.6)$$

where the convergence on the right hand side is meant in $\mathcal{L}^2(-\pi, \pi]$. Then if $f(\lambda)$ is non-negative almost everywhere with respect to Lebesque measure then $f(\cdot)$ is a spectral density.

Proof. Observe that the series on the right hand side of (6.6) is convergent in $\mathcal{L}^2(-\pi,\pi]$ as functions $e^{-ik\lambda}$ are orthogonal (which follows from the equality $\int_{-\pi}^{\pi} e^{ik\lambda}d\lambda = 2\pi I\{k=0\}$) and the assumption that $\gamma(\cdot)\in\ell^2$. Taking scalar product of both sides of (6.6) with $e^{-ih\lambda}$ and using orthogonality again we check that equation (6.3) is satisfied. As $f(\cdot)$ is non-negative in view of uniqueness property in Herglotz's theorem it is the spectral density and its integral $\int_{-\pi}^{\cdot} f(\omega)\,d\omega$ is its spectral distribution function. Non-negativity of f in the following inversion theorem is guaranteed when a slightly stronger condition is imposed on covariance function, namely that it is absolutely summable. Namely, we have

Theorem 6.2.2 *(Inversion theorem)*
If $\gamma(\cdot)$ is the covariance function such that $\sum_{i=-\infty}^{\infty}|\gamma(i)|<\infty$, then the spectral density f of $\gamma(\cdot)$ exists and is given by (6.6).

We remark that it follows from the absolute summability of $\gamma(\cdot)$ and continuity of $e^{i\lambda}$ that f given by inversion formula (6.6) is continuous.
Proof. The proof is very similar to the previous one. We note that the series on the right side of (6.6) converges almost everywhere in view of absolute summability of $\gamma(\cdot)$ We have

$$\begin{aligned}\int_{-\pi}^{\pi} f(\lambda)e^{ik\lambda}d\lambda &= \int_{-\pi}^{\pi}\{\frac{1}{2\pi}\sum_{h=-\infty}^{\infty} e^{-ih\lambda}\gamma(h)\}e^{ik\lambda}d\lambda \\ &= \sum_{h=-\infty}^{\infty}\frac{1}{2\pi}\int e^{i(k-h)\lambda}d\lambda\gamma(h)=\gamma(k), \qquad (6.7)\end{aligned}$$

where changing the order of integration is feasible as $\int_{\pi}^{\pi}\sum_h |e^{i(k-h)\lambda}\gamma(h)|d\lambda<\infty$. Moreover, $f(\lambda)\geqslant 0$, as function $f_N(\cdot)$ defined in the proof of Theorem 6.1.2 converges to f a.s. Namely,

$$0\leqslant f_N(\lambda)=\frac{1}{2\pi N}\sum_{r,s=1}^{n} e^{-irs}\gamma(r-s)e^{is\lambda}=\frac{1}{2\pi}\sum_{|h|<N}\big(1-\frac{|h|}{N}\big)e^{-ih\lambda}\gamma(h)\to f(\lambda)$$

for $\lambda\in(-\pi,\pi]$.
Variant of this result provides alternative condition for a function $\gamma(\cdot)\colon \mathbb{Z}\to\mathbb{C}$ to be an autocovariance.

Corollary 6.2.3 *If $\gamma(\cdot)\colon\mathbb{Z}\to\mathbb{C}$ is absolutely summable and*

$$f(\lambda)=\frac{1}{2\pi}\sum_{h=-\infty}^{\infty} e^{-ih\lambda}\gamma(h)\ \geqslant 0$$

then $\gamma(\cdot)$ is an autocovariance and $f(\cdot)$ its spectral density.
As in the previous proof we check that (6.3) is satisfied and the property follows from non-negativity of f and uniqueness in Herglotz's theorem.

Example 6.2.4 *(i) Consider an autocovariance function of a* $WN(0,\sigma^2)$ *process,*
$\gamma(h)=\sigma^2 I\{h=0\}$ *. Obviously, we have* $\gamma(\cdot)\in\ell^1$*. Using Theorem 6.2.2 we have*

$$f(\lambda)=\frac{1}{2\pi}\sum_{k=-\infty}^{\infty}\gamma(k)e^{-ik\lambda}=\frac{\gamma(0)}{2\pi}=\frac{\sigma^2}{2\pi},\quad \text{for } \lambda\in(-\pi,\pi),$$

thus all frequencies are represented in spectral density in the same degree.
(ii) Consider now AR(1) process, where $|\varphi|<1$*. Then*

$$\gamma(h)=\frac{\varphi^{|h|}\sigma^2}{1-\varphi^2}\ \in\ell^1$$

and applying Theorem 6.2.2 we have

$$f(\lambda)=\frac{1}{2\pi}\ \frac{\sigma^2}{1-\varphi^2}\big(1+\sum_{h=1}^{\infty}\varphi^h(e^{i\lambda h}+e^{-i\lambda h})\big)$$

$$=\frac{\sigma^2}{2\pi(1-\varphi^2)}\big(1+\frac{\varphi e^{i\lambda}}{1-\varphi e^{i\lambda}}+\frac{\varphi e^{-i\lambda}}{1-\varphi e^{-i\lambda}}\big)=$$

$$=\frac{1}{2\pi}\ \frac{\sigma^2}{1-\varphi^2}\Big(\frac{1+\varphi^2-\varphi e^{i\lambda}-\varphi e^{-i\lambda}+\varphi e^{i\lambda}-\varphi^2+\varphi e^{-i\lambda}-\varphi^2}{|1-\varphi e^{-i\lambda}|^2}\Big)$$

$$=\frac{\sigma^2}{2\pi|1-\varphi e^{-i\lambda}|^2}=\frac{\sigma^2}{2\pi(1-2\varphi\cos\lambda+\varphi^2)}.$$

It is easily seen that the maximal value of f *is taken at 0.*

In the next section we show that the result of the last example can be obtained as a special case of a general procedure without resorting to the inversion theorem.

6.2.1 Spectral properties of a linear process

Theorem 6.2.5 *Let* $(Y_t)_{t\in\mathbb{Z}}$ *be a complex-valued mean-zero weakly stationary process with spectral distribution function* $F_Y(\cdot)$*. Assume that* $(\psi_j)\in\ell^1$ *and*

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

Then spectral distribution function F_X *of* (X_t) *equals*

$$F_X(\lambda)=\int_{-\pi}^{\lambda}\psi(e^{-i\nu})dF_Y(\nu),$$

where

$$\psi(z) = \sum_{j=-\infty}^{\infty} \psi_j z^j \quad \textit{for } |z| \leqslant 1. \tag{6.8}$$

In particular, if F_Y has a spectral density f_Y, then F_X has also a spectral density and it equals

$$f_X(\lambda) = \left|\psi(e^{-i\lambda})\right|^2 f_Y(\lambda).$$

Proof. We have

$$\gamma_X(h) = \sum_{j,k=-\infty}^{\infty} \psi_j \overline{\psi_k} \gamma_Y(h-j+k) = \sum_{j,k=-\infty}^{\infty} \psi_j \overline{\psi_k} \int_{-\pi}^{\pi} e^{i\lambda(h-j+k)} dF_Y(\lambda)$$

$$= \int_{-\pi}^{\pi} e^{ih\lambda} \Big(\sum_{j=-\infty}^{\infty} \psi_j e^{-ij\lambda} \Big) \overline{\Big(\sum_{k=-\infty}^{\infty} \psi_k e^{-ik\lambda} \Big)} dF_Y(\lambda)$$

$$= \int_{-\pi}^{\pi} e^{ih\lambda} \Big| \sum_{j=-\infty}^{\infty} \psi_j e^{-ij\lambda} \Big|^2 dF_Y(\lambda)$$

where the third equality follows from absolute summability of (ψ_i).

Definition 15 *Function $\psi(e^{-i\cdot})$ is called a transfer function, function $|\psi(e^{-i\cdot})|^2$ a power transfer function, and $\Psi(B) = \sum_{-\infty}^{\infty} \psi_i B^i$ is a linear filter.*

Observe that the above result yields a powerful tool for signal modulation: by using an appropriate power transfer function we can suppress unwanted frequencies and boost desired ones.

Corollary 6.2.6 *Let X_t be purely non-deterministic process such that $\gamma(\cdot) \in \ell^2$ and $X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}$ its Wold decomposition. Then*

$$|\psi(e^{-i\lambda})|^2 \sigma^2 = \sum_{k=-\infty}^{\infty} \gamma(k) e^{-ik\lambda} \tag{6.9}$$

in $\mathcal{L}^2[-\pi, \pi]$. Moreover,

$$\gamma(k) = \frac{\sigma^2}{2\pi} \int_{-\pi}^{\pi} |\psi(e^{-i\lambda})|^2 e^{ik\lambda} \, d\lambda \tag{6.10}$$

and

$$\sum_{k=-\infty}^{\infty} |\gamma(k)|^2 = \frac{\sigma^4}{2\pi} \int_{-\pi}^{\pi} |\psi(e^{-i\lambda})|^4 \, d\lambda. \tag{6.11}$$

The second equality follows from the Parseval theorem after noting that (6.10) implies that $(\gamma(\cdot))_k$ are Fourier coefficients of the function $\sigma^2|\psi(e^{-i\cdot})|^2$ with respect to the orthogonal basis $(e^{ik\cdot})_k$.

Example 6.2.7 *Consider ARMA(p, q) process $\varphi(B)X_t = \theta(B)Z_t$, where $Z_t \sim WN(0, \sigma^2)$ and such that $\varphi(z) \neq 0$ for $|z| = 1$. Then we proved that*

$$X_t = \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}, \tag{6.12}$$

where (ψ_j) are absolutely summable coefficients in the expansion of $\psi(z) = \theta(z)/\varphi(z)$ for $|z| = 1$. But this means that (6.8) holds and it follows that

$$f_X(\lambda) = \frac{\sigma^2}{2\pi} \frac{|\theta(e^{-i\lambda})|^2}{|\varphi(e^{-i\lambda})|^2}. \tag{6.13}$$

In a special case of AR(p) process we obtain that

$$f_X(\lambda) = \frac{\sigma^2}{2\pi} |1 - \varphi_1 - \varphi_1^{-ip\lambda} - \varphi_2 e^{-i2\lambda} - \cdots - \varphi_p e^{-ip\lambda}|^2 \tag{6.14}$$

In particular (6.13) implies the form of a spectral density for AR(1) time series we computed before. For MA(1) process we obtain that spectral density equals $(\sigma^2/2\pi)(1 + 2\theta\cos\lambda + \theta^2)$.

Example 6.2.8 *Consider the usual structural equation $\varphi(B)X_t = \theta(B)Z_t$ for a given white noise (Z_t). We will show now that in the case of ARMA(p, q) time series if polynomial $\varphi(z)$ does not have zeros on the unit circle $|z| = 1$ then there exists polynomial $\tilde{\varphi}$ such that the solution $\tilde{X}_t$ of the structural equation $\tilde{\varphi}(B)\tilde{X}_t = \theta(B)\tilde{Z}_t$ with a certain white noise $(\tilde{Z}_t)$ is causal and $\tilde{X}_t$ has the same covariance structure as the solution to the original structural equation. We will use spectral domain approach which in this example clearly shows advantages of employing it. We represent $\varphi(z)$ as $\varphi(z) = \prod_{i=1}^{p}(1 - a_i^{-1}z)$, where a_i are roots of $\varphi(\cdot)$ and the representation uses the fact that the constant term in $\varphi(\cdot)$ is one. Without loss of generality we can assume that $a_1, \ldots, a_s$ are roots lying outside the unit circle and $a_{s+1}, \ldots, a_p$ inside it. Note that no roots lie on the unit circle itself. Let*

$$\tilde{\varphi}(z) = \prod_{i=1}^{s}(1 - a_i^{-1}z) \prod_{i=s+1}^{p}(1 - \bar{a}_i z)$$

and consider a stationary solution (W_t) to the structural equation $\tilde{\varphi}(B)W_t = \theta(B)Z_t$. The roots of $\tilde{\varphi}(z)$ are $a_1, \ldots a_s, \bar{a}_{s+1}^{-1}, \ldots, \bar{a}_p^{-1}$. Note that as $\bar{a}_i^{-1} = (a_i/|a_i|)|a_i|^{-1}$ and thus $|\bar{a}_i^{-1}| > 1$ for $i = s+1, \ldots, p$, then we have that all roots of $\tilde{\varphi}(z)$ lie outside the unit circle and the causal solution to the last structural equation exists. Moreover as

$$|1 - \bar{a}_j e^{-i\lambda}| = |e^{i\lambda} - \bar{a}_j| = |e^{-i\lambda} - a_j| = |a_j||1 - a_j^{-1}e^{-i\lambda}|$$

in view of (6.13) we have that

$$f_X(\lambda) = \Big(\prod_{j=s+1}^{p} |a_j|^2 \Big) f_W(\lambda),$$

where $f_W(\lambda)$ denotes the spectral density of (W_t). Thus changing white noise Z_t which is $WN(0, \sigma^2)$ to white noise $WN(0, \sigma^2 \prod_{j=s+1}^{p} |a_j|^2)$ and denoting it by $\tilde{Z}_t$ we have that causal solution to structural equation $\tilde{\varphi}(B)\tilde{X}_t = \theta(B)\tilde{Z}_t$ has the same spectral density and whence autocovariance function as (X_t).
Analogously, if we assume that both autoregressive and moving average polynomials $\varphi(\cdot)$ and $\theta(\cdot)$ do not have zeros for $|z| = 1$ reasoning analogously we can find causal and invertible ARMA process having the same covariance structure as the original time series.

We consider now the problem of optimal linear prediction in frequency domain. We note that the problem of 1-step prediction for mean-zero stationary real-valued process

$$\text{argmin}_{a_1,\ldots,a_n \in \mathbb{R}} ||X_{n+1} - \sum_{i=1}^{n} a_i X_{n+1-i}||^2$$

can be rephrased in the following way. Criterion function equals

$$\begin{aligned} &< X_{n+1} - \sum_{i=1}^{n} a_i X_{n+1-i}, X_{n+1} - \sum_{i=1}^{n} a_i X_{n+1-i} > \\ &= \gamma(0) + \sum_{1 \leqslant k,l \leqslant n} a_k a_l \gamma(k-l) - 2 \sum_{1 \leqslant k \leqslant n} a_k \gamma(k) \\ &= \int_{-\pi}^{\pi} |1 - \sum_{k=1}^{n} e^{ik\lambda} a_k|^2 dF(\lambda). \end{aligned} \tag{6.15}$$

Thus in frequency domain optimal prediction is equivalent to analytic problem of finding a minimum of the above integral which is a squared distance between 1 and trigonometric polynomial in $L^2([-\pi, \pi], dF)$.

6.3 The Kolmogorov–Szegő theorem

In this section we state and prove in the partial case a deep result known as the Kolmogorov–Szegő theorem which shows how the prediction error relates to the spectral density. Actually, Szegő proved that $\int \log f(\lambda)\, d\lambda > -\infty$ implies that the process is nondeterministic, which is a corollary to celebrated formula (6.19) proved by Kolmogorov. We will discuss the case when the spectral density is continuous and strictly positive. Before that we prove results on AR and MA approximations which are interesting in their own light and they are also

used in the proof of the main result of this section. We will prove that a continuous spectral density can be approximated by a spectral density of a causal autoregressive process and a spectral density of invertible moving average process. Unfortunately, the result does not specify orders of these processes which ensure desired accuracy of approximation.
The following results holds true, with $||\cdot||_\infty$ denoting sup norm on $[-\pi, \pi]$.

Theorem 6.3.1 *Let f be a spectral density of a weakly stationary process and assume that it is continuous. For any $\varepsilon > 0$ there exist $p, q \in \mathbb{N}$ and a causal autoregressive process (U_t) of order p and invertible moving average process (V_t) of order q such that their pertaining spectral densities f_U and f_V satisfy*

$$\max(||f - f_U||_\infty, ||f - f_V||_\infty) \leqslant \varepsilon.$$

We will first prove the crucial lemma.

Lemma 6.3.2 *Let Let f be a continuous spectral density. For any $\varepsilon > 0$ there exists polynomial $a(z) = \sum_{i=0}^{p} a_i z^i$, $a_i \in \mathbb{R}$, $a_0 = 1$ with roots outside the unit circle such that*

$$||A|a(e^{i\cdot})|^2 - f(\cdot)||_\infty < \varepsilon, \tag{6.16}$$

where $A = (2\pi(\sum_{i=0}^{p} a_i^2))^{-1} \int_{-\pi}^{\pi} f(\lambda)\, d\lambda$.

Proof of the Lemma 6.3.2 (outline). We first note that if $C(z) = \sum_{k=-p}^{p} c_k z^k$ is such that $c_k = c_{-k}$, $c_0 = 1$, $c_k \in \mathbb{R}$ and that $C(z) \neq 0$ for $|z| = 1$ then

$$C(z) = Ka(z)a(z^{-1}), \tag{6.17}$$

where $a(z)$ satisfies assumptions of the Lemma and K is some constant. In particular $C(e^{i\lambda}) = K|a(e^{-i\lambda})|^2$. Indeed, $c_k = c_{-k}$ implies that $c_k(z^{-1})^k = c_{-k}z^{-k}$ and it follows that z_0 is the root of $C(z)$ only if z_0^{-1} is also a root. Thus polynomial $z^p C(z)$ of order $2p$ can be represented as $K\prod_{j=1}^{p}(1 - z/\eta_j)(1 - z\eta_j)$, where η_j, η_j^{-1} are its roots and $|\eta_j| > 1$. Obvious manipulations yield

$$\begin{aligned} C(z) &= K\prod_{j=1}^{p}(1 - \frac{z}{\eta_j})(\frac{1}{z} - \eta_j) = (-1)^p K\prod_{j=1}^{p}\eta_j(1 - \frac{z}{\eta_j})(1 - \frac{1}{\eta_j z}) \\ &= \tilde{K}a(z)a(z^{-1}), \end{aligned}$$

where

$$a(z) = \prod_{j=1}^{p}(1 - \frac{z}{\eta_j}) = \sum_{i=1}^{p} a_i z^i \quad \text{and} \quad \tilde{K} = (-1)^p K\prod_{i=1}^{p}\eta_j.$$

Note that if $\mathrm{Im}(\eta_j) \neq 0$ for some $1 \leqslant j \leqslant p$ then the conjugate root $\bar{\eta}_j$ satisfies $\bar{\eta}_j = \eta_k$ for some $1 \leqslant k \leqslant p$ as obviously $|\bar{\eta}_j| = |\eta_j| > 1$. Since $P_w(z) = (1 - wz)(1 - \bar{w}z)$ for any $w \in \mathbb{C}$ has real coefficients it readily follows that coefficients $a_1, \ldots, a_p$ of $a(\cdot)$ are real.
In order to prove (6.16) note first that by truncating $f := \max(f, \delta)$ for any

$\delta >$ we may assume that $f \geqslant \delta$. Moreover as f is continuous on $[-\pi, \pi]$ and $f(-\pi) = f(\pi)$, Ces'aro means of it Fourier expansion converge uniformly to f, namely

$$||n^{-1} \sum_{i=0}^{n-1} S_i f - f||_\infty \to 0, \tag{6.18}$$

where $S_k f(\lambda) = \sum_{|j| \leqslant k} f_j e^{ij\lambda}$ and $f_j = (2\pi)^{-1} \int_{-\pi}^{\pi} f e^{ij\lambda}\, d\lambda$. Note that the approximand of f in (6.18) can be written as

$$W_n = n^{-1} \sum_{j=0}^{n-1} \sum_{|k| \leqslant j} \big(1 - \frac{|k|}{n}\big) f_k \exp(-ik\lambda).$$

For $n \geqslant n_0(\delta)$ such that $||W_n - f||_\infty < \delta$ we easily check that W_n satisfies (6.17) (note that $f \geqslant \delta$ is used to ensure that $W_n(z) \neq 0$ for $|z| = 1$). Thus equating constants on both sides of $W_n(z) = \tilde{K} a(z) a(z^{-1})$ we have

$$\tilde{K}(1 + a_1^2 + \cdots + a_p^2) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(\lambda)\, d\lambda.$$

From this and (6.18) the lemma readily follows as in view of $a_i \in \mathbb{R}$ we have $a(e^{-i\lambda}) a(e^{i\lambda}) = |a(e^{-i\lambda})|^2$.
To prove the theorem for autoregressive approximation apply the Lemma above to f_ε^{-1}, where $f_\varepsilon = \max(f, \varepsilon/2)$. Note that f_ε^{-1} is a bounded continuous spectral density. Then it readily follows by elementary manipulations that f is uniformly approximated by $K^{-1}|a(e^{-i\lambda})|^{-2}$, where $a(z) \neq 0$ for $|z| \leqslant 1$, which is a spectral density of a causal AR(p) process pertaining to polynomial $a(\cdot)$ and WN$(0, 2\pi/K)$ in view of (6.13).

A deep result expressing error of prediction based on the whole past in terms of spectral density is given by the Kolmogorov–Szegő's theorem . The assumption on continuity strict positivity of the spectral density is unnecessary and is imposed only to make proof more accessible, for the general case see Grenander and Szegő (1958). In general f in Kolmogorov's formula below is replaced by the density f^a of the absolutely continuous part of the spectral measure.

Theorem 6.3.3 *Let $(X_t)_{t \in \mathbb{Z}}$ be a weakly stationary process with continuous spectral density $f(\cdot)$ bounded away from 0. Then*

$$\sigma^2 = \| X_t - P_{H_{t-1}} X_t \|^2 = 2\pi \exp\big\{ \frac{1}{2\pi} \int_{-\pi}^{\pi} \log f(\lambda) d\lambda \big\} \tag{6.19}$$

We will check first that the result holds for causal AR(p) process. In view of (6.13) we have

$$\int_{-\pi}^{\pi} \log f(\lambda) d\lambda = 2\pi \log \frac{\sigma^2}{2\pi} - \sum_{j=1}^{p} \int_{-\pi}^{\pi} \log |1 - a_j e^{-ij\lambda}|^2 d\lambda, \tag{6.20}$$

where a_j are reciprocals of the roots of $\varphi(\cdot)$ and since the process is causal we have $|a_j| < 1$. Using expansion $\log(1-z) = -\sum_{j=1}^{\infty} z^j/j$ valid for $|z| < 1$ we have further for the summands in the above expression

$$\begin{aligned}\int_{-\pi}^{\pi} \log|1 - a_j e^{-ij\lambda}|^2 d\lambda &= \int_{-\pi}^{\pi} \big(\log(1 - a_j e^{-ij\lambda}) + \log(1 - \bar{a}_j e^{ij\lambda})\big) d\lambda \\ &= -\int_{-\pi}^{\pi} \Big(\sum_{k=1}^{\infty} \frac{a_j^k e^{ik\lambda}}{k} + \sum_{k=1}^{\infty} \frac{\bar{a}_j^k e^{ik\lambda}}{k}\Big) d\lambda = 0\end{aligned}$$

as the change of summation and integration is valid in view of absolute summability of $(a_j^k/k)_k$. Thus only the first term in is nonzero and the theorem holds in this case.

Consider now the general case when f is continuous spectral density such that $\inf_{\lambda\in[-\pi,\pi]} f(\lambda) = \varepsilon > 0$. Using the previous result to approximate spectral densities $f + \varepsilon/2$ and $f - \varepsilon/2$ with accuracy $\varepsilon/2$ we construct two spectral densities $g_{1,\varepsilon}$ and $g_{2,\varepsilon}$ of causal autoregressive processes such that

$$f(\lambda) - \varepsilon \leqslant g_{1,\varepsilon}(\lambda) \leqslant f(\lambda) \leqslant g_{2,\varepsilon}(\lambda) \leqslant f(\lambda) + \varepsilon$$

for $\lambda \in [-\pi, \pi]$. It follows from (6.15) that the corresponding prediction error satisfy

$$\sigma_n^2(g_{1,\varepsilon}) \leqslant \sigma_n^2(f) \leqslant \sigma_n^2(g_{2,\varepsilon})$$

and the bounds converge to prediction errors based on the whole past $\sigma^2(g_{1,\varepsilon})$ and $\sigma^2(g_{2,\varepsilon})$, respectively, which are given by

$$\sigma^2(g_{i,\varepsilon}) = 2\pi \exp\big\{\frac{1}{2\pi}\int_{-\pi}^{\pi} \log g_{i,\varepsilon}(\lambda) d\lambda\big\},$$

for $i = 1, 2$. As spectral densities $g_{i,\varepsilon}$ converge uniformly to f when $\varepsilon \to 0$ and as they are bounded away from 0 it also easily follows that $\sigma^2(g_{2,\varepsilon}) - \sigma^2(g_{1,\varepsilon}) \to 0$ which proves the result.

Corollary 6.3.4 *Under conditions of Theorem 6.3.3 we have for $n \to \infty$*

$$\frac{1}{n}\sum_{i=1}^{n} \log \tilde{\lambda}_i \to \frac{1}{2\pi}\int_{-\pi}^{\pi} \log f(\lambda)\, d\lambda,$$

where $\tilde{\lambda}_i$, $i = 1, \ldots, n$ are eigenvalues of matrix $\mathbf{\Gamma}_n/2\pi$.

Proof. The corollary follows from (5.13) which can be restated as

$$\frac{1}{n}\sum_{i=1}^{n} \log \tilde{\lambda}_i \to \log(\sigma^2/2\pi).$$

Corollary 6.3.5 *Assume that $\mathbb{X}_t$ is a stationary Gaussian sequence with a spectral density satisfying assumptions of Theorem 6.3.3. Then*

$$h = \lim_{n\to\infty} \frac{H(X_1, \ldots, X_n)}{n} = \frac{1}{2}\log(2\pi e) + \log\sigma. \tag{6.21}$$

Proof. Proof readily follows from Theorems 6.3.3 and 2.6.2 together with equality (3.30).
We refer to Gray (2006) for a review of results in this vein. Theorem 6.3.3 also yields a useful sufficient condition for checking that the process is deterministic.

Corollary 6.3.6 *If*

$$\int_{-\pi}^{\pi} \log f(\lambda) d\lambda = -\infty \tag{6.22}$$

then the process (X_t) is deterministic.

Proof. We note that if a set $\{\lambda \in (-\pi, \pi), f(\lambda) = 0\}$ has positive Lebesgue measure, or, equivalently measure of the support of f is less then 2π then (6.22) holds. Indeed, as $\log f(\lambda) \leqslant f(\lambda)$ we have

$$\int_{\{\lambda: f(\lambda)>0\}} \log f(\lambda) d\lambda \leqslant \int_{\{\lambda: f(\lambda)>0\}} f(\lambda) d\lambda = \gamma(0) \ < \ \infty. \tag{6.23}$$

thus

$$\int \log f(\lambda) d\lambda = \int_{\{\lambda:\, f(\lambda)>0\}} \log f(\lambda) d\lambda + \int_{\{\lambda:\, f(\lambda)=0\}} \log f(\lambda) d\lambda = -\infty,$$

since the first integral is bounded from above and the second is equal to $-\infty$. Thus if a set $\{\omega \in (-\pi, \pi) : \ f(\omega) = 0\}$ has positive Lebesgue measure then it follows from Kolmogorov–Szegő's theorem that a process having spectral density $f(\cdot)$ is deterministic.

6.4 Spectral representation of a weakly stationary time series

We briefly discuss an important analogue of Herglotz's representation for a weakly stationary time series itself, namely the equality

$$X_t = \int_{(-\pi,\pi]} e^{it\lambda}\, dZ(\lambda), \qquad t \in \mathbb{Z}, \tag{6.24}$$

where $Z(\lambda)$ for $\lambda \in [-\pi, \pi]$ is a zero-mean process with orthogonal (uncorrelated) increments i.e. such that

$$< Z(\lambda_4) - Z(\lambda_3), Z(\lambda_2) - Z(\lambda_1) >= 0, \quad \text{where } \lambda_1 \leqslant \lambda_2 \leqslant \lambda_3 \leqslant \lambda_4$$

and which moreover is right-continuous i.e. $Z(\lambda+\delta) \to Z(\lambda)$ in $\mathcal{L}^2$ when $\delta \to 0^+$.

Definition 16 *Equality (6.24) is called the spectral representation of a weakly stationary process* $(X_t)_{t\in\mathbb{Z}}$.

First we make it clear what is meant by the right hand side of (6.24). For a given process $(Z(\lambda))_{\lambda\in[-\pi,\pi]}$ which is right-continuous with orthogonal increments we define

$$F(\lambda) = ||Z(\lambda) - Z(-\pi)||^2$$

and check that F satisfies properties of a spectral distribution function listed in Herglotz's theorem, namely that it is bounded, non-decreasing and right-continuous function on $[-\pi,\pi]$ such that $F(-\pi) = 0$. Note that e.g. for the Brownian motion B starting at $-\pi$ such that $\mathrm{Var}B(\lambda) = (\lambda+\pi)\sigma^2$ we have $F(\lambda) = (\lambda+\pi)\sigma^2$. Thus F defines a measure on $[-\pi,\pi]$ for which F is a cumulative distribution. Moreover, it turns out that for $g \in \mathcal{L}^2([-\pi,\pi],F)$ we can define an integral $I(g) = \int_{(-\pi,\pi]} g(\lambda)\,dZ(\lambda)$ by a continuous extension of a natural definition of the integral for simple functions, namely

$$I(g) = \sum_{j=1}^{p} a_i(Z(\lambda_{i+1}) - Z(\lambda_i)), \quad \text{where} \quad g(\lambda) = \sum_{j=1}^{p} a_j I_{(\lambda_j,\lambda_{j+1}]}(\lambda).$$

One then proves that $I(h)$ is a linear transformation of $\mathcal{L}^2([-\pi,\pi],F)$ onto the closed subspace of $\mathcal{L}^2$ such that

$$< I(g), I(h) >= \mathrm{Cov}(I(g), I(h)) =< g, h >_F := \int_{-\pi}^{\pi} g(\lambda)\bar{h}(\lambda)\,dF(\lambda). \qquad (6.25)$$

Now in order to prove (6.24) we need to associate with any weakly stationary process $(X_t)_{t\in\mathbb{Z}}$ a specific process $(Z(\lambda))_{\lambda\in[-\pi,\pi]}$ with orthogonal increments such that its pertaining distribution function F coincides with the spectral distribution function F_{spec} of the considered time series. To this end we define so called Kolmogorov isomorfism, which transforms time domain onto frequency domain. It is a transform T defined on $\mathcal{H} = \bar{sp}(X_t)$ onto the space $\mathcal{L}^2([-\pi,\pi],F_{spec})$ defined by a continuous extension of its definition for linear combinations of values of time series at given time points

$$T(\sum_{j=1}^{p} g_j X_{t_j}) = \sum_{j=1}^{p} g_j e^{it_j\cdot}.$$

Then defining

$$Z(\lambda) = T^{-1}(I_{(-\pi,\lambda]}(\cdot))$$

it is shown that $(Z(\lambda))_{\lambda\in[-\pi,\pi]}$ is a zero-mean process with orthogonal increments such that $T^{-1}(g) = I(g)$. Then from the definition of I we obtain

$$I(e^{it\cdot}) = \int_{(-\pi,\pi]} e^{it\lambda}\,dZ(\lambda)$$

and on the other hand obviously we have that $T^{-1}(e^{it\cdot}) = X_t$. As $I(e^{it\cdot}) = T^{-1}(e^{it\cdot})$ we obtain (6.24). From (6.25) and uniqueness property in Herglotz's representation it also follows that $F = F_{spec}$.
It can be also shown that as a spectral distribution function is uniquely determined, process $(Z(\lambda))$ appearing in (6.24) is unique in the sense that for any two processes $(Z_1(\cdot))$ and $(Z_2(\cdot))$ giving rise to representation (6.24) random variables $(Z_1(\lambda))$ and $(Z_2(\lambda))$ should coincide with probability one for any $\lambda \in [-\pi, \pi]$.
Intuitively, representation (6.24) yields decomposition of X_t into a sum of sinusoids $e^{it\lambda}dZ(\lambda)$ where $dZ(\lambda)s$ are random amplitudes which are uncorrelated for different values of λ.
Spectral representation (6.24) can be applied to study properties of sample characteristics of the process. We give one representative example.

Example 6.4.1 *If (X_t) is weakly stationary zero-mean process then*

$$\bar{X}_n = n^{-1}\sum_{t=1}^{n} X_t \to Z(0)$$

in $\mathcal{L}^2$, where $Z(\lambda)_{\lambda\in[-\pi,\pi]}$ is a process defined in (6.24). Since $||Z(0)||^2 = F(0) - F(0^-)$ it follows that when F is continuous at 0 the sample mean converges to the mean of marginal distribution. We will return to this subject in the next chapter. In order to prove the convergence note that the spectral representation yields

$$n^{-1}\sum_{t=1}^{n} X_t = \int n^{-1}\sum_{i=1}^{n} e^{it\lambda}dZ(\lambda) = \int n^{-1}e^{i\lambda}\frac{1-e^{in\lambda}}{1-e^{i\lambda}}dZ(\lambda).$$

The integrand is bounded by 1 and converges pointwise to $I\{\lambda = 0\}$. Thus the result follows from the Lebesque dominated convergence theorem.

6.5 Problems

1. Check that for uncorrelated time series X_t i Y_t we have $F_{X+Y} = F_X + F_Y$, where F_X denotes spectral distribution function of (X_t).
2. Find autocorrelation function for a time series with a spectral density

$$f(\lambda) = \frac{\pi - |\lambda|}{\pi^2} I\{|\lambda| \leqslant \pi\}$$

3. Prove Wiener's theorem:

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}|\gamma(h)|^2 = \sum_{\lambda_i}(F(\lambda_i) - F(\lambda_i^-))^2,$$

where (λ_i) denote jumps of spectral distribution function F.
In particular it follows from this problem that if F does not have jumps then $\sum_{i=1}^{n} |\gamma(h)|^2 = o(n)$.
4. Let X_t i Y_t have spectral densities f_X i f_Y and autocovariance matrices $\Gamma_{n,X}$ and $\Gamma_{n,Y}$, respectively. Prove that if $f_X(\lambda) \geqslant f_Y(\lambda)$ for $\lambda \in [-\pi, \pi]$ then:
(i) $\Gamma_{n,X}$ -$\Gamma_{n,Y} \geqslant 0$;
(ii) $\mathrm{Var}(b'X) \geqslant \mathrm{Var}(b'Y)$, where $X = (X_1, \ldots, X_n)'$ and $b = (b_1, \ldots, b_n)'$.
5. Let $X_t = A\cos(\pi t/3) + B\sin(\pi t/3)$, where A, B are uncorrelated zero mean random variables with the common variance σ^2. Prove that X_t does not have a spectral density.
6. Let (X_t) be time series with autocovariance $\gamma(h) = 2\sin h/h$ for $h \neq 0$ and $\gamma(0) = 2$.
Find its spectral density.
7. Find a spectral density of moving average of order 3 of X_t that is of the process $Y_t = (X_{t-1} + X_t + X_{t+1})/3$ in terms of a spectral density of X_t. In the case when (X_t) is $\mathrm{WN}(0, \sigma^2)$ find minima and maxima of the spectral density of (Y_t).
8. Dirichlet kernel is defined as $D_n(\lambda) = \sum_{k=-n}^{n} e^{ik\lambda}$. Check the following equality

$$D_n(\lambda) = \frac{1 - e^{i(n+1)\lambda}}{1 - e^{i\lambda}} + \frac{1 - e^{-i(n+1)\lambda}}{1 - e^{-i\lambda}} - 1 = \frac{\sin(n + 1/2)\lambda}{\sin(\lambda/2)} \tag{6.26}$$

for $\lambda \neq 0$ and $D_n(0) = 2n + 1$.
9. Fejér kernel $F_n(\lambda)$ is defined as an arithmetic mean of first n Dirichlet kernels devided by 2π. Check that

$$2\pi F_n(\lambda) = \frac{(D_0(\lambda) + D_1(\lambda) + \ldots + D_{n-1}(\lambda))}{n} = \frac{1}{n}\frac{1 - \cos n\lambda}{1 - \cos\lambda} = \frac{1}{n}\frac{\sin^2 n\lambda/2}{\sin^2 \lambda/2}$$

for $\lambda \neq 0$ and $2\pi F_n(0) = n$.
10. Prove using the inversion formula that if $\gamma(\cdot)$ is absolutely summable then the pertaining spectral density is continuous.
11. Prove that for $\mathrm{MA}(q)$ process with pertaining polynomial $\theta(z) = \sum_{i=0}^{q} \theta_i z^i$, where $\theta_0 = 0$ its spectral density equals

$$f(\lambda) = \frac{\sigma^2}{2\pi}\Big(\sum_{i=0}^{q} \theta_i^2 + \sum_{k=1}^{q}\sum_{i=k}^{q} \theta_i\theta_{i-k}\cos k\lambda\Big).$$

12. Prove that if F is the spectral cumulative distribution of the process with $\gamma(0) = 1$ and ξ is a random variable on $[-\pi, \pi]$ distributed according to F then F is the spectral function of the process $X_t = e^{it\xi}$. Note that this indicates that definition of deterministic process does not cover many processes which intuitively should be treated as such.

7

Estimation of the mean and the correlation function

In this chapter we discuss estimation of the first and the second order characteristics of a weakly stationary process. Asymptotic distributions for a sample mean and correlation function will be proved under assumption that the underlying time series is a linear process with independent innovations.

7.1 Estimation of the mean

Assume that $(X_t)_{t\in\mathbb{N}}$ is a weakly stationary time series and that a block consisting of the first n observations $X_1, \ldots, X_n$ is observable. Let $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$ be a sample mean. Our aim is to investigate the properties of $\bar{X}_n$ as the estimator of the mean μ of marginal distribution. The crucial property here is stationarity, in particular the fact that all observable random variables $X_1, \ldots, X_n$ have mean μ and thus $E\bar{X}_n = \mu$. Recall also that due to ergodic theorem $\bar{X}_n \to \mu$ a.s. provided $(X_t)_{t\in\ N}$ is ergodic. However, the inference, in particular construction of confidence interval for μ, is complicated by the dependence between variables which may lead to a large, and not easily estimable variance of $\bar{X}_n$. Recall that for an iid sequence provided that $EX_1^2 < \infty$, we have

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{\mathcal{D}} N(0, \mathrm{Var}X_1)$$

which makes it possible to construct confidence interval for μ and test hypotheses $\mu = \mu_0$. We will check under what conditions the result above, with possibly different asymptotic variance, can be extended to stationary time series.
Observe that since the block of indices $\{1, 2, \ldots, n\}$ for $1 \leqslant h \leqslant n$ contains exactly $n - h$ pairs (i, j) such that $j - i = h$, we have (see figure below)

$$\mathrm{Var}\bar{X}_n = n^{-2}\mathrm{Var}\big(\sum_{i=1}^n X_i\big) = n^{-2}\big(\sum_{i=1}^n \mathrm{Var}X_i + 2\sum_{1\leqslant i<j\leqslant n} \mathrm{Cov}(X_i, X_j)\big) =$$
$$n^{-2}\big(n\gamma(0) + 2\sum_{h=1}^{n-1}\gamma(h)(n-h)\big) = \frac{\gamma(0)}{n} + \frac{2}{n^2}\sum_{h=1}^{n-1}(n-h)\gamma(h). \tag{7.1}$$

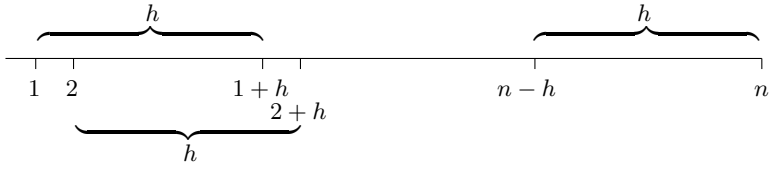


Fig. 7.1. There are $(n-h)$ pairs (j,i) such that $j>i$ and $j-i=h$

Thus

$$n\mathrm{Var}(\bar{X}_n)=\gamma(0)+2\sum_{h=1}^{n-1}(1-\frac{h}{n})\gamma(h) \tag{7.2}$$

Thus the natural question is under what conditions we have

$$2\sum_{h=1}^{n-1}\Big(1-\frac{h}{n}\Big)\gamma(h)\ \approx\ 2\sum_{h=1}^{\infty}\gamma(h).$$

This is answered by the following proposition

Proposition 7.1.1 *If*

$$\sum_{h=1}^{\infty}|\gamma(h)|<\infty \tag{7.3}$$

then

$$n\,Var(\bar{X}_n)\longrightarrow\gamma(0)+2\sum_{h=1}^{\infty}\gamma(h).$$

The limit equals $2\pi f(0)$*, where* $f(\lambda)=(2\pi)^{-1}\sum_{k=-\infty}^{\infty}\gamma(k)e^{-ik\lambda}$ *is a spectral density.*

Proof. From (7.2) it follows that it is enough to prove that (7.3) implies

$$\sum_{h=1}^{n-1}\frac{h}{n}\gamma(h)\longrightarrow 0.$$

For any $\varepsilon>0$

$$|\sum_{h=1}^{n-1}\frac{h}{n}\gamma(h)\,|\leqslant\sum_{h=1}^{[n\varepsilon]}|\,\frac{h}{n}\gamma(h)\,|+\sum_{h=[n\varepsilon]+1}^{\infty}|\gamma(h)|\leqslant\varepsilon\sum_{h=1}^{[n\varepsilon]}|\gamma(h)|+\sum_{h=[n\varepsilon]+1}^{\infty}|\gamma(h)| \tag{7.4}$$

For $\eta>0$ we let $\varepsilon=\eta/2\sum_{h=1}^{\infty}|\gamma(h)|$, and note the first sum is not larger than $\eta/2$. For such ε we choose $n_0(\varepsilon)$ such that the second sum is less than $\eta/2$ for $n\geqslant n_0(\varepsilon)$. Then the left hand side of (7.4) is less than η.

Remark 7.1.2 *For a linear process*

$$X_t = m + \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}, \quad (Z_t)_{t\in\mathbb{Z}} - WN(0,\sigma^2), \quad \sum_{j=-\infty}^{\infty} |\psi_j| < \infty$$

we have (i)

$$\sum_{h=-\infty}^{\infty} \gamma(h) = \sigma^2 \Big(\sum_{j=-\infty}^{\infty} \psi_j \Big)^2. \tag{7.5}$$

Moreover,

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

Namely,

$$\gamma(h) = \sigma^2 \Big(\sum_{i=-\infty}^{\infty} \psi_i \psi_{i+h} \Big)$$

and thus

$$\sum_{h=-\infty}^{\infty} \gamma(h) = \sigma^2 \Big(\sum_{h=-\infty}^{\infty} \sum_{i=-\infty}^{\infty} \psi_i \psi_{i+h} \Big) = \sigma^2 \Big(\sum_{j=-\infty}^{\infty} \psi_i \Big)^2$$

which validates (i). Inequality in (ii) is justified analogously.

7.2 Asymptotic distribution of $\bar{X}_n$ for the linear process

We first state Ibragimov-Linnik theorem from which asymptotic distribution of the mean for the linear process follows easily.

Theorem 7.2.1 *(Ibragimow and Linnik (1971))*
Assume that $X_t = \sum_{i=-\infty}^{\infty} a_j \varepsilon_{t-j}$, *where* $(a_j) \in \ell^2$ *and* (ε_j) *is the strong white noise with the finite second moment. Assume that* $\sigma_n^2 = \mathrm{Var}(S_n) \to \infty$. *Then*

$$S_n/\sigma_n \xrightarrow{\mathcal{D}} N(0,1)$$

when $n \to \infty$.

Note that σ_n^2 is not assumed to be of order n and because of this the last result is a useful tool for studying long-range dependent sequences.
Proof. Note that $S_n = \sum_{j=1}^n X_j = \sum_{k=-\infty}^{\infty} \big(\sum_{j=1}^n a_{j-k} \big) \varepsilon_k$ and thus

$$\sigma_n^2 = \mathrm{Var}(S_n) = \sum_{k=-\infty}^{\infty} \big(\sum_{j=1}^n a_{j-k} \big)^2.$$

Let $w_{kn} = \sum_{j=1}^n a_{j-k}$. We will first prove that

$$\frac{w_{kn}^2}{\sigma_n^2} \leqslant \frac{4\sum_{k=-\infty}^{\infty} a_k^2}{\sigma_n}\Big(1+\frac{1}{2\sigma_n}\Big). \tag{7.6}$$

Then it will follow in view of assumptions that $\max_k w_{kn}^2/\sigma_n^2 \to 0$ when $n \to \infty$. In order to prove (7.6) note that as $w_{k-l,n} = w_{k-l-1,n} - a_{n-(k-l-1)} + a_{1-(k-l)}$ we have

$$\frac{w_{k-l,n}^2}{\sigma_n^2} = \frac{a_{l+1-k}^2 + a_{n+l+1-k}^2}{\sigma_n^2} + \frac{2(a_{l+1-k} - a_{n+l+1-k})w_{k-l-1,n}}{\sigma_n^2} + \frac{w_{k-l-1,n}^2}{\sigma_n^2}.$$

Analogously,

$$\frac{w_{k-l+1,n}^2}{\sigma_n^2} = \frac{a_{l-k}^2 + a_{n+l-k}^2}{\sigma_n^2} + \frac{2(a_{l-k} - a_{n+l-k})w_{k-l,n}}{\sigma_n^2} + \frac{w_{k-l,n}^2}{\sigma_n^2},$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$\frac{w_{k,n}^2}{\sigma_n^2} = \frac{a_{1-k}^2 + a_{n+1-k}^2}{\sigma_n^2} + \frac{2(a_{1-k} - a_{n+1-k})w_{k-1,n}}{\sigma_n^2} + \frac{w_{k-1,n}^2}{\sigma_n^2}.$$

Using the above inequalities sequentially from the last to the first and applying to the middle term inequalities $(x+y)^2 \leqslant 2(x^2+y^2)$ and $|w_{k-i}|/\sigma_n \leqslant 1$ we obtain

$$\frac{w_{k,n}^2}{\sigma_n^2} \leqslant \frac{2\sum_{k=-\infty}^{\infty} a_k^2}{\sigma_n^2} + \frac{4\sum_{k=-\infty}^{\infty} a_k^2}{\sigma_n} + \frac{w_{k-l-1,n}^2}{\sigma_n^2}.$$

As the term $w_{k-l-1,n}^2/\sigma_n^2$ can be made arbitrarily small by choice of sufficiently large $l = l(k)$ (possibly depending on k), (7.6) is proved.
Let $a_{kn} = w_{kn}/\sigma_n$. Thus we have

$$\sigma_n^{-1} S_n = \sigma_n^{-1}(X_1 + \cdots + X_n) =: \sum_{k=-\infty}^{\infty} a_{kn}\varepsilon_k,$$

where $\sum_k a_{kn}^2 = 1$ and $\max_k a_{kn}^2 \to 0$ when $n \to \infty$. It is now easy to see that Lindeberg's condition is satisfied. Indeed, we write

$$\sigma_n^{-1} S_n = \sum_{i=1}^{2N+2} \xi_{ni},$$

where $\xi_{n1} = \sum_{|k|>N} a_{kn}\varepsilon_k$ and for $2 \leqslant i \leqslant 2N+2$, $\xi_{ni} = a_{i-N-2,n}\varepsilon_{i-N-2}$, where $N = N(n)$ is chosen such that $\sum_{|k|>N} a_{kn}^2 \leqslant \eta_n$ and η_n is some sequence tending to 0. Now observe that with c_n^2 denoting the upper bound in (7.6) we have

$$\sum_{i=1}^{2N+2} E(\xi_{ni}^2 I\{|\xi_{ni}| \geqslant \gamma\}) \leqslant \sum_{k=2}^{2N+2} a_{kn}^2 E(I\{|\varepsilon_1| \geqslant \gamma/c_n\}\varepsilon_1^2) + \eta_n^2 = o(1)$$

in view of $\varepsilon_1 \in \mathcal{L}^2$ and $\sum_k a_{kn}^2 = 1$. Thus Lindeberg's condition is satisfied.

Theorem 7.2.2 *Assume that (X_t) is a linear process*

$$X_t = \mu + \sum_{k=-\infty}^{\infty} \psi_k \varepsilon_{t-k},$$

where (ε_t) is a strong WN$(0, \sigma^2)$ and moreover

$$\sum_{k=-\infty}^{\infty} |\psi_k| < \infty, \qquad \sum_{k=-\infty}^{\infty} \psi_k \neq 0.$$

Then

$$n^{1/2}(\bar{X}_n - \mu) \xrightarrow{\mathcal{D}} N(0, v),$$

where

$$v = \sigma^2 \Big(\sum_{i=-\infty}^{\infty} \psi_i \Big)^2.$$

The proof of the result follows from the Ibragimov-Linnik theorem and Proposition 7.1.1. The assumption $\sum_{k=-\infty}^{\infty} \psi_k \neq 0$ implies that $v > 0$, if it is not satisfied we have $n^{1/2}(\bar{X}_n - \mu) \xrightarrow{\mathcal{D}} 0$. Moreover, due to Proposition 7.1.1 and (7.5) asymptotic variance v equals to the limit of $n\mathrm{Var}(\bar{X}_n)$.
Below we give the second derivation of the asymptotic distribution of the mean which is based on small block-large block method devised by Bernstein, which will be also used to derive Bartlett's formula below.
Proof (outline). We will use the following lemma (cf. Billingsley (1968), Theorem 4.2)

Lemma 7.2.3 *Let X_{mn}, (X_m) i (Y_n) be sequences of random variables such that $X_{mn} \xrightarrow{\mathcal{D}} X_m$ when $n \to \infty$, $X_m \xrightarrow{\mathcal{D}} X$ when $m \to \infty$ and moreover*

$$\lim_{m\to\infty} \limsup_{n\to\infty} P(|X_{mn} - Y_n| \geqslant \varepsilon) = 0.$$

Then

$$Y_n \xrightarrow{\mathcal{D}} X.$$

The result can be depicted in the following diagram.

$$\begin{array}{ccc} X_{mn} & \xrightarrow{n\to\infty} & X_m \\ & & \Big\downarrow \mathcal{D}\ m\to\infty \\ \underset{(Y_n \text{ close to } X_{mn})}{Y_n} & \xrightarrow[n\to\infty]{} & X \end{array}$$

Proof of theorem. We consider truncated linear process ($m \in \mathbb{N}$).

$$X_t^m = m + \sum_{j=-m}^{m} \psi_j \varepsilon_{t-j}$$

Obviously X_t^m is not observable and it is used only in this proof. Observe that $(X_t^m)_{t\in N}$ is $(2m)$–dependent, which means that X_t^m, X_s^m are independent for $|s-t| > 2m$, which follows from the fact that

$$\{\varepsilon_{t-j}, j = -m, \ldots, m\} \cap \{\varepsilon_{s-i}, i = -m, \ldots, m\} = \emptyset$$

and (ε_t) is the *strong* white noise.
We use the lemma for $X_{mn} := n^{1/2}(n^{-1}\sum_{t=1}^{n} X_t^m - \mu)$. We prove that

$$X_{mn} = \underset{n\to\infty}{\xrightarrow{\mathcal{D}}} N\big(0, \sigma^2(\sum_{j=-m}^{m} \psi_j)^2\big) \underset{m\to\infty}{\xrightarrow{\mathcal{D}}} N(0, v). \tag{7.7}$$

The first convergence in (7.7) is implied by the Central Limit Theorem for $(2m)$–dependent variables, stating that

$$n^{1/2}(\bar{X}_n - \mu) \xrightarrow{\mathcal{D}} N(0, \sum_{j=-2m}^{2m} \gamma(j)). \tag{7.8}$$

This is proved in the following way by Bernstein's large block-small block method: consider $k >> 2m$ and divide the observations into blocks: $k - 2m$ observations of the first block, then the gap consisting of $2m$ observations, then the second block consisting of $k-2m$ observations and so on. Note that $l = [n/k]$ is the index of the last block. Let Σ_i denote the sum of the observations of ith block.

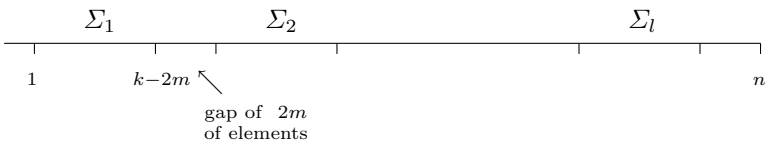


Fig. 7.2. Division of $X_1, \ldots, X_n$ into blocks and gaps

Note that $\Sigma_1, \ldots, \Sigma_l$ are iid variables having the same distribution and we use the CLT for iid observations applied to $(\Sigma_1 + \cdots + \Sigma_l)/n$. Note that when $n \to \infty$

$$n\mathrm{Var}\Big(\frac{\Sigma_1+\cdots+\Sigma_l}{n}\Big)=\frac{(k-2m)l}{n}(k-2m)\mathrm{Var}\Big(\frac{\Sigma_1}{k-2m}\Big)\to\sum_{j=-2m}^{2m}\gamma(j) \quad (7.9)$$

in view of Proposition 7.1.1. Moreover it can be proved that the average of observations outside the blocks becomes negligible. Namely, there are $[n/k]$ inner gaps consisting of $2m$ observations each and a (possible) final 'leftover' consisting of $n-k[n/k]$ observations. Observe that that the normalized variance of the average of observations from inner gaps equals

$$\frac{[n/k]\sum_{|j|<2m}(2m-|j|)\gamma(j)}{n}\to 0$$

when $k\to\infty$. Thus indeed under this condition the average pertaining to the inner blocks becomes negligible in view of the Chebyshev inequality. Similarly we show that the analogous average of elements from the final leftover part becomes negligible. This proves the first convergence in (7.7).
For the second convergence in (7.7) it is enough to note that $(\sum_{j=-m}^{m}\psi_j)^2$ converges to v.
We check the conditions of the lemma for $Y_n:=n^{1/2}(\bar{X}_n-\mu)$. Observe that

$$\mathrm{Var}(Y_n-X_{mn})=n\mathrm{Var}\Big(\frac{1}{n}\sum_{t=1}^{n}\sum_{j:|j|>m}\psi_j\varepsilon_{t-j}\Big)\underset{n\to\infty}{\longrightarrow}\Big(\sum_{j:|j|>m}\psi_j\Big)^2\sigma^2.$$

Thus

$$\lim_{m\to\infty}\limsup_{n\to\infty}\mathrm{Var}(Y_n-X_{mn})=0.$$

Now the last condition of the lemma follows from the Chebyshev inequality, since

$$P(|Y_n-X_{mn}|\geqslant\varepsilon)\leqslant\frac{\mathrm{Var}(Y_n-X_{mn})}{\varepsilon^2}$$

Remark 7.2.4 *We derive now asymptotic distribution of $\bar{X}_n$ for the process AR(1)*

$$X_t-m=\varphi(X_{t-1}-\mu)+\varepsilon_t,$$

where (ε_t) is a strong WN$(0,\sigma^2)$.
As $\gamma(h)=\varphi^{|h|}\sigma^2/(1-\varphi^2)$ we have

$$v=\big(1+2\sum_{h=1}^{\infty}\varphi^h\big)\frac{\sigma^2}{(1-\varphi^2)}=\big(1+\frac{2\varphi}{(1-\varphi)}\big)\frac{\sigma^2}{(1-\varphi^2)}=\frac{\sigma^2}{(1-\varphi)^2}.$$

Confidence interval for μ is thus

$$\Big(\bar{X}_n-\frac{z_{1-\alpha/2}\ \sigma}{(1-\varphi)\sqrt{n}},\quad \bar{X}_n+\frac{z_{1-\alpha/2}\ \sigma}{(1-\varphi)\sqrt{n}}\Big)$$

Note that the length of the interval equal $2z_{1-\alpha/2}\,\sigma/(1-\varphi)\sqrt{n}$ *is increasing function of* φ*, thus the larger value the longer the confidence interval. We compare the length of this interval for AR(1) process with the length of an analogous interval for a strong white noise (iid sequence). To be fair we consider the strong white noise having the same marginal variance as the autoregressive process, namely* $\sigma^2/(1-\varphi^2)$*. In this case, length of the confidence interval equals*

$$\Big(\bar{X}_n - \frac{z_{1-\alpha/2}\;\sigma}{\sqrt{1-\varphi^2}\sqrt{n}},\quad \bar{X}_n + \frac{z_{1-\alpha/2}\;\sigma}{\sqrt{1-\varphi^2}\sqrt{n}}\Big)$$

It is easy to see that since $1+\varphi < 1-\varphi$ *for* $\varphi < 0$*, for such* φ *confidence interval for independent case is actually longer than for AR(1), thus in this case dependence helps in more precise estimation of the mean. This is obviously due to the fact the asymptotic variance of the mean in this case is smaller than for white noise and is related to oscillations of AR(1) for* $\varphi < 0$ *which are averaged out when the mean is calculated.*

7.3 Estimation of the covariance and correlation function

We deal now with estimation of the covariance function $\gamma(h)$ for a weakly stationary process

$$\gamma(h) = E((X_t - \mu)(X_{t+|h|} - \mu))$$

In order to estimate $\gamma(h)$ we replace the expected value in the above expression by an average of $(X_t - \bar{X}_n)(X_{t+|h|} - \bar{X}_n)$ for all possible pairs $(X_t, X_{t+|h|})$ such that $1 \leqslant t \leqslant n,\ 1 \leqslant t+|h| \leqslant n$.

Fig. 7.3. Pairs $(X_i, X_{i+h}), i = 1, \ldots, n-|h|$

As the number of such pairs is $n-|h|$ the usual estimator of $\gamma(h)$ is

$$\widetilde{\gamma}(h) = \frac{1}{n-|h|}\sum_{t=1}^{n-|h|}(X_{t+|h|} - \bar{X}_n)(X_t - \bar{X}_n)$$

However, we show below that in order to obtain an estimator which is non-negative definite we have to change the norming factor to n^{-1}, namely

$$\widehat{\gamma}(h) = \frac{1}{n} \sum_{t=1}^{n-|h|} (X_{t+|h|} - \bar{X}_n)(X_t - \bar{X}_n) \tag{7.10}$$

Definition 17 *Estimator $\widehat{\gamma}(h)$ defined in (7.10) will be called empirical covariance for lag h.*

Recall that $\mathbf{\Gamma}_n = \big(\gamma(i-j)\big)_{1\leqslant i,j\leqslant n}$ is non-negative definite, and it is desirable that the same property holds for its estimator. Indeed such property holds for $\widehat{\mathbf{\Gamma}}_n = \big(\widehat{\gamma}(i-j)\big)_{1\leqslant i,j\leqslant n}$ but not necessarily for $\widetilde{\mathbf{\Gamma}}_n = \big(\widetilde{\gamma}(i-j)\big)_{1\leqslant i,j\leqslant n}$. In order to see why $\widehat{\mathbf{\Gamma}}_n$ is non-negative definite observe that we put $Y_i = X_i - \bar{X}$ and define

$$\mathbf{T}_{n\times 2n-1} = \begin{pmatrix} \overbrace{0 \ \ldots\ldots \ 0}^{n-1} \ Y_1 \ Y_2 \ \ldots \ Y_{n-1} \ Y_n \\ \underbrace{0 \ \ldots\ldots \ 0}_{n-2} \ Y_1 \ Y_2 \ Y_3 \ \ldots \ \ Y_n \ \ 0 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ Y_1 \ldots Y_{n-1} \ Y_n \ 0 \ \ldots \ \ 0 \ \ 0 \end{pmatrix}$$

then we have

$$\widehat{\mathbf{\Gamma}}_n = n^{-1}\mathbf{T}\mathbf{T}' \geqslant 0.$$

Remark 7.3.1 *One may prove that $\widehat{\mathbf{\Gamma}}_n$ is strictly positive definite and thus invertible provided that $\hat{\gamma}(0)$ is positive. This always happens if not all observations are equal. This follows from the fact that if $\gamma(\cdot)$ is such that $\gamma(0) > 0$ and $\lim_{h\to\infty} \gamma(h) = 0$, then for any n covariance matrix $\mathbf{\Gamma}_n$ is strictly positive definite.*

Note also that as the number $n - |h|$ of pairs $(X_t, X_{t+|h|})$ gets smaller for large h, it is unwise to use $\widehat{\gamma}(h)$ for large h. Rules of thumb similar to the following ones: estimate $\widehat{\gamma}(h)$ only for $h :\leqslant n/4$ or $n \geqslant 50$, $h \leqslant \sqrt{n}$ are frequently used without much theoretical justification. Also note that it follows from the definition that $\widehat{\gamma}(h)$ and $\widehat{\gamma}(j)$ are dependent random variables. The same is also true for empirical autocorrelation function defined by

Definition 18 *Estimator*

$$\widehat{\rho}(h) = \frac{\widehat{\gamma}(h)}{\widehat{\gamma}(0)}. \tag{7.11}$$

is called empirical correlation for lag h.

7.4 Bartlett's theorem

Define vector of h first autocorrelations beginning from the autocorrelation at lag 1 as follows

$$\boldsymbol{\rho}(h) = (\rho(1), \ldots, \rho(h))'$$

and the corresponding vector of empirical autocorrelations (compare (7.11)) as

$$\widehat{\boldsymbol{\rho}}(h) = (\widehat{\rho}(1), \ldots, \widehat{\rho}(h))'$$

We state now Bartlett's theorem which states the asymptotic distribution of centred and normalized vector $\widehat{\boldsymbol{\rho}}(h)$.

Theorem 7.4.1 *Let (X_t) be the linear process*

$$X_t = m + \sum_{j=-\infty}^{\infty} \psi_j \varepsilon_{t-j}$$

such that (Z_t) is a strong white noise $WN(0, \sigma^2)$, $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$ and $EZ_t^4 < \infty$.
Then

$$n^{1/2}(\widehat{\boldsymbol{\rho}}(h) - \boldsymbol{\rho}(h)) \xrightarrow{\mathcal{D}} N(0, \mathbf{W}),$$

where $\mathbf{W} = (w_{ij})_{i,j \leqslant h}$,

$$w_{ij} = \sum_{k=1}^{\infty} \lambda_{ki} \lambda_{kj}.$$

and

$$\lambda_{ki} = \rho(k+i) + \rho(k-i) - 2\rho(k)\rho(i) \tag{7.12}$$

Proof (outline). We assume that $EX_t = 0$ and we first prove the result with $\hat{\gamma}(h)$ replaced by $\gamma_0(h) = n^{-1} \sum_{t=1}^{n} X_t X_{t+h}$. Let $\kappa = E\varepsilon_t^4/(E\varepsilon^2)$ denote the kurtosis of ε_i. We note that

$$E(\varepsilon_s \varepsilon_t \varepsilon_u \varepsilon_v) = \begin{cases} \kappa\sigma^4 & \text{if } |\{s,t,u,v\}| = 1 \\ \sigma^4 & \text{if } |\{s,t,u,v\}| = 2 \\ 0 & \text{otherwise} \end{cases} \tag{7.13}$$

Moreover

$$E(X_t X_{t+p} X_{t+h+p} X_{t+h+p+q}) = \sum_{i,j,k,l} \psi_i \psi_{j+p} \psi_{k+h+p} \psi_{l+h+p+q} E(\varepsilon_{t-i} \varepsilon_{t-j} \varepsilon_{t-k} \varepsilon_{t-l}). \tag{7.14}$$

Observe that the subsum of the (7.14) corresponding to $i = j = k = l$ equals

$$\kappa\sigma^4 \sum_i \psi_i \psi_{i+p} \psi_{i+h+p} \psi_{i+h+p+q} =: \kappa\sigma^4 A,$$

corresponding to $i = j \neq k = l$ equals

$$\sigma^4 \sum_{i \neq k} \psi_i \psi_{i+p} \psi_{k+p+h} \psi_{k+p+h+q} = \gamma(p)\gamma(q) - \sigma^4 A,$$

corresponding to $i = k \neq j = l$ equals

$$\sigma^4 \sum_{i \neq j} \psi_i \psi_{i+p+h} \psi_{j+p} \psi_{j+p} \psi_{j+p+h+q} = \gamma(p+h)\gamma(q+h) - \sigma^4 A$$

and corresponding to $i = l \neq j = k$ equals

$$\sigma^4 \sum_{i \neq j} \psi_i \psi_{i+p+h+q} \psi_{j+p} \psi_{j+p+h} = \gamma(p+h+q)\gamma(h) - \sigma^4 A.$$

Thus (7.14) equals

$$(\kappa-3)\sigma^4 \sum_i \psi_i \psi_{i+p} \psi_{i+h+p} \psi_{i+h+p+q} + \gamma(p)\gamma(q) + \gamma(p+h)\gamma(q+h) + \gamma(p+h+q)\gamma(h)$$

and we have

$$E(\gamma_0(p)\gamma_0(q)) = n^{-2} \sum_{s=1}^{n} \sum_{t=1}^{n} E(X_t X_{t+p} X_s X_{s+q}) =$$
$$n^{-2} \sum_{s=1}^{n} \sum_{t=1}^{n} [\gamma(p)\gamma(q) + \gamma(s-t)\gamma(s-t-p+q) + \gamma(s-t+q)\gamma(s-t-p) +$$
$$+ (\kappa-3)\sigma^4 \sum_i \psi_i \psi_{i+p} \psi_{i+s-t} \psi_{i+s-t-q}.$$

As the right hand side of the above expression depends on s and t only through $s-t$ we let $k = s-t$ and obtain after subtracting $E(\gamma_0(p))E(\gamma_0(q)) = \gamma(p)\gamma(q)$,

$$n\mathrm{Cov}(\gamma_0(p)\gamma_0(q)) = \sum_{k=-n+1}^{n-1} (1 - |k|/n) T_k,$$

where

$$T_k = \gamma(k)\gamma(k-p+q) + \gamma(k+q)\gamma(k-p) + (\kappa-3)\sigma^4 \sum_i \psi_i \psi_{i+p} \psi_{i+k} \psi_{i+k+q}.$$

It is easy to show that $(T_k)_{-\infty}^{\infty}$ are summable and thus

$$\lim_{n\to\infty} n\mathrm{Cov}(\gamma_0(p)\gamma_0(q)) = \sum_{k=-\infty}^{\infty} T_k$$
$$= (\kappa-3)\gamma(p)\gamma(q) + \sum_{k=-\infty}^{\infty} \gamma(k)\gamma(k-p+q) + \gamma(k+q)\gamma(k-p).$$

Using Bernstein's block method as in the proof of asymptotic normality of the mean we show that

$$n^{1/2}(\gamma_0(0)-\gamma(0),\ldots,\gamma_0(0)-\gamma(0)) \xrightarrow{\mathcal{D}} N(0,V),$$

where $V=(v_{pq})$ is $(h+1)\times(h+1)$ covariance matrix such that v_{pq} is given by $\sum_{k=-\infty}^{\infty} T_k$ above. The next step is to show that for each $0 \leqslant p \leqslant h$, $n^{1/2}(\hat{\gamma}(p)-\gamma_0)=o_P(1)$. The last step is to use the delta method with the function $f:\mathbb{R}^{p+1}\to\mathbb{R}^p$ equal $f(x_0,x_1,\ldots,x_p)=(x_1/x_0,\ldots,x_p/x_0)$ and to note that $f(\hat{\gamma}(0),\ldots,\hat{\gamma}(h))=(\hat{\rho}(1),\ldots,\hat{\rho}(h))$. It follows that the resulting limiting covariance is of the form $\tilde{W}=DVD'$ and $D=\gamma(0)^{-1}[-\boldsymbol{\rho}(h),I_h]$. Some algebra shows that $\tilde{w}_{ij}$ coincide with w_{ij} given in the statement of the theorem. This ends the proof.

Corollary 7.4.2 *Assume that $(Z_t)_{t\in\mathbb{Z}}$ is a strong white noise $WN(0,\sigma^2)$ such that $EZ_t^4<\infty$. Then*

$$n^{1/2}(\widehat{\boldsymbol{\rho}}(h)-\rho(h))=n^{1/2}\widehat{\boldsymbol{\rho}}(h) \xrightarrow{\mathcal{D}} N(0,I) \tag{7.15}$$

Proof of the corollary. Note first that white noise satisfies the assumptions of the theorem with $\psi_j=I\{j=0\}$. Moreover, $\rho(k)=I\{k=0\}$. Thus it follows from the definition of λ_{ki} that it is non-zero for $k\geqslant 1$ only in the case when $k=i$ as then $\lambda_{ki}=\rho(k-i)=1$. Thus it follows that $\lambda_{ki}\lambda_{kj}\neq 0$ only in the case when $k=i=j$ and $w_{ij}=I\{i=j\}$. The conclusion of the corollary then obviously follows.

Asymptotic convergence in (7.15) is used to construct heuristic test of the hypothesis

$$H_0:\ (X_t)_{t\in\mathbb{Z}} \ \text{ is strong } \ \mathrm{WN}(0,\sigma^2)$$

by means of test statistic

$$T=\#\{i:\ i=1,\ldots,h \quad |\widehat{\rho}(i)|>z_{1-\alpha/2}n^{-1/2}\}=:\sum_{i=1}^{h} J_i$$

with rejection region $\mathcal{C}=\{T>\alpha h\}$. Value $\alpha=0.05$ is usually used. Note that under null hypothesis H_0 indicators J_i are binary random variables equal 1 with probability approximately α, thus $ET\approx h\alpha$. Frequently, a quick and dirty test is performed which rejects the hypothesis when number T of correlations falling outside the respective confidence interval exceeds its expected value. Obviously this is not, even approximately, the test on the level α. Such tests can be constructed using the observation that the distribution of $\sum_{i=1}^{h} I\{\sqrt{n}|\hat{\rho}(i)|>c\}$ converges in distribution to random variable $W_c=\sum_{i=1}^{h} I\{|Z_i|>c\}$, where Z_i are independent standard normal random variables. This follows by continuous mapping theorem by noting that the transform is discontinuous only on the set of measure zero for limiting distribution. Then for each $k<h$ we determine threshold c_α as $c_\alpha=\inf_c\{W_c\geqslant k\}\leqslant\alpha$. The problem here is that the distribution of W_c is discrete and it is not evident how to choose an appropriate value of k.

Let us also note that individual test statistic J_i with rejection region $\{J_i=1\}$

does have approximately level α for testing $\rho(i) = 0$ against the alternative $\rho(i) \neq 0$, however test statistic $\max_{i=1,\ldots,h} J_i$ with rejection region $\{\max_{i=1,\ldots,h} J_i = 1\}$ *does not* have level α even approximately. It is obviously related to multiple testing problem, as this test is equivalent to testing h individual hypotheses and rejecting H_0 when at least one of them is rejected when the critical level for each test was calculated for the single hypothesis case.

Consider now $MA(1)$ with innovations being strong white noise and calculate asymptotic variances w_{ii}. By an easy calculation we have

$$w_{11} = (1 - 2\rho^2(1))(1 - 2\rho^2(1)) + \rho^2(1) = 1 - 3\rho^2(1) + 4\rho^4(1),$$

when the first summand corresponds to $k = 1$ in (7.12) and the second to $k = 2$. Analogously for $i \geqslant 2$

$$w_{ii} = 1 + \rho^2(1) + \rho^2(1)$$

when the first summand corresponds to $k = i$ in (7.12) and the second and the third to $k = i - 1$ and $k = i + 1$.
Thus for $i \geqslant 2$, $\widehat{\rho}(i)$ belongs to confidence interval $\pm 1.96 n^{-1/2}(1 + 2\widehat{\rho}^2(1))^{1/2}$ with probability approximately 0.95. Analogously, for MA(q) series

$$w_{ii} = 1 + 2(\rho^2(1) + \cdots + \rho^2(q)), \quad \text{for } i > q \tag{7.16}$$

as for $k = i - q, \ldots, i + q$ we get non-zero summands $\rho(k - i)$ in (7.12) equal $\rho(-q), \ldots, \rho(0), \ldots \rho(q)$.

Remark 7.4.3 *Frequently instead of testing the hypothesis H_0 that the proces (X_t) is white noise we test the hypothesis $H_0^{'}$: (X_t) is $MA(q-1)$. Thus when $H_0^{'}$ is satisfied, in view of 7.16, squared standard error of $\widehat{\rho}(q)$ can be defined as.*

$$SE^2_{\widehat{\rho}(q)} = \frac{1 + 2(\widehat{\rho}^2(1) + \cdots + \widehat{\rho}^2(q-1))}{n}. \tag{7.17}$$

7.5 Problems

1. Using Bartlett's theorem justify that for AR(1) time series it holds

$$\lim_{n\to\infty} n\mathrm{Var}(\hat{\rho}(k)) \to \frac{(1 - \varphi^{2k})(1 + \varphi^2)}{(1 - \varphi^2)} - 2i\varphi^{2k}.$$

2. (i) Using problem 1 and delta method calculate the asymptotic variances of the estimators of autoregressive coefficient φ defined as $\hat{\varphi} = \hat{\rho}(1)$ and $\tilde{\varphi} = (\hat{\rho}(3))^{1/3}$.
(ii) Check that when $\varphi \to 1$ ratio of the asymptotic variances of the estimators $\tilde{\varphi}$ and $\hat{\varphi}$ tends to infinity.
3. Check that $\sum_{i=0}^{n-1} \hat{\gamma}(h) = 0$, where $\widehat{\gamma}(h) = \frac{1}{n}\sum_{t=1}^{n}(X_{t+h} - \bar{X}_n)(X_t - \bar{X}_n)$.
4. Prove that for the weakly stationary time series we have

(i) $E(\bar{X}_n^2) \leqslant \gamma(0)$;
(ii) If the autocorrelation function is nonnegative then $\lim_n nE(\bar{X}_n^2)$ exists (but is possibly infinite);
(iii) Construct an example of the process that $\lim_{n\to\infty} nE(\bar{X}_n^2) = 0$.
5. Show that if $\gamma(h) \to 0$ when $h \to \infty$ then $\mathrm{Var}(\bar{X}_n) \to 0$ when $n \to \infty$.
6. Fill in the missing details in the proof of CLT of $2m$-dependent variables, in particular justify (7.9).

8

Parameter estimation for ARMA(p, q) time series

We discuss parameter estimation of ARMA(p, q) time series assuming that orders p and q are known, or, in the the case of AR(p) process that the upper bound of p can be given. We start with a discussion of the latter.

8.1 Estimation for AR(p) time series

We will describe two basic methods of estimation of autoregressive parameters and variance of innovations: the Yule-Walker and Burg's estimators.

8.1.1 The Yule-Walker estimators

We consider again zero-mean ARMA(p, q) time series $(X_t)_{t\in\mathbb{Z}}$ described by $p + q + 1$ unknown parameters: $(\varphi_1, \ldots, \varphi_p)$, $(\theta_1, \ldots, \theta_q)$ and σ^2. We now describe basic methods of estimation of these parameters. We start with a particular situation when $q = 0$ and the process reduces to autoregressive time series AR(p). We assume additionally that $(X_t)_{t\in\mathbb{Z}}$ is causal and, without loss of generality that the mean is zero. The simplest method of estimation of its parameters is based on the Yule–Walker equations. They are obtained when covariances of $X_{t-i}, i = 0, 1, \ldots, p$ with both sides of structural equation

$$X_t - \sum_{i=1}^{p} \varphi_i X_{t-i} = \varepsilon_t$$

are calculated and equalled. Then we obtain using causality

$$\begin{cases} \gamma(0) - \varphi_1\gamma(1) - \cdots - \varphi_p\gamma(p) = \sigma^2 \\ \qquad \gamma(i) - \sum_{j=1}^{p} \varphi_i\gamma(i-j) = 0 \quad i = 1, 2, \ldots, p. \end{cases} \tag{8.1}$$

Remark 8.1.1 *We note that the above equations coincide with the Yule-Walker equations for the optimal linear predictor considered in Chapter 2. The sole difference is that in the above equations coefficients of polynomial $\varphi(z)$ appear instead of the coefficients of the best linear predictor. This is because those two sets of coefficients coincide in the case of causal AR(p) time series. Namely, it follows that $(\varphi_{p1}, \ldots, \varphi_{pp}) = (\varphi_1, \ldots, \varphi_p)$ and in general for $m \geqslant p$ vector $(\varphi_{m1}, \ldots, \varphi_{mp})$ equals $(\varphi_1, \ldots, \varphi_p, 0, \ldots, 0)$, where the number of appended zeros is $m - p$.*

We can write (8.1) in the matrix form

$$\boldsymbol{\Gamma}_p \boldsymbol{\varphi} = \boldsymbol{\gamma}_p$$

$$\sigma^2 = \gamma(0) - \boldsymbol{\varphi}' \boldsymbol{\gamma}_p$$

$$\boldsymbol{\gamma}_p = (\gamma(1), \gamma(2), \ldots, \gamma(p))'$$

where, as usual, $\boldsymbol{\Gamma}_p = (\gamma(i-j))_{1 \leqslant i,j \leqslant p}$ and $\boldsymbol{\varphi}_p = (\varphi_{p1}, \ldots, \varphi_{pp})'$.
If $\boldsymbol{\Gamma}_p$ is invertible then

$$\boldsymbol{\varphi} = \boldsymbol{\Gamma}_p^{-1} \boldsymbol{\gamma}_p \qquad \sigma^2 = \gamma(0) - \boldsymbol{\gamma}_p' \boldsymbol{\Gamma}_p^{-1} \boldsymbol{\gamma}_p. \tag{8.2}$$

The Yule-Walker estimators are plug-in estimators, obtained when in the above equations $\gamma(i)$, $i = 0, 1, \ldots, p$ are replaced by $\widehat{\gamma}(i)$. Thus having observed a part $X_1, \ldots, X_n$ of a sample path we calculate empirical covariances $\widehat{\gamma}(i)$ and then we construct $\widehat{\boldsymbol{\varphi}}_p = \widehat{\boldsymbol{\varphi}}_p^{(n)}$. In this way we obtain

$$\hat{\boldsymbol{\varphi}}_p = \hat{\boldsymbol{\Gamma}}_p^{-1} \hat{\boldsymbol{\gamma}}_p \qquad \hat{\sigma}^2 = \hat{\gamma}(0) - \hat{\boldsymbol{\gamma}}_p' \hat{\boldsymbol{\Gamma}}_p^{-1} \hat{\boldsymbol{\gamma}}_p. \tag{8.3}$$

Definition 19 *Estimators (8.3) are called the Yule-Walker estimators of AR(p) parameters.*

In problem 8.1 we will show that the polynomial $\hat{\varphi}(z) = 1 - \hat{\varphi}_1 z - \cdots - \varphi_p z^p \neq 0$ for $|z| \leqslant 1$ thus AR(p) process W_t satisfying $\hat{\varphi}(B) W_t = Z_t$, where (Z_t) is WN$(0, \hat{\sigma}^2)$, is causal. It follows from causality that its autocovariance $\gamma_W(\cdot)$ satisfies

$$\gamma_W(h) - \sum_{i+1}^{p} \hat{\varphi}_i \gamma_W(h-i) = \sigma^2 I\{h = 0\}$$

for $h = 0, 1, \ldots, p$. Consider the $p + 1$ equations

$$\gamma(h) - \sum_{i+1}^{p} \hat{\varphi}_i \gamma(h-i) = \sigma^2 I\{h = 0\}$$

with $\gamma(0), \ldots \gamma(p)$ unknown. It follows from the above considerations that the solution are empirical autocovariances $\hat{\gamma}(0), \ldots, \hat{\gamma}(p)$. As the solution is unique we obtain that $\gamma_W(h) = \hat{\gamma}(h)$ for $h = 0, \ldots, p$, thus we have constructed AR(p) with theoretical autocovariances $\gamma(0), \ldots \gamma(p)$ coinciding with empirical autocovariances $\hat{\gamma}(0), \ldots \hat{\gamma}(p)$. Note that the above reasoning yield that for any nondegenerate $\boldsymbol{\Gamma}_{p+1}$ we can construct AR(p) process having autocovariances $\gamma(0), \ldots \gamma(p)$. In problem 8.3 to this chapter we will check that if $\gamma(0) > 0$ and $\gamma(h) \to 0$ when $h \to \infty$ then $\boldsymbol{\Gamma}_n$ is positive definite for any $n \in \mathbb{N}$. From this it follows that $\hat{\boldsymbol{\Gamma}}_p^{-1}$ exists under the only condition that empirical variance $\hat{\gamma}(0)$ is positive which happens always when there are at least two different observations among

$X_1, \ldots, X_n$. Indeed, note that if we additionally define $\hat{\gamma}(h) = 0$ for $|h| > n$ then $\hat{\gamma}(\cdot)$ is non-negative definite. Thus it follows exactly as in the proof that $\hat{\mathbf{\Gamma}}_n$ is non-negative definite by defining $Y_i = 0$ for $i > n$, there and noting that the proof holds for any $k \in \mathbb{N}$ replacing n. As $\hat{\gamma}(\cdot)$ is non-negative definite there exists stationary sequence $(\tilde{X}_t)$ such that its covariance coincides with $\hat{\gamma}(\cdot)$ and we apply aforementioned result to this sequence. As $\hat{\gamma}(h) = 0$ for $|h| > n$ both assumptions of the result are satisfied. Thus $\hat{\mathbf{\Gamma}}_n^{-1}$ exists and in particular $\hat{\mathbf{\Gamma}}_p^{-1}$ exists for $p \leqslant n$.

It is sometimes more convenient to express the Yule-Walker estimators in terms of sample correlations instead of sample covariances. Remembering that matrix of empirical correlations $\widehat{\mathbf{R}}_p = \widehat{\mathbf{\Gamma}}_p/\hat{\gamma}(0)$ and correlation vector $\hat{\boldsymbol{\rho}}_p = \hat{\boldsymbol{\gamma}}_p/\hat{\gamma}(0)$, by dividing both sides of (8.3) by $\hat{\gamma}(0)$ we obtain

$$\widehat{\boldsymbol{\varphi}}_p = \widehat{\mathbf{R}}_p^{-1}\widehat{\boldsymbol{\rho}}_p$$

$$\widehat{\sigma}^2 = \widehat{\gamma}(0)(1 - \widehat{\boldsymbol{\rho}}_p'\mathbf{R}_p^{-1}\widehat{\boldsymbol{\rho}}_p).$$

Theorem 8.1.2 *(Asymptotic distribution of Y-W estimators for AR(p)) Let be $(X_t)_{t\in\mathbb{Z}}$ be causal AR(p) with innovations $(\varepsilon_t)_{t\in\mathbb{Z}}$ being strong WN$(0, \sigma^2)$. Then when $n \to \infty$*

$$n^{1/2}(\widehat{\boldsymbol{\varphi}}_p^{(n)} - \boldsymbol{\varphi}_p) \xrightarrow{\mathcal{D}} N(0, \sigma^2\mathbf{\Gamma}_p^{-1}),$$

where $\widehat{\boldsymbol{\varphi}}_p^{(n)} = (\widehat{\varphi}_{p1}^{(n)}, \ldots, \widehat{\varphi}_{pp}^{(n)})'$ are Y-W estimators.

Proof (outline). Assume that (X_t) is a mean-zero process. Let $\bar{\boldsymbol{\varphi}} = (X'X)^{-1}X'Y$, where $Y = (X_1, \ldots, X_n)'$, $X_i = 0$ for $i \leqslant 0$ and

$$X = \begin{pmatrix} X_0 & \ldots & X_{1-p} \\ X_1 & \ldots & X_{2-p} \\ X_{n-1} & \ldots & X_{n-p} \end{pmatrix}$$

is the matrix of predictors' values $(X_{t-1}, \ldots, X_{t-p})'$. Note that $\bar{\boldsymbol{\varphi}}$ differs from $\widehat{\boldsymbol{\varphi}}$ in that $\hat{\boldsymbol{\Gamma}}_p$ is replaced by $X'X$ and $\hat{\boldsymbol{\gamma}}_p$ by $X'Y$. The first step is to show that the difference between $\bar{\boldsymbol{\varphi}}$ and $\widehat{\boldsymbol{\varphi}}$ is $o_P(n^{-1/2})$ meaning that $n^{1/2}(\widehat{\boldsymbol{\varphi}} - \bar{\boldsymbol{\varphi}}) = o_P(1)$. This is essentially due to the form of $\widehat{\boldsymbol{\varphi}}$ and the observation that under the assumptions of the theorem we have $X'X/n \to \mathbf{\Gamma}_p$ and $X'Y/n \to \boldsymbol{\gamma}_p$ in probability. Thus it is enough to prove the result for $\bar{\boldsymbol{\varphi}}$. However,

$$\begin{aligned} n^{1/2}(\bar{\boldsymbol{\varphi}} - \boldsymbol{\varphi}) &= n^{1/2}((X'X)^{-1}X'(X\boldsymbol{\varphi} + \boldsymbol{\varepsilon}) - \boldsymbol{\varphi}) \\ &= n^{1/2}(X'X)^{-1}X'\boldsymbol{\varepsilon} = (X'X/n)^{-1}n^{1/2}X'\boldsymbol{\varepsilon}. \end{aligned}$$

Moreover, observe that with $\mathbf{X}_t = (X_t, \ldots, X_{t-p+1})'$ we have

$$n^{1/2}X'\boldsymbol{\varepsilon} = n^{-1/2}\sum_{t=1}^{n}\varepsilon_t\mathbf{X}_{t-1} \xrightarrow{\mathcal{D}} N(0,\sigma^2\boldsymbol{\Gamma}_p).$$

This follows by truncation technique and using CLT for m-dependent random variables after noting that due to causality ε_t and $\mathbf{X}_{t-1}$ are uncorrelated. Thus the covariance matrix of $\varepsilon_t\mathbf{X}_{t-1}$ equals $\sigma^2\boldsymbol{\Gamma}_p$. The last step is to note that $X'X/n \to \boldsymbol{\Gamma}_p$ and use Slutzky's lemma.

The main problem with using this result in practice is that p is almost always unknown. What happens if AR(m) process is fitted when $m > p$? As AR(p) time series is also AR(m) this should lead to a reasonable procedure. It turns out that

Theorem 8.1.3 *If $(X_t)_{t\in\mathbb{Z}}$ is causal AR(p) with innovations $(\varepsilon_t)_{t\in\mathbb{Z}}$ being strong white noise WN$(0,\sigma^2)$ and $\widehat{\boldsymbol{\varphi}}_m^{(n)} = (\hat{\varphi}_{m1}^{(n)},\dots,\hat{\varphi}_{mm}^{(n)})' = \widehat{\mathbf{R}}_m^{-1}\widehat{\boldsymbol{\rho}}_m$, $m \geqslant p$, then*

$$n^{1/2}(\widehat{\boldsymbol{\varphi}}_m^{(n)} - \boldsymbol{\varphi}) \xrightarrow{\mathcal{D}} N(0,\sigma^2\boldsymbol{\Gamma}_m^{-1})$$

when $n\to\infty$.
In particular for $m > p$

$$n^{1/2}\widehat{\boldsymbol{\varphi}}_{mm}^{(n)} \xrightarrow{\mathcal{D}} N(0,1)$$

Proof. Only the last part of the theorem needs proving. Recall that we proved that prediction error $\sigma_i^2 = |\boldsymbol{\Gamma}_{i+1}|/|\boldsymbol{\Gamma}_i|$, and whence $|\boldsymbol{\Gamma}_m| = \sigma_0^2\sigma_1^2\cdots\sigma_{m-1}^2$. In particular, for AR$(p)$ and $m \geqslant p$ it follows that $|\boldsymbol{\Gamma}_m| = |\boldsymbol{\Gamma}_p|\sigma^{2(m-p)}$ and thus $(\boldsymbol{\Gamma}_m^{-1})_{mm} = \sigma^{-2}$. The last equality is valid in view of the fact that $(\boldsymbol{\Gamma}_m^{-1})_{mm}$ is equal to the ratio of minor M_{mm} obtained when mth row and mth column is removed from $\boldsymbol{\Gamma}_m$ and determinant $|\boldsymbol{\Gamma}_m|$. However, $M_{mm} = |\boldsymbol{\Gamma}_{m-1}|$ and thus $(\boldsymbol{\Gamma}_m^{-1})_{mm} = |\boldsymbol{\Gamma}_{m-1}|/|\boldsymbol{\Gamma}_m| = \sigma^{-2}$.

We note that fitting AR(m) does not create any new difficulties. Namely, needed coefficients $\hat{\boldsymbol{\varphi}}_m$ are coefficients of the best prediction for the process with covariance matrix $\widehat{\boldsymbol{\Gamma}}_m$ and covriance vector $\widehat{\gamma}_m$. As $\widehat{\boldsymbol{\Gamma}}_{m+1}$ is positive definite, such process exists, moreover, it is causal (see problem 8.1). Thus, in particular, the vector $\hat{\varphi}_m$ may be determined from the Durbin-Levinson equations.
Moreover, from the theorem above it follows the form of confidence region for vector of coefficients. Namely, recalling that $\sigma^2 = \sigma_p^2$ for AR(p) time series and plugging in $\hat{\sigma}_p^2$ instead of σ^2 and $\widehat{\boldsymbol{\Gamma}}_p$ instead of $\boldsymbol{\Gamma}_p$ we obtain that the confidence region with approximate level of confidence $1-\alpha$ equals

$$\{\varphi:\ n(\hat{\varphi}_p-\varphi)'\widehat{\boldsymbol{\Gamma}}_p(\hat{\varphi}_p-\varphi) \leqslant \hat{\sigma}_p^2\chi_{p,1-\alpha}^2\},$$

where $\chi_{p,1-\alpha}^2$ is quantile of order $1-\alpha$ of χ^2 distribution with p degrees of freedom.

Moreover, the last result is used to justify the following heuristic method of choosing the order of autoregressive process.
Namely, we fit AR(m) with large m

$$\widehat{\boldsymbol{\varphi}}_m = \widehat{\mathbf{R}}_m^{-1}\widehat{\boldsymbol{\rho}}_m, \qquad \widehat{\boldsymbol{\varphi}}_m^{(n)} = (\widehat{\varphi}_{m1}^{(n)}, \ldots, \widehat{\varphi}_{mm}^{(n)})'$$

If $m > p$ then in view of Theorem 8.1.3 $\widehat{\varphi}_{mm}^{(n)} \sim N(0, 1/n)$ approximately for large n, where $\widehat{\varphi}_{mm}^{(n)} = \widehat{\alpha}_m$ sample partial correlation coefficient. Thus as the order of the model we may choose the largest $p \in N$ such that $|\widehat{\alpha}_p| > z_{1-\alpha/2}/\sqrt{n}$.

8.1.2 Burg's estimators

We know that estimation of autoregressive coefficients of AR(p) process is equivalent to estimation of coefficients $\boldsymbol{\varphi} = (\varphi_{p1}, \ldots, \varphi_{pp})'$ of the best linear predictor of X_t based on $X_{t-1}, \ldots, X_{t-p}$. We first derive relation between prediction based on the past and the future observations which is essentially restatement of the relation used in the Durbin-Levinson algorithm.
Consider (backward) predictor $\widehat{X}^b_{n+1+i-t}$ of $X_{n+1+i-t}, t \geqslant i$ based on i past observations and the corresponding residuals $u_i(t)$:

$$u_i(t) = X_{n+1+i-t} - \widehat{X}^b_{n+1+i-t}, \qquad t = i+1, \ldots, n.$$

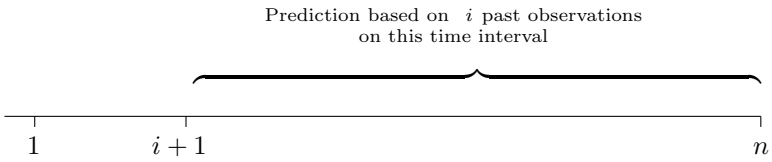


Fig. 8.1. Prediction based on the past observations

Now consider also (forward) predictor $\widehat{X}^f_{n+1+i-t}$ of X_{n+1-t}, $t \geqslant i$ based on i future observations and its residuals $v_i(t)$:

$$v_i(t) = X_{n+1-t} - \widehat{X}^f_{n+1-t}, \qquad t = i+1, \ldots, n.$$

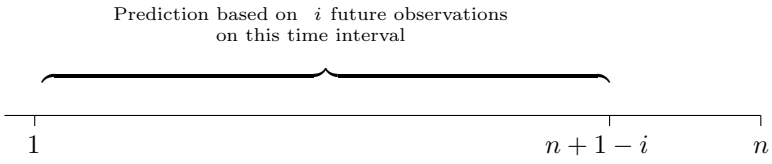


Fig. 8.2. Prediction based on future observations

In particular we thus have $u_0(t) = X_{n+1-t}$ and $v_0(t) = X_{n+1-t}$. The following equalities relating $u_i(t)$ and $v_i(t)$ hold

$$\begin{aligned} u_i(t) &= u_{i-1}(t) - \varphi_{ii} v_{i-1}(t) \\ v_i(t) &= v_{i-1}(t) - \varphi_{ii} u_{i-1}(t). \end{aligned} \tag{8.4}$$

Indeed, using definitions of the residuals we rewrite the first equality in (8.4) as

$$\begin{aligned} &X_{n+1+i-t} - \sum_{k=1}^{i} \varphi_{ik} X_{n+1+i-t-k} \\ &= X_{n+1+i-t} - \sum_{k=1}^{i-1} \varphi_{i-1,k} X_{n+i-t-k} - \varphi_{ii}(X_{n+1-t} - \sum_{k=1}^{i-1} \varphi_{i-1,k} X_{n+1-t+k}). \end{aligned}$$

We check directly that the coefficients corresponding to X_{n+1-t} for $k = i$ coincide, whereas for $k = 1, \ldots, i-1$ comparison of coefficients corresponding to $X_{n+1+i-t-k}$ on both sides yields $\varphi_{ik} = \varphi_{i-1,k} - \varphi_{ii}\varphi_{i-1,i-k}$, what follows from the Durbin-Levinson algorithm (3.19). The second equlity is checked analogously. Burg estimators $\widetilde{\varphi}_{11}, \ldots, \widetilde{\varphi}_{pp}$ are calculated sequentially. First we define

$$\widetilde{\varphi}_{11} = \arg\min \sigma_1^2(\varphi_{11}),$$

where

$$\sigma_1^2 = \frac{1}{2(n-1)} \sum_{t=2}^{n} \{u_1^2(t) + v_1^2(t)\}.$$

Note that from (8.4) σ_1^2 depends only on φ_{11} and the known values $v_0(t)$ i $u_0(t)$. Then using (8.4) we calculate $\hat{v}_1(t)$ i $\hat{u}_1(t)$. Analogously

$$\widetilde{\varphi}_{22} = \arg\min \sigma_2^2,$$

where

$$\sigma_2^2 = \frac{1}{2(n-2)} \sum_{t=3}^{n} \{\hat{u}_2^2(t) + \hat{v}_2^2(t)\}$$

and so on. We obtain $\widetilde{\varphi}_{11}, \ldots, \widetilde{\varphi}_{pp}$ as the result. Estimators φ_{pj} for $1 \leqslant j \leqslant p-1$ are obtained from the Durbin–Levinson algorithm after plugging in $\widetilde{\varphi}_{kk}$ for φ_{kk}. Variance σ^2 is estimated from the formula for σ_p^2 with φ_{pp} replaced by its estimator $\widetilde{\varphi}_{pp}$.

We have the following result on asymptotic distribution of Burg's estimators. As in Theorem 8.1.2 we assume that p is known.

Theorem 8.1.4 *Assume that the assumptions of the previous theorem hold and let* $\widetilde{\boldsymbol{\varphi}}^{(n)} = (\widetilde{\varphi}_{p1}^{(n)}, \ldots, \widetilde{\varphi}_{pp}^{(n)})'$ *be Burg's estimator based on* n *observations. Then*

$$n^{1/2}(\widetilde{\boldsymbol{\varphi}}^{(n)} - \boldsymbol{\varphi}) \xrightarrow{\mathcal{D}} N_p(0, \sigma^2 \boldsymbol{\Gamma}_p^{-1}).$$

We note that the asymptotic distribution coincides with that for Y-W estimator.

8.2 Preliminary estimation of parameters for ARMA(p,q) time series

We describe now two methods which can be used for estimation of ARMA(p,q) time series. Usually, they are used only to determine a starting point for the ML estimator which is calculated iteratively via innovation algorithm.

8.2.1 Yule-Walker estimators for ARMA(p,q) time series

We write the Yule-Walker equations in more general situation when the order of the moving average part is larger than 0. If the process is causal with representation $X_t = \sum_{i=0}^{\infty} \psi_i \varepsilon_{t-i}$ we have, letting with $\theta_0 = 1$

$$\begin{aligned}\operatorname{Cov}(X_{t-k}, \sum_{i=0}^{q} \theta_i \varepsilon_{t-i}) &= \operatorname{Cov}(\sum_{i=0}^{\infty} \psi_i \varepsilon_{t-k-i}, \sum_{i=0}^{q} \theta_i \varepsilon_{t-i}) \\ &= \operatorname{Cov}(\sum_{i=k}^{\infty} \psi_{i-k} \varepsilon_{t-i}, \sum_{i=0}^{q} \theta_i \varepsilon_{t-i}) = \sigma^2 \sum_{i=k}^{q} \theta_j \psi_{i-k}.\end{aligned}$$

Calculating the covariance of X_{t-k} with both sides of structural equation we thus obtain

$$\gamma(k) - \sum_{i=1}^{p} \varphi_i \gamma(k-i) = \sigma^2 \sum_{j=k}^{q} \theta_j \psi_{j-k}.$$

We need $p+q+1$ equations and thus usually the equations above are considered for $0 \leqslant k \leqslant p+q$. Then we have $p+q+1$ equations for $p+q+1$ parameters $\psi_1, \ldots, \psi_q, \theta_1, \ldots, \theta_q, \sigma^2$. The Yule-Walker estimators are obtained as a solution to these equations when $\gamma(i)$ are replaced by their sample counterparts. The problem is that the solution may be not unique or may not even exist. Moreover, such estimators are usually much more variable than ML estimators described later on. For example, if we write down the equations for MA(1) time series. As in this case we obviously have $\psi_0 = 1$ i $\psi_1 = \theta_1$, we get

$$\begin{aligned}\gamma(0) &= \sigma^2(1 + \theta_1 \psi_1) = \sigma^2(1 + \theta_1^2) \\ \gamma(1) &= \sigma^2 \theta_1,\end{aligned}$$

which leads to the equations $\hat{\gamma}(0) = \hat{\sigma}^2(1 + \hat{\theta}_1^2)$, $\hat{\gamma}(1) = \hat{\sigma}^2 \hat{\theta}_1$. It is easy to check that solution to these equations exists only when $\hat{\rho}(1) = \hat{\gamma}(1)/\hat{\gamma}(0) \leqslant 0.5$.

8.2.2 Preliminary estimation using the Durbin-Levinson algorithm

In order to obtain preliminary estimators of parameters for ARMA(p, q) we can use an alternative solution. Namely, we recall from Chapter 4 that when ARMA(p, q) process is causal and has representation $X_t = \sum_{i=0}^{\infty} \psi_i \varepsilon_{t-i}$ then $\boldsymbol{\varphi}$ and $\boldsymbol{\theta}$ satisfy equalities $\psi_0 = 1$ and

$$\psi_j = \theta_j + \sum_{0<i<j} \varphi_i \psi_{j-i} \quad j = 1, 2, \ldots,$$

where $\theta_j = 0$ for $j > q$ i $\varphi_j = 0$ for $j > p$. In order to obtain $p + q$ equations to solve for φ and θ we consider the above equations for $j = 1, \ldots, p+q$ and replace ψ_j for $j = 1, \ldots, p + q$ by their estimators. How to obtain estimators of ψ_j ? Note that for large m $X_t \approx \sum_{i=0}^{m} \psi_i \varepsilon_{t-i}$, and thus estimators $\hat{\theta}_{m,1}, \ldots, \hat{\theta}_{m,p+q}$ of coefficients of MA(m) model fitted by means of innovation algorithm may serve as estimators of ψ_i. It is known that when $m \to \infty$ when k is fixed and under appropriate conditions on causal and invertible process $(\hat{\theta}_{m,1}, \ldots, \hat{\theta}_{m,k})$ is consistent estimator of $(\psi_1, \ldots, \psi_k)$ (Brockwell and Davis (1991)). Thus the discussed procedure should lead to reasonable approximation of unknown parameters. Estimators of φ_i, θ_j are obtained from the equations

$$\hat{\theta}_{m,j} = \theta_j + \sum_{i=1}^{\min(j,p)} \varphi_i \hat{\theta}_{m,j-i} \quad j = 1, 2, \ldots, p + q,$$

obtained when innovation estimators $\hat{\theta}_{m,j}$ are plugged in the previous equation in the place of ψ_j. Then considering first p last equations for which $\theta_j = 0$ and calculating estimators of $\varphi_1, \ldots, \varphi_p$ and then plugging them into the first q equations and calculating estimators of $\theta_1, \ldots, \theta_q$ we obtain desired estimators. As an estimator of σ^2 we consider its approximation from innovation algorithm.

8.2.3 The Hannan–Rissanen method

The starting point of the Hannan–Rissanen method is noting that structural equation of AR(p) process

$$X_t = \varphi_1 X_{t-1} + \cdots + \varphi_p X_{t-p} + Z_t \tag{8.5}$$

is nothing else than linear regression with vector of predictors equal $\mathbf{X} = (X_{t-1}, \ldots, X_{t-p})'$ and response X_t. In the case of ARMA(p, q) time series the situation is analogous:

$$X_t = \varphi_1 X_{t-1} + \cdots + \varphi_p X_{t-p} + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q} + Z_t.$$

The only difference is that vector of regressors $(X_{t-1},\dots,X_{t-p},Z_{t-1},\dots,Z_{t-q})$ is not wholly observable. The idea is to replace $Z_{t-1},\dots,Z_{t-q}$ with residuals from an appropriate autoregression model. The algorithm consists of the following steps. First, take $m>\max(p,q)$ and fit autoregressive model AR(m) with resulting estimators of coefficients equal $\widehat{\varphi}_{m,1},\dots,\widehat{\varphi}_{m,m}$. Calculate residuals

$$\widehat{Z}_t = X_t - \widehat{\varphi}_{m,1}X_{t-1} - \cdots - \widehat{\varphi}_{m,m}X_{t-m}, \quad t=m+1,\dots,n.$$

Secondly, estimate $\boldsymbol{\beta}=(\varphi_1,\dots,\varphi_p,\theta_1,\dots,\theta_q)$ in the regression model with response X_t and predictors $X_{t-1},\dots,X_{t-p},\widehat{Z}_{t-1},\dots,\widehat{Z}_{t-q}$. The main problem here is to understand why residuals $\widehat{Z}_{t-1},\dots,\widehat{Z}_{t-q}$ may serve as proxies for $Z_{t-1},\dots,Z_{t-q}$. To justify, note that if (X_t) is invertible then $Z_t=\sum_{i=0}^{\infty}\eta_i X_{t-i}$ and thus for large m $Z_t\approx\sum_{i=0}^{m}\eta_i X_{t-i}$ or, in other words, $(\eta_0=1)$ $X_t\approx Z_t-\sum_{i=1}^{m}\eta_i X_{t-i}$. Thus approximating X_t with process AR(m) we obtain $\varphi_{m,i}\approx-\eta_i$ and it follows that $\hat{Z}_t$ is heuristically sound approximation of Z_t. Finally, we let $\widehat{\sigma}^2=S(\widehat{\beta})/(n-m)$.

8.3 The Maximum likelihood estimators for Gaussian ARMA(p,q) time series

Innovation algorithm is crucial in writing down the likelihood function for Gaussian ARMA(p,q) time series. We recall that the innovations representation is representation of $\widehat{X}_{n+1}=P_{sp\{X_1,\dots,X_n\}}X_{n+1}$ in the form

$$\widehat{X}_{n+1}=\sum_{j=1}^{n}\theta_{nj}(X_{n+1-j}-\widehat{X}_{n+1-j})$$

where $\widehat{X}_t=P_{sp\{X_1,\dots,X_{t-1}\}}X_t$ denotes the best linear predictor based on all available observations up to time $t-1$. Recall also that the summands in the above decomposition are orthogonal: $sp(X_l,l\leqslant k)\ni X_k-\widehat{X}_k\perp X_s-\widehat{X}_s$ for $s>k$. Moreover, decomposition

$$X_n=(X_n-\widehat{X}_n)+\widehat{X}_n$$

consists of two orthogonal components. The two last equalities can be written in the matrix form as follows (cf 3.35)

$$\begin{pmatrix}X_1\\X_2\\\vdots\\X_n\end{pmatrix}=\underbrace{\begin{pmatrix}1&\dots&0&0\\\theta_{11}&\dots&0&0\\\vdots&&\vdots&\vdots\\\theta_{n-1,n-1}&\dots&\theta_{n-1,1}&1\end{pmatrix}}_{\boldsymbol{\Theta}_n}\begin{pmatrix}X_1-\widehat{X}_1\\X_2-\widehat{X}_2\\\vdots\\X_n-\widehat{X}_n\end{pmatrix} \tag{8.6}$$

and using orthogonality of innovations we have that

$$\mathbf{D}_n := \mathbf{\Sigma}_{X_n - \widehat{X}_n} = \mathrm{diag}(v_0, \ldots, v_{n-1}).$$

In view of (8.6)

$$\mathbf{\Gamma}_n = \mathbf{\Theta}_n \mathbf{D}_n \mathbf{\Theta}_n',$$

where $\mathbf{\Theta}_n$ is lower triangular matrix with unit diagonal defined above. Thus determinant $|\mathbf{\Theta}_n| = 1$ and we have that $|\mathbf{\Gamma}_n| = v_0 \cdots v_{n-1}$. Moreover, we have $\mathbf{\Gamma}_n^{-1} = \mathbf{\Theta}'^{-1}_n \mathbf{D}_n^{-1} \mathbf{\Theta}_n^{-1}$. Define $\mathbf{X}_n = (X_1, \ldots, X_n)'$ and $\widehat{\mathbf{X}}_n = (\hat{X}_1, \ldots, \hat{X}_n)'$. Then in view of $\mathbf{X}_n = \mathbf{\Theta}_n(\mathbf{X}_n - \widehat{\mathbf{X}}_n)$

$$\mathbf{X}_n' \mathbf{\Gamma}_n^{-1} \mathbf{X}_n = (\mathbf{X}_n - \widehat{\mathbf{X}}_n)' \mathbf{D}_n^{-1} (\mathbf{X}_n - \widehat{\mathbf{X}}_n) = \sum_{i=1}^{n} (X_i - \widehat{X}_i)^2 / v_{i-1}. \tag{8.7}$$

Assume now that (X_t) is a zero-mean causal ARMA(p,q) process with Gaussian innovations. From the causal representation then it follows that (X_t) is Gaussian and thus $X_1, \ldots, X_n$ is zero-mean Gaussian vector. Whence it distribution is $N(0, \mathbf{\Gamma}_n)$ and the likelihood in view of (8.7) is

$$L(\boldsymbol{\varphi}, \boldsymbol{\theta}, \sigma^2, X_1, \ldots, X_n) = (2\pi)^{-1/2} (\det \mathbf{\Gamma}_n)^{-1/2} \exp\{-\frac{1}{2} \mathbf{X}_n' \mathbf{\Gamma}_n^{-1} \mathbf{X}_n\} \tag{8.8}$$

$$= \frac{1}{(2\pi)^{n/2} (v_0 \cdots v_{n-1})^{1/2}} \exp\big(-\frac{1}{2} \sum_{i=1}^{n} (X_i - \widehat{X}_i)^2 / v_{i-1}\big) \tag{8.9}$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2} (r_0 \cdots r_{n-1})^{1/2}} \exp\big(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \widehat{X}_i)^2 / r_{i-1}\big), \tag{8.10}$$

where $r_i = v_i/\sigma^2, \quad i = 0, \ldots, n-1$. We recall from Section 4.2.2 that for ARMA(p,q) time series $\widehat{X}_i$ i r_i do not depend on σ^2.

Definition 20 *Estimators satisfying*

$$(\widehat{\boldsymbol{\varphi}}, \widehat{\boldsymbol{\theta}}, \hat{\sigma}^2) = \mathrm{argmax}_{\boldsymbol{\varphi}, \boldsymbol{\theta}, \sigma^2} L(\boldsymbol{\varphi}, \boldsymbol{\theta}, \sigma^2, X_1, \ldots, X_n)$$

are called maximum likelihood estimators of ARMA(p,q) parameters.

Let $\mathcal{L}(\boldsymbol{\varphi}, \boldsymbol{\theta}, \sigma^2) = \log L(\boldsymbol{\varphi}, \boldsymbol{\theta}, \sigma^2)$ and note that differentiating $\mathcal{L}$ with respect to σ^2 we obtain that $\hat{\sigma}^2$ satisfies

$$-\frac{n}{2\hat{\sigma}^2} + \frac{1}{2(\hat{\sigma}^2)^2} \sum_{i=1}^{n} \frac{(X_i - \widehat{X}_i)^2}{r_{i-1}} = 0$$

and thus

$$\widehat{\sigma}^2(\boldsymbol{\varphi},\boldsymbol{\theta}) = \underbrace{\frac{1}{n}\sum_{i=1}^{n}\frac{(X_i-\widehat{X}_i)^2}{r_{i-1}}}_{S(\boldsymbol{\varphi},\boldsymbol{\theta})}. \tag{8.11}$$

Let $\mathcal{L}(\boldsymbol{\varphi},\boldsymbol{\theta}) = \mathcal{L}(\boldsymbol{\varphi},\boldsymbol{\theta},\hat{\sigma}^2)$ be a reduced log-likelihood and define $l(\boldsymbol{\varphi},\boldsymbol{\theta}) = -2\log\mathcal{L}(\boldsymbol{\varphi},\boldsymbol{\theta})/n$. Then

$$l(\boldsymbol{\varphi},\boldsymbol{\theta}) = 2\pi + 1 + \log(n^{-1}S(\boldsymbol{\varphi},\boldsymbol{\theta})) + n^{-1}\sum_{j=1}^{n}\log r_{j-1} \tag{8.12}$$

Note that the maximum likelihood estimator of $\boldsymbol{\varphi},\boldsymbol{\theta}$ is a stationary point of $l(\boldsymbol{\varphi},\boldsymbol{\theta})$. Solution is found by iterative methods starting from some preliminary estimates of $\boldsymbol{\varphi},\boldsymbol{\theta}$ estimating σ^2 using (8.11) and then using iterative search to find minimum of $l(\boldsymbol{\varphi},\boldsymbol{\theta})$. Note that predictors $\hat{X}_i$ and r_i depend only on $\boldsymbol{\varphi},\boldsymbol{\theta}$ and can be found using the innovation algorithm. It is important to ensure that the initial estimates correspond to causal time series as otherwise iterative process may not converge. Usually, for the AR(p) process Yule-Walker or Burg's estimators are considered as initial estimators, for the general ARMA(p,q) we use the Hannan-Rissanen or estimators described in Section 8.2.2.

8.3.1 Weighted least squares estimators

Note that after omitting constant term and $R_n := n^{-1}\sum_{j=1}^{n}\log r_{j-1}$ in (8.12) we are left with increasing function of

$$\widehat{\sigma}^2 = \underbrace{\frac{1}{n}\sum_{i=1}^{n}\frac{(X_i-\widehat{X}_i)^2}{r_{i-1}}}_{S(\boldsymbol{\varphi},\boldsymbol{\theta})} \tag{8.13}$$

which is weighted least squares criterion, a standard criterion used in linear regression analysis which accounts for heteroscedasticity of errors. Estimators of $\boldsymbol{\varphi}$ and $\boldsymbol{\theta}$ obtained by minimization of (8.13) are called the weighted least squares estimators. Corresponding estimator of σ^2 is defined as

$$\hat{\sigma}^2_{WLS} = \frac{S(\hat{\boldsymbol{\varphi}},\hat{\boldsymbol{\theta}})}{n-p-q}.$$

Justification of the method follows from the fact that for invertible process $n^{-1}\sum_{j=1}^{n}\log r_{j-1}$ tend to 0 in probability since $v_i \to \sigma^2$. Recall namely that for $\sigma^2 = \mathrm{Var}(\varepsilon_t)$ we have

$$\sigma^2 = \| X_t - P_{H_{t-1}}X_t \|^2 .$$

Indeed, for an invertible representation of ε_t

$$\varepsilon_t = \sum_{i=0}^{\infty}\pi_i X_{t-1}$$

it follows by comparing coefficients corresponding to constants in $\pi(z) = \varphi(z)/\theta(z)$ that $\pi_0 = 1$. Thus

$$X_t = \varepsilon_t - \sum_{i=1}^{\infty} \pi_i X_{t-1}$$

and $P_{H_{t-1}} X_t = \sum_{i=1}^{\infty} \pi_i X_{t-1}$. Thus for invertible ARMA(p,q) time series innovation variance coincides with the error of prediction based on the whole past. Since $v_i \to \sigma^2$ it follows that $(v_0 v_1 \cdots v_{n-1})^{1/n} \longrightarrow \sigma^2$ and thus

$$n^{-1} \sum_{j=1}^{n} \log r_{j-1} \longrightarrow 0.$$

Intuitively, if process is invertible weighted least squares estimators should behave similarly to maximum likelihood estimators.

8.3.2 Likelihood function in the spectral domain

The considerations of the previous section lead to the simple approximation of the log-likelihood function in the spectral domain. Namely, consider zero mean one-sided invertible linear process $X_t = \sum_{i=0}^{\infty} a_i \varepsilon_{t-i}$. We have

$$f_X(\lambda) = \frac{\sigma_\varepsilon^2}{2\pi} |A(e^{-i\lambda})|^2 =: \frac{\sigma_\varepsilon^2}{2\pi} h(\lambda)$$

and then $(-2/n)\mathcal{L}(\mathbf{x})$ with $\mathbf{x} = (x_1, \ldots, x_n)'$ equals, up to the constant term,

$$\frac{1}{n} \log |\mathbf{\Gamma}_{n,f}| + \frac{1}{n} \mathbf{x}' \mathbf{\Gamma}_{n,f}^{-1} \mathbf{x} = \log \sigma_\varepsilon^2 + \frac{1}{n} \log |\mathbf{\Gamma}_{n,h}| + \frac{2\pi}{\sigma_\varepsilon^2} \frac{1}{n} \mathbf{x}' \mathbf{\Gamma}_{n,h}^{-1} \mathbf{x}. \quad (8.14)$$

As previously, we show that $n^{-1} \log |\mathbf{\Gamma}_{n,h}| \to 0$ for invertible linear process and thus (8.14) can be approximated by

$$\log \sigma_\varepsilon^2 + \frac{2\pi}{\sigma_\varepsilon^2} \frac{1}{n} \mathbf{x}' \mathbf{\Gamma}_{n,h}^{-1} \mathbf{x}. \quad (8.15)$$

Note that usually, as in the case of ARMA processes, $h = h(\theta)$ is parametrized by parameter θ which does not depend on σ_ε^2. Approximation (8.15) will be further simplified in Chapter 12 using so-called Whittle's approximation to construct estimator of parameter of long-range dependence.

8.3.3 Asymptotic distribution of estimatiors of parameters for ARMA(p,q) time series

In order to state asymptotic distribution of ARMA(p,q) time series we assume that its innovations (Z_t) are strong WN$(0,\sigma^2)$. Let $\widehat{\boldsymbol{\beta}} = (\widehat{\varphi}_1, \ldots, \widehat{\varphi}_p, \widehat{\theta}_1, \ldots, \widehat{\theta}_q)'$ be a vector of ML estimators of $\boldsymbol{\beta} = (\varphi_1, \ldots, \varphi_p, \theta_1, \ldots, \theta_q)'$.

Theorem 8.3.1 *Consider causal and invertible* $(X_t)_{t\in\mathbb{Z}}$ *proces* $ARMA(p,q)$ *such that* $\varphi(\cdot)$ *i* $\theta(\cdot)$ *do not have common roots. Then*

$$n^{1/2}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \xrightarrow{\mathcal{D}} N(0,V(\boldsymbol{\beta}))$$

where

$$V(\boldsymbol{\beta}) = \sigma^2\, \Sigma^{-1}_{(\mathbf{U}_t',\mathbf{V}_t')'}$$

and

$$\mathbf{U}_t=(U_t,U_{t-1},\ldots,U_{t-p})', \quad \mathbf{V}_t=(V_t,V_{t-1},\ldots,V_{t-q})'$$

satisfy autoregressive structural equations

$$\varphi(B)U_t=Z_t, \quad \theta(B)V_t=Z_t.$$

Note that for $q=0$ asymptotic covariance $V(\boldsymbol{\beta})=\sigma^2\,\Sigma^{-1}_{\mathbf{U}_t}=\sigma^2\boldsymbol{\Gamma}_p^{-1}$ coincides with asymptotic covariance of Y-W estimators. In particular, asymptotic variances of Y-W and ML estimators are the same. This is remarkable, as usually variances of moment estimators are substantially larger than for ML estimators.
Examples. For $p=1$ we obtain that $\widehat{\varphi}$ is $\mathrm{AN}(\varphi, n^{-1}(1-\varphi^2))$. For $p=2$

$$\widehat{\boldsymbol{\varphi}} \text{ is } \mathrm{AN}\Big(\begin{bmatrix}\varphi_1\\ \varphi_2\end{bmatrix}, n^{-1}\begin{bmatrix}1-\varphi_2^2 & -\varphi_1(1+\varphi_2)\\ -\varphi_1(1+\varphi_2) & 1-\varphi_2^2\end{bmatrix}\Big)$$

For $\mathrm{MA}(q)$ process asymptotic covariance of ML estimators is $V(\boldsymbol{\theta})=\sigma^2(\boldsymbol{\Gamma}_q^*)^{-1}$ where $\boldsymbol{\Gamma}_q^*$ is covariance matrix of (V_t) satisfying

$$V_t+\theta_1V_{t-1}+\cdots+\theta_qV_{t-q}=Z_t$$

which is $\mathrm{AR}(q)$ time series with $\boldsymbol{\varphi}=-\boldsymbol{\theta}$. In particular, for $q=2$ asymptotic covariance is

$$n^{-1}\begin{pmatrix}1-\theta_2^2 & \theta_1(1-\theta_2)\\ \theta_1(1-\theta_2) & 1-\theta_2^2\end{pmatrix}$$

8.4 Problems

1. For any weakly stationary time series (X_t) let covariance matrix $\boldsymbol{\Gamma}_p$ be strictly positive definite and $\boldsymbol{\varphi}_p=\boldsymbol{\Gamma}_p^{-1}\boldsymbol{\gamma}_p$ be a vector of projection coefficients. Let $\varphi(z)=1-\varphi_1z-\ldots-\varphi_pz^p$ be an associated autoregressive polynomial and denote by w reciprocal of any root of it i.e. $\varphi(z)=(1-wz)\varphi_1(z)$.
(i) Prove that $w=\mathrm{cor}(Y_t,Y_{t+1})=\rho$, where $Y_t=\varphi_1(B)(X_t)$ by noting that for $\tilde{\varphi}(z)=(1-\rho z)\varphi_1(z)$ we have $||\tilde{\varphi}(B)X_{t+1}||^2\leqslant||\varphi(B)X_{t+1}||^2$.
(ii) Prove that all roots of $\varphi(z)$ lie outside the unit circle. Thus $\mathrm{AR}(p)$ process with $\boldsymbol{\varphi}$ given by Y-W estimator yields causal series.
2. Using innovation algorithm devise a method of simulating n observations from

Gaussian vector $X_1, \ldots, X_n$ with mean 0 and covariance matric $\mathbf{\Gamma}_n$ having at your disposal iid $N(0, 1)$ sequence of length n.
3. State the algorithm for solving Problem 3 using the Cholesky decomposition and indicate a main difference between the two.
4. Show that if covariance function $\gamma(\cdot)$ is such that $\gamma(0) > 0$ and $\gamma(h) \to 0$ when $\to \infty$ then $\mathbf{\Gamma}_n$ is invertible for any n. Hint: reasoning by contradiction note that any X_n with $n > s$ will be linear combination of the first s observations when s is the last index such that $\mathbf{\Gamma}_s$ is invertible.
5. Consider the sample path $X_1, \ldots, X_n$ of causal AR(p) series with Gaussian innovations and $n > p$. Show that (8.8) implies that the likelihood for $X_1, \ldots, X_n$ equals

$$L(\boldsymbol{\varphi}, \sigma^2) = (2\pi\sigma^2)^{-n/2} |\mathbf{W}_p|^{-1/2} \times \exp\Big\{ -\frac{1}{2}\sigma^{-2}\Big[\mathbf{X}_p' \mathbf{W}_p^{-1} \mathbf{X}_p + \sum_{t=p+1}^{n} (X_t - \varphi_1 X_{t-1} - \cdots - \varphi_p X_{t-p})^2\Big]\Big\},$$

where $\mathbf{X}_p = (X_1, \ldots, X_p)'$ and $\mathbf{W}_p = \sigma^{-2}\mathbf{\Gamma}_p$.

9

Modelling using time series

In this chapter we consider various aspects of modelling and diagnostics for time series including model selection for ARMA(p,q) processes, diagnostics of fit for such time series and white noise tests as well as basic models of nonstationary time series and their fitting.

9.1 Model selection of ARMA(p,q) time series

We discussed before two simple heuristic tools which may be used to select order of autoregressive or moving average process. They are based on two properties. The first proved in Theorem 4.2.8 states that:
If $(X_t)_{t\in\mathbb{Z}}$ is weakly stationary nondeterministic process such that $\alpha(i)=0$ for $i>p$ then (X_t) is AR(p). Analogous property related to covariance function is (cf Theorem 4.2.9):
If $\gamma(i)=0$ for $i>q$ then (X_t) is MA(q).
Moreover, the following results hold:
(i) $n^{1/2}\widehat{\alpha}(i)\xrightarrow{\mathcal{D}}N(0,1)$ for $i>p$ if $\widehat{\alpha}(i)$ is the Yule-Walker estimator of $\alpha(i)=0$ for AR(p) process with iid innovations based on a sample path $X_1,\ldots,X_n$.
and
(ii) $n^{1/2}\widehat{\rho}(i)\xrightarrow{\mathcal{D}}N(0,1)$ for an estimator $\widehat{\rho}(i)$ of $\rho(i)=0$ for strong white noise based on a sample path $X_1,\ldots,X_n$.
Property (i) is used in heuristic identification of the order of AR(p) series: we look for the first k such that for all $i>k$, $|\widehat{\alpha}(i)|\leqslant z_{1-\alpha/2}/\sqrt{n}$ and then $p=k$ is considered as possible order of AR process. Similarly, the analogous procedure is used to identify white noise, and sometimes this procedure is extended to identify order of MA process. However, property (ii) holds for white noise only and does not hold for MA processes. As we mentioned in the discussion of Bartlett's theorem for MA processes we can use the test of hypothesis that the underlying process is MA(q) by testing $\rho(q+1)=0$ using the test statistic $\hat{\rho}(q+1)/SE_{\hat{\rho}(q+1)}$, where $SE_{\hat{\rho}(q+1)}$ is the standard error defined in (7.17).
In the case of AR(p) processes another possibility would be to base determination of its order on testing the sequence of hypotheses $H_{0j}:\varphi_j=0$ using as a test statistic $T_j=\hat{\varphi}_{jm}/\hat{\sigma}_{mj}$ obtained from the fitting of AR(m) model. Here, $\hat{\sigma}^2_{mj}$ is the j^{th} diagonal element of the estimated limiting covariance matrix in Theorem 8.1.3. The problems arising here are that on the one hand we need to take large m to ensure that it is larger than unknown p on the other hand we need to

account for multiple testing problem arising in such a case.
We turn now to discussing theoretically justified methods of choosing the order of ARMA(p,q) time series which are based on criteria approach to model selection. Suppose that we would like to fit the model ARMA(p,q) to our data and we have to decide what p and q to choose. We focus on the most frequently applied case of Gaussian ARMA(p,q) model. If we proceed naively by looking at $(p,q) = \arg\min\{-\log L(\widehat{\varphi}_p, \widehat{\theta}_q, \widehat{\sigma})\}$ this will invariably lead to choice of maximal p and q considered. This is due to overfitting as the data have been already used to estimate the parameters and now we would like to use it again to estimate the dimension of the model. Note also that for large p and q we can always approximate a given sample path by sample path of ARMA(p,q) or even AR(p) process. Recall that we have proved in section 8.1.1 that for *any* nondegenerate covariance matrix $\mathbf{\Gamma}_p = (\gamma(i-j))_{1\leqslant i,j\leqslant p+1}$ there exists AR(p) process with such covariances. The natural way to avoid this effect is to incorporate the complexity penalty into the criterion function i.e. the cost of fitting a model containing many parameters. Let us discuss shortly the most popular penalties:
(i) Akaike criterion AIC (Akaike (1970))

$$\mathrm{AIC} = -2\log L(\hat{\boldsymbol{\varphi}}_p, \hat{\boldsymbol{\theta}}_q, \widehat{\sigma}) \ + 2(p+q+1), \tag{9.1}$$

where vector $\hat{\boldsymbol{\varphi}}_p, \hat{\boldsymbol{\theta}}_q, \widehat{\sigma}$ is the vector of ML estimators. Note that $p+q+1$ is the number of parameters of the potential candidate ARMA(p,q).
(ii) Corrected AIC criterion AICC (cf. Hurvich and Tsai (1989)):

$$\mathrm{AICC} = -2\ln L(\hat{\boldsymbol{\varphi}}_p, \hat{\boldsymbol{\theta}}_q, \widehat{\sigma}) + \frac{2(p+q+1)n}{n-p-q-2}. \tag{9.2}$$

(iii) Bayesian Information Criterion BIC (Schwarz (1978))

$$\mathrm{BIC} = -2\log L((\widehat{\boldsymbol{\varphi}}_p, \widehat{\boldsymbol{\theta}}_q, \widehat{\sigma})) + (p+q+1)\log n. \tag{9.3}$$

Akaike Information criterion is an estimator of Kullback-Leibler divergence beteween the fitted model and the true one and AICC is based on bias correction of this estimate. Namely, it is shown in Akaike (1970) that the number of parameters of the model divided by the sample size is approximately the bias of $n^{-1}\sum_{j=1}^{n}\log f(X_j, \hat{\boldsymbol{\varphi}}_p, \hat{\boldsymbol{\theta}}_q, \widehat{\sigma})$ as an estimator of $E_g \log f(X, \boldsymbol{\varphi}_p, \boldsymbol{\theta}_q, \sigma)$, where g is the true model density. Note that -$E_g \log f(X, \boldsymbol{\varphi}_p, \boldsymbol{\theta}_q, \sigma)$ is the part of of Kullback-Leibler divergence $D(g||f)$ depending on f. Thus

$$-E_g \log f(X, \boldsymbol{\varphi}_p, \boldsymbol{\theta}_q, \sigma) + \frac{1}{n}\sum_{j=1}^{n}\log f(X_j, \hat{\boldsymbol{\varphi}}_p, \hat{\boldsymbol{\theta}}_q, \widehat{\sigma}) \approx \frac{p+q+1}{n}.$$

Whence the dimension of the model is added to loglikelihood in order to account for the bias. Optimum model which is chosen by AIC criterion is a trade-off between two terms in (9.1).
In contrast BIC criterion (9.3) approximates aposteriori probability of a model

under consideration given the data. Note that for n such that $\log n > 2$ i.e. $n \geqslant 8$ BIC penalty is larger than AIC penalty and thus size of the model chosen by BIC is not larger than size of the model selected by AIC. It is known (cf Hannan (1997)) that when $\hat{p}_{BIC}$ and $\hat{q}_{BIC}$ are the orders chosen by BIC and the data is sampled from ARMA(p,q) time series with $n \to \infty$, then BIC is consistent in the sense that $P(\hat{p}_{BIC} = p\,, \hat{q}_{BIC} = q) \to 1$.
The method discussed below is used to choose the order of AR(p) process.
FPE (Final Prediction Error) In order to derive the form of the criterion suppose that we have two independent trajectories $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ of causal AR(p) process and we will forecast the observations of the second sample path using parameter estimates based on the first sample path $X_1, \ldots, X_n$. FPE criterion is an approximation of the mean squared error of the forecast.
We have that

$$E(Y_{n+1} - \widehat{\varphi}_1 Y_n - \cdots - \widehat{\varphi}_p Y_{n+1-p})^2 \approx \sigma^2(1 + \frac{p}{n}), \tag{9.4}$$

where $\widehat{\varphi}_1, \ldots, \widehat{\varphi}_p$ are the Yule-Walker or ML estimators of parameters based on $X_1, \ldots, X_n$. In order to justify this approximate equality note that as (Y_t) is causal then in the representation $Y_{n+1} = \varepsilon_{n+1} + \sum_{i=1}^{p} \varphi_i Y_{n+1-i}$ variable ε_{n+1} is uncorrelated with Y_t for $t \leqslant n$. Thus

$$\begin{aligned} &E((Y_{n+1} - \sum_{i=1}^{p} \hat{\varphi}_i Y_{n+1-i})^2 | X_1, \ldots, X_n) \\ &= E(\varepsilon_{n+1} - \sum_{i=1}^{p} (\hat{\varphi}_i - \varphi_i) Y_{n+1-i})^2 | X_1, \ldots X_n) = \sigma^2 + (\hat{\varphi} - \varphi)' \mathbf{\Gamma}_p (\hat{\varphi} - \varphi) \end{aligned}$$

and since $n^{1/2}(\hat{\varphi} - \varphi) \to (0, \sigma^2 \Gamma_p^{-1})$ then using continuous mapping theorem we have that $n(\hat{\varphi} - \varphi)' \mathbf{\Gamma}_p (\hat{\varphi} - \varphi)$ is approximately distributed as $\sigma^2 \chi_p^2$, where χ_p^2 stands for chi square distribution with p degrees of freedom. Calculating expectations of both sides of the above equation and recalling that the expected value of χ^2 with p degrees of freedom is p we obtain justification of (9.4). Final Prediction Error is an estimator of the right hand side of the above equality when we use $\hat{\sigma}^2 = (n/(n-p))\hat{\sigma}^2_{ML}$ as an approximation of σ^2. After plugging $\hat{\sigma}^2$ in (9.4) in we obtain

$$\text{FPE} = \hat{\sigma}^2 \frac{n+p}{n} = \widehat{\sigma}^2_{ML} \frac{n+p}{n-p},$$

where $\widehat{\sigma}^2_{ML}$ is ML estimator of σ^2. The order p of autoregressive model which minimizes FPE defined above is chosen. It can be shown that in the case of AR(p) process FPE criterion is close to AIC. The reason for this is an approximate equality holding for small values of $2p/(n-p)$

$$\log(FPE) = \log(\widehat{\sigma}^2_{ML}) + \log(1 + \frac{2p}{n-p}) \approx \log(\widehat{\sigma}^2_{ML}) + \frac{2p}{n-p},$$

as the first term corresponds to twice loglikelihood divided by n.
There are several other criteria, we mention e.g. CAT criterion introduced by Parzen and Hannanan-Quinn's criterion. AIC and BIC are the most used ones.

9.1.1 Diagnostics of ARMA(p, q) model fit

Let us note that even for fit of AR(p) process using unstandardised residuals is far-fetched as they are heteroscedastic. As a usual tool of checking whether ARMA time series is well fitted empirical counterparts of *standardized* innovations are used. Usually we consider

$$\widehat{e}_t = (X_t - \widehat{X}_t(\widehat{\boldsymbol{\varphi}}, \widehat{\boldsymbol{\theta}}))/\{r_{t-1}(\widehat{\boldsymbol{\varphi}}, \widehat{\boldsymbol{\theta}})\}^{1/2},$$

where $r_t = v_t/\sigma^2$. $\widehat{X}_t(\widehat{\boldsymbol{\varphi}}, \widehat{\boldsymbol{\theta}})$ denotes prediction of X_t for ARMA time series with parameters $(\widehat{\boldsymbol{\varphi}}, \widehat{\boldsymbol{\theta}})$. Obviously, $(\widehat{e}_t)$ are only approximation of white noise

$$e_t = (X_t - \widehat{X}_t(\boldsymbol{\varphi}, \boldsymbol{\theta}))/\{r_{t-1}(\boldsymbol{\varphi}, \boldsymbol{\theta})\}^{1/2}.$$

More specifically, $(\hat{e}_t)$ would have been white noise, provided X_t were generated from ARMA$(\widehat{\boldsymbol{\varphi}}, \widehat{\boldsymbol{\theta}})$ process. Plot of $\hat{e}_t$ against time is the basic diagnostic tool for ARMA processes.
For empirical covariances we can use the following result due to Box and Pierce (1970) which takes the dependence between empirical autocorrelations into account.

Assume that ARMA series is causal and invertible. Consider the product of $\varphi(z)$ and $\theta(z)$ and let

$$\widetilde{\varphi}(z) = \varphi(z)\theta(z) = 1 - \widetilde{\varphi}_1 z - \cdots - \widetilde{\varphi}_{p+q} z^{p+q}.$$

Moreover, define expansion of reciprocal of $\widetilde{\varphi}(z)$ into power series

$$a(z) = (\widetilde{\varphi}(z))^{-1} = \sum_{j=0}^{\infty} a_j z^j.$$

Let (Y_t) be AR$(p+q)$ time series pertaining to polynomial $\widetilde{\varphi}(\cdot)$ with $\sigma^2 = 1$ and $\widetilde{\boldsymbol{\Gamma}}_{p+q}$ covariance matrix of $(Y_1, \ldots, Y_{p+q})$.
Note that

$$\widetilde{\boldsymbol{\Gamma}}_{p+q} = \Big[\sum_{k=0}^{\infty} a_k a_{k+|i-j|}\Big]_{i,j=1}^{p+q},$$

since $a(z)$ defined above yields causal representation of (Y_t). For $h \geqslant p+q$ denote by $\widehat{\boldsymbol{\rho}}_h = (\widehat{\rho}(1), \ldots, \widehat{\rho}(h))'$ a vector of empirical autocorrelations of $\widehat{e}_t$ and

$$\mathbf{T}_h = [a_{i-j}]_{\substack{1 \leqslant i \leqslant h \\ 1 \leqslant j \leqslant p+q}}.$$

Box-Pierce correction is based on the fact that

$$n^{1/2}\widehat{\boldsymbol{\rho}}_h \xrightarrow{\mathcal{D}} N(0, \mathbf{I} - \mathbf{Q}),$$

where $\mathbf{Q} = \mathbf{T}_h \widetilde{\mathbf{\Gamma}}_{p+q}^{-1} \mathbf{T}_h' = [q_{ij}]_{i,j=1}^h$. Whence asymptotic variance of $\widehat{\rho}_r(i)$ equals $n^{-1}(1 - q_{ii}))$, where $q_{ii} = q_{ii}(\boldsymbol{\varphi}, \boldsymbol{\theta})$. For construction of confidence intervals $q_{ii}(\widehat{\boldsymbol{\varphi}}, \widehat{\boldsymbol{\theta}})$ are used.
For AR(1) it is easily checked that $\widetilde{\mathbf{\Gamma}}_{p+q} = (1 - \varphi^2)^{-1}$ and $q_{ii} = q_{ii}(\varphi) = \varphi^{2(i-1)}(1 - \varphi^2)$.
In diagnostics portmanteau tests discussed below for hypothesis that h first correlations are zero are used. It turns out that to account for the fact that we deal with residuals from the fitted model we change the number of degrees of freedom of asymptotic distribution of $n\widehat{\boldsymbol{\rho}}_h' \widehat{\boldsymbol{\rho}}_h$, changes from h to $h - p - q$. For details we refer to Box and Pierce (1970).
Sometimes the following simple approach is useful. Suppose that we fit ARMA process

$$\varphi(B)X_t = \theta(B)Z_t$$

to the data and the residuals r_t indicate that the model does *not* fit. However the analysis of residuals suggests that Z_t might satisfy ARMA structural equation

$$\varphi_2(B)Z_t = \theta_2(B)W_t,$$

where W_t is a white noise. Then the two above equations imply that

$$\varphi_2(B)\varphi(B)X_t = \varphi_2(B)\theta(B)Z_t = \theta(B)\varphi_2(B)Z_t = \theta(B)\theta_2(B)W_t = \theta_2(B)\theta(B)W_t$$

and then we can again fit ARMA process to the original data with polynomials $\varphi_2\varphi$ and $\theta_2\theta$.

9.1.2 Testing white noise hypothesis using empirical correlations

We recall from the previous section that when $(\varepsilon_i)_{i\in\mathbb{Z}}$ is a strong white noise $\mathrm{WN}(0, \sigma^2)$ then $\widehat{\boldsymbol{\rho}}(h) = (\widehat{\rho}(1), \ldots, \widehat{\rho}(h))'$ has approximately $N(0, n^{-1}\ I)$ distribution as $\sqrt{n}\widehat{\boldsymbol{\rho}}(h) \xrightarrow{\mathcal{D}} N(0, I)$. Using the continuous mapping theorem with $T(x_1, \ldots, x_h) = \sum_{i=1}^h x_i^2$ we have

$$T(\sqrt{n}\widehat{\boldsymbol{\rho}}(h)) \xrightarrow{\mathcal{D}} T(N(0, I))$$

or, equivalently,

$$Q = n \parallel \widehat{\boldsymbol{\rho}}(h) \parallel^2 \xrightarrow{\mathcal{D}} \chi_h^2 \tag{9.5}$$

The left hand side of (9.5) is a frequently used test statistic for testing the hypothesis : $H_0 :$ $(\varepsilon_i)_{i\in\mathbb{Z}}$ is white noise.

We fix $h \in \mathbb{N}$ and use (9.5) as a test statistic with critical region $\{Q > \chi^2_{1-\alpha,h}\}$. This is so called portmanteau test (or 'trunk' test, as it 'contains' all tests $\rho(i) \equiv 0$ for $i = 1, \ldots, h$). Ljung-Box test is modification of the above procedure.

Ljung–Box test

$$Q_{LB} \;=\; n(n+2) \sum_{j=1}^{h} \widehat{\rho}^2(j) \;/\; (n-j) \tag{9.6}$$

The motivation of this modification is based on the fact that distribution Q_{LB} is more adequately approximated by χ^2_h distribution than Q.

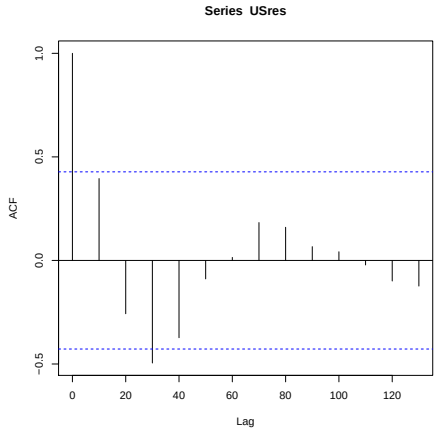
McLeod–Li test This is Ljung–Box test applied to $X_1^2, \ldots, X_n^2$. The idea is that sometimes nonlinear dependence between random variables is revealed when some (nonlinear) transformation is applied to them.

Example 9.1.1 *We calculate empirical correlations and perform Ljung-Box test with* $h = 10$ *for residuals of uspop.dat when the quadratic curve is fitted (compare Lecture 1).*

```
acf(USres)
Box.test(USres, lag = 10, type="Ljung"
 Box-Ljung test


data:  USres
X-squared = 18.5999, df = 10, p-value = 0.04565
```

Although at the level 0.05 we reject the null hypothesis that all correlations $\rho(i)\,, i = 1, \ldots, 10$ but it is a close call as p-value approximately equals 0.05. Moreover, judging on the autocorrelations alone we do not see obvious reasons why we should reject the null hypothesis: all autocorrelations but one for $1 \leqslant h \leqslant 10$ are contained in confidence intervals for white noise (note that time unit here is 10 years and that is why autocorrelations are computed for multiplies of 10).



9.1.3 Various white noise tests

There is a legion of tests for whiteness. We consider only few of them:
(i) Change of tendency (or turning points) test. The starting point of this test is the observation that for 6 different arrangements of the triple of distinct values only two are monotone and the remaining four contain change of tendency (turning point). Thus, if we denote by T be the number of triples (X_i, X_{i+1}, X_{i+2}), $1 \leqslant i \leqslant n-2$, such that X_{i+1} is a turning point, then for a strong white noise (X_t) we have

$$ET = 2(n-2)/3.$$

Moreover, one obtains

$$\sigma_T^2 = (16n - 29)/90.$$

Using the property above and asymptotic normality of $(T - ET)/\sigma_T$ it is easy to derive a form of rejection region of white noise test using T as a test statistic.
(ii) Sign of differences test. Let

$$S = \#\{2 \leqslant i \leqslant n : X_i > X_{i-1}\}.$$

For a strong white noise (X_t)

$$ES = (n-1)/2 \qquad \sigma_S^2 = (n+1)/12$$

which as before, as S is asymptotically normal, leads to construction of a rejection region for test statistic S. Note, however, that when the time series has a cyclic component then S may be close to $n/2$ and in such a case null hypothesis will not be rejected.
(iii) The rank test

$$P = \#\{n \geqslant j > i \geqslant 1 : X_j > X_i\}.$$

For a strong white noise we have

$$EP = n(n-1)/4 \qquad \sigma_P^2 = n(n-1)(2n+5)/72.$$

(iv) R/S test. We consider also a test which is based on the rescaled range of partial sums (cf Mandelbrot (1991)), frequently used in the case of financial returns. We define the rescaled range of the cumulative sums of X_t as

$$M_n = [\max_{1\leqslant i\leqslant n} \sum_{t=1}^{i}(X_t - \bar{X}_t)] - [\min_{1\leqslant i\leqslant n} \sum_{t=1}^{i}(X_t - \bar{X}_t)].$$

R/S statistic is

$$R/S = \sqrt{n}\frac{M_n}{\sigma}, \tag{9.7}$$

where $\hat{\sigma}$ is standard deviation of X_t. Application of this test to detect long-range dependence will be discussed in Chapter 12.
It is necessary to remember that if we use several of these tests then level of significance of a resulting joint test increases and we have to account for that applying e.g. Bonferroni correction.

9.2 Modelling nonstationary time series

In this section we will discuss the main modelling techniques, parametric as well as nonparametric, which are used to model time series which are nonstationary. We will first focus on a case of time series which after removal of deterministic part becomes weakly stationary. Most commonly used model of such process is as follows

$$X_t = m_t + s_t + Y_t, \qquad t \in \mathbb{Z}, \tag{9.8}$$

where Y_t is a weakly stationary zero-mean time series, m_t is a deterministic trend component and s_t is a is a deterministic seasonal component , which means that there exists $d \in \mathbb{N}$ such that for all $t \in \mathbb{Z}$ we have $s_{t+d} = s_t$.
Modelling of such time series usually consists in removing of a trend and a seasonal component and then fitting a model of stationary time series to the remaining part. Removal of both components is usually based on their preliminary estimation or appropriate differencing of time series. Alternatively, we can at once try to fit SARIMA process described below to the original time series but it is always worthwile to see how the trajectory of differenced process (with a lag 1 or equal to a suspected period of a process) looks like and whether it resembles a stationary one. Features of such process should include no obvious traits of periodicity or a trend and quickly decaying correlations. Trajectories of weakly stationary time series usually have these features but one should keep in mind that there are stationary processes, such as long memory processes, which trajectories resemble trajectories of nonstationary processes: they exhibit *local* trends and slowly decaying correlations.
We first start with parametric models (9.8) for which sesasonal component $s_t \equiv 0$ and then generalize them to include periodic pattern.

9.2.1 ARIMA and SARIMA processes

We first define ARIMA(p, d, q) process where additional parameter $d \in \mathbb{N}$.

Definition 21 *$(X_t)_{t\in\mathbb{Z}}$ is called ARIMA(p, d, q) process if its d^{th} difference $(1-B)^d X_t$ is causal ARMA(p, q) process, or equivalently*

$$\varphi(B)(1-B)^d X_t = \theta(B)Z_t, \tag{9.9}$$

where $(Z_t)_{t\in\mathbb{Z}}$ is WN$(0, \sigma^2)$ and $\varphi(z) \neq 0$ for $|z| \leqslant 1$.

Note that for $d > 0$ the process $(X_t)_{t\in\mathbb{Z}}$ satisfying structural equation (9.9) is nonstationary as 1 is a root of order d of $\varphi^*(z) = (1-z)^d\varphi(z)$. Moreover note that for such $(X_t)_{t\in\mathbb{Z}}$, series $(X_t)_{t\in\mathbb{Z}} + W(t)$, where W_t is a polynomial of order less than d, is also ARIMA(p, d, q). It is important feature of ARIMA(p, d, q) time series as it allows to model processes with a trend using them. However it also creates problems when making predictions for such processes, and without additional assumptions this is frequently impossible (see Problem 9.1). Putting it differently, for ARIMA(p, d, q) process we can determine covariance structure of $\varphi(B)(1-B)^d X_t$ but not that of (X_t). Note that the estimation of parameters of $\varphi(\cdot)$ and $\theta(\cdot)$ in such setting is a delicate problem as it may happen that e.g. process ARIMA$(1, 1, 0)$, which after differencing is AR(1), may be identified as stationary $AR(2)$ with one of the roots of polynomial $\varphi(\cdot)$ lying close to the unit circle. Usual practical procedure of fitting such processes consists in differencing of the original process several (usually no more than two or three times) and checking for traits of nonstationaritity for differenced series and then, if the process resembles stationary one, fitting ARMA time series to it.
ARIMA stands for Integrated AutoRegressive Moving Average process. Adjective 'integrated' is explained by the following example.

Example 9.2.1 *(i) ARIMA$(1, 1, 0)$ process or equivalently IAR$(1, 1)$. The process satisfies*

$$(1-\varphi B)(1-B)X_t = Z_t$$

which implies that with $Y_t = X_t - X_{t-1}$ we have

$$X_t - X_{t-1} = Y_t, \qquad X_t = X_0 + \sum_{j=1}^{t} Y_j.$$

As a cummulative sum above is a discrete equivalent of integral this explains the name of the process.
(ii) ARIMA$(0, 1, 1)$ or IMA$(1, 1)$ process.

$$(1-B)X_t = W_t - \theta W_{t-1} = (1-\theta B)W_t,$$

where $|\theta| < 1$ and W_t is WN$(0, \sigma^2)$. Thus

$$\frac{(1-B)}{(1-\theta B)}X_t = W_t,$$

$$(1-B)\sum_{i=0}^{\infty}\theta^i B^i X_t = X_t + \sum_{i=1}^{\infty}\theta^i X_{t-i} - \sum_{i=1}^{\infty}\theta^{i-1}X_{t-i} = W_t,$$

and equivalently

$$X_t = \sum_{i=1}^{\infty}(1-\theta)\theta^{i-1}X_{t-i} + W_t,$$

Note that if $W_t \perp H_{t-1}$ *then the term* $\sum_{i=1}^{\infty}(1-\theta)\theta^{i-1}X_{t-i}$ *in the decomposition above is an optimal predictor of* X_t*. Moreover in such a case*

$$\begin{aligned}(1-\theta)X_t + \theta\widehat{X}_t &= (1-\theta)X_t + \sum_{i=1}^{\infty}(1-\theta)\theta^i X_{t-i}\\ &= (1-\theta)X_t + \sum_{i=2}^{\infty}(1-\theta)\theta^{i-1}X_{t+1-i} = \widehat{X}_{t+1}. \quad (9.10)\end{aligned}$$

This is a recursive equation of exponential smoothing which we will encounter later on in connection with the Holt-Winters estimators.

9.2.2 SARIMA processes

Now we generalize ARIMA process to a broader concept of SARIMA series which incorporates non-zero seasonal component. To this end we introduce an operator of seasonal differencing with period s

$$(1-B^s)X_t = X_t - X_{t-s}.$$

Parameter $s \in \mathbb{N}$ and is equal to assumed or observed period of seasonal component in the data. Thus $s = 7$ corresponds to weekly period when t corresponds to days, $s = 4$ corresponds to yearly period when t corresponds to consecutive quarters of the year and so on. Note that operator $(1-B^s)$ used above is different from $(1-B)^s$ which corresponds to s-fold differencing. First we introduce a notion of ARMA$(P,Q)_s$ which is ARMA process for the time points $\{ks+s_0\}_{k\in\mathbb{Z}}$, where $0 \leqslant s_0 < s$ or, equivalently, it satisfies structural equation

$$\varphi(B^s)X_t = \Theta(B^s)\varepsilon_t, \quad (9.11)$$

where (ε_t) is WN$(0,\sigma^2)$. Consider the following example.

Example 9.2.2 *Assume that data is collected for* r *years every month, which gives* $12r$ *data points, and is arranged in the form of a table*

	Month			
	1	*2*	...	*12*
1	X_1	X_2	...	X_{12}
Rok 2	X_{13}	X_{14}	...	X_{24}
...	...	...	...	...
r	$X_{1+12(r-1)}$	$X_{2+12(r-1)}$	...	$X_{12+12(r-1)}$

We assume that each column is generated by ARMA(P,Q) *series:*
$X_{j+12t} = \varphi_1 X_{j+12(t-1)} + \cdots + \varphi_P X_{j+12(t-P)} + \varepsilon_{j+12t} + \Theta_1 \varepsilon_{j+12(t-1)} + \cdots + \Theta_Q \varepsilon_{j+12(t-Q)}, \quad j = 1, \ldots, 12, \ t = 1, \ldots, r$
For each $0 < j \leqslant 12$ ε_{j+12t} *is assumed to be white noise WN*$(0, \sigma^2)$*. Thus we impose independence of noise columnwise but not necessarily between columns. If we additionally assume that* (ε_t) *is white noise WN*$(0, \sigma^2)$ *we obtain ARMA*$(P,Q)_{12}$ *process. In particular, if* $P = 1$ *and* $Q = 0$ *ACF of the process satisfies* $\rho(12i) = \varphi^{|i|}$.

Definition 22 *The process which satisfies structural equality (9.11) with* (ε_t) *fulfilling*

$$\varphi(B)\varepsilon_t = \theta(B)Z_t,$$

where (Z_t) *is WN*$(0, \sigma^2)$ *is called multiplicative ARMA*$(p,q) \times (P,Q)_s$.

As operators $\varphi(B^s)$ and $\varphi(B)$ commute, we get

$$\begin{aligned} \varphi(B^s)\varphi(B)X_t &= \varphi(B)\varphi(B^s)X_t = \varphi(B)\Theta(B^s)\varepsilon_t \\ &= \Theta(B^s)\varphi(B)\varepsilon_t = \Theta(B^s)\theta(B)Z_t. \end{aligned}$$

A form of a structural equation explains the name of the process. Note that $\varphi(z)\varphi(z^s)$ is a polynomial of order $p + sP$ being a product of two polynomials of order p and sP for which some coefficients are zero and analogous observation is valid for $\theta(z)\Theta(z^s)$.

9.2.3 Nonparametric methods

We now briefly describe estimation of a trend and a seasonal component without assuming additional structure of these terms. As before, assume initially that $s_t \equiv 0$ and note that m_t may be estimated by any of many nonparametric regression estimators, such as kernel estimators or polynomial smoothers. In order to see why it works consider a simple estimator, a running mean

$$\widehat{m}_t = \frac{1}{2q+1} \sum_{|j| \leqslant q} X_{t-j} = \sum_{j=-\infty}^{\infty} a_j X_{t-j}, \tag{9.12}$$

where

$$a_j = \begin{cases} \frac{1}{2q+1}, & \text{for } |j| \leqslant q, \\ 0, & \text{otherwise,} \end{cases}$$

and we enlarge the sample path by letting $X_t = X_1$ for $t < 1$, $X_t = X_n$ for $t > n$. Thus running mean is a special linear filter of the observed data. Note that when (9.8) holds

$$\widehat{m}_t = \frac{1}{2q+1} \sum_{|j| \leqslant q} m_{t-j} + \frac{1}{2q+1} \sum_{|j| \leqslant q} Y_{t-j}$$

and if $q = q_n$ is not too large the first term in the decomposition above is a reasonable approximation to m_t provided m_t does not change rapidly. On the other hand, if q_n is not to small the second term will be close to 0, if process (Y_t) is ergodic in view of ergodic theorem. It follows that for moderate t $\widehat{m}_t$ should be a reasonable estimator of m_t.
Consider now the case that seasonal component is present and assume that

$$s_{t+d} = s_t \qquad \text{and} \qquad \sum_{j=1}^{d} s_j = 0.$$

Note that the second assumption is not restrictive as we can always center (s_t) by the mean value of the period $\bar{s}$ and augment m_t by this value. We describe now the stepwise procedure of estimating s_t and m_t. Throughout d is assumed known.

1) The first step is preliminary estimation of m by the running median with number of terms equal to d. More specifically, for $d = 2q + 1$ we consider the ordinary running mean (9.12) with parameter q, whereas for $d = 2q$ we define

$$\widehat{m}_t = \frac{0.5X_{t-q} + X_{t-q+1} + \cdots + X_{t+q-1} + 0.5X_{t+q}}{d}, \qquad q \leqslant t \leqslant n - q.$$

2) In the second step we estimate seasonal component s_k for $k \leqslant d$ $(s_{k+d} = s_k)$ as follows. Let

$$\tilde{s}_{k+jd} = X_{k+jd} - \widehat{m}_{k+jd}, \qquad j\colon q \leqslant k + jd \leqslant n - q,$$

Let w_k be the mean of $\{\tilde{s}_{k+jd}\}$ and finally

$$\widehat{s}_k = w_k - \bar{w}.$$

3) The third step is simply deseasonalisation of X_t

$$d_t = X_t - \widehat{s}_t.$$

4) In the fourth step m_t is estimated parametrically or nonparametrically by $\widehat{m}_t$.
5) In the last step we fit a stationary process to $Y_t = d_t - \widehat{m}_t$,
We recall that for $s_t = 0$, polynomial trend can be eliminated by d-fold differentiation by taking into account that if $m_t = \sum_{j=0}^{k} \psi_j t^j$ then

$$\bigtriangledown^k X_t = \bigtriangledown^k Y_t + k! c_k.$$

Thus $d = k+1$ is sufficient to remove polynomial trend of order k. However, the problem is that we eliminate and not estimate trend,the form of which is usually of interest.

9.2.4 The Holt–Winters method

We discusss now the most popular method of estimation of a trend and seasonal components. It is quite ingenious albeit simple. We describe the method first for the case $s_t \equiv 0$. Instead of estimating only trend m_t we aim at simultaneous estimation of

$$(m_t, b_t),$$

where b_t denotes the change of the trend in the moment t. This is done recursively. Let $0 \leqslant \alpha, \beta \leqslant 1$ denote parameters of the method. Define

$$\begin{cases} \widehat{m}_{n+1} = (\widehat{m}_n + \widehat{b}_n)(1-\alpha) + \alpha X_{n+1}, & 0 < \alpha < 1, \\ \widehat{b}_{n+1} = (\widehat{m}_{n+1} - \widehat{a}_n)\beta + (1-\beta)\widehat{b}_n, & 0 < \beta < 1, \end{cases} \tag{9.13}$$

and we let $\widehat{m}_2 = X_2, \widehat{b}_2 = (X_2 - X_1)$, Thus the trend at moment $n+1$ is estimated by the convex combination of the observation X_{n+1} at this moment and the estimator of the trend at this moment, namely $m_n + b_n h$ for $h = 1$. The equations (9.13) are recursively solved for $i = 3, \ldots, n$.
Note that recurrence is performed timewise: values of a trend and its change approximated at t are used to estimate these parameters at the later time points.
Consider now the general case given by (9.8). Now the aim is to estimate

$$(m_t, b_t, s_t),$$

where s_t has period d. The Holt-Winter equations are as follows

$$\begin{cases} \widehat{m}_{n+1} = (\widehat{m}_n + \widehat{b}_n)(1-\alpha) + \alpha(X_{n+1} - \widehat{s}_{n+1-d}), & 0 < \alpha < 1, \\ \widehat{b}_{n+1} = (\widehat{m}_{n+1} - \widehat{m}_n)\beta + (1-\beta)\widehat{b}_n, & 0 < \beta < 1, \\ \widehat{c}_{n+1} = (1-\gamma)\widehat{s}_{n+1-d} + \gamma(X_{n+1} - \widehat{m}_{n+1}), & 0 < \gamma < 1. \end{cases}$$

For the first equation we used the fact that seasonal component s_{n+1} at time $n+1$, the estimator of which is not available at this stage, can be replaced by $\widehat{s}_{n+1-d}$. We let:
$\widehat{m}_{d+1} = X_{d+1}$,
$\widehat{b}_{d+1} = (X_{d+1} - X_1)/d$,
$\widehat{s}_i = Y_i - (Y_1 + \widehat{b}_{d+1}(i-1)), \quad i = 1, 2, \ldots, d.$
Usual method of choosing α and β is

$$(\alpha_0, \beta_0) := \arg\min_{\alpha,\beta,\gamma} \sum_{i=d+1}^{n} (X_i - \widehat{X}_i)^2.$$

Note that the Holt-Winters estimators yield assumptions-free predictors of X_{n+h}, namely

$$\widehat{X}_{n+h} = \widehat{a}_n + \widehat{b}_n h + \widehat{c}_{n+h-d}.$$

Observe that in the simplest case when we let $\beta = \gamma = 0$ and use only α above we obtain:

$$\widehat{m}_{n+1} = \widehat{m}_n(1-\alpha) + \alpha X_{n+1}.$$

This is so called exponential smoothing method. Explicit solution to thee above equation is

$$\widehat{m}_{t+1} = \sum_{j=0}^{t-1} \alpha(1-\alpha)^j X_{t+1-j} + (1-\alpha)^t X_1.$$

Note that for estimation of m_{t+1} the influence of X_{t+1-j} is exponentially decaying in j, whence the name.

9.2.5 Problems

1. Let X_t dla $t \geqslant 0$ is defined as $X_t = X_0 + \sum_{i=1}^{t} Y_i$, where Y_i ARMA(p,q) series. X_t is thus ARIMA$(p,1,q)$. Let $H_t = sp(X_0, X_1, \ldots, X_t)$. Prove that

$$P_{H_t} X_{t+1} = X_t + P_{H_t} Y_{t+1}$$

and if X_0 and Y_i for $i \geqslant 0$ are uncorrelated then $P_{H_t} Y_{t+1} = P_{sp(Y_1,\ldots,Y_t)} Y_{t+1}$.
2. Justify the following statement: if for a certain data AIC and BIC for ARMA(p,q) process have unique minima and models corresponding to the minima have the same number of parameters than those models coincide.
3. Construct an example of a weak white noise $(\varepsilon_t)_{t\in\mathbb{Z}}$ such that (ε_t^2) are correlated, generate trajectories of length $n = 100, 500, 100$ and check how Ljung-Box and McLeod Li test works for them.
4. Prove that if $X_t = Y_t + m_t$, where $m_t = \sum_{j=0}^{k} \psi_j t^j$ then

$$\bigtriangledown^k X_t = \bigtriangledown^k Y_t + k! c_k.$$

5. Suppose that we fit AR(k) model with $\sigma^2 = 1$ to the data where $1 \leqslant k \leqslant p$ and k denotes the number of lags considered $X_{t-i_1}, \ldots, X_{t-i_k}$. BIC approach assumes uniform distribution on the family $\mathcal{M}$ of all possible 2^p models and let U be a random variable having this distribution i.e. $P(U = m) = 1/2^p$ for any $m \in \mathcal{M}$. Show that the number of elements $|U|$ has binomial distribution Bin$(p, 1/2)$ and thus $E|U| = p/2$. This indicates that when the number of true legs is much smaller than $p = p_n$ BIC criterion may have, similarly to AIC, the tendency to select too large a model.

10

Estimation of the spectral density

The chapter is devoted to discussion of properties of two basic estimators of a spectral density, namely a periodogram and a smoothed periodogram.
Before we consider nonparametric estimators of the spectral density note that in parametric case, in particular for ARMA processes we can use the form of their spectral densities derived in Chapter 5 for more general case of linear processes. In particular, in view of (6.14) a plug-in estimator of the spectral density of AR(p) is

$$\hat{f}(\lambda) = \frac{\hat{\sigma}^2}{2\pi}|1 - \hat{\varphi}_1 e^{-i\lambda} - \ldots - \hat{\varphi}_p e^{-ip\lambda}|^2, \tag{10.1}$$

where $\hat{\varphi}_1, \cdots, \hat{\varphi}_p$ and $\hat{\sigma}^2$ are the Yule-Walker estimators. This estimator has an interesting property: it maximises entropy among spectral density estimators for which pertaining variances are equal $\hat{\gamma}(0), \ldots \hat{\gamma}(p)$ (cf. Brockwell and Davis (1991), Section 10.6). Estimator (10.1) can be extended in an obvious way to estimate a spectral density of ARMA process in view of (6.14) using e.g. ML estimators.

10.1 Periodogram

Assume that the underlying process $(X_t)_{t\in\mathbb{N}}$ is real valued weakly stationary having mean μ and such that its covariance function is absolutely summable : $\sum_{h=-\infty}^{\infty} |\gamma(h)| < \infty$.
We recall that inversion theorem (cf Theorem 6.2.2) states that in this case

$$f(\lambda) = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} \gamma(h) e^{-ih\lambda}, \qquad \lambda \in [-\pi, \pi]. \tag{10.2}$$

We show now that the definition of the simplest estimator of the spectral density, namely a periodogram, is directly based on (10.2). Suppose that we want to estimate $f(\lambda)$ after having observed $X_1, \ldots, X_n$.

Definition 23 *An estimator of $f(\cdot)$ based on $X_1, \ldots, X_n$ and defined as*

$$I_n(\lambda) = \frac{1}{2\pi n} \mid \sum_{t=1}^{n} X_t e^{it\lambda} \mid^2 = \frac{1}{2\pi n}\Big((\sum_{t=1}^{n} X_t \cos t\lambda)^2 + (\sum_{t=1}^{n} X_t \sin t\lambda)^2 \Big) \tag{10.3}$$

for $\lambda \in [-\pi, \pi]$, is called periodogram.

Note that as X_t is real valued we have $I_n(\lambda) = I_n(-\lambda)$ i.e. periodogram is an even function. Moreover, for $\lambda = 0$, $I_n(0) = n(\bar{X}_n)^2/2\pi$, thus $I_n(0)$ can be calculated once the empirical mean is evaluated.
Justification of (10.3) rests on the following fact. Consider first the situation when $\mu = 0$ and modify the definition of an empirical covariance to

$$\bar{\gamma}(h) = \frac{1}{n - |h|} \sum_{t=1}^{n} X_t X_{t+|h|}$$

Note that the centring by $\bar{X}_n$ in the definition of empirical covariance is omitted as the mean is zero and is assumed known. Using the definition of $\widehat{\gamma}(h)$ we can write the definition of the periodogram as

$$I_n(\lambda) = \frac{1}{2\pi n} \sum_{1 \leqslant t,s \leqslant n} X_t X_s e^{i(t-s)\lambda} = \frac{1}{2\pi} \sum_{h:|h|<n} \bar{\gamma}(h) e^{-ih\lambda} \tag{10.4}$$

which bears a close resemblance to (10.2).

Definition 24 *Fourier frequencies are defined as $\lambda_j = 2\pi j/n$, where $j \in F_n$ and*

$$F_n := \{-[\frac{n-1}{2}], [\frac{n-2}{2}] \ldots, [\frac{n}{2}]\},$$

with $[y]$ denoting the integer part of y.

Note that the set $\{\lambda_j\}_{j \in F_n}$ consists of n frequencies contained in $(-\pi, \pi]$. For odd $n = 2k + 1$, F_n is symmetric with respect to 0 and corresponding frequencies are $-2\pi k/n, \ldots, 2\pi k/n$, whereas for $n = 2k$ the frequencies are $-2\pi(k-1)/n, \ldots, 2\pi k/n$. The largest frequency for even n does not have its mirror image in F_n. We remark that periodogram calculated at a Fourier frequency is the squared modulus of Discrete Fourier Transform of $(X_t)_{t=1}^n$ which can be efficiently calculated using Fast Fourier Transform.
Observe that for an arbitrary mean μ and using the usual definition of an empirical covariance

$$\widehat{\gamma}(h) = \frac{1}{n} \sum_{t=1}^{n-|h|} (X_t - \bar{X})(X_{t+|h|} - \bar{X}) \tag{10.5}$$

the representation (10.4) is valid for Fourier frequencies $\lambda_j \neq 0$. Namely, taking advantage of

$$\sum_{t=1}^{n} e^{i\lambda_j t} = e^{i\lambda_j} \sum_{t=0}^{n-1} e^{i\lambda_j t} = \frac{e^{i\lambda_j} - e^{in\lambda_j}}{1 - e^{i\lambda_j}} = 0, \quad j \neq 0, \tag{10.6}$$

we have

$$I_n(\lambda_j) = \frac{1}{2\pi n} \mid \sum_{t=1}^{n} X_t e^{it\lambda_j} \mid^2$$
$$= \frac{1}{2\pi n} \mid \sum_{t=1}^{n} (X_t - \bar{X}) e^{it\lambda_j} \mid^2 = \frac{1}{2\pi} \sum_{|h|<n} \widehat{\gamma}(h) e^{-ih\lambda_j},$$

where $\widehat{\gamma}(h)$ is defined in (10.5).

Remark 10.1.1 *We can also interpret the periodogram differently. To this end, define*

$$\mathbb{C}^n \ni \mathbf{e}_j = n^{-1/2}(e^{i\lambda_j}, e^{i2\lambda_j}, \ldots, e^{in\lambda_j})'$$

Note that $(\mathbf{e}_j)_{j\in F_n}$ *form an orthonormal basis in* $\mathbb{C}^n$.
Thus for $\mathbf{X} = (X_1, X_2, \ldots, X_n)'$ *we have for certain* $\alpha_1, \ldots, \alpha_n$

$$\mathbf{X} = \sum_{j\in F_n} \alpha_j \mathbf{e}_j, \tag{10.7}$$

and

$$\alpha_j = < X, \mathbf{e}_j > = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} X_k e^{-ik\lambda_j}.$$

The sequence $\{\alpha_k\}$ *is called a discrete Fourier transform of* $\mathbf{X}$. *Series* (X_t) *is decomposed into a linear combination of periodic waves* $\mathbf{e}_j$ *with frequencies* λ_j *and, moreover, a large coefficient* α_j *corresponds to a large contribution of the pertaining component. This can be clearly seen by the following development. Note that* $I_n(-\lambda_j)$ *is related to squared scalar product of* $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ *and* $\mathbf{e}_j$. *Specifically,*

$$I_n(-\lambda_j) = \frac{1}{2\pi} \mid < \mathbf{X}, \mathbf{e}_j > \mid^2 = \frac{1}{2\pi} \mid \sum_{i=1}^{n} X_t e^{-it\lambda_j} \mid^2 = \frac{|\alpha_j|^2}{2\pi}.$$

Since $(\mathbf{e}_j)_{j\in F_n}$ *is orthonormal basis we have*

$$\parallel \mathbf{X} \parallel^2 = \sum_{k=-[(n-1)/2]}^{[n/2]} |\alpha_k|^2 = 2\pi \sum_{j\in F_n} I(\lambda_j). \tag{10.8}$$

The above equality shows the decomposition of $\parallel \mathbf{X} \parallel^2$, *which is the total variability of* $\mathbf{X}$, *into components related to different Fourier frequencies.*
Note also that as series X_t *is real-valued (10.7) can be written as*

$$X_t = \sum_{j\in F_n} \alpha_j e^{it\lambda_j} =$$
$$= \frac{1}{\sqrt{n}}\alpha_0 + \frac{2}{\sqrt{n}} \sum_{k=1}^{[(n-1)/2]} (\alpha_{j1}\cos(\lambda_j t) + \alpha_{j2}\sin(\lambda_j t)) + (-1)^{n/2}\frac{1}{\sqrt{n}}\alpha_{n/2},$$

where the last term above is defined as 0 for an odd n.

Sometimes a different definition of periodogram is used which defines it as a piecewise constant function with heights of steps equal to $I_n(\lambda_0), \ldots, I_n(\lambda_{[n/2]})$. Namely, for $k = 0, \ldots, [n/2]$ define

$$I_n(\omega) = \begin{cases} I_n(\lambda_k), & \lambda_k - \pi/n < \lambda \leqslant \lambda_k + \pi/n, \quad 0 \leqslant \lambda \leqslant \pi, \\ I_n(-\lambda), & \lambda \in [-\pi, 0), \end{cases}$$

Here throughout we will consider (10.3) as the definition of the periodogram.

10.2 Basic properties of periodogram

Theorem 10.2.1 *Let $(X_t)_{t\in\mathbb{N}}$ be a weakly stationary process with mean μ and $\gamma(\cdot) \in \ell^1$. Then for $\lambda = 0$*

$$EI_n(0) - n\mu^2/2\pi \longrightarrow f(0) \tag{10.9}$$

and for $\lambda \in [-\pi, \pi] \setminus \{0\}$ we have

$$EI_n(\lambda) \longrightarrow f(\lambda) \tag{10.10}$$

Proof. In order to prove (10.9) note that $EI_n(0) = nE(\bar{\mathbf{X}}^2)/2\pi$, and thus left-hand side of (10.9) equals

$$\frac{1}{2\pi}\,(nE(\bar{\mathbf{X}}^2) - n\mu^2) = \frac{1}{2\pi}\,n\mathrm{Var}(\bar{\mathbf{X}}) \longrightarrow f(0),$$

where the last convergence follows as $\gamma(\cdot) \in \ell^1$. We prove (10.10) for $\mu = 0$. In this case

$$EI_n(\lambda) \;=\; \frac{1}{2\pi} \sum_{|h|<n} \frac{n - |h|}{n} \gamma(h) e^{-ih\lambda} \longrightarrow f(\lambda).$$

For general μ we prove that for any sequence of Fourier frequencies $\lambda_{k,n} \to \lambda$ we have $EI_n(\lambda_{k,n}) \to f(\lambda)$ by using (10.6) and then the proof is accomplished by taking into account continuity of $EI_n(\cdot)$.
Thus for a zero mean process with summable covariances $I_n(\lambda)$ is an asymptotically unbiased estimator of $f(\lambda)$ and it is also easy to see that in this case the convergence of $EI_n(\lambda)$ to $f(\lambda)$ is uniform in $\lambda \in [-\pi, \pi]$.
However, it turns out that the variability of the periodogram does not tend to 0 with the growing sample size, namely $\mathrm{Var} I_n(\lambda) \to C \neq 0$, when $n \to \infty$ and thus $I_n(\lambda)$ is *not* consistent estimator of $f(\lambda)$. This in particular follows from the theorem below asserting that the asymptotic distribution of the periodogram is exponential. Before we state the result we give an intuition why it holds.

Example 10.2.2 *Distribution of periodogram for Gaussian white noise.*
Vectors

$\mathbf{c}_j =< \cos\lambda_j, \cos 2\lambda_j, \ldots, \cos n\lambda_j >' \cdot (2/n)^{1/2}$,
$\mathbf{s}_j =< \sin\lambda_j, \sin 2\lambda_j, \ldots, \sin n\lambda_j >' \cdot (2/n)^{1/2}, \qquad j = 1, \ldots, [(n-1)/2]$
are orthonormal and from the previous remark it follows that

$$I_n(\lambda_j) = \frac{1}{2\pi}\left\{\frac{< \mathbf{X}, \mathbf{c}_j >^2 + < \mathbf{X}, \mathbf{s}_j >^2}{2}\right\}.$$

Thus if (X_t) is a Gaussian white noise $N(0, \sigma^2)$ then $< \mathbf{X}, \mathbf{c}_j >$ and $< \mathbf{X}, \mathbf{s}_j >$ are distributed as $N(0, \sigma^2)$ and they are independent. Indeed, in particular

$$E(< \mathbf{X}, \mathbf{c}_j >^2) = E\Big(\frac{2}{n}\sum_{k=1}^{n} X_k^2 \cos^2 k\lambda_j\Big) = \frac{2\sigma^2}{n}\sum_{k=1}^{n}\cos^2 k\lambda_j = \sigma^2.$$

Thus $< \mathbf{X}, \mathbf{c}_j >^2 + < \mathbf{X}, \mathbf{s}_j >^2$ has χ^2_2 distribution which coincides with the exponential distribution $\mathcal{E}((2\sigma^2)^{-1})$ with parameter $\lambda = (2\sigma^2)^{-1}$ and expected value $2\sigma^2$. Thus it follows that $I_n(\lambda_j) \sim \mathcal{E}(2\pi/\sigma^2) = \mathcal{E}(f(\lambda_j)^{-1})$ and has expected value $f(\lambda_j)$.

Note that in view of this intuition set F_n in (10.8) may be replaced by the set λ_i, $i = 1, \ldots, [(n-1)/2]$ with added values $\lambda_0 = 0$ and $\lambda_{n/2} = \pi$ (the last value only for even n). Contribution of the last values to the decomposition of $||\mathbf{X}||^2$ equals $I(\lambda_0)$ i $I(\lambda_{n/2})$ respectively, whereas the contribution of the remaining frequencies should be counted twice and equals $2I(\lambda_i)$.
We first state the proposition asserting that the variance of the periodogram converges to the constant value. However, periodograms at different points become asymptotically uncorrelated. In order to see this in the special case let $(\varepsilon_t)_{t\in\mathbb{Z}}$ be a strong WN$(0, \sigma^2)$ with fourth cumulant κ_4. Reasoning similarly to the proof of Theorem 7.4.1,we have for $\lambda_1 \neq \lambda_2$

$$\begin{aligned} E(I_n(\lambda_1)I_n(\lambda_2)) &= E\Big[\frac{1}{4\pi^2 n}\sum_v\sum_u\sum_t\sum_s \varepsilon_s\varepsilon_t\varepsilon_u\varepsilon_v e^{i\{(s-t)\lambda_1+(u-v)\lambda_2\}}\Big] \\ &= \frac{1}{4\pi^2 n}\Big[\sigma^4\Big\{n^2 + \sum_t\sum_s e^{i(s-t)(\lambda_1+\lambda_2)} + \sum_t\sum_s e^{i(s-t)(\lambda_1-\lambda_2)}\Big\} + n\kappa_4\Big] \\ &= \frac{\kappa_4}{4\pi^2 n} + \frac{\sigma^4}{4\pi^2} + \frac{\sigma^4}{4\pi^2 n^2}\Big\{\Big(\frac{\sin\frac{1}{2}n(\lambda_1+\lambda_2)}{\sin\frac{1}{2}(\lambda_1+\lambda_2)}\Big)^2 + \Big(\frac{\sin\frac{1}{2}n(\lambda_1-\lambda_2)}{\sin\frac{1}{2}(\lambda_1-\lambda_2)}\Big)^2\Big\} \quad (10.11) \end{aligned}$$

Performing similar calculations for $\lambda_1 = \lambda_2$ we thus obtain

Proposition 10.2.3 *Let $(X_t)_{t\in N}$ be strong white noise WN$(0, \sigma^2)$ and $\kappa_4 = cum(X_t, X_t, X_t, X_t)$.*
For $0 < \lambda < \pi$

$$VarI_n(\lambda) = \frac{\sigma^4}{4\pi^2} + \frac{\kappa_4}{4\pi^2 n} + o(\frac{1}{n})$$

and for $\lambda = 0, \pi$

$$VarI_n(\lambda) = \frac{\sigma^4}{2\pi^2} + \frac{\kappa_4}{4\pi^2 n} + o(\frac{1}{n}),$$

whereas for $\lambda_1 \neq \lambda_2$

$$Cov(I_n(\lambda_1), I_n(\lambda_2)) = \frac{\kappa_4}{4\pi^2 n} + o(\frac{1}{n}).$$

Thus it follows that if $\kappa_4 \neq 0$ $\mathrm{Cov}(I(\lambda_1), I(\lambda_2))$ is of order $O(n^{-1})$ whereas for $\kappa_4 = 0$ is $O(n^{-2})$. Moreover, variance of $I_n(\lambda)$ is aproximately twice as large as for $I_n(\lambda)$ for $\lambda \neq 0$.
The main result stated below confirms the intuition that periodogram is asymptotically exponential. Moreover, for distinct frequencies periodograms become asymptotically independent.

Theorem 10.2.4 *(i) Let* $X_t = \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}$, (Z_t) *be a strong* $WN(0, \sigma^2)$, $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$ *and* ω_i, $i = 1, \ldots, m$ *such that* $f(\lambda) > 0$ *for* $\lambda \in [-\pi, \pi]$, $0 < \omega_1 < \omega_2 < \cdots < \omega_m < \pi$. *Then*

$$(I_n(\omega_1), \ldots, I_n(\omega_m)) \xrightarrow{\mathcal{D}} (\varepsilon_1, \ldots, \varepsilon_m),$$

where ε_i *are independent exponentially distributed* $\mathcal{E}(f(\omega_i)^{-1})$.
(ii) If moreover $EZ_t^4 < \infty$ *and* $\sum |j|^{1/2} |\psi_j| < \infty$ *then* $\mathrm{Cov}(I_n(\lambda_k), I_n(\lambda_j)) = O(n^{-1})$, *where* $\lambda_j = 2\pi j/n, \lambda_k = 2\pi k/n$ *are two different Fourier frequencies for fixed* $j, k \in \mathbb{N}$.

The result shows a momentous advantage of moving from state domain to frequency domain. By doing so we change a complicated dependence structure of a weakly stationary process to asymptotically independent periodogram values from which the original dependence structure can be recovered.
In particular it follows that the variables $U_n(\omega) = (I_n(\omega) - f(\omega))/f(\omega)$ are approximately $\mathcal{E}(1)$ distributed and $U_n(\omega)$ i $U_n(\omega')$ are asymptotically independent for $\omega \neq \omega'$. Also from the last property together with the fact that asymptotic variance is approximately the same for close points it follows that the averaging of periodogram over close frequencies should diminish the variance of the estimate. Moreover, if the frequencies over which averaging is done are sufficiently close, we should not loose asymptotic unbiasedness of the periodogram. These observations will be used below in construction of smoothed periodogram.

10.2.1 Prewhitening

The idea is parallel to the tranformation method in density estimation and is based on the observation that the original spectral density might be difficult to estimate because e.g. of sharp peaks and it is much easier to estimate regular spectral density without much variation. Thus we transform appropriately the original data, estimate the spectral density of transformed sample and then transform it back. The transformation is usually the linear filter as the resulting spectral density is easily derived. If $Y_t = \sum_{k=0}^{\infty} \psi_k X_{t-k}$ is the chosen filter and $\hat{f}_Y$ spectral density estimator for the filtered data then prewhitened estimator is

$$\hat{f}_X(\lambda) = |\Psi(e^{-i\lambda})|^{-2}\hat{f}_Y(\lambda),$$

where $\Psi(e^{-i\lambda}$ is a transfer function of the filter. The crucial question is how to choose a filter. Note that the most natural proposal would be to transform (X_t) to a white noise but this obviously requires knowledge of spectral density f_X. Frequently, coefficients of AR(p) approximation are chosen as coefficients of the filter, when the approximation is obtained e.g. by considering the least squares fit of lagged observations $X_{t-1}, \ldots, X_{t-p}$ to X_t.

10.3 White noise tests using periodograms

10.3.1 Test of cumulative periodogram

As we have shown above that for Gaussian white noise periodograms at Fourier frequencies $I_n(\lambda_j)$ have exponential distribution $\mathcal{E}(\sigma^2/2\pi)$ and are independent, the following proposition can be used to construct test statistic of whiteness of Gaussian sequence.

Proposition 10.3.1 *Assume that $Z_1, \ldots, Z_q$ are independent random variables with exponential $\mathcal{E}(\lambda)$ distribution, then*

$$Y_i = \Big(\sum_{k=1}^{i} Z_k\Big)/\Big(\sum_{k=1}^{q} Z_k\Big), \qquad i = 1, \ldots, q-1$$

have the same joint distribution as order statistics $U_{1:q-1}, \ldots, U_{q-1:q-1}$ of uniform iid random variables $U_1, \ldots, U_{q-1}$

$$(Y_1, \ldots, Y_{q-1}) \ \sim \ (U_{1:q-1}, \ldots, U_{q-1:q-1}).$$

Note in particular that if F_{q-1} denotes cumulative distribution function of $Y_i, i = 1, \ldots, q-1$ defined in the proposition, then its distribution coincides with that of cumulative distribution of a random sample consisting of $q-1$ independent uniform random variables. Thus Kolmogorov–Smirnov statistic can be used to construct test whether the underlying sample is a Gaussian iid sample. We whence define with $q = [(n-1)/2]$

$$T_{KS} = \sup_{s\in[0,1]} \sqrt{q-1} \ \mid \hat{F}_{q-1}(s) - s \mid .$$

Critical region of the test is built using the corollary of Donsker's theorem

$$T_{KS} \xrightarrow{\mathcal{D}} \sup_{s\in[0,1]} \mid B^0(s) \mid,$$

under H_0, where B^0 is the Brownian bridge on $[0,1]$. Thus $\hat{F}_{q-1}$ should lie in a strip $y = x \pm w_{1-\alpha/2}(q-1)^{-1/2}$, where $w_{1-\alpha}$ is the quantile of order $1-\alpha$ of the

distribution of the $\sup_{s\in[0,1]} |\, B^0(s)\,|$. In order to apply this test $q = [(n-1)/2]$ should be at least 30.
Note that similar test for the hypothesis $H{:}f = f_0$, where f_0 is a fixed spectral density may be constructed by replacing $I(\lambda_k)$ above by $I(\lambda_k)/f_0(\lambda_k)$, which under H_0 have approximately $\mathcal{E}(1)$ distribution.

10.3.2 Fisher's test

The last proposition can be used to construct many test statistics for the whitness of Gaussian sequence having the form $T = f(Y_1, \ldots, Y_{q-1})$ provided that the distribution of T when $Y_1, \ldots, Y_{q-1}$ is the uniform random sample is known. Fisher's test considers the maximal spacing

$$M_q = \max_{1\leqslant i\leqslant q} (Y_i - Y_{i-1}) \;=\; \max_{1\leqslant i\leqslant q} \frac{I(\lambda_i)}{\sum\limits_{k=1}^{q} I(\lambda_k)}, \qquad Y_0 = 0,\ Y_q = 1.$$

Its distribution is

$$P(M_q \leqslant x) = \sum_{j=1}^{q} (-1)^j \binom{q}{j} (1 - ja)_+^{q-1},$$

(cf. Feller (1971), p. 29) and is used to determine threshold of the critical region. Note that qM_q can be used to test for hidden periodicities as M_q is the maximal value of the periodogram among Fourier frequencies relative to their average. Thus M_q is well suited to detect the situation when (X_t) is Gaussian white noise with added periodic component.

10.4 Smoothed periodograms

We discuss now smoothed periodograms which in contrast to original periodogram are consistent estimators of the spectral density. This method stems from a recognition that partial sums $(2\pi)^{-1} \sum_{k=-p}^{p} \gamma(k) e^{-ik\lambda}$ might not be a good approximation to a spectral density and the convergence is improved by introducing weight factors into the latter yielding

$$\frac{1}{2\pi} \sum_{k=-p}^{p} \Big(1 - \frac{|k|}{p}\Big) \gamma(k) e^{-ik\lambda}.$$

This idea is now generalized and applied to periodogram. To this end consider a kernel w defined as a function such that $w(x) = 0$ for $x \in [-1, 1]^c$, which is symmetric, bounded by 1 and satisfies $w(0) = 1$. It is also sometimes called data window or a taper. Consider also integer $r = r(n)$ such that $r \leqslant n$. Value $w(h/r)$ for $h \leqslant r$ will play the role of the weight for $\hat{\gamma}(h)$.

Definition 25 *We define smoothed periodogram as*

$$\widehat{f}(\lambda) = \frac{1}{2\pi} \sum_{|h| \leqslant r} w(h/r)\ \widehat{\gamma}(h)\ e^{-ih\lambda}. \tag{10.12}$$

Note that for a typical weight function $w(\cdot)$ such that $w(\cdot)$ is decreasing on $[0, 1]$, $w(h/r)$ down-weighs $\widehat{\gamma}(h)$ for larger h and the maximal downweighing correspond to h such that $h \approx r$. Estimator $\widehat{f}(\lambda)$ is sometimes called the lag window spectral density estimator and w is called the lag window. We show now that $\widehat{f}(\lambda)$ is smooothed version of the periodogram in the sense that is obtained as convolution of the periodogram with so called spectral window. Namely, we define for $\lambda \in [-\pi, \pi]$

$$\widetilde{I}_n(\omega) = \frac{1}{2\pi} \sum_{|h|<n} \widehat{\gamma}(h)\ e^{-ih\omega}.$$

$\widetilde{I}_n(\cdot)$ is an approximation to the periodogram which we considered justifying its definition. We have that $\widetilde{I}_n(\lambda_j) = I_n(\lambda_j)$, moreover

$$\widehat{\gamma}(h) = \int_{-\pi}^{\pi} e^{ih\omega} \widetilde{I}_n(\omega) d\omega.$$

Thus smoothed periodogram (10.12) equals

$$\begin{aligned}
\widehat{f}(\lambda) &= \frac{1}{2\pi} \sum_{|h| \leqslant r} w(h/r) \int_{-\pi}^{\pi} e^{-ih(\lambda-\omega)}\ \widetilde{I}_n(\omega)\ d\omega\ = \\
&= \int_{-\pi}^{\pi} \frac{1}{2\pi} \sum_{|h| \leqslant r} w(h/r)\ e^{-ih(\lambda-\omega)}\ \widetilde{I}_n(\omega)\ d\omega\ = \\
&= \int_{-\pi}^{\pi} W_r(\lambda - \omega)\ \widetilde{I}_n(\omega)\ d\omega = \int_{-\pi-\lambda}^{\pi-\lambda} W(\omega)\ \widetilde{I}_n(\omega + \lambda)\ d\omega,
\end{aligned}$$

where

$$W_r(\omega) = \frac{1}{2\pi} \sum_{|h| \leqslant r} w(h/r)\ e^{-ih\omega},$$

is so called spectral window. Thus with $*$ denoting the convolution we have

$$\widehat{f}(\lambda) = W_r * \widetilde{I}_n(\lambda)$$

It is seen that $\hat{f}$ is approximately equal to $(2\pi/n) \sum_{|j| \leqslant [n/2]} W_r(\lambda_j) \widetilde{I}_n(\lambda + \lambda_j)$ thus can be regarded as a weighted average of values $I_n(\lambda + \lambda_j)$. We also note that

$$\int_{-\pi}^{\pi} W_r(\lambda)\, d\lambda = \frac{1}{2\pi} \int_{-\pi}^{\pi} w(0)\, d\lambda = w(0) = 1.$$

Although $W_r(\cdot)$ integrates to 1 it can have negative values and thus does not have to be a density.
We discuss now basic asymptotic properties of smoothed periodogram.

Proposition 10.4.1 *(i) If* $r = r(n) \to \infty$ *then* $\lim_{n\to\infty} E\widehat{f}(\lambda) = f(\lambda)$, *(ii) When* $r(n)/n \to 0$ *then*

$$\frac{n}{r(n)}\, Var\widehat{f}(\lambda) \longrightarrow \begin{cases} 2f^2(\lambda) \int\limits_{-1}^{1} w^2(s)ds, & \lambda \in \{0, \pi\}, \\ f^2(\lambda) \int\limits_{-1}^{1} w^2(s)ds, & 0 < \lambda < \pi. \end{cases}$$

Examples Consider w being the uniform kernel on $(-1, 1)$.

$$w(x) = \begin{cases} 1, & |x| \leqslant 1, \\ 0, & |x| > 0, \end{cases}$$

$$W_r(\omega) = \frac{1}{2\pi} \sum_{|h|\leqslant r} e^{-ih\omega} = \frac{1}{2\pi}\, \frac{\sin\left((r + 1/2)\omega\right)}{\sin\left(\omega/2\right)}.$$

Spectral window obtained in this case is called Dirichlet kernel. It follows from Proposition 10.4.1 that

$$\mathrm{Var}\widehat{f}(\lambda) \;\sim\; \frac{2r}{n}\; f^2(\lambda), \qquad 0 < \lambda \leqslant \pi.$$

We consider now triangular, or the Bartlett window

$$w(x) = \begin{cases} 1 - |x|, & |x| \leqslant 1, \\ 0, & \text{w otherwise}, \end{cases}$$

Some algebraic manipulations yield

$$W_r(\lambda) = \frac{1}{2\pi r}\; \frac{\sin^2\left(r\lambda/2\right)}{\sin^2\left(\lambda/2\right)}, \tag{10.13}$$

Spectral window is thus the Féjer kernel (cf Problem 6.8) in this case. From the Proposition 10.4.1 we have

$$\mathrm{Var}\widehat{f}(\lambda) \;\sim\; \frac{r}{n}\; f^2(\lambda) \int w^2(s)ds = \; \frac{2r}{3n}\; f^2(\lambda).$$

Thus we obtain smaller asymptotic variance for triangular window than for rectangular one.

10.5 Problems

1. Consider the process $X_t = \mu + A\cos\lambda_k t + B\sin\lambda_k t + Z_t$, where Z_t is a Gaussian white noise, where A i B are constants and Fourier frequency $\lambda_k = 2\pi k/n \in (0,\pi)$ is known. We observe values X_t for $t = 1, 2, \ldots, n$ for which a linear model with explanatory variables $\mathbf{c}_k$ and $\mathbf{s}_k$ defined in Section 10.2 is fitted. We want to test whether a periodic component is significant i.e. we want to test H_0: $A = B = 0$. Express statistics $F = (SSR/2)/(SSE/(n-3))$ in terms of periodogram and using the results of this chapter show that under H_0 it has F Snedecor distribution with parameters $(2, n-3)$.
Hint. Note that $SSR = ||P_{sp(\mathbf{c}_k,\mathbf{s}_k)}\mathbf{X}||^2 = 2I(\lambda_k) \sim \sigma^2\chi^2(2)$, and $SSE = ||\mathbf{X} - P_{sp(\mathbf{1}_n,\mathbf{c}_k,\mathbf{s}_k)}\mathbf{X}||^2$, where $\mathbf{1}_n = n^{-1/2}(1,\cdots,1)'$, equals $||\mathbf{X}||^2 - I(0) - 2I(\lambda_k) \sim \sigma^2\chi^2(n-3)$.
2. Check validity of (10.13).
3. In connection with derivation in Section 10.2 check that for independent standard normal variables Z_1 and Z_2, $W = (Z_1^2 + Z_2^2)/2$ has exponential distribution with parameter 1.
4. Consider Blackman-Tuckey weight function $w_a(x) = (1 - 2a + 2a\cos x)I\{|x| \leqslant 1\}$ where $a > 0$. Find the corresponding spectral window.
5. Let (ε_t) be a white noise $(0, \sigma^2)$ and,κ_4 stand for fourth cumulant. Prove that for $\lambda \in \{0, \pi\}$

$$\mathrm{Var} I_n(\lambda) = \frac{\sigma^4}{2\pi^2} + \frac{\kappa_4}{4\pi^2 n} + o(\frac{1}{n}).$$

11

Nonlinear processes ARCH and GARCH

We introduce and discuss here two important classes of conditionally heteroscedastic nonlinear time series: ARCH and GARCH processes. First, however, we review some properties of finacial indices which were motivation of introducing such classes.

11.1 Returns of financial indices and stylized facts about them

In this section we define returns of financial indices and discuss some stylized facts about them which provide motivation for defining important classes of nonlinear processes ARCH and GARCH. Exhaustive treatment of GARCH modelling is given in Francq and Zakoian (2010), we also refer to Taylor (2005) and Ruppert (2011).

11.1.1 Financial returns

Let Y_t denote value of some financial index recorded at time t.

Definition 26 *A simple return of Y_t at the moment t is defined as*

$$\tilde{R}_t = \frac{Y_t - Y_{t-1}}{Y_{t-1}}$$

Thus the simple return is a change of Y_t, $\triangle Y_t = Y_t - Y_{t-1}$ relative to Y_{t-1}. Much more frequently used return is a logarithmic return (for the unit time interval)

Definition 27 *The logarithmic return of Y_t at the moment t is defined as*

$$R_t = \log(Y_t/Y_{t-1}) = (1-B)\log Y_t = \triangle \log Y_t = \log(1 + \frac{Y_t - Y_{t-1}}{Y_{t-1}}) \approx \tilde{R}_t,$$

where the last approximate equality holds when $\tilde{R}_t$ is small.

Analogously, we define a logarithmic return for time interval h as $R_{t,h} = \log(Y_t/Y_{t-h})$. There are several useful properties of logarithmic returns which explain why they are preferred to simple returns. First, if we sum up the logarithmic

returns for the consecutive unit time intervals we obtain the logarithmic return for the time interval h. Indeed,

$$R_t + \ldots + \ldots R_{t-h+1} = \log(Y_t/Y_{t-h}).$$

Moreover, logarithmic returns are symmetric in the sense that occurrence of consecutive positive and negative returns of the same absolute value means that the index returned to its former value. For example, if

$$R_t = \log(\frac{Y_t}{Y_{t-1}}) = 0,5$$

$$R_{t+1} = \log(\frac{Y_{t+1}}{Y_t}) = -0,5$$

then $Y_{t+1} = \exp(-0,5)Y_t = \exp(-0,5)\exp(0,5)Y_{t-1} = Y_{t-1}$.
However, this is not true for simple returns as if

$$\tilde{R}_t = \frac{(Y_t - Y_{t-1})}{Y_{t-1}} = 0.5$$

$$\tilde{R}_{t+1} = \frac{(Y_{t+1} - Y_t)}{Y_t} = -0.5,$$

then $Y_{t+1} = 0.5Y_t = 0,5 \times 1,5Y_{t-1} = 0.75Y_{t-1}$.
Consider now a classical hypothesis H on returns stating that (R_t) is a sequence of independent random variables having mean μ and variance σ^2. It is called the random walk hypothesis.
Although the random walk hypothesis turned out to be an oversimplified assumption about returns, it is interesting to see what are its consequences. Consider the yearly return R on some index performing from moment n to moment $n+N$ (usually N is considered equal to $21 \times 12 = 252$ (time is calculated in business days on a stock-exchange). R is thus the simple return for the horizon N :

$$R = \frac{Y_{n+N} - Y_n}{Y_n} = \frac{Y_{n+N}}{Y_n} - 1 = \exp\{\sum_{h=1}^{N} R_{n+h}\} - 1.$$

We want to approximate expected value ER using random walk hypothesis. We apply Central Limit Theorem to (R_t) and have that

$$\sum_{i=1}^{N} R_{n+h} \quad \text{has w approximately distribution } \mathrm{N}(\mathrm{N}\mu, \mathrm{N}\sigma^2),$$

and precisely this distribution if the returns are normal. Thus given that H is satisfied $R+1$ has approximately lognormal distribution and in view of the properties of the lognormal

$$ER \approx \exp(N\mu + \frac{1}{2}N\sigma^2) - 1.$$

Thus ER may be estimated by

$$\widehat{ER} = \exp(N\bar{R} + \frac{1}{2}NS^2) - 1,$$

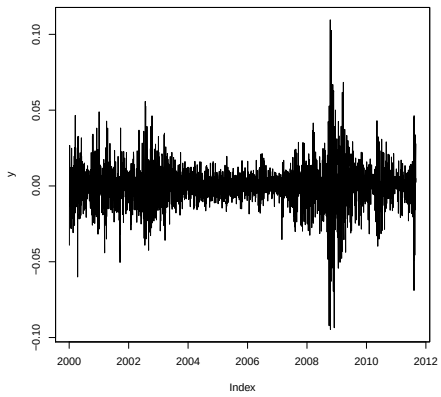
where $\bar{R}$ is the empirical mean of R_{n+h}, and S^2 its empirical variance.
Taylor (2005) uses this approach to calculate return on cacao for the period of 1971-1980 ($N = 2441$ observations). He finds that $10^4\bar{R} = 9.57$, $10^2S^2 = 2.03$, and thus according to the last equality $\widehat{ER} = 0.34\%$.

11.1.2 Stylized properties of financial returns

Analysis of financial data leads to formulation of three stylized properties on financial indices:

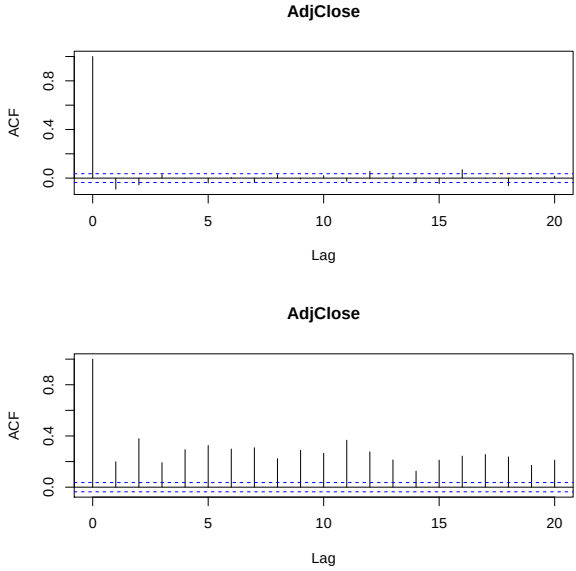
- Tails of distributions of financial returns R_t decrease more slowly than that of $N(0, 1)$,
- R_t are uncorrelated whereas R_t^2 are correlated,
- Large changes of consecutive values of R_t frequently follow previous large changes (volatility clustering).

Note that the second and the third of the stylized facts directly contradict the random walk hypothesis.
Let us define volatility as some measure of variability of R_t, usually a standard deviation, either unconditional or conditional one given the past of the process. The plot below shows logarithmic returns of Y being $S\&P500$ index from the begining of this century to August 31, 2012 (based on adjusted closing prices). Volatility clustering is evident.

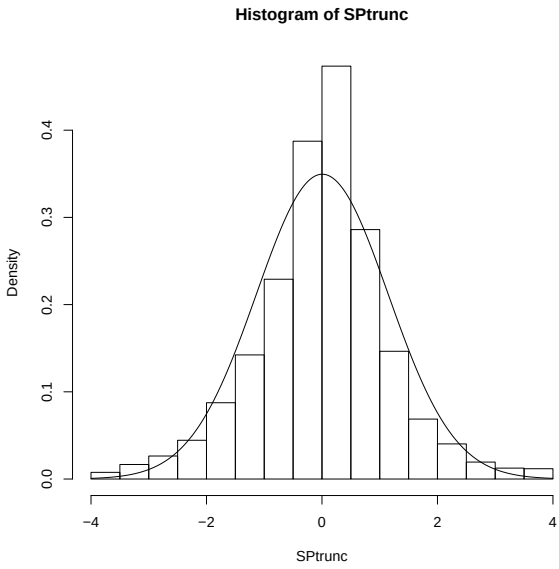


The second stylized fact asserts that returns are nonlinearly dependent as in the following example. If we take random variable X symmetric with respect to 0

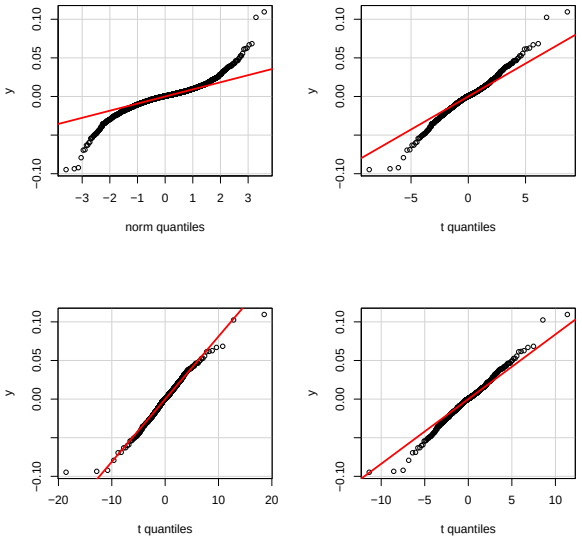
and consider pair of rvs: X and $Y = X^2$ then they are evidently dependent but uncorrelated: $\rho(X, Y) = 0$ as $\mathrm{Cov}(X, Y) = EX \times X^2 - EXEX^2 = 0$. In such situations the correlation coefficient will not detect their dependence.
A similar phenomenon is observed for financial returns. Consider the plots of autocorrelations for $Y = S\&P500$ and their squares below. The bands on the plots correspond to confidence intervals for ACFs of white noise. ACF for returns suggests that they are uncorrelated. However, persistent dependence for squared returns is visible on the second plot.



Consider now the first stylized fact in the case $S\&P500$ index. For this time series $\min R_t = -23\%$. Under normality, and considering the empirical standard deviation as an adequate estimator of σ, probability that such value or larger is observed is $2,23 \times 10^{-97}$ and would happen once in 10^{95} years. This strongly contradicts normality of financial returns. Below we show the histogram of $S\&P$ index (in percents and truncated at 4%) with superimposed normal density with the same empirical mean and the variance.



Note that a peak of returns' distribution is higher than the peak of the reference normal distribution whereas its tails are heavier than for the normal. Also the 'shoulders' of the empirical distribution are lower than for the reference distribution. The two first observations correspond to the fact that there are relatively more days with small and large volatility than predicted by the normal model. modelling of distribution of financial returns is a challenging task. Whether the proposed distribution is adequate or not may be also visualised with the aid of quantile plot with a respect to postulated distribution. Note that the most frequently used normal plot is the quantile plot with the respect to the normal distribution. Consider quantile plots of $S\&P$ with respect to the normal distribution and t distribution with number of degrees of freedom 3,4 and 5 respectively shown clockwise on the plot.



It is seen that in this case the best fit among four considered distributions is obtained for the t distribution with 4 degrees of freedom.

11.2 Nonlinear processes ARCH and GARCH

11.2.1 ARCH(1) processes

Observe that for (X_t) a causal AR(p) process we have that $\mathrm{Var}(X_t|X_s, s<t) = \sigma_\varepsilon^2$, thus the conditional variance given the past is constant and the process is conditionally homoscedastic.We define now the class of conditionally heteroscedastic processes i.e. processes for which the conditional variance depends on the previous values, which also have other properties exhibited by returns. We start with the simplest process of this kind, namely autoregressive conditionally heteroscedastic process of order 1, ARCH(1).

Definition 28 *$(X_t)_{t\in\mathbb{Z}}$ is ARCH(1) process when it satisfies the equations*

$$X_t = \sigma_t Z_t, \quad \sigma_t^2 = \alpha_0 + \alpha_1 X_{t-1}^2, \tag{11.1}$$

where Z_t is a strong WN$(0,1)$ such that Z_t is independent of $(X_s)_{s\leqslant t-1}$ and $\alpha_0, \alpha_1 \geqslant 0$.

Note that in contrast to ARMA(p,q) processes weak stationarity is *not* required in the definition of ARCH(1) and for general ARCH(p) proceses defined later.

A sufficient property for existence of stationary version of ARCH(1), namely that $\alpha_1 < 1$ is discussed below. We have the following two basic properties of ARCH(1) process:

$$E(X_t \mid X_s, s < t) = E(X_t \mid X_{t-1}) = \sigma_t E(Z_t \mid X_{t-1}) = \sigma_t E(Z_t) = 0$$

and

$$\mathrm{Var}(X_t \mid X_s, s < t) = \mathrm{Var}(X_t \mid X_{t-1}) = E(\sigma_t^2 Z_t^2 | X_{t-1}) = \sigma_t^2 E Z_t^2 = \alpha_0 + \alpha_1 X_{t-1}^2,$$

thus σ_t^2 in (11.1) is the conditional variance of X_t given the past of the process. This means, in view of definition (11.1) that the predictive dstribution of X_t i.e. conditional distribution of X_t given $X_s, s < t$ is a scaled distribution of Z_t where the scaling depends on the value of the last observation. We first indicate one of the most important properties of *weakly stationary* ARCH(1) process, namely that it is a weak white noise. Indeed, note that as $E(X_t \mid X_{t-1}) = 0$ we have that $EX_t = 0$. Similarly for $h > 0$ we have

$$\begin{aligned} E(X_t \cdot X_{t+h} \mid X_t \ldots, X_{t+h-1}) &= E(X_t \sigma_{t+h} Z_{t+h} \mid X_t, \ldots, X_{t+h-1}) \\ &= X_t (\alpha_0 + \alpha_1 X_{t+h-1}^2)^{1/2} E Z_{t+h} = 0 \end{aligned}$$

and taking expectations of both sides with respect to $X_t \ldots, X_{t+h-1}$ we obtain that

$$E(X_t \cdot X_{t+h}) = 0.$$

Note that we tacitly assumed that the conditional expectation $E(X_t \cdot X_{t+h} \mid X_t, \ldots, X_{t+h-1})$ exists which follows from the fact that $X_t \cdot X_{t+h}$ is integrable. It turns out that the condition

$$\alpha_1 < 1$$

is sufficient and necessary condition for existence of strictly stationary solution to (11.1) for which EX_t^2 is finite. Thus in this case (X_t) is a weak white noise (but obviously not a strong one as the process is conditionally heteroscedastic). We also discuss below a condition under which EX_t^4 is finite. Why such a property is important ? Namely, we show that under this condition X_t^2 is AR(1) process, more specifically it is a weakly stationary process satisfying

$$X_t^2 - \alpha_0 - \alpha_1 X_{t-1}^2 = \varepsilon_t, \tag{11.2}$$

where (ε_t) is a weak white noise. Indeed, note that in view of the definition of the process we have $X_t^2 - \alpha_0 - \alpha_1 X_{t-1}^2 = X_t^2 - \sigma_t^2$ and whence $\varepsilon_t = \sigma_t^2 (Z_t^2 - 1)$. We check that (ε_t) is indeed a weak white noise. We have

$$E\varepsilon_s = E(E(\varepsilon_s \mid X_w, w < s)) = E(\sigma_s^2 E(Z_s^2 - 1)) = 0,$$

and analogously

$$E(\varepsilon_s \cdot \varepsilon_t) = E(E(\varepsilon_s \varepsilon_t \mid X_w, w < t, Z_w, w < t)) = 0, \qquad s < t,$$

where in the last equality we used the property that Z_t is independent of $X_s, s < t$. Finiteness of EX_t^4 was needed to ensure existence of $E(\varepsilon_s\varepsilon_t \mid X_w, w < t, Z_w, w < t)$. Note that from (11.2) it follows that provided $\alpha_1 < 1$ we have that

$$EX_t^2 = \alpha_0/(1-\alpha_1)$$

(for $\alpha_0 = 0$ we have that $X_t \equiv 0$ for all $t \in \mathbb{Z}$). As $EX_t =$ this implies

$$\mathrm{Var}(X_t) = \alpha_0/(1-\alpha_1).$$

Note also that for $k > 0$ we have

$$\begin{aligned}
&\mathrm{Var}(X_{t+k}|X_{t-i}, i \geqslant 0) = E(\sigma_{t+k}^2 Z_{t+k}^2 | X_{t-i}, i \geqslant 0) \\
&= \alpha_0 + \alpha_1 E(X_{t+k-1}^2 Z_{t+k}^2 | X_{t-i}, i \geqslant 0) \ldots = \alpha_0 + \alpha_0\alpha_1 \\
&+ \ldots + \alpha_0\alpha_1^{k-1} + \alpha_1^k E(Z_{t+k}^2 Z_{t+k-1}^2 \ldots Z_t^2 X_t^2 | X_{t-i}, i \geqslant 0) = \\
&\frac{\alpha_0(1-\alpha_1^k)}{(1-\alpha_1)} + \alpha_1^k X_t^2,
\end{aligned} \tag{11.3}$$

which shows that a large absolute value of X_t will influence variability of X_{t+k} for several steps k ahead. Note that from the above reasoning it follows that

$$\mathrm{Var}(X_{t+k}|X_t) = \mathrm{Var}(X_{t+k}|X_{t-i}, i \geqslant 0).$$

In order to ensure that $EX_t^4 < \infty$ it is sufficient to assume that

$$EZ_t^4 < \infty \qquad \text{and} \qquad \max(1, (EZ_t^4)^{1/2}\alpha_1) < 1,$$

We remark that that when the above condition is satisfied X_t^2 is weakly stationary AR(1) process and moreover in view of the expression for EX_t^2 we can write (11.2) in the form

$$X_t^2 - EX_t^2 = \alpha_1(X_{t-1}^2 - EX_{t-1}^2) + \varepsilon_t.$$

Thus it follows from the properties of AR(1) time series that

$$\rho_{X^2}(h) = \alpha_1^{|h|}.$$

Note that for Gaussian N(0,1) noise we have that $EZ_t^4 = 3$ and the condition on finiteness of EX_t^4 states that $3\alpha_1^2 < 1$. We calculate kurtosis of X_t under this condition and show that it is larger than 3, which indicates that the marginal distribution of X_t is *not* Gaussian. Indeed, multiplying both sides of (11.2) by X_t^2 and taking expectation of both sides yields

$$EX_t^4 = \alpha_0 EX_t^2 + \alpha_1 E(X_{t-1}^2 X_t^2) + E(X_t^2\varepsilon_t). \tag{11.4}$$

Using independence of Z_t and X_{t-1} it follows that $E(X_t^2\varepsilon_t) = E\varepsilon_t^2$. Moreover, as $\alpha_0 = (1-\alpha_1)EX_t^2$ and $E(Z_t^2 - 1)^2 = 2$ we have

$$\begin{aligned} E\varepsilon_t^2 &= E((Z_t^2-1)^2(\alpha_0+\alpha_1 X_{t-1}^2)^2) \\ &= 2[(1-\alpha_1)^2(EX_t^2)^2+\alpha_1^2 EX_t^4+2\alpha_1(1-\alpha_1)(EX_t^2)^2] \end{aligned}$$

and

$$\alpha_1 E(X_{t-1}^2 X_t^2) = (1-\alpha_1)(EX_t^2)^2+\alpha_1 EX_t^4.$$

Plugging the last two equations into (11.4) after some easy calculations and using $\alpha_0 = (1-\alpha_1)EX_t^2$ again we obtain that

$$(1-3\alpha_1^2)EX_t^4 = 3(1-\alpha_1^2)(EX_t^2)^2$$

and whence

$$\kappa_X = \frac{EX_t^4}{(EX_t^2)^2} = \frac{3(1-\alpha_1^2)}{(1-3\alpha_1^2)} > 3 = \kappa_Z.$$

Thus the tails of distribution of X_t are heavier than the tail of normal Z_t.

11.2.2 ARCH(p) processes

We define now the class of ARCH processes of an arbitrary order p introduced by Engle (1982) to model variability of inflation rates in the U.K.

Definition 29 *We say that (X_t) is an ARCH(p) process if*

$$X_t = \sigma_t Z_t, \qquad \sigma_t^2 = \alpha_0+\alpha_1 X_{t-1}^2+\cdots+\alpha_p X_{t-p}^2, \qquad \alpha_0,\alpha_1,\ldots,\alpha_p \geqslant 0, \quad (11.5)$$

where Z_t is strong WN$(0,1)$ and Z_t is independent of X_s, $s<t$.

As before by conditioning we prove that

$$EX_t = 0$$

and

$$\mathrm{Var}(X_t \mid X_s, s<t) = \alpha_0+\alpha_1 X_{t-1}^2+\cdots+\alpha_p X_{t-p}^2.$$

Thus the conditional variance $\mathrm{Var}(X_t \mid X_s, s<t)$ is an affine combination of the squares of p previous values of the process with non-negative coefficients.
We state now the theorem which specifies the conditions under which a strictly stationary version of (X_t) with finite second moment exists and when its fourth moment is finite.

Theorem 11.2.1 *(i) Sufficient and necessary condition for existence of strictly stationary time series (X_t) satisfying (11.5) such that $EX_t^2 < \infty$ is*

$$\sum_{j=1}^{p} \alpha_j < 1.$$

(it is also necessary condition for existence of weakly stationary solution). Moreover, we then have

$$EX_t = 0, \qquad EX_t^2 = \frac{\alpha_0}{(1 - \sum_{i=1}^{p} \alpha_i)} \tag{11.6}$$

and $X_t \equiv 0$ *if* $\alpha_0 = 0$.
(ii) If $EZ_t^4 < \infty$ *and*

$$\max(1, (EZ_t^4)^{1/2}) \sum_{j=1}^{p} \alpha_j < 1, \tag{11.7}$$

then fourth moment $EX_t^4 < \infty$.

Below we list several properties of ARCH(p) time series which follow from the result.
1) Under assumptions of Theorem 11.2.1(i) weakly stationary process ARCH(p) is a weak WN$(0, \sigma_X^2)$, where $\sigma_X^2 = \alpha_0/(1 - \sum_{i=1}^{p} \alpha_i)$.
2) Moreover, note that we can write

$$X_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i}^2 + \varepsilon_t, \tag{11.8}$$

where

$$\varepsilon_t = (Z_t^2 - 1)\{\alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i}^2\}. \tag{11.9}$$

Under assumptions of Theorem 11.2.1(ii) (ε_t) is WN$(0, \sigma_\varepsilon^2)$, where

$$\sigma_\varepsilon^2 = \mathrm{Var}(Z_t^2) \cdot E\{\alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i}^2\}^2.$$

Also note that in view of (11.6) structural equation (11.8) for X_t^2 can be written as

$$X_t^2 - EX_t^2 = \sum_{i=1}^{p} \alpha_i (X_{t-i}^2 - EX_{t-i}^2) + \varepsilon_t,$$

thus $(X_t^2)_{t \in \mathbb{Z}}$ is AR(p) process. Moreover,
3) If condition (11.7) holds then X_t^2 is a causal process AR(p). This property follows from the last equation after noting that

$$\sup_{|z| \leqslant 1} \Big| \sum_{j=1}^{p} \alpha_j z^j \Big| \leqslant \sum_{j=1}^{p} |\alpha_j| < 1$$

for $|z| \leqslant 1$, thus

$$\varphi(z) = 1 - \sum_{j=1}^{p} \alpha_j z^j > 0, \qquad |z| \leqslant 1.$$

4) If $E(X_t^2 \mid X_{t-1}, \dots, X_{t-p})$ is not constant, or, equivalently, $\sum_{i=1}^p \alpha_i > 0$, then

$$\kappa_X > \kappa_Z.$$

Indeed,

$$\begin{aligned} E(X_t^4 \mid X_{t-1}, \dots, X_{t-p}) &= \sigma_t^4 E Z_t^4 \\ &= \kappa_Z \sigma_t^4 (E Z_t^2)^2 = \kappa_Z \{E(X_t^2 \mid X_{t-1}, \dots, X_{t-p})\}^2. \end{aligned}$$

Thus

$$EX_t^4 = \kappa_Z E(E(X_t^2 \mid X_{t-1}, \dots, X_{t-p}))^2 \geqslant \kappa_Z (EX_t^2)^2$$

The last inequality follows from $EY^2 \geqslant (EY)^2$ applied to $Y = E(X_t^2 \mid X_{t-1}, \dots, X_{t-p})$, note that equality holds only for $Y \equiv c$ a.s.
Finally we define ARCH(∞) process as follows: $Y_t = \rho_t \xi_t$ is ARCH(∞) if ξ_t is a sequence of nonnegative iid random variables with mean 1 and

$$\rho_t = \alpha_0 + \sum_{i=1}^{\infty} \alpha_i Y_{t-i},$$

where $\alpha_i \geqslant 0$ for $i \geqslant 0$. Note that for any ARCH(p) its square satisfies the conditions for ARCH(∞) process and additionally $\alpha_j = 0$ for $j \geqslant p$. We will prove that stationary version of ARCH(∞) process exists provided $\sum_{j=1}^{\infty} \alpha_j < 1$. This will be used to check stationarity conditions of GARCH(p, q) process in the next section. Namely, using the definition of Y_t recursively we obtain

$$\begin{aligned} Y_t &= \alpha_0 \xi_t + \sum_{i=1}^{\infty} \alpha_i \xi_t \xi_{t-i} \rho_{t-i} = \alpha_0 \xi_t + \alpha_0 \sum_{i=1}^{\infty} \alpha_i \xi_t \xi_{t-i} + \sum_{i,j=1}^{\infty} \alpha_i \alpha_j \xi_t \xi_{t-i} Y_{t-i-j} \\ &= \alpha_0 \xi_t + \alpha_0 \sum_{l=1}^{k} \sum_{1 \leqslant j_1, \dots, j_l} \alpha_{j_1} \cdots \alpha_{j_l} \xi_t \xi_{t-j_1} \cdots \xi_{t-j_1-\cdots j_l} \\ &+ \sum_{1 \leqslant j_1, j_2 \dots, j_{k+1}} \alpha_{j_1} \cdots \alpha_{j_{k+1}} \xi_t \xi_{t-j_1} \dots \xi_{t-j_1-\cdots j_k} Y_{t-j_1-\cdots-j_{k+1}} \end{aligned}$$

We define now Y_t' by omitting the last term in the decomposition of Y_t and replacing k by ∞ , namely

$$Y_t' = \alpha_0 \xi_t + \alpha_0 \sum_{l=1}^{\infty} \sum_{1 \leqslant j_1, \dots, j_l} \alpha_{j_1} \cdots \alpha_{j_l} \xi_t \xi_{t-j_1} \cdots \xi_{t-j_1-\cdots j_l}$$

Observe that as the summands of the inner sum consist of products of independent random variables with mean 1, we have that

$$E\Big(\sum_{1 \leqslant j_1, \dots, j_l} \alpha_{j_1} \cdots \alpha_{j_l} \xi_t \xi_{t-j_1} \dots \xi_{t-j_1-\cdots j_l} \Big) = \sum_{1 \leqslant j_1, \dots, j_l} \alpha_{j_1} \cdots \alpha_{j_l} = \Big(\sum_{j=1}^{\infty} \alpha_j \Big)^l$$

Thus in view of the condition on α_j we have that $E(Y_t') = \alpha_0/(1 - \sum_{j=1}^{\infty} \alpha_j)$ and since Y_t' is nonegative it follows that it is finite almost surely. In view of its definition Y_t' is stationary and it is easily seen to satisfy definition of ARCH(∞). Moreover, it can be shown that Y_t' is the unique stationary solution.

11.2.3 GARCH(p, q) process

We consider now the most natural, and also most commonly used, generalization of ARCH(p) process which consists in incorporation of moving average part into the definition of conditional variance of X_t. It was introduced independently by T. Bollerslev and S. Taylor in 1986.

Definition 30 *$(X_t)_{t\in\mathbb{Z}}$ is called GARCH(p, q) process (Generalized ARCH) if*

$$X_t = \sigma_t Z_t, \tag{11.10}$$

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2, \tag{11.11}$$

where, as before, Z_t is independent of $X_s, s < t$, $p, q \in \mathbb{N}$, $p \geqslant 1$ and all coefficients $\alpha_0, \alpha_i, \beta_j$ are nonnegative.

Note that for $q = 0$, the process reduces to ARCH(p). By conditioning we have that

$$\mathrm{Var}(X_t | X_{t-s}, \sigma_{t-s}^2, s > 0) = \sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2,$$

and depends linearly not only on previous squared values of the process, as in the case of ARCH(p), by also on q previous values of the conditional variances. Moreover, note that it follows from the definition of σ_t^2 that σ_{t-j}^2 for $j > 0$ is a measurable function of X_{t-k}^2 with $k \geqslant j$ (cf. equality 11.16) below) and thus Z_t is independent of σ_{t-j}^2 . We note that σ_t^2 does not depend on the sign of X_{t-i}. For nonsymmetric modelling of volatility we refer to Nelson (1990).

As before, we do not require that (X_t) is weakly stationary. Existence of stationary solutions is dealt with in the next result.

Theorem 11.2.2 *(i) Strictly stationary solution of (11.10) with a finite second moment exists if and only if*

$$\sum_{i=1}^{p} \alpha_i + \sum_{j=1}^{q} \beta_j \ < \ 1. \tag{11.12}$$

Moreover, in this case X_t is a weak WN$(0, \sigma_X^2)$, where

$$\sigma_X^2 = \frac{\alpha_0}{1 - \sum_{i=1}^{p} \alpha_i - \sum_{j=1}^{q} \beta_j}. \tag{11.13}$$

(ii) If

$$\max\{1,(EZ_t^4)^{1/2}\}\big(\sum_{j=1}^{p}\alpha_j/(1-\sum_{j=1}^{q}\beta_j)\big) \;<\; 1 \tag{11.14}$$

then $EX_t^4<\infty$.

We discuss now why the condition $\sum_{i=1}^{p}\alpha_i+\sum_{j=1}^{q}\beta_j \;<\; 1$ stated in (11.12) is sufficient for existence of weakly stationary solution. We can write

$$(1-\sum_{j=1}^{q}\beta_jB^j)\sigma_t^2=\alpha_0+\sum_{i=1}^{p}\alpha_iB^iX_t^2.$$

As the condition (11.12) in particular implies that $\sum_{j=1}^{q}\beta_j \;<\; 1$ and thus autoregressive polynomial $1-\sum_{j=1}^{q}\beta_jz^j$ does not have zeros in the unit disc, we have

$$\sigma_t^2=(1-\sum_{j=1}^{q}\beta_jB^j)^{-1}\Big\{\alpha_0+\sum_{i=1}^{p}\alpha_iB^iX_t^2\Big\}=d_0+\sum_{i=1}^{\infty}d_iX_{t-i}^2, \tag{11.15}$$

where $(1-\sum_{j=1}^{q}\beta_jB^j)d_0=\alpha_0$ and thus $d_0=\alpha_0/(1-\sum_{j=1}^{q}\beta_j)$. Moreover, d_i for $i\geqslant 1$ satisfy the equation $\sum_{i=1}^{\infty}d_iz^i=\sum_{i=1}^{p}\alpha_iz^i/(1-\sum_{i=1}^{q}\beta_iz^i)$. Taking $z=1$ we obtain

$$\sum_{i=1}^{\infty}d_i=\sum_{i=1}^{p}\alpha_i/(1-\sum_{i=1}^{q}\beta_i).$$

Also $d_i=\beta_i+\sum_{k=1}^{i-1}\alpha_id_{i-k}$, where undefined β_i and α_j are taken equal 0, from which it follows that $d_i\geqslant 0$ for all i. It follows that (X_t^2) has representation as ARCH(∞) and is strictly stationary in view of the last comment of the previous section provided that $\sum_{i=1}^{\infty}d_i<1$ which is equivalent to the condition stated in the theorem.

We also note that by induction it can be proved that

$$\begin{aligned}\sigma_t^2=&\frac{\alpha_0}{(1-\sum_{j=1}^{q}\beta_j)}+\sum_{i=1}^{p}\alpha_iX_{t-i}^2+\\&+\sum_{i=1}^{p}\alpha_i\sum_{k=1}^{\infty}\sum_{j_1=1}^{q}\cdots\sum_{j_k=1}^{q}\beta_{j_1}\cdots\beta_{j_k}X_{t-i-j_1-\cdots-j_k}^2\end{aligned} \tag{11.16}$$

Now we establish an important relation of GARCH processes with ARMA(p,q) processes. Note that

$$X_t^2=\alpha_0+\sum_{i=1}^{p}\alpha_iX_{t-i}^2+\sum_{j=1}^{q}\beta_j\sigma_{t-j}^2+\varepsilon_t,$$

where

$$\varepsilon_t = X_t^2 - \sigma_t^2$$

Thus substituting $\varepsilon_{t-j} = X_{t-j}^2 - \sigma_{t-j}^2$, we obtain

$$X_t^2 = \alpha_0 + \sum_{i=1}^{p\vee q}(\alpha_i + \beta_j)X_{t-j}^2 + \varepsilon_t - \sum_{j=1}^{q}\beta_j\varepsilon_{t-j},$$

where $\alpha_{p+j} = \beta_{q+j} = 0$, for $j \geqslant 1$ and $p \vee q = \max(p, q)$. Under conditions of Theorem 11.2.2 (ii) we prove that (ε_t) is weak white noise and thus X_t^2 is ARMA$(p \vee q, q)$, which is causal nad invertible due to the condition $\sum_{i=1}^{p}\alpha_i + \sum_{j=1}^{q}\beta_j < 1$.

Usually in practice when modelling autocovariance structure of (X_t^2) with the use of ARCH(p) process, we need large order of this process. Hovever, as causal ARMA process is AR(∞) and we know that (X_t^2) is ARMA when (X_t) is an GARCH process, one can try to use instead of ARCH process of large order GARCH(p, q) process with moderate p and q. Actually, very frequently GARCH(1,1) process yields adequate approximation of financial time series.

We have proven above that stationary ARCH(p) process has kurtosis $\kappa_X > \kappa_Z$ provided its conditional variance is not constant. This extends to GARCH processes by conditioning with respect to all lagged X_{t-i} in the proof instead of p last values.

Case of GARCH(1,1). We now consider in a greater detail the case of GARCH(1,1) time series. Theorem 11.2.2 gives sufficient and necessary condition for existence of strictly stationary GARCH(p, q) process with the finite second moment. We now study the question of existence of strictly GARCH(1,1) with no condition on its second moment imposed. Let $a(z) = \alpha_1 z^2 + \beta_1$.

Theorem 11.2.3 *If*

$$\gamma = E(\log(\alpha_1 Z_t^2 + \beta_1)) = E(\log a(Z_t)) < 0 \tag{11.17}$$

then

$$h_t = \alpha_0(1 + \sum_{i=1}^{\infty} a(Z_{t-1})\dots a(Z_{t-i}))$$

converges and is finite almost surely. The process $X_t = \sqrt{h_t}Z_t$ *is the unique stationary GARCH(1,1) process. If* $\gamma \geqslant 0$ *and* $\alpha_0 > 0$ *then no stationary solution exists.*

Note that if $\alpha_1 + \beta_1 < 1$ holds which is sufficient condition for existence of strictly stationary solution with the finite second moment, we have in view of the concavity of the logarithmic function that

$$E(\log a(Z_t)) \leqslant \log(E(a(Z_t))) = \log(\alpha_1 + \beta_1) < 0$$

thus the condition (11.17) is weaker than $\alpha_1 + \beta_1 < 1$. We also note that in contrast to the condition $\alpha_1 + \beta_1 < 1$, (11.17) is not symmetric in α_1 and β_1 and

depends on the distribution of ε_1.
Proof.We will prove that condition (11.17) is sufficient for existence of stationary solution. The key observation here is to note that for the GARCH(1,1) processes definiton of σ_t^2 can be written in the form

$$\sigma_t^2 = \alpha_0 + a(Z_{t-1})\sigma_{t-1}^2$$

and iterating we obtain that σ_t^2 has to satisfy the following equation

$$\sigma_t^2 = \alpha_0(1 + \sum_{i=1}^{N} a(Z_{t-1}) \cdots a(Z_{t-i})) + a(Z_{t-1}) \cdots a(Z_{t-N})\sigma_{t-N-1}^2$$

Let $h_t(N) = \alpha_0(1 + \sum_{i=1}^{N} a(Z_{t-1}) \cdots a(Z_{t-i}))$ and note that when $N \to \infty$ $h_t(N)$ converges to h_t almost surely as summands of the sum are nonnegative. Thus it is natural to consider h_t as a candidate for σ_t^2. We will prove that under condition (11.17) h_t is finite almost surely. We check the Cauchy rule for series with nonnegative summands. We have

$$[a(Z_{t-1}) \cdots (Z_{t-n})]^{1/n} = \exp\{\frac{1}{n}\sum_{i=1}^{n} \log a(Z_{t-i})\} \to e^{\gamma} < 1$$

almost surely in view of strong law of large numbers. Thus h_t is finite almost surely (by the same token for $\gamma > 1$ we show that $h_t = \infty$ almost surely and thus stationary solution does not exist in this case). Note, moreover, that the definition of h_t yields

$$h_t = \alpha_0 + h_{t-1}a(Z_{t-1}).$$

Thus *defining* $X_t = \sqrt{h_t}Z_t$ in view of the last equality we have that

$$\begin{aligned} X_t &= (\alpha_0 + h_{t-1}a(Z_{t-1}))^{1/2}Z_t = (\alpha_0 + \alpha_1 h_{t-1}Z_{t-1}^2 + \beta_1 h_{t-1})^{1/2}Z_t \\ &= (\alpha_0 + \alpha_1 X_{t-1}^2 + \beta_1 h_{t-1})^{1/2}Z_t \end{aligned}$$

which yields the structural equation of GARCH(1,1) process with $\sigma_t^2 = h_t$.
In order to check uniqueness, assume that $\tilde{X}_t = \tilde{\sigma}_t\eta_t$ is another solution. Then reasoning as before we obtain

$$\tilde{\sigma}_t^2 - h_t = (h_t(N) - h_t) + a(Z_{t-1}) \cdots a(Z_{t-N})\tilde{\sigma}_{t-N-1}^2$$

Note that as h_t converges to a finite limit a.s. thus $a(Z_{t-1}) \cdots a(Z_{t-N})$ converges to 0 for $N \to \infty$ and as in view of stationarity the distribution of σ_{t-N-1}^2 does not depend on N, the second term in the last equation tends to 0 and as the first term also tends to 0, the uniqueness follows.

11.3 Estimation for ARCH(p) and GARCH(p, q) processes

The most frequently used method of estimation is method of maximizing conditional likelihood. The reason for this is that by conditioning we express it in terms of menageable conditional densities and without relying on marginal density of X_t. In order to ilustrate the method we start with ARCH(1) process with Gaussian innovations. We have that the density $f_X(X_1, \ldots, X_n)$ of $(X_1, \ldots, X_n)$ can be written as

$$f_X(X_1, \ldots, X_n) = f(X_1) f(X_2 \mid X_1) \cdots f(X_n \mid X_1, \ldots, X_{n-1}).$$

Thus

$$f_X(X_1, \ldots, X_n \mid X_1) = f(X_2 \mid X_1) \cdots f(X_n \mid X_1, \ldots, X_{n-1}) = \prod_{t=2}^{n} \frac{1}{\{2\pi(\alpha_0 + \alpha_1 X_{t-1}^2)\}^{1/2}} \exp\big(-\frac{X_t^2}{2(\alpha_0 + \alpha_1 X_{t-1}^2)}\big). \tag{11.18}$$

Conditional ML estimator is defined as

$$(\hat{\alpha}_0, \hat{\alpha}_1) = \underset{\alpha_0, \alpha_1}{\operatorname{argmax}} f_X(X_1, \ldots, X_n \mid X_1).$$

This simple reasoning can be extended to ARCH(p) processes, and, with some modifications, to GARCH(p, q). If (X_t) is ARCH(p) with Gaussian innovations, we obtain in a similar manner for $n > p$ that

$$-2 \log f_X(X_1, \ldots, X_n \mid X_1, \ldots, X_p) = (n-p) \log(2\pi) + \sum_{t=p+1}^{n} \log \sigma_t^2 + \frac{X_t^2}{\sigma_t^2}, \tag{11.19}$$

where $\sigma_t^2 = \alpha_0 + \alpha_1 X_{t-1}^2 + \cdots + \alpha_p X_{t-p}^2$. Thus, as before treating loglikelihood in (11.19) as the function of $\alpha_0, \ldots, \alpha_p$ we look for its maximisers being conditional maximum likelihood estimators.

We can also use (11.19) for GARCH(p, q) series. The essential difference is now that in view of (11.16) σ_t^2 depends now on infinite number of previous X_{t-j}^2. As we have only $X_1, \ldots, X_n$ at our disposal, truncation of the infinite sum in (11.16) is necessary. Namely, we define

$$\bar{\sigma}_t^2 = \frac{\alpha_0}{(1 - \sum_{j=1}^{q} \beta_j)} + \sum_{i=1}^{p} \alpha_i X_{t-i}^2 + \\ + \sum_{i=1}^{p} \alpha_i \sum_{k=1}^{\infty} \sum_{j_1=1}^{q} \cdots \sum_{j_k=1}^{q} \beta_{j_1} \cdots \beta_{j_k} X_{t-i-j_1-\cdots-j_k}^2 I\{t - i - j_1 - \cdots - j_k \geqslant 1\}$$

for $t > p$ and we plug $\bar{\sigma}_t^2$ in the loglikelihood above in place of σ_t^2.

The method of conditional maximum likelihood can be extended to the case when Z_t have known, but not necessarily normal distribution. Then minimization of (11.19) is replaced by minimization of

$$\sum_{t=p+1}^{n} \log \sigma_t^2 - \log f(X_t/\sigma_t)$$

and in the case of GARCH(p, q) process σ_t is replaced by $\tilde{\sigma}_t$ defined as above. Apart from normal distribution, distributions with heavier tails than normal are often used as e.g. the t-distribution with ν degrees of freedom, when ν can be any number larger than 2, or Generalized Gaussian Distribution (GED).
Another possibility of estimating GARCH(p, q) parameters is provided by Whittle's estimator . It relies on representation (11.15). In view of it the squared process $Y_t := X_t^2 = \sigma_t^2 Z_t^2$ can be under assumptions of Theorem 11.12 (ii) represented as AR(∞) series

$$Y_t = \frac{\alpha_0}{1 - \sum_{j=1}^{q} \beta_j} + \sum_{j=1}^{\infty} d_j Y_{t-j} + \varepsilon_t,$$

where

$$\varepsilon_t = (Z_t^2 - 1)\Big\{ \frac{\alpha_0}{1 - \sum_{j=1}^{q} \beta_j} + \sum_{j=1}^{\infty} d_j Y_{t-j} \Big\}.$$

Then it follows that the spectral density f_Y of (Y_t) is

$$f_Y(\lambda) = \frac{\sigma_\varepsilon^2}{2\pi} \Big| 1 - \sum_{j=1}^{\infty} d_j e^{-ij\lambda} \Big|^2,$$

where $\sigma_\varepsilon^2 = \mathrm{Var}(\varepsilon_t)$. Note that σ_ε^2 and (d_i) are nonlinear functions of $\alpha_0, \alpha_1, \ldots, \alpha_p$ and $\beta_1, \ldots \beta_q$. Whittle's estimator choses parameters $\alpha_0, \alpha_1, \ldots, \alpha_p$ and $\beta_1, \ldots \beta_q$ which minimize Whittle's criterion function

$$W = \sum_{j=1}^{T-1} \frac{I_n(\lambda_j)}{f_Y(\lambda_j)},$$

see Giraitis and Robinson (2001).

11.3.1 Testing for ARCH(p)

Detecting heteroscedasticity in data is an important task as disregarding it may lead to errors in testing procedures, most frequently to overrejection of conventional tests. We mention only two tests for ARCH effects. The hypothesis we want to test is $H_0 : \alpha = (\alpha_1, \ldots, \alpha_p) = 0$ and H_1 its complement. The first is an application of the conditional likelihood ratio test and assumes that we know the density f of innovations. In this case suppose that $(\hat{\alpha}_0, \hat{\alpha})$ is unconstrained ML estimator of model parameters and let $\bar{\alpha}_0$ be ML estimator of α_0 under null hypothesis. Conditional LRT statistic is

$$LRT = 2\log\Big\{\prod_{t=p+1}^{n}\frac{\sigma_t(\hat{\alpha}_0,\hat{\alpha})^{-1}f(X_t/\sigma_t(\hat{\alpha}_0,\hat{\alpha}))}{\sigma_t(\bar{\alpha}_0,0)^{-1}f(X_t/\sigma_t(\bar{\alpha}_0,0))}\Big\}.$$

Under appropriate regularity conditions LRT has approximately chi squared distribution with p degrees of freedom. Two other tests, frequently used in similar contexts, namely the Wald test and the score test can be also used. The advantage of the score test is that the fitting of the ARCH(p) model is not necessary. It was shown by Engle (1982) that in the case of normal innovations the score test is equivalent to the test based on R^2 statistic. The latter uses the property that X_t^2 is AR(p) process provided that fourth moment of X_t exists. Thus we can fit the model

$$X_t^2 = \alpha_0 + \sum_{i=1}^{p}\alpha_i X_{t-i}^2 + Z_t,$$

using as estimates ML conditional likelihood estimators and calculate the squared coefficient of determination R^2 of the fit. Under H_0 R^2 has approximately chi squared distribution with p degrees of freedom.

11.3.2 Problems

1. Consider time series $(X_t)_{t\in\mathbb{Z}}$ given by

$$X_t = \sin(\varepsilon_t\pi)\sin(\varepsilon_{t-1}\pi),$$

where (ε_t) is a strong $WN(0,\sigma^2)$ such that for $t\in\mathbb{Z}$ ε_t is uniformly distributed on $[-1,1]$. Prove that $(X_t))_{t\in\mathbb{Z}}$ is a weak $WN(0,\sigma^2)$ and calculate σ^2.
2. Assume that (11.12) holds and check that $EX_t = 0$ and EX_t^2 satisfies $EX_t^2 = \alpha_0 + (\sum_{i=1}^{p}\alpha_i + \sum_{i=1}^{q}\beta_j)EX_t^2$ and thus (11.13) holds.
3. Show that under assumptions of Theorem 11.2.1 (i) the sequence (ε_t) defined in (11.9) is a martingale difference with respect to the whole past i.e. $E(X_t|X_{t-i}, i > 0) = 0$.
4. Check that under assumption (11.12) GARCH(p,q) is an invertible ARMA$(p\vee q, q)$ process.
5. Consider GARCH$(1,1)$ series and prove that under condition (11.14) we have

$$\mathrm{Var}(X_t|X_{t-i}, i\geqslant 1) = \frac{\alpha_0}{1-\beta_1} + \alpha_1\sum_{j=0}^{\infty}\beta_1^j X_{t-j-1}^2.$$

12

Long-range dependent time series

We review properties of stationary time series with non-summable covariances, known as long-range dependent or long memory processes. It turns out such processes arise smong others as increments of self-similar processes and the estimators of their characteristics exhibit different behaviour than in the case of weakly dependent data. Moreover, regression estimation when the regression function is contaminated with long-range dependent errors is discussed.

12.1 Strongly dependent processes

In many places in this book we assumed that covariances of the time series are absolutely summable: in particular this was the sufficient condition ensuring existence of the spectral density. Many time series models enjoy this feature, in particular for causal ARMA processes their autocovariance decays to 0 exponentially fast i.e. there exists $D > 0$ and $0 < r < 1$ such that

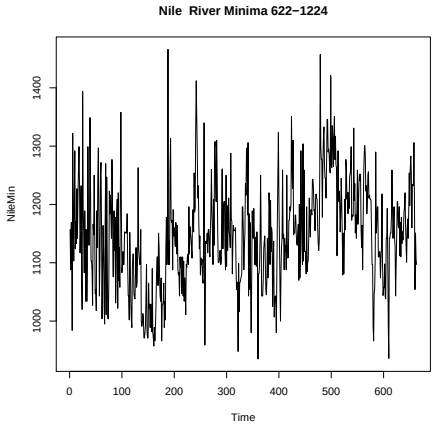
$$|\gamma(k)| \leqslant Dr^k.$$

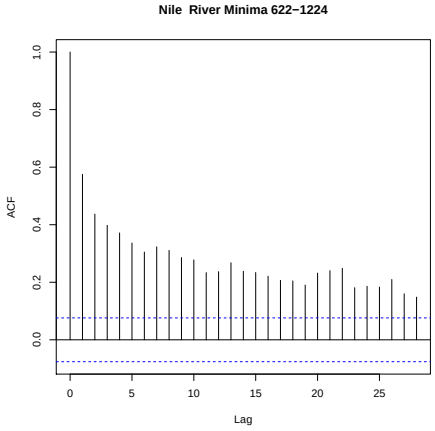
However, there are many examples of financial, hydrological, geophysical data and data in various different domains which indicate that strong dependence persists even for observations which are recorded at distant time points. Sample paths of such data frequently show many periods of apparent trends, which also may change their direction. Consider the classical data of minimal levels on the Nile river between 622-1224 analised by Hurst and Mandelbrot shown below. Autocovariances for Nile data decay quickly to around 0.5 and then the decay becomes very slow. This typical feature of long-ranging dependence between observations became frequently studied and data having this property is called *long-range dependent* or *long-memory* data. On the theorethical level long-range dependence is defined as follows.

Definition 31 *We call weakly stationary series long-range dependent (LRD) if*

$$\sum_{h=1}^{\infty} \gamma_X(h) = \infty, \tag{12.1}$$

that is if the sum of covariances diverges. In the case when the sum of covariances converges and its sum is strictly positive the sequence is said to be short-range dependent. The case when the sum of covariances is zero is called antipersistent.





Note that long-range dependence is one of many possible quantifications of strong dependence. We remark also that without some assumptions on regularity of the decay, condition (12.1) is practically unverifiable based on a finite part of trajectory as we may have long stretch of small and quickly decaying correlations followed by a very slow decay later on. That is why in the LRD models presented below some regularity of the decay of the autocovariance is imposed, most commonly that it is, up to a slowly varying factor, an hyperbolically decaying function.

We stress that it is one of the possible definitions of LRD. It makes a perfect sense to define LRD in terms of non-summable *partial* autocorrelations as in Dębowski (2007) or Inoue (2008). Alas, this theory is much less developed than for LRD in sense of (12.1).

We also note that frequently slightly modified definition of long-range dependence is used in which instead of condition (12.1) the condition of *absolute* summability of autocovariance is imposed. However, for models considered later,

autocovariance will have the form $\gamma(h) = L(h)h^{-\alpha}$, where $0 < \alpha < 1$ and $L(\cdot)$ is a slowly varying function. As $L(\cdot)$ has constant sign in infinity (cf. problem 12.6) there is no much difference between these two possible definitions.
There is an important link between long-range dependent sequences and self-similarity property which partly explains frequent occurrence of the former. We recall the following.

Definition 32 *A process $(Y_t)_{t\in\mathbb{R}}$ is self-similar with parameter H, where $0 < H < 1$, if*

$$(Y_{cu})_{u\in\mathbb{R}} \stackrel{\mathcal{D}}{=} c^H (Y_u)_{u\in\mathbb{R}},$$

where $\stackrel{\mathcal{D}}{=}$ denotes equality in distribution.

It is known that the limits of normalized partial sums of ergodic sequences $\sum_{k=1}^{[nt]} X_k$ are necessarily self-similar (the Lamperti theorem) which shows that such processes in exact or approximate form are ubiquitous. The definition also implies that the sample paths of the self-similar processes aggregated at different time scales should behave essentially the same. Such feature is observed e.g. for many characteristics of internet traffic (see e.g. Willinger et al. (1995)) and its connection with LRD for underlying phenomenon is intensively explored.
Now we show that the increments $X_k = Y_k - Y_{k-1}$ of the self-similar process with stationary increments and parameter H such that $1/2 < H < 1$ are LRD. Namely, it is easy to see that $(X_k)_{k\in\mathbb{N}}$ is weakly stationary series and as $\mathrm{Var}(Y_u) = u^{2H}\sigma^2$ for $u > 0$, where $\sigma^2 = \mathrm{Var}(Y_1)$ we have

$$\gamma_X(k) = \mathrm{Cov}((Y_{t+k} - Y_{t+k-1}), (Y_t - Y_{t-1})) = \frac{\sigma^2}{2}[|k+1|^{2H} + |k-1|^{2H} - 2|k|^{2H}].$$

Transforming it further

$$\gamma_X(k) = \frac{\sigma^2}{2} k^{2H-2} \frac{[|1+\frac{1}{k}|^{2H} + |1-\frac{1}{k}|^{2H} - 2]}{\frac{1}{k^2}} \sim \sigma^2 H(2H-1)k^{2H-2}, \quad (12.2)$$

where the equivalence follows from expanding x^{2H} around 1. Since for $1/2 < H < 1$ we have that $-1 < 2H-2 < 0$, (X_k) is LRD. The Gaussian self-similar process with parameter H and stationary increments such that $\mathrm{Var}(B_H(1)) = 1$ is called the fractional Brownian motion (fBm) with self-similarity (or Hurst) parameter H and is usually denoted by B_H. Note that for $H = 1/2$ the ordinary Brownian motion $B(\cdot)$ is obtained. $B_H(\cdot)$ has the following integral representation with respect to $B(\cdot)$

$$B_H(u) = \frac{1}{C_H} \int_{-\infty}^{+\infty} [(u-x)_+^{H-1/2} - (-x)_+^{H-1/2}]\, dB(x), \quad (12.3)$$

where the normalizing constant equals

$$C^2(H) = \left[\frac{1}{2H} + \int_0^\infty (1+v)^{H-1/2} - v^{H-1/2}\right]^2 dv.$$

Its increments $X_k = B_H(k) - B_H(k-1)$ discussed above are called the Fractional Gaussian Noise (FGN).

12.1.1 Hyperbolically decaying covariances

Typical regularity assumption on the behaviour of the covariance function of LRD sequence (X_t) is that

$$\gamma_X(k) = L(k)k^{-\alpha}, \tag{12.4}$$

$0 < \alpha < 1$ and $L(\cdot)$ is ultimately positive and slowly varying function at infinity in Karamata's sense i.e. such that $\lim_{x\to\infty} L(wx)/L(x) \to 1$ for any $w > 0$. Typical slowly varying function, besides an arbitrary constant obviously, is $L(x) = \log(x)$. It follows from the properties of slowly varying functions that similarily to the case when $L \equiv C$ for which $\sum_{k=1}^{\infty} \gamma_X(k)$ is infinite, this also holds for a general $L(\cdot)$, thus $(X_t)_{t\in\mathbb{N}}$ with autocovariance (12.4) is LRD. Note that due to asymptotic equivalence (12.2) the FGN satisfies (12.4). We will need also some assumptions on the structure of the process itself, such as Gaussian or linear subordination, to study its asymptotic behaviour, however some properties can be derived directly from (12.4). In particular this concerns behaviour of the spectral density at 0. To this end we define a slightly different and stronger concept of slow variability in Zygmund's sense which stipulates that the function $L(\cdot)$ is ultimately positive and for any $\delta > 0$ the functions $x^{\delta}L(x)$ and $x^{-\delta}L(x)$ are ultimately monotone. Then the following result holds (see e.g. Theorem 1.3 in Beran et al. (2013) and references therein)

Theorem 12.1.1 *(i) Assume that (12.4) is satisfied with* $0 < \alpha < 1$ *and* $L(\cdot)$ *slowly varying in Zygmund's sense. Then the spectral density* f *exists and*

$$f(\lambda) \sim L_f(\lambda)\lambda^{\alpha-1}, \quad \lambda \to 0, \tag{12.5}$$

where $L_f(\lambda) = L(\lambda^{-1})\Gamma(1-\alpha)\sin(\pi\alpha/2)$.
(ii) If

$$f(\lambda) = L_f(\lambda)\lambda^{-2d}, \quad 0 < \lambda < \pi, \tag{12.6}$$

where $0 < d < 1/2$ *and* $L_f(\lambda)$ *is slowly varying at 0 in Zygmund's sense then*

$$\gamma(k) \sim L_\gamma(k)|k|^{2d-1}, \quad |k| \to \infty,$$

where $L_\gamma(k) = 2L_f(k^{-1})\Gamma(1-2d)\sin\pi d$.

We discuss now several other models beside FGN of long-range dependent series with covariance structure (12.4). For an exhaustive treatment of the subject we refer to Beran et al. (2013).

12.1.2 Subordinated Gaussian processes

Let $(Z_t)_{t\in\mathbb{N}}$ be a stationary Gaussian sequence such that Z_ts are standard normal and (12.4) is satisfied. In order to make the model more flexible and allow for an arbitrary marginal distribution we consider so-called subordinated Gaussian process.

Definition 33 $(X_t)_{t\in\mathbb{N}}$ *defined as*

$$X_t = G(Z_t), \qquad t = 1, 2 \ldots, \tag{12.7}$$

where $G(\cdot) \in \mathcal{L}^2(\mathbb{R}, \varphi)$ *i.e.* $\int G^2(z)\varphi(z)\,dz < \infty$, $EG(Z_1) = 0$ *and* φ *is the standard normal density is called subordinated Gaussian process .*

$\mathcal{L}^2(\mathbb{R}, \varphi)$ is a $\mathcal{L}^2$ space with weight φ in which a scalar product is defined as $< F, G >= \int F(s)G(s)\varphi(s)\,ds$. The question when $(X_t)_{t\in\mathbb{N}}$ is also LRD is most conveniently answered by considering Fourier expansion of G in $\mathcal{L}^2(\mathbb{R}, \varphi)$ with respect to the Hermite polynomials:

$$G(z) = \sum_{k=1}^{\infty} \frac{J_k}{k!} H_k(x), \tag{12.8}$$

where

$$H_k(x) = (-1)^k e^{x^2/2} \frac{d^k}{dx^k}(e^{-x^2/2})$$

is k^{th} Hermite polynomial: $H_0(x) \equiv 1$, $H_1(x) = x$, $H_2(x) = x^2 - 1, \ldots$. Formula (12.8) follows from the fact that $(H_k)_{k=0}^{\infty}$ form an orthogonal basis in $\mathcal{L}^2(\mathbb{R}, \varphi)$ and moreover

$$\int H_k(z)H_l(z)\varphi(z)\,dz = I\{k = l\}k!.$$

Note that coefficient J_0 is missing in the expansion (12.8) due to $EG(Z_1) = 0$. Let $m = \min\{k : J_k \neq 0\}$ be the Hermite rank of the function G i.e. the index of the first non-zero term in the expansion (12.8). Obviously, the Hermite rank of H_k is k. The Hermite rank plays pivotal role in determining whether (X_t) is LRD. Namely, the following equality known as the Mercer formula holds

$$\mathrm{Cov}(H_k(Z_1), H_l(Z_2)) = k!\mathrm{Cov}(Z_1, Z_2)^k I\{k = l\}.$$

This is proved using the expansion of bivariate normal density with respect to products of Hermite polynomials $H_m(x)H_m(y)$. Note that it follows that $(H_k(Z_t))$ is long-range dependent sequence provided $k\alpha < 1$. This statement can be generalized to arbitrary G. Namely, the ensuing equality based on (12.8)

$$\mathrm{Cov}(G(Z_1), G(Z_2)) = \sum_{k=m}^{\infty} \frac{J_k^2}{k!} \mathrm{Cov}(Z_1, Z_2)^k \tag{12.9}$$

and properties of the slowly varying functions imply the following Lemma.

Lemma 12.1.2 *Assume that $m\alpha < 1$, where m is the Hermite rank of G. Then (i)*

$$\gamma_X(j) \sim \frac{J_m^2}{m!}\gamma_Z^m(j) \tag{12.10}$$

when $j \to \infty$ and (X_t) is LRD.
(ii) Moreover,

$$\mathrm{Var}(X_1 + \ldots + X_n) \sim \frac{2m!}{(1-m\alpha)(2-m\alpha)} L^m(n) n^{2-m\alpha} \tag{12.11}$$

Proof (i). Rewriting (12.9) we have

$$\gamma_X(j) = \sum_{k=m}^{\infty} \frac{J_k^2}{k!}\gamma_Z^k(j).$$

Now note that

$$\begin{aligned}\sum_{k=m}^{\infty} \frac{J_k^2}{k!}\gamma_Z^k(j) &= \frac{J_m^2}{m!}\gamma_Z^m(j) + \gamma_Z^{m+1}(j) \sum_{k=m+1}^{\infty} \frac{J_k^2}{k!}\gamma_Z^{k-m-1}(j) \\ &= \frac{J_m^2}{k!}\gamma_Z^m(j)(1+o(1)),\end{aligned} \tag{12.12}$$

as $\gamma_Z(j) \to 0$ when $j \to \infty$ and $\sum_{k=m+1}^{\infty} J_k^2 \gamma_Z^{k-m-1}(j)/k!$ is summable in view of summability of $\sum_k J_k^2/k!$ for j such that $|\gamma_Z(j)| < 1$.
Proof (ii). We asssume more generally that $\gamma_X(k) \sim L_1(k)k^{-\beta}$, where L_1 is ultimately positive slowly varying function and $0 < \beta < 1$. Observe that denoting $S_n = X_1 + \ldots + X_n$ and letting $L(-k) = L(k)$ we have

$$\begin{aligned}\mathrm{Var}(S_n) &= n \sum_{|k|<n} \Big(1 - \frac{|k|}{n}\Big)\gamma_X(k) \\ &\sim n \sum_{|k|<n, k\neq 0} L(k)|k|^{-\beta} - \sum_{|k|<n, k\neq 0} L(k)|k|^{-\beta+1} \\ &= L(n)n^{2-\beta}\Big[\frac{1}{n}\sum_{|k|<n, k\neq 0} \frac{L(k)}{L(n)}\Big(\frac{|k|}{n}\Big)^{-\beta} - \frac{1}{n}\sum_{|k|<n, k\neq 0} \frac{L(k)}{L(n)}\Big(\frac{|k|}{n}\Big)^{-\beta+1}\Big] \\ &\sim 2L(n)n^{2-\beta}\Big(\int_0^1 u^{-\beta}\,du - \int_0^1 u^{-\beta+1}\,du\Big) \\ &= \frac{2}{(1-\beta)(2-\beta)} L(n)n^{2-\beta}.\end{aligned} \tag{12.13}$$

The last equivalence above follows from the properties of slowly varying functions and definition of the integral. We skip its formal proof. Taking $\beta = m\alpha$ and $L_1(z) = L^m(z)J_m^2/m!$ we obtain the proof of (ii).
Observe that it follows from the derivation above that in the case of LRD that the sum of individual variances is negligible compared to the variance of the sum

$$\sum_{i=1}^{n} \mathrm{Var}(X_i) = o(\mathrm{Var}(S_n)).$$

Note also that for the FGN series defined above variance of the mean can be exactly calculated

$$\mathrm{Var}(n^{-1}\sum_{i=1}^{n} Y_i) = \mathrm{Var}(n^{-1}B_H(n)) = \sigma^2 n^{2H-2}.$$

It can also be proved that when $m\alpha > 1$ covariances of the gausssian subordinated series are summable i.e. $(X_t)_{t\in\mathbb{N}}$ is short-range dependent.
Let H be Hurst exponent defined by the equality $2H = 2 - m\alpha$ and note that for the case $m\alpha < 1$ considered in the lemma we have that $1 - 1/(2m) < H < 1$. Note that the lemma asserts that for LRD subordinated Gaussian series standard deviation of the partial sum S_n is, up to the slowly varying function, of order n^H and it increases more quickly than the rate $n^{1/2}$ of standard deviation of independent and weakly dependent variables. Note also that the behaviour of S_n is the same as the leading term of its expansion namely $\sum_{k=m}^{\infty}(J_m/m!)H_m(Z_i)$. We state now two fundamental results concerning asymptotic behaviour of partial sums of LRD series and empirical processes based on them.

Theorem 12.1.3 *Let $(X_t)_{t\in\mathbb{N}}$ be defined in (12.7) and $m\alpha < 1$. Let $L_m(n) = C_m L^m(n)$ and $C_m = 2/(1 - m\alpha)(2 - m\alpha)$. Then*

$$n^{-H} L_m^{-1/2}(n) S_n \xrightarrow[n\to\infty]{\mathcal{D}} \frac{J_m}{m!} Z_{m,H}(1), \tag{12.14}$$

where $Z_{m,H}(1)$ is the value at 1 of the Hermite process $Z_{m,H}(t)$. For $m = 1$ $Z_{m,H}(\cdot)$ is the fractional Brownian noise and $Z_m(1)$ is standard normal. For $m > 1$ $Z_m(1)$ is not Gaussian.

Hermite processes are defined e.g. in Beran et al. (2013), definition 3.24. There are two remarkable features of the result. The first one is normalization of the partial sum which is smaller than the usual $n^{-1/2}$ norming and the second is non-normality of the limit. This shows that the long-range dependent processes behave qualitatively and quantitatively differently from the weakly dependent ones. This has profound consequences for the estimators of parameters of such series and will be discussed later.
Wee now discuss the result for empirical process and show that it also essentially differs from the analogous result for the weakly dependent processses.
Let F be cumulative distribution of $G(Z_1)$ and denote by $m(x)$ the Hermite rank of $G_x(s) = I\{G(s) \leqslant x\} - F(x)$ i.e. the first nonzero term in the expansion

$$I\{G(s) \leqslant x\} - F(x) = \sum_{k=1}^{\infty} \frac{J_k(x)}{k!} H_k(s),$$

where $J_k(x) = E(I\{G(X) \leqslant x\}H_k(X))$. Moreover, define the Hermite rank m of the class of functions $\{I\{G(\cdot) \leqslant x\} - F(x)\}_{x\in\mathbb{R}}$ as $m = \inf_{s\in\mathbb{R}} m(x)$ i.e. the smallest possible Hermite rank for these functions. Let $F_n(x) = n^{-1}\sum_{i=1}^{n} I\{X_i \leqslant x\}$

be the empirical distribution function of $X_1, \ldots, X_n$.We are now in the position to state

Theorem 12.1.4 *Let $(X_t)_{t\in\mathbb{N}}$ be defined in (12.7) and $m\alpha < 1$, where m is the Hermite rank of the family $\{I\{G(\cdot) \leqslant x\} - F(x)\}_{s\in\mathbb{R}}$. Then in $\mathcal{D}[-\infty, \infty]$*

$$n^{1-H} L_m^{-1/2}(n)(F_n(\cdot) - F(\cdot)) \xrightarrow[n\to\infty]{\mathcal{D}} \frac{J_m(x)}{m!} Z_m(1), \tag{12.15}$$

where $L_m(\cdot)$ is defined in the previous theorem.

Note that in contrast to the case of independent and identically distributed observations, where the limit of $n^{1/2}(F_n(\cdot) - F(\cdot))$ is the Brownian bridge, here the limit turns out to be a degenerate process being a product of a single variable and the deterministic function. Moreover, note that the limit is zero for all points x for which the Hermite rank of the function $G_x(\cdot)$ is larger than m.

12.1.3 Subordinated linear processes

The similar construction to subordinated Gaussian processes is possible starting from LRD linear sequences. Namely let

$$\varepsilon_t = \sum_{i=0}^{\infty} a_i \eta_{t-i},$$

where $a_i = \tilde{L}(i) i^{-\beta}$, $1/2 < \beta < 1$ and $\tilde{L}(\cdot)$ is a function slowly varying at infinity. Moreover, assume that η_i are zero mean independent and identically distributed random variables such that $\mathrm{e}\eta_i^2 = 1$. Using properties of slowly varying functions again it can be proved that in such a case (ε_t) is LRD. Namely, the following result holds.

Lemma 12.1.5 *Assume that $1/2 < \beta < 1$. Then (i)*

$$\gamma_\varepsilon(k) \sim L(k) k^{-\alpha}, \tag{12.16}$$

where $\alpha = 2\beta - 1$ and $L(k) = C_\beta \tilde{L}^2(k)$, $C_\beta = \int_0^\infty (x + x^2)^{-\beta}\, dx$.
(ii) Moreover, for $S_n = \varepsilon_1 + \cdots + \varepsilon_n$ we have

$$\mathrm{Var}(S_n) \sim \frac{2}{(2-2\beta)(3-2\beta)} L^2(n) n^{3-2\beta}.$$

Note that the exponent of decay of $\gamma_\varepsilon(k)$ satisfies $0 < \alpha < 1$ thus (ε_t) is LRD.
Proof (i). The proof is similar to that of the previous lemma. Namely,

$$\begin{aligned}\gamma_X(k) &= \sum_{k=0}^{\infty} a_j a_{j+k} \sim \sum_{k=1}^{\infty} L(j) j^{-\beta} L(j+k)(j+k)^{-\beta} \\ &= L^2(k) k^{1-2\beta} \frac{1}{k} \sum_{j=1}^{\infty} \frac{L(j)}{L(k)} \Big(\frac{j}{k}\Big)^{-\beta} \Big(\frac{L(j+k)}{L(k)} (1 + \frac{j}{k}\Big)^{-\beta}\end{aligned}$$

$$= L^2(k)k^{1-2\beta}\int_0^\infty x^{-\beta}(1+x)^{-\beta}\,dx. \tag{12.17}$$

Again, the last equivalence follows from the properties of the slowly varying functions and the definition of the integral. Note that $\int_0^\infty x^{-\beta}(1+x)^{-\beta}\,dx$ exists as $1/2 < \beta$.
The proof of (ii) follows from the proof of (ii), Lemma 12.1.2.

Remark 12.1.6 *Note that in view of the above results the following relations hold between parameter of decay α of covariance function γ, β of coefficients for linear process a_i and the order $2d$ of the pole of the spectral density at 0. We have*

$$\alpha = 2\beta - 1, \quad \beta = 1 - d, \quad \alpha = 1 - 2d.$$

Moreover, Hurst parameter H satisfies $2H = 2 - \alpha$.

Important LRD linear process is fractionally differenced white noise denoted as FARIMA$(0, d, 0)$, where $0 < d < 1/2$. Recall that d-fold differencing with $d \in \mathbb{N}$ corresponds to the operator $(1 - B)^d$, where

$$\triangle^d = (1 - B)^d = \sum_{k=0}^{d} \binom{d}{k}(-B)^d.$$

We generalize this definition to arbitrary d replacing $\binom{d}{k}$ by $\Gamma(d+1)/(\Gamma(k+1)\Gamma(d-k+1))$ and letting

$$(1 - B)^d = \sum_{k=0}^{\infty}(-1)^k \frac{\Gamma(d+1)}{\Gamma(k+1)\Gamma(d-k+1)}B^k.$$

Consider the process FARIMA$(0, d, 0)$ which is the solution of the equation

$$(1 - B)^d X_t = \eta_t$$

and (η_t) is a strong white noise such that $E\eta_j^2 - 1$. We have $X_t = A(B)\eta_t$, with

$$A(z) = (1 - z)^{-d} = \sum_{j=0}^{\infty} a_j z^j = \sum_{j=0}^{\infty} \binom{-d}{j}(-1)^j z^j,$$

for $|z| < 1$. Using $x\Gamma(x) = \Gamma(x+1)$ it follows

$$a_j = (-1)^j \binom{-d}{j} = (-1)^j \frac{\Gamma(-d+1)}{\Gamma(j+1)\Gamma(-d-j+1)} = \frac{\Gamma(j+d)}{\Gamma(j+1)\Gamma(d)}.$$

Moreover, using Stirling's formula we obtain

$$a_j \sim \frac{1}{\Gamma(d)} j^{d-1}$$

and thus β in the definition of LRD linear process equals $\beta = 1 - d$ and is contained in the interval $(0, 1/2)$. Observe also that the spectral density of FARIMA$(0, d, 0)$ process exists and is given by

$$f_X(\lambda) = \frac{1}{2\pi}|1 - e^{-i\lambda}|^{-2d} \sim \frac{1}{2\pi}|\lambda|^{-2d}$$

(recall that $\mathrm{Var}(\eta_t) = 1$). This is common feature of LRD proceses in general, namely under some technical conditions their spectral densities exist and have a pole at 0 of order λ^{1-2H}.

Definition 34 *We define general FARIMA(p, d, q) process as a process such that its d-fold differencing yields ARMA(p, q) series. Thus FARIMA(p, d, q) process has representation*

$$X_t = \frac{\theta(B)}{\varphi(B)}(\varepsilon_t), \tag{12.18}$$

where (ε_t) is FARIMA$(0, d, 0)$ process.

The spectral density of FARIMA(p, d, q) satisfies

$$f_X(\lambda) = \frac{\sigma^2}{2\pi}\Big|\frac{\theta(e^{-i\lambda})}{\varphi(e^{-i\lambda})}\Big|^2|1 - e^{-i\lambda}|^{-2d} \sim \frac{\sigma^2}{2\pi}\Big|\frac{\theta(1)}{\varphi(1)}\Big|^2\lambda^{-2d}$$

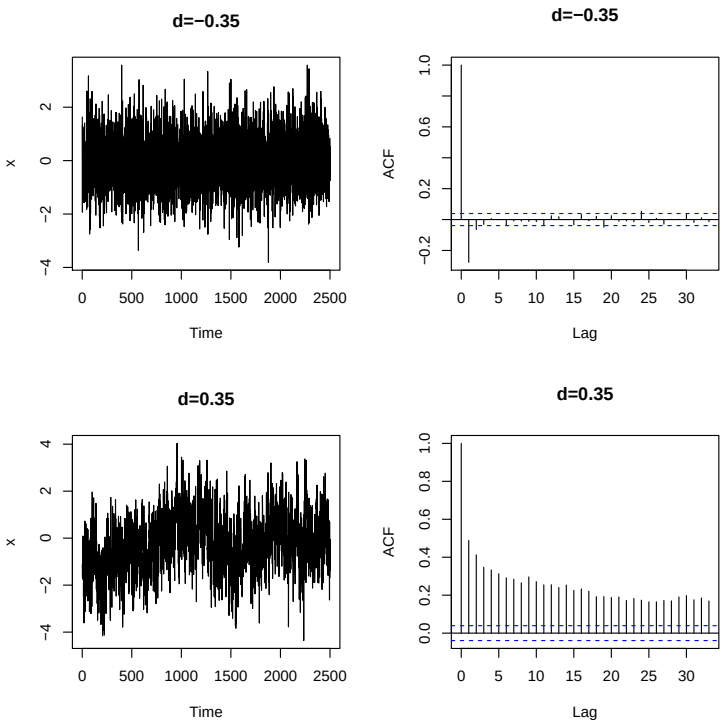
when $\lambda \to 0$.
Below we show sample paths and empirical ACF of FARIMA$(0, d, 0)$ series for $d = \pm 0.35$, that is of the process which after d-fold differentiation becomes white noise.
It is seen that sample paths and autocorrelations differ greatly. For $d = 0.35$ autocorrelations decay very slowly in contrast to the case $d = -0.35$. Sample path for $d = -0.35$ does not exhibit persisting local trends in contrast to $d = 0.35$. Note that in representation (12.18) FARIMA$(0, d, 0)$ process can be replaced by any LRD sequence, e.g. by FGN process, and we still obtain LRD series as the result. This yields a flexible method of obtaining new models for LRD. Generalization of this is subordination approach.

Definition 35 *We define a series subordinated to the linear process as*

$$X_t = G(\varepsilon_t),$$

where (ε_t) is the linear process defined above.

We now state the result analogous to Theorem 12.1.3. First we define the power rank of function G which is analogue of the Hermite rank of G which we used in connection with Gaussian subordinated processes.



The power rank of the function G with respect to ε_1, where G is such that $EG(\varepsilon_1) = 0$, is the smallest integer m such that m-th derivative $G_\infty^{(m)}(0) \neq 0$ where $G_\infty(x) = EG(\varepsilon_1 + x)$. Note that power rank depends on the function G and distribution of ε_1.

Theorem 12.1.7 *(Ho and Hsing (1980)) Assume that power rank m of G satisfies $m(2\beta - 1) < 1$, $E|\eta_1|^{4+\delta} < \infty$ and moreover*

$$\max_{r=1,\ldots,m} \sup_{y\in\mathbb{R}} |G_\infty^{(r)}(y)| < \infty.$$

Then

$$n^{-H} L_m^{-1/2}(n) S_n \xrightarrow[n\to\infty]{\mathcal{D}} G_\infty^{(m)}(0) Z_{m,H}(1), \tag{12.19}$$

where

$$L_m(\cdot) = \frac{2m!}{(1 - m(2\beta - 1))(2 - m(2\beta - 1))} L^m(\cdot),$$

$L(\cdot)$ is defined below (12.16), $2H = 2 - m(2\beta - 1)$ and where $Z_{m,H}(1)$ is the value at 1 of the Hermite process $Z_{m,H}(t)$.

Note that for the quadratic function centred at 0 $G(x) = x^2 - E\varepsilon_1^2$ it is easy to see that $G_\infty^{(1)}(0) = 2E\varepsilon_1 = 0$ and $G_\infty^{(2)}(0) = 2$. Thus the power rank of G is 2

and the limit in this case coincides with $2Z_{2,H}(1)$.
It is also interesting to observe that if k is an integer much smaller than n and we sample every k^{th} observation from the LRD sequence $X_1, \ldots, X_n$ with covariance function $\gamma_X(n) = L(n)n^{-\alpha}$ with $0 < \alpha < 1$ and consider the mean of such subsample

$$\bar{X}_n(k) = \frac{1}{[n/k]} \sum_{i=1}^{[n/k]} X_{ik} \tag{12.20}$$

then, asymptotically, its variance does not depend on k in sharp contrast to independent case. Namely,

$$\operatorname{Var}(\bar{X}_n(k)) \sim C_\alpha L(n) n^{-\alpha}, \tag{12.21}$$

where $C_\alpha = 2[(1-\alpha)(2-\alpha)]^{-1}$. This follows from (12.13) after noticing that the covariance function for $X_k, X_{2k}, \ldots, X_{[n/k]k}$ is $\gamma_k(i) = \tilde{L}(i)i^{-\alpha}$, where $\tilde{L}(i) = L(ik)k^{-\alpha}$.
On the more sombre note observe that the results above indicate that a construction of a confidence interval for the mean is much more complicated for LRD data than for independent data. However, the main difficulty is not dependence of the asymptotic variance on the unknown parameters which may be estimated, but the fact that the asymptotic distribution of the mean may vary for the given value of the Hurst parameter H. Namely, consider empirical mean of $(H_m^m(Z_i))_{i=1}^n$, where $(Z_i^{(m)})$ is the Gaussian sequence with decay of covariance function equal α/m for $0 < \alpha < 1$. Then, the Hurst parameter for subordinated Gaussian process $(H_m^m(Z_i))$ with $G = H_m^m$ is α regardless of m but e.g. for $m = 1, 2$ the limit of the mean is proportional to $Z_1(1)$ and $Z_2(1)$ respectively which is Gaussian in the first case and differs from Gaussian in the second. This is unresolved problem and no definitive answer to this is known, although some promising methods based on resampling exist (cf. Hall et al. (1998)).

12.2 Estimation of long-range dependence parameter

We discuss shortly the problem of estimation of long-range dependence parameter. Usually the main parameter of interest is the Hurst parameter i.e. H such that the partial sums of the sample path S_n behave like n^{2H}, possibly up to some slowly varying function. This immediately yields the first method of estimation of H, namely the estimator based on the variance plot.
Variance plot Let $k \in \mathbb{N}$ such that $2 \leqslant k \leqslant n/2$ and consider m_k subsequences of length k of $X_1, \ldots, X_n$. These can either be built as disjoint $[n/k]$ subsequences or partially overlapping $n-k$ subsequences with varying first observation. Let $\bar{X}_1(k), \ldots, \bar{X}_{m_k}(k)$ be the sample means of these subseries and denote by $s^2(k)$ their sample variances. Variance plot is the scatterplot of $\log s^2(k)$ against $\log k$ for $k = 2, \ldots, [n/2]$. For long-range dependent sequences the least squares regression line fitted to the plot has a slope approximating $2H - 2 > -1$. Thus

quick and dirty test of occurrence of LRD would be to draw a line with a slope -1 and see whether the slope of the regression line is less steep than this reference line. Alternatively, we can create a correlogram, that is a plot of $\log \hat{\rho}(k)$ against k and fit regression line to this plot. Again its slope should be close to $2H-2$. Another statistic which was used to discover long-range dependent phenomenon for Nile data is the rescaled range statistic R/S defined in (9.7).

The R/S estimator Let $Q(t,k) = R(t,k)/S(t,k)$ be R/S statistic based on $X_t, \ldots, X_{t+k-1}$ observations and consider the scatter plot of $n-k$ values of $\log Q(\cdot,k)$ against $\log k$. Then it is known (Mandelbrot (1991)) that for weakly dependent ergodic data such that the normalized partial sums $t^{-1/2}\sum_{s=1}^{t} X_s$ converge to the Brownian motion we have that $k^{-1/2}Q(t,k) \to \xi$, where ξ is a nondegenerate random variable whereas for LRD ergodic data when $t^{-H}\sum_{s=1}^{t} X_s$ converge to the fractional Brownian motion the correct normalization to obtain nondegenerate limit in distribution is k^{-H}. Thus the least squres regression line fitted to the constructed R/S plot will have approximate slope H. Lo (1991) introduced a modification of R/S statistic which used different estimator of the variance of the cumulative sum. Drawbacks of the procedures above is that they are not robust against the departures of stationarity, in particular against slowly decaying trends, that is simple nonstationary processes consisting of a trend contaminated by a white noise can be confused with stationary ones exhibiting LRD (cf Bhattacharya et al. (1983)). Moreover, the two described heuristic procedures rely on asymptotic behaviour of the involved statistics and this means that the initial part of the respective plot under consideration should be discarded. The corresponding cut-off point is usually chosen based on visual inspection but this is subjective and, what is even more significant, its choice has large impact on the value of the estimator.

Estimation of the Hurst parameter in the spectral domain The other possibility is to exploit behaviour of the spectral density of LRD sequences at 0, namely that

$$f(\lambda) \sim c_f |\lambda|^{1-2H}, \tag{12.22}$$

or more generally, similarity in (12.22) is up to a slowly varying function. We plot $I(\lambda_{j,n})$ against $\lambda_{j,n} = 2\pi j/n$ for $l \leqslant j \leqslant m$, where $l, m \in \mathbb{N}$ are parameters of the procedure and let $\hat{H} = (1-\hat{\beta}_1)/2$, where $\hat{\beta}_1$ is the slope of LS line fitted to such data. This is known as Geweke-Porter-Hudak (GPH) or log-periodogram regression estimator of H. Then the following result due to Robinson (1995) holds

Theorem 12.2.1 *Assume that (12.22) holds with $1/2 < H < 1$ and that $m, l \to \infty$ are such that $m = o(n^{4/5})$, $\log n = o(m^{1/2})$, $l = o(m)$ and $m^{1/2}\log m = o(l)$. Then under regularity conditions on f we have*

$$m^{1/2}(\hat{H} - H) \to N(0, \frac{\pi^2}{24}).$$

Note that the norming is $m^{1/2}$ and not $n^{1/2}$ thus one can expect considerable variability of $\hat{H}$. Choice of boundaries m and l is highly nontrivial and may influ-

ence the value of the estimator. We refer to Beran (1994) for a short discussion of this problem.
We present approximation of the maximum likelihood estimator based on (8.15) and Whittle's approximation to it. Namely, Whittle's approximation consists in replacing $\mathbf{\Gamma}_n^{-1}(\theta)$ by

$$W_n(\theta) = \big[\frac{1}{(2\pi)^2}\int_{-\pi}^{\pi} e^{i(r-s)}\lambda \frac{1}{f_X(\lambda,\theta)}\,d\lambda\big]_{r,s=1,\ldots,n},$$

where $f_X(\lambda,\theta)$ is a postulated parametric spectral density and defining Whittle's estimate of θ as the minimizer of the respective quadratic form

$$\hat{\theta}_W = \operatorname{argmin} \mathbf{x}' W_n(\theta)\mathbf{x}.$$

W_n is an asymptotic inverse of $\mathbf{\Gamma}_n$ in the sense defined in Beran (1994), lemma 5.3.
Consider the situation when θ is one dimensional and equal to the decay parameter of the spectral density at 0, namely $f_X(\lambda) \sim c_f|\lambda|^{-2d}$ when $\lambda \to 0$. Then we have (cf. Giraitis and Surgailis (1990))

Theorem 12.2.2 *Assume that $(X_t)_{t\in\mathbb{Z}}$ is a linear process with a spectral density $f_X(\lambda) \sim c_f|\lambda|^{-2d}$ when $\lambda \to 0$ for $-1/2 < d < 1/2$ and $\theta = d$. Then under appropriate regularity conditions we have*

$$n^{1/2}(\hat{d}_W - d) \to N(0, 4\pi V^{-1}),$$

where $V = \int_{-\pi}^{\pi}(f_X'(\lambda)/f_X(\lambda))^2\,d\lambda$.

Note that the integrand is the square of the logarithmic derivative of f_X. An amazing fact about this result is that even in the case of LRD ($0 < d < 1/2$) we have that normalizing factor is $n^{1/2}$ and the limit is normal. We also note that in some important cases the limiting variance does not depend on LRD parameter. Indeed, note that for FARIMA$(0,d,0)$ series we have

$$f_X(\lambda) = \frac{\sigma^2}{2\pi}|1-e^{-i\lambda}|^2 = \frac{\sigma^2}{2\pi}(1-2\cos\lambda)^{-d}$$

and it easily follows that its logarithmic derivative equals $-\log(2-2\cos\lambda)$ and the limiting variance $V = \pi^2/6$.

12.3 Fixed-design regression

We discuss now one specific problem of nonparametric estimation which nicely shows how different the inference for LRD data is in comparison with independent or weakly dependent case. We also indicate some ways how to account for this problem. Namely, we consider in greater detail the problem of nonparametric estimation of regression function in the fixed-design regression model (FDR)

with LRD errors. FDR model with uniform design stipulates that the observations are given as

$$Y_{i,n} = g(i/n) + \varepsilon_{i,n}, \quad i = 1, \ldots, n, \tag{12.23}$$

where $g : [0,1] \to \mathbb{R}$ is an unknown function to be estimated using $Y_{n,1}, \ldots, Y_{n,n}$. Triangular array $(\varepsilon_{i,n})$ is LRD in the sense that for each n sequence $(\varepsilon_{i,n})_{i=1}^n$ is stationary with zero mean and covariance function $r(\cdot)$ which does not depend on n and such that $\sum r(h)$ is infinite. In the following we will suppress in the notation dependence of $Y_{i,n}$ and $\varepsilon_{i,n}$ on n. We will only discuss one of possible regression estimators, namely Priestley-Chao estimator defined as

$$\hat{g}_n(x) = \frac{1}{nb_n} \sum_{i=1}^{n} K\big(\frac{x - i/n}{b_n}\big) Y_i \tag{12.24}$$

Although other, more sophisticated estimators, as e.g. local linear smoothers, are frequently considered, simplicity of Priestley-Chao estimator makes it easier to show how LRD errors influence inference here. Behaviour of MSE of $\hat{g}_n(x)$ was derived by Hall and Hart (1990) who proved

Theorem 12.3.1 *Assume that* $r_\varepsilon(k) \sim Ck^{-\alpha}$ *when* $k \to \infty$ *for some* $C > 0$ *and* $0 < \alpha \leqslant 1$. *Moreover,* $\sup_{0<x<1} |g^{(i)}(x)| < \infty$ *for* $i = 0, 1, 2$. *Then for any* $\delta > 0$ *we have*

$$\begin{aligned} MSE(\hat{g}_n(x)) &= \frac{C}{(nb_n)^\alpha} \int\int |x - y|^{-\alpha} K(x) K(y)\, dx\, dy \\ &\quad + \frac{b_n^4}{4} \Big(\int s^2 K(s)\, ds \Big)^2 g''^2(s) + o((nb_n)^{-\alpha} + b_n^4), \end{aligned} \tag{12.25}$$

uniformly in $x \in (\delta, 1 - \delta)$.

The first term in (12.25) corresponds to the variance and the second to the squared bias of $\hat{g}_n(x)$. We readily calculate the value of the minimizer of the two main terms in the decomposition of MSE which yields

$$b_n = (\alpha D_1 / D_2)^{1/(4+\alpha)} n^{-\alpha/(4+\alpha)},$$

where $D_1(\alpha) = C \int\int |x - y|^{-\alpha} K(x) K(y)\, dx\, dy$ and $D_2 = \big(\int s^2 K(s)\, ds\big)^2 g''^2(s)$. This shows that asymptotically optimal bandwith for estimating $g(x)$ depends on parameter α of hyperbolic decay which is unknown and this complicates greatly using its empirical version as plug-in versions exhibit significant variability. It is also known that in the case of one-sided linear process with covariance function $\gamma_\varepsilon(k) = L(k)k^{-\alpha}$ we have that $\hat{g}_n(x)$ is asymptotically normal when the normalization $a_n^* = (nb_n)^\alpha / L^{1/2}(nb_n)$ is used.

Observe that the design points are consecutively sampled from the uniform grid and thus the errors related to points which are close will be highly correlated. This leads to a natural question what happens when the design is still uniform but random. This question is partially answered by Csörgő and Mielniczuk (2000). Namely, consider the random-design regression model (RDR)

$$Y_{i,n} = g(X_i) + \varepsilon_{i,n}, \quad i = 1, \ldots, n, \tag{12.26}$$

where X_i are independent uniformly distributed on [0,1] and also independent from $\varepsilon_{i,n}$. Define Priestley-Chao estimator for this model analogously to (12.24) by

$$\hat{g}_n(x) = \frac{1}{nb_n} \sum_{i=1}^{n} K\big(\frac{x - X_i}{b_n}\big) Y_i. \tag{12.27}$$

Csörgő and Mielniczuk (2000) have shown that for one-sided LRD moving average (ε_i) we have

$$\min\big((nb_n)^{1/2}, \frac{n^{\alpha/2}}{L^{1/2}(n)}\big)\big(\hat{g}_n(x) - g(x)\big) \to \rho(x)Z, \tag{12.28}$$

where Z has the standard normal distribution and $\rho(x)$ is a deterministic function which form depends on which norming above actually prevails. Note also that the norming factor in (12.27) is *larger* than a_n^* suggesting that $\hat{g}_n(x)$ for RDR design is less variable than in the case of FDR design. This can be intuitively justified by seeing that $\hat{g}_n(x)$ is based on observations Y_i corresponding to observations X_i or i/n which fall into small neighborhood of x, where for simplicity we assume that K is compactly supported. In the case of FDR these observations are a block of consecutive observations and thus are strongly dependent. However, in the case of RDR the observations falling in the vicinity of x have random indices and thus their mutual dependence is much smaller. This leads to idea of using randomization for the FDR design. Namely, we define $\sigma = \sigma_n$ to be random permutation of $\{1, 2, \ldots, n\}$ and consider the randomized fixed-design regression model (RFDR) as

$$Y_{i,n} = g\Big(\frac{\sigma_n(i)}{n}\Big) + \varepsilon_{i,n}, \quad i = 1, \ldots, n, \tag{12.29}$$

which corresponds to the situation when in the i^{th} step we randomly sample the design point $\sigma_n(i)/n$ and sample the i^{th} observation Y_i there. We stress that the model (12.29) is adequate if the dependence between errors is due to the length of time interval which elapsed between respective observations. Obviously, in such a case we have the following form of Priestley-Chao estimator

$$\hat{g}_n(x) = \frac{1}{nb_n} \sum_{i=1}^{n} K\big(\frac{x - \sigma(i)/n}{b_n}\big) Y_i. \tag{12.30}$$

To see that randomization helps consider momentarily simple linear regression model $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1, \ldots, n$, where $x_i = i/n$ and errors are positively correlated i.e $r(i) > 0$ for $i \in \mathbb{N}$. Let $\hat{\beta}_1$ be LS estimator of the slope. Now we randomise the design points in which we take observations and we obtain the observations $Y_i = \beta_0 + \beta_1 x_{\sigma(i)} + \varepsilon_i$. Let $\tilde{\beta}_1$ be the slope of the LS line

fitted to these observations. Thus we have that $\hat{\beta}_1 = \sum_{i=1}^n Y_i(x_i - \bar{x})/S$ where $S = \sum_{i=1}^n (x_i - \bar{x})^2$ and

$$\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma_\varepsilon^2}{S} + \frac{\sum_{i \neq j} r(|i-j|)(x_i - \bar{x})(x_j - \bar{x})}{S^2},$$

where $\sigma_\varepsilon^2 = \gamma(0)$, whereas

$$\mathrm{Var}(\tilde{\beta}_1) = \frac{\sigma_\varepsilon^2}{S} + \frac{\sum_{i \neq j} \bar{r}(|i-j|)(x_i - \bar{x})(x_j - \bar{x})}{S^2},$$

and $\bar{r}(i-j) = \bar{r} = (n(n-1))^{-1} \sum_{k \neq l} r(|k-l|)$.
In the case of positively correlated errors we have that $\bar{r} > 0$ and thus as $\sum_{i \neq j} \bar{r}(|i-j|)(x_i - \bar{x})(x_j - \bar{x}) = -S < 0$. Thus

$$\mathrm{Var}(\tilde{\beta}_1) < \frac{\sigma_\varepsilon^2}{S} \sim n^{-1},$$

where the equivalence is due to property that $S \sim n/12$ in probability. However, in the case when spectral density f of ε_i satisfies $f(\lambda) \sim \lambda^{\alpha-1}$ for $\lambda \to 0$, Yajima (1988) proved that $n^\alpha \mathrm{Var}(\hat{\beta}_1) \to c > 0$ thus $\mathrm{Var}(\hat{\beta}_1)$ is of order $n^{-\alpha}$ and thus larger than n^{-1}.
MSE analysis of $\hat{g}_n(x)$ is based on the decomposition

$$\begin{aligned} \hat{g}_n(x) &= \frac{1}{n} \sum_{i=1}^n K_b\big(x - \frac{\sigma(i)}{n}\big) g\big(\frac{\sigma(i)}{n}\big) + \sum_{i=1}^n K_b\big(x - \frac{\sigma(i)}{n}\big)\varepsilon_i \\ &= \frac{1}{nb} \sum_{i=1}^n k_i(x) g\big(\frac{i}{n}\big) + \frac{1}{nb} \sum_{i=1}^n k_i(x) \varepsilon_{\sigma^{-1}(i)}, \end{aligned} \tag{12.31}$$

where $K_b(x) = b^{-1}K(x/b)$ and $k_i(x) = K((x - i/n)/b)$. Thus the random component of $\hat{g}_n(x)$ is a weighted sum of $\bar{\varepsilon}_{in} = \varepsilon_{\sigma^{-1}}(i)$ and it is easy to see that $\bar{\varepsilon}_{in}$ are exchangeable random variables such that

$$\mathrm{Cov}(\bar{\varepsilon}_{in}, \bar{\varepsilon}_{jn}) \sim \frac{2L(n)n^{-\alpha}}{(1-\alpha)(2-\alpha)}$$

for $i \neq j$. This clearly indicates why randomization in case of estimation of *local* parameters is beneficial. Namely, covariance $L(h)h^{-\alpha}$ of Y_i and Y_{i+h} is replaced by $CL(n)n^{-\alpha}$ which for h much smaller than n (and observations corresponding to such h are dominant for local estimators) is significantly smaller. Thus by an easy check we obtain the analogue of Theorem 12.3.1.

Theorem 12.3.2 *Assume that conditions of Theorem 12.3.1 are satisfied. Then*

$$\begin{aligned} MSE(\hat{g}_n(x)) &= \frac{C_\alpha}{n^\alpha} \int\int |x-y|^{-\alpha} K(x)K(y)\,dx\,dy + \frac{1}{nb_n}\sigma_\varepsilon^2 \int K^2(s)\,ds \\ &= +\frac{b_n^4}{4}\big(\int s^2 K(s)\,ds\big)^2 g''^2(s) + o(n^{-\alpha} + (nb_n)^{-1} + b_n^4), \end{aligned} \tag{12.32}$$

uniformly in $x \in (\delta, 1-\delta)$, *where* $C_\alpha = 2C/(1-\alpha)(2-\alpha)$.

It is thus seen that by using randomized design we have diminished the variance of Priestley-Chao estimator. Note also that for small bandwidths the term $(nb_n)^{-1}$ is larger than $n^{-\alpha}$ and the MSE behaves as in the case of independent data. For large bandwidths, however, the term $n^{-\alpha}$ which does not depend on b_n prevails. This is example of co-called smoothing dichotomy, for other examples see e.g. Csörgő and Mielniczuk (1995) and Wu and Mielniczuk (2002).

12.4 Problems

1. Check the asymptotic validity of (12.21) using the discussion in the text.
2. Prove Theorem 12.1.7 for $G(s) = s$ using Ibragimov-Linnik theorem.
3. Consider a linear process $X_t = \sum_{i=0}^{\infty} a_i \eta_{t-i}$ with zero mean square integrable innovations and such that $(a_i) \in \ell^2$ and $\sum_{i=0}^{\infty} a_i \neq 0$. Show that (X_t) is short-range dependent and that $n^{-1/2}(X_1 + \cdots + X_n) \to N(0, v)$ in distribution, where $v = \sum_{k=-\infty}^{\infty} \gamma_X(k)$.
4. Let Z_1, Z_2 be the standard normal random variables. Using the diagram formula (Theorem 2.1.2) show that

$$\mathrm{Cov}(Z_1^2, Z_2^2) = 2\mathrm{Cov}^2(Z_1, Z_2)$$

i.e. Mercer formula for $m = 2$.
5. Show the conclusion of Problem 4 for the LRD linear process with innovations η_i such that $E\eta_i^4 < \infty$.
6. Show that if $L(\cdot)$ is slowly varying in Karamata's sense then sign of $L(\cdot)$ is ultimately constant. Hint: reasoning by contradiction take $p, q > t_0$ such that $L(p) > 0$ and $L(q) < 0$ where t_0 is such that $L(2t)/L(t) > 0$ when $t \geqslant t_0$. Show that then for any $n \in \mathbb{N}$ $L(2^n p) > 0$ and $L(2^n q) < 0$.

References

H. Akaike. Statistical predictor identification. *Annals of the Institute for Statistical Mathematics*, 22:203–217, 1970.

D. Andrews. Nonstrong mixing autoregressive processes. *Journal of Applied Probability*, 21:930–934, 1984.

B. Baxter. An asymptotic result for the finite predictor. *Math. Scand.*, 10: 137–144, 1960.

J. Beran. *Statistics for Long-Memory Processes.* Chapman and Hall, New York, 1994.

J. Beran, Y. Feng, S. Ghosh, and R. Kulik. *Long-Memory Processes.* Springer, New York, 2013.

R. Bhattacharya, V. Gupta, and E. Waymire. The Hurst effect under trends. *Journal of Applied Probability*, 20:649–662, 1983.

P. Billingsley. *Convergence of Probability Measures.* Wiley, New York, 1968.

G. Box and D. Pierce. Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *J. Amer. Statist. Assoc.*, 65: 1509–1526, 1970.

G. Box, G. Jenkins, and G. Reinsel. *Time Series Analysis: Forecasting and Control.* Wiley, New York, 2008.

R. Bradley. Basic properties of strong mixing conditions. a survey and some open questions. *Probability Surveys*, 2:107–144, 2005.

P. Brockwell and R. Davis. *Time Series: Theory and Methods.* Springer, New York, 1991.

B. Brown. Martingale central limit theorems. *Ann. Math. Stat.*, 42:59–66, 1971.

C. Chatfield. *The Analysis of Time Series: the Introduction.* Chapman and Hall, London, 2003.

R. Coqburn. Asymptotic properties of stationary sequences. *Univ. Calif. Publ. Statist.*, 3:99–146, 1960.

T. Cover and J. Thomas. *Elements of Information Theory.* Wiley, New York, 2006.

J. Crutchfield and D. Feldman. Regularities unseen, randomness observed: levels of entropy convergence. *Chaos*, 3:25–54, 2003.

J. Cryer and K.S. Chan. *Time Series Analysis.* Wiley, New York, 2008.

S. Csörgő and J. Mielniczuk. The smoothing dichotomy in random-design regression with long-memory errors based on moving averages. *Statistica Sinica*, 10:771–787, 2000.

S. Csörgő and J. Mielniczuk. Density estimation under long-range dependence. *Annals of Statistics*, 23:990–999, 1995.

J. Dedecker, P. Doukhan, G. Lang, R. Leon, S. Louhichi, and C. Prieur. *Weak Dependence: with Examples and Applications*. Springer, New York, 2007.

P. Diaconis and D. Freedman. Iterated random functions. *SIAM Review*, 41: 41–76, 1999.

Ł. Dębowski. *Information Theory and Statistics*. Institute of Computer Science, PAS, Warsaw, 2013.

Ł. Dębowski. On processes with summable partial autocorrelations. *Statistics and Probability Letters*, 77:752–759, 2007.

J. Doob. *Stochastic Processes*. Wiley, New York, 1953.

P. Doukhan and S. Louhichi. A new weak dependence condition with application to moment inequalities. *Stochastic Processes and Their Applications*, 84:313–342, 1999.

R. Engle. Autoregressive conditional heteroscedasticity with estimates of U.K. inflation. *Econometrica*, 50:987–1008, 1982.

J. Fan and Q. Yao. *Nonlinear Time Series: Parametric and Nonparametric Methods*. Springer, New York, 2003.

W. Feller. *An Introduction to Probability Theory and its Applications. Vol. 2*. Springer, New York, 1971.

C. Francq and J.M. Zakoian. *GARCH Models*. Wiley, New York, 2010.

L. Giraitis and P. Robinson. Whittle estimation of ARCH models. *Econometric Theory*, 17:608–623, 2001.

L. Giraitis and D. Surgailis. A central limit theorem for quadratic forms in strongly dependent linear variables and its application to asymptotic normality of Whittle's estimate. *Probability Theory and Related Fields*, 86:87–104, 1990.

M. Gordin. The central limit theorem for stationary processes. *Dokl. Akad. Nauk SSSR*, 239:392–393, 1969.

R. Gray. Toeplitz and circulant matrices: a review. *Foundations and Trends in Communications and Information Theory*, 2:155–239, 2006.

U. Grenander and Szegő. *Toeplitz Forms and Their Applications*. University of California Press, Berkeley, 1958.

P. Hall and J. Hart. Nonparametric regression with long-range dependence. *Stochatic Proceess. Appl.*, 36:339–351, 1990.

P. Hall, B.-Y. Jing, and S. Lahiri. On the sampling window method for long-range dependent data. *Statistica Sinica*, 8:1189–1204, 1998.

J. Hamilton. *Time Series Analysis*. Princeton, Princeton, 1994.

E. Hannan. *Time Series Analysis*. Methuen, London, 1967.

E. Hannan. The central limit theorem for time series regression. *Stochastic Process. Appl.*, 36:339–351, 1979.

E. Hannan. The estimation of the order of an ARMA process. *Annals of Probability*, 25:1636–1669, 1997.

H.C. Ho and T. Hsing. Limit theorems for functionals of moving averages. *Ann. Statist.*, 8:1071–1081, 1980.

H. Hurd and A. Miamee. *Periodically Correlated Random Sequences.* New York, New York, 2007.

C. Hurvich and C. Tsai. Regression and time series model selection in small samples. *Biometrika*, 76:297–307, 1989.

I. Ibragimow and Yu. Linnik. *Independent and Stationary Sequences of Random Variables.* Wolters-Noordhoff, Groningen, 1971.

I. Ibragimow and Yu. Rozanov. *Gaussian Random Processes.* Springer, New York, 1978.

A. Inoue. AR and MA representations of partial autocorrelation functions, with applications. *Probability Theory and Related Fields*, 140:523–551, 2008.

A. Kolmogorov and Yu. Rozanov. On strong mixing conditions for stationary gaussian sequences. *Theor.Probab.Appl*, 5:204–208, 1960.

S. Lahiri. *Resampling Methods for Dependent Data.* Springer, New York, 2003.

J. Lindsay. *Statistical Analysis of Stochastic Processes in Time.* Cambridge University Press, Cambridge, 2006.

A. Lo. Long-term memory in stock market prices. *Econometrica*, 59:1279–1313, 1991.

H. Lütkepohl. *New Introduction to Multiple Time Series Analysis.* Wiley, New York, 2007.

S. Makridakis, S. Wheelwright, and R. Hyndman. *Forecasting. Methods and Applications.* Wiley, New York, 1998.

B. Mandelbrot. Limit-theorems on the self-normalized range for weakly and strongly dependent processes. *Zeitschrift fuer Wahrscheinlichkeitstheorie und Verwandene Gebiete*, 59:271–285, 1991.

A. McQuarrie and C. Tsai. *Regression and Time Series Model Selection.* World Scientific, Singapore, 1998.

D. Nelson. Conditional heroscedasticity in aasset pricing: A new approach. *Econometrica*, 59:347–370, 1990.

M. Pourahmadi. *Foundations of Time Series Analysis and Prediction Theory.* Wiley, New York, 2001.

P. Robinson. Log-periodogram regression of time series with long-range dependence. *Annals of Statistics*, 23:1048–1072, 1995.

M. Rosenblatt. *Stationary sequences and random fields.* Birkhauser, Boston, 1985.

D. Ruppert. *Statistics and Data Analysis for Financial Engineering.* Springer, New York, 2011.

G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6: 461–464, 1978.

R. Shumway and D. Stoffer. *Time Series Analysis and Its Applications.* Springer, New York, 2006.

M. Taniguchi and Y. Kakizawa. *Asymptotic Theory of Statistical Inference for Time Series.* Springer, New York, 2000.

S. Taylor. *Asset Price Dynamics, Volatility and Prediction*. Princeton University Press, Princeton, 2005.

R. Tsay. *Analysis of Financial Time Sereies*. Wiley, New York, 2010.

W. Willinger, M. Taqqu, W. Leland, and D. Wilson. Self-similarity in high-speed packet traffic: analysis and modeling of ethernet traffic measurements. *Statistical Science*, 10:67–85, 1995.

S. Wood. *Generalized Additive Models*. Chapman and Hall, Boca Raton, 2006.

W.B. Wu. Nonlinear system theory: another look at dependence. *Proceedings of the National Academy of Sciences*, 102:14150–14154, 2005.

W.B. Wu and J. Mielniczuk. Kernel density estimation for linear process. *Annals of Statistics*, 30:1441–1459, 2002.

Y. Yajima. On estimation of a regresssion model with long-memory errors. *Annals of Statistics*, 16:633–655, 1988.

Index

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

The Project is co-financed by the European Union from resources of the European Social Fund

ISBN 978-83-63159-16-0
e-ISBN 978-83-63159-17-7